

Manifold MCMC for Mixture Models

Vassilios Stathopoulos Mark A. Girolami

Department of Statistical Science
University College London

April 2011

- 1 Manifold MCMC
 - Metropolis-Hastings
 - Metropolis Adjusted Langevin Algorithm
 - Manifold Metropolis Adjusted Langevin Algorithm
 - Manifold Hamiltonian Monte Carlo

- 2 Finite Gaussian Mixture Models
 - Model re-parameterisation
 - Approximations of the Metric Tensor
 - Examples

- 3 Ad-Mixtures
 - Model re-parameterisation
 - Metropolis within Gibbs scheme
 - Example

- 4 Conclusions and Discussion

Aim: sample from intricate distribution $\pi(\theta)$ – Bayesian: $\pi(\theta) \propto p(\theta)p(x|\theta)$.

Metropolis-Hastings Algorithm

At initialization $t = 0$,

- 1 set θ^0 arbitrarily

At each iteration $t \geq 1$

- 1 sample (θ^*) from $q(\cdot|\theta^{t-1})$
- 2 with probability

$$\min \left\{ 1, \frac{\pi(\theta^*)}{\pi(\theta^{t-1})} \frac{q(\theta^{t-1}|\theta^*)}{q(\theta^*|\theta^{t-1})} \right\}$$

accept: set $\theta^k = \theta^*$, else reject: set $\theta^k = \theta^{k-1}$.

In the simplest case $q(\theta^*|\theta^{t-1}) = \mathcal{N}(\theta^*|\theta^{t-1}, \epsilon \mathbf{I})$

- [Roberts & Tweedie1996], for $\theta \in \mathbb{R}^D$ with density $\pi(\theta)$, $\mathcal{L}(\theta) \equiv \log \pi(\theta)$, define Langevin diffusion

$$d\theta(t) = \frac{1}{2} \nabla_{\theta} \mathcal{L}(\theta(t)) dt + d\mathbf{b}(t)$$

- First order Euler-Maruyama discrete integration of diffusion

$$\theta(\tau + \epsilon) = \theta(\tau) + \frac{\epsilon^2}{2} \nabla_{\theta} \mathcal{L}(\theta(\tau)) + \epsilon \mathbf{z}(\tau)$$

- Proposal

$$q(\theta^* | \theta) = \mathcal{N}(\theta^* | \mu(\theta, \epsilon), \epsilon^2 \mathbf{I}) \quad \text{with} \quad \mu(\theta, \epsilon) = \theta + \frac{\epsilon^2}{2} \nabla_{\theta} \mathcal{L}(\theta)$$

- Isotropic diffusion inefficient, employ pre-conditioning

$$\theta^* = \theta + \epsilon^2 \mathbf{M} \nabla_{\theta} \mathcal{L}(\theta) / 2 + \epsilon \sqrt{\mathbf{M}} \mathbf{z}$$

- How to set \mathbf{M} systematically? Tuning in transient & stationary phases

[Girolami & Calderhead 2011], discretised Langevin diffusion on manifold defines proposal mechanism

$$\theta_d^* = \theta_d + \frac{\epsilon^2}{2} \left(\mathbf{G}^{-1}(\theta) \nabla_{\theta} \mathcal{L}(\theta) \right)_d - \epsilon^2 \sum_{i,j}^D \mathbf{G}(\theta)_{i,j}^{-1} \Gamma_{i,j}^d + \epsilon \left(\sqrt{\mathbf{G}^{-1}(\theta)} \mathbf{z} \right)_d$$

$$q(\theta^* | \theta^{t-1}) = \mathcal{N} \left(\theta^* | \mu(\theta^{t-1}, \epsilon), \epsilon^2 \mathbf{G}^{-1}(\theta^{t-1}) \right)$$

Where

$$\mu(\theta, \epsilon)_d = \theta_d + \frac{\epsilon^2}{2} \left(\mathbf{G}^{-1}(\theta) \nabla_{\theta} \mathcal{L}(\theta) \right)_d - \epsilon^2 \sum_{i,j}^D \mathbf{G}(\theta)_{i,j}^{-1} \Gamma_{i,j}^d$$

Metric tensor $\mathbf{G}(\theta) = \text{cov}(\nabla_{\theta} \mathcal{L}(\theta))$ is the expected FI and $\Gamma_{i,j}^d$ are the Christoffel symbols

$$\Gamma_{i,j}^d = \frac{1}{2} \sum_m \mathbf{G}^{-1}(\theta)_{d,m} \left(\frac{\partial \mathbf{G}(\theta)_{m,i}}{\partial \theta_j} + \frac{\partial \mathbf{G}(\theta)_{m,j}}{\partial \theta_i} - \frac{\partial \mathbf{G}(\theta)_{i,j}}{\partial \theta_m} \right)$$

[Girolami & Calderhead 2011], discretised Langevin diffusion on manifold defines proposal mechanism

$$\theta_d^* = \theta_d + \frac{\epsilon^2}{2} \left(\mathbf{G}^{-1}(\theta) \nabla_{\theta} \mathcal{L}(\theta) \right)_d - \epsilon^2 \sum_{i,j}^D \mathbf{G}(\theta)_{i,j}^{-1} \Gamma_{i,j}^d + \epsilon \left(\sqrt{\mathbf{G}^{-1}(\theta)} \mathbf{z} \right)_d$$

$$q(\theta^* | \theta^{t-1}) = \mathcal{N} \left(\theta^* | \mu(\theta^{t-1}, \epsilon), \epsilon^2 \mathbf{G}^{-1}(\theta^{t-1}) \right)$$

Where

$$\mu(\theta, \epsilon)_d = \theta_d + \frac{\epsilon^2}{2} \left(\mathbf{G}^{-1}(\theta) \nabla_{\theta} \mathcal{L}(\theta) \right)_d - \epsilon^2 \sum_{i,j}^D \mathbf{G}(\theta)_{i,j}^{-1} \Gamma_{i,j}^d$$

Metric tensor $\mathbf{G}(\theta) = \text{cov}(\nabla_{\theta} \mathcal{L}(\theta))$ is the expected FI and $\Gamma_{i,j}^d$ are the Christoffel symbols

$$\Gamma_{i,j}^d = \frac{1}{2} \sum_m \mathbf{G}^{-1}(\theta)_{d,m} \left(\frac{\partial \mathbf{G}(\theta)_{m,i}}{\partial \theta_j} + \frac{\partial \mathbf{G}(\theta)_{m,j}}{\partial \theta_i} - \frac{\partial \mathbf{G}(\theta)_{i,j}}{\partial \theta_m} \right)$$

Simplified MMALA assumes that the metric is not changing locally

[Girolami & Calderhead 2011], design proposal mechanism that follows direct paths - geodesics

Introduce auxiliary variable $\mathbf{p} \sim \mathcal{N}(0, \mathbf{G}(\theta))$. Hamiltonian defined on Riemann manifold using the negative joint log likelihood is

$$H(\theta, \mathbf{p}) = \underbrace{-\mathcal{L}(\theta) + \frac{1}{2} \log(2\pi)^D |\mathbf{G}(\theta)|}_{\text{Potential energy}} + \underbrace{\frac{1}{2} \mathbf{p}^T \mathbf{G}(\theta)^{-1} \mathbf{p}}_{\text{Kinetic energy}}$$

Sample using a Metropolis within Gibbs scheme

$$\mathbf{p}^t \sim \mathcal{N}(0, \mathbf{G}(\theta^{t-1})) \quad \theta^t \sim p(\theta | \mathbf{p}^t)$$

Second step involves integrating

$$\frac{d\theta}{dt} = \frac{\partial}{\partial \mathbf{p}} H(\theta, \mathbf{p}) \quad \frac{d\mathbf{p}}{dt} = -\frac{\partial}{\partial \theta} H(\theta, \mathbf{p})$$

using a time reversible, volume preserving numerical integration such as generalised leapfrog.

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- Gibbs sampler is straightforward using a data augmentation scheme and conjugate priors. [McLachlan & Peel2000]
- Poor mixing especially when components overlap.[Celeux et al.2000]
- Attracted by local modes of the posterior [Marin et al.2005].
- Both problems related to the conditional dependence of latent variables and model parameters.
- Simulated tempering with Metropolis-Hastings base kernels [Celeux et al.2000, Jarsa et. al.2005].
- Manifold MCMC algorithms can be used as base kernels for simulated tempering and population MCMC.

- Univariate mixtures of Gaussians

$$\pi_k = \frac{e^{\alpha_k}}{\sum_{k'=1}^K e^{\alpha_{k'}}}, \quad p(\alpha_k) = \mathcal{G}(e^{\alpha_k} | \lambda, 1) e^{\alpha_k}$$

$$\sigma_k^2 = e^{\gamma_k}, \quad p(\gamma_k) = \mathcal{IG}(\sigma_k^2 | b, c) e^{\gamma_k}$$

- Univariate mixtures of Gaussians

$$\pi_k = \frac{e^{\alpha_k}}{\sum_{k'=1}^K e^{\alpha_{k'}}}, \quad p(\alpha_k) = \mathcal{G}(e^{\alpha_k} | \lambda, 1) e^{\alpha_k}$$

$$\sigma_k^2 = e^{\gamma_k}, \quad p(\gamma_k) = \mathcal{IG}(\sigma_k^2 | b, c) e^{\gamma_k}$$

- Multivariate mixtures of Gaussians, [Pineiro & Bates1996]

$$\Sigma = \mathbf{L}\mathbf{L}^T, \quad \mathbf{L}_k = \begin{bmatrix} e^{((\mathbf{B}_k)_{1,1})} & 0 & 0 & 0 \\ (\mathbf{B}_k)_{2,1} & \ddots & 0 & 0 \\ \vdots & \ddots & e^{((\mathbf{B}_k)_{d,d})} & 0 \\ (\mathbf{B}_k)_{D,1} & \dots & (\mathbf{B}_k)_{D,D-1} & e^{((\mathbf{B}_k)_{D,D})} \end{bmatrix}$$

$$p(\mathbf{B}_k) = \prod_{d=1}^D \mathcal{G}(e^{(\mathbf{B}_k)_{d,d}} | \omega, \psi) e^{(\mathbf{B}_k)_{d,d}} \prod_{d' \neq d} \mathcal{N}((\mathbf{B}_k)_{d',d} | 0, \beta)$$

- For mixture models the expected FI is not explicit.
- Empirical FI using a finite sample estimate, [McLachlan & Peel2000]

$$\mathbf{G}(\theta) = \mathbf{S}^T \mathbf{S} - \frac{1}{N} \bar{\mathbf{s}} \bar{\mathbf{s}}^T \xrightarrow{N \rightarrow \infty} \text{cov}(\nabla_{\theta} \mathcal{L}(\theta))$$

$$\frac{\partial \mathbf{G}(\theta)}{\partial \theta_d} = \left(\frac{\partial \mathbf{S}^T}{\partial \theta_d} \mathbf{S} + \mathbf{S}^T \frac{\partial \mathbf{S}}{\partial \theta_d} \right) - \frac{1}{N} \left(\frac{\partial \bar{\mathbf{s}}}{\partial \theta_d} \bar{\mathbf{s}}^T + \bar{\mathbf{s}} \frac{\partial \bar{\mathbf{s}}^T}{\partial \theta_d} \right)$$

where $\mathbf{S}_{i,d} = \frac{\partial \log p(x_i|\theta)}{\partial \theta_d}$ and $\bar{\mathbf{s}} = \sum_{n=1}^N \mathbf{S}_{i,d}$.

- For mixture models the expected FI is not explicit.
- Empirical FI using a finite sample estimate, [McLachlan & Peel2000]

$$\mathbf{G}(\theta) = \mathbf{S}^T \mathbf{S} - \frac{1}{N} \bar{\mathbf{s}} \bar{\mathbf{s}}^T \xrightarrow{N \rightarrow \infty} \text{cov}(\nabla_{\theta} \mathcal{L}(\theta))$$

$$\frac{\partial \mathbf{G}(\theta)}{\partial \theta_d} = \left(\frac{\partial \mathbf{S}^T}{\partial \theta_d} \mathbf{S} + \mathbf{S}^T \frac{\partial \mathbf{S}}{\partial \theta_d} \right) - \frac{1}{N} \left(\frac{\partial \bar{\mathbf{s}}}{\partial \theta_d} \bar{\mathbf{s}}^T + \bar{\mathbf{s}} \frac{\partial \bar{\mathbf{s}}^T}{\partial \theta_d} \right)$$

where $\mathbf{S}_{i,d} = \frac{\partial \log p(x_i|\theta)}{\partial \theta_d}$ and $\bar{\mathbf{s}} = \sum_{n=1}^N \mathbf{S}_{i,d}$.

- Alternative metric between densities, [Basu et al.1998]

$$L_2 = \int |p(\mathbf{x}|\theta) - p(\mathbf{x}|\theta + \delta\theta)|^2 d\mathbf{x}$$

$$\mathbf{G}(\theta) = \int \nabla_{\theta} p(\mathbf{x}|\theta) \nabla_{\theta}^T p(\mathbf{x}|\theta) d\mathbf{x}$$

$$p(x|\mu, \sigma^2) = 0.7 \times \mathcal{N}(x|0, \sigma^2) + 0.3 \times \mathcal{N}(x|\mu, \sigma^2)$$

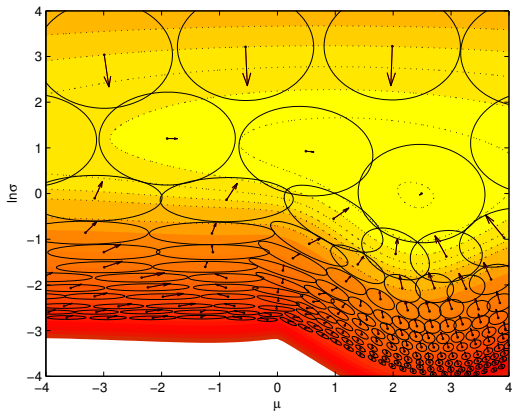


Figure: Arrows correspond to the gradients and ellipses to the inverse metric tensor. Dashed lines are isocontours of the joint log density

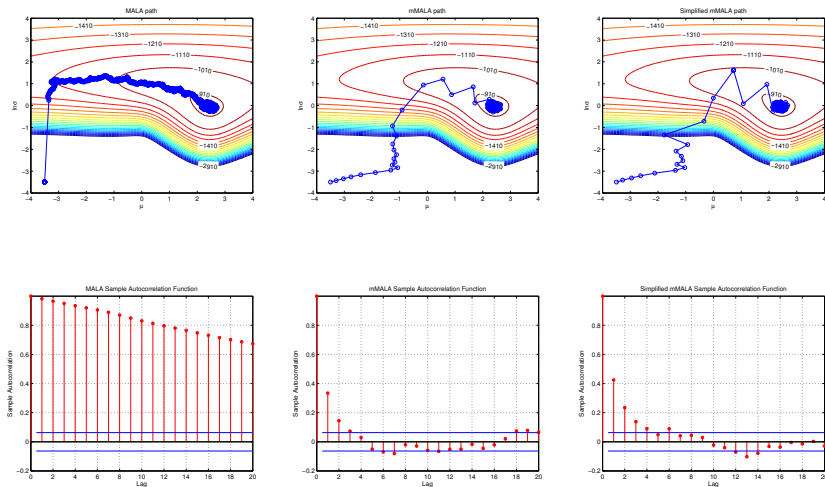


Figure: Comparison of MALA (left), mMALA (middle) and simplified mMALA (right) convergence paths and autocorrelation plots. Autocorrelation plots are from the stationary chains, i.e. once the chains have converged to the stationary distribution.

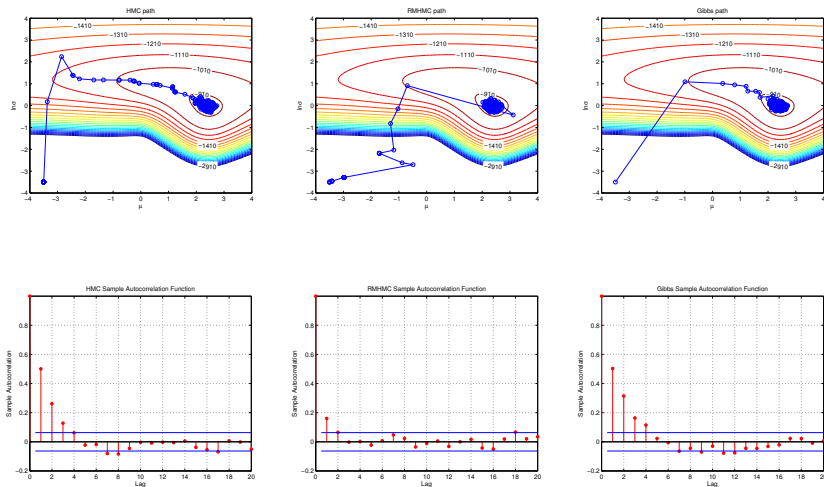


Figure: Comparison of HMC (left), RMHMC (middle) and GIBBS (right) convergence paths and autocorrelation plots. Autocorrelation plots are from the stationary chains, i.e. once the chains have converged to the stationary distribution.

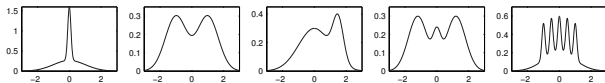
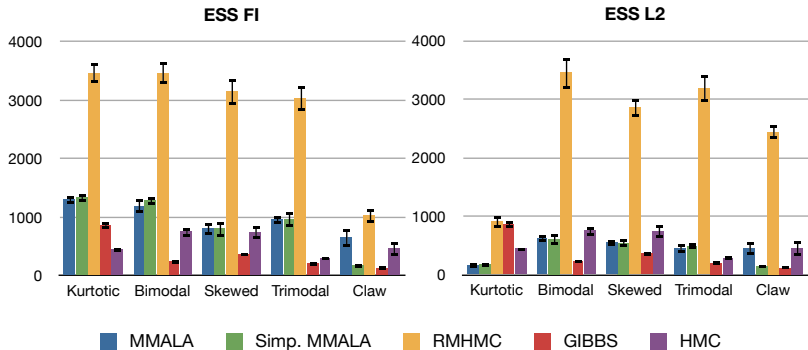


Figure: Densities used to generate synthetic datasets.



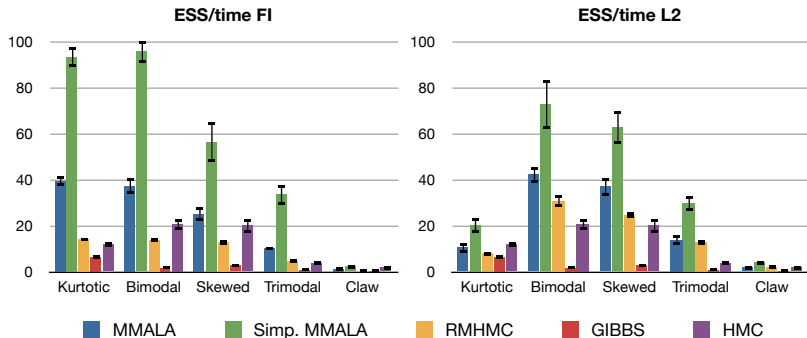
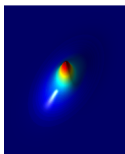


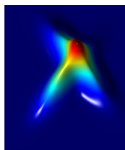
Table: Densities used to generate synthetic datasets



$$\pi_1 = 0.8 \quad \pi_2 = 0.2$$

$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = [0, 0]^T$$

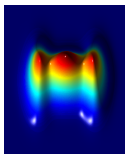
$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.01 \end{bmatrix}$$



$$\pi_1 = 0.5 \quad \pi_2 = 0.5$$

$$\boldsymbol{\mu}_1 = [0, 0]^T \quad \boldsymbol{\mu}_2 = [0.8, 0.8]^T$$

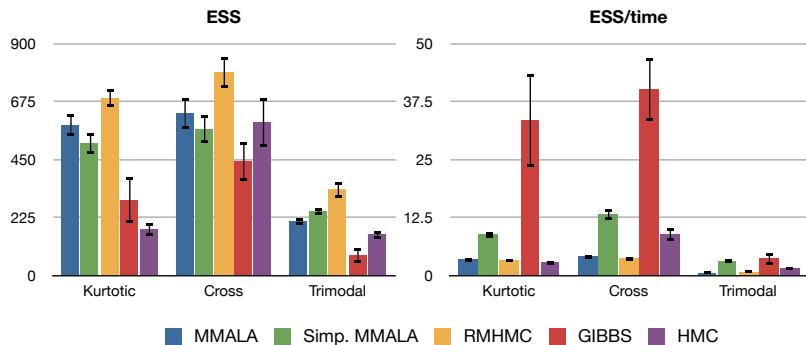
$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$

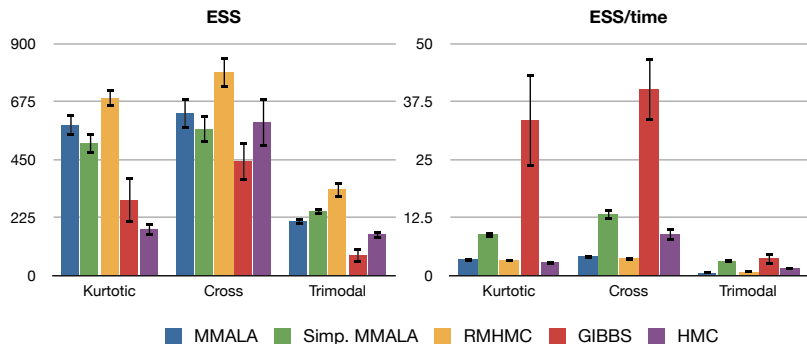


$$\pi_1 = 0.55 \quad \pi_2 = 0.225 \quad \pi_3 = 0.225$$

$$\boldsymbol{\mu}_1 = [0, 0]^T \quad \boldsymbol{\mu}_2 = [1.8, 0]^T \quad \boldsymbol{\mu}_3 = [-1.8, 0]^T$$

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0.5 \end{bmatrix} \quad \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \begin{bmatrix} 0.2 & 0 \\ 0 & 1 \end{bmatrix}$$





Computational complexity of manifold MCMC, $(K \times [D + D \times (D + 1)/2 + 1])^3$
 Gibbs sampler, $K[D + (D(D + 1)/2)^3 + 1]$

- Popular models in population genetics.
- The genotype of an individual is an ad-mixture of K unknown sub-populations.
- Goal: given the genotype of a sample of individuals at specific loci infer their ancestry proportions and sub-population allele frequencies.
- Applications in document analysis
- A document is an ad-mixture of K unknown topics.
- Give a collection of documents infer topic distributions and topic proportions for each document.

- θ_n mixing proportions for n^{th} observation.
- ϕ_k parameters of the k^{th} sub-population/topic.

The generative process for ad-mixtures is

$$\theta_n \sim \mathcal{D}_K(\alpha), \quad n \in \{1, \dots, N\}$$

$$M_n \sim \mathcal{P}(\lambda), \quad n \in \{1, \dots, N\}$$

$$\phi_k \sim \mathcal{D}_T(\beta), \quad k \in \{1, \dots, K\}$$

$$\mathbf{x}_n \sim \mathcal{M}(\theta_n \Phi, M_n), \quad n \in \{1, \dots, N\}$$

The likelihood can be written as

$$p(\mathbf{X} | \Theta, \Phi) = \prod_n \prod_t \left(\sum_k \theta_{n,k} \phi_{k,t} \right)^{x_{n,t}}$$

- θ_n mixing proportions for n^{th} observation.
- ϕ_k parameters of the k^{th} sub-population/topic.

The generative process for ad-mixtures is

$$\theta_n \sim \mathcal{D}_K(\alpha), \quad n \in \{1, \dots, N\}$$

$$M_n \sim \mathcal{P}(\lambda), \quad n \in \{1, \dots, N\}$$

$$\phi_k \sim \mathcal{D}_T(\beta), \quad k \in \{1, \dots, K\}$$

$$\mathbf{x}_n \sim \mathcal{M}(\theta_n \Phi, M_n), \quad n \in \{1, \dots, N\}$$

The likelihood can be written as

$$p(\mathbf{X} | \Theta, \Phi) = \prod_n \prod_t \left(\sum_k \theta_{n,k} \phi_{k,t} \right)^{x_{n,t}}$$

Re-parameterisation

$$\theta_{n,k} = \frac{e^{\gamma_{n,k}}}{\sum_{k'} e^{\gamma_{n,k'}}}, \quad \phi_{k,t} = \frac{e^{\psi_{k,t}}}{\sum_{k'} e^{\psi_{k,t'}}}$$

- Number of parameters increases linearly with the observations.
 $K \times (N + T)$
- Exploit conditional independence

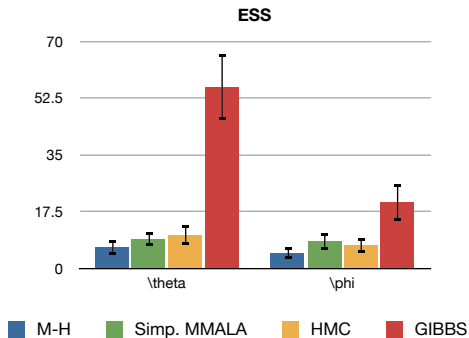
$$\mathbb{E} \left[\frac{\partial \mathcal{L}(\Gamma, \Psi)}{\partial \gamma_{n,k}} \frac{\partial \mathcal{L}(\Gamma, \Psi)}{\partial \gamma_{i,j}} \right] = 0, \quad \forall i \neq n$$

- Metropolis within Gibbs scheme

$$\begin{aligned} \gamma_n | \Gamma_{/n}, \Psi &\sim p(\gamma_n | \Gamma_{/n}, \Psi), \quad \forall n \\ \Psi | \Gamma &\sim p(\Psi | \Gamma) \end{aligned}$$

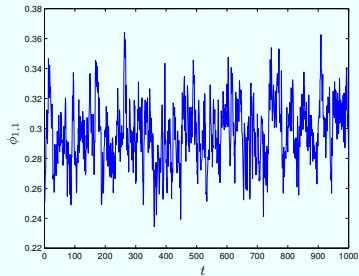
- Each step is a simplified MMALA sampler

- 200 observations with average length 25, from 6 sub-populations with 9 markers / terms, $\alpha = 0.5$, $\beta = 0.1$.

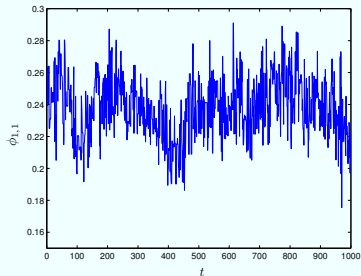


Synthetic data example

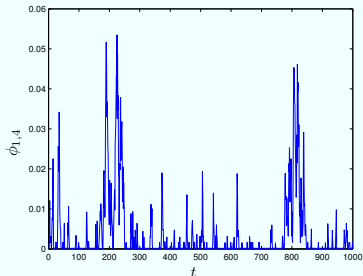
Gibbs Trace Plot



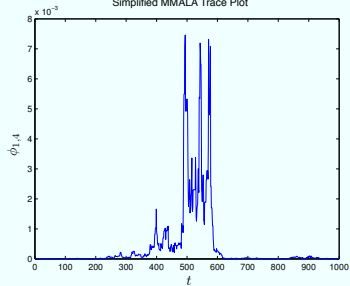
Simplified MMALA Trace Plot



Gibbs Trace Plot




Simplified MMALA Trace Plot



- Approximations of the expected FI.
- Observed FI is effective but computationally expensive.
- Simplified MMALA as effective as MMALA.
- Problematic on skew posteriors.
- Algorithms scale as $\mathcal{O}(D^3)$, D the number of parameters.
- Cumbersome derivations, consider automatic differentiation.

- Appropriate geometry for finite mixtures ?
- Geometry for mixtures with unknown number of components and applications in MCMC ?
- Overcoming computational cost ?
- Alternative re-parameterisations ?

-  Ayanendranath Basu, R. Harris, Ian, Nils L. Hjort, and M. C. Jones.
Robust and efficient estimation by minimising a density power divergence.
Biometrika, 85(3):549–559, 1998.
-  Gilles Celeux, Merrilee Hurn, and Christian P. Robert.
Computational and inferential difficulties with mixture posterior distributions.
Journal of the American Statistical Association, 95(451):957–970, 2000.
-  Mark Girolami and Ben Calderhead.
Riemann manifold Langevin and Hamiltonian Monte Carlo methods.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(2):123–214, 2011.
-  A. Jasra, C. Holmes, and D. A. Stephens
MCMC and the label switching problem in Bayesian mixture models.
Statistical Science, 20:50–67, 2005.
-  M. J. Marin, K. Mengersen, and P. Robert, C.
Bayesian Modelling and Inference on Mixtures of Distributions.
Handbook of Statistics, pages 15840–15845, 2005.



Geoffrey McLachlan and David Peel.

Finite Mixture Models.

Wiley Series in Probability and Statistics. Wiley-Interscience, October 2000.



C. Pinheiro, Jose and M. Bates, Douglas.

Unconstrained parametrizations for variance-covariance matrices.

Statistics and Computing, 6(3):289–296, 1996.



G. O. Roberts and R. O. Tweedie

Exponential convergence of Langevin distributions and their discrete approximations

Bernoulli, 2(4):341–363, 1996