Atm S 552 Lecture 1
Course Introduction and Review of Probability

Winter 2015

1.1 Introduction

Course goal: Students will learn a computational toolbox to analyze and explore large datasets coming from geophysical applications.

In this class, we will focus on data 'structured' in terms of a set of *numerical characteristics* describing each **data entity** . Data entities may have a meaningful ordering (e. g. in terms of space and/or time coordinates, such as the pixel grid of a rectangular image, or in terms of connections in a network or tree). In other case, no such ordering exists (e. g. the students of a class). The form of the data affects what data analysis tools will be appropriate.

We will assume the data can be represented as a one- or multi-dimensional array of numbers. Some dimensions may index multiple numerical characteristics of each entity, and other dimensions may index multiple entities. For instance, consider two entities, the sea level pressure [hPa] and sea-surface temperature [K] fields, tabulated on a longitude-latitude grid at a sequence of times. This might be stored as an $m \times n \times p \times 2$ 4D array of real numbers, where m indexes the longitude, n the latitude, p the time, and the 4th dimension indexes the field (1 for SLP, 2 for SST).

Data arrays may take diverse forms, e. g.:

- time series (univariate, multivariate)
- gridded multidimensional array (image, spatial map)
- space-time (global weather observations)
- other connectivities (networks)

Usually, we have some analysis goals, often connected with comparing a model (physical or statistical) with the data or using the data to develop a simple model or decision-making algorithm.

- Detecting an oscillation or trend
- Filtering out a 'signal' from residual 'noise'

- Statistically characterizing time or space variability in the data.
- Finding a reduced-complexity approximation to the data.
- Estimating uncertain parameters, model-data fusion and data assimilation

The mathematics that we will use involve

- Elementary probability and statistics
- Linear algebra
- Fourier analysis

and are implemented in Matlab (used here), Python, R and other popular software packages for data manipulation, data analysis and statistics.

1.2 Review of probability concepts/definitions

We start with a quick summary of probability and statistics. Our learning goal is to understand, formulate and test simple statistical models of data. To begin, we introduce basic probability concepts/definitions:

Sample space Set S of all possible outcomes of a trial or experiment; can be discrete or continuous., e. g. the four outcomes $\{(H/T, H/T)\}$ for sequentially flipping two coins.

Event Some subset E of the sample space, e. g. $\{(H,T),(T,H)\}$ is the event in which exactly one head is tossed.

Probability of an event E The proportion of the time $P(E)$ that E is realized if the trial is repeated ad infinitum.

Conditional probability Probability that event E_2 occurs given that event E_1 also occurs, $P(E_2|E_1) = P(E_2E_1)/P(E_1)$, e. g. the probability of the event E_2 that two coins both come up heads given the event E_1 at least one of them comes up heads is 1/3.

Independent events Events E_1 and E_2 are independent if $P(E_1E_2) = P(E_1)P(E_2)$, i.e. if the probability of event E_1 is unaffected by the co-occurrence of event E_2. Example: E_1 is coin 1 coming up heads, E_2 is coin 2 coming up heads.

Random variable (RV) The undetermined outcome of some repeatable experiment that can be described in terms of events within a known sample space. For instance, the number of heads N obtained when flipping two fair coins is a random variable with possible values 0, 1, 2, and $P(N = \{0,1,2\}) = \{1/4, 1/2, 1/4\}$.

Discrete vs. continuous RV A discrete RV can take a finite or countable set of values; a continuous RV can take any value over an uncountable (usually continuous) range of values.

Cumulative distribution function (CDF) of a RV X is $F(a) = P(X \leq a)$.

Probability density function (PDF) of a continuous RV is

$$f(x) = \lim_{\epsilon \to 0} \frac{P(x - \epsilon/2 \leq X \leq x + \epsilon/2)}{\epsilon}$$

Its integral over the range of X is 1. The PDF of a discrete RV X can be regarded as a sum of delta functions at each possible outcome x_n weighted by the probability $p(x_n)$ of that outcome.