Bayesian Model Comparison

PhD course in Statistical Inference

Mattias Villani

Department of Statistics
Stockholm University
and
Department of Computer and Information Science
Linköping University

- **Bayesian model comparison**
- **Marginal likelihood**
- **Log Predictive Score**

Bayesian model comparison

- Known parameters, just use the **likelihood ratio**

$$\frac{p_1(\mathbf{y}|\theta_1)}{p_2(\mathbf{y}|\theta_2)}$$

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

$$p_k(\mathbf{y}) = \int p_k(\mathbf{y}|\theta_k)p_k(\theta_k)d\theta_k.$$

- θ_k is 'removed' by the prior. **Not a silver bullet**. **Priors matter!**
- The **Bayes factor**

$$B_{12}(y) = \frac{p_1(\mathbf{y})}{p_2(\mathbf{y})}.$$

- **Posterior model probabilities**

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

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

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

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

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

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

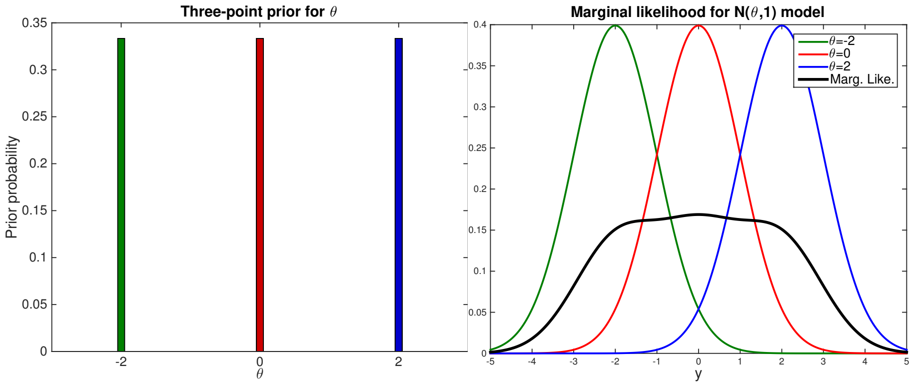
  where $B(\cdot, \cdot)$ is the **Beta function**.
- Posterior model probabilities

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

- The **Bayes factor**

$$BF(M_0; M_1) = \frac{p(x_1, \ldots, x_n | H_0)}{p(x_1, \ldots, x_n | H_1)} = \frac{\theta_0^s (1-\theta_0)^f B(\alpha, \beta)}{B(\alpha+s, \beta+f)}.$$

Priors matter



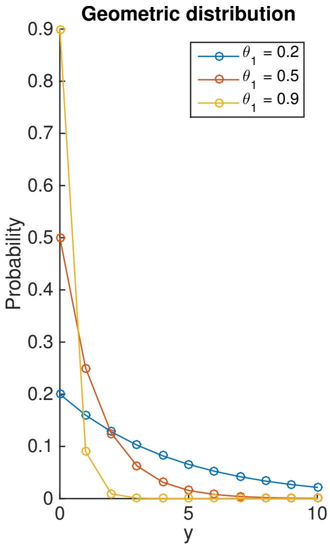
Example: Geometric vs Poisson

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

$$\begin{aligned} p_1(y_1, ..., y_n) &= \int p_1(y_1, ..., y_n | \theta_1) p(\theta_1) d\theta_1 \\ &= \frac{\Gamma(\alpha_1 + \beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \frac{\Gamma(n + \alpha_1)\Gamma(n\bar{y} + \beta_1)}{\Gamma(n + n\bar{y} + \alpha_1 + \beta_1)} \end{aligned}$$

- **Marginal likelihood** for M_2

$$p_2(y_1, ..., y_n) = \frac{\Gamma(n\bar{y} + \alpha_2)\beta_2^{\alpha_2}}{\Gamma(\alpha_2)(n + \beta_2)^{n\bar{y} + \alpha_2}} \frac{1}{\prod_{i=1}^{n} y_i!}$$


Geometric distribution
Probability
y
θ_1 = 0.2
θ_1 = 0.5
θ_1 = 0.9
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
5
10
Poisson distribution
Probability
y
θ_2 = 0.5
θ_2 = 1
θ_2 = 5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
5
10

Geometric vs Poisson, cont.

- Priors match prior predictive means:

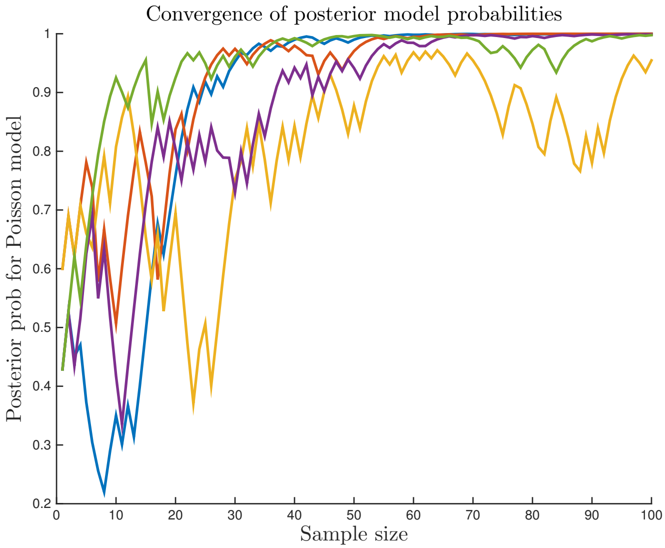
$$E(y_i|M_1) = E(y_i|M_2) \iff \alpha_1\alpha_2 = \beta_1\beta_2$$

- **Data**: $y_1 = 0,\ y_2 = 0$.

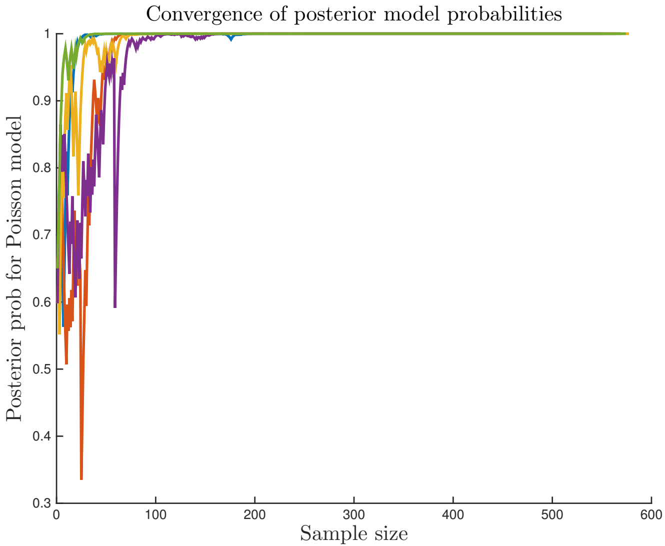
	$\alpha_1 = 1, \beta_1 = 2$ $\alpha_2 = 2, \beta_2 = 1$	$\alpha_1 = 10, \beta_1 = 20$ $\alpha_2 = 20, \beta_2 = 10$	$\alpha_1 = 100, \beta_1 = 200$ $\alpha_2 = 200, \beta_2 = 100$
BF_{12}	1.5	4.54	5.87
$\Pr(M_1\|\mathbf{y})$	0.6	0.82	0.85
$\Pr(M_2\|\mathbf{y})$	0.4	0.18	0.15

- **Data**: $y_1 = 3,\ y_2 = 3$.

	$\alpha_1 = 1, \beta_1 = 2$ $\alpha_2 = 2, \beta_2 = 1$	$\alpha_1 = 10, \beta_1 = 20$ $\alpha_2 = 20, \beta_2 = 10$	$\alpha_1 = 100, \beta_1 = 200$ $\alpha_2 = 200, \beta_2 = 100$
BF_{12}	0.26	0.29	0.30
$\Pr(M_1\|\mathbf{y})$	0.21	0.22	0.23
$\Pr(M_2\|\mathbf{y})$	0.79	0.78	0.77


Convergence of posterior model probabilities
Posterior prob for Poisson model
Sample size
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
10
20
30
40
50
60
70
80
90
100

Geometric vs Poisson for Pois(1) data



- **Multivariate time series**

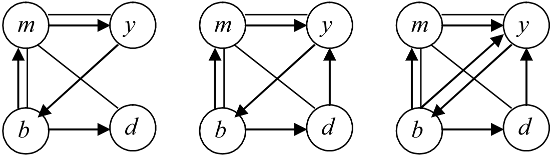
$$\mathbf{x}_t = \alpha\beta'\mathbf{z}_t + \Phi_1\mathbf{x}_{t-1} + \ldots\Phi_k\mathbf{x}_{t-k} + \Psi_1 + \Psi_2 t + \Psi_3 t^2 + \varepsilon_t$$

- 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

- **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 the elements of the inverse covariance matrix of the errors.



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

- Coherence of pair-wise comparisons

$$B_{12} = B_{13} \cdot B_{32}$$

- **Consistency** when true model is in $\mathcal{M} = \{M_1, ..., M_K\}$

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

- "KL-consistency" when $M_{TRUE} \notin \mathcal{M}$

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

  M^* minimizes **KL divergence** between $p_M(\mathbf{y})$ and $p_{TRUE}(\mathbf{y})$.
- Smaller models always win when priors are very vague.
- **Improper priors** cannot be used for model comparison.

Marginal likelihood measures out-of-sample predictive performance

- The **marginal likelihood** can be **decomposed** as

$$p(y_1, ..., y_n) = p(y_1)p(y_2|y_1) \cdots p(y_n|y_1, y_2, ..., y_{n-1})$$

- Assume that y_i is independent of $y_1, ..., y_{i-1}$ conditional on θ:

$$p(y_i|y_1, ..., y_{i-1}) = \int p(y_i|\theta)p(\theta|y_1, ..., y_{i-1})d\theta$$

- **Prediction of y_1** is based on the prior of θ. Sensitive to prior.
- **Prediction of y_n** uses almost all the data to infer θ. Not sensitive to prior when n is not small.

Normal example

- **Model**: $y_1, \ldots, y_n | \theta \sim N(\theta, \sigma^2)$ with σ^2 known.
- **Prior**: $\theta \sim N(0, \kappa^2 \sigma^2)$.
- **Intermediate posterior** at time $i-1$

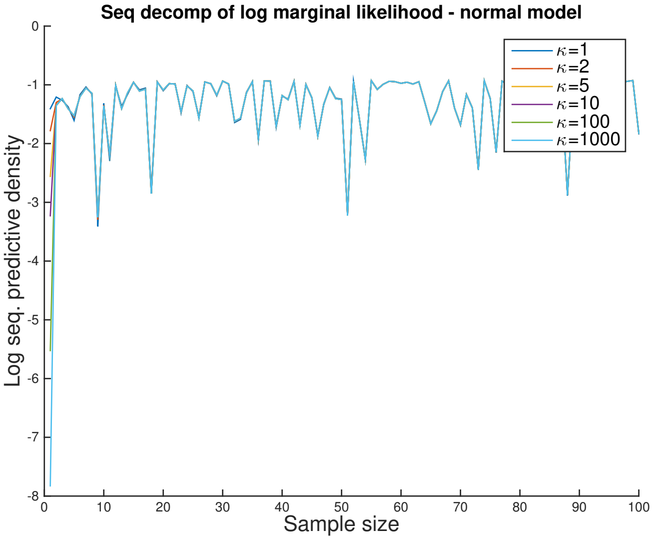
$$\theta | y_1, \ldots, y_{i-1} \sim N\left[w_i(\kappa) \cdot \bar{y}_{i-1}, \frac{\sigma^2}{i-1+\kappa^{-2}}\right]$$

  where $w_i(\kappa) = \frac{i-1}{i-1+\kappa^{-2}}$.
- **Intermediate predictive density** at time $i-1$

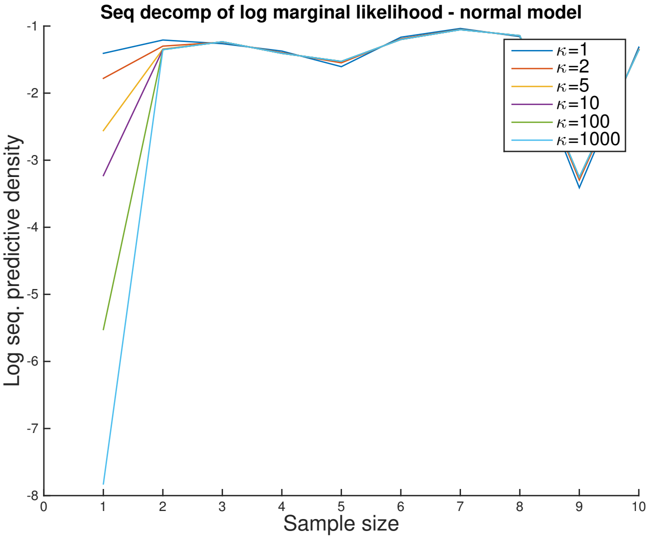
$$y_i | y_1, \ldots, y_{i-1} \sim N\left[w_i(\kappa) \cdot \bar{y}_{i-1}, \sigma^2\left(1 + \frac{1}{i-1+\kappa^{-2}}\right)\right]$$

- For $i = 1$, $y_1 \sim N\left[0, \sigma^2\left(1 + \frac{1}{\kappa^{-2}}\right)\right]$ can be very sensitive to κ.
- For large i: $y_i | y_1, \ldots, y_{i-1} \overset{approx}{\sim} N\left(\bar{y}_{i-1}, \sigma^2\right)$, not sensitive to κ.

First observation is sensitive to κ



First observation is sensitive to κ - zoomed



Log Predictive Score - LPS

- Reduce sensitivity to the prior: sacrifice n^* observations to train the prior into a posterior.
- **Predictive (Density) Score** (**PS**). Decompose $p(y_1, ..., y_n)$ as

$$\underbrace{p(y_1)p(y_2|y_1)\cdots p(y_{n^*}|y_{1:(n^*-1)})}_{training}\ \underbrace{p(y_{n^*+1}|y_{1:n^*})\cdots p(y_n|y_{1:(n-1)})}_{test}$$

- Usually report on log scale: **Log Predictive Score** (**LPS**).
- Time-series: obvious which data are used for training.
- Cross-sectional data: training-test split by **cross-validation**:

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

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.

- Let γ be a quanitity with an interpretation which stays the same across the two models.
- Example: Prediction $\gamma = (y_{T+1}, ..., y_{T+h})'$.
- The marginal posterior distribution of γ reads

$$p(\gamma|\mathbf{y}) = p(M_1|\mathbf{y})p_1(\gamma|\mathbf{y}) + p(M_2|\mathbf{y})p_2(\gamma|\mathbf{y}),$$

  where $p_k(\gamma|\mathbf{y})$ is the marginal posterior of γ conditional on model k.
- Predictive distribution includes **three sources of uncertainty**:
  - **Future errors**/disturbances (e.g. the ε's in a regression)
  - **Parameter uncertainty** (the predictive distribution has the parameters integrated out by their posteriors)
  - **Model uncertainty** (by model averaging)