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

Shaun McDonald, Dave Campbell, Haoxuan Zhou

June 9, 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

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

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

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

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

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

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 (or perhaps more accurately, the shape assumptions underpinning it) 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 n required to estimate a d-dimensional integral to within some error tolerance increases exponentially in d `[citation needed, especially for BQ]`. This creates an unfortunate computational bottleneck in BQ, as the main cost is inverting the $n \times n$ matrix $C_0^t\left(\boldsymbol{s}, \boldsymbol{s}\right)$. 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. Presumably this goal can be achieved with 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 $t \mapsto f\left(Gt + \hat{t}\right)$.

As it pertains to the selection of interrogation points $\boldsymbol{s}$, this transformation serves two purposes. The first is to "rotate" the domain of f so that its direction of strongest curvature corresponds to one of the axes[6]. Heuristically, this means that the values of f along the axes will offer the most pertinent information about its shape. The second purpose is scaling and shifting, which allows interrogation points to be defined in a very intuitive way. This is best explained with an example: in two dimensions, when $H = -I$ and $\hat{t} = (0, 0)$, an interrogation point at $(m, 0)$ corresponds to a point that is m "standard deviations" (of the bivariate Normal distribution with density proportional to m_0^t) from the mode along the x-axis. With these ideas in mind, we propose to use a d-dimensional "cross-shaped" grid of interrogation points[7] consisting of the mode and additional points placed at regular spacings along each axis. Such grids will be characterized by the distances between consecutive points along the axes and the distance between the mode and the extremal points, both in terms of "standard deviations of m_0^t" as described above. For instance, one may wish to place interrogation points at half-integer multiples of the standard deviation, up to a maximum of three standard deviations, along each axis. This corresponds to points of the form $\pm\frac{m}{2}e_i$ where $m = 0, 1, \ldots, 6$ and e_i is a standard basis vector in $\mathbb{R}^d$, $i = 1, \ldots, d$. Although the use of these ideas requires a somewhat costly eigendecomposition of H, we believe that the alignment of shape information with the axes outweighs any such costs: the proposed cross-shaped grids grow linearly in size with d, bypassing much of the computational cost associated with more involved quadrature techniques. For instance, the grid given as an example above consists of only $n = 12d + 1$ points, and is expected to convey enough shape information to make the diagnostic work.

5 A finite-integral covariance kernel

The choice is a covariance kernel is important in determining the behaviour of a probabilistic quadrature method.Chkrebtii et al. [1], and subsequently Zhou [2], used a self-convolution of the popular squared exponential kernel `[sources/further info on sq exp kernel could be inserted if needed]`:

$$C_0^t(t, u) = \left(\frac{\sqrt{\pi}\lambda}{\alpha}\right)^d \exp\left[-\frac{\|t - u\|^2}{4\lambda^2}\right], \tag{10}$$

where, respectively, the *length-scale* and *precision* hyperparameters λ and α control the sample smoothness and spread of the GP.

A problem arises if one wishes to use this kernel without modification: its integral over $\mathbb{R}^d \times \mathbb{R}^d$ diverges, so the prior distribution assigned to F will

[6] This point can be made clear with some linear algebra and multivariate calculus. First note that the second directional derivative of $logf$ at the mode is always negative and is minimized along the direction of some eigenvector of H. Finally observe that the Hessian of f at the mode has the same eigenvectors as H, and the rotational part of the transformation maps them to standard basis vectors.

[7] If I recall correctly, somebody made this suggestion to Dave at a conference in early 2018.

have infinite variance. Some practitioners avoid this problem by integrating over finite regions rather than the whole of $\mathbb{R}^d$: Chkrebtii et al. [1] considered ODE's defined no compact intervals, and Zhou [2] took integrals over a region bounded by the extremal interrogation points. It is perhaps more common in BQ literature `[todo: add citations]` to integrate with respect to a probability measure Π on $\mathbb{R}^d$. In that case, the object of interest is $\int_{\mathbb{R}^d} f(t)\mathrm{d}\Pi(t)$, for which the prior mean and variance are defined by, respectively, the integral of m_0^t w.r.t Π and the integral of C_0^t w.r.t. the product measure $\Pi \times \Pi$. With this framework, all of the necessary integrals converge, in contrast to our setting where integrals are taken w.r.t. the Lebesgue measure.

To solve the problem of infinite variance, we take a different approach[8]: adding a "decay" factor to the covariance kernel so that it has finite integral over $\mathbb{R}^d \times \mathbb{R}^d$. The modified kernel used throught this manuscript is

$$C_0^t(t,u) = f\left(\hat{t}\right)^2 \left(\frac{\sqrt{\pi}\lambda}{\alpha}\right)^d \exp\left[-\frac{\|t-u\|^2}{4\lambda^2}\right] \exp\left[-\frac{\|t\|^2+\|u\|^2}{4\gamma^2}\right]. \tag{11}$$

The new hyperparameter γ controls the rate at which the GP prior variance, $C_0^t(t,t)$, decays as t moves away from the origin. Intuitively it is reasonable to impose such behaviour on the GP prior: any function f to which this diagnostic applies would certainly decay to 0 in the tails, so the prior assumption that uncertainty/variability in its values decreases with distance from the mode (assumed to be at the origin, as per Section 4) is appropriate[9]. With this modification, the covariance kernel (11) is proportional to a $2d$-dimensional Gaussian density. Figure 1 shows a visual comparison between this kernel and the one given by (10) in the one-dimensional case. Note that (10) is the limiting case of (11) as $\gamma \to \infty$. Using this modified kernel with $\gamma \in (0, \infty)$ ensures that the scalar Normal prior on F will have finite variance `[todo: make sure all the GP convergence stuff is actually valid for this]` given by

$$C_0 = f\left(\hat{t}\right)^2 \left[\frac{4\gamma^2\lambda^2\sqrt{\pi^3}}{\alpha\sqrt{2\gamma^2+\lambda^2}}\right]^d. \tag{12}$$

The factor of $f\left(\hat{t}\right)^2$ ensures scale invariance in the diagnostic. The posterior variance of F (6) depends only on the placement of interrogation points `[this is well-known in BQ, so could maybe provide source]`. In turn, interrogation point locations are indirectly based on the shape of $\log f$ as described in Section 4. Because the shape of $\log(Mf)$ is the same for any scaling constant $M > 0$, the posterior variance will be unchanged by such scaling even though the difference between the posterior and prior means is $M(m_1 - L(f))$. Therefore, without explictly incorporating scale information into the covariance, the rejection criteria for the diagnostic will be scale-dependent: an undesirable trait given that the shape of the true function and the *proportional* difference

[8] Should Richard be a last author? Modifying the covariance kernel to decay was his idea (not specifically *how* to modify it, though)

[9] This was also a point made by Richard originally.

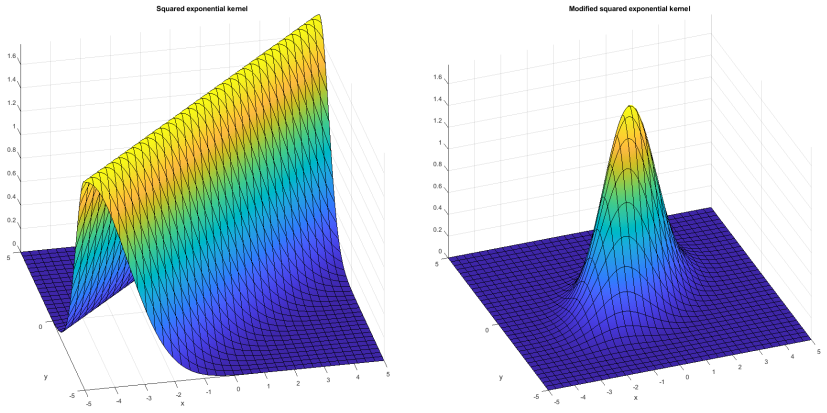

Figure 1: For $d = 1$, the usual squared exponential kernel with $\lambda = \alpha = 1$ (left) vs. the modified decaying version with $\lambda = \alpha = \gamma = 1$ (right).

between its integral and the Laplace approximation do not depend on scale[10]. Incorporating the factor of $f\left(\hat{t}\right)^2$ into the covariance kernel ensures that the Laplace approximation is rejected for f iff it is rejected for Mf for all $M > 0$.

6 Optimization and Experiments

Literature on BQ contains several approaches to optimizing both interrogation point placements and covariance hyperparameters `[sources and examples coming soon]`. However, these approaches are usually aimed towards a high-accuracy, low-variance estimate of the integral, which differs from the goal of the diagnostic. In particular, the diagnostic should *fail* to reject any function whose shape is sufficiently close to Gaussian for the Laplace approximation to be reasonable. Thus, our approach to optimization will be based on somewhat heuristic calibrations.

Let $T_{d,\nu}$ denote the density of the d-variate T distribution with ν degrees of freedom. Such a density has heavier tails than a d-dimensional Gaussian density, and so its integral is underestimated by the Laplace approximation. However, the Gaussian is the limiting case of the T density as $\nu \to \infty$. Therefore, for some large value of ν, the shape of $T_{d,\nu}$ is "sufficiently Gaussian" and the diagnostic should be calibrated *not* to reject its Laplace approximation, which will be close to the true integral value of 1. We will expand on this shortly, but first it is useful for exploratory purposes to see how the posterior mean (5) and variance (6) depend on the hyperparameters of the covariance kernel. For this preliminary

[10] Another thing briefly mentioned by Richard and only recently remembered/implemented by me.

visualization, we use true functions of $T_{1,1}$ and $T_{2,1}$ (for which the tails are heavy enough to warrant a definite rejection of the Laplace approximation), a grid of interrogation points placed at $0, \pm 1, \pm 2$ standard deviations along each axis, and $\alpha = 1$. The latter choice is made because the only effect of α is on the scale of the posterior variance, and so we are more interested in the effect of λ and γ on diagnostic behaviour.

Figure 2 shows variation in the posterior mean and variance over a range of λ and γ values. Jagged edges in the plots at high λ's are likely indicative of numerical instability due to oversmoothing, which pushes the matrix $C_0^t(\boldsymbol{s}, \boldsymbol{s})$ towards singularity. The posterior reduction in variance increases with λ, but evidently to a lesser extent than the prior variance, so that the posterior variance is ultimately higher for large λ. The effect of λ will be explored later in more detail, but the effect of γ is interesting enough to warrant discussion here. Although γ appears to influence the variance "correction" (the difference between prior and posterior variance) and posterior mean at low values (and causes some possible numerical instability for the latter), its effect levels off considerably beyond a certain threshold. Indeed, when γ is sufficiently large, its primary contribution is towards the scale of the prior variance (12). The behaviour is similar for the 2-dimensional case, as shown in Figure 3. The main difference is that the posterior variance increases more dramaticlly with both γ and λ, which is to be expected due to the prior variance being $\mathcal{O}\left(\lambda^{2d}\right)$ and $\mathcal{O}\left(\gamma^{2d}\right)$. There also appears to be a greater range of variability with respect ot λ in the posterior mean, and perhaps in the variance correction as well.

Let us now return to the issue of calibration. Recall that the limiting behaviour of $T_{d,\nu}$ with respect to ν makes it close to a Gaussian shape for high degrees of freedom. This ensures that the Laplace approximation,

$$L\left(T_{d,\nu}\right) = \left(\frac{2}{\nu + d}\right)^{\frac{d}{2}} \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}, \tag{13}$$

is increasing in ν and approaches 1 as $\nu \to \infty$. For a given dimension d, let ν_d be the smallest integer such that $L\left(T_{d,\nu_d}\right) \geq 0.95$

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`.

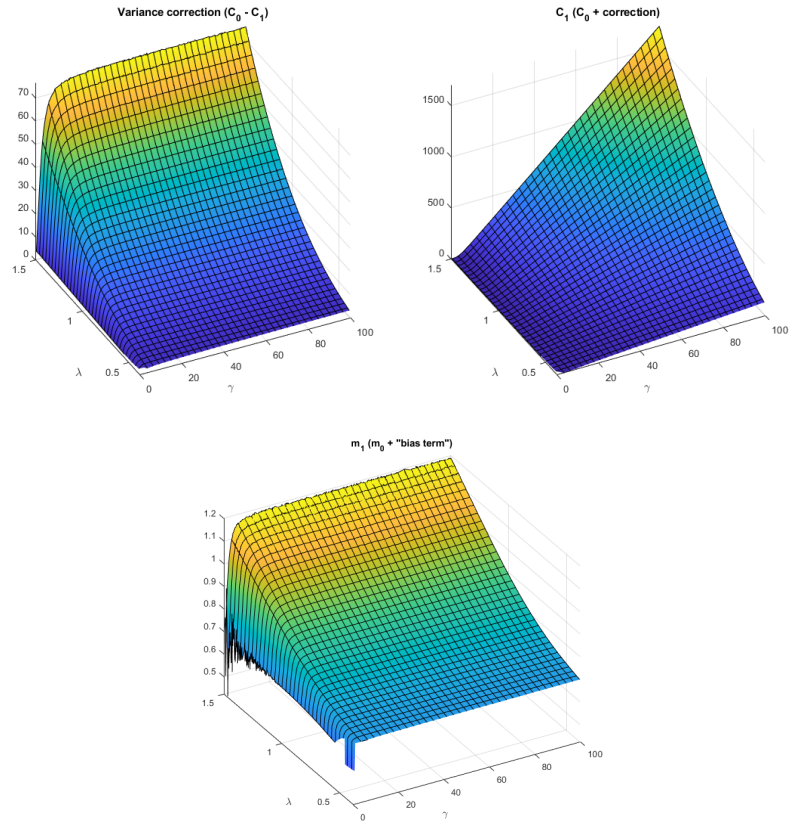

Figure 2: Behaviour of the diagnostic for a 1-dimensional T-distribution with 1 degree of freedom. Top left: variance correction term (difference between C_0 and C_1). Top right: posterior variance. Bottom: posterior mean.

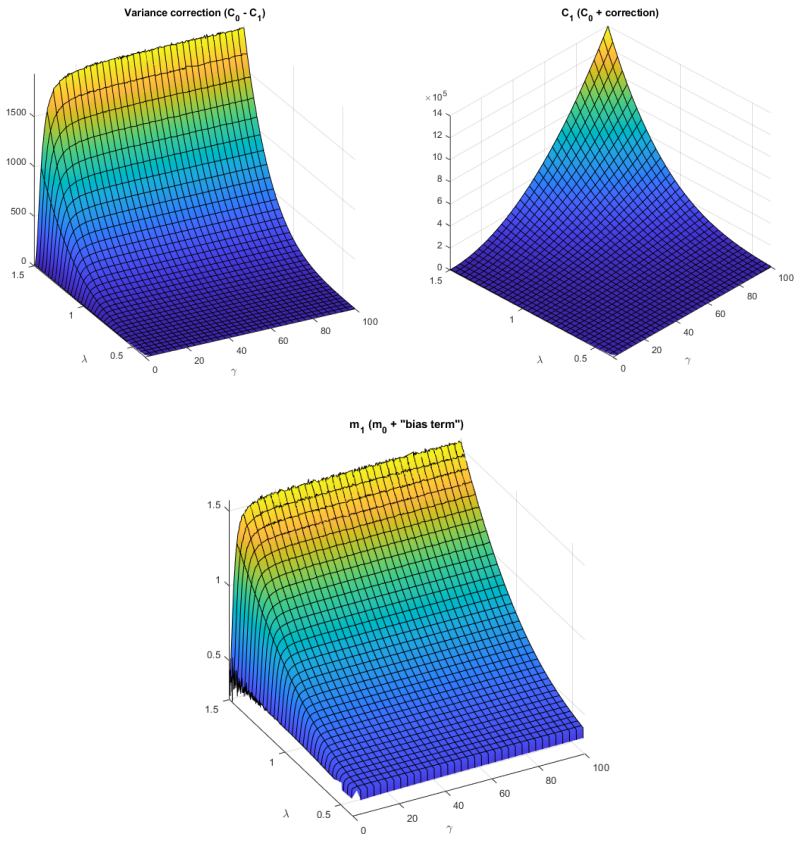

Figure 3: Behaviour of the diagnostic for a 2-dimensional T-distribution with 1 degree of freedom. Top left: variance correction term. Top right: posterior variance. Bottom: posterior mean.