Bayesian Learning

Lecture 10 - Probabilistic programming and Bayesian model comparison

Mattias Villani

Department of Statistics
Stockholm University

mattiasvillani.com @matvil @matvil mattiasvillani

Overview

- Probabilistic programming with Stan
- Bayesian model comparison
- Marginal likelihood
- Model averaging

Probabilistic programming with Stan

- See this Quarto notebook for an introduction to `Stan` and the `loo` package for model evaluation using iid Normal model as the running example.

Bayesian model comparison

- **Posterior model probabilities**

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

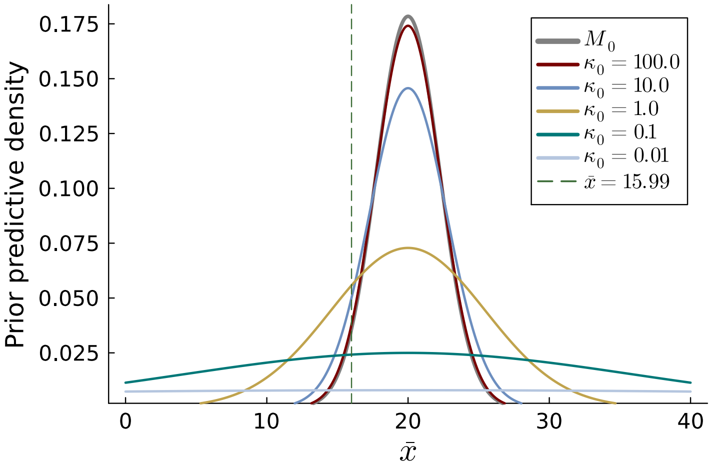
- The **marginal likelihood** for model M_k with parameters θ_k

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

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

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

Internet speed data - prior predictive density



- See this interactive notebook for an example with the iid normal model with known variance.

Laplace approximation

- Taylor approximation of the log posterior around the posterior mode $\tilde{\theta}$ gives:

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

- So integrating both sides with respect to θ gives

$$
p(\mathbf{y}) = \int p(\mathbf{y}|\theta)p(\theta)d\theta = p(\mathbf{y}|\tilde{\theta})p(\tilde{\theta})(2\pi)^{p/2}\left|J_{\mathbf{y}}^{-1}(\tilde{\theta})\right|^{1/2}
$$

Laplace approximation

- **The Laplace approximation** of the log marginal likelihood:

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

where p is the number of unrestricted parameters.

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

Model averaging

- Two models M_1 and M_2, each giving a predictive density:
  - Predictive density Model M_1: $p_1(\tilde{y}|\mathbf{y})$
  - Predictive density Model M_2: $p_2(\tilde{y}|\mathbf{y})$
- Instead of selecting one model, we can **marginalize over the models** using the posterior model probabilities $\Pr(M_k|\mathbf{y})$ in **model averaging**:

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

- 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)