Bayesian Statistics

Fabrizio Ruggeri

Istituto di Matematica Applicata e Tecnologie Informatiche
Consiglio Nazionale delle Ricerche
Via Alfonso Corti 12, I-20133, Milano, Italy, European Union
fabrizio@mi.imati.cnr.it
www.mi.imati.cnr.it/fabrizio/

MONTE CARLO SIMULATION

- Consider $X \sim \mathcal{E}(\lambda)$ and a prior $\lambda \sim G(\alpha, \beta)$
- How can we compute the prior predictive density at a value x?
  i.e. $f(x) = \int_0^\infty f(x|\lambda)\pi(\lambda)d\lambda$?
- $\Rightarrow f(x) = \displaystyle\int_0^\infty \lambda e^{-\lambda x} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta} d\lambda = \alpha \frac{\beta^\alpha}{(\beta + x)^{\alpha+1}}$
- But what about choosing a Weibull prior $\lambda \sim \mathcal{W}(\alpha, \beta)$?
- Weibull density: $\pi(\lambda|\alpha, \beta) = \dfrac{\beta}{\alpha} \left(\dfrac{\lambda}{\alpha}\right)^{\beta-1} e^{(-\lambda/\alpha)^\beta},\ \alpha, \beta > 0$
- In this case we are unable to compute (at least easily) the integral so that we need to resort to a Monte Carlo simulation method
- Here I will not discuss about the (rate of) convergence and errors committed in evaluating the integral

MONTE CARLO SIMULATION

- How to approximate the prior predictive density for $x > 0$?
- Choose a grid of (equally spaced or not) points $x_j,\ j = 1, \ldots, M$
- Draw a ("large") sample $\lambda_1, \lambda_2, \ldots, \lambda_N$ from the Weibull prior
- For each x_j compute $f(x_j) = \int_0^\infty f(x_j|\lambda)\pi(\lambda)d\lambda \approx \sum_{i=1}^{N} \frac{f(x_j|\lambda_i)}{N} = \tilde{f}(x_j)$
- $\Rightarrow$ Approximation of $f(x)$, based on $\tilde{f}(x_1), \ldots, \tilde{f}(x_M)$, using splines, etc.
- Many sources of uncertainty: N, M, grid, fitted function
- We omit a discussion about those uncertainties
- MC for posterior expectations $\int h(\lambda)\pi(\lambda|y)d\lambda$, e.g. posterior mean for $h(\lambda) = \lambda$

IMPORTANCE SAMPLING

- Bayes Theorem: $\pi(\lambda|y) = \dfrac{f(y|\lambda)\pi(\lambda)}{\int f(y|\theta)\pi(\theta)d\theta}$
- $f(y|\lambda)$ and $\pi(\lambda)$ are known (at least in our course) but $f(y) = \int f(y|\theta)\pi(\theta)d\theta$ might not be
- The inability of computing the normalising constant $f(y)$ has been a huge problem before the MCMC era started (still it is a problem!)
- $\Rightarrow$ We know only $q(\lambda|y) = f(y|\lambda)\pi(\lambda)$ and we are neither able to compute the posterior in closed form nor to draw a sample from it
- We are interested in computing $E\left[h(\lambda)|y\right] = \displaystyle\int h(\lambda)\pi(\lambda|y)d\lambda$
- Choose an "adequate" *proposal density* $g(\lambda)$

  (An optimal, but not always possible, choice for $g(\lambda)$ could be such that $\dfrac{f(y|\lambda)\pi(\lambda)}{g(\lambda)}$ is roughly constant)

IMPORTANCE SAMPLING

- We are interested in

$$\begin{aligned} E\left[h(\lambda)|y\right] &= \int h(\lambda)\pi(\lambda|y)d\lambda \\ &= \frac{\int h(\lambda)f(y|\lambda)\pi(\lambda)d\lambda}{\int f(y|\lambda)\pi(\lambda)d\lambda} \\ &= \frac{\int h(\lambda)q(\lambda|y)d\lambda}{\int q(\lambda|y)d\lambda} \\ &= \frac{\int h(\lambda)q(\lambda|y)/g(\lambda)\cdot g(\lambda)d\lambda}{\int q(\lambda|y)/g(\lambda)\cdot g(\lambda)d\lambda} \end{aligned}$$

- We draw a sample $\lambda_1, \ldots, \lambda_N$ from $g(\lambda)$
- $\Rightarrow E\left[h(\lambda)|y\right] \approx \frac{\frac{1}{N}\sum_{i=1}^{N} h(\lambda_i)w(\lambda_i)}{\frac{1}{N}\sum_{i=1}^{N} w(\lambda_i)}$
- $w(\lambda_i) = \frac{q(\lambda_i|y)}{g(\lambda_i)}$: *importance weights*

IMPORTANCE SAMPLING

- In general, it is suggested to use the same random draws for both numerator and denominator

- Importance sampling is not a useful method if the importance weights vary substantially

- The worst possible scenario occurs when the importance weights are small with high probability but with a low probability are huge, which happens, e.g., if q has wide tails compared to g, as a function of λ

- In general, without some form of mathematical analysis of the exact and approximate densities, there is always the realistic possibility of missing some extremely large but rare importance weights

MARKOV CHAIN MONTE CARLO

- A Markov chain is a sequence of r.v.'s $\{X_n\}$ such that the distribution of any X_n depends on the past only through X_{n-1}
- $\mathbb{P}(X_n|X_{n-1}, X_{n-2}, \ldots, X_1) = \mathbb{P}(X_n|X_{n-1}),\ \forall n$
- MCMC used to draw samples "converging" towards posterior $\pi(\theta|\underline{X})$
- Name MCMC due to simulations based on transition distributions $p(\theta^i|\theta^{i-1})$
- Many works dealt with the theory justifying MCMC, ensuring theoretical convergence to the posterior distribution: we will not discuss them except for mentioning that the posterior distribution is the stationary distribution of an appropriate Markov chain
- Many works addressed the issue of empirically guaranteeing the practical convergence: we will discuss them briefly
- Many MCMC methods: here only Gibbs sampling and Metropolis-Hastings algorithm
- No mention of other simulation methods, like Variational Bayes and Approximate Bayesian Computation

MARKOV CHAIN MONTE CARLO

- Sample $\underline{X}$ and parameter $\theta = (\theta_1, \ldots, \theta_n)$
- Notation $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$, for $i = 1, \ldots, n$
- *Gibbs sampling* is used when $\pi(\theta|\underline{X})$ is not available but all $\pi(\theta_i|\theta_{-i}, \underline{X})$, $i = 1, \ldots, n$, are
- Example seen earlier: $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu \in \Re$ and $\sigma^2 > 0$ unknown
  - Prior $\pi(\mu, \sigma^2) = \pi(\mu|\sigma^2)\pi(\sigma^2)$
  - $\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \tau^2\sigma^2)$
  - $\sigma^2 \sim \mathcal{IG}(\alpha, \beta)$ Inverse gamma
  - $\mu|\sigma^2, \underline{X} \sim \mathcal{N}\left(\frac{\sum_{i=1}^n X_i + \mu_0/\tau^2}{n + 1/\tau^2}, \frac{\sigma^2}{n + 1/\tau^2}\right)$
  - $\sigma^2|\mu, \underline{X} \sim \mathcal{IG}(\alpha + (n+1)/2, \beta + \sum_{i=1}^n (X_i - \mu)^2/2 + (\mu - \mu_0)^2/(2\tau^2))$

MARKOV CHAIN MONTE CARLO

- In words, Gibbs sampling consists of a "sufficient" number of steps in which each parameter θ_i is sequentially drawn from its *full conditional distribution* $\pi(\theta_i|\theta_{-i}, \underline{X})$, where θ_{-i} contains the values of $\theta_1, \ldots, \theta_{i-1}$ generated at the current step and those of $\theta_{i+1}, \ldots, \theta_n$ generated at the previous step

- Algorithm

  1. Set $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_n^{(0)})$ and $j = 0$
  2. Set $j = j + 1$
  3. For $i = 1, \ldots, n$, draw $\theta_i^{(j)}$ from $\pi(\theta_i|\theta_1^{(j)}, \ldots, \theta_{i-1}^{(j)}, \theta_{i+1}^{(j-1)}, \ldots, \theta_n^{(j-1)}, \underline{X})$
  4. If $j < N$ (set a priori) then go back to (2)
  5. $\Rightarrow \theta^{(j)}, j = 1, \ldots, N$, used to get a sample from the posterior distribution

- Some $\theta^{(j)}$'s might be discarded, e.g. initial ones (more later)

MARKOV CHAIN MONTE CARLO *

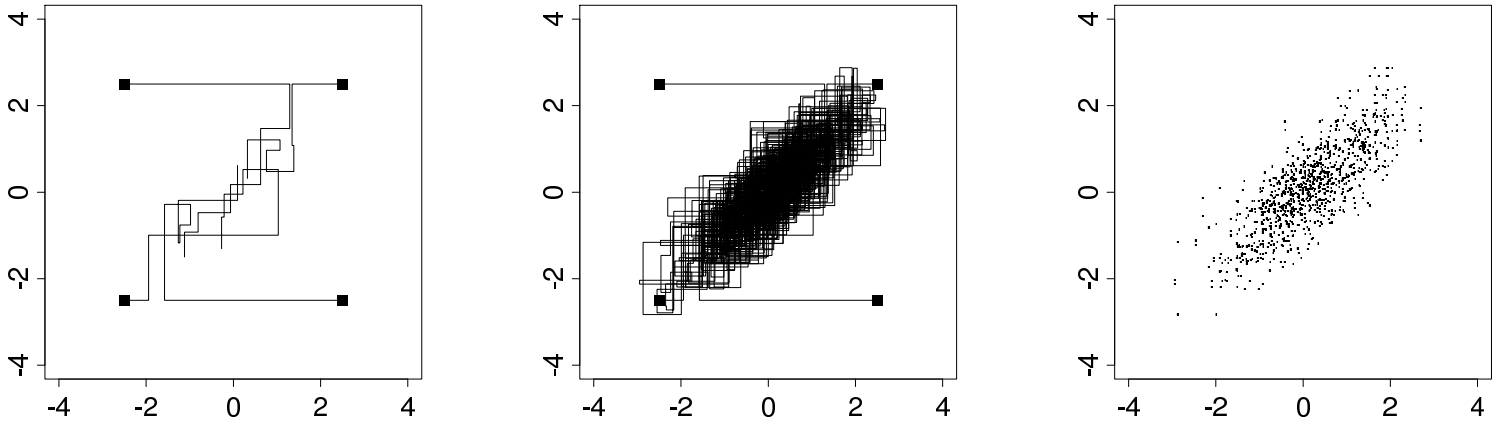
- Consider a single observation (y_1, y_2) from a bivariate Gaussian with unknown mean $\theta = (\theta_1, \theta_2)$ and known covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$
- Uniform prior on θ: $\pi(\theta) \propto c,\ c > 0$
- $\Rightarrow$ Posterior $\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} | y \sim \mathcal{N}\left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$
- Although it is simple to draw directly from the joint posterior distribution of (θ_1, θ_2), for the purpose of exposition we demonstrate the Gibbs sampler here
- Simulate (alternating) from known full conditional distributions
  - $\theta_1 | \theta_2, y \sim \mathcal{N}(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2)$
  - $\theta_2 | \theta_1, y \sim \mathcal{N}(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2)$

*Example from Gelman et al., *Bayesian Data Analysis, Third Edition*, freely available at `http://www.stat.columbia.edu/~gelman/book/BDA3.pdf`

MARKOV CHAIN MONTE CARLO

- Take $\rho = 0.8$ and $(y_1, y_2) = (0, 0)$
- $\Rightarrow$ Posterior $\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} | y \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} \right)$
- Four independent sequences starting at $(\pm 2.5, \pm 2.5)$ to remove dependence on initial point
- Sequences run until convergence to the posterior is achieved (more later on checking for convergence)
- By convergence we mean that the drawn samples are from an approximating distribution close to the posterior one (our target)
- Use of just part of the data, removing the initial ones since they might not be in the approximating distribution (this operation is called *burn-in*)
- Sometimes one searches to reduce correlation between samples so that just 1 every m is kept

MARKOV CHAIN MONTE CARLO



- Left: First 10 iterations for four independent sequences starting at $(\pm 2.5, \pm 2.5)$
- Center: After 500 iterations, the sequences have reached approximate convergence
- Right: The points from the second halves of the sequences, discarding the first 250 samples values of each sequence (burn-in)
- Often just one sequence is drawn but for longer time
- Note how the samples are around $(0, 0)$ and showing a strong positive correlation, as expected knowing the exact joint posterior

MARKOV CHAIN MONTE CARLO

- In Gibbs sampling we assumed that it was always possible to get the full conditional $\pi(\theta_i|\theta_{-i}, \underline{X})$ for all i's but is not always the case
- Sometimes we know only $\pi(\theta_i|\theta_{-i}, \underline{X}) \propto q(\theta_i|\theta_{-i}, \underline{X})$ where $q(\cdot)$ is not a density function
- It is a similar case to what seen before when we considered $q(\lambda|y) = f(y|\lambda)\pi(\lambda)$ known, unlike its integral w.r.t. λ which is the normalising constant in Bayes Theorem
- In this case we will use *Metropolis-Hastings steps within Gibbs*
- The Metropolis-Hastings algorithm allows to draw a value θ_i^* from a proposal density $p(\theta_i)$ and accept either it or $\theta_i^{(j-1)}$ as $\theta_i^{(j)}$ with probabilities depending on both p and q
- The proposal density for θ_i^* could be chosen, e.g., either as the same for each iteration or as dependent on the previous $\theta_i^{(j-1)}$

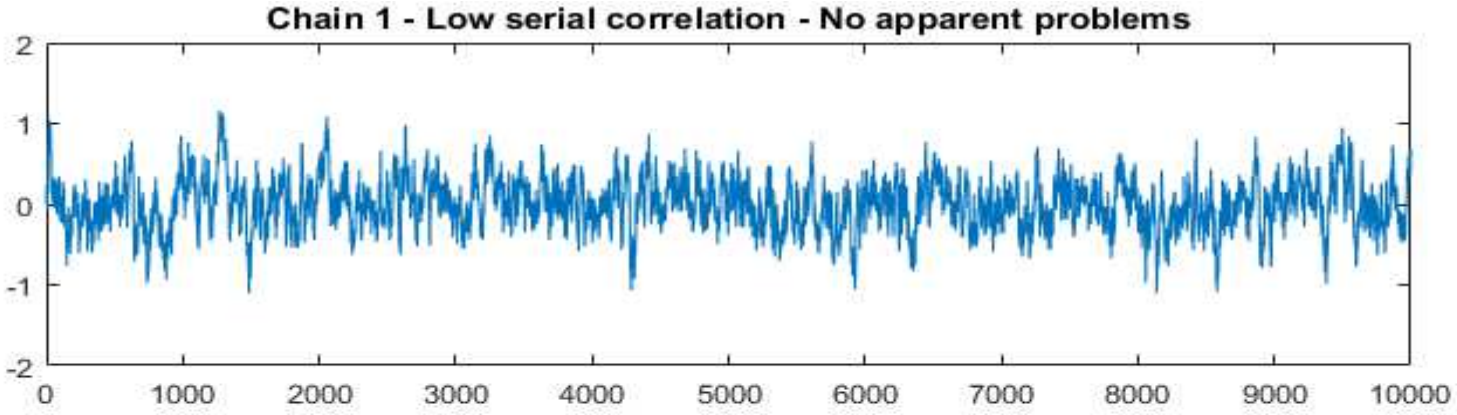
MARKOV CHAIN MONTE CARLO

- For simplicity of notation, let us remove the index i and the conditioning on the other parameters and consider just θ
- Suppose that $\pi(\theta|y)$ is known just up to a constant, i.e. $\pi(\theta|y) \propto q(\theta|y)$, or $\pi(\theta|y) = \dfrac{q(\theta|y)}{\int q(\theta|y)d\theta}$
- We start with an initial value $\theta^{(0)}$ s.t. $\pi(\theta^{(0)}|y) > 0$
- For each iteration $j = 1, \ldots, N$, generate a proposal θ^* from a proposal density $p_j(\theta|\theta^{(j-1)})$
- Compute the ratio $r = \dfrac{q(\theta^*|y)/p(\theta^*|\theta^{(j-1)})}{q(\theta^{(j-1)}|y)/p(\theta^{(j-1)}|\theta^*)}$
- Set $\theta^{(j)} = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta^{(j-1)} & \text{otherwise} \end{cases}$

MARKOV CHAIN MONTE CARLO

- We already saw that running more than one simulation at the time and removing the initial values should reduce the dependence on the initial values
- The proposal distributions are often chosen depending on the value at the previous iteration, e.g. a Gaussian distribution centered at it, or independently from it, possibly the same at all iterations, e.g. Gaussians with the same mean
- Many tools developed to check convergence of the sequence to the true distribution
- The simplest, graphical, tool to assess convergence is to check if the plot of the sample mean stabilises as the iterations grow (if not, then no convergence)
- Given a sample $\theta^{(S+1)}, \ldots, \theta^{(N)}$, with a burn-in of size S, then estimators of $E(h(\theta)|y)$ are given by $\frac{\sum_{j=S+1}^{N} h(\theta^{(j)})}{N-S}$, like
  - $E(\theta|y) \approx \frac{\sum_{j=S+1}^{N} \theta^{(j)}}{N-S}$
  - $\mathbb{P}(\theta \in A|y) \approx \frac{\#\{\theta^{(j)} \in A\}_{j=S+1}^{N}}{N-S}$

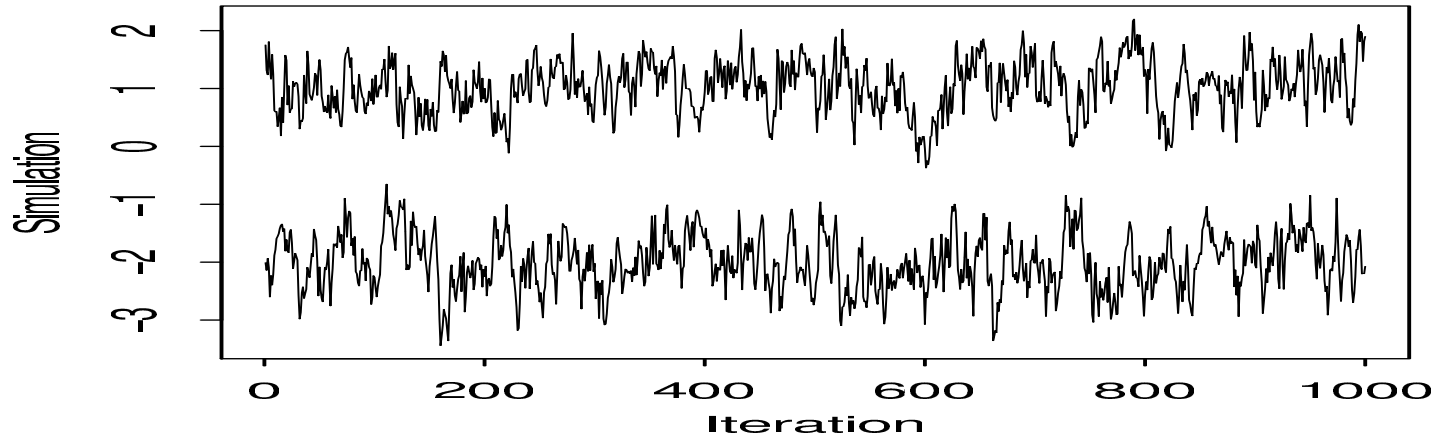
MARKOV CHAIN MONTE CARLO*



- Trace plots are heuristic tools, widely used to check convergence of the MCMC
- They plot the values of each parameter for all the iterations
- They are ”good” when the plot keeps jumping within a set which denotes where the posterior density is concentrated
- The trace plot in the figure is a good one, unlike the next ones

*Plots from `www.statlect.com`

MARKOV CHAIN MONTE CARLO

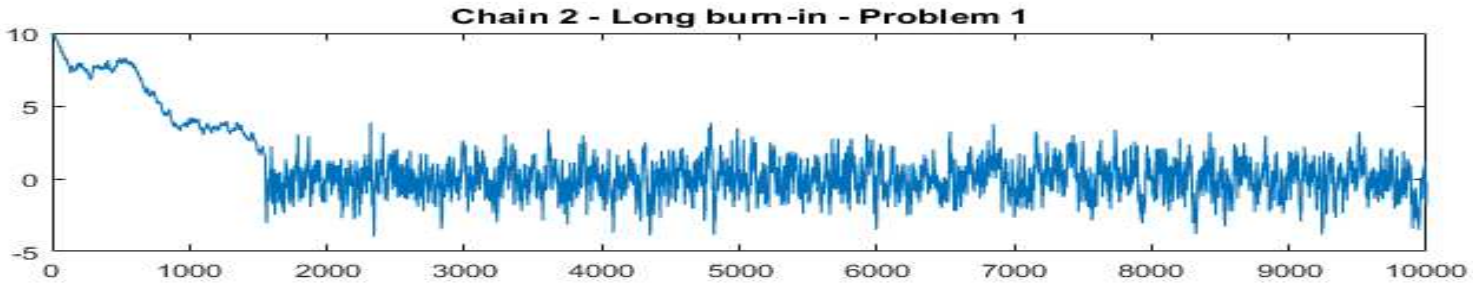


- Here two sequences have been running and both of them are converging but to two different values

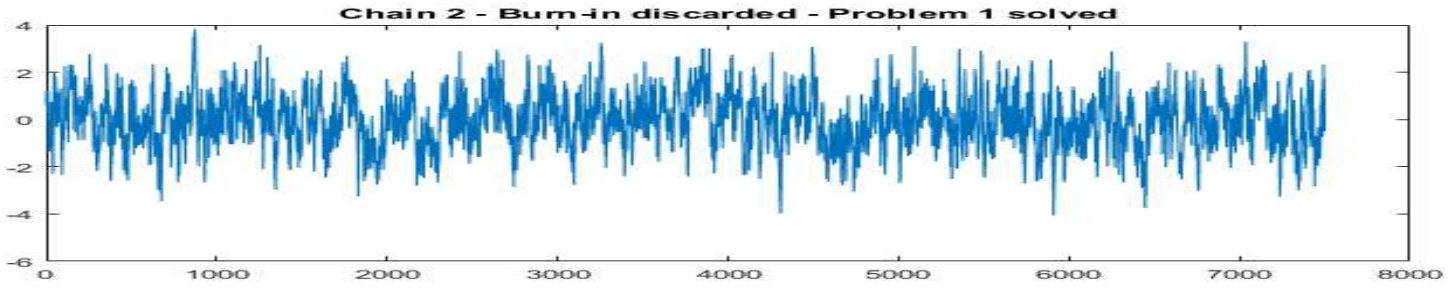
- In general, a plot like this is not desirable since it does not give a clear indication about where the posterior density is, unless the density is bimodal

- In the latter case one would expect the chain to jump from one mode to another

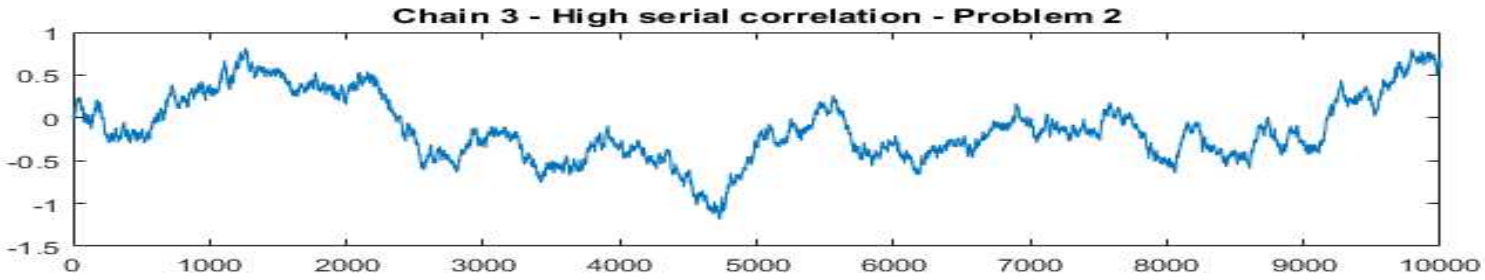
MARKOV CHAIN MONTE CARLO



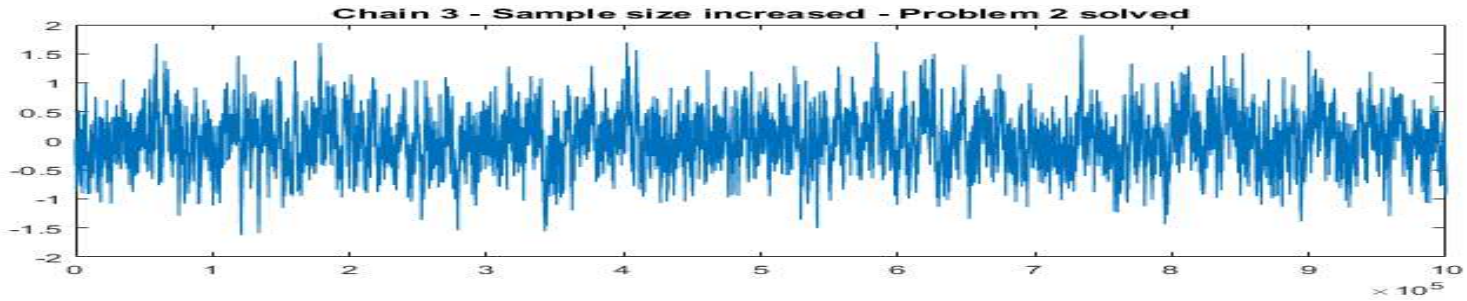
- The first part of the sample looks very different from the remaining part.
- Most likely, the initial distribution and the distributions of the subsequent terms of the chain were very different from the target distribution, but then the chain slowly converged to the target distribution
- The problem can be solved by removing the initial values (burn-in)



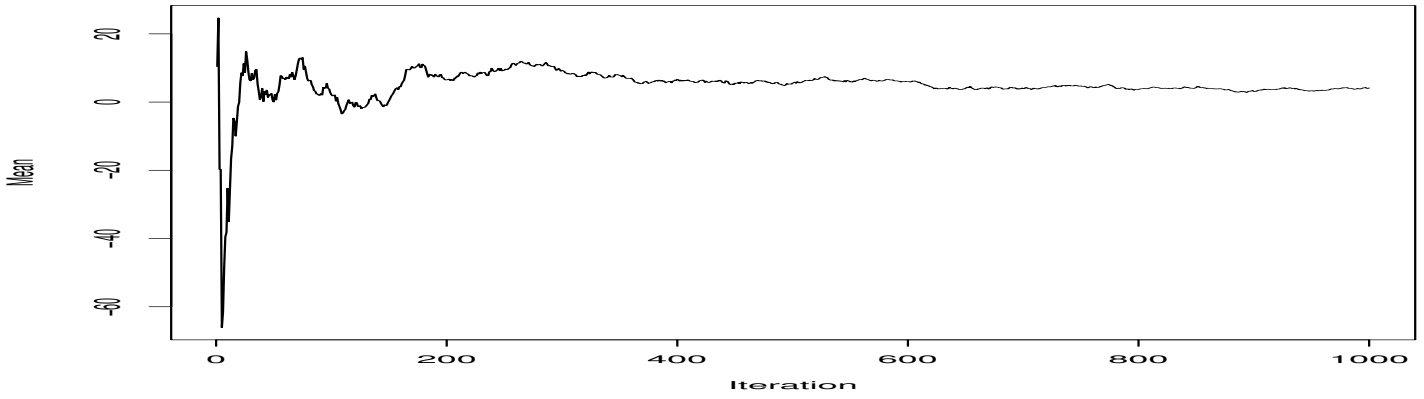
MARKOV CHAIN MONTE CARLO



- A lot of autocorrelation between the draws ($\Rightarrow$ lack of independence)
- Chain very slow in exploring the sample space, explored only few times
- The problem could be due to a small number of iterations $\Rightarrow$ run longer and, possibly, take one draw out of m to avoid large sample size and remove autocorrrelation



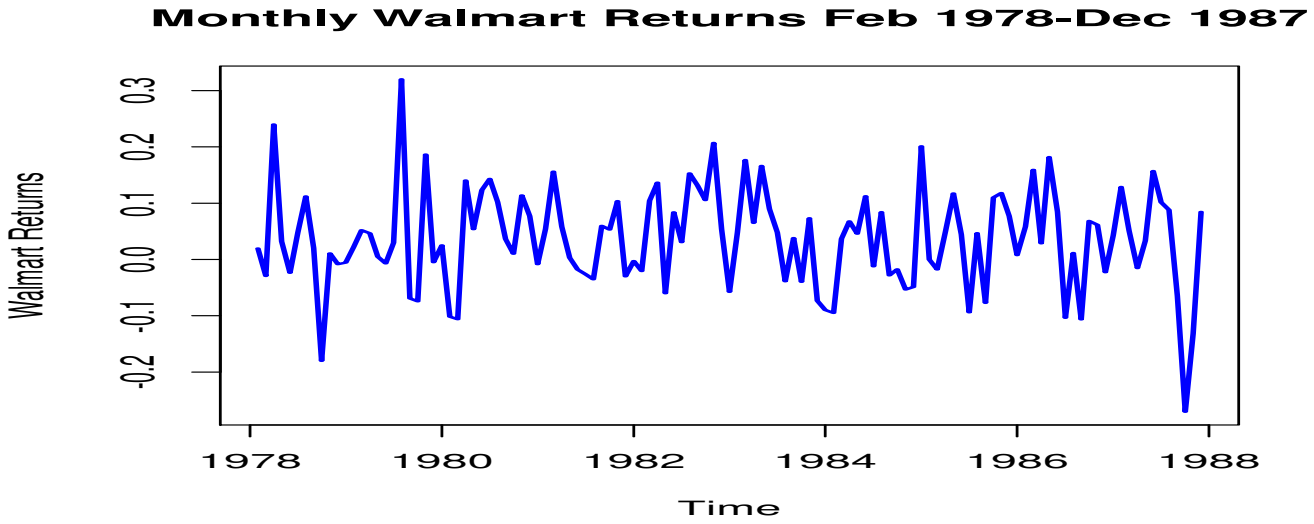
MARKOV CHAIN MONTE CARLO



- We can also show how estimators, like posterior mean, are evolving as the number of iterations increases nd if they are stabiling around a value

- In this case the plot denotes the evolution of the sample mean, estimator of the parameter, as the iterations increase

MARKOV CHAIN MONTE CARLO (MCMC)



- Return: (current month value - past month value)/past month value
- File cina24-walmart.txt
  - Monthly Walmart Returns February 1978 - December 1987
  - Monthly S&P500 Returns February 1978 - December 1987
- I leave it to you to consider the S&P500 data

MARKOV CHAIN MONTE CARLO (MCMC)

- Visual inspection suggests no serial correlation (confirmed by tests)

- Seems to be given by a constant mean with some uncorrelated error

```
# Read a file, from working directory, with labels in the first line
setwd("D:") # Careful: OK if file in drive D: (e.g. USB)
setwd("C:/Users/fabru/Desktop/cina24") # for me
all=read.table("cina24-walmart.txt", header=TRUE)
attach(all) # Call WMART a column of data instead of all$WMART $
head(all) # Shows the first lines in the file
# Define data in WMART as time series object, starting at 2/1978
# frequency of 12 as number of observations per unit of time (year)
wmart=ts(data=WMART, start=c(1978,2),frequency=12)
ts.plot(wmart,col="blue",lwd="2",ylab="Walmart Returns",
main="Monthly Walmart Returns Feb 1978-Dec 1987")
mean(wmart); sd(wmart); 1/(sd(wmart))^2
```

- For S&P500 replace WMART/wmart with SP500/sp500

MARKOV CHAIN MONTE CARLO (MCMC)*

```
a=2;b=5;N=10000 # Try others, e.g. a=0;b=0
muN=rep(0,N+1) # R starts from 1; muN[1] initial value
tauN=rep(0,N+1)
tauN[1]=100 # First element cannot be 0
meanW=mean(WMART); lenW=length(WMART)
library(LaplacesDemon) # Needed for rnormp (Gaussian with precision)
for (i in 1:N) {muN[i+1]=rnormp(1,meanW,lenW*tauN[i]);
tauN[i+1]=rgamma(1,a+lenW/2, b+sum((WMART-muN[i+1])^2)/2)}
mean(muN);mean(WMART);mean(tauN);1/var(WMART)
par(mfrow=c(2,1))
hist(muN);hist(tauN)
plot(density(muN));plot(density(tauN))
meanM=rep(0,N+1);meanT=rep(0,N+1)
for (i in 1:(N+1)) {meanM[i]=mean(muN[1:i]); meanT[i]=mean(tauN[1:i])}
plot(meanM[(N/2):(N+1)],type='l');plot(meanT[(N/2):(N+1)],type='l')
```

*Most R codes from Albert's book

HIERARCHICAL MODELS

- Consider the number of car accidents over 30 years by a driver (M) in Milano and one (R) in Roma
- We can consider two persons, randomly selected or not, or the average of (a subset of) the population in the two cities but then we round up to an integer
- The event is rare and takes only integer values $\Rightarrow$ Poisson distribution
- $X \sim \mathcal{P}(\lambda) \rightarrow \mathbb{P}(X = x) = \frac{\lambda^x}{x!} e^{-\lambda};\ x \in \mathbb{Z}$
- How should we model our data and prior for M and R?
- We should think if the behaviour of the two drivers is the same, completely different or there are similarities
- How do we transform those situations into a statistical model?

HIERARCHICAL MODELS

- n_M and n_R number of accidents for M and R
- λ_M and λ_R parameters for Poisson distribution for n_M and n_R
- **Equal**: If the two drivers are behaving in the same way, we model the data independently but with a common λ, with gamma prior $\mathcal{G}(\alpha, \beta)$
  - $\Rightarrow \pi(\lambda|n_M, n_R) \propto \lambda^{n_M} e^{-\lambda} \cdot \lambda^{n_R} e^{-\lambda} \cdot \lambda^{\alpha-1} e^{-\lambda\beta}$
  - $\Rightarrow \lambda|n_M, n_R \sim \mathcal{G}(\alpha + n_M + n_R, \beta + 2)$
- **Completely different**: If the two drivers are behaving in a completely different way, we model the data not only independently but also with different λ's, and independent gamma priors
  - $n_M \sim \mathcal{P}(\lambda_M)$ and $\lambda_M \sim \mathcal{G}(\alpha_M, \beta_M) \Rightarrow \lambda|n_M \sim \mathcal{G}(\alpha_M + n_M, \beta_M + 1)$
  - $n_R \sim \mathcal{P}(\lambda_R)$ and $\lambda_R \sim \mathcal{G}(\alpha_R, \beta_R) \Rightarrow \lambda|n_R \sim \mathcal{G}(\alpha_R + n_R, \beta_R + 1)$

HIERARCHICAL MODELS

- **Similar**: If the two drivers are behaving in a similar way, we model the data independently, with different λ's, but drawn from the same exponential (for simplicity) prior, dependent on a parameter θ
  - $\Rightarrow \pi(\lambda_M, \lambda_R | n_M, n_R, \theta) \propto \lambda_M^{n_M} e^{-\lambda_M} \cdot \lambda_R^{n_R} e^{-\lambda_R} \cdot \theta e^{-\lambda_M \theta} \cdot \theta e^{-\lambda_R \theta}$
  - $\Rightarrow \lambda_M | n_M, n_R, \theta \sim \mathcal{G}(n_M + 1, \theta + 1)$ and $\lambda_R | n_M, n_R, \theta \sim \mathcal{G}(n_R + 1, \theta + 1)$
- Two independent gamma posteriors for known θ but what about if unknown?
- We could consider a gamma prior $\theta \sim \mathcal{G}(a, b)$
- $\Rightarrow \pi(\lambda_M, \lambda_R, \theta | n_M, n_R) \propto \lambda_M^{n_M} e^{-\lambda_M} \cdot \lambda_R^{n_R} e^{-\lambda_R} \cdot \theta e^{-\lambda_M \theta} \cdot \theta e^{-\lambda_R \theta} \cdot \theta^{a-1} e^{-b\theta}$
- Gibbs sampling:
  - $\lambda_M | \lambda_R, \theta, n_M, n_R \sim \mathcal{G}(\theta + n_M + 1, \theta + 1)$
  - $\lambda_R | \lambda_M, \theta, n_M, n_R \sim \mathcal{G}(\theta + n_R + 1, \theta + 1)$
  - $\theta | \lambda_M, \lambda_R, n_M, n_R \sim \mathcal{G}(a + 2, b + \lambda_M + \lambda_R)$

HIERARCHICAL MODELS

- We have to integrate out θ if we are just interested in the full conditionals of each λ given the other

$$\begin{aligned}
\pi(\lambda_M, \lambda_R | n_M, n_R) &= \int \pi(\lambda_M, \lambda_R, \theta | n_M, n_R) d\theta \\
&\propto \lambda_M^{n_M} e^{-\lambda_M} \lambda_R^{n_R} e^{-\lambda_R} \int \theta^{a+1} e^{-(b+\lambda_M+\lambda_R)\theta} d\theta \\
&\propto \frac{\lambda_M^{n_M} e^{-\lambda_M} \lambda_R^{n_R} e^{-\lambda_R}}{(b+\lambda_M+\lambda_R)^{a+2}}
\end{aligned}$$

- $\Rightarrow$ We can use Gibbs sampling with Metropolis steps within
  - $\pi(\lambda_M | \lambda_R, n_M, n_R) \propto \dfrac{\lambda_M^{n_M} e^{-\lambda_M}}{(b+\lambda_M+\lambda_R)^{a+2}}$
  - $\pi(\lambda_R | \lambda_M, n_M, n_R) \propto \dfrac{\lambda_R^{n_R} e^{-\lambda_R}}{(b+\lambda_M+\lambda_R)^{a+2}}$
- As proposal distributions we could use $\mathcal{G}(n_M+1, 1)$ and $\mathcal{G}(n_R+1, 1)$, respectively

HIERARCHICAL MODELS

- Empirical Bayes is a practical, although not properly Bayesian, alternative to the choice of a prior on θ
- The idea is to find the value of θ maximising the probability of the data and plug it into the formulas
- The critical aspect, from a strict Bayesian viewpoint, is that data are used twice, first to find a value of θ and then computing the posterior distribution: priors should be independent from the data!
- We have to look for $\widehat{\theta} = \arg\max_\theta f(n_M, n_R|\theta)$
- With the same computations as before for θ known, we plug in $\widehat{\theta}$
  $\Rightarrow \lambda_M|n_M, n_R, \widehat{\theta} \sim \mathcal{G}(n_M + 1, \widehat{\theta} + 1)$ and $\lambda_R|n_M, n_R, \widehat{\theta} \sim \mathcal{G}(n_R + 1, \widehat{\theta} + 1)$

HIERARCHICAL MODELS

$$
\begin{aligned}
f(n_M, n_R|\theta) &= \int f(n_M, n_R|\lambda_M, \lambda_R)\pi(\lambda_M, \lambda_R|\theta)d\lambda_M d\lambda_R \\
&\propto \int \lambda_M^{n_M} e^{-\lambda_M} \cdot \lambda_R^{n_R} e^{-\lambda_R} \cdot \theta e^{-\lambda_M \theta} \cdot \theta e^{-\lambda_R \theta} d\lambda_M d\lambda_R \\
&\propto \theta^2 \int \lambda_M^{n_M} e^{-(\theta+1)\lambda_M} d\lambda_M \int \lambda_R^{n_R} e^{-(\theta+1)\lambda_R} d\lambda_R \\
&\propto \theta^2 \frac{\Gamma(n_M+1)}{(\theta+1)^{n_M+1}} \frac{\Gamma(n_R+1)}{(\theta+1)^{n_R+1}} \\
&\propto \frac{\theta^2}{(\theta+1)^{n_M+n_R+2}} \\
&= h(n_M, n_R, \theta)
\end{aligned}
$$

- $\dfrac{\partial \log h(n_M, n_R, \theta)}{\partial \theta} = \dfrac{2}{\theta} - \dfrac{n_M + n_R + 2}{\theta + 1}$
- $\dfrac{\partial \log h(n_M, n_R, \theta)}{\partial \theta} = 0 \Leftrightarrow \widehat{\theta} = \dfrac{2}{n_M + n_R}$

HIERARCHICAL MODELS

- Is $\widehat{\theta} = \frac{2}{n_M + n_R}$ surprising? Not much!
- We are considering an event described by a Poisson distribution with parameter λ
- For $X \sim \mathcal{P}(\lambda)$ we know that $E(X) = \lambda$
- For $\lambda \sim \mathcal{E}(\theta)$ we know that $E(\lambda) = 1/\theta$
- Since we use $\widehat{\theta} = \frac{2}{n_M + n_R}$, we can think of X somehow approximated (with some mathematical imprecision) by $\frac{n_M + n_R}{2}$, which is very reasonable under our assumptions

HIERARCHICAL MODELS

- In Italy every year students in some grades are taking tests on their knowledge about Italian language and Mathematics. The results of the tests could be affected by the school attended by the students so that it is reasonable to assume that the outcome for students of the same school are modelled by the same distribution while there should be a difference between schools.

- The same model could be used for batches of the same item but produced in different factories or survival times of patients in different hospitals

- We suppose that we observe data from n different groups, with n_i, $i = 1, \ldots, n$, elements in each of them

- Therefore the data are Y_{ij_i}, $i = 1, \ldots, n$ and $j_i = 1, \ldots, n_i$, although we will use Y_{ij} for simplicity

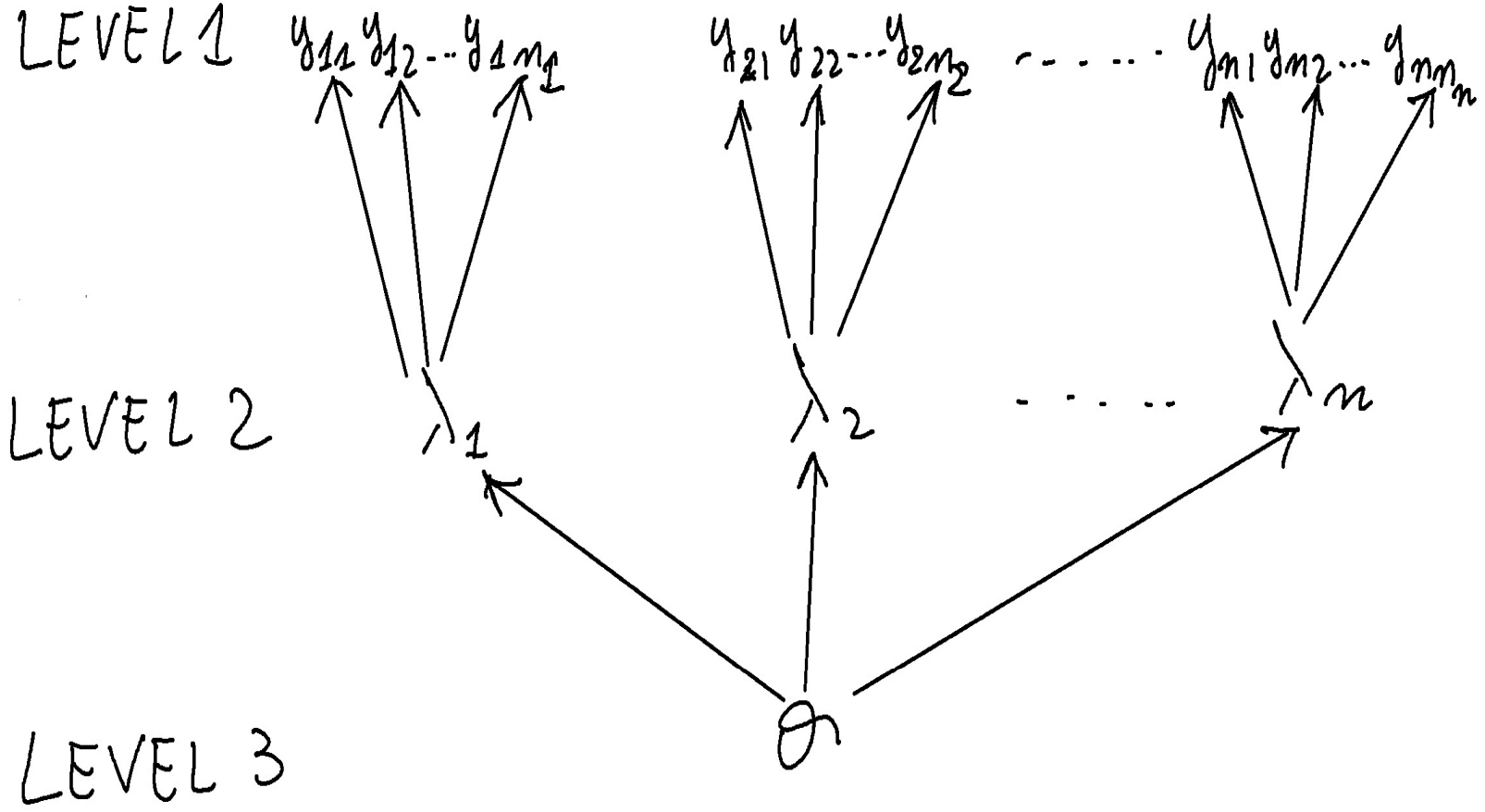
- Notation: $\underline{Y}_i = \{Y_{i1}, \ldots, Y_{i,n_i}\}$, $i = 1, \ldots, n$ data for i-th group

- Hierarchical models related to the notion of exchangeability, i.e. $\mathbb{P}(X_1, \ldots, X_n)$ invariant w.r.t. permutations (but we will not discuss it)

HIERARCHICAL MODELS

- Each group has its own distribution with a common parameter, i.e., the density of Y_{ij} is $f(y_{ij}|\lambda_i), i = 1, \ldots, n, j = 1, \ldots, n_i$
- This assumption implies a common behaviour within the group
- We assume that the functional form of f is not changing between groups (but it could)
- All the parameters λ_i's are supposed different (although sometimes some groups might have the same parameter)
- This assumption implies that the behaviour changes between groups
- All λ_i's come from the same distribution, i.e. $g(\lambda_i|\theta)$, where θ is a parameter in common
- This assumption implies that the behaviour of the groups, although different, is actually similar
- As before, a prior could be chosen for θ or a value could be plugged in, using, e.g., Empirical Bayes

HIERARCHICAL MODELS



HIERARCHICAL MODELS

- $\{y_{i1}, \ldots, y_{in_i} | \lambda_i\} \sim$ i.i.d. $f(y|\lambda_i), i = 1, \ldots, n$

  Within group sampling variability

- $\{\lambda_1, \ldots, \lambda_n\} \sim$ i.i.d. $g(\lambda|\theta)$

  Between groups sampling variability

- $\theta \sim \pi(\theta|\omega)$

  Prior distribution with hyperparameter ω

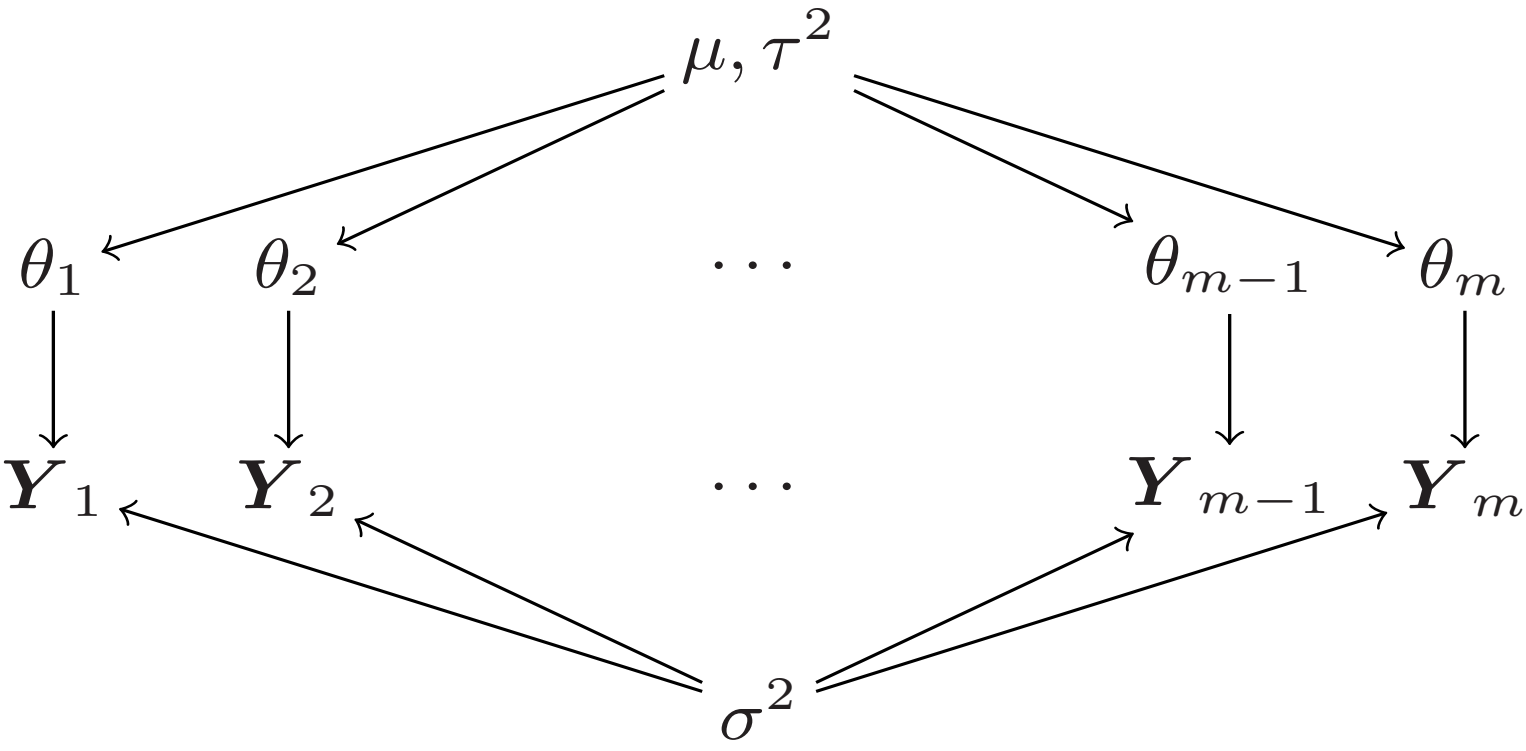
- Sometimes both $f(y|\lambda_i)$ and $g(\lambda|\theta)$ are called *sampling distributions*

- A popular model to describe heterogeneity of means across several populations is a hierarchical Normal model where both sampling distributions are Gaussian

HIERARCHICAL MODELS

- Observations in group $j, j = 1, \ldots, m$: $Y_{ji} \sim \mathcal{N}(\theta_j, \sigma^2)$ (Within group variability)
- Mean of group $j, j = 1, \ldots, m$: $\theta_j \sim \mathcal{N}(\mu, \tau^2)$ (Between groups variability)
- Independent priors on (μ, τ^2, σ^2) : $\pi(\mu)\pi(\tau^2)\pi(\sigma^2)$
  - $\mu \sim \mathcal{N}(\mu_0, \gamma_0^2)$
  - $\tau^2 \sim \mathcal{IG}(\eta_0/2, \eta_0\tau_0^2/2)$
  - $\sigma^2 \sim \mathcal{IG}(\nu/2, \nu\sigma_0^2/2)$
- Note that we assume the same variance for all the observations, while the mean is the same within a group but it changes between groups
- As seen graphically in the next slide, (μ, τ^2) provide information on Y's but, once θ is known, the distributions of Y's do not depend on (μ, τ^2)

HIERARCHICAL MODELS*



*From Hoff (2009), *A First Course in Bayesian Statistical Methods*, Springer

HIERARCHICAL MODELS

- Notation: $Y_i = (Y_{j1}, \ldots, Y_{jn_j})$, $j = 1, \ldots, m$
- $Y = (Y_1, \ldots, Y_m)$ and $\theta = (\theta_1, \ldots, \theta_m)$
- Joint posterior distribution

$$\begin{aligned}\pi(\theta, \mu, \tau^2, \sigma^2|Y) &\propto \pi(\mu, \tau^2, \sigma^2) g(\theta|\mu, \tau^2, \sigma^2) f(Y|\theta, \mu, \tau^2, \sigma^2) \\ &\propto \pi(\mu)\pi(\tau^2)\pi(\sigma^2) \left\{ \prod_{j=1}^{m} g(\theta_j|\mu, \tau^2) \right\} \left\{ \prod_{j=1}^{m} \prod_{i=1}^{n_j} f(y_{ji}|\theta_j, \sigma^2) \right\}\end{aligned}$$

- Full conditionals for μ and τ^2: $\pi(\mu, \tau^2|\theta, \sigma^2, Y) \propto \pi(\mu)\pi(\tau^2) \prod_{j=1}^{m} g(\theta_j|\mu, \tau^2)$
- $\pi(\mu|\theta, \tau^2, \sigma^2, Y) \propto \pi(\mu) \prod_{j=1}^{m} g(\theta_j|\mu, \tau^2)$
- $\pi(\tau^2|\theta, \mu, \sigma^2, Y) \propto \pi(\tau^2) \prod_{j=1}^{m} g(\theta_j|\mu, \tau^2)$

HIERARCHICAL MODELS

- The two full conditionals look very familiar!
  - Sample $(\theta_, \ldots, \theta_m)$ from $\mathcal{N}(\mu, \tau^2)$
  - $\mu \sim \mathcal{N}(\mu_0, \gamma_0^2)$
  - $\tau^2 \sim \mathcal{IG}(\eta_0/2, \eta_0\tau_0^2/2)$

- $\mu|\theta, \tau^2, Y \sim \mathcal{N}\left(\dfrac{m\overline{\theta}/\tau^2 + \mu_0/\gamma_0^2}{m/\tau^2 + 1/\gamma_0^2}, \left[m/\tau^2 + 1/\gamma_0^2\right]^{-1}\right)$

- $\tau^2|\theta, \mu, Y \sim \mathcal{IG}\left(\dfrac{\eta_0 + m}{2}, \dfrac{\eta_0\tau_0^2 + \sum_{j=1}^m (\theta_j - \mu)^2}{2}\right)$

- Here $\overline{\theta} = \displaystyle\sum_{j=1}^m \theta_j/m$

HIERARCHICAL MODELS

- Regarding θ, we can compute the full conditional for each θ_j, as dependent on $\mu, \tau^2, \sigma^2, Y_j$ since it is independent from the other θ_k's and the data from other groups
- $g(\theta_j|\mu, \tau^2, \sigma^2, Y_j) \propto g(\theta_j|\mu, \tau^2) \prod_{i=1}^{n_j} f(y_{ji}|\theta_j, \sigma^2), j = 1, \ldots, m$
- We have the product of Gaussian densities (already done, although in a simpler case)
- $\Rightarrow \theta_j|\mu, \tau^2, \sigma^2, Y_j \sim \mathcal{N}\left(\frac{n_j\overline{y}_j/\sigma^2 + 1/\tau^2}{n_j/\sigma^2 + 1/\tau^2}, \left[n_j/\sigma^2 + 1/\tau^2\right]^{-1}\right)$
- Here $\overline{y}_j = \sum_{i=1}^{n_j} y_{ji}/n_j$

HIERARCHICAL MODELS

- Full conditional of σ^2

$$\begin{aligned}\pi(\sigma^2|\theta, Y) &\propto \pi(\sigma^2)\left\{\prod_{j=1}^{m} g(\theta_j|\mu,\tau^2)\right\}\left\{\prod_{j=1}^{m}\prod_{i=1}^{n_j} f(y_{ji}|\theta_j,\sigma^2)\right\} \\ &\propto (\sigma^2)^{-\nu_0/2+1} e^{-\nu_0\sigma_0^2/(2\sigma^2)} \cdot (\sigma^2)^{-\sum_{j=1}^{m} n_j/2} e^{-\sum_{j=1}^{m}\sum i=1^{n_j}(y_{ij}-\theta_j^2)/(2\sigma^2)}\end{aligned}$$

- $\Rightarrow \sigma^2|\theta, Y \sim \mathcal{IG}\left((\nu_0 + \sum_{j=1}^{m} n_j)/2, (\nu_0\sigma_0^2 + \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij}-\theta_j)^2)/2\right)$

- We use the Gibbs algorithm to get a sample from the posterior distribution since all the conditional distributions are properly specified

HIERARCHICAL MODELS*

- We are interested in learning about the mortality rates due to heart transplant surgery for 94 hospitals
- The number of deaths within 30 days of heart transplant surgery is recorded for each hospital
- It is the same problem, but addressed differently, we considered when checking if there was discrepancy between observed values and predictive distribution based on the select pair sampling model/prior
- Each hospital has a true mortality rate λ_i, and so one wishes to simultaneously estimate the 94 rates $\lambda_1, \ldots, \lambda_{94}$
- It is reasonable to believe a priori that the true rates are similar in size, which implies a dependence structure between the parameters
- If one is told some information about a particular hospital's true rate, that information would likely affect one's belief about the location of a second hospital's rate

*Example from Albert's book

HIERARCHICAL MODELS

- In addition, we record for each hospital an expected number of deaths called the exposure, denoted by e
- We let y_i and e_i denote the respective observed number of deaths and exposure for the i-th hospital
- A standard model assumes that the number of deaths y_i follows a Poisson distribution with mean $e_i\lambda_i$ and the objective is to estimate the mortality rate per unit exposure λ_i
- The fraction y_i/e_i is the number of deaths per unit exposure and can be viewed as an estimate of the death rate for the i-th hospital
- Suppose we are interested in simultaneously estimating the true mortality rates $\{\lambda_i\}$ for all hospitals
- One option is simply to estimate the true rates by using the individual death rates: $y_1/e_1, \ldots, y_{94}/e_{94}$

HIERARCHICAL MODELS

- Unfortunately, the individual rates y_i/e_i's can be poor estimates, especially for the hospitals with small exposures
- Some of those hospitals did not experience any deaths and the individual death rate $y_i/e_i = 0$ would likely underestimate the hospital's true mortality rate
- Since the individual death rates can be poor, it seems desirable to combine the individual estimates in some way to obtain improved estimates
- Suppose we can assume that the true mortality rates are equal across hospitals, i.e. $\lambda_1 = \ldots = \lambda_{94}$
- Under this "equal-means" Poisson model, the estimate of the mortality rate for the i-th hospital would be the pooled estimate $\frac{\sum_{i=1}^{94} y_j}{\sum_{i=1}^{94} e_j}$
- The pooled estimate is based on the strong assumption that the true mortality rate is the same across hospitals but this is questionable since one would expect some variation in the true rates

HIERARCHICAL MODELS

- We have discussed two possible estimates for the mortality rate of the i-th hospital: the individual estimate y_i/e_i and the pooled estimate $\frac{\sum_{i=1}^{94} y_j}{\sum_{i=1}^{94} e_j}$

- A third possibility is the compromise estimate $(1-\varepsilon)\frac{y_i}{e_i} + \varepsilon\frac{\sum_{i=1}^{94} y_j}{\sum_{i=1}^{94} e_j}$

- One could consider equal mortality rate so that $y_i \sim \mathcal{P}(e_i\lambda)$

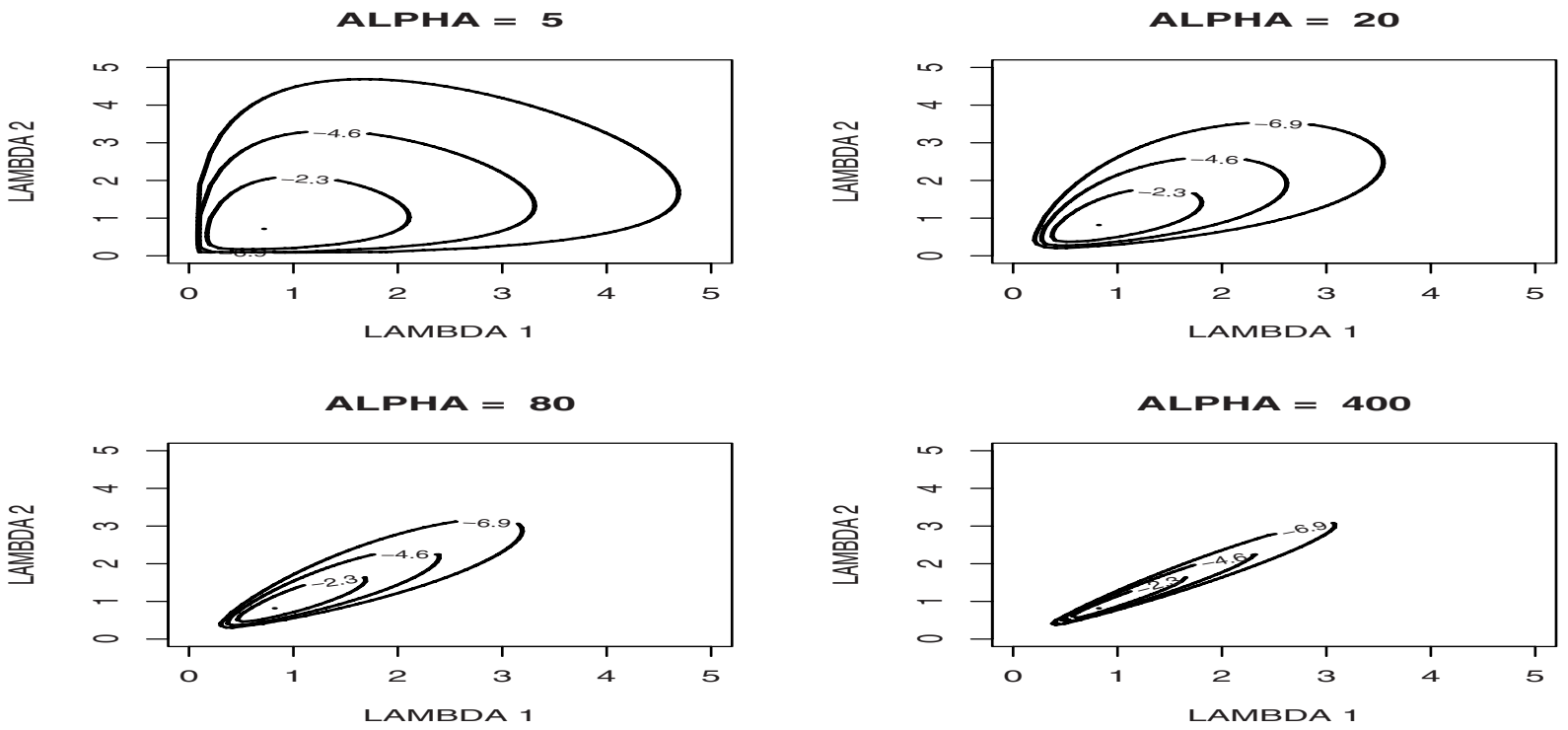
- The gamma prior for λ is conjugate w.r.t. the model, as already seen earlier in similar situations

- We leave the computations as an exercise: Albert considered a non informative prior $\pi(\lambda) \propto 1/\lambda$ but a proper gamma prior can be used

HIERARCHICAL MODELS

- $y_i \sim \mathcal{P}(e_i\lambda_i)$, $i = 1, \ldots, 94$
- $\lambda_1, \ldots, \lambda_{94} \sim \mathcal{G}(\alpha, \alpha/\mu)$, with mean μ and variance μ^2/α
  $g(\lambda(\alpha, \mu) = \dfrac{(\alpha/\mu)^\alpha \lambda^{\alpha-1} e^{-\alpha\lambda/\mu}}{\Gamma(\alpha)}$
- Consider the hyperparameters μ and α as independent
- $\mu \sim \mathcal{IG}(a, b)$ and $\pi(\alpha)$ for α
- If we consider a Dirac prior at α_0 for α and just the first two hospitals, we get
  $g(\lambda_1, \lambda_2|\alpha_0) \propto \dfrac{(\lambda_1\lambda_2)^{\alpha_0-1}}{(\alpha_0(\lambda_1 + \lambda_2) + b)^{2\alpha_0+a}}$
- With $\mu \sim \mathcal{IG}(10, 10)$ its mean is 1 and (λ_1, λ_2) will be around $(1, 1)$
- We consider different values of α_0

HIERARCHICAL MODELS*



- Contour graphs of the exchangeable prior on (λ_1, λ_2)

*From Albert's book

HIERARCHICAL MODELS

- We now provide just a sketch of the analysis performed by Albert in his book: more details and R codes can be found in it
- Consider $g(\mu) \propto 1/\mu$ and $g(\alpha) = \dfrac{z_0}{(\alpha + z_0)^2}$
- Conditional on μ and α, the λ_i's have independent posterior distributions: $\lambda_i|\alpha, \mu, y_i \sim \mathcal{G}(y_i + \alpha, e_i + \alpha/\mu)$
- $\Rightarrow E(\lambda_i|\alpha, \mu, y_i) = \dfrac{y_i + \alpha}{e_i + \alpha/\mu}$
- The posterior on α and μ is given, for a constant K, by

$$p(\alpha, \mu|\text{data}) = K \frac{1}{\Gamma^{94}(\alpha)} \prod_{j=1}^{94} \left[\frac{(\alpha/\mu)^\alpha \Gamma(\alpha + y_i)}{(\alpha/\mu + e_i)^{(\alpha+y_i)}} \right] \frac{z_0}{(\alpha + z_0)^2} \frac{1}{\mu},$$

HIERARCHICAL MODELS

- Having the posterior conditionals on the λ_i's in closed form (i.e. a Gamma distribution) and knowing the posterior conditional for (α, μ) apart from a constant, it is possible to use a Gibbs sampling algorithm with Metropolis-Hastings steps within

- At this point we can compare hospitals

- If we look for the "best hospital" then we should look for the one with the lowest estimated posterior mean of λ_i's

- If we want to compare two hospitals, i and j, then we can estimate $\mathbb{P}(\lambda_i < \lambda_j)$ by simply counting the frequency of the samples $\lambda_i^{(s)} < \lambda_j^{(s)}$

- We say that hospital i is better than hospital j if $\mathbb{P}(\lambda_i < \lambda_j) > 0.5$