From Data to Insights - Exercise sheet 5

discussed May 23 and May 24

May 16, 2024

1 Impact of noisy covariances on parameter constraints

In some cosmological experiment scientists are analysing a measured data vector $\boldsymbol{d}$ of $N_{\text{data}} = 400$ data points in order to derive constraints on a vector $\boldsymbol{\pi}$ of overall $N_{\text{param}} = 30$ model parameters.

Assume that the likelihood of $\boldsymbol{d}$ is well approximated by a multi-variate Gaussian PDF. If you want to estimate the covariance matrix of $\boldsymbol{d}$ from N_{sim} simulated observations, then:

A) How big does N_{sim} need to be in order for you to get the width of the parameter contours right to $\sim 1\%$?

B) How big does N_{sim} need to be in order for you to get the location of the parameter contours right to $\sim 1\%$?

C) Assume that you only have the computing resources to generate $N_{\text{sim}} = 1000$ simulations. By how much do you need to inflate your parameter constraints in order to take into account the true uncertainty of your analysis? What percentage of your final parameter uncertainties is caused only by the fact that your covariance is noisy?

2 Noise properties of first order precision matrix expansion (PME)

Consider the 1st oder PME for the full covariance matrix $\mathbf{C}$ (i.e. we are not splitting the covariance into two contributions $\mathbf{A}$ and $\mathbf{B}$).

- Assume that our covariance model $\mathbf{M}$ is pretty good, i.e. $\mathbf{M} \approx \mathbf{C}$. Show that in this case

$$\text{Cov}\left(\hat{\Psi}_{1\text{st},ij}, \hat{\Psi}_{1\text{st},kl}\right) \approx \frac{1}{N_{\text{sim}} - 1} \left(\Psi_{1\text{st},ik}\Psi_{1\text{st},jl} + \Psi_{1\text{st},il}\Psi_{1\text{st},jk}\right) \ . \tag{1}$$

  This means that the noise of the 1st order PME behaves similar to the noise of the standard covariance estimator (and not like that of the inverse covariance estimator).

- In general, the covariance model will be different from the true covariance. Find at least two problems/criticisms that PME should be facing in this more general situation.

3 Convince yourself that Bayes=frequentist for linear, Gaussian likelihoods

- Step 1 - play God

  1.a: Make up some mean vector $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{C}$ for some data dimension $N_{\text{data}} > 3$. (In your eternal wisdom, you make sure that $\boldsymbol{C}$ is symmetric and positive definite.)

  1.b: Create 1000 parallel Universes in which Human scientists measure random realisations $\boldsymbol{d}$ drawn from a Gaussian distribution $p(\boldsymbol{d}|\boldsymbol{\mu}_0, \boldsymbol{C})$.

1.c: Then write a holy book in which you tell the Humans that their data vectors were drawn from a Gaussian distribution with covariance $\boldsymbol{C}$ and expectation value

$$\boldsymbol{\mu}(\alpha) = \boldsymbol{\mu}_0 + \alpha \cdot \boldsymbol{v} \ , \tag{2}$$

where $\boldsymbol{v}$ is some vector of the same dimension as $\boldsymbol{\mu}_0$. Tell the Humans what $\boldsymbol{v}$ is but DO NOT tell them that $\alpha = 0$!

- Step 2 - descent from the heavens and become Human

  2.a: In each Universe, use your knowledge from the holy book to carry out a Bayesian analysis and derive a posterior distribution $p(\alpha|\boldsymbol{d})$ (assuming a wide, flat prior for α).

  2.b In each Universe, determine whether $\alpha = 0$ is inside the 1-σ interval of your posterior (i.e. the smallest interval that contains 68.3% of the posterior's probability).

- Step 3 - Judgement day

  How often was $\alpha = 0$ in your 1-σ interval?

4 Discussion questions

- In which situations could a noisy covariance estimate be useful / necessary?
- List at least 4 different criteria by which one could judge the quality of a covariance estimate or a covariance model. Why could these criteria be useful?
- Does every multi-variate distribution have a covariance matrix?