BML. 2 Overview of Bayesian computational methods

DataLab CSIC

Brief description

The intro presented key methods in Bayesian inference and basic models and faced 'severe' computational problems quite rapidly (beta-binomial, logistic regression)

We overview and start with computational strategies to deal with those problems.

Here:

- Strategies based on asymptotic behaviour: asymptotic normality of posterior
- Monte Carlo strategies; including intro to MCMC

Later:

- Detailed intro to MCMC, Variational Bayes, INLA, SG-MCMC and hybrid forms

Serves also to further understand the behaviour of the posterior (and Bayesian inference at large)

Schedule

Recall computational problem

Overview strategies

Methods based on asymptotics

Monte Carlo

Lab 2.1 Comparing rough approximation and asymptotic approximation (bioassay)

Lab 2.2 MC, MC vs analytic vs asymptotic, Basic Gibbs sampler example

Sources:

arxiv 2112.10342

French and DRI (2000) Ch6 and 7

BDA3 (2014) Ch 4+App B

Overview

Computational problems in Bayesian analysis

Computing the posterior

$$p(\theta|x) = \frac{p(x|\theta)\, p(\theta)}{p(x)} = \frac{p(x|\theta)\, p(\theta)}{\int p(x|\theta)\, p(\theta)\, d\theta} \propto p(x|\theta)\, p(\theta)$$

Computing the predictive

$$p(y|x) = \int p(y|\theta)\, p(\theta|x)\, d\theta$$

Finding the optimal alternative

$$\max_a \int u(a,\theta)\, p(\theta|x)\, d\theta$$

$$\max_a \int u(a,y)\, p(y|x)\, dy$$

Computational problems in Bayesian analysis

Computing the posterior

$$p(\theta|x) = \frac{p(x|\theta)\, p(\theta)}{p(x)} = \frac{p(x|\theta)\, p(\theta)}{\int p(x|\theta)\, p(\theta)\, d\theta} \propto p(x|\theta)\, p(\theta)$$

Computing the predictive

$$p(y|x) = \int p(y|\theta)\, p(\theta|x)\, d\theta$$

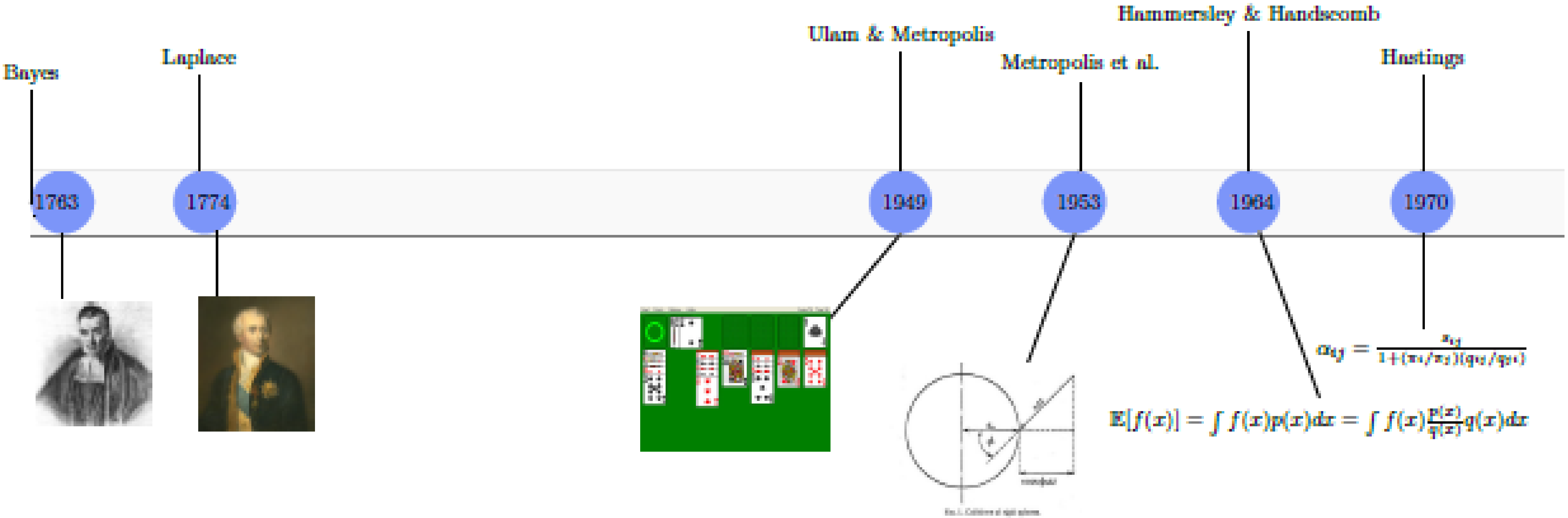
Finding the optimal alternative

$$\max_a \int u(a,\theta)\, p(\theta|x)\, d\theta$$

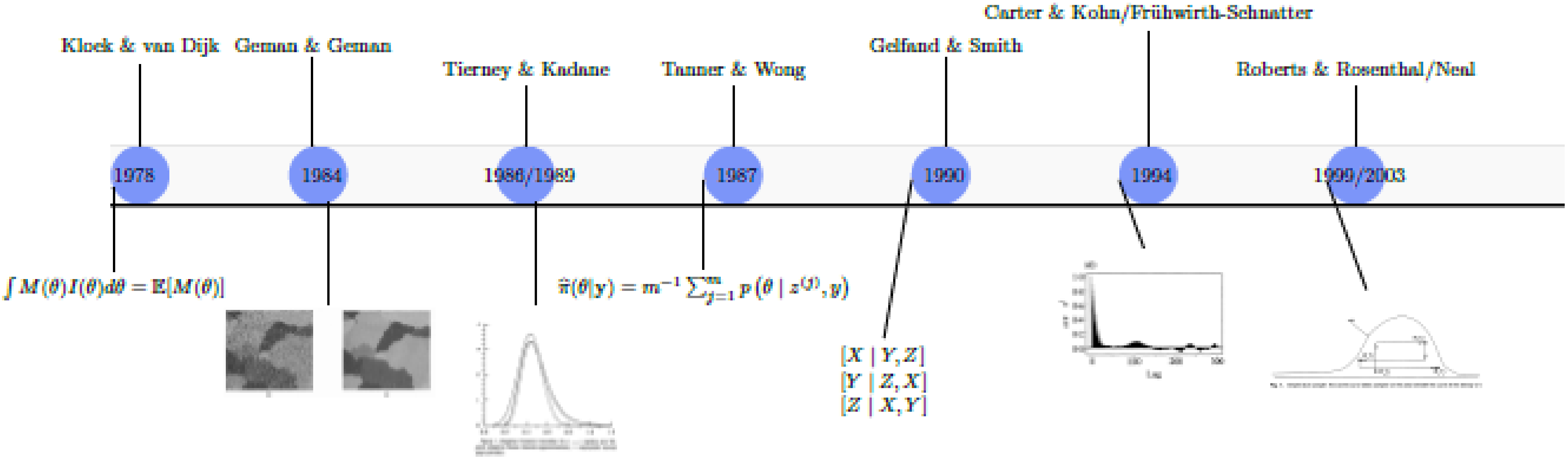
$$\max_a \int u(a,y)\, p(y|x)\, dy$$

Forget the last two for the momento!!!

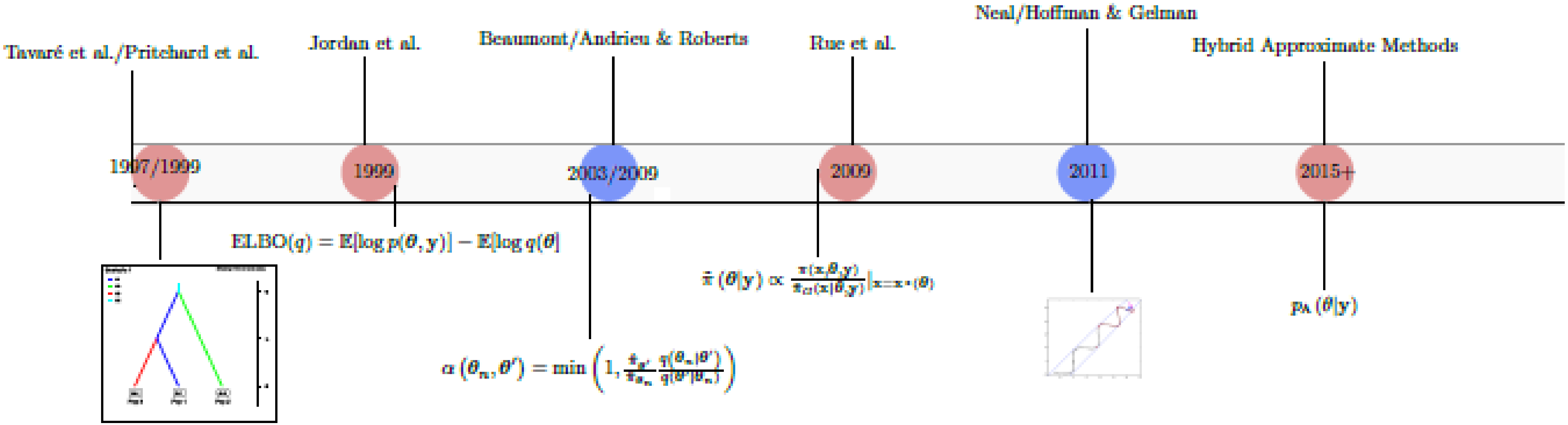
Quick tour over computational strategies



Quick tour over computational strategies



Quick tour over computational strategies



Conjugate models

Most examples so far conjugate

Prior and posterior from the same family

- Beta-binomial. Prior and posterior are beta
- Normal-normal. Prior and posterior are normal

Detailed table in

https://en.wikipedia.org/wiki/Conjugate_prior

The exponential family

Model

$$p(x_i \mid \theta) = h(x_i)\, g(\theta)\, e^{\phi(\theta)^t u(x_i)}$$

Likelihood

(incl. sufficient statistic)

$$p(x \mid \theta) = \left(\prod_{i=1}^{n} h(x_i)\right) g(\theta)^n \exp\left(\phi(\theta)^t \sum_{i=1}^{n} u(x_i)\right)$$

$$\propto g(\theta)^n \exp\left(\phi(\theta)^t\, t(x)\right) \qquad t(x) = \sum_{i=1}^{n} u(x_i)$$

Prior

$$p(\theta) \propto g(\theta)^\nu \exp\left(\phi(\theta)^t z\right)$$

Posterior

$$p(\theta \mid x) \propto p(\theta)\, p(x \mid \theta) \propto g(\theta)^{n+\nu} \exp\left(\phi(\theta)^t \left(t(x) + z\right)\right)$$

Example with beta-Bernoulli

$$f(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i} = 1 \cdot (1-\theta) \cdot \left(\frac{\theta}{1-\theta}\right)^{x_i}$$

$$= (1-\theta)\, e^{\phi x_i} \qquad \phi = \log\frac{\theta}{1-\theta}$$

$$f(x|\theta) = (1-\theta)^n \exp\left(\phi \sum_{i=1}^{n} x_i\right) \qquad t(x) = \sum_{i=1}^{n} x_i$$

$$f(\theta) \propto (1-\theta)^{\nu} \exp\left(\phi(\theta)^t z\right)$$

$$f(\theta|x) \propto (1-\theta)^{n+\nu} \exp\left(\phi(\theta)^t \left(\sum x_i + z\right)\right)$$

$$(1-\theta)^{\nu} \exp\left(\log\frac{\theta}{1-\theta}\, z\right) = (1-\theta)^{\nu} \frac{\theta^z}{(1-\theta)^z} = \theta^z (1-\theta)^{\nu - z}$$

Numerical example with beta for later reference

A computer manufacturer claims that 95% of its products do not require maintenance during the first year warranty. We buy 20 and 12 do require maintenance in the first year. Give point and interval estimates for the quality and test whether or not it is bigger than 95% with 0-1 utility

$\theta \sim \mathcal{B}e\,(4.75, 0.25)$ $\quad E(\theta) = .95$ $\quad \sigma(\theta) \simeq .08$

$mode(\theta) = 1$

$f(8|\theta) \propto \theta^{8}\,(1-\theta)^{12}$

$f(\theta|8) \propto \theta^{12.75}\,(1-\theta)^{12.25}$ $\quad \theta|8 \sim \mathcal{B}e\,(13.75, 13.25)$

$E(\theta|8) = \dfrac{13.75}{27} \simeq 0.51$ $\quad \sigma(\theta|8) \simeq 0.0009$

$med(\theta|8) \simeq 0.51$ $\quad mode(\theta|8) = 0.51$

$Pr(\theta \geq .95|x) \simeq 0 \longrightarrow$ REJECT H_0!!

Asymptotics and Laplace integration

Asymptotics of posterior: recall beta-binomial model

Point estimation

Posterior mean

$$E(\theta|x) = \frac{\alpha+x}{\alpha+\beta+n}$$

Mix of prior and data

$$\frac{n}{\alpha+\beta+n}\,\frac{x}{n} + \frac{\alpha+\beta}{\alpha+\beta+n}\,\frac{\alpha}{\alpha+\beta}$$

What if n grows??

$$\xrightarrow[n]{} \approx \frac{x}{n} \qquad Var(\theta|x) = \frac{(\alpha+x)(\beta+n-x)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} \xrightarrow[n]{} 0$$

Posterior median

$$med(\theta|x) \simeq \frac{\alpha+x-\frac{1}{3}}{\alpha+\beta+n-\frac{2}{3}} \qquad qbeta(0.5, \alpha+x, \alpha+\beta+n-x)$$

Posterior mode

$$mode(\theta|x) = \frac{\alpha+x-1}{\alpha+\beta+n-2}$$

Asymptotics of posterior: recall normal-normal model

Point estimation

Posterior mean

Mix of prior and data

What if n grows??

$$\frac{\frac{\sum x_i}{\sigma^2} + \frac{\theta_0}{\sigma_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} \frac{\sum x_i}{n} + \frac{\frac{1}{\sigma_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} \theta_0$$

$n \to \infty \quad \approx \quad \bar{x}$

Posterior median

SAME

Posterior mode

SAME

Asymptotics: Discrete case

Assume parameter space is discrete and there is a true underlying distribution.

For the given parametric model, consider the parameter minimising Kullback-Leibler divergence to the true distribution.

If minimiser has positive prior probability, the posterior concentrates over the minimiser as data accumulates.

$$KL(p\|q) = -\int p(x) \ln\left\{\frac{q(x)}{p(x)}\right\} dx = E\left(-\ln \frac{p(x)}{q(x)}\right)$$

$$= -\int p(x) \ln q(x)\, dx - \int -p(x) \ln p(x)\, dx$$

$$KL(p\|q) \geq 0 \ . \quad KL(p\|q) = 0 \iff p = q$$

$$KL(p\|q) \neq KL(q\|p)$$

Asymptotics: Discrete case

$\Theta = \{\theta_1, \dots, \theta_K\}$ $\quad x = (x_1, \dots, x_n)$

TRUE DISTRIBUTION $f(x) = \prod_{i=1}^{n} f(x_i)$

PRIOR $f(\theta)$ $\quad$ MODEL $f(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$

EXPECTED LOGARITHMIC SCORE $L(\theta) = -\int \log\left(f(x|\theta)\right) f(x)\, dx$

$\theta_0 = \arg\min L(\theta)$, EXISTS AND IS UNIQUE (ASS.)

$$\theta \neq \theta_0 \quad \log \frac{f(\theta|x)}{f(\theta_0|x)} = \log \frac{f(\theta)}{f(\theta_0)} + \sum_{i=1}^{n} \log \frac{f(x_i|\theta)}{f(x_i|\theta_0)} \quad (1)$$

$$E\left(\log \frac{f(x_i|\theta)}{f(x_i|\theta_0)}\right) = -L(\theta) + L(\theta_0) < 0$$

$$\text{SLLN} \quad \log \frac{f(\theta|x)}{f(\theta_0|x)} \xrightarrow[n]{} -\infty \Rightarrow \frac{f(\theta|x)}{f(\theta_0|x)} \xrightarrow[n]{} 0 \Rightarrow f(\theta|x) \xrightarrow[n]{} 0$$

$$f(\theta_0|x) = 1 - \sum_{\theta \neq \theta_0} f(\theta|x) \xrightarrow[n]{} 1.$$

Asymptotics: Continuous case

If parameter space is compact and A is a neigborhood of minimiser with positive prior probability, the posterior concentrates in A as data accumulates.

Convert it to the discrete case

Asymptotics: normality with posterior mode I

Under certain regularity conditions, the posterior approaches normality around the posterior mode as data accumulates

$\hat{\theta}$ = mode $p(\theta|x)$ UNIMODAL, APP. SYMMETRIC AROUND $\hat{\theta}$

$$\log p(\theta|x) = \log p(\hat{\theta}|x) + \frac{1}{2}(\theta-\hat{\theta})'\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(\theta|x)\right)_{\theta=\hat{\theta}}(\theta-\hat{\theta}) + r(\theta)$$

$$p(\theta|x) \underset{\sim}{\propto} \exp\left(-\frac{1}{2}(\theta-\hat{\theta})'\left(-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(\theta|x)\right)(\theta-\hat{\theta})\right) \approx N(\hat{\theta}, V)$$

V INVERSE OF $-$HESSIAN OF $\log p(\theta|x)$ AT $\hat{\theta}$

AS $n \to \infty$, $p(\theta|x)$ CONCENTRATES AROUND $\hat{\theta} \to \theta_0$, $r(\theta) \to 0$

$$-\frac{\partial^2}{\partial\theta_j\theta_k}\log p(\theta|x)\Big|_{\theta=\hat{\theta}} = \underbrace{-\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log p(\hat{\theta})}_{\sigma} - \sum_{i=1}^{n}\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log p(x_i|\theta)\Big|_{\theta=\hat{\theta}}$$

$$\downarrow$$

$$nI(\theta_0|X)^{-1}$$

Asymptotics: normality with posterior mode II

Parameter space in R^p

Mode unique

Mode interior to parameter space

Density continuous and positive at mode

Log-likelihood twice differentiable at neighbourhood of mode

Asymptotics: normality with MLE

As sample size increases, likelihood dominates (a non-degenerate) prior

Prior behaves as approximately constant

Posterior mode approximated by MLE

Fisher's info matrix at MLE approximates variance

Back to the beta numerical example

$$\theta \sim \mathcal{B}e(4.75, 0.25) \longrightarrow \theta | x \sim \mathcal{B}e(13.75, 13.25)$$

$$\Pr\left(\theta \in [0.325, 0.692]\right) = .95$$

$$\text{MLE} = \frac{x}{n} = \frac{12}{20} = 0.6$$

$$\sigma^2 = \frac{1}{n}\frac{x}{n}\left(1-\frac{x}{n}\right) = \frac{1}{20}\frac{12}{20}\left(1-\frac{12}{20}\right) = .012$$

$$\left[\frac{x}{n} \pm z_{\alpha/2}\sqrt{\frac{1}{n}\frac{x}{n}\left(1-\frac{x}{n}\right)}\right] \simeq [.39, .81]$$

Usage

Approximate posterior by normal

Moments and regions as in normal

A beta numerical example

$\theta \sim \mathcal{B}e(1,1)$

$n = 10, \quad x = 3$

MLE $\hat{\theta} = 3/10$

$\frac{\partial^2}{\partial \theta^2} f(x|\theta) \quad = \frac{3}{\theta^2} + \frac{7}{(1-\theta)^2}$

$\text{Var} = 21/1000$

$\mathcal{B}e(4,8) \quad \text{vs} \quad N(0.3, 0.021)$

Laplace integration

$g(\theta) > 0$

$$E_{\theta|x}(g(\theta)) = \int g(\theta)\, f(\theta|x)\, d\theta = \frac{\int g(\theta)\, f(x|\theta)\, f(\theta)\, d\theta}{\int f(x|\theta)\, f(\theta)\, d\theta}$$

$$\hat{\theta} = \arg\max G(\theta)$$

$$G(\theta) = \log\left(f(x|\theta)\, f(\theta)\right)$$

$$\theta^* = \arg\max G^*(\theta)$$

$$G^*(\theta) = \log\left(g(\theta)\, f(x|\theta)\, f(\theta)\right)$$

$$E_{\theta|x}(g(\theta)) \simeq \sqrt{\frac{\det(V^*)}{\det(V)}}\; \frac{g(\theta^*)\, f(x|\theta^*)\, f(\theta^*)}{f(x|\hat{\theta})\, f(\hat{\theta})}$$

$$V^*_{ij} = -\left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} G^*(\theta)\right)_{\theta^*} \qquad V_{ij} = -\left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} G(\theta)\right)_{\hat{\theta}}$$

Summary

In some contexts, as data accumulates posterior may be approximated
by a normal distribution

Sometimes, but only sometimes, this provides sufficiently good approximations of the required posterior or predictive integrals. At least serves as a benchmark…. Check the lab

… We still need more powerful approaches

…. But we shall recover these ideas when talking about variational Bayes

Numerical integration

Strategies so far

- Conjugate models
- Posterior asymptotics to normality
- Laplace integration

Insufficient for modern statistics and machine learning

Numerical integration. Brief recall

Problem

$$I_s = \int_{[0,1]^s} f(u)\, du$$

s-dimensional trapezium rule

$$I_s \simeq \sum_{n_1=0}^{m} \cdots \sum_{n_s=0}^{m} w_{n_1} \cdots w_{n_s}\, f\left(\frac{n_1}{m}, \dots, \frac{n_s}{m}\right)$$

$$w_0 = w_m = 1/2m \qquad w_n = 1/m \quad 1 \le n \le m-1$$

error analysis

$$\frac{\partial^2 f}{\partial u_i^2} \text{ cont in } [0,1]^s \longrightarrow O(m^{-2})$$

$$N = (m+1)^s \longrightarrow O(N^{-2/s})$$

$$s = 5,\ \varepsilon \le 10^{-2} \Longrightarrow N = 10^5\ !!!$$

Dependence of error bound on dimension is typical!!!!

Monte Carlo integration

Monte Carlo integration. Brief recall

Problem

$$I_s = \int_{[0,1]^s} f(u)\, du$$

Deterministic problem recast as stochatic (Monte Carlo)

$$I_s = E(f)$$

Monte Carlo integration. Brief recall

Suggested strategy

Sample $u_1, \dots, u_N \sim \mathcal{U}[0,1]^s$

Do $\hat{I}_s = \frac{1}{N} \sum_{i=1}^{N} f(u_i)$

Monte Carlo integration. Brief recall

Analysis. SLLN

$$\hat{I}_s = \frac{1}{N}\sum_{i=1}^{N} f(u_i) \xrightarrow{a.s.} E(f) = \boxed{I_s}$$

$$\int_{[0,1]^d} (\hat{I}_s - I_s)^2 \, du = \frac{\sigma^2(f)}{N} = Var(\hat{I}_s)$$

Error bounds

$$\sigma^2(f) = \int_{[0,1]^d} (f - E(f))^2 \, du \qquad \boxed{\int_{[0,1]^d} f^2(u)\, du}$$

CLT prob. error bounds

$$Pr\left(\frac{c_1\sigma(f)}{\sqrt{N}} \le \hat{I}_s - I_s \le \frac{c_2\sigma(f)}{\sqrt{N}}\right) \xrightarrow[N]{} \Phi(c_2) - \Phi(c_1)$$

SE

$$EE(\hat{I}_s) = \frac{1}{\sqrt{N}} \sqrt{\frac{\sum (f(u_i) - \hat{I}_s)^2}{N-1}}$$

MC vs trapezium

$$O\left(N^{-\frac{1}{2}}\right) \quad \text{vs} \quad O\left(N^{-\frac{2}{s}}\right)$$

This is general. As dimension grows, numerical gets less efficient… but MC's efficiency is dimension independent!!!

MC. Generalization

Problem

$$I_g = \int f(u)\, g(u)\, du = E_g(f)$$

Strategy

Sample $u_1, \dots, u_N \sim g$

Do $\hat{I}_g = \frac{1}{N} \sum_{i=1}^{N} f(u_i)$

MC. Generalization. Example

Problem

$$I = \int_{-\infty}^{\infty} (x+x^2)\, g(x)\, dx \qquad g \sim N(\mu=1, \sigma=2)$$

$$I = E(x+x^2) = E(x)+E(x^2) = 1+(\sigma^2+\mu^2) = 6$$

```
x <- rnorm(30,1,2)
int <- mean(x)
err <- 1/sqrt(30) * sd(x)
```

MC. Random variate generation

Inversion method

$\mathrm{cdf}(X) \longrightarrow F$

$U \sim \mathcal{U}(0,1) \longrightarrow F^{-1}(U) \sim X$

$P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$

Sample $U \sim \mathcal{U}(0,1)$
Do $X = F^{-1}(U)$
Output X

Example

$X|\lambda \sim \mathcal{E}xp(\lambda)$

$F(x|\lambda) = \begin{cases} 0, & x \leq 0 \\ 1-e^{-\lambda x}, & x \geq 0 \end{cases}$

$X = -\frac{1}{\lambda} \ln(1-U)$

Sample $U \sim \mathcal{U}(0,1)$
Do $X = -\ln U / \lambda$
Output X

MC. Random variate generation

Inversion method

Use of transformations

Rejection

Ratio of uniforms

Pretests

Composition

At the core of, e.g., r***dist*** functions in R…. but not sufficient

Markov chain Monte Carlo intro

General idea

Objective

$$I_g = \int f(x)\, g(x)\, dx = E_g(f)$$

Difficult or inefficient to sample from g

General idea

Markov chain X_n with same state space and convergent to target distribution g

$$X_n \xrightarrow{d} g$$

Strategy

Initialise x_0, $n=1$

Until convergence, Generate $X_n \mid X_{n-1}=x_{n-1}$, $n=n+1$, n^*

Until n^*+m, Generate and collect $X_n \mid X_{n-1}=x_{n-1}$, $n=n+1$

$$\hat{I}_g \approx \frac{1}{m}\sum_{n^*}^{n^*+m} f(x_i)$$

Problem

So how do we ‘invent’ such Markov chains?

Motivating Gibbs sampler

(X,Y) Bernoulli variables with joint distribution

X	Y	$P(X,Y)$
0	0	p_1
1	0	p_2
0	1	p_3
1	1	p_4

Compute the marginals of X and Y

Compute the conditionals

Motivating Gibbs sampler

The conditionals are characterised by

$$A_{yx} = \begin{pmatrix} P(Y=0|X=0) & P(Y=1|X=0) \\ P(Y=0|X=1) & P(Y=1|X=1) \end{pmatrix} = \begin{pmatrix} \frac{p_1}{p_1+p_3} & \frac{p_3}{p_1+p_3} \\ \frac{p_2}{p_2+p_4} & \frac{p_4}{p_2+p_4} \end{pmatrix}$$

$$A_{xy} = \begin{pmatrix} \frac{p_1}{p_1+p_2} & \frac{p_2}{p_1+p_2} \\ \frac{p_3}{p_3+p_4} & \frac{p_4}{p_3+p_4} \end{pmatrix}$$

Motivating Gibbs sampler

Consider the sampling scheme

Initialise $Y_0 = y_0$, $i = 1$

Do

  Sample $X_i \sim X \mid Y = y_{i-1}$

  Sample $Y_i \sim Y \mid X = x_i$

  $i = i + 1$

Motivating Gibbs sampler

X_n a Markov chain with transition matrix

$$A = A_{yx} A_{xy}$$

Motivating Gibbs sampler

Convergence of X_n

$$X_n \xrightarrow{d} X$$

$$(p_1+p_3 \quad p_2+p_4) = (p_1+p_3 \quad p_2+p_4)\, A$$

Similarly,

$$Y_n \xrightarrow{d} Y \qquad (X_n, Y_n) \xrightarrow{d} (X, Y)$$

Motivating Gibbs sampler

General Gibbs sampler

$$X = (X_1, \dots, X_p) \sim \pi$$

Sample from $X_s \mid X_{-s} = (X_1, \dots, X_{s-1}, X_{s+1}, \dots, X_p)$

Initialize $X_1^0, \dots, X_p^0$, $i=1$

Iterate

Sample $X_1^i \sim X_1 \mid X_2^{i-1}, \dots, X_p^{i-1}$

Sample $X_2^i \sim X_2 \mid X_1^i, X_3^{i-1}, \dots, X_p^{i-1}$

...

Sample $X_p^i \sim X_p \mid X_1^i, X_2^i, \dots, X_p^{i-1}$

$i = i+1$

Motivating Gibbs sampler

Example

$$\pi(x_1, x_2) = \frac{1}{\pi} e^{-x_1(1+x_2^2)} \qquad (x_1, x_2) \in (0, \infty) \times (-\infty, \infty)$$

Motivating Gibbs sampler

Example

$$\pi(x_1, x_2) = \frac{1}{\pi} e^{-x_1(1+x_2^2)}$$

$$(x_1, x_2) \in (0, \infty) \times (-\infty, \infty)$$

$$\pi(x_1|x_2) = \frac{\pi(x_1, x_2)}{\pi(x_2)} \propto \pi(x_1, x_2) \propto e^{-x_1(1+x_2^2)}$$

$$X_1|X_2 = x_2 \sim \mathcal{E}xp(1 + x_2^2)$$

$$\pi(x_2|x_1) \propto \pi(x_1, x_2) \propto e^{-x_1 x_2^2},$$

$$X_2|X_1 = x_1 \sim \mathcal{N}\left(0, \sigma^2 = \frac{1}{2x_1}\right)$$

Motivating Gibbs sampler

Example

Choose initial x_2^0, $i=1$

Until convergence

Sample $x_1^i \sim Exp(1+(x_2^{i-1})^2)$

Sample $x_2^i \sim N\left(0, \frac{1}{2x_1^i}\right)$

$i=i+1$

Summary and to be seen

We have described the basic MCMC strategy with an intro to Gibbs sampling

Gibbs sampling, Metropolis-Hastings, Hybrid

Hamiltonian Monte Carlo

Reversible jumper

…. But we shall require other strategies with very large problems