Bayesian Learning

Lecture 11 - Bayesian Model Comparison. Variable selection.

Bertil Wegmann

Department of Computer and Information Science
Linköping University

LINKÖPING UNIVERSITY

Overview

- Computing the marginal likelihood
- Information criteria
- Bayesian variable selection
- Model averaging

Marginal likelihood in conjugate models

- **Marginal likelihood**: $\int p(\mathrm{y}|\theta)p(\theta)d\theta$. Integration!
- Short cut for **conjugate models**:

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

- 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

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

- **Monte Carlo estimate**. Draw from the prior $\theta^{(1)}, ..., \theta^{(N)}$ and

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

  Unstable when posterior is different from prior.
- **Importance sampling**. Let $\theta^{(1)}, ..., \theta^{(N)}$ be draws from $g(\theta)$.

$$\int p(\mathrm{y}|\theta)p(\theta)d\theta = \int \frac{p(\mathrm{y}|\theta)p(\theta)}{g(\theta)}g(\theta)d\theta \approx N^{-1}\sum_{i=1}^{N}\frac{p(\mathrm{y}|\theta^{(i)})p(\theta^{(i)})}{g(\theta^{(i)})}$$

- **Modified Harmonic mean**: $g(\theta) = N(\tilde{\theta}, \tilde{\Sigma}) \cdot I_c(\theta)$, where $\tilde{\theta}$ and $\tilde{\Sigma}$ is the posterior mean and covariance matrix estimated from MCMC, and $I_c(\theta) = 1$ if $(\theta - \tilde{\theta})'\tilde{\Sigma}^{-1}(\theta - \tilde{\theta}) \leq c$.

Laplace approximation

- Taylor approximation of the log likelihood

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

so

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

- **The Laplace approximation**:

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

where p is the number of unrestricted parameters.

BIC approximation

- The Laplace approximation:

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

- $\hat{\theta}$ and $J_{\hat{\theta},\mathrm{y}}$ can be obtained with optimization.
- The observed information at the mode can be written as $J_{\hat{\theta},\mathrm{y}} = n\overline{J_{\hat{\theta}}}$, where $\overline{J_{\hat{\theta}}}$ is the average observed information per observation. This gives $\frac{1}{2}\ln\left|J_{\hat{\theta},\mathrm{y}}^{-1}\right| = -\frac{p}{2}\ln n + \frac{1}{2}\ln\left|\overline{J_{\hat{\theta}}}^{-1}\right|$.
- The BIC approximation assumes for large samples that the small terms $\frac{1}{2}\ln\left|\overline{J_{\hat{\theta}}}^{-1}\right|$, $\frac{p}{2}\ln(2\pi)$ and $\ln p(\hat{\theta})$ are ignored.

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

Information criteria and AIC

- **Information criteria**: measures the predictive accuracy of a model.

- **Based on information theory**. Theoretical estimate of the relative out-of-sample (test sample) KL divergence from a model to the "perfect" model.

- **AIC**: approximates the predictive accuracy of a model and works in the following settings:
  1. flat priors
  2. posterior is approximately normal
  3. The number of observations n is much larger than the number of parameters p, i.e. $n >> p$.

$$AIC = -2 \log p \left(y | \hat{\theta}_{ML} \right) + 2p$$

Information criteria: DIC

- **DIC**: "Bayesian version" of AIC. OK with informative priors, but conditions 2. and 3. for AIC also applies to DIC.

$$DIC = -2 \log p\left(y|\hat{\theta}_{Bayes}\right) + 2p_{DIC}$$

$$p_{DIC} = 2\left[\log p\left(y|\hat{\theta}_{Bayes}\right) - \mathrm{E}_{post}\left[\log\left(p\left(y|\theta\right)\right)\right]\right]$$

$$\mathrm{E}_{post}\left[\log\left(p\left(y|\theta\right)\right)\right] = \frac{1}{S}\sum_{s=1}^{S} \log p\left(y|\theta^{(s)}\right)$$

.

Information criteria: WAIC

- **WAIC**: Widely Applicable Information Criteria. More general than AIC and DIC.

$$WAIC = -2\text{lppd} + 2p_{WAIC}$$

$$\text{lppd} = \sum_{i=1}^{n} \log\left(\text{E}_{post}\left[p\left(y_i|\theta\right)\right]\right)$$

$$p_{WAIC} = \sum_{i=1}^{n} V_{s=1}^{S}\left(\log p\left(y_i|\theta^{(s)}\right)\right),$$

where $V_{s=1}^{S}$ is the sample variance.

- Requires the data to be partitioned in n pieces. Might be problematic for structured-data settings, e.g. time series, spatial, and network data.

Information criteria: Bayesian LOO-CV

- **Bayesian leave-one-out cross validation (LOO-CV)**: WAIC is asymptotically (as $n \to \infty$) equal to Bayesian LOO-CV.

$$LOO - CV = -2\mathrm{lppd}_{loo-cv} = -2\mathrm{lppd} + 2p_{loo-cv}$$

$$\mathrm{lppd}_{loo-cv} = \sum_{i=1}^{n} \log\left(\mathrm{E}_{post}\left[p\left(y_i|\theta^{(i)}\right)\right]\right)$$

$$p_{loo-cv} = \mathrm{lppd} - \mathrm{lppd}_{loo-cv}$$

- As for WAIC: requires the data to be partitioned in n pieces.

Bayesian variable selection

- Linear regression:

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

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

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

- 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:

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

- 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 ($\mathcal{I}$)

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

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**

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

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

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

Bayesian variable selection via Gibbs sampling

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

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

- Simulate from $p(\mathcal{I}|\mathrm{y}, \mathrm{X})$ using **Gibbs sampling**:
  - Draw $I_1|\mathcal{I}_{-1}, \mathrm{y}, \mathrm{X}$
  - Draw $I_2|\mathcal{I}_{-2},\mathrm{y}, \mathrm{X}$
  - ...
  - Draw $I_p|\mathcal{I}_{-p}, \mathrm{y}, \mathrm{X}$
- Note that: $Pr(I_i = 0|\mathcal{I}_{-i}, \mathrm{y}, \mathrm{X}) \propto Pr(I_i = 0, \mathcal{I}_{-i}|\mathrm{y}, \mathrm{X})$.
- Compute $p(\mathcal{I}|\mathrm{y}, \mathrm{X}) \propto p(\mathrm{y}|\mathrm{X}, \mathcal{I}) \cdot p(\mathcal{I})$ for $I_i = 0$ and for $I_i = 1$.
- **Model averaging** in a single simulation run.
- Simulate from $p(\beta, \sigma^2|\mathcal{I}, \mathrm{y},\mathrm{X})$ for each draw of $\mathcal{I}$.

Simple general Bayesian variable selection

- The previous algorithm only works when we can compute

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

- **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:

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

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

Variable selection in more complex models

Wegmann and Villani (2011)

Journal of Business & Economic Statistics, July 2011

Table 5. Comparing the posterior inference for the eBay data from the Gaussian and Gamma models

		Mean		SD		Incl. prob.	
Parameter	Covariate	Gauss	Gamma	Gauss	Gamma	Gauss	Gamma
κ/τ	–	5.499	2.997	0.772	0.111	1.000	1.000
μ	Const	28.273	28.307	0.245	0.304	1.000	1.000
	Book_d	0.740	0.747	0.010	0.012	1.000	1.000
	$\text{Book} \cdot \text{Power}_d$	0.033	0.046	0.015	0.018	0.064	0.107
	Book · ID	0.128	0.052	0.036	0.039	0.900	0.017
	Book · Sealed	0.372	0.488	0.029	0.051	1.000	1.000
	Book · MinBlem	−0.022	0.002	0.021	0.028	0.010	0.008
	Book · MajBlem	−0.252	−0.269	0.030	0.040	1.000	1.000
	Book · LargNeg	−0.003	−0.020	0.018	0.025	0.004	0.009
$\log(\sigma^2)$	Const	3.997	4.314	0.071	0.038	1.000	1.000
	LBook_d	1.262	1.276	0.038	0.026	1.000	1.000
	LBook · Power	0.043	0.069	0.018	0.020	0.220	1.000
	LBook · ID	0.042	0.032	0.040	0.067	0.481	0.011
	LBook · Sealed	0.211	0.362	0.027	0.019	1.000	1.000
	LBook · MinBlem	−0.028	−0.057	0.027	0.026	0.012	0.039
	LBook · MajBlem	0.036	0.063	0.040	0.049	0.007	0.017
	LBook · NegScore	0.035	0.042	0.021	0.027	0.017	0.050
$\log(\lambda)$	Const	1.193	1.234	0.021	0.022	1.000	1.000
	Power	0.009	−0.028	0.035	0.029	0.005	0.012
	ID	−0.177	−0.197	0.110	0.078	0.030	0.048
	Sealed	0.323	0.331	0.048	0.048	1.000	1.000
	MinBlem	−0.049	−0.042	0.048	0.048	0.008	0.009
	MajBlem	−0.151	−0.115	0.085	0.097	0.019	0.015
	NegScore	0.055	0.086	0.049	0.047	0.012	0.022
	LBook_d	−0.038	−0.036	0.027	0.021	0.018	0.031
	MinBidShare_d	−1.433	−1.380	0.056	0.059	1.000	1.000

NOTE: The approximate bid function was used for both models. $c = m, \bar{\kappa} = 0.25, g = 4$, and $\pi = 0.2$. $x_1 \cdot x_2$ denotes the interaction of x_1 and x_2.

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

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

  $p_k(\gamma|\mathrm{y})$ 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)