Bayesian Learning

Lecture 12 - Predictive model comparison methods and variable selection

Mattias Villani 👨

Department of Statistics
Stockholm University

mattiasvillani.com @matvil @matvil mattiasvillani

Overview

- Log predictive scores for model comparison
- Bayesian variable selection
- Model averaging
- Posterior predictive analysis

Marginal likelihood measures out-of-sample predictive performance

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

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

  a product of **intermediate predictive densities**

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

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

- **Prediction of x_1** is based on the prior of $\boldsymbol{\theta}$. Sensitive to prior.
- **Prediction of x_n** uses almost all the data to infer $\boldsymbol{\theta}$. Not sensitive to prior when n is not small.

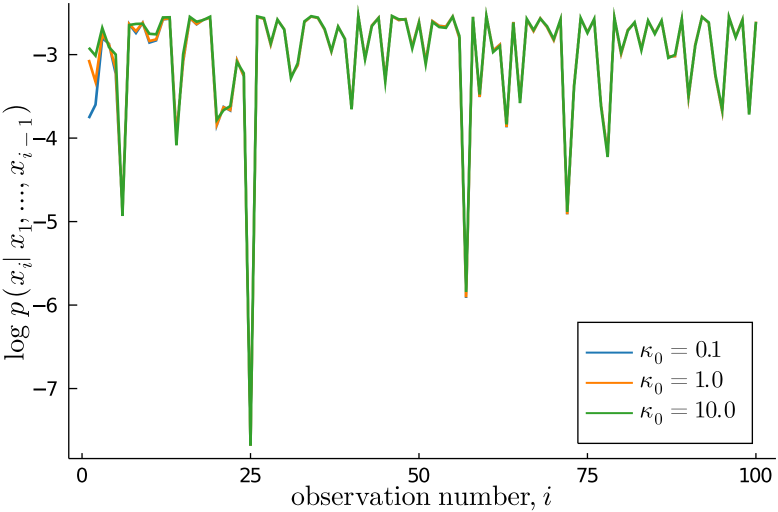
Normal example

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

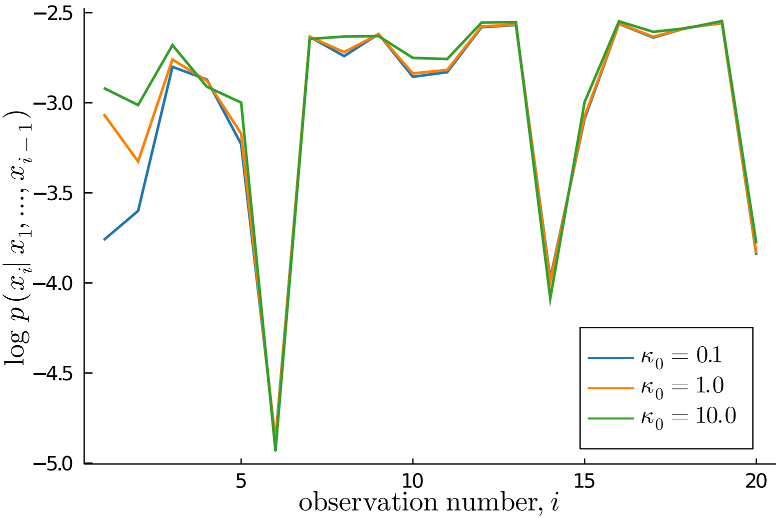
$$x_i | x_1, \ldots, x_{i-1} \sim \mathrm{N}\left(\mu_{i-1}, \sigma^2\left(1 + \frac{1}{i-1+\kappa_0}\right)\right),$$

  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: $x_i | x_1, ..., x_{i-1} \overset{\text{approx}}{\sim} N\left(\bar{x}_{i-1}, \sigma^2\right)$, not sensitive to κ_0.

First observations are sensitive to κ_0



First observations are sensitive to κ_0 - zoomed



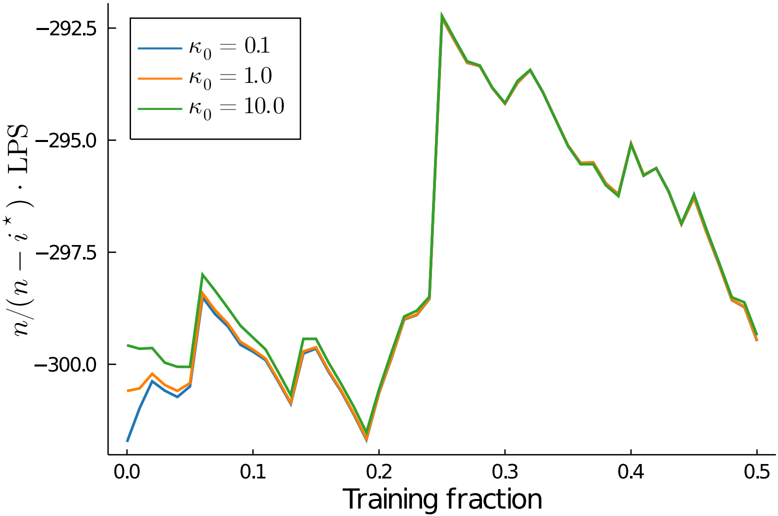
Log Predictive Score - LPS

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

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

- Time-series: obvious which data are used for training.

LPS not sensitive to κ_0



Bayesian variable selection

- Linear regression:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \varepsilon.$$

- Which variables have **non-zero** coefficient?

$$\begin{aligned} H_0 &: \beta_0 = \beta_1 = ... = \beta_p = 0 \\ H_1 &: \beta_1 = 0 \\ H_2 &: \beta_1 = \beta_2 = 0 \end{aligned}$$

- Introduce **variable selection indicators** $\mathcal{I} = (I_1, ..., I_p)$.
- Example: $\mathcal{I} = (1, 1, 0)$ means that $\beta_1 \neq 0$ and $\beta_2 \neq 0$, but $\beta_3 = 0$, so x_3 drops out of the model.

Bayesian variable selection

- Model inference, just crank the Bayesian machine:

$$p(\mathcal{I}|\mathbf{y},\mathbf{X}) \propto p(\mathbf{y}|\mathbf{X},\mathcal{I}) \cdot p(\mathcal{I})$$

- The prior $p(\mathcal{I})$ is typically taken to be

$$I_1, ..., I_p|\theta \overset{iid}{\sim} Bernoulli(\theta)$$

- θ is the **prior inclusion probability**.

- Challenge: Computing the **marginal likelihood** for each model $(\mathcal{I})$

$$p(\mathbf{y}|\mathbf{X},\mathcal{I}) = \int p(\mathbf{y}|\mathbf{X},\mathcal{I},\beta)p(\beta|\mathbf{X},\mathcal{I})d\beta$$

Bayesian variable selection

- Let $\beta_{\mathcal{I}}$ denote the **non-zero** coefficients under $\mathcal{I}$.
- Prior:

$$\beta_{\mathcal{I}}|\sigma^2 \sim N\left(0, \sigma^2\Omega_{\mathcal{I},0}^{-1}\right)$$
$$\sigma^2 \sim Inv-\chi^2\left(\nu_0, \sigma_0^2\right)$$

- **Marginal likelihood**

$$p(\mathbf{y}|\mathbf{X},\mathcal{I}) \propto \left|\mathbf{X}_{\mathcal{I}}'\mathbf{X}_{\mathcal{I}} + \Omega_{\mathcal{I},0}^{-1}\right|^{-1/2} \left|\Omega_{\mathcal{I},0}\right|^{1/2} \left(\nu_0\sigma_0^2 + RSS_{\mathcal{I}}\right)^{-(\nu_0+n-1)/2}$$

where $\mathbf{X}_{\mathcal{I}}$ is the covariate matrix for the subset selected by $\mathcal{I}$.

- $RSS_{\mathcal{I}}$ is (almost) the residual sum of squares for model with $\mathcal{I}$

$$RSS_{\mathcal{I}} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}_{\mathcal{I}}\left(\mathbf{X}_{\mathcal{I}}'\mathbf{X}_{\mathcal{I}} + \Omega_{\mathcal{I},0}\right)^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{y}$$

Bayesian variable selection via Gibbs sampling

- But there are 2^p model combinations to go through! *Ouch*!
- ... but most have essentially zero posterior probability. *Phew*!
- **Simulate** from the joint posterior distribution:

$$p(\beta, \sigma^2, \mathcal{I}|\mathbf{y},\mathbf{X}) = p(\beta, \sigma^2|\mathcal{I}, \mathbf{y},\mathbf{X})p(\mathcal{I}|\mathbf{y}, \mathbf{X}).$$

- Simulate from $p(\mathcal{I}|\mathbf{y}, \mathbf{X})$ using **Gibbs sampling**:
  - Draw $I_1|\mathcal{I}_{-1}, \mathbf{y}, \mathbf{X}$
  - Draw $I_2|\mathcal{I}_{-2},\mathbf{y}, \mathbf{X}$
  - ...
  - Draw $I_p|\mathcal{I}_{-p}, \mathbf{y}, \mathbf{X}$
  - Draw β, σ^2 from $p(\beta, \sigma^2|\mathcal{I}, \mathbf{y},\mathbf{X})$.
- Compute $p(\mathcal{I}|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\mathbf{X}, \mathcal{I}) \cdot p(\mathcal{I})$ for $I_i = 0$ and for $I_i = 1$, and normalize.
- **Model averaging** in a single simulation run.

Simple general Bayesian variable selection

- The previous algorithm only works when we can compute

$$p(\mathcal{I}|\mathbf{y}, \mathbf{X}) = \int p(\beta, \sigma^2, \mathcal{I}|\mathbf{y}, \mathbf{X}) d\beta d\sigma$$

- **MH** - **propose** β and $\mathcal{I}$ jointly from the proposal distribution

$$q(\beta_p|\beta_c, \mathcal{I}_p) q(\mathcal{I}_p|\mathcal{I}_c)$$

- Main difficulty: how to propose the non-zero elements in β_p?
- Simple approach:
  - Approximate posterior with **all** variables in the model:

$$\beta|\mathbf{y}, \mathbf{X} \overset{approx}{\sim} N\left[\hat{\beta}, J_{\mathbf{y}}^{-1}(\hat{\beta})\right]$$

  - Propose β_p from $N\left[\hat{\beta}, J_{\mathbf{y}}^{-1}(\hat{\beta})\right]$, conditional on the zero restrictions implied by $\mathcal{I}_p$. Formulas are available.

Variable selection in more complex models

Posterior summary of the one-component split-t model.[a]

Parameters	Mean	Stdev	Post.Incl.
Location μ			
Const	0.084	0.019	–
Scale ϕ			
Const	0.402	0.035	–
LastDay	−0.190	0.120	0.036
LastWeek	**−0.738**	**0.193**	**0.985**
LastMonth	**−0.444**	**0.086**	**0.999**
CloseAbs95	0.194	0.233	0.035
CloseSqr95	0.107	0.226	0.023
MaxMin95	**1.124**	**0.086**	**1.000**
CloseAbs80	0.097	0.153	0.013
CloseSqr80	0.143	0.143	0.021
MaxMin80	−0.022	0.200	0.017
Degrees of freedom ν			
Const	2.482	0.238	–
LastDay	0.504	0.997	0.112
LastWeek	**−2.158**	**0.926**	**0.638**
LastMonth	0.307	0.833	0.089
CloseAbs95	0.718	1.437	0.229
CloseSqr95	1.350	1.280	0.279
MaxMin95	1.130	1.488	0.222
CloseAbs80	0.035	1.205	0.101
CloseSqr80	0.363	1.211	0.112
MaxMin80	−1.672	1.172	0.254
Skewness λ			
Const	−0.104	0.033	–
LastDay	−0.159	0.140	0.027
LastWeek	−0.341	0.170	0.135
LastMonth	−0.076	0.112	0.016
CloseAbs95	−0.021	0.096	0.008
CloseSqr95	−0.003	0.108	0.006
MaxMin95	0.016	0.075	0.008
CloseAbs80	0.060	0.115	0.009
CloseSqr80	0.059	0.111	0.010
MaxMin80	0.093	0.096	0.013

Model averaging

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

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

  $p_k(\gamma|\mathbf{y})$ is the marginal posterior of γ conditional on M_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)

Posterior predictive analysis

- If $p(y|\theta)$ is a 'good' model, then the data actually observed should not differ 'too much' from simulated data from $p(y|\theta)$.
- Bayesian: simulate data from the **posterior predictive distribution**:

$$p(y^{rep}|y) = \int p(y^{rep}|\theta)p(\theta|y)d\theta.$$

- Difficult to compare y and y^{rep} because of dimensionality.
- Solution: compare **low-dimensional statistic** $T(y,\theta)$ to $T(y^{rep},\theta)$.
- Evaluates the full probability model consisting of both the likelihood *and* prior distribution.

Posterior predictive analysis

- **Algorithm** for simulating from the posterior predictive density $p[T(y^{rep})|y]$:

1. Draw a $\theta^{(1)}$ from the posterior $p(\theta|y)$.
2. Simulate a data-replicate $y^{(1)}$ from $p(y^{rep}|\theta^{(1)})$.
3. Compute $T(y^{(1)})$.
4. Repeat steps 1-3 a large number of times to obtain a sample from $T(y^{rep})$.

- We may now compare the observed statistic $T(y)$ with the distribution of $T(y^{rep})$.
- **Posterior predictive p-value**: $\Pr[T(y^{rep}) \geq T(y)]$
- Informal graphical analysis.

Posterior predictive analysis - Normal model, max statistic