GWDA Lectures

Sanjeev Dhurandhar

August 11, 2014

Contents

1 Statistical Detection of GW Signals in Noise **3**
- 1.1 Introduction . . . 3
- 1.2 Random variables and their distributions . . . 4
  - 1.2.1 Distribution and density functions . . . 4
  - 1.2.2 Multivariate random variables . . . 5

2 The Matched Filter **10**
- 2.1 Introduction . . . 10
- 2.2 White, coloured and stationary noise . . . 11
- 2.3 The matched filter . . . 13

3 The statistics and geometry of detection **17**
- 3.1 Hypothesis testing . . . 17
- 3.2 The 1 dimensional case . . . 18
- 3.3 2 dimensional case . . . 19
- 3.4 Neyman-Pearson criterion . . . 20

4 Composite hypothesis and geometry **22**
- 4.1 Composite hypothesis and maximum likelihood detection . . . 22
- 4.2 Ambiguity function and the metric . . . 23

Topics

- Random variables, distributions of random variables, uniform distribution, univariate Gaussian or normal distribution, multivariate Gaussian distribution.
- Noise as a random variable, statistical characterisation of noise, autocorrelation function, wide sense stationary noise, power spectral density of noise, white and coloured noise, Gaussian noise, matched filter.
- Hypothesis testing, binary hypothesis, Neyman-Pearson criterion, likelihood ratio, false alarm and detection probabilities.
- Composite hypothesis, geometric formulation, ambiguity function, metric, template banks, computational costs.

Books

1. A. Papoulis: *Probability, Random Variables and Stochastic Processes*, Mc Graw-Hill book co.
2. C. W. Helstrom: *Statistical Theory of Signal Detection*, Pergamon Press.
3. L. A. Wainstein and V. D. Zubakov: *Extraction of Signals from Noise*, Prentice Hall.

Lecture 1

Statistical Detection of GW Signals in Noise

1.1 Introduction

Ground based as well as space based gravitational wave (GW) detectors will continuously produce data. Since gravity couples to matter weakly, the gravitational wave signals are expected to be weak and so will be buried in the noise. The aim of the data analyst is to extract the signal buried in the noise by detecting it and measuring its parameters. The signal that the detectors are expected to see is a metric perturbation. For simplicity if we take the background metric η_{ik} to be flat, the full metric g_{ik} can be written as:

$$g_{ik} = \eta_{ik} + h_{ik}\,, \tag{1.1}$$

where h_{ik} is a tiny perturbation in the metric describing the gravitational wave. This perturbation may be due to compact binary star coalescences, supernovae, asymmetric spinning neutron stars, stochastic GW background, etc. We denote a typical component of the perturbation by simply h. We disregard the indices which essentially provide directional information. The metric (and the perturbation) is a dimensionless quantity and typically $h \sim 10^{-23}$ or smaller. For a GW source, h can be estimated from the well-known Landau-Lifschitz quadrupole formula. The GW amplitude h is related to the second time derivative of the quadrupole moment (this quantity has the dimensions of energy) of the source:

$$h \sim \frac{4}{r}\,\frac{G}{c^4}\,E^{\rm kinetic}_{\rm nonspherical}, \tag{1.2}$$

where r is the distance to the source, G is the gravitational constant, c is the speed of light and $E^{\rm kinetic}_{\rm nonspherical}$ is the kinetic energy in the *nonspherical* motion of the source. If we consider $E^{\rm kinetic}_{\rm nonspherical}/c^2$ a fraction of a solar mass and the distance to the source ranging from galactic scale of tens of kpc to cosmological distances of Gpc, then h ranges from 10^{-17} to 10^{-23}. These numbers then set the scale for the sensitivities at which the GW detectors must operate. The GW is detected by measuring the change in armlength of a laser interferometric detector which is observed as a phase shift on a photodiode. If the change in the armlength L of the detector is δL, then,

$$\delta L \sim hL\,. \tag{1.3}$$

The noise on the other hand is of the same order or larger. Therefore one needs to extract the GW signal given by the function $h(t)$ where t denotes the time, out of the noise $n(t)$. The noise $n(t)$ is a stochastic process and is random. Any detection is a statistical problem. One can

never be sure of a 100 % detection. Because there is noise in the data it is possible that the noise appears like a signal and it is only the noise that one is looking at. One can only assign probabilities to the detection process and provide confidence levels for claiming a detection.

In these lectures I will start with random variables and their probability distributions, the most impostant of these being the Gaussian random variables which are most common because of the central limit theorem. Then I will talk about matched filters and their optimality. The Neyman-Pearson criterion is most appropriate in the context of GW detection. I will discuss this and the Neyman-Pearson lemma in the context of hypothesis testing. This will bring us to the likelihood ratio and maximum likelihood. Finally I will discuss how geometry can be used in statistical detection problems and also how it provides an elegant framework of manifolds and the metric.

1.2 Random variables and their distributions

Our approach here will be more intuitive than rigorous. We will explain concepts via examples rather than rigorous definitions. Consider a sample space denoted by Ω on which a probability measure is defined. A simple example is $\Omega = \{H, T\}$ where H and T are respectively heads and tails in a coin tossing experiment. The (measurable) subsets of Ω are called events. A random variable is a (measurable) function from a sample space Ω to $\mathcal{R}$ the set of real numbers. For example we may define $X(H) = 1$ and $X(T) = 0$, then X is an example of a random variable. The random variables we are interested in, are more complex than this simple example; it is the data generated by the detector which contains noise.

1.2.1 Distribution and density functions

Given the random variable X (we will often abridge random variable by r.v.), we denote the set $X^{-1}((-\infty, x])$ simply by $\{X \leq x\}$. Then the distribution function of X is defined as:

$$F_X(x) = P\{X \leq x\}\,, \tag{1.4}$$

where P denotes the probability of the event under consideration. We will require such functions when we talk about false alarm and detection probabilities later. F_X has the following properties:

1. F_X is a non-decereasing function of x i.e. whenever $x_1 < x_2$, we have $F_X(x_1) \leq F_X(x_2)$;
2. $F_X(-\infty) = 0, \quad F_X(\infty) = 1$;
3. F_X is right continuous.

Usually we deal with the derivative of F_X which is called the *probability density function* or in short just 'pdf'. We have:

$$p_X(x) = \frac{dF_X}{dx}\,, \tag{1.5}$$

where we have denoted the pdf by p_X. From the above properties of F_X, we must have $p_X(x) \geq 0$ and

$$\int_{-\infty}^{\infty} p_X(x)dx = 1\,. \tag{1.6}$$

When F_X is discontinuous at some point we obtain a delta function at that point for p_X.

Some simple examples of distributions are the uniform, Gaussian etc. The phase ϕ could be uniformly distributed between 0 and 2π. What this means is that $p(\phi)$ (we drop the subscript

on p) is a constant. This constant can be evaluated by noting that its integral over the range $[0, 2\pi]$ must be 1. Hence we must have $p(\phi) = 1/2\pi$. An important distribution is the Gaussian or the normal distribution - its pdf is given by:

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad -\infty < x < \infty \,. \tag{1.7}$$

It is defined over the real line. Its mean is zero and variance is unity. The mean is the first moment and its variance is the second moment. They are given by:

$$\begin{aligned} \mu &= \langle x \rangle = \int_{-\infty}^{\infty} x p(x) dx = 0 \\ \sigma^2 &= \langle x^2 \rangle - \mu^2 = \int_{-\infty}^{\infty} x^2 p(x) dx = 1 \,. \end{aligned} \tag{1.8}$$

Such a distribution is also described by the statement $X \sim N(0, 1)$, where N stands for normal; 0 is the mean and 1 the variance. We can give frequentist's interpretation to this pdf. Let a large number N of trials be taken of X, then we ask the question how many of these values lie between say 0.1 and 0.2. Then, on an average $N \times P\{0.1 \leq X \leq 0.2\}$ values of the trials will lie between 0.1 and 0.2. The probability $P\{0.1 \leq X \leq 0.2\}$ is the number,

$$P\{0.1 \leq X \leq 0.2\} = \int_{0.1}^{0.2} p_X(x) dx \,. \tag{1.9}$$

We can easily transform this distribution by translating and scaling it. Let,

$$y = \frac{x - \mu}{\sigma} \,, \tag{1.10}$$

then the r.v. Y has the pdf,

$$p_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} \, e^{-(y-\mu)^2/2\sigma^2} \,. \tag{1.11}$$

Note that the Jacobian factor of the transformation ($dy = dx/\sigma$) is included in the pdf - in this case $1/\sigma$. This distribution of the r.v. Y has mean μ and variance σ^2 and is Gaussian distributed. So we may write, $Y \sim N(\mu, \sigma^2)$. The square root of the variance is σ and is called the standard deviation of the r.v. Y.

1.2.2 Multivariate random variables

The motivation for studying multivariate random variables comes from the fact that noise is a stochastic process and we will be considering data trains containing samples which are all r.v.s. Normally, we have a data segment $[0, T]$ and it is sampled uniformly with N points i.e. $\Delta = T/N$ is the sampling interval between two consecutive samples and is constant. So we have samples of the data $x(t_k = k\Delta) = x_k, \quad k = 0, 1, 2, ...N-1$. We can look upon the samples x_k as components of a N-dimensional vector, say a column vector, which we denote by $\mathbf{x}$. Then each of the x_k are r.v.s and $\mathbf{x}$ is a random vector distributed according to a N dimensional multivariate distribution. If the stochastic process is Gaussian, then $\mathbf{x}$ is a multivariate Gaussian. Therefore we need to understand how the r.v. of one variable generalises to a random vector of N variables.

To gain insight into general multivariate distributions it is best to begin with a two dimensional distribution or a bivariate distribution. The bivariate distribution has the advantage that one can easily visualise it and at the same time it has interesting features which are missing in

the univariate (1 dimensional) distribution. For example two random variables can be correlated and this fact is encoded in their covariance. We will first analyse a 2 dimensional Gaussian distribution.

Let us take two r.v.s X and Y. Their joint pdf is $p(x,y)$ which is defined as follows: for any event $A \subset \mathcal{R}^2$,

$$\int_A p(x,y)\ dxdy = P\{(x,y) \in A\}\,. \tag{1.12}$$

The distribution function is defined by the generalisation:

$$F(x,y) = P\{X \leq x,\ Y \leq y\} = \int_{-\infty}^{x}\int_{-\infty}^{y} dxdy\ p(x,y)\,. \tag{1.13}$$

A simple example is the Gaussian distribution:

$$p(x,y) = \frac{1}{2\pi\sigma^2}\ e^{-\frac{1}{2\sigma^2}(x^2+y^2)}\,. \tag{1.14}$$

One observes that $p(x,y)$ can be factorised into a product of two univariate Gaussians $p_X(x) \times p_Y(y)$ each of which is $N(0,1)$. It immediately follows from this that the r.v.s X and Y are independent. One can similarly factorise the distribution function as well, $F(x,y) = F_X(x)F_Y(y)$ into distribution functions of individual r.v.s X and Y. The covariance between X and Y is zero. One can easily check this by computing:

$$\langle xy\rangle = \frac{1}{2\pi\sigma^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dxdy\ xy\ e^{-\frac{1}{2\sigma^2}(x^2+y^2)} = 0\,. \tag{1.15}$$

Because the means vanish, $\langle x\rangle = \langle y\rangle = 0$, the covariance $\langle xy\rangle - \langle x\rangle\langle y\rangle = 0$. Similarly one can show by computing the corresponding integrals, that the variances are $\langle x^2\rangle = \langle y^2\rangle = \sigma^2$.

A more interesting example is the distribution when the r.v.s X, Y are correlated. Then the Gaussian distribution is given by:

$$p(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\ e^{-Q(x,y)}\,, \tag{1.16}$$

where $Q(x,y)$ is the quadratic,

$$Q(x,y) = \frac{1}{2(1-\rho^2)}\left[\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - 2\frac{\rho xy}{\sigma_x\sigma_y}\right]\,. \tag{1.17}$$

Here σ_x, σ_y are respectively the standard deviations of X and Y and ρ is the correlation coefficient which satisfies $-1 < \rho < 1$. When $\rho > 0$, X, Y are positively correlated while if $\rho < 0$ they are negatively correlated or anticorrelated. We may calculate the first and second moments of this distribution directly by integration. The means as before are zero, $\langle x\rangle = \langle y\rangle = 0$. The second moments are $\langle x^2\rangle = \sigma_x^2,\ \langle y^2\rangle = \sigma_y^2,\ \langle xy\rangle = \rho\sigma_x\sigma_y$. Computing the second moments requires some involved integration. The second moments can be collected together in a covariance matrix which we denote by C. Then:

$$C \equiv \begin{bmatrix} \langle x^2\rangle & \langle xy\rangle \\ \langle xy\rangle & \langle y^2\rangle \end{bmatrix} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}\,. \tag{1.18}$$

If we write the random vector:

$$\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}\,, \tag{1.19}$$

then we can write in short $C = \langle \mathbf{x}\mathbf{x}^T \rangle$, where $\mathbf{x}^T$ denotes the transpose of $\mathbf{x}$ and is therefore a row vector. The quadratic in terms of C can be written in short as $Q(x,y) = \mathbf{x}^T C^{-1} \mathbf{x}$. Note that the inverse of the covariance matrix, namely, C^{-1} makes an appearance in the pdf. C^{-1} is called the *Fischer information matrix*. Explicitly,

$$C^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} \frac{1}{\sigma_x^2} & -\frac{\rho}{\sigma_x \sigma_y} \\ -\frac{\rho}{\sigma_x \sigma_y} & \frac{1}{\sigma_y^2} \end{bmatrix} . \tag{1.20}$$

Note that if $\rho \neq 0$ then the r.v.s X and Y are correlated. To gain further insight we can plot contours of $p(x,y) = \text{const.}$ and understand what correlation means. In order to do this it helps to rotate the axes by 45°. Just to keep the algebra simple we will set $\sigma_x = \sigma_y = 1$, then we have,

$$p(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-Q(x,y)} , \tag{1.21}$$

where Q simplifies to,

$$Q(x,y) = \frac{1}{2(1-\rho^2)} (x^2 + y^2 - 2\rho xy) . \tag{1.22}$$

We make the transformations:

$$u = \frac{1}{\sqrt{2}}(x-y), \qquad v = \frac{1}{\sqrt{2}}(x+y) , \tag{1.23}$$

where the corresponding r.v.s U, V are related to X, Y by identical expressions. With these transformations, we have,

$$Q(u,v) = \frac{1}{2}\left(\frac{u^2}{1-\rho} + \frac{v^2}{1+\rho}\right) \equiv \frac{1}{2}\left(\frac{u^2}{\sigma_u^2} + \frac{v^2}{\sigma_v^2}\right) . \tag{1.24}$$

where $\sigma_u^2 = 1-\rho$ and $\sigma_v^2 = 1+\rho$. The corresponding pdf in terms of u, v is given by:

$$p(u,v) = \frac{1}{2\pi\sigma_u\sigma_v} \; e^{-\frac{1}{2}\left(\frac{u^2}{\sigma_u^2} + \frac{v^2}{\sigma_v^2}\right)} , \tag{1.25}$$

which factorises into,

$$p(u,v) = \frac{1}{\sqrt{2\pi}\sigma_u} \; e^{-\frac{1}{2}\frac{u^2}{\sigma_u^2}} \times \frac{1}{\sqrt{2\pi}\sigma_v} \; e^{-\frac{1}{2}\frac{v^2}{\sigma_v^2}} \equiv p_U(u) p_V(v) . \tag{1.26}$$

Each of them are univariate Gaussians. This also shows that the r.v.s U and V are independent. We may draw the contours of equal probability density which are essentially given by $Q(u,v) = \text{const.}$ These are just ellipses whose axes are the (u,v) axes or in the (x,y) plane these are ellipses whose axes are rotated by 45° with respect to the (x,y) axes. We show them below in the figure. For $\rho > 0$ the ellipses have the v-axis as the major axis. In fact when $\rho \longrightarrow 1$, the ellipses collapse onto the v-axis and all the probability density is concentrated along the v-axis; since then $\sigma_u \longrightarrow 0$, the pdf $p_U(u) \longrightarrow \delta(u)$, the delta function. Since $x = y$ on the v-axis, if X has taken the value x_0, the probability is very high that Y also will take the same value x_0 or a value very close to it, since the pdf is concentrated around this line. The r.v.s X and Y are then positively correlated.

On the other hand when $\rho < 0$, the contours are ellipses with the u-axis as the major axis. If now X takes the value x_0, then it is most likely that Y will take values near $-x_0$, because now the pdf is concentrated around the u axis which is just $y = -x$. Then r.v.s X and Y are said to be anticorrelated. The figures for $\rho > 0$ and $\rho < 0$ are shown below.

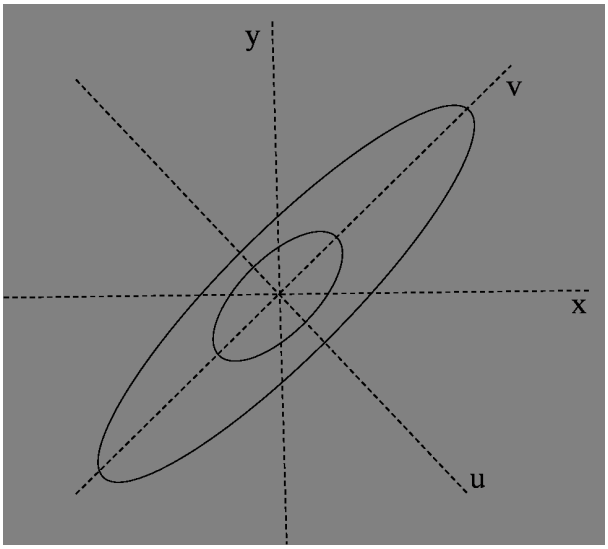

Figure 1.1: $\rho > 0$: The constant pdf contours are ellipses with the v-axis being the major axis. The contours are shown with unbroken curves while the axes are shown as dashed lines.

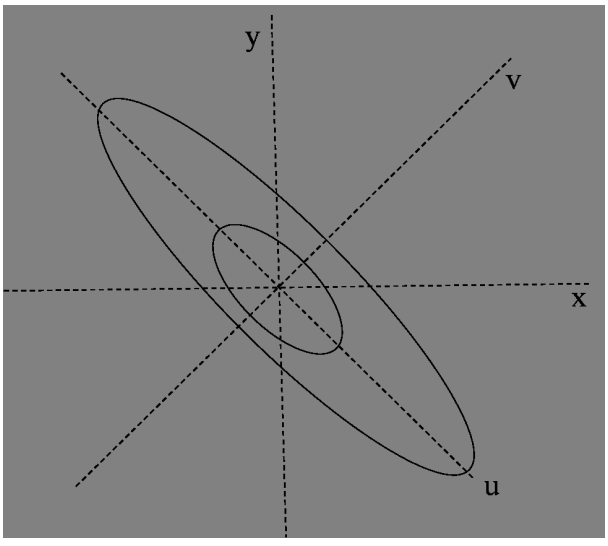

Figure 1.2: $\rho < 0$: The constant pdf contours are ellipses with the u-axis being the major axis. The contours are shown with unbroken curves while the axes are shown as dashed lines.

One can now generalise to a N dimensional random vector. Now the random vector $\mathbf{x}^T = (x_0, x_1, ..., x_{N-1})$ has N components and we have a $N \times N$ covariance matrix which we continue to denote by C. The zero mean multivariate Gaussian pdf is then given by:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{N/2}\sqrt{\det C}}\, e^{-\frac{1}{2}\mathbf{x}^T C^{-1}\mathbf{x}}\,, \tag{1.27}$$

where $C = \langle \mathbf{x}\mathbf{x}^T \rangle$. One can show this by direct integration i.e. $\langle x_i x_k \rangle = C_{ik}$ (one may employ certain tricks like differentiation with respect to a parameter to show this in an efficient and elegant way). If the noise vector $\mathbf{n}$ has such a pdf, then the noise is called Gaussian. Also in this case the noise has zero mean. We can always do this for detector noise by subtracting out the DC component and thus making the mean zero. Also if this is time series data, the nearby samples in time can be correlated. In the subsequent lectures we will discuss white and coloured noise where this issue is relevant.

Lecture 2

The Matched Filter

2.1 Introduction

When a known signal is embedded in the noise, matched filtering is the optimal method in extracting the signal from the noise. We will discuss here in what sense the matched filter is optimal. To give an example consider a sinusoidal signal say,

$$h(t) = A\cos 2\pi f_0 t\,, \tag{2.1}$$

where f_0 is a constant frequency and A the amplitude. The signal is embedded in the noise $n(t)$. We will consider here the time t as continuous and the signal and the noise as functions of the continuous variable t. When handling data one must of course sample the data and then one gets a discrete time series. We will switch from discrete to continuous time domains and vice-versa without any cause for confusion. We take the noise to be additive, then the data which we denote by $x(t)$ is given by,

$$x(t) = n(t) + h(t)\,. \tag{2.2}$$

The figure below depicts the situation:

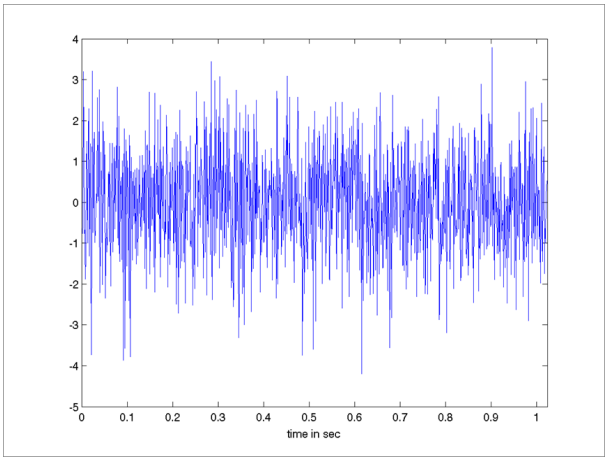

Figure 2.1: The time series of a sinusoidal signal embedded in the noise. The amplitude of the signal is small compared with the noise so that the signal is not visible.

We now apply the matched filter to this data. In this case the matched filter is also a sinusoid and is easily implemented by taking the Fourier transform of the data. We compute the statistic $C(f)$:

$$C(f) = |\tilde{x}(f)|, \qquad \tilde{x}(f) = \int_{-\infty}^{\infty} dt\ x(t)\ e^{-2\pi i f t}\,, \tag{2.3}$$

where $\tilde{x}(f)$ is the Fourier transform of $x(t)$ as defined above. The statistic $C(f)$ is just the modulus of the Fourier transform. This is shown in the figure below:

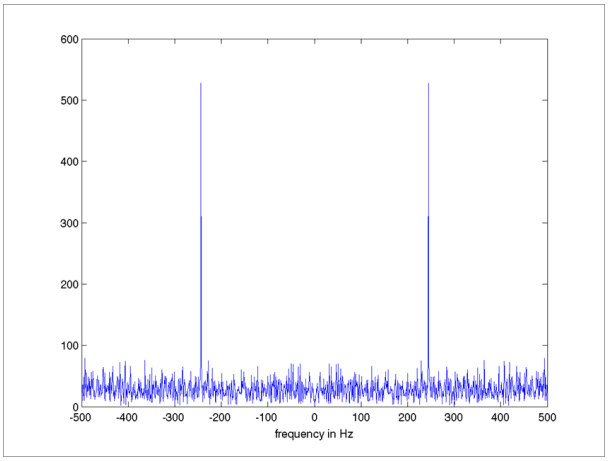

Figure 2.2: The statistic $C(f)$ is plotted versus the frequency f. Two peaks are seen at $\pm f_0$.

There are two peaks occuring at $\pm f_0$ because $\cos 2\pi f_0 t = (e^{-2\pi i f_0 t} + e^{2\pi i f_0 t})/2$. If we set a threshold sufficiently high so that it is unlikely that the noise crosses this threshold we would claim a detection of the signal. Apart from detection, we also recover parameters of the signal from the statistic such as the amplitude, the frequency and the phase.

This example shows how matched filtering works. We will now discuss this method in detail.

2.2 White, coloured and stationary noise

We will consider the noise to be a zero mean stochastic process - at each time t, $n(t)$ is a r.v. Also any DC which is there in the noise is removed from the data. At each time t we write,

$$\langle n(t) \rangle = 0\,, \tag{2.4}$$

where now the angular brackets denote *ensemble* average. One imagines a large number of identical detectors producing data. Then at a given time t we take the average of $n(t)$ over the ensemble of detectors. We characterise the noise by its autocorrelation function K defined as follows:

$$\langle n(t) n(t') \rangle = K(t, t')\,. \tag{2.5}$$

Clearly from the definition $K(t, t')$ is symmetric in its arguments, namely, $K(t, t') = K(t', t)$. We will normally put restrictions on this function and hence the noise. The noise is called stationary if for any N samples, the distribution function or the pdf is invariant under time translations. That is,

$$F(n(t_0), n(t_1), ..., n(t_{N-1})) = F(n(t_0 + \tau), n(t_1 + \tau), ..., n(t_{N-1} + \tau))\,, \tag{2.6}$$

for all time translations τ. We will put a weaker restriction on the noise requiring only the first two moments to be invariant under time translations. That is:

$$\langle n(t)\rangle = \langle n(t+\tau)\rangle, \quad \langle n(t)n(t')\rangle = \langle n(t+\tau)n(t'+\tau)\rangle\,, \tag{2.7}$$

for all time translations τ. Such a stochastic process is called wide sense stationary or in short WSS. We will assume that the noise we have is WSS. This is actually not the case, but if the signals are for short durations, then the noise for that duration can be considered to satisfy the WSS conditions.

When the noise is WSS, we have $K(t,t') = K(t+\tau, t'+\tau)$ for all τ. We may now choose $\tau = -t'$, then $K(t,t') = K(t-t',0)$ or K becomes a function of only one variable $t-t'$. Without any cause for confusion, we denote it by the same symbol $K(t-t')$. Physically, this means that the mean and the covariance between noise samples do not depend on absolute time but only on the time difference.

Because of the time translation symmetry, in Fourier space we have a simple representation of the autocorrelation function. In the Fourier space, we do the following calculation:

$$\begin{aligned}
\langle \tilde{n}(f)\tilde{n}^\star(f')\rangle &= \int\int dtdt'\ \langle n(t)n(t')\rangle\ e^{-2\pi ift+2\pi if't'} \\
&= \int\int dtdt'\ K(t-t')\ e^{-2\pi ift+2\pi if't'} \\
&= \int\int dtd\tau\ K(\tau)\ e^{-2\pi ift+2\pi if'(t+\tau)} \\
&= \int d\tau\ K(\tau)\ e^{2\pi if'\tau} \times \delta(f-f') \\
&= S(f)\ \delta(f-f')\,.
\end{aligned} \tag{2.8}$$

Here the star denotes the complex conjugate of the quantity. Note that $\tilde{n}(f)$ is a complex r.v. for each frequency f. We have also changed variables by defining $t' = t+\tau$ in the integrals and written $S(f)$ for the Fourier transform of K. Because of stationarity we get a delta function $\delta(f-f')$ in the frequency space. Since $K(\tau) = K(-\tau)$, that is, it is an even function, its Fourier transform $S(f)$ is real. Also $S(f) > 0$ or it is positive definite. It is called the power spectral density (PSD) of the noise. $S(f)$ is also an even function, namely, $S(-f) = S(f)$ and it can be folded on itself and defined only for $f \geq 0$. Then it is called the one sided PSD. We will not discuss this here. If the time is measured in units of seconds, $S(f)$ has dimensions of Hz^{-1}.

We can characterise the noise as white or coloured. If $S(f) = S_0$ a constant, then the noise is called white; otherwise it is coloured. In case of white noise, the autocorrelation function becomes,

$$K(\tau) = S_0 \int df\ e^{2\pi if\tau} \equiv S_0\delta(\tau)\,. \tag{2.9}$$

So then we have the relations:

$$\langle n(t)n(t')\rangle = S_0\delta(t-t'), \qquad \langle \tilde{n}(f)\tilde{n}^\star(f)\rangle = S_0\delta(f-f')\,. \tag{2.10}$$

The noise samples at different times are uncorrelated however small the difference $t-t'$ is. White noise is an idealisation. Noise can be considered white in some limited band where the function $S(f)$ is more or less flat.

We now define Gaussian white noise. Consider a data segment $[0,T]$ uniformly sampled with N samples and sampling interval Δ; we then have $T = N\Delta$ and $t_k = k\Delta,\ k = 0,1,...,N-1$. We integrate $\langle n(t)n(t_k)\rangle$ over a sample width:

$$\int_{t_k-\Delta/2}^{t_k+\Delta/2} dt\ \langle n(t)n(t_k)\rangle = S_0 \int_{t_k-\Delta/2}^{t_k+\Delta/2} dt\ \delta(t-t_k)\,. \tag{2.11}$$

The LHS gives $\langle n_k^2 \rangle \Delta$, while the RHS is just S_0. Thus we have the relation:

$$\langle n_k^2 \rangle = \frac{S_0}{\Delta} \equiv \sigma_\Delta^2 \,. \tag{2.12}$$

Thus the pdf for the noise vector $\mathbf{n}^T = (n_0, n_1, ..., n_{N-1})$ is:

$$p(\mathbf{n}) = \frac{1}{(\sqrt{2\pi}\sigma_\Delta)^N} \; e^{-\frac{1}{2}\frac{|\mathbf{n}|^2}{\sigma_\Delta^2}} \,, \tag{2.13}$$

where $|\mathbf{n}|^2 = \mathbf{n}^T\mathbf{n} = \sum_{k=0}^{N-1} n_k^2$ which is in fact the square of the Euclidean norm. This is the pdf of white Gaussian noise.

We can extend this discussion to coloured noise. We define the covariance matrix of the noise vector as:

$$C_{ik} = K(t_i - t_k) = K((i-k)\Delta) \,. \tag{2.14}$$

Then the noise pdf is:

$$p(\mathbf{n}) = \frac{1}{(2\pi)^{N/2}\sqrt{\det C}} \; e^{-\frac{1}{2}\mathbf{n}^T C^{-1}\mathbf{n}} \,. \tag{2.15}$$

We may define the matrix μ as the inverse of C, then in the exponent of the pdf we get the quantity $\mu_{ik} n_i n_k$. This motivates the definition of a scalar product. We define a scalar product of two data vectors $\mathbf{x}$ and $\mathbf{y}$ as:

$$(\mathbf{x}, \mathbf{y}) = \mu_{ik} x_i y_k \,, \tag{2.16}$$

then we can write:

$$p(\mathbf{n}) = \frac{(\det \mu)^{1/2}}{(2\pi)^{N/2}} \; e^{-\frac{1}{2}(\mathbf{n},\mathbf{n})} \,. \tag{2.17}$$

If we call the set of data trains as $\mathcal{D}$ which is essentially $\mathcal{R}^N$, then with the scalar product so defined, $\mathcal{D}$ is a Hilbert space. For the special case of white noise the metric μ_{ik} reduces to δ_{ik}, the Kronecker delta.

In the Fourier space in case of WSS noise, the metric μ_{ik} is diagonalised; $\mu_{ik} \longrightarrow 1/S(f)$. In fact, the equation (2.16) goes over to,

$$(\mathbf{x}, \mathbf{y}) = \int_{-\infty}^{\infty} df \; \frac{\tilde{x}^\star(f)\tilde{y}(f)}{S(f)} \,. \tag{2.18}$$

Since the vectors $\mathbf{x}, \mathbf{y}$ are real in the time domain, the scalar product is real.

The noise in a real detector is coloured and a detector has a bandwidth in which it is sensitive to signals. Any real detector has a finite response time in which it reacts to a signal. This response time determines the upper limit frequency of the frequency band in which the detector is sensitive.

2.3 The matched filter

Consider a data segment $[0, T]$ and a known signal $h(t)$. This signal is buried in the noise whose auto correlation function is $K(t, t')$. In general we have not put any restrictions of stationarity at this stage. Then the matched filter $q(t)$ is defined as the solution of the integral equation:

$$\int_0^T K(t, t') q(t') \; dt' = h(t) \,. \tag{2.19}$$

If we sample the data segment then this becomes a matrix equation. The matrix must be inverted to obtain the filter vector **q**.

If the noise is WSS, we have the equation:

$$\int_0^T K(t-t')q(t')\ dt' = h(t)\,. \tag{2.20}$$

The integral is a convolution and we can readily obtain q by going over to the frequency domain by taking Fourier transforms. Indeed in the Fourier space the above equation reduces to $\tilde{q}(f)S(f) = \tilde{h}(f)$ and since h is known, we immediately have the solution for the matched filter **q**:

$$\tilde{q}(f) = \frac{\tilde{h}(f)}{S(f)}\,. \tag{2.21}$$

Below the figure shows the plot of the correlation when a matched filter is applied to the data containing the GW binary signal. It is difficult to see the signal in the data. A peak in the correlation is observed at the time of arrival of the signal.

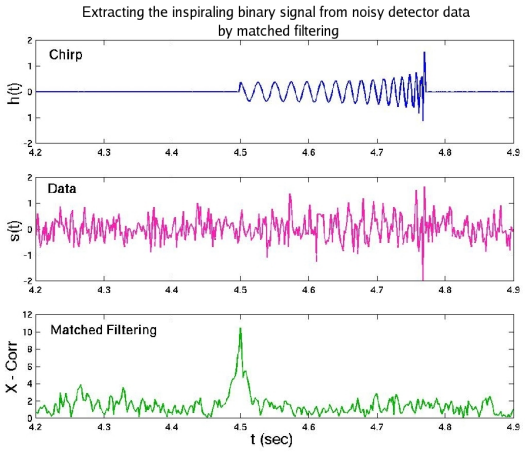

Figure 2.3: The operation of matched filtering is shown in the figure. The top shows the signal. The middle shows the signal buried in the noise. The bottom shows the output of the matched filter.

We now prove that the matched filter produces the highest signal to noise ratio (SNR) among all linear filters. This is in one way the matched filter is optimal. This statement is true irrespective whether the noise is Gaussian or not. First we must define the SNR of a given statistic. In Fourier space we have:

$$\tilde{x}(f) = \tilde{h}(f) + \tilde{n}(f)\,. \tag{2.22}$$

Apply the filter $\tilde{q}(f)$ to the data then we have the statistic c given by,

$$c = \int \tilde{q}^{\star}(f)\tilde{x}(f)\ df\,. \tag{2.23}$$

The integrals go from $-\infty$ to ∞ which we have not explicitly written. We note that c is real for real filters and real data. Note that c is also a r.v. since it depends on the data **x** which is a

r.v. The SNR is the ratio of the mean of c to its standard deviation. The mean of c is denoted by μ_c and it is given by:

$$\begin{aligned}\mu_c &= \langle c \rangle = \left\langle \int \tilde{q}^\star(f)\tilde{x}(f)\ df \right\rangle \\ &= \left\langle \int \tilde{q}^\star(f)\tilde{h}(f)\ df\rangle + \int \tilde{q}^\star(f)\tilde{n}(f)\ df \right\rangle \\ &= \int \tilde{q}^\star(f)\tilde{h}(f)\ df\,. \end{aligned} \tag{2.24}$$

The last line follows since the second term is zero because the mean of the noise is zero and the first term is deterministic. The variance of c, say σ_c^2, is obtained from taking the expected value of the filter operating on the noise only. Thus we have:

$$\begin{aligned}\sigma_c^2 &= \left\langle \int \tilde{q}(f)\tilde{n}^\star(f)\ df \int \tilde{q}^\star(f')\tilde{n}(f')\ df' \right\rangle \\ &= \int df df'\ \tilde{q}(f)\tilde{q}^\star(f')\ \langle \tilde{n}^\star(f)\tilde{n}(f')\rangle \\ &= \int df df'\ \tilde{q}(f)\tilde{q}^\star(f')\ S(f)\ \delta(f-f') \\ &= \int df\ |\tilde{q}(f)|^2 S(f)\,. \end{aligned} \tag{2.25}$$

The SNR which we denote by ρ is just μ_c/σ_c and is given by,

$$\rho(q) = \frac{\int \tilde{q}^\star(f)\tilde{h}(f)\ df}{[\int df\ |\tilde{q}(f)|^2 S(f)]^{1/2}}\,. \tag{2.26}$$

We observe that ρ depends on the filter q we apply which is itself a function, while the SNR is a number. Thus ρ maps a function to a real number - it is called a functional. Secondly, we also observe that scaling q, that is, $q \longrightarrow Aq$, where A is a constant, does not change ρ. Thus ρ is invariant under scalings of the filter q. We can use this fact to simplify the algebra in the calculations. Accordingly we set,

$$\int df\ |\tilde{q}(f)|^2 S(f) = 1\,. \tag{2.27}$$

We have then normalised the filter. The SNR for the normalised filter q is given by:

$$\rho = \int \tilde{q}^\star(f)\tilde{h}(f)\ df\,. \tag{2.28}$$

We now maximise $\rho(q)$ subject to the condition $\int df\ |\tilde{q}(f)|^2 S(f) = 1$. This is most conveniently done by invoking the Schwarz inequality. Write,

$$\int \tilde{q}^\star(f)\tilde{h}(f)\ df = \int \left(\tilde{q}^\star(f)\sqrt{S(f)}\right)\left(\frac{\tilde{h}(f)}{\sqrt{S(f)}}\right)\ df\,. \tag{2.29}$$

Then we have from the Schwarz inequality,

$$\begin{aligned}\left|\int \tilde{q}^\star(f)\tilde{h}(f)\ df\right| &\leq \left|\int |\tilde{q}(f)|^2 S(f)\ df\right|^{\frac{1}{2}} \left|\int df\ \frac{|\tilde{h}(f)|^2}{S(f)}\right|^{\frac{1}{2}} \\ &= \left|\int df\ \frac{|\tilde{h}(f)|^2}{S(f)}\right|^{\frac{1}{2}}\,. \end{aligned} \tag{2.30}$$

The maximisation is attained when we have proportionality between the two vectors or when the vectors point in the same direction. That is when,

$$\tilde{q}(f)\sqrt{S(f)} = A\frac{\tilde{h}(f)}{\sqrt{S(f)}}\,, \tag{2.31}$$

where A is a constant. The matched filter is determined upto a scaling. A can be fixed by requiring the filter $\tilde{q}(f)$ to have unit norm. Thus,

$$\tilde{q}(f) = A\frac{\tilde{h}(f)}{S(f)}\,. \tag{2.32}$$

The q constructed in this way maximises the SNR ρ.

$\mathbf{q}$ can also be obtained from calculus of variations if one maximises $\rho(\mathbf{q})$ subject to

$$\int df\ |\tilde{q}(f)|^2 S(f) = 1\,. \tag{2.33}$$

One must use the method of Lagrange multipliers and set the functional derivative of

$$F(\mathbf{q}) \equiv \rho(\mathbf{q}) - \lambda\left(\int df\ |\tilde{q}(f)|^2 S(f) - 1\right)\,, \tag{2.34}$$

to zero, where λ is a Lagrange multiplier. One then obtains Eq.(2.32).

Note that this result further generalises to the case of even nonstationary noise. One then must consider Eq. (2.19). If we discretise this equation, it becomes a matrix equation $C_{ik}q_k = h_i$. Assuming that C_{ik} is invertible we can solve for q_i, namely, $q_i = \mu_{ik}h_k$. One then follows on the same lines as above and identical result can be established for this general case, namely, that the matched filter maximises the SNR.

Lecture 3

The statistics and geometry of detection

3.1 Hypothesis testing

Given a data vector $\mathbf{x}$ we must decide whether a signal is present or absent in the data. The data vector $\mathbf{x}$ is a vector random variable with a multivariate probability distribution. The probability distributions will differ depending on whether there is signal present in the data or not. Given $\mathbf{x}$ we must decide which of these distributions the vector $\mathbf{x}$ belongs to; when there is no signal in the data, the data is just noise $\mathbf{x} = \mathbf{n}$, while if there is a signal $\mathbf{s}$ in the data then, $\mathbf{x} = \mathbf{s} + \mathbf{n}$. The hypotheses are called H_0 and H_1 respectively. The multivariate probability distributions are denoted by $p_0(\mathbf{x})$ when the hypothesis is H_0 and $p_1(\mathbf{x})$ when the hypothesis is H_1. Given $\mathbf{x}$ we must decide between the hypotheses H_0 and H_1. Since we have assumed additive noise we also have $p_1(\mathbf{x}) = p_0(\mathbf{x} - \mathbf{s})$.

To decide between the hypotheses, we device a test. A test amounts to partitioning the range of $\mathbf{x}$ into two disjoint regions R and R^c, that is if the range of $\mathbf{x}$ is $\mathcal{R}^N$, then $\mathcal{R}^N = R + R^c$, where the plus sign represents disjoint union. Thus if $\mathbf{x} \in R$, the signal is present or H_1 is true or else $\mathbf{x} \in R^c$, the signal is absent. Generally, the region R is far away from the origin of $\mathcal{R}^N$ while R^c is closer to the origin. The boundary of R namely, ∂R is called the decision surface D. Usually for well behaved probability distributions like Gaussian for example, D is a smooth $N-1$ dimensional hypersurface of $\mathcal{R}^N$. So now the problem of deciding whether H_0 or H_1 is true amounts to determining the region R according to some specific criteria. In order to do this we define the false alarm and detection probabilities as follows:

$$\begin{aligned} P_F(R) &= \int_R p_0(\mathbf{x})\ d^N x\,, \\ P_D(R) &= \int_R p_1(\mathbf{x})\ d^N x\,. \end{aligned} \tag{3.1}$$

P_F is called the false alarm probability and P_D is called the detection probability. Further $1 - P_D$ is called the false dismissal probability while $1 - P_F$ is called the confidence. The false alarm probability is the probability of the noise masquerading as a signal, that is, choosing H_1 when H_0 is in fact true. While the false dismissal probability is the probability of saying there is no signal in the data, when there is in fact a signal, that is choosing H_0 over H_1 when H_1 in fact is true.

We will see how we can use these probabilities in deciding R. First we discuss the one dimensional case when the data is just one quantity x. This is the simplest case and so easy to understand which nevertheless illustrates some important aspects of the problem. However, we

are not in this situation - we have a data train and consequently a data vector. We will next look at the 2-dimensional case which brings out features of the problem which are not present in the 1 dimensional case, but is relatively simpler than the N dimensional case. Finally, we will discuss the full N-dimensional case.

3.2 The 1 dimensional case

In discussing this case we will take Gaussian distributions. We will take Λ as the r.v then we have,

$$\begin{aligned} p_0(\Lambda) &= \frac{1}{\sqrt{2\pi}}\, e^{-\frac{\Lambda^2}{2}}\,, \\ p_1(\Lambda) &= \frac{1}{\sqrt{2\pi}}\, e^{-\frac{(\Lambda-A)^2}{2}}\,, \end{aligned} \tag{3.2}$$

where A is the amplitude of the signal. Implicit is the assumption that the noise is additive. We now fix the false alarm probability to be some small number say α. Normally, α is chosen small so that we will not often make the mistake of saying there is a signal when there is not. For GWDA, we must choose the false alarm probability that we can tolerate. Typically, $\alpha \sim 10^{-10}$ or 10^{-14} an extremely small number. It is usually decided by the event rate we expect from astrophysics. The false alarm rate must be much smaller than the event rate if we are to have confidence in our detection. Accordingly, we choose a value of Λ, say Λ_0, so that we have,

$$P_F = \frac{1}{\sqrt{2\pi}} \int_{\Lambda_0}^{\infty} d\Lambda\; e^{-\frac{\Lambda^2}{2}} = \alpha\,. \tag{3.3}$$

Thus fixing α determines Λ_0. The quantity Λ_0 is called the threshold. We say that a signal is present if $\Lambda \geq \Lambda_0$ otherwise we say that the signal is absent. Thus Λ_0 determines the region R, namely, $R = (\Lambda_0, \infty)$ while $R^c = (-\infty, \Lambda_0]$. The detection probability is then also determined because we have,

$$P_D = \frac{1}{\sqrt{2\pi}} \int_{\Lambda_0}^{\infty} d\Lambda\; e^{-\frac{(\Lambda-A)^2}{2}}\,. \tag{3.4}$$

The figure below depicts this situation:

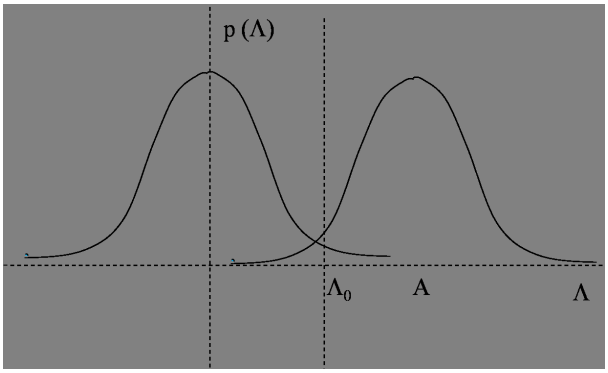

Figure 3.1: The probability distributions $p_0(\Lambda)$ and $p_1(\Lambda)$ are shown. The threshold Λ_0 is also shown which is fixed by the false alarm probability $P_F = \alpha$.

In fact in the one dimensional case, since the region R gets fixed by the false alarm probability, the detection probability is also determined. It is important to realise that this is not

the situation for vector data of more than 1 dimension; the false alarm proability does not fix the region R. We will discuss this situation in the next section.

We observe that the detection probability also depends on the signal amplitude A. The greater the amplitude, higher is the detection probability. Thus in this case this is how one handles the situation. Normally, we fix $P_F = \alpha$ and $P_D = \beta$, say 90%, then both these quantities fix A which we call $A_{\rm critical}$. We then take the decision of observing the signal with $P_F = \alpha$ and with $P_D > \beta$ if $A > A_{\rm critical}$.

3.3 2 dimensional case

When the data vector has more than one component, additional features of the problem come into play. The simplest case here is the two dimensional case where these features can be understood and pictures can be drawn to analyse the situation. We again consider Gaussian distributions for p_0 and p_1 and also for simplicity we will consider white noise, that is, the samples of $\mathbf{x} = (x_0, x_1)$ are uncorrelated and identically distributed. Thus we take,

$$\begin{aligned} p_0(\mathbf{x}) &= \frac{1}{2\pi}\, e^{-\frac{1}{2}(x_0^2+x_1^2)}\,, \\ p_1(\mathbf{x}) &= \frac{1}{2\pi}\, e^{-\frac{1}{2}[(x_0-s_0)^2+(x_1-s_1)^2]}\,. \end{aligned} \tag{3.5}$$

The signal now is a 2 dimensional vector $\mathbf{s} = (s_0, s_1)$. We could have considered correlation between x_0 and x_1 but that does not change the salient features while on the other hand it makes the algebra more complex. We will remark later on how one can generalise to this case. The false alarm and detection probabilities are defined as,

$$\begin{aligned} P_F(R) &= \int_R p_0(x_0, x_1)\; dx_0 dx_1\,, \\ P_D(R) &= \int_R p_1(x_0, x_1)\; dx_0 dx_1\,. \end{aligned} \tag{3.6}$$

where the region R is to be determined. We may fix the false alarm probability to some value α, but now fixing the false alarm probability *does not fix the region R nor the detection probability.* In 2 dimensions, the decision surface or the boundary ∂R is a *curve* and has more degrees of freedom. An infinity of curves and regions R are possible which yield the same false alarm probability α but different detection probabilities. This is the crucial difference between the one dimensional case and multidimensional case.

So then how does one go about the problem? One way is to assign costs for each type of decision. Then we have a cost matrix c_{ij}, where c_{ij} is the cost incurred by choosing hypothesis H_i, when H_j is true. A criterion could be to minimise the average cost given the prior probabilities for the hypotheses H_0 and H_1. This is called Bayes criterion. However, in GWDA one cannot assign costs in any meaningful way; there are no tangible costs to saying there is a GW signal, when there is no signal and vice-versa. The criterion that seems to be most appropriate in this situation is the Neyman-Pearson criterion. Here we fix the false alarm probability to some small value that we can tolerate and maximise the detection probability with respect to this constraint. This is a well posed problem and the solution is given by the Neyman-Pearson lemma which then determines the region R. We will discuss this later in detail. For now we focus on the 2 dimensional case.

Let us fix the false alarm probability to some value α. Now there is a wide choice for the region R and the decision surface - which is a curve. For simplicity, we take these curves as straight lines. Note that in principle it could be any complicated curve. We will see later from

the Neyman-Pearson lemma that the optimal curve turns out to be in fact a straight line. Given the circular symmetry of $p_0(\mathbf{x})$ around the origin, it is easy to check that the decision curves are any of the straight lines tangent to the circle of radius Λ_0, where the threshold Λ_0 is the solution to the equation:

$$\frac{1}{\sqrt{2\pi}} \int_{\Lambda_0}^{\infty} dx \; e^{-\frac{x^2}{2}} = \alpha \,. \tag{3.7}$$

See figure below: Since $p_1(\mathbf{x})$ is centred around the signal vector $\mathbf{s}$, the different regions R

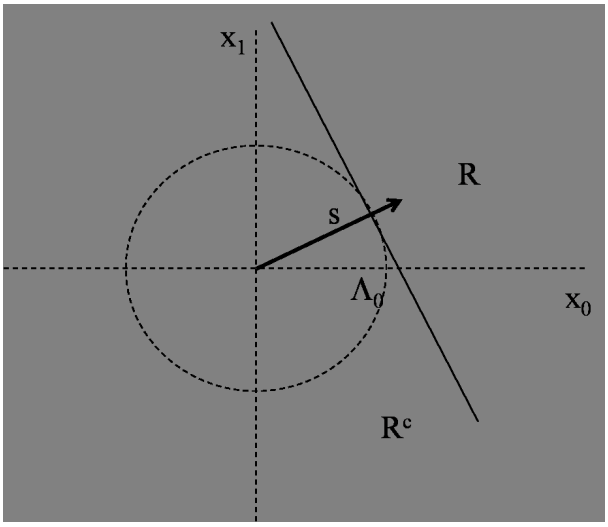

Figure 3.2: The regions R are the half planes whose boundaries are straight lines tangent to the circle of radius Λ_0 centred at the origin. We choose the region R which maximises the detection probability, namely when ∂R is furthest away from from the point $\mathbf{s}$, that is when the tangent is orthogonal to the direction of $\mathbf{s}$.

give different detection probabilities. The maximum detection probability is obtained when the decision surface ∂R is furthest away from the signal position vector $\mathbf{s}$. This occurs when the decision line is orthogonal to the direction of $\mathbf{s}$. Any other line, tangent to the circle will be closer to $\mathbf{s}$ and consequently will yield lower detection probability. Of course at this point we are only considering straight lines as the decision curves. How do we know that there is no other decision curve (and region R) which does not yield a higher detection probability? The answer is found from the Neyman- Pearson lemma which asserts that in this case the decision curve must be a straight line.

Also if we choose to fix the detection probability, then it must be less than this maximum detection probability. If we do fix it to be less than the maximum, there could be multiple solutions for R. In fact if we restrict to straight lines as decision curves, there are two tangents which are equidistant from the position vector $\mathbf{s}$, which will yield the same detection probabilties. In higher dimensions the choice is even more, since there are more degrees of freedom. In the next section we go over to the general case of multidimensions and discuss the Neyman-Pearson criterion and the lemma.

3.4 Neyman-Pearson criterion

When events are rare and the costs of making mistakes unquantifiable the Neyman-Pearson criterion is appropriate. It is so for the GW detection. We fix the false alarm probability

$P_F(R) = \alpha$ to a small number that can be tolerated. It is fixed essentially by the event rate that one expects. If one expects 40 or 50 events (say compact binary coalescences) in a year, then one may fix the false alarm rate to one false alarm per year. We may afford the mistake of making one false detection among 50 true events.

The Neyman-Pearson criterion states:

Maximise the detection probability $P_D(R)$ subject to $P_F(R) = \alpha$.

The question essentially is how do we find R? The answer is supplied by the Neyman-Pearson Lemma (which we do not prove). It essentially provides a prescription to find R, which is the following:

Define the likelihood ratio:

$$\Lambda(\mathbf{x}) = \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})}\,, \tag{3.8}$$

then the regions which maximise P_D are of the form $R(\Lambda_0) = \{\mathbf{x}/\Lambda(\mathbf{x}) \geq \Lambda_0\}$ where Λ_0 is a parameter.

Now we choose the parameter Λ_0 such that $P_F(R(\Lambda_0)) = \alpha$. This fixes Λ_0 which is called the threshold. The optimal region R in the sense of Neyman-Pearson is also fixed as well as the decision surface D which is then given by the equation $\Lambda(\mathbf{x}) = \Lambda_0$.

We illustrate the situation for Gaussian noise which in general could be coloured. Then the pdfs for the signal absent and signal present case are the following:

$$\begin{aligned} p_0(\mathbf{x}) &= A_N\, e^{-\frac{1}{2}\mu_{ik}x_i x_k}\,, \\ p_1(\mathbf{x}) &= A_N\, e^{-\frac{1}{2}\mu_{ik}(x_i-s_i)(x_k-s_k)}\,. \end{aligned} \tag{3.9}$$

Here $\mathbf{s}$ is the N dimensional signal vector with components $s_i, i = 0, 1, 2, ...N-1$, A_N is the normalisation constant which normalises the pdfs, μ_{ik} is the inverse of the covariance matrix and Einstein summation convention is used for repeated indices. We may now compute $\Lambda(\mathbf{x})$ which is the ratio of the two pdfs. But it is more convenient to compute the log likelihood ratio because of the exponentials in the pdfs and because the logarithm is a monotonic function of its argument. We can then set the threshold on the log likelihood ratio instead of the likelihood. We then have:

$$\begin{aligned} \ln\Lambda(\mathbf{x}) &= \mu_{ik}x_i s_k - \frac{1}{2}\mu_{ik}s_i s_k \\ &\equiv (\mathbf{x},\mathbf{s}) - \frac{1}{2}\parallel \mathbf{s} \parallel^2\,. \end{aligned} \tag{3.10}$$

The monotonicity of the functions involved allows us to translate the equation, $\Lambda(\mathbf{x}) \geq \Lambda_0$ to $\rho \geq \rho_0$ where,

$$\rho = (\mathbf{x},\mathbf{s}) = \mu_{ik}x_i s_k\,. \tag{3.11}$$

Thus we just need to compute the scalar product of the data vector with the signal vector. We may write,

$$\rho = q_i x_i, \qquad q_i = \mu_{ik}s_k\,, \tag{3.12}$$

where we recognise that q_i is just the matched filter!

Thus in Gaussian noise we have shown that the matched filter is also optimal in the Neyman-Pearson sense - it maximises the detection probability for a given false alarm probability. The decision surface is just $\mathbf{q}\cdot\mathbf{x} = \rho_0$ which is a hyperplane in $\mathcal{D}$.

Lecture 4

Composite hypothesis and geometry

4.1 Composite hypothesis and maximum likelihood detection

Although the above discussion is simple and easy to apply, we are not in this situation where there is a *single* known signal buried in the data. What we have in fact is a family of signals, say $h(t;\lambda^\alpha)$ which depend on a number of parameters $\lambda^\alpha,\ \alpha = 1, 2, ...p$. For example in the case coalescing binary sources, the signal depends on the individual masses, m_1, m_2, the amplitude A and other kinematical parameters like the time of arrival, initial phase etc. All these constitute the λ^α in this case and they span some subset $U \subset \mathcal{R}^p$. We collectively denote the λ^α by the vector λ. Thus $\lambda \in U$, where U is the domain of the parameters. Also we will denote the class of signals as $\mathbf{h}(\lambda)$, where now $\mathbf{h}(\lambda) \in \mathcal{D}$, the space of data trains. Thus we have a mapping from $U \longrightarrow \mathcal{D}$ where $\lambda \in U$ is mapped to $\mathbf{h}(\lambda) \in \mathcal{D}$. The image of U under this mapping is a p dimensional manifold - a submanifold of $\mathcal{D}$. This manifold we call the *signal* manifold. The signal parameters λ^α can be regarded as coordinates on the signal manifold. This why we denote them with the superscript α.

We have now to deal with a composite hypotheses:

- H_0: No signal present, that is, $\mathbf{x} = \mathbf{n}$, noise only and the corresponding pdf is $p_0(\mathbf{x})$.
- H_1: One among the family of signals $\mathbf{h}(\lambda)$ is present, that is, $\mathbf{x} = \mathbf{n} + \mathbf{h}(\lambda)$ with the corresponding pdf now denoted by $p_1(\mathbf{x};\lambda)$.

We have now a more complex situation here because now we have a family of pdfs $p_1(\mathbf{x};\lambda)$.

It turns out that we can still apply the Neyman-Pearson criterion - but now it is applicable to the average detection probability which we denote by $\langle P_D\rangle$, namely, that it is this average detection probability which is maximised for a given false alarm $P_F = \alpha$. The average is defined in the following way:

Consider a prior on the parameter space, which we denote by $z(\lambda)$. We take it to be normalised:

$$\int_U z(\lambda)\ d\lambda = 1\,. \tag{4.1}$$

Then the average detection probability is given by,

$$\langle P_D\rangle = \int_U z(\lambda)\ P_D(\lambda)\ d\lambda = \int_U d\lambda\ z(\lambda) \int_R d^N\mathbf{x}\ p_1(\mathbf{x};\lambda)\,. \tag{4.2}$$

Then the $\langle P_D\rangle$ is maximised for a given $P_F = \alpha$ for the region R which is now defined via the average likelihood ratio:

$$\langle\Lambda\rangle(\mathbf{x}) = \int_U z(\lambda)\ \Lambda(\mathbf{x};\lambda)\ d\lambda\,, \tag{4.3}$$

where $\Lambda(\mathbf{x};\lambda) = p_1(\mathbf{x};\lambda)/p_0(\mathbf{x})$. The regions $R(\Lambda_0)$ are now of the form $\langle\Lambda\rangle(\mathbf{x}) \geq \Lambda_0$. We fix Λ_0 by requiring $P\{\langle\Lambda\rangle(\mathbf{x}) \geq \Lambda_0\} = \alpha$. Then the region R is fixed and the average detection probability $\langle P_D\rangle$ is maximised.

If $z(\lambda)$ is a slowly varying function of the parameters λ and the likelihood ratio $\Lambda(\mathbf{x};\lambda)$ has a sharp peak at some $\lambda = \lambda_{\max}$, then the dominant contribution to the average likelihood ratio $\langle\Lambda\rangle$ comes from this peak. Therefore we have,

$$\langle\Lambda\rangle \propto \max_{\lambda} \Lambda(\mathbf{x};\lambda)\,. \tag{4.4}$$

The detection then can be based on the statistic $\Lambda_{\max}$, where,

$$\Lambda_{\max} = \max_{\lambda} \Lambda(\mathbf{x};\lambda)\,. \tag{4.5}$$

This is called maximum likelihood detection (MLD). Therefore we compute $\Lambda(\mathbf{x};\lambda)$ for all $\lambda \in U$ and then take the maximum over λ and compare the same with a pre-assigned threshold to decide on the detection.

Apart from deciding on the detection, we also obtain the estimates of the parameters of the signal, that is, the parameters $\lambda_{\max}$ at which the likelihood ratio $\Lambda(\mathbf{x};\lambda)$ is maximised. These are called maximum likelihood estimates (MLE).

4.2 Ambiguity function and the metric

Sometimes if some of the parameters of the signal are kinematical such as the amplitude, time of arrival or the phase, they can be easily dealt with in various quick ways. If the statistic is a linear function of the data such as the matched filter, we can analytically maximise over the amplitude. We take the compact binary inspiral as a typical example. In order to search over amplitude, we may define a normalised template $\mathbf{s}(\lambda)$ such that $\parallel \mathbf{s}(\lambda) \parallel = 1$, where now λ represents all parameters other than the amplitude. The signal is then $\mathbf{h}(\lambda) = A\ \mathbf{s}(\lambda)$ where A is the amplitude of the signal. The data is then:

$$\mathbf{x} = \mathbf{h}(\lambda) + \mathbf{n} = A\ \mathbf{s}(\lambda) + \mathbf{n}\,. \tag{4.6}$$

Correlating with $\mathbf{s}(\lambda)$ we have,

$$c = \langle \mathbf{s}, \mathbf{x}\rangle = A + \langle \mathbf{s}, \mathbf{n}\rangle\,. \tag{4.7}$$

The expected value of c is A, since the second term vanishes. The point is that we do not need to search over the amplitude. Moreover, we also obtain an estimate of the amplitude of the signal.

Similarly we can use the fast Fourier transform to search a transient signal over the time of arrival. The phase can be maximised using quadratures and so on. But there could be other parameters on which the signal depends, which cannot be searched over so easily. For the inspiraling binaries these are the masses, spins etc. Here one needs to densely cover the parameter space with templates - a bank of templates - so that one does not as far as is possible miss any signal. On the other hand since computing resources are finite and limited we have to seek a compromise on the number of templates, so their number must be minimised by placing them as far apart as is possible. This number is decided by the maximum mismatch between the signal and the template that can be tolerated. If we are prepared to lose say 10 % of the GW sources, then a maximum mismatch of 3 % in the correlation can be tolerated; a 3 % loss in SNR leads to 10 % loss in the detection volume as the volume goes as the cube of the distance

to which we can observe the source. The mismatch is quantified in terms of what is called the ambiguity function. We define the ambiguity function as:

$$\mathcal{H}(\lambda, \lambda') = (\mathbf{s}(\lambda), \mathbf{s}(\lambda'))\,, \tag{4.8}$$

where the signals $\mathbf{s}(\lambda)$ are normalised as mentioned before and the brackets denote the scalar product as defined earlier. One immediately deduces that $|\mathcal{H}(\lambda, \lambda')| \leq 1$. This is immediate from the Schwarz inequality,

$$\mathcal{H}(\lambda, \lambda') = (\mathbf{s}(\lambda), \mathbf{s}(\lambda')) \leq \| \mathbf{s}(\lambda) \| \| \mathbf{s}(\lambda') \| = 1\,. \tag{4.9}$$

Since the templates have norm unity, they lie on a submanifold of a hypersphere of $N-1$ dimensions. The ambiguity function can be thought of as a cosine of the angle between the unit vectors $\mathbf{s}(\lambda)$ and $\mathbf{s}(\lambda')$.

We can now place the templates with the help of the ambiguity function. If the maximum mismatch is taken to be small like 3 % for example, then the template parameters do not differ much, so that we may write, $\lambda' = \lambda + \Delta\lambda$, where $\Delta\lambda$ is small. Then $\mathcal{H}$ is Taylor expanded to the second order; the first derivative vanishes because when the parameters match, $\mathcal{H}$ is maximum. Thus,

$$\begin{aligned} \mathcal{H}(\lambda, \lambda + \Delta\lambda) &\simeq 1 + \frac{1}{2}\frac{\partial^2 \mathcal{H}}{\partial\lambda^\alpha \partial\lambda^\beta}\Delta\lambda^\alpha \Delta\lambda^\beta\,, \\ &\equiv 1 - g_{\alpha\beta}\Delta\lambda^\alpha\Delta\lambda^\beta\,, \end{aligned} \tag{4.10}$$

where we have defined a metric on the parameter space as,

$$g_{\alpha\beta} = -\frac{1}{2}\frac{\partial^2 \mathcal{H}}{\partial\lambda^\alpha\partial\lambda^\beta}\,. \tag{4.11}$$

This is how it has been defined in some of the literature. However, geometrically, if $\Delta\theta$ is the angle between $\mathbf{s}(\lambda)$ and $\mathbf{s}(\lambda + \Delta\lambda)$, then $\mathcal{H} = \cos\Delta\theta \simeq 1 - \Delta\theta^2/2$. And because the angle $\Delta\theta$ is geometrically the distance on a unit (hyper)sphere, we should actually define the metric as:

$$\mathcal{H}(\lambda, \lambda + \Delta\lambda) = 1 - \frac{1}{2} g_{\alpha\beta}\Delta\lambda^a\Delta\lambda^\beta\,, \tag{4.12}$$

where $g_{\alpha\beta}\Delta\lambda^a\Delta\lambda^\beta$ is now the square of the distance $\Delta\theta^2$ on the unit hypersphere. Further, this is also the induced metric on the parameter space where the parameter space is regarded as a submanifold of $\mathcal{D}$ on which we already have the metric μ_{ik}. Then it is easy to show that,

$$g_{\alpha\beta} = \mu_{ik}\frac{\partial s^i}{\partial\lambda^\alpha}\frac{\partial s^k}{\partial\lambda^\beta}\,. \tag{4.13}$$

However, since this factor of 2 in the definitions does not affect the physical results when used consistently, it may be regarded as a convention. We therefore follow the first formula given in Eq.(4.10) so as to be in tune with the literature.

Since some of the parameters, namely, the kinematical parameters like the time of arrival or the phase can be dealt with in a quick manner, the remaining parameters must be searched over with the help of templates. We therefore maximise the match over the kinematical parameters and redefine the ambiguity function over these remaining parameters and call it the reduced ambiguity function. We will still denote the parameters by λ with the understanding that now λ does not include the kinematical parameters and also continue to denote the reduced ambiguity function by $\mathcal{H}$ and the corresponding metric by $g_{\alpha\beta}$.

If we call the maximum mismatch as ϵ then we require that $\mathcal{H}(\lambda, \lambda') \geq 1 - \epsilon \equiv MM$, where then MM is called the minimal match. Thus for the typical case we have considered, $\epsilon = 0.03$ and $MM = 0.97$. The placement of templates is governed by the equation:

$$g_{\alpha\beta}\Delta\lambda^{\alpha}\Delta\lambda^{\beta} = \epsilon\,. \tag{4.14}$$

Typically the contours traced out by $\Delta\lambda^{\alpha}$ are hyperellipsoids, because $g_{\alpha\beta}\Delta\lambda^{a}\Delta\lambda^{\beta}$ is a positive definite quadratic form - this is because $\mathcal{H}$ attains a true maximum when the parameters coincide. For points lying inside the ellipsoid, the mismatch is less than ϵ. The figure below shows the two dimensional case. The contours now are ellipses as seen in the figure.

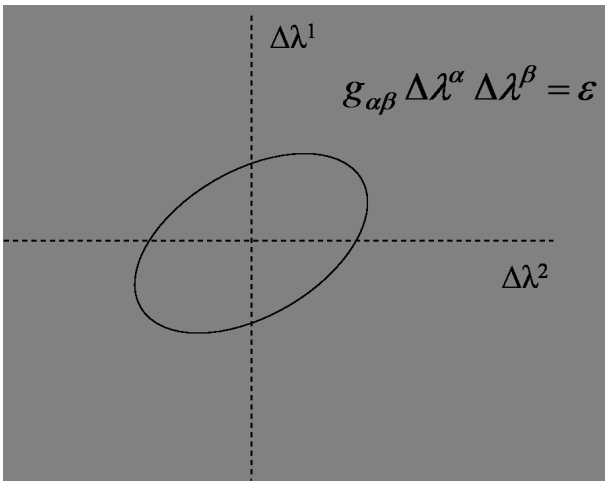

Figure 4.1: A plot of $g_{\alpha\beta}\Delta\lambda^{\alpha}\Delta\lambda^{\beta} = \epsilon$ is shown for the parameters λ^1 and λ^2 where only two dimensions are considered. The origin of the coordinate system has been shifted to the coordinates of the template. The contour is an ellipse.

Given the parameter space we can also compute the number of templates required to cover the parameter space with a given mismatch ϵ in terms of the metric. This formula is due to Owen. Let Δl be the side of the hypercube which fits inside the hyperellipsoid. The template is at the centre of the hypercube. The vertex furthest away from the centre is at distance whose square is $N(\Delta l/2)^2$. This results from the repeated application of Pythagorus theorem. Thus we have the relations:

$$g_{\alpha\beta}\Delta\lambda^{\alpha}\Delta\lambda^{\beta} = N\left(\frac{\Delta l}{2}\right)^2 = \epsilon = 1 - MM\,. \tag{4.15}$$

The number of templates $N_{\text{templates}}$ is the volume of the parameter space divided by the volume of the hypercube, that is,

$$N_{\text{templates}} = \frac{\int_U d\lambda\sqrt{\det(g)}}{\Delta l^p}\,. \tag{4.16}$$

Solving for Δl in terms of the mismatch gives us the result:

$$N_{\text{templates}} = \frac{\int_U d\lambda\sqrt{\det(g)}}{(2\sqrt{1 - MM/p})^p}\,. \tag{4.17}$$

The above formula gives a rough estimate of the number of the templates required to search the parameter space. However, the problem is really that of 'tiling' the parameter space. That is the

templates must span the parameter space with the given minimal match, that is leave no 'holes' in the parameter space, namely, at no point in the parameter space the match should fall below the minimal match. This is one condition. The opposing condition is that one must be able to achieve this with a minimum number of templates in order to minimise the computational cost. These are two opposing criteria which then govern the placement of templates. There are significant overlaps among the neighbouring templates which must be minimised. So the packing of templates matters. Also there are boundary effects where templates placed near the boundary of the parameter space 'spill' outside the boundary - for example the parameter space can become narrow in certain regions, nevertheless templates are required in order to span the space but then they spill out in other directions outside the parameter space. These considerations affect the number of templates required.

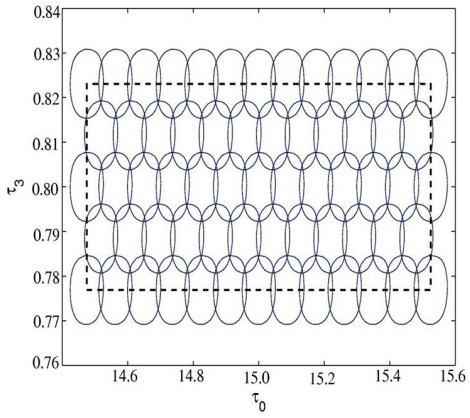

Figure 4.2: An hexagonal packing of templates for the nonspinning binary inspiral parameter space. Hexagonal packing is more efficient than square packing by about 23 %.

Coming back to the arrangement or the patterns, in two dimensions, the hypercube (in this case a square) is not the optimal way to tile. A hexagonal packing reduces the overlap significantly. This reduces the number of templates by about 23 %. In the figure below we show how templates can be packed in a hexagonal pattern in the case of the inspiraling binary signals. The parameters are the chirp times τ_0 and τ_3 which are functions of the two masses m_1 and m_2.