MAS368: Decision making and Bayes linear methods

Module leader: Dr Darren J Wilkinson

email: d.j.wilkinson@ncl.ac.uk

Office: M515, **Phone:** 7320

Every day, we are all faced with the problem of having to make decisions. Suppose that on your way to university you pass two stottie shops, which sell similarly priced stotties. You wish to buy yourself a cheese stottie for lunch from one of the two shops. You know that the second shop sells nicer stotties, but is sometimes sold out by the time you are going past. Should you stop at the first shop and make do with an inferior stottie, or go to the second and risk ending up with nothing? How can you quantify the non-monetary costs, benefits and uncertainties in the above problem? How can you decide on the correct decision for you? The first part of the course is concerned with problems such as these.

Suppose now that you decided to go to the second shop, and managed to get a superior stottie, but that whilst still better than the stotties from the first shop, was not quite as good as you had expected. How can you now revise your beliefs about the quality of stotties from the second shop tomorrow? The second part of the course will be concerned with such inferential problems, and how to use a computing package, $[B/D]$, for the adjustment of limited belief specifications.

Contents

Chapter 1

Decision Making

1.1 Introduction

Making decisions when there is no uncertainty is relatively straight forward — simply write down all of your possible options, and decide which you like the best! Making decisions becomes more interesting when there are uncertainties — do you go for a "safe" option (not much uncertainty), or a "risky" option (lots of uncertainty), supposing that the "risky" option could be better or worse than the safe option, but you don't know which? Before we can hope to tackle such questions we need a language for uncertainty — probability.

1.2 Subjective probability

1.2.1 Introduction

What is probability? What is the meaning of the following statement?

The probability that it rains tomorrow is 0.25

The classical interpretation of probability is based on the concept of equally likely events, but there is no obvious partition of tomorrow's weather into equally likely events. The frequentist interpretation of probability is based on the concept of long-run relative frequencies, but tomorrow will only happen once. The statement cannot be given a classical or frequentist interpretation, but it must mean something — we intuitively know what it means.

The subjective interpretation of probability is based around the notion of *fair betting odds*, which provides a natural translation of the statement into the following.

The fair betting odds for rain tomorrow are 3 to 1 against

The problem then becomes: who decides which betting odds are fair? The answer (though some don't like it) is *You*.

1.2.2 Assessment of probability

How do we assess probabilities under the subjective interpretation?

Definition

> *The probability, p, of an event, E, is the price £p you consider fair for the bet which pays you £1 if E occurs, and £0 if it does not.*

For example, if you consider £0.25 to be the fair price for a gamble which pays £1 if it rains tomorrow, then *your* probability for rain tomorrow is 0.25. There are three obvious questions one might ask about such a definition of probability.

1. Does such a definition lead to the usual axioms of probability, so that we can treat subjective probability mathematically, in the same way as we usually do?

2. Can people have any probabilities they want for collections of events?

3. Can we ensure that people make honest probabilistic assessments?

The answers are *Yes, No, Yes* respectively, and the concept of *coherence* is the key to understanding why.

1.2.3 Coherence

The principle of coherence (also known as *rationality*) simply states that an individual should never do anything which leads to certain loss. Consequently, when assessing probabilities, one should envisage the following scenario: When you announce your probability for an event, an adversary gets to choose whether or not you are required to *place* or *host* a bet for a £1 prize at the odds you quote as fair.

Thus, if you quote a probability higher than the "true" probability of the event, and your adversary spots this, he will require you to *place* the bet, and you will be paying more than you should for the gamble. If you declare a probability smaller than the "true" value, your adversary will pay you the amount you declare, and require you to pay him £1 if the event occurs. Under such a scenario, coherence requires that you quote honest probabilities. This answers question 3, so what about the probability axioms?

One of the axioms states that for any event E, $p \equiv \mathrm{P}(E) \leq 1$. What if you assess $p > 1$? Your adversary will require you to pay this amount for a gamble which can only pay £1 — incoherent! Now consider $\mathrm{P}(A \cup B)$ for $A \cap B = \emptyset$. You will pay $\mathrm{P}(A)$ for a gamble on A, and $\mathrm{P}(B)$ for a gamble on B, but since $A \cap B = \emptyset$, at most one of A and B will occur, and so you have paid $\mathrm{P}(A) + \mathrm{P}(B)$ for a gamble on $A \cup B$. *ie.* $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$ for a coherent individual. The other axioms and standard results of probability may be derived in a similar way.

If you do not declare probabilities for a partition of events according to the laws of probability, then a clever adversary can always arrange a combination of bets for you to take which is guaranteed to loose money, whatever event occurs.

A combination of bets that is guaranteed to loose money is known (in Britain!), as a *Dutch Book*. Someone prepared to take such bets is known as a *Money pump*. To avoid being a money pump, you must be coherent!

Consistency with classicism

Subjective probability is quite consistent with the classical interpretation of probability, since if an individual judges a collection of events to be equally likely, then it would be incoherent to assign to the events anything other than equal probability. However, subjectivists maintain that the assessment of whether or not a collection of events is equally likely, is necessarily a subjective one.

1.2.4 Rewards, preferences and gambles

In the previous sections, we have been thinking about gambles where the potential reward was £1. However, there are many decision situations where the outcomes, rewards, payoffs, *etc.*, are non-monetary, and yet we do not immediately have a quantitative scale with which we may compare them. Before we can construct such a *utility* scale for a set of rewards, we need to know the *preference ordering* for that set.

Preferences

Given a choice between two rewards, A and B (*eg.* A might be a cream cake, and B might be a chocolate cake), we say that *you prefer* A *to* B, and write $B \prec^\star A$, if you would pay an amount of money (however small), in order to swap B for A. We say that you are *indifferent* between A and B, written, $A \sim^\star B$, if neither of $A \prec^\star B$ or $B \prec^\star A$ hold. We say that you do not prefer B to A, written $B \preceq^\star A$, if one of $B \prec^\star A$ or $B \sim^\star A$ holds.

For a *coherent* individual, $\preceq^\star$ is a *transitive* logical relation. That is, for three rewards, A, B and C, if $A \preceq^\star B$ and $B \preceq^\star C$, then we must have $A \preceq^\star C$.

Suppose this were not the case, so that $A \preceq^\star B$, $B \preceq^\star C$ and $C \prec^\star A$, and further suppose that you start with reward A. By the first relation, you would be happy to swap A for B. By the second relation, you would be happy to swap B for C, and by the third relation, you would pay money to swap C for A. You now have the same reward you started with, and are poorer — you are an incoherent money pump!

It follows that for a collection of n rewards, there is a labelling, $R_1, R_2, \ldots, R_n$ such that

$$R_1 \preceq^\star R_2 \preceq^\star \cdots \preceq^\star R_n$$

This is known as the *preference ordering* for the reward set. In particular, there is always a *best* reward, R_n, and a *worst* reward, R_1, though these are not necessarily unique.

Gambles

A *gamble*, G, is simply a random reward. *eg.* Toss a coin; if it is heads, the reward is a chocolate cake, and is a cream cake otherwise. Note that despite being random, this may still be regarded as a reward, over which preferences may be held. More formally, we write

$$G = p_1 R_1 +_g p_2 R_2 +_g \cdots +_g p_m R_m$$

for the gamble which returns R_1 with probability p_1, R_2 with probability p_2, *etc.* It is important to understand that this is a notational convenience — you are not "multiplying" chocolate cakes by probabilities, or "adding" bits of chocolate cakes to cream cakes! The use of $+_g$ rather than just $+$ helps emphasise this.

So, if chocolate cake is denoted by H, and cream cake by R, then the gamble above can be written

$$G = \frac{1}{2} H +_g \frac{1}{2} R$$

It is clear that if $H \preceq^\star R$, then we will have $H \preceq^\star G \preceq^\star R$. Gambles provide the link between probability, preference and utility.

An algebra for gambles

At first the $+_g$ notation seems a little strange, but it is in fact very powerful, and really comes into its own when you start to think about gambles on gambles, *etc.*

Consider rewards A and B. The gambles G, H and K are defined in the following way.

$$\begin{aligned} G =& pA +_g (1-p)B \\ H =& qA +_g (1-q)B \\ K =& rG +_g (1-r)H \end{aligned}$$

G and H are simple gambles over rewards A and B, but K is a gamble whose outcomes are themselves gambles over the rewards A and B. Ultimately, the gamble K has to result in the reward A or B, and so must be equivalent to a simple gamble over A and B — but which one?

$$\begin{aligned} K &= rG +_g (1-r)H \\ &= r[pA +_g (1-p)B] +_g (1-r)[qA +_g (1-q)B] \\ &\sim^\star [rp + (1-r)q]A +_g [r(1-p) + (1-r)(1-q)]B \end{aligned}$$

Study the use of $=$, $\sim^\star$, $+$ and $+_g$ carefully!

1.3 Utility

1.3.1 Introduction

Preference orderings are not quantitative — they tell you (say) that B is preferred to A and that C is preferred to B, but they cannot tell you whether or not a 50-50 gamble between A and C is preferred to B. For that we need a numerical scale — *utility*.

A utility function, $u(\cdot)$, is a function mapping rewards/gambles to the real line in a way that preserves preference over those rewards/gambles. *ie.* if $A \preceq^\star B$, then $u(A) \leq u(B)$, *etc.* The formal definition is as follows

Definition

*A **utility function**, $u(\cdot)$, on gambles*

$$G = p_1R_1 +_g p_2R_2 +_g \cdots +_g p_mR_m$$

over rewards, $R_1, \ldots, R_m$, assigns a real number $u(G)$ to each gamble subject to the following conditions:

1. *If $G_1 \prec^\star G_2$, then $u(G_1) < u(G_2)$, and if $G_1 \sim^\star G_2$, then $u(G_1) = u(G_2)$.*
2. *For any $p \in [0,1]$, and rewards A, B, we have*

$$u(pA +_g [1-p]B) = pu(A) + [1-p]u(B)$$

It follows immediately that

$$u(G) = p_1u(R_1) + \cdots + p_mu(R_m)$$

and hence that

$$u(G) = \mathrm{E}(u(G)).$$

1.3.2 Making decisions

In principle, the problem of decision making is now completely solved. Condition 1 tells us that utilities agree with preferences, and so one should always choose the reward with the largest utility, and condition 2 tells us that the utility of a gamble is the *expected* utility of the gamble. The two conditions together tell us that given a collection of gambles over a set of rewards, we should always choose the one with the highest expected utility.

Now all we need to know is a method for assessing utilities, analogous to the method we devised for assessing subjective probabilities.

Assessing utility

For an ordered set of rewards, $R_1 \preceq^\star R_2 \preceq^\star \cdots \preceq^\star R_n$, define $u(R_1) = 0$ and $u(R_n) = 1$. Then for integer i such that $1 < i < n$, define $u(R_i)$ to be the probability, p, such that

$$R_i \sim^\star (1-p)R_1 +_g pR_n$$

For example, if A, B and C are Alabama, Banana and Chocolate cakes respectively, and you hold $A \prec^\star B \prec^\star C$, then put $u(A) = 0$ and $u(C) = 1$, and consider the gamble $pC +_g (1-p)A$. The value of p which makes you indifferent between this gamble and B is your utility for B.

We need to show that this assessment method does indeed give a utility function the way we defined it earlier.

We first need to show that for gambles G_1 and G_2, such that $G_1 \prec^\star G_2$, we have $u(G_1) < u(G_2)$. By the construction of our function, the utilities of G_1 and G_2 are such that

$$\begin{aligned} G_1 &\sim^\star u(G_1)R_n +_g [1-u(G_1)]R_1 \\ G_2 &\sim^\star u(G_2)R_n +_g [1-u(G_2)]R_1 \end{aligned}$$

But since $R_1 \prec^\star R_n$, for G_2 to be preferred to G_1, the probability of R_n in the equivalent gamble must be bigger for G_2. That is, $u(G_1) < u(G_2)$.

Now we need to show that $u(pG_1 +_g (1-p)G_2) = pu(G_1) + (1-p)u(G_2)$ Well, the gamble $pG_1 +_g (1-p)G_2$ is equivalent to

$$\begin{aligned} p\{u(G_1)R_n +_g [1-u(G_1)]R_1\} +_g (1-p)\{u(G_2)R_n +_g [1-u(G_2)]R_1\} \\ \sim^\star \big\{pu(G_1) + (1-p)u(G_2)\big\}R_n +_g \big\{p[1-u(G_1)] + (1-p)[1-u(G_2)]\big\}R_1 \end{aligned}$$

and so the utility of the gamble must be $pu(G_1) + (1-p)u(G_2)$.

1.3.3 Uniqueness of utility

Is utility unique? That is, is there only one set of numbers which appropriately represent your preferences for gambles over a set of rewards? The answer is yes and no!

Suppose that $u(\cdot)$ is a utility function. Define a new function $v(\cdot)$ such that $v(\cdot) = a\,u(\cdot) + b$, where b is any real number, and a is any strictly positive real number. In fact, $v(\cdot)$ is a perfectly valid utility function, entirely equivalent to $u(\cdot)$. Let us now demonstrate this.

If $R_1 \prec^\star R_2$, then $u(R_1) < u(R_2)$, since $u(\cdot)$ is a utility. Now, since $a > 0$, $au(R_1) < au(R_2)$ and $au(R_1) + b < au(R_2) + b$, so that $v(R_1) < v(R_2)$. Clearly $v(\cdot)$ satisfies the first condition of utility. Also,

$$\begin{aligned} v(pR_1 +_g [1-p]R_2) &= a\,u(pR_1 +_g [1-p]R_2) + b \\ &= ap\,u(R_1) + a[1-p]u(R_2) + b \\ &= p\big[au(R_1) + b\big] + [1-p]\big[au(R_2) + b\big] \\ &= p\,v(R_1) + [1-p]v(R_2) \end{aligned}$$

So, $v(\cdot)$ is an equivalent utility function. This tells us that if we have a utility function, all positive linear scalings of it are valid, equivalent utility functions. Are there any equivalent utility functions which are not positive linear scalings? The answer is no.

Suppose that $u(\cdot)$ and $v(\cdot)$ are equivalent utility functions. Then we want to show that there exists $a > 0$ and b such that $v(\cdot) = a\,u(\cdot) + b$. Let R_n be the best reward, and R_1 be the worst reward. We know that $u(R_n) > u(R_1)$ and $v(R_n) > v(R_1)$, since $u(\cdot)$ and $v(\cdot)$ are utility functions. So, there exist $a > 0$ and b such that

$$\begin{aligned} a\,u(R_n) + b =& v(R_n) \\ a\,u(R_1) + b =& v(R_1) \end{aligned}$$

since these are just two linear equations in two unknowns. So, we now have our candidates for a and b, but we still need to show that they work for all rewards, R_i. First find p such that

$$R_i \sim^\star [1-p]R_1 +_g pR_n$$

Then,

$$\begin{aligned} v(R_i) =& v([1-p]R_1 +_g pR_n) \\ =& [1-p]v(R_1) + p\,v(R_n) \\ =& [1-p]\big[a\,u(R_1) + b\big] + p\big[a\,u(R_n) + b\big] \\ =& a\big[[1-p]u(R_1) + p\,u(R_n)\big] + b \\ =& a\,u(R_i) + b \end{aligned}$$

Therefore utility is unique up to a positive linear transform. That is, any linear scaling of a utility function gives an equivalent utility function, but if two utility functions are not linear scalings of one another, then they are not equivalent, and there are gambles for which preferences are different.

1.4 Utility for money

1.4.1 Introduction

Utility was motivated as a quantitative scale for comparison of gambles over rewards which were not necessarily quantitative. However, utility is not just *a* quantitative scale — it is the *correct* quantitative scale. In particular, it turns out that money, despite being a quantitative scale, is inappropriate for making decisions regarding financial gambles. Somewhat ironically, it is in the field of decision making for financial gambles that utility theory has proved most powerful.

1.4.2 Example

Suppose you have a computer worth £1,500. The probability it is stolen this year is 0.02. It will cost you £42 to insure it against theft. Do you?

The insurance company will pay £1,500 with probability 0.02, and £0 otherwise. *ie.* from the viewpoint of the insurance company, they are offering the gamble

$$G = 0.02£1,500 +_g 0.98£0$$

The *Expected Monetary Value* (EMV) of this gamble is

$$\text{EMV}(G) = £0.02 \times 1500 = £30$$

This is less than the £42 premium, so perhaps you should not insure? But insurance companies make money — their premiums are always bigger than the expected payoff! Does this mean that no-one should ever insure anything? No — it simply means that money is not a utility scale.

1.4.3 The St. Petersburg Paradox

(due to Bernoulli)

Consider a game where you toss a coin until you get a head. With every tail you toss, your prize doubles. For example, if you make n tosses, you receive $£2^n$.

$$\begin{aligned}G =& \frac{1}{2^1}£2^1 +_g \frac{1}{2^2}£2^2 +_g \frac{1}{2^3}£2^3 +_g \cdots \\ \text{EMV}(G) =& £\left(\frac{2^1}{2^1} + \frac{2^2}{2^2} + \frac{2^3}{2^3} + \cdots\right) \\ =& £(1 + 1 + 1 + \cdots) \\ =& £\infty\end{aligned}$$

What a great game — infinite expected payoff! But how much would *you* pay to play the game? We will return to this later...

1.4.4 Attitude to risk

Consider a monetary gamble

$$G = p_1£R_1 +_g p_2£R_2 +_g \cdots +_g p_n£R_n$$

There are two "expectations" of interest.

$$\text{EMV}(G) = £\big(p_1R_1 + p_2R_2 + \cdots + p_nR_n\big)$$

and

$$u(G) = p_1u(R_1) + p_2u(R_2) + \cdots + p_nu(R_n)$$

The first expectation is an amount of money, and the second is a utility. In order to be able to compare these, we need the following.

Definitions

For a monetary gamble, G, the *certainty equivalent* of G, $£R$, is the amount of money such that

$$£R \sim^\star G$$

or equivalently, such that $u(£R) = u(G)$.

The *risk premium* of G is the amount, Π_G, where

$$\Pi_G = \text{EMV}(G) - £R$$

A decision maker is *risk averse* if, for any monetary gamble, G, $\Pi_G > 0$, and *risk prone* if $\Pi_G < 0,\ \forall G$. If $\Pi_G = 0,\ \forall G$, then the decision maker is *risk neutral* and can make optimal decisions by maximising EMV.

[N.B. Many people are risk prone for small gambles and risk averse for large gambles. There is nothing incoherent about this as long as they are consistent (which most are not!).

Also, in the OR courses, such as MAS237 (OR 1B), you will make decisions by maximising EMV. This is fine for risk-neutral decision makers. For complex decision problems in a non-risk-neutral situation, you can use the techniques from MAS237, such as decision trees, replacing monetary rewards by utilities — the algorithms are identical.]

1.4.5 Risk averse utility functions

What does a risk-averse utility function look like? Choose any two monetary rewards, $£R_1$ and $£R_2$ (where $R_1 < R_2$), and consider gambles on those rewards.

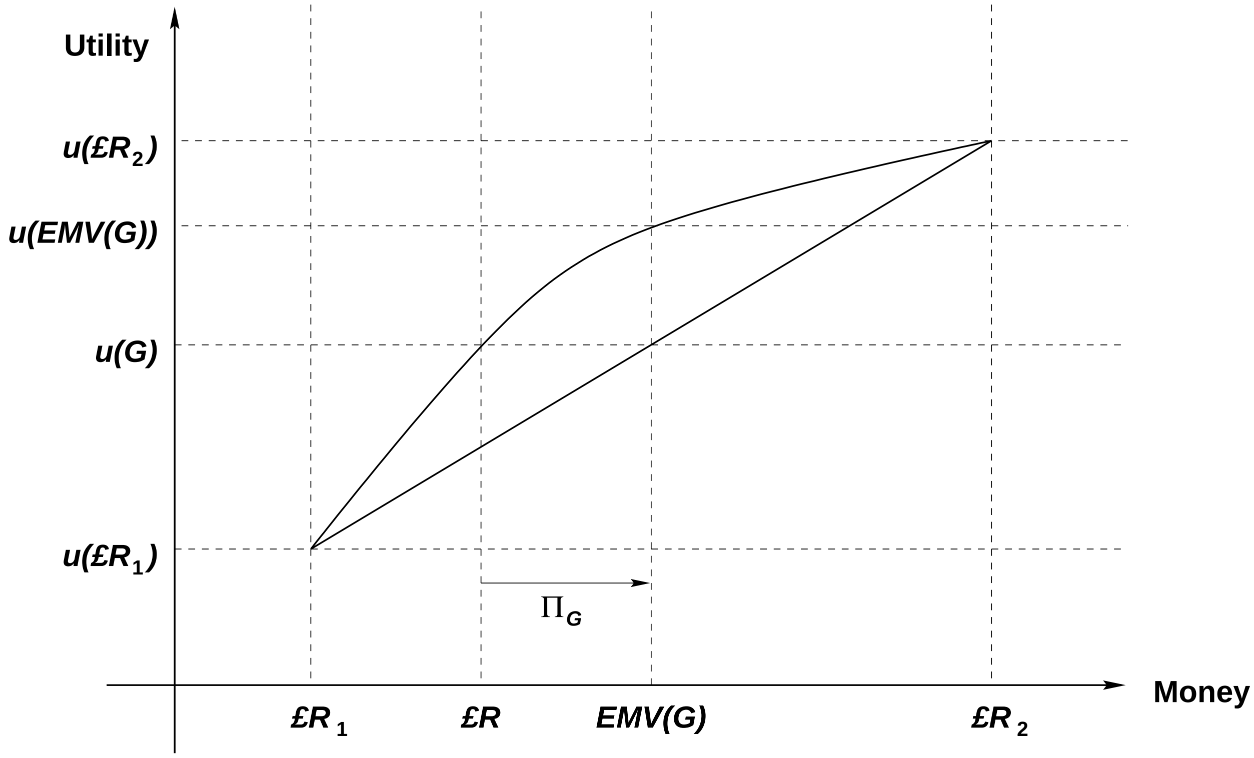
$$\begin{aligned} G =& p£R_1 +_g (1-p)£R_2 \\ u(G) =& p\,u(£R_1) + (1-p)u(£R_2) \\ u(\text{EMV}(G)) =& u(£(pR_1 + (1-p)R_2)) \end{aligned}$$

Now, risk averseness implies that $u(\text{EMV}(G)) \geq u(G)$, so

$$u(£(pR_1 + (1-p)R_2)) \geq p\,u(£R_1) + (1-p)u(£R_2)$$

That is, the utility function lies above the chord between any two points — this is the definition of a *concave* function!

Risk averse	$\iff$	Concave utility for money
Risk neutral	$\iff$	Linear utility for money
Risk prone	$\iff$	Convex utility for money



The above figure shows utility, EMV and certainty equivalent for the gamble $G = \frac{1}{2}£R_1 +_g \frac{1}{2}£R_2$, together with a concave utility for money.

We now know that risk averse individuals have concave utilities for money. Does the reverse also hold? That is, does every monotonic increasing concave function represent a risk-averse utility function? Yes! To show this, we need to use *Jensen's inequality*, which states that for any monotonic increasing strictly concave function, $f(\cdot)$, and any (non-trivial) random variable, X, we have

$$\mathrm{E}(f(X)) < f(\,\mathrm{E}(X))$$

So, suppose we have a candidate monotonic increasing concave function, $u(\cdot)$. The fact that it is monotonic increasing makes it a valid utility for money, so we just need to show that it represents a risk-averse utility function. Suppose we have any monetary gamble, G, with uncertain payoff $£X$. Clearly

$$u(\text{EMV}(G)) = u(\,\mathrm{E}(X))$$

and

$$u(G) = \mathrm{E}(u(X))$$

so Jensen's inequality tells us that

$$u(G) < u(\text{EMV}(G))$$

This means that the risk premium for the gamble, Π_G must be positive, and hence this utility function is risk-averse.

So, given some parameterised classes of concave functions, one can simply pick the class which best represents your utility profile, and then fix parameters according to your wealth, degree of risk aversion *etc.*, and then make decisions for arbitrary monetary gambles using this utility function.

1.4.6 Logarithmic utility for money

A good example of a concave utility function (which also has some other desirable properties) is the log function. That is, a class of utility functions for money, with a single tunable parameter, c, is given by

$$u(x) = \log(x + c)$$

The parameter c controls the degree of risk aversion. Note that if we fix c and look at positive linear transforms of this utility function,

$$u(x) = a \log(x + c) + b$$

varying $a > 0$ and b does not change the properties of the utility function in any way (so, in particular, the base of the logarithm will not change the properties of the utility function). Note that using the second form of the utility function, fixing utilities for three monetary values will give us three equations in three unknowns, which can be solved to give a, b and c.

In general, this is tricky to solve analytically, so another way to fix c is to consider a small reward, £R_1, and a large reward, £R_3. Small and large means they should cover most of the likely range of gambles under consideration. Then consider the reward £R_2, such that

$$£R_2 \sim^\star \frac{1}{2}£R_1 +_g \frac{1}{2}£R_3$$

From this we get that

$$2\log(R_2 + c) = \log(R_1 + c) + \log(R_3 + c)$$

and solving this for c we get

$$c = \frac{R_2^2 - R_1 R_3}{R_1 + R_3 - 2R_2}$$

— Exercise!

For example, if it is the case that for you,

$$£30 \sim^\star \frac{1}{2}£0 +_g \frac{1}{2}£100$$

and if you think your utility for money can be reasonably well approximated as logarithmic, then the tuning parameter, c, is given by 22.5. That is, your utility function is

$$u(x) = \log(x + 22.5)$$

Another way to fix the parameter c, is to regard it as a measure of the decision maker's total wealth. That is, we consider the decision maker's utility for having total wealth x to be

$$u(x) = \log(x)$$

and utilities for gambles are assessed by considering their effect on total wealth. Note however, that there are no tuning parameters in such a situation, and so it is important to assess whether or not such a utility function does match the decision maker's preferences for monetary gambles.

Twice-differentiable utility functions

For twice-differentiable functions, convexity and concavity can be determined from the sign of the second derivative, and so we have the following alternative characterisation.

Risk averse	$\iff$	$u''(x) < 0, \ \forall x$
Risk neutral	$\iff$	$u''(x) = 0, \ \forall x$
Risk prone	$\iff$	$u''(x) > 0, \ \forall x$

1.4.7 St. Petersburg paradox (revisited)

Suppose for simplicity that your utility for reward £X is $\log(X)$. How much then would you pay to play the game? *ie.* what is your certainty equivalent for this game?

$$
\begin{aligned}
u(G) =& \frac{1}{2^1}u(2^1) + \frac{1}{2^2}u(2^2) + \frac{1}{2^3}u(2^3) + \cdots \\
=& \frac{1}{2^1}\log(2^1) + \frac{1}{2^2}\log(2^2) + \frac{1}{2^3}\log(2^3) + \cdots \\
=& \frac{1}{2^1}\log(2) + \frac{2}{2^2}\log(2) + \frac{3}{2^3}\log(2) + \cdots \\
=& \log(2)\sum_{r=1}^{\infty} r\left(\frac{1}{2}\right)^r \\
=& \log(2) \times 2 \qquad\qquad \text{(Exercise!)} \\
=& 2\log(2)
\end{aligned}
$$

So our certainty equivalent is the amount, £X such that $u(X) = 2\log(2)$, so the game is worth £4.

By using $u(x) = \log(x)$ rather than $u(x) = \log(x + c)$ for some $c > 0$, we have made the calculations easy, but $\log(x)$ is rather too risk averse for most people — I would certainly pay more than £4 to play this game!

1.4.8 Measuring risk attitudes

How can you measure the degree of risk aversion of an individual? You might be tempted to use the second derivative, $u''(x)$, but this changes when you change the scaling of the utility function. Consequently, the following measure is usually used.

$$
r(x) = \frac{-u''(x)}{u'(x)}
$$

$r(x)$ is your *local risk aversion* at x. The sign of $r(x)$ has the following interpretations.

Risk averse	$\iff$	$r(x) > 0, \ \forall x$
Risk neutral	$\iff$	$r(x) = 0, \ \forall x$
Risk prone	$\iff$	$r(x) < 0, \ \forall x$

We still need to confirm that for utilities $u(\cdot)$ and $v(\cdot) = a\,u(\cdot) + b$, the risk aversion function, is the same. Let $r(\cdot)$ be the risk aversion for $u(\cdot)$ and $s(\cdot)$ be the risk aversion for $v(\cdot)$. Then

$$\begin{aligned} s(x) =& \frac{-v''(x)}{v'(x)} \\ =& \frac{-(a\,u(x)+b)''}{(a\,u(x)+b)'} \\ =& \frac{-a\,u''(x)}{a\,u'(x)} \\ =& r(x) \end{aligned}$$

So risk aversion *is* a scale-free measure.

1.4.9 Decreasing risk-aversion

The richer you get, the less risk-averse you become. This is essentially why big companies have approximately linear utility for relatively small amounts of money, and hence can make decisions by maximising EMV.

Risk-aversion for logarithmic utility

Let us consider the risk-aversion function for the logarithmic utility function, $u(x) = \log(x + c)$. First we need the first and second derivatives of this utility function.

$$\begin{aligned} u'(x) =& \frac{1}{x+c} \\ u''(x) =& \frac{-1}{(x+c)^2} \end{aligned}$$

So the risk-aversion function is

$$r(x) = \frac{1}{x+c}$$

Notice that this function tends to zero (risk-neutrality) as *either* the wealth parameter, c, *or* the amount of money, x, tends to infinity.

Of course, this is simply because the log function flattens out — but this highlights another desirable feature of this parametric form for monetary utility. However, a possibly undesirable feature is that the log function is unbounded — some people consider this unrealistic. There are other parametric forms regularly used for representing monetary utility — see Exercises 2.

1.4.10 Buying and selling gambles

When considering the monetary gamble, G, the certainty equivalent, $£R$, such that

$$£R \sim^\star G$$

has been of use in understanding the value of the gamble. If you are offered a choice between $£R$ and G, you would be indifferent between them. It is tempting to interpret this as the maximum amount you would pay in order to *buy* the gamble. In fact, this is not the case — it is actually the minimum amount you would *sell* the gamble for, if you owned it.

It is easy to see why, intuitively, these amounts are not necessarily the same. *eg.* it is easy to construct a gamble which will have a certainty equivalent of £1,000,000 for you, but that doesn't mean that you would pay that amount for it — you may not have that much money to spend! It simply means that you would be indifferent between having the gamble and having £1,000,000. *ie.* if you owned the gamble, the minimum amount you would sell it for is £1,000,000.

1.4.11 Example

Suppose that an individual has utility for money

$$u(x) = \sqrt{x + 100}$$

and is considering the gamble

$$G = \frac{2}{3}£0 +_g \frac{1}{3}£50$$

First note that the EMV of the gamble is £16.67, but this is bigger than both the buy and sell price of the gamble. We now calculate the utility of the gamble.

$$u(G) = \frac{2}{3}\sqrt{100} + \frac{1}{3}\sqrt{150} = 10.749$$

So the certainty equivalent of the gamble is given by

$$R = 10.749^2 - 100 = 15.54$$

So if the individual "owned" the gamble, they would not sell it for less than £15.54.

Now let's think about how much the individual would actually *pay* in order to *buy* the rights to the gamble.

Suppose that the individual is considering whether or not to pay an amount, £p, in order to buy the gamble. Then the individual has a choice between buying (G_1) and not buying (G_2) the gamble.

$$\begin{aligned} G_1 =& \frac{2}{3}£(-p) +_g \frac{1}{3}£(50 - p) \\ G_2 =& £0 \end{aligned}$$

The individual will choose the gamble with the highest utility. The maximum p for which the individual will buy the gamble occurs when $G_1 \sim^\star G_2$, that is, when $u(G_1) = u(G_2)$. The utilities of the gambles are

$$\begin{aligned} u(G_1) =& \frac{2}{3}\sqrt{100 - p} + \frac{1}{3}\sqrt{150 - p} \\ u(G_2) =& \sqrt{100} = 10 \end{aligned}$$

So, the maximum purchase price, p, is given by a solution to

$$10 = \frac{2}{3}\sqrt{100 - p} + \frac{1}{3}\sqrt{150 - p}$$

Squaring up and re-arranging gives

$$9p^2 + 7500p - 117500 = 0$$

The positive root is 15.38. So, the maximum amount of money you would *pay* in order to *buy* the gamble, is £15.38.

1.4.12 Repeated games

When we have been considering choices between gambles, we have been thinking about a single decision for a "one-off" choice. Sometimes however, gambles can be repeated — this is particularly true in the context of "games of chance".

Consider a game which pays £10 with probability 1/2, and nothing otherwise.

The gamble can be written

$$G = \frac{1}{2}£0 +_g \frac{1}{2}£10$$

Then the EMV of G is £5. That is, on average this game will pay out £5 per go.

However, this game can be played more than once. What form does that gamble take if we are to play the game twice? We will write this as G^2. Again, this is just a notational convenience — we are not "multiplying" the gamble by itself!

$$\begin{aligned} G^2 &= \left(\frac{1}{2}£0 +_g \frac{1}{2}£10\right)^2 \\ &= \frac{1}{4}£0 +_g \frac{1}{2}£10 +_g \frac{1}{4}£20 \end{aligned}$$

Of course, the EMV of G^2 is £10, twice the EMV of G. EMV behaves very nicely in this regard. However, utility (and therefore preference) does not behave so nicely.

In fact, your certainty equivalent for the repeated gamble tends to the EMV of the repeated gamble, as the number of repetitions increases. This is what *expected* payoff is all about. In the context of financial gambles repeated a great many times, EMV is a perfectly acceptable criterion for making decisions. It is in the context of financial decisions made once or a small number of times that utility for money is so important.

1.4.13 Continuous payoffs

So far we have only considered financial gambles where the payoff is a discrete random variable. The theory works equally well for decision problems where the payoff is a continuous random variable.

We already know that if your utility for money is $u(x)$, and a gamble, G has payoff £X, where X is a (discrete or continuous) random variable, then the utility of the gamble, $u(G)$, is given by $\mathrm{E}(u(X))$, the expected utility of G.

So, given a choice between gambles $G_1, G_2, \ldots$ with payoffs £X_1, £$X_2, \ldots$ respectively, simply calculate $u(G_1) = \mathrm{E}(u(X_1)), u(G_2) = \mathrm{E}(u(X_2)), \ldots$ and choose the gamble with the largest utility.

1.4.14 Example

Consider the gamble, G, which has random payoff £X, where X is a continuous random variable uniformly distributed on $[10, 50]$. Suppose that your utility for money is given by $u(x) = \log(x)$. What is your risk premium, Π_G, for G?

The random variable X has pdf $f(x)$, where

$$f(x) = \begin{cases} \frac{1}{40} & 10 \le x \le 50 \\ 0 & \text{otherwise.} \end{cases}$$

Also, the expected value of X is $(10 + 50)/2 = 30$, so $\mathrm{EMV}(G) = £30$.

Now

$$\begin{aligned} u(G) =& \ \mathrm{E}(u(X)) \\ &= \int u(x) f(x)\, dx \\ &= \int_{10}^{50} \log(x) \times \frac{1}{40}\, dx \\ &= \frac{1}{40} \big[x \log(x) - x \big]_{10}^{50} \\ &= 3.314 \end{aligned}$$

So the certainty equivalent is £27.50 and the risk premium, Π_G is £2.50.

The problem with continuous payoffs is that the necessary integrals become difficult to tackle analytically if the payoff has a distribution other than uniform and utility is non-linear. However, the necessary expectations are easily calculated *via* simulation on a computer.

1.4.15 Example

If your utility for money is $u(x) = \log(x + 1)$ and the uncertain payoff is an exponential random variable with parameter 1/4, what is the certainty equivalent of the gamble?

Well, the exponential random variable has pdf $f(x)$ where

$$f(x) = \begin{cases} \frac{1}{4} \exp(-x/4) & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

and so the expected utility is given by

$$\begin{aligned} \mathrm{E}(u(X)) &= \int u(x) f(x)\, dx \\ &= \int_0^\infty \frac{1}{4} \log(x + 1) \exp(-x/4)\, dx \end{aligned}$$

This integral isn't very easy to do analytically, but the Maple command

```
int(0.25*log(x+1)*exp(-0.25*x),x=0..infinity);
```

returns 1.341. However, in some ways it is more instructive to think about obtaining this answer *via* simulation in Minitab. First we can simulate 10000 exponential random variables.

```
MTB > random 10000 c1;
SUBC> exponential 4.
MTB > mean c1
```

Note that Minitab parameterises the exponential distribution with its mean, rather than the reciprocal. The sample mean of these simulated values will be close to the expectation of the random variable, X, which is 4. Now, we are actually interested in the expected value of the random variable $u(X)$, as this is the utility of the gamble. We can simulate these with

```
MTB > let c2=log(c1+1)
MTB > mean c2
```

Now c2 contains simulated values of the random variable $u(X)$, and so the sample mean of these will be close to the expectation of $u(X)$, the utility of the gamble, 1.341.

The certainty equivalent is then just $\exp(1.341) - 1 = 2.82$. So the gamble is worth £2.82 which is less than the EMV of £4, as it should be.

This "Monte Carlo" simulation method for solving decision problems "scales up" much better than the direct integration approach, in the context of a complex decision situation such as repeated games or a multi-stage decision problem.

1.5 Multi-attribute utility

1.5.1 Introduction

Suppose that you regularly travel around the UK, sometimes by train, and sometimes by coach. There are two factors (known as *attributes*) you want to take into account when deciding how to travel: journey time (T) and cost (C). Generally speaking, the train is faster, but costs more. Also, both modes of transport are subject to uncertain delays.

In order to decide on which mode of transport is best, a utility function for all t and c combinations is required. That is, the *multi-attribute utility function*, $u(t, c)$ needs to be specified. How do we do this?

Clearly it would be reasonably straight forward to specify $u(t)$ and $u(c)$ separately. We already know how to specify $u(c)$ — it is just a utility for money. $u(t)$ could be specified similarly, using some simple parametric form. However, given these two functions, is it possible to deduce the form of $u(t, c)$? Answer: Sort of, sometimes!

1.5.2 Utility independence

Let $u(x, y)$ be a joint utility function over two reward attributes, X and Y. We say that X is *utility independent* of Y if $u(x, y_0)$ and $u(x, y_1)$ are *equivalent* utility functions over the reward attribute X, irrespective of the choice of y_0 and y_1.

N.B. We are *not* saying that $u(x, y_1) = u(x, y_0)$! Remember that utility functions are equivalent up to a positive linear scaling, so all we are actually saying is that

$$u(x, y_1) = a\,u(x, y_0) + b$$

for some $a > 0$ and b, if X is utility independent of Y.

Returning to our example for time and money, we would say that the cost attribute, C, is utility independent of the time attribute, T, if your utility function for cost remains equivalent as you change journey time.

For example, if you hold

$$£22 \sim^{\star} \frac{1}{2}£10 +_g \frac{1}{2}£30$$

for the cost attribute when the journey time is 1 hour, you will also hold this if the journey time is 2 hours.

However, we actually need a bit more than this "one-sided" independence in order to make progress with finding a convenient form for $u(t, c)$.

1.5.3 Mutual utility independence

We say that reward attributes X and Y are *mutually utility independent* if X is utility independent of Y, *and* Y is utility independent of X.

Mutual utility independence is the key property required in order to make progress with utility function representation. If all of the attributes of a utility function are mutually utility independent, then there turns out to be a neat way to represent the multi-attribute utility function in terms of the marginal utility functions.

If mutual utility independence does not hold, then there is no neat representation, and utilities over all possible reward combinations must be elicited directly.

Let us now examine the implication of mutual utility independence for a multi-attribute utility function, $u(x, y)$ over two attributes, X and Y. First, it tells us that X is utility independent of Y. So, for any y_0 and y_1, we will have

$$u(x, y_1) = a\, u(x, y_0) + b$$

for some $a > 0$ and b. However, if we let y_1 vary, a and b may change. That is, a and b are both functions of y_1, but crucially, are *not* functions of x. That is

$$u(x, y_1) = a(y_1)u(x, y_0) + b(y_1)$$

ie. for any chosen y_0, the utility function takes the form

$$u(x, y) = a(y)u(x, y_0) + b(y)$$

By a similar argument, the fact that Y is utility independent of X allows us to deduce that there exist functions $c(x)$ and $d(x)$ such that

$$u(x, y) = c(x)u(x_0, y) + d(x)$$

for any chosen x_0. Together, these representations are enough to allow us to deduce the following theorem.

Theorem (two-attribute representation theorem)

If reward attributes X and Y are mutually utility independent, then the two-attribute utility function, $u(x, y)$, takes the form

$$\boxed{u(x, y) = u(x, y_0) + u(x_0, y) + k\, u(x, y_0)u(x_0, y)}$$

for some real number k and arbitrarily chosen origin (x_0, y_0) such that $u(x_0, y_0) = 0$.

N.B. This says that for some $\alpha, \beta > 0$ and γ, we have

$$u_{X,Y}(x, y) = \alpha\, u_X(x) + \beta\, u_Y(y) + \gamma\, u_X(x)u_Y(y)$$

So, if we have mutual utility independence, we just need to specify a few constants in addition to the marginal utility functions, in order to fix the full form of the multi-attribute utility function.

Proof

X is utility independent of Y and Y is utility independent of X together imply the existence of functions $a(y), b(y), c(x), d(x)$ such that

$$u(x, y) = a(y)u(x, y_0) + b(y) \tag{1}$$

and

$$u(x, y) = c(x)u(x_0, y) + d(x) \tag{2}$$

Putting $x = x_0$ into (1) and $y = y_0$ in (2) gives

$$u(x_0, y) = b(y) \tag{3}$$
$$u(x, y_0) = d(x) \tag{4}$$

Now putting (3) in (1) and (4) in (2) gives

$$u(x, y) = a(y)u(x, y_0) + u(x_0, y) \tag{5}$$
$$u(x, y) = c(x)u(x_0, y) + u(x, y_0) \tag{6}$$

Now we put (5)=(6) and rearrange to separate x and y.

$$\begin{aligned}
a(y)u(x, y_0) + u(x_0, y) =& c(x)u(x_0, y) + u(x, y_0) \\
\Rightarrow a(y)u(x, y_0) - u(x, y_0) =& c(x)u(x_0, y) - u(x_0, y) \\
\Rightarrow [a(y) - 1]u(x, y_0) =& [c(x) - 1]u(x_0, y) \\
\Rightarrow \frac{a(y) - 1}{u(x_0, y)} =& \frac{c(x) - 1}{u(x, y_0)} \qquad (7)
\end{aligned}$$

Now the LHS is a function of y only, and the RHS is a function of x only. Yet x and y are allowed to vary independently. Thus, the equality can only hold if both sides are equal to some constant, k. In particular, we must have

$$\begin{aligned}
\frac{a(y) - 1}{u(x_0, y)} =& k \\
\Rightarrow a(y) =& k\, u(x_0, y) + 1 \qquad (8)
\end{aligned}$$

Finally, putting (8) in (5) gives

$$\begin{aligned}
u(x, y) =& [k\, u(x_0, y) + 1]u(x, y_0) + u(x_0, y) \\
=& u(x, y_0) + u(x_0, y) + k\, u(x, y_0)u(x_0, y)
\end{aligned}$$

□

How do we interpret k?

Consider continuous attributes X and Y such that more X is preferred to less, and more Y is preferred to less. *ie.* The marginal utility functions are increasing. For example, X might be "money" and Y might be "fun".

Consider the following rewards (where $x_0 < x_1$ and $y_0 < y_1$):

A	(x_0, y_0)	no money or fun
B	(x_0, y_1)	fun but no money
C	(x_1, y_0)	money but no fun
D	(x_1, y_1)	money and fun

Clearly we have $A \prec^\star B \prec^\star D$ and $A \prec^\star C \prec^\star D$. We can't say whether or not you prefer B to C, but this doesn't matter. Now consider the following two gambles on those rewards.

$$
\begin{aligned}
G =& \frac{1}{2}A +_g \frac{1}{2}D \\
H =& \frac{1}{2}B +_g \frac{1}{2}C
\end{aligned}
$$

So with gamble G you get all or nothing, and with gamble H you get one of money or fun.

If X and Y are mutually utility independent, then by the two-attribute representation theorem,

$$
\begin{aligned}
u(D) =& u(x_1, y_1) \\
=& u(x_1, y_0) + u(x_0, y_1) + k\, u(x_1, y_0)u(x_0, y_1)
\end{aligned}
$$

and we know $u(A) = 0$, so the utility of G is

$$
\begin{aligned}
u(G) =& \frac{1}{2}u(A) + \frac{1}{2}u(D) \\
=& \frac{1}{2}[u(x_1, y_0) + u(x_0, y_1) \\
& \qquad + k\, u(x_1, y_0)u(x_0, y_1)]
\end{aligned}
$$

and

$$
\begin{aligned}
u(H) =& \frac{1}{2}[u(B) + u(C)] \\
=& \frac{1}{2}[u(x_1, y_0) + u(x_0, y_1)]
\end{aligned}
$$

So

$$
u(G) - u(H) = \frac{k}{2}u(x_1, y_0)u(x_0, y_1)
$$

So, if $k > 0$, $H \prec^\star G$. If $k < 0$, $G \prec^\star H$, and if $k = 0$, $G \sim^\star H$. What does this mean intuitively, in terms of our example?

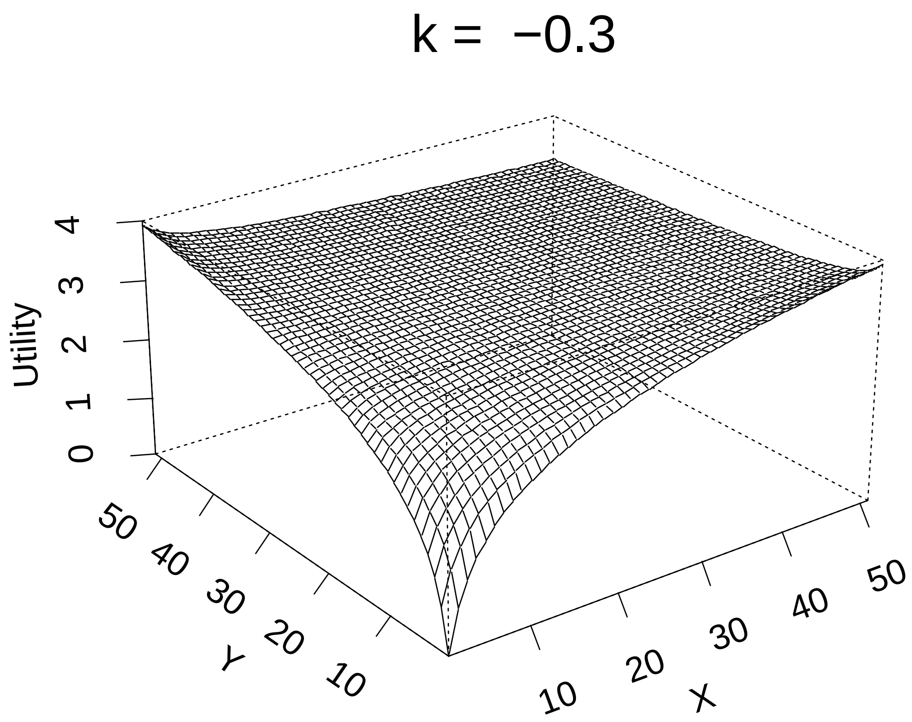
- If $k > 0$, then it is important to have *both* money and fun — having lots of one or the other isn't a good substitute. In this case, the attributes are said to be *complimentary*.
- If $k < 0$ having one of money or fun is almost as good as having both money and fun. In this case, the attributes are said to be *substitutes*.
- If $k = 0$, then preferences for money and fun add together in a simple way, and the attributes are said to be *preference unrelated*.

$k > 0$	X, Y are complimentary
$k < 0$	X, Y are substitutes
$k = 0$	X, Y are preference unrelated

Let us now consider what the utility surfaces look like for different values of k. We will look at the case where the marginal utilities are logarithmic and increasing. Suppose that $u(x, 0) = \log(x + 1)$ and $u(0, y) = \log(y + 1)$. Then we have

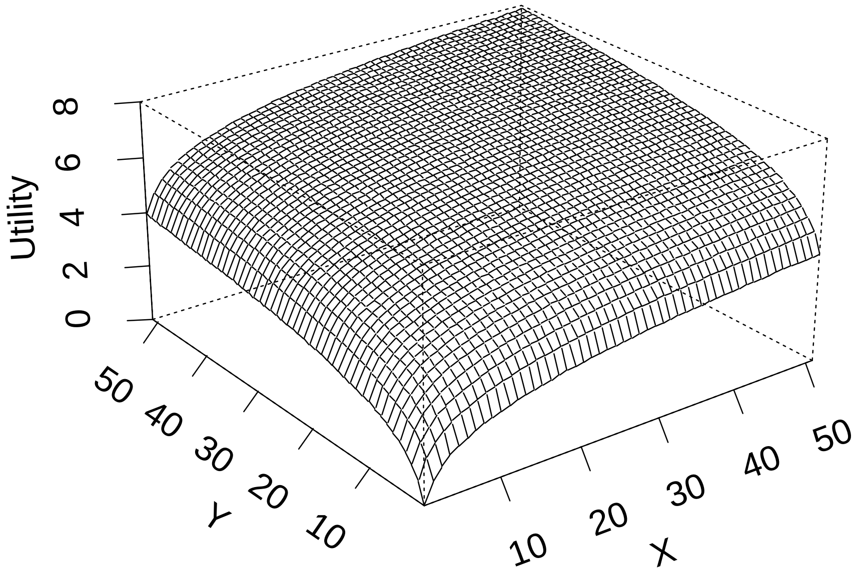
$$\begin{aligned} u(x, y) =& u(x, 0) + u(0, y) + k\, u(x, 0)u(0, y) \\ =& \log(x + 1) + \log(y + 1) \\ &+ k \log(x + 1) \log(y + 1) \end{aligned}$$

We can plot this for different values of k.

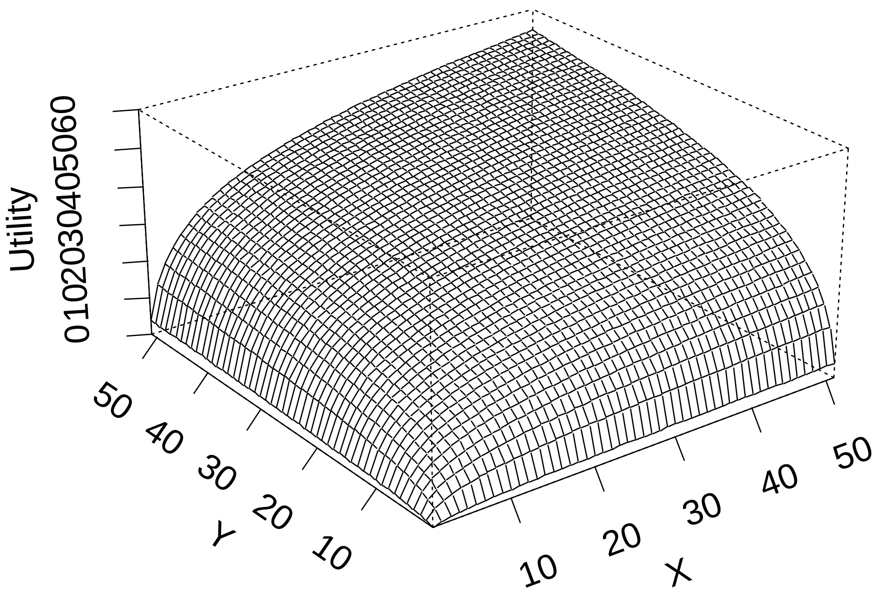


Above, X and Y are substitutes, so that X or Y is as good (if not better) than X and Y together. Below we have X and Y preference unrelated, so that the utilities just add up. That is, the height at $(50, 50)$ is the height at $(50, 0)$ plus the height at $(0, 50)$.

k = 0


Utility
0 2 4 6 8
50 40 30 20 10
Y
10 20 30 40 50
X

k = 3


Utility
0102030405060
50 40 30 20 10
Y
10 20 30 40 50
X

Here we have X and Y complimentary. In this case, utility increases very slowly in the direction of the x- or y-axes, but comparatively fast in the direction $y = x$. The positive interaction means that both attributes need to be high in order to have a high utility.

1.5.4 Specification of two-attribute utility functions

We now know that if two attributes, X and Y are mutually utility independent, then the joint utility function is determined by the utility functions for the marginals, and a few constants. How do we fix those constants?

Suppose that we have assessed $u_X(x)$ such that $u_X(x_0) = 0$ and $u_X(x_1) = 1$ (it may be necessary to re-scale the utility function so that x_0 and x_1 are convenient reward attributes to consider). Similarly, we have $u_Y(y)$ such that $u_Y(y_0) = 0$ and $u_Y(y_1) = 1$. The two-attribute utility function takes the form:

$$\boxed{u(x, y) = k_X u_X(x) + k_Y u_Y(y) + (1 - k_X - k_Y)u_X(x)u_Y(y)}$$

and is scaled so that $u(x_0, y_0) = 0$ and $u(x_1, y_1) = 1$. We just need to fix the constants k_X and k_Y. Now, k_X is just $u(x_1, y_0)$, so k_X is the probability which makes you indifferent between (x_1, y_0) and a gamble on (x_0, y_0) and (x_1, y_1). That is

$$(x_1, y_0) \sim^\star (1 - k_X)(x_0, y_0) +_g k_X(x_1, y_1)$$

Similarly, k_Y is just $u(x_0, y_1)$, so k_Y is such that

$$(x_0, y_1) \sim^\star (1 - k_Y)(x_0, y_0) +_g k_Y(x_1, y_1)$$

So, assessing the two probabilities which make you indifferent fixes the constants, and hence the full form of the two-attribute utility function.

Given that we need to consider the above gambles, it will often be necessary to first re-scale the marginal utility functions, $u_X(x)$ and $u_Y(y)$ so that the gambles are more natural quantities to consider. Also note that the interaction term, k, is given by $(1 - k_X - k_Y)/(k_X k_Y)$ and so the sign of k is the same as the sign of $1 - k_X - k_Y$.

1.5.5 Example

Let's return the the example of specifying a joint utility function for cost (C) and journey time (T) for UK travel by coach or train. The joint utility function will be denoted $u(c, t)$.

Suppose that a person judges C and T to be mutually utility independent, and specifies the marginal utilities to have the form $\log(150 - c)$ and $-t$ respectively. In addition, they specify the following equivalences.

$$(0, 10) \sim^\star 0.6(100, 10) +_g 0.4(0, 0)$$
$$(100, 0) \sim^\star 0.7(100, 10) +_g 0.3(0, 0)$$

This is sufficient to determine the complete form of the joint utility function, so what is it? The gambles given tell us that we need to re-scale the marginal utility functions so that $u_C(100) =$

$u_T(10) = 0$ and $u_C(0) = u_T(0) = 1$. So, putting

$$\begin{aligned} u_C(c) =& a\ \log(150 - c) + b \\ u_C(100) = 0 \Rightarrow & a\ \log(50) + b = 0 \\ u_C(0) = 1 \Rightarrow & a\ \log(150) + b = 1 \end{aligned}$$

Solving for a and b and substituting back in gives

$$u_C(c) = \frac{\log(3 - \frac{c}{50})}{\log(3)}$$

Re-scaling the marginal utility for time similarly gives

$$u_T(t) = 1 - \frac{t}{10}$$

Now that our marginals are scaled appropriately, the equivalences given tell us that $k_C = 0.4$ and $k_T = 0.3$, so

$$u(c,t) = 0.4\frac{\log(3 - \frac{c}{50})}{\log 3} + 0.3\left(1 - \frac{t}{10}\right) + (1 - 0.4 - 0.3)\frac{\log(3 - \frac{c}{50})}{\log 3}\left(1 - \frac{t}{10}\right)$$

and since $1 - 0.4 - 0.3 = 0.3 > 0$, C and T are *complimentary*.

1.5.6 Example (ctd.)

Suppose now that you have to make a journey to London by coach or train. The train should take 3 hours and cost £35, and the coach should take 5 hours, and cost £20. From experience, you think that the train will be delayed by 1 hour with probability 1/3, and the coach will be delayed by 1 hour with probability 1/4. Do you go by train (G) or coach (H)?

$$\begin{aligned} G =& \frac{2}{3}(35,3) +_g \frac{1}{3}(35,4) \\ H =& \frac{3}{4}(20,5) +_g \frac{1}{4}(20,6) \end{aligned}$$

so

$$\begin{aligned} u(G) =& \frac{2}{3}u(35,3) + \frac{1}{3}u(35,4) = 0.66 \\ u(H) =& \frac{3}{4}u(20,5) + \frac{1}{4}u(20,6) = 0.61 \end{aligned}$$

Therefore $H \prec^\star G$ so you should go by *train*.

1.5.7 More than two attributes

Suppose now that we wish to specify a utility function over more than two attributes. How can we proceed? In fact, this is a rather complex area, which we won't cover in full detail in this module. We will look briefly at two possibilities, both of which represent simple generalisations of the two-attribute case.

Additive independence

Recall the form of the two-attribute mutually utility independent representation which we used for specification

$$u(x, y) = k_X u_X(x) + k_Y u_Y(y) + (1 - k_X - k_Y)u_X(x)u_Y(y)$$

In the special case where attributes X and Y are preference unrelated, the coefficient of the interaction term $(1 - k_X - k_Y)$ is zero, so $k_X + k_Y = 1$, and the utility function becomes

$$u(x, y) = k_X u_X(x) + k_Y u_Y(y)$$

This has a simple generalisation to three attributes as

$$u(x, y, z) = k_X u_X(x) + k_Y u_Y(y) + k_Z u_Z(z)$$

where $k_X + k_Y + k_Z = 1$. The generalisation to n attributes is

$$u(\mathbf{x}) = \sum_{i=1}^{n} k_i u_i(x_i)$$

where

$$\sum_{i=1}^{n} k_i = 1$$

so providing that all attributes are preference unrelated, the joint utility function can be constructed from marginal specifications and a set of weights k_i, which can be assessed as the utility of obtaining the ith attribute with utility 1 in the marginal, and the value of all other attributes with utility zero.

1.5.8 Example

Four attributes, all preference unrelated, with identical marginals

$$u_i(x_i) = \frac{\log(x_i + 1)}{\log 11},\ i = 1, 2, 3, 4.$$

So, $u_i(0) = 0$ and $u_i(10) = 1$. We know, therefore, that

$$u(x_1, x_2, x_3, x_4) = \sum_{i=1}^{4} k_i \frac{\log(x_i + 1)}{\log 11}$$

but what are the k_i?

$$\begin{aligned} u(10, 0, 0, 0) =& k_1 \\ u(0, 10, 0, 0) =& k_2 \end{aligned}$$

etc. So, for example, k_1 can be assessed by considering your utility for $(10, 0, 0, 0)$. *ie.* k_1 is such that

$$(10, 0, 0, 0) \sim^\star (1 - k_1)(0, 0, 0, 0) +_g k_1(10, 10, 10, 10)$$

If, having assessed the k_i's, you find that they do not sum to one, then the assumption that the attributes are preference unrelated is invalid! However, if the k_i's sum to something very close to one, you may wish to adjust the k_i slightly so that they do. N.B. This is a bit analogous to adjusting subjective probabilities for a partition so that they sum to one.

Of course, the assumption of additive independence is very strong, and we already know from our examination of the two-attribute case that it usually does not hold.

Multiplicative utility

We now consider a possible solution for the case where the k_i's, as assessed above, do not sum to one. This arises from a representation for a particularly strong form of mutual utility independence over the attributes. Consider once again, the two-attribute representation.

$$u(x, y) = k_X u_X(x) + k_Y u_Y(y) + (1 - k_X - k_Y)u_X(x)u_Y(y)$$

$(1 - k_X - k_Y)$ is the constant which deforms the utility surface so that the utility is one at (x_1, y_1), whilst preserving the marginal specifications. It will be more convenient for us to now write this as $k\,k_X k_Y$, where k is a constant which ensures this. So, in this case of two attributes, $k = (1 - k_X - k_Y)/(k_X k_Y)$. Now,

$$\begin{aligned} u(x, y) =& k_X u_X(x) + k_Y u_Y(y) + k\,k_X k_Y u_X(x)u_Y(y) \\ \Rightarrow 1 + k\,u(x, y) =& 1 + k\,k_X u_X(x) + k\,k_Y u_Y(y) + k^2 k_X u_X(x) k_Y u_Y(y) \\ =& (1 + k\,k_X u_X(x))(1 + k\,k_Y u_Y(y)) \end{aligned}$$

Now this has a simple generalisation to three attributes as

$$1 + k\,u(x, y, z) = (1 + k\,k_X u_X(x))(1 + k\,k_Y u_Y(y))(1 + k\,k_Z u_Z(z))$$

Note that as before, $u(x_0, y_0, z_0) = 0$, $u(x_1, y_0, z_0) = k_X$, $u(x_0, y_1, z_0) = k_Y$ and $u(x_0, y_0, z_1) = k_Z$. So, k_X, k_Y and k_Z can be specified as before. k is a constant ensuring that $u(x_1, y_1, z_1) = 1$. Evaluating our utility at (x_1, y_1, z_1) gives

$$1 + k = (1 + k\,k_X)(1 + k\,k_Y)(1 + k\,k_Z)$$

So, once k_X, k_Y and k_Z have been specified, k can be obtained as the solution to the above equation. Although it looks like a cubic, it boils down to a quadratic very simply.

1.5.9 Example

Three attributes X, Y and Z with marginal utility functions

$$u_X(x) = \frac{\log(x + 1)}{\log 11}$$

with $u_Y(y)$ and $u_Z(z)$ defined identically. The constants $k_X = 1/3$, $k_Y = 1/4$ and $k_Z = 1/4$ are specified. Now, $k_X + k_Y + k_Z = 5/6$, so additive independence cannot be assumed. Using the multiplicative form of the utility function, what is the utility function over the three attributes?

k is a solution to the equation

$$\begin{aligned}1+k =& (1+\frac{1}{3}k)(1+\frac{1}{4}k)^2\\ \Rightarrow 48+48k =& (3+k)(4+k)^2\\ \Rightarrow 48+48k =& 48+24k+3k^2+16k+8k^2+k^3\\ \Rightarrow 48 =& 24+3k+16+8k+k^2\\ \Rightarrow 0 =& k^2+11k-8\end{aligned}$$

So $k = -11.684$ or 0.684. Obviously, we want the root 0.684. So, the utility function is (equivalent to)

$$u(x, y, z) = \left(1 + \frac{0.684}{3}\frac{\log(x+1)}{\log 11}\right) \times \left(1 + \frac{0.684}{4}\frac{\log(y+1)}{\log 11}\right) \times \left(1 + \frac{0.684}{4}\frac{\log(z+1)}{\log 11}\right)$$

If the sum of the k_i's is less than one, we want a positive root, and if it is bigger than one, we want a negative root. The required root will usually be the closest root to zero satisfying these constraints.

1.5.10 Full multiplicative form

For n attributes, the multiplicative utility function takes the form

$$1 + k\,u(\mathbf{x}) = \prod_{i=1}^{n}(1 + k\,k_i u(x_i))$$

where the k_i have the usual interpretation, and k is a root of the equation

$$1 + k = \prod_{i=1}^{n}(1 + k\,k_i)$$

Whilst there are generalisations of this form for utility, in most situations this form has sufficient flexibility to be able to model most multi-attribute functions reasonably well.

It is here that we shall leave the topic of multi-attribute utility, and turn our attention from multiple attributes to the particularly thorny problem of multiple decision makers.

1.6 Group decision making

1.6.1 Introduction

So far, we have been concerned with the problem of how an individual decision maker (*"you"*) can make decisions under uncertainty. We have seen that this is necessarily a subjective process, as individuals assess uncertainties and rewards differently. Very often, however, decisions need to be made which are "best" for a *group* of individuals. The group may be a couple, a family, people attending a meeting, a company, a nation, or even the world!

Recall that one of the first things we did when starting to think about decision making under uncertainty was to establish a preference ordering for a set of rewards. Similarly, we would like to be able to establish a preference ordering for a group of individuals, either as a goal in its own right, or as a precursor to further decision modelling.

A natural starting point is to see if we can find a way of combining the preferences of individuals into a preference ordering for the group.

Notation

We already have a symbol representing preference for an individual ($\preceq^\star$). For a collection of decision makers, $i = 1, \ldots, n$, we will use $\preceq^\star_i$ for the preference of individual i, and $\preceq^\star_G$ for the preference of the group.

We need a way of combining individual preferences into group preferences in a sensible way. For example, if $B \preceq^\star_i A$ is true for every individual i, we will probably want $B \preceq^\star_G A$ to hold for the group.

1.6.2 Majority rule

Majority rule is a simple way of combining individual preferences in order to give the preferences of a group. It simply states that if more people prefer A to B, then the group should prefer A to B. *ie.* If more individuals hold $B \prec^\star_i A$ than $A \prec^\star_i B$, then we must have $B \prec^\star_G A$. Majority rule sounds great, but it doesn't work! Consider the following group of individuals, and their ratings of political parties.

Conservative	$\prec^\star_1$	Labour	$\prec^\star_1$	Liberal
Labour	$\prec^\star_2$	Liberal	$\prec^\star_2$	Conservative
Liberal	$\prec^\star_3$	Conservative	$\prec^\star_3$	Labour

According to majority rule, we must have Conservative $\prec^\star_G$ Labour $\prec^\star_G$ Liberal $\prec^\star_G$ Conservative! So $\prec^\star_G$ isn't transitive! Clearly this is unacceptable from the point of view of making decisions based on group preferences. We need a way of combining preferences in such a way that the resulting preference ordering is well behaved — it must be defined, and it must be transitive. Such well-behaved rules for combining preferences are known as *Social Welfare Functions*.

1.6.3 Social Welfare Functions

A *Social Welfare Function* (SWF), is a function which operates on preference orderings of individuals, and returns a group preference profile, $\preceq^\star_G$, which is always defined, complete and transitive.

Example 1

Ignore all of the individual preferences, and simply return the rewards in alphabetical order.

This is undesirable because we want the group preferences to somehow relate to the preferences of the individuals.

Example 2

Return the preference ordering of one of the individuals, ignoring all others.

This is undesirable because only the preferences of one individual are taken into account.

1.6.4 Constraints on SWFs

Below are four "reasonable" conditions which we might want our SWF to obey.

Unrestricted domain (U) For any set of individual preferences over rewards A and B, the SWF will return one of $A \prec^\star B$, $B \prec^\star A$ or $A \sim^\star B$.

No dictatorship (D) There is no individual i for whom $\preceq_i^\star$ automatically becomes $\preceq_G^\star$.

Pareto's condition (P) If every member of the group holds $A \prec_i^\star B$ then the group will hold $A \prec_G^\star B$.

Independence of irrelevant alternatives (I) If a reward is deleted, then individual preferences over the remaining rewards will not change, and so group preferences over the remaining rewards should not change.

1.6.5 Arrow's impossibility theorem

Theorem

Provided there are at least three rewards and at least two individuals, then there is no SWF satisfying all of (U), (P), (D) and (I).

Note 1

This means that for any SWF, we can find a collection of individual preferences which will cause at least one of the conditions to be broken.

Note 2

The theorem requires there to be at least three rewards. If there are only two rewards, then simple *majority rule* satisfies all of (U), (P), (D) and (I)!

Note 3

We will prove the theorem by showing that if we have a SWF satisfying (U), (P) and (I), then there is a dictator, so (D) cannot hold.

Before we can attempt to prove the theorem, we need some concepts, and some preliminary results.

1.6.6 Definitions

A subgroup, V of individuals is said to be *decisive* for reward A over reward B if whenever we have

$$B \prec_i^\star A, \quad \forall i \in V$$

we have $B \prec_G^\star A$, irrespective of the preferences of those individuals not in V (for majority rule, any group containing over half of the individuals is decisive for every pair of rewards).

N.B. An *individual* who is decisive over every pair of rewards is a *dictator*.

A subgroup V of individuals is said to be *almost decisive* for reward A over reward B if whenever we have

$$B \prec_i^\star A, \quad \forall i \in V$$

and

$$A \prec_j^\star B, \quad \forall j \notin V$$

we have $B \prec_G^\star A$.

A *minimal almost decisive* (MAD) group, V, for reward A over reward B, is an almost decisive group for A over B, such that no subgroup of V is almost decisive for *any* C over *any* D.

Lemma 1 *Given a SWF satisfying (U), (P) and (I), there exists a MAD group.*

Clearly given any A and B, the whole group is decisive, by (P), so just keep discarding individuals until we reach a MAD group for some C over some D.

Lemma 2 *Given a SWF satisfying (U), (P) and (I), all MAD groups consist of a single individual.*

We know that MAD groups exist, so choose one, V, and suppose it is almost decisive for A over B. Suppose that V contains more than one person, and choose an individual, j from V. Let W be the remaining members of V and let U be everyone not in V. Let C be any reward other than A or B, and suppose that the individuals hold the following preferences.

$$\begin{aligned} &C \prec_j^\star B \prec_j^\star A \\ &B \prec_i^\star A \prec_i^\star C, \quad \forall i \in W \\ &A \prec_k^\star C \prec_k^\star B, \quad \forall k \in U \end{aligned}$$

Now $B \prec_i^\star A, \forall i \in V$, and $A \prec_k^\star B, \forall k \notin V$, so

$$B \prec_G^\star A$$

as V is MAD for A over B. Now $\forall i \in W$ we have $B \prec_i^\star C$, and $\forall k \notin W$ we have $C \prec_k^\star B$. But since $W \subset V$ and V is MAD, W can't be almost decisive for C over B, so we must have

$$C \preceq_G^\star B$$

but since group preference is transitive, we know that

$$C \prec_G^\star A$$

However, this would mean that j was almost decisive for C over A, as $C \prec_j^\star A$ and $A \prec_i^\star C, \forall i \neq j$, which cannot be true, as V is MAD. Therefore our supposition that V contained more than one person must have been false.

Lemma 3 *If we have a SWF satisfying (U), (P) and (I), and a MAD group for A over B consisting of a single individual j, then j is decisive for every reward C over every reward D.* ie. *j is a dictator.*

Step 1: Show that j is decisive for A over any $D \neq B$.

Suppose that $D \prec_j^\star B \prec_j^\star A$ and let U be the group of everyone other than j. Suppose further that $\forall i \in U$, $A \prec_i^\star B$ and $D \prec_i^\star B$ (note that this says nothing about the relative preferences for A and D). Now since j is almost decisive for A over B, we have

$$B \prec_G^\star A$$

and since everyone has $D \prec_i^\star B$, from (P) we know that

$$D \prec_G^\star B$$

and since group preference is transitive, we know

$$D \prec_G^\star A$$

so j is decisive for A over D.

Step 2: Show that j is decisive for any $C \neq A$ over any $D \neq A$.

Suppose that $D \prec_j^\star A \prec_j^\star C$ and that $\forall i \in U$, $A \prec_i^\star C$ (note that this says nothing about the relative preferences for C and D). Now since j is decisive for A over D we have

$$D \prec_G^\star A$$

Also, (P) gives

$$A \prec_G^\star C$$

so

$$D \prec_G^\star C$$

and so j is decisive for C over D.

Step 3: Show that j is decisive for any C over A.

Suppose that $A \prec_j^\star D \prec_j^\star C$ and that $\forall i \neq j$ we have $A \prec_i^\star D$. Then we get $D \prec_G^\star C$ and $A \prec_G^\star D$ which together give $A \prec_G^\star C$, so j is decisive for C over A.

Step 4: Show that j is decisive for A over B.

Suppose that $B \prec_j^\star D \prec_j^\star A$ and that $\forall i \neq j$ we have $B \prec_i^\star D$. Then we get $D \prec_G^\star A$ and $B \prec_G^\star D$ which together give $B \prec_G^\star A$, so j is decisive for A over B.

Step 5: Steps 1, 2, 3 and 4 together prove the Lemma.

Proof of Theorem

Let us assume we have a SWF satisfying (U), (P) and (I).

- By Lemma 1, there is a MAD group for some A over some B.
- Lemma 2 tells us that this MAD group consists of a single individual j.
- Lemma 3 tells us that this individual is a dictator.
- Consequently, (D) cannot hold for the SWF.

This proves the Theorem.

1.6.7 Utilitarianism

As we have already seen, preference orderings do not capture the opinions of a decision maker sufficiently precisely to allow preferences over gambles to be determined. Maybe this is the problem with attempting to combine preferences of individuals in order to obtain the preferences of the group. Perhaps we should really be thinking about combining *utilities*.

Utilitarianism is a sociological philosophy which suggests that

> *"Social choice should attempt to maximise the well-being of the citizenry".*

This sounds laudable enough, but what is meant by "well-being"? Is it happiness? Pleasure? The philosophy is: given a choice between social options, society should rank highest the option which gives the most "pleasure" to its citizens. Is there any justification for using utility as a measure of "pleasure", and adding up utilities across the population? It turns out that the answer is "Yes"!

Suppose we have a society consisting of m citizens, who need to rank n social choices, $R_1, R_2, \ldots, R_n$. Each citizen in society is assumed to be individually rational, and hence can specify utilities for the rewards

$$u_i(R_j), \quad 1 \leq i \leq m, \quad 1 \leq j \leq n.$$

Each individual scales their utilities so that they assign 0 to their least favourite reward, and 1 to their favourite reward. Thus, all utilities specified are between 0 and 1 inclusive. For each reward, we can form a vector of the utilities specified by the citizens

$$\mathbf{u}(R_j), \quad 1 \leq j \leq n.$$

Thus, *society* specifies a *vector* of utilities for each reward.

In order to be able to think about a utility function for society, we need to think of society as being a rational individual, so with this in mind, we introduce the concept of a *planner*. The *planner*, P, must always take the decision for society. P is individually rational, and so has a utility function, $w(\cdot)$, which will be based on society's vector function, $\mathbf{u}(\cdot)$.

P is a selfless individual, who simply wishes to make the best choice for society. We therefore assume that P obeys the following two conditions for *"social rationality"*:

Anonymity (A) P would have the same utility under any permutation of the individuals.

Strong Pareto (SP) If every citizen is indifferent between two rewards, so is P. If no citizen holds $A \prec^\star B$, and some citizens hold $B \prec^\star A$, then the planner holds $B \prec^\star A$.

Is there a way to satisfy these constraints? Yes! Do we need any more constraints? No!

> (A) and (SP) uniquely determine $w(\cdot)$.

First we note that $w(\cdot)$ must be some function of $\mathbf{u}(\cdot)$. That is,

$$w(R_j) = f(\mathbf{u}(R_j))$$

for some function $f(\cdot)$, defined over m-dimensional vectors. For example, suppose that for rewards R_i and R_j, society specifies

$$\begin{aligned}\mathbf{u}(R_i) =& \mathbf{u}(R_j)\\ &\iff u_k(R_i) = u_k(R_j), \quad 1 \le k \le m\\ &\iff R_i \sim_k^\star R_j\end{aligned}$$

Now this and (SP) together give

$$\mathbf{u}(R_i) = \mathbf{u}(R_j) \Rightarrow R_i \sim_P^\star R_j \iff w(R_i) = w(R_j)$$

Thus, $\mathbf{u}(\cdot)$ determines $w(\cdot)$. So, if $w(R_j) = f(\mathbf{u}(R_j))$, what is $f(\cdot)$?

Theorem

If P obeys (A) and (SP), then $f(\cdot)$ must take the form

$$f\begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = x_1 + x_2 + \cdots + x_m$$

ie. *for any reward, R_j,*

$$w(R_j) = u_1(R_j) + u_2(R_j) + \cdots + u_m(R_j)$$

Proof

Step 1: Show that the given $f(\cdot)$ satisfies (A) and (SP).

Clearly (A) is satisfied, as addition is commutative.

If every individual is indifferent between A and B, then $u_k(A) = u_k(B), 1 \le k \le m$. So

$$\begin{aligned}w(A) =& u_1(A) + \cdots + u_m(A)\\ =& u_1(B) + \cdots + u_m(B)\\ =& w(B)\end{aligned}$$

and P is indifferent between A and B. Also, if no-one holds $u_k(A) < u_k(B)$ and some people hold $u_k(A) > u_k(B)$, we will have $w(A) > w(B)$ and so P will prefer A to B. Therefore we have (SP).

Step 2: Show that the given $f(\cdot)$ is the only possible function.

We will only prove this for a society consisting of two citizens. The general proof is similar but a bit messy. As $w(\cdot)$ is a utility, it is unique only up to a positive linear scaling, so we will fix the lowest value at 0. Consider any reward, R_0, which is rated worst by both citizens. By (SP), this is rated worst by P, so $w(R_0) = 0$.

Now consider some other reward, R, such that $\mathbf{u}(R) = \binom{x_1}{x_2}$, and think about the gamble, G, where

$$G = p\,R +_g (1-p)R_0$$

Clearly

$$\begin{aligned} u_1(G) =& p\, u_1(R) + (1-p) u_1(R_0) \\ =& p\, x_1 \end{aligned}$$

and similarly $u_2(G) = p\, x_2$, so

$$\mathbf{u}(G) = \begin{pmatrix} p\, x_1 \\ p\, x_2 \end{pmatrix} = p \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Also, $w(\cdot)$ is a utility function, so

$$\begin{aligned} w(G) =& p\, w(R) + (1-p) w(R_0) \\ =& p\, w(R) = p\, f(\mathbf{u}(R)) = p\, f \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \end{aligned}$$

But

$$w(G) = f(\mathbf{u}(G)) = f\left(p \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right)$$

Therefore

$$f\left(p \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right) = p\, f \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \qquad \text{(*)}$$

Again, utility is unique only up to a positive linear scaling, and we have so far fixed only one value, so we are still free to fix another. Suppose there was a reward, R_*, rated best by the first citizen, and worst by the second. We will fix P's utility for such a reward to be 1. That is $w(R_*) = 1 = f(\mathbf{u}(R_*)) = f\binom{1}{0}$. Therefore $f\binom{1}{0} = 1$, and by (A), we must also have $f\binom{0}{1} = 1$. Now $(*)$ gives $f\binom{x}{0} = x\, f\binom{1}{0} = x$ and similarly

$$f \begin{pmatrix} 0 \\ x \end{pmatrix} = x \qquad (\dagger)$$

Now consider the gamble

$$H = \frac{1}{2} A +_g \frac{1}{2} B$$

where $\mathbf{u}(A) = \binom{x_1}{0}$ and $\mathbf{u}(B) = \binom{0}{x_2}$. Clearly,

$$\begin{aligned} u_1(H) =& \frac{1}{2} u_1(A) + \frac{1}{2} u_1(B) = \frac{1}{2} x_1 \\ u_2(H) =& \frac{1}{2} u_2(A) + \frac{1}{2} u_2(B) = \frac{1}{2} x_2 \end{aligned}$$

So, $\mathbf{u}(H) = \frac{1}{2}\binom{x_1}{x_2}$ and

$$\begin{aligned} w(H) =& f(\mathbf{u}(H)) \\ =& f\left(\frac{1}{2} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right) \\ =& \frac{1}{2} f \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \end{aligned}$$

by $(*)$. But $w(\cdot)$ is a utility, so

$$\begin{aligned} w(H) =& \frac{1}{2}w(A) + \frac{1}{2}w(B) \\ =& \frac{1}{2}f(\mathbf{u}(A)) + \frac{1}{2}f(\mathbf{u}(B)) \\ =& \frac{1}{2}f\begin{pmatrix} x_1 \\ 0 \end{pmatrix} + \frac{1}{2}f\begin{pmatrix} 0 \\ x_2 \end{pmatrix} \\ =& \frac{1}{2}x_1 + \frac{1}{2}x_2 \qquad \text{by } (\dagger) \end{aligned}$$

So $\frac{1}{2}f\binom{x_1}{x_2} = \frac{1}{2}x_1 + \frac{1}{2}x_2$ and hence

$$f\left(\begin{array}{c} x_1 \\ x_2 \end{array}\right) = x_1 + x_2$$

This proves the theorem. □

1.6.8 Example

The local council have enough money to build only one of the following: Swimming pool (S), Library (L), Car park (C), Museum (M) or Nothing, (N).
The town has two citizens. Citizen 1 assigns:
$u_1(S) = 1$, $u_1(L) = 0.5$, $u_1(C) = 0$, $u_1(M) = 0.5$, $u_1(N) = 0$.
Citizen 2 assigns:
$u_2(S) = 0.1$, $u_2(L) = 0$, $u_2(C) = 1$, $u_2(M) = 0$, $u_2(N) = 0$.
The town planner therefore assigns:
$w(S) = 1.1$, $w(L) = 0.5$, $w(C) = 1$, $w(M) = 0.5$, $w(N) = 0$.
Consequently, the town's preferred option is the Swimming pool.

1.7 Summary and conclusions

1.7.1 Summary of Decision Making

In this part of the course, we have seen how it is possible to quantify uncertainty using subjective probability, and how one can use preferences over gambles (uncertain rewards) in order to determine a utility function over a set of rewards. We have seen how this utility function can be used in order to assist decision making under uncertainty. We looked in particular at the case of monetary rewards, and how one can specify risk-averse utility for money using parameterised classes of concave functions.

Next we looked at specifying multi-attribute utility — that is, utilities for rewards which have more than one (continuous) attribute associated with them. Finally we have examined the problem of group decision making, and seen that there is no satisfactory way of combining preferences of individuals into the preferences of the group, but that there is a sensible way of combining utilities, related to the concept of utilitarianism.

1.7.2 Bayes linear methods

In the next part of the course, we will start to get a little more "statistical", and think about quantifying personal opinion about random quantities, and how we can use data in order to update that opinion as new information comes to light.

Chapter 2

Bayes Linear Methods

2.1 Introduction

In the "Decision making" part of the course, we looked in some detail at the problem of quantifying uncertainty for random events using subjective probability, and preferences for random rewards (gambles) using utility. We then looked at how those specifications could be exploited, in order to aid decision making under uncertainty.

In this part of the course, we will look at how we can quantify our uncertainty for random quantities (some of which are to be observed), and then exploit those specifications in order to aid inference for the unobserved quantities.

In general, the "Bayesian" analysis proceeds as follows:

1. An investigator is interested in the value of some uncertain quantities which cannot be measured directly.
2. The investigator conceives of an experiment resulting in measurements of quantities which are in some way informative for the quantities of interest.
3. The investigator quantifies his/her opinions regarding the value of the quantities of interest, the quantities to be measured, and the relationships between them.
4. The experiment is carried out, and the measurable quantities are measured.
5. The data are used to "update" the opinions of the investigator regarding the quantities of interest.

For those people who didn't do the Bayesian Statistics reading module (most of you), the idea that inference might be a subjective procedure (dependent on the opinions of the investigator before the experiment takes place) probably seems a bit strange, if not a little disturbing!

However, a moment's thought reveals that such a procedure (in a more or less formal sense) is at the heart of *the scientific method*. In any case, so-called "objective" statistical analyses are always inherently subjective, dependent as they are on modelling and distributional assumptions, choices of null hypotheses, critical values *etc.*

2.1.1 Notation

We suppose that we are interested in a collection, B, of random quantities

$$B = \{B_1, B_2, \dots , B_m\}$$

which we cannot observe directly. However, we can conduct an experiment which will yield a collection, D, of observations

$$D = \{D_1, D_2, \ldots D_n\}$$

which are in some way informative for B. The collection of quantities in D will at some point be observed, but at present are unknown, and hence (in some sense) random. We let

$$A = B \cup D$$

be the collection of all random quantities in our problem.

2.1.2 Example

Suppose we are interested in the average weekly alcohol consumption of Newcastle University students. We will call this quantity B_1, our quantity of interest. So here, $B = \{B_1\}$.

B_1 would be very difficult to find out exactly, but it would be quite easy to get 25 students to record what they drank for one particular week. This would result in measurements $D_1, D_2, \ldots, D_{25}$, which are currently unknown. So here, $D = \{D_1, D_2, \ldots, D_{25}\}$. Our set of all quantities, A, is just

$$A = \{B_1, D_1, D_2, \ldots, D_{25}\}.$$

Even before we carry out the experiment, we have some idea about the likely value of B_1 and the range of the responses, $D_1, \ldots, D_{25}$. Ultimately, we could carry out the experiment, and obtain the responses, $d_1, d_2, \ldots, d_{25}$, and (somehow) use them in order to update what we think about B_1.

2.2 Subjective expectation

2.2.1 Introduction

We must begin our analysis by quantifying our opinions and uncertainties about the random quantities in our problem. We know how we can use subjective probability in order to quantify opinion regarding uncertain events, but we need a generalisation of this concept for random quantities. The key to doing this is to use *subjective expectation* (sometimes known as *prevision*).

2.2.2 Example

Let X denote the amount of rain (in millimeters) which will fall in Newcastle during the calendar month April 2000. What is the meaning of the statement: " $\mathrm{E}(X) = 160$"? Once again, it clearly means something, but it is hard to give a classical or frequentist interpretation of the statement.

The quantity X is unknown, and hence random, so what is its *expectation*, $\mathrm{E}(X)$? A natural first attempt at defining $\mathrm{E}(X)$ subjectively is to say it is the *price*, £r which you consider fair for the gamble which will pay £X when X becomes known.

However, we know from the first part of the course that this will only work if your utility for money is linear. We also know that utility for money is not linear for most people over anything other than a small range of values. One way around this is to define it as the value r you would quote if you were to be subjected to a *loss*, L of

$$L = k(X - r)^2$$

where k is some constant ensuring that the loss is in units of utility. Usually, k will be some small monetary value, *eg.* £0.01, ensuring that the range of possible losses is sufficiently small for your utility for money to be approximately linear. This way, your loss, L, can be thought of as a loss of utility.

Since we make decisions by *maximising* expected utility, we will also make decisions by *minimising* expected loss, provided that the loss is measured on a utility scale. What value of r will minimise the expected loss, $\mathrm{E}(L)$?

Differentiate and equate to zero:

$$\begin{aligned}\frac{d}{dr}\ \mathrm{E}(L) &= 0 \\ \Rightarrow \frac{d}{dr}\ \mathrm{E}\big(k[X-r]^2\big) &= 0 \\ \Rightarrow -2k\ \mathrm{E}(X-r) &= 0 \\ \Rightarrow r &= \mathrm{E}(X)\end{aligned}$$

so in order to be *coherent*, you *must* choose r to be $\mathrm{E}(X)$.

Of course in practice, thinking about quadratic loss functions is difficult, and subjective expectations will usually be specified as some sort of point estimate or "best guess" for X. The quadratic loss function is just the way in which "poor guesses" are penalised, and ensures that choosing the expected value of X is the optimal strategy.

2.2.3 Axioms and coherence

When we defined subjective probability, we showed that for rational individuals, subjective probability must satisfy the usual axioms of probability that we are all familiar with. Expectation also has some important properties which we are all used to relying on (such as its linearity), and so we need to demonstrate that for rational individuals, these familiar properties still hold.

2.2.4 Linearity

The properties of expectation that make it so useful and convenient to work with all derive from its linearity. For example:

$$\mathrm{E}(aX + bY + c) = a\ \mathrm{E}(X) + b\ \mathrm{E}(Y) + c$$

If we can't rely on this property for subjective expectation, it will not be of much use to us. Fortunately, for rational individuals, coherence forces the above condition to hold. We look now at the key coherence requirements for expectation specification.

2.2.5 Convexity

Property 1

For a rational individual,

$$\inf X \leq\ \mathrm{E}(X) \leq \sup X$$

That is, if X is bounded above or below, then these bounds constrain the possible assignments for $\mathrm{E}(X)$. We will just examine the case $\inf X \leq \mathrm{E}(X)$, as the case $\mathrm{E}(X) \leq \sup X$ follows trivially by looking at $-X$.

Suppose that you assign $\mathrm{E}(X) = e_X$ where $e_X < \inf X$, then we must show that you are incoherent. Put $e_X^\star = \inf X$. Now

$$L(r) = k(X - r)^2$$

is the loss to be imposed. You have chosen to suffer the loss

$$\begin{aligned} L(e_X) &= k(X - e_X)^2 \\ &= k(X - e_X^\star + e_X^\star - e_X)^2 \\ &= k[(X - e_X^\star)^2 + (e_X^\star - e_X)^2 + 2(X - e_X^\star)(e_X^\star - e_X)] \\ &> k[(X - e_X^\star)^2 + (e_X^\star - e_X)^2] \\ &> k(X - e_X^\star)^2 \\ &= L(e_X^\star) \end{aligned}$$

That is, $L(e_X)$ must be bigger than $L(e_X^\star)$ *whatever value of X occurs*. Thus the individual is facing a certain loss of at least $k(e_X - e_X^\star)^2$ more than necessary, and hence is incoherent.

2.2.6 Scaling

Property 2

For a rational individual,

$$\mathrm{E}(\lambda X) = \lambda\, \mathrm{E}(X)\,.$$

Putting $Y = \lambda X$, we need to specify $\mathbf{e} = (e_X, e_Y)$ for $\mathbf{W} = (X, Y)$, and we want to show that a coherent individual will always make specifications so that $e_Y = \lambda e_X$. Suppose he does not, so that $\mathbf{e}$ is not on the line $Y = \lambda X$. Let $\mathbf{e}^\star$ be the point on the line closest to $\mathbf{e}$ (so that $\mathbf{e}^\star$ is the orthogonal projection of $\mathbf{e}$ onto the line). Then

$$\begin{aligned} L(\mathbf{e}) &= k(X - e_X)^2 + k(Y - e_Y)^2 \\ &= k\|\mathbf{W} - \mathbf{e}\|^2 \\ &= k\|\mathbf{W} - \mathbf{e}^\star + \mathbf{e}^\star - \mathbf{e}\|^2 \\ &= k[\|\mathbf{W} - \mathbf{e}^\star\|^2 + \|\mathbf{e}^\star - \mathbf{e}\|^2 + 2(\mathbf{W} - \mathbf{e}^\star) \cdot (\mathbf{e}^\star - \mathbf{e})] \\ &= k[\|\mathbf{W} - \mathbf{e}^\star\|^2 + \|\mathbf{e}^\star - \mathbf{e}\|^2] \\ &< k\|\mathbf{W} - \mathbf{e}^\star\|^2 \\ &= L(\mathbf{e}^\star) \end{aligned}$$

By choosing $\mathbf{e}$ rather than $\mathbf{e}^\star$, the individual faces a loss of $k\|\mathbf{e}^\star - \mathbf{e}\|^2$ more than necessary, which is incoherent.

2.2.7 Additivity

Property 3

For a rational individual,

$$\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)\,.$$

Putting $Z = X + Y$, we need to specify $\mathbf{e} = (e_X, e_Y, e_Z)$ for $\mathbf{W} = (X, Y, Z)$. We want to show that a coherent individual will always make specifications such that $e_Z = e_X + e_Y$. Suppose that she does not, so that $\mathbf{e}$ does not lie on the plane $Z = X + Y$. Let $\mathbf{e}^*$ be the orthogonal projection of $\mathbf{e}$ onto the plane. Exactly the same reasoning as before shows that $L(\mathbf{e}) > L(\mathbf{e}^\star)$, and so she is incoherent.

2.2.8 Expectation as a primitive concept

In the first part of the course we defined subjective probability. Conventional treatments of probability theory start with probability and use it to define random quantities, and then define expectation in terms of probability. So, we could have defined subjective expectation for a discrete random quantity as

$$\mathrm{E}(X) = \sum_{i=1}^{\infty} x_i \,\mathrm{P}(X = x_i)$$

where $\mathrm{P}(X = x_i)$ is subjective probability for the event $(X = x_i)$. Indeed, this matching up of expectation and probability so that they are mutually consistent is important (in fact, it is a coherence requirement).

However, in Bayes linear methods, we take the view that assigning probabilities to all of the possible outcomes of a random quantity is too arduous a task to do by default, and so we make our definition of subjective expectation the fundamental quantity of interest.

2.2.9 Probability via expectation

If probability and expectation need to be consistent with one another, and we are taking subjective expectation to be the fundamental quantity of interest, how can we get back a theory of subjective probability? In fact, it's rather easy! Associate with every event, F, an *indicator* random variable, I_F as the random quantity

$$I_F = \begin{cases} 1 & \text{if } F \text{ is true} \\ 0 & \text{if } F \text{ is false} \end{cases}$$

Then I_F is a random quantity for which we can specify a subjective expectation, $\mathrm{E}(I_F)$, and so we define subjective probability via

$$\mathrm{P}(F) = \mathrm{E}(I_F)$$

since

$$\begin{aligned} \mathrm{E}(I_F) &= 0 \times \mathrm{P}(I_F = 0) + 1 \times \mathrm{P}(I_F = 1) \\ &= \mathrm{P}(I_F = 1) \\ &= \mathrm{P}(F) \end{aligned}$$

We can now see how our axioms for subjective probability follow directly from our properties of subjective expectation.

First note that for any event, F, $\inf I_F \geq 0$ and $\sup I_F \leq 1$. Property 3 then gives $0 \leq \mathrm{E}(I_F) \leq 1$ and hence $0 \leq \mathrm{P}(F) \leq 1$. Now consider two disjoint events F and $F^\star$. $F \cap F^\star = \emptyset \Rightarrow I_{F \cup F^\star} = I_F + I_{F^\star}$. Consequently,

$$\begin{aligned} \mathrm{P}(F \cup F^\star) &= \mathrm{E}(I_{F \cup F^\star}) \\ &= \mathrm{E}(I_F + I_{F^\star}) \\ &= \mathrm{E}(I_F) + \mathrm{E}(I_{F^\star}) \\ &= \mathrm{P}(F) + \mathrm{P}(F^\star) \end{aligned}$$

To tie this all up properly, there are a few more details to fill in. However, we won't concern ourselves unnecessarily with such philosophical niceties. All we need to understand is that we can consider subjective expectation to be a fundamental quantity, which we specify directly, and that if we need to think about probability for any reason, we can define it simply in terms of expectation.

2.2.10 Subjective variance and covariance

Subjective expectation allows us to make a statement about the *location* of a quantity for which we have uncertainty, but it tells us nothing about the *spread* of the quantity — the *degree of uncertainty* we have about the value. Nor does it tell us anything about the relationship between different random quantities.

Fortunately, expectation is a very powerful tool, and we can use it to define subjective variance and covariance directly, as

$$\begin{aligned} \mathrm{Var}\,(X) &= \mathrm{E}\big([X - \mathrm{E}(X)]^2\big) \\ &= \mathrm{E}\big(X^2\big) - \mathrm{E}(X)^2 \\ \mathrm{Cov}\,(X, Y) &= \mathrm{E}([X - \mathrm{E}(X)][Y - \mathrm{E}(Y)]) \\ &= \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) \end{aligned}$$

where $\mathrm{E}(\cdot)$ denotes subjective expectation. However, assessing these directly from the definitions is difficult, so we use more convenient methods for assessment.

Assessment of variance

In order to assess the variance of a random quantity, X, we think about the range of likely values of X. If we think about an interval, $[a, b]$ for which we are 95% certain it contains X, then assuming that X is vaguely Normal, $b - a$ will be approximately 4 standard deviations, and so putting

$$\mathrm{Var}\,(X) = \left(\frac{b - a}{4}\right)^2$$

should be fairly reasonable.

Assessment of covariance

If we wish to think about the assessment of covariance between two random quantities, X and Y, we need to think about how knowledge of one gives information about the other.

This can be done by thinking about your uncertainty for Y in terms of a 95% interval which has length l, and then thinking about the length of interval $l^\star$ you would have if you knew a particular value of X (*ie.* $X = x$). From standard bivariate Normal theory, we know that

$$\frac{l^\star}{l} = \frac{\sigma_{Y|X}}{\sigma_Y} = \sqrt{1-\rho^2}$$

where ρ is the *subjective correlation* between X and Y, so

$$\text{Cov}\,(X,Y) = \sqrt{\text{Var}\,(X)\,\text{Var}\,(Y)\left[1 - \frac{{l^\star}^2}{l^2}\right]}$$

should be reasonable if the joint distribution of X and Y is vaguely bivariate Normal. In fact, this assessment procedure will make a lot more sense later, when we know a bit more about Bayes linear methods! In any case, it is important to distinguish between the formal definitions of subjective variance and covariance, and the "rough and ready" techniques you might employ in order to specify them.

2.3 Belief specification

2.3.1 Introduction

For the Bayes linear analysis we are going to carry out, we need to specify our prior beliefs about all of the quantities in the problem. So, if we have $B = \{B_1, \ldots, B_m\}$, $D = \{D_1, \ldots, D_n\}$ and $A = B \cup D$, then we need to specify an expectation for every quantity in A, a variance for every quantity in A, and a covariance between every pair of quantities in A.

This sounds very difficult when presented in this general way, and in fact is quite demanding in the most general situation. However, for the standard "statistical" situation (simple random sampling, *etc.*), the symmetries present make specification much more straight forward than you might think. We'll come back to this later — for now, assume that someone else has done the specification for us. Bayes linear methods are a set of "tools" for analysing and using these *second-order* specifications for inference.

2.3.2 Belief structures

A *belief structure* is simply the term given to the collection of all prior specifications relating the quantities in A. It is convenient to arrange these in a tabular form. For example, if $B = \{B_1\}$, $D = \{D_1, D_2\}$, $\text{E}(B_1) = 1$, $\text{E}(D_1) = 2$, $\text{E}(D_2) = 3$, $\text{Var}\,(B_1) = 4$, $\text{Var}\,(D_1) = 5$, $\text{Var}\,(D_2) = 6$, $\text{Cov}\,(B_1, D_1) = 1$, $\text{Cov}\,(B_1, D_2) = 1$, $\text{Cov}\,(D_1, D_2) = 2$, then the belief structure can be presented as a table.

B_1	1	4		
D_1	2	1	5	
D_2	3	1	2	6

The column of numbers in the middle is the *expectation vector* for the quantities in A, and the collection of numbers on the right is the lower triangle of the *variance matrix* for the quantities in A. In order to be able to deal with these, we need to be familiar with the analysis of vector random quantities.

2.3.3 Vector random quantities

The expectation of a vector of random quantities is defined to be the vector of expectations of the components. That is

$$\mathrm{E}\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} \mathrm{E}(X_1) \\ \vdots \\ \mathrm{E}(X_n) \end{pmatrix}$$

It is clear that if $\mathbf{X}$ and $\mathbf{Y}$ are random vectors, then $\mathrm{E}(\mathbf{X}+\mathbf{Y}) = \mathrm{E}(\mathbf{X}) + \mathrm{E}(\mathbf{Y})$. Also, if $\mathbf{X}$ is a random vector, A is a constant matrix and $\mathbf{b}$ is a constant vector, it is easy to show that

$$\mathrm{E}(A\mathbf{X}+\mathbf{b}) = A\ \mathrm{E}(\mathbf{X}) + \mathbf{b}$$

This is how the key linearity properties of expectation carry over to random vectors. The *variance matrix* of a random vector $\mathbf{X}$ is defined by

$$\begin{aligned} \mathrm{Var}\,(\mathbf{X}) &= \mathrm{E}\big([\mathbf{X} - \mathrm{E}(\mathbf{X})][\mathbf{X} - \mathrm{E}(\mathbf{X})]^{\mathrm{T}}\big) \\ &= \mathrm{E}\big(\mathbf{X}\mathbf{X}^{\mathrm{T}}\big) - \mathrm{E}(\mathbf{X})\ \mathrm{E}(\mathbf{X})^{\mathrm{T}} \end{aligned}$$

Note that the (i,j)th element of this matrix is $\mathrm{Cov}\,(X_i, X_j)$, and so *inter alia*, the variance matrix is symmetric. The *covariance matrix* between random vectors $\mathbf{X}$ and $\mathbf{Y}$ is defined by

$$\begin{aligned} \mathrm{Cov}\,(\mathbf{X},\mathbf{Y}) &= \mathrm{E}\big([\mathbf{X} - \mathrm{E}(\mathbf{X})][\mathbf{Y} - \mathrm{E}(\mathbf{Y})]^{\mathrm{T}}\big) \\ &= \mathrm{E}\big(\mathbf{X}\mathbf{Y}^{\mathrm{T}}\big) - \mathrm{E}(\mathbf{X})\ \mathrm{E}(\mathbf{Y})^{\mathrm{T}} \end{aligned}$$

Note that the (i,j)th element of this matrix is $\mathrm{Cov}\,(X_i, Y_j)$ and that $\mathrm{Var}\,(\mathbf{X}) = \mathrm{Cov}\,(\mathbf{X},\mathbf{X})$.

Variance and covariance matrices have many useful properties which derive from the linearity of expectation. The most useful properties are given below:

$$\begin{aligned} \mathrm{Var}\,(A\mathbf{X}+\mathbf{b}) &= A\mathrm{Var}\,(\mathbf{X})\,A^{\mathrm{T}} \\ \mathrm{Cov}\,(\mathbf{X}+\mathbf{Y},\mathbf{Z}) &= \mathrm{Cov}\,(\mathbf{X},\mathbf{Z}) + \mathrm{Cov}\,(\mathbf{Y},\mathbf{Z}) \\ \mathrm{Cov}\,(\mathbf{Y},\mathbf{X}) &= \mathrm{Cov}\,(\mathbf{X},\mathbf{Y})^{\mathrm{T}} \\ \mathrm{Cov}\,(A\mathbf{X}+\mathbf{b},\mathbf{Y}) &= A\mathrm{Cov}\,(\mathbf{X},\mathbf{Y}) \\ \mathrm{Cov}\,(\mathbf{X},A\mathbf{Y}+\mathbf{b}) &= \mathrm{Cov}\,(\mathbf{X},\mathbf{Y})\,A^{\mathrm{T}} \\ \mathrm{Var}\,(\mathbf{X}+\mathbf{Y}) &= \mathrm{Var}\,(\mathbf{X}) + \mathrm{Cov}\,(\mathbf{X},\mathbf{Y}) \\ &\quad + \mathrm{Cov}\,(\mathbf{Y},\mathbf{X}) + \mathrm{Var}\,(\mathbf{Y}) \\ \mathrm{Var}\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} &= \begin{pmatrix} \mathrm{Var}\,(\mathbf{X}) & \mathrm{Cov}\,(\mathbf{X},\mathbf{Y}) \\ \mathrm{Cov}\,(\mathbf{Y},\mathbf{X}) & \mathrm{Var}\,(\mathbf{Y}) \end{pmatrix} \end{aligned}$$

2.3.4 Example

Suppose that we have the following belief structure

B_1	3	4	
D_1	4	1	9

and wish to add to the belief structure the unobservable quantity $B_2 = B_1 + 2D_1$.

A belief structure determines beliefs over all linear combinations of its elements, as well as the elements themselves, so no extra specification is required.

$$\begin{aligned}(B_2) =&(1,2)\begin{pmatrix} B_1 \\ D_1 \end{pmatrix} \\ \mathrm{E}(B_2) =&(1,2)\begin{pmatrix} 3 \\ 4 \end{pmatrix} = (11) \\ \mathrm{Var}\,(B_2) =&(1,2)\begin{pmatrix} 4 & 1 \\ 1 & 9 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \end{pmatrix} = (44) \\ \mathrm{Cov}\left((B_2), \begin{pmatrix} B_1 \\ D_1 \end{pmatrix}\right) =&(1,2)\begin{pmatrix} 4 & 1 \\ 1 & 9 \end{pmatrix} = (6, 19)\end{aligned}$$

Adding these to our tableau gives

B_1	3	4		
D_1	4	1	9	
B_2	11	6	19	44

or

B_1	3	4		
B_2	11	6	44	
D_1	4	1	19	9

2.4 Bayes linear adjusted expectation

2.4.1 Introduction

We motivated subjective expectation as a point estimate which minimises an *expected quadratic loss*. This is quite a key idea in statistics, and we will now use it to derive *adjusted expectation*. You should note however, that there are many other justifications for using adjusted expectation — in particular, it is the *conditional expectation vector* in multivariate normal theory.

Motivation

If we want a point estimate of a random variable, X, we might use the value, α which minimises the expected value of $(X-\alpha)^2$. Clearly the more different α is from X, the bigger the loss will be. Since we don't yet know the value of X, we can't just choose X! So, by minimising the expected loss, we get a "best guess" for X, before X is known. By differentiating and equating to zero, we find that $\alpha = \mathrm{E}(X)$. That is, the point estimate, α, which minimises the expected loss is just $\mathrm{E}(X)$.

Using data

Now suppose that we are interested in B_1 — the first unobservable in our belief structure. If we wanted a best guess for B_1, we would just use $\mathrm{E}(B_1)$. But we are going to observe a collection of data, $D = \{D_1, \ldots, D_n\}$ which is perhaps informative for B_1, so how can we use it? One way is to allow our best guess to depend on the data which we have not yet observed. This will give us a best guess which is random, but it will become known once the data is observed. Allowing the best guess to be any function of the data turns out to be difficult, and so we will just allow our best guess to depend on a linear function of the data. That is, our best guess will be of the form

$$\alpha + \beta_1 D_1 + \beta_2 D_2 + \cdots + \beta_n D_n$$

and we will choose $\alpha, \beta_1, \ldots, \beta_n$ in order to minimise

$$\mathrm{E}\big((B_1 - \alpha - \beta_1 D_1 - \beta_2 D_2 - \ldots - \beta_n D_n)^2\big)$$

We call the observable random quantity which minimises this the *adjusted expectation for B_1 given D*, and denote it $\mathrm{E}_D(B_1)$. So, if we put $C = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}$, we want to find the constant, α and the vector $\boldsymbol{\beta}$ which minimise the loss, L, where

$$L = \mathrm{E}\big([B_1 - C]^2\big)$$

It is more convenient to re-write C in the form

$$\begin{aligned} C &= \alpha + \boldsymbol{\beta}^{\mathrm{T}}[\mathbf{D} - \mathrm{E}(\mathbf{D}) + \mathrm{E}(\mathbf{D})] \\ &= \alpha + \boldsymbol{\beta}^{\mathrm{T}}\,\mathrm{E}(\mathbf{D}) + \boldsymbol{\beta}^{\mathrm{T}}[\mathbf{D} - \mathrm{E}(\mathbf{D})] \\ &= \alpha^\star + \boldsymbol{\beta}^{\mathrm{T}}[\mathbf{D} - \mathrm{E}(\mathbf{D})] \end{aligned}$$

where $\alpha^\star = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\,\mathrm{E}(\mathbf{D})$. As we are about to see, things just work out a bit simpler if we write C as a constant plus a mean zero random term. We can now think about minimising L wrt $\alpha^\star$ and $\boldsymbol{\beta}$.

$$\begin{aligned} L &= \mathrm{E}\big([B_1 - C]^2\big) \\ &= [\,\mathrm{E}(B_1 - C)]^2 + \mathrm{Var}\,(B_1 - C) \\ &= \big[\,\mathrm{E}\big(B_1 - \alpha^\star - \boldsymbol{\beta}^{\mathrm{T}}[\mathbf{D} - \mathrm{E}(\mathbf{D})]\big)\big]^2 + \mathrm{Var}\,\big(B_1 - \alpha^\star - \boldsymbol{\beta}^{\mathrm{T}}[\mathbf{D} - \mathrm{E}(\mathbf{D})]\big) \\ &= [\,\mathrm{E}(B_1) - \alpha^\star]^2 + \mathrm{Var}\,\big(B_1 - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}\big) \end{aligned}$$

This is now the sum of a positive term involving $\alpha^\star$ only and a positive term involving $\boldsymbol{\beta}$ only, and so we can do the minimisations separately.

The first term is clearly minimised by choosing $\alpha^\star = \mathrm{E}(B_1)$, as then this term is zero. To find the minimum of the second term, we need to differentiate it wrt $\boldsymbol{\beta}$, and equate to zero. Now

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}} \mathrm{Var}\,\big(B_1 - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}\big) &= \frac{\partial}{\partial \boldsymbol{\beta}}\big[\mathrm{Var}\,(B_1) + \mathrm{Var}\,\big(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}\big) - 2\mathrm{Cov}\,\big(B_1, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}\big)\big] \\ &= \frac{\partial}{\partial \boldsymbol{\beta}}\big[\mathrm{Var}\,(B_1) + \boldsymbol{\beta}^{\mathrm{T}}\mathrm{Var}\,(\mathbf{D})\,\boldsymbol{\beta} - 2\mathrm{Cov}\,(B_1, \mathbf{D})\,\boldsymbol{\beta}\big] \\ &= 2\mathrm{Var}\,(\mathbf{D})\,\boldsymbol{\beta} - 2\mathrm{Cov}\,(\mathbf{D}, B_1) \end{aligned}$$

so

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}} \mathrm{Var}\,\big(B_1 - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}\big) &= 0 \\ \Rightarrow \mathrm{Var}\,(\mathbf{D})\,\boldsymbol{\beta} &= \mathrm{Cov}\,(\mathbf{D}, B_1) \\ \Rightarrow \boldsymbol{\beta} &= \mathrm{Var}\,(\mathbf{D})^{-1}\,\mathrm{Cov}\,(\mathbf{D}, B_1) \\ \Rightarrow \boldsymbol{\beta}^{\mathrm{T}} &= \mathrm{Cov}\,(B_1, \mathbf{D})\,\mathrm{Var}\,(\mathbf{D})^{-1} \end{aligned}$$

Exercise — check that this is a minimum. So, our adjusted expectation is given by

$$\mathrm{E}_D(B_1) = \mathrm{E}(B_1) + \mathrm{Cov}\,(B_1, \mathbf{D})\,\mathrm{Var}\,(\mathbf{D})^{-1}\,[\mathbf{D} - \mathrm{E}(\mathbf{D})]$$

For a vector, $\mathbf{B}$, the *adjusted expectation vector*, $\mathrm{E}_D(\mathbf{B})$ is just the vector of adjusted expectations of the components, and hence

$$\boxed{\mathrm{E}_D(\mathbf{B}) = \mathrm{E}(\mathbf{B}) + \mathrm{Cov}\,(\mathbf{B},\mathbf{D})\,\mathrm{Var}\,(\mathbf{D})^{-1}\,[\mathbf{D} - \mathrm{E}(\mathbf{D})]}$$

Bayes linear adjusted expectation

2.4.2 Observed adjusted expectation

We now know how to work out the form of adjusted expectation. This is an observable random quantity which will become known when $\mathbf{D}$ is observed. Suppose that $\mathbf{D}$ is observed to be $\mathbf{d}$. Then the *observed adjusted expectation,* $\mathrm{E}_d(\mathbf{B})$, is just $\mathrm{E}_D(\mathbf{B})$ evaluated at $\mathbf{d}$.

$$\mathrm{E}_d(\mathbf{B}) = \mathrm{E}(\mathbf{B}) + \mathrm{Cov}\,(\mathbf{B},\mathbf{D})\,\mathrm{Var}\,(\mathbf{D})^{-1}\,[\mathbf{d} - \mathrm{E}(\mathbf{D})]$$

This is a constant (vector), and represents our expectation for $\mathbf{B}$ after seeing the data. Thus, our prior expectation, $\mathrm{E}(\mathbf{B})$ is altered, depending on the outcome of the data, in order to get a "new" expectation, $\mathrm{E}_d(\mathbf{B})$.

2.4.3 Example (A)

Consider the following belief structure for $\mathbf{B} = (B_1, B_2)^{\mathrm{T}}$ and $\mathbf{D} = (D_1, D_2)^{\mathrm{T}}$.

B_1	1	4			
B_2	2	2	9		
D_1	1	2	1	4	
D_2	2	1	2	2	9

What is $\mathrm{E}_D(\mathbf{B})$?

$$\begin{aligned}
\mathrm{E}_D(\mathbf{B}) &= \mathrm{E}(\mathbf{B}) + \mathrm{Cov}\,(\mathbf{B},\mathbf{D})\,\mathrm{Var}\,(\mathbf{D})^{-1}\,[\mathbf{D} - \mathrm{E}(\mathbf{D})] \\
&= \begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 4 & 2 \\ 2 & 9 \end{pmatrix}^{-1} \begin{pmatrix} D_1 - 1 \\ D_2 - 2 \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{2}D_1 + \frac{1}{2} \\ \frac{5}{32}D_1 + \frac{3}{16}D_2 + \frac{47}{32} \end{pmatrix}
\end{aligned}$$

Now suppose that the data is observed to be $d_1 = 2$, $d_2 = 6$. To get the observed adjusted expectation, just put $D_1 = 2$ and $D_2 = 6$ into the above expression for adjusted expectation in order to get

$$\mathrm{E}_d(\mathbf{B}) = \begin{pmatrix} 1\frac{1}{2} \\ 2\frac{29}{32} \end{pmatrix}$$

2.4.4 Adjusted variance

Before observing the data, $\mathrm{E}_D(\mathbf{B})$ is a random quantity, and hence has a variance, known as the *resolved variance*. We can compute it as follows

$$\begin{aligned}
\mathrm{Var}\,(\mathrm{E}_D(\mathbf{B})) &= \mathrm{Var}\left(\mathrm{Cov}\,(\mathbf{B},\mathbf{D})\,\mathrm{Var}\,(\mathbf{D})^{-1}\,\mathbf{D}\right) \\
&= \mathrm{Cov}\,(\mathbf{B},\mathbf{D})\,\mathrm{Var}\,(\mathbf{D})^{-1}\,\mathrm{Cov}\,(\mathbf{D},\mathbf{B})
\end{aligned}$$

It also has a covariance with the quantities of interest, which we compute as

$$\begin{aligned}\mathrm{Cov}\left(\mathbf{B},\ \mathrm{E}_D(\mathbf{B})\right) &= \mathrm{Cov}\left(\mathbf{B}, \mathrm{Cov}\left(\mathbf{B},\mathbf{D}\right)\mathrm{Var}\left(\mathbf{D}\right)^{-1}\mathbf{D}\right)\\ &= \mathrm{Cov}\left(\mathbf{B},\mathbf{D}\right)\mathrm{Var}\left(\mathbf{D}\right)^{-1}\mathrm{Cov}\left(\mathbf{D},\mathbf{B}\right)\end{aligned}$$

This tells us that

$$\mathrm{Var}\left(\ \mathrm{E}_D(\mathbf{B})\right) = \mathrm{Cov}\left(\mathbf{B},\ \mathrm{E}_D(\mathbf{B})\right)$$

which means that

$$\mathrm{Cov}\left(\ \mathrm{E}_D(\mathbf{B}), \mathbf{B} -\ \mathrm{E}_D(\mathbf{B})\right) = 0$$

so if we write $\mathbf{B}$ as

$$\mathbf{B} =\ \mathrm{E}_D(\mathbf{B}) + (\mathbf{B} -\ \mathrm{E}_D(\mathbf{B}))$$

we get

$$\mathrm{Var}\left(\mathbf{B}\right) = \mathrm{Var}\left(\ \mathrm{E}_D(\mathbf{B})\right) + \mathrm{Var}\left(\mathbf{B} -\ \mathrm{E}_D(\mathbf{B})\right)$$

This last equations shows how our uncertainty for $\mathbf{B}$ is decomposed into two parts — a part which is explainable by the data, and a part which is "left over" after observing the data. As previously mentioned, the part explained by the data is known as the resolved variance, and the part left over is the *adjusted variance*, written $\mathrm{Var}_D\left(\mathbf{B}\right)$, where

$$\begin{aligned}\mathrm{Var}_D\left(\mathbf{B}\right) =& \mathrm{Var}\left(\mathbf{B} -\ \mathrm{E}_D(\mathbf{B})\right)\\ =& \mathrm{Var}\left(\mathbf{B}\right) - \mathrm{Cov}\left(\mathbf{B},\mathbf{D}\right)\mathrm{Var}\left(\mathbf{D}\right)^{-1}\mathrm{Cov}\left(\mathbf{D},\mathbf{B}\right)\end{aligned}$$

Thus, our prior variance matrix, $\mathrm{Var}\left(\mathbf{B}\right)$ is updated after observing the data to give a "new" variance matrix, $\mathrm{Var}_D\left(\mathbf{B}\right)$, the old variance matrix adjusted in the light of the data.

A crucial thing to note is that the adjusted variance matrix does not depend explicitly on the observed data, $\mathbf{D}$, and so we can evaluate our adjusted variance matrix *before* observing the data. This has important consequences for Bayes linear design problems.

2.4.5 Example (A)

Returning to our example, we can evaluate $\mathrm{Var}_D\left(\mathbf{B}\right)$ as follows

$$\begin{aligned}\mathrm{Var}_D\left(\mathbf{B}\right) &= \begin{pmatrix} 4 & 2 \\ 2 & 9 \end{pmatrix} - \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 4 & 2 \\ 2 & 9 \end{pmatrix}^{-1}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\\ &= \begin{pmatrix} 3 & 1\frac{1}{2} \\ 1\frac{1}{2} & 8\frac{15}{32} \end{pmatrix}\end{aligned}$$

So, observing the data changes the expectation vector for $\mathbf{B}$ from

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix} \text{ to } \begin{pmatrix} 1\frac{1}{2} \\ 2\frac{29}{32} \end{pmatrix}$$

and the variance matrix from

$$\begin{pmatrix} 4 & 2 \\ 2 & 9 \end{pmatrix} \text{ to } \begin{pmatrix} 3 & 1\frac{1}{2} \\ 1\frac{1}{2} & 8\frac{15}{32} \end{pmatrix}.$$

Note that the adjusted expectations of the quantities of interest are just the elements of the adjusted expectation vector, and that the adjusted variances of the quantities of interest are just the diagonal elements of the adjusted variance matrix.

Once we know the adjusted expectation vector and variance matrix, we can work out the adjusted mean and variance of any linear combination of the quantities of interest.

2.4.6 Resolution

Consider a quantity of interest, $B^\star$, which is a linear combination of the quantities in B. Once we have specified the belief structure (and before we observe the data), we can work out the prior and adjusted variance of $B^\star$. One way of summarising the effect of the adjustment process on $B^\star$ is to look at the proportion of prior variance explained by the data. Consequently, we define the *resolution* for the adjustment of $B^\star$ by D, written $\mathrm{r}_D\left(B^\star\right)$, in the following way.

$$\begin{aligned} \mathrm{r}_D\left(B^\star\right) =& \frac{\mathrm{Var}\left(\ \mathrm{E}_D(B^\star)\right)}{\mathrm{Var}\left(B^\star\right)} \\ =& \frac{\mathrm{Var}\left(B^\star\right) - \mathrm{Var}_D\left(B^\star\right)}{\mathrm{Var}\left(B^\star\right)} \\ =& 1 - \frac{\mathrm{Var}_D\left(B^\star\right)}{\mathrm{Var}\left(B^\star\right)} \end{aligned}$$

2.4.7 Example (A)

Consider $B^\star = B_1 + B_2$ *ie.* $B^\star = (1,1)\mathbf{B}$. So

$$\begin{aligned} \mathrm{Var}\left(B^\star\right) =& (1,1)\begin{pmatrix} 4 & 2 \\ 2 & 9 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ =& 17 \end{aligned}$$

Similarly

$$\begin{aligned} \mathrm{Var}_D\left(B^\star\right) &= (1,1)\begin{pmatrix} 3 & 1\frac{1}{2} \\ 1\frac{1}{2} & 8\frac{15}{32} \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= 14\frac{15}{32} \end{aligned}$$

So,

$$\begin{aligned} \mathrm{r}_D\left(B^\star\right) &= 1 - \frac{14\frac{15}{32}}{17} \\ &= \frac{81}{544} \approx 0.15 \end{aligned}$$

That is, approximately 15% of the "uncertainty" for $B^\star$ has been/will be "resolved" by observing the data. Note that we can write the relationship between prior and adjusted variances in several different ways. For example

$$\text{Var}\,(\,\text{E}_D(B^\star)) = \text{r}_D\,(B^\star)\,\text{Var}\,(B^\star)$$

and

$$\text{Var}_D\,(B^\star) = [1 - \text{r}_D\,(B^\star)]\text{Var}\,(B^\star)$$

So $1 - \text{r}_D\,(B^\star)$ is the scale factor, mapping prior variance to adjusted variance. If we are adjusting a collection of quantities, B, by a collection D, a natural question to ask is:

Which linear combinations of the quantities of interest have we learned most and least about?

That is, if we have $B^\star = \boldsymbol{\alpha}^{\text{T}}\mathbf{B}$ (for some vector $\boldsymbol{\alpha}$), what choice of $\boldsymbol{\alpha}$ will lead to the smallest and largest values of $\text{r}_D\,(B^\star)$? In order to answer this, we need to think a little more generally about the transformation from prior to adjusted variance matrices.

2.4.8 Introduction to $[B/D]$

Before looking in more detail at some of the underlying Bayes linear theory, we will examine a computer software package designed especially for carrying out Bayes linear computations. The package is known as $[B/D]$ (the name represents the adjustment of B by D). The package allows you to carry out a Bayes linear analysis in a natural way, without having to worry about the underlying matrix theory. Reconsider Example (A). We can declare the prior belief specifications and data as follows.

```
BD>element b.1=1,b.2=2
BD>base b=b.$
BD>element d.1=1,d.2=2
BD>base d=d.$
BD>base a=b,d
BD>var v(1,a)
BD*4
BD*2 9
BD*2 1 4
BD*1 2 2 9
BD>data d.1(1)=2
BD>data d.2(1)=6
```

$[B/D]$ allows many different specifications to be stored. By default, initial expectation declarations go into store 1, but the expression `v(1,a)` is needed to refer to variance store 1 for the whole of base `a`. We can check we have made our specifications correctly using the `look` command.

```
BD>look (b)
Base
A  :  B, D
B  :  B.1, B.2
```

```
D  :  D.1, D.2

BD>look (e1)
Store number 1
Element      Expectation      SD
B.1            1.00000       2.0000
B.2            2.00000       3.0000
D.1            1.00000       2.0000
D.2            2.00000       3.0000

BD>look (v1)
Covariances in store (1) :
B.1 :    4.0000   2.0000   2.0000   1.0000
B.2 :    2.0000   9.0000   1.0000   2.0000
D.1 :    2.0000   1.0000   4.0000   2.0000
D.2 :    1.0000   2.0000   2.0000   9.0000

BD>look (d)
Data storage   :
          d.1         d.2
  1      2.000       6.000
```

We can now get $[B/D]$ to do Bayes linear adjustments for us. For example, if we wish to keep adjusted expectations and variances in store 2, and adjust B by D we do

```
BD>control ae=2
BD>control ac=2
BD>adjust [b/d]
```

We can then examine the results.

```
BD>look (e2)b
Store number 2
Element      Expectation      SD
B.1            1.50000       1.7321
B.2            2.90625       2.9101

BD>look (v2)b
Covariances in store (2) :
B.1 :    3.0000   1.5000
B.2 :    1.5000   8.4688

BD>show e
Element         Adjusted Expectation
B.1 :         0.5000 D.1  + 0.0000 D.2  + 0.5000
B.2 :         0.1562 D.1  + 0.1875 D.2  + 1.4688

BD>show v+
              ------------ Variation -------------
Element       Initial      Adjustment     Resolved    Resolution
B.1              4.0000        3.0000        1.0000     0.2500
```

```
B.2                 9.0000         8.4688         0.5313     0.0590
[B]                 2              1.7148         0.2852     0.1426
```

Next we will add a linear combination, $B^{\star}$, to the belief structure, and examine the effect of the adjustment on that quantity.

```
BD>build bstar = (1)b.1 + (1)b.2
BD>look (e1)
Store number 1
Element      Expectation      SD
B.1            1.00000       2.0000
B.2            2.00000       3.0000
Bstar          3.00000       4.1231
D.1            1.00000       2.0000
D.2            2.00000       3.0000

BD>look (v1)
Covariances in store (1) :
B.1   :    4.0000   2.0000   6.0000   2.0000   1.0000
B.2   :    2.0000   9.0000  11.0000   1.0000   2.0000
Bstar :    6.0000  11.0000  17.0000   3.0000   3.0000
D.1   :    2.0000   1.0000   3.0000   4.0000   2.0000
D.2   :    1.0000   2.0000   3.0000   2.0000   9.0000

BD>adjust [bstar/d]
BD>show v+
              ------------ Variation -------------
Element       Initial      Adjustment      Resolved     Resolution
Bstar           17.0000        14.4688         2.5313     0.1489
[Bstar]         1               0.8511         0.1489     0.1489
```

For more details, see the $[B/D]$ Manual, available locally on the WWW via the page for this course, or using the address `http://www.mas.ncl.ac.uk/~ndjw1/bdman/man.html`

2.5 Variance and resolution transforms

2.5.1 Definitions

Let's look again at the adjusted variance matrix:

$$\begin{aligned}
\text{Var}_D(\mathbf{B}) &= \text{Var}(\mathbf{B}) - \text{Cov}(\mathbf{B},\mathbf{D})\,\text{Var}(\mathbf{D})^{-1}\,\text{Cov}(\mathbf{D},\mathbf{B}) \\
&= \text{Var}(\mathbf{B}) - \text{Cov}(\mathbf{B},\mathbf{D})\,\text{Var}(\mathbf{D})^{-1}\,\text{Cov}(\mathbf{D},\mathbf{B})\,\text{Var}(\mathbf{B})^{-1}\,\text{Var}(\mathbf{B}) \\
&= \left(\text{I} - \text{Cov}(\mathbf{B},\mathbf{D})\,\text{Var}(\mathbf{D})^{-1}\,\text{Cov}(\mathbf{D},\mathbf{B})\,\text{Var}(\mathbf{B})^{-1}\right)\text{Var}(\mathbf{B}) \\
&= \left(\text{I} - \text{T}_{[D]}^{\{B\}}\right)\text{Var}(\mathbf{B})
\end{aligned}$$

where

$$\boxed{\text{T}_{[D]}^{\{B\}} = \text{Cov}(\mathbf{B},\mathbf{D})\,\text{Var}(\mathbf{D})^{-1}\,\text{Cov}(\mathbf{D},\mathbf{B})\,\text{Var}(\mathbf{B})^{-1}}$$

$\mathrm{T}_{[D]}^{\{B\}}$ is known as the *resolution transform for the adjustment of* $\mathbf{B}$ *by* $\mathbf{D}$. Also, if we put $\mathrm{S}_{[D]}^{\{B\}} = \mathrm{I} - \mathrm{T}_{[D]}^{\{B\}}$, then we have

$$\boxed{\mathrm{Var}_D\left(\mathbf{B}\right) = \mathrm{S}_{[D]}^{\{B\}}\mathrm{Var}\left(\mathbf{B}\right)}$$

and $\mathrm{S}_{[D]}^{\{B\}}$ is known as the *variance transform for the adjustment of* $\mathbf{B}$ *by* $\mathbf{D}$. The variance transform is the matrix (or linear operator) which transforms the prior variance matrix into the adjusted variance matrix. It therefore contains a great deal of information about the adjustment process.

Clearly an alternative expression for the variance transform is

$$\mathrm{S}_{[D]}^{\{B\}} = \mathrm{Var}_D\left(\mathbf{B}\right)\mathrm{Var}\left(\mathbf{B}\right)^{-1}$$

This is useful if you've already calculated $\mathrm{Var}_D\left(\mathbf{B}\right)$, or if you wish to check your matrix calculations.

2.5.2 Example (A)

$$\begin{aligned}\mathrm{T}_{[D]}^{\{B\}} &= \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 4 & 2 \\ 2 & 9 \end{pmatrix}^{-1}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 4 & 2 \\ 2 & 9 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \frac{1}{4} & 0 \\ \frac{55}{512} & \frac{9}{256} \end{pmatrix}\end{aligned}$$

So, we can use this to work out $\mathrm{S}_{[D]}^{\{B\}}$ as

$$\begin{aligned}\mathrm{S}_{[D]}^{\{B\}} &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} \frac{1}{4} & 0 \\ \frac{55}{512} & \frac{9}{256} \end{pmatrix} \\ &= \begin{pmatrix} \frac{3}{4} & 0 \\ -\frac{55}{512} & \frac{247}{256} \end{pmatrix}\end{aligned}$$

Alternatively, we can calculate $\mathrm{S}_{[D]}^{\{B\}}$ directly using $\mathrm{Var}_D\left(\mathbf{B}\right)$, which we calculated earlier.

$$\begin{aligned}\mathrm{S}_{[D]}^{\{B\}} &= \begin{pmatrix} 3 & \frac{3}{2} \\ \frac{3}{2} & \frac{271}{32} \end{pmatrix}\begin{pmatrix} 4 & 2 \\ 2 & 9 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \frac{3}{4} & 0 \\ -\frac{55}{512} & \frac{247}{256} \end{pmatrix}\end{aligned}$$

2.5.3 Interpreting the transforms

First let us look at the case of updating a single random quantity, $B^\star$. We know that

$$\mathrm{Var}_D\left(B^\star\right) = \left(1 - \mathrm{T}_{[D]}^{\{B^\star\}}\right)\mathrm{Var}\left(B^\star\right)$$

and so

$$\mathrm{T}_{[D]}^{\{B^\star\}} = 1 - \frac{\mathrm{Var}_D\left(B^\star\right)}{\mathrm{Var}\left(B^\star\right)}$$

So

$$\mathrm{T}_{[D]}^{\{B^\star\}} = \mathrm{r}_D\left(B^\star\right)$$

A one-dimensional resolution transform is therefore easy to interpret. In general, the resolution transform is so named because it contains information about the resolutions for all linear combinations of quantities of interest. In fact, the eigenvalues of the (transposed) resolution transform matrix correspond to the resolutions of the linear combinations represented by the corresponding eigenvector.

2.5.4 Eigenstructure of the resolution transform

Suppose that $\boldsymbol{\alpha}$ is an eigenvector of $\mathrm{T}_{[D]}^{\{B\}\mathrm{T}}$, and that λ is the corresponding eigenvalue, so that

$$\mathrm{T}_{[D]}^{\{B\}\mathrm{T}}\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha}$$

Let's look at $\mathrm{r}_D\left(B^\star\right)$, the resolution of $B^\star = \boldsymbol{\alpha}^\mathrm{T}\mathbf{B}$.

$$\begin{aligned}
\mathrm{r}_D\left(B^\star\right) =& 1 - \frac{\mathrm{Var}_D\left(B^\star\right)}{\mathrm{Var}\left(B^\star\right)} \\
=& 1 - \frac{\boldsymbol{\alpha}^\mathrm{T}\mathrm{Var}_D\left(\mathbf{B}\right)\boldsymbol{\alpha}}{\boldsymbol{\alpha}^\mathrm{T}\mathrm{Var}\left(\mathbf{B}\right)\boldsymbol{\alpha}} \\
=& 1 - \frac{\boldsymbol{\alpha}^\mathrm{T}\left(\mathrm{I} - \mathrm{T}_{[D]}^{\{B\}}\right)\mathrm{Var}\left(\mathbf{B}\right)\boldsymbol{\alpha}}{\boldsymbol{\alpha}^\mathrm{T}\mathrm{Var}\left(\mathbf{B}\right)\boldsymbol{\alpha}} \\
=& \frac{\boldsymbol{\alpha}^\mathrm{T}\mathrm{T}_{[D]}^{\{B\}}\mathrm{Var}\left(\mathbf{B}\right)\boldsymbol{\alpha}}{\boldsymbol{\alpha}^\mathrm{T}\mathrm{Var}\left(\mathbf{B}\right)\boldsymbol{\alpha}} \\
=& \frac{\left(\mathrm{T}_{[D]}^{\{B\}\mathrm{T}}\boldsymbol{\alpha}\right)^\mathrm{T}\mathrm{Var}\left(\mathbf{B}\right)\boldsymbol{\alpha}}{\boldsymbol{\alpha}^\mathrm{T}\mathrm{Var}\left(\mathbf{B}\right)\boldsymbol{\alpha}} \\
=& \frac{\left(\lambda\boldsymbol{\alpha}\right)^\mathrm{T}\mathrm{Var}\left(\mathbf{B}\right)\boldsymbol{\alpha}}{\boldsymbol{\alpha}^\mathrm{T}\mathrm{Var}\left(\mathbf{B}\right)\boldsymbol{\alpha}} \\
=& \lambda
\end{aligned}$$

That is, the random quantity, $B^\star$, represented by the eigenvector, $\boldsymbol{\alpha}$, has resolution equal to the corresponding eigenvalue, λ.

It turns out that belief transforms always have a full set of real eigenvalues and eigenvectors. Since the eigenvalues all correspond to resolutions, we also know that they must all be between zero and one inclusive. Eigenvectors with zero eigenvalues correspond to quantities we have learned nothing about. Eigenvectors with unit eigenvalues correspond to quantities we have actually observed.

The random quantities corresponding to the eigenvectors of the resolution transform are known as the *canonical quantities*. The corresponding eigenvalues are known as the *canonical resolutions*.

2.5.5 Example (A)

The transpose of the resolution transform for the problem is

$$\mathrm{T}_{[D]}^{\{B\}\mathrm{T}} = \begin{pmatrix} \frac{1}{4} & 0 \\ \frac{55}{512} & \frac{9}{256} \end{pmatrix}^{\mathrm{T}} = \begin{pmatrix} \frac{1}{4} & \frac{55}{512} \\ 0 & \frac{9}{256} \end{pmatrix}$$

Since this matrix is upper triangular, we can read the eigenvalues off the diagonal: the canonical resolutions are 1/4 and 9/256. N.B. *This only works for upper or lower triangular matrices!*

The eigenvector corresponding to 1/4 is $(1, 0)^{\mathrm{T}}$. So B_1 is the first canonical quantity, and has resolution 1/4. It represents the linear combination of the quantities of interest for which the data has been most informative.

The eigenvector corresponding to 9/256 is $(1, -2)^{\mathrm{T}}$. So $B_1 - 2B_2$ is the second canonical quantity, and has resolution 9/256. As it is the last canonical quantity, it represents the linear combination of the quantities of interest for which the data have been least informative.

Note that $\mathrm{Cov}\,(B_1, B_1 - 2B_2) = 0$.

A useful summary of the overall effect of the adjustment is the sum of the canonical resolutions divided by its maximum possible value (the dimension of $\mathbf{B}$). Since the sum of the eigenvalues of a matrix is just the trace of the matrix, we define

$$\mathrm{r}_D\,(\mathbf{B}) = \frac{\mathrm{Tr}\left(\mathrm{T}_{[D]}^{\{B\}}\right)}{\mathrm{rank}\,(\mathrm{Var}\,(\mathbf{B}))}$$

2.5.6 Example (A)

For our example, $\mathrm{Var}\,(\mathbf{B})$ has full rank (*ie.* 2), and so, since

$$\mathrm{T}_{[D]}^{\{B\}} = \begin{pmatrix} \frac{1}{4} & 0 \\ \frac{55}{512} & \frac{9}{256} \end{pmatrix}$$

we have

$$\mathrm{r}_D\,(\mathbf{B}) = \frac{\frac{1}{4} + \frac{9}{256}}{2} = \frac{73}{512} = 0.1426$$

2.5.7 Resolutions in $[B/D]$

Returning to the $[B/D]$ session from before, we can access information relating the the resolution transform for the adjustment as follows.

```
BD>adjust [b/d]
BD>show v+
               ------------ Variation -------------
Element        Initial      Adjustment      Resolved    Resolution
B.1               4.0000        3.0000        1.0000      0.2500
B.2               9.0000        8.4688        0.5313      0.0590
[B]               2             1.7148        0.2852      0.1426
```

We can now understand the bottom line of this output. The first number is the rank of $\mathrm{Var}\,(\mathbf{B})$, the third is the trace of the resolution transform, and the last is the resolution of the adjustment.

We can also obtain the canonical directions and their resolutions, as follows.

```
BD>show cd+

 Canonical Directions  :
                 1           2
Resolution     0.2500      0.0352

B.1            0.5000      0.1768
B.2            0.0000     -0.3536
Constant      -0.5000      0.5303
```

The constant is just minus the expected value. Note that it is the direction of the eigenvector which matters — the magnitude is arbitrary.

Finally, if we want to see the raw resolution transform in full, we can do this as follows.

```
BD>show rm

Vectors of the resolution matrix :
  1 :   0.2500 B.1  + 0.0000 B.2
  2 :   0.1074 B.1  + 0.0352 B.2
```

The numeric terms give the elements of the resolution transform matrix.

2.5.8 Eigenstructure of the variance transform

Recall that the variance transform and the resolution transform are related via

$$\mathrm{S}_{[D]}^{\{B\}} = \mathrm{I} - \mathrm{T}_{[D]}^{\{B\}}$$

The eigenstructure of the two matrices is also closely related. For suppose that $\boldsymbol{\alpha}$ is an eigenvector of $\mathrm{T}_{[D]}^{\{B\}\mathrm{T}}$ with corresponding eigenvalue λ. Then

$$\begin{aligned}\mathrm{S}_{[D]}^{\{B\}\mathrm{T}}\boldsymbol{\alpha} =& \left(\mathrm{I} - \mathrm{T}_{[D]}^{\{B\}\mathrm{T}}\right)\boldsymbol{\alpha} \\ =& \boldsymbol{\alpha} - \mathrm{T}_{[D]}^{\{B\}\mathrm{T}}\boldsymbol{\alpha} \\ =& \boldsymbol{\alpha} - \lambda\boldsymbol{\alpha} \\ =& (1-\lambda)\boldsymbol{\alpha}\end{aligned}$$

So $\boldsymbol{\alpha}$ is an eigenvector of $\mathrm{S}_{[D]}^{\{B\}\mathrm{T}}$ with eigenvalue $1-\lambda$. Consequently, all of the eigenvectors of $\mathrm{S}_{[D]}^{\{B\}\mathrm{T}}$ and $\mathrm{T}_{[D]}^{\{B\}\mathrm{T}}$ are the same, and the corresponding eigenvalues are one minus each other. So *inter alia*, all eigenvalues of $\mathrm{S}_{[D]}^{\{B\}\mathrm{T}}$ are between zero and one, and small eigenvalues correspond to random quantities for which the data has been highly informative.

2.5.9 Example (A)

Without calculation, we can write down the eigenstructure of

$$\mathrm{S}_{[D]}^{\{B\}\mathrm{T}} = \begin{pmatrix} \frac{3}{4} & -\frac{55}{512} \\ 0 & \frac{247}{256} \end{pmatrix}$$

It has eigenvector $(1,0)^{\mathrm{T}}$ with eigenvalue $3/4$ and eigenvector $(1,-2)^{\mathrm{T}}$ with eigenvalue $247/256$. The smaller eigenvalue ($3/4$), corresponds to the random quantity for which the data were most informative.

2.6 Exchangeability

2.6.1 Introduction

Exchangeability is the subjective statistical equivalent of the concept of "simple random sampling". Strict exchangeability is phrased in terms of probability distributions. A collection of random quantities are *exchangeable* if the joint probability distribution remains invariant under an arbitrary permutation of the random quantities. However, this property is rather stronger than we need, since we are only interested in expectations, variances and covariances. Weak (second-order) exchangeability is therefore defined as follows.

Definition

A collection of random quantities $X_1, \dots, X_n$ is ***second-order exchangeable*** *if all expectation, variance and covariance specifications remain invariant under an arbitrary permutation of the random quantities, $X_{(1)}, \dots, X_{(n)}$.*

If the expectation specifications have to remain the same when random quantities are permuted, then the expectation specifications must all be the same. Similarly, all variance specifications must be the same. Also, all covariances between different random quantities must be the same. This leads to the following

Lemma

If the quantities $X_1, \dots, X_n$ are second-order exchangeable, then the belief structure for $\mathbf{X} = (X_1, \dots, X_n)^{\mathrm{T}}$ has the following form

X_1	e	v			
X_2	e	c	v		
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\ddots$	
X_n	e	c	$\cdots$	c	v

So, the full belief structure for an arbitrarily large collection of second-order exchangeable quantities is determined by the specification of just 3 numbers: e, v and c.

The assumption of exchangeability is qualitatively different to that usually used in modelling data (independence, identically distributed, *etc.*). However, the exchangeability assumption does imply just such a statistical model.

Consider the following model

$$X_k = \Theta + \epsilon_k, \quad k = 1, 2, \dots, n$$

where $\mathrm{E}(\Theta) = \mu_0$, $\mathrm{Var}\,(\Theta) = \sigma_0^2$, $\mathrm{E}(\epsilon_k) = 0$, $\mathrm{Var}\,(\epsilon_k) = \sigma^2$, $\mathrm{Cov}\,(\Theta, \epsilon_k) = 0$, $\mathrm{Cov}\,(\epsilon_j, \epsilon_k) = 0,\ \forall j \neq k$.

If Θ was "fixed but unknown", rather than "random", this would be the standard statistical model for simple random variation. What is the belief structure over $\mathbf{X} = (X_1, \dots, X_n)^{\mathrm{T}}$ for this model? Well,

$$\mathrm{E}(X_k) = \mathrm{E}(\Theta) + \mathrm{E}(\epsilon_k) = \mu_0,$$

$$\mathrm{Var}\,(X_k) = \mathrm{Var}\,(\Theta) + \mathrm{Var}\,(\epsilon_k) = \sigma_0^2 + \sigma^2$$

and

$$\begin{aligned}\text{Cov}\,(X_j, X_k) =&\text{Cov}\,(\Theta + \epsilon_j, \Theta + \epsilon_k)\\ =&\text{Cov}\,(\Theta, \Theta)\\ =&\text{Var}\,(\Theta) = \sigma_0^2\end{aligned}$$

So, the belief structure is

$$\begin{array}{c|c|ccccc} X_1 & \mu_0 & \sigma_0^2+\sigma^2 & & & \\ X_2 & \mu_0 & \sigma_0^2 & \sigma_0^2+\sigma^2 & & \\ \vdots & \vdots & \vdots & \ddots & \ddots & \\ X_n & \mu_0 & \sigma_0^2 & \cdots & \sigma_0^2 & \sigma_0^2+\sigma^2 \end{array}$$

This is exactly the form of the belief structure for exchangeable random quantities, where $e = \mu_0$, $c = \sigma_0^2$ and $v = \sigma_0^2 + \sigma^2$. So, random quantities modelled in this way are second-order exchangeable. Remarkably, however, the implication also goes the other way.

Theorem (second-order exchangeability representation)

If random quantities, $X_1, \ldots, X_n$ may be regarded as a finite sample from a potentially infinite collection of random quantities, $X_1, X_2, \ldots$, which are second-order exchangeable, then the $\{X_k\}$ may be represented in the form

$$X_k = \Theta + \epsilon_k, \quad k = 1, 2, \ldots$$

where $\Theta, \epsilon_1, \epsilon_2, \ldots$ are random quantities with the properties

$$\text{E}(\Theta) = \text{E}(X_k) = \mu_0,$$

$$\text{Var}\,(\Theta) = \text{Cov}\,(X_j, X_k) = \sigma_0^2,$$

$$\text{E}(\epsilon_k) = 0,$$

$$\text{Var}\,(\epsilon_k) = \text{Var}\,(X_k) - \sigma_0^2 = \sigma^2,$$

$$\text{Cov}\,(\Theta, \epsilon_k) = 0,$$

$$\text{Cov}\,(\epsilon_k, \epsilon_k) = 0, \forall j \neq k.$$

Sketch proof:

Assuming that all appropriate limits exist, it can be shown that $X_k = \Theta + \epsilon_k$ by taking

$$\Theta = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} X_k, \quad \epsilon_k = X_k - \Theta, \quad \forall k.$$

The properties of Θ and ϵ_k are then easily deduced. For example

$$\begin{aligned}\text{Var}\,(\Theta) =& \text{Var}\left(\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}X_k\right)\\ =& \lim_{N\to\infty}\text{Var}\left(\frac{1}{N}\sum_{k=1}^{N}X_k\right)\\ =& \lim_{N\to\infty}\frac{1}{N^2}\text{Var}\left(\sum_{k=1}^{N}X_k\right)\\ =& \lim_{N\to\infty}\frac{1}{N^2}\big[N\text{Var}\,(X_k)\\ &\quad + N(N-1)\text{Cov}\,(X_j, X_k)\,\big]\\ =& \text{Cov}\,(X_j, X_k)\end{aligned}$$

Other properties may be derived similarly.

So, if a collection of random quantities are judged to be second-order exchangeable, then they should be "modelled" as $X_k = \Theta + \epsilon_k$, where Θ denotes their "underlying mean", and ϵ_k denotes "individual variation". In particular, $\text{Var}\,(\Theta) = \sigma_0^2$ represents uncertainty about the underlying mean, and $\text{Var}\,(\epsilon_k) = \sigma^2$ represents a specification for the variability of the X_k. So, for $\mathbf{X} = (X_1, \ldots, X_n)^{\text{T}}$ we have

$$\begin{aligned}\text{Var}\,(\mathbf{X}) =& \begin{pmatrix}\sigma_0^2+\sigma^2 & \sigma_0^2 & \cdots & \sigma_0^2\\ \sigma_0^2 & \sigma_0^2+\sigma^2 & \ddots & \vdots\\ \vdots & \ddots & \ddots & \sigma_0^2\\ \sigma_0^2 & \cdots & \sigma_0^2 & \sigma_0^2+\sigma^2\end{pmatrix}\\ =& \sigma^2\text{I} + \sigma_0^2\text{J},\end{aligned}$$

where J is an $n \times n$ matrix of 1's. Similarly,

$$\begin{aligned}\text{Cov}\,(\Theta, \mathbf{X}) =& (\text{Cov}\,(\Theta, X_1), \ldots, \text{Cov}\,(\Theta, X_n))\\ =& (\text{Var}\,(\Theta), \ldots, \text{Var}\,(\Theta))\\ =& (\sigma_0^2, \ldots, \sigma_0^2)\\ =& \sigma_0^2(1, \ldots, 1)\\ =& \sigma_0^2\mathbf{1}^{\text{T}}\end{aligned}$$

where $\mathbf{1}$ is an n-dimensional vector of 1's. So, if we wish to make inferences about the underlying mean, Θ, based on a sample, $X_1, \ldots, X_n$, we calculate $\text{E}_X(\Theta)$, our adjusted expectation for the underlying mean of the population, based on our sample of size n, as

$$\begin{aligned}\text{E}_X(\Theta) =& \text{E}(\Theta)\\ &+ \text{Cov}\,(\Theta, \mathbf{X})\,\text{Var}\,(\mathbf{X})^{-1}\,[\mathbf{X} - \text{E}(\mathbf{X})]\\ =& \mu_0 + \sigma_0^2\mathbf{1}^{\text{T}}\left[\sigma^2\text{I} + \sigma_0^2\text{J}\right]^{-1}[\mathbf{X} - \mu_0\mathbf{1}]\end{aligned}$$

since $\text{E}(\mathbf{X}) = (\mu_0, \ldots, \mu_0)^{\text{T}}$. Now we need to invert the matrix $\sigma^2\text{I} + \sigma_0^2\text{J}$. It is easy to verify (by multiplication) that

$$\left[\sigma^2\text{I} + \sigma_0^2\text{J}\right]^{-1} = \frac{1}{\sigma^2}\text{I} - \frac{\sigma_0^2}{\sigma^2(\sigma^2 + n\sigma_0^2)}\text{J}$$

Therefore,

$$
\begin{aligned}
\mathrm{E}_X(\Theta) =& \mu_0 + \\
& \sigma_0^2 \mathbf{1}^{\mathrm{T}} \left[\frac{1}{\sigma^2} \mathrm{I} - \frac{\sigma_0^2}{\sigma^2(\sigma^2 + n\sigma_0^2)} \mathrm{J} \right] [\mathbf{X} - \mu_0 \mathbf{1}] \\
=& \mu_0 + \\
& \sigma_0^2 \left[\frac{1}{\sigma^2} - \frac{n\sigma_0^2}{\sigma^2(\sigma^2 + n\sigma_0^2)} \right] \mathbf{1}^{\mathrm{T}} [\mathbf{X} - \mu_0 \mathbf{1}] \\
=& \mu_0 + \frac{\sigma_0^2}{\sigma^2 + n\sigma_0^2} [n\bar{X} - n\mu_0] \\
=& \frac{\sigma^2}{\sigma^2 + n\sigma_0^2} \mu_0 + \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2} \bar{X}
\end{aligned}
$$

Notice that although we have conditioned on the whole of $\mathbf{X}$, our adjusted expectation depends only on

$$\bar{X} = \frac{1}{n} \sum_{k=1}^{n} X_k.$$

So, in some sense (to be made precise shortly), if we wish to adjust Θ by $\mathbf{X}$, it is "sufficient" to know only the single number $\bar{X}$, rather than every X_k.

2.6.2 Example (B)

Suppose that we are interested in the average weekly alcohol consumption of male Newcastle University students. We are to take a random sample of 20 such individuals, and they will each measure their alcohol consumption (in units) for a particular week. The units drunk by each individual will be denoted $X_1, \ldots, X_{20}$. Before the measurements are obtained, these are random quantities about which specifications may be made.

Now, since *a priori* we have no individual-specific information about the drinking habits of the students in our sample, it will be reasonable to judge our sample of 20 as a sample from an (essentially infinite) second-order exchangeable population, $X_1, X_2, \ldots$ of possible measurements. Consequently, we can apply the representation theorem, and write

$$X_k = \Theta + \epsilon_k$$

where Θ represents the population average. We need to make specifications $\mathrm{E}(\Theta) = \mu_0$, $\mathrm{Var}(\Theta) = \sigma_0^2$, and $\mathrm{Var}(\epsilon_k) = \sigma^2$. We can make specifications for Θ by thinking about the population average alcohol consumption. Earlier in this part of the course, we thought about making specifications for this quantity. I specified $\mu_0 = 12$ and $\sigma_0^2 = 39$.

Specifications for σ^2 can be made by thinking about the uncertainty you would have about a particular X_k, if there was no uncertainty about Θ (*ie.* $\sigma_0^2 = 0$). So suppose we were told that the population average was exactly 10 units (chosen arbitrarily, but in line with the values we have specified for μ_0 and σ_0^2). What would be your uncertainty about a particular X_k measurement?

I would be 95% sure that the value of X_k would lie in $[0, 40]$, and this is compatible with a specification of $\sigma^2 = 100$ (taking $4\sigma = 40$). Now since $n = 20$, we have

$$\begin{aligned} E_X(\Theta) =& \frac{\sigma^2}{\sigma^2 + n\sigma_0^2}\mu_0 + \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}\bar{X} \\ =& \frac{100}{100 + 20 \times 39}12 + \frac{20 \times 39}{100 + 20 \times 39}\bar{X} \\ =& 1.36 + 0.87\bar{X} \end{aligned}$$

We see that given our specification, a sample of size 20 leads to almost ignoring μ_0 in favour of $\bar{X}$. Thus our adjusted expectation for Θ is not overly sensitive to the specifications we have made, and depends mainly on the data. Suppose now that we carry out the experiment, and obtain $\bar{x} = 15$. Then

$$E_x(\Theta) = 1.36 + 0.87 \times 15 = 14.7$$

This is the sample mean, shrunk very slightly towards the prior mean.

In general, it is true that the adjusted expectation is a weighted average of the prior and sample means, since

$$E_X(\Theta) = \alpha_n\mu_0 + (1 - \alpha_n)\bar{X}$$

where

$$\alpha_n = \frac{\sigma^2}{\sigma^2 + n\sigma_0^2}$$

Note that as $n \to \infty$, $\alpha_n \to 0$. So, for large samples, $E_X(\Theta) \approx \bar{X}$, and the prior mean is ignored. However, for small samples, α_n can be close to 1, corresponding to almost ignoring the data in favour of the prior mean.

2.6.3 Sufficiency

We already have an intuitive understanding of what sufficiency is, but we formalise it in the context of Bayes linear methods as follows.

Definition

A random vector $\mathbf{X}^\star$ is Bayes linear sufficient for the adjustment of a vector $\boldsymbol{\Theta}$, by a vector $\mathbf{X}$, if

$$E_{X^\star}(\boldsymbol{\Theta}) = E_X(\boldsymbol{\Theta})$$

and

$$E_X(\mathbf{X}^\star) = \mathbf{X}^\star$$

The first condition corresponds to our intuition about what sufficiency means. That is, adjusting by the sufficient statistic $\mathbf{X}^\star$, has the same effect as adjusting by all the data, $\mathbf{X}$. The second condition ensures that $\mathbf{X}^\star$ is linearly determined by $\mathbf{X}$. It is clear that the second condition is satisfied if all the elements of $\mathbf{X}^\star$ are linear combinations of elements of $\mathbf{X}$.

Therefore, in order to show that in the case of simple exchangeable adjustments, $\bar{X}$ is Bayes linear sufficient for the adjustment of Θ by **X**, it remains only to confirm that $\mathrm{E}_{\bar{X}}(\Theta) = \mathrm{E}_{X}(\Theta)$, since we know that $\bar{X}$ is a linear combination of the elements of **X**. First note that,

$$\begin{aligned}\bar{X} =& \frac{1}{n}\sum_{k=1}^{n} X_k \\ =& \frac{1}{n}\sum_{k=1}^{n}(\Theta + \epsilon_k) \\ =& \Theta + \frac{1}{n}\sum_{k=1}^{n}\epsilon_k.\end{aligned}$$

So,

$$\begin{aligned}\mathrm{Var}\left(\bar{X}\right) =& \mathrm{Var}\left(\Theta\right) + \frac{1}{n^2} \times n\mathrm{Var}\left(\epsilon_k\right) \\ =& \sigma_0^2 + \frac{1}{n}\sigma^2\end{aligned}$$

and

$$\begin{aligned}\mathrm{Cov}\left(\Theta, \bar{X}\right) =& \mathrm{Cov}\left(\Theta, \Theta\right) \\ =& \mathrm{Var}\left(\Theta\right) = \sigma_0^2,\end{aligned}$$

which gives

$$\begin{aligned}\mathrm{E}_{\bar{X}}(\Theta) =& \mathrm{E}(\Theta) + \frac{\mathrm{Cov}\left(\Theta, \bar{X}\right)}{\mathrm{Var}\left(\bar{X}\right)}[\bar{X} - \mathrm{E}\left(\bar{X}\right)] \\ =& \mu_0 + \frac{\sigma_0^2}{\sigma_0^2 + \frac{1}{n}\sigma^2}(\bar{X} - \mu) \\ =& \mathrm{E}_X(\Theta)\,.\end{aligned}$$

This confirms the sufficiency of $\bar{X}$. Let us now look at the adjusted variance for simple exchangeable adjustments. First note that

$$\begin{aligned}\mathrm{Var}_X\left(\Theta\right) =& \mathrm{Var}\left(\Theta - \mathrm{E}_X(\Theta)\right) \\ =& \mathrm{Var}\left(\Theta - \mathrm{E}_{\bar{X}}(\Theta)\right) \\ =& \mathrm{Var}_{\bar{X}}\left(\Theta\right)\end{aligned}$$

Notice that in this case, the adjusted variance is the same as the variance adjusted by a sufficient statistic. Indeed, this is the case in general. The latter is clearly easier to calculate, so

$$\begin{aligned}
\mathrm{Var}_X(\Theta) =& \mathrm{Var}_{\bar{X}}(\Theta) \\
=& \mathrm{Var}(\Theta) - \frac{\mathrm{Cov}\left(\Theta, \bar{X}\right)^2}{\mathrm{Var}\left(\bar{X}\right)} \\
=& \mathrm{Var}(\Theta) - \frac{\mathrm{Var}(\Theta)^2}{\mathrm{Var}\left(\bar{X}\right)} \\
=& \sigma_0^2 - \frac{\sigma_0^4}{\sigma_0^2 + \frac{\sigma^2}{n}} \\
=& \frac{\sigma^2}{\sigma^2 + n\sigma_0^2}\sigma_0^2 \\
=& \alpha_n \sigma_0^2
\end{aligned}$$

where $\alpha_n = \sigma^2/(\sigma^2 + n\sigma_0^2)$, as before. Note that if we put $\sigma_1^2 = \mathrm{Var}_X(\theta)$, then we have

$$\begin{aligned}
\sigma_1^2 &= \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2} \\
\Rightarrow \frac{1}{\sigma_1^2} &= \frac{\sigma^2 + n\sigma_0^2}{\sigma^2\sigma_0^2} \\
&= \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}
\end{aligned}$$

That is, one can obtain the adjusted *precision* by summing the prior precision and the precision of the data.

2.6.4 Example (B)

$$\begin{aligned}
\mathrm{Var}_X(\Theta) =& \frac{100}{100 + 20 \times 39} \times 39 \\
=& 4.43
\end{aligned}$$

So the uncertainty about the underlying mean has been reduced considerably.

2.6.5 Resolution

Now we can calculate an expression for the resolution of the adjustment of Θ by $\mathbf{X}$.

$$\begin{aligned}
\mathrm{r}_X(\Theta) =& 1 - \frac{\mathrm{Var}_X(\Theta)}{\mathrm{Var}(\Theta)} \\
=& 1 - \frac{\sigma^2}{\sigma^2 + n\sigma_0^2} \\
=& 1 - \alpha_n
\end{aligned}$$

Thus, as $n \to \infty$, $\alpha_n \to 0$ and so $\mathrm{r}_X(\Theta) \to 1$. That is, we are able to resolve as much uncertainty as desired for Θ, by observing increasingly large samples.

2.6.6 Example (B)

For our sample of size 20, the resolution is

$$\mathrm{r}_X\left(\Theta\right) = 1 - \alpha_n = 1 - 0.11 = 0.89$$

How large a sample would we need in order to resolve at least 95% of the prior variance (make $\mathrm{r}_X\left(\Theta\right)$ at least 0.95)? Well, we need to make $\alpha_n < 0.05$. That is

$$\begin{aligned}
&\frac{\sigma^2}{\sigma^2 + n\sigma_0^2} < 0.05 \\
\Rightarrow& 0.95\sigma^2 < 0.05n\sigma_0^2 \\
\Rightarrow& n > \frac{19\sigma^2}{\sigma_0^2} = \frac{19 \times 100}{39} = 48.7
\end{aligned}$$

Thus a sample of size 49 would be required. This is an example of Bayes linear experimental design. It works because adjusted variance does not depend explicitly on the data.

2.6.7 Predictive adjustments

Suppose that we are not just interested in inference for the population mean, Θ, but also in predicting a future observable X_{n+1} (or a collection of future observations). It turns out that $\bar{X}$ is Bayes linear sufficient for $\mathbf{X}$ in this case also (proof almost identical). So

$$\begin{aligned}
\mathrm{E}_X(X_{n+1}) =& \ \mathrm{E}(X_{n+1}) + \frac{\mathrm{Cov}\left(X_{n+1}, \bar{X}\right)}{\mathrm{Var}\left(\bar{X}\right)}\left[\bar{X} - \ \mathrm{E}\big(\bar{X}\big)\right] \\
=& \ \mu_0 + \frac{\mathrm{Var}\left(\Theta\right)}{\sigma_0^2 + \frac{1}{n}\sigma^2}\left[\bar{X} - \ \mathrm{E}\big(\bar{X}\big)\right] \\
=& \ \alpha_n\mu_0 + (1 - \alpha_n)\bar{X}
\end{aligned}$$

as before. That is,

$$\mathrm{E}_X(X_{n+1}) = \ \mathrm{E}_X(\Theta)$$

However, the adjusted variance is slightly different. The variance of X_{n+1} adjusted by $\mathbf{X}$ is given by

$$\begin{aligned}
\mathrm{Var}_X\left(X_{n+1}\right) =& \mathrm{Var}_{\bar{X}}\left(X_{n+1}\right) \\
=& \mathrm{Var}\left(X_{n+1}\right) - \frac{\mathrm{Cov}\left(X_{n+1}, \bar{X}\right)^2}{\mathrm{Var}\left(\bar{X}\right)} \\
=& \sigma_0^2 + \sigma^2 - \frac{[\sigma_0^2]^2}{\sigma_0^2 + \frac{\sigma^2}{n}} \\
=& \frac{(n+1)\sigma_0^2 + \sigma^2}{n\sigma_0^2 + \sigma^2}\sigma^2
\end{aligned}$$

Note in particular that

$$\mathrm{Var}_X\left(X_{n+1}\right) \longrightarrow \sigma^2 \text{ as } n \longrightarrow \infty$$

So the uncertainty corresponding to ϵ_t will always remain, no matter how much data we have. Only the aspect of our uncertainty corresponding to Θ will be resolved.

2.6.8 Maximal resolution

Since $\mathrm{Var}_X\left(X_{n+1}\right)$ decreases to σ^2, as n becomes large, it is clear that the corresponding resolution, $\mathrm{r}_X\left(X_{n+1}\right)$ increases to a limiting value smaller than one. That is

$$\begin{aligned}\mathrm{r}_X\left(X_{n+1}\right) \longrightarrow & 1-\frac{\sigma^2}{\sigma^2+\sigma_0^2} \\ = & \frac{\sigma_0^2}{\sigma^2+\sigma_0^2}\end{aligned}$$

as n becomes large. This limiting resolution is clearly the maximum possible, for any size sample. That is, no matter how large a sample we obtain, we will always have uncertainty about future observables, due to individual variation.

This maximal resolution is important from a design perspective, as it represents an upper bound on what we can hope to learn, against which we may compare alternatives based on finite sample sizes.

2.6.9 Example (B)

For our example with $n=20$, we have

$$\mathrm{Var}_X\left(X_{n+1}\right)=\frac{21\times 39+100}{20\times 39+100}\times 100=104.3$$

which is not far from the limiting adjusted variance of 100. Similarly,

$$\mathrm{r}_X\left(X_{n+1}\right)=1-\frac{104.3}{139}=0.25$$

which is not far from the maximal resolution of

$$1-\frac{100}{139}=0.28$$

2.6.10 Exchangeable adjustments in $[B/D]$

Of course, we can analyse exchangeable random quantities in the same way as any other quantities in $[B/D]$. However, because exchangeable random quantities arise very frequently in statistical contexts, $[B/D]$ has some special features which make analysis for exchangeable data very simple and convenient.

Let's take our Example (B), supposing that we have only 5 observations, $10, 15, 11, 30, 20$. We can declare the prior specifications in $[B/D]$ as follows.

```
BD>element x=12
BD>var v(1,x)=39
BD>var v(2,x)=139
BD>control priorvar=1,betweenvar=1
BD>control infovar=2,modelvar=1
```

Notice that we put $\mathrm{Var}\left(\Theta\right)$ in store 1 and $\mathrm{Var}\left(X_k\right)$ in store 2. The control statements tell $[B/D]$ what variance specifications are in which store. These controls are as follows:

`priorvar`	$\text{Var}(\Theta)$
`betweenvar`	$\text{Cov}(\Theta, X_k)$
`infovar`	$\text{Var}(X_k)$
`modelvar`	$\text{Cov}(X_j, X_k)$

We therefore need `infovar` to point to store 2, and the other controls to point to store 1. Next we declare the data as follows.

```
BD>data x(1)=10
BD>data x(2)=15
BD>data x(3)=11
BD>data x(4)=30
BD>data x(5)=20
```

Obviously, there are ways to read in data from files, but they are a bit clumsy, and we don't need to use them for small amounts of data. Adjustment is done as before.

```
BD>adjust [x/x]
BD>show a
Element     Adjusted Expectation
X              15.4373

BD>show e
Element         Adjusted Expectation
X :        0.6610 X  + 4.0678

BD>show v+
          ---------- Variation -------------              Maximal
Element   Initial    Adjustment    Resolved  Resolution  Resolution
X           39.0000     13.2203      25.7797    0.6610     1.0000
[X]          1            0.3390      0.6610    0.6610     1.0000
```

The output summarises the exchangeable adjustment. Note that the output from `show v+` includes a new column, representing the maximal resolution for the quantity of interest. Clearly this is 1 for Θ.

2.6.11 Predictive adjustment in $[B/D]$

If our interest is in a future X_{n+1}, rather than Θ, then we just need to change one control.

```
BD>control priorvar=2
```

Our prior variance is now that of X_{n+1} rather than Θ. Adjustment takes place as before.

```
BD>adjust [x/x]
BD>show a
Element     Adjusted Expectation
X              15.4373

BD>show e
Element         Adjusted Expectation
```

```
X :        0.6610 X  + 4.0678

BD>show v+
          ---------- Variation ----------                Maximal
Element   Initial    Adjustment  Resolved   Resolution  Resolution
X          139.0000    113.2203   25.7797    0.1855      0.2806
[X]          1            0.8145    0.1855    0.1855      0.2806
```

Notice that the resolutions for X_{n+1} are smaller than for Θ, and that the maximal resolution for X_{n+1} agrees with the calculation we did by hand earlier.

2.7 Summary

The key idea in Bayes linear methods is that given limited specifications, we can update our opinions about collections of random quantities in the light of informative data. Further, in the most general setting, no modelling or distributional assumptions are required.

We started by thinking about how we can quantify means variances and covariances, and how to use some simple matrix theory to work with linear combinations of random quantities, and how we can display specifications in the form of a “belief structure”. We then thought about how we can use data in order to define “adjusted expectation”, which is a kind of “best guess” that is allowed to depend on the outcome of an experiment. We next saw how this adjusted expectation decomposed uncertainty into resolved and adjusted variance matrices, and how we can understand the transition from prior to adjusted variance by examining the eigenstructure of the “resolution transform”.

Finally we turned our attention to “exchangeability”, which links Bayes linear methods to more conventional statistical theory by providing a representation of the problem in terms of a simple statistical model for random sampling. We saw how we can use the exchangeable representation to make inferences for the “population mean” and future observables, and to understand the “large sample” behaviour of such inferential procedures.

We’ve also seen how to use some computer software, $[B/D]$, in order to analyse Bayes linear problems without the user having to bother with the underlying matrix theory or calculations.