1.010 Uncertainty in Engineering
Fall 2008

Application Example 5

(Bernoulli Trial Sequence and Dependence in Binary Time Series)

IS THE SERIES OF RAINY/NON-RAINY DAYS A BERNOULLI TRIAL SEQUENCE?

Note: The shaded text in this note involves the concept of correlation function for a random sequence. This concept will be encountered later in the course. In a fist reading, you may skip that text.

The concept of dependent random variables finds an important application in so-called time series, which are models of how a random quantity varies in time. In the case when time is discrete or discretized, as for example happens when one considers variables with daily, monthly or annual values (daily close of the stock market, monthly average temperatures, annual sales, score of i^{th} baseball game, etc.), the time series is simply a discrete sequence of random variables X_i. An important issue in modeling such sequences is the probabilistic dependence among different variables.

Here we illustrate the concepts of dependence in time series by considering the simplest case, which is a series of indicator variables $\{I_i, i = 0, \pm 1, \pm 2, ...\}$. Such variables can be used to indicate whether or not an event of interest occurs ($I_i = 1$) or does not occur ($I_i = 0$) at “time” i. For example, we can take the event of interest to be the fact that day i is rainy. Again to keep the illustration simple, we consider the case when the sequence of rainy/non-rainy days is stationary. Stationarity means that the sequence has everywhere the same statistical properties. Hence, the probability $P[I_i = 1]$ does not depend on i, the probability $P[(I_i = 1) \cap (I_j = 1)]$ depends only on the separating distance $|i - j|$, and so on. The assumption of stationarity is realistic in many cases and greatly simplifies the characterization of a time series.

An implication of stationarity for a random sequence $\{I_i, i = 0, \pm 1, \pm 2, ...\}$ is that the correlation function ρ_{ij} depends only on the time lag $|i-j|$.

The joint distribution of the daily rain/no-rain indicators I_i and I_j has 4 probability masses:

- a probability mass p_{00} at (0,0), which gives the probability that both days are dry,
- a probability mass p_{11} at (1,1), which gives the probability that both days are wet,
- probability masses p_{01} and p_{10} at (0,1) and (1,0), which give the probability that day i is dry and day j is wet and, viceversa, that day i is wet and day j is dry.

These four probabilities must add to unity. Moreover, due to stationarity, $p_{01} = p_{10}$, meaning that the relative frequency of the two "transitions" (dry → wet and wet → dry) must be the same. This means that the joint distribution of I_i and I_j is completely specified by just two probabilities, which we take to be p_{00} and p_{11}. The other probabilities must be $p_{01} = p_{10} = (1 - p_{00} - p_{11})/2$.

Since p_{00} and p_{11} determine the joint distribution of I_i and I_j, they also determine the marginal distribution of I_i (which, due to stationarity, does not depend on i) and the conditional distributions of $(I_i|I_j)$ and $(I_j|I_i)$ (which, due to stationarity, are the same and depend only on the time lag |i - j|). In fact,

$$\begin{array}{ll} P[I_i = 0] = p_0 = p_{00} + p_{01} & P[I_i = 1] = p_1 = 1 - p_0 = p_{11} + p_{01} \\ P[I_i = 0|\, I_j = 0] = p_{00}/(\, p_{00} + p_{01}) & P[I_i = 1|\, I_j = 0] = 1 - P[I_i = 0|\, I_j = 0] \qquad (1) \\ P[I_i = 0|\, I_j = 1] = p_{01}/(\, p_{11} + p_{01}) & P[I_i = 1|\, I_j = 1] = 1 - P[I_i = 0|\, I_j = 1] \end{array}$$

Notice the following:

- The marginal probabilities p_0 and $p_1 = 1 - p_0$ describe the overall dryness/wetness of the region. In fact, these probabilities give the long-term fraction of days when it rains or does not rain. However, they say nothing about the pattern of rain. For example, for a given fraction p_1 of rainy days, such days might in some regions come as long runs of uninterrupted rain followed by long dry spells, whereas in other regions dry and wet day may be much more "mixed". The following are examples of rain/no-rain patterns with the same values of p_0 and p_1 but very different "clustering characteristics":

50-day record from Climate 1:

00000000011111110000000000000000000111000000000000

(2)

50-day record from Climate 2:

00010001100000000100100000010000000110001001000000

In both cases, there are 10 rainy days out of 50 and reasonable estimates of p_0 and p_1 are p_0 = 40/50 = 0.8 and and p_1 = 10/50 = 0.2.

- To understand how the differences in the patterns of Eq. 2 are reflected in our indicator model, consider the joint distribution of rain/no-rain in two consecutive days, i and j = i + 1. We estimate the probabilities p_{00} and p_{11} of the joint distribution as the relative frequencies in the samples of pairs of consecutive dry (00) and consecutive wet (11) days, respectively. This gives:

$$\text{for Climate 1: } p_{00} = 37/49 = 0.7551, \qquad p_{11} = 8/49 = 0.1633$$
$$p_{01} = p_{10} = (1 - p_{00} - p_{11})/2 = 0.0408$$

(3)

$$\text{for Climate 2: } p_{00} = 31/49 = 0.6327, \qquad p_{11} = 2/49 = 0.0407$$
$$p_{01} = p_{10} = (1 - p_{00} - p_{11})/2 = 0.1633$$

Notice that the marginal probabilities p_0 and p_1, obtained using Eq. 1, are the same in the two cases. What is different is that these probabilities are contributed in different amounts by the "diagonal terms" p_{00} and p_{11} and the "off-diagonal terms" p_{01} and p_{10}: in Climate 1, the off-diagonal terms are nearly zero, meaning that the probability of the weather changing from one day to the next is very small (this is consistent with the observed long spells of wet and dry periods), whereas in Climate 2 the probability of a weather change is much higher.

Limiting cases with respect to weather variability are joint distributions of the following types:

- extreme permanence of weather conditions: $p_{11} = 1 - p_{00}$ and $p_{01} = p_{10} = 0$. In this case, extremely long dry periods alternate with extremely long wet periods. The ratio between the average lengths of dry and wet periods equals p_{00}/p_{11}.

- extreme variability of weather conditions: either p_{11} or p_{00} or both are zero and $p_{01} = p_{10} \neq 0$. For example, suppose that $p_{11} = p_{00} = 0$ and $p_{01} = p_{10} = 0.5$. This means that weather alternates in a perfectly predictable manner between rainy and non-rainy days, as 01010101...

In both cases, the weather can be deterministically predicted from one day to the next. In fact, the conditional distribution of $(I_{i+1}|I_i)$ has in both cases a unit mass at 0 or at 1. This is consistent with the intuitive notion that, in the first case, the weather tomorrow is with probability 1 the same as the weather today and, in the second case, the weather tomorrow is with probability one opposite to the weather today. An intermediate case between these two extremes is when the weather in different days is independent. The condition of independence is that

$$p_{ij} = p_i \, p_j, \quad \text{for all } i,j = 0,1 \qquad (4)$$

In this case of independence, it is easy to verify that the conditional distributions are identical to the marginal distribution. In a sense, this is the case of maximum prediction uncertainty, because information from weather in the past is useless to predict future weather conditions.

The previous examples illustrate the limitation of providing only the marginal distribution of variables that form dependent time series - or for that matter any set of dependent variables - and the additional information provided by the joint distributions. Now we turn to the calculation of first and second-moment properties, i.e. mean values, variances, and covariance and correlation functions.

The mean value and variance of I_i can be calculated from the marginal distribution and are:

$$m = (0)P[I_i = 1] + (1)P[I_i = 1] = P[I_i = 1] = p_1$$

$$\sigma^2 = E[I_i^2] - m^2 = p_1 - p_1^2 \qquad (5)$$

The covariance may be found using the relation $\mathrm{Cov}[I_i,I_j] = E[I_iI_j] - E[I_i]\, E[I_j]$. This gives

$$\mathrm{Cov}[I_i,I_j] = p_{11} - p_1^2 \qquad (6)$$

Therefore, the correlation function is given by

$$\rho_{ij} = (p_{11} - p_1^2)/(\, p_1 - p_1^2) \qquad (7)$$

This correlation is zero for $p_{11} = p_1^2$, which for the indicator variables I_j and I_j is also a condition of independence (see Eq. 4).

A first key issue when modeling an indicator time series $\{I_i\}$ is whether the series satisfies the conditions of stationarity and independence, i.e. whether it may be regarded as a Bernoulli Trial Sequences (BTS). To illustrate, we consider a series of 84 daily rain/no-rain indicators collected at a station in Illinois during the summer season. The recorded series is

100001011100000000011011111100000111000000000100010111100000011111
100000000001010000

The marginal probabilities of rainy and non-rainy conditions are estimated to be $p_1 = P[I_i = 1] = 30/84 = 0.36$ and $p_0 = P[I_i = 0] = 54/84 = 0.64$.

In order to determine whether the sequence is independent or not, we estimate the probabilities p_{00} and p_{11} for different time lags $|i-j|$, the correlation function $\rho_{|i-j|} = (p_{11}-p_1^2)/(p_1-p_1^2)$, and the one-day-lag conditional probabilities

$$p_{0|0} = P[I_{i+1} = 0|I_i = 0]. \qquad p_{1|0} = P[I_{i+1} = 1|I_i = 0]$$
$$p_{0|1} = P[I_{i+1} = 0|I_i = 1]. \qquad p_{11} = P[I_{i+1} = 1|I_i = 1] \qquad (8)$$

Results are as follows:

	|i-j|		
	1	**2**	**3**
p_{00}	0.50	0.48	0.42
p_{11}	0.20	0.20	0.14
ρ_{ij}	0.40	0.29	0.03
$p_{0\mid 0}$	0.78	0.75	0.65
$p_{0\mid 1}$	0.40	0.47	0.62

Problem 5.1 *Based on the above statistics and on direct inspection of the rain/no-rain record, discuss whether the BTS assumption is realistic or not.*

From the previous analysis, you have probably concluded that the assumption of independence is not tenable. Then the problem is to formulate a suitable model with dependence. Perhaps the simplest random sequences $\{I_i\}$ with dependence are so-called Markov chains. The Markov property is a limited-memory property, which in our case says that the weather in the future depends on the weather in the past only through the most recent (the present) weather conditions. Therefore, in predicting whether tomorrow it will rain or shine, one needs only consider whether it rains or shines today. While such an assumption may not be entirely verified by actual weather patterns, it provides a convenient first step towards modeling dependence.

A way to appreciate the simplicity of Markov models is to set up a weather simulation procedure. From the rain/shine record above, you have estimated the marginal probabilities p_0 and p_1 and the one-day conditional probabilities in Eq. 8. This is all one needs to simulate Markov sequences of dry/wet weather patterns. Start by simulating the weather state (rain or shine) in day 1 by randomly drawing from the marginal distribution. Then, for day 2, draw a random variable from the conditional distribution given the state in day 1. Continue this operation for successive days. Notice that this is analogous to an “urn model” in which there are two urns with different proportions of white and black balls. White balls signify shine and black balls signify rain. The proportions of black and white balls in the two urns correspond to the conditional probabilities of rain and shine given that the previous day was sunny (urn 1) or rainy (urn 2). To generate weather sequences, one draws randomly, with replacement, from the urn that corresponds to the weather of the preceding day.

Problem 5.2 *Use the above Markov (urn) model and the marginal and conditional probabilities extracted from the Illinois sequence to produce a long synthetic weather record. Compare statistics of the record with those of the actual weather sample.*