

Bayesian Statistics I

Lecture 11 - Bayesian Model Comparison

Mattias Villani

Department of Statistics
Stockholm University

Department of Computer and Information Science
Linköping University



mattiasvillani.com



@matvil



[mattiasvillani](https://github.com/mattiasvillani)

Overview

- Bayesian model comparison
- Marginal likelihood
- Log Predictive Score

Using likelihood for model comparison

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

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

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

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

- 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

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

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

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

■ θ_k is 'removed' by the averaging wrt prior. **Priors matter!**

■ The **Bayes factor**

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

Jeffreys scale of evidence for the Bayes factor

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

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**

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

Normal example

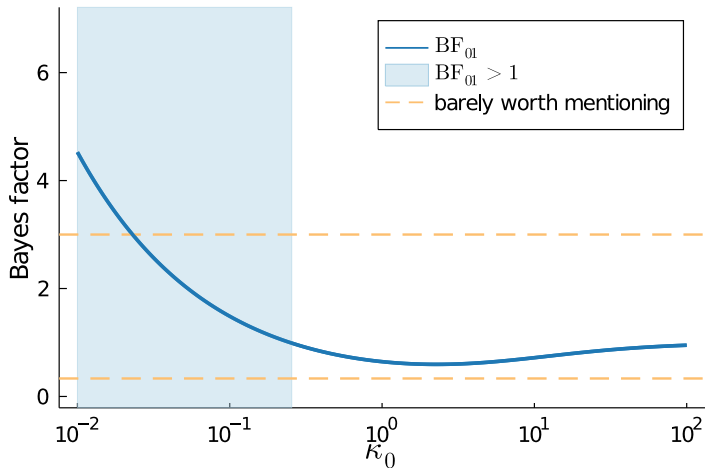
- **Model:** $x_1, \dots, x_n \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2)$, σ^2 known.
- **Prior:** $\theta \sim N(\mu_0, \sigma^2/\kappa_0)$.
- **Likelihood:** \bar{x} is **sufficient** for θ and $\bar{x}|\theta \sim N(\theta, \sigma^2/n)$.
- **Marginal likelihood:** $p(\bar{x}|M_1) = N(\mu_0, \sigma^2(1/n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

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

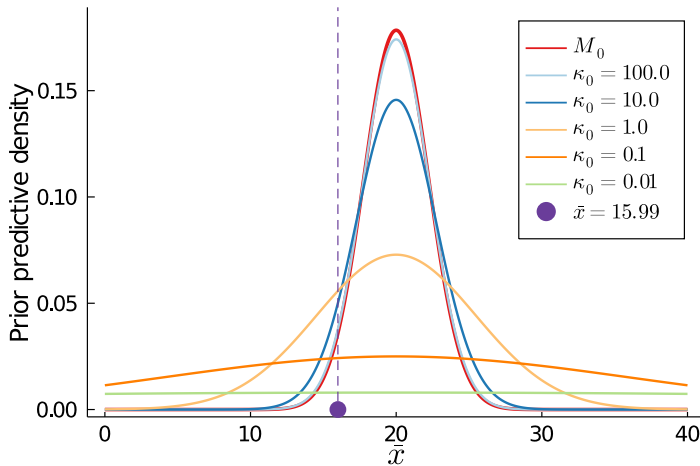
$$\log \frac{p(\bar{x}|M_0)}{p(\bar{x}|M_1)} = -\frac{1}{2} \log \left(\frac{\kappa_0}{\kappa_0 + n} \right) - \frac{n(\bar{x} - \mu_0)^2}{2\sigma^2} \left(\frac{n}{\kappa_0 + n} \right)$$

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

Internet speed data - Bayes factor




Internet speed data - prior predictive density

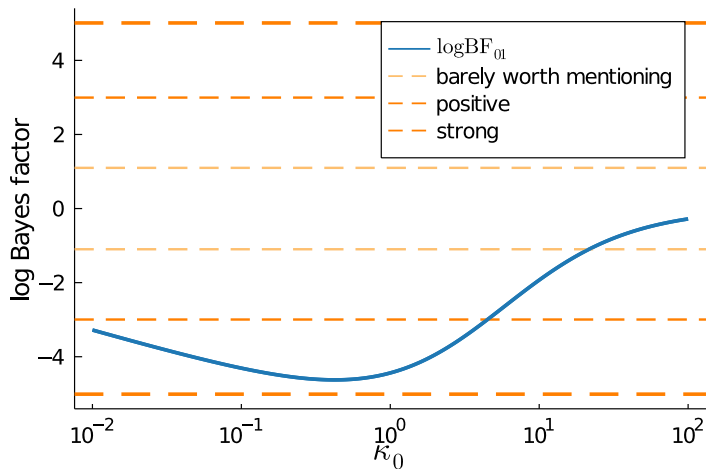


Vague priors for marginal likelihoods is a bad idea

- 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:

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

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

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

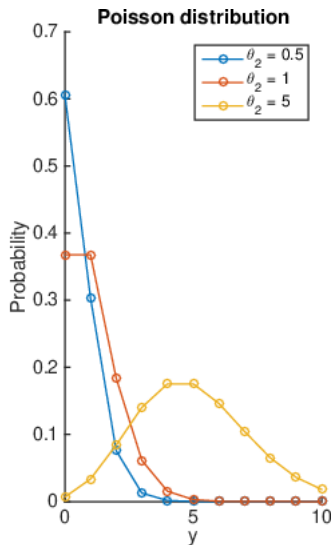
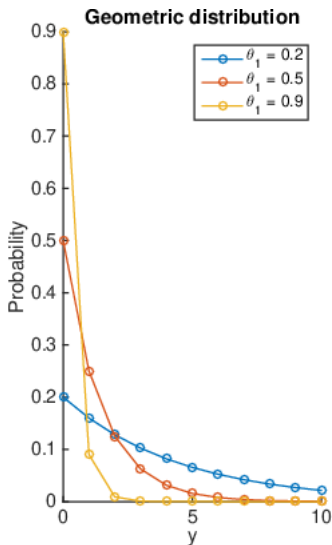
- **Marginal likelihood** for M_1

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

- **Marginal likelihood** for M_2

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

Geometric and Poisson



Geometric vs Poisson

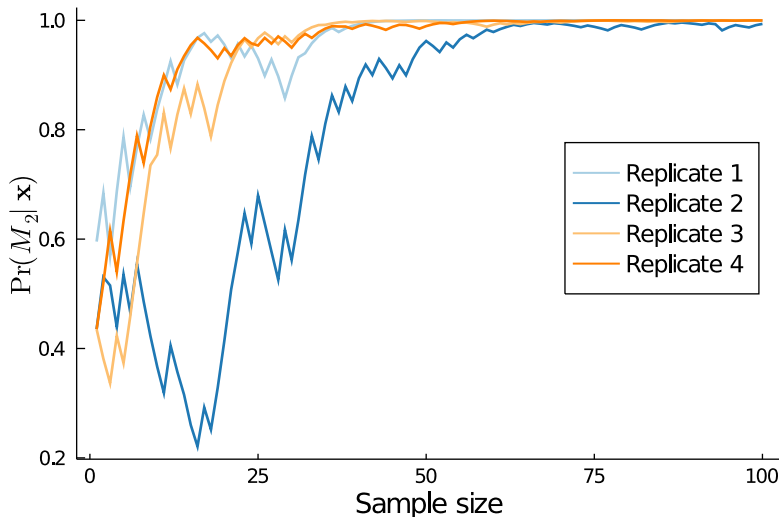
- Use priors to match prior predictive means:

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

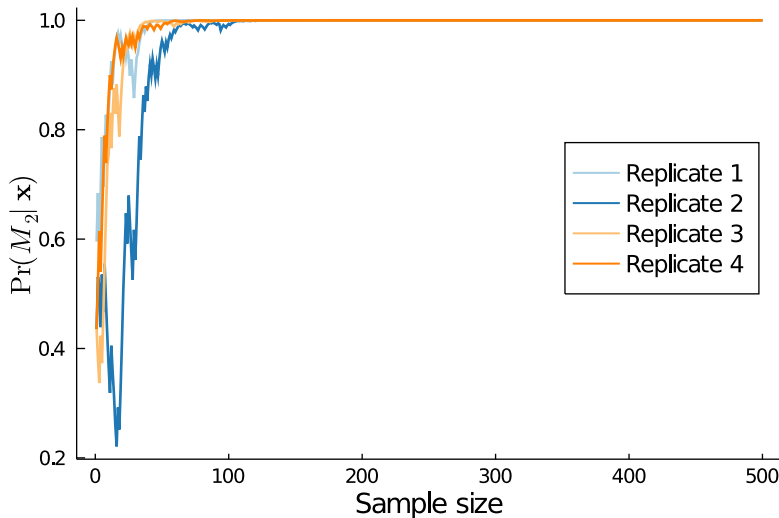
- Geometric model: $\alpha_1 = 10, \beta_1 = 20$.
- Poisson model: $\alpha_2 = 20, \beta_2 = 10$.

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

Geometric vs Poisson for Pois(1) data



Geometric vs Poisson for Pois(1) data



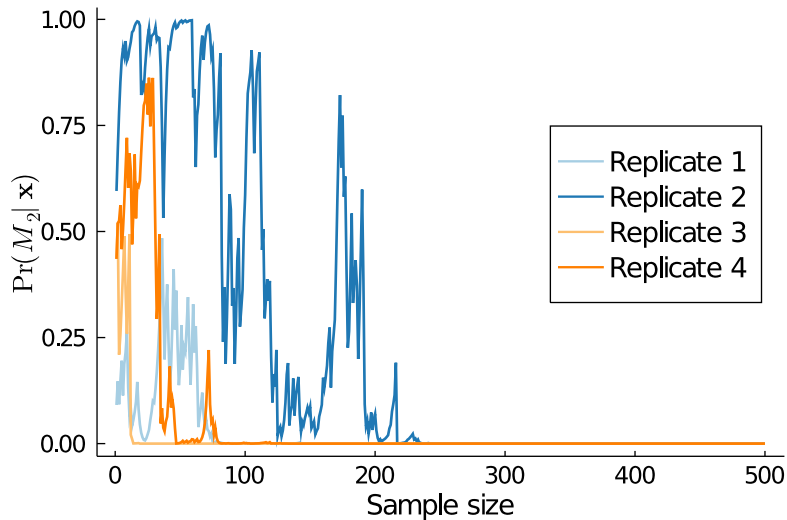
Asymptotic properties of marginal likelihood

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

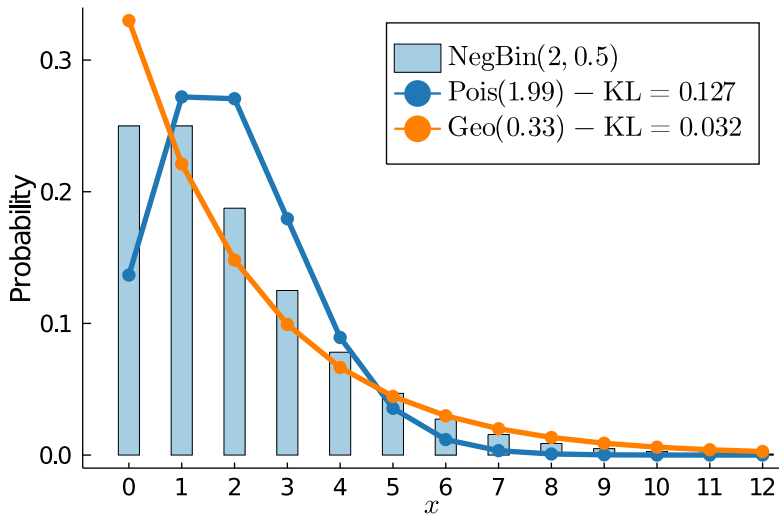
$$\Pr(M = M^\star | y) \rightarrow 1 \quad \text{as } n \rightarrow \infty$$

- \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^* is **not** in \mathcal{M} .

■ **KL-consistency**: when $M^* \notin \mathcal{M}$

$$\Pr(M = \tilde{M} | y) \rightarrow 1 \quad \text{as } n \rightarrow \infty,$$

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

$$\text{KL}(M^*, M) = \int \log \frac{p(y|M^*)}{p(y|\hat{\theta}_M, M)} p(y|M^*) dy$$

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

Model choice in multivariate time series¹

■ Multivariate time series

$$x_t = \alpha\beta'z_t + \Phi_1x_{t-1} + \dots\Phi_kx_{t-k} + \Psi_1 + \Psi_2t + \Psi_3t^2 + \varepsilon_t$$

■ Need to choose:

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

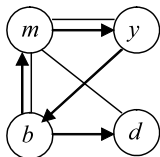
THE MOST PROBABLE (k, r, s) COMBINATIONS IN THE DANISH MONETARY DATA.

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

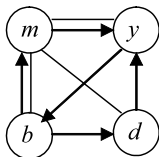
¹Corander and Villani (2004). Statistica Neerlandica.

Graphical models for multivariate time series²

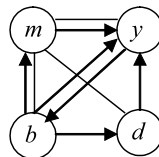
- **Graphical models** for multivariate time series.
- Zero-restrictions on the effect from time series i on time series j , for all lags. (**Granger Causality**).
- Zero-restrictions on inverse covariance matrix of the errors. Contemporaneous conditional independence.



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



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



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

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

Laplace approximation

- Taylor approximation of the log likelihood

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

so

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

- The Laplace approximation:

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

where p is the number of unrestricted parameters.

■ The Laplace approximation:

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

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

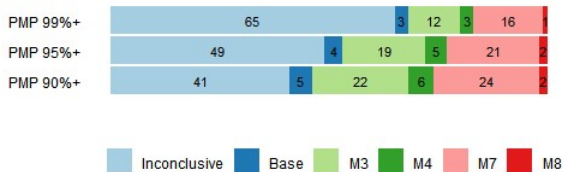
■ The **BIC approximation** assumes that $J_{\hat{\theta},y}$ behaves like $n \cdot I_p$ in large samples and the small term $\frac{p}{2} \ln(2\pi)$ is ignored

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

$\Pr(M_k|y)$ can be overfident - macroeconomics³

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

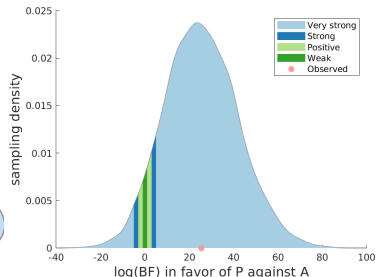
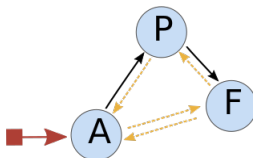
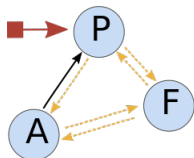


³Oelrich et al (2020). When are Bayesian model probabilities overconfident?

$\Pr(M_k|y)$ can be overfident - neuroscience⁴

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



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

Marginal likelihood measures out-of-sample predictive performance

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

$$p(x_1, \dots, x_n) = p(x_1)p(x_2|x_1) \cdots p(x_n|x_1, x_2, \dots, x_{n-1})$$

a product of **intermediate predictive densities**

$$p(x_i|x_1, \dots, x_{i-1}) = \int p(x_i|x_1, \dots, x_{i-1}, \theta)p(\theta|x_1, \dots, x_{i-1})d\theta$$

and $p(\theta|x_1, \dots, x_{i-1})$ is the **intermediate posterior**.

- **Prediction of x_1** is based on the prior of θ . Sensitive to prior.
- **Prediction of x_n** uses almost all the data to infer θ . Not sensitive to prior when n is not small.

Normal example

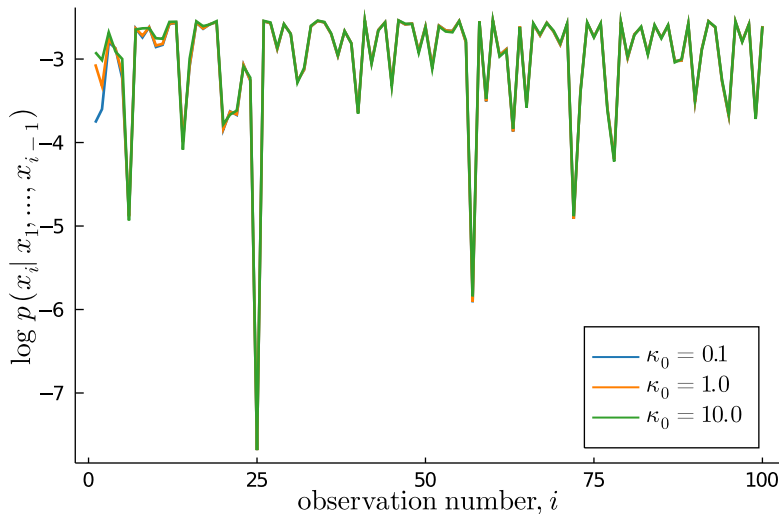
- **Model:** $x_1, \dots, x_n | \theta \sim N(\theta, \sigma^2)$ with σ^2 known.
- **Prior:** $\theta \sim N(0, \sigma^2 / \kappa_0)$.
- **Intermediate predictive density** at time $i - 1$

$$x_i | x_1, \dots, x_{i-1} \sim N \left(\mu_{i-1}, \sigma^2 \left(1 + \frac{1}{i-1 + \kappa_0} \right) \right),$$

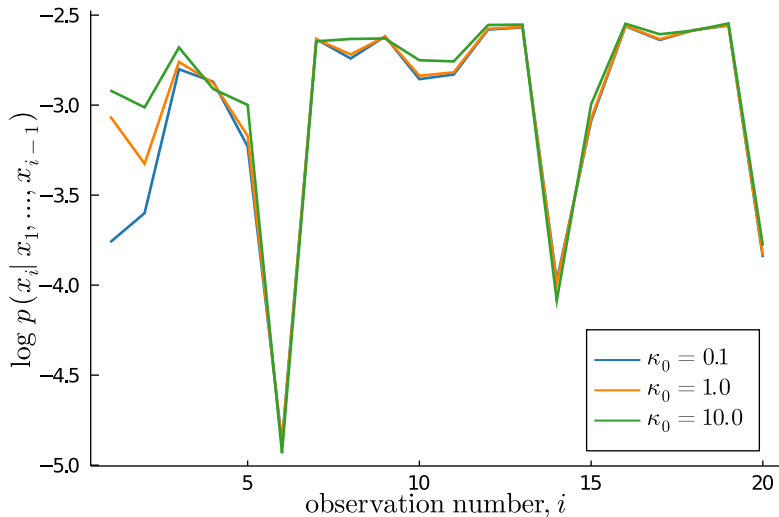
where

- ▶ $\mu_{i-1} = w_{i-1} \bar{x}_{i-1} + (1 - w_{i-1}) \mu_0$
- ▶ \bar{x}_{i-1} is the sample mean of the first $i - 1$ obs
- ▶ $w_{i-1} = (i - 1) / (i - 1 + \kappa_0)$
- $i = 1$, $x_1 \sim N \left[0, \sigma^2 \left(1 + \frac{1}{\kappa_0} \right) \right]$ can be very sensitive to κ_0 .
- Large i : $x_i | x_1, \dots, x_{i-1} \stackrel{\text{approx}}{\sim} N(\bar{x}_{i-1}, \sigma^2)$, not sensitive to κ_0 .

First observations are sensitive to κ_0



First observations are sensitive to κ_0 - zoomed



Log Predictive Score - LPS

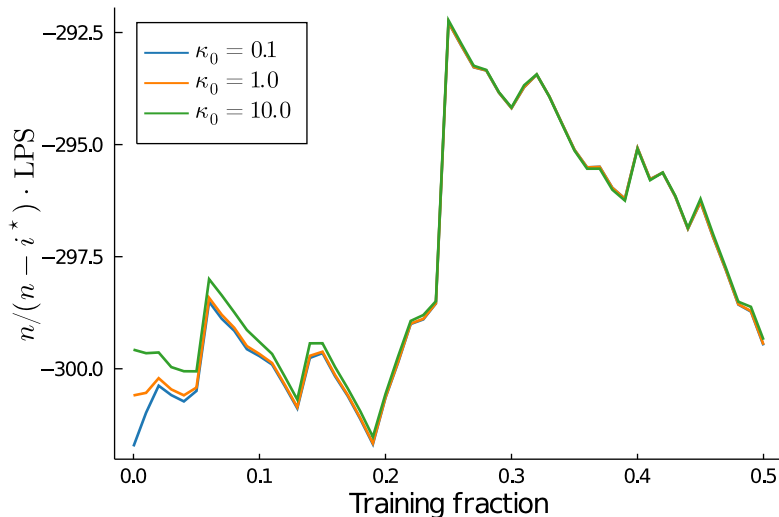
- Reduce prior sensitivity: use n^* observations to train the prior.
- (Log) Predictive (Density) Score (PS):

$$\underbrace{p(x_1)p(x_2|x_1)\cdots p(x_{n^*}|x_{1:(n^*-1)})}_{\text{training}} \underbrace{p(x_{n^*+1}|x_{1:n^*})\cdots p(x_n|x_{1:(n-1)})}_{\text{test}}$$

- Time-series: obvious which data are used for training.
- Cross-sectional data: training-test split by **cross-validation**:

	n data observations				
	$1, 2, \dots, n-1, n$				
Split 1:	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2:	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 3:	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 4:	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 5:	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

LPS not sensitive to κ_0



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.