Bayesian Learning - Lecture 1

Mattias Villani

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Course overview

- Three audiences:
  - **Master students in Statistics and Data Mining** (732A91)
  - **Engineering students** (TDDE07)
  - a few **PhD students**
- Course **webpage** is **here**. Course **syllabus** is **here**.
- Modes of teaching:
  - **Lectures** (Mattias Villani)
  - **Mathematical exercises** (Per Sidén)
  - **Computer labs** (Per Sidén)
- **Modules**:
  - The **Bayesics**, single- and multiparameter models
  - **Regression** and **Classification models**
  - **Advanced models** and **Posterior Approximation** methods
  - **Model Inference, Model evaluation** and **Variable Selection**
- **Examination**
  - Lab reports, 3 credits
  - Computer exam (using R), 3 credits

Lecture overview

- The **likelihood function**
- **Bayesian inference**
- **Bernoulli model**
- **Normal model** with known variance

THE LIKELIHOOD FUNCTION - BERNOULLI TRIALS

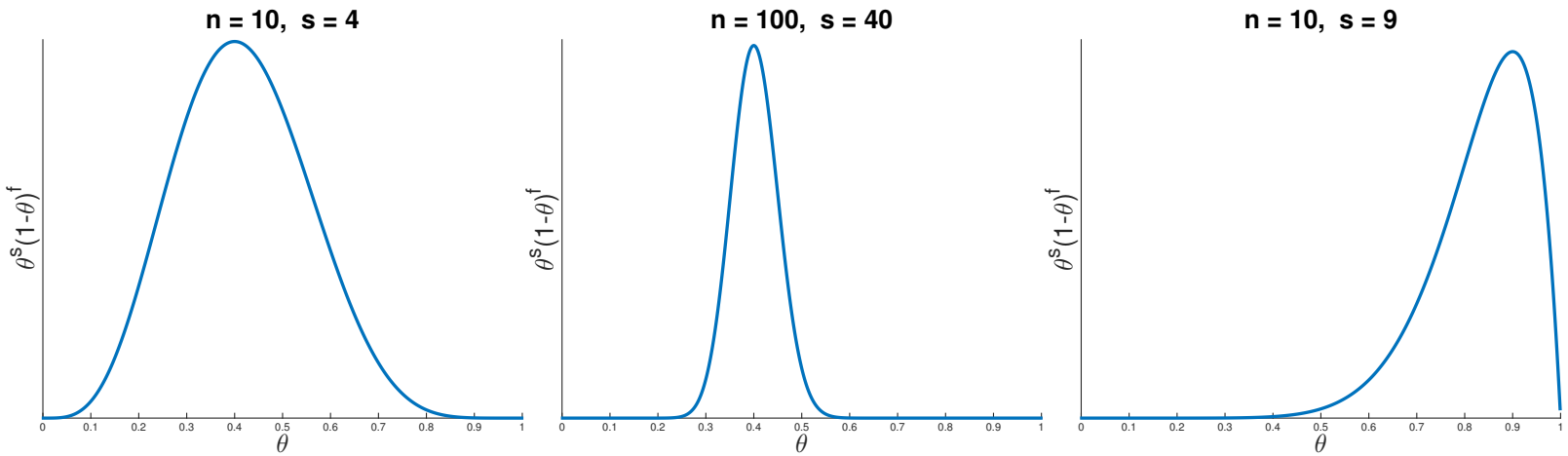
- **Bernoulli trials**:

$$X_1, ..., X_n|\theta \overset{iid}{\sim} Bern(\theta).$$

- **Likelihood** from $s = \sum_{i=1}^{n} x_i$ successes and $f = n - s$ failures.

$$p(x_1, ..., x_n|\theta) \quad = \quad p(x_1|\theta) \cdots p(x_n|\theta) = \theta^s(1-\theta)^f$$

- **Maximum likelihood estimator** $\hat{\theta}$ maximizes $p(x_1, ..., x_n|\theta)$.
- Given the data $x_1, ..., x_n$, we may plot $p(x_1, ..., x_n|\theta)$ as a function of θ.



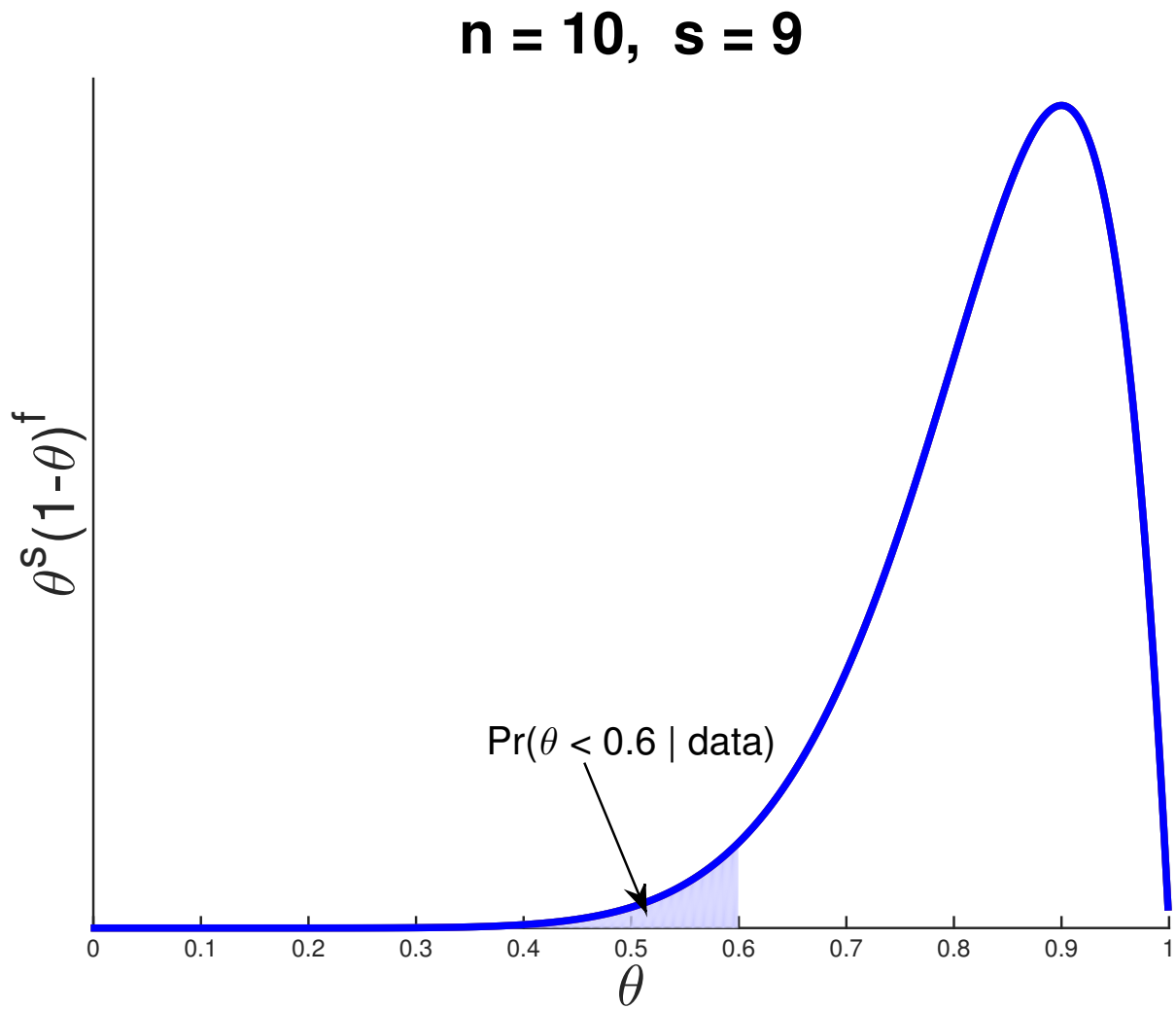
THE LIKELIHOOD FUNCTION

- Say it out loud:

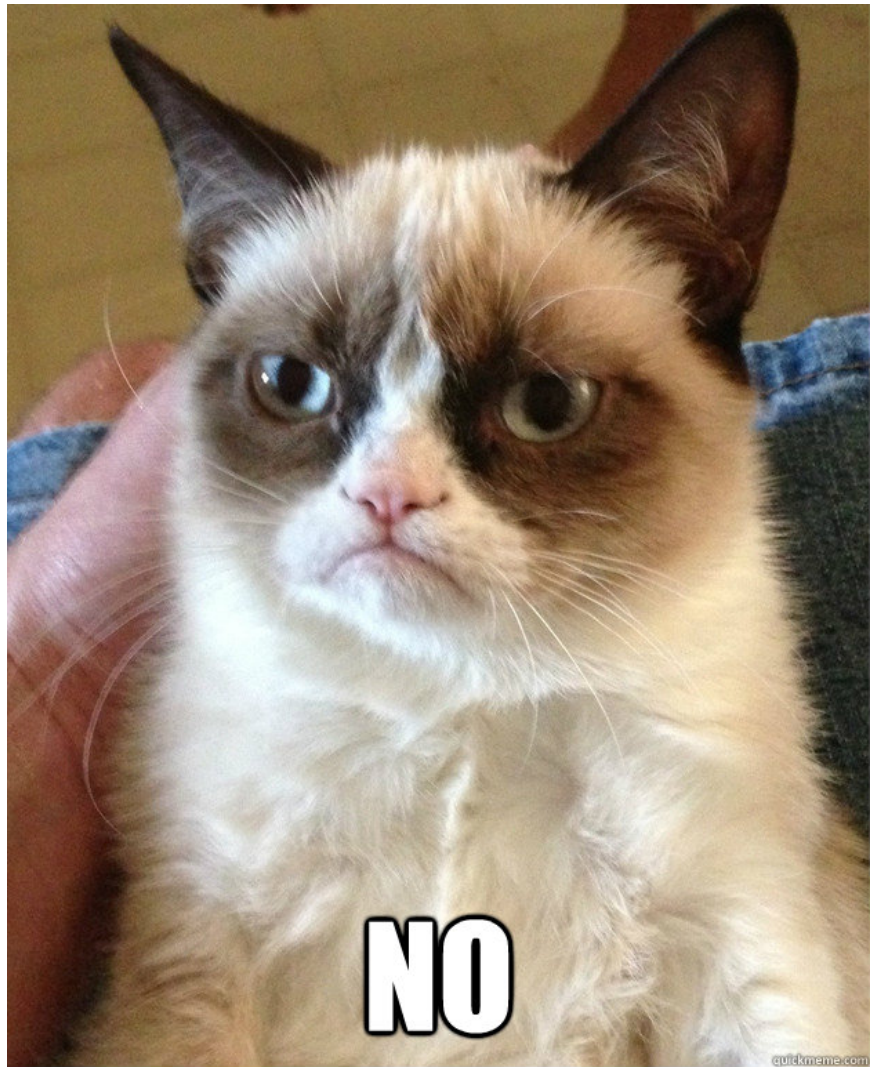
  The likelihood function is
  the probability of the observed data
  considered as a function of the parameter.

- The symbol $p(x_1, ..., x_n|\theta)$ plays two different roles:
- **Probability distribution** for the data.
  - The data $\mathbf{x} = (x_1, ..., x_n)$, are random.
  - θ is fixed.
- **Likelihood function** for the parameter
  - The data $\mathbf{x} = (x_1, ..., x_n)$ are fixed.
  - $p(x_1, ..., x_n|\theta)$ is function of θ.

PROBABILITIES FROM THE LIKELIHOOD!!

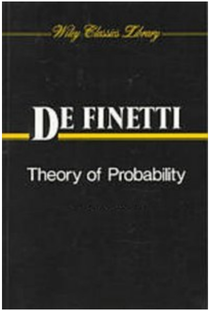


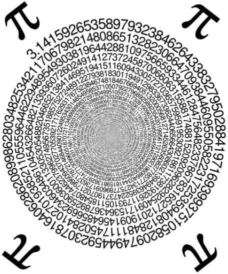
PROBABILITIES FROM THE LIKELIHOOD!!



UNCERTAINTY AND SUBJECTIVE PROBABILITY

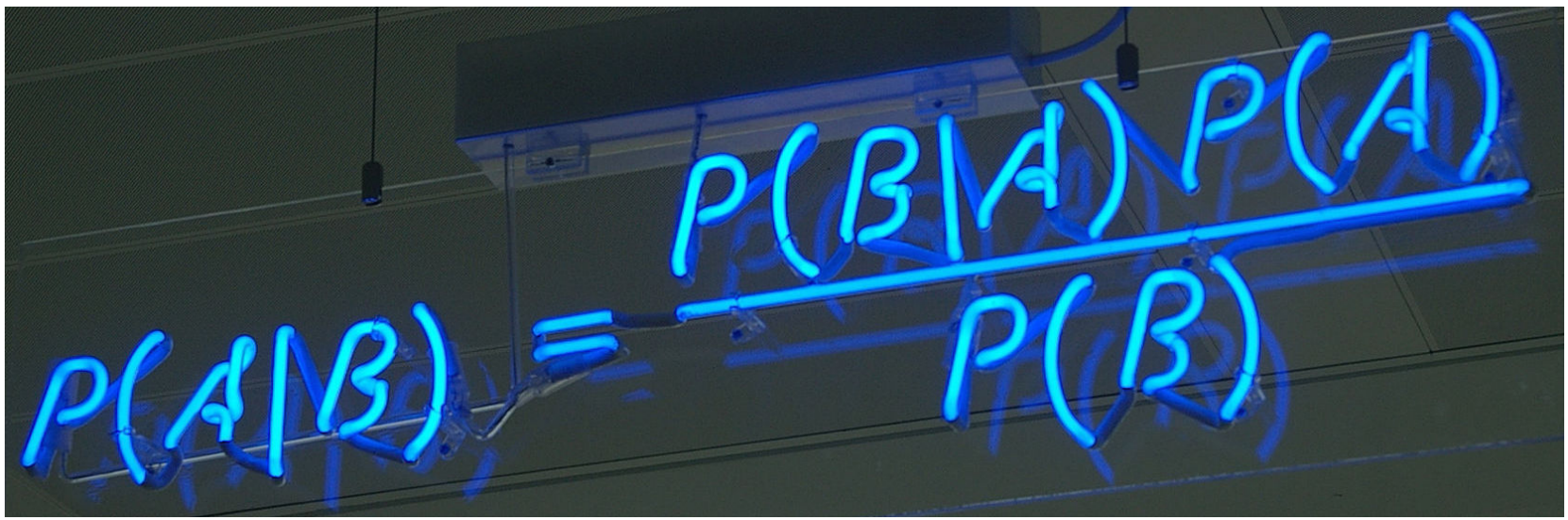
- Statements like $\Pr(\theta < 0.6|\text{data})$ only make sense if θ is random.
- But θ may be a fixed natural constant?
- **Bayesian: doesn't matter if θ is fixed or ('intrinsically') random.**
- Do **You** know the value of θ or not?
- $p(\theta)$ reflects Your knowledge/**uncertainty** about θ.
- **Subjective probability**.
- The statement $p(\text{10th decimal of } \pi = 9) = 0.1$ makes sense.





Bayesian learning

- **Bayesian learning** about a model parameter θ:
  - state your **prior** knowledge about θ as a probability distribution $p(\theta)$.
  - **collect data x** and form the **likelihood** function $p(\mathbf{x}|\theta)$.
  - **combine** your prior knowledge $p(\theta)$ with the data information $p(\mathbf{x}|\theta)$.
- **How to combine** the two sources of information? **Bayes' theorem**.



Learning from data - Bayes' theorem

- How do we **update** from the **prior** $p(\theta)$ to the **posterior** $p(\theta|Data)$?
- **Bayes' theorem** for events A and B

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

- Bayes' Theorem for a model parameter θ

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

- The prior $p(\theta)$ is the hero that converts the likelihood function $p(Data|\theta)$ into a posterior probability density $p(\theta|Data)$.
- A probability distribution for θ is extremely useful. **Decision making**.
- **No prior - no posterior - no useful inferences - no fun**.

Bayes' theorem for medical diagnosis

- A = {Horrible and very rare disease}, B ={Positive medical test}.
- $p(A) = 0.0001$. $p(B|A) = 0.9$. $p(B|A^c) = 0.05$.
- Probability of being sick given a positive test:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|A^c)p(A^c)} \approx 0.001797.$$

- Probably not sick, but 18 times more probable than before the test.

- Morale of the story: If you want $p(A|B)$ then $p(B|A)$ does not tell the whole story. The prior probability $p(A)$ is also very important.

"You can't enjoy the Bayesian omelette without breaking the Bayesian eggs"

Leonard Jimmie Savage

The normalizing constant is not important

- Bayes theorem

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

- The integral $p(Data) = \int_\theta p(Data|\theta)p(\theta)d\theta$ can make you cry.
- $p(Data)$ 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

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

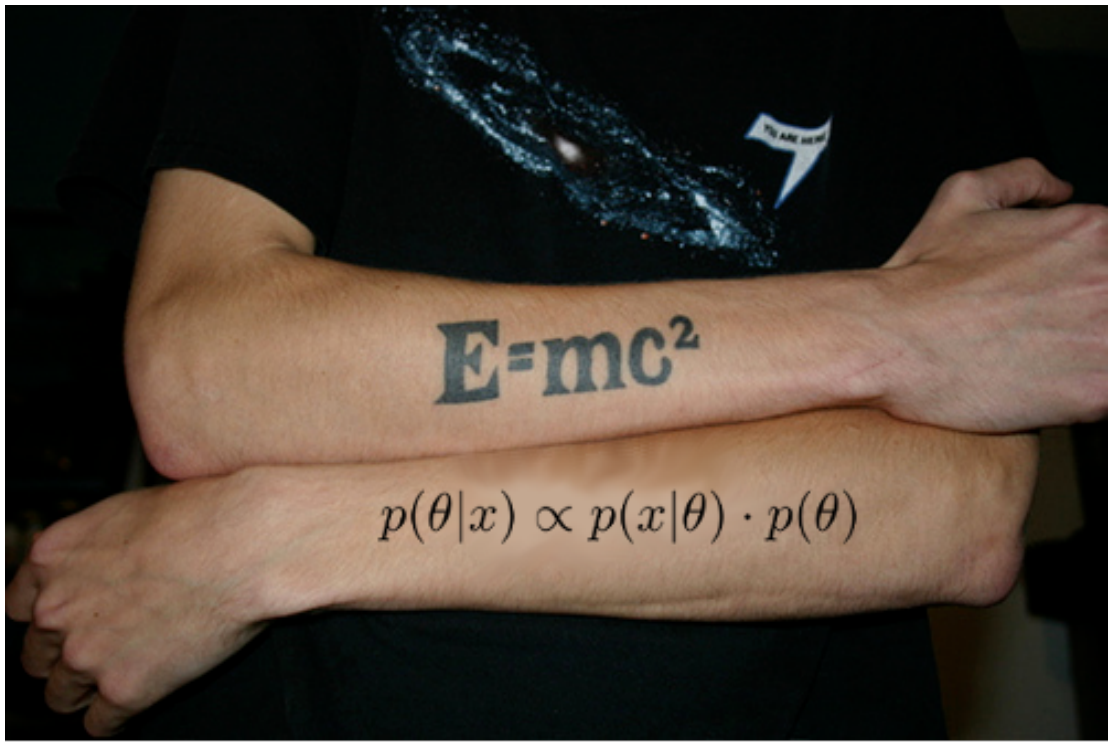
Great theorems make great tattoos

- All you need to know:

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

or

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



Bernoulli trials - Beta prior

- **Model**

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

- **Prior**

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

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

- **Posterior**

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

- This is proportional to the $Beta(\alpha+s, \beta+f)$ density.
- The **prior-to-posterior** mapping reads

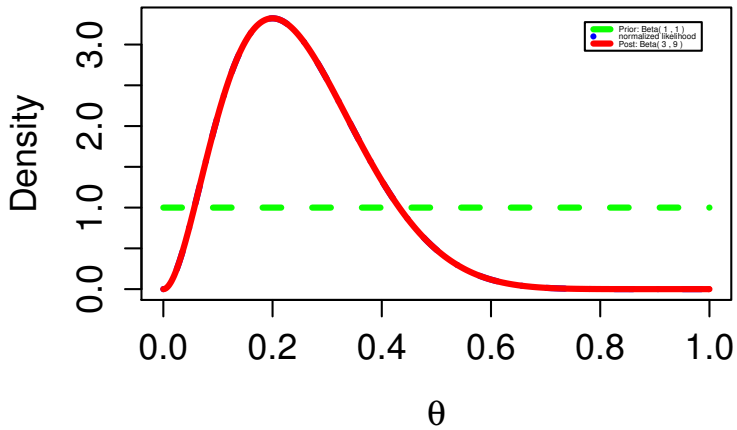
$$\theta \sim Beta(\alpha, \beta) \overset{x_1,...,x_n}{\Longrightarrow} \theta|x_1, ..., x_n \sim Beta(\alpha+s, \beta+f).$$

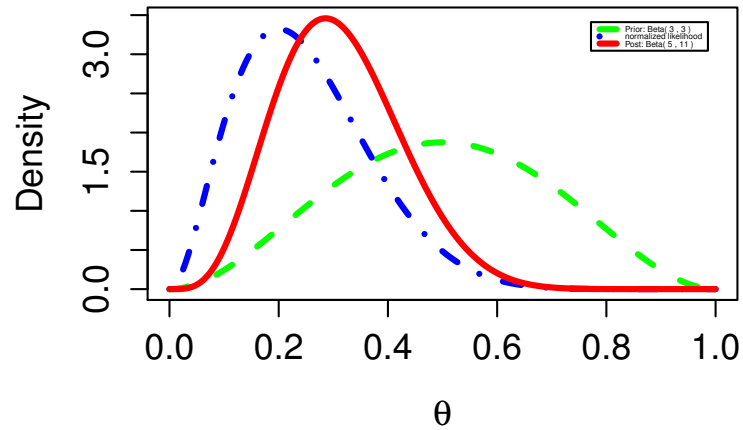
Bernoulli example: spam emails

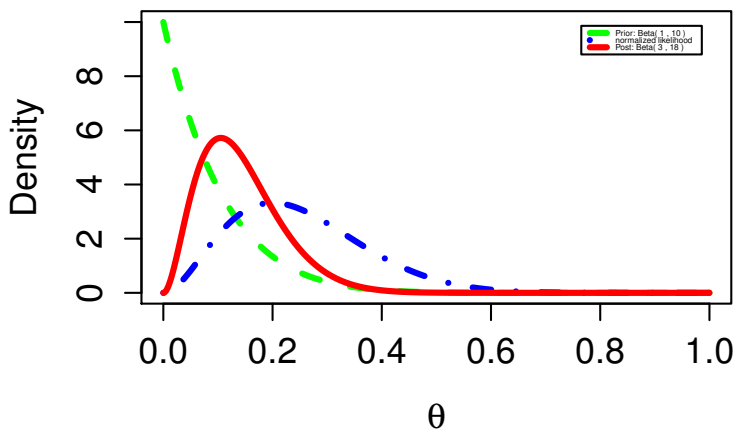
- George has gone through his collection of 4601 e-mails. He classified 1813 of them to be spam.
- Let $x_i = 1$ if i:th email is spam. Assume $x_i|\theta \overset{iid}{\sim} Bernoulli(\theta)$ and $\theta \sim \text{Beta}(\alpha, \beta)$.
- Posterior

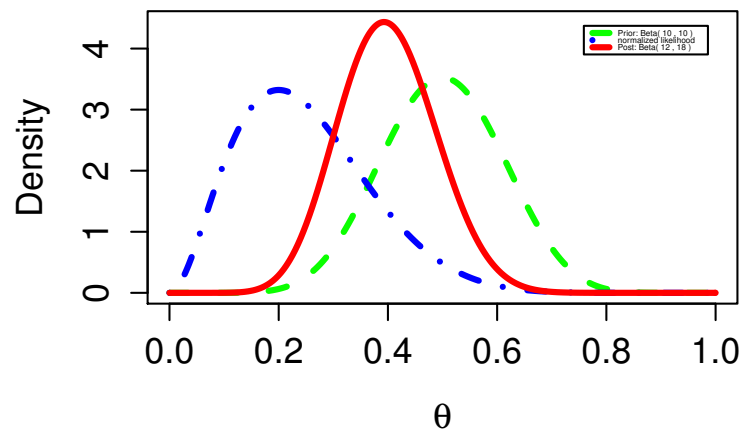
$$\theta|x \sim Beta(\alpha + 1813, \beta + 2788)$$

SPAM DATA (N=10): PRIOR SENSITIVITY

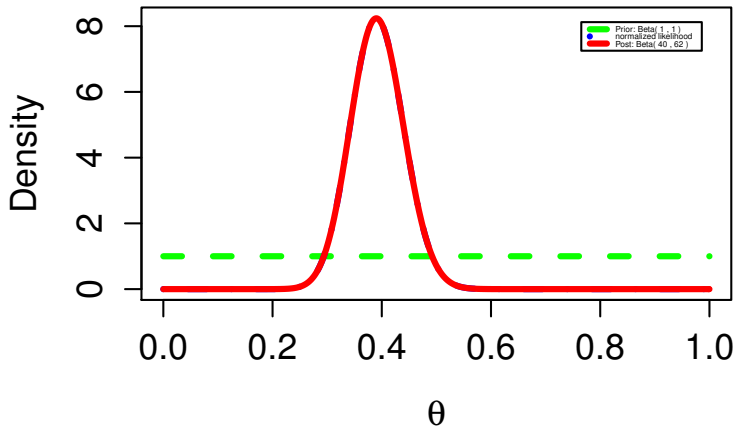


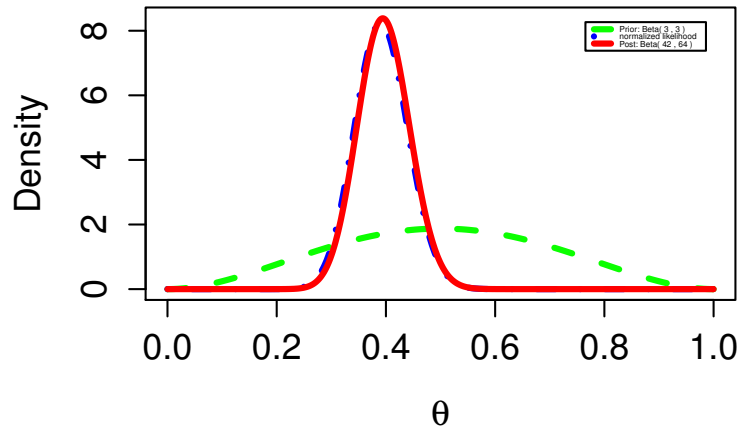


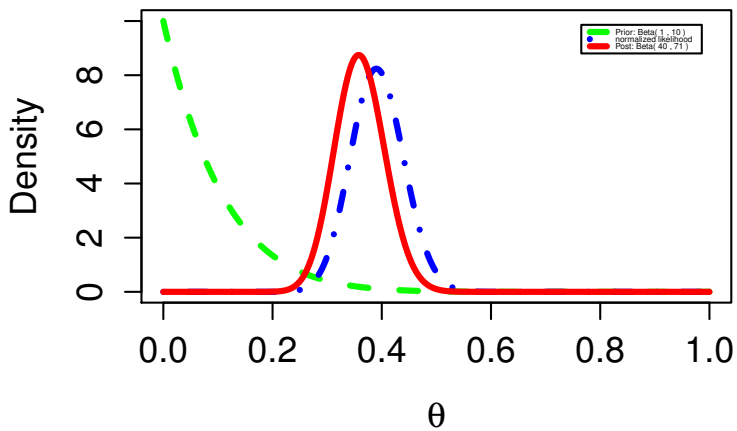


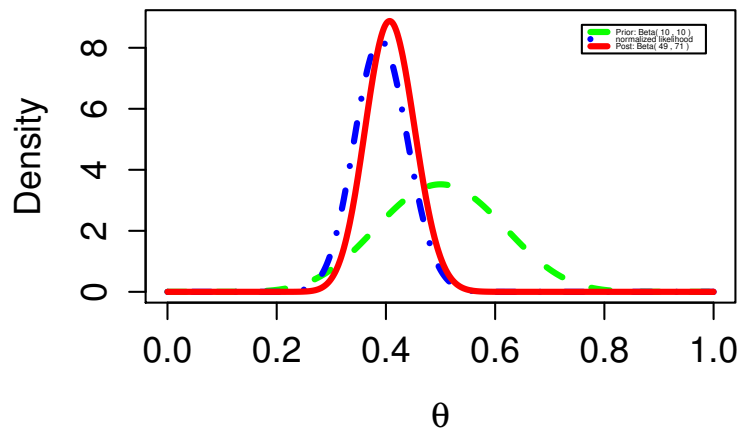


SPAM DATA (N=100): PRIOR SENSITIVITY

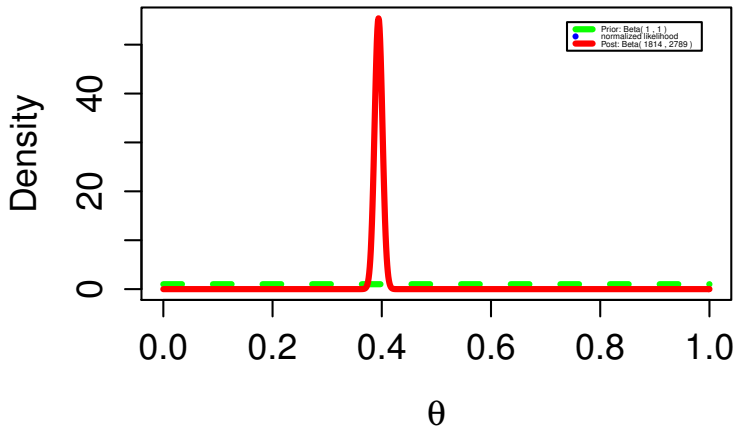


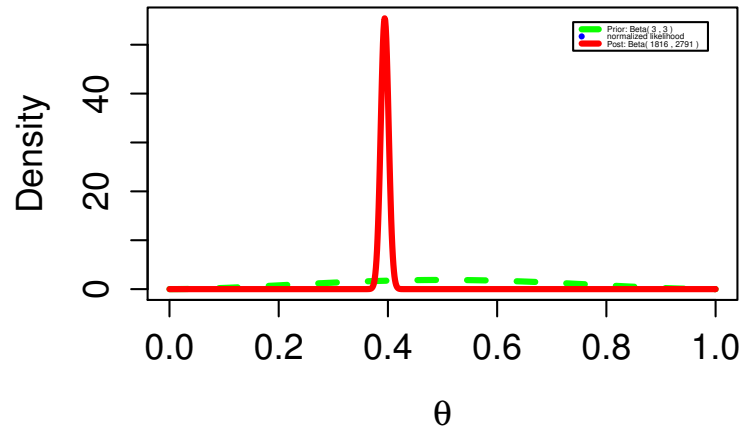


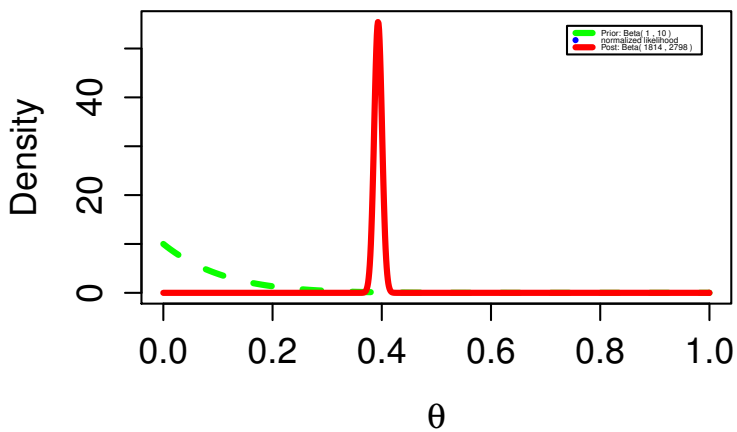


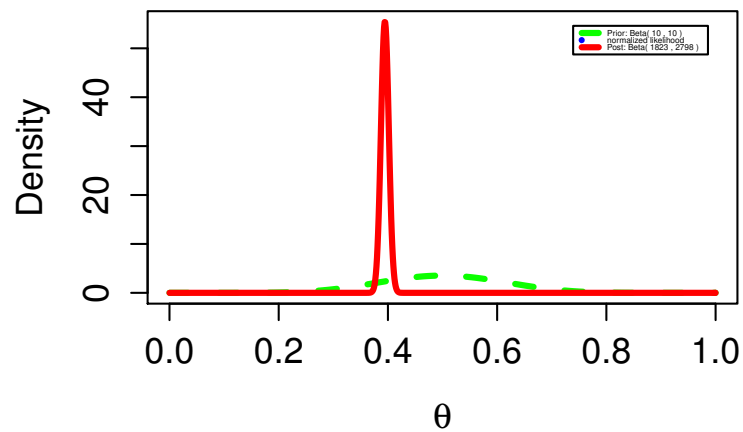


SPAM DATA (N=4601): PRIOR SENSITIVITY

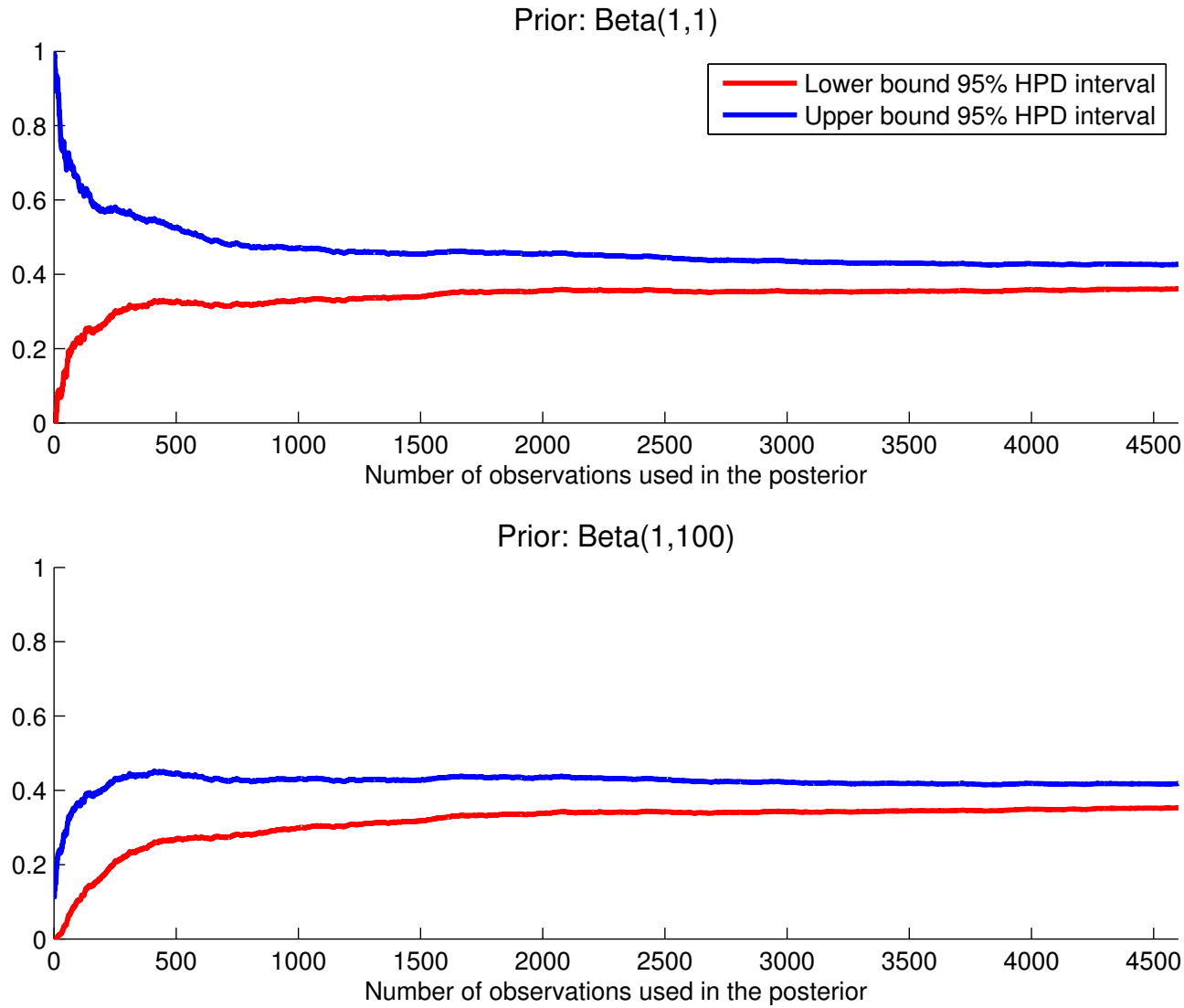








Spam data: Posterior convergence



NORMAL DATA, KNOWN VARIANCE - UNIFORM PRIOR

- Model:

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

- Prior:

$$p(\theta) \propto c \text{ (a constant)}$$

- Likelihood

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

- Posterior

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

Normal data, known variance - normal prior

- Prior

$$\theta \sim N(\mu_0, \tau_0^2)$$

- Posterior

$$\begin{aligned} p(\theta|x_1, ..., x_n) &\propto p(x_1, ..., x_n|\theta, \sigma^2)p(\theta) \\ &\propto N(\theta|\mu_n, \tau_n^2), \end{aligned}$$

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 data, known variance - normal prior

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

$$\text{Posterior precision} = \text{Data precision} + \text{Prior precision}$$

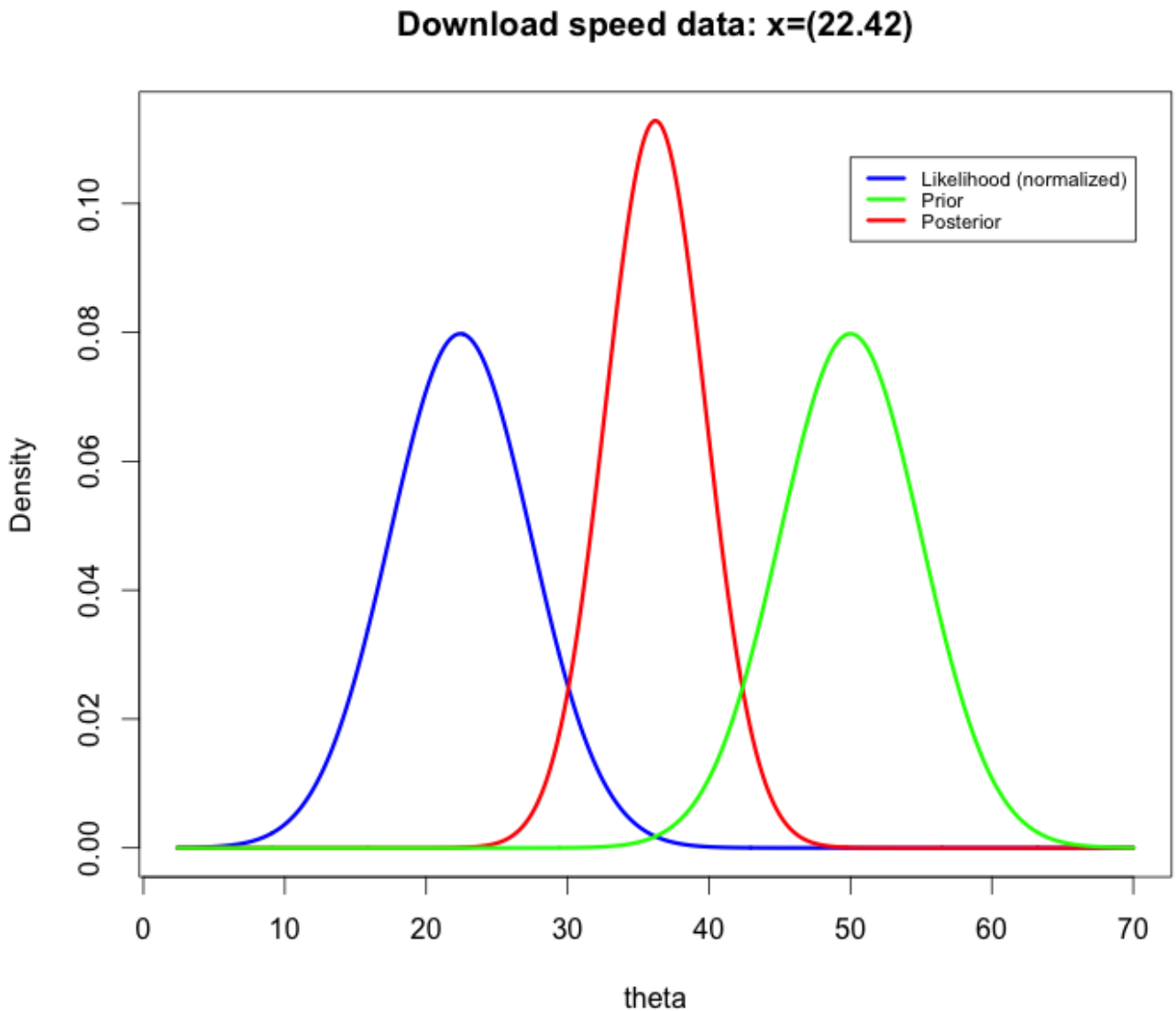
$$\text{Posterior mean} =$$

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

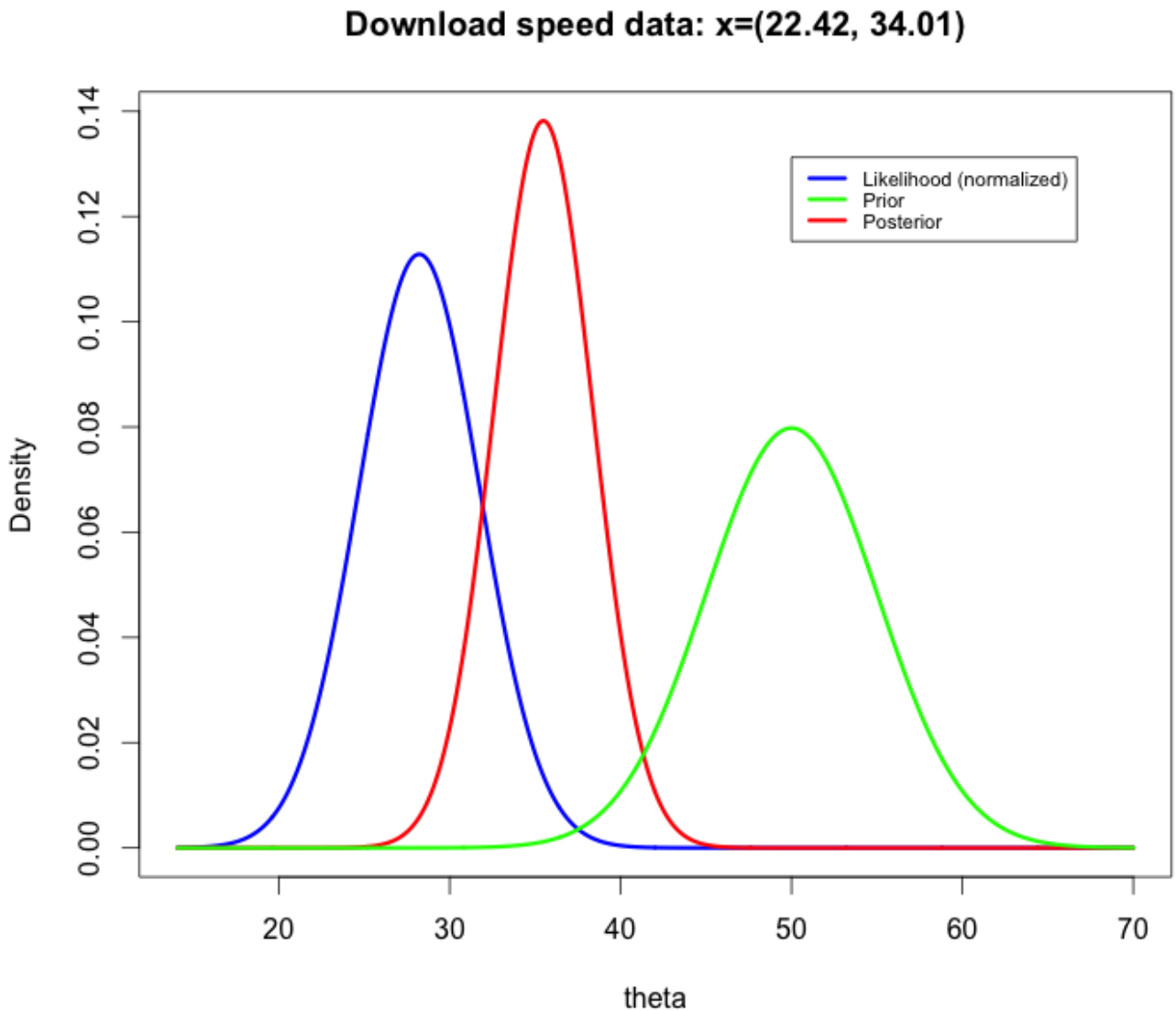
Download speed

- Data: $\mathrm{x} = (22.42, 34.01, 35.04, 38.74, 25.15)$ Mbit/sec.
- Model; $X_1, ..., X_5 \sim N(\theta, \sigma^2)$.
- Assume $\sigma = 5$ (measurements can vary $\pm$10MBit with 95% probability)
- My prior: $\theta \sim N(50, 5^2)$.

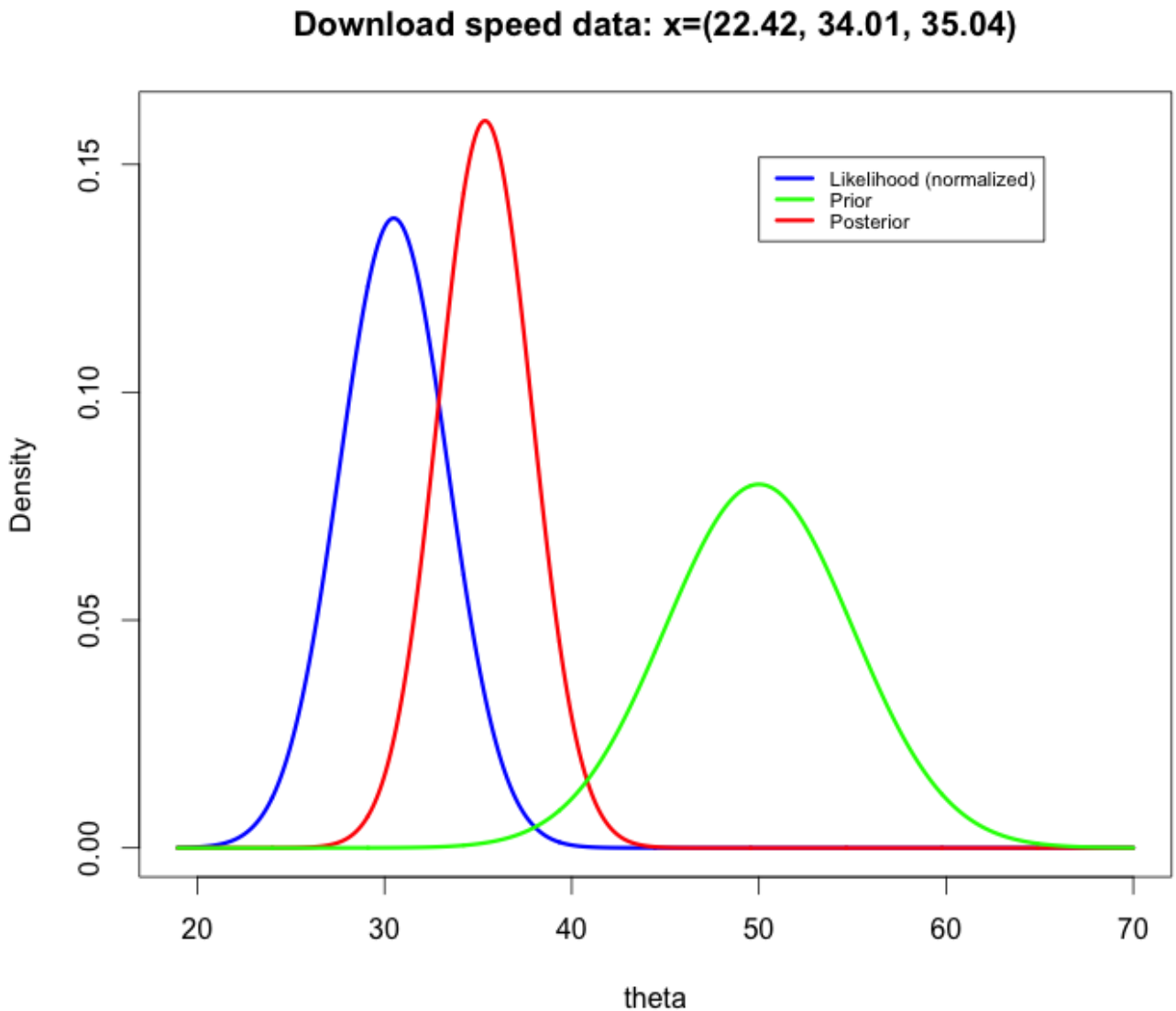
Download speed n=1



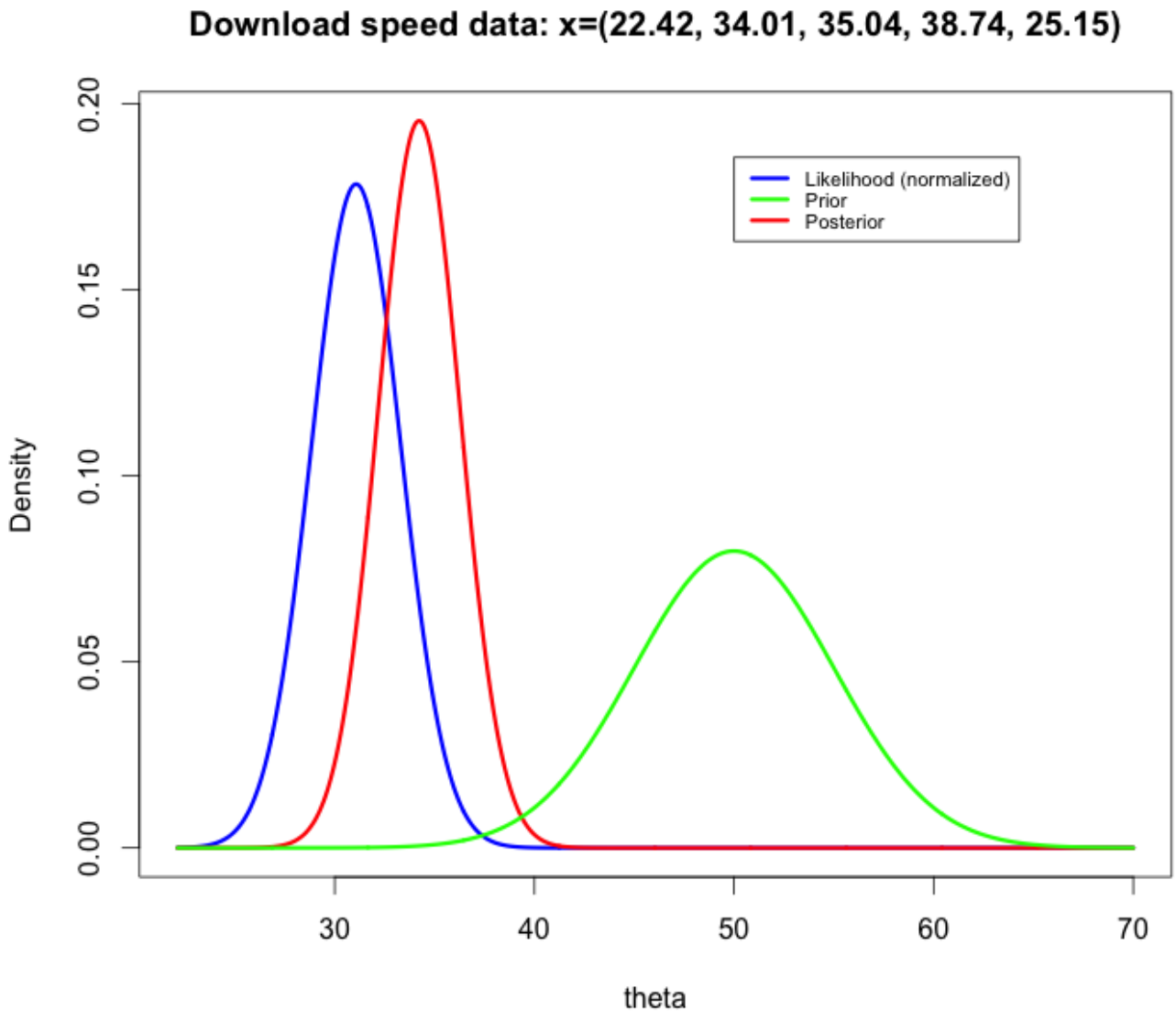
DOWNLOAD SPEED N=2



DOWNLOAD SPEED N=3

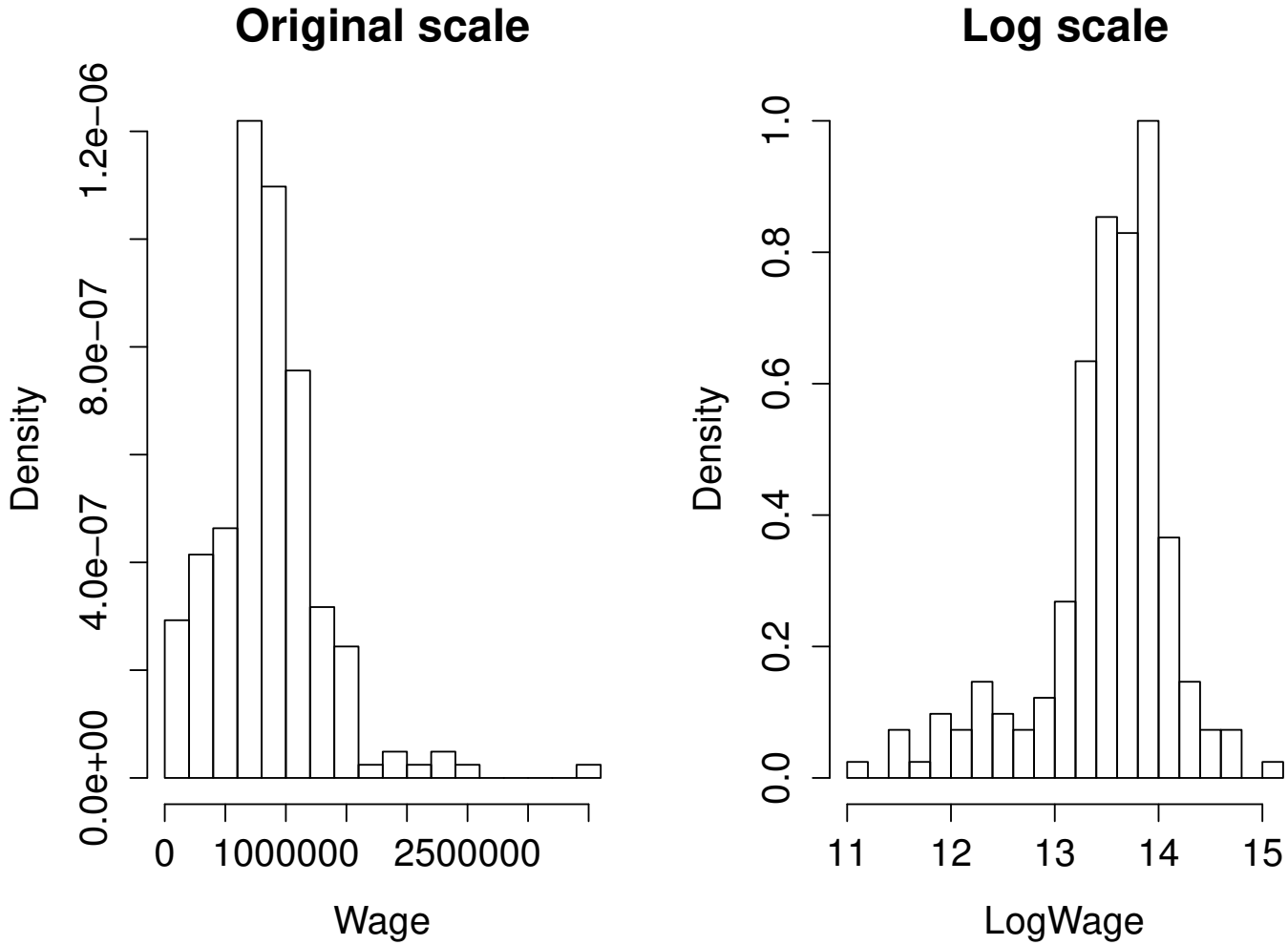


Download speed n=5



Canadian wages data

- Data on wages for 205 Canadian workers.



Canadian wages

- Model

$$X_1, ..., X_n|\theta \sim N(\theta, \sigma^2), \ \sigma^2 = 0.4$$

- Prior

$$\theta \sim N(\mu_0, \tau_0^2), \ \mu_0 = 12 \text{ and } \tau_0 = 10$$

- Posterior

$$\theta|x_1, ..., x_n \sim N\left(\mu_n, \tau_n^2\right),$$

where $\mu_n = w\bar{x} + (1-w)\mu_0$.

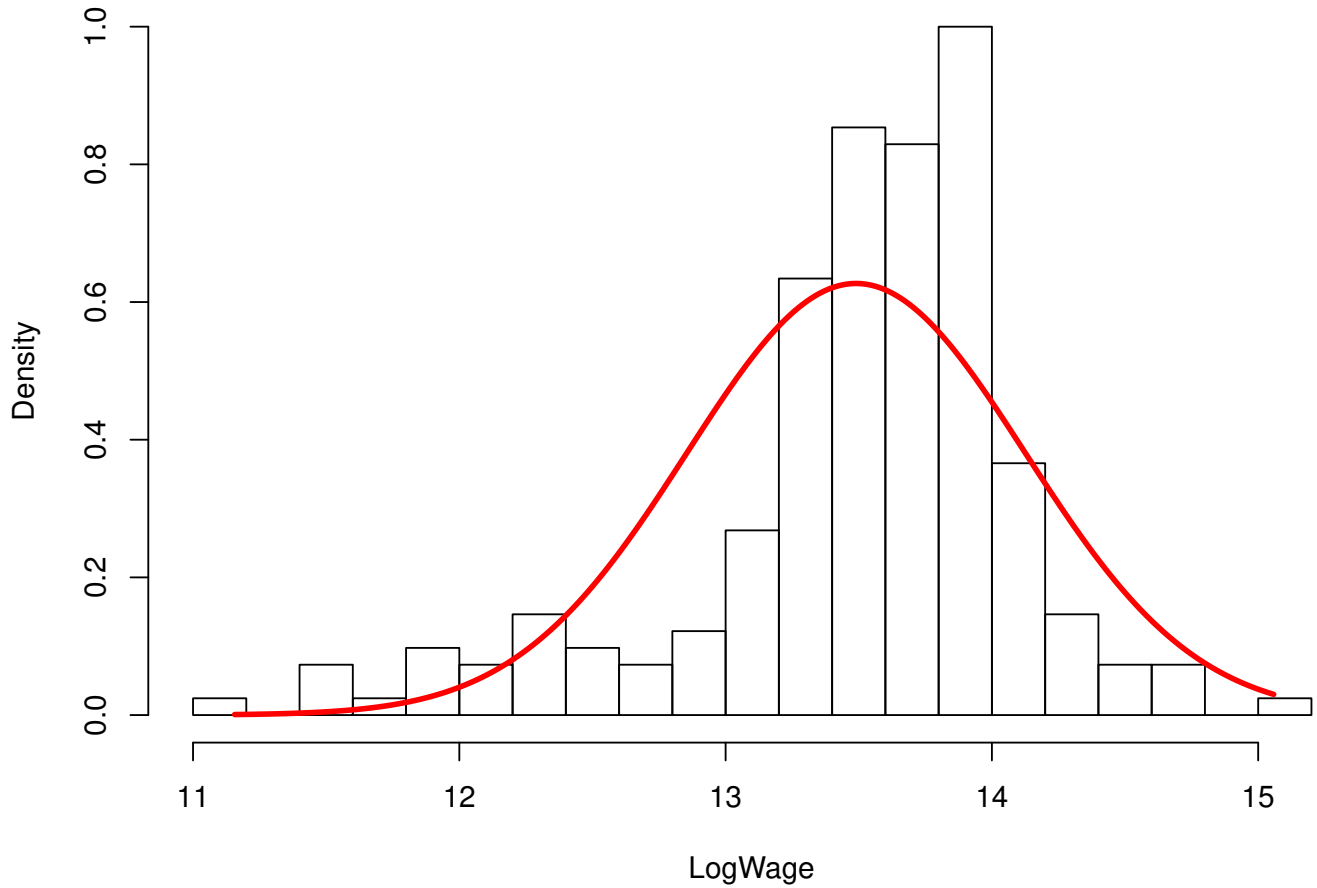
- For the Canadian wage data:

$$w = \frac{\sigma^{-2}n}{\sigma^{-2}n + \tau_0^{-2}} = \frac{2.5 \cdot 205}{2.5 \cdot 205 + 1/100} = 0.999.$$

$$\mu_n = w\bar{x} + (1-w)\mu_0 = 0.999 \cdot 13.489 + (1-0.999) \cdot 12 \approx 13.489$$

$$\tau_n^2 = (2.5 \cdot 205 + 1/100)^{-1} = 0.00195$$

CANADIAN WAGES DATA - MODEL FIT



Bayesian Learning - Lecture 2

Mattias Villani

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Lecture overview

- The **Poisson model**
- **Conjugate priors**
- **Prior elicitation** - how to come up with a prior.
- **Non-informative priors**

Poisson model

- **Model**:

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

- **Poisson distribution**

$$p(y) = \frac{\theta^y e^{-\theta}}{y!}$$

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

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

- ***Prior***:

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

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

Poisson model, cont.

- **Posterior**

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

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

- **Prior-to-Posterior mapping**:

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

Poisson example - 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=$\bar{y} = 537/576 \approx 0.9323$.

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

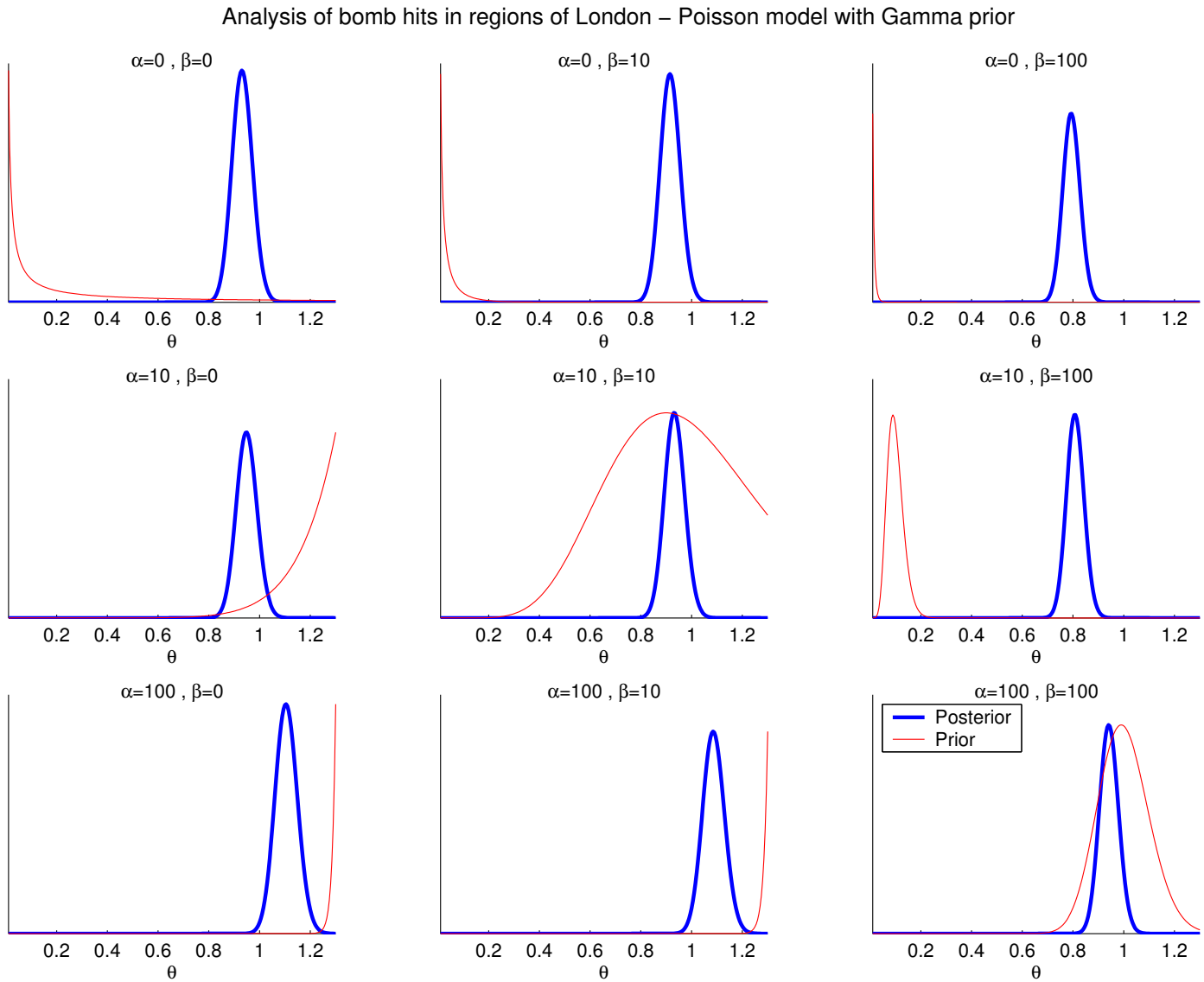
$$E(\theta|y) = \frac{\alpha + \sum_{i=1}^{n} y_i}{\beta + n} \approx \bar{y} \approx 0.9323,$$

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



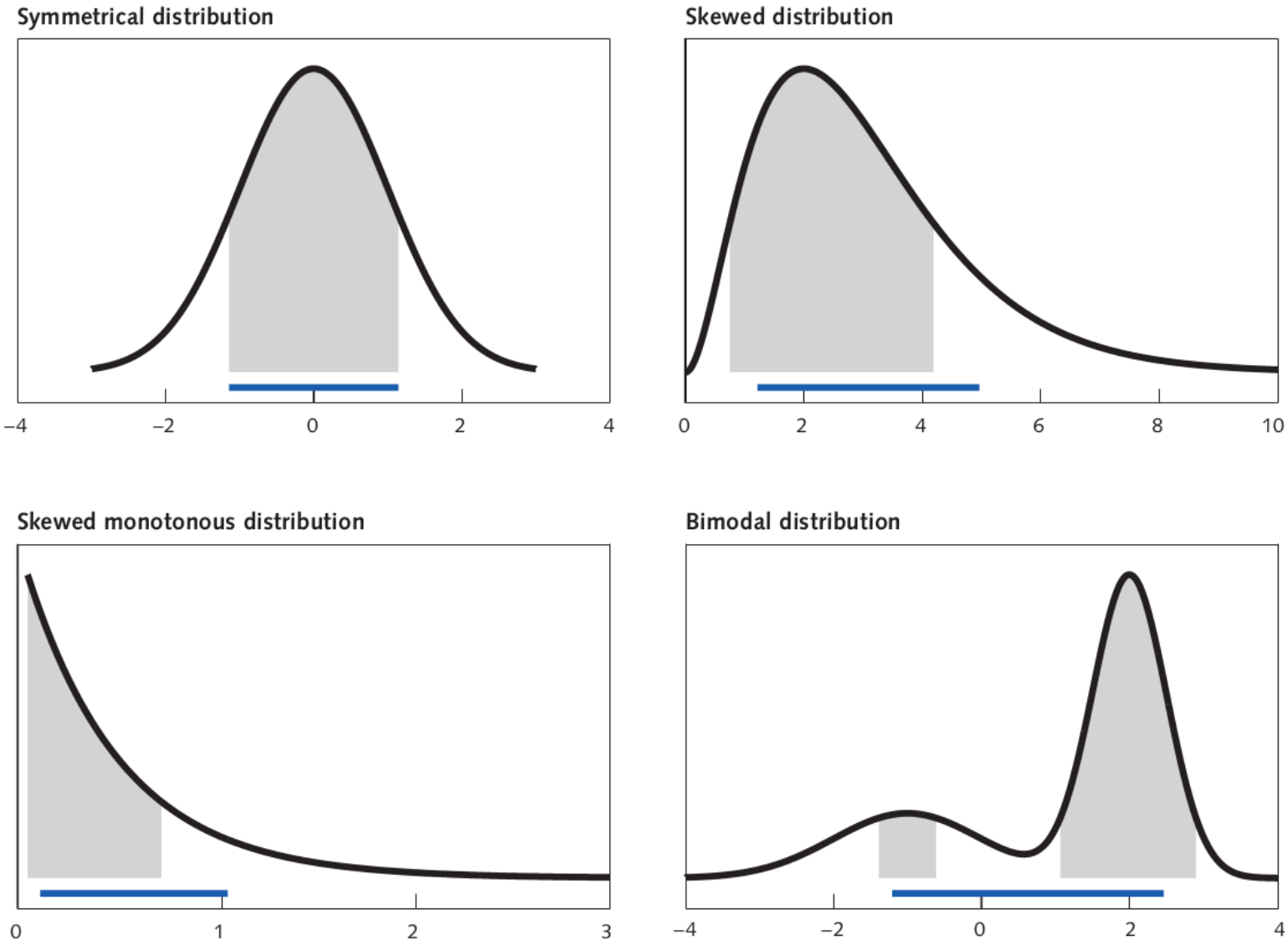
Poisson example - posterior intervals

- **Bayesian 95% credible 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 $\alpha = \beta = 0$)
- **Highest Posterior Density** (**HPD**) interval contains the θ values with highest pdf.
- An exact Highest Posterior Density (HPD) interval is $[0.8525; 1.0144]$. Obtained numerically, assuming $\alpha = \beta = 0$.

Illustration of different interval types



Conjugate priors

- Normal likelihood: Normal prior$\rightarrow$Normal posterior. (posterior belongs to the same distribution family as prior)
- Bernoulli likelihood: Beta prior$\rightarrow$Beta posterior.
- Poisson likelihood: Gamma prior$\rightarrow$Gamma posterior.
- **Conjugate priors**: A prior is conjugate to a model (likelihood) if the prior and posterior belong to the same distributional family.
- Formal definition: 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 $p(y|\theta) \in \mathcal{F}$.

Prior elicitation

- The prior should be determined (elicited) by an **expert**. Typically, expert$\neq$statistician.
- Elicit the prior on a **quantity that she knows well** (maybe log odds $\ln \frac{\theta}{1-\theta}$ when the model is $Bern(\theta)$). The statistician can always compute the implied prior on other quantities after the elicitation.
- Elicit the prior by asking the expert probabilistic questions:
  - $E(\theta) = ?$
  - $SD(\theta) = ?$
  - $Pr(\theta < c) = ?$
  - $Pr(y > c) = ?$
- **Show the expert some consequences** of her elicitated prior. If she does not agree with these consequences, iterate the above steps until she is happy.
- **Beware of psychological effects**, such as anchoring.

Prior elicitation - AR(p) example

- **Autoregressive process** or order p

$$y_t = \phi_1(y_{t-1} - \mu) + ... + \phi_p(y_{t-p} - \mu) + \varepsilon_t, \ \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

- Informative prior on the unconditional mean: $\mu \sim N(\mu_0, \tau_0^2)$. Usually, μ_0 and τ_0^2 can be specified accurately.
- "Noninformative" prior on σ^2: $p(\sigma^2) \propto 1/\sigma^2$
- Assume for simplicity that all $\phi_i, i = 1, ..., p$ are independent a priori, and $\phi_i \sim N(\mu_i, \psi_i)$
- Prior on $\phi = (\phi_1, ..., \phi_p)$ centered on persistent AR(1) process: $\mu_1 = 0.8, \mu_2 = ... = \mu_p = 0$
- Prior variance of the ϕ_i decay towards zeros: $Var(\phi_i) = \frac{c}{i^\lambda}$, so that "longer" lags are more likely to be zero a priori. λ is a parameter that can be used to determine the rate of decay.

Different types of prior information

- Real **expert information**. Combo of previous studies and experience.
- Vague prior information, or even **noninformative priors**.
- **Reporting priors**. Easy to understand the information they contain.
- **Smoothness priors**. Regularization. Shrinkage. Big thing in modern statistics/machine learning.

'Non-informative' priors

- **Subjective consensus**: when extreme priors give essentially the same posterior.

$$p(\theta|\mathbf{x}) \to N\left(\hat{\theta}, J^{-1}_{\hat{\theta},\mathbf{x}}\right) \text{ for all } p(\theta) \text{ as } n \to \infty,$$

where $J_{\theta,\mathbf{x}}$ is the **observed information**

$$J_{\theta,\mathbf{x}} = -\frac{\partial^2 \ln L(\theta;\mathbf{x})}{\partial\theta^2}$$

- A common non-informative prior is **Jeffreys' prior**

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

where I_θ is the **Fisher information**

$$I_\theta = E_{\mathbf{x}|\theta}\left(J_{\theta,\mathbf{x}}\right)$$

Jeffreys' prior for Bernoulli trial data

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

$$\ln p(\mathbf{x}|\theta) = s \ln \theta + f \ln(1-\theta)$$

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

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

$$I(\theta) = \frac{E_{\mathbf{x}|\theta}(s)}{\theta^2} + \frac{E_{\mathbf{x}|\theta}(f)}{(1-\theta)^2} = \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}$$

Thus, the Jeffreys' prior is

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

Bayesian Learning - Lecture 3

Mattias Villani

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Lecture overview

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

Marginalization

- Models with multiple parameters $\theta_1, \theta_2, ...$
- Examples: $x_i \overset{iid}{\sim} N(\theta, \sigma^2)$; multiple regression ...
- **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:

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

- Complicated to graph the joint posterior.
- Some of the parameters may not be of direct interest (**nuisance**).
- Integrate out (**marginalize**) all nuisance parameters.
- Example: $\theta = (\theta_1, \theta_2)'$, θ_2 is a nuisance. **Marginal posterior** of θ_1

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

Normal model with unknown variance - Uniform prior

- **Model**

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

- **Prior**

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

- **Posterior**

$$\theta|\sigma^2, \mathbf{x} \sim N\left(\bar{x}, \frac{\sigma^2}{n}\right)$$

$$\sigma^2|\mathbf{x} \sim \text{Inv} - \chi^2(n-1, s^2),$$

where

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

is the usual sample variance.

Normal model with unknown variance - Uniform prior

- **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 Inv-$\chi^2(n-1, s^2)$)
  3. Draw a θ from $N\left(\bar{x}, \frac{\sigma^2}{n}\right)$ 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

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

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**:

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

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

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

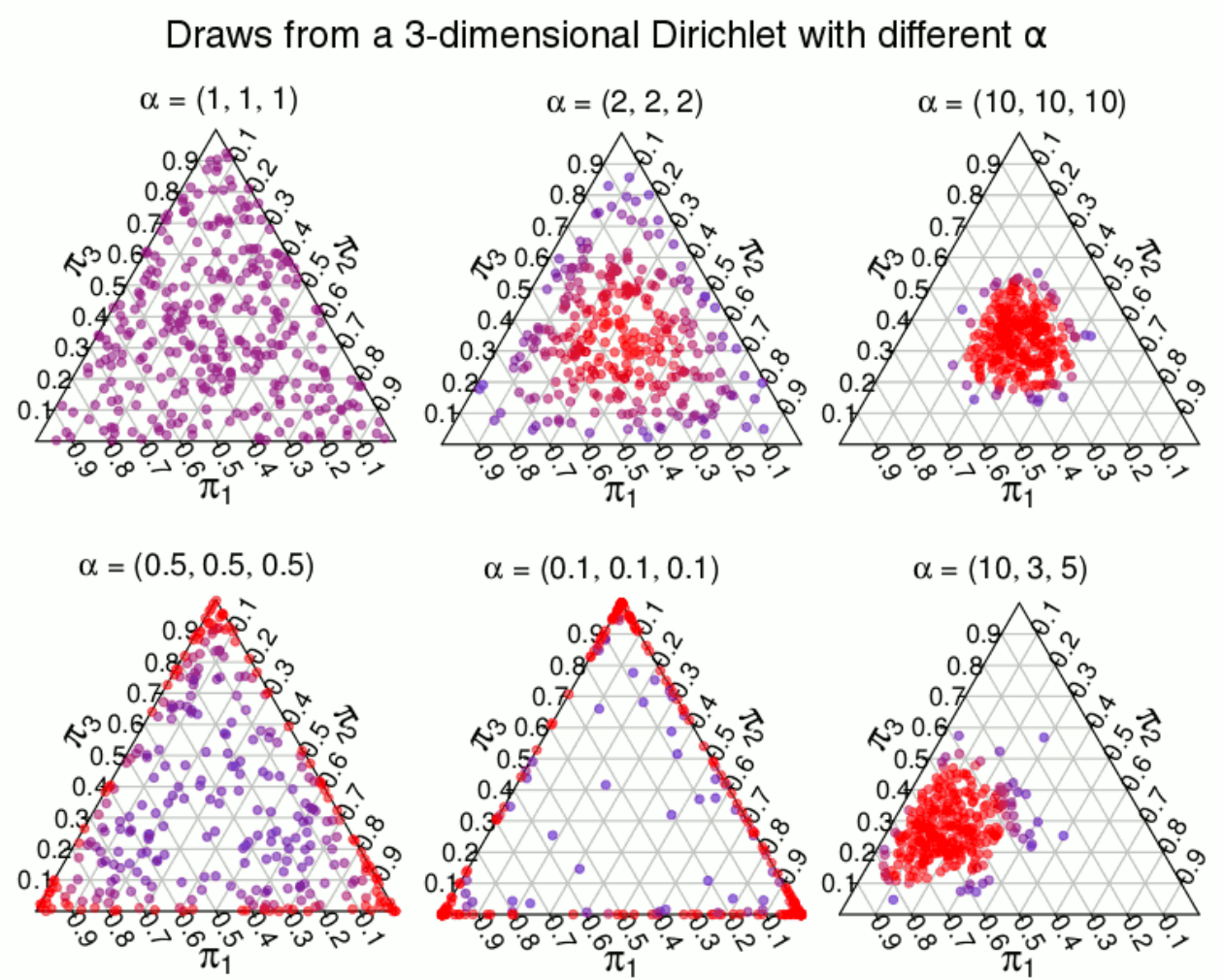
- Moments of $\theta = (\theta_1, ..., \theta_K)' \sim \textit{Dirichlet}(\alpha_1, ..., \alpha_K)$

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

$$\text{V}(\theta_k) = \frac{\text{E}(\theta_k)\left[1 - \text{E}(\theta_k)\right]}{1 + \sum_{j=1}^{K} \alpha_j}$$

- Note that $\sum_{k=1}^{K} \alpha_k$ is a precision parameter.

Dirichlet distribution



Multinomial model with Dirichlet prior

- 'Non-informative': $\alpha_1 = ... = \alpha_K = 1$ (uniform and proper).

- **Simulating** from the Dirichlet distribution:
  - Generate $x_1 \sim Gamma(\alpha_1, 1), ..., x_K \sim Gamma(\alpha_K, 1)$.
  - Compute $y_k = x_k / (\sum_{j=1}^{K} x_j)$.
  - $y = (y_1, ..., y_K)$ is a draw from the $\text{Dirichlet}(\alpha_1, ..., \alpha_K)$ distribution.

- ***Prior-to-Posterior updating***:

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

Example: market shares

- A recent survey among consumer smartphones owners in the U.S. showed that among the 513 respondents:
  - 180 owned an iPhone
  - 230 owned an Android phone
  - 62 owned a Blackberry phone
  - 41 owned some other mobile phone.
- Previous survey: iPhone 30%, Android 30%, Blackberry 20% and Other 20%.
- Pr(Android has largest share | Data)
- Prior: $\alpha_1 = 15, \alpha_2 = 15, \alpha_3 = 10$ and $\alpha_4 = 10$ (prior info is equivalent to a survey with only 50 respondents)
- Posterior: $(\theta_1, \theta_2, \theta_3, \theta_4)|\mathbf{y} \sim \text{Dirichlet}(195, 245, 72, 51)$

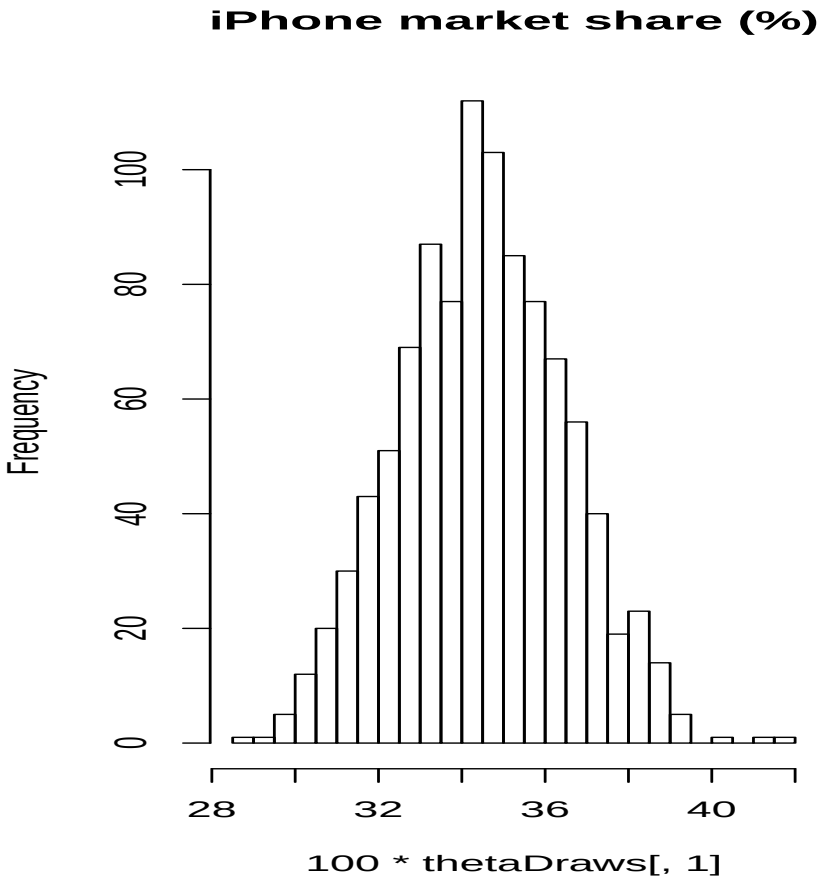
R CODE FOR MARKET SHARE EXAMPLE

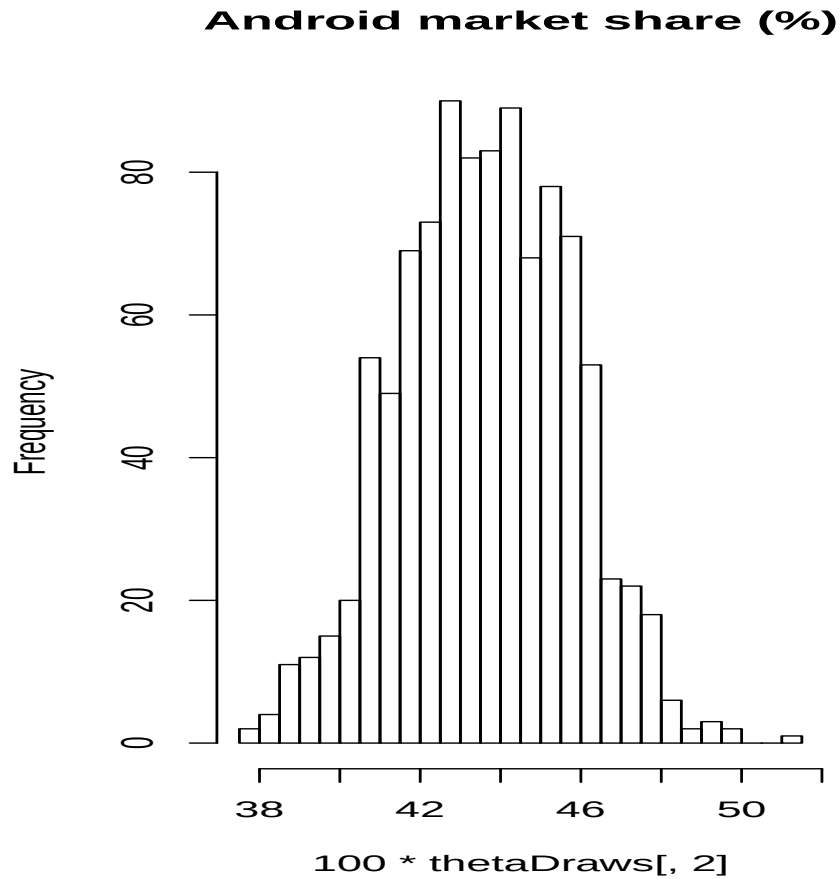
```r
# Setting up data and prior
y <- c(180,230,62,41) # The cell phone survey data (K=4)
alpha <- c(15,15,10,10) # Dirichlet prior hyperparameters
nIter <- 1000 # Number of posterior draws

# Defining a function that simulates from a Dirichlet distribution
SimDirichlet <- function(nIter, param){
  nCat <- length(param)
  thetaDraws <- as.data.frame(matrix(NA, nIter, nCat)) # Storage.
  for (j in 1:nCat){
    thetaDraws[,j] <- rgamma(nIter,param[j],1)
  }
  for (i in 1:nIter){
    thetaDraws[i,] = thetaDraws[i,]/sum(thetaDraws[i,])
  }
  return(thetaDraws)
}

# Posterior sampling from Dirichlet posterior
thetaDraws <- SimDirichlet(nIter,y + alpha)
```

R CODE FOR MARKET SHARE EXAMPLE, CONT

```r
# Posterior mean and standard deviation of Androids share (in %)
message(mean(100*thetaDraws[,2]))


## 43.6037640347044

message(sd(100*thetaDraws[,2]))

## 2.12799854134713

# Computing the posterior probability that Android is the largest
PrAndroidLargest <- sum(thetaDraws[,2]>apply(thetaDraws[,c(1,3,4)],1,max))/nIter
message(paste('Pr(Android has the largest market share) = ', PrAndroidLargest))

## Pr(Android has the largest market share) =  0.993
```

R CODE FOR MARKET SHARE EXAMPLE, CONT





Multivariate normal - known Σ

- **Model**

$$y_1, ..., y_n \overset{iid}{\sim} N_p(\mu, \Sigma)$$

where Σ is a known covariance matrix.

- **Density**

$$p(y|\mu, \Sigma) = |2\pi\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right)$$

- **Likelihood**

$$p(y_1, ..., y_n|\mu, \Sigma) \propto |\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)'\Sigma^{-1}(y_i-\mu)\right)$$
$$= |\Sigma|^{-n/2} \exp\left(-\frac{1}{2} tr\Sigma^{-1}S_\mu\right)$$

where $S_\mu = \sum_{i=1}^{n}(y_i-\mu)(y_i-\mu)'$.

Multivariate normal - known Σ

- **Prior**

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

- **Posterior**

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

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: $\Lambda_0^{-1} \to 0$.

Bayesian Learning - Lecture 4

Mattias Villani

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Lecture overview

- **Prediction**
  - Normal model
  - More complex examples
- **Decision theory**
  - The elements of a decision problem
  - The Bayesian way
  - Point estimation as a decision problem

Prediction/Forecasting

- **Posterior predictive distribution** for future $\tilde{y}$ given observed data y

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

- If $p(\tilde{y}|\theta, y) = p(\tilde{y}|\theta)$ [not true for time series], then

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

- The **parameter uncertainty** is represented in $p(\tilde{y}|y)$ by **averaging over** $p(\theta|y)$.

Prediction - Normal data, known variance

- Under the uniform prior $p(\theta) \propto c$, then

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

where

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

Prediction - Normal data, known variance

- Under the uniform prior $p(\theta) \propto c$, then

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

where

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

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

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

$$\tilde{y}|y \sim N\left[\bar{y}, \sigma^2\left(1 + \frac{1}{n}\right)\right]$$

Predictive distribution - Normal model and normal prior

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

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

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 $\tilde{y}$ (conditional variance formula):

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

- In summary:

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

Bayesian prediction in more complex models

- **Autoregressive process**

$$y_t = \phi_1(y_{t-1} - \mu) + ... + \phi_p(y_{t-p} - \mu) + \varepsilon_t, \ \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

- Simulate a draw from $p(\phi_1, \phi_2, ..., \phi_p, \mu, \sigma | y)$
  - Conditional on that draw $\theta^{(1)} = (\phi_1^{(1)}, \phi_2^{(1)}, ..., \phi_p^{(1)}, \mu^{(1)}, \sigma^{(1)})$, simulate
  - $\tilde{y}_{T+1} \sim p(y_{T+1} | y_T, y_{T-1}, ..., y_{T-p}, \theta^{(1)})$
  - $\tilde{y}_{T+2} \sim p(y_{T+2} | \tilde{y}_{T+1}, y_T, ..., y_{T-p}, \theta^{(1)})$
  - and so on.
- Repeat for new θ draws.

Bayesian prediction in more complex models

- **Autoregressive process**

$$y_t = \phi_1(y_{t-1} - \mu) + ... + \phi_p(y_{t-p} - \mu) + \varepsilon_t, \ \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

- Simulate a draw from $p(\phi_1, \phi_2, ..., \phi_p, \mu, \sigma | y)$
  - Conditional on that draw $\theta^{(1)} = (\phi_1^{(1)}, \phi_2^{(1)}, ..., \phi_p^{(1)}, \mu^{(1)}, \sigma^{(1)})$, simulate
  - $\tilde{y}_{T+1} \sim p(y_{T+1} | y_T, y_{T-1}, ..., y_{T-p}, \theta^{(1)})$
  - $\tilde{y}_{T+2} \sim p(y_{T+2} | \tilde{y}_{T+1}, y_T, ..., y_{T-p}, \theta^{(1)})$
  - and so on.
- Repeat for new θ draws.

- **Regression trees**.
  - Uncertainty on which variables to split on, and the split point.
  - For given draw of splitting variables and split points, simulate a response. Repeat for many different draws.

Predicting auction prices on eBay

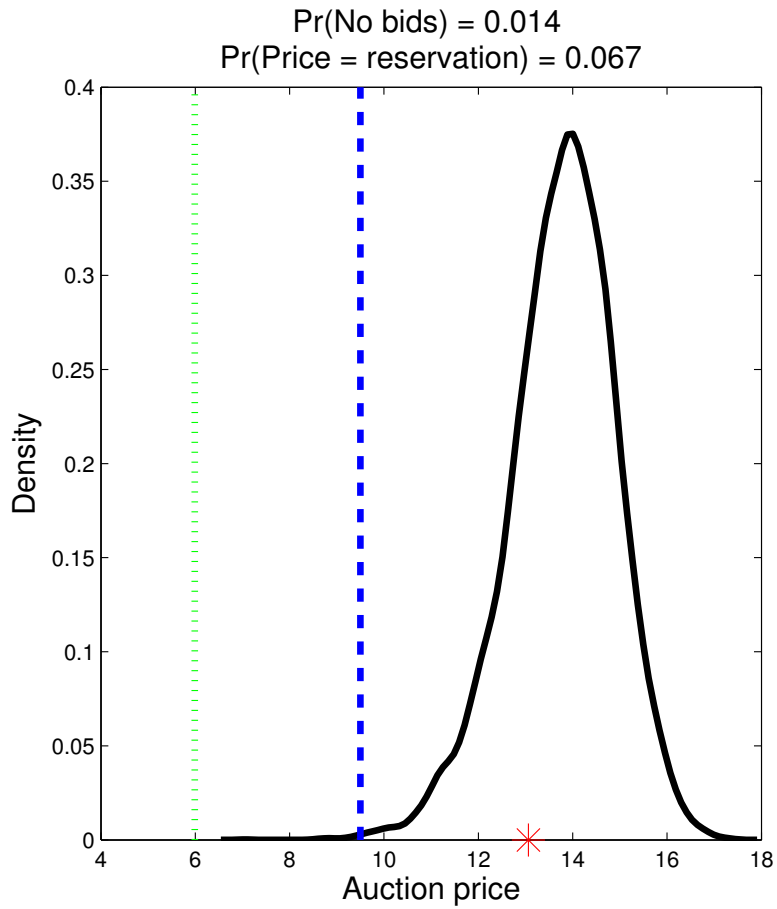
- Problem: Predicting the auctioned price in eBay coin auctions.
- Data: Bid from 1000 auctions on eBay.
  - The highest bid is not observed.
  - The lowest bids are also not observed because of the seller's reservation price.
- Covariates: auction-specific, e.g. Book value from catalog, seller's reservation price, quality of sold object, rating of seller, powerseller, verified seller ID etc
- Buyers are strategic. Their bids does not fully reflect their valuation. Game theory. **Very** complicated likelihood.

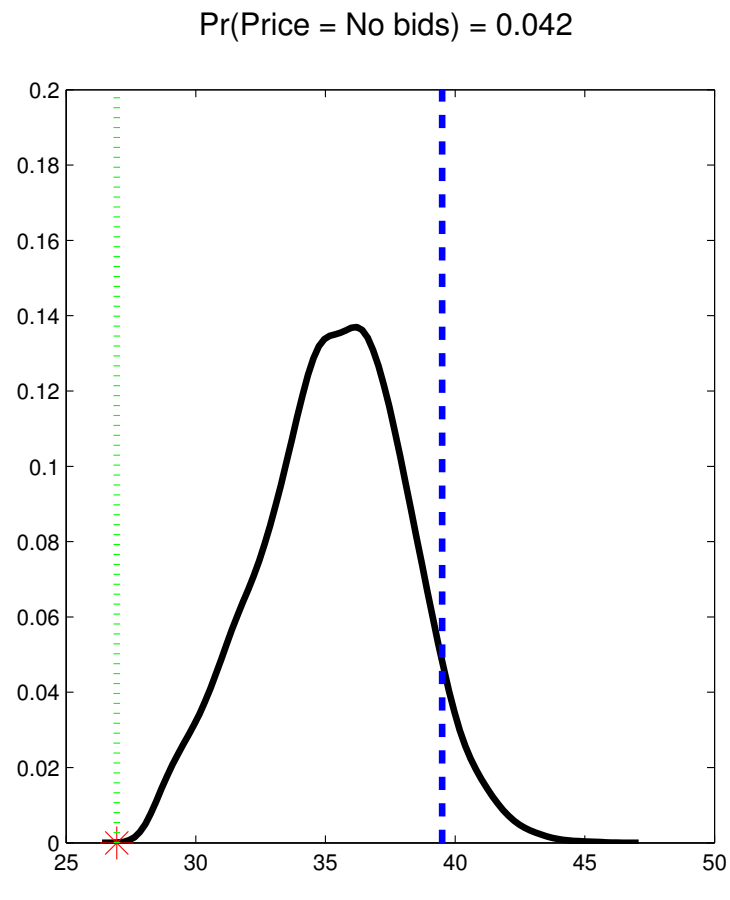
Simulating auction prices on eBay, cont.

- A draw from the posterior predictive distibution of an auction's price:

1. Simulate a draw $\theta^{(1)}$ from the posterior of the model parameters θ (using MCMC)
2. Simulate the number of bidders conditional on θ (Poisson process)
3. Simulate the bidders' valuations.
4. Simulate a complete auction bid sequence, $\mathbf{b}^{(1)}$, conditional on the valuations and $\theta = \theta^{(1)}$.
5. For the bid sequence $\mathbf{b}^{(1)}$, return the next to largest bid (eBay's proxy bidding system).

PREDICTING AUCTION PRICES ON EBAY, CONT.





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, Surgery.
- 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:

	Rainy	Sunny
Umbrella	20	10
No umbrella	50	0

Decision Theory, cont.

- Example **loss functions** when both a and θ are continuous:
  - Linear: $L(a, \theta) = |a - \theta|$
  - Quadratic: $L(a, \theta) = (a - \theta)^2$
  - Lin-Lin:

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

- Example:
  - θ is the number of items demanded of a product
  - a is the number of items in stock
  - Utility

$$U(a, \theta) = \begin{cases} p \cdot \theta - c_1(a - \theta) & \text{if } a > \theta \text{ [too much stock]} \\ p \cdot a - c_2(\theta - a)^2 & \text{if } a \leq \theta \text{ [too little stock]} \end{cases}$$

Optimal decision

- Ad hoc decision rules:
  - *Minimax*. Choose the decision that minimizes the maximum loss.
  - *Minimax-regret* ... bla bla bla ...
- Bayesian **theory**: Just maximize the **posterior expected utility**:

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

  where $E_{p(\theta|y)}$ denotes the posterior expectation.
- Using simulated draws $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ from $p(\theta|y)$:

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

- **Separation principle**:

1. First obtain $p(\theta|y)$
2. then form $U(a, \theta)$ and finally
3. choose a that maximes $E_{p(\theta|y)}[U(a, \theta)]$.

Choosing a point estimate is a decision

- 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** $\rightarrow$ $c_2/(c_1+c_2)$ quantile of the posterior is optimal
  - **Zero-one loss** $\rightarrow$ Posterior mode is optimal

Bayesian Learning - Lecture 5

Mattias Villani

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Lecture overview

- Normal model with conjugate prior
- The linear regression model
- Non-linear regression
- Regularization priors

Normal model - normal prior

- Model

$$y_1, ..., y_n | \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$$

- Conjugate prior

$$\theta | \sigma^2 \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right)$$
$$\sigma^2 \sim Inv\text{-}\chi^2(\nu_0, \sigma_0^2)$$

Normal model with normal prior

- Posterior

$$\theta|y,\sigma^2 \sim N\left(\mu_n, \frac{\sigma^2}{\kappa_n}\right)$$
$$\sigma^2|y \sim \textit{Inv-}\chi^2(\nu_n, \sigma_n^2).$$

where

$$\begin{aligned}
\mu_n &= \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)^2.
\end{aligned}$$

- Marginal posterior

$$\theta \sim t_{\nu_n}\left(\mu_n, \sigma_n^2/\kappa_n\right)$$

The Linear Regression Model

- The ordinary linear regression model:

$$\begin{aligned} y_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \\ & \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2). \end{aligned}$$

- Parameters $\theta = (\beta_1, \beta_2, ..., \beta_k, \sigma^2)$.
- Assumptions:
  - $E(y_i) = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$ (linear function)
  - $Var(y_i) = \sigma^2$ (homoscedasticity)
  - $Corr(y_i, y_j | X, \beta, \sigma^2) = 0$, $i \neq j$.
  - Normality of ε_i.
  - The x's are assumed known (non-random).

Linear regression in matrix form

- The linear regression model in matrix form

$$\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times k)(k\times 1)}{\mathbf{X}\beta} + \underset{(n\times 1)}{\varepsilon}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \ \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

- Usually $x_{i1} = 1$, for all i. β_1 is the intercept.
- Likelihood for the full sample

$$\mathbf{y}|\beta, \sigma^2, \mathbf{X} \sim N(\mathbf{X}\beta, \sigma^2 I_n)$$

LINEAR REGRESSION - UNIFORM PRIOR

- Standard non-informative prior: uniform on $(\beta, \log \sigma^2)$

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

- Joint posterior of β and σ^2:

$$\begin{aligned} \beta|\sigma^2, \mathbf{y} &\sim N\left[\hat{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right] \\ \sigma^2|\mathbf{y} &\sim \textit{Inv-}\chi^2(n-k, s^2) \end{aligned}$$

  where $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $s^2 = \frac{1}{n-k}(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})$.
- Simulate from the joint posterior by iteratively simulating from
  - $p(\sigma^2|\mathbf{y})$
  - $p(\beta|\sigma^2, \mathbf{y})$
- Marginal posterior of β :

$$\beta|\mathbf{y} \sim t_{n-k}\left[\hat{\beta}, s^2(X'X)^{-1}\right]$$

Linear regression - conjugate prior

- Joint prior for β and σ^2

$$\begin{aligned}\beta|\sigma^2 &\sim N\left(\mu_0, \sigma^2\Omega_0^{-1}\right)\\ \sigma^2 &\sim Inv-\chi^2\left(\nu_0, \sigma_0^2\right)\end{aligned}$$

- Posterior

$$\begin{aligned}\beta|\sigma^2, \mathbf{y} &\sim N\left[\mu_n, \sigma^2\Omega_n^{-1}\right]\\ \sigma^2|\mathbf{y} &\sim Inv-\chi^2\left(\nu_n, \sigma_n^2\right)\end{aligned}$$

$$\begin{aligned}\mu_n &= \left(\mathbf{X}'\mathbf{X}+\Omega_0\right)^{-1}\left(\mathbf{X}'\mathbf{X}\hat{\beta}+\Omega_0\mu_0\right)\\ \Omega_n &= \mathbf{X}'\mathbf{X}+\Omega_0\\ \nu_n &= \nu_0+n\\ \nu_n\sigma_n^2 &= \nu_0\sigma_0^2+\left(\mathbf{y}'\mathbf{y}+\mu_0'\Omega_0\mu_0-\mu_n'\Omega_n\mu_n\right)\end{aligned}$$

Polynomial regression

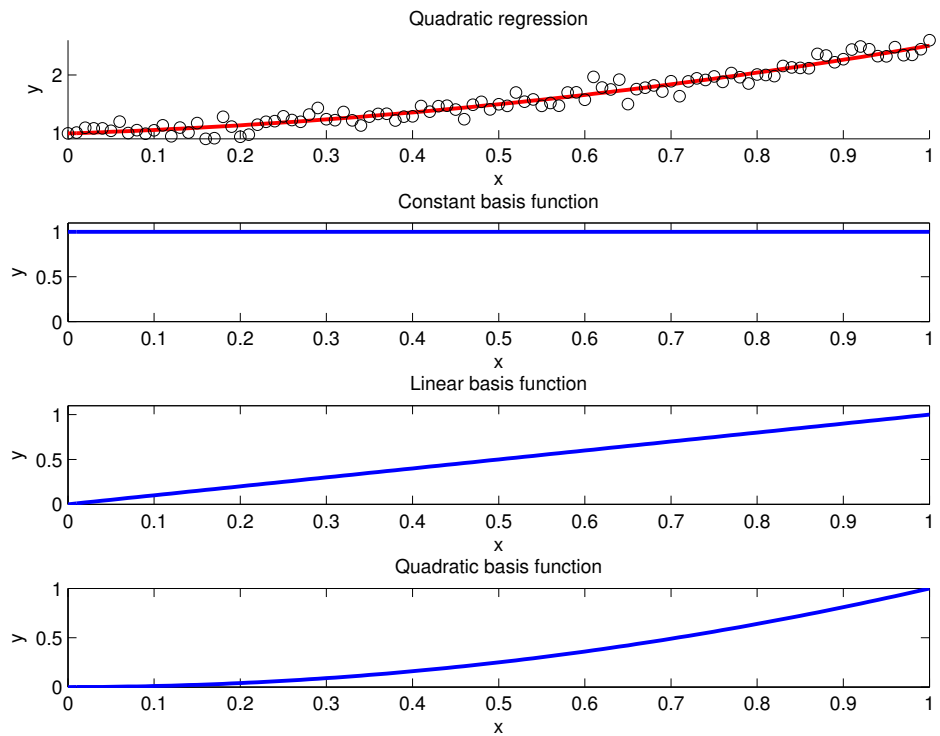
- **Polynomial regression**

$$f(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_k x_i^k.$$

$$\mathbf{y} = \mathbf{X}_P \beta + \varepsilon,$$

where

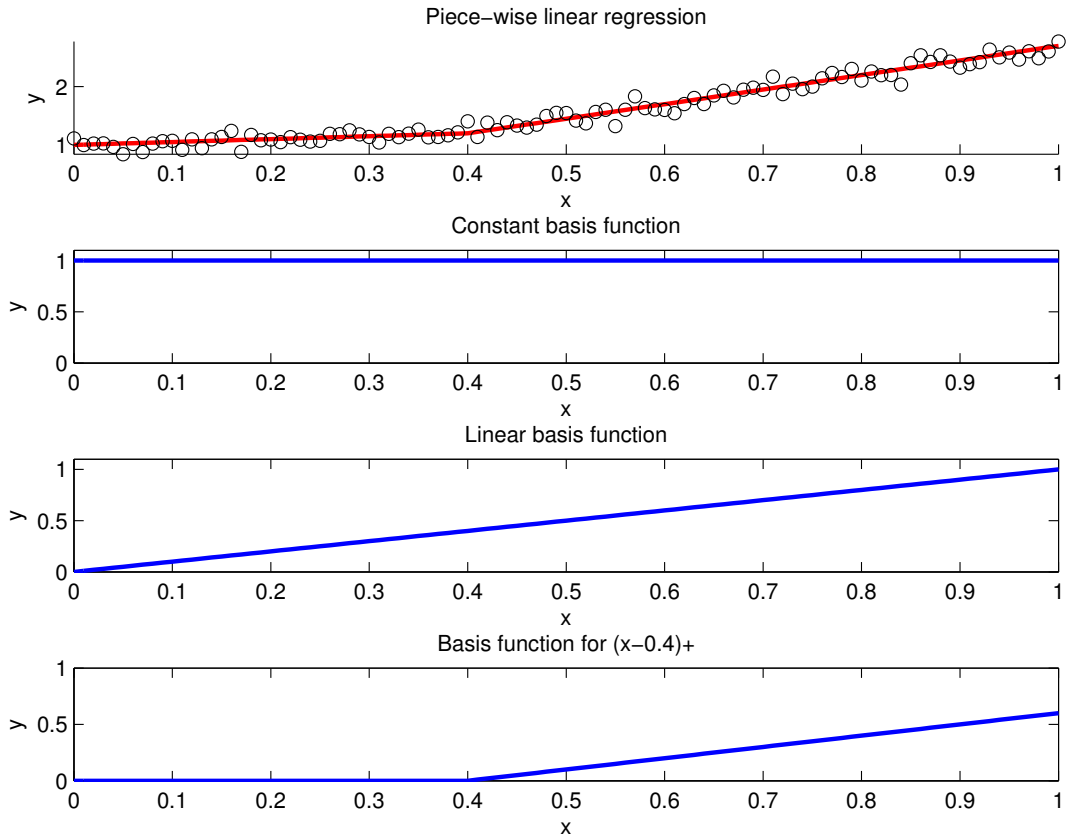
$$\mathbf{X}_P = (1, x, x^2, ..., x^k).$$



Spline regression

- Polynomials are too global. Need more local basis functions.
- ***Truncated power splines given knot locations*** $k_1, \ldots, k_m$

$$b_{ij} = \begin{cases} (x_i - k_j)^p & \text{if } x_i > k_j \\ 0 \text{ otherwise} \end{cases}$$



Splines, cont.

- Note: given the knots, the non-parametric spline regression model is a linear regression of y on the m 'dummy variables' b_j

$$\mathbf{y} = \mathbf{X}_b\beta + \varepsilon,$$

where X_b is the basis regression matrix

$$\mathbf{X}_b = (b_1, ..., b_m).$$

- It is also common to include an intercept and the linear part of the model separately. In this case we have

$$\mathbf{X}_b = (1, x, b_1, ..., b_m).$$

Smoothness prior for splines

- Problem: too many knots leads to **over-fitting**.
- Solution: **smoothness/shrinkage/regularization prior**

$$\beta_i|\sigma^2 \overset{iid}{\sim} N\left(0, \frac{\sigma^2}{\lambda}\right)$$

- Larger λ gives smoother fit. Note: here we have $\Omega_0 = \lambda I$.
- Equivalent to a penalized likelihood:

$$-2 \cdot \log p(\beta|\sigma^2, \mathbf{y}, \mathbf{X}) \propto RSS(\beta) + \lambda\beta'\beta$$

- Posterior mean gives **ridge regression** estimator

$$\tilde{\beta} = \left(\mathbf{X}'\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}'\mathbf{y}$$

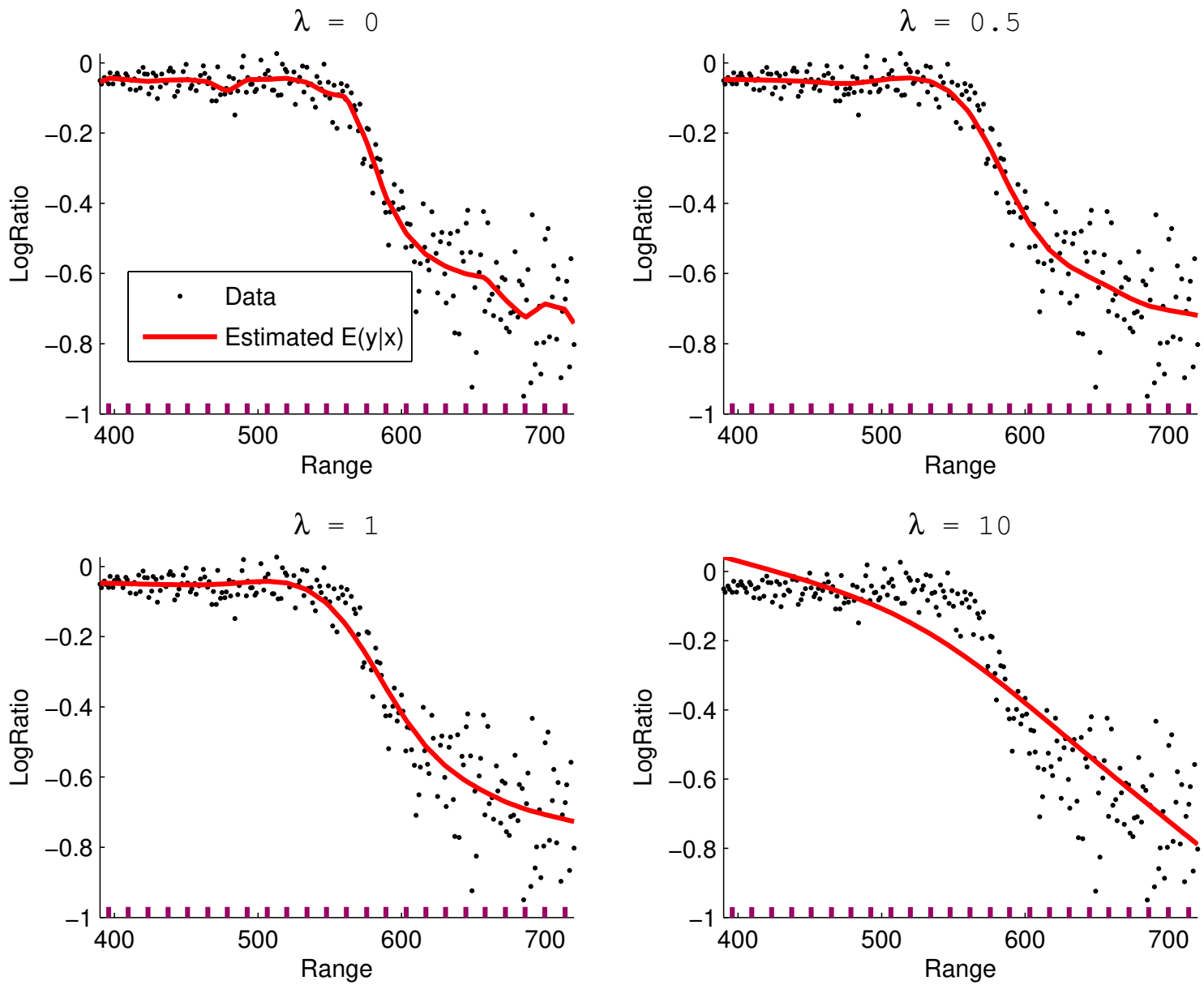
- **Shrinkage** toward zero

$$\text{As } \lambda \to \infty,\ \tilde{\beta} \to 0$$

- When $\mathbf{X}'\mathbf{X} = I$

$$\tilde{\beta} = \frac{1}{1+\lambda}\hat{\beta}_{OLS}$$

BAYESIAN SPLINE WITH SMOOTHNESS PRIOR



Smoothness prior for splines, cont.

- The famous **Lasso** variable selection method is equivalent to using the posterior mode estimate under the prior:

$$\beta_i|\sigma^2 \overset{iid}{\sim} \text{Laplace}\left(0, \frac{\sigma^2}{\lambda}\right)$$

with density

$$p(\beta_i) = \frac{\lambda}{2\sigma^2} \exp\left(-\frac{\lambda\,|\beta_i|}{\sigma^2}\right)$$

- The Bayesian shrinkage prior is **interpretable**. **Not ad hoc**.
- Laplace distribution have heavy tails.
- Laplace: many β_i are close to zero, but some β_i may be very large.
- Normal distribution have light tails.
- Normal prior: most β_i are fairly equal in size, and no single β_i can be very much larger than the other ones.

Estimating the shrinkage

- How do we determine the degree of smoothness, λ? Cross-validation is one possible approach.
- Bayesian: λ is unknown $\Rightarrow$ use a prior for λ.
- One possibility: $\lambda \sim Inv - \chi^2(\eta_0, \lambda_0)$. The user specifies η_0 and λ_0.
- Alternative approach: specify the prior on the *degrees of freedom*.
- Hierarchical setup:

$$\begin{aligned} \mathbf{y}|\beta, \mathbf{X} &\sim N(\mathbf{X}\beta, \sigma^2 I_n) \\ \beta|\sigma^2, \lambda &\sim N\left(0, \sigma^2 \lambda^{-1} I_m\right) \\ \sigma^2 &\sim Inv - \chi^2(\nu_0, \sigma_0^2) \\ \lambda &\sim Inv - \chi^2(\eta_0, \lambda_0) \end{aligned}$$

so $\Omega_0 = \lambda I_m$ in the previous notation.

Regression with estimated shrinkage

- The joint posterior of β, σ^2 and λ is

$$\begin{aligned}
\beta|\sigma^2, \lambda, \mathbf{y} &\sim N\left(\mu_n, \Omega_n^{-1}\right) \\
\sigma^2|\lambda, \mathbf{y} &\sim Inv-\chi^2\left(\nu_n, \sigma_n^2\right) \\
p(\lambda|\mathbf{y}) &\propto \sqrt{\frac{|\Omega_0|}{|\mathbf{X}^T\mathbf{X}+\Omega_0|}}\left(\frac{\nu_n\sigma_n^2}{2}\right)^{-\nu_n/2} \cdot p(\lambda)
\end{aligned}$$

where $\Omega_0 = \lambda I_m$, and $p(\lambda)$ is the prior for λ, and

$$\begin{aligned}
\mu_n &= \left(\mathbf{X}^T\mathbf{X}+\Omega_0\right)^{-1}\mathbf{X}^T\mathbf{y} \\
\Omega_n &= \mathbf{X}^T\mathbf{X}+\Omega_0 \\
\nu_n &= \nu_0 + n \\
\nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + \mathbf{y}^T\mathbf{y} - \mu_n^T\Omega_n\mu_n
\end{aligned}$$

More complexity

- The **location of the knots** can be treated as unknown, and estimated from the data. Joint posterior

$$p(\beta, \sigma^2, \lambda, k_1, ..., k_m | \mathbf{y}, \mathbf{X})$$

- The marginal posterior for $\lambda, k_1, ..., k_m$ is a nightmare.

- MCMC can be used to simulate from the joint posterior. Li and Villani (2013, SJS).
- The basic spline model can be extended with:
  - **Heteroscedastic errors** (also modelled with a spline)
  - **Non-normal errors** (student-t or mixture distributions)
  - **Autocorrelated/dependent errors** (AR process for the error term)
- MCMC can again be used to simulate from the joint posterior.

Bayesian Learning - Lecture 6

Mattias Villani

Division of Statistics
Department of Computer and Information Science
Linköping University

Lecture overview

- **Classification**
- **Naive Bayes**
- **Normal approximation** of posterior
- **Logistic regression** - demo in R

Bayesian classification

- **Classification: output is a discrete label**. Examples:
  - binary (0-1). Spam/Ham.
  - Multi-class. ($c = 1, 2, ..., C$). $\{iPhone, Android, Windows, Other\}$.
- **Bayesian classification**

$$\underset{c \in \mathcal{C}}{\operatorname{argmax}}\, p(c|\mathbf{x})$$

  where $\mathbf{x} = (x_1, ..., x_p)$ is a covariate/feature vector.
- **Discriminative models** - model $p(c|\mathbf{x})$ directly.
- Examples: logistic regression, support vector machines.
- **Generative models** - Use Bayes' theorem

$$p(c|\mathbf{x}) \propto p(\mathbf{x}|c)p(c)$$

  and model class-conditional distribution $p(\mathbf{x}|c)$ and prior $p(c)$.
- Examples: discriminant analysis, naive Bayes.

Naive Bayes

- By Bayes' theorem
$$p(c|\mathbf{x}) \propto p(\mathbf{x}|c)p(c)$$
- $p(c)$ can be estimated by Multinomial-Dirichlet analysis.
- $p(\mathbf{x}|c)$ can be $N(\theta_c, \Sigma_c)$ or mixture of normals (see last module).
- $p(\mathbf{x}|c)$ can be very high-dimensional and hard to estimate.
- Even with binary features, the outcome space of $p(\mathbf{x}|c)$ can be huge.
- Naive Bayes: **features are assumed independent**
$$p(\mathbf{x}|c) = \prod_{j=1}^{n} p(x_j|c)$$
- Naive Bayes solution
$$p(c|\mathbf{x}) \propto \left[\prod_{j=1}^{n} p(x_j|c)\right] p(c)$$

Classification with logistic regression

- Response is assumed to be **binary** ($y = 0$ or 1).
- Example: Spam ($y = 1$) or Ham ($y = 0$). Covariates: \$-symbols, etc.
- **Logistic regression**

$$\Pr(y_i = 1 \mid x_i) = \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)}.$$

- Likelihood

$$p(y|X, \beta) = \prod_{i=1}^{n} \frac{[\exp(x_i'\beta)]^{y_i}}{1 + \exp(x_i'\beta)}.$$

- Prior $\beta \sim N(0, \tau^2 I)$. Posterior is non-standard (see demo in R later).
- Alternative: **Probit regression**

$$Pr(y_i = 1|x_i) = \Phi(x_i'\beta)$$

- **Multi-class** ($c = 1, 2, ..., C$) logistic regression

$$\Pr(y_i = c \mid x_i) = \frac{\exp(x_i'\beta_c)}{\sum_{k=1}^{C} \exp(x_i'\beta_k)}$$

Large sample approximate posterior

- **Taylor expansion of log-posterior** around the posterior mode $\theta = \tilde{\theta}$:

$$\ln p(\theta|y) = \ln p(\tilde{\theta}|y) + \frac{\partial \ln p(\theta|y)}{\partial \theta}|_{\theta=\tilde{\theta}}(\theta - \tilde{\theta}) + \frac{1}{2!}\frac{\partial^2 \ln p(\theta|y)}{\partial \theta^2}|_{\theta=\tilde{\theta}}(\theta - \tilde{\theta})^2 + ...$$

- From the definition of the posterior mode:

$$\frac{\partial \ln p(\theta|y)}{\partial \theta}|_{\theta=\tilde{\theta}} = 0$$

- So, in **large samples** (where we can ignore higher order terms):

$$p(\theta|y) \approx p(\tilde{\theta}|y) \exp\left(-\frac{1}{2}J_{\mathbf{y}}(\tilde{\theta})(\theta - \tilde{\theta})^2\right)$$

  where $J_{\mathbf{y}}(\tilde{\theta}) = -\frac{\partial^2 \ln p(\theta|y)}{\partial \theta^2}|_{\theta=\tilde{\theta}}$ is the **observed information**.

- **Approximate posterior**

$$\theta|y \overset{approx}{\sim} N\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$$

Example: gamma posterior

- Poisson model: $\theta|y_1, ..., y_n \sim Gamma(\alpha + \sum_{i=1}^n y_i, \beta + n)$

$$\log p(\theta|y_1, ..., y_n) \propto (\alpha + \sum_{i=1}^n y_i - 1) \log \theta - \theta(\beta + n)$$

- First derivative of log density

$$\frac{\partial \ln p(\theta|y)}{\partial \theta} = \frac{\alpha + \sum_{i=1}^n y_i - 1}{\theta} - (\beta + n)$$

$$\tilde{\theta} = \frac{\alpha + \sum_{i=1}^n y_i - 1}{\beta + n}$$

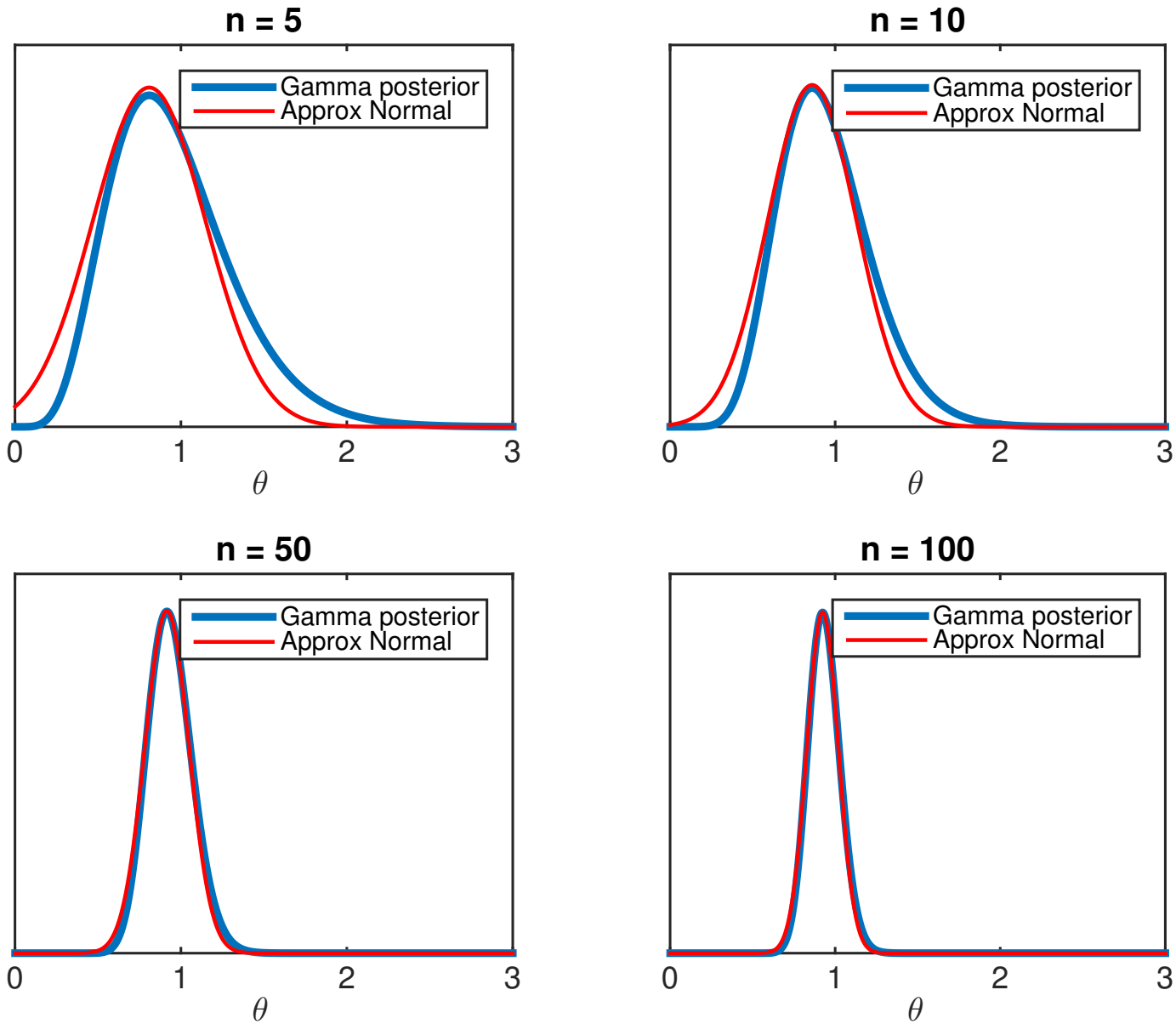
- Second derivative at mode $\tilde{\theta}$

$$\frac{\partial^2 \ln p(\theta|y)}{\partial \theta^2}|_{\theta=\tilde{\theta}} = -\frac{\alpha + \sum_{i=1}^n y_i - 1}{\left(\frac{\alpha+\sum_{i=1}^n y_i - 1}{\beta+n}\right)^2} = -\frac{(\beta + n)^2}{\alpha + \sum_{i=1}^n y_i - 1}$$

- So, the normal approximation is

$$N\left[\frac{\alpha + \sum_{i=1}^n y_i - 1}{\beta + n}, \frac{\alpha + \sum_{i=1}^n y_i - 1}{(\beta + n)^2}\right]$$

Example: gamma posterior



Normal approximation of posterior

- $\theta|y \overset{approx}{\sim} N\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$ works also when θ is a vector.
- How to compute $\tilde{\theta}$ and $J_{\mathbf{y}}(\tilde{\theta})$?
- Standard **optimization routines** may be used. (`optim.r`).
  - **Input**: an expression proportional to $\log p(\theta|y)$ and initial values.
  - **Output**: $\log p(\tilde{\theta}|y)$, $\tilde{\theta}$ and Hessian matrix $(-J_{\mathbf{y}}(\tilde{\theta}))$.
- **Re-parametrization** may improve normal approximation. [Don't forget the **Jacobian**!]
  - If $\theta \geq 0$ use $\phi = \log(\theta)$.
  - If $0 \leq \theta \leq 1$, use $\phi = \ln[\theta/(1-\theta)]$.
- **Heavy tailed approximation**: $\theta|y \overset{approx}{\sim} t_v\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$ for suitable degrees of freedom v.

Example: gamma posterior - reparam.

- Poisson model revisited. Reparameterize to $\phi = \log(\theta)$.
- Use change-of-variables formula from a basic probability course

$$\log p(\phi|y_1, ..., y_n) \propto (\alpha + \sum_{i=1}^{n} y_i - 1)\phi - \exp(\phi)(\beta + n) + \phi$$

- Taking first and second derivatives and evaluating at $\tilde{\phi}$ gives

$$\tilde{\phi} = \log\left(\frac{\alpha + \sum_{i=1}^{n} y_i}{\beta + n}\right) \text{ and } \frac{\partial^2 \ln p(\phi|y)}{\partial \phi^2}|_{\phi=\tilde{\phi}} = \alpha + \sum_{i=1}^{n} y_i$$

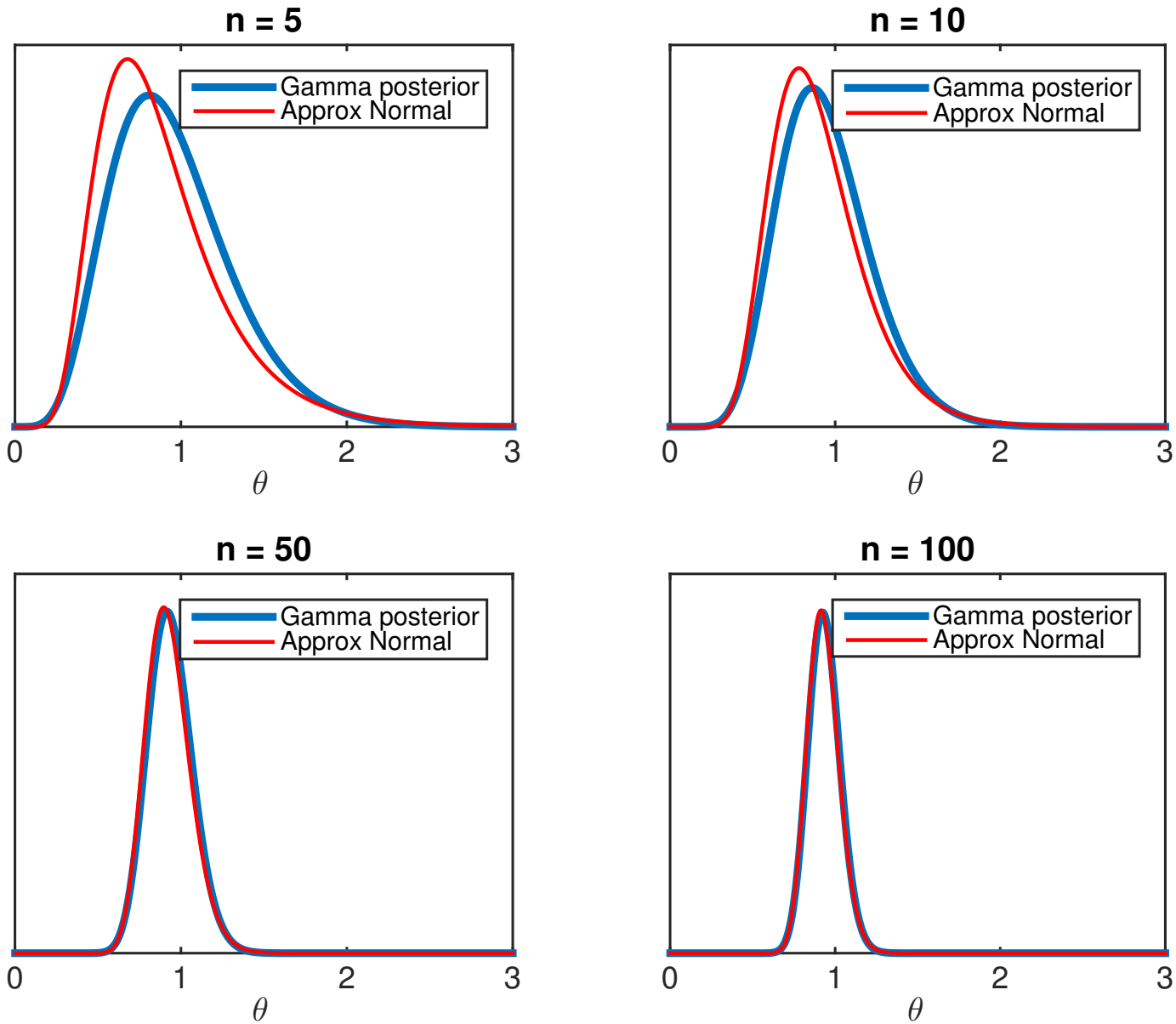
- So, the normal approximation for $p(\phi|y_1, ...y_n)$ is

$$\phi = \log(\theta) \sim N\left[\log\left(\frac{\alpha + \sum_{i=1}^{n} y_i}{\beta + n}\right), \frac{1}{\alpha + \sum_{i=1}^{n} y_i}\right]$$

which means that $p(\theta|y_1, ...y_n)$ is log-normal:

$$\theta|y \sim LN\left[\log\left(\frac{\alpha + \sum_{i=1}^{n} y_i}{\beta + n}\right), \frac{1}{\alpha + \sum_{i=1}^{n} y_i}\right]$$

EXAMPLE: GAMMA POSTERIOR - REPARAMETERIZED



Normal approximation of posterior

- Even if the posterior of θ is approx normal, **interesting functions** of $g(\theta)$ may not be (e.g. predictions).
- But approximate posterior of $g(\theta)$ can be obtained by **simulating** from $N\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$.
- **Example**: Posterior of Gini coefficient.
  - Model: $x_1, ..., x_n | \mu, \sigma^2 \sim LN(\mu, \sigma^2)$.
  - Let $\phi = \log(\sigma^2)$. And $\theta = (\mu, \phi)$.
  - Joint posterior $p(\mu, \phi)$ may be approximately normal: $\theta | y \overset{approx}{\sim} N\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$.
  - Simulate $\theta^{(1)}, ..., \theta^{(N)}$ from $N\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$. Compute $\sigma^{(1)}, ..., \sigma^{(N)}$.
  - Compute $G^{(i)} = 2\Phi\left(\sigma^{(i)}/\sqrt{2}\right)$ for $i = 1, ..., N$.

Bayesian Learning - Lecture 7

Mattias Villani and Per Sidén

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Lecture overview

- Monte Carlo simulation and random number generation
- Gibbs sampling
- Data augmentation
  - Mixture models
  - Probit regression
- Regularized regression revisited

Monte Carlo sampling

- If $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ is an *iid* **sequence** from a distribution $p(\theta)$, then

$$\frac{1}{N}\sum_{t=1}^{N}\theta^{(t)} \quad \rightarrow \quad E(\theta)$$

$$\frac{1}{N}\sum_{t=1}^{N}g(\theta^{(t)}) \quad \rightarrow \quad E[g(\theta)]$$

  where $g(\theta)$ is some well-behaved function.

- Easy to compute **tail probabilities** $\Pr(\theta \leq c)$ by letting

$$g(\theta) = I(\theta \leq c)$$

  and

$$\frac{1}{N}\sum_{t=1}^{N}g(\theta^{(t)}) = \frac{\#\ \theta\text{-draws smaller than } c}{N}.$$

Direct sampling by the inverse CDF method

- How to **simulate** from a distribution?
- Let $f(x)$ be the density function of a stochastic variable. CDF: $F(x)$. **Inverse CDF method**:
  1. Generate u from the uniform distribution on $[0, 1]$.
  2. Compute $x = F^{-1}(u)$.
- Example 1: **Exponential distribution**:

$$u = F(x) = 1 - \exp(-\lambda x)$$

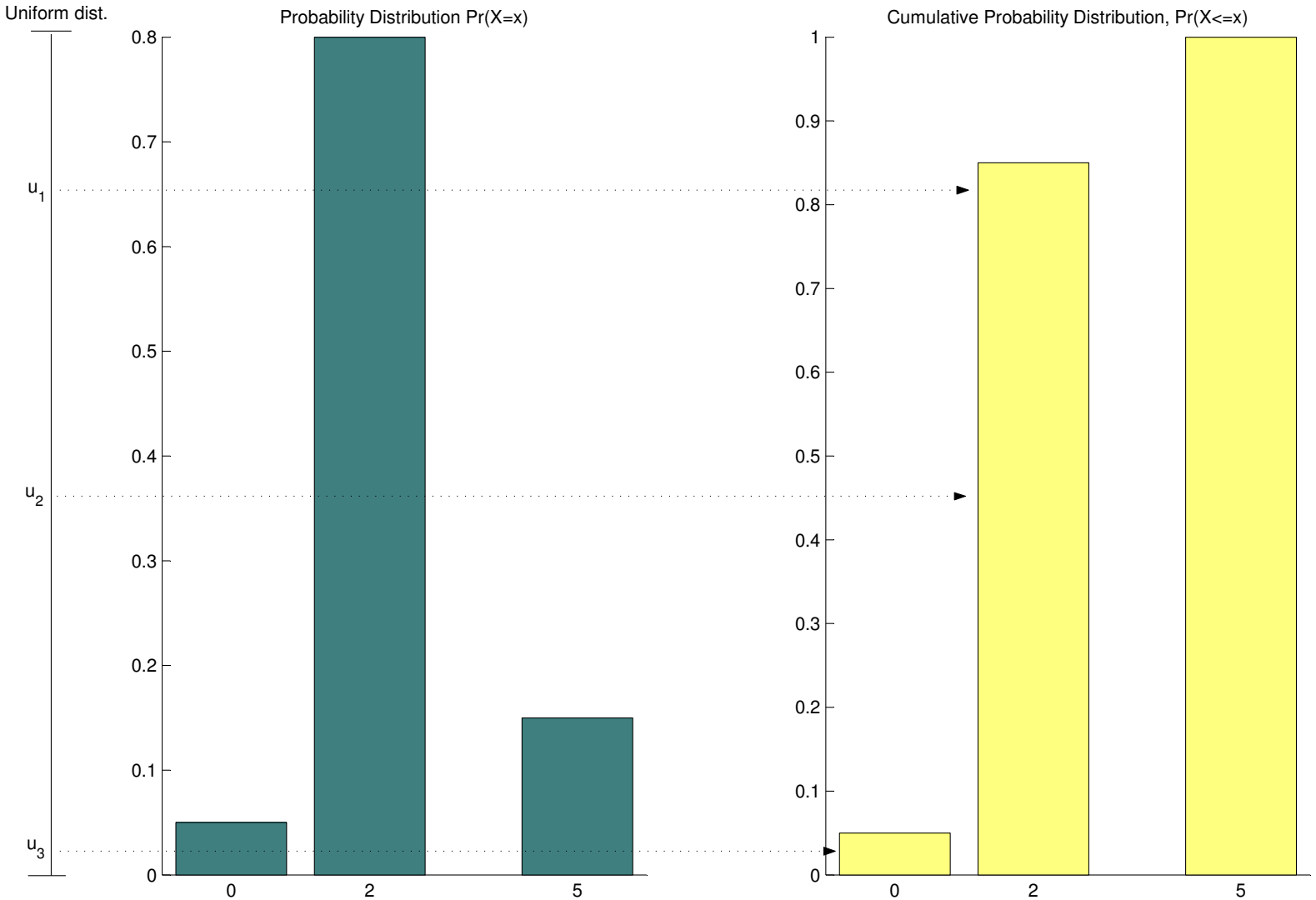
Inverting gives

$$x = -\ln(1 - u)/\lambda$$

But $1 - u$ is also uniformly distributed on [0,1]. So:

- If $x = -(\ln u)/\lambda$ where $u \sim \mathit{Unif}(0, 1)$, then $x \sim \mathit{Expon}(\lambda)$.

Inverse CDF method, discrete case



Direct sampling by the inverse CDF method

- Example 2: **Cauchy distribution**:

$$\begin{aligned} f(x) &= \frac{1}{\pi}\frac{1}{1+x^2} \\ u &= F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x) \end{aligned}$$

  Inverting ...

$$x = \tan[\pi(u - 1/2)].$$

- We can also use relations between distribution to sample from distributions.
- Cauchy-example, cont. If y and z are independent $N(0,1)$ variables, then $z = \frac{y}{z} \sim \mathit{Cauchy}$.
- Example: **Chi-square**. If $x_1, ..., x_v \overset{iid}{\sim} N(0,1)$, then $y = \sum_{i=1}^{v} x_i^2 \sim \chi_v^2$.

Gibbs sampling

- Easily implemented methods for **sampling from multivariate distributions**, $p(\theta_1, ..., \theta_k)$.

- Requirements: Easily sampled **full conditional posteriors**:
  - $p(\theta_1|\theta_2, \theta_3..., \theta_k)$
  - $p(\theta_2|\theta_1, \theta_3, ..., \theta_k)$
  - $\vdots$
  - $p(\theta_k|\theta_1, \theta_2, ..., \theta_{k-1})$
- Started out in the early 80's in the image analysis literature.
- Gibbs sampling is a **special case of Metropolis-Hastings** (see Lecture 8)
- Metropolis-Hastings is a Markov Chain Monte Carlo (MCMC) algorithm.

The Gibbs sampling algorithm

A: Choose initial values $\theta_2^{(0)}, \theta_3^{(0)}, ..., \theta_k^{(0)}$.

B:
- B_1 Draw $\theta_1^{(1)}$ from $p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, ..., \theta_k^{(0)})$
- B_2 Draw $\theta_2^{(1)}$ from $p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, ..., \theta_k^{(0)})$
- $\vdots$
- B_n Draw $\theta_k^{(1)}$ from $p(\theta_k|\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_{k-1}^{(1)})$

C: Repeat Step B N times.

GIBBS SAMPLING, CONT.

- The Gibbs draws $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ are **dependent** (autocorrelated), but **arithmetic means converge to expected values**

$$\frac{1}{N}\sum_{t=1}^{N}\theta_j^{(t)} \quad \rightarrow \quad E(\theta_j)$$

$$\frac{1}{N}\sum_{t=1}^{N}g(\theta^{(t)}) \quad \rightarrow \quad E[g(\theta)]$$

- $\theta^{(1)}, ..., \theta^{(N)}$ **converges in distribution** to the target $p(\theta)$.
- $\theta_j^{(1)}, ..., \theta_j^{(N)}$ converge to the marginal distribution of θ_j, $p(\theta_j)$.
- **Dependent** draws $\rightarrow$ **less efficient** than iid sampling.
- Compare sampling from:
  - $x_t \overset{iid}{\sim} N(0, \sigma^2)$
  - $x_t = 0.9x_{t-1} + \varepsilon_t$ with $\varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$.

Gibbs sampling multivariate normal

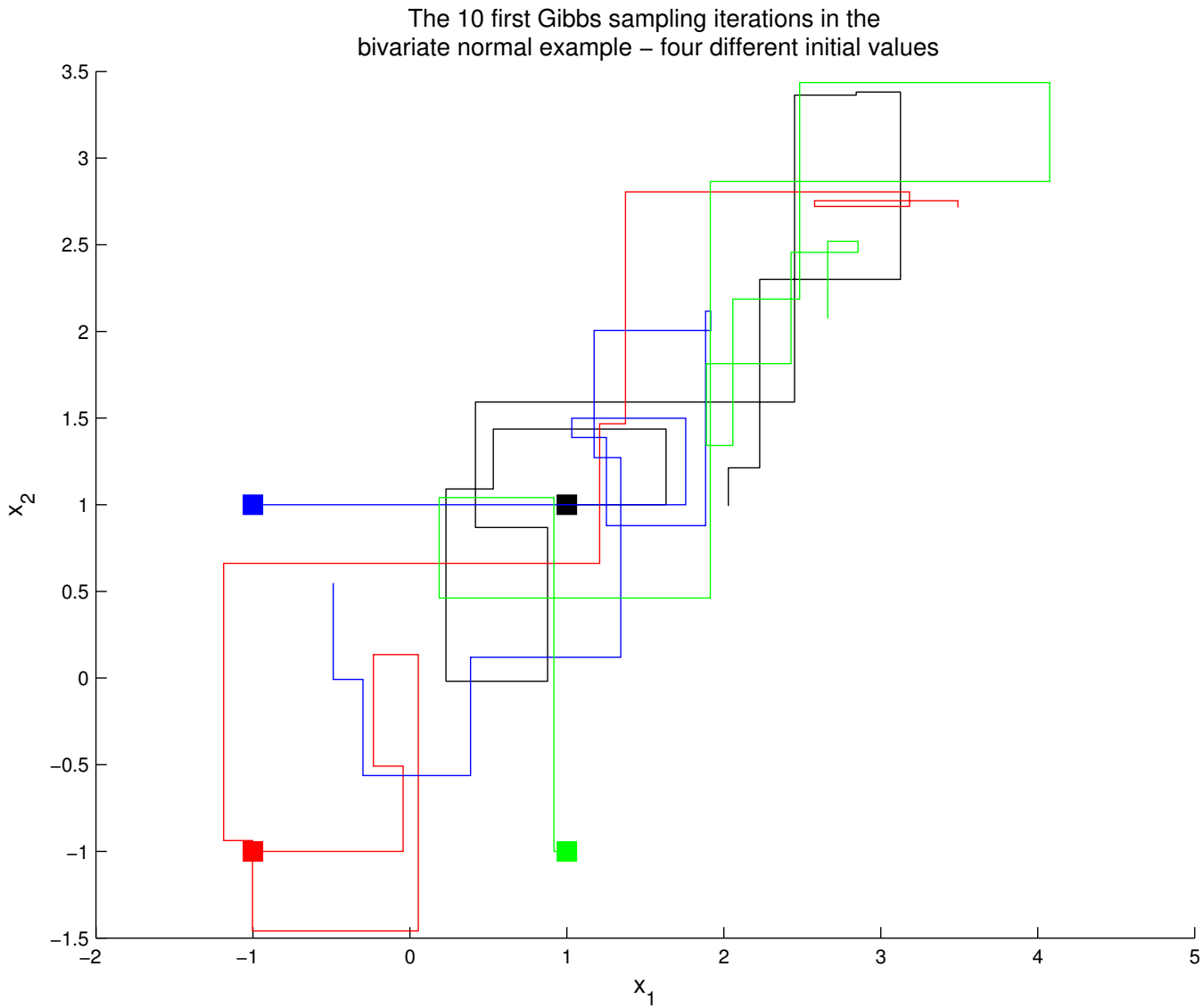
- Bivariate normal:
  - Joint distribution

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]$$

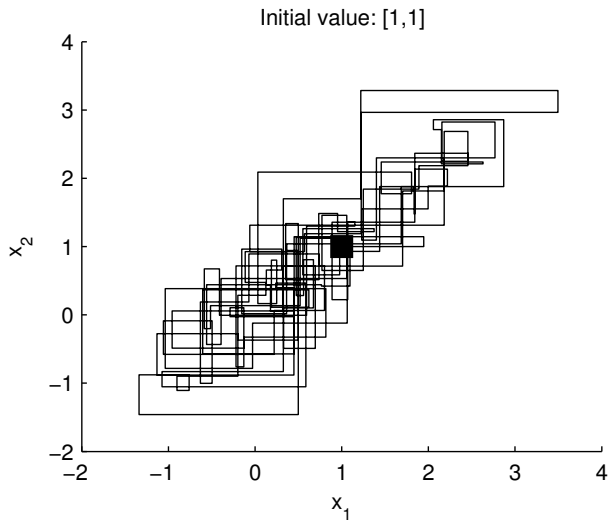
  - Full conditional posteriors:

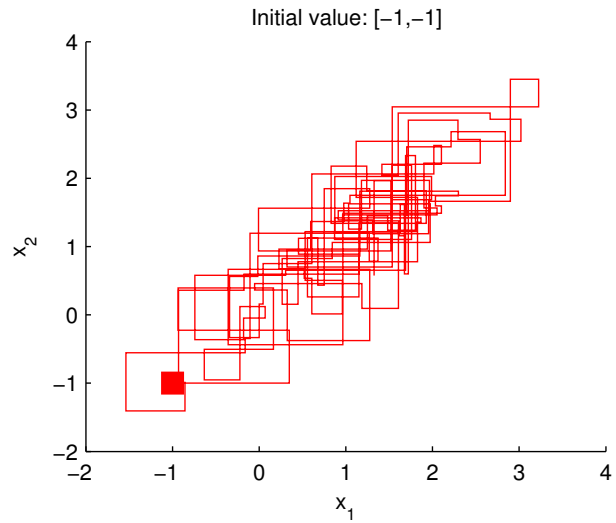
$$\begin{aligned} \theta_1|\theta_2 &\sim N[\mu_1 + \rho(\theta_2 - \mu_2), 1 - \rho^2] \\ \theta_2|\theta_1 &\sim N[\mu_2 + \rho(\theta_1 - \mu_1), 1 - \rho^2] \end{aligned}$$

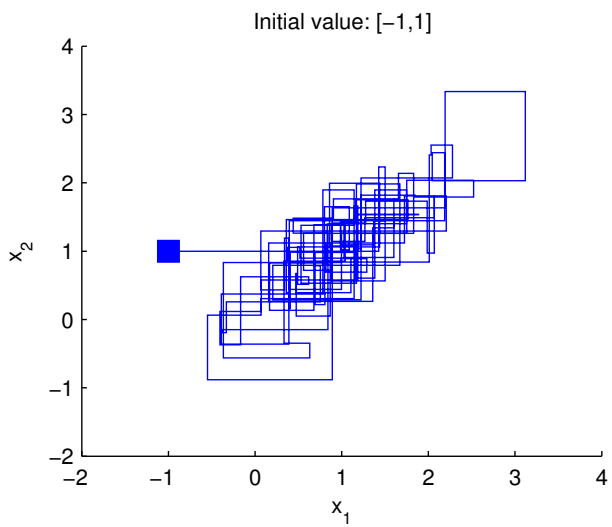
GIBBS SAMPLING - BIVARIATE NORMAL

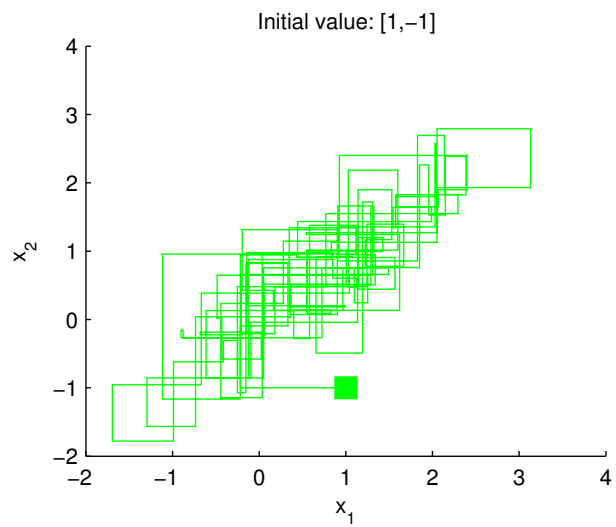


Gibbs sampling - Bivariate normal

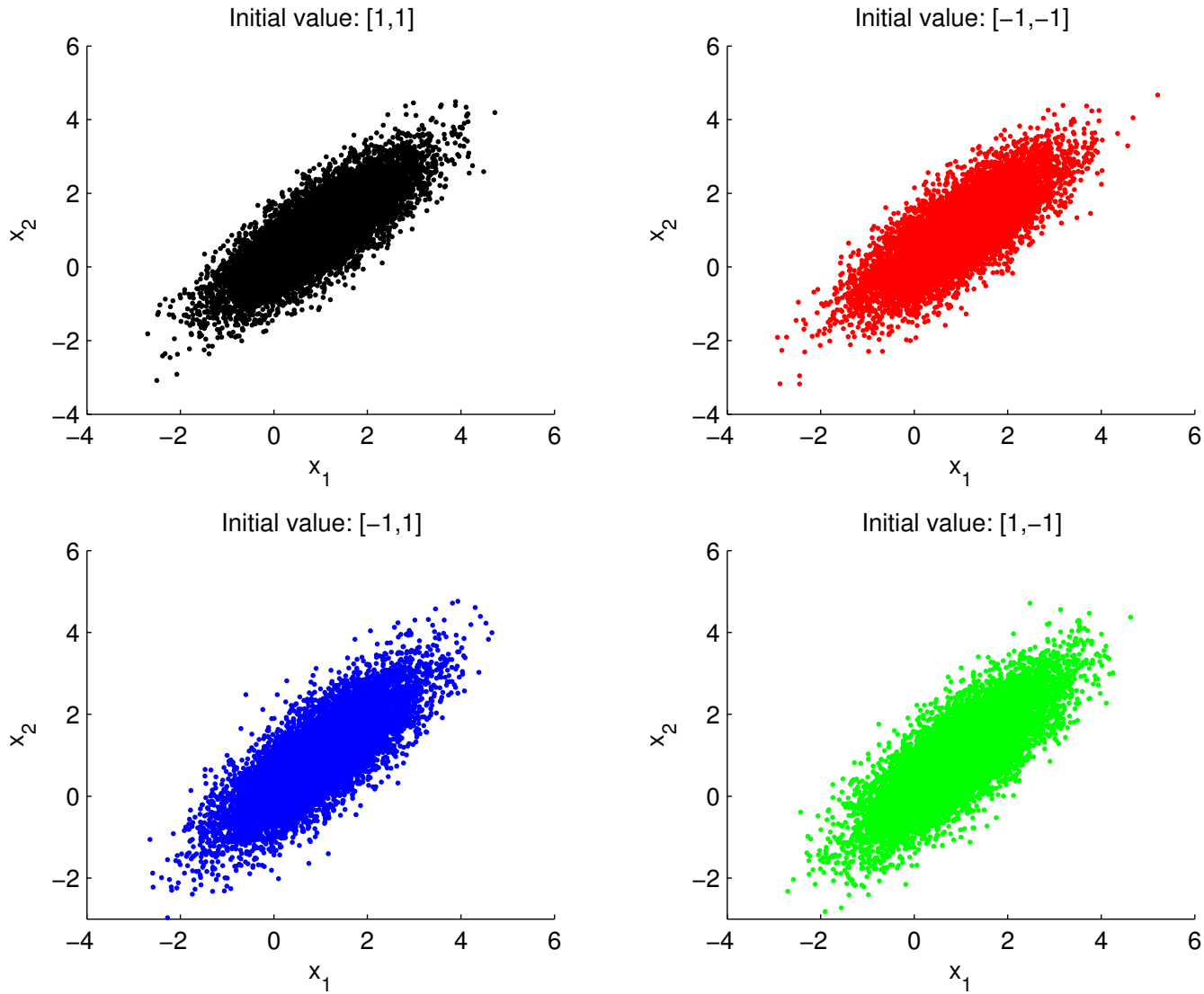




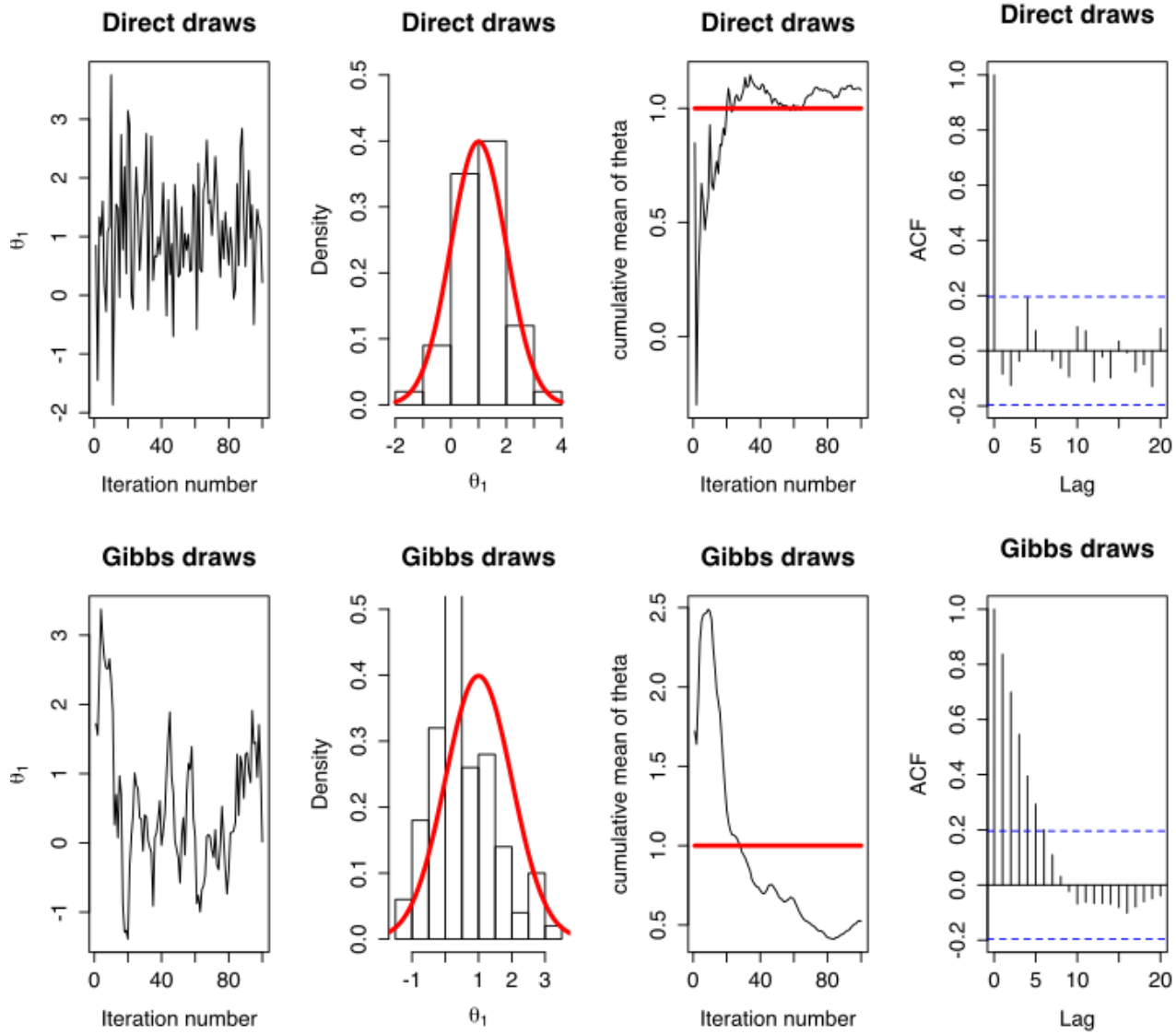




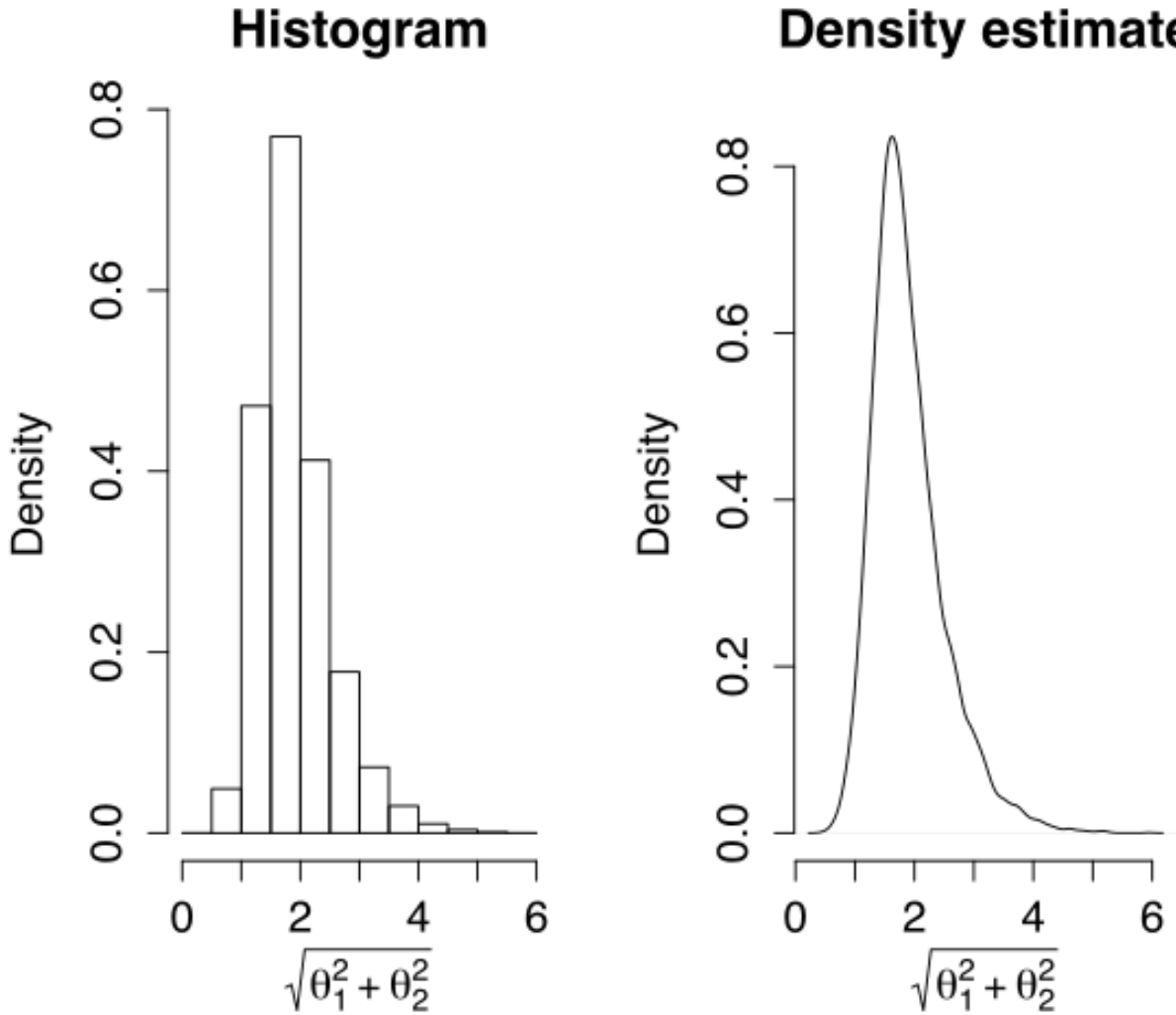
Gibbs sampling - Bivariate normal



Direct sampling vs Gibbs sampling



ESTIMATING THE DENSITY OF $g(\theta_1, \theta_2) = \sqrt{\theta_1^2 + \theta_2^2}$

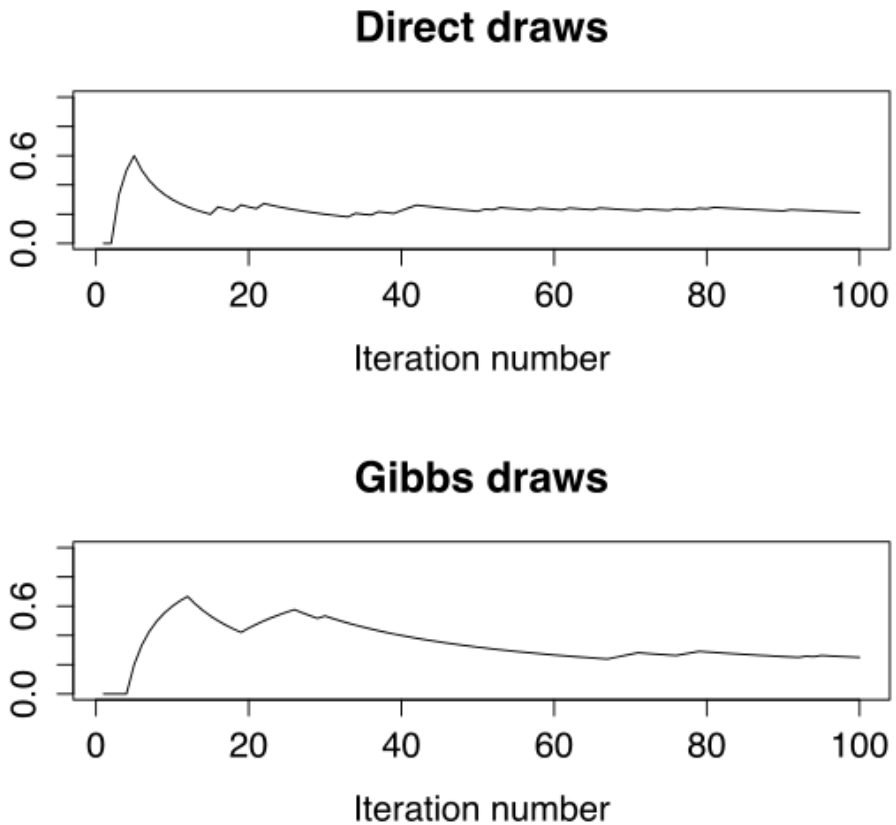


Estimating $Pr(\theta_1 > 0, \theta_2 > 0)$

- We can estimate a joint probability by counting:

$$Pr(\theta_1 > 0, \theta_2 > 0) \approx N^{-1} \sum_{i=1}^{N} 1(\theta_1^{(i)} > 0, \theta_2^{i)} > 0)$$

.



Gibbs sampling for normal model with non-conjugate prior

- Normal model with semi-conjugate prior

$$\mu \sim N(\mu_0, \tau_0^2)$$
$$\sigma^2 \sim Inv - \chi^2(\nu_0, \sigma_0^2)$$

- Conditional posteriors

$$\mu|\sigma^2, x \sim N\left(\mu_n, \tau_n^2\right)$$
$$\sigma^2|\mu, x \sim Inv - \chi^2\left(\nu_n, \frac{\nu_0\sigma_0^2 + \sum_{i=1}^n \left(x_i - \mu\right)^2}{n + \nu_0}\right)$$

with μ_n and τ_n^2 defined the same as when σ^2 is known (Lecture 1).

Gibbs sampling for AR processes

- **AR(p) process**

$$x_t = \mu + \phi_1(x_{t-1} - \mu) + ... + \phi_p(x_{t-p} - \mu) + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2).$$

- Let $\phi = (\phi_1, ..., \phi_p)'$.
- **Prior**:
  - μ ~Normal
  - ϕ ~Multivariate Normal
  - σ^2 ~Scaled Inverse χ^2.
- The **posterior** can be simulated by Gibbs sampling:
  - $\mu|\phi, \sigma^2, x \sim$ Normal
  - $\phi|\mu, \sigma^2, x \sim$ Multivariate Normal
  - $\sigma^2|\mu, \phi, x \sim$ Scaled Inverse χ^2

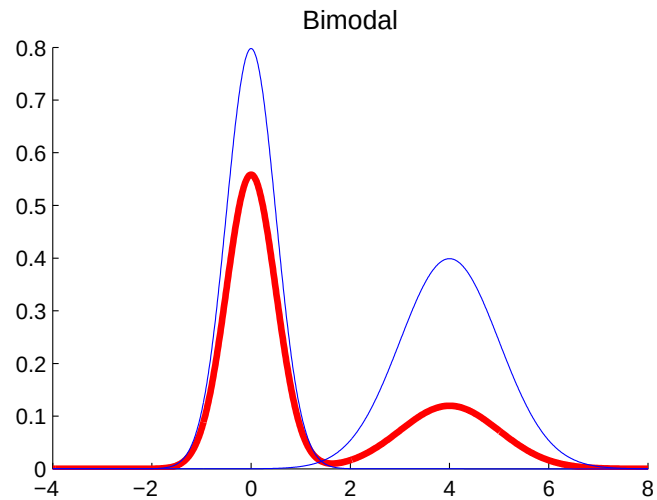
Data augmentation - Mixture distributions

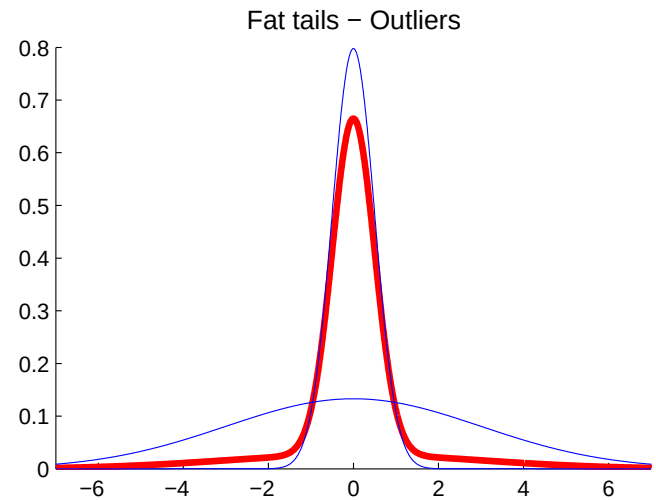
- Let $\phi(x|\mu, \sigma^2)$ denotes the **PDF** of a **normal** variable $x \sim N(\mu, \sigma^2)$.
- **Two-component mixture of normals** [MN(2)]

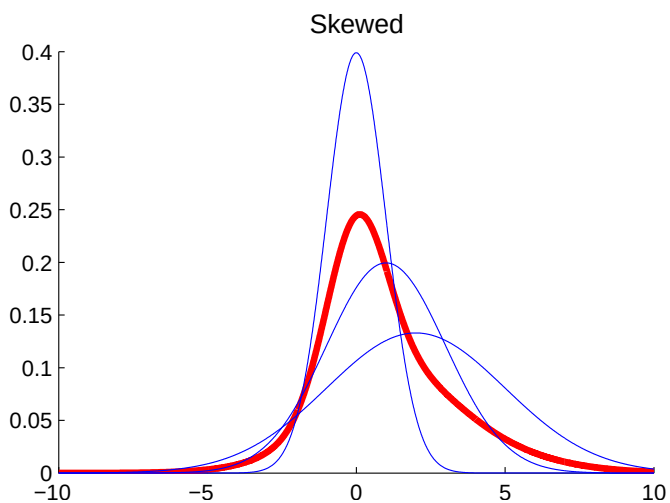
$$p(x) = \pi \cdot \phi(x|\mu_1, \sigma_1^2) + (1 - \pi) \cdot \phi(x|\mu_2, \sigma_2^2)$$

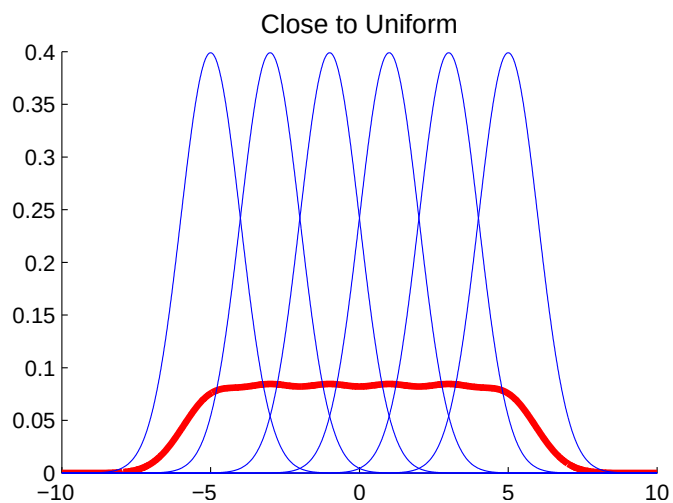
- **Simulate** from a MN(2):
  - Simulate an indicator $I \in \{1, 2\}$: $I \sim Bern(\pi)$.
  - If $I = 1$, simulate x from $N(\mu_1, \sigma_1^2)$
  - If $I = 2$, simulate x from $N(\mu_2, \sigma_2^2)$.

Illustration of mixture distributions









Mixture distributions, cont.

- Not easy to estimate directly - the likelihood is a product of sums.
- **Assume** that we knew which of the two densities each observation came from.

$$I_i = \begin{cases} 1 \text{ if } x_i \text{ came from Density 1} \\ 2 \text{ if } x_i \text{ came from Density 2} \end{cases}.$$

- Armed with knowledge of $I_1, ..., I_n$ it is now easy to estimate π, $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2$ by separating the sample according to the I's.
- But we do **not** know $I_1, ..., I_n$!

Gibbs sampling for mixture distributions

- Prior: $\pi \sim Beta(\alpha_1, \alpha_2)$. Conjugate prior for (μ_j, σ_j^2), see Lecture 5.
- Define: $n_1 = \sum_{i=1}^{n} (I_i = 1)$ and $n_2 = n - n_1$.
- **Gibbs sampling**:
  - $\pi \mid \mathbf{I}, \mathbf{x} \sim Beta(\alpha_1 + n_1, \alpha_2 + n_2)$
  - $\sigma_1^2 \mid \mathbf{I}, \mathbf{x} \sim Inv\text{-}\chi^2(\nu_{n_1}, \sigma_{n_1}^2)$ and $\mu_1 | \mathbf{I}, \sigma_1^2, \mathbf{x} \sim N\left(\mu_{n_1}, \frac{\sigma_1^2}{\kappa_{n_1}}\right)$
  - $\sigma_2^2 \mid \mathbf{I}, \mathbf{x} \sim Inv\text{-}\chi^2(\nu_{n_2}, \sigma_{n_2}^2)$ and $\mu_2 | \mathbf{I}, \sigma_2^2, \mathbf{x} \sim N\left(\mu_{n_2}, \frac{\sigma_2^2}{\kappa_{n_2}}\right)$
  - $I_i \mid \pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \mathbf{x} \sim Bern(\theta_i)$, $i = 1, ..., n$,

$$\theta_i = \frac{(1-\pi)\phi(x_i; \mu_2, \sigma_2^2)}{\pi\phi(x_i; \mu_1, \sigma_1^2) + (1-\pi)\phi(x_i; \mu_2, \sigma_2^2)}.$$

Gibbs sampling for mixture distributions

- K-component mixture of normals

$$p(x) = \sum_{k=1}^{K} \pi_k \phi(x; \mu_k, \sigma_k^2),$$

  where $\sum_{k=1}^{K} \pi_k = 1$.
- **Multi-class indicators**: $I_i = k$ if observation i comes from density k.
- Gibbs sampling with
  - $(\pi_1, ..., \pi_K) \mid \mathbf{I}, \mathbf{x} \sim \textit{Dirichlet}(\alpha_1 + n_1, \alpha_2 + n_2, ..., \alpha_K + n_K)$
  - $\sigma_k^2 \mid \mathbf{I}, \mathbf{x} \sim \textit{Inv-}\chi^2$ and $\mu_k | \mathbf{I}, \sigma_k^2, \mathbf{x} \sim \textit{Normal}, \quad \textit{for } k = 1, ..., K,$
  - $I_i \mid \pi, \mu, \sigma^2, \mathbf{x} \sim \textit{Multinomial}(\theta_{i1}, ..., \theta_{iK}), \textit{ for } i = 1, ..., n,$

$$\theta_{ij} = \frac{\pi_j \phi(x_i; \mu_j, \sigma_j^2)}{\sum_{r=1}^{k} \pi_r \phi(x_i; \mu_r, \sigma_r^2)}.$$

- Gibbs sampling is very powerful for **missing data** problems. **Semi-supervised learning**.

Data augmentation - Probit regression

- **Probit** model:

$$\Pr(y_i = 1 \mid x_i) = \Phi(x_i^T \beta)$$

- **Random utility formulation** of the probit:

$$\begin{aligned} u_i &\sim N(x_i^T \beta, 1) \\ y_i &= \begin{cases} 1 & \text{if } u_i > 0 \\ 0 & \text{if } u_i \leq 0 \end{cases}. \end{aligned}$$

- Check: $\Pr(y_i = 1 \mid x_i) = \Pr(u_i > 0) = 1 - \Pr(u_i \leq 0) = 1 - \Pr(u_i - x_i^T \beta < -x_i^T \beta) = 1 - \Phi(-x_i^T \beta) = \Phi(x_i^T \beta)$.
- If $u = (u_1, ..., u_n)$ were observed, then β could be analyzed by traditional linear regression. But, u is **not observed**. Gibbs sampling to the rescue!

Gibbs sampling for the Probit regression

- Simulate from joint posterior $p(u, \beta|y)$ iterating between the **full conditional posteriors**:
  - $p(\beta|u, y)$, which is multivariate normal (this is just a linear regression)
  - $p(u_i|\beta, y)$, $i = 1, ..., n$.
- The full conditional posterior distribution of u_i is:

$$p(u_i|\beta, y) \propto p(y_i|\beta, u_i)p(u_i|\beta)$$

$$= \begin{cases} N(u_i|x_i'\beta, 1) & \text{truncated to } u_i \in (-\infty, 0] \text{ if } y_i = 0 \\ N(u_i|x_i'\beta, 1) & \text{truncated to } u_i \in (0, \infty) \text{ if } y_i = 1 \end{cases}$$

- Collect the β-draws. A histogram of these draws approximates $p(\beta|y) = \int p(u, \beta|y)du$.

Regularized regression with Gibbs

- Recap: The joint posterior of β, σ^2 and λ is

$$\beta|\sigma^2, \lambda, \mathbf{y}, \mathbf{X} \sim N\left(\mu_n, \Omega_n^{-1}\right)$$
$$\sigma^2|\lambda, \mathbf{y}, \mathbf{X} \sim Inv-\chi^2\left(\nu_n, \sigma_n^2\right)$$
$$p(\lambda|\mathbf{y}, \mathbf{X}) \propto \sqrt{\frac{|\Omega_0|}{|X'X+\Omega_0|}}\left(\frac{\nu_n\sigma_n^2}{2}\right)^{-\nu_n/2} \cdot p(\lambda)$$

  where $p(\lambda)$ is the $Inv-\chi^2$ prior for λ.
- This is the **conditional-marginal decomposition**

$$p(\beta, \sigma^2, \lambda|\mathbf{y}, \mathbf{X}) = p(\beta|\sigma^2, \lambda, \mathbf{y}, \mathbf{X})p(\sigma^2|\lambda, \mathbf{y}, \mathbf{X})p(\lambda|\mathbf{y}, \mathbf{X})$$

- **Gibbs sampling** can instead be used:
  - Sample $\beta|\sigma^2, \lambda, \mathbf{y}, \mathbf{X}$ from Normal
  - Sample $\sigma^2|\beta, \lambda, \mathbf{y}, \mathbf{X}$ from Inv-χ^2
  - Sample $\lambda|\beta, \sigma^2, \mathbf{y}, \mathbf{X}$ from Inv-χ^2
- Note that λ is now **easy** to simulate **once we condition** on β and σ^2.

Improving the efficiency of the Gibbs sampler

- ***Efficient blocking***. Correlated parameters should ideally be included in the same updating block.
- ***Reparametrization***. Convergence can improve dramatically in alternative parametrizations.
- ***Data augmentation***. Bring in latent (unobserved) variables that make the full conditional posteriors more easily sampled (Probit, Mixture models etc). Downside: Typically increases the autocorrelation between draws.
- ***Parameter expansion***. Introducing (non-sense) parameters in the model may break the dependence between the original parameters (Example probit).

Bayesian Learning - Lecture 8

Mattias Villani

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Lecture overview

- Markov Chain Monte Carlo - the general idea
- Metropolis-Hastings
- MCMC in practice

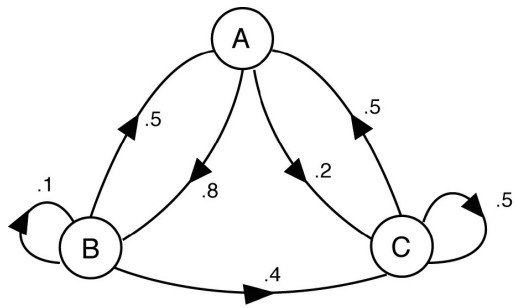
Markov chains

- Let $\mathcal{S} = \{s_1, s_2, ..., s_k)$ be a finite set of **states**.
  - Weather: $\mathcal{S} = \{\text{sunny, rain}\}$.
  - Journal rankings: $\mathcal{S} = \{A+, A, B, C, D, E\}$
- **Markov chain** is a stochastic process $\{X_t\}_{t=1}^{T}$ with random **state transitions**

$$p_{ij} = \Pr(X_{t+1} = s_j | X_t = s_i)$$

- Example realization journal ranking:
  $X_1 = C, X_2 = C, X_3 = B, X_4 = A+, X_5 = B$.
- **Transition matrix** for weather example

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 0.9 & 0.1 \\ 0.7 & 0.3 \end{pmatrix}$$



Stationary distribution

- *h*-step transition probabilities

$$P_{ij}^{(h)} = \Pr(X_{t+h} = s_j | X_t = s_i)$$

- *h*-step transition matrix

$$P^{(h)} = P^h$$

- The chain has a **unique equilibrium** stationary distribution $\pi = (\pi_1, ..., \pi_k)$ if it is
  - **irreducible** (possible to get from any state from any state)
  - **aperiodic** (does not get stuck in predictable cycles)
  - **positive recurrent** (expected time of returning to any state is finite)
- Limiting (long-run) distribution

$$P^t \to \begin{pmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_k \\ \pi_1 & \pi_2 & \cdots & \pi_k \\ \vdots & \vdots & & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_k \end{pmatrix} \text{ as } t \to \infty$$

STATIONARY DISTRIBUTION, CONT.

- Limiting (long-run) distribution

$$P^t \to \begin{pmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_k \\ \pi_1 & \pi_2 & \cdots & \pi_k \\ \vdots & \vdots & & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_k \end{pmatrix} \text{ as } t \to \infty$$

- **Stationary distribution**

$$\pi = \pi P$$

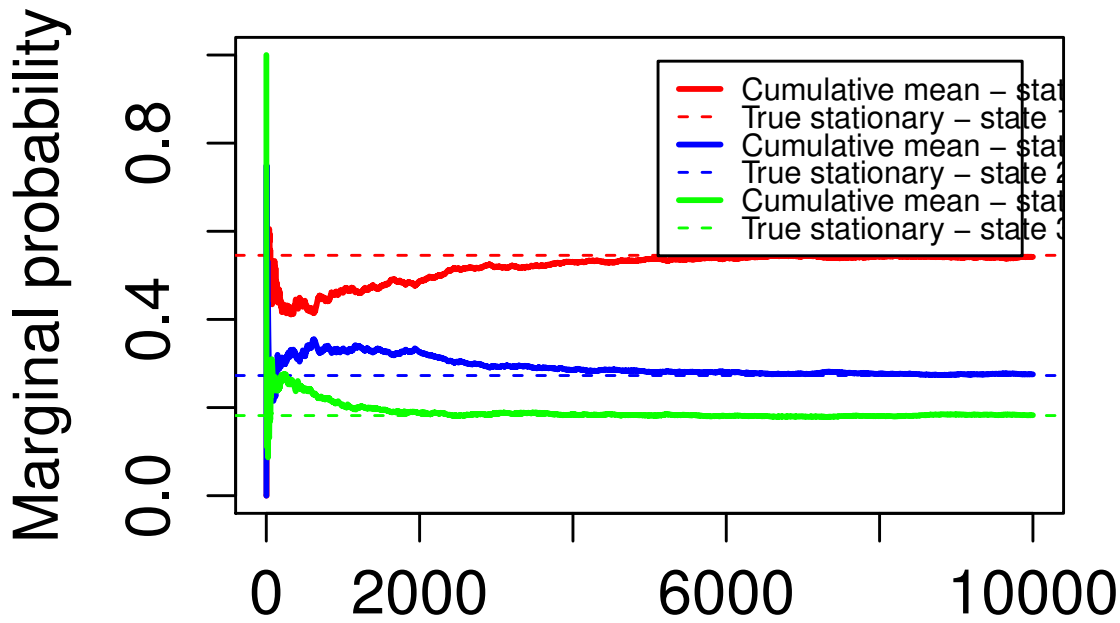
- Example:

$$P = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.4 \end{pmatrix}$$

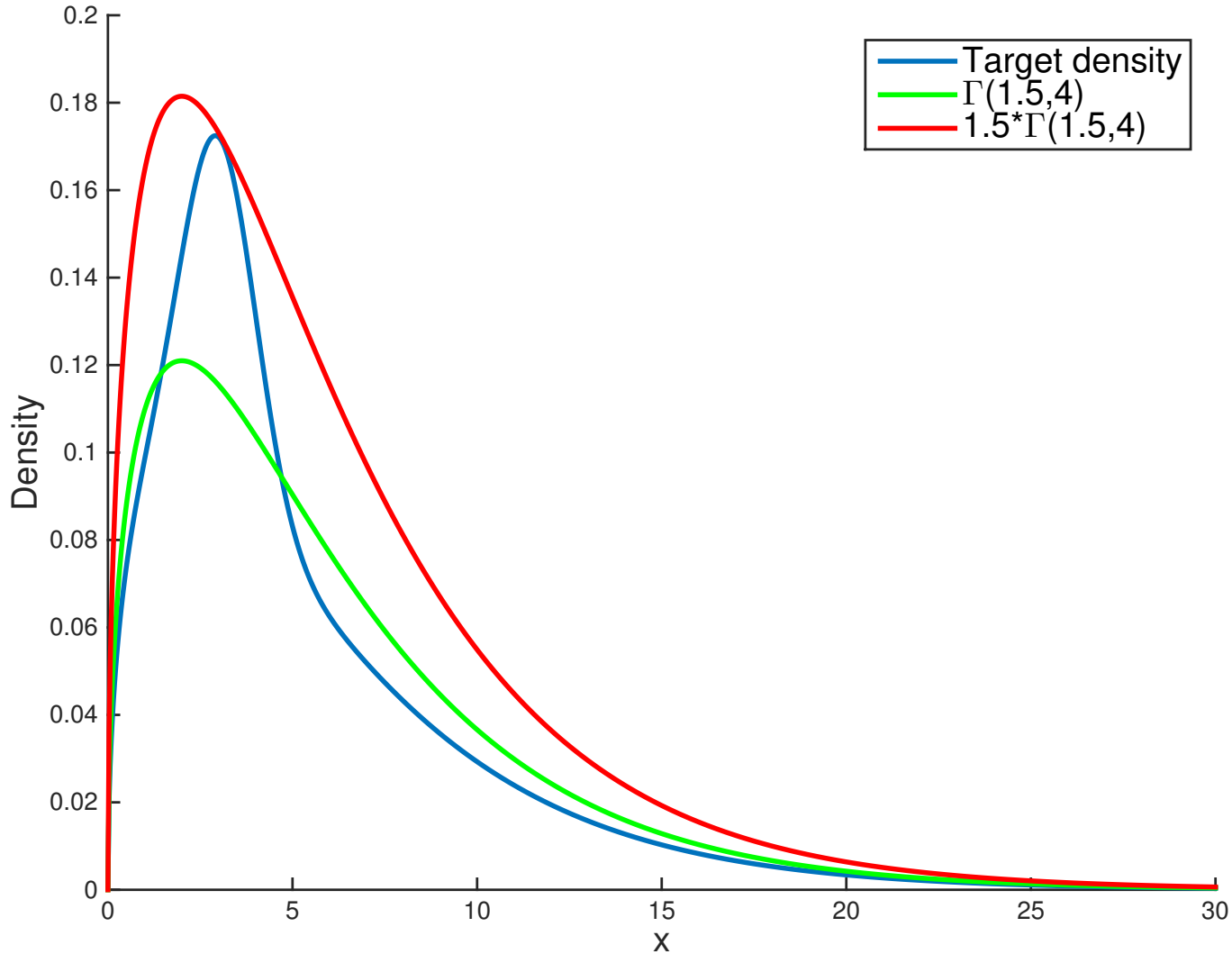
$$\pi = (0.545, 0.272, 0.181)$$

THE BASIC MCMC IDEA

- Aim: to simulate from a discrete distribution $p(x)$ when $x \in \{s_1, s_2, ..., s_k\}$.
- **MCMC**: **simulate a Markov Chain** with a **stationary distribution** that is exactly $p(x)$.
- How to set up the transition matrix P? **Metropolis-Hastings**!



Rejection sampling



Random walk Metropolis algorithm

- Initialize $\theta^{(0)}$ and iterate for $i = 1, 2, ...$

  1. Sample $\theta_p | \theta^{(i-1)} \sim N\left(\theta^{(i-1)}, c \cdot \Sigma\right)$ (the **proposal distribution**)

  2. Compute the **acceptance probability**

$$\alpha = \min\left(1, \frac{p(\theta_p|\mathbf{y})}{p(\theta^{(i-1)}|\mathbf{y})}\right)$$

  3. With probability α set $\theta^{(i)} = \theta_p$ and $\theta^{(i)} = \theta^{(i-1)}$ otherwise.

Random walk Metropolis, cont.

- Assumption: we can compute $p(\theta_p|\mathbf{y})$ for any θ.
- Proportionality constant in $p(\theta_p|\mathbf{y})$ does not matter. It will cancel in α

$$\alpha = \min\left(1, \frac{c \cdot p(\theta_p|\mathbf{y})}{c \cdot p(\theta^{(i-1)}|\mathbf{y})}\right) = \min\left(1, \frac{p(\theta_p|\mathbf{y})}{p(\theta^{(i-1)}|\mathbf{y})}\right)$$

- So we many use tattoo-version: $p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta)$

$$\alpha = \min\left(1, \frac{p\left(\mathbf{y}|\theta_p\right) p\left(\theta_p\right)}{p\left(\mathbf{y}|\theta^{(i-1)}\right) p\left(\theta^{(i-1)}\right)}\right)$$

- We can generalize the proposal $\theta_p|\theta^{(i-1)} \sim N\left(\theta^{(i-1)}, c \cdot \Sigma\right)$ to

$$\theta_p|\theta^{(i-1)} \sim q\left(\cdot|\theta^{(i-1)}\right)$$

where $q\left(\cdot|\theta^{(i-1)}\right)$ is symmetric is its arguments

$$q\left(y|x\right) = q\left(x|y\right)$$

Random walk Metropolis, cont.

- Common choices of Σ in proposal $N\left(\theta^{(i-1)}, c \cdot \Sigma\right)$:
  - $\Sigma = I$ (may propose 'off the cigar')
  - $\Sigma = J^{-1}_{\hat{\theta},\mathbf{y}}$ (propose 'along the cigar')
  - Adaptive. Start with $\Sigma = I$ and then recompute Σ from an initial simulation run.
- c is set so that average acceptance probability is roughly 25-30%.
- A good proposal:
  - **Easy to sample**
  - **Easy to compute** α
  - Proposals should take reasonably **large steps** in θ-space
  - Proposals should **not be reject too often**.

The Metropolis-Hastings algorithm

- Generalization when the proposal density is not symmetric.

- Initialize $\theta^{(0)}$ and iterate for $i = 1, 2, ...$

  1. Sample $\theta_p \sim q\left(\cdot|\theta^{(i-1)}\right)$ (the **proposal distribution**)

  2. Compute the **acceptance probability**

$$\alpha = \min\left(1, \frac{p(\mathbf{y}|\theta_p)p(\theta_p)}{p(\mathbf{y}|\theta^{(i-1)})p(\theta^{(i-1)})} \frac{q\left(\theta^{(i-1)}|\theta_p\right)}{q\left(\theta_p|\theta^{(i-1)}\right)}\right)$$

  3. With probability α set $\theta^{(i)} = \theta_p$ and $\theta^{(i)} = \theta^{(i-1)}$ otherwise.

The independence sampler

- **Independence sampler**: $q\left(\theta_p|\theta^{(i-1)}\right) = q\left(\theta_p\right)$.
- Proposal is independent of previous draw.
- Example:

$$\theta_p \sim t_v\left(\hat{\theta}, J^{-1}_{\hat{\theta},\mathbf{y}}\right),$$

  where $\hat{\theta}$ and , $J_{\hat{\theta},\mathbf{y}}$ are computed by numerical optimization.
- Can be very **efficient**, but has a tendency to **get stuck**.
- Make sure that $q\left(\theta_p\right)$ has **heavier tails** than $p(\theta|\mathbf{y})$.

Metropolis-Hastings within Gibbs

- **Gibbs sampling** from $p(\theta_1, \theta_2, \theta_3|\mathbf{y})$
  - Sample $p(\theta_1|\theta_2, \theta_3, \mathbf{y})$
  - Sample $p(\theta_2|\theta_1, \theta_3, \mathbf{y})$
  - Sample $p(\theta_3|\theta_1, \theta_2, \mathbf{y})$
- When a **full conditional is not easily sampled** we can simulate from it using MH.
- Example: at ith iteration, propose θ_2 from $q(\theta_2|\theta_1, \theta_3, \theta_2^{(i-1)}, \mathbf{y})$. Accept/reject.
- Gibbs sampling is a special case of MH when $q(\theta_2|\theta_1, \theta_3, \theta_2^{(i-1)}, \mathbf{y}) = p(\theta_2|\theta_1, \theta_3, \mathbf{y})$, which gives $\alpha = 1$. Always accept.

The efficiency of MCMC

- $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ are **dependent** (autocorrelated).
- How efficient is my MCMC compared to iid sampling?
- If $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ are iid with variance σ^2, then

$$\text{Var}(\bar{\theta}) = \frac{\sigma^2}{N}.$$

- If $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ are generated by MCMC

$$\text{Var}(\bar{\theta}) = \frac{\sigma^2}{N}\left(1 + 2\sum_{k=1}^{\infty}\rho_k\right)$$

  where $\rho_k = Corr(\theta^{(i)}, \theta^{(i+k)})$ is the autocorrelation at lag k.
- **Inefficiency factor**

$$\text{IF} = 1 + 2\sum_{k=1}^{\infty}\rho_k$$

- **Effective sample size** from MCMC

$$\text{ESS} = N/\text{IF}$$

Burn-in and convergence

- **How long burn-in?**
- **How long to sample** after burn-in?
- To **thin** of not to thin? Only keeping every h draw reduces autocorrelation.
- **Convergence diagnostics**
  - Raw plots of simulated sequences (trajectories)
  - CUSUM plots + Local means
  - Potential scale reduction factor, R.

Bayesian Learning - Lecture 9

Mattias Villani and Per Sidén

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Lecture overview

- Hamiltonian Monte Carlo
- Stan
- Variational Bayes

Hamiltonian Monte Carlo

- **Motivation:** Assume that $\theta = (\theta_1, ..., \theta_p)$. If p is large, then most of the mass of $p(\theta|y)$ is usually located on some subregion in $\mathbb{R}^p$ with complicated geometry.
- Finding a good proposal distribution $q\left(\cdot|\theta^{(i-1)}\right)$ for the MH algorithm might be hard
  $\Rightarrow$ Use very small step sizes or few accepted proposed samples.

Hamiltonian Monte Carlo

- **Motivation:** Assume that $\theta = (\theta_1, ..., \theta_p)$. If p is large, then most of the mass of $p(\theta|y)$ is usually located on some subregion in $\mathbb{R}^p$ with complicated geometry.
- Finding a good proposal distribution $q\left(\cdot|\theta^{(i-1)}\right)$ for the MH algorithm might be hard
  $\Rightarrow$ Use very small step sizes or few accepted proposed samples.
- **Hamiltonian Monte Carlo** (**HMC**) borrows ideas from physics to allow more rapid movements in the posterior distribution.
- HMC adds an auxiliary **momentum** parameter $\phi = (\phi_1, \ldots, \phi_p)$ and samples from $p(\theta, \phi|y) = p(\theta|y)\, p(\phi)$.

Hamiltonian Monte Carlo

- Background from physics: **Hamiltonian** system $H(\theta, \phi) = U(\theta) + K(\phi)$, where U is the potential energy and K is the kinetic energy.
- Dynamics:

$$\frac{d\theta_i}{dt} = \frac{\partial H}{\partial \phi_i} = \frac{\partial K}{\partial \phi_i},$$
$$\frac{d\phi_i}{dt} = -\frac{\partial H}{\partial \theta_i} = -\frac{\partial U}{\partial \theta_i}$$

- Use $U(\theta) = -\log[p(\theta)\, p(y|\theta)]$.
- Use $\phi \sim N(0, M)$ and $K(\phi) = -\log[p(\phi)] = \frac{1}{2}\phi^T M^{-1}\phi + \text{const}$, where M is the mass matrix (often diagonal).

Hamiltonian Monte Carlo

- This gives the system:

$$\frac{d\theta_i}{dt} = \left[M^{-1}\phi\right]_i,$$
$$\frac{d\phi_i}{dt} = \frac{\partial \log p\left(\theta|y\right)}{\partial \theta_i}$$

which can be simulated using the **leapfrog algorithm**

$$\phi_i\left(t+\frac{\varepsilon}{2}\right) = \phi_i\left(t\right) - \frac{\varepsilon}{2}\frac{\partial \log p\left(\theta(t)|y\right)}{\partial \theta_i},$$
$$\theta\left(t+\varepsilon\right) = \theta\left(t\right) + \varepsilon M^{-1}\phi(t),$$
$$\phi_i\left(t+\varepsilon\right) = \phi_i\left(t+\frac{\varepsilon}{2}\right) - \frac{\varepsilon}{2}\frac{\partial \log p\left(\theta(t)|y\right)}{\partial \theta_i},$$

where ε is the step size.

The Hamiltonian Monte Carlo algorithm

- Initialize $\theta^{(0)}$ and iterate for $i = 1, 2, ...$

  1. Sample the starting momentum $\phi_s \sim N(0, M)$
  2. Simulate new values for (θ_p, ϕ_p) by iterating the leapfrog algorithm L times, starting in $\left(\theta^{(i-1)}, \phi_s\right)$.
  3. Compute the **acceptance probability**

$$\alpha = \min\left(1, \frac{p(y|\theta_p)p(\theta_p)}{p(y|\theta^{(i-1)})p(\theta^{(i-1)})}\frac{p(\phi_p)}{p(\phi_s)}\right)$$

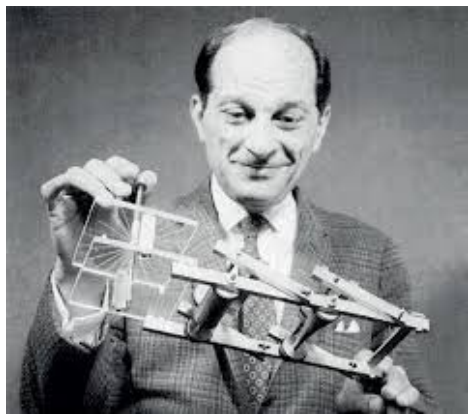
  4. With probability α set $\theta^{(i)} = \theta_p$ and $\theta^{(i)} = \theta^{(i-1)}$ otherwise.

- Imagine a hockey pluck sliding over a friction-less surface: illustration.
- The stepsize ε, number of leapfrog iterations L and mass matrix M are tuning parameters that can be tuned during the burn-in phase.

Stan

- **Stan** is a probabilistic programming language based on HMC.
- Allows for Bayesian inference in many models with automatic implementation of the MCMC sampler.
- Named after Stanislaw Ulam (1909-1984), co-inventor of the Monte Carlo algorithm.
- Written in C++ but can be run from R using the package `rstan`

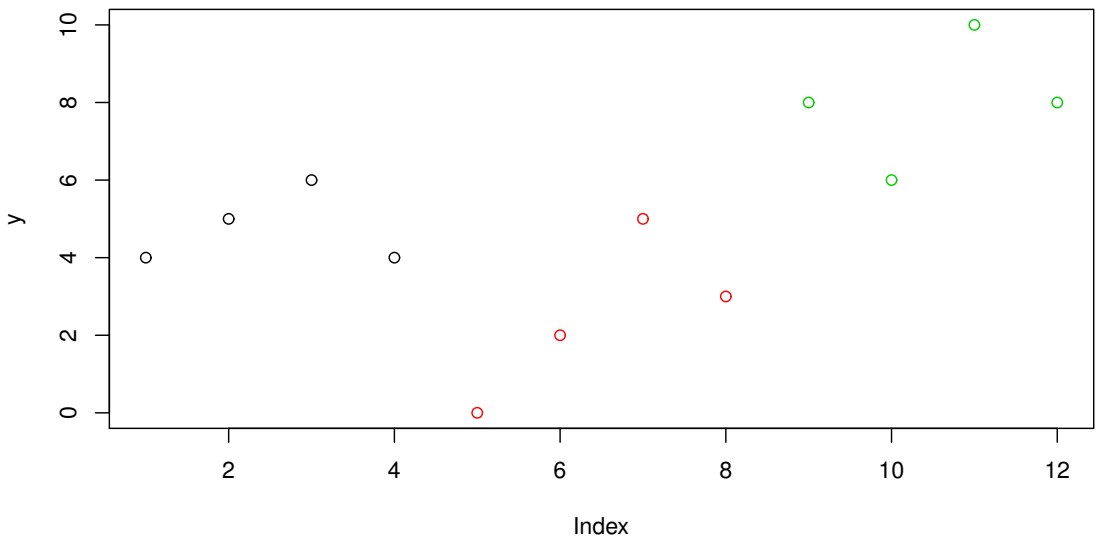
Stan logo



Stanislaw Ulam

STAN - TOY EXAMPLE: THREE PLANTS

- Three plants were observed for four months, measuring the number of flowers



Stan Model 1: iid normal

$$y_i \overset{iid}{\sim} N\left(\mu, \sigma^2\right)$$

```
library(rstan)
y = c(4,5,6,4,0,2,5,3,8,6,10,8)
N = length(y)

StanModel = '
data {
  int<lower=0> N; // Number of observations
  int<lower=0> y[N]; // Number of flowers
}
parameters {
  real mu;
  real<lower=0> sigma2;
}
model {
  mu ~ normal(0,100); // Normal with mean 0, st.dev. 100
  sigma2 ~ scaled_inv_chi_square(1,2); // Scaled-inv-chi2 with nu 1, sigma 2
  for(i in 1:N)
    y[i] ~ normal(mu,sqrt(sigma2));
}'
```

STAN MODEL 2: MULTILEVEL NORMAL

$$y_{i,p} \sim N\left(\mu_p, \sigma_p^2\right), \quad \mu_p \sim N\left(\mu, \sigma^2\right)$$

```
StanModel = '
data {
  int<lower=0> N; // Number of observations
  int<lower=0> y[N]; // Number of flowers
  int<lower=0> P; // Number of plants
}
transformed data {
  int<lower=0> M; // Number of months
  M = N / P;
}
parameters {
  real mu;
  real<lower=0> sigma2;
  real mup[P];
  real sigmap2[P];
}
model {
  mu ~ normal(0,100); // Normal with mean 0, st.dev. 100
  sigma2 ~ scaled_inv_chi_square(1,2); // Scaled-inv-chi2 with nu 1, sigma 2
  for(p in 1:P){
    mup[p] ~ normal(mu,sqrt(sigma2));
    for(m in 1:M)
      y[M*(p-1)+m] ~ normal(mup[p],sqrt(sigmap2[p]));
  }
}'
```

Stan Model 3: multilevel Poisson

$$y_{i,p} \sim Poisson\left(\mu_p\right), \quad \mu_p \sim logN\left(\mu, \sigma^2\right)$$

```
StanModel = '
data {
  int<lower=0> N; // Number of observations
  int<lower=0> y[N]; // Number of flowers
  int<lower=0> P; // Number of plants
}
transformed data {
  int<lower=0> M; // Number of months
  M = N / P;
}
parameters {
  real mu;
  real<lower=0> sigma2;
  real mup[P];
}
model {
  mu ~ normal(0,100); // Normal with mean 0, st.dev. 100
  sigma2 ~ scaled_inv_chi_square(1,2); // Scaled-inv-chi2 with nu 1, sigma 2
  for(p in 1:P){
    mup[p] ~ lognormal(mu,sqrt(sigma2)); // Log-normal
    for(m in 1:M)
      y[M*(p-1)+m] ~ poisson(mup[p]); // Poisson
  }
}'
```

STAN: FIT MODEL AND ANALYZE OUTPUT

```
data = list(N=N, y=y, P=P)
burnin = 1000
niter = 2000
fit = stan(model_code=StanModel,data=data,
           warmup=burnin,iter=niter,chains=4)

# Print the fitted model
print(fit,digits_summary=3)

# Extract posterior samples
postDraws <- extract(fit)

# Do traceplots of the first chain
par(mfrow = c(1,1))
plot(postDraws$mu[1:(niter-burnin)],type="l",ylab="mu",main="Traceplot")

# Do automatic traceplots of all chains
traceplot(fit)

# Bivariate posterior plots
pairs(fit)
```

Stan - useful links

- Getting started with RStan
- RStan vignette
- Stan Modeling Language User's Guide and Reference Manual
- Stan Case Studies

Variational Bayes

- Let $\theta = (\theta_1, ..., \theta_p)$. Approximate the posterior $p(\theta|y)$ with a (simpler) distribution $q(\theta)$.
- We have already seen: $q(\theta) = N\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$.
- **Mean field Variational Bayes (VB)**

$$q(\theta) = \prod_{i=1}^{p} q_i(\theta_i)$$

- **Parametric VB**, where $q_\lambda(\theta)$ is a parametric family with parameters λ.
- Find the $q(\theta)$ that **minimizes the Kullback-Leibler distance** between the true posterior p and the approximation q:

$$KL(q, p) = \int q(\theta) \ln \frac{q(\theta)}{p(\theta|y)} d\theta = E_q \left[\ln \frac{q(\theta)}{p(\theta|y)} \right].$$

Mean field approximation

- Factorization

$$q(\theta) = \prod_{i=1}^{p} q_i(\theta_i)$$

- **No specific functional forms are assumed** for the $q_i(\theta)$.
- **Optimal densities** can be shown to satisfy:

$$q_i(\theta) \propto \exp\left(E_{-\theta_i} \ln p(\mathbf{y}, \theta)\right)$$

  where $E_{-\theta_i}(\cdot)$ is the expectation with respect to $\prod_{i \neq j} q_j(\theta_j)$.
- **Structured mean field approximation**. Group subset of parameters in tractable blocks. Similar to Gibbs sampling.

Mean field approximation - algorithm

- Initialize: $q_2^*(\theta_2), ..., q_M^*(\theta_p)$
- Repeat until convergence:
  - $q_1^*(\theta_1) \leftarrow \frac{\exp\left[E_{-\theta_1} \ln p(\mathbf{y}, \theta)\right]}{\int \exp\left[E_{-\theta_1} \ln p(\mathbf{y}, \theta)\right] d\theta_1}$
  - $\vdots$
  - $q_p^*(\theta_p) \leftarrow \frac{\exp\left[E_{-\theta_p} \ln p(\mathbf{y}, \theta)\right]}{\int \exp\left[E_{-\theta_p} \ln p(\mathbf{y}, \theta)\right] d\theta_p}$
- Note: we make no assumptions about parametric form of the $q_i(\theta)$, but the optimal $q_i(\theta)$ often turn out to be parametric (normal, gamma etc).
- The updates above then boil down to just updating of hyperparameters in the optimal densities.

Mean field approximation - Normal model

- **Model**: $X_i|\theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$.
- **Prior**: $\theta \sim N(\mu_0, \tau_0^2)$ **independent** of $\sigma^2 \sim Inv - \chi^2(\nu_0, \sigma_0^2)$.
- **Mean-field approximation**: $q(\theta, \sigma^2) = q_\theta(\theta) \cdot q_{\sigma^2}(\sigma^2)$.
- Optimal densities

$$q_\theta^*(\theta) \propto \exp\left[E_{q(\sigma^2)} \ln p(\theta, \sigma^2, \mathbf{x})\right]$$
$$q_{\sigma^2}^*(\sigma^2) \propto \exp\left[E_{q(\theta)} \ln p(\theta, \sigma^2, \mathbf{x})\right]$$

Normal model - VB algorithm

- **Variational density for σ^2**

$$\sigma^2 \sim Inv - \chi^2\left(\tilde{\nu}_n, \tilde{\sigma}_n^2\right)$$

where $\tilde{\nu}_n = \nu_0 + n$ and $\tilde{\sigma}_n = \frac{\nu_0\sigma_0^2 + \sum_{i=1}^n (x_i - \tilde{\mu}_n)^2 + n\cdot\tilde{\tau}_n^2}{\nu_0 + n}$

- **Variational density for θ**

$$\theta \sim N\left(\tilde{\mu}_n, \tilde{\tau}_n^2\right)$$

where

$$\tilde{\tau}_n^2 = \frac{1}{\frac{n}{\tilde{\sigma}_n^2} + \frac{1}{\tau_0^2}}$$

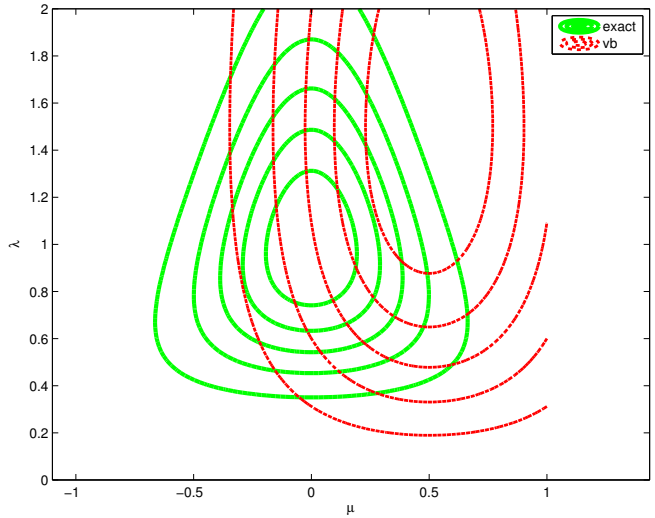
$$\tilde{\mu}_n = \tilde{w}\bar{x} + (1 - \tilde{w})\mu_0,$$

where

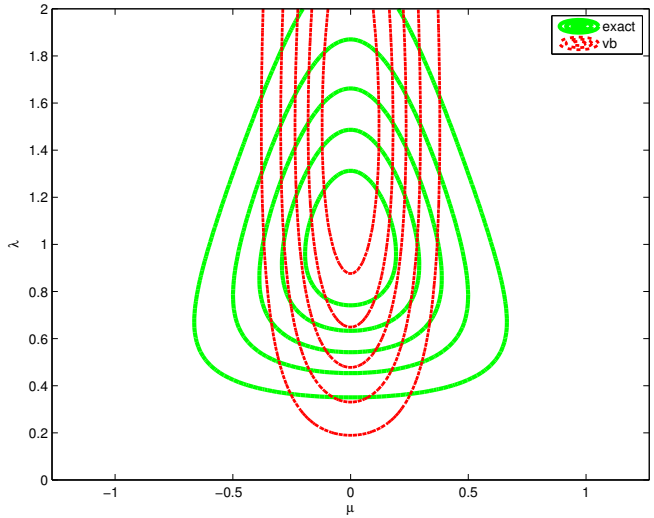
$$\tilde{w} = \frac{\frac{n}{\tilde{\sigma}_n^2}}{\frac{n}{\tilde{\sigma}_n^2} + \frac{1}{\tau_0^2}}$$

Normal example from Murphy ($\lambda = 1/\sigma^2$)

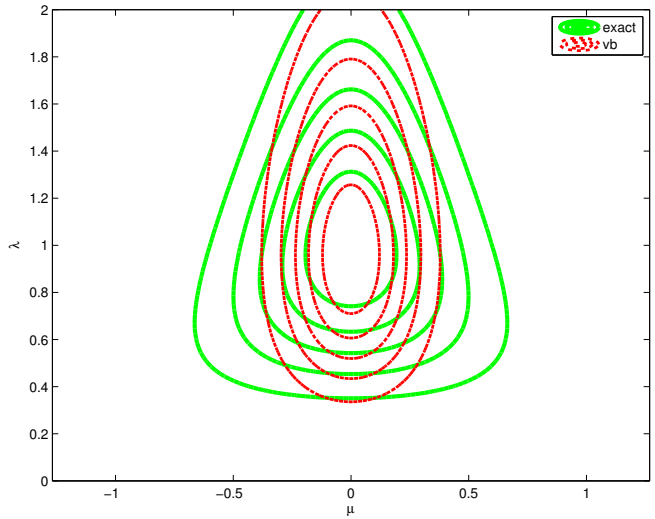
Initial values



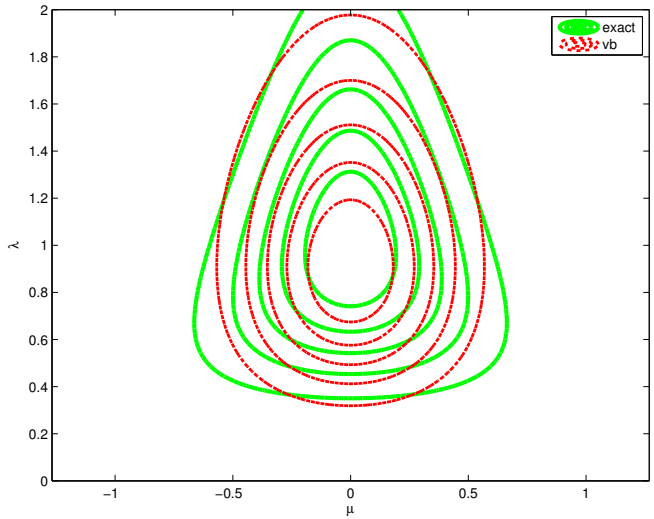
After updating q_μ



After updating q_{σ^2}



At convergence



Probit regression

- **Model**:

$$\Pr(y_i = 1|\mathbf{x}_i) = \Phi(\mathbf{x}_i^T \beta)$$

- **Prior**: $\beta \sim N(0, \Sigma_\beta)$. For example: $\Sigma_\beta = \tau^2 I$.
- **Latent variable formulation** with $u = (u_1, ..., u_n)'$

$$\mathbf{u}|\beta \sim N(\mathbf{X}\beta, 1)$$

and

$$y_i = \begin{cases} 0 & \text{if } u_i \leq 0 \\ 1 & \text{if } u_i > 0 \end{cases}$$

- Factorized **variational approximation**

$$q(\mathbf{u}, \beta) = q_{\mathbf{u}}(\mathbf{u}) q_\beta(\beta)$$

VB FOR PROBIT REGRESSION

- VB posterior

$$\beta \sim N\left(\tilde{\mu}_\beta, \left(\mathbf{X}^T\mathbf{X} + \Sigma_\beta^{-1}\right)^{-1}\right)$$

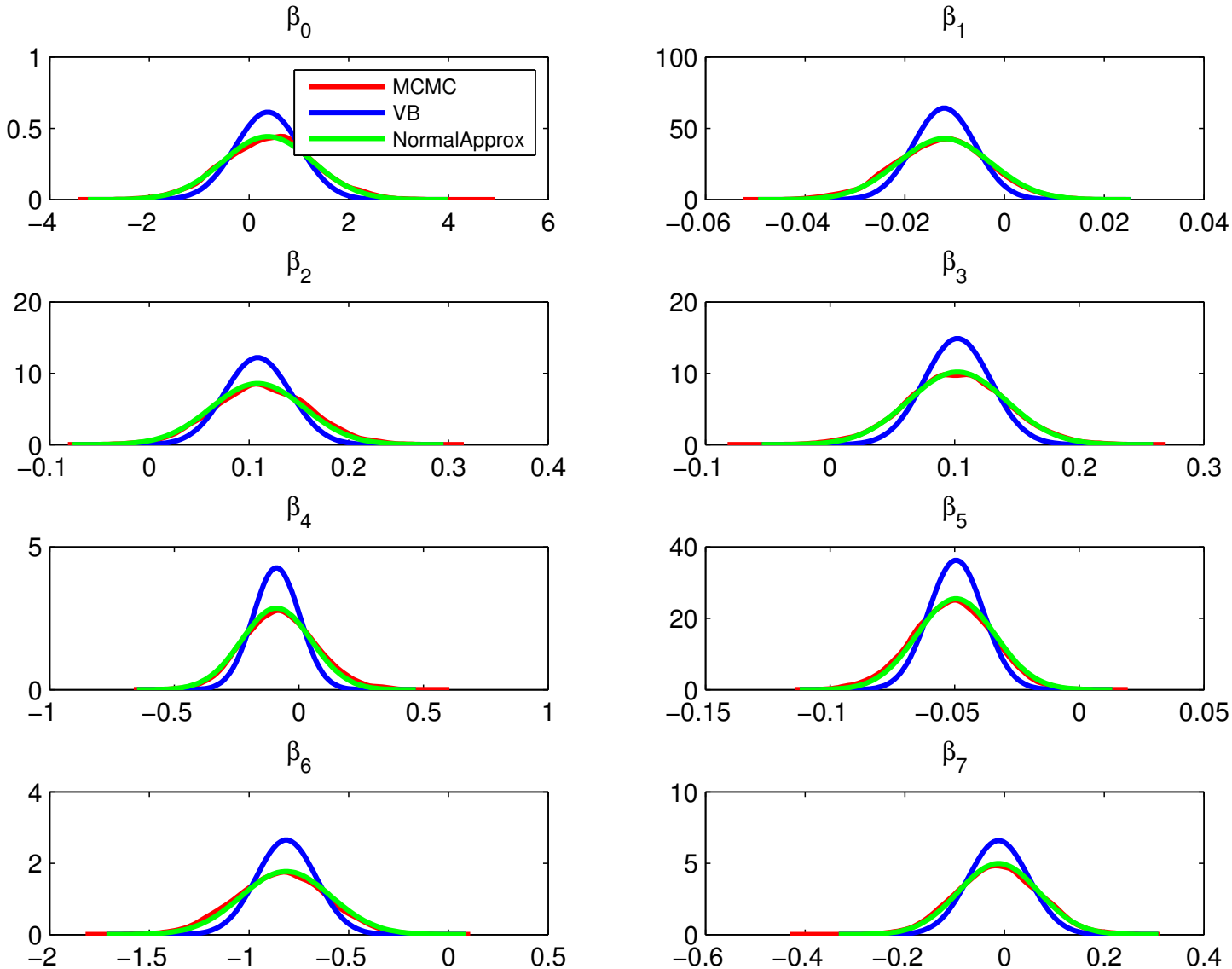
where

$$\tilde{\mu}_\beta = \left(\mathbf{X}^T\mathbf{X} + \Sigma_\beta^{-1}\right)^{-1}\mathbf{X}^T\tilde{\mu}_\mathbf{u}$$

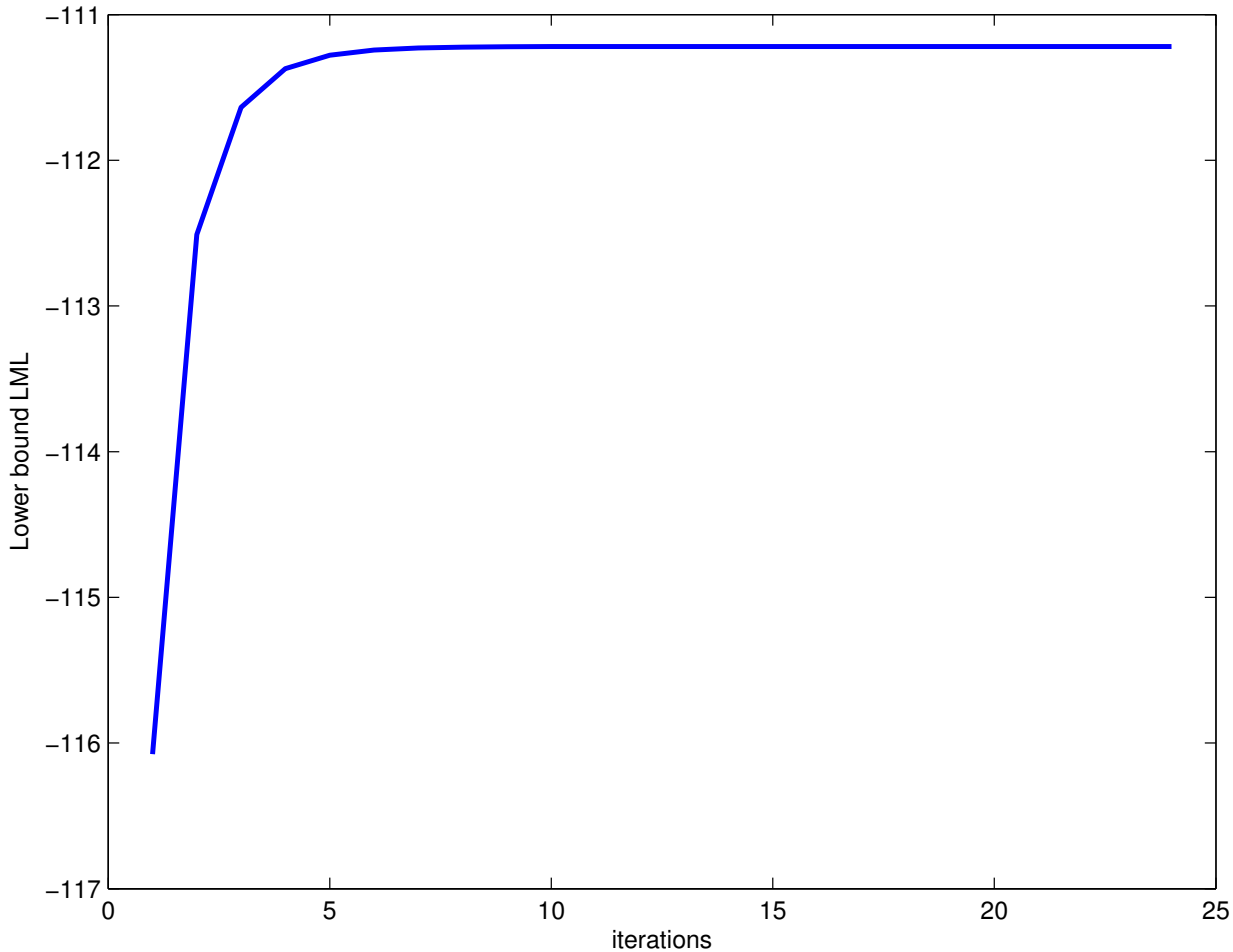
and

$$\tilde{\mu}_\mathbf{u} = \mathbf{X}\tilde{\mu}_\beta + \frac{\phi\left(\mathbf{X}\tilde{\mu}_\beta\right)}{\Phi\left(\mathbf{X}\tilde{\mu}_\beta\right)^\mathbf{y}\left[\Phi\left(\mathbf{X}\tilde{\mu}_\beta\right) - \mathbf{1}_n\right]^{\mathbf{1}_n - \mathbf{y}}}.$$

PROBIT EXAMPLE (N=200 OBSERVATIONS)



PROBIT EXAMPLE



Bayesian Learning - Lecture 10

Mattias Villani

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Overview

- Bayesian model comparison
- Marginal likelihood

USING LIKELIHOOD FOR MODEL COMPARISON

- Consider two models for the data $\mathbf{y} = (y_1, ..., y_n)$: M_1 and M_2.
- Let $p_i(\mathbf{y}|\theta_i)$ denote the data density under model M_i.
- If know θ_1 and θ_2, the **likelihood ratio** is useful

$$\frac{p_1(\mathbf{y}|\theta_1)}{p_2(\mathbf{y}|\theta_2)}.$$

- The **likelihood ratio** with ML estimates plugged in:

$$\frac{p_1(\mathbf{y}|\hat{\theta}_1)}{p_2(\mathbf{y}|\hat{\theta}_2)}.$$

- Bigger models always win in estimated likelihood ratio.
- **Hypothesis tests** are problematic for non-nested models. End results are not very useful for analysis.

Bayesian model comparison

- Just use your priors $p_1(\theta_1)$ och $p_2(\theta_2)$.
- The **marginal likelihood** for model M_k with parameters θ_k

$$p_k(y) = \int p_k(y|\theta_k)p_k(\theta_k)d\theta_k.$$

- θ_k is removed by the prior. **Not a silver bullet**. **Priors matter!**
- The **Bayes factor**

$$B_{12}(y) = \frac{p_1(y)}{p_2(y)}.$$

- **Posterior model probabilities**

$$\underbrace{\Pr(M_k|\mathbf{y})}_{\text{posterior model prob.}} \propto \underbrace{p(\mathbf{y}|M_k)}_{\text{marginal likelihood}} \cdot \underbrace{\Pr(M_k)}_{\text{prior model prob.}}$$

Bayesian hypothesis testing - Bernoulli

- **Hypothesis testing** is just a special case of model selection:

$$M_0 : x_1, ..., x_n \overset{iid}{\sim} Bernoulli(\theta_0)$$

$$M_1 : x_1, ..., x_n \overset{iid}{\sim} Bernoulli(\theta), \theta \sim Beta(\alpha, \beta)$$

$$p(x_1, ..., x_n | M_0) = \theta_0^s (1-\theta_0)^f,$$

$$\begin{aligned} p(x_1, ..., x_n | M_1) &= \int_0^1 \theta^s (1-\theta)^f B(\alpha, \beta)^{-1} \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta \\ &= B(\alpha + s, \beta + f) / B(\alpha, \beta), \end{aligned}$$

  where $B(\cdot, \cdot)$ is the **Beta function**.

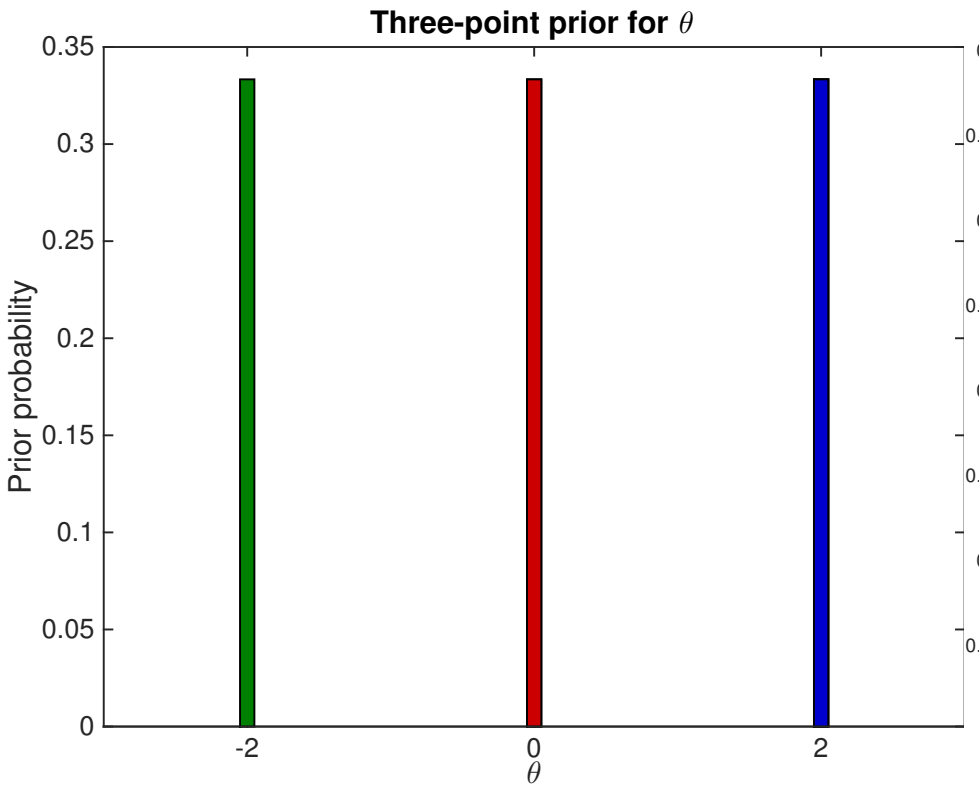
- Posterior model probabilities

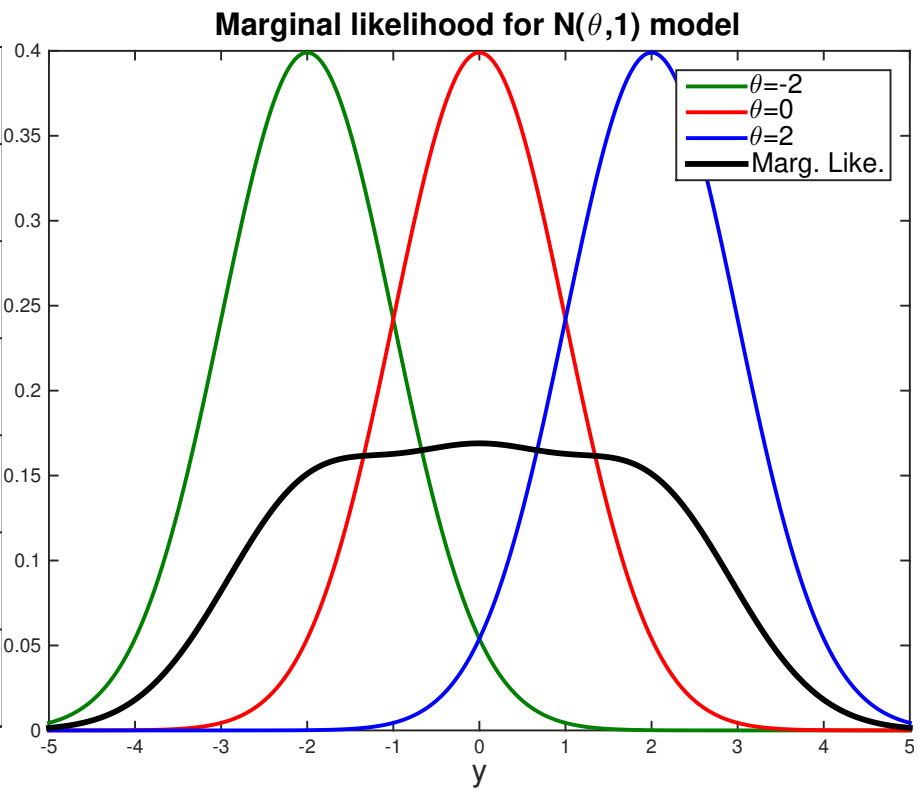
$$Pr(M_k | x_1, ..., x_n) \propto p(x_1, ..., x_n | M_k) Pr(M_k), \text{ for } k = 0, 1.$$

- The Bayes factor

$$BF(M_0; M_1) = \frac{p(x_1, ..., x_n | H_0)}{p(x_1, ..., x_n | H_1)} = \frac{\theta_0^s (1-\theta_0)^f B(\alpha, \beta)}{B(\alpha + s, \beta + f)}.$$

PRIORS MATTER





Example: Geometric vs Poisson

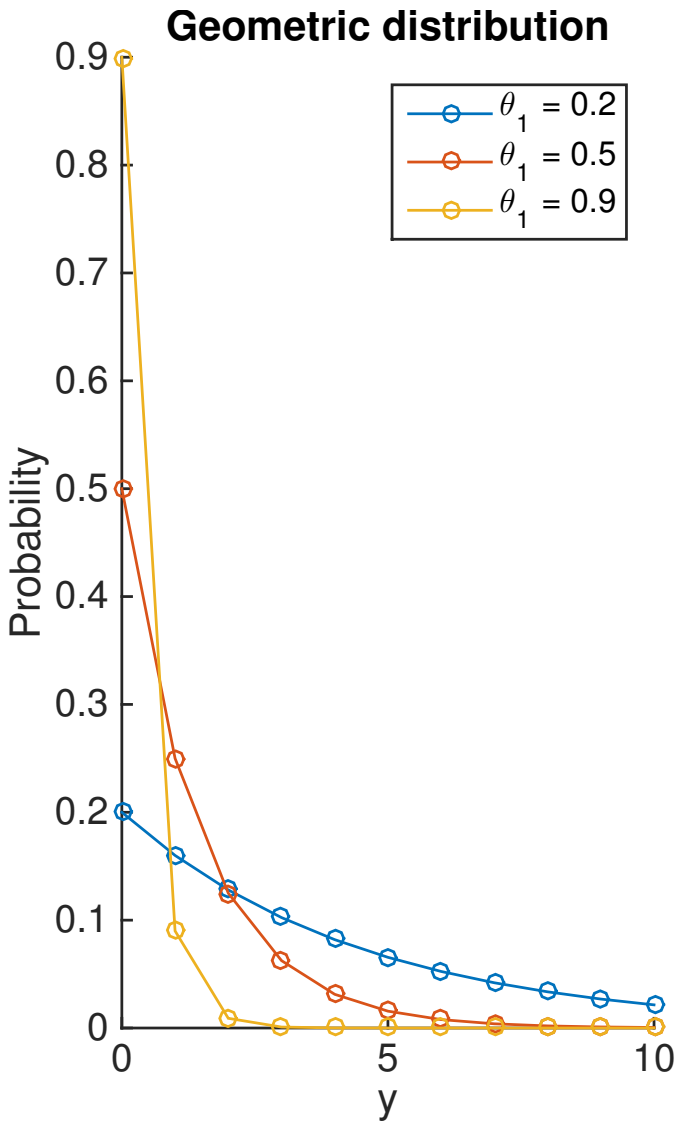
- Model 1 - **Geometric** with Beta prior:
  - $y_1, ..., y_n|\theta_1 \sim Geo(\theta_1)$
  - $\theta_1 \sim Beta(\alpha_1, \beta_1)$
- Model 2 - **Poisson** with Gamma prior:
  - $y_1, ..., y_n|\theta_2 \sim Poisson(\theta_2)$
  - $\theta_2 \sim Gamma(\alpha_2, \beta_2)$
- Marginal likelihood for M_1

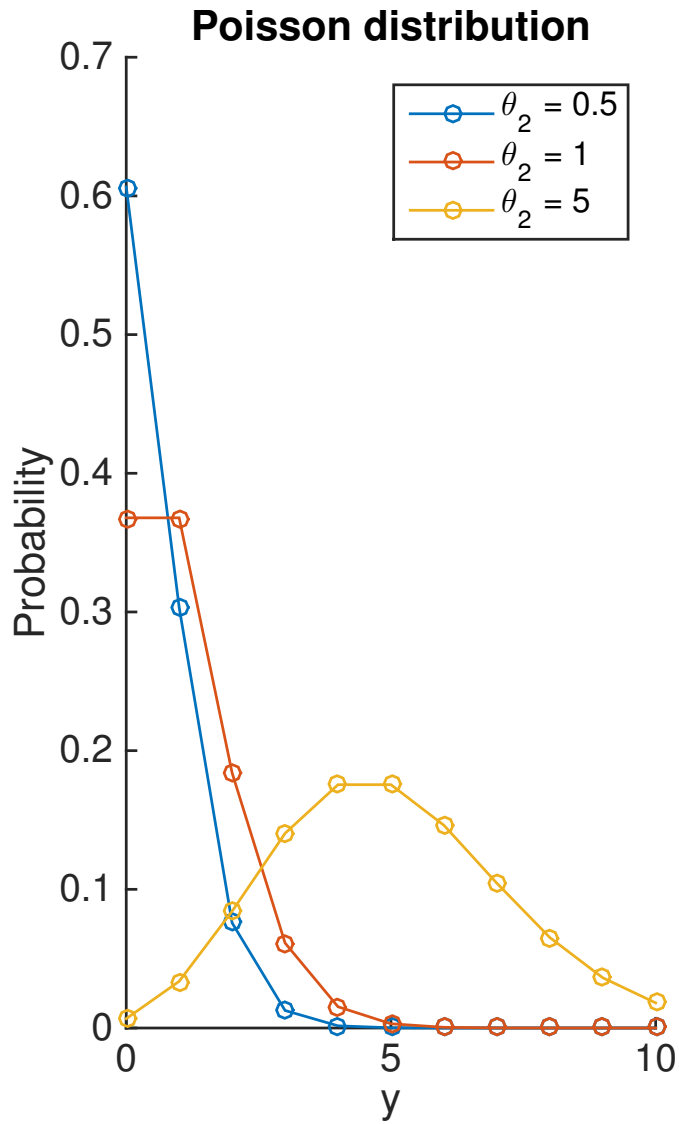
$$p_1(y_1, ..., y_n) = \int p_1(y_1, ..., y_n|\theta_1)p(\theta_1)d\theta_1$$
$$= \frac{\Gamma(\alpha_1 + \beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \frac{\Gamma(n + \alpha_1)\Gamma(n\bar{y} + \beta_1)}{\Gamma(n + n\bar{y} + \alpha_1 + \beta_1)}$$

- Marginal likelihood for M_2

$$p_2(y_1, ..., y_n) = \frac{\Gamma(n\bar{y} + \alpha_2)\beta_2^{\alpha_2}}{\Gamma(\alpha_2)(n + \beta_2)^{n\bar{y}+\alpha_2}} \frac{1}{\prod_{i=1}^n y_i!}$$

Geometric and Poisson





Geometric vs Poisson, cont.

- Priors match prior predictive means:

$$E(y_i|M_1) = E(y_i|M_2) \quad \Longleftrightarrow \quad \alpha_1\alpha_2 = \beta_1\beta_2$$

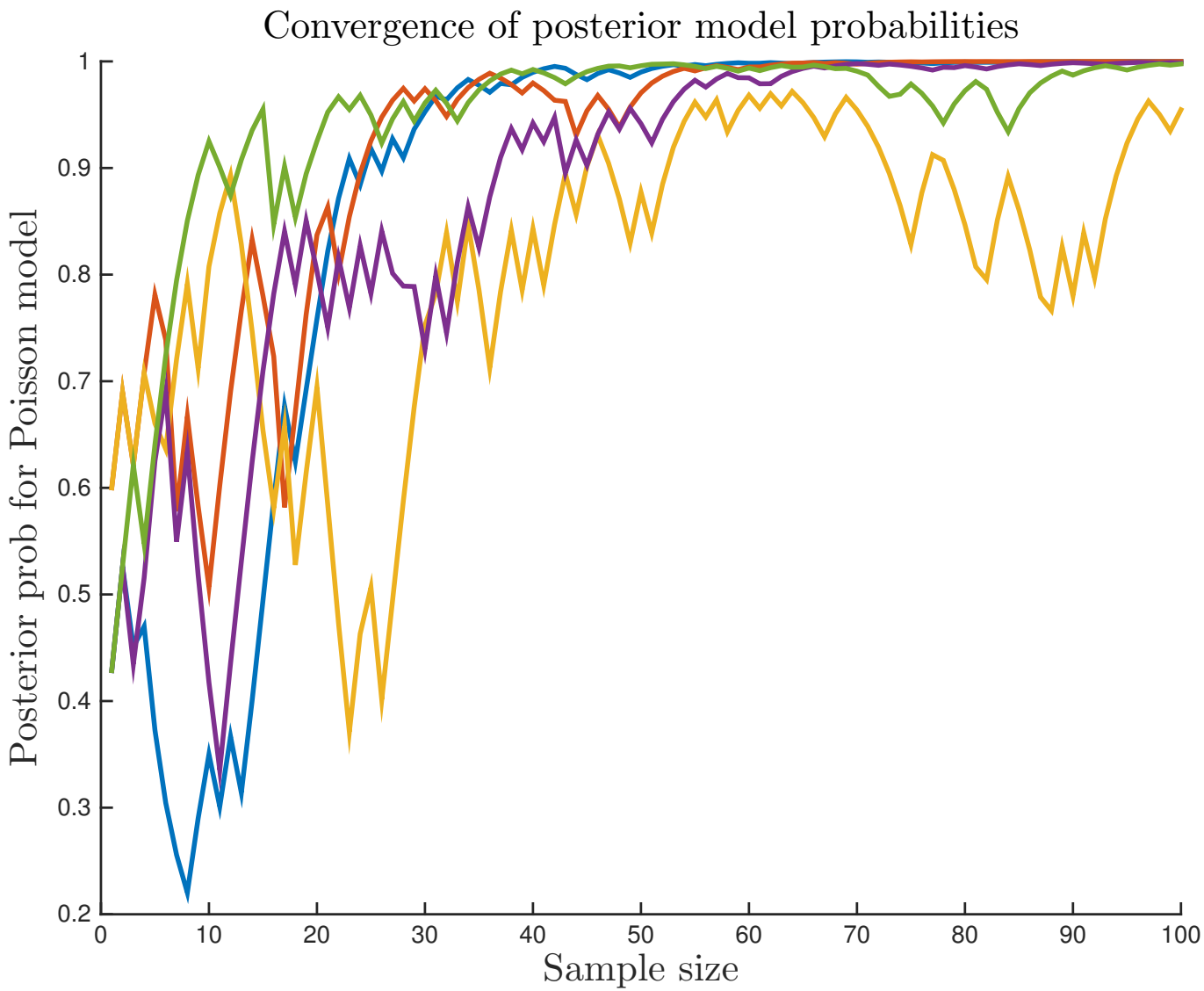
- **Data**: $y_1 = 0,\ y_2 = 0$.

	$\alpha_1 = 1, \beta_1 = 2$ $\alpha_2 = 2, \beta_2 = 1$	$\alpha_1 = 10, \beta_1 = 20$ $\alpha_2 = 20, \beta_2 = 10$	$\alpha_1 = 100, \beta_1 = 200$ $\alpha_2 = 200, \beta_2 = 100$
BF_{12}	1.5	4.54	5.87
$\Pr(M_1\|\mathbf{y})$	0.6	0.82	0.85
$\Pr(M_2\|\mathbf{y})$	0.4	0.18	0.15

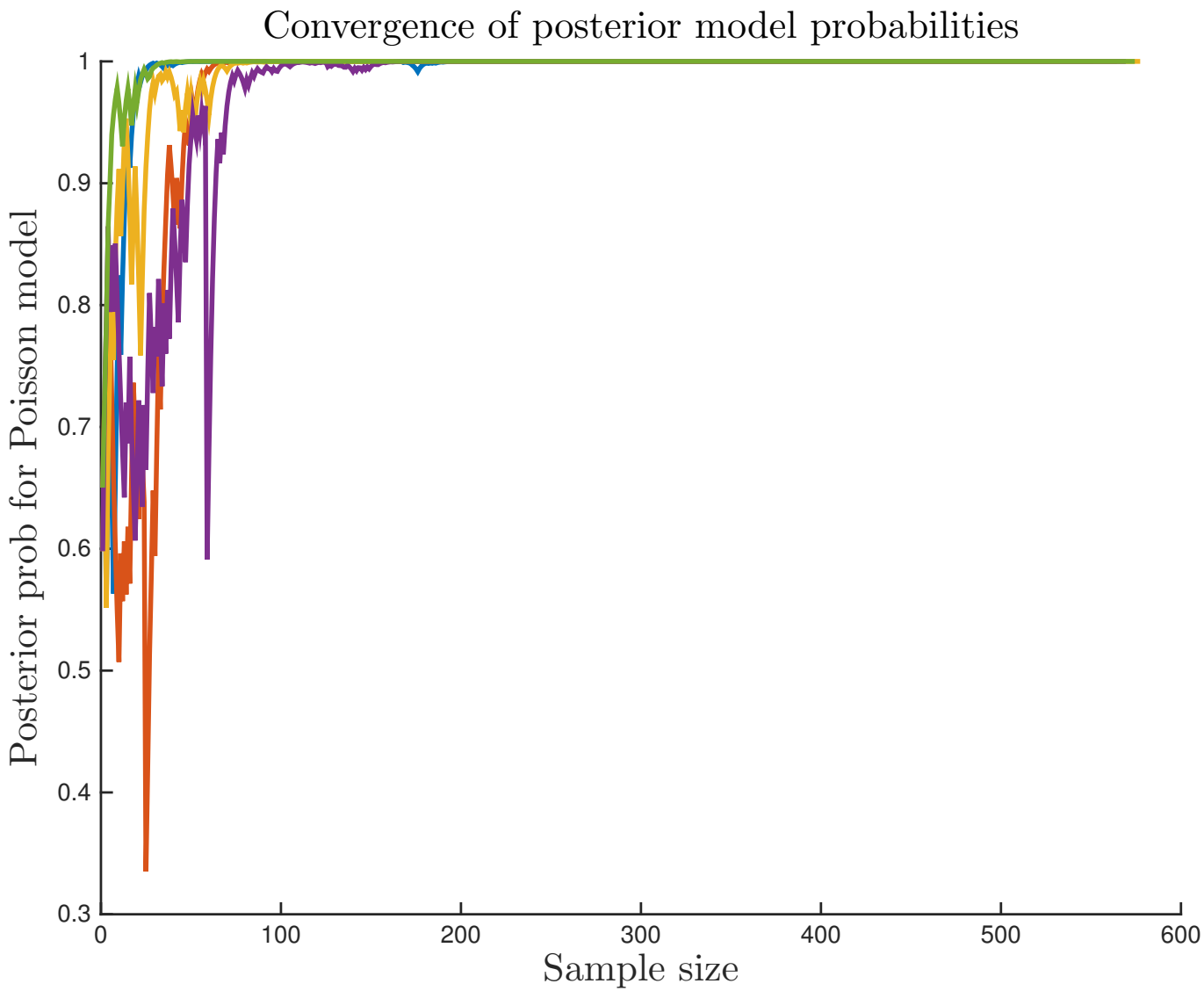
- **Data**: $y_1 = 3,\ y_2 = 3$.

	$\alpha_1 = 1, \beta_1 = 2$ $\alpha_2 = 2, \beta_2 = 1$	$\alpha_1 = 10, \beta_1 = 20$ $\alpha_2 = 20, \beta_2 = 10$	$\alpha_1 = 100, \beta_1 = 200$ $\alpha_2 = 200, \beta_2 = 100$
BF_{12}	0.26	0.29	0.30
$\Pr(M_1\|\mathbf{y})$	0.21	0.22	0.23
$\Pr(M_2\|\mathbf{y})$	0.79	0.78	0.77

Geometric vs Poisson for Pois(1) data



Geometric vs Poisson for Pois(1) data



Model choice in multivariate time series

- Multivariate time series

$$\mathbf{x}_t = \alpha\beta'\mathbf{z}_t + \Phi_1\mathbf{x}_{t-1} + ...\Phi_k\mathbf{x}_{t-k} + \Psi_1 + \Psi_2 t + \Psi_3 t^2 + \varepsilon_t$$

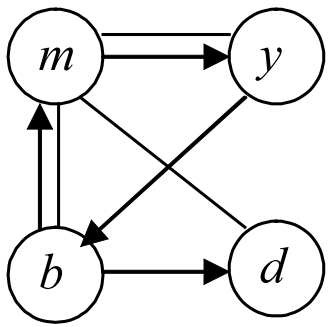
- Need to choose:
  - **Lag length**, $(k = 1, 2.., 4)$
  - **Trend model** $(s = 1, 2, ..., 5)$
  - **Long-run (cointegration) relations** $(r = 0, 1, 2, 3, 4)$.

The most probable (k, r, s) combinations in the Danish monetary data.

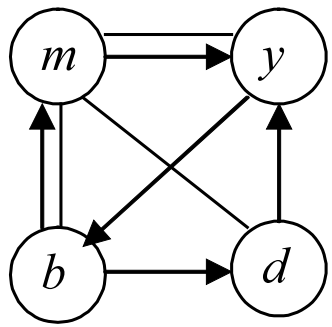
k	1	1	1	1	1	1	1	1	0	1
r	3	3	2	4	2	1	2	3	4	3
s	3	2	2	2	3	3	4	4	4	5
$p(k, r, s\|y, x, z)$	.106	.093	.091	.060	.059	.055	.054	.049	.040	.038

Graphical models for multivariate time series

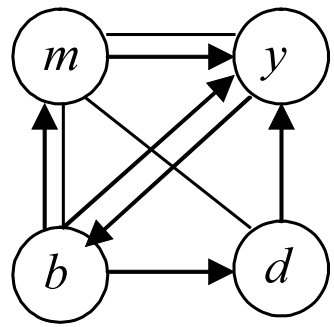
- **Graphical models** for multivariate time series.
- Zero-restrictions on the effect from time series i on time series j, for all lags. (**Granger Causality**).
- Zero-restrictions on the elements of the inverse covariance matrix of the errors.



$p(G|\mathbf{X}) = 0.0033$



$p(G|\mathbf{X}) = 0.0028$



$p(G|\mathbf{X}) = 0.0025$

Properties of Bayesian model comparison

- Coherence of pair-wise comparisons

$$B_{12} = B_{13} \cdot B_{32}$$

- **Consistency** when true model is in $\mathcal{M} = \{M_1, ..., M_K\}$

$$\Pr(M = M_{TRUE}|\mathbf{y}) \to 1 \quad \text{as} \quad n \to \infty$$

- "KL-consistency" when $M_{TRUE} \notin \mathcal{M}$

$$\Pr(M = M^*|\mathbf{y}) \to 1 \quad \text{as} \quad n \to \infty$$

  where M^* is the model that minimizes Kullback-Leibler distance between $p_M(\mathbf{y})$ and $p_{TRUE}(\mathbf{y})$.
- Smaller models always win when priors are very vague.
- **Improper priors** cannot be used for model comparison.

Marginal likelihood measures out-of-sample predictive performance

- The marginal likelihood can be decomposed as

$$p(y_1, ..., y_n) = p(y_1)p(y_2|y_1) \cdots p(y_n|y_1, y_2, ..., y_{n-1})$$

- If we assume that y_i is independent of $y_1, ..., y_{i-1}$ conditional on θ:

$$p(y_i|y_1, ..., y_{i-1}) = \int p(y_i|\theta)p(\theta|y_1, ..., y_{i-1})d\theta$$

- The prediction of y_1 is based on the prior of θ, and is therefore sensitive to the prior.
- The prediction of y_n uses almost all the data to infer θ. Very little influenced by the prior when n is not small.

Normal example

- **Model**: $y_1, ..., y_n|\theta \sim N(\theta, \sigma^2)$ with σ^2 known.
- **Prior**: $\theta \sim N(0, \kappa^2\sigma^2)$.
- Intermediate posterior at time $i-1$

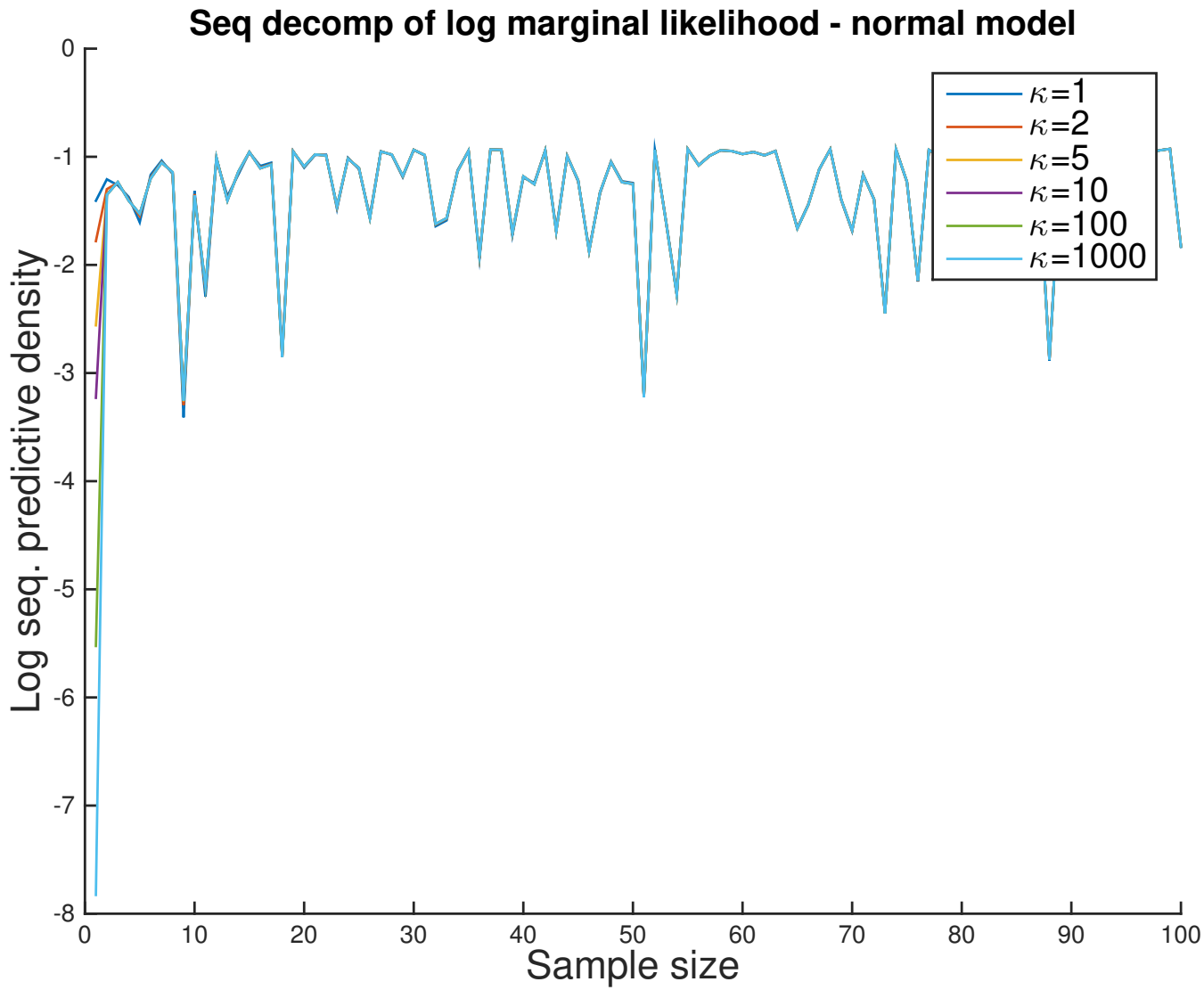
$$\theta|y_1, ..., y_{i-1} \sim N\left[w_i(\kappa) \cdot \bar{y}_{i-1}, \frac{\sigma^2}{i-1+\kappa^{-2}}\right]$$

  where $w_i(\kappa) = \frac{i-1}{i-1+\kappa^{-2}}$.
- Predictive density at time $i-1$

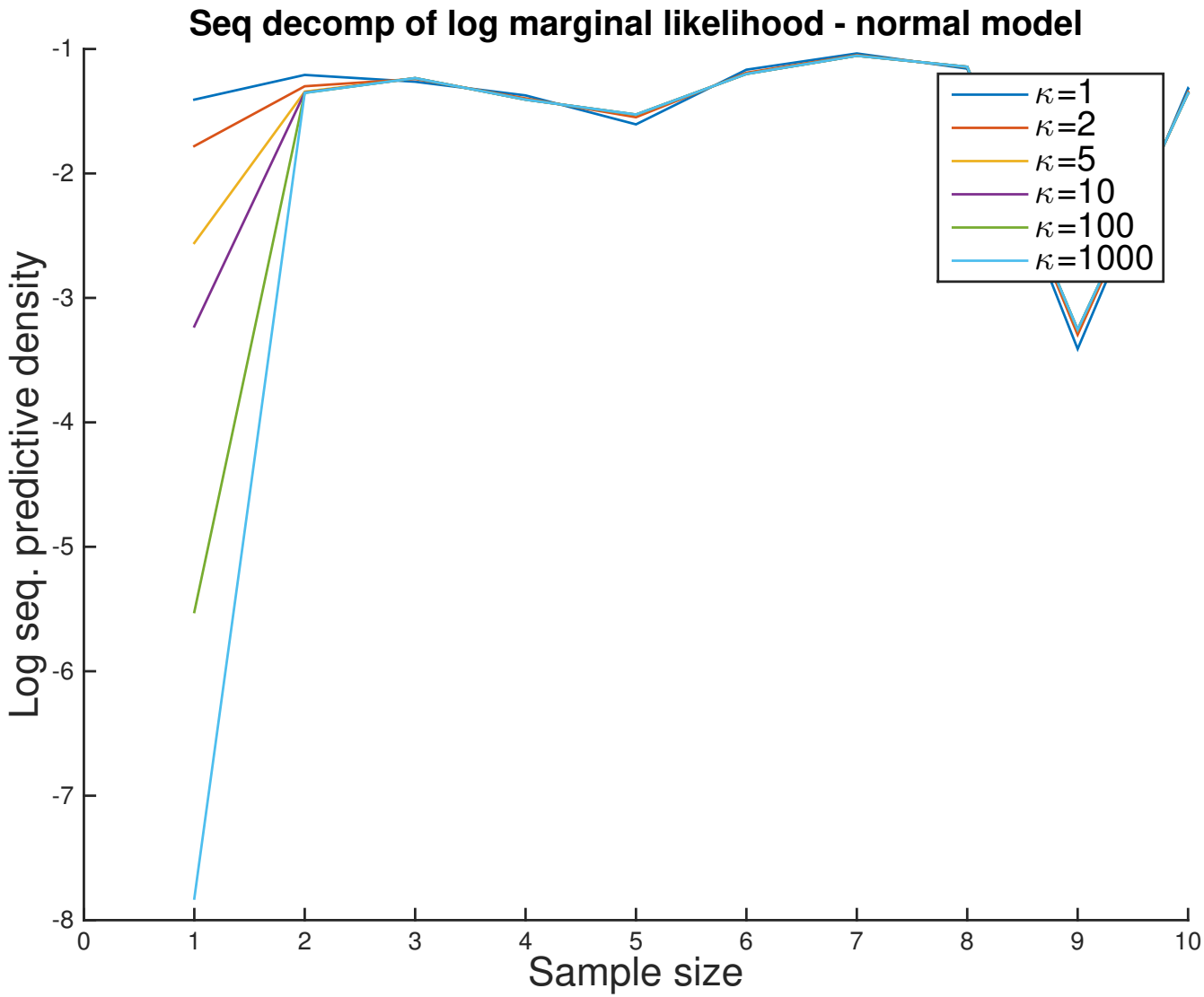
$$y_i|y_1, ..., y_{i-1} \sim N\left[w_i(\kappa) \cdot \bar{y}_{i-1}, \sigma^2\left(1 + \frac{1}{i-1+\kappa^{-2}}\right)\right]$$

- Terms with i large: $y_i|y_1, ..., y_{i-1} \overset{approx}{\sim} N\left(\bar{y}_{i-1}, \sigma^2\right)$, not sensitive to κ
- For $i = 1$, $y_1 \sim N\left[0, \sigma^2\left(1 + \frac{1}{\kappa^{-2}}\right)\right]$ can be very sensitive to κ.

First observation is sensitive to κ



First observation is sensitive to κ



Log Predictive Score - LPS

- To reduce sensitivity to the prior: sacrifice n^* observations to train the prior into a better posterior.
- Predictive density score (PS). Decompose $p(y_1, ..., y_n)$ as

$$\underbrace{p(y_1)p(y_2|y_1)\cdots}_{\text{training}} \underbrace{p(y_{n^*+1}|y_1, ..., y_{n^*})\cdots p(y_n|y_1, y_2, ..., y_{n-1})}_{\text{test}}$$

- Usually report on log scale: **Log Predictive Score** (**LPS**).
- But which observations to train on (and which to test on)?
- Straightforward for time series.
- Cross-sectional data: **cross-validation**.

And hey! ... let's be careful out there.

- Be especially careful with Bayesian model comparison when
  - The compared models are
    - very different in structure
    - severly misspecified
    - very complicated (black boxes).
  - The priors for the parameters in the models are
    - not carefully elicited
    - only weakly informative
    - not matched across models.
  - The data
    - has outliers (in all models)
    - has a multivariate response.

Bayesian Learning - Lecture 11

Mattias Villani

Division of Statistics
Department of Computer and Information Science
Linköping University

Overview

- Computing the marginal likelihood
- Bayesian variable selection
- Model averaging

Marginal likelihood in conjugate models

- Computing the marginal likelihood requires integration w.r.t. θ.
- Short cut for conjugate models by rearragement of Bayes' theorem:

$$p(y) = \frac{p(y|\theta)p(\theta)}{p(\theta|y)}$$

- Bernoulli model example

$$p(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$
$$p(y|\theta) = \theta^{s}(1-\theta)^{f}$$
$$p(\theta|y) = \frac{1}{B(\alpha+s,\beta+f)}\theta^{\alpha+s-1}(1-\theta)^{\beta+f-1}$$

- Marginal likelihood

$$p(y) = \frac{\theta^{s}(1-\theta)^{f}\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\frac{1}{B(\alpha+s,\beta+f)}\theta^{\alpha+s-1}(1-\theta)^{\beta+f-1}} = \frac{B(\alpha+s,\beta+f)}{B(\alpha,\beta)}$$

Computing the marginal likelihood

- Usually difficult to evaluate the integral

$$p(\mathbf{y}) = \int p(\mathbf{y}|\theta)p(\theta)d\theta = E_{p(\theta)}[p(\mathbf{y}|\theta)].$$

- Draw from the prior $\theta^{(1)}, ..., \theta^{(N)}$ and use the Monte Carlo estimate

$$\hat{p}(\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N} p(\mathbf{y}|\theta^{(i)}).$$

  Unstable if the posterior is somewhat different from the prior.
- **Importance sampling**. Let $\theta^{(1)}, ..., \theta^{(N)}$ be iid draws from $g(\theta)$.

$$\int p(\mathbf{y}|\theta)p(\theta)d\theta = \int \frac{p(\mathbf{y}|\theta)p(\theta)}{g(\theta)}g(\theta)d\theta \approx N^{-1}\sum_{i=1}^{N} \frac{p(\mathbf{y}|\theta^{(i)})p(\theta^{(i)})}{g(\theta^{(i)})}$$

- **Modified Harmonic mean**: $g(\theta) = N(\tilde{\theta}, \tilde{\Sigma}) \cdot I_c(\theta)$, where $\tilde{\theta}$ and $\tilde{\Sigma}$ is the posterior mean and covariance matrix estimated from an MCMC chain, and $I_c(\theta) = 1$ if $(\theta - \tilde{\theta})'\tilde{\Sigma}^{-1}(\theta - \tilde{\theta}) \leq c$.

Computing the marginal likelihood, cont.

- Rearrangement of Bayes' theorem: $p(\mathbf{y}) = p(\mathbf{y}|\theta)p(\theta)/p(\theta|\mathbf{y})$.
- We must know the posterior, **including** the normalization constant.
- But we only need to know $p(\theta|\mathbf{y})$ in a single point θ_0.
- **Kernel density estimator** to approximate $p(\theta_0|\mathbf{y})$. Unstable.
- Chib (1995, JASA) provide better solutions for **Gibbs sampling**.
- Chib-Jeliazkov (2001, JASA) generalizes to **MH algorithm** (good for IndepMH, terrible for RWM).
- **Reversible Jump MCMC** (RJMCMC) for model inference.
  - MCMC methods that moves in model space.
  - Proportion of iterations spent in model k estimates $\Pr(M_k|\mathbf{y})$.
  - Usually hard to find efficient proposals. Sloooow convergence.
- **Bayesian nonparametrics** (e.g. Dirichlet process priors).

Laplace approximation

- Taylor approximation of the log likelihood

$$\ln p(\mathbf{y}|\theta) \approx \ln p(\mathbf{y}|\hat{\theta}) - \frac{1}{2} J_{\hat{\theta},\mathbf{y}} (\theta - \hat{\theta})^2,$$

so

$$\begin{aligned} p(\mathbf{y}|\theta)p(\theta) &\approx p(\mathbf{y}|\hat{\theta}) \exp\left[-\frac{1}{2} J_{\hat{\theta},\mathbf{y}} (\theta - \hat{\theta})^2\right] p(\hat{\theta}) \\ &= p(\mathbf{y}|\hat{\theta}) p(\hat{\theta}) (2\pi)^{p/2} \left|J_{\hat{\theta},\mathbf{y}}^{-1}\right|^{1/2} \\ &= \times \underbrace{(2\pi)^{-p/2} \left|J_{\hat{\theta},\mathbf{y}}^{-1}\right|^{-1/2} \exp\left[-\frac{1}{2} J_{\hat{\theta},\mathbf{y}} (\theta - \hat{\theta})^2\right]}_{\text{multivariate normal density}} \end{aligned}$$

- **The Laplace approximation**:

$$\ln \hat{p}(\mathbf{y}) = \ln p(\mathbf{y}|\hat{\theta}) + \ln p(\hat{\theta}) + \frac{1}{2} \ln \left|J_{\hat{\theta},\mathbf{y}}^{-1}\right| + \frac{p}{2} \ln(2\pi),$$

where p is the number of unrestricted parameters in the model.

BIC

- **The Laplace approximation**:

$$\ln \hat{p}(\mathbf{y}) = \ln p(\mathbf{y}|\hat{\theta}) + \ln p(\hat{\theta}) + \frac{1}{2} \ln \left| J^{-1}_{\hat{\theta},\mathbf{y}} \right| + \frac{p}{2} \ln(2\pi).$$

- Note that $\hat{\theta}$ and $J_{\hat{\theta},\mathbf{y}}$ can be obtained with **numerical optimization**.

- The **BIC approximation** is a large sample (large n) approximation obtained when $J_{\hat{\theta},\mathbf{y}}$ behaves like $n \cdot I_p$ in large samples and the small term $+\frac{p}{2} \ln(2\pi)$ is ignored

$$\ln \hat{p}(\mathbf{y}) = \ln p(\mathbf{y}|\hat{\theta}) + \ln p(\hat{\theta}) - \frac{p}{2} \ln n.$$

Bayesian variable selection

- Linear regression:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \varepsilon.$$

- Which variables have **non-zero** coefficient? Example of hypotheses:

$$\begin{aligned} H_0 &: \beta_0 = \beta_1 = ... = \beta_p = 0 \\ H_1 &: \beta_1 = 0 \\ H_2 &: \beta_1 = \beta_2 = 0 \end{aligned}$$

- Introduce **variable selection indicators** $\mathcal{I} = (I_1, ..., I_p)$.
- Example: $\mathcal{I} = (1, 1, 0)$ means that $\beta_1 \neq 0$ and $\beta_2 \neq 0$, but $\beta_3 = 0$, so x_3 drops out of the model.

Bayesian variable selection, cont.

- Model inference, just crank the Bayesian machine:

$$p(\mathcal{I}|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\mathbf{X}, \mathcal{I}) \cdot p(\mathcal{I})$$

- The prior $p(\mathcal{I})$ is typically taken to be $I_1, ..., I_p|\theta \overset{iid}{\sim} Bernoulli(\theta)$.
- θ is the **prior inclusion probability**.
- Challenge: Computing the **marginal likelihood** for each model ($\mathcal{I}$)

$$p(\mathbf{y}|\mathbf{X}, \mathcal{I}) = \int p(\mathbf{y}|\mathbf{X}, \mathcal{I}, \beta) p(\beta|\mathbf{X}, \mathcal{I}) d\beta$$

Bayesian variable selection, cont.

- Let $\beta_{\mathcal{I}}$ denote the **non-zero** coefficients under $\mathcal{I}$.
- Prior:

$$\beta_{\mathcal{I}}|\sigma^2 \sim N\left(0, \sigma^2\Omega_{\mathcal{I},0}^{-1}\right)$$
$$\sigma^2 \sim Inv-\chi^2\left(\nu_0, \sigma_0^2\right)$$

- **Marginal likelihood**

$$p(\mathbf{y}|\mathbf{X}, \mathcal{I}) \propto \left|\mathbf{X}_{\mathcal{I}}'\mathbf{X}_{\mathcal{I}} + \Omega_{\mathcal{I},0}^{-1}\right|^{-1/2} |\Omega_{\mathcal{I},0}|^{1/2} \left(\nu_0\sigma_0^2 + RSS_{\mathcal{I}}\right)^{-(\nu_0+n-1)/2}$$

where $\mathbf{X}_{\mathcal{I}}$ is the covariate matrix for the subset selected by $\mathcal{I}$.

- $RSS_{\mathcal{I}}$ is (almost) the residual sum of squares under model implied by $\mathcal{I}$

$$RSS_{\mathcal{I}} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}_{\mathcal{I}}\left(\mathbf{X}_{\mathcal{I}}'\mathbf{X}_{\mathcal{I}} + \Omega_{\mathcal{I},0}\right)^{-1}\mathbf{X}_{\mathcal{I}}'\mathbf{y}$$

Bayesian variable selection via Gibbs sampling

- But there are 2^p model combinations to go through! Ouch!
- ... but most will have essentially zero posterior probability. Phew!
- **Simulate** from the joint posterior distribution:

$$p(\beta, \sigma^2, \mathcal{I}|\mathbf{y},\mathbf{X}) = p(\beta, \sigma^2|\mathcal{I}, \mathbf{y},\mathbf{X})p(\mathcal{I}|\mathbf{y}, \mathbf{X}).$$

- Simulate from $p(\mathcal{I}|\mathbf{y}, \mathbf{X})$ using **Gibbs sampling**:
  - Draw $I_1|\mathcal{I}_{-1}, \mathbf{y}, \mathbf{X}$
  - Draw $I_2|\mathcal{I}_{-2}, \mathbf{y}, \mathbf{X}$
  - ...
  - Draw $I_p|\mathcal{I}_{-p}, \mathbf{y}, \mathbf{X}$
- Only need to compute $Pr(I_i = 0|\mathcal{I}_{-i}, \mathbf{y}, \mathbf{X})$ and $Pr(I_i = 1|\mathcal{I}_{-i}, \mathbf{y}, \mathbf{X})$.
- Automatic model averaging, all in one simulation run.
- If needed, simulate from $p(\beta, \sigma^2|\mathcal{I}, \mathbf{y},\mathbf{X})$ for each draw of $\mathcal{I}$.

Pseudo code for Bayesian variable selection

0 Initialize $\mathcal{I}^{(0)} = (I_1^{(0)}, I_2^{(0)} ..., I_p^{(0)})$

1 Simulate σ^2 and β from [Note: $\nu_n, \sigma_n^2, \mu_n, \Omega_n$ all depend on $\mathcal{I}^{(0)}$]

- $\sigma^2 | \mathcal{I}^{(0)}, \mathbf{y}, \mathbf{X} \sim Inv - \chi^2 \left(\nu_n, \sigma_n^2 \right)$
- $\beta | \sigma^2, \mathcal{I}^{(0)}, \mathbf{y}, \mathbf{X} \sim N \left[\mu_n, \sigma^2 \Omega_n^{-1} \right]$

2.1 Simulate $I_1 | \mathcal{I}_{-1}, \mathbf{y}, \mathbf{X}$ by [define $\mathcal{I}_{prop}^{(0)} = (1 - I_1^{(0)}, I_2^{(0)} ..., I_p^{(0)})$]

- compute marginal likelihoods: $p(\mathbf{y} | \mathbf{X}, \mathcal{I}^{(0)})$ and $p(\mathbf{y} | \mathbf{X}, \mathcal{I}_{prop}^{(0)})$
- Simulate $I_1^{(1)} \sim Bernoulli(\kappa)$ where

$$\kappa = \frac{p(\mathbf{y} | \mathbf{X}, \mathcal{I}^{(0)}) \cdot p(\mathcal{I}^{(0)})}{p(\mathbf{y} | \mathbf{X}, \mathcal{I}^{(0)}) \cdot p(\mathcal{I}^{(0)}) + p(\mathbf{y} | \mathbf{X}, \mathcal{I}_{prop}^{(0)}) \cdot p(\mathcal{I}_{prop}^{(0)})}$$

2.2 Simulate $I_2 | \mathcal{I}_{-2}, \mathbf{y}, \mathbf{X}$ as in Step 2.1, but $\mathcal{I}^{(0)} = (I_1^{(1)}, I_2^{(0)}, ..., I_p^{(0)})$

⋮

2.p Simulate $I_p | \mathcal{I}_{-p}, \mathbf{y}, \mathbf{X}$ as in Step 2.1, but $\mathcal{I}^{(0)} = (I_1^{(1)}, I_2^{(1)}, ..., I_p^{(0)})$

3 Repeat Steps 1-2 many times.

Simple general Bayesian variable selection

- The previous algorithm only works when we can integrate out all the model parameters to obtain

$$p(\mathcal{I}|\mathbf{y},\mathbf{X}) = \int p(\beta,\sigma^2,\mathcal{I}|\mathbf{y},\mathbf{X})d\beta d\sigma$$

- MH - **propose** β and $\mathcal{I}$ jointly from the proposal distribution

$$q(\beta_p|\beta_c,\mathcal{I}_p)q(\mathcal{I}_p|\mathcal{I}_c)$$

- Main difficulty: how to propose the non-zero elements in β_p?
- Simple approach:
  - Approximate posterior with all variables in the model: $\beta|\mathbf{y},\mathbf{X} \overset{approx}{\sim} N\left[\hat{\beta}, J_{\mathbf{y}}^{-1}(\hat{\beta})\right]$
  - Propose β_p from $N\left[\hat{\beta}, J_{\mathbf{y}}^{-1}(\hat{\beta})\right]$, conditional on the zero restrictions implied by $\mathcal{I}_p$. Formulas are available.

VARIABLE SELECTION IN MORE COMPLEX MODELS

Posterior summary of the one-component split-t model.[a]

Parameters	Mean	Stdev	Post.Incl.
Location μ			
Const	0.084	0.019	–
Scale ϕ			
Const	0.402	0.035	–
LastDay	−0.190	0.120	0.036
LastWeek	**−0.738**	**0.193**	**0.985**
LastMonth	**−0.444**	**0.086**	**0.999**
CloseAbs95	0.194	0.233	0.035
CloseSqr95	0.107	0.226	0.023
MaxMin95	**1.124**	**0.086**	**1.000**
CloseAbs80	0.097	0.153	0.013
CloseSqr80	0.143	0.143	0.021
MaxMin80	−0.022	0.200	0.017
Degrees of freedom v			
Const	2.482	0.238	–
LastDay	0.504	0.997	0.112
LastWeek	**−2.158**	**0.926**	**0.638**
LastMonth	0.307	0.833	0.089
CloseAbs95	0.718	1.437	0.229
CloseSqr95	1.350	1.280	0.279
MaxMin95	1.130	1.488	0.222
CloseAbs80	0.035	1.205	0.101
CloseSqr80	0.363	1.211	0.112
MaxMin80	−1.672	1.172	0.254
Skewness λ			
Const	−0.104	0.033	–
LastDay	−0.159	0.140	0.027
LastWeek	−0.341	0.170	0.135
LastMonth	−0.076	0.112	0.016
CloseAbs95	−0.021	0.096	0.008
CloseSqr95	−0.003	0.108	0.006
MaxMin95	0.016	0.075	0.008
CloseAbs80	0.060	0.115	0.009
CloseSqr80	0.059	0.111	0.010
MaxMin80	0.093	0.096	0.013

Model averaging

- Let γ be a quanitity with an interpretation which stays the same across the two models.
- Example: Prediction $\gamma = (y_{T+1}, ..., y_{T+h})'$.
- The marginal posterior distribution of γ reads

$$p(\gamma|\mathbf{y}) = p(M_1|\mathbf{y})p_1(\gamma|\mathbf{y}) + p(M_2|\mathbf{y})p_2(\gamma|\mathbf{y}),$$

  where $p_k(\gamma|\mathbf{y})$ is the marginal posterior of γ conditional on model k.
- Predictive distribution includes three sources of uncertainty:
  - **Future errors**/disturbances (e.g. the ε's in a regression)
  - **Parameter uncertainty** (the predictive distribution has the parameters integrated out by their posteriors)
  - **Model uncertainty** (by model averaging)

Bayesian Learning - Lecture 12

Mattias Villani

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Overview

- Model evaluation - Posterior predictive analysis
- Course summary and discussion

Models - why?

- We now know how to **compare** models.
- But how do we know if any given model is 'any good'?
- George Box: '**All models are false, but some are useful**'.

What is your model for really?

- **Prediction**.
  - Interpretation not a concern
  - Black-box approach may be ok.
  - Extrapolation?
  - Model averaging may be a good idea.
- Abstraction to **aid in thinking** about a phenomena.
  - Prediction accuracy of less concern.
  - Model averaging may be a bad idea.
- Model as a **compact description of a complex phenomena**.
  - Computational cost of model evaluation may be a concern.
  - Online/real-time analysis.

Posterior predictive analysis

- If $p(y|\theta)$ is a 'good' model, then the data actually observed should not differ 'too much' from simulated data from $p(y|\theta)$.

- Bayesian: simulate data from the **posterior predictive distribution**:

$$p(y^{rep}|y) = \int p(y^{rep}|\theta)p(\theta|y)d\theta.$$

- Difficult to compare y and y^{rep} because of dimensionality.
- Solution: compare **low-dimensional statistic** $T(y, \theta)$ to $T(y^{rep}, \theta)$.
- Evaluates the full probability model consisting of both the likelihood *and* prior distribution.

Posterior predictive analysis, cont.

- **Algorithm** for simulating from the posterior predictive density $p[T(y^{rep})|y]$:

1. Draw a $\theta^{(1)}$ from the posterior $p(\theta|y)$.
2. Simulate a data-replicate $y^{(1)}$ from $p(y^{rep}|\theta^{(1)})$.
3. Compute $T(y^{(1)})$.
4. Repeat steps 1-3 a large number of times to obtain a sample from $T(y^{rep})$.

- We may now compare the observed statistic $T(y)$ with the distribution of $T(y^{rep})$.
- **Posterior predictive p-value**: $\Pr[T(y^{rep}) \geq T(y)]$
- Informal **graphical analysis**.

Posterior predictive analysis - Examples

- Ex. 1. Model: $y_1, ..., y_n \overset{iid}{\sim} N(\mu, \sigma^2)$. $T(y) = \max_i |y_i|$.
- Ex. 2. Assumption of no reciprocity in networks.
  $y_{ij}|\theta \overset{iid}{\sim} Bernoulli(\theta)$. $T(y) =$proportion of reciprocated node pairs.
- Ex. 3. ARIMA-process. $T(y)$ may be the autocorrelation function.
- Ex. 4. Poisson regression. $T(y)$ frequency distribution of the response counts. Proportions of zero counts.

Posterior predictive analysis - Normal model, max statistic