My close encounter with STAN

Christiaan van Dorp

March 21, 2016

Motivation — 'flu sampler'

- 45 years of influenza-like illness (ILI) reporting by general practitioners (GPs).
- Each year, what is the fraction of susceptibles to influenza, in different age groups?
- Problems:
  - ILI is not influenza: 50% due to pathogens like respiratory syncytial virus (RSV), rhinovirus, parainfluenza, ...
  - Not everybody with ILI consults a GP (under reporting)
  - ILI GP consulting decreases over time.
- Method: Bayesian hierarchical model + SIR model + MCMC
  - Many parameters: $k > 45 \times 6 = 270$
  - Observations: $n = 13333$

Bayesian inference

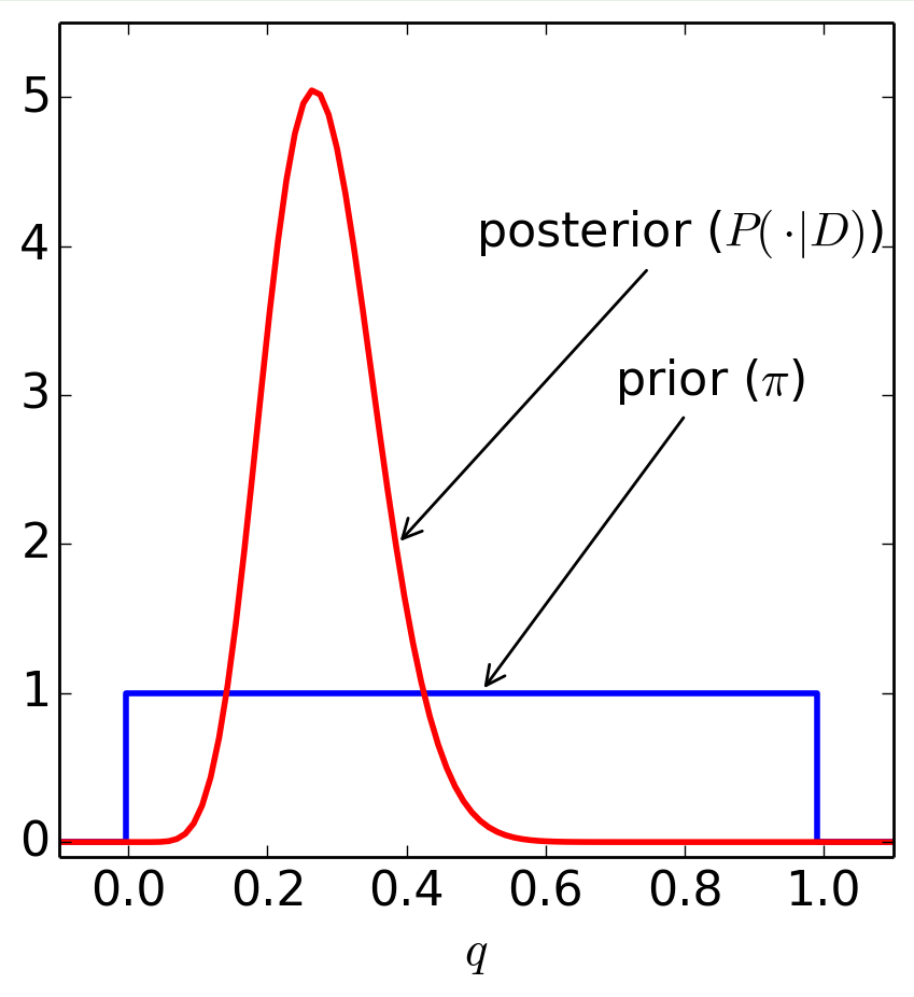
- Parameters are random variables (and have a probability density).
- Prior belief about parameters + data $\longrightarrow$ 'updated' (posterior) belief about parameters
- i.e. Adding data to your model (hopefully) narrows the parameter distribution (parameter estimation).

Example

- Data: $D = \{k_1 = 3, k_2 = 4, k_3 = 1\}$
- Model: $M : k_i \sim \mathrm{Binom}(n, q)$ with $n = 10$ (likelihood of observing D, given q: $P(D|q)$)
- Prior: $q \sim \mathrm{Beta}(1, 1) = \mathrm{Uniform}(0, 1)$ (PDF for q is $\pi(q) = 1$)
- Posterior:
  $P(q|D) = \frac{P(D|q)\pi(q)}{\int P(D|q)\pi(q)dq} = c \cdot q^{k_1+k_2+k_3}(1-q)^{3n-(k_1+k_2+k_3)}$.
- Hence, after observing D, the updated distribution of q is $\mathrm{Beta}(9, 23)$.

Bayesian inference

Example (continued)



- **Maximum a posterior** (MAP) estimate: 0.27 (cf. ML estimate)
- 95% **credibility** interval: $[0.14, 0.45]$ (cf. 95% CI; eas**ier** interpretation).
- The prior and posterior are in the same 'family' of distributions: The Beta distribution is a **conjugate prior** of the Binomial distribution.

Bayesian inference — Model selection

- p-values are hard to interpret; Bayesian equivalent: *Bayes factor*.
- **Marginal likelihood** of M: $P(D|M) = \int \pi(q)P(q|D,M)dq$
- **Posterior probability** of M: $P(M|D) = \frac{P(D|M)\pi(M)}{P(D)}$.
- Two models: M_1 and M_0, Bayes factor: $K = \frac{P(D|M_1)}{P(D|M_0)} = \frac{P(M_1|D)\pi(M_1)}{P(M_0|D)\pi(M_0)}$
- If a priori M_1 and M_0 are equally likely, then K tells you **how much more likely M_1 is compared to M_0 after observing D**.
- Of course, *someone* came up with arbitrary (and widely used) thresholds (cf. $p < 0.05$)

Definition

$2\log(K)$	K	Strength of evidence
0 to 2	1 to 3	not worth more than a bare mention
2 to 6	3 to 20	positive
6 to 10	20 to 150	strong
> 10	> 150	very strong

Bayesian inference — Model selection

Example (fair or biased coin)

- M_0: The coin is fair: probability of heads is $\frac{1}{2}$.
- M_1: the coin could be biased: probability of heads is $q \sim \mathrm{Beta}(1,1)$.
- D: 100 coin tosses, 65 are head.
- $P(D|M_0) = \binom{100}{65}(\frac{1}{2})^{65}(1-\frac{1}{2})^{35} \approx 0.001$
- $P(D|M_1) = \binom{100}{65}\int_0^1 q^{65}(1-q)^{35}dq = \binom{100}{65}B(66,36) \approx 0.01$
- $2\log(K) \approx 4.9 \implies$ "positive evidence" for a biased coin

- In most real-life examples, Bayes factors are notoriously hard to compute
- Approximations: BIC, DIC (cf. AIC)

$$\mathrm{BIC} = -2 \cdot \log(\hat{L}) + k \cdot \log(n)$$

with $\hat{L}$ the maximum of the likelihood function.

MCMC

- Hard to compute Bayes factor, because it is hard to compute $\int P(D|q)\pi(q)dq$
- Hence: posterior only known up to a constant: $P(q|D) \propto \pi(q)p(D|q)$.
- Alternative: **Monte Carlo** integration.

Example (rejection sampling)

Approximate $\int_0^\infty f(x)dx$

1. find a distribution P on $[0, \infty)$ that is easy to sample from (!).
2. Find a constant M such that $f(x) \leq M \cdot P(x)$ for all $x \in [0, \infty)$ (!).
3. take a sample x from P and **accept** x with probability $f(x)/(M \cdot P(x))$.
4. repeat until N accepted samples $x_1, \ldots, x_N$.
5. $\int_0^\infty f(x)dx \approx \frac{1}{N} \sum_{i=1}^N f(x_i)$

MCMC

- Previous example: method for approximate integration, but also for sampling from the distribution $\bar{f}(x) = \frac{f(x)}{\int_0^\infty f(x')dx'}$.
- May not be very efficient: For instance when M needs to be large, many samples (from P) will be rejected.
- Alternative: **Markov Chain** Monte Carlo. Approximate $\bar{f}$ with the stationary distribution of a conveniently chosen Markov chain.

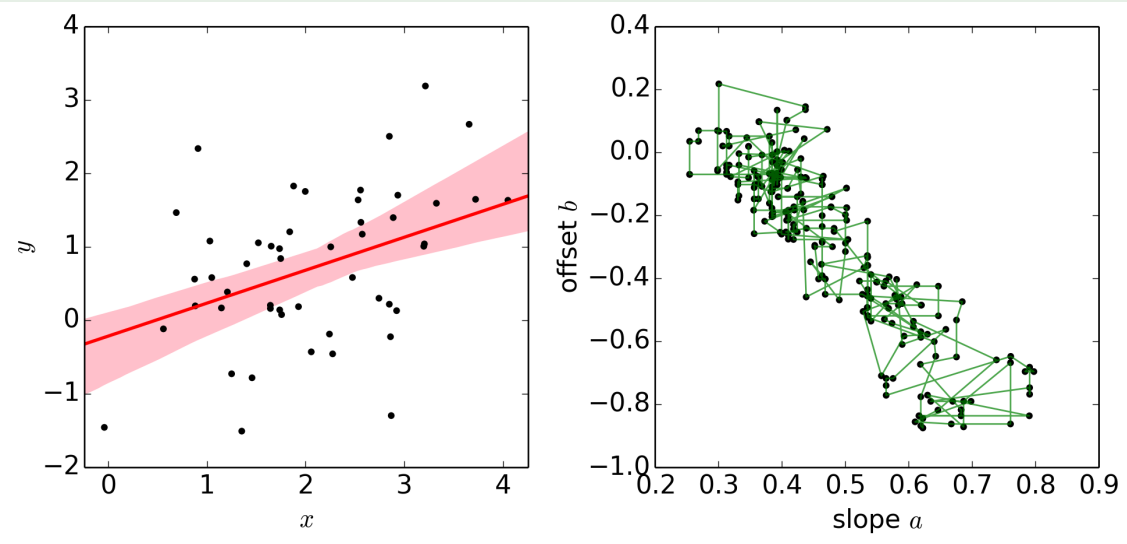
Example (Metropolis algorithm)

1. Given a sample q_t, sample a point q'_t from a **proposal distribution** $Q(\cdot|q_t)$. *Symmetry:* $Q(x|y) = Q(y|x)$.
2. With probability $\min(1, P(q'_t|D)/P(q_t|D))$, let $q_{t+1} = q'_t$ (accept), else $q_{t+1} = q_t$ (reject).

Symmetry of the proposal and particular choice for the acceptance probability $\implies$ **'detailed balance'** $\implies$ convergence to stationary distribution (given uniqueness).

Example (Gibbs sampling)

- Method for sampling from joint distributions; $q = (q^1, q^2, \ldots, q^k)$
- In each step, sample q^i_{t+1} from the distribution of q^i conditioned on $q^1_{t+1}, \ldots, q^{i-1}_{t+1}, q^{i+1}_t, \ldots, q^k_t$ (possibly using Metropolis)



Gibbs and Metropolis are special cases of the **Metropolis-Hastings** algorithm.

HMC

- Gibbs sampling is **not invariant** under rotation of the parameter space (sensitive to model parameterization).
- Has major problems with **correlated parameters** (e.g. background ILI and ILI reporting)
- A random walk is not the most efficient mode of transportation.
- **Hamiltonian Monte Carlo** (HMC) solves these problems...
- The parameter vector q represents a 'particle' subject to Hamiltonian dynamics (time reversible + conservation of 'volume').
- The particle has momentum p (for every q_i, introduce a p_i).
- $P(q|D)$ determines the 'potential energy'.
- HMC samples from a joint distribution of (q, p), then q is obtained by marginalizing.

HMC

1. At every step, sample p from a (Normal) distribution, independent from q
2. integrate the dynamics of (p, q) along a geodesic

$$\begin{aligned} \frac{dp}{dt} &= -\frac{\partial H}{\partial q} \\ \frac{dq}{dt} &= \frac{\partial H}{\partial p} \end{aligned} \tag{1}$$

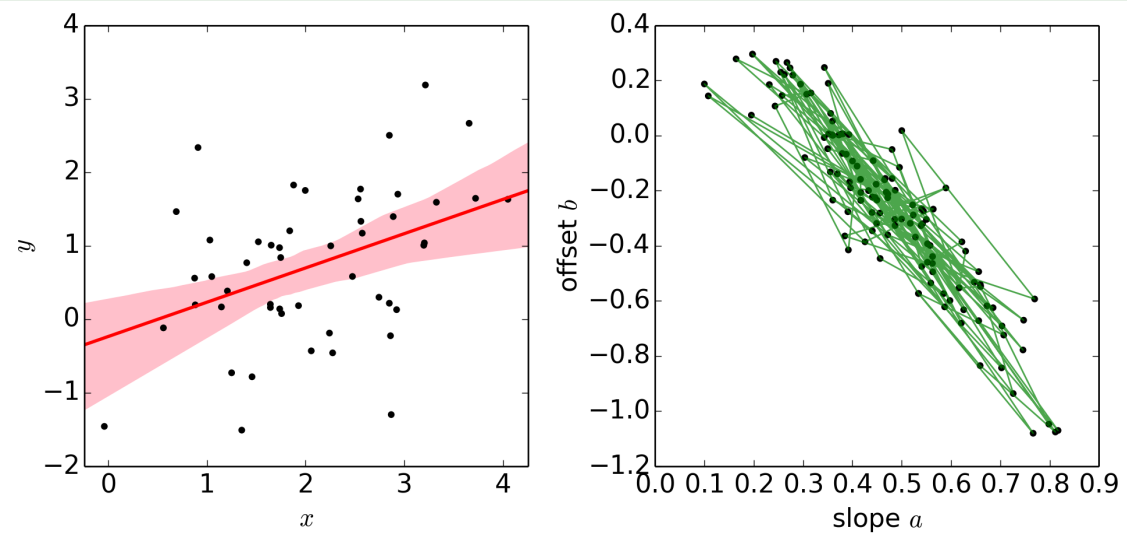
   where $H(q, p) = V(q) + T(p)$ (potential + kinetic energy) and $V(q) = -\log(P(q|D))$
3. Make a deleberate 'numerical error' disrupting conservation of energy, and accept the new state using a Metropolis step.

HMC needs a lot of fine-tuning to work properly. Needs gradient of likelihood function (!)

STAN

- STAN: software package for HMC; PyStan, RStan,...
- Uses **Automatic Differentiation** to calculate the gradient of the log-likelihood
- Built-in ODE integrator (work in progress)

Example



Much less "random walk behavior"; better **mixing**

STAN — code example

Example

```
data { /* D */
    int<lower=0> N;
    vector[N] xs;
    vector[N] ys;
}
parameters { /* q */
    real slope;
    real offset;
    real<lower=0> sigma;
}
model { /* pi and P */
    offset ~ normal(0, 1000);
    slope ~ normal(0, 1000);
    sigma ~ normal(0, 1000); /* sigma > 0 => Half-Normal */
    ys ~ normal(slope * xs + offset, sigma); /* vectorization */
}
```

Software

- BUGS (Bayesian inference Using Gibbs Sampling `http://www.mrc-bsu.cam.ac.uk/software/bugs/`
- JAGS (Just Another Gibbs Sampler) `http://mcmc-jags.sourceforge.net/`
- STAN (after Stanislaw Ulam) `http://mc-stan.org/`