Core statistics for bioinformatics

Woon Wei Lee

March 12, 2003

Contents

1 Introduction

1.1 What is Bioinformatics?

Bioinformatics is a newly coined term and refers to a novel branch of science straddling the traditional domains of biology and informatics, which is itself a new area of research. Hence, bioinformatics is primarily concerned with the creation and application of information-based methodologies to the analysis of biological data sets and the subsequent exploitation of the information contained therein. The widespread adoption of a range of technologies such as microarrays as well as large scale genome sequencing projects has resulted in a situation where a large amount of data is being generated on a daily basis - too large, in fact, for manual examination and subsequent exploitation. Hence, the development of a range of suitable informatics tools for automated feature extraction and analysis of these data sets is required. The tools provided by bioinformatics are intended to fill this gap.

In addition, biological systems are intrinsically noisy. Fundamentally, biological systems and the processes driving them are "fuzzy" in nature. As a result, any data or observations derived thence will inevitably be equally fuzzy. Due to this inherently noisy nature, the mathematical techniques used to deal with biological datasets must be able to deal with the uncertainty that is invariably present in the data. Statistical methods are the natural solution to this problem.

Hence, it is clear that the effective use of bioinformatics necessitates a sound mastery of the underlying mathematical and in particular statistical principles. This short course has been designed to provide a suitable starting point from which the bioinformatics course may be more effectively attacked. The objective is to introduce all the relevant statistical concepts so that the algorithms and methodologies used in bioinformatics can be more readily understood and more effectively applied.

1.2 The story so far..

At the basic level, statistics is typically taught as a collection of quantities which are calculated based on either the results of an experiment, or on a sample of values taken from a population which is of interest to the researcher. The most common examples are the mean, median and mod of a sample. In one way or another, these three quantities approximate the typical values expected of the data set, though the slight differences in the way in which this is achieved means that different aspects of what is "typical" are emphasised. Other statistics may characterise the spread in values of the elements of the dataset. The most commonly quoted example is the standard deviation (and variance) of the dataset. This is the square root of the mean squared deviation from the sample mean. Other less commonly

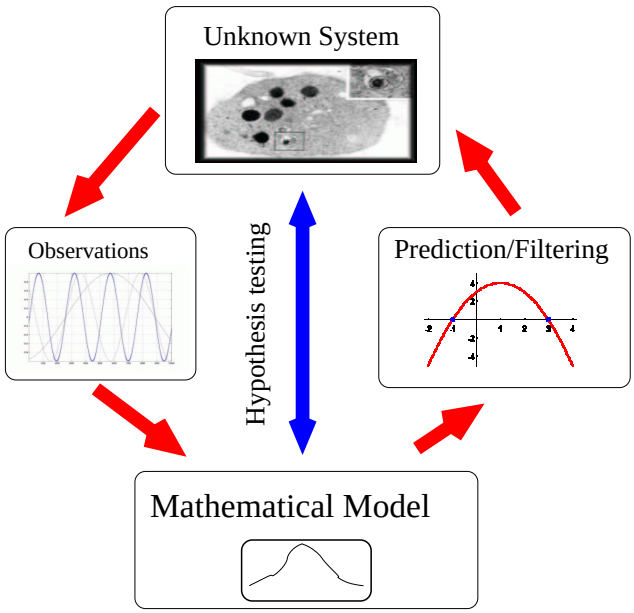

Figure 1: Statistical learning process

used statistics describe higher order properties of the distribution and come with such exotic names as the "skew" or the "kurtosis" of the distribution.

While these statistics provide a convenient means by which a dataset may be easily characterised, their widespread use has obscured a lot of the "meat" associated with the study of statistics. Proper use of statistical theory requires that we approach the subject from a probabilistic perspective, as only then can a more profound understanding and appreciation be gained regarding the data and its underlying causes. Such a firm grounding is certainly essential for successful mastery and exploitation of the many tools offered by bioinformatics.

Roughly, the process by which statistics is used to elucidate an unknown system may be summarised by the graph in figure 1. In general, the system of interest if invariably unknown (otherwise, it wouldn't be very challenging!). However, it is still possible to learn about the system by making indirect, and inevitably noisy observations of its underlying state. The challenge then is to generate a mathematical model which can effectively account for these observations, and there are a number of algorithms by which this can be achieved. Due to the uncertainty in the data, it is imperative that any such model has the capability to deal with uncertainty - hence a probabilistic model suggests itself. Note that the uncertainty in a system can originate from two sources:

1. Uncertainty due to actual random processes affecting the data, such

as mutations in DNA,

2. uncertainty due to incomplete information, where the model must be able to account for our belief in the current state of the data

Once a suitable model has been devised, it is helpful to use *hypothesis tests* to determine the validity of the model, i.e.: its faithfulness to the actual data generator. This is a statistical process and only provides us with a specified degree of confidence in the model - it can never confirm a model with 100% certainty. Finally, and only if the validity of the model can be ascertained with a reasonable degree of certainty, a range of activities can be carried out including prediction, inference, filtering and so on, which allow us to indirectly deduce the state of the system of interest, thus completing the cycle.

1.3 Introduction to random variables and probability distributions

Firstly, we need to make some informal definitions for key phrases which will be used liberally throughout this course.

Random experiment - Experiments for which the outcome cannot be predicted with certainty.

Random variable - The outcome of a random experiment. Conventionally written with uppercase symbols e.g.: X,Y,etc

Discrete random variable - A numerical quantity that randomly assumes a value drawn from a finite set of possible outcomes. For example, the outcome of a dice throw is a discrete random variable with a solution space: {1,2,3,4,5,6}

Continuous random variable - Similar to the discrete case, but this time the solution space consist of a range of possible values, with (in principle) infinite resolution

Probability distribution - This is a function, $P_X(x)$ over the solution space of the random variable, yielding the probability of occurence for each potential outcome. Again this can be differentiated into *discrete* and *continuous* instances. Probability distributions are constrained by the following condition:

$$\int_x P_X(x)dx = 1 \tag{1}$$

For a discrete random variable X, the probability distribution is often represented in the form of a table containing all possible values which the variable can take, accompanied by the corresponding probabilities. In

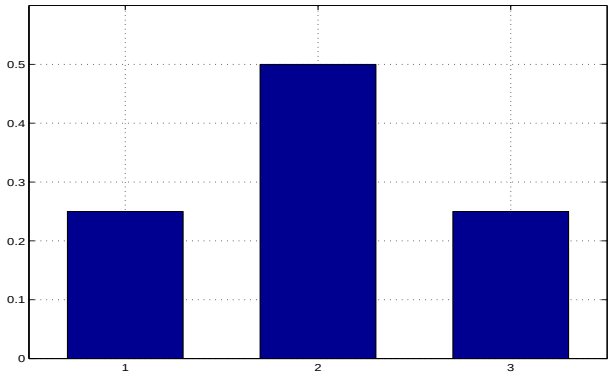

Figure 2: Probability distribution for coin toss experiment

the conventional view, these are interpreted as the relative frequencies of occurence of the various values. In later sections, we will be covering an alternative perspective known as the *Bayesian framework*, in which the probabilities are treated as subjective measures of belief in certain outcomes of the random experiment.

A good example is the result of a coin toss experiment, which is performed by tossing a fair coin twice, and recording the number of heads observed. Assuming this experiment is performed a large number of times, we can expect that the results will occur approximately according to the following frequencies:

Number of heads (X)	Relative frequency
0	0.25
1	0.5
2	0.25

In later sections, we will examine how this can be calculated analytically. Clearly, the most straightforward way in which the probability distribution may be obtained is by repeating an experiment a large number of times, then compiling and tabulating the results. These may then be presented as a histogram depicting the probabilities of each of the outcomes. For our experiment above, an idealised graph is shown in figure 2.

2 Probability distribution functions

A probability distribution function or PDF is simply a function defined over the entire solution space (i.e.: the space of all possible values which the

random variable is able to return) which allows the probability or probability density at each potential solution to be determined analytically.

Many such functions have been proposed, corresponding to a variety of theorised situations. However, in real life experimental conditions such conditions are rarely achieved exactly, which means that the actual distributions from which real-life data is sampled often deviate from these idealised distribution functions. Nevertheless, for practical reasons and mathematical tractability, it is the accepted practice to model real life distributions by *fitting* one of the existing classes of distribution functions to match the data.

We now study some of these functions.

2.1 One Bernoulli Trial

A *Bernoulli trial* is a single trial with two possible outcomes, often called "success" and "failure". The probability of success is denoted by p and the probabilty of failure,q, is simply given by 1-p, since there are no other possible outcomes of the experiment.

Hence, if we label a "success" as a 1 and a "failure" as a 0, we obtain the following formula for the outcome of a bernoulli trial:

$$P_X(x) = p^x(1-p)^{1-x}, x = \{0, 1\} \tag{2}$$

2.2 The Binomial distribution

A Binomial random variable is the number of successes obtained after repeating a given Bernoulli trial n number of times, where the probabilities p and q are fixed for the duration of the experiment. There is also the added condition that the outcome of the successive Bernoulli trials be independent of one another.

For a Binomial random variable X, the probability distribution P_X is given by:

$$P_X(x) =^n C_r p^x(1-p)^{n-x}, x = 1, 2, ..., n \tag{3}$$

where ${}^nC_r$ is the combination operator, which gives the number of ways in which you can select r items from a collection of n. Note that the distribution is described by two *parameters*, n and p, which together determine the characteristics of the resulting distribution function.

In the case where $n = 20$ and $p = 0.5$, the resulting binomial distribution is shown in figure 3.

2.3 The Poisson distribution

One commonly encountered scenario is where the event of interest occurs a finite number of times within a given time interval. Commonly quoted

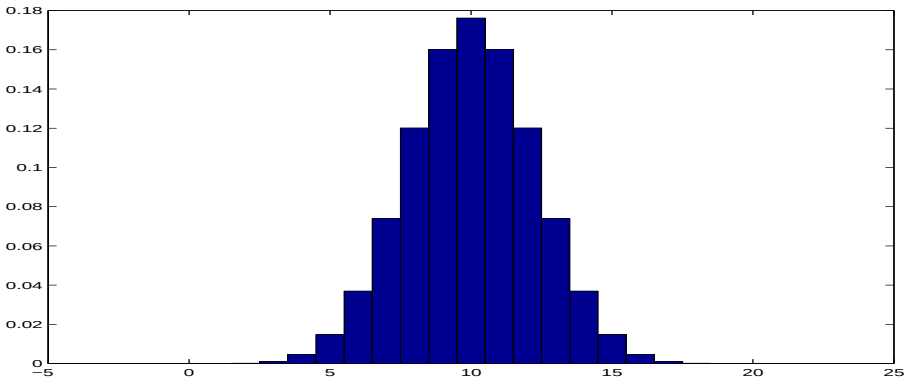

Figure 3: Binomial distribution with $n = 20$ and $p = 0.5$

examples are the number of car accidents during a fixed period, the number of phone calls received, and so on. In such cases, while there is an infinitesimally small probability of the event occurring at a particular time instant, the actual number of time "instants" is extremely large (as the time duration is continuous, the number of instants is effectively infinite). Hence, it is often sufficient just to know the *mean* number of occurrences in a fixed time interval. The probability distribution for the number of occurrences can then be well approximated by the *Poisson distribution*, given by the following function:

$$f(x|\lambda) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & \text{for } \; x = 0, 1, 2, ... \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Where λ is the mean number of occurrences in the time period of interest, and x is the actual number of occurrences. Clearly in this expression, e^{λ} serves as a normalising factor (since it does not depend on x), and the value of the probability is determined by the expression $\frac{\lambda^x}{x!}$.

2.4 The uniform distribution

Perhaps the most straightforward distribution function is the uniform distribution. A random variable is said to have a uniform distribution if the density function is constant over a given range (and zero elsewhere), i.e.: all possible values within the accepted range of values have equal probability. For the range $a \geq x \geq b$, this is expressed as:

$$P(x) = \begin{cases} \frac{1}{a-b} & \text{for } \; a \geq x \geq b \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

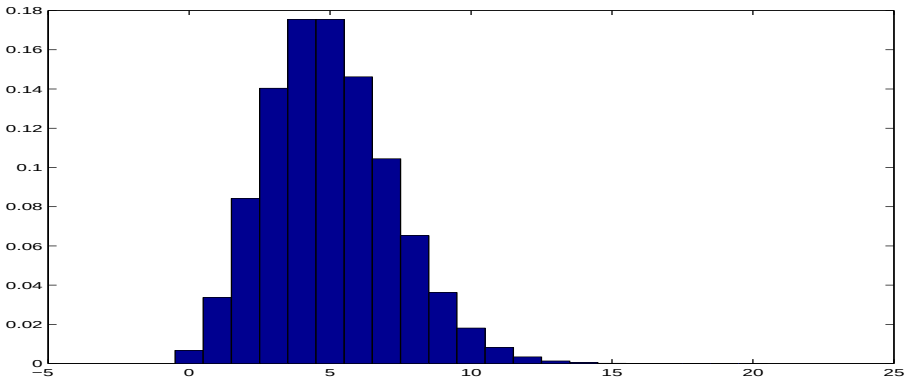

Figure 4: Poisson distribution with $\lambda = 5$

2.5 The normal distribution

By far the most common, and certainly the most important probability distribution that we will be studying is the normal or Gaussian distribution, shown in figure 5. A continuous random variable X drawn from a normal distribution has a density function:

$$P_X(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6}$$

The density function is parameterised by the quantities σ and μ which represent the standard deviation and mean of the distribution respectively.

For a number of reasons, both theoretical and practical, the gaussian is the distribution of choice for many applications. However, two factors in particular account for its pre-eminence:

1. The mathematical properties of the density function for the normal distribution make it extremely easy to work with. The mean and variance of the distribution are immediately evident from the function - as a matter of fact, a gaussian is completely described by the mean and variance. Higher order cumulants of the distribution are zero.

2. Many naturally occurring phemomena often have distributions that are approximately normal. This is a direct consequence of the *central limit theorem* which states that the composite distribution resulting from the combination of a large number of independent random variables will converge to a normal distribution.

2.6 Characteristics of a random variable

The distribution of a random variable X contains all the information regarding the stochastic properties of X. However, in many cases, it is difficult to

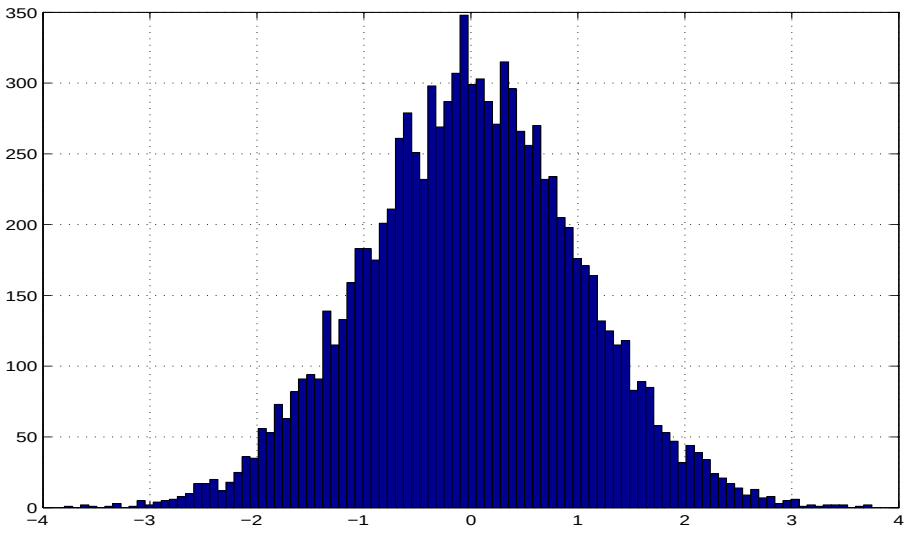

Figure 5: Histograms for samples drawn from a zero mean, unit variance normal distribution

represent this information as the PDFs of real random variables can often be complex and not easily characterised by one of the existing families of analytical distributions. In this section we study ways by which a distribution may be "summarised" to give a general idea of its key properties without having to describe the entire distribution.

2.6.1 Expectation

One of the most commonly invoked quantities is the *expectation* of a random variable. Denoted by $E[x]$, this is defined as:

$$E_X[x] = \int_{x=-\infty}^{\infty} x.P_X(x)dx \tag{7}$$

The number $E[x]$ is also called the *expected* value of X or often simply the *mean* of X. Note that it is analogous to the centre of gravity of a physical object (as taught in earlier tutorials!). Effectively, the mean is the centre of mass of the probability density function.

From this discussion it is evident that the mean should in fact be distinguished from the *arithmetic average* of a sample of points, also known as the sample mean. For a sample size n, this is:

$$\overline{X_n} = \frac{1}{n}(X_1, X_2, ..., X_n) \tag{8}$$

The mean is related to the actual underlying distribution from which the data is sampled, whereas the average is a statistical measure that is derived

from the samples themselves. In fact, it can be proven that if a *collection* of populations were drawn from a given distribution, the averages of the individual populations would themselves be distributed according to a normal distribution, with a mean and variance determined by the number of points in the samples. In fact, the exact values for these parameters may be easily determined thus:

$$\begin{aligned} E[\overline{X_n}] &= E[\frac{1}{n}(x_1 + x_2 + ... + x_n)] \\ &= \frac{1}{n}.nE[X_i] \\ &= \mu \end{aligned} \tag{9}$$

The variance of the sample means can be predicted in a similar fashion thus:

$$\begin{aligned} Var(\overline{X_n}) &= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} x_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \ \ (x\text{'s are independent of one another}) \\ &= \frac{1}{n^2}.n\sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned} \tag{10}$$

What these these two results tell us is that, while the *expected value* of the sample mean will be the mean of the underlying distribution, this is only an estimate and varies with a given variance which is inversely proportional to the sample size (i.e.: the larger the sample size, the more accurate the estimate).

2.6.2 Moments of a distribution

The mean and variance are special cases of the *moments* of a probability distribution.

For a random variable X, the rth moment M_r (where r is any positive integer, is defined as:

$$\begin{aligned} M_r &= E_X[x^r] \\ &= \int_{-\infty}^{\infty} x^r P_X(x) dx \end{aligned} \tag{11}$$

It is also cannot be assumed that a certain moment of a given distribution exists. If a distribution is bounded (i.e.: if the PDF integrates out to one), then it is necessarily true that all moments exist. However, while it is possible for all moments to exist even if the PDF is not bounded, this is not necessarily true. It can be shown that if the rth moment of X exists, then all moments of lower order must also exist.

2.6.3 Moment generating functions

Given the density function, how can find the moments of a distribution? In many cases, this can be obtained directly but often it can be quite challenging. One approach by which a given moment may sometimes be conveniently calculated is via a *moment generating function*.

3 Distribution functions of more than one random variable

It is possible to combine PDFs from separate random variables to form composite distributions. In such cases, it is useful to be able to classify these according to their respective functions. These help to clarify what a distribution function says about a pair (or more) of random variable. In particular, we identify three common classes into which composite PDFs may fall.

3.1 Joint distributions

For this, and all proceeding examples in this section, we will concentrate on the case where there are two random variables, X and Y, which are not necessarily independent. All examples can easily be generalised to the case of multiple random variables.

Consider the case where we sample simultaneously from X and Y, i.e.: we conduct a *joint experiment*. What is the probability of observing a particular pair of outcomes? In this case, we can formulate the answer as a new composite distribution function which extends over the combination of the solution spaces of the two random variables.

To help visualise this, let us assume that X and Y are two discrete random variables with the solution space defined by $X, Y \in \{1, 2, 3, 4, 5\}$. In this case, the possible combinations of values which the joint random variable (X, Y) can assume are shown in figure 6. For each of the points in the grid, we can now assign a probability of the corresponding outcome of the joint experiment. These probability values are denoted by $P(X, Y)$, and are given by the *joint probability distribution* of X and Y.

3.2 Conditional distributions

Suppose that we already know the outcome of experiment Y. Clearly this would greatly limit the number of possible outcomes in the joint solution space. In our current example, since we are only dealing with two variables, this effectively reduces the solution space to one dimensional. It is clear that, for any given value of Y, the corresponding probability for a particular

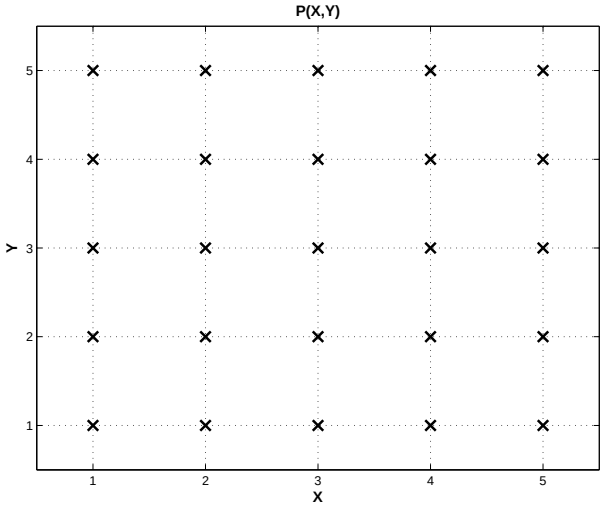

Figure 6: Possible combinations of values for discrete random variables X and Y

value of X can be obtained simply by reading along the particular row corresponding to the incident value of Y.

We call this new distribution the *conditional* distribution of X given Y. Equivalently, it is normal to speak of the probability of X conditional upon a certain value of Y. Mathematically, this is written as $P(X|Y)$, and is derived from:

$$P(X|Y) = \frac{P(X,Y)}{P(Y)} \tag{12}$$

Note how it can easily be seen from figure 6 that this corresponds to the joint distribution values for the required values of X (along the row corresponding to the incident value of Y), normalised by the sum of all the joint probability values along the row.

3.3 Marginal distributions

In the final example, consider the situation where we are not interested in the outcome for experiment Y, i.e.: we are only interested in the outcome of X. For a given $X_n = x$, we can obtain the unconditional probability by summing over $P(X,Y)$ for all the possible values of Y. This is a process called *marginalisation* and is written as follows:

$$P(X) = \sum_{Y} P(X,Y) \tag{13}$$

The resulting distribution, $P(X)$, is then called the *marginal distribution*.

3.4 Independent random variables

Before proceeding further, this is a suitable point for the introduction of the concept of *statistical independence* when applied to random variables. In many cases, "independence" as used in statistics corresponds well with the general meaning of the word, as used in everyday situations. i.e., a given random experiment is independent of another random experiment if the associated random variables do not depend on each other in any way. For example, the result of two successive coin-tosses occur completely randomly and are independent of one another.

However, it is still useful for a formal definition be given. We say that two random variables X and X are considered *statistically independent* if, and only if, the *joint distribution* of the two variables is equal to the product of the two *marginal distributions*, written as:

$$P(X, Y) = P(X)P(Y) \tag{14}$$

In such a case, the grid in figure 6 becomes a multiplication matrix - where the values associated with the vertices can be found from the product of the unconditional probabilities $P(X)$ and $P(Y)$.

4 Estimation theory

So far, we have covered some of the basic concepts of probability which provide the basis upon which the study of statistics is built. In particular, we would like to consider real world data as observations of some underlying generator. As was mentioned earlier in the notes, it is almost always impossible to study this underlying generator directly. However, what is commonly possible is to learn about its properties based on indirect observations.

From the previous sections we have seen that one way in which we can reason about this underlying probability distribution in a sensible way is if we assume some parametric distribution for it. For example, if we want to learn about the distribution of heights in the population of Malaysia, we can assume that it is drawn from a gaussian distribution (and in fact, it does, approximately!). The process by which we learn about the mean and variance of this distribution is a crucial activity in statistics and is widely referred to as *estimation*. As a loose guide, an estimator is some function or algorithm by which the realisations of a random variable are mapped to an estimate of the parameters of the underlying generator. Simply averaging a dataset provides a good example of an estimator that is very commonly used. It can be shown that the arithmetical average of a set of data provides an unbiased estimate of the expectation of the underlying distribution from which the data was drawn.

4.1 Maximum likelihood estimation

Broadly speaking, there are two approaches to statistics which, while actually sharing a lot of common ground, are widely regarded as being from opposing camps. One on hand, there is the "Frequentist" position, and on the other we have the Bayesian framework.

One popular method taken from the frequentist camp, is that maximum likelihood estimation. This is the procedure for estimating the parameters of the unknown model, by maximising the likelihood of the observed data. That is to say we would like to find:

$$\theta_{ML} = \text{argmax}_\theta \left[P(Y|\theta) \right] \tag{15}$$

Here, θ represents the parameters of the model which we would like to estimate, whereas Y denotes the available observations.

4.1.1 Example: linear regression

Suppose we have a set of paired values, x and y, which we assume are linearly correlated. Accordingly, we assume that the two are related by the expression $y = Mx$. Hence, we would like to estimate the value of the parameter M, which in this case is the gradient of the line obtained by plotting x vs y on an $x - y$ plane. Finally, to obtain a maximum likelihood solution, we also need to assume some kind of noise model. This is necessary because real data is never exact - otherwise, we can obtain M simply by evaluating:

$$M = \frac{y_1 - y_2}{x_1 - x_2}$$

(which wouldn't be very interesting!). A commonly used assumption is that of gaussian noise. That is to say:

$$y = Mx + \nu \tag{16}$$

where $\nu \sim \text{N}(\mu, \sigma)$. Hence, the distribution of y conditional upon x is given by:

$$\begin{aligned} P(y|x) &\sim N(Mx, \sigma) \\ &\propto exp\left[\left(\frac{y - Mx}{\sigma} \right)^2 \right] \end{aligned} \tag{17}$$

To simplify the maximisation of the likelihood, we now take the *logarithm* of the expression above. Note that this is acceptable since logarithm is a monotonic function - i.e.: it only increases in one direction, such that $log(x_1) > log(x_2)$ necessarily implies that $x_1 > x_2$. Taking the logarithm of $P(y|x)$ yields the log-likelihood term:

$$-\mathbf{log} P(y|x) = -\left(\frac{y - Mx}{\sigma} \right)^2$$

Note that we take the *negative* log likelihood - the reason for this will become clear shortly. We can now easily differentiate this with respect to M, and set to zero, to obtain:

$$\begin{aligned} \frac{d[-\mathbf{log}P(y|x)]}{dM} &= \frac{2}{\sigma^2}(y - Mx).x = 0 \\ \therefore \quad & x^T y = x^T x M \\ \therefore \quad & M = (x^T x)^{-1} x^T y \end{aligned} \tag{18}$$

Note the final left hand expression, $(x^T x)^{-1} x^T$. This is called the *pseudo-inverse* of x, and is commonly denoted as $x^\dagger$. The evaluation of the pseudo-inverse of a matrix is a common function which is widely available in statistics/mathematical packages - enabling maximum likelihood fitting of this sort to be performed with great ease. Nevertheless, it is useful and conceptually important to be aware of the underlying model - i.e.: that linear regression is actually equivalent to fitting a linear gaussian noise observation model to the data.

4.2 Bayesian framework

The maximum likelihood method discussed above has proved very useful for many applications. However, it also has some shortcomings. In particular, by maximising over the parameter space, it is discarding all the possible model parameters in favour of one "optimal" solution. While this is a practical strategy in many instances it is also sensitive to the shape of the likelihood function. Take, for example, the likelihood function depicted in figure 7. This is an example of a bimodal distribution - in fact a mixture of two gaussian distributions. However, one of the density functions has a much smaller variance and as such is a lot more peaked. In fact however, the probability mass of the first distribution is only half that of the flatter distribution. This means that, while the maximum likelihood solution will be the peak of the first distribution, it is far more likelihood that the "optimal" parameters will lie somewhere in the region defined by the second distribution.

The Bayesian framework helps to overcome this problem by attempting to consider the entire PDF of the solution space, rather than just the mode. It is based on Bayes' theorem, which is given by:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{19}$$

What this provides, in very general terms, is a means by which the conditional probability of a given model, X, given the available data, can be linked to the conditional probability of observing the data if the model were correct. In practice, the significance of this is that it gives a broad relationship for estimating the parameters θ of a proposed model provided based on

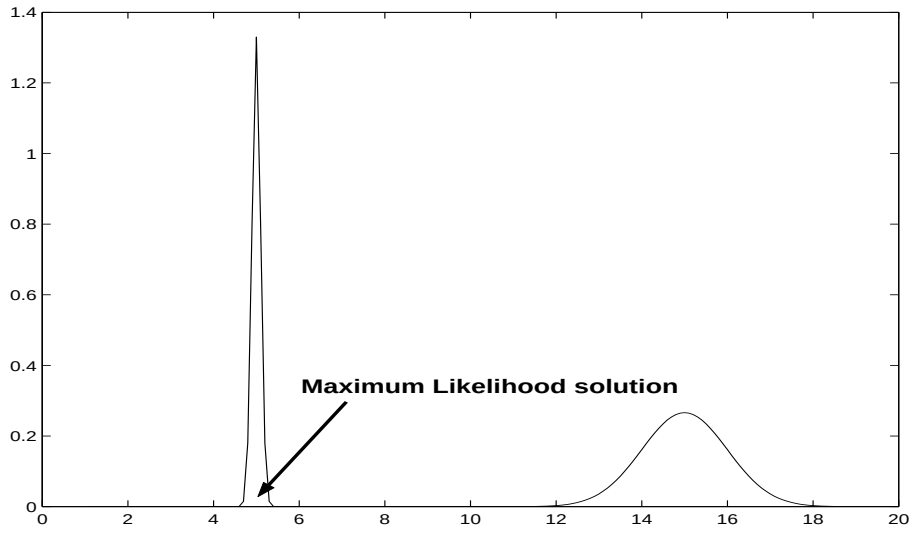

Figure 7: PDF consisting of a mixture of two gaussian distributions with $\mu_1 = 5$ and $\mu_2 = 15$ and $\sigma_1 = 0.1$ and $\sigma_2 = 1$ respectively

observations derived from the true model. This is because it is much easier to estimate $P(Y|X)$ than the other way around.

In Bayesian terminology, the terms in equation 19 are often referred to as follows:

1. In common with frequentist terminology, $P(Y|X)$ is called the likelihood function. This is basically the probability of observing the experimental data, given the proposed model,

2. $P(X)$ is the prior distribution for the model. This terms allows any prior information regarding the model parameters to be incorporated into the inference. If no prior information is available, a suitable "initial guess" can be provided,

3. $P(X|Y)$ is the posterior distribution. This reflects the knowledge we have regarding the model after having incorporated information contained in the observations,

4. finally, the $P(Y)$ in the denominator on the right hand side is called the evidence term. This is obtained by marginalising out X - and is thus constant for all values of X. Hence, its main function is as a normalising term.

The key concept in Bayesian analysis is that the distributions described above are constantly updated, to reflect the increase in our knowledge about the system being studied as new observations become available (it is also

possible that very noisy observations will actually increase the degree of uncertainty by "spreading" the distributions). However, in general, the process of updating the distributions accompanies an increase in our knowledge of the system.

The general process of Bayesian learning is as follows:

1. Start by making an initial guess on the state of the system. This is the prior distribution $P(X)$,

2. when we receive (or make) an observation, Y, we can calculate the likelihood $P(Y|X)$ using the model X that we have assumed,

3. the combination of prior and likelihood can then be used to calculate the updated distribution for X, i.e.: the posterior distribution $P(X|Y)$,

4. the whole process can be repeated whenever a new observation (or set of observations) is obtained. However, at each step, the posterior of the previous step is used as the new prior distribution.

In this way, the Bayesian methodology also carries certain philosophical implications as well. In particular, it describes a systematic framework in which we may explicitly specify our belief in the parameters of a model, and a procedure through which this belief may be updated by comparison with observed data.

5 Markovian dynamics

Hitherto, we have looked at probability distributions that do not change in time. The models that have been examined in the preceeding sections specify a static density function over the solution space, and it is assumed that observations may be made indefinitely without changing the probabilistic structure of the data.

In this section we introduce a modelling paradigm which allows for changes in the statistical properties of the data over time. Such dynamic models allow a much more general range of phenomena to be modelled.

5.1 Dynamical processes

A dynamical process is basically one which changes over time. Essentially, these processes are regarded as being composed of an underlying "state" which evolves in time according to some dynamic evolution rule, often containing stochastic components. This is illustrated in figure 8, where x_t represents the state of the system at time t, and y_t the observation (also at time t).

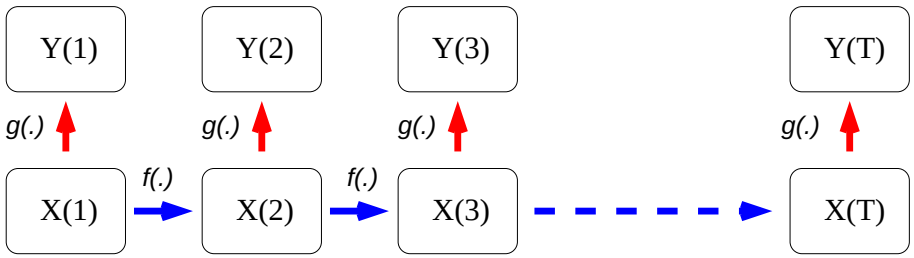

Figure 8: Block diagram depicting the evolution of a generic dynamical process through time

5.2 Markov processes

The key elements in this model are the two functions $f(.)$ and $g(.)$. The function $f(.)$ is known as the transition or evolution function and determines how the system changes over time. In general we would like to model cases where the following two relationships hold true:

$$x_t = f(x_{t-1}) , \quad \text{and} \tag{20}$$

$$y_t = g(y_t) \tag{21}$$

Equation 20 is of particular significance as it indicates that the state of the system at time t is dependent only on the state of the system at time $t-1$. This is known as the Markov property and any system in which this applies is a Markov process. Equation 21 defines the relationship between the state of the system and the observations generated from it. Again, note that the observations at time t only depend on the state of the system at time t.