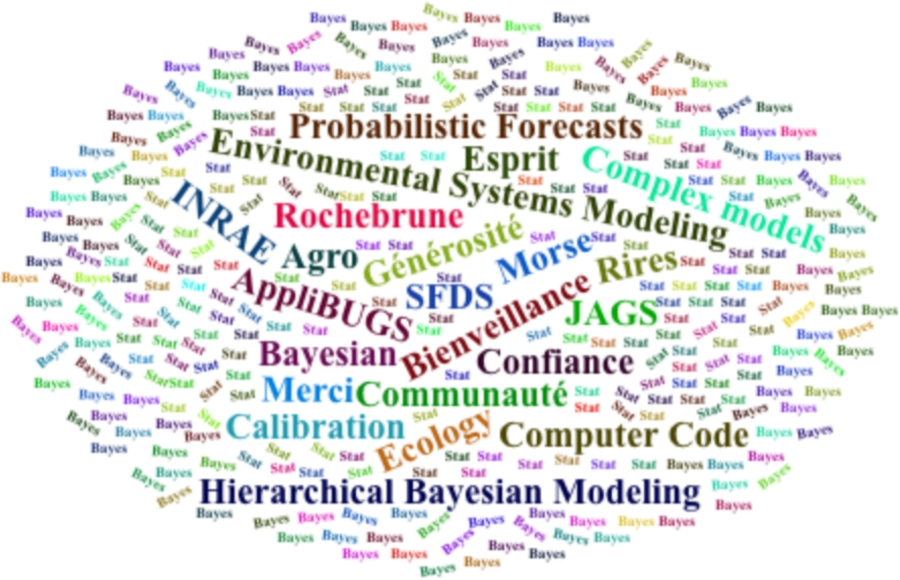
Séminaire de statistique bayésienne en l'honneur d'Éric Parent
Uncertainty quantification for marginal computations

Jean-Michel Marin

University of Montpellier, CNRS
Alexander Grothendieck Montpellier Institute

September 2022

Merci Éric


Probabilistic Forecasts
Environmental Systems Modeling
Esprit
Complex models
INRAE
Rochebrune
Générosité
Morse
Agro
Rires
AppliBUGS
SFDS
Bienveillance
JAGS
Bayesian
Confiance
Merci
Communauté
Calibration
Ecology
Computer Code
Hierarchical Bayesian Modeling
Bayes
Stat

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)

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

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 this 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 marginal likelihood approximation

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

Critiques of the marginal likelihood often note its inability to manage improper priors for hypothesis testing, sensitivity to prior assumptions, lack of uncertainty representation over hyperparameters, and its potential misuse in advocating for models with fewer parameters **Gelman et al. (2013)**

Alternative to the marginal likelihood approach: the literature contains a number of approaches intended to evaluate the quality of any given model based on their predictive capacity **Vehtari, Gelman and Gabry (2017)**

Introduction

We recall some 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

Bridge sampling techniques

Bennett (1976), 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

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

$\theta_0^{(1)}, \ldots, \theta_0^{(N)}$ is an N-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

$$h(\theta) = \frac{C}{f(\mathbf{y}|\theta)\pi(\theta)/2 + g(\theta)m/2}$$

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

$$\hat{m}^{(t+1)} = \frac{\displaystyle\frac{1}{N}\sum_{i=1}^{N} \frac{f(\mathbf{y}|\theta_0^{(i)})\pi(\theta_0^{(i)})}{f(\mathbf{y}|\theta_0^{(i)})\pi(\theta_0^{(i)})/2 + g(\theta_0^{(i)})\hat{m}^{(t)}/2}}{\displaystyle\frac{1}{N}\sum_{i=1}^{N} \frac{g(\theta_1^{(i)})}{f(\mathbf{y}|\theta_1^{(i)})\pi(\theta_1^{(i)})/2 + g(\theta_1^{(i)})\hat{m}^{(t)}/2}}$$

Bridge sampling techniques

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

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

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

Some others alternatives

Large set of approximations for marginal likelihood
or Bayes factors

- The **Chib (1995)** proposal
- **Gelman and Meng (1998)** Thermodynamic integration, path sampling
- 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

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

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

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

$$\mathbb{P}(z=0) = 1/2 \quad ; \quad \mathbb{P}(z=1) = 1/2$$

Noise-contrastive estimation

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

The *pseudo likelihood*

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

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

Noise-contrastive estimation

The *pseudo log-likelihood*

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

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

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

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

Noise-contrastive estimation

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

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

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

Optimal bridge estimator solution of

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

Noise-contrastive estimation

Let $c = -\log(m)$

Logistic regression approximation

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

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

An asymptotic result

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

$$m^* = \int f(\mathbf{y}|\theta)\pi(\theta)d\theta \quad ; \quad c^* = -\log(m^*)$$

Bennett (1976), Meng and Wong (1996)
Gutmann and Hyvarinen (2012)

$$\sqrt{N}(\hat{c} - c^*) \longrightarrow N\left(0, \left[\int \frac{g(\theta)\pi(\theta|\mathbf{y})}{\pi(\theta|\mathbf{y}) + g(\theta)} d\theta\right]^{-1} - 2\right)$$

$$\hat{c} = -\log(\hat{m})$$

Weighted likelihood bootstrap for noise-contrastive estimation

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

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

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

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 pseudo log-likelihood

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

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

Weighted likelihood bootstrap for noise-contrastive estimation

Corrected version

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

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

Discussion

We propose a very simple and low cost approach based on a modified Weighted Likelihood Bootstrap (WLB) for which we prove that the uncertainty quantification is asymptotically valid

Interest of WLB compared with the estimation of the variance and using Gaussian quantiles

$$\sqrt{N}(\hat{c} - c^*) \longrightarrow N\left(0, \left[\int \frac{g(\theta)\pi(\theta|\mathbf{y})}{\pi(\theta|\mathbf{y}) + g(\theta)} d\theta\right]^{-1} - 2\right)$$

- difficulty to estimate the integral of the same nature as the computation of c
- we propose a method which does not require such a computation