MAST90125: Bayesian Statistical learning

Lecture 17: Empirical Bayes and chains comparison

Feng Liu and Guoqi Qian

What have we done so far?

In the previous three lectures, we have studied regression models in Bayesian context. We did this for both cases where the response variable has

- a normal distribution conditional on the linear predictor $\mathbf{x}'\boldsymbol{\beta}$ (linear models),
- its distribution belonging to an exponential family conditional on the linear predictor $\mathbf{x}'\boldsymbol{\beta}$ (generalised linear models).

In particular, we focused on cases where Gibbs sampling was possible. In so doing, we ended up with cases in which we implicitly considered either

- Empirical Bayes
- Data augmentation.

Meanwhile, you might realize that different choices will lead to different results. So

- How to compare those chains with different initialization?

In this lecture, we will first introduce Empirical Bayes. We will then study how to compare chains (different proposed conditional distributions).

A reminder

Consider Bayes Theorem,

$$p(\boldsymbol{\theta}|y) = \frac{p(\boldsymbol{\theta}, y)}{p(y)} = \frac{p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(y)} \quad \text{where } y \text{ is the data of size } n \geq 1.$$

- Now assume that $\boldsymbol{\theta}$ is multivariate and partition the vector as $(\boldsymbol{\theta}_1, \ \boldsymbol{\theta}_2)$. With this, we can re-write Bayes theorem as follows:

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|y) = \frac{p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, y)}{p(y)} = \frac{p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, y)p(\boldsymbol{\theta}_2, y)}{p(y)} = p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, y)p(\boldsymbol{\theta}_2|y)$$

- This may look uninteresting, as it just shows that the joint posterior corresponds to the conditional posterior of $\boldsymbol{\theta}_1$ multiplied by the marginal posterior of $\boldsymbol{\theta}_2$. However, now consider the following:

A reminder

- We know that as n becomes large, $p(\boldsymbol{\theta}|y) \to \mathcal{N}(\hat{\boldsymbol{\theta}}_{MAP}, I(\hat{\boldsymbol{\theta}}_{MAP})^{-1})$, where $I(\hat{\boldsymbol{\theta}}_{MAP})_{ij} = -\frac{\partial^2 \log(p(\boldsymbol{\theta},y))}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{MAP}}$.
- Furthermore, we know that the posterior variance $\text{Var}(\boldsymbol{\theta}|y)$ will decrease with n. Now assume that $\text{Var}(\boldsymbol{\theta}_2|y) \to 0$ at a faster rate than $\text{Var}(\boldsymbol{\theta}_1|y)$.
- In this scenario, $p(\boldsymbol{\theta}_2|y)$ has a sharper peak than $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, y)$, and we can write

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|y) = p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, y)p(\boldsymbol{\theta}_2|y) \approx p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^*, y)$$

  where $\boldsymbol{\theta}_2^*$ is the mode of $p(\boldsymbol{\theta}_2|y)$.
- Here, rather than undertake a full Bayesian analysis, we replace some parameters by their certain point estimates, and we name this method *Empirical Bayes*.

Where have you used Empirical Bayes before?

- It turns out we have implicitly used Empirical Bayes in this subject before.
  - In our example of Bayesian LASSO in lecture 14, fitted in Lab 7 to the US Judge ratings data, we fixed γ while estimating:

$$\boldsymbol{\beta}, \sigma_e^2, \sigma_j^2; \quad j = 1, \ldots p.$$

  - In the description of the linear mixed model in lecture 14, we assumed $\mathbf{K}$ was known up to proportionality.
  - In Lecture 12, we determined a Gibbs sampler for estimating σ^2, μ_1, μ_2 when $\mathbf{y}$ was bivariate normal with mean $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and variance-covariance matrix $\sigma^2\mathbf{R}$, where $\mathbf{R}$ is a correlation matrix. However we did not estimate ρ at the very beginning.

Potential impact?

- What do you think is the potential impact of using an empirical Bayes, rather than a fully Bayesian approach?
  - If we reduce the dimensionality of the problem, which is what we are doing when using Empirical Bayes estimation, we would
    - artificially reduce the level of uncertainty in parameter estimation. In Bayesian inference, this would imply narrower posterior distributions.
  - To confirm this, let's re-examine some of the plots from the Bivariate normal example in Lecture 12.

Bivariate Normal Example in Lecture 13

In this example, assume $\mathbf{y}_i = (y_{1i},\ y_{2i})'$, $i = 1, \cdots, n$, are drawn from a bivariate normal distribution where each component marginally has the same variance.

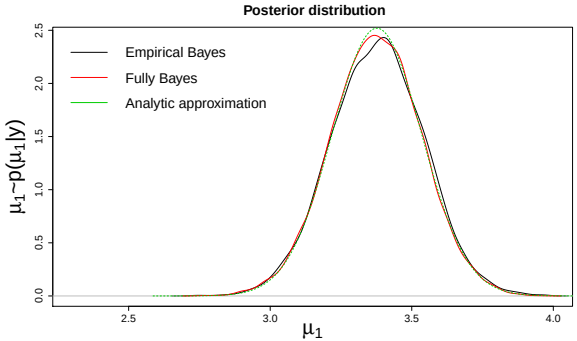
$$p(\mathbf{y}|\boldsymbol{\mu}, \sigma^2, \rho) = \frac{1}{2\pi\sigma^2\sqrt{(1-\rho^2)}} e^{-\frac{(\mathbf{y}-\boldsymbol{\mu})'\mathbf{R}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{2\sigma^2}},$$

where $\boldsymbol{\mu} = (\mu_1,\ \mu_2)'$, $\mathbf{R} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \Rightarrow \mathbf{R}^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$.

In the following, we first assume that the parameters requiring estimating are μ_1, μ_2, σ^2, and ρ is given. We then consider the situation where ρ also needs to be estimated. We use Gibbs (possibly plus MH) sampling to perform Bayesian analysis. Note if ρ is known, Bayesian analysis can be performed analytically. We will consider a flat prior for μ_1, μ_2 and assume $p(\sigma^2) \propto (\sigma^2)^{-1}$.

Example 1: μ_1

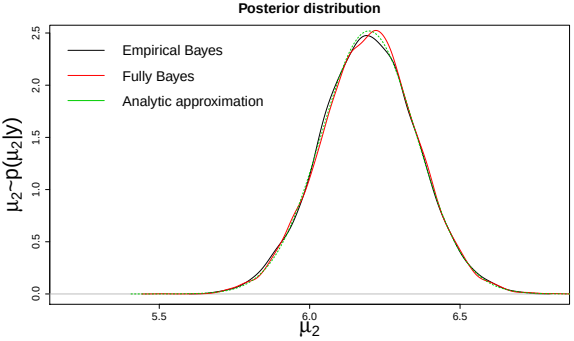
- In the following slides, we display the estimated posterior pdf's of μ_1, μ_2 and σ^2 obtained by
  - using Gibbs sampling with ρ fixed,
  - using Gibbs sampling with a Metropolis step to sample from the posterior of ρ.



- For μ_1, there appears to be little difference between empirical Bayes (ρ fixed) and full Bayesian (ρ estimated) estimates of the marginal posterior.

Example 1: μ_2

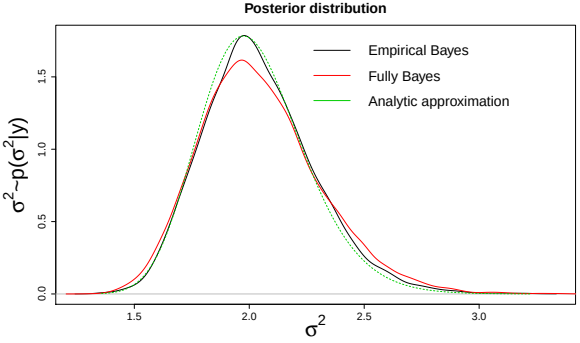
- In the following slides, we display the estimated posterior pdf's of μ_1, μ_2 and σ^2 obtained by
  - using Gibbs sampling with ρ fixed,
  - using Gibbs sampling with a Metropolis step to sample from the posterior of ρ.



- For μ_2, again there appears to be little difference between empirical Bayes (ρ fixed) and full Bayesian (ρ estimated) estimates of the marginal posterior.

Example 1: σ^2

- In the following slides, we display the estimated posterior pdf's of μ_1, μ_2 and σ^2 obtained by
  - using Gibbs sampling with ρ fixed,
  - using Gibbs sampling with a Metropolis step to sample from the posterior of ρ.



- For σ^2, fixing ρ has a noticeable impact on the posterior inference. The estimated posterior from the full Bayesian approach is less peaked with wider tails than the empirical Bayes counterpart.

Comparing chains

- We have considered a binomial likelihood with a Beta prior before. We have seen that different proposal distributions all lead to chains that converged to $p(\theta|y)$.

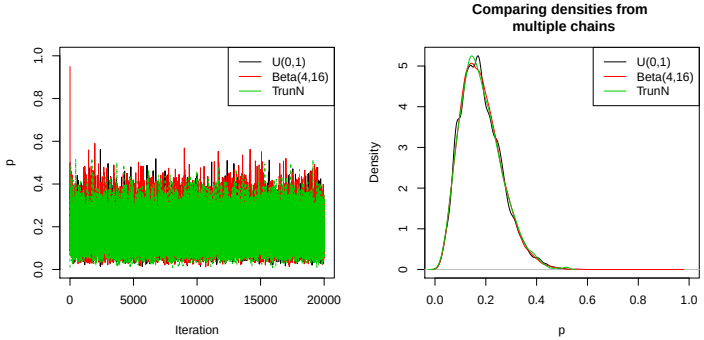


Figure: Comparison the chains using good $J(\theta^*|\cdot)$'s.

- Despite using different proposal distributions (resulting in different transitional pdf's $J(\theta^*|\cdot)$), all chains converge to $p(\theta|y)$. Cf. the Figure. Note the left plot is more commonly used.

Example 2: Poisson likelihood, Gamma prior

- Consider an example where we know the exact posterior. Assume we have count data for which a plausible likelihood is Poisson,

$$p(y_1, \dots y_n|\lambda) = \prod_{i=1}^{n} p(y_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} = \frac{\lambda^{n\bar{y}} e^{-n\lambda}}{\prod_{i=1}^{n} y_i!}.$$

If we assume a Gamma prior for λ

$$p(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda},$$

we can show the posterior is Gamma $(\alpha + n\bar{y},\ \beta + n)$ by finding the kernel of the joint distribution,

$$p(y_1, \dots y_n, \lambda) = p(y_1, \dots y_n|\lambda)p(\lambda) \propto \lambda^{n\bar{y}} e^{-n\lambda} \times \lambda^{\alpha-1} e^{-\beta\lambda}$$

Example 2: Poisson likelihood, Gamma prior

- Now pretend you cannot sample directly from the exact posterior, and you decide to implement a Metropolis Hastings algorithm to get around this, which requires us to determine a proposal distribution.
- You decide that a good proposal distribution is

$$J(\lambda^*|\lambda^{(t-1)}) = \log \mathcal{N}(\hat{\mu}, c^2\hat{\sigma}^2/n),$$

where $\hat{\mu}$ is the sample mean of $\log(y_i + 0.1); i = 1, \ldots, n$, $\hat{\sigma^2}$ is the sample variance of $\log(y_i + 0.1); i = 1, \ldots, n$, and c is some constant.

  - Though by default the proposal distribution of λ^* depends on $\lambda^{(t-1)}$, it does not have to, as in this example.

- On the next slide, you will find the effect of using different proposal distributions.

Example 2: Poisson likelihood, Gamma prior

- Effect of poorly chosen proposal distributions.

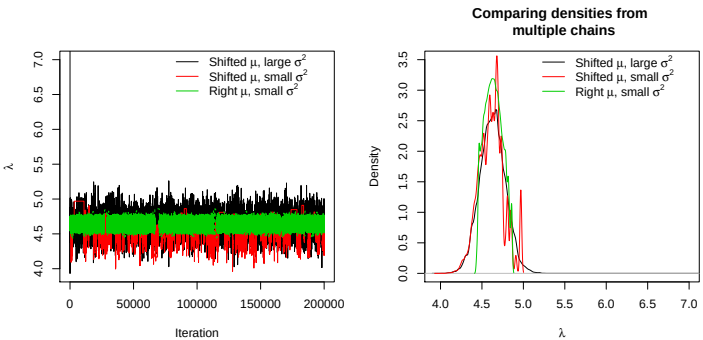


Figure: Comparing chains with poorly chosen $J(\theta^*|\cdot)$.

- Iteration plots of the three chains do not completely overlap, unlike the previous Binomial+Beta case.
- Determining the best proposal distribution is often very difficult. Instead, people generate multiple chains using the same proposal distribution but with different initial values.

A numerical criteria for convergence

- Would you trust visual inspection that chains have converged and/or mixed?
  - Would your answer change with the dimension of $\boldsymbol{\theta}$?

- To provide a numerical measure of convergence/mixing, Gelman and Rubin (1992) proposed the following diagnostic method, where for each element θ_j of the parameter vector $\boldsymbol{\theta}$ (notation $\boldsymbol{\theta}_{ijk}$ is used here, representing the j-th element in $\boldsymbol{\theta}$ in the k-th chain and i-th iteration), calculate

$$\hat{R}_j = \frac{(n-1)W_j/n + B_j/n}{W_j}$$

where $W_j = \frac{\sum_{k=1}^{m} s_{jk}^2}{m}$, $s_{jk}^2 = \frac{\sum_{i=1}^{n}(\psi_{ijk}-\bar{\psi}_{jk})^2}{n-1}$, $B_j = \frac{n\sum_{k=1}^{m}(\bar{\psi}_{jk}-\bar{\psi}_j)^2}{m-1}$, m is the number of chains, n is the number of iterations per chain, and $\psi_{ijk} = f(\boldsymbol{\theta}_{ijk})$.

- But what do we learn from this.

A numerical criteria for convergence

- In the formula for $\hat{R}_j$, s_{jk}^2 is the estimated posterior variance of parameter ψ_j in chain k, meaning W_j is the average within-chain posterior variance. B_j is the estimated variance between chain posterior chain. When the chains have converged to the posterior distribution, we would expect ...
  - The between and within chain variances to be the same $\Rightarrow \hat{R}_j \to 1$.
- If this is not the case, for instance if initial values were over-dispersed, we would expect $B_j > W_j \Rightarrow \hat{R}_j > 1$.
  - Note: Using multiple chains slit from one long chain may mitigate the over-dispersion effect.
  - The formula for $\hat{R}_j$ is motivated by ANOVA like problems. One ANOVA assumption is normally distributed errors. Hence the function f in $\psi_{ijk} = f(\boldsymbol{\theta}_{ijk})$, is often chosen to transform a highly non-normal parameter to something approximately normal.

Example of convergence diagnostic

- For p in the binomial likelihood, we use the transformation $\text{logit}(p)$. For λ in the Poisson likelihood, we use the transformation $\log(\lambda)$. We will discard the first 5 % of iterations from analysis, split the remaining chains exactly in half and perform no thinning.

Table: Gelman-Rubin diagnostic for two examples

Model	Parameter	$\hat{R}$	Upper limit C.I.
Binomial likelihood, Beta prior	p	1.0001	1.0002
(Well chosen $J(p\|\cdot)$, 20000 iterations)	$\text{logit}(p)$	1.0015	1.0019
Poisson likelihood, Gamma prior	λ	1.0203	1.0412
(Poorly chosen $J(\lambda\|\cdot)$, 200000 iterations)	$\log(\lambda)$	1.0206	1.0418

- The Gelman-Rubin diagnostics suggests that the chains of p converge to $p(p|\mathbf{y})$ more quickly than the chains of λ converge to $p(\lambda|\mathbf{y})$.

Effective sample size

- Because of auto-correlation, the number of iterations is not a measure of the true sample size. The effective sample size is defined as

$$n_{\text{eff}} = \frac{mn}{1 + 2\sum_{t=1}^{\infty} \rho_t},$$

  where m is the number of chains, n is the number of iterations per chain, and ρ_t is the auto-correlation at lag t. In order to estimate n_{eff}, the infinite sum is truncated at T, where T is the first positive integer such that $\hat{\rho}_{T+1} + \hat{\rho}_{T+2} \leq 0$.
- This means that the estimated effective sample size, $\hat{n}_{\text{eff}}$ is:

$$\hat{n}_{\text{eff}} = \frac{mn}{1 + 2\sum_{t=1}^{T} \hat{\rho}_t}.$$

Example Effective sample size

- Let's now calculate the effective sample size for each chain of the two examples.

Table: Effective sample size for two examples

Model	Parameter	$\hat{n}_{\text{eff}}$		
		Chain 1	Chain 2	Chain 3
Binomial likelihood, Beta prior	p	5408.63	5637.29	12599.14
(Well chosen $J(p\|\cdot)$, 20000 iterations)	$\text{logit}(p)$	6055.75	6774.92	11232.21
Poisson likelihood, Gamma prior	λ	1441.92	3405.03	431.44
(Poorly chosen $J(\lambda\|\cdot)$, 200000 iterations)	$\log(\lambda)$	1443.56	3424.521	438.76

Note for Binomial likelihood, $J(|)$ was $U(0,1)$ (chain 1), $\text{Be}(4,16)$ (chain 2), and truncated normal (chain 3). For Poisson likelihood, $J(|)$ was lognormal with shifted mean and large variance (chain 1), shifted mean and small variance (chain 2), and correct mean and small variance (chain 3).

- $\hat{n}_{\text{eff}}$ is lower than the number of iterations, especially in the Poisson case.
- Note the effective sample size and Gelman-Rubin diagnostics were calculated using the functions `effectiveSize` and `gelman.diag` from the `R` package `coda`.