On some computational methods for Bayesian model choice

Christian P. Robert

CREST-INSEE and Université Paris Dauphine
http://www.ceremade.dauphine.fr/~xian

Warwick EPSRC Symposium 2008/2009
Joint works with M. Beaumont, N. Chopin, J.-M. Cornuet, and J.-M. Marin

Outline

Construction of Bayes tests

Definition (Test)

Given an hypothesis $H_0 : \theta \in \Theta_0$ on the parameter $\theta \in \Theta_0$ of a statistical model, a **test** is a statistical procedure that takes its values in $\{0, 1\}$.

Theorem (Bayes test)

The Bayes estimator associated with π and with the $0-1$ loss is

$$\delta^\pi(x) = \begin{cases} 1 & \text{if } \pi(\theta \in \Theta_0|x) > \pi(\theta \notin \Theta_0|x), \\ 0 & \text{otherwise,} \end{cases}$$

Construction of Bayes tests

Definition (Test)

Given an hypothesis $H_0 : \theta \in \Theta_0$ on the parameter $\theta \in \Theta_0$ of a statistical model, a **test** is a statistical procedure that takes its values in $\{0, 1\}$.

Theorem (Bayes test)

The Bayes estimator associated with π and with the $0-1$ loss is

$$\delta^\pi(x) = \begin{cases} 1 & \text{if } \pi(\theta \in \Theta_0|x) > \pi(\theta \notin \Theta_0|x), \\ 0 & \text{otherwise,} \end{cases}$$

Bayes factor

Definition (Bayes factors)

For testing hypotheses $H_0 : \theta \in \Theta_0$ vs. $H_{\mathsf{a}} : \theta \notin \Theta_0$, under prior

$$\pi(\Theta_0)\pi_0(\theta) + \pi(\Theta_0^c)\pi_1(\theta)\,,$$

central quantity

$$B_{01} = \frac{\pi(\Theta_0|x)}{\pi(\Theta_0^c|x)} \Big/ \frac{\pi(\Theta_0)}{\pi(\Theta_0^c)} = \frac{\displaystyle\int_{\Theta_0} f(x|\theta)\pi_0(\theta)\mathrm{d}\theta}{\displaystyle\int_{\Theta_0^c} f(x|\theta)\pi_1(\theta)\mathrm{d}\theta}$$

[Jeffreys, 1939]

Self-contained concept

Outside decision-theoretic environment:

- eliminates impact of $\pi(\Theta_0)$ but depends on the choice of (π_0, π_1)
- Bayesian/marginal equivalent to the likelihood ratio
- Jeffreys' scale of evidence:
  - if $\log_{10}(B^\pi_{10})$ between 0 and 0.5, evidence against H_0 *weak*,
  - if $\log_{10}(B^\pi_{10})$ 0.5 and 1, evidence *substantial*,
  - if $\log_{10}(B^\pi_{10})$ 1 and 2, evidence *strong* and
  - if $\log_{10}(B^\pi_{10})$ above 2, evidence *decisive*
- Requires the computation of the marginal/evidence under both hypotheses/models

Model choice and model comparison

Choice between models

Several models available for the same observation

$$\mathfrak{M}_i : x \sim f_i(x|\theta_i), \qquad i \in \mathfrak{I}$$

where $\mathfrak{I}$ can be finite or infinite

Replace hypotheses with models but keep marginal likelihoods and Bayes factors

Model choice and model comparison

Choice between models

Several models available for the same observation

$$\mathfrak{M}_i : x \sim f_i(x|\theta_i), \qquad i \in \mathfrak{I}$$

where $\mathfrak{I}$ can be finite or infinite

Replace hypotheses with models but keep marginal likelihoods and Bayes factors

Bayesian model choice

Probabilise the entire model/parameter space

- allocate probabilities p_i to all models $\mathfrak{M}_i$
- define priors $\pi_i(\theta_i)$ for each parameter space Θ_i
- compute

$$\pi(\mathfrak{M}_i|x) = \frac{p_i \int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)\mathrm{d}\theta_i}{\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)\mathrm{d}\theta_j}$$

- take largest $\pi(\mathfrak{M}_i|x)$ to determine "best" model, or use averaged predictive

$$\sum_j \pi(\mathfrak{M}_j|x) \int_{\Theta_j} f_j(x'|\theta_j)\pi_j(\theta_j|x)\mathrm{d}\theta_j$$

Bayesian model choice

Probabilise the entire model/parameter space

- allocate probabilities p_i to all models $\mathfrak{M}_i$
- define priors $\pi_i(\theta_i)$ for each parameter space Θ_i
- compute

$$\pi(\mathfrak{M}_i|x) = \frac{p_i \int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)\mathrm{d}\theta_i}{\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)\mathrm{d}\theta_j}$$

- take largest $\pi(\mathfrak{M}_i|x)$ to determine "best" model, or use averaged predictive

$$\sum_j \pi(\mathfrak{M}_j|x) \int_{\Theta_j} f_j(x'|\theta_j)\pi_j(\theta_j|x)\mathrm{d}\theta_j$$

Bayesian model choice

Probabilise the entire model/parameter space

- allocate probabilities p_i to all models $\mathfrak{M}_i$
- define priors $\pi_i(\theta_i)$ for each parameter space Θ_i
- compute

$$\pi(\mathfrak{M}_i|x) = \frac{p_i \int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)\mathrm{d}\theta_i}{\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)\mathrm{d}\theta_j}$$

- take largest $\pi(\mathfrak{M}_i|x)$ to determine "best" model, or use averaged predictive

$$\sum_j \pi(\mathfrak{M}_j|x) \int_{\Theta_j} f_j(x'|\theta_j)\pi_j(\theta_j|x)\mathrm{d}\theta_j$$

Bayesian model choice

Probabilise the entire model/parameter space

- allocate probabilities p_i to all models $\mathfrak{M}_i$
- define priors $\pi_i(\theta_i)$ for each parameter space Θ_i
- compute

$$\pi(\mathfrak{M}_i|x) = \frac{p_i \int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)\mathrm{d}\theta_i}{\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)\mathrm{d}\theta_j}$$

- take largest $\pi(\mathfrak{M}_i|x)$ to determine "best" model, or use averaged predictive

$$\sum_j \pi(\mathfrak{M}_j|x) \int_{\Theta_j} f_j(x'|\theta_j)\pi_j(\theta_j|x)\mathrm{d}\theta_j$$

Evidence

All these problems end up with a similar quantity, the *evidence*

$$\mathfrak{Z}_k = \int_{\Theta_k} \pi_k(\theta_k) L_k(\theta_k) \, \mathrm{d}\theta_k,$$

aka the marginal likelihood.

Bayes factor approximation

When approximating the Bayes factor

$$B_{01} = \frac{\int_{\Theta_0} f_0(x|\theta_0)\pi_0(\theta_0)\mathrm{d}\theta_0}{\int_{\Theta_1} f_1(x|\theta_1)\pi_1(\theta_1)\mathrm{d}\theta_1}$$

use of importance functions ϖ_0 and ϖ_1 and

$$\widehat{B}_{01} = \frac{n_0^{-1}\sum_{i=1}^{n_0} f_0(x|\theta_0^i)\pi_0(\theta_0^i)/\varpi_0(\theta_0^i)}{n_1^{-1}\sum_{i=1}^{n_1} f_1(x|\theta_1^i)\pi_1(\theta_1^i)/\varpi_1(\theta_1^i)}$$

Bridge sampling

Special case:
If

$$\begin{array}{rcl} \pi_1(\theta_1|x) & \propto & \tilde{\pi}_1(\theta_1|x) \\ \pi_2(\theta_2|x) & \propto & \tilde{\pi}_2(\theta_2|x) \end{array}$$

live on the same space ($\Theta_1 = \Theta_2$), then

$$B_{12} \approx \frac{1}{n} \sum_{i=1}^{n} \frac{\tilde{\pi}_1(\theta_i|x)}{\tilde{\pi}_2(\theta_i|x)} \qquad \theta_i \sim \pi_2(\theta|x)$$

[Gelman & Meng, 1998; Chen, Shao & Ibrahim, 2000]

Bridge sampling variance

The bridge sampling estimator does poorly if

$$\frac{\text{var}(\widehat{B}_{12})}{B_{12}^2} \approx \frac{1}{n}\,\mathbb{E}\left[\left(\frac{\pi_1(\theta)-\pi_2(\theta)}{\pi_2(\theta)}\right)^2\right]$$

is large, i.e. if π_1 and π_2 have little overlap...

Bridge sampling variance

The bridge sampling estimator does poorly if

$$\frac{\text{var}(\widehat{B}_{12})}{B_{12}^2} \approx \frac{1}{n}\,\mathbb{E}\left[\left(\frac{\pi_1(\theta)-\pi_2(\theta)}{\pi_2(\theta)}\right)^2\right]$$

is large, i.e. if π_1 and π_2 have little overlap...

(Further) bridge sampling

In addition

$$B_{12} = \frac{\int \tilde{\pi}_2(\theta|x)\alpha(\theta)\pi_1(\theta|x)d\theta}{\int \tilde{\pi}_1(\theta|x)\alpha(\theta)\pi_2(\theta|x)d\theta} \qquad \forall\ \alpha(\cdot)$$

$$\approx \frac{\frac{1}{n_1}\sum_{i=1}^{n_1} \tilde{\pi}_2(\theta_{1i}|x)\alpha(\theta_{1i})}{\frac{1}{n_2}\sum_{i=1}^{n_2} \tilde{\pi}_1(\theta_{2i}|x)\alpha(\theta_{2i})} \qquad \theta_{ji} \sim \pi_j(\theta|x)$$

An infamous example

When

$$\alpha(\theta) = \frac{1}{\tilde{\pi}_1(\theta)\tilde{\pi}_2(\theta)}$$

harmonic mean approximation to B_{12}

$$\widehat{B_{21}} = \frac{\frac{1}{n_1}\sum_{i=1}^{n_1} 1/\tilde{\pi}_1(\theta_{1i}|x)}{\frac{1}{n_2}\sum_{i=1}^{n_2} 1/\tilde{\pi}_2(\theta_{2i}|x)} \qquad \theta_{ji} \sim \pi_j(\theta|x)$$

[Newton & Raftery, 1994]

Infamous: Most often leads to an infinite variance!!!

[Radford Neal's blog, 2008]

An infamous example

When

$$\alpha(\theta) = \frac{1}{\tilde{\pi}_1(\theta)\tilde{\pi}_2(\theta)}$$

harmonic mean approximation to B_{12}

$$\widehat{B_{21}} = \frac{\frac{1}{n_1}\sum_{i=1}^{n_1} 1/\tilde{\pi}_1(\theta_{1i}|x)}{\frac{1}{n_2}\sum_{i=1}^{n_2} 1/\tilde{\pi}_2(\theta_{2i}|x)} \qquad \theta_{ji} \sim \pi_j(\theta|x)$$

[Newton & Raftery, 1994]

Infamous: Most often leads to an infinite variance!!!

[Radford Neal's blog, 2008]

"The Worst Monte Carlo Method Ever"

"The good news is that the Law of Large Numbers guarantees that this estimator is consistent ie, it will very likely be very close to the correct answer if you use a sufficiently large number of points from the posterior distribution.

The bad news is that the number of points required for this estimator to get close to the right answer will often be greater than the number of atoms in the observable universe. The even worse news is that itws easy for people to not realize this, and to naively accept estimates that are nowhere close to the correct value of the marginal likelihood."

[Radford Neal's blog, Aug. 23, 2008]

"The Worst Monte Carlo Method Ever"

"The good news is that the Law of Large Numbers guarantees that this estimator is consistent ie, it will very likely be very close to the correct answer if you use a sufficiently large number of points from the posterior distribution.
The bad news is that the number of points required for this estimator to get close to the right answer will often be greater than the number of atoms in the observable universe. The even worse news is that itws easy for people to not realize this, and to naively accept estimates that are nowhere close to the correct value of the marginal likelihood."

[Radford Neal's blog, Aug. 23, 2008]

Optimal bridge sampling

The optimal choice of auxiliary function is

$$\alpha^\star = \frac{n_1 + n_2}{n_1\pi_1(\theta|x) + n_2\pi_2(\theta|x)}$$

leading to

$$\widehat{B}_{12} \approx \frac{\dfrac{1}{n_1}\displaystyle\sum_{i=1}^{n_1} \dfrac{\tilde{\pi}_2(\theta_{1i}|x)}{n_1\pi_1(\theta_{1i}|x) + n_2\pi_2(\theta_{1i}|x)}}{\dfrac{1}{n_2}\displaystyle\sum_{i=1}^{n_2} \dfrac{\tilde{\pi}_1(\theta_{2i}|x)}{n_1\pi_1(\theta_{2i}|x) + n_2\pi_2(\theta_{2i}|x)}}$$

Back later!

Optimal bridge sampling (2)

Reason:

$$\frac{\mathrm{Var}(\widehat{B}_{12})}{B_{12}^2} \approx \frac{1}{n_1 n_2} \left\{ \frac{\int \pi_1(\theta)\pi_2(\theta)[n_1\pi_1(\theta)+n_2\pi_2(\theta)]\alpha(\theta)^2\,\mathrm{d}\theta}{\left(\int \pi_1(\theta)\pi_2(\theta)\alpha(\theta)\,\mathrm{d}\theta\right)^2} - 1 \right\}$$

(by the δ method)

Drag: Dependence on the unknown normalising constants solved iteratively

Optimal bridge sampling (2)

Reason:

$$\frac{\mathrm{Var}(\widehat{B}_{12})}{B_{12}^2} \approx \frac{1}{n_1 n_2} \left\{ \frac{\int \pi_1(\theta)\pi_2(\theta)[n_1\pi_1(\theta) + n_2\pi_2(\theta)]\alpha(\theta)^2 \,\mathrm{d}\theta}{\left(\int \pi_1(\theta)\pi_2(\theta)\alpha(\theta)\,\mathrm{d}\theta\right)^2} - 1 \right\}$$

(by the δ method)
Drag: Dependence on the unknown normalising constants solved iteratively

Ratio importance sampling

Another identity:

$$B_{12} = \frac{\mathbb{E}_\varphi \left[\tilde{\pi}_1(\theta)/\varphi(\theta)\right]}{\mathbb{E}_\varphi \left[\tilde{\pi}_2(\theta)/\varphi(\theta)\right]}$$

for any density φ with sufficiently large support

[Torrie & Valleau, 1977]

Use of a single sample $\theta_1, \ldots, \theta_n$ from φ

$$\widehat{B}_{12} = \frac{\sum_{i=1} \tilde{\pi}_1(\theta_i)/\varphi(\theta_i)}{\sum_{i=1} \tilde{\pi}_2(\theta_i)/\varphi(\theta_i)}$$

Ratio importance sampling

Another identity:

$$B_{12} = \frac{\mathbb{E}_\varphi \left[\tilde{\pi}_1(\theta)/\varphi(\theta)\right]}{\mathbb{E}_\varphi \left[\tilde{\pi}_2(\theta)/\varphi(\theta)\right]}$$

for any density φ with sufficiently large support

[Torrie & Valleau, 1977]

Use of a single sample $\theta_1, \ldots, \theta_n$ from φ

$$\widehat{B}_{12} = \frac{\sum_{i=1} \tilde{\pi}_1(\theta_i)/\varphi(\theta_i)}{\sum_{i=1} \tilde{\pi}_2(\theta_i)/\varphi(\theta_i)}$$

Ratio importance sampling (2)

Approximate variance:

$$\frac{\mathsf{var}(\widehat{B}_{12})}{B_{12}^2} \approx \frac{1}{n}\,\mathbb{E}_\varphi\left[\left(\frac{(\pi_1(\theta)-\pi_2(\theta))^2}{\varphi(\theta)^2}\right)^2\right]$$

Optimal choice:

$$\varphi^*(\theta) = \frac{|\,\pi_1(\theta)-\pi_2(\theta)\,|}{\int |\,\pi_1(\eta)-\pi_2(\eta)\,|\;\mathrm{d}\eta}$$

[Chen, Shao & Ibrahim, 2000]

Ratio importance sampling (2)

Approximate variance:

$$\frac{\mathsf{var}(\widehat{B}_{12})}{B_{12}^2} \approx \frac{1}{n}\,\mathbb{E}_\varphi\left[\left(\frac{(\pi_1(\theta)-\pi_2(\theta))^2}{\varphi(\theta)^2}\right)^2\right]$$

Optimal choice:

$$\varphi^*(\theta) = \frac{|\,\pi_1(\theta)-\pi_2(\theta)\,|}{\int |\,\pi_1(\eta)-\pi_2(\eta)\,|\;\mathsf{d}\eta}$$

[Chen, Shao & Ibrahim, 2000]

Improving upon bridge sampler

Theorem 5.5.3: The asymptotic variance of the optimal ratio importance sampling estimator is smaller than the asymptotic variance of the optimal bridge sampling estimator

[Chen, Shao, & Ibrahim, 2000]

Does not require the normalising constant

$$\int | \pi_1(\eta) - \pi_2(\eta) | \, \mathrm{d}\eta$$

but a simulation from

$$\varphi^*(\theta) \propto | \pi_1(\theta) - \pi_2(\theta) | \, .$$

Improving upon bridge sampler

Theorem 5.5.3: The asymptotic variance of the optimal ratio importance sampling estimator is smaller than the asymptotic variance of the optimal bridge sampling estimator

[Chen, Shao, & Ibrahim, 2000]

Does not require the normalising constant

$$\int | \pi_1(\eta) - \pi_2(\eta) | \, \mathrm{d}\eta$$

but a simulation from

$$\varphi^*(\theta) \propto | \pi_1(\theta) - \pi_2(\theta) | \, .$$

Generalisation to point null situations

When

$$B_{12} = \frac{\int_{\Theta_1} \tilde{\pi}_1(\theta_1)\mathrm{d}\theta_1}{\int_{\Theta_2} \tilde{\pi}_2(\theta_2)\mathrm{d}\theta_2}$$

and $\Theta_2 = \Theta_1 \times \Psi$, we get $\theta_2 = (\theta_1, \psi)$ and

$$B_{12} = \mathbb{E}_{\pi_2}\left[\frac{\tilde{\pi}_1(\theta_1)\omega(\psi|\theta_1)}{\tilde{\pi}_2(\theta_1,\psi)}\right]$$

holds for any conditional density $\omega(\psi|\theta_1)$.

X-dimen'al bridge sampling

Generalisation of the previous identity:
For any α,

$$B_{12} = \frac{\mathbb{E}_{\pi_2}\left[\tilde{\pi}_1(\theta_1)\omega(\psi|\theta_1)\alpha(\theta_1,\psi)\right]}{\mathbb{E}_{\pi_1\times\omega}\left[\tilde{\pi}_2(\theta_1,\psi)\alpha(\theta_1,\psi)\right]}$$

and, for any density φ,

$$B_{12} = \frac{\mathbb{E}_{\varphi}\left[\tilde{\pi}_1(\theta_1)\omega(\psi|\theta_1)/\varphi(\theta_1,\psi)\right]}{\mathbb{E}_{\varphi}\left[\tilde{\pi}_2(\theta_1,\psi)/\varphi(\theta_1,\psi)\right]}$$

[Chen, Shao, & Ibrahim, 2000]

Optimal choice: $\omega(\psi|\theta_1) = \pi_2(\psi|\theta_1)$

[Theorem 5.8.2]

X-dimen'al bridge sampling

Generalisation of the previous identity:
For any α,

$$B_{12} = \frac{\mathbb{E}_{\pi_2}\left[\tilde{\pi}_1(\theta_1)\omega(\psi|\theta_1)\alpha(\theta_1,\psi)\right]}{\mathbb{E}_{\pi_1\times\omega}\left[\tilde{\pi}_2(\theta_1,\psi)\alpha(\theta_1,\psi)\right]}$$

and, for any density φ,

$$B_{12} = \frac{\mathbb{E}_{\varphi}\left[\tilde{\pi}_1(\theta_1)\omega(\psi|\theta_1)/\varphi(\theta_1,\psi)\right]}{\mathbb{E}_{\varphi}\left[\tilde{\pi}_2(\theta_1,\psi)/\varphi(\theta_1,\psi)\right]}$$

[Chen, Shao, & Ibrahim, 2000]

Optimal choice: $\omega(\psi|\theta_1) = \pi_2(\psi|\theta_1)$

[Theorem 5.8.2]

Approximating $\mathfrak{Z}_k$ from a posterior sample

Use of the [harmonic mean] identity

$$\mathbb{E}^{\pi_k}\left[\frac{\varphi(\theta_k)}{\pi_k(\theta_k)L_k(\theta_k)}\middle| x\right] = \int \frac{\varphi(\theta_k)}{\pi_k(\theta_k)L_k(\theta_k)} \frac{\pi_k(\theta_k)L_k(\theta_k)}{\mathfrak{Z}_k}\,\mathrm{d}\theta_k = \frac{1}{\mathfrak{Z}_k}$$

no matter what the proposal $\varphi(\cdot)$ is.

[Gelfand & Dey, 1994; Bartolucci et al., 2006]

Direct exploitation of the MCMC output

RB-RJ

Approximating $\mathfrak{Z}_k$ from a posterior sample

Use of the [harmonic mean] identity

$$\mathbb{E}^{\pi_k}\left[\left.\frac{\varphi(\theta_k)}{\pi_k(\theta_k)L_k(\theta_k)}\right| x\right] = \int \frac{\varphi(\theta_k)}{\pi_k(\theta_k)L_k(\theta_k)} \frac{\pi_k(\theta_k)L_k(\theta_k)}{\mathfrak{Z}_k}\,\mathrm{d}\theta_k = \frac{1}{\mathfrak{Z}_k}$$

no matter what the proposal $\varphi(\cdot)$ is.

[Gelfand & Dey, 1994; Bartolucci et al., 2006]

Direct exploitation of the MCMC output

RB-RJ

Comparison with regular importance sampling

Harmonic mean: Constraint opposed to usual importance sampling constraints: $\varphi(\theta)$ must have lighter (rather than fatter) tails than $\pi_k(\theta_k)L_k(\theta_k)$ for the approximation

$$\widehat{\mathfrak{Z}_{1k}} = 1 \Big/ \frac{1}{T} \sum_{t=1}^{T} \frac{\varphi(\theta_k^{(t)})}{\pi_k(\theta_k^{(t)})L_k(\theta_k^{(t)})}$$

to have a finite variance.

E.g., use finite support kernels (like Epanechnikov's kernel) for φ

Comparison with regular importance sampling

Harmonic mean: Constraint opposed to usual importance sampling constraints: $\varphi(\theta)$ must have lighter (rather than fatter) tails than $\pi_k(\theta_k)L_k(\theta_k)$ for the approximation

$$\widehat{\mathfrak{Z}_{1k}} = 1 \Big/ \frac{1}{T} \sum_{t=1}^{T} \frac{\varphi(\theta_k^{(t)})}{\pi_k(\theta_k^{(t)})L_k(\theta_k^{(t)})}$$

to have a finite variance.
E.g., use finite support kernels (like Epanechnikov's kernel) for φ

Comparison with regular importance sampling (cont'd)

Compare $\widehat{\mathfrak{Z}_{1k}}$ with a standard importance sampling approximation

$$\widehat{\mathfrak{Z}_{2k}} = \frac{1}{T} \sum_{t=1}^{T} \frac{\pi_k(\theta_k^{(t)}) L_k(\theta_k^{(t)})}{\varphi(\theta_k^{(t)})}$$

where the $\theta_k^{(t)}$'s are generated from the density $\varphi(\cdot)$ (with fatter tails like t's)

Approximating $\mathfrak{Z}_k$ using a mixture representation

Bridge sampling redux

Design a specific mixture for simulation [importance sampling] purposes, with density

$$\widetilde{\varphi}_k(\theta_k) \propto \omega_1 \pi_k(\theta_k) L_k(\theta_k) + \varphi(\theta_k)\,,$$

where $\varphi(\cdot)$ is arbitrary (but normalised)

Note: ω_1 is not a probability weight

Approximating $\mathfrak{Z}_k$ using a mixture representation

Bridge sampling redux

Design a specific mixture for simulation [importance sampling] purposes, with density

$$\widetilde{\varphi}_k(\theta_k) \propto \omega_1 \pi_k(\theta_k) L_k(\theta_k) + \varphi(\theta_k)\,,$$

where $\varphi(\cdot)$ is arbitrary (but normalised)
Note: ω_1 is not a probability weight

Approximating $\mathfrak{Z}$ using a mixture representation (cont'd)

Corresponding MCMC (=Gibbs) sampler

At iteration t

1. Take $\delta^{(t)} = 1$ with probability

$$\omega_1 \pi_k(\theta_k^{(t-1)}) L_k(\theta_k^{(t-1)}) \Big/ \left(\omega_1 \pi_k(\theta_k^{(t-1)}) L_k(\theta_k^{(t-1)}) + \varphi(\theta_k^{(t-1)})\right)$$

   and $\delta^{(t)} = 2$ otherwise;

2. If $\delta^{(t)} = 1$, generate $\theta_k^{(t)} \sim \text{MCMC}(\theta_k^{(t-1)}, \theta_k)$ where $\text{MCMC}(\theta_k, \theta_k')$ denotes an arbitrary MCMC kernel associated with the posterior $\pi_k(\theta_k|x) \propto \pi_k(\theta_k) L_k(\theta_k)$;

3. If $\delta^{(t)} = 2$, generate $\theta_k^{(t)} \sim \varphi(\theta_k)$ independently

Approximating $\mathfrak{Z}$ using a mixture representation (cont'd)

Corresponding MCMC (=Gibbs) sampler

At iteration t

1. Take $\delta^{(t)} = 1$ with probability

$$\omega_1 \pi_k(\theta_k^{(t-1)}) L_k(\theta_k^{(t-1)}) \Big/ \left(\omega_1 \pi_k(\theta_k^{(t-1)}) L_k(\theta_k^{(t-1)}) + \varphi(\theta_k^{(t-1)})\right)$$

   and $\delta^{(t)} = 2$ otherwise;

2. If $\delta^{(t)} = 1$, generate $\theta_k^{(t)} \sim \text{MCMC}(\theta_k^{(t-1)}, \theta_k)$ where $\text{MCMC}(\theta_k, \theta_k')$ denotes an arbitrary MCMC kernel associated with the posterior $\pi_k(\theta_k|x) \propto \pi_k(\theta_k) L_k(\theta_k)$;

3. If $\delta^{(t)} = 2$, generate $\theta_k^{(t)} \sim \varphi(\theta_k)$ independently

Approximating $\mathfrak{Z}$ using a mixture representation (cont'd)

Corresponding MCMC (=Gibbs) sampler

At iteration t

1. Take $\delta^{(t)} = 1$ with probability

$$\omega_1 \pi_k(\theta_k^{(t-1)}) L_k(\theta_k^{(t-1)}) \Big/ \left(\omega_1 \pi_k(\theta_k^{(t-1)}) L_k(\theta_k^{(t-1)}) + \varphi(\theta_k^{(t-1)})\right)$$

   and $\delta^{(t)} = 2$ otherwise;

2. If $\delta^{(t)} = 1$, generate $\theta_k^{(t)} \sim \text{MCMC}(\theta_k^{(t-1)}, \theta_k)$ where $\text{MCMC}(\theta_k, \theta_k')$ denotes an arbitrary MCMC kernel associated with the posterior $\pi_k(\theta_k|x) \propto \pi_k(\theta_k) L_k(\theta_k)$;

3. If $\delta^{(t)} = 2$, generate $\theta_k^{(t)} \sim \varphi(\theta_k)$ independently

Evidence approximation by mixtures

Rao-Blackwellised estimate

$$\hat{\xi} = \frac{1}{T} \sum_{t=1}^{T} \omega_1 \pi_k(\theta_k^{(t)}) L_k(\theta_k^{(t)}) \Big/ \omega_1 \pi_k(\theta_k^{(t)}) L_k(\theta_k^{(t)}) + \varphi(\theta_k^{(t)}) \,,$$

converges to $\omega_1 \mathfrak{Z}_k / \{\omega_1 \mathfrak{Z}_k + 1\}$
Deduce $\hat{\mathfrak{Z}}_{3k}$ from $\omega_1 \hat{\mathfrak{Z}}_{3k} / \{\omega_1 \hat{\mathfrak{Z}}_{3k} + 1\} = \hat{\xi}$ ie

$$\hat{\mathfrak{Z}}_{3k} = \frac{\sum_{t=1}^{T} \omega_1 \pi_k(\theta_k^{(t)}) L_k(\theta_k^{(t)}) \Big/ \omega_1 \pi(\theta_k^{(t)}) L_k(\theta_k^{(t)}) + \varphi(\theta_k^{(t)})}{\sum_{t=1}^{T} \varphi(\theta_k^{(t)}) \Big/ \omega_1 \pi_k(\theta_k^{(t)}) L_k(\theta_k^{(t)}) + \varphi(\theta_k^{(t)})}$$

[Bridge sampler]

Evidence approximation by mixtures

Rao-Blackwellised estimate

$$\hat{\xi} = \frac{1}{T} \sum_{t=1}^{T} \omega_1 \pi_k(\theta_k^{(t)}) L_k(\theta_k^{(t)}) \Big/ \omega_1 \pi_k(\theta_k^{(t)}) L_k(\theta_k^{(t)}) + \varphi(\theta_k^{(t)}) \,,$$

converges to $\omega_1 \mathfrak{Z}_k / \{\omega_1 \mathfrak{Z}_k + 1\}$
Deduce $\hat{\mathfrak{Z}}_{3k}$ from $\omega_1 \hat{\mathfrak{Z}}_{3k} / \{\omega_1 \hat{\mathfrak{Z}}_{3k} + 1\} = \hat{\xi}$ ie

$$\hat{\mathfrak{Z}}_{3k} = \frac{\sum_{t=1}^{T} \omega_1 \pi_k(\theta_k^{(t)}) L_k(\theta_k^{(t)}) \Big/ \omega_1 \pi(\theta_k^{(t)}) L_k(\theta_k^{(t)}) + \varphi(\theta_k^{(t)})}{\sum_{t=1}^{T} \varphi(\theta_k^{(t)}) \Big/ \omega_1 \pi_k(\theta_k^{(t)}) L_k(\theta_k^{(t)}) + \varphi(\theta_k^{(t)})}$$

[Bridge sampler]

Chib's representation

Direct application of Bayes' theorem: given $\mathbf{x} \sim f_k(\mathbf{x}|\theta_k)$ and $\theta_k \sim \pi_k(\theta_k)$,

$$\mathfrak{Z}_k = m_k(\mathbf{x}) = \frac{f_k(\mathbf{x}|\theta_k)\,\pi_k(\theta_k)}{\pi_k(\theta_k|\mathbf{x})}$$

Use of an approximation to the posterior

$$\widehat{\mathfrak{Z}}_k = \widehat{m_k}(\mathbf{x}) = \frac{f_k(\mathbf{x}|\theta_k^*)\,\pi_k(\theta_k^*)}{\hat{\pi}_k(\theta_k^*|\mathbf{x})}\,.$$

Chib's representation

Direct application of Bayes' theorem: given $\mathbf{x} \sim f_k(\mathbf{x}|\theta_k)$ and $\theta_k \sim \pi_k(\theta_k)$,

$$\mathfrak{Z}_k = m_k(\mathbf{x}) = \frac{f_k(\mathbf{x}|\theta_k)\,\pi_k(\theta_k)}{\pi_k(\theta_k|\mathbf{x})}$$

Use of an approximation to the posterior

$$\widehat{\mathfrak{Z}}_k = \widehat{m_k}(\mathbf{x}) = \frac{f_k(\mathbf{x}|\theta_k^*)\,\pi_k(\theta_k^*)}{\hat{\pi}_k(\theta_k^*|\mathbf{x})}\,.$$

Case of latent variables

For missing variable $\mathbf{z}$ as in mixture models, natural Rao-Blackwell estimate

$$\widehat{\pi_k}(\theta_k^*|\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} \pi_k(\theta_k^*|\mathbf{x}, \mathbf{z}_k^{(t)})\,,$$

where the $\mathbf{z}_k^{(t)}$'s are Gibbs sampled latent variables

Label switching

A mixture model [special case of missing variable model] is invariant under permutations of the indices of the components. E.g., mixtures

$$0.3\mathcal{N}(0,1) + 0.7\mathcal{N}(2.3,1)$$

and

$$0.7\mathcal{N}(2.3,1) + 0.3\mathcal{N}(0,1)$$

are **exactly** the same!

© The component parameters θ_i are not identifiable marginally since they are exchangeable

Label switching

A mixture model [special case of missing variable model] is invariant under permutations of the indices of the components. E.g., mixtures

$$0.3\mathcal{N}(0,1) + 0.7\mathcal{N}(2.3,1)$$

and

$$0.7\mathcal{N}(2.3,1) + 0.3\mathcal{N}(0,1)$$

are **exactly** the same!

© The component parameters θ_i are not identifiable marginally since they are exchangeable

Connected difficulties

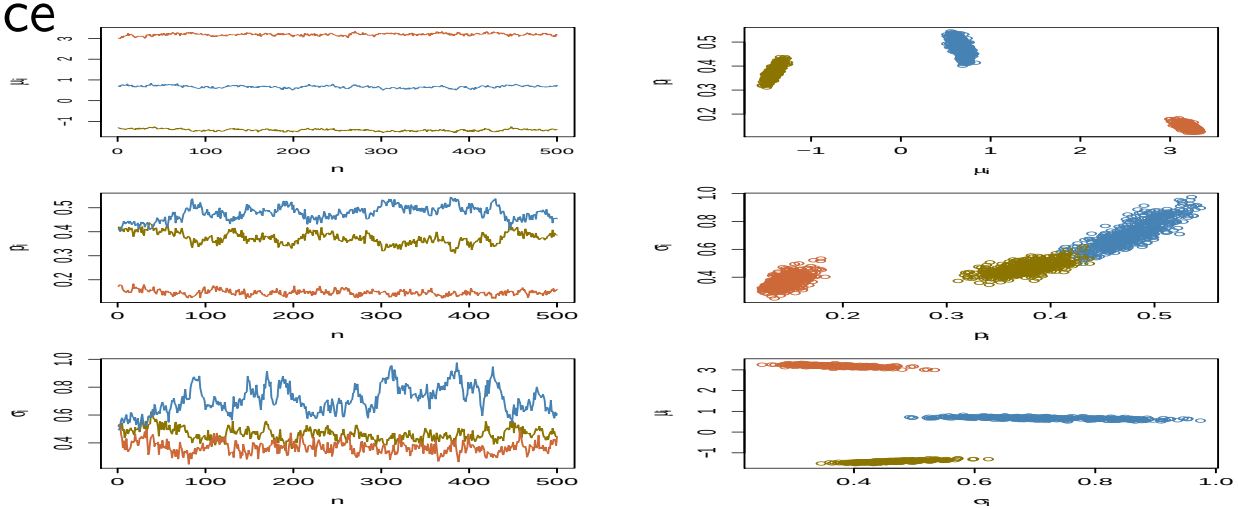
1. Number of modes of the likelihood of order $\mathrm{O}(k!)$:
   © Maximization and even [MCMC] exploration of the posterior surface harder
2. Under exchangeable priors on $(\theta, \mathbf{p})$ *[prior invariant under permutation of the indices]*, all posterior marginals are identical:
   © Posterior expectation of θ_1 equal to posterior expectation of θ_2

Connected difficulties

1. Number of modes of the likelihood of order $\mathrm{O}(k!)$:
   © Maximization and even [MCMC] exploration of the posterior surface harder
2. Under exchangeable priors on $(\theta, \mathbf{p})$ *[prior invariant under permutation of the indices]*, all posterior marginals are identical:
   © Posterior expectation of θ_1 equal to posterior expectation of θ_2

License

Since Gibbs output does not produce exchangeability, the Gibbs sampler has not explored the whole parameter space: it lacks energy to switch simultaneously enough component allocations at once



Label switching paradox

We should observe the exchangeability of the components [label switching] to conclude about convergence of the Gibbs sampler.
If we observe it, then we do not know how to estimate the parameters.
If we do not, then we are uncertain about the convergence!!!

Label switching paradox

We should observe the exchangeability of the components [label switching] to conclude about convergence of the Gibbs sampler. If we observe it, then we do not know how to estimate the parameters.

If we do not, then we are uncertain about the convergence!!!

Label switching paradox

We should observe the exchangeability of the components [label switching] to conclude about convergence of the Gibbs sampler.
If we observe it, then we do not know how to estimate the parameters.
If we do not, then we are uncertain about the convergence!!!

Compensation for label switching

For mixture models, $\mathbf{z}_k^{(t)}$ usually fails to visit all configurations in a balanced way, despite the symmetry predicted by the theory

$$\pi_k(\theta_k|\mathbf{x}) = \pi_k(\sigma(\theta_k)|\mathbf{x}) = \frac{1}{k!}\sum_{\sigma\in\mathfrak{S}} \pi_k(\sigma(\theta_k)|\mathbf{x})$$

for all σ's in $\mathfrak{S}_k$, set of all permutations of $\{1,\ldots,k\}$.
Consequences on numerical approximation, biased by an order $k!$
Recover the theoretical symmetry by using

$$\widetilde{\pi_k}(\theta_k^*|\mathbf{x}) = \frac{1}{T\,k!}\sum_{\sigma\in\mathfrak{S}_k}\sum_{t=1}^{T} \pi_k(\sigma(\theta_k^*)|\mathbf{x},\mathbf{z}_k^{(t)})\,.$$

[Berkhof, Mechelen, & Gelman, 2003]

Compensation for label switching

For mixture models, $\mathbf{z}_k^{(t)}$ usually fails to visit all configurations in a balanced way, despite the symmetry predicted by the theory

$$\pi_k(\theta_k|\mathbf{x}) = \pi_k(\sigma(\theta_k)|\mathbf{x}) = \frac{1}{k!}\sum_{\sigma\in\mathfrak{S}} \pi_k(\sigma(\theta_k)|\mathbf{x})$$

for all σ's in $\mathfrak{S}_k$, set of all permutations of $\{1,\ldots,k\}$.
Consequences on numerical approximation, biased by an order $k!$
Recover the theoretical symmetry by using

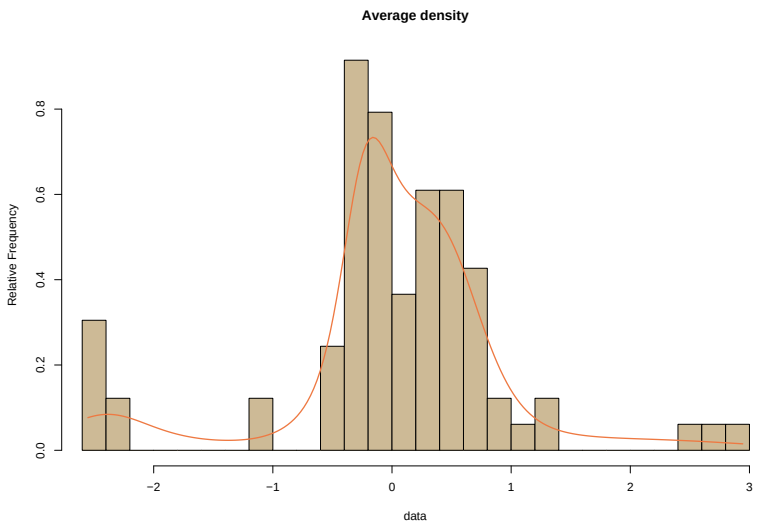
$$\widetilde{\pi_k}(\theta_k^*|\mathbf{x}) = \frac{1}{T\,k!}\sum_{\sigma\in\mathfrak{S}_k}\sum_{t=1}^{T} \pi_k(\sigma(\theta_k^*)|\mathbf{x},\mathbf{z}_k^{(t)})\,.$$

[Berkhof, Mechelen, & Gelman, 2003]

Galaxy dataset

$n = 82$ galaxies as a mixture of k normal distributions with both mean and variance unknown.

[Roeder, 1992]



Galaxy dataset (k)

Using only the original estimate, with θ_k^* as the MAP estimator,

$$\log(\hat{m}_k(\mathbf{x})) = -105.1396$$

for $k = 3$ (based on 10^3 simulations), while introducing the permutations leads to

$$\log(\hat{m}_k(\mathbf{x})) = -103.3479$$

Note that

$$-105.1396 + \log(3!) = -103.3479$$

k	2	3	4	5	6	7	8
$m_k(\mathbf{x})$	-115.68	-103.35	-102.66	-101.93	-102.88	-105.48	-108.44

Estimations of the marginal likelihoods by the symmetrised Chib's approximation (based on 10^5 Gibbs iterations and, for $k > 5$, 100 permutations selected at random in $\mathfrak{S}_k$).

[Lee, Marin, Mengersen & Robert, 2008]

Galaxy dataset (k)

Using only the original estimate, with θ_k^* as the MAP estimator,

$$\log(\hat{m}_k(\mathbf{x})) = -105.1396$$

for $k = 3$ (based on 10^3 simulations), while introducing the permutations leads to

$$\log(\hat{m}_k(\mathbf{x})) = -103.3479$$

Note that

$$-105.1396 + \log(3!) = -103.3479$$

k	2	3	4	5	6	7	8
$m_k(\mathbf{x})$	-115.68	-103.35	-102.66	-101.93	-102.88	-105.48	-108.44

Estimations of the marginal likelihoods by the symmetrised Chib's approximation (based on 10^5 Gibbs iterations and, for $k > 5$, 100 permutations selected at random in $\mathfrak{S}_k$).

[Lee, Marin, Mengersen & Robert, 2008]

Galaxy dataset (k)

Using only the original estimate, with θ_k^* as the MAP estimator,

$$\log(\hat{m}_k(\mathbf{x})) = -105.1396$$

for $k = 3$ (based on 10^3 simulations), while introducing the permutations leads to

$$\log(\hat{m}_k(\mathbf{x})) = -103.3479$$

Note that

$$-105.1396 + \log(3!) = -103.3479$$

k	2	3	4	5	6	7	8
$m_k(\mathbf{x})$	-115.68	-103.35	-102.66	-101.93	-102.88	-105.48	-108.44

Estimations of the marginal likelihoods by the symmetrised Chib's approximation (based on 10^5 Gibbs iterations and, for $k > 5$, 100 permutations selected at random in $\mathfrak{S}_k$).

[Lee, Marin, Mengersen & Robert, 2008]

Bayesian variable selection

Example of a regression setting: one dependent random variable y and a set $\{x_1, \ldots, x_k\}$ of k explanatory variables.

Question: Are all x_i's involved in the regression?

Assumption: every subset $\{i_1, \ldots, i_q\}$ of q $(0 \leq q \leq k)$ explanatory variables, $\{\mathbf{1}_n, x_{i_1}, \ldots, x_{i_q}\}$, is a proper set of explanatory variables for the regression of y [intercept included in every corresponding model]

Computational issue

2^k models in competition...

[Marin & Robert, Bayesian Core, 2007]

Bayesian variable selection

Example of a regression setting: one dependent random variable y and a set $\{x_1, \ldots, x_k\}$ of k explanatory variables.

Question: Are all x_i's involved in the regression?

Assumption: every subset $\{i_1, \ldots, i_q\}$ of q $(0 \leq q \leq k)$ explanatory variables, $\{\mathbf{1}_n, x_{i_1}, \ldots, x_{i_q}\}$, is a proper set of explanatory variables for the regression of y [intercept included in every corresponding model]

Computational issue

2^k models in competition...

[Marin & Robert, Bayesian Core, 2007]

Bayesian variable selection

Example of a regression setting: one dependent random variable y and a set $\{x_1, \ldots, x_k\}$ of k explanatory variables.

Question: Are all x_i's involved in the regression?

Assumption: every subset $\{i_1, \ldots, i_q\}$ of q $(0 \leq q \leq k)$ explanatory variables, $\{\mathbf{1}_n, x_{i_1}, \ldots, x_{i_q}\}$, is a proper set of explanatory variables for the regression of y [intercept included in every corresponding model]

Computational issue

2^k models in competition...

[Marin & Robert, Bayesian Core, 2007]

Bayesian variable selection

Example of a regression setting: one dependent random variable y and a set $\{x_1, \ldots, x_k\}$ of k explanatory variables.

Question: Are all x_i's involved in the regression?

Assumption: every subset $\{i_1, \ldots, i_q\}$ of q $(0 \leq q \leq k)$ explanatory variables, $\{\mathbf{1}_n, x_{i_1}, \ldots, x_{i_q}\}$, is a proper set of explanatory variables for the regression of y [intercept included in every corresponding model]

Computational issue

2^k models in competition...

[Marin & Robert, Bayesian Core, 2007]

Reversible jump

Idea: Set up a proper measure–theoretic framework for designing moves *between* models $\mathfrak{M}_k$

[Green, 1995]

Create a **reversible kernel** $\mathfrak{K}$ on $\mathfrak{H} = \bigcup_k \{k\} \times \Theta_k$ such that

$$\int_A \int_B \mathfrak{K}(x, dy)\pi(x)dx = \int_B \int_A \mathfrak{K}(y, dx)\pi(y)dy$$

for the invariant density π [x is of the form $(k, \theta^{(k)})$]

Reversible jump

Idea: Set up a proper measure–theoretic framework for designing moves *between* models $\mathfrak{M}_k$

[Green, 1995]

Create a **reversible kernel** $\mathfrak{K}$ on $\mathfrak{H} = \bigcup_k \{k\} \times \Theta_k$ such that

$$\int_A \int_B \mathfrak{K}(x, dy)\pi(x)dx = \int_B \int_A \mathfrak{K}(y, dx)\pi(y)dy$$

for the invariant density π [x is of the form $(k, \theta^{(k)})$]

Local moves

For a move between two models, $\mathfrak{M}_1$ and $\mathfrak{M}_2$, the Markov chain being in state $\theta_1 \in \mathfrak{M}_1$, denote by $\mathfrak{K}_{1\to2}(\theta_1, d\theta)$ and $\mathfrak{K}_{2\to1}(\theta_2, d\theta)$ the corresponding kernels, under the *detailed balance condition*

$$\pi(d\theta_1)\,\mathfrak{K}_{1\to2}(\theta_1, d\theta) = \pi(d\theta_2)\,\mathfrak{K}_{2\to1}(\theta_2, d\theta)\,,$$

and take, wlog, $\dim(\mathfrak{M}_2) > \dim(\mathfrak{M}_1)$.

Proposal expressed as

$$\theta_2 = \Psi_{1\to2}(\theta_1, v_{1\to2})$$

where $v_{1\to2}$ is a random variable of dimension $\dim(\mathfrak{M}_2) - \dim(\mathfrak{M}_1)$, generated as

$$v_{1\to2} \sim \varphi_{1\to2}(v_{1\to2})\,.$$

Local moves

For a move between two models, $\mathfrak{M}_1$ and $\mathfrak{M}_2$, the Markov chain being in state $\theta_1 \in \mathfrak{M}_1$, denote by $\mathfrak{K}_{1\to2}(\theta_1, d\theta)$ and $\mathfrak{K}_{2\to1}(\theta_2, d\theta)$ the corresponding kernels, under the *detailed balance condition*

$$\pi(d\theta_1)\,\mathfrak{K}_{1\to2}(\theta_1, d\theta) = \pi(d\theta_2)\,\mathfrak{K}_{2\to1}(\theta_2, d\theta)\,,$$

and take, wlog, $\dim(\mathfrak{M}_2) > \dim(\mathfrak{M}_1)$.
Proposal expressed as

$$\theta_2 = \Psi_{1\to2}(\theta_1, v_{1\to2})$$

where $v_{1\to2}$ is a random variable of dimension $\dim(\mathfrak{M}_2) - \dim(\mathfrak{M}_1)$, generated as

$$v_{1\to2} \sim \varphi_{1\to2}(v_{1\to2})\,.$$

Local moves (2)

In this case, $\mathfrak{q}_{1\to2}(\theta_1, d\theta_2)$ has density

$$\varphi_{1\to2}(v_{1\to2}) \left| \frac{\partial \Psi_{1\to2}(\theta_1, v_{1\to2})}{\partial(\theta_1, v_{1\to2})} \right|^{-1},$$

by the Jacobian rule.

◂ Reverse importance link

If probability $\varpi_{1\to2}$ of choosing move to $\mathfrak{M}_2$ while in $\mathfrak{M}_1$, acceptance probability reduces to

$$\alpha(\theta_1, v_{1\to2}) = 1 \wedge \frac{\pi(\mathfrak{M}_2, \theta_2)\, \varpi_{2\to1}}{\pi(\mathfrak{M}_1, \theta_1)\, \varpi_{1\to2}\, \varphi_{1\to2}(v_{1\to2})} \left| \frac{\partial \Psi_{1\to2}(\theta_1, v_{1\to2})}{\partial(\theta_1, v_{1\to2})} \right| .$$

If several models are considered simultaneously, with probability $\varpi_{1\to2}$ of choosing move to $\mathfrak{M}_2$ while in $\mathfrak{M}_1$, as in

$$\mathfrak{K}(x, B) = \sum_{m=1}^{\infty} \int \rho_m(x, y) \mathfrak{q}_m(x, dy) + \omega(x) \mathbb{I}_B(x)$$

Local moves (2)

In this case, $\mathfrak{q}_{1\to2}(\theta_1, d\theta_2)$ has density

$$\varphi_{1\to2}(v_{1\to2}) \left| \frac{\partial \Psi_{1\to2}(\theta_1, v_{1\to2})}{\partial(\theta_1, v_{1\to2})} \right|^{-1},$$

by the Jacobian rule.

◂ Reverse importance link

If probability $\varpi_{1\to2}$ of choosing move to $\mathfrak{M}_2$ while in $\mathfrak{M}_1$, acceptance probability reduces to

$$\alpha(\theta_1, v_{1\to2}) = 1 \wedge \frac{\pi(\mathfrak{M}_2, \theta_2)\, \varpi_{2\to1}}{\pi(\mathfrak{M}_1, \theta_1)\, \varpi_{1\to2}\, \varphi_{1\to2}(v_{1\to2})} \left| \frac{\partial \Psi_{1\to2}(\theta_1, v_{1\to2})}{\partial(\theta_1, v_{1\to2})} \right| .$$

If several models are considered simultaneously, with probability $\varpi_{1\to2}$ of choosing move to $\mathfrak{M}_2$ while in $\mathfrak{M}_1$, as in

$$\mathfrak{K}(x, B) = \sum_{m=1}^{\infty} \int \rho_m(x, y) \mathfrak{q}_m(x, dy) + \omega(x) \mathbb{I}_B(x)$$

Local moves (2)

In this case, $\mathfrak{q}_{1\to2}(\theta_1, d\theta_2)$ has density

$$\varphi_{1\to2}(v_{1\to2}) \left| \frac{\partial \Psi_{1\to2}(\theta_1, v_{1\to2})}{\partial(\theta_1, v_{1\to2})} \right|^{-1},$$

by the Jacobian rule.

◂ Reverse importance link

If probability $\varpi_{1\to2}$ of choosing move to $\mathfrak{M}_2$ while in $\mathfrak{M}_1$, acceptance probability reduces to

$$\alpha(\theta_1, v_{1\to2}) = 1 \wedge \frac{\pi(\mathfrak{M}_2, \theta_2)\, \varpi_{2\to1}}{\pi(\mathfrak{M}_1, \theta_1)\, \varpi_{1\to2}\, \varphi_{1\to2}(v_{1\to2})} \left| \frac{\partial \Psi_{1\to2}(\theta_1, v_{1\to2})}{\partial(\theta_1, v_{1\to2})} \right| .$$

If several models are considered simultaneously, with probability $\varpi_{1\to2}$ of choosing move to $\mathfrak{M}_2$ while in $\mathfrak{M}_1$, as in

$$\mathfrak{K}(x, B) = \sum_{m=1}^{\infty} \int \rho_m(x, y) \mathfrak{q}_m(x, dy) + \omega(x) \mathbb{I}_B(x)$$

Generic reversible jump acceptance probability

Acceptance probability of $\theta_2 = \Psi_{1\to2}(\theta_1, v_{1\to2})$ is

$$\alpha(\theta_1, v_{1\to2}) = 1 \wedge \frac{\pi(\mathfrak{M}_2, \theta_2)\, \varpi_{2\to1}}{\pi(\mathfrak{M}_1, \theta_1)\, \varpi_{1\to2}\, \varphi_{1\to2}(v_{1\to2})} \left| \frac{\partial \Psi_{1\to2}(\theta_1, v_{1\to2})}{\partial(\theta_1, v_{1\to2})} \right|$$

while acceptance probability of θ_1 with $(\theta_1, v_{1\to2}) = \Psi^{-1}_{1\to2}(\theta_2)$ is

$$\alpha(\theta_1, v_{1\to2}) = 1 \wedge \frac{\pi(\mathfrak{M}_1, \theta_1)\, \varpi_{1\to2}\, \varphi_{1\to2}(v_{1\to2})}{\pi(\mathfrak{M}_2, \theta_2)\, \varpi_{2\to1}} \left| \frac{\partial \Psi_{1\to2}(\theta_1, v_{1\to2})}{\partial(\theta_1, v_{1\to2})} \right|^{-1}$$

©Difficult calibration

Generic reversible jump acceptance probability

Acceptance probability of $\theta_2 = \Psi_{1\to2}(\theta_1, v_{1\to2})$ is

$$\alpha(\theta_1, v_{1\to2}) = 1 \wedge \frac{\pi(\mathfrak{M}_2, \theta_2)\, \varpi_{2\to1}}{\pi(\mathfrak{M}_1, \theta_1)\, \varpi_{1\to2}\, \varphi_{1\to2}(v_{1\to2})} \left| \frac{\partial \Psi_{1\to2}(\theta_1, v_{1\to2})}{\partial(\theta_1, v_{1\to2})} \right|$$

while acceptance probability of θ_1 with $(\theta_1, v_{1\to2}) = \Psi^{-1}_{1\to2}(\theta_2)$ is

$$\alpha(\theta_1, v_{1\to2}) = 1 \wedge \frac{\pi(\mathfrak{M}_1, \theta_1)\, \varpi_{1\to2}\, \varphi_{1\to2}(v_{1\to2})}{\pi(\mathfrak{M}_2, \theta_2)\, \varpi_{2\to1}} \left| \frac{\partial \Psi_{1\to2}(\theta_1, v_{1\to2})}{\partial(\theta_1, v_{1\to2})} \right|^{-1}$$

Generic reversible jump acceptance probability

Acceptance probability of $\theta_2 = \Psi_{1\to2}(\theta_1, v_{1\to2})$ is

$$\alpha(\theta_1, v_{1\to2}) = 1 \wedge \frac{\pi(\mathfrak{M}_2, \theta_2)\, \varpi_{2\to1}}{\pi(\mathfrak{M}_1, \theta_1)\, \varpi_{1\to2}\, \varphi_{1\to2}(v_{1\to2})} \left| \frac{\partial \Psi_{1\to2}(\theta_1, v_{1\to2})}{\partial(\theta_1, v_{1\to2})} \right|$$

while acceptance probability of θ_1 with $(\theta_1, v_{1\to2}) = \Psi^{-1}_{1\to2}(\theta_2)$ is

$$\alpha(\theta_1, v_{1\to2}) = 1 \wedge \frac{\pi(\mathfrak{M}_1, \theta_1)\, \varpi_{1\to2}\, \varphi_{1\to2}(v_{1\to2})}{\pi(\mathfrak{M}_2, \theta_2)\, \varpi_{2\to1}} \left| \frac{\partial \Psi_{1\to2}(\theta_1, v_{1\to2})}{\partial(\theta_1, v_{1\to2})} \right|^{-1}$$

©Difficult calibration

Green's sampler

Algorithm

Iteration t $(t \geq 1)$**:** if $x^{(t)} = (m, \theta^{(m)})$,

1. Select model $\mathfrak{M}_n$ with probability π_{mn}
2. Generate $u_{mn} \sim \varphi_{mn}(u)$ and set $(\theta^{(n)}, v_{nm}) = \Psi_{m \to n}(\theta^{(m)}, u_{mn})$
3. Take $x^{(t+1)} = (n, \theta^{(n)})$ with probability

$$\min\left(\frac{\pi(n, \theta^{(n)})}{\pi(m, \theta^{(m)})} \, \frac{\pi_{nm}\varphi_{nm}(v_{nm})}{\pi_{mn}\varphi_{mn}(u_{mn})} \left|\frac{\partial \Psi_{m \to n}(\theta^{(m)}, u_{mn})}{\partial(\theta^{(m)}, u_{mn})}\right|, 1\right)$$

and take $x^{(t+1)} = x^{(t)}$ otherwise.

Interpretation

The representation puts us back in a fixed dimension setting:

- $\mathfrak{M}_1 \times \mathfrak{V}_{1\to 2}$ and $\mathfrak{M}_2$ in one-to-one relation.
- reversibility imposes that θ_1 is derived as

$$(\theta_1, v_{1\to 2}) = \Psi^{-1}_{1\to 2}(\theta_2)$$

- appears like a *regular* Metropolis–Hastings move from the couple $(\theta_1, v_{1\to 2})$ to θ_2 when stationary distributions are $\pi(\mathfrak{M}_1, \theta_1) \times \varphi_{1\to 2}(v_{1\to 2})$ and $\pi(\mathfrak{M}_2, \theta_2)$, and when proposal distribution is *deterministic*

Interpretation

The representation puts us back in a fixed dimension setting:

- $\mathfrak{M}_1 \times \mathfrak{V}_{1\to 2}$ and $\mathfrak{M}_2$ in one-to-one relation.
- reversibility imposes that θ_1 is derived as

$$(\theta_1, v_{1\to 2}) = \Psi_{1\to 2}^{-1}(\theta_2)$$

- appears like a *regular* Metropolis–Hastings move from the couple $(\theta_1, v_{1\to 2})$ to θ_2 when stationary distributions are $\pi(\mathfrak{M}_1, \theta_1) \times \varphi_{1\to 2}(v_{1\to 2})$ and $\pi(\mathfrak{M}_2, \theta_2)$, and when proposal distribution is *deterministic*

Alternative

Saturation of the parameter space $\mathfrak{H} = \bigcup_k \{k\} \times \Theta_k$ by creating

- $\theta = (\theta_1, \ldots, \theta_D)$
- a model index M
- pseudo-priors $\pi_j(\theta_j | M = k)$ for $j \neq k$

[Carlin & Chib, 1995]

Validation by

$$\mathbb{P}(M = k|x) = \int P(M = k|x, \theta)\pi(\theta|x)\mathrm{d}\theta = \mathfrak{Z}_k$$

where the (marginal) posterior is [not π_k!]

$$\pi(\theta|x) = \sum_{k=1}^{D} \mathbb{P}(\theta, M = k|x)$$

$$= \sum_{k=1}^{D} p_k \, \mathfrak{Z}_k \, \pi_k(\theta_k|x) \prod_{j \neq k} \pi_j(\theta_j|M = k) \, .$$

Alternative

Saturation of the parameter space $\mathfrak{H} = \bigcup_k \{k\} \times \Theta_k$ by creating

- $\theta = (\theta_1, \ldots, \theta_D)$
- a model index M
- pseudo-priors $\pi_j(\theta_j | M = k)$ for $j \neq k$

[Carlin & Chib, 1995]

Validation by

$$\mathbb{P}(M = k|x) = \int P(M = k|x, \theta)\pi(\theta|x)\mathrm{d}\theta = \mathfrak{Z}_k$$

where the (marginal) posterior is [not π_k!]

$$\begin{aligned}\pi(\theta|x) &= \sum_{k=1}^{D} \mathbb{P}(\theta, M = k|x) \\ &= \sum_{k=1}^{D} p_k \, \mathfrak{Z}_k \, \pi_k(\theta_k|x) \prod_{j \neq k} \pi_j(\theta_j|M = k) \, .\end{aligned}$$

MCMC implementation

Run a Markov chain $(M^{(t)}, \theta_1^{(t)}, \ldots, \theta_D^{(t)})$ with stationary distribution $\pi(\theta, M|x)$ by

1. Pick $M^{(t)} = k$ with probability $\pi(\theta^{(t-1)}, k|x)$
2. Generate $\theta_k^{(t-1)}$ from the posterior $\pi_k(\theta_k|x)$ [or MCMC step]
3. Generate $\theta_j^{(t-1)}$ $(j \neq k)$ from the pseudo-prior $\pi_j(\theta_j|M = k)$

Approximate $\mathbb{P}(M = k|x) = \beth_k$ by

$$\check{p}_k(x) \propto p_k \sum_{t=1}^{T} f_k(x|\theta_k^{(t)})\, \pi_k(\theta_k^{(t)}) \prod_{j \neq k} \pi_j(\theta_j^{(t)}|M = k)$$

$$\Big/ \sum_{\ell=1}^{D} p_\ell\, f_\ell(x|\theta_\ell^{(t)})\, \pi_\ell(\theta_\ell^{(t)}) \prod_{j \neq \ell} \pi_j(\theta_j^{(t)}|M = \ell)$$

MCMC implementation

Run a Markov chain $(M^{(t)}, \theta_1^{(t)}, \ldots, \theta_D^{(t)})$ with stationary distribution $\pi(\theta, M|x)$ by

1. Pick $M^{(t)} = k$ with probability $\pi(\theta^{(t-1)}, k|x)$
2. Generate $\theta_k^{(t-1)}$ from the posterior $\pi_k(\theta_k|x)$ [or MCMC step]
3. Generate $\theta_j^{(t-1)}$ $(j \neq k)$ from the pseudo-prior $\pi_j(\theta_j|M = k)$

Approximate $\mathbb{P}(M = k|x) = \beth_k$ by

$$\check{p}_k(x) \propto p_k \sum_{t=1}^{T} f_k(x|\theta_k^{(t)})\, \pi_k(\theta_k^{(t)}) \prod_{j \neq k} \pi_j(\theta_j^{(t)}|M = k)$$

$$\Big/ \sum_{\ell=1}^{D} p_\ell\, f_\ell(x|\theta_\ell^{(t)})\, \pi_\ell(\theta_\ell^{(t)}) \prod_{j \neq \ell} \pi_j(\theta_j^{(t)}|M = \ell)$$

MCMC implementation

Run a Markov chain $(M^{(t)}, \theta_1^{(t)}, \ldots, \theta_D^{(t)})$ with stationary distribution $\pi(\theta, M|x)$ by

1. Pick $M^{(t)} = k$ with probability $\pi(\theta^{(t-1)}, k|x)$
2. Generate $\theta_k^{(t-1)}$ from the posterior $\pi_k(\theta_k|x)$ [or MCMC step]
3. Generate $\theta_j^{(t-1)}$ $(j \neq k)$ from the pseudo-prior $\pi_j(\theta_j|M = k)$

Approximate $\mathbb{P}(M = k|x) = \mathfrak{Z}_k$ by

$$\check{p}_k(x) \propto p_k \sum_{t=1}^{T} f_k(x|\theta_k^{(t)})\, \pi_k(\theta_k^{(t)}) \prod_{j \neq k} \pi_j(\theta_j^{(t)}|M = k)$$

$$\Big/ \sum_{\ell=1}^{D} p_\ell\, f_\ell(x|\theta_\ell^{(t)})\, \pi_\ell(\theta_\ell^{(t)}) \prod_{j \neq \ell} \pi_j(\theta_j^{(t)}|M = \ell)$$

Scott's (2002) proposal

Suggest estimating $\mathbb{P}(M = k|x)$ by

$$\widetilde{\mathfrak{Z}}_k \propto p_k \sum_{t=1}^{T} \left\{ f_k(x|\theta_k^{(t)}) \Big/ \sum_{j=1}^{D} p_j \, f_j(x|\theta_j^{(t)}) \right\} ,$$

based on D simultaneous and independent MCMC chains

$$(\theta_k^{(t)})_t \, , \qquad 1 \leq k \leq D \, ,$$

with stationary distributions $\pi_k(\theta_k|x)$ [instead of above joint!!]

Scott's (2002) proposal

Suggest estimating $\mathbb{P}(M = k|x)$ by

$$\widetilde{\mathfrak{Z}}_k \propto p_k \sum_{t=1}^{T} \left\{ f_k(x|\theta_k^{(t)}) \Big/ \sum_{j=1}^{D} p_j \, f_j(x|\theta_j^{(t)}) \right\} ,$$

based on D simultaneous and independent MCMC chains

$$(\theta_k^{(t)})_t \, , \qquad 1 \leq k \leq D \, ,$$

with stationary distributions $\pi_k(\theta_k|x)$ [instead of above joint!!]

Congdon's (2006) extension

Selecting flat [prohibited!] pseudo-priors, uses instead

$$\widehat{\mathfrak{Z}}_k \propto p_k \sum_{t=1}^{T} \left\{ f_k(x|\theta_k^{(t)})\pi_k(\theta_k^{(t)}) \Big/ \sum_{j=1}^{D} p_j\, f_j(x|\theta_j^{(t)})\pi_j(\theta_j^{(t)}) \right\},$$

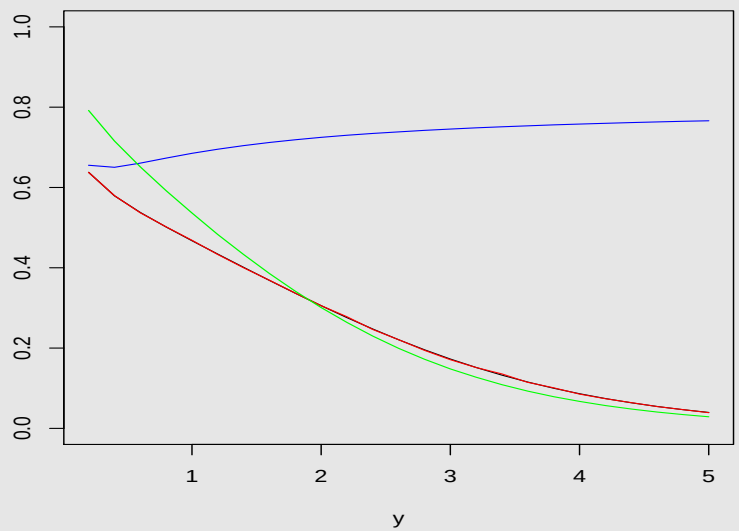
where again the $\theta_k^{(t)}$'s are MCMC chains with stationary distributions $\pi_k(\theta_k|x)$

Examples

Example (Model choice)

Model $\mathfrak{M}_1 : x|\theta \sim \mathcal{U}(0,\theta)$ with prior $\theta \sim \mathcal{E}xp(1)$ is versus model $\mathfrak{M}_2 : x|\theta \sim \mathcal{E}xp(\theta)$ with prior $\theta \sim \mathcal{E}xp(1)$. Equal prior weights on both models: $\varrho_1 = \varrho_2 = 0.5$.

Approximations of $\mathbb{P}(M = 1|x)$: Scott's (2002) (blue), and Congdon's (2006) (red) [$N = 10^6$ simulations].

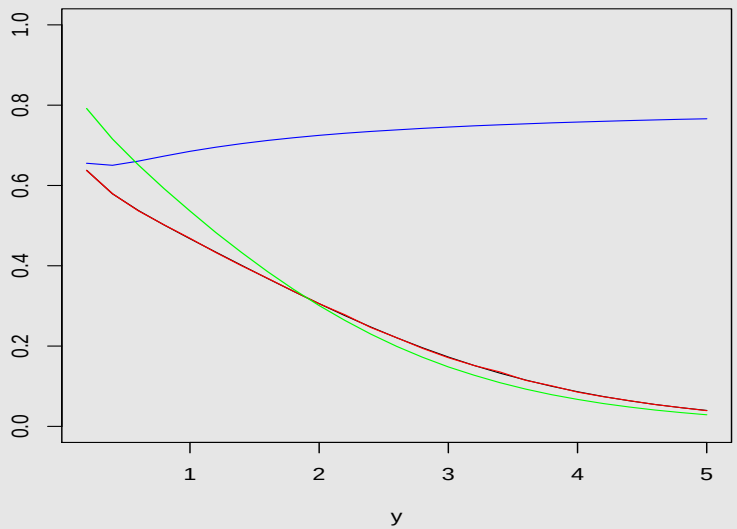


Examples

Example (Model choice)

Model $\mathfrak{M}_1 : x|\theta \sim \mathcal{U}(0,\theta)$ with prior $\theta \sim \mathcal{E}xp(1)$ is versus model $\mathfrak{M}_2 : x|\theta \sim \mathcal{E}xp(\theta)$ with prior $\theta \sim \mathcal{E}xp(1)$. Equal prior weights on both models: $\varrho_1 = \varrho_2 = 0.5$.

Approximations of $\mathbb{P}(M = 1|x)$:
Scott's (2002) (blue), and
Congdon's (2006) (red)
[$N = 10^6$ simulations].

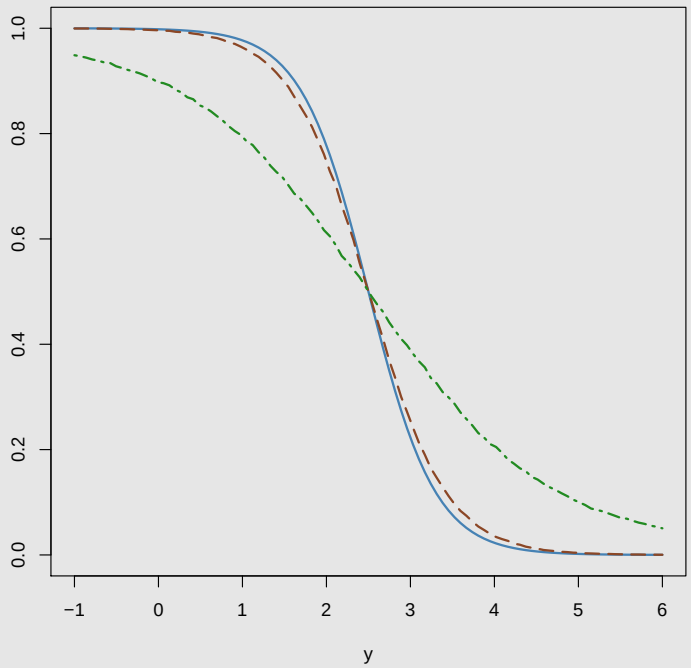


Examples (2)

Example (Model choice (2))

Normal model $\mathfrak{M}_1 : x \sim \mathcal{N}(\theta, 1)$ with $\theta \sim \mathcal{N}(0, 1)$ vs. normal model $\mathfrak{M}_2 : x \sim \mathcal{N}(\theta, 1)$ with $\theta \sim \mathcal{N}(5, 1)$

Comparison of both approximations with $\mathbb{P}(M = 1|x)$: Scott's (2002) (green and mixed dashes) and Congdon's (2006) (brown and long dashes) [$N = 10^4$ simulations].

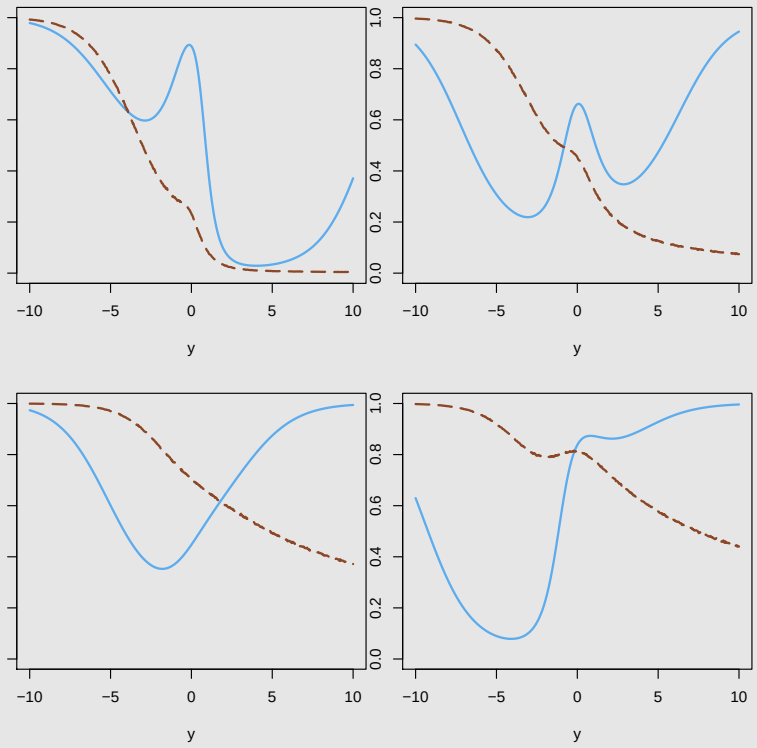


Examples (3)

Example (Model choice (3))

Model $\mathfrak{M}_1 : x \sim \mathcal{N}(0, 1/\omega)$ with $\omega \sim \mathcal{E}xp(a)$ vs.
$\mathfrak{M}_2 : \exp(x) \sim \mathcal{E}xp(\lambda)$ with $\lambda \sim \mathcal{E}xp(b)$.

Comparison of Congdon's (2006) (brown and dashed lines) with $\mathbb{P}(M = 1|x)$ when (a, b) is equal to $(.24, 8.9)$, $(.56, .7)$, $(4.1, .46)$ and $(.98, .081)$, resp. [$N = 10^4$ simulations].



Nested sampling: Goal

Skilling's (2007) technique using the one-dimensional representation:

$$\mathfrak{Z} = \mathbb{E}^{\pi}[L(\theta)] = \int_0^1 \varphi(x)\,\mathrm{d}x$$

with

$$\varphi^{-1}(l) = P^{\pi}(L(\theta) > l).$$

Note; $\varphi(\cdot)$ is intractable in most cases.

Nested sampling: First approximation

Approximate $\mathfrak{Z}$ by a Riemann sum:

$$\widehat{\mathfrak{Z}} = \sum_{i=1}^{j} (x_{i-1} - x_i)\varphi(x_i)$$

where the x_i's are either:

- deterministic: $x_i = e^{-i/N}$
- or random:

$$x_0 = 0, \quad x_{i+1} = t_i x_i, \quad t_i \sim \mathcal{B}e(N, 1)$$

so that $\mathbb{E}[\log x_i] = -i/N$.

Extraneous white noise

Take

$$\mathfrak{Z} = \int e^{-\theta}\,\mathrm{d}\theta = \int \frac{1}{\delta}\, e^{-(1-\delta)\theta}\, e^{-\delta\theta} = \mathbb{E}_\delta\left[\frac{1}{\delta}\, e^{-(1-\delta)\theta}\right]$$

$$\hat{\mathfrak{Z}} = \frac{1}{N}\sum_{i=1}^{N} \delta^{-1}\, e^{-(1-\delta)\theta_i}\,(x_{i-1} - x_i)\,, \quad \theta_i \sim \mathcal{E}(\delta)\,\mathbb{I}(\theta_i \leq \theta_{i-1})$$

N	deterministic	random
50	4.64	10.5
	4.65	10.5
100	2.47	4.9
	2.48	5.02
500	.549	1.01
	.550	1.14

Comparison of variances and MSEs

Extraneous white noise

Take

$$\mathfrak{Z} = \int e^{-\theta}\,\mathrm{d}\theta = \int \frac{1}{\delta}\, e^{-(1-\delta)\theta}\, e^{-\delta\theta} = \mathbb{E}_\delta\left[\frac{1}{\delta}\, e^{-(1-\delta)\theta}\right]$$

$$\hat{\mathfrak{Z}} = \frac{1}{N}\sum_{i=1}^{N} \delta^{-1}\, e^{-(1-\delta)\theta_i}\,(x_{i-1} - x_i)\,, \quad \theta_i \sim \mathcal{E}(\delta)\,\mathbb{I}(\theta_i \leq \theta_{i-1})$$

N	deterministic	random
50	4.64	10.5
	4.65	10.5
100	2.47	4.9
	2.48	5.02
500	.549	1.01
	.550	1.14

Comparison of variances and MSEs

Extraneous white noise

Take

$$\mathfrak{Z} = \int e^{-\theta}\,\mathrm{d}\theta = \int \frac{1}{\delta}\, e^{-(1-\delta)\theta}\, e^{-\delta\theta} = \mathbb{E}_\delta\left[\frac{1}{\delta}\, e^{-(1-\delta)\theta}\right]$$

$$\hat{\mathfrak{Z}} = \frac{1}{N}\sum_{i=1}^{N} \delta^{-1}\, e^{-(1-\delta)\theta_i}\,(x_{i-1} - x_i)\,, \quad \theta_i \sim \mathcal{E}(\delta)\,\mathbb{I}(\theta_i \leq \theta_{i-1})$$

N	deterministic	random
50	4.64	10.5
	4.65	10.5
100	2.47	4.9
	2.48	5.02
500	.549	1.01
	.550	1.14

Comparison of variances and MSEs

Nested sampling: Second approximation

Replace (intractable) $\varphi(x_i)$ by φ_i, obtained by

Nested sampling

Start with N values $\theta_1, \ldots, \theta_N$ sampled from π

At iteration i,

1. Take $\varphi_i = L(\theta_k)$, where θ_k is the point with smallest likelihood in the pool of θ_i's
2. Replace θ_k with a sample from the prior constrained to $L(\theta) > \varphi_i$: the current N points are sampled from prior constrained to $L(\theta) > \varphi_i$.

Nested sampling: Second approximation

Replace (intractable) $\varphi(x_i)$ by φ_i, obtained by

Nested sampling

Start with N values $\theta_1, \ldots, \theta_N$ sampled from π
At iteration i,

1. Take $\varphi_i = L(\theta_k)$, where θ_k is the point with smallest likelihood in the pool of θ_i's
2. Replace θ_k with a sample from the prior constrained to $L(\theta) > \varphi_i$: the current N points are sampled from prior constrained to $L(\theta) > \varphi_i$.

Nested sampling: Second approximation

Replace (intractable) $\varphi(x_i)$ by φ_i, obtained by

Nested sampling

Start with N values $\theta_1, \ldots, \theta_N$ sampled from π
At iteration i,

1. Take $\varphi_i = L(\theta_k)$, where θ_k is the point with smallest likelihood in the pool of θ_i's
2. Replace θ_k with a sample from the prior constrained to $L(\theta) > \varphi_i$: the current N points are sampled from prior constrained to $L(\theta) > \varphi_i$.

Nested sampling: Third approximation

Iterate the above steps until a given stopping iteration j is reached: e.g.,

- observe very small changes in the approximation $\widehat{\mathfrak{Z}}$;
- reach the maximal value of $L(\theta)$ when the likelihood is bounded and its maximum is known;
- truncate the integral $\mathfrak{Z}$ at level ϵ, i.e. replace

$$\int_0^1 \varphi(x)\,\mathrm{d}x \qquad \text{with} \qquad \int_\epsilon^1 \varphi(x)\,\mathrm{d}x$$

Approximation error

$$\text{Error} = \widehat{\mathfrak{Z}} - \mathfrak{Z}$$

$$= \sum_{i=1}^{j} (x_{i-1} - x_i)\varphi_i - \int_0^1 \varphi(x)\,\mathrm{d}x = -\int_0^\epsilon \varphi(x)\,\mathrm{d}x$$

$$+ \left[\sum_{i=1}^{j} (x_{i-1} - x_i)\varphi(x_i) - \int_\epsilon^1 \varphi(x)\,\mathrm{d}x\right] \quad \text{(Quadrature Error)}$$

$$+ \left[\sum_{i=1}^{j} (x_{i-1} - x_i)\left\{\varphi_i - \varphi(x_i)\right\}\right] \quad \text{(Stochastic Error)}$$

[Dominated by Monte Carlo!]

A CLT for the Stochastic Error

The (dominating) stochastic error is $O_P(N^{-1/2})$:

$$N^{1/2}\{\text{Stochastic Error}\} \xrightarrow{\mathcal{D}} \mathcal{N}(0, V)$$

with

$$V = -\int_{s,t\in[\epsilon,1]} s\varphi'(s)t\varphi'(t)\log(s\vee t)\,\mathrm{d}s\,\mathrm{d}t.$$

[Proof based on Donsker's theorem]

The number of simulated points equals the number of iterations j, and is a multiple of N: if one stops at first iteration j such that $e^{-j/N} < \epsilon$, then: $j = N\lceil -\log\epsilon\rceil$.

A CLT for the Stochastic Error

The (dominating) stochastic error is $O_P(N^{-1/2})$:

$$N^{1/2}\{\text{Stochastic Error}\} \xrightarrow{\mathcal{D}} \mathcal{N}(0, V)$$

with

$$V = -\int_{s,t\in[\epsilon,1]} s\varphi'(s)t\varphi'(t)\log(s\vee t)\,\mathrm{d}s\,\mathrm{d}t.$$

[Proof based on Donsker's theorem]

The number of simulated points equals the number of iterations j, and is a multiple of N: if one stops at first iteration j such that $e^{-j/N} < \epsilon$, then: $j = N\lceil -\log\epsilon\rceil$.

Curse of dimension

For a simple Gaussian-Gaussian model of dimension $\dim(\theta) = d$, the following 3 quantities are $\mathrm{O}(d)$:

1. asymptotic variance of the NS estimator;
2. number of iterations (necessary to reach a given truncation error);
3. cost of one simulated sample.

Therefore, CPU time necessary for achieving error level e is

$$\mathrm{O}(d^3/e^2)$$

Curse of dimension

For a simple Gaussian-Gaussian model of dimension $\dim(\theta) = d$, the following 3 quantities are $\mathrm{O}(d)$:

1. asymptotic variance of the NS estimator;
2. number of iterations (necessary to reach a given truncation error);
3. cost of one simulated sample.

Therefore, CPU time necessary for achieving error level e is

$$\mathrm{O}(d^3/e^2)$$

Curse of dimension

For a simple Gaussian-Gaussian model of dimension $\dim(\theta) = d$, the following 3 quantities are $\mathrm{O}(d)$:

1. asymptotic variance of the NS estimator;
2. number of iterations (necessary to reach a given truncation error);
3. cost of one simulated sample.

Therefore, CPU time necessary for achieving error level e is

$$\mathrm{O}(d^3/e^2)$$

Curse of dimension

For a simple Gaussian-Gaussian model of dimension $\dim(\theta) = d$, the following 3 quantities are $\mathrm{O}(d)$:

1. asymptotic variance of the NS estimator;
2. number of iterations (necessary to reach a given truncation error);
3. cost of one simulated sample.

Therefore, CPU time necessary for achieving error level e is

$$\mathrm{O}(d^3/e^2)$$

Sampling from constr'd priors

Exact simulation from the constrained prior is intractable in most cases!

Skilling (2007) proposes to use MCMC, but:

- this introduces a bias (stopping rule).
- if MCMC stationary distribution is unconst'd prior, more and more difficult to sample points such that $L(\theta) > l$ as l increases.

If implementable, then slice sampler can be devised at the same cost!

[Thanks, Gareth!]

Sampling from constr'd priors

Exact simulation from the constrained prior is intractable in most cases!

Skilling (2007) proposes to use MCMC, but:

- this introduces a bias (stopping rule).
- if MCMC stationary distribution is unconst'd prior, more and more difficult to sample points such that $L(\theta) > l$ as l increases.

If implementable, then slice sampler can be devised at the same cost!

[Thanks, Gareth!]

Sampling from constr'd priors

Exact simulation from the constrained prior is intractable in most cases!

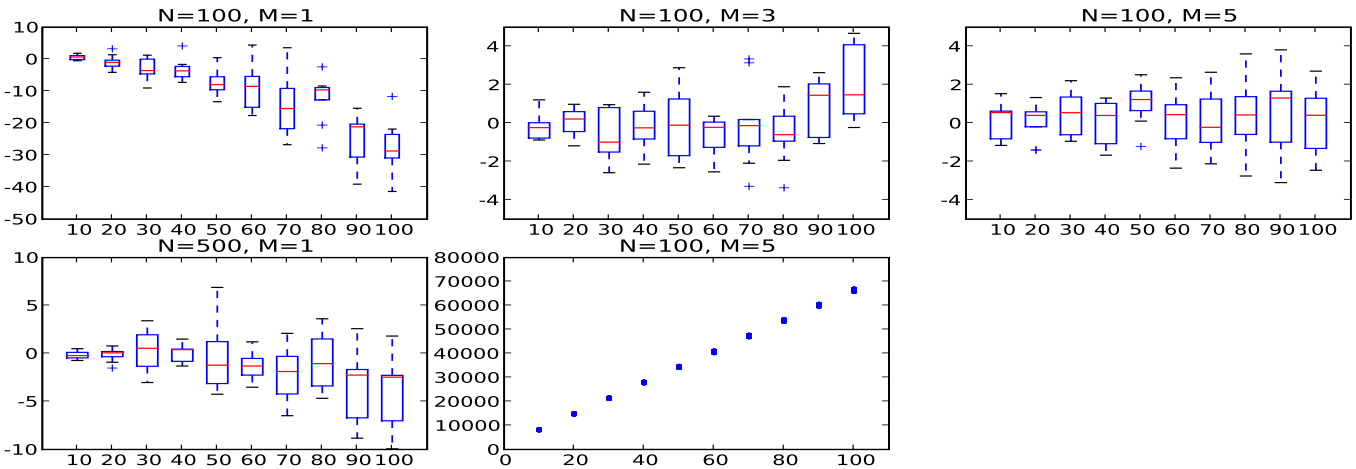
Skilling (2007) proposes to use MCMC, but:

- this introduces a bias (stopping rule).
- if MCMC stationary distribution is unconst'd prior, more and more difficult to sample points such that $L(\theta) > l$ as l increases.

If implementable, then slice sampler can be devised at the same cost!

[Thanks, Gareth!]

Illustration of MCMC bias



Log-relative error against d *(left)*, avg. number of iterations *(right)* vs dimension d, for a Gaussian-Gaussian model with d parameters, when using $T = 10$ iterations of the Gibbs sampler.

A IS variant of nested sampling

Consider instrumental prior $\widetilde{\pi}$ and likelihood $\tilde{L}$, weight function

$$w(\theta) = \frac{\pi(\theta)L(\theta)}{\widetilde{\pi}(\theta)\widetilde{L}(\theta)}$$

and weighted NS estimator

$$\widehat{\mathfrak{Z}} = \sum_{i=1}^{j} (x_{i-1} - x_i)\varphi_i w(\theta_i).$$

Then choose $(\widetilde{\pi}, \widetilde{L})$ so that sampling from $\widetilde{\pi}$ constrained to $\widetilde{L}(\theta) > l$ is easy; e.g. $\mathcal{N}(c, I_d)$ constrained to $\|c - \theta\| < r$.

A IS variant of nested sampling

Consider instrumental prior $\widetilde{\pi}$ and likelihood $\tilde{L}$, weight function

$$w(\theta) = \frac{\pi(\theta)L(\theta)}{\widetilde{\pi}(\theta)\widetilde{L}(\theta)}$$

and weighted NS estimator

$$\widehat{\mathfrak{Z}} = \sum_{i=1}^{j}(x_{i-1} - x_i)\varphi_i w(\theta_i).$$

Then choose $(\widetilde{\pi}, \widetilde{L})$ so that sampling from $\widetilde{\pi}$ constrained to $\widetilde{L}(\theta) > l$ is easy; e.g. $\mathcal{N}(c, I_d)$ constrained to $\|c - \theta\| < r$.

Benchmark: Target distribution

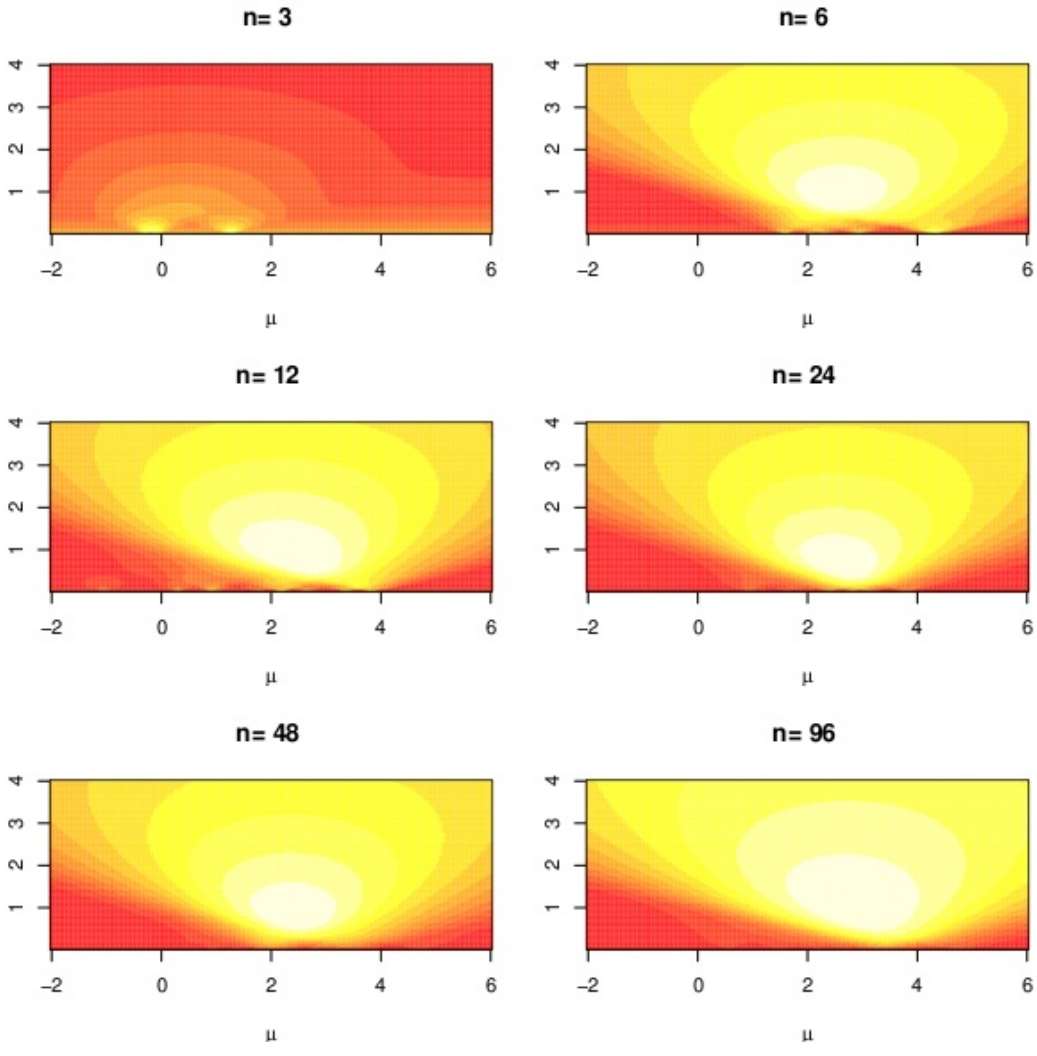
Posterior distribution on (μ, σ) associated with the mixture

$$p\mathcal{N}(0,1) + (1-p)\mathcal{N}(\mu,\sigma)\,,$$

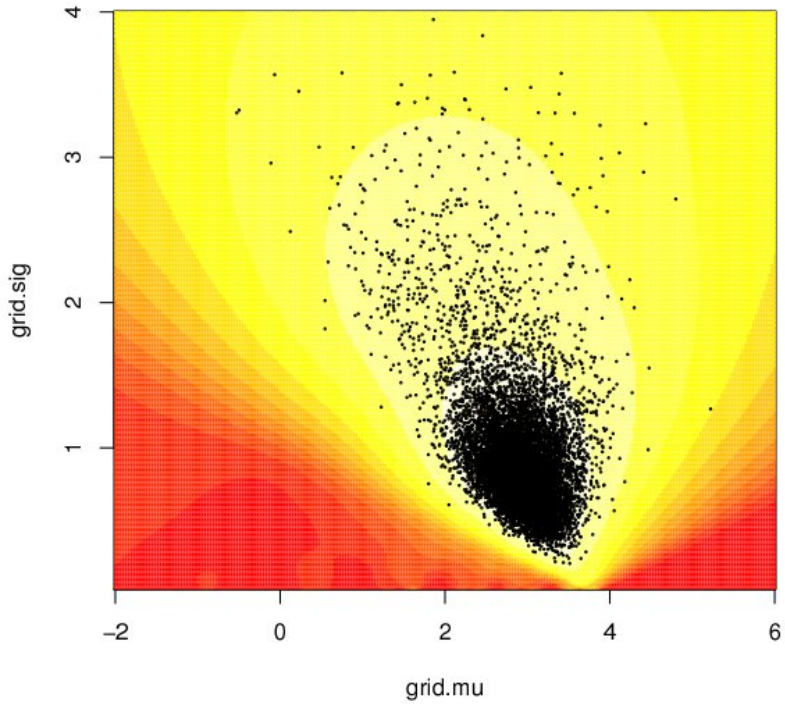
when p is known

Experiment

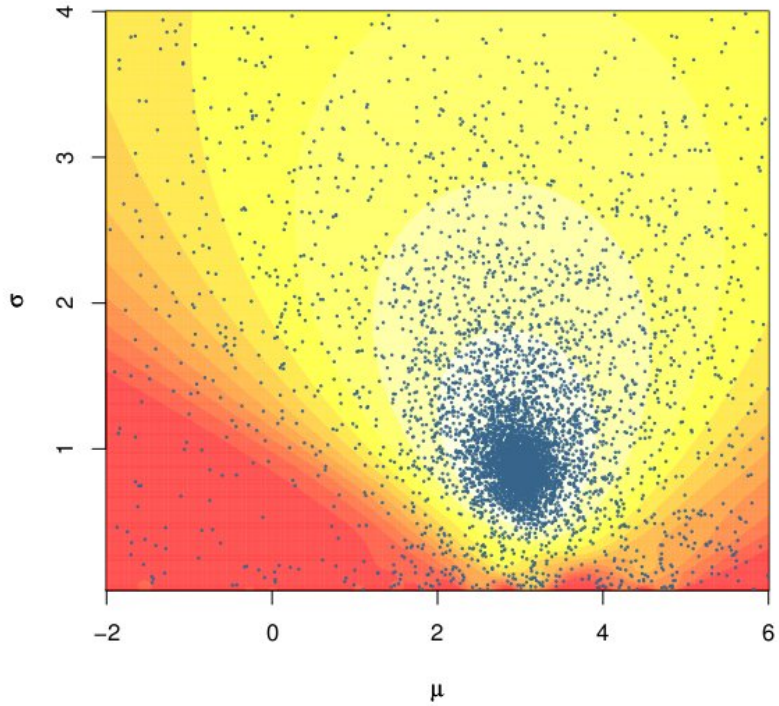
- n observations with $\mu = 2$ and $\sigma = 3/2$,
- Use of a uniform prior both on $(-2, 6)$ for μ and on $(.001, 16)$ for $\log \sigma^2$.
- occurrences of posterior bursts for $\mu = x_i$
- computation of the various estimates of $\mathfrak{Z}$



Experiment (cont'd)

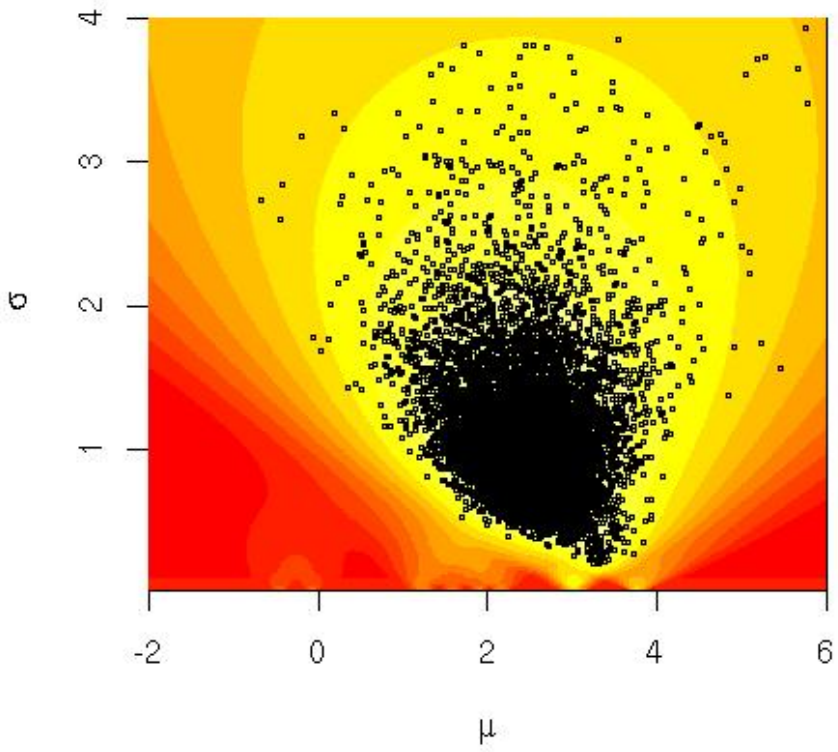


MCMC sample for $n = 16$ observations from the mixture.

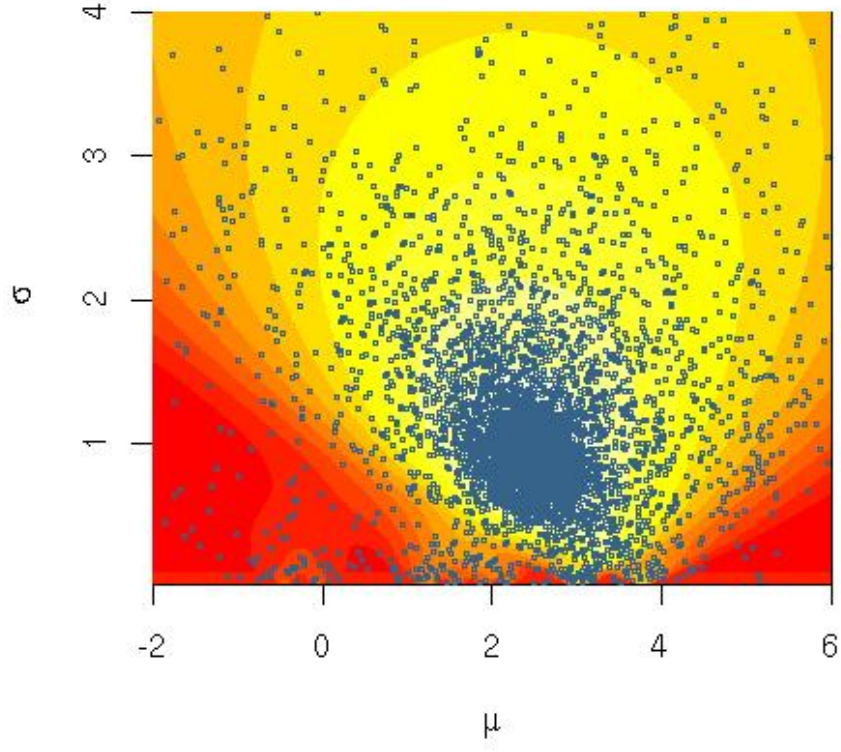


Nested sampling sequence with $M = 1000$ starting points.

Experiment (cont'd)



MCMC sample for $n = 50$ observations from the mixture.

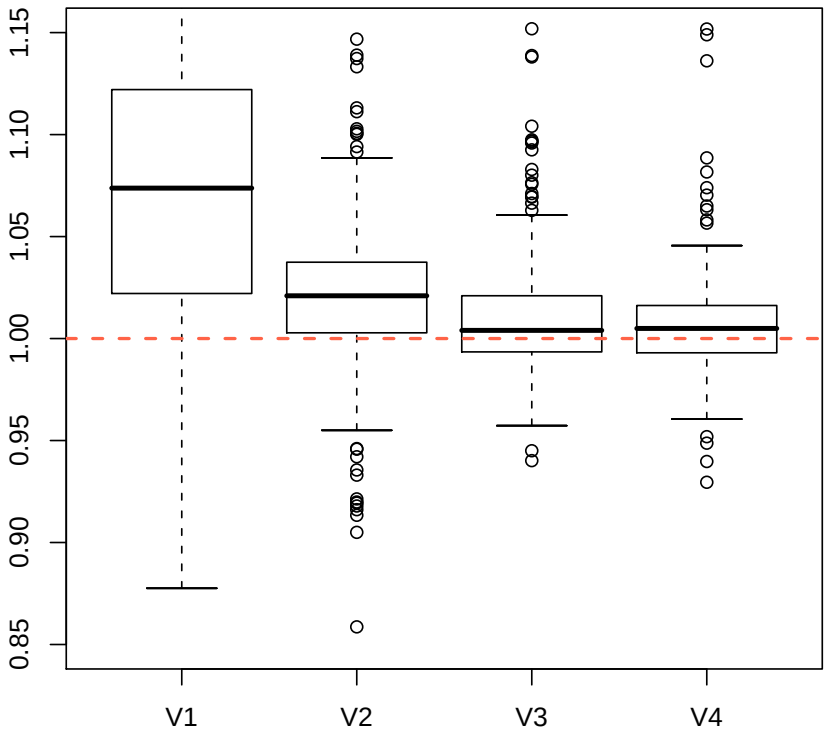


Nested sampling sequence with $M = 1000$ starting points.

Comparison

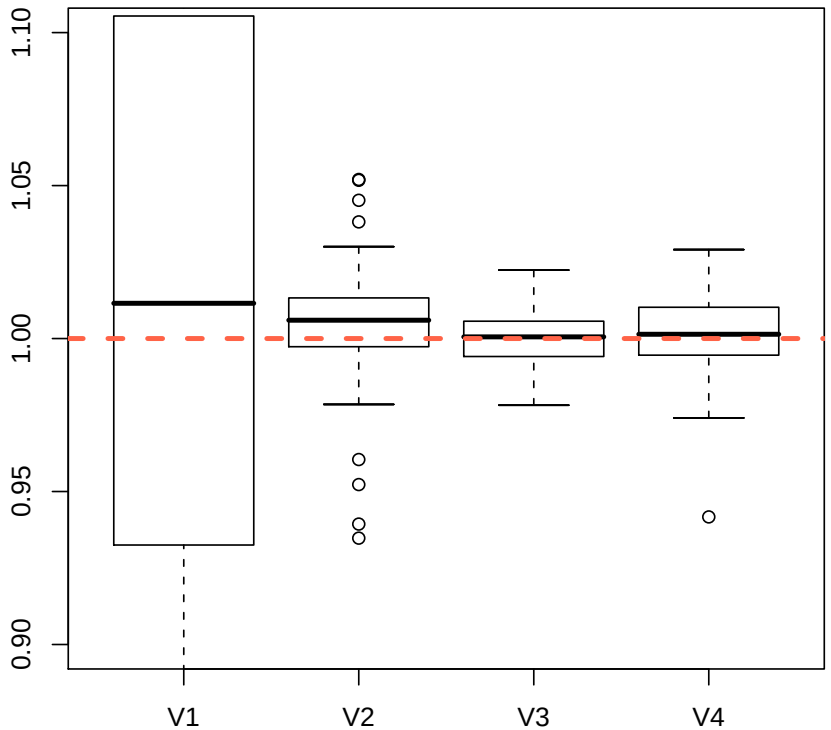
Monte Carlo and MCMC (=Gibbs) outputs based on $T = 10^4$ simulations and numerical integration based on a 850×950 grid in the (μ, σ) parameter space.
Nested sampling approximation based on a starting sample of $M = 1000$ points followed by at least 103 further simulations from the constr'd prior and a stopping rule at 95% of the observed maximum likelihood.
Constr'd prior simulation based on 50 values simulated by random walk accepting only steps leading to a lik'hood higher than the bound

Comparison (cont'd)



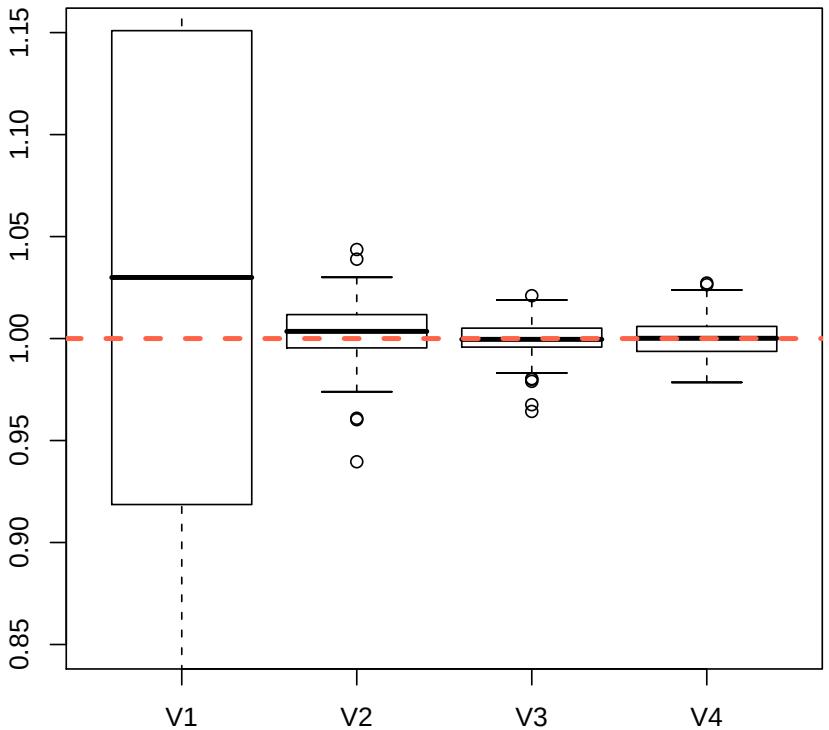
Graph based on a sample of 10 observations for $\mu = 2$ and $\sigma = 3/2$ (150 replicas).

Comparison (cont'd)



Graph based on a sample of 50 observations for $\mu = 2$ and $\sigma = 3/2$ (150 replicas).

Comparison (cont'd)



Graph based on a sample of 100 observations for $\mu = 2$ and $\sigma = 3/2$ (150 replicas).

Comparison (cont'd)

Nested sampling gets less reliable as sample size increases
Most reliable approach is mixture $\widehat{\mathfrak{Z}}_3$ although harmonic solution $\widehat{\mathfrak{Z}}_1$ close to Chib's solution [taken as golden standard]
Monte Carlo method $\widehat{\mathfrak{Z}}_2$ also producing poor approximations to $\mathfrak{Z}$
(Kernel ϕ used in $\widehat{\mathfrak{Z}}_2$ is a t non-parametric kernel estimate with standard bandwidth estimation.)

Approximate Bayesian Computation

Bayesian setting: target is $\pi(\theta)f(x|\theta)$

When likelihood $f(x|\theta)$ not in closed form, likelihood-free rejection technique:

ABC algorithm

For an observation $y \sim f(y|\theta)$, under the prior $\pi(\theta)$, keep *jointly* simulating

$$\theta' \sim \pi(\theta)\,, x \sim f(x|\theta')\,,$$

until the auxiliary variable x is equal to the observed value, $x = y$.

[Pritchard et al., 1999]

Approximate Bayesian Computation

Bayesian setting: target is $\pi(\theta)f(x|\theta)$
When likelihood $f(x|\theta)$ not in closed form, likelihood-free rejection technique:

ABC algorithm

For an observation $y \sim f(y|\theta)$, under the prior $\pi(\theta)$, keep *jointly* simulating

$$\theta' \sim \pi(\theta)\,, x \sim f(x|\theta')\,,$$

until the auxiliary variable x is equal to the observed value, $x = y$.

[Pritchard et al., 1999]

Approximate Bayesian Computation

Bayesian setting: target is $\pi(\theta)f(x|\theta)$
When likelihood $f(x|\theta)$ not in closed form, likelihood-free rejection technique:

ABC algorithm

For an observation $y \sim f(y|\theta)$, under the prior $\pi(\theta)$, keep *jointly* simulating

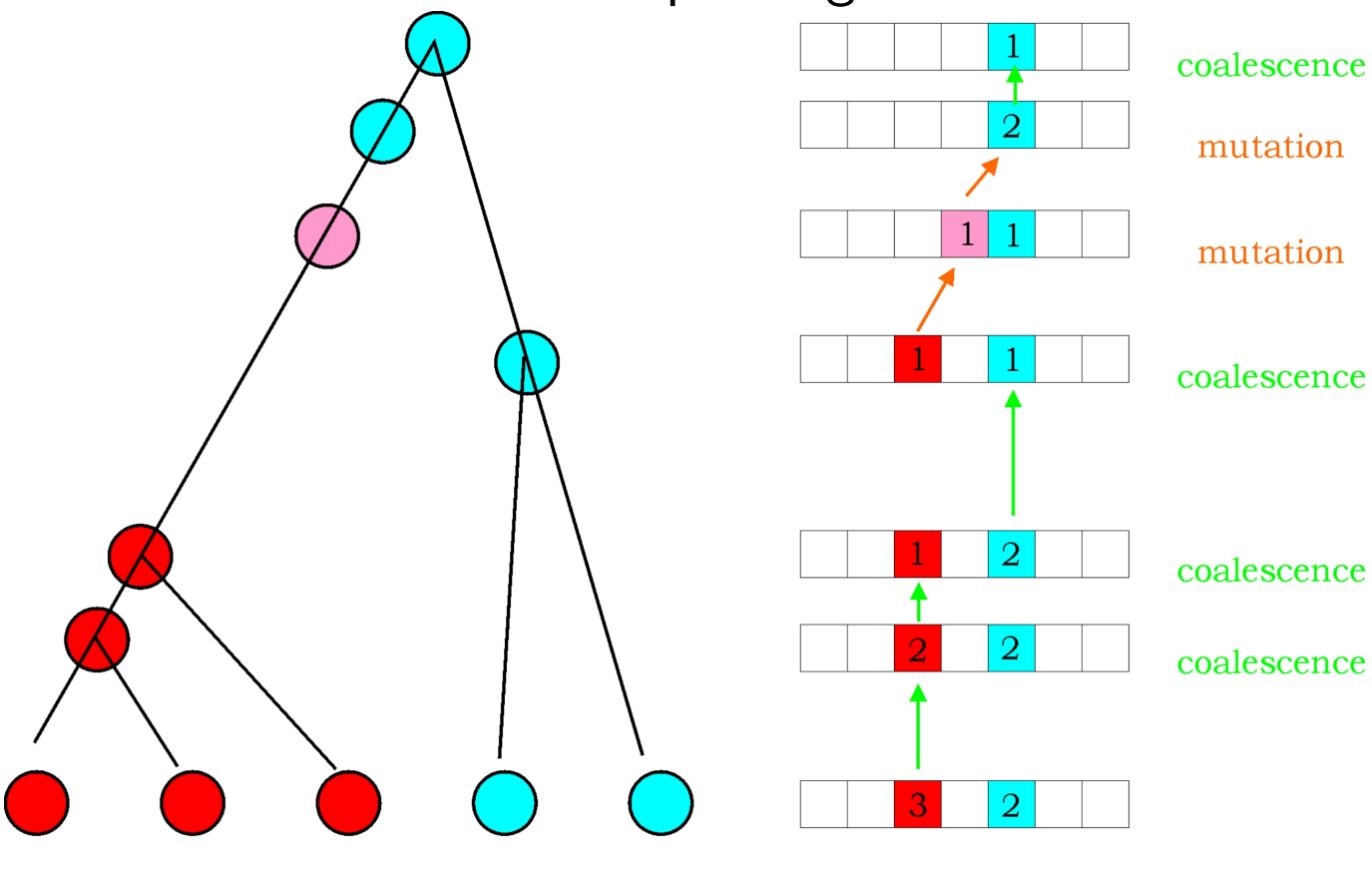
$$\theta' \sim \pi(\theta)\,, x \sim f(x|\theta')\,,$$

until the auxiliary variable x is equal to the observed value, $x = y$.

[Pritchard et al., 1999]

Population genetics example

Tree of ancestors in a sample of genes



A as approximative

When y is a continuous random variable, equality $x = y$ is replaced with a tolerance condition,

$$\varrho(x, y) \leq \epsilon$$

where ϱ is a distance between summary statistics

Output distributed from

$$\pi(\theta)\, P_\theta\{\varrho(x, y) < \epsilon\} \propto \pi(\theta|\varrho(x, y) < \epsilon)$$

A as approximative

When y is a continuous random variable, equality $x = y$ is replaced with a tolerance condition,

$$\varrho(x, y) \leq \epsilon$$

where ϱ is a distance between summary statistics
Output distributed from

$$\pi(\theta)\, P_\theta\{\varrho(x, y) < \epsilon\} \propto \pi(\theta|\varrho(x, y) < \epsilon)$$

ABC improvements

Simulating from the prior is often poor in efficiency

Either modify the proposal distribution on θ to increase the density of x's within the vicinity of y...

[Marjoram et al, 2003; Bortot et al., 2007, Sisson et al., 2007]

...or by viewing the problem as a conditional density estimation and by developing techniques to allow for larger ϵ

[Beaumont et al., 2002]

ABC improvements

Simulating from the prior is often poor in efficiency
Either modify the proposal distribution on θ to increase the density of x's within the vicinity of y...

[Marjoram et al, 2003; Bortot et al., 2007, Sisson et al., 2007]

...or by viewing the problem as a conditional density estimation and by developing techniques to allow for larger ϵ

[Beaumont et al., 2002]

ABC improvements

Simulating from the prior is often poor in efficiency
Either modify the proposal distribution on θ to increase the density of x's within the vicinity of y...

[Marjoram et al, 2003; Bortot et al., 2007, Sisson et al., 2007]

...or by viewing the problem as a conditional density estimation and by developing techniques to allow for larger ϵ

[Beaumont et al., 2002]

ABC-MCMC

Markov chain $(\theta^{(t)})$ created via the transition function

$$\theta^{(t+1)} = \begin{cases} \theta' \sim K(\theta'|\theta^{(t)}) & \text{if } x \sim f(x|\theta') \text{ is such that } x = y \\ & \text{and } u \sim \mathcal{U}(0,1) \leq \dfrac{\pi(\theta')K(\theta^{(t)}|\theta')}{\pi(\theta^{(t)})K(\theta'|\theta^{(t)})}\,, \\ \theta^{(t)} & \text{otherwise,} \end{cases}$$

has the posterior $\pi(\theta|y)$ as stationary distribution

[Marjoram et al, 2003]

ABC-MCMC

Markov chain $(\theta^{(t)})$ created via the transition function

$$\theta^{(t+1)} = \begin{cases} \theta' \sim K(\theta'|\theta^{(t)}) & \text{if } x \sim f(x|\theta') \text{ is such that } x = y \\ & \text{and } u \sim \mathcal{U}(0,1) \leq \dfrac{\pi(\theta')K(\theta^{(t)}|\theta')}{\pi(\theta^{(t)})K(\theta'|\theta^{(t)})}\,, \\ \theta^{(t)} & \text{otherwise,} \end{cases}$$

has the posterior $\pi(\theta|y)$ as stationary distribution

[Marjoram et al, 2003]

ABC-PRC

Another sequential version producing a sequence of Markov transition kernels K_t and of samples $(\theta_1^{(t)}, \ldots, \theta_N^{(t)})$ $(1 \leq t \leq T)$

ABC-PRC Algorithm

1. Pick a $\theta^\star$ is selected at random among the previous $\theta_i^{(t-1)}$'s with probabilities $\omega_i^{(t-1)}$ $(1 \leq i \leq N)$.
2. Generate
$$\theta_i^{(t)} \sim K_t(\theta|\theta^\star)\,, x \sim f(x|\theta_i^{(t)})\,,$$
3. Check that $\varrho(x, y) < \epsilon$, otherwise start again.

[Sisson et al., 2007]

ABC-PRC

Another sequential version producing a sequence of Markov transition kernels K_t and of samples $(\theta_1^{(t)}, \ldots, \theta_N^{(t)})$ $(1 \leq t \leq T)$

ABC-PRC Algorithm

1. Pick a $\theta^\star$ is selected at random among the previous $\theta_i^{(t-1)}$'s with probabilities $\omega_i^{(t-1)}$ $(1 \leq i \leq N)$.
2. Generate
$$\theta_i^{(t)} \sim K_t(\theta|\theta^\star)\,, x \sim f(x|\theta_i^{(t)})\,,$$
3. Check that $\varrho(x, y) < \epsilon$, otherwise start again.

[Sisson et al., 2007]

ABC-PRC weight

Probability $\omega_i^{(t)}$ computed as

$$\omega_i^{(t)} \propto \pi(\theta_i^{(t)}) L_{t-1}(\theta^\star|\theta_i^{(t)}) \{\pi(\theta^\star) K_t(\theta_i^{(t)}|\theta^\star)\}^{-1} ,$$

where L_{t-1} is an arbitrary transition kernel.

In case

$$L_{t-1}(\theta'|\theta) = K_t(\theta|\theta') ,$$

all weights are equal under a uniform prior.

Inspired from Del Moral et al. (2006), who use backward kernels L_{t-1} in SMC to achieve unbiasedness

ABC-PRC weight

Probability $\omega_i^{(t)}$ computed as

$$\omega_i^{(t)} \propto \pi(\theta_i^{(t)}) L_{t-1}(\theta^\star|\theta_i^{(t)}) \{\pi(\theta^\star) K_t(\theta_i^{(t)}|\theta^\star)\}^{-1} ,$$

where L_{t-1} is an arbitrary transition kernel.
In case

$$L_{t-1}(\theta'|\theta) = K_t(\theta|\theta') ,$$

all weights are equal under a uniform prior.
Inspired from Del Moral et al. (2006), who use backward kernels L_{t-1} in SMC to achieve unbiasedness

ABC-PRC weight

Probability $\omega_i^{(t)}$ computed as

$$\omega_i^{(t)} \propto \pi(\theta_i^{(t)}) L_{t-1}(\theta^\star|\theta_i^{(t)}) \{\pi(\theta^\star) K_t(\theta_i^{(t)}|\theta^\star)\}^{-1} ,$$

where L_{t-1} is an arbitrary transition kernel.
In case

$$L_{t-1}(\theta'|\theta) = K_t(\theta|\theta') ,$$

all weights are equal under a uniform prior.
Inspired from Del Moral et al. (2006), who use backward kernels L_{t-1} in SMC to achieve unbiasedness

ABC-PRC bias

Lack of unbiasedness of the method

Joint density of the accepted pair $(\theta^{(t-1)}, \theta^{(t)})$ proportional to

$$\pi(\theta^{(t-1)}|y) K_t(\theta^{(t)}|\theta^{(t-1)}) f(y|\theta^{(t)}) \,,$$

For an arbitrary function $h(\theta)$, $\mathbb{E}[\omega_t h(\theta^{(t)})]$ proportional to

$$\begin{aligned}
&\iint h(\theta^{(t)}) \, \frac{\pi(\theta^{(t)}) L_{t-1}(\theta^{(t-1)}|\theta^{(t)})}{\pi(\theta^{(t-1)}) K_t(\theta^{(t)}|\theta^{(t-1)})} \, \pi(\theta^{(t-1)}|y) K_t(\theta^{(t)}|\theta^{(t-1)}) f(y|\theta^{(t)}) \mathrm{d}\theta^{(t-1)} \mathrm{d}\theta^{(t)} \\
&\quad \propto \iint h(\theta^{(t)}) \, \frac{\pi(\theta^{(t)}) L_{t-1}(\theta^{(t-1)}|\theta^{(t)})}{\pi(\theta^{(t-1)}) K_t(\theta^{(t)}|\theta^{(t-1)})} \, \pi(\theta^{(t-1)}) f(y|\theta^{(t-1)}) \\
&\qquad \times K_t(\theta^{(t)}|\theta^{(t-1)}) f(y|\theta^{(t)}) \mathrm{d}\theta^{(t-1)} \mathrm{d}\theta^{(t)} \\
&\quad \propto \int h(\theta^{(t)}) \pi(\theta^{(t)}|y) \left\{ \int L_{t-1}(\theta^{(t-1)}|\theta^{(t)}) f(y|\theta^{(t-1)}) \mathrm{d}\theta^{(t-1)} \right\} \mathrm{d}\theta^{(t)} \,.
\end{aligned}$$

ABC-PRC bias

Lack of unbiasedness of the method

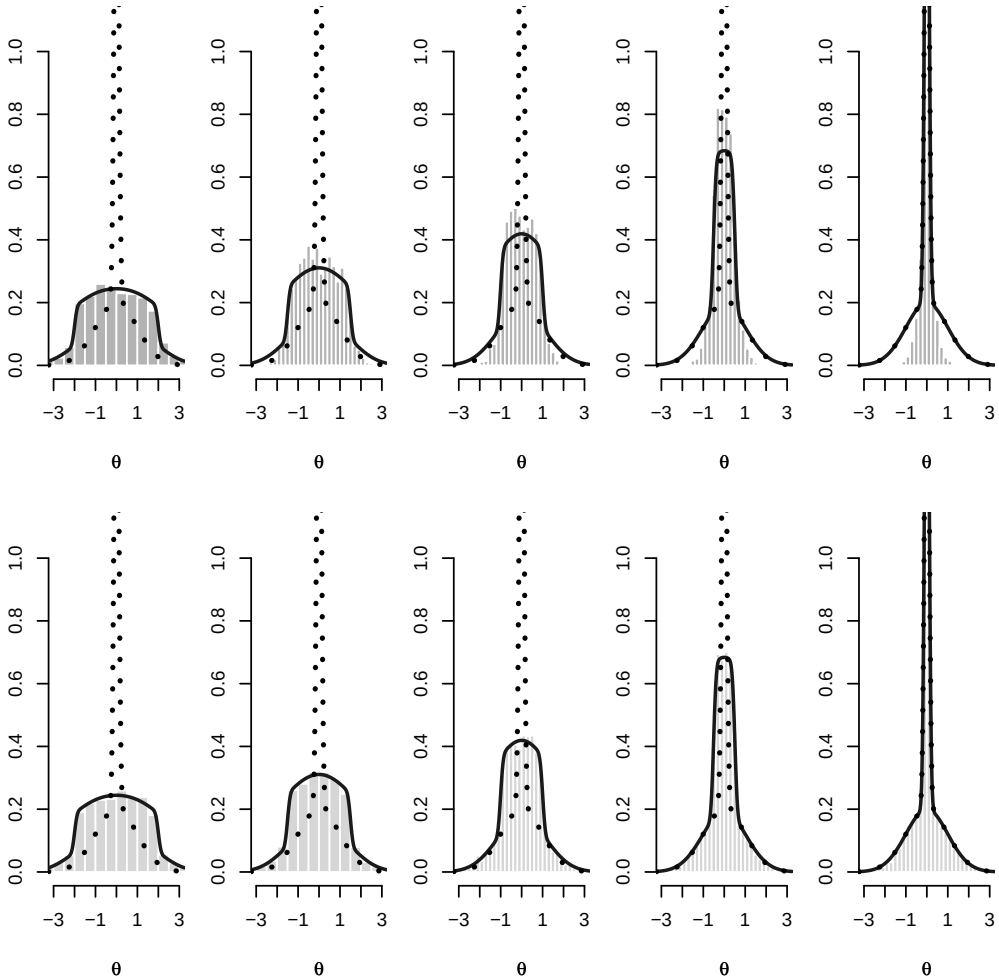
Joint density of the accepted pair $(\theta^{(t-1)}, \theta^{(t)})$ proportional to

$$\pi(\theta^{(t-1)}|y) K_t(\theta^{(t)}|\theta^{(t-1)}) f(y|\theta^{(t)}) \,,$$

For an arbitrary function $h(\theta)$, $\mathbb{E}[\omega_t h(\theta^{(t)})]$ proportional to

$$\begin{aligned}
&\iint h(\theta^{(t)}) \frac{\pi(\theta^{(t)}) L_{t-1}(\theta^{(t-1)}|\theta^{(t)})}{\pi(\theta^{(t-1)}) K_t(\theta^{(t)}|\theta^{(t-1)})} \pi(\theta^{(t-1)}|y) K_t(\theta^{(t)}|\theta^{(t-1)}) f(y|\theta^{(t)}) \mathrm{d}\theta^{(t-1)} \mathrm{d}\theta^{(t)} \\
&\quad \propto \iint h(\theta^{(t)}) \frac{\pi(\theta^{(t)}) L_{t-1}(\theta^{(t-1)}|\theta^{(t)})}{\pi(\theta^{(t-1)}) K_t(\theta^{(t)}|\theta^{(t-1)})} \pi(\theta^{(t-1)}) f(y|\theta^{(t-1)}) \\
&\qquad \times K_t(\theta^{(t)}|\theta^{(t-1)}) f(y|\theta^{(t)}) \mathrm{d}\theta^{(t-1)} \mathrm{d}\theta^{(t)} \\
&\quad \propto \int h(\theta^{(t)}) \pi(\theta^{(t)}|y) \left\{ \int L_{t-1}(\theta^{(t-1)}|\theta^{(t)}) f(y|\theta^{(t-1)}) \mathrm{d}\theta^{(t-1)} \right\} \mathrm{d}\theta^{(t)} \,.
\end{aligned}$$

A mixture example



Comparison of $\tau = 0.15$ and $\tau = 1/0.15$ in K_t

A PMC version

Use of the same kernel idea as ABC-PRC but with IS correction
Generate a sample at iteration t by

$$\hat{\pi}_t(\theta^{(t)}) \propto \sum_{j=1}^{N} \omega_j^{(t-1)} K_t(\theta^{(t)}|\theta_j^{(t-1)})$$

modulo acceptance of the associated x_t, and use an importance weight associated with an accepted simulation $\theta_i^{(t)}$

$$\omega_i^{(t)} \propto \pi(\theta_i^{(t)}) / \hat{\pi}_t(\theta_i^{(t)}) \,.$$

© **Still likelihood free**

[Beaumont et al., 2008, arXiv:0805.2256]

The ABC-PMC algorithm

Given a decreasing sequence of approximation levels $\epsilon_1 \geq \ldots \geq \epsilon_T$,

1. At iteration $t = 1$,

   For $i = 1, ..., N$

   Simulate $\theta_i^{(1)} \sim \pi(\theta)$ and $x \sim f(x|\theta_i^{(1)})$ until $\varrho(x, y) < \epsilon_1$
   Set $\omega_i^{(1)} = 1/N$

   Take τ^2 as twice the empirical variance of the $\theta_i^{(1)}$'s

2. At iteration $2 \leq t \leq T$,

   For $i = 1, ..., N$, repeat

   Pick $\theta_i^\star$ from the $\theta_j^{(t-1)}$'s with probabilities $\omega_j^{(t-1)}$
   generate $\theta_i^{(t)}|\theta_i^\star \sim \mathcal{N}(\theta_i^\star, \sigma_t^2)$ and $x \sim f(x|\theta_i^{(t)})$

   until $\varrho(x, y) < \epsilon_t$
   Set $\omega_i^{(t)} \propto \pi(\theta_i^{(t)}) / \sum_{j=1}^N \omega_j^{(t-1)} \varphi\left(\sigma_t^{-1}\left\{\theta_i^{(t)} - \theta_j^{(t-1)}\right\}\right)$

   Take τ_{t+1}^2 as twice the weighted empirical variance of the $\theta_i^{(t)}$'s

A mixture example (0)

Toy model of Sisson et al. (2007): if

$$\theta \sim \mathcal{U}(-10, 10)\,, \quad x|\theta \sim 0.5\,\mathcal{N}(\theta, 1) + 0.5\,\mathcal{N}(\theta, 1/100)\,,$$

then the posterior distribution associated with $y = 0$ is the normal mixture

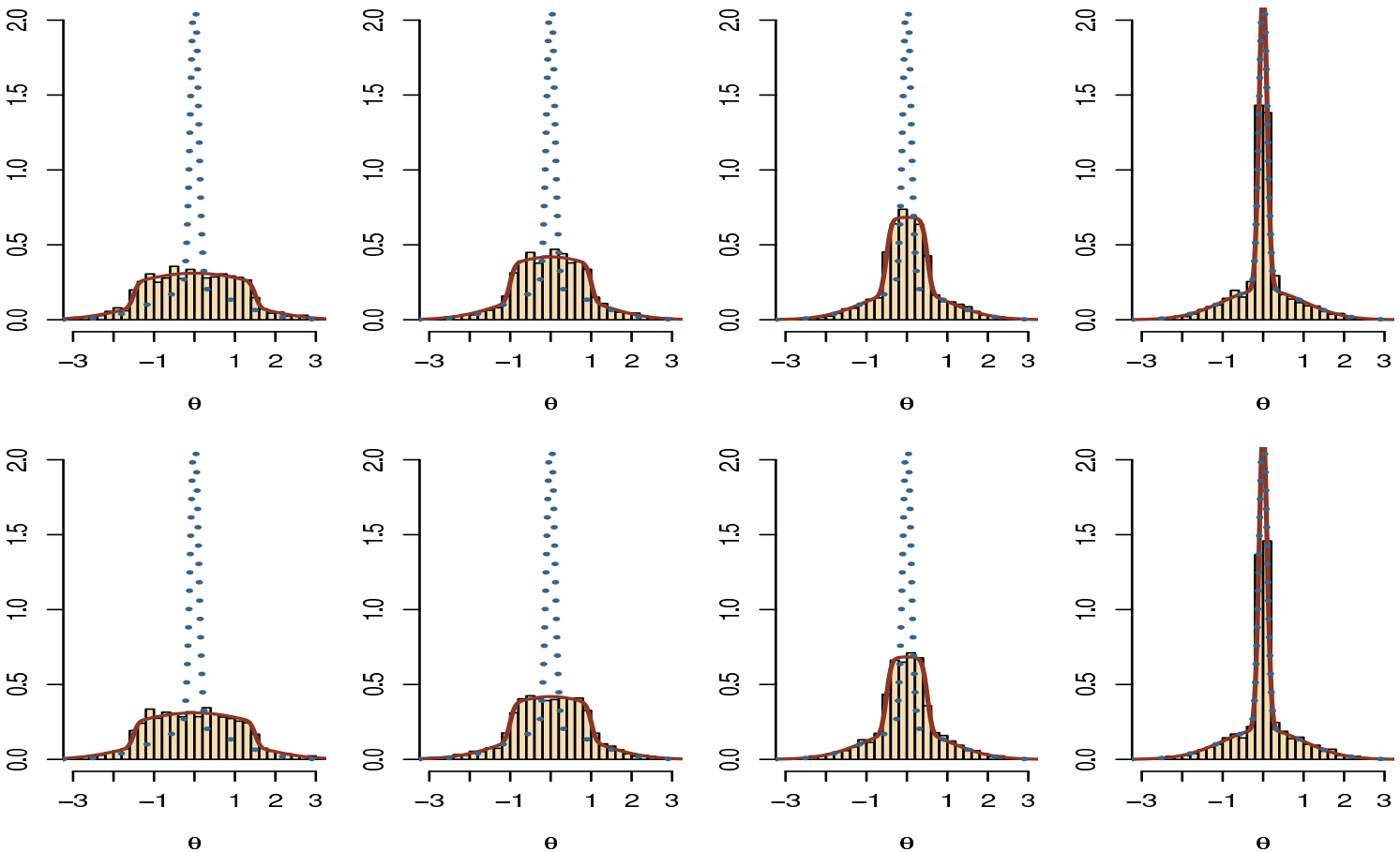
$$\theta|y = 0 \sim 0.5\,\mathcal{N}(0, 1) + 0.5\,\mathcal{N}(0, 1/100)$$

restricted to $[-10, 10]$.
Furthermore, true target available as

$$\pi(\theta||x| < \epsilon) \propto \Phi(\epsilon - \theta) - \Phi(-\epsilon - \theta) + \Phi(10(\epsilon - \theta)) - \Phi(-10(\epsilon + \theta))\,.$$

A mixture example (2)

Recovery of the target, whether using a fixed standard deviation of $\tau = 0.15$ or $\tau = 1/0.15$, or a sequence of adaptive τ_t's.



Gibbs random fields

Gibbs distribution

The rv $\mathbf{y} = (y_1, \ldots, y_n)$ is a **Gibbs random field** associated with the graph $\mathfrak{G}$ if

$$f(\mathbf{y}) = \frac{1}{\mathfrak{Z}} \exp\left\{ -\sum_{c \in \mathcal{C}} V_c(\mathbf{y}_c) \right\},$$

where $\mathfrak{Z}$ is the normalising constant, $\mathcal{C}$ is the set of cliques of $\mathfrak{G}$ and V_c is any function also called **potential**
$U(\mathbf{y}) = \sum_{c \in \mathcal{C}} V_c(\mathbf{y}_c)$ is the **energy function**

© $\mathfrak{Z}$ is usually unavailable in closed form

Gibbs random fields

Gibbs distribution

The rv $\mathbf{y} = (y_1, \ldots, y_n)$ is a **Gibbs random field** associated with the graph $\mathfrak{G}$ if

$$f(\mathbf{y}) = \frac{1}{\mathfrak{Z}} \exp\left\{ -\sum_{c \in \mathcal{C}} V_c(\mathbf{y}_c) \right\},$$

where $\mathfrak{Z}$ is the normalising constant, $\mathcal{C}$ is the set of cliques of $\mathfrak{G}$ and V_c is any function also called **potential**
$U(\mathbf{y}) = \sum_{c \in \mathcal{C}} V_c(\mathbf{y}_c)$ is the **energy function**

© **$\mathfrak{Z}$ is usually unavailable in closed form**

Potts model

Potts model

$V_c(\mathbf{y})$ is of the form

$$V_c(\mathbf{y}) = \theta S(\mathbf{y}) = \theta \sum_{l \sim i} \delta_{y_l = y_i}$$

where $l \sim i$ denotes a neighbourhood structure

In most realistic settings, summation

$$Z_\theta = \sum_{\mathbf{x} \in \mathcal{X}} \exp\{\theta^\mathsf{T} S(\mathbf{x})\}$$

involves too many terms to be manageable and numerical approximations cannot always be trusted

[Cucala, Marin, CPR & Titterington, 2009]

Potts model

Potts model

$V_c(\mathbf{y})$ is of the form

$$V_c(\mathbf{y}) = \theta S(\mathbf{y}) = \theta \sum_{l \sim i} \delta_{y_l = y_i}$$

where $l \sim i$ denotes a neighbourhood structure

In most realistic settings, summation

$$Z_\theta = \sum_{\mathbf{x} \in \mathcal{X}} \exp\{\theta^\mathsf{T} S(\mathbf{x})\}$$

involves too many terms to be manageable and numerical approximations cannot always be trusted

[Cucala, Marin, CPR & Titterington, 2009]

Bayesian Model Choice

Comparing a model with potential S_0 taking values in $\mathbb{R}^{p_0}$ versus a model with potential S_1 taking values in $\mathbb{R}^{p_1}$ can be done through the Bayes factor corresponding to the priors π_0 and π_1 on each parameter space

$$\mathfrak{B}_{m_0/m_1}(\mathbf{x}) = \frac{\int \exp\{\theta_0^\mathsf{T} S_0(\mathbf{x})\}/Z_{\theta_0,0}\pi_0(\mathrm{d}\theta_0)}{\int \exp\{\theta_1^\mathsf{T} S_1(\mathbf{x})\}/Z_{\theta_1,1}\pi_1(\mathrm{d}\theta_1)}$$

Use of Jeffreys' scale to select most appropriate model

Bayesian Model Choice

Comparing a model with potential S_0 taking values in $\mathbb{R}^{p_0}$ versus a model with potential S_1 taking values in $\mathbb{R}^{p_1}$ can be done through the Bayes factor corresponding to the priors π_0 and π_1 on each parameter space

$$\mathfrak{B}_{m_0/m_1}(\mathbf{x}) = \frac{\int \exp\{\theta_0^\mathsf{T} S_0(\mathbf{x})\}/Z_{\theta_0,0}\pi_0(\mathrm{d}\theta_0)}{\int \exp\{\theta_1^\mathsf{T} S_1(\mathbf{x})\}/Z_{\theta_1,1}\pi_1(\mathrm{d}\theta_1)}$$

Use of Jeffreys' scale to select most appropriate model

Neighbourhood relations

Choice to be made between M neighbourhood relations

$$i \overset{m}{\sim} i' \qquad (0 \leq m \leq M-1)$$

with

$$S_m(\mathbf{x}) = \sum_{i \overset{m}{\sim} i'} \mathbb{I}_{\{x_i = x_{i'}\}}$$

driven by the posterior probabilities of the models.

Model index

Formalisation via a model index $\mathcal{M}$ that appears as a new parameter with prior distribution $\pi(\mathcal{M} = m)$ and $\pi(\theta|\mathcal{M} = m) = \pi_m(\theta_m)$

Computational target:

$$\mathbb{P}(\mathcal{M} = m|\mathbf{x}) \propto \int_{\Theta_m} f_m(\mathbf{x}|\theta_m)\pi_m(\theta_m)\,\mathrm{d}\theta_m\,\pi(\mathcal{M} = m)\,,$$

Model index

Formalisation via a model index $\mathcal{M}$ that appears as a new parameter with prior distribution $\pi(\mathcal{M} = m)$ and $\pi(\theta|\mathcal{M} = m) = \pi_m(\theta_m)$
Computational target:

$$\mathbb{P}(\mathcal{M} = m|\mathbf{x}) \propto \int_{\Theta_m} f_m(\mathbf{x}|\theta_m)\pi_m(\theta_m)\,\mathrm{d}\theta_m\,\pi(\mathcal{M} = m)\,,$$

Sufficient statistics

By definition, if $S(\mathbf{x})$ sufficient statistic for the joint parameters $(\mathcal{M}, \theta_0, \ldots, \theta_{M-1})$,

$$\mathbb{P}(\mathcal{M} = m|\mathbf{x}) = \mathbb{P}(\mathcal{M} = m|S(\mathbf{x}))\,.$$

For each model m, own sufficient statistic $S_m(\cdot)$ and $S(\cdot) = (S_0(\cdot), \ldots, S_{M-1}(\cdot))$ also sufficient.
For Gibbs random fields,

$$\begin{aligned} x|\mathcal{M} = m \sim f_m(\mathbf{x}|\theta_m) &= f_m^1(\mathbf{x}|S(\mathbf{x}))\, f_m^2(S(\mathbf{x})|\theta_m) \\ &= \frac{1}{n(S(\mathbf{x}))}\, f_m^2(S(\mathbf{x})|\theta_m) \end{aligned}$$

where

$$n(S(\mathbf{x})) = \sharp\{\tilde{\mathbf{x}} \in \mathcal{X} : S(\tilde{\mathbf{x}}) = S(\mathbf{x})\}$$

© $S(\mathbf{x})$ **is therefore also sufficient for the joint parameters**
[Specific to Gibbs random fields!]

Sufficient statistics

By definition, if $S(\mathbf{x})$ sufficient statistic for the joint parameters $(\mathcal{M}, \theta_0, \ldots, \theta_{M-1})$,

$$\mathbb{P}(\mathcal{M} = m|\mathbf{x}) = \mathbb{P}(\mathcal{M} = m|S(\mathbf{x}))\,.$$

For each model m, own sufficient statistic $S_m(\cdot)$ and $S(\cdot) = (S_0(\cdot), \ldots, S_{M-1}(\cdot))$ also sufficient.

For Gibbs random fields,

$$\begin{aligned} x|\mathcal{M} = m \sim f_m(\mathbf{x}|\theta_m) &= f_m^1(\mathbf{x}|S(\mathbf{x}))\, f_m^2(S(\mathbf{x})|\theta_m) \\ &= \frac{1}{n(S(\mathbf{x}))}\, f_m^2(S(\mathbf{x})|\theta_m) \end{aligned}$$

where

$$n(S(\mathbf{x})) = \sharp\{\tilde{\mathbf{x}} \in \mathcal{X} : S(\tilde{\mathbf{x}}) = S(\mathbf{x})\}$$

© $S(\mathbf{x})$ **is therefore also sufficient for the joint parameters**

[Specific to Gibbs random fields!]

Sufficient statistics

By definition, if $S(\mathbf{x})$ sufficient statistic for the joint parameters $(\mathcal{M}, \theta_0, \ldots, \theta_{M-1})$,

$$\mathbb{P}(\mathcal{M} = m|\mathbf{x}) = \mathbb{P}(\mathcal{M} = m|S(\mathbf{x}))\,.$$

For each model m, own sufficient statistic $S_m(\cdot)$ and $S(\cdot) = (S_0(\cdot), \ldots, S_{M-1}(\cdot))$ also sufficient.
For Gibbs random fields,

$$\begin{aligned} x|\mathcal{M} = m \sim f_m(\mathbf{x}|\theta_m) &= f_m^1(\mathbf{x}|S(\mathbf{x}))f_m^2(S(\mathbf{x})|\theta_m) \\ &= \frac{1}{n(S(\mathbf{x}))} f_m^2(S(\mathbf{x})|\theta_m) \end{aligned}$$

where

$$n(S(\mathbf{x})) = \sharp\{\tilde{\mathbf{x}} \in \mathcal{X} : S(\tilde{\mathbf{x}}) = S(\mathbf{x})\}$$

© $S(\mathbf{x})$ **is therefore also sufficient for the joint parameters**

[Specific to Gibbs random fields!]

ABC model choice Algorithm

ABC-MC

- Generate m^* from the prior $\pi(\mathcal{M} = m)$.
- Generate $\theta^*_{m^*}$ from the prior $\pi_{m^*}(\cdot)$.
- Generate x^* from the model $f_{m^*}(\cdot|\theta^*_{m^*})$.
- Compute the distance $\rho(S(\mathbf{x}^0), S(\mathbf{x}^*))$.
- Accept $(\theta^*_{m^*}, m^*)$ if $\rho(S(\mathbf{x}^0), S(\mathbf{x}^*)) < \epsilon$.

[Cornuet, Grelaud, Marin & Robert, 2008]

Note When $\epsilon = 0$ the algorithm is exact

ABC approximation to the Bayes factor

Frequency ratio:

$$\begin{aligned}\overline{BF}_{m_0/m_1}(\mathbf{x}^0) &= \frac{\hat{\mathbb{P}}(\mathcal{M}=m_0|\mathbf{x}^0)}{\hat{\mathbb{P}}(\mathcal{M}=m_1|\mathbf{x}^0)} \times \frac{\pi(\mathcal{M}=m_1)}{\pi(\mathcal{M}=m_0)} \\ &= \frac{\sharp\{m^{i*}=m_0\}}{\sharp\{m^{i*}=m_1\}} \times \frac{\pi(\mathcal{M}=m_1)}{\pi(\mathcal{M}=m_0)}\,,\end{aligned}$$

replaced with

$$\widehat{BF}_{m_0/m_1}(\mathbf{x}^0) = \frac{1+\sharp\{m^{i*}=m_0\}}{1+\sharp\{m^{i*}=m_1\}} \times \frac{\pi(\mathcal{M}=m_1)}{\pi(\mathcal{M}=m_0)}$$

to avoid indeterminacy (also Bayes estimate).

ABC approximation to the Bayes factor

Frequency ratio:

$$\begin{aligned}\overline{BF}_{m_0/m_1}(\mathbf{x}^0) &= \frac{\hat{\mathbb{P}}(\mathcal{M}=m_0|\mathbf{x}^0)}{\hat{\mathbb{P}}(\mathcal{M}=m_1|\mathbf{x}^0)} \times \frac{\pi(\mathcal{M}=m_1)}{\pi(\mathcal{M}=m_0)} \\ &= \frac{\sharp\{m^{i*}=m_0\}}{\sharp\{m^{i*}=m_1\}} \times \frac{\pi(\mathcal{M}=m_1)}{\pi(\mathcal{M}=m_0)}\,,\end{aligned}$$

replaced with

$$\widehat{BF}_{m_0/m_1}(\mathbf{x}^0) = \frac{1+\sharp\{m^{i*}=m_0\}}{1+\sharp\{m^{i*}=m_1\}} \times \frac{\pi(\mathcal{M}=m_1)}{\pi(\mathcal{M}=m_0)}$$

to avoid indeterminacy (also Bayes estimate).

Toy example

iid Bernoulli model versus two-state first-order Markov chain, i.e.

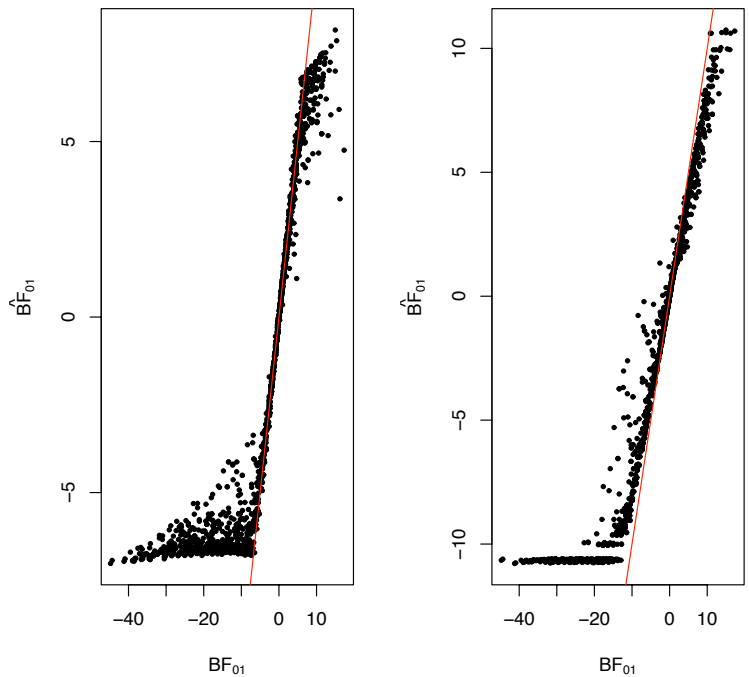
$$f_0(\mathbf{x}|\theta_0) = \exp\left(\theta_0 \sum_{i=1}^{n} \mathbb{I}_{\{x_i=1\}}\right) / \{1+\exp(\theta_0)\}^n \,,$$

versus

$$f_1(\mathbf{x}|\theta_1) = \frac{1}{2}\exp\left(\theta_1 \sum_{i=2}^{n} \mathbb{I}_{\{x_i=x_{i-1}\}}\right) / \{1+\exp(\theta_1)\}^{n-1} \,,$$

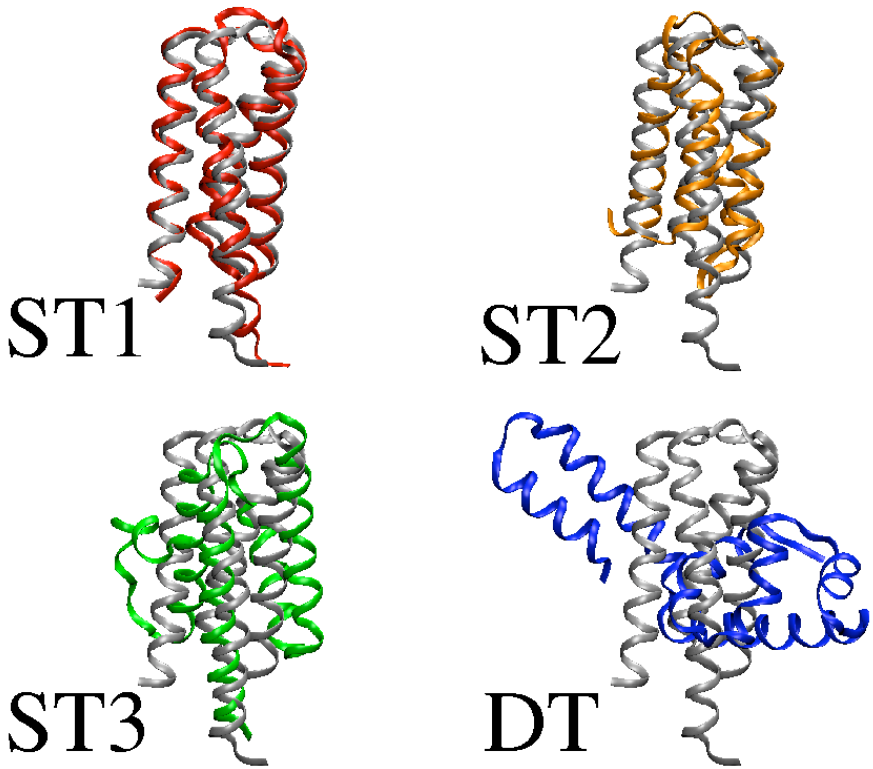
with priors $\theta_0 \sim \mathcal{U}(-5,5)$ and $\theta_1 \sim \mathcal{U}(0,6)$ (inspired by "phase transition" boundaries).

Toy example (2)



(left) Comparison of the true $BF_{m_0/m_1}(\mathbf{x}^0)$ with $\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)$ (in logs) over $2,000$ simulations and 4.10^6 proposals from the prior. *(right)* Same when using tolerance ϵ corresponding to the 1% quantile on the distances.

Protein folding



Superposition of the native structure *(grey)* with the **ST1** structure *(red.)*, the **ST2** structure *(orange)*, the **ST3** structure *(green)*, and the **DT** structure *(blue)*.

Protein folding (2)

	% seq . Id.	TM-score	FROST score
1i5nA (**ST1**)	32	0.86	75.3
1ls1A1 (**ST2**)	5	0.42	8.9
1jr8A (**ST3**)	4	0.24	8.9
1s7oA (**DT**)	10	0.08	7.8

Characteristics of dataset. *% seq. Id.*: percentage of identity with the query sequence. *TM-score.*: similarity between predicted and native structure (uncertainty between 0.17 and 0.4) *FROST score*: quality of alignment of the query onto the candidate structure (uncertainty between 7 and 9).

Protein folding (3)

	NS/ST1	**NS/ST2**	**NS/ST3**	**NS/DT**
$\widehat{BF}$	1.34	1.22	2.42	2.76
$\widehat{\mathbb{P}}(\mathcal{M} = \textbf{NS}\|\mathbf{x}^0)$	0.573	0.551	0.708	0.734

Estimates of the Bayes factors between model **NS** and models **ST1**, **ST2**, **ST3**, and **DT**, and corresponding posterior probabilities of model **NS** based on an ABC-MC algorithm using $1.2\,10^6$ simulations and a tolerance ϵ equal to the 1% quantile of the distances.