Beyond MCMC: two case studies and a few thoughts

Nicolas Chopin
(joint work with Tonly Lelièvre, Gabriel Stoltz, Judith Rousseau, Brunero Liseo)
nicolas.chopin@ensae.fr

ENSAE-CREST, Malakoff, FRANCE

Limitations of MCMC

- slow: error is $0(N^{-1/2})$, correlation between successive steps.
- local exploration: how to assess convergence if posterior is not unimodal?
- tedious to tune $\Rightarrow$ these algorithms are not robust.
- bugs are difficult to detect.

Alternatives

- fast approximations: variational Bayes, Expectation-Propagation, INLA (see Havard's talk).
- adaptive MCMC;
- adaptive biasing; e.g. Wang-Landau.
- Sequential Monte Carlo, see e.g. Doucet et al. (2006).

First case study

Gaussian Mixture model:

$$p(y_i|\theta) = \sum_{k=1}^{K} q_k \sqrt{\frac{v_k}{2\pi}} \exp\left\{-\frac{v_k}{2}(y_i - \mu_k)^2\right\}$$

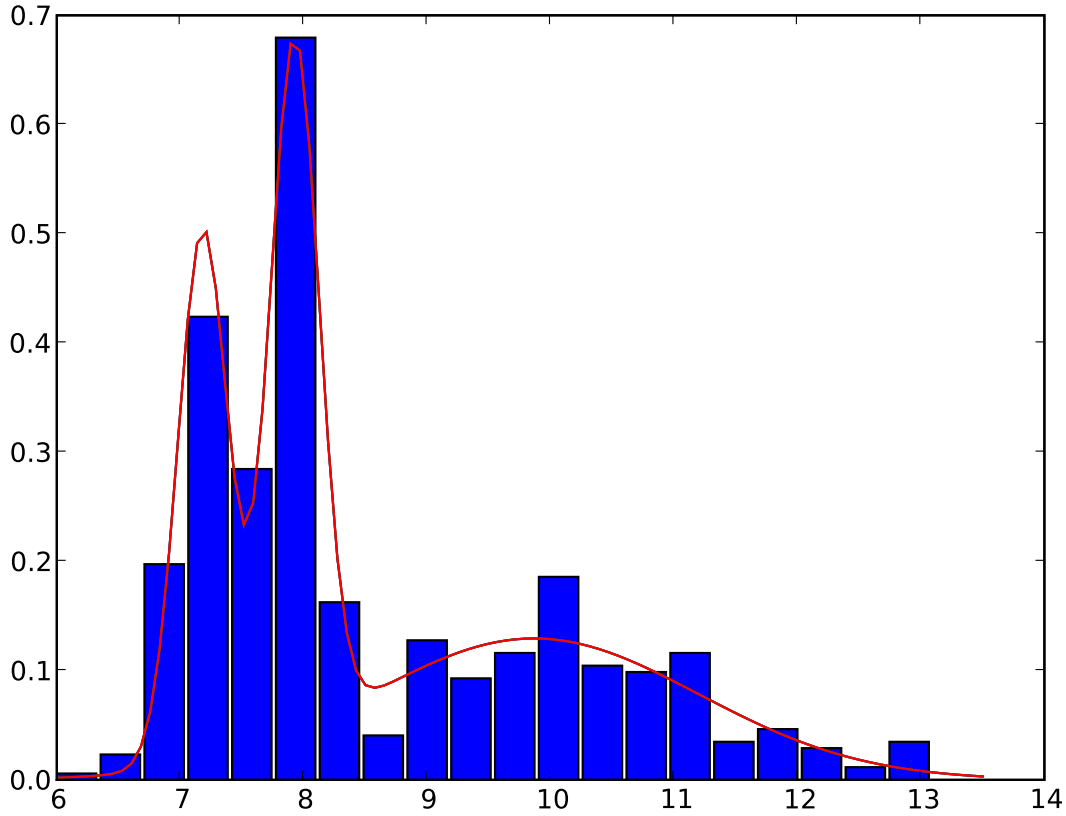
$$\mu_1, \ldots, \mu_K \sim N(m, s^2), \quad v_1, \ldots, v_K \sim G(a, b)$$

$$q_1, \ldots, q_K \sim Dir(1, \ldots, 1)$$

or equivalently

$$q_k = \omega_k / (\omega_1 + \ldots + \omega_K), \quad \omega_1, \ldots, \omega_K \sim Exp(1).$$

Hidalgo stamps dataset



Quotes

Jasra et al. (2005): " We feel that the Gibbs sampler run with completion is often not worth programming [...] since the chance of it falling to converge is too high."

Celeux et al. (2000): "Although somewhat presumptuous, we consider that almost the entirety of MCMC implemented for mixture models has failed to converge."

Molecular simulation

Concerns simulation of some continuous-time dynamics, e.g.

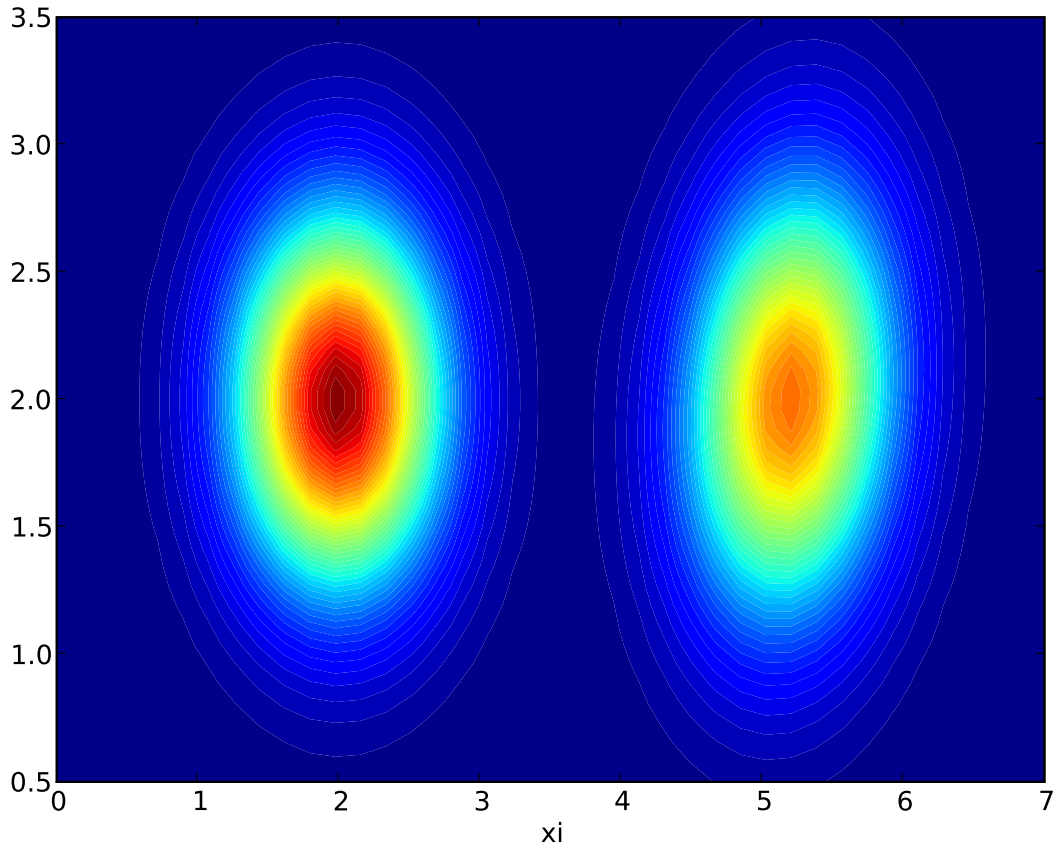
$$dX_t = -\nabla V(X_t)dt + (\beta/2)^{-1/2}dW_t$$

with associated Boltzmann-Gibbs density:

$$p(x) \propto \exp\{-\beta V(x)\}$$

(In practice, approximated by long runs of random walk Hastings-Metropolis.)

Metastability



Adaptive biased sampling

Aim is to sample from biased potential function

$$\widetilde{V} = V - F(\xi)$$

where F is the free energy, i.e.

$$\widetilde{p}(x) \propto \exp\left\{-\beta\widetilde{V}\right\}$$

is such that the margin of ξ is uniform.
At iteration t, perform Hastings-Metropolis move w.r.t.

$$\widetilde{V}_t = V - F_t(\xi)$$

and adjust F_t on the fly.

Wang-Landau (ABP)

At each iteration t, compute, for ξ in small bin $[e_k, e_{k+1}]$

$$F_t(\xi) = \frac{1}{N} \sum_{i=1}^{n} I\{\xi(x_i) \in [e_k, e_{k+1}]\}$$

i.e. penalises progressively already visited regions, see e.g. Atchadé and Liu (2004).

Adaptive Biasing Force (ABF)

$$E_{\widetilde{V}}\left[\frac{\partial V}{\partial \xi} - \frac{\partial F}{\partial \xi}\right] = 0$$

Progressively adapt the 'force' $\frac{\partial F}{\partial \xi}$, i.e. for $\xi \in [e_k, e_{k+1}]$

$$\frac{\partial F_t}{\partial \xi}(\xi) = \frac{1}{N}\sum_{i=1}^{n}\frac{\partial V}{\partial \xi}(x_i) I\{\xi(x_i) \in [e_k, e_{k+1}]\}$$

Back to mixture problems

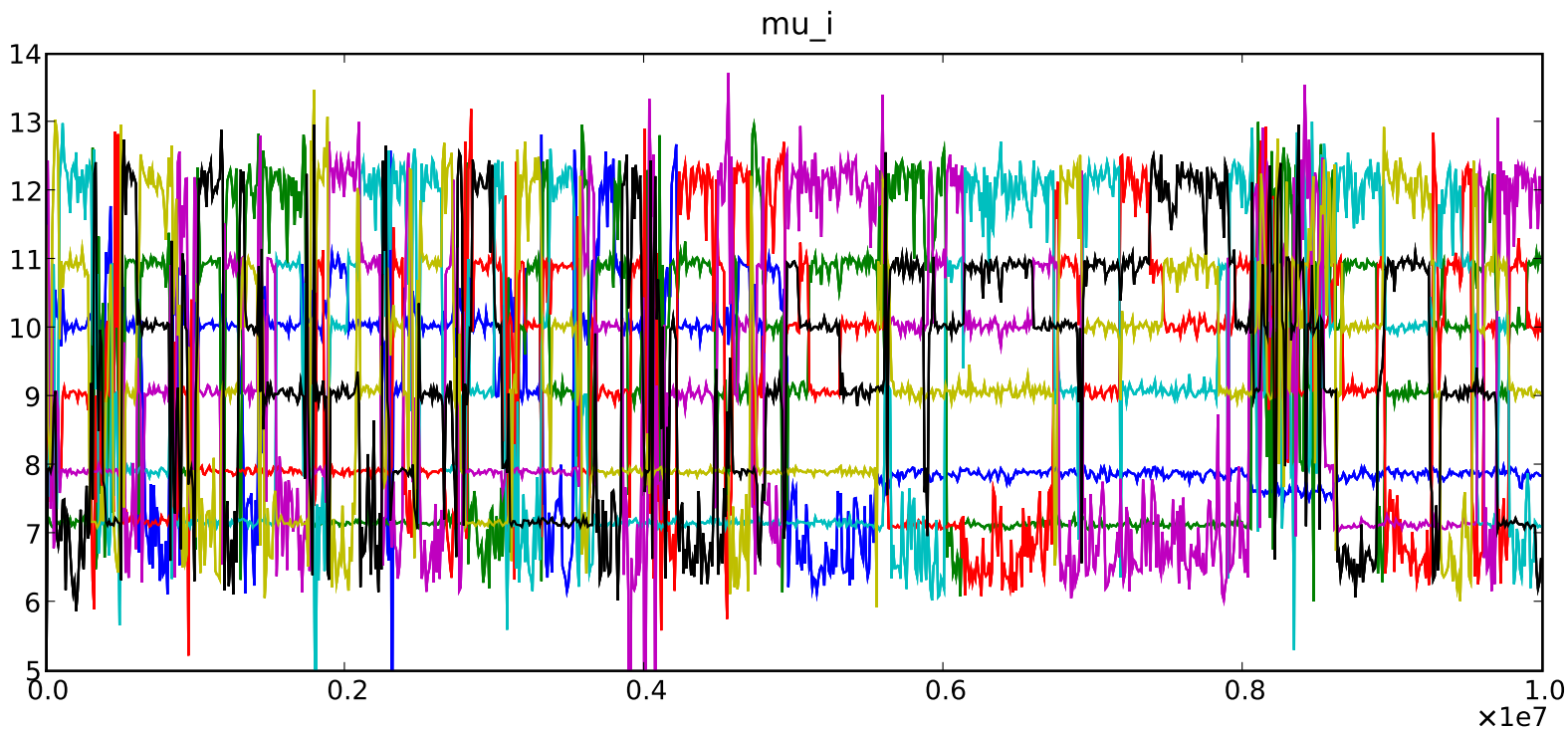
Our findings:

1. ABF outperforms ABP.
2. 'temperature' not necessarily the best ξ; certainly not the easiest to interpret.
3. seems much easier to find 'stable' directions, i.e. all symmetric functions of θ are 'stable'.
4. but careful with metaphors.

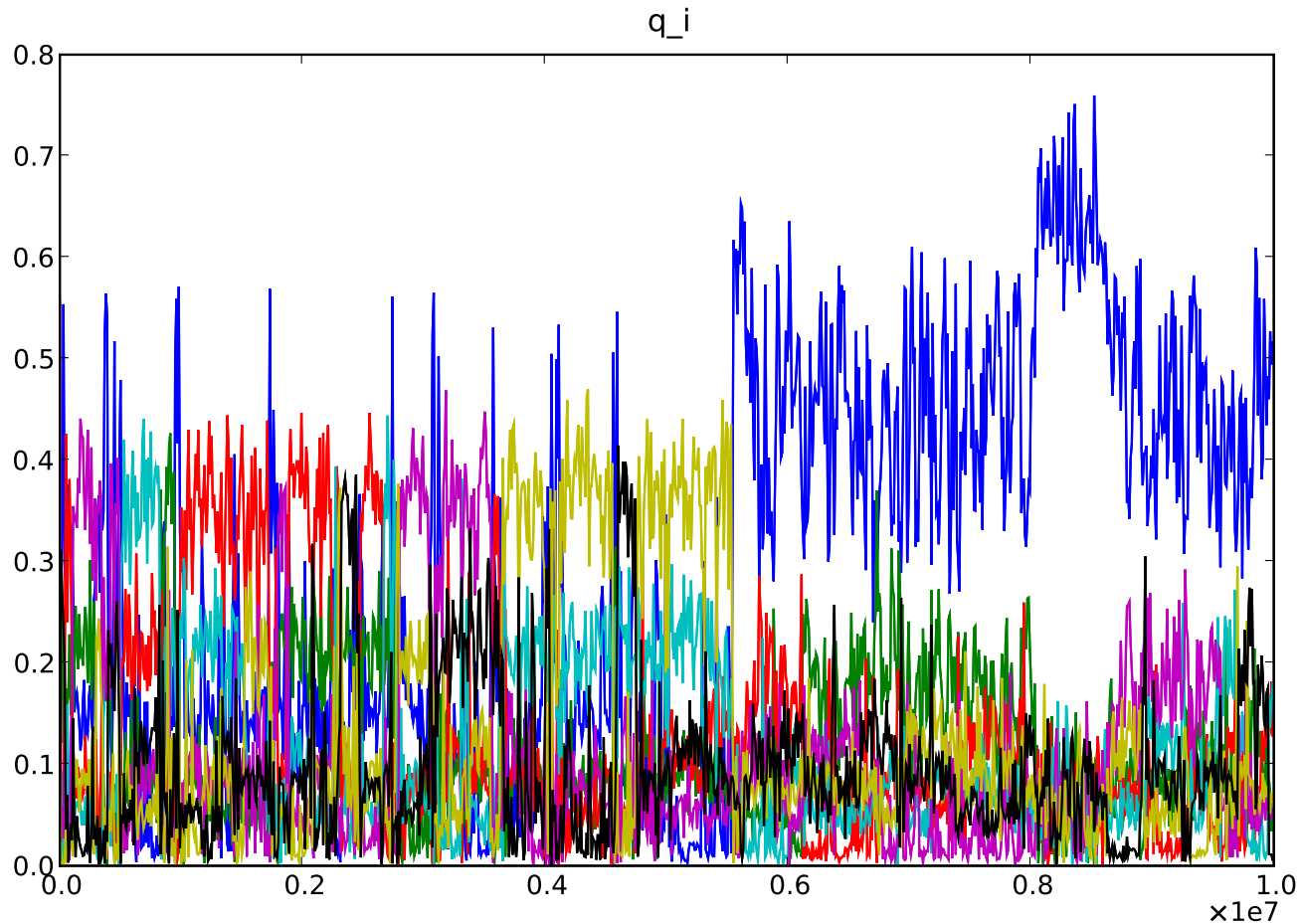
$\xi = q_1$

1. not symmetric;
2. constrained to $[0, 1]$;
3. Forcing q_1 close to one empty other components, and helps component switching, in a symmetric way;
4. Forcing q_1 close to zero: more later.

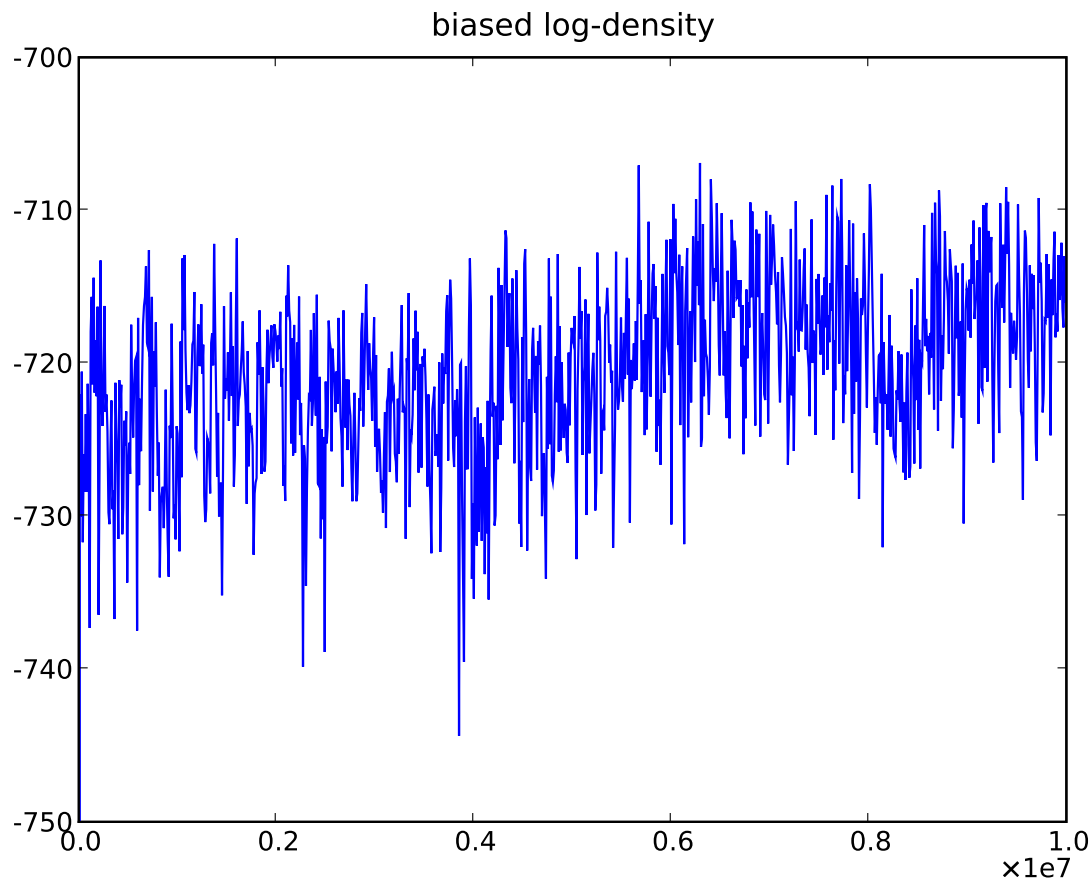
Hidalgo, $K = 7$ (I)



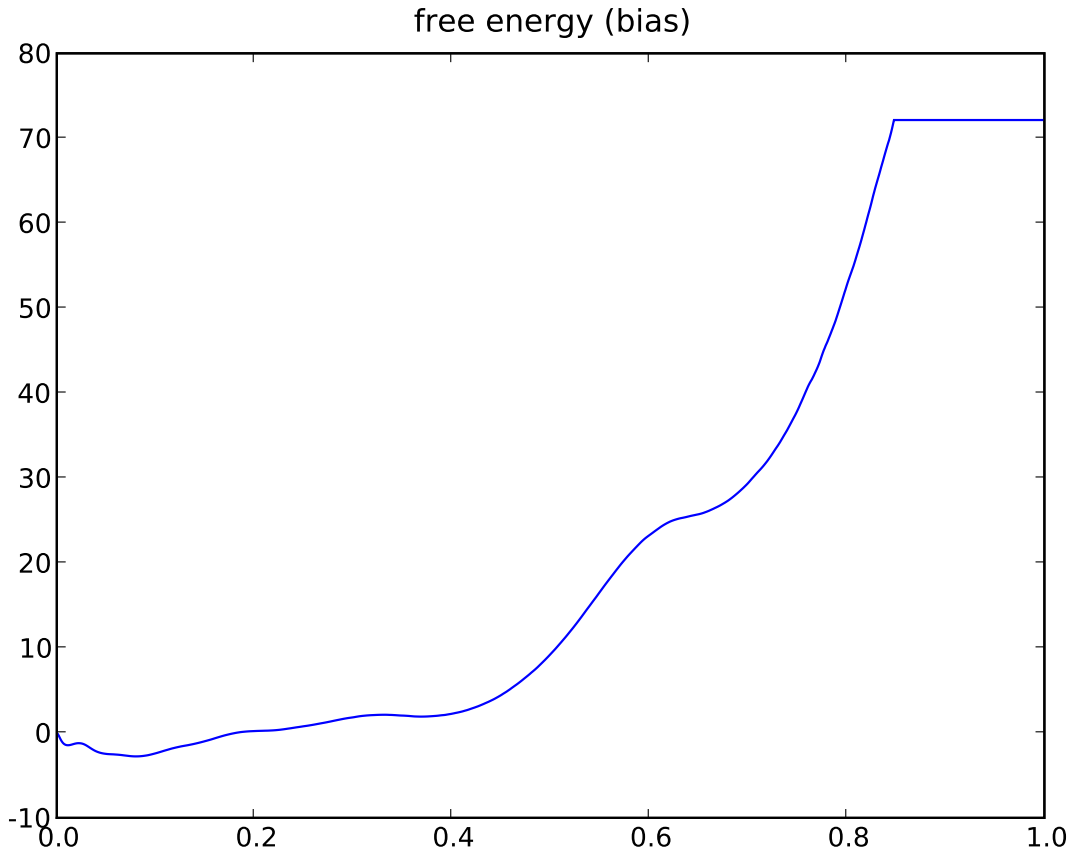
Hidalgo, $K = 7$ (II)



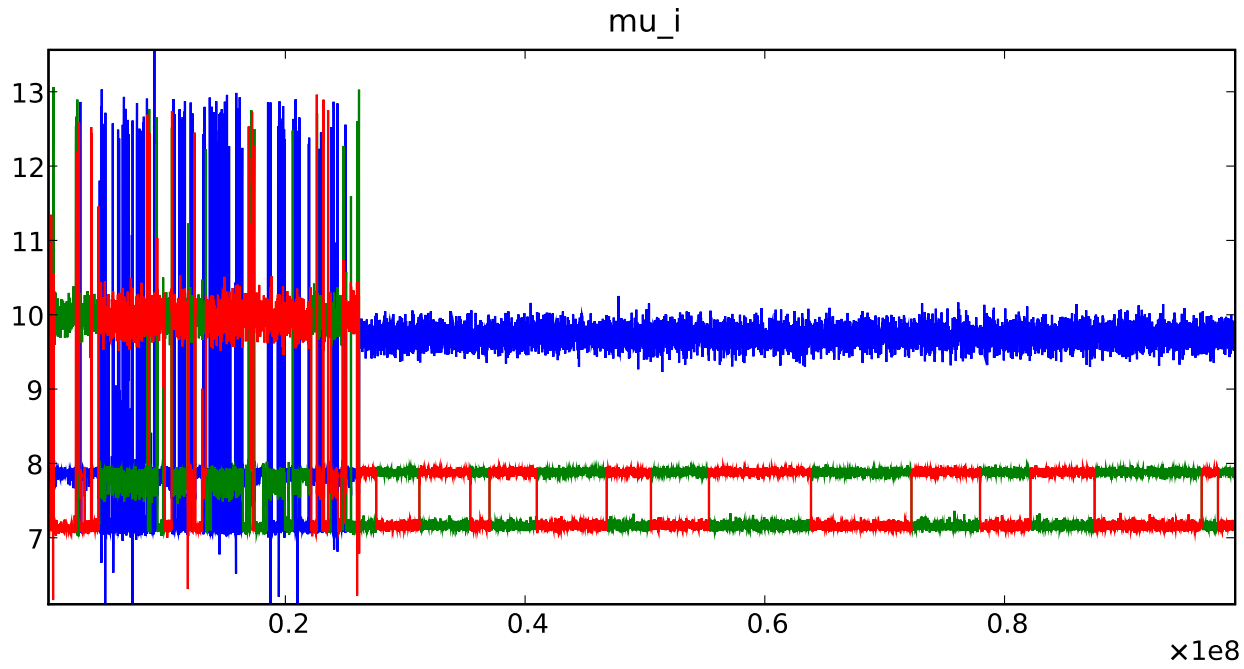
Hidalgo, $K = 7$ (III)



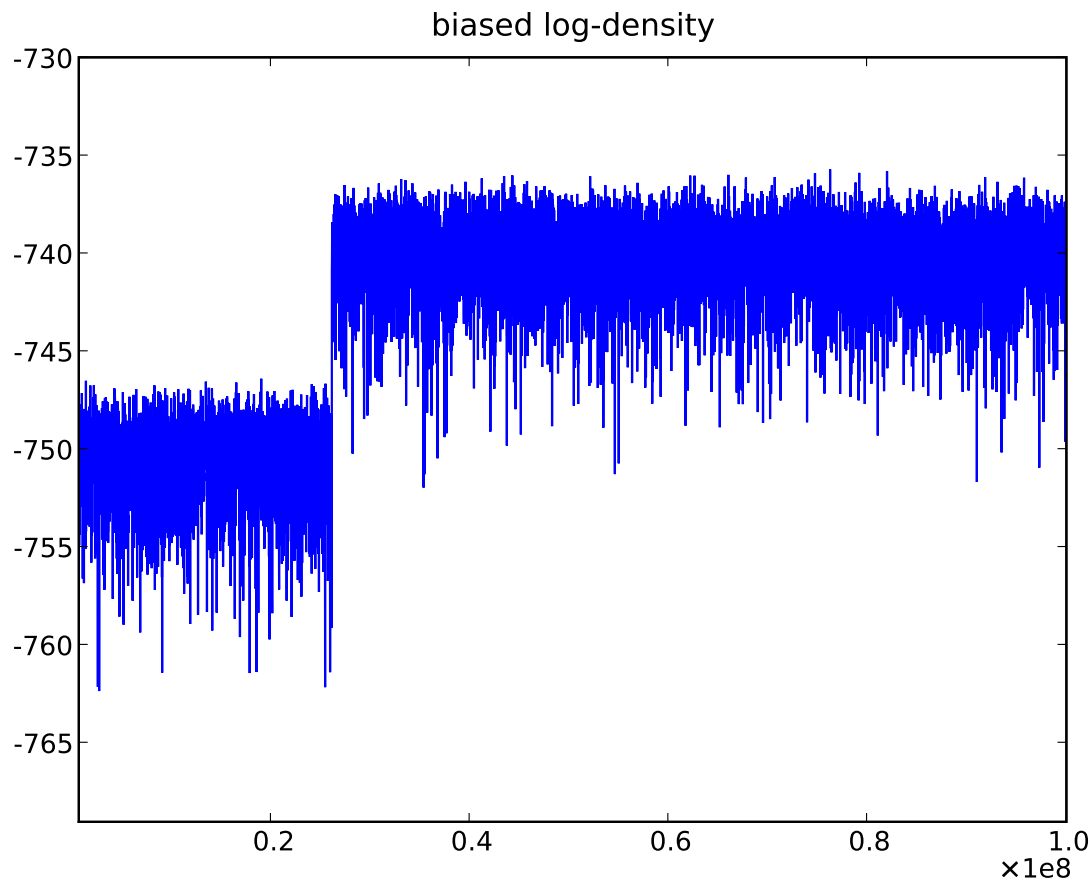
Hidalgo, $K = 7$ (IV)



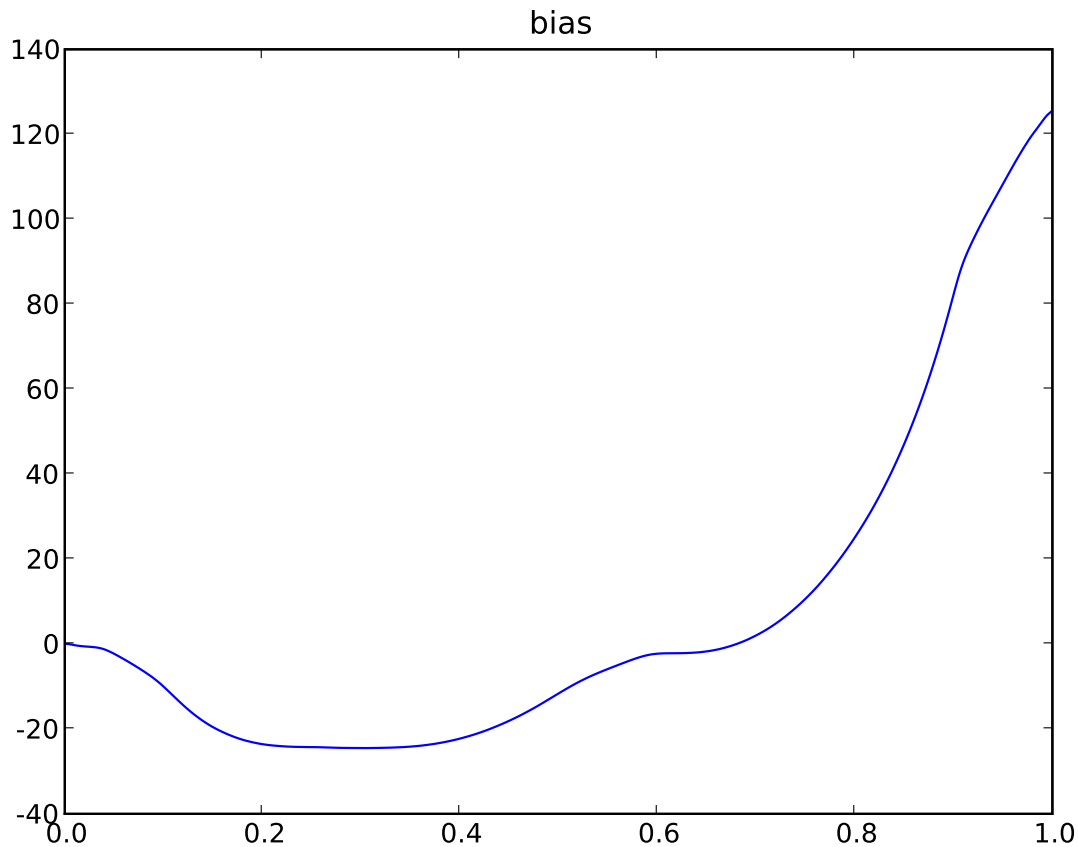
Hidalgo, $K = 3$ (I)



Hidalgo, $K = 3$ (II)



Hidalgo, $K = 3$ (III)



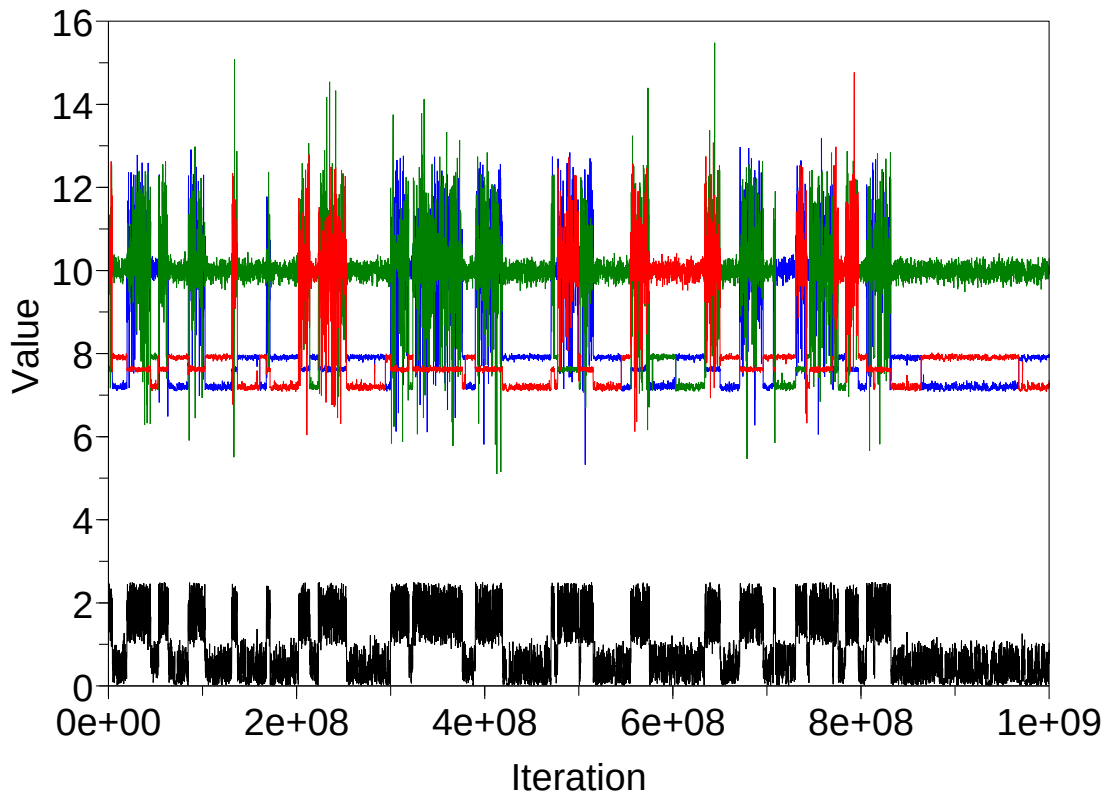
$\xi = \beta$

In the prior for inverse-variances:

$$v_i \sim G(a, \beta), \quad \beta \sim G(g, h)$$

1. symmetric;
2. constrained to $[0, 10 * b]$;
3. Large values of β penalise small variances.

Hidalgo, $K = 3$, $\xi = \beta$



Visiting $q_i \approx 0$ regions

allows us to estimate $p(y|K)/p(y|K-1)$:

$$\frac{p(y|K)}{\widetilde{p}(y)} = E_{\widetilde{V}}\left[\exp\{F(\xi)\}\right]$$

$$\frac{p(y|K-1)}{\widetilde{p}(y)} = \frac{1}{K}\sum_{i=1}^{K} E_{\widetilde{V}}\left[\frac{p(y|\theta)}{p(y|\theta[q_i=0])}\exp\{F(\xi)\}\right]$$

using importance sampling, or reversible jump (with birth-death steps).

Second case study

Nonparametric inference from long memory Gaussian processes:

$$Y|f \sim N\left(0, T(f)\right)$$

$$T(f)[k,l] = \gamma(k-l) = \int_{-\pi}^{\pi} e^{i(k-l)\omega} f(\omega)\, d\omega$$

with a nonparametric prior for f, e.g. (Rousseau et al., 2009)

$$f(\omega) = \left|1 - e^{i\omega}\right|^{-2d} \exp\left\{\sum_{j=0}^{K} \theta_j \cos(j\omega)\right\}$$

with $d \sim U(0, 1/2)$, $\theta_j \sim N(0, 1/j)$, $K \sim Poisson(1)$.

Computational difficulties

$$p(y|f) = (2\pi)^{-n/2} |T(f)|^{-1/2} \exp\left\{-\frac{1}{2} y^T T(f)^{-1} y\right\}$$

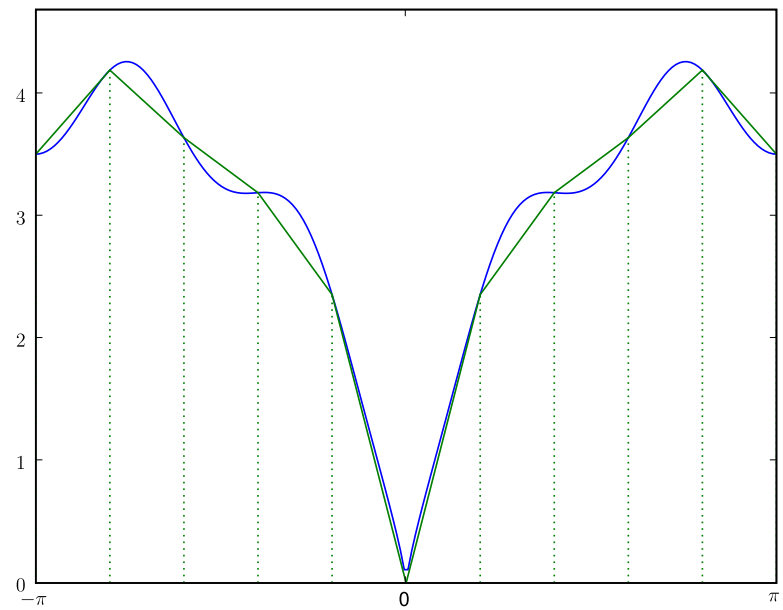
where $|T(f)|$ is easy to approximate, and

$$T(f)^{-1} \approx T(g), \quad g = \frac{1}{4\pi^2 f}$$

but we still have n integrals to compute, for $k = 1, \ldots, n$;

$$\int_{-\pi}^{\pi} e^{ik\omega} g(\omega)\, d\omega$$

Interpolation



If g is replaced by a linear interpolation, all the integrals can be computed exactly using one FFT of the points $g(\omega_k)$, $k = 1, \ldots, M$; the number of bins M must be $\geq 4N$. Cost is $O(M)$.

Sequential Monte Carlo

Consider the 'intermediate' likelihood functions $p_m(y|f)$ corresponding to

$$T_m[k,l] = \begin{cases} T(g)[k,l] & \text{if} |k-l| \leq 2^m \\ 0 & \text{otherwise} \end{cases}$$

then apply SMC to sequence:

$$p_t(f) \propto p(f) p_m(y|f)^{r/a} p_{m+1}(y|f)^{1-r/a}$$

for $t = am + r, r \in [0, a-1]$.
(Additional benefit: use Divide and Conquer structure of FFT.)

SMC steps

0. Draw $f_j \sim p(f)$, $j = 1, \ldots, N$. Set $t = 0$, $w_j = 1$.
1. Re-weight

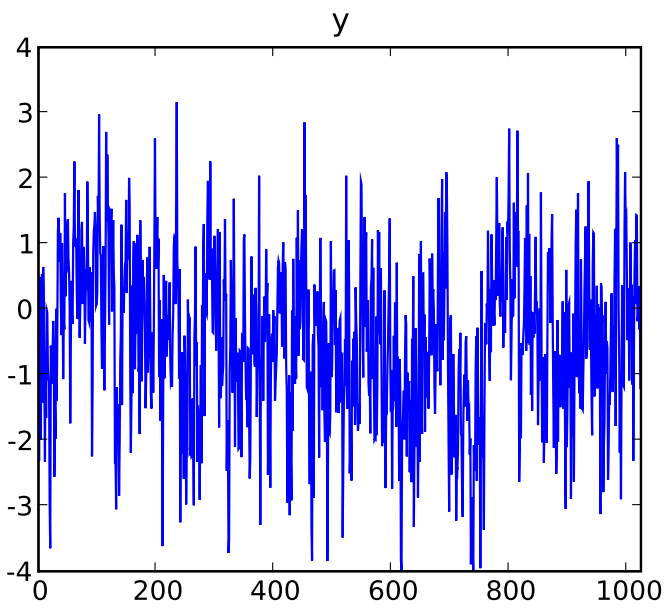
$$w_j = w_j \frac{p_{t+1}(f_j)}{p_t(f_j)}$$

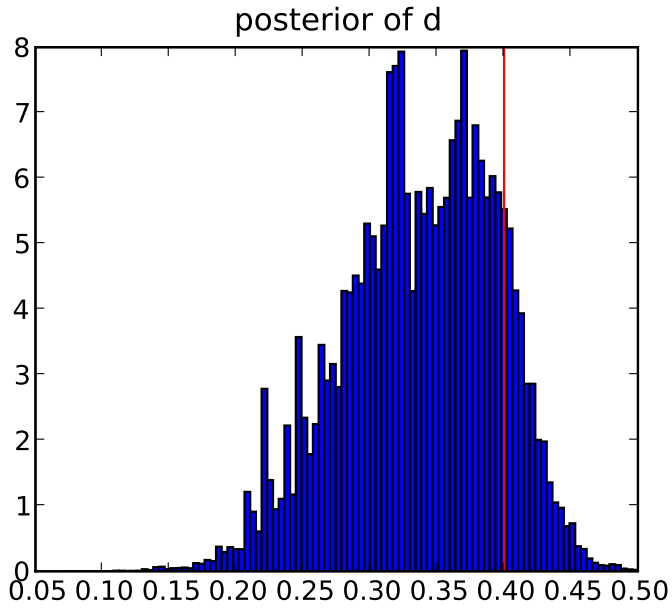
1. If Variance(weights)>γ,
   (a) resample
   (b) apply MCMC step w.r.t. p_{t+1}.
2. $t \leftarrow t + 1$. Go to 1.

Automatic tuning of MCMC

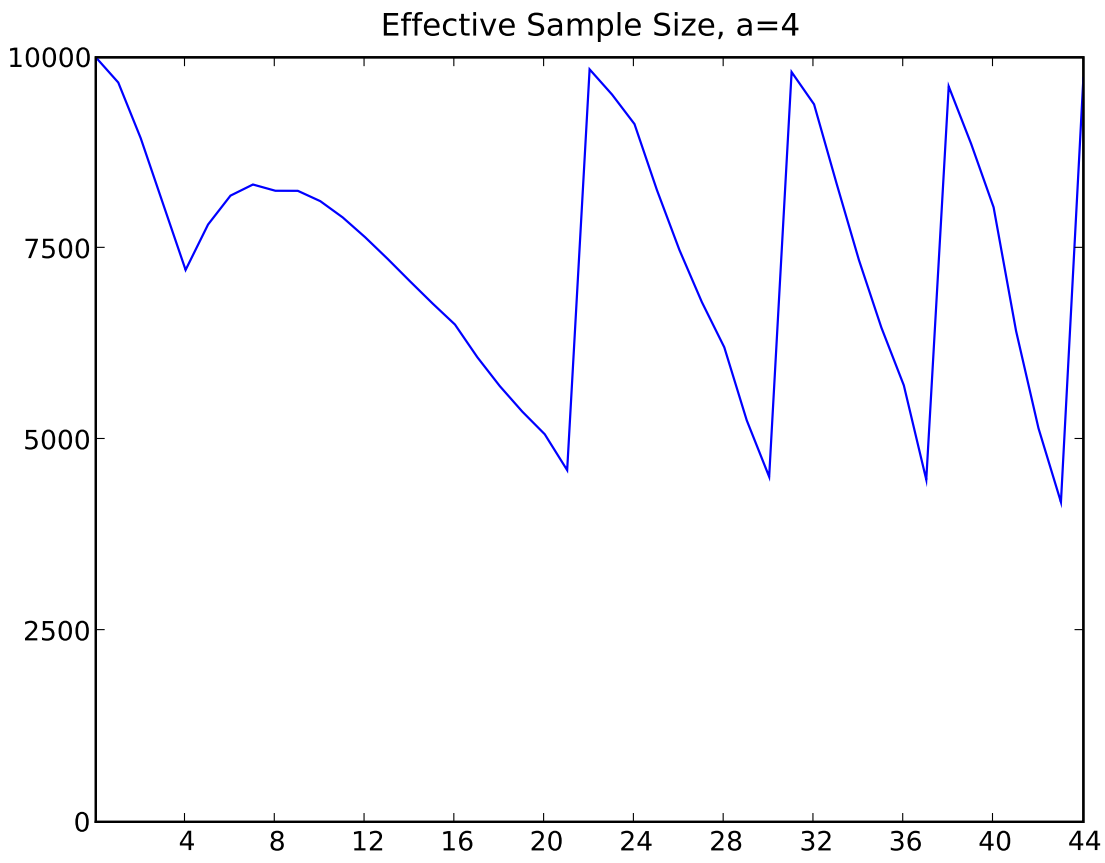
- Conditional on K, random walk HM steps, with proposal covariance tuned to current particle system.
- To move K, use reversible jump with Green (2003, HSS book)'s proposal.
- Or just use positive discrimination (Chopin, 2007), i.e. Wang-Landau biasing w.r.t. K.

A simulated example

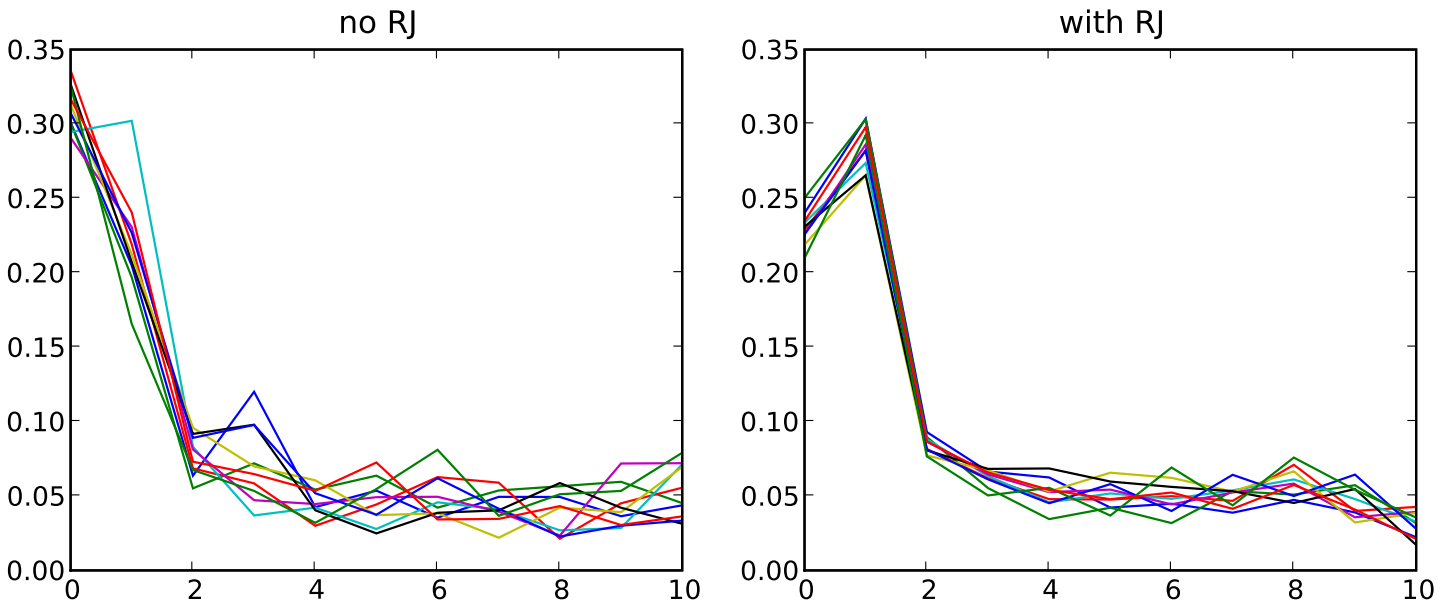




A simulated example (II)



A simulated example (III)



Conclusions

- thriving to provide automatic/robust algorithms.
- SMC, together with biasing methods, seem excellent candidate for this purpose:
  1. Biasing (a) makes target easier to explore the target (b) facilitates construction of proposals
  2. in SMC, tuning proposals is trivial.
- Currently working on a SMC version of ABF.