On approximating Bayes factors

Christian P. Robert

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

CRiSM workshop on model uncertainty
U. Warwick, May 30, 2010
Joint works with N. Chopin, J.-M. Marin, K. Mengersen & D. Wraith

Outline

Model choice as 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

Model choice as 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

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,

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,

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,

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,

Bayes factor

Definition (Bayes factors)

For comparing model $\mathfrak{M}_0$ with $\theta \in \Theta_0$ vs. $\mathfrak{M}_1$ with $\theta \in \Theta_1$, under priors $\pi_0(\theta)$ and $\pi_1(\theta)$, central quantity

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

[Jeffreys, 1939]

Evidence

Problems using 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.

[Jeffreys, 1939]

A comparison of importance sampling solutions

[Marin & Robert, 2010]

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

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

Diabetes in Pima Indian women

Example (R benchmark)

"A population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix (AZ), was tested for diabetes according to WHO criteria. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases."

200 Pima Indian women with observed variables

- plasma glucose concentration in oral glucose tolerance test
- diastolic blood pressure
- diabetes pedigree function
- presence/absence of diabetes

Probit modelling on Pima Indian women

Probability of diabetes function of above variables

$$\mathbb{P}(y = 1|x) = \Phi(x_1\beta_1 + x_2\beta_2 + x_3\beta_3)\,,$$

Test of $H_0 : \beta_3 = 0$ for 200 observations of Pima.tr based on a g-prior modelling:

$$\beta \sim \mathcal{N}_3(0, n\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1})$$

[Marin & Robert, 2007]

Probit modelling on Pima Indian women

Probability of diabetes function of above variables

$$\mathbb{P}(y = 1|x) = \Phi(x_1\beta_1 + x_2\beta_2 + x_3\beta_3)\,,$$

Test of $H_0 : \beta_3 = 0$ for 200 observations of Pima.tr based on a g-prior modelling:

$$\beta \sim \mathcal{N}_3(0, n\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1})$$

[Marin & Robert, 2007]

MCMC 101 for probit models

Use of either a random walk proposal

$$\beta' = \beta + \epsilon$$

in a Metropolis-Hastings algorithm (since the likelihood is available) or of a Gibbs sampler that takes advantage of the missing/latent variable

$$z|y, x, \beta \sim \mathcal{N}(x^{\mathsf{T}}\beta, 1)\left\{\mathbb{I}^{y}_{z\geq 0} \times \mathbb{I}^{1-y}_{z\leq 0}\right\}$$

(since $\beta|y, X, z$ is distributed as a standard normal)

[Gibbs three times faster]

MCMC 101 for probit models

Use of either a random walk proposal

$$\beta' = \beta + \epsilon$$

in a Metropolis-Hastings algorithm (since the likelihood is available) or of a Gibbs sampler that takes advantage of the missing/latent variable

$$z|y, x, \beta \sim \mathcal{N}(x^{\mathsf{T}}\beta, 1) \left\{ \mathbb{I}^{y}_{z\geq 0} \times \mathbb{I}^{1-y}_{z\leq 0} \right\}$$

(since $\beta|y, X, z$ is distributed as a standard normal)

[Gibbs three times faster]

MCMC 101 for probit models

Use of either a random walk proposal

$$\beta' = \beta + \epsilon$$

in a Metropolis-Hastings algorithm (since the likelihood is available) or of a Gibbs sampler that takes advantage of the missing/latent variable

$$z|y, x, \beta \sim \mathcal{N}(x^{\mathsf{T}}\beta, 1) \left\{ \mathbb{I}^{y}_{z\geq 0} \times \mathbb{I}^{1-y}_{z\leq 0} \right\}$$

(since $\beta|y, X, z$ is distributed as a standard normal)

[Gibbs three times faster]

Importance sampling for the Pima Indian dataset

Use of the importance function inspired from the MLE estimate distribution

$$\beta \sim \mathcal{N}(\hat{\beta}, \hat{\Sigma})$$

R Importance sampling code

```
model1=summary(glm(y~-1+X1,family=binomial(link="probit")))
is1=rmvnorm(Niter,mean=model1$coeff[,1],sigma=2*model1$cov.unscaled)
is2=rmvnorm(Niter,mean=model2$coeff[,1],sigma=2*model2$cov.unscaled)
bfis=mean(exp(probitlpost(is1,y,X1)-dmvlnorm(is1,mean=model1$coeff[,1],
     sigma=2*model1$cov.unscaled))) / mean(exp(probitlpost(is2,y,X2)-
     dmvlnorm(is2,mean=model2$coeff[,1],sigma=2*model2$cov.unscaled)))
```

Importance sampling for the Pima Indian dataset

Use of the importance function inspired from the MLE estimate distribution

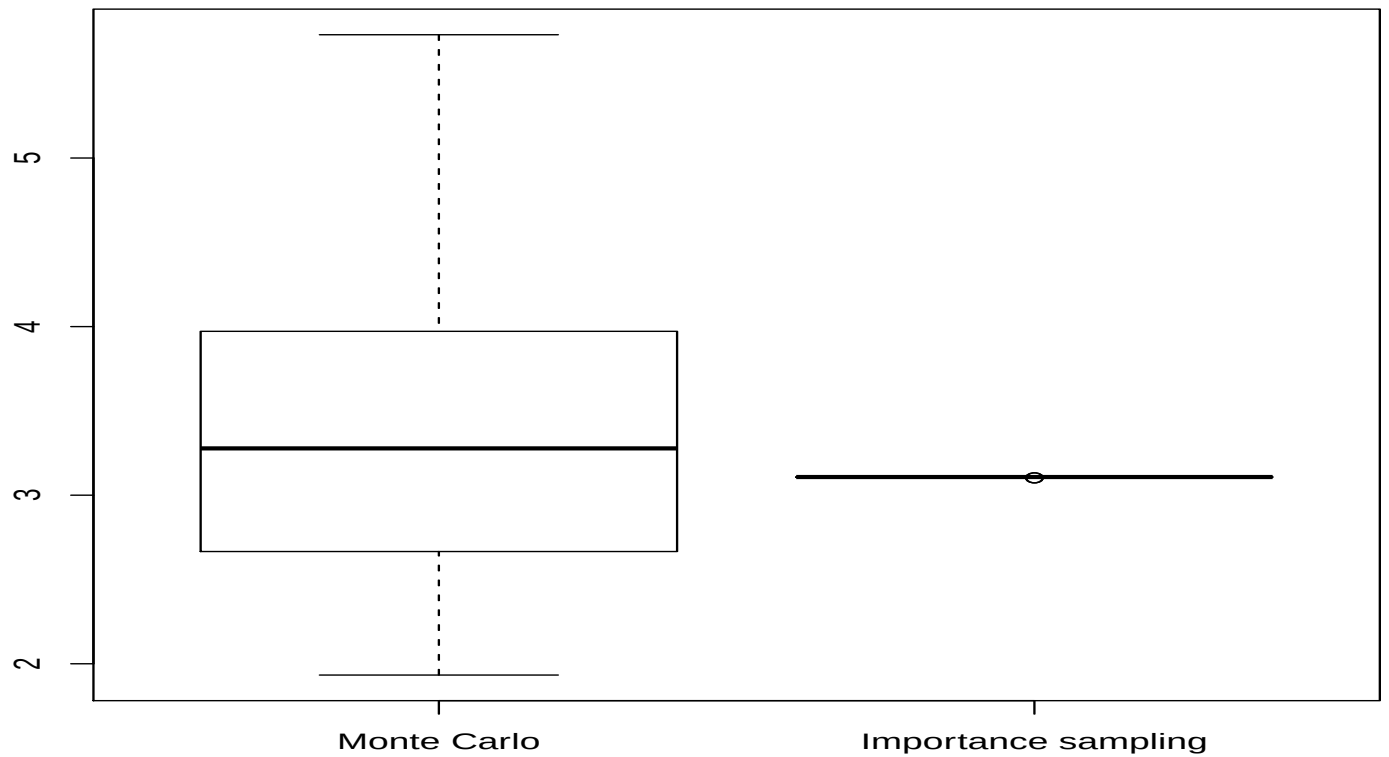
$$\beta \sim \mathcal{N}(\hat{\beta}, \hat{\Sigma})$$

R Importance sampling code

```
model1=summary(glm(y~-1+X1,family=binomial(link="probit")))
is1=rmvnorm(Niter,mean=model1$coeff[,1],sigma=2*model1$cov.unscaled)
is2=rmvnorm(Niter,mean=model2$coeff[,1],sigma=2*model2$cov.unscaled)
bfis=mean(exp(probitlpost(is1,y,X1)-dmvlnorm(is1,mean=model1$coeff[,1],
     sigma=2*model1$cov.unscaled))) / mean(exp(probitlpost(is2,y,X2)-
     dmvlnorm(is2,mean=model2$coeff[,1],sigma=2*model2$cov.unscaled)))
```

Diabetes in Pima Indian women

Comparison of the variation of the Bayes factor approximations based on 100 replicas for $20,000$ simulations from the prior and the above MLE importance sampler



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

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

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

Bridge sampling variance

The bridge sampling estimator does poorly if

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

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

(Further) bridge sampling

General identity:

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

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

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)

Drawback: 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)

Drawback: Dependence on the unknown normalising constants solved iteratively

Extension to varying dimensions

When $\dim(\Theta_1) \neq \dim(\Theta_2)$, e.g. $\theta_2 = (\theta_1, \psi)$, introduction of a *pseudo-posterior density*, $\omega(\psi|\theta_1, x)$, augmenting $\pi_1(\theta_1|x)$ into joint distribution

$$\pi_1(\theta_1|x) \times \omega(\psi|\theta_1, x)$$

on Θ_2 so that

$$B_{12} = \frac{\int \tilde{\pi}_1(\theta_1|x)\alpha(\theta_1,\psi)\pi_2(\theta_1,\psi|x)\mathrm{d}\theta_1\omega(\psi|\theta_1,x)\,\mathrm{d}\psi}{\int \tilde{\pi}_2(\theta_1,\psi|x)\alpha(\theta_1,\psi)\pi_1(\theta_1|x)\mathrm{d}\theta_1\,\omega(\psi|\theta_1,x)\,\mathrm{d}\psi}$$

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

for any conditional density $\omega(\psi|\theta_1)$ and any joint density φ.

Extension to varying dimensions

When $\dim(\Theta_1) \neq \dim(\Theta_2)$, e.g. $\theta_2 = (\theta_1, \psi)$, introduction of a *pseudo-posterior density*, $\omega(\psi|\theta_1, x)$, augmenting $\pi_1(\theta_1|x)$ into joint distribution

$$\pi_1(\theta_1|x) \times \omega(\psi|\theta_1, x)$$

on Θ_2 so that

$$B_{12} = \frac{\int \tilde{\pi}_1(\theta_1|x)\alpha(\theta_1,\psi)\pi_2(\theta_1,\psi|x)\mathrm{d}\theta_1\omega(\psi|\theta_1,x)\,\mathrm{d}\psi}{\int \tilde{\pi}_2(\theta_1,\psi|x)\alpha(\theta_1,\psi)\pi_1(\theta_1|x)\mathrm{d}\theta_1\,\omega(\psi|\theta_1,x)\,\mathrm{d}\psi}$$

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

for any conditional density $\omega(\psi|\theta_1)$ and any joint density φ.

Illustration for the Pima Indian dataset

Use of the MLE induced conditional of β_3 given (β_1, β_2) as a pseudo-posterior and mixture of both MLE approximations on β_3 in bridge sampling estimate

R bridge sampling code

```
cova=model2$cov.unscaled
expecta=model2$coeff[,1]
covw=cova[3,3]-t(cova[1:2,3])%*%ginv(cova[1:2,1:2])%*%cova[1:2,3]

probit1=hmprobit(Niter,y,X1)
probit2=hmprobit(Niter,y,X2)
pseudo=rnorm(Niter,meanw(probit1),sqrt(covw))
probit1p=cbind(probit1,pseudo)

bfbs=mean(exp(probitlpost(probit2[,1:2],y,X1)+dnorm(probit2[,3],meanw(probit2[,1:2]),
     sqrt(covw),log=T))/ (dmvnorm(probit2,expecta,cova)+dnorm(probit2[,3],expecta[3],
     cova[3,3])))/ mean(exp(probitlpost(probit1p,y,X2))/(dmvnorm(probit1p,expecta,cova)+
     dnorm(pseudo,expecta[3],cova[3,3])))
```

Illustration for the Pima Indian dataset

Use of the MLE induced conditional of β_3 given (β_1, β_2) as a pseudo-posterior and mixture of both MLE approximations on β_3 in bridge sampling estimate

R bridge sampling code

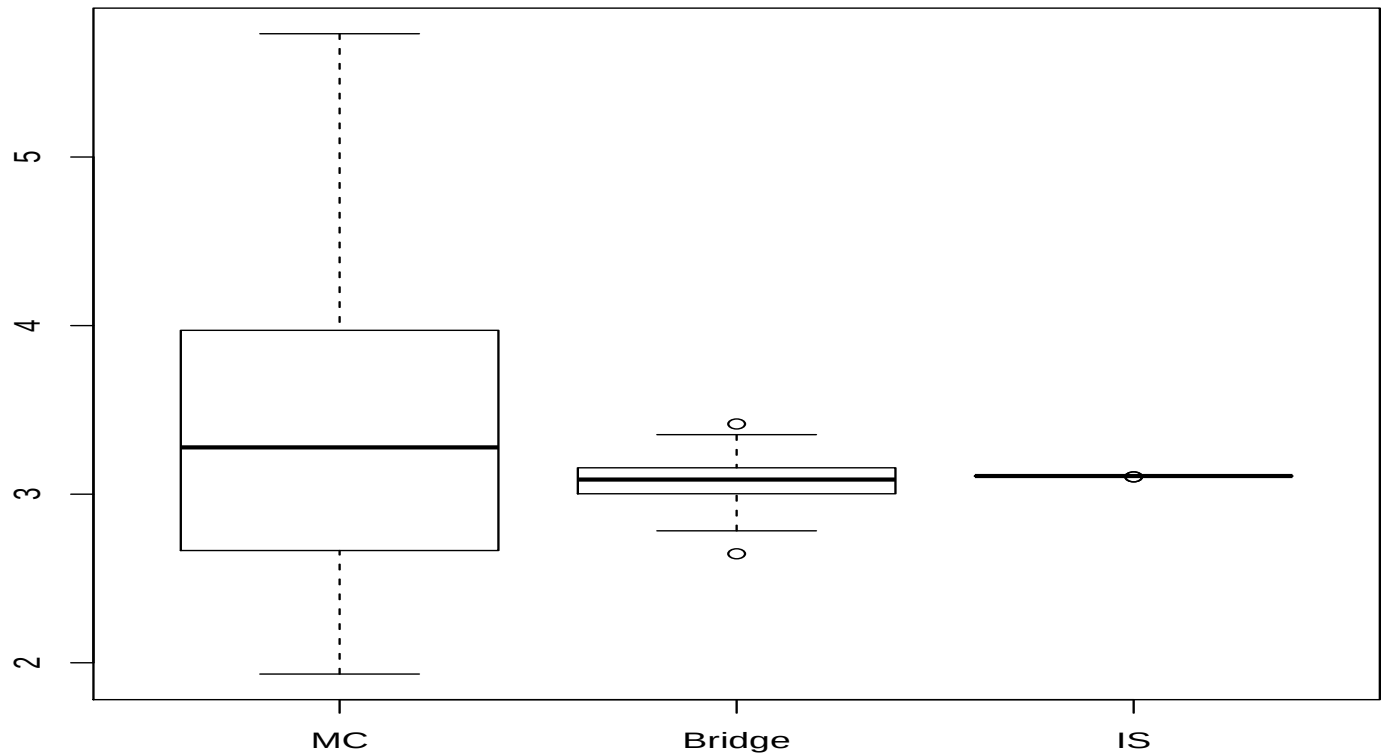
```
cova=model2$cov.unscaled
expecta=model2$coeff[,1]
covw=cova[3,3]-t(cova[1:2,3])%*%ginv(cova[1:2,1:2])%*%cova[1:2,3]

probit1=hmprobit(Niter,y,X1)
probit2=hmprobit(Niter,y,X2)
pseudo=rnorm(Niter,meanw(probit1),sqrt(covw))
probit1p=cbind(probit1,pseudo)

bfbs=mean(exp(probitlpost(probit2[,1:2],y,X1)+dnorm(probit2[,3],meanw(probit2[,1:2]),
     sqrt(covw),log=T))/ (dmvnorm(probit2,expecta,cova)+dnorm(probit2[,3],expecta[3],
     cova[3,3])))/ mean(exp(probitlpost(probit1p,y,X2))/(dmvnorm(probit1p,expecta,cova)+
     dnorm(pseudo,expecta[3],cova[3,3])))
```

Diabetes in Pima Indian women (cont'd)

Comparison of the variation of the Bayes factor approximations based on $100 \times 20,000$ simulations from the prior (MC), the above bridge sampler and the above importance sampler



The original harmonic mean estimator

When $\theta_{ki} \sim \pi_k(\theta|x)$,

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{L(\theta_{kt}|x)}$$

is an unbiased estimator of $1/m_k(x)$

[Newton & Raftery, 1994]

Highly dangerous: Most often leads to an infinite variance!!!

The original harmonic mean estimator

When $\theta_{ki} \sim \pi_k(\theta|x)$,

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{L(\theta_{kt}|x)}$$

is an unbiased estimator of $1/m_k(x)$

[Newton & Raftery, 1994]

Highly dangerous: Most often leads to an infinite variance!!!

"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 it's easy for people to not realize this, and to naïvely 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 it's easy for people to not realize this, and to naïvely accept estimates that are nowhere close to the correct value of the marginal likelihood."

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

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

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

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)

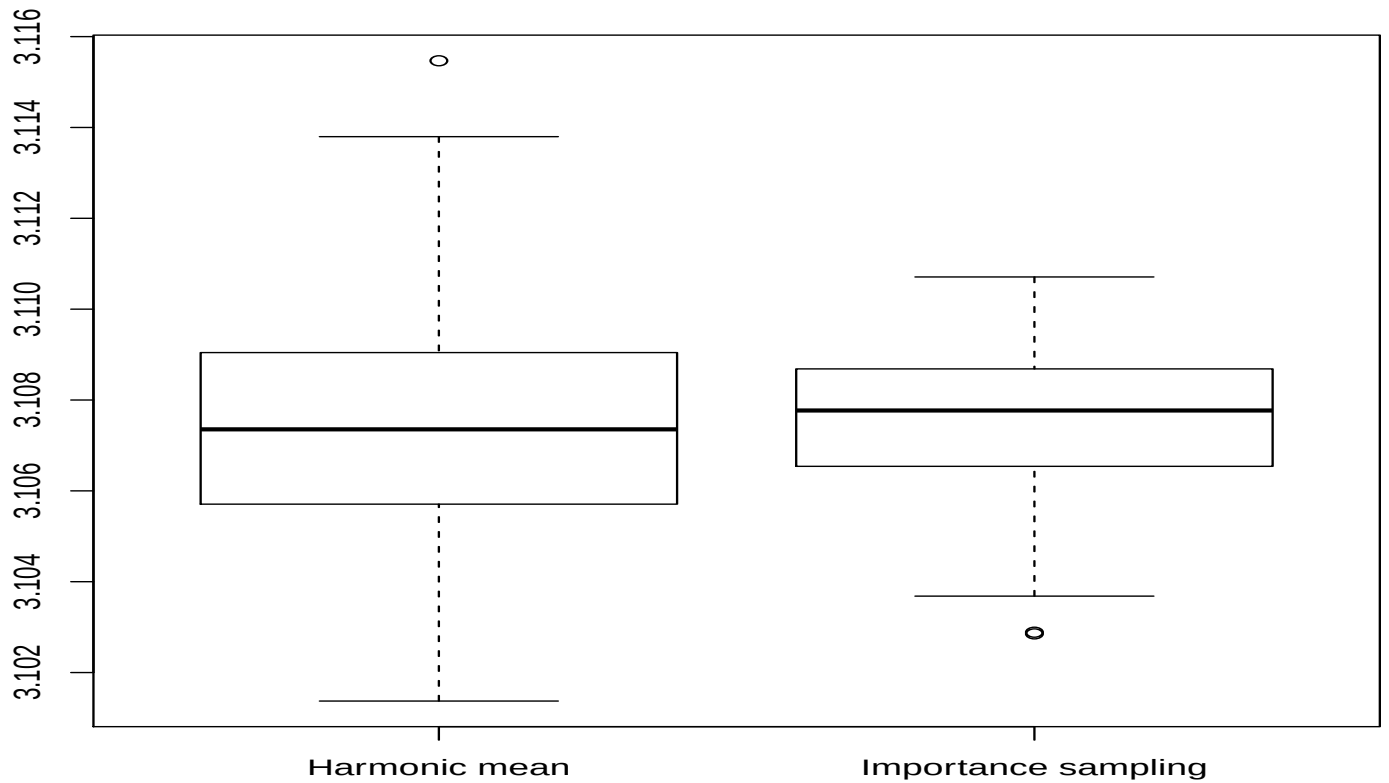
HPD indicator as φ

Use the convex hull of MCMC simulations corresponding to the 10% HPD region (easily derived!) and φ as indicator:

$$\varphi(\theta) = \frac{10}{T} \sum_{t \in \text{HPD}} \mathbb{I}_{d(\theta,\theta^{(t)}) \leq \epsilon}$$

Diabetes in Pima Indian women (cont'd)

Comparison of the variation of the Bayes factor approximations based on 100 replicas for $20,000$ simulations for a simulation from the above harmonic mean sampler and importance samplers



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

[Chopin & Robert, 2010]

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

[Chopin & Robert, 2010]

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

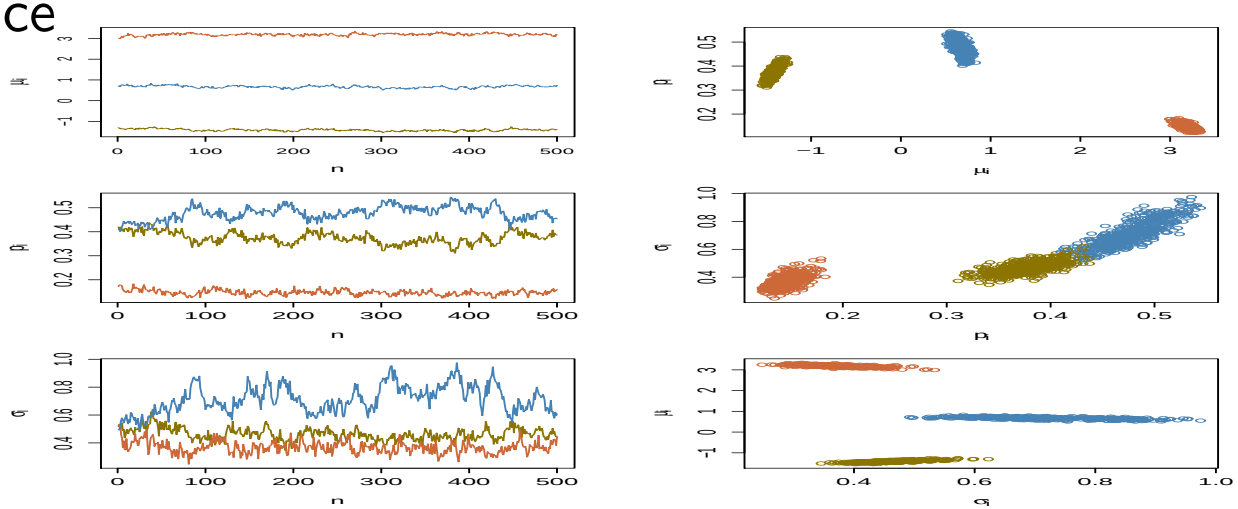
1. Number of modes of the likelihood of order $\mathrm{O}(k!)$:
   © Maximization and even [MCMC] exploration of the posterior surface harder
2. Under exchangeable priors on $(\theta, \mathbf{p})$ *[prior invariant under permutation of the indices]*, all posterior marginals are identical:
   © Posterior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

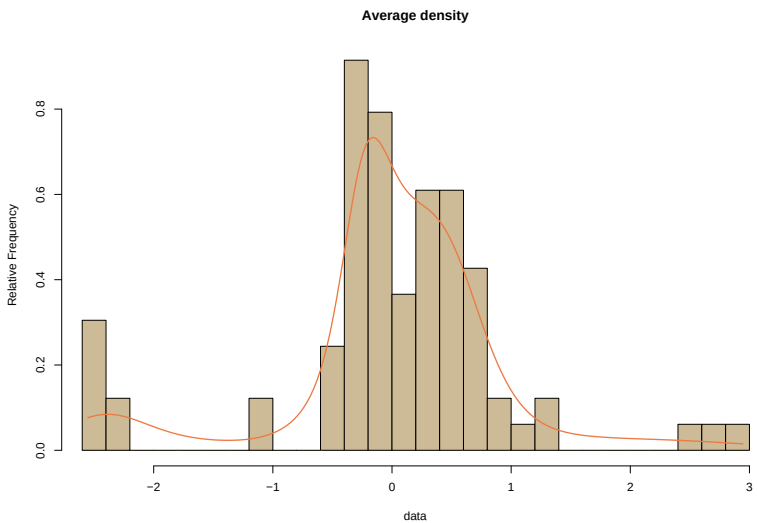
$$\widetilde{\pi_k}(\theta_k^*|\mathbf{x}) = \frac{1}{T\,k!}\sum_{\sigma\in\mathfrak{S}_k}\sum_{t=1}^{T} \pi_k(\sigma(\theta_k^*)|\mathbf{x},\mathbf{z}_k^{(t)})\,.$$

[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]

Case of the probit model

For the completion by z,

$$\hat{\pi}(\theta|x) = \frac{1}{T} \sum_t \pi(\theta|x, z^{(t)})$$

is a simple average of normal densities

R Bridge sampling code

```
gibbs1=gibbsprobit(Niter,y,X1)
gibbs2=gibbsprobit(Niter,y,X2)
bfchi=mean(exp(dmvlnorm(t(t(gibbs2$mu)-model2$coeff[,1]),mean=rep(0,3),
       sigma=gibbs2$Sigma2)-probitlpost(model2$coeff[,1],y,X2)))/
     mean(exp(dmvlnorm(t(t(gibbs1$mu)-model1$coeff[,1]),mean=rep(0,2),
       sigma=gibbs1$Sigma2)-probitlpost(model1$coeff[,1],y,X1)))
```

Case of the probit model

For the completion by z,

$$\hat{\pi}(\theta|x) = \frac{1}{T}\sum_t \pi(\theta|x, z^{(t)})$$

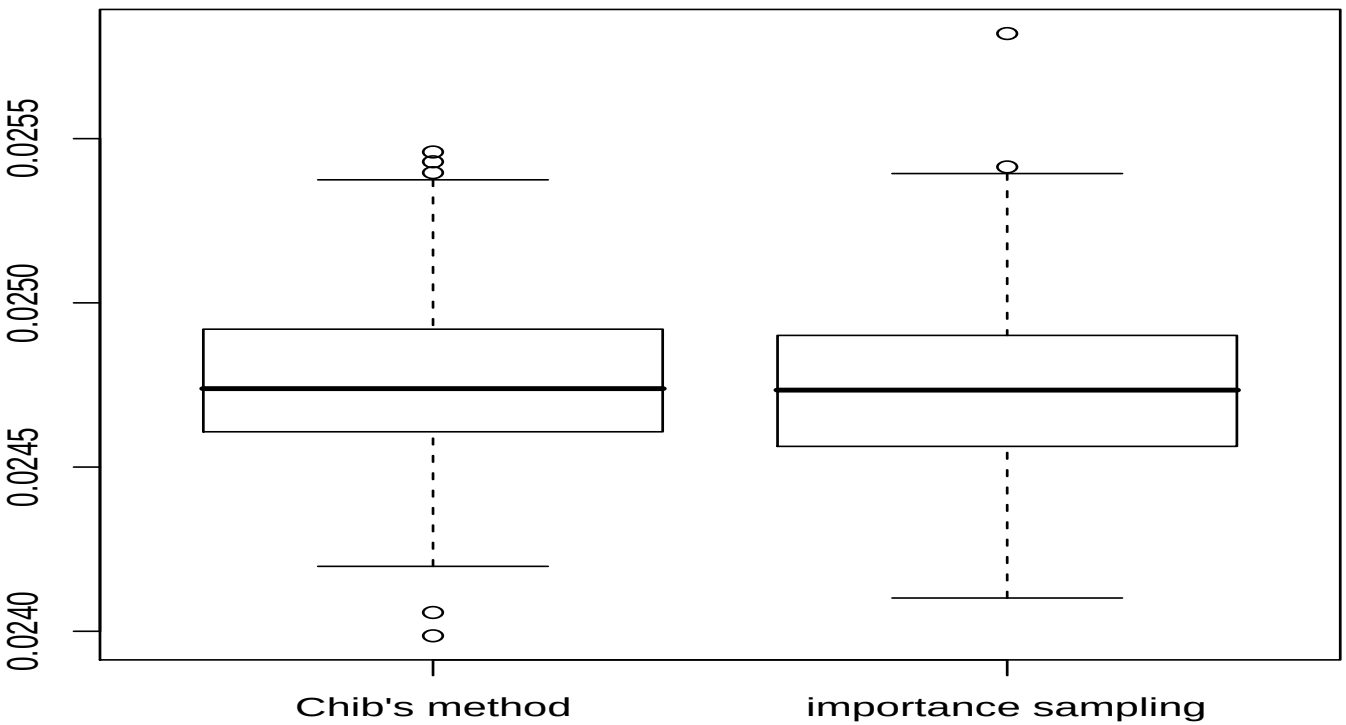
is a simple average of normal densities

R Bridge sampling code

```
gibbs1=gibbsprobit(Niter,y,X1)
gibbs2=gibbsprobit(Niter,y,X2)
bfchi=mean(exp(dmvlnorm(t(t(gibbs2$mu)-model2$coeff[,1]),mean=rep(0,3),
        sigma=gibbs2$Sigma2)-probitlpost(model2$coeff[,1],y,X2)))/
      mean(exp(dmvlnorm(t(t(gibbs1$mu)-model1$coeff[,1]),mean=rep(0,2),
        sigma=gibbs1$Sigma2)-probitlpost(model1$coeff[,1],y,X1)))
```

Diabetes in Pima Indian women (cont'd)

Comparison of the variation of the Bayes factor approximations based on 100 replicas for $20,000$ simulations for a simulation from the above Chib's and importance samplers



The Savage–Dickey ratio

Special representation of the Bayes factor used for simulation

Original version (Dickey, AoMS, 1971)

integrands.

Define for all η, $P'(\eta \mid \overline{H}) = \int f(\eta, \zeta)\, d\zeta$, and define for all η, ζ, $P'(\eta, \zeta \mid \overline{H}, \mathbf{D}) = \varphi(\mathbf{D} \mid \eta, \zeta) \cdot f(\eta, \zeta)/\Phi(\mathbf{D} \mid \overline{H})$, motivating the quite natural definition for all η, $P'(\eta \mid \overline{H}, \mathbf{D}) = \int \varphi(\mathbf{D} \mid \eta, \zeta) \cdot f(\eta, \zeta)\, d\zeta/\Phi(\mathbf{D} \mid \overline{H})$.

THEOREM (Savage's Density Ratio). *If*

$$g(\zeta) = f(\eta_H, \zeta)/\int f(\eta_H, \tilde{\zeta})\, d\tilde{\zeta}, \tag{3.8}$$

then

$$L_D(H) = P'(\eta_H \mid \overline{H}, \mathbf{D})/P'(\eta_H \mid \overline{H}). \tag{3.9}$$

Savage's density ratio theorem

Given a test $H_0 : \theta = \theta_0$ in a model $f(x|\theta, \psi)$ with a nuisance parameter ψ, under priors $\pi_0(\psi)$ and $\pi_1(\theta, \psi)$ such that

$$\pi_1(\psi|\theta_0) = \pi_0(\psi)$$

then

$$B_{01} = \frac{\pi_1(\theta_0|x)}{\pi_1(\theta_0)}\,,$$

with the obvious notations

$$\pi_1(\theta) = \int \pi_1(\theta, \psi)\mathrm{d}\psi\,, \quad \pi_1(\theta|x) = \int \pi_1(\theta, \psi|x)\mathrm{d}\psi\,,$$

[Dickey, 1971; Verdinelli & Wasserman, 1995]

Rephrased

"Suppose that $f_0(\theta) = f_1(\theta|\phi = \phi_0)$. As $f_0(x|\theta) = f_1(x|\theta, \phi = \phi_0)$,

$$f_0(x) = \int f_1(x|\theta, \phi = \phi_0) f_1(\theta|\phi = \phi_0)\, \mathrm{d}\theta = f_1(x|\phi = \phi_0)\,,$$

i.e., the denumerator of the Bayes factor is the value of $f_1(x|\phi)$ at $\phi = \phi_0$, while the denominator is an average of the values of $f_1(x|\phi)$ for $\phi \neq \phi_0$, weighted by the prior distribution $f_1(\phi)$ under the augmented model. Applying Bayes' theorem to the right-hand side of *[the above]* we get

$$f_0(x) = f_1(\phi_0|x) f_1(x)/f_1(\phi_0)$$

and hence the Bayes factor is given by

$$B = f_0(x)/f_1(x) = f_1(\phi_0|x)/f_1(\phi_0)\,.$$

the ratio of the posterior to prior densities at $\phi = \phi_0$ under the augmented model."

[O'Hagan & Forster, 1996]

Rephrased

"Suppose that $f_0(\theta) = f_1(\theta|\phi = \phi_0)$. As $f_0(x|\theta) = f_1(x|\theta, \phi = \phi_0)$,

$$f_0(x) = \int f_1(x|\theta, \phi = \phi_0) f_1(\theta|\phi = \phi_0)\, \mathrm{d}\theta = f_1(x|\phi = \phi_0)\,,$$

i.e., the denumerator of the Bayes factor is the value of $f_1(x|\phi)$ at $\phi = \phi_0$, while the denominator is an average of the values of $f_1(x|\phi)$ for $\phi \neq \phi_0$, weighted by the prior distribution $f_1(\phi)$ under the augmented model. Applying Bayes' theorem to the right-hand side of *[the above]* we get

$$f_0(x) = f_1(\phi_0|x) f_1(x) / f_1(\phi_0)$$

and hence the Bayes factor is given by

$$B = f_0(x)/f_1(x) = f_1(\phi_0|x)/f_1(\phi_0)\,.$$

the ratio of the posterior to prior densities at $\phi = \phi_0$ under the augmented model."

[O'Hagan & Forster, 1996]

Measure-theoretic difficulty

Representation depends on the choice of versions of conditional densities:

$$B_{01} = \frac{\int \pi_0(\psi) f(x|\theta_0,\psi)\,\mathrm{d}\psi}{\int \pi_1(\theta,\psi) f(x|\theta,\psi)\,\mathrm{d}\psi\mathrm{d}\theta} \quad \text{[by definition]}$$

$$= \frac{\int \pi_1(\psi|\theta_0) f(x|\theta_0,\psi)\,\mathrm{d}\psi\,\pi_1(\theta_0)}{\int \pi_1(\theta,\psi) f(x|\theta,\psi)\,\mathrm{d}\psi\mathrm{d}\theta\,\pi_1(\theta_0)} \quad \text{[specific version of } \pi_1(\psi|\theta_0)$$

$$\text{and arbitrary version of } \pi_1(\theta_0)]$$

$$= \frac{\int \pi_1(\theta_0,\psi) f(x|\theta_0,\psi)\,\mathrm{d}\psi}{m_1(x)\pi_1(\theta_0)} \quad \text{[specific version of } \pi_1(\theta_0,\psi)]$$

$$= \frac{\pi_1(\theta_0|x)}{\pi_1(\theta_0)} \quad \text{[version dependent]}$$

Measure-theoretic difficulty

Representation depends on the choice of versions of conditional densities:

$$B_{01} = \frac{\int \pi_0(\psi) f(x|\theta_0,\psi)\,\mathrm{d}\psi}{\int \pi_1(\theta,\psi) f(x|\theta,\psi)\,\mathrm{d}\psi\mathrm{d}\theta} \quad \text{[by definition]}$$

$$= \frac{\int \pi_1(\psi|\theta_0) f(x|\theta_0,\psi)\,\mathrm{d}\psi\,\pi_1(\theta_0)}{\int \pi_1(\theta,\psi) f(x|\theta,\psi)\,\mathrm{d}\psi\mathrm{d}\theta\,\pi_1(\theta_0)} \quad \text{[specific version of } \pi_1(\psi|\theta_0) \text{ and arbitrary version of } \pi_1(\theta_0)\text{]}$$

$$= \frac{\int \pi_1(\theta_0,\psi) f(x|\theta_0,\psi)\,\mathrm{d}\psi}{m_1(x)\pi_1(\theta_0)} \quad \text{[specific version of } \pi_1(\theta_0,\psi)\text{]}$$

$$= \frac{\pi_1(\theta_0|x)}{\pi_1(\theta_0)} \quad \text{[version dependent]}$$

Choice of density version

© Dickey's (1971) condition is not a condition:

If

$$\frac{\pi_1(\theta_0|x)}{\pi_1(\theta_0)} = \frac{\int \pi_0(\psi) f(x|\theta_0,\psi)\,\mathrm{d}\psi}{m_1(x)}$$

is chosen as a version, then Savage–Dickey's representation holds

Choice of density version

© Dickey's (1971) condition is not a condition:

If

$$\frac{\pi_1(\theta_0|x)}{\pi_1(\theta_0)} = \frac{\int \pi_0(\psi) f(x|\theta_0,\psi)\,\mathrm{d}\psi}{m_1(x)}$$

is chosen as a version, then Savage–Dickey's representation holds

Savage–Dickey paradox

Verdinelli-Wasserman extension:

$$B_{01} = \frac{\pi_1(\theta_0|x)}{\pi_1(\theta_0)}\, \mathbb{E}^{\pi_1(\psi|x,\theta_0,x)} \left[\frac{\pi_0(\psi)}{\pi_1(\psi|\theta_0)} \right]$$

similarly depends on choices of versions...

...but Monte Carlo implementation relies on specific versions of all densities *without making mention of it*

[Chen, Shao & Ibrahim, 2000]

Savage–Dickey paradox

Verdinelli-Wasserman extension:

$$B_{01} = \frac{\pi_1(\theta_0|x)}{\pi_1(\theta_0)}\, \mathbb{E}^{\pi_1(\psi|x,\theta_0,x)} \left[\frac{\pi_0(\psi)}{\pi_1(\psi|\theta_0)} \right]$$

similarly depends on choices of versions...

...but Monte Carlo implementation relies on specific versions of all densities *without making mention of it*

[Chen, Shao & Ibrahim, 2000]

Computational implementation

Starting from the (new) prior

$$\tilde{\pi}_1(\theta, \psi) = \pi_1(\theta)\pi_0(\psi)$$

define the associated posterior

$$\tilde{\pi}_1(\theta, \psi|x) = \pi_0(\psi)\pi_1(\theta)f(x|\theta, \psi)/\tilde{m}_1(x)$$

and impose

$$\frac{\tilde{\pi}_1(\theta_0|x)}{\pi_0(\theta_0)} = \frac{\int \pi_0(\psi)f(x|\theta_0, \psi)\,\mathrm{d}\psi}{\tilde{m}_1(x)}$$

to hold.

Then

$$B_{01} = \frac{\tilde{\pi}_1(\theta_0|x)}{\pi_1(\theta_0)}\,\frac{\tilde{m}_1(x)}{m_1(x)}$$

Computational implementation

Starting from the (new) prior

$$\tilde{\pi}_1(\theta, \psi) = \pi_1(\theta)\pi_0(\psi)$$

define the associated posterior

$$\tilde{\pi}_1(\theta, \psi|x) = \pi_0(\psi)\pi_1(\theta)f(x|\theta, \psi)/\tilde{m}_1(x)$$

and impose

$$\frac{\tilde{\pi}_1(\theta_0|x)}{\pi_0(\theta_0)} = \frac{\int \pi_0(\psi)f(x|\theta_0, \psi)\,\mathrm{d}\psi}{\tilde{m}_1(x)}$$

to hold.
Then

$$B_{01} = \frac{\tilde{\pi}_1(\theta_0|x)}{\pi_1(\theta_0)}\,\frac{\tilde{m}_1(x)}{m_1(x)}$$

First ratio

If $(\theta^{(1)}, \psi^{(1)}), \ldots, (\theta^{(T)}, \psi^{(T)}) \sim \tilde{\pi}(\theta, \psi|x)$, then

$$\frac{1}{T} \sum_t \tilde{\pi}_1(\theta_0|x, \psi^{(t)})$$

converges to $\tilde{\pi}_1(\theta_0|x)$ (if the right version is used in θ_0).

$$\tilde{\pi}_1(\theta_0|x, \psi) = \frac{\pi_1(\theta_0) f(x|\theta_0, \psi)}{\int \pi_1(\theta) f(x|\theta, \psi) \, \mathrm{d}\theta}$$

Rao–Blackwellisation with latent variables

When $\tilde{\pi}_1(\theta_0|x,\psi)$ unavailable, replace with

$$\frac{1}{T}\sum_{t=1}^{T}\tilde{\pi}_1(\theta_0|x,z^{(t)},\psi^{(t)})$$

via data completion by latent variable z such that

$$f(x|\theta,\psi)=\int\tilde{f}(x,z|\theta,\psi)\,\mathrm{d}z$$

and that $\tilde{\pi}_1(\theta,\psi,z|x)\propto\pi_0(\psi)\pi_1(\theta)\tilde{f}(x,z|\theta,\psi)$ available in closed form, including the normalising constant, based on version

$$\frac{\tilde{\pi}_1(\theta_0|x,z,\psi)}{\pi_1(\theta_0)}=\frac{\tilde{f}(x,z|\theta_0,\psi)}{\int\tilde{f}(x,z|\theta,\psi)\pi_1(\theta)\,\mathrm{d}\theta}\,.$$

Rao–Blackwellisation with latent variables

When $\tilde{\pi}_1(\theta_0|x,\psi)$ unavailable, replace with

$$\frac{1}{T}\sum_{t=1}^{T}\tilde{\pi}_1(\theta_0|x,z^{(t)},\psi^{(t)})$$

via data completion by latent variable z such that

$$f(x|\theta,\psi)=\int\tilde{f}(x,z|\theta,\psi)\,\mathrm{d}z$$

and that $\tilde{\pi}_1(\theta,\psi,z|x)\propto\pi_0(\psi)\pi_1(\theta)\tilde{f}(x,z|\theta,\psi)$ available in closed form, including the normalising constant, based on version

$$\frac{\tilde{\pi}_1(\theta_0|x,z,\psi)}{\pi_1(\theta_0)}=\frac{\tilde{f}(x,z|\theta_0,\psi)}{\int\tilde{f}(x,z|\theta,\psi)\pi_1(\theta)\,\mathrm{d}\theta}\,.$$

Bridge revival (1)

Since $\tilde{m}_1(x)/m_1(x)$ is unknown, apparent failure!

bridge identity

$$\mathbb{E}^{\tilde{\pi}_1(\theta,\psi|x)}\left[\frac{\pi_1(\theta,\psi)f(x|\theta,\psi)}{\pi_0(\psi)\pi_1(\theta)f(x|\theta,\psi)}\right] = \mathbb{E}^{\tilde{\pi}_1(\theta,\psi|x)}\left[\frac{\pi_1(\psi|\theta)}{\pi_0(\psi)}\right] = \frac{m_1(x)}{\tilde{m}_1(x)}$$

to (biasedly) estimate $\tilde{m}_1(x)/m_1(x)$ by

$$T \Big/ \sum_{t=1}^{T} \frac{\pi_1(\psi^{(t)}|\theta^{(t)})}{\pi_0(\psi^{(t)})}$$

based on the same sample from $\tilde{\pi}_1$.

Bridge revival (1)

Since $\tilde{m}_1(x)/m_1(x)$ is unknown, apparent failure!

bridge identity

$$\mathbb{E}^{\tilde{\pi}_1(\theta,\psi|x)}\left[\frac{\pi_1(\theta,\psi)f(x|\theta,\psi)}{\pi_0(\psi)\pi_1(\theta)f(x|\theta,\psi)}\right] = \mathbb{E}^{\tilde{\pi}_1(\theta,\psi|x)}\left[\frac{\pi_1(\psi|\theta)}{\pi_0(\psi)}\right] = \frac{m_1(x)}{\tilde{m}_1(x)}$$

to (biasedly) estimate $\tilde{m}_1(x)/m_1(x)$ by

$$T\Big/\sum_{t=1}^{T}\frac{\pi_1(\psi^{(t)}|\theta^{(t)})}{\pi_0(\psi^{(t)})}$$

based on the same sample from $\tilde{\pi}_1$.

Bridge revival (2)

Alternative identity

$$\mathbb{E}^{\pi_1(\theta,\psi|x)}\left[\frac{\pi_0(\psi)\pi_1(\theta)f(x|\theta,\psi)}{\pi_1(\theta,\psi)f(x|\theta,\psi)}\right] = \mathbb{E}^{\pi_1(\theta,\psi|x)}\left[\frac{\pi_0(\psi)}{\pi_1(\psi|\theta)}\right] = \frac{\tilde{m}_1(x)}{m_1(x)}$$

suggests using a second sample $(\bar{\theta}^{(1)}, \bar{\psi}^{(1)}, z^{(1)}), \ldots,$ $(\bar{\theta}^{(T)}, \bar{\psi}^{(T)}, z^{(T)}) \sim \pi_1(\theta, \psi|x)$ and the ratio estimate

$$\frac{1}{T}\sum_{t=1}^{T} \pi_0(\bar{\psi}^{(t)}) \Big/ \pi_1(\bar{\psi}^{(t)}|\bar{\theta}^{(t)})$$

Resulting unbiased estimate:

$$\widehat{B_{01}} = \frac{1}{T}\frac{\sum_t \tilde{\pi}_1(\theta_0|x, z^{(t)}, \psi^{(t)})}{\pi_1(\theta_0)} \frac{1}{T}\sum_{t=1}^{T} \frac{\pi_0(\bar{\psi}^{(t)})}{\pi_1(\bar{\psi}^{(t)}|\bar{\theta}^{(t)})}$$

Bridge revival (2)

Alternative identity

$$\mathbb{E}^{\pi_1(\theta,\psi|x)}\left[\frac{\pi_0(\psi)\pi_1(\theta)f(x|\theta,\psi)}{\pi_1(\theta,\psi)f(x|\theta,\psi)}\right] = \mathbb{E}^{\pi_1(\theta,\psi|x)}\left[\frac{\pi_0(\psi)}{\pi_1(\psi|\theta)}\right] = \frac{\tilde{m}_1(x)}{m_1(x)}$$

suggests using a second sample $(\bar{\theta}^{(1)}, \bar{\psi}^{(1)}, z^{(1)}), \ldots,$ $(\bar{\theta}^{(T)}, \bar{\psi}^{(T)}, z^{(T)}) \sim \pi_1(\theta, \psi|x)$ and the ratio estimate

$$\frac{1}{T}\sum_{t=1}^{T} \pi_0(\bar{\psi}^{(t)}) \Big/ \pi_1(\bar{\psi}^{(t)}|\bar{\theta}^{(t)})$$

Resulting unbiased estimate:

$$\widehat{B_{01}} = \frac{1}{T}\frac{\sum_t \tilde{\pi}_1(\theta_0|x, z^{(t)}, \psi^{(t)})}{\pi_1(\theta_0)} \frac{1}{T}\sum_{t=1}^{T} \frac{\pi_0(\bar{\psi}^{(t)})}{\pi_1(\bar{\psi}^{(t)}|\bar{\theta}^{(t)})}$$

Difference with Verdinelli–Wasserman representation

The above leads to the representation

$$B_{01} = \frac{\tilde{\pi}_1(\theta_0|x)}{\pi_1(\theta_0)}\, \mathbb{E}^{\pi_1(\theta,\psi|x)} \left[\frac{\pi_0(\psi)}{\pi_1(\psi|\theta)} \right]$$

which shows how our approach differs from Verdinelli and Wasserman's

$$B_{01} = \frac{\pi_1(\theta_0|x)}{\pi_1(\theta_0)}\, \mathbb{E}^{\pi_1(\psi|x,\theta_0,x)} \left[\frac{\pi_0(\psi)}{\pi_1(\psi|\theta_0)} \right]$$

Difference with Verdinelli–Wasserman approximation

In terms of implementation,

$$\widehat{B_{01}}^{\mathsf{MR}}(x) = \frac{1}{T}\frac{\sum_t \tilde{\pi}_1(\theta_0|x, z^{(t)}, \psi^{(t)})}{\pi_1(\theta_0)}\,\frac{1}{T}\sum_{t=1}^{T}\frac{\pi_0(\bar{\psi}^{(t)})}{\pi_1(\bar{\psi}^{(t)}|\bar{\theta}^{(t)})}$$

formaly resembles

$$\widehat{B_{01}}^{\mathsf{VW}}(x) = \frac{1}{T}\sum_{t=1}^{T}\frac{\pi_1(\theta_0|x, z^{(t)}, \psi^{(t)})}{\pi_1(\theta_0)}\,\frac{1}{T}\sum_{t=1}^{T}\frac{\pi_0(\tilde{\psi}^{(t)})}{\pi_1(\tilde{\psi}^{(t)}|\theta_0)}\,.$$

But the simulated sequences differ: first average involves simulations from $\tilde{\pi}_1(\theta, \psi, z|x)$ and from $\pi_1(\theta, \psi, z|x)$, while second average relies on simulations from $\pi_1(\theta, \psi, z|x)$ and from $\pi_1(\psi, z|x, \theta_0)$,

Difference with Verdinelli–Wasserman approximation

In terms of implementation,

$$\widehat{B_{01}}^{\mathsf{MR}}(x) = \frac{1}{T}\frac{\sum_t \tilde{\pi}_1(\theta_0|x, z^{(t)}, \psi^{(t)})}{\pi_1(\theta_0)}\,\frac{1}{T}\sum_{t=1}^{T}\frac{\pi_0(\bar{\psi}^{(t)})}{\pi_1(\bar{\psi}^{(t)}|\bar{\theta}^{(t)})}$$

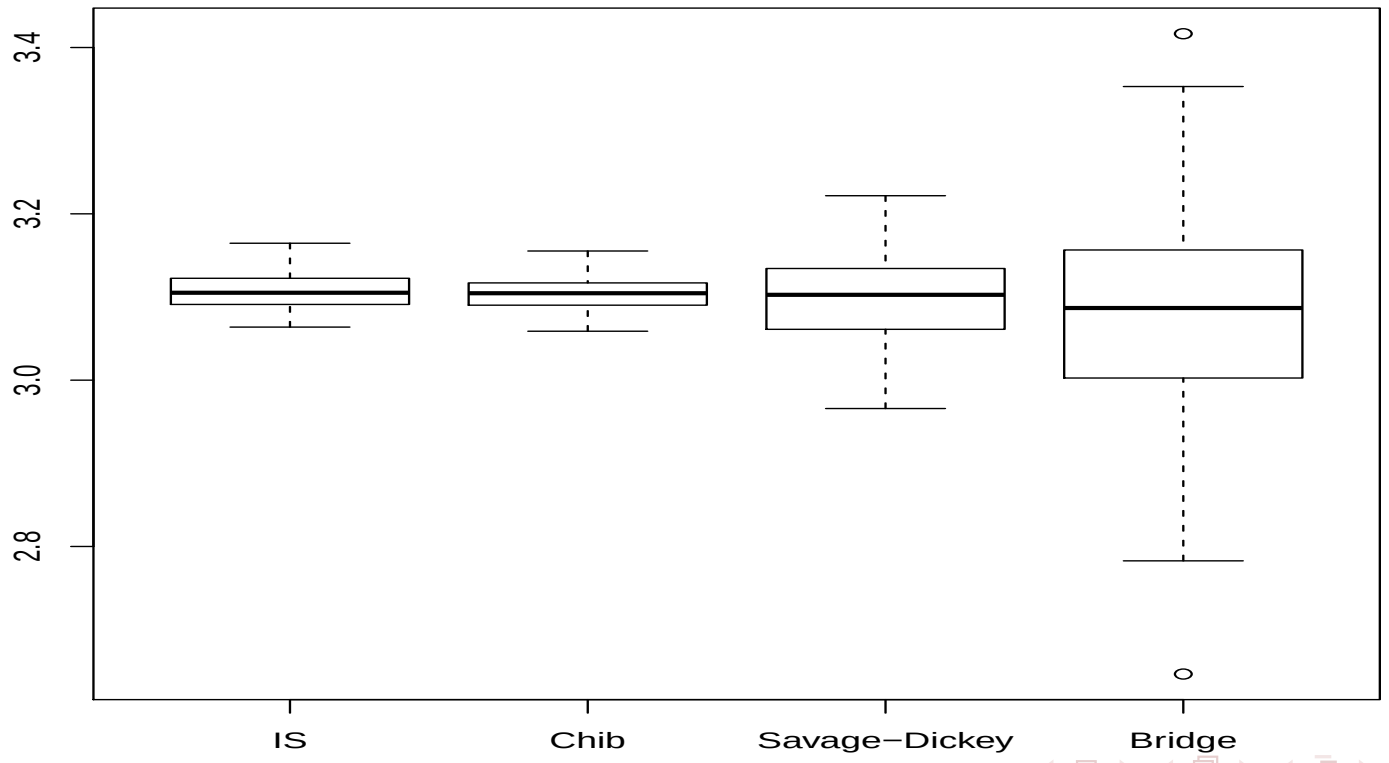
formaly resembles

$$\widehat{B_{01}}^{\mathsf{VW}}(x) = \frac{1}{T}\sum_{t=1}^{T}\frac{\pi_1(\theta_0|x, z^{(t)}, \psi^{(t)})}{\pi_1(\theta_0)}\,\frac{1}{T}\sum_{t=1}^{T}\frac{\pi_0(\tilde{\psi}^{(t)})}{\pi_1(\tilde{\psi}^{(t)}|\theta_0)}\,.$$

But the simulated sequences differ: first average involves simulations from $\tilde{\pi}_1(\theta, \psi, z|x)$ and from $\pi_1(\theta, \psi, z|x)$, while second average relies on simulations from $\pi_1(\theta, \psi, z|x)$ and from $\pi_1(\psi, z|x, \theta_0)$,

Diabetes in Pima Indian women (cont'd)

Comparison of the variation of the Bayes factor approximations based on 100 replicas for $20,000$ simulations for a simulation from the above importance, Chib's, Savage–Dickey's and bridge samplers



Properties of nested sampling

[Chopin & Robert, 2010]

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:

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

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

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

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

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

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!]

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

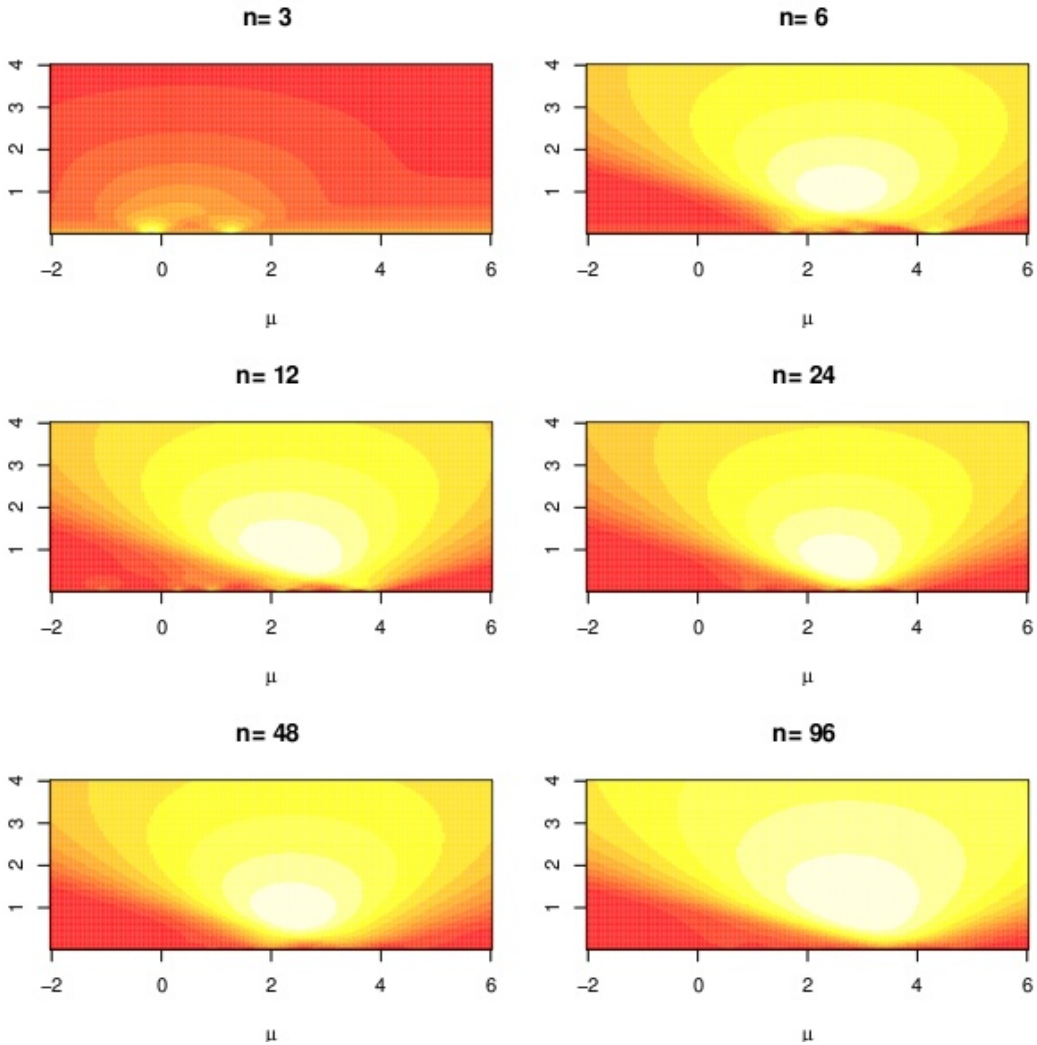
Posterior distribution on (μ, σ) associated with the mixture

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

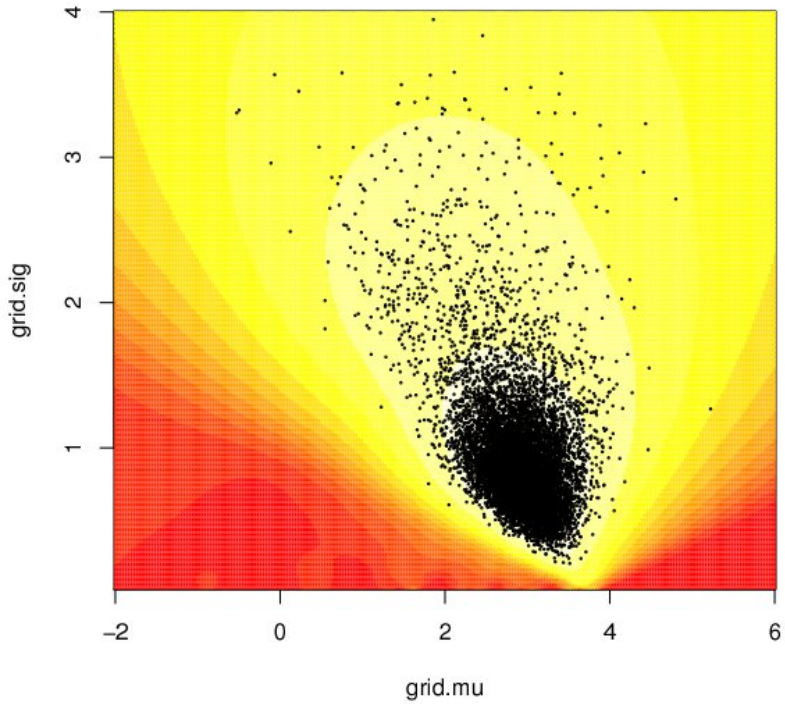
when p is known

Experiment

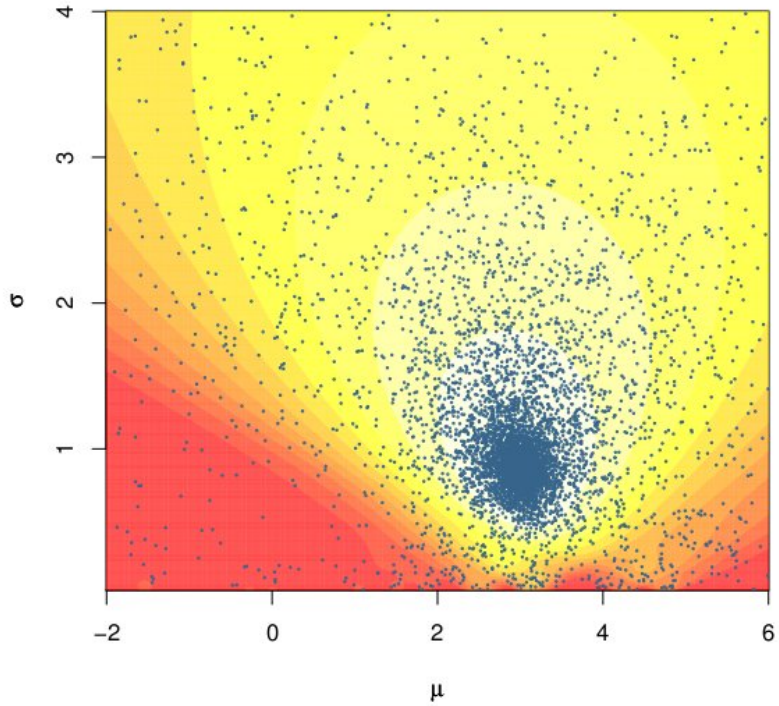
- n observations with $\mu = 2$ and $\sigma = 3/2$,
- Use of a uniform prior both on $(-2, 6)$ for μ and on $(.001, 16)$ for $\log \sigma^2$.
- occurrences of posterior bursts for $\mu = x_i$
- computation of the various estimates of $\mathfrak{Z}$



Experiment (cont'd)

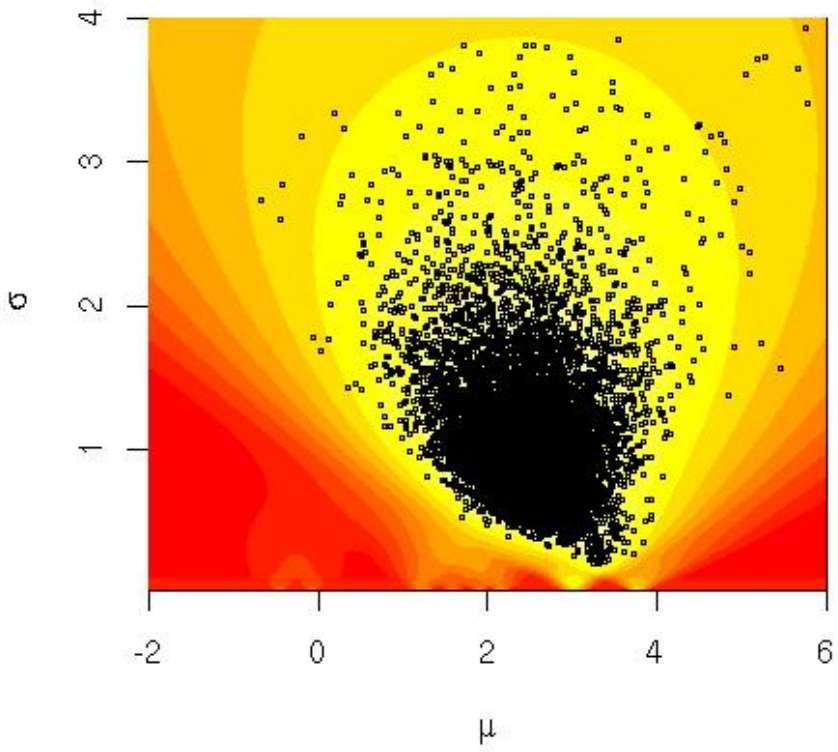


MCMC sample for $n = 16$ observations from the mixture.

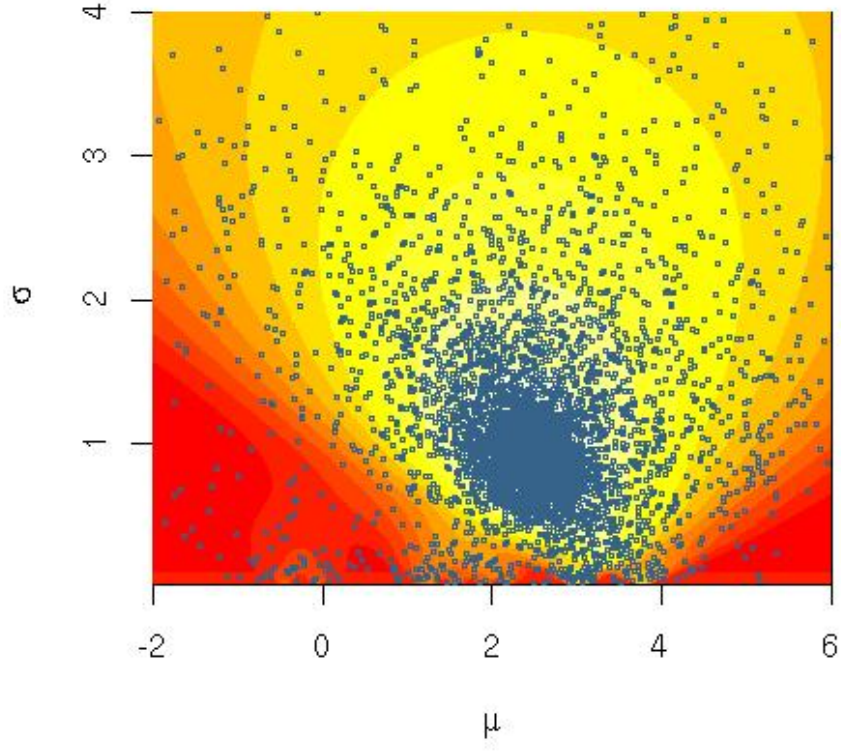


Nested sampling sequence with $M = 1000$ starting points.

Experiment (cont'd)



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

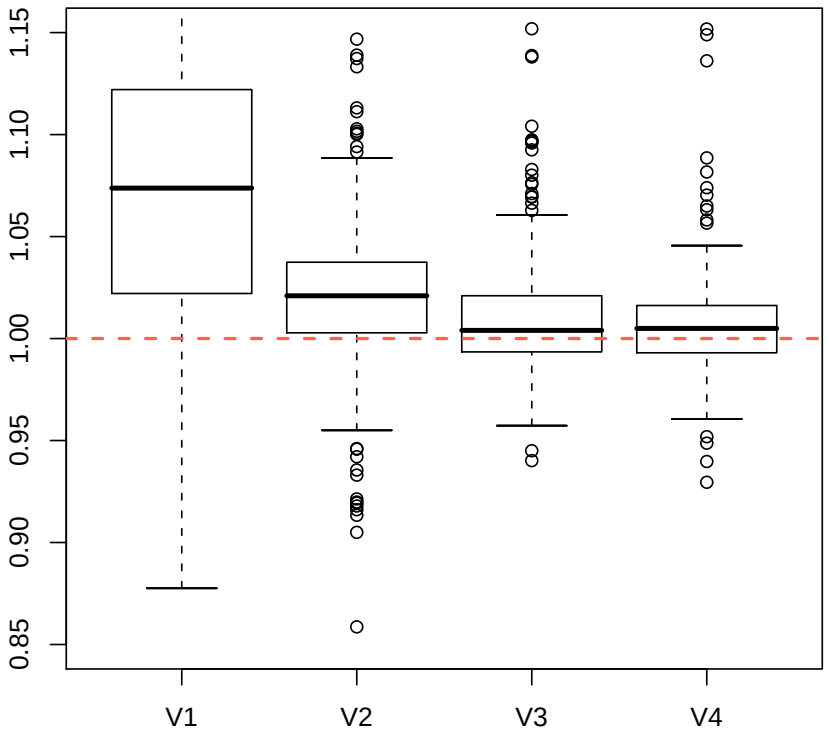


Nested sampling sequence with $M = 1000$ starting points.

Comparison

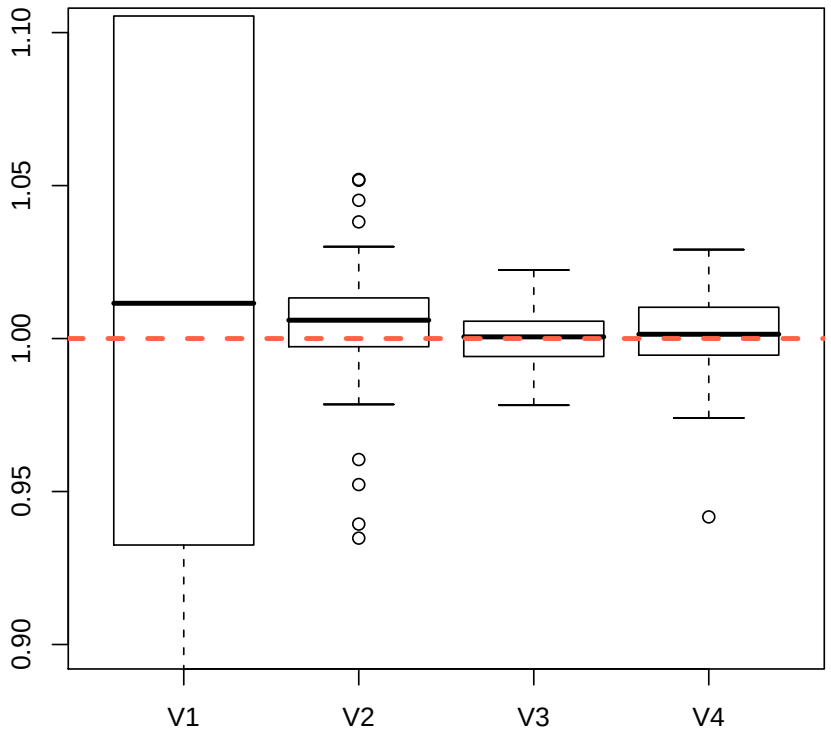
Monte Carlo and MCMC (=Gibbs) outputs based on $T = 10^4$ simulations and numerical integration based on a 850×950 grid in the (μ, σ) parameter space.
Nested sampling approximation based on a starting sample of $M = 1000$ points followed by at least 103 further simulations from the constr'd prior and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



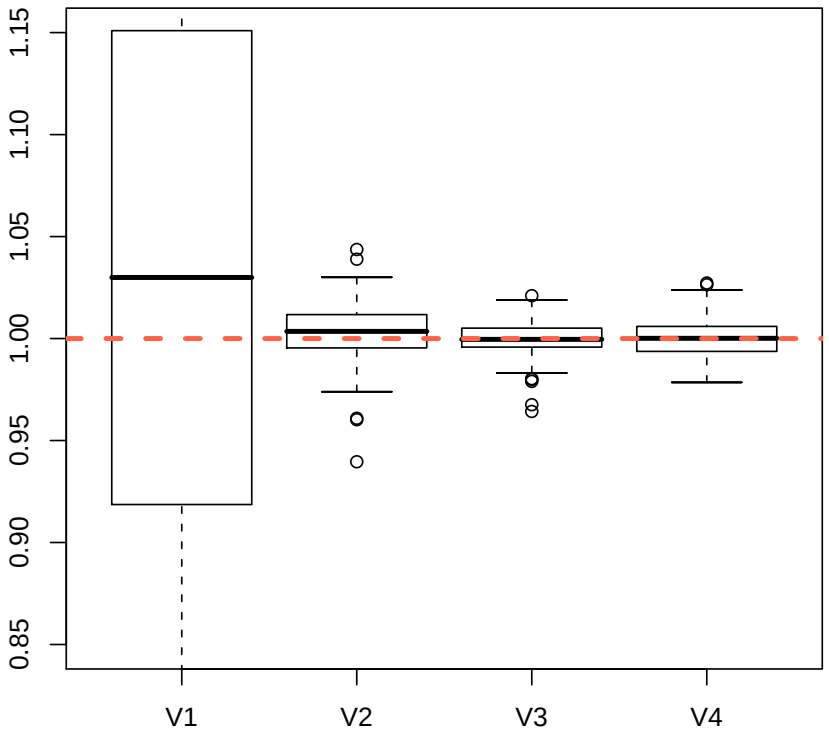
Graph based on a sample of 10 observations for $\mu = 2$ and $\sigma = 3/2$ (150 replicas).

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.)