STAT 432: Basics of Statistical Learning

Bayesian Regularized Linear Regression I

Shiwei Lan, Ph.D. <shiwei@illinois.edu>

http://shiwei.stat.illinois.edu/stat432.html

March 1, 2019

University of Illinois at Urbana-Champaign

Motivation

- Frequentist penalized linear regression
  - Based on optimization;
  - No natural way to quantify uncertainty of estimates.
- Bayesian regularized linear regression
  - Based on posterior estimate/inference;
  - Comes with uncertainty quantification;
  - Penalization can be naturally viewed as priors.

Outline

- Brief introduction to Bayesian Statistics.
- Bayesian regularized linear regressions with conjugate priors.
- Brief introduction to MCMC.
  - Metropolis-Hastings Algorithm.
  - Hamiltonian Monte Carlo*.
  - Spherical Hamiltonian Monte Carlo*.
- Bayesian regularized linear regressions with general priors*.

Intro to Bayesian Statistics

The role of statistics in science

- Statistics methods are mainly inspired by applied scientific problems.
- The overall goal of statistical analysis is to provide a robust framework for
  - designing scientific studies,
  - collecting empirical evidence, and
  - analyzing the data,
- in order to
  - understand unknown phenomena,
  - answer scientific questions, and
  - make decisions.
- To this end, we rely on the observed data as well as our *domain knowledge*.

Domain knowledge

- Our domain knowledge, which we refer to as our *prior* information, is itself based on previous empirical evidence.
- For example, if we are interested in the average normal body temperature, we would of course measure body temperature of samples from the population.
- But we also know, based on previous empirical evidence, that this average is a number close to $98.6\,^{\circ}\mathrm{F}$.
- In this case, our prior knowledge asserts that values around 98 are more plausible compared to values around 90 for example.

Prior or no prior: this is a question.

- Most frequentist methods attempt to minimize our reliance on prior information.
- They use the domain knowledge for example to decide which characteristics of the population are relevant to our scientific problem (e.g., we might not include height as a risk factor for cancer), but avoid using priors when making inference.
- Bayesian methods on the other hand provide a mathematical framework to incorporate prior knowledge in the process of making inference.
- This is based on the philosophy that if the prior is in fact informative, this should lead to more accurate inference and better decision.

Likelihood-based inference

- Define the underlying mechanism that generates data, y, using a probability model, $P(y|\theta)$, which depends on the unknown parameter of interest, θ.
- The *likelihood function* is defined by plugging-in the observed data $\mathbf{y}$ in the probability distribution and expressing it as a function of model parameters, i.e., $f(\theta, \mathbf{y})$.
- Frequentist methods estimate model parameters by maximizing the likelihood function (MLE).
  - Under weak regularity conditions, the MLE demonstrates attractive properties as the sample size $n \to \infty$, including asymptotic normality, consistency, and efficiency.
  - We can also use the likelihood function to device standard tests (Wald test, score test, and likelihood ratio test) to perform hypothesis testing.

Issue? Violation of the strong likelihood principle

- Strong likelihood principle: Denote the observed sample from a random variable, X, with $p_1(x|\theta)$ as $\mathbf{x}$, and the observed sample from another random variable, Y, with $p_2(y|\theta)$ as $\mathbf{y}$. If the corresponding likelihood functions are proportional, $f_1(\theta, \mathbf{x}) \propto f_2(\theta, \mathbf{y})$, then inference for θ should be the same whether we observe $\mathbf{x}$ or $\mathbf{y}$.
- The following example is discussed by David MacKay.
- A scientist has just received a grant to examine whether a specific coin is fair (i.e., $P(H) = P(T) = 0.5$) or not.
- He sets up a lab to toss the coin. Of course, due to his limited budget, he can only toss the coin a finite number of times.
- He hires a frequentist statistician to estimate the p-value hoping that the result could be published in one of the journals that only publish if p-value is less than 0.05!

Violation of the strong likelihood principle

- The scientist tosses the coin 12 times, of which only 3 are heads.
- The statistician says: “you tossed the coin 12 times and you got 3 heads. The one-sided p-value is 0.07”.
- The scientist says: “Well, it wasn’t exactly like that... I actually repeated the coin tossing experiment until I got 3 heads and then I stopped”.
- The statistician say: “In that case, your p-value is 0.03”.
- What’s wrong?

Violation of the strong likelihood principle

- The scientist tosses the coin 12 times, of which only 3 are heads.
- The statistician says: “you tossed the coin 12 times and you got 3 heads. The one-sided p-value is 0.07”.
- The scientist says: “Well, it wasn’t exactly like that... I actually repeated the coin tossing experiment until I got 3 heads and then I stopped”.
- The statistician say: “In that case, your p-value is 0.03”.
- What’s wrong?
- Note that in the first scenario, we use a binomial model, and in the second scenario, we use a negative-binomial model with the following likelihood functions respectively,

$$\begin{aligned} f_1(\theta, x) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \\ f_2(\theta, x) &= \binom{n-1}{x-1}\theta^x(1-\theta)^{n-x} \end{aligned}$$

Bayesian inference

- Bayesian inference is making statements about unknown quantities in terms of probabilities given the observed data and our prior knowledge.
- Our *prior* knowledge represents the extent of our belief and uncertainty regarding the value of unobservables. We express our prior using probability models.
- We also use probability models to define the underlying mechanism that has generated the data.
- To make inference, we update our prior opinion about unobservables given the observed data. We refer to this updated opinion as our *posterior* opinion, which itself is expressed in terms of probabilities.

Probability for Bayesian: It’s subjective!

- In the Bayesian paradigm, probability is a measure of uncertainty.
    - “Coins don’t have probabilities, people have probabilities”, Persi Diaconis.
    - “The only relevant thing is uncertainty–the extent of our own knowledge and ignorance,” Bruno deFinetti.
- In this view, all that matters is uncertainty, and all uncertainties are expressed in terms of probability.
- Therefore, we use probability models for:
    - random variables that change and
    - those that might not change (e.g., the population mean) but we are uncertain about their value .

Probability for frequentist: It’s objective!

- Consider the well-known coin tossing example. What is the probability of head in one toss?
- In the frequentist view, probability is assigned to an event by regarding it as a class of individual events (i.e., trials) all equally probable and stochastically independent.
- For the coin tossing example, we assume a sequence of *iid* tosses, and the probability of head is 1/2 since the number of times we observe head divided by the number of trials reaches 1/2 as the number of trials grows.
- But... “what is the probability that the price of this car in the picture is less than $5,000?”.

Prior probability

- Bayesians feel comfortable to assign probabilities to events that are not repeatable.
- We use probability not only for the data, y, where our certainty is due to data variation, but also for model parameters, θ, where our uncertainty is due to the fact that θ is the population parameter, and it is almost always unknown.
- Therefore, before doing statistical inference, we need to specify the extent of our belief and our uncertainty about the possible values of the parameter using prior probability $P(\theta)$.
- We usually use our (or other's) domain knowledge, which is accumulated based on previous scientific studies.
- We almost always have such information, although it could be vague.

Prior probability: body temperature

- Let's denote the average normal body temperature for the population as θ. We know that θ should be close to $98.6\,^{\circ}\mathrm{F}$; that is, values close to $98.6\,^{\circ}\mathrm{F}$ are more plausible than values close to $90\,^{\circ}\mathrm{F}$ for example.
- We assume that as we move away from the $98.6\,^{\circ}\mathrm{F}$ the values become less likely in a symmetric way (i.e., it does not matter if we go higher or lower). We can set $\theta \sim N(98.6, \tau^2)$.
- In the above prior, τ^2 determines how certain we are about the average normal body temperature being around $98.6\,^{\circ}\mathrm{F}$.
- If we believe that it is almost impossible that the average normal body temperature is above 113.6 and below 83.6, we can set $\tau = 5$ so the approximate 99.7% interval includes all the plausible values from 83.6 to 113.6.
- A general advice is that we should keep an open mind, consider all possibilities, and avoid using very restrictive priors.

Bayes' theorem

- For two events, A and B, *Bayes' theorem* can be simply presented as follows:

$$P(B|A) \quad = \quad \frac{P(B)P(A|B)}{P(A)}$$

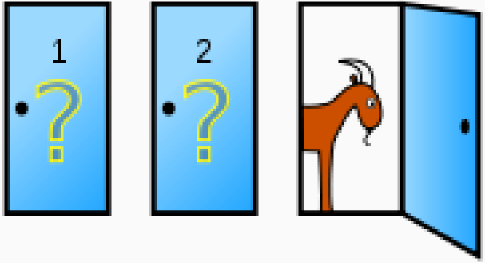
- Also, recall that if $B = (B_1, B_2, ..., B_n)$ are a set of events that partition the sample space, Ω, using the law of total probability, we have:

$$\begin{aligned} P(A) &= P(A|B_1)P(B_1) + ... + P(A|B_n)P(B_n) \\ P(B_i|A) &= \frac{P(B_i)P(A|B_i)}{\sum_{i'}^{n} P(B_{i'})P(A|B_{i'})} \end{aligned}$$

- This simple formula which is known as *Bayes' theorem* is the basis of Bayesian analysis. (However, using this theorem does not automatically make you a Bayesian!)

Monty Hall problem

- The problem is based on a TV game show hosted by Monty Hall. In this show, contestants have the chance to win cars. Before each show, a car is put behind one of three closed doors. The other two doors have goats behind them. The contestant is asked to choose one of the three doors. Then, Monty Hall opens one of the other two doors, which he knows does not have a car behind it. After that, the contestant can choose to switch to the remaining unopened door, or stay with his original selection.
- The question is: should he switch?



Monty Hall problem

- At the beginning, the car can be behind any of the three doors with equal probability. That is,

$$P(C_1) = P(C_2) = P(C_3) = 1/3$$

- Let's say we choose door number 1, and Monty open door number 3.
- Now let's write down the conditional probability of opening, O_3, given the three possibilities (i.e., C_1, C_2, and C_3):

$$\begin{aligned} P(O_3|C_1) &= 1/2 \\ P(O_3|C_2) &= 1 \\ P(O_3|C_3) &= 0 \end{aligned}$$

- Now using the law of total probability we can find the marginal probability for opening door number 3:

$$P(O_3) = 1/3 \times 1/2 + 1/3 \times 1 + 1/3 \times 0 = 1/2$$

Monty Hall problem

- Using Bayes' theorem, we have:

$$\begin{aligned}
P(C_1|O_3) &= \frac{P(C_1)P(O_3|C_1)}{P(O_3)} \\
&= \frac{1/3 \times 1/2}{1/2} = 1/3 \\
P(C_2|O_3) &= \frac{P(C_2)P(O_3|C_2)}{P(O_3)} \\
&= \frac{1/3 \times 1}{1/2} = 2/3
\end{aligned}$$

- Therefore, probability of winning doubles if we switch.
- You can try this using a penny and three cups.

Bayesian inference

- Assume exchangeability for observations, we have
  - The conditional distribution of y given θ is

$$P(y|\theta) = P(y_1, y_2, ..., y_n|\theta) = \prod_{i=1}^{n} P(y_i|\theta)$$

  - There exists a *prior* probability distribution $P(\theta)$ over the parameters of the model such that we can find the unconditional (or marginal) joint distribution of observations as follows:

$$P(y) = P(y_1, y_2, ..., y_n) = \int_{\Omega} \prod_{i=1}^{n} P(y_i|\theta) P(\theta) d\theta$$

- Therefore, we first need to specify the model $P(y|\theta)$ for the observed data, and the prior $P(\theta)$ for the parameter of the model.
- The next step in Bayesian inference is to find *posterior distribution*, $P(\theta|y)$.

Bayesian inference

- Bayes' theorem provides a mathematical formula for obtaining $P(\theta|y)$ based on $P(\theta)$ and $P(y|\theta)$:

$$P(\theta|y) \quad = \quad \frac{P(\theta)P(y|\theta)}{P(y)}$$

- Since $P(y)$ does not depend on θ, we can use the following unnormalized form of the posterior distribution:

$$P(\theta|y) \quad \propto \quad P(\theta)P(y|\theta)$$

- This simple formula is the essential part of Bayesian analysis.
- It is used not only for expressing our updated belief about model parameters (i.e., θ), but also for making decisions (e.g., accepting or rejecting a hypothesis) and predicting unknown observable (e.g., for future cases).

Posterior predictive distribution

- Sometimes our objective is to use the posterior distribution to predict future observations.
- That is, after observing some data, $y = (y_1, y_2, ..., y_n)$, we want to predict the next observation, y_{n+1}, which we denote as $\tilde{y}$.
- Note that we still have uncertainty regarding our prediction, and therefore we express such prediction in the form of probability distribution, called *posterior predictive distribution*, which is obtained by summing (or integrating) over posterior distribution of θ, i.e., $P(\theta|y)$:

$$P(\tilde{y}|y) \quad = \quad \int_\theta P(\tilde{y}|\theta, y)P(\theta|y)d\theta$$

Since $\tilde{y}$ is independent of y given θ, we have

$$P(\tilde{y}|y) \quad = \quad \int_\theta P(\tilde{y}|\theta)P(\theta|y)d\theta$$

Bayesian Models with Conjugate Priors

Parametric models

- Next, we will discuss some simple models commonly used for typical random variables.
- The focus in this model is on one single parameter, which represents the population mean.
- If there are other parameters in the model, we would regard them as nuisance parameters.
- Based on the likelihood-prior pair, we consider
    - Binomial – Beta
    - Poisson – Gamma
    - Normal – Normal

Binomial model

- Consider a sequence of n independent Bernoulli trials with success (e.g. getting 'head' in tossing coins) probability θ.
- Then the number of success y has a Binomial(n, θ) distribution:

$$P(y|n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

- Assuming the prior $P(\theta)$, the marginal distribution of y can be obtained as

$$P(y) = \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} P(\theta) d\theta$$

- We therefore obtain the posterior distribution as follows:

$$P(\theta|y) \quad = \quad \frac{\binom{n}{y} \theta^y (1-\theta)^{n-y} P(\theta)}{\int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} P(\theta) d\theta}$$

Binomial model

- Let's say we are quite ignorant about the possible value of θ. That is to say, we think θ is uniformly distributed in [0, 1], i.e., $P(\theta) = 1, 0 \leq \theta \leq 1$.
- Then we have

$$P(y) \quad = \quad \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} d\theta = \frac{1}{n+1}$$

- The posterior distribution simplifies to

$$P(\theta|y) \quad = \quad \frac{(n+1)!}{y!(n-y)!} \theta^y (1-\theta)^{n-y}$$

- This is a Beta($y+1, n-y+1$) distribution with expectation

$$E(\theta|y) = \frac{y+1}{n+2}$$

Prediction in binomial model

- For the above model, since $P(\tilde{y} = 1|\theta) = \theta$, the posterior predictive distribution can be obtained as

$$P(\tilde{y} = 1|y) \quad = \quad \int_0^1 \theta P(\theta|y) d\theta$$

- Recall that the posterior distribution of $\theta|y$ is a Beta($y + 1, n - y + 1$) distribution.
- Therefore, $\int_0^1 \theta P(\theta|y) d\theta$ is the posterior expectation of θ:

$$P(\tilde{y} = 1|y) = \frac{y + 1}{n + 2}$$

- Exercise: Use the above results and find the probability that sun rises tomorrow.

Predicting the election result

- We want to predict which one of two candidates, A or B, will win the election.
- Let's denote the probability that A wins as θ, and we assume *a priori* the probability of winning for candidate A has a uniform distribution.
- We ask 10 people which candidate they would choose in this election. Of 10 people surveyed, 3 people said they are going to vote for A.
- Our updated belief in A's winning has now a Beta(4, 8) distribution.
- The posterior expectation of A's winning is $\frac{4}{12} = 0.33$, which is also the probability that the next person we survey votes for A.
- Note that this is almost the same as the maximum likelihood estimation $\frac{3}{10} = 0.3$ (the similarity is superficial however since the underlying philosophies are different).

Conjugate priors

- In the above example, the derivation of posterior distribution was quite simple since it had a closed form.
- This was due to our choice of prior, i.e., uniform distribution.
- Note that uniform prior on $[0, 1]$ is in fact Beta(1, 1) distribution.
- Therefore, for the above binomial model, both prior and posterior are Beta distributions.
- This is called “conjugacy” and the prior is called a “conjugate” prior.
- Conjugacy is informally defined as a situation where the prior distribution $P(\theta)$ and the corresponding posterior distribution, $P(\theta|y)$ belong to the same distributional family.
- Using conjugate priors makes sampling and Bayesian inference much easier compared to using non-conjugate priors.

Exponential family

- A large class of distributions, called *exponential family*, have the following form:

$$P(y_i|\theta) = h(y_i)g(\theta)\exp(\phi(\theta)^T s(y_i))$$

- Many widely used distributions such as normal, bernoulli, and Poisson belong to the exponential family.
- $\phi(\theta)$ is called the “natural parameter” of the family.
- The joint distribution for a set of conditionally (given θ) independent observations, $y = (y_1, y_2, ..., y_n)$ is

$$P(y|\theta) = \Big[\prod_i h(y_i)\Big] g(\theta)^n \exp(\phi(\theta)^T \sum_i s(y_i))$$

- $t(y) = \sum_i s(y_i)$ is a *sufficient statistic* for θ.

Conjugate priors (formal definition)

- If for an exponential family we define our prior as

$$P(\theta) \propto g(\theta)^{\eta} \exp(\phi(\theta)^T \nu)$$

  then the posterior would have a similar form:

$$P(\theta|y) \propto g(\theta)^{\eta+n} \exp(\phi(\theta)^T (\nu + t(y)))$$

- In this case, $P(\theta)$ is a conjugate prior.
- For example, in the binomial distribution

$$\begin{aligned} g(\theta) &= (1-\theta) \\ \phi(\theta) &= \log(\frac{\theta}{1-\theta}) \end{aligned}$$

Binomial model

- Recall that a conjugate prior is proportional to

$$P(\theta) \propto g(\theta)^{\eta} \exp(\phi(\theta)^T \nu)$$

- Therefore, the conjugate prior for the Binomial model has the following form:

$$P(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- This is a Beta(α, β) distribution.
- We can interpret this prior as observing $\alpha - 1$ prior success and $\beta - 1$ prior failure. That is, the prior acts as additional data.
- The posterior distribution is also Beta, with parameters $\alpha + y$ and $\beta + n - y$, i.e., Beta$(\alpha + y, \beta + n - y)$
- Note that the uniform distribution we previously used is in fact a special cases of Beta distribution where $\alpha = \beta = 1$.

The election example

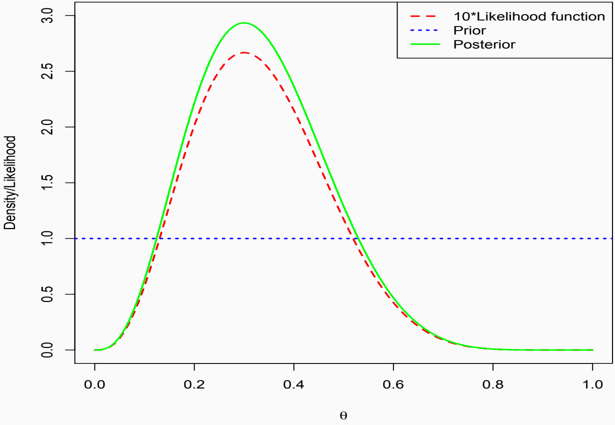


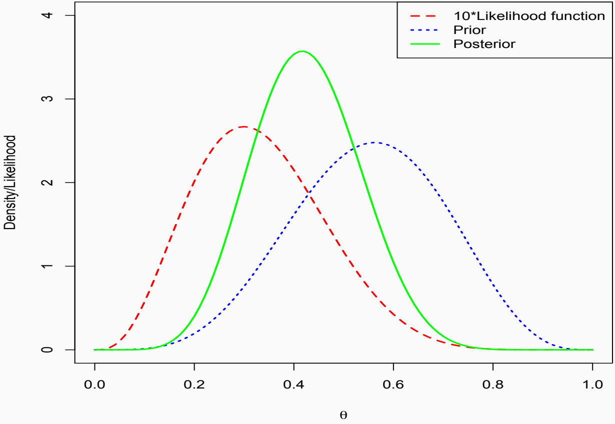
Figure 1: The likelihood function, the uniform prior, and the posterior distribution for the election example.

The election example with informative prior

- As before, assume that we have surveyed 10 people and 3 of them are going to vote for candidate A.
- This time, however, we know that candidate A belongs to the party that in the previous elections won about 55% of votes.
- Instead of a uniform prior, we could use a more informative Beta prior which reflects such prior information.
- For example, we could choose a Beta prior whose mean is $\frac{\alpha}{\alpha+\beta} = 0.55$, and it is broad enough to reflect the extent of our uncertainty. We choose a $P(\theta) = \mathrm{Beta}(5.5, 4.5)$ as our prior.
- We should always use a reasonably broad prior. As Savage (1954) said: “Keep the mind open, or at least ajar”.

The election example with informative prior

Figure 2: The posterior distribution of $\theta|y$ is $\mathrm{Beta}(8.5, 11.5)$. So while the MLE is 0.3, the posterior expectation is 0.425, which is a compromise between the observed data and the prior.



The election example when more data are observed

- Now assume that we have obtained additional budget to survey 20 more people. The result shows that 12 out 20 are going to vote for candidate A.
- It makes sense to update our opinion based on this new information. It is also reasonable not to ignore the previous data.
- However, we do not need to start our analysis from the beginning. We can use the previous posterior distribution, $P(\theta) = \mathrm{Beta}(8.5, 11.5)$, as our new prior and obtain a new posterior based on the more recent data.
- Our new posterior is therefore $P(\theta|y) = \mathrm{Beta}(20.5, 19.5)$.
- The posterior expectation ($\frac{20.5}{20.5+19.5} = 0.51$) and the MLE ($15/30 = 0.5$) are now getting closer as the amount of data increases.

The election example with informative prior

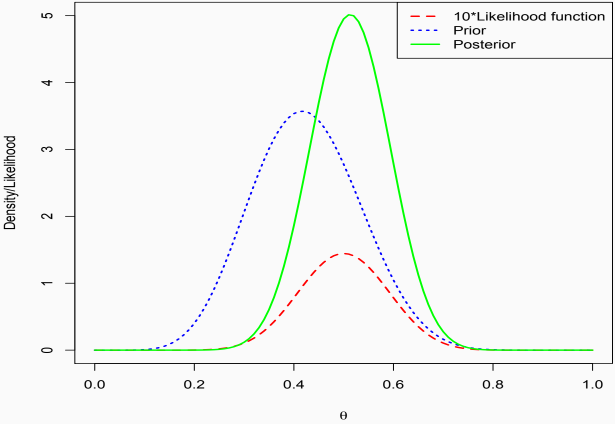


Figure 3: The likelihood function, the Beta prior (based on our previous posterior distribution), and the new posterior distribution for the election example with more data.

Poisson model

- Poisson model is another member of exponential family and is commonly used for count data.
- Assume we have observed $y = (y_1, y_2, ..., y_n)$:

$$\begin{aligned} P(y|\theta) &= \prod_i \frac{\theta^{y_i} \exp(-\theta)}{y_i!} \\ &\propto \exp(-n\theta) \exp(\log(\theta) \sum y_i) \end{aligned}$$

- The conjugate prior would have the following form:

$$\begin{aligned} P(\theta) &\propto (\exp(-\theta))^\eta \exp(\nu \log(\theta)) \\ &\propto \exp(-\eta\theta)\theta^\nu \end{aligned}$$

- Using $P(\theta) \propto \exp(-\beta\theta)\theta^{\alpha-1}$, which is a Gamma($\alpha, \beta$) distribution, as our prior, we obtain the following posterior:

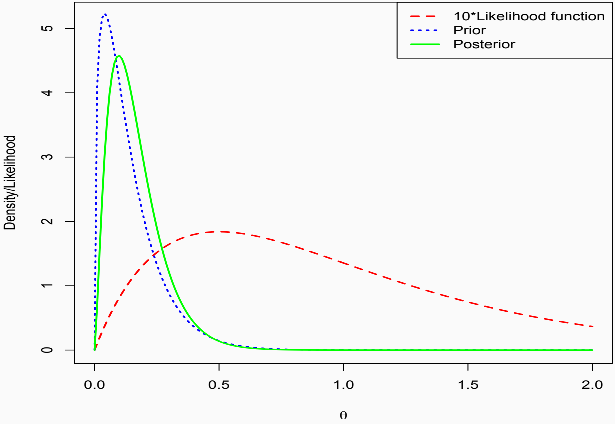
$$\theta|y \sim \text{Gamma}(\alpha + \sum_{i=1}^{n} y_i, \beta + n)$$

MLS example

- When David Beckham joined LA Galaxy, he scored one goal in his first two MLS games.
- After that the manager of LA Galaxy wanted to predict the number of goals Beckham would score in the remaining games.
- We model the number of goals, y_i, he scores in a game using a Poisson model with parameter θ.
- The maximum likelihood is $\hat{\theta} = 0.5$.
- Now let's use a Gamma(α, β) prior for θ.
- Alternatively, we might want to use Beckham's history in Real Madrid to build a prior opinion.
- When in Madrid, Beckham scored 3 goals in 22 games (i.e., $3/22 = 0.14$ on average) during 06-07 season.
- For our example, we could use the conjugate Gamma(1.4, 10) prior with mean $1.4/10 = 0.14$.

MLS example

- Since Gamma is a conjugate prior for the parameter of poisson model, the posterior also has a Gamma distribution, which in this case is a Gamma(1.4+1, 10+2) distribution.
- The expected number of goals is therefore $2.4/12 = 0.2$



MLS example

- Posterior is again a compromise between the prior and the data (likelihood).
- In this example, as shown in the graph, the posterior is more similar to the prior than the likelihood.
- This is due to the fact that the amount of data is small.
- As the amount of data increases the influence of prior on posterior decreases while the effect of likelihood increases.
- In 2008-2009, Beckham played 25 games and scored 5 goals. This is a 0.2 average, which is much closer to our estimate 0.19 compared to the MLE, which is 0.5

Univariate normal model

- The normal distribution is also a member of exponential families.
- We first consider a situation where there is only one observation and the variance is known.

$$\begin{aligned} y &\sim N(\theta, \sigma^2) \\ P(y|\theta,\sigma) &= \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{\theta^2}{2\sigma^2})\exp(-\frac{y^2}{2\sigma^2}+\frac{\theta y}{\sigma^2}) \end{aligned}$$

- So the general form of a conjugate prior is

$$P(\theta) \propto \exp(a\theta^2 + b\theta)$$

which can be parameterized as

$$P(\theta) \propto \exp(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2)$$

which is a $N(\mu_0, \tau_0^2)$ distribution.

Univariate normal model

- As the result, the posterior distribution would be

$$P(\theta|\sigma, y) \propto \exp(-\frac{1}{2\sigma^2}(y-\theta)^2 - \frac{1}{2\tau_0^2}(\theta-\mu_0)^2)$$

- When you complete the square, the posterior would also become a normal distribution:

$$P(\theta|\sigma, y) \propto \exp(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2)$$

  which is a $N(\mu_1, \tau_1^2)$ distribution with

$$\mu_1 = \frac{\frac{\mu_0}{\tau_0^2} + \frac{y}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}} \qquad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

- For n observations, we write the model for $\bar{y}$; therefore,

$$\mu_n = \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \qquad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

Derivations for a single observation

$$
\begin{aligned}
P(\theta|\sigma, y) &\propto \exp(-\frac{1}{2\sigma^2}(y-\theta)^2 - \frac{1}{2\tau_0^2}(\theta-\mu_0)^2) \\
&= \exp[-\frac{1}{2}(\theta^2(\frac{1}{\sigma^2}+\frac{1}{\tau_0^2}) - 2\theta(\frac{y}{\sigma^2}+\frac{\mu_0}{\tau_0^2}) + (\frac{y^2}{\sigma^2}+\frac{\mu_0^2}{\tau_0^2}))] \\
&= \exp[-\frac{a}{2}(\theta^2 - 2\frac{b}{a}\theta + \frac{c}{a})] \\
&= \exp[-\frac{a}{2}(\theta^2 - 2\frac{b}{a}\theta + \frac{b^2}{a^2} - \frac{b^2}{a^2} + \frac{c}{a})] \\
&= \exp[-\frac{a}{2}(\theta - \frac{b}{a})^2]\exp[-\frac{a}{2}(\frac{c}{a} - \frac{b^2}{a^2})] \\
&\propto \exp[-\frac{a}{2}(\theta - \frac{b}{a})^2]
\end{aligned}
$$

This is a normal distribution with mean b/a and variance $1/a$.

Students' height example

- Let's assume that the height (in inch) of students in this class follows a normal distribution $N(\theta, 16)$.
- We use a $\theta \sim N(65, 9)$ prior.
- We measure the hight of three students: $y_1 = 72$, $y_2 = 75$, and $y_3 = 70$.
- The posterior distribution of θ is also a normal distribution $N(\mu_n, \tau_n^2)$, where

$$\mu_n = \frac{\frac{65}{9} + \frac{217}{16}}{\frac{1}{9} + \frac{3}{16}} \qquad \frac{1}{\tau_n^2} = \frac{1}{9} + \frac{3}{16}$$

- Therefore, $\theta | y \sim N(69.6, 3.4)$
- The role of prior is substantial here due to the small sample size. The prior modifies the likelihood based estimate (i.e., $\bar{y} = 72.3$), which could have been misleading since all the observed data points happened to be from tall people.

Students’ height example

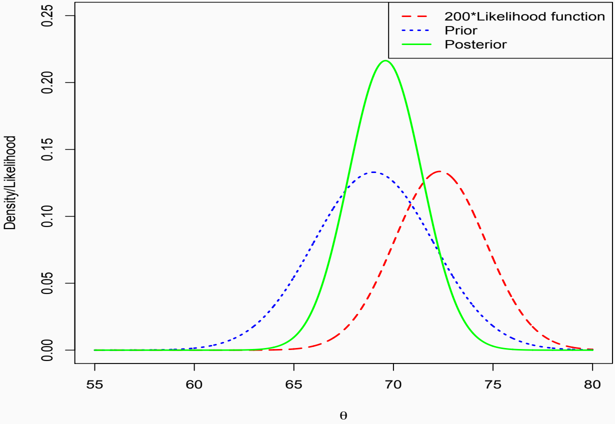


Figure 4: Again, the posterior distribution could be interpreted as a compromise between the prior and the likelihood.

Prediction in univariate normal model

- In the above example, what would be our prediction for the height of next person we observe?
- Denote our prediction as $\tilde{y}$, and the corresponding distribution as $P(\tilde{y}|y)$, i.e., the posterior predictive probability. As before,

$$P(\tilde{y}|y) = \int P(\tilde{y}|\theta)P(\theta|y)d\theta$$

- By integrating out θ, the conditional distribution of $\tilde{y}$ given y is normal with the following mean and variance:

$$\begin{aligned} E(\tilde{y}|y) &= E(E(\tilde{y}|\theta, y)|y) = E(\theta|y) = \mu_n \\ Var(\tilde{y}|y) &= E(var(\tilde{y}|\theta, y)|y) + Var(E(\tilde{y}|\theta, y)|y) \\ &= E(\sigma^2|y) + Var(\theta|y) \\ &= \sigma^2 + {\tau_n}^2 \end{aligned}$$

Prediction in univariate normal model

- We could use μ_n, the posterior expectation of θ, as our single point estimate for $\tilde{y}$.
- The variation around this estimate (i.e., our uncertainty) comes from two difference sources: σ^2, the sampling variation (which is assumed fixed here) of data according to the model, and τ_n^2, the posterior variation of the model parameter, θ, given the observed data.
- In the height example, our guess for the height of the forth student can be expressed by a $N(69.6, 19.4)$ distribution.

Multivariate normal model

- For multivariate normal distribution with known covariance, Σ, we assume

$$\begin{aligned} \mathbf{x} &\sim N_p(\boldsymbol{\mu}, \Sigma) \\ \boldsymbol{\mu} &\sim N_p(\boldsymbol{\mu}_0, \Sigma_0) \end{aligned}$$

- The posterior distribution of $\boldsymbol{\mu}$ given n observations is also a multivariate normal distribution,

$$\boldsymbol{\mu}|\bar{\mathbf{x}} \sim N_p(\boldsymbol{\mu}_n, \Sigma_n)$$

  where

$$\begin{aligned} \boldsymbol{\mu}_n &= (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}(\Sigma_0^{-1}\boldsymbol{\mu}_0 + n\Sigma^{-1}\bar{\mathbf{x}}) \\ \Sigma_n &= (\Sigma_0^{-1} + n\Sigma^{-1})^{-1} \end{aligned}$$

- One can read more about conjugate priors here.

Bayesian regularized linear regressions with conjugate priors

Bayesian Linear regression models

- Consider the following liner regression model:

$$y|x, \beta, \sigma^2 \sim N(x\beta, \sigma^2 I_n)$$

- y is a column vector of n outcome observations, x is an $n \times (p+1)$ matrix of predictors with its first column being all 1's.
- β is a column vector with $p+1$ elements $(\beta_0, \beta_1, ..., \beta_p)$ where β_0 is the intercept and β_j is the effect of the j^{th} predictor x_j on y.
- In Bayesian analysis, a common prior for parameters are

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$$
$$\beta|\mu_0, \Lambda_0 \sim N_{p+1}(\mu_0, \Lambda_0)$$

where $\mu_0 = (\mu_{00}, \mu_{01}, ..., \mu_{0p})$ typically set to zero (unless we believe otherwise), and $\Lambda_0 = \text{diag}(\tau_0^2, \tau_1^2, ..., \tau_p^2)$ should be sufficiently broad.

Posterior distributions of β

- The posterior distributions of β has the following closed form:

$$\begin{aligned}
\beta|x, y, \sigma^2 &\sim N(\mu_n, \Lambda_n) \\
\mu_n &= (x_*' \Sigma_*^{-1} x_*)^{-1} x_*' \Sigma_*^{-1} y_* \\
\Lambda_n &= (x_*' \Sigma_*^{-1} x_*)^{-1} \\
x_* &= \begin{pmatrix} x \\ I_{p+1} \end{pmatrix} \qquad y_* = \begin{pmatrix} y \\ \mu_0 \end{pmatrix} \qquad \Sigma_* = \begin{pmatrix} \sigma^2 I_n & 0 \\ 0 & \Lambda_0 \end{pmatrix}
\end{aligned}$$

- Looking at it this way, the prior plays the role of extra data with $x_{\beta = I_{p+1}}$, $y_\beta = \mu_0$ and the covariance Λ_0.
- That's why Bayesian models do not break down when $p > n$.

Connection to ridge regression!

- Let's take a closer look at the maximum a posterior (MAP)

$$\begin{aligned}\mu_n &= (x_*'\Sigma_*^{-1}x_*)^{-1}x_*'\Sigma_*^{-1}y_* \\ &= (\sigma^{-2}x'x + \Lambda_0^{-1})^{-1}(\sigma^{-2}x'y + \Lambda_0^{-1}\mu_0)\end{aligned}$$

- Let $\mu_0 = 0$, $\sigma^2\Lambda_0^{-1} = \lambda I$, then we have

$$\mu_n = \hat{\beta}^{\text{ridge}} = (x'x + \lambda I)^{-1}x'y$$

- This is exactly the solution to ridge regression!
- Indeed, if we write down the negative logarithm of posterior density of β we have

$$\begin{aligned}-\log P(\beta|x,y,\sigma^2) &= -\log P(y|x,\beta,\sigma^2) - \log P(\beta) \\ &= \frac{1}{2}\sigma^{-2}\|y - x\beta\|_2^2 + \frac{1}{2}\beta'\Lambda_0^{-1}\beta \\ &= \frac{1}{2}\sigma^{-2}(\|y - x\beta\|_2^2 + \lambda\|\beta\|_2^2)\end{aligned}$$

Posterior distributions of σ^2

- Now, we want to obtain the posterior distribution of σ^2
- Given β, again we have a simple normal model with observations y_i with known mean ($x\beta$), unknown variance σ^2, and conditionally conjugate prior Inv-$\chi^2(\nu_0, \sigma_0^2)$.
- As we saw before, the posterior distribution of $\sigma^2|x, y, \beta$ is also scaled Inv-χ^2

$$\begin{aligned} \sigma^2|x, y, \beta &\sim \text{Inv-}\chi^2(\nu_0 + n, \frac{\nu_0\sigma_0^2 + n\nu}{\nu_0 + n}) \\ \nu &= \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i\beta)^2 \end{aligned}$$

Improper priors

- If we do not have an informative priors, we can instead use the following prior:

$$p(\beta, \sigma^2 | x) \propto \sigma^{-2}$$

- For β this is equivalent (in limit) to taking all $\tau_j^2 \to \infty$.
- The posterior distribution therefore becomes

$$\begin{aligned} \beta | y, \sigma^2 &\sim N(\hat{\beta}, V_\beta \sigma^2) \\ \hat{\beta} &= (x'x)^{-1} x'y, \qquad V_\beta = (x'x)^{-1} \end{aligned}$$

- $\hat{\beta}$ is exactly the OLS solution!
- The posterior distribution of σ^2 also has a closed form

$$\begin{aligned} \sigma^2 | x, y, \hat{\beta} &\sim \text{Inv-}\chi^2(n-p-1, s^2) \\ s^2 &= \frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - x_i \hat{\beta})^2 \end{aligned}$$

Example: Children’s test score

- Consider the children’s test score example discussed by Gelman and Hill (2007).
- In this example, we are interested in the effect of mother’s education (mhsg) and her IQ (miq) on the cognitive test score of 3 to 4 year old children.
- For our Bayesian model, we use the following broad priors

$$\sigma^2 \sim \text{Inv-}\chi^2(1, 0.5)$$
$$\beta \sim N_{p+1}(0, 100^2 I)$$

- We used the Gibbs sampler to obtain 10000 samples and discarded the first 1000.

Example: Children's test score

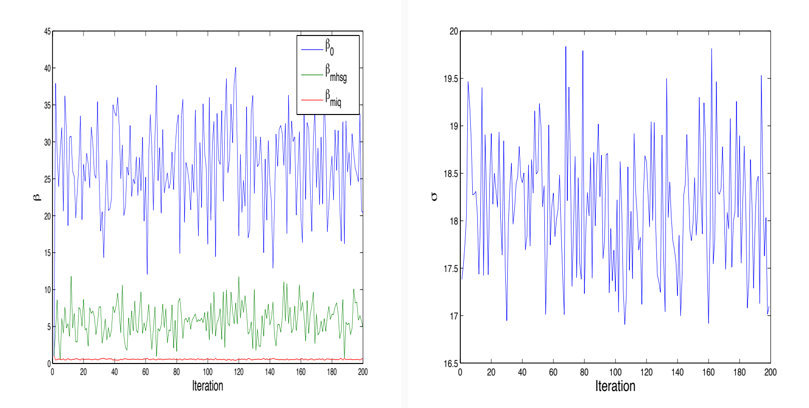


Figure 5: The trace plots of posterior samples for β's (left) and σ (right).

Example: Children's test score

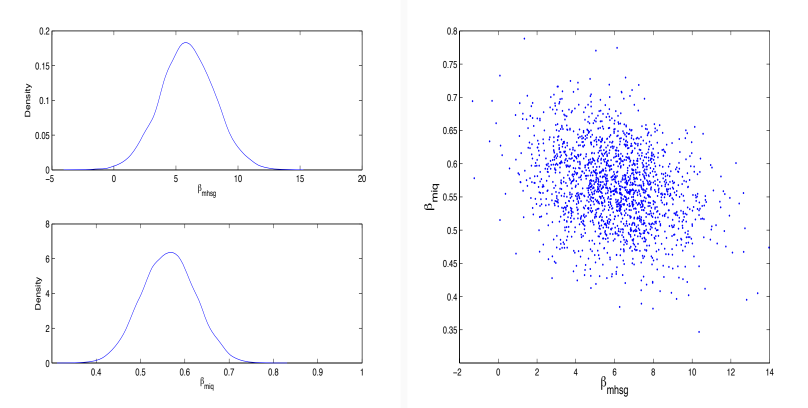


Figure 6: Marginal posterior distributions for β's (left) and the scatter plot of posterior samples for β_{mhsg} and β_{miq} (right).

Example: Children’s test score

Table 1: The posterior estimates and 95% intervals for the regression parameters in the children’s test score example.

Parameter	Posterior expectation	95% Probability Interval
β_0	25.7939	[14.4, 37.2]
β_{mhsg}	5.9278	[1.6, 10.3]
β_{miq}	0.5633	[0.4, 0.7]
σ	18.2	[16.9, 19.4]

Bayesian Lasso

- How about Lasso?

$$\operatorname*{argmin}_{\beta} \|y - x\beta\|_2^2 + \lambda\|\beta\|_1$$

- Can we come up the similar Bayesian version as ridge regression?
- That is, can we have some prior for β, such that the ℓ_1-penalization corresponds to the log-prior?

$$\begin{aligned} -\log P(\beta|x, y, \sigma^2) &= -\log P(y|x, \beta, \sigma^2) - \log P(\beta) \\ &\propto \|y - x\beta\|_2^2 + \lambda\|\beta\|_1 \end{aligned}$$

- Actually, this is called *Laplace* distribution $P(\beta) \propto \exp(-\lambda\|\beta\|_1)$.

Bayesian Lasso

- More generally, we use the following (conditional) Laplace prior

$$P(\beta|\sigma^2) = \prod_{j=0}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_j|/\sqrt{\sigma^2}}$$

  where σ^2 can be given some non-informative prior $1/\sigma^2$.
- This distribution has the following representation as a scale mixture of normals with an exponential mixing density

$$\frac{a}{2} e^{-a|z|} \quad = \quad \int_0^\infty \frac{1}{\sqrt{2\pi s}} e^{-z^2/(2s)} \frac{a^2}{2} e^{-a^2 s/2} ds, \quad a > 0$$

- Denote $\Lambda_0 = \mathrm{diag}(\tau_0^2, \tau_1^2, ..., \tau_p^2)$. Then we use the following priors

$$\begin{aligned} \beta|\sigma^2, \Lambda_0 \quad &\sim \quad N_{p+1}(0, \sigma^2 \Lambda_0) \\ \tau_j \quad &\overset{iid}{\sim} \quad \mathrm{Exp}(\lambda^2/2), \quad j = 0, 1, \cdots, p \end{aligned}$$

Bayesian Lasso

- Then we can have conditional conjugacy and the full conditional posteriors are

$$\begin{aligned}
\beta | x, y, \sigma^2, \Lambda_0^2 &\sim N(\mu_n, \Lambda_n) \\
\mu_n &= (x'x + \Lambda_0^{-1})^{-1} x'y \\
\Lambda_n &= \sigma^2 (x'x + \Lambda_0^{-1})^{-1} \\
1/\tau_j^2 &\overset{iid}{\sim} \mathrm{IG}(\mu', \lambda'), \quad j = 0, \cdots, p \\
\mu' = \sqrt{\frac{\lambda^2 \sigma^2}{\beta_j^2}} &, \quad \lambda' = \lambda^2
\end{aligned}$$

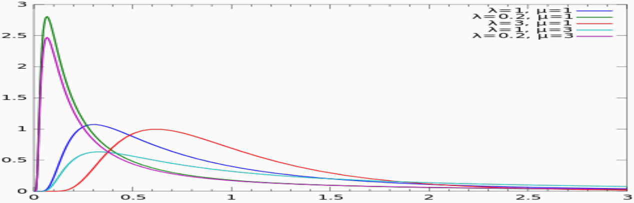
where the inverse-Gaussian distribution $\mathrm{IG}(\mu', \lambda')$ has the following density

$$f(x) = \sqrt{\frac{\lambda'}{2\pi}} x^{-3/2} \exp\left\{ -\frac{\lambda'(x - \mu')^2}{2(\mu')^2 x} \right\}, \quad x > 0$$

Connection to OLS and Ridge?

- Let's take a closer look at the conditional MAP

$$
\begin{aligned}
\mu_n &= (x'x + \Lambda_0^{-1})^{-1} x'y \\
1/\tau_j^2 &\overset{iid}{\sim} \mathrm{IG}(\mu', \lambda'), \quad j = 0, \cdots, p \\
\mu' = \sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}} &, \quad \lambda' = \lambda^2
\end{aligned}
$$



- Let $\lambda \to 0$, $1/\tau_j \to 0$, then $\mu_n \to (x'x)^{-1}x'y$, which is OLS!
- Let $\lambda \to \infty$, $1/\tau_j \sim \lambda$, then $\mu_n \sim (x'x + \lambda I)^{-1}x'y$ which behaves similarly as ridge solution!

Diabetes Data

- Now we consider for the diabetes data of Efron et. al (2004) which has $n = 442$ and $p = 10$.
- We compare Bayesian Lasso with Frequentist Lasso and ridge regression for the entire solution path for λ.

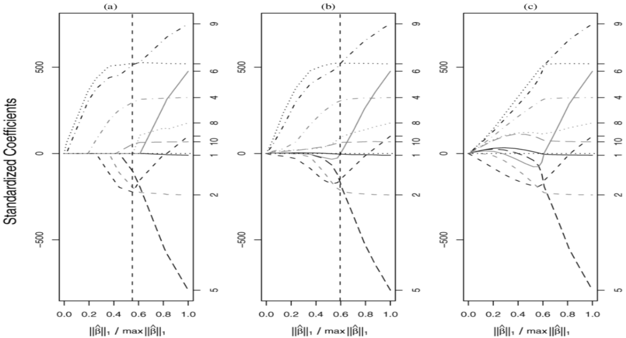


Figure 1. Lasso (a), Bayesian Lasso (b), and ridge regression (c) trace plots for estimates of the diabetes data regression parameters versus the relative L_1 norm, with vertical lines for the Lasso and Bayesian Lasso indicating the estimates chosen by n-fold cross-validation and marginal maximum likelihood. The Bayesian Lasso estimates were posterior medians computed over a grid of λ values, using 10,000 consecutive iterations of the Gibbs sampler of Section 2 (after 1,000 burn-in iterations) for each λ.

Bayesian bridge regression

- We can consider the following more general 'bridge' regression

$$\underset{\beta}{\text{argmin}}\ \|y - x\beta\|_2^2 + \lambda\|\beta\|_q^q, \qquad \|\beta\|_q := \left(\sum_{j=0}^{p} |\beta_j|^q\right)^{1/q}$$

- One can consider the following (conditional) prior

$$P(\beta|\sigma^2) \quad \propto \quad \prod_{j=0}^{p} e^{-\lambda(|\beta_j|/\sqrt{\sigma^2})^q}$$

- And construct a similar mixture representation which is much more involved.
- Read *The Bayesian Lasso* (2008) by Park and Casella.
- In the next lecture, we will introduce a special MCMC algorithm that can handle the probability distributions defined on the domains constrained by q-norms. It will enable us to use more general priors in a natural way.