MVE550 2020 Lecture 1.1
Introduction to stochastic processes and Bayesian inference

Petter Mostad

Chalmers University

November 1, 2020

Examples of things one might want to study

- Growth of bacterial colony (example from our textbook Dobrow).
- The price of a stock in the stockmarket.
- The evolution of a "trait" (or property) in a phylogenetic tree.
- The spread of an infections disease.

Some common features in the examples

- There is a time involved: Observations can be “indexed” with a specific time.
- Our *goal* can be to “understand” something or to *make predictions*.
- My opinion: Prediction is the central goal!
  - To “understand” something usually means to create some kind of underlying *model*.
  - Any model is a scientific model only if it makes *predictions*, and it can only be evaluated in terms of the correctness of its predictions.

Deterministic and stochastic models

- In some cases, i.e., basic physics, it makes sense to make exact predictions, without uncertainty.
- Example: $F = ma$.
- We may call such models *deterministic* models.
- In most cases, it is more reasonable to make probabilistic predictions.
- All our examples above are of this type.
- Stochastic models = probabilistic models, making probability predictions.

Stochastic processes

- A stochastic process is a collection of *random variables* $\{X_t, t \in I\}$.
- The set I is the *index set* of the process. I most often represents a set of *specific times*.
- The random variables are defined on a common *state space* $\mathcal{S}$. This set represents the *possible values* the random variables X_t can have.
- In our four examples, the state spaces might be
  - A non-negative count.
  - A non-negative real number.
  - A set of species, with descriptions of their relevant genetic sequences and their relevant traits.
  - Some description of the amount of infections (and possibly immunity) in the population.
- Some further examples:
  - A vector of real numbers.
  - A grid of numbers (representing an image?)
  - A 3D grid of numbers (representing the stresses in a building?)
  - An infinte sequence of numbers.
  - A continuous function from $[0, 1]$ to real numbers.

The Markov property

- For us, the index set I will (generally) be some subset of the real numbers (representing time).
- Generally, for any $t_0 \in I$, the probabilities for outcomes for X_t, where $t > t_0$, may depend on the values of X_s for all $s \leq t_0$.
- The process fulfills the *Markov property* if, for any $t_0 \in I$, whenever X_{t_0} is known, X_t (with $t > t_0$) is independent of the values for X_s for all $s < t_0$.
- More or less all the stochastic processes we will deal with in this course will have this property.

What is a Random Variable?

Intuitive definition:

- A *variable* which has possible values in some *state space* $\mathcal{S}$. We will generally assume that the state space is a subset of the real numbers.
- Examples of state spaces used in the course:
  - $\mathcal{S} = \{1, 2, 3, 4\}$.
  - $\mathcal{S}$ is all positive integers: $\{1, 2, 3, 4, 5, \ldots, \}$.
  - $\mathcal{S}$ is all non-negative real numbers.
- There are probabilities assigned to values and sets of values in the state space.
- We separate between *discrete* and *continuous* random variables.
- For discrete random variables, we assign a probability to each single value in the state space.
- For continous random variables, we focus instead of assigning probabilities to *intervals* of values in the state space.
- (We will return shortly with more precise definitions.)

Main types of stochastic processes in this course

Dobrow Chapters	Time (I)	State space ($\mathcal{S}$)
2&3: Discrete Markov chains	Discrete	Discrete
4: Branching processes	Discrete	Discrete
5: Markov chain Monte Carlo	Discrete	Continuous/Discrete
6: Poisson processes	Continuous	Discrete
7: Continuous-time Markov chains	Continuous	Discrete
8: Brownian motion	Continuous	Continuous

What do we want to do with the models?

- Easiest approach: Set up model based on general knowledge, make predictions from models.
- Examples:
  - Throwing a dice.
  - Predictions about a card game.
  - Other types of game predictions.
- More useful situation:
  1. You have *data*.
  2. You want find a model so that the data could reasonably be produced by it.
  3. You want to use this model for predictions of future observations.
- **Using data in this way is called *inference*.**

How to find a model that might have produced the data?

Two (main) alternatives:

- Classical inference (or "frequentist" inference):
  1. Find *estimates* for *parameters* of the model, using the data.
  2. To find *estimates*, use *estimators* that have *desireable properties*.
  3. Plug the estimates into the models and make predictions from resulting models.
- Bayesian inference:
  1. Set up a stochastic model making predictions of *observed data* and *possible future data*.
  2. Find the *conditional probability* for the future predictions given the values of the observed data.

MVE550 2020 Lecture 1.2
Random variables and conditional probability

Petter Mostad

Chalmers University

November 2, 2020

Outline

- Review: Dobrow Appendices A, B, C, D
- Random variables
- Computer simulations
- Conditional probability / conditional distributions
- Laws of total expectation and total variance

Dobrow Appendices A, B, C, D

- These appendices contain material that you (in principle) should know already.
- I *strongly recommend* that you look through these, at least to find out how much of them you know and how much and what you don't know.
- Appendix A: Getting started with R.
- Appendix B: Probability review.
- Appendix C: Summary of common probability distributions.
- Appendix D: Matrix algebra review.

Review: Random variables

A random variable X with state space S is a real-valued function on S together with a *probability* $\Pr(\cdot)$ on S. The probability $\Pr(\cdot)$ satisfies

- $0 \leq \Pr(A) \leq 1$ for all *measurable* subsets $A \subseteq S$.
- $\Pr(S) = 1$
- $\Pr(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \Pr(A_i)$ when the A_i are disjoint.

- These are the Kolmogorov axioms for probability.
- Measurable subsets are called *events*.
- What is a *measurable* subset?

Measurable subsets

Let S be any set.

- A *sigma-algebra* Ω on S is a set of subsets of S such that
  - Ω includes S
  - If $A \in \Omega$ then $A^c = S \setminus A \in \Omega$.
  - If $A_1, A_2, \ldots, \in \Omega$ then $\cup_{i=1}^{\infty} A_i \in \Omega$
- The *measurable sets* are those that are in an appropriately defined sigma-algebra.
- What you need to know for this course: When S is finite or countable, all subsets will be measurable. When S is some interval of real numbers, there will exist subsets that are not measurable, but we will not be concerned with them.

Computer simulation and probability

- Note: Many random variables can be represented with a computer program which *simulates* random output.
- The output is then *pseudo-random*, we may return to this.
- We may then use

$$\text{Frequency of computer output} \approx \text{Probability of output}$$

- Making this precise yields powerful computational methods, some of which we will use and/or study in this course.

Conditional probability

- Given events A and B, the *conditional probability* for A given B is

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

- Events A and B are *independent* if $\Pr(A \cap B) = \Pr(A)\Pr(B)$.
- Law of total probability: Let $B_1, \ldots, B_k$ be a sequence of events that *partitions* S. Then

$$\Pr(A) = \sum_{i=1}^{k} \Pr(A \cap B_i) = \sum_{i=1}^{k} \Pr(A \mid B_i)\Pr(B_i).$$

- Bayes law for probabilities follows from definitions:

$$\Pr(B \mid A) = \frac{\Pr(A \mid B)\Pr(B)}{\Pr(A)}$$

Notation for discrete probability distributions

- For a discrete random variable X we may write $\Pr(X = x)$ for $\Pr(\{x : X = x\})$.
- For a joint distribution for two discrete random variables X and Y we may write $\Pr(X = x, Y = y)$ for $\Pr(\{x : X = x\} \cap \{y : Y = y\})$ and $\Pr(X = x \mid Y = y)$ for $\Pr(\{x : X = x\} \mid \{y : Y = y\})$
- The formulas of the previous overhead can then be written

$$
\begin{aligned}
\Pr(X = x \mid Y = y) &= \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)} \\
\Pr(X = x) &= \sum_y \Pr(X = x \mid Y = y) \Pr(Y = y) \\
\Pr(Y = y \mid X = x) &= \frac{\Pr(X = x \mid Y = y) \Pr(Y = y)}{\Pr(X = x)}
\end{aligned}
$$

The generic π-notation

We may use the *generic* π-notation as a shorthand:

- Write $\pi(x)$ for $\Pr(X = x)$, $\pi(x, y)$ for $\Pr(X = x, Y = y)$ and $\pi(x \mid y)$ for $\Pr(X = x \mid Y = y)$.
- The formulas of the previous overhead can then be written

$$\begin{aligned}
\pi(x \mid y) &= \frac{\pi(x, y)}{\pi(y)} \\
\pi(x) &= \sum_y \pi(x \mid y)\pi(y) \\
\pi(y \mid x) &= \frac{\pi(x \mid y)\pi(y)}{\pi(x)}
\end{aligned}$$

- (This notation will be used in the Lecture Notes, but is not used in Dobrow).

Conditional densities for continuous distributions

- For a continuous random variable X, we will write its *density function* as $\pi(x)$, extending the generic π notation.
- If we have a joint distribution for continuous random variables X and Y, the joint density function may be written $\pi(x, y)$.
- We get formulas like

$$\int \pi(x)\, dx = 1 \qquad \text{and} \qquad \int \pi(x, y)\, dy = \pi(x).$$

- We may *define* the conditional density as

$$\pi(y \mid x) = \frac{\pi(x, y)}{\pi(x)}.$$

- We get similar formulas as for discrete variables:

$$\begin{aligned} \pi(x) &= \int_y \pi(x \mid y)\pi(y)\, dy \\ \pi(y \mid x) &= \frac{\pi(x \mid y)\pi(y)}{\pi(x)} \end{aligned}$$

Expectation and conditional expectation

- Recall, the expectation of a discrete random variable is

$$\mathrm{E}(Y) = \sum_y y\pi(y)$$

and of a continuous random variable

$$\mathrm{E}(Y) = \int_y y\pi(y)\,dy.$$

- The conditional expectation in the discrete case is

$$\mathrm{E}(Y \mid X = x) = \sum_y y\pi(y \mid x)$$

and in the continous case

$$\mathrm{E}(Y \mid X = x) = \int_y y\pi(y \mid x)\,dy.$$

Law of total expectation

- If X is a discrete random variable, we get that

$$\mathrm{E}\,(Y) = \sum_x \mathrm{E}\,(Y \mid X = x)\,\pi(x).$$

- If X is a continuous random variable we get

$$\mathrm{E}\,(Y) = \int_x \mathrm{E}\,(Y \mid X = x)\,\pi(x)\,dx$$

- In both cases this can be written as

$$\mathrm{E}\,(Y) = \mathrm{E}\,(\mathrm{E}\,(Y \mid X))\,.$$

Law of total variance

- Recall that, by definition,

$$\mathrm{Var}\left(Y\right) = \mathrm{E}\left(\left(Y - \mathrm{E}\left(Y\right)\right)^2\right) = \mathrm{E}\left(Y^2\right) - \mathrm{E}\left(Y\right)^2.$$

- Similarly, we have for the conditional variance

$$\mathrm{Var}\left(Y \mid X = x\right) = \mathrm{E}_{Y|X=x}\left(\left(Y - \mathrm{E}\left(Y \mid X = x\right)\right)^2\right)$$

- With these definitions, we can now show the law of total variance:

$$\mathrm{Var}\left(Y\right) = \mathrm{E}\left(\mathrm{Var}\left(Y \mid X\right)\right) + \mathrm{Var}\left(\mathrm{E}\left(Y \mid X\right)\right)$$