

# BAYESIAN LEARNING - LECTURE 10

Mattias Villani

**Division of Statistics and Machine Learning  
Department of Computer and Information Science  
Linköping University**

# OVERVIEW

- ▶ Bayesian model comparison
- ▶ Marginal likelihood

# USING LIKELIHOOD FOR MODEL COMPARISON

- ▶ Consider two models for the data  $\mathbf{y} = (y_1, \dots, y_n)$ :  $M_1$  and  $M_2$ .
- ▶ Let  $p_i(\mathbf{y}|\theta_i)$  denote the data density under model  $M_i$ .
- ▶ If know  $\theta_1$  and  $\theta_2$ , the **likelihood ratio** is useful

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

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

$$\frac{p_1(\mathbf{y}|\hat{\theta}_1)}{p_2(\mathbf{y}|\hat{\theta}_2)}.$$

- ▶ 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

- ▶ Just use your priors  $p_1(\theta_1)$  och  $p_2(\theta_2)$ .
- ▶ The **marginal likelihood** for model  $M_k$  with parameters  $\theta_k$

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

- ▶  $\theta_k$  is removed by the prior. **Not a silver bullet. Priors matter!**
- ▶ The **Bayes factor**

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

- ▶ **Posterior model probabilities**

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

# BAYESIAN HYPOTHESIS TESTING - BERNOULLI

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

$$M_0 : x_1, \dots, x_n \stackrel{iid}{\sim} \text{Bernoulli}(\theta_0)$$

$$M_1 : x_1, \dots, x_n \stackrel{iid}{\sim} \text{Bernoulli}(\theta), \theta \sim \text{Beta}(\alpha, \beta)$$

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

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

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

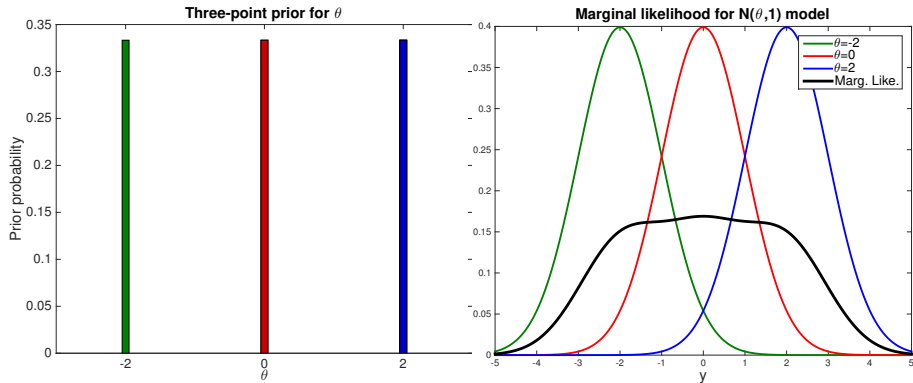
- Posterior model probabilities

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

- The Bayes factor

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

# PRIORS MATTER



## EXAMPLE: GEOMETRIC VS POISSON

- ▶ Model 1 - **Geometric** with Beta prior:

- ▶  $y_1, \dots, y_n | \theta_1 \sim \text{Geo}(\theta_1)$
- ▶  $\theta_1 \sim \text{Beta}(\alpha_1, \beta_1)$

- ▶ Model 2 - **Poisson** with Gamma prior:

- ▶  $y_1, \dots, y_n | \theta_2 \sim \text{Poisson}(\theta_2)$
- ▶  $\theta_2 \sim \text{Gamma}(\alpha_2, \beta_2)$

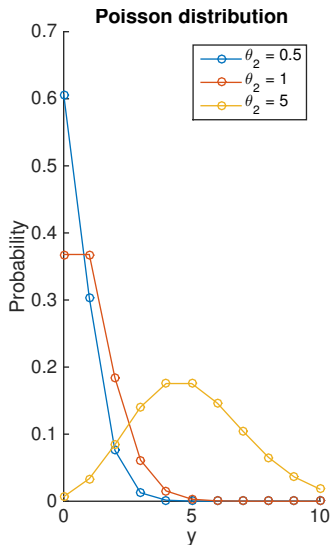
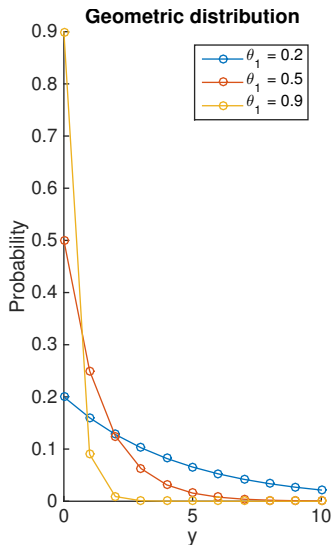
- ▶ Marginal likelihood for  $M_1$

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

- ▶ Marginal likelihood for  $M_2$

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

# GEOMETRIC AND POISSON





## GEOMETRIC VS POISSON, CONT.

- Priors match prior predictive means:

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

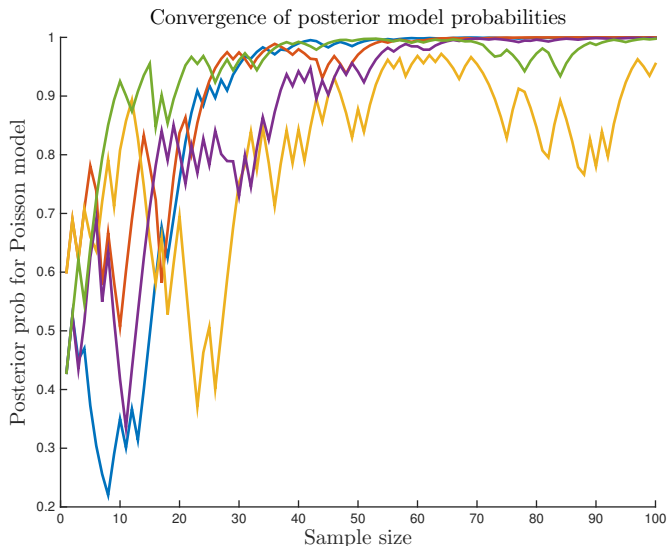
- Data:**  $y_1 = 0, y_2 = 0$ .

	$\alpha_1 = 1, \beta_1 = 2$ $\alpha_2 = 2, \beta_2 = 1$	$\alpha_1 = 10, \beta_1 = 20$ $\alpha_2 = 20, \beta_2 = 10$	$\alpha_1 = 100, \beta_1 = 200$ $\alpha_2 = 200, \beta_2 = 100$
$BF_{12}$	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

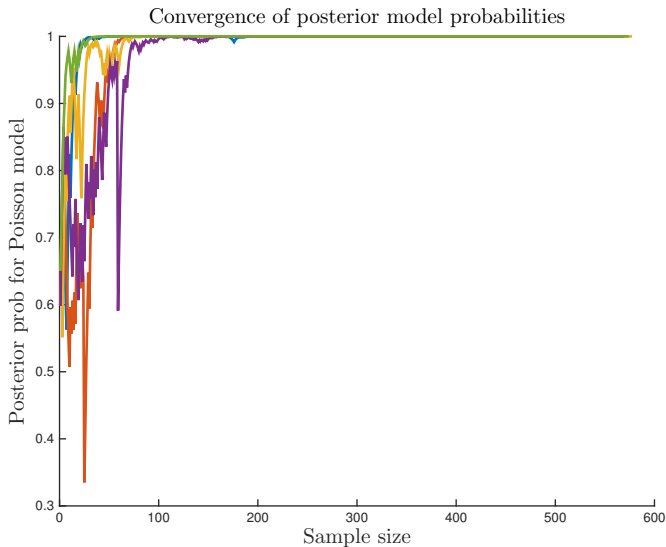
- Data:**  $y_1 = 3, y_2 = 3$ .

	$\alpha_1 = 1, \beta_1 = 2$ $\alpha_2 = 2, \beta_2 = 1$	$\alpha_1 = 10, \beta_1 = 20$ $\alpha_2 = 20, \beta_2 = 10$	$\alpha_1 = 100, \beta_1 = 200$ $\alpha_2 = 200, \beta_2 = 100$
$BF_{12}$	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

# GEOMETRIC VS POISSON FOR POIS(1) DATA



# GEOMETRIC VS POISSON FOR POIS(1) DATA



# MODEL CHOICE IN MULTIVARIATE TIME SERIES

- ▶ Multivariate time series

$$\mathbf{x}_t = \alpha\beta'\mathbf{z}_t + \Phi_1\mathbf{x}_{t-1} + \dots\Phi_k\mathbf{x}_{t-k} + \Psi_1 + \Psi_2t + \Psi_3t^2 + \varepsilon_t$$

- ▶ Need to choose:

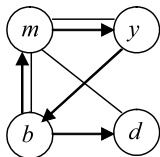
- ▶ **Lag length**, ( $k = 1, 2, \dots, 4$ )
- ▶ **Trend model** ( $s = 1, 2, \dots, 5$ )
- ▶ **Long-run (cointegration) relations** ( $r = 0, 1, 2, 3, 4$ ).

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

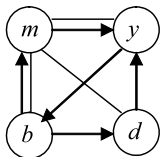
$k$	1	1	1	1	1	1	1	1	0	1
$r$	3	3	2	4	2	1	2	3	4	3
$s$	3	2	2	2	3	3	4	4	4	5
$p(k, r, s y, x, z)$	.106	.093	.091	.060	.059	.055	.054	.049	.040	.038

# GRAPHICAL MODELS FOR MULTIVARIATE TIME SERIES

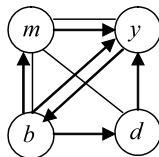
- ▶ **Graphical models** for multivariate time series.
- ▶ Zero-restrictions on the effect from time series  $i$  on time series  $j$ , for all lags. (**Granger Causality**).
- ▶ Zero-restrictions on the elements of the inverse covariance matrix of the errors.



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



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



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

# PROPERTIES OF BAYESIAN MODEL COMPARISON

- ▶ Coherence of pair-wise comparisons

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

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

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

- ▶ “KL-consistency” when  $M_{TRUE} \notin \mathcal{M}$

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

where  $M^*$  is the model that minimizes Kullback-Leibler distance 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

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

- ▶ If we assume that  $y_i$  is independent of  $y_1, \dots, y_{i-1}$  conditional on  $\theta$ :

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

- ▶ The prediction of  $y_1$  is based on the prior of  $\theta$ , and is therefore sensitive to the prior.
- ▶ The prediction of  $y_n$  uses almost all the data to infer  $\theta$ . Very little influenced by the prior when  $n$  is not small.

## NORMAL EXAMPLE

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

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

where  $w_i(\kappa) = \frac{i-1}{i-1+\kappa^{-2}}$ .

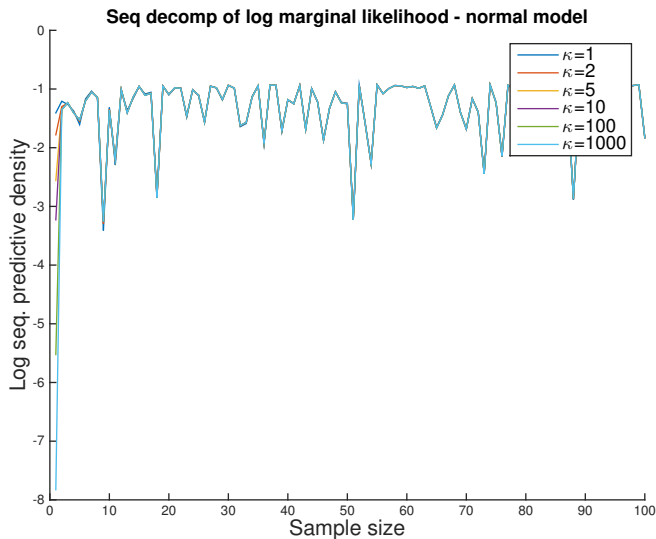
- ▶ Predictive density at time  $i - 1$

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

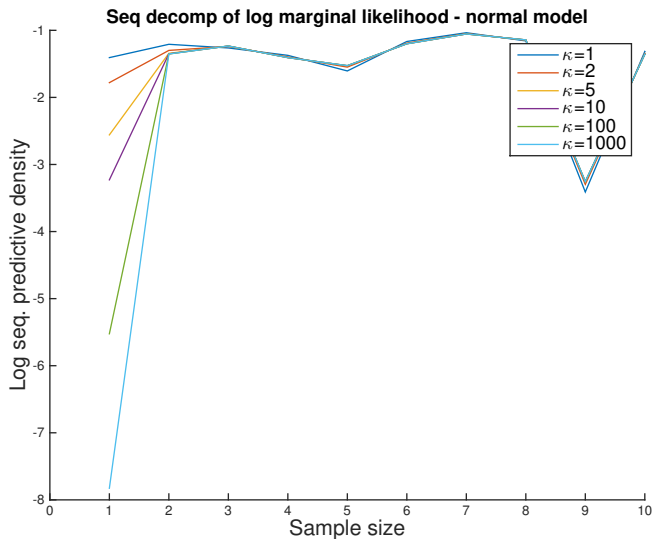
- ▶ Terms with  $i$  large:  $y_i | y_1, \dots, y_{i-1} \overset{\text{approx}}{\sim} N(\bar{y}_{i-1}, \sigma^2)$ , not sensitive to  $\kappa$
- ▶ For  $i = 1$ ,  $y_1 \sim N \left[ 0, \sigma^2 \left( 1 + \frac{1}{\kappa^{-2}} \right) \right]$  can be very sensitive to  $\kappa$ .



# FIRST OBSERVATION IS SENSITIVE TO $\kappa$



# FIRST OBSERVATION IS SENSITIVE TO $\kappa$



# LOG PREDICTIVE SCORE - LPS

- ▶ To reduce sensitivity to the prior: sacrifice  $n^*$  observations to train the prior into a better posterior.
- ▶ Predictive density score (PS). Decompose  $p(y_1, \dots, y_n)$  as

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

- ▶ Usually report on log scale: **Log Predictive Score (LPS)**.
- ▶ But which observations to train on (and which to test on)?
- ▶ Straightforward for time series.
- ▶ Cross-sectional data: **cross-validation**.

# AND HEY! ... LET'S BE CAREFUL OUT THERE.

- ▶ Be especially careful with Bayesian model comparison when
  - ▶ The compared models are
    - ▶ very different in structure
    - ▶ severely 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.