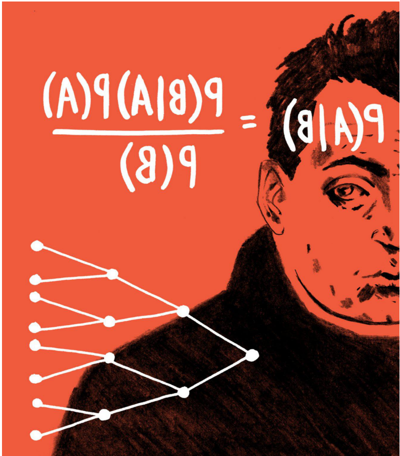
A MODERN INFERENCE TOOLBOX

Towards fast, scalable Bayes

12. Dezember 2014



Dr Ewan Cameron

University of Oxford (Zoology)
SEEG / **Malaria Atlas Project**

OVERVIEW

1. Where we are: Bayes in astronomy

2. Why is Bayes computationally challenging?

3. Towards fast, scalable Bayes

WHERE WE ARE: BAYES IN ASTRONOMY

WHERE WE ARE IN ASTRONOMY

- ⋆ the Bayesian approach is becoming ever more popular: a trend largely shaped by in-field textbooks such as
  - [Bayesian Logical Data Analysis for the Physical Sciences]] (Gregory, 2010)
  - [Practical Statistics for Astronomers] (Wall & Jenkins, 2012)
  - [Statistics, Data Mining, and Machine Learning in Astronomy] (Ivezic et al., 2014)

- ⋆ and occasionally by more general references: e.g.
  - [Bayesian Data Analysis] (Gelman et al., 2004/2013)
  - [Gaussian Processes for Machine Learning] (Rasmussen & Williams, 2006)

WHERE WE ARE IN ASTRONOMY

- ⋆ astrostatistics themed conferences & summer schools are now a thing:
    - the [Statistical Challenges in Modern Astronomy'] series at Penn State
    - [Statistical Challenges in 21st C. Cosmology] (IAUS306)
    - [Bayesian astrophysics: XXVI Canary Islands Winter School of Astrophysics]
- ⋆ as are astrostatistics sessions at statistics conferences:
    - at the ISI World Statistics Congress (e.g. HK 2013)
    - at ISBA (e.g. Kyoto 2013)

WHERE WE ARE IN ASTRONOMY

- early adopters have had great success applying established techniques: e.g.
  - hierarchical modelling (e.g. SNa fitting: Mandel et al., 2011; stellar eccentricities: Hogg et al., 2010)
  - Gaussian processes for spatial fields (e.g. Wandelt et al., 2004; Jasche & Kitaura, 2010)
  - Gaussian processes as non-parametric noise models (e.g. Gibson et al., 2012)
  - Bayesian model averaging / model selection (e.g. cosmology: Trotta 2008; exoplanets: Feroz et al., 2011)
  - Approximate Bayesian Computation (e.g. Cameron & Pettitt, 2012; Weyant et al., 2013)

WHERE WE ARE IN ASTRONOMY

- ⋆ nevertheless, few graduates will have received any statistical training by the end of their PhD
- ⋆ most astro-statisticians have only a superficial knowledge of mathematical statistics: this limits the potential for cross-disciplinary exchange
- ⋆ we're rarely at the forefront of methodology (with the exception of nested sampling, perhaps)
    - cf. ``What we talk about when we talk about fields'' (Cameron, 2014: http://arxiv.org/abs/1406.6371)
- ⋆ in-depth (i.e., full-time / funded) collaborations with statisticians are rare

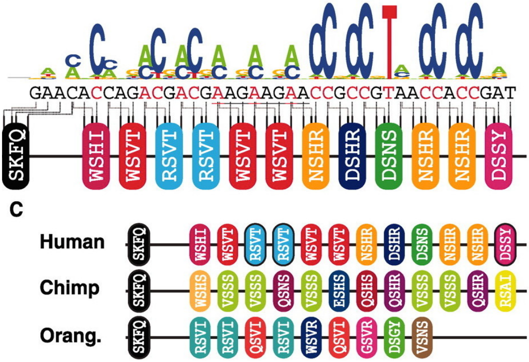
BY CONTRAST: IN GENETICS, BIOLOGY, AND EPIDEMIOLOGY:

⋆ in the ``bio-sciences'' long-term collaborations with top statisticians are fostered to develop cutting edge techniques for domain-specific applications ...

... with exciting results & high impact publications!

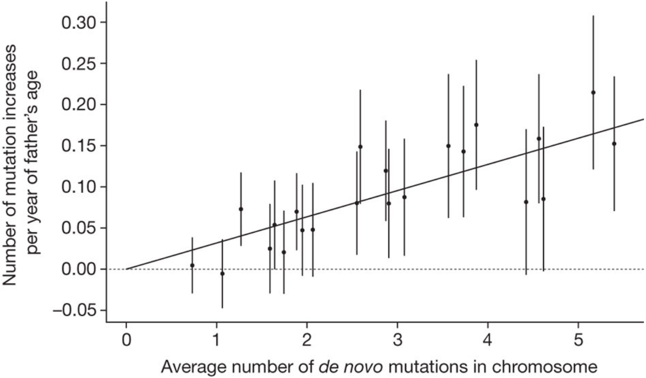
BY CONTRAST: IN GENETICS, BIOLOGY, AND EPIDEMIOLOGY:

- ⋆ Myers et al. (2010) in Science: ``Drive Against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination''



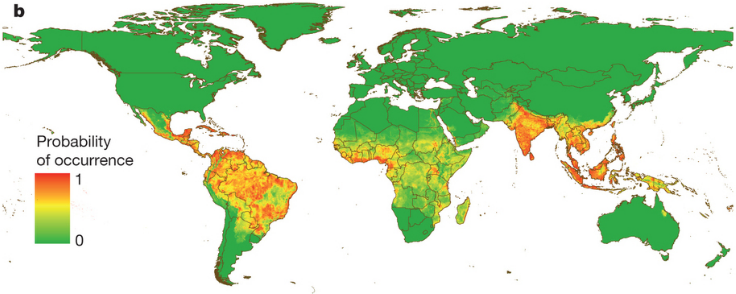
BY CONTRAST: IN GENETICS, BIOLOGY, AND EPIDEMIOLOGY:

- ⋆ Kong et al. (2012) in Nature: ``Rate of de novo mutations and the importance of father's age to disease risk''



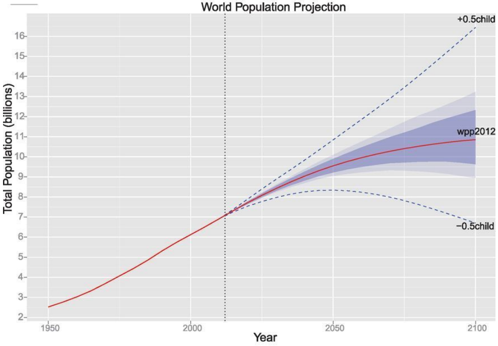
BY CONTRAST: IN GENETICS, BIOLOGY, AND EPIDEMIOLOGY:

- ⋆ Bhatt et al. (2013) in Nature: ``The global distribution and burden of dengue''



BY CONTRAST: IN GENETICS, BIOLOGY, AND EPIDEMIOLOGY:

⋆ Gerland & Raftery et al. (2014) in Science: ``World population stabilization unlikely this century''



WHERE WE ARE IN GRAVITATIONAL WAVE ASTRONOMY

- ⋆ we're at Type I civilization (on the Kardashev scale): we use all the resources on our planet, but haven't ventured further afield!
  - Bayesian model selection is well used: typically implemented via nested sampling with `multinest`
  - hierarchical models solve routine inference problems like dealing with selection bias
  - some novel techniques for speeding up likelihood evaluations are being explored

WHERE WE ARE IN GRAVITATIONAL WAVE ASTRONOMY

- ⋆ we're at Type I civilization (on the Kardashev scale): we use all the resources on our planet, but haven't ventured further afield!
  - Bayesian model selection is well used: typically implemented via nested sampling with `multinest` [but no sign of SMC or particle filtering]
  - hierarchical models solve routine inference problems like dealing with selection bias [but no sign of JAGS/STAN, or non-parametrics]
  - some novel techniques for speeding up likelihood evaluations are being explored [but no sign of ``Russian roulette'' series truncations, the unscented transform, or subset posteriors]

WHY IS BAYES COMPUTATIONALLY CHALLENGING?

WHY IS BAYES COMPUTATIONALLY CHALLENGING?

- ⋆ the complex models required for `real-world' problems rarely allow for analytic solutions or offer low-dimensional sufficient statistics
- ⋆ approximate inference via Markov Chain Monte Carlo is usually possible (unless the likelihood is unavailable: hence, doubly intractable $\rightarrow$ ABC), but:
  - posterior estimates convergence as $O(N^{-1/2})$ (cf. Tierney, 1994): so ``it is easy to get rough estimates, but nearly impossible to get accurate ones'' (Rue et al., 2008)
  - the design of an effective (efficient) MCMC transition kernel becomes more and more difficult as the model dimension increases

``MO DATA, MO PROBLEMS''

⋆ MCMC is not a natural tool for Big Data inference:

- for iid observations the cost of likelihood function evaluation goes as $O(n)$;
- while for more complex models (e.g. Gaussian processes; cf. Neal 1997) it can be as bad as $O(n^3)$;

- it's non-sequential: every time we add new data we need to recompute the posterior more-or-less `from scratch'
- it's difficult to parallelise efficiently without further approximation (cf. Scott et al. 2013)

TOWARDS FAST, SCALABLE BAYES

PREVIEW

[i] Hamiltonian MCMC

[ii] Sequential Monte Carlo

[iii] Pseudo-marginal MCMC

[iv] Consensus/Median Posteriors

HAMILTONIAN MCMC

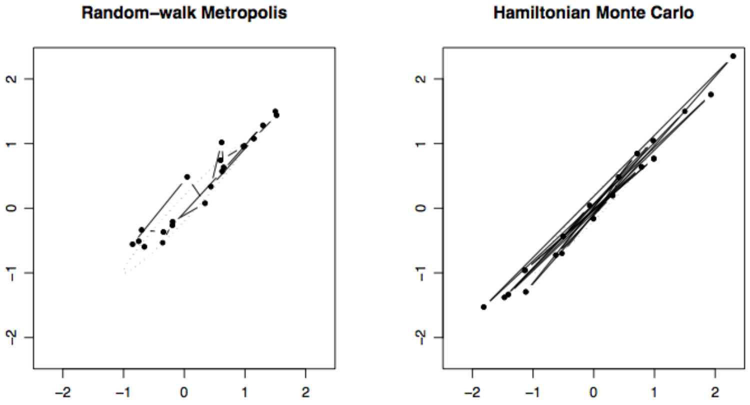
⋆ Hamiltonian MCMC = exploring the posterior with Hamiltonian dynamics:

- introduces a ``momentum'' term to the posterior: $H(\theta, p) \propto \exp\left(\log\left(\pi(\theta|y)\right) - \frac{1}{2} p \cdot p\right)$
- improves the efficiency scaling with dimension: $\sim O(d^{\frac{5}{4}})$ compared to $O(d^2)$ for random walk MCMC
- but requires `tuning' and computation/estimation of the gradient of H;

- see for reference: Neal (2012) [arXiv:1206.1901]
- first use in GW astronomy? Lentati et al., 2013

HAMILTONIAN MCMC

⋆ RW MCMC vs HMC



HAMILTONIAN MCMC

- ⋆ to-date there have been few astronomical applications except to large-scale structure (cf. Jasche & Kitaura 2010)
  - presumably because it's non-trivial to code from scratch and then tune to a given problem

- ⋆ hopefully this will change thanks to STAN (mc-stan.org):
  - an open source package for MCMC sampling with HMC
  - includes the self-tuning NUTS (No-U-Turn) HMC sampler (Hoffmann & Gelman, 2011)
  - a BUGS/JAGS style programming interface lets the user quickly build and sample from hierarchical models

HAMILTONIAN MCMC

- Example STAN code

```
model {
  matrix[N,N] Sigma;
  // off-diagonal elements
  for (i in 1:(N-1)) {
    for (j in (i+1):N) {
      Sigma[i,j] <- eta_sq * exp(-rho_sq * pow(x[i] - x[j],2));
      Sigma[j,i] <- Sigma[i,j];
    }
  }
  // diagonal elements
  for (k in 1:N)
    Sigma[k,k] <- eta_sq + sigma_sq;  // + jitter

  eta_sq ~ cauchy(0,5);
  rho_sq ~ cauchy(0,5);
  sigma_sq ~ cauchy(0,5);

  y ~ multi_normal(mu,Sigma);
}
```

A NOTE ON HIERARHICAL MODELS

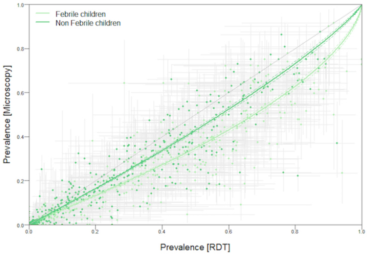
- ⋆ IMO, 90% of astrostatistical problems can be solved within hours (at most, a day) simply by writing the model out in hierarchical form & coding it up in STAN (or JAGS: Just Another Gibbs Sampler)
- ⋆ example from epidemiology: an EIV probit regression

$$
\begin{aligned}
n_i^{\mathrm{Mic}} &\sim \mathrm{Binom}(p_i^{\mathrm{Mic}}, n_i^{\mathrm{tot}}) \\
n_i^{\mathrm{RDT}} &\sim \mathrm{Binom}(p_i^{\mathrm{RDT}}, n_i^{\mathrm{tot}}) \\
\Phi^{-1}(p_i^{\mathrm{Mic}}) &= \alpha + \beta \times \Phi^{-1}(p_i^{\mathrm{RDT}}) \\
\{\Phi^{-1}(p_i^{\mathrm{RDT}})\}\ (i = 1, \ldots, n^{\mathrm{obs}}) &\sim F \\
F &\sim \mathrm{DP}(G_\Theta, m) \\
\alpha,\ \beta,\ \Theta,\ m &\sim \pi
\end{aligned}
$$

A NOTE ON HIERARHICAL MODELS

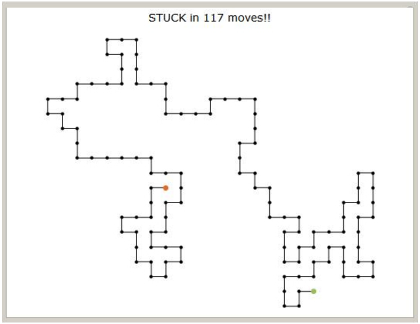
- ⋆ example from epidemiology: an EIV probit regression (JAGS code)

```
model.jags <- "model {
for (i in 1:N) {
nx[i] ~ dbinom(px[i],n[i]);
ny[i] ~ dbinom(py[i],n[i]);
}
for (i in 1:N) {
py[i] <- phi(alpha+beta*probit(px[i]));
}
alpha ~ dnorm(0,1);
beta ~ dnorm(0,1);
for (i in 1:N) {
px[i] <- phi(gx[i]);
}
for (i in 1:N) {
gx[i] ~ dnorm(mu[label[i]],precision[label[i]]);
}
for (i in 1:K) {
mu[i] ~ dnorm(0,1);
precision[i] ~ dgamma(1,1);
}
for (i in 1:N) {
label[i] ~ dcat(compprobs);
}
compprobs ~ ddirch(eta);
}"
```



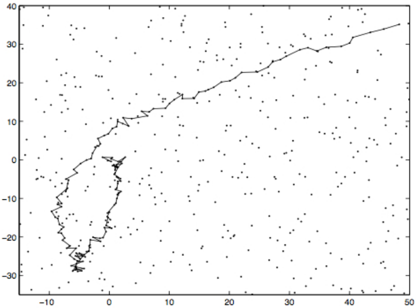
SEQUENTIAL MONTE CARLO

- Sequential Monte Carlo (SMC) techniques first appeared in the 1950s in the study of self-avoiding random walks (cf. Hammersley & Morton, 1954)
  - originally dismissed as ``poor man's Monte Carlo''



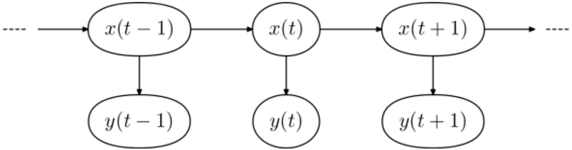
SEQUENTIAL MONTE CARLO

- ⋆ rediscovered in the 1990s as powerful solutions to the problems of:
  - missing data imputation (Kong et al., 1994)
  - automated target tracking (Gordon et al., 1994)



SEQUENTIAL MONTE CARLO

- ⋆ nowadays SMC methods are just as valuable as MCMC to the applied statistician
- ⋆ used almost exclusively for state space models: e.g. when we have noisy observations of a random process evolving in time



SEQUENTIAL MONTE CARLO

- ⋆ but SMC methods are also ubiquitous as tools for `ordinary' Bayesian inference:
  - e.g. posterior sampling via `data tempering' (Chopin, 2001)
  - e.g. marginal likelihood estimation: including two cosmological applications (Wraith et al, 2009; Kilbinger et al., 2010)

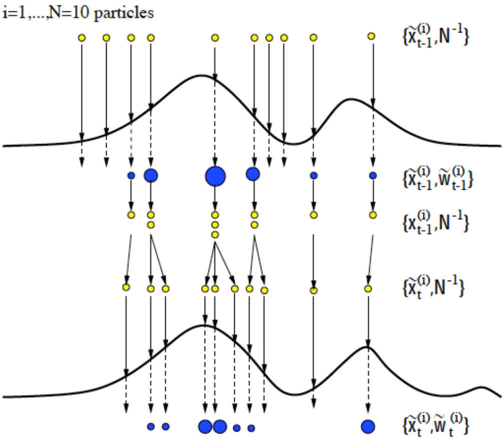
- ⋆ indeed any hierarchical model that can be written as a graphical model can be structured for SMC (Naesseth et al., 2014)

SEQUENTIAL MONTE CARLO

- ⋆ whereas MCMC iterates a single `particle' through a given parameter space such that its path converges towards the target density, SMC iterates a whole population of weighted particles
    - in many applications the iteration will be from the prior to the posterior via `data tempering' or the thermodynamic path
    - importantly: the SMC posterior can be updated `online': just add new data and proceed (without having to start all over)
    - the key steps of SMC are: re-weighting, resampling, & refreshment

SEQUENTIAL MONTE CARLO

- ⋆ example: re-weighting, resampling, & refreshment



SEQUENTIAL MONTE CARLO

⋆ keyword & references for further reading:

- particle filtering, particle Gibbs, sequential importance sampling, population Monte Carlo
- introduction/review: Doucet, de Freitas, & Gordon (2012)
- http://www.stats.ox.ac.uk/∼doucet/smc_resources.html
- introductory book: [Monte Carlo Strategies in Scientific Computing] Liu (2002)
- advanced: Del Moral et al., 2006

- extensions: [particle Gibbs w/ ancestor sampling] Lindsten et al., 2012
- extensions: [sequential quasi-Monte Carlo] Gerber & Chopin, 2014 [read at the RSS on Wednesday!]

PSEUDO-MARGINAL MCMC

- ⋆ a wide class of Bayesian procedures have recently been `discovered' based on the observation that MCMC can still target the correct posterior if the acceptance ratio,

$$z = \frac{L(y|\theta')\pi(\theta')f(\theta|\theta')}{L(y|\theta)\pi(\theta)f(\theta'|\theta)}$$

is replaced with unbiased estimates of $L(y|\cdot)$,

$$z = \frac{\hat{L}(y|\theta')\pi(\theta')f(\theta|\theta')}{\hat{L}(y|\theta)\pi(\theta)f(\theta'|\theta)}$$

(cf. Beaumont, 2003; Andrieu & Roberts, 2009; Doucet et al., 2012)

PSEUDO-MARGINAL MCMC

⋆ these `pseudo-marginal' MCMC algorithms can be useful when the likelihood is known conditional upon some unknown latent variables:

$$L(y|\theta) = \int L(y|z)\pi(z|\theta)dz$$

which suggests the importance sampling estimator,

$$\hat{L}(y|\theta) = \sum_{i=1}^{N} \frac{L(y|z_i)\pi(z_i|\theta)}{g(z_i)} \quad \text{for } z_i \sim g(\cdot)$$

- e.g. when the likelihood is known conditional upon some noisily-measured covariates
- some previous (`unwitting') applications of this form in astronomy: Hogg et al. (2010); Schneider et al. (2014)
- value of understanding context is access to statistical results on estimator choice for maximum efficiency

PSEUDO-MARGINAL MCMC

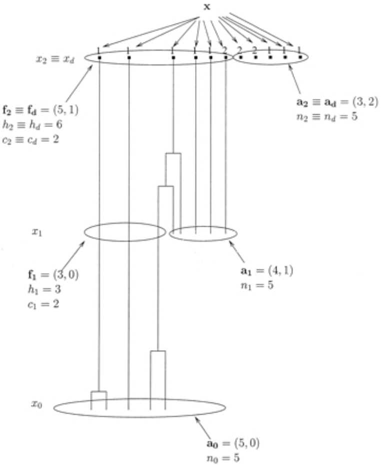
⋆ another case is where $\pi(z|\theta)$ is actually unknown (e.g., stochastic output from a computational simulation)

$$\hat{L}(y|\theta) = \sum_{i=1}^{N} L(y|z_i) \quad \text{for } z_i \sim \pi(z|\theta)$$

- this type of pseudo-marginal algorithm encompasses Approximate Bayesian Computation
- again, some previous (`unwitting') applications of this form in astronomy: `Bayesian simulation sampling' (Fardal et al., 2013)
- classic example is simulation of ancestor histories in population genetics

PSEUDO-MARGINAL MCMC

⋆ example: ancestor history simulations (Beaumont et al., 2003)



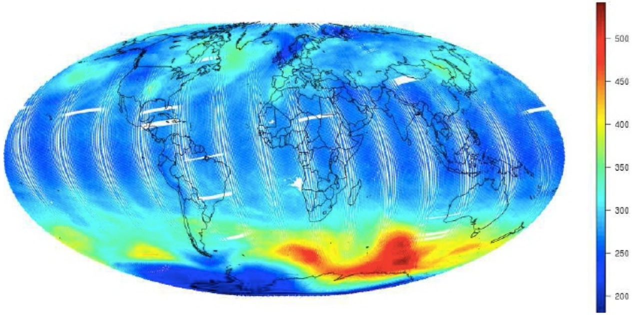
PSEUDO-MARGINAL MCMC

- ⋆ yet another case is where $\pi(y|\theta) = f(y|\theta)/Z(\theta)$ is known only up to an intractible normalizing constant
  - e.g. the Ising model, spatial point processes, massive Gaussian Markov Random Fields

- ⋆ unbiased estimators for $1/Z(\theta)$ can be constructed using a ``Russian roulette'' approach (cf. Lyne et al., 2014)
  - write the unknown term as an infinite series expansion, but only evaluate to a random finite number of terms
  - parts of this idea come from the Quantum Chromodynamics literature!
  - [applicable to some problems in GW? Canizares et al., 2013; Lentati et al., 2014]

PSEUDO-MARGINAL MCMC

- ⋆ example: ``Russian roulette'' GMRF (Lyne et al., 2014)



CONSENSUS MCMC

- in the `Big Data' regime it can easily happen that we have a simple model with iid observations yet the sheer volume of data means that likelihood evaluation becomes a limiting step for MCMC:

$$y = \{y_1, y_2, \ldots, y_n\} \sim f(\cdot|\theta^*) \text{ i.e., } L(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

- if we try to parallelise by splitting the data into m subsamples and distributing to m cores, we still have to wait for the slowest core to return its likelihood evalution before we can decide whether to accept or reject the proposed θ'

CONSENSUS MCMC

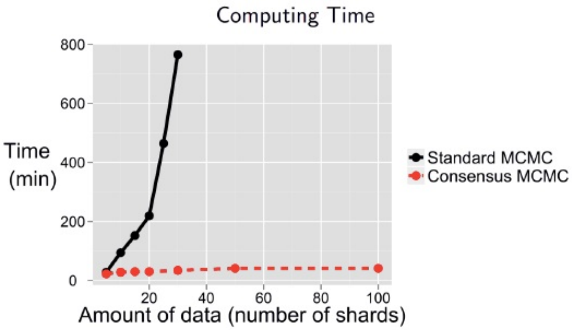
- a `naive' solution is to allow MCMC to run separate chains on the subsetted data give to each core, and then recombine the posteriors as a product of Normals fitted to each (Scott et al., 2013):

$$\hat{\pi}(\theta|y) \propto \prod_{j=1}^{m} \hat{N}_{[\pi(\{y\}_m|\theta)\pi(\theta)]}(\theta)$$

- if we try to parallelise by splitting the data into m subsamples and distributing to m cores, we still have to wait for the slowest core to return its likelihood evalution before we can decide whether to accept or reject the proposed θ'

CONSENSUS MCMC

- this can give huge speed ups, but assumes the each subset posterior is well-approximated by a Gaussian

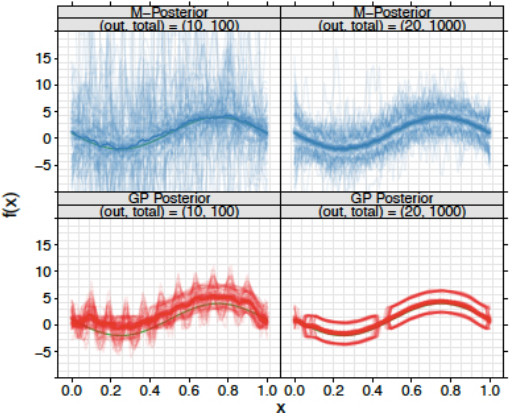


MEDIAN OF SUBSET POSTERIORS

- ⋆ a more sophisticated solution for combining subset posteriors is to take a median
- ⋆ but this requires some mathematical deliberation since a median of probability measures is not trivial to define
  - Minsker et al. (2014) solve this problem by introducing a kernel-based metric distance which has a median recovered via the Wieszfeld algorithm
  - available as the Mposteriors package in R
  - suggested to `power up' each subset posterior as $L(\{y\}_m|\theta)^m\pi(\theta)$

MEDIAN OF SUBSET POSTERIORS

⋆ GP example from Minsker et al. (2014):



AND MANY MORE

⋆ Other contemporary techniques left undiscussed:

- Bayesian non-parametrics with the Dirichlet processes
  [e.g. semi-parametric error models; Cameron & Pettitt, 2013: fine structure constant paper II]
  [e.g. online classification via the Mondrian process; Lakshminarayan et al., 2014]
- the unscented transform
  [e.g. for building fast approximate likelihood functions; e.g. Goldberger et al., 2008]
- Approximate Bayesian Computation
  [for intractible likelihoods: e.g., Cameron & Pettitt, 2013; Weyant et al., 2013]
- INLA (the Integrated Nested Lapalce Approximation) and the SPDE approach to random fields
  [for fast GP fitting: e.g., Rue et al., 2008; Lindgren et al., 2011]