3F3 Statistical Signal Processing: First Draft

'Need to Know...'

Mrinank Sharma

January 6, 2018

This is an attempt to condense everything which **must be memorised** for this course and hence does not include data-book content. However, more involved proofs are included here.

1. We call (Ω, P) the *probability space* where Ω is the sample space (the set of all possible outcomes) and P is a mapping from events to a number in the interval $[0, 1]$.

   Probability Space

2. A probability P assigns each event $E, E \subset \Omega$, a number in $[0, 1]$. P must satisfy the following axioms.

   Probability Axioms

   - $P(\Omega) = 1$.
   - For disjoint $A_1, A_2, ...$ we can state $P(\bigcup\limits_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

   All the rules of probability follow from the above axioms. It is often useful to draw Venn diagrams.

3. Define the indicator function for an event E.

   Indicator Function

$$\mathbb{I}_E(t) = \begin{cases} 1 & t \in E \\ 0 & t \notin E \end{cases}$$

4. Conditional probabilities are themselves proper probability distributions and follow the axioms of probability (by considering $P(\Omega|G)$ and $P(A \cup B|G)$ for disjoint A, B). The above definition gives rise to the probability chain rule.

   Conditional Probability

5. For independent events,

   Independent Events

$$P(A, B) = P(A \cap B) = P(A)P(B) \Leftrightarrow P(A|B) = P(A)$$

6. Given a probability space, (Ω, P), a random variable is a function $X : \Omega \rightarrow \mathbb{R}$. There is a sample space behind every random variable and the probabilities depend both on the probability space and the mapping.

   Random Variable Definition

   A random variable is discrete if its range is a finite or countable set. A set is countable if a one-to-one mapping from the set to the integers can be defined. For these variables a **probability mass function** is used.

   Discrete RV

   Continuous random variables have probability density functions, not mass functions. For these variables there must exist a non-negative function, $f_X(x)$ such that $\int_{-\infty}^{\infty} f_X(x)\ dx = 1$ and for $a < b$, $Pr(a \leq X \leq b) = \int_a^b f_X(x)\ dx$. A continuous RV has no concentration of probability at particular points; the probability of a single value is meaningless.

   Continuous RV

7. The CDF is defined as

   Cumulative Density Function

$$F_X(x) = Pr(X \leq x)$$

   From the axioms of probability, the following properties hold.

   - $0 \leq F_X(x) \leq 1$.
   - $F_X(x)$ is non-decreasing as x increases.
   - $Pr(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1)$.

- $\lim_{x\to-\infty} F_X(x) = 0$ & $\lim_{x\to\infty} F_X(x) = 1$
- $F_X(x)$ is right-continuous.

For a continuous RV,

$$F_x(x) = \int_{-\infty}^{x} f(t)\ dt \quad f_X(x) = \frac{dF_X(x)}{dx}$$

The CDF can be used to transform random variables. Let X be a random variable and let $Y = r(X)$.

CDF Method for Transforming RV

(a) Find the set $A_y = \{x : r(x) \leq y\}$.

(b) Find the CDF of Y.

$$\begin{aligned} F_Y(y) &= Pr(r(X) \leq y) = Pr(X \in A_y) \\ &= \int_{-\infty}^{\infty} \mathbb{I}_{A_y} f_X(x)\ dx \end{aligned}$$

(c) Differentiate to form the probability density function.

8. The probability density function of the sum of two independent random variables is the convolution between their probability density functions. $f_Y(y) = \int_{-\infty}^{\infty} f_1(x_1)f_2(y - x_1)\ dx_1$

Sum of independent RV

9. For random variables X and Y,

Rule of Iterated Expectation

$$\mathbb{E}_{X,Y}[r(X,Y)] = \mathbb{E}_Y[\mathbb{E}_X[r(X,Y)|Y]]$$

10. For independent $X_1, X_2, ..., X_n$

Expection of the Product of Independent RVs

$$\mathbb{E}[\Pi_{i=1}^{n} X_i] = \Pi_{i=1}^{n}\mathbb{E}[X_i]$$

This result can be shown by expanding the expectation formula and factorising the joint pdf.

11. For **any** $X_1, X_2, ..., X_n$

Sum of RVs

$$\mathbb{E}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n}\mathbb{E}[X_i]$$

To prove this, write the expectation as an integral.

12. For one-to-one or many-to-one mappings, the *Jacobian* can be used to calculate transformed probability densities. For a transformation with **random vectors** $\mathbf{x}, \mathbf{y}$ and $\mathbf{y} = G(\mathbf{x})$ The Jacobian can be expressed as follows:

Jacobian

$$f_Y(\mathbf{y}) = \sum_i f_x(G_i^{-1}(\mathbf{y}))||\mathbf{J}|| \quad |\mathbf{J}| = \frac{\partial(x_1, ..., x_n)}{\partial(y_1, ..., y_n)}$$

where $||\mathbf{J}||$ is the modulus of the Jacobian matrix and serves to normalise the density. The form of the Jacobian can be recalled (in a hand-wavey manner) by noting that $f_y dy = f_x dx$ and extending this to vectors.

13. Define the characteristic function of a random vector, $\mathbf{x}$ as follows

Characteristic Function

$$\varphi(\mathbf{t}) = \mathbb{E}[exp(j\mathbf{t}^T\mathbf{x})], \quad \mathbf{t} \in \mathbb{R}^n$$

The characteristic function fully defined the probability distribution (there is a one-to-one mapping, e.g. Fourier transforms). The characteristic function is useful for calculating $\mathbb{E}[X^N]$ (for a random variable, consider differentiating the characteristic function and substituting $t = 0$).

For the sum of independent random variables, the overall characteristic function is the product of each characteristic function (by expectation of the product of independent random variables).

14. Let the process $\{X_n\}_{n\geq 0}$ be a collection of discrete random variables taking values in $S = \{1, ..., L\}$. *Markov Chains*

    - The pmf of X_0 is $\boldsymbol{\lambda} = (\lambda_i : i \in S)$ i.e. $p_{X_0}(X_0 = i) = \lambda_i$ where $\lambda_i \geq 0$ & $\sum_i \lambda_i = 1$.
    - Let $\mathbf{P} = (P_{i,j} : i, j \in S)$ be a non-negative matrix with each row being a pmf on S (i.e. each row normalises and is non-negative). For $n > 0$, $p(i_n|i_0, ..., i_{n-1}) = \mathbf{P}_{i_{n-1},i_n}$. This is an example of **limited memory** and is known as the Markov property.

    Defining the pair $(\boldsymbol{\lambda}, \mathbf{P})$ defines a Markov chain where $\boldsymbol{\lambda}$ is the initial distribution of the chain and $\mathbf{P}$ is the transition probability matrix.

    The marginals of a Markov chain are easy to compute; after starting at the initial distribution, multiplying by the transition matrix gives the probability distribution for the next index. The following result can be seen by induction.

$$p_{X_n}(i_n) = (\boldsymbol{\lambda}\mathbf{P}^n)_{i_n}$$

    For an *irreducible* Markov chain (all states communicate), the sample average converges to $\sum_{i\in S} \pi_i r(i)$ for any function r *Ergodic Theorem*

    Setting $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$ and solving i.e. finding the eigenvectors of $\mathbf{P}$ gives the **stationary distribution of a Markov chain**.

15. A discrete time process, $\{X_n\}_{n\geq 0}$ is strictly stationary iff *Strict Sense Stationarity*

$$Pr(X_0 \in A_0, ..., X_k \in A_k) = Pr(X_{0+m} \in A_0, ..., X_{k+m} \in A_k)$$

    i.e. any two sections of the process are statistically indistinguishable. If a Markov process is initialised with initial probability (row vector) $\boldsymbol{\pi}$ such that $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$ then it is strict sense stationarity.

16. Let $\{X_i\}_{i\geq 0}$ be a sequence of iid variables with $\mathbb{E}[X_i] = 0$. Then *Law of Large Numbers*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} X_i = 0$$

    The non-zero mean case is considered by considering the iid sequence $\{X_i - \mu\}_{i\geq 0}$

17. Let $\{W_i\}_{i\in\mathbb{Z}}$ be a sequence of random variables with mean 0 and variance σ^2 and $\mathbb{E}[W_iW_j] = 0, i \neq j$. The auto-regressive process of order p (AR(p)) $\{X_i\}_{i\in\mathbb{Z}}$ is *Auto-Regressive Process*

$$X_i = \sum_{k=1}^{p} a_k X_{n-k} + W_i$$

    where $a_1, ..., a_p$ are constants. The AR(1) process can be considered to be an IIR L.T.I. filter with $h(k) = a^k$

18. A discrete time process, $\{X_i\}_{i\in\mathbb{Z}}$ is wide sense stationary iff *Wide Sense Stationarity*

    (a) $\mathbb{E}[X_i] = \mu$ i.e. constant mean.

    (b) $\mathbb{E}[X_iX_j] = \mathbb{E}[X_{i+k}X_{j+k}]$ for any i, j, k. (Assuming zero mean; otherwise transform random variables to give zero mean)

    Define the auto-correlation functionıfor a wide sense stationary process as follows

$$r_{xx}[k] = \mathbb{E}[X_nX_{n+k}],\ k \in \mathbb{Z}$$

    This is an even function!

19. Let $\{W_i\}_{i\in\mathbb{Z}}$ be a sequence of random variables with mean 0 and variance σ^2 and $\mathbb{E}[W_iW_j] = 0, i \neq j$. The moving average process of order q (MA(q)) $\{X_i\}_{i\in\mathbb{Z}}$ *Moving Average Process*

$$X_i = \sum_{k=1}^{q} b_k W_{n-k} + W_i$$

    where $b_1, ..., b_q$ are constants.

20. Let $\{W_i\}_{i\in\mathbb{Z}}$ be a sequence of random variables with mean 0 and variance σ^2 and $\mathbb{E}[W_iW_j] = 0, i \neq j$. The ARMA$(p,q)$ process $\{X_i\}_{i\in\mathbb{Z}}$ satisfies: *ARMA Process*

$$X_i = \sum_{k=1}^{p} a_k X_{n-k} + \sum_{k=1}^{q} b_k W_{n-k} + W_i$$

i.e. a combination of the AR and MA processes. It can be interpreted as a *causal filter* applied to the input.

21. For a WSS random process, $\{X_n\}$, the power spectrum is defined as the DTFT of the discrete autocorrelation function i.e. *Power Spectra*

$$\mathcal{S}_X(e^{j\Omega}) = \sum_{m=-\infty}^{\infty} r_{XX}[m]e^{-jm\Omega}$$

$$r_{XX}[m] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{S}_X(e^{j\Omega})e^{jm\Omega}d\Omega$$

where the normalised frequency, $\Omega = \omega T$, is in radians per sample. If $\Omega = 2\pi$, this is the sampling frequency).

the power spectrum is a real, positive, even and periodic function of frequency. It can be interpreted as a density spectrum as the mean-squared signal value across different frequencies can be obtained through integration over these frequencies. *Properties*

22. White noise is a WSS process for which *White Noise*

$$c_{XX}[n] = \mathbb{E}[(X_n - \mu)(X_{n+m} - \mu)] = \sigma_X^2\delta[m]$$

If the noise is zero mean, then $c_{XX}[n] = r_{XX}[n]$ and σ_X^2 is the mean-squared value, sometimes referred to as the power. The **spectrum of a zero mean white noise process is flat** i.e. equal across all frequencies.

23. If the input, $\{X_i\}_{i\in\mathbb{Z}}$, to a discrete time LTI system is WSS then the output, $\{Y_i\}_{i\in\mathbb{Z}}$, is also WSS. *LTI Systems*

Proof: Noting that $Y_n = \sum_{k=-\infty}^{\infty} h_{n-k}X_k$ (i.e. using the convolution representation) consider the mean and auto-correlation of Y.

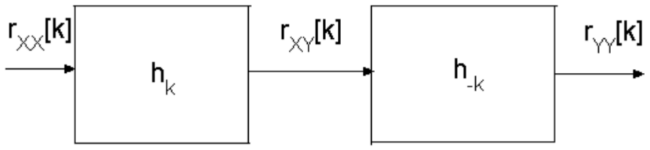

Figure 1: Correlation functions for a WSS input to a LTI system i.e. $r_{YY}[k] = r_{XX}[k] * h_k * h_{-k}$. Shown easily through the convolution representation.

Converting to the frequency spectrum,

$$\mathcal{S}_Y(e^{j\omega T}) = |\mathcal{H}(e^{j\omega T})|^2\mathcal{S}_X(e^{j\omega T})$$

Note that $\mathcal{H}(e^{j\omega T}) = \mathcal{Z}[h_k]|_{z=e^{j\omega T}}$; the Z transform and the DTFT are inherently linked.

24. Ergodic processes are WSS processes for which statistical characteristics can be obtained through time averages e.g. *Ergodicity*

- Mean Ergodic: $\mu = \mathbb{E}[X_n] = \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1} x_n$
- Correlation Ergodic: $r_{XX}[k] = \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1} x_n x_{n+k}$

and hence the above methods can be used for estimation for sufficiently large N. IT is hard to determine whether a process is ergodic; the following rules are used:

(a) Necessary and sufficient condition for mean ergodicity:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} c_{XX}[k] = 0$$

where $c_{XX}[k]$ is the auto-covariance function.

(b) A sufficient condition for mean ergodicity is that $c_{XX}[0] < \infty$ and $\lim_{N\to\infty} c_{XX}[n] = 0$.

Unless otherwise stated, assume that signals in this course are both wide-sense stationary and ergodic but this is often not true in practice!!

25. The Wiener filter is optimally adapted to the statistical characteristics of a random process. A desired signal is hidden is noise: *The Wiener Filter*

$$x_n = d_n + v_n$$

The Wiener filter is the optimal linear filter for estimation of d_n using x_n with some assumptions regarding the processes. It is assumed that $\{x_n\}$ **and** $\{d_n\}$ **are jointly WSS** i.e. all correlation functions are a function of only the time difference.

Filter derivation: *Derivation*

(a) Depending on the type of filter to be used i.e. order of filter, choose an impulse response for the filter h_p.

(b) Define $J = \mathbb{E}[\epsilon_n^2]$ where $e_n = d_n - \hat{d_n}$ where $\hat{d_n} = h_p * x_n$.

(c) Set $\frac{\partial J}{\partial h_q} = 0$ **simultaneously for all filter values**. This leads gives the **orthogonality principle**;

$$\mathbb{E}[\epsilon_n x_{n-q}] = 0 \ \forall \ q$$

(d) Substitute for ϵ_n.
If the filter is finite, this gives an infinite number of equations and is most easily solved in the frequency domain: *Infinite Solution*

$$\mathcal{H}(e^{j\Omega})\mathcal{S}_X(e^{j\Omega}) = \mathcal{S}_{XD}(e^{j\Omega})$$

Where $\mathcal{S}_{XD}(e^{j\Omega})$ is the cross-power spectrum, generally complex valued and measuring the coherence between two processes at a particular frequency. Note that:

$$\mathcal{S}_{XD}(e^{j\Omega}) = \mathcal{S}_{DX}^*(e^{j\Omega})$$

For an FIR filter, there are a finite number of simultaneous equations which can be solved (e.g. calculating the matrix inverse). *FIR Filter*

$$\mathbf{R_x h} = \mathbf{r_{xd}}$$

(e) Substitute back into J, using the orthogonality principle. It is easier to integrate the power spectrum of the error signal over frequencies.

For an infinite filter with uncorrelated noise, the equations form:

$$\mathcal{H}(e^{j\Omega}) = \frac{\mathcal{S}_D(e^{j\Omega})}{\mathcal{S}_D(e^{j\Omega}) + \mathcal{S}_V(e^{j\Omega})}$$

26. The correlation matrix is often formed when solving the Weiner-Hoff equations. *Correlation Matrix*

$$\mathbf{R}_X = \begin{bmatrix} r_{XX}[0] & r_{XX}[1] & \cdots & r_{XX}[N] \\ r_{XX}[1] & r_{XX}[0] & \cdots & r_{XX}[N-1] \\ \vdots & \vdots & \ddots & \vdots \\ r_{XX}[N] & r_{XX}[N-1] & \cdots & r_{XX}[0] \end{bmatrix}$$

Since $r_{XX}[k] = r_{XX}[-k]$, $\mathbf{R}_X$ is a symmetric matrix and also has constant diagonals. For a generated sequence, $r_{dd}[k]$, the autocorrelation matrix formed must be valid. A necessary condition is that $\mathbf{R}_X$ is non-negative definite i.e.

$$\mathbf{a^T R_X a} \geq 0 \; \forall \mathbf{a}$$

Proof: Consider $\mathbf{x_n} = [x_n x_{n-1} ... x_{n-N}]$ noting that $(\mathbf{a^T x_n})^2 = \mathbf{a^T (x_n x_n^T) a} \geq 0$. $\mathbb{E}[x_n x_n^T] = \mathbf{R_x}$ and take the expectation of the expanded formula where $\mathbf{a}$ is not random.

27. Whilst the Wiener filter extracts a random filter from noise, **matched filters are used to detect a known signal buried in noise**. *Matched Filters*

$$\mathbf{x = s + v}$$

For a FIR filter with length N, the output at $N-1$ can be obtained as a dot product where a tilde represents a time reversed vector. The output has a component due to the deterministic signal as well as the noise.

$$y_{N-1} = \mathbf{h^T \tilde{s}} + \mathbf{h^T \tilde{v}}$$

In order to give the best change of detecting the signal, the SNR is maximised at the output of the filter at this time.

$$\text{SNR} = \frac{|\mathbf{h^T \tilde{s}}|^2}{\mathbb{E}[|\mathbf{h^T \tilde{v}}|^2]} = \frac{\mathbf{h^T (\tilde{s}\tilde{s}^T) h}}{\mathbb{E}[|\mathbf{h^T \tilde{v}}|^2]}$$

The matrix $\mathbf{M} = \mathbf{\tilde{s}\tilde{s}^T}$ has eigenvector $\frac{\tilde{\mathbf{s}}}{|\tilde{\mathbf{s}}|}$ with eigenvalue $\mathbf{\tilde{s}^T\tilde{s}}$. Any other vector orthogonal to this eigenvector i.e. $\mathbf{\tilde{s}^T x} = 0$ will have eigenvalue zero (by substitution). Hence an orthonormal basis can be formed and using this property simplification made.

$$\begin{aligned} \mathbf{h^T (\tilde{s}\tilde{s}^T) h} &= \mathbf{h^T} \alpha \; (\mathbf{\tilde{s}^T \tilde{s}}) \frac{\tilde{\mathbf{s}}}{|\tilde{\mathbf{s}}|} \\ &= \alpha^2 (\mathbf{\tilde{s}^T \tilde{s}}) \end{aligned}$$

For a white, zero mean noise process with variance σ_v^2

$$\mathbb{E}[|\mathbf{h^T \tilde{v}}|^2] = \sigma_v^2 \mathbf{h^T h}$$

Noting that scaling h would not change the SNR, set $|\mathbf{h}| = 1$.

$$\text{SNR} = \frac{\alpha^2 (\tilde{s}\tilde{s}^T)}{\sigma_v^2}$$

Where α is the the projection of $\mathbf{h}$ on to $\tilde{\mathbf{s}}$. The maximum value of α is chosen (1). Since the deterministic correlation function is always at a maximum with zero lag, the signal output term for the optimal filter is always less than that computed for $n = N-1$ and hence a maximum is looked for.

Hence the optimal filter is the (normalised), time reversed version of the desired signal. Without using the filter, the maximum achievable SNR is $\frac{\text{max. signal val}}{\sigma_v^2}$ and hence the matched filter gives a significant benefit.

28. For a time windowed signal of length $2N+1$, $\mathbf{x_n^N} = \mathbf{w_n^N x_n}$ where *Einstein-Wiener-Khinchin Theorem*

$$\mathbf{w_n^N} = \begin{cases} 1 & -N \leq n \leq N \\ 0 & \text{otherwise} \end{cases}$$

The DTFT of the windowed signal is effectively the DFT evaluated at continuous frequencies with a origin shift which introduces a linear phase shift. Note that

$$\begin{aligned}
\text{DTFT}\{\mathbf{x_n^N} * \mathbf{x_{-n}^N}\} &= X^N(e^{j\Omega})X^N(e^{j\Omega})^* = |X^N(e^{j\Omega})|^2 \\
\{\mathbf{x_n^N} * \mathbf{x_{-n}^N}\}(m) &= \sum_n \mathbf{x_n w_n^N x_{n-m} w_{n-m}^N} \\
\therefore \text{DTFT}\{\mathbb{E}[\frac{1}{2N+1}\sum_n \mathbf{x_n w_n^N x_{n-m} w_{n-m}^N}]\} &= \mathbb{E}[\frac{1}{2N+1}|X^N(e^{j\Omega})|^2] \\
&= r_{XX[m]}\frac{1}{2N+1}\sum_n \mathbf{w_n^N w_{n-m}^N} \\
&= r_{XX}[m]t[m]
\end{aligned}$$

noting the linearity of the DTFT and expectation. $t[m]$ is the deterministic autocorrelation function of the window function, w_n. As n increases, $t[m]$ gives wider and flatter and hence $T(e^{j\Omega})$ tends towards a delta function.

$$\begin{aligned}
\text{DTFT}\{r_{XX}[m]t[m]\} = \mathcal{S}_X(e^{j\Omega}) * T(e^{j\Omega}) = \mathbb{E}[\frac{1}{2N+1}|X^N(e^{j\Omega})|^2] \\
\lim_{N\to\infty} \mathbb{E}[\frac{1}{2N+1}|X^N(e^{j\Omega})|^2] = \mathcal{S}_X(e^{j\Omega})
\end{aligned}$$

Hence the power spectrum can be understood as the expected value of (time-normalised) DTFT squared of the signal values. Hence, the a random DFT/DTFT should look similar to the underlying power spectrum of the process. *Interpretation*

29. In general, a set of unknown parameters are inferred from a a number of measurements i.e. where $P << N$ *Estimation Setup*

$$\mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \end{bmatrix} \qquad \boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_{P-1} \end{bmatrix}$$

We will assume that we are able to model the probability of receiving the given data, $\mathbf{x}$, given the parameters i.e. $p(\mathbf{x}|\boldsymbol{\theta})$ is known. Note that **estimation seeks to estimate a single value whilst the task of inference is to infer the entire probability distribution**.

30. There are multiple methods to estimating a parameter such as the mean of a population. In order to assess estimators, we must consider: *Assessing Estimator*

    (a) Estimator **bias**. If $\hat{\mu}$ is an estimator, consider whether

    $$\mathbb{E}[\hat{\mu}] = \mu = \mathbb{E}[X]$$

    If true, the estimator is said to be unbiased.

    (b) Estimator **variance**

    $$\text{Var}[\hat{\mu}] = \mathbb{E}[\hat{\mu}^2] - \mathbb{E}[\hat{\mu}]^2$$

    The variance measures how the likely spread of the estimator; if the variance is large, even if unbiased, a typical estimate may be far from the true value.

An unbiased estimator whose variance tends to 0 as $N \to \infty$ is said to be **consistent** and given enough data will get the 'correct' estimate. The best estimator for a specific problem has no bias and the minimum variance is the *MVU (minimum variance ubiased) estimator*.

31. In the linear model, it is assumed that the data is generated as some linear function of the *The Geneal Linear Model*

parameters with an added noise term i.e.

$$x_n = \mathbf{g_n^T}\boldsymbol{\theta} + e_n \qquad \text{or } \mathbf{x} = \mathbf{G}\boldsymbol{\theta} + \mathbf{e} \text{ where}$$

$$\mathbf{G} = \begin{bmatrix} \mathbf{g_0^T} \\ \mathbf{g_1^T} \\ \vdots \\ \mathbf{g_{N-1}^T} \end{bmatrix}$$

$\mathbf{G}$ can be chosen to give a wide of possible models and may be fixed after the data has been observed (but in this case, the data cannot be generated from the model). The order of the model is the number of parameters in the model.

Ordinary Least Squares Estimation

32. The OLS estimator finds a 'best fit' model which matches the data by minimising

$$J = \mathbf{e}^T\mathbf{e}$$

i.e. the least squared error. For the general linear model, a pseudo-inverse solution is reached

$$\boldsymbol{\theta}^{OLS} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{x}$$

which can be obtained by setting $\frac{\partial J}{\partial \boldsymbol{\theta}} = \mathbf{0}$ (note the vector $\mathbf{0}$) or by completing the square which shows that the OLS estimator is **globally** optimal.

Covariance of OLS

It can be shown that the OLS estimator is unbiased easily through substitution into the relevant equations. The variance is more difficult to consider. Define the OLS matrix term as $\mathbf{C} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T$ and examine the variance of **any other unbiased estimator**, $\hat{\boldsymbol{\theta}}$.

$$\hat{\boldsymbol{\theta}} = \mathbf{D}\mathbf{x} \qquad \mathbf{D} = \mathbf{C} + \boldsymbol{\Delta}$$

$\boldsymbol{\Delta}$ is some matrix perturbation away from the OLS solution.

$$\mathbb{E}[\mathbf{D}\mathbf{x}] = \mathbb{E}[(\mathbf{C} + \boldsymbol{\Delta})\mathbf{x}] = \boldsymbol{\theta} + \boldsymbol{\Delta}\mathbf{G}\boldsymbol{\theta} = \boldsymbol{\theta}$$
$$\therefore \boldsymbol{\Delta}\mathbf{G} = \mathbf{0} \text{ since true for all } \boldsymbol{\theta}$$

Define the covariance matrix for the estimator

$$\text{cov}(\hat{\boldsymbol{\theta}}) = \mathbb{E}[(\hat{\boldsymbol{\theta}} - \mathbb{E}[\hat{\boldsymbol{\theta}}])(\hat{\boldsymbol{\theta}} - \mathbb{E}[\hat{\boldsymbol{\theta}}])^T] = \mathbb{E}[\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^T] - \boldsymbol{\theta}\boldsymbol{\theta}^T$$

note that the (i,i)th element is the variance of $\hat{\theta}_i$ and the simplification is due to the estimators being unbiased. Assume that the error is **zero mean white noise with variance** σ_e^2. Using this, substituting in $\hat{\boldsymbol{\theta}}, \mathbf{D}$ and using $\boldsymbol{\Delta}\mathbf{G} = \mathbf{0}$

$$\text{cov}(\hat{\boldsymbol{\theta}}) = \mathbb{E}[\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^T] - \boldsymbol{\theta}\boldsymbol{\theta}^T = \sigma_e^2\mathbf{D}\mathbf{D}^T$$

Expanding $\mathbf{D}$ out and noting that $\boldsymbol{\Delta}\mathbf{C}^T = \mathbf{C}\boldsymbol{\Delta}^T = \mathbf{0}$ we form

$$\begin{aligned} \text{cov}(\hat{\boldsymbol{\theta}}) &= \sigma_e^2((\mathbf{G}^T\mathbf{G})^{-1} + \boldsymbol{\Delta}\boldsymbol{\Delta}^T) \\ &= \text{cov}(\boldsymbol{\theta}^{OLS}) + \sigma_e^2\boldsymbol{\Delta}\boldsymbol{\Delta}^T \end{aligned}$$

The diagonal elements of $\boldsymbol{\Delta}\boldsymbol{\Delta}$ are ≥ 0 with equality corresponding to the OLS solution. Therefore var$(\hat{\theta}_i) \geq$ var(θ_i^{OLS}) and hence the OLS estimator is the minimum variance unbiased estimator of $\boldsymbol{\theta}$ and is termed a **best linear unbiased estimator (BLUE)**. *Blue*

Likelihood Estimation

33. If he error sequence is assumed drawn from an iid source whose probability distribution is known, likelihood estimation may be used.

$$p(\mathbf{e}) = p_e(e_0)p_1(e_1)\dots p_e(e_{N-1})$$

If p_e is the zero-mean normal distribution, the model is equivalent to the Linear Gaussian model. **Choose the value of θ which maximises the likelihood of the observed data, x** i.e.

$$\boldsymbol{\theta}^{ML} = \arg\max_{\boldsymbol{\theta}}\{p(\mathbf{x}|\boldsymbol{\theta})\}$$

This can be achieved using standard differential calculus. It is often convenient to maximise the log-likelihood function since it is a monotonically increasing function. In the case of the Linear Gaussian model, the **the least squares solution is equivalent to the maximum likelihood solution**.

The noise variance can also be estimated by considering the log-likelihood function at the optimal parameter estimate. The noise level is the mean-squared error at the ML parameter solution.

34. If the parameters are treated as random vectors with pdfs assigned to the parameters which **represent some prior knowledge about the relative probability of different parameters before the data is observed**. If nothing is known *a priori* then the prior distributions should express no intial preferences. *Bayesian Method*

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})} \propto p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

$p(\boldsymbol{\theta}|\mathbf{x})$ is termed the posterior probability whilst $p(\boldsymbol{\theta})$ is termed the prior. The posterior probability is made up of the 'new information' from the data and the prior belief of the data which may even be highly subjective. The **prior should be chosen carefully**. Note that the marginal likelihood, $p(\mathbf{x})$ can be obtained through marginalisation.

Note that we are also implicitly conditioning on many pieces of additional prior information e.g. the form of the data generation process. To reflect this, the prior is occasionally written as $p(\boldsymbol{\theta}|\mathbf{x}, \mathcal{M})$ where $\mathcal{M}$ denotes all of the additional modelling and distribution assumptions made.

$$\boldsymbol{\theta}^{\mathsf{MAP}} = \arg\max_{\boldsymbol{\theta}}\{p(\boldsymbol{\theta}|\mathbf{x})\}$$

The ML estimate can be interpreted as the MAP estimate with a uniform prior. As the sample size tends towards infinity, the Bayesian solution tends towards the maximum likelihood solution.

Note that the Gaussian is said to be a 'conjugate' prior since it's form makes Bayesian calculations straightforward and available in closed form.

35. Extending the Bayesian approach, alternatively, a cost function could be minimised, $C(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ which expresses the cost of an incorrect estimation. This is usually a non-negative function with $C(\boldsymbol{\theta}, \boldsymbol{\theta}) = 0$. Then solve *MMSE*

$$\min_{\hat{\boldsymbol{\theta}}} \mathbb{E}[C(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})] \qquad \hat{\boldsymbol{\theta}}(\mathbf{x})$$

If only one parameter is being estimated and $C(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, differentiation leads to the following result:

$$\hat{\theta}^{\mathsf{MMSE}} = \mathbb{E}[\theta|\mathbf{x}]$$

For the linear gaussian model, the estimator coincides with the *maximum a posteriori* estimator but this is not always the case.