A probabilistic diagnostic tool to assess Laplace approximations: proof of concept and non-asymptotic experimentation

Shaun McDonald, Dave Campbell, Haoxuan Zhou

June 3, 2020

Abstract

In many statistical models, we need to integrate functions that may be high-dimensional. Such integrals may be impossible to compute exactly, or too expensive to compute numerically. Instead, we can use the *Laplace approximation* for the integral. This approximation is exact if the function is proportional to the density of a normal distribution; therefore, its effectiveness may depend intimately on the true shape of the function. To assess the quality of the approximation, we use *probabilistic numerics*: recasting the approximation problem in the framework of probability theory. In this probabilistic approach, uncertainty and variability don't come from a frequentist notion of randomness, but rather from the fact that the function may only be partially known. We use this framework to develop a diagnostic tool for the Laplace approximation, modelling the function and its integral as a Gaussian process and devising a "test" by conditioning on a finite number of function values.We will discuss approaches for designing and optimizing such a tool and demonstrate it on known sample functions, highlighting in particular the challenges one may face in high dimensions.

1 Introduction

Coming soon. Some combination of abstract (above) and framework (below). Specifically mention:

1. That we are building on the work of Zhou [2]
2. That this is non-asymptotic and not intended as a substitute for full-on MC integration or BQ - rather as a "middle-ground" amount of effort.
3. The goal is to "test" the assumptions underlying the Laplace approximation (e.g. "how Gaussian is this function?"). The Laplace approximation may still hold for a non-Gaussian shape, but such a function should be

rejected by our diagnostic ("sufficiently non-Gaussian things warrant further attention), at which point a more involved integration would show that the approximation was fine after all.

2 Framework and notation

Consider a positive function $f : \mathbb{R}^d \to \mathbb{R}_{>0}$. The object of interest for the diagnostic is the integral $F = \int_{\mathbb{R}^d} f(t)\mathrm{d}t$. In practical applications `[citation needed]`, typically f and F are actually functions with an additional argument vector of structural parameters θ, with $t \in \mathbb{R}^d$ a vector of nuisance parameters to be marginalized. For instance, f may be a joint probability density for (θ, t), in which case F would be the marginal distribution of θ after integrating over t. To reflect this common setting, Zhou [2] called f and F the *full* and *target* functions, respectively. For the present discussion, non-marginalized arguments θ are not relevant, so any dependence on them is omitted and f and F are simply called the *true* function and integral, respectively.

Suppose now that f has all second-order partial derivatives[1], and a (local) maximum at some point $\hat{t} \in \mathbb{R}^d$. To reflect the use case where f is a density, $\hat{t}$ is called a *mode*. Let H be the Hessian of $\log f$ at $\hat{t}$, and suppose that it is negative-definite (i.e. that f is log-concave at the mode). The first step in arriving at the Laplace approximation `[citation needed]` for F is to take a second-order Taylor expansion of $\log f$ about $\hat{t}$. Noting that all first-order partial derivatives of $\log f$ are equal to zero at the mode, this approximation is

$$\log f(t) \approx \log f\left(\hat{t}\right) + \frac{1}{2}\left(t - \hat{t}\right)^\top H \left(t - \hat{t}\right). \tag{1}$$

Exponentiating the right side of (1) gives an approximation for f in the form of (up to normalizing constants) a Gaussian density centered at $\hat{t}$ with covariance matrix $-H^{-1}$. In turn, integrating this exponentiated function produces the *Laplace approximation*[2]

$$\begin{aligned} F \approx L\left(f\right) &:= f\left(\hat{t}\right) \int_{R^d} \exp\left[\frac{1}{2}\left(t - \hat{t}\right)^\top H \left(t - \hat{t}\right)\right] \mathrm{d}t \\ &= f\left(\hat{t}\right) \sqrt{(2\pi)^d \det\left(-H^{-1}\right)}. \end{aligned} \tag{2}$$

The Laplace approximation is exact (or "true") if f is itself proportional to a Gaussian density. There are other function shapes for which this may be the case, but such instances may be thought of as "coincidence". Certainly, the construction of the Laplace approximation via (1) is based on an assumption of approximately Gaussian shape, and this is assumption is our primary interest in developing a diagnostic.

[1] TODO: check actual assumptions for Laplace. Are third derivatives necessary?

[2] TODO: get citation for this. In particular, there are a couple of variations I've seen in the form for "Laplace's method" in the Bayesian literature. For instance, sometimes $\log f$ is multiplied by a constant (usually sample size) before exponentiating.

3 Probabilistic numerics and Bayesian quadrature

[3]Broadly speaking, probabilistic numerics is the use of probability theory, from a somewhat Bayesian perspective, to simultaneously perform estimation and uncertainty quantification in standard numerical problems `[citation needed]`. For instance, Chkrebtii et al. [1] developed a probabilistic solver for differential equations. For a given equation, they jointly modelled the function and its derivatives with a Gaussian process prior, then sequentially conditioned on true derivative values to conduct posterior inference on the entire solution.

The approach briefly described above - using Gaussian process priors and finitely many function values to obtain posteriors for the functions and quantities of interest - is at the core of many probabilistic numerical methods. In particular, it is the standard framework with which *Bayesian quadrature* (BQ) is usually conducted. `[COMING SOON: citations and context for BQ. As ''original'' as possible]`

The machinery of BQ can be used to develop a probabilistic diagnostic for the Laplace approximation, as in [2]. Recalling the notation of Section 2, f is modelled with a Gaussian process prior. The mean function of the GP prior, m_0^t, is taken to be the Gaussian function underpinning (1) and (2):

$$m_0^t(t) := f\left(\hat{t}\right) \exp\left[\frac{1}{2}\left(t-\hat{t}\right)^\top H\left(t-\hat{t}\right)\right], t \in \mathbb{R}^d. \tag{3}$$

The covariance operator for the GP is a (positive-definite) kernel C_0^t on $\mathbb{R}^d \times \mathbb{R}^d$ to be defined later.

By the projection property of Gaussian processes `[citation/clarification needed - will fill in later]`, such a prior on f induces a scalar Normal prior on F with mean $m_0 := \int_{R^d} m_0^t(t)\mathrm{d}t = L(f)$ and variance $C_0 := \int_{R^d}\int_{R^d} C_0^t(t,u)\mathrm{d}t\mathrm{d}u$, provided all relevant quantities exist and are finite.[4]

In what follows, let $\boldsymbol{s} = (s_1, \ldots, s_n)^\top \in \mathbb{R}^{d\times n}$ be a row-wise concatenation of n vectors in $\mathbb{R}^d$. Then, for instance, the notation $f(\boldsymbol{s})$ will refer to the column vector $(f(s_1), \ldots, f(s_n))^\top \in \mathbb{R}^n$, and $C_0^t(\boldsymbol{s}, \boldsymbol{s})$ will denote the $n \times n$ matrix with $(i,j)^{\text{th}}$ entry $C_0^t(s_i, s_j)$. As in [2][5], one may use true function values at the *interrogation points* $\boldsymbol{s}$ to obtain a posterior distribution for f (with a slight abuse of notation):

$$f \mid [f(\boldsymbol{s})] \sim \mathcal{GP}\left(m_1^t, C_t^1\right), \tag{4}$$

$$m_1^t(t) = m_0^t(t) + C_t^0(t,\boldsymbol{s})^\top \left[C_t^0(\boldsymbol{s},\boldsymbol{s})\right]^{-1}\left(f(\boldsymbol{s}) - m_0^t(\boldsymbol{s})\right), \tag{5}$$

$$C_1^t(t,u) = C_0^t(t,u) - C_t^0(t,\boldsymbol{s})^\top \left[C_t^0(\boldsymbol{s},\boldsymbol{s})\right]^{-1} C_t^0(u,\boldsymbol{s}). \tag{6}$$

[3]TODO: pad out the introductory PN/BQ stuff.

[4]TODO: check this and find citations. Need to check the conditions under which it holds on an infinite domain and state those more precisely

[5]And maybe another PN/BQ citation, since this is the standard thing to do. A citation for the posterior update of a GP may be good too.

In turn, the posterior distribution on the integral F is

$$F \mid [f(\boldsymbol{s})] \sim \mathcal{N}(m_1, C_1), \tag{7}$$

$$m_1 = L(f) + \left[\int_{\mathbb{R}^d} C_t^0(t, \boldsymbol{s})\mathrm{d}z\right]^\top \left[C_t^0(\boldsymbol{s}, \boldsymbol{s})\right]^{-1} \left(f(\boldsymbol{s}) - m_0^t(\boldsymbol{s})\right), \tag{8}$$

$$C_1 = C_0 - \left[\int_{\mathbb{R}^d} C_t^0(t, \boldsymbol{s})\mathrm{d}z\right]^\top \left[C_t^0(\boldsymbol{s}, \boldsymbol{s})\right]^{-1} \left[\int_{\mathbb{R}^d} C_t^0(t, \boldsymbol{s})\mathrm{d}z\right], \tag{9}$$

where the integrals are row-wise over $\boldsymbol{s}$:

$$\int_{\mathbb{R}^d} C_t^0(t, \boldsymbol{s})\mathrm{d}z = \left(\int_{\mathbb{R}^d} C_t^0(t, s_1)\,\mathrm{d}z, \ldots, \int_{\mathbb{R}^d} C_t^0(t, s_n)\,\mathrm{d}z\right)^\top.$$

The posterior (7) will serve as the diagnostic for the Laplace approximation. Borrowing from the traditional notion of hypothesis testing, one may deem the Laplace approximation acceptable or valid if $L(f)$ falls within the range spanned by the $(0.025, 0.975)$ quantiles of (7) (the 95% "confidence interval" centered at the posterior mean). Conversely, if $L(f)$ is outside of this interval, the Laplace approximation would be deemed inappropriate, and one would proceed to use a more involved method to estimate F. Traditionally `[add old BQ citation]`, the goal of BQ is convergence to the true integral: choosing the covariance kernel and interrogation points such that (8) and (9) are close to F and 0, respectively. This is not our main goal in designing the diagnostic, which is intended to be decidedly non-asymptotic: rather, it should be able to effectively facilitate the aforementioned "hypothesis" test with as little computational cost as possible, whether or not that results in a good integral estimate.

4 Placement of interrogation points

Typically, the numer of points required to estimate a d-dimensional integral to within some error tolerance increases exponentially in d `[citation needed, especially for BQ]`. However, the goal of this diagnostic is to efficiently test the Gaussian shape assumption underpinning the Laplace approximation, with accurate integral estimation as an afterthought. Thus, it should require fewer interrogation points than a full BQ, allowing in principle for easier scailng to high dimensions.

First, assume without loss of generality that $\hat{t}$ is at the origin and $H = -I$. This ensures that the Gaussian approximation to f, m_0^t, is proportional to a standard Normal density. If this not the case, recall that H is negative definite, so there exists a matrix G such that $G^\top H G = -I$. For instance, $G = V\left[\sqrt{-D}\right]^{-1}$ from the eigendecomposition $H = VDV^\top$ serves this purpose. Then the aforementioned assumptions may be enforced by replacing f with the function $t \mapsto f\left(Gt + \hat{t}\right)$.

References

[1] Oksana A Chkrebtii, David A Campbell, Ben Calderhead, and Mark A Girolami. Bayesian Solution Uncertainty Quantification for Differential Equations. *Bayesian Analysis*, 11(4):1239–1267, 2016. doi: 10.1214/16-BA1036. URL `https://projecteuclid.org/download/pdfview{_}1/euclid.ba/1473276259`.

[2] Haoxuan Zhou. Bayesian Integration for Assessing the Quality of the Laplace Approximation. Master's thesis, Simon Fraser University, nov 2017. URL `http://summit.sfu.ca/item/17765`.