On some computational methods for Bayesian model choice

Christian P. Robert

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

Joint work with Nicolas Chopin and Jean-Michel Marin

Outline

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

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

[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 $\log_{10}(B_{10}^{\pi})$ between 0 and 0.5, evidence against H_0 *weak*,
  - if $\log_{10}(B_{10}^{\pi})$ 0.5 and 1, evidence *substantial*,
  - if $\log_{10}(B_{10}^{\pi})$ 1 and 2, evidence *strong* and
  - if $\log_{10}(B_{10}^{\pi})$ 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

Bayesian resolution

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 resolution

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 resolution

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

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

- 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 resolution

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

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

- 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*

$$\mathfrak{Z} = \int \pi(\theta) L(\theta) \, \mathrm{d}\theta,$$

aka the marginal likelihood.

Bridge sampling

If

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

live on the same space, 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]

(Further) bridge sampling

In addition

$$
\begin{aligned}
B_{12} &= \frac{\displaystyle\int \tilde\pi_2(\theta|x)\alpha(\theta)\pi_1(\theta|x)d\theta}{\displaystyle\int \tilde\pi_1(\theta|x)\alpha(\theta)\pi_2(\theta|x)d\theta} \qquad \forall\ \alpha(\cdot) \\
&\approx \frac{\displaystyle\frac{1}{n_1}\sum_{i=1}^{n_1}\tilde\pi_2(\theta_{1i}|x)\alpha(\theta_{1i})}{\displaystyle\frac{1}{n_2}\sum_{i=1}^{n_2}\tilde\pi_1(\theta_{2i}|x)\alpha(\theta_{2i})} \qquad \theta_{ji}\sim\pi_j(\theta|x)
\end{aligned}
$$

Optimal bridge sampling

The optimal choice of auxiliary function α

$$\alpha^\star = \frac{n_1 + n_2}{n_1 \tilde{\pi}_1(\theta|x) + n_2 \tilde{\pi}_2(\theta|x)}$$

leading to

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

Back later!

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

Use of the identity

$$\mathbb{E}^\pi \left[\frac{\varphi(\theta)}{\pi(\theta)L(\theta)} \middle| x \right] = \int \frac{\varphi(\theta)}{\pi(\theta)L(\theta)} \frac{\pi(\theta)L(\theta)}{\mathfrak{Z}} \,\mathrm{d}\theta = \frac{1}{\mathfrak{Z}}$$

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

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

Direct exploitation of MCMC output

▸ RE-RJ

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

Use of the identity

$$\mathbb{E}^{\pi}\left[\left.\frac{\varphi(\theta)}{\pi(\theta)L(\theta)}\right| x\right] = \int \frac{\varphi(\theta)}{\pi(\theta)L(\theta)} \frac{\pi(\theta)L(\theta)}{\mathfrak{Z}} \, \mathrm{d}\theta = \frac{1}{\mathfrak{Z}}$$

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

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

Direct exploitation of 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 $\pi(\theta)L(\theta)$ for the approximation

$$\widehat{\mathfrak{Z}_1} = 1 \Bigg/ \frac{1}{T} \sum_{t=1}^{T} \frac{\varphi(\theta^{(t)})}{\pi(\theta^{(t)})L(\theta^{(t)})}$$

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 $\pi(\theta)L(\theta)$ for the approximation

$$\widehat{\mathfrak{Z}_1} = 1 \Bigg/ \frac{1}{T} \sum_{t=1}^{T} \frac{\varphi(\theta^{(t)})}{\pi(\theta^{(t)})L(\theta^{(t)})}$$

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 $\widehat{\mathfrak{Z}_1}$ with a standard importance sampling approximation

$$\widehat{\mathfrak{Z}_2} = \frac{1}{T} \sum_{t=1}^{T} \frac{\pi(\theta^{(t)})L(\theta^{(t)})}{\varphi(\theta^{(t)})}$$

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

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

Bridge sampling redux

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

$$\tilde{\varphi}(\theta) \propto \omega_1 \pi(\theta) L(\theta) + \varphi(\theta)\,,$$

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

Note: ω_1 is not a probability weight

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

Bridge sampling redux

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

$$\tilde{\varphi}(\theta) \propto \omega_1 \pi(\theta) L(\theta) + \varphi(\theta) \,,$$

where $\varphi(\theta)$ 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

$$\omega_1 \pi(\theta^{(t-1)}) L(\theta^{(t-1)}) \Big/ \left(\omega_1 \pi(\theta^{(t-1)}) L(\theta^{(t-1)}) + \varphi(\theta^{(t-1)}) \right)$$

   and $\delta^{(t)} = 2$ otherwise;
2. If $\delta^{(t)} = 1$, generate $\theta^{(t)} \sim \text{MCMC}(\theta^{(t-1)}, \theta^{(t)})$ where $\text{MCMC}(\theta, \theta')$ denotes an arbitrary MCMC kernel associated with the posterior $\pi(\theta|x) \propto \pi(\theta) L(\theta)$;
3. If $\delta^{(t)} = 2$, generate $\theta^{(t)} \sim \varphi(\theta)$ independently

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

Corresponding MCMC (=Gibbs) sampler

At iteration t

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

$$\omega_1 \pi(\theta^{(t-1)}) L(\theta^{(t-1)}) \Big/ \left(\omega_1 \pi(\theta^{(t-1)}) L(\theta^{(t-1)}) + \varphi(\theta^{(t-1)}) \right)$$

   and $\delta^{(t)} = 2$ otherwise;
2. If $\delta^{(t)} = 1$, generate $\theta^{(t)} \sim \text{MCMC}(\theta^{(t-1)}, \theta^{(t)})$ where $\text{MCMC}(\theta, \theta')$ denotes an arbitrary MCMC kernel associated with the posterior $\pi(\theta|x) \propto \pi(\theta) L(\theta)$;
3. If $\delta^{(t)} = 2$, generate $\theta^{(t)} \sim \varphi(\theta)$ independently

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

Corresponding MCMC (=Gibbs) sampler

At iteration t

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

$$\omega_1 \pi(\theta^{(t-1)}) L(\theta^{(t-1)}) \Big/ \left(\omega_1 \pi(\theta^{(t-1)}) L(\theta^{(t-1)}) + \varphi(\theta^{(t-1)})\right)$$

   and $\delta^{(t)} = 2$ otherwise;
2. If $\delta^{(t)} = 1$, generate $\theta^{(t)} \sim \text{MCMC}(\theta^{(t-1)}, \theta^{(t)})$ where $\text{MCMC}(\theta, \theta')$ denotes an arbitrary MCMC kernel associated with the posterior $\pi(\theta|x) \propto \pi(\theta) L(\theta)$;
3. If $\delta^{(t)} = 2$, generate $\theta^{(t)} \sim \varphi(\theta)$ independently

Evidence approximation by mixtures

Rao-Blackwellised estimate

$$\hat{\xi} = \frac{1}{T} \sum_{t=1}^{T} \omega_1 \pi(\theta^{(t)}) L(\theta^{(t)}) \Big/ \omega_1 \pi(\theta^{(t)}) L(\theta^{(t)}) + \varphi(\theta^{(t)}) \,,$$

converges to $\omega_1 \mathfrak{Z} / \{\omega_1 \mathfrak{Z} + 1\}$
Deduce $\hat{\mathfrak{Z}}_3$ from $\omega_1 \hat{\mathfrak{Z}}_3 / \{\omega_1 \hat{\mathfrak{Z}}_3 + 1\} = \hat{\xi}$ ie

$$\hat{\mathfrak{Z}}_3 = \frac{\sum_{t=1}^{T} \omega_1 \pi(\theta^{(t)}) L(\theta^{(t)}) \Big/ \omega_1 \pi(\theta^{(t)}) L(\theta^{(t)}) + \varphi(\theta^{(t)})}{\sum_{t=1}^{T} \varphi(\theta^{(t)}) \Big/ \omega_1 \pi(\theta^{(t)}) L(\theta^{(t)}) + \varphi(\theta^{(t)})}$$

[Bridge sampler]

Evidence approximation by mixtures

Rao-Blackwellised estimate

$$\hat{\xi} = \frac{1}{T} \sum_{t=1}^{T} \omega_1 \pi(\theta^{(t)}) L(\theta^{(t)}) \Big/ \omega_1 \pi(\theta^{(t)}) L(\theta^{(t)}) + \varphi(\theta^{(t)}) \,,$$

converges to $\omega_1 \mathfrak{Z} / \{\omega_1 \mathfrak{Z} + 1\}$
Deduce $\hat{\mathfrak{Z}}_3$ from $\omega_1 \hat{\mathfrak{Z}}_3 / \{\omega_1 \hat{\mathfrak{Z}}_3 + 1\} = \hat{\xi}$ ie

$$\hat{\mathfrak{Z}}_3 = \frac{\sum_{t=1}^{T} \omega_1 \pi(\theta^{(t)}) L(\theta^{(t)}) \Big/ \omega_1 \pi(\theta^{(t)}) L(\theta^{(t)}) + \varphi(\theta^{(t)})}{\sum_{t=1}^{T} \varphi(\theta^{(t)}) \Big/ \omega_1 \pi(\theta^{(t)}) L(\theta^{(t)}) + \varphi(\theta^{(t)})}$$

[Bridge sampler]

Chib's representation

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

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

Use of an approximation to the posterior

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

Chib's representation

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

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

Use of an approximation to the posterior

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

Case of latent variables

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

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

where the $\mathbf{z}_k^{(t)}$'s are the latent variables simulated by a Gibbs sampler.

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

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

[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

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

[Berkhof, Mechelen, & Gelman, 2003]

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 $\mathfrak{K}_{1\to 2}(\theta_1, d\theta)$ and $\mathfrak{K}_{2\to 1}(\theta_2, d\theta)$ the corresponding kernels, under the *detailed balance condition*

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

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

Proposal expressed as

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

where $v_{1\to 2}$ is a random variable of dimension $\dim(\mathfrak{M}_2) - \dim(\mathfrak{M}_1)$, generated as

$$v_{1\to 2} \sim \varphi_{1\to 2}(v_{1\to 2})\,.$$

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 $\mathfrak{K}_{1\to 2}(\theta_1, d\theta)$ and $\mathfrak{K}_{2\to 1}(\theta_2, d\theta)$ the corresponding kernels, under the *detailed balance condition*

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

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

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

where $v_{1\to 2}$ is a random variable of dimension $\dim(\mathfrak{M}_2) - \dim(\mathfrak{M}_1)$, generated as

$$v_{1\to 2} \sim \varphi_{1\to 2}(v_{1\to 2})\,.$$

Local moves (2)

In this case, $\mathfrak{q}_{1\to 2}(\theta_1, d\theta_2)$ has density

$$\varphi_{1\to 2}(v_{1\to 2}) \left| \frac{\partial \Psi_{1\to 2}(\theta_1, v_{1\to 2})}{\partial(\theta_1, v_{1\to 2})} \right|^{-1},$$

by the Jacobian rule.

Reverse importance link

If probability $\varpi_{1\to 2}$ of choosing move to $\mathfrak{M}_2$ while in $\mathfrak{M}_1$, acceptance probability reduces to

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

©Difficult calibration

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

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

Local moves (2)

In this case, $\mathfrak{q}_{1\to 2}(\theta_1, d\theta_2)$ has density

$$\varphi_{1\to 2}(v_{1\to 2}) \left| \frac{\partial \Psi_{1\to 2}(\theta_1, v_{1\to 2})}{\partial(\theta_1, v_{1\to 2})} \right|^{-1},$$

by the Jacobian rule.

Reverse importance link

If probability $\varpi_{1\to 2}$ of choosing move to $\mathfrak{M}_2$ while in $\mathfrak{M}_1$, acceptance probability reduces to

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

©Difficult calibration

Alternative

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

- a model index M
- pseudo-priors $\pi_j(\theta_j|M=k)$ for $j \neq k$

[Carlin & Chib, 1995]

Validation by

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

where the (marginal) posterior is

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

$$= \sum_{k=1}^{D} \varrho_k \, m_k(y) \, \pi_k(\theta_k|y) \prod_{j\neq k} \pi_j(\theta_j|M=k) \,.$$

Alternative

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

- a model index M
- pseudo-priors $\pi_j(\theta_j | M = k)$ for $j \neq k$

[Carlin & Chib, 1995]

Validation by

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

where the (marginal) posterior is

$$\begin{aligned}\pi(\theta|y) &= \sum_{k=1}^{D} \pi(\theta, M = k|y) \\ &= \sum_{k=1}^{D} \varrho_k \, m_k(y) \, \pi_k(\theta_k|y) \prod_{j\neq k} \pi_j(\theta_j|M = k)\,.\end{aligned}$$

MCMC implementation

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

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

Approximate $\pi(M = k|y) = \mathfrak{z}_k$ by

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

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

MCMC implementation

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

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

Approximate $\pi(M = k|y) = \mathfrak{Z}_k$ by

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

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

MCMC implementation

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

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

Approximate $\pi(M = k|y) = \mathfrak{z}_k$ by

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

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

Scott's (2002) proposal

Suggest estimating $P(M = k|y)$ by

$$\tilde{\varrho}_k(y) \propto \varrho_k \sum_{t=1}^{T} \left\{ f_k(y|\theta_k^{(t)}) \Big/ \sum_{j=1}^{D} \varrho_j \, f_j(y|\theta_j^{(t)}) \right\} ,$$

based on D simultaneous and independent MCMC chains

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

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

Scott's (2002) proposal

Suggest estimating $P(M = k|y)$ by

$$\tilde{\varrho}_k(y) \propto \varrho_k \sum_{t=1}^{T} \left\{ f_k(y|\theta_k^{(t)}) \Big/ \sum_{j=1}^{D} \varrho_j \, f_j(y|\theta_j^{(t)}) \right\} ,$$

based on D simultaneous and independent MCMC chains

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

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

Congdon's (2006) extension

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

$$\hat{\varrho}_k(y) \propto \varrho_k \sum_{t=1}^{T} \left\{ f_k(y|\theta_k^{(t)})\pi_k(\theta_k^{(t)}) \Big/ \sum_{j=1}^{D} \varrho_j \, f_j(y|\theta_j^{(t)})\pi_j(\theta_j^{(t)}) \right\},$$

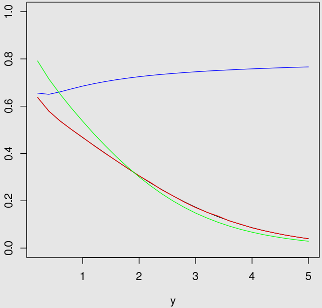
where again the $\theta_k^{(t)}$'s are MCMC chains with stationary distributions $\pi_k(\theta_k|y)$

Examples

Example (Model choice)

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

Approximations of $\pi(M = 1|y)$: Scott's (2002) (green), and Congdon's (2006) (brown) ($N = 10^6$ simulations).

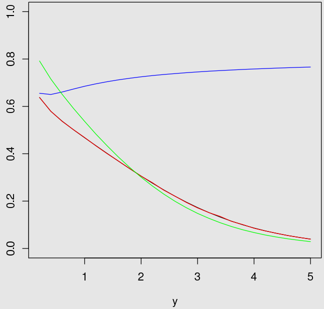


Examples

Example (Model choice)

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

Approximations of $\pi(M = 1|y)$:
Scott's (2002) (green), and
Congdon's (2006) (brown)
($N = 10^6$ simulations).

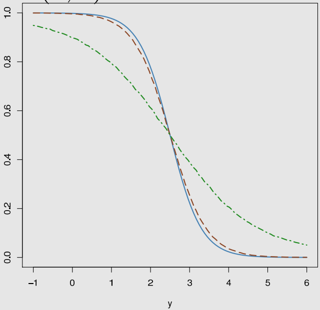


Examples (2)

Example (Model choice (2))

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

Comparison of both approximations with $\pi(M = 1|y)$: Scott's (2002) (green and mixed dashes) and Congdon's (2006) (brown and long dashes) ($N = 10^4$ simulations).

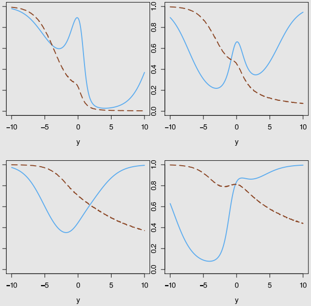


Examples (3)

Example (Model choice (3))

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

Comparison of Congdon's (2006) (brown and dashed lines) with $\pi(M = 1|y)$ 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: $\quad x_i = e^{-i/N}$
- 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

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

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

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

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

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

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

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

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

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

$$
\begin{aligned}
\text{Error} &= \widehat{\mathfrak{Z}} - \mathfrak{Z} \\
&= \sum_{i=1}^{j} (x_{i-1} - x_i)\varphi_i - \int_0^1 \varphi(x)\,\mathrm{d}x = -\int_0^{\epsilon} \varphi(x)\,\mathrm{d}x \\
&+ \left[\sum_{i=1}^{j} (x_{i-1} - x_i)\varphi(x_i) - \int_{\epsilon}^{1} \varphi(x)\,\mathrm{d}x \right] \qquad \text{(Quadrature Error)} \\
&+ \left[\sum_{i=1}^{j} (x_{i-1} - x_i)\left\{\varphi_i - \varphi(x_i)\right\} \right] \qquad \text{(Stochastic Error)}
\end{aligned}
$$

[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 $O_P(N^{-1/2})$:

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

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 $O(d)$:

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

$$O(d^3/e^2)$$

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!

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!

Sampling from constr'd priors

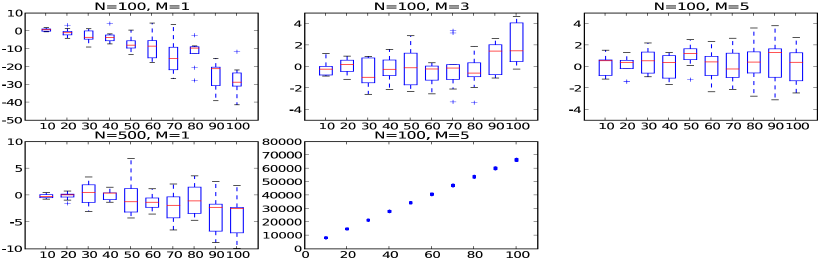
Exact simulation from the constrained prior is intractable in most cases!

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!

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

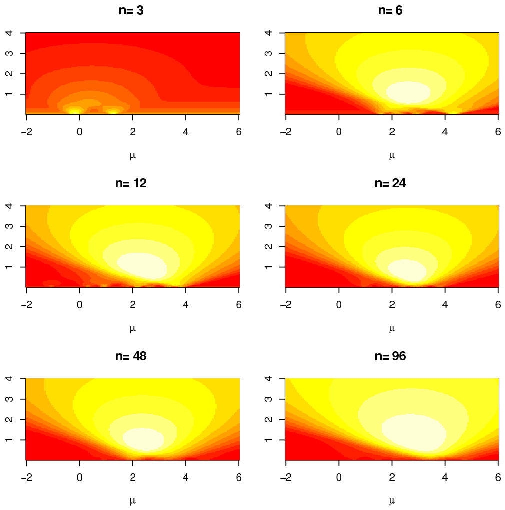
Posterior distribution on (μ, σ) associated with the mixture

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

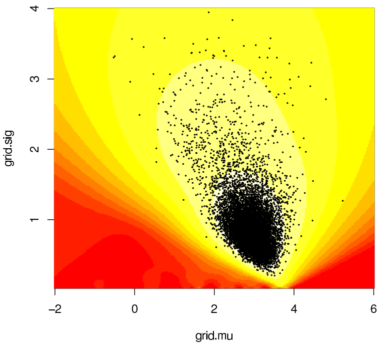
when p is known

Experiment

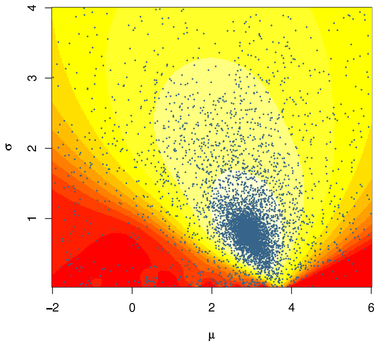
- n observations with $\mu = 2$ and $\sigma = 3/2$,
- Use of a uniform prior both on $(-2, 6)$ for μ and on $(.001, 16)$ for $\log \sigma^2$.
- occurrences of posterior bursts for $\mu = x_i$
- computation of the various estimates of $\mathfrak{Z}$



Experiment (cont'd)

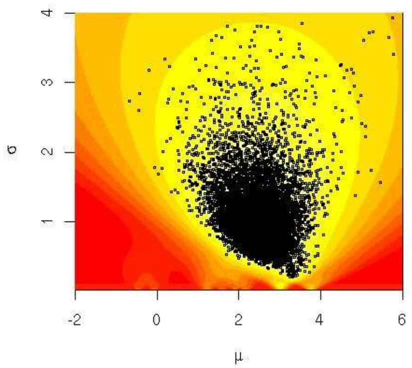


MCMC sample for $n = 16$ observations from the mixture.

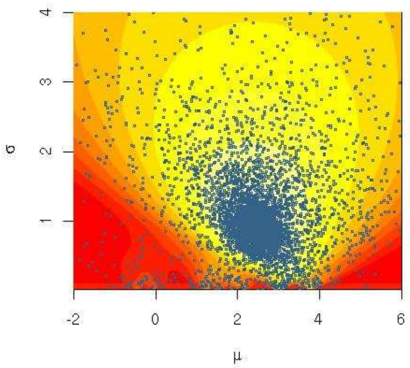


Nested sampling sequence with $M = 1000$ starting points.

Experiment (cont'd)



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

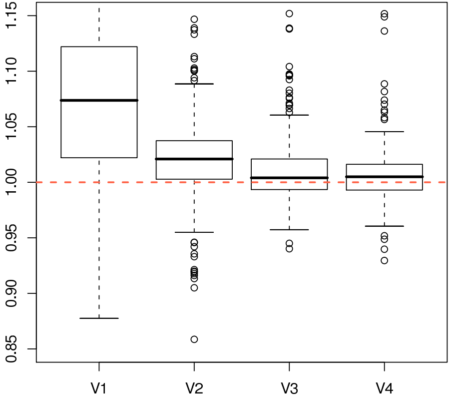


Nested sampling sequence with $M = 1000$ starting points.

Comparison

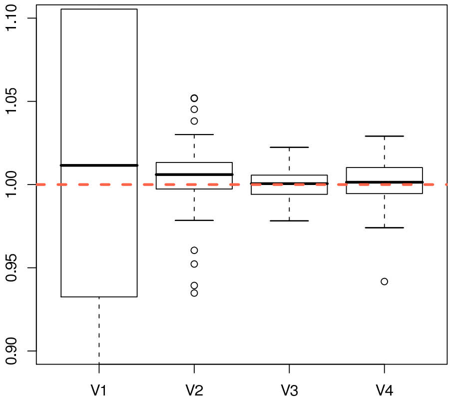
Monte Carlo and MCMC (=Gibbs) outputs based on $T = 10^4$ simulations and numerical integration based on a 850×950 grid in the (μ, σ) parameter space.
Nested sampling approximation based on a starting sample of $M = 1000$ points followed by at least 103 further simulations from the constr'd prior and 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)



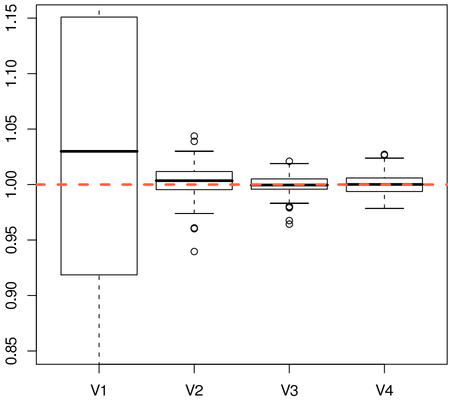
Graph based on a sample of 10 observations for $\mu = 2$ and $\sigma = 3/2$ (150 replicas).

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.)