

# BAYESIAN LEARNING - LECTURE 10

Mattias Villani

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

# OVERVIEW

- ▶ Model comparison
- ▶ Model checking

# LIKELIHOOD IS NO GOOD FOR MODEL COMPARISON

- ▶ Consider two models:  $M_1$  and  $M_2$ . Let  $p_i(y|\theta_i)$  denote the data density under model  $M_i$ . If we knew the values of  $\theta_1$  and  $\theta_2$ , then the likelihood ratio

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

could be used to compare the models.

- ▶ What if the model parameters are unknown? The estimated likelihood ratio:

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

where  $\hat{\theta}_1$  and  $\hat{\theta}_2$  are the maximum likelihood estimates.

- ▶ Estimated likelihood ratio is useless in itself as the larger model always has larger likelihood. Comparison with sampling distribution of the estimated likelihood ratio is one solution.

# ENTER BAYES

- ▶ Bayesian: use your priors  $p_1(\theta_1)$  och  $p_2(\theta_2)$  and compute the **marginal likelihood**, or **prior predictive density**, for each model

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

- ▶ The **Bayes factor** can be used to compare to models

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

- ▶ The marginal likelihoods may be converted into posterior probabilities of the models ( $M_1, M_2$ ):

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

where  $B_{12}(y)$  is the Bayes factor in favor of  $M_1$ .

Posterior model odds ratio = Prior model odds ratio · Bayes factor

# 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

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

# BAYESIAN HYPOTHESIS TESTING - BERNOULLI EXAMPLE

- ▶ This is equivalent to the posterior under the following 'spike-and-slab' prior:

$$p(\theta) = \pi l_{\theta_0}(\theta) + (1 - \pi) \text{Beta}(\alpha, \beta)$$

- ▶ Note: data can now *support* a null hypothesis (not only reject it). This is all due to the introduction of a prior.

## BAYESIAN HYPOTHESIS TESTING, CONT.

- ▶ Bayes tests are consistent (not true for frequentist test)

$$p(H_k|\mathbf{x}) \rightarrow 1 \text{ as } n \rightarrow \infty \text{ if } H_k \text{ is true.}$$

- ▶ The priors must be proper. Example: Let  $x_1, \dots, x_n$  be an independent sample from  $N(\theta, 1)$ .

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta \neq \theta_0, \text{ with prior } N(\mu_0, \tau_0^2) \text{ if } H_1 \text{ holds.}$$

Then it can be shown that:

$$p(H_0|\mathbf{x}) \rightarrow 1 \text{ as } \tau_0^2 \rightarrow \infty,$$

regardless of which hypothesis is the true one.

- ▶ This result is entirely in the logic of Bayesian testing!

## EXAMPLE - VARIABLE SELECTION

- ▶ Linear regression:

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

Which variables have non-zero coefficient? Example of hypotheses:

$$H_0 : \beta_0 = \beta_1 = \dots \beta_p = 0$$

$$H_1 : \beta_1 = 0$$

$$H_2 : \beta_1 = \beta_2 = 0$$

we could consider all possible subsets of  $\beta$  coefficients to be zero.  
Easy! Just compute the marginal likelihood of each hypothesis.

- ▶ MCMC sampling algorithms for variable selection. Introduce variable indicators:

$$I_j = \begin{cases} 0 & \text{if } \beta_j = 0 \\ 1 & \text{if } \beta_j \neq 0 \end{cases}$$

- ▶ Sample from the joint posterior  $p(\beta_0, \beta_1, \dots, \beta_p, I_1, I_2, \dots, I_p | y, x)$  using Gibbs sampling (linear Gaussian regression) or Metropolis-Hastings (everything else).



# EXAMPLE: MODEL CHOICE IN MULTIVARIATE TIME SERIES

- ▶ Multivariate time series

$$x_t = \alpha z_t + \Phi_1 x_{t-1} + \dots \Phi_k x_{t-k} + \Psi_1 + \Psi_2 t + \Psi_3 t^2 + \varepsilon_t$$

- ▶ Need to choose:

- ▶ **Lag length**,  $k = 1, 2, \dots, 4$ .
- ▶ **Trend model** ( $s = 1, 2, \dots, 5$ )
- ▶ Number of **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

# MODEL AVERAGING

- ▶ Let  $\gamma$  be a quantity with an interpretation which stays the same across the two models (for example a future value of the data  $\tilde{y}$ ). The marginal posterior distribution of  $\gamma$  reads

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

where  $p_k(\gamma|y)$  is the marginal posterior of  $\gamma$  conditional on model  $k$ .

- ▶ Prediction:  $\gamma = (y_{T+1}, \dots, y_{T+h})'$ .
- ▶ Predictive distribution includes three sources of uncertainty:
  - ▶ Future errors/disturbances (e.g. the  $\varepsilon$ 's in a regression)
  - ▶ Parameter uncertainty (the predictive distribution  $p(\tilde{y}|y)$  has the parameters integrated out by their posteriors)
  - ▶ Model uncertainty (by model averaging)

# MARGINAL LIKELIHOOD AS MEASURE OF OUT-OF-SAMPLE PREDICTIVE PERFORMANCE

- ▶ The marginal likelihood of a sample  $y_1, \dots, y_T$  can be expressed as

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

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

where we assume that  $y_t$  is independent of  $y_1, \dots, y_{t-1}$  conditional on  $\theta$ .

- ▶ The prediction of  $y_1$  is based on the prior of  $\theta$ , and is therefore sensitive to the prior.
- ▶ The prediction of  $y_T$  uses almost all the data to infer  $\theta$ . Very little influenced by the prior when  $T$  is not small.
- ▶ Log Predictive Score.
- ▶ Cross-validation.

# MARGINAL LIKELIHOOD IN CONJUGATE MODELS

- ▶ Computing the marginal likelihood requires integration w.r.t.  $\theta$ .
- ▶ Short cut for conjugate models by rearrangement of Bayes' theorem:

$$p(y) = \frac{p(y|\theta)p(\theta)}{p(\theta|y)}$$

- ▶ Bernoulli model example

$$p(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

$$p(y|\theta) = \theta^s (1-\theta)^f$$

$$p(\theta|y) = \frac{1}{B(\alpha+s, \beta+f)} \theta^{\alpha+s-1} (1-\theta)^{\beta+f-1}$$

- ▶ Marginal likelihood

$$p(y) = \frac{\theta^s (1-\theta)^f \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}}{\frac{1}{B(\alpha+s, \beta+f)} \theta^{\alpha+s-1} (1-\theta)^{\beta+f-1}} = \frac{B(\alpha+s, \beta+f)}{B(\alpha, \beta)}$$

# COMPUTING THE MARGINAL LIKELIHOOD

- Usually difficult to evaluate the integral

$$p(y) = \int p(y|\theta)p(\theta)d\theta = E_{p(\theta)}[p(y|\theta)].$$

- A (naive) first try is to draw from the prior  $\theta^{(1)}, \dots, \theta^{(N)}$  and estimating the marginal likelihood by the average likelihood

$$\hat{p}(y) = \frac{1}{N} \sum_{i=1}^N p(y|\theta^{(i)}).$$

Unstable if the posterior is very different from the prior.

- Importance sampling. Let  $\theta^{(1)}, \dots, \theta^{(N)}$  be iid draws from some density  $g(\theta)$ .

$$\begin{aligned} p(y) &= \int p(y|\theta)p(\theta)d\theta = \int \frac{p(y|\theta)p(\theta)}{g(\theta)}g(\theta)d\theta \\ &= E_g \left[ \frac{p(y|\theta)p(\theta)}{g(\theta)} \right] \approx N^{-1} \sum_{i=1}^N \frac{p(y|\theta^{(i)})p(\theta^{(i)})}{g(\theta^{(i)})}. \end{aligned}$$

# COMPUTING THE MARGINAL LIKELIHOOD, CONT.

- ▶ Rearrangement of Bayes' theorem:

$$p(y) = \frac{p(y|\theta)p(\theta)}{p(\theta|y)}.$$

- ▶ Problem: we must know the posterior, **including** the normalization constant. The  $\propto$  trick does not work here.
- ▶ But we only need to know  $p(\theta|y)$  in a single point  $\theta_0$ .
- ▶ Kernel density estimator may be used to approximate  $p(\theta_0|y)$ . Unstable. Chib (1995, JASA) and Chib-Jeliazkov (2001, JASA) provide better solutions.
- ▶ Reversible Jump MCMC (RJMCMC) for model inference.
  - ▶ MCMC methods can be extended to not only move in the parameter space for a given model, but also jumping between models.
  - ▶ The proportion of iterations spent in model  $k$  is an estimate of  $\Pr(M_k|y)$ .

# APPROXIMATE MARGINAL LIKELIHOODS

- ▶ Taylor approximation of the log posterior

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

$$p(\mathbf{y}|\theta)p(\theta) \approx p(\mathbf{y}|\hat{\theta})p(\hat{\theta}) \exp \left[ -\frac{1}{2}J_{\hat{\theta},\mathbf{y}}(\theta - \hat{\theta})^2 \right],$$

which can be integrated analytically w.r.t.  $\theta$ , using properties of the multivariate normal pdf.

- ▶ The Laplace approximation:

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

where  $p$  is the number of unrestricted parameters in the model.

- ▶ Cruder version of the Laplace: The SBC (BIC) approximation

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

# BAYESIAN MODEL INFERENCE - A CRITIQUE

- ▶ Bayes factors (model probabilities) are very sharp inference objects. Handle with care.
- ▶ Minor differences in the prior can lead to large differences in the Bayes factor, especially in high-dimensional non-linear models.
- ▶ Continuous model expansion is usually a better alternative, when feasible.
- ▶ Improper priors cannot be used to compute Bayes factors. Several tricks have been developed to handle this, but they are non-Bayesian.
- ▶ Bayes factors are relative measures, all models under consideration may be bad approximations to the data.
- ▶ Bayes model probabilities essentially assume the true data generating process (DGP) is among the compared models. Box: All models are false, but some are useful.
- ▶ When none of the compared models are true:  $Pr(M_i|y) \rightarrow 1$  for the model which is closest to the DGP in the Kullback-Leibler sense. Putting all eggs in one basket is not always a good idea.'



# MODELS - WHY?

- ▶ We now know how to **compare** models.
- ▶ But how do we know if any given model is 'any good'?
- ▶ George Box: '**All models are false, but some are useful**'.

# WHAT IS YOUR MODEL FOR REALLY?

## ► Prediction.

- Interpretation not a concern
- Black-box approach may be ok.
- Extrapolation?
- Model averaging may be a good idea.

## ► Abstraction to **aid in thinking** about a phenomena.

- Prediction accuracy of less concern.
- Model averaging may be a bad idea.

## ► Model as a **compact description of a complex phenomena**.

- Computational cost of model evaluation may be a concern.
- Online/real-time analysis.

# 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.

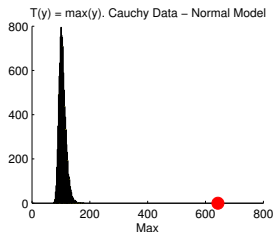
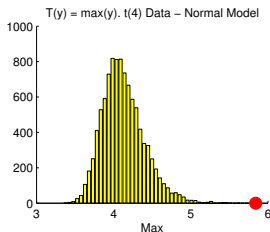
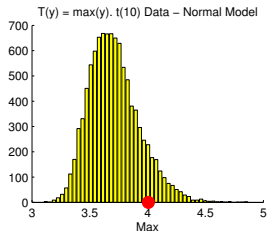
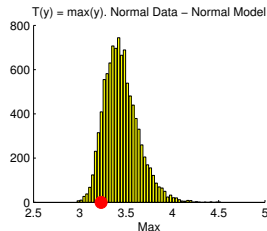
## POSTERIOR PREDICTIVE ANALYSIS, CONT.

- ▶ **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 - EXAMPLES

- ▶ Ex. 1. Model:  $y_1, \dots, y_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$ .  $T(y) = \max_i |y_i|$ .
- ▶ Ex. 2. Assumption of no reciprocity in networks.  
 $y_{ij} | \theta \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$ .  $T(y)$  = proportion of reciprocated node pairs.
- ▶ Ex. 3. ARIMA-process.  $T(y)$  may be the autocorrelation function.
- ▶ Ex. 4. Poisson regression.  $T(y)$  frequency distribution of the response counts. Proportions of zero counts.

# POSTERIOR PREDICTIVE ANALYSIS - NORMAL MODEL, MAX STATISTIC



# THE FIT OF A MIXTURE OF POISSON REGRESSIONS

