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

Shaun McDonald, Dave Campbell, Haoxuan Zhou

July 22, 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 and its underlying shape assumptions, 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. Use cases including state space models
2. That we are building on the work of Zhou [2]
3. 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.
4. 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. Of course, the Laplace approximation will not hold for a non-Gaussian shape, so tl;dr the diagnostic should reject if (*but not only if*) the Laplace approximation is wrong.

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 (hereafter called the *Gaussian approximation to* f) produces the *Laplace approximation*[2]

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

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

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 approximation of f 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 $\mathbb{R}^d \times \mathbb{R}^d$ defined in Section 5.

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 $\boldsymbol{s} = (s_1, \ldots, s_n)^\top \in \mathbb{R}^{n\times d}$ 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

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]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

abuse of notation):

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

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

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

In turn, the posterior distribution on the integral F is

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

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

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

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

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

It is useful to think of the posterior means and variances as their prior counterparts modified by the addition or subtraction of some "correction term".

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 motivating 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 ("rejection"), 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 $C_0^t(\boldsymbol{s}, \boldsymbol{s})$. However,

[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.

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. In this section, we aim to design the interrogation grid $\boldsymbol{s}$ with this in mind.

Recalling that H is negative-definite, consider its eigendecomposition $H = VDV^\top$ and let $G = V\left[\sqrt{-D}\right]^{-1}$. An $n \times d$ matrix of "preliminary" interrogation points $\boldsymbol{s}^* = (s_1^*, \ldots, s_n^*)^\top$ will be defined as described below, and ultimately the points comprising $\boldsymbol{s}$ will be of the form $s_i = Gs_i^* + \hat{t}$, $i = 1, \ldots, n$.

This transformation serves two purposes. The first is a "rotation", as G maps standard basis vectors to eigenvectors of H. Thus, points along one of the axes are transformed to align with the direction of strongest curvature in f at the mode[6]. Heuristically, this means preliminary points on the axes offer the most pertinent information about the shape of f. The second purpose is scaling and shifting, which allows interrogation points to be defined in a very intuitive way with respect to the Gaussian approximation of f. Specifically, consider a preliminary point $s_i^* = me_j$, with $m \in \mathbb{R}$ and e_j equal to the j^{th} standard basis vector in $\mathbb{R}^d$. Then the resulting interrogation point s_i is m "standard deviations" - w.r.t. the multivariate Normal density proportional to m_0^t, in the direction of the j^{th} eigenvector of H and with scale defined by the corresponding eigenvalue - away from the mode.

With these ideas in mind, we propose to use a d-dimensional "cross-shaped" grid of preliminary points[7] consisting of the origin 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 origin and the extremal points, in terms of "standard deviations of m_0^t in each direction" as described above. For instance, one may wish to place interrogation points at the mode and half-integer multiples of the standard deviation, up to a maximum of three standard deviations, along each "principal" direction or axis defined by the (orthogonal) eigenvectors of H. With the scheme defined above, this corresponds to preliminary points of the form $\pm\frac{m}{2}e_j$ where $m = 0, 1, \ldots, 6$ and $j = 1, \ldots, d$. 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 should convey enough shape information to make the diagnostic work, depending on the contours of the true function (a caveat which will be expanded upon in Section 6.3).

[6] This point can be made clear with some linear algebra and multivariate calculus. First note that the second directional derivative of $\log f$ 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.

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

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 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

$$C_0^t(t,u) = \kappa\left(G^{-1}\left(t-\hat{t}\right), G^{-1}\left(u-\hat{t}\right)\right), \tag{11}$$

$$\kappa(t,u) = f\left(\hat{t}\right)^2 \left(\frac{\sqrt{\pi}\lambda}{\alpha}\right)^d \exp\left[-\frac{\|t-u\|^2}{4\lambda^2}\right] \exp\left[-\frac{\|t\|^2+\|u\|^2}{4\gamma^2}\right]. \tag{12}$$

The transformation applied to the arguments is the inverse of that which was used to transform the preliminary points into interrogation points in Section 4. The new hyperparameter γ controls the rate at which $\kappa(t,t)$ decays as t moves away from the origin. With the aforementioned transformation of the arguments, the prior GP variance $C_0^t(t,t)$ is therefore highest at the mode, and decays in each direction at a rate depending on the curvature of f at $\hat{t}$ in that direction. Intuitively it is reasonable to impose such behaviour on the prior: any function f to which this diagnostic applies would certainly be nonnegative and decay to 0 in the tails, so the prior assumption that uncertainty/variability in its values decreases with distance from the mode is appropriate[9]. With this modification, the covariance kernel (12) is proportional to a $2d$-dimensional Gaussian density. Figure 1 shows a visual comparison between this kernel and

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

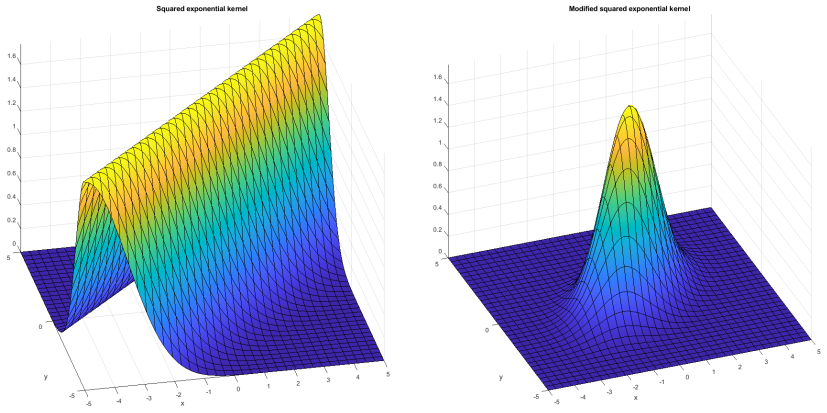

Figure 1: For $d = 1$, the usual squared exponential kernel with $\lambda = \alpha = 1$ (left) vs. the modified decaying version with $\lambda = \alpha = \gamma = 1$ (right).

the one given by (10) in the one-dimensional case. Note that (10) is the limiting case of (12) 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

$$C_0 = f\left(\hat{t}\right)^2 \det\left(G\right)^2 \left[\frac{4\gamma^2\lambda^2\sqrt{\pi^3}}{\alpha\sqrt{2\gamma^2+\lambda^2}}\right]^d. \tag{13}$$

The factors of $f\left(\hat{t}\right)$ (included explicitly in the kernel κ) and $\det(G) = \sqrt{\det\left(-H^{-1}\right)}$ (induced by integrating C_0^t over either of its arguments) induce an important kind of invariance in the diagnostic. Without them, for any f the posterior variance of F (9) would be independent of the function itself, depending only on the placement of interrogation points `[this is well-known in BQ, so could maybe provide source]`. To see why this is undesirable, consider a function f_{Trans} defined by scaling such an f and linearly transforming its domain: $f_{\text{Trans}} : t \mapsto af\left(At\right)$, for some $a > 0$ and $A \in \mathbb{R}^{n\times n}$ with $\det(A) \neq 0$. Because the Hessian of $\log f_{\text{Trans}}$ is $A^\top HA$, it follows that $L\left(f_{\text{Trans}}\right) = a\left|\det(A)\right| L\left(f\right)$. Thus, *without* the function-dependent factors included in (11–13), the rejection behaviour for the diagnostic - determined by the posterior mean and variance (8–9) - would be different when applied to f_{Trans} than it would be for f. This arguably runs counter to the purpose of the diagnostic, which is to determine how close a function's shape is to that of its Gaussian approximation. Intuitively, it is easy to reason that the idea of "closeness in shape" should be independent of linear transformations to the function and its domain (and indeed, the *proportional* distance between the true integral and the Laplace approximation

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

is invariant to such transformations)[10]. The function-dependent factors in our covariance structure ensure that the posterior mean and standard deviation of the diagnostic applied[11]to f_{Trans} are both proportional to $a\,|\det(A)|$, thus encoding this notion of "invariance": for a fixed set of hyperparameters $(\lambda, \gamma, \alpha)$, the diagnostic results in a rejection when applied to f iff it rejects when applied to f_{Trans} for *any* positive a and invertible A.

6 Calibration and Simulated 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 of a given function, 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. Because the diagnostic uses less information about the true function than a normal BQ (as per Section 4), is also desirable to have a "one-size-fits-all" set of hyperparameters so that it can be easily applied without the need for function-specific recalibration. Our approach to optimization will be based on somewhat heuristic calibrations with these considerations in mind.

Let $T_{\nu,d}$ 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 $T_{\nu,d}$ 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 (8) and variance (9) depend on the hyperparameters of the covariance kernel. For this preliminary visualization, we use true functions of $T_{1,1}$ and $T_{1,2}$ (for which the tails are heavy enough to warrant a definite rejection of the Laplace approximation), a grid of preliminary points placed at $0, \pm 1, \pm 2$ units 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

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

[11] To be clear, when applying the diagnostic to the transformed function, the $f(\hat{t})$ and G terms in the covariance structure (which correspond to an arbitrary f) would be replaced by the equivalent terms corresponding to f_{Trans}. More accurate notation for C_0^t would reflect its dependence on the function f, but this would be cumbersome.

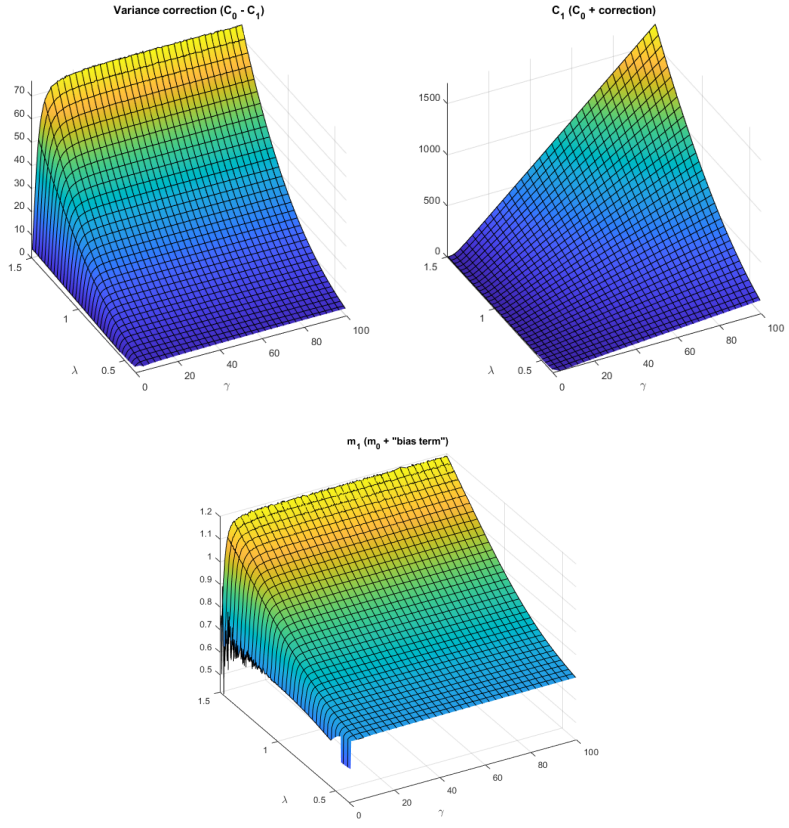

Figure 2: Behaviour of the diagnostic for a 1-dimensional T-distribution with 1 degree of freedom. Top left: variance correction term (difference between C_0 and C_1). Top right: posterior variance. Bottom: posterior mean.

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 (13). 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 to λ in the posterior mean, and perhaps in the variance correction as well.

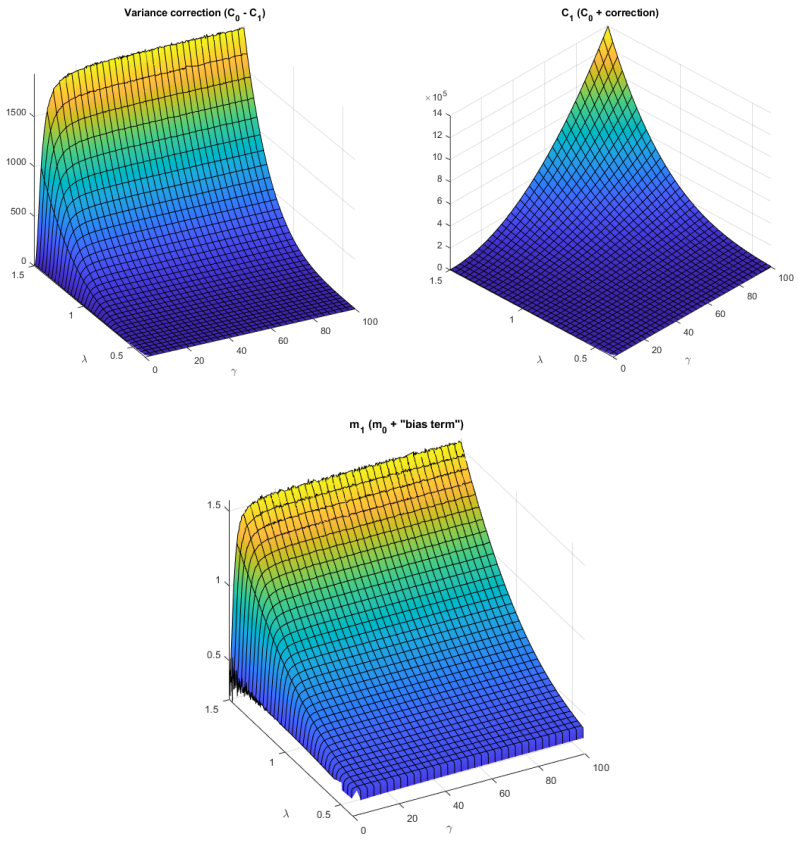

Figure 3: Behaviour of the diagnostic for a 2-dimensional T-distribution with 1 degree of freedom. Top left: variance correction term. Top right: posterior variance. Bottom: posterior mean.

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,

$$L\left(T_{\nu,d}\right) = \left(\frac{2}{\nu+d}\right)^{\frac{d}{2}} \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}, \tag{14}$$

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 $L\left(T_{\nu_d,d}\right) \geq 0.95$. T densities with more than ν_d degrees of freedom are close enough in shape to Gaussians that their Laplace approximations are within 5% of the true integral value; conversely, those with lower degrees of freedom have heavier tails and Laplace approximations that underestimate the true integral by over 5%. Thus, $T_{\nu_d,d}$ will serve as the "edge case" function used to calibrate the diagnostic. Specifically, for a given dimension d and grid of interrogation points $\boldsymbol{s}$,

1. α should be set so that $L\left(T_{\nu_d,d}\right)$ is on the boundary of the non-rejection region (the interval spanned by the (0.025, 0.975) quantiles of the posterior $\mathcal{N}\left(m_1, C_1\right)$ distribution); and

2. λ and γ should be set to minimize discrepancy between the true function $T_{\nu_d,d}$ and the posterior mean function m_1^t (5).

The first requirement ensures that $T_{\nu_d,d}$ acts as a sort of "threshold", as intended. With α set in this way, $T_{\nu_d,d}$ is "just Gaussian enough" for the Laplace approximation to be deemed reasonable. The second requirement ensures that the posterior GP provides a good fit to this threshold function. Although the diagnostic is not intended to produce high-accuracy approximations in general, at the very least it can be calibrated to do so for this special case. This is particularly important as the choice of hyperparameters can compromise inference in ways that require more in-depth scrutiny to understand, as will be discussed in the following subsection.

6.1 Calibrating in two dimensions

Recall that the diagnostic is intended to test whether or not the use of the Laplace approximation is justified by the shape of the true function. To this end, the actual values produced by the integral posterior do not necessarily tell the whole story. Note that the posterior integral mean m_1 (8) and GP mean m_1^t (5) are related by $m_1 = \int_{\mathbb{R}^d} m_1^t(t)\mathrm{d}t$. Thus, the hyperparameters λ and γ can affect inference in ways not immediately obvious from Figures 2 and 3, by influencing the shape of the GP posterior mean.

With these considerations in mind, let us attempt to calibrate the diagnostic in two dimensions using the bivariate T distribution with $\nu_2 = 38$ degrees of freedom. The effects of λ and γ on m_1^t and the resulting inference for the integral will be assessed, with α set to put the Laplace approximation on the edge of the rejection region as described above. The interrogation grid will be expanded

slightly from the one used for Figure 3, adding preliminary points at ± 3 along each axis for a total of 13 interrogation points.

As in Figures 2-3, we found that γ had a diminishing effect on the shape of m_1^t as it increased beyond a certain threshold (not shown). Thus, we initially fix the moderately high (and somewhat arbitrary) value $\gamma = 0.25$ and vary λ, which controls the smoothness of the GP as it does with the conventional squared exponential kernel. For a low value of λ, Figure 4 shows the difference $m_1^t - T_{38,2}$ and the posterior inference for the integral. By (5), it holds that $m_1^t(\boldsymbol{s}) = f(\boldsymbol{s})$ for any f and any combination of hyperparameter values. However, at any other point t, the extent to which $m_1^t(t)$ updates from $m_0^t(t)$ is determined by the "weights" $C_t^0(t, \boldsymbol{s})^\top \left[C_t^0(\boldsymbol{s}, \boldsymbol{s})\right]^{-1}$. When λ is small, there is almost no prior dependence between GP values at distinct points, so these weights are close to zero for $t \notin \boldsymbol{s}$. This explains the spikes in Figure 4: the posterior mean is forced to equal $T_{38,2}$ at the interrogation points, but everywhere else it is virtually unchanged from the prior mean m_0^t (which has lighter tails than the true funtion, hence the negativity of the difference at non-interrogation points). Thus, the posterior mean of the integral, m_1, barely updates from the prior guess $L(T_{38,2})$. In this case, there is not much point in calibrating α as described above.

Figure 5 shows the effect of a moderate increase in λ. This induces more dependence between values at distinct points, and therefore more smoothness, in the GP. Thus, m_1^t updates to a greater extent along the lines defined by the interrogation points, and the value of m_1 moves closer to the true F. However, there are still large valleys in the difference $m_1^t - T_{38,2}$, centered around the main diagonals of the plane and within the boundaries of the interrogation grid. Recalling that the Gaussian approximation is lower than the true function in these regions, it is clear that the change from m_0^t to m_1^t is not as large as it should be there. With a moderate λ-value the influence of the interrogation points is relatively weak in these regions, and one may say that the GP is failing to *interpolate*. To allow the interrogations to exert sufficient influence in these regions, it is necessary to further increase λ as in Figure 6. However, this comes at a price: increasing λ to the extent necessary for good interpolation in the "diagonal regions" causes undesirable extrapolation effects due to oversmoothing. Indeed, in all four directions just beyond the extremal interrogation points, m_1^t dips well below $T_{38,2}$. As a result, m_1 is *lower* than the Laplace approximation, let alone the true integral value. Oversmoothing causes the weights $C_t^0(t, \boldsymbol{s})^\top \left[C_t^0(\boldsymbol{s}, \boldsymbol{s})\right]^{-1}$ to have unpredictable effects at t beyond the boundaries of the interrogation grid, depending on the spread and density of $\boldsymbol{s}$ as well as the shape of f. In some cases, the "valleys" seen in Figure 6 may be replaced by large "hills", causing m_1 to significantly overestimate the value of F (not shown). In any case, it seems in this case that good interpolation *within* the interrogation boundaries comes at the expense of poor extrapolation *beyond* them.

Of course, such extrapolation issues would not matter if we simply integrated only over the rectangle enclosed by the extremal interrogation points as in Zhou

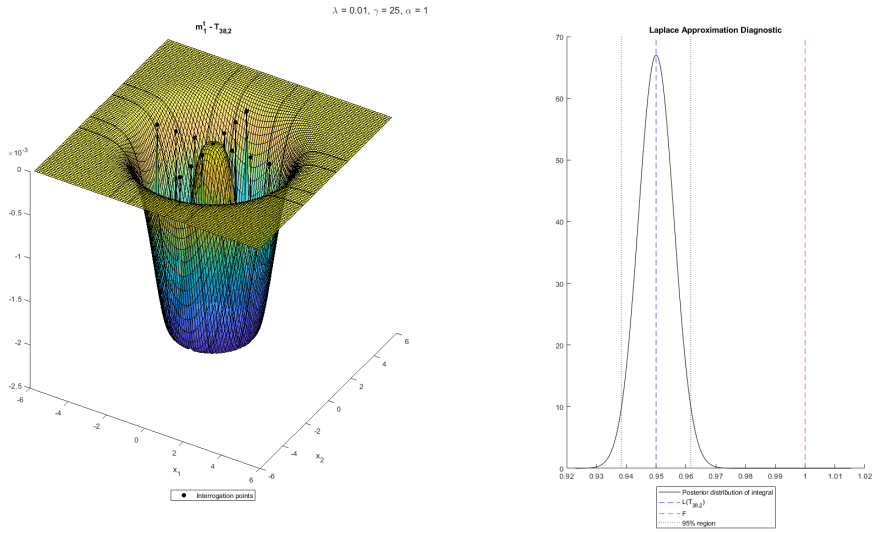

Figure 4: Posterior results in 2 dimensions with a high γ value and low λ value.

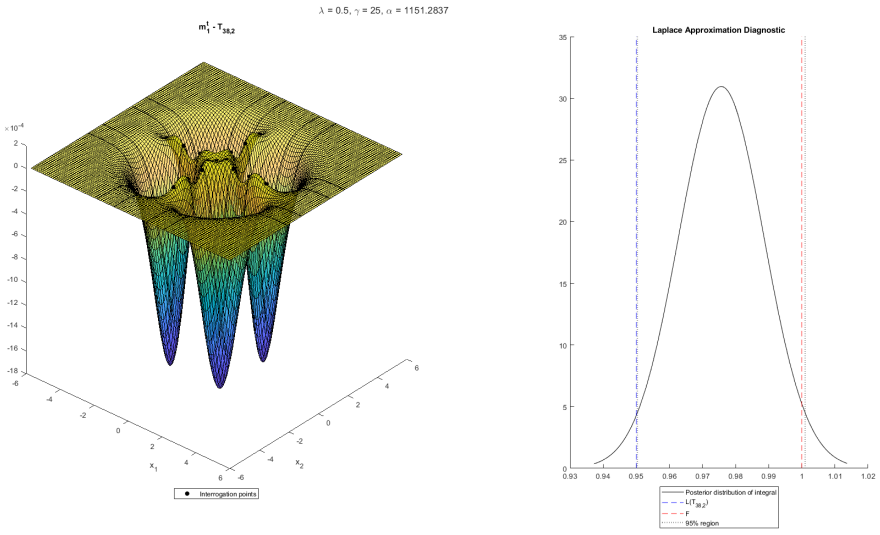

Figure 5: Posterior results in 2 dimensions with a high γ value and low λ value.

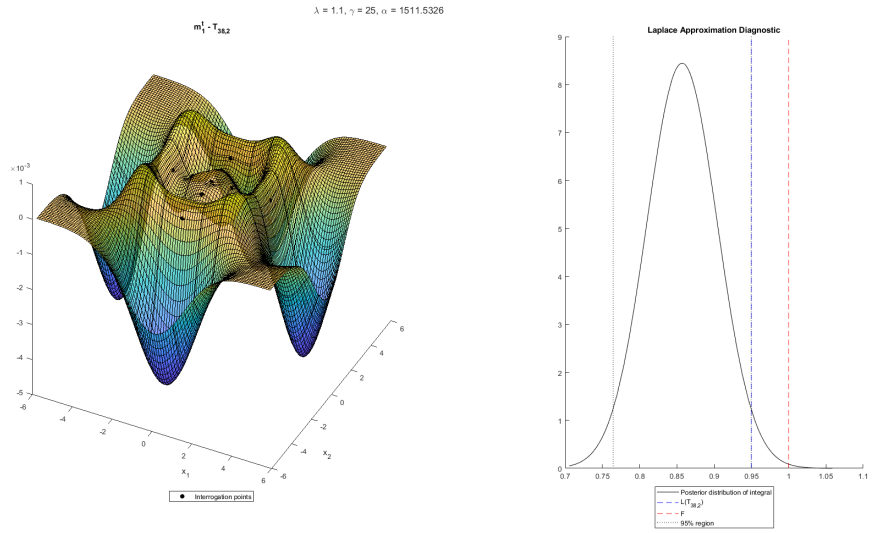

Figure 6: Posterior results in 2 dimensions with a high γ value and high λ value.

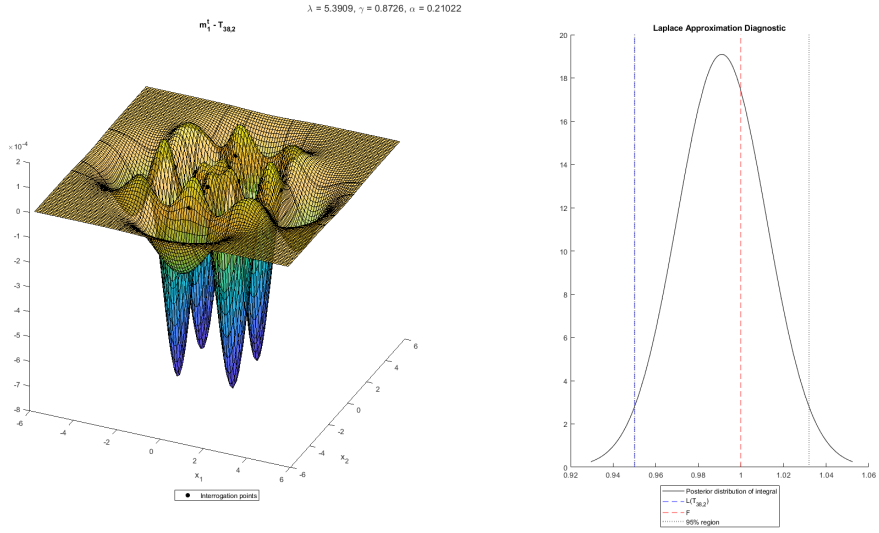

Figure 7: Posterior results in 2 dimensions with a low γ value and high λ value.

[2]. One may also wonder if the effects of oversmoothing could be mitigated by using a kernel which enforces less smoothness, such as the Matérn or uniform kernel `[sources]`. However, we have not yet exhausted all possible behaviours for the kernel (12), as the decay parameter γ can still be altered. Recalling that γ controls the covariance of points far away from the origin, it is reasonable to hypothesize that a lower γ-value would reduce extreme behaviour in the weights at points beyond the interrogation boundaries. In this case, good interpolation could then be achieved by increasing λ *without* sacrificing extrapolation. Indeed, attempting to minimize the L^2-error $\int_{\mathbb{R}^2} \left(m_1^t(t) - T_{38,2}(t)\right)^2 \mathrm{d}t$ with respect to λ and γ (see Section 6.2) leads to a suggestion of $(\lambda, \gamma) \approx (5.3909, 0.8726)$. Figure 7 shows the results with these parameters. Some of the undesirable beahviour associated with both interpolation and extrapolation is present, but both types of behaviour are reduced in magnitude compared to Figures 5 and 6. As a result, the posterior inference for the integral is more accurate than it was with any other hyperparameter combination considered thus far.

6.1.1 A banana-shaped function

`[source and context]` considered a function with "banana-" or "boomerang-shaped" contours, defined by "twisting" one coordinate of a Gaussian density. Letting $\phi(\cdot; \Sigma)$ denote a bivariate Normal density with covariance matrix Σ, the version of the function used here is

$$\beta(t) := \phi\left(t_1, t_2 - \frac{1}{2}\left(t_1^2 - 3\right); \mathrm{diag}(3, 1)\right). \tag{15}$$

From Figure 8, it is clear that β has a decidedly non-Gaussian shape. However, it turns out that the Laplace approximation is indeed true for this function: $L(\beta) = \int_{\mathbb{R}^2} \beta(t)\mathrm{d}t = 1$. As discussed in Section 2, this may be viewed as "coincidence", as $\log \beta$ is clearly not globally similar to its second-order Taylor approximation (it can be shown that its Taylor expansion contains third- and fourth-order terms). The banana-shaped function therefore represents an interesting test case for the diagnostic: although the Laplace approximation is technically valid here, it should still be rejected due to its shape. In practical terms, the diagnostic should signify that such a function is sufficiently non-Gaussian to warrant further attention, at which point a more involved integration scheme could be used to verify that the Laplace approximation was valid after all.

6.2 On quantitative optimization of hyperparameters

It was mentioned above that hyperparameters for the two-dimensional diagnostic were chosen to approximately minimize the L^2-error between the callibration function and the posterior mean. An analytic expression for this error may not exist - in fact, it is not worth the effort to even attempt deriving one, as for all but the smallest interrogation grids the analytic form of $\left[C_0^t(\boldsymbol{s}, \boldsymbol{s})\right]^{-1}$ is prohibitively complicated[12]. Instead, the integral of $\left(m_1^t - T_{38,2}\right)^2$ over $\mathbb{R}^2$ was approximated using a simple Riemann sum over the grid of points $\{-10, -9.99, -9.98, \ldots, 9.99, 10\}^2$.

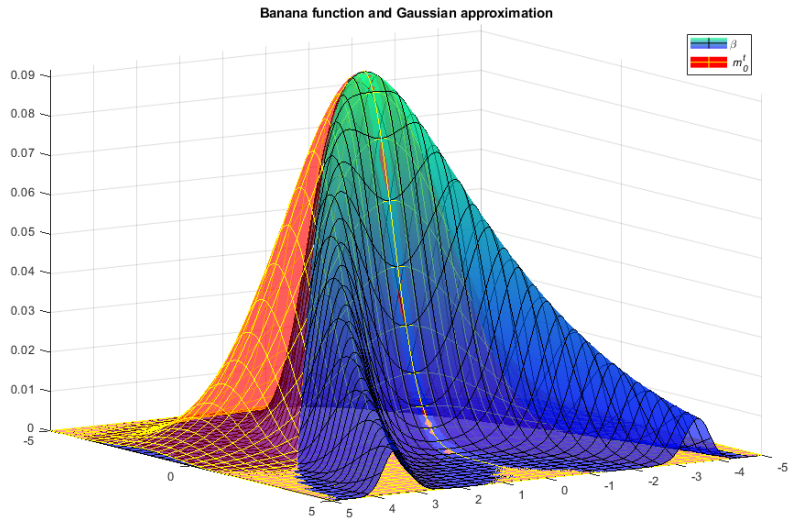

Figure 8: A two-dimensional "banana-shaped" function alongside its Gaussian approximation.

Minimization was carried out with the `fminunc` function in MATLAB `[source]`, using the BFGS algorithm `[source]` with pre-supplied gradient functions. The results of this optimization should be regarded with some skepticism - partially because of the rather simplistic approach to numerical integration, but also because the ill-conditioning of $\left[C_0^t\left(\boldsymbol{s},\boldsymbol{s}\right)\right]^{-1}$ for high λ-values appears to result in a fair amount of numerical instability. In particular, we found that finite-difference approximations to the gradients near the determined "optimal" value were wildly inaccurate. Although we supplied functions to the optimizer to (presumably) compute the gradients to a higher degree of accuracy, it is possible that similar issues affected the computation of second-order derivatives used in the algorithm.

One could apply this approach to optimization to higher dimensions, produce a table of recommended (λ,γ) pairs for each d, and perhaps attempt to infer an approximate relationship between optimal hyperparameter values and dimensionality. We have not pursued this here, as the computational cost of such optimization increases exponentially in d. Instead, we use a heuristic visual approach in higher dimensions: selecting a (λ,γ) pair that appears upon visual inspection to minimize the difference $m_1^t - T_{\nu_d,d}$ as uniformly as possible, balancing problems with interpolation and extrapolation. Due to the symmetry of the multivariate T distribution, this can be done in $d > 2$ dimensions by

[12] As evidenced by the failure of MATLAB's Symbolic Toolkit `[source]` to produce it after several hours running on a laptop with 16 GB of RAM, 3GB of dedicated GPU memory, and four Intel i5-9300H 2.40GHz CPU cores.

looking at a 2-dimensional "slice" of the difference, with $d-2$ of its arguments set to zero (as the mode is at the origin).

Another possible approach to hyperparameter optimization is to maximize the likelihood of the interrogation values $T_{\nu_d,d}(\boldsymbol{s})$ with respect to the GP prior. Using the log-likelihood to optimize is much less computationally difficult than using the L^2-error, and is a common method in more "conventional" BQ literature [sources]. Once again using the BFGS algorithm with pre-supplied gradient functions, we found the (approximately) likelihood-optimal hyperparameters in two dimensions were $(\lambda, \gamma, \alpha) = (0.7042, 1.5268, 580.8387)$. Note that unlike the L^2-error, the likelihood *does* depend on the prior precision α, which therefore cannot be set to ensure non-rejection for the threshold function. Indeed, the high likelihood-optimal α-value in two dimensions results in a narrow posterior for the integral of $T_{38,2}$, so that the Laplace approximation is rejected. As one may expect based on the (λ, γ)-values, the posterior mean m_1^t (not shown) has similar characteristics to those shown in Figure 5, but the interpolation biases are smaller in magnitude. Ultimately, the hyperparameters used in Figure 7 still appear to be a better choice for estimation of $T_{38,2}$ and its integral. Likelihood maximization provides the best fit for a finite set of function values, and is likely more suitable for BQ when the interrogation points thoroughly cover the domain of integration. Given that our interrogation grid is deliberately sparse, it is not surprising that this approach does not provide the best results for our purposes.

6.3 The curse of dimensionality

Recall that we use a T-distribution with ν_d degrees of freedom to calibrate the d-dimensional diagnostic using the Laplace approximation formula (14). Although (14) increases towards 1 as $\nu \to \infty$ for fixed d, it is decreasing in d for fixed ν. Therefore, ν_d is an *increasing* function of d, as higher degrees of freedom are needed to ensure $L(\nu_d, d) \geq 0.95$ when d is large. Put another way, in higher dimensions, a T density must be closer in shape to a Gaussian for the Laplace approximation to be within 5% of the true integral.

This is a special case of a more general phenomenon: small differences in function shapes have more opportunity to influence their integrals in high dimensions. In heuristic terms, for large d differences between a function f and its Gaussian approximation m_0^t are compounded over a higher number of dimensions, thus increasing the discrepancy between their integrals. Therefore, by necessity the diagnostic must be more shape-sensitive at high dimensions, as the Laplace approximation and its associated assumptions are reasonable for a diminished variety of function shapes. These high-dimensional considerations can result in some unusual diagnostic behaviour, as explained below.

Consider the diagnostic in 30 dimensions, calibrated with a T density with $\nu_{30} = 4660$ degrees of freedom. Rather than attempting another optimization based on L^2-error (which would involve a 30-dimensional numerical integral in this case), we simply used herusitics and visual inspection as described in Section 6.2. Maintaing the γ value of 0.8726 used in two dimensions, λ was

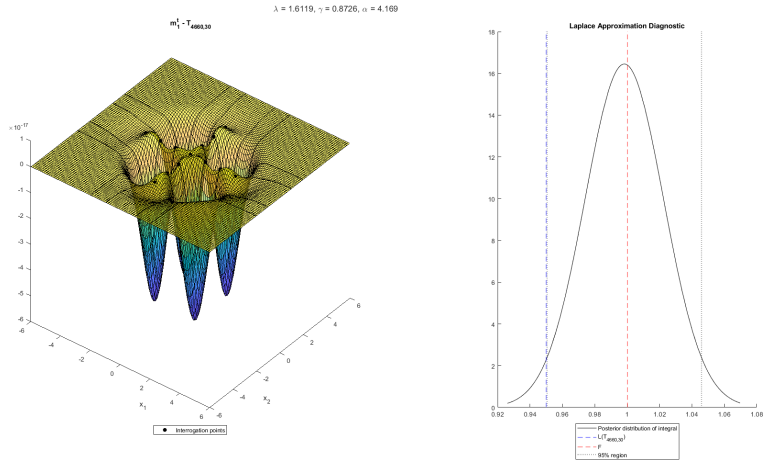

Figure 9: Posterior results for the empirically calibrated diagnostic in 30 dimensions. The left plot is a two-dimensional slice of $m_1^t - T_{4660,30}$, for which all 28 other coordinates were taken to be 0.

selected such m_1^t appeared as close as possible to $T_{4660,30}$, ultimately resulting in $(\lambda, \gamma, \alpha) = (1.6119, 0.8726, 4.169)$ (Figure 9). As in Figures 5 and 7, we see valleys in the diagonal regions within the interrogation boundaries. As before, the prior mean m_0^t is lower than the true function in these regions, and the posterior mean "correction" $m_1^t - m_0^t$ (which, as explained in Section 6.1, decays with distance from the interrogation points to an extent controlled by λ and γ) is too small to fully correct underestimation. However, these differences are small enough that the inference is reasonably good (and as before, the only way to remove them entirely is to incur unacceptably large extrapolation errors).

Note that the multivariate T density used thus far corresponds to the joint distribution whose components are uncorrelated but *not* independent. The independent case results in a different density function: a product of d univariate T distributions,

$$f_{\nu,d}(t) = \prod_{i=1}^{d} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}} \left(1 + \frac{t_i^2}{\nu}\right)^{-\frac{\nu+1}{2}}. \tag{16}$$

It turns out that fewer degrees of freedom are required to bring the Laplace approximation of this function within 5% of the true integral (which is equal to 1): in particular, $L\left(f_{439,30}\right) = 0.95$. Figure 10 shows the surprising results of the diagnostic applied to $f_{439,30}$, using the empirically-determined hyperparameters stated above. The valleys typically seen in the "diagonal regions" are now hills, causing m_1 to significantly overestimate F. This is an excellent example of the aforementioned "compounding" of small differences: the height of the hills is on the order of 10^{-16}, but even such a miniscule difference can lead to a large bias when integrated over a 30-dimensional space. This is seemingly

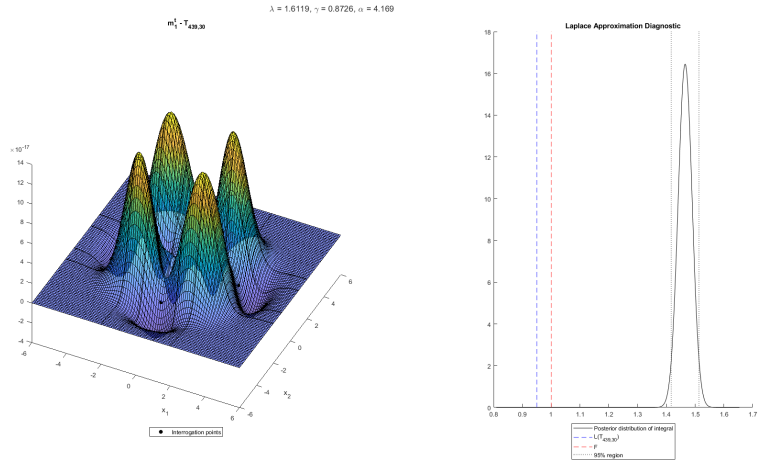

Figure 10: Posterior results for the empirically calibrated diagnostic in 30 dimensions. The left plot is a two-dimensional slice of $m_1^t - f_{439,30}$, for which all 28 other coordinates were taken to be 0.

unfortunate behaviour, given that the Laplace approximation is within 5% of the true integral and $f_{439,30}$ appears extremely similar to a Gaussian.

There are sensible reasons for this behaviour. Unlike the true multivariate T density, the product of independent densities (16) does *not* have spherical contours: on a given d-dimensional sphere, its value is highest at the points on the axes. In mathematical terms, given $r > 0$ and $t \in \mathbb{R}^d$ with $\|t\| = r$, $T_{\nu,d}(t) = T_{\nu,d}(r, 0, \ldots, 0)$, but $f_{\nu,d}(t) \leq f_{\nu,d}(r, 0, \ldots, 0)$ with equality iff t is on an axis `[todo: verify the three-line ‘‘proof’’ I just did in my head]`. Consequently, $f_{439,30}$ is closer to its Gaussian approximation in the "diagonal regions" than it is on the axes. Because all interrogation points are on the axes[13], the correction term for the posterior mean (5) - a weighted sum of the differences $f_{439,30}(\boldsymbol{s}) - m_0^t(\boldsymbol{s})$ - is *too large* in the diagonal regions, whereas it was slightly too small there for the diagnostic applied to $T_{4660,30}$.

For a contrast, consider a different function

$$g_{\nu,d}(t) = \left[\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}}\right]^d \left(1 + \frac{\|t\|^2}{\nu}\right)^{-\frac{\nu+1}{2}}. \tag{17}$$

This function is equal to $f_{\nu,d}$ on the axes, and the Hessian of $\log g_{\nu,d}$ at the origin is equal to that of $\log f_{\nu,d}$. Thus, $L(g_{\nu,d}) = L(f_{\nu,d})$, and the diagnostic applied to $g_{\nu,d}$ will produce the exact same posterior (for both the function and its integral) as the diagnostic applied to $f_{\nu,d}$. However, $g_{\nu,d}$ has spherical contours, so the true value of its integral is greater than that of $f_{\nu,d}$. In particular, the

[13]Recall from Section 4 that the interrogation points are placed on axes corresponding to eigenvectors of the log function's Hessian. However, this Hessian is diagonal for all functions considered here, so these axes correspond to the usual choice of the standard basis.

integral of $g_{439,30}$ over $\mathbb{R}^{30}$ is equal to 1.661. The diagnostic posterior shown on the right of Figure 10 is reasonably close to *this* value, which is far enough away from the Laplace approximation of 0.95 to warrant a rejection.

These examples illustrate a potential downside of our approach to interrogation point selection, particularly in high dimensions. By design, the diagnostic uses a limited amount of "information" about the true function: its local curvature at the mode, and some of its values along the "principal" axes defined by the eigenvectors of H (hereafter in this paragraph, the word "axes" will refer to these). If this information is consistent with the function's behaviour in regions away from the axes - as is the case with the spherical-contoured functions considered thus far - then our empirical evidence suggests that the diagnostic can be calibrated to have reasonably good accuracy and behaviour[14]. If this is not the case, then the diagostic can produce misleading estimates as it does not take any off-axis information into account. In cases such as the two-dimensional "banana function" this is not necessarily a problem: the primary goal of detecting its highly non-Gaussian shape was achieved, regardless of the inaccuracy of the posterior integral inference. However, in higher dimensions, large biases can arise from even seemingly subtle differences between the function's behaviour on and off the axes, simply because (in somewhat informal terms) the off-axis regions comprise a greater "proportion of the volume" over which we integrate. Such was the case for $f_{439,30}$: to the naked eye its contours appear nearly spherical (not shown), and it looks close enough to a Gaussian that one may expect a non-rejection from the diagnostic. Comparing its behaviour to that of $g_{439,30}$ highlights how influential small differences can be in high dimensions when limited information is used. Results like these which contradict visual intuition may be an inevitable consequence of the need for greater shape sensitivity in higher dimensions.

Of course, these issues could be eliminated with a larger interrogation grid including points off the axes. Recall the construction of the preliminary point set in Section 4 as line-shaped lattices along each axis of $\mathbb{R}^d$. This construction can be easily generalized to provide larger grids using (in a slight abuse of terminology) Cartesian products: square lattices in the orthogonal planes spanned by each pair of axes, cubic lattices in the spaces spanned by each triple of axes, and so on. However, this would be antithetical to the goal of using a small number of informative points for computationally efficient shape assessment. In particular, we wish to avoid the trap of interrogation grids with size exponentially increasing in d, as this offers no advantage over full-on Monte Carlo integration. A possible middle-ground approach would be to increase the size of the grid according to some set of dimensionality thresholds, such that its growth remained sub-exponential. For instance, one could use the line lattices for $d = 1$ to, say, 10; square lattices for $d = 10$ to 30, and so on. Fur-

[14]There is somewhat more nuance than this. The examples considered thus far show that *some* bias remains in the posterior GP even when the true function has elliptical contours. Indeed, the contours of the posterior mean function are not elliptical in general. The best possible accuracy would be achieved for a function which was equal to the posterior GP mean of its own diagnostic, but this is an unlikely scenario in practice.

ther experimentation would be needed to determine the thresholds providing the best tradeoff between computation and accuracy, but this is not pursued here. Other strategies for interrogation point selection, particularly sequential methods, are popular in BQ literature `[TODO: add examples with sources and maybe move this discussion to Section 4]`. Like most BQ concepts, these are typically applied towards the goal of obtaining high-accuracy and low-variance integral estimates, and it remains to be seen how they would suit our purposes.

Another consideration in high dimensions is the issue of hyperparameter sensitivity. Small changes in γ and λ create small changes in the shape of m_1^t, but as discussed, these can create very large fluctuations in the value of the posterior mean integral m_1 when compounded over a large number of dimensions. The posterior variance C_1 (9) can also change quite dramatically in the high-dimensional setting with small perturbations to the hyperparameter values, depending on the values themselves. From (12) and (13), it is clear that C_1 is $\mathcal{O}\left(\alpha^{-d}\right)$, but its relationship to λ and γ with respect to d is more complicated `[could math it out and see if I can make a big-O statement beyond an appeal to numerical evidence]`. In any case, the higher the dimensionality, the more certainty one must have in the "optimal" hyperparameter values.

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`.