MAST90125: Bayesian Statistical learning

Lecture 14: Implementing Bayesian computation for regression

Feng Liu and Guoqi Qian

THE UNIVERSITY OF MELBOURNE

What have we done so far

So far in this subject, we have seen the cases where posterior distributions are known analytically.

In the last few lectures, we have learned several computational techniques designed to produce random samples from the posterior distribution, $p(\theta|\mathbf{y})$, when the posterior is not in a form easy for sampling.

However we have not specified any major statistical models where Bayesian computing can play a significant role in data analysis and information retrieval. In this and the following lectures, we will learn how to perform Bayesian computing and inference for regression types of models.

Linear regression

- A linear regression model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

where $\mathbf{X}$ is an $n \times p$ matrix of covariate values, $\boldsymbol{\beta}$ is a coefficient vector of length p, and $\mathbf{y}, \boldsymbol{\epsilon}$ are vectors of length n containing responses and residuals respectively.

Bayesian view of linear regression

- The sample mean analysis can be viewed as a special case of regression with $\mathbf{X} = \mathbf{1}$ and $\boldsymbol{\beta} = \mu$.
- Let's go back to the Bayesian context. In Bayesian statistics, what should we do?

Bayesian view of linear regression

- The sample mean analysis can be viewed as a special case of regression with $\mathbf{X} = \mathbf{1}$ and $\boldsymbol{\beta} = \mu$.

- Let's go back to the Bayesian context. In Bayesian statistics, what should we do?

- Full probabilistic modelling. We can assume that $p(\boldsymbol{\beta}) \propto 1$ and $p(\sigma^2) \propto (\sigma^2)^{-1}$ or we can let $\tau = (\sigma^2)^{-1}$ thus $p(\tau) \propto \tau^{-1}$.

Bayesian view of linear regression

- With these priors, the joint pdf $p(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \tau)$ is,

$$
\begin{aligned}
&= \frac{\tau^{n/2}}{(2\pi)^{n/2}} e^{-\frac{\tau(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2}} \times \tau^{-1} \\
&= \frac{\tau^{n/2-1}}{(2\pi)^{n/2}} e^{-\frac{\tau(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}+\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}+\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\hat{\boldsymbol{\beta}})}{2}}
\end{aligned}
$$

Bayesian view of linear regression

- With these priors, the joint pdf $p(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \tau)$ is,

$$
\begin{aligned}
&= \frac{\tau^{n/2}}{(2\pi)^{n/2}} e^{-\frac{\tau(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2}} \times \tau^{-1} \\
&= \frac{\tau^{n/2-1}}{(2\pi)^{n/2}} e^{-\frac{\tau(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}+\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}+\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\hat{\boldsymbol{\beta}})}{2}} \\
&= \frac{\tau^{n/2-1}}{(2\pi)^{n/2}} e^{-\frac{\tau(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})}{2}} e^{-\frac{\tau(\mathbf{X}\boldsymbol{\beta}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{X}\boldsymbol{\beta}-\mathbf{X}\hat{\boldsymbol{\beta}})}{2}} e^{\tau(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{X}\boldsymbol{\beta}-\mathbf{X}\hat{\boldsymbol{\beta}})} \\
&= \frac{\tau^{n/2-1}}{(2\pi)^{n/2}} e^{-\frac{\tau(n-p)s^2}{2}} e^{-\frac{\tau(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{2}} \qquad \text{as } \mathbf{X}'(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0} \text{ by definition}
\end{aligned}
$$

$$\text{and } s^2 = \frac{(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})}{n-p}.$$

Bayesian view of linear regression

- If we extract the kernel of β from the joint distribution,

$$e^{-\frac{\tau(\beta-\hat{\beta})'(\mathbf{X}'\mathbf{X})(\beta-\hat{\beta})}{2}}$$

  we can deduce that $p(\beta|\tau,\hat{\beta})$ is normal with mean $\hat{\beta}$ and variance $\frac{(\mathbf{X}'\mathbf{X})^{-1}}{\tau}$.

- If we wish to avoid using a Gibbs sampler, we can marginalise β out of the joint distribution,

$$p(\tau,\mathbf{y},\mathbf{X}) = \frac{\tau^{\frac{n}{2}-1}e^{-\frac{\tau(n-p)s^2}{2}}}{(2\pi)^{n/2}}\int e^{-\frac{\tau(\beta-\hat{\beta})'(\mathbf{X}'\mathbf{X})(\beta-\hat{\beta})}{2}}d\beta = \frac{\tau^{\frac{n}{2}-1}e^{-\frac{\tau(n-p)s^2}{2}}}{(2\pi)^{n/2}}\left(\frac{2\pi}{\tau}\right)^{\frac{p}{2}}\det(\mathbf{X}'\mathbf{X})^{-\frac{1}{2}}$$
$$\propto \quad \tau^{(n-p)/2-1}e^{-\frac{\tau(n-p)s^2}{2}}$$

  we can deduce that $p(\tau|s^2) = \text{Ga}(\frac{n-p}{2},\frac{(n-p)s^2}{2})$.

Bayesian view of linear regression

- Then, we need to remove the dependency on τ in the posterior for $\boldsymbol{\beta}$, we can marginalise τ out of the joint distribution,

$$
\begin{aligned}
p(\boldsymbol{\beta}, \mathbf{y}, \mathbf{X}) &= \int \frac{\tau^{\frac{n}{2}-1} e^{-\frac{\tau(n-p)s^2}{2}}}{(2\pi)^{n/2}} e^{-\frac{\tau(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{2}} d\tau \\
&= \frac{\Gamma(n/2)((n-p)s^2/2 + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})/2)^{-n/2}}{(2\pi)^{n/2}} \\
&\propto \left(1 + \frac{1}{n-p} \frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{s^2}\right)^{-\frac{(n-p)+p}{2}}
\end{aligned}
$$

which indicates that $p(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, s^2)$ is multivariate t with location parameter $\hat{\boldsymbol{\beta}}$ and scale parameter $s^2(\mathbf{X}'\mathbf{X})^{-1}$.

Blocked Gibbs sampler for linear regression

- We have just shown how least squares estimates of regression coefficients $\hat{\beta}$ and error variance s^2 get involved in Bayesian analysis, including specifying the posterior distributions which allow us to draw independent samples from.

- Now let's say we decide to draw dependent samples using Gibbs sampler. A problems arises as which conditional posterior pdf's need to be sequentially updated in Gibbs sampler.

  - In a blocked Gibbs sampler for linear regression, the parameters $(\boldsymbol{\beta}, \tau)$ are partitioned into two groups: $\boldsymbol{\beta}$ and τ; and these two groups are updated one by one sequentially in Gibbs sampling steps.

Blocked Gibbs sampler for linear regression

- We have already determined the conditional posterior for $\boldsymbol{\beta}$,

$$p(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, \tau, \mathbf{X}) = \mathcal{N}(\hat{\boldsymbol{\beta}}, (\mathbf{X}'\mathbf{X})^{-1}/\tau). \tag{1}$$

- Returning to the full joint distribution,

$$p(\boldsymbol{\beta}, \tau, \mathbf{y}, \mathbf{X}) \quad = \quad \frac{\tau^{n/2}}{(2\pi)^{n/2}} e^{-\frac{\tau(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2}} \times \tau^{-1}$$

  we can see that this contains the kernel of a Gamma distribution with respect to τ such that

$$p(\tau|\boldsymbol{\beta}, \mathbf{y}, \mathbf{X}) = \mathrm{Ga}(n/2, (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/2) \tag{2}$$

- We then would run a Gibbs sampler by cycling between drawing from the conditional posteriors (1) and (2).

Computational aside

- Linear regression is an example where a single block of parameters, $\boldsymbol{\beta}$ exists.
- By updating in blocks, we need to draw sample from multivariate distributions, which rapidly becomes computationally expensive. Calculating $(\mathbf{X}'\mathbf{X})^{-1}$ is of order $\mathcal{O}(n^3)$ for instance. Wherever possible, we want to minimise computational cost.
- While later in this lecture, we will discuss unblocked samplers, at this point we will discuss some tricks you could use to reduce the cost of a blocked sampler in regression.

Computational aside

- You may be familiar with the singular value decomposition (SVD),

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}'$$

  where $\mathbf{U}$ is an $n \times r$ matrix such that $\mathbf{U}'\mathbf{U} = \mathbf{I}_r$, $\boldsymbol{\Lambda}$ is a diagonal matrix of size r with $\boldsymbol{\Lambda}_{ii} \neq 0, \forall i$, and $\mathbf{V}$ is a $p \times r$ matrix such that $\mathbf{V}'\mathbf{V} = \mathbf{I}_r$ and $r \leq p$ with equality holding if $\mathbf{X}$ is full rank $\Leftrightarrow$ $(\mathbf{X}'\mathbf{X})$ is invertible.

- Using the SVD, $(\mathbf{X}'\mathbf{X})$ is,

$$\mathbf{X}'\mathbf{X} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{U}'\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}' = \mathbf{V}\boldsymbol{\Lambda}^2\mathbf{V}'$$

  which means that $(\mathbf{X}'\mathbf{X})^{-1}$ is

$$(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{V}\boldsymbol{\Lambda}^{-2}\mathbf{V}'$$

Computational aside

- You may be familiar with the singular value decomposition (SVD),

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}'$$

where $\mathbf{U}$ is an $n \times r$ matrix such that $\mathbf{U}'\mathbf{U} = \mathbf{I}_r$, $\boldsymbol{\Lambda}$ is a diagonal matrix of size r with $\boldsymbol{\Lambda}_{ii} \neq 0, \forall i$, and $\mathbf{V}$ is a $p \times r$ matrix such that $\mathbf{V}'\mathbf{V} = \mathbf{I}_r$ and $r \leq p$ with equality holding if $\mathbf{X}$ is full rank $\Leftrightarrow$ $(\mathbf{X}'\mathbf{X})$ is invertible.

- Using the SVD, $(\mathbf{X}'\mathbf{X})$ is,

$$\mathbf{X}'\mathbf{X} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{U}'\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}' = \mathbf{V}\boldsymbol{\Lambda}^2\mathbf{V}'$$

which means that $(\mathbf{X}'\mathbf{X})^{-1}$ is

$$(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{V}\boldsymbol{\Lambda}^{-2}\mathbf{V}' \Longrightarrow \mathbf{V}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{V} = \boldsymbol{\Lambda}^{-2} \tag{3}$$

Computational aside

- Based on Eq. (3), we can see the conditional posterior of $\tilde{\boldsymbol{\beta}} = \mathbf{V}'\boldsymbol{\beta}$ is

$$p(\mathbf{V}'\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, \tau, \mathbf{X}) = \mathcal{N}(\mathbf{V}'\hat{\boldsymbol{\beta}}, \boldsymbol{\Lambda}^{-2}/\tau).$$

  with elements being independent, as the variance-covariance matrix is diagonal.
- We would then back transform using $\boldsymbol{\beta} = \mathbf{V}\tilde{\boldsymbol{\beta}}$ to obtain an estimate of $\boldsymbol{\beta}$ for a particular iteration.
- Why does this procedure consume less resources?

Computational aside

- Based on Eq. (3), we can see the conditional posterior of $\tilde{\boldsymbol{\beta}} = \mathbf{V}'\boldsymbol{\beta}$ is

$$p(\mathbf{V}'\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, \tau, \mathbf{X}) = \mathcal{N}(\mathbf{V}'\hat{\boldsymbol{\beta}}, \boldsymbol{\Lambda}^{-2}/\tau).$$

  with elements being independent, as the variance-covariance matrix is diagonal.
- We would then back transform using $\boldsymbol{\beta} = \mathbf{V}\tilde{\boldsymbol{\beta}}$ to obtain an estimate of $\boldsymbol{\beta}$ for a particular iteration.
- Why does this procedure consume less resources?
- We estimate less parameters and these parameters are independent.

Un-blocked Gibbs sampler for linear regression

- Alternatively, we use Gibbs sampler to generate samples of $\boldsymbol{\beta}$.
- We may prefer to update elements of $\boldsymbol{\beta}$ one by one (or one sub-block by one sub-block). To work out the resulting conditional posterior, let's derive the kernel for sub-block $\boldsymbol{\beta}_i$,

$$\begin{aligned} e^{-\frac{\tau(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2}} &= e^{\frac{-\tau(\mathbf{y}-\mathbf{X}_i\boldsymbol{\beta}_i-\mathbf{X}_{-i}\boldsymbol{\beta}_{-i})'(\mathbf{y}-\mathbf{X}_i\boldsymbol{\beta}_i-\mathbf{X}_{-i}\boldsymbol{\beta}_{-i})}{2}} \\ &\propto e^{\frac{-\tau(\boldsymbol{\beta}_i'(\mathbf{X}'\mathbf{X})_{ii}\boldsymbol{\beta}_i+2\boldsymbol{\beta}_i'\mathbf{X}_i'(\mathbf{y}-\mathbf{X}_{-i}\boldsymbol{\beta}_{-i}))}{2}} \\ &= e^{\frac{-\tau(\boldsymbol{\beta}_i'(\mathbf{X}'\mathbf{X})_{ii}\boldsymbol{\beta}_i+2\boldsymbol{\beta}_i'(\mathbf{X}'\mathbf{X})_{ii}(\mathbf{X}'\mathbf{X})_{ii}^{-1}\mathbf{X}_i'(\mathbf{y}-\mathbf{X}_{-i}\boldsymbol{\beta}_{-i}))}{2}} \end{aligned}$$

from which we can deduce the conditional posterior for a particular element (or sub-block) of $\boldsymbol{\beta}$ is

$$p(\boldsymbol{\beta}_i|\boldsymbol{\beta}_{-i},\mathbf{y},\tau) = \mathcal{N}\left((\mathbf{X}'\mathbf{X})_{ii}^{-1}\mathbf{X}_i'(\mathbf{y}-\mathbf{X}_{-i}\boldsymbol{\beta}_{-i}),\ \ (\mathbf{X}'\mathbf{X})_{ii}^{-1}/\tau\right).$$

Notes

- Remember what we have done on the previous slides are different methods for obtaining parameter estimates.

- As part of any analysis, we should remember to examine the model fit, which can be tackled by some of the following methods
    - Residual plots. Note different realisations of residuals are produced for different samples generated from the posterior pdf.
    - Posterior predictive checking. In the case of linear regression, this consists of generating $y^{\text{rep}} = \mathbf{X}\beta + e^{\text{rep}}$, where $e^{rep} \sim \mathcal{N}(0, \mathbf{I}/\tau)$ and β is a posterior sample, and constructing appropriate test statistics.
    - Model comparison.

Prior choices

- We have shown how linear regression analysis can be proceeded with in Bayesian context.
- It should be noted that the prior distributions for $\boldsymbol{\beta}$ and σ^2 can be chosen in multiple ways. However, all the prior distributions should meet the specific restrictions on $\boldsymbol{\beta}$ and σ^2.
  - $\boldsymbol{\beta} \in \mathbb{R}^p$. Therefore any prior that marginally for each component of $\boldsymbol{\beta}$ has range $(-\infty, \infty)$ is acceptable.
  - $\sigma^2 \in \mathbb{R}^+$. Therefore any prior for σ^2 that has range $(0, \infty)$ is acceptable.

Extensions

- We know that a normal prior and normal likelihood are conjugate, so if we choose $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_0, \sigma^2_\beta \mathbf{K})$ *a priori*, there will exist nice computational properties in estimating these parameters. Moreover, special cases of this prior correspond to several well-known extensions to linear regression.
  - If $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma^2_\beta \mathbf{K})$, the resultant model is known as a random effect regression model.
  - Often $\mathbf{K}$ is constrained to $\mathbf{I}_p$. If $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma^2_\beta \mathbf{I}_p)$, the resultant model is a regression model having independent variance components.
  - Different parts of $\boldsymbol{\beta}$ can be assigned different priors independently. For example, if we partition $\boldsymbol{\beta}$ as $(\boldsymbol{\beta}_1 \; \boldsymbol{\beta}_2)$ such that $p(\boldsymbol{\beta}_1) \propto 1, p(\boldsymbol{\beta}_2) = \mathcal{N}(\mathbf{0}, \sigma^2_\beta \mathbf{K})$, we will obtain a linear mixed model.