Advanced Bayesian Learning

Regularization and Variable Selection - Lecture 2

Mattias Villani

Department of Statistics

Lecture overview

- Spike-and-slab variable selection regression
- Polya-Gamma augmentation for logistic regression
- Extensions

Bayesian variable selection

- Linear regression

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

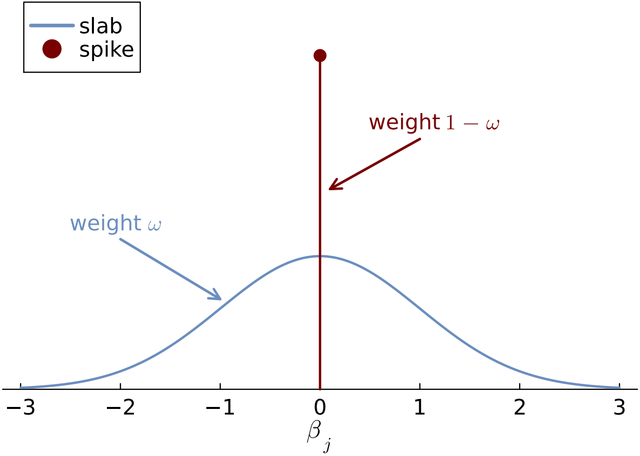
- Which variables have non-zero coefficients?
- Introduce **variable selection indicators** $\boldsymbol{z} = (z_1, \ldots, z_p)$.
- Example: $\boldsymbol{z} = (1, 1, 0)$ means that $\beta_1 \neq 0$ and $\beta_2 \neq 0$, but $\beta_3 = 0$, so x_3 drops out of the model.
- **Spike-and-slab** prior

$$z_1, \ldots, z_p \sim \mathrm{Bernoulli}(\omega)$$

$$\beta_j | z_j \sim \begin{cases} N(0, \tau^2\sigma^2) & \text{if } z_j = 1 \\ = 0 & \text{if } z_j = 0 \end{cases}$$

- **Prior inclusion probability** ω.

Spike-and-slab prior



,

Bayesian variable selection

- Posterior

$$p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{z}|\boldsymbol{y}, \boldsymbol{X}) = p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{X})p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{X})$$

$$p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{X}) \propto p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{z}) \cdot p(\boldsymbol{z})$$

- The prior $p(\boldsymbol{z})$ is $z_1, \ldots, z_p \sim \mathrm{Bernoulli}(\omega)$ as before.
- Need the **marginal likelihood** $p(\mathrm{y}|\mathrm{X}, \boldsymbol{z})$ for each model $\boldsymbol{z}$.

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{z}) = \int p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{z})p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{z})d\boldsymbol{\beta}$$

- For **linear Gaussian regression** the marginal likelihood is

$$t_{\nu_{0,\boldsymbol{z}}}\left(\boldsymbol{y}|0, \sigma^2_{0,\boldsymbol{z}}(\boldsymbol{I}_n + \boldsymbol{X}_{\boldsymbol{z}}\boldsymbol{\Omega}^{-1}_{0,\boldsymbol{z}}\boldsymbol{X}^{\top}_{\boldsymbol{z}})\right)$$

  where $t_{\nu_{0,\boldsymbol{z}}}$ is the multivariate-t density and $\boldsymbol{X}_{\boldsymbol{z}}$ is the matrix of covariates selected by $\boldsymbol{z}$.
- Prior hyperparameters ν_0, σ^2_0 and $\boldsymbol{\Omega}_0$ allowed to depend on $\boldsymbol{z}$.

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:

$$p(\boldsymbol{\beta}, \sigma^2, \mathbf{z}|\mathbf{y}, \mathbf{X}) = p(\boldsymbol{\beta}, \sigma^2|\mathbf{z}, \mathbf{y}, \mathbf{X})p(\mathbf{z}|\mathbf{y}, \mathbf{X})$$

- Simulate from $p(\mathbf{z}|\mathbf{y}, \mathbf{X})$ using **Gibbs sampling**:
  - Draw $z_1|\mathbf{z}_{-1}, \mathbf{y}, \mathbf{X}$
  - Draw $z_2|\mathbf{z}_{-2}, \mathbf{y}, \mathbf{X}$
  - ...
  - Draw $z_p|\mathbf{z}_{-p}, \mathbf{y}, \mathbf{X}$
  - Draw $\boldsymbol{\beta}, \sigma^2$ from $p(\boldsymbol{\beta}, \sigma^2|\mathbf{z}, \mathbf{y}, \mathbf{X})$.
- Compute $p(\mathbf{z}|\mathrm{y}, \mathrm{X}) \propto p(\mathrm{y}|\mathrm{X}, \mathbf{z}) \cdot p(\mathbf{z})$ for $z_j = 0$ and for $z_j = 1$, and normalize.
- **Model averaging** in a single simulation run.

Bayesian variable selection algorithm

Gibbs sampling for Bayesian variable selection in regression

Input: $n \times p$ matrix with p covariates as columns $\mathbf{X}$
vector $\mathbf{y}$ with response observations
slab variance τ^2
prior inclusion probability ω
initial variable indicators $\mathbf{z}^{(0)} = (z_1^{(0)}, \ldots, z_p^{(0)})$
number of posterior draws m.

for j in *1:m* **do**

// Update regression parameters

Draw $(\sigma^2)^{(j)} | \mathbf{y}, \mathbf{X}_{\mathbf{z}^{(j-1)}} \sim \text{ScaledInv}-\chi^2(\nu_n, \sigma_n^2)$

Draw $\boldsymbol{\beta}_{\mathbf{z}^{(j-1)}}^{(j)} | (\sigma^2)^{(j)}, \mathbf{y}, \mathbf{X}_{\mathbf{z}^{(j-1)}} \sim N(\boldsymbol{\mu}_n, (\sigma^2)^{(j)} \Omega_n^{-1})$

Set $\boldsymbol{\beta}^{(j)}[\mathbf{z}^{(j-1)}] = \boldsymbol{\beta}_{\mathbf{z}^{(j-1)}}^{(j)}$ and $\boldsymbol{\beta}^{(j)}[\texttt{Not}(\mathbf{z}^{(j-1)})] = 0$

// Update mixture allocations

Set $\tilde{\mathbf{z}} = \mathbf{z}^{(j-1)}$

for k in *1:p* **do**

Set $\tilde{\mathbf{z}}_0$ to $\tilde{\mathbf{z}}$ but with kth element equal to 0

Set $\tilde{\mathbf{z}}_1$ to $\tilde{\mathbf{z}}$ but with kth element equal to 1

Compute $\tilde{\omega}_{k,0} \propto (1-\omega) \cdot p(\mathbf{y}|\mathbf{X}_{\tilde{\mathbf{z}}_0})$

Compute $\tilde{\omega}_{k,1} \propto \omega \cdot p(\mathbf{y}|\mathbf{X}_{\tilde{\mathbf{z}}_1})$

Normalize $\tilde{\omega}_{k,0}$ and $\tilde{\omega}_{k,1}$ to sum to one

Simulate allocation $z_k^{(j)} \sim \text{Bernoulli}(\tilde{\omega}_{k,1})$

Update $\tilde{\mathbf{z}}$ with the new allocation $z_k^{(j)}$

end

Set $\mathbf{z}^{(j)} = \tilde{\mathbf{z}}$

end

Bayesian variable selection algorithm

for j in *1:m* **do**

// Update regression parameters

Draw $(\sigma^2)^{(j)}|\mathbf{y}, \mathbf{X}_{\mathbf{z}^{(j-1)}} \sim \text{ScaledInv}-\chi^2(\nu_n, \sigma_n^2)$

Draw $\boldsymbol{\beta}^{(j)}_{\mathbf{z}^{(j-1)}}|(\sigma^2)^{(j)}, \mathbf{y}, \mathbf{X}_{\mathbf{z}^{(j-1)}} \sim N(\boldsymbol{\mu}_n, (\sigma^2)^{(j)}\Omega_n^{-1})$

Set $\boldsymbol{\beta}^{(j)}[\mathbf{z}^{(j-1)}] = \boldsymbol{\beta}^{(j)}_{\mathbf{z}^{(j-1)}}$ and $\boldsymbol{\beta}^{(j)}[\texttt{Not}(\mathbf{z}^{(j-1)})] = 0$

// Update mixture allocations

Set $\tilde{\mathbf{z}} = \mathbf{z}^{(j-1)}$

for k in *1:p* **do**

Set $\tilde{\mathbf{z}}_0$ to $\tilde{\mathbf{z}}$ but with kth element equal to 0

Set $\tilde{\mathbf{z}}_1$ to $\tilde{\mathbf{z}}$ but with kth element equal to 1

Compute $\tilde{\omega}_{k,0} \propto (1-\omega) \cdot p(\mathbf{y}|\mathbf{X}_{\tilde{\mathbf{z}}_0})$

Compute $\tilde{\omega}_{k,1} \propto \omega \cdot p(\mathbf{y}|\mathbf{X}_{\tilde{\mathbf{z}}_1})$

Normalize $\tilde{\omega}_{k,0}$ and $\tilde{\omega}_{k,1}$ to sum to one

Simulate allocation $z_k^{(j)} \sim \text{Bernoulli}(\tilde{\omega}_{k,1})$

Update $\tilde{\mathbf{z}}$ with the new allocation $z_k^{(j)}$

end

Set $\mathbf{z}^{(j)} = \tilde{\mathbf{z}}$

end

Simple general Bayesian variable selection

- The previous algorithm only works when we can compute

$$p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{X}) = \int p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{z}|\boldsymbol{y}, \boldsymbol{X}) d\boldsymbol{\beta} d\sigma$$

- MH - **propose** $\boldsymbol{\beta}$ and $\boldsymbol{z}$ jointly from the proposal distribution

$$q(\boldsymbol{\beta}_p|\boldsymbol{\beta}_c, \boldsymbol{z}_p) q(\boldsymbol{z}_p|\boldsymbol{z}_c)$$

- Main difficulty: how to propose the non-zero elements in $\boldsymbol{\beta}_p$?
- Simple approach:
  - Approximate posterior with **all** variables in the model:

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

  - Propose $\boldsymbol{\beta}_p$ from $N\left[\hat{\boldsymbol{\beta}}, J_{\mathrm{y}}^{-1}(\hat{\boldsymbol{\beta}})\right]$, conditional on the zero restrictions implied by $\boldsymbol{z}_p$. 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 v			
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

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

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

Pólya-Gamma augmentation for logistic regression

- **Logistic regression**

$$\Pr(y = y_i | \boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_i^\top \beta)^{y_i}}{1 + \exp(\boldsymbol{x}_i^\top \beta)} \quad \text{for } y_i \in \{0, 1\}$$

$$\Pr(y_1, \dots, y_n | \boldsymbol{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \Pr(y = y_i | \boldsymbol{x}_i, \boldsymbol{\beta})$$

- The key identity

$$\frac{(e^\psi)^a}{(1 + e^\psi)^b} = 2^{-b} e^{\kappa \psi} \int_0^\infty e^{\omega \psi^2 / 2} p(\omega) d\omega,$$

where $\kappa = a - b/2$ and $p(\omega)$ is the density of the Pólya-Gamma distribution

$$\omega \sim \mathrm{PG}(b, 0)$$

Pólya-Gamma augmentation for logistic regression

- So for each term in the likelihood function:

$$\frac{\exp(\boldsymbol{x}_i^\top \beta)^{y_i}}{1+\exp(\boldsymbol{x}_i^\top \beta)} = \frac{1}{2} e^{\kappa_i \boldsymbol{x}_i^\top \beta} \int_0^\infty e^{\omega_i (\boldsymbol{x}_i^\top \beta)^2/2} p(\omega_i) d\omega_i,$$

- The likelihood conditional on $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$ is

$$\prod_{i=1}^n \frac{\exp(\boldsymbol{x}_i^\top \beta)^{y_i}}{1+\exp(\boldsymbol{x}_i^\top \beta)} \propto \prod_{i=1}^n e^{\kappa_i \boldsymbol{x}_i^\top \beta} e^{\omega_i (\boldsymbol{x}_i^\top \beta)^2/2}$$

$$= \exp\left(\sum_{i=1}^n \kappa_i \boldsymbol{x}_i^\top \beta + \frac{\omega_i (\boldsymbol{x}_i^\top \beta)^2}{2} \right)$$

which is an exponential of a quadratic form in $\boldsymbol{\beta}$.

- Hence

$$\boldsymbol{\beta} \sim N(\mu_0, \Sigma_0) \quad \Longrightarrow \quad \boldsymbol{\beta} | \boldsymbol{\omega}, \boldsymbol{y}, \boldsymbol{X} \sim N(\mu_n, \Sigma_n)$$

Pólya-Gamma distribution

Pólya-Gamma distribution

$X \sim \mathrm{PG}(b, c)$ for $X > 0$.

A Pólya-Gamma is defined as a infinite weighted sum (convolution) of iid Gamma distributed variables

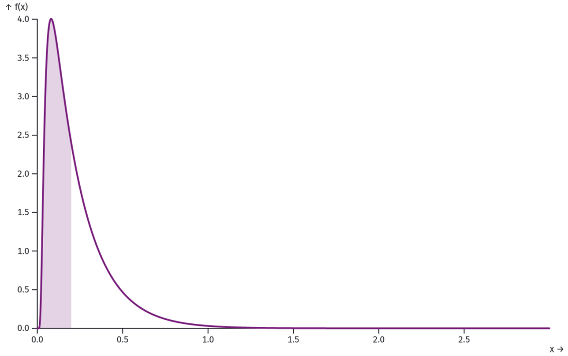
$$X \overset{d}{=} \sum_{k=1}^{\infty} v_k Y_k$$

where $\overset{d}{=}$ mean equality in distribution, the weights are

$$v_k = \frac{1}{2(k-1/2)^2\pi^2 + c^2/2}$$

and $Y_k \overset{\mathrm{iid}}{\sim} \mathrm{Gamma}(b, 1)$.

Pólya-Gamma distribution



Pólya-Gamma augmentation for logistic regression

Gibbs sampling for logistic regression using Pólya-Gamma augmentation

Input: response vector $\mathbf{y} = (y_1, \ldots, y_n)^\top$
matrix $(n \times p)$ with covariates $\mathbf{X}$
initial value $\boldsymbol{\beta}^{(0)}$
number of posterior draws m.

$\boldsymbol{\kappa} \leftarrow (y_1 - 1/2, \ldots, y_n - 1/2)^\top$

for k in $1{:}m$ **do**

// Update Pólya-Gamma variables

for i in $1{:}n$ **do**

$\omega_i^{(k)} | \boldsymbol{\beta}^{(k-1)}, \mathbf{y}, \mathbf{x}_i \sim \mathrm{PG}(1, \mathbf{x}_i^\top \boldsymbol{\beta}^{(k-1)})$

end

$\boldsymbol{\Omega}^{(k)} \leftarrow \mathrm{Diag}(\omega_1^{(k)}, \ldots, \omega_n^{(k)})$

// Update $\boldsymbol{\beta}$

$\boldsymbol{\Sigma}_n \leftarrow (\mathbf{X}^\top \boldsymbol{\Omega}^{(k)} \mathbf{X} + \boldsymbol{\Sigma}_0^{-1})^{-1}$

$\boldsymbol{\mu}_n \leftarrow \boldsymbol{\Sigma}_n (\mathbf{X}^\top \boldsymbol{\kappa} + \boldsymbol{\Sigma}_0^{-1} \mu_0)$

Draw $\boldsymbol{\beta}^{(k)} | \boldsymbol{\omega}$ from $N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$

end

Output: m draws $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(m)}$ from the posterior distribution $p(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X})$.