Probability distribution functions

Everything there is to know about random variables

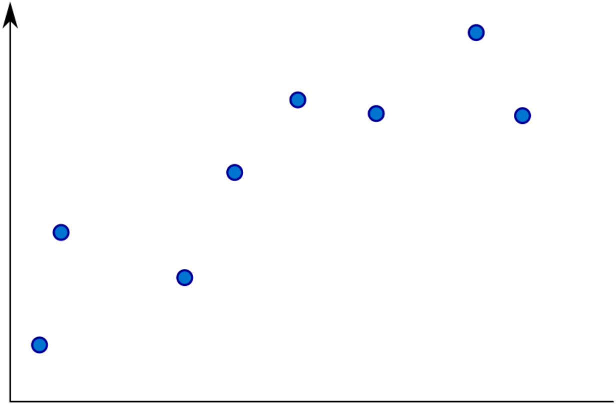
Imperial Data Analysis Workshop (2018)

Elena Sellentin

Sterrewacht
Universiteit Leiden, NL

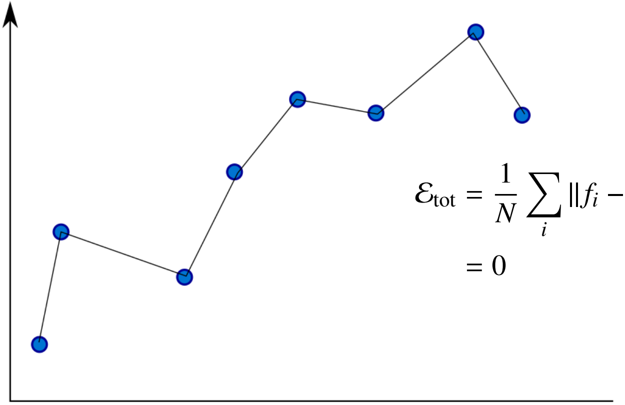
Cross validation

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



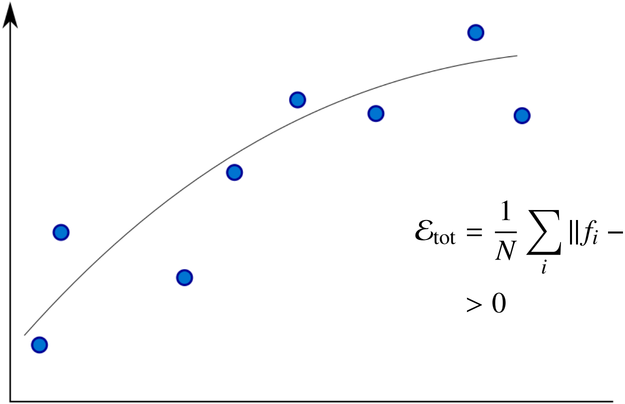
Cross validation

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



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

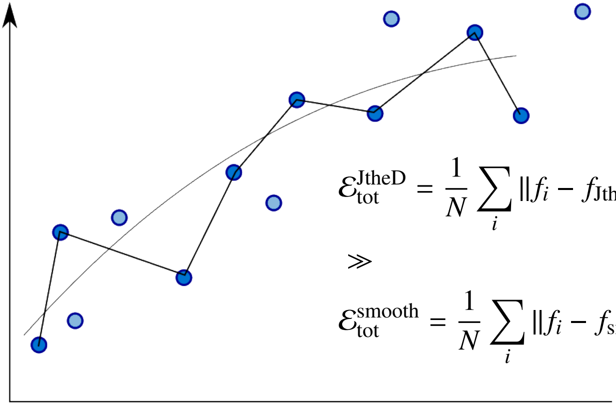
Cross validation



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

$$> 0$$

Cross validation

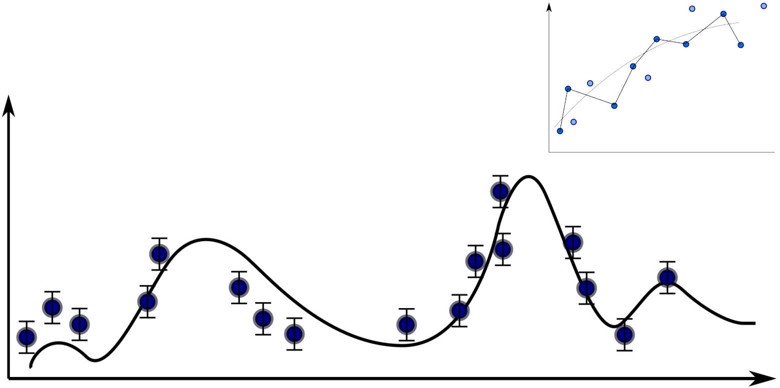


$$\mathcal{E}_{\text{tot}}^{\text{JtheD}} = \frac{1}{N}\sum_i \|f_i - f_{\text{JtheD}}(\mathbf{x})\|_2^2$$

$$\gg$$

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

An improved distance



$$\chi^2 = (\mathbf{x} - \mu)^\top \mathsf{C}^{-1} (\mathbf{x} - \mu)$$

Cross validation

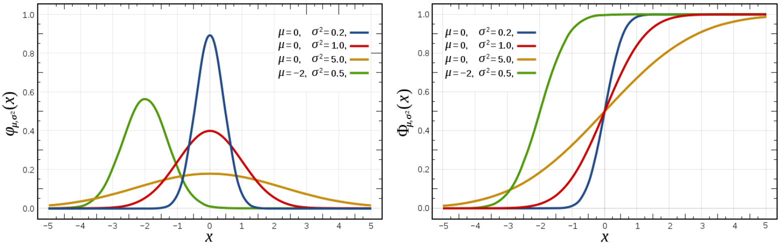
- We saw that join-the-dots 'perfectly' explains one data set, but then failed catastrophically on the repeated measurement.
- The fitted curve explained the first dataset somewhat 'worse', but then correctly predicted the outcome of the second measurement.

The best model minimizes a DISTANCE, e.g....

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

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

Probability distributions



- Left: Probability density functions $\mathcal{P}(x)$
- Right: Their cumulative distribution functions $C(x) = \int_{-\infty}^{x} \mathcal{P}(x')\mathrm{d}x'$.

Why moments lose information

What are...

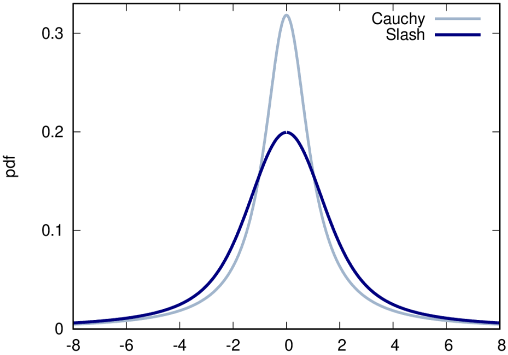
...mean, variance, skewness, curtosis, moments, cumulants?

The m-th moment is defined as

$$\langle x^m \rangle = \int x^m \, \mathrm{pdf}(x) \, \mathrm{d}x. \tag{2}$$

- Moments are scalars (or tensors).
- Pdfs are full functions ($\Rightarrow$ more information).
- From a pdf, moments can be computed. The inverse is not automatically true.

Advantages of pdfs



- For both: No mean, no variance, no skewness, no exc. curtosis, no moment-generating function.
- But the full pdfs exist!
- For some distribs, not even their zeroth moment (the normalization) exists $\Rightarrow$ improper priors.

Say you know $\mathcal{P}(x)$. But theory can only explain $y = f(x)$, some function of x.

$\mathcal{P}(x)$ contains all information about x. But how do you find $\mathcal{P}(y)$?

EXAMPLE:

$x = E$, the energy of particles (fermions and bosons) in a star.
Observable is $y = L$, the luminosity of the star.
Aim: To learn about the inner structure and composition of the star.
The function $f(x)$ is then 'nuclear physics + radiative transfer'.

Ways to find $\mathcal{P}(y)$

Case 1: y depends only on one random variable x.

$\Rightarrow$ Variable transformation.

Case 2: y depends on more random variables, e.g. u and v.

1. Product distribs: u, v independent, find $\mathcal{P}(y)$ for $y = uv$?
2. Ratio distributions: u, v independent, find $\mathcal{P}(y)$ for $y = u/v$?
3. Distributions of sums: $\mathcal{P}(y), y = \sum_i u_i$? $\Rightarrow$ convolutions.

Ways to find $\mathcal{P}(y)$: one r.v.

Analytically:

1. Via transformation of variables $\mathcal{P}(x)\mathrm{d}x = \mathcal{P}(y)\mathrm{d}y$.
2. Via the cumulative distribution function.
3. Via moment-generating functions ($\Rightarrow$ literature.)

Numerically:

1. Via sampling ($\Rightarrow$ Daniel Mortlock, Andrew Jaffe)

Ways to find $\mathcal{P}(y)$: many r.vs.

$y = f(u, v, x, ...)$, all random.
$\Rightarrow$ Finding $\mathcal{P}(y)$ always means
'From joint distribution of u, v, x... to distrib of y'.

1. What is $\mathcal{P}(u, v, x, ...)$?
2. If all are independent, then it is $\mathcal{P}(u)\mathcal{P}(v)\mathcal{P}(x)$....
3. If they are *not independent* $\Rightarrow$ (Bayesian) Hierarchical Models.
4. Marginalize over everything but y.

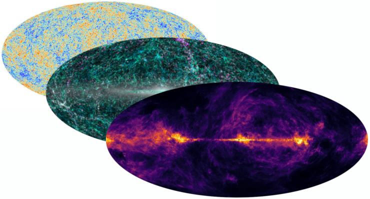
Maximum Likelihood estimation

- Common knowledge: Best fitting parameters $\leftrightarrow$ minimum χ^2
- Gaussian likelihood: $L(\mathbf{x}|\boldsymbol{\theta}) \propto \exp(-\chi^2/2)$
- $\Rightarrow$ Minimum-χ^2 is the maximum of the Gaussian likelihood
- If you don't have a *Gaussian* noise process? $\mathbf{x} \sim \mathcal{D}$ and $\mathcal{D} \neq \mathcal{G}$?

Maximum likelihood estimation is more general!

- $\mathbf{x}, \boldsymbol{\theta}, \mathcal{D}(\mathbf{x})$
- Then the likelihood is $L(\mathbf{x}|\boldsymbol{\theta}) = \mathcal{D}(\mathbf{x}|\boldsymbol{\theta})$
- The best-fitting params are then where $\nabla_{\boldsymbol{\theta}} L(\mathbf{x}|\boldsymbol{\theta}) = 0$

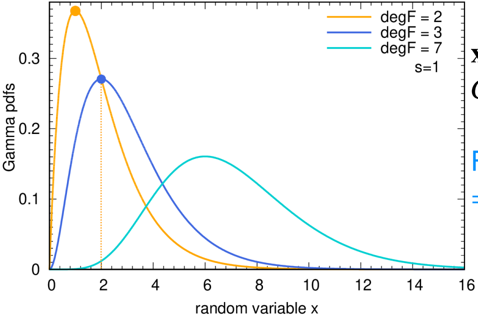
Power spectra



Maximum Likelihood:

Find $\boldsymbol{\theta}$ for which

$\nabla_\theta L(\mathbf{x}|\boldsymbol{\theta}) = 0$



$\mathbf{x} = \hat{\mathbf{C}}(l)$,

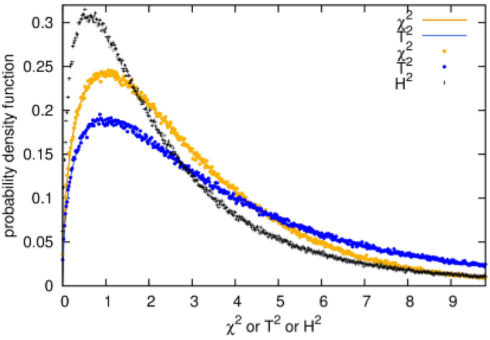
$\hat{C}(l) = \frac{1}{2\ell+1} \sum_m a_{lm} a^*_{lm}$

From $a_{lm} \sim \mathcal{G}(0, C_\ell)$

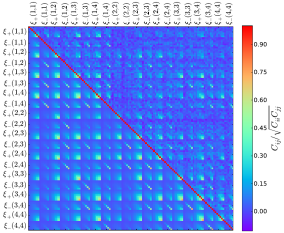
$\Rightarrow \hat{C}_\ell \sim \Gamma(\nu, C_\ell)$

Other common distributions

- Poisson distribution (galaxy clusters)
- Chi-squared and T^2-distributions;
  $\Rightarrow \chi^2/degF \approx 1$-test.



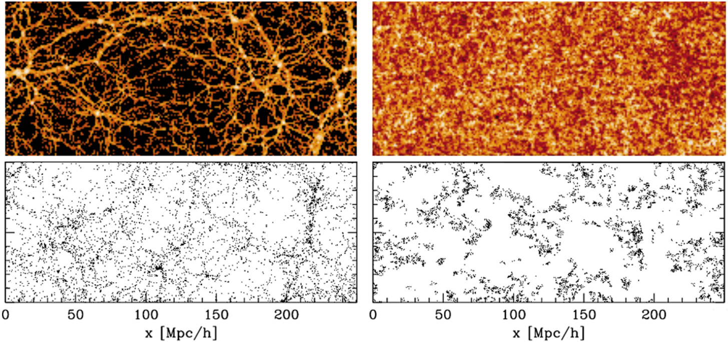
Sellentin & Heavens, MNRAS (2016)



Hildebrandt et al. (KiDS team)

Other common distributions

- Uniform distribution (Fourier phases)

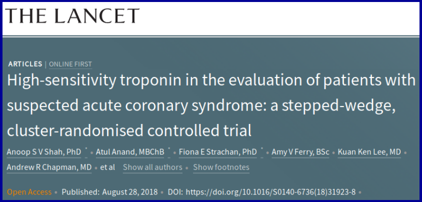


Coles & Chiang (2000) and Sefusatti & Scoccimarro (2005)

From likelihood to posterior

MNRAS, A&A,
JCAP, Phys. Rev.
astro in general:

$$\mathcal{P}(\boldsymbol{\theta}|\mathbf{x}) = \frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{x})}$$


THE LANCET

ARTICLES | ONLINE FIRST

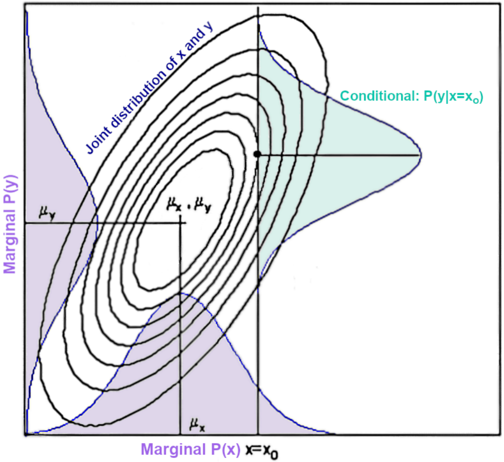
High-sensitivity troponin in the evaluation of patients with suspected acute coronary syndrome: a stepped-wedge, cluster-randomised controlled trial

Anoop S V Shah, PhD • Atul Anand, MBChB • Fiona E Strachan, PhD • Amy V Ferry, BSc • Kuan Ken Lee, MD • Andrew R Chapman, MD • et al Show all authors • Show footnotes

Open Access • Published: August 28, 2018 • DOI: https://doi.org/10.1016/S0140-6736(18)31923-8 •

- $\pi(\theta_1) = \delta_D(\theta_1)$ (Conditional distributions.)
- $\pi(\mathbf{x}) = \int L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}^n\theta$ (Evidence)
- $\mathcal{P}(\boldsymbol{\theta}|\mathbf{x})\pi(\mathbf{x}) = L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. $\Rightarrow$ $\pi(\mathbf{x})$: have you manipulated your data taking process? Is there a selection effect hiding somewhere?
- $\pi(\boldsymbol{\theta})$ does not necessarily need to be normalizable: “Improper priors” (still need a proper posterior).

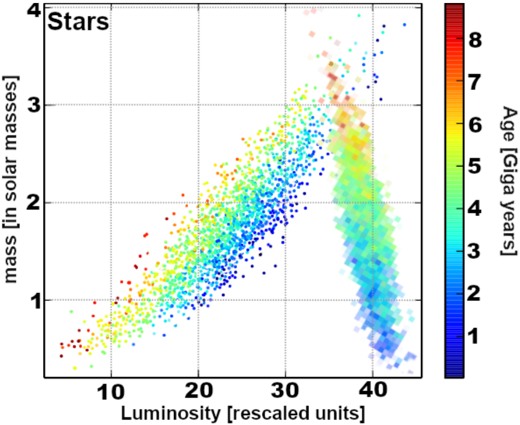
Multivariate distributions



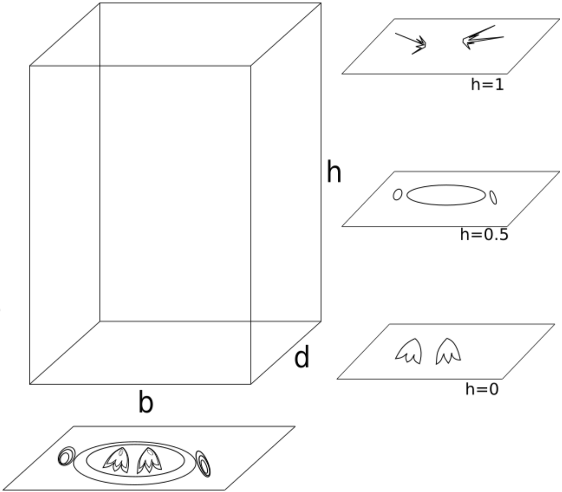
→ Reducing the dimensionality: (1) taking conditionals and
(2) taking marginals

Conditional distributions of stars

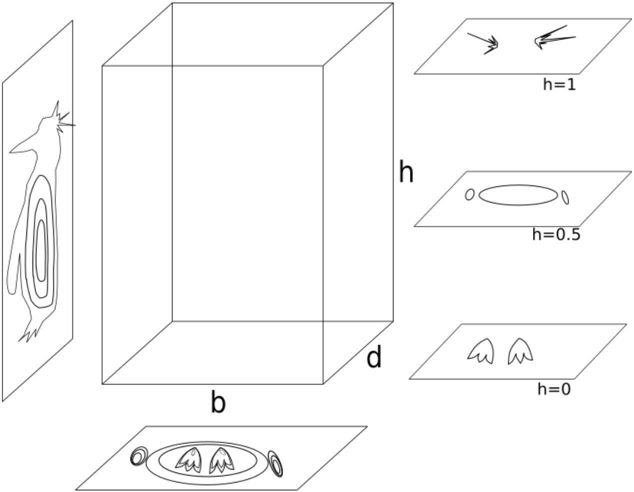
- Inferred variables: stellar mass, (absolute) luminosity, age, stellar population-type (A or B, say)



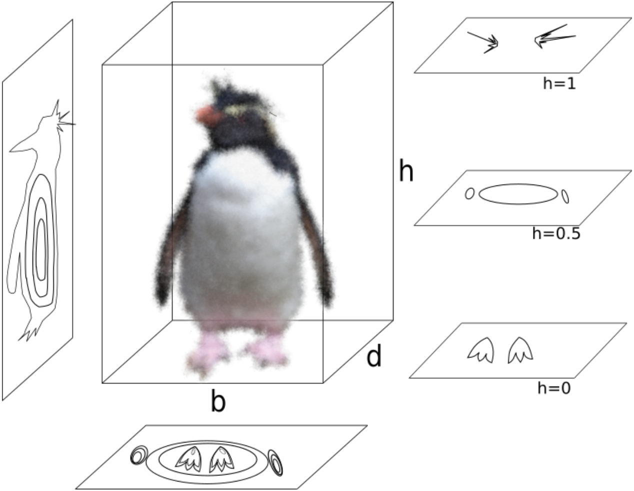
Multivariate projections



Multivariate projections



Multivariate penguin



Cosmological parameters from Planck

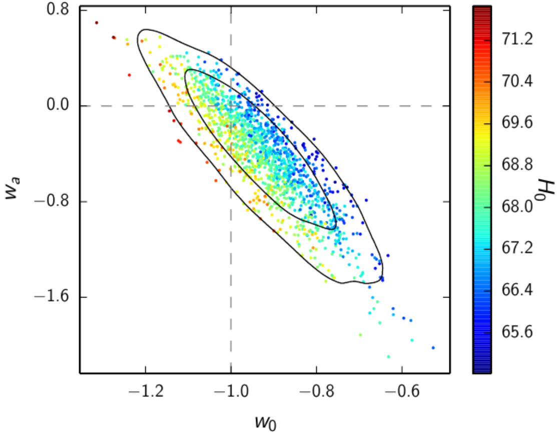


Fig. 27. Samples from the distribution of the dark energy parameters w_0 and w_a using *Planck* TT+lowP+BAO+JLA data, colour-coded by the value of the Hubble parameter H_0. Contours show the corresponding 68 % and 95 % limits. Dashed grey lines intersect at the point in parameter space corresponding to a cos-

An easy (non-astronomical) example

Alex

Data: Alex is British; The average number of children a British woman gives birth to is 1.8.

Model: $C \sim \text{Poisson}(C; \lambda = 1.8)$ (Inspired by kids being integers.)

- Abbreviate: Number of children as C, Alex as A.
- What does $\mathcal{P}(C|A)$ express?
- What does $\int C\, \mathcal{P}(C|A)\mathrm{d}C$ express?
- What is $\int C\, \mathcal{P}(C|A)\mathrm{d}C$ in numbers?

Hint: The Poisson distribution is $\text{Poisson}(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$.
Careful: It looks smooth but has only integer support, since x is the integer-valued random variate.

Alex

Does Alex abbreviate Alexander or Alexandra?

- New variable: $G = \{XX, XY\}$.
- What does $\pi(XX|A)$ express? What could it numerically be?
- What does $\pi(A|XX)$ express? What could it numerically be?
- What does $\pi(A|XX) = 0$ imply?

Alex

Does Alex abbreviate Alexander or Alexandra?

For $\pi(A|XX) = 0$, what is the meaning and the numerical value of $\int C\, \mathcal{P}(C|A)\pi(A|XX)\mathrm{d}C$?