From Data to Insights - Exercise sheet 4

discussed May 16 and 17

May 10, 2024

1 Something wrong with the inverse covariance

In the file `https://cloud.physik.lmu.de/index.php/s/H63R9kfz5KRtmzC` you find the covariance matrix $\mathbf{C}$ of some cosmological signal $\boldsymbol{\xi}$ [1] which consists of overall 15 data points (i.e. $\mathbf{C}$ is a 15×15 matrix).

Often, the covariance of a given signal is not known analytically but instead has to be estimated from simulated data. If one has N measurements $\boldsymbol{\xi}_1$, ... , $\boldsymbol{\xi}_N$ then the standard estimator $\hat{\mathbf{C}}$ for the covariance has the elements

$$\hat{C}_{ij} = \frac{1}{N-1} \sum_{n=1}^{N} \left(\boldsymbol{\xi}_n - \bar{\boldsymbol{\xi}}\right)_i \left(\boldsymbol{\xi}_n - \bar{\boldsymbol{\xi}}\right)_j \ , \tag{1}$$

where $\bar{\boldsymbol{\xi}} = \frac{1}{N} \sum_n \boldsymbol{\xi}_n$ is the average of all simulated measurements. In this exercise we will investigate a potential problem with the above covariance estimator.

A) Using e.g. `numpy.random.multivariate_normal` , draw $N_{\mathrm{sim}} = 50$ random vectors from a Gaussian distribution with the covariance matrix given by the file in the above link. (For this exercise, the mean of the Gaussian distribution doesn't matter, and you can set it to zero.)

B) From the 50 data vectors, obtain a covariance estimate $\hat{\mathbf{C}}$ (e.g. using `nu py.cov`). Plot it and compare it to the true covariance matrix $\mathbf{C}$.

C) Invert the true covariance to obtain the so called *precision matrix* $\boldsymbol{\Psi} = \mathbf{C}^{-1}$. Then compute the product of $\boldsymbol{\Psi}$ with the covariance estimate, i.e. consider the matrix

$$\hat{\mathbf{A}} := \boldsymbol{\Psi} \cdot \hat{\mathbf{C}} \ . \tag{2}$$

What is the trace of this matrix? Repeat that calculation a number of times (with different random seeds, which you can e.g. enforce with `numpy.random.seed`) and demonstrate that on average you get a trace of 15. This is to be expected, because

$$\begin{aligned} \langle\hat{\mathbf{C}}\rangle &= \mathbf{C} \\ \Rightarrow \langle\hat{\mathbf{A}}\rangle &= \mathbf{1} \\ \Rightarrow \langle\mathrm{Tr}\hat{\mathbf{A}}\rangle &= 15 \ . \end{aligned} \tag{3}$$

D) To evaluate a Gaussian likelihood, we don't need an estimate of the covariance, but an estimate of the inverse covariance (since the inverse covariance enters the Gaussian PDF)! Compute the inverse estimated covariance matrix, $\hat{\boldsymbol{\Psi}} = \hat{\mathbf{C}}^{-1}$. Then compute the product of $\hat{\boldsymbol{\Psi}}$ with the true covariance, i.e. consider the matrix

$$\hat{\mathbf{B}} := \hat{\boldsymbol{\Psi}} \cdot \mathbf{C} \ . \tag{4}$$

E) Is the trace of $\hat{\mathbf{B}}$ consistent with a value of 15 (the trace of the unit matrix in 15 dimensions). To appropriately answer this question you will need an error bar for Trace($\hat{\mathbf{B}}$). Obtain that by again repeating the above calculation a number of times.

E) Why is $\langle\mathrm{Trace}(\hat{\mathbf{B}})\rangle \neq 15$??

[1] It doesn't matter here what kind of signal, but for completeness: it is the covariance of a measurement of the correlation function of the gravitational lensing field induced by cosmic matter density fluctuations

2 Confidence intervals vs. credible intervals

Using e.g. `python` code, set up a version of the above experiment, where the data vector $\hat{\boldsymbol{d}} = (\hat{d})$ is only one-dimensional, and where scientists use a model $\mu(\alpha) = \alpha^3$. Assume that the true value of α is 0. How many scientists include this value within their 68.3% confidence regions? What happens if the true value of α is 1 instead? Hint: given a measurement $\hat{d}$ the best-fitting parameter is $\alpha_{\mathrm{BF}} = \hat{d}^{1/3}$. Create an array of possible α values around this location (e.g. using `numpy.linspace`) and calculate the posterior density of α on this grid of points. To know how wide your array should be and how small your steps in α should be, you can use standard error propagation. (recall that the standard deviation of $\hat{d}$ is σ, which you can take to be 0.5 as before.)

3 Discussion question

This time, let's do 'open mic': Bring up any questions / comments about frequentist and Bayesian statistics that you would like to discuss.