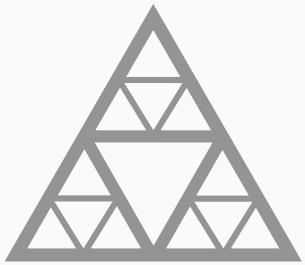


École des Ponts

ParisTech

Hamiltonian Monte Carlo methods

A Riemannian geometry perspective

Emile Mathieu, Kimia Nadjahi

April 11, 2017

Ecole des Ponts ParisTech

Problem Statement

Problem Statement

Intractable density:

$p(\boldsymbol{\beta}) = \tilde{p}(\boldsymbol{\beta}) / \int \tilde{p}(\boldsymbol{\beta}) d\boldsymbol{\beta}$

Metropolis-Hastings:

- Define an ergodic Markov process with stationary distribution $p(\boldsymbol{\beta})$.
- Transitions $\boldsymbol{\beta} \mapsto \boldsymbol{\beta}^*$ proposed with density $q(\boldsymbol{\beta}^* | \boldsymbol{\beta})$ accepted with probability

$$\alpha(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = \min \left\{ 1, \frac{\tilde{p}(\boldsymbol{\beta}^*) q(\boldsymbol{\beta} | \boldsymbol{\beta}^*)}{\tilde{p}(\boldsymbol{\beta}) q(\boldsymbol{\beta}^* | \boldsymbol{\beta})} \right\}$$

What proposal distribution q ?

Typically, random walk: $q(\boldsymbol{\beta}^* | \boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}^* | \boldsymbol{\beta}, \boldsymbol{\Lambda})$

Low $\|\boldsymbol{\Lambda}\|$	**High** $\|\boldsymbol{\Lambda}\|$
High acceptance rate	Low acceptance rate
Highly correlated samples	Not so many correlated samples

Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (I)

Goal:

Make large transitions accepted with high probability.

How to:

- Independent auxiliary variable: $\boldsymbol{p} \sim \mathcal{N}(\boldsymbol{p}|\mathbf{0}, \boldsymbol{M})$
- Joint density: $p(\boldsymbol{\beta}, \boldsymbol{p}) = p(\boldsymbol{\beta})p(\boldsymbol{p}) = p(\boldsymbol{\beta})\mathcal{N}(\boldsymbol{p}|\mathbf{0}, \boldsymbol{M})$
- The negative joint log-probability is (with $\mathcal{L}(\boldsymbol{\beta}) = \log\{p(\boldsymbol{\beta})\}$)

$$\boldsymbol{H}(\boldsymbol{\beta}, \boldsymbol{p}) = \underbrace{-\mathcal{L}(\boldsymbol{\beta})}_{\text{potential energy}} + \frac{1}{2}\log\{(2\pi)^D|\boldsymbol{M}|\} + \underbrace{\frac{1}{2}\boldsymbol{p}^T\boldsymbol{M}^{-1}\boldsymbol{p}}_{\text{kinetic energy}} \qquad (1)$$

where $\mathcal{L}(\boldsymbol{\beta}) = \log\{p(\boldsymbol{\beta})\}$

Hamiltonian Monte Carlo (II)

- Hamilton's equations:

$$\frac{d\boldsymbol{\beta}}{d\tau} = \frac{\partial \boldsymbol{H}}{\partial \boldsymbol{p}} = \boldsymbol{M}^{-1}\boldsymbol{p}, \quad \frac{d\boldsymbol{p}}{d\tau} = -\frac{\partial \boldsymbol{H}}{\partial \boldsymbol{\beta}} = \nabla_{\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta}) \tag{2}$$

- Numerical integrator (Stormer-Verlet or leapfrog):

$$\boldsymbol{p}^{\tau+\frac{\epsilon}{2}} = \boldsymbol{p}^{\tau} + \frac{\epsilon}{2}\nabla_{\boldsymbol{\beta}}\mathcal{L}\{\boldsymbol{\beta}^{\tau}\} \tag{3}$$

$$\boldsymbol{\beta}^{\tau+\epsilon} = \boldsymbol{\beta}^{\tau} + \epsilon \boldsymbol{M}^{-1}\boldsymbol{p}^{\tau+\frac{\epsilon}{2}} \tag{4}$$

$$\boldsymbol{p}^{\tau+\epsilon} = \boldsymbol{p}^{\tau+\frac{\epsilon}{2}} + \frac{\epsilon}{2}\nabla_{\boldsymbol{\beta}}\mathcal{L}\{\boldsymbol{\beta}^{\tau+\epsilon}\} \tag{5}$$

- Acceptance probability: $\min(1, \exp\{-H(\boldsymbol{\beta}^*, \boldsymbol{p}^*) + H(\boldsymbol{\beta}, \boldsymbol{p})\})$

Hyperparameters:

- Step size ϵ and number of integration steps: via acceptance rate
- Yet, choice of mass matrix $\boldsymbol{M}$ is critical.

Geometric concepts

Geometric concepts

Goal:

Automatically determine $\boldsymbol{M}$ at each step.

Fisher-Rao metric:

Distance between parametrized density functions

$$\mathrm{KL}(p(\boldsymbol{y};\beta)||p(\boldsymbol{y};\beta+\delta\beta)) \simeq \delta\beta^T \boldsymbol{G}(\beta)\delta\beta$$

where $\boldsymbol{G}(\beta) = -\mathbb{E}_{y|\beta}\left[\frac{\partial^2}{\partial\beta^2}\log\{p(\boldsymbol{y}|\beta)\}\right]$

Fisher information matrix $\boldsymbol{G}(\beta)$ is p.d. metric defining a Riemann manifold.

General metric tensor

With a Bayesian perspective:

$$\boldsymbol{G}(\beta) = -\mathbb{E}_{y|\beta}\left[\frac{\partial^2}{\partial\beta^2}\log\{p(\boldsymbol{y},\beta)\}\right]$$

Riemann manifold Hamiltonian Monte Carlo

Riemann manifold Hamiltonian Monte Carlo (I)

- The Hamiltonian is

$$\boldsymbol{H}(\boldsymbol{\beta}, \boldsymbol{p}) = -\mathcal{L}(\boldsymbol{\beta}) + \frac{1}{2}\log\{(2\pi)^D|\boldsymbol{G}(\boldsymbol{\beta})|\} + \frac{1}{2}\boldsymbol{p}^T\boldsymbol{G}(\boldsymbol{\beta})^{-1}\boldsymbol{p} \quad (6)$$

- Marginal density:

$$p(\boldsymbol{\beta}) \propto \int \exp\{-\boldsymbol{H}(\boldsymbol{\beta}, \boldsymbol{p})\}d\boldsymbol{p} = \exp\{\mathcal{L}(\boldsymbol{\beta})\}$$

- Joint density: $p(\boldsymbol{\beta}, \boldsymbol{p}) = p(\boldsymbol{\beta})p(\boldsymbol{p}|\boldsymbol{\beta}) = p(\boldsymbol{\beta})\mathcal{N}(\boldsymbol{p}|\boldsymbol{0}, \boldsymbol{G}(\boldsymbol{\beta}))$
- Hamilton's equations:

$$\frac{d\beta_i}{d\tau} = \frac{\partial \boldsymbol{H}}{\partial p_i} = \{\boldsymbol{G}(\boldsymbol{\beta})^{-1}\boldsymbol{p}\}_i \quad (7)$$

$$\frac{dp_i}{d\tau} = -\frac{\partial \boldsymbol{H}}{\partial \beta_i} = \frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_i} - \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{G}(\boldsymbol{\beta})^{-1}\frac{\partial \boldsymbol{G}(\boldsymbol{\beta})}{\partial \beta_i}\right\} + \frac{1}{2}\boldsymbol{p}^T\boldsymbol{G}(\boldsymbol{\beta})^{-1}\frac{\partial \boldsymbol{G}(\boldsymbol{\beta})}{\partial \beta_i}\boldsymbol{G}(\boldsymbol{\beta})^{-1}\boldsymbol{p} \quad (8)$$

Riemann manifold Hamiltonian Monte Carlo (II)

Numerical integrator (generalized leapfrog):

$$\boldsymbol{p}^{\tau+\frac{\epsilon}{2}} = \boldsymbol{p}^{\tau} - \frac{\epsilon}{2}\nabla_{\beta}\boldsymbol{H}\left\{\beta^{\tau}, \boldsymbol{p}^{\tau+\epsilon/2}\right\} \quad (9)$$

$$\beta^{\tau+\epsilon} = \beta^{\tau} + \frac{\epsilon}{2}\left[\nabla_{\boldsymbol{p}}\boldsymbol{H}\left\{\beta^{\tau}, \boldsymbol{p}^{\tau+\frac{\epsilon}{2}}\right\} + \nabla_{\boldsymbol{p}}\boldsymbol{H}\left\{\beta^{\tau+\epsilon}, \boldsymbol{p}^{\tau+\frac{\epsilon}{2}}\right\}\right] \quad (10)$$

$$\boldsymbol{p}^{\tau+\epsilon} = \boldsymbol{p}^{\tau+\frac{\epsilon}{2}} - \frac{\epsilon}{2}\nabla_{\beta}\boldsymbol{H}\left\{\beta^{\tau+\epsilon}, \boldsymbol{p}^{\tau+\frac{\epsilon}{2}}\right\} \quad (11)$$

Acceptance probability:

$$\min(1, \exp\{-H(\beta^{*}, \boldsymbol{p}^{*}) + H(\beta, \boldsymbol{p})\})$$

Application: Bayesian logistic regression

Bayesian logistic regression (BLR)

Probabilistic model

- Likelihood:

$$p(\boldsymbol{y} = \mathbf{1}|\boldsymbol{X}) = \eta(\boldsymbol{X}^T\beta)$$

$$\text{with } \eta : x \mapsto (1 + e^{-x})^{-1}$$

- Prior: $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \alpha\boldsymbol{I})$
- Metric tensor:

$$\boldsymbol{G}(\beta) = \boldsymbol{X}^T\boldsymbol{\Lambda}\boldsymbol{X} + \alpha^{-1}\boldsymbol{I}$$

$$\text{with } \Lambda_{i,i} = \eta(\boldsymbol{\beta}^T\boldsymbol{X}_{n,.}^T)\{1 - \eta(\boldsymbol{\beta}^T\boldsymbol{X}_{n,.}^T)\}$$

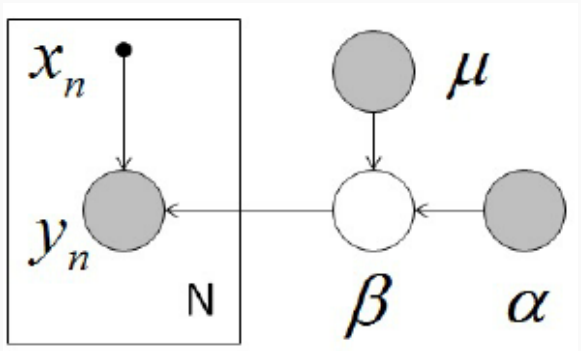


Figure 1: DAG for the Bayesian logistic regression.

Iterative Weighted Least Squares (IWLS)

- MLE of BLR obtained with Newton-Raphson method:

$$\hat{\boldsymbol{\beta}}^{(t)} = ((\alpha \boldsymbol{I})^{-1} + \boldsymbol{X}^T \boldsymbol{\Lambda} \boldsymbol{X})^{-1} (\boldsymbol{X}^T \boldsymbol{\Lambda} \tilde{\boldsymbol{y}}(\hat{\boldsymbol{\beta}}^{(t-1)}))$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}^{(t)} = \boldsymbol{G}(\boldsymbol{\beta})^{-1} = ((\alpha \boldsymbol{I})^{-1} + \boldsymbol{X}^T \boldsymbol{\Lambda} \boldsymbol{X})^{-1}$$

  where $\tilde{\boldsymbol{y}}(\hat{\boldsymbol{\beta}}^{(t-1)}) = \boldsymbol{X} \hat{\boldsymbol{\beta}}^{(t-1)} + \boldsymbol{\Lambda}^{-1}(\boldsymbol{y} - \eta(\boldsymbol{\beta}^T \boldsymbol{X}))$

- Combination of the MCMC and IWLS iteration schemes:
  1. Initialization: $\beta = \beta^{(0)}$, $t = 1$
  2. Sample β_{new} from $\mathcal{N}(\hat{\beta}^{(t)}, \hat{\Sigma}_{\beta}^{(t)})$
  3. Accept it with probability $\alpha(\beta^{(t-1)}, \beta_{new})$ and set $\beta^{(t)} = \beta_{new}$; otherwise, $\beta^{(t)} = \beta^{(t-1)}$
  4. Do $t := t + 1$ and return to Step 2

Auxiliary variable Gibbs sampler

Representation of BLR with auxiliary variables [Helmes and Hold, 2005]:

$$
\begin{aligned}
y_i &= sgn(z_i) \\
z_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \epsilon_i \\
\epsilon_i &\sim \mathcal{N}(0, \lambda_i) \\
\lambda_i &= (2\psi_i)^2 \\
\psi_i &\sim KS \\
\boldsymbol{\beta} &\sim \mathcal{N}(\mathbf{0}, \alpha\boldsymbol{I})
\end{aligned}
$$

$\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{\lambda} \sim \mathcal{N}(\boldsymbol{B}, \boldsymbol{V})$ with $\boldsymbol{B}, \boldsymbol{V}$ as in WLS, $z_i|\boldsymbol{\beta}, \boldsymbol{X}_i, y_i, \lambda_i \sim$ truncated normal, $\lambda_i|z_i, \boldsymbol{\beta}$ sampled with rejection sampling

Block Gibbs sampler:

1. Update $\{\boldsymbol{z}, \boldsymbol{\beta}\}$ jointly given $\boldsymbol{\lambda}$
2. Update $\boldsymbol{\lambda}|\boldsymbol{z}, \boldsymbol{\beta}$

Adaptive Metropolis-Hastings

Component-wise random symmetric walk

At iteration i for component k:

1. Sample $\tilde{\beta}^k_{i+1} \sim \mathcal{N}(\beta^k_i, \sigma_k)$
2. Set $\beta^k_{i+1} = \tilde{\beta}^k_{i+1}$ with probability $\tilde{p}(\tilde{\boldsymbol{\beta}}_{i+1})/\tilde{p}(\boldsymbol{\beta}_i)$
3. Otherwise $\beta^k_{i+1} = \beta^k_i$

Component-wise adaptive variance

Every 100 samples after burn-in:

- If $currentAcceptanceRate^k > \gamma_{max}$
  $$\sigma_k = 1.2 * \sigma_k$$
- Else if $currentAcceptanceRate^k < \gamma_{min}$
  $$\sigma_k = 0.8 * \sigma_k$$

Results

Effective Sample Size (ESS)

$$ESS = \frac{N}{1 + 2\sum_k \gamma(k)}$$

where N number of samples after burn-in, $\gamma(k)$ autocorrelation of lag k
Ideally, $ESS = N$.

Trade-off time and quality: ratio time / min(ESS)

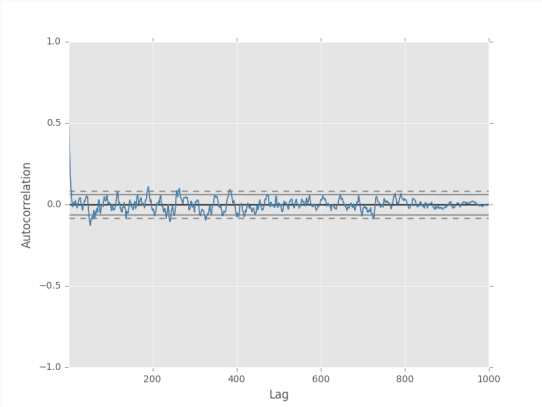
Results

Averaged results: sampling experiments repeated 10 times

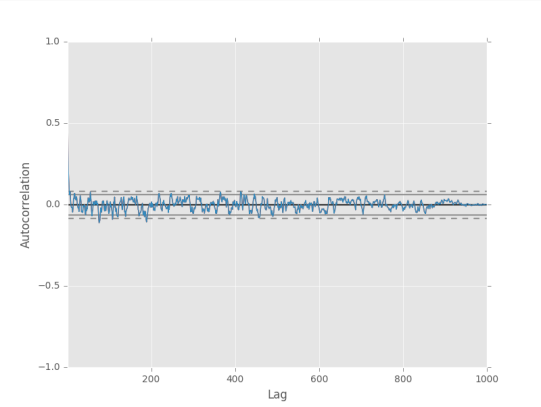
Method	*Time*	*ESS (min, median, max)*	*time / min(ESS)*
Metropolis	10.4	(305.5, 653.6, 801)	0.034
Auxiliary variables	589.4	(710.6, 1199.4, 1715.2)	0.83
HMC	44.9	(3349, 3634, 4141)	0.0134
IWLS	17.9	(21.78, 96.34, 322.4)	0.83
RMHMC	164.1	(4865, 5000, 5000)	0.034

Table 1: Heart data set – comparison of sampling methods

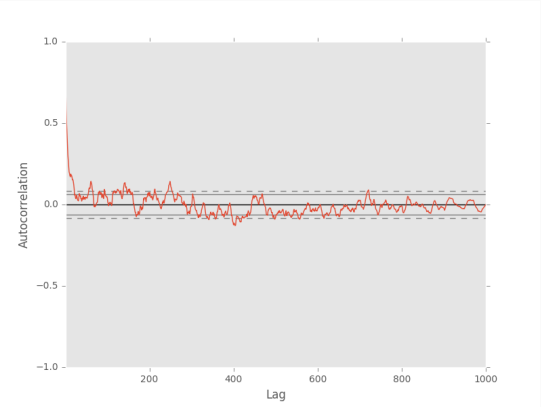
Results



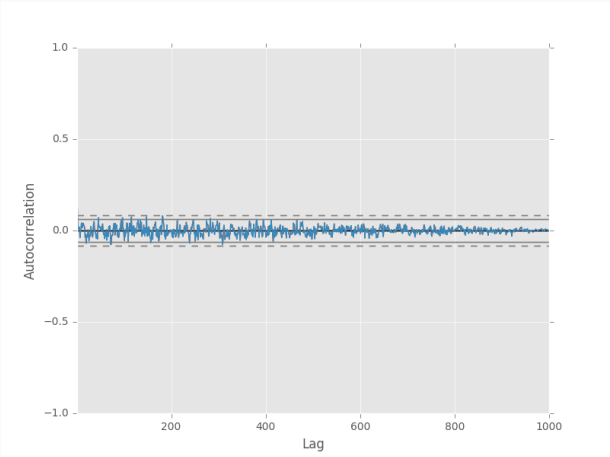
(a) Adaptive MH



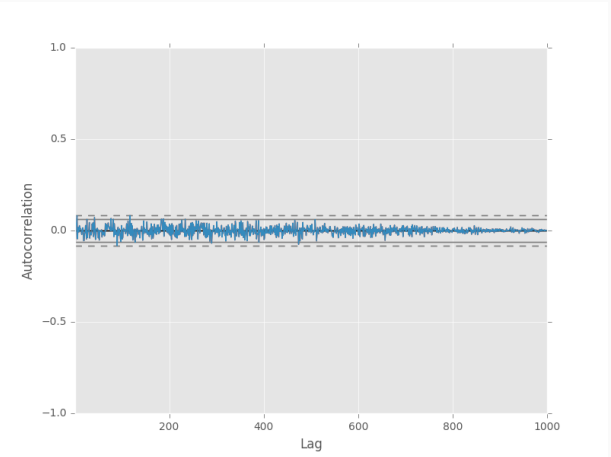
(b) Auxiliary Gibbs



(c) IWLS



(d) HMC



(e) RMHMC

Figure 2: Autocorrelation plots: 1st covariate, 1000 samples of Heart data

Conclusion

Thank you for your attention !

Questions?