Parallel Tempering

Mateusz Łącki and **Błażej Miasojedow**
Wydział Matematyki, Informatyki i Mechaniki
University of Warsaw, Poland
mateusz.lacki@biol.uw.edu.pl
B.Miasojedow@mimuw.edu.pl

Bayesian Inference in Bioinformatics

- Suppose we can measure some quantity y. Assume, that parameter α describes y's distribution
  - → let both be random and their joint density $g(y, \alpha)$ factorise so that $g(y, \alpha) = h(y|\alpha)f(\alpha)$, where f is *a priori* distribution on the parameter
  - → f might result from an underlying physical theory
- Real sample points $\mathfrak{y} = [y_1, \dots, y_{\mathrm{M}}]$ are observed
- The *a posteriori* distribution of α given the sample $\mathfrak{y}$, $f(\alpha|\mathfrak{y})$, describes how our knowledge about the studied quantity x is influenced by empirical evidence collected in $\mathfrak{y}$
  - → Obtain it via the Bayes Formula

$$f(\alpha|\mathfrak{y}) = \frac{h(y_1|\alpha)\dots h(y_{\mathrm{M}}|\alpha)f(\alpha)}{\int h(y_1|\beta)\dots h(y_{\mathrm{M}}|\beta)f(\beta)\mathrm{d}\,\beta}$$

Applications

- Hierarchical modelling for identification of co-expression patterns in microarray data by cluster analysis (Medvedovic *et al.*, 2004; Stingo and M., 2010)
- Assessing the importance of explanatory variables (Stingo and M., 2010)
- Model Selection

MCMC

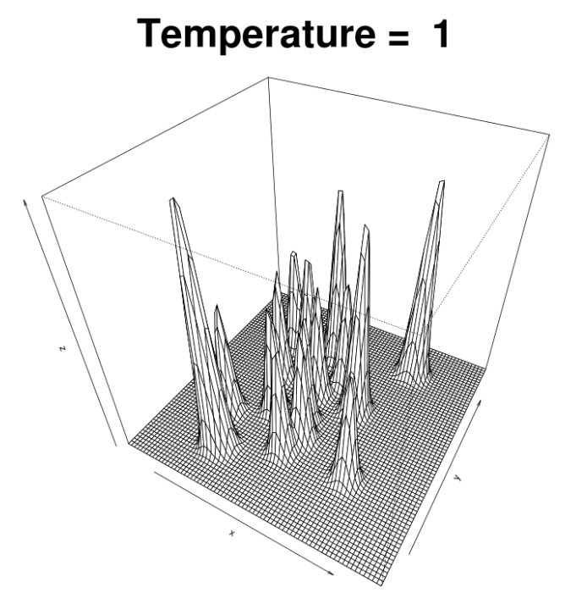
- MCMC algorithms are used to simulate samples out of analytically untractable posterior distributions
  - → Most popular algorithm: Green-Metropolis-Hastings (Geyer, 2012)
  - → Generates a sequence of points that are thought of as being an instantiation of a Markov Chain, $X \equiv \{X^{[k]}\}_{k=0}^{\infty}$
  - → Each point $X^{[k]}$ is generated by accepting or rejecting at random a step proposal from the chain's last position $X^{[k-1]}$
  - → Approximates, thanks to Ergodic Theory, integrals

$$\mathcal{E}g(X) = \int_{\Omega} g(x)\pi(x)\mathrm{d}\,x \approx \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} g(X^{[i]}),$$

where π is the density of a posteriori distribution. In particular: approximates probabilities of any measurable set, $\mathcal{P}(A)$

- GMH estimates may suffer from poor mixing
  - → Chain X restricted to user-provided number of iterations N could get stuck in a probability cluster
  - → Multimodial priors result in multimodial posteriors
  - → Multimodial priors are selected when we suspect that the phenomenon under study is not concentrated around a particular point

Example



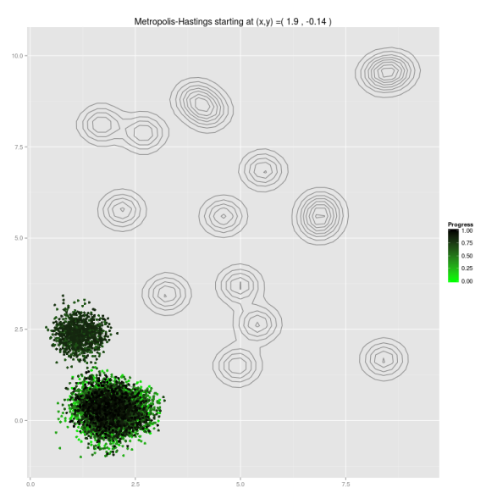
- Let π be a mixture of normal distributions

$$\pi(x) = \sum_{i=1}^{20} \frac{\omega_i}{\sigma_i\sqrt{2\pi}} \exp\Big(- \frac{(x-\mu_i)^{\mathrm{t}}(x-\mu_i)}{2\sigma_i^2}\Big)$$

where σ_i are standard deviations, ω_i are weights, and μ_i are means (Baragatti *et al.*, 2013)

- Some of the peaks mingle together to form bigger ones

- GMH draws sample points from only two modes that are not far away from the starting points drawn at random from the region of the suspected probability concentration
- Estimates of probabilities and moments are totally fallacious



¿Question?

How can we enhance mixing so that the State Space is better searched for probability clusters?

Parallel Tempering a.k.a. Replica Monte Carlo

- Foundations of PT laid by Swendsen and Wang (1986)
- Generates several chains $X = [X_1, \dots, X_{\mathrm{L}}]$ and consists of two phases

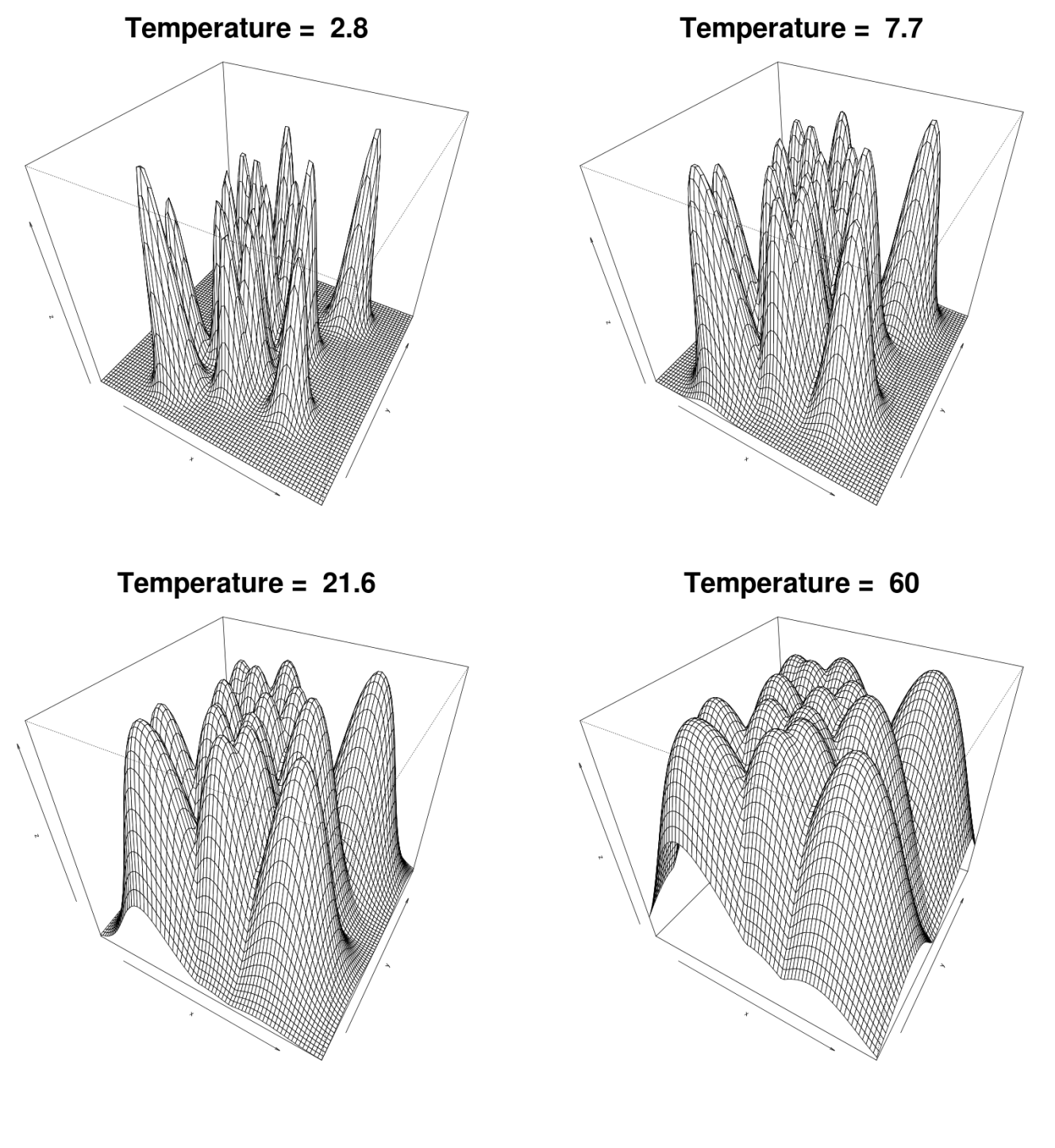
Ph I Drawing a point $\tilde{X}_l$ from π^{β_l}, where $1 = \beta_1 > \dots > \beta_{\mathrm{L}} > 0$ are called inverse temperatures (note that first coordinate corresponds to our initial problem)

Ph II Swaping some of $\tilde{X}$ coordinates at random: the Swap Strategy

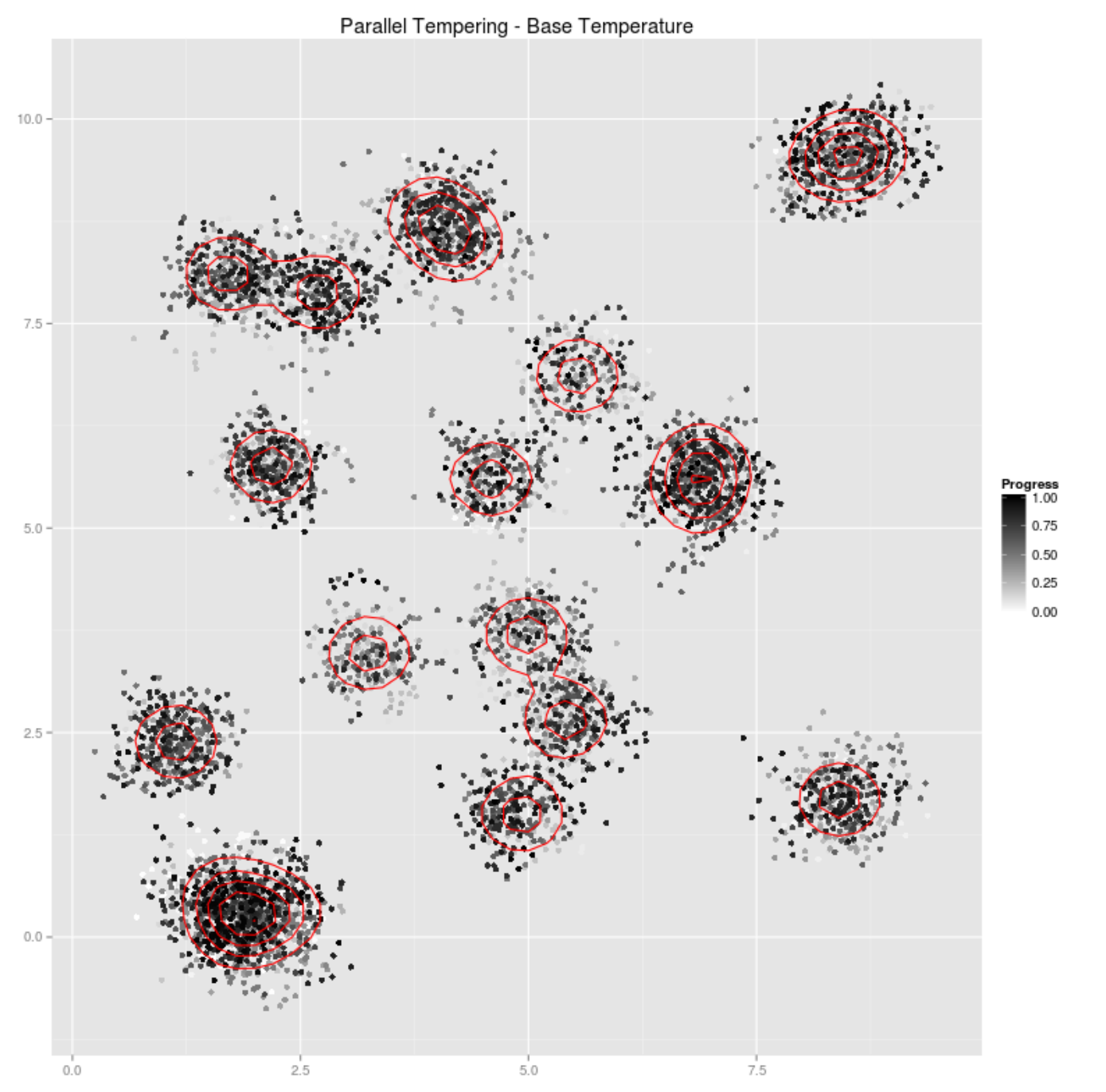
- Ph I mitigates the impact of multimodiality enlarging the probability of accepting steps from regions which the GMH would judge unlikely to draw from
- Ph II allows the passing of information from different chains: otherwise they would operate independently and first chain would gain nothing from other coordinates

Parallel Tempering at work

- More tempered chains are at ease when passing from one mode to another…



- …and Ph II assures that the base temperature chain explores more modes



Different Swap Strategies

- In Ph II one can implement a multitude of Swap Strategies
- Suppose that $\tilde{X}^{[k]} = x$. Then the distribution on the indices can be described as $p(i, j|x)$. We explored the following strategies ($\propto$ denotes proportionality and $\wedge$ - the minimum)

$$p(i,j|x) \propto \frac{\pi(x_j)}{\pi(x_i)} \wedge \frac{\pi(x_i)}{\pi(x_j)}$$

Strategy 1 promotes swaps between coordinates or relatively the same level, i.e. $\pi(x_j) \approx \pi(x_i)$

Strategy 2 breaks the symmetry of the previous one, giving more attention to swaps into regions of higher probability

$$p(i,j|x) \propto \frac{\pi(x_j)}{\pi(x_i)} \wedge 1$$

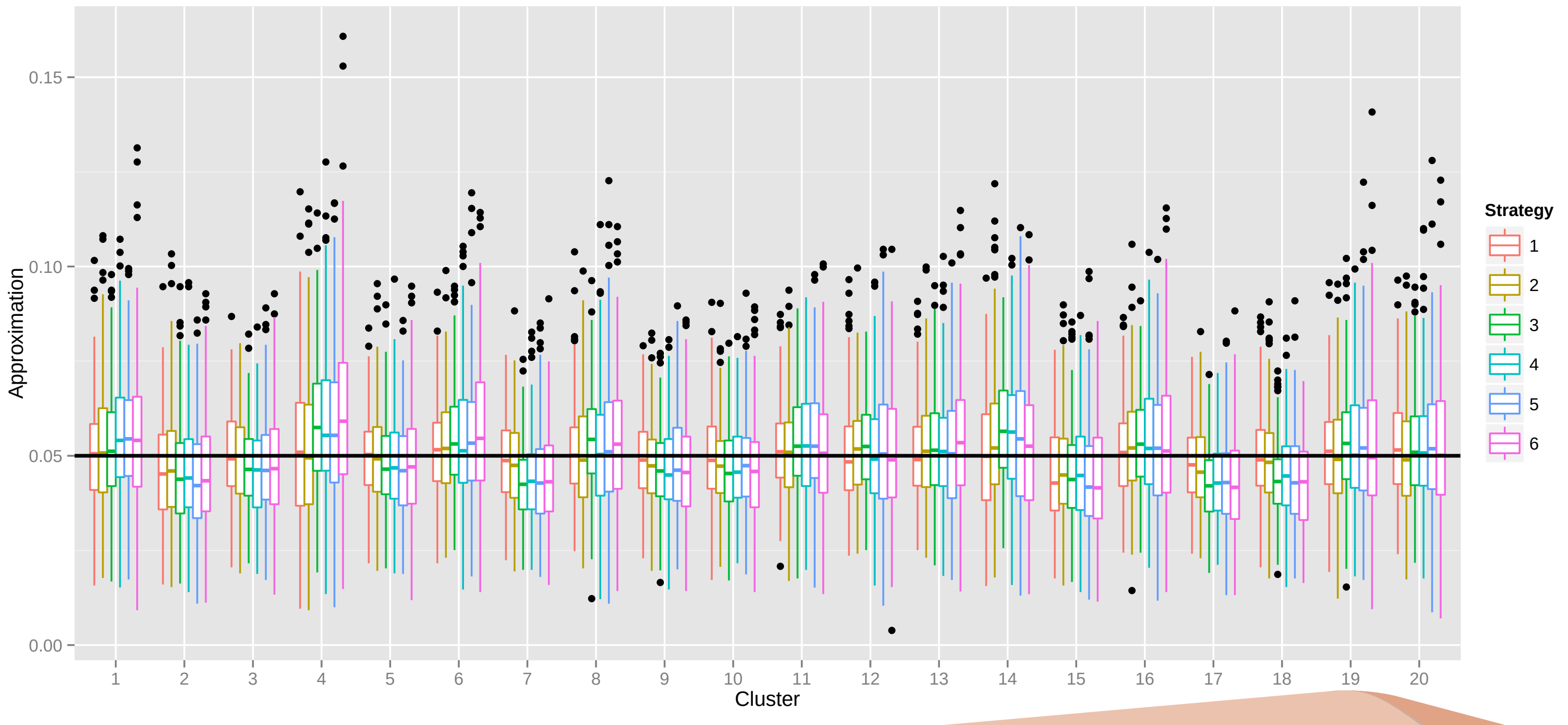
$$p(i,j|x) \propto \Big(\frac{\pi(x_j)}{\pi(x_i)} \wedge \frac{\pi(x_i)}{\pi(x_j)}\Big)^{\beta_i - \beta_j}$$

Strategy 3 softens the requirement that $\pi(x_j) \approx \pi(x_i)$ for similarly tempered coordinates, i.e. where $\beta_i - \beta_j \approx 0$: swaps between adjacent chains get more probable

Strategy 4 generalises the last one favouring more distant choices: ρ might any metric (*e.g.* euclidean)

$$p(i,j|x) \propto \Big(\frac{\pi(x_j)}{\pi(x_i)} \wedge \frac{\pi(x_i)}{\pi(x_j)}\Big)^{\frac{\beta_i - \beta_j}{1+\rho(x_i,x_j)}}$$

- All the above strategies explicitly refer to values of π in points drawn in Ph I making the draws computationally cheap
- Strategies 5 and 6 are independent of evaluations of π giving equal probability to all possible and all neighbouring swaps respectively
- Beneath we represent results obtained by all the strategies when trying to approximate the values of different modes: they should be equal roughly to 0.05, π being a mixture of 20 equally probable normal distributions. Results were obtained after 240 runs of PT for each Strategy, with 2500 steps of burn-in and 7500 steps of simulations. Note that evaluation-independent strategies have more extreme outliers



¡Good Software Available Soon!

- An R package, under the working name of StochasticSimulations, will be soon available for widespread use for free
- Among its features
  - → Division of the simulations into modules: Algorithm, State Space, Target Measure will provide a logic for the implementation of different models
  - → Implementation of the most common choices for State Space: $\mathbb{R}^{\mathrm{S}}$ and Discrete state space
  - → Implementation of the GMH and PT algorithms with the above-mentioned Strategies
  - → GGPLOT 2 based visualisations
- …and much, much more: stay tuned!

References

Baragatti, M., Grimaud, A. and Pommeret, D. (2013). Parallel tempering with equi-energy moves. *Statistics and Computing*, **23** (3), 323–339.

Geyer, C. J. (2012). *Markov Chain Monte Carlo Lecture Notes*. Unpublished.

Medvedovic, M., Yeung, K. and Bumgarner, R. (2004). Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, **20** (8), 1222–1232.

Miasojedow, B., Moulines, E. and Vihola, M. (2013). An adaptive parallel tempering algorithm. *Journal of Computational and Graphical Statistics*, **I** (january).

Stingo, F. and M., V. (2010). *Bayesian Statistics 9*, Oxford University Press, chap. Bayesian Models for Variable Selection that Incorporate Biological Information.

Swendsen, R. and Wang, J.-S. (1986). *Replica Monte Carlo Simulation of Spin-Glasses, Parallel Tempering*.