Basics of statistics
Getting used to noise

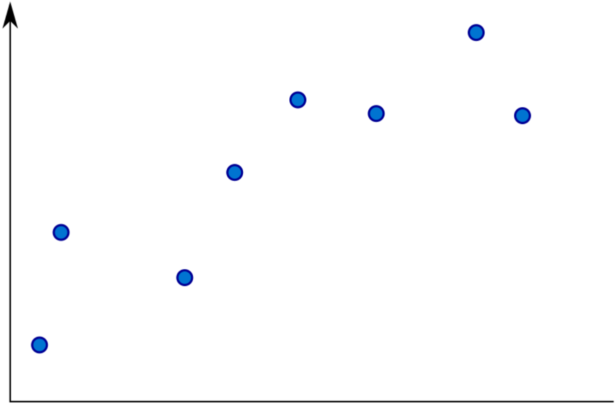
Elena Sellentin

Université de Genève
Département de Physique Théorique

MACSS Summer school
8th-12th May 2017

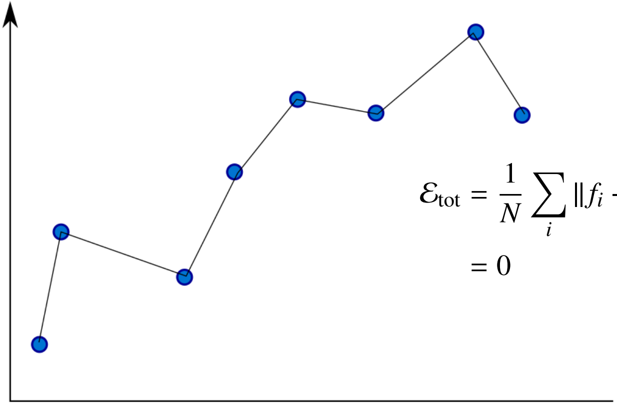
Cross validation

Given the following data points, how do we prove which curve describes the data best?



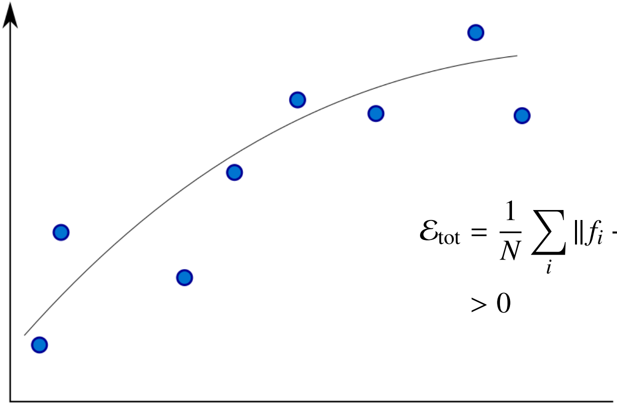
Cross validation

We all know this is wrong, but how do we prove it?



$$\mathcal{E}_{\rm tot} = \frac{1}{N}\sum_i \|f_i - f(\boldsymbol{x})\|_2^2$$
$$= 0$$

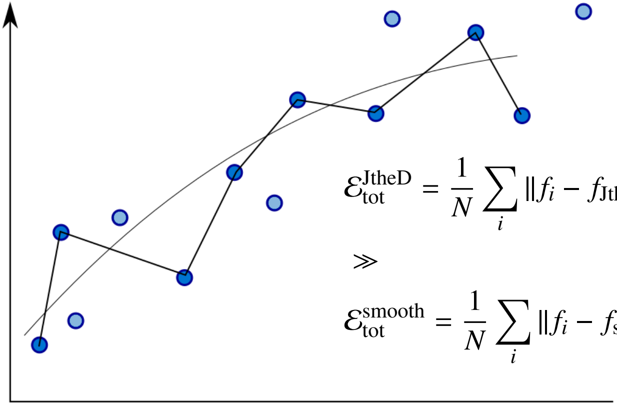
Cross validation



$$\mathcal{E}_{\mathrm{tot}} = \frac{1}{N}\sum_i \|f_i - f(\boldsymbol{x})\|_2^2$$

$$> 0$$

Cross validation



$$\mathcal{E}_{\text{tot}}^{\text{JtheD}} = \frac{1}{N}\sum_i \|f_i - f_{\text{JtheD}}(\boldsymbol{x})\|_2^2$$

$$\gg$$

$$\mathcal{E}_{\text{tot}}^{\text{smooth}} = \frac{1}{N}\sum_i \|f_i - f_{\text{smooth}}(\boldsymbol{x})\|_2^2$$

Cross validation

- We saw that join-the-dots ‘perfectly’ explains one data set, but then failed catastrophically on the repeated measurement.
- The fitted curve explained the first dataset somewhat ‘worse’, but then correctly predicted the outcome of the second measurement.

The best model minimizes...

$$\mathcal{E}_{\text{tot}} = \frac{1}{N} \sum_i ||f_i - f(\boldsymbol{x})||_2^2 \tag{1}$$

... for current and future measurements.
$\Rightarrow$ It minimizes the distance to taken data, **and** not taken but statistically iid data.

Measuring in noisy data space

We know how to calculate the distance with a Euclidean metric.

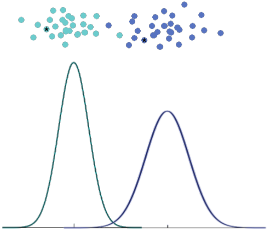
$$\boldsymbol{x}_1, \boldsymbol{x}_2 \Rightarrow \quad \sqrt{D(\boldsymbol{x}_1, \boldsymbol{x}_2)} = \sqrt{(x_1^1 - x_2^1)^2 + (x_1^2 - x_2^2) + \ldots + (x_1^n - x_2^n)^2} \tag{2}$$

$$D(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\boldsymbol{x}_1 - \boldsymbol{x}_2)^T \mathbb{I} (\boldsymbol{x}_1 - \boldsymbol{x}_2). \tag{3}$$

Now generalize this concept to a noisy space....

Measuring in noisy data space

Noise $\Rightarrow$ Distances must become uncertain.
Dimensionful parameters $\Rightarrow$ metric must measure in those units.



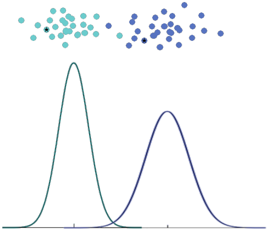
The covariance matrix:

$$\mathsf{C} = \langle (\boldsymbol{x} - \langle \boldsymbol{x} \rangle)(\boldsymbol{x} - \langle \boldsymbol{x} \rangle)^T \rangle \tag{4}$$

$$D(\boldsymbol{x}, \boldsymbol{\mu}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \mathsf{C}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}). \tag{5}$$

Mahalanobis distance: measure in units of the expected scatter.

Measuring in noisy data space



$$D(\boldsymbol{x}, \boldsymbol{\mu}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \mathsf{C}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}). \quad (6)$$

⇒ 'Distance' now measures compatibility, expectations and suprises.

Imagine you have taken one measurement $\boldsymbol{x}_1$ and worked out your measurement errors. Now you repeat and get $\boldsymbol{x}_2$.

- Huge $D(\boldsymbol{x}_1, \boldsymbol{x}_2)$: Eeew, that came unexpectedly!
- Tiny $D(\boldsymbol{x}_1, \boldsymbol{x}_2)$: Someone must be fooling you...

Consequences of Mahalanobis distances

Introduce parameters: trying to explain the data with a parametric model.

$$D(\boldsymbol{x}, \boldsymbol{\mu}(\boldsymbol{p})) = (\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{p}))^T \mathsf{C}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{p})). \tag{7}$$

- Which values do the parameters need to take, such that my model is most compatible with the data?
  $\Rightarrow$ minimize $D(\boldsymbol{x}, \boldsymbol{\mu}(\boldsymbol{p}))$ $\Rightarrow$ 'Fitting' procedure.
- How much wiggle room is then still left? $\Rightarrow$ $\hat{\boldsymbol{p}} \pm \Delta \boldsymbol{p}$.
- Which $D_{\mathrm{crit}} = D_{\mathrm{min}}(\boldsymbol{x}, \boldsymbol{\mu}(\hat{\boldsymbol{p}})) + \Delta D(\boldsymbol{x}, \boldsymbol{\mu}(\boldsymbol{p}))$ is impossibly far away?

$\rightarrow$ where $\Delta\chi^2$ comes from (just **one** example).
$\rightarrow$ depends on the chosen metric: $\mathsf{C}, \mathbb{I}, \mathsf{F}^{-1}$

Likelihoods

If $\boldsymbol{x}$ is random, then any $f(\boldsymbol{x})$ is random, hence $D(\boldsymbol{x}, \boldsymbol{\mu}(\boldsymbol{p}))$ is also a random variable.

$\Rightarrow$ If possible, work with a full likelihood.

Most famously:

$$\mathcal{G}(\boldsymbol{x}, \boldsymbol{\mu}(\boldsymbol{p}), \mathsf{C}) = \frac{1}{\sqrt{|2\pi\mathsf{C}|}} \exp\left(-\frac{1}{2}[\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{p})]^T \mathsf{C}^{-1} [\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{p})]\right). \tag{8}$$

Minimizing $\chi^2 \leftrightarrow$ maximizing a Gaussian likelihood.

Rules for the Game of Noise

- Random variables are drawn from probability distribution functions. We write $x \sim \mathcal{P}(x)$ in the univariate case, $\boldsymbol{x} \sim \mathcal{P}(\boldsymbol{x})$ in the multivariate case and $\mathsf{S} \sim \mathcal{P}(\mathsf{S})$ in the matrix-variate case.
- For any function f, if x is a random variable, so is $f(x)$.
- $P \geq 0$
- $\sum_i P_i = 1, \quad \text{especially } P(A) + P(\bar{A}) = 1$, or $\int \mathcal{P}(x)\mathrm{d}x = 1$
- $P(A, B|C) = P(A|C)P(B|A, C) = P(B|C)P(A|B, C)$,
- Exchange of variables:

$$g(\boldsymbol{y}) = f(\boldsymbol{x} = \tilde{\boldsymbol{y}}(\boldsymbol{x}))|J| \tag{9}$$

  where the Jacobian is

$$|J| = \det\left(\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}}\right). \tag{10}$$

- Bayes' Theorem: $P(A|B, C) = \frac{P(A|C)P(B|A,C)}{P(B|C)}$

Consequences

- Estimators, e.g. best-fitting parameters, are random variables (they are a function of the data).
- Sometimes you must get, and will get, 'untypical' measurements. They come from the tail of the distributions $\mathcal{P}(x)$.
- Non-linear functions of Gaussian random variables follow non-Gaussian distributions, and the CLT agrees with this.
  - the ratio of two Gaussian rv follows a Cauchy distribution
  - the exponential of a Gaussian rv is log-normally distributed
  - the absolute value of Fourier modes from a Gaussian random field follow a Rayleigh distribution
  - many more $\Rightarrow$ Exercises.

Bayesian Inference

Bayesian Inference

Inverse Problem: Given data $\boldsymbol{x}$, how do we explain them?

- Given: a dataset $\boldsymbol{x}$ and a model $\mathcal{M}(\boldsymbol{p})$
- To which posterior do the data $\boldsymbol{x}$ constrain the parameters $\boldsymbol{p}$?
- Given two models $\mathcal{M}_1$ and $\mathcal{M}_2$, which one explains the data better?

$$\mathcal{P}(\boldsymbol{\theta}, \mathcal{M}|\boldsymbol{x}) = \frac{L(\boldsymbol{x}|\boldsymbol{\theta}, \mathcal{M})\pi(\boldsymbol{\theta})}{\epsilon(\boldsymbol{x}|\mathcal{M})} \tag{11}$$

Bayesian Inference

$$\mathcal{P}(\boldsymbol{\theta}, \mathcal{M}|\boldsymbol{x}) = \frac{L(\boldsymbol{x}|\boldsymbol{\theta}, \mathcal{M})\pi(\boldsymbol{\theta})}{\epsilon(\boldsymbol{x}|\mathcal{M})} \tag{12}$$

- $\mathcal{P}(\boldsymbol{\theta}, \mathcal{M}|\boldsymbol{x})$: the posterior.
- $L(\boldsymbol{x}|\boldsymbol{\theta}, \mathcal{M})$: the likelihood.
- $\pi(\boldsymbol{\theta})$: the priors.
- $\epsilon(\boldsymbol{x}|\mathcal{M})$: the evidence ('marginal likelihood').

→ Even for a known likelihood, the posterior can be difficult to obtain. Need sampling techniques (MCMC, Gibbs...) & DALI.