Probability

Adam Boult (www.bou.lt)

January 22, 2021

Contents

Preface

This is a live document, and is full of gaps, mistakes, typos etc.

Part I

Probability

Chapter 1

Events, the probability function and the Kolgomorov axioms

1.1 Events

1.1.1 Elementary events

We have a sample space, Ω consisting of elementary events.

All elementary events are disjoint sets.

1.1.2 Non-elementary events

We have a σ-algebra over Ω called F. A σ-algebra takes a set a provides another set containing subsets closed under complement. The power set is an example.

All events E are subsets of Ω

$\forall E \in F E \subseteq \Omega$

1.1.3 Mutually exclusive events

Events are mutually exclusive if they are disjoint sets.

1.1.4 Complements

For each event E, there is a complementary event E^C such that:

$E \vee E^C = \Omega$

$E \wedge E^C = \varnothing$

This exists by construction in the measure space.

1.1.5 Union and intersection

As events are sets, we can define algebra on sets. For example for two events E_i and E_j we can define:

- $E_i \wedge E_j$
- $E_i \vee E_j$

1.2 Kolmogorov axioms

1.2.1 The probability function

For all events E in F, the probability function P is defined.

1.2.2 Measure space

This gives us the following measure space:

(Ω, F, P)

1.2.3 First Kolmogorov axiom

First axiom

The probability of all events is a non-negative real number.

$\forall E \in F[(P(E) \geq 0) \wedge (P(E) \in \mathbb{R})]$

1.2.4 Second Kolmogorov axiom

The probability of one of the elementary events occuring is 1.

The probability of the outcome set is 1.

$P(\Omega) = 1$

1.2.5 Third Kolmogorov axiom

The probability of union for mutually exclusive events is:

$P(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$

1.3 Basic results

1.3.1 Probability of null

$P(\Omega) = 1$

$P(\Omega \vee \varnothing) = 1$

$P(\Omega) + P(\varnothing) = 1$

$P(\varnothing) = 0$

1.3.2 Monotonicity

Consider $E_i \subseteq E_j$:

$E_j = E_i \vee E_k$

$P(E_j) = P(E_i \vee E_k)$

Disjoint so:

$P(E_j) = P(E_i) + P(E_k)$

We know that $P(E_k) \geq 0$ from axiom 1 so:

$P(E_j) \geq P(E_i)$

1.3.3 Bounds of probabilities

As all events are subsets of the sample space:

$P(\Omega) \geq P(E)$

$1 \geq P(E)$

From axiom 1 then know:

$\forall E \in F[0 \leq P(E) \leq 1]$

1.3.4 Union and intersection for null and universal

$P(E \wedge \varnothing) = P(\varnothing) = 0$

$P(E \vee \Omega) = P(\Omega) = 1$

$P(E \vee \varnothing) = P(E)$

$P(E \wedge \Omega) = P(E)$

1.3.5 Separation rule

Firstly:

$P(E_i) = P(E_i \wedge \Omega)$

$P(E_i) = P(E_i \wedge (E_j \vee E_j^C))$

$P(E_i) = P((E_i \wedge E_j) \vee (E_i \wedge E_j^C))$

As the latter are disjoint:

$P(E_i) = P((E_i \wedge E_j) + (E_i \wedge E_j^C))$

1.3.6 Addition rule

We know that:

$P(E_i \vee E_j) = P((E_i \vee E_j) \wedge (E_j \vee E_j^C))$

By the distributive law of sets:

$P(E_i \vee E_j) = P((E_i \wedge E_j^C) \vee E_j)$

$P(E_i \vee E_j) = P((E_i \wedge E_j^C) \vee (E_j \wedge (E_i \vee E_i^C))$

By the distributive law of sets:

$P(E_i \vee E_j) = P((E_i \wedge E_j^C) \vee (E_j \wedge E_i) \vee (E_j \wedge E_i^C))$

As these are disjoint:

$P(E_i \vee E_j) = P(E_i \wedge E_j^C) + P(E_j \wedge E_i) + P(E_j \wedge E_i^C)$

From the separation rule:

$P(E_i \vee E_j) = P(E_i) - P(E_i \wedge E_j) + P(E_j \wedge E_i) + P(E_j) - P(E_j \wedge E_i)$

$P(E_i \vee E_j) = P(E_i) + P(E_j) - P(E_i \wedge E_j)$

1.3.7 Probability of complements

From the addition rule:

$P(E_i \vee E_j) = P(E_i) + P(E_j) - P(E_i \wedge E_j)$

Consider E and E^C:

$P(E \vee E^C) = P(E) + P(E^C) - P(E \wedge E^C)$

We know that E and E^C are disjoint, that is:

$E \wedge E^C = \varnothing$

Similarly by construction:

$E \vee E^C = \Omega$

So:

$P(\Omega) = P(E) + P(E^C) - P(\varnothing)$

$1 = P(E) + P(E^C)$

1.4 Other

1.4.1 Odds

Given a set of outcomes for a variable, the odds of the outcome are defined as:

$o_f = \dfrac{P(E)}{P(E^C)}$

For example, the odds of rolling a 6 are $\dfrac{1}{5}$.

1.4.2 Discrete and continous probability

We know that:

$\sum_y P(X \wedge Y) = P(X)$

So for the continuous case

$P(X) = \int_{-\infty}^{\infty} P(X \wedge Y)dy$

This behaves like the probability for a single event, or multiple events with one fewer event if there were more than 2 events to start with.

1.4.3 Marginalisation

Chapter 2

Conditional probability and Bayes' theorem

2.1 Introduction

2.1.1 Conditional probability

We define conditional probability

$$P(E_i|E_j) := \frac{P(E_i \wedge E_j)}{P(E_j)}$$

We can show this is between 0 and 1.

$$P(E_j) = P(E_i \wedge E_j) + P(\bar{E}_i \wedge E_j)$$

$$P(E_i|E_j) := \frac{P(E_i \wedge E_j)}{P(E_i \wedge E_j) + P(\bar{E}_i \wedge E_j)}$$

We know:

$$P(x_i|y_j) := \frac{P(x_i \wedge y_j)}{P(y_j)}$$

$$P(y_j|x_i) := \frac{P(x_i \wedge y_j)}{P(x_i)}$$

So:

$$P(x_i|y_j)P(y_j) = P(y_j|x_i)P(x_i)$$

$$P(x_i|y_j) = \frac{P(y_j|x_i)P(x_i)}{P(y_j)}$$

Note that this is undefined when $P(y_j) = 0$

Note that for the same event,

$P(x_i|x_j) = \frac{P(x_i \wedge x_j)}{P(x_j)}$

$P(x_i|x_j) = 0$

For the same outcome:

$P(x_i|x_i) = \frac{P(x_i \wedge x_i)}{P(x_i)}$

$P(x_i|x_i) = \frac{P(x_i)}{P(x_i)}$

$P(x_i|x_i) = 1$

2.1.2 Bayes' theorem

From the definition of conditional probability we know that:

$P(E_i|E_j) := \frac{P(E_i \wedge E_j)}{P(E_j)}$

$P(E_j|E_i) := \frac{P(E_i \wedge E_j)}{P(E_i)}$

So:

$P(E_i \wedge E_j) = P(E_i|E_j)P(E_j)$

$P(E_i \wedge E_j) = P(E_j|E_i)P(E_i)$

So:

$P(E_i|E_j)P(E_j) = P(E_j|E_i)P(E_i)$

2.1.3 Independent variables

Events are independent if:

$P(E_i|E_j) = P(E_i)$

Note that:

$P(E_i \wedge E_j) = P(E_i|E_j)P(E_j)$

And so for independent events:

$P(E_i \wedge E_j) = P(E_i)P(E_j)$

2.1.4 Conjugate priors

If the prior $P(\theta)$ and the posterior $P(\theta|X)$ are in the same family of distributions (eg both Gaussian), then the prior and posterior are conjugate distributions

Chapter 3

Entropy

3.1 Entropy

3.1.1 Information

Criteria

Self information measures surprise of outcome. also called a surprisal.

When we observe an outcome we get information. We can develop a measure for how much information is associated with a specific measurement.

Rule 1: Information is always positive

Rule 2: If $P(x) = 1$, the the information for $I(P(x)) = 0$.

Rule 3: If two events are independent, then their information is additive.

- $P(C) = P(A)P(B)$
- $I(P(C)) = I(P(A)P(B))$
- $I(P(A)) + I(P(B)) = I(P(A)P(B))$

Choice of function

A function which satisifes this is $I(P(A)) = -\log(P(A))$

Any base can be used. 2 is most common, information is in units of bit then.

3.1.2 Entropy

Introduction

Entropy measures the expected amount of information produced by a source.

$H(P(x)) = E(I(P(x))$

Entropy is similar to variance, is the sense that both measure uncertainty.

Entropy, however, has no references to specific values of x. If all values were multiplied by 100, or if parts of the distribution were cut up and swapped, entropy would be unaffected.

For a probability function $p(z)$, its entropy is :

$H(p) = -\int p(z) \ln p(z) dz.$

This is a measure of the spread of a distribution.

Negative infinity means no uncertainty

For a multivariate gaussian $H = d/2ln(2\pi e|\Sigma)$.

Part II

Variables

Chapter 4

Variables

4.1 Variables

4.1.1 Random variables

Defining variables

We have a sample space, Ω. A random variable X is a mapping from the sample space to the real numbers:

$X : \Omega \to \mathbb{R}$

We can then define the set of elements in Ω. As an example, take a coin toss and a die roll. The sample space is:

$\{H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6\}$

A random variable could give us just the die value, such that:

$X(H1) = X(T1) = 1$

We can define this more precisely using set-builder notation, by saying the following is defined for all $c \in \mathbb{R}$:

$\{\omega | X(\omega) \leq c\}$

That is, for any number random variable map X, there is a corresponding subset of Ω containing the ωs in Ω which map to less than c.

Multiple variables

Multiple variables can be defined on the sample space. If we rolled a die we could define variables for

- Whether it was odd/even
- Number on the die
- Whether it was less than 3

With more die we could add even more variables

Derivative variables

If we define a variable X, we can also define another variable $Y = X^2$.

4.1.2 Probability mass functions

$P(X = x) = P(\omega | X(\omega) = x)$

For discrete probability, this is a helpful number. For example for rolling a die.

This is not helpful for continuous probability, where the chance of any specific outcome is 0.

4.1.3 Cumulative distribution functions

Definition

Random variables all valued as real numbers, and so we can write:

$P(X \leq x) = P(\omega | X(\omega) \leq x)$

Or:

$F_X(x) = \int_{-\infty}^{x} f_X(u)du$

$F_X(x) = \sum_{x_i \leq x} P(X = x_i)$

Partitions

$P(X \leq x) + P(X \geq x) - P(X = x) = 1$

Interval

$P(a < X \leq b) = F_X(b) - F_X(a)$

4.1.4 Probability density functions

Definition

If continuous, probability at any point is 0. We instead look at probability density.

Derived from cumulative distribution function:

$F_X(x) = \int_{-\infty}^{x} f_X(u)du$

The density function is $f_X(x)$.

Conditional probability distributions

For probability mass functions:

$$P(Y = y|X = x) = \frac{P(Y = y \wedge X = x)}{P(X = x)}$$

For probability density functions:

$$f_Y(y|X = x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

4.2 Multiple variables

4.2.1 Joint and marginal probability

Joint probability

$P(X = x \wedge Y = y)$

Marginal probability

$P(X = x) = \sum_y P(X = x \wedge Y = y)$

$P(X = x) = \sum_y P(X = x|Y = y)P(Y = y)$

4.2.2 Independence and conditional independence

Independence

x is independent of y if:

$\forall x_i \in x, \forall y_j \in y(P(x_i|y_j) = P(x_i)$

If $P(x_i|y_j) = P(x_i)$ then:

$P(x_i \wedge y_j) = P(x_i).P(y_j)$

This logic extends beyond just two events. If the events are independent then:

$P(x_i \wedge y_j \wedge z_j) = P(x_i).P(y_j \wedge z_k) = P(x_i).P(y_j).P(z_k)$

Note that because:

$$P(x_i|y_j) = \frac{P(x_i \wedge y_j)}{P(y_j)}$$

If two variables are independent

$$P(x_i|y_j) = \frac{P(x_i)P(y_j)}{P(y_j)}$$

$P(x_i|y_j) = P(x_i)$

Conditional independence

$P(A \wedge B|X) = P(A|X)P(B|X)$

This is the same as:

$P(A|B \wedge X) = P(A|X)$

Chapter 5

Expected value, conditional expectation and Jensen's inequality

5.1 Moments

5.1.1 Functionals of probabilities

$\phi(P) \in \mathbb{R}$ is a functional on $P(X)$.

Examples include the expectation and variance.

We can define derivatives on these functionals.

$\phi(P) \approx \phi(P^0) + D_\phi(P - P^0)$

Where D_ϕ is linear.

5.1.2 Expected value

Definition

For a random variable (or vector of random variables), x, we define the expected value of $f(x)$ as :

$E[f(x)] := \sum f(x_i)P(x_i)$

The expected value of random variable x is therefore this where $f(x) = x$.

$E(x) = \sum_i x_i P(x_i)$

Linearity of expectation

We can show that $E(x+y) = E(x) + E(y)$:

$E[x+y] = \sum_i \sum_j (x_i + y_j) P(x_i \wedge y_j)$

$E[x+y] = \sum_i \sum_j x_i [P(x_i \wedge y_j)] + \sum_i \sum_j [y_j P(x_i \wedge y_j)]$

$E[x+y] = \sum_i x_i \sum_j [P(x_i \wedge y_j)] + \sum_j y_j \sum_i [P(x_i \wedge y_j)]$

$E[x+y] = \sum_i x_i P(x_i) + \sum_j y_j P(y_j)$

$E[x+y] = E[x] + E[y]$

Expectations of multiples

Expectations

$E(cx) = \sum_i cx P(x_i)$

$E(cx) = c \sum_i x P(x_i)$

$E(cx) = cE(x)$

Expectations of constants

$E(c) = \sum_i c_i P(c_i)$

$E(c) = cP(c)$

$E(c) = c$

Conditional expectation

If Y is a variable we are interested in understanding, and X is a vector of other variables, we can create a model for Y given X.

This is the conditional expectation.

$E[Y|X]$

$E[P(Y|X)Y]$

In the continuous case this is

$E(Y|X) = \int_{-\infty}^{\infty} yP(y|X)dy$

We can then identify an error vector.

$\epsilon := Y - E(Y|X)$

So:

$Y = E(Y|X) + \epsilon$

Here Y is called the dependent variable, and X is called the dependent variable.

Iterated expectation

$E[E[Y]] = E[Y]$

$E[E[Y|X] = E[Y]$

5.1.3 Jensen's inequality

If ϕ is convex then:

$\phi(E[X]) \geq E[\phi(X)])$

Chapter 6

Variance and covariance

6.1 Introduction

6.1.1 Variance

Definition

The variance of a random variable is given by:

$Var(x) = E((x - E(x))^2)$

$Var(x) = E(x^2 + E(x)^2 - 2xE(x))$

$Var(x) = E(x^2) + E(E(x)^2) - E(2xE(x))$

$Var(x) = E(x^2) + E(x)^2 - 2E(x)^2$

$Var(x) = E(x^2) - E(x)^2$

Variance of a constant

$Var(c) = E(c^2) - E(c)^2$

$Var(c) = c^2 - c^2$

$Var(c) = 0$

Variance of multiple

$Var(cx) = E((cx)^2) - E(cx)^2$

$Var(cx) = E(c^2x^2) - [\sum_i cxP(x_i)]^2$

$Var(cx) = c^2E(x^2) - c^2[\sum_i xP(x_i)]^2$

$Var(cx) = c^2[E(x^2) - E(x)^2]$

$Var(cx) = c^2Var(x)$

Link between variance of expectation

$E(x)^2 + Var(x) = E(x)^2 + E((x - E(x))^2)$

$E(x)^2 + Var(x) = E(x)^2 + E(x^2 + E(x)^2 - 2xE(x))$

$E(x)^2 + Var(x) = E(x)^2 + E(x^2) + E(E(x)^2) - E(2xE(x))$

$E(x)^2 + Var(x) = E(x)^2 + E(x^2) + E(x)^2 - 2E(x)E(x))$

$E(x)^2 + Var(x) = E(x^2)$

Covariance

$Var(x + y) = E((x + y)^2) - E(x + y)^2$

$Var(x + y) = E(x^2 + y^2 + 2xy) - E(x + y)^2$

$Var(x + y) = E(x^2) + E(y^2) + E(2xy) - E(x + y)^2$

$Var(x + y) = E(x^2) + E(y^2) + E(2xy) - [E(x) + E(y)]^2$

$Var(x + y) = E(x^2) + E(y^2) + E(2xy) - E(x)^2 - E(y)^2 - 2E(x)E(y)]$

$Var(x + y) = [E(x^2) - E(x)^2] + [E(y^2) - E(y)^2] + E(2xy) - 2E(x)E(y)$

$Var(x + y) = Var(x) + Var(y) + 2[E(xy) - E(x)E(y)]$

We then define:

$Cov(x, y) := E(xy) - E(x)E(y)$

Noting that:

$Cov(x, x) = E(xx) - E(x)E(x)$

$Cov(x, x) = Var(x)$

So:

$Var(x + y) = Var(x) + Var(y) + 2Cov(x, y)$

$Var(x + y) = Cov(x, x) + Cov(x, y) + Cov(y, x) + Cov(y, y)$

$Cov(x, c) = E(xc) - E(x)E(c)$

$Cov(x, c) = cE(x) - cE(x)$

$Cov(x, c) = 0$

6.1.2 Covariance matrix

With multiple events, covariance can be defined between each pair of events, including the event with itself.

The covariance between 2 variables is:

$Cov(x_i, x_j) := E(x_i x_j) - E(x_i)E(x_j)$

Which is equal to:

$Cov(x_i, x_j) = E[x_i - E(x_i)][x_j - E(x_j)]$

We can therefore generate a covariance matrix through:

$\sum = E[(X - E[X])(X - E[X])^T]$

Chapter 7

Higher moments

7.1 Introduction

7.1.1 Moments

Moments

The nth moment of variable X is defined as:

$E[X^n] = \sum_i x_i^n P(x_i)$

The mean is the first moment.

Central moments

The nth central moment of variable X is defined as:

$\mu_n = E[(X - E[X])^n] = \sum_i (x_i - E[X])^n P(x_i)$

The variance is the second central moment.

Standardised moments

The nth standardised moment of variable X is defined as:

$$\frac{E[(X - E[X])^n]}{(E[(X - E[X])^2]^{\frac{n}{2}}} = \frac{\mu_n}{\sigma^n}$$

Kertosis

Kertosis is the third standardised moment.

Skew

Skew is the fourth standardised moment.

Chapter 8

Markov's inequality and Chebyshev's inequality

8.1 Other

8.1.1 Markov's inequality and Chebyshev's inequality

Lemma 1

$E[I_{X \geq a}] = P(X \geq a)$

Consider the indicator function.

$I_{X \geq a}$

This is equal to 0 if X is below a and 1 otherwise.

We can take expectations of this.

$E[I_{X \geq a}] = P(X \geq a).1 + P(X < a).0 = P(X \geq a)$

$E[I_{X \geq a}] = P(X \geq a)$

Lemma 2

$aI_{X \geq a} \leq X$

While X is below a the left side is equal to 0, which holds.

While X is equal to a the left side is equal to X, which holds.

While X is above a the left side is equal to a, which holds.

Markovs inequality

$P(X \geq a) \leq \frac{\mu}{a}$

From above:

$aI_{X \geq a} \leq X$

We can take expectations of both sides:

$E[aI_{X \geq a}] \leq E[X]$

$aP(X \geq a) \leq E[X]$

$P(X \geq a) \leq \frac{\mu}{a}$

Chebyshevs inequality

We know from Markovs inequality that:

$P(X \geq a) \leq \frac{\mu}{a}$

Lets take the variable X to be $(X - \mu)^2$

$P((X - \mu)^2 \geq a) \leq \frac{E[(X - \mu)^2]}{a}$

$P((X - \mu)^2 \geq a) \leq \frac{\sigma^2}{a}$

$P(|X - \mu| \geq \sqrt{a}) \leq \frac{\sigma^2}{a}$

Take a to be a multiple k^2 of the variance σ^2.

$a = k^2\sigma^2$

$P(|X - \mu| \geq k\sigma) \leq \frac{\sigma^2}{k^2\sigma^2}$

$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$

Chapter 9

Characteristic functions

9.1 Characteristic functions

9.1.1 Characteristic functions

Transformations

Summary

Cumulative probability function

$F = \int_{-\infty}^{\infty} xP(x)$

Moment generating function

$F = \int_{-\infty}^{\infty} e^{tx}P(x)$

Characteristic function

$F = \int_{-\infty}^{\infty} e^{itx}P(x)$

Moment generating function

Take random variable X. This has moments we wish to calculate.

We can transform our function in other forms which maintain all of the required information. For example we could also use the cumulative probability function to calculate moments. We now look for an alternative form of the probability density function which allows us to easily calculate moments.

One method is to use the probability density function and the definitions of moments, but there are other options. For example, consider the function:

$E[e^{tX}]$

Which expands to:

$E[e^{tX}] = \sum_{j=1}^{\infty} \frac{t^j E[X^j]}{j!}$

By taking the mth derivative of this, we get

$E[X^m] + \sum_{j=m+1}^{\infty} \frac{t^j E[X^j]}{j!}$

We can then set $t = 0$ to get

$E[X^m]$

Alternatively, see that differentiating m times gets us

$E[X^m e^{tX}]$

If we can get this function, we can then easily generate moments.

The function we need to get is:

$E[e^{tX}]$

In the discrete case this is:

$E[e^{tX}] = \sum_{i=1} e^{tx_i} p_i$

In the continuous case:

$E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} P(x) dx$

Characteristic function

It may not be possible to calculate the integral for the moment generating function. We now look for an alternative formula with which we can generate the same moments.

Consider

$E[e^{itX}]$

As this can be broken down into sinusoidal functions it can more readily be integrated.

This expands to

$E[e^{itX}] = \sum_{j=1}^{\infty} \frac{i^j t^j E[X^j]}{j!}$

By taking the mth derivative we get.

$E[X^m] i^m + \sum_{j=m+1}^{\infty} \frac{t^j E[X^j]}{j!}$

By setting $t = 0$ we then get:

$E[X^m]i^m$

Alternatively see that differentiating m times gets us

$E[(iX)^m e^{itX}]$

So we can get the moment by differentiating m times, and multiplying by i^{-m}.

Inverses of these functions

Moment generating function

Characteristic function

Moments of constants added to variables

$\phi_{X+c}(t) = E[e^{it(X+c)}]$

$\phi_{X+c}(t) = E[e^{itX}e^{i}tc]$

$\phi_{X+c}(t) = e^{itc}E[e^{itX}]$

$\phi_{X+c}(t) = e^{itc}\phi_X(t)$

$\phi_X(t) = e^{-itc}\phi_{X+c}(t)$

Moments of constants multiplied by events

$\phi_{cX}(t) = E[e^{itcX}]$

$\phi_{cX}(t) = \phi_X(ct)$

Taylor series of a characteristic function

$\phi_X(t) = E[e^{itX}]$

$\phi_X(t) = \sum_{j=0}^{\infty} \frac{\phi_X^j(a)(t-a)}{j!}$

Around $a = 0$

$\phi_X(t) = \sum_{j=0}^{\infty} \frac{\phi_X^j(0)(t)}{j!}$

The characteristic function is now given in terms of its moments.

We know:

$\phi_X^j(0) = E[X^j]i^j$

So:

$$\phi_X(t) = \sum_{j=0}^{\infty} \frac{E[X^j] i^j (t)^j}{j!}$$

$$\phi_X(t) = \sum_{j=0}^{\infty} \frac{E[X^j] (it)^j}{j!}$$

We know:

$$\frac{E[X^0](it)^0}{0!} = E[1] = 1$$

$$\frac{E[X^1](it)^1}{1!} = E[X](it) = it\mu_X$$

$$\frac{E[X^2](it)^2}{2!} = \frac{-E[X^2]t^2}{2} = \frac{-(\mu_X + \sigma_X^2)t^2}{2}$$

So:

$$\phi_X(t) = 1 + it\mu_X - \frac{(\mu_X + \sigma_X^2)t^2}{2} + \sum_{j=3}^{\infty} \frac{E[X^j](it)^j}{j!}$$

Part III

Simple probability distributions

Chapter 10

Single observation discrete distributions

10.1 Bernoulli distribution

10.1.1 Introduction

The outcome of a Bernoulli trial is either 0 or 1. We can describe it as:

$P(1) = p$

$P(0) = 1 - p$

With a single parameter p.

10.1.2 Moments of the Bernoulli distribution

The mean of a Bernoulli trial is $E[X] = (1-p)(0) + (p)(1) = p$.

The variance of a Bernoulli trial is $E[(X-\mu)^2] = (1-p)(0-\mu)^2 + (p)(1-\mu)^2 = (1-p)p^2 + p(1-p)^2] = p(1-p)$.

10.2 Discrete

10.2.1 The categorical distribution

Bernoulli with three or more discrete possible outcomes.

10.2.2 Degenerate distribution

Chapter 11

Simple continous distributions

11.1 Continous distributions

11.1.1 Exponential distribution

11.1.2 Weibull distribution

11.1.3 Power law

$P(X) = \frac{\alpha - 1}{a}(\frac{x}{a})^{-\alpha}$

Where a is the lower bound.

$P(X) = 0$ for $X < a$.

Moments of the power law

$E[X^m] = \frac{\alpha - 1}{\alpha - 1 - m}a$

If $m \geq \alpha - 1$ then this is not well defined.

Higher order moments, such that the variance, cannot be identified.

11.1.4 Logistic distribution

The logistic distribution has the cumulative distribution function:

$$F(x) = \frac{1}{1 + e^{-\frac{x-\mu}{s}}}$$

11.1.5 Lvy distribution

Definition

The Lvy distribution is a continuous probability distribution.

The marginal probability is:

$$P(X) = \sqrt{\frac{c}{2\pi}} \frac{e^{-\frac{c}{2(x-\mu)}}}{(x-\mu)^{\frac{3}{2}}}$$

11.2 Other

11.2.1 Discrete uniform distribution

There is a set s such that:

$P(x \in s) = p$

$P(x \notin s) = 0$

Moments of the uniform distribution

The mean is the mean of the set s.

If the set is all numbers of the real line between two values, a and b, then:

The mean is $\frac{1}{2}(a + b)$.

The variance is $\frac{(b-a)^2}{12}$ in the continuous case.

The variance is $\frac{(b-a+1)^2 - 1}{12}$ in the discrete case.

11.2.2 Dirac distribution

11.2.3 Empirical distribution

11.2.4 Laplace distribution

11.2.5 Split-normal distribution

Part IV

Statistics

Chapter 12

Independent and identically distributed variables

12.1 Identically Independently Distributed variables (IID)

12.1.1 IID

Identically distributed

x is identically distributed to y if:

$\forall i(\exists x_i \rightarrow P(x_i) = P(y_i))$

Covariance matrix of IID variables

For IID varaibles, the covariance matrix is:

$\Sigma = \sigma^2 I$

Chapter 13

The weak law of large numbers

13.1 Weak law of large numbers

13.1.1 Weak law of large numbers

The sample mean is:

$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

The variance of this is:

$Var[\bar{X}_n] = Var[\frac{1}{n}\sum_{i=1}^{n} X_i]$

$Var[\bar{X}_n] = \frac{1}{n^2} nVar[X]$

$Var[\bar{X}_n] = \frac{\sigma^2}{n}$

We know from Chebyshevs inequality:

$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$

Use $\bar{X}_n$ as X:

$P(|\bar{X}_n - \mu| \geq \frac{k\sigma}{\sqrt{n}}) \leq \frac{1}{k^2}$

Update k so $k := \frac{k\sqrt{n}}{\sigma}$

$$P(|\bar{X}_n - \mu| \geq k) \leq \frac{\sigma^2}{nk^2}$$

As n increases, the chance that the sample mean lies outside a given distance from the population mean approaches 0.

Chapter 14

Levy's continuity theorem

14.1 Lvy's continuity theorem

14.1.1 Lvy's continuity theorem

Chapter 15

The central limit theorem and the gaussian/normal distribution

15.1 Central limit theorem

15.1.1 Central limit theorem

Generalise weak law of large numbers

Characteristic function of summed IID events

$Z = \sum_{i=1}^{n} Y_i$

$\phi_Z(t) = E[e^{itZ}]$

$\phi_Z(t) = E[e^{it \sum_{i=1}^{n} Y_i}]$

$\phi_Z(t) = E[e^{itY}]^n$

$\phi_Z(t) = \phi_Y(t)^n$

Taylor series: first moments dominate with means

$Z = \sum_{i=1}^{n} Y_i$

$Y = \frac{X}{n}$

$\phi_Z(t) = \phi_Y(t)^n$

$$\phi_Z(t) = \phi_{\frac{X}{n}}(t)^n$$

$$\phi_Z(t) = \phi_X(\frac{t}{n})^n$$

$$\phi_X(t) = 1 + it\mu_X - \frac{(\mu_X + \sigma_X^2)t^2}{2} + \sum_{j=3}^{\infty} \frac{E[X^j](it)^j}{j!}$$

$$\phi_X(\frac{t}{n}) = 1 + i\frac{t\mu_X}{n} - \frac{(\mu_X + \sigma_X^2)(\frac{t}{n})^2}{2} + \sum_{j=3}^{\infty} \frac{E[X^j](i\frac{t}{n})^j}{j!}$$

$$\phi_X(\frac{t}{n}) = 1 + i\frac{t\mu_X}{n} - \frac{(\mu_X + \sigma_X^2)t^2}{2n^2} + \sum_{j=3}^{\infty} \frac{E[X^j](i\frac{t}{n})^j}{j!}$$

Eliminating the imaginary term

We want μ to be 0.

$$Z = \sum_{i=1}^{n} Y_i$$

$$Y = \frac{X - \mu_X}{n}$$

$$\phi_Y(t) = 1 + it\mu_Y - \frac{(\mu_Y + \sigma_Y^2)t^2}{2} + \sum_{j=3}^{\infty} \frac{E[Y^j](it)^j}{j!}$$

$$\mu_Y = E[\frac{X - \mu_X}{n}] = \mu_X - \mu_X n = 0$$

$$\phi_Y(t) = 1 - \frac{\sigma_Y^2 t^2}{2} + \sum_{j=3}^{\infty} \frac{E[Y^j](it)^j}{j!}$$

$$\sigma_Y^2 = E[(\frac{X - \mu_X}{n})^2]$$

$$\sigma_Y^2 = E[\frac{X^2 + \mu_X^2 - 2X\mu_X}{n^2}]$$

$$\sigma_Y^2 = \frac{E[X^2] + E[\mu_X^2] - E[2X\mu_X]}{n^2}] \; \sigma_Y^2 = \frac{E[X^2] - \mu_X^2}{n^2}]$$

$$\sigma_Y^2 = \frac{\sigma_X^2}{n^2}$$

$$\phi_Y(t) = 1 - \frac{\sigma_X^2 t^2}{2n^2} + \sum_{j=3}^{\infty} \frac{E[(\frac{X - \mu}{n})^j](it)^j}{j!}$$

$$\phi_Z(t) = \phi_Y(t)^n$$

$$\phi_Z(t) = [1 - \frac{\sigma_X^2 t^2}{2n^2} + \sum_{j=3}^{\infty} \frac{E[(\frac{X-\mu}{n})^j](it)^j}{j!}]^n$$

$$\phi_Z(t) = [1 - \frac{\sigma_X^2 t^2}{2n^2}]^n$$

Eliminating σ^2

$$Z = \sum_{i=1}^{n} Y_i$$

$$Y = \frac{X - \mu_X}{\sigma n}$$

$$\phi_Y(t) = 1 + it\mu_Y - \frac{(\mu_Y + \sigma_Y^2)t^2}{2} + \sum_{j=3}^{\infty} \frac{E[Y^j](it)^j}{j!}$$

$$\mu_Y = E[\frac{X - \mu_X}{\sigma_X n}] = \mu_X - \mu_X \sigma_X n = 0$$

$$\phi_Y(t) = 1 - \frac{\sigma_Y^2 t^2}{2} + \sum_{j=3}^{\infty} \frac{E[Y^j](it)^j}{j!}$$

$$\sigma_Y^2 = E[(\frac{X - \mu_X}{\sigma n})^2]$$

$$\sigma_Y^2 = E[\frac{X^2 + \mu_X^2 - 2X\mu_X}{\sigma^2 n^2}]$$

$$\sigma_Y^2 = \frac{E[X^2] + \mu_X^2 - 2E[X]\mu_X}{\sigma^2 n^2}$$

$$\sigma_Y^2 = \frac{E[X^2] - \mu_X^2}{\sigma^2 n^2}$$

$$\sigma_Y^2 = \frac{\sigma_X^2}{\sigma^2 n^2}$$

$$\sigma_Y^2 = \frac{1}{n^2}$$

$$\phi_Y(t) = 1 - \frac{t^2}{2n^2} + \sum_{j=3}^{\infty} \frac{E[(\frac{X-\mu}{\sigma n})^j](it)^j}{j!}$$

$$\phi_Z(t) = \phi_Y(t)^n$$

$$\phi_Z(t) = [1 - \frac{t^2}{2n^2} + \sum_{j=3}^{\infty} \frac{E[(\frac{X-\mu}{\sigma n})^j](it)^j}{j!}]^n$$

$$\phi_Z(t) = [1 - \frac{t^2}{2n^2}]^n$$

Preparing for exponential expansion

We know that

$[1 + \frac{x}{n}]^n = e^x$

As $n \to \infty$.

With:

$Z = \sum_{i=1}^{n} Y_i$

$Y = \frac{X - \mu_X}{\sigma n}$

We have:

$\phi_Z(t) = [1 - \frac{t^2}{2n^2}]^n$

With:

$Z = \sum_{i=1}^{n} Y_i$

$Y = \frac{X - \mu_X}{\sigma \sqrt{n}}$

We have:

$\phi_Z(t) = [1 - \frac{t^2}{2n}]^n$

Which tends towards

$\phi_Z(t) = e^{-\frac{1}{2}t^2}$

Rescaling

The average of random variables, less their mean, and divided by their standard deviation multiplied by the square root of the sample size, follows a normal distribution as n increases.

What does this say about the actual distribution of sample averages?

$Z = \sum_{i=1}^{n} Y_i$

$Y_i = \frac{X_i - \mu_X}{\sigma_X \sqrt{n}}$

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

$Y = \frac{X}{n}$

Let's create Q.

$$Q = \frac{Z\sigma_X}{\sqrt{n}} + \mu_X$$

$$Q = \frac{(\sum_{i=1}^{n} Y_i)\sigma_X}{\sqrt{n}} + \mu_X$$

$$Q = \frac{(\sum_{i=1}^{n} (\frac{X_i - \mu_X}{\sigma_X \sqrt{n}}))\sigma_X}{\sqrt{n}} + \mu_X$$

$$Q = \sum_{i=1}^{n} (\frac{X_i - \mu_X}{n}) + \mu_X$$

$$Q = \sum_{i=1}^{n} (\frac{X_i - \mu_X}{n} + \frac{\mu_X}{n})$$

$$Q = \sum_{i=1}^{n} (\frac{X_i}{n})$$

This is the sample average.

$$\phi_Q(t) = \phi_{\frac{Z\sigma_X}{\sqrt{n}} + \mu_X}(t)$$

$$\phi_Q(t) = \phi_Z(\frac{t\sigma_X}{\sqrt{n}}) e^{it\mu_X}$$

$$\phi_Z(\frac{t\sigma_X}{\sqrt{n}}) = e^{-\frac{1}{2}(\frac{t\sigma_X}{\sqrt{n}})^2}$$

$$\phi_Z(\frac{t\sigma_X}{\sqrt{n}}) = e^{-\frac{1}{2}\frac{t^2\sigma_X^2}{n}}$$

$$\phi_Q(t) = e^{-\frac{1}{2}\frac{t^2\sigma_X^2}{n}} e^{it\mu_X}$$

Normal distribution

We name the normal distribution this function when $n = 1$

$$N(\mu_X, \sigma_X^2) = e^{-\frac{1}{2}\frac{t^2\sigma_X^2}{n}} e^{it\mu_X}$$

$$N(\mu_X, \sigma_X^2) = e^{-\frac{1}{2}t^2\sigma_X^2} e^{it\mu_X}$$

Getting the probability distribution function

$\phi_X(t) = e^{-\frac{1}{2}t^2\sigma_X^2} e^{it\mu_X}$

$\phi_X(t) = e^{-\frac{1}{2}t^2\sigma_X^2} [\cos(t\mu_X) + i\sin(t\mu_X)]$

15.2 Convergence

15.2.1 Convergence in distribution (converge weakly)

15.2.2 Convergence in probability and o-notation

Introduction

Converges in probability

$P(distance(X_n, X) > \epsilon) \to 0$

For all ϵ.

$X_n \to^P X$

Little o notation

Little o notation is used to describe convergence in probability.

$X_n = o_p(a_n)$

mean that

$\frac{X_n}{a_n}$

Converges to 0 and n approaches something

Can be wrtiten:

$\frac{X_n}{a_n} = o_p(1)$

Big O notation

Big O notation is used to describe boundedness.

$X_n = O_p(a_n)$

means that:

If something is little o, it is big O.

15.2.3 Almost sure convergence

X_n converges almost surely to X if:

$d(X_n, X) \to 0$

Where $d(X_n, X)$ is a distance metric.

$X_n \to^{as} X$

15.3 Gaussian distributions

15.3.1 Gaussian

$$f_x = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma}}$$

15.3.2 The error function and the complementary error function

15.3.3 Multivariable Gaussian distribution

Definition

For univariate:

$x \sim N(\mu, \sigma^2)$

We define the multivariate gaussian distribution as the distribution where any linear combination of components are gaussian.

For multivariate:

$X \sim N(\mu, \Sigma)$

Where μ is now a vector, and Σ is the covariance matrix.

Density function is :

$$f_x = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

For normal gaussian it is:

$$f_x = \frac{1}{\sqrt{2\pi|\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2)}$$

This is the same wher $n = 1$.

Singular Gaussians

Need det $|\Sigma|$ and Σ^{-1}. These rely on the covariance matrix not being degenerate.

If the covariance matrix is degenerate we can instead use the pseudo inverse, and the pseudo determinant.

Chapter 16

Statistics

16.1 Creating statistics

16.1.1 Creating statistics

We take a sample from the distribution.

$x = (x_1, x_2, ..., x_n)$

A statistic is a function on this sample.

$S = S(x_1, x_2, ..., x_n)$.

16.2 Moments of statistics

16.2.1 Bias from single and joint estimation

Bias from single estimation

$\mathbf{x}_i$ and $\mathbf{z}_i$ are not independent, so we cannot estimate just $y_i = \mathbf{x}_i\theta$.

Bias from joint estimation

We could estimate our equation with a single ML algorithm.

$y_i = f(\mathbf{x}_i, \theta) + g(\mathbf{z}_i) + \epsilon_i$

For example, using LASSO.

However this would introduce bias into our estimates for θ.

Bias from iterative estimation

We could iteratively estimate both θ and $g(\mathbf{z}_i)$.

For example iteratvely doing OLS for θ and random forests for z_i.

This would also introduce bias into θ.

16.3 Asymptotic properties of statistics

16.3.1 Asymptotic distributions

$f(\hat{\theta}) \rightarrow^d G$

Where G is some distribution.

16.3.2 Asymptotic mean and variance

16.3.3 Asymptotic normality

Many statistics are asymptotically normally distribution.

This is a result of the central limit theorem.

For example:

$\sqrt{n}S \rightarrow^d N(s, \sigma^2)$

Confidence intervals for asymptotically normal statistics

We have the mean and variance, and know the distribution. This allows us to calculare confidence intervals.

Chapter 17

Order statistics

17.1 Order statistics

17.1.1 Order statistics

Defining order statistics

The kth order statistic is the kth smallest value in a sample.

$x_{(1)}$ is the smallest value in a sample, the minimum.

$x_{(n)}$ is the largest value in a sample, the maximum.

Probability distributions of order statistics

The probability distribution of order statistics depends on the underlying probability distribution.

Probability distribution of sample maximum

If we have:

$Y = \max \mathbf{X}$

The probability distribution is:

$P(Y \leq y) = P(X_1 \leq y, X_2 \leq y, ..., X_n \leq y)$

If these are iid we have:

$P(Y \leq y) = \prod_i P(X_i \leq y)$

$F_y(y) = F_X(y)^n$

The density function is:

$f_y(y) = nF_X(y)^{n-1} f_x(y)$

Probability distribution of the sample minimum

If we have:

$Y = \min \mathbf{X}$

The probability distribution is:

$P(Y \leq y) = P(X_1 \geq y, X_2 \geq y, ..., X_n \geq y)$

If these are iid we have:

$P(Y \leq y) = \prod_i P(X_i \geq y)$

$F_y(y) = [1 - F_X(y)]^n$

The density function is:

$f_y(y) = -n[1 - F_X(y)]^{n-1} f_x(y)$

Part V

More probability distributions

Chapter 18

Repeated observations discrete distributions

18.1 Binomial

18.1.1 Binomial distribution

If we repeat a Bernoulli trials with the same parameter and sum the results, we have the binomial distribution.

We therefore have two parameters, p and n.

$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

18.1.2 Moments of the binomial distribution

The mean is np, which can be seen as the trials are independent.

Similarly, the variances can be addeded together giving $np(1-p)$.

18.1.3 Multinomial distribution

The mass function for the binomial case is:

$f(x) = \frac{n!}{x!(n-x)!} p^k (1-p)^{n-k}$

18.1.4 The multinomial distribution

This generalises the binomial distribution where there are more than 2 outcomes.

$f(x_1, ..., x_n) = \frac{n!}{\prod_i x_i!} \prod_i p_i^{x_i}$

18.2 Poisson

18.2.1 Poisson distribution

18.2.2 Definition

We can use the Poisson distribution to model the number of indepedent events that occur in an a time period.

For a very short time period the chance of us observing an event is a Bernoulli trial.

$P(1) = p$

$P(0) = 1 - p$

18.2.3 Chance of no observations

Let's consider the chance of repeatedly getting 0: $P(0; t)$.

We can see that: $P(0; t + \delta t) = P(0; t)(1 - p)$.

And therefore:

$P(0; t + \delta t) - P(0; t) = -pP(0; t))$

By setting $p = \lambda \delta t$:

$\frac{P(0; t + \delta t) - P(0; t)}{\delta t} = -\lambda P(0; t))$

$\frac{\delta P(0; t)}{\delta t} = -\lambda P(0; t)$

$P(0; t) = Ce^{-\lambda t}$

If $t = 0$ then $P(0; t) = 0$ and so $C = 1$.

$P(0; t) = e^{-\lambda t}$

18.2.4 Deriving the Poisson distribution

Chapter 19

Extreme value distributions

19.1 Extreme value distributions

19.1.1 Type-I - Gumbel distribution

The probability function is:

$$f(x) = \frac{1}{\beta} e^{-(\frac{x-\mu}{\beta} + e^{-\frac{x-\mu}{\beta}})}$$

We can use:

$$z = \frac{x-\mu}{\beta}$$

To get:

$$f(x) = \frac{1}{\beta} e^{-(z+e^{-z})}$$

Link to the logistic function

The difference between two draws from a Gumbel distribution is drawn from the logistic function.

19.1.2 Type-II - Frechet distribution

19.1.3 Type-III - Reversed Weibull distribution

Chapter 20

Mixture distributions

20.1 Mixture models

20.1.1 Gaussian Mixture Models

Mixture models

We have a latent variable which is part of the process

The variable is distributed according to parametric distribution, but parameters are different for differnet latent classes.

There are K latent classes, and so K sets of parameters.

The population is weighted into the K classes.

We have a distribution, but we have different parameters for the distribution for different populations.

For example we could observe the height of men and women, where both are normally distributed but with different parameters.

Where there is a normal distribution, this is a Gaussian mixture model.

If there is more than one variable to observe, this is a multivariate Gaussian mixture model.

Gaussian Mixture Models (GMM)

In a Gaussian Mixture Model each non latent variable has a normal distriubtion with a mean and variance. For multiple variables there is a covariance matrix.

Part VI

Stochastic processes

Chapter 21

Stochastic processes and their moments

21.1 Introduction to processes

21.1.1 Stochastic processes

In a stochastic process we have a mapping from a variable (time) to a random variable.

Discrete and continuous time

Time could be discrete, or continuous.

Temperature over time is a stochastic process, as is the number of cars sold each day.

Discrete and continous state space

The state space for temperature is continous, the number of people on the moon is discrete.

21.1.2 Stochastic evolution

We can describe processes by their evolution.

$p(x_t|x_{t-1}...)$

21.1.3 Gaussian processes

21.1.4 Moments of stochastic processes

21.1.5 Autocovariance and autocorrelation

Autocovariance

$AC(a, b) = cov(X_a, X_b)$

Autocorrelation

The autocorrelation between two time periods is their covariance, normlised by their variances

$$AC(a, b) = \frac{E[(X_a - \mu_a)(X_b - \mu_b)]}{\sigma_a \sigma_b}$$

This is also called serial correlation.

Chapter 22

White noise, and weak- and wide-sense stationarity

22.1 Stationarity

22.1.1 Weak- and wide-sense stationarity

Unconditional probabilities don't change over time.

So GDP would not be stationary, but random noise would. A random walk is not stationary, because the variance increases over time.

22.1.2 Weak-sense stationary

Mean and autocovarinance don't change over time.

22.1.3 Wide-sense stationary

All moments are the same.

22.1.4 Unit roots

22.2 Introduction

22.2.1 White noise

Variables at each time are indepdendent.

Chapter 23

Random walks

23.1 Random walks

23.1.1 Random walks

Chapter 24

Martingale processes

24.1 Introduction

24.1.1 Martingale property

For a process with the Martingale property, the expected value of all future variables is the current state.

This only restricts expectations.

$E(X_{n+1}|X_0, ..., X_n) = X_n$

Chapter 25

Markov processes

25.1 Introduction

25.1.1 Markov property

For a process with the Markov property, only the current state matters for all probability distributions.

$P(x_{t+n}|x_t) = P(x_{t+n}|x_t, x_{t-1}...)$

25.2 Markov chains

25.2.1 Finite state Markov chains

Transition matrices

This shows the probability for moving between discrete states.

We can show the probability of being in a state by multiplying the vector state by the transition matrix.

Mv

Time-homogenous Markov chains

For time-homogenous Markov chains the transition matrix is independent of time.

For these we can calculate the probability of being in any given state in the future:

$M^n v$

This becomes independent of v as we tend to infinity. The initial starting state does not matter for long term probabilities.

How to find steady state probability?

$Mv = v$

The eigenvectors! With associated eigenvector 1. There is only one eigenvector. We can find it by iteratively multiplying any vector by M.

25.2.2 Infinite state Markov chains

Markov model description We can represent the transition matrix as a series of rules to reduce the number of dimensions $P(x_t|y_{t-1}) = f(x, y)$

can represent states as number, rather than atomic. could be continuous, or even real.

in more complex, can use vectors.

25.3 Hidden Markov Models

25.3.1 Introduction

As well as the Markov process X, we have another process Y which depends on X.

25.4 Dynamic Bayesian networks

25.4.1 Introduction

Chapter 26

Multivariate time series

26.1 Multiple time series

26.1.1 Cointegration

If we have multiple variables, we can explore the order of integration of linear combinations.

If two series have time trends, a linear combination of them could remove this.

26.1.2 Exogeneity

Contemporaneous exogeneity

$Cov(x_{it}, u_{it}) = 0$

Strict exogeneity

$Cov(x_{is}, u_{it}) = 0)$

This is stronger than contemporeous, all periods.

Shocks don't affect future outcomes.

Sequential exogeneity

Sequential exogeneity: a bit looser than strict exogeneity. only holds when $s \leq t$.

So shocks can affect, but only in future.

26.1.3 Introduction

Weak stationary processes can be decomposed to a deterministic and a stochastic component.

Chapter 27

Bayesian networks

27.1 Bayesian networks

27.1.1 Bayesian networks

Chapter 28

Survival functions

28.1 Introduction

28.1.1 Survival functions

Part VII

Discrete-time stochastic processes

Chapter 29

Orders of integration

29.1 Introduction

29.1.1 Orders of integration

How many diffs do you need to do to get a stationary process?

If something is first order integrated it is $I(1)$.

29.1.2 Trend stationary

If we can remove the trend as a function, eg linear or non-linear growth, and the rest is stationary, then the process is trend stationary

29.1.3 Seasonal and non-seasonal trends

We can model the process as:

$y_t = \mu_t + f(t) + \epsilon_t$

29.1.4 Cyclical fluctuations

We can have shocks having effects over time.

This is separate to trends.

Chapter 30

Auto-Regressive processes, Moving-Average processes and Wold's theorem

30.1 Autoregressive model

30.1.1 Autoregressive models (AR)

AR(1)

Our basic model was:

$x_t = \alpha + \epsilon_t$

We add an autoregressive component by adding a lagged observation.

$x_t = \alpha + \beta x_{t-1} + \epsilon_t$

AR(p)

AR(p) has p previous dependent variables.

$x_t = \alpha + \sum_{i=1}^{p} \beta_i x_{t-i}$

Propagation of shocks

A shock bumps up the output variable, which bumps up output variables forever, at a decreasing rate.

30.1.2 Testing for stationarity with Dickey-Fuller (DF) and Augmented Dicky-Fuller (ADF)

Stationarity

Unit roots

Integration order

Dickey-Fuller

The Dickey-Fuller test tests if there is a unit root.

The AR(1) model is:

$y_t = \alpha + \beta y_{t-1} + \epsilon_t$

We can rewrite this as:

$\Delta y_t = \alpha + (\beta - 1) y_{t-1} + \epsilon_t$

We test if $\beta - 1) = 0$.

If the coefficient on the last term is 1 we have a random walk, and the process is non-stationary.

If the last term is < 1 then we have a stationary process.

Variation: Removing the drift

If our model has no intercept it is:

$y_t = \beta y_{t-1} + \epsilon_t$

$\Delta y_t = (\beta - 1) y_{t-1} + \epsilon_t$

Variation: Adding a deterministic trend

If our model has a time trend it is:

$y_t = \alpha\beta y_{t-1} + \gamma t + \epsilon_t$

$\Delta y_t = \alpha + (\beta - 1) y_{t-1} + \gamma t + \epsilon_t$

Augmented Dickey-Fuller

We include more lagged variables.

$y_t = \alpha + \beta t + \sum_i^p \theta_i y_{t-i} + \epsilon_t$

If no unit root, can do normal OLS?

30.1.3 Autoregressive Conditional Heteroskedasticity (ARCH)

Variance of the AR(1) model

The standard AR(1) model is:

$y_t = \alpha + \beta y_{t-1} + \epsilon_t$

The variance is:

$Var(y_t) = Var(\alpha + \beta y_{t-1} + \epsilon_t)$

$Var(y_t)(1 - \beta^2) = Var(\epsilon_t)$

Assuming the errors are IID we have:

$$Var(y_t)) = \frac{\sigma^2}{1 - \beta^2}$$

This is independent of historic observations, which may not be desirable.

Conditional variance

Consider the alternative formulation:

$y_t = \epsilon_t f(y_{t-1})$

This allows for conditional heteroskedasticity.

30.2 Moving average models

30.2.1 Moving Average models (MA)

We add previous error terms as input variables

MA(q) has q previous error terms in the model

Unlike AR models, the effects of any shocks wear off after q terms.

This is harder to fit the OLS, the error terms themselves are not observed.

30.3 Autoregressive Moving Average models

30.3.1 Autoregressive Moving Average models (ARMA)

We include both AR and MA

Estimted using Box-Jenkins

30.3.2 Autoregressive Integrated Moving Average models (ARIMA)

Uses differences to remove non statiority

Also estiamted with box-jenkins

30.3.3 Seasonal ARIMA

30.4 Wold's theorem

30.4.1 Introduction

Chapter 31

Vector Autoregression (VAR)

31.1 Vector Autoregression (VAR)

31.1.1 Vector Autoregression (VAR)

We consider a vector of observables, not just one

Autoregressive (AR) model for a vector.

VAR(p) looks p back.

The AR(p) model is:

$y_t = \alpha + \sum_{i=1}^{p} \beta y_{t-i} + \epsilon_t$

VAR(p) generalises this to where y_t is a vector. We define VAR(p) as:

y_t

$y_t = c + \sum_{i=1}^{p} A_i y_{t-i} + \epsilon_t$

31.1.2 VAR impulse response

31.1.3 Bayesian VAR

31.2 Structural models

31.2.1 Autoregressive Distributed Lag (ARDL) model

Include lagged y and lagged x (and current x)

Chapter 32

ARMAX

32.1 ARMAX

32.1.1 ARMAX

32.1.2 ARIMAX

32.1.3 SARIMA

Chapter 33

Partial Adjustment Model (PAM)

33.1 Partial Adjustment Model

33.1.1 Partial Adjustment Model

Estimating a static model

We start by estimating a static model.

$y_t = \alpha + \theta x_t + \gamma_t$

Equilibrium

We then use this form an equilibrium for y_t, y_t^*.

$y_t^* = \hat{\alpha} + \hat{\theta} x_t$

The process depends on the difference from this equilibrium.

$y_t - y_{t-1} = \beta(y_t^* - y_{t-1}) + \epsilon_t$

$y_t - y_{t-1} = \beta(\hat{\alpha} + \hat{\theta} x_t - y_{t-1}) + \epsilon_t$

$y_t = \beta\hat{\alpha} + \beta\hat{\theta} x_t + (1-\beta) y_{t-1} + \epsilon_t$

$y_t = \alpha y_{t-1} + (1-\beta)(y_t^* - y_{t-1}) + \epsilon$

The higher β, the slower the adjustment.

If stationary, can we can use OLS.

Chapter 34

Error Correction Model

34.1 Error Correction Model

34.1.1 Error Correction Model

Static model

Like PAM we start with static estimator.

The ECM

The ECM does a regression with first differences, and includes lagged error terms.

We start with a basic first-difference model.

$\Delta y_t = \Delta x_t$

We could also expand this to include laggs for both x and y. Here we don't.

We know that long term $y_t = \theta x_t$. We use the error from this in a first difference model.

$\Delta y_t = \alpha \Delta x_t + \beta(y_{t-1} - \theta x_{t-1})$

Page on identifying error terms

Also, page on Vector Error Correction Model (VECM)

Part VIII

Continuous-time stochastic processes

Chapter 35

Wiener processes and Brownian motion

35.1 Wiener proceses

35.1.1 Independent increments

The changes in any non-overlapping time increments are independent.

Formally:

$t_0 < t_1 < t_2 < ... < t_m$

With X_t

$X_{t_1} - X_{t_0}$ is indepentent from $X_{t_2} - Xt_1$ etc.

35.1.2 Wiener processes

A Wiener process is a process W_t with independent increments, which: + Is continuous + Has normally distributed increments.

Can be constructed as limit of random walk. Can also be constructed as integral of Gaussian noise?

35.2 Brownian motion

35.2.1 Brownian motion

brownian motion in stats. given we start at a, what is chance be end up at b? normal. do 1d then multi d

Chapter 36

Stochastic differential equations

Part IX

Other

Chapter 37

Redundant Array of Independent Disks (RAID)

37.1 Introduction

37.1.1 Introduction

Take physical disk drives, create logical disk drives.

37.1.2 Striping

A single file is spread over multiple disks.

Can be done at bit/byte/block level.

37.1.3 Mirroring

Same data on multiple drives

37.1.4 Parity

Used for error detection.

In protocol for sending/saving we say that all bits must be even (eg 1100) (or odd)

For given bit of info, we add parity bit to guarantee that bit is indeed even/odd. 1100 becomes 11000; 1000 becomes 10000

Cannot correct errors, just detect them. only detects if odd number of errors.

Parity can be stored on dedicate disk, distributed.

alternative to parity bit: hamming code

37.1.5 RAID levels

RAID 0: Uses striping across disks, but no redunency. allows for improved read/write times. if any drive fails, all fail RAID 1: Data written identically to two drives. read times increased, as with raid 0, due to mirroring. writing is slower. no parity or striping RAID 2: bit level striping and hamming code. rarely used RAID 3: rarely used. byte level stripping. Dedicated parity disk RAID 4: dedicated parity disk RAID 5: block level striping, distributed parity RIAD 6: double distributed parity

37.1.6 ZFS

RAID-Z: Under ZFC, similar to RAID 5

37.1.7 Off-site backups

Part X

Sampling

Chapter 38

Rejection sampling

38.1 Direct sampling

38.1.1 Density estimation through direct sampling

I THINK THE STUFF HERE IS LIMITATIONS TO REJECTION SAMPLING??

DIRECT SAMPLING IS DOING PHYSICAL SAMPLES, MANUALLY PICKING BALLS FROM URL ETC?

There is distribution $P(x)$ which we want to know more about.

If the function was closed, we could estimate it by using values of x.

38.1.2 Limitations of direct sampling

However if the function does not have such a form, we cannot do that.

We can't plug in values, because the function is complex.

Sometimes we may know a function of the form:

$f(x) = cP(x)$

That is, a multiple of the function.

This can happen from Bayes' theorem:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

We may be able to estimate $P(x|y)$ and $P(y)$, but not $P(x)$

This means be have

$P(y|x) = cP(x|y)P(y)$

38.2 Acceptance-rejection sampling

38.2.1 Introduction

Used to sample from propability distribution function.

Useful when can't use direct sampling, because no closed form.

MORE GENERALLY FRAME THESE FIRST AS SAMPLING FROM PROBABILITY FUNCTION.

Generate pairs of (x, y). If $y < P(x)$ then keep x.

Metropolis-Hastings and Gibb's sampling are extensions of this.

Chapter 39

Markov chain Monte Carlo sampling

39.1 Markov Chain Monte Carlo (MCMC) methods

39.2 Metropolis-Hastings algorithm

39.2.1 The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm creates a set of samples x such that the distribution of the samples approaches the goal distribution.

Initialisation

The algorithm takes an arbitrary starting sample x_0. It then must decide which sample to consider next.

Generation

It does this using a Markov chain. That is, there is a map $g(x_j, x_i)$.

This distribution is generally a normal distribution around x_i, making the process a random walk.

Acceptance

Now we have a considered sample, we can either accept or reject it. It is this step that makes the end distribution approximage the function.

We accept if $\frac{f(x_j)}{f(x_i)} > u$, where u is a random variable between 0 and 1, generated each time.

We can calculate this because we know this function.

Properties

39.3 Gibb's sampling

39.3.1 Gibb's sampling

Introduction

As with Metropolis-Hastings, we want to generate samples for $P(X)$ and use this to approximate its form.

We do this by using the conditional distribution. If X is a vector then we also have:

$P(x_j|x_0, ..., x_{j-1}, x_{j+1}, ..., x_n)$

We use our knowledge of this distribution.

Start with vector x_0.

This has components $x_{0,j}$

To form the next vector x_1 we loop through each component.

$P(x_{1,0}|x_{0,0}, x_{0,1}, ..., x_{0,n})$

We use this to form $x_{1,0}$

However after th the first component we update this so it uses the updated variables.

$P(x_{1,k}|x_{1,0}, ..., x_{1,k-1}, x_{0,k}, ..., x_{0,n}$

This means we only need to know the conditional distributions.

Chapter 40

Sampling from processes

40.1 Introduction

Part XI

Stochastic methods

Chapter 41

Stochastic methods for integration

41.1 Introduction

Chapter 42

Stochastic optimisation

42.1 Random search

42.1.1 Random search

We start with a random set of parameters, x.

We then loop through the following:

- We define a search space local to our current selection.
- We randomly select a point from this space.
- We compare the new point to our current point. If the new point is better we move to that.

42.1.2 Random optimisation

This is similar to random search, however we use a multivariate Gaussian distribution around our current point rather than a hypersphere.

42.1.3 Simulated annealing

Introduction

We can use a version of Metropolis-Hastings to find the global maximum of a function $f(x)$.

We start with an arbitrary point x_0.

We move randomly from this to identiy a candidate point x_c.

We accept this with probability depending on the the relationship between x_0 and x_c.

This process will converge on the global maximum.

Hyperparameter

There is a hyperparameter for selection. At the extreme this becomes a greedy function.

42.2 Bayesian optimisation

42.2.1 Bayesian optimisation

Introduction

If we have sampled from the hyperparameter space we know something about the shape.

Can we use this to inform where we should next look?

The shape of the function is $y = f(\mathbf{x})$

We have observations $\mathbf{X}$ and $\mathbf{y}$.

So what's our posterior, $P(y|\mathbf{X}, \mathbf{y})$?

Exploration and exploitation

The can be a tradeoff between:

- Exploring - which gives us a better shape for $y = f(x)$; and
- Exploiting - which gives us a better estimate for the global optimum.

The surrogate function

We do not know $y = f(x)$, but we model it as:

$z(x) = y(x) + \epsilon$

We can then maximise z

Proposing new candidates

We want an algorithm which maps from our history of observations to a new candidate.

There are different approaches:

- Probability of improvement - Choosing one with the highest chance of a more optimal value
- Expected improvement - Choosing one with the biggest expected increase in the optimal value
- Entropy search - choosing one which reduces uncertainty about the global maximum.

42.3 Evolutionary algorithms

42.3.1 Evolutionary algorithms

Initialisation

We generate a set of candidate parameter values, x.

Evaluate using the fitness function

We evaluate each of these against a fitness function (the function we are optimising).

We assign fitness values to each individual.

Crossover and mutation

We generate a second generation. We select "parents" randomly using the fitness values as weightings.

The values of the new individual are a function of the values of the parents, and noise (mutation).

We do this for each member in the next generation.

We iterate this process across successive generations.

42.4 Differential evolution

42.4.1 Differential evolution

42.5 Particle swarms

42.5.1 Particle swarms

Chapter 43

Calculus of stochastic processes

43.1 Introduction

43.1.1 Ito integrals

43.1.2 Stochastic differential equations

Chapter 44

Lossy compression

44.1 Lossy compression