Uncertainty quantification for marginal computations

Jean-Michel Marin

University of Montpellier, CNRS
Alexander Grothendieck Montpellier Institute

June 2022

Joint work with

- **Christian Robert**
  University Paris Dauphine and University of Warwick

- **Judith Rousseau**
  University of Oxford

Introduction

M Bayesian parametric models in competition

$$f_m(\mathbf{y}|\theta_m) \quad \pi_m(\theta_m) \quad m = 1, \ldots, M$$

Prior probabilities in the model space $\quad \mathbb{P}(\mathscr{M} = m)$

Target: the model's posterior probabilities

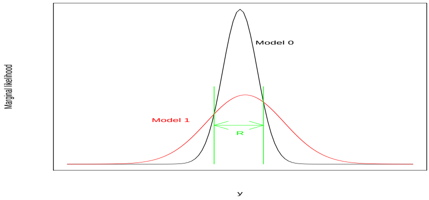
$$\mathbb{P}(\mathscr{M} = m|\mathbf{y}) \propto \mathbb{P}(\mathscr{M} = m) \int f_m(\mathbf{y}|\theta_m)\pi_m(\theta_m)d\theta_m$$

Introduction

A key quantity the marginal likelihood (the evidence)

$$\int f_m(\mathbf{y}|\theta_m)\pi_m(\theta_m)d\theta_m$$

Bayesian inference embodies Occam's razor



A simple model, like Model 0, makes only a limited range of predictions; a more powerful model, like Model 1, is able to predict a greater variety of data sets

If the data set falls in region R, the less powerful model will be the more probable model

Introduction

The marginal likelihood corresponds to a **penalized** likelihood

The BIC information criterium **Schwarz (1978)** comes from an asymptotic Laplace approximation of the marginal likelihood

Drton and Plummer (2017) Very nice extensions for singular model selection problems

Bayes factor for models M_1 and M_0

$$B_{10} = \frac{\int f_1(\mathbf{y}|\theta_1)\pi_1(\theta_1)d\theta_1}{\int f_0(\mathbf{y}|\theta_0)\pi_0(\theta_0)d\theta_0}$$

Introduction

Difficulties with the Bayesian model choice paradigm

Prior difficulties

- How to choose the prior distributions on the parameters of each model in a compatible way?
- What about the prior distribution in the models's space?

We do not address these crucial questions in this talk

Introduction

Computational difficulties

- How to approximate the marginal likelihoods?
- When the number of models in consideration is huge, how to explore the models's space?

We consider the case of a limited number of models and not address trans-dimensional sampling solutions, like the reversible jump algorithm

Introduction

We concentrate on the crucial question: how to approximate the marginal likelihood

$$m = \int f(\mathbf{y}|\theta)\pi(\theta)d\theta = \mathbb{E}_\pi\left[f(\mathbf{y}|\theta)\right]$$

We consider the case where the calculating of the likelihood is tractable

Introduction

We recall the main approximation techniques

We highlight the link between the Bridge sampling method and the noise-contrastive strategy

We show how to skillfully use the Weighted Likelihood Bootstrap technique to evaluate the associated error

Standard Monte Carlo approximation

$$m = \int f(\mathbf{y}|\theta)\pi(\theta)d\theta = \mathbb{E}_\pi\left[f(\mathbf{y}|\theta)\right]$$

$\theta^{(1)}, \ldots, \theta^{(N)}$ is an N-sample from $\pi(\cdot)$

$$\hat{m} = \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{y}|\theta^{(i)})$$

When the prior is far from the posterior
$\Longrightarrow$ very high variance

Importance sampling approximation

$g(\cdot)$ such that $g(\theta) > 0$ when $f(\mathbf{y}|\theta)\pi(\theta) > 0$

$$m = \int f(\mathbf{y}|\theta)\pi(\theta)d\theta = \mathbb{E}_g\left[f(\mathbf{y}|\theta)\frac{\pi(\theta)}{g(\theta)}\right]$$

$\theta^{(1)}, \ldots, \theta^{(N)}$ is an N-sample from $g(\cdot)$

$$\hat{m} = \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{y}|\theta^{(i)})\frac{\pi(\theta^{(i)})}{g(\theta^{(i)})}$$

Problem specific and curse of dimensionality

Chib's solution

Chib (1995)

$$m = \frac{f(\mathbf{y}|\theta)\,\pi(\theta)}{\pi(\theta|\mathbf{y})}$$

For an arbitrary value θ^* of $\theta \Longrightarrow$

$$\hat{m} = \frac{f(\mathbf{y}|\theta^*)\,\pi(\theta^*)}{\hat{\pi}(\theta^*|\mathbf{y})}$$

$\hat{\pi}(\theta|\mathbf{y})$ may be the Gaussian approximation based on the MLE

Chib's solution

Approximation based on a preliminary MCMC sample

Latent variables models $\Longrightarrow$ natural approximation to $\pi_k(\theta^*|\mathbf{y})$

$$\hat{\pi}(\theta^*|\mathbf{y}) = \frac{1}{T}\sum_{t=1}^{T} \pi(\theta^*|\mathbf{y}, \mathbf{z}^{(t)})$$

$\mathbf{z}^{(t)}$ the latent variables simulated by the MCMC sampler

High variance and curse of dimensionality

Bridge sampling techniques

Meng and Wong (1996), Meng and Schilling (2002)

$$m = \frac{\int f(\mathbf{y}|\theta)\pi(\theta)h(\theta)g(\theta)d\theta}{\int g(\theta)h(\theta)\pi(\theta|\mathbf{y})d\theta} = \frac{\mathbb{E}_g\left[f(\mathbf{y}|\theta)\pi(\theta)h(\theta)\right]}{\mathbb{E}_\pi\left[h(\theta)g(\theta)|\mathbf{y}\right]}$$

$g(\theta)$ a proposal distribution
$h(\theta)$ the bridge function

$$\hat{m} = \frac{\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} f(\mathbf{y}|\theta_0^{(i)})\pi(\theta_0^{(i)})h(\theta_0^{(i)})}{\frac{1}{N}\sum_{i=1}^{N} h(\theta_1^{(i)})g(\theta_1^{(i)})}$$

$\theta_0^{(1)}, \ldots, \theta_0^{(\alpha N)}$ is an αN-sample from $g(\cdot)$
$\theta_1^{(1)}, \ldots, \theta_1^{(N)}$ is an N-sample from $\pi(\cdot|\mathbf{y})$

Bridge sampling techniques

Gronau, Singmann, Wagenmakers (2020)
Nice R library bridgesampling

Overstall and Forster (2010) a convenient proposal
Gaussian distribution with its first two moments chosen to match those of the posterior distribution

Optimal bridge function

$$h(\theta) = \frac{C}{\left(\frac{1}{1+\alpha}\right) f(\mathbf{y}|\theta)\pi(\theta) + \left(\frac{\alpha}{1+\alpha}\right) g(\theta)m}$$

Optimal in the sense that it minimizes the relative squared error

The constant C cancels

Bridge sampling techniques

The optimal bridge function depends on m
$\Longrightarrow$ iterative scheme

$$\hat{m}^{(t+1)} = \frac{\displaystyle\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} \frac{f(\mathbf{y}|\theta_0^{(i)})\pi(\theta_0^{(i)})}{\left(\frac{1}{1+\alpha}\right) f(\mathbf{y}|\theta_0^{(i)})\pi(\theta_0^{(i)}) + \left(\frac{\alpha}{1+\alpha}\right) g(\theta_0^{(i)})\hat{m}^{(t)}}}{\displaystyle\frac{1}{N}\sum_{i=1}^{N} \frac{g(\theta_1^{(i)})}{\left(\frac{1}{1+\alpha}\right) f(\mathbf{y}|\theta_1^{(i)})\pi(\theta_1^{(i)}) + \left(\frac{\alpha}{1+\alpha}\right) g(\theta_1^{(i)})\hat{m}^{(t)}}}$$

Bridge sampling techniques

$$h_{1,(i)} = \frac{f(\mathbf{y}|\theta_1^{(i)})\pi(\theta_1^{(i)})}{g(\theta_1^{(i)})} \quad h_{0,(i)} = \frac{f(\mathbf{y}|\theta_0^{(i)})\pi(\theta_0^{(i)})}{g(\theta_0^{(i)})}$$

$$\hat{m}^{(t+1)} = \frac{\dfrac{1}{\alpha}\displaystyle\sum_{i=1}^{\alpha N} \frac{h_{0,(i)}}{h_{0,(i)} + \alpha\hat{m}^{(t)}}}{\displaystyle\sum_{i=1}^{N} \frac{1}{h_{1,(i)} + \alpha\hat{m}^{(t)}}}$$

$$\alpha\hat{m}^{(t+1)} \sum_{i=1}^{N} \frac{1}{h_{1,(i)} + \alpha\hat{m}^{(t)}} = \sum_{i=1}^{\alpha N} \frac{h_{0,(i)}}{h_{0,(i)} + \alpha\hat{m}^{(t)}}$$

Some others alternatives

Large set of approximations for marginal likelihood or Bayes factors

- Annealed Importance Sampling by **Neal (2001)**
- Sub-product of Sequential Monte Carlo samplers **Del Moral, Doucet and Jasra (2006)**
- The Savage–Dickey ratio **Verdinelli and Wasserman (1995), Marin and Robert (2010)**
- ...

Noise-contrastive estimation

Idea: reduce an estimation problem to a classification problem
Several versions:

- Logistic regression for density estimation: **Hastie et al. (2003)**
- Intensity estimation: **Baddeley et al. (2010)**
- Logistic regression for estimation in unnormalised models: **Geyer (1994) and Gutmann and Hyvarinen (2012)**

Noise-contrastive estimation

$$f_0(\theta|\mathbf{y}, z=0) = g(\theta) \quad ; \quad f_1(\theta|\mathbf{y}, z=1) = \frac{f(\mathbf{y}|\theta)\pi(\theta)}{m} = \pi(\theta|\mathbf{y})$$

$$\mathbb{P}(z=1|\mathbf{y},\theta) = \frac{\mathbb{P}(z=1)\dfrac{f(\mathbf{y}|\theta)\pi(\theta)}{m}}{\mathbb{P}(z=1)\dfrac{f(\mathbf{y}|\theta)\pi(\theta)}{m} + \mathbb{P}(z=0)g(\theta)}$$

$$\mathbb{P}(z=0|\mathbf{y},\theta) = \frac{\mathbb{P}(z=0)g(\theta)}{\mathbb{P}(z=1)\dfrac{f(\mathbf{y}|\theta)\pi(\theta)}{m} + \mathbb{P}(z=0)g(\theta)}$$

$$\mathbb{P}(z=0) \propto \alpha N \quad ; \quad \mathbb{P}(z=1) \propto N \quad ; \quad \frac{\mathbb{P}(z=0)}{\mathbb{P}(z=1)} = \alpha$$

Noise-contrastive estimation

$\theta_0^{(1)}, \ldots, \theta_0^{(\alpha N)}$ is an αN-sample from $g(\cdot)$
$\theta_1^{(1)}, \ldots, \theta_1^{(N)}$ is an N-sample from $\pi(\cdot|\mathbf{y})$

The *pseudo likelihood*

$$\prod_{i=1}^{N} \left\{ \frac{\frac{f(\mathbf{y}|\theta_1^{(i)})\pi(\theta_1^{(i)})}{m}}{\frac{f(\mathbf{y}|\theta_1^{(i)})\pi(\theta_1^{(i)})}{m} + \alpha g(\theta_1^{(i)})} \right\} \times$$

$$\prod_{i=1}^{\alpha N} \left\{ \frac{\alpha g(\theta_0^{(i)})}{\frac{f(\mathbf{y}|\theta_0^{(i)})\pi(\theta_0^{(i)})}{m} + \alpha g(\theta_0^{(i)})} \right\}$$

Noise-contrastive estimation

The *pseudo log-likelihood*

$$q(m) = \text{cst} - N\log(m) - \sum_{i=1}^{N} \log\left(\frac{f(\mathbf{y}|\theta_1^{(i)})\pi(\theta_1^{(i)})}{m} + \alpha g(\theta_1^{(i)})\right) -$$

$$\sum_{i=1}^{\alpha N} \log\left(\frac{f(\mathbf{y}|\theta_0^{(i)})\pi(\theta_0^{(i)})}{m} + \alpha g(\theta_0^{(i)})\right)$$

$$mq'(m) = -N + \sum_{i=1}^{\alpha N} \frac{h_{0,(i)}}{h_{0,(i)} + \alpha m} + \sum_{i=1}^{N} \frac{h_{1,(i)}}{h_{1,(i)} + \alpha m}$$

$$mq'(m) = \sum_{i=1}^{\alpha N} \frac{h_{0,(i)}}{h_{0,(i)} + \alpha m} - \sum_{i=1}^{N} \frac{\alpha m}{h_{1,(i)} + \alpha m}$$

Noise-contrastive estimation

Let $\hat{m}$ be the solution of $\hat{m}q'(\hat{m}) = 0$

$$\iff \alpha m \sum_{i=1}^{N} \frac{1}{h_{1,(i)} + \alpha m} = \sum_{i=1}^{\alpha N} \frac{h_{0,(i)}}{h_{0,(i)} + \alpha m}$$

$\hat{m}$ is equivalent to the optimal bridge estimator if m

Optimal bridge estimator solution of

$$\alpha \hat{m}^{(t+1)} \sum_{i=1}^{N} \frac{1}{h_{1,(i)} + \alpha \hat{m}^{(t)}} = \sum_{i=1}^{\alpha N} \frac{h_{0,(i)}}{h_{0,(i)} + \alpha \hat{m}^{(t)}}$$

Noise-contrastive estimation

Let $c = -\log(m)$

Logistic regression approximation

$$\log\left(\frac{\mathbb{P}(z=1|\mathbf{y},\theta)}{\mathbb{P}(z=0|\mathbf{y},\theta)}\right) = -\log(m) + \log\left(\frac{f(\mathbf{y}|\theta)\pi(\theta)}{\alpha g(\theta)}\right)$$

$$\log\left(\frac{\mathbb{P}(z=1|\mathbf{y},\theta)}{\mathbb{P}(z=0|\mathbf{y},\theta)}\right) = c + \log\left(\frac{f(\mathbf{y}|\theta)\pi(\theta)}{\alpha g(\theta)}\right)$$

Noise-contrastive estimation

Let m^* be the true value of m that is

$$m^* = \int f(\mathbf{y}|\theta)\pi(\theta)d\theta$$

Pseudo likelihood paradigm $\Longrightarrow$

$$\sqrt{N}(\hat{c} - c^*) \longrightarrow$$

$$N\left(0, \left[\int \frac{\alpha g(\theta)\pi(\theta|\mathbf{y})}{\pi(\theta|\mathbf{y}) + \alpha g(\theta)} d\theta\right]^{-1} - (1 + 1/\alpha)\right)$$

$$c = -\log(m) \quad ; \quad c^* = -\log(m^*)$$

Pseudo likelihood paradigm $\Longrightarrow$ **Weighted likelihood bootstrap to estimate the variance of** $\hat{c}$

Weighted likelihood bootstrap for noise-contrastive estimation

In all of the following, many regularity conditions are assumed

$\ell(\theta) = f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$ a parametric family with $\theta \in \mathbb{R}$

$\pi(\theta)$ a prior distribution

Le Cam (1956) **Bernstein-von Mises theorem**

$$\boxed{\left(\theta - \hat{\theta}_n\right)|\mathbf{x} \approx \mathcal{N}(0, \hat{\sigma}_n) \quad (\text{for } n \text{ large})}$$

$\hat{\theta}_n$ is the MLE of θ and $\hat{\sigma}_n = \left(-\frac{\partial^2 \left(\sum_{i=1}^{n} \log f(x_i|\theta)\right)}{(\partial\theta)^2}(\hat{\theta}_n)\right)^{-1}$

Weighted likelihood bootstrap for noise-contrastive estimation

Newton and Raftery (1994)

Let $\omega = (\omega_1, \ldots, \omega_n)$ has a uniform Dirichlet distribution
The associated weighted likelihood function is

$$\tilde{\ell}(\theta) = \prod_{i=1}^{n} f(x_i|\theta)^{\omega_i}$$

$\tilde{\theta}_n$ is the maximum value of $\tilde{\ell}(\theta)$

The conditional distribution of $\tilde{\theta}_n$ is a good approximation of the posterior posterior distribution of θ

$$\boxed{\left(\tilde{\theta}_n - \hat{\theta}_n\right)|\mathbf{x} \approx \mathscr{N}(0, \hat{\sigma}_n) \quad (\text{for } n \text{ large})}$$

Weighted likelihood bootstrap for noise-contrastive estimation

Finally, recall that

$$\boxed{\left(\hat{\theta}_n - \theta\right) \approx \mathcal{N}(0, \hat{\sigma}_n) \quad \text{(for } n \text{ large)}}$$

$\Longrightarrow \mathbb{V}(\hat{\theta}_n) \approx \hat{\sigma}_n$

The variance of the MLE can be approximate by using the empirical variance of $\tilde{\theta}_n$

Weighted likelihood bootstrap for noise-contrastive estimation

Sample the ω_i independently from an exponential distribution with parameter equal to 1 and renormalize

Calculate $\tilde{\theta}_n$ (the maximum value of $\tilde{\ell}(\theta)$)

Repeat the two previous steps several times and estimate the variance of $\hat{\theta}_n$ with the empirical variance of the $\tilde{\theta}_n$

As we are in a specific pseudo likelihood context some corrections are needed

Weighted likelihood bootstrap for noise-contrastive estimation

The basic Weighted Likelihood bootstrap would be based on the following weighted likelihood

$$-N\log(m) - \sum_{i=1}^{N} \omega_{i,1} \log\left(\frac{f(\mathbf{y}|\theta_1^{(i)})\pi(\theta_1^{(i)})}{m} + \alpha g(\theta_1^{(i)})\right) -$$

$$\sum_{i=1}^{\alpha N} \omega_{i,0} \log\left(\frac{f(\mathbf{y}|\theta_0^{(i)})\pi(\theta_0^{(i)})}{m} + \alpha g(\theta_0^{(i)})\right)$$

Weighted likelihood bootstrap for noise-contrastive estimation

Corrected version

$$-N\log(m)-\sum_{i=1}^{N}\frac{N\omega_{i,1}}{\sum_{i=1}^{N}\omega_{i,1}}\log\left(\frac{f(\mathbf{y}|\theta_1^{(i)})\pi(\theta_1^{(i)})}{m}+\alpha g(\theta_1^{(i)})\right)-$$

$$\sum_{i=1}^{\alpha N}\frac{N\omega_{i,0}}{\sum_{i=1}^{\alpha N}\omega_{i,0}}\log\left(\frac{f(\mathbf{y}|\theta_0^{(i)})\pi(\theta_0^{(i)})}{m}+\alpha g(\theta_0^{(i)})\right)$$

A toy example

$y|\theta \sim \mathcal{N}(\theta, 1)$
$\theta \sim \mathcal{N}(0, 1)$

In such a case

$$m = \exp(-y^2/4)/\sqrt{4\pi}$$
$$\theta|y \sim \mathcal{N}(y/2, \sqrt{1/2})$$

```
y <- 5
target <- -dnorm(y,0,sqrt(2),log=TRUE)
target
[1] 7.515512
```

A toy example

```
# lh  <- log(f(y|theta))+log(pi(theta))-g(theta)

mqprime <- function(const,Nsim,lh)
{
-Nsim+sum(exp(lh)/(exp(lh)+exp(-const)))
}

mqprimew <- function(const,Nsim,lh,w)
{
-sum(w[1:Nsim])+sum(w*exp(lh)/(exp(lh)+exp(-const)))
}
```

A toy example

```
Nsim <- 10^5 ; m <- 3 ; sig <- 0.8

thetapost <- rnorm(Nsim,mean=y/2,sd=sqrt(1/2))
thetag <- rnorm(Nsim,mean=m,sd=sqrt(sig))
zeta <- c(thetapost,thetag)

lh  <- dnorm(zeta,mean=y,log=TRUE)+
dnorm(zeta,log=TRUE)-
dnorm(zeta,mean=m,sd=sqrt(sig),log=TRUE)

bridge <- uniroot(mqprime,Nsim=Nsim,lh=lh,
c(target-1,target+1),tol=.Machine$double.eps^0.5)$root

bridge
[1] 7.515067
```

A toy example

Variability of the bridge estimate via Monte Carlo replicates

```
Nsim <- 10^5 ; N <- 100 ; monte.carlo <- rep(0,N)
for (i in 1:N)
{
thetapost <- rnorm(Nsim,mean=y/2,sd=sqrt(1/2))
thetag <- rnorm(Nsim,mean=m,sd=sqrt(sig))
zeta <- c(thetapost,thetag)
lh  <- dnorm(zeta,mean=y,log=TRUE)+
dnorm(zeta,log=TRUE)-
dnorm(zeta,mean=m,sd=sqrt(sig),log=TRUE)
monte.carlo[i] <- uniroot(mqprime,Nsim=Nsim,lh=lh,
c(target+1,target-1),tol=.Machine$double.eps^0.5)$root
}
sd(sqrt(Nsim)*monte.carlo)
[1] 0.4721598
```

A toy example

Variability of the bridge estimate via Weighted Likelihood Bootstrap

```
Nsim <- 10^5
thetapost <- rnorm(Nsim,mean=y/2,sd=sqrt(1/2))
thetag <- rnorm(Nsim,mean=m,sd=sqrt(sig))
zeta <- c(thetapost,thetag)
lh  <- dnorm(zeta,mean=y,log=TRUE)+
dnorm(zeta,log=TRUE)-
dnorm(zeta,mean=m,sd=sqrt(sig),log=TRUE)
N <- 100 ; wlb <- rep(0,N) ; for (i in 1:N) {
w1 <- rexp(Nsim) ; w1 <- w1/sum(w1)*Nsim
w2 <- rexp(Nsim) ; w2 <- w2/sum(w2)*Nsim
w <- c(w1,w2)
wlb[i] <- uniroot(mqprimew,Nsim=Nsim,lh=lh,w=w,
c(target+1,target-1),tol=.Machine$double.eps^0.5)$root
}
sd(sqrt(Nsim)*wlb)
[1] 0.4648654
```