Statistical Techniques

Tarun Deep Saini

Indian Institute of Science, Bangalore, India
tarun@physics.iisc.ernet.in

1. A random process is one whose outcome does not *seem* to follow a deterministic pattern. Of course, it could well be that the underlying laws are deterministic, and even simple to state. However, our knowledge of the parameters of the system is often limited, and if the system is sufficiently sensitive to those parameters, it virtually becomes impossible to predict the outcome of the process.

2. A crucial aspect of a random process is the label that differentiates one outcome from another. A specific outcome of a random process can in principle have several labels that can apply to it. For example, a process in which a random fruit is picked can have labels such as 'sweet', 'sour', 'red', 'green', 'big', 'small', etc. The labels should be unambiguous to ensure mutual exclusivity, therefore, if a fruit is red and green, then we shall choose red-green as a distinct label to red or green. A simple random process is defined to be such that it has a single label and a compound process as one that can have multiple labels. Moreover, the labels could be discrete as well as continuous. The labels even or odd are discrete, but the label T that gives the temperature at a given point in space is a continuous label.

3. A class of random processes is statistically deterministic, that is although we are not able to predict a single outcome, we are still able to predict the relative frequencies of distinct outcomes. This is done in a probabilistic manner as follows. Let us consider a simple discrete random process X with outcomes belonging to a set U, that is $X_i \in U \equiv \{X_1, X_2, X_3, \ldots\}$. Clearly, X_is are the mutually exclusive labels of the random outcomes of the process. By repeatedly drawing outcomes[1], we can construct frequencies P_{X_i} through
$$P_{X_i} = \lim_{N_X \to \infty} \frac{N_{X_i}}{N_X}, \tag{1}$$
where N_{X_i} is the number of times X_i occurs in the experiment and N_X is the total number—regardless of the X label—of random events in the experiment. If this limit exists and is well defined then P_{X_i} is called the probability of obtaining X_i and $P_X = \{P_{X_i}\}$ is called the probability distribution function (PDF).

   Examples: Throwing a dice, drawing cards randomly from a deck, tossing a coin, etc.

4. The expectation value of a function F of X_i is defined as
$$\langle F \rangle = \sum_i P_{X_i} F(X_i). \tag{2}$$

[1] Drawing outcomes means repeating the process to obtain many random samples. For example, throwing a coin repeatedly generates a random sequence of Heads and Tails.

From the definition of probability, the frequency interpretation of the expectation value is

$$\langle F \rangle = \lim_{N_X \to \infty} \sum_i \frac{N_{X_i} F(X_i)}{N_X}, \tag{3}$$

which we see to be the sample mean (average) of a very large sample. We see below that we can also think of this as average over an ensemble.

5. From a logical point of view it is useful to define the concept of an *ensemble* of identically prepared systems. The limit over N_X could then be thought of as limit over the ensemble. For example, instead of throwing a single coin a thousand times, we can throw a thousand coins a single time. This makes it possible to talk about the time evolution of the PDF. For example in the case of throwing a coin, if the coin suffers wear and tear on each throw, the likelihood of obtaining a Head or a Tail may change with time. An ensemble of identical coins can capture this phenomenon easily.

6. If the PDF for a system is invariant with respect to time then the process is called stationary. For stationary processes, limit over ensemble is identical to the limit over time. As an example of a non-stationary process consider a dice that accumulates dirt and grime unevenly over a period of time. In this case its PDF might evolve with time.

7. An ensemble is a useful theoretical device, however, often it is impossible to realize it in practice. The assumption of stationarity then makes it possible to estimate the PDFs for such systems. In other cases there could be theoretical reasons to believe that the distribution has a certain form. In such cases the PDF is a theoretical model for the random process and has to be verified by experiment.

Compound processes

8. Let us consider a random processes with two labels, X and Y. To determine the probability that the event E_{XY} has the outcome $E_{X_iY_j}$, we consider the limit

$$P_{X_iY_j} = \lim_{N_{XY} \to \infty} \frac{N_{X_iY_j}}{N_{XY}}, \tag{4}$$

which can be written as

$$P_{X_iY_j} = \lim_{N \to \infty} \frac{N_{X_iY_j}}{N_{XY_j}} \times \frac{N_{XY_j}}{N_{XY}}. \tag{5}$$

Here, the ratio N_{XY_j}/N_{XY} denotes the frequency of obtaining events E_{XY_j}, where X is unspecified (it can have an arbitrary value). Mathematically,

$$E_{XY_j} = \sum_i E_{X_iY_j}. \tag{6}$$

Later we will learn that this quantity is called the marginalised distribution of Y. Clearly, this is the probability of obtaining Y_j. The first ratio, similarly, assumes that Y_j is already the label of any of the outcomes, and the ratio that we are seeking is called the conditional probability of obtaining X_i given that Y_j is already a label of the event. Both together can be written as

$$P_{X_iY_j} = P(X_i|Y_j)P(Y_j)\,. \tag{7}$$

More informally, omitting the subscripts, this takes the form

$$P(XY) = P(X|Y)P(Y)\,. \tag{8}$$

The left hand side is called the 'joint probability' of events X and Y. More specifically:

$$\begin{aligned} P(XY) &\equiv \text{Joint probability of } X \text{and } Y & (9)\\ P(X|Y) &\equiv \text{Conditional probability of } X \text{given } Y. & (10) \end{aligned}$$

By interchanging the roles of X and Y, it is easy to prove the Bayes theorem:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}\,. \tag{11}$$

Example: A medical test for a rare disease has an efficacy of 90 percent. One percent population of a country is infected with the disease. If an individual is tested positive then what is the probability that he is infected? (8 per cent)

Example: Solve the above problem assuming that the person tests positive in two repeated tests.

Example: Solve the Monty Hall problem using Bayes' theorem.

9. If $P(Y|X) = P(Y)$, or in other words the probability of Y does not depend on X, then the labels X and Y are said to be independent.

   Example: Give up to five examples of independent labels.

   Example: Look up `http://en.wikipedia.org/wiki/Sally_Clark` and write a short essay on the case, elucidating the aspects of probability relevant to the case.

10. It is important to note that the multiple labels do no mean that the process is necessarily multi-dimensional. For example, in a process where a natural number is picked randomly, multiple labels can be produced by demanding different properties of natural numbers, X could be the set of even numbers and Y could be the set of primes. In this case, the number two is a member of both the sets and we can non trivially talk about the probability of finding a number that satisfies both properties. However, we can also imagine processes where two distinct random processes are used as joint labels, for example, on the earth every point has a temperature and pressure, then, a joint event could be that at a given place the temperature and pressure are in a certain range.

Continuous random process

11. A continuous random process X is one whose outcome is continuous. In this case the PDF $P(X)$ is a probability density in the following sense. The probability of obtaining an outcome in the range X and $X + dX$ is given by

$$\Delta p = P(X)\Delta X, \tag{12}$$

where ΔX is small.

12. By generalization, the expectation value of a function $F(X)$ is

$$\langle F\rangle = \int P(X)F(X)dX \tag{13}$$

In particular the expectation value of $F(X) = X$, called the mean, is

$$\mu = \langle X\rangle = \int XP(X)dX \tag{14}$$

A random process that has zero mean is called *centered*.

13. The variance of random variable X is defined as

$$\sigma^2 = \langle (X-\mu)^2\rangle = \int (X-\mu)^2 P(X)dX, \tag{15}$$

and σ is called the standard deviation.

14. It should be noted that the usual notion of mean and variance are defined using samples. The definitions above give the ensemble mean and variance which are differentiated from the other usage by appending the word 'sample'. Thus, one talks of sample mean and sample variance.

Measuring simple physical quantities

15. Measurement of any physical quantity is always accompanied by the introduction of random noise due to a variety of reasons. Let us consider the process of measuring the time period of a pendulum. Imagine that the measurement is done in a manner that produces continuous, non-discrete numbers . This requires imagining a clock that generates output with an infinite number of significant figures, although we do not demand that the readings be accurate. Imagine that the clock is stopped by a person or a device every time the pendulum crosses zero, and the readings on the clock are registered. The readings $\{T_i\}$ then are such that

$$T_i = T_0 + \epsilon_i, \tag{16}$$

where T_0 is the intrinsic time period and ϵ_i is a random number, changing from measurement to measurement. Let us assume that the noise is on the average zero, i.e., $\langle \epsilon_i\rangle = 0$. If we take N such measurements, then how shall we estimate the true time period T_0 of the pendulum? We will learn later how to do this more formally, here we shall go by the intuitive notion of taking the mean of the measurements. That is, the required estimate is given by

$$\bar{T} = \frac{1}{N}\sum_i T_i, \tag{17}$$

where $\bar{T}$ is the sample mean/average of sample $\{T_i\}$. It is clear that this estimate will be close to the true value T_0 but in general will not coincide with it. Therefore, a more appropriate estimate should quote a likely range and not just a number. We typically give as range the variance of the random quantity (we shall later on see how to choose the range on the basis of probability theory).

16. By the presence of ϵ_i in each measurement, the estimate $\bar{T}$ is a random variable. Therefore, we can compute its mean and variance by taking the expectation value of appropriate quantities. It is easy to see that since $\langle\epsilon_i\rangle = 0$

$$\langle\bar{T}\rangle = T_0. \tag{18}$$

If the standard deviation of ϵ is σ, then we calculate the variance of $\bar{T}$ through

$$\sigma_{\bar{T}}^2 = \langle(\bar{T} - \langle\bar{T}\rangle)^2\rangle) = \langle\bar{T}^2\rangle - T_0^2. \tag{19}$$

Consider the term

$$\langle\bar{T}^2\rangle = \frac{1}{N^2}\sum_{ij}\langle T_0^2 + (\epsilon_i + \epsilon_j)T_0 + \epsilon_i\epsilon_j\rangle = \frac{1}{N^2}\sum_{ij}(T_0^2 + \sigma^2\delta_{ij}) = T_0^2 + \frac{\sigma^2}{N}. \tag{20}$$

Using this it is easy to see that the standard deviation

$$\sigma_{\bar{T}} = \frac{\sigma}{\sqrt{N}}. \tag{21}$$

17. The canonical quote for the estimate of the time period of the pendulum is then

$$T = \bar{T} \pm \frac{\sigma}{\sqrt{N}} \tag{22}$$

18. We see that as N becomes large, the estimate becomes more and more precise. The notion of precision is related to the smallness of the standard deviation of the estimate.

19. If $\langle\epsilon_i\rangle \neq 0$ then you can show that the previous argument for standard deviation of $\bar{T}$ still holds, but the expectation value of $\bar{T}$ fails to coincide with the true value T_0. Such random noise is called systematic and such measurements give a biased estimate of the measured quantity.

$$\langle\bar{T}\rangle = T_0 + T_b \tag{23}$$

and the quoted value above would have the unknown quantity T_b sitting inside $\bar{T}$, making the estimate useless unless one can estimate T_b independently.

Example: Show that the result for the standard deviation of $\bar{T}$ remains unchanged for biased noise.

20. An estimate is precise if its standard deviation is small compared to the measured quantity. It is also accurate if no bias is present, i.e., $T_b = 0$. If bias is present then an estimate could be precise but not accurate.

The characteristic function; PDFs of means of distributions

21. The mean and variance are the most commonly used statistical descriptors of random processes. In general we define the mth moment of the random variable X as

$$\langle X^m\rangle = \int_{-\infty}^{\infty} X^m P(X) dX. \tag{24}$$

It is easy to show that $\sigma^2 = \langle X^2\rangle - \langle X\rangle^2$. The moments of a probability distribution determine the distribution uniquely. This can be seen more clearly below where we show that the moments of a distribution determine the Fourier transform of the PDF uniquely.

22. The characteristic function of $P(X)$ is defined as the expectation value of $\exp(iKX)$

$$\hat{P}(K) = \langle \exp(iKX) \rangle = \int_{-\infty}^{\infty} \exp(iKX)P(X)dX. \tag{25}$$

The characteristic function is the Fourier transform of the PDF. Since the PDF is normalized, we see that the characteristic function always exist. However, the moments of the random process (derived below) may not exist.

23. It is easy to see that

$$\langle X^m \rangle = \left(\frac{1}{i}\right)^m \frac{d^m}{dK^m} \hat{P}(K)\Big|_{K=0}. \tag{26}$$

This shows that the moments of a PDF determine the Fourier transform of the PDF through a Taylor expansion uniquely.

24. A centered **Gaussian** random process is one for which

$$\hat{G}(K) = \exp\left(-\frac{1}{2}\sigma^2 K^2\right). \tag{27}$$

The PDF for a centered Gaussian process is easily seen to be

$$G(X) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{X^2}{2\sigma^2}\right). \tag{28}$$

25. **Cauchy distribution**: Another distribution of great interest due to its properties is

$$P(X) = \frac{1}{\pi} \frac{1}{1+X^2}. \tag{29}$$

By contour integration, it is easy to see that the characteristic function for the centered Cauchy distribution is given by

$$\hat{P}(K) = \exp(-|K|). \tag{30}$$

Prove this result.

26. It is easy to show that if the characteristic function of a centered distribution is $\hat{P}(K)$ then the characteristic function of this distribution when it is shifted to $X = \mu$ is given by $\exp(iK\mu)\hat{P}(K)$.

27. **Change of variables**: Let $P_Y(Y)$ be the PDF of Y. Let us determine the PDF of X, where $Y = Y(X)$. By definition

$$dP = P_X(Y)dX = P_Y(Y)dY = P_Y(Y(X))\frac{dY}{dX}dX. \tag{31}$$

Therefore,

$$P_X(X) = P_Y(Y(X)) \times \frac{dY}{dX}. \tag{32}$$

It is easy to generalize this to a multidimensional PDF, where instead of a simple derivative we will find the Jacobian of the transformation.

28. **Generating random numbers that follow** $P(X)$: Let $C(X)$ be the cumulative distribution function of $P(X)$

$$C(X) = \int_{-\infty}^{X} P(Z)dZ \tag{33}$$

if Y is a uniform random number in the range 0—1 then

$$X = C^{-1}(Y) \tag{34}$$

is distributed as $P(X)$. **Prove this using the change of variable rule**.

29. **Characteristic function of a scaled variable**: Let us consider $Y = NX$, where N is an arbitrary number (not necessarily a natural number)

$$\begin{aligned} \hat{P}_X(K) &= \int_{-\infty}^{\infty} P_X(X)\exp(iKX)dX & (35) \\ &= \int_{-\infty}^{\infty} P_Y(NX)\exp\left[i\left(\frac{K}{N}\right)(NX)\right]d(NX) = \hat{P}_Y(K/N). & (36) \end{aligned}$$

30. When we consider the mean of a sample to determine the parameter that defines the centre of the distribution (this need not be the mean in case it is undefined, for instance in the case of Cauchy distribution), a relevant question is how effective is it? The result quoted above about the variance of the mean of a sample shows that the variance falls with the number of samples if the mean and variance exists. This can fail to be true for distributions where the mean and variance are not defined. For example, for the Cauchy distribution we have seen that the characteristic function is given by $\exp(-|K|)$. Since the derivatives of this function are not defined at $K = 0$, all the moments of this function are ill-defined. Consider the mean of a sample $\{X_i\}$

$$\bar{X} = \frac{1}{N}\sum X_i. \tag{37}$$

It is clear that $\bar{X}$ is a random number. How do we calculate its distribution function? It can be shown by using the convolution theorem (to be proved later) that if the characteristic function of X_i is $\hat{P}_X(K)$ then

$$\hat{P}_{N\bar{X}}(K) = \hat{P}_X^N(K). \tag{38}$$

By the result of previous calculation this implies that

$$\hat{P}_{\bar{X}}(K) = \hat{P}_X^N(K/N). \tag{39}$$

Applying this result to the Gaussian distribution we obtain

$$\hat{P}_{\bar{X}}(K) = \exp\left(-\frac{1}{2}\frac{\sigma^2}{N}K^2\right), \tag{40}$$

and for the Cauchy distribution

$$\hat{P}_{\bar{X}}(K) = \exp(-|K|). \tag{41}$$

The result for the Gaussian distribution is consistent with our previous calculation. However, for the Cauchy distribution we find that the mean follows the same distribution as the sample distribution. Therefore, the mean is no better than choosing a random element of the sample! We shall later see how to estimate the centre of the Cauchy distribution from a sample using the maximum likelihood estimation.

Example: Distributions with tails fatter than the Gaussian distribution are called heavy tailed distributions. Cauchy distribution is an example of this. Review the following pages and write a page long summary.

http://en.wikipedia.org/wiki/L%C3%A9vy_flight
http://en.wikipedia.org/wiki/Fat_tail
http://en.wikipedia.org/wiki/Black_swan_theory

Propagation of errors

31. We are often called upon to quote the estimate of a derived quantity. For concreteness, let us consider the simple pendulum again. If we can measure its time period and its length then it is possible to estimate the acceleartion due to gravity g through the relation

$$T = 2\pi\sqrt{\frac{l}{g}}. \tag{42}$$

Rearranging this gives

$$g = \frac{4\pi^2 l}{T^2}. \tag{43}$$

32. To quote an estimate for g, let us first consider the general problem of estimating $f(X_i)$, where X_i are the N physical measured parameters. The result for the estimate is actually very simple

$$\bar{f} = f(X_{i0}), \tag{44}$$

that is, the estimate of f is the function $f(X_i)$ evaluated at the estimate X_{i0}. We next consider the range that we should quote for the estimate.

The first step is to linearize the function around the measured estimates X_{i0}

$$f(X_i) = f(X_{i0}) + \sum_i^N \frac{\partial f}{\partial X_i}(X_i - X_{i0}). \tag{45}$$

The variance of f is given by

$$\sigma_f^2 = \langle (f(X_i) - f(X_{i0})^2 \rangle = \sum_i^N \sum_j^N \frac{\partial f}{\partial X_i}\frac{\partial f}{\partial X_j}\langle (X_i - X_{i0})(X_j - X_{j0})\rangle. \tag{46}$$

We define

$$\langle (X_i - X_{i0})(X_j - X_{j0})\rangle = \sigma_{ij} \tag{47}$$

as the covariance matrix. If the variables X_is are independently measured then it is clear that the covariance matrix is diagonal and is given by $\sigma_{ij} = \sigma_i^2\delta_{ij}$, where σ_i^2 is the variance of parameter X_i. For this case

$$\sigma_f^2 = \sum_i^N \left(\frac{\partial f}{\partial X_i}\right)^2 \sigma_i^2. \tag{48}$$

By dividing both sides by $\bar{f}^2$, this can be written as

$$\frac{\sigma_f^2}{\bar{f}^2} = \sum_i^N \left(\frac{\partial \ln f}{\partial X_i}\right)^2 \sigma_i^2. \tag{49}$$

Applying this to the problem of measuring the acceleration due to gravity by a simple pendulum we obtain

$$\frac{\sigma_g^2}{g^2} = \left(\frac{\sigma_l}{l}\right)^2 + \left(\frac{2\sigma_T}{T}\right)^2. \tag{50}$$

A result that can more simply be obtained using the elementary notion of propagation of errors using the log rule. This derivation, in fact, proves the log rule.

33. **The Central Limit Theorem**: Let a random process Z be the sum of a large number, N, of random variables X_i

$$Z = \sum_i^N X_i, \tag{51}$$

and P_i is the centered PDF corresponding to X_i. By the convolution theorem the characteristic function for Z is given by

$$\hat{P_Z}(K) = \exp\left(\sum_i iK\mu_i\right)\prod_i \hat{P_i}(K). \tag{52}$$

Then, Taylor expanding $\hat{P_i}(K)$ we obtain

$$\hat{P_Z}(K) = \exp\left(\sum_i iK\mu_i\right)\prod_i \left(1 - \frac{1}{2}\sigma_i^2K^2 + \mathcal{O}(K^3)\right). \tag{53}$$

If the mean and the variance of Z are well defined and finite then it follows that $\sigma_i \ll 1$, otherwise the sum of σ_i^2 (which is the variance of Z) would diverge. Therefore, the leading term in the above product is the second order term. Using the approximation $1 + \Delta \cong \exp(\Delta)$ if $\Delta \ll 1$, we get

$$\hat{P_Z}(K) = \exp\left(iK\mu\right)\exp\left(-\frac{1}{2}\sigma^2K^2\right), \tag{54}$$

where μ and σ^2 are the sum of means and variances of X_is respectively. This is the characteristic function for a Gaussian process and thus it proves the central limit theorem.

34. In this proof the number of terms has been assumed to be very large. In case the number of terms is small, the higher terms start contributing as well. The central behaviour of the distribution of Z, in fact, turns out to be close to a Gaussian even for small number of terms, but the tails converge to Gaussian more slowly.

35. The utility of CLT is that it allows us to assume many distributions to be Gaussian if physical arguments exist to show that the random process is created by a sum of a large number of tiny effects. For example, when we receive light from a distant star, the photons are randomly scattered by tiny amounts due to fluctuations in the refractive index of air due to turbulence. The sum of a large number of tiny kicks through CLT then produces a Gaussian distribution of angles of deflection, leading to the standard Gaussian form of the point spread function.

Multivariate Gaussian random processes

36. It the outcome of a random process is not a scalar but a vector $\boldsymbol{X} = \{X_1, X_2, X_3, \ldots\}$, where X_is are continuous variables, then it is called a multivariate random process.

37. A multivariate random process is Gaussian if for any arbitrary vector $\mathbf{K}$, $\mathbf{K} \cdot \mathbf{X}$ is a Gaussian random variable.

38. The characteristic function is given by

$$\hat{G}(\mathbf{K}) = \langle \exp\left(i\,\mathbf{K} \cdot \mathbf{X}\right) \rangle = \exp\left(-\frac{1}{2}\langle \mathbf{K} \cdot \mathbf{X} \rangle^2\right) = \exp\left(-\frac{1}{2} K_i K_j \langle X_i X_j \rangle\right). \qquad (55)$$

39. The distribution function is the inverse Fourier transform of this and is given by

$$G(\mathbf{X}) = \frac{1}{(2\pi)^{N/2}\sqrt{\det(C)}} \exp\left(-\frac{1}{2} C_{ij}^{-1} X_i X_j\right), \qquad (56)$$

where we have defined $C_{ij} = \langle X_i X_j \rangle$ as the covariance matrix. We note that the distribution function of a multivariate random process is completely specified by the covariance matrix. In fact, all the higher moments of a Gaussian random field can be derived from the second moment.

40. If the random variables X_is are statistically independent then the covariance matrix is diagonal.

The Likelihood function

41. Till now we have focused our attention on simple measurements where the quantity being measured is either the quantity of interest or can be derived from it straightforwardly. Let us revisit the problem of measuring the time period of a simple pendulum. More formally, this problem can formulated in the following manner: A given measurement

$$T_i = T_0 + \epsilon_i. \qquad (57)$$

If the PDF of the noise ϵ is a known function $P_\epsilon(\epsilon) = f(\epsilon)$, then the above equation gives us $P_T(T) = f(T - T_0)$, that is, we can determine the PDF of the time measurements T. Clearly,

$$\int_{-\infty}^{\infty} f(T - T_0) dT = 1. \tag{58}$$

Note that till now this distribution is not completely determined due to the fact that T_0 is not known. Given $P_T(T)$, we can determine the probability that the time periods will be in the range T to $T + \Delta T$ as $\Delta p = P_T(T)\Delta T$. Let there be N measurements of time periods, T_i. The probability of data can be obtained by asking for the probability for the data to be in the range T_i and $T_i + \Delta T_i$. If the measurements are assumed to be independent then this is given by

$$dp = \prod_i^N f(T_i - T_0) d^N T \equiv L(T_i; T_0) d^N T, \tag{59}$$

where we have defined the *Likelihood* function $L(T_i; T_0)$. The dependence of this function is on the data T_i and the unknown parameter T_0. Clearly,

$$\int L(T_i; T_0) d^N T = 1. \tag{60}$$

42. The expectation value of an arbitrary function of data is given by

$$\langle f(T_i) \rangle = \int f(T_i) L(T_i; T_0) d^N T. \tag{61}$$

We use this below to obtain the variance of an estimator.

43. We have earlier estimated the parameter T_0 by computing the average of T_0. In general, an *estimator* of temperature T_0 is denoted as $\hat{T}(T_i)$ (not to be confused with the characteristic function). Therefore, our earlier calculation can be written as

$$\hat{T}(T_i) = \frac{1}{N} \sum_i T_i. \tag{62}$$

The estimator $\hat{T}$ takes as its input data T_i and produces an estimate of the quantity of interest T_0, which itself is not directly measured.

44. **Unbiased estimator**: This estimator of T_0 is unbiased if

$$\int L(T_i; T_0) \hat{T}(T_i) d^N T = T_0, \tag{63}$$

and it is easy to see that if the PDF $f_\epsilon(\epsilon)$ has a well defined mean, and if the mean is zero then this will be true.

45. **Cramér-Rao inequality**: Given an estimator its it is clear that as the data changes the estimate would jump around randomly. There is a very important bound that can be

obtained on the variance of an arbitrary estimator in the following manner. Differentiate Eq 63 w.r.t. T_0 to obtain

$$\int \frac{dL(T_i;T_0)}{dT_0}\hat{T}(T_i)d^NT = 1, \tag{64}$$

Differentiating Eq. 60 w.r.t. T_0 and then multiplying it with T_0 gives

$$\int T_0 \frac{dL(T_i;T_0)}{dT_0} d^NT = 0, \tag{65}$$

The previous two equation can be combined to give

$$\int \frac{dL(T_i;T_0)}{dT_0}\left[\hat{T}(T_i) - T_0\right] d^NT = 1, \tag{66}$$

This can be written as

$$\int \left\{\sqrt{L(T_i;T_0)}\frac{d\log L(T_i;T_0)}{dT_0}\right\}\left\{\sqrt{L(T_i;T_0)}\left[\hat{T}(T_i) - T_0\right]\right\} d^NT = 1, \tag{67}$$

Using Cauchy-Schwarz inequality

$$\left[\int f(x)g(x)dx\right]^2 \leq \int f(x)^2dx \int g(x)^2dx \tag{68}$$

we obtain

$$1 \leq \int L(T_i;T_0)\left(\frac{d\log L(T_i;T_0)}{dT_0}\right)^2 d^NT \int L(T_i;T_0)\left[\hat{T}(T_i) - T_0\right]^2 d^NT \tag{69}$$

Or in other words

$$\sigma^2_{\hat{T}} \geq \left[\left\langle\left(\frac{d\log L(T_i;T_0)}{dT_0}\right)^2\right\rangle\right]^{-1} \tag{70}$$

Example: Prove that

$$\left\langle\left(\frac{d\log L(T_i;T_0)}{dT_0}\right)^2\right\rangle = -\left\langle\frac{d^2\log L(T_i;T_0)}{dT_0^2}\right\rangle \tag{71}$$

Maximum likelihood estimator

46. The likelihood function furnishes a very simple way to obtain an estimator for a parameter by demanding that the probability for data be maximum for that estimate. For the present case, the estimator for T_0 is obtained by solving for T_0 in the equation

$$\frac{d}{dT_0}L(T_i;T_0) = 0. \tag{72}$$

47. Let us use this idea on a specific PDF. If

$$f(\epsilon) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right) \tag{73}$$

then the likelihood function is given by

$$L(T_i; T_0) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2}\sum_i \frac{(T_i - T_0)^2}{\sigma^2}\right). \tag{74}$$

It is clear that maximizing the likelihood function is equivalent to minimizing

$$\chi^2 = \sum_i \frac{(T_i - T_0)^2}{\sigma^2} \tag{75}$$

Carrying out the algebra we find that the estimator for T_0 for a Gaussian random noise is the same as the sample mean of T_i. That is

$$\hat{T}_{\text{ML}} = \bar{T}. \tag{76}$$

Example: Using the Cramér-Rao inequality show that for an arbitrary estimator $\hat{T}$

$$\sigma^2_{\hat{T}} \geq \frac{\sigma^2}{N}. \tag{77}$$

Also verify Eq. 71. By the results of our previous calculations it becomes clear that $\hat{T}_{\text{ML}}$ given above satisfies the equality sign and is therefore the best possible estimator (it has the minimum variance).

Example: Find the maximum likelihood estimator for T_0 if the noise PDF $f(\epsilon)$ is given by the Cauchy distribution and show that it is not the same as the sample average.

Example: In the above example we have assumed that the noise distribution is fully known to us. If all that we know is that the distribution is a centred Gaussian but we do not know σ then ML method can be used to determine σ as well. Find the required estimator. Also show that if in addition the mean of noise is not zero then the bias cannot be measured using the ML technique.

48. More generally, the measured quantity is a known function of parameters of a theory

$$y = g(x, \theta_i) + \epsilon, \tag{78}$$

where x is a single control parameter; θ_is are the parameters of the theory; and ϵ is the noise. Data typically consists of N measurements of y at various values of the control parameter $\{x_i, y_i\}$. The control parameter too is a result of an experiment and is therefore likely to be noisy. However, it turns out that in general handling errors on both is intractable. Therefore, below we assume that the control parameter is, in fact, measured without any error. The likelihood function in this case takes the form

$$L(y_i, x_i; \theta_j) = \prod_i^N f(y_i - g(x_i; \theta_j)). \tag{79}$$

Example: A simple example is Ohm's law $V = IR$, where I, the current, could be the control parameter that we can vary at will, and the measurements of V then forms the data $\{I_i, V_i\}$

49. For Gaussian noise with known properties this leads to the ML estimator through the minimization of the χ^2 function

$$\chi^2 = \sum_i^N \frac{(y_i - g(x_i;\theta_j))^2}{\sigma_i^2}, \tag{80}$$

where we have further generalized to the case where each measurement can have its own separate variance σ_i. Since CLT states that noise resulting from a large number of small additive random events tends to the Gaussian distribution, maximum lieklihood method is usually encapsulated in the form of minimizing the χ^2 function, which is the de facto method of choice for most parameter estimation problems.

The ML estimators for θ_j can be obtained by solving for θ_js in the following equations

$$\frac{\partial \chi^2}{\partial \theta_j} = 0. \tag{81}$$

Note that this procedure gives as many equations as there are parameters θ_js, and the solution seems to be well determined. However, the equations may not all be independent if the number of data points N is smaller than the the number of parameters. For example, for a two-dimensional linear fit, it is obvious that a single data point does not determine a unique fit since an infinity of lines can pass through a point.

50. It is often convenient to rescale the model by defining $y_i' = y_i/\sigma_i$ and $y'(x_i) = y(x_i)/\sigma_i$ to obtain

$$\chi^2 = \sum_i^N (y_i' - y'(x_i;\theta_j))^2. \tag{82}$$

This makes the χ^2 a sum of terms with expectation value unity. We will often assume a scaling of this form for convenience.

51. **Linear Model**: The ML estimator for a linear model can be computed analytically. We consider a linear model of the form

$$y(x) = \sum_{\alpha=1}^m \theta_\alpha g_\alpha(x), \tag{83}$$

where θ_α are the m parameters of the linear model, and $g_\alpha(x)$ are the m arbitrary functions. If the data consists of N entries of the form $\{x_i, y_i\}$ then

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \theta_1 g_1(x_1) & \theta_2 g_2(x_1) & \ldots & \theta_m g_m(x_1) \\ \theta_1 g_1(x_2) & \theta_2 g_2(x_2) & \ldots & \theta_m g_m(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ \theta_1 g_1(x_N) & \theta_2 g_2(x_N) & \ldots & \theta_m g_m(x_N) \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}.$$

This can be written as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} g_1(x_1) & g_2(x_1) & \ldots & g_m(x_1) \\ g_1(x_2) & g_2(x_2) & \ldots & g_m(x_2) \\ \vdots & \vdots & \ldots & \vdots \\ g_1(x_N) & g_2(x_N) & \ldots & g_m(x_N) \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_N \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix},$$

which in the matrix form takes on a neat expression

$$\mathrm{y} = \mathrm{G}\theta + \epsilon, \tag{84}$$

where the square matrices are denoted with a capital letter and the parameters θ are the true values. Using this χ^2 can be written as

$$\chi^2 = (\mathrm{y}^{\mathrm{T}} - \theta^{\mathrm{T}}\mathrm{G}^{\mathrm{T}})(\mathrm{y} - \mathrm{G}\theta). \tag{85}$$

Note that here we have assumed the scaling desribed above. The estimator for θ can be computed from this function by minimizing this function and leads to the following equation

$$\mathrm{G}^{\mathrm{T}}\mathrm{G}\theta = \mathrm{G}^{\mathrm{T}}\mathrm{y}, \tag{86}$$

which can be solved to obtain

$$\hat{\theta} = (\mathrm{G}^{\mathrm{T}}\mathrm{G})^{-1}\mathrm{G}^{\mathrm{T}}\mathrm{y}; \quad \hat{\theta}^{\mathrm{T}} = \mathrm{y}^{\mathrm{T}}\mathrm{G}(\mathrm{G}^{\mathrm{T}}\mathrm{G})^{-1}. \tag{87}$$

As expected the estimator is a function of data vector y and is therefore a random vector.

52. **The expectation value of χ^2_{min}**: To determine the quality of fit it is necessary to consider the numerical value of the residual χ^2 function. This can be easily done analytically for the linear model considered above

$$\mathrm{y} = \mathrm{G}\theta + \epsilon. \tag{88}$$

It is convenient to choose the true parameters $\theta_i = 0$ without any loss of generality.

$$\mathrm{y} = \epsilon. \tag{89}$$

We are interested in evaluating the χ^2 at the best fit parameters $\hat{\theta}$. Upon expanding the χ^2 function we obtain

$$\chi^2 = \mathrm{y}^{\mathrm{T}}\mathrm{y} + \theta^{\mathrm{T}}\mathrm{G}^{\mathrm{T}}\mathrm{G}\theta - 2\mathrm{y}^{\mathrm{T}}\mathrm{G}\theta \tag{90}$$

now we substitute $\hat{\theta}$ for θ to obtain

$$\chi^2_{\mathrm{min}} = \mathrm{y}^{\mathrm{T}}\mathrm{y} + \mathrm{y}^{\mathrm{T}}\mathrm{G}(\mathrm{G}^{\mathrm{T}}\mathrm{G})^{-1}\mathrm{G}^{\mathrm{T}}\mathrm{G}(\mathrm{G}^{\mathrm{T}}\mathrm{G})^{-1}\mathrm{G}^{\mathrm{T}}\mathrm{y} - 2\mathrm{y}^{\mathrm{T}}\mathrm{G}(\mathrm{G}^{\mathrm{T}}\mathrm{G})^{-1}\mathrm{G}^{\mathrm{T}}\mathrm{y} \tag{91}$$

Use the fact that $(\mathrm{G}^{\mathrm{T}}\mathrm{G})^{-1}\mathrm{G}^{\mathrm{T}}\mathrm{G} = \mathrm{I}$, where I is an $m \times m$ identity matrix. It is easy to see that $\mathrm{I}\mathrm{G}^{\mathrm{T}} = \mathrm{G}^{\mathrm{T}}$ and $\mathrm{G}\mathrm{I} = \mathrm{G}$, therefore, $\mathrm{G}(\mathrm{G}^{\mathrm{T}}\mathrm{G})^{-1}\mathrm{G}^{\mathrm{T}}\mathrm{G} = \mathrm{G}$. This simplies the above expression to

$$\chi^2_{\mathrm{min}} = \mathrm{y}^{\mathrm{T}}\mathrm{y} - \mathrm{y}^{\mathrm{T}}\mathrm{G}(\mathrm{G}^{\mathrm{T}}\mathrm{G})^{-1}\mathrm{G}^{\mathrm{T}}\mathrm{y} \tag{92}$$

Now use the fact that $\mathrm{x}^{\mathrm{T}}\mathrm{A}\mathrm{x} = \mathrm{tr}[\mathrm{A}(\mathrm{x}\mathrm{x}^{\mathrm{T}})]$ to recast the previous equation in the form

$$\chi^2_{\mathrm{min}} = \mathrm{y}^{\mathrm{T}}\mathrm{y} - \mathrm{tr}[(\mathrm{G}^{\mathrm{T}}\mathrm{G})^{-1}\mathrm{G}^{\mathrm{T}}\mathrm{y}\mathrm{y}^{\mathrm{T}}\mathrm{G}] \tag{93}$$

Taking the expectation value we get

$$\langle\chi^2_{\mathrm{min}}\rangle = \langle\mathrm{y}^{\mathrm{T}}\mathrm{y}\rangle - \mathrm{tr}[(\mathrm{G}^{\mathrm{T}}\mathrm{G})^{-1}\mathrm{G}^{\mathrm{T}}\langle\mathrm{y}\mathrm{y}^{\mathrm{T}}\rangle\mathrm{G}] \tag{94}$$

Noting that $\langle \mathrm{y}^{\mathrm{T}}\mathrm{y}\rangle = N$ and $\langle \mathrm{y}\mathrm{y}^{\mathrm{T}}\rangle = \mathrm{J}$, where J is a $N \times N$ identity matrix, and also $\mathrm{JG} = \mathrm{G}$, we obtain

$$\langle \chi^2_{\mathrm{min}}\rangle = N - \mathrm{tr}[(\mathrm{G}^{\mathrm{T}}\mathrm{G})^{-1}\mathrm{G}^{\mathrm{T}}\mathrm{G}] \tag{95}$$

As noted earlier $(\mathrm{G}^{\mathrm{T}}\mathrm{G})^{-1}\mathrm{G}^{\mathrm{T}}\mathrm{G} = \mathrm{I}$

$$\langle \chi^2_{\mathrm{min}}\rangle = N - \mathrm{tr}[I] = N - m = \nu \equiv \mathrm{d.o.f.} \tag{96}$$

where d.o.f. is an abbreviation of degrees of freedom. It is also customary to define the reduced chi-square, $\chi^2_\nu = \chi^2_{\mathrm{min}}/(N-m)$. The above result then says that $\langle \chi^2_\nu\rangle = 1$

Example: Prove that if we do not assume $\theta_{i0} = 0$ then the above algebra still goes through, thus showing that the result is independent of value of parameters.

53. Although proved for the linear model, this result approximatley holds for non-linear models too if the errors are small since a non-linear function can always be linearized around the best fit, and small errors ensure that the linear approximation holds good.

Bayesian Statistics

54. Let us write the likelihood function using the abstract notation

$$P(\mathrm{D}|\boldsymbol{\theta};\mathrm{M},\mathrm{I}) = \mathrm{L}(\boldsymbol{y};\boldsymbol{\theta}), \tag{97}$$

where D denotes data $\boldsymbol{y}$ (containing N entries); the symbol $\boldsymbol{\theta}$ denotes the m parameters θ_j; the symbol M denotes the model being considered and I is whatever information is available to us. Using the Bayes theorem we can write

$$P(\boldsymbol{\theta}|\mathrm{D};\mathrm{M},\mathrm{I}) = \frac{\mathrm{P}(\mathrm{D}|\boldsymbol{\theta};\mathrm{M},\mathrm{I})\mathrm{P}(\boldsymbol{\theta};\mathrm{M},\mathrm{I})}{\mathrm{P}(\mathrm{D};\mathrm{M},\mathrm{I})} \tag{98}$$

Let us define the various symbols in this equation in words

$$\begin{aligned}
P(\mathrm{D}|\boldsymbol{\theta};\mathrm{M},\mathrm{I}) &\equiv \text{Probability of data given } \boldsymbol{\theta} \text{ (likelihood function)}\\
P(\boldsymbol{\theta};M,I) &\equiv \text{Prior probability distribution of parameters}\\
P(\boldsymbol{\theta}|\mathrm{D};\mathrm{M},\mathrm{I}) &\equiv \text{Posterior probability distribution of parameters}\\
P(\mathrm{D};\mathrm{M},\mathrm{I}) &\equiv \text{Probability of data given the model M}
\end{aligned}$$

The auxiliary symbols M,I denote the context in which the inversion is carried out. We shall return to this usage in a later section. However, for present, it is clear what the context is and we can omit the symbols for brevity.

$$P(\boldsymbol{\theta}|\mathrm{D}) = \frac{\mathrm{P}(\mathrm{D}|\boldsymbol{\theta})\mathrm{P}(\boldsymbol{\theta})}{\mathrm{P}(\mathrm{D})} \tag{99}$$

Also, note that $P(\mathrm{D})$ is a number that determines the overall normalization, which can be fixed after the calculation. Inserting the likelihood function we obtain

$$P(\boldsymbol{\theta}|\mathrm{D}) \propto \mathrm{L}(\boldsymbol{y};\boldsymbol{\theta})\mathrm{P}(\boldsymbol{\theta}) \tag{100}$$

55. The prior probability $P(\boldsymbol{\theta})$ requires careful considertation of the type of parameter that one is dealing with. However there is one choice that makes the Bayesian method give results identical to what one obtains through the maximum likelihood function. Note that all probability statements made in the likelihood method pertained to the data and not to the parameters of the theory. But with this inversion we now have the probability distribution of over parameters themselves. If $P(\boldsymbol{\theta}) =$ uniform then

$$P(\boldsymbol{\theta}|\mathrm{D}) \propto \mathrm{L}(\boldsymbol{y};\boldsymbol{\theta}) \tag{101}$$

56. To see how this can be used to obtain common sensical results (already obtained by different methods before) we now consider our old example of the problem of measuring the time period of a simple pendulum.

$$P(T_0|\mathrm{D}) \propto \mathrm{L}(\mathrm{T_i};\mathrm{T_0}) \propto \exp\left(-\frac{1}{2}\sum_{\mathrm{i}}\frac{(\mathrm{T_i}-\mathrm{T_0})^2}{\sigma^2}\right). \tag{102}$$

Expanding gives

$$P(T_0|\mathrm{D}) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{\mathrm{i}}(\mathrm{T_i^2}+\mathrm{T_0^2}-2\mathrm{T_i}\mathrm{T_0})\right]. \tag{103}$$

Any term not containing T_0 can be absorbed in the normalization, therefore,

$$P(T_0|\mathrm{D}) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{\mathrm{i}}(\mathrm{T_0^2}-2\mathrm{T_i}\mathrm{T_0})\right] \propto \exp\left[-\frac{1}{2\sigma^2}\left(\mathrm{N}\mathrm{T_0^2}-2\mathrm{T_0}\sum_{\mathrm{i}}\mathrm{T_i}\right)\right]. \tag{104}$$

Rearranging terms and adding a constant (not a function of T_0) to the exponent gives

$$P(T_0|\mathrm{D}) \propto \exp\left[-\frac{\mathrm{N}}{2\sigma^2}\left(\mathrm{T_0^2}-2\mathrm{T_0}\bar{\mathrm{T}}+\bar{\mathrm{T}}^2\right)\right]. \tag{105}$$

which can be written as

$$P(T_0|\mathrm{D}) = \sqrt{\frac{\mathrm{N}}{2\pi\sigma^2}}\exp\left[-\frac{\mathrm{N}}{2\sigma^2}\left(\mathrm{T_0}-\bar{\mathrm{T}}\right)^2\right]. \tag{106}$$

We see that the probability distribution for T_0 is centred at the sample mean and is a Gaussian distribution with variance $\sigma_{T_0}^2 = \sigma^2/N$. This is what we deduced earlier using the ML method.

57. Consider a general problem where there are multiple parameters $\boldsymbol{\theta}$

$$P(\boldsymbol{\theta}|\mathrm{D}) \propto \mathrm{L}(\boldsymbol{y}|\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\chi^2(\boldsymbol{y},\boldsymbol{\theta})\right) \tag{107}$$

The probability peaks at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, and it is obvious that $\nabla_\theta\chi^2(\boldsymbol{y},\boldsymbol{\theta})|_{\theta_0} = 0$. Then Taylor expanding χ^2 around $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ we obtain

$$P(\boldsymbol{\theta}|\mathrm{D}) \propto \mathrm{L}(\boldsymbol{y}|\boldsymbol{\theta}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\cdot\mathbf{F}\cdot(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right], \tag{108}$$

where constant terms have been absorbed in the normalization. In terms of index notion this can be written as

$$P(\boldsymbol{\theta}|\mathrm{D}) \propto \mathrm{L}(\boldsymbol{y}|\boldsymbol{\theta}) \propto \exp\left[-\frac{1}{2}\sum_{\mathrm{i,j}}(\theta_{\mathrm{i}}-\theta_{0\mathrm{i}})\mathrm{F}_{\mathrm{ij}}(\theta_{\mathrm{j}}-\theta_{\mathrm{j}0})\right] \tag{109}$$

The matrix F is called the Fisher matrix and it is given by

$$F_{ij} = \frac{1}{2}\frac{\partial^2\chi^2}{\partial\theta_i\partial\theta_j}\bigg|_{\boldsymbol{\theta}_0} = -\frac{\partial^2\log L}{\partial\theta_i\partial\theta_j}\bigg|_{\boldsymbol{\theta}_0} \tag{110}$$

The covariance matrix that encodes the errors on parameters is given by the inverse of this matrix

$$\mathrm{C}_{\mathrm{ij}} = \mathrm{F}^{-1}_{\mathrm{ij}} \tag{111}$$

This is proved below.

58. In Eq. 109 above translate the coordinate system so the origin coincides with the best fit $\boldsymbol{\theta}_0$, thereby simplifying it to

$$P(\boldsymbol{\theta}|\mathrm{D}) \propto \exp\left(-\frac{1}{2}\sum_{\mathrm{i,j}}\theta_{\mathrm{i}}\mathrm{F}_{\mathrm{ij}}\theta_{\mathrm{j}}\right). \tag{112}$$

By diagonalizing the Fisher matrix and carrying out the integration over parameters it can be easily shown that

$$P(\boldsymbol{\theta}|\mathrm{D}) = \sqrt{\frac{\det(\mathrm{F})}{(2\pi)^{\mathrm{N}}}}\exp\left(-\frac{1}{2}\sum_{\mathrm{i,j}}\theta_{\mathrm{i}}\mathrm{F}_{\mathrm{ij}}\theta_{\mathrm{j}}\right) = \sqrt{\frac{\det(\mathrm{F})}{(2\pi)^{\mathrm{N}}}}\exp\left(-\frac{1}{2}\theta^{\mathrm{T}}\mathrm{F}\theta\right). \tag{113}$$

Since F is a symmetric matrix it can be diagonalized with an orthogonal transformation. Let $\mathcal{O}$ be that transformation then from properties of orthogomal transformations we know that the inverse transformation is given by $\mathcal{O}^{\mathrm{T}}$. Let the transformed coordinates be $\eta = \mathcal{O}\theta$, then the probability distribution function for η can be written as

$$P(\boldsymbol{\eta}|\mathrm{D}) = \sqrt{\frac{\det(\mathrm{G})}{(2\pi)^{\mathrm{N}}}}\exp\left(-\frac{1}{2}\eta^{\mathrm{T}}\mathrm{G}\eta\right) = \sqrt{\frac{\det(\mathrm{G})}{(2\pi)^{\mathrm{N}}}}\exp\left[-\frac{1}{2}\sum_{\mathrm{i}}\mathrm{G}_{\mathrm{ii}}\eta_{\mathrm{i}}^2\right], \tag{114}$$

where the diagonal matrix $G = \mathcal{O}\mathrm{F}\mathcal{O}^{\mathrm{T}}$. Recall that the Jacobian of this transformation is unity due to orthogonality condition. Consider the direct product matrix $[\eta\eta^{\mathrm{T}}]_{ij} = \eta_i\eta_j$. Given the $P(\boldsymbol{\eta}|\mathrm{D})$ above, it follows that

$$\langle\eta_i\eta_j\rangle = \frac{1}{\mathrm{G}_{\mathrm{ii}}}\delta_{ij} = \mathrm{G}^{-1}_{ij}, \tag{115}$$

which in the matrix notation becomes

$$\langle\eta\eta^{\mathrm{T}}\rangle = \mathcal{O}\langle\theta\theta^{\mathrm{T}}\rangle\mathcal{O}^{\mathrm{T}} = \mathrm{G}^{-1} \tag{116}$$

which implies

$$\langle\theta\theta^{\mathrm{T}}\rangle = \mathcal{O}^{\mathrm{T}}\mathrm{G}^{-1}\mathcal{O} = \mathrm{F}^{-1} \tag{117}$$

This proves the result given in Eq. 111

59. **Marginalization over parameters**: In the case of multi-parameter fitting we often need to quote the results for a single parameter. For example, in cosmology a data set could be used to determine the amount of dark matter in the universe, however, the fitting function will often contain other parameters, such as those pertaining to dark energy and Hubble parameter. In this case the posterior probability of a single parameter can be obtained by integrating over the other parmeters. Formally, if the theory has parameters $\theta_1, \theta_2, \cdots, \theta_k$ and if we are interested only in parameter θ_1 then

$$P_m(\theta_1|\mathrm{D}) = \int \mathrm{P}(\theta_1, \cdots, \theta_\mathrm{k}|\mathrm{D})\mathrm{d}\theta_2\mathrm{d}\theta_3\cdots\mathrm{d}\theta_\mathrm{k}, \tag{118}$$

where the subscript m denotes marginalized . It is easy to see that this is a genuine PDF and integrates out to unity. This PDF is called the marginalized distribution of θ_1. It is easy to see graphically that the one sigma range for the marginalized distribution is smaller than that of the original distribution.

60. **Marginalization using the Fisher matrix**: The above prescription gives the most general method for marginalizing over nuisance parameters. Often when the data quality is good, the Gaussian approximation (Eq. 109) holds good. In this situation the Fisher matrix can be used to determine the marginalized distribution for any subset of parameters. As an example let us say we are interested in parameters θ_1 and θ_3 but not in others, in this case we should be working with

$$P_m(\theta_1, \theta_2) = \int P(\theta_1, \theta_2, \theta_3 \cdots, \theta_k|\mathrm{D})\mathrm{d}\theta_2\mathrm{d}\theta_4\cdots\mathrm{d}\theta_\mathrm{k} \tag{119}$$

Under the assumption of small errors, this distribution is close to a Gaussian and is fully determined by the correlations $\langle\theta_1^2\rangle$, $\langle\theta_2^2\rangle$ and $\langle\theta_1\theta_2\rangle$. The two dimensional Fisher matrix describing the Gaussian distribution for θ_1 and θ_2 is given by

$$\mathrm{F}_{ij}^{(2)} = \left[\begin{array}{cc} \langle\theta_1^2\rangle & \langle\theta_1\theta_2\rangle \\ \langle\theta_1\theta_2\rangle & \langle\theta_2^2\rangle \end{array}\right]^{-1}. \tag{120}$$

By Eq. 117, this can be obtained by inverting the projected full Fisher matrix F on the θ_1, θ_2 plane.

61. **Model Selection**: If there are several models capable of explaining the data, the Bayesian statistics can be used to select the best amongst them. Let the models be denoted as $\{M_i\}$. Let us rewrite the full expression of posterior probability for the parameters of model M_i with intrinsic parameters $\boldsymbol{\theta}$.

$$P(\boldsymbol{\theta}|\mathrm{D}; \mathrm{M_i}, \mathrm{I}) = \frac{\mathrm{P}(\mathrm{D}|\boldsymbol{\theta}; \mathrm{M_i}, \mathrm{I})\mathrm{P}(\boldsymbol{\theta}; \mathrm{M}, \mathrm{I})}{\mathrm{P}(\mathrm{D}; \mathrm{M_i}, \mathrm{I})} \tag{121}$$

Integrating both sides with respect to $\boldsymbol{\theta}$ and recalling that the posterior probability on the left hand side integrates out to unity gives

$$P(\mathrm{D}; \mathrm{M_i}, \mathrm{I}) = \int \mathrm{P}(\mathrm{D}|\boldsymbol{\theta}; \mathrm{M_i}, \mathrm{I})\mathrm{P}(\boldsymbol{\theta}; \mathrm{M}, \mathrm{I})\mathrm{d}^\mathrm{N}\boldsymbol{\theta} \tag{122}$$

The quantity on the left is called the ’evidence’. Specifically, it is the evidence for model M_i. Consider the expression

$$P(M_i|\mathrm{D},\mathrm{I}) = \frac{\mathrm{P(D|M_i,I)P(M_i|I)}}{\mathrm{P(D|I)}} \tag{123}$$

Now, it is clear that $P(\mathrm{D|M_i,I}) = \mathrm{P(D;M_i,I)}$ is the evidence calculated above. Also, note that

$$\sum_i P(M_i|\mathrm{D},\mathrm{I}) = 1 \tag{124}$$

by definition. Therefore, the denominator $P(\mathrm{D|I})$ can be calculated by imposing this normalization condition. The quanitity $P(M_i|I)$ is the prior probability for our ’faith’ in model M_i, and in principle we can assume equal faith in all models. This implies that the model with the greatest evidence fits the data the best and is selected by it.

62. Random Fields

63. Till now we have talked of only scalar and vector random processes where the outcome of an experiment is either a scalar or a vector. In general random processes could be functions of space as well as time.

64. Let us first consider a random scalar field $\Phi(\mathbf{x}, t)$. What it means is that at a given space point, Φ is a random variable that has different values are different times. And at a fixed time it has different values at different space points.

65. The PDF for such a process is a more complicated object called a probability functional that assigns probability for the process returning a value between $\Phi(\mathbf{x}, t)$ and $\Phi(\mathbf{x}, t) + d\Phi(\mathbf{x}, t)$. As we have learnt, for Gaussian random processes all the information is encoded in the two point correlation function that we had earlier called the covariance tensor; which in this case is defined as

$$\xi(\mathbf{x}, \mathbf{x}', t, t') = \langle \Phi(\mathbf{x}, t)\Phi(\mathbf{x}', t') \rangle \tag{125}$$

66. If the correlation function depends only on the difference $\mathbf{x} - \mathbf{x}'$ then the process is called homogenous, similarly, if the correlation function depends only on the difference in time $t - t'$ then the process is called stationary. For a moment let us imagine a random process that is independent of time. Then for a homogenous process

$$\xi(\mathbf{x} - \mathbf{x}') = \langle \Phi(\mathbf{x})\Phi(\mathbf{x}') \rangle \tag{126}$$

67. Furthermore, if the correlation function depends only on the length $r = |\mathbf{x} - \mathbf{x}'|$ then the process is called isotropic. Thus, for an isotropic process

$$\xi(|\mathbf{x} - \mathbf{x}'|) = \langle \Phi(\mathbf{x})\Phi(\mathbf{x}') \rangle \tag{127}$$

68. <u>Power spectrum:</u> Consider the Fourier transform of $\Phi(\mathbf{x})$

$$\Phi_{\mathbf{k}} = \int_{R^3} \Phi(\mathbf{x}) e^{i\mathbf{k}\cdot\mathbf{x}}\, d^3r \tag{128}$$

and the inverse transform as

$$\Phi(\mathbf{x}) = \frac{1}{(2\pi)^3} \int_{R^3} \Phi_{\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{x}}\, d^3k \tag{129}$$

We should note here that for a homogenous random field, which extends over infinite space, the Fourier transform does not exist. We must assume a finite support for the process. However, taking the volume to be large our formal results, derived below through the Fourier transform, still hold. For a finite volume Fourier integral is replaced with the Fourier series leading to similar results.

69. Let us consider the quantity

$$\begin{aligned}
\langle \Phi_{\mathbf{k}} \Phi^{\star}_{\mathbf{k}'} \rangle &= \int_{R^3} \int_{R^3} \langle \Phi(\mathbf{x}) \Phi(\mathbf{x}') \rangle e^{i(\mathbf{k}\cdot\mathbf{x} - \mathbf{k}'\cdot\mathbf{x}')} d^3r d^3r' \\
&= \int_{R^3} \int_{R^3} \xi(\mathbf{x} - \mathbf{x}') e^{i(\mathbf{k}\cdot\mathbf{x} - \mathbf{k}'\cdot\mathbf{x}')} d^3r d^3r'
\end{aligned}$$

By changing coordinates to $\mathbf{u} = \mathbf{x} - \mathbf{x}'$ and $\mathbf{x}' = \mathbf{x}'$, this can be cast in the form

$$\langle \Phi_{\mathbf{k}} \Phi^{\star}_{\mathbf{k}'} \rangle = \int_{R^3} \int_{R^3} \xi(\mathbf{u}) e^{i[(\mathbf{k}-\mathbf{k}')\cdot\mathbf{x}' + \mathbf{k}\cdot\mathbf{u}]} d^3u d^3r' \tag{130}$$

This change of coordinates has to be considered carefully. We have to ensure that the Jacobian is properly taken into account and the tiling of the coordinate space is done properly. For details look for Faltung theorem in any standard book.

70. Performing the integration over $\mathbf{x}'$ we obtain

$$\langle \Phi_{\mathbf{k}} \Phi^{\star}_{\mathbf{k}'} \rangle = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}') \int_{R^3} \xi(\mathbf{u}) e^{i\mathbf{k}\cdot\mathbf{u}} d^3u = (2\pi)^3 \delta^3(\mathbf{k} - \mathbf{k}')\, P(\mathbf{k}) \tag{131}$$

where $P(\mathbf{k})$ is the power spectrum. For an isotropic process $P(\mathbf{k}) = P(k)$, and is given by the Fourier transform of the two point correlation function $\xi(r)$. This remarkable result is known as Wiener-Khinchin theorem.

71. **An example from cosmology**: The matter density field in the universe $\rho(\mathbf{x}, t)$ is a function of space and time. On average the Universe is homogenous and isotropic and has a mean density

$$\bar{\rho} = \langle \rho(\mathbf{x}) \rangle \tag{132}$$

where we note that the average density is independent of the position. The precise meaning of this statement is in terms of the initial conditions in the universe due to inflationary expansion which leaves the universe smooth but with tiny density fluctuations

$$\delta(\mathbf{x}) = \frac{\rho(\mathbf{x}) - \bar{\rho}}{\bar{\rho}} \tag{133}$$

it is easy to see that $\langle \delta(\mathbf{x}) \rangle = 0$.

72. Models of inflation inform us that the density fluctuations started out as Gaussian random field. The probability distribution function is determined completely by specifying the two point correlation function

$$\xi(r) = \langle \delta(\mathbf{x}_1)\delta(\mathbf{x}_2)\rangle \tag{134}$$

where $r = |\mathbf{x}_1 - \mathbf{x}_2|$

73. Consider two points separated by r and having density contrasts δ_1 and δ_2. We have

$$\langle \delta_1^2\rangle = \langle \delta_2^2\rangle = \xi(0), \quad \langle \delta_1\delta_2\rangle = \xi(r) \tag{135}$$

To obtain the PDF for δ_1 and δ_2, we invert the covariance matrix to obtain the probability distribution function

$$P(\delta_1, \delta_2) = \frac{1}{2\pi\sqrt{\xi(0)^2 - \xi(r)^2}} \exp\left[-\frac{1}{2}\frac{\xi(0)\delta_1^2 + \xi(0)\delta_2^2 - 2\xi(r)\delta_1\delta_2}{\xi(0)^2 - \xi(r)^2}\right]. \tag{136}$$

And the conditional probabilty for δ_1 given δ_2 can be written as

$$P(\delta_1|\delta_2) \propto \exp\left[-\frac{\xi(0)}{2(\xi(0)^2 - \xi(r)^2)}\left(\delta_1 - \frac{\xi(r)}{\xi(0)}\delta_2\right)^2\right]. \tag{137}$$

The mean value of δ_1 has now shifted to

$$\langle \delta_1\rangle = \frac{\xi(r)}{\xi(0)}\delta_2. \tag{138}$$

This result is noted to explain the excess clustering of galaxies and clusters of galaxies since these are likely to form in the overdense regions. And since around a galaxy (large δ value) the mass distribution is biased towards large values, such regions are favorable for the formation of even more galaxies, making the galaxy distribution more clustered then the underlying mass distribution.

Two-point correlator for homogenous, incompressible turbulent velocity field

74. Let us now consider a more complicated process such as turbulence. Turbulence produces a random velocity field $\mathbf{v}(\mathbf{x}, t)$ at every point in space.

75. Consider the tensorial two-point correlation function

$$\Gamma_{ij}(\mathbf{x}, \mathbf{x}', t, t') = \langle v_i(\mathbf{x}, t) v_j(\mathbf{x}', t') \rangle \tag{139}$$

Let us specialize to homogenous and isotropic turbulence. Then

$$\Gamma_{ij}(r) = \langle v_i(\mathbf{x}, t) v_j(\mathbf{x} + \boldsymbol{r}, t) \rangle \tag{140}$$

We shall suppress writing t, t' in expressions since all operations below pertain to space. The only isotropic tensor that we can construct from r_i is

$$\Gamma_{ij}(r) = A(r)\delta_{ij} + B(r) r_i r_j + C(r)\epsilon_{ijk} r_k \tag{141}$$

One can show that this expressions is covariant under spatial rotation. However, the last term flips sign under spatial inversion. Without any loss of generality we can decompose this in terms of longitudinal and normal components

$$\Gamma_{ij}(r) = T_N(r)\left(\delta_{ij} - \frac{r_i r_j}{r^2}\right) + T_L \frac{r_i r_j}{r^2} + C(r)\epsilon_{ijk} r_k \tag{142}$$

76. Incompressibility condition is equivalent to $\partial_i \Gamma_{ij} = 0$, which leads to the following constraint

$$T_N = \frac{r T_L'}{2} + T_L = \frac{r^2 T_L'}{2r} \tag{143}$$

Fourier Techniques

77. Fourier methods are useful in a variety of contexts. The Fourier transform of a function is a linear map

$$F(k) = \int f(x) \mathrm{e}^{\mathrm{i}kx} dx \tag{144}$$

with the inverse transform given by

$$f(x) = \frac{1}{2\pi} \int F(k) \mathrm{e}^{-\mathrm{i}kx} dk \tag{145}$$

78. Fourier transforms can be used effectively to solve partial differential equations. For example, given the wave equation