Ch4. Bayesian Methods

ST4240, 2014/2015

Version 0.2

Alexandre Thiéry

Department of Statistics and Applied Probability

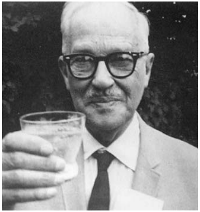
Inverse Probabilities

- Thomas Bayes (18th century)
  $\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A|B) \times \mathbb{P}(B)}{\mathbb{P}(A)}$

- Pierre-Simon Laplace (18th century)
  Bayes' Theroem, 11 years later
  Laplacian inference
  Laplace approximation

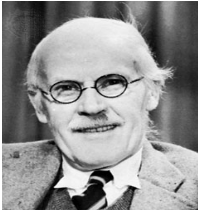
Frequentist Statistics

- Ronald A. Fisher (1890-1962)
- Jerzy Neyman (1894 -1981)

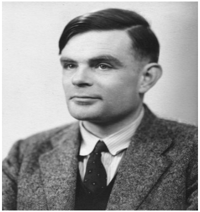


Revival of Bayesian Statistics

- Harold Jeffreys (1891 - 1989)



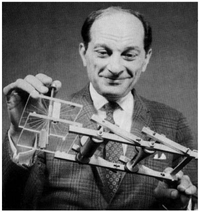
- Alan Turing (1912 - 1954)



Alan Turing's life

Monte Carlo methods: 1940's

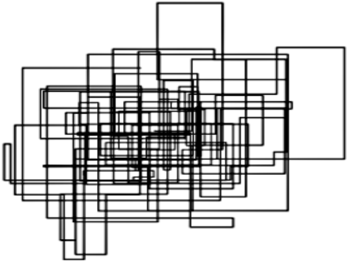
- John Von Neuman (1903 - 1957)
- Stan Ulam (1909 - 1984)

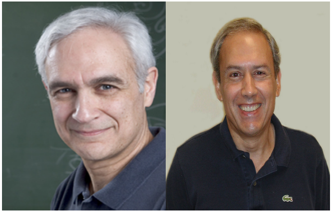


Gibbs Sampling

- Valentin Fedorovich Turchin
  Invents the Gibbs algorithm (1971)

- Stuart and Donald Geman
  Rediscovery, 13 years later!





Outline

Bayes' formula

- For two events A and B,

$$\mathbb{P}(B \mid A) \;=\; \frac{\mathbb{P}(A \mid B) \times \mathbb{P}(B)}{\mathbb{P}(A)}$$

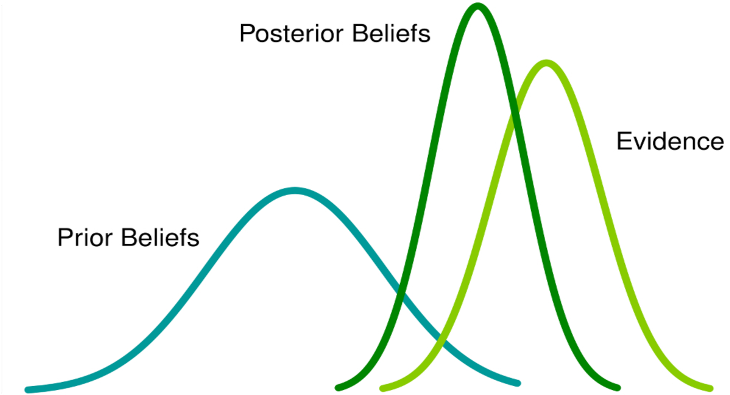
- In Bayesian Statistics,

$$p(\theta \mid \mathrm{data}) = \frac{p(\mathrm{data} \mid \theta) \times \pi_0(\theta)}{p(\mathrm{data})}$$

- or equivalently

$$(\mathrm{posterior}) \;\propto\; (\mathrm{likelihood}) \times (\mathrm{prior})$$

Bayes' Law



Deadly disease!

- A doctor has a bad news for you:
- The test for this deadly disease that you have done last week is positive!
- The test is 99% accurate:
  - $\mathbb{P}\{(\text{test positive}) \mid (\text{sick})\} = 99\%$
  - $\mathbb{P}\{(\text{test negative}) \mid (\text{non sick})\} = 99\%$
- The disease is rare: $1/100000$ of population!
- **[Exercise]** how worried should you be?

Noisy observation: Gaussian case

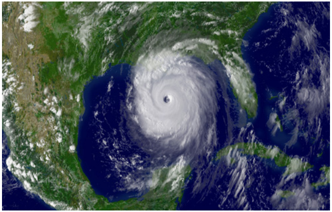
- $x \in \mathbb{R}$ is an unknown quantity.
- A radar gives a noisy estimate x,

$$y = x + \mathbf{N}\left(0, \sigma^2_{\text{(noise)}}\right)$$

- Prior distribution on position of x,

$$\pi_0(x) \sim \mathbf{N}\left(0, \sigma^2_{\text{(prior)}}\right)$$

- **[Exercise]** posterior for x ?
- Leads to the Kalman Filter.



Gaussian setting: multivariate case

- Consider a prior distribution on $\beta \in \mathbb{R}^{p+1}$ given by

$$\pi_0(\beta) \sim \mathbf{N}\left(0, \tau^2\, I_{p+1}\right)$$

- For a design matrix $X \in \mathbb{R}^{n,p+1}$ we observe

$$y \sim X\,\beta + \mathbf{N}\left(0, \sigma^2\, I_n\right)$$

- **[Exercise]** what is the posterior distribution π for β? Write down the mean and covariance of π.

Outline

Gibbs sampling: the setting

- Suppose that one is interested in the target distribution

$$\pi\left(x^{(1)}, x^{(2)}, \ldots, x^{(d)}\right)$$

- To be able to use the Gibbs sampler, one needs to be able to **simulate from the conditional distributions**

$$X^{(i)} \mid X^{(-i)} = x^{(-i)}$$

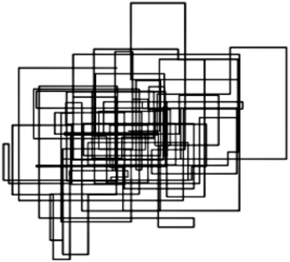
where $x^{(-i)} = \left(x^{(1)}, \ldots, x^{(i-1)}, x^{(i+1)}, \ldots, x^{(d)}\right)$

Gibbs sampling: the algorithm

- Choose a coordinate i at random
- replace $x^{(i)}$ by a sample of the conditional law

$$X^{(i)} \mid X^{(-i)} = x^{(-i)}$$

- Iterate!

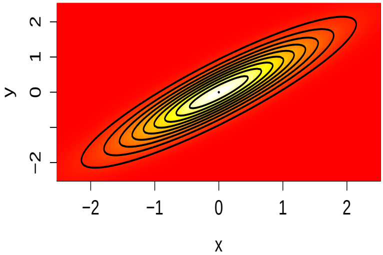


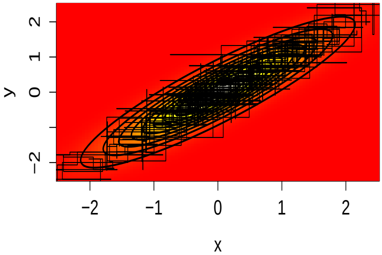
Gaussian example

- A bi-variate centred Gaussian random variable (X_1, X_2) with $\sigma^2_{X_1} = \sigma^2_{X_2} = 1$ and $\mathrm{Corr}(X_1, X_2) = \rho$ has density

$$\pi(x_1, x_2) \propto \exp\left\{ -\frac{x_1^2 + x_2^2 - 2\rho x_1 x_2}{2(1-\rho^2)} \right\}$$

- **[Exercise]** Gibbs sampler for π?





Slice Sampling: the trick

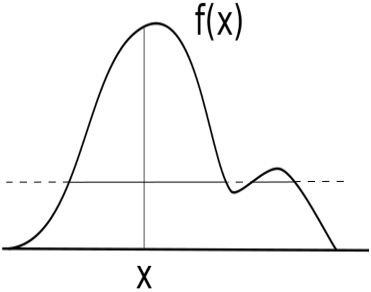
- Goal: Gibbs sampling for $\pi(x) \propto f(x)$ for $x \in \mathbb{R}^d$
- Data augmentation trick: the density $\widetilde{\pi}$ on $\mathbb{R}^d \times \mathbb{R}$

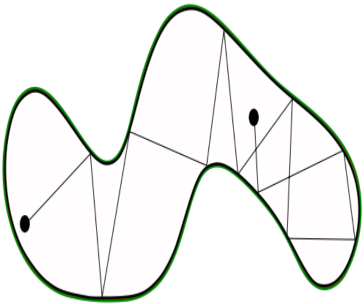
$$\widetilde{\pi}(x, u) \propto \mathbb{I}\Big(0 \leq u \leq f(x)\Big)$$

  has same marginal distribution as π
- In other words

$$\int_{u=0}^{\infty} \widetilde{\pi}(x, u)\, du = \pi(x)$$

Slice sampling: the intuition





Slice sampling: the algorithm

- Sample $u \mid x$:

$$u \sim \mathrm{Unif}\Big\{[0, f(x)]\Big\}$$

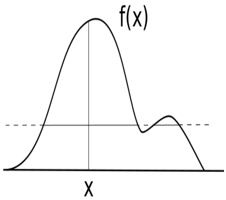
- Sample $x \mid u$:

$$x \sim \mathrm{Unif}\Big\{S(u)\Big\}$$

where $S(u)$ is the slice

$$S(u) = \Big\{x \in \mathbb{R}^d \; : \; f(x) \geq u\Big\}$$

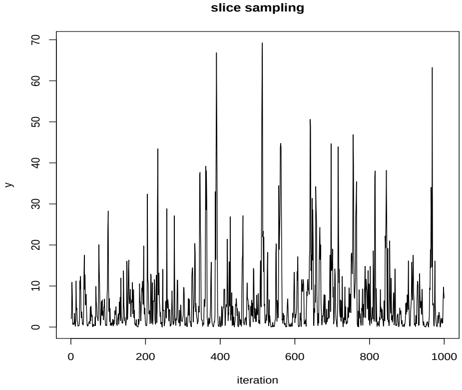
- Iterate!



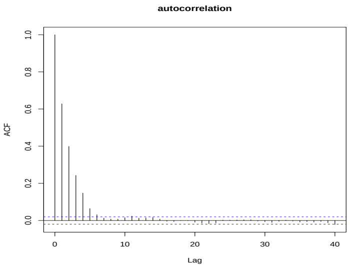
Slice sampling: example

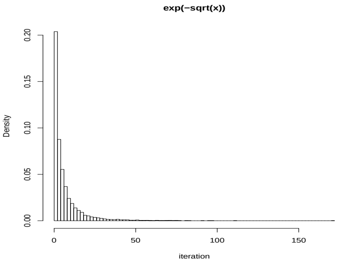
- **[Exercise]** use slice sampling for sampling from the distribution on $x \in (0, \infty)$ defined by

$$\pi(x) \propto e^{-\sqrt{x}}$$



Slice sampling: example





Outline

The model

- Consider a prior distribution on $\beta \in \mathbb{R}^{p+1}$ given by

$$\pi_0(\beta) \sim \mathbf{N}\left(0, \tau^2 I_{p+1}\right)$$

- For a design matrix $X \in \mathbb{R}^{n,p+1}$ we observe

$$y \sim X\beta + \mathbf{N}\left(0, \sigma^2 I_n\right)$$

- We assume a (Jeffrey) prior on σ^2,

$$\pi_0(\sigma^2) \propto 1/\sigma^2$$

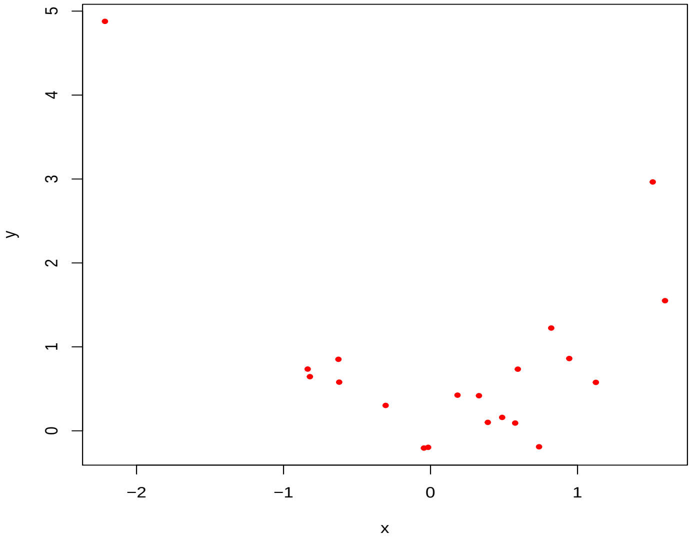
The model

- For clarity, let us set $v \equiv \sigma^2$
- **[Exercise]** Find the conditional distribution $\beta \mid y, v$.
- **[Exercise]** Prove that if $Z \sim \Gamma(\alpha, \beta) \propto z^{\alpha-1}e^{-\beta z}$ then

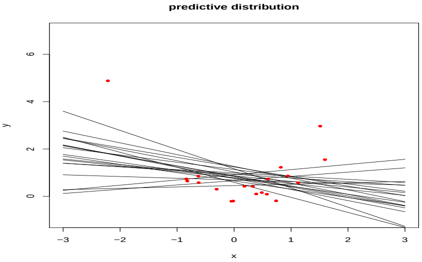
$$1/Z \equiv Y \sim \mathrm{IG}(\alpha, \beta) \propto \frac{e^{-\beta/y}}{y^{\alpha+1}}$$

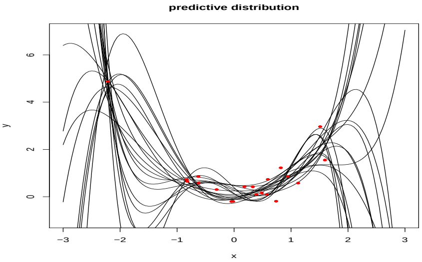
- **[Exercise]** Find the conditional distribution $v \mid y, \beta$
- Deduce a Gibbs sampling algorithm for sampling from he Bayesian ridge regression model.

Bayesian regression: the data

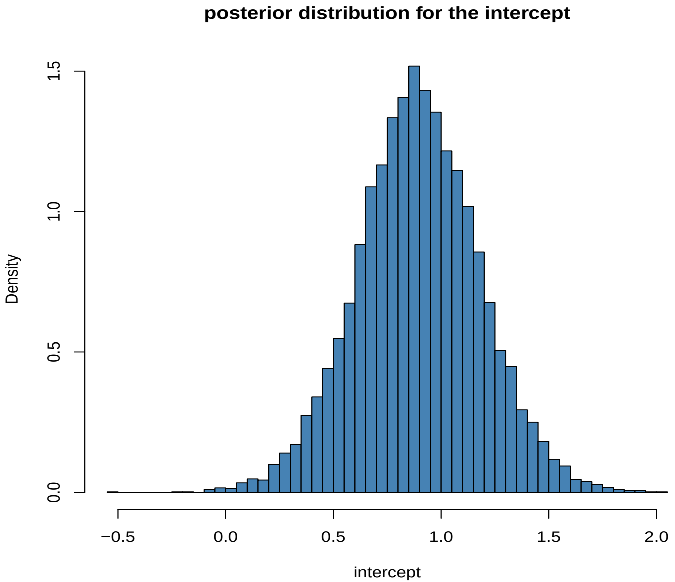


Bayesian regression: fitting





Bayesian regression: uncertainty quantification



Bayesian prediction – predictive

- Goal: estimate $\varphi(\theta)$ given some data $y = (y_1, \ldots, y_n)$ and a probabilistic model describing how the data are generated,

$$p\,(y \mid \theta)$$

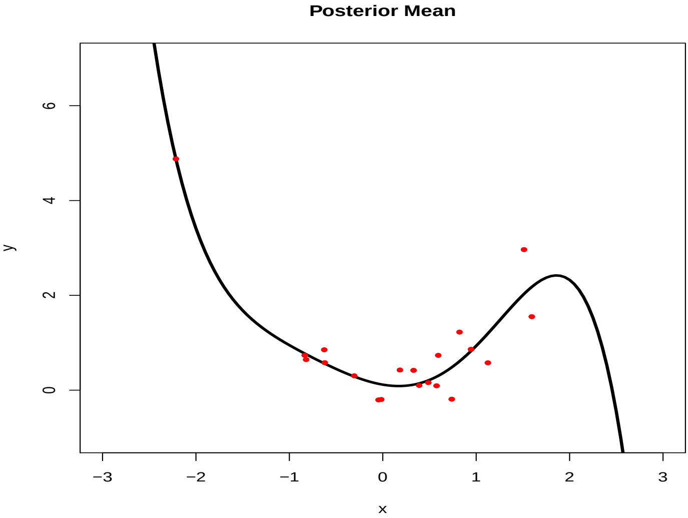
- The posterior mean is the estimator that minimises the squared error

$$\mathbb{E}\,[\varphi(\theta) |\, y]$$

- In practice: a long MCMC simulation $(\theta_1, \ldots, \theta_N)$ from the posterior distribution π

$$\widehat{\varphi(\theta)} \equiv N^{-1} \sum_{i=1}^{N} \varphi(\theta_i)$$

Posterior mean



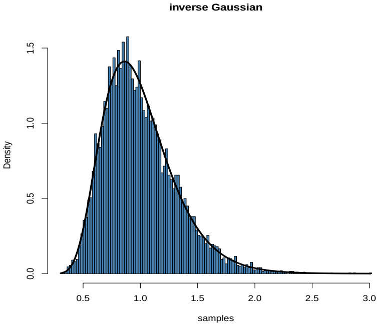
Outline

Bayesian Toolbox I

- Inverse Gaussian distribution $\mathrm{InvGauss}(\mu, \lambda)$ on $x \in (0, \infty)$,

$$\left[\frac{\lambda}{2\pi x^3}\right]^{1/2} \exp\left\{-\frac{\lambda (x-\mu)^2}{2\mu^2 x}\right\}$$

  has mean μ and variance μ^3/λ. The quantity $1/\lambda$ is sometimes called the dispersion parameter



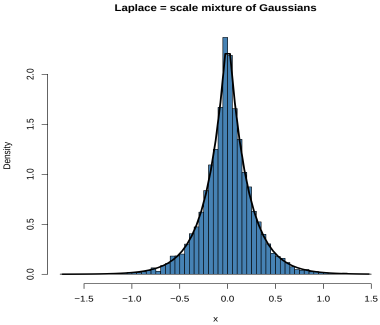
Bayesian Toolbox II

- Laplace distribution $\mathrm{Laplace}(\lambda)$ with rate λ on $x \in (-\infty, \infty)$,

$$\frac{\lambda}{2}\, e^{-\lambda\,|x|}$$

- Laplace distribution $X \sim \mathrm{Laplace}(\lambda)$ as scale mixture of Normal

$$\tau \sim \mathrm{Exp}(\lambda^2/2)$$
$$X \sim \mathbf{N}\,(0, \tau)$$



Generalized Double Pareto model

- Data $y \in \mathbb{R}^n$ are collected and modelled by a linear model

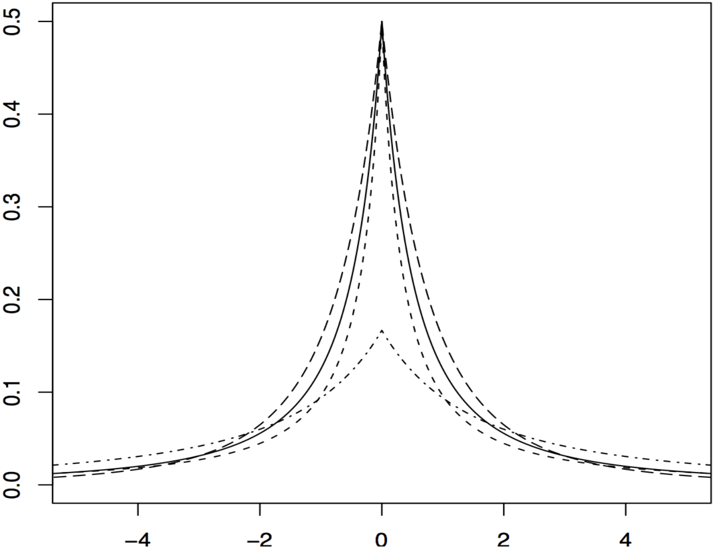
$$y = X\beta + \mathbf{N}\left(0, \sigma^2 I_n\right)$$

for a known design matrix $X \in \mathbb{R}^{n,p}$.
- Goal: estimate coefficient $\beta \in \mathbb{R}^p$ and noise intensity $\sigma^2 > 0$.
- We assume a (Jeffrey) prior on $v \equiv \sigma^2$ so that $\pi_0(v) \propto 1/v$
- We assume the following sparsity inducing prior for $\beta \in \mathbb{R}^p$

$$\begin{aligned} \beta_j &\sim \mathbf{N}\left(0, \sigma^2 \tau_j\right) \\ \tau_j &\sim \mathrm{Exp}(\lambda_j^2/2) \\ \lambda_j &\sim \Gamma(\alpha, \eta) \end{aligned}$$

for some fixed parameter $\alpha > 0$ and $\eta > 0$.
- Note that $\overline{\tau} = (\tau_1, \ldots, \tau_p) \in \mathbb{R}^p$ and $\overline{\lambda} = (\lambda_1, \ldots, \lambda_p) \in \mathbb{R}^p$

Generalized Double Pareto prior



Gibbs Sampling

- To design a Gibbs sampler, one needs to find the conditional distributions
  - $\beta \mid (y, v, \overline{\tau}, \overline{\lambda})$ which also equals $\beta \mid (y, v, \overline{\tau})$
  - $v \mid (y, \beta, \overline{\tau}, \overline{\lambda})$ which also equals $v \mid (y, \beta)$
  - $\overline{\tau} \mid (y, \beta, v, \overline{\lambda})$ and it suffices to study $\tau_j \mid (y, \beta_j, \lambda_j)$
  - $\overline{\lambda} \mid (y, \beta, v, \overline{\tau})$ and it suffices to study $\lambda_j \mid (y, \tau_j)$
- If one can integrate out a variable, one should do it since this yields to a faster algorithm. In our case, it turns out that it is possible to write down an expression for $\overline{\lambda} \mid (y, \beta, v)$ instead of $\overline{\tau} \mid (y, \beta, v, \overline{\lambda})$.

Gibbs Sampling

- **[Exercise]** Prove that $\beta \mid (y, v, \overline{\tau})$ is a Gaussian and find its mean and variance.
- **Answer:** The conditional distribution is Gaussian with parameters

$$\Sigma = v\ \left(X^T X + T^{-1}\right)^{-1} \qquad \text{and} \qquad \mu = \left(X^T X + T^{-1}\right)^{-1} X^T y$$

where $T = \mathrm{Diag}(\tau_1, \ldots, \tau_p)$ is the square matrix $T \in \mathbb{R}^{p \times p}$ with $(\tau_1, \ldots, \tau_p)$ on the diagonal.

Gibbs Sampling

- **[Exercise]** Prove that $v \mid (y, \beta)$ is an inverse Gamma distribution and find its parameters.
- **Answer:** The conditional distribution is an inverse Gamma distribution with parameters

$$\mathrm{IG}\left([n+p]/2, \|y - X\,\beta\|^2/2 + \left\langle \beta, T^{-1}\,\beta \right\rangle/2\right)$$

Gibbs Sampling

- **[Exercise]** Prove that $\tau_j^{-1} \mid (y, \beta_j, \lambda_j)$ has the law of an inverse Gaussian distribution and find its parameters.
- **Answer:** the conditional distribution of $\omega_j \equiv \tau_j^{-1}$ is an inverse Gaussian distribution with mean $\lambda_j\sqrt{v}/|\beta_j|$ and dispersion parameter $1/\lambda_j^2$.

Gibbs Sampling

- **[Exercise]** Prove that $\lambda_j \mid (y, \tau_j)$ has a Gamma distribution and find its parameters.
- **Answer:** the conditional distribution $\lambda_j \mid (y, \tau_j)$ is

$$\Gamma(\alpha + 1, \eta + |\beta_j|/\sqrt{v})$$