Notes on Sequence Modelling

G.A. Jarrad

June 23, 2015

1 Introduction

blah, blah, blah

2 Random Process Sequences

Consider, in general terms, a random process R that generates a sequence of variables, $R_1, R_2, R_3, \ldots$, where the index i gives the discrete *stage* in the sequence, and each variable R_i randomly takes some value $r_i \in \mathcal{R}$. Then, for some arbitrary sequence length n, we define

$$\vec{R}_n \quad = \quad (R_1, R_2, \ldots, R_n) \tag{2.1}$$

to be a length-n process sequence, i.e. $|\vec{R}_n| = n$, and further define

$$\vec{r}_n \quad = \quad (r_1, r_2, \ldots, r_n) \in \mathcal{R}^n \tag{2.2}$$

to be a corresponding length-n observation sequence of values. The probability (for a discrete-value process) or probability density (for a continuous-value process) of a given sequence $\vec{r}_n$ is then defined as

$$p(\vec{R}_n = \vec{r}_n) \quad = \quad p(R_1 = r_1, \ldots, R_n = r_n)\,. \tag{2.3}$$

Hence, note that if R is a discrete-value process then we must have

$$\sum_{\vec{r}_n \in \mathcal{R}^n} p(\vec{R}_n = \vec{r}_n) \quad = \quad \sum_{r_1 \in \mathcal{R}} \cdots \sum_{r_n \in \mathcal{R}} p(R_1 = r_1, \ldots, R_n = r_n) \quad = \; 1\,. \tag{2.4}$$

Similarly, if R is a continuous-value process, then we must instead have

$$\int_{\mathcal{R}^n} p(\vec{R}_n = \vec{r}_n)\, d\vec{r}_n \; = \; \int_{\mathcal{R}} \cdots \int_{\mathcal{R}} p(R_1 = r_1, \ldots, R_n = r_n)\, dr_1 \cdots dr_n = 1\,. \tag{2.5}$$

In other words, given the sequence length n, the set $\mathcal{R}^n$ of all possible sequences $\vec{r}_n$ covers the entire probability space.

This summation property causes modelling problems if we do not know in advance the exact length of a sequence. For example, if the set of all length-1 sequences already covers the entire probability space, and so too does the set of all length-2 sequences, then the set of all length-1 and length-2 sequences covers the space twice over. This problem, however, only exists due to our ambiguous

notion of a sequence. In practice, suppose we have observed a given sequence $\vec{r}_n$. How do we know if the underlying process R has terminated, or will instead continue to generate another observed value r_{n+1}, leading to the extended sequence $\vec{r}_{n+1}$? Similarly, how do we know that the first observed value r_1 was not in fact part of a longer, unobserved sequence of values $\ldots, r_{-2}, r_{-1}, r_0$?

In order to handle such difficulties, we distinguish between a so-called *incomplete* sequence $\vec{R}_n$, and a *complete* sequence $\langle \vec{R}_n \rangle$ that has definite stages of initiation and termination. Thus, we see that each length-2 incomplete sequence $\vec{r}_2$ starts with a length-1 incomplete sequence $\vec{r}_1$, so that measuring the set of all length-1 and length-2 incomplete sequences amounts to double counting.

A complete sequence may be specified by introducing indicator variables that define the start and end of the sequence. Thus, indicator ι_0 takes a value of 1 if the sequence starts at stage 0 (i.e. just prior to R_1), or a value of 0 if it does not. Similarly, indicator τ_{n+1} takes a value of 1 if the sequence terminates at stage $n+1$ (i.e. immediately after R_n), or a value of 0 if it does not. We may now also consider *partially complete* sequences, namely the *start sequence* $\langle \vec{R}_n$, which was initiated at stage 0 but not yet definitely terminated (i.e. $\iota_0 = 1$ but τ_{n+1} is unknown), and the *end sequence* $\vec{R}_n \rangle$, which was terminated at stage $n+1$ but not definitely initiated at stage 0 (i.e. $\tau_{n+1} = 1$ but ι_0 is unknown).

The probability or probability density of a given complete sequence $\langle \vec{r}_n \rangle$ is now defined as

$$p(\langle \vec{r}_n \rangle) \quad = \quad p(\iota_0 = 1, R_1 = r_1 \ldots, R_n = r_n, \tau_{n+1} = 1)\,; \tag{2.6}$$

likewise

$$p(\langle \vec{r}_n) \quad = \quad p(\iota_0 = 1, R_1 = r_1 \ldots, R_n = r_n) \tag{2.7}$$

for a start sequence, and

$$p(\vec{r}_n \rangle) \quad = \quad p(R_1 = r_1 \ldots, R_n = r_n, \tau_{n+1} = 1) \tag{2.8}$$

for an end sequence. In the special cases where we know in advance that the start sequence $\langle \vec{r}_n$ definitely does not terminate at stage $n+1$ (i.e. $\tau_{n+1} = 0$), and the end sequence $\vec{r}_n \rangle$ definitely does not initiate at stage 0 (i.e. $\iota_0 = 0$), we may write

$$p(\langle \vec{r}_n !) \quad = \quad p(\iota_0 = 1, R_1 = r_1 \ldots, R_n = r_n, \tau_{n+1} = 0)\,, \tag{2.9}$$

and

$$p(!\vec{r}_n \rangle) \quad = \quad p(\iota_0 = 0, R_1 = r_1 \ldots, R_n = r_n, \tau_{n+1} = 1)\,. \tag{2.10}$$

Hence, for a discrete-value process, the probability mass of all complete length-n sequences is given by

$$\sum_{\vec{r}_n \in \mathcal{R}^n} p(\langle \vec{r}_n \rangle) \quad = \quad p(\iota_0 = 1, \tau_{n+1} = 1) \;\doteq\; P(N = n)\,, \tag{2.11}$$

where we have introduced the random variable N to denote the length of an arbitrary complete sequence. For the corresponding continuous-value process, we likewise deduce that

$$\int_{\mathcal{R}^n} p(\langle \vec{r}_n \rangle)\, d\vec{r}_n \quad = \quad p(\iota_0 = 1, \tau_{n+1} = 1) \;=\; P(N = n)\,. \tag{2.12}$$

We therefore deduce that the set $\mathcal{R}^* = \bigcup_{n=0}^{\infty} \mathcal{R}^n$ of all complete sequences of arbitrary length covers the entire probability space exactly once, since for the discrete-value case we have

$$\sum_{\vec{r}_* \in \mathcal{R}^*} p(\langle \vec{r}_* \rangle) \;=\; \sum_{n=0}^{\infty} \sum_{\vec{r}_n \in \mathcal{R}^n} p(\langle \vec{r}_n \rangle) \;=\; \sum_{n=0}^{\infty} P(N = n) \;=\; 1\,, \quad (2.13)$$

and for the continuous-value case we have

$$\int_{\mathcal{R}^*} p(\langle \vec{r}_* \rangle)\, d\vec{r}_* \;=\; \sum_{n=0}^{\infty} \int_{\mathcal{R}^n} p(\langle \vec{r}_n \rangle)\, d\vec{r}_n \;=\; \sum_{n=0}^{\infty} P(N = n) \;=\; 1\,. \quad (2.14)$$

3 Markov Process Sequences

In the previous section we defined random processes and the sequences they generate. It is generally assumed that a random process R obeys strict causality, such that the values generated by the process form a *temporal* sequence, where the distribution of values for variable R_i, at stage i, depends only upon the values generated previously in the sequence at stages $i-1, i-2, \ldots, 1$. In addition, for a complete sequence the distribution of the variable R_1 at the initial stage depends strongly upon being first in the sequence, and likewise the distribution of the variable R_n, for a length-n sequence, depends strongly upon both the past values in the sequence and on the fact that it is the final stage.

Causality therefore indicates that a probability model of the sequences generated by a random process R can be factored in terms of conditioning, at each stage i, the random variable R_i on the values of the previous variables $R_{i-1}, R_{i-2}, \ldots, R_1$. Thus, for any given complete sequence $\langle \vec{r}_n \rangle$, we obtain the model

$$\begin{aligned} p(\langle \vec{r}_n \rangle) \;=\;& p(\iota_0 = 1)\, p(R_1 = r_1 \mid \iota_0 = 1) \\ & p(R_2 = r_2 \mid R_1 = r_1, \iota_0 = 1) \cdots \\ & p(R_n = r_n \mid R_{n-1} = r_{n-1}, \ldots, R_1 = r_1, \iota_0 = 1) \\ & p(\tau_{n+1} = 1 \mid R_n = r_n, \ldots, R_1 = r_1, \iota_0 = 1) \qquad (3.1) \\ =\;& p(\iota_0 = 1) \prod_{i=1}^{n} p(R_i = r_i \mid \vec{R}_{i-1} = \vec{r}_{i-1}, \iota_0 = 1) \\ & p(\tau_{n+1} = 1 \mid \vec{R}_n = \vec{r}_n, \iota_0 = 1)\,. \qquad (3.2) \end{aligned}$$

Recall from Section 2 that we may adjust this model to allow for partially complete or incomplete sequences by suitably modifying the corresponding boundary conditions for ι_0 and τ_{n+1}.

In practice, this factored model is usually simplified further by limiting the conditional dependency on past values to a small number. In particular, one might assume the (first-order) Markov property, such that the value at any stage i depends only upon the immediate past value at stage $i-1$, namely

$$\begin{aligned} p(\langle \vec{r}_n \rangle) \;=\;& p(\iota_0 = 1)\, p(R_1 = r_1 \mid \iota_0 = 1) \\ & \prod_{i=2}^{n} p(R_i = r_i \mid R_{i-1} = r_{i-1}) \\ & p(\tau_{n+1} = 1 \mid R_n = r_n)\,. \qquad (3.3) \end{aligned}$$

Alternatively, equation (3.2) may be generalised to an m-th order model by limiting the dependency of R_i to the previous m values, namely

$$\begin{aligned} & p(R_i = r_i \mid \vec{R}_{i-1} = \vec{r}_{i-1}, \iota_0 = 1) \\ & \quad = \quad \begin{cases} p(R_i = r_i \mid \vec{R}_{i-m,i-1} = \vec{r}_{i-m,i-1}) & \text{if } i > m\,, \\ p(R_i = r_i \mid \vec{R}_{i-1} = \vec{r}_{i-1}, \iota_0 = 1) & \text{if } i \leq m\,. \end{cases} \quad , \end{aligned} \tag{3.4}$$

where

$$\vec{R}_{i,j} \quad = \quad (R_i, R_{i+1}, \ldots, R_j)\,. \tag{3.5}$$