# Machine Learning - Waseda University
# Markov Chain Monte Carlo Methods

AD

July 2011

# Introduction

- Bayesian model: likelihood $f(x|\theta)$ and prior distribution $\pi(\theta)$.
- Bayesian inference is based on the posterior distribution

$$\pi(\theta|x) = \frac{\pi(\theta) f(x|\theta)}{\pi(x)}$$

where

$$\pi(x) = \int_{\Theta} \pi(\theta) f(x|\theta) d\theta.$$

- Many point estimates require computing additional integrals, e.g.

$$\mathbb{E}[\varphi(\theta)|x] = \int_{\Theta} \varphi(\theta) \pi(\theta|x) d\theta$$

- Assume $\theta = (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2$ then, if $\theta_2$ are some nuisance parameters, we are only interested in the marginal posterior distribution

$$\pi(\theta_1 | x) = \int_{\Theta_2} \pi(\theta_1, \theta_2 | x) \, d\theta_2.$$

- Assume $Y | \theta \sim g(y | \theta)$ then, given the observations $x$, the predictive distribution is

$$\pi(y | x) = \int_{\Theta} g(y | \theta) \, \pi(\theta | x) \, d\theta.$$

- Athlough Bayesian inference is conceptually simple, it requires being able to compute potentially high-dimensional integrals.
- In previous lectures, we have discussed mostly examples where these calculations could be performed analytically. However, analytic tractability seriously restricts the class of models we can work with.
- Approximations such as the Laplace's approximation and BIC are rather crude and require large sample sizes. Consequently until the beginning of the 90's, Bayesian inference for all but simple models could not be implemented.
- Alternative deterministic Bayesian approximation techniques include variational and expectation-propagation but make strong assumptions.

- MCMC are a class of powerful simulation-based algorithms that allows us to sample (approximately) from any high-dimensional probability distribution.
- The availability of these algorithms has truly revolutionized many fields, allowing people o fit complex models.
- MCMC are now widely used in bioinformatics, machine learning, econometrics, applied statistics, genetics, etc.

# Introduction to Monte Carlo

- Assume you are interested in approximating an high-dimensional pdf $\pi(\theta|x)$.
- A Monte Carlo approximation consists of sampling a large number $N$ of i.i.d. random variables $\theta^{(i)} \sim \pi(\theta|x)$ and build the following approximation

$$\widehat{\pi}_N(\theta|x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta^{(i)}}(\theta)$$

where $\delta_a(\theta)$ is the delta-Dirac mass which is such that

$$\int_A \delta_a(\theta)\, d\theta = \begin{cases} 1 & \text{if } a \in A, \\ 0 & \text{otherwise.} \end{cases}$$

- This approach is in contrast with what is usually done in parametric statistics, *i.e.* start with samples and then introduce a distribution with an algebraic representation for the underlying population.

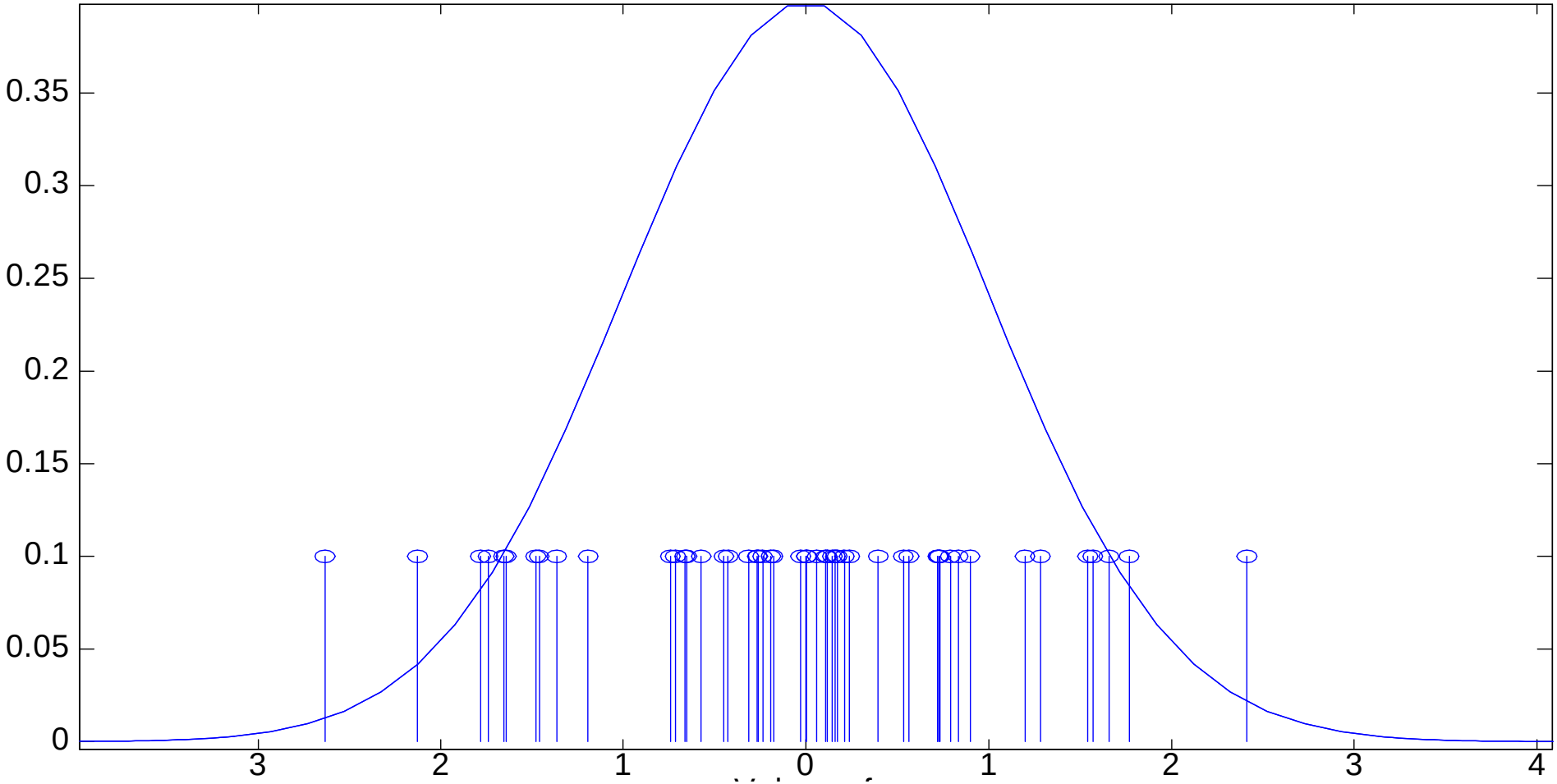



Figure: Normal $\mathcal{N}(0,1)$ and Monte Carlo approximation using $N = 50$ samples

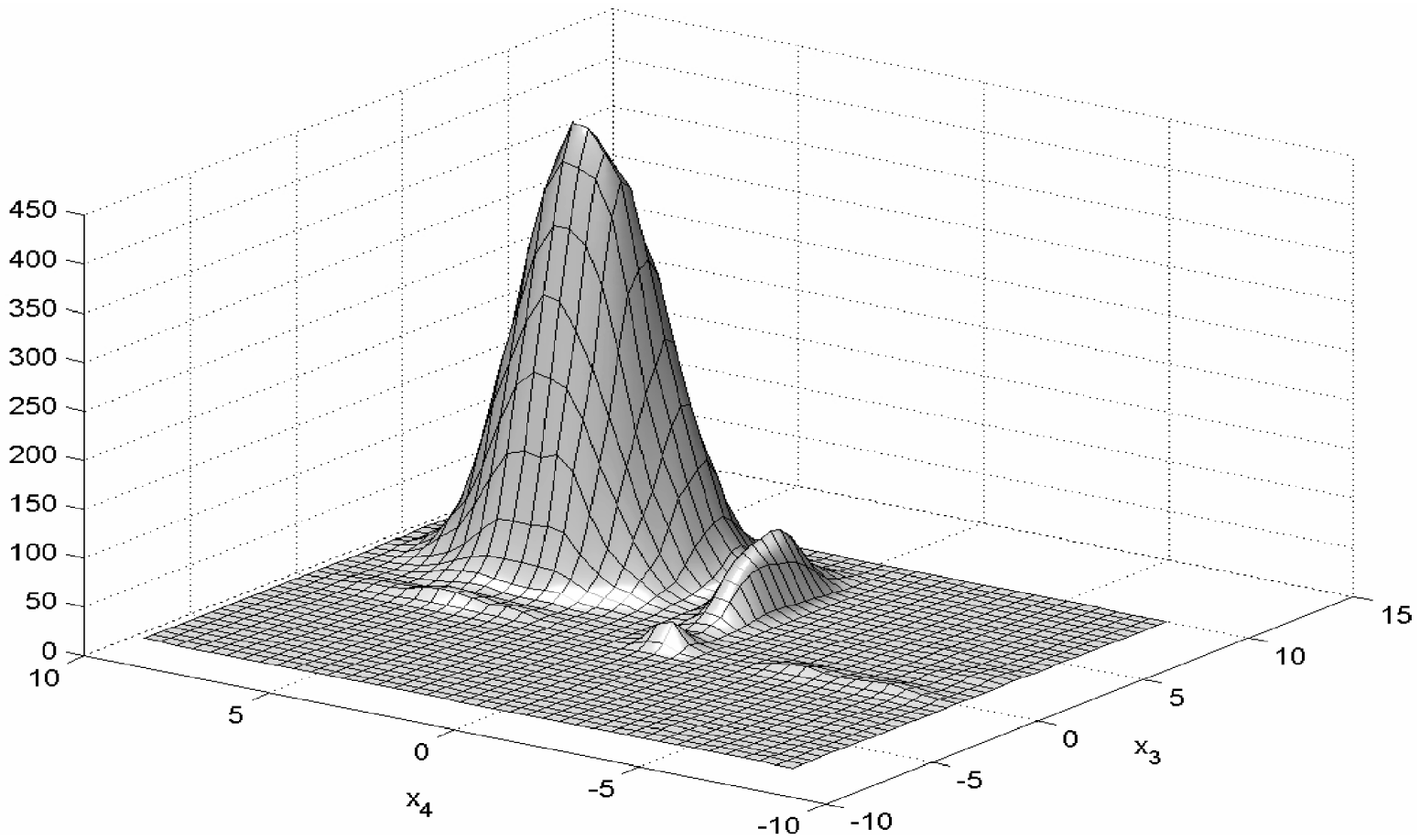



Figure: Bivariate non-standard probability density function

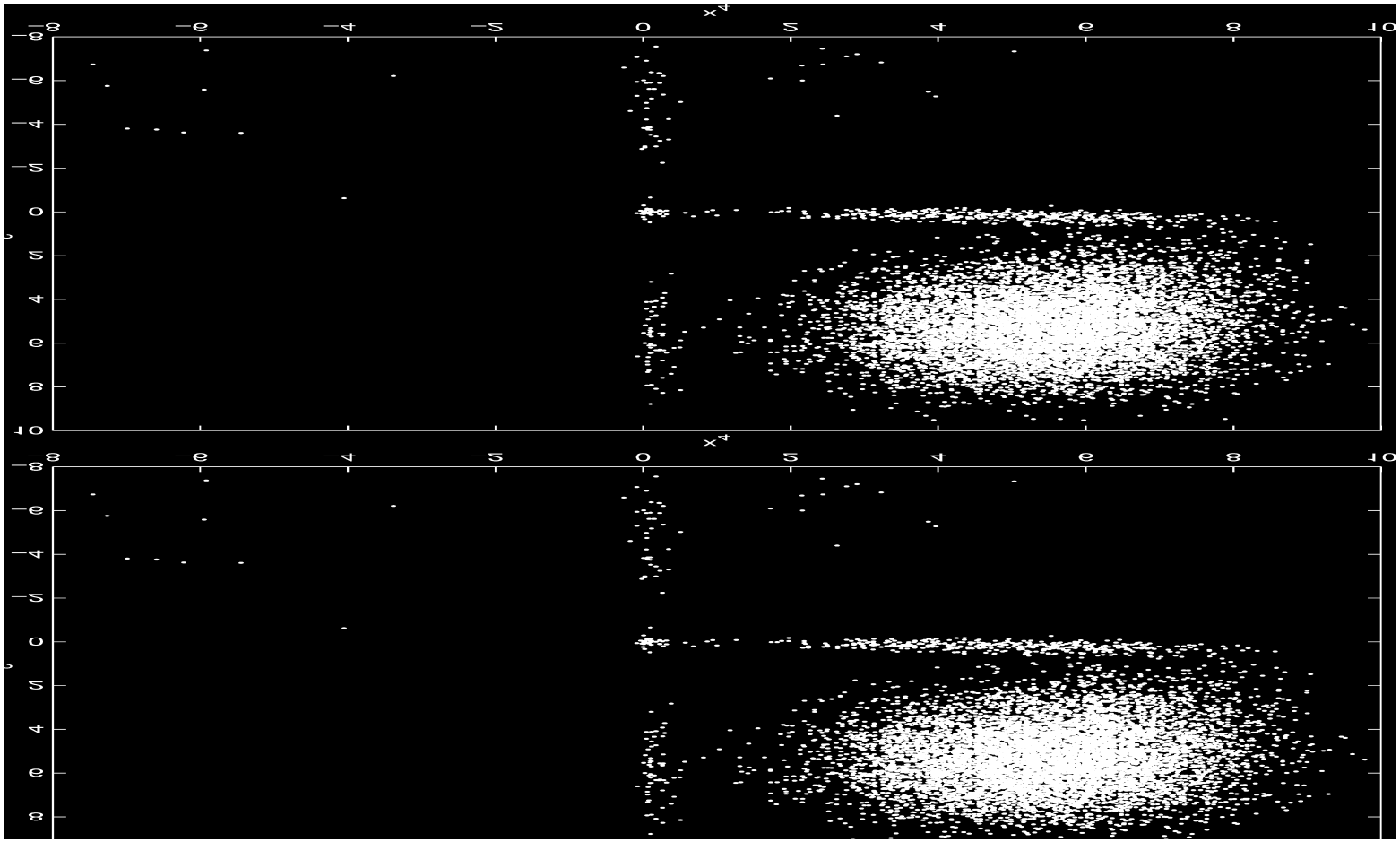


Figure: Scatterplot of $N = 1000$ random samples

- You can think of the Monte Carlo method as an 'clever' discretization of the space where automatically the samples concentrate themselves in regions of high probability mass.
- An alternative deterministic approximation would consist of discretizing $\Theta$ using a regular grid and then computing

$$\widetilde{\pi}_N(\theta) = \frac{\sum_{i=1}^{N} \pi\left(\theta^{(i)} \middle| x\right) \delta_{\theta^{(i)}}(\theta)}{\sum_{j=1}^{N} \pi\left(\theta^{(j)} \middle| x\right)}.$$

- Such an approach will be extremely inefficient in high-dimensional spaces. If $\Theta = \mathbb{R}^d$ and you discretize say each dimension using $p$ values, then the total number of points in the grid will be $p^d$. Given we are routinely interested in problems where $d = 100$, even the crudest discretization $p = 2$ would require $2^{100} >> 1$ points.

# Properties of Monte Carlo Estimates

- Now consider the problem of estimating

$$\mathbb{E}_{\pi}(\varphi) = \int_{\Theta} \varphi(\theta)\, \pi(\theta|\, x)\, d\theta$$

  using Monte Carlo where $\varphi : \Theta \longrightarrow \mathbb{R}$.
- We substitute to $\pi(\theta|\, x)$ its Monte Carlo approximation $\widehat{\pi}_N$ and obtain

$$\begin{aligned}\mathbb{E}_{\widehat{\pi}_N}(\varphi) &= \int_{\Theta} \varphi(\theta) \left( \sum_{i=1}^{N} \frac{1}{N} \delta_{\theta^{(i)}}(\theta) \right) d\theta = \sum_{i=1}^{N} \frac{1}{N} \int_{\Theta} \varphi(\theta)\, \delta_{\theta^{(i)}}(\theta)\, d\theta \\ &= \frac{1}{N} \sum_{i=1}^{N} \varphi(\theta^{(i)}).\end{aligned}$$

- This estimate is unbiased

$$\mathbb{E}_{\{\theta^{(i)}\}} \left[ \mathbb{E}_{\widehat{\pi}_N}(\varphi) \right] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\theta^{(i)}} \left( \varphi(\theta^{(i)}) \right) = \mathbb{E}_{\pi}(\varphi).$$

- The variance is given by

$$\mathbb{V}_{\left\{\theta^{(i)}\right\}}\left[\mathbb{E}_{\widehat{\pi}_N}(\varphi)\right] = \frac{1}{N}\mathbb{V}_{\pi}\left(\varphi(\theta)\right).$$

- The CLT yields

$$\sqrt{N}\left(\mathbb{E}_{\widehat{\pi}_N}(\varphi) - \mathbb{E}_{\pi_N}(\varphi)\right) \xrightarrow{\mathsf{d}} \mathcal{N}\left(0, \mathbb{V}_{\pi}\left(\varphi(\theta)\right)\right).$$

- The most remarkable property of the MC estimate is that the rate of convergence is independent of the dimension of $\Theta$.
- It is sometimes said that the MC beats the *curse of dimensionality*. This is not quite true.

- MC can easily be used to compute marginal distributions. We approximate

$$\begin{aligned}\pi\left(\theta_1 \middle| x\right) &= \int_{\Theta_2} \pi\left(\theta_1, \theta_2 \middle| x\right) d\theta_2 \approx \int_{\Theta_2} \widehat{\pi}_N\left(\theta_1, \theta_2 \middle| x\right) d\theta_2 \\ &= \frac{1}{N} \int \sum_{i=1}^{N} \delta_{\theta_1^{(i)}, \theta_2^{(i)}}\left(\theta_1, \theta_2\right) d\theta_2 \\ &= \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta_1^{(i)}}\left(\theta_1\right) = \widehat{\pi}_N\left(\theta_1 \middle| x\right)\end{aligned}$$

- Similarly the predictive distribution can be approximated easily

$$\begin{aligned}\pi\left(y \middle| x\right) &= \int_{\Theta} g\left(y \middle| \theta\right) \pi\left(\theta \middle| x\right) d\theta \approx \int_{\Theta} g\left(y \middle| \theta\right) \widehat{\pi}_N\left(\theta \middle| x\right) d\theta \\ &= \frac{1}{N} \sum_{i=1}^{N} g\left(y \middle| \theta^{(i)}\right).\end{aligned}$$

- However, the marginal likelihood $\pi\left(x\right)$ cannot be estimated easily using samples from $\pi\left(\theta \middle| x\right)$!

# Sampling from complex distributions

- There are standard methods to sample from classical distributions such as Beta, Gamma, Normal, Poisson etc. We will not detail them here.
- We are interested in problems where $\pi(\theta|x)$ is not standard and is only known up to a normalizing constant; i.e.

$$\pi(\theta|x) = \frac{\pi(\theta) f(x|\theta)}{\pi(x)}$$

  where $\pi(\theta) f(x|\theta)$ is known pointwise whereas $\pi(x)$ is unknown.
- There is no method available able to sample from any high dimensional probability distribution.

# Accept-Reject Method

- The rejection method relies on a so-called proposal distribution $q(\theta|x)$ which is selected such that it is easy to sample from it.
- We have $q(\theta|x) \propto q^*(\theta|x)$ where $q^*(\theta|x)$ is known pointwise.
- We need $q^*(\theta|x)$ to 'dominate' $\pi(\theta) f(x|\theta)$; i.e.

$$C = \sup_{\theta \in \Theta} \frac{\pi(\theta) f(x|\theta)}{q^*(\theta|x)} < +\infty$$

- This implies $\pi(\theta|x) > 0 \Rightarrow q^*(\theta|x) > 0$ but also that the tails of $q(\theta|x)$ must be thicker than the tails of $\pi(\theta|x)$.

Consider $M \geq C$. Then the accept/reject procedure proceeds as follows.

1. Sample $\theta^* \sim q(\theta | x)$ and $U \sim \mathcal{U}[0,1]$.
2. If $U < \frac{\pi(\theta^*) f(x|\theta^*)}{M q^*(\theta^*|x)}$ then return $\theta^*$; otherwise return to step 1.

# Proof of Validity

- We have for any $\theta \in \Theta (= \mathbb{R}$, this is only to simplify notation)

$$\begin{aligned}
& \Pr\left(\theta^* \leq \theta \text{ and } \theta^* \text{ accepted}\right) \\
= & \int_{\Theta} \int_0^1 \mathbb{I}\left(\theta^* \leq \theta\right) \mathbb{I}\left(u \leq \frac{\pi\left(\theta^*\right) f\left(x \middle| \theta^*\right)}{M q^*\left(\theta^* \middle| x\right)}\right) q\left(\theta^* \middle| x\right) \times 1 du d\theta^* \\
= & \int_{-\infty}^{\theta} \frac{\pi\left(\theta^*\right) f\left(x \middle| \theta^*\right)}{M q^*\left(\theta^* \middle| x\right)} q\left(\theta^* \middle| x\right) d\theta^* \\
= & \frac{\int_{-\infty}^{\theta} \pi\left(\theta^*\right) f\left(x \middle| \theta^*\right) d\theta^*}{M \int_{\Theta} q^*\left(\theta^* \middle| x\right) d\theta^*}.
\end{aligned}$$

- The probability of being accepted is simply

$$\Pr\left(\theta^* \text{ accepted}\right) = \frac{\int_{\Theta} \pi\left(\theta^*\right) f\left(x \middle| \theta^*\right) d\theta^*}{M \int_{\Theta} q^*\left(\theta^* \middle| x\right) d\theta^*} = \frac{\pi(x)}{M \int_{\Theta} q^*\left(\theta^* \middle| x\right) d\theta^*}.$$

- Finally, we have

$$\begin{aligned}
\Pr\left(\theta^* \leq \theta \middle| \theta^* \text{ accepted}\right) &= \frac{\Pr\left(\theta^* \leq \theta \text{ and } \theta^* \text{ accepted}\right)}{\Pr\left(\theta^* \text{ accepted}\right)} \\
&= \frac{\int_{-\infty}^{\theta} \pi\left(\theta^*\right) f\left(x \middle| \theta^*\right) d\theta^*}{\pi\left(x\right)} \\
&= \int_{-\infty}^{\theta} \pi\left(\theta^* \middle| x\right) d\theta^*.
\end{aligned}$$

- The acceptance probability $\Pr\left(\theta^{*} \text{ accepted}\right)$ is a measure of efficiency.
- The number of trials before accepting a candidate follows a geometric distribution

$$\begin{aligned} \Pr\left(k^{\text{th}} \text{ proposal accepted}\right) &= (1-\rho)^{k-1}\rho \\ \text{where } \rho &= \left(\frac{\pi(x)}{M\int_{\Theta} q^{*}\left(\theta^{*}\,|\,x\right)d\theta^{*}}\right) \end{aligned}$$

thus its expected value is

$$\sum_{k=0}^{\infty} k\left(1-\rho\right)^{k-1}\rho = \frac{1}{\rho} = \frac{1}{\Pr\left(\theta^{*} \text{ accepted}\right)}.$$

# Toy Example

- Consider the following distribution

$$\pi\left(\theta|\, x\right) \propto \exp\left(-\theta^2/2\right)\left(\sin\left(\theta x\right)^2 + 3\cos\left(\theta x\right)^2 \sin\left(4x\right)^2 + 1\right)$$

- We use $q^*\left(\theta|\, x\right) = q^*\left(\theta\right) = \exp\left(-\theta^2/2\right)$, that is $q\left(\theta\right) = \mathcal{N}\left(\theta; 0, 1\right)$.
- This proposal is easy to sample and

$$\frac{\exp\left(-\theta^2/2\right)\left(\sin\left(\theta x\right)^2 + 3\cos\left(\theta x\right)^2 \sin\left(4x\right)^2 + 1\right)}{\exp\left(-\theta^2/2\right)} \leq 1 + 3 + 1 = 5$$

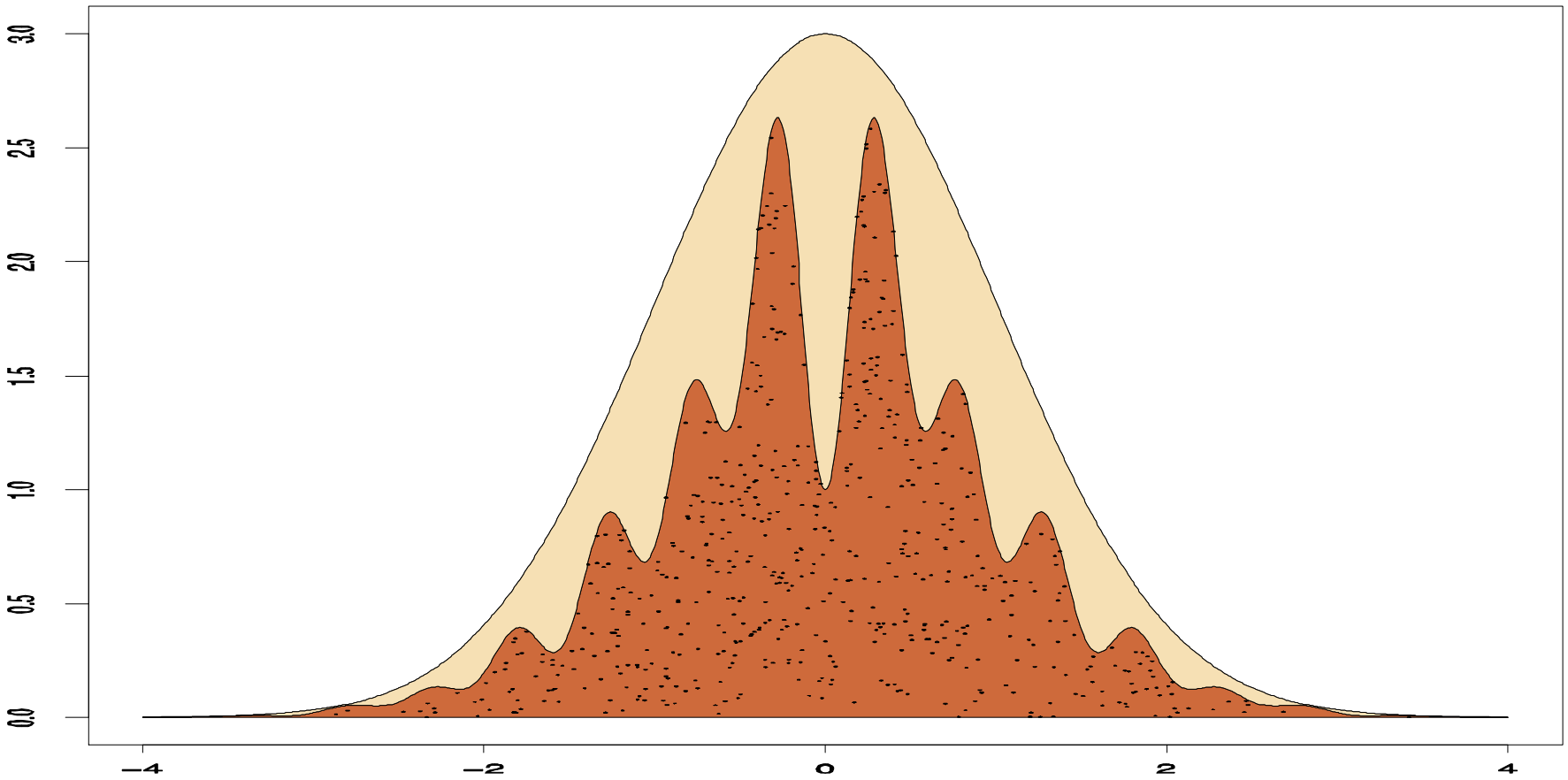



Figure: $5q^{*}(\theta)$ and unnormalized target distribution $\exp\left(-\theta^{2}/2\right)\left(\sin(\theta x)^{2}+3\cos(\theta x)^{2}\sin(4x)^{2}+1\right)$

# Using the Prior as Proposal

- A simple choice consists of selecting $q(\theta|x) = q^*(\theta|x) = \pi(\theta)$.
- This is possible if the likelihood is upper bounded as

$$\sup_{\theta\in\Theta} \frac{\pi(\theta) f(x|\theta)}{q^*(\theta|x)} = \sup_{\theta\in\Theta} f(x|\theta) \leq M$$

- In this case, expected value before acceptance is $\rho^{-1}$ where

$$\rho = \frac{\int_\Theta \pi^*(\theta^*) f(x|\theta^*) d\theta^*}{M \int_\Theta q^*(\theta^*|x) d\theta^*} = \frac{\pi(x)}{M}$$

and provides us with an estimate of the marginal likelihood.

# Inefficiency of the Accept-Reject Strategy

- Consider $\theta \sim \mathcal{N}\left(0, \tau^2\right)$ and $X_i | \theta \sim \mathcal{N}\left(\theta, \sigma^2\right)$.
- In this case, we have

$$\pi\left(\theta | x_{1:n}\right) = \mathcal{N}\left(\theta; m, v^2\right)$$

where $v^2 = \left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)^{-1}$, $m = v^2\left(\frac{\sum_{i=1}^n x_i}{\sigma^2}\right)$ and

$$\pi\left(x_{1:n}\right) = \frac{v}{\left(2\pi\sigma^2\right)^{n/2}\tau} \exp\left(\frac{m^2}{2v^2} - \frac{\sum_{i=1}^n x_i^2}{2\sigma^2}\right).$$

whereas the likelihood is bounded by $\left(2\pi\sigma^2\right)^{-n/2}$ so

$$\rho = \frac{\pi\left(x_{1:n}\right)}{M} = \frac{v}{\tau} \exp\left(\frac{m^2}{2v^2} - \frac{\sum_{i=1}^n x_i^2}{2\sigma^2}\right).$$

- For $\tau^2 >> 1$ and $\sigma^2 << 1$, we have $v^2 \approx \frac{\sigma^2}{n}$, $m \approx \overline{x}$

$$\rho \approx \frac{\sigma^2}{n\tau} \exp\left(\frac{n\left(\overline{x}^2 - \overline{x^2}\right)}{2\sigma^2}\right) \to 0$$

# Application to Genetic Linkage Model

- Consider the following genetic linkage model where observations

$$(X_1, X_2, X_3, X_4) \sim \mathcal{M}\left(n; \frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right)$$

  with $\theta \in (0,1)$.
- We set an uniform prior $\pi(\theta) = \mathcal{U}[0,1](\theta)$ and want to estimate $\pi(\theta | x_{1:4})$ where $(x_1, x_2, x_3, x_4) = (125, 18, 20, 34)$.
- We have

$$\pi(\theta | x_{1:4}) \propto (2+\theta)^{x_1} (1-\theta)^{x_2+x_3} \theta^{x_4} 1_{(0,1)}(\theta).$$

- This univariate distribution is not standard and we sample from it using Accept-Reject.

- We propose to use $q(\theta|x) = q^*(\theta|x) = \mathcal{U}[0,1](\theta)$.
- To apply accept-reject, we need to be able to upper bound over $\theta \in (0,1)$ the function

$$g(\theta) = (2+\theta)^{x_1} (1-\theta)^{x_2+x_3} \theta^{x_4}.$$

- Using a simple optimization algorithm (direct search or EM), we get

$$g(\theta) \leq g(\theta_{\max})$$

where $\theta_{\max} = 0.6268$ and $g(\theta_{\max}) = \exp(67.3841)$.

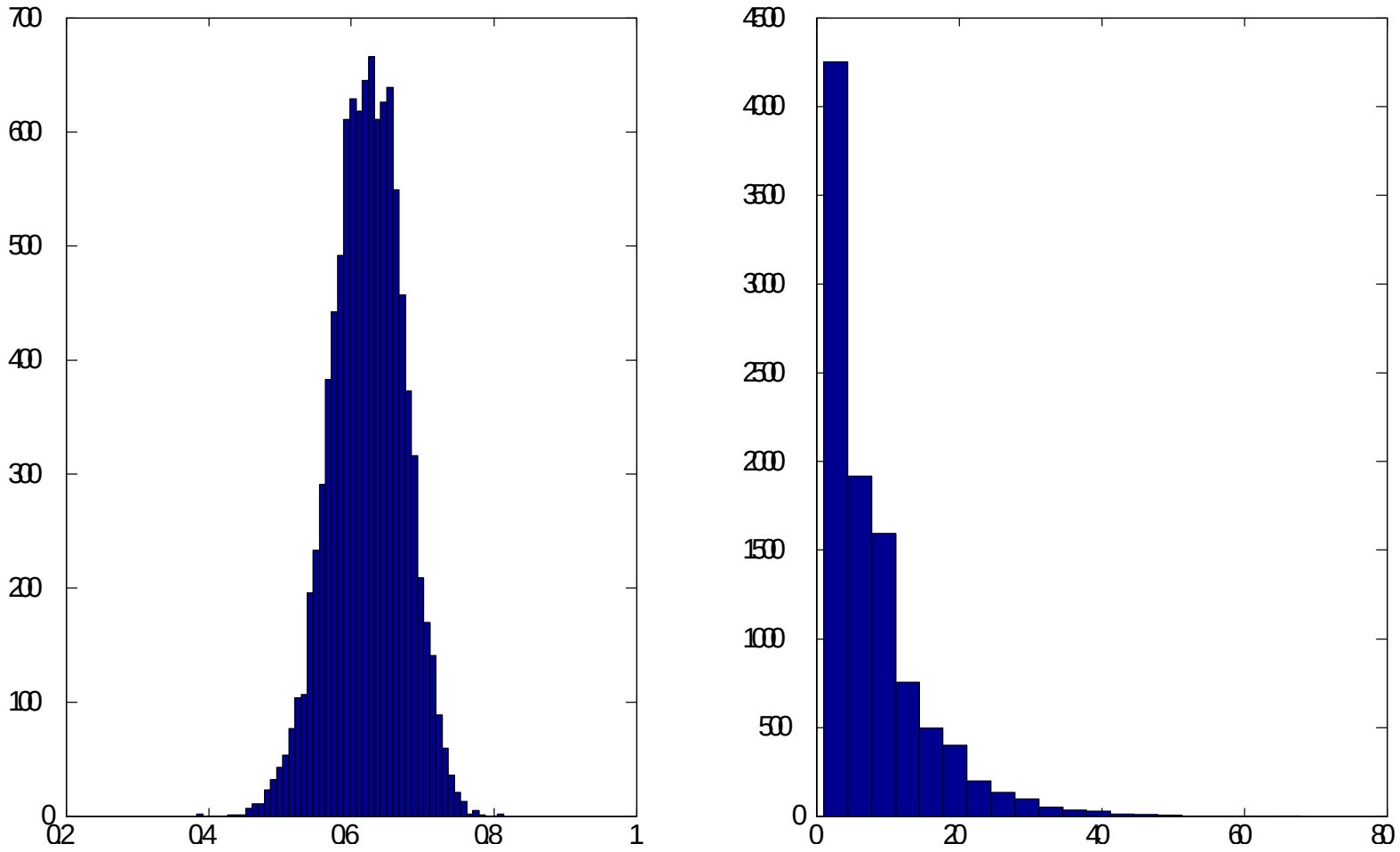



Figure: Histogram approximation of $\pi\left(\theta \middle| x_{1:4}\right)$ (left) and histogram approximation of waiting time distribution before acceptance (mean 7.8) (right)

# Limitations of Accept-Reject

- For complex problems, it will be impossible to bound the ratio between the (unnormalized) target and the proposal.
- Even if it is possible to obtain such a bound bound, the computational complexity typically increases exponentially fast with the dimension of the problem.
- We need to use more powerful techniques.

# The Gibbs Sampler

- The Gibbs sampler is an iterative popular method to sample from high dimensional probability distributions which has found numerous applications in Bayesian statistics.
- For sake of simplicity, consider first that we are interested in sampling from $\pi(\theta)$ where $\theta = (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2$.
- The Gibbs sampler relies on the fact that, although it is impossible to sample from $\pi(\theta)$, it is often possible to sample from the conditional distributions

$$\pi(\theta_1 | \theta_2) \text{ and } \pi(\theta_2 | \theta_1).$$

- Note that here $\pi(\theta)$ denotes a generic pdf and could be the posterior!

# Two Component Gibbs Sampler aka Data Augmentation

- To sample from $\pi(\theta_1, \theta_2)$, the Gibbs sampler generates a Markov chain $\left(\theta_1^{(i)}, \theta_2^{(i)}\right)$ as follows.
- Initialization: Set $\left(\theta_1^{(0)}, \theta_2^{(0)}\right)$ deterministically or randomly.
- Iteration $i$; $i \geq 1$.
  - Sample $\theta_1^{(i)} \sim \pi\left(\theta_1 | \, \theta_2^{(i-1)}\right)$
  - Sample $\theta_2^{(i)} \sim \pi\left(\theta_2 | \, \theta_1^{(i)}\right)$.

- Under weak assumptions, after many such iterations (usually several hundred or thousand are required), the sampling distribution of $\left(\theta_1^{(i)}, \theta_2^{(i)}\right)$ will approximate closely $\pi\left(\theta_1, \theta_2\right)$.
- We should discard the first hundred/thousand simulated samples, say $N_0$, whose distribution might be 'far' from $\pi\left(\theta_1, \theta_2\right)$.
- The samples $\left(\theta_1^{(i)}, \theta_2^{(i)}\right)$ are not independent but it is still valid to consider the approximation

$$\widehat{\pi}_N\left(\theta_1, \theta_2\right) = \frac{1}{N_0 - N + 1} \sum_{i=N_0}^{N} \delta_{\left(\theta_1^{(i)}, \theta_2^{(i)}\right)}\left(\theta_1, \theta_2\right).$$

- Note that a deterministic version of this algorithm: $\theta_1^{(i)} = \arg\max \pi\left(\theta_1 | \theta_2^{(i-1)}\right)$ and $\theta_2^{(i)} = \arg\max \pi\left(\theta_2 | \theta_1^{(i)}\right)$ would not converge typically to the global maximum of $\pi\left(\theta_1, \theta_2\right)$.

# Toy Example

- Consider the distribution

$$\begin{aligned}\pi(\theta_1, \theta_2) &= \mathcal{N}\left((\theta_1, \theta_2), \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \\ &\propto \exp\left(-\frac{1}{2(1-\rho^2)}\left(\theta_1^2 - 2\rho\theta_1\theta_2 + \theta_2^2\right)\right)\end{aligned}$$

  where $|\rho| < 1$.
- To sample from $\pi(\theta_1, \theta_2)$, the Gibbs sampler would sample iteratively and successively from

$$\begin{aligned}\pi(\theta_1 | \theta_2) &= \mathcal{N}\left(\theta_1; \rho\theta_2, 1-\rho^2\right), \\ \pi(\theta_2 | \theta_1) &= \mathcal{N}\left(\theta_2; \rho\theta_1, 1-\rho^2\right).\end{aligned}$$

- As $|\rho| \longrightarrow 1$ the algorithm will converge more slowly.
- Generally speaking, the more correlated $\theta_1$ and $\theta_2$ are the slowler the algorithm converges.

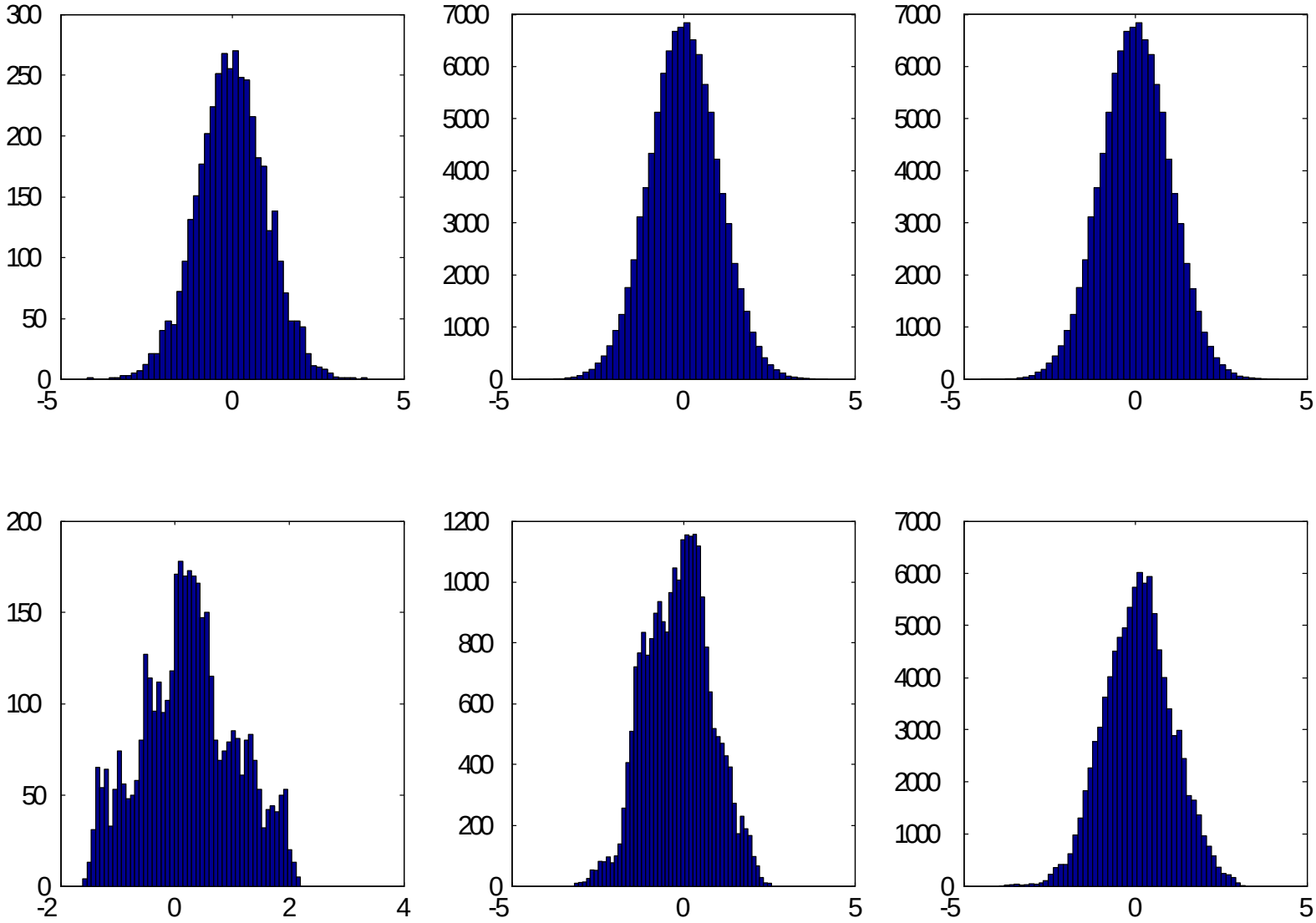



Figure: Estimation of $\pi(\theta_1) = \mathcal{N}(\theta_1, 0, 1)$ using the Gibbs sampler for $N = 5000$ (left), 25000 (center) and 100000 (right) with $N_0 = 1000$ and $\rho = 0.2$ (top), $\rho = 0.999$ (bottom)

# Application to Genetic Linkage Model

- In this example, we have

$$(X_1, X_2, X_3, X_4) \sim \mathcal{M}\left(n; \frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right)$$

  and the target posterior distribution is given by

$$\pi(\theta | x_{1:4}) \propto (2+\theta)^{x_1} (1-\theta)^{x_2+x_3} \theta^{x_4} 1_{(0,1)}(\theta).$$

- We cannot use the Gibbs sampler in this case.
- Now assume we introduce the missing data $(Z_1, Z_2)$ such that $Z_1 + Z_2 = X_1$ and

$$(Z_1, Z_2, X_2, X_3, X_4) \sim \mathcal{M}\left(n; \frac{1}{2}, \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right).$$

- This complete likelihood is given by

$$g(z_{1:2}, x_{2:4} | \theta) \propto \theta^{z_2+x_4} (1-\theta)^{x_2+x_3}.$$

- In the EM, we introduced these missing data to ease the maximization of the likelihood. Here we use them to ease the simulation from the posterior.
- We will now use the Gibbs sampler to generate sample from

$$\pi\left(\theta, z_1, z_2 \middle| x_{1:4}\right).$$

To implement the Gibbs sampler, we need to be able to sample from $\pi\left(\theta \middle| x_{1:4}, z_{1:2}\right)$ and $\pi\left(z_{1:2} \middle| x_{1:4}, \theta\right)$.

- We have

$$\begin{aligned}\pi\left(\theta \middle| x_{1:4}, z_{1:2}\right) &\propto g\left(z_{1:2}, x_{2:4} \middle| \theta\right)\pi\left(\theta\right) \propto \theta^{z_2+x_4}\left(1-\theta\right)^{x_2+x_3} \\ &= \mathcal{B}eta\left(\theta; z_2+x_4+1, x_2+x_3+1\right).\end{aligned}$$

- We have

$$\pi\left(z_{1:2} \middle| x_{1:4}, \theta\right) = \mathcal{M}\left(z_{1:2}; x_1, \frac{\frac{1}{2}}{\frac{1}{2}+\frac{\theta}{4}}, \frac{\frac{\theta}{4}}{\frac{1}{2}+\frac{\theta}{4}}\right)$$

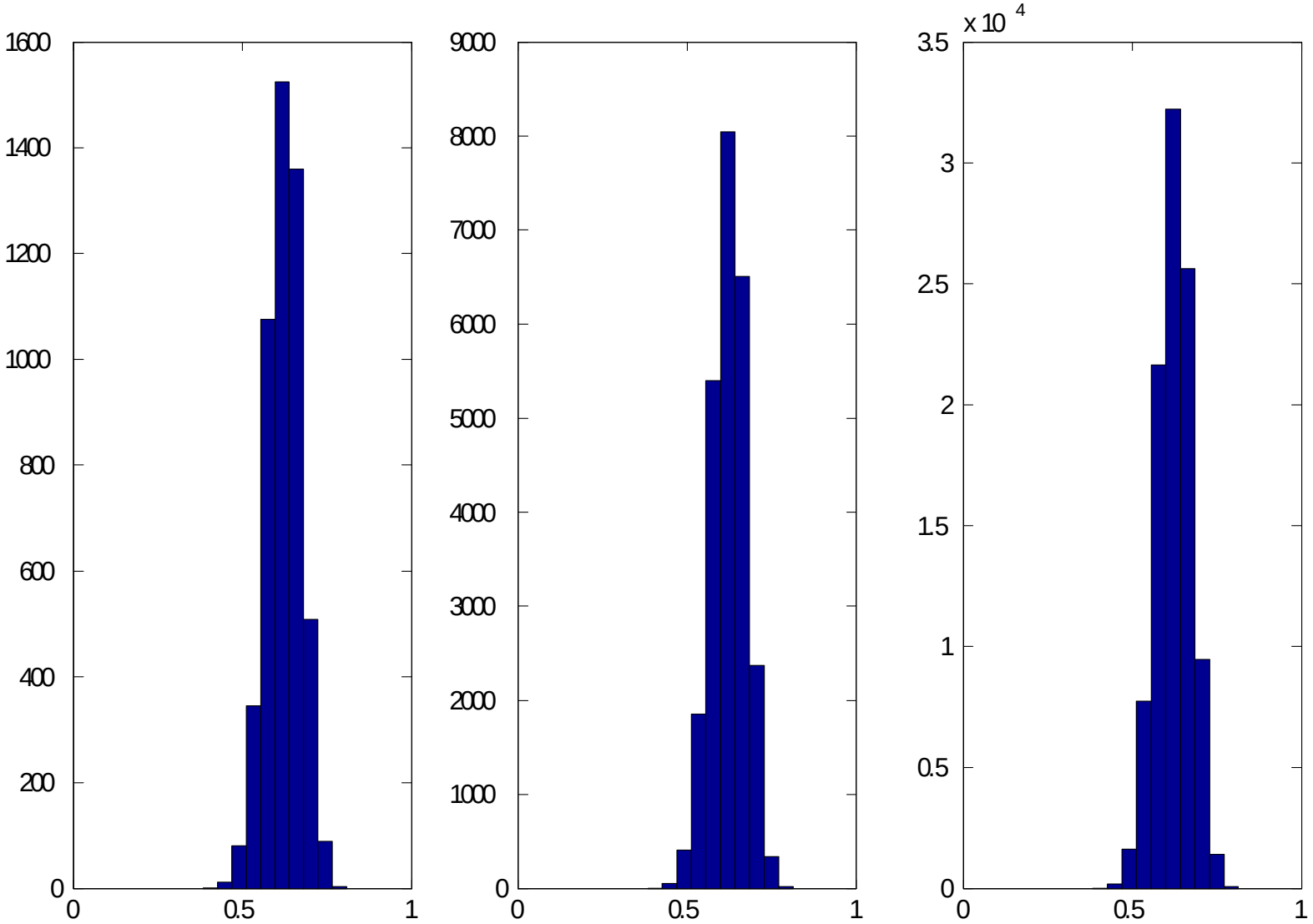



Figure: Estimation of $\pi(\theta|x_{1:4})$ using the Gibbs sampler for $N = 5000$ (left), 25000 (center) and 100000 (right) with $N_0 = 1000$.

- It is beyond the scope of this course to establish convergence results for the Gibbs sampler.
- **Proposition**: If $\left(\theta_1^{(i-1)}, \theta_2^{(i-1)}\right) \sim \pi\left(\theta_1, \theta_2\right)$ then $\left(\theta_1^{(i)}, \theta_2^{(i)}\right) \sim \pi\left(\theta_1, \theta_2\right)$.
- *Proof.* The joint distribution of $\left(\left(\theta_1^{(i-1)}, \theta_2^{(i-1)}\right), \left(\theta_1^{(i)}, \theta_2^{(i)}\right)\right)$ is given by

$$\pi\left(\theta_1^{(i-1)}, \theta_2^{(i-1)}\right) \pi\left(\theta_1^{(i)} \middle| \theta_2^{(i-1)}\right) \pi\left(\theta_2^{(i)} \middle| \theta_1^{(i)}\right)$$

so

$$\begin{aligned}
&\int \int \pi\left(\theta_1^{(i-1)}, \theta_2^{(i-1)}\right) \pi\left(\theta_1^{(i)} \middle| \theta_2^{(i-1)}\right) \pi\left(\theta_2^{(i)} \middle| \theta_1^{(i)}\right) d\theta_1^{(i-1)} d\theta_2^{(i-1)} \\
= &\int \pi\left(\theta_2^{(i-1)}\right) \pi\left(\theta_1^{(i)} \middle| \theta_2^{(i-1)}\right) \pi\left(\theta_2^{(i)} \middle| \theta_1^{(i)}\right) d\theta_2^{(i-1)} \\
= &\int \pi\left(\theta_1^{(i)}, \theta_2^{(i-1)}\right) \pi\left(\theta_2^{(i)} \middle| \theta_1^{(i)}\right) d\theta_2^{(i-1)} \\
= &\pi\left(\theta_1^{(i)}\right) \pi\left(\theta_2^{(i)} \middle| \theta_1^{(i)}\right) = \pi\left(\theta_1^{(i)}, \theta_2^{(i)}\right).
\end{aligned}$$

# Gibbs Sampler

- We can extend straightforwardly the Gibbs sampler to the case where $\theta = (\theta_1, ..., \theta_p)$. To sample from $\pi(\theta)$, the Gibbs sampler generates a Markov chain $\left(\theta_1^{(i)}, \theta_2^{(i)}, ..., \theta_p^{(i)}\right)$ as follows.
- Initialization: Set $\left(\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_p^{(0)}\right)$ deterministically or randomly.
- Iteration $i$; $i \geq 1$.
  - For $k = 1$ to $p$

    Sample $\theta_k^{(i)} \sim \pi\left(\theta_k \middle| \theta_{-k}^{(i)}\right)$ where

    $\theta_{-k}^{(i)} = \left(\theta_1^{(i)}, ..., \theta_{k-1}^{(i)}, \theta_{k+1}^{(i-1)}, ..., \theta_p^{(i-1)}\right)$.

# Bayesian Hierarchical Models

- Hierarchical models allow us to combine information and to combine the results of several studies addressing a set of related research hypotheses (meta-analysis).
- Consider a set of experiments/studies, $j = 1, ..., J$, in which experiment $j$ has data (vector) $y_j$ and parameter (vector) $\theta_j$, with likelihood $p(y_j|\theta_j)$. [the method applies equally well to nonexperimental data].
- If no observation -other than the data $\{y_j\}$- is available to distinguish any of the $\theta_j$'s from any of the others, and no ordering/grouping of the parameters can be made, one must assume symmetry among the parameters in their prior; i.e. the parameters $(\theta_1, \theta_2, ..., \theta_J)$ should be exchangeable. This means that $p(\theta_1, \theta_2, ..., \theta_J)$ is invariant to permutation of the indexes $(1, 2, ..., J)$.

# Exchangeability

- The simplest form of exchangeability is

$$p\left(\theta_1, \theta_2, ..., \theta_J | \phi\right) = \prod_{j=1}^{J} p\left(\theta_j | \phi\right).$$

- In general $\phi$ is unknown and we consider it random so that

$$p\left(\theta_1, \theta_2, ..., \theta_J\right) = \int \left( \prod_{j=1}^{J} p\left(\theta_j | \phi\right) \right) p\left(\phi\right) d\phi.$$

- As a consequence, we have

$$p\left(\theta_j | y_1, y_2, ..., y_J\right) \neq p\left(\theta_j | y_j\right).$$

- This model allows us to borrow information from different but related datasets.

# Example: Study of the effectiveness of cardiac treatments

- Assume we have $J$ hospitals.
- Let $y_j$ be data corresponding say to the number of patients having survived a cardiac treatment.
- Let $\theta_j$ be the survival probability for patients in hospital $j$.
- For example, we could have a model like

$$Y_j | (n_j, \theta_j) \sim \mathcal{B}in (n_j, \theta_j)$$

and

$$\theta_j \sim \mathcal{B}eta (\alpha, \beta) .$$

- This model reflects the fact that the $\theta_j$ are different but related to each other.
- Further, we put a prior distribution on the hyperparameters $(\alpha, \beta)$.

# Example: Baseball data

- We have the statistics of $J = 17$ players in pre-season exhibition matches.
- The data $y_j$ for the player $j$ corresponds to the number of home runs in $n_j$ times at the bat modelled through

$$Y_j | (n_j, p_j) \sim \mathcal{B}in (n_j, p_j) .$$

-

| $j$ | McGw. | Sosa | Griffey | Castilla | Gonz. | Gala. | Palm. | Vaughn |
|---|---|---|---|---|---|---|---|---|
| $y_j$ | 7 | 9 | 4 | 7 | 3 | 6 | 2 | 10 |
| $n_j$ | 58 | 59 | 74 | 84 | 69 | 63 | 60 | 54 |

| $j$ | Bond | Bag. | Piaz. | Thom. | Thom. | Mart. | Wal. | Burks | Buhner |
|---|---|---|---|---|---|---|---|---|---|
| $y_j$ | 2 | 2 | 4 | 3 | 2 | 5 | 3 | 2 | 6 |
| $n_j$ | 53 | 60 | 66 | 66 | 72 | 64 | 42 | 38 | 58 |

- Following Efron & Morris (JASA, 1975), we define

$$X_j = f_{n_j}\left(Y_j / n_j\right)$$

where

$$f_a\left(u\right) = a^{1/2} \sin^{-1}\left(2u - 1\right).$$

- This is a variance stabilising transformation. It can be shown that we now have approximately

$$X_j | \, \theta_j \sim \mathcal{N}\left(\theta_j, 1\right)$$

where

$$\theta_j = f_{n_j}\left(p_j\right).$$

- *Remark*: We only use this transformation to compare our results to the James-Stein estimate discussed later during the lectures.

# Bayesian Model for Baseball data

- We set an exchangeable prior distribution of the form

$$\theta_j | \left(\mu, \tau^2\right) \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mu, \tau^2\right)$$

and

$$\pi\left(\mu, \tau^2\right) = \pi\left(\mu\right) \pi\left(\tau^2\right)$$

with

$$\pi\left(\tau^2\right) = \mathcal{IG}\left(\tau^2; \frac{a}{2}, \frac{b}{2}\right), \ \pi\left(\mu\right) \propto 1.$$

where $a = b = 0.001$.

- Note that if you are a specialist of baseball, an exchangeable prior might not be appropriate.

- The full posterior distribution is given by

$$\begin{aligned}
& \pi\left(\mu, \tau^2, \theta_{1:J} \middle| x_{1:J}\right) \\
\propto \quad & \pi\left(\mu, \tau^2\right) \prod_{j=1}^{J} \pi\left(\theta_j \middle| \mu, \tau^2\right) \prod_{j=1}^{J} f\left(x_j \middle| \theta_j\right) \\
\propto \quad & \frac{1}{\tau^{J+a+2}} \exp\left(-\frac{b}{2\tau^2} - \sum_{j=1}^{J}\left(\frac{\left(\theta_j-\mu\right)^2}{2\tau^2} + \frac{\left(x_j-\theta_j\right)^2}{2}\right)\right)
\end{aligned}$$

- This distribution does not admit a closed-form expression and we are going to use the Gibbs sampler by decomposing the parameter space in 3 blocks $\mu$, $\tau^2$ and $\theta_{1:J}$.
- The Gibbs sampler will require being able to be able to sample from $\pi\left(\mu \middle| x_{1:J}, \tau^2, \theta_{1:J}\right)$, $\pi\left(\tau^2 \middle| x_{1:J}, \mu, \theta_{1:J}\right)$ and $\pi\left(\theta_{1:J} \middle| x_{1:J}, \mu, \tau^2\right)$.

# Gibbs Sampling for Baseball Data

- Full conditional distribution for $\mu$

$$\begin{aligned}\pi\left(\mu \middle| x_{1:J}, \tau^2, \theta_{1:J}\right) &\propto \exp\left(-\sum_{j=1}^{J} \frac{\left(\theta_j - \mu\right)^2}{2\tau^2}\right) \\ &= \mathcal{N}\left(\mu; J^{-1}\sum_{j=1}^{J}\theta_j, J^{-1}\tau^2\right).\end{aligned}$$

- Full conditional distribution for $\tau^2$

$$\begin{aligned}\pi\left(\tau^2 \middle| x_{1:J}, \mu, \theta_{1:J}\right) &\propto \frac{1}{\tau^{J+a+2}} \exp\left(-\frac{b}{2\tau^2} - \sum_{j=1}^{J} \frac{\left(\theta_j - \mu\right)^2}{2\tau^2}\right) \\ &= \mathcal{IG}\left(\tau^2; \frac{J+a}{2}, \frac{b + \sum_{j=1}^{J}\left(\theta_j - \mu\right)^2}{2}\right).\end{aligned}$$

- Full conditional distribution for $\pi\left(\theta_{1:J}|\, x_{1:J}, \mu, \tau^2\right)$

$$\pi\left(\theta_{1:J}|\, x_{1:J}, \mu, \tau^2\right) = \prod_{j=1}^{J} \pi\left(\theta_j|\, x_j, \mu, \tau^2\right)$$

where

$$\begin{aligned}\pi\left(\theta_j|\, x_j, \mu, \tau^2\right) &\propto \exp\left(-\left(\frac{(\theta_j-\mu)^2}{2\tau^2} + \frac{(x_j-\theta_j)^2}{2}\right)\right) \\ &= \mathcal{N}\left(\theta_j; \frac{\mu+\tau^2 x_j}{1+\tau^2}, \frac{\tau^2}{1+\tau^2}\right).\end{aligned}$$

- These three distributions can be sampled using standard procedures and the simulated samples $\left(\mu^{(i)}, \tau^{2(i)}, \theta_{1:J}^{(i)}\right)$ are (asymptotically) distributed from the posterior.
- We applied the Gibbs sampling using $N = 10000$ iterations and compared it to a simpler empirical Bayesian analysis.

# Empirical Bayesian Analysis for Baseball Data

- In an empirical Bayesian analysis of the data, we just obtain a point estimate $\left(\widehat{\mu}, \widehat{\tau}^2\right)$ of $\left(\mu, \tau^2\right)$ using the data.
- For example, given that

$$\begin{aligned} f\left(x_j \middle| \mu, \tau^2\right) &= \int f\left(x_j \middle| \theta_j\right) \pi\left(\theta_j \middle| \mu, \tau^2\right) d\theta_j \\ &= \mathcal{N}\left(x_j; \mu, 1+\tau^2\right) \end{aligned}$$

  when can select

$$\widehat{\mu} = \frac{1}{J}\sum_{j=1}^{J} x_j, \ \widehat{\tau^2} = \frac{1}{J}\sum_{j=1}^{J}\left(x_j - \widehat{\mu}\right)^2 - 1$$

- This is *not* a Bayesian calculation because it is not based on any specified full probability model. The selection of the point estimate is somewhat arbitrary (although principled) and no uncertainty about $\mu, \tau^2$ is taken into account.

- Moreover, the distribution

$$\pi\left(\theta_j \middle| x_j, \widehat{\mu}, \widehat{\tau^2}\right) = \mathcal{N}\left(\theta_j; \frac{\widehat{\mu} + \widehat{\tau^2} x_j}{1 + \widehat{\tau^2}}, \frac{\widehat{\tau^2}}{1 + \widehat{\tau^2}}\right)$$

  are necessarily Gaussian and we somehow use the data twice.
- The main advantage of the empirical Bayes approach is that it is very easy to implement compared to the 'full' Bayesian approach which relies no MCMC.

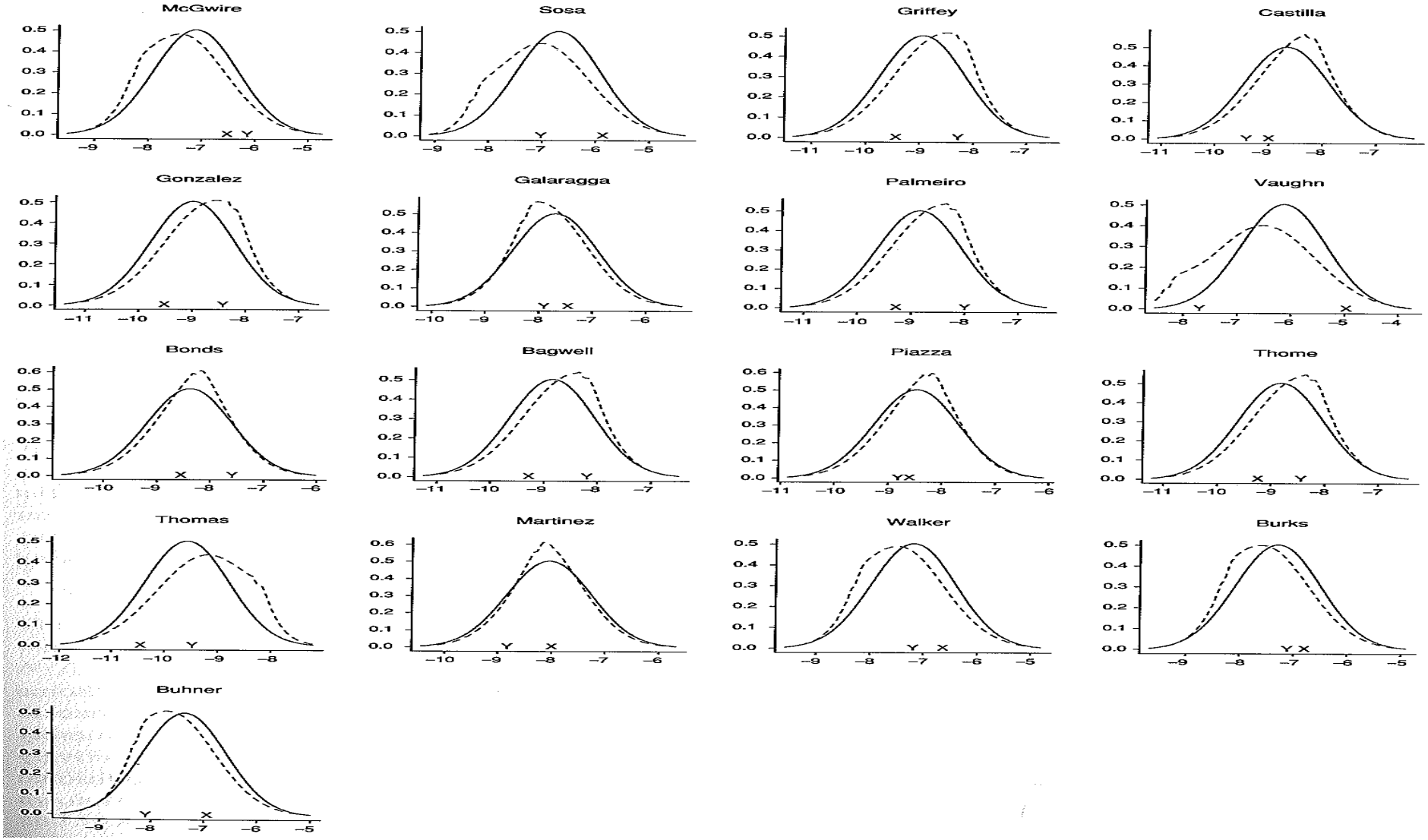



Figure: Posterior distributions $\pi\left(\theta_j \middle| x_{1:J}\right)$ (dashed line) estimated using MCMC and $\pi\left(\theta_j \middle| x_j, \widehat{\mu}, \widehat{\tau^2}\right)$ (solid line) estimated using Empirical Bayes.

- The estimate of $\pi\left(\theta_j \middle| x_j\right)$ was not obtained by smoothing the histogram of the simulated values $\theta_j^{(i)}$ but we use the fact that

$$\pi\left(\theta_j \middle| x_{1:J}\right) = \int \pi\left(\theta_j \middle| x_j, \mu, \tau^2\right) \pi\left(\mu, \tau^2 \middle| x_{1:J}\right) d\mu d\tau^2$$

- So using the Monte Carlo approximation of $\pi\left(\mu, \tau^2 \middle| x_{1:J}\right)$

$$\widehat{\pi}\left(\mu, \tau^2 \middle| x_{1:J}\right) = \frac{1}{N - N_0 + 1} \sum_{i=N_0}^{N} \delta_{\mu^{(i)}, \tau^{2(i)}}\left(\mu, \tau^2\right)$$

we obtain

$$\begin{aligned} \widehat{\pi}\left(\theta_j \middle| x_{1:J}\right) &= \frac{1}{N - N_0 + 1} \sum_{i=N_0}^{N} \pi\left(\theta_j \middle| x_j, \mu^{(i)}, \tau^{2(i)}\right) \\ &= \frac{1}{N - N_0 + 1} \sum_{i=N_0}^{N} \mathcal{N}\left(\theta_j; \frac{\mu + \tau^{2(i)} x_j}{1 + \tau^{2(i)}}, \frac{\tau^{2(i)}}{1 + \tau^{2(i)}}\right). \end{aligned}$$

- This is a so-called Rao-Blackwellised estimate.

# Example: Nuclear Pump Data

- Multiple failures in a nuclear plant

| Pump $j$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| # Failures $p_j$ | 5 | 1 | 5 | 14 | 3 |
| Times $t_j$ | 94.32 | 15.72 | 62.88 | 125.76 | 5.24 |

| Pump $j$ | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|
| # Failures $p_j$ | 19 | 1 | 1 | 4 | 22 |
| Times $t_j$ | 31.44 | 1.05 | 1.05 | 2.10 | 10.48 |

- Model: Failures of the $j-$th pump follow a Poisson process with parameter $\lambda_j$ $(1 \leq j \leq 10)$. For an observed time $t_j$, the number of failures $p_j$ is thus a Poisson $\mathcal{P}(\lambda_j t_j)$ random variable.
- The unknown parameters consist of $\theta = (\lambda_1, \ldots, \lambda_{10}, \beta)$.

- Hierarchical model

$$\lambda_j | (\alpha, \beta) \overset{\text{iid}}{\sim} \mathcal{G}(\alpha, \beta) \text{ and } \beta \sim \mathcal{G}(\gamma, \delta)$$

  with $\alpha = 1.8$ and $\gamma = 0.01$ and $\delta = 1$.

- The posterior distribution is proportional to

$$\pi(\lambda_{1:10}, \beta | p_{1:10}, t_{1:10}) \propto \pi(\beta) \prod_{j=1}^{10} \pi(\lambda_j | \beta) \prod_{j=1}^{10} f(p_j | \lambda_j, t_j)$$

$$\propto \quad \beta^{\gamma-1} \exp(-\delta\beta) \prod_{j=1}^{10} \beta^{\alpha} \lambda_j^{\alpha-1} \exp(-\beta\lambda_j) \prod_{j=1}^{10} (\lambda_j t_j)^{p_j} \exp(-\lambda_j t_j)$$

- This multidimensional distribution is rather complex. It is not obvious how the rejection method or importance sampling could be used in this context.

- The conditionals have a familiar form

$$\pi\left(\lambda_{1:10}\middle|\, p_{1:10}, t_{1:10}, \beta\right) = \prod_{j=1}^{10} \pi\left(\lambda_j\middle|\, p_j, t_j, \beta\right)$$

where

$$\begin{aligned} \pi\left(\lambda_j\middle|\, p_j, t_j, \beta\right) &\propto \lambda_j^{p_j+\alpha-1} \exp\{-(t_j+\beta)\lambda_j)\} \\ &= \mathcal{G}(\lambda_j; p_j+\alpha, t_j+\beta) \end{aligned}$$

- For the hyperparameter

$$\begin{aligned} \pi\left(\beta\middle|\, p_{1:10}, t_{1:10}, \lambda_{1:10}\right) &\propto \beta^{10\alpha+\gamma-1} \exp\left(-\left(\delta + \sum_{j=1}^{10} \lambda_j\right)\beta\right) \\ &= \mathcal{G}(\beta; \gamma+10\alpha, \delta+\sum_{j=1}^{10}\lambda_j). \end{aligned}$$

- Gibbs sampling is once more easily feasible in such cases.

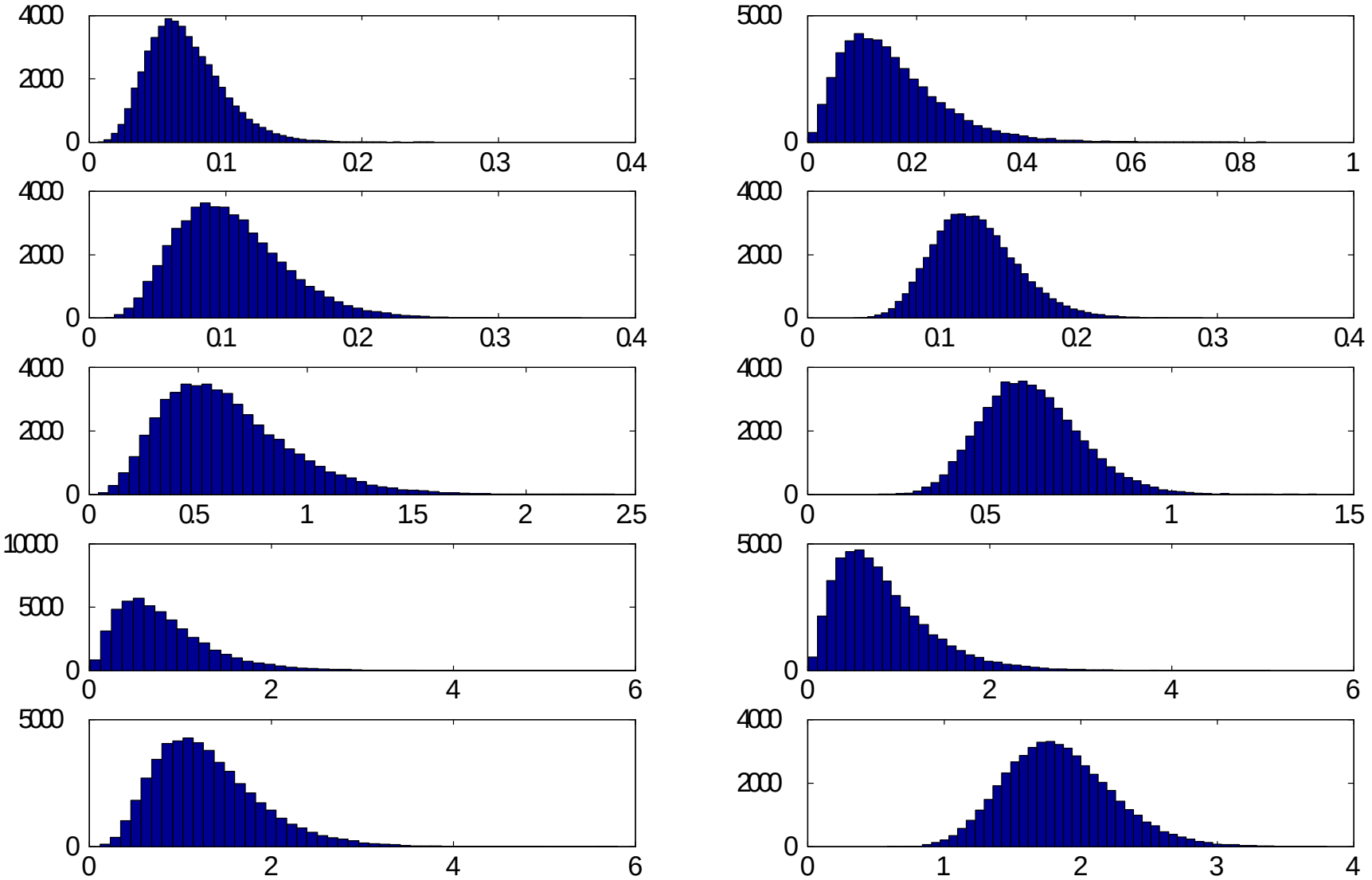



Figure: Histogram approximations of $\pi\left(\lambda_j \middle| t_{1:10}, p_{1:10}\right)$ obtained using $N = 50,000$ and $N_0 = 1000$.

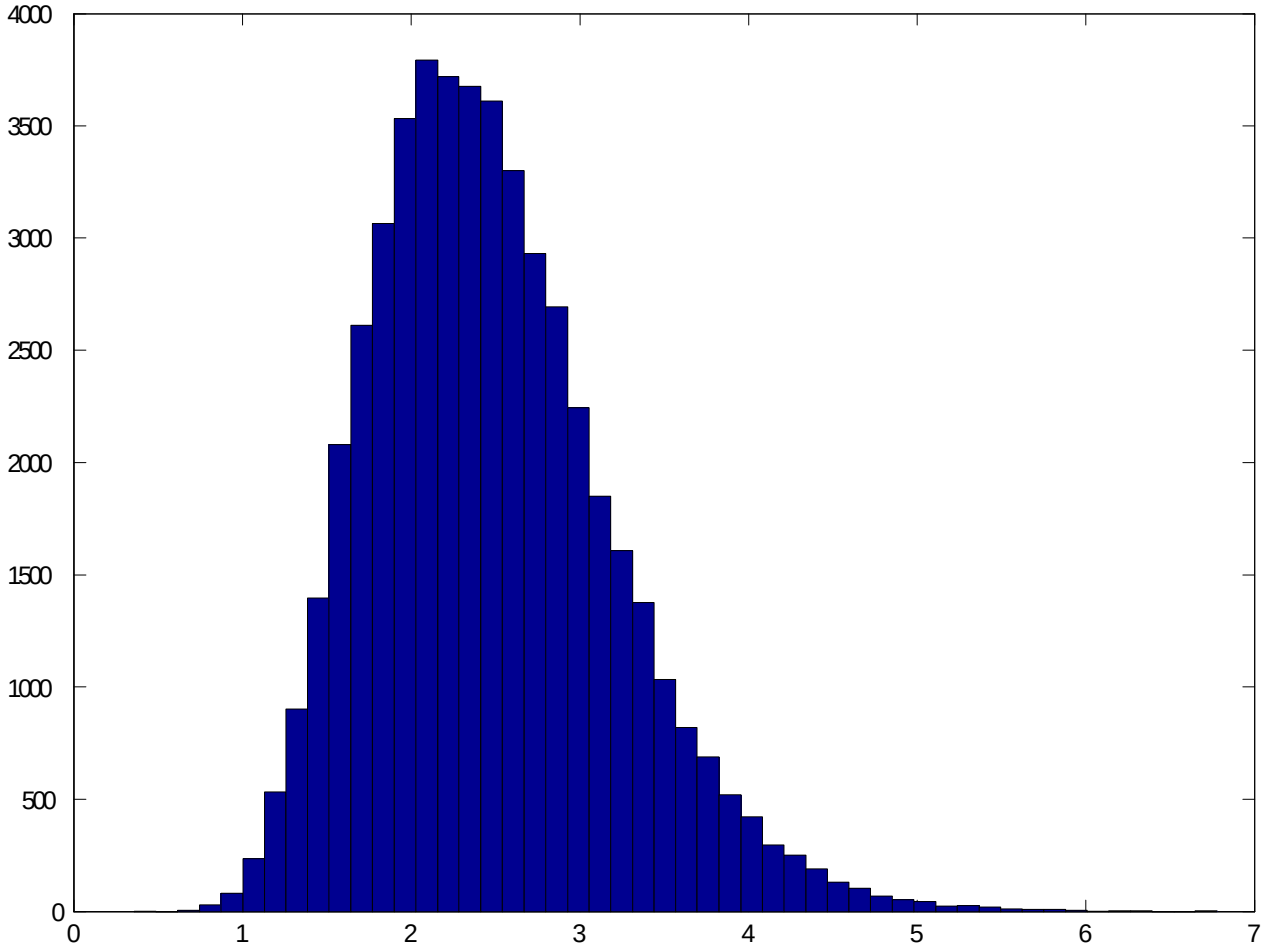



Figure: Histogram approximation of $\pi\left(\beta \middle| p_{1:10}, t_{1:10}\right)$.

- The posterior distribution of the hyperparameter is quite diffuse. Hence the results we obtained using a full Bayesian approach are significantly different from an empirical Bayes approach.

# Limitation of Gibbs Sampling

- The Gibbs sampler requires sampling from the full conditional distributions
$$\pi\left(\theta_k \middle| \theta_{-k}\right).$$
- For many complex models, it is impossible to sample from several of these "full" conditional distributions.
- Even if it is possible to implement the Gibbs sampler, the algorithm might be very inefficient because the variables are very correlated or sampling from the full conditionals is extremely expensive/inefficient.

# Metropolis-Hastings Algorithm

- The Metropolis-Hastings algorithm is an alternative algorithm to sample from any probability distribution $\pi(\theta)$ known up to a normalizing constant.
- This can be interpreted as the basis of all MCMC algorithms: It provides a generic way to build a Markov kernel admitting $\pi(\theta)$ as an invariant distribution.
- The Metropolis algorithm was named the "Top algorithm of the 20th century" by computer scientists, mathematicians, physicists.

- Introduce a proposal distribution/kernel $q\left(\theta' \middle| \theta\right)$, i.e.

$$\int q\left(\theta' \middle| \theta\right) d\theta' = 1 \text{ for any } \theta.$$

- The basic idea of the MH algorithm is to propose a new candidate $\theta'$ based on the current state of the Markov chain $\theta$.
- We only accept this algorithm with respect to a probability $\alpha\left(\theta, \theta'\right)$ which ensures that the invariant distribution of the transition kernel is the target distribution $\pi\left(\theta\right)$.

- Initialization: Select deterministically or randomly $\theta^{(0)}$.
- Iteration $i$; $i \geq 1$:
  - Sample $\theta^* \sim q\left(\theta | \, \theta^{(i-1)}\right)$ and compute

$$\alpha\left(\theta^{(i-1)}, \theta^*\right) = \min\left(1, \frac{\pi\left(\theta^*\right) q\left(\theta^{(i-1)} \middle| \theta^*\right)}{\pi\left(\theta^{(i-1)}\right) q\left(\theta^* | \, \theta^{(i-1)}\right)}\right).$$

  - With probability $\alpha\left(\theta^{(i-1)}, \theta^*\right)$, set $\theta^{(i)} = \theta^*$; otherwise set $\theta^{(i)} = \theta^{(i-1)}$.
- Simulated annealing is an extremely popular optimization algorithm: it is a simple nonhomogeneous version of MH where at iteration $i$ the target distribution is $\pi_i(\theta) \propto [\pi(\theta)]^{\gamma_i}$ where $\gamma_i$ is an increasing sequence going to $\infty$.

- It is not necessary to know the normalizing constant of $\pi(\theta)$ to implement the algorithm.
- This algorithm is extremely general: $q\left(\theta' \middle| \theta\right)$ can be any proposal distribution. So in practice, we can select it so that it is easy to sample from it.
- There is much more freedom than in the Gibbs sampler where the proposal distributions are fixed once we have partitioned the vector parameter.

# Random Walk Metropolis

- The original Metropolis algorithm corresponds to the following choice for $q\left(\theta' \middle| \theta\right)$

$$\theta' = \theta + Z \text{ where } Z \sim g;$$

  i.e. this is a so-called *random walk proposal*.

- The distribution $g\left(z\right)$ is the distribution of the random walk increments $Z$ and

$$q\left(\theta' \middle| \theta\right) = g\left(\theta' - \theta\right) \quad \Rightarrow \quad \alpha\left(\theta, \theta'\right) = \min\left(1, \frac{\pi\left(\theta'\right) g\left(\theta - \theta'\right)}{\pi\left(\theta\right) g\left(\theta' - \theta\right)}\right).$$

- If $g\left(z\right) = g\left(-z\right)$ - e.g. $Z \sim \mathcal{N}\left(0, \Sigma\right)$- then

$$\alpha\left(\theta, \theta'\right) = \min\left(1, \frac{\pi\left(\theta'\right)}{\pi\left(\theta\right)}\right).$$

- There is no clear guideline how to select the proposal distribution.
- When the variance of the random walk increments (if it exists) is very small then the acceptance rate can be expected to be around 0.5-0.7.
- You would like to scale the random walk moves such that it is possible to move reasonably fast in regions of positive probability masses under $\pi$.

- In the Bayesian framework where $\pi(\theta) = p(\theta|x) \propto f(x|\theta)\, p(\theta)$ and $q\left(\theta' \middle| \theta\right) = q\left(\theta \middle| \theta'\right) = g\left(\theta' - \theta\right)$ then

$$\alpha\left(\theta, \theta'\right) = \min\left(1, \frac{f\left(x \middle| \theta'\right) p\left(\theta'\right)}{f\left(x \middle| \theta\right) p\left(\theta\right)}\right).$$

- Assuming $g(z) = \mathcal{N}(z; 0, \Sigma)$ then the selection of $\Sigma$ will be difficult.
- When the Fisher/observed information matrix at $\theta$ is available, then we typically select $\Sigma$ as the inverse of it.

# Toy Example

- Consider the case where

$$\pi(\theta) \propto \exp\left(-\frac{\theta^2}{2}\right).$$

- We implement the MH algorithm for

$$\begin{aligned} q_1\left(\theta' \middle| \theta\right) &\propto \exp\left(-\frac{\left(\theta'-\theta\right)^2}{2\left(0.2\right)^2}\right), \\ q_2\left(\theta' \middle| \theta\right) &\propto \exp\left(-\frac{\left(\theta'-\theta\right)^2}{2\left(5\right)^2}\right), \\ q_3\left(\theta' \middle| \theta\right) &\propto \exp\left(-\frac{\left(\theta'-\theta\right)^2}{2\left(0.02\right)^2}\right). \end{aligned}$$

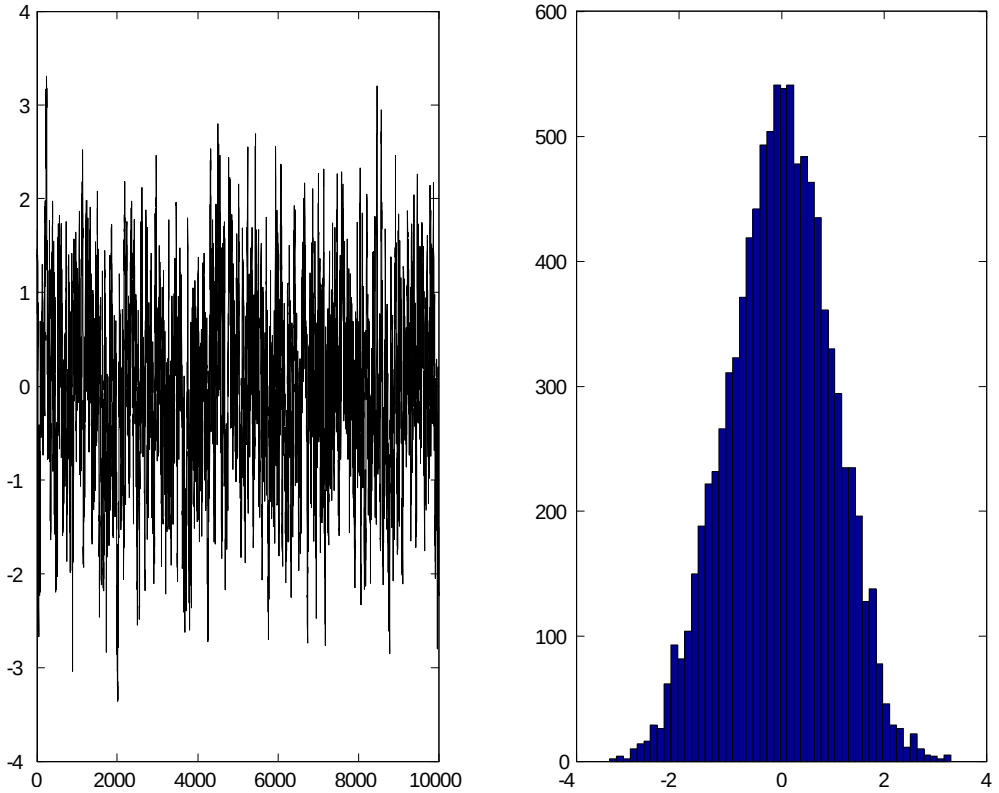



Figure: MCMC output for $q_1$, we estimate $\mathbb{E}(\theta) = -0.02$ and $\mathbb{V}(\theta) = 0.99$

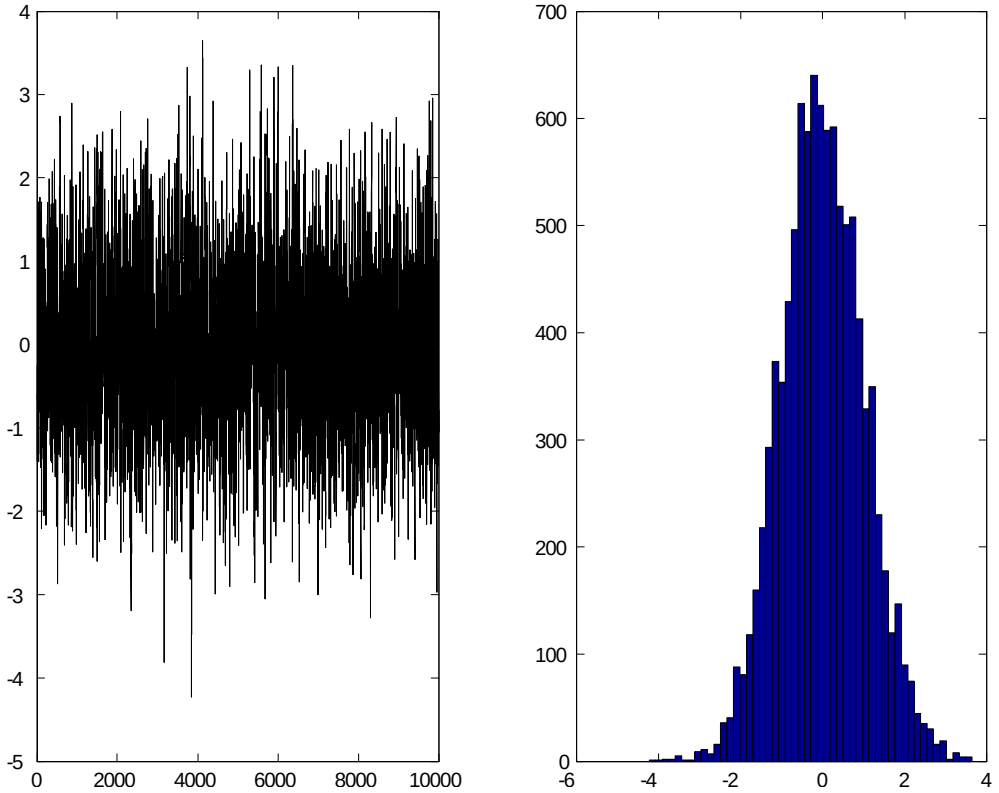



Figure: MCMC output for $q_2$, we estimate $\mathbb{E}(\theta) = 0.00$ and $\mathbb{V}(\theta) = 1.02$.

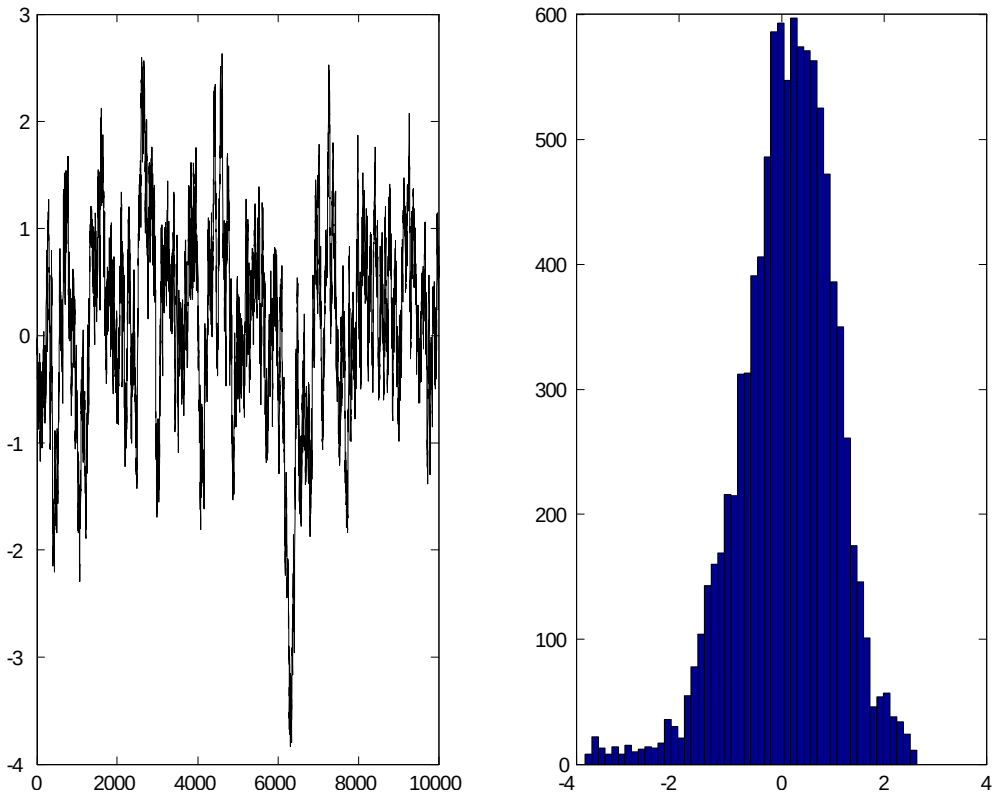



Figure: MCMC output for $q_3$, we estimate $\mathbb{E}(\theta) = 0.10$ and $\mathbb{V}(\theta) = 0.92$.

# Independent Metropolis-Hastings

- The Hastings' generalization corresponds to the following choice for $q\left(\theta' \middle| \theta\right)$

$$q\left(\theta' \middle| \theta\right) = q\left(\theta'\right);$$

  i.e. this is a so-called *independent proposal*.
- In this case, the acceptance probability is given by

$$\alpha\left(\theta, \theta'\right) = \min\left(1, \frac{\pi\left(\theta'\right) q\left(\theta\right)}{q\left(\theta'\right) \pi\left(\theta\right)}\right).$$

- The ratio $\pi\left(\theta\right) / q\left(\theta\right)$ also appears the Accept/Reject method.
- The optimal independent proposal is clearly $q\left(\theta\right) = \pi\left(\theta\right)$!

- In the Bayesian framework where $\pi(\theta) = p(\theta|x) \propto f(x|\theta)\, p(\theta)$ and $q(\theta) = p(\theta)$ then

$$\alpha(\theta, \theta') = \min\left(1, \frac{f(x|\theta')}{f(x|\theta)}\right).$$

- Like the accept/reject method, this approach will be inefficient if the prior and the posterior are very different from each other.
- The MH is very flexible and we could use for example

$$q(\theta) = \mathcal{N}\left(\theta; \widehat{\theta}_{MLE}, \sigma^2\right)$$

  or a distribution with thicker tails.
- Any heuristic can be made rigourous using the MH algorithm.

# Toy example

- **Example**: Consider the case where

$$\pi(\theta) \propto \exp\left(-\frac{\theta^2}{2}\right).$$

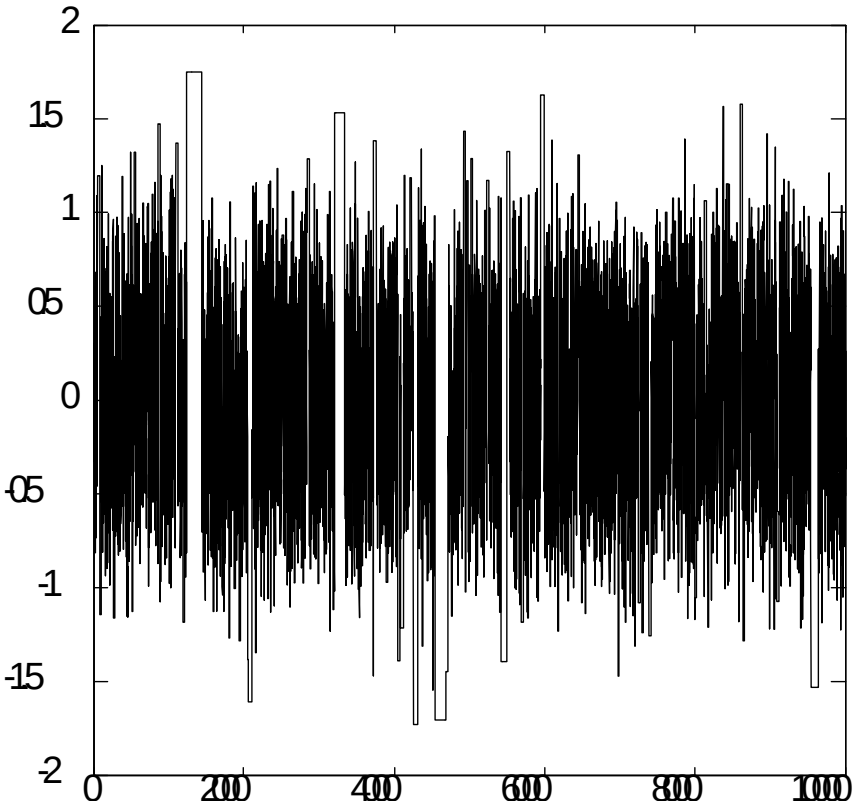
- We implement the MH algorithm for

$$q_1(\theta) \propto \exp\left(-\frac{\theta^2}{2(0.2)^2}\right)$$

so $\pi(\theta)/q_1(\theta) \to \infty$ as $\theta \to \infty$ and for

$$q_2(\theta) \propto \exp\left(-\frac{\theta^2}{2(5)^2}\right)$$

so $\pi(\theta)/q_2(\theta) \leq C < \infty$ for all $\theta$.



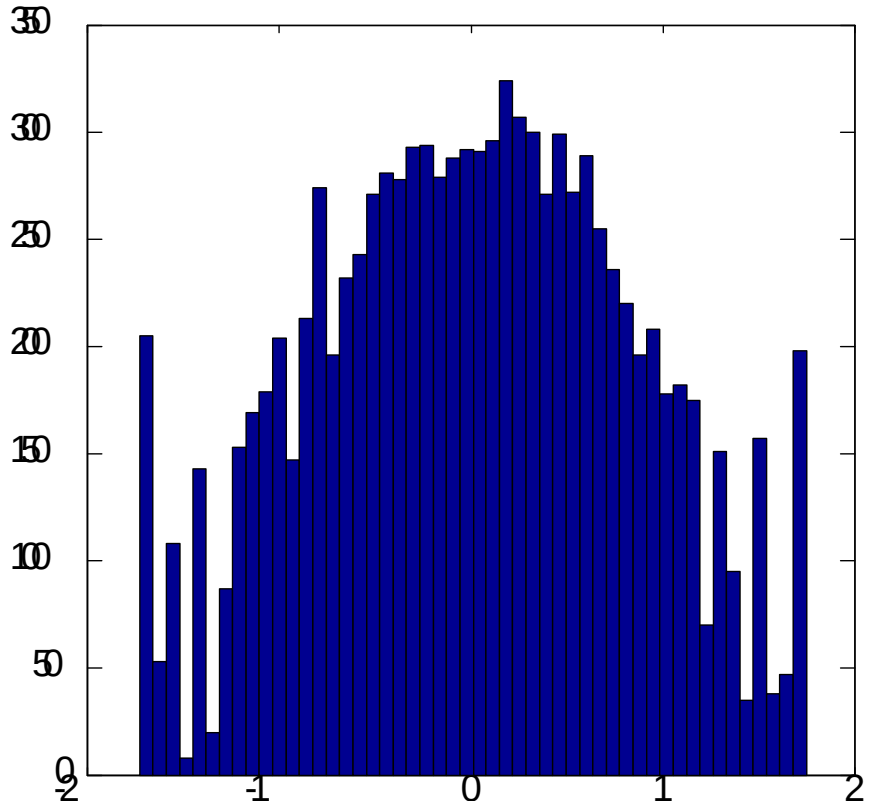
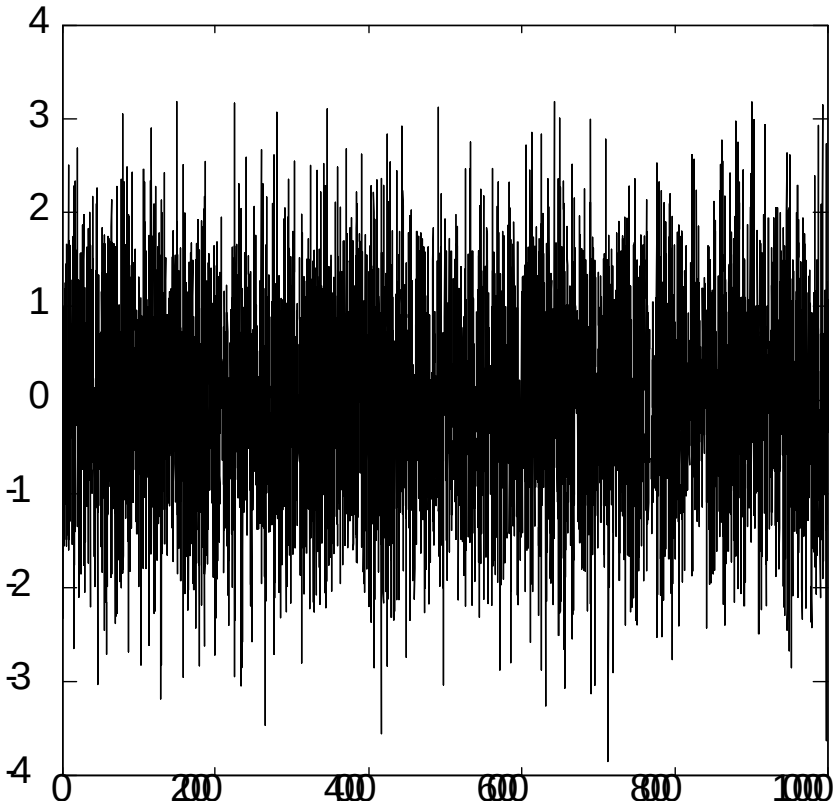




Figure: MCMC output for $q_1$, we estimate $\mathbb{E}(\theta) = 0.0206$ and $\mathbb{V}(\theta) = 0.83$.



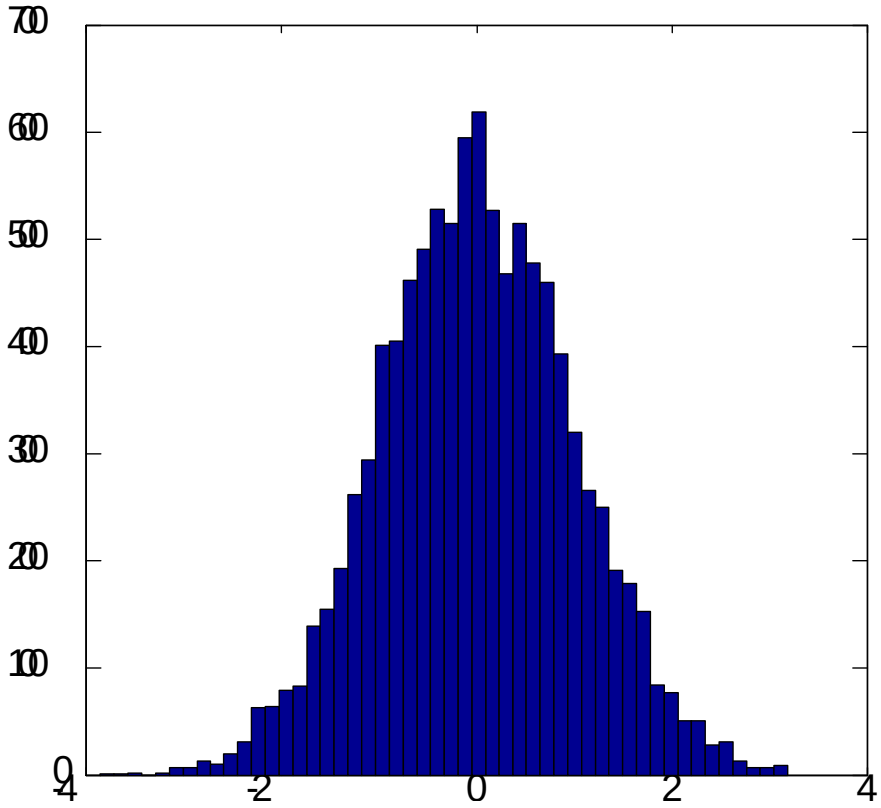




Figure: MCMC output for $q_2$, we estimate $\mathbb{E}(\theta) = -0.004$ and $\mathbb{V}(\theta) = 1.00$.

# Ensuring convergence

- To establish that the MH chain converges towards the required target, we need to establish that
  - If $\theta^{(i-1)} \sim \pi(\theta)$ then $\theta^{(i)} \sim \pi(\theta)$, i.e. $\pi$ is an invariant distribution of the Markov kernel associated to the MH algorithm.
  - The Markov chain is irreducible; i.e. we can reach any set $A$ such that $\pi(A) > 0$.
  - The Markov chain is aperiodic; i.e. we do not visit the state-space in a periodic way.

# Metropolis Hastings kernel

- The transition kernel associated to the MH algorithm can be rewritten as

$$K\left(\theta' \middle| \theta\right) = \alpha\left(\theta, \theta'\right) q\left(\theta' \middle| \theta\right) + \underbrace{\left(1 - \int \alpha\left(\theta, u\right) q\left(u \middle| \theta\right) du\right)}_{\text{rejection probability}} \delta_\theta\left(\theta'\right)$$

- Clearly we have

$$\begin{aligned} \int K\left(\theta' \middle| \theta\right) d\theta' &= \int \alpha\left(\theta, \theta'\right) q\left(\theta' \middle| \theta\right) d\theta' \\ &\quad + \left(1 - \int \alpha\left(\theta, u\right) q\left(u \middle| \theta\right) du\right) \int \delta_\theta\left(\theta'\right) d\theta' \\ &= 1. \end{aligned}$$

- **Proposition**: If $\theta^{(i-1)} \sim \pi(\theta)$ then $\theta^{(i)} \sim \pi(\theta)$.
- To prove the result, we are going to establish the stronger *reversibility property*: for any $\theta, \theta'$

$$\pi(\theta) K\left(\theta' \middle| \theta\right) = \pi\left(\theta'\right) K\left(\theta \middle| \theta'\right);$$

  i.e. the probability of being in $A$ and moving to $B$ is equal to the probability of being in $B$ and moving to $A$.
- Indeed the reversibility condition implies that

$$\begin{aligned} \int \pi(\theta) K\left(\theta' \middle| \theta\right) d\theta &= \int \pi\left(\theta'\right) K\left(\theta \middle| \theta'\right) d\theta \\ &= \pi\left(\theta'\right) \int K\left(\theta \middle| \theta'\right) d\theta \\ &= \pi\left(\theta'\right) \end{aligned}$$

- By definition of the kernel, we have

$$\begin{aligned}\pi(\theta) K\left(\theta' \middle| \theta\right) &= \pi(\theta) \alpha\left(\theta, \theta'\right) q\left(\theta' \middle| \theta\right) \\ &\quad + \pi(\theta)\left(1 - \int \alpha(\theta, u) q(u|\theta)\, du\right) \delta_\theta\left(\theta'\right).\end{aligned}$$

- Then

$$\begin{aligned}\pi(\theta) \alpha\left(\theta, \theta'\right) q\left(\theta' \middle| \theta\right) &= \pi(\theta) q\left(\theta' \middle| \theta\right) \min\left(1, \frac{\pi\left(\theta'\right) q\left(\theta \middle| \theta'\right)}{\pi(\theta) q\left(\theta' \middle| \theta\right)}\right) \\ &= \min\left(\pi(\theta) q\left(\theta' \middle| \theta\right), \pi\left(\theta'\right) q\left(\theta \middle| \theta'\right)\right) \\ &= \pi\left(\theta'\right) q\left(\theta \middle| \theta'\right) \min\left(1, \frac{\pi(\theta) q\left(\theta' \middle| \theta\right)}{\pi\left(\theta'\right) q\left(\theta \middle| \theta'\right)}\right) \\ &= \pi\left(\theta'\right) \alpha\left(\theta', \theta\right) q\left(\theta \middle| \theta'\right).\end{aligned}$$

- We have obviously

$$\begin{aligned} & \left(1-\int \alpha\left(\theta, u\right) q\left(u \middle| \theta\right) du\right) \delta_{\theta}\left(\theta'\right) \pi\left(\theta\right) \\ = & \left(1-\int \alpha\left(\theta', u\right) q\left(u \middle| \theta'\right) du\right) \delta_{\theta'}\left(\theta\right) \pi\left(\theta'\right). \end{aligned}$$

- It follows that

$$\pi\left(\theta\right) K\left(\theta' \middle| \theta\right) = \pi\left(\theta'\right) K\left(\theta \middle| \theta'\right).$$

- Hence, $\pi$ is the invariant distribution of the transition kernel $K$.

# Irreducibility and Aperiodicity

- To ensure irreducibility, a sufficient but not necessary condition is that

$$\pi\left(\theta'\right) > 0 \Rightarrow q\left(\theta' \middle| \theta\right) > 0 \text{ for any } \theta$$

- Aperiodicity is automatically ensured as there is always a strictly positive probability to reject the candidate.
- Theoretically, the MH algorithm converges under very weak assumptions to the target distribution $\pi$. In practice, this convergence can be so slow that the algorithm is useless.

# Application to Genetic Linkage Model

- In this example, we have

$$(X_1, X_2, X_3, X_4) \sim \mathcal{M}\left(n; \frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right)$$

  and the target posterior distribution is given by

$$\pi(\theta | x_{1:4}) \propto (2+\theta)^{x_1} (1-\theta)^{x_2+x_3} \theta^{x_4} 1_{(0,1)}(\theta).$$

- Accept/Reject requires maximizing the likelihood whereas Gibbs sampling requires introducing missing data.
- Alternatively, we can use a simple MH algorithm with proposal distribution $q(\theta' | \theta) = \mathcal{N}(\theta'; \theta, \sigma^2)$. See computer simulations.

# Application to Probit Regression

- We consider the following example: we take 4 measurements from 100 genuine Swiss banknotes and 100 counterfeit ones (Marin & Robert, *Bayesian Core*, Springer-Verlag, 2007).
- The response variable $y$ is 0 for *genuine* and 1 for *counterfeit* and the explanatory variables are
  - $x^1$ the length,
  - $x^2$: the width of the left edge
  - $x^3$: the width of the right edge
  - $x^4$: the bottom margin witdth

- As $Y \in \{0,1\}$, we cannot have a linear model

$$\begin{aligned} Y &= x^1\beta_1 + \ldots + x^4\beta_4 + \varepsilon \\ &= \mathbf{x}^{\mathsf{T}}\beta + \varepsilon \end{aligned}$$

where $\mathbf{x} = \begin{pmatrix} x^1 & x^2 & x^3 & x^4 \end{pmatrix}^{\mathsf{T}}$ and $\beta = \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 & \beta_4 \end{pmatrix}^{\mathsf{T}}$.

- We select a probit link here

$$\Pr\left(Y = 1 \middle| \beta, \mathbf{x}\right) = 1 - \Pr\left(Y = 0 \middle| \beta, \mathbf{x}\right) = \Phi\left(\mathbf{x}^{\mathsf{T}}\beta\right)$$

where

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp\left(-\frac{v^2}{2}\right) dv$$

- For $n = 200$ data, the likelihood is then given by

$$f\left(y_{1:n} \middle| \beta, \mathbf{x}_{1:n}\right) = \prod_{i=1}^{n} \Phi\left(\mathbf{x}_i^{\mathsf{T}}\beta\right)^{y_i} \left(1 - \Phi\left(\mathbf{x}_i^{\mathsf{T}}\beta\right)\right)^{1-y_i}.$$

- We assume a vague prior $p(\beta) = \mathcal{N}(\beta; 0, 100\mathrm{I}_4)$.
- We use a simple random walk sampler where $q(\beta' | \beta) = \mathcal{N}(\beta'; \beta, \tau^2\Sigma)$ with $\Sigma = \Omega^{-1}$, $\Omega_{i,j} = \left[ -\frac{\partial^2 \log f(y_{1:n}|\beta, \mathbf{x}_{1:n})}{\partial\beta_i \partial\beta_j} \right] \Big|_{\beta_{\mathrm{MLE}}}$.
- The algorithm is thus simply given at iteration $i$ by
  - Sample $\beta^* \sim \mathcal{N}\left(\beta^{(i-1)}, \tau^2\Sigma\right)$ and compute

$$\begin{aligned} \alpha\left(\beta^{(i-1)}, \beta^*\right) &= \min\left(1, \frac{p(\beta^* | y_{1:n}, \mathbf{x}_{1:n})}{p\left(\beta^{(i-1)} \middle| y_{1:n}, \mathbf{x}_{1:n}\right)}\right) \\ &= \min\left(1, \frac{f(y_{1:n} | \beta^*, \mathbf{x}_{1:n}) p(\beta^*)}{f\left(y_{1:n} | \beta^{(i-1)}, \mathbf{x}_{1:n}\right) p\left(\beta^{(i-1)}\right)}\right). \end{aligned}$$

  - Set $\beta^{(i)} = \beta^*$ with probability $\alpha\left(\beta^{(i-1)}, \beta^*\right)$ and $\beta^{(i)} = \beta^{(i-1)}$ otherwise.
- Algorithm tested with $\tau^2 = 1$, 10 and 0.1. Best results obtained with $\tau^2 = 1$.

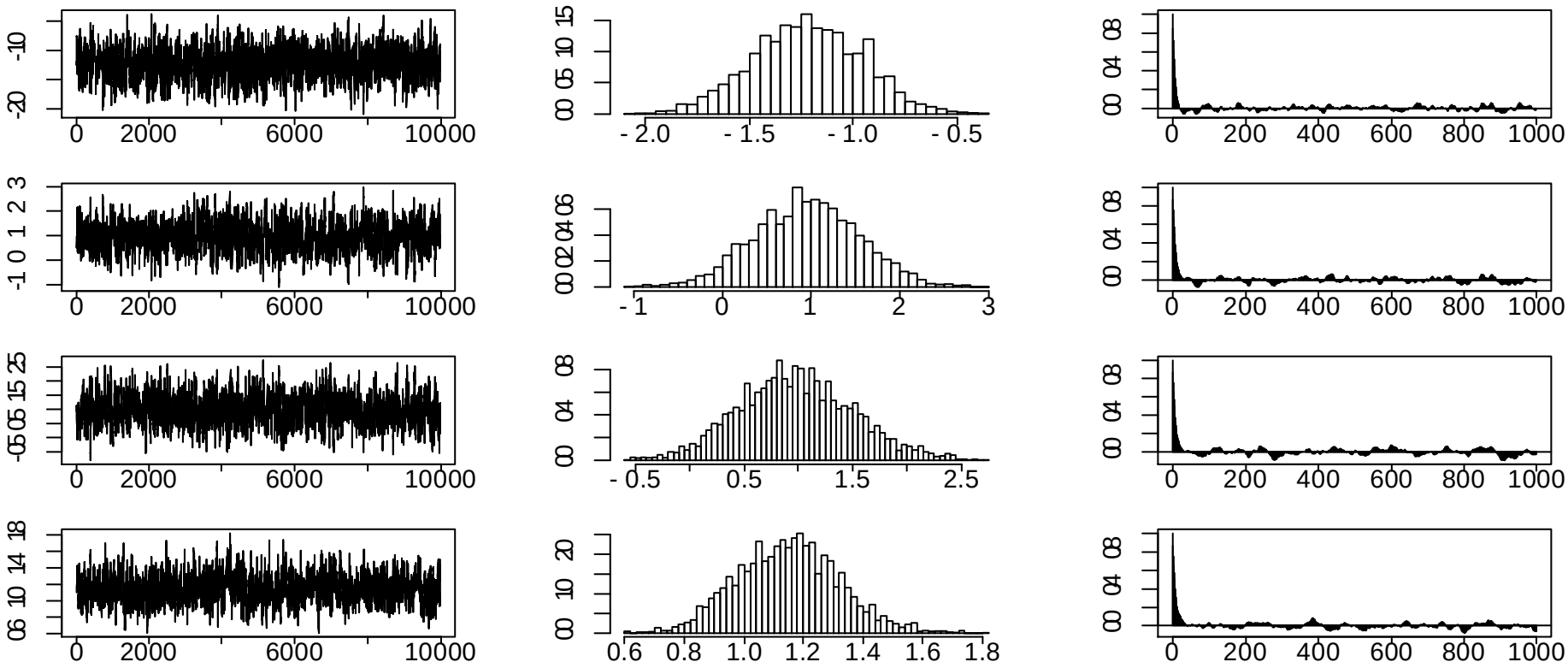



Figure: Traces (left), Histograms (middle) and Autocorrelations (right) for $\left(\beta_1^{(i)}, .., \beta_4^{(i)}\right)$ for $N = 10000$ and $N_0 = 1000$.

- We found for $\widehat{\beta} = \mathbb{E}\left(\beta \middle| y_{1:n}, \mathbf{x}_{1:n}\right) = (-1.22, 0.95, 0.96, 1.15)$ so a simple plug-in estimate of the predictive probability of a counterfeit bill is

$$\Pr\left(Y = 1 \middle| \mathbf{x}, \widehat{\beta}\right) = \Phi\left(\mathbf{x}^{\mathsf{T}}\widehat{\beta}\right)$$

- The predictive distribution is given by

$$\Pr\left(Y = 1 \middle| \mathbf{x}, \mathbf{x}_{1:n}\right) = \int \Phi\left(\mathbf{x}^{\mathsf{T}}\beta\right) \pi\left(\beta \middle| y_{1:n}, \mathbf{x}_{1:n}\right) d\beta.$$

- We rerun the algorithm on only $n = 100$ randomly selected data (50 genuine and 50 counterfeit) and use the results to classify the remaining 100 banknotes. The missclassification rate was 13% for the plug-in classifier and 7% for the full Bayesian approach.

# Limitations of the MH algorithm

- The MH algorithm is a simple and very general algorithm to sample from a target distribution $\pi(\theta)$.
- In practice, the choice of the proposal distribution has a crucial impact on the performance of the algorithm.
- In high dimensional problems, a simple MH algorithm will be useless. It will be necessary to use a combination of MH kernels.

# Towards more flexible algorithms

- It is possible to combine several MH kernels, say using

$$\begin{aligned} K\left(\theta'\middle|\theta\right) &= \lambda K_1\left(\theta'\middle|\theta\right) + (1-\lambda) K_2\left(\theta'\middle|\theta\right), \\ K\left(\theta'\middle|\theta\right) &= \int K_1\left(\theta'\middle|\theta''\right) K_2\left(\theta''\middle|\theta\right) d\theta'' \end{aligned}$$

  where $K_1$ (resp. $K_2$) is an MH algorithm of proposal $q_1$ (resp. $q_2$).
- Each proposal can modify only a subset of the components of $\theta$. That is, if $\theta = (\theta_1, \theta_2)$, then we can have $q_1\left(\theta'\middle|\theta\right) = q_1\left(\theta_1'\middle|\theta\right)\delta_{\theta_2}\left(\theta_2'\right)$ and $q_2\left(\theta'\middle|\theta\right) = \delta_{\theta_1}\left(\theta_1'\right) q_2\left(\theta_2'\middle|\theta\right)$.
- This algorithm satisfies

$$\begin{aligned} & \int \pi(\theta) K\left(\theta'\middle|\theta\right) d\theta \\ = \; & \lambda \int \pi(\theta) K_1\left(\theta'\middle|\theta\right) d\theta + (1-\lambda) \int \pi(\theta) K_2\left(\theta'\middle|\theta\right) d\theta \\ = \; & \lambda\pi\left(\theta'\right) + (1-\lambda)\pi\left(\theta'\right) = \pi\left(\theta'\right). \end{aligned}$$

# Metropolis one-at-a time

- To sample from $\pi(\theta)$ where $\theta = (\theta_1, ..., \theta_p)$, we can use the following algorithm at iteration $i$.
- Iteration $i$; $i \geq 1$:
  - For $k = 1 : p$
    - Sample $\theta_k^{(i)}$ using an MH step of proposal distribution $q_k\left(\theta_k | \left(\theta_{-k}^{(i)}, \theta_k^{(i-1)}\right)\right)$ and target $\pi\left(\theta_k | \theta_{-k}^{(i)}\right)$ where $\theta_{-k}^{(i)} = \left(\theta_1^{(i)}, ..., \theta_{k-1}^{(i)}, \theta_{k+1}^{(i-1)}, ..., \theta_p^{(i-1)}\right)$.
- *Remark*: It is possible to rewrite each MH step as a proposal $q_k\left(\theta_k | \left(\theta_{-k}^{(i)}, \theta_k^{(i-1)}\right)\right) \delta_{\theta_{-k}^{(i)}}(\theta_{-k})$ and target $\pi(\theta_k, \theta_{-k})$.

- The acceptance ratio of the MH step updating $\theta_k$ is given by

$$\begin{aligned}
& \alpha\left(\theta_{-k}^{(i)}, \theta_k^{(i-1)}, \theta_k^*\right) \\
= & \min\left(1, \frac{\pi\left(\theta_k^* \middle| \theta_{-k}^{(i)}\right) q_k\left(\theta_k^{(i-1)} \middle| \left(\theta_{-k}^{(i)}, \theta_k^*\right)\right)}{\pi\left(\theta_k^{(i-1)} \middle| \theta_{-k}^{(i)}\right) q_k\left(\theta_k^* \middle| \left(\theta_{-k}^{(i)}, \theta_k^{(i-1)}\right)\right)}\right) \\
= & \min\left(1, \frac{\pi\left(\theta_k^*, \theta_{-k}^{(i)}\right) q_k\left(\theta_k^{(i-1)} \middle| \left(\theta_{-k}^{(i)}, \theta_k^*\right)\right)}{\pi\left(\theta_k^{(i-1)}, \theta_{-k}^{(i)}\right) q_k\left(\theta_k^* \middle| \left(\theta_{-k}^{(i)}, \theta_k^{(i-1)}\right)\right)}\right)
\end{aligned}$$

- If we select $q_k\left(\theta_k \middle| \left(\theta_{-k}^{(i)}, \theta_k^{(i-1)}\right)\right) = \pi\left(\theta_k \middle| \theta_{-k}^{(i)}\right)$ then $\alpha\left(\theta_{-k}^{(i)}, \theta_k^{(i-1)}, \theta_k^*\right) = 1$ and we are back to the Gibbs sampler.
- **Example**: Assume we have $\pi(\theta_1, \theta_2)$ where it is easy to sample from $\pi(\theta_1 | \theta_2)$ and $\pi(\theta_2 | \theta_1)$ is not standard. We then use a Gibbs step to update $\theta_1$ and an MH step to update $\theta_2$.

# Example: Baseball data revisited

- We have the statistics of $J = 17$ players in pre-season exhibition matches where the data $y_j$ for the player $j$ corresponds to the number of home runs in $n_j$ times at the bat modelled through

$$Y_j | (n_j, p_j) \sim \mathcal{B}in (n_j, p_j) .$$

- We use the logit transformation and write

$$\theta_j = \log \left( \frac{p_j}{1 - p_j} \right) \Leftrightarrow p_j = \frac{\exp (\theta_j)}{1 + \exp (\theta_j)}$$

  which translates the parameter range $(0, 1)$ for $p_j$ to $(-\infty, \infty)$ for $\theta_j$.
- We set an exchangeable prior where

$$\theta_j | (\mu, \tau) \overset{\text{i.i.d.}}{\sim} \mathcal{N} (\mu, \tau^2) .$$

  and the hyperprior is selected as before with $a = b = 0.001$.

$$\pi (\tau^2) = \mathcal{IG} \left( \tau^2; \frac{a}{2}, \frac{b}{2} \right) , \ \pi (\mu) \propto 1.$$

- The full posterior distribution is given by

$$\begin{aligned}
& \pi\left(\mu, \tau^2, \theta_{1:J} \middle| y_{1:J}\right) \\
\propto \quad & \pi\left(\mu, \tau^2\right) \prod_{j=1}^{J} \pi\left(\theta_j \middle| \mu, \tau^2\right) \prod_{j=1}^{J} f\left(y_j \middle| \theta_j, n_j\right) \\
\propto \quad & \frac{1}{\tau^{J+a+2}} \exp\left(-\frac{b}{2\tau^2} - \sum_{j=1}^{J} \frac{\left(\theta_j - \mu\right)^2}{2\tau^2}\right) \\
& \times \prod_{j=1}^{J} \left(\frac{\exp\left(\theta_j\right)}{1+\exp\left(\theta_j\right)}\right)^{y_j} \left(\frac{1}{1+\exp\left(\theta_j\right)}\right)^{n_j - y_j}
\end{aligned}$$

- This distribution does not admit a closed-form expression and we are going to use the Gibbs sampler by decomposing the parameter space in $J+2$ blocks $\mu$, $\tau^2$, $\theta_1$, $\theta_2$, ...., $\theta_j$.
- The Gibbs sampler will require being able to be able to sample from $\pi\left(\mu \middle| y_{1:J}, \tau^2, \theta_{1:J}\right)$, $\pi\left(\tau^2 \middle| y_{1:J}, \mu, \theta_{1:J}\right)$ and $\pi\left(\theta_j \middle| y_{1:J}, \mu, \tau^2\right)$.

- Full conditional distribution for $\mu$

$$\begin{aligned}\pi\left(\mu\middle|\, y_{1:J}, \tau^2, \theta_{1:J}\right) &= \pi\left(\mu\middle|\, \tau^2, \theta_{1:J}\right) \\ &= \mathcal{N}\left(\mu; J^{-1}\sum_{j=1}^{J}\theta_j, J^{-1}\tau^2\right).\end{aligned}$$

- Full conditional distribution for $\tau^2$

$$\begin{aligned}\pi\left(\tau^2\middle|\, y_{1:J}, \mu, \theta_{1:J}\right) &= \pi\left(\tau^2\middle|\, \mu, \theta_{1:J}\right) \\ &= \mathcal{IG}\left(\tau^2; \frac{a+J}{2}, \frac{b+\sum_{j=1}^{J}\left(\theta_j-\mu\right)^2}{2}\right).\end{aligned}$$

- These two distributions are standard.

- Full conditional distribution for $\theta_j$

$$\pi\left(\theta_j \middle| y_{1:J}, \tau^2, \mu\right) \propto \frac{\exp\left(\theta_j\right)^{y_j}}{\left(1+\exp\left(\theta_j\right)\right)^{n_j}} \exp\left(-\frac{\left(\theta_j-\mu\right)^2}{2\tau^2}\right).$$

- This distribution does not admit a closed-form expression so we use a Metropolis algorithm where

$$q\left(\theta_j' \middle| \theta_j\right) = \mathcal{U}\left(\theta_j'; [\theta_j-\delta, \theta_j+\delta]\right)$$

  where $\delta$ is a parameter of the algorithm.
- We run $N = 100,000$ iterations of the Metropolis-Hastings one-at-a time and discard the first $N_0 = 10,000$ samples. The acceptance rates for the Metropolis algorithm was given by

| $\delta$ | 0.1 | 0.5 | 1 | 2 | 10 |
|---|---|---|---|---|---|
| Average acceptance proba. | 0.87 | 0.55 | 0.35 | 0.17 | 0.04 |

- Based on these results we computed the predictive distribution of $Z_i$ the number of homes runs in the full season where $m_j$ is the number of bats in the full seasons. We assume

$$Z_i \sim \mathcal{B}in\left(m_j, p_j\right).$$

- The predictive distribution is given by

$$\Pr\left(Z_i = k \middle| y_{1:J}, n_{1:J}, m_{1:J}\right) = \int \Pr\left(Z_i = k \middle| m_j, p_j\right) \pi\left(p_j \middle| y_{1:J}, n_{1:J}\right) dp_j$$

and can be approximated through the ouput of the MCMC algorithm

$$\frac{1}{N - N_0 + 1} \sum_{i=N_0}^{N} \Pr\left(Z_i = k \middle| m_j^{(i)}, p_j^{(i)}\right).$$

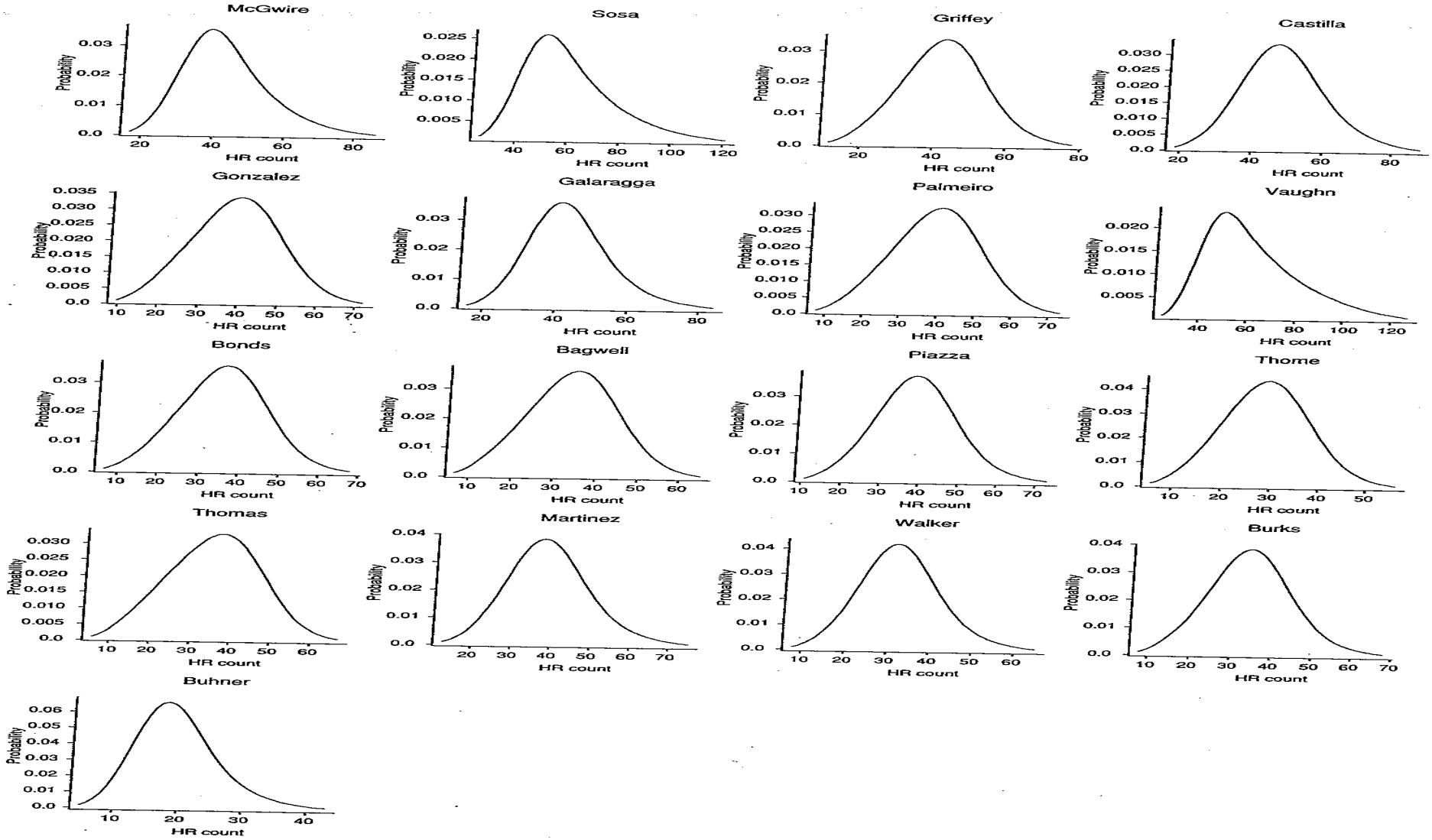



Figure: Predictive distribution for the number of home runs scored by each batter.