On Model Selection in Cosmology

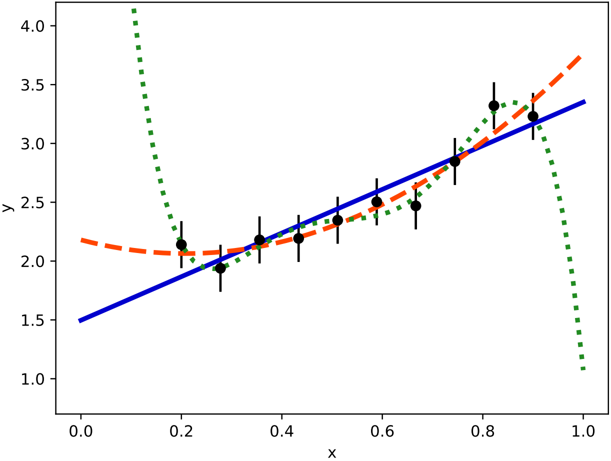
Martin Kerscher

Fakultät für Physik
Ludwig–Maximilians–Universität München

Bayes Forum — May 3, 2019

M.Kerscher, J.Weller: `https://arxiv.org/abs/1901.07726`

Numquam ponenda est pluralitas sine necessitate.



$$f(x, \boldsymbol{\theta}) = \theta_0 + \theta_1 x$$

$$g(x, \boldsymbol{\phi}) = \phi_0 + \phi_1 x + \phi_2 x^2$$

$$h(x, \boldsymbol{\psi}) = \psi_0 + \psi_1 x + \psi_2 x^2 + \psi_3 x^3 + \psi_4 x^4 + \psi_5 x^5$$

Parameter estimation (briefly)

Model selection

- Goodness of fit
- Likelihood ratio test
- Bayesian model comparison
- Information theoretic approach: classical and Bayesian

The expansion history of the Universe

Discussion

Parameter estimation

- Measurements $\boldsymbol{d} = (x_i, y_i)_{i=1}^N$.
- Model $f(x, \boldsymbol{\theta})$ with parameters $\boldsymbol{\theta} \in A \subset \mathbb{R}^K$.
- Objective: determine $\boldsymbol{\theta}^*$ such that $f(x, \boldsymbol{\theta}^*)$ is the *best* approximation to the data $\boldsymbol{d}$, such that $f(x_i, \boldsymbol{\theta}^*)$ is approximation y_i.

All methods start with the likelihood – here a Gaussian likelihood:

$$p_f(\boldsymbol{d} \,|\, \boldsymbol{\theta}) = \Big((2\pi)^N \det(\Sigma)\Big)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{y} - \boldsymbol{f})^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{f})\right)$$

$\boldsymbol{y} = (y_1, \ldots, y_N)^T$, $\boldsymbol{f} = (f(x_1, \boldsymbol{\theta}), \ldots, f(x_N, \boldsymbol{\theta}))^T$, covariance matrix Σ.

Maximum likelihood and least square

- **Maximum likelihood:** Find $\boldsymbol{\theta}^*$ maximising $p_f(\boldsymbol{d} \mid \boldsymbol{\theta}^*)$.

  In choosing the parameter $\boldsymbol{\theta}^*$, the data points become the most probable data points given the model f.

- **Least square:** simplified maximum likelihood with Gaussian likelihood and diagonal Σ.

  Searching for maximum of the log-likelihood

  $$\log p_f(\boldsymbol{d} \mid \boldsymbol{\theta}) = \text{const.} - \tfrac{1}{2} \sum_{i=1}^{N} \frac{(y_i - f(x_i, \boldsymbol{\theta}))^2}{\sigma_i^2}$$

  is equivalent to searching for the minimum of

  $$\chi_f^2 = \sum_{i=1}^{N} \frac{(y_i - f(x_i, \boldsymbol{\theta}))^2}{\sigma_i^2}.$$

Bayesian parameter estimation

- **Prior distribution** $p_f(\boldsymbol{\theta})$ of the parameters for model $f(x, \boldsymbol{\theta})$.
- **Posterior distribution** $p_f(\boldsymbol{\theta} \mid \boldsymbol{d})$ from Bayes theorem

$$p_f(\boldsymbol{\theta} \mid \boldsymbol{d}) = \frac{p_f(\boldsymbol{d} \mid \boldsymbol{\theta}) \; p_f(\boldsymbol{\theta})}{p_f(\boldsymbol{d})}$$

- **Evidence** (or marginal likelihood):

$$p_f(\boldsymbol{d}) = \int p_f(\boldsymbol{d} \mid \boldsymbol{\theta}) \; p_f(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}.$$

- Maximum posterior estimate (MAP): $\boldsymbol{\theta}^*$ maximising $p_f(\boldsymbol{\theta}^* \mid \boldsymbol{d})$.

Model selection

- Consider two models

$$f(x, \boldsymbol{\theta}) \text{ with parameters } \boldsymbol{\theta} \in A \subset \mathbb{R}^K,$$
$$g(x, \phi) \text{ with parameters } \phi \in B \subset \mathbb{R}^L.$$

- Determine $\boldsymbol{\theta}^*$ and ϕ^* (with your favourite method).
- Which one is better, $f(x, \boldsymbol{\theta}^*)$ or $g(x, \phi^*)$?

Goodness of fit

- With the best fit parameters $\boldsymbol{\theta}^*$ calculate

$$\chi_f^2 = \sum_{i=1}^{N} \frac{(y_i - f(x_i, \boldsymbol{\theta}^*))^2}{\sigma_i^2}$$

- Reduced–χ^2

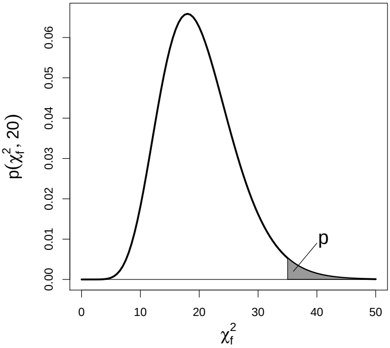
$$\chi_{f,\mathrm{red}}^2 = \frac{\chi_f^2}{n_{\mathrm{df}}}$$

typically $n_{\mathrm{df}} = N - K$.

- $\chi_{f,\mathrm{red}}^2 \approx 1$ a good fit,
- $\chi_{f,\mathrm{red}}^2 > 1$ a bad fit, and
- $\chi_{f,\mathrm{red}}^2 < 1$ an overfit.

Hypothesis test and p–value

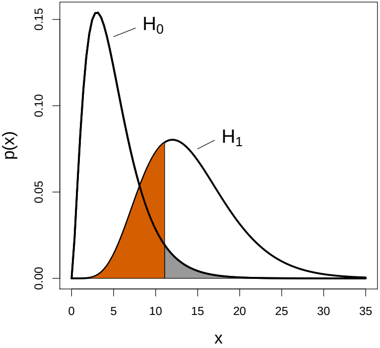
N data points independent and with Gaussian error, then χ_f^2 follows χ^2–distribution with $N-1$ degrees of freedom.



- p–value: $p = 1 - G_{N-1}\left(\chi_f^2\right)$ indicates how incompatible the data are with our model (the null hypothesis)
- Level of significance α, often $\alpha = 0.05$.
- If $p < \alpha$ our model, the null-hypothesis gets rejected.

Given the model (the null-hypothesis), a small p-value allows us to reject the model but we do not learn anything about the false negative rate.

Error of the first and of the second kind



- Our model H_0 and the alternative model H_1
- H_0 is **not** rejected ("accepted") at the significance level $\alpha = 0.05$.
- Assume H_1 is true, then the false negative rate is $\beta = 0.32$.

α type I error, false positive rate (grey)
β type II error, false negative rate (red)

Likelihood ratio test

- Nested models: f is a special case of g (i.e. $A \subsetneq B$ and $f|_A \equiv g|_A$).

  Null hypothesis: "f is the true model with $\boldsymbol{\theta}^\star \in A$"
  Alternative hypothesis: "g is the true model with $\boldsymbol{\phi}^\star \in B$"

- Likelihood ratio

$$L = \frac{p_f(\boldsymbol{d} \,|\, \boldsymbol{\theta}^*)}{p_g(\boldsymbol{d} \,|\, \boldsymbol{\phi}^*)}$$

  - large samples: $-2 \log L$ is χ^2-distributed with d.f. $\nu = \dim B - \dim A$.
  - $p = 1 - G_\nu(-2 \log L)$, discard null–hypothesis if $p < \alpha$ with significance level α

- Neyman-Pearson Lemma: the test based on the likelihood ratio is minimising the false negative rate.

Bayesian model comparison

- Probability of model f *and* data $\boldsymbol{d}$: $\quad p(f \text{ and } \boldsymbol{d}) = p_f(\boldsymbol{d})\ \pi_f$
  similarly for model g: $\quad p(g \text{ and } \boldsymbol{d}) = p_g(\boldsymbol{d})\ \pi_g$

  π_f and π_g prior probabilities of our models.

- Bayes factor

  $$\frac{p(f \text{ and } \boldsymbol{d})}{p(g \text{ and } \boldsymbol{d})} = \frac{p_f(\boldsymbol{d})}{p_g(\boldsymbol{d})} \equiv B_{fg}$$

  (assuming $\pi_f = \pi_g$).

- $B_{fg} > 1$ then favour model f over model g
  (compare Jeffreys' scale).

- Evidence (marginal Likelihood)

  $$p_f(\boldsymbol{d}) = \int p_f(\boldsymbol{d} \mid \boldsymbol{\theta})\ p_f(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$$

  (nested sampling, Chib's method, population MC, direct integration)

- BIC: asymptotic for large N
  (Schwarz 1978)

Information theoretic approach

- Compare the predictive distribution $p_{p,f}(d)$ to the *true* probability $p_T(d)$ using **Kulback–Leibler** (KL) divergence (relative entropy)

$$
\begin{aligned}
D(p_T \,|\, p_{p,f}) &= \int p_T(d) \log \frac{p_T(d)}{p_{p,f}(d)} \, \mathrm{d}d \\
&= \underbrace{\int p_T(d) \log p_T(d) \, \mathrm{d}d}_{\text{independent of model } f} - \int p_T(d) \log p_{p,f}(d) \, \mathrm{d}d
\end{aligned}
$$

- Classical approach: use **predictive likelihood** (marginalised Likelihood):

$$
p_{p,f}(d_i) \equiv p_f(d_i \,|\, \boldsymbol{\theta}) = \int p_f(\boldsymbol{d} \,|\, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{d}_{[i]}
$$

with $\boldsymbol{d}_{[i]} = ((x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \ldots, (x_N, y_N))^T$.

Classical information theoretic approach I

- compare predictive likelihood $p_f(d \mid \boldsymbol{\theta}^*)$ with parameter $\boldsymbol{\theta}^*$ to true distribution $p_T(d)$:

$$D(p_T \mid p_f) = \underbrace{\text{const}}_{\text{independent from f}} - \underbrace{\int p_T(d) \log p_f(d \mid \boldsymbol{\theta}^*) \, \mathrm{d}d}_{\eta(f)}$$

- The expected log likelihood:

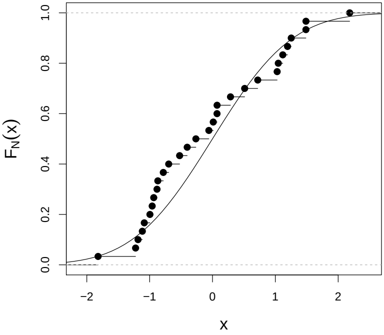
$$\eta(f) = \int \log p_f(d \mid \boldsymbol{\theta}^*) \, \mathrm{d}F_T(d),$$

- Expectation over the true distribution F_T — **not available**.

Insertion: Empirical distribution function

Cumulative distribution function

$$F(x) = \int_{-\infty}^{x} p(x')\mathrm{d}x'$$



- $\{x_1, \ldots, x_N\}$ i.i.d. from F.
- Empirical distribution function

$$F_N(x) = \frac{1}{N}\sum_{i=1}^{N} I_{[x_i,\infty)}(x)$$

- Glivenko-Cantelli:
  $\|F_N - F\|_\infty \to 0 \quad$ uniformly

Classical information theoretic approach II

$$D(p_T \mid p_f) = \mathrm{const}(f) - \underbrace{\int \log p_f(d \mid \boldsymbol{\theta}^*) \, \mathrm{d}F_T(d)}_{\eta(f)}$$

- Estimate $\eta(f)$ by replacing $F_T(d)$ with $F_{T,N}(d)$

$$\widehat{\eta}(f) = \int \log p_f(d \mid \boldsymbol{\theta}^*) \, \mathrm{d}F_{T,N}(d) = \frac{1}{N} \sum_{i=1}^{N} \log p_f(d_i \mid \boldsymbol{\theta}^*)$$

- However $\widehat{\eta}(f)$ is biased ($\boldsymbol{\theta}^*$ is point estimate) with

$$b(f) = \int \left(\widehat{\eta}(f) - \eta(f) \right) \mathrm{d}F_T.$$

AIC and beyond

- $b(f)$ asymptotically goes like K/N (Akaike, 1972)

$$\mathrm{AIC}(f) \equiv -2N\,(\widehat{\eta} - K/N) = -2\sum_{i=1}^{N} \log p_f(d_i \,|\, \boldsymbol{\theta}^*) + 2\,K,$$

  the model with a smaller value of the AIC is favoured.

- Estimate $\widetilde{b}(f)$ using bootstrap, then

$$\mathrm{EIC}(f) \equiv -2N\left(\widehat{\eta}(f) - \widetilde{b}(f)\right),$$

  the model with a smaller value of the EIC is favoured.

Bayesian information theoretic approach

Posterior predictive distribution:

$$p_{\mathrm{ppd},f}(d) = \int p_f(d \,|\, \boldsymbol{\theta})\, p_f(\boldsymbol{\theta} \,|\, \boldsymbol{d})\, \mathrm{d}\boldsymbol{\theta}.$$

Compare $p_{\mathrm{ppd},f}(d)$ to true $p_T(d)$ with KL-divergence

$$D(p_T \,|\, p_{\mathrm{ppd},f}) = \mathrm{const} - \underbrace{\int \log p_{\mathrm{ppd},f}(d)\, \mathrm{d}F_T(d)}_{\kappa(f)}$$

- Estimate $\kappa(f)$ using empirical distribution $F_{T,N}$

$$\widehat{\kappa}(f) = \frac{1}{N} \sum_{i=1}^{N} \log \Big(\underbrace{\int p_f(d_i \,|\, \boldsymbol{\theta})\, p_f(\boldsymbol{\theta} \,|\, \boldsymbol{d})\, \mathrm{d}\boldsymbol{\theta}}_{\text{estimate from the MC chain}} \Big)$$

- **Bayesian Predictive Information Criterium**

$$\mathrm{BPIC}(f) \equiv -2N\, \widehat{\kappa}(f)$$

The methods

- goodness of fit: χ^2_{red}, p–value
- likelihood ratio test: p–value
- Bayesian: Bayes factor and BIC
- Classical information theoretic approach: AIC, EIC
- Bayesian information theoretic approach: BPIC

The expansion history of the Universe

- SN Ia's: redshift z and magnitude $\rightarrow$ distance modulus μ
  Union 2.1 (Suzuki et al. 2011) $N = 580$ data points $d_i = (z_i, \mu_i)$
- Distance modulus – redshift relation

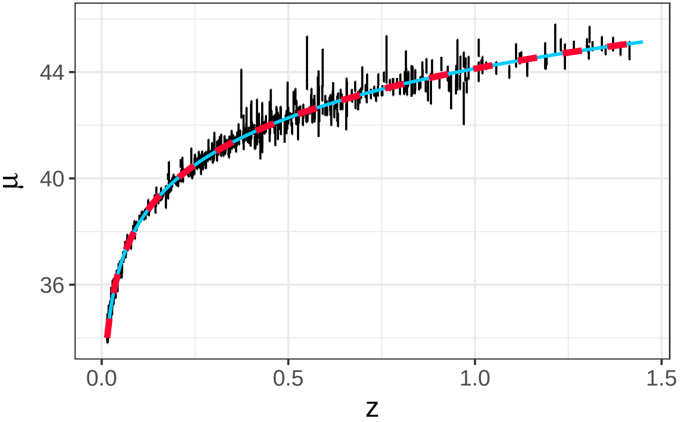
$$\mu(z, \boldsymbol{\theta}) = 5 \log_{10} d_L(z, \boldsymbol{\theta}) + 25$$

  model dependence through luminosity distance $d_L(z, \boldsymbol{\theta})$.
- ΛCDM (with $\Omega_\Lambda = 1 - \Omega_m$)

$$d_L(z, \Omega_m) = d_H\,(1+z) \int_o^z \frac{\mathrm{d}z'}{\sqrt{\Omega_m(1+z')^3 + \Omega_\Lambda}} \ ,$$

- wCDM model (with constant e.o.s. $p = w\varrho$)

$$d_L(z, \Omega_m, w) = d_H\,(1+z) \int_o^z \frac{\mathrm{d}z'}{\sqrt{\Omega_m(1+z')^3 + \Omega_\Lambda(1+z')^{3(1+w)}}} \ .$$



- ΛCDM (blue): $\Omega_m = 0.278 \pm 0.007$
- wCDM (red): $\Omega_m = 0.279 \pm 0.06$ and $w = -1.0 \pm 0.13$

Goodness of fit:

$\chi^2_{\text{red}}(\Lambda) = 0.971,\ p_\Lambda = 0.68$ $\qquad$ $\chi^2_{\text{red}}(w) = 0.973,\ p_w = 0.67$

Likelihood ratio test:

$p = 1 - G_1(-2\log L) = 0.975$

Bayesian approach:

$\text{BIC}_\Lambda = -231.1$ $\qquad$ $\text{BIC}_w = -224.8$

$B_{\Lambda w} = \frac{p_\Lambda(\boldsymbol{d})}{p_w(\boldsymbol{d})} = 5.45$ substantial evidence for ΛCDM (Jeffreys' scale)

Classical information theoretic approach:

$\text{AIC}_\Lambda = -235.5$ $\qquad$ $\text{AIC}_w = -233.5$

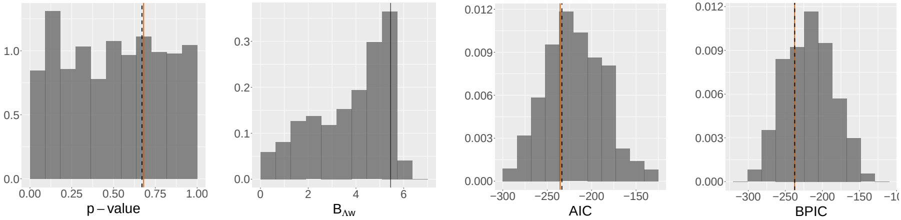
$\text{EIC}_\Lambda = -239.2$ $\qquad$ $\text{EIC}_w = -241.0$

Bayesian information theoretic approach:

$\text{BPIC}_\Lambda = -237.5$ $\qquad$ $\text{BPIC}_w = -237.3$

Is there a substantial difference?

Generate random artificial data sets assuming ΛCDM:



midspread:

$\Delta_{p_\Lambda} = 0.46$, $\Delta_{\mathrm{AIC}_\Lambda} = 42$, $\Delta_{\widetilde{\eta}_\Lambda} = 0.034$, $\Delta_{B_{\Lambda w}} = 3.4$, $\Delta_{\mathrm{BIC}_\Lambda} = 42$, $\Delta_{\mathrm{BPIC}_\Lambda} = 43$

Using the Union 2.1 data set we cannot decide wether the ΛCDM or the wCDM model should be preferred.

Summary of the methods

- With the goodness of fit you rank models according to their ability to fit the data points.
- The likelihood ratio allows you to discard a given model (your null hypothesis) in favour of the alternative model.
- In a Bayesian model comparison you use the evidence ratio to compare the joint probabilities of the models and the data.
- In the classical information theoretic approach you measure how good the best fitting models are at predicting new data.
- In the Bayesian information theoretic approach you measure how good the posterior predictive distributions of the models are at predicting new data.

Discussion

- Two questions $\cdots$
  - Which model, with sufficient data, can be identified as the true model?
  - Based on the data, which model lies closest to the true model?

  $\cdots$ two answers?

- Both likelihood ratio test and the Bayesian model selection go for the *truth*: Either you discard the false models via tests, or you determine the most probable model.
- With the information theoretic approach one tries to identify the model which is closest to the true distribution and most effective in predictions.

- **BUT:** All models are wrong (G. Box).

AddOn

Bootstrap for $b(f)$

- Explicating the dependence on the data $\boldsymbol{d}$

$$b(f) = \mathbb{E}_{F_T}\left[\eta\left(f;\boldsymbol{\theta}^\star(\boldsymbol{d}), F_T\right) - \eta\left(f;\boldsymbol{\theta}^\star(\boldsymbol{d}), F_{T,N,\boldsymbol{d}}\right)\right].$$

- generate bootstrap samples $\widetilde{\boldsymbol{d}} = (\tilde{x}_i, \tilde{y}_i)_{i=1}^N$ from the data by repeatedly drawing from $\boldsymbol{d}$ with putting back (sampling from $F_{T,N,\boldsymbol{d}}$).
- With bootstrap samples $\widetilde{\boldsymbol{d}}$ from $\boldsymbol{y}$ estimate $b(f)$:

$$\widetilde{b}(f) = \mathbb{E}_{T,N,\boldsymbol{d}}\left[\eta\left(f;\boldsymbol{\theta}^\star(\widetilde{\boldsymbol{d}}), F_{T,N,\boldsymbol{d}}\right) - \eta\left(f;\boldsymbol{\theta}^\star(\widetilde{\boldsymbol{d}}), F_{T,N,\widetilde{\boldsymbol{d}}}\right)\right].$$

- Again an asymptotic result, variance reduction possible.

Priors

- subjective priors
- objective priors, non-informative priors, reference priors
- priors suggested by preceding results (regress?)

- Two models for generating mocks:
  - $\Omega_m = 0.278, w = -1$ (red)
  - $\Omega_m = 0.171, w = -0.8$ (blue)
- For each mock calculate

$$\Delta\mathrm{AIC} = \mathrm{AIC}_\Lambda - \mathrm{AIC}_w, \qquad \Delta\mathrm{BPIC} = \mathrm{BPIC}_\Lambda - \mathrm{BPIC}_w.$$

- Empirical distribution