Bayesian Statistics

Fabio Sigrist

ETH Zurich, Autumn Semester 2019

Today's topics

- Bayesian linear regression model
- Bayesian variable selection

Bayesian linear regression model

We consider here the **linear regression model**

$$y = \alpha \mathbf{1} + X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

where

- y is an $n \times 1$ vector of responses
- $\mathbf{1} = (1, \ldots, 1)^T$
- X is the $n \times p$ design matrix
  - the j-th column of X contains the values of the j-th explanatory variable
  - we assume that all columns are centered: $X^T \mathbf{1} = 0$
- α is the intercept, β the $p \times 1$ regression parameter
- ε is the $n \times 1$ vector of errors

Model selection

We also want to do **selection of explanatory variables**.

- We denote by γ an element of $\{0, 1\}^p$ where $\gamma_j = 1$ iff the j-th variable is selected
- Let β_γ be the sub-vector that contains only the selected components and X_γ the corresponding submatrix
- The number of explanatory variables included in the model γ is denoted by $|\gamma|$

Then the model indexed by γ is

$$y = \alpha\mathbf{1} + X_\gamma\beta_\gamma + \varepsilon$$

Bayesian inference

- Our goal is to compute the posterior of the unknowns $(\gamma, \beta_\gamma, \alpha, \sigma^2)$
- For doing this, we need to specify a likelihood and a prior
- In the following, we first assume that γ is fixed (i.e., no model selection)

Bayesian linear regression model

Bayesian linear regression model: likelihood

- The **likelihood** is given by

$$(\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(y - \alpha\mathbf{1} - X_\gamma\beta_\gamma)^T(y - \alpha\mathbf{1} - X_\gamma\beta_\gamma)\right)$$

- This can also be written as

$$(\sigma^2)^{-n/2} \exp\left(-\frac{s_\gamma^2 + n(\widehat{\alpha} - \alpha)^2 + (\beta_\gamma - \widehat{\beta}_\gamma)^T X_\gamma^T X_\gamma(\beta_\gamma - \widehat{\beta}_\gamma)}{2\sigma^2}\right)$$

where

  - $\widehat{\alpha}$ and β_γ are the MLEs:

$$\widehat{\alpha} = \bar{y}, \quad \widehat{\beta}_\gamma = (X_\gamma^T X_\gamma)^{-1} X_\gamma^T y$$

  - s_γ^2 is the residual sum of squares

$$s_\gamma^2 = (y - \widehat{\alpha}\mathbf{1} - X_\gamma\widehat{\beta}_\gamma)^T(y - \widehat{\alpha}\mathbf{1} - X_\gamma\widehat{\beta}_\gamma)$$

See blackboard for derivation

Bayesian linear regression model: prior

As **prior**, we choose

$$\pi(\beta_\gamma, \alpha, \sigma^2) = \pi(\alpha)\pi(\sigma^2)\pi(\beta_\gamma \mid \sigma^2) \propto \pi(\beta_\gamma \mid \sigma^2)\sigma^{-2}$$

- α independent from σ^2 and β_γ
- Univariate Jeffreys priors for α and σ^2
- For β_γ, use the so-called ***g*-prior** of Zellner

$$\beta_\gamma \mid \sigma^2 \sim \mathcal{N}(\beta_\gamma^0, g\sigma^2(X_\gamma^T X_\gamma)^{-1})$$

  - β_γ^0 is the prior mean. Often, one uses $\beta_\gamma^0 = 0$

Comments on the *g*-prior

- Since the design matrix is considered to be known and fixed, we can use it for the prior
- $g > 0$ is a hyperparameter which can be interpreted as a measure of the amount of information available in the prior relative to the data
- The g-prior arises as the posterior from a flat prior and a response vector $y = X_\gamma \beta_\gamma^0$ (i.e., $y = 0$ if $\beta_\gamma^0 = 0$) with the same design matrix X_γ, no intercept, and error variance $g\sigma^2$
- We cannot use a flat prior if we want to do model selection later because this would leave posterior probabilities of different models γ undefined

Bayesian linear regression model: posterior

Combining the prior and likelihood leads to the posterior

$$
\begin{aligned}
&\pi(\beta_\gamma, \alpha, \sigma^2 \mid y) \\
&\propto (\sigma^2)^{-n/2-1} \exp\left(-\frac{s_\gamma^2}{2\sigma^2}\right) (g\sigma^2)^{-|\gamma|/2} \det(X_\gamma^T X_\gamma)^{1/2} \\
&\quad \cdot \exp\left(-\frac{n(\widehat{\alpha}-\alpha)^2 + \left(\beta_\gamma - \widehat{\beta}_\gamma\right)^T X_\gamma^T X_\gamma \left(\beta_\gamma - \widehat{\beta}_\gamma\right) + \frac{1}{g}\left(\beta_\gamma - \beta_\gamma^0\right)^T X_\gamma^T X_\gamma \left(\beta_\gamma - \beta_\gamma^0\right)}{2\sigma^2}\right)
\end{aligned}
$$

Bayesian linear regression model: posterior

We obtain the following marginal and conditional posteriors:

- $\beta_\gamma \mid y, \sigma^2 \sim \mathcal{N}\left(\frac{g}{g+1}\widehat{\beta}_\gamma + \frac{1}{g+1}\beta_\gamma^0, \frac{g\sigma^2}{g+1}(X_\gamma^T X_\gamma)^{-1}\right)$
- $\alpha \mid y, \sigma^2 \sim N(\bar{y}, \frac{\sigma^2}{n})$
- $\sigma^{-2} \mid y \sim \mathrm{Gamma}\left(\frac{n-1}{2}, \frac{s_\gamma^2 + \frac{1}{g+1}\left(\widehat{\beta}_\gamma - \beta_\gamma^0\right)^T X_\gamma^T X_\gamma \left(\widehat{\beta}_\gamma - \beta_\gamma^0\right)}{2}\right)$

See blackboard for derivation

Bayesian variable selection

Posterior model probabilities for model selection

For Bayesian model selection, we put a prior on the set of possible models γ and compute the **posterior model probability**

$$\pi(\gamma \mid y) = \frac{\pi(\gamma) f(y \mid \gamma)}{\sum_{\gamma'} \pi(\gamma') f(y \mid \gamma')}$$

where the marginal likelihood $f(y \mid \gamma)$ is

$$f(y \mid \gamma) = \int f(y \mid \beta_\gamma, \alpha, \sigma^2) \pi(\beta_\gamma, \alpha, \sigma^2) d\beta_\gamma d\alpha d\sigma^2$$

Model selection and improper priors

- We **cannot use an improper prior for β_γ** (or any model specific parameter in general) since $f(y \mid \gamma)$ is only defined up to an arbitrary constant which does not cancel in $\pi(\gamma \mid y)$ because this constant differs for different models. This leads to **indeterminate model probabilities and Bayes factors**
- An improper prior for α and σ^2 is allowed because these two parameters are shared by all models

Posterior model probabilities for g-prior

For the g-prior*, $f(y \mid \gamma)$ can be computed in closed form:

$$f(y \mid \gamma) \propto \frac{(1+g)^{(n-1-|\gamma|)/2}}{(1+g(1-R_\gamma^2))^{(n-1)/2}}$$

where

- $R_\gamma^2 = 1 - \frac{s_\gamma^2}{s_0^2}$,
  and where $s_0^2 = (y - \bar{y}\mathbf{1})^T(y - \bar{y}\mathbf{1})$ is the sum of squared errors in the null model $\gamma = 0$
- "$\propto$" means up to factors which contain neither γ nor g

See blackboard

*For the sake of simplicity, we assume $\beta_\gamma^0 = 0$ in the following

Choice of a prior $\pi(\gamma)$ for $\gamma \in \{0, 1\}^p$

- The simplest choice is the **uniform prior** $\boldsymbol{\pi(\gamma) = 2^{-p}}$ for all γ. I.e., each explanatory variable is included with probability $\frac{1}{2}$, independently of the other. For large p, this is however **informative for the size of the model** because with high prior probability $|\gamma| \approx \frac{p}{2}$.

- A **uniform prior for** $|\gamma|$ is obtained by assuming that each explanatory variable is included with probability r where r is unknown and uniform on $(0, 1)$

- If the number of variables p is large, then computing $\pi(\gamma \mid y)$ for all γ is difficult. In such a situation, stochastic search algorithms are preferable

Bayes factor for model selection

- The posterior model probabilities $\pi(\gamma \mid y)$ depend on the prior $\pi(\gamma)$

- One can avoid this if one uses the Bayes factor which is independent of the prior:

$$
\begin{aligned}
B(\gamma, \gamma') &= \frac{\pi(\gamma \mid y)}{\pi(\gamma' \mid y)} \frac{\pi(\gamma')}{\pi(\gamma)} \\
&= \frac{f(y \mid \gamma)}{f(y \mid \gamma')} \\
&= \underbrace{(1+g)^{(|\gamma'|-|\gamma|)/2}}_{\substack{\text{"Complexity penalty"} \\ \text{(decreases with } |\gamma|)}} \underbrace{\left(\frac{1+g(1-R^2_{\gamma'})}{1+g(1-R^2_{\gamma})}\right)^{(n-1)/2}}_{\substack{\text{"Goodness of fit"} \\ \text{(increases with } R^2_\gamma)}}
\end{aligned}
$$

- The Bayes factor for comparing γ with the null model is

$$
B(\gamma, 0) = \frac{(1+g)^{(n-|\gamma|-1)/2}}{(1+g(1-R^2_\gamma))^{(n-1)/2}}
$$

Bayesian model averaging

Bayesian model averaging

- For predicting a new observation y_{n+1} for a given vector x_{n+1} of explanatory variables, **Bayesian model averaging** is an alternative to model selection.

- Bayesian model averaging works by
  1. making predictions under each model,
  2. averaging all predictions according to the posterior probability of each model.

- For instance, the prediction of the mean of y_{n+1} is (for known g) given by

$$\mathbb{E}(y_{n+1} \mid y) = \bar{y} + \frac{g}{g+1} \sum_{\gamma} x_{n+1,\gamma}^T \widehat{\beta}_\gamma \pi(\gamma \mid y).$$

Unknown g

Choosing g

- Bayes factors (and also posterior distributions) depend on the choice of g
- As ***g* tends to infinity**, the prior becomes non-informative. However, as $g \to \infty$, $B(\gamma, 0) \to 0$ for any $\gamma \neq 0$. I.e., we always choose the null model (**“Bartlett’s paradox”**)
- Choosing any **fixed value for *g*** also leads to problems: if g is fixed and $R^2_\gamma \to 1$ then $B(\gamma, 0) \to (1+g)^{(n-1-\gamma)/2}$ which is finite although one would expect that that this goes to infinity (**“information paradox”**)

Choosing g

- In order to avoid these paradoxes, we make use of hierarchical models and consider g to be unknown
- In addition, when choosing g, an often desirable frequentist property is the so called **model selection consistency**: the probability that the true model γ' is selected must converge to 1 as the number of samples n goes to infinity:

$$\pi(\gamma' \mid y) \xrightarrow{p} 1 \text{ as } n \to \infty$$

Choosing g: empirical Bayes

- In an **empirical Bayes approach**, we can determine $\widehat{g}$ either separately for each model γ or globally for all models together

- **Separately**:

$$\begin{aligned}\widehat{g} &= \arg\max\left((n-1-|\gamma|)\log(1+g)-(n-1)\log(1+g(1-R_\gamma^2))\right)\\ &= \max\left(\frac{(n-1-|\gamma|)R_\gamma^2}{|\gamma|(1-R_\gamma^2)}-1, 0\right)\end{aligned}$$

  The above ratio is the standard F-test statistics for the null hypothesis $\beta_\gamma = 0$

- **Globally**:

$$\widehat{g} = \arg\max\sum_{\gamma}\pi(\gamma)f(y \mid \gamma)$$

  This has to be computed numerically

Unknown g: empirical Bayes

- In both cases, one can show that the information paradox does not occur any more.
- The empirical Bayes approaches do have model selection consistency except if the true model is the null model

Unknown g: fully Bayesian

- In a **fully Bayesian approach**, one can both avoid the above paradoxes and have model selection consistency for all true models

- It is desirable to have a prior $\pi(g)$ such that

$$f(y \mid \gamma) \propto \int \frac{(1+g)^{(n-1-|\gamma|)/2}}{(1+g(1-R_\gamma^2))^{(n-1)/2}} \pi(g) dg$$

  can be computed easily

- In order to avoid the information paradox, it is sufficient to have

$$\int (1+g)^{(n-1-|\gamma|)/2} \pi(g) dg = \infty \quad (|\gamma| \le p)$$

Priors on g

- **Zellner-Siow prior**

$$\pi(g) \propto g^{-3/2} \exp(-n/(2g)) \qquad (\text{i.e., } g \sim IG(1/2, n/2))$$

  - Has the model selection consistency property for all true models

- **Hyper-g prior**

$$\pi(g) \propto (1+g)^{-a/2} \quad (a < 2 \leq 3)$$

  - "Only" has the model selection consistency property for all true models except the null model