

Advanced Time Series Analysis

Probability:

A *very* quick summary.

A probability space (Ω, \mathcal{F}, P) (a model of a random experiment) consists of:

1. The sample space, Ω - the outcomes of a random experiment.
2. The set of random events, \mathcal{F} – subsets of Ω .
3. The probability measure P , a function that associates a real number $P(A)$ to every element A of \mathcal{F} .

Points:

- \mathcal{F} contains all the events to which probabilities can be assigned - called *measurable sets*.
- \mathcal{F} is a σ -field. In other words,
 1. $\Omega \in \mathcal{F}$.
 2. If $A \in \mathcal{F}$ then $\Omega - A \in \mathcal{F}$. (Closed under complementation)
 3. If $A_i \in \mathcal{F}$ for $i = 1, 2, 3, \dots$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$. (Closed under countable union.)

- Given a collection of subsets C , $\sigma(C)$ denotes the smallest σ -field containing C ; "the σ -field generated by C ".
- P must obey the axioms of probability:

1. $P(A) \geq 0$, all $A \in \mathcal{F}$.
2. $P(\Omega) = 1$.
3. (Additivity) If A and B are disjoint (no points in common) then
$$P(A \text{ or } B) = P(A) + P(B).$$

Further, (*countable* additivity): if $A_i \in \mathcal{F}$ for $i = 1, 2, 3, \dots$ are a disjoint collection, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

- Going to the limit of infinite collections of events is a key feature of a useful probability model.

Examples:

Random Variables

Take $\Omega = \mathbb{R}$ - the real line $(-\infty, +\infty)$.

Then a probability space is defined by

$$(\mathbb{R}, \mathcal{B}, \mu)$$

where \mathcal{B} denotes the Borel field of \mathbb{R} , and μ denotes a probability measure.

- \mathcal{B} is the σ -field generated by the half-lines $(-\infty, x]$, all $x \in \mathbb{R}$.
 - That is, it is the smallest collection of sets containing the half lines and closed under complementation and countable union.
 - \mathcal{B} includes all open sets of \mathbb{R} !
 - \mathcal{B} is defined this way as the largest collection of sets of real numbers to which probabilities can be consistently assigned.
- Write X as the random variable having this distribution – taking values $x \in \mathbb{R}$.
- μ is usually represented by a non-decreasing function

$$F(x) = \mu(\{X \leq x\})$$

called the cumulative distribution function (c.d.f.).

- μ and F are equivalent representations of the distribution.
- If F is everywhere differentiable, the distribution of X is said to be continuous and can be represented by the probability density function

$$f(x) = \frac{dF}{dx}$$

such that

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Random Vectors

Take $\Omega = \mathbb{R}^k$, the Cartesian product of k copies of \mathbb{R} .

- The probability space for a random vector $\mathbf{X} = (X_1, \dots, X_k)$ is defined by

$$(\mathbb{R}^k, \mathcal{B}^k, \mu).$$

- \mathcal{B}^k is the σ -field generated by the "measurable rectangles" of \mathbb{R}^k
 - That is, the sets constructed as the Cartesian products of k sets from \mathcal{B} .

- μ is a probability measure defined by the multivariate c.d.f.

$$F(x_1, \dots, x_k) = \mu(X_1 \leq x_1, \dots, X_k \leq x_k)$$

- A random vector is more than just a collection of random variables.
 - The *joint* distribution of the coordinates is what matters.

Time Series

Definition:

A time series is a time-ordered sequence of observations indexed by dates (integers)

$$x_t, \quad t = 1, \dots, n.$$

We aim to make inferences about the way in which the series evolves, and so forecast its future path.

- The theoretical framework is probability theory - we treat the (whole) sequence as a random element of a probability space.
- We also treat it as a finite segment of a potentially infinite sequence, $\{x_t, t = \dots, -1, 0, 1, 2, \dots, n, \dots\}$
 - The probability space of interest to us is therefore the set of all possible infinite sequences (points of \mathbb{R}^∞).
 - Write $\{x_t\}$ to denote the sequence with coordinates x_t .
- Let \mathcal{C} denote the collection of *finite dimensional cylinder sets* of \mathbb{R}^∞ - the sets of which at most a finite number of co-ordinates are restricted to sets of \mathcal{B} .
 - In effect, the elements of sets of \mathcal{C} have a representation as random vectors – points in \mathbb{R}^k , where k is the highest restricted co-ordinate.

- The Borel field \mathcal{B}^∞ is defined as the smallest σ -field containing \mathcal{C} .
- It remains to show that distributions of the form $(\mathbb{R}^\infty, \mathcal{B}^\infty, \mu)$ exist – a non-trivial question!.

Kolmogorov Consistency Theorem.

- The "fidis" (finite dimensional distributions) of a random sequence are the joint distributions of finite collections of sequence coordinates.
 - The fidis are just the distributions of random k -vectors, in effect, whose properties are known.
- The Consistency Theorem states that given a family of fidis μ_k , specified for every finite k , then subject to the *consistency condition*

$$\mu_k(E) = \mu_m(E \times \mathbb{R}^{m-k}) \text{ for } E \in \mathcal{B}^k \text{ and } m > k > 0$$

there exists a unique, probability measure μ , such that the infinite random sequence is distributed according to $(\mathbb{R}^\infty, \mathcal{B}^\infty, \mu)$, where the μ_k are the fidis of the sequence.

- This fundamental result underpins all statistical work in time series, by assuring us that the concept of an infinite-dimensional distribution (essential for asymptotic arguments) is well-founded.

Terminology

- A stochastic sequence is one type of *stochastic process* - the general term encompasses processes in continuous time, etc. etc.
- An actual (observed) time series is called a *realization* of the stochastic process - just one of the many "histories" that chance could have thrown up.

The Fundamental Problem...

If statistical averaging operations are applied to a realization (calculation of sample means, variances, correlations etc.), can we use the standard statistical theory to analyse and interpret these quantities?

- Remember - the standard theory invokes *the random sampling* assumption. It views a sample as a collection of independent random variables.
- Realizations of time series are not independent samples, in general.
- They exhibit local or global trends, cycles, etc. We can often use their recent histories to forecast future out-turns. Not features of a random sample!

Since all time series analysis depends on such averaging, we have to know when the procedure is legitimate:

Will it yield useful estimates of the underlying parameters of the distribution?

Some important concepts...

1. Stationarity

Definition: Consider a finite segment of a time series, x_t, \dots, x_{t+m} . (arbitrary t and m). If the *joint distribution* of these random variables is the same as that of $x_{t+k}, \dots, x_{t+k+m}$ for any k , then we say the time series is stationary.

Without stationarity, we cannot be sure that averaging two different segments of a realization will yield the same result.

Example: Consider $E(x_t)$.

- Expected value of the series coordinate at time t . An attribute of the distribution of the sequence *as a whole*.
- If we sample many realizations of the process, and look at the values at time t , $E(x_t)$ is the typical value (central tendency) of x_t *across realizations*. May depend on t .
- If the process is stationary, then $E(x_t)$ is the same for every t . We can write $E(x_t) = E(x_1)$, all t .

2. Ergodicity

Definition

Let $\{x_t\}$ be a stationary sequence. The sequence is *ergodic* if for all *time-invariant* random events $E \in \mathcal{F}$, either $P(E) = 0$ or $P(E) = 1$

- Time-invariance means that replacing t by $t + 1$ in the definition of E yields an event E' that is identical to E with probability 1. In other words, $P(E\Delta E') = 0$.
 - Note, an invariant event must involve *all* sequence coordinates!
- If a sequence is ergodic, it can be shown that for any pair of measurable sets of real numbers, $A \in \mathcal{B}$ and $B \in \mathcal{B}$,

$$\frac{1}{n} \sum_{k=1}^n P(x_1 \in A \text{ and } x_k \in B) \rightarrow P(x_1 \in A)P(x_1 \in B)$$

as $n \rightarrow \infty$.

- If $\text{Var}(x_t)$ exists, ergodicity can further be shown to imply

$$\frac{1}{n} \sum_{k=1}^n \text{Cov}(x_1, x_k) \rightarrow 0.$$

Ergodic Theorem: (a strong law of large numbers)

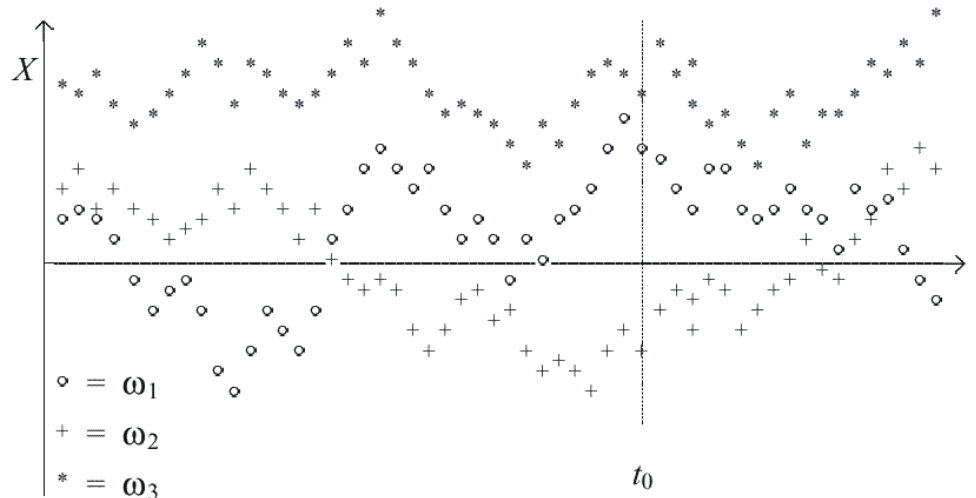
If $\{x_t, t = 1, 2, 3, \dots\}$ is stationary and ergodic then

$$\frac{1}{n} \sum_{t=1}^n x_t \rightarrow E(x_1) \text{ with probability 1.}$$

That is, the average of any realization converges to the same constant as $n \rightarrow \infty$ – except possibly for cases arising with zero probability.

A counter-example:

If $\{y_t, t = 1, 2, 3, \dots\}$ is stationary and ergodic, and Z is a random variable (not fixed across realizations) and $x_t = y_t + Z$, then the sequence $\{x_t, t = 1, 2, 3, \dots\}$ is *not* ergodic.



- In the sketch, suppose $Z(\omega_1) = Z(\omega_2) = 0$ but $Z(\omega_3) = 1$.
- If (say) $E(x_t) = 0$ and Z is Bernoulli-distributed with $P(Z = 1) = p$ and $P(Z = 0) = 1 - p$ then
 - The *ensemble mean* is $E(x_t) = p$.
 - $\frac{1}{n} \sum_{t=1}^n x_t(\omega_j) \rightarrow_{as} \mu_j$ where $\mu_1 = \mu_2 = 0$ but $\mu_3 = 1$.

3. Independence

For any finite collection of sequence coordinates, $x_{t_1}, x_{t_2}, \dots, x_{t_m}$, we can consider their joint distribution, and also their marginal distributions (treating each as a random variable in the usual way).

The question is this: although part of the same realization, do they behave like a random sample?

Definition: If any such collection is *totally independent*, such that for any measurable sets $A_1 \in \mathcal{B}$, $A_2 \in \mathcal{B}$, ..., $A_m \in \mathcal{B}$,

$$P(x_{t_1} \in A_1, x_{t_2} \in A_2, \dots, x_{t_m} \in A_m) = \\ P(x_{t_1} \in A_1)P(x_{t_2} \in A_2) \cdots P(x_{t_m} \in A_m)$$

for all collections (t_1, \dots, t_m) , and all $m > 1$, then we say that the sequence is serially independent.

- If the coordinates are continuously distributed, such that the distribution of $x_{t_1}, x_{t_2}, \dots, x_{t_m}$ is described by a probability density function $f(x_{t_1}, x_{t_2}, \dots, x_{t_m})$, then independence implies

$$f(x_{t_1}, x_{t_2}, \dots, x_{t_m}) = f(x_{t_1})f(x_{t_2}) \cdots f(x_{t_m})$$

4. Independent and identically distributed (i.i.d.)

i.e., both independent and stationary.

- This is like a true random sample, drawn from a fixed distribution.
- Not often found in ‘nature’, but an important building block of many time series models.

5. Mixing

Time series that are not independent may still exhibit limited *memory*. That is, if coordinates t_1, t_2, \dots, t_m are widely separated in time, then $x_{t_1}, x_{t_2}, \dots, x_{t_m}$ are "nearly" independent.

Definition (slightly simplified):

Let $\{x_t\}$ be a stationary process. The process is said to be mixing if any pair of measurable sets of real numbers, $A \in \mathcal{B}$ and $B \in \mathcal{B}$, satisfy

$$\left| P(\{x_t \in A\} \text{ and } \{x_{t+m} \in B\}) - P(x_t \in A)P(x_{t+m} \in B) \right| \rightarrow 0$$

as $m \rightarrow \infty$.

Points:

- Compare definition of ergodicity. A stationary mixing process is ergodic.
- By stationarity, the indicated probabilities are the same for any t . However...
- The mixing concept can be generalized to the nonstationary case. There are several different ways to define mixing in this case e.g.
 - strong mixing (α -mixing)
 - uniform mixing (ϕ -mixing)
(n.b. uniform mixing is a stronger concept than strong mixing!)
- May need to specify the rate of convergence to zero (mixing size).

Autocorrelation

A natural measure of memory in a stationary time series is the sequence of *autocovariances*:

$$\gamma_j = \text{Cov}(x_t, x_{t+j}), \quad j = 0, 1, 2, 3, \dots$$

The *correlogram* is the name given the normalized sequence

$$\rho_j = \frac{\gamma_j}{\gamma_0}, \quad j = 0, 1, 2, 3, \dots$$

– the *autocorrelations*.

Note that $|\rho_j| \leq 1$ for all j . This measure of memory is scale-independent.

- For a mixing process, $\gamma_j \rightarrow 0$.
- $\gamma_j \rightarrow 0$ does not imply mixing!

Definition:

- ■ A process is called *white noise* if

$$E(x_t) = 0, \quad \text{Cov}(x_t, x_{t-j}) = \begin{cases} \sigma^2, & j = 0 \\ 0 & j \neq 0 \end{cases}$$

- Note: this is an engineering term, describing the spectrum of the process (see later).

Technical Note:

Distinction between independence and uncorrelatedness:

Consider random variables x and y .

1. If $E(xy) = E(x)E(y)$ then x and y are uncorrelated.
2. If $E(\phi(x)\psi(y)) = E(\phi(x))E(\psi(y))$ for *all* measurable, integrable functions $\phi(\cdot)$ and $\psi(\cdot)$, then x and y are independent.

Hence, independence implies uncorrelatedness, but not the other way round.

- If (and only if) x and y are *jointly normally distributed* (Gaussian) the covariance completely represents their dependence. In this case only, the two concepts are equivalent.

Wide-Sense Stationarity

Definition:

If $E(x_t)$ exists and is the same for every t , $\text{Cov}(x_t, x_{t+j})$ exists and depends only on j , the time series $\{x_t\}$ is said to be wide-sense stationary (or covariance stationary, or weakly stationary).

Points:

- This definition is conventional, but a bit tricky. A series can be strictly stationary (as defined above) and yet not wide-sense stationary, because $\text{Var}(x_t)$ is undefined. This does not mean that its distribution depends on time.
- It is convenient, however, because a range of results can be stated for wide-sense stationary processes, although they needn't be stationary in the strict sense.
- A white-noise process is wide-sense stationary by definition. However, it need not be strictly stationary to satisfy the definition.

Weak and Strong Dependence

Definition:

A wide-sense stationary process is said to have *short memory* (or, is *weakly dependent*) if

$$\sum_{j=0}^{\infty} |\gamma_j| < \infty.$$

If this sum diverges, it has *long memory* (strongly dependent)

- Obviously, this definition is relevant to those cases where the autocorrelations contain the important information about dependence.
- If $\gamma_j = O(j^{-1-\varepsilon})$ for $\varepsilon > 0$, the summability condition holds, but note that

$$\sum_{j=1}^m j^{-1} = O(\log m).$$

- To explain why this “break-even” point is the interesting one will require a bit more theory. We return to this later on.

Note: “ $O(\cdot)$ ” = order of magnitude. We say that $\gamma_j = O(j^a)$ if

$$\frac{|\gamma_j|}{j^a} \leq B < \infty$$

for some positive constant B and all values of j .

Wold's Decomposition

Let $\{x_t, -\infty < t < \infty\}$ be a wide-sense stationary process with $E(x_t) = 0$. Then, there exists a sequence of constants $\theta_j, j = 0, 1, 2, 3, \dots$ with $\theta_0 = 1$ and $\sum_{j=0}^{\infty} \theta_j^2 < \infty$, and a white noise process $\{\varepsilon_t\}$, such that

$$x_t = \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j} + v_t$$

where $E(v_t \varepsilon_{t-j}) = 0$ for all j , and there exist constants $\alpha_0, \alpha_1, \alpha_2, \dots$ such that $\text{Var}\left(\sum_{j=0}^{\infty} \alpha_j v_{t-j}\right) = 0$.

- The condition on the fixed weights θ_j is called *square-summability*.
- Note that by the definition, v_t is *perfectly* predictable one step ahead, since

$$v_{t+1} = -\frac{1}{\alpha_0} \sum_{j=1}^{\infty} \alpha_j v_{t-j+1} \text{ with probability 1.}$$

v_t is called the *deterministic component* of the process.

- The representation of the non-deterministic component is called a *moving average* of infinite order (MA(∞)).
- Wold's decomposition is often cited as a justification for using *linear models* to represent the non-deterministic component of a time series.

Linear Processes

Definition:

A linear process is one having the representation

$$x_t = \sum_{j=0}^{\infty} \theta_j u_{t-j}$$

where the process $\{u_t\}$ is i.i.d.

The special feature of a linear process is that its dependence structure is entirely determined by the sequence of weights θ_j .

- However, i.i.d. is a stronger condition than white noise, even if the process is strictly stationary – as noted before, independence implies uncorrelatedness, but uncorrelatedness does not imply independence.
- The exception to this rule is the normal (Gaussian) process, that is, any finite collection of coordinates is jointly normally distributed. In this case, the dependence is entirely represented by the pairwise autocorrelations.
- Hence: the Wold decomposition is a much stronger result for Gaussian processes than more generally.

Spectral Analysis

Think of the *spectrum* (or *spectral density*) of a time series as the Fourier transform of the correlogram.

$$f(\omega) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma_j e^{-ij\omega} \quad -\pi \leq \omega \leq \pi$$

Since

- $\gamma_{-j} = \gamma_j$ (stationarity)
- $e^{ix} = \cos x + i \sin x$ (Euler's formula)
- $\cos(-x) = \cos x$, $\sin(-x) = -\sin x$

an equivalent form is

$$f(\omega) = \frac{\gamma_0}{2\pi} + \frac{1}{\pi} \sum_{j=1}^{\infty} \gamma_j \cos(\omega j)$$

Note, $f(\omega) = f(-\omega)$.

f represents the dependence in the *frequency domain*. We can speak of dependence at high and low frequencies.

- Estimating the spectrum is an important issue in some applications, but we will not deal with it in this course.

- In a white noise process, $f(\omega) = \frac{\gamma_0}{2\pi}$ at all ω – amplitude is the same at all frequencies.. This is a ‘flat spectrum’.

- In a weakly dependent process,

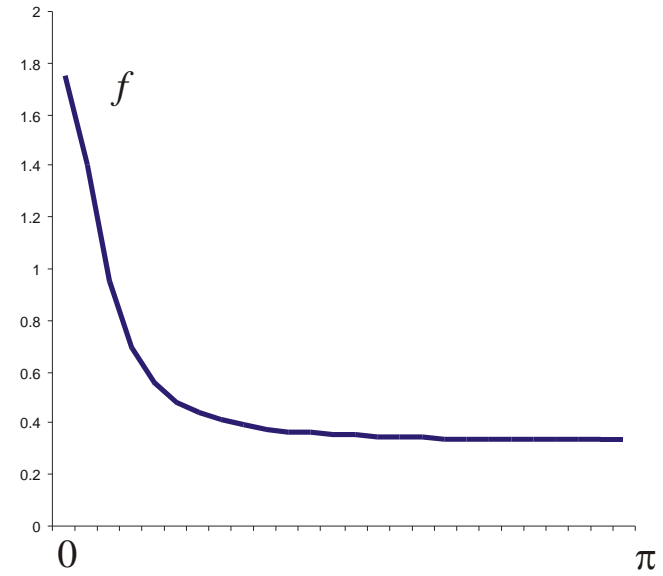
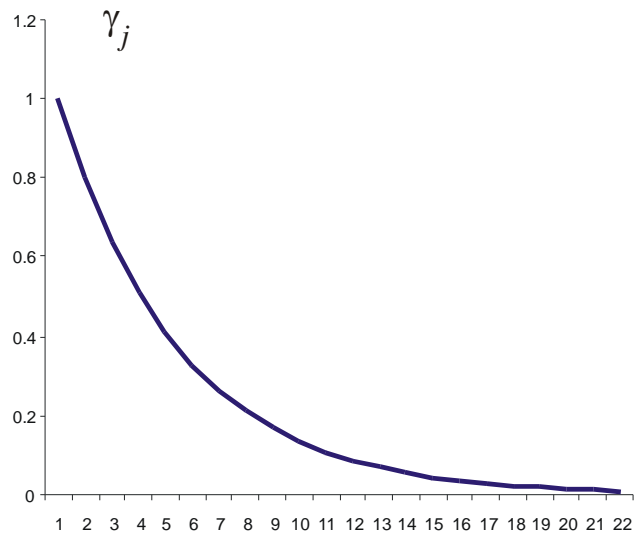
$$f(0) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma_j < \infty$$

We say the spectrum is ‘bounded at the origin’.

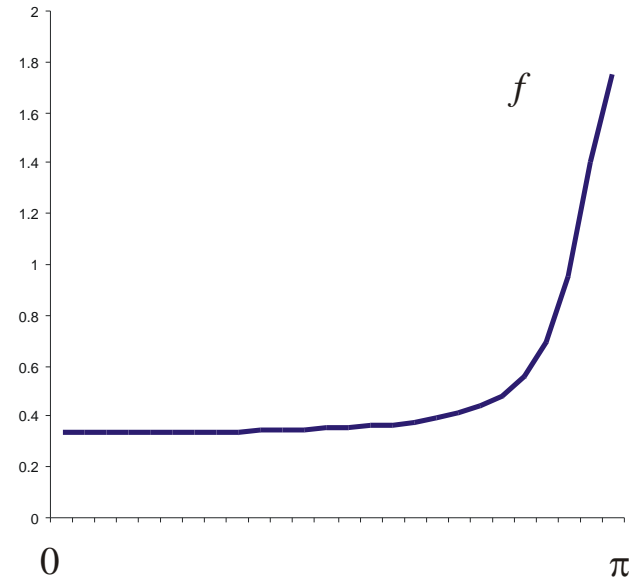
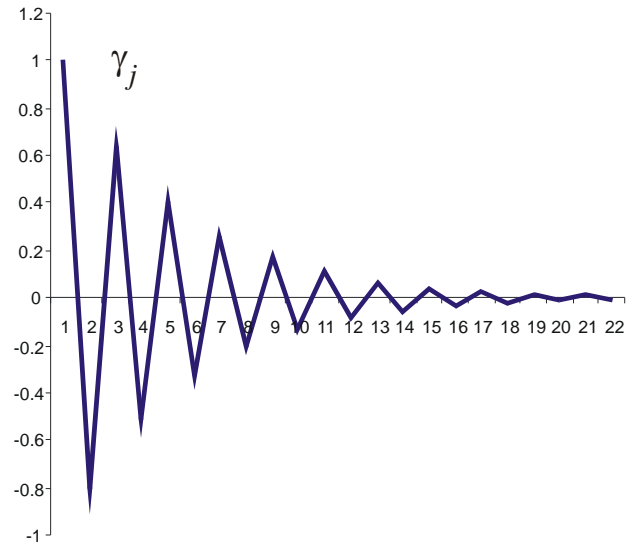
- Economic time series typically have largest spectral density at low frequencies (long cycles).

The following examples show contrasting cases ...

1. $\gamma_j = 0.8^j$ - positive autocorrelation, spectral density concentrated at low frequencies.

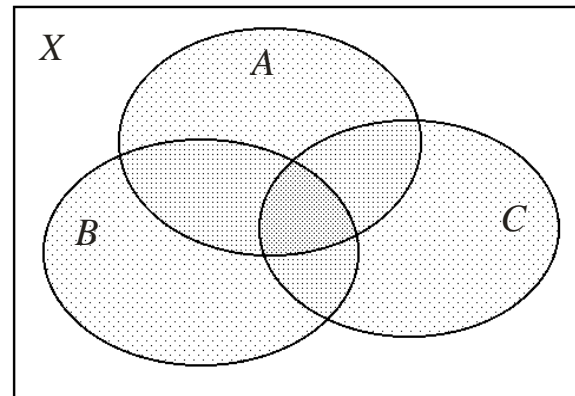


2. $\gamma_j = (-0.8)^j$ - negative autocorrelation, spectral density concentrated at high frequencies.



Basic Facts about Sets

Venn Diagram:



- *Universe of Discourse: X*
- *Set: $A \subseteq X$ (subset of X)*
- *Element: $x \in X$.*
- *Union: $A \cup B = \{x \in X : x \in A \text{ or } x \in B\}$*
- *Intersection: $A \cap B = \{x \in X : x \in A \text{ and } x \in B\}$*
- *Complement: $A^c = \{x \in X : x \notin A\}$*
- *Difference: $A - B = \{x \in X : x \in A \text{ and } x \notin B\} = A \cap B^c$ – thus, $A^c = X - A$.*
- *Symmetric difference: $A \Delta B = (A - B) \cup (B - A)$*
- *Cartesian Product: $A \times B = \{(x, y) : x \in A, y \in B\}$ - a set of pairs.(e.g. points in a plane).*

Set Algebra:

Unions and intersections obey associative, commutative and distributive laws.

De Morgan's Laws:

$$(A \cup B)^c = A^c \cap B^c$$

$$(A \cap B)^c = A^c \cup B^c$$

And in general,

$$\left(\bigcup_{i=1}^{\infty} A_i \right)^c = \bigcap_{i=1}^{\infty} A_i^c$$

$$\left(\bigcap_{i=1}^{\infty} A_i \right)^c = \bigcup_{i=1}^{\infty} A_i^c.$$

- Note, all set operations reduce to unions and complements.