Monte Carlo methods for model selection and other variable-dimension problems

Peter Green

University of Bristol
and UTS, Sydney

Bogotá, 25/26 de julio de 2013

Outline

Model selection in a Bayesian setting

One of the attractions of the Bayesian approach to modelling and inference is that uncertainty in the model (and in truth, there is always is uncertainty) is naturally dealt with in the same paradigm.

To a Bayesian, uncertainty in

- data
- parameters
- model

may be different things philosophically, but they can all be treated in the same way mathematically, through probability distributions.

Model selection in a Bayesian setting

One of the attractions of the Bayesian approach to modelling and inference is that uncertainty in the model (and in truth, there is always is uncertainty) is naturally dealt with in the same paradigm.

To a Bayesian, uncertainty in

- data
- parameters
- model

may be different things philosophically, but they can all be treated in the same way mathematically, through probability distributions.

Hierarchical model

A natural set-up consists of

- a prior $p(k)$ over models k in a countable set $\mathcal{K}$, and
- for each k
  - a prior distribution $p(\theta_k|k)$, and
  - a likelihood $p(Y|k,\theta_k)$ for the data Y.

For definiteness and simplicity, suppose that $p(\theta_k|k)$ is a density in n_k dimensions, and that there are no other parameters, so that where there are parameters common to all models these are subsumed into each $\theta_k \in \mathcal{R}^{n_k}$.

Additional parameters, perhaps in additional layers of a hierarchy, are easily dealt with. All probability distributions are proper.

Hierarchical model

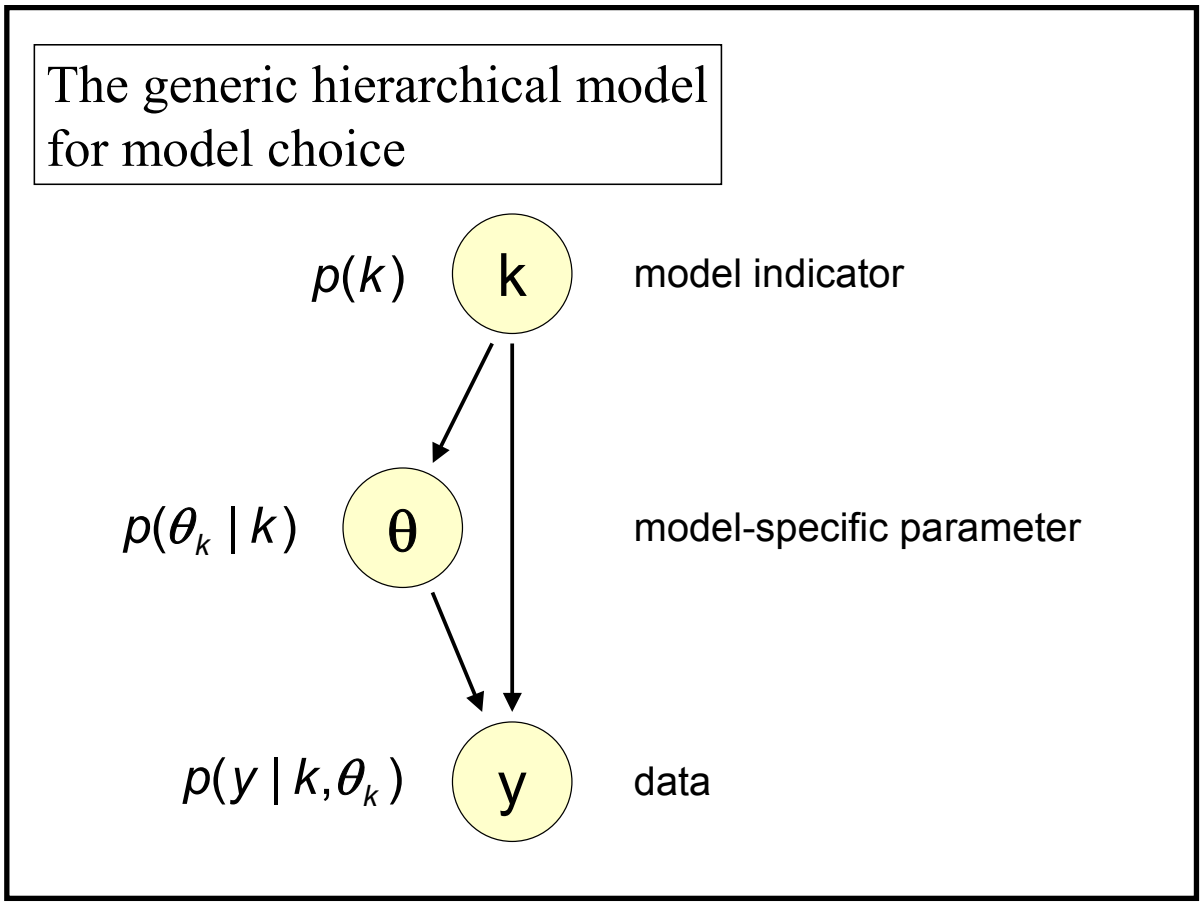
A natural set-up consists of

- a prior $p(k)$ over models k in a countable set $\mathcal{K}$, and
- for each k
  - a prior distribution $p(\theta_k|k)$, and
  - a likelihood $p(Y|k,\theta_k)$ for the data Y.

For definiteness and simplicity, suppose that $p(\theta_k|k)$ is a density in n_k dimensions, and that there are no other parameters, so that where there are parameters common to all models these are subsumed into each $\theta_k \in \mathcal{R}^{n_k}$.

Additional parameters, perhaps in additional layers of a hierarchy, are easily dealt with. All probability distributions are proper.

Hierarchical model as DAG



Note the generality of this basic formulation: it embraces both

- genuine model-choice situations, where the variable k indexes the collection of discrete models under consideration, but also
- settings where there is really a single model, but one with a variable dimension parameter, for example a functional representation such as a series whose number of terms is not fixed (in which case, k is unlikely to be of direct inferential interest).

A simple model choice problem in regression

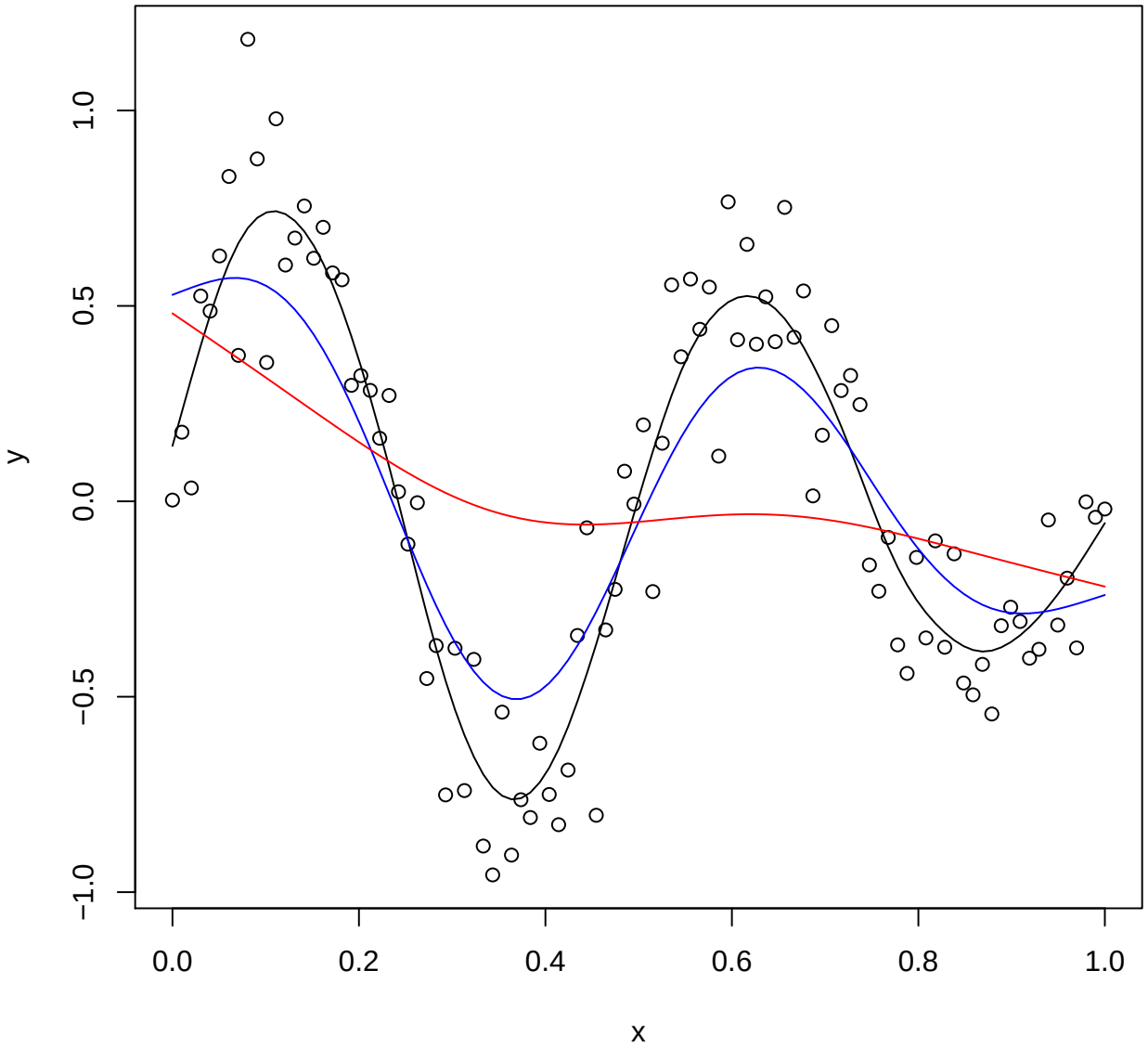
We have data $(x_i, y_i), i = 1, 2, \ldots, n$ and we entertain two alternative models, equally probable a priori:

1. $k = 1$: $Y_i | x_i \sim N(\alpha + \beta x_i, \sigma^2)$
2. $k = 2$: $Y_i | x_i \sim N(\gamma + \delta e^{\epsilon x_i}/(1 + e^{\epsilon x_i}), \sigma^2)$

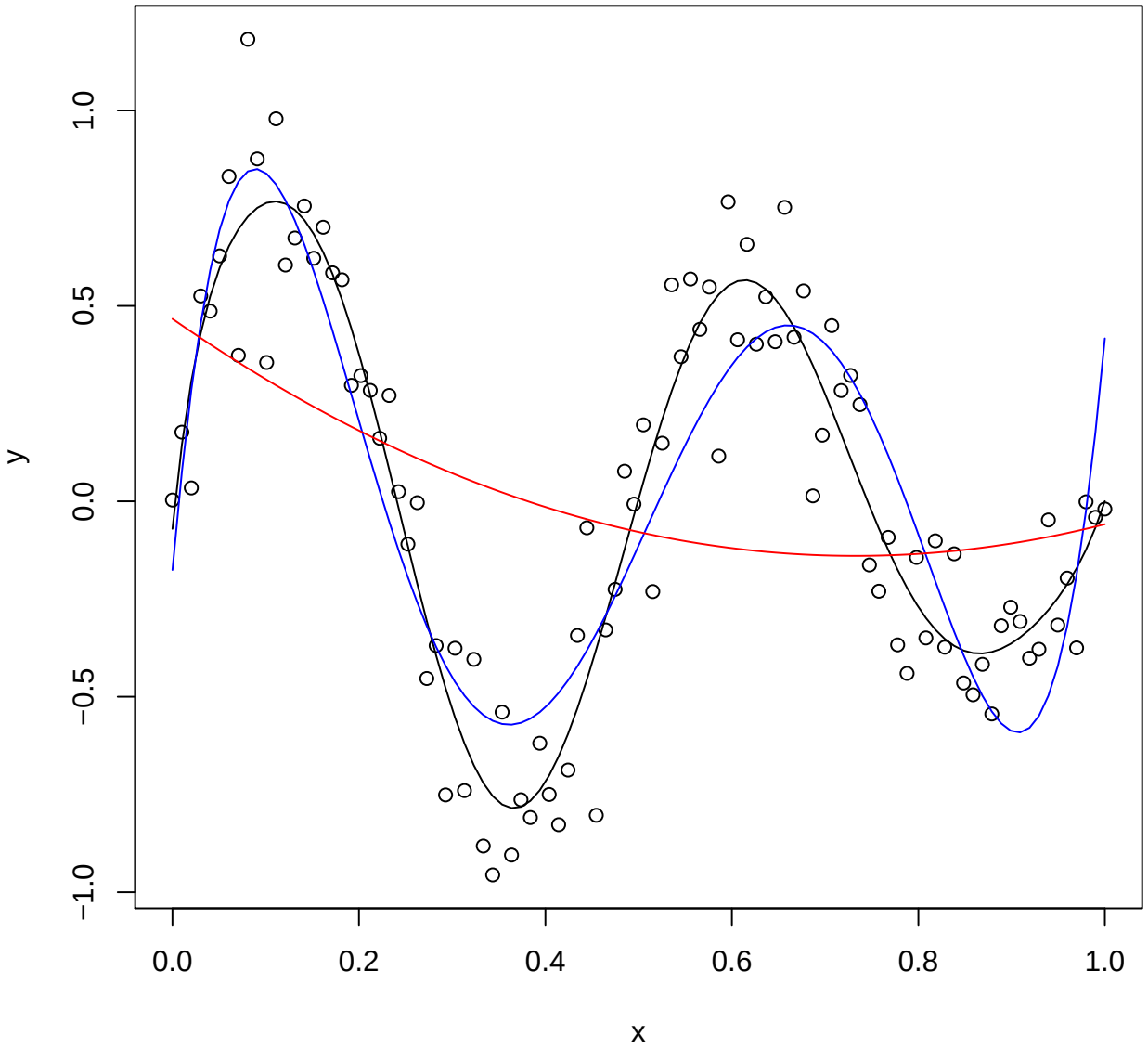
Note that one model has 3 parameters, the other 4.

How can we make Bayesian inference about (k, θ_k), where $\theta_1 = (\alpha, \beta, \sigma)$ and $\theta_2 = (\gamma, \delta, \epsilon, \sigma)$?

Splines



Polynomials



Trans-dimensional problems

What if the number of things you don't know is one of the things you don't know?

- variable selection
- finite mixture estimation
- change-point analysis
- nonparametric hazard estimation
- automatic curve fitting (degree, discontinuities)
- ARIMA model fitting
- graphical model determination
- ion channel data, quantitative trait locus analysis
- model-based CART
- image segmentation, object recognition
- restoring old movies

Trans-dimensional problems

What if the number of things you don't know is one of the things you don't know?

- variable selection
- finite mixture estimation
- change-point analysis
- nonparametric hazard estimation
- automatic curve fitting (degree, discontinuities)
- ARIMA model fitting
- graphical model determination
- ion channel data, quantitative trait locus analysis
- model-based CART
- image segmentation, object recognition
- restoring old movies

Recent real non-academic application areas

Applications in Geophysical sciences

Geophysical inversion
Geophysical source reconstruction
Geophysical electrical resistivity
Ground flow models
Air pollution, greenhouse gases, remote sensing
Air pollution, change point models
Climate and land models

Applications in Ecology and the Environment

Geophysical inversion
Phylogenetics and biodiversity
Animal abundance
Ecology of wildlife
Ecology of salmon
Ecology, conservation, environment
Remote Sensing land use

Recent real non-academic application areas

Agricultural and Medical applications

QTLs in agriculture
Protein-DNA interactions and medical implications
Genetics and disease aetiology
Spatial epidemiology of Chagas disease

Social and commercial applications

Health Insurance claim data
Financial modelling
Criminology

Applied image analysis and computer vision

Computer vision - object tracking
Image and classification using LiDaR
Imaging of geosynchronous orbits, managing space debris
Mixture modelling, image analysis
NMR
fMRI

Reporting inference about model and parameter

The joint posterior

$$p(k, \theta_k|Y) = \frac{p(k)p(\theta_k|k)p(Y|k, \theta_k)}{\sum_{k' \in \mathcal{K}} \int p(k')p(\theta'_{k'}|k')p(Y|k', \theta'_{k'})\mathrm{d}\theta'_{k'}}$$

can always be factorised as

$$p(k, \theta_k|Y) = p(k|Y)p(\theta_k|k, Y)$$

– very often the basis for reporting the inference, and in some of the methods mentioned below is also the basis for computation.

Reporting inference about model and parameter

The joint posterior

$$p(k, \theta_k|Y) = \frac{p(k)p(\theta_k|k)p(Y|k, \theta_k)}{\sum_{k' \in \mathcal{K}} \int p(k')p(\theta'_{k'}|k')p(Y|k', \theta'_{k'})\mathrm{d}\theta'_{k'}}$$

can always be factorised as

$$p(k, \theta_k|Y) = p(k|Y)p(\theta_k|k, Y)$$

– very often the basis for reporting the inference, and in some of the methods mentioned below is also the basis for computation.

Bayes factors, Evidence

Perhaps you didn't think you wanted to report $p(k|Y)$?
Of course, with these posterior probabilities, we can report Bayes Factors

$$B_{kl} = \frac{p(Y|k)}{p(Y|l)} = \frac{p(k|Y)}{p(l|Y)} \div \frac{p(k)}{p(l)}$$

for pairwise comparison of models.

For some, the Evidence (= marginal likelihood) $p(Y|k)$ has an intrinsic meaning and interpretation.

Bayesian model averaging, prediction

We can do model averaging

$$E(F|Y) = \sum_k \int F(k, \theta_k) p(k, \theta_k|Y) \mathrm{d}\theta_k$$

for any function F with the same interpretation in each model; the expectation can be estimated simply by averaging F along the entire run, essentially ignoring the model indicator k.

As an example, we can do prediction:

$$p(Y^+|Y) = \sum_k p(Y^+|k, Y) p(k|Y)$$

a posterior-weighted mixture of the within-model-k predictions

$$p(Y^+|k, Y) = \int p(Y^+|k, \theta_k) p(\theta_k|k, Y) \mathrm{d}\theta_k$$

Bayesian model averaging, prediction

We can do model averaging

$$E(F|Y) = \sum_k \int F(k, \theta_k) p(k, \theta_k|Y) \mathrm{d}\theta_k$$

for any function F with the same interpretation in each model; the expectation can be estimated simply by averaging F along the entire run, essentially ignoring the model indicator k.

As an example, we can do prediction:

$$p(Y^+|Y) = \sum_k p(Y^+|k, Y) p(k|Y)$$

a posterior-weighted mixture of the within-model-k predictions

$$p(Y^+|k, Y) = \int p(Y^+|k, \theta_k) p(\theta_k|k, Y) \mathrm{d}\theta_k$$

What do we need to compute?

So, however we wish to report our inference about model selection and parameter estimation, the computation we need to do is equivalent to computing the joint posterior $p(k, \theta_k | Y) = p(k|Y)p(\theta_k|k, Y)$.

The only absolutely general methods we know for computing the joint posterior use simulation, and MCMC is by far the most important of these in current practice.

What do we need to compute?

So, however we wish to report our inference about model selection and parameter estimation, the computation we need to do is equivalent to computing the joint posterior $p(k, \theta_k | Y) = p(k|Y)p(\theta_k|k, Y)$.

The only absolutely general methods we know for computing the joint posterior use simulation, and MCMC is by far the most important of these in current practice.

What is MCMC?

Monte Carlo = using the Law of Large Numbers (LoLN) to do calculation.

$$\frac{1}{N}\sum_{t=1}^{N} X^{(t)} \to \mu$$

if $X^{(1)}, X^{(2)}, \ldots$ are i.i.d. with $E(X^{(t)}) = \mu$.

Or more generally,

$$\frac{1}{N}\sum_{t=1}^{N} g(X^{(t)}) \to \mu$$

if $X^{(1)}, X^{(2)}, \ldots$ are i.i.d. with $E(g(X^{(t)})) = \mu$.

What is MCMC?

Monte Carlo = using the Law of Large Numbers (LoLN) to do calculation.

$$\frac{1}{N}\sum_{t=1}^{N} X^{(t)} \to \mu$$

if $X^{(1)}, X^{(2)}, \ldots$ are i.i.d. with $E(X^{(t)}) = \mu$.

Or more generally,

$$\frac{1}{N}\sum_{t=1}^{N} g(X^{(t)}) \to \mu$$

if $X^{(1)}, X^{(2)}, \ldots$ are i.i.d. with $E(g(X^{(t)})) = \mu$.

What is MCMC?

Markov chain Monte Carlo = using the LoLN for Markov chains to do calculation.

$$\frac{1}{N}\sum_{t=1}^{N} g(X^{(t)}) \to \mu$$

if $X^{(1)}, X^{(2)}, \ldots$ form an ergodic Markov chain whose invariant distribution π has $E(g(X^{(t)})) = \int g(x)\pi(dx) = \mu$.

Why is it interesting?

- Central to computational Bayesian inference (and computation in many other fields, e.g. statistical physics, spatial stochastic processes, etc...) is the need to calculate expectations and probabilities under complex high-dimensional distributions
- It is much easier to construct and simulate a Markov chain with a specified invariant ('target') distribution than an independent random sample from that distribution

Analysis of Normal random sample

Data $Y_1, Y_2, \ldots, Y_n$ are a random sample from $N(\mu, \sigma^2)$.

Independent priors on μ and σ:

$$\begin{aligned} \mu &\sim N(\xi, \kappa^{-1}) \\ \sigma^{-2} &\sim \Gamma(\alpha, \beta) \end{aligned}$$

(these are only conditionally conjugate).

Joint posterior:

$$\begin{aligned} p(\mu, \sigma^2 | Y) &\propto (\sigma^2)^{-\alpha-1-n/2} \\ &\times \exp\left\{ -\frac{\beta}{\sigma^2} - \frac{\kappa(\mu-\xi)^2}{2} - \frac{\sum(Y_i-\mu)^2}{2\sigma^2} \right\} \end{aligned}$$

which is not of standard form.

Analysis of Normal random sample

'Full conditionals' are easily found:

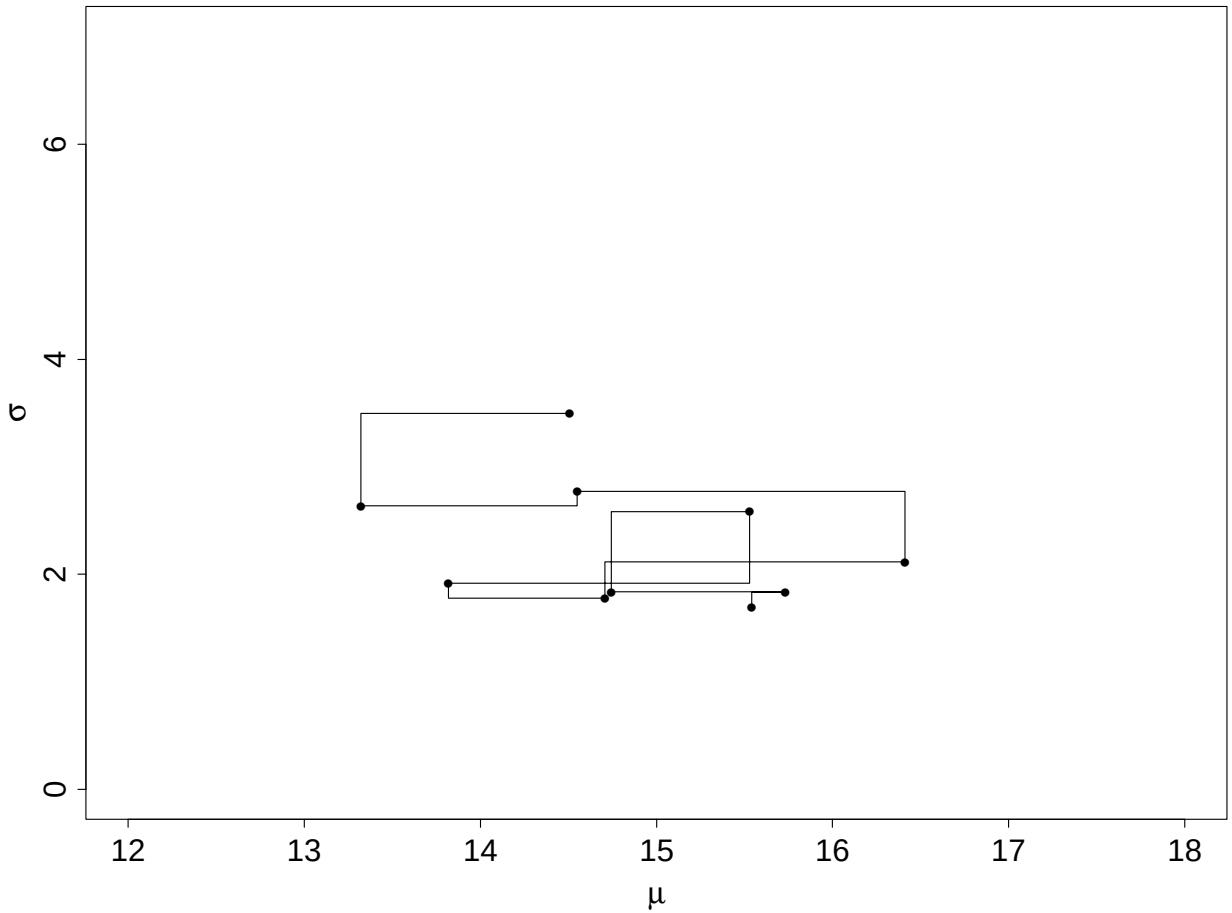
$$\begin{aligned} \mu|\sigma, Y &\sim N\left(\frac{\sigma^{-2}\sum Y_i + \kappa\xi}{\sigma^{-2}n + \kappa}, \frac{1}{\sigma^{-2}n + \kappa}\right) \\ \sigma^{-2}|\mu, Y &\sim \Gamma(\alpha + n/2, \beta + \sum(Y_i - \mu)^2/2) \end{aligned}$$

and we can implement the so-called *Gibbs sampler* by alternately drawing μ and σ^{-2} from these distributions.

This defines a Markov chain with states $X^{(t)} = (\mu^{(t)}, \sigma^{(t)})$, whose transition probabilities depend on Y.

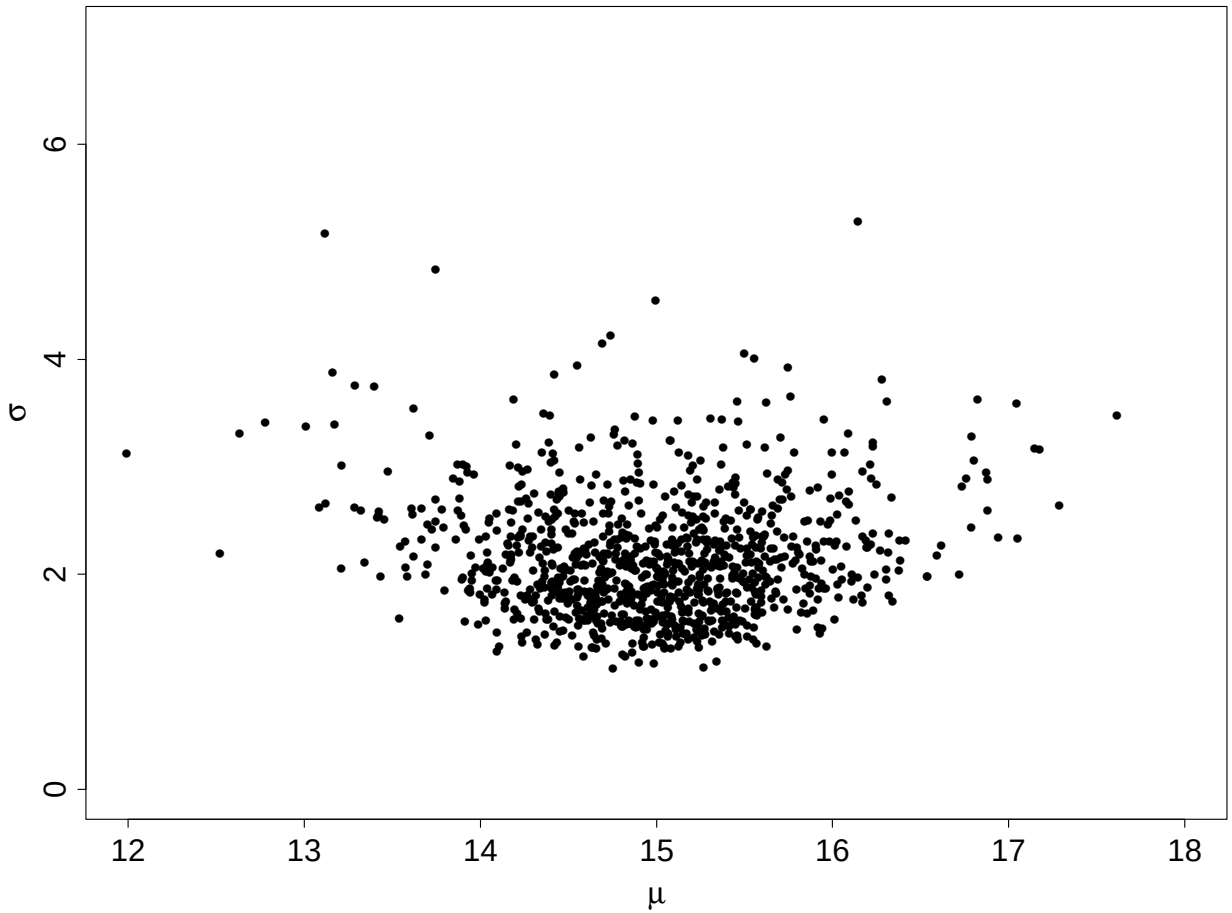
Analysis of Normal random sample

Gibbs sample of size $N = 10$. Posterior sample of (μ, σ) from data with $n = 10$, $\overline{Y} = 15$, $s_Y^2 = 4$. Uninformative prior.



Analysis of Normal random sample

Gibbs sample of size $N = 1000$. Posterior sample of (μ, σ) from data with $n = 10$, $\overline{Y} = 15$, $s_Y^2 = 4$. Uninformative prior.



Analysis of Normal random sample

We can show that the equilibrium distribution of this chain $X^{(t)} = (\mu^{(t)}, \sigma^{(t)}), t = 1, 2, \ldots$ is exactly $p(\mu, \sigma | Y)$, the posterior distribution of the parameters given the data Y.

And the chain is ergodic, so this is also the limiting distribution of the chain

$$(\mu^{(t)}, \sigma^{(t)}) \xrightarrow{d} p(\mu, \sigma | Y)$$

(note that the limit is a distribution, not a point).

The law of large numbers applies, and for any function g,

$$\frac{1}{N} \sum_{t=1}^{N} g(\mu^{(t)}, \sigma^{(t)}) \to \int g(\mu, \sigma) p(\mu, \sigma | Y) d\mu d\sigma$$

Analysis of Normal random sample

We can show that the equilibrium distribution of this chain $X^{(t)} = (\mu^{(t)}, \sigma^{(t)}), t = 1, 2, \ldots$ is exactly $p(\mu, \sigma | Y)$, the posterior distribution of the parameters given the data Y.

And the chain is ergodic, so this is also the limiting distribution of the chain

$$(\mu^{(t)}, \sigma^{(t)}) \xrightarrow{d} p(\mu, \sigma | Y)$$

(note that the limit is a distribution, not a point).

The law of large numbers applies, and for any function g,

$$\frac{1}{N} \sum_{t=1}^{N} g(\mu^{(t)}, \sigma^{(t)}) \to \int g(\mu, \sigma) p(\mu, \sigma | Y) d\mu d\sigma$$

Analysis of Normal random sample

We can show that the equilibrium distribution of this chain $X^{(t)} = (\mu^{(t)}, \sigma^{(t)}), t = 1, 2, \ldots$ is exactly $p(\mu, \sigma | Y)$, the posterior distribution of the parameters given the data Y.

And the chain is ergodic, so this is also the limiting distribution of the chain

$$(\mu^{(t)}, \sigma^{(t)}) \xrightarrow{d} p(\mu, \sigma | Y)$$

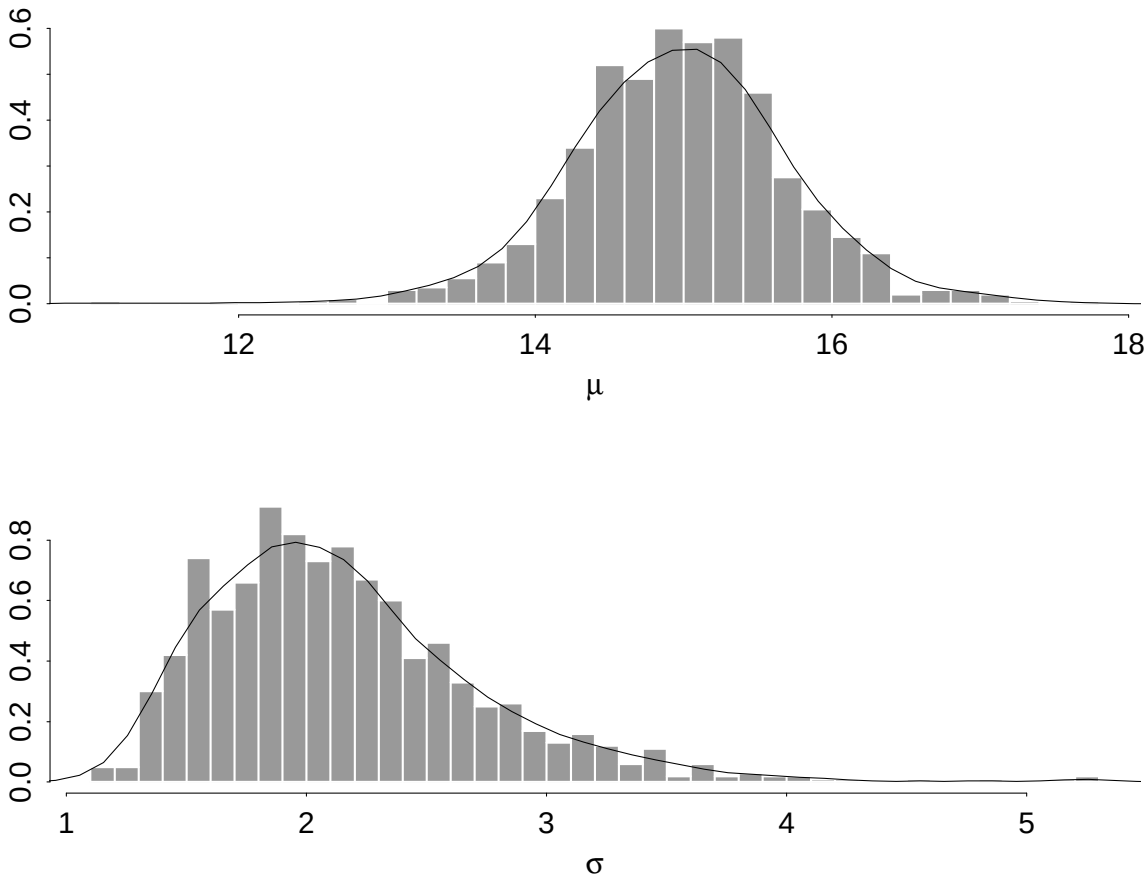
(note that the limit is a distribution, not a point).

The law of large numbers applies, and for any function g,

$$\frac{1}{N} \sum_{t=1}^{N} g(\mu^{(t)}, \sigma^{(t)}) \to \int g(\mu, \sigma) p(\mu, \sigma | Y) d\mu d\sigma$$

Analysis of Normal random sample

Marginal posterior distributions of μ and σ from data with $n = 10$, $\overline{Y} = 15$, $s_Y^2 = 4$. Uninformative prior.



Central limit theorem

Under weak conditions, ergodic Markov chains also satisfy a central limit theorem

$$\frac{\sqrt{N}}{\sigma}\left(\frac{1}{N}\sum_{t=1}^{N} g(X^{(t)}) - \mu\right) \xrightarrow{d} N(0,\tau)$$

as $N \to \infty$, where $\sigma^2 = \mathrm{var}(g(X))$ under the invariant distribution, and τ is the integrated autocorrelation time, $\tau = \sum_{t=-\infty}^{\infty} \gamma_t$ where γ_t is the lag-t equilibrium autocorrelation of $\{g(X^{(t)})\}$.

Note that unlike conventional numerical quadrature methods, the error does not scale adversely as the dimension d increases – we always have the "square root law" (although to be fair, typically τ may increase with d).

Central limit theorem

Under weak conditions, ergodic Markov chains also satisfy a central limit theorem

$$\frac{\sqrt{N}}{\sigma}\left(\frac{1}{N}\sum_{t=1}^{N} g(X^{(t)}) - \mu\right) \xrightarrow{d} N(0,\tau)$$

as $N \to \infty$, where $\sigma^2 = \mathrm{var}(g(X))$ under the invariant distribution, and τ is the integrated autocorrelation time, $\tau = \sum_{t=-\infty}^{\infty} \gamma_t$ where γ_t is the lag-t equilibrium autocorrelation of $\{g(X^{(t)})\}$.

Note that unlike conventional numerical quadrature methods, the error does not scale adversely as the dimension d increases – we always have the "square root law" (although to be fair, typically τ may increase with d).

Convergence and efficiency

To evaluate performance of an MCMC method, we need to understand (and discriminate between) two different things:

1. Convergence: how quickly does $X^{(t)} \xrightarrow{d} \pi$? (uniform /geometric ergodicity, bounds on TV norm, etc.)
2. Efficiency: determined by autocorrelation times τ for functionals $g(X)$ of interest (which can be estimated empirically to find MCSE's).

One MCMC method may dominate another on one criterion and not the other.

How do you construct a Markov chain for a given target distribution?

We have a target distribution π on a space Ω, perhaps $\mathcal{R}^d$. We want to construct and simulate a Markov chain $X^{(1)}, X^{(2)}, \ldots$ that is ergodic and has invariant distribution π. Suppose the transition matrix is P.

We need two things:

1. π is invariant for P, i.e. $\sum_x \pi(x)P(x,y) = \pi(y)$ (in discrete distribution notation)
2. P is irreducible

How do you construct a Markov chain for a given target distribution?

We have a target distribution π on a space Ω, perhaps $\mathcal{R}^d$. We want to construct and simulate a Markov chain $X^{(1)}, X^{(2)}, \ldots$ that is ergodic and has invariant distribution π. Suppose the transition kernel is P, $P(x, A) = P(X^{(t+1)} \in A | X^{(t)} = x)$.

We need two things:

1. π is invariant for P, i.e. $\int \pi(dx) P(x, A) = \pi(A)$ (in general measure-theoretic notation)
2. P is irreducible

Detailed balance and reversibility

In practice, rather than check invariance, we usually look for detailed balance, which is stronger but easier to check:
$\pi(x)P(x,y) = \pi(y)P(y,x)$ for all x, y, or
$\pi(dx)P(x,dy) = \pi(dy)P(y,dx)$ for all x, y.

If this holds, the chain is reversible. Most, but not quite all, Markov chains of use in MCMC are reversible, or are built from reversible pieces.

Myths about MCMC 1

We need aperiodicity – we don't, since we do not need convergence in distribution of $g(X^{(t)})$, only of the ergodic averages $N^{-1}\sum_{t=1}^{N} g(X^{(t)})$.

Myths about MCMC 1

We need aperiodicity – we don't, since we do not need convergence in distribution of $g(X^{(t)})$, only of the ergodic averages $N^{-1}\sum_{t=1}^{N} g(X^{(t)})$.

Metropolis-Hastings

We will explain this in a discrete distribution notation: $\pi(x)$ is target distribution; $P(x, y)$ is the transition 'matrix'.

When current state is $X^{(t)} = x$:

1. Propose a new state x', drawn from distribution $Q(x, x')$
2. With probability $\alpha(x, x')$, given by

$$\alpha(x, x') = \min\left\{1, \frac{\pi(x')Q(x', x)}{\pi(x)Q(x, x')}\right\}$$

set new state $X^{(t+1)} = x'$, otherwise stay where you are: $X^{(t+1)} = x$

For $x' \neq x$, $P(x, x') = Q(x, x')\alpha(x, x')$;
$P(x, x) = Q(x, x) + \sum_{x' \neq x} Q(x, x')(1 - \alpha(x, x'))$.

Metropolis-Hastings

We will explain this in a discrete distribution notation: $\pi(x)$ is target distribution; $P(x, y)$ is the transition 'matrix'.

When current state is $X^{(t)} = x$:

1. Propose a new state x', drawn from distribution $Q(x, x')$
2. With probability $\alpha(x, x')$, given by

$$\alpha(x, x') = \min\left\{1, \frac{\pi(x')Q(x', x)}{\pi(x)Q(x, x')}\right\}$$

set new state $X^{(t+1)} = x'$, otherwise stay where you are: $X^{(t+1)} = x$

For $x' \neq x$, $P(x, x') = Q(x, x')\alpha(x, x')$;
$P(x, x) = Q(x, x) + \sum_{x' \neq x} Q(x, x')(1 - \alpha(x, x'))$.

Metropolis-Hastings: proof of detailed balance

We have to show that $\pi(x)P(x,x') = \pi(x')P(x',x)$; for $x = x'$ there is nothing to prove.

For $x \neq x'$,

$$\begin{aligned}\pi(x)P(x,x') &= \pi(x)Q(x,x')\alpha(x,x') \\ &= \pi(x)Q(x,x')\min\left\{1, \frac{\pi(x')Q(x',x)}{\pi(x)Q(x,x')}\right\} \\ &= \min\{\pi(x')Q(x',x), \pi(x)Q(x,x')\}\end{aligned}$$

but this is symmetric in x and x' so we are done.

Metropolis-Hastings: proof of detailed balance

We have to show that $\pi(x)P(x,x') = \pi(x')P(x',x)$; for $x = x'$ there is nothing to prove.

For $x \neq x'$,

$$\begin{aligned}\pi(x)P(x,x') &= \pi(x)Q(x,x')\alpha(x,x') \\ &= \pi(x)Q(x,x')\min\left\{1, \frac{\pi(x')Q(x',x)}{\pi(x)Q(x,x')}\right\} \\ &= \min\{\pi(x')Q(x',x), \pi(x)Q(x,x')\}\end{aligned}$$

but this is symmetric in x and x' so we are done.

Multiple moves, schedules, irreducibility

In practice, it is useful to have not one type of transition or 'move' but several, so we have a 'dictionary' of transition matrices $\{P_m(x, x')\}$ all in detailed balance with respect to $\pi(x)$, e.g. MH matrices for different proposals Q_m. These can be combined in many ways to yield an overall transition matrix $P(x, x')$, e.g.

- Deterministically, e.g. $P = P_1 P_2 \dots P_M$
- Randomly, e.g. $P = M^{-1} \sum_m P_m$
- Palindromically: $P = P_1 P_2 \dots P_M P_M P_{M-1} \dots P_1$

In all these cases, and others, π remains an invariant distribution, but detailed balance may be lost.

We seek a combination so that P is irreducible (even if individual P_m are not), and with good performance (see later).

Multiple moves, schedules, irreducibility

In practice, it is useful to have not one type of transition or 'move' but several, so we have a 'dictionary' of transition matrices $\{P_m(x, x')\}$ all in detailed balance with respect to $\pi(x)$, e.g. MH matrices for different proposals Q_m. These can be combined in many ways to yield an overall transition matrix $P(x, x')$, e.g.

- Deterministically, e.g. $P = P_1 P_2 \dots P_M$
- Randomly, e.g. $P = M^{-1} \sum_m P_m$
- Palindromically: $P = P_1 P_2 \dots P_M P_M P_{M-1} \dots P_1$

In all these cases, and others, π remains an invariant distribution, but detailed balance may be lost.

We seek a combination so that P is irreducible (even if individual P_m are not), and with good performance (see later).

Important special cases of Metropolis-Hastings

- Metropolis: $Q(x, x') = Q(x', x)$
- Random walk Metropolis: $Q(x, x') = q(x' - x)$
- Independence MH: $Q(x, x') = q(x') \, \forall \, x$
- Gibbs: $Q_m(x, x') = \pi(x'_m | x_{-m}) \, I[x'_i = x_i, i \neq m]$ (the conditional distribution of x_m holding other components fixed)

Note that the last of these is the only one referring at all to the target – for the others, the target is only used to calculate the acceptance probability, not in random number generation.

Important special cases of Metropolis-Hastings

- Metropolis: $Q(x, x') = Q(x', x)$
- Random walk Metropolis: $Q(x, x') = q(x' - x)$
- Independence MH: $Q(x, x') = q(x') \, \forall \, x$
- Gibbs: $Q_m(x, x') = \pi(x'_m | x_{-m}) \, I[x'_i = x_i, i \neq m]$ (the conditional distribution of x_m holding other components fixed)

Note that the last of these is the only one referring at all to the target – for the others, the target is only used to calculate the acceptance probability, not in random number generation.

Myths about MCMC 2

Gibbs is the method of choice, we only use anything else if the full conditionals $\pi(x'_m|x_{-m})$ are not easy to sample from – there are no general reasons why Gibbs is best, and often it is very bad; the only clear advantage is that you have fewer choices to make.

Myths about MCMC 2

Gibbs is the method of choice, we only use anything else if the full conditionals $\pi(x'_m|x_{-m})$ are not easy to sample from – there are no general reasons why Gibbs is best, and often it is very bad; the only clear advantage is that you have fewer choices to make.

Why has MCMC been such a tremendous boon to Bayesian statistics?

- Algorithm generated by assumed model
- Modularity
- Product structure facilitates calculation of $\pi(x')/\pi(x)$ ratios, used in

$$\alpha(x, x') = \min\left\{1, \frac{\pi(x')Q(x', x)}{\pi(x)Q(x, x')}\right\}$$

- We only need know π up to a normalising constant

Myths about MCMC 3

We should sub-sample (or thin) the MCMC sample $X^{(1)}, X^{(2)}, \ldots$ in order to reduce/eliminate autocorrelation among $\{g(X^{(t)})\}$ – we shouldn't, subsampling the sequence increases the asymptotic variance of the ergodic average

Myths about MCMC 3

We should sub-sample (or thin) the MCMC sample $X^{(1)}, X^{(2)}, \ldots$ in order to reduce/eliminate autocorrelation among $\{g(X^{(t)})\}$ – we shouldn't, subsampling the sequence increases the asymptotic variance of the ergodic average

Myths about MCMC 4

Careful use of MCMC diagnostics gives you reassurance that your chain has converged and that you are successfully sampling the target distribution – nothing in the past history of the process can tell you that it will not jump to a new part of the sample space in the future

Myths about MCMC 4

Careful use of MCMC diagnostics gives you reassurance that your chain has converged and that you are successfully sampling the target distribution – nothing in the past history of the process can tell you that it will not jump to a new part of the sample space in the future

Some strengths of sample-based Bayesian computation using algorithms generated by the model

- Freedom in modelling
  - in principle, no limits
  - well-adapted for models defined on sparse graphs
- Freedom in inference
  - in principle, no limits
  - can address questions only posed after simulation completed (e.g. ranking and selection)
  - opportunities for simultaneous inference
- Allows/encourages sensitivity analysis
- Model comparison/criticism/choice
- Coherently integrates uncertainty
- Only available method for complex problems

Some weaknesses and dangers

- Order $\sqrt{N}$ precision
- Possibility of slow convergence, especially when not diagnosable (meta-stability)
- Risk that fitting technology runs ahead of statistical science
- Risk of undisciplined, selective presentation
- Difficulty of validating code

Adaptive MCMC

Except in small/easy problems, the choice of proposal distribution $Q(x, x')$ is crucial in designing Metropolis-Hastings methods that perform well enough.

An intuitive approach would be to adjust $Q(x, x')$ at time t based on the history $X^{(1)}, X^{(2)}, \ldots, X^{(t)}$ – but if we did that, $(X^{(t)})$ would not be a Markov chain! Then the straightforward limit theory no longer applies.

We can legitimately:

- Perform pilot runs – experiment until performance acceptable, then run again to collect samples
- Only start collecting samples after adjustment ceases
- Prove new ergodic theorems!

Adaptive MCMC

Except in small/easy problems, the choice of proposal distribution $Q(x, x')$ is crucial in designing Metropolis-Hastings methods that perform well enough.

An intuitive approach would be to adjust $Q(x, x')$ at time t based on the history $X^{(1)}, X^{(2)}, \ldots, X^{(t)}$ – but if we did that, $(X^{(t)})$ would not be a Markov chain! Then the straightforward limit theory no longer applies.

We can legitimately:

- Perform pilot runs – experiment until performance acceptable, then run again to collect samples
- Only start collecting samples after adjustment ceases
- Prove new ergodic theorems!

Adaptive MCMC

Except in small/easy problems, the choice of proposal distribution $Q(x, x')$ is crucial in designing Metropolis-Hastings methods that perform well enough.

An intuitive approach would be to adjust $Q(x, x')$ at time t based on the history $X^{(1)}, X^{(2)}, \ldots, X^{(t)}$ – but if we did that, $(X^{(t)})$ would not be a Markov chain! Then the straightforward limit theory no longer applies.

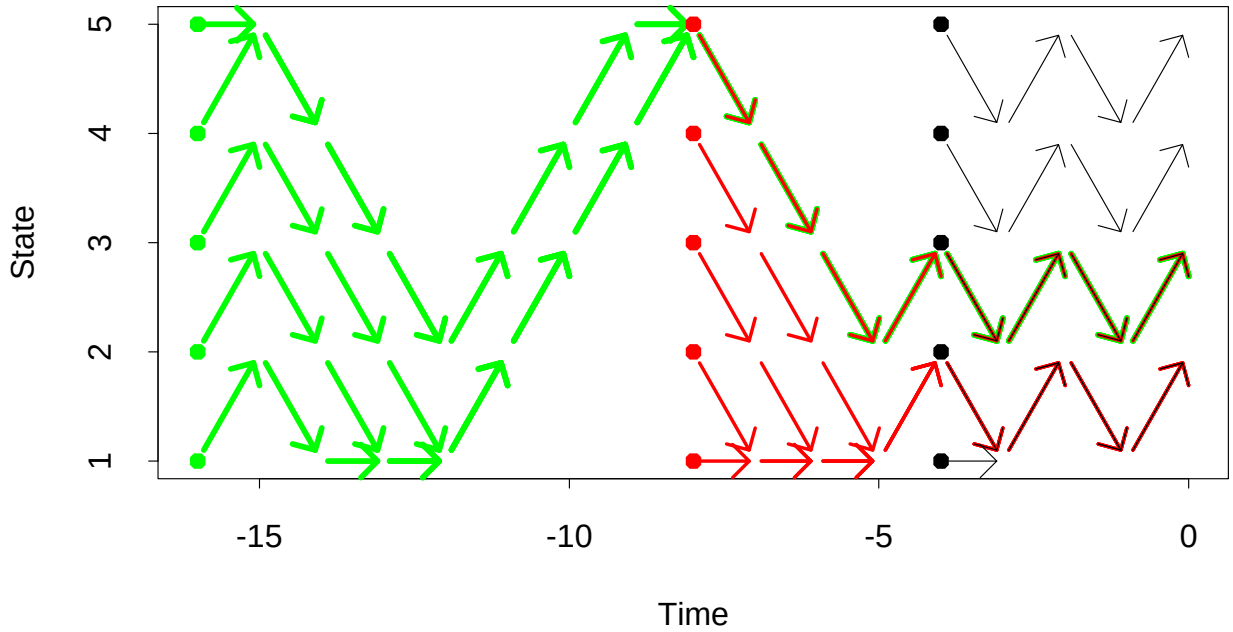
We can legitimately:

- Perform pilot runs – experiment until performance acceptable, then run again to collect samples
- Only start collecting samples after adjustment ceases
- Prove new ergodic theorems!

Coupling from the past (CFTP)

Here is a beautiful idea due to Propp and Wilson (1995): a way of organising a Markov chain simulation so that after a finite but random amount of work, it exactly delivers a sample from the target distribution!

For example, look at this partial simulation of a symmetric random walk with reflecting barriers:



ABC – approximate Bayesian computation

In some complex statistical models (e.g. in population genetics, spatial statistics, . . .), not only is the posterior/target distribution $\pi(x)$ only known up to an unknown normalising constant, it cannot be evaluated tractably at all.

ABC methods provide a way to conduct MCMC-like sampling providing you can simulate data from the assumed model for any parameter value.

ABC – approximate Bayesian computation

In some complex statistical models (e.g. in population genetics, spatial statistics, . . .), not only is the posterior/target distribution $\pi(x)$ only known up to an unknown normalising constant, it cannot be evaluated tractably at all.

ABC methods provide a way to conduct MCMC-like sampling providing you can simulate data from the assumed model for any parameter value.

Sequential Monte Carlo

a.k.a. particle filtering.

Dynamic models (time series, stochastic volatility models, state-space models, hidden Markov chains, . . .) form a huge class of statistical models for which computational inference is needed – but the interaction between the real time scale of the process and the artificial time scale of the MCMC simulation poses problems.

These are being effectively addressed by particle filter methods – not MCMC but iterative importance-sampling.

Tricks have been devised to deal with 'static' parameters, and there are now reliable SMC approaches even for non-dynamic Bayesian problems.

Sequential Monte Carlo

a.k.a. particle filtering.

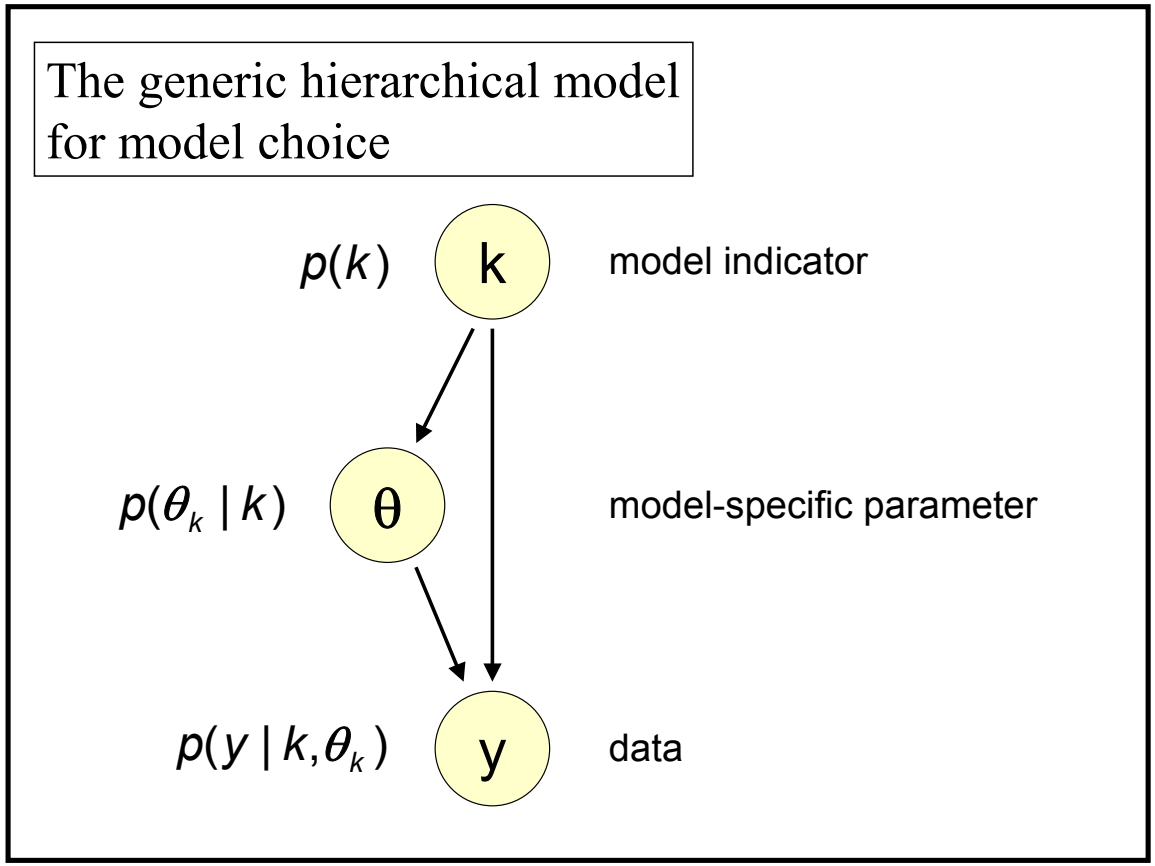
Dynamic models (time series, stochastic volatility models, state-space models, hidden Markov chains, . . .) form a huge class of statistical models for which computational inference is needed – but the interaction between the real time scale of the process and the artificial time scale of the MCMC simulation poses problems.

These are being effectively addressed by particle filter methods – not MCMC but iterative importance-sampling.

Tricks have been devised to deal with 'static' parameters, and there are now reliable SMC approaches even for non-dynamic Bayesian problems.

Trans-dimensional problems

Most 'trans-dimensional' problems can be set up as hierarchical models



Trans-dimensional statistical inference: a simple example

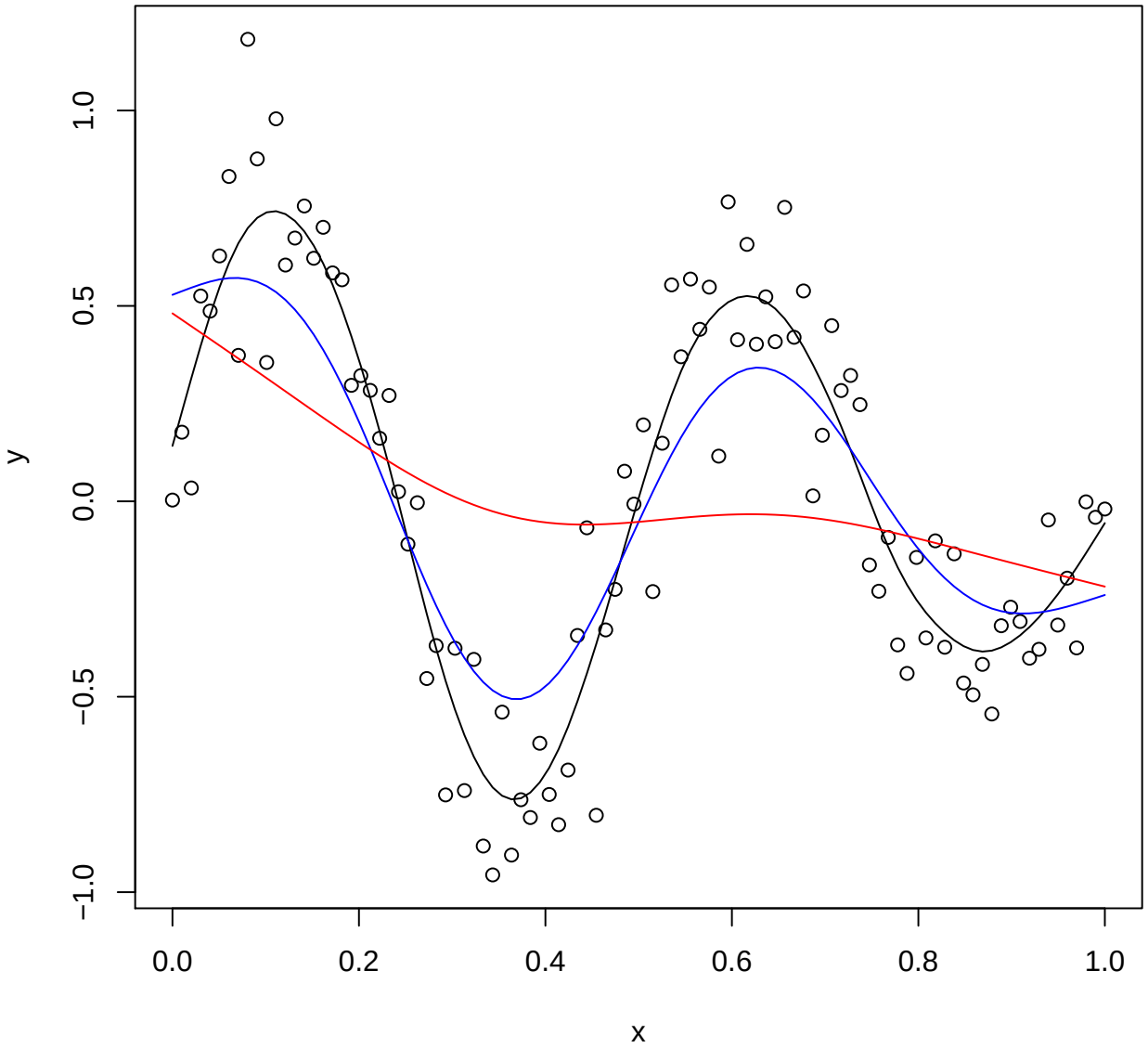
We have data $(x_i, y_i), i = 1, 2, \ldots, n$ and we entertain two alternative models, equally probable a priori:

1. $k = 1$: $Y_i | x_i \sim N(\alpha + \beta x_i, \sigma^2)$
2. $k = 2$: $Y_i | x_i \sim N(\gamma + \delta e^{\epsilon x_i}/(1 + e^{\epsilon x_i}), \sigma^2)$

Note that one model has 3 parameters, the other 4.

How can we make Bayesian inference about (k, θ_k), where $\theta_1 = (\alpha, \beta, \sigma)$ and $\theta_2 = (\gamma, \delta, \epsilon, \sigma)$, using MCMC?

Splines



Across- and within-model simulation

How to compute $p(k, \theta_k|Y)$?

Two main approaches using MCMC:

- across: one MCMC simulation with states of the form $(k, \theta_k) \approx p(k, \theta_k|Y)$
- within: separate simulations of $\theta_k \approx p(\theta_k|k, Y)$ for each k.

and beyond straight MCMC:

- particle filter: SMC in place of MCMC
- ABC: 'likelihood-free' methods where $Y|k, \theta_k$ can be simulated but $p(Y|k, \theta_k)$ cannot be evaluated.
- nested sampling
- variational methods

Across- and within-model simulation

How to compute $p(k, \theta_k | Y)$?

Two main approaches using MCMC:

- across: one MCMC simulation with states of the form $(k, \theta_k) \approx p(k, \theta_k | Y)$
- within: separate simulations of $\theta_k \approx p(\theta_k | k, Y)$ for each k.

and beyond straight MCMC:

- particle filter: SMC in place of MCMC
- ABC: 'likelihood-free' methods where $Y | k, \theta_k$ can be simulated but $p(Y | k, \theta_k)$ cannot be evaluated.
- nested sampling
- variational methods

Across-model simulation: Reversible jump MCMC

The across-model MCMC simulator follows the ideal Bayesian, and treats (k, θ_k) as a single unknown. The state space for an across-model simulation is $\{(k, \theta_k)\} = \bigcup_{k \in \mathcal{K}} (\{k\} \times \mathcal{R}^{n_k})$.

Mathematically, this is not a particularly awkward object. But at least a little non-standard, when n_k varies with k.

We use Metropolis-Hastings to build a suitable reversible chain.

On the face of it, this requires measure-theoretic notation, which may be unwelcome! The point of the 'reversible jump' framework is to render the measure theory invisible, by means of a construction using only ordinary densities. Even the fact that we are jumping dimensions becomes essentially invisible.

Across-model simulation: Reversible jump MCMC

The across-model MCMC simulator follows the ideal Bayesian, and treats (k, θ_k) as a single unknown. The state space for an across-model simulation is $\{(k, \theta_k)\} = \bigcup_{k \in \mathcal{K}} (\{k\} \times \mathcal{R}^{n_k})$.

Mathematically, this is not a particularly awkward object. But at least a little non-standard, when n_k varies with k.

We use Metropolis-Hastings to build a suitable reversible chain.

On the face of it, this requires measure-theoretic notation, which may be unwelcome! The point of the 'reversible jump' framework is to render the measure theory invisible, by means of a construction using only ordinary densities. Even the fact that we are jumping dimensions becomes essentially invisible.

Metropolis-Hastings

We will explain this first in a discrete distribution notation: $\pi(x)$ is target distribution; $P(x, y)$ is the transition 'matrix'.

When current state is $X^{(t)} = x$:

1. Propose a new state x', drawn from distribution $Q(x, x')$
2. With probability $\alpha(x, x')$, given by

$$\alpha(x, x') = \min\left\{1, \frac{\pi(x')Q(x', x)}{\pi(x)Q(x, x')}\right\}$$

set new state $X^{(t+1)} = x'$, otherwise stay where you are: $X^{(t+1)} = x$

For $x' \neq x$, $P(x, x') = Q(x, x')\alpha(x, x')$;
$P(x, x) = Q(x, x) + \sum_{x' \neq x} Q(x, x')(1 - \alpha(x, x'))$.

Metropolis-Hastings on general state spaces

π is target distribution; $P(x, B)$ is the transition kernel.

When current state is $X^{(t)} = x$:

1. Choose move type m with probability $c_m(x)$
2. Propose a new state x', drawn from distribution $Q_m(x, dx')$
3. With probability $\alpha_m(x, x')$, given by

$$\alpha_m(x, x') = \min\left\{1, \frac{\pi(dx')c_m(x')Q_m(x', dx)}{\pi(dx)c_m(x)Q_m(x, dx')}\right\}$$

(formally, this is a Radon-Nikodym derivative) set new state $X^{(t+1)} = x'$, otherwise stay where you are: $X^{(t+1)} = x$

For $B \not\ni x$, $P(x, B) = \int_{x' \in B} \sum_m c_m(x) Q_m(x, dx') \alpha_m(x, x') dx'$.

Metropolis-Hastings on general state spaces

π is target distribution; $P(x, B)$ is the transition kernel.

When current state is $X^{(t)} = x$:

1. Choose move type m with probability $c_m(x)$
2. Propose a new state x', drawn from distribution $Q_m(x, dx')$
3. With probability $\alpha_m(x, x')$, given by

$$\alpha_m(x, x') = \min\left\{1, \frac{\pi(dx')c_m(x')Q_m(x', dx)}{\pi(dx)c_m(x)Q_m(x, dx')}\right\}$$

(formally, this is a Radon-Nikodym derivative) set new state $X^{(t+1)} = x'$, otherwise stay where you are: $X^{(t+1)} = x$

For $B \not\ni x$, $P(x, B) = \int_{x' \in B} \sum_m c_m(x) Q_m(x, dx') \alpha_m(x, x') dx'$.

Metropolis-Hastings: proof of detailed balance

We have to show that $\int_{x \in A} \pi(dx) P(x, B) = \int_{x' \in B} \pi(dx) P(x', A)$ for all A, B.

The LHS is the integral over $A \times B$ of

$$\pi(dx) P(x, dx') = \pi(dx) \sum_m c_m(x) Q_m(x, dx') \alpha_m(x, x')$$

(assuming $A \cap B = \emptyset$) and the choice of $\alpha_m(x, x')$ ensures that this is symmetric in x and x'.

Metropolis-Hastings: proof of detailed balance

We have to show that $\int_{x \in A} \pi(dx) P(x, B) = \int_{x' \in B} \pi(dx) P(x', A)$ for all A, B.

The LHS is the integral over $A \times B$ of

$$\pi(dx) P(x, dx') = \pi(dx) \sum_m c_m(x) Q_m(x, dx') \alpha_m(x, x')$$

(assuming $A \cap B = \emptyset$) and the choice of $\alpha_m(x, x')$ ensures that this is symmetric in x and x'.

Reversible jump MCMC

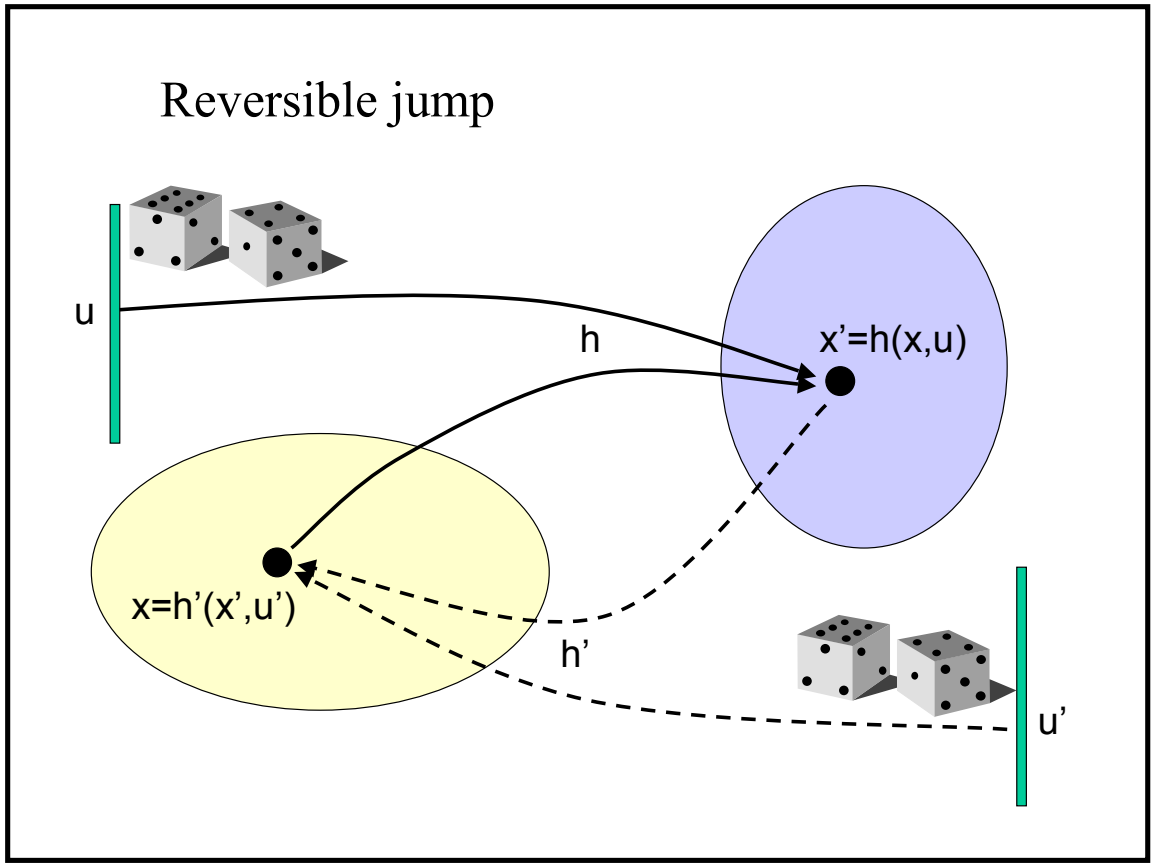
But we shouldn't need to use (or even know) measure theory to do numerical simulation!

A solution is to 'model the program'. How do we simulate any Markov chain? Given the current state $X^{(t)}$,

1. Generate some (uniform) random numbers u
2. Calculate the new state as a deterministic function of the old state and the random numbers: $X^{(t+1)} = h(X^{(t)}, u)$

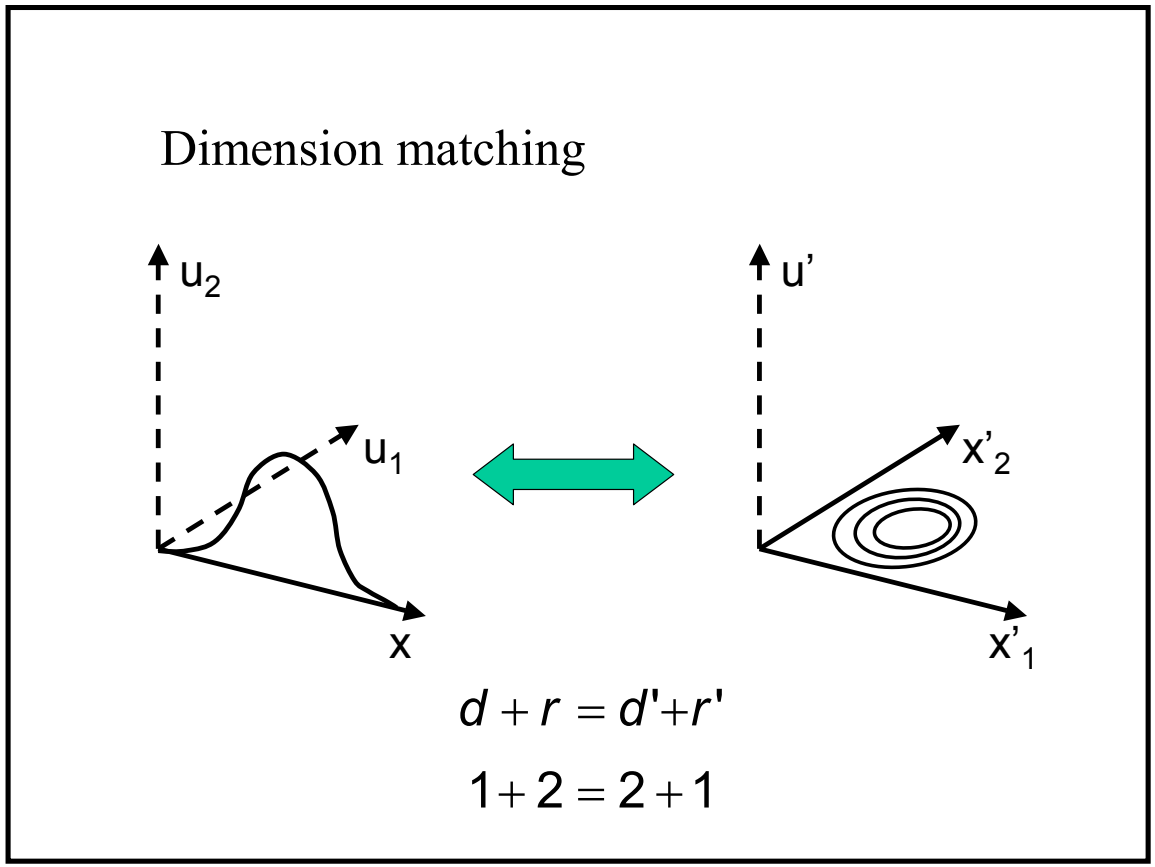
Reversible jump MCMC

Reversible jump MCMC is a Metropolis-Hastings method, employing on-the-fly auxiliary random variables to make difficult jumps between values of $x = (k, \theta)$ in different spaces



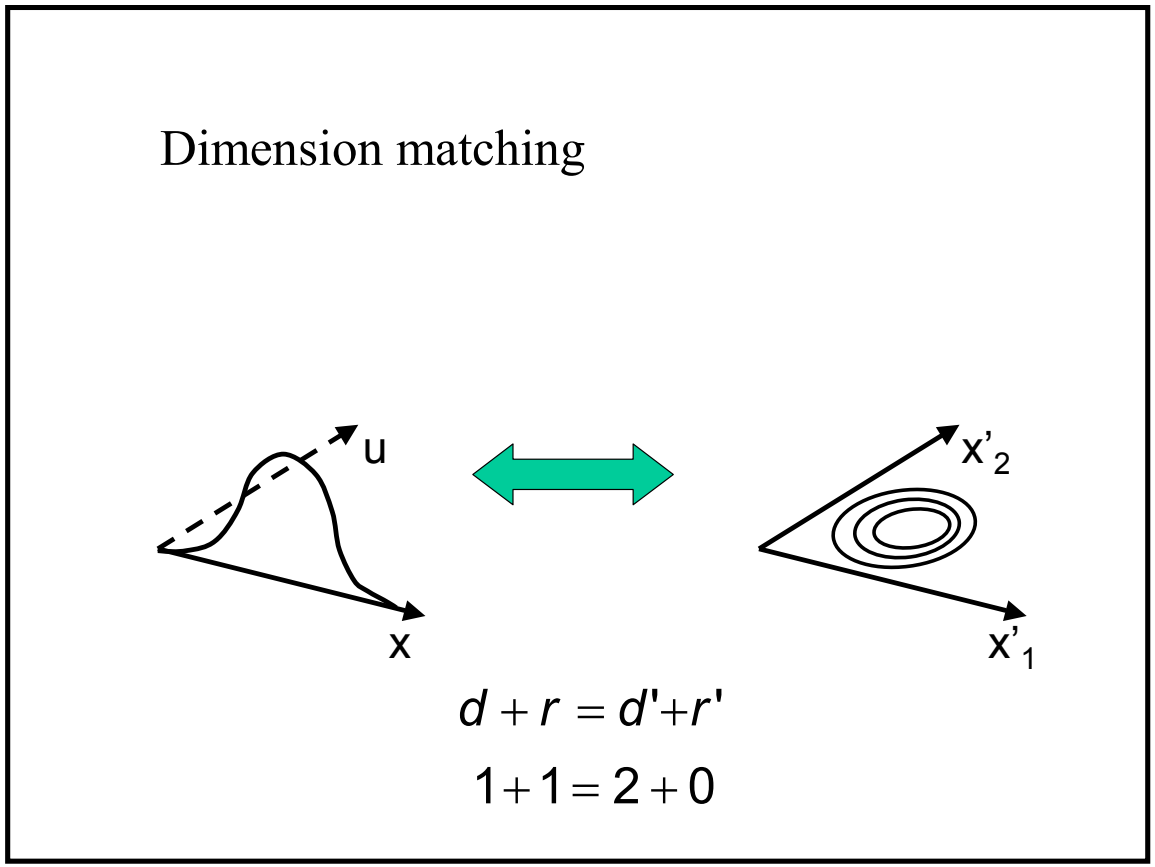
Reversible jump MCMC

The total dimensions of state and auxiliary variables must be preserved



Reversible jump MCMC

The total dimensions of state and auxiliary variables must be preserved



Reversible jump MCMC

The acceptance probability can then be calculated explicitly:

$$\alpha_m(x, x') = \min\left\{1, \frac{\pi(dx')c_m(x')Q_m(x', dx)}{\pi(dx)c_m(x)Q_m(x, dx')}\right\}$$
$$= \min\left\{1, \frac{\pi(x')c_m(x')g_m(u')}{\pi(x)c_m(x)g_m(u)}\left|\frac{\partial(x', u')}{\partial(x, u)}\right|\right\}$$

using ordinary densities.

Reversible jump MCMC

The acceptance probability can then be calculated explicitly:

$$\alpha_m(x, x') = \min\left\{1, \frac{\pi(dx')c_m(x')Q_m(x', dx)}{\pi(dx)c_m(x)Q_m(x, dx')}\right\}$$
$$= \min\left\{1, \frac{\pi(x')c_m(x')g_m(u')}{\pi(x)c_m(x)g_m(u)}\left|\frac{\partial(x', u')}{\partial(x, u)}\right|\right\}$$

using ordinary densities.

Toy example

..... of no statistical use at all!
Suppose x lies in $\mathcal{R} \cup \mathcal{R}^2$: $\pi(x)$ is a mixture:
with probability $p_1 = 0.4$, $x \sim f_B = Beta(2,3)$,
with probability $p_2 = 0.6$, $(x_1, x_2, 1 - x_1 - x_2) \sim f_D = Dirichlet(2,3,4)$.

I will use three moves:

(1) within $\mathcal{R}$: $x \to U(x - \epsilon, x + \epsilon)$.

(2) within $\mathcal{R}^2$: $(x_1, x_2) \to (x_2, x_1)$.

(3) between $\mathcal{R}$ and $\mathcal{R}^2$

In $\mathcal{R}$, choose (1) or (3) with probabilities $1 - r_1, r_1 = 0.7$.
In $\mathcal{R}^2$, choose (2) or (3) with probabilities $1 - r_2, r_2 = 0.4$.
Thus $c_3(x) = r_1$ for all $x \in \mathcal{R}$ and $c_3(x') = c_2$ for all $x' \in \mathcal{R}^2$.

Toy example

..... of no statistical use at all!
Suppose x lies in $\mathcal{R} \cup \mathcal{R}^2$: $\pi(x)$ is a mixture:
with probability $p_1 = 0.4$, $x \sim f_B = Beta(2, 3)$,
with probability $p_2 = 0.6$, $(x_1, x_2, 1 - x_1 - x_2) \sim f_D = Dirichlet(2, 3, 4)$.

I will use three moves:

(1) within $\mathcal{R}$: $x \to U(x - \epsilon, x + \epsilon)$.

(2) within $\mathcal{R}^2$: $(x_1, x_2) \to (x_2, x_1)$.

(3) between $\mathcal{R}$ and $\mathcal{R}^2$

In $\mathcal{R}$, choose (1) or (3) with probabilities $1 - r_1, r_1 = 0.7$.
In $\mathcal{R}^2$, choose (2) or (3) with probabilities $1 - r_2, r_2 = 0.4$.
Thus $c_3(x) = r_1$ for all $x \in \mathcal{R}$ and $c_3(x') = c_2$ for all $x' \in \mathcal{R}^2$.

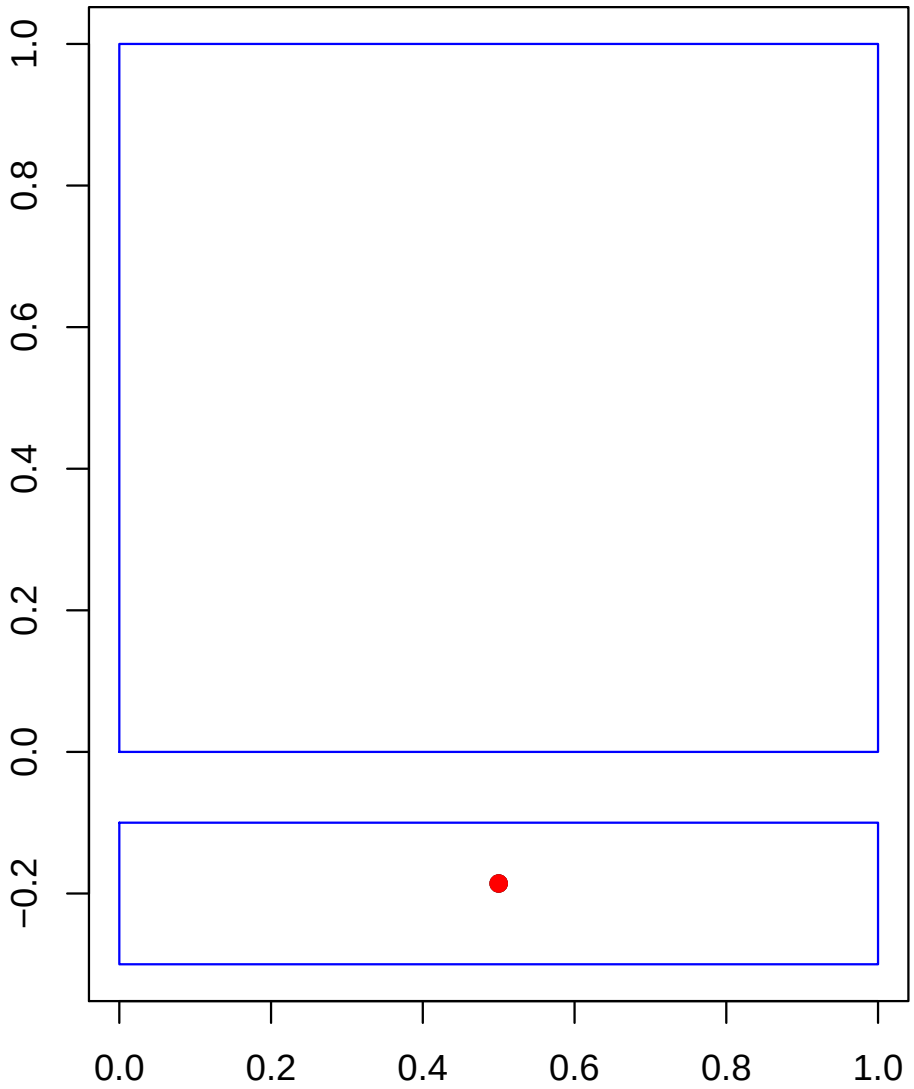
Dimension-changing with move (3)

Proposal:
To go from $x \in \mathcal{R}$ to $(x_1, x_2) \in \mathcal{R}^2$, draw u from $U(0,1)$ [so $g_3(u) = 1$ if $0 < u < 1$] and propose $(x_1, x_2) = (x, u)$. For reverse move, no u' required [write $g_3'(u') \equiv 1$] and set $x = x_1$. This certainly gives a bijection: $(x, u) \leftrightarrow (x_1, x_2)$, with Jacobian $= 1$.
Acceptance decision:

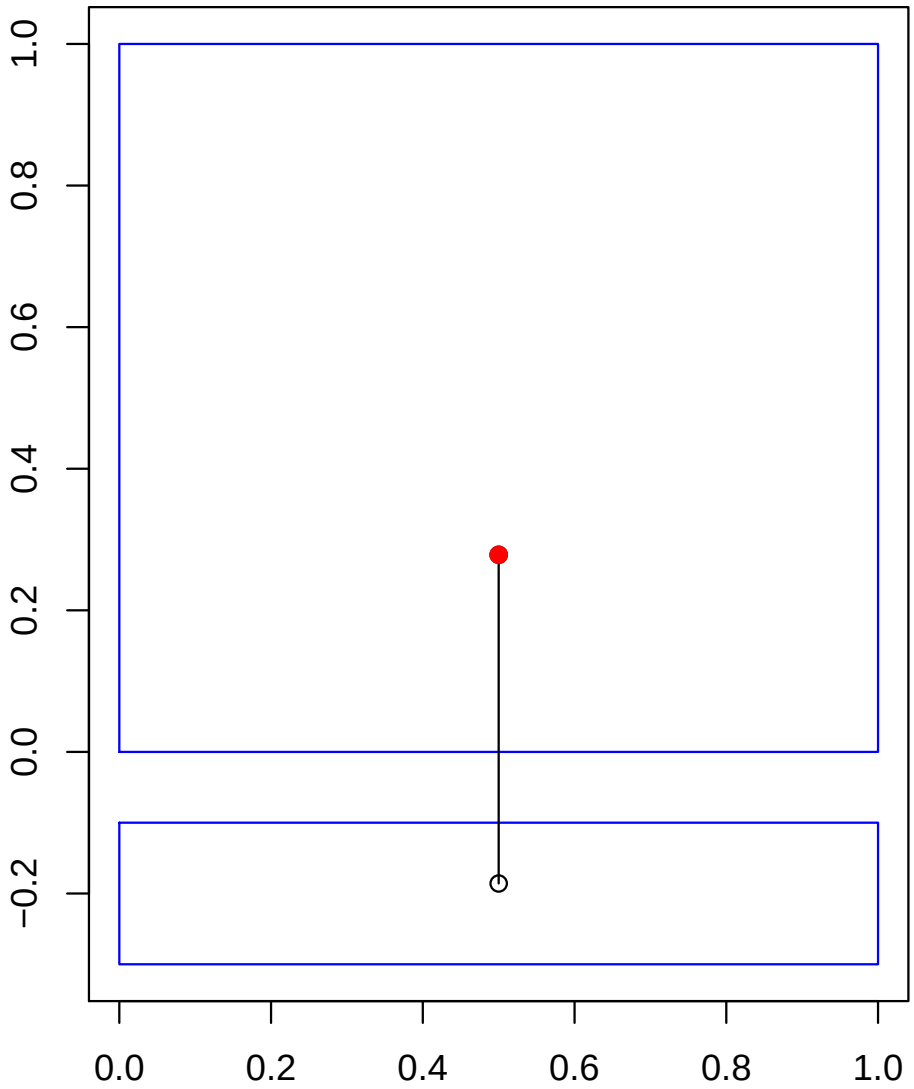
$$\begin{aligned}
\alpha &= \min\left\{1, \frac{\pi(x')}{\pi(x)} \frac{c_3(x')}{c_3(x)} \frac{g_3'(u')}{g_3(u)} \left|\frac{\partial(x', u')}{\partial(x, u)}\right|\right\} \\
&= \min\left\{1, \frac{p_2 f_D(x, u)}{p_1 f_B(x)} \frac{r_2}{r_1} \frac{1}{1} |1|\right\} \\
&= \min\left\{1, \frac{p_2 r_2 f_D(x, u)}{p_1 r_1 f_B(x)}\right\}
\end{aligned}$$

For reverse move, $\alpha = \min\{1, (p_1 r_1 f_B(x))/(p_2 r_2 f_D(x, u))\}$.

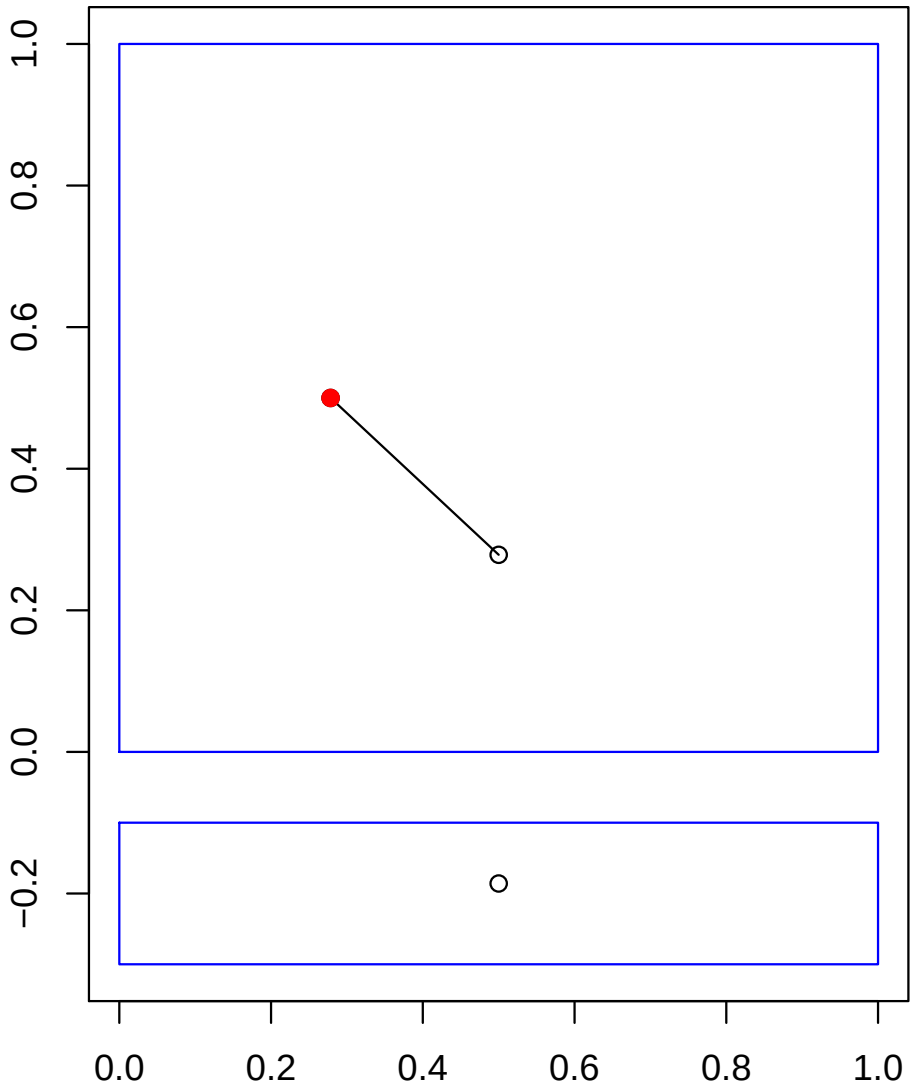
Toy example



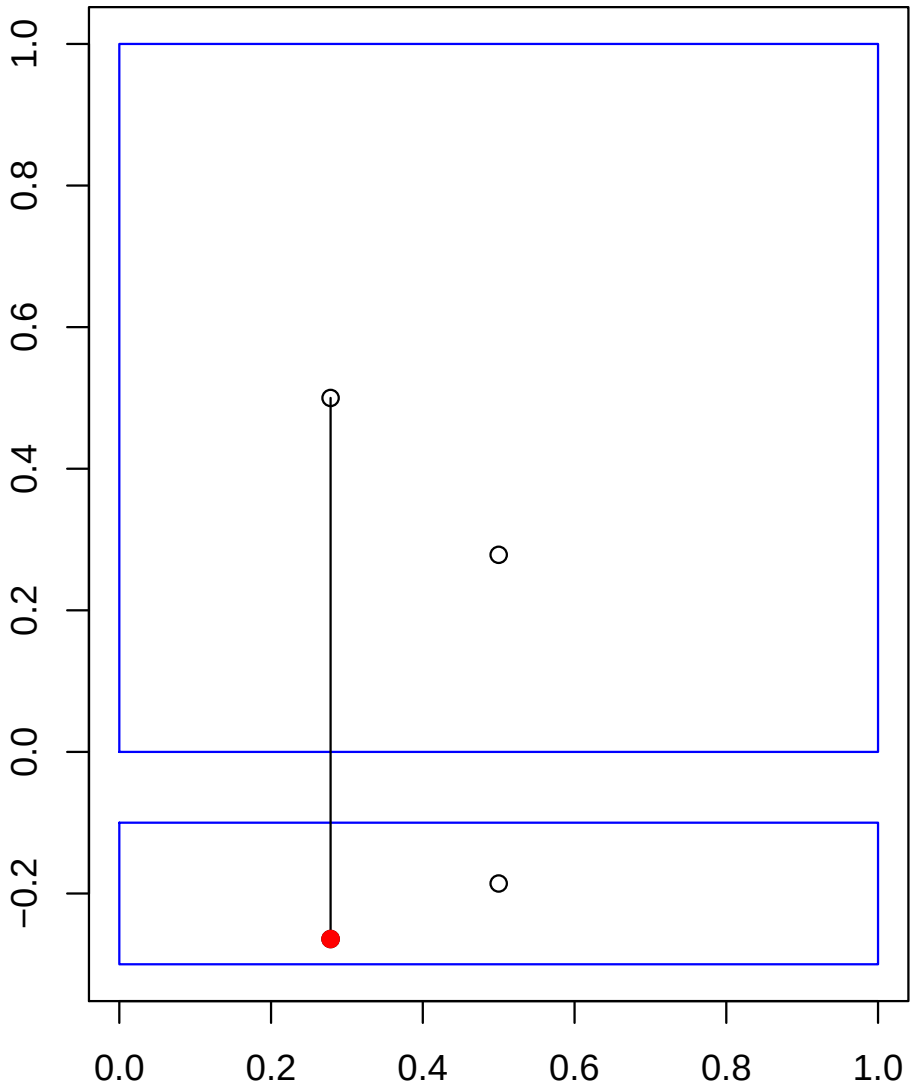
Toy example



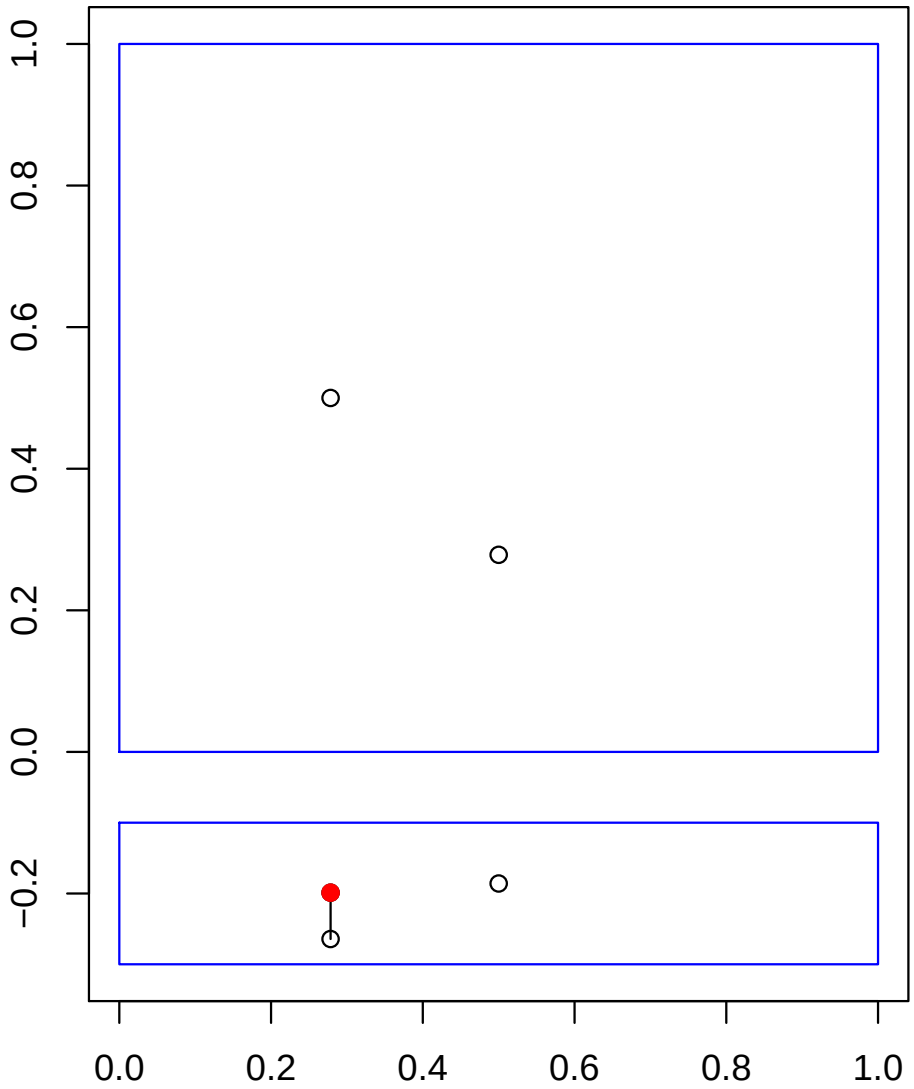
Toy example



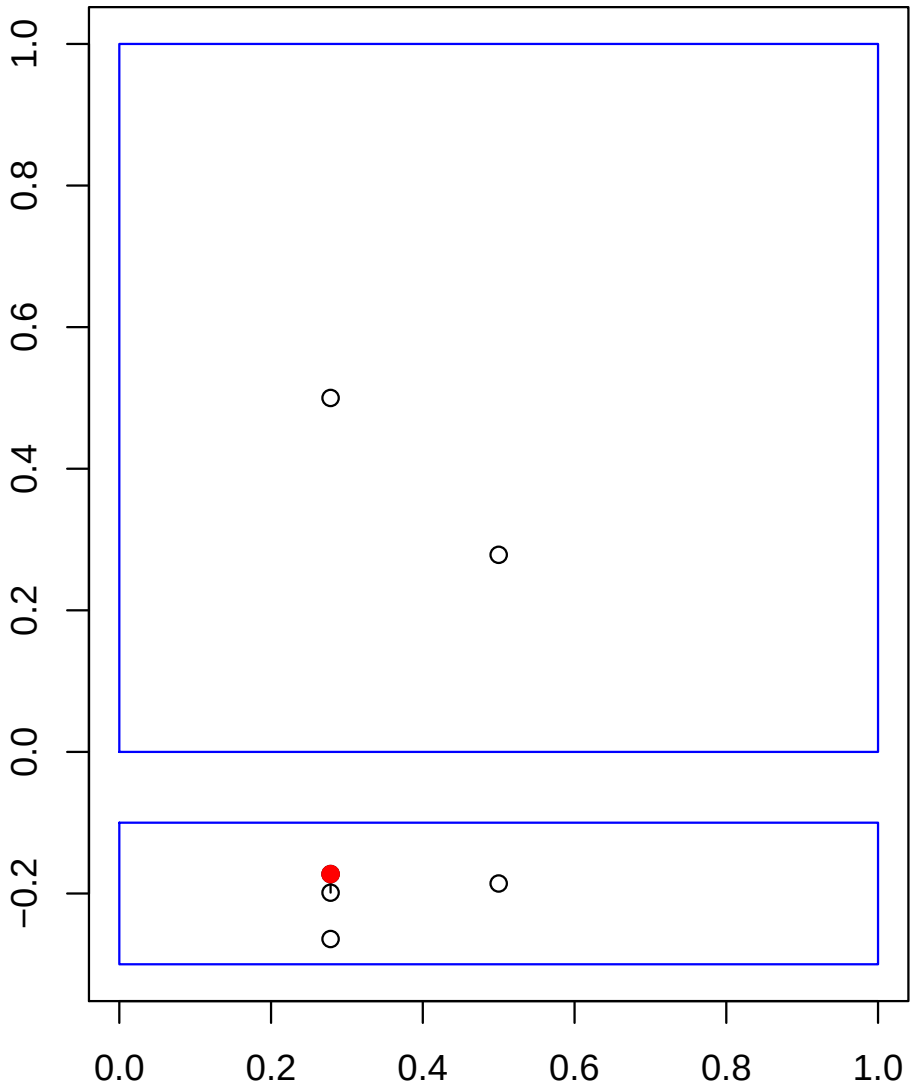
Toy example



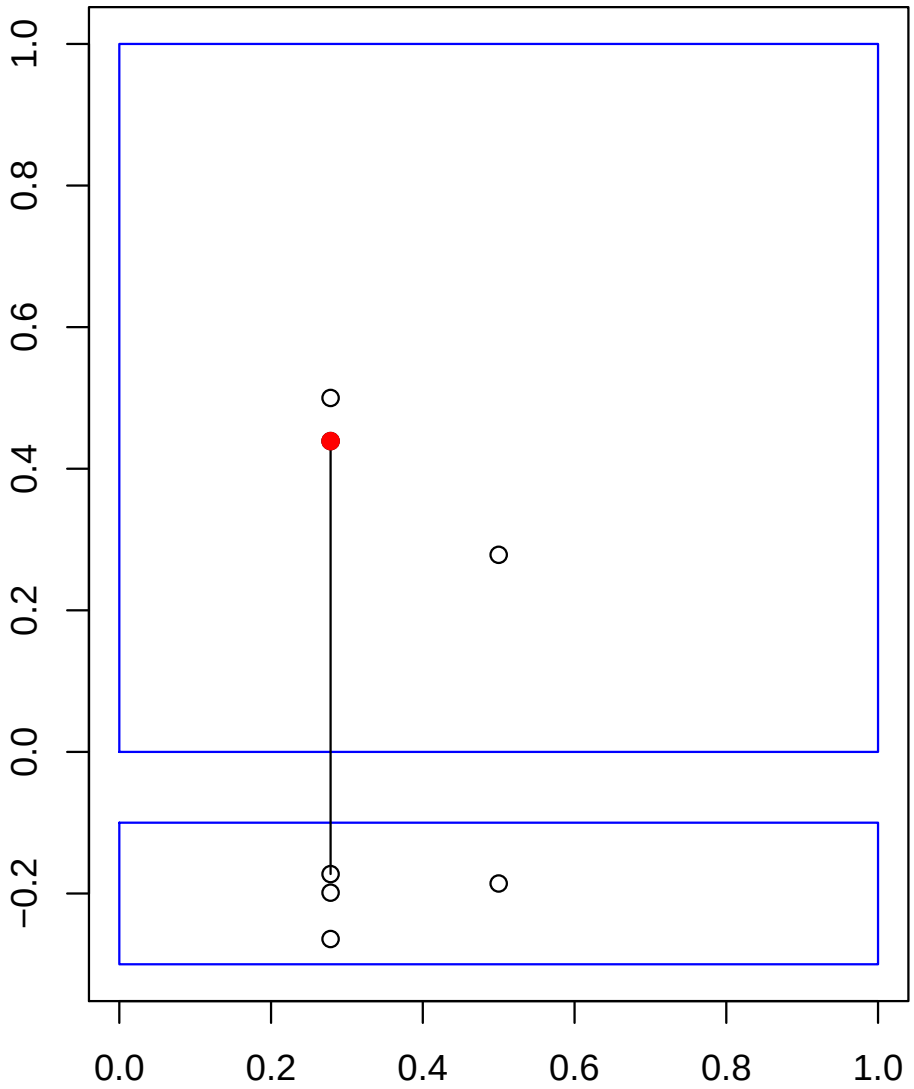
Toy example



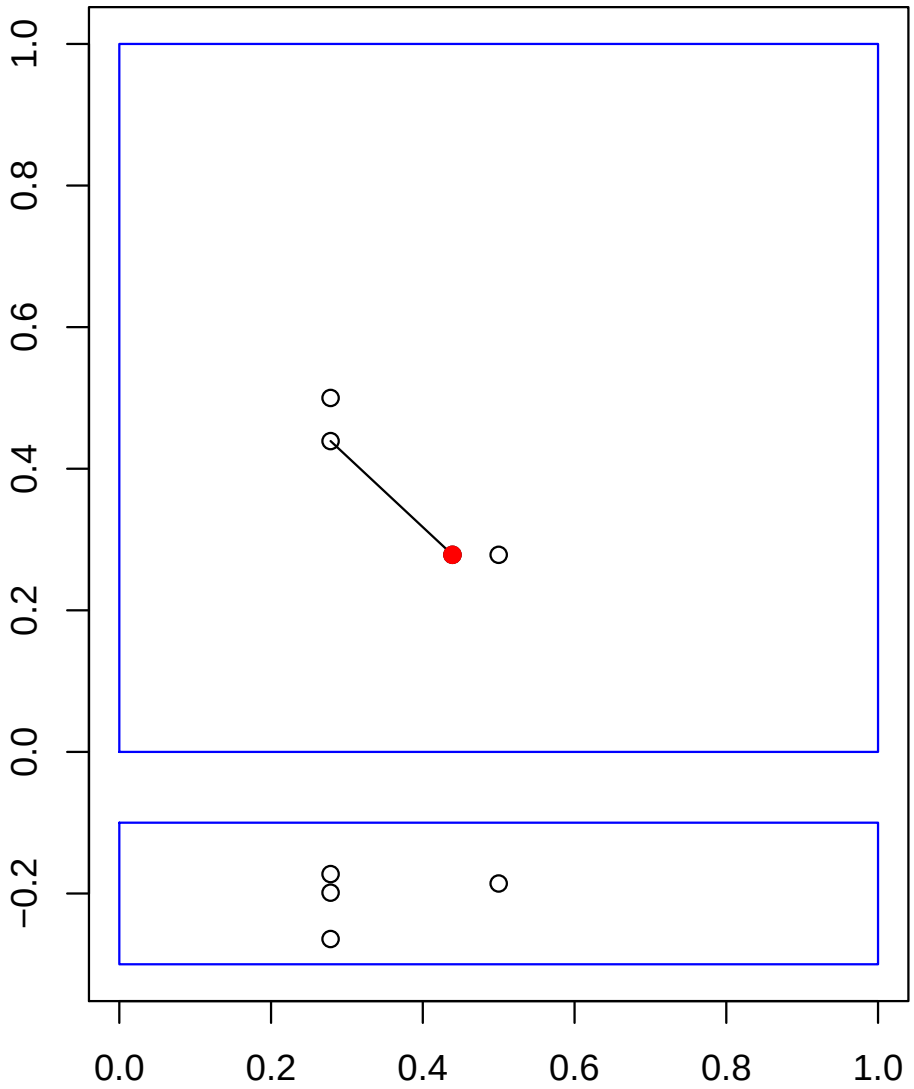
Toy example



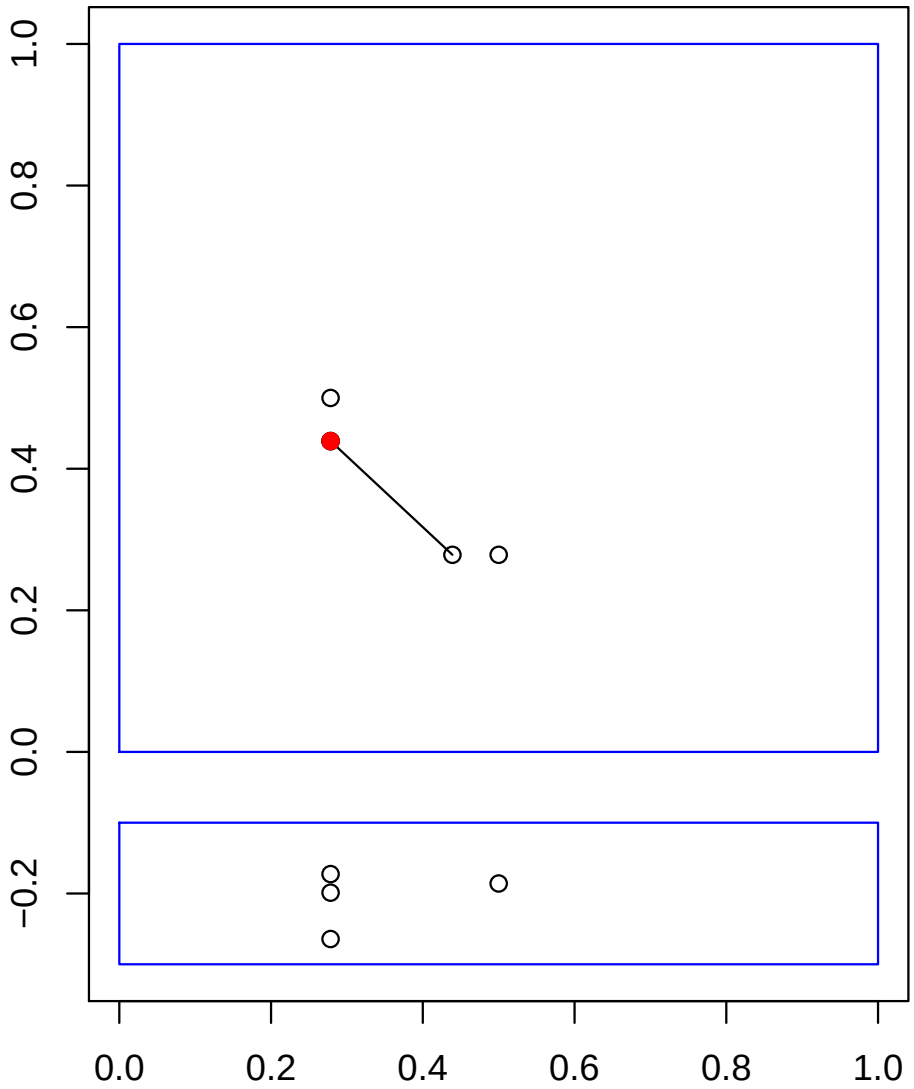
Toy example



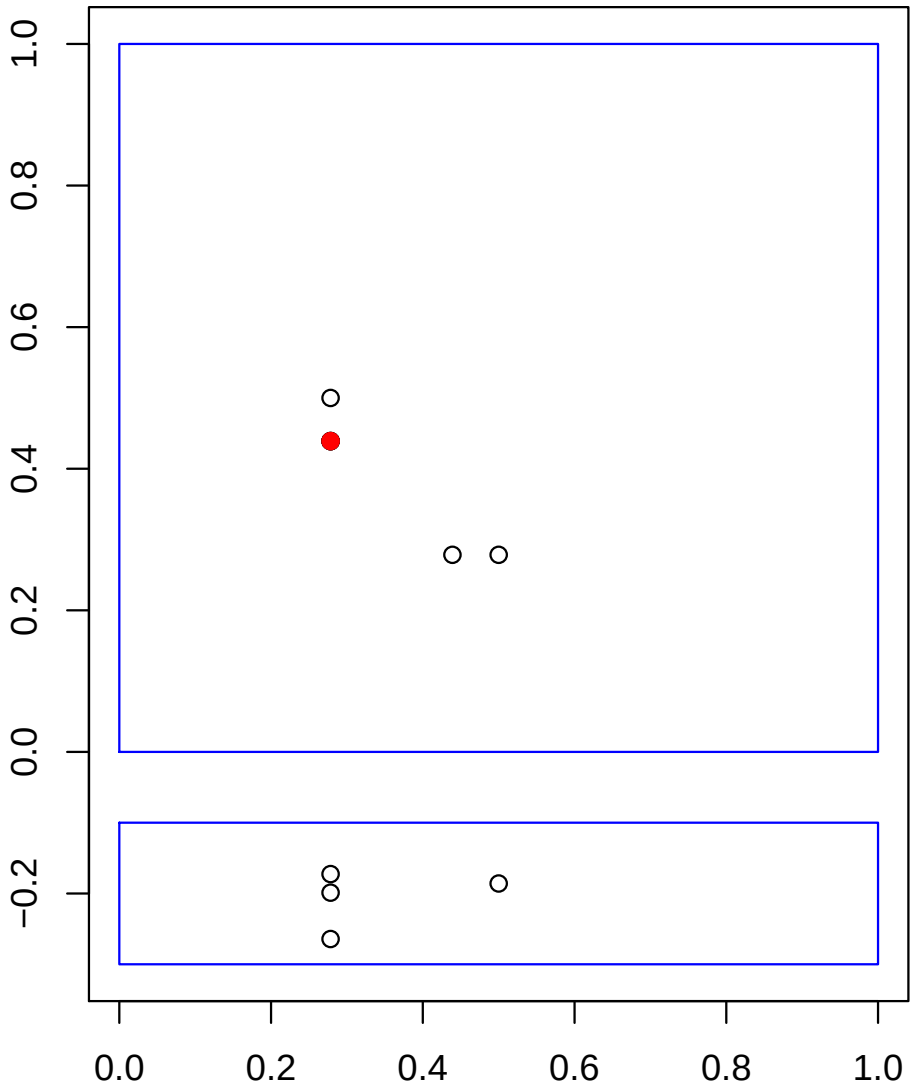
Toy example



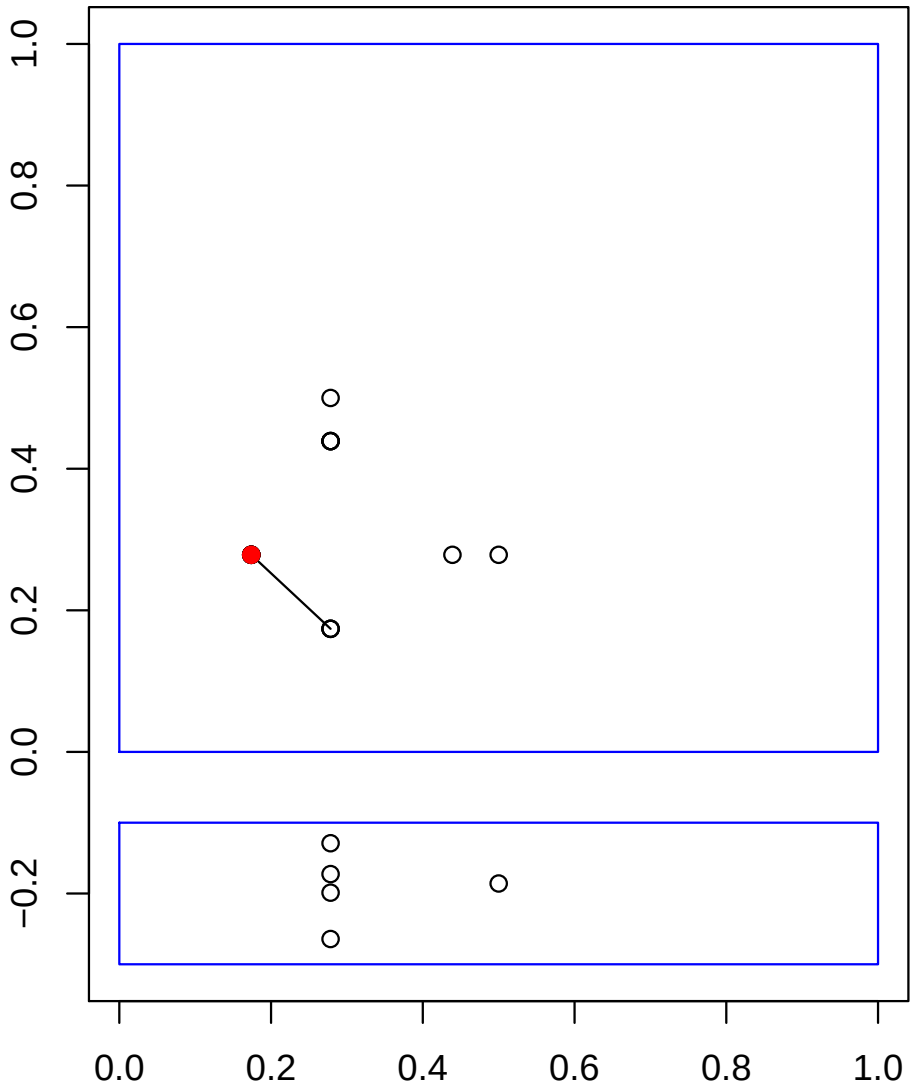
Toy example



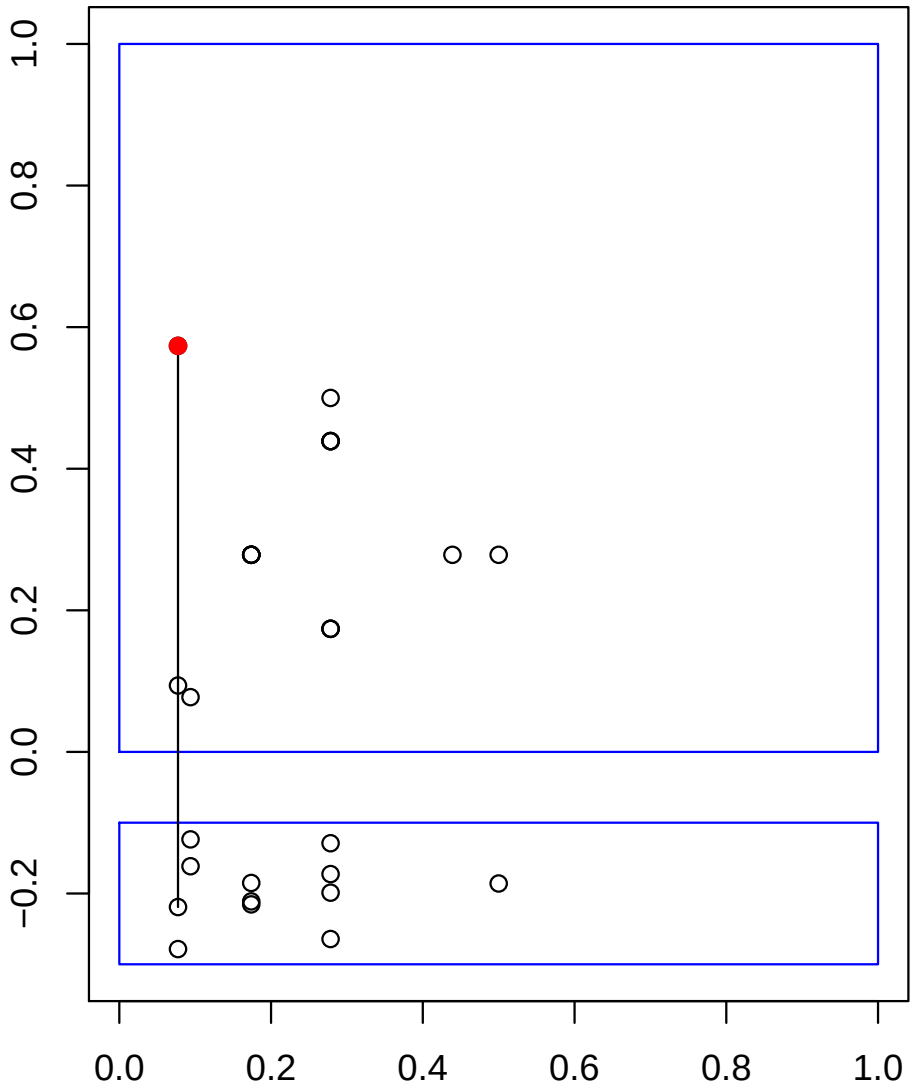
Toy example



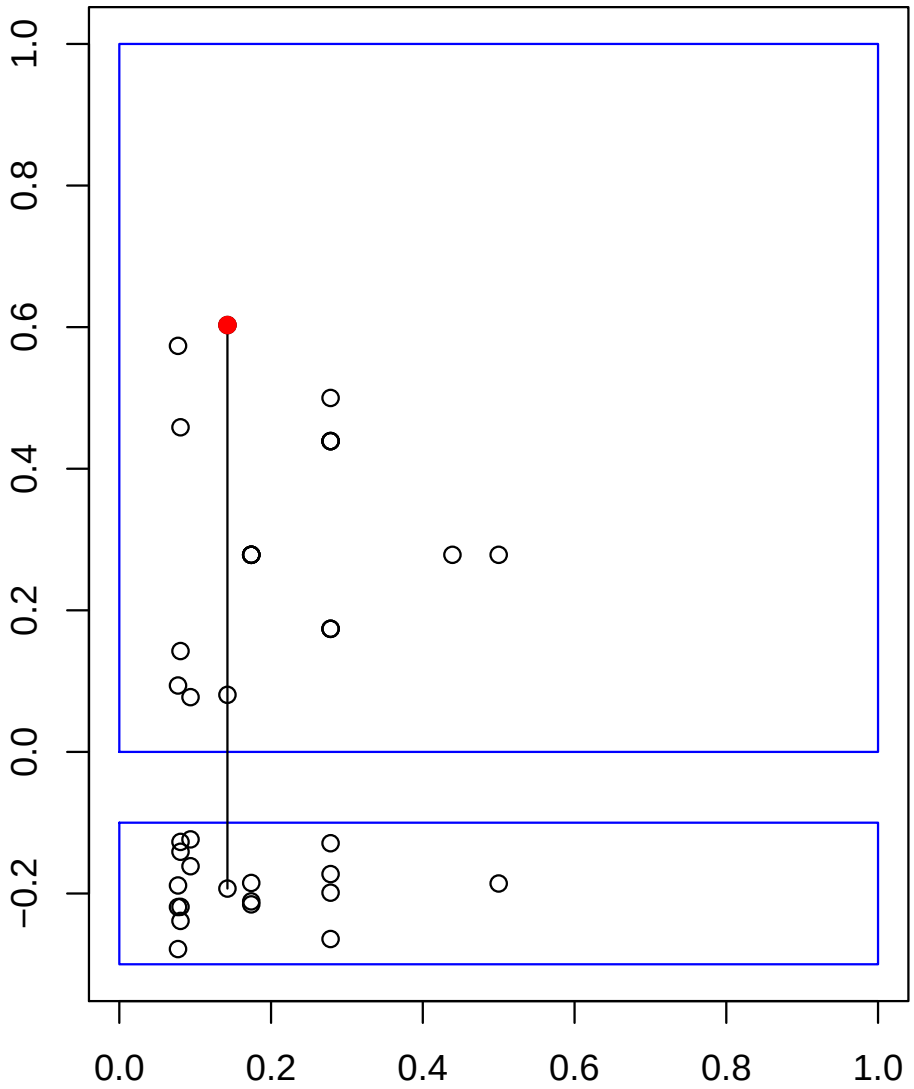
Toy example



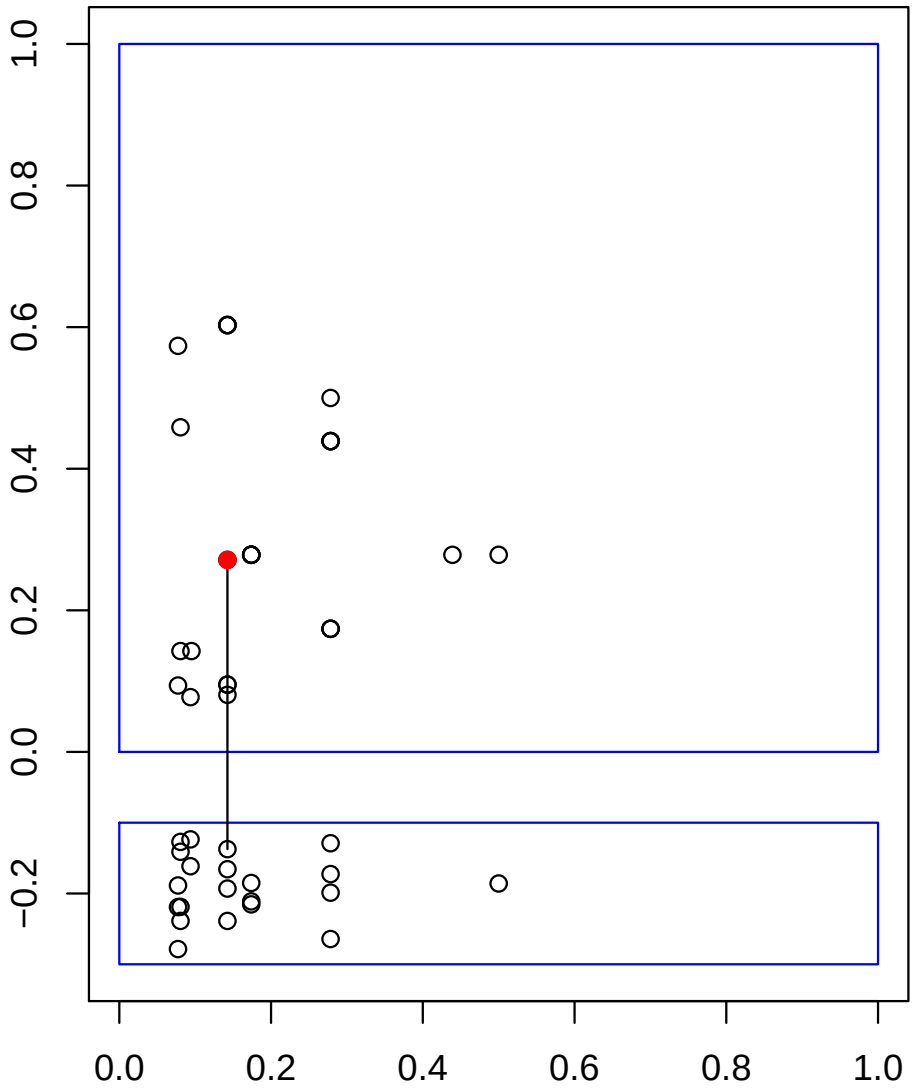
Toy example



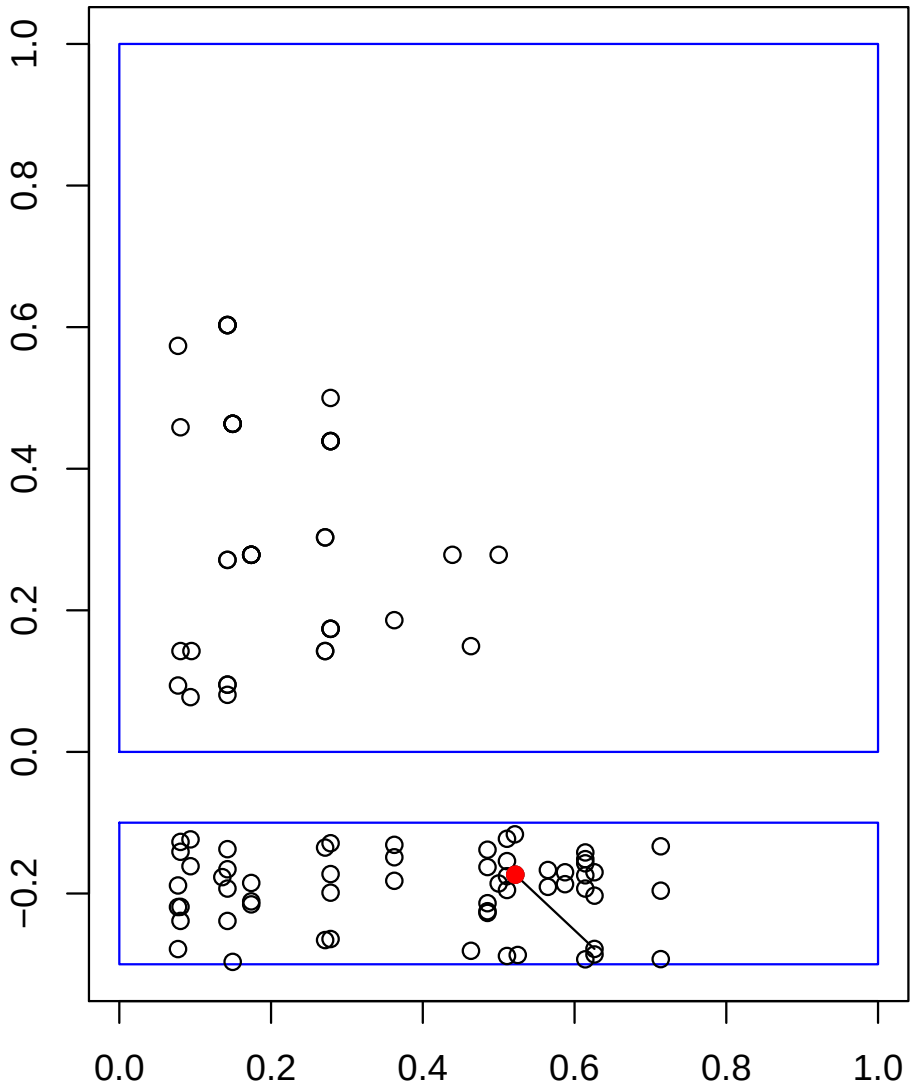
Toy example



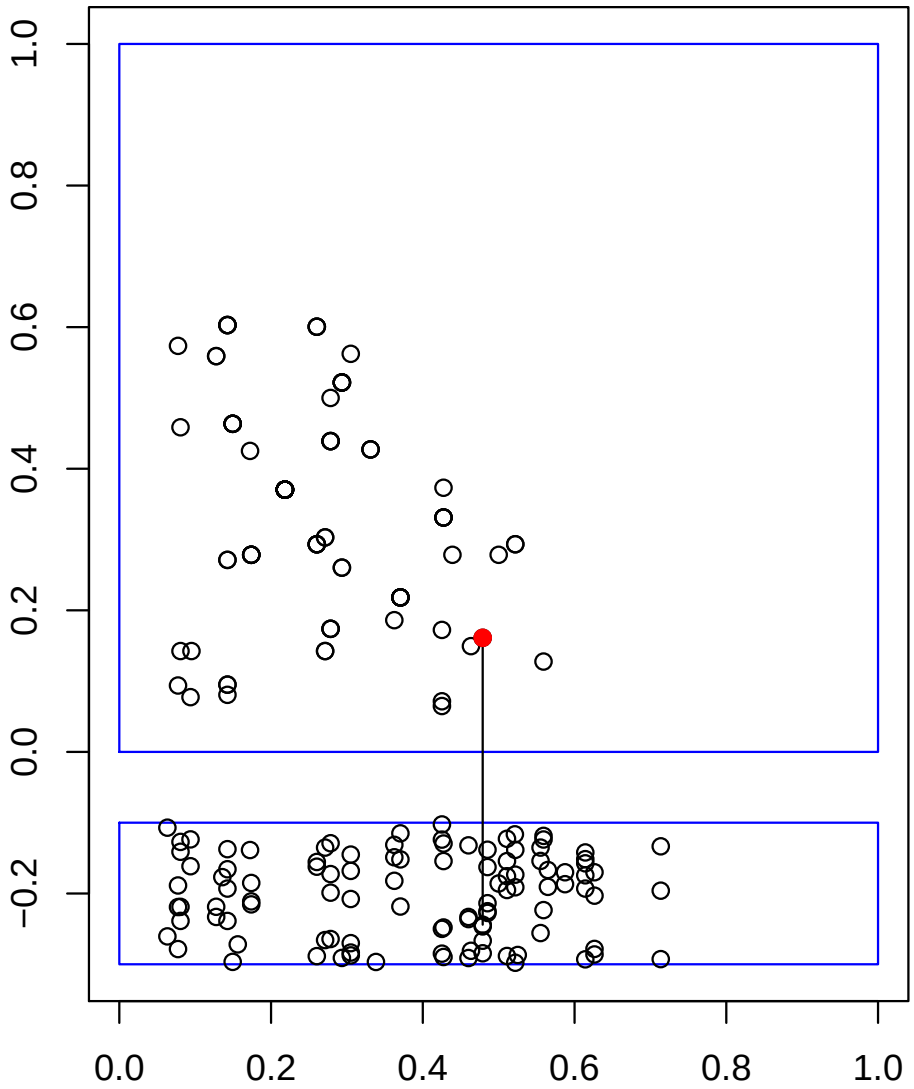
Toy example



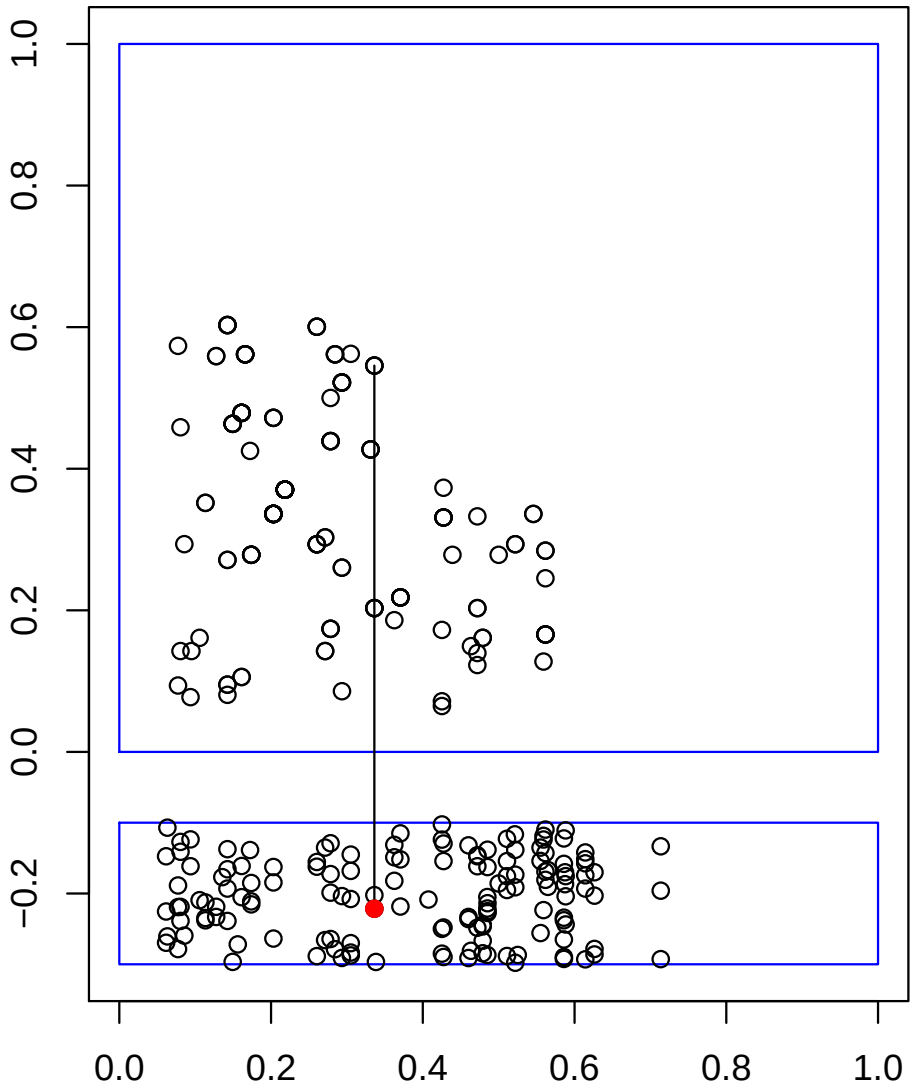
Toy example



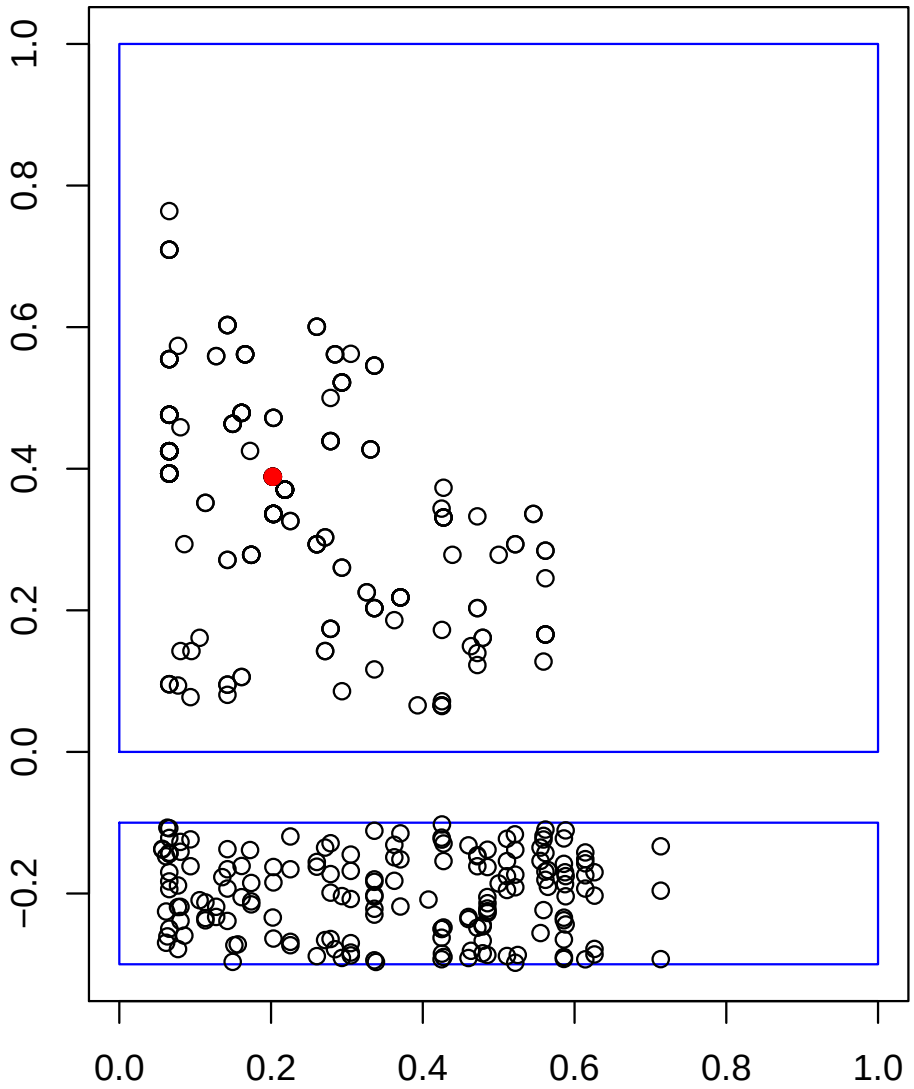
Toy example



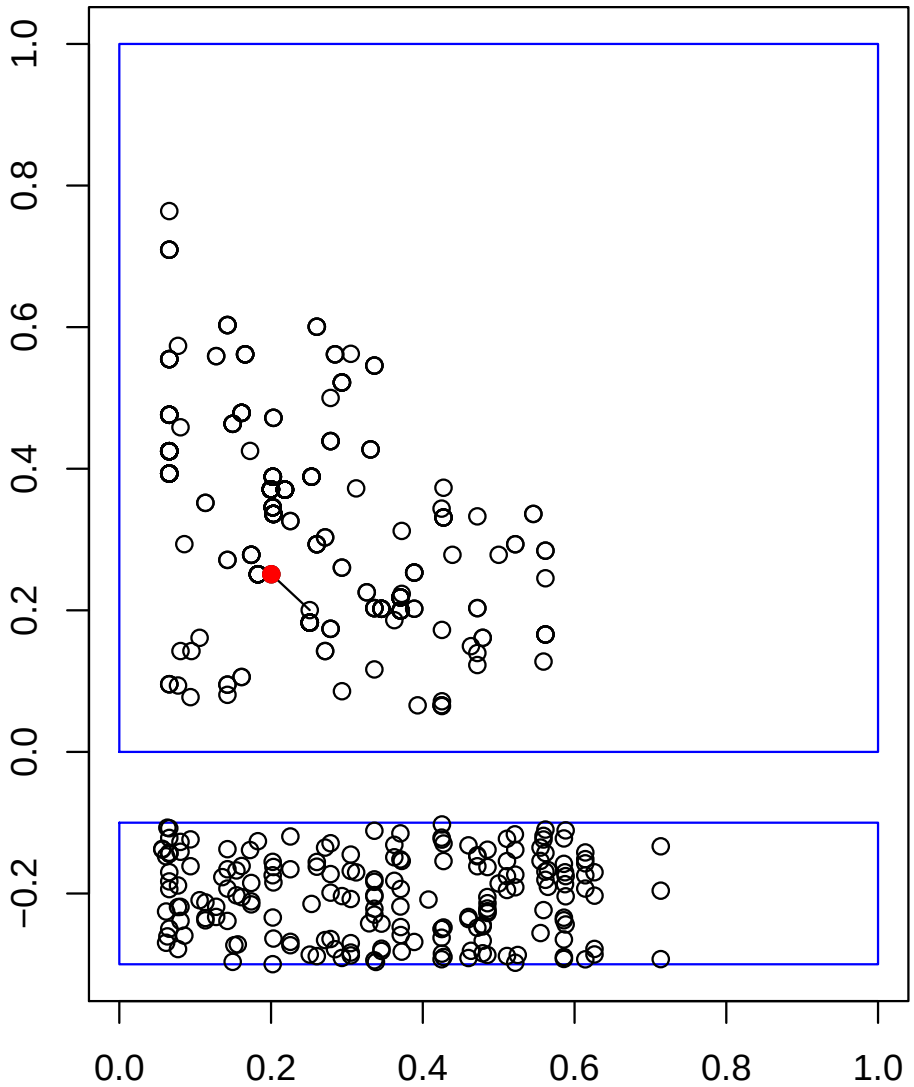
Toy example



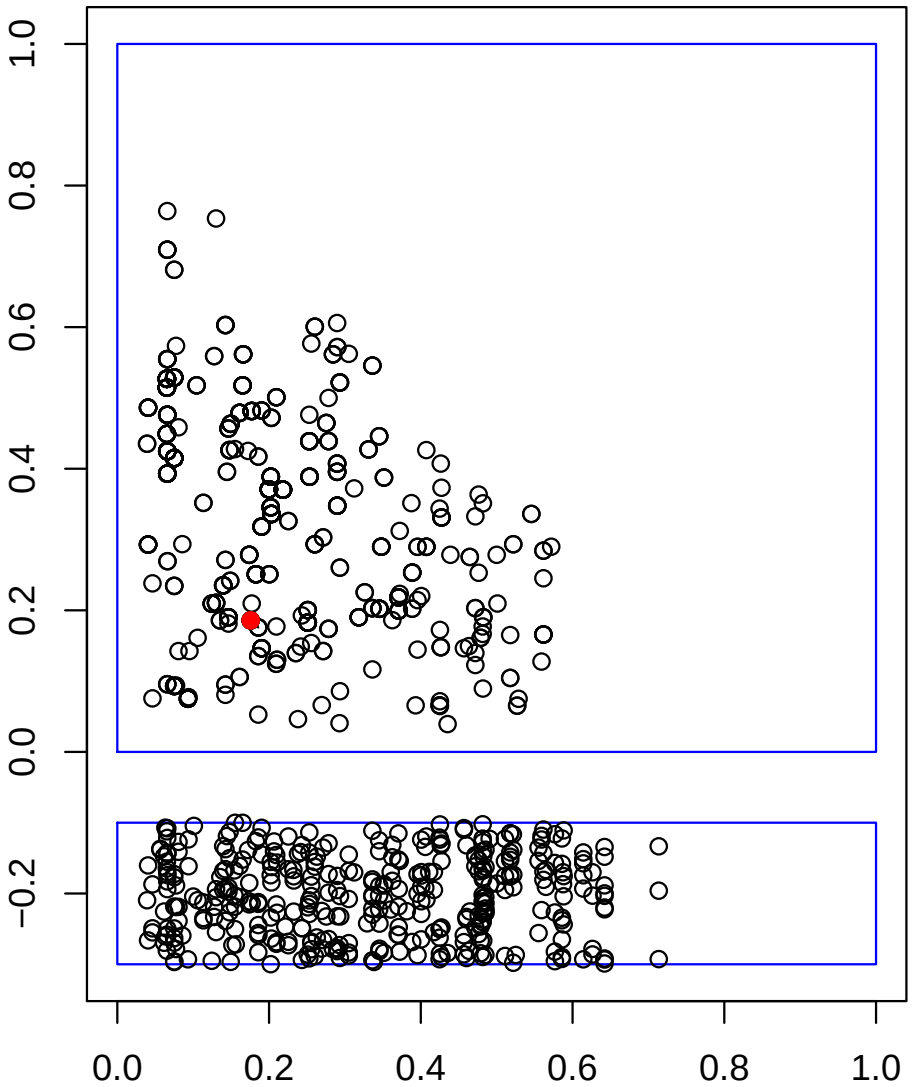
Toy example



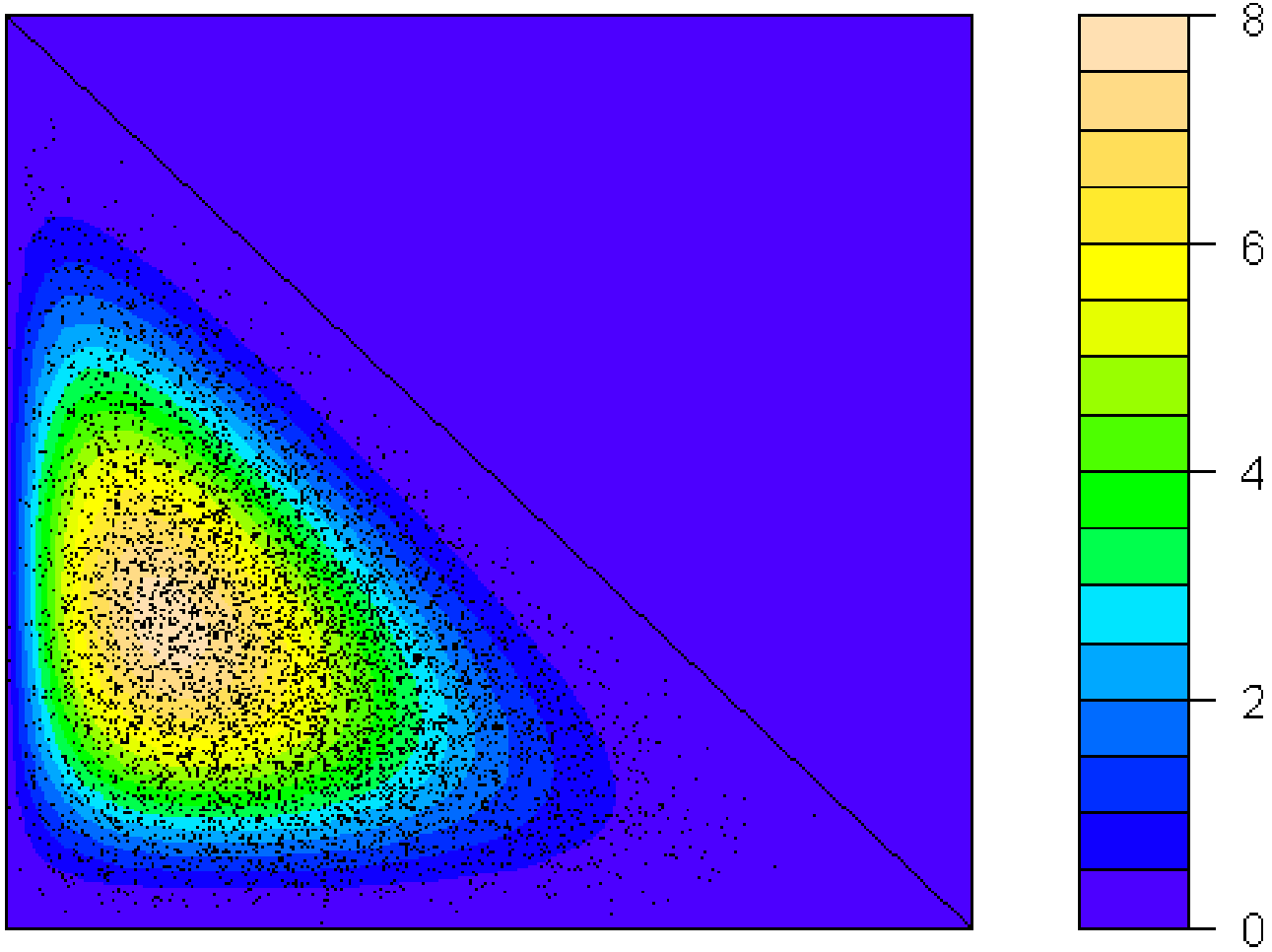
Toy example



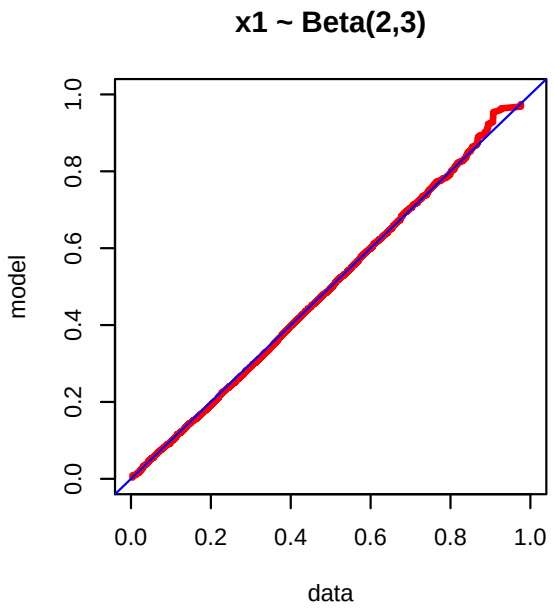
Toy example

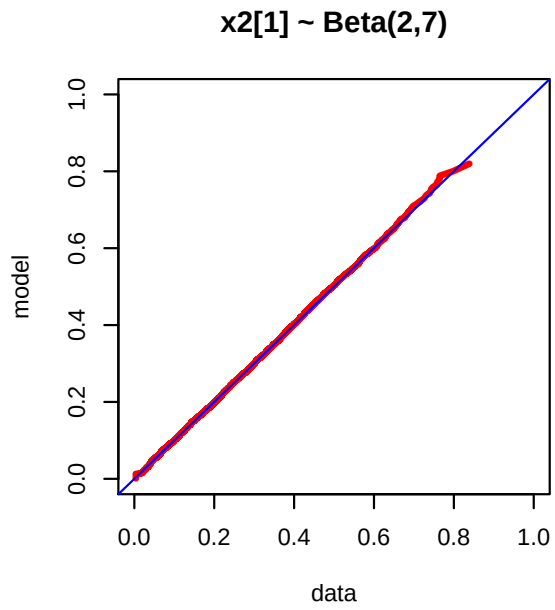


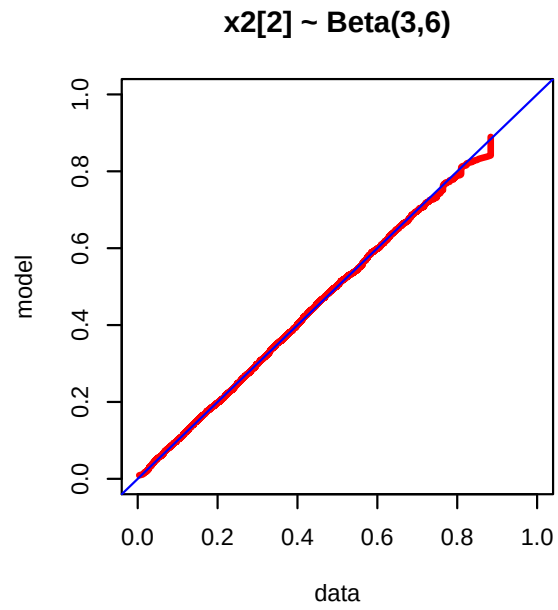
Toy



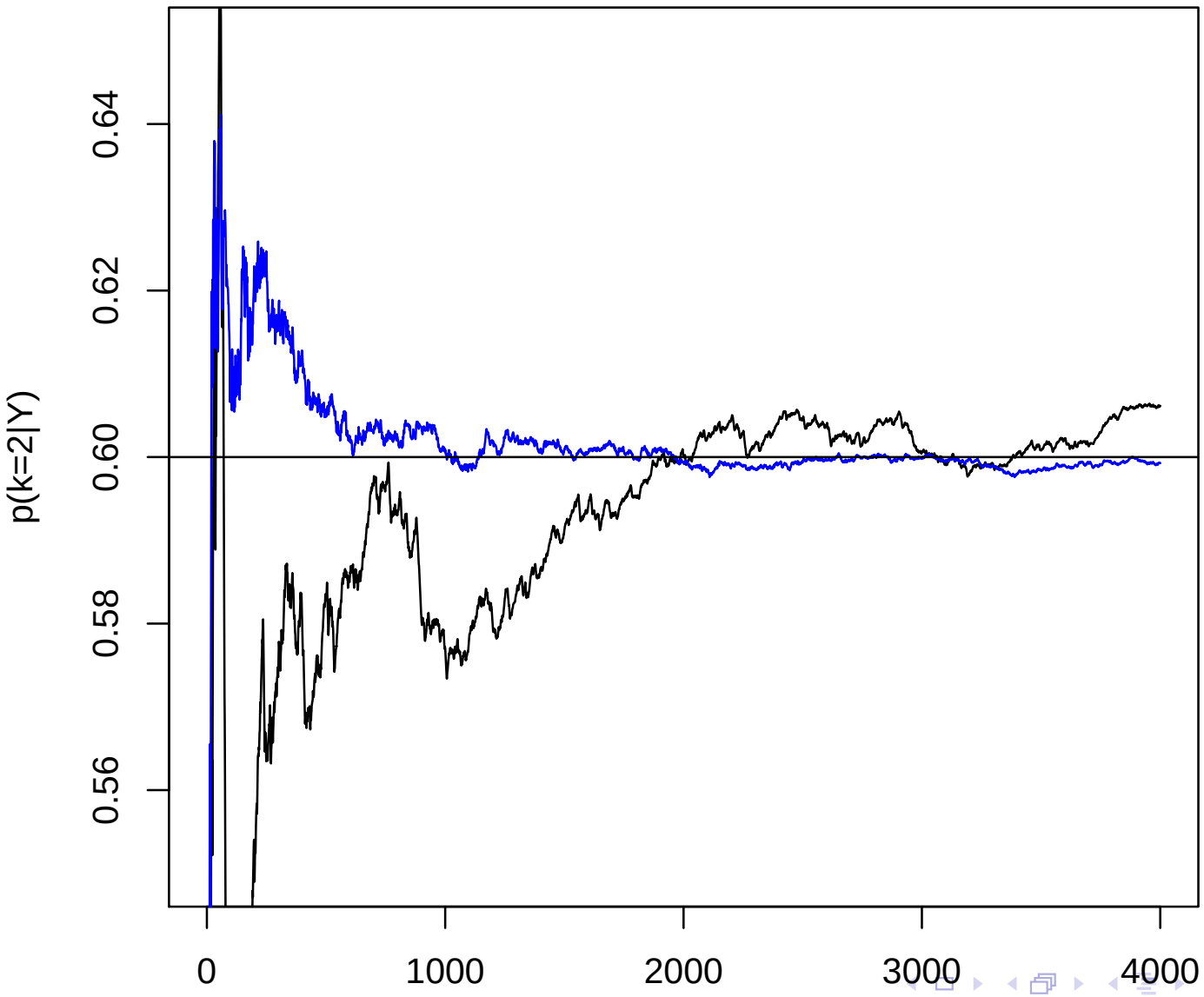
Toy example


x1 ~ Beta(2,3)
model
data
0.0
0.2
0.4
0.6
0.8
1.0


x2[1] ~ Beta(2,7)
model
data
0.0
0.2
0.4
0.6
0.8
1.0


x2[2] ~ Beta(3,6)
model
data
0.0
0.2
0.4
0.6
0.8
1.0

Toy example



Alternatives to joint model-parameter sampling

- When marginal likelihoods $p(Y|k) = \int p(Y|k,\theta_k)p(\theta_k|k)d\theta_k$ are tractable, it's usually a good idea to compute them (thus marginalising over θ_k) then conduct search/sampling only over the model indicator k.
- Marginal likelihoods via within-model sampling.

Within-model sampling

It is possible to use a run of a MCMC sampler for the within-model posterior $p(\theta_k|Y,k)$ to estimate the marginal likelihoods $p(Y|k)$, separately, for each model k.

Then these marginal likelihoods can be combined to compute Bayes factors

$$B_{kl} = \frac{p(Y|k)}{p(Y|l)}$$

and hence joint model-parameter posteriors

$$p(k,\theta_k|Y) = p(k|Y)p(\theta_k|k,Y) = \frac{B_{kl}p(k)}{\sum_{k'} B_{k'l}p(k')}p(\theta_k|k,Y)$$

Estimating marginal likelihoods by MCMC is a subtle matter, and the subject of much ongoing research.

Within-model sampling

It is possible to use a run of a MCMC sampler for the within-model posterior $p(\theta_k|Y,k)$ to estimate the marginal likelihoods $p(Y|k)$, separately, for each model k.

Then these marginal likelihoods can be combined to compute Bayes factors

$$B_{kl} = \frac{p(Y|k)}{p(Y|l)}$$

and hence joint model-parameter posteriors

$$p(k,\theta_k|Y) = p(k|Y)p(\theta_k|k,Y) = \frac{B_{kl}p(k)}{\sum_{k'} B_{k'l}p(k')}p(\theta_k|k,Y)$$

Estimating marginal likelihoods by MCMC is a subtle matter, and the subject of much ongoing research.

Within-model sampling

It is possible to use a run of a MCMC sampler for the within-model posterior $p(\theta_k|Y,k)$ to estimate the marginal likelihoods $p(Y|k)$, separately, for each model k.

Then these marginal likelihoods can be combined to compute Bayes factors

$$B_{kl} = \frac{p(Y|k)}{p(Y|l)}$$

and hence joint model-parameter posteriors

$$p(k,\theta_k|Y) = p(k|Y)p(\theta_k|k,Y) = \frac{B_{kl}p(k)}{\sum_{k'} B_{k'l}p(k')}p(\theta_k|k,Y)$$

Estimating marginal likelihoods by MCMC is a subtle matter, and the subject of much ongoing research.

Within-model sampling

It is possible to use a run of a MCMC sampler for the within-model posterior $p(\theta_k|Y,k)$ to estimate the marginal likelihoods $p(Y|k)$, separately, for each model k.

Then these marginal likelihoods can be combined to compute Bayes factors

$$B_{kl} = \frac{p(Y|k)}{p(Y|l)}$$

and hence joint model-parameter posteriors

$$p(k,\theta_k|Y) = p(k|Y)p(\theta_k|k,Y) = \frac{B_{kl}p(k)}{\sum_{k'} B_{k'l}p(k')}p(\theta_k|k,Y)$$

Estimating marginal likelihoods by MCMC is a subtle matter, and the subject of much ongoing research.

Some issues in choosing a sampler

- How many models are there?
- Do we want results across k, within each k, or for one k of interest?
- Do we need the Evidence (marginal likelihood) values $p(Y|k)$ absolutely, or only relatively?
- Jumping between models as an aid to mixing (c.f. simulated tempering: mixing may be better in the 'other' model)
- Are samplers for individual models already written and tested?
- Are standard strategies like split/merge likely to work?
- Trade-off between remembering and forgetting θ_k when leaving model k

The different ways that models can interconnect

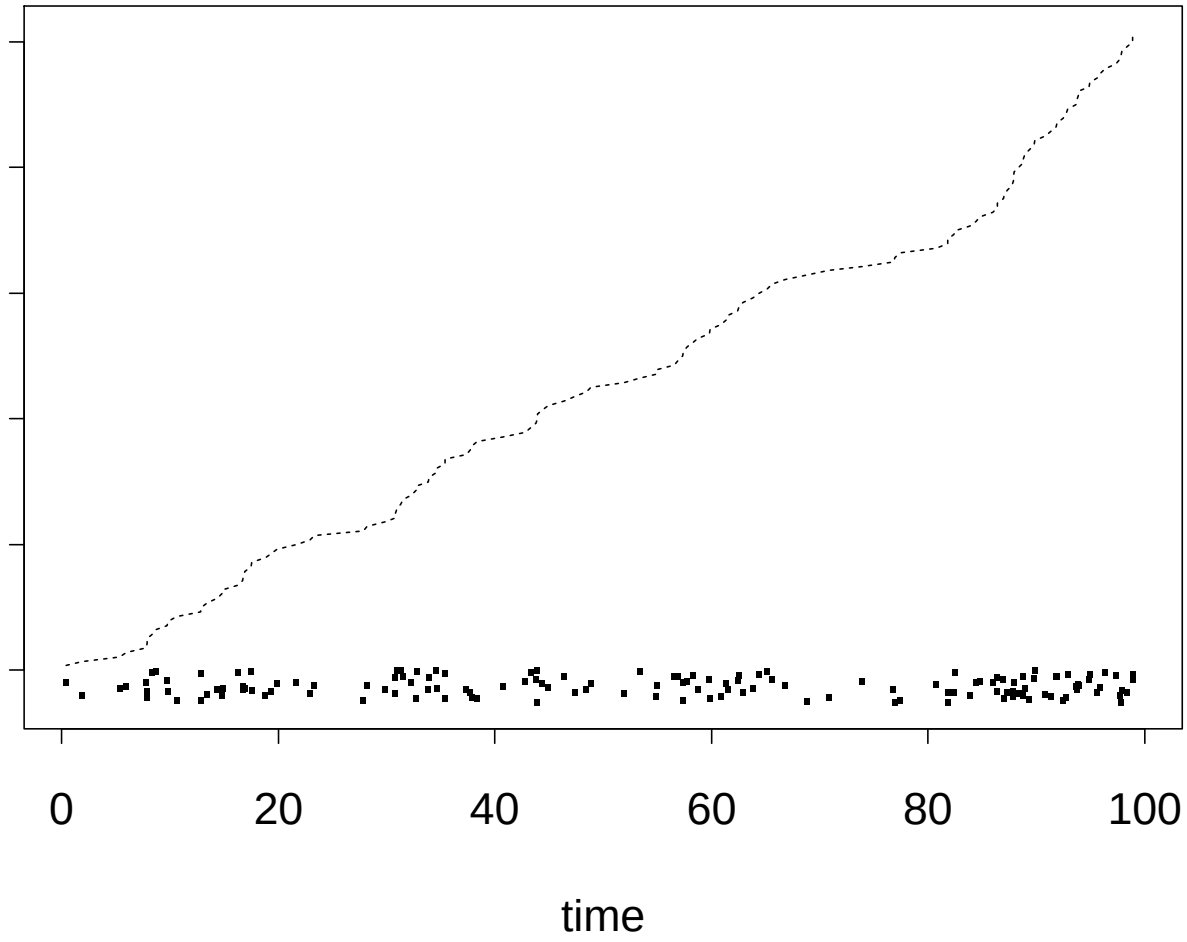
- completely unrelated
- nested in some irregular way
- mixture models (nesting and exchangeability)
- variable selection (factorial structure)
- graphical models (possibly with constraints such as decomposability)

Methodological connections and extensions

- Jump–diffusion (Grenander & Miller, 1994, Phillips & Smith, 1996)
- Point process representations and samplers (Geyer & Møller, 1994, Stephens, 2000)
- Product-space samplers (Carlin & Chib, 1995, Green & O'Hagan, 1998, Dellaportas *et al.*, 2002)
- Delayed rejection (Green & Mira, 2001)
- Connections between reversible jump and continuous time birth-and-death samplers (Cappé, Robert & Rydén, 2001)
- Composite model space framework (Godsill, 2001)
- Efficient construction of proposals (Brooks, Giudici & Roberts, 2003)
- Automatic RJ sampler (Hastie, 2005)
- Population RJMCMC and Interacting SMC (Jasra *et al*, 2007/8)

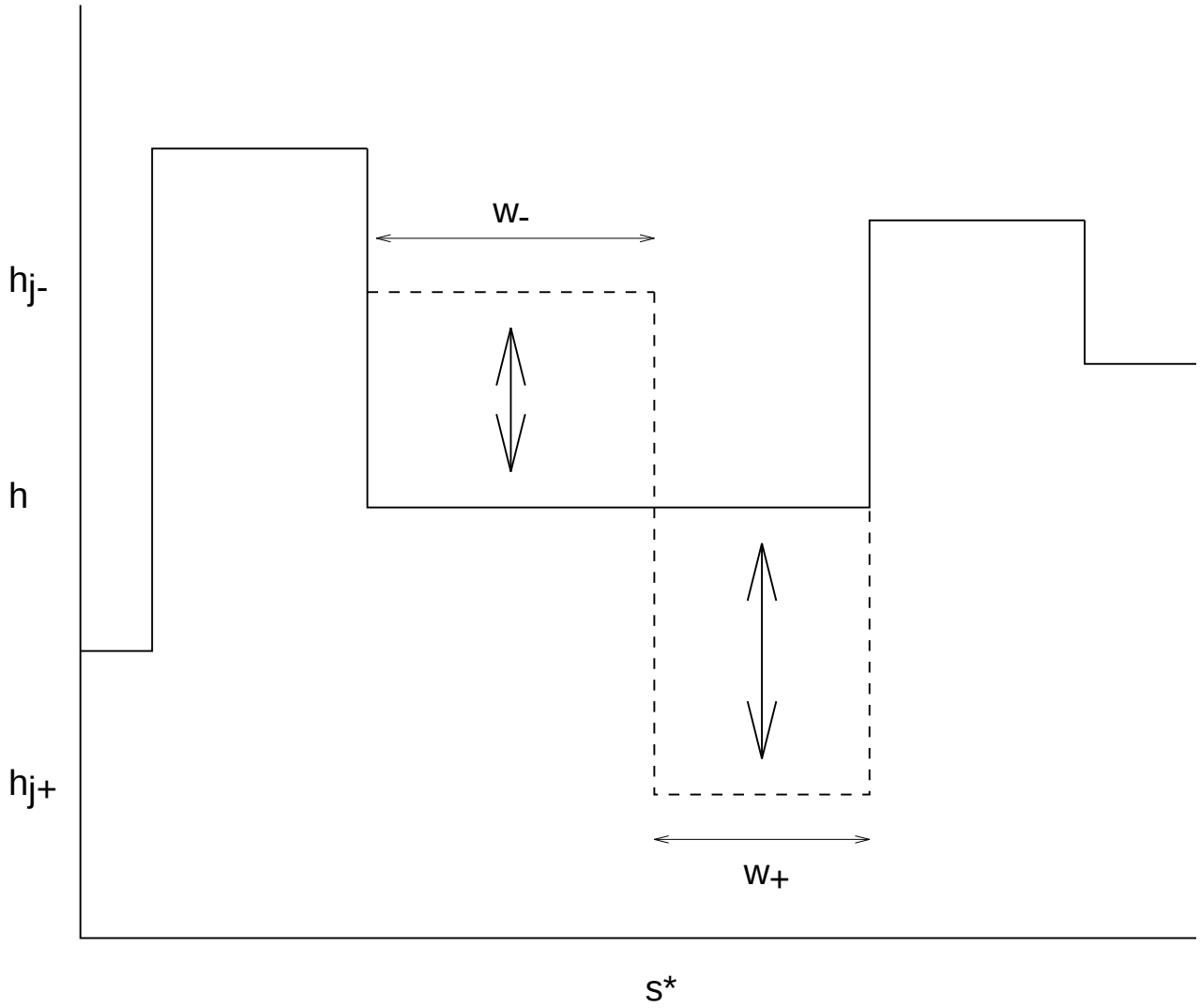
Cyclones hitting the Bay of Bengal

141 cyclones over a period of 100 years (a cyclone is a storm with winds > 88 km h^{-1})



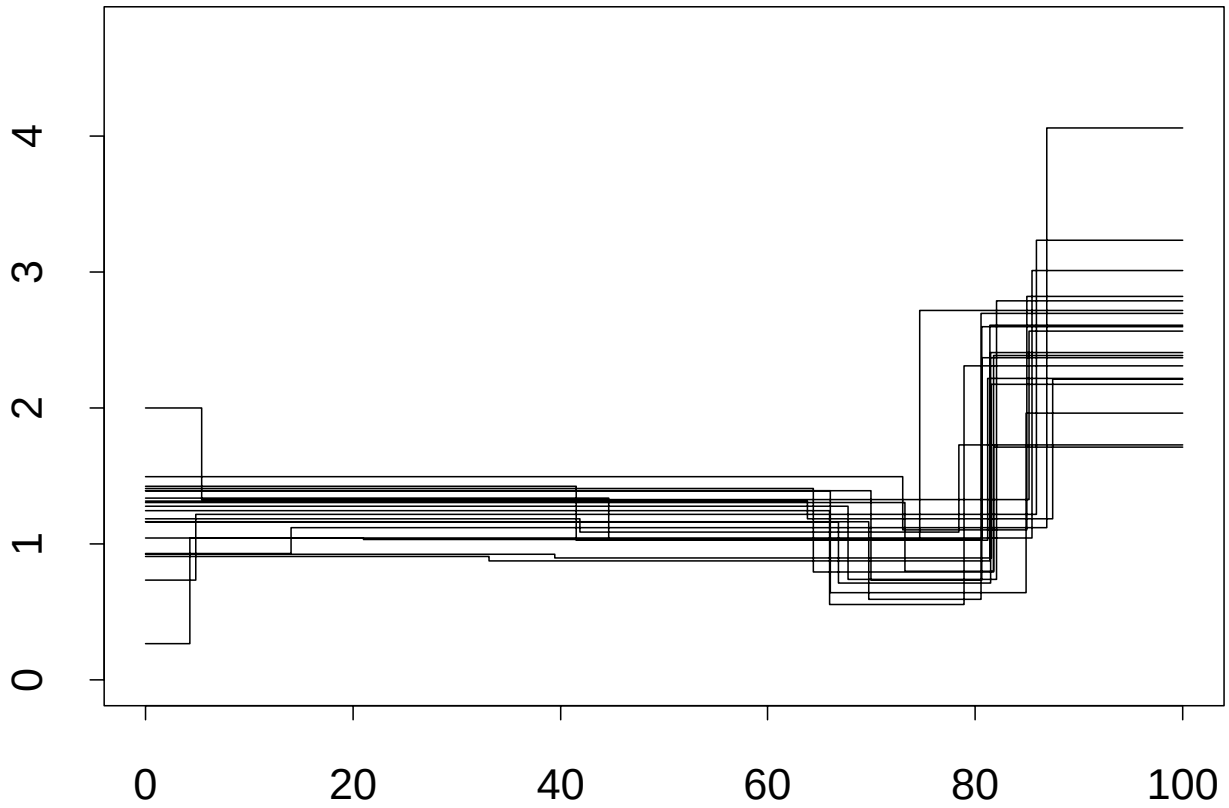
Cyclones hitting the Bay of Bengal

We model this point process as an inhomogeneous Poisson process, whose intensity is a step function with an unknown number of steps; the dimension-changing move requites splitting and merging steps



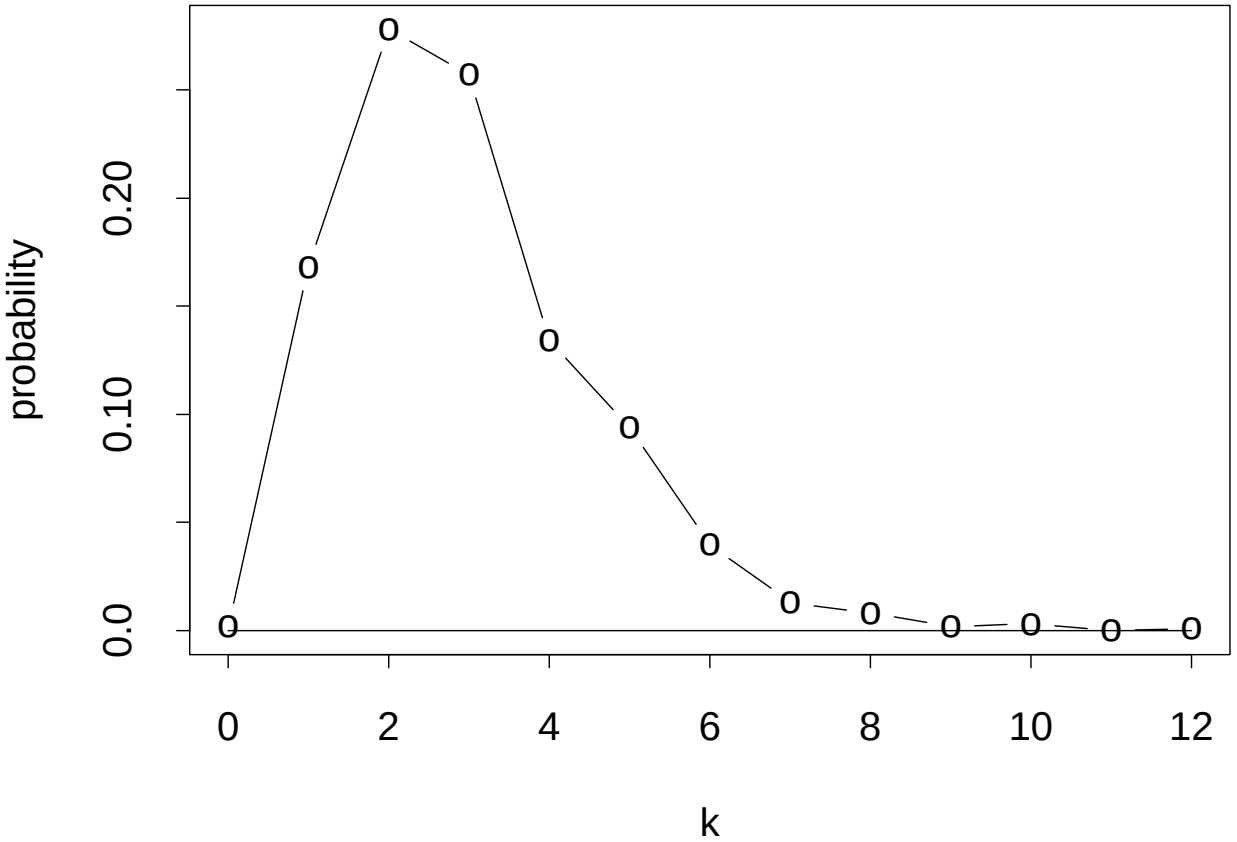
Cyclones hitting the Bay of Bengal

Sample from the posterior over step functions.



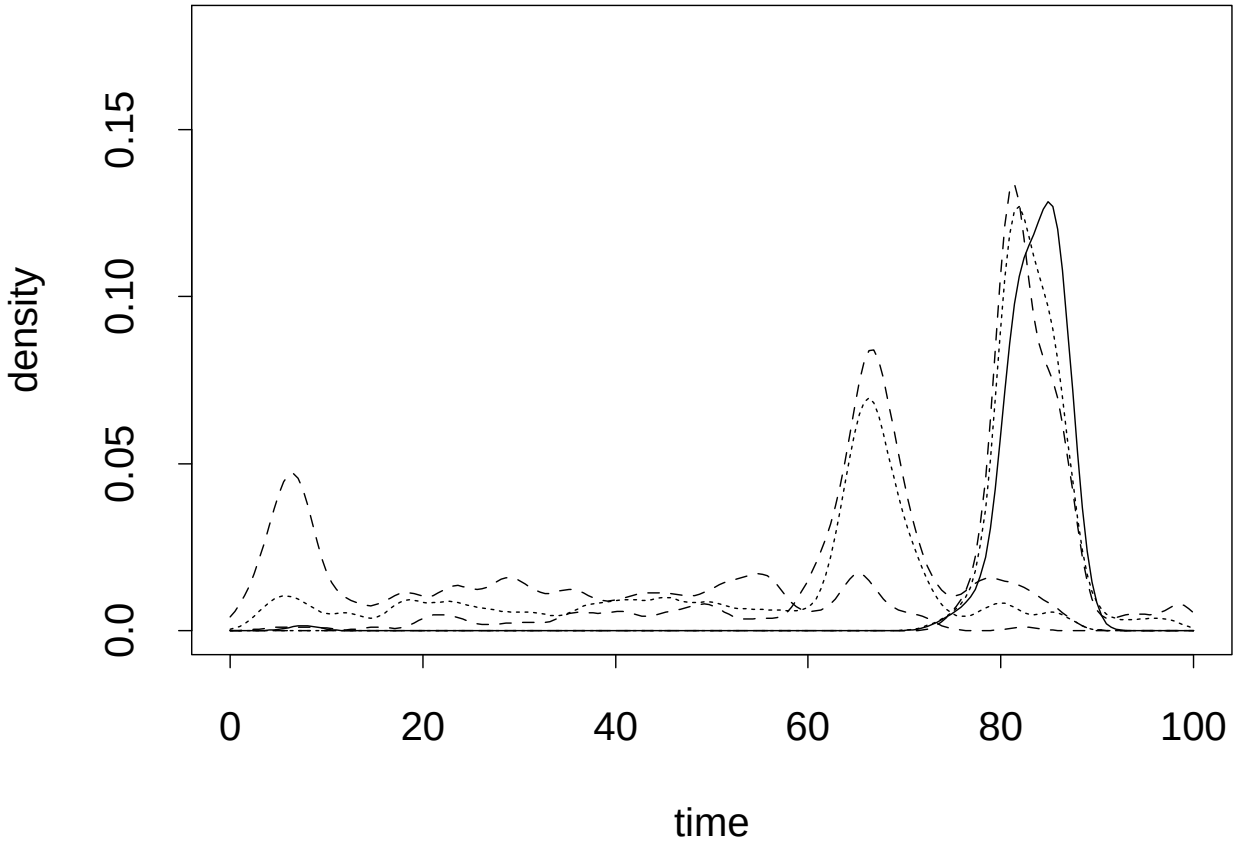
Cyclones hitting the Bay of Bengal

Posterior for the number of change points k: zero change points is ruled out; $k = 1$ or 2 more probable than under the prior.



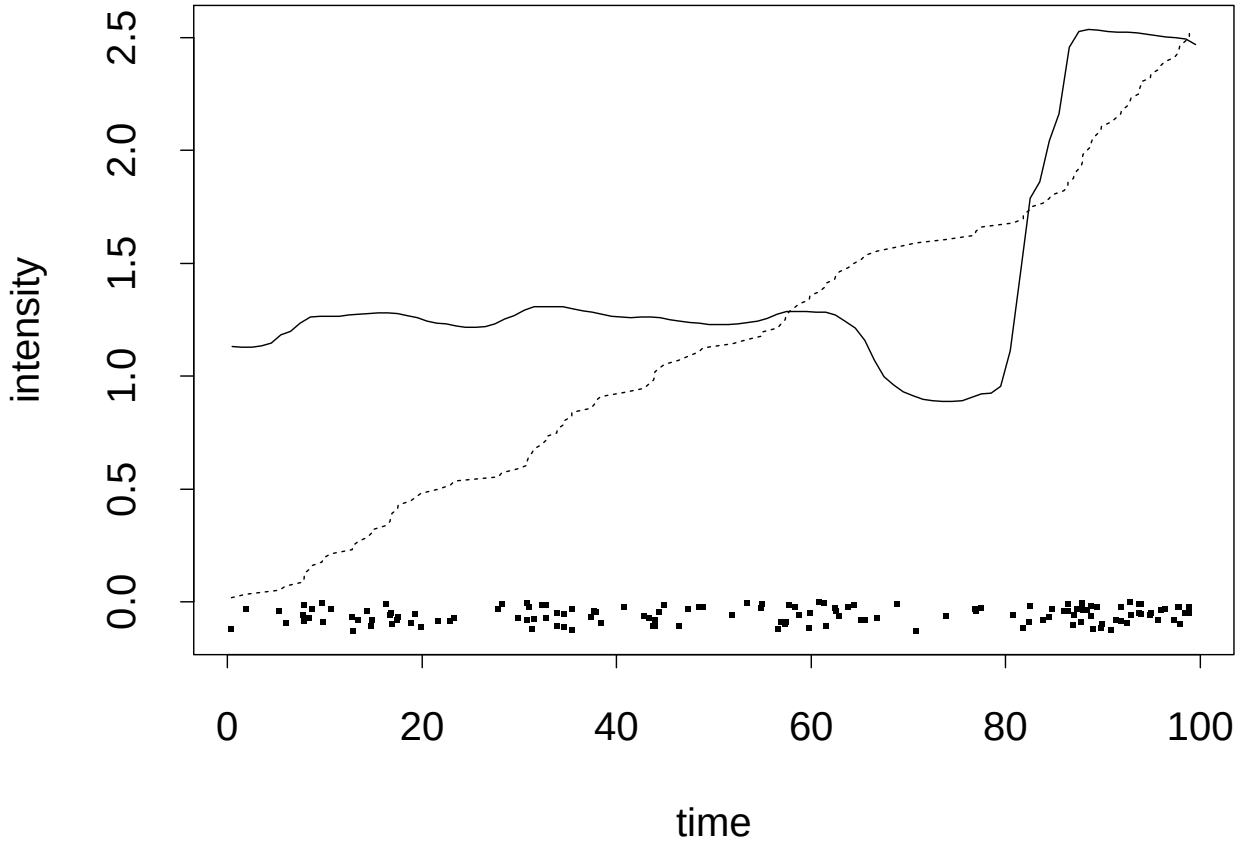
Cyclones hitting the Bay of Bengal

Posterior density estimates for change-point positions



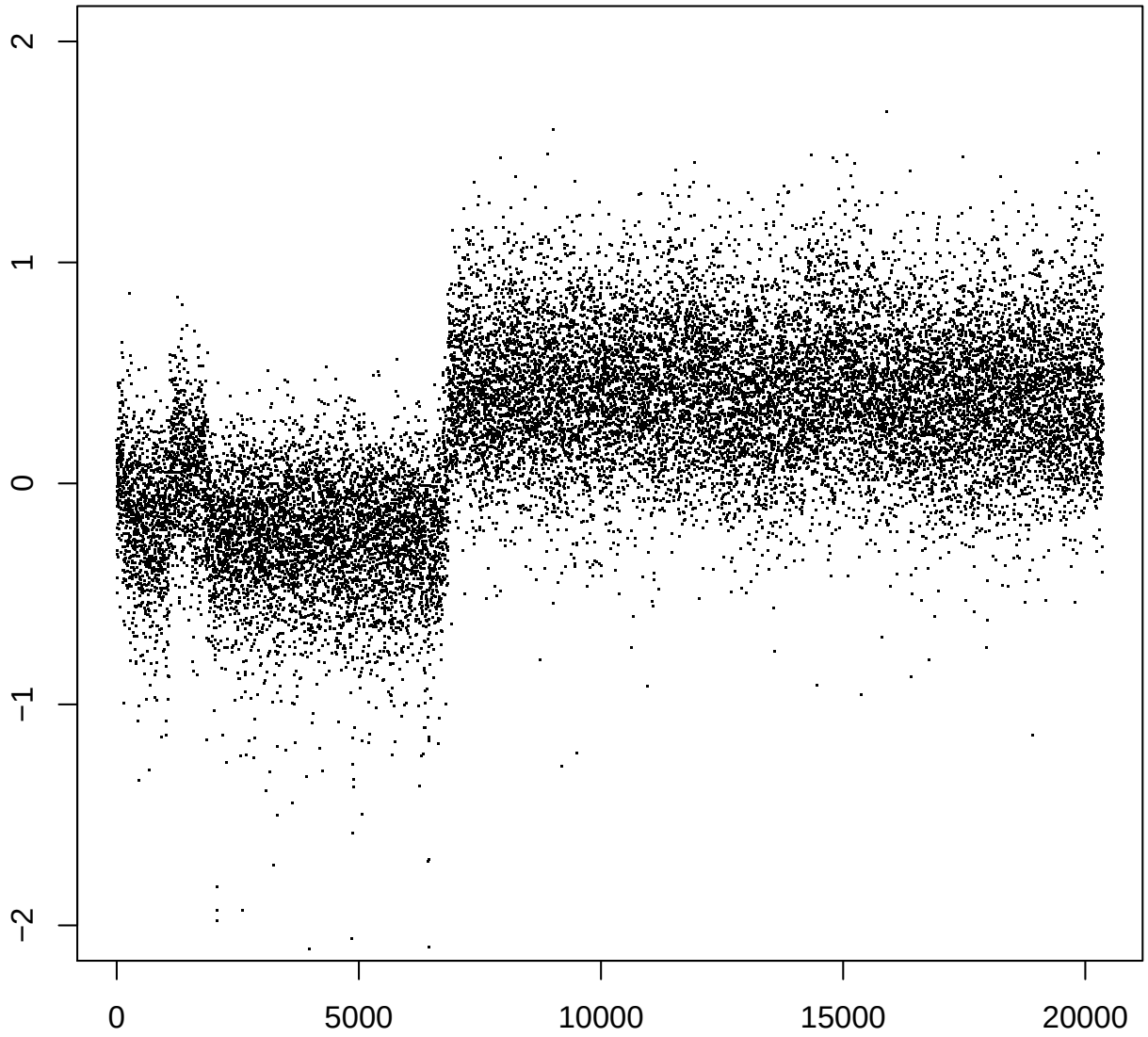
Cyclones hitting the Bay of Bengal

Model-averaged estimate: posterior expectation of intensity (the expectation of a random step function is not a step function!).

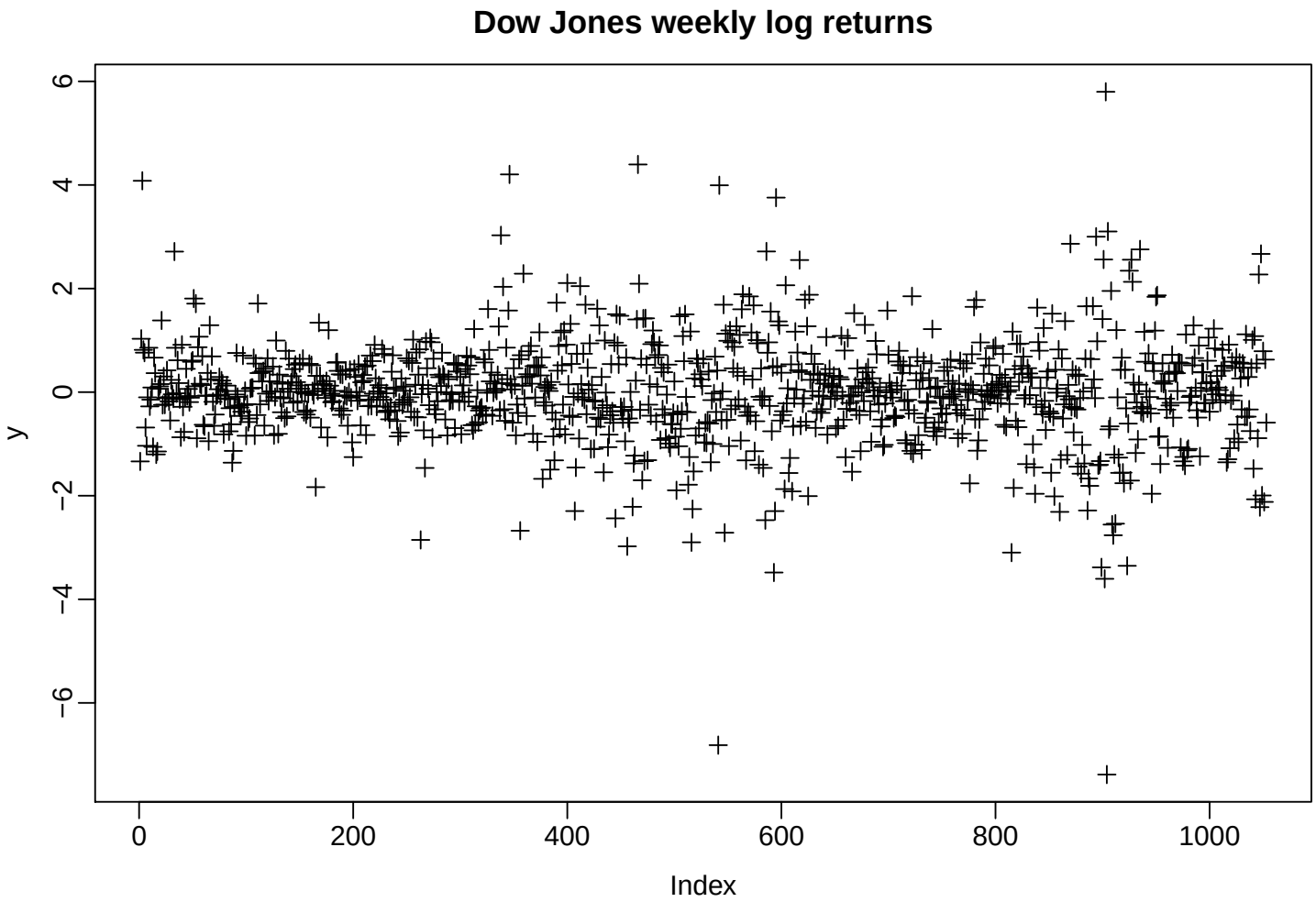


CNV data

Log R ratio trace reflects variation in copy number along the genome: data from Chris Yau, with thanks.



Dow Jones data



Graphical models

The conditional independence graph G of a multivariate distribution (for a random vector X, say) tells us much about the structure of the distribution. $G = (V, E)$ where the vertices V index the components of X, and there is an (undirected) edge between vertices i and j, written $i \sim j$

unless $\quad X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}$

Bayesian graphical model determination

Given n i.i.d. samples $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from a multivariate distribution on $\mathcal{R}^v$ parameterised by the graph G and parameters θ, a typical formulation takes the form

$$p(G, \theta, \mathbf{X}) = p(G)p(\theta|G)p(\mathbf{X}|G, \theta)$$

and we perform joint structural/quantitative learning by computing the posterior $p(G, \theta|\mathbf{X}) \propto p(G, \theta, \mathbf{X})$.

Graphical models

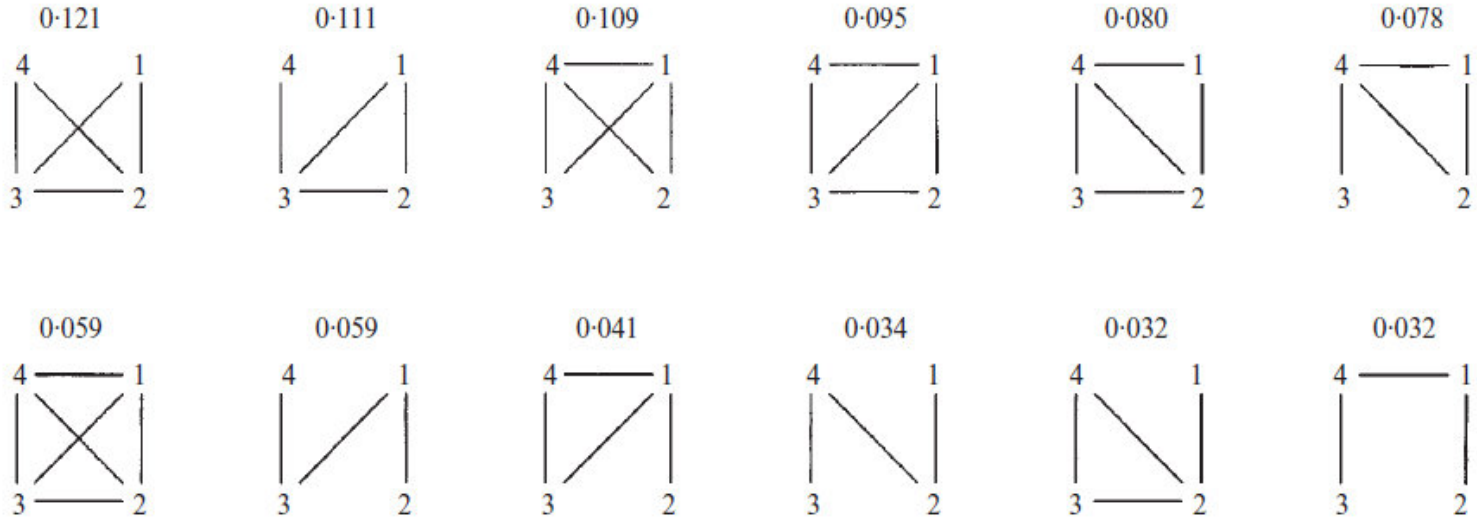


Fig. 2. Most probable graphs for Fret's dataset, together with the associated probabilities.

Frets' head data:
posterior probability on graphs (for highest probability graphs).

Graphical models

Frets' head data: (model-averaged) marginal posteriors on partial correlations.

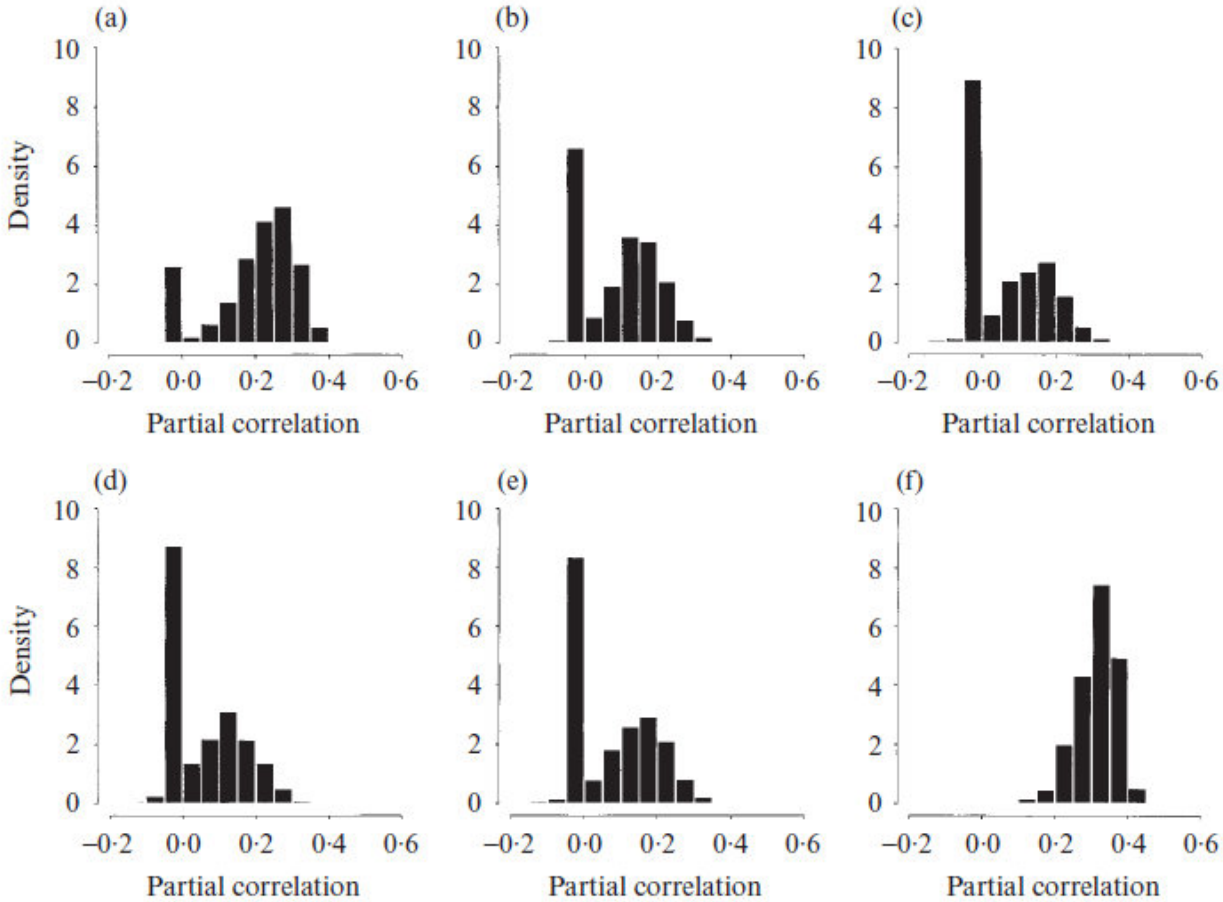
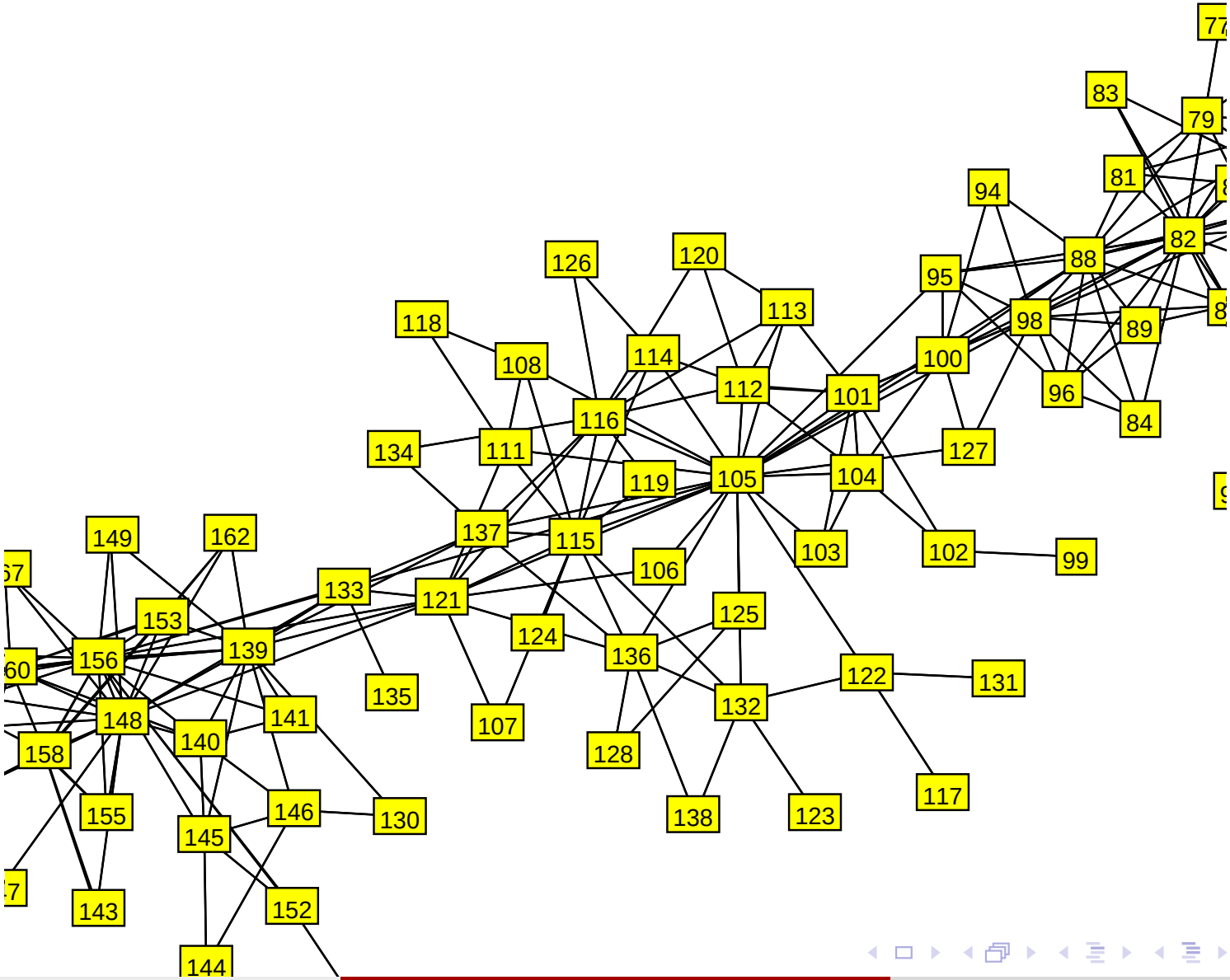


Fig. 3. Posterior distributions of the partial correlation coefficients for Fret's dataset. (a) For variables (1, 2), (b) for (1, 3), (c) for (1, 4), (d) for (2, 3), (e) for (2, 4), (f) for (3, 4).

Linkage Disequilibrium ($p = 500$, $n = 60$)



Bayesian models for mixtures – context 1

$$y_i \sim \sum_{j=1}^{k} w_j f(\cdot|\theta_j) \quad \text{independently for } i = 1, 2, \ldots, n$$

- $f(\cdot|\theta)$ is a given parametric family
- $\{y_i\}$ observed
- $\{\theta_j\}, \{w_j\}, k$ unknown

Heterogeneous population: Groups $j = 1, 2, \ldots, k$, sizes $\propto w_j$.
Observation y_i drawn from unknown group z_i: latent *allocation variable*.

$$p(z_i = j) = w_j \quad \text{independently for } i = 1, 2, \ldots, n$$

$$y_i|z \sim f(\cdot|\theta_{z_i}) \quad \text{independently for } i = 1, 2, \ldots, n$$

Bayesian models for mixtures – context 2

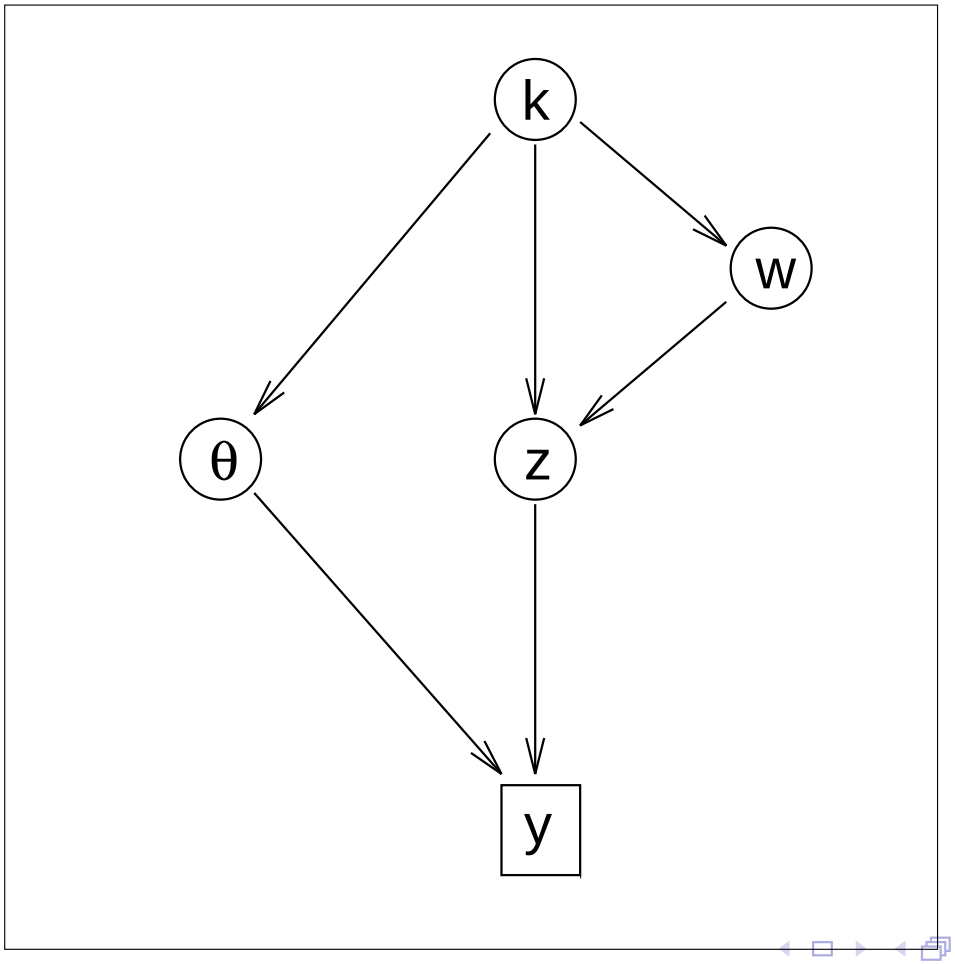
$$y_i \sim \sum_{j=1}^{k} w_j f(\cdot|\theta_j) \quad \text{independently for } i = 1, 2, \ldots, n$$

- $f(\cdot|\theta)$ is a given parametric family
- $\{y_i\}$ observed
- $\{\theta_j\}, \{w_j\}, k$ unknown

Semi-parametric density estimation: (not prime focus here) use same representation, but $\{z_i\}$ now artificial.

Hierarchical model

$$p(k, \theta, w, z, y) = p(k)p(\theta|k)p(w|k)p(z|w, k)p(y|\theta, z)$$



Application of reversible jump MCMC to mixtures

We use two dimension-changing moves:

- splitting/combining components
- birth/death of empty components

(the former is essential, the latter is introduced simply to improve mixing in some rather extreme cases)

Split/combine move, for univariate normal mixtures

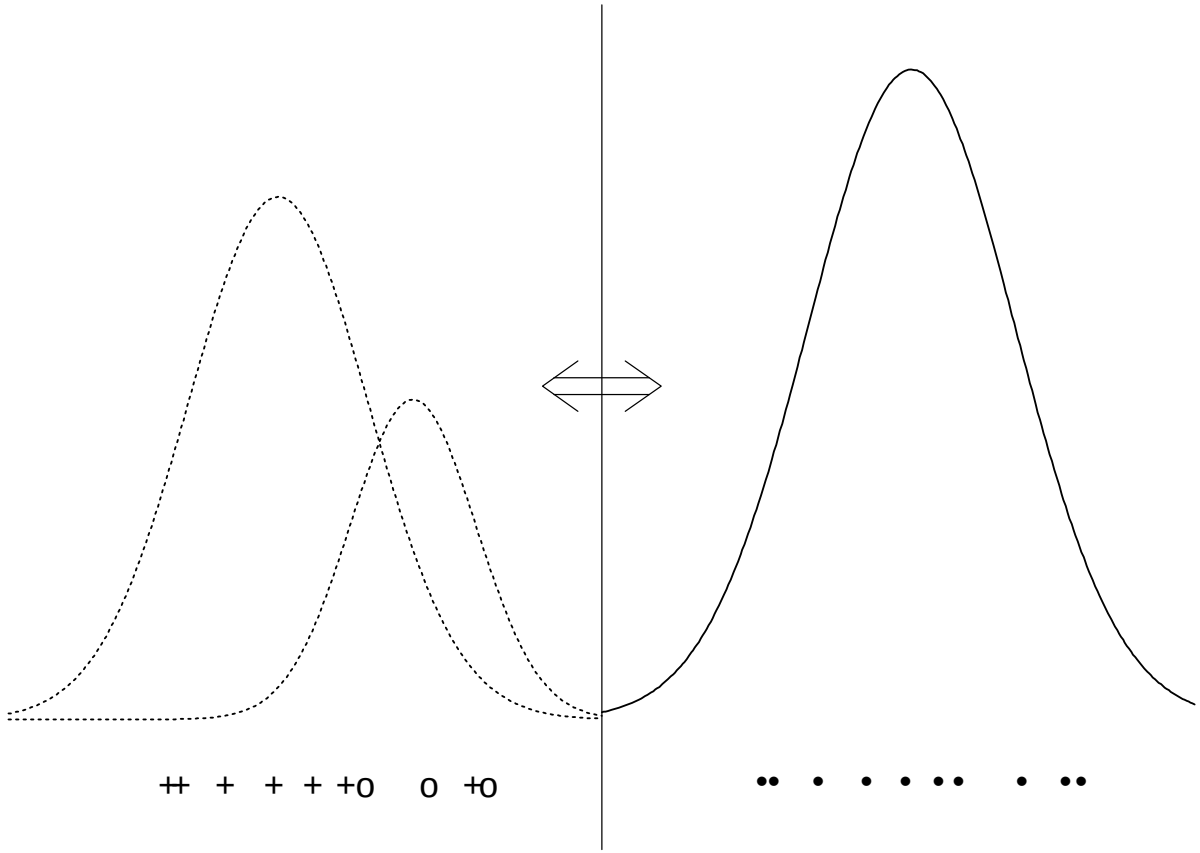
Propose to split a randomly chosen component ($k \to k+1$) or combine two *adjacent* randomly chosen components ($k \to k-1$), *and* reallocate affected observations.

$$(k, w, \mu, \sigma, z) \to (k \pm 1, w', \mu', \sigma', z')$$

Propose a parameter set in the new subspace that is *intuitively* roughly as well supported by the posterior as the old set. We preserve combined weight, mean and variance:

$$\begin{aligned} w_{j*} &= w_{j_1} + w_{j_2} \\ w_{j*}\mu_{j*} &= w_{j_1}\mu_{j_1} + w_{j_2}\mu_{j_2} \\ w_{j*}(\mu_{j*}^2 + \sigma_{j*}^2) &= w_{j_1}(\mu_{j_1}^2 + \sigma_{j_1}^2) + w_{j_2}(\mu_{j_2}^2 + \sigma_{j_2}^2) \end{aligned}$$

Illustration of split/combine proposal



Acceptance probability for split move

For the split move the probability is $\min(1, A)$, where A is

$$
\begin{aligned}
& (\text{likelihood ratio}) \times \frac{p(k+1)}{p(k)} \times (k+1) \\
\times \quad & \frac{w_{j_1}^{\delta-1+l_1} w_{j_2}^{\delta-1+l_2}}{w_{j^*}^{\delta-1+l_1+l_2} B(\delta, k\delta)} \times \sqrt{\frac{\kappa}{2\pi}} \\
\times \quad & \exp\left[-\frac{1}{2}\kappa\{(\mu_{j_1}-\xi)^2 + (\mu_{j_2}-\xi)^2 - (\mu_{j^*}-\xi)^2\}\right] \\
\times \quad & \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{\sigma_{j_1}^2 \sigma_{j_2}^2}{\sigma_{j^*}^2}\right)^{-\alpha-1} \exp\left(-\beta(\sigma_{j_1}^{-2} + \sigma_{j_2}^{-2} - \sigma_{j^*}^{-2})\right) \\
\times \quad & \frac{d_{k+1}}{b_k P_{\text{alloc}}} \times \{g_{2,2}(u_1) g_{2,2}(u_2) g_{1,1}(u_3)\}^{-1} \\
\times \quad & \frac{w_{j^*} |\mu_{j_1} - \mu_{j_2}| \sigma_{j_1}^2 \sigma_{j_2}^2}{u_2(1-u_2^2) u_3(1-u_3) \sigma_{j^*}^2}
\end{aligned}
$$

Posterior distribution of k

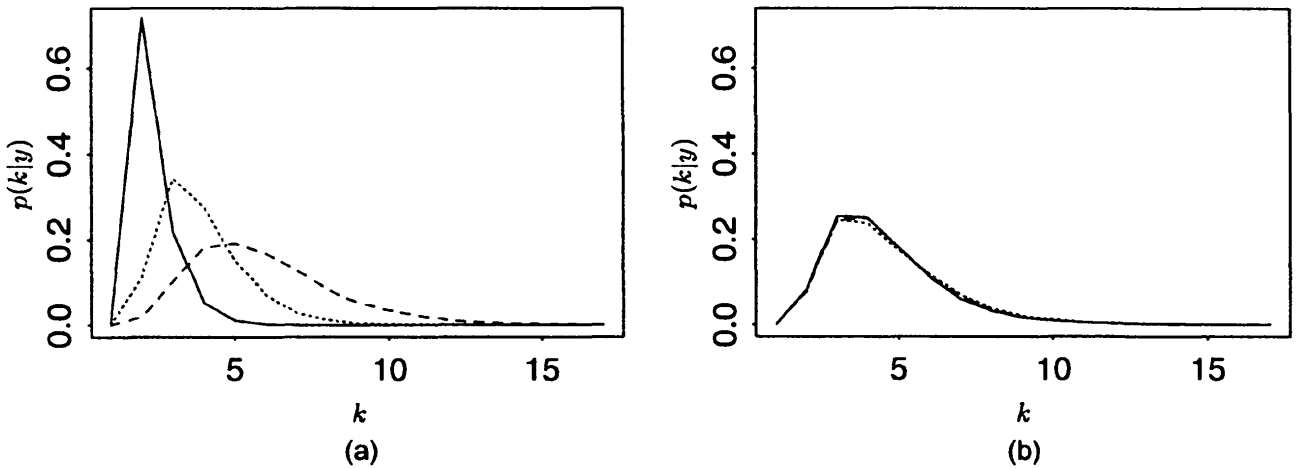


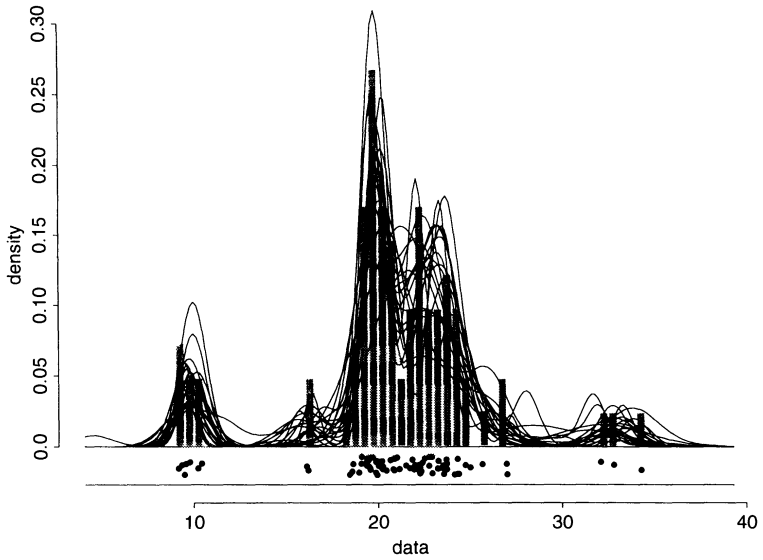
Fig. 6. Posterior distributions of k: comparison of sensitivity to hyperparameters between fixed and random β models: (a) fixed β, $\alpha = 2$ and $\sqrt{(\beta/\alpha)}$ varying between $R/5$ (———), $R/10$ (·····) and $R/20$ (- - - -); (b) random β, $\alpha = 2$, $g = 0.2$ and $\sqrt{(g/h\alpha)}$ varying between $R/5$ (———), $R/10$ (·····) and $R/20$ (- - - -)

Sample from the posterior distribution of the mixture density

Each sweep:

$$(k, w, \theta) \longrightarrow f(\cdot|k, w, \theta) = \sum_{j=1}^{k} w_j f(\cdot|\theta_j)$$

For Galaxy data:



Predictive densities from three data sets

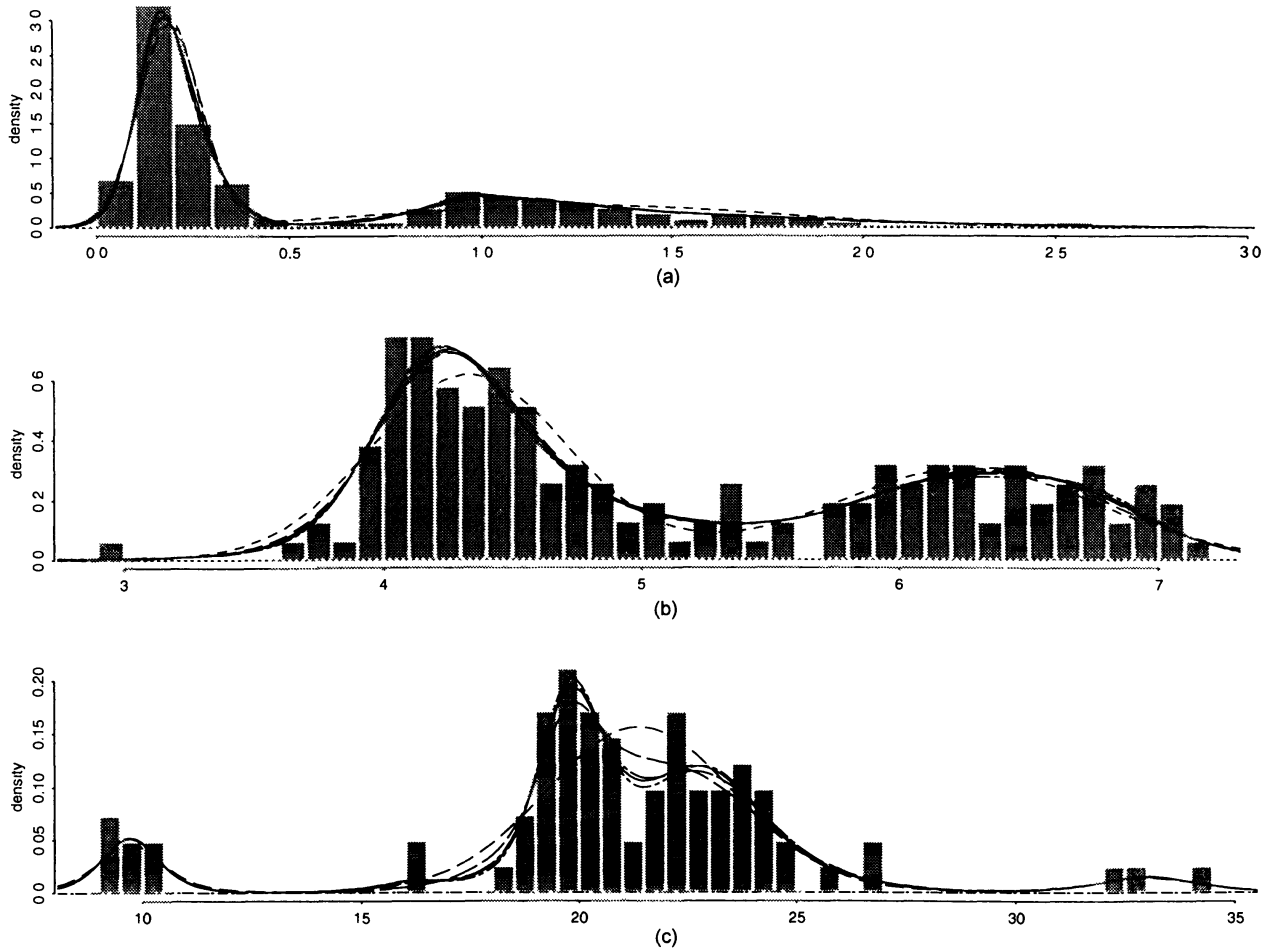


Fig. 2. Predictive densities for (a) the enzyme, (b) the acidity and (c) the galaxy data sets, unconditionally (———) and conditionally (– – – –) on various values of k: the curves displayed are for $k = 2$–6, except for the galaxy data, where they are for $k = 3$–6; in each case note that it is only the smallest k shown that gives an appreciably different estimate

MCMC performance. Jump moves

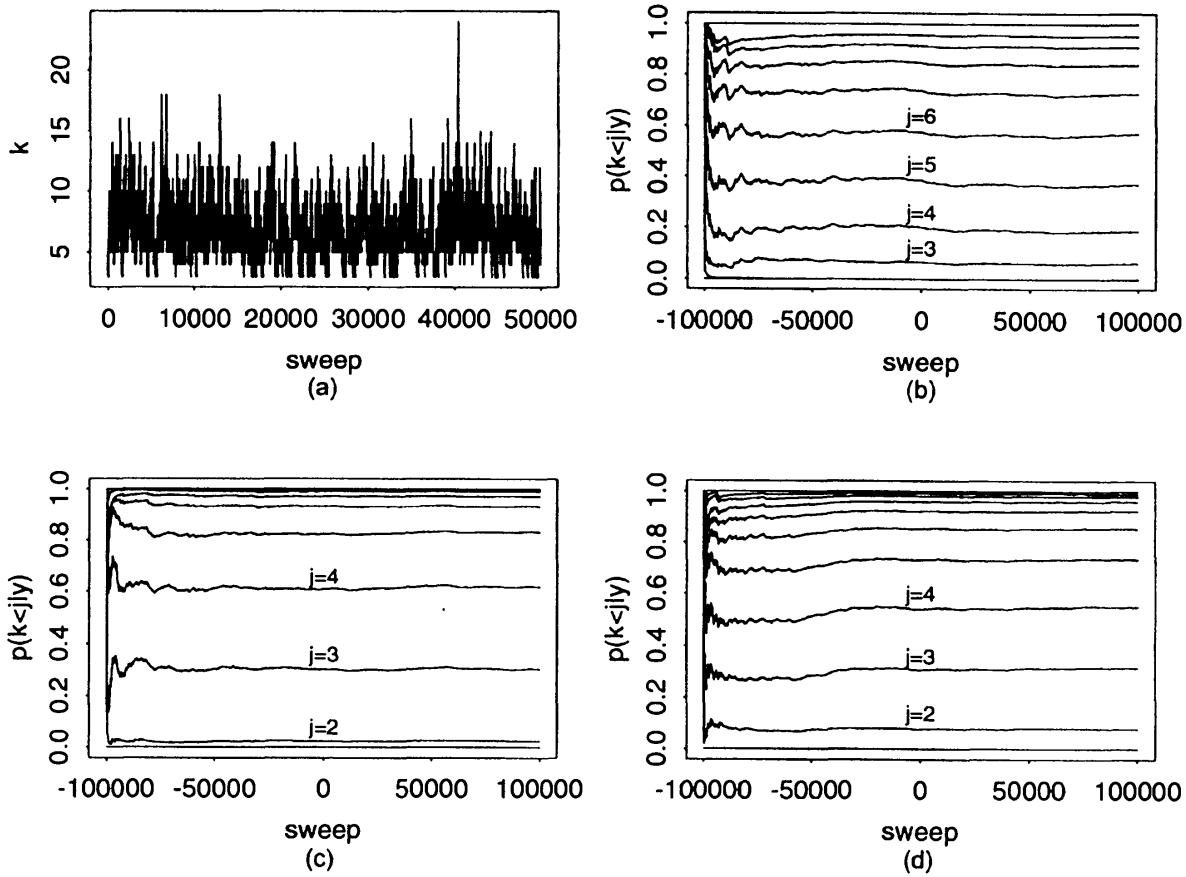


Fig. 7. (a) Example of a trace of k for the galaxy data set, for 50000 sweeps after burn-in, and cumulative occupancy fractions for (b) the galaxy, (c) the enzyme and (d) the acidity data sets, for a complete run including burn-in

MCMC performance. Parameter moves

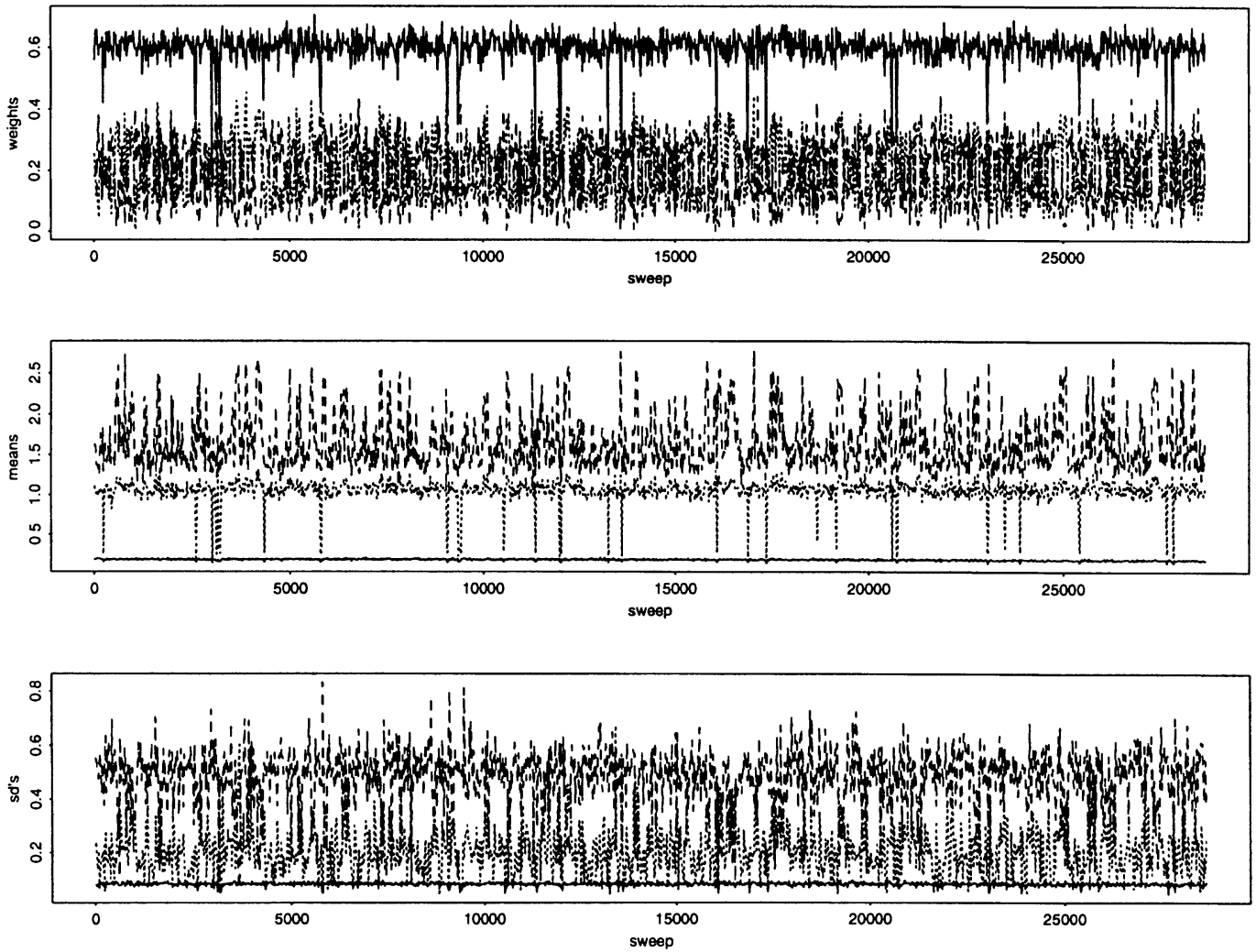
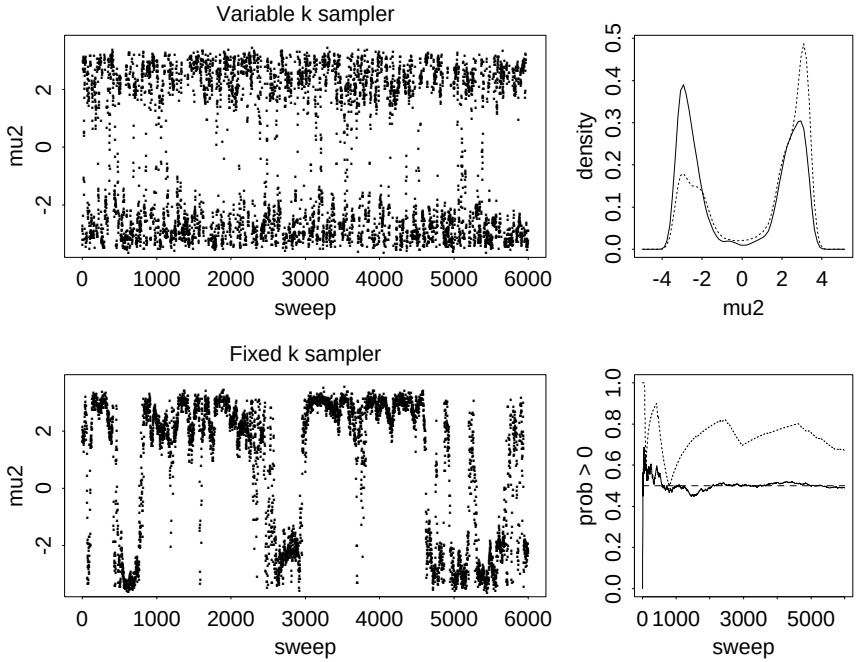


Fig. 8. Traces of parameter estimates against visits to $k = 3$, enzyme data

Effective mixing within k

- no trapping states
- multimodality explored satisfactorily
  $\longrightarrow$ better mixing than a fixed-k sampler in case where posterior has several well separated modes



Extensions

Other univariate distributions

- discrete distributions
- 2 parameter density suitably reparametrised
- adding an overall shape parameter to model mixtures of skewed distributions

Other mixture models

- modelling the weights or allocations (genetic, spatial applications...)
- adapting the algorithm to mixture models based on a Dirichlet process prior: alternative to the incremental sampler

Robust prior modelling in Bayesian analyses use of mixtures with k unknown as a prior model arises in many application contexts, e.g. measurement error problems, random effects, frailty models.

Software

WinBugs: The WinBugs system – respected and very widely used software for Bayesian analysis, and its scope has been recently extended to support fitting of a wide range of trans-dimensional models, including variable selection, automatic curve-fitting using splines, Bayesian MARS and CART, normal mixture analysis, spatial epidemiology clustering models and variable-order Markov chains.

Winbugs example: variable selection


```
model {                                                          #1
  for (i in 1:n) {                                               #2
    Z[i] ~ dnorm(psi[i], tau)                                    #3
  }                                                              #4
  psi[1:n] <- jump.lin.pred(X[1:n, 1:Q], k, beta.prec)           #5
  id <- jump.model.id(psi[1:n])                                  #6
  beta.prec <- tau / lambda                                      #7
  tau ~ dgamma(a, b)                                             #8
  k ~ dbin(0.5, Q)                                               #9
}                                                                #10
```

In reality, it's not that easy...

- Where does our set of candidate models come from?
  We cannot say anything probabilistic about models that are not even considered.

In reality, it's not that easy...

- Where does our set of candidate models come from?
  We cannot say anything probabilistic about models that are not even considered.

In reality, it's not that easy...

- Prior model probabilities may be fictional
  The ideal Bayesian (or her/his scientist colleague) has real prior probabilities reflecting scientific judgement/belief across the model space; not very common in practice!

  Arbitrariness in prior model probabilities may not affect Bayes factors, but it sabotages Bayesian model averaging!

In reality, it's not that easy...

- Prior model probabilities may be fictional
  The ideal Bayesian (or her/his scientist colleague) has real prior probabilities reflecting scientific judgement/belief across the model space; not very common in practice!

  Arbitrariness in prior model probabilities may not affect Bayes factors, but it sabotages Bayesian model averaging!

In reality, it's not that easy...

- Prior model probabilities may be fictional
  The ideal Bayesian (or her/his scientist colleague) has real prior probabilities reflecting scientific judgement/belief across the model space; not very common in practice!

  Arbitrariness in prior model probabilities may not affect Bayes factors, but it sabotages Bayesian model averaging!

In reality, it's not that easy...

- **No chance of passing the test of a sensitivity analysis**
  In ordinary parametric problems we commonly find that inferences are rather insensitive to moderately large variations in prior assumptions, except when there are very few data (indeed, the opposite case, of high sensitivity, poses a challenge to the non-Bayesian – perhaps the data carry less information than hoped?). But it's clear that a test of sensitivity to model probabilities will always fail:

$$\frac{p^\star(k|Y)}{p^\star(l|Y)} = \frac{p(k|Y)}{p(l|Y)} \times \left(\frac{p^\star(k)}{p^\star(l)} \div \frac{p(k)}{p(l)} \right)$$

In reality, it's not that easy...

- No chance of passing the test of a sensitivity analysis
  In ordinary parametric problems we commonly find that inferences are rather insensitive to moderately large variations in prior assumptions, except when there are very few data (indeed, the opposite case, of high sensitivity, poses a challenge to the non-Bayesian – perhaps the data carry less information than hoped?). But it's clear that a test of sensitivity to model probabilities will always fail:

$$\frac{p^\star(k|Y)}{p^\star(l|Y)} = \frac{p(k|Y)}{p(l|Y)} \times \left(\frac{p^\star(k)}{p^\star(l)} \div \frac{p(k)}{p(l)} \right)$$

In reality, it's not that easy...

- No chance of passing the test of a sensitivity analysis
  In ordinary parametric problems we commonly find that inferences are rather insensitive to moderately large variations in prior assumptions, except when there are very few data (indeed, the opposite case, of high sensitivity, poses a challenge to the non-Bayesian – perhaps the data carry less information than hoped?). But it's clear that a test of sensitivity to model probabilities will always fail:

$$\frac{p^\star(k|Y)}{p^\star(l|Y)} = \frac{p(k|Y)}{p(l|Y)} \times \left(\frac{p^\star(k)}{p^\star(l)} \div \frac{p(k)}{p(l)} \right)$$

In reality, it's not that easy...

- Improper parameter priors problems
  In ordinary parametric problems it is commonly true that it is safe to use improper priors – when posterior distributions are well-defined as limits based on a sequence of approximating proper priors (and not usually sensitive to what that sequence is).

  But improper parameter priors make Bayes factors indeterminate (since improper priors can only be defined up to arbitrary normalising constants, which persist into marginal likelihoods).

  And proper but vague/diffuse priors fail to solve the problem, since the Bayes factors will depend on the arbitrary degree of vagueness used.

In reality, it's not that easy...

- Improper parameter priors problems
  In ordinary parametric problems it is commonly true that it is safe to use improper priors – when posterior distributions are well-defined as limits based on a sequence of approximating proper priors (and not usually sensitive to what that sequence is).

  But improper parameter priors make Bayes factors indeterminate (since improper priors can only be defined up to arbitrary normalising constants, which persist into marginal likelihoods).

  And proper but vague/diffuse priors fail to solve the problem, since the Bayes factors will depend on the arbitrary degree of vagueness used.

In reality, it's not that easy...

- Improper parameter priors problems
  In ordinary parametric problems it is commonly true that it is safe to use improper priors – when posterior distributions are well-defined as limits based on a sequence of approximating proper priors (and not usually sensitive to what that sequence is).

  But improper parameter priors make Bayes factors indeterminate (since improper priors can only be defined up to arbitrary normalising constants, which persist into marginal likelihoods).

  And proper but vague/diffuse priors fail to solve the problem, since the Bayes factors will depend on the arbitrary degree of vagueness used.

In reality, it's not that easy...

- Improper parameter priors problems
  In ordinary parametric problems it is commonly true that it is safe to use improper priors – when posterior distributions are well-defined as limits based on a sequence of approximating proper priors (and not usually sensitive to what that sequence is).

  But improper parameter priors make Bayes factors indeterminate (since improper priors can only be defined up to arbitrary normalising constants, which persist into marginal likelihoods).

  And proper but vague/diffuse priors fail to solve the problem, since the Bayes factors will depend on the arbitrary degree of vagueness used.

In reality, it's not that easy...

- Improper parameter priors problems, continued

  In certain circumstances, ideas such as Intrinsic or Fractional Bayes factors, or Expected Posterior priors, can be applied, essentially based on tying together improper priors across different models. These ideas lose much of the appeal of ideal Bayes arguments, have arbitrary aspects, and are not widely accepted.

In reality, it's not that easy...

- Improper parameter priors problems, continued

  In certain circumstances, ideas such as Intrinsic or Fractional Bayes factors, or Expected Posterior priors, can be applied, essentially based on tying together improper priors across different models. These ideas lose much of the appeal of ideal Bayes arguments, have arbitrary aspects, and are not widely accepted.

Model uncertainty? Yes, but do we have to choose?

Model uncertainty is a fact of life.

- When can we quantify it?
- When can we eliminate it?
- When can we accommodate it?

Why choose?

- Prediction
- Scientific understanding
- Presentation
- Policy
- Defence

An illusion of unity?

Is ‘model’ too much of a catch-all?

- different scientific mechanisms
- selection of predictors in regression
- number of components in mixture
- order of AR model
- complexity of polynomial or spline

All the other criteria

- Bayesian hypothesis testing, and 'alternative-prior carpentry'
- AIC, BIC, DIC, DIC+, MDL, C_p
- Decision theory
- Bayesian p-values
- Posterior predictive checks

Acknowledgements

Thanks to all my collaborators on trans-dimensional MCMC problems: Carmen Fernández, Paolo Giudici, Miles Harkness, David Hastie, Matthew Hodgson, Antonietta Mira, Agostino Nobile, Marco Pievatolo, Sylvia Richardson, Luisa Scaccia and Claudia Tarantola.

References and preprints available from

http://www.stats.bris.ac.uk/∼peter/Research.html
http://www.stats.bris.ac.uk/∼peter/papers/HastieGreenR1.pdf,
Statistica Neerlandica, 66, 309-338.
doi:10.1111/j.0039-0402.2011.00516.x
P.J.Green@bristol.ac.uk