The Randomness of the Real World

In the real-world, there are many random (formally, stochastic) systems changing over time which can be modelled very accurately through sequences of dependent random variables than independent variables. For instance, simply consider the outdoor atmospheric temperature on successive days or the value of Microsoft share at the conclusion of succeeding trading days. For these systems, it is not implausible to assume the probability of transitioning from one state to another is dependent solely upon the system's present state, and thus is uninfluenced by extra data relating to previous states. The statistical model which incorporates this attribute is known as the Markov chain. The ideas of state and state transition constitute the foundation of analysis for Markov chain models, and are particularly valuable in the exploration of a wide range of real-world applications.

They are named for Andrei Andreyevich Markov, a prominent Russian mathematician, who first developed the theoretical basis for the statistical model to help mathematically explore the alternation of consonant and vowel letters in Alexander Sergeyevich Pushkin's verse novel "EugineOnegin". His mathematical effort significantly helped in ushering stochastic process (formally, they are a collection of random variables, dependent upon some parameter, usually time) theory into mathematical consciousness.

The Markov chain's defining feature is that the process' recall (formally, memory) reaches back only to its most immediate state. Information relating to system's present state alone is enough to describe the process' forthcoming evolution. This statistical model is the most elementary model for stochastic systems dynamically evolving over time where the system's succeeding states are not independently related. However, even though this model significantly simplifies reality, it is these simplifying assumptions that allow practical real-world problems to be mathematically examined in an analytical way. Stochastic processes, particularly Markov chains, have a wide range of application in a number of different fields ranging from the natural sciences, social sciences (particularly economics), engineering and computer science, operations research all the way to sport.

In my research on Markov chain theory however, the majority of the emphasis was placed on the application of theory to predict the future behaviour of the system under exploration. However, the important empirically-based mathematical procedure of how to impose the Markov chain model on real-life data for a practical application was important never considered. Therefore my focused research question into helping resolve this is:

"How can the Markov chain model be imposed on data so that a real-world stochastic process can be mathematically modelled?"

In my IB Mathematics HL studies, I find that there is not much of an emphasis on stochastic processes and the application of mathematical theory to modelling them. This extended essay and my research questions helped in a significantly way in filling this void.

2. Mathematics of Markov

Before the research question can be directly answered, however, we need to introduce the mathematical concepts and notation underpinning stochastic processes and Markov chains that will be used for the remainder of the essay.

2.1. Stochastic Processes:

(The reference information upon which I have used as the basis for the material in this section comes from [1, p1], but the examples are my own).

The term stochastic process is used to denote a collection of random variables, dependent upon some parameter, (usually time or space), where belongs to an index set . Given a certain value of the parameter , the corresponding values that can take are then known as the states of the stochastic process at . The set of all these possible states is known as the state space.

Depending on'scardinality , a stochastic process can be classified as either continuous or discrete parameter. If is uncountable , then it is a continuous-parameter stochastic process. Examples of this include sequences of event that occur randomly over time, such as the number of hospital admissions. This can be modelled as a continuous parameter stochastic process where the random variable represents the number of admission after hours. The corresponding state space of is or equivalently.

If is finite or countablyinfinite , then it is a discrete parameter stochastic process. As an example of this, consider an elementary simulation of tossing a fair coin, whereby heads represents a win and tails a loss. This can be modelled as a discrete parameter stochastic process, where is the random variable denoting the outcome of thetoss of the coin. The corresponding state space of is thus , where denotes tails and denotes heads.

A stochastic process can be further classified as continuous or discrete space, depending on 's cardinality. If is uncountable, then it is continuous-space, and if is finite or countably infinite, it is discrete-space.

For this extended essay, only one type of stochastic process will be considered in detail. The restrictions in this case are that firstly the stochastic process is discrete parameter, with the parameter being time, and secondly that it is discrete space. Therefore the stochastic processes considered will be of the following form:

, with the corresponding state space taken to be , 2.1 Markov Chains:

(The reference information upon which I have used as the basis for the material in this section comes from [2, pg)

We can now formally define a Markov chain as follows:

A stochastic process with state-space is defined as

a discrete-time Markov Chain, if for each time point and for all the possible states:

This equation can be more clearly interpreted as saying that the conditional distribution of any future state , given that the present state is and has been reached via states , is dependent only upon the present state . This is also known as the discrete time Markovian property.

Time-homogenous Markov chains have the property that the transition probability,

, ,

does not depend on the time-parameter .

Therefore, for all :

.

The other conditions thatmust satisfy are:

These probabilities represent the one-step transition probabilities and are equal for all . Now, if is finite, and the Markov chain is time-homogenous, then the set of one-step transition probabilities can be represented in terms of a transition matrix, , where the element in the row and column of this matrix respectively represents the probability of the system going from state to in one-step:

The process also needs an initial probability distribution to specify the initial state. This is expressed in vector form as follows:

The conditions that must be satisfied are:

Both the transition probability matrix and the initial probability distribution comprise fully the Markov chain model.

3. Connection to the Real-World

Now that the relevant concepts and notation behind what constitutes a discrete-time homogenous Markov chain has been covered, we are now in a position to link this probabilistic model with relevant real life situations. However the aforementioned limitations regarding modelling random processes through Markov chains is that it represents a simplification, because it relies on the assumption that probability of moving from one state to another is dependent solely on the present.

In some particular instances though, the assumption of the Markovian property holds perfectly. As a case in point, consider the trivial example of the game Snakes and Ladders. In this game, the probability of reaching another position clearly depends upon only the player’s current position and the outcome of the next die toss, and not on how that position was reached. A more generalised and applicable example however relates to that of the random walk, which can be most simply considered to be a process which takes successive random steps.

For this extended essay, the real-life application I will focus on is where the Markovian property does not intrinsically apply. Rather it is assumed, and the data is used to provide estimates for different quantities, such as initial probability distributions and transition probabilities. In real-life, however, many other factors are likely to be at play which cannot be considered because the assumption of the discrete time Markovian property is restrictive. But given the complex nature of the real-world, simplifications must be made so that it is mathematically workable. This allows for a variety of calculations and observations to be made that can then be used to predict and better understand the random process.

The application I focus on includes the modelling of rainfall patterns. This will be considered after the section relating to how transition probabilities can be estimated from data.

4. Estimation of transition probabilities:

In this section, the first implied part of the research question will be answered, namely how to estimate the various probabilities of transition in stochastic process. This is arguably the most important steps in the construction of a discrete time Markov chain model for a certain application, because is it here that an empirical basis is provided for the model.

To do this, I will derive a mathematical formula that will provide a method of estimating the transition probabilities that best fit the Markov Chain model imposed on the data used. Finding appropriate literature at a suitable level was particularly difficult. Either the mathematics was much too demanding or the result was stated as a theorem without explication. However, two particular sources, [3] and [4], presented as particularly useful on this topic. In [3], the problem regarding how to best estimate the transition probabilities was presented as an exercise. Preliminary definitions and hints on the appropriate method and the theorem itself were given to the reader, but the actual substantive proof was left for the reader to undertake. In [4], a very brief outline of the proof was provided, but several non-trivial intermediate steps were omitted as assumed. This can be attributed to the fact the book is targeted at upper-level undergraduate students and first-year graduate students.

For this extended essay, I have written my own solution out to the problem presented [3], using [4] as a reference, and with all intermediate steps included and fully explained. The definitions and notation in [3] and [4] have been changed to maintain consistency with previous sections.

Maximum likelihood estimation of transition probabilities:

The method suggested in [3] and outlined in [4] was the maximum likelihood estimation method. A brief overview will be given before it is applied to the Markov chain model

Essentially maximum likelihood estimation is a mathematical method for finding the optimal fit of a statistical model to the given data through estimating the parameters which maximise the probability of obtaining the given data, given the assumed model. This probability is also known as the likelihood function and is expressed as a function of the parameters of the statistical model. When the likelihood function is a maximum, the values of the parameter at that point are known the maximum likelihood estimates.

Application to Markov chains

The first step is to assume that the stochastic process under consideration is determined by a Markov chain. Then, having defined and specified the states,, we have the Markov chain, , where the number of states is known, but the transition probability matrix: , and initial probability distribution: is not.

Now having empirically observed the Markov chain during the times , we wish to estimate , the transition probabilities. However, it is not possible to accurately estimate, because we have only observed once.

If the sequence of statesis observed, then we can write the likelihood function as:

.

Now, in order to proceed we must manipulate and simply this expression by utilising the Markovian property. This is a necessary step because what we are eventually trying to do is maximise the likelihood function, given the key assumptions of a discrete time Markov chain model. This will then allow us to find out the formula for the values of the transition probabilities that result in the likelihood function being a maximum.

To simplify, we first make use of the formula for conditional probability, which is:.

Using this formula and letting and , we can rewrite as:

From the assumption of the Markovian property, therefore we can make the simplification:.

Thus, we can rewrite the above equation:

Again using the formula for conditional probability and the Markovian property, and letting , we can further simplify the above:

Successively iterating as above, we derive:

We now take the logarithm of the likelihood function expressed in this form. The reason for this is that the logarithmic form is often mathematically more treatable, particularly when optimisation methods in calculus are applied. Maximising an expression which is a product is often much more difficult than maximising an expression which is a sum. Since however both and reach their maximum at the same point, maximising either will give the maximum likelihood estimator of , i.e. the equation for .

Therefore taking logarithms of both sides:

In order to proceed, we need to express in terms of transition probabilities:

where counts the number of transitions from state to state that are observed in a set of data. In the above, the term simply represents an indicator function, which assigns the value one for every transition from to and zero otherwise (thus counting the transitions). Therefore:

It is at this point that we are actually ready to proceed to the most important step, i.e., the optimisation step. This is the step where we maximise the likelihood function in order to find out the corresponding formula for that best fits the Markov chain model that is imposed on the data. Note that the “caret” symbol here represents the fact that is the maximum likelihood estimate of .

Thus, formally setting out the optimisation problem:

,

The above constraint is a consequence of the Markov chain model assumption (refer to Section 2.2). To maximise this function, the hint given in [3] to use the method of Lagrange multipliers is taken.

The reason the method of Lagrange multipliers is suggested is because it allows for the maxima of a function to be calculated, while also accounting for any constraints that might be present. Therefore, we need to write the optimisation problem out, but this time in its Lagrangian form as follows, where is the Lagrange multiplier:

To maximise this, we first need to take the partial derivative of , but with respect to :

By setting the resulting expression to equal zero, we can thus find out the values of for which the likelihood function is a maximum:

, giving the formula for as .

Now, from the constraint mentioned earlier (), we find that the value of the Langrage multiplier when the likelihood function is a maximum is:

, yielding .

Therefore, substituting in into the equation , we get the following expression for the maximum likelihood estimate for as:

Application to Rainfall Data:

In this section, I will apply the mathematical theory derived above to a real-life application (predicting rainfall patterns). The data for this analysis will be Sydney's rainfall data for May 2010 and is attached in the index.

Firstly, the states of the Markov chain need to be specified. For this model, let state:

represent rainfall

Therefore , and the parameter is time (days), where. Now over the 31 days, the following sequence was observed:

The likelihood function of this sequence, given the assumptions of a Markov chain model is:

Taking the logarithm of both sides gives the log-likelihood function:

To formulate the Markov chain model, we need to calculate the maximum likelihood estimate for the transition probabilities.

These are the values for when the log-likelihood function is a maximum. From Section, we know that this is when

Thus to proceed we need the values for and :

Now, we can calculate the transition probabilities as follows:

The one-step transition probability matrix for this random process is thus:

To complete this Markov chain model, we need the initial probability distribution of its

initial state . However in the context of this example, the initial probability distribution is not particularly important or significant because the weather is a continual process; the weather cannot simply start anew. There is always a previous state that needs to be considered. But for the sake of completeness, assuming that the weather on a current day is sunny, we can write the initial probability distribution simply as:

.

Generalizing the problem to higher-order Markov chains:

Although the Markovian property is a powerful simplifying function that allows real-life data to be mathematically modelled, the assumption that the memory of the system goes back only one time-unit is perhaps too simplistic in many real-life systems. Even in the case of rainfall, there are a variety of factors that would mean the system's memory extends back further than simply one day. For example, the existence of cold fronts and other weather phenomena could mean that the actual probability of having certain combinations of states is higher than would be predicted by a Markov chain model. To account for the possibility of such factors, we can extend the problem by considering how a order Markov chain model can be imposed on real-world data. The here refers to how many time-units the memory of the system goes back. Thus a 2nd order Markov chain as applied to the rainfall pattern application of would mean that the memory of the weather system goes back two days. Mathematically, we can express the order Markovian property as:

In this section, I will now consider the maximum likelihood estimation method to derive a formula for the transition probabilities that maximises the likelihood function of the observed sequence for when a order Markov chain is assumed:

As previously, the states of the order Markov chain are defined, but the transition probability matrix , and initial probability distribution: are not known

Having empirically observed the Markov chain during the times , we wish to estimate the transition probabilities of .

If the sequence of statesis observed, then we can write the likelihood function is also:

To simplify, we again use of the formula for conditional probability, which is:.

Using this formula and letting and , we can rewrite as:

From the assumption of the order Markovian property, therefore we can make the simplification:

Thus, we can rewrite the above equation as:

Again using the formula for conditional probability and the Markovian property, and letting , we can further simplify the above:

Successively iterating as above, we derive:

Taking logarithms of both sides:

In order to proceed, we now express in terms of transition probabilities:

where counts the number of transitions from state to states that are observed in a set of data. In the above, the term is an indicator function, assigning the value one for every transition from to and zero otherwise. Therefore:

To proceed, we formally set out the optimisation problem as follows:

Expressing this in Lagrangian form:

We now take the partial derivative of , with respect to :

Setting the resulting expression to equal zero, we can now find the value of for which the likelihood function is a maximum:

, giving the formula for as:.

From the constraint mentioned earlier , we find that the value of the Langrage multiplier when the likelihood function is a maximum is:

, yielding .

Therefore, substituting in into the equation, we get the following expression for the maximum likelihood estimate of :

.

Application to Rainfall Data:

In this section, I will apply the above theory to the rainfall data, but severely limit the number of states because of the increased complexity.

Firstly, the states of the Markov chain need to be specified. For this model, let state:

represent

Therefore , and the parameter is time (days), where. Now over the 31 days, the following sequence was observed:

The likelihood function of this sequence, given the assumptions of a Markov chain model is:

Taking the logarithm of both sides gives the log-likelihood function:

To formulate the Markov chain model, we need to calculate the maximum likelihood estimate for the transition probabilities:

These values correspond to when the log-likelihood function is a maximum. From, we know that when

Now, from the data, we have:

and:

Therefore, we can calculate the transition probabilities as follows:

To complete the 2nd-order Markov chain model, we provide the initial probability distribution as follows:

Conclusion and evaluation:

In this essay, I have answered the research question in both a theoretical and an empirical sense. In the theoretical sense, statistical theory was developed that allowed a determination of how to a Markov chain model can be imposed on real-world data in a general sense. This was achieved using the maximum likelihood estimation method to derive the formula for the maximum likelihood estimate of probabilities for both a Markov chain model, and the generalised order Markov chain model. The formulas are respectively and .

Maximising the likelihood function represented the most significant challenge encountered because it was both multivariate and incorporated constraints. However this was achieved through the optimization method of Lagrange multipliers.

In the empirical sense, I showed how to use the mathematical theory derived in actually applying it to real-world data. Imposing both the first-order Markov chain model and order Markov chain model allowed me to notice the strengths and weaknesses of both methods. The first-order Markov chain model had the advantage of simpler mathematically to implement, as only one-step transition probabilities are necessary to be considered. For the second order Markov chain model two-step transition probabilities have to be considered. Thus for the second-order Markov chain in Section 112 , even when only two states were considered, 8 two-step transition probabilities had to be calculated. The complexity increases considerably as and increase. Both methods however suffer when there is a limited data range, as with the rainfall pattern application in this essay. A small sample can be considerably biased, leading to inaccurate estimates of both and. Overall though

There are several unresolved questions however. Firstly is the maximum likelihood estimation method the best method available? From my research no other method was encountered, but this remains a valid door for enquiry. Secondly, how do we estimate the initial probability vector for applications where this is important? The rainfall pattern is an example of where finding this would be trivial, but in applications such as sport and language modelling, this is more important. But because the initial distribution is observed once, not much information can be derived. [7, p60] however provides two possible solutions, but they considerably outside this essay’s scope. Lastly the problem of which