Bayesian Learning

Lecture 11 - Computations. Variable selection.

Mattias Villani

Department of Statistics
Stockholm University

Department of Computer and Information Science
Linköping University

LINKÖPING UNIVERSITY

Overview

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

Marginal likelihood in conjugate models

- **Marginal likelihood**: $\int p(\mathbf{y}|\theta)p(\theta)d\theta$. 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

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

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

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

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

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

- **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 $(\theta - \tilde{\theta})'\tilde{\Sigma}^{-1}(\theta - \tilde{\theta}) \leq c$.

Computing the marginal likelihood, cont.

- To use $p(\mathbf{y}) = p(\mathbf{y}|\theta)p(\theta)/p(\theta|\mathbf{y})$ we need $p(\theta|\mathbf{y})$.
- But we only need to know $p(\theta|\mathbf{y})$ in a single point θ_0.
- Kernel density estimator to approximate $p(\theta_0|\mathbf{y})$. Unstable.
- Chib's method (1995, JASA). Great, but only **Gibbs sampling**.
- Chib-Jeliazkov (2001, JASA) generalizes to **MH algorithm** (good for IndepMH, terrible for RWM).
- Reversible Jump MCMC (RJMCMC) for model inference.
  - MCMC methods that moves in model space.
  - Proportion of iterations spent in model k estimates $\Pr(M_k|\mathbf{y})$.
  - Usually hard to find efficient proposals. Sloooow convergence.
- Bayesian nonparametrics (e.g. Dirichlet process priors).

Laplace approximation

- Taylor approximation of the log likelihood

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

so

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

- **The Laplace approximation**:

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

where p is the number of unrestricted parameters.

BIC

- **The Laplace approximation**:

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

- $\hat{\theta}$ and $J_{\hat{\theta},\mathbf{y}}$ can be obtained with **optimization/autodiff**.
- The **BIC approximation** assumes that $J_{\hat{\theta},\mathbf{y}}$ behaves like $n \cdot I_p$ in large samples and the small term $\frac{p}{2}\ln(2\pi)$ is ignored

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

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 $I_p|\mathcal{I}_{-p}, \mathbf{y}, \mathbf{X}$
- Note that: $Pr(I_i = 0|\mathcal{I}_{-i}, \mathbf{y}, \mathbf{X}) \propto Pr(I_i = 0, \mathcal{I}_{-i}|\mathbf{y}, \mathbf{X})$.
- 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 $I_i = 1$.
- **Model averaging** in a single simulation run.
- If needed, simulate from $p(\beta, \sigma^2|\mathcal{I}, \mathbf{y},\mathbf{X})$ for each draw of $\mathcal{I}$.

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)