Bayesian Data Analysis & MCMC

Jian Zeng

September 28, 2016

What is Bayesian statistics?

Definition

Bayesian statistics, named for Thomas Bayes (1701–1761), is a theory in the field of statistics in which the evidence about the **true state of the world** is expressed in terms of 'degrees of belief' called **Bayesian probabilities**. – *Wikipedia*

- Fallacy: Bayesian methods depend on totally subjective interpretations of probability
- Truth: Bayesians share the same viewpoint of the world with Frequentists
- The true state of nature is embodied in a fixed but unknown parameter value that governs the distribution of observable quantities
- If we know everything about all physical relations in the world, we would know the values that would be assumed by observable quantities with certainty

Meaning of probability

Frequentist

- The probability of an event is the limiting value of its frequency in a large number of trials

Bayesian

- Probabilities are used to quantify our beliefs or knowledge about possible values of unknowns (parameters)

This is the fundamental difference between Bayesian and Frequentist statistics

What is fixed? What is random?

Frequentist

- Data are repeatable random samples – *random variables*
- Underlying parameters remain constant during the repeatable process
- Parameters are fixed

Bayesian

- Data are observed from the realized sample
- Data are fixed
- Parameters are unknown and described probabilistically
- Not necessary to define random variable

Bayesian probability

- It is legitimate to write

$$\Pr(t_1 < \theta < t_2) = c$$

with θ, t_1, t_2 and c all being constants
- Not a statement a random quantity or random variable
- It is a statement about our *knowledge* that θ lies in the interval (t_1, t_2)

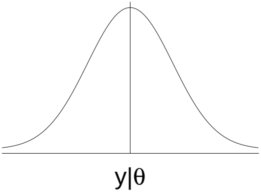
Example

- What is the probability that $h^2 > 0.5$?
- What is the probability that height is controlled by more than 1000 loci?

How to make inference?

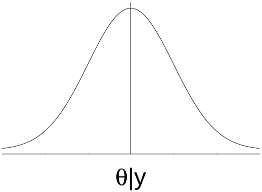
Frequentist

- Maximum likelihood



Bayesian

- Posterior probability



Bayes Theorem

The conditional probability of X given Y is

$$\Pr(X \mid Y) = \frac{\Pr(X, Y)}{\Pr(Y)} = \frac{\Pr(Y \mid X)\Pr(X)}{\Pr(Y)}$$

where $\Pr(X, Y)$ is the joint probability of X and Y, $\Pr(X)$ is the probability of X, and $\Pr(Y)$is the probability of Y.

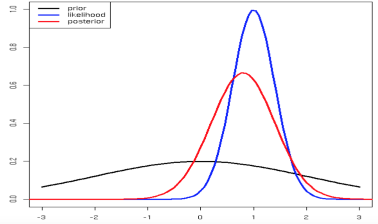
Essential of Bayesian inference

- **Prior** probabilities quantify beliefs about parameters before the data are analyzed
- Parameters are related to the data through the model or "**likelihood**" which is the conditional probability density for the data given the parameters
- The prior and the likelihood are combined using Bayes theorem to obtain **posterior** probabilities, which are conditional probabilities for the parameters given the data
- Inferences about parameters are based on the posterior

Bayesian theorem in Bayesian inference

- Let $f(\boldsymbol{\theta})$ denote the prior probability density for $\boldsymbol{\theta}$
- Let $f(\mathbf{y}|\boldsymbol{\theta})$ denote the likelihood
- Then, the posterior probability of $\boldsymbol{\theta}$ is

$$\begin{aligned} f(\boldsymbol{\theta}|\mathbf{y}) &= \frac{f(\mathbf{y}|\boldsymbol{\theta})\, f(\boldsymbol{\theta})}{f(\mathbf{y})} \\ &\propto f(\mathbf{y}|\boldsymbol{\theta})\, f(\boldsymbol{\theta}) \end{aligned}$$



Example: the conjugate prior for the normal distribution

Suppose

$$y_i \,|\mu \sim N\left(\mu, \sigma^2\right) \, i.i.d. \text{ and } \mu \sim N(\mu_0, \sigma_0^2)$$

where σ^2, μ_0 and σ_0^2 are known. Then:

$$\mu \,|\mathbf{y} \sim N\left(\frac{\sigma_0^2}{\frac{\sigma^2}{n} + \sigma_0^2}\bar{y} + \frac{\sigma^2}{\frac{\sigma^2}{n} + \sigma_0^2}\mu_0, \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\right)$$

- With no observations, the posterior mean is the prior mean
- As the number of observations becomes large, the posterior mean $\approx \bar{y}$

Equivalence to BLUP

The *i.i.d.* observations can be represented by the model:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{e}$$

with a prior knowledge that $\mu = \mu_0$ with uncertainty σ_0^2. Thus, the linear model with the additional (prior) data:

$$\begin{bmatrix} \mathbf{y} \\ \mu_0 \end{bmatrix} = \begin{bmatrix} \mathbf{1} \\ 1 \end{bmatrix} \mu + \begin{bmatrix} \mathbf{e} \\ \boldsymbol{\epsilon} \end{bmatrix} \text{ with } \mathit{Var} \begin{bmatrix} \mathbf{y} \\ \mu_0 \end{bmatrix} = \begin{bmatrix} \mathbf{I}\sigma^2 & 0 \\ 0 & \sigma_0^2 \end{bmatrix}$$

OLS equations:

$$\begin{aligned} \begin{bmatrix} \mathbf{1}' & 1 \end{bmatrix} \begin{bmatrix} \mathbf{I}\sigma^2 & 0 \\ 0 & \sigma_0^2 \end{bmatrix} \begin{bmatrix} \mathbf{1} \\ 1 \end{bmatrix} \hat{\mu} &= \begin{bmatrix} \mathbf{1}' & 1 \end{bmatrix} \begin{bmatrix} \mathbf{I}\sigma^2 & 0 \\ 0 & \sigma_0^2 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mu_0 \end{bmatrix} \\ \left(\frac{\mathbf{1}'\mathbf{1}}{\sigma^2} + \frac{1}{\sigma_0^2} \right) \hat{\mu} &= \frac{\mathbf{1}'\mathbf{y}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \end{aligned}$$

Computing posteriors

- Often no closed form for $f(\boldsymbol{\theta}\,|\mathbf{y})$
  - Non-conjugate prior: e.g. mixture prior for SNP effects
- Further, even if computing $f(\boldsymbol{\theta}\,|\mathbf{y})$ is feasible, obtaining $f(\theta_j\,|\mathbf{y})$ would require integrating over many dimensions, e.g.

$$f(\theta_1\,|\mathbf{y}) = \int f(\theta_1\,|\theta_2, \mathbf{y})\, f(\theta_2\,|\mathbf{y})\,\mathrm{d}\theta_2$$

- Thus, in many situations, inferences are mad using the empirical posterior constructed by drawing samples from $f(\boldsymbol{\theta}\,|\mathbf{y})$
- MCMC (Markov chain Monte Carlo) techniques are widely used for drawing samples from posteriors and for making inferences

Monte Carlo integration

Consider evaluating the integral

$$E_f\left[h\left(X\right)\right] = \int h\left(x\right) f\left(x\right) \mathrm{d}x$$

Using the Monte Carlo estimate

$$\hat{h} = \frac{1}{T} \sum_{t=1}^{T} h\left(x^{(t)}\right)$$

where $x^{(t)} \sim i.i.d. f\left(x\right)$.

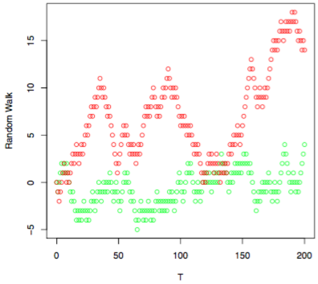
- Now, integration problem solved! But how to draw sample from $f\left(x\right)$, namely $f\left(\boldsymbol{\theta} \left| \mathbf{y} \right.\right)$?

Markov chain

- Stochastic process is a sequence of random variable $\{X(t), t \in T\}$
  - $X(t)$ is the state of the process at time t
  - T is the set of time points at which we observe $X(t)$
  - The state space is the set of possible values of $X(t)$
- A stochastic process has the *Markov property* if, given the present, the future does not depend on the past
- A stochastic process satisfies the Markov property is called *Markov chain*

Markov chain

- A simple example of a Markov chain is the random walk. At each time point, move right one step with probability p or move left one step with probability $1 - p$
- Starting at $X(0) = 0$ move left or right by 1 with probability $p = 0.5$ over $T = 200$ steps



Inference from Markov chain

Can show that samples obtained from a Markov chain can be used to draw inferences from the joint posterior distribution provided the chain is:

- **Irreducible** (Ergodic): can move from any state i to any other state j
- **Positive recurrent** (aperiodic): return time to any state has finite expectation
- *Markov Chains*, J. R. Norris (1997)

MCMC sampling techniques

- Gibbs sampler
- Metropolis-Hastings sampler

Gibbs sampler

- Want to draw samples from $f(x_1, x_2, \ldots, x_n)$
- Even though it may be possible to compute $f(x_1, x_2, \ldots, x_n)$, it is difficult to draw samples directly from $f(x_1, x_2, \ldots, x_n)$
- Gibbs:
  - Get valid a starting point $\boldsymbol{x}^0$
  - Draw sample $\boldsymbol{x}^t$ as:

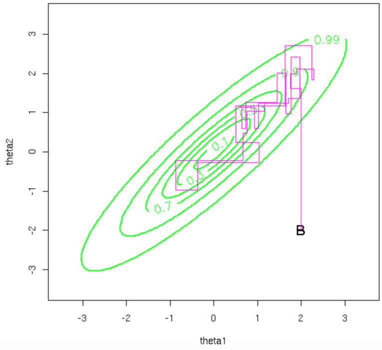
$$
\begin{array}{lll}
x_1^t & \text{from} & f(x_1 | x_2^{t-1}, x_3^{t-1}, \ldots, x_n^{t-1}) \\
x_2^t & \text{from} & f(x_2 | x_1^{t}, x_3^{t-1}, \ldots, x_n^{t-1}) \\
x_3^t & \text{from} & f(x_3 | x_1^{t}, x_2^{t}, \ldots, x_n^{t-1}) \\
\vdots & & \vdots \\
x_n^t & \text{from} & f(x_n | x_1^{t}, x_2^{t}, \ldots, x_{n-1}^{t})
\end{array}
$$

- The sequence $\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^n$ is a Markov chain with stationary distribution $f(x_1, x_2, \ldots, x_n)$

Why Gibbs sampling works

Gibbs sampling can be thought of as a practical implementation of the fact that knowledge of the conditional distributions is sufficient to determine a joint distribution.

– Casella and George



Metropolis-Hastings sampler

- Sometimes may not be able to draw samples directly from $f(x_i|\boldsymbol{x}_{i\_})$
- Convergence of the Gibbs sampler may be too slow
- Metropolis-Hastings (MH) for sampling from $f(x)$:
  - a candidate sample, y, is drawn from a proposal distribution $q(y|x^{t-1})$
  - $$x^t = \begin{cases} y & \text{with probability } \alpha \\ x^{t-1} & \text{with probability } 1-\alpha \end{cases}$$
  - $$\alpha = \min(1, \frac{f(y)q(x^{t-1}|y)}{f(x^{t-1})q(y|x^{t-1})})$$
- The samples from MH is a Markov chain with stationary distribution $f(x)$

Proposal distributions

Two main types:

- Approximations of the target density: $f(x)$
  - Not easy to find approximation that is easy to sample from
  - High acceptance rate is good!
- Random walk type: stay close to the previous sample
  - Generally easy to construct proposal
  - High acceptance rate may indicate that candidate is too close to previous sample
  - Intermediate acceptance rate is good

Applications in whole-genome analyses

- Prediction
  - predicting phenotypes, polygenic sores of individual risk
- Estimation of quantities of interest
  - SNP effects, genetic variance
  - SNP-based heritability
- Hypothesis test
  - Bayesian GWAS

Popular Bayesian methods for whole-genome analyses

$$y_i = \mu + \sum_j X_{ij}\alpha_j + e_i$$

Priors:

- $\mu \propto$constant (not proper, but posterior is proper)
- $e_i \sim i.i.d.N\left(0, \sigma_e^2\right); \sigma_e^2 \sim \nu_e S_e^2 \chi_{\nu_e}^{-2}$
- Different priors for α_j

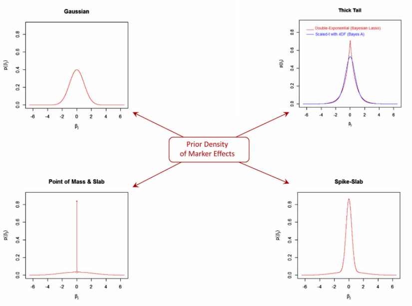


Figure 1 Commonly used prior densities of marker effects (all with zero mean and unit variance). The densities are organized in a way that, starting from the Gaussian in the top left corner, as one moves clockwise, the amount of mass at zero increases and tails become thicker and flatter.

Priors for SNP effects

- BayesA; BayesB (Meuwissen et al. 2001)
  - univariate-t prior; a mixture of zero with a given prob. π and t-distribution with prob. $1 - \pi$
- BayesC; BayesCπ (Habier et al. 2011)
  - a mixture of zero and normal distribution with unknown π
- BayesR (Erbe et al. 2012); BayesRC (Macleod et al. 2016)
  - a mixture of normals; can incorporate functional information
- BayesLasso (Park and Casella, 2008)
  - double exponential distribution
- BSLMM (Zhou et al. 2013); BOLT-LMM (Loh et al. 2015)
  - BayesC$\pi +$ polygenic component; efficient variational Bayes

Advantages and disadvantages of Bayesian methods

Advantages:

- Simultaneously fit all SNPs in the model
- Incorporate prior knowledge, e.g. mixture prior for SNP effects
- Appealing interpretation of results
- Simultaneous discovery, estimation and prediction analysis

Disadvantages:

- Computational cost
- Does not guarantee converge