COLUMBIA UNIVERSITY
Statistical Inference and Modeling
Lecture Note 15

Prof. Marco Avella
Scribe: Harin Sanghirun
Spring 2019

Bayesian Statistics

Motivation

- Incorporate prior belief into statistical analysis
- Introduce key ideas of Bayesian approach to statistics
- Understand main pros and cons of Bayesian statistics
- Elements of frequentist analysis of Bayesian methods

Frequentist V. Bayesian approach

Frequentist:

- Observe data assumed to be generated randomly (by nature, by measurements, by design of experiment, etc...)
- Make assumptions of the generating process (e.g. iid normal, Markov chain, linear time series, GLM, etc.)
- The generating process is specified by some parameter of interest.
- Parameters are fixed but unknown. They are estimable from the data and we test hypothesis related to them

Bayesian:

- Still observe randomly generated data generated by some process specified by some unknown parameter.
- We have some prior belief about the parameters of a distribution.

  example. If we believe that our data follows

  - a Poisson distribution, we might have some belief about its mean (which defines the distribution).
  - a Normally distributed data, then we might have some belief about its mean and variance.

- We want to update our believe using the data and transform it into a posterior believe.

example. Women proportion
Let p be the proportion of women in the population. Sample n people at random with replacement in the population and denote by $X_1, \dots, X_n$ their sex (1 for woman, 0 otherwise).

In frequentist approach we estimate p with MLE and construct some confidence interval. We do some hypothesis testing e.g. $H_0 : p = 1/2$ versus $H_1 : p \neq 1/2$.

Before analyzing the data, we may believe that p is likely to be close to $1/2$ and the Bayesian approach is a way to update our believe using the data.

We can model our believe using a distribution for p, hence treating p as a random variable! In this case the true parameter is not random but it is modeled as a random variable to model our belief.

We could for example assume that $p \sim Beta(a, a)$ for some $a > 0$. This distribution is called the prior distribution. Recall that a $Beta(a, b)$ distribution gives a probability of a value between 0 and 1 with the expected value $\frac{a}{a+b}$. So a $Beta(a, a)$ distribution has the expected value of $1/2$, the value we belief that the proportion of women should lie around.

Since Bayesian statistics relies on the Bayes' Theorem to do inference after we have seen new evidence, it is useful to review the Bayes' Theorem.

Definition. Bayes' theorem
Let $A_1, \dots, A_k$ be events that partition a sample space, let B be an arbitrary event on that space for which $\mathbb{P}(B) > 0$. Then Bayes' theorem is

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}$$

This reverses the order of conditioning by expressing $\mathbb{P}(A_j|B)$ in terms of $\mathbb{P}(B|A_i)$ and the marginal probability $\mathbb{P}(B)$ in the denominator.

- For continuous random variables Y and Z

$$f_{Z|Y}(z|y) = \frac{f_{Y|Z}(y|z)f_Z(z)}{\int f_{Y|Z}(y|z)f_Z(z)\mathrm{d}z},$$

provided the marginal density $f(y) > 0$. The integration is replaced by summation for discrete random variables.

The Posterior Distribution

Combining our belief and the information we gain from observation gives us the posterior distribution, which represents our belief about the parameter of a distribution after we have seen the data.

Let θ be the parameter that we have some prior belief on, $X_1, \dots, X_n$ be the information we observed, and $\pi(\cdot)$ a density function, then applying the Bayes' Theorem we have

$$\pi(\theta|X_1, \dots, X_n) = \frac{\pi(\theta) f(X_1, \dots, X_n|\theta)}{f(X_1, \dots, X_n)} = \frac{\pi(\theta) f(X_1, \dots, X_n|\theta)}{\int \pi(\theta) f(X_1, \dots, X_n|\theta) \mathrm{d}\theta}, \text{ for all } \theta \in \Theta.$$

We call the density of our prior belief conditioning on the data $\pi(\theta|X_1, \dots, X_n)$ the posterior distribution.

Since the normalizing constant $f(X_1, \dots, X_n)$ does not depend on θ, we can also write

$$\pi(\theta|X_1, \dots, X_n) \propto \pi(\theta) f(X_1, \dots, X_n|\theta),$$

Bayes' Theorem tells us that $\pi(\theta|X_1, \dots, X_n)$ is a probability distribution i.e. it integrates to one, so we do not need to be concerned about the variables that is not the input of our function since it is only there to make the function integrates to one. We can match the numerator with a form of distribution that we know. Once we have a matching distribution, we can infer the normalizing constant from the distribution's density without having to evaluate the denominator (which often times cannot even be solved).

example. Let $X_1, \dots, X_n|\theta \sim N(\theta, 1)$ with prior $\theta \sim N(0, 1)$. The numerator of the posterior distribution, as a function of θ is

$$\begin{aligned} p(\theta|X_1, \dots, X_n) &\propto p(\theta) p(X_1, \dots, X_n|\theta) \\ &\propto \frac{1}{\sqrt{2\pi}} \exp\{\frac{-\theta^2}{2}\} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\Big\{ - \frac{(X_i - \theta)^2}{2} \Big\} \\ &\propto \exp\Big(- \frac{1}{2} \sum_i X_i^2 + \theta \sum_i X_i - \frac{n}{2}\theta^2 - \frac{1}{2}\theta^2 \Big) \\ &\propto \exp\Big(n\theta\bar{X} - \frac{(n+1)}{2}\theta^2 \Big) \\ &\propto \exp\Big[- \frac{\{\theta - n\bar{X}/(n+1)\}^2}{2/(n+1)} \Big] \end{aligned}$$

Therefore the posterior distribution of θ is $N(\frac{1}{n+1}\sum_i X_i, \frac{1}{n+1})$.

Ignorance

Non informative priors

Sometimes, we lack prior information, we want to use a prior that provides as little information as possible. A good candidate is the prior $\pi(\theta) \propto 1$, i.e. constant pdf on parameter space Θ.

If Θ is bounded, this is a *uniform prior* on Θ. If Θ is unbounded, it does not define a proper pdf since it is not integrable!

An improper prior on θ still defines a posterior distribution using Bayes' rule.

example. if $\pi(\theta) = 1$ for all $\theta \in \mathbb{R}$ and given θ, $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$:

$$\begin{aligned}\pi(\theta|X_1,\ldots,X_n) &\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2\right\}\\ &\propto \exp\left\{n\bar{X}\theta - n\theta^2/2\right\}\\ &\propto \exp\left\{-\frac{1}{2n^{-1}}(\theta-\bar{X})^2\right\}\end{aligned}$$

i.e. the posterior distribution is $N(\bar{X}, \frac{1}{n})$.

Jeffreys' prior

Making ignorance invariant under reparametrization

- Paradox: The probability of success in a Bernoulli trial lies in $[0, 1]$. If we are completely ignorant of its true value, the obvious prior is $\pi(\theta) = 1,\ 0 \leq \theta \leq 1$. But if we are completely ignorant of θ , we are also completely ignorant of $\psi = \log(1/\theta)$. However the density implied by the uniform prior for θ implies that $\psi \sim Exp(1)$, which is far from expressing ignorance about ψ.

  To prove that the reparametrized ψ follows the claimed distribution, we can use the following fact from Section 1.5 of Knight's Mathematical Statistics.

Suppose a random variable $X \sim F_x$ is known and $Y = g(X)$ where g is a monotone function then

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{\partial g^{-1}(y)}{\partial y}\right|$$

Proof.

$$X \sim \text{Uniform}(0, 1)$$

$$Y = \log(1/\theta) = -\log(\theta) = g(x) \iff g^{-1}(Y) = e^{-Y} = \theta$$

$$f_Y(y) = 1 \cdot |-e^{-y}| = e^{-y} \iff Y \sim Exp(1)$$

□

- The problem above is that the representation of ignorance is not invariant under reparametrization. A solution to this is Jeffreys' prior

  $$\pi_J(\theta) \propto \sqrt{\det I(\theta)},$$

  where $I(\theta)$ is the Fisher information matrix of the statistical model associated with $X_1, \ldots, X_n$ in frequentist approach.

Conjugate densities

Closed form posteriors

- In a statistical model, when the prior $\pi(\theta)$ and $\pi(\theta|X_1, \ldots, X_n)$ belong to the same family of distributions, it is called conjugate prior.
- They are particularly appealing because they lead to closed form posterior distributions.

Examples

1. Normal prior and normal sampling for a normal posterior.
2. Beta prior and binomial sampling for a Beta posterior.
3. Gamma prior and Poisson sampling for Gamma posterior.

example. Gamma Prior and Poisson Sampling

Let $X_1, \ldots, X_n|\lambda \sim \text{Poisson}(\lambda)$ and $\lambda \sim \text{Gamma}(\alpha, \nu)$

$$f(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!},\ \pi(\lambda) = \frac{\nu^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\nu\lambda},\ \lambda > 0$$

$$\begin{aligned}\pi(\lambda|X_1,\ldots,X_n) &\propto \pi(\lambda)p(X_1,\ldots,X_n|\lambda)\\ &\propto \lambda^{\alpha-1}e^{-\nu\lambda}\prod_{i=1}^{n}e^{-\lambda}\lambda^{x_i}\\ &\propto \lambda^{\alpha-\sum_i x_i-1}e^{-\nu\lambda-n\lambda}\\ &\propto \lambda^{\alpha-\sum_i x_i-1}e^{-(\nu+n)\lambda}\end{aligned}$$

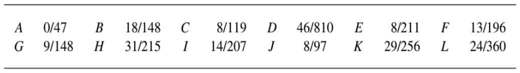
Since a Gamma(a,b) is defined by the pdf $f(x)\propto x^{a-1}e^{-bx}$, we can see that our posterior is also a Gamma with $a=\alpha-\sum_i x_i$ and $b=\nu+n$, in another words

$$\pi(\lambda|X_1,\ldots,X_n)\sim \text{Gamma}(\alpha-\sum_i x_i,\nu+n)$$

If we were to write it out with the normalizing constant, we would have

$$\pi(\lambda|X_1,\ldots,X_n)=\frac{(\nu+n)^{\alpha+\sum_i X_i}\lambda^{\alpha+\sum_i X_i-1}}{\Gamma(\nu+\sum_i X_i)}e^{-(\nu+n)\lambda}$$

example. A simple Bayesian approach
Data on the mortality levels for cardiac surgery on babies at 12 hospitals. A simple model treats the number of deaths r as binomial with mortality rate θ and denominator m. At hospital A, for example, $m=47$ and $r=\theta$, giving maximum likelihood estimate $\theta_A=0/47=0$, which seems too optimistic. If we take a beta prior density with $a=b=1$, the posterior density is beta with parameters $a+r=1$ and $b+m-r=48$.



A	0/47	*B*	18/148	*C*	8/119	*D*	46/810	*E*	8/211	*F*	13/196
G	9/148	*H*	31/215	*I*	14/207	*J*	8/97	*K*	29/256	*L*	24/360

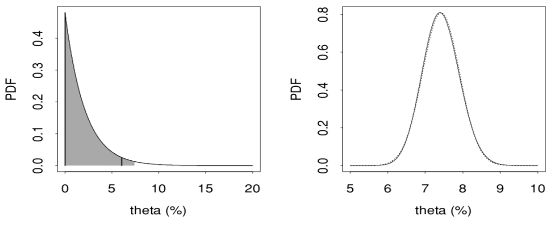


Figure 1: Bottom left panel: posterior density for θ_A, showing boundaries of 0.95 highest posterior credible interval (vertical lines) and region between posterior 0.025 and 0.975 quantiles of $\pi(\theta_A|y)$ (shaded). Botton right panel: exact posterior beta density for overall mortality rate θ (solid) and normal approximation (dots).

A Bayesian hierarchical model

When our data consist of sampling from different groups which we might suspect to have different characteristic, we may use a Bayesian hierarchical model which model between-group variation and within-group variation through a hierarchy of distribution.

example. Continuing with our cardiac arrest data.

Although the number of operations and the death rates vary, we have no further knowledge of the hospitals and hence no basis for treating them other than entirely symmetrically, suggesting the hierarchical model

$$r_j|\theta_j \overset{ind}{\sim} Bin(m_j, \theta_j), \quad j = A, \dots, L, \quad \theta_A, \dots, \theta_L|\xi \overset{iid}{\sim} f(\theta|\xi), \quad \xi \sim \pi(\xi)$$

where

- r_j is the number of death at hospital j
- θ_j is the probability of dying at hospital j
- m_j the number of operations

A simple model formulation is to suppose that given $\xi = (\mu, \sigma^2)$ the log odds of death have a normal distribution, i.e.

$$\beta_j = \log\{\theta_j/(1-\theta_j)\} \sim N(\mu, \sigma^2)$$

and additionally assume that $\mu \sim N(0, c^2)$ and $\sigma^2 \sim \text{InvGamma}(a, b)$ for some specified a, b, c. Note that normal and inverse gamma distribution makes the conjugate prior for a normal sampling model, so we are choosing it out of mathematical convenience. We can summarize our hierarchical model with the follow diagram, where the arrow points from the variable that is dependent on the the dependent variable.

$$\begin{tikzcd}
\mu \sim N(0,c^2), \sigma^2 \sim InvGamma(a,b) & & \mu, \sigma^2 \arrow[dl] \arrow[dr] & \\
\theta_j|\mu, \sigma^2 \sim F(\mu, \sigma^2) & \theta_1 \arrow[d] & \cdots & \theta_J \arrow[d] \\
r_{i,j}|\theta_j \sim \text{Bern}(\theta_j) & r_{1,1}, \dots, r_{m_1,1} & \cdots & r_{1,J}, \dots, r_{m_J,J}
\end{tikzcd}$$

where F is some distribution that is conditional on μ and σ^2, but not Normal, since we assume that a transformation of θ_j is Normal.

The interpretation of this model is that we think that each hospital has a probability of dying from operation θ_j, but θ_j varies across the hospital through some transformation of the normal distribution.

One benefit of a hierarchical model is that information across hospitals are shared through μ and σ^2, this mean that if we have small data point on some hospital, we will still be able to make reasonable inference about it, without having extreme results such as the probability of death at some hospital being 0 or 1.

Finally, we have the joint density

$$\prod_j \frac{m_j}{(m_j - r_j)!r_j!} \frac{e^{r_j \beta_j}}{(1 + e^{\beta_j})^{m_j}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ - \frac{1}{2\sigma^2}(\beta_j - \mu)^2 \right\} \times \pi(\mu)\pi(\sigma^2)$$

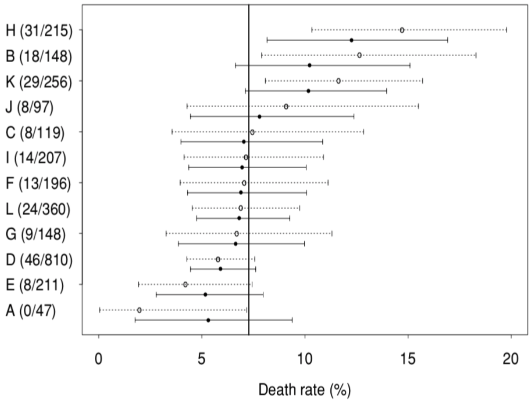


Figure 2: Posterior summaries for mortality rates for cardiac surgery data. Posterior means and 0.95 equitailed credible intervals for separate analyses for each hospital are shown by hollow circles and dotted lines, while blobs and solid lines show the corresponding quantities for a hierarchical model. Note the shrinkage of the estimates for the hierarchical model towards the overall posterior mean rate, shown as the solid vertical line; the hierarchical intervals are slightly shorter than those for the simpler model.

Bayesian Estimation

- The posterior distribution $\pi(\theta|X_1, \ldots, X_n)$ is a random probability distribution on the parameter space Θ and can be used to address statistical questions about θ.

- Posterior mean estimator

$$\hat{\theta} = \mathbb{E}_{\theta|X}[\theta] = \int_\Theta \theta\pi(\theta|X_1, \ldots, X_n)\mathrm{d}\theta$$

  Note that this estimator depends on the prior π through the posterior distribution.

- More generally, one can define a Bayes estimator as the minimizer of expected posterior loss i.e.

  $$\hat{\theta}_{Bayes} = \text{argmin}_{\theta\in\Theta}\mathbb{E}_{\theta|X}[L(X,\theta)].$$

  The posterior mean minimizes the expected squared loss $L_2(X,\theta) = (X-\theta)^2$.

Bayesian Inference

- The Bayesian analogue of a $(1-\alpha)$ confidence interval is a $(1-\alpha)$ credible set defined as a random set $\mathcal{C} \subset \Theta$ such that

  $$1-\alpha = \mathbb{P}(\theta \in \mathcal{C}|X_1,\dots,X_n) = \int_{\mathcal{C}} \pi(\theta|X_1,\dots,X_n)\mathrm{d}\theta$$

- The Bayesian analogue of test statistics for comparing a null hypothesis H_0 with an alternative hypothesis H_1 is to look at

  $$\underbrace{\frac{\mathbb{P}(H_1|X_1,\dots,X_n)}{\mathbb{P}(H_0|X_1,\dots,X_n)}}_{\text{Posterior odds}} = \underbrace{\frac{\mathbb{P}(X_1,\dots,X_n|H_1)}{\mathbb{P}(X_1,\dots,X_n|H_0)}}_{\text{Bayes factor}} \times \underbrace{\frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)}}_{\text{Prior odds}}$$

  Therefore the change in prior to posterior odds for H_1 relative to H_0 depends on the data only through the Bayes factor.

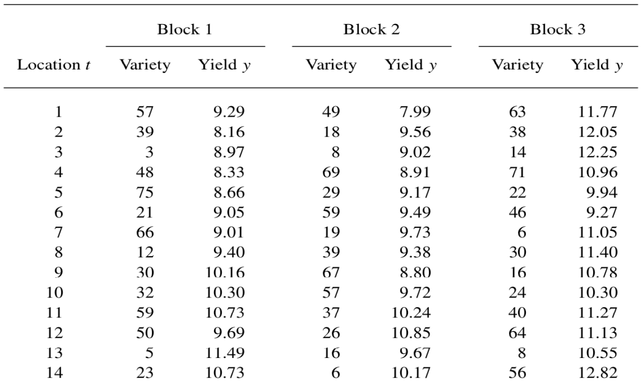
example. Spring Barley data

GAM example revisited

A more flexible model that allows the yield for the νth variety in the bth block to depend on its location $t_{\nu b}$ is

$$y_{\nu b} = g_b(\tau_b) + \beta_\nu + \varepsilon_{\nu b},\ \ \nu = 1,\dots,75,\ \ b = 1,2,3,$$

where $\varepsilon \overset{iid}{\sim} N(0,\sigma^2)$ and $g_b(\cdot)$ is smooth function that determines how the fertility pattern in block b depends on the location t.



	Block 1		Block 2		Block 3	
Location t	Variety	Yield y	Variety	Yield y	Variety	Yield y
1	57	9.29	49	7.99	63	11.77
2	39	8.16	18	9.56	38	12.05
3	3	8.97	8	9.02	14	12.25
4	48	8.33	69	8.91	71	10.96
5	75	8.66	29	9.17	22	9.94
6	21	9.05	59	9.49	46	9.27
7	66	9.01	19	9.73	6	11.05
8	12	9.40	39	9.38	30	11.40
9	30	10.16	67	8.80	16	10.78
10	32	10.30	57	9.72	24	10.30
11	59	10.73	37	10.24	40	11.27
12	50	9.69	26	10.85	64	11.13
13	5	11.49	16	9.67	8	10.55
14	23	10.73	6	10.17	56	12.82

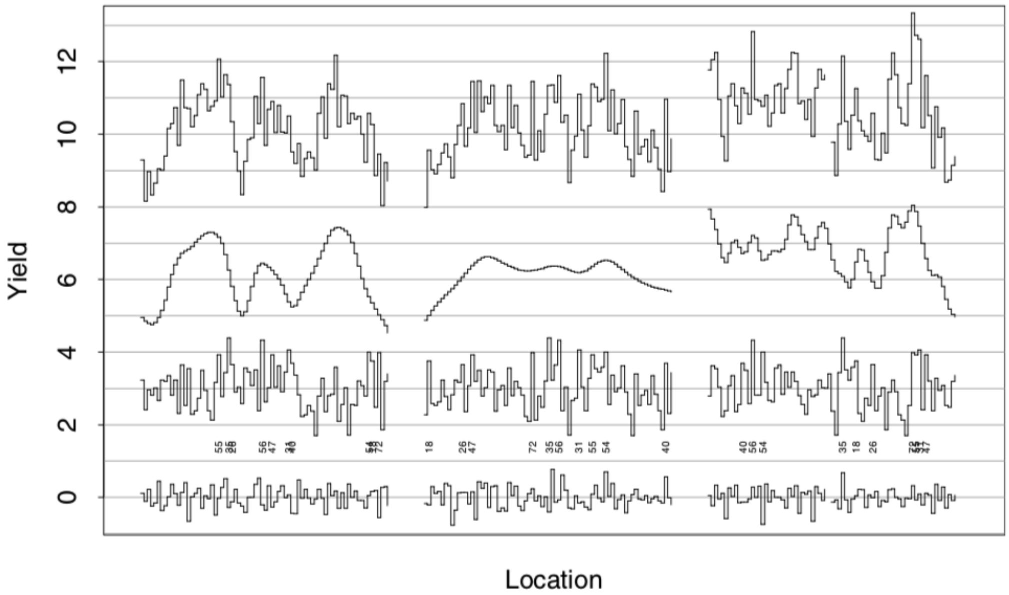


Figure 3: Spring barley data analysis. Block 1 is shown on the left and block 3 on the right. The panel shows, from the top, the original yields y, the fertility trend and variety effect estimates $\hat{g}_b(t)$ and β_ν, both offset for display, and the crude residuals. The varieties with the ten largest β_ν a remarked.

A Bayesian hierarchical model

- Let $\mathbf{y} = (y_1, \ldots, y_n)^T$ denote the yields in the $n = 225$ plots and let ψ_j denote the unknown fertility of plot j. Further let $\mathbf{X}$ denote the $n \times p$ design matrix showing which of the $p = 75$ variety parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ have been allocated to the plots.

- Assume the model

$$\mathbf{y}|\boldsymbol{\beta}, \psi, \lambda_y \sim N_n(\psi + \mathbf{X}\boldsymbol{\beta}, \lambda_y^{-1}\mathbf{I}_n),$$

with $\lambda_y \sim \text{Gamma}(a, b)$, a, b known and the hierarchical prior for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} \sim N_p(\mathbf{0}, \lambda_\beta^{-1}\mathbf{I}_p),\ \ \lambda_\beta \sim \text{Gamma}(c, d), \quad c, d \text{ known}.$$

To account for fertility patterns for the different blocks, one can use the normal Markov field

$$\pi(\psi|\lambda_\psi) \propto \lambda_\psi^{n/2} \exp\Big\{-\frac{1}{2}\lambda_\psi \sum_{i\sim j}(\psi_i - \psi_j)^2\Big\} = \lambda_\psi^{n/2} \exp\Big\{-\frac{1}{2}\lambda_\psi \psi^T \mathbf{W}\psi\Big\},$$

where the summation is taking over pairs of neighboring plots and $\lambda_\psi \sim \text{Gamma}(g,h)$, with g and h specified.

The above hierarchical model with conjugate priors leads to a closed form joint posterior density for $\boldsymbol{\beta}, \psi, \boldsymbol{\lambda}$. Setting $a = c = g = 1$ and $b = d = h = 0.005$, and simulating from the above model seems to lead to a reasonable fit.

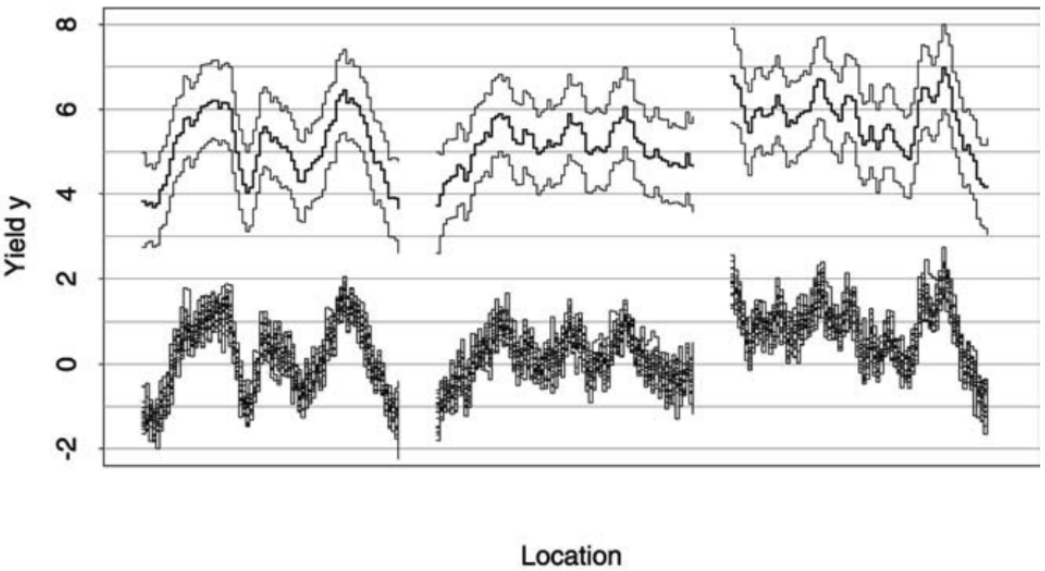
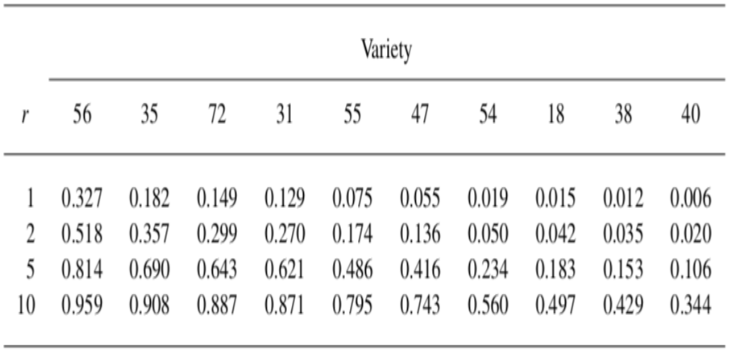


Figure 4: Figure 11.12 Posterior summaries for fertility trend ψ for the three blocks of spring barley data, shown from left to right. Above: median trend (heavy) and overall 0.9 posterior credible bands. Below: 20 simulated trends from Gibbs sampler output.



	Variety									
r	56	35	72	31	55	47	54	18	38	40
1	0.327	0.182	0.149	0.129	0.075	0.055	0.019	0.015	0.012	0.006
2	0.518	0.357	0.299	0.270	0.174	0.136	0.050	0.042	0.035	0.020
5	0.814	0.690	0.643	0.621	0.486	0.416	0.234	0.183	0.153	0.106
10	0.959	0.908	0.887	0.871	0.795	0.743	0.560	0.497	0.429	0.344

Figure 5: Posterior probabilities that a variety is ranked among the best r varieties, estimated from 10,000 iterations of Gibbs sampler.

Normal linear model

Ridge regression

- Assuming a random sample of pairs (X_i, Y_i) is such that

$$Y_i = \mathbf{X}_i^T\boldsymbol{\beta} + \varepsilon_i, \ i = 1, \dots, n,$$

$$\mathrm{cov}(\varepsilon_i, \mathbf{X}_i) = 0, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \ \text{ and } \ \boldsymbol{\beta} \sim N_p(\mathbf{0}, \tau^2\mathbf{I}_p).$$

- Posterior distribution of $\boldsymbol{\beta}$:

$$\pi(\boldsymbol{\beta}|\mathbf{X},\mathbf{Y}) = f(\mathbf{X},\mathbf{Y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta}) \propto \exp\Big\{-\frac{1}{2\sigma^2}\Big(\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{\sigma^2}{\tau^2}\|\boldsymbol{\beta}\|_2^2\Big)\Big\}$$

  and some manipulations show that

$$\boldsymbol{\beta}|\mathbf{X},\mathbf{Y} \sim N_p\Big(\big(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I}\big)^{-1}\mathbf{X}^T\mathbf{Y}, \big(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I}\big)^{-1}\Big)$$

- Bayes posterior mean estimator:

$$\hat{\boldsymbol{\beta}} = \mathbb{E}[\boldsymbol{\beta}|\mathbf{X},\mathbf{Y}] = \big(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I}\big)^{-1}\mathbf{X}^T\mathbf{Y}$$

Spline smoothing

Bayesian view of roughness penalty

Remember that in the roughness penalty approach to nonparametric regression we find the penalized least squares estimator that minimizes

$$\hat{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}}\Big\{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\boldsymbol{\beta}^T\mathbf{P}\boldsymbol{\beta}\Big\}$$

where

$$\mathbf{X} = \begin{bmatrix} b_1(x_1) & \dots & b_K(x_1) \\ b_1(x_2) & \dots & b_K(x_2) \\ \vdots & & \vdots \\ b_1(x_n) & \dots & b_K(x_n) \end{bmatrix} \text{ and } \mathbf{P} = \begin{bmatrix} \int b_1''(x)b_1''(x)\mathrm{d}x & \dots & \int b_1''(x)b_K''(x)\mathrm{d}x \\ \int b_2''(x)b_1''(x)\mathrm{d}x & \dots & \int b_2''(x)b_K''(x)\mathrm{d}x \\ \vdots & & \vdots \\ \int b_K''(x)b_1''(x)\mathrm{d}x & \dots & \int b_K''(x)b_K''(x)\mathrm{d}x \end{bmatrix}$$

- Bayesian view: penalization as a prior put on wiggliness of the unknown function. In particular, it is equivalent to $\beta \sim N(0, \mathbf{P}^{-}\sigma^2/\lambda)$, where $\mathbf{P}^{-}$ is a generalized inverse of $\mathbf{P}$.

- Posterior distribution:

$$\boldsymbol{\beta}|\mathbf{x},\mathbf{y} \sim N(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{P})^{-1}).$$

Bayesian computation

Monte Carlo techniques

Suppose we want to evaluate the integral

$$\mu = \int f(x|\theta)\pi(\theta)\mathrm{d}\theta$$

- Monte Carlo method: generate large sample $\theta_1, \ldots, \theta_S$ from $\pi(\theta)$. Then by LLN we have that $\mu \approx \frac{1}{S}\sum_{s=1}^{S} f(x|\theta_s)$.
- Importance Sampling: generate sample $\theta_1, \ldots, \theta_S$ from density $h(\theta)$ whose support includes that of $\pi(\theta)$ and compute

  $$\hat{\mu} = \frac{1}{S}\sum_{s=1}^{S} f(x|\theta_s) \underbrace{\frac{\pi(\theta_s)}{h(\theta_s)}}_{w(\theta_s)}$$

  or the more commonly used importance sampling ratio estimator

  $$\hat{\mu}_{rat} = \frac{\sum_{s=1}^{S} f(x|\theta_s)w(\theta_s)}{\sum_{s=1}^{S} w(\theta_s)}$$

Markov chain Monte Carlo

Often time, we are not able to sample from the posterior distribution directly, so Monte Carlo sampling does not work. Instead we need to use Markov chain Monte Carlo.

Example: Gibbs sampler

Idea: construct a Markov chain that will, if run for an infinitely long period, generate samples from a posterior distribution, specified implicitly and known only up to a normalizing constant.

- Let $U = (U_1, \ldots, U_k)$ be a random variable of dimension k whose joint density $\pi(u)$ is unknown. We suppose that we can simulate observations from the full *conditional densities* $\pi(u_i|u_{-i})$, where $u_{-i} = (u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_k)$.
- Gibbs sampling is successive simulation from the $\pi(u_i|u_{-i})$ according to:

  1. Initialize by taking arbitrary values of $U_1^{(0)}, \ldots, U_k^{(0)}$.
  2. For $i = 1, \ldots, I$,
     (a) generate $U_1^{(i)}$ from $\pi(u_1|u_2 = U_2^{(i-1)}, \ldots, u_k = U_k^{(i-1)})$,
     (b) generate $U_2^{(i)}$ from $\pi(u_2|u_1 = U_1^{(i-1)}, u_3 = U_3^{(i-1)}, \ldots, u_k = U_k^{(i-1)})$,
     $\vdots$
     (c) generate $U_k^{(i)}$ from $\pi(u_1|u_2 = U_2^{(i-1)}, \ldots, u_{k-1} = U_{k-1}^{(i-1)})$

Markov chain Monte Carlo

Example: Gibbs sampler for bivariate normal with means 0, variances 1 and correlation ρ

Discussion

- Bayesian methods are naturally useful in frequentist world when considering time series of previous studies are available.

- Natural connections with Linear Mixed Models

  $$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon},\ \mathbf{b} \sim N_q(\mathbf{0}, \boldsymbol{\Psi}),\ \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}),$$

  where $\boldsymbol{\beta}$ are fixed effects and $\mathbf{b}$ are random effects.

- Rich literature dedicated to computational aspects such as Laplace approximations and Markov chain Monter Carlo (MCMC).

- Bernstein-von Mises theorem: the posterior distribution behaves for large n like a normal distribution centered at the MLE i.e.

  $$\pi(\theta|X, \ldots, X_n) \approx N(\hat{\theta}_{MLE}, I(\theta)^{-1}/n).$$

  This guarantess nice frequentist behavior of certain Bayes procedures.

References

- A.C. Davison (2003). “Statistical Models” *Chapter 11.1–11.4*

- A. Gelman, H.S. Stern, J.B. Carlin, D.B. Dunson, A. Vehtari and D.B. Rubin (2013). “Bayesian data analysis”