How to optimally quantify the uncertainty of the stepping-stone sampling estimate

John Siryj

26 November 2019

Our motivating problem I

Common pursuit in statistics:

- Model selection

  Bayesian $\Longrightarrow$ marginal likelihood

Want measure of uncertainty in marginal likelihood estimate

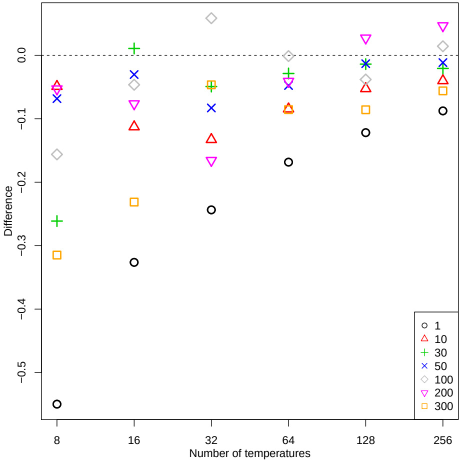
- Independent marginal likelihood estimates

  Problem: impractical
- Independent bootstrap

  Problem: underestimates when data is dependent
- Moving block bootstrap

  Problem: need to choose block length λ

Our motivating problem II



Statement of Bayes rule

Bayes rule can be written as

$$p(\boldsymbol{\theta}|\boldsymbol{X}, M) = \frac{\mathcal{L}(\boldsymbol{X}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)}{z}$$

where

- $p(\cdot)$ is the posterior
- $\mathcal{L}(\cdot)$ is the likelihood
- $\pi(\cdot)$ is the prior
- M is the model
- $\boldsymbol{\theta}$ are the parameters
- $\boldsymbol{X}$ is the data
- z is the **marginal likelihood**

Calculating the marginal likelihood

Need to solve the following

$$z = \int_{\Theta} \mathcal{L}(\boldsymbol{X}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)d\boldsymbol{\theta}$$

Modern methods often employed include:

- Power posterior methods
  - Stepping-stone sampling (Xie et al., 2010)
  - Thermodynamic integration (Gelman and Meng, 1994)
  - Generalised stepping-stone sampling (Fan et al., 2010)
- Nested sampling (Skilling, 2004)

Power posterior

Note that we can modify Bayes rule as

$$p_\beta(\boldsymbol{\theta}|\boldsymbol{X}, M) = \frac{\mathcal{L}(\boldsymbol{X}|\boldsymbol{\theta}, M)^\beta \pi(\boldsymbol{\theta}|M)}{z_\beta}$$

where $z_\beta = \int_\Theta \mathcal{L}(\boldsymbol{X}|\boldsymbol{\theta}, M)^\beta \pi(\boldsymbol{\theta}|M) d\boldsymbol{\theta}$

Note that

- $\beta = 0 \implies$ prior
- $\beta = 1 \implies$ posterior

Thus, it defines a path between prior and posterior

Stepping-stone sampling

Marginal likelihood can be seen as the ratio z_1/z_0. Expand out as a telescopic product

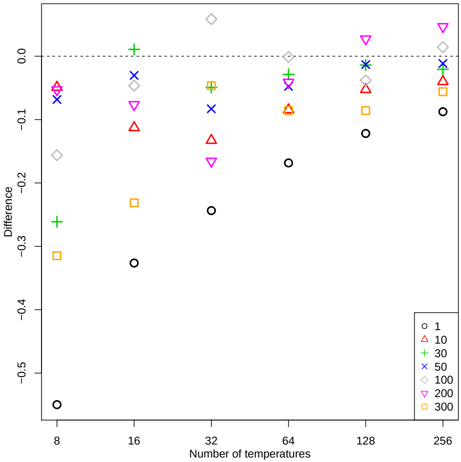
$$z = \frac{z_1}{z_0} = \frac{z_{\beta_1}}{z_{\beta_0}}\frac{z_{\beta_2}}{z_{\beta_1}}\cdots\frac{z_{\beta_{K-2}}}{z_{\beta_{K-3}}}\frac{z_{\beta_{K-1}}}{z_{\beta_{K-2}}} = \prod_{k=1}^{K-1} r_k$$

where $r_k = z_{\beta_k}/z_{\beta_{k-1}}$

Approximated by the Monte Carlo estimator

$$\hat{z}_{SS} = \prod_{k=1}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\boldsymbol{X}|\boldsymbol{\theta}^i_{\beta_{k-1}}, M)^{\beta_k-\beta_{k-1}}$$

Bootstrapping for dependent data I



Bootstrapping for dependent data II

On dependent data the bootstrap approach of Efron (1979)

- destroys dependence structure
- underestimates uncertainty

Block bootstrap approaches like that of Künsch (1989) are preferred

Moving block bootstrap

With a block bootstrap approach

- Sample blocks of consecutive points
- Many different approaches exist to divide data
- Depends on length parameter λ

Moving block bootstrap example

Suppose we have data $\{1, 3, 7, 2, 9, 8\}$ and $\lambda = 2$.

The (Künsch) blocks are then

$$B_1 = \{1, 3\},\ B_2 = \{3, 7\},\ B_3 = \{7, 2\},\ B_4 = \{2, 9\},\ B_5 = \{9, 8\}$$

If we resample blocks B_4, B_3, and B_1 then we have a new bootstrap dataset

$$\{B_4, B_3, B_1\} = \{2, 9, 7, 2, 1, 3\}$$

Motivation

One major question remains, how do we choose the "best" block length λ for our series?

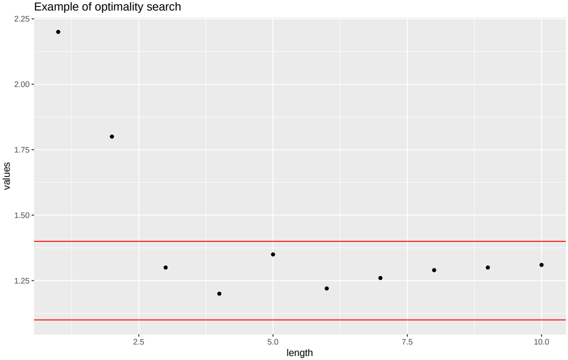
For AR and MA models:

- See Hall et al. (1995)
- See Lahiri (1999)

Our approach:

- Empirical approach

Our Approach



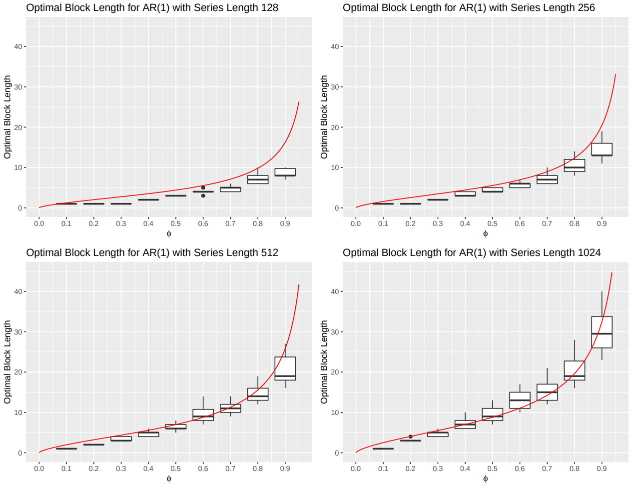
Problem:
Try to find minimal block length which produces uncertainty within a certain range of the standard deviation distribution.

Example I

We considered an AR(1) model with varied parameter ϕ

Wanted to consider optimal block length for the calculation of the uncertainty in the estimate of $\hat{\phi}$

Example II


Optimal Block Length for AR(1) with Series Length 128
Optimal Block Length for AR(1) with Series Length 256
Optimal Block Length for AR(1) with Series Length 512
Optimal Block Length for AR(1) with Series Length 1024
Optimal Block Length
ϕ
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
0 10 20 30 40

Tempered Gaussian model I

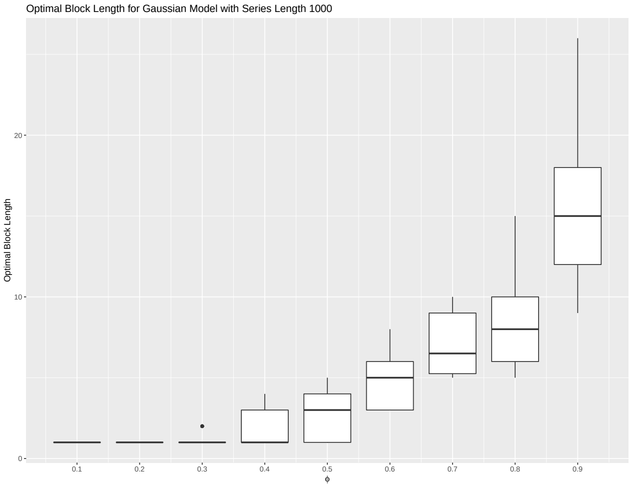
Consider a Gaussian model parametrised by

- Prior: $\theta_i \overset{\text{iid}}{\sim} \mathrm{N}(0,1)$
- Likelihood: $L(\boldsymbol{\theta}) = \prod_{i=1}^{d} \exp(-\theta_i^2/2\nu)$, ν fixed parameter
- Power posterior: $\mathrm{N}(0, \nu/(\nu+\beta))$, for inverse temperature β

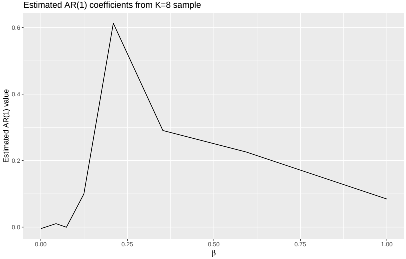
This has exact marginal likelihood $z = (\nu/(1+\nu))^{d/2}$

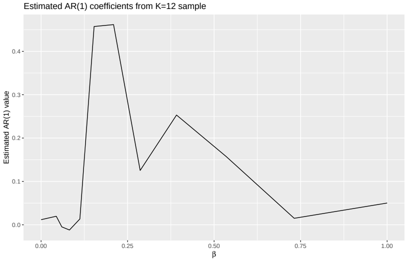
Could sample independently, but want to test dependent case. Thus sample AR(1) with corresponding scale factors. Approach is easier than Metropolis-Hastings.

Parallel tempered Gaussian with significant dependence



More realistic scenario





Gravitational wave data background

Simulated black hole coalescence signal in the Advanced LIGO and Advanced Virgo GW detectors. The specifications:

- Masses: 25 and 13 $M_\odot$
- Luminosity distance: 614 Mpc
- Signal-to-noise ratio: 17.9 in the 3 detector network

Gravitational wave data

In Maturana-Russel et al. (2019) the authors used the previous data.

They employed

- Stepping-stone sampling for marginal likelihood estimation
- Random grid search for optimal block length
- 1000 independent block bootstrap samples per estimate

They reported most conservative uncertainty estimate out of all block bootstraps

Our results

Wanting to improve on the results of Maturana-Russel et al. (2019) we have applied the optimal block length strategy and have gotten the following results.

K	$\hat{z}$	25%	Median	75%	Original
8	$-5730.064\pm$	0.340	0.340	0.344	0.40
12	$-5729.999\pm$	0.144	0.160	0.176	0.32

Table: Estimates for the marginal likelihood ± 1 SD of uncertainty across inverse temperatures $K = 8, 12$ using optimal block length

Conclusion

As we saw above using a random grid search approach could possibly overestimate the uncertainty in a marginal likelihood calculation in dependent data.

Future Work /Extensions

We have identified the following as possible extensions or uses of the approaches considered

- Generalised stepping-stone sampling algorithm estimates (Fan et al., 2010)
- Use on penalty term of DIC (Gelman et al., 2004)
- Application on direct Bayes factor calculation (Baele et al., 2013)

References I

Baele, G., Lemey, P., and Vansteelandt, S. (2013). Make the most of your samples: Bayes factor estimators for high-dimensional models of sequence evolution. *BMC Bioinformatics*, 14(1):85.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Fan, Y., Wu, R., Chen, M.-H., Kuo, L., and Lewis, P. O. (2010). Choosing among partition models in bayesian phylogenetics. *Molecular biology and evolution*, 28(1):523–532.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

References II

Gelman, A. and Meng, X. (1994). Path sampling for computing normalizing constants: identities and theory. *University of Chicago Department of Statistics Technical Report*, 1(377).

Hall, P., Horowitz, J. L., and Jing, B.-Y. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika*, 82(3):561–574.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241.

Lahiri, S. N. (1999). Theoretical comparisons of block bootstrap methods. *The Annals of Statistics*, 27(1):386–404.

References III

Lartillot, N. and Philippe, H. (2006). Computing bayes factors using thermodynamic integration. *Systematic biology*, 55(2):195–207.

Maturana-Russel, P., Meyer, R., Veitch, J., and Christensen, N. (2019). Stepping-stone sampling algorithm for calculating the evidence of gravitational wave models. *Physical Review D*, 99(8):084006.

Skilling, J. (2004). Nested sampling. In *AIP Conference Proceedings*, volume 735, pages 395–405. AIP.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., and Chen, M.-H. (2010). Improving Marginal Likelihood Estimation for Bayesian Phylogenetic Model Selection. *Systematic Biology*, 60(2):150–160.

Our algorithm I

1. Suppose we are given a collection of $\mathcal{N}$ estimates $\hat{\theta}_i$, $i \in \{1, \ldots, \mathcal{N}\}$ for some unknown parameter of interest θ
2. Using our $\hat{\theta}_i$'s we are able to get an estimate of the uncertainty in our collection call it $\sigma_{\hat{\theta}}$
3. Repeatedly obtaining collections of $\hat{\theta}_i$'s a reasonable amount of times allows up to build up likely bounds $\left(\sigma_{\hat{\theta}}^{lower}, \sigma_{\hat{\theta}}^{upper}\right)$ for $\sigma_{\hat{\theta}}$
4. Now obtain a single $\hat{\theta}$
5. Starting at block length λ_0 and applying a (moving) block bootstrap with block length λ allows us to produce a bootstrap sample $\hat{\theta}^*$

Our algorithm II

6. Repeat the last step a reasonable amount of times to produce an uncertainty estimate $\hat{\sigma}_{\hat{\theta}|\lambda}$
7. If the estimate $\hat{\sigma}_{\hat{\theta}|\lambda}$ lies with the bounds $\left(\sigma_{\hat{\theta}}^{lower}, \sigma_{\hat{\theta}}^{upper}\right)$ for $\sigma_{\hat{\theta}}$ output $\hat{\sigma}_{\hat{\theta}|\lambda}$ as the optimal block length
8. Otherwise, increment λ and repeat steps 5-7 until an optimal block length is output or the global maximum search value is reached
   - It is important to have some reasonable upper bound on the largest value to be considered in case the algorithm manages to miss the convergence window
   - We must make sure this bound is high enough to not be encountered by the vast majority of samples, yet low enough so as to help restart any transient samples

Our algorithm III

9. Repeat this process until a specified large number of optimal block length have been found
10. Apply the median to the collection of optimal block lengths to find the "true" optimal block length
    - One might want to output quantiles or confidence intervals for the median to give a more complete picture of the "true" optimal block length