Quantifying uncertainty and correlation in complex systems

Problem sheet 1

(counts $20/50$ homework marks)

1.1 (For credit) Graphical analysis of data

Download the datasets `data11.dat`, `data12.dat` and `data13.dat`.

(a) Use Q-Q-plots to see if any two of the datasets are from the same distribution.

(b) Identify the type of distribution of each set, e.g. by using Q-Q-plots against theoretical distributions. You do not have to include those plots, but support your identification with one convincing plot in each case.
For exponentials or power laws plot the empirical tail in a suitable scaling and compare to the CDF, for Gaussians or uniform distributions use `ksdensity` and compare to the PDF on a suitable scale. Infer relevant parameters for CDF or PDF from the data (such as mean, variance or range of the distribution).

1.2 (For credit) Characteristic functions

Compute the chracteristic function $\phi_X(t)$ for the exponential distribution with $f_X(x) = \lambda e^{-\lambda x}$, $x \geq 0$, and for the uniform distribution $f_X(x) = 1/(2\pi)$ for $x \in [0, 2\pi)$.

1.3 (For credit) Temperatures in the Midlands

Download the file `Midlandstemperatures.zip`. The .txt files contain the original data with explanations, but you can focus on the two sheets in the .xls file.

(a) For mean temperatures (worksheet Midlands_mean) test whether the annual means and the monthly means of your favourate month a normally distributed using Q-Q-plots.
Produce a smoothed histogram with `ksdensity` and compare to the expected Gaussian PDFs, estimating mean and variance from the data.

(b) For maximum temperatures (worksheet Midlands_max) first compute the annual maximum over all months for each year. What distribution do you expect for the the yearly maxima?
Estimate the relevant parameter values from the data, and compare a smoothed histogram with the PDF of the expected distribution. (Use `gevfit` and `gevpdf` as done in class.)

1.4 (For credit) Conditional probabilities and Bayes' Rule

Let us revisit the problem of false positives discussed in the lectures. Suppose a disease affects 1 person in 10000 of the population. As part of a government monitoring programme, you are randomly selected to be tested for the disease and visit the doctor to take the test. The doctor tells you that the test returns an accurate result 99% of the time (that is to say its sensitivity and specificity are 99%). Your test comes back positive.

(a) Before you took the test, what would be your estimate of the probability that you have the disease?

(b) After your test comes back positive, what is your estimate of the probability that you have the disease?

(c) How accurate would the test have to be before you would be seriously worried about a positive outcome (say 50% certain that you have the disease)?

1.5 Correlations

(a) **(Not for credit)** Generate a sample of m 2-dimensional Gaussians with mean $(0, 1)$ and covariance matrix $\left(\begin{smallmatrix} 1 & -1 \\ -1 & 2 \end{smallmatrix}\right)$ using `mvnrnd(MU,SIGMA,m)`. Use `cov(X)` to estimate the covariance matrix from the data and produce a scatter plot to visualize the dependence. What happens if the sign of the covariances changes from -1 to 1.?

(b) **(for credit)** Consider the mean temperatures in the Midlands from the file in Q1.3. Produce scatter plots to see if the temperatures in January and February are correlated, as well as January and July, and July and August.
Use `cov(X)` and `corrcoef(X)` to compute covariance matrices and correlations, and discuss the results in a few lines.

1.6 (For credit) Maximum likelihood estimator for exponential

Suppose that you observe a sequence of N random variables $x_1, x_2, \ldots, x_N$, and you have reason to believe that they are iid exponential.

(a) What is the likelihood function of the sequence $x_1, x_2, \ldots, x_N$?

(b) Show that the Maximum Likelihood Estimator for the mean μ is

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i \, .$$

(c) Is this estimator biased or unbiased? Justify your answer by direct computation.

(d) Download the file `exponential.dat`. It is claimed that this file contains iid exponential random variables. Compute the Maximum Likelihood Estimate of the mean $\hat{\mu}$ and compare the empirical tail of the data with the expected theoretical one (`1-expcdf(x,`$\hat{\mu}$`)`).

1.7 (Not for credit) CLT and extreme value statistics

Generate $m = 500$ samples of a sequence of $n = 10, 100, 1000$ iid random variables from a uniform (`rand(m,n)`), exponential (`exprnd(`μ`,m,n)`) and generalized pareto power law (`gprnd(k,`σ`,`θ`,[m,n])`).

(a) Confirm the CLT by plotting a smoothed histogram of the rescaled sums $S_n/\sqrt{n}$ or the appropriate rescaling for the pareto law, comparing with the theoretical prediction.

(b) Confirm convergence to extreme value distributions by plotting a smoothed histogram of the appropriately rescaled maxima M_n, comparing with the theoretical prediction.

1.8 (Not for credit) Entropy maximization The entropy S assigns a number $S(\pi)$ to a probability distribution, interpreted as the '*uncertainty*' (degree randomness) of π. Let f be the PDF for a continuous random variable X, then the entropy is given by

$$S(f) := -\int_D f(x) \ln f(x)\, dx\ , \quad \text{where } X \text{ is taking values in } D \subseteq \mathbb{R}\ .$$

We want to maximize the entropy over PDFs f, i.e. find the PDF with maximal uncertainty. We 'Taylor-expand' S around some PDF f using the variational formula

$$S(f + \epsilon f_1) = S(f) + \epsilon \frac{\delta S(f)}{\delta f}(f_1) + o(\epsilon)$$

for some fixed function $f_1 : D \to \mathbb{R}$ and $\epsilon \searrow 0$. The variational derivative is just given by the partial derivative (fixing everything except ϵ)

$$\frac{\delta S(f)}{\delta f}(f_1) := \frac{\partial}{\partial \epsilon} S(f + \epsilon f_1)\Big|_{\epsilon=0}\ .$$

(a) Show that

$$\frac{\delta S(f)}{\delta f}(f_1) = -\int_D f_1(x)\big(1 + \ln f(x)\big)dx$$

(b) Let $D = [0, 1]$. Use the method of Lagrange multipliers to maximize $S(f)$ where variations of f should fulfill the constraint

$$\int_D \big(f(x) + \epsilon f_1(x)\big)dx = 1\ .$$

The maximizer should be the PDF of the uniform distribution.

(c) Now let $D = [0, \infty]$ and maximize S under the additional constraint

$$\int_D x\big(f(x) + \epsilon f_1(x)\big)dx = \mu\ .$$

So we maximize among PDFs with a given mean $\mu > 0$, which leads to an exponential distribution.

(d) Now let $D = \mathbb{R}$ and maximze S among distributions with given mean $\mu = 0$ and variance σ^2. The resulting PDF is that of a Gaussian.