Metropolis-Hastings Algorithms with Adaptive Proposals

joint work with Francois Perron and Bo Cai

Dept of Maths and Statistics, University of Montreal, Canada
Dept of Biostatistics, University of South Carolina, USA

Introduction

Seminal papers on MCMC: Geman and Geman (1984), Gelfand and Smith (1990)

Introduction

Seminal papers on MCMC: Geman and Geman (1984), Gelfand and Smith (1990)

General strategy:

- generate samples $\{X_i, i = 0, 1, ...\}$ from target density π on $D \subseteq \mathbb{R}^n$
- approximate $I(g) = \int_D g(x)\pi(x)dx$
- by $\hat{I}_N(g) = \frac{1}{N}\sum_{i=1}^{N} g(X_i)$
- provided that Markov chain is ergodic

Introduction, continued

Building block: Metropolis-Hastings (MH) algorithm

- proposal distributions $q(.|x), x \in D$, generating possible transitions of the Markov chain from x to y
- accepted (otherwise rejected) with probability

$$\alpha(x, y) = \min \left\{ 1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)} \right\}$$

Introduction, continued

- performance of MH depends on choice of proposal densities
- optimal acceptance rates known for various specific MCMC algorithms (Roberts and Rosenthal, 2001)
- but tuning by hand is time-consuming
- adaptive MCMC: automatic tuning "on the fly"
- Warning: adaptation can easily perturb ergodicity

Review of Adaptive Techniques

- pilot adaptation

Review of Adaptive Techniques

- pilot adaptation
- regeneration times (Brockwell and Kadane 2005)

Review of Adaptive Techniques

- pilot adaptation
- regeneration times (Brockwell and Kadane 2005)
- finite horizon, kernel density estimate
  Normal Kernel Coupler, Warnes (2001)

Review of Adaptive Techniques

- pilot adaptation
- regeneration times (Brockwell and Kadane 2005)
- finite horizon, kernel density estimate
  Normal Kernel Coupler, Warnes (2001)
- infinite adaptation of random walk MH
  using all previous states, Haario et al. (2001)

Review of Adaptive Techniques

- pilot adaptation
- regeneration times (Brockwell and Kadane 2005)
- finite horizon, kernel density estimate
  Normal Kernel Coupler, Warnes (2001)
- infinite adaptation of random walk MH
  using all previous states, Haario et al. (2001)
- coupling and ergodicity of adaptive MCMC
  Roberts and Rosenthal (2007)

Review of Adaptive Techniques

- pilot adaptation
- regeneration times (Brockwell and Kadane 2005)
- finite horizon, kernel density estimate
  Normal Kernel Coupler, Warnes (2001)
- infinite adaptation of random walk MH
  using all previous states, Haario et al. (2001)
- coupling and ergodicity of adaptive MCMC
  Roberts and Rosenthal (2007)
- Andrieu and Thoms (2008): vanishing adaptation

MH with Adpative Proposals

Idea:

- proposal as close to the target π as possible
- last m samples provide info about π
- use kernel density estimate based on last m samples
- finite horizon technique

Normal Kernel Coupler

Suggested by Warnes (2001):

Let $x_1^{(t)}, \ldots, x_m^{(t)}$ be set of m current states

- Select component $x_i^{(t)}$ to update
- Propose new candidate y_i:

$$q(y_i|x^{(t)}) = \frac{1}{m}\sum_{j=1}^{m} N(y_i|x_j^{(t)}, h^2 V)$$

Here: MH-within-Gibbs

- new target pdf on product space D^m
  $\boldsymbol{\pi}(x_1, \ldots, x_m) = \prod_{i=1}^{m} \pi(x_i)$

Here: MH-within-Gibbs

- new target pdf on product space D^m
  $\boldsymbol{\pi}(x_1, \ldots, x_m) = \prod_{i=1}^m \pi(x_i)$
- Gibbs sampler: in cycle t, iteration i draw from
  $\boldsymbol{\pi}(x_i | x_{-i}^{(t)}) = \pi(x_i)$ using MH

Here: MH-within-Gibbs

- new target pdf on product space D^m
  $\boldsymbol{\pi}(x_1, \ldots, x_m) = \prod_{i=1}^m \pi(x_i)$
- Gibbs sampler: in cycle t, iteration i draw from
  $\boldsymbol{\pi}(x_i | x_{-i}^{(t)}) = \pi(x_i)$ using MH
- generate candidate y from proposal $q(.|x_{-i}^{(t)})$

Here: MH-within-Gibbs

- new target pdf on product space D^m
  $\boldsymbol{\pi}(x_1, \ldots, x_m) = \prod_{i=1}^m \pi(x_i)$
- Gibbs sampler: in cycle t, iteration i draw from
  $\boldsymbol{\pi}(x_i | x_{-i}^{(t)}) = \pi(x_i)$ using MH
- generate candidate y from proposal $q(.|x_{-i}^{(t)})$
- accept with probability

$$\alpha(x_i^{(t)}, y) = \min\left\{1, \frac{\pi(y) q(x_i^{(t)} | x_{-i}^{(t)})}{\pi(x_i^{(t)}) q(y | x_{-i}^{(t)})}\right\}$$

as $\boldsymbol{\pi}(y | x_{-i}^{(t)}) = \pi(y)$

Ergodicity

Roberts and Rosenthal (2006)
Harris recurrence of MH-within-Gibbs algorithms:

- $r-$dim. integral of $\boldsymbol{\pi}$ has finite Lebesque integral over every r-dim. coordinate hyperplane of D^m, $1 \leq r \leq m$
- full chain and all subchains are ϕ-irreducible

1-Dimensional Target

Here: 2 novel principles

- ATRIMS:
  Adaptive Triangular Metropolis Sampling
- ATRAMS:
  Adaptive Trapezoidal Metropolis Sampling

1-Dimensional Target

Here: 2 novel principles

- ATRIMS:
  Adaptive Triangular Metropolis Sampling
- ATRAMS:
  Adaptive Trapezoidal Metropolis Sampling

Compare these to

- ARMS (Gilks et al. 1995): Adaptive Rejection Metropolis Sampling
  standard black-box technique
- NKC (Warnes, 2001): Normal Kernel Coupler

ATRIMS

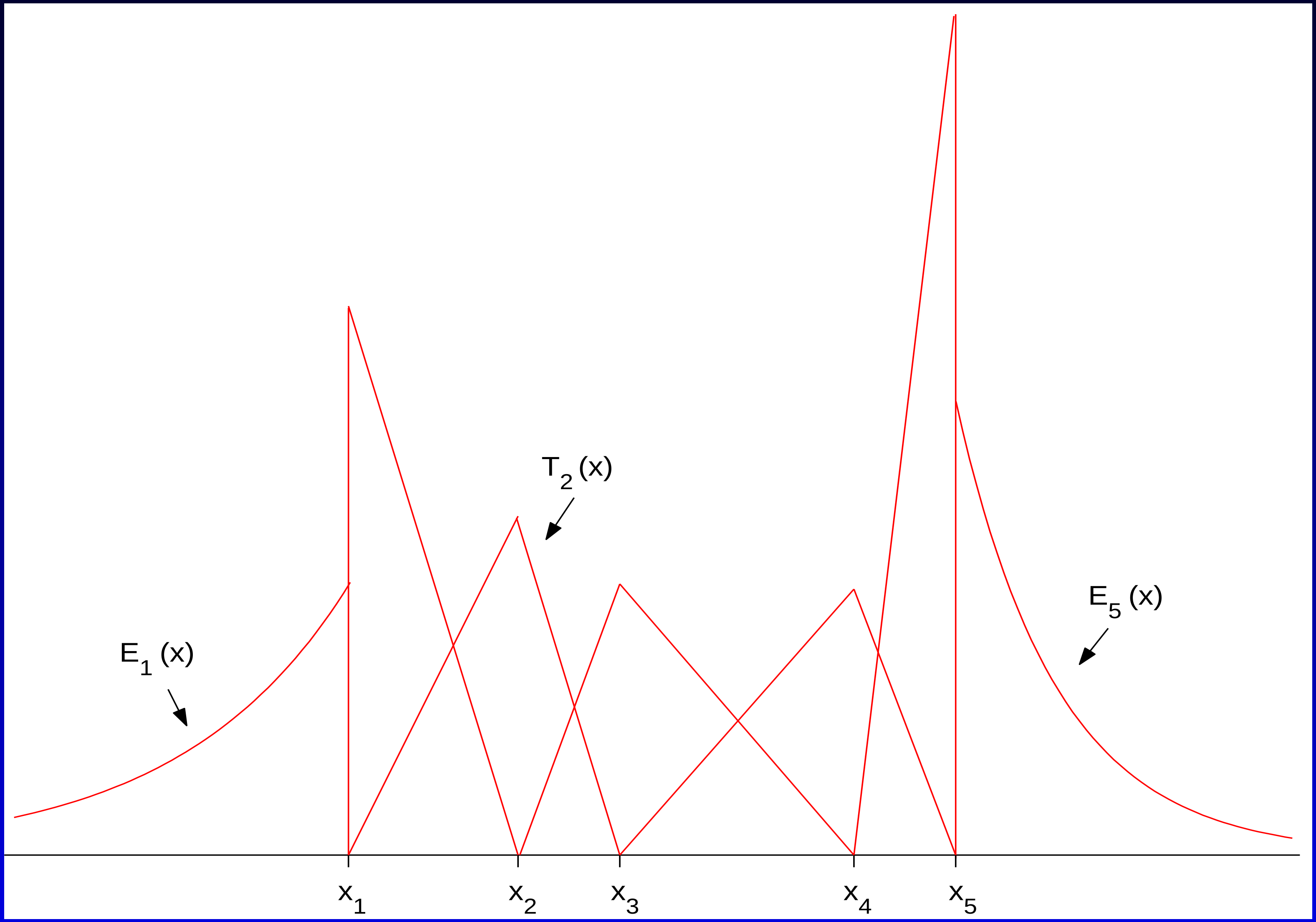
Say π target pdf on 1-dim. space D

already sampled: $x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \ldots, x_m^{(t)}$

denote those by: $x_1, \ldots, x_{m-1}$

Perron and Mengersen (2001):
any pdf can be approximated by a mixture of triangular distributions

Triangular Densities



ATRIMS

$$q(x|x_1, \ldots, x_{m-1}) =$$

$$\begin{cases} w_0 E_1(x), & x \in (-\infty, x_1), \\ w_i T_i(x) + w_{i+1} T_{i+1}(x), & x \in [x_i, x_{i+1}), \\ w_m E_{m-1}(x), & x \in [x_{m-1}, \infty). \end{cases}$$

with weights:

$$w = \{\underbrace{\frac{1}{m}, \frac{1}{2m}, \frac{1}{m}, \ldots, \frac{1}{m}, \frac{1}{2m}, \frac{1}{m}}_{m+1}\}.$$

ATRIMS

Advantages over NKC and ARMS:

- sampling from triangular/exponential is fast

ATRIMS

Advantages over NKC and ARMS:

- sampling from triangular/exponential is fast
- evaluation of proposal only requires evaluation of 2 mixture components

ATRIMS

Advantages over NKC and ARMS:

- sampling from triangular/exponential is fast
- evaluation of proposal only requires evaluation of 2 mixture components
- increasing the horizon m won't increase evaluation effort

ATRIMS

Advantages over NKC and ARMS:

- sampling from triangular/exponential is fast
- evaluation of proposal only requires evaluation of 2 mixture components
- increasing the horizon m won't increase evaluation effort
- starting knots in ARMS may not depend on previously sampled points (Gilks et al, 1997)

ATRIMS

Advantages over NKC and ARMS:

- sampling from triangular/exponential is fast
- evaluation of proposal only requires evaluation of 2 mixture components
- increasing the horizon m won't increase evaluation effort
- starting knots in ARMS may not depend on previously sampled points (Gilks et al, 1997)
- ARMS requires finite support interval

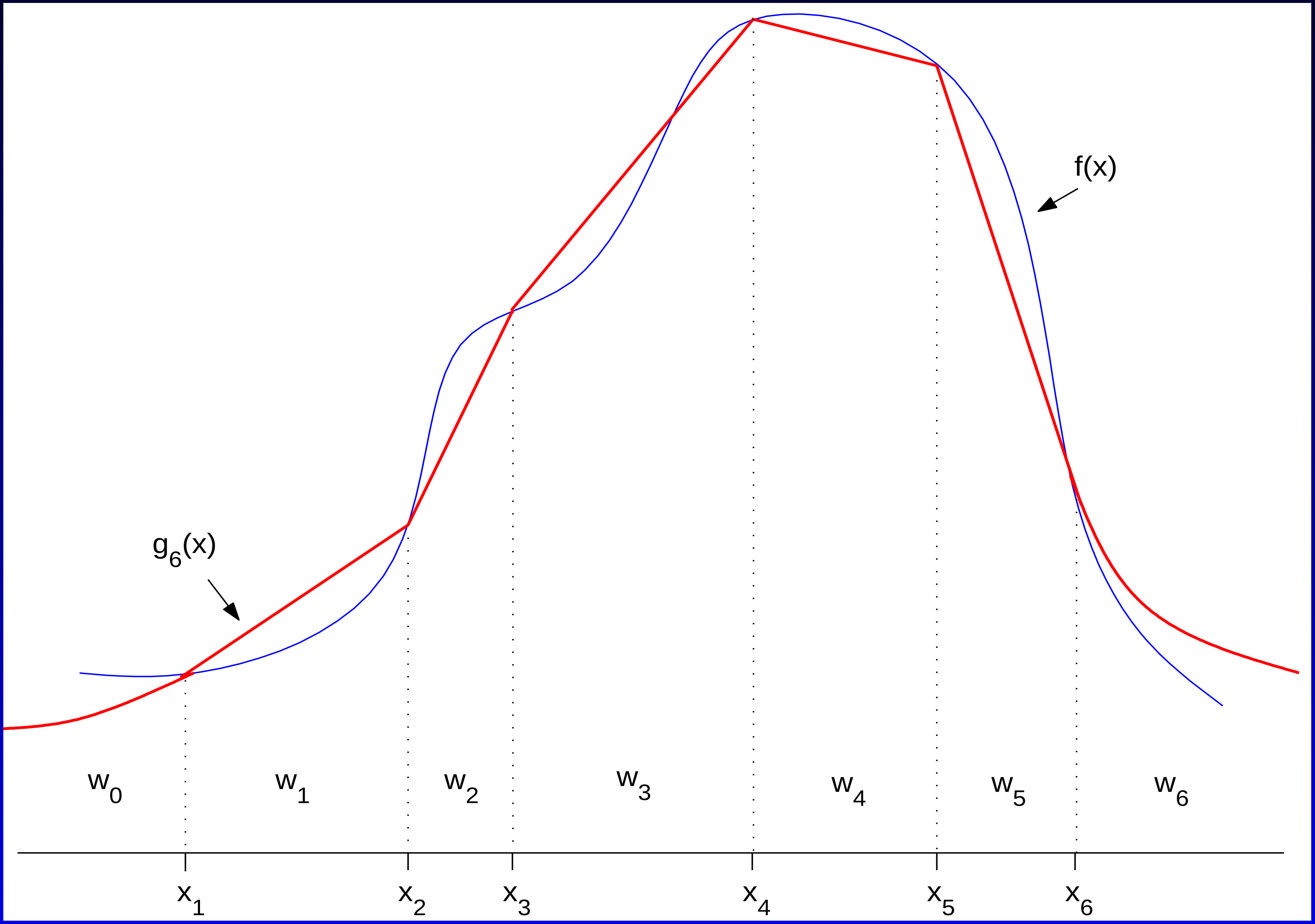
ATRAMS

Idea: use that

- the functional form of target pdf is known
- the target pdf already evaluated at $m - 1$ points

$\Longrightarrow$ piecewise linear/trapezoidal approximation

Trapezoidal Densities



ATRAMS

$$q(x|x_1,\dots,x_{m-1}) =$$

$$\begin{cases} w_0 E_1'(x), & \text{if } x \in (-\infty, x_1), \\ w_i T_i'(x), & \text{if } x \in [x_i, x_{i+1}), i = 1, 2, \dots, m - \\ w_{m-1} E_{m-1}'(x), & \text{if } x \in [x_{m-1}, \infty). \end{cases}$$

with weights:

$$w_i = \begin{cases} \frac{1}{m}, & \text{for } i = 0, \\ \frac{m-2}{m}\frac{s_i}{S}, & \text{for } i = 1, 2, \dots, m-2, \\ \frac{1}{m}, & \text{for } i = m-1. \end{cases}$$

ATRAMS

Remarks:

- sampling from piecewise linear/exponential is fast

ATRAMS

Remarks:

- sampling from piecewise linear/exponential is fast
- increasing the horizon m won't increase evaluation effort

ATRAMS

Remarks:

- sampling from piecewise linear/exponential is fast
- increasing the horizon m won't increase evaluation effort
- abscissae in subsequent steps are based on previous Gibbs iterations

ATRAMS

Remarks:

- sampling from piecewise linear/exponential is fast
- increasing the horizon m won't increase evaluation effort
- abscissae in subsequent steps are based on previous Gibbs iterations

Extension to multivariate target: ATRIMS/ATRAMS within Gibbs sampling

Simulation Study

Compare ATRIMS and ATRAMS to ARMS and NKC to sample from univariate distributions:

- Gumbel(0,10)
- Logistic(0,2)
- $0.3 * N(5, 1^2) + 0.7 * N(10, 3^2)$
- $0.3 * N(2, 1^2) + 0.6 * N(20, 3^2) + 0.1 * N(35, 1^2)$

Simulation Study, continued

and to sample from multivariate distributions:

- $$0.5N_2\left(\begin{pmatrix}-1.5\\1.5\end{pmatrix},\begin{pmatrix}1 & -0.95\\-0.95 & 1\end{pmatrix}\right)+$$
  $$0.5N_2\left(\begin{pmatrix}1.5\\1.5\end{pmatrix},\begin{pmatrix}1 & 0.95\\0.95 & 1\end{pmatrix}\right)$$
- $d=4$ and $d=20$-dim. Normal:
  $$\begin{pmatrix}X_1\\X_2\\\vdots\\X_d\end{pmatrix}\sim N_d\left(\begin{pmatrix}0\\0\\\vdots\\0\end{pmatrix},\begin{pmatrix}1 & \rho & \dots & \rho^{d-1}\\ \rho & 1 & \dots & \rho^{d-2}\\ \vdots & \vdots & \ddots & \vdots\\ \rho^{d-1} & \rho^{d-2} & \dots & 1\end{pmatrix}\right)$$

Simulation Study, continued

Target Distribution	Integrated Autocorrelation Times of ARMS	NKC	ATRIMS	ATRAMS
Gumbel(0,10)	0.97	1.21	1.65	1.83
Logistic(0,2)	1.31	1.99	2.46	2.04
Bimodal Normal	1.10	1.25	1.41	1.56
Trimodal Normal	1.79	1.70	1.60	1.81
Bivariate Mixed Normal				
X	2.94	3.02	3.05	3.24
Y	2.96	3.00	3.07	3.21
4-D Narrow Normal				
X_1	2.93	2.92	2.96	3.05
X_2	2.89	2.89	2.94	3.01
X_3	2.93	2.91	2.91	3.04
X_4	2.91	2.92	2.96	3.07
20-D Narrow Normal	2.91	2.95	2.98	3.04

Simulation Study, continued

Target Distribution	CPU time in seconds for 10,000 samples of ARMS	NKC	ATRIMS	ATRAMS
Gumbel(0,10)	1.28	9.9	0.71	1.03
Logistic(0,2)	0.96	10.1	0.74	0.85
Bimodal Normal	1.60	10.2	0.95	1.33
Trimodal Normal	4.51	9.8	1.34	1.66
Bivariate Mixed Normal	10.06	16.51	4.99	6.04
4-D Narrow Normal	10.00	18.62	3.33	4.41
20-D Narrow Normal	95.67	141.35	68.84	70.20

Case Study

General state-space model

Kuensch (2001), West and Harrison (1997)

Observation equation:

$$y_t = h_t(\theta_t, v_t)$$

State equation:

$$\theta_t = g_t(\theta_{t-1}, u_t)$$

Computation

Carlin et al. (1992): Gibbs sampler for nonlinear non-Gaussian state-space models

Gibbs sampler combined with ARMS used to fit population dynamics models for fisheries stock assessment, e.g. Meyer and Millar (1999)

Fisheries Stock Assessment

Yellowfin tuna data from Pella and Tomlinson (1969)

Year	Catch	CPUE
1934	60.9	10361
1935	72.3	11484
1936	78.4	11571
1937	91.5	11116
1938	78.3	11463
1939	110.4	10528
1940	114.6	10609
1941	76.8	8018
1942	42.0	7040
⋮	⋮	⋮
1965	180.1	4166
1966	182.3	4513
1967	178.9	5292

Biomass Dynamics Model

new biomass

- $=$ old biomass

Biomass Dynamics Model

new biomass

- $=$ old biomass
- $+$ growth

Biomass Dynamics Model

new biomass

- $=$ old biomass
- $+$ growth
- $+$ recruitment

Biomass Dynamics Model

new biomass

- $=$ old biomass
- $+$ growth
- $+$ recruitment
- $-$ natural mortality

Biomass Dynamics Model

new biomass

- $=$ old biomass
- $+$ growth
- $+$ recruitment
- $-$ natural mortality
- $-$ catch

Delay Difference Model

Observation equation:

$$y_t = Q\theta_t + v_t \qquad v_t \sim N(0, w_t\tau^2)$$

State equation:

$$\begin{aligned}\theta_{t+1} &= (1+\rho)e^{-M}(\theta_t - kC_t) - \\ &\quad \rho\, e^{-2M}\frac{(\theta_t - kC_t)}{\theta_t}(\theta_{t-1} - kC_{t-1}) + \\ &\quad r\left(1 - \rho\,\omega\, e^{-M}\frac{(\theta_t - kC_t)}{\theta_t}\right) + u_{t+1},\ u_{t+1} \sim N(0, \sigma^2)\end{aligned}$$

Model Parameters

39 unknown parameters to be estimated

- relative biomasses θ_t, $t = 1, \ldots, N = 34$
- population parameters $k, Q, r, \sigma^2, \tau^2$

Informative priors

Full Conditionals

E.g., full conditional posterior density for r:

$$
\begin{aligned}
p(r \mid \theta_t, k, Q, \sigma^2, \tau^2) &\propto p(r) \prod_{t=2}^{N} p(\theta_t \mid \theta_{t-1}, \theta_{t-2}, k, r, \sigma^2) \\
&\propto \frac{1}{r} \exp\left(-\frac{(\log r - \mu_r)^2}{2\sigma_r^2} - \frac{1}{2\sigma^2} \sum_{t=2}^{N} (\theta_t - g(\theta_t))^2 \right)
\end{aligned}
$$

Comparison

250,000 cycles of the Gibbs sampler

CPU time for

- ARMS: 124.86

Comparison

250,000 cycles of the Gibbs sampler

CPU time for

- ARMS: 124.86
- NKC: 301.03

Comparison

250,000 cycles of the Gibbs sampler

CPU time for

- ARMS: 124.86
- NKC: 301.03
- ATRIMS: 78.48

Comparison

250,000 cycles of the Gibbs sampler

CPU time for

- ARMS: 124.86
- NKC: 301.03
- ATRIMS: 78.48
- ATRAMS: 85.66

Comparison

250,000 cycles of the Gibbs sampler

CPU time for

- ARMS: 124.86
- NKC: 301.03
- ATRIMS: 78.48
- ATRAMS: 85.66

Discussion

- general class of adaptive MH algorithms

Discussion

- general class of adaptive MH algorithms
- finite horizon techniques, use m previous samples

Discussion

- general class of adaptive MH algorithms
- finite horizon techniques, use m previous samples
- adaptivity via MH-within-Gibbs

Discussion

- general class of adaptive MH algorithms
- finite horizon techniques, use m previous samples
- adaptivity via MH-within-Gibbs
- for univariate target: ATRIMS and ATRAMS

Discussion

- general class of adaptive MH algorithms
- finite horizon techniques, use m previous samples
- adaptivity via MH-within-Gibbs
- for univariate target: ATRIMS and ATRAMS
- local support of triangular/trapezoidal density results in less function evaluations than NKC

Discussion

- general class of adaptive MH algorithms
- finite horizon techniques, use m previous samples
- adaptivity via MH-within-Gibbs
- for univariate target: ATRIMS and ATRAMS
- local support of triangular/trapezoidal density results in less function evaluations than NKC
- starting knots of proposal allowed to depend on previously sampled points in contrast to ARMS

Discussion

- general class of adaptive MH algorithms
- finite horizon techniques, use m previous samples
- adaptivity via MH-within-Gibbs
- for univariate target: ATRIMS and ATRAMS
- local support of triangular/trapezoidal density results in less function evaluations than NKC
- starting knots of proposal allowed to depend on previously sampled points in contrast to ARMS
- work in progress: multivariate proposals

Reference

Meyer R, Cai B, Perron F (2008): Adaptive rejection Metropolis sampling using Lagrange interpolation polynomials of degree 2, *Computational Statistics and Data Analysis*, 52, 3408-3423.

Cai, B., Meyer, R., Perron, F. (2008): Metropolis-Hastings Algorithms with Adaptive Proposals. *Statistics and Computing*, 18, 421-433.