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

Shaun McDonald, Dave Campbell, Haoxuan Zhou

February 8, 2022

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

Many statistical models assume the existence of "unseen" variables which influence the actual observed data, but are distinct from the model parameters that are of interest for inference. One such model is the *state-space model (SSM)*, which has become a staple of ecological modelling [e.g. 2, and references therein] and will serve as a motivating example throughout this paper. Briefly, the SSM assumes that (possibly vector-valued) data y_t are observed at discrete time steps $t = 1, \ldots, T$. At a given time t, the distribution of y_t depends on an unobserved or "hidden" state $x_t \in \mathbb{R}^q$ (typically the dimensionality of x_t is the same for all t, but it may differ from the dimensionality of the y_t's). In turn, the distribution of x_t depends on the previous hidden state, x_{t-1}. The reader may recognize this as the structure of a *hidden Markov model (HMM)*, although that term is

typically used when the domain of the hidden states is discrete [e.g 11]. Here, they are assumed to be continuous and possibly multivariate.

In mathematical terms, the SSM is characterized by the joint likelihood[1]

$$p_{x,y}(\boldsymbol{x}, \boldsymbol{y} \mid \theta) = p(x_1 \mid \theta) \left[\prod_{t=2}^{T} p(x_t \mid x_{t-1}, \theta)\right] \left[\prod_{t=1}^{T} p(y_t \mid x_t, \theta)\right], \qquad (1)$$

where $\boldsymbol{x} = (x_1, \ldots, x_T)$ is a vector of dimension $d = qT$ concatenating the hidden states, $\boldsymbol{y}$ is defined analogously, and θ is a vector of model parameters. These parameters are conceptually different from the hidden states even though both are unobserved: θ represents the *fixed effects* of the model, whereas $\boldsymbol{x}$ represents *random effects*[2].

There are a variety of methods for both frequentist and Bayesian inference with SSM's [e.g. 14, 43, and references therein]. In the frequentist framework, one typically wishes to estimate θ by maximizing the marginal likelihood of the data,

$$p_y(\boldsymbol{y} \mid \theta) = \int_{\mathbb{R}^d} p_{x,y}(\boldsymbol{x}, \boldsymbol{y} \mid \theta) \, \mathrm{d}\boldsymbol{x}. \qquad (2)$$

Unfortunately, the necessary integral over the hidden states is of dimension $d = qT$, and as such the marginal likelihood cannot realistically be computed - much less optimized - in most cases. Instead, frequentist inference methods for SSM's typically rely on approximate methods to obtain a suitable estimate of θ. One common example is the use of the *Laplace approximation (LA)*.The Laplace approximation of the marginal likelihood is reasonably easy to compute and optimize as a function of θ, but it is based on certain assumptions about the shape of the joint likelihood as a function of $\boldsymbol{x}$: namely, that it is well approximated by a d-dimensional Gaussian density. If this assumption is not satisfied, the LA may not be suitable, and different methods for SSM inference may need to be invoked.

The example of the SSM provides motivation for the broader goal of this manuscript, which is to develop a diagnostic tool to check the assumptions underpinning the LA In particular, our interest is in assessing whether or not a given function is "close enough" to the Gaussian shape to justify using the Laplace approximation of its integral. In making this assessment, we strive for a "middle ground" of computational effort: the diagnostic will naturally be more complex than the LA itself, but much less expensive than a full-fledged

[1] There are several possible formulations for the distribution of the first hidden state (the $p(x_1 \mid \theta)$ term in (1)). Some literature assumes it to depend on an "initial state" x_0 which is given its own prior in turn [e.g. 33] or simply point estimated [e.g. 43].The latter is essentially equivalent to specifying an "unconditional" distribution for x_1, another common approach [e.g. 11, 28]. Some authors omit the $p(x_1 \mid \theta)$ term entirely, thereby implictly assigning x_1 an "improper uniform prior" [e.g. 34, which is the formulation used in Section 7.1]. The general model form given in (1) will suffice for the purposes of this manuscript.

[2] Of course, in a Bayesian setting, both model components are given priors and essentially treated in the same way. In that case, the difference between them is more of a "philosphical" matter.

numerical estimate of the integral. Expanding on the work of Zhou [47], here we describe such a diagnostic tool based on the machinery of *probabilistic numerics*, a burgeoning field which exploits probability theory to tackle numerical problems. The tool is an application of the probabilistic numerical technique of *Bayesian quadrature (BQ)*, which allows for both estimation and inference of unknown integrals. Unlike "conventional" BQ, however, the actual integral value is of secondary importance, as the tool is primarily intended to capture as much information as possible about the *shape of the integrand*. In keeping with the aforementioned objective of "medium effort", the tool is also decidedly non-asymptotic: it is meant to deliver as much information as possible with a modest amount of computation, without consideration of any type of limiting behaviour. The goal is a fast, informal method that can be readily deployed to determine if additional modelling efforts are needed beyond the LA.

The remainder of the manuscript proceeds as follows. Section 2 defines the LA and establishes the notation used throughout the paper, while Section 3 provides more detail about the workings of probabilistic numerics and BQ in particular. Sections 4–5 provide technical details about the design of our diagnostic tool, and Sections 6–7 show applications and challenges in the high-dimensional setting.

2 Framework and notation

Consider a positive function $f : \mathbb{R}^d \to \mathbb{R}_{>0}$ and its integral $F = \int_{\mathbb{R}^d} f(x)\mathrm{d}x$. More rigorous treatments of the Laplace approximation are available in, for instance, De Bruijn [13] and Barndorff-Nielsen et al. [3], but for this exposition it suffices to assume that all second-order partial derivatives of f exist and are continuous, and that f attains a maximum at some point $\hat{x} \in \mathbb{R}^d$. To reflect the common use case where f is a density or likelihood, $\hat{x}$ is called a *mode*. Let H be the Hessian of $\log f$ at $\hat{x}$ and suppose that it is negative definite. Taking a second-order Taylor expansion of $\log f$ about $\hat{x}$ gives the approximation

$$\log f(x) \approx \log f\left(\hat{x}\right) + \frac{1}{2}\left(x - \hat{x}\right)^\top H \left(x - \hat{x}\right), \tag{3}$$

since all first-order partial derivatives of $\log f$ are equal to zero at the mode. Exponentiating (3) gives an approximation for f in the form of (up to normalizing constants) a Gaussian density centered at $\hat{x}$ with covariance matrix $-H^{-1}$. In turn, integrating this exponentiated function (hereafter called the *Gaussian approximation to* f) produces the *Laplace approximation* to F:

$$\begin{aligned} F \approx L\left(f\right) &:= f\left(\hat{x}\right) \int_{\mathbb{R}^d} \exp\left[\frac{1}{2}\left(x - \hat{x}\right)^\top H \left(x - \hat{x}\right)\right] \mathrm{d}x \\ &= f\left(\hat{x}\right) \sqrt{\left(2\pi\right)^d \det\left(-H^{-1}\right)}. \end{aligned} \tag{4}$$

The LA has a long history of use in statistics [e.g. 30, 44]. It is exact (or "true") if the integrand 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 derivation of the LA via (3) is based on an assumption of approximately Gaussian shape (insofar as it assumes that the second-order Taylor series is a reasonable approximation to $\log f$), and as noted in Section 1 this assumption is our main interest.

Before proceeding to further details about the construction of the diagnostic tool, it is worthwhile to connect these concepts to the SSM example described in Section 1. For given observations $\boldsymbol{y}$ and parameter values θ, the joint likelihood $p_{xy}(\cdot, \boldsymbol{y} \mid \theta)$ takes the role of the integrand, viewed as a function of the hidden states $\boldsymbol{x} \in \mathbb{R}^d$. In turn, one can see from (2) that the marginal likelihood $p_y(\boldsymbol{y} \mid \theta)$ takes the role of the integral over $\mathbb{R}^d$ to be approximated by $L(p_{xy})$. Note, however, that this approximation is itself a function of $\boldsymbol{y}$ and θ, as both

$$\hat{x} = \underset{\boldsymbol{x}}{\operatorname{argmax}}\, p_{xy}(\boldsymbol{y}, \boldsymbol{x} \mid \theta) \quad \text{and} \quad H = \left.\frac{\partial^2 \log p_{xy}}{\partial \boldsymbol{x}^2}\right|_{(\boldsymbol{y}, \hat{x}, \theta)}$$

may depend on these quantities. Indeed, one of the most common ways to "fit an SSM" in the frequentist sense is to maximize $L(p_{xy})$ with respect to θ (given observed $\boldsymbol{y}$), typically using standard numerical algorithms. Fitting the model in this way becomes a matter of *nested* optimization, since in each iteration $\hat{x} = \hat{x}(\theta, \boldsymbol{y})$ must be (numerically) calculated for the current θ-value [see 29, for instance].

Implicit in the use of such methods for SSM's is the assumption that the LA is reasonably accurate given $\boldsymbol{y}$ and for each θ-value calculated during the optimization steps. If the shape of p_{xy} with respect to $\boldsymbol{x}$ is not "sufficiently Gaussian" at a given iteration, then the ultimate estimate of θ may not be close to the actual MLE for the marginal likelihood. Therefore, it would be desirable to check the validity of the LA at each step, using the diagnostic tool detailed below.

3 Probabilistic numerics and Bayesian quadrature

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 [19]. For instance, Chkrebtii et al. [10] 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 evaluations of the true derivative to conduct posterior inference on the entire solution.

The approach briefly described above - using Gaussian process priors and finitely many function evaluations 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 [see 7, 12, and references therein]. As the name

suggests, BQ is a probabilistic analogue to standard numerical integration that uses a combination of prior belief and gathered informaton about a function. The remainder of this section, in which the diagnostic for the LA is developed, will also serve as an explanation of the mathematical machinery underpinning BQ.

Literature on BQ commonly assumes that the integral of interest is with respect to a probability (i.e. finite) measure G on the domain [e.g 7], and a standard choice for $\mathbb{R}^d$ is a d-dimensional Gaussian measure [35, 25]. Accordingly, we use an "importance weighting trick" [26, 40, 36] to re-express the integral of interest. Recalling the notation of Section 2, the integral of f over $\mathbb{R}^d$ is

$$F = \int_{\mathbb{R}^d} f(x)\mathrm{d}x = \int_{\mathbb{R}^d} r(x)g(x)\mathrm{d}x = \int_{\mathbb{R}^d} r(x)\mathrm{d}G(x), \tag{5}$$

where $r := f/g$ and g is the density of the aforementioned Gaussian measure G, the parameters of which will be discussed later. It is this "re-weighted" function r that is modelled with a Gaussian process prior [26]. The mean function of the GP prior, m_0^x, is taken to be the (similarly re-weighted) Gaussian approximation of f underpinning (3) and (4):

$$m_0^x(x) := \frac{f\left(\hat{x}\right)\exp\left[\frac{1}{2}\left(x-\hat{x}\right)^\top H\left(x-\hat{x}\right)\right]}{g(x)}, x \in \mathbb{R}^d. \tag{6}$$

The covariance operator for the GP is a (positive-definite) kernel C_0^x on $\mathbb{R}^d \times \mathbb{R}^d$, defined in Section 4.2. Because integration is a linear projection, such a prior on g induces a univariate normal prior on F with mean $m_0 := \int_{\mathbb{R}^d} m_0^x(x)\mathrm{d}G(x) = L(f)$ and variance $C_0 := \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} C_0^x(x,z)\mathrm{d}G(x)\mathrm{d}G(z)$ [e.g. 40, 19][3].

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 (transposed) vectors in $\mathbb{R}^d$ (we will sometimes call it a "*grid*" of n "points" in $\mathbb{R}^d$). Then, for instance, the notation $r\left(\boldsymbol{s}\right)$ will refer to the column vector $\left(r\left(s_1\right), \ldots, r\left(s_n\right)\right)^\top \in \mathbb{R}^n$, and $C_0^x\left(\boldsymbol{s},\boldsymbol{s}\right)$ will denote the $n \times n$ matrix with $(i,j)^{\text{th}}$ entry $C_0^x\left(s_i, s_j\right)$. Using standard GP identities [e.g. 41], one may use true function values at the *interrogation points* $\boldsymbol{s}$ to obtain a posterior distribution for g (with another slight abuse of notation):

$$r \mid r(\boldsymbol{s}) \sim \mathcal{GP}\left(m_1^x, C_1^x\right), \tag{7}$$

$$m_1^x(x) = m_0^x(x) + C_0^x(x,\boldsymbol{s})^\top\left[C_0^x(\boldsymbol{s},\boldsymbol{s})\right]^{-1}\left(r(\boldsymbol{s}) - m_0^x(\boldsymbol{s})\right), \tag{8}$$

$$C_1^x(x,z) = C_0^x(x,z) - C_0^x(x,\boldsymbol{s})^\top\left[C_0^x(\boldsymbol{s},\boldsymbol{s})\right]^{-1} C_0^x(z,\boldsymbol{s}). \tag{9}$$

[3] Unsure whether this needs a citation, or can just be taken as ''common knowledge''. Some BQ papers do, some don't. Alternative citations could include O'Hagan 1991 and Briol et al 2019, or Michael Osborne's thesis? Note that all of these only state that the *posterior* for the integral is normal. If I need a different citation to talk about its *prior*, I'm not sure what to use.

In turn, the posterior distribution on the integral F is [e.g. 7, or, indeed, virtually any BQ paper]

$$F \mid r(\boldsymbol{s}) \sim \mathcal{N}\left(m_1, C_1\right), \tag{10}$$

$$m_1 = L(f) + \left[\int_{\mathbb{R}^d} C_0^x(z, \boldsymbol{s})\mathrm{d}G(z)\right]^\top \left[C_0^x(\boldsymbol{s}, \boldsymbol{s})\right]^{-1}\left(r(\boldsymbol{s}) - m_0^x(\boldsymbol{s})\right), \tag{11}$$

$$C_1 = C_0 - \left[\int_{\mathbb{R}^d} C_0^x(x, \boldsymbol{s})\mathrm{d}G(x)\right]^\top \left[C_0^x(\boldsymbol{s}, \boldsymbol{s})\right]^{-1}\left[\int_{\mathbb{R}^d} C_0^x(x, \boldsymbol{s})\mathrm{d}G(x)\right]; \tag{12}$$

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

$$\int_{\mathbb{R}^d} C_0^x(x, \boldsymbol{s})\mathrm{d}G(x) = \left(\int_{\mathbb{R}^d} C_0^x\left(x, s_1\right)\mathrm{d}G(x), \ldots, \int_{\mathbb{R}^d} C_0^x\left(x, s_n\right)\mathrm{d}G(x)\right)^\top.$$

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 (10) 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 (10): 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 could proceed to use a more involved method to estimate F.

4 Design decisions

In broad terms, there are three major categories of "design" choices one must make in order to conduct BQ, each of which will be explored in the following subsections. First, we must decide where to place interrogation points $\boldsymbol{s}$; second, a covariance kernel C_0^x must be chosen for the GP prior; and finally, we must specify the measure G against which to integrate. The latter two involve setting some *hyperparameters* that will govern the behaviour of the Gaussian process; this will be deferred to Section 5.

Recall that the diagnostic is intended to quickly - and somewhat heuristically - determine whether a given function f is "sufficiently Gaussian" to justify the LA for its integral. In particular, it should expend only as much computational effort as is necessary to reliably make this determination, with actual estimation of the integral F being a *secondary* goal. In this respect, its objectives are different from those of "traditional" BQ, in which interrogation points may be chosen to minimize the posterior variance of the integral [35, 32, 20] or the entropy of the integrand [17]; and hyperparameters may be chosen by some goodness-of-fit criterion [7, 41] or approximately marginalized [37], with both approaches depending on the "observations" $r\left(\boldsymbol{s}\right)$. The computational

costs arising from such methods would be antithetical to the "quick" nature of the diagnostic. Instead, it should be "one-size-fits-all" so that it can be quickly applied to any suitable function. Although "ad hoc" design choices are made in some BQ papers [e.g 25], the fact remains that the usual goal is to obtain an accurate integral estimate with low uncertainty. Beyond the issue of computation, there is a more fundamental difference between our goals and those of "traditional" BQ, or, indeed, the usual principles of inference more broadly. Typically, one may wish to maximize the *power* of their inference, ensuring that any true deviation from some null hypothesis will be found with sufficient data. In the present context, this would mean embracing the standard BQ goal of high accuracy and low uncertainty, so that even the smallest deviation from the LA could be rejected if there are enough well-placed interrogation points. However, such a diagnostic would not be very useful in practice. Harkening back to the SSM example from Section 1, in all but the simplest models it will be known in advance that the joint likelihood is not *exactly* Gaussian, and the LA not exactly met. The pertinent question is whether the joint likelihood is Gaussian *enough*, and a diagnostic that answered this question in the negative for every nonlinear model would be trivial and useless. Thus, the usual aim of high "power" is actually *not* desirable here: the diagnostic should be calibrated such that it *fails to reject* any function which is "close enough" to Gaussian, in a sense explained below. In this way, the design choices detailed in the following sections target an unconventional notion of "*good-enough-ness of fit*".

4.1 Placement of interrogation points

The selection of interrogation points (or "nodes", as they are commonly known in the literature) is the defining feature of any quadrature method. Much has been written about the asymptotic error rates (as number of points $n \to \infty$) of various quadrature methods, and the ways in which they depend on the dimensionality of the domain d and the smoothness of the integrand [e.g. 24, 7]. However, none of these considerations are relevant to the development of a quick, one-size-fits-all tool intended to determine if a function is "Gaussian enough" for the LA to be reasonable. Thus, the grid of interrogation points must provide as much pertinent information as possible about the *shape* of f, and (particularly in high dimensions, as explained below) how this shape influences the validity of the LA. Importantly, it must do this with as small a grid as possible in order to be "medium-effort"; in particular, the grid size must grow at a reasonable rate with respect to d. One hopes that the goals of the diagnostic can be accomplished with less computation than it takes to conduct a more accurate BQ.

To begin with, let $\boldsymbol{s}^* = (s_1^*, \ldots, s_n^*)^\top \in \mathbb{R}^{n \times d}$ be a grid of "preliminary" interrogation points. Ostensibly the preliminary grid should not depend on any properties of the function f, but considerations such as dimensionality can certainly inform its construction. We will assume that the grid is a union of *fully symmetric sets*, as considered by Karvonen and Särkkä [25]. Briefly, this means that if we take an arbitrary vector s_i^* from the grid, any vector obtained

via permutation or sign changes of its coordinates is also in the grid [ibid.]. We also assume that the grid contains multiples of the standard basis vectors of $\mathbb{R}^d$ (i.e. points are placed "along the axes") and that its centroid is the origin (the origin may be included in the grid, but this is not strictly necessary). No further restrictions will be placed on the preliminary grid, but some type of sparsity is desirable for the computational reasons mentioned above. The sparse grid methods desribed by Karvonen and Särkkä [25], or modifications thereof, are particularly useful to this end.

Now, recalling that H is negative-definite, consider the eigendecomposition $-H^{-1} = VDV^\top$ (where V is orthogonal and D is diagonal) and let $T := V\sqrt{D}$. The vectors comprising the actual interrogation grid $\boldsymbol{s}$ used in the diagnostic will be affine transformations of the preliminary grid vectors: $s_i = Ts_i^* + \hat{x}$, $i = 1, \ldots, n$. This transformation serves three purposes. The first is a translation so that the centroid of the grid is $\hat{x}$, the mode of f. Since f will be a density or likelihood in most applications, it makes sense for the grid to be oriented around the region of highest density. In contrast, a grid centered at the origin may be "off-center" for some integrands, capturing only limtied tail behaviour and certainly not enough "shape information". The second purpose for the transformation is a rotation, as T maps standard basis vectors to eigenvectors of H (which are the same as those of $-H^{-1}$). Thus, by placing some of the preliminary points along the "standard axes" of $\mathbb{R}^d$, we ensure that the corresponding interrogation points are aligned along the directions in which the "curvature" of f at the mode is most extreme[4]. Because H completely characterizes the shape of f under the "null hypothesis" that it (approximately) satisfies the assumptions of the LA, heuristically it makes sense to say that, *a priori*, one would expect such interrogation points to contain the most pertinent "shape information". Finally, the transformation "stretches" its inputs in the direction of each eigenvector V_i by a factor of $\sqrt{D_{ii}}$ (D_{ii} being the eigenvalue associated with V_i). Thus, if H is such that the Gaussian approximation to f (and, presumably, f itself) has different scales in different directions, the grid will capture this appropriately. In summary, this transformation turns a preliminary grid of the type stipulated above into an interrogation grid that is adapted to the contours of the Gaussian approximation to f. In this respect, it can be assumed - *a priori* or "under the null hypothesis" of Gaussian shape - that the grid so obtained is, in some informal sense, "optimal" for obtaining the necessary information about f.

There is another, perhaps more intuitive interpretation of interrogration grids generated in this way. Let X be a multivariate normal random variable with density proportional to the Gaussian approximation to f, i.e. $X \sim \mathcal{N}\left(\hat{x}, -H^{-1}\right)$. Then the i^{th} component of the vector VX is the i^{th} *principal component*, or PC, of X, and has marginal variance equal to D_{ii} [22]. Thus,

[4]This point can be formalized and 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 maximized (resp. minimized) in the direction of the first (resp. last) eigenvector of H. Finally observe that this statement must also be true for f itself since is always positive and its gradient is zero at $\hat{x}$.

the affine transformation of the preliminary grid is centered at the mean of X, aligned with its "principal axes", and scaled according to the scales of its PC's. For example, recall that for $i = 1, \ldots, d$, the preliminary grid contains points of the form $\pm m e_i$, where $m > 0$ and e_i is the i^{th} standard basis vector of $\mathbb{R}^d$. The corresponding interrogation points, $\pm m\sqrt{D_{ii}}V_i + \hat{x}$, are "$m$ standard deviations (of the i^{th} PC of X) away from the mode (in the direction of that PC)".

4.2 Form of covariance kernel

The covariance structure of the diagnostic will be based on the *squared exponential kernel*:

$$\kappa(x, z) = \alpha^{-d} \exp\left[-\frac{\|x - z\|^2}{2\lambda^2}\right], \tag{13}$$

a common choice in BQ [e.g 35, 25, 7]. The hyperparameter α controls the *precision* of the GP, serving as a scaling factor for its variance and for that of its integral. It is more common in literature to parameterize the kernel in terms of scale as opposed to precision, replacing α^{-d} in (13) with α^2 [e.g. 35, 17], but the practical difference between these choices is purely notational. The parameterization in (13) is the same as that used by Chkrebtii et al. [10], and the fact that α is raised to the power of $-d$ in (13) reflects their notion that the d-dimensional kernel can be viewed as a pointwise product of d univariate kernels. The hyperparameter λ is the *length-scale*, which controls the size of fluctuations in GP values between distinct points [41]. In an informal sense[5], λ therefore controls the "smoothness" of the GP.

The actual covariance function used in the diagnostic is a modification of (13) based on the function of interest f. It is

$$C_0^x(x, z) = f(\hat{x})^2 \det\left(-H^{-1}\right) \kappa\left(T^{-1}x, T^{-1}z\right), \tag{14}$$

where the matrix T was defined in Section 4.1. Because $\|T^{-1}x - T^{-1}z\|^2 = (x-z)^\top(-H)(x-z)$, the prior covariance of the GP at distinct points depends on the distance between these points in a linear transformation of Euclidean space, with the transformation depending on the "curvature" of $\log f$ at $\hat{x}$. Equivalently, the prior GP covariance function (14) is a (scaled) *Mahalanobis kernel* [1].

4.3 Choice of measure

In Section 3, we used an importance re-weighting trick to express F as an integral w.r.t. a Gaussian measure G. O'Hagan [35] and Kennedy [26] considered BQ for $r = f/g$ with a constant GP prior mean and noted that results would be most

[5] In a *formal* sense, a GP with squared exponential covariance kernel is infinitely differentiable, in the mean square sense, regardless of the value of λ [41]. "Smoothness" as informally used above simply means an absence of "wiggles" at small scales in functions sampled from the GP.

accurate if the density g closely approximated the shape of f, i.e. if r was roughly constant. The latter noted an analogy with importance sampling (IS), in which F is also modelled as the integral of r w.r.t. G and the shape of g should match that of the integrand [e.g. 46]. Although our GP prior mean (6) is not constant, we still found in preliminary experiments that g had to be a fairly good "fit" to f in order for the diagnostic to behave reasonably. Within the convenient class of Gaussian measures, remarks by O'Hagan and Kennedy suggest that g proportional to the Gaussian approximation to f, i.e. $G = \mathcal{N}\left(\hat{x}, -H^{-1}\right)$, would be a reasonable "starting point". The measure ultimately used for the diagnostic is a slight modification of this:

$$G = \mathcal{N}\left(\hat{x}, -\gamma^2 H^{-1}\right), \tag{15}$$

where the new hyperparameter $\gamma > 0$ controls the spread of G and will be discussed in Section 5.

4.4 Invariance of diagnostic behaviour

At first glance, it may seem that these function-specific design choices are antithetical to the intended "one-size-fits-all" nature of the diagnostic. On the contrary, our design ensures a few kinds of advantageous "invariance". Recall that the interrogation points are obtained from the function-agnostic preliminary grid as $s_i = Ts_i^* + \hat{x}$, $i = 1, \ldots, n$. Plugging any two interrogation points s_i, s_j into (14) therefore gives $C_0^x\left(s_i, s_j\right) \propto \kappa\left(s_i^*, s_j^*\right)$. Note also that analogous results can be shown to hold for the integral terms in (11 – 12)[6] and for the prior mean interrogations $m_0^x\left(\boldsymbol{s}\right)$. Therefore, in principle the interrogations should provide the same quality and quantity of "information" for *any* f. Now, recall that the diagnostic rejects the LA for f iff it is not contained in the central 95% interval of the integral posterior, i.e. iff $L(f) \notin \left(m_1 - 1.96\sqrt{C_1}, m_1 + 1.96\sqrt{C_1}\right)$. Note that $\sqrt{C_1}$ is equal to $L(f) \propto f\left(\hat{x}\right)\sqrt{\det\left(-H^{-1}\right)}$ times a factor depending only on $\boldsymbol{s}^*$ and the hyperparameters $(\lambda, \alpha, \gamma)$ (by (12) and (14)); similarly, m_1 is equal to $L(f)$ times a factor depending only on $\boldsymbol{s}^*$, the hyperparameters, and the "normalized" function values $f\left(\boldsymbol{s}\right)/f\left(\hat{x}\right)$ (by (6), (11), and the definition of r). Thus, the necessary and sufficient condition for rejection does not depend on the actual values of $\hat{x}$, $f\left(\hat{x}\right)$ and $\det\left(-H^{-1}\right)$: *it is invariant to any scaling of the function or affine transformation of its domain.* More formally, for a fixed set of hyperparameters, the diagnostic rejects the LA when applied to f iff it rejects the LA when applied to any function of the form $f_{\text{Trans}} : x \mapsto af\left(Ax + b\right)$ with $a > 0$, $A \in \mathbb{R}^{d \times d}$ with $\det(A) \neq 0$, and $b \in \mathbb{R}^d$. The only way in which f affects the result of the diagnostic is through the *relative* differences between its values at the interrogation points and those of its Gaussian approximation. Because the diagnostic seeks only to determine whether f is "sufficiently Gaussian in shape", this is precisely the appropriate behaviour for it to have.

[6] To see this, note that the density g has a multiplicative factor of $\sqrt{\det\left(-H\right)} = \left|\det\left(T^{-1}\right)\right|$, and integrate (14) w.r.t. G by substitution. This is another reason why the choice of measure (15) makes sense.

Note the "standardized" design developed in Sections 4.1–4.3 is not without precedent in the BQ literature. For instance, Särkkä et al. [42] adopted the idea of *stochastic decoupling* from sigma-point methodology: to integrate a function r against some Gaussian measure $\mathcal{N}(\mu, P)$, they placed a GP prior with the standard squared exponential covariance kernel (13) on the function $r_{\text{Trans}} : x \mapsto r\left(\mu + \sqrt{P}x\right)$ and used a standardized set of "unit" interrogation points. Such an approach is essentially equivalent (possibly up to variance scaling factors) to our design; indeed, the authors made note of its invariance to affine transformations. However, their main interest was in deriving BQ-based methods for filtering and smoothing in nonlinear SSM's, in which μ and P are computed for each necessary integral according to their algorithms [42].

5 Hyperparameter calibration

It remains to select values for $(\lambda, \alpha, \gamma)$. As discussed above, the design of the interrogation grid and covariance kernel serve to "standardize" the input and output scales of the GP, so it is not necessary to consider these factors when setting the hyperparameters. Indeed, for a given dimension d and preliminary grid $\boldsymbol{s}^*$, the same hyperparameter values should be used for *any* f to ensure the aforementioned diagnostic invariance. Recall from the beginning of Section 4 that the intent is to test "good-enough-ness of fit": the diagnostic should reject the LA for functions with a substantially non-Gaussian shape, but should *not* be so "powerful" that it rejects functions which are close enough to Gaussian. With this in mind, we propose to set the hyperparameters in a somewhat heuristic way based on a predetermined *calibration* or *test function* τ. Such a function should have a shape fairly close to Gaussian in order to serve as the "edge case" for the diagnostic. Specifically, given a preliminary grid $\boldsymbol{s}^*$ and test function τ, the hyperparameters for the d-dimensional diagnostic should be set such that the following conditions are met when the diagnostic is applied to τ.

(1) The LA $L(\tau)$ should be on the boundary of the rejection region (i.e. equal to one of the endpoints of the 95% central interval for the integral posterior), and

(2a) the discrepancy between τ and the "un-weighted" posterior GP mean, $m_1^x \cdot g$, should be as small as possible throughout the domain, or at the very least

(2b) the posterior integral mean m_1 is as close as possible to the true integral of τ.

Either version of the second condition should ensure that the diagnostic is reasonably accurate when applied to τ. Of course, in general accurate estimation is still an ancillary goal, but at the very least it should be achieved for the test function to ensure that the diagnostic uses interrogations in a sensible way. Condition (2a) is the more desirable version since it directly targets the shape of

the function and also implies (2b) by design, but in high dimensions with large interrogation grids it may only be possible to ensure that (2b) is met. The first condition establishes τ as the "borderline" function: any function that is "less Gaussian" will have its LA rejected, and any function "at least as Gaussian" will not. To see this, consider the normalized posterior "correction term"[7]

$$\Delta(f) := \frac{\sqrt{\det(-H)}}{f(\hat{x})} \left[\int_{\mathbb{R}^d} C_0^x(z, \boldsymbol{s}) \mathrm{d}G(z) \right]^\top [C_0^x(\boldsymbol{s}, \boldsymbol{s})]^{-1} (r(\boldsymbol{s}) - m_0^x(\boldsymbol{s})), \quad (16)$$

which, as per (11), is (up to the scaling factors in front) the difference between the prior and posterior integral means when the diagnostic is applied to a function f. It can be shown that the rejection criterion for the diagnostic is equivalent to $f(\hat{x}) \sqrt{\det(-H^{-1})} |\Delta(f)| > 1.96\sqrt{C_1}$. Recall from Section 4.4 that C_1 only depends on f through scaling factors $f(\hat{x})^2$ and $\det\left(-H^{-1}\right)$, so the rejection criteria is equivalent to $|\Delta(f)| > \epsilon$, where the number $\epsilon > 0$ depends only on $\boldsymbol{s}^*, \lambda, \alpha$, and γ. Now, to meet condition (1) for the test function τ is to have $|\Delta(\tau)| = \epsilon$. Therefore, with this calibration scheme a function f will have its LA rejected iff $|\Delta(f)| > |\Delta(\tau)|$. Again, all that matters are the *relative differences* between a function and its Gaussian approximation at the interrogation points - specifically, whether the weighted sum of these as given by (16) (with the weights depending on $\boldsymbol{s}^*, \lambda$, and γ) is larger in magnitude than it is for the predetermined "borderline Gaussian" τ.

A natural choice for a test function is the density of a d-dimensional multivarite Student's t distribution with ν degrees of freedom, mean at the origin, and scale matrix equal to the identity. Denote this density by $\tau_{\nu,d}$, so

$$\tau_{\nu,d}(x) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}} \left(1 + \frac{\|x\|^2}{\nu}\right)^{-\frac{\nu+d}{2}}, \quad (17)$$

and note that it has heavier tails than a d-dimensional Gaussian density, so the LA, given by the formula

$$L(\tau_{\nu,d}) = \left(\frac{2}{\nu+d}\right)^{\frac{d}{2}} \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}, \quad (18)$$

underestimates the true integral (which is always equal to 1). However, $\tau_{\nu,d}$ approaches a standard multivariate normal density in the limit $\nu \to \infty$, and therefore $L(\tau_{\nu,d}) \to 1$ as well. Therefore, for some large value of ν, the shape of $\tau_{\nu,d}$ may be said to be "sufficiently Gaussian" to warrant non-rejection of the LA. Denote such a value by ν_d to reflect the fact (discussed further in Section 7) that the specific choice of test function should depend on the dimension d. One option that works reasonably well (at least, in low and moderate dimensions)

[7]To avoid any possible confusion, it should be reiterated that all of the quantities in these definitions - namely, $G, r, \boldsymbol{s}, m_0^x, m_1^x, C_0$, and C_1 - technically depend on f through the constructions detailed in Sections 3–4.3. More accurate notation would reflect this explicitly, but such notation would be cumbersome.

is to let ν_d be the smallest integer such that $L(\tau_{\nu_d,d}) \geq 0.95$. The densities of multivariate t variables 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 LA's that underestimate the true integral by over 5%.

With the family of test functions established, it is now possible to discuss how one may set the hyperparameters to satisfy the conditions listed above. First note that the precision parameter α does not actually affect the posterior mean; as a scaling factor, it serves only to ensure that condition (1) is met. Thus, it suffices to find good values for λ and γ, after which α can simply be chosen to scale the posterior variance C_1 such that $|\Delta(\tau_{\nu_d,d})| = \epsilon$.

The fact that λ affects the shape of the GP mean is obvious since, as noted in Section 4.2, it determines the "smoothness" of functions sampled from the GP and is therefore a "shape parameter" in some sense. What is perhaps more surprising is the effect of γ, the scaling factor for the underlying measure G. Recall from Section 4.3 that G is analogous to the proposal distribution in IS. It is well-known that the performance of an importance sampler will be poor if the density g has lighter tails than f, and it is therefore better to err on the side of caution by taking g to have slightly heavier tails [e.g. 46]. In our context, this corresponds to setting γ slightly larger than 1, and in our experiments we use a value of

$$\gamma = \sqrt{1.5 \frac{\nu_d + d}{\nu_d + d - 3}}. \tag{19}$$

The heuristic motivation for this choice is as follows. Consider d-dimensional random vectors $Y \sim \tau_{\nu_d,d}$ and $X \sim g$, where $g = g(\tau_{\nu_d,d})$ is the density corresponding to (15) for the choice of function $f = \tau_{\nu_d,d}$. The γ-value given by (19) ensures that $1.5 \times \mathrm{Var}[Y_1 \mid Y_2 = 0, \ldots, Y_d = 0] = \mathrm{Var}[X_1 \mid X_2 = 0, \ldots, X_d = 0]$ - in words, the univariate conditional densities (with all other coordinates fixed at the origin) of the t distribution used for calibration have variance equal to two thirds of those of the "approximating Gaussian density" g [26]. Here the analogy with IS becomes somewhat strained, as it can be easily shown that *any* Gaussian proposal distribution will result in an importance sampler with infinite variance when applied to a t density. In fact, taking G itself as a t distribution is often a good choice in IS due to the heaviness of the tails [45, and references therein]. Prüher et al. [38] considered this choice of G in BQ, but noted that the kernel integrals in (11–12) would not have closed forms. For computational convenience we will retain our choice of a Gaussian measure, but note that, unlike IS, the posterior variance of the integral is still guaranteed to be finite here.

5.1 Calibrating in two dimensions

Using these ideas, we will now demonstrate how calibration can work for the diagnostic in $d = 2$ dimensions. The test function will be a bivariate t density

with $\nu_2 = 38$ degrees of freedom, as $L(\tau_{38,2}) = 0.95$. The preliminary interrogation grid $\boldsymbol{s}^*$ will consist of evenly-spaced points in a "cross-shaped" formation "on the axes" of $\mathbb{R}^2$:

$$\boldsymbol{s}^* = \{(0,0)\} \cup \{\pm m e_i : m = 1,2,3, i = 1,2\}, \tag{20}$$

where e_i is the i^{th} standard basis vector of $\mathbb{R}^2$. Such "cross-shaped grids" are appealing, at least in low dimensions, because the number of points n scales lineary with d. Here, we have $n = 13$.

In order to heuristically understand how hyperparameter choices affect the behaviour of the diagnostic, it will be useful to plot the difference between the test function $\tau_{38,2}$ and the "un-weighted" GP posterior mean $m_1^x \cdot g$ for various (λ, γ)-values. Note that the "optimal" hyperparameters will depend on the dimensionality of the domain, the specific test function used, and the preliminary grid chosen. In particular, if one wishes to use the diagnostic in 2 dimensions with a different preliminary grid from the one considered here, it should not necessarily be assumed that the λ value given below is suitable for the new grid.

Choosing γ according to (19) with $d = 2$ and $\nu_d = 38$ results in a value of $\gamma = 1.2734$. In this low-dimensional setting with a small interrogation grid, it is possible to crudely approximate an analytic method to find an "optimal" λ: given the aforementioned γ-value, we approximate the "L^2 error" $\int_{\mathbb{R}^2} \left(m_1^x(x) g(x) - \tau_{38,2}(x)\right)^2 \mathrm{d}x$ and its derivative w.r.t. λ by simple Riemann sums over the grid of points $\{-10, -9.99, -9.98, \ldots, 9.99, 10\}^2$. This approximate error is then minimized w.r.t. λ using the BFGS algorithm as implemented in the `fminunc` function in the MATLAB Optimization Toolbox [31], resulting in a value of $\lambda = 4.2241$.

Figure 1 shows results for the diagnostic applied to $\tau_{38,2}$ with these design choices. The difference $m_1^x \cdot g - \tau_{38,2}$ is very small among the lines defined by the interrogation grid, but there are deep valleys centered around the "main diagonals" of the plane and within the boundaries of the interrogation grid. Since the heavy-tailed t density is larger than its Gaussian approximation in these regions, it is clear that there is not much difference between the prior and posterior GP means there. The interrogation points are too far from these regions to exert much influence on the posterior mean there - in this respect, one may say that the GP is failing to *interpolate* to these areas. A more mathematical explanation of this behaviour can be extracted from (8), the definition of m_1^x. By this definition, it holds that $m_1^x(\boldsymbol{s})\, g(\boldsymbol{s}) = f(\boldsymbol{s})$ for any f and any combination of hyperparameter values. However, at any other point x, the extent to which $m_1^x(x)$ updates from the prior GP mean $m_0^x(x)$ is determined by the "weights" $C_0^x(x, \boldsymbol{s})^\top \left[C_0^x(\boldsymbol{s}, \boldsymbol{s})\right]^{-1}$. These weights tend to decrease in magnitude as x moves away from the points in $\boldsymbol{s}$, to an extent determined by λ and γ. When λ is small, there is almost no prior dependence between GP values at distinct points, so these weights are close to zero for $x \notin \boldsymbol{s}$. This can be seen in Figure 2: the posterior GP mean is forced to equal $\tau_{38,2}$ at the interrogation points, but everywhere else it is virtually unchanged from the prior mean m_0^x).

Thus, in this case m_1 is very close to the prior value $m_0 = L(\tau_{38,2}) = 0.95$. In contrast, the "optimal" λ-value results in a posterior integral estimate of $m_1 = 0.99095$, quite close to the true value of 1. As mentioned above, α is chosen to ensure that the test function is on the boundary between rejection and non-rejection, resulting in a posterior variance of $C_1 = 4.3653 \times 10^{-4}$ for the "optimal" λ and 5.7369×10^{-8} for the lower one.

The effect of γ is less easily explained than that of λ. In fact, their effects counterbalance each other to some degree: we found that it was still possible to approximate an "optimal" λ with the method described above even for different fixed values of γ, with lower γ-values resulting in higher required λ-values and vice-versa. In principle, this suggests that the diagnostic will not be too sensitive to the use of different γ-values, since any possible negative effect on its performance could be mitigated by adjusting λ in the opposite direction. However, there is a limit to this in practice, and γ-values that are either too low or too high can still be problematic. With a lower value of $\gamma = 1$, it became difficult to find an optimal λ, as the BFGS algorithm was quite sensitive to the choice of initial value. Although the results of differently-initialized BFGS runs were not consistent with each other, they all resulted in final λ-values over 9. At length-scales this large, the *Gram matrix* $C_0^x(\boldsymbol{s}, \boldsymbol{s})$ is poorly conditioned (for instance, with $\boldsymbol{s}^*$ given by (20), its reciprocal condition number is 7.7885×10^{-14} when $\lambda = 9$, as opposed to 7.1579×10^{-10} when $\lambda = 4.2241$), so numerical stability becomes a concern. Furthermore, even with λ-values this high, the posterior integral mean m_1 was around 0.986: not as close to 1 as it was with the slightly larger γ-value and its "optimal" λ. The fact that these difficulties exist for $\gamma = 1$ is noteworthy since this corresponds to using an integrating measure whose density is proportional to the Gaussian approximation to the true function.

Concerns about numerical accuracy do not exist with an even larger γ-value, as the corresponding optimal λ-value will be smaller and the Gram matrix will therefore be better conditioned. However, sensitivity becomes a problem in this situation: when γ is high, even a relatively small deviation from the optimal λ can change the diagnostic's behaviour quite dramatically. This will be of particular concern in higher dimensions, in which it is not viable to approximate and optimize the L^2 error numerically. In the current 2-dimensional setting, with $\gamma = 3$, the approximately-optimal λ-value is 1.1953, and the results with these hyperparameters (not shown) are fairly similar to those in Figure 1. A modest increase to $\lambda = 1.3$ creates a noticeably different outcome, as shown in Figure 3. The "interpolation valleys" seen in Figure 1 are slightly smaller in size, as the larger length-scale increases dependence between distinct points in the GP, thereby allowing the interrogations to exert more influence at faraway points. However, this slight improvement in interpolation comes at a price: However, this comes at a cost: undesirable *extrapolation* effects due to oversmoothing. Indeed, in all four directions just beyond the extremal interrogation points, m_1^x dips well below the true function $\tau_{38,2}$. As a result, $m_1 = 0.98108$ is farther from the true integral than it was with the hyperparameter values in Figure 1. Oversmoothing causes the weights $C_0^x(x, \boldsymbol{s})^\top \left[C_0^x(\boldsymbol{s}, \boldsymbol{s})\right]^{-1}$ to have unpredictable

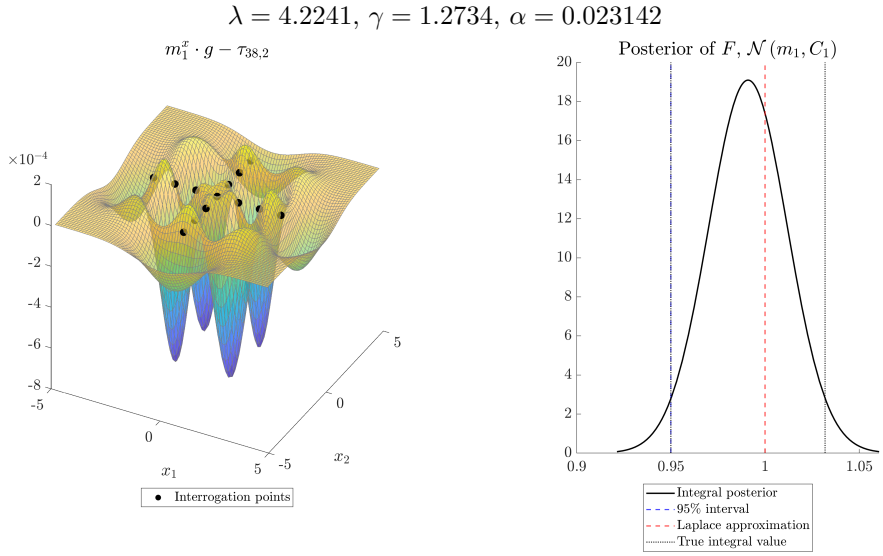

Figure 1: Results for the diagnostic applied to the 2-dimensional test function $\tau_{38,2}$, with an "optimal" λ, γ obtained from (19), and α set to ensure that the LA is on the boundary of the "rejection region". Left: the difference between the un-weighted posterior GP mean and the true function. Right: the posterior distribution for the integral F.

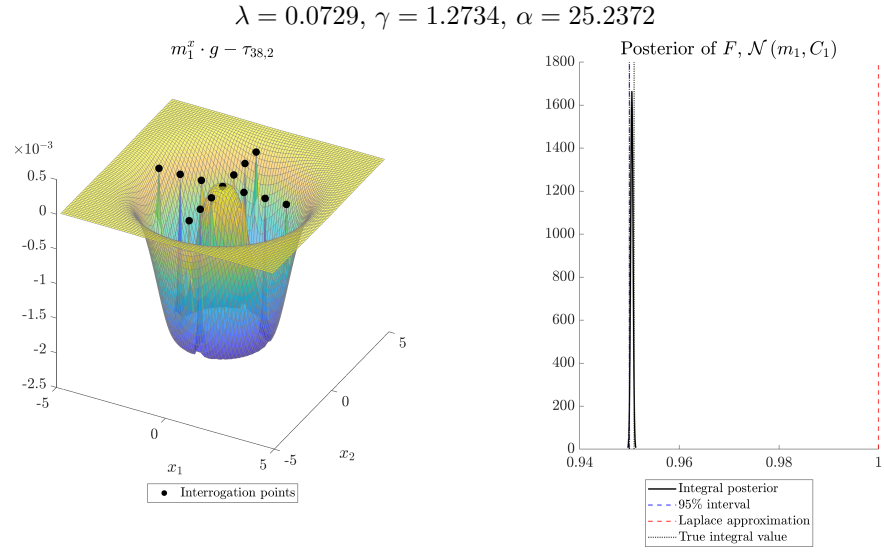

Figure 2: Results for the diagnostic applied to $\tau_{38,2}$ with a low λ-value.

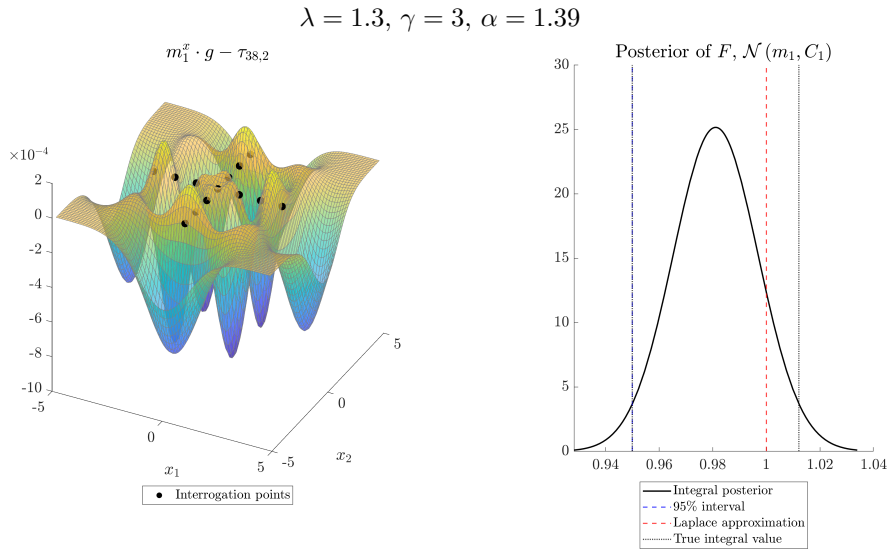

Figure 3: Results for the diagnostic applied to $\tau_{38,2}$ with a higher γ-value and a λ-value that is somewhat larger than the optimum.

effects at x beyond the boundaries of the interrogation grid, depending on the spread and density of $\boldsymbol{s}$ as well as the shape of the integrand. In some cases, the "extrapolation valleys" seen in Figure 3 may be replaced by large "hills", causing m_1 to significantly overestimate the value of F (not shown). It is now clear that the original hyperparameter values in Figure 1 provide the best "tradeoff", balancing the interpolation errors of undersmoothing with the extrapolation errors of oversmoothing.

6 Example: a banana-shaped function

In a paper on MCMC algorithms, Haario et al. [18] considered a function with "banana-shaped" contours, defined by "twisting" one coordinate of a Gaussian density. Letting $\varphi(\cdot; \Sigma)$ denote a bivariate Gaussian density with mean at the origin and covariance matrix Σ, the version of the function used here is

$$\beta(x) := \varphi\left(x_1, x_2 - \frac{1}{2}\left(x_1^2 - 3\right); \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}\right).$$

It turns out that the Laplace approximation is true for this function: $L(\beta) = \int_{\mathbb{R}^2} \beta(x) \mathrm{d}x = 1$. As discussed in Section 2, this may be viewed as "coincidence", as it is clear from Figure 4 that β is not well-approximated by a Gaussian shape. In this way, the function β represents an interesting test case for the diagnostic: although its LA is technically valid, it is *not* "Gaussian enough" and should therefore be rejected. Indeed, with the preliminary interrogation grid (20) and

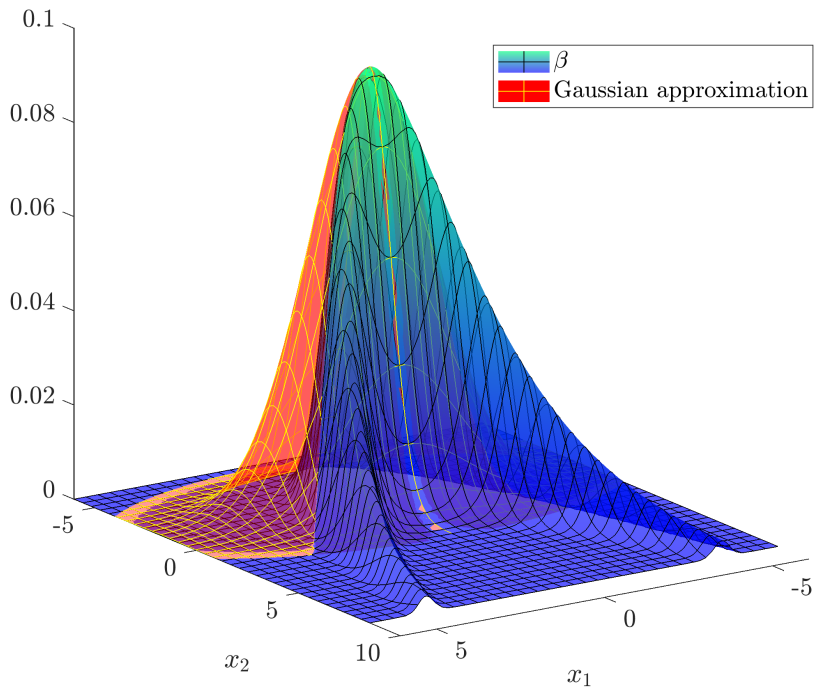

Figure 4: A two-dimensional "banana-shaped" function alongside its Gaussian approximation.

corresponding (approximately) "optimal" hyperparameters (see Figure 1), this is precisely what the diagnostic does, as shown in Figure 5. The un-weighted GP posterior mean now accurately captures the light tails of β along the line $x_2 = 0$, although it does not capture the large ridges defining the "banana" shape since there are no interrogation points along these ridges. As a result, the posterior integral estimate m_1 is 0.3658 - well below the true value and the LA. Note also that there are small oscillations between the interrogation points along the x_1-axis, perhaps signifying a small amount of oversmoothing. Finally, observe that the posterior variance is small enough to result in a rejection of the LA, which is well above the 97.5% quantile for the posterior distribution of F. These design choices would certianly be poor ones if accurate integral estimation was the main goal. In this framework, however, they are clearly suitable - the shape information captured by the diagnostic suggests that β is not Gaussian enough to justify using the LA outright. In this type of scenario, a practitioner could subsquently employ a different method to estimate the integral. Presumably, they would then discover that the LA was correct all along - but *not* because of the quality of the Taylor approximation (3) underpinning its use.

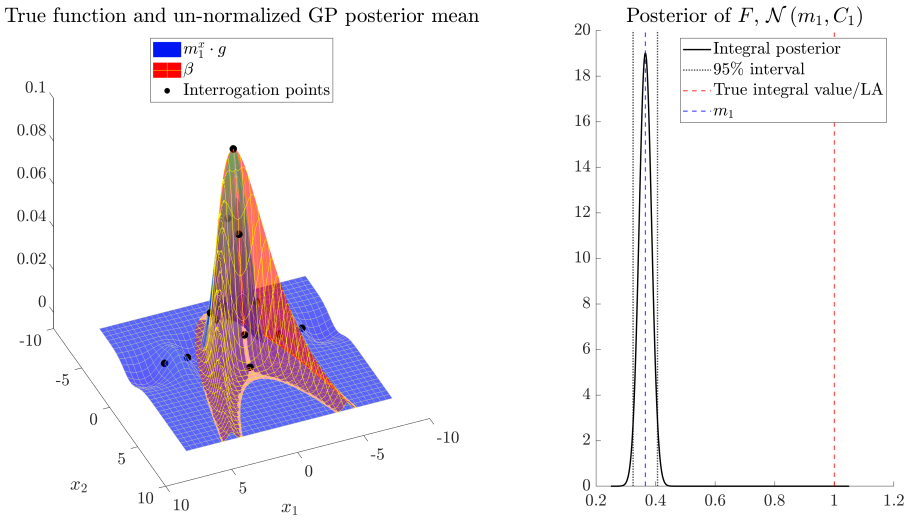

Figure 5: Results from applying the diagnostic to the two-dimensional banana-shaped function, using the same design choices as in Figure 1. Note that the colours in the left plot are reversed from those in Figure 4 for easier visualization.

7 High-dimensional considerations & examples

The low-dimensional experiments of Sections 5.1 and 6 were useful for exposition, but ultimately our main interest is in applying the diagnostic to higher-dimensional functions. Unsurprisingly, for large dimensions d it is more challenging to ensure that the diagnostic behaves well. Recall from Section 5.1 that we found an approximately "optimal" length-scale λ by minimizing a type of L^2 error associated with the calibration function $\tau_{\nu_d,d}$. This required the numerical approximation of an integral over $\mathbb{R}^d$, which is not computationally feasible in high dimensions (if it was, there would be no need for the LA or for this very diagnostic). It is also not viable to seek a closed-form expression for the L^2 error: doing so would, in turn, require an analytic expression for the inverse of the Gram matrix $C_0^x(\boldsymbol{s}, \boldsymbol{s})$, which will be prohibitively complicated for all but the smallest of interrogation grids. In moderate dimensions $d > 2$, condition (2a) from Section 5 can be assessed with a heuristic visual approach: viewing a 2-dimensional "slice" of the difference $m_1^x \cdot g - \tau_{\nu_d,d}$ with $x_3, \ldots, x_d$ all set to 0 (exploiting the symmetry of the t density and the fact that its mode is at the origin), one can adjust λ so as to make this difference appear as uniformly small as possible, attempting to balance issues with interpolation and extrapolation. Unfortunately, even this approach ceases to be viable when d is large, so that Condition (2b) is all that can be ensured. The reasons for this depend on the structure of the preliminary grid $\boldsymbol{s}^*$; in turn, this structure should be chosen to mitigate the challenges that arise in high dimensions. More details on some possible choices are given below. We found in preliminary experiments that

grids of the form (20) - that is, those with multiple evenly-spaced points along each axis - did not work very well when generalized to higher dimensions. Note that, although the points along any given axis are equally spaced in such grids, the distances between points on *different* axes will be larger. We conjecture that this variation in interrogation point distances becomes problematic in high dimensions as more axes and points are added.

Fundamentally, the issue in high dimensions is that a function's "shape information" - of the type described in the preceding sections - becomes more divorced from the value of its integral, making it more difficult to test the notion of "sufficiently Gaussian shape to justify the LA". There are a few different possible causes for this. The first is a well-known "curse of dimensionality" affecting certain high-dimensional probability density functions: most of their mass is in the tails, far away from the high-density region directly surrounding the mode [e.g. 8, 5]. Essentially, this happens because the neighbourhood around the mode is of a much smaller (Lebesgue) volume than the region encompassing the tails, so that most of the mass contributing to the integral is in a "shell" where the *product* of density and volume is high [ibid.]. For instance, if X is a d-dimensional standard normal random variable, the *Gaussian annulus theorem* [6, Theorem 2.9] states that, with high probability, X will be in a spherical shell of width $\mathcal{O}(1)$ and distance $\mathcal{O}\left(\sqrt{d}\right)$ from the origin.

This poses an unfortunate challenge for the diagnostic: when the integrand f is a high-dimensional density, its shape is easiest to visually assess around the mode where its values are relatively large, but its integral (and its LA, which is the integral of the Gaussian approximation to f) may be determined farther away where f is much smaller. For example, consider the case $d = 72$ (the dimensionality of the real-data examples in Section 7.1), for which (as explained in Section 5) we take the calibration function τ to be a multivariate t density with $\nu_{72} = 25921$ degrees of freedom because $L\left(\tau_{25921,72}\right) = 0.95$. The top plot of Figure 6 shows the integral of this density - and that of its Gaussian approximation ($m_0^x \cdot g$, in the notation of Section 3) - over the 72-dimensional ball $\{x : \|x\| < r\}$ as the radius r varies. Observe that both τ and its Gaussian approximation have most of their mass between distances 7–10 from the origin. Furthermore, the difference between the integrals does not start to become apparent until the radius of integration is at least 8 (note that, as $r \to \infty$, the integrals of τ and its Gaussian approximation converge to 1 and the LA, respectively). This affirms the idea that most of the important information about the integral (in particular, its closeness to the LA) is quite far from the mode, in a region that authors such as Betancourt [5] call the *typical set*. In contrast, the region of maximal *shape difference* between τ and its Gaussian approximation occurs much closer to the origin, where there is almost no mass. This can be seen in the bottom plot of Figure 6, which shows that τ differs most from its Gaussian approximation at a distance of approximately 2 from the origin. Even there, the largest difference between them is only about 0.002% of τ's value at the mode. Further out in the aforementioned "typical set", the two functions are visually indistinguishable.

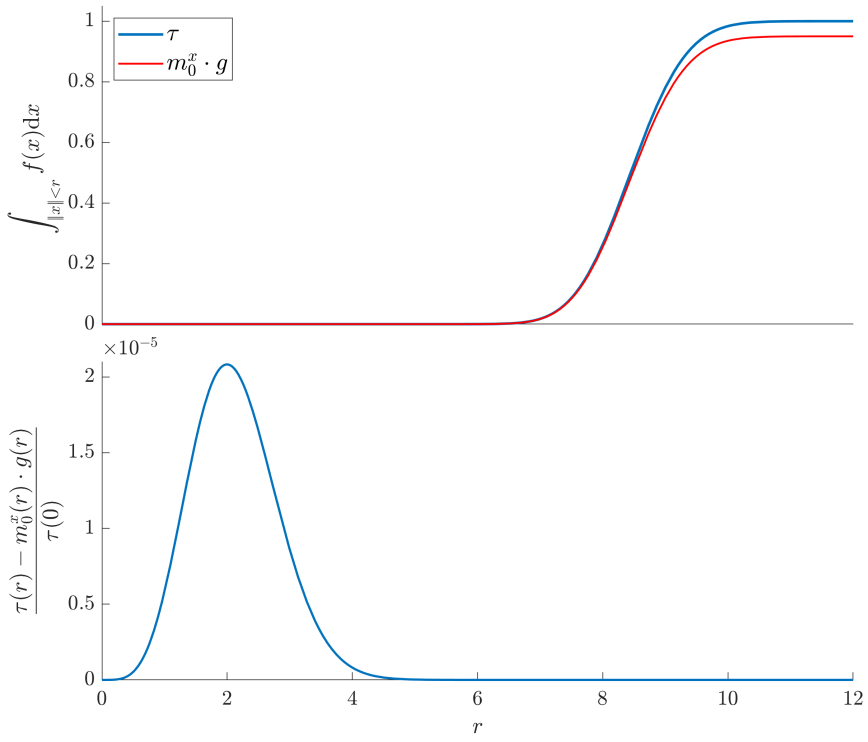

Figure 6: Top: the amount of mass enclosed by $\tau = \tau_{25921,72}$ and its Gaussian approximation over a ball of radius r centered at the origin. Bottom: the difference between τ and its Gaussian approximation at a distance of radius r from the origin, normalized by the value of τ at the origin.

There is another interesting point to be made here about the high-dimensional diagnostic. It was stated in Section 5 that ν_d, the degrees of freedom for the calibration function in d dimensions, would depend on d itself. Indeed, the Laplace approximation (18) for a multivariate t density is decreasing in d for fixed ν. Thus, if ν_d is defined, as previously suggested, to be the smallest integer such that $L\left(\tau_{\nu_d,d}\right) \geq 0.95$, then ν_d is necessarily an increasing function of d. Put another way, in higher dimensions a t density must be closer in shape to a Gaussian for its LA to be within 5% of the true integral value. Indeed, using this definition of ν_d in 72 dimensions resulted in the extremely high value $\nu_{72} = 25921$. The difference between the resulting t density and its Gaussian approximation is small enough to be virtually invisible, but because this difference is compounded over a (typical) set of extremely high volume, it results in a sizable difference between integrals.

In light of these ideas, our first suggested design for a high-dimensional diagnostic uses a preliminary grid $\boldsymbol{s}^* = \{\mathbf{0}\} \cup \left\{\pm\sqrt{d}e_i : i = 1, \ldots, d\right\}$, where $\mathbf{0}$ denotes the origin and e_i once again denotes the i^{th} standard basis vector of $\mathbb{R}^d$. This will result in $2d+1$ interrogation points: one at the mode, and two at distances of $\mathcal{O}\left(\sqrt{d}\right)$ away from it along each "principal axis" (see Section 4.1). Per the discussion above, if a function f is assumed *a priori* to have similar shape to a Gaussian density, then it is reasonable to expect this type of design to provide the most pertinent information about its integral. As described by Särkkä et al. [42], this choice of $\boldsymbol{s}^*$ in BQ creates a connection with *sigma-point methods*, in which such grids are used to estimate integrals for filtering and smoothing in nonlinear SSM's [e.g. 23]. In particular, aside from the inclusion of the origin this choice of $\boldsymbol{s}^*$ is identical to the point set used in the *cubature Kalman filter* (CKF) of Ienkaran and Haykin [21].

With this preliminary grid in $d = 72$ dimensions, we use $\tau_{25921,72}$ as our calibration function and once again take the hyperparameter γ as in (19), resulting in a value of $\gamma = 1.2248$. As alluded to above, here λ cannot be selected to visually ensure that Condition (2a) is met as in the low-dimensional experiments of Section 5.1. Because the differences between the calibration function and its Gaussian approximation are so small at the chosen interrogation points, adjusting λ does not produce any visible change in the difference $m_1^x \cdot g - \tau_{25921,72}$ (not shown). Thus, we must rely on the weaker Condition 2(b): selecting λ to produce a reasonable posterior integral estimate m_1. We found $\lambda = 3.7$ to be a good choice for this, giving a posterior integral mean of $m_1 = 0.998$. Finally, $\alpha = 0.1565$ is once again chosen to ensure that the calibration curve's LA (equal to 0.95) is on the boundary of the rejection region. Note that, although we were unable to use shape information as directly as we did in the low-dimensional experiments, the diagnostic's rejection criterion still depends solely on the "correction term" (16), itself a measure of deviation between a function and its Gaussian approximation. It could be said that the high-dimensional diagnostic, as it is configured here, determines whether a function is sufficiently Gaussian *in the tails* to justify the LA.

7.1 Example: North Sea cod modelling

This section returns to the SSM discussed in Sections 1 – 2. Recall that, given observed data $\boldsymbol{y}$, such a model can be fit by maximizing the Laplace-approximated marginal likelihood (integrating over hidden states $\boldsymbol{x}$) with respect to parameters θ. These methods are increasingly common in fisheries science, where they are used for *stock assessment*: to infer population dynamics for various species of fish given observations from surveys and commercial catches [2]. SSM's applied to stock assessment are often called *state-space assessment models (SSAM's)* [ibid.] and serve as a natural context to test our diagnostic: although the LA is commonly used in practice for these models, if the joint likelihood (1) is not "sufficiently Gaussian" in shape, then the LA may not be a suitable proxy for the marginal likelihood (2) and the resulting inferences may be incorrect.

To investigate the performance of our diagnostic in this "real-world" setting, we use a dataset containing multiyear measurements of cod stocks in the North Sea and fit SSAM's to various subsets of this data following Aeberhard et al. [2]. The observations y_t are taken on an annual basis over the span of several decades ($t = 1963, \ldots, 2015$). Briefly, for a given year t, y_t is a vector comprising the amounts of cod of different ages observed during surveys and commercial catches conducted that year[8]. The hidden state x_t contains, for each age group, the "true" abundance and fishing mortality rate for cod in that age group in year t. Finally, θ represents a variety of "global" parameters such as scaling factors and variances. The SSAM used here [see 34, and references therein] is highly nonlinear, with complex dependencies between the age-specific components of x_t and x_{t-1}. For the sake of brevity further details are omitted here, but they are available in the appendix of Aeberhard et al. [2]. All models were fit using the `stockassessment` R package [34, 4], which is in turn built on the TMB package [29].

Two SSAM's are considered here, each corresponding to a different subset of the available data: one fit to the data collected from 1970 to 1975 (hereafter the "1970 model"), and another to the data from 2005–2011 (the "2005 model"). Since each hidden state x_t is of dimension 12, using these six-year "windows" results in a latent dimensionality of $d = 12 \times 6 = 72$ for each model: fairly modest (and computationally convenient) compared to the 636 dimensions associated with the full dataset [2], but still large enough that any non-LA approach to marginalizing the likelihood would be far from trivial[9]. As stated above, the Laplace-approximated marginal likelihood $L(p_{xy})$ is maximized numerically w.r.t. θ, and ideally we would like to use our diagnostic at each step of this optimization to ensure it remains accurate throughout. For simplicity in these experiments, we only apply the diagnostic at the last optimization step,

[8] Note that the dimensionality of y_t is not constant with t, as the time ranges of the commercial catches and surveys only partially overlap. However, "missing observations" are not a problem for either model fitting or the diagnostic.

[9] We also found that, with smaller time windows, there was not sufficient data to guarantee model convergence. Even six-year windows besides the ones used here typically did not converge without careful selection of algorithm settings.

seeking to determine *only for the final parameter values* $\hat{\theta}$ whether $p_{xy}\left(\cdot, \boldsymbol{y} \mid \hat{\theta}\right)$ is "Gaussian enough" to justify the LA.

In order to assess the performance of the diagnostic, it is desirable to have some other estimate of the marginal likelihood $p_y\left(\boldsymbol{y} \mid \hat{\theta}\right)$ to serve as an approximate "ground truth". Since standard numerical integration is completely nonviable in 72 dimensions, we instead obtain such estimates via importance sampling [e.g. 16, and references therein]. For both models, samples were taken from a noncentral multivariate t distribution with mean $\hat{x}$, scale matrix $-H^{-1}$, and 5 degrees of freedom [15]. The joint likelihoods of both models appear to have light tails in $\boldsymbol{x}$ (see below), so this choice of importance distribution should mitigate the risk of infinite variance in theory [46, 45]. However, because we can only assess the tail behaviour of the models in finitely many directions, we cannot rule out the possibility that, somewhere in the 72-dimensional space, they have a tail even heavier than that of a t density. We conjecture that this is not the case, although the existence of such a tail could certainly result in a sampler with infinite variance. A more pressing concern is that poor finite-sample performance can still occur even with theoretical guarantees. Nevertheless, importance sampling is not the main concern here - it is intended only as a convenient, if somewhat informal, check on the LA diagnostic.

This diagnostic is not the only way to check the LA for a SSM — the `checkConsistency` function in the TMB package [29] provides another method[10]. It is essentially a *score test* [39] for the Laplace-approximated marginal likelihood: by simulating many separate data sets $\boldsymbol{y}^* \sim p_y\left(\cdot \mid \hat{\theta}\right)$ (which can be done by simulating $\boldsymbol{x}^* \sim p_x$, then $\boldsymbol{y}^* \sim p_{y|x}$), it constructs a test statistic to test the hypothesis $\mathbb{E}_y\left[\nabla_\theta \log L\left(p_{xy}\right)\big|_{\hat{\theta}}\right] = 0$, under which the test statistic is asymptotically χ^2-distributed. Since the true marginal score function has mean zero, a rejection of this hypothesis means that the LA is *not* a suitable approximation for the marginal likelihood p_y. It will be useful to compare this method to our diagnostic, but it should be noted that there is a key conceptual difference between them. The `checkConsistency` methodology views $L\left(p_{xy}\right)$ and p_y as *functions of* $\boldsymbol{y}$; with this view, it seeks to determine whether the marginal likelihood is well approximated by the LA, and what effects this approximation could have on the bias of the estimated $\hat{\theta}$. In contrast, our diagnostic is focused on shape of the joint likelihood p_{xy} when viewed *as a function of* $\boldsymbol{x}$: in particular, whether this shape warrants the use of the LA to fit the model *for the observed* (i.e. *fixed*) $\boldsymbol{y}$.

Figure 7 shows results (from the diagnostic, as well as the aforementioned importance sampler with differing numbers of sample) for the 1970 model. For the importance samplers, 95% confidence intervals were obtained with a Gaussian approximation, using the sample variance of the IS weights. The central limit theorem dictates that for a well-behaved importance sampler, the width of

[10] Refer to the source code at https://github.com/kaskr/adcomp/blob/master/TMB/R/checker.R for further detail. Notes provided by Anders Nielsen in personal correspondence also helped to inform this discussion.

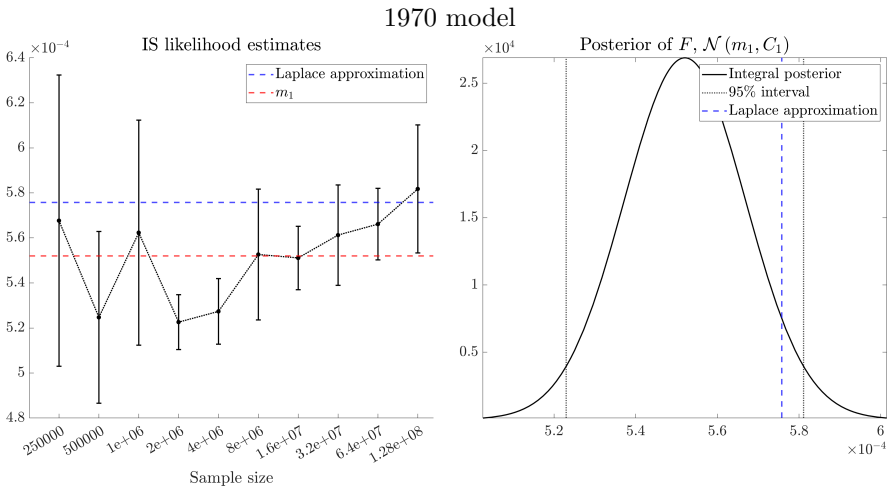

Figure 7: Results of the diagnostic applied to the 1970 SSAM. Left: IS estimates of $p_y\left(\boldsymbol{y} \mid \hat{\theta}\right)$ at various sample sizes (black dots) with estimated 95% confidence intervals (vertical line segments), with the Laplace approximation (blue dashed line) and the posterior integral mean (red dashed line) for reference. Right: the posterior distribution for the marginal likelihood, obtained from the diagnostic.

these intervals should be roughly $\mathcal{O}\left(S^{-1/2}\right)$, where S is the number of samples. The left plot of Figure 7 indicates that this may not be the case. Indeed, the score test of Koopman et al. [27] rejected the hypothesis that these samplers had finite variance. These rejections are typically the result of a few large weights, which seemingly indicate that in a few directions the tails of $p_{xy}\left(\cdot, \boldsymbol{y} \mid \hat{\theta}\right)$ are too heavy relative to those of the proposal density. However, further numerical evidence indicated that the tails of the squared joint likelihood were eventually dominated by its Gaussian approximation in those directions. In mathematical terms, at all sampled points $x \in \mathbb{R}^d$ for which the importance weights were large, it appeared that, for sufficiently large $r > 0$,

$$\left[p_{xy}\left(\hat{x} + rz, \boldsymbol{y} \mid \hat{\theta}\right)\right]^2 = o\left(\phi\left(\hat{x} + rz\right)\right) \tag{21}$$

as functions of r, where ϕ is the Gaussian approximation to $p_{xy}\left(\cdot, \boldsymbol{y} \mid \hat{\theta}\right)$ and z is a unit vector in the direction of $x - \hat{x}$. Since the ratio of a Gaussian density and a Student's t density is certainly integrable over $\mathbb{R}^d$, this provides some limited indication that the integral defining the variance of the importance sampler [e.g. 15] may indeed be finite after all. This is a very informal check on the validity of IS, and it certainly does not guarantee finite-sample stability. However, their use as a heuristic reference against which to check the diagnostic does not seem unreasonable here.

Most of the importance samplers include the LA within their 95% confidence

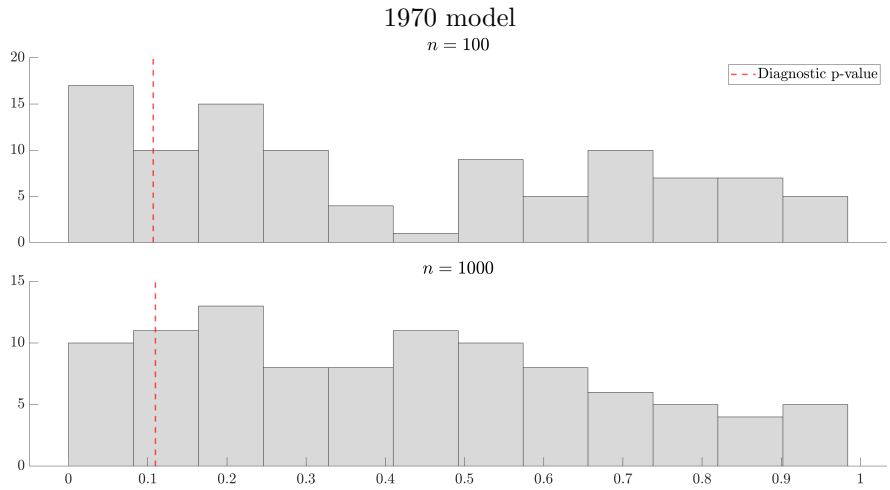

Figure 8: Histograms of p-vals from repeated runs (100 runs each for simulation sizes $n = 100$ and $n = 1000$) of `checkConsistency` on the fitted 1970 SSAM. The "p-value" given by the diagnostic is shown as a dashed red line on each histogram.

intervals, suggesting it is not excessively far from the true marginal likelihood value. The fact that most of the IS estimates are below the LA suggests that the latter is perhaps a slight overestimate of the true value (i.e. that the tails of the joint likehood, as a function of x, tend to be lighter than those of its Gaussian approximation). Our diagnostic produces a similar conclusion: the posterior integral mean is slightly lower than the LA, but not to a degree that warrants rejection. With respect to our notion of "good-enough-ness-of-fit", it seems that the LA is a reasonable approximation to the marginal likelihood for this model, at least for the parameter values $\hat{\theta}$.

Since the diagnostic is based on a Gaussian "confidence interval" for the integral (see Section 3), its behaviour can be equivalently described in terms of "p-values": recalling from (10) that the integral posterior is $F \mid r(\boldsymbol{s}) \sim \mathcal{N}(m_1, C_1)$, it is straightforward to show that the diagnostic rejects the LA iff

$$2\left[1 - \Phi\left(\frac{|m_1 - L(f)|}{\sqrt{C_1}}\right)\right] < 0.05,$$

where Φ is the cdf of a standard Normal random variable, and the quantity on the left-hand side has a natural interpretation as a sort of "p-value". This facilitates some comparison between the diagnostic and the `checkConsistency` method. Recall that the latter simulates n separate data sets to construct a test statistic that is asymptotically χ^2-distributed when $\mathbb{E}_y\left[\nabla_\theta \log L\left(p_{xy}\right)\big|_{\hat{\theta}}\right] = 0$. This test statistic induces a p-value; if this is below some threshold (say, 0.05), we reject the hypothesis that the marginal likelihood and the LA are the same (as functions of $\boldsymbol{y}$). In Figure 8, we have performed the `checkConsistency`

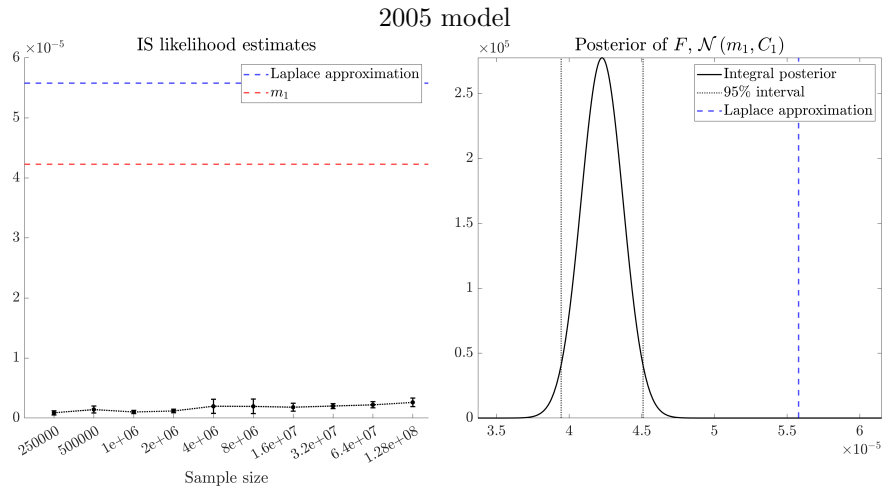

Figure 9: Results of the diagnostic applied to the 2005 SSAM. Left: IS estimates of $p_y\left(\boldsymbol{y} \mid \hat{\theta}\right)$ at various sample sizes (black dots) with estimated 95% confidence intervals (vertical line segments), with the Laplace approximation (blue dashed line) and the posterior integral mean (red dashed line) for reference. Right: the posterior distribution for the marginal likelihood, obtained from the diagnostic.

test 100 times each for two simulation sizes ($n = 100$ and $n = 1000$) in order to see how the p-value distribution changes with the number of simulated data sets and how it relates to the p-value of the diagnostic. If the null hypothesis of `checkConsistency` is true (i.e. the LA is the true marginal likelihood), then the p-value of the corresponding test should be uniformly distributed over $(0, 1)$. Although the histograms in Figure 8 show some deviation from uniformity, it is not severe. The p-value associated with the diagnostic is just above 0.1, consistent with non-rejection of the LA (see Figure 7). It is interesting to see from Figure 8 that the diagnostic and `checkConsistency` seem to lead to similar conclusions — that the LA may deviate slightly from the true marginal likelihood, but not to a problematic extent — despite the fundamental difference in the questions addressed by each method.

The results are markedly different for the 2005 model, as shown in Figure 9. IS stability considerations apply here as they did for the 1970 model: Koopman et al.'s score test [27] rejected the hypothesis of finite variance for the largest sample sizes, but (21) held in the directions of all the largest weights, potentially indicating a finite (but possibly quite large) variance. All IS estimates are far lower than the LA, suggesting that the joint likelihood is, for the most part, substantially lighter-tailed than its Gaussian approximation. Accordingly, the diagnostic strongly rejects the LA, which is well above the upper bound of the posterior 95% confidence interval. Note that there is still substantial disagreement between the diagnostic and the importance samplers as it pertains

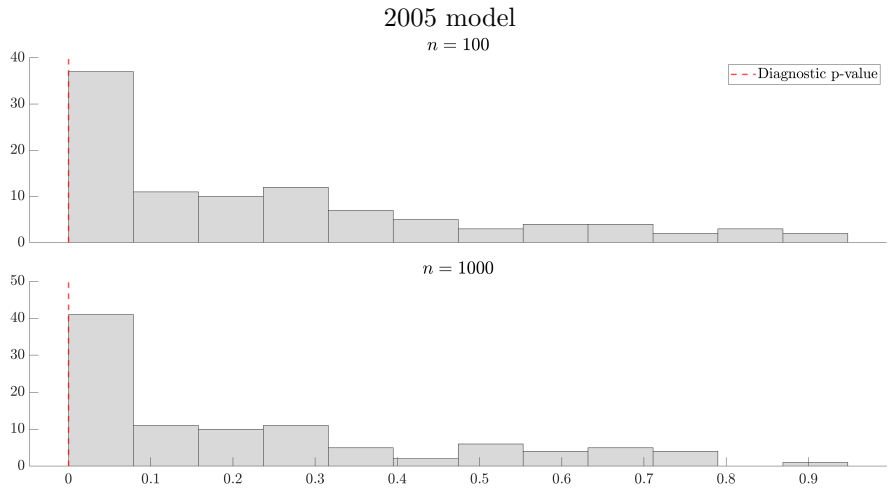

Figure 10: Histograms of p-vals from repeated runs (100 runs each for simulation sizes $n = 100$ and $n = 1000$) of `checkConsistency` on the fitted 2005 SSAM. The "p-value" given by the diagnostic is shown as a dashed red line on each histogram

to estimation of the true marginal likelihood. Thus, the posterior integral mean from the diagnostic should not be taken as a high-quality estimate, but what is important is that both methods agree on rejection of the LA.

As before, we also conduct repeated runs of `checkConsistency` and compare the resulting p-value distributions to the one associated with the diagnostic. The latter is numerically indistinguishable from zero, and for both simulation sizes the p-value distribution is decidedly non-uniform. As was the case with the 1970 model, both methods appear to agree that the LA is an unsuitable approximation to the marginal likelihood, despite asking this question in different ways.

Differing philosophies aside, one clear advantage the diagnostic has over `checkConsistency` is computation time. Using the `checkConsistency` replications shown in Figures 8 and 10, as well as 100 repeated computations of the diagnostic itself, Table 1 shows median computation times — along with median absolute deviations — for each method applied to each model. Note that the time cost for the diagnostic includes the evaluation of function interrogations, the eigendecomposition of the Hessian, and the calculation of all the necessary kernel terms for BQ (the latter step was sped up substantially using the methods of Karvonen and Särkkä [25], as explained in Section 7.2. It is also interesting to note the differences in computational times between models: across all methods, the times for the 2005 model are longer than those for the 1970 model. Presumably, this is because of the "inner" numerical optimization [29] used to calculate the mode $\hat{x} = \hat{x}\left(\boldsymbol{y}, \hat{\theta}\right)$, which may require more iterations

Time (seconds)	1970 model	2005 model
`checkConsistency`, $n = 100$	2.511 ± 0.035	7.367 ± 0.136
`checkConsistency`, $n = 1000$	25.115 ± 0.152	73.584 ± 0.489
Diagnostic	0.009 ± 0.007	0.012 ± 0.0003

Table 1: Table showing median computation times (along with median absolute deviations) of each method, applied to each model.

for the 2005 model than the 1970 model due to differences in their respective joint likelihoods. This would also explain why the difference is so much more pronounced for the `checkConsistency` runs, which require repeated (and possibly even more demanding) inner optimizations to find $\hat{x} = \hat{x}\left(\boldsymbol{y}^*, \hat{\theta}\right)$ for each simulated dataset $\boldsymbol{y}^*$. In any case, the diagnostic is by far the fastest method of assessing the LA[11].

7.2 Higher-order interrogation grids

The interrogation grids used thus far have been quite simple, consisting of $\mathcal{O}(d)$ preliminary points placed along the axes of $\mathbb{R}^d$ in a d-dimensional "cross" shape. As noted in the introduction of Section 7, there is precedent in the literature for the use of such simple grids [42, 21]. They seem to be a reasonable choice here as well, allowing us to calibrate the diagnostic in such a way that appropriate results are obtained for a variety of "toy" and real-world examples. However, one potential drawback of such grids is that they only allow the diagnostic to use information about a function's shape along its "principal axes" (see Section 4.1). If this is not indicative of the function's behaviour in the rest of the domain, it is conceivable that the diagnostic could produce misleading results. For instance, consider the d-dimensional function

$$f_{\nu,d}(x) = \prod_{i=1}^{d} \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}} \left(1 + \frac{x_i^2}{\nu}\right)^{-\frac{\nu+d}{2}}. \tag{22}$$

Like the multivariate t density (17), this function has a mode at the origin. The functions are equal there, as are the Hessians of their logs. Furthermore, they are equal along the axes of $\mathbb{R}^d$. Thus, their LA's are the same, and the diagnostic would give the same results for both functions using any of the "cross-shaped" interrogation grids considered above. However, the functions differ on the rest of their domain, and their integrals are different as a result. Wheres the integral of $\tau_{\nu,d}$ over $\mathbb{R}^d$ is equal to 1 for all (ν, d), the integral of $f_{\nu,d}$ is

$$\frac{\Gamma\left(\frac{\nu+d-1}{2}\right)^d}{\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{\nu+d}{2}\right)^{d-1}}.$$

[11] IS computation times are not shown, as these were not replicated. However, they behaved largely as expected: computation times were roughly linear in the number of samples, and universally longer than those for the diagnostic.

In particular, for $d = 72, \nu = \nu_{72} = 25921$ (the values used to calibrate the 72-dimensional diagnostic at the beginning of this section), $\int_{\mathbb{R}^{72}} f_{25921,72}(x)\mathrm{d}x = 0.952$. Thus the integral of $f_{25921,72}$ is quite a bit closer to the LA (0.95) than that of the calibration function $\tau_{25921,72}$, but the diagnostic calibrated with a "cross-shaped" grid will treat both of them identically, so that the LA is on the boundary of the rejection region for each function. One could argue that this is undesirable: the values of $f_{25921,72}$ "off the axes" are lower (and therefore, closer to the Gaussian approximation) than those of the calibration function, causing its integral to be closer to the LA, so perhaps the diagnostic should produce a more definitive non-rejection for this function. For this to be possible, we must be able to capture the differences between $f_{\nu,d}$ and $\tau_{\nu,d}$, for which a *higher-order* interrogation grid is required.

Here, a grid of "order" s is one whose size scales as $\mathcal{O}(d^s)$ for some fixed power $s > 1$ (the grids used throughout the manuscript thus far had $s = 1$). In order to use such grids without an excessive increase in computation time (which would defeat the purpose of the diagnostic), we use *fully symmetric kernel quadrature* (FSKQ), as detailed by Karvonen and Särkkä [25]. Briefly, because the squared exponential kernel is isotropic, using fully symmetric preliminary grids (as described in Section 4.1) reduces the number of *unique* quadrature weights that need to be calculated, allowing for significant algebraic and computational simplifications in BQ.

Here, we recalibrate the 72-dimensional diagnostic using a *sparse Gauss-Hermite grid of order 2* — the two-dimensional version of which is shown in Figure 11 — as the preliminary grid. Following Karvonen and Särkkä [25], we remove the origin, as its quadrature weight tends to be a large negative value for most hyperparameter combinations. Furthermore, because a function is always equal to its Gaussian approximation at the mode, the origin does not actually contribute to the diagnostic beyond its effect on the inverted Gram matrix. We also multiply each point in the Gauss-Hermite grid by 3.6, thereby ensuring that they are far enough away from the origin to cover the "typical set" discussed at the beginning of this section. The final preliminary grid in 72 dimensions is of size $n = 10512$, and as with the original "cross-shaped" preliminary grid (which, for reference, contained $n = 145$ points) we calibrate the diagnostic using the t density $\tau_{25921,72}$ and taking the hyperparameter $\gamma = 1.2248$. As before, it is not possible to calibrate with respect to Condition (2a) from 5. Here, this is because of the size of the grid: the computational simplifications of FSKQ are only applicable to the integral of the GP, not to the GP posterior mean function (8) itself. As such, the visual calibration of Section 5.1 is not viable: even though we would only need to view a 2-dimensional slice of $m_1^x \cdot g - \tau_{25921,72}$, every change to the hyperparameter λ would still necessitate the recalculation and inversion of the 10512×10512 Gram matrix, which is too slow for minute visual adjustments. Instead, we once again calibrate with respect to Conditions (1) and (2b), resulting in hyperparameters $(\lambda, \alpha) = (3.7, 0.1349)$ and a posterior integral mean of $m_1 = 0.9945$ for the calibration function.

Applying the new calibrated diagnostic with the larger preliminary grid to the SSAM's from Section 7.1 reveals that the use of higher-order grids does not

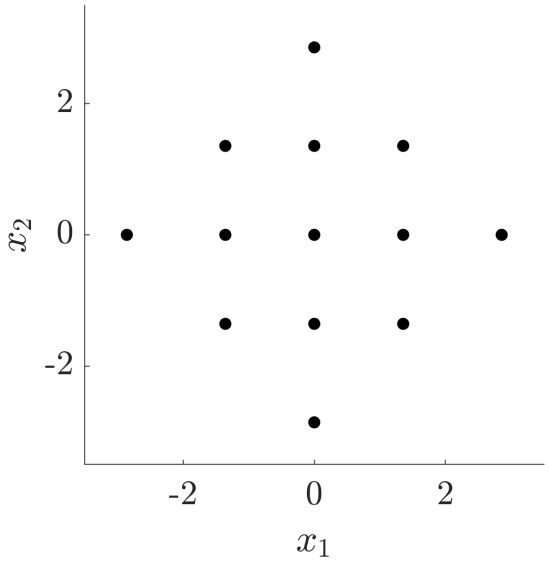

Figure 11: A sparse Gauss-Hermite quadrature grid of order 2 in $d = 2$ dimensions.

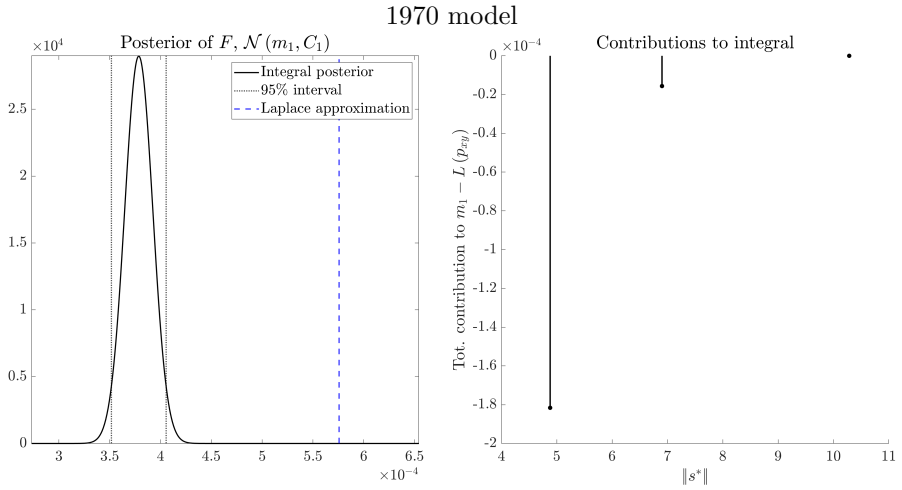

Figure 12: Results of the diagnostic with a higher-order interrogation grid applied to the 1970 SSAM. Left: the posterior distribution for the marginal likelihood, obtained from the diagnostic. Right: the total mass contributions to the quadrature estimate made by interrogations as a function of the distance between the corresponding preliminary points and the origin.

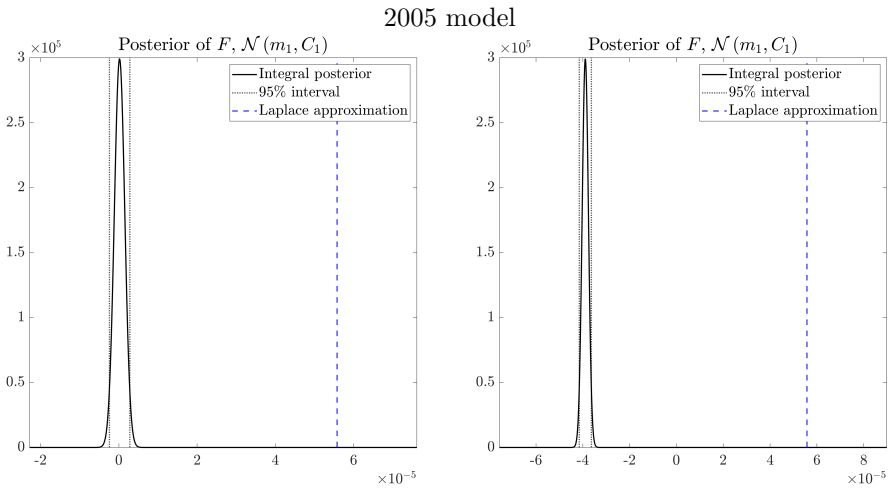

Figure 13: Results of the diagnostic with a higher-order interrogation grid applied to the 1970 SSAM. Left: the posterior distribution for the marginal likelihood, obtained from the diagnostic. Right: the total mass contributions to the quadrature estimate made by interrogations as a function of the distance between the corresponding preliminary points and the origin.

automatically cause an improvement in the diagnostic's behaviour — indeed, the opposite may occur. The left plot of Figure 12 shows that, in contrast to the results in Section 7.1, this version of the diagnostic rejects the LA for the 1970 model. Initially, this may suggest that the tails of the joint likelihood are substantially lighter than those of its Gaussian approximation in directions besides its "principal axes", which would not have been observable using the smaller grid. However, this is at odds with the results of the importance samplers and `checkConsistency`, both of which suggested that the LA was *not* very far from the true marginal likelihood and neither of which is constrained to the use of information on the principal axes of the joint likelihood. Furthermore, the right plot of Figure 12 reveals that the largest overall contribution to the lowered integral estimate comes from the interrogation points which are closest to the mode. This is despite the fact that there are only 144 such points in the Gauss-Hermite grid. In contrast, the points further from the origin — of which there are 10368 — collectively contribute a much smaller amount to the estimate. As discussed in the introduction to Section 7, most of a high-dimensional function's mass is in its tails; ideally this would be reflected when using a preliminary grid with most of its points far away from the origin. In light of these considerations, it seems reasonable to conclude that this version of the diagnostic is not providing accurate inference on the integral, or on the function shape information most pertinent to it.

The new diagnostic exhibits a different problem when applied to the 2005 model, as seen in Figure 13. The left plot shows that the LA is once again

definitively rejected, although the actual integral posterior differs quite noticeably from the one in Figure 9. However, as it turns out, there is one interrogation point s where the weighted difference $r(s) - m_0^x(s)$ (recalling the notation and terminology of Section 3) is far larger than it is for any of the other points. Removing this point from the grid, but keeping the hyperparameters fixed[12], results in a surprisingly large change in the posterior, shifting its mean from a small positive value to a larger negative value (which is nonsensical, given that the integral is a likelihood and must therefore be nonnegative). Although the diagnostic achieves its primary goal in both cases for this model — namely, determining that the joint likelihood's shape (as a function of $\boldsymbol{x}$) is too non-Gaussian to justify the LA — it is certainly undesirable for one interrogation point to have such a large impact. A given function's LA could be rejected based solely on the inclusion or exclusion of a single point at which it deviates significantly from its Gaussian approximation, thereby rendering the diagnostic too sensitive to be useful for nontrivial high-dimensional applications (see the discussion at the beginning of Section 4).

The computation times for the diagnostic with the higher-order grid are predictably higher than they were for the original diagnostic, although it is still much faster than `checkConsistency`. The median time was 0.4154 seconds for the 1970 model (MAD: 0.0105 seconds) and 0.5422 seconds for the 2005 model (MAD: 0.0104 seconds). Nevertheless, given the difficulties encountered above, the simpler CKF-style grid used in Section 7.1 seems to be a better choice.

8 Discussion

In this manuscript, we have built on the work of Zhou [47] to develop a non-asymptotic diagnostic tool for assessing the viability of Laplace approximations to integrals. More specifically and accurately, the diagnostic assesses whether a function's shape is close enough to the Gaussian approximation used to justify the LA. It does so using the method of Bayesian quadrature, but in multiple ways it is structured differently than a more "conventional" BQ application. Namely, we avoid design choices that would ensure accurate, low-uncertainty estimates for the integral of a specific function, opting instead for a "one-size-fits-all" approach: relatively simple interrogation grids intended to capture the most pertinent information about a function's behaviour, hyperparameters chosen heuristically using calibration functions, and a covariance structure that ensures the diagnostic is invariant to all properties of the integrand besides its shape. More broadly, the diagnostic is based on a notion of "good-enough-ness-of-fit" that stands in stark contrast to a more conventional, power-focused approach to statistical inference. Indeed, such an approach would render the diagnostic useless, causing it to prioritize the detection of *any* deviation from Gaussian

[12] Note that deleting the corresponding preliminary interrogation point did not produce a sizeable change in the diagnostic's behaviour when applied to the calibration function (not shown), despite not adjusting the hyperparameters for the altered grid. Thus, there is no concern about miscalibration here.

shape and likely producing rejections in almost all non-trivial applications.

Challenges arise when using the diagnostic in high dimensions, although they are not insurmountable. Compared to low-dimensional settings, it is more difficult to make conclusions about a function's integral given limited information about its shape - either because a high-dimensional function's mass tends to be far away from the regions with the most notable "shape information" (the curse of dimensionality), or because a single direction of non-Gaussian shape (which, intuitively, seems more likely to occur in high dimensinos) can affect the diagnostic's behaviour to an unreasonable extent. Because of these challenges, more consideration must be given in high-dimensional spaces when choosing the preliminary interrogation grid and setting the hyperparameters, and the focus must be on the function's shape in its tail regions, assumed to correspond to its "typical set". If this is done carefully, the diagnostic can be calibrated to produce reasonable and useful results on real-world examples, as shown in Section 7.1.

Given SSAM's that had already been fit (producing parameter estimates $\hat{\theta}$), we applied the diagnostic to their joint likelihoods $p_{xy}\left(\cdot, \boldsymbol{y} \mid \hat{\theta}\right)$. While this served the purposes of this manuscript (namely, a proof-of-concept for the diagnostic itself), it ignores the fact that the parameter estimate itself depends on the use of Laplace approximations: specifically, that it is obtained by maximizing the LA $L\left(p_{xy}\left(\cdot, \boldsymbol{y} \mid \theta\right)\right)$ with respect to θ. Given the low computational cost of the diagnostic, it would be desirable to fold it directly into a model-fitting workflow, checking at each iteration of numerical optimization whether or not the LA is justified, thereby indicating if other methods need to be invoked to correct any incurred bias in the estimated model parameters.

Despite the promising initial performance of the diagnostic, there are opportunities for future potential improvements. The difficulties of using higher-order grids encountered in Section 7.2 should be further explored, as their resolution could result in improved diagnostic behaviour on a wider variety of functions. The methods of choosing interrogation points cited in the introduction of Section 4 may be a useful starting point to this end, but care must be taken to modify these methods in a way that preserves the quick, "one-size-fits-all" nature of the diagnostic. Another aspect of the diagnostic that remains unaddressed is the prior structure: speficially, that our use of a GP prior is *technically* inappapropriate given that most applications involve likelihoods, which are nonnegative. It is worth investigating other prior speficiations proposed in the BQ literature [e.g. 17, 9], which preserve nonnegativity of the integrand at the expense of inducing a non-analytic distribution on the integral which must be approximated.

As a final note, we conjecture that the methods developed here may be more broadly applicable beyond the assessment of Laplace approximations. Indeed, a great deal of statistical methods are based on an assumption that some function is well approximated by a Gaussian shape, which is precisely the assumption that the diagnostic is designed to check. The general idea of using non-asymptotic methods to diagnose the use of asymptotic methods is one that certainly warrants further consideration and study.

References

[1] Shigeo Abe. Training of support vector machines with Mahalanobis kernels. In *Proc. International Conference on Artificial Neural Networks (ICANN 2005)*, pages 571–576. Springer, Berlin, Heidelberg, 2005. ISBN 3540287558. doi: 10.1007/11550907_90. URL `http://www2.eedept.kobe-u.ac.jp/$\sim$abe`.

[2] William H Aeberhard, Joanna Mills Flemming, and Anders Nielsen. Review of State-Space Models for Fisheries Science. *Annual Review of Statistics and Its Application*, 5:215–235, 2018. doi: 10.1146/annurev-statistics. URL `https://doi.org/10.1146/annurev-statistics-`.

[3] O E Barndorff-Nielsen, D R Cox, and H.F.D.R. Cox. *Asymptotic Techniques for Use in Statistics*. Asymptotic Techniques for Use in Statistics. Springer US, 1989. ISBN 9780412314001. URL `https://books.google.ca/books?id=UQ9yIrZpMToC`.

[4] Casper W. Berg and Anders Nielsen. Accounting for correlated observations in an age-based state-space stock assessment model. *ICES Journal of Marine Science: Journal du Conseil*, 73:1788–1797, 2016. doi: 10.1093/icesjms/fsw046.

[5] Michael Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo. Technical report, 2018.

[6] Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of Data Science*. Cambridge University Press, jan 2020. ISBN 9781108755528. doi: 10.1017/9781108755528. URL `https://www-cambridge-org.proxy.lib.sfu.ca/core/books/foundations-of-data-science/6A43CE830DE83BED6CC5171E62B0AA9E`.

[7] François-Xavier Briol, Chris J. Oates, Mark Girolami, Michael A. Osborne, and Dino Sejdinovic. Probabilistic Integration: A Role in Statistical Computation? *Statistical Science*, 34(1):1–22, 2019. URL `http://www.`

[8] Bob Carpenter. Typical sets and the curse of dimensionality, 2017. URL `https://mc-stan.org/users/documentation/case-studies/curse-dims.html`.

[9] Henry Chai and Roman Garnett. Improving Quadrature for Constrained Integrands. Technical report, 2019.

[10] 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`.

[11] Nicolas Chopin and Omiros Papaspiliopoulos. *An Introduction to Sequential Monte Carlo*. Springer Series in Statistics. Springer International Publishing, Cham, 2020. ISBN 978-3-030-47844-5. doi: 10.1007/978-3-030-47845-2. URL `https://link.springer.com/10.1007/978-3-030-47845-2`.

[12] Jon Cockayne, Chris Oates, Tim Sullivan, and Mark Girolami. Bayesian Probabilistic Numerical Methods. *SIAM Review*, 61(4):756–789, feb 2019. URL `http://arxiv.org/abs/1702.03673`.

[13] Nicolaas Govert De Bruijn. *Asymptotic methods in analysis*, volume 4. Courier Corporation, 1981.

[14] Perry de Valpine. Review of methods for fitting time-series models with process and observation error and likelihood calculations for nonlinear, non-Gaussian state-space models. *Bulletin of Marine Science*, 70(2):455–471, 2002.

[15] Michael Evans and Tim Swartz. Methods for Approximating Integrals in Statistics with Special Emphasis on Bayesian Integration Problems. *Statistical Science*, 10(3):254–272, 1995.

[16] John Geweke. Bayesian Inference in Econometric Models Using Monte Carlo Integration. *Econometrica*, 57(6):1317–1339, 1989.

[17] Tom Gunter, Michael A Osborne, Roman Garnett, Philipp Hennig, and Stephen J Roberts. Sampling for Inference in Probabilistic Models with Fast Bayesian Quadrature. In *Advances in Neural Information Processing Systems*, pages 2789–2797, 2014.

[18] Heikki Haario, Eero Saksman, and Johanna Tamminen. Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*, 14:375–395, 1999.

[19] Philipp Hennig, Michael A. Osborne, and Mark Girolami. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179):20150142, jul 2015. ISSN 1364-5021. doi: 10.1098/rspa.2015.0142. URL `https://royalsocietypublishing.org/doi/10.1098/rspa.2015.0142`.

[20] Ferenc Huszár and David Duvenaud. Optimally-weighted herding is bayesian quadrature. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 377–386, 2012.

[21] Ienkaran and Simon Haykin. Cubature kalman filters. In *IEEE Transactions on Automatic Control*, volume 54, pages 1254–1269, 2009. doi: 10.1109/TAC.2009.2019800.

[22] I.T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag New York, New York, 2 edition, 2002. ISBN 0-387-95442-2. doi: 10.1007/b98835. URL `http://link.springer.com/10.1007/b98835`.

[23] Simon Julier and Jeffrey K. Uhlmann. A General Method for Approximating Nonlinear Transformations of Probability Distributions. Technical report, University of Oxford, Oxford, 1996.

[24] Vesa Kaarnioja. Smolyak Quadrature. Master's thesis, University of Helsinki, 2013.

[25] Toni Karvonen and Simo Särkkä. Fully symmetric kernel quadrature. *SIAM Journal on Scientific Computing*, 40(2):A697–A720, mar 2018. ISSN 10957197. doi: 10.1137/17M1121779.

[26] Marc Kennedy. Bayesian quadrature with non-normal approximating functions. *Statistics and Computing*, 8:365–375, 1998.

[27] Siem Jan Koopman, Neil Shephard, and Drew Creal. Testing the assumptions behind importance sampling. *Journal of Econometrics*, 149: 2–11, 2009. doi: 10.1016/j.jeconom.2008.10.002. URL `www.elsevier.com/locate/jeconom`.

[28] Shinsuke Koyama, Lucia Castellanos Pérez-bolde, Cosma Rohilla Shalizi, and Robert E. Kass. Approximate Methods for State-Space Models. *Journal of the American Statistical Association*, 105(489):170–180, 2010. doi: 10.1198/jasa.2009.tm08326.

[29] Kasper Kristensen, Anders Nielsen, Casper W. Berg, Hans Skaug, and Bradley M. Bell. TMB: Automatic differentiation and laplace approximation. *Journal of Statistical Software*, 70(1):1–21, apr 2016. ISSN 15487660. doi: 10.18637/jss.v070.i05. URL `https://www.jstatsoft.org/index.php/jss/article/view/v070i05/v70i05.pdfhttps://www.jstatsoft.org/index.php/jss/article/view/v070i05`.

[30] D. V. Lindley. The Use of Prior Probability Distributions in Statistical Inference and Decisions. In *Proc. 4th Berkeley Symp. on Math. Stat. and Prob.*, volume 4, pages 453–468. University of California Press, jan 1961.

[31] The MathWorks, Inc. *MATLAB Optimization Toolbox*. Natick, MA, USA, 2019. URL `https://www.mathworks.com/help/pdf_doc/optim/optim.pdf`.

[32] Thomas P. Minka. Deriving quadrature rules from Gaussian processes. Technical report, Carnegie Mellon University, 2000.

[33] Lawrence M. Murray. Bayesian state-space modelling on high-performance hardware using LibBi. *Journal of Statistical Software*, 67(10), oct 2015. ISSN 15487660. doi: 10.18637/jss.v067.i10.

[34] Anders Nielsen and Casper W. Berg. Estimation of time-varying selectivity in stock assessments using state-space models. *Fisheries Research*, 158:96–101, 2014. doi: 10.1016/j.fishres.2014.01.014.

[35] A. O'Hagan. Bayes-Hermite quadrature. *Journal of Statistical Planning and Inference*, 29:245–260, 1991. doi: 10.1016/0378-3758(91)90002-V.

[36] Michael Osborne. *Bayesian Gaussian Processes for Sequential Prediction, Optimisation and Quadrature*. PhD thesis, University of Oxford, 2010.

[37] Michael A Osborne, David Duvenaud, Roman Garnett, Carl E Rasmussen, Stephen J Roberts, and Zoubin Ghahramani. Active Learning of Model Evidence Using Bayesian Quadrature. In *Advances in neural information processing systems*, pages 46–54, 2012.

[38] Jakub Prüher, Filip Tronarp, Toni Karvonen, Simo Särkkä, and Ondřej Straka. Student-t process quadratures for filtering of non-linear systems with heavy-tailed noise. In *20th International Conference on Information Fusion*, Xi'an, China, aug 2017. Institute of Electrical and Electronics Engineers Inc. doi: 10.23919/ICIF.2017.8009742.

[39] C Radhakrishna Rao. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 44, pages 50–57. Cambridge University Press, 1948.

[40] Carl Edward Rasmussen and Zoubin Ghahramani. Bayesian Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 489–496, 2003. URL `http://www.gatsby.ucl.ac.uk`.

[41] Carl Edward. Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006. ISBN 9780262182539.

[42] Simo Särkkä, Jouni Hartikainen, Lennart Svensson, and Fredrik Sandblom. On the relation between Gaussian process quadratures and sigma-point methods. Technical report, 2015.

[43] Hans J Skaug and David A Fournier. Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. *Computational Statistics & Data Analysis*, 51:699–709, 2006. doi: 10.1016/j.csda.2006.03.005. URL `www.elsevier.com/locate/csda`.

[44] Luke Tierney and Joseph B. Kadane. Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

[45] Surya T. Tokdar and Robert E Kass. Importance sampling: a review. *Advanced Review*, 2:54–60, 2009. doi: 10.1002/wics.56.

[46] Changhe Yuan and Marek J. Druzdzel. Theoretical analysis and practical insights on importance sampling in Bayesian networks. *International Journal of Approximate Reasoning*, 46:320–333, 2007. doi: 10.1016/j.ijar.2006.09.006. URL `www.elsevier.com/locate/ijar`.

[47] 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`.