Journal of Mathematical Psychology ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Tutorial

A tutorial on bridge sampling

Quentin F. Gronau [a,*], Alexandra Sarafoglou [a], Dora Matzke [a], Alexander Ly [a], Udo Boehm [a], Maarten Marsman [a], David S. Leslie [b], Jonathan J. Forster [c], Eric-Jan Wagenmakers [a], Helen Steingroever [a]

[a] Department of Psychology, University of Amsterdam, The Netherlands
[b] Department Mathematics and Statistics, Lancaster University, UK
[c] Mathematical Science, University of Southampton, UK

HIGHLIGHTS

- We provide a tutorial on bridge sampling for estimating marginal likelihoods.
- We use the beta-binomial model as a running example.
- We estimate the marginal likelihood for the Expectancy Valence (EV) model.
- We obtain accurate results for individual-level and hierarchical EV model versions.

ARTICLE INFO

Article history:
Received 17 March 2017
Received in revised form 31 August 2017
Available online xxxx

Keywords:
Hierarchical model
Normalizing constant
Marginal likelihood
Bayes factor
Predictive accuracy
Reinforcement learning

ABSTRACT

The marginal likelihood plays an important role in many areas of Bayesian statistics such as parameter estimation, model comparison, and model averaging. In most applications, however, the marginal likelihood is not analytically tractable and must be approximated using numerical methods. Here we provide a tutorial on bridge sampling (Bennett, 1976; Meng & Wong, 1996), a reliable and relatively straightforward sampling method that allows researchers to obtain the marginal likelihood for models of varying complexity. First, we introduce bridge sampling and three related sampling methods using the beta-binomial model as a running example. We then apply bridge sampling to estimate the marginal likelihood for the Expectancy Valence (EV) model—a popular model for reinforcement learning. Our results indicate that bridge sampling provides accurate estimates for both a single participant and a hierarchical version of the EV model. We conclude that bridge sampling is an attractive method for mathematical psychologists who typically aim to approximate the marginal likelihood for a limited set of possibly high-dimensional models.

Bayesian statistics has become increasingly popular in mathematical psychology (Andrews & Baguley, 2013; Bayarri, Benjamin, Berger, & Sellke, 2016; Poirier, 2006; Vanpaemel, 2016; Verhagen, Levy, Millsap, & Fox, 2015; Wetzels, Tutschkow, Dolan, van der Sluis, Dutilh, & Wagenmakers, 2016). The Bayesian approach is conceptually simple, theoretically coherent, and easily applied to relatively complex problems. These problems include, for instance, hierarchical modeling (Matzke, Dolan, Batchelder, & Wagenmakers, 2015; Matzke & Wagenmakers, 2009; Rouder & Lu, 2005; Rouder, Lu, Speckman, Sun, & Jiang, 2005; Rouder, Lu, Sun, Speckman, Morey, & Naveh-Benjamin, 2007) or the comparison of non-nested models (Lee, 2008; Pitt, Myung, & Zhang, 2002; Shiffrin, Lee, Kim, & Wagenmakers, 2008). Three major applications of Bayesian statistics concern parameter estimation, model comparison, and Bayesian model averaging. In all three areas, the marginal likelihood – that is, the probability of the observed data given the model of interest – plays a central role (see also Gelman & Meng, 1998).

First, in parameter estimation, we consider a single model and aim to quantify the uncertainty for a parameter of interest θ after having observed the data y. This is realized by means of a posterior distribution that can be obtained using Bayes theorem:

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{\int p(y \mid \theta')\, p(\theta')\, \mathrm{d}\theta'} = \frac{\overbrace{p(y \mid \theta)}^{\text{likelihood}}\ \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(y)}_{\text{marginal likelihood}}}. \quad (1)$$

* Correspondence to: Department of Psychology, PO Box 15906, 1001 NK Amsterdam, The Netherlands.
E-mail address: Quentin.F.Gronau@gmail.com (Q.F. Gronau).

https://doi.org/10.1016/j.jmp.2017.09.005

Here, the marginal likelihood of the data $p(y)$ ensures that the posterior distribution is a proper probability density function (PDF) in the sense that it integrates to 1. This illustrates why in parameter estimation the marginal likelihood is referred to as a normalizing constant.

Second, in model comparison, we consider m ($m \in \mathbb{N}$) competing models, and are interested in the relative plausibility of a particular model $\mathcal{M}_i$ ($i \in \{1, 2, \ldots, m\}$) given the prior model probability and the evidence from the data y (see three special issues on this topic in the *Journal of Mathematical Psychology*: Mulder & Wagenmakers, 2016; Myung, Forster, & Browne, 2000; Wagenmakers & Waldorp, 2006). This relative plausibility is quantified by the so-called posterior model probability $p(\mathcal{M}_i \mid y)$ of model $\mathcal{M}_i$ given the data y (Berger & Molina, 2005):

$$p(\mathcal{M}_i \mid y) = \frac{p(y \mid \mathcal{M}_i)\, p(\mathcal{M}_i)}{\sum_{j=1}^{m} p(y \mid \mathcal{M}_j)\, p(\mathcal{M}_j)}, \tag{2}$$

where the denominator is the sum of the marginal likelihood times the prior model probability of all m models. In model comparison, the marginal likelihood for a specific model is also referred to as the model evidence (Didelot, Everitt, Johansen, & Lawson, 2011), the integrated likelihood (Kass & Raftery, 1995), the predictive likelihood of the model (Gamerman & Lopes, 2006, chapter 7), the predictive probability of the data (Kass & Raftery, 1995), or the prior predictive density (Ntzoufras, 2009). Note that conceptually the marginal likelihood of Eq. (2) is the same as the marginal likelihood of Eq. (1). However, for the latter equation we droped the model index because in parameter estimation we consider only one model.

If only two models $\mathcal{M}_1$ and $\mathcal{M}_2$ are considered, Eq. (2) can be used to quantify the relative posterior model plausibility of model $\mathcal{M}_1$ compared to model $\mathcal{M}_2$. This relative plausibility is given by the ratio of the posterior probabilities of both models, and is referred to as the posterior model odds:

$$\underbrace{\frac{p(\mathcal{M}_1 \mid y)}{p(\mathcal{M}_2 \mid y)}}_{\substack{\text{posterior}\\ \text{odds}}} = \underbrace{\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}}_{\substack{\text{prior}\\ \text{odds}}} \times \underbrace{\frac{p(y \mid \mathcal{M}_1)}{p(y \mid \mathcal{M}_2)}}_{\substack{\text{Bayes}\\ \text{factor}}}. \tag{3}$$

Eq. (3) illustrates that the posterior model odds are the product of two factors: The first factor is the ratio of the prior probabilities of both models—the prior model odds. The second factor is the ratio of the marginal likelihoods of both models—the so-called Bayes factor (Etz & Wagenmakers, in press; Jeffreys, 1961; Ly, Verhagen, & Wagenmakers, 2016a, b; Robert, 2016). The Bayes factor plays an important role in model comparison and is referred to as the "standard Bayesian solution to the hypothesis testing and model selection problems" (Lewis & Raftery, 1997, p. 648) and "the primary tool used in Bayesian inference for hypothesis testing and model selection" (Berger, 2006, p. 378).

Third, the marginal likelihood plays an important role in Bayesian model averaging (BMA; Hoeting, Madigan, Raftery, & Volinsky, 1999) where aspects of parameter estimation and model comparison are combined. As in model comparison, BMA considers several models; however, it does not aim to identify a single best model. Instead it fully acknowledges model uncertainty. Model averaged parameter inference can be obtained by combining, across all models, the posterior distribution of the parameter of interest weighted by each model's posterior model probability, and as such depends on the marginal likelihood of the models. This procedure assumes that the parameter of interest has identical interpretation across the different models. Model averaged predictions can be obtained in a similar manner.

A problem that arises in all three areas – parameter estimation, model comparison, and BMA – is that an analytical expression of the marginal likelihood can be obtained only for certain restricted examples. This is a pressing problem in Bayesian modeling, and in particular in mathematical psychology where models can be nonlinear and equipped with a large number of parameters, especially when the models are implemented in a hierarchical framework. Such a framework incorporates both commonalities and differences between participants of one group by assuming that the model parameters of each participant are drawn from a group-level distribution (for advantages of the Bayesian hierarchical framework see Ahn, Krawitz, Kim, Busemeyer, & Brown, 2011; Navarro, Griffiths, Steyvers, & Lee, 2006; Rouder & Lu, 2005; Rouder et al., 2005; Rouder, Lu, Morey, Sun, & Speckman, 2008; Scheibehenne & Pachur, 2015; Shiffrin et al., 2008; Wetzels, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2010). For instance, consider a four-parameter Bayesian hierarchical model with four group-level distributions each characterized by two parameters and a group size of 30 participants; this then results in 30×4 individual-level parameters and 2×4 group-level parameters for a total of 128 parameters. In sum, even simple models quickly become complex once hierarchical aspects are introduced and this frustrates the derivation of the marginal likelihood.

To overcome this problem, several Monte Carlo sampling methods have been proposed to approximate the marginal likelihood. In this tutorial we focus on four such methods: the bridge sampling estimator (Bennett, 1976, Chapter 5 of Chen, Shao, & Ibrahim, 2012; Meng & Wong, 1996), and its three commonly used special cases—the naive Monte Carlo estimator, the importance sampling estimator, and the generalized harmonic mean estimator (for alternative methods see Gamerman & Lopes, 2006, Chapter 7; and for alternative approximation methods relevant to model comparison and BMA see Carlin & Chib, 1995; Green, 1995).[1] As we will illustrate throughout this tutorial, the bridge sampler is accurate, efficient, and relatively straightforward to implement (e.g., DiCiccio, Kass, Raftery, & Wasserman, 1997; Frühwirth-Schnatter, 2004; Meng & Wong, 1996).

The goal of this tutorial is to bring the bridge sampling estimator to the attention of mathematical psychologists. We aim to explain this estimator and facilitate its application by suggesting a step-by-step implementation scheme. To this end, we first show how bridge sampling and the three special cases can be used to approximate the marginal likelihood in a simple beta-binomial model. We begin with the naive Monte Carlo estimator and progressively work our way up – via the importance sampling estimator and the generalized harmonic mean estimator – to the most general case considered: the bridge sampling estimator. This order was chosen such that key concepts are introduced gradually and estimators are of increasing complexity and sophistication. The first three estimators are included in this tutorial with the sole purpose of facilitating the reader's understanding of bridge sampling. In the second part of this tutorial, we outline how the bridge sampling estimator can be used to derive the marginal likelihood for the Expectancy Valence (EV; Busemeyer & Stout, 2002) model—a popular, yet relatively complex reinforcement-learning model for the Iowa gambling task (Bechara, Damasio, Damasio, & Anderson, 1994). We apply bridge sampling to both an individual-level and a hierarchical implementation of the EV model.

Throughout the article, we use the software package R to implement the bridge sampling estimator for the various models. The interested reader is invited to reproduce our results by downloading the code and all relevant materials from our Open Science Framework folder at osf.io/f9cq4.

[1] The Appendix provides a derivation showing that the first three estimators are indeed special cases of the bridge sampler.