Basics tutorial (2019)

Exercise 1: Matrix manipulations and numerics warm-up

The file 'DataFiles/SN_covmat.txt' contains the vectorized version of the covariance matrix of an old supernova data set (the 'JLA data set').

1. Read in the covariance matrix, find its dimension, and convert it from a vector to a matrix.
2. Plot the covariance matrix.
3. Compute and plot the correlation matrix.
4. Which two data points are the most correlated of them all? Which ones are the least correlated?
5. If you were forced to throw out 10 data points, but you wish to lose as little information as possible, which data points would you throw out?
6. Which data point has the largest error bar, as it would usually be plotted when showing a plot of the data and their error bars?
7. Compute the determinant of the covariance matrix. Think of the properties of the determinant in linear maps as they appear in linear algebra. Which meaning does the determinant of the covariance matrix have?
8. Invert the covariance matrix, and plot the inverse (which is often called 'precision matrix').

Exercise 2: Simple and/or interesting questions

1. What is the probability that a six-sided dice falls with the '5' facing upwards?
2. What is the probability that a fair coin lands 'heads'?
3. After having thrown a dice infinitely many times, what is the average score?
4. When rolling a dice 1000 times, what is the distribution of how often sixes were thrown?
5. If colour blindness is inherited via the X-chromosome, and a family of four has a colour-blind father, a colour-seeing mother, and a colour-blind son, what is the probability that the family's daughter is colour-blind as well?
6. Upon birth, the gender of a new-born is not always determined through genetics. Animal species exist which determine gender through other factors than genetics, e.g. breeding temperature. However, essentially all species where gender is genetically determined (i.e. via equivalents of the human X- and Y-chromosomes) have a gender ratio of 50:50. Why is that so? Would the species not breed faster if there were many more females than males?
7. You play a game of hangman, meaning you have to guess a word, indicated e.g. as '__ __ __ ... __'. Which letter of the alphabet is the most informative?

Exercise 3: Bargaining for funding

You wish to fund an experiment of yours. If you have more money, then you can afford more data points. The more data points you have, the more precise will your measurement be. Sadly, the more money you want, the less inclined to support you will the funding agency be. In the end, you have to give scaling relations.

1. Imagine your data points are iid, drawn from a Gaussian distribution of unknown mean but known variance. You wish to estimate the mean. How does the error of the mean decrease with the number of data points funded?

2. Imagine your data points are again iid, but this time drawn from a Uniform distribution of unknown upper limit. You wish to estimate the upper bound. How does the error on the upper bound decrease with the number of data points funded?

Exercise 4: Background estimation: where would you measure next?

Consider the situation in Fig. 1. The data points there shown have standard deviations which are constant per box, where the boxes are demarked by the vertical dashed lines. The standard deviations are indicated in each box. The noise is additive Gaussian noise and the points are uncorrelated. You think the data points come from a constant offset c, and you wish to measure this constant as precisely as possible. You have enough money to measure once more.

1. To which box would you add a data point?

2. What is the variance of your estimate of c?

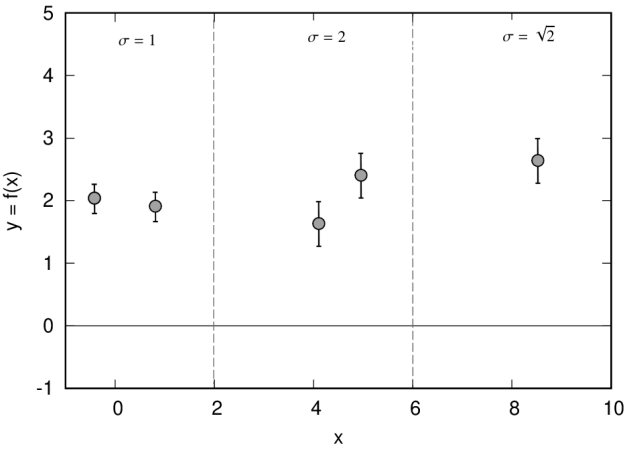

Fig. 1: Where would you add a new data point?

Exercise 5: Being conservative: where would you measure now?

Consider the situation in Fig. 2. Unphysical regions where there cannot be any data are shaded. You have control over the variable x: by chosing x you can decide where to measure next. Data points which you already have are indicated. The standard deviations σ per box refer to the error bars attached to the data shown. All errors are uncorrelated.
You do not know the functional relation between x and y. You therefore replace it by a spline and you wish to estimate the position of the spline points. These are indicated by black squares. The connecting line is the then resulting spline. The squares can move up and down, but not left and right.

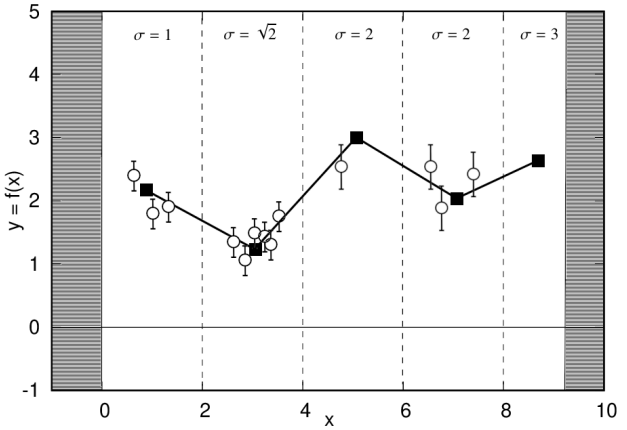

Fig. 2: Where would you add a new data point?

- To learn most about the function, where do you measure next?

- Sketch possible shapes of the function.

Exercise 6: Experimental design

Consider the situation in Fig. 3. The black squares are again spline points whose position you wish to estimate. They can only move vertically.
You currently have money to measure, but some Brexit-related politician might soon decide to not fund your research anymore. You therefore wish to learn as quickly as possible about the splined function, not knowing when funding will stop, i.e. not knowing which is the last data point you can take.

- Design the optimal sequence in which to measure. Stop after 10 data points.

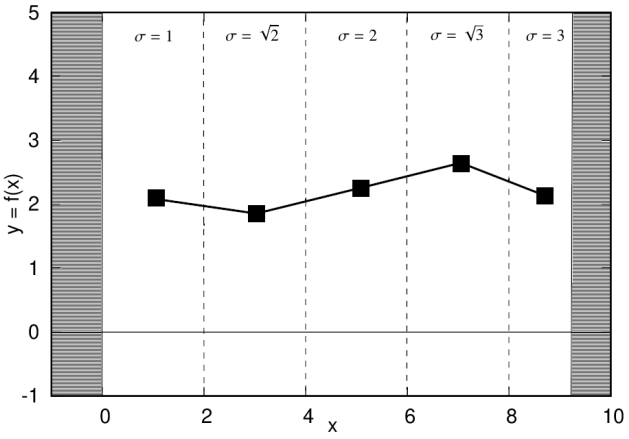

Fig. 3: In which sequence would you put down 10 data points?

Exercise 7: The pdf of astronomical magnitudes

In the lecture we saw that non-linear functions of Gaussian random variables do not follow a Gaussian distribution anymore. We now apply this to astronomy.
Consider a telescope that observes a star. It will receive individual photons, which are Poisson distributed, and in the limit of bright stars the Poisson distribution tends to a Gaussian distribution. Here, we can therefore realistically assume a Gaussian distribution for the photon counts. However, astronomers conventially do not work with photon counts, but with magnitudes instead.

- If the number n_o of photons received (per time interval Δt, with a detector of area A) follows the Gaussian distribution

$$n_o \sim \mathcal{G}(n_t, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(n_o - n_t)^2}{\sigma^2}\right),$$

  which distribution does the star's flux F have?

- Apparent magnitudes are related to the ratio of two fluxes by

$$m_1 - m_2 = -\frac{5}{2}\log_{10}\left(\frac{F_1}{F_2}\right).$$

  - If you observe $n_1 \sim \mathcal{G}(n_{t_1}, \sigma_1)$, and $n_2 \sim \mathcal{G}(n_{t_2}, \sigma_2)$ photons per time interval from two different stars, does the difference $\Delta m = m_1 - m_2$ in apparent magnitudes still follow a Gaussian distribution?
  - Assume all magnitudes m_1 are measured relative to a precisely known magnitude m_2 (e.g. Vega or the North-Polar Sequence; you do not need to look up these numbers for the exercise). Which distribution do the magnitudes m_2 then follow?

- Sample this process: Begin by drawing Gaussian random numbers for the photon counts, and compute the histograms of the fluxes and the magnitudes. How well is the uncertainty of magnitudes modelled by a Gaussian distribution, if
  - one star's magnitude (the reference magnitude) is precisely known

- the photon counts of both stars need to be measured and follow two Gaussian distributions of different standard deviations $\sigma_{1,2}$
- the photon counts of both stars need to be measured but follow the same Gaussian distribution with standard deviation σ. This could e.g. happen if your detector is extremely noisy.
- If time permits: Tune the brightnesses of the stars by varying n_{t_1} and n_{t_2}, and vary the standard deviations σ_1 and σ_2. Can you reach limits in which the magnitude's distribution is extremely (non-)Gaussian?

Exercise 8: Absolute values of Gaussians

Using the univariate Gaussian distribution for a random variate x,

$$x \sim \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right),$$

derive from the law of total probability, $P(A) + P(\bar{A}) = 1$, the distribution of the absolute value $|x|$.

Exercise 9: When tests fail tests

Any test is only ever as good as the mathematics inside it: It does precisely what its internal equations tell it to do, but one frequently meets the attitude that a result must be true, because some test was applied. In this exercise, we will hence break a series of well-known tests, i.e. construct situations where the tests themselves fail.

1. Consider the **Kolmogorov-Smirnov test**. It is often used to test the hypothesis that a set of random samples originate from a hypothesized distribution. Look at how it works: it uses cumulative distributions, and their differences.
   - What does the test react to?
   - What does the test not react to?
   - Use this to create a situation where the test fails.

   The failure mode is here that the test tells you your samples are compatible with a chosen distribution, even though it is the wrong distribution.

2. Consider the **Chi-squared over Degrees-of-Freedom test**. It is used to judge whether a fitted model is probably the correct model. It computes χ^2/degF at the best-fitting point, and if

   $$\chi^2/\mathrm{degF} \approx 1,$$

   then the model is considered to be the correct one. Often, one also hears that the entire analysis must be bias-free, otherwise one would not have gotten a $\chi^2/\mathrm{degF} \approx 1$.
   - Look at how the test works. What does it compute?
   - Where are sources for mistakes which the test then misinterprets?
   - Which assumptions of the test can be broken?
   - Give 3 ways in which the $\chi^2/\mathrm{degF} \approx 1$ test can fail.

   The failure mode is here that you believe you did everything correctly because you get a reduced-χ^2 of approximately unity, but in reality you either took the wrong model or you made any other crucial mistake.