Bayesian Learning

Lecture 11 - Bayesian model comparison

Mattias Villani

Department of Statistics
Stockholm University

mattiasvillani.com @matvil @matvil mattiasvillani

Overview

- **Bayesian model comparison**
- **Marginal likelihood**

Using likelihood for model comparison

- Consider two models for the data $\mathbf{y} = (y_1, ..., y_n)$: M_1 and M_2.
- Let $p(\mathbf{y}|\theta_k, M_k)$ denote the data density under model M_k.
- If we know θ_1 and θ_2, the **likelihood ratio** is useful

$$\frac{p(\mathbf{y}|\theta_1, M_1)}{p(\mathbf{y}|\theta_2, M_2)}.$$

- The **likelihood ratio** with **ML estimates** plugged in:

$$\frac{p(\mathbf{y}|\hat{\theta}_1, M_1)}{p(\mathbf{y}|\hat{\theta}_2, M_2)}.$$

- Bigger models always win in estimated likelihood ratio.
- **Hypothesis tests** are problematic for non-nested models. End results are not very useful for analysis.

Bayesian model comparison

- **Posterior model probabilities**

$$\underbrace{\Pr(M_k|\mathbf{y})}_{\text{posterior model prob.}} \propto \underbrace{p(\mathbf{y}|M_k)}_{\text{marginal likelihood}} \cdot \underbrace{\Pr(M_k)}_{\text{prior model prob.}}$$

- The **marginal likelihood** for model M_k with parameters θ_k

$$\underbrace{p(\mathbf{y}|M_k)} = \int p(\boldsymbol{y}|\theta_k, M_k)p(\theta_k|M_k)d\theta_k.$$

- θ_k is 'removed' by the averaging wrt prior. **Priors matter!**
- The **Bayes factor**

$$B_{12}(\boldsymbol{y}) = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)}$$

Jeffreys scale of evidence for the Bayes factor

- Barely worth mentioning: $1 < \mathrm{BF} \leq 3$
- Positive: $3 < \mathrm{BF} \leq 20$
- Strong: $20 < \mathrm{BF} \leq 150$
- Very strong: > 150

Bayesian hypothesis testing - Bernoulli

- **Hypothesis testing** is just a special case of model selection:

$$M_0 : x_1, ..., x_n \overset{iid}{\sim} \mathrm{Bernoulli}(\theta_0)$$

$$M_1 : x_1, ..., x_n \overset{iid}{\sim} \mathrm{Bernoulli}(\theta), \theta \sim \mathrm{Beta}(\alpha, \beta)$$

$$p(x_1, ..., x_n | M_0) = \theta_0^s (1-\theta_0)^f,$$

$$\begin{aligned} p(x_1, ..., x_n | M_1) &= \int_0^1 \theta^s (1-\theta)^f B(\alpha, \beta)^{-1} \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta \\ &= B(\alpha+s, \beta+f)/B(\alpha, \beta), \end{aligned}$$

where $B(\cdot, \cdot)$ is the Beta function.

- **Posterior model probabilities**

$$\Pr(M_k | x_1, ..., x_n) \propto p(x_1, ..., x_n | M_k) \Pr(M_k), \text{ for } k = 0, 1.$$

- The **Bayes factor**

$$\mathrm{BF}(M_0; M_1) = \frac{p(x_1, ..., x_n | M_0)}{p(x_1, ..., x_n | M_1)} = \frac{\theta_0^s (1-\theta_0)^f B(\alpha, \beta)}{B(\alpha+s, \beta+f)}.$$

Normal example

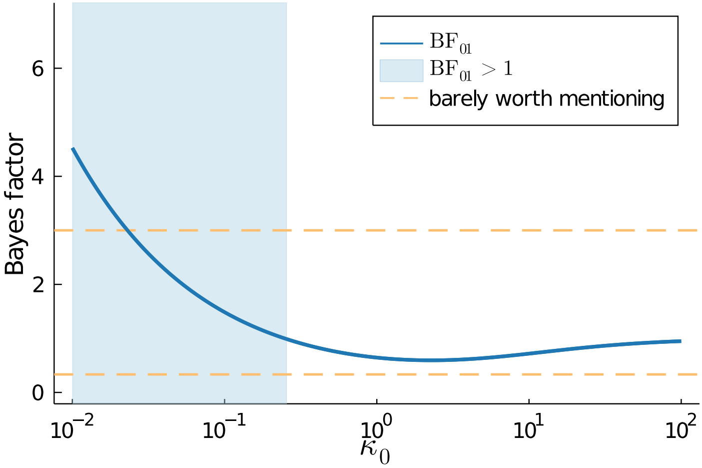
- **Model**: $x_1, \ldots, x_n \overset{\text{iid}}{\sim} N(\theta, \sigma^2)$, σ^2 known.
- **Prior**: $\theta \sim N(\mu_0, \sigma^2/\kappa_0)$.
- **Likelihood**: $\bar{x}$ is **sufficient** for θ and $\bar{x}|\theta \sim N(\theta, \sigma^2/n)$.
- **Marginal likelihood**: $p(\bar{x}|M_1) = N\left(\mu_0, \sigma^2(1/n + 1/\kappa_0)\right)$.
- Testing a **sharp null**: $M_0 : \theta = \mu_0$ vs $M_1 : \theta \neq \mu_0$.

$$B_{01} = \frac{p(\bar{x}|M_0)}{p(\bar{x}|M_1)} = \frac{N\left(\bar{x}|\mu_0, \sigma^2/n\right)}{N\left(\bar{x}|\mu_0, \sigma^2(1/n + 1/\kappa_0)\right)}$$

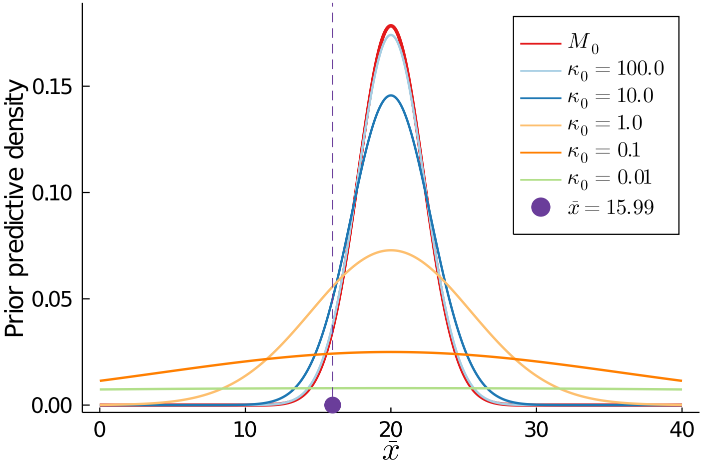
$$\log \frac{p(\bar{x}|M_0)}{p(\bar{x}|M_1)} = -\frac{1}{2}\log\left(\frac{\kappa_0}{\kappa_0 + n}\right) - \frac{n(\bar{x} - \mu_0)^2}{2\sigma^2}\left(\frac{n}{\kappa_0 + n}\right)$$

- $\kappa_0 \to \infty$ then $B_{01} \to 1$ (prior under M_1 is a point mass at 0)
- $\kappa_0 \to 0$ then $B_{01} \to \infty$ ($p(\bar{x}|M_1)$ is average $p(\bar{x}|\theta)$ wrt prior)

Internet speed data - Bayes factor



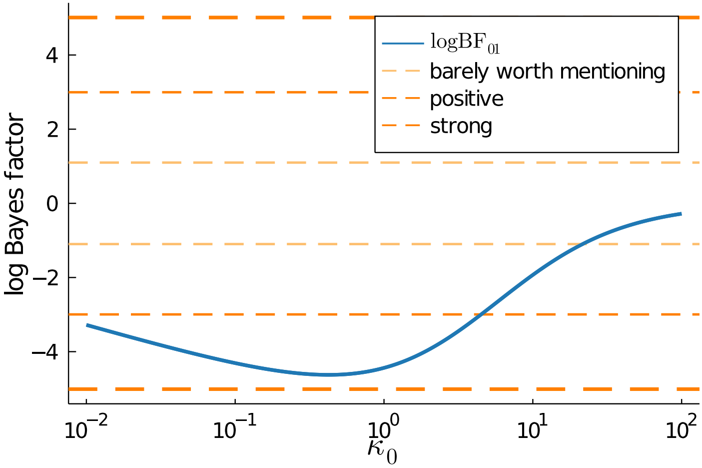
Internet speed data - prior predictive density



Vague priors for marginal likelihoods is a bad idea

- Smaller models always win when priors are very vague.
- **Improper priors** cannot be used for model comparison. ⚠

Internet speed data with $\bar{x} = 12$



Example: Geometric vs Poisson

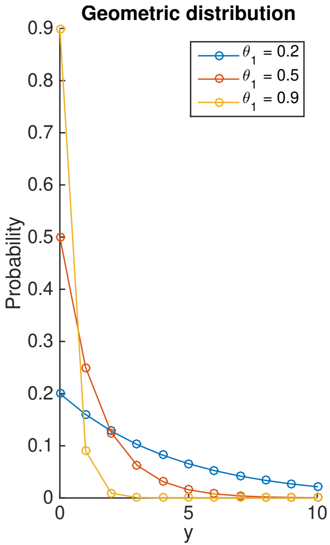
- Model 1 - **Geometric** with Beta prior:
  - $y_1, ..., y_n|\theta_1 \sim \mathrm{Geo}(\theta_1)$
  - $\theta_1 \sim \mathrm{Beta}(\alpha_1, \beta_1)$
- Model 2 - **Poisson** with Gamma prior:
  - $y_1, ..., y_n|\theta_2 \sim \mathrm{Poisson}(\theta_2)$
  - $\theta_2 \sim \mathrm{Gamma}(\alpha_2, \beta_2)$
- **Marginal likelihood** for M_1

$$
\begin{aligned}
p(y_1, ..., y_n|M_1) &= \int p(y_1, ..., y_n|\theta_1, M_1)p(\theta_1|M_1)d\theta_1 \\
&= \frac{\Gamma\left(\alpha_1+\beta_1\right)}{\Gamma\left(\alpha_1\right)\Gamma\left(\beta_1\right)} \frac{\Gamma\left(n+\alpha_1\right)\Gamma\left(n\bar{y}+\beta_1\right)}{\Gamma\left(n+n\bar{y}+\alpha_1+\beta_1\right)}
\end{aligned}
$$

- **Marginal likelihood** for M_2

$$
p(y_1, ..., y_n|M_2) = \frac{\Gamma(n\bar{y}+\alpha_2)\beta_2^{\alpha_2}}{\Gamma(\alpha_2)(n+\beta_2)^{n\bar{y}+\alpha_2}} \frac{1}{\prod_{i=1}^n y_i!}
$$

Geometric and Poisson



Geometric vs Poisson

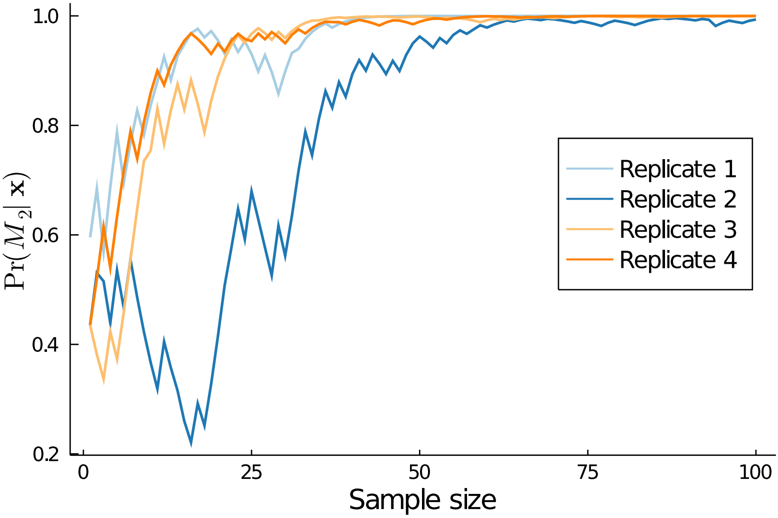
- Use priors to match prior predictive means:

$$E(y|M_1) = E(y|M_2) \quad \iff \quad \alpha_1\alpha_2 = \beta_1\beta_2$$

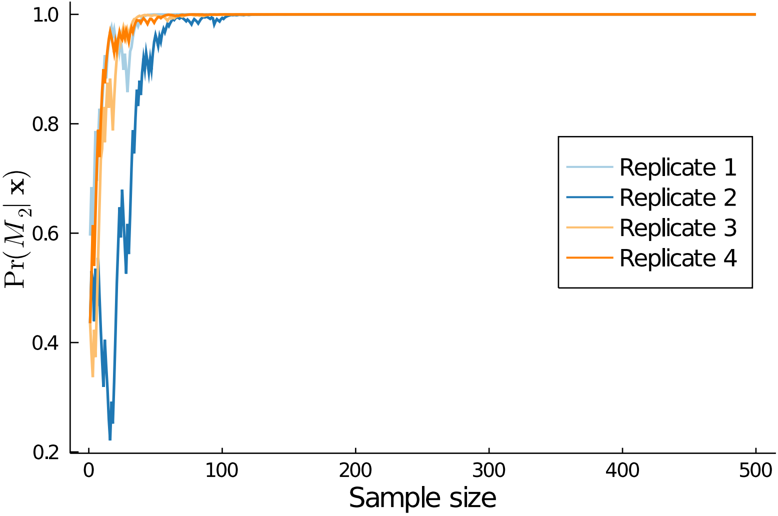
- Geometric model: $\alpha_1 = 10, \beta_1 = 20$.
- Poisson model: $\alpha_2 = 20, \beta_2 = 10$.

	$y_1 = 0,\ y_2 = 0$	$y_1 = 3,\ y_2 = 3.$
BF_{12}	4.54	0.29
$\Pr(M_1|\mathbf{y})$	0.82	0.22
$\Pr(M_2|\mathbf{y})$	0.18	0.78

Geometric vs Poisson for Pois(1) data



Geometric vs Poisson for Pois(1) data



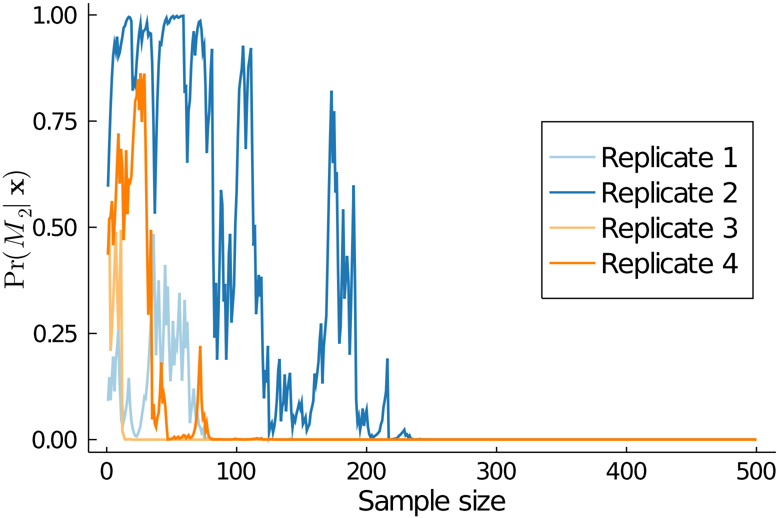
Asymptotic properties of marginal likelihood

- Set of compared models: $\mathcal{M} = \{M_1, ..., M_K\}$.
- **$\mathcal{M}$-closed**: data generating process $M^\star$ is in $\mathcal{M}$.
- $\mathcal{M}$-closed **consistency**:

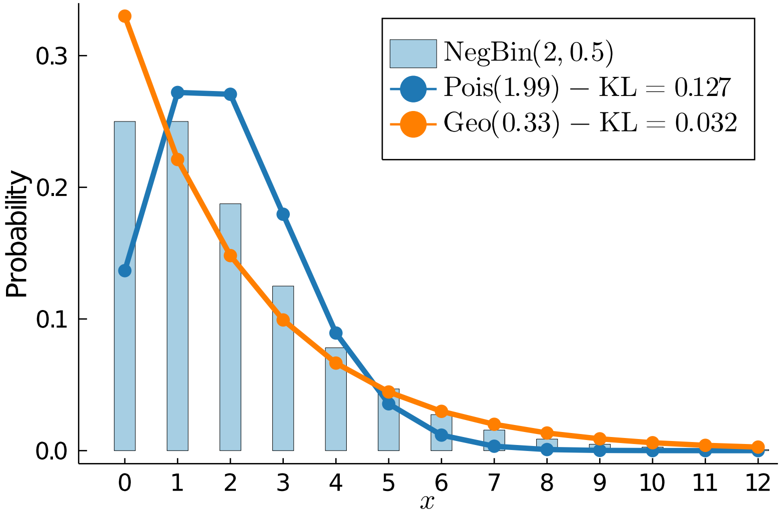
$$\Pr(M = M^\star | \mathbf{y}) \to 1 \quad \text{as} \quad n \to \infty$$

- **$\mathcal{M}$-open**: data generating process $M^\star$ is **not** in $\mathcal{M}$.
- **$\mathcal{M}$-open** is the realistic case.
- George Box: all models are false but some are useful.
- Where do posterior model probabilities go in **$\mathcal{M}$-open**?

Geometric vs Poisson for NegBin(2,0.5) data



Geometric vs Poisson for NegBin(2,0.5) data



Marginal likelihood is KL-consistent in $\mathcal{M}$-open

- **$\mathcal{M}$-open**: data generating process $M^{\star}$ is **not** in $\mathcal{M}$.
- **KL-consistency**: when $M^{\star} \notin \mathcal{M}$

$$\Pr\left(M = \tilde{M} | \mathbf{y}\right) \to 1 \quad \text{as} \quad n \to \infty,$$

- $\tilde{M}$ minimizes **KL divergence** between $p(\mathbf{y}|M)$ and $p(\mathbf{y}|M^{\star})$:

$$\mathrm{KL}(M^{\star}, M) = \int \log \frac{p(\mathbf{y}|M^{\star})}{p(\mathbf{y}|\hat{\theta}_M, M)} p(\mathbf{y}|M^{\star}) d\boldsymbol{y}$$

- $\hat{\theta}_M$ - model parameter that makes M as KL-close as possible to $M^{\star}$.

Model choice in multivariate time series[1]

- **Multivariate time series**

$$\mathbf{x}_t = \alpha\beta'\mathbf{z}_t + \Phi_1\mathbf{x}_{t-1} + ...\Phi_k\mathbf{x}_{t-k} + \Psi_1 + \Psi_2 t + \Psi_3 t^2 + \varepsilon_t$$

- Need to choose:
  - **Lag length**, $(k = 1, 2.., 4)$
  - **Trend model** $(s = 1, 2, ..., 5)$
  - **Long-run (cointegration) relations** $(r = 0, 1, 2, 3, 4)$.

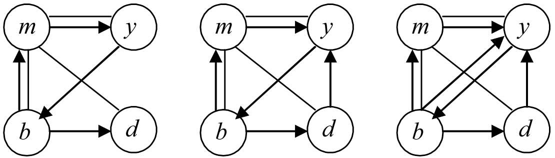
THE MOST PROBABLE (k, r, s) COMBINATIONS IN THE DANISH MONETARY DATA.

k	1	1	1	1	1	1	1	1	0	1
r	3	3	2	4	2	1	2	3	4	3
s	3	2	2	2	3	3	4	4	4	5
$p(k, r, s \mid y, x, z)$	.106	.093	.091	.060	.059	.055	.054	.049	.040	.038

[1] Corander and Villani (2004). Statistica Neerlandica.

Graphical models for multivariate time series[2]

- **Graphical models** for multivariate time series.
- Zero-restrictions on the effect from time series i on time series j, for all lags. (**Granger Causality**).
- Zero-restrictions on inverse covariance matrix of the errors. Contemporaneous conditional independence.



$p(G\backslash\mathbf{X}) = 0.0033$ $p(G\backslash\mathbf{X}) = 0.0028$ $p(G\backslash\mathbf{X}) = 0.0025$

[2] Corander and Villani (2004). Journal of Time Series Analysis.

Laplace approximation

- Taylor approximation of the log likelihood

$$\ln p(\mathbf{y}|\theta) \approx \ln p(\mathbf{y}|\hat{\theta}) - \frac{1}{2} J_{\hat{\theta},\mathbf{y}} (\theta - \hat{\theta})^2,$$

so

$$p(\mathbf{y}|\theta)p(\theta) \approx p(\mathbf{y}|\hat{\theta}) \exp\left[-\frac{1}{2} J_{\hat{\theta},\mathbf{y}} (\theta - \hat{\theta})^2\right] p(\hat{\theta})$$

$$= p(\mathbf{y}|\hat{\theta}) p(\hat{\theta}) (2\pi)^{p/2} \left| J^{-1}_{\hat{\theta},\mathbf{y}} \right|^{1/2}$$

$$\times \underbrace{(2\pi)^{-p/2} \left| J^{-1}_{\hat{\theta},\mathbf{y}} \right|^{-1/2} \exp\left[-\frac{1}{2} J_{\hat{\theta},\mathbf{y}} (\theta - \hat{\theta})^2\right]}_{\text{multivariate normal density}}$$

- **The Laplace approximation**:

$$\ln \hat{p}(\mathbf{y}) = \ln p(\mathbf{y}|\hat{\theta}) + \ln p(\hat{\theta}) + \frac{1}{2} \ln \left| J^{-1}_{\hat{\theta},\mathbf{y}} \right| + \frac{p}{2} \ln(2\pi),$$

where p is the number of unrestricted parameters.

BIC

- **The Laplace approximation**:

$$\ln \hat{p}(\mathbf{y}) = \ln p(\mathbf{y}|\hat{\theta}) + \ln p(\hat{\theta}) + \frac{1}{2}\ln\left|J_{\hat{\theta},\mathbf{y}}^{-1}\right| + \frac{p}{2}\ln(2\pi).$$

- $\hat{\theta}$ and $J_{\hat{\theta},\mathbf{y}}$ can be obtained with **optimization/autodiff**.

- The **BIC approximation** assumes that $J_{\hat{\theta},\mathbf{y}}$ behaves like $n \cdot I_p$ in large samples and the small term $\frac{p}{2}\ln(2\pi)$ is ignored

$$\ln \hat{p}(\mathbf{y}) = \ln p(\mathbf{y}|\hat{\theta}) + \ln p(\hat{\theta}) - \frac{p}{2}\ln n.$$

$\Pr(M_k|\mathbf{y})$ can be overfident - macroeconomics[3]

Table: Posterior model probabilities - Smets-Wouters DSGE model

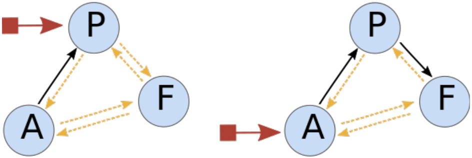
Base	M1	M2	M3	M4	M5	M6	M7	M8
0.01	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00

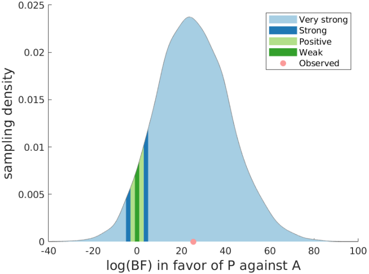
[3] Oelrich et al (2020). When are Bayesian model probabilities overconfident?

$\Pr(M_k|\boldsymbol{y})$ can be overfident - neuroscience[4]

Table: Posterior model probabilities - Dynamic Causal Models

A	F	P	AF	PA	PF	PAF
0.00	0.00	1.00	0.00	0.00	0.00	0.00





[4] Oelrich et al (2020). When are Bayesian model probabilities overconfident?

And hey! ... let's be careful out there

- Be especially **careful** with Bayesian model comparison when
  - The **compared models** are
    - very different in structure
    - severly misspecified
    - very complicated (black boxes).
  - The **priors** for the parameters in the models are
    - not carefully elicited
    - only weakly informative
    - not matched across models.
  - The **data**
    - has outliers (in all models)
    - has a multivariate response.