Bayesian Essentials

Feng Li
feng.li@cufe.edu.cn

School of Statistics and Mathematics
Central University of Finance and Economics

NOTE: Most of the contents are from Bayesian Statistics course taught by Professor Mattias Villani

http://www.mattiasvillani.com/teaching/bayesian-statistics/

Overview

- Likelihood
- Bayesian inference
- The Bernoulli model
- The Normal model

The likelihood function

- Bernoulli trials:

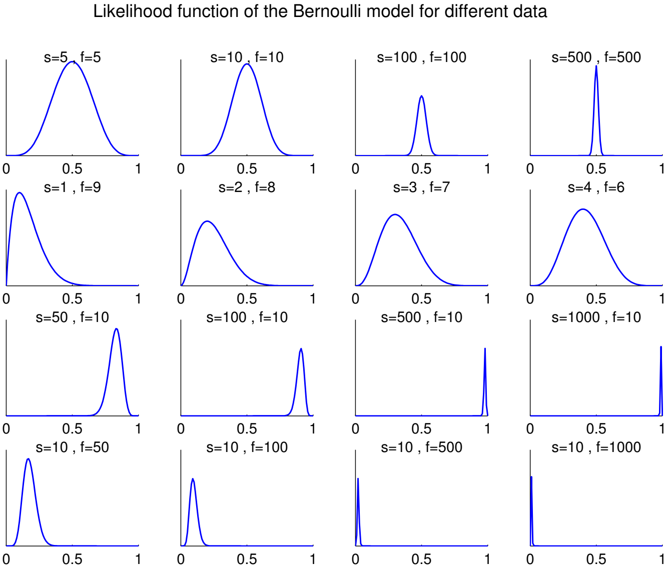
$$x_1, ..., x_n|\theta \overset{iid}{\sim} Bern(\theta).$$

- Likelihood:

$$\begin{aligned} p(x_1, ..., x_n|\theta) &= p(x_1|\theta) \cdots p(x_n|\theta) \\ &= \theta^s (1-\theta)^f, \end{aligned}$$

  where $s = \sum_{i=1}^n x_i$ is the number of successes in the Bernoulli trials and $f = n - s$ is the number of failures.
- Given the data $x_1, ..., x_n$, we may plot $p(x_1, ..., x_n|\theta)$ as a function of θ.

The likelihood function from Bernoulli trials



Uncertainty and subjective probability

- Will the likelihood give us un idea of which values of θ that should be regarded as probable (in some sense)? Kind of, but ... No!
- In order to say that one value of θ is more probable than another we clearly must think of θ as random. But θ may be something that we know is non-random, *e.g.* a fixed natural constant.
- **Bayesian: doesn't matter if θ is fixed or random**. What matters is whether or not You know the value of θ. If θ is uncertainty to You, then You can assign a probability distribution to θ which reflects Your knowledge about θ. **Subjective probability**.

Learning from data - Bayes' theorem

- Given that you have formulated a distribution for θ, $p(\theta)$, how can we learn from data? That is, how do we make the transition from $p(\theta) \to p(\theta|Data)$? Bayes' theorem is the key.
- One form of Bayes' theorem reads (A and B are events)

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}.$$

  So that Bayes' theorem 'reverses the conditioning', i.e. takes us from $p(B|A)$ to $p(A|B)$.
- Let $A = \theta$ and $B = Data$

$$p(\theta|Data) = \frac{p(Data|\theta)p(\theta)}{p(Data)}.$$

- Interpreting the likelihood function as a probability density for θ is just as wrong as ignoring the factor $p(A)/p(B)$ in Bayes' theorem.

Generalized Bayes' theorem

- From your basic statistics textbook:

$$p(A_i|B) = \frac{p(B|A_i)p(A_i)}{p(B)} = \frac{p(B|A_i)p(A_i)}{\sum_{i=1}^{k} p(B|A_i)p(A_i)}.$$

- Let $\theta_1, ..., \theta_k$ be k different values on a parameter θ. Bayes' Theorem:

$$p(\theta_i|Data) = \frac{p(Data|\theta_i)p(\theta_i)}{p(Data)} = \frac{p(Data|\theta_i)p(\theta_i)}{\sum_{i=1}^{k} p(Data|\theta_i)p(\theta_i)}.$$

- If θ takes on a continuum of values

$$p(\theta|Data) = \frac{p(Data|\theta)p(\theta)}{\int_\theta p(Data|\theta)p(\theta)d\theta}.$$

The joy of ignoring a normalizing constant

- When *Data* is known, $p(Data)$ in Bayes' theorem is just a constant that makes $p(\theta|Data)$ integrate to one. Example: $x \sim N(\mu, \sigma^2)$

$$p(x) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right].$$

- We may write

$$p(x) \propto \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right].$$

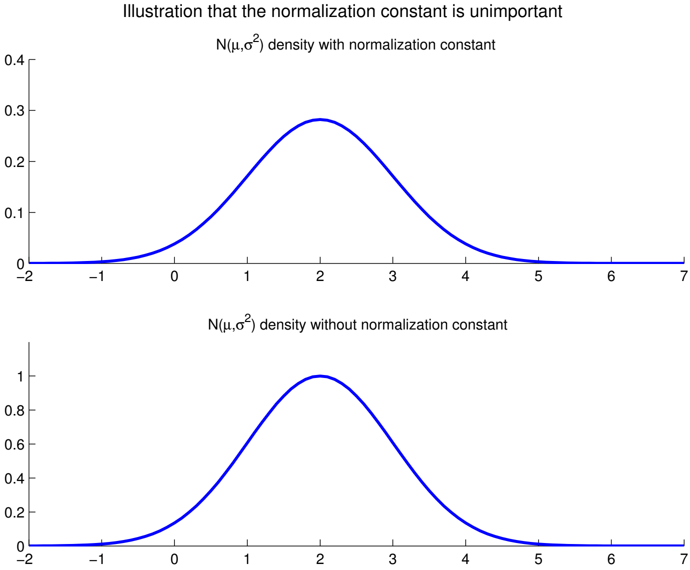
- Short form of Bayes' theorem

$$p(\theta|Data) \propto p(Data|\theta)p(\theta)$$

or

$$\text{Posterior} \propto \text{ Likelihood } \cdot \text{ Prior}$$

Normalization constant is not important



Bayesian updating

- Suppose: you already have $x_{1,} x_2, \ldots, x_n$ data points, and the corresponding posterior $p(\theta|x_1, \ldots, x_n)$
- Now, a fresh additional data point x_{n+1} arrives.
- The posterior based on all available data is

$$p(\theta|x_{1,} \ldots, x_{n+1}) \propto p(x_{n+1}|\theta, x_1, \ldots, x_n) p(\theta|x_1, \ldots, x_n).$$

  - The following are therefore equivalent:
  - Analyzing the likelihood of all data $x_{1,} \ldots, x_{n+1}$ with the prior based on no data $p(\theta)$
  - Analyzing the likelihood of the fresh data point x_{n+1} with the 'prior' equal to the posterior based on the old data $p(\theta|x_1, \ldots, x_n)$.
- Yesterday's posterior is today's prior.

Bernoulli trials - Beta prior

- Model:

$$x_1, \ldots, x_n|\theta \overset{iid}{\sim} Bern(\theta)$$

- Prior:

$$\theta \sim Beta(\alpha, \beta)$$

$$p(y) = \frac{\Gamma(\alpha, \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \quad \text{for } 0 \leq y \leq 1.$$

- Posterior

$$\begin{aligned} p(\theta|x_1, \ldots, x_n) &\propto p(x_1, \ldots, x_n|\theta)p(\theta) \\ &= \theta^s(1-\theta)^f\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}. \end{aligned}$$

- But this is recognized as proportional to the $Beta(\alpha+s, \beta+f)$ density. That is, the prior-to-posterior mapping reads

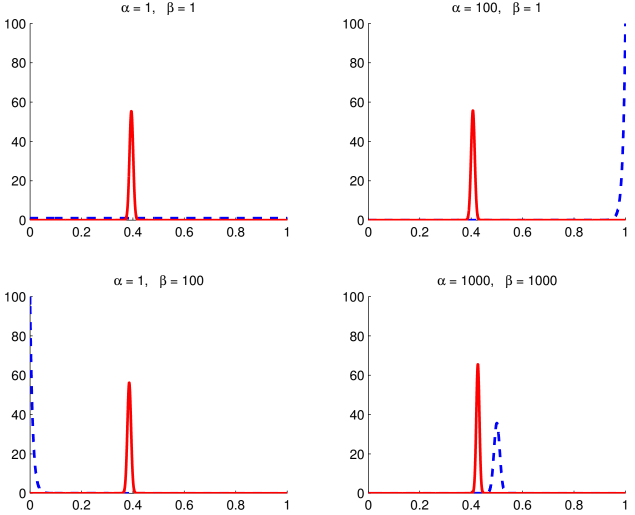
$$\theta \sim Beta(\alpha, \beta) \overset{x_1, \ldots, x_n}{\Longrightarrow} \theta|x_1, \ldots, x_n \sim Beta(\alpha+s, \beta+f).$$

Bernoulli example: spam emails

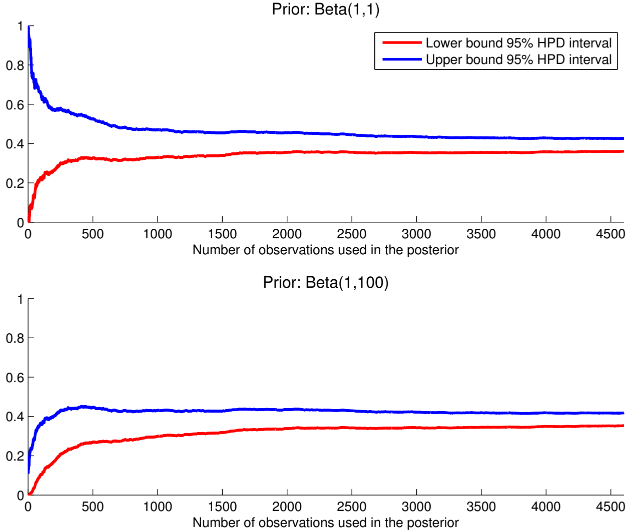
- George has gone through his collection of 4601 e-mails. He classified 1813 of them to be spam.
- Let $x_i = 1$ if i:th email is spam. Assume $x_i|\theta \overset{iid}{\sim} \mathit{Bernoulli}(\theta)$ and $\theta \sim \mathrm{Beta}(\alpha, \beta)$.
- Posterior

$$\theta|x \sim \mathit{Beta}(\alpha + 1813, \beta + 2788)$$

Spam data: The effect of different priors



Spam data: Posterior convergence



Normal data with known variance - uniform prior

- Model:

$$x_1, ..., x_n | \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2).$$

- Prior:

$$p(\theta) \propto c$$

- Likelihood (see Technical Appendix A):

$$\begin{aligned} p(x_1, ..., x_n | \theta, \sigma^2) &= \Pi_{i=1}^n (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(x_i - \theta)^2\right] \\ &\propto \exp\left[-\frac{1}{2(\sigma^2/n)}(\theta - \bar{x})^2\right]. \end{aligned}$$

- Posterior

$$\theta | x_1, ..., x_n \sim N(\bar{x}, \sigma^2/n)$$

Normal with known variance - normal prior

- Prior
$$\theta \sim N(\mu_0, \tau_0^2)$$
- Posterior (see Technical Appendix A)
$$\begin{aligned} p(\theta|x_1, ..., x_n) &\propto p(x_1, ..., x_n|\theta, \sigma^2)p(\theta) \\ &\propto N(\theta|\mu_n, \tau_n^2), \end{aligned}$$

where
$$\frac{1}{\tau_n^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2},$$
$$\mu_n = w\bar{x} + (1 - w)\mu_0,$$

and
$$w = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}.$$

Normal with known variance - normal prior, cont.

$$\theta \sim N(\mu_0, \tau_0^2) \overset{x_1,\ldots,x_n}{\Longrightarrow} \theta | x \sim N(\mu_n, \tau_n^2).$$

Posterior precision = Data precision + Prior precision

Posterior mean =

$\frac{\text{Data precision}}{\text{Posterior precision}}$(Data mean) + $\frac{\text{Prior precision}}{\text{Posterior precision}}$(Prior mean)

Overview Lecture 2

- Conjugate priors
- Poisson model
- 'Non-Informative' priors
- Jeffreys' prior

Conjugate priors

- Normal likelihood: Normal prior$\rightarrow$Normal posterior. (posterior belongs to the same distribution family as prior)
- Binomial likelihood: Beta prior$\rightarrow$Beta posterior.
- *Conjugate priors*: Let $\mathcal{F} = \{p(y|\theta), \theta \in \Theta\}$ be a class of sampling distributions. A family of distributions $\mathcal{P}$ is conjugate for $\mathcal{F}$ if

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|x) \in \mathcal{P}$$

  holds for all $p(x|\theta) \in \mathcal{F}$.
- *Natural conjugate prior*: $p(\theta) = c \cdot p(y_1, ..., y_n|\theta)$ for some constant c, i.e. the prior is of the same functional form as the likelihood.

Poisson model

- Likelihood from iid Poisson sample $y = (y_1, ..., y_n)$

$$p(y|\theta) = \left[\prod_{i=1}^{n} p(y_i|\theta)\right] \propto \theta^{(\sum_{i=1}^{n} y_i)} \exp(-\theta n),$$

  so that the sum of counts $\sum_{i=1}^{n} y_i$ is a sufficient statistic for θ.
- *Natural conjugate prior for Poisson parameter θ*

$$p(\theta) \propto \theta^{\alpha-1} \exp(-\theta\beta) \propto Gamma(\alpha, \beta)$$

  which contains the info: $\alpha - 1$ counts in β observations.

Poisson model, cont.

- *Posterior for Poisson parameter* θ. Multiplying the poisson likelihood and the Gamma prior gives the posterior

$$\begin{aligned} p(\theta|y_1,...,y_n) &\propto \left[\prod_{i=1}^{n} p(y_i|\theta)\right] p(\theta) \\ &\propto \theta^{\sum_{i=1}^{n} y_i} \exp(-\theta n)\theta^{\alpha-1} \exp(-\theta\beta) \\ &= \theta^{\alpha+\sum_{i=1}^{n} y_i-1} \exp[-\theta(\beta+n)], \end{aligned}$$

  which is proportional to the $Gamma(\alpha + \sum_{i=1}^{n} y_i, \beta + n)$ distribution.
- In summary

$$\begin{aligned} \text{Model: } & y_1,...,y_n|\theta \overset{iid}{\sim} Po(\theta) \\ \text{Prior: } & \theta \sim Gamma(\alpha,\beta) \\ \text{Posterior: } & \theta|y_1,...,y_n \sim Gamma(\alpha + \sum_{i=1}^{n} y_i, \beta + n). \end{aligned}$$

Poisson example - Number of bomb hits in London

$n = 576$, $\sum_{i=1}^{n} y_i = 229 \cdot 0 + 211 \cdot 1 + 93 * 2 + 35 * 3 + 7 * 4 + 1 \cdot 5 = 537$.

Average number of hits per region$=\bar{y} = 537/576 \approx 0.9323$.

$$p(\theta|y) \propto \theta^{\alpha+537-1} \exp[-\theta(\beta+576)]$$

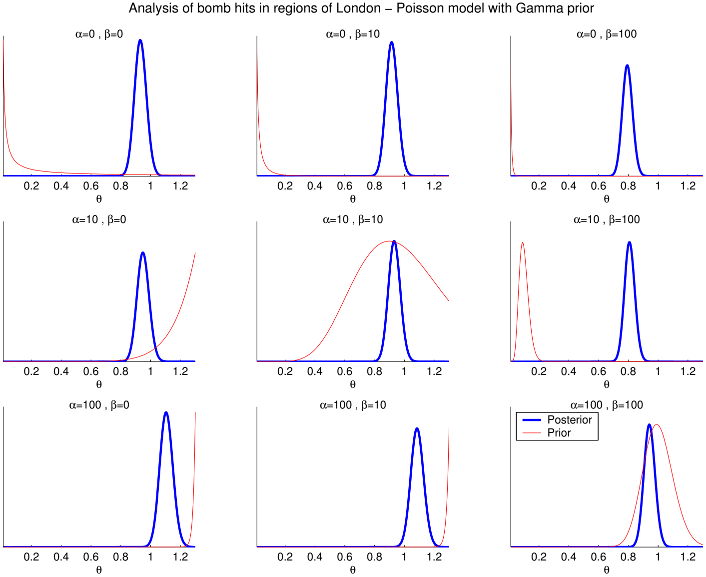
$$E(\theta|y) = \frac{\alpha + \sum_{i=1}^{n} y_i}{\beta + n} \approx \bar{y} \approx 0.9323,$$

and

$$SD(\theta|y) = \left(\frac{\alpha + \sum_{i=1}^{n} y_i}{(\beta+n)^2}\right)^{1/2} = \frac{(\alpha + \sum_{i=1}^{n} y_i)^{1/2}}{(\beta+n)} \approx \frac{(537)^{1/2}}{576} \approx 0.0402.$$

if α and β are small compared to $\sum_{i=1}^{n} y_i$ and n.

Poisson bomb hits in London



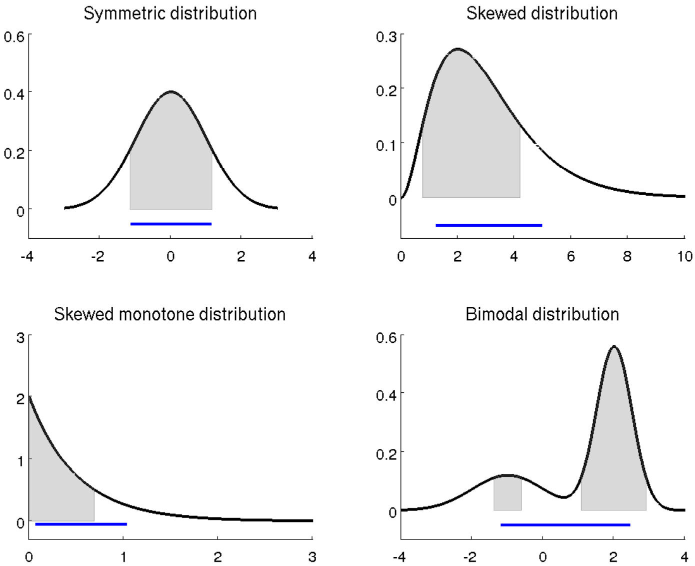
Poisson example - posterior probability intervals

- Bayesian 95% interval: the probability that the unknown parameter θ lies in the interval is 0.95. What a relief!
- Approximate 95% credible interval for θ (for small α and β):

$$E(\theta|y) \pm 1.96 \cdot SD(\theta|y) = [0.8535; 1.0111]$$

- An exact 95% equal-tail interval is $[0.8550;\ 1.0125]$ (assuming $\alpha = \beta = 0$)
- An exact Highest Posterior Density (HPD) interval is $[0.8525;\ 1.0144]$. Obtained numerically, assuming $\alpha = \beta = 0$.

Illustration of different interval types



Non-informative priors

- ... do not exist!
- ... may be improper and still lead to proper posterior
- Regularization priors
- Ideal communication. Present the posterior distributions for all possible priors.
- Practical communication - Reference priors.
- Cannot demand that users specify priors on high-dimensional in detail. Model the prior in terms of a few hyperparameters.
- Subjective consensus: when extreme priors give essentially the same posterior. This will happen, given enough data as

$$p(\theta|y) \to N\left[\hat{\theta}, I^{-1}\right] \text{ for all } p(\theta) \text{ as } n \to \infty.$$

Jeffreys' prior

- A common non-informative prior is Jeffreys' prior

$$p(\theta) = |I(\theta)|^{1/2},$$

where

$$J(\theta) = -E_{y|\theta}\left[\frac{d^2 \ln p(y|\theta)}{d\theta^2}\right]$$

is the expected Fisher information.

Jeffreys' prior for Bernoulli trial data

$$y_1, \ldots, y_n | \theta \overset{iid}{\sim} Bern(\theta).$$

$$\ln p(y|\theta) = s \ln \theta + f \ln(1-\theta)$$

$$\frac{d \ln p(y|\theta)}{d\theta} = \frac{s}{\theta} - \frac{f}{(1-\theta)}$$

$$\frac{d^2 \ln p(y|\theta)}{d\theta^2} = -\frac{s}{\theta^2} - \frac{f}{(1-\theta)^2}$$

$$J(\theta) = \frac{E_{y|\theta}(s)}{\theta^2} + \frac{E_{y|\theta}(f)}{(1-\theta)^2} = \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}$$

Thus, the Jeffreys' prior is

$$p(\theta) = |J(\theta)|^{1/2} \propto \theta^{-1/2}(1-\theta)^{-1/2} \propto Beta(\theta|1/2, 1/2).$$

Jeffreys' prior Binomial vs Negative binomial sampling

- Bernoulli experiment: Perform n independent trials with success probabilty θ and count the number of successes. Here

$$y|\theta \sim Bin(\theta)$$

- Inverse Bernoulli experiment: Perform independent trials with success probabilty θ until you have observed y successes. Here

$$y|\theta \sim NegBin(\theta)$$

- Exercise: Suppose you performed both of the two experiments and that in both cases you ended up doing n trials and observed y successes. Show that the likelihood function conveys the same information on θ in both cases, but that Jeffreys prior is not the same in both models. Is this reasonable?

Properties of Jeffreys prior

- Invariant to 1:1 transformations of θ. Doesn't matter which parametrization we derive the prior, it always contains the same info.
- Two models with identical likelihood functions (up to constant) can yield different Jeffreys' prior. Jeffreys' prior does not respect the likelihood principle. The crux of the matter is the expectation with respect to the sampling distribution.
- Jeffreys' prior may be a very complicated (non-conjugate) distribution.
- Problematic in multivariate problems. Dubious results in many standard models.

Lecture overview

- Multiparameter models
- Marginalization
- Normal model with unknown variance
- Bayesian analysis of multinomial data
- Bayesian analysis of multivariate normal data

Marginalization

- Models usually contains several parameter $\theta_1, \theta_2, \ldots$. Examples: $x_i \overset{iid}{\sim} N(\theta, \sigma^2)$; multiple regression ...
- The Bayesian computes the joint posterior distribution

$$p(\theta_1, \theta_2, ..., \theta_p | y) \propto p(y|\theta_1, \theta_2, ..., \theta_p) p(\theta_1, \theta_2, ..., \theta_p).$$

  ... or in vector form:

$$p(\theta) \propto p(y|\theta)p(\theta).$$

- Complicated to graph the joint posterior.
- Some of the parameters may not be of direct interest (nuisance parameters), but are nevertheless needed in the model.
- No problem: just integrate them out (marginalize with respect to, average over) all nuisance parameters.
- Example: $\theta = (\theta_1, \theta_2)'$, where θ_2 is a nuisance. We are interested in the marginal posterior of θ_1

$$p(\theta_1|y) \quad = \quad \int p(\theta_1, \theta_2|y) d\theta_2 = \int p(\theta_1|\theta_2, y) p(\theta_2|y) d\theta_2.$$

Normal model with unknown variance - Uniform prior

- Model:

$$y, ..., y_n \overset{iid}{\sim} N(\mu, \sigma^2)$$

- Prior

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$$

- Posterior:

$$\mu|\sigma^2, y \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right)$$
$$\sigma^2|y \sim \mathrm{Inv} - \chi^2(n-1, s^2),$$

where

$$s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$$

is the usual sample variance.

Normal model with unknown variance - Uniform prior, cont.

- Simulating the posterior of the normal model with non-informative prior:
  1. Draw $X \sim \chi^2(n-1)$
  2. Compute $\sigma^2 = \frac{(n-1)s^2}{X}$ (this a draw from $\text{Inv-}\chi^2(n-1, s^2)$)
  3. Draw a μ from $N\left(\bar{y}, \frac{\sigma^2}{n}\right)$ conditional on the previous draw σ^2
  4. Repeat step 1-3 many times.
- The sampling is implemented in the R program NormalNonInfoPrior.R
- We may derive the marginal posterior analytically as

$$\mu | y \sim t_{n-1}\left(\bar{y}, \frac{s^2}{n}\right).$$

Multinomial model with Dirichlet prior

- *Data*: $y = (y_1, ...y_K)$, where y_k counts the number of observations in the kth category. $\sum_{k=1}^K y_k = n$. Example: brand choices.
- Multinomial model:

$$p(y|\theta) \propto \prod_{k=1}^{K} \theta_k^{y_k}, \text{ where } \sum_{k=1}^{K} \theta_j = 1.$$

- *Conjugate prior*: $\mathrm{Dirichlet}(\alpha_1, ..., \alpha_K)$

$$p(\theta) \propto \prod_{k=1}^{K} \theta_j^{\alpha_j - 1}.$$

- Moments of $\theta = (\theta_1, ..., \theta_K)' \sim \mathit{Dirichlet}(\alpha_1, ..., \alpha_K)$

$$\mathrm{E}(\theta_k) = \frac{\alpha_k}{\sum_{j=1}^K \alpha_j}$$

$$\mathrm{V}(\theta_k) = \frac{\mathrm{E}(\theta_k)\left[1 - \mathrm{E}(\theta_k)\right]}{1 + \sum_{k=1}^K \alpha_k}$$

- Note that $\sum_{k=1}^K \alpha_k$ is the precision (inverse variance).

Multinomial model with Dirichlet prior, cont.

- 'Non-informative': $\alpha_1 = ... = \alpha_K = 1$ (uniform and proper).
- Simulating from the Dirichlet distribution:
  - Generate $x_1 \sim Gamma(\alpha_1, \beta), ..., x_K \sim Gamma(\alpha_K, \beta)$, independently. Any β will do as long it is the same for all x_i.
  - Compute $y_k = x_k / (\sum_{j=1}^{K} x_j)$.
  - $y = (y_1, ..., y_K)$ is a draw from the $\mathrm{Dirichlet}(\alpha_1, ..., \alpha_K)$ distribution.
- *Prior-to-Posterior updating*:

$$\begin{aligned}
\textit{Model}: \quad & y = (y_1, ...y_K) \sim \mathrm{Multin}(n; \theta_1, ..., \theta_K) \\
\textit{Prior}: \quad & \theta = (\theta_1, ..., \theta_K) \sim \mathrm{Dirichlet}(\alpha_1, ..., \alpha_K) \\
\textit{Posterior}: \quad & \theta | y \sim \mathrm{Dirichlet}(\alpha_1 + y_1, ..., \alpha_K + y_K).
\end{aligned}$$

Multivariate normal - known covariance matrix

- Model:
$$y_1, ..., y_n \overset{iid}{\sim} N_p(\mu, \Sigma)$$
where Σ is a known covariance matrix.
- Density
$$p(y|\mu, \Sigma) = |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right)$$
- Likelihood:
$$\begin{aligned} p(y_1, ..., y_n|\mu, \Sigma) &\propto |\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)'\Sigma^{-1}(y_i-\mu)\right) \\ &= |\Sigma|^{-n/2} \exp\left(-\frac{1}{2} tr\Sigma^{-1} S_\mu\right) \end{aligned}$$
where $S_\mu = \sum_{i=1}^{n}(y_i-\mu)(y_i-\mu)'$.

Multivariate normal - known covariance matrix, cont.

- Prior:

$$\mu \sim N_p(\mu_0, \Lambda_0)$$

- Posterior:

$$\mu \sim N(\mu_n, \Lambda_n)$$

  where

$$\begin{aligned} \mu_n &= (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y}) \\ \Lambda_n^{-1} &= \Lambda_0^{-1} + n\Sigma^{-1} \end{aligned}$$

- Note how the posterior mean is (matrix) weighted average of prior and data information.
- Noninformative prior: let the precision go to zero: $\Lambda^{-1} \to 0$.

Multivariate normal - Conjugate prior

- Conjugate prior is Normal-IW$(\mu_0, \kappa_0, \Lambda_0, \nu_0)$

$$\begin{aligned} \Sigma &\sim Inv-Wishart(\Lambda_0, \nu_0) \\ \mu|\Sigma &\sim N(\mu_0, \kappa_0^{-1}\Sigma) \end{aligned}$$

- Density:

$$|\Sigma|^{-[(\nu_0+d)/2+1]} \exp\left(-\frac{1}{2}tr(\Lambda_0\Sigma^{-1}) - \frac{\kappa_0}{2}(\mu-\mu_0)'\Sigma^{-1}(\mu-\mu_0)\right)$$

- Posterior is also Normal IW

$$\begin{aligned} \mu_n &= \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y} \\ \kappa_n &= \kappa_0 + n \\ \nu_n &= \nu_0 + n \\ \Lambda_n &= \Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)(\bar{y}-\mu_0)' \end{aligned}$$

Lecture overview

- Bayesian prediction
- Decision theory

Prediction

- We may use the estimated model for forecasting a future observation $\tilde{y}$.
- *Posterior predictive distribution* (y denotes available data at the time of forecasting)

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

  where the last step holds if $p(\tilde{y}|\theta, y) = p(\tilde{y}|\theta)$.
- The uncertainty that comes from not knowing θ is represented in $p(\tilde{y}|y)$ by averaging over $p(\theta|y)$.

Prediction Bernoulli data

- Let $y = \sum_{i=1}^{n} y_i$ and $\tilde{y}$ the outcome of the next trial

$$
\begin{aligned}
p(\tilde{y} &= 1|y) = \int_\theta p(\tilde{y} = 1|\theta) p(\theta|y) d\theta \\
&= \int_\theta \theta p(\theta|y) d\theta = E_{\theta|y}(\theta) = \frac{\alpha + y}{\alpha + \beta + n}.
\end{aligned}
$$

- Uniform prior ($\alpha = \beta = 1$)

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

Prediction Normal data with known variance

- Assume the uniform prior $p(\theta) \propto c$.

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

where

$$\begin{aligned} \theta|y &\sim N(\bar{y}, \sigma^2/n) \\ \tilde{y}|\theta &\sim N(\theta, \sigma^2) \end{aligned}$$

Simulate from the predictive distribution - Normal model

1. Generate a posterior draw of θ ($\theta^{(1)}$) from $N(\bar{y}, \sigma^2/n)$
2. Generate a draw of $\tilde{y}$ ($\tilde{y}^{(1)}$) from $N(\theta^{(1)}, \sigma^2)$ (note the mean)
3. Repeat steps 1 and 2 a large number of times (N) with the result:
   - Sequence of posterior draws: $\theta^{(1)}, ..., \theta^{(N)}$
   - Sequence of predictive draws: $\tilde{y}^{(1)}, ..., \tilde{y}^{(N)}$.

Predictive distribution - Normal model and uniform prior

- $\theta^{(1)} = \bar{y} + \varepsilon^{(1)}$, where $\varepsilon^{(1)} \sim N(0, \sigma^2/n)$. (Step 1).
- $\tilde{y}^{(1)} = \theta^{(1)} + \upsilon^{(1)}$, where $\upsilon^{(1)} \sim N(0, \sigma^2)$. (Step 2).
- $\tilde{y}^{(1)} = \bar{y} + \varepsilon^{(1)} + \upsilon^{(1)}$.
- $\varepsilon^{(1)}$ and $\upsilon^{(1)}$ are independent.
- The sum of two normal random variables follows a normal distribution, so $\tilde{y}$ follows a normal distribution with

$$
\begin{aligned}
E(\tilde{y}|y) &= E(\tilde{y}|y) = \bar{y} \\
V(\tilde{y}|y) &= \frac{\sigma^2}{n} + \sigma^2 = \sigma^2\left(1 + \frac{1}{n}\right).
\end{aligned}
$$

- Note that the estimation uncertainty (σ^2/n) is typically much less important than the intrinsic population uncertainty, σ^2.

Predictive distribution - Normal model and normal prior

- It easy to see that the predictive distribution is normal.
- The mean can be obtained from

$$E_{\tilde{y}|\theta}(\tilde{y}|\theta) = \theta$$

  and then remove the conditioning on θ by averaging over θ

$$E(\tilde{y}|y) = E_{\theta|y}(\theta) = \mu_n \text{ (Posterior mean of } \theta).$$

- The predictive variance of $\tilde{y}$ can be obtained from the conditional variance formula

$$\begin{aligned} V(\tilde{y}|y) &= E_{\theta|y}[V_{\tilde{y}|\theta}(\tilde{y}|\theta)] + V_{\theta|y}[E_{\tilde{y}|\theta}(\tilde{y}|\theta)] \\ &= E_{\theta|y}(\sigma^2) + V_{\theta|y}(\theta) \\ &= \sigma^2 + \tau_n^2 \\ &= \text{(Population variance + Posterior variance of } \theta). \end{aligned}$$

- In summary:

$$\tilde{y}|y \sim N(\mu_n, \sigma^2 + \tau_n^2).$$

Decision Theory

- Let θ be an unknown quantity. State of nature. Examples: Future inflation, Global temperature, Disease.
- Let $a \in \mathcal{A}$ be an action. Ex: Interest rate, Energy tax, Operation.
- Choosing action a when state of nature turns out to be θ gives utility

$$U(a, \theta)$$

- Alternatively loss $L(a, \theta) = -U(a, \theta)$.
- Loss table:

	θ_1	θ_2
a_1	$L(a_1, \theta_1)$	$L(a_1, \theta_2)$
a_2	$L(a_2, \theta_1)$	$L(a_2, \theta_2)$

- Example utility functions:
  - Linear: $L(a, \theta) = |a - \theta|$
  - Quadratic: $L(a, \theta) = (a - \theta)^2$
  - Lin-Lin:

$$L(a, \theta) = \begin{cases} c_1 & \text{if } a \leq \theta \\ c_2 & \text{if } a > \theta \end{cases}$$

Optimal decision

- Ad hoc decision rules:
  - *Minimax*. Choose the decision that minimizes the maximum loss.
  - *Minimax-regret*: Choose the decision rule that gives you least regret when you eventually find out the true value of θ.
- Bayesian axiomatic theory gives you the rule: Choose the action that maximizes the (posterior) expected utility:

$$a_{bayes} = \mathrm{argmax}_{a \in \mathcal{A}} \; E_{p(\theta|y)}[L(a, \theta)],$$

where $E_{p(\theta|y)}$ denotes the posterior expectation.

- Using simulated draws $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ from $p(\theta|y)$:

$$E_{p(\theta|y)}[L(a, \theta)] \approx N^{-1} \sum_{i=1}^{N} L(a, \theta^{(i)})$$

- *Separation principle*: The analysis of uncertainty (i.e. the posterior of θ) is completely separated from the utilities of the choices.

Point and interval estimation

- Choosing a point estimator is a decision problem.
- Which to choose: posterior median, mean or mode?
- It depends on your loss function:
  - Linear loss $\rightarrow$ Posterior median is optimal
  - Quadratic loss $\rightarrow$ Posterior mean is optimal
  - Lin-Lin loss $\rightarrow c_1/(c_1+c_2)$ quantile of the posterior is optimal
  - Zero-one loss $\rightarrow$ Posterior mode is optimal
- Similar analysis can be used to select interval type: symmetric or HPD?

Combining expert judgement

- Available: K unbiased expert forecast/judgement: $\mathbf{y} = (y_1, y_2, ..., y_K)'$ and
$$\mathbf{y}|\theta \sim N(\theta \cdot \mathbf{1}, \Sigma)$$
where Σ is assumed to be known.
- Assuming a uniform prior for θ, the posterior distribution is of the form
$$\theta|\mathbf{y} \sim N(\mathbf{w}'\mathbf{y}, \psi^2)$$
where
$$\mathbf{w} = \frac{\mathbf{1}' \cdot \Sigma^{-1}}{\mathbf{1}' \cdot \Sigma^{-1} \cdot \mathbf{1}}$$
$$\psi = \frac{1}{\sqrt{\mathbf{1}' \cdot \Sigma^{-1} \cdot \mathbf{1}}}$$
- Weights can be negative, and it makes sense!
- Correlation between expert's forecasts are important. A poor expert can ge a sizeable weight if she is negatively correlated with the rest.
- Estimate Σ with past forecast errors. Difficulty: estimate Σ precisely. A decent prior on Σ can help!