Introduction to Bayesian Inference

Mohan Delampady
Statistics and Mathematics Unit
Indian Statistical Institute, Bangalore

July, 2008

Contents

1 What is Statistical Inference? **2**

2 Frequentist Statistics **4**

3 Conditioning on Data **5**

4 The Bayesian Recipe **6**

5 Inference for Binomial proportion **7**

6 Inference With Normals/Gaussians **10**

7 Bayesian Computations **16**

8 Monte Carlo Sampling **17**

9 Markov Chain Monte Carlo Methods **20**

10 Empirical Bayes Methods for High Dimensional Problems **32**

1 What is Statistical Inference?

It is an inverse problem as in 'Toy Example':
Example 1 (Toy). Suppose a million candidate stars are examined for the presence of planetary systems associated with them. If 272 'successes' are noticed, how likely that the success rate is 1%, 0.1%, 0.01%, $\cdots$ for the entire universe?

Probability models for observed data involve *direct* probabilities:
Example 2. An astronomical study involved 100 galaxies of which 20 are Seyfert galaxies and the rest are starburst galaxies. To illustrate generalization of certain conclusions, say 10 of these 100 galaxies are randomly drawn. How many galaxies drawn will be Seyfert galaxies?

This is exactly like an artificial problem involving an urn having 100 marbles of which 20 are red and the rest blue. 10 marbles are drawn at random with replacement (repeatedly, one by one, after replacing the one previously drawn and mixing the marbles well). How many marbles drawn will be red?

Data and Models

X = number of Seyfert galaxies (red marbles) in the sample (out of sample size $n = 10$)

$$P(X = k|\theta) = \binom{n}{k}\theta^k(1-\theta)^{(n-k)}, \quad k = 0, 1, \ldots n \tag{1}$$

In (1) θ is the proportion of Seyfert galaxies (red marbles) in the urn, which is also the probability of drawing a Seyfert galaxy at each draw. In Example 2, $\theta = \frac{20}{100} = 0.2$ and $n = 10$. So,

$P(X = 0|\theta = 0.2) = 0.8^{10}$, $P(X = 1|\theta = 0.2) = 10 \times 0.2 \times 0.8^9$, and so on.

In practice, as in 'Toy Example', θ is unknown and inference about it is the question to solve.

In the Seyfert/starburst galaxy example, if θ is not known and 3 galaxies out of 10 turned out to be Seyfert, one could ask:

how likely is $\theta = 0.1$, or 0.2 or 0.3 or $\ldots$?

Thus inference about θ is an inverse problem:

Causes (parameters) $\longleftarrow$ Effects (observations)

How does this *inversion* work?

The direct probability model $P(X = k|\theta)$ provides a *likelihood function* for the unknown *parameter* θ when data $X = x$ is observed:

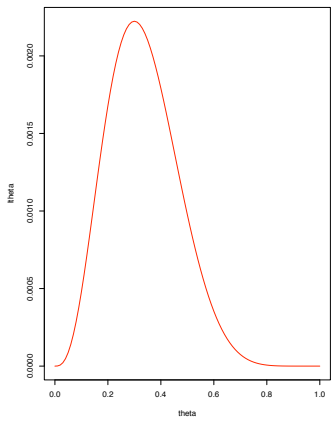
$l(\theta|x) = f(x|\theta)$ $(= P(X = x|\theta)$ when X is a discrete random variable) as function of θ for given x.

Interpretation: $f(x|\theta)$ says how likely x is under different θ or the model $P(.|\theta)$, so if x is observed, then $P(X = x|\theta) = f(x|\theta) = l(\theta|x)$ should be able to indicate what the likelihood of different θ values or $P(.|\theta)$ are for that x.

As a function of x for fixed θ $P(X = x|\theta)$ is a probability mass function or density, but as a function of θ for fixed x, it has no such meaning, but just a measure of likelihood.

After an experiment is conducted and seeing data x, the only entity available to convey the information about θ obtained from the experiment is $l(\theta|x)$.

For the Urn Example we have $l(\theta|X = 3) \propto \theta^3(1 - \theta)^7$:



Maximum Likelihood Estimation (MLE): If $l(\theta|x)$ measures the likelihood of different θ (or the corresponding models $P(.|\theta)$), just find that $\theta = \hat{\theta}$ which maximizes the likelihood.

For model (1)

$$\hat{\theta} = \hat{\theta}(x) = x/n = \text{ sample proportion of successes .}$$

This is only an estimate. How good is it? What is the possible error in estimation?

Likelihood function $l(\theta|x)$ has nothing to say about these.

2 Frequentist Statistics

Consider repeating this experiment again and again. Then one can look at all possible sample data. i.e. all possible x values. Utilize *long-run average behaviour* of the MLE. i.e. treat $\hat{\theta}$ as a random quantity by replacing x by X in $\hat{\theta}(x)$. i.e. look at X/n where X can take all possible values, 0, 1, ... n.

$X \sim$ Binomial(n, θ) with the probability model (1). Noting that the variance of such an X is $n\theta(1-\theta)$, one obtains the variance of X/n to be $\theta(1-\theta)/n$, which can be estimated by $\hat{\theta}(1-\hat{\theta})/n$. A measure of estimation error of $\hat{\theta}$ is the estimated standard deviation of X/n, namely, $\sqrt{\hat{\theta}(1-\hat{\theta})/n}$. For further development we need large n, so that we can apply the *Law of Large Numbers* and the *Central Limit Theorem* to X/n. Then, the estimator will be close to the true θ probabilistically and also, it is approximately distributed like a Gaussian random variable with mean θ and variance $\theta(1-\theta)/n$.

Confidence Statements

Specifically, for large n, approximately

$$\frac{\hat{\theta} - \theta}{\sqrt{\theta(1-\theta)/n}} \sim N(0,1),$$

or

$$\frac{\hat{\theta} - \theta}{\sqrt{\hat{\theta}(1-\hat{\theta})/n}} \sim N(0,1). \tag{2}$$

From (2), an approximate 95% confidence interval for θ (when n is large) is

$$\hat{\theta} \pm 2\sqrt{\hat{\theta}(1-\hat{\theta})/n}.$$

What Does This Mean?

Simply, if we sample again and again, in about 19 cases out of 20 this random interval

$$\left(\hat{\theta}(X) - 2\sqrt{\hat{\theta}(X)(1-\hat{\theta}(X))/n}, \hat{\theta}(X) + 2\sqrt{\hat{\theta}(X)(1-\hat{\theta}(X))/n}\right)$$

will contain the true unknown value of θ.

Fine, but what can we say about the one interval that we can construct for the given sample or data x?

Nothing; either θ is inside $(0.3 - 2\sqrt{0.3 \times 0.7/10}, 0.3 + 2\sqrt{0.3 \times 0.7/10})$ or it is outside.

Can we say $0.3 - 2\sqrt{0.3 \times 0.7/10} \leq \theta \leq 0.3 + 2\sqrt{0.3 \times 0.7/10}$ with 95% chance?

Not in this approach. If θ is treated as fixed unknown constant, conditioning on the given data $X = x$ is meaningless.

3 Conditioning on Data

- What other approach is possible, then?
- How does one condition on data?
- How does one talk about probability of a model or a hypothesis?

Example 3.(not from physics but medicine) Consider a blood test for a certain disease; result is *positive* ($x = 1$) or *negative* ($x = 0$). Suppose θ_1 denotes *disease is present*, θ_2 *disease not present*.

Test is not confirmatory. Instead the probability distribution of X for different θ is:

	$x = 0$	$x = 1$	What does it say?
θ_1	0.2	0.8	Test is +ve 80% of time if 'disease present'
θ_2	0.7	0.3	Test is −ve 70% of time if 'disease not present'

If for a particular patient the test result comes out to be 'positive', what should the doctor conclude?

What is the Question?

What is to be answered is 'what are the chances that the *disease is present* given that the test is positive?' i.e., $P(\theta = \theta_1 | X = 1)$.

What we have is $P(X = 1|\theta = \theta_1)$ and $P(X = 1|\theta = \theta_2)$.

We have the 'wrong' conditional probabilities. They need to be 'reversed'. But how?

4 The Bayesian Recipe

Recall Bayes Theorem: If A and B are two events,

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

assuming $P(B) > 0$. Therefore, $P(A \text{ and } B) = P(A|B)P(B)$, and by symmetry $P(A \text{ and } B) = P(B|A)P(A)$. Consequently, if $P(B|A)$ is given and $P(A|B)$ is desired, note

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}.$$

Rule of total probability says,

$$\begin{aligned} P(B) = P(B \text{ and } \Omega) &= P(B \text{ and } A) + P(B \text{ and } A^c) \\ &= P(B|A)P(A) + P(B|A^c)(1 - P(A)), \text{ so} \end{aligned}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)(1 - P(A))} \tag{3}$$

Bayes Theorem allows one to invert a certain conditional probability to get a certain other conditional probability. How does this help us?

In our example we want $P(\theta = \theta_1|X = 1)$. From (3),

$$\begin{aligned} &P(\theta = \theta_1 \mid X = 1) \\ &= \frac{P(X = 1 \mid \theta = \theta_1)P(\theta = \theta_1)}{P(X = 1 \mid \theta_1)P(\theta = \theta_1) + P(X = 1 \mid \theta_2)P(\theta = \theta_2)}. \end{aligned} \tag{4}$$

So, all we need is $P(\theta = \theta_1)$, which is simply the probability that a randomly chosen person has this disease, or just the 'prevalence' of this disease in the concerned population. The good doctor most likely has this information from his experience in the field. But this is not part of the experimental data. This is pre-experimental information or *prior* information. If we have this, and are willing to incorporate it in the analysis, we get the post-experimental information or *posterior* information in the form of $P(\theta|X = x)$.

In our example, if we take $P(\theta = \theta_1) = 0.05$ or 5%, we get

$$P(\theta = \theta_1 \mid X = 1) = \frac{0.8 \times 0.05}{0.8 \times 0.05 + 0.3 \times 0.95} = \frac{0.04}{0.325} = 0.123$$

which is only 12.3% and $P(\theta = \theta_2 \mid X = 1) = 0.877$ or 87.7%.

Formula (4) which shows how to 'invert' the given conditional probabilities, $P(X = x \mid \theta)$ into the conditional probabilities of interest, $P(\theta \mid X = x)$ is an instance of the Bayes Theorem, and hence the *Theory of Inverse Probability* (usage at the time of Bayes and Laplace, late eighteenth century and even by Jeffreys), is known these days as *Bayesian inference*.

Ingredients of Bayesian inference:

likelihood function, $l(\theta|x)$; θ can be a parameter vector

prior probability, $\pi(\theta)$

Combining the two, one gets the posterior probability density or mass function

$$\pi(\theta \mid x) = \begin{cases} \frac{\pi(\theta)l(\theta|x)}{\sum_j \pi(\theta_j)l(\theta_j|x)} & \text{if } \theta \text{ is discrete;} \\ \frac{\pi(\theta)l(\theta|x)}{\int \pi(u)l(u|x)\,du} & \text{if } \theta \text{ is continuous.} \end{cases} \tag{5}$$

5 Inference for Binomial proportion

Example 2 contd. Suppose we have no special information available on θ. Then assume θ is uniformly distributed on the interval $(0, 1)$. i.e., the prior density is $\pi(\theta) = 1, \quad 0 < \theta < 1$.

This is a choice of *non-informative* or *vague* or *reference* prior. Often, Bayesian inference from such a prior coincides with classical inference.

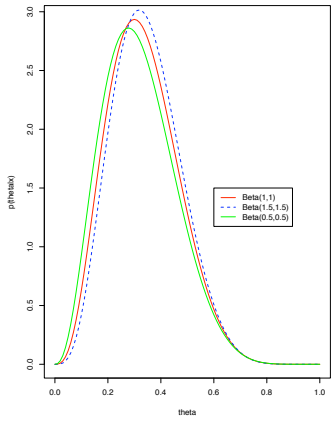
In the Example then the posterior density of θ given x is

$$\begin{aligned}\pi(\theta|x) &= \frac{\pi(\theta)l(\theta|x)}{\int \pi(u)l(u|x)\,du} \\ &= \frac{(n+1)!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}, \quad 0<\theta<1.\end{aligned}$$

As a function of θ, this is the same as the likelihood function $l(\theta|x) \propto \theta^x(1-\theta)^{n-x}$, and so maximizing the posterior probability density will give the same estimate as the maximum likelihood estimate!

Influence of the Prior

If we had some knowledge about θ which can be summarized in the form of a Beta prior distribution with parameters α and γ, the posterior will also be Beta with parameters $x+\alpha$ and $n-x+\gamma$. Such priors which result in posteriors from the same 'family' are called 'natural conjugate priors'. Robustness?



Objective Bayesian Analysis:

Invariant priors: Jeffreys

Reference priors: Bernardo, Jeffreys

Maximum entropy priors: Jaynes

In Example 2, what $\pi(\theta|x)$ says is that the uncertainty in θ can now be described in terms of an actual probability distribution concentrated around the maximum likelihood estimate $\hat{\theta} = x/n$. However, the interpretation of $\hat{\theta}$ as an estimate of θ is quite different. It is the most probable value of the unknown parameter θ conditional on the sample data x; it is called the 'maximum a posteriori estimate (MAP)' or the 'highest posterior density estimate (HPD)'.

There is no need to mimic the MLE anymore. We have a genuine probability distribution, namely, the posterior distribution to quantify our post-experimental knowledge about θ. Indeed the usual Bayes estimate is the mean of the posterior distribution which minimizes the posterior dispersion:

$$E[(\theta - \hat{\theta}_B)^2|x] = \min_a E[(\theta - a)^2|x],$$

when $\hat{\theta}_B = E(\theta|x)$.

If we choose $\hat{\theta}_B$ as the estimate of θ, we get a natural measure of variability of this estimate in the form of the posterior variance: $E[(\theta - E(\theta|x))^2|x]$. Therefore the posterior standard deviation is a natural measure of estimation error. i.e., our estimate is $\hat{\theta}_B \pm \sqrt{E[(\theta - E(\theta|x))^2|x]}$.

In fact, we can say much more. For any interval around $\hat{\theta}$ we can compute the (posterior) probability of it containing the true parameter θ. In other words, a statement such as

$$P(\hat{\theta}_B - k_1 \leq \theta \leq \hat{\theta}_B + k_2|x) = 0.95$$

is perfectly meaningful.

All these inferences are conditional on the given data.

In Example 2, if the prior is a Beta distribution with parameters α and γ, then $\theta|x$ will have a Beta$(x+\alpha, n-x+\gamma)$ distribution, so the Bayes estimate of θ will be

$$\hat{\theta}_B = \frac{(x+\alpha)}{(n+\alpha+\gamma)} = \frac{n}{n+\alpha+\gamma}\frac{x}{n} + \frac{\alpha+\gamma}{n+\alpha+\gamma}\frac{\alpha}{\alpha+\gamma}.$$

This is a convex combination of sample mean and prior mean, with the weights depending upon the sample size and the strength of the prior information as measured by the values of α and γ.

Bayesian inference relies on the conditional probability language to revise one's knowledge. In the above example, prior to the collection of sample data one had some (vague, perhaps) information on θ. Then came the sample data. Combining the model density of this data with the prior density one gets the posterior density, the conditional density of θ given the data. From now on until further data is available, this posterior distribution of θ is the only relevant information as far as θ is concerned.

6 Inference With Normals/Gaussians

Gaussian PDF:

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{over } [-\infty,\infty] \tag{6}$$

Common abbreviated notation: $X \sim N(\mu,\sigma^2)$

Parameters:

$$\begin{aligned} \mu &= E(X) \equiv \langle X \rangle \equiv \int x\, f(x|\mu,\sigma^2)\, dx \\ \sigma^2 &= E(X-\mu)^2 \equiv \langle (X-\mu)^2 \rangle \equiv \int (x-\mu)^2\, f(x|\mu,\sigma^2)\, dx \end{aligned}$$

Inference About a Normal Mean

Example 4. Fit a normal/Gaussian model to the 'globular cluster luminosity functions' data. The set-up is as follows.

Our data consist of n measurements, $X_i = \mu + \epsilon_i$.
Suppose the noise contributions are independent, and $\epsilon_i \sim N(0,\sigma^2)$. Denoting by $\mathbf{x}$, the random sample $(x_1, \ldots x_n)$,

$$\begin{aligned} f(\mathbf{x}|\mu,\sigma^2) &= \prod_i f(x_i|\mu,\sigma^2) \\ &= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \\ &= (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2} \\ &= (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n (x_i-\bar{x})^2 + n(\bar{x}-\mu)^2\right]}. \end{aligned}$$

Note $(\bar{X}, s^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1))$ is *sufficient* for the parameters (μ, σ^2). This is a very substantial data compression.

Inference About a Normal Mean, σ^2 known

(Not useful, but easy to understand.)

$$l(\mu|\mathbf{x}) \propto f(\mathbf{x}|\mu, \sigma^2) \propto e^{-\frac{n}{2\sigma^2}(\mu-\bar{x})^2},$$

so that $\bar{X}$ is sufficient. Also, $\bar{X}|\mu \sim N(\mu, \sigma^2/n)$. If an informative prior, $\mu \sim N(\mu_0, \tau^2)$ is chosen for μ,

$$\begin{aligned}
\pi(\mu|\mathbf{x}) &\propto l(\mu|\mathbf{x})\pi(\mu) \\
&\propto e^{-\frac{1}{2}\left[\frac{n(\mu-\bar{x})^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{\tau^2}\right]} \\
&\propto e^{-\frac{\tau^2+\sigma^2/n}{2\tau^2\sigma^2/n}\left(\mu - \frac{\tau^2\sigma^2/n}{\tau^2+\sigma^2/n}(\frac{\mu_0}{\tau^2} + \frac{n\bar{x}}{\sigma^2})\right)^2}.
\end{aligned}$$

i.e., $\mu|\mathbf{x} \sim N(\hat{\mu}, \delta^2)$:

$$\begin{aligned}
\hat{\mu} &= \frac{\tau^2\sigma^2/n}{\tau^2+\sigma^2/n}(\frac{\mu_0}{\tau^2} + \frac{n\bar{x}}{\sigma^2}) \\
&= \frac{\tau^2}{\tau^2+\sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\tau^2+\sigma^2/n}\mu_0.
\end{aligned}$$

$\hat{\mu}$ is the Bayes estimate of μ, which is just a weighted average of sample mean $\bar{x}$ and prior mean μ_0.

δ^2 is the posterior variance of μ and

$$\delta^2 = \frac{\tau^2\sigma^2/n}{\tau^2+\sigma^2/n} = \frac{\sigma^2}{n}\frac{\tau^2}{\tau^2+\sigma^2/n}.$$

Therefore $\hat{\mu} \pm \delta$ is our estimate for μ and $\hat{\mu} \pm 2\delta$ is a 95% HPD (Bayesian) credible interval for μ.

What happens as $\tau^2 \to \infty$, or as the prior becomes more and more flat?

$$\hat{\mu} \to \bar{x}, \quad \delta \to \frac{\sigma}{\sqrt{n}}$$

i.e., Jeffreys' prior $\pi(\mu) = C$ reproduces frequentist inference.

Inference About a Normal Mean, σ^2 unknown

Our observations $X_1, \ldots X_n$ is a random sample from a Gaussian population with both mean μ and variance σ^2 unknown.
We are only interested in μ.
How do we get rid of the nuisance parameter σ^2?
Bayesian inference uses posterior distribution which is a probability distribution, so σ^2 should be integrated out from the joint posterior distribution of μ and σ^2.

$$l(\mu, \sigma^2|\mathbf{x}) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n (x_i-\bar{x})^2 + n(\mu-\bar{x})^2\right]}.$$

Start with $\pi(\mu, \sigma^2)$ and get

$$\pi(\mu, \sigma^2|\mathbf{x}) \propto \pi(\mu, \sigma^2) l(\mu, \sigma^2|\mathbf{x})$$

and then get

$$\pi(\mu|\mathbf{x}) = \int_0^\infty \pi(\mu, \sigma^2|\mathbf{x})\, d\sigma^2.$$

Use Jeffreys' prior $\pi(\mu, \sigma^2) \propto 1/\sigma^2$: Flat prior for μ which is a location or translation parameter, and an independent flat prior for $\log(\sigma)$ which is again a location parameter, being the log of a scale parameter.

$$\pi(\mu, \sigma^2|\mathbf{x}) \propto \frac{1}{\sigma^2} l(\mu, \sigma^2|\mathbf{x})$$

$$\begin{aligned}
\pi(\mu|\mathbf{x}) &\propto \int_0^\infty (\sigma^2)^{-(n+1)/2} e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n (x_i-\bar{x})^2 + n(\mu-\bar{x})^2\right]} d\sigma^2 \\
&\propto \left[(n-1)s^2 + n(\mu-\bar{x})^2\right]^{-n/2} \\
&\propto \left[1 + \frac{1}{n-1}\frac{n(\mu-\bar{x})^2}{s^2}\right]^{-n/2} \\
&\propto \text{density of Students } t_{n-1}.
\end{aligned}$$

$$\frac{\sqrt{n}(\mu-\bar{x})}{s} | \text{ data} \sim t_{n-1}$$

$$P(\bar{x} - t_{n-1}(0.975)\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{n-1}(0.975)\frac{s}{\sqrt{n}} \mid \text{ data}) = 95\%$$

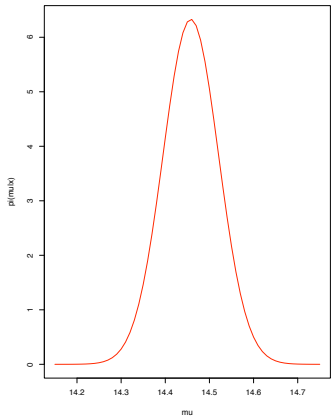
i.e., the Jeffreys' translation-scale invariant prior reproduces frequentist inference.

What if there are some constraints on μ such as $-A \leq \mu \leq B$, for example, $\mu > 0$? We will get a truncated t_{n-1} instead, but the procedure will go through with minimal change.

Example 4 contd. (GCL Data) $n = 360$, $\bar{x} = 14.46$, $s = 1.19$.

$$\frac{\sqrt{360}(\mu - 14.46)}{1.19} | \text{ data} \sim t_{359}$$

$$\mu | \text{ data} \sim N(14.46, 0.063^2) \text{ approximately.}$$



Estimate for mean GCL is 14.46 ± 0.063 and 95% HPD credible interval is $(14.33, 14.59)$.

Comparing two Normal Means

Example 5. Check whether the mean distance indicators in the two populations of LMC datasets are different. Model as follows:

$X_1, \ldots X_{n_1}$ is a random sample from $N(\mu_1, \sigma_1^2)$.
$Y_1, \ldots Y_{n_2}$ is a random sample from $N(\mu_2, \sigma_2^2)$.
Samples are independent.
Unknown parameters: $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$
Quantity of interest: $\eta = \mu_1 - \mu_2$
Nuisance parameters: σ_1^2 and σ_2^2

Case 1. $\sigma_1^2 = \sigma_2^2$. Then sufficient statistic for (μ_1, μ_2, σ^2) is $\left(\bar{X}, \bar{Y}, s^2 = \frac{1}{n_1+n_2-2}\left(\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{j=1}^{n_2}(Y_j - \bar{Y})^2\right)\right)$

$\bar{X}|\mu_1, \mu_2, \sigma^2 \sim N(\mu_1, \sigma^2/n_1)$, $\bar{Y}|\mu_1, \mu_2, \sigma^2 \sim N(\mu_2, \sigma^2/n_2)$, $(n_1+n_2-2)s^2|\mu_1, \mu_2, \sigma^2 \sim \sigma^2\chi^2_{n_1+n_2-2}$.

These three are independently distributed.

$\bar{X} - \bar{Y}|\mu_1, \mu_2, \sigma^2 \sim N(\eta, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2}))$, $\eta = \mu_1 - \mu_2$
Use Jeffreys' location-scale invariant prior $\pi(\mu_1, \mu_2, \sigma^2) \propto 1/\sigma^2$

$$\eta|\sigma^2, \mathbf{x}, \mathbf{y} \sim N(\bar{x} - \bar{y}, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})), \text{ and}$$
$$\pi(\eta, \sigma^2|\mathbf{x}, \mathbf{y}) \propto \pi(\eta|\sigma^2, \mathbf{x}, \mathbf{y})\pi(\sigma^2|s^2), \tag{7}$$

Integrate out σ^2 from (7) as in the previous example to get

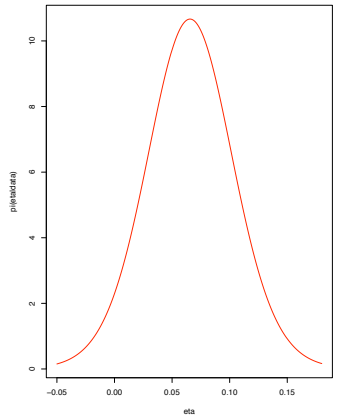
$$\frac{\eta - (\bar{x} - \bar{y})}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \mid \mathbf{x}, \mathbf{y} \sim t_{n_1+n_2-2}.$$

95% HPD credible interval for $\eta = \mu_1 - \mu_2$ is

$$\bar{x} - \bar{y} \pm t_{n_1+n_2-2}(0.975)s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

same as frequentist t-interval.

Example 5 contd. We have $\bar{x} = 18.539$, $\bar{y} = 18.473$, $n_1 = 13$, $n_2 = 12$ and $s^2 = 0.0085$. $\hat{\eta} = \bar{x} - \bar{y} = 0.066$, $s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 0.037$, $t_{23}(0.975) = 2.069$.
95% HPD credible interval for $\eta = \mu_1 - \mu_2$: $(0.066 - 2.069 \times 0.037, 0.066 + 2.069 \times 0.037) = (-0.011, 0.142)$.



Case 2. σ_1^2 and σ_2^2 are not known to be equal.

From the one-sample normal example, note that
$(\bar{X}, s_X^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2)$ sufficient for (μ_1, σ_1^2), and
$(\bar{Y}, s_Y^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_j - \bar{Y})^2)$ sufficient for (μ_2, σ_2^2).

Making inference on $\eta = \mu_1 - \mu_2$ when σ_1^2 and σ_2^2 are not assumed to be equal is called the Behrens-Fisher problem for which the frequentist solution is not very straight forward, but the Bayes solution is.

$\bar{X}|\mu_1, \sigma_1^2 \sim N(\mu_1, \sigma_1^2/n_1)$, $(n_1 - 1)s_X^2|\mu_1, \sigma_1^2 \sim \sigma^2\chi^2_{n_1-1}$, and are independently distributed.

$\bar{Y}|\mu_2, \sigma_2^2 \sim N(\mu_2, \sigma_2^2/n_2)$, $(n_2 - 1)s_Y^2|\mu_2, \sigma_2^2 \sim \sigma^2\chi^2_{n_2-1}$, and are independently distributed.

X and **Y** samples are independent.

Use Jeffreys' prior $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \propto 1/\sigma_1^2 \times 1/\sigma_2^2$

Calculations similar to those in one-sample case give:

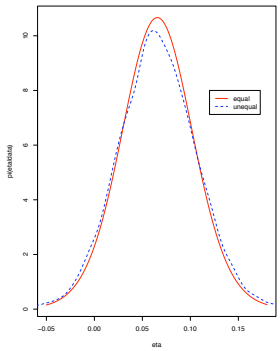
$$\begin{aligned} \frac{\sqrt{n_1}(\mu_1 - \bar{x})}{s_X} | \text{ data} &\sim t_{n_1-1}, \\ \frac{\sqrt{n_2}(\mu_2 - \bar{y})}{s_Y} | \text{ data} &\sim t_{n_2-1}, \end{aligned} \tag{8}$$

and these two are independent.

Posterior distribution of $\eta = \mu_1 - \mu_2$ given the data is non-standard (differ-

ence of two independent t variables) but not difficult to get.
Use Monte-Carlo Sampling: Simply generate (μ_1, μ_2) repeatedly from (8) and construct a histogram for $\eta = \mu_1 - \mu_2$

Example 5 (LMC) contd. Looks slightly different.



Posterior mean of $\eta = \mu_1 - \mu_2$ is

$$\hat{\eta} = E(\mu_1 - \mu_2 | \text{ data}) = \begin{cases} 0.0656 & \text{equal variance;} \\ 0.0657 & \text{unequal variance.} \end{cases} \tag{9}$$

95% HPD credible interval for $\eta = \mu_1 - \mu_2$ is

$$= \begin{cases} (-0.011, 0.142) & \text{equal variance;} \\ (-0.014, 0.147) & \text{unequal variance.} \end{cases} \tag{10}$$

7 Bayesian Computations

Bayesian analysis requires computation of expectations and quantiles of probability distributions (posterior distributions). Most often posterior distributions will not be standard distributions. Then posterior quantities of inferential interest cannot be computed in closed form. Special techniques are needed.

Example M1. Suppose $X_1, X_2, \ldots, X_k$ are observed number of certain type of stars in k similar regions. Model them as independent Poisson counts:

$X_i \sim \text{Poisson}(\theta_i)$. θ_i are *a priori* considered related. $\nu_i = \log(\theta_i)$ is the ith element of $\boldsymbol{\nu}$ and suppose

$$\boldsymbol{\nu} \sim N_k\left(\mu\mathbf{1}, \tau^2\left\{(1-\rho)I_k + \rho\mathbf{1}\mathbf{1}'\right\}\right),$$

where $\mathbf{1}$ is the k-vector with all elements being 1, and μ, τ^2 and ρ are known constants. Then

$$f(\boldsymbol{x}|\boldsymbol{\nu}) = \exp\left(-\sum_{i=1}^{k}\{e^{\nu_i} - \nu_i x_i\}\right) / \prod_{i=1}^{k} x_i!.$$

$$\pi(\boldsymbol{\nu}) \propto \exp\left(-\frac{1}{2\tau^2}(\boldsymbol{\nu} - \mu\mathbf{1})'\left((1-\rho)I_k + \rho\mathbf{1}\mathbf{1}'\right)^{-1}(\boldsymbol{\nu} - \mu\mathbf{1})\right)$$

$\pi(\boldsymbol{\nu}|\boldsymbol{x}) \propto \exp\left\{-\sum_{i=1}^{k}\{e^{\nu_i} - \nu_i x_i\} - \frac{(\boldsymbol{\nu}-\mu\mathbf{1})'((1-\rho)I_k+\rho\mathbf{1}\mathbf{1}')^{-1}(\boldsymbol{\nu}-\mu\mathbf{1})}{2\tau^2}\right\}$.

To obtain the posterior mean of θ_j, compute

$$E^{\pi}(\theta_j|x) = E^{\pi}(\exp(\nu_j)|x) = \frac{\int_{\mathcal{R}^k}\exp(\nu_j)g(\boldsymbol{\nu}|\boldsymbol{x})\,d\boldsymbol{\nu}}{\int_{\mathcal{R}^k} g(\boldsymbol{\nu}|\boldsymbol{x})\,d\boldsymbol{\nu}},$$

where $g(\boldsymbol{\nu}|\boldsymbol{x}) = \exp\left\{-\sum_{i=1}^{k}\{e^{\nu_i} - \nu_i x_i\} - \frac{(\boldsymbol{\nu}-\mu\mathbf{1})'((1-\rho)I_k+\rho\mathbf{1}\mathbf{1}')^{-1}(\boldsymbol{\nu}-\mu\mathbf{1})}{2\tau^2}\right\}$.

This is a ratio of two k-dimensional integrals, and as k grows, the integrals become less and less easy to work with. Numerical integration techniques fail to be an efficient technique in this case. This problem, known as the *curse of dimensionality*, is due to the fact that the size of the part of the space that is not relevant for the computation of the integral grows very fast with the dimension. Consequently, the error in approximation associated with this numerical method increases as the power of the dimension k, making the technique inefficient.

The recent popularity of Bayesian approach to statistical applications is mainly due to advances in statistical computing. These include the E-M algorithm and the Markov chain Monte Carlo (MCMC) sampling techniques.

8 Monte Carlo Sampling

Consider an expectation that is not available in closed form. To estimate a population mean, gather a large sample from this population and consider

the corresponding sample mean. The Law of Large Numbers guarantees that the estimate will be *good* provided the sample is large enough. Specifically, let f be a probability density function (or a mass function) and suppose the quantity of interest is a finite expectation of the form

$$E_f h(\boldsymbol{X}) = \int_{\mathcal{X}} h(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x} \tag{11}$$

(or the corresponding sum in the discrete case). If i.i.d. observations $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ can be generated from the density f, then

$$\bar{h}_m = \frac{1}{m} \sum_{i=1}^{m} h(\boldsymbol{X}_i) \tag{12}$$

converges in probability to $E_f h(\boldsymbol{X})$. This justifies using $\bar{h}_m$ as an approximation for $E_f h(\boldsymbol{X})$ for large m.

To provide a measure of accuracy or the extent of error in the approximation, compute the standard error. If $\mathrm{Var}_f h(\boldsymbol{X})$ is finite, then $\mathrm{Var}_f(\bar{h}_m) = \mathrm{Var}_f h(\boldsymbol{X})/m$. Further, $\mathrm{Var}_f h(\boldsymbol{X}) = E_f h^2(\boldsymbol{X}) - (E_f h(\boldsymbol{X}))^2$ can be estimated by

$$s_m^2 = \frac{1}{m} \sum_{i=1}^{m} (h(\boldsymbol{X}_i) - \bar{h}_m)^2,$$

and hence the standard error of $\bar{h}_m$ can be estimated by

$$\frac{1}{\sqrt{m}} s_m = \frac{1}{m} \Big(\sum_{i=1}^{m} (h(\boldsymbol{X}_i) - \bar{h}_m)^2\Big)^{1/2}.$$

Confidence intervals for $E_f h(\boldsymbol{X})$: Using CLT

$$\frac{\sqrt{m}\,\big(\bar{h}_m - E_f h(\boldsymbol{X})\big)}{s_m} \underset{m\to\infty}{\longrightarrow} N(0,1), \text{ so}$$

$(\bar{h}_m - z_{\alpha/2} s_m/\sqrt{m}, \bar{h}_m + z_{\alpha/2} s_m/\sqrt{m})$ can be used as an approximate $100(1-\alpha)\%$ confidence interval for $E_f h(\boldsymbol{X})$, with $z_{\alpha/2}$ denoting the $100(1-\alpha/2)\%$ quantile of standard normal.

If we want to approximate the posterior mean, try to generate i.i.d. observations from the posterior distribution and consider the mean of this sample. This is rarely useful because most often the posterior distribution will be a non-standard distribution which may not easily allow sampling from it. What are some other possibilities?

Example M2. Suppose X is $N(\theta, \sigma^2)$ with known σ^2 and a Cauchy(μ, τ) prior on θ is considered appropriate. Then

$$\pi(\theta|x) \propto \exp\left(-(\theta - x)^2/(2\sigma^2)\right)\left(\tau^2 + (\theta - \mu)^2\right)^{-1},$$

and hence the posterior mean is

$$\begin{aligned} E^\pi(\theta|x) &= \frac{\int_{-\infty}^{\infty} \theta \exp\left(-\frac{(\theta-x)^2}{2\sigma^2}\right)\left(\tau^2 + (\theta-\mu)^2\right)^{-1} d\theta}{\int_{-\infty}^{\infty} \exp\left(-\frac{(\theta-x)^2}{2\sigma^2}\right)\left(\tau^2 + (\theta-\mu)^2\right)^{-1} d\theta} \\ &= \frac{\int_{-\infty}^{\infty} \theta \left\{\frac{1}{\sigma}\phi\left(\frac{\theta-x}{\sigma}\right)\right\}\left(\tau^2 + (\theta-\mu)^2\right)^{-1} d\theta}{\int_{-\infty}^{\infty} \left\{\frac{1}{\sigma}\phi\left(\frac{\theta-x}{\sigma}\right)\right\}\left(\tau^2 + (\theta-\mu)^2\right)^{-1} d\theta}, \end{aligned}$$

where ϕ denotes the density of standard normal.

$E^\pi(\theta|x)$ is the ratio of expectation of $h(\theta) = \theta/(\tau^2 + (\theta-\mu)^2)$ to that of $h(\theta) = 1/(\tau^2 + (\theta-\mu)^2)$, both expectations being with respect to the $N(x, \sigma^2)$ distribution. Therefore, we simply sample $\theta_1, \theta_2, \ldots$ from $N(x, \sigma^2)$ and use

$$\widehat{E^\pi(\theta|x)} = \frac{\sum_{i=1}^m \theta_i \left(\tau^2 + (\theta_i - \mu)^2\right)^{-1}}{\sum_{i=1}^m \left(\tau^2 + (\theta_i - \mu)^2\right)^{-1}}$$

as our Monte Carlo estimate of $E^\pi(\theta|x)$. Note that (11) and (12) are applied separately to both the numerator and denominator, but using the same sample of θ's. It is unwise to assume that the problem has been completely solved. The sample of θ's generated from $N(x, \sigma^2)$ will tend to concentrate around x, whereas to satisfactorily account for the contribution of the Cauchy prior to the posterior mean, a significant portion of the θ's should come from the tails of the posterior distribution.

Why not express the posterior mean in the form

$$E^\pi(\theta|x) = \frac{\int_{-\infty}^{\infty} \theta \exp\left(-\frac{(\theta-x)^2}{2\sigma^2}\right)\pi(\theta)\, d\theta}{\int_{-\infty}^{\infty} \exp\left(-\frac{(\theta-x)^2}{2\sigma^2}\right)\pi(\theta)\, d\theta},$$

and then sample θ's from Cauchy(μ, τ) and use the approximation

$$\widehat{E^\pi(\theta|x)} = \frac{\sum_{i=1}^m \theta_i \exp\left(-\frac{(\theta_i-x)^2}{2\sigma^2}\right)}{\sum_{i=1}^m \exp\left(-\frac{(\theta_i-x)^2}{2\sigma^2}\right)}?$$

However, this is also not satisfactory because the tails of the posterior distribution are not as heavy as those of the Cauchy prior, and there will be excess sampling from the tails relative to the center. So the convergence of the approximation will be slower resulting in a larger error in approximation (for a fixed m). Ideally, therefore, sampling should be from the posterior distribution itself. With this view in mind, a variation of the above theme, called Monte Carlo importance sampling has been developed.

Consider (11) again. Suppose that it is difficult or expensive to sample directly from f, but there exists a probability density u that is very *close* to f from which it is easy to sample. Then we can rewrite (11) as

$$\begin{aligned} E_f h(\boldsymbol{X}) &= \int_{\mathcal{X}} h(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x} = \int_{\mathcal{X}} h(\boldsymbol{x}) \frac{f(\boldsymbol{x})}{u(\boldsymbol{x})} u(\boldsymbol{x})\, d\boldsymbol{x} \\ &= \int_{\mathcal{X}} \{h(\boldsymbol{x}) w(\boldsymbol{x})\}\, u(\boldsymbol{x})\, d\boldsymbol{x} = E_u \{h(\boldsymbol{X}) w(\boldsymbol{X})\}, \end{aligned}$$

where $w(\boldsymbol{x}) = f(\boldsymbol{x})/u(\boldsymbol{x})$. Now apply (12) with f replaced by u and h replaced by hw. In other words, generate i.i.d. observations $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ from the density u and compute

$$\overline{hw}_m = \frac{1}{m} \sum_{i=1}^{m} h(\boldsymbol{X}_i) w(\boldsymbol{X}_i).$$

The sampling density u is called the *importance* function. [1]

9 Markov Chain Monte Carlo Methods

A severe drawback of the standard Monte Carlo sampling/ importance sampling: complete determination of the functional form of the posterior density is needed for implementation.

Situations where posterior distributions are incompletely specified or are specified indirectly cannot be handled: joint posterior distribution of the vector of parameters is specified in terms of several conditional and marginal distributions, but not directly.

This covers a large range of Bayesian analysis because a lot of Bayesian modeling is hierarchical so that the joint posterior is difficult to calculate but

[1] Rest of these notes was not covered in the lectures and may be omitted at first reading.

the conditional posteriors given parameters at different levels of hierarchy are easier to write down (and hence sample from).

Markov Chains. A sequence of random variables $\{X_n\}_{n\geq 0}$ is a *Markov chain* if for any n, given the current value, X_n, the *past* $\{X_j, j \leq n-1\}$ and the *future* $\{X_j : j \geq n+1\}$ are *independent*. In other words,

$$P(A \cap B|X_n) = P(A|X_n)P(B|X_n), \tag{13}$$

where A and B are events defined respectively in terms of the past and the future.

Important subclass: Markov chains with time homogeneous or *stationary transition probabilities*: the probability distribution of X_{n+1} given $X_n = x$, and the past, $X_j : j \leq n-1$ depends only on x and does not depend on the values of $X_j : j \leq n-1$ or n.

If the set S of values $\{X_n\}$ can take, known as the *state space*, is countable, this reduces to specifying the transition probability matrix $P \equiv ((p_{ij}))$ where for any two values i, j in S, p_{ij} is the probability that $X_{n+1} = j$ given $X_n = i$, i.e., of moving from state i to state j in one time unit.

For state space S that is not countable, specify a *transition kernel* or *transition function* $P(x, \cdot)$ where $P(x, A)$ is the probability of moving from x into A in one step, i.e., $P(X_{n+1} \in A|X_n = x)$.

Given the transition probability and the probability distribution of the initial value X_0, one can construct the joint probability distribution of $\{X_j : 0 \leq j \leq n\}$ for any finite n. i.e.,

$$\begin{aligned}
&P(X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i_n) \\
&= P(X_n = i_n|X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) \\
&\quad \times P(X_0 = i_0, X_1 = i_1, \ldots X_{n-1} = i_{n-1}) \\
&= p_{i_{n-1}i_n} P(X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) \\
&= P(X_0 = i_0) p_{i_0 i_1} p_{i_1 i_2} \ldots p_{i_{n-1} i_n}.
\end{aligned}$$

A probability distribution π is called *stationary* or *invariant* for a transition probability P or the associated Markov chain $\{X_n\}$ if it is the case that when the probability distribution of X_0 is π then the same is true for X_n for all $n \geq 1$. Thus in the countable state space case a probability distribution $\pi = \{\pi_i : i \in S\}$ is stationary for a transition probability matrix P if for

each j in S,

$$\begin{aligned} P(X_1 = j) &= \sum_i P(X_1 = j | X_0 = i) P(X_0 = i) \\ &= \sum_i \pi_i p_{ij} = P(X_0 = j) = \pi_j. \end{aligned} \tag{14}$$

In vector notation it says $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots)$ is a left eigenvector of the matrix P with eigenvalue 1 and

$$\boldsymbol{\pi} = \boldsymbol{\pi} P. \tag{15}$$

Similarly, if S is a continuum, a probability distribution π with density $p(x)$ is *stationary* for the transition kernel $P(\cdot, \cdot)$ if

$$\pi(A) = \int_A p(x)\, dx = \int_S P(x, A) p(x)\, dx$$

for all $A \subset S$.

A Markov chain $\{X_n\}$ with a countable state space S and transition probability matrix $P \equiv ((p_{ij}))$ is said to be *irreducible* if for any two states i and j the probability of the Markov chain visiting j starting from i is positive, i.e., for some $n \geq 1, p_{ij}^{(n)} \equiv P(X_n = j | X_0 = i) > 0$.

A similar notion of *irreducibility*, known as Harris or Doeblin irreducibility exists for the general state space case also.

Theorem (Law of Large Lumbers for Markov Chains). $\{X_n\}_{n \geq 0}$ is a Markov chain with a countable state space S and a transition probability matrix P. Suppose it is *irreducible* and has a stationary probability distribution $\pi \equiv (\pi_i : i \in S)$ as defined in (14). Then, for any bounded function $h : S \to R$ and for any initial distribution of X_0

$$\frac{1}{n} \sum_{i=0}^{n-1} h(X_i) \to \sum_j h(j) \pi_j \tag{16}$$

in probability as $n \to \infty$.

A similar law of large numbers (LLN) holds when the state space S is not countable. The limit value in (16) will be the integral of h with respect to the stationary distribution π. A sufficient condition for the validity of this LLN

is that the Markov chain $\{X_n\}$ be Harris irreducible and have a stationary distribution π.

How is this Useful?

A probability distribution π on a set S is given. Want to compute the "integral of h with respect to π", which reduces to $\sum_j h(j)\pi_j$ in the countable case.

Look for an irreducible Markov chain $\{X_n\}$ with state space S and stationary distribution π. Starting from some initial value X_0, run the Markov chain $\{X_j\}$ for a period of time, say $0, 1, 2, \ldots n-1$ and consider as an estimate

$$\mu_n = \frac{1}{n}\sum_{0}^{n-1} h(X_j). \tag{17}$$

By the LLN (16), μ_n will be close to $\sum_j h(j)\pi_j$ for large n.

This technique is called *Markov chain Monte Carlo* (MCMC).

To approximate $\pi(A) \equiv \sum_{j\in A}\pi_j$ for some $A \subset S$ simply consider

$$\pi_n(A) \equiv \frac{1}{n}\sum_{0}^{n-1} I_A(X_j) \to \pi(A),$$

where $I_A(X_j) = 1$ if $X_j \in A$ and 0 otherwise.

An irreducible Markov chain $\{X_n\}$ with a countable state space S is called *aperiodic* if for some $i \in S$ the greatest common divisor, g.c.d. $\{n : p_{ii}^{(n)} > 0\} = 1$. Then, in addition to the LLN (16), the following result on the convergence of $P(X_n = j)$ holds.

$$\sum_j |P(X_n = j) - \pi_j| \to 0 \tag{18}$$

as $n \to \infty$, for any initial distribution of X_0. In other words, for large n the probability distribution of X_n will be close to π. There exists a result similar to (18) for the general state space case also.

This suggests that instead of doing one run of length n, one could do N independent runs each of length m so that $n = Nm$ and then from the i^{th}

run use only the m^{th} observation, say, $X_{m,i}$ and consider the estimate

$$\tilde{\mu}_{N,m} \equiv \frac{1}{N} \sum_{i=1}^{N} h(X_{m,i}). \tag{19}$$

Metropolis-Hastings Algorithm

Very general MCMC method with wide applications. Idea is not to directly simulate from the given target density (which may be computationally difficult), but to simulate an easy Markov chain that has this target density as the stationary distribution.

Let π be the target probability distribution on S, a finite or countable set. Let $Q \equiv ((q_{ij}))$ be a transition probability matrix such that for each i, it is computationally easy to generate a sample from the distribution $\{q_{ij} : j \in S\}$. Generate a Markov chain $\{X_n\}$ as follows. If $X_n = i$, first sample from the distribution $\{q_{ij} : j \in S\}$ and denote that observation Y_n. Then, choose X_{n+1} from the two values X_n and Y_n according to

$$P(X_{n+1} = Y_n | X_n, Y_n) = \rho(X_n, Y_n) = 1 - P(X_{n+1} = X_n | X_n, Y_n),$$

where the "acceptance probability" $\rho(\cdot, \cdot)$ is given by

$$\rho(i,j) = \min \left\{ \frac{\pi_j}{\pi_i} \frac{q_{ji}}{q_{ij}}, 1 \right\} \text{ for all } (i,j) \text{ such that } \pi_i q_{ij} > 0.$$

$\{X_n\}$ is a Markov chain with transition probability matrix $P = ((p_{ij}))$ given by

$$p_{ij} = \begin{cases} q_{ij}\rho_{ij} & j \neq i, \\ 1 - \sum\limits_{k \neq i} p_{ik}, & j = i. \end{cases} \tag{20}$$

Q is called the "proposal transition probability" and ρ the "acceptance probability". A significant feature of this transition mechanism P is that P and π satisfy

$$\pi_i p_{ij} = \pi_j p_{ji} \quad \text{for all } i, j. \tag{21}$$

This implies that for any j

$$\sum_i \pi_i p_{ij} = \pi_j \sum_i p_{ji} = \pi_j, \tag{22}$$

or, π is a stationary probability distribution for P.

Suppose S is irreducible with respect to Q and $\pi_i > 0$ for all i in S. It can then be shown that P is irreducible, and because it has a stationary distribution π, LLN (16) is available. This algorithm is thus a very flexible and useful one. The choice of Q is subject only to the condition that S is irreducible with respect to Q. A sufficient condition for the aperiodicity of P is that $p_{ii} > 0$ for some i or equivalently

$$\sum_{j \neq 1} q_{ij}\rho_{ij} < 1.$$

A sufficient condition for this is that there exists a pair (i, j) such that $\pi_i q_{ij} > 0$ and $\pi_j q_{ji} < \pi_i q_{ij}$.

Recall that if P is aperiodic, then both the LLN (16) and (18) hold.

If S is not finite or countable but is a continuum and the target distribution $\pi(\cdot)$ has a density $p(\cdot)$, then one proceeds as follows: Let Q be a transition function such that for each x, $Q(x, \cdot)$ has a density $q(x, y)$. Then proceed as in the discrete case but set the "acceptance probability" $\rho(x, y)$ to be

$$\rho(x, y) = \min\left\{\frac{p(y)q(y, x)}{p(x)q(x, y)}, 1\right\}$$

for all (x, y) such that $p(x)q(x, y) > 0$.

A particularly useful feature of the above algorithm is that it is enough to know $p(\cdot)$ upto a multiplicative constant as the "acceptance probability" $\rho(\cdot, \cdot)$ needs only the ratios $p(y)/p(x)$ or π_i/π_j.

This assures us that in Bayesian applications it is not necessary to have the normalizing constant of the posterior density available for computation of the posterior quantities of interest.

Gibbs Sampling

Most of the new problems that Bayesians are asked to solve are high-dimensional: e.g. micro-arrays, image processing. Bayesian analysis of such problems involve target (posterior) distributions that are high-dimensional multivariate distributions.

In image processing, typically one has $N \times N$ square grid of pixels with $N = 256$ and each pixel has $k \geq 2$ possible values. Each configuration has

$(256)^2$ components and the state space S has $k^{(256)^2}$ configurations. How does one simulate a random configuration from a target distribution over such a large S?

Gibbs sampler is a technique especially suitable for generating an irreducible aperiodic Markov chain that has as its stationary distribution a target distribution in a high-dimensional space having some special structure.

The most interesting aspect of this technique: to run this Markov chain, it suffices to generate observations from *univariate distributions*.

The *Gibbs sampler* in the context of a bivariate probability distribution can be described as follows. Let π be a target probability distribution of a bivariate random vector (X, Y). For each x, let $P(x, \cdot)$ be the conditional probability distribution of Y given $X = x$. Similarly, let $Q(y, \cdot)$ be the conditional probability distribution of X given $Y = y$. Note that for each x, $P(x, \cdot)$ is a univariate distribution, and for each y, $Q(y, \cdot)$ is also a univariate distribution. Now generate a bivariate Markov chain $Z_n = (X_n, Y_n)$ as follows:

Start with some $X_0 = x_0$. Generate an observation Y_0 from the distribution $P(x_0, \cdot)$. Then generate an observation X_1 from $Q(Y_0, \cdot)$. Next generate an observation Y_1 from $P(X_1, \cdot)$ and so on. At stage n if $Z_n = (X_n, Y_n)$ is known, then generate X_{n+1} from $Q(Y_n, \cdot)$ and Y_{n+1} from $P(X_{n+1}, \cdot)$.

If π is a discrete distribution concentrated on $\{(x_i, y_j) : 1 \le i \le K, 1 \le j \le L\}$ and if $\pi_{ij} = \pi(x_i, y_j)$ then $P(x_i, y_j) = \pi_{ij}/\pi_{i\cdot}$ and $Q(y_j, x_i) = \pi_{ij}/\pi_{\cdot j}$, where $\pi_{i\cdot} = \sum_j \pi_{ij}$, $\pi_{\cdot j} = \sum_i \pi_{ij}$. Thus the transition probability matrix $R = ((r_{(ij),(k\ell)}))$ for the $\{Z_n\}$ chain is given by

$$\begin{aligned} r_{(ij),(k\ell)} &= Q(y_j, x_k) P(x_k, y_\ell) \\ &= \frac{\pi_{kj}}{\pi_{\cdot j}} \frac{\pi_{k\ell}}{\pi_{k\cdot}}. \end{aligned}$$

Verify that this chain is irreducible, aperiodic, and has π as its stationary distribution. Thus LLN (16) and (18) hold in this case. Thus for large n, Z_n can be viewed as a sample from a distribution that is close to π and one can approximate $\sum_{i,j} h(i,j)\pi_{ij}$ by $\sum_{1=1}^{n} h(X_i, Y_i)/n$.

Illustration: Consider sampling from $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$. The

conditional distribution of X given $Y = y$ and that of Y given $X = x$ are

$$X|Y = y \sim N(\rho y, 1 - \rho^2) \text{ and } Y|X = x \sim N(\rho x, 1 - \rho^2). \qquad (23)$$

Using this property, Gibbs sampling proceeds as follows: Generate (X_n, Y_n), $n = 0, 1, 2, \ldots$, by starting from an arbitrary value x_0 for X_0, and repeat the following steps for $i = 0, 1, \ldots, n$.

1. Given x_i for X, draw a random deviate from $N(\rho x_i, 1-\rho^2)$ and denote it by Y_i.
2. Given y_i for Y, draw a random deviate from $N(\rho y_i, 1-\rho^2)$ and denote it by X_{i+1}.

The theory of Gibbs sampling tells us that if n is large, then (x_n, y_n) is a random draw from a distribution that is close to $N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$.

Multivariate extension: π is a probability distribution of a k-dimensional random vector $(X_1, X_2, \ldots, X_k)$. If $\boldsymbol{u} = (u_1, u_2, \ldots, u_k)$ is any k-vector, let $\boldsymbol{u}_{-i} = (u_1, u_2, \ldots, u_{i-1}, u_{i+1}, \ldots, u_k)$ be the $k-1$ dimensional vector resulting by dropping the ith component u_i. Let $\pi_i(\cdot|\boldsymbol{x}_{-i})$ denote the univariate conditional distribution of X_i given that $\boldsymbol{X}_{-i} \equiv (X_1, X_2, X_{i-1}, X_{i+1}, \ldots, X_k) = \boldsymbol{x}_{-i}$. Starting with some initial value for $\boldsymbol{X}_0 = (x_{01}, x_{02}, \ldots, x_{0k})$ generate $\boldsymbol{X}_1 = (X_{11}, X_{12}, \ldots, X_{1k})$ sequentially by generating X_{11} according to the univariate distribution $\pi_1(\cdot|\boldsymbol{x}_{0_{-1}})$ and then generating X_{12} according to $\pi_2(\cdot|(X_{11}, x_{03}, x_{04}, \ldots, x_{0k})$ and so on.

The most important feature to recognize here is that all the univariate conditional distributions, $X_i|\boldsymbol{X}_{-i} = \boldsymbol{x}_{-i}$, known as *full conditionals* should easily allow sampling from them. This is the case in most hierarchical Bayes problems. Thus, the Gibbs sampler is particularly well adapted for Bayesian computations with hierarchical priors.

Rao-Blackwellization

The variance reduction idea of the famous *Rao-Blackwell theorem* in the presence of auxiliary information can be used to provide improved estimators when MCMC procedures are adopted.

Theorem (Rao-Blackwell) Let $\delta(X_1, X_2, \ldots, X_n)$ be an estimator of θ with finite variance. Suppose that T is sufficient for θ, and let $\delta^*(T)$, defined

by $\delta^*(t) = E(\delta(X_1, X_2, \ldots, X_n)|T = t)$, be the conditional expectation of $\delta(X_1, X_2, \ldots, X_n)$ given $T = t$. Then

$$E(\delta^*(T) - \theta)^2 \leq E(\delta(X_1, X_2, \ldots, X_n) - \theta)^2.$$

The inequality is strict unless $\delta = \delta^*$, or equivalently, δ is already a function of T.

By the property of iterated conditional expectation,

$$E(\delta^*(T)) = E\left[E(\delta(X_1, X_2, \ldots, X_n)|T)\right] = E(\delta(X_1, X_2, \ldots, X_n)).$$

Therefore, to compare the mean squared errors (MSE) of the two estimators, compare their variances only. Now,

$$\begin{aligned} \mathrm{Var}(\delta(X_1, X_2, \ldots, X_n)) &= \mathrm{Var}\left[E(\delta|T)\right] + E\left[\mathrm{Var}(\delta|T)\right] \\ = \mathrm{Var}(\delta^*) + E\left[\mathrm{Var}(\delta|T)\right] &> \mathrm{Var}(\delta^*), \end{aligned}$$

unless $\mathrm{Var}(\delta|T) = 0$, which is the case only if δ is a function of T.

The Rao–Blackwell theorem involves two key steps: variance reduction by conditioning and conditioning by a sufficient statistic. The first step is based on the *analysis of variance* formula: For any two random variables S and T, because

$$\mathrm{Var}(S) = \mathrm{Var}(E(S|T)) + E(\mathrm{Var}(S|T)),$$

one can reduce the variance of a random variable S by taking conditional expectation given some auxiliary information T. This can be exploited in MCMC.

$(X_j, Y_j), j = 1, 2, \ldots, N$: a single run of the Gibbs sampler algorithm with a target distribution of a bivariate random vector (X, Y). Let $h(X)$ be a function of the X component of (X, Y) and let its mean value be μ. Goal is to estimate μ. A first estimate is the sample mean of the $h(X_j), j = 1, 2, \ldots, N$. From the MCMC theory, as $N \to \infty$, this estimate will converge to μ in probability. The computation of variance of this estimator is not easy due to the (Markovian) dependence of the sequence $\{X_j, j = 1, 2, \ldots, N\}$. Suppose we make n independent runs of Gibbs sampler and generate $(X_{ij}, Y_{ij}), j = 1, 2, \ldots, N; i = 1, 2, \ldots, n$. Suppose that N is sufficiently large so that (X_{iN}, Y_{iN}) can be regarded as a sample from the limiting target distribution of the Gibbs sampling scheme. Thus $(X_{iN}, Y_{iN}), i = 1, 2, \ldots, n$ form a random sample from the target distribution. Consider a second estimate of μ—the sample mean of $h(X_{iN}), i = 1, 2, \ldots, n$.

This estimator ignores part of the MCMC data but has the advantage that the variables $h(X_{iN})$, $i = 1, 2, \ldots, n$ are independent and hence the variance of their mean is of order n^{-1}. Now applying the variance reduction idea of the Rao-Blackwell theorem by using the auxiliary information Y_{iN}, $i = 1, 2, \ldots, n$, one can improve this estimator as follows:

Let $k(y) = E(h(X)|Y = y)$. Then for each i, $k(Y_{iN})$ has a smaller variance than $h(X_{iN})$ and hence the following third estimator,

$$\frac{1}{n}\sum_{i=1}^{n} k(Y_{iN}),$$

has a smaller variance than the second one. A crucial fact to keep in mind here is that the exact functional form of $k(y)$ be available for implementing this improvement.

(Example M2 continued.) $X|\theta \sim N(\theta, \sigma^2)$ with known σ^2 and $\theta \sim$ Cauchy(μ, τ). Simulate θ from the posterior distribution, but sampling directly is difficult.

Gibbs sampling: Cauchy is a scale mixture of normal densities, with the scale parameter having a Gamma distribution.

$$\begin{aligned}\pi(\theta) &\propto \left(\tau^2 + (\theta-\mu)^2\right)^{-1} \\ &\propto \int_0^\infty (\frac{\lambda}{2\pi\tau^2})^{1/2} \exp\left(-\frac{\lambda}{2\tau^2}(\theta-\mu)^2\right) \lambda^{1/2-1} \exp(-\frac{\lambda}{2})\, d\lambda,\end{aligned}$$

so that $\pi(\theta)$ may be considered the marginal prior density from the joint prior density of (θ, λ) where

$$\theta|\lambda \sim N(\mu, \tau^2/\lambda) \text{ and } \lambda \sim \text{Gamma}(1/2, 1/2).$$

This implicit hierarchical prior structure implies: $\pi(\theta|x)$ is the marginal density from $\pi(\theta, \lambda|x)$.

Full conditionals of $\pi(\theta, \lambda|x)$ are standard distributions:

$$\theta|\lambda, x \sim N\left(\frac{\tau^2}{\tau^2+\lambda\sigma^2}x + \frac{\lambda\sigma^2}{\tau^2+\lambda\sigma^2}\mu, \frac{\tau^2\sigma^2}{\tau^2+\lambda\sigma^2}\right), \tag{24}$$

$$\lambda|\theta, x \sim \lambda|\theta \sim \text{Exponential}\left(\frac{\tau^2+(\theta-\mu)^2}{2\tau^2}\right). \tag{25}$$

Thus, the Gibbs sampler will use (24) and (25) to generate (θ, λ) from $\pi(\theta, \lambda|x)$.

Example M5. X = number of defectives in the daily production of a product. $(X \mid Y, \theta) \sim$ binomial(Y, θ), where Y, a day's production, is Poisson with known mean λ, and θ is the probability that any product is defective. The difficulty is that Y is not observable, and inference has to be made on the basis of X only. Prior: $(\theta \mid Y = y) \sim$ Beta(α, γ), with known α and γ independent of Y. Bayesian analysis here is not difficult because the posterior distribution of $\theta|X = x$ can be obtained as follows. First, $X|\theta \sim$ Poisson$(\lambda\theta)$. Next, $\theta \sim$ Beta(α, γ). Therefore,

$$\pi(\theta|X = x) \quad \propto \quad \exp(-\lambda\theta)\theta^{x+\alpha-1}(1-\theta)^{\gamma-1}, 0 < \theta < 1. \tag{26}$$

This is not a standard distribution, and hence posterior quantities cannot be obtained in closed form. Instead of focusing on $\theta|X$ directly, view it as a marginal component of $(Y, \theta \mid X)$. Check that the full conditionals of this are given by
$Y|X = x, \theta \sim x +$ Poisson$(\lambda(1-\theta))$, and
$\theta|X = x, Y = y \sim$ Beta$(\alpha + x, \gamma + y - x)$
both of which are standard distributions.

Example M5 continued. It is actually possible here to sample from the posterior distribution using the *accept-reject* Monte Carlo method:

Let $g(\boldsymbol{x})/K$ be the target density, where K is the possibly unknown normalizing constant of the unnormalized density g. Suppose $h(\boldsymbol{x})$ is a density that can be simulated by a known method and is close to g, and suppose there exists a known constant $c > 0$ such that $g(\boldsymbol{x}) < ch(\boldsymbol{x})$ for all $\boldsymbol{x}$. Then, to simulate from the target density, the following two steps suffice. Step 1. Generate $\boldsymbol{Y} \sim h$ and $U \sim U(0,1)$;
Step 2. Accept $\boldsymbol{X} = \boldsymbol{Y}$ if $U \leq g(\boldsymbol{Y})/\{ch(\boldsymbol{Y})\}$; return to Step 1 otherwise.
The optimal choice for c is $\sup\{g(\boldsymbol{x})/h(\boldsymbol{x})$.

In Example M5, from (26),

$$g(\theta) = \exp(-\lambda\theta)\theta^{x+\alpha-1}(1-\theta)^{\gamma-1}I\{0 \leq \theta \leq 1\},$$

so that $h(\theta)$ may be chosen to be the density of Beta$(x+\alpha, \gamma)$. Then, with the above-mentioned choice for c, if $\theta \sim$ Beta$(x+\alpha, \gamma)$ is generated in Step 1, its 'acceptance probability' in Step 2 is simply $\exp(-\lambda\theta)$.

Even though this method works here, let us see how the Metropolis-Hastings algorithm can be applied.

The required Markov chain is generated by taking the transition density $q(z, y) = q(y|z) = h(y)$, independently of z. Then the acceptance probability

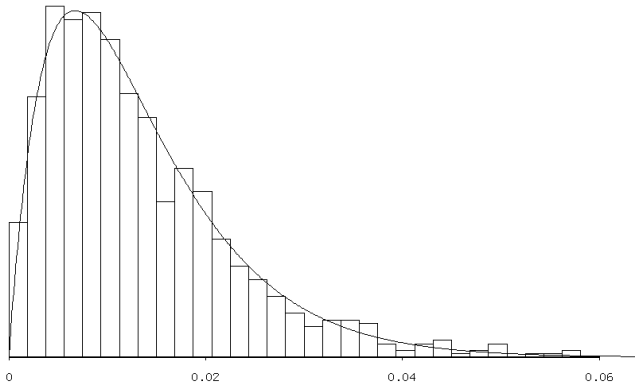


Figure 1: M-H frequency histogram and true posterior density.

is

$$\begin{aligned}\rho(z,y) &= \min\left\{\frac{g(y)h(z)}{g(z)h(y)},1\right\} \\ &= \min\left\{\exp\left(-\lambda(y-z)\right),1\right\}.\end{aligned}$$

The steps involved in this "independent" M-H algorithm are:

Start at $t=0$ with a value x_0 in the support of the target distribution; in this case, $0<x_0<1$. Given x_t, generate the next value in the chain as given below.

(a) Draw Y_t from Beta$(x+\alpha,\gamma)$.

(b) Let

$$x_{(t+1)} = \begin{cases} Y_t & \text{with probability } \rho_t \\ x_t & \text{otherwise,} \end{cases}$$

where $\rho_t = \min\{\exp\left(-\lambda(Y_t - x_t)\right),1\}$.

(c) Set $t=t+1$ and go to step (a).

Run this chain until $t=n$, a suitably chosen large integer. In our example, for $x=1$, $\alpha=1$, $\gamma=49$ and $\lambda=100$, we simulated such a Markov chain. The resulting frequency histogram is shown in Figure below, with the true posterior density super-imposed on it.

10 Empirical Bayes Methods for High Dimensional Problems

This is becoming popular again, this time for 'high dimensional' problems. Astronomers routinely estimate characteristics of millions of similar astronomical objects – distance, radial velocity whatever. Consider the data:

$$(\mathbf{X}_1 = \begin{pmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1n} \end{pmatrix}, \mathbf{X}_2 = \begin{pmatrix} X_{21} \\ X_{22} \\ \vdots \\ X_{2n} \end{pmatrix}, \cdots \mathbf{X}_p = \begin{pmatrix} X_{p1} \\ X_{p2} \\ \vdots \\ X_{pn} \end{pmatrix}).$$

$\mathbf{X}_j$ represents n repeated independent observations on the jth object, $j = 1, 2, \ldots p$. The important point is n is small, 2, 5, or 10, whereas p is large, such as a million.

Suppose $X_{j1}, \ldots X_{jn}$ measure μ_j with variability σ^2.

Problem: Maximum likelihood can give wrong estimates

Take $n = 2$ and suppose

$$\begin{pmatrix} X_{j1} \\ X_{j2} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_j \\ \mu_j \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}\right), \quad j = 1, 2, \ldots p.$$

i.e., we measure μ_j with 2 independent measurements, each coming with a $N(0, \sigma^2)$ error added to it; we do this for a very large number p of objects. What is the MLE of σ^2?

$$\begin{aligned} l(\mu_1, \ldots \mu_p; \sigma^2 | \mathbf{x}_1, \ldots \mathbf{x}_p) &= f(\mathbf{x}_1, \ldots \mathbf{x}_p | \mu_1, \ldots \mu_p; \sigma^2) \\ &= \prod_{j=1}^{p} \prod_{i=1}^{2} f(x_{ji} | \mu_j, \sigma^2) \\ &= (2\pi\sigma^2)^{-p} \exp(-\frac{1}{2\sigma^2} \sum_{j=1}^{p} \sum_{i=1}^{2} (x_{ji} - \mu_j)^2) \\ &= (2\pi\sigma^2)^{-p} \exp(-\frac{1}{2\sigma^2} \sum_{j=1}^{p} \left[\sum_{i=1}^{2} (x_{ji} - \bar{x}_j)^2 + 2(\bar{x}_j - \mu_j)^2 \right]). \end{aligned}$$

$\hat{\mu}_j = \bar{x}_j = (x_{j1} + x_{j2})/2$ and

$$\begin{aligned}
\hat{\sigma^2} &= \frac{1}{2p}\sum_{j=1}^{p}\sum_{i=1}^{2}(x_{ji} - \bar{x}_j)^2 \\
&= \frac{1}{2p}\sum_{j=1}^{p}\left[\left(x_{j1} - \frac{x_{j1}+x_{j2}}{2}\right)^2 + \left(x_{j2} - \frac{x_{j1}+x_{j2}}{2}\right)^2\right] \\
&= \frac{1}{2p}\sum_{j=1}^{p} 2\frac{(x_{j1}-x_{j2})^2}{4} = \frac{1}{4p}\sum_{j=1}^{p}(x_{j1}-x_{j2})^2.
\end{aligned}$$

Since $X_{j1} - X_{j2} \sim N(0, 2\sigma^2)$, $j = 1, 2\ldots$,

$$\frac{1}{p}\sum_{j=1}^{p}(X_{j1}-X_{j2})^2 \xrightarrow[p\to\infty]{P} 2\sigma^2, \text{ so that}$$

$$\hat{\sigma^2} = \frac{1}{4p}\sum_{j=1}^{p}(X_{j1}-X_{j2})^2 \xrightarrow[p\to\infty]{P} \frac{\sigma^2}{2}, \text{ and not } \sigma^2.$$

Good estimates for σ^2 do exist, for example,

$$\frac{1}{2p}\sum_{j=1}^{p}(X_{j1}-X_{j2})^2 \xrightarrow[p\to\infty]{P} 2\sigma^2.$$

What is going wrong here?

This is not a *small* p, *large* n problem, but a *small* n, *large* p problem. i.e. a high dimensional problem, so needs care!

As $p \to \infty$, there are too many parameters to estimate and the likelihood function is unable to see where information lies, so tries to distribute it everywhere.

What is the way out? Go Bayesian!

There is a lot of information available on σ^2 (note $\sum_{j=1}^{p}(X_{j1}-X_{j2})^2 \sim 2\sigma^2\chi^2_p$) but very little on individual μ_j. However, if μ_j are 'similar', there is a lot of information on where they come from, because we get to see p samples, p large.

Suppose we are interested in μ_j. How can we use the above information? Model as follows:

$\bar{X}_j|\mu_j, \sigma^2 \sim N(\mu_j, \sigma^2/2)$, $j = 1, \ldots p$, independent observations.
σ^2 may be assumed known, since a reliable estimate $\hat{\sigma^2} = \frac{1}{2p}\sum_{j=1}^{p}(X_{j1} - X_{j2})^2$ is available. Express the information that μ_j are 'similar' in the form: $\mu_j, j = 1, \ldots p$ is a random sample (collection) from $N(\eta, \tau^2)$. Where do we get the η and τ^2, the prior mean and prior variance?

Marginally (or in predictive sense) $\bar{X}_j, j = 1, \ldots p$ is a random sample from $N(\mu_0, \tau^2 + \sigma^2/2)$. Use this random sample.

Estimate η by $\hat{\eta} = \bar{\bar{X}} = \frac{1}{p}\sum \bar{X}_j$ and τ^2 by $\hat{\tau^2} = \left(\frac{1}{p-1}\sum_{j=1}^{p}(\bar{X}_j - \bar{\bar{X}})^2 - \sigma^2/2\right)^+$.

Now one could pretend that the prior for $(\mu_1, \ldots \mu_p)$ is $N(\hat{\eta}, \hat{\tau^2})$ and compute the Bayes estimates for μ_j:

$$E(\mu_j|\mathbf{X}_1, \ldots \mathbf{X}_p) = (1 - \hat{B})\bar{X}_j + \hat{B}\bar{\bar{X}},$$

where $\hat{B} = \frac{\sigma^2/2}{\sigma^2/2 + \hat{\tau^2}}$. If instead of 2 observations, each sample has n observations, replace 2 by n. This is called *Empirical Bayes* since the prior is estimated using data. There is also a fully Bayesian counter-part called Hierarchical Bayes.