Statistics and Computing (2023) 33:64
https://doi.org/10.1007/s11222-023-10225-3

ORIGINAL PAPER

Variational Tobit Gaussian Process Regression

Marno Basson[1] · Tobias M. Louw[1] · Theresa R. Smith[2]

Received: 1 September 2022 / Accepted: 21 February 2023 / Published online: 31 March 2023

Abstract
We propose a variational inference-based framework for training a Gaussian process regression model subject to censored observational data. Data censoring is a typical problem encountered during the data gathering procedure and requires specialized techniques to perform inference since the resulting probabilistic models are typically analytically intractable. In this article we exploit the variational sparse Gaussian process inducing variable framework and local variational methods to compute an analytically tractable lower bound on the true log marginal likelihood of the probabilistic model which can be used to perform Bayesian model training and inference. We demonstrate the proposed framework on synthetically-produced, noise-corrupted observational data, as well as on a real-world data set, subject to artificial censoring. The resulting predictions are comparable to existing methods to account for data censoring, but provides a significant reduction in computational cost.

Keywords Gaussian process regression · Tobit regression · Bayesian statistics · Censored data · Variational inference · Local variational methods

1 Introduction

Central to any data analysis procedure is data gathering. A practical problem that typically arises during the data gathering process is censoring, which occurs when we partially observe a measurement. An example of data censoring occurs when the measured value falls outside the sensitivity range of the measurement device (e.g. a temperature sensor). Specialized inference techniques are required to address the problems that arise from censored data.

Tobit models are a popular class of censored regression models, tracing back to the work of Tobin (1958). Subsequently, Amemiya (1984) provided a detailed survey and taxonomy of the different parametric variations of Tobit approaches. These models have been adapted and applied to numerous settings. For example, Allik et al. (2016) use a parametric type I Tobit model to develop a formulation of the Kalman filter suitable for censored observations. Recent censored regression frameworks have focused more on combining censored models with flexible architectures that can capture the underlying nonlinear relationships in data. These, for example, include deep neural networks (Wu et al. 2018), random forests (Hutter et al. 2013; Li and Bradic 2020) and Gaussian process models (Ertin 2007; Groot and Lucas 2012; Chen et al. 2013; Gammelli et al. 2020a, b).

Gaussian processes (GPs) provide a fully Bayesian nonparametric approach for performing inference for nonlinear functions and have become increasingly more popular in the machine learning community (MacKay 2004; Rasmussen and Williams 2006; Bishop 2009; Titsias and Lawrence 2010). Using the GP regression framework, we can derive the full Bayesian predictive density for such functions, allowing us to estimate a mean function and quantify uncertainty around the mean estimate (Snelson et al. 2004; Groot and Lucas 2012). In GP regression, point estimates for the unknown kernel function parameters are often obtained by maximizing the log marginal likelihood of the observed data or, in variational methods, a lower bound on the log marginal likelihood. However, the presence of the censored observations means that this marginal likelihood cannot be computed in closed-form.

Ertin (2007) proposed a censored GP regression framework, within the context of censored wireless sensor readings, by treating the censored variable as a mixture of a binary and a Gaussian random variable followed by defining a GP prior over the latent function values. Ertin (2007) circumvents the analytical intractability of the posterior density

✉ Tobias M. Louw
tmlouw@sun.ac.za

[1] Department of Process Engineering, Stellenbosch University, Stellenbosch 7600, South Africa

[2] Department of Mathematical Sciences, University of Bath, BA2 7AY, Bath, UK

and the marginal likelihood of this model by approximating the posterior density with a Laplace approximation (Bishop 2009).

Groot and Lucas (2012) then extended the censored GP regression framework to include the type I Tobit model (see Amemiya 1984). They circumvent the analytically intractable posterior density by applying expectation propagation (Minka 2001a, b) with the goal to approximate the type I Tobit likelihood terms by local likelihood factors using non-normalized Gaussian density functions. This work has been applied to wind power forecasting (Chen et al. 2013), predicting clinical scores from neuro-imaging data (Rao et al. 2016), and modeling the demand for shared transport services while allowing for time-varying detection limits (Gammelli et al. 2020a).

Gammelli et al. (2020b) propose an extension of the work of Groot and Lucas (2012) by {1} incorporating a non-constant heteroskedastic observation model, {2} using a multi-output GP prior to exploit information from potentially correlated outputs to enable better modeling of the censored data, and {3} circumventing the analytical intractability that arises from the proposed framework by developing a variational lower bound on the log marginal likelihood which they optimized with stochastic variational inference (Hoffman et al. 2013; Blei et al. 2017).

In this article, we provide a mathematical tool that allows us to derive a closed-form variational lower bound on the log marginal likelihood of the original probabilistic model by applying variational sparse GP regression in conjunction with local variational methods. Our proposed methodology is closely related to the work of Ertin (2007) and Groot and Lucas (2012) and, similar to Gammelli et al. (2020b), relies on variational methods to perform approximate inference.

A key development in our approach is that we maximize a secondary variational lower bound on the Tobit model which relies on {1} the variational sparse GP regression framework developed by Titsias (2008, 2009) and {2} local variational methods which aim to lower bound the Tobit likelihood factors instead of approximating these factors (see Jordan et al. 1999; Nickisch and Rasmussen 2008; Bishop 2009). The use of the variational sparse GP framework results in a reduction in time complexity (Titsias 2009), thereby enabling us to perform inference on larger censored data sets previously intractable to an analysis by GP regression models. To the best of our knowledge, such an implementation does not yet exist in the current censored Gaussian process regression literature. We demonstrate that our variational inference-based framework computationally outperforms the competing benchmarks while maintaining comparable prediction accuracy.

The remainder of the article is structured as follows. Section 2 focuses on the theoretical development of the Tobit GP regression model and Section 3 introduces the variational approximations that allow us to derive a closed-form variational lower bound that can be used for Bayesian model training and inference. In Section 4 we derive the required equations for the latent function predictive posterior density, while Section 5 demonstrates the ability of the proposed framework to learn a latent function representation from observational data subject to artificial censoring. In Section 6 we end with a discussion followed by making explicit some of the limitations associated with the proposed framework.

2 The Tobit Gaussian Process Regression Model

In this section, we briefly review the standard GP regression model and then introduce the theoretical framework for Tobit GP regression.

Suppose we have a data set consisting of pairs $\{(x_i, y_i)\}_{i=1}^{N}$. We assume that each observation y_i is a noisy, independent realization of an unknown latent function $f_i = f(x_i)$ at scalar input x_i, with additive noise from a zero mean Gaussian density with unknown variance σ_y^2:

$$y_i = f_i + \epsilon_i; \quad \epsilon_i \sim \mathcal{N}(\epsilon_i | 0, \sigma_y^2) \tag{1}$$

This induces a joint Gaussian likelihood function of the form

$$p(\boldsymbol{y}|\boldsymbol{f}, \sigma_y) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{f}, \sigma_y^2 \boldsymbol{I}_{NN}) \tag{2}$$

We denote with $\mathcal{N}(\cdot)$ the Gaussian density function, $\boldsymbol{y} \in \mathbb{R}^{N\times1}$ the vector of observed data, and $\boldsymbol{f} \in \mathbb{R}^{N\times1}$ the vector of latent function values at the training input locations $\boldsymbol{x} \in \mathbb{R}^{N\times1}$. The matrix $\boldsymbol{I}_{NN}$ denotes the $N \times N$ identity matrix. Next, we specify a zero mean GP prior with kernel function $k(x_i, x_j)$ such that

$$f \sim \mathcal{GP}(0, k(x_i, x_j)) \tag{3}$$

For the finite set of training input locations $\boldsymbol{x}$ associated with $\boldsymbol{f}$, the GP follows a multivariate Gaussian density with covariance matrix $\boldsymbol{K}_{NN}$, the $N \times N$ covariance matrix which is constructed using the user-specified kernel function $k(x_i, x_j)$ on the training input locations:

$$p(\boldsymbol{f}|\boldsymbol{\theta}_k) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K}_{NN}) \tag{4}$$

where $\boldsymbol{\theta}_k$ collectively denotes the typically unknown kernel function parameters.

Point estimates for the unknown kernel parameters $\boldsymbol{\theta}_k$ and unknown noise variance σ_y^2, which we collectively denote by the parameter vector $\boldsymbol{\theta}$, can be obtained by using gradient-based optimization to maximize the log marginal likelihood of the model which is given by

$$\ln p(\boldsymbol{y}) = \ln \left[\int_{\boldsymbol{f}} p(\boldsymbol{y}|\boldsymbol{f}, \sigma_y^2) p(\boldsymbol{f}|\boldsymbol{\theta}_k) d\boldsymbol{f} \right]$$

For the Gaussian likelihood function in Eq. (2), the marginal likelihood of the model can be computed analytically as

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\theta}) &= \mathcal{N}(\boldsymbol{y}|\boldsymbol{0}, \boldsymbol{C}_{NN}) \\ \boldsymbol{C}_{NN} &= \boldsymbol{K}_{NN} + \sigma_y^2 \boldsymbol{I}_{NN} \end{aligned} \tag{5}$$

Refer to Rasmussen and Williams (2006) and Bishop (2009) for a detailed overview of the Gaussian process regression framework.

The Tobit Gaussian process regression model can be thought of as an extended version of the standard GP regression model as applied to censored observational data. For censored data, the standard GP regression likelihood function (see Eq. (2)) is no longer valid due to limitations that arise from our measurement sensitivity range.

Suppose that the detection limits for the measurement of interest are known in advance and constant with respect to time. When we observe that $y_i = l_b$, where l_b corresponds to the lower detection limit, we only know an upper bound on the corresponding observation for y_i, i.e., $y_i \in (-\infty, l_b]$, rendering the Gaussian assumption inappropriate (Groot and Lucas 2012).

To account for the limitation associated with the sensitivity range, we alter the way we construct our likelihood function. In latent function regions where we observe data, we retain the base GP architecture as outlined by Eqs. (1) to (2). However, in latent function regions where, for example, the measurement instrument/analysis procedure transforms (or reports) the data as the corresponding censored detection limit, we ask ourselves the following additional question:

What is the probability that the data, i.e., the random variable Y_i that is associated with marginal density $p(y_i|f_i)$, falls either (scenario 1) above the upper detection limit u_b or (scenario 2) below the lower detection limit l_b?

In other words, when we consider the marginal density associated with the random variable Y_i, we want to answer the following two questions (subject to which censoring scenario we consider)

$$\mathbb{P}(Y_i \geq u_b) = 1 - \int_{-\infty}^{u_b} p(y_i|f_i) dy_i \tag{6}$$

$$\mathbb{P}(Y_i \leq l_b) = \int_{-\infty}^{l_b} p(y_i|f_i) dy_i \tag{7}$$

Note that $\mathbb{P}(\cdot)$ denotes the probability value whereas $p(\cdot)$ denotes the probability density function, associated with the random variable Y_i, which we derive from Eq. (1) as

$$p(y_i|f_i) = \mathcal{N}(y_i|f_i, \sigma_y^2) \tag{8}$$

From Eqs. (6), (7) and (8) we can construct a piece-wise defined likelihood, i.e., a mixed-likelihood, which we will denote with the symbol $p_o(\cdot)$, that accounts for data censoring as follows

$$p_o(y_i|f_i) = \begin{cases} \Phi(l_b|f_i, \sigma_y^2) & \text{if } y_i = l_b \\ \mathcal{N}(y_i|f_i, \sigma_y^2) & \text{if } l_b < y_i < u_b \\ 1 - \Phi(u_b|f_i, \sigma_y^2) & \text{if } y_i = u_b \end{cases} \tag{9}$$

We denote with $\Phi(\cdot)$ the Gaussian cumulative distribution function (cdf). Furthermore, note that we implicitly assumed that the latent function is corrupted by noise and that the noise-corrupted data value is then censored and reported (Groot and Lucas 2012). For notational convenience we use $\Phi(u_b|f_i, \sigma_y^2)$ to imply $\Phi(\frac{u_b - f_i}{\sigma_y})$.

Gammelli et al. (2020b) draws an interesting connection between heteroskedastic regression and censored observation models. The authors provide a qualitative understanding of the reasons why they suggest the use of input-dependent noise models and show, from a simulation-based perspective and real-world data sets, how heteroskedasticity can allow one to more accurately model the censored observations associated with Tobit-based likelihood functions.

As noted by Gammelli et al. (2020b), the likelihood variance parameter σ_y^2 directly controls the slope of the Gaussian cdf factors (see Eq. (9)) and would enforce the same amount of overestimation for all of the censored observations (refer to Appendix A, Section A.1). However, the amount of overestimation can be regulated/adjusted with a heteroskedastic parameterization for the variance. This would allow the Tobit model to automatically tune the amount of overestimation resulting in improved predictive performance. Consequently, we augment each Gaussian cdf factor in Eq. (9) with an additional variance parameter and construct an adjusted mixed-likelihood, which we will denote with the symbol $p_m(\cdot)$, that assigns the following probability/density function portions conditioned on the training input location

$$p_m(y_i|f_i) = \begin{cases} \Phi(l_b|f_i, \sigma_y^2 + \sigma_{l_b}^2) & \text{if } y_i = l_b \\ \mathcal{N}(y_i|f_i, \sigma_y^2) & \text{if } l_b < y_i < u_b \\ 1 - \Phi(u_b|f_i, \sigma_y^2 + \sigma_{u_b}^2) & \text{if } y_i = u_b \end{cases} \tag{10}$$

Note that for training input locations associated with the lower detection limit l_b we assume a constant (with respect to the input x_i) heteroskedastic noise model with a total variance contribution which is the sum of the original mixed-likelihood variance in Eq. (9) and a regulating variance parameter. A similar argument holds for the upper detection limit u_b (refer to Appendix A Section A.2 for more details). Note that the variance parameter for the uncensored observations remains the same as in Eq. (9). Given a censored data set with a total of N entries, and assuming independence, we can construct our mixed-likelihood function as follows

$$\prod_{i=1}^{N} p_m(y_i|f_i) = \prod_{y_i=l_b} [1 - \Phi(f_i|l_b, \sigma_y^2 + \sigma_{l_b}^2)] \times \prod_{l_b<y_i<u_b} \mathcal{N}(y_i|f_i, \sigma_y^2) \prod_{y_i=u_b} \Phi(f_i|u_b, \sigma_y^2 + \sigma_{u_b}^2) \quad (11)$$

We arrived at Eq. (11) by using Eq. (10) and the Gaussian cdf property $\Phi(y|x, \sigma^2) = 1 - \Phi(x|y, \sigma^2)$ (see Pishro-Nik 2014). Note that Eq. (11) is known as the Tobit likelihood function, or the type I Tobit model, and comprises a mixture of Gaussian density and Gaussian cdf likelihood terms (see Amemiya 1984; Groot and Lucas 2012). From here on we drop the dependence on any model parameters for notational convenience. We also abuse notation and use the following definition for notational convenience

$$p_m(\boldsymbol{y}|\boldsymbol{f}) = \prod_{i=1}^{N} p_m(y_i|f_i) \quad (12)$$

Inference about the latent function proceeds along the same line as for the standard GP regression model. We start with our mixed-likelihood function, as given by Eq. (12), and define a zero mean GP prior over $\boldsymbol{f}$ with kernel function $k(x_i, x_j)$ (see Eqs. (3) and (4)). Using Bayes' rule, we can compute a posterior density as follows

$$p(\boldsymbol{f}|\boldsymbol{y}) = \frac{p_m(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})}{p(\boldsymbol{y})} \quad (13)$$

Regardless of whether the factor $p_m(y_i|f_i)$ or $p_o(y_i|f_i)$ is used in the mixed-likelihood function, we refer to Eq. (13) as the Tobit Gaussian process regression (T-GPR) model. The marginal likelihood of the censored data set is given by

$$p(\boldsymbol{y}) = \int_{\boldsymbol{f}} p_m(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})d\boldsymbol{f} \quad (14)$$

However, unlike the standard GP regression model, the marginal likelihood given by Eq. (14) is analytically intractable due to the mixture of Gaussian density and Gaussian cdf likelihood terms (Ertin 2007; Groot and Lucas 2012).

3 Lower Bounding the Log Marginal Likelihood

Next, we propose circumventing this analytical intractability by adopting a variational inference-based framework, which will allow us to compute an analytically tractable lower bound on the true log marginal likelihood of the original probabilistic model given by Eq. (13). This new lower bound can then be used to perform Bayesian model training and inference.

3.1 Applying the Variational Sparse Gaussian Process Framework

The application of the standard GP regression model to large data sets has always been challenging due to the need to invert the $N \times N$ covariance matrix $\boldsymbol{C}_{NN}$ (see Eq. (5)) which requires time complexity that scales as $\mathcal{O}(N^3)$ where N is the number of data entries. For large data sets, the (numerical) inversion process becomes prohibitively slow rendering the standard GP regression model computationally intractable. Consequently, practitioners have resorted to using approximate or sparse methodologies to address the limitations associated with the (numerical) inversion process. Much research has primarily focused on advanced sparse GP methodologies where a smaller set of M function points are used as support/inducing variables. For example, see the work of Csató and Opper (2002), Seeger et al. (2003) and Snelson and Ghahramani (2005). For a detailed and unifying view of sparse approximate GP regression, refer to the work of Quiñonero-Candela and Rasmussen (2005).

The variational sparse GP regression framework proposed by Titsias (2008, 2009) has sparked significant interest. The proposed methodology, with time complexity that scales as $\mathcal{O}(NM^2)$, allows practitioners to circumvent the computational demands associated with inverting the required covariance matrix while also offering a formulation whereby practitioners can maximize a variational lower bound to select the inducing variable input locations and the model hyperparameters. Although the variational sparse GP regression framework was originally proposed for computational speedups, the framework has also been used as a mathematical tool to induce an analytically tractable lower bound for various state-of-the-art probabilistic models such as {1} the (B)GP-LVM (Titsias and Lawrence 2010; Damianou et al. 2016) and {2} deep Gaussian processes (Damianou and Lawrence 2013).

We adopt the variational sparse GP regression framework developed by Titsias (2008, 2009) for the following reasons: {1} We exploit the sparse framework for its original intent which is to offer computational speedups for large data sets, and {2} the sparse framework allows us to derive a variational lower bound on the log marginal likelihood of the T-GPR model in Eq. (13) which remains intractable. We will induce analytical tractability by exploiting local variational methods which result in a framework that can be used for model training and inference. Note that the variational lower bound of our proposed framework can also be used as a steppingstone to gain access to the (B)GP-LVM (see Titsias and Lawrence 2010; Damianou et al. 2016) as applied to censored observational data. It is worth pointing out that the standard GP latent variable model (Lawrence 2004), as well as the (B)GP-LVM counterpart (Titsias and Lawrence 2010), are typically applied in the context of uncensored observational data. See, for example, the applications in Urtasun et al. (2006), Lawrence (2007), Wang et al. (2008), Titsias and Lawrence (2010), Campbell and Yau (2015) and Zhang et al. (2017). However, we have yet to find any sparse GP inducing variable-based or (B)GP-LVM frameworks that explicitly incorporate the type I Tobit likelihood function to account for censoring in regression settings. The closest related literature we could find stems from the survival analysis branch of statistics and includes the work of Barrett and Coolen (2016), Saul et al. (2016), and Alaa and van der Schaar (2017). Another related approach includes the work of Lázaro-Gredilla (2012) who applied the Bayesian warped GP framework to censored data without explicitly accounting for the censoring mechanism in the likelihood function.

The synthesis of our proposed approach finds inspiration in the work of Saul et al. (2016), which itself builds on the ideas of Hensman et al. (2013) and Hensman et al. (2015). However, instead of resorting to numerical integration to address the intractability which arises from the non-Gaussian likelihood function (which is the type I Tobit likelihood function in our case), we exploit local variational methods (see Section 3.2).

In principle, the variational sparse GP regression framework developed by Titsias (2008, 2009) aims to minimize, in the Kullback-Leibler ($\mathcal{KL}$) divergence sense, the dissimilarity between the approximate posterior and exact posterior density. Within the context of the Tobit GP regression model in Eq. (13), we start by augmenting the prior density with inducing variables $\boldsymbol{u}$ such that

$$p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y}) = \frac{p_m(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u})}{p(\boldsymbol{y})} \tag{15}$$

Note that Eq. (15) is equivalent to the original T-GPR model since we can recover Eq. (13) by marginalizing out the inducing variables $\boldsymbol{u}$. However, the reason we allow for the augmented inducing variables $\boldsymbol{u}$ stems from the fact that these variables allow us to produce analytically tractable (and computationally efficient) approximations. Our goal is to minimize the $\mathcal{KL}$-divergence given by

$$\begin{aligned}&\mathcal{KL}[q(\boldsymbol{f}, \boldsymbol{u})||p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y})]\\&= \iint\limits_{\boldsymbol{u}\,\boldsymbol{f}} q(\boldsymbol{f}, \boldsymbol{u}) \ln \frac{q(\boldsymbol{f}, \boldsymbol{u})}{p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y})} d\boldsymbol{f} d\boldsymbol{u}\end{aligned} \tag{16}$$

We expand Eq. (16) by using Eq. (15) to obtain

$$\begin{aligned}\ln p(\boldsymbol{y}) &= \mathcal{KL}[q(\boldsymbol{f}, \boldsymbol{u})||p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y})]\\&\quad + \iint\limits_{\boldsymbol{u}\,\boldsymbol{f}} q(\boldsymbol{f}, \boldsymbol{u}) \ln \frac{p_m(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u})}{q(\boldsymbol{f}, \boldsymbol{u})} d\boldsymbol{f} d\boldsymbol{u}\end{aligned}$$

Next, we recall that the $\mathcal{KL}$-divergence satisfies Gibb's inequality (MacKay 2004), i.e.,

$$\mathcal{KL}[q(\boldsymbol{f}, \boldsymbol{u})||p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y})] \geq 0$$

Therefore, we conclude that

$$\ln p(\boldsymbol{y}) \geq \mathcal{F}[q(\boldsymbol{f}, \boldsymbol{u})] \tag{17}$$

The quantity $\mathcal{F}[q(\boldsymbol{f}, \boldsymbol{u})]$ is given by

$$\mathcal{F}[q(\boldsymbol{f}, \boldsymbol{u})] = \iint\limits_{\boldsymbol{u}\,\boldsymbol{f}} q(\boldsymbol{f}, \boldsymbol{u}) \ln \frac{p_m(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u})}{q(\boldsymbol{f}, \boldsymbol{u})} d\boldsymbol{f} d\boldsymbol{u} \tag{18}$$

We refer to the quantity in Eq. (18) as the variational lower bound. Other common names for this bound include the Evidence Lower BOund (ELBO), see Blei et al. (2017), or the variational free energy (MacKay 2004). Next, we note that maximizing the variational lower bound given by Eq. (18) is equivalent to minimizing the $\mathcal{KL}$-divergence in Eq. (16). Following Titsias (2009), we select the following approximating variational posterior density

$$q(\boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{f}|\boldsymbol{u})q(\boldsymbol{u}) \tag{19}$$

From Eq. (19) we see that under the selected variational approximation, the only free-form density we can optimize for is $q(\boldsymbol{u})$ since $p(\boldsymbol{f}|\boldsymbol{u})$ corresponds to the conditional GP prior density under the augmented probability model (for further details, see Titsias 2009). Furthermore, since $p(\boldsymbol{f}|\boldsymbol{u})$ does not have an explicit dependence on the data $\boldsymbol{y}$, the only way for the data $\boldsymbol{y}$ to affect $\boldsymbol{f}$ is through the inducing variables $\boldsymbol{u}$, i.e., $\boldsymbol{u}$ acts as a *summary statistic*, which is how we build sparsity into the model since $M \ll N$ (see Bui and Turner 2014). The symbol M denotes the number of user-specified inducing variables.

With Eq. (19) we can simplify Eq. (18) to obtain the following variational lower bound

$$\mathcal{F}[q(\boldsymbol{u})] = \int_{\boldsymbol{u}} q(\boldsymbol{u}) \left[\int_{\boldsymbol{f}} p(\boldsymbol{f}|\boldsymbol{u}) \ln p_m(\boldsymbol{y}|\boldsymbol{f}) d\boldsymbol{f} + \ln \frac{p(\boldsymbol{u})}{q(\boldsymbol{u})} \right] d\boldsymbol{u} \tag{20}$$

However, we note that Eq. (20) contains the following analytically intractable expectation

$$\mathbb{E}_{p(\boldsymbol{f}|\boldsymbol{u})}[\ln p_m(\boldsymbol{y}|\boldsymbol{f})] = \int_{\boldsymbol{f}} p(\boldsymbol{f}|\boldsymbol{u}) \ln p_m(\boldsymbol{y}|\boldsymbol{f}) d\boldsymbol{f} \tag{21}$$

The analytical intractability (see Eqs. (22) and (23) below) arises from the presence of the Gaussian cdf factors in the likelihood function. We note that

$$\begin{aligned} &\mathbb{E}_{p(\boldsymbol{f}|\boldsymbol{u})}[\ln p_m(\boldsymbol{y}|\boldsymbol{f})] \\ &= \int_{\boldsymbol{f}} p(\boldsymbol{f}|\boldsymbol{u}) \ln \prod_{i=1}^{N} p_m(y_i|f_i) d\boldsymbol{f} \end{aligned} \tag{22}$$

From Eqs. (11) and (22) we have that

$$\begin{aligned} &\mathbb{E}_{p(\boldsymbol{f}|\boldsymbol{u})}[\ln p_m(\boldsymbol{y}|\boldsymbol{f})] \\ &= \int_{\boldsymbol{f}_{l_b}} p(\boldsymbol{f}_{l_b}|\boldsymbol{u}) \ln \left\{ \prod_{y_i=l_b} [1 - \Phi(f_i|l_b, \sigma_y^2 + \sigma_{l_b}^2)] \right\} d\boldsymbol{f}_{l_b} \\ &\quad + \int_{\boldsymbol{f}_{u_n}} p(\boldsymbol{f}_{u_n}|\boldsymbol{u}) \ln \left\{ \prod_{l_b<y_i<u_b} \mathcal{N}(y_i|f_i, \sigma_y^2) \right\} d\boldsymbol{f}_{u_n} \\ &\quad + \int_{\boldsymbol{f}_{u_b}} p(\boldsymbol{f}_{u_b}|\boldsymbol{u}) \ln \left\{ \prod_{y_i=u_b} \Phi(f_i|u_b, \sigma_y^2 + \sigma_{u_b}^2) \right\} d\boldsymbol{f}_{u_b} \end{aligned} \tag{23}$$

Note that we used the marginalization property of the multivariate Gaussian density to arrive at Eq. (23). We denote with symbol $\boldsymbol{f}_{l_b}$ the vector of latent function values associated with the lower bound l_b censored observations. A similar argument holds for the latent function vector $\boldsymbol{f}_{u_b}$. Symbol $\boldsymbol{f}_{u_n}$ denotes the vector associated with the uncensored observations.

3.2 Local Variational Methods: Lower Bounding the Censored Variables

We circumvent the analytical intractability in Eq. (23) by implementing an alternative 'local' lower bounding strategy that shares similarities with the variational framework we have been working with. The variational inference framework we have been considering, within the context of the work of Titsias (2008, 2009), and in general, can be interpreted as a 'global' method in the sense that we directly seek an approximation to the entire posterior density over all the model random variables of interest. 'Local' variational methods provide an alternative approach and involve finding local bounds (either upper or lower) on functions over individual or groups of variables within the model (Gibbs and MacKay 2000 and Bishop 2009).

From Eq. (23) we see that the functions of interest, i.e., the functions that result in the expectation being analytically intractable, correspond to the Gaussian cdf likelihood factors. If we can construct local lower bounds for each Gaussian cdf factor present in Eq. (23), we can use the corresponding local lower bounds, in conjunction with Eq. (20), to develop a secondary variational lower bound on the log marginal likelihood, which we can use for Bayesian model training and inference about the latent function of interest. Following the approach outlined in Nickisch and Rasmussen (2008), we propose the following quadratic local lower bound on each Gaussian cdf likelihood factor in the logarithmic domain. Here we provide an example for the censored variables associated with l_b.

$$\begin{aligned} &\ln [1 - \Phi(f_i|l_b, \sigma_y^2 + \sigma_{l_b}^2)] \\ &\geq \frac{1}{\sigma_y^2 + \sigma_{l_b}^2} \left[-\frac{1}{2} f_i^2 + b_i(f_i - l_b) + c_i \right] \end{aligned} \tag{24}$$

We compute the required local likelihood lower bound parameters b_i and c_i by requiring that, at some arbitrary (and freely optimizable variational) point ζ_i, the following conditions must hold

$$\begin{aligned} &\ln [1 - \Phi(f_i|l_b, \sigma_y^2 + \sigma_{l_b}^2)] \Big|_{f_i=\zeta_i} \\ &= \frac{1}{\sigma_y^2 + \sigma_{l_b}^2} \left[-\frac{1}{2} f_i^2 + b_i(f_i - l_b) + c_i \right] \Bigg|_{f_i=\zeta_i} \\ &\frac{d}{df_i} \left(\ln [1 - \Phi(f_i|l_b, \sigma_y^2 + \sigma_{l_b}^2)] \right) \Big|_{f_i=\zeta_i} \\ &= \frac{d}{df_i} \left(\frac{1}{\sigma_y^2 + \sigma_{l_b}^2} \left[-\frac{1}{2} f_i^2 + b_i(f_i - l_b) + c_i \right] \right) \Bigg|_{f_i=\zeta_i} \end{aligned} \tag{25}$$

Using Eqs. (24) to (25) we can show that

$$\begin{aligned} &1 - \Phi(f_i|l_b, \sigma_y^2 + \sigma_{l_b}^2) \\ &\geq \exp \left\{ \frac{1}{\sigma_y^2 + \sigma_{l_b}^2} \left[-\frac{1}{2} f_i^2 + b_i(f_i - l_b) + c_i \right] \right\} \end{aligned} \tag{26}$$

$$b_i = \zeta_i - (\sigma_y^2 + \sigma_{l_b}^2) \frac{\mathcal{N}(\zeta_i|l_b, \sigma_y^2 + \sigma_{l_b}^2)}{1 - \Phi(\zeta_i|l_b, \sigma_y^2 + \sigma_{l_b}^2)} \tag{27}$$

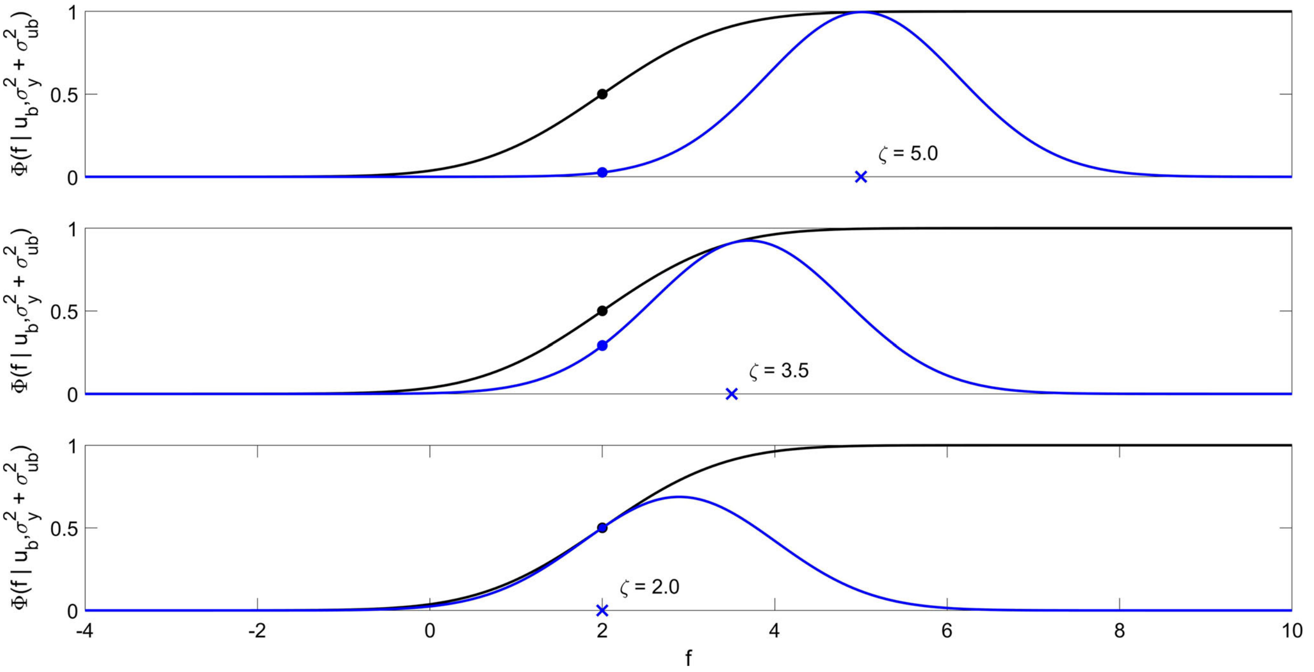

Fig. 1 The black curve depicts the Gaussian cdf factor associated with the upper detection limit u_b, viewed as a function of f, together with the local likelihood lower bound (depicted in blue, see Eqs. (29) to (30)) for various values of the freely optimizable variational parameter ζ (blue cross). The black dot represents the Gaussian cdf factor output at the latent function test point ($f_t = 2$). We see that by adjusting the parameter ζ we can control the quality of the local likelihood lower bound output (blue dot). We also note that the local lower bound output becomes tight, i.e., exact, when $\zeta = f_t$. (Color figure online)

$$
\begin{aligned}
c_i = &\frac{1}{2}\zeta_i^2 - b_i(\zeta_i - l_b)\\
&+ (\sigma_y^2 + \sigma_{l_b}^2)\ln\left[1 - \Phi(\zeta_i|l_b, \sigma_y^2 + \sigma_{l_b}^2)\right]
\end{aligned}
\tag{28}
$$

Note that a similar argument holds for the Gaussian cdf factor associated with $\Phi(f_i|u_b, \sigma_y^2 + \sigma_{u_b}^2)$. We can show that

$$
\begin{aligned}
&\Phi(f_i|u_b, \sigma_y^2 + \sigma_{u_b}^2)\\
&\quad\geq \exp\left\{\frac{1}{\sigma_y^2 + \sigma_{u_b}^2}\left[-\frac{1}{2}f_i^2 + b_i(f_i - u_b) + c_i\right]\right\}
\end{aligned}
\tag{29}
$$

$$
\begin{aligned}
b_i &= \zeta_i + (\sigma_y^2 + \sigma_{u_b}^2)\frac{\mathcal{N}(\zeta_i|u_b, \sigma_y^2 + \sigma_{u_b}^2)}{\Phi(\zeta_i|u_b, \sigma_y^2 + \sigma_{u_b}^2)}\\
c_i &= \frac{1}{2}\zeta_i^2 - b_i(\zeta_i - u_b)\\
&\quad + (\sigma_y^2 + \sigma_{u_b}^2)\ln\Phi(\zeta_i|u_b, \sigma_y^2 + \sigma_{u_b}^2)
\end{aligned}
\tag{30}
$$

Refer to Fig. 1 for an illustration of the proposed local likelihood lower bound approach as applied to the Gaussian cdf factor associated with the upper detection limit u_b (see Eqs. (29) to (30)). Observe that the local lower bound parameters b_i and c_i only depend on the freely optimizable parameter ζ_i. In other words, we can merely adjust the parameter ζ_i to improve the quality of the local lower bound. Next, we observe from Eqs. (11) and (26) to (30) that

$$
\prod_{y_i = l_b}\left[1 - \Phi(f_i|l_b, \sigma_y^2 + \sigma_{l_b}^2)\right] \geq g(\boldsymbol{f}_{l_b}|\boldsymbol{\zeta}_{l_b}, l_b, \sigma_y^2, \sigma_{l_b}^2)
\tag{31}
$$

$$
\prod_{y_i = u_b}\Phi(f_i|u_b, \sigma_y^2 + \sigma_{u_b}^2) \geq g(\boldsymbol{f}_{u_b}|\boldsymbol{\zeta}_{u_b}, u_b, \sigma_y^2, \sigma_{u_b}^2)
\tag{32}
$$

We note that

$$
\begin{aligned}
&g(\boldsymbol{f}_{l_b}|\boldsymbol{\zeta}_{l_b}, l_b, \sigma_y^2, \sigma_{l_b}^2) =\\
&\quad\exp\left\{\frac{1}{\sigma_y^2 + \sigma_{l_b}^2}\left[-\frac{1}{2}\boldsymbol{f}_{l_b}^T\boldsymbol{f}_{l_b} + \boldsymbol{b}_{l_b}^T(\boldsymbol{f}_{l_b} - l_b\mathbf{1}_{l_b}) + \boldsymbol{c}_{l_b}^T\mathbf{1}_{l_b}\right]\right\}
\end{aligned}
\tag{33}
$$

$$
\begin{aligned}
&g(\boldsymbol{f}_{u_b}|\boldsymbol{\zeta}_{u_b}, u_b, \sigma_y^2, \sigma_{u_b}^2) =\\
&\quad\exp\left\{\frac{1}{\sigma_y^2 + \sigma_{u_b}^2}\left[-\frac{1}{2}\boldsymbol{f}_{u_b}^T\boldsymbol{f}_{u_b} + \boldsymbol{b}_{u_b}^T(\boldsymbol{f}_{u_b} - u_b\mathbf{1}_{u_b}) + \boldsymbol{c}_{u_b}^T\mathbf{1}_{u_b}\right]\right\}
\end{aligned}
\tag{34}
$$

We denote with N_{l_b} the number of censored lower bound observations. The $N_{l_b} \times 1$ vectors $\boldsymbol{b}_{l_b}$ and $\boldsymbol{c}_{l_b}$ collect the element-wise entries, as calculated using Eqs. (27) and (28), for each element of the vector $\boldsymbol{f}_{l_b}$ (each of which is associated with a freely optimizable variational parameter ζ_i, which we collectively denote by the $N_{l_b} \times 1$ vector $\boldsymbol{\zeta}_{l_b}$). The symbol $\mathbf{1}_{l_b}$ denotes the $N_{l_b} \times 1$ vector of ones. A similar argument holds for $\boldsymbol{f}_{u_b}$. Next, from Eqs. (21), (31) and (32) we can show that

$$\begin{aligned} &\int_{\boldsymbol{f}} p(\boldsymbol{f}|\boldsymbol{u}) \ln p_m(\boldsymbol{y}|\boldsymbol{f}) d\boldsymbol{f} \\ &\quad \geq \int_{\boldsymbol{f}} p(\boldsymbol{f}|\boldsymbol{u}) \ln p_l(\boldsymbol{y}|\boldsymbol{f}) d\boldsymbol{f} \end{aligned} \tag{35}$$

We denote with $p_l(\boldsymbol{y}|\boldsymbol{f})$ the following

$$\begin{aligned} &p_l(\boldsymbol{y}|\boldsymbol{f}) = g(\boldsymbol{f}_{l_b}|\boldsymbol{\zeta}_{l_b}, l_b, \sigma_y^2, \sigma_{l_b}^2) \\ &\quad \times \left[\prod_{l_b < y_i < u_b} \mathcal{N}(y_i|f_i, \sigma_y^2) \right] g(\boldsymbol{f}_{u_b}|\boldsymbol{\zeta}_{u_b}, u_b, \sigma_y^2, \sigma_{u_b}^2) \end{aligned} \tag{36}$$

Observe that by our local likelihood lower bound design, we have that $\ln g(\boldsymbol{f}_{l_b}|\boldsymbol{\zeta}_{l_b}, l_b, \sigma_y^2, \sigma_{l_b}^2)$ and $\ln g(\boldsymbol{f}_{u_b}|\boldsymbol{\zeta}_{u_b}, u_b, \sigma_y^2, \sigma_{u_b}^2)$ are quadratic in the logarithmic domain. Consequently, we can analytically evaluate each Gaussian expectation on the right-hand side of the inequality in Eq. (35), circumventing the original analytical intractability that arose in Eq. (21) as a result of the presence of the Gaussian cdf likelihood factors. Using Eqs. (17) to (20) and (35) we also observe that

$$\begin{aligned} &\ln p(\boldsymbol{y}) \geq \mathcal{F}[q(\boldsymbol{u})] \\ &\quad \geq \int_{\boldsymbol{u}} q(\boldsymbol{u}) \left[\int_{\boldsymbol{f}} p(\boldsymbol{f}|\boldsymbol{u}) \ln p_l(\boldsymbol{y}|\boldsymbol{f}) d\boldsymbol{f} + \ln \frac{p(\boldsymbol{u})}{q(\boldsymbol{u})} \right] d\boldsymbol{u} \end{aligned} \tag{37}$$

From Eq. (37) we see that by lower bounding each Gaussian cdf factor we have implicitly developed a secondary variational lower bound to the original lower bound $\mathcal{F}[q(\boldsymbol{u})]$ (see Eq. (20)) stemming from the Kullback–Leibler divergence framework (which is itself a lower bound to the log marginal likelihood of the original probabilistic model). We denote our secondary variational lower bound as follows

$$\begin{aligned} &\mathcal{F}^*[q(\boldsymbol{u})] \\ &\geq \int_{\boldsymbol{u}} q(\boldsymbol{u}) \left[\int_{\boldsymbol{f}} p(\boldsymbol{f}|\boldsymbol{u}) \ln p_l(\boldsymbol{y}|\boldsymbol{f}) d\boldsymbol{f} + \ln \frac{p(\boldsymbol{u})}{q(\boldsymbol{u})} \right] d\boldsymbol{u} \end{aligned} \tag{38}$$

3.3 Deriving the Optimal $q(\boldsymbol{u})$ Density and the 'Collapsed' Lower Bound

Next, we analytically maximize our secondary lower bound in Eq. (38) and note that we have the following integral constraint

$$\int_{\boldsymbol{u}} q(\boldsymbol{u}) d\boldsymbol{u} = 1 \tag{39}$$

We construct our Lagrangian, subject to the integral constraint in Eq. (39), as follows (for more details, see Logan 2006)

$$\mathcal{L}[q(\boldsymbol{u}), \lambda] = q(\boldsymbol{u}) \left[\Psi(\boldsymbol{u}) + \ln \frac{p(\boldsymbol{u})}{q(\boldsymbol{u})} \right] + \lambda q(\boldsymbol{u}) \tag{40}$$

We denote with symbol λ the Lagrange multiplier. Furthermore, we define $\Psi(\boldsymbol{u})$ as follows

$$\Psi(\boldsymbol{u}) = \int_{\boldsymbol{f}} p(\boldsymbol{f}|\boldsymbol{u}) \ln p_l(\boldsymbol{y}|\boldsymbol{f}) d\boldsymbol{f}$$

According to the Euler–Lagrange equation, the stationary condition for the optimal density $q(\boldsymbol{u})$ satisfies

$$\frac{\partial \mathcal{L}[q(\boldsymbol{u}), \lambda]}{\partial q(\boldsymbol{u})} = 0 \tag{41}$$

From Eqs. (40) and (41) we can show that the optimal $q(\boldsymbol{u})$ corresponds to

$$q(\boldsymbol{u}) = \frac{p(\boldsymbol{u}) \exp\{\Psi(\boldsymbol{u})\}}{\int_{\boldsymbol{u}} p(\boldsymbol{u}) \exp\{\Psi(\boldsymbol{u})\} d\boldsymbol{u}} \tag{42}$$

We back-substitute Eq. (42) into Eq. (38) to derive the corresponding optimal 'collapsed' secondary lower bound as

$$\mathcal{F}^*(\boldsymbol{\theta}) = \ln \int_{\boldsymbol{u}} p(\boldsymbol{u}) \exp\{\Psi(\boldsymbol{u})\} d\boldsymbol{u} \tag{43}$$

Note that after marginalizing over the inducing variables $\boldsymbol{u}$, the resulting 'collapsed' secondary lower bound depends on the remaining model parameters, which we collectively denote by the parameter vector $\boldsymbol{\theta}$. From Eq. (42) we can analytically derive the optimal $q(\boldsymbol{u})$ and show that the density corresponds to a multivariate Gaussian parameterized by

$$\begin{aligned} q(\boldsymbol{u}) &= \mathcal{N}(\boldsymbol{u}|\boldsymbol{\mu}_u, \boldsymbol{S}_u) \\ \boldsymbol{\mu}_u &= \boldsymbol{K}_{MM} \boldsymbol{Q}^{-1} \boldsymbol{K}_{MN}^l \boldsymbol{\Sigma}_{y_l}^{-1} \boldsymbol{y}_l \\ \boldsymbol{S}_u &= \boldsymbol{K}_{MM} \boldsymbol{Q}^{-1} \boldsymbol{K}_{MM} \\ \boldsymbol{Q} &= \boldsymbol{K}_{MM} + \boldsymbol{K}_{MN}^l \boldsymbol{\Sigma}_{y_l}^{-1} \boldsymbol{K}_{NM}^l \end{aligned} \tag{44}$$

The matrix $\boldsymbol{K}_{MM}$, which stems from the augmented probability model in Eq. (15), requires evaluating the user-specified kernel function between the freely optimizable inducing input locations. Furthermore, we note that

$$\boldsymbol{y}_l = \begin{bmatrix} \boldsymbol{b}_{l_b} \\ \boldsymbol{y}_o \\ \boldsymbol{b}_{u_b} \end{bmatrix} ; \; \boldsymbol{K}^l_{NM} = \begin{bmatrix} \boldsymbol{K}_{N_{l_b}M} \\ \boldsymbol{K}_{N_{yo}M} \\ \boldsymbol{K}_{N_{u_b}M} \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{yl} = diag \begin{bmatrix} (\sigma_y^2 + \sigma_{l_b}^2)\boldsymbol{I}_{N_{l_b}N_{l_b}} \\ \sigma_y^2 \boldsymbol{I}_{N_{yo}N_{yo}} \\ (\sigma_y^2 + \sigma_{u_b}^2)\boldsymbol{I}_{N_{u_b}N_{u_b}} \end{bmatrix} \tag{45}$$

We denote with Eq. (45) the block diagonal matrix $\boldsymbol{\Sigma}_{yl}$. The symbol N_{u_b} refers to the number of censored upper bound observations. The symbol N_{y_o} denotes the number of noise-corrupted observations, collectively denoted by the vector $\boldsymbol{y}_o \in \mathbb{R}^{N_{yo}\times 1}$, that are not censored. The matrix $\boldsymbol{K}_{N_{l_b}M}$ requires evaluating the user-specified kernel function between the training input locations associated with the vector $\boldsymbol{f}_{l_b}$ and the freely optimizable inducing input locations. A similar argument holds for matrix $\boldsymbol{K}_{N_{u_b}M}$. The matrix $\boldsymbol{K}_{N_{yo}M}$ requires evaluating the kernel function between the training input locations associated with the vector $\boldsymbol{f}_{u_n}$ and the inducing input locations. We also note that $\boldsymbol{K}^l_{MN} = (\boldsymbol{K}^l_{NM})^T$. After some algebraic manipulation of Eq. (43), we arrive at the following secondary variational lower bound

$$\begin{aligned} \mathcal{F}^*(\boldsymbol{\theta}) &= \ln \left\{ \frac{|\boldsymbol{K}_{MM}|^{\frac{1}{2}}}{(2\pi)^{\frac{N_{yo}}{2}} (\sigma_y^2)^{\frac{N_{yo}}{2}} |\boldsymbol{Q}|^{\frac{1}{2}}} \exp\{\mathcal{A}_{\mathcal{F}^*}\} \right\} \\ &\quad - \frac{1}{2}\text{tr}\left\{ \boldsymbol{\Sigma}_{yl}^{-1} \left[\boldsymbol{K}^l_{NN} - \boldsymbol{K}^l_{NM} \boldsymbol{K}^{-1}_{MM} \boldsymbol{K}^l_{MN} \right] \right\} \\ \mathcal{A}_{\mathcal{F}^*} &= -\frac{1}{2}\boldsymbol{y}_l{}^T \boldsymbol{A} \boldsymbol{y}_l + \frac{1}{2}\boldsymbol{b}^T \boldsymbol{\Sigma}_c^{-1} \boldsymbol{b} \\ &\quad + \boldsymbol{c}^T \boldsymbol{\Sigma}_c^{-1} \boldsymbol{1}^* - \boldsymbol{b}^T \boldsymbol{\Sigma}_c^{-1} \boldsymbol{d} \end{aligned} \tag{46}$$

Refer to Section A.3 in Appendix A for details on the derivation of the secondary variational lower bound. Recall that $\boldsymbol{\theta}$ collectively denotes the model parameters, which include the kernel function parameters, the variance parameters from the adjusted mixed-likelihood function, the inducing variable input locations, and the local likelihood lower bound parameters. Matrices $\boldsymbol{A}$, $\boldsymbol{\Sigma}_c$ and $\boldsymbol{K}^l_{NN}$ can be computed as follows

$$\boldsymbol{A} = \boldsymbol{\Sigma}_{yl}^{-1} - \boldsymbol{\Sigma}_{yl}^{-1} \boldsymbol{K}^l_{NM} \boldsymbol{Q}^{-1} \boldsymbol{K}^l_{MN} \boldsymbol{\Sigma}_{yl}^{-1}$$

$$\boldsymbol{\Sigma}_c = diag \begin{bmatrix} (\sigma_y^2 + \sigma_{l_b}^2)\boldsymbol{I}_{N_{l_b}N_{l_b}} \\ (\sigma_y^2 + \sigma_{u_b}^2)\boldsymbol{I}_{N_{u_b}N_{u_b}} \end{bmatrix} \tag{47}$$

$$\boldsymbol{K}^l_{NN} = diag \begin{bmatrix} \boldsymbol{K}_{N_{l_b}N_{l_b}} \\ \boldsymbol{K}_{N_{yo}N_{yo}} \\ \boldsymbol{K}_{N_{u_b}N_{u_b}} \end{bmatrix} \tag{48}$$

We denote with Eqs. (47) and (48) the block diagonal matrices $\boldsymbol{\Sigma}_c$ and $\boldsymbol{K}^l_{NN}$, respectively. Vectors $\boldsymbol{b}$, $\boldsymbol{c}$, $\boldsymbol{1}^*$ and $\boldsymbol{d}$ are defined as follows

$$\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_{l_b} \\ \boldsymbol{b}_{u_b} \end{bmatrix}$$

$$\boldsymbol{c} = \begin{bmatrix} \boldsymbol{c}_{l_b} \\ \boldsymbol{c}_{u_b} \end{bmatrix}$$

$$\boldsymbol{1}^* = \begin{bmatrix} \boldsymbol{1}_{l_b} \\ \boldsymbol{1}_{u_b} \end{bmatrix}$$

$$\boldsymbol{d} = \begin{bmatrix} (l_b) \times \boldsymbol{1}_{l_b} \\ (u_b) \times \boldsymbol{1}_{u_b} \end{bmatrix}$$

Furthermore, we note that Eq. (46) is a valid secondary variational lower bound on the log marginal likelihood of the probabilistic model (see Eq. (15)) which can be maximized, using gradient-based optimization, to find point estimates for $\boldsymbol{\theta}$. This allows us to perform variational Bayesian model training and inference. We, therefore, refer to our proposed methodology as the Variational Tobit Gaussian process regression (VT-GPR) framework.

It is worth pointing out that one common criticism of $\mathcal{KL}$-divergence-based variational inference (see Eq. (16)) is its tendency to underestimate the posterior density variance (Blei et al. 2017). However, simulation-based studies performed by Titsias (2009, see Figures 1 and 2) indicate that with enough inducing/support variables, the variational sparse GP framework is able to match the standard GP model prediction results. In this regard, $\mathcal{KL}$-divergence-based variational inference does not necessarily underestimate the posterior density variance. Furthermore, when we set $M = N$ inducing variables and place them at the training input locations, i.e., $\boldsymbol{u} = \boldsymbol{f}$, the variational sparse GP framework of Titsias (2008, 2009) reduces to the standard GP regression framework (Hensman et al. 2013). However, due to the presence of censored observations, we do expect that the VT-GPR framework will display deteriorated prediction performance in censored latent function regions as a result of our proposed local variational method providing limited domain support for each Gaussian cdf factor (see Fig. 1).

Note that for a numerically stable implementation of the secondary variational lower bound, we propose following the idea outlined in Titsias (2008) which relies on the addition of "jitter" to the main diagonal elements of matrix $\boldsymbol{K}_{MM}$ to stabilize the optimization routine. Furthermore, it is also worth pointing out that in the absence of any censored observations, our proposed secondary variational lower bound reduces to the variational sparse GP lower bound derived in Titsias (2008, 2009).

4 VT-GPR Model Predictions

Model predictions about the latent function f, which we collectively denote with the latent prediction vector $\boldsymbol{f}^*$, at unsampled locations $\boldsymbol{x}^*$ are in line with the framework proposed by Titsias (2008, 2009). Starting from the joint density we have that

$$p(\boldsymbol{f}^*|\boldsymbol{y}) = \iint\limits_{\boldsymbol{u}\,\boldsymbol{f}} p(\boldsymbol{f}^*, \boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y}) d\boldsymbol{f} d\boldsymbol{u}$$
$$p(\boldsymbol{f}^*|\boldsymbol{y}) = \iint\limits_{\boldsymbol{u}\,\boldsymbol{f}} p(\boldsymbol{f}^*|\boldsymbol{f}, \boldsymbol{u}, \boldsymbol{y}) p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y}) d\boldsymbol{f} d\boldsymbol{u}$$

Given that $\boldsymbol{f}^*$ is conditionally independent of $\boldsymbol{f}$ and $\boldsymbol{y}$ given $\boldsymbol{u}$ we have that

$$p(\boldsymbol{f}^*|\boldsymbol{y}) = \iint\limits_{\boldsymbol{u}\,\boldsymbol{f}} p(\boldsymbol{f}^*|\boldsymbol{u}) p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y}) d\boldsymbol{f} d\boldsymbol{u}$$

From our variational approximation in Eq. (19), we know that

$$p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y}) \approx p(\boldsymbol{f}|\boldsymbol{u}) q(\boldsymbol{u})$$

Therefore, we have that

$$p(\boldsymbol{f}^*|\boldsymbol{y}) \approx \iint\limits_{\boldsymbol{u}\,\boldsymbol{f}} p(\boldsymbol{f}^*|\boldsymbol{u}) p(\boldsymbol{f}|\boldsymbol{u}) q(\boldsymbol{u}) d\boldsymbol{f} d\boldsymbol{u}$$
$$p(\boldsymbol{f}^*|\boldsymbol{y}) \approx q(\boldsymbol{f}^*) = \int\limits_{\boldsymbol{u}} p(\boldsymbol{f}^*|\boldsymbol{u}) q(\boldsymbol{u}) d\boldsymbol{u} \tag{49}$$

We note that

$$p(\boldsymbol{f}^*|\boldsymbol{u}) = \mathcal{N}(\boldsymbol{f}^*|\boldsymbol{K}_{N^*M}\boldsymbol{K}_{MM}^{-1}\boldsymbol{u}, \boldsymbol{\Sigma})$$
$$\boldsymbol{\Sigma} = \boldsymbol{K}_{N^*N^*} - \boldsymbol{K}_{N^*M}\boldsymbol{K}_{MM}^{-1}\boldsymbol{K}_{MN^*} \tag{50}$$

From Eqs. (44), (49) and (50) we can show that the latent function predictive density $q(\boldsymbol{f}^*)$ takes the form of a multivariate Gaussian density parameterized by

$$q(\boldsymbol{f}^*) = \mathcal{N}(\boldsymbol{f}^*|\boldsymbol{\mu}_{\boldsymbol{f}^*}, \boldsymbol{\Sigma}_{\boldsymbol{f}^*})$$
$$\boldsymbol{\mu}_{\boldsymbol{f}^*} = \boldsymbol{K}_{N^*M}\boldsymbol{Q}^{-1}\boldsymbol{K}_{MN}^{l}\boldsymbol{\Sigma}_{y_l}^{-1}\boldsymbol{y}_l$$
$$\boldsymbol{\Sigma}_{\boldsymbol{f}^*} = \boldsymbol{K}_{N^*N^*} - \boldsymbol{K}_{N^*M}\boldsymbol{K}_{MM}^{-1}\boldsymbol{K}_{MN^*} + \boldsymbol{K}_{N^*M}\boldsymbol{Q}^{-1}\boldsymbol{K}_{MN^*}$$

5 Experiments

To demonstrate the VT-GPR framework, we now consider its application to two synthetic examples and a real-world data set. For each synthetic example, we generate noise-corrupted observational data, which is then subjected to artificial censoring. Furthermore, throughout all our experiments we use the exponentiated quadratic kernel function (see Eq. (51)) with signal variance σ_f^2 and length scale l.

$$k(x_i, x_j) = \sigma_f^2 \exp\left\{\frac{-(x_j - x_i)^2}{2l^2}\right\} \tag{51}$$

5.1 Synthetic Data: Example 1

In our first experiment, we reproduce the artificial example outlined in the work of Groot and Lucas (2012). They created a data set consisting of 30 equally spaced inputs on the interval [0, 1] and generated latent function outputs from

$$f(x) = (6x - 2)^2 \sin(2(6x - 2)) \tag{52}$$

The data is then artificially contaminated by adding zero mean Gaussian distributed noise with variance $\sigma_y^2 = 0.1$. Groot and Lucas (2012) censor 40% of the observations by calculating the 40th percentile of the data and use the corresponding value as the lower detection limit l_b. We repeat this procedure for a randomly generated data set, using the available implementation of Groot and Lucas (2012).

We then train the T-GPR model on the censored data set using {1} expectation propagation (EP) (Groot and Lucas 2012), {2} the Laplace approximation (LA) (Ertin 2007) and {3} our proposed VT-GPR framework using gradient-based optimization. We used an in-house implementation of {1}, and {2} we use the implementation in the publicly available GPstuff MATLAB toolbox (Vanhatalo et al. 2013). For the VT-GPR training procedure we fixed (i.e., implicitly assumed) 15 as the number of inducing variables (see Section B.1 in Appendix B for more details).

To compare the T-GPR frameworks, we simulated 1000 additional independent data sets and trained the LA and EP-based T-GPR frameworks, as well as our VT-GPR framework, on each data set. For all data sets, we calculate the 40th percentile and use the corresponding value as the lower detection limit l_b. We report {1} the root mean squared error (RMSE, Eq. (53)), {2} the mean absolute error (MAE, Eq. (54)), and {3} the mean negative log-loss (MNLL, Eq. (55)) to compare the model predictions from the various T-GPR frameworks to the true latent function. For all three criteria, smaller values imply better performance. We define the error measures as follows (see Rasmussen and Williams 2006; Lázaro-Gredilla et al. 2010; Groot and Lucas 2012):

$$\text{RMSE}(f, \boldsymbol{\mu}_{\boldsymbol{f}^*}) = \sqrt{\frac{1}{N^*}\sum_{i=1}^{N^*}\left(f_i - \left(\boldsymbol{\mu}_{\boldsymbol{f}^*}\right)_i\right)^2} \tag{53}$$

$$\text{MAE}(f, \boldsymbol{\mu}_{\boldsymbol{f}^*}) = \frac{1}{N^*}\sum_{i=1}^{N^*}\left|f_i - \left(\boldsymbol{\mu}_{\boldsymbol{f}^*}\right)_i\right| \tag{54}$$

$$\text{MNLL}(f, \boldsymbol{\mu}_{f^*}) = \frac{1}{N^*}\sum_{i=1}^{N^*}\left[\frac{1}{2}\ln\left(2\pi\sigma_i^2\right) + \frac{\left(f_i - \left(\boldsymbol{\mu}_{f^*}\right)_i\right)^2}{2\sigma_i^2}\right] \quad (55)$$

The symbol N^* denotes the total number of predicted latent function values. We denote with symbol $f_i = f(x_i)$ the true underlying latent function value (see Eq. (52)), as indexed by x_i, whereas the vector $\boldsymbol{\mu}_{f^*}$ denotes the mean model prediction. The symbol σ_i^2 denotes the marginal predictive variance associated with $\left(\boldsymbol{\mu}_{f^*}\right)_i$.

To illustrate the scalability of the proposed VT-GPR framework, we carry out a further experiment by training {1} the standard GP, {2} the LA-based, {3} the EP-based, and {4} the VT-GPR frameworks on increasingly larger data sets, holding fixed the other aspects of the example described above. For each data set and proposed framework, we initiate 10 randomly generated parameter starting points and then calculate the average run time per starting point. This procedure was repeated 10 times for each data set size, followed by averaging across the average computational run times.

Figure 2 shows the results from the reproduced example in Groot and Lucas (2012), where we censored based on a lower bound of $l_b = -0.1185$. Qualitatively, from Fig. 2, we observe that all three T-GPR frameworks have the ability to learn an underlying representation that is consistent with the original latent function given by Eq. (52) from the censored data set. The interested reader is referred to the work of Groot and Lucas (2012, Figure 2) for comparisons of the LA and EP-based T-GPR frameworks against the standard GPR model (see Section 2) when the censored data are either treated as missing values or as uncensored observations exactly equal to the detection limit.

In Fig. 3 we depict and compare the distribution of the generated RMSE, MAE and MNLL results across the 1000 data sets using box plots. The additional dashed red line in Fig. 3 depicts the mean value of the generated results. Qualitatively, for the RMSE (left panel) and MAE (middle panel) results in Fig. 3, we observe that there is no significant difference in the predictive performance results across the T-GPR frameworks. Arguably, we can state that the Laplace-based T-GPR framework marginally outperforms the EP-based and VT-GPR frameworks. However, when we consider the MNLL performance measure results (right panel) we observe that the proposed VT-GPR framework performs worse when compared to the LA and EP-based frameworks.

To understand the discrepancy in the predictive performance behaviour we observe in Fig. 3, we stratify the error measures, relative to the underlying latent function values calculated from Eq. (52), by the lower detection limit l_b such that

$$\text{MSE}(f, \boldsymbol{\mu}_{f^*}) = \frac{1}{N^*}\sum_{f(x)\leq l_b}\left(f_i - \left(\boldsymbol{\mu}_{f^*}\right)_i\right)^2 + \frac{1}{N^*}\sum_{f(x)>l_b}\left(f_i - \left(\boldsymbol{\mu}_{f^*}\right)_i\right)^2$$

Note that instead of using the RMSE we opted for the MSE (mean squared error) for convenience. This stratification procedure partitions each error measure into two different contributing error components. The first stratified error component considers the predictive performance in lower bound censored latent function regions, whereas the second error component considers predictive performance in uncensored latent function regions. The MAE and MNLL stratified error measures are constructed in a similar fashion. The bold values in Tables 1, 2, and, 3 indicate the best-performing framework, for each performance measure, for the example under consideration.

Table 1 summarises the mean error component contributions for each of the stratified error measures across the 1000 additional independently generated data sets. We observe that, in the lower bound censored latent function regions, the quantitative performance measure contributions for the VT-GPR framework are, on average, larger when compared to the LA and EP-based frameworks, indicating worse predictive performance. The deteriorated performance is especially noticeable from the MNLL performance measure. We suspect that the worse performance is a result of the single regulating variance parameter ($\sigma_{l_b}^2$) that we introduced in the adjusted mixed-likelihood (refer to Appendix B Section B.2 for more details).

However, in the uncensored latent function regions, the VT-GPR framework provides predictive performance results that are, on average, comparable to the LA and EP-based frameworks. Table 1 also highlights that, on average, for the example under consideration, the Laplace-based T-GPR framework outperforms the EP-based and VT-GPR frameworks.

Turning our attention to the scalability results, in Fig. 4 we depict and compare the average computational run time for the various frameworks, across the 10 repeated experiments, together with error bars corresponding to three standard deviations. From Fig. 4 we see a clear separation, i.e., no overlapping error bars, between the computational run times for the standard GP and the VT-GPR framework at a data set size of approximately 3500 points. Thus, at a data set size of approximately 3500 data points, the VT-GPR framework becomes computationally less demanding and starts outperforming the standard GP model. We also observe that after approximately 1500 data points, the VT-GPR framework

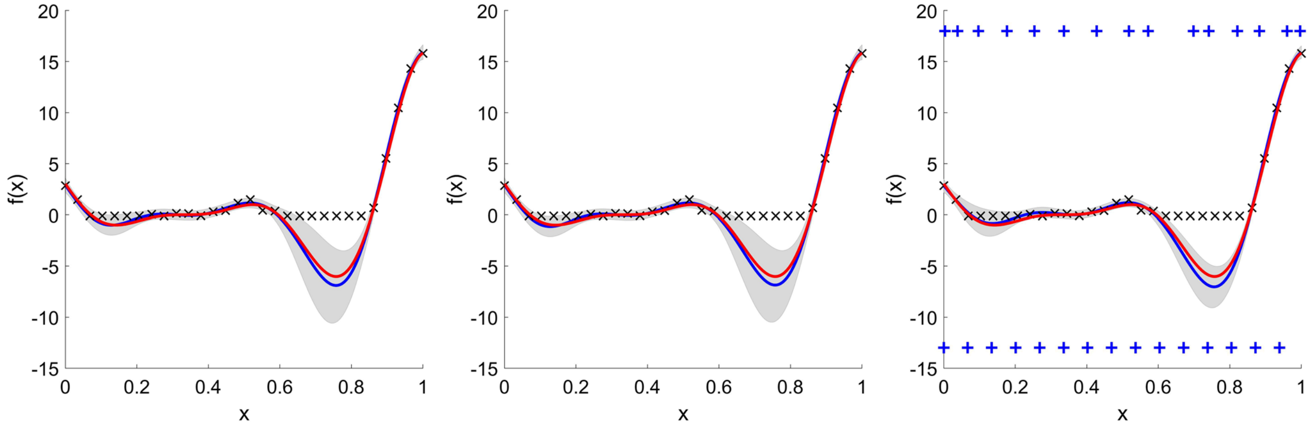

Fig. 2 Tobit Gaussian process regression results with l_b set to the 40th percentile of the uncensored observational data. Left Panel: T-GPR latent function predictive results using the Laplace approximation. Middle Panel: T-GPR latent function predictive results using the expectation propagation framework. Right Panel: VT-GPR latent function predictive results. Additional Information: The black '×'-sign denotes the observational data (noisy and/or censored), the red line denotes the underlying latent function (see Eq. (52)) while the blue curve denotes the mean model prediction (model MAP estimate). The corresponding grey shaded area depicts the 99% point-wise credibility interval. Furthermore, the blue '+'-sign at the bottom of the right panel depicts the initial inducing input locations, which we initialized to evenly spaced input points across the function domain, while the optimized inducing input locations are depicted at the top of the right panel. We arbitrarily selected 15 as the number of inducing variables for our VT-GPR implementation (see Section B.1 in Appendix B for more details). (Color figure online)

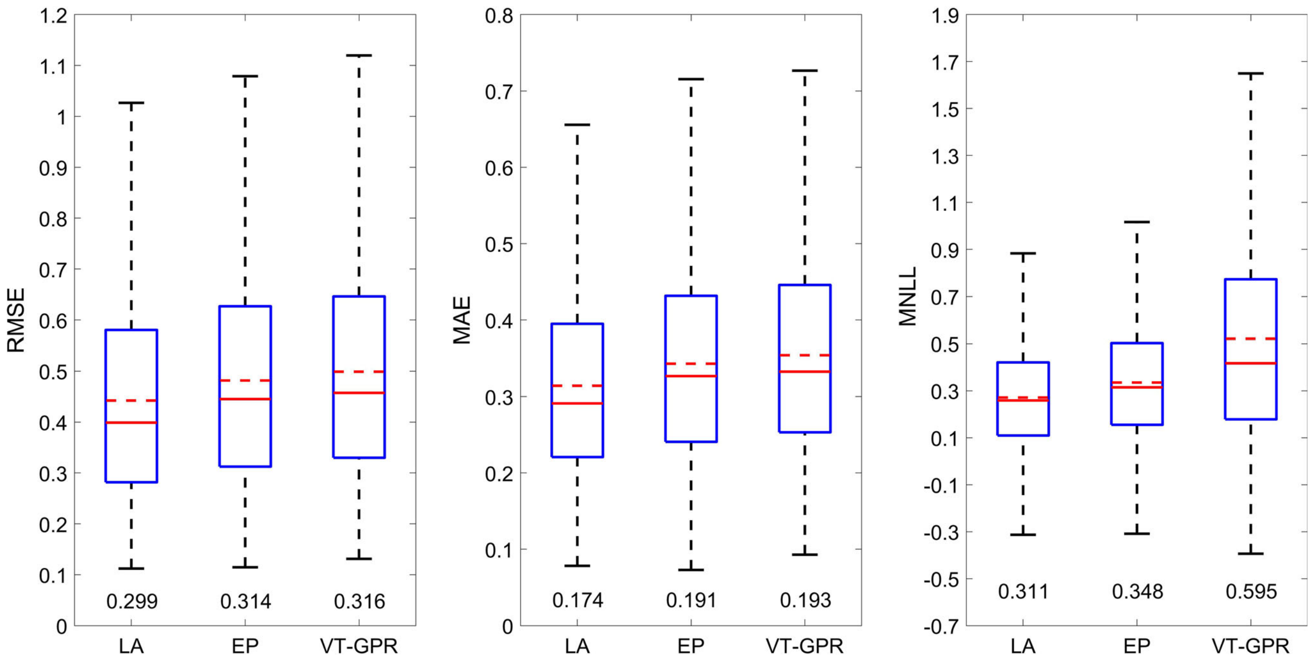

Fig. 3 Box plot visualization for the generated RMSE (left panel), MAE (middle panel) and MNLL (right panel) results, respectively, for each T-GPR framework. The dashed red line depicts the mean value for each quantitative performance measure across the 1000 additional independently generated data sets. The interquartile range is denoted at the bottom whisker of each box plot . (Color figure online)

Table 1 Summary of the mean contributions to each quantitative performance measure for the various T-GPR frameworks for Example 1

	Tobit GP (LA)	Tobit GP (EP)	VT-GPR
	Stratify by: $f(x) \le l_b$		
Mean MSE	**0.2062**	0.2418	0.2603
Mean MAE	**0.2096**	0.2347	0.2458
Mean MNLL	**0.3267**	0.3694	0.5162
	Stratify by: $f(x) > l_b$		
Mean MSE	**0.0312**	0.0339	0.0338
Mean MAE	**0.1047**	0.1083	0.1084
Mean MNLL	**−0.0558**	− 0.0345	0.0041

starts to computationally outperform the LA and EP-based approaches.

5.2 Synthetic Data: Example 2

For our second experiment, we expand the example outlined in the work of Groot and Lucas (2012) by introducing an upper detection limit based on the 90th percentile of the uncensored observational data, thereby increasing the percentage of censored observations from 40% (Example 1) to 50%.

We create a data set consisting of 30 equally spaced inputs on the interval [0, 1.15] and generate latent function outputs from Eq. (52). Following this, we artificially contaminate the latent function outputs by adding zero mean Gaussian distributed noise with variance $\sigma_y^2 = 0.1$. We censor the observations by calculating a lower detection limit l_b equal to the 40th percentile of the data and the upper detection limit u_b is calculated from the 90th percentile of the uncensored observational data.

Plots of the latent function predictive results from the single simulation, as well as box plots comparing the performance metrics across 1000 independently generated data sets, for the various T-GPR frameworks, are shown in Figs. 11 and 12, respectively, in Section B.3 of Appendix B. Note that, similar to Example 1, we fixed the number of inducing variables to 15 for all the VT-GPR training procedures. As in Example 1, all the T-GPR frameworks have the ability to learn a good underlying representation of the latent function from the censored data. However, the LA and EP-based T-GPR frameworks produce more conservative, i.e., larger, credibility intervals compared to our proposed VT-GPR framework. Contrasting the results from Example 1, the VT-GPR framework marginally outperforms the LA and EP-based frameworks in terms of RMSE and MAE; however, we again see that the VT-GPR framework performs worse on the MNLL.

As before, we stratify the error measures, relative to the underlying latent function values calculated from Eq. (52), by the lower detection limit l_b and the upper detection limit u_b. Refer to Table 2 for a summary the mean error component contributions for each of the stratified error measures across the 1000 additional independently generated data sets.

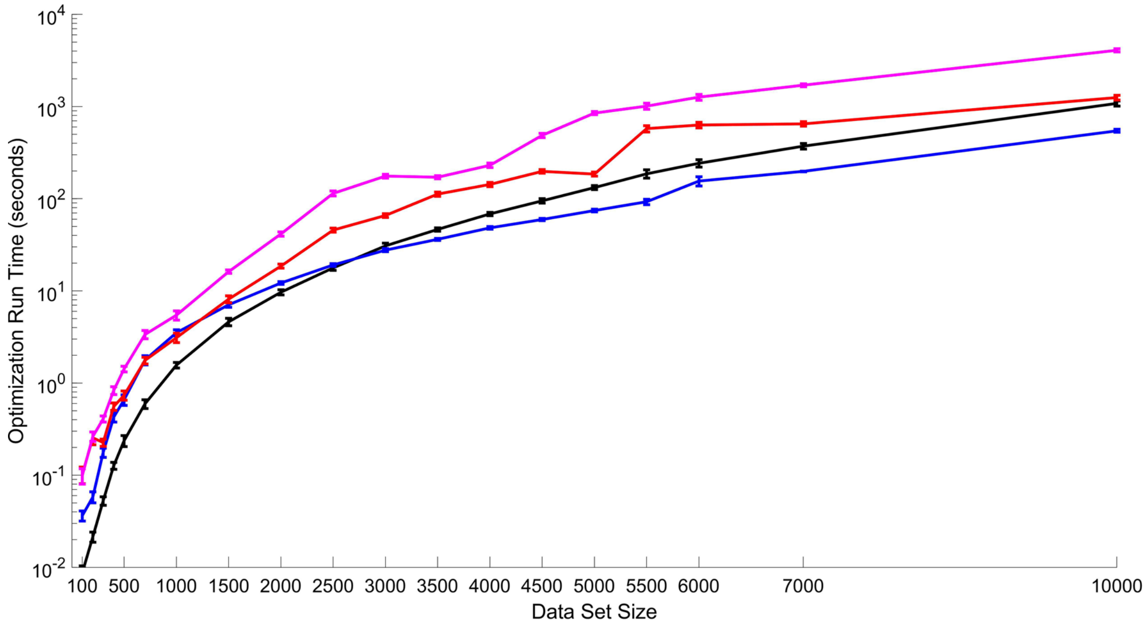

Fig. 4 Average optimization run time for {1} the standard GP model (black), {2} the LA-based T-GPR framework (red), {3} the EP-based approach (magenta), and {4} the VT-GPR framework (blue). The various curves depict and compare the average log (base 10) computational run time for the different frameworks, across the 10 repeated experiments, together with three standard deviation error bars. (Color figure online)

Table 2 Summary of the mean contributions to each quantitative performance measure for the various T-GPR frameworks for Example 2

	Tobit GP (LA)	Tobit GP (EP)	VT-GPR
	Stratify by: $f(x) \leq l_b$		
Mean MSE	0.4758	0.5530	**0.4224**
Mean MAE	0.3271	0.3434	**0.3071**
Mean MNLL	**0.5323**	0.5781	0.5976
	Stratify by: $l_b < f(x) < u_b$		
Mean MSE	**0.0470**	0.0641	0.0509
Mean MAE	**0.1154**	0.1278	0.1167
Mean MNLL	**0.0596**	0.1236	0.1846
	Stratify by: $f(x) \geq u_b$		
Mean MSE	0.0643	0.0742	**0.0557**
Mean MAE	0.0668	0.0732	**0.0617**
Mean MNLL	**0.1124**	0.1232	0.1366

From Table 2 we observe that, on average, for the example under consideration, the VT-GPR framework seems to provide slightly better mean model prediction results (see the mean MSE and MAE) in the censored latent function regions and comparable results in the uncensored latent function regions. However, similar to Example 1, we observe that the MNLL error measure contributions for the VT-GPR framework are larger when compared to the LA and EP-based frameworks, indicating worse predictive performance. This arises due to the less conservative credibility intervals produced by the proposed VT-GPR framework relative to the competing benchmarks.

5.3 Real-World Data: Example 3

Our third experiment focuses on the application of the T-GPR frameworks on a real-world data set. We sourced an electrical conductivity (EC) data set, for the Vaal River at Groot Vadersbosch/Buffelsfontein, from the Department of Water and Sanitation, South Africa (DWS 2019). The electrical conductivity of water is a measure of its ability to conduct electrical current and is affected by the presence of positively and negatively charged ions from dissolved salts and other chemicals.

Water bodies, like the Vaal River, tend to have a constant EC range. Once the EC range has been established, it can be used as a baseline for comparison with future EC measurements. If we observe significant changes in the electrical conductivity, relative to the baseline, it can be an indicator that some source of pollution has entered the water body. Thus, we can think of EC as a useful measure of water quality where, generally speaking, lower EC values indicate better water quality (EPA 2022).

The Vaal River EC data, measured in milli-siemens per meter (mS/m), was collected between 03 January 1984 and 11 July 1997, with a manual EC sample taken from Monday to Friday (excluding holidays). To perform our regression analysis, we convert the EC measurement dates into serial numbers, where, by default, 01 January 1900 corresponds to serial number 1. We then subtract the serial number associated with 03 January 1984 such that the first EC entry in the Vaal River data set corresponds to taking an EC measurement on day 0 (our reference time stamp).

We create our data set by extracting the last 150 EC measurements and the corresponding time stamps. Next, we calculate the 10th and 90th percentile of the 150 data points and use the calculated values as the artificial lower detection limit l_b and upper detection limit u_b, respectively. For the 150 EC data points, we find that $l_b = 26.5$ mS/m and $u_b = 43.5$ mS/m. From an implementation perspective, recall that we used the publicly available GPstuff MATLAB toolbox to train the LA-based T-GPR framework. We selected the 150 EC measurements as this resulted in a numerically stable implementation for each of the T-GPR frameworks. Furthermore, the 150 measurements also capture enough interesting latent function behaviour to provide a fair predictive performance comparison.

Next, we train the following regression frameworks on the artificially censored EC data: $\{1\}$ the standard Gaussian process regression (GPR) model, $\{2\}$ the standard GPR model with the censored observations treated as missing (i.e., removing the censored data), $\{3\}$ the LA-based T-GPR framework, $\{4\}$ the EP-based approach, and $\{5\}$ the VT-GPR framework. The latent function predictive results are shown in Fig. 5.

From Fig. 5 we qualitatively observe that the T-GPR frameworks (c, d, and e) produce fairly similar prediction results. Interestingly enough, the standard GPR model (a) produces prediction results that are in line with the results obtained from the various T-GPR frameworks in the uncensored latent function regions.

However, the standard model appears to directly interpolate the censored observations (note how the MAP estimate peaks at the censored limits, also see panel (f) for this behaviour) in the censored latent function regions. This behaviour arises due to the censored observational data forming part of the observation vector $\boldsymbol{y}$ (see Section 2) which is a direct consequence of the standard GPR model's inability to account for the data censoring mechanism.

The standard missing data GP regression model (b) also produces prediction results that are quite consistent with the various T-GPR frameworks but, contrasting the previous GP model depicted in (a), directly interpolates the missing data values. We also observe that the interpolating behaviour is accompanied by more conservative, i.e., larger, point-wise credibility intervals indicating that the model is less confident

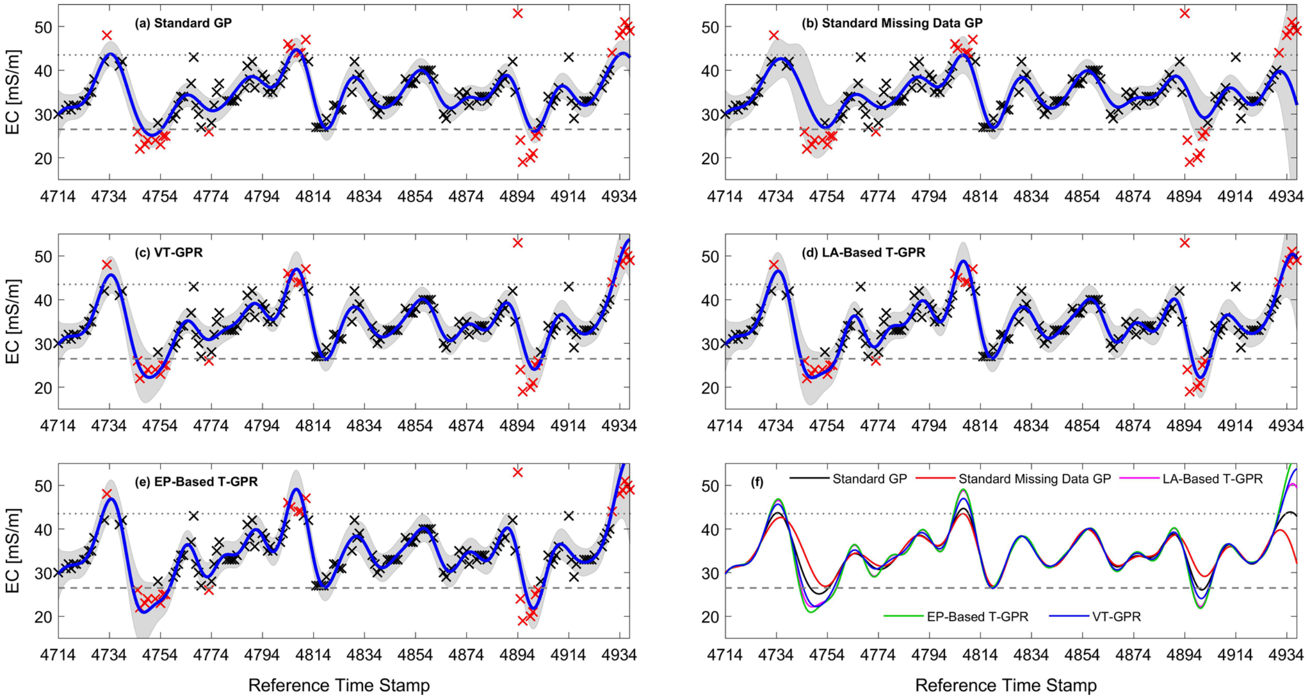

Fig. 5 Gaussian process regression results with l_b (grey dashed line) and u_b (grey dotted line) set to the 10th and 90th percentile of the EC observational data, respectively. For panels (**a**) to (**e**) the blue curve denotes the mean model prediction (model MAP estimate) whereas the grey shaded area depicts the 99% point-wise credibility interval. The black '×'-sign denotes the noisy uncensored EC data whereas the red '×'-sign denotes the noisy uncensored EC data that is artificially censored during the model training procedures. Note that for panel (**b**) the artificially censored observations are treated as missing values. Panel (**f**) depicts the mean model prediction for the various regression frameworks shown in panels (**a**) to (**e**). We arbitrarily selected 35 as the number of inducing variables for our VT-GPR implementation. (Color figure online)

Table 3 Summary of the mean quantitative performance measures for the various Gaussian process regression frameworks for Example 3

	RMSE	MAE	MNLL
Standard GP Regression	3.9479 ±0.0013	2.3243 ±0.0017	7.6342 ±0.0326
Missing Data GP Regression	5.0758 ±0.0034	2.9115 ±0.0023	4.7436 ±0.0139
Tobit GP (LA)	**3.3753** ±0.0007	**1.8683** ±0.0010	4.1794 ±0.0025
Tobit GP (EP)	3.5169 ±0.0062	1.9598 ±0.0078	**4.1218** ±0.0038
VT-GPR	3.4454 ±0.0076	1.9226 ±0.0046	4.2958 ±0.0349

about the behaviour of the underlying latent function in the missing data regions. For a visual comparison of the mean model estimates across the various GPR frameworks, refer to panel (f) in Fig. 5.

Table 3 reports the mean quantitative performance measures for each of the 5 regression frameworks, together with three standard deviation error bars, obtained from 10 independent training runs where each training run was optimized across 1000 randomly generated parameter starting points. From an error measure perspective, we observe that the T-GPR frameworks outperform the standard GPR models. We also observe that the VT-GPR MNLL is higher when compared to the LA and EP-based frameworks.

This, again, emphasizes that the proposed VT-GPR framework produces less conservative, i.e., smaller, credibility intervals relative to the competing LA and EP-based benchmarks. However, when we consider the mean model prediction results (see panel (f), the RMSE, and the MAE), we observe that the T-GPR frameworks produce quite comparable performance results with the LA-based framework slightly outperforming the EP-based and VT-GPR frameworks.

6 Discussion and Limitations

In this article, we introduced a variational inference-based framework for training a GP regression model subject to censored data. Our proposed framework relies on the variational sparse GP inducing variable framework and local variational methods which allow us to variationally integrate out the latent function values associated with the Gaussian cdf factors (which would otherwise be analytically intractable). We demonstrated the proposed VT-GPR framework on synthetically produced, as well as a real-world, data set subject to artificial censoring and found that the framework can produce results comparable to other methodologies presented in the literature. However, the proposed VT-GPR framework computationally outperforms the standard GPR model, as well as the competing benchmarks, for larger data sets (refer to Fig. 4).

Furthermore, the proposed framework can also be used as a mathematical tool to gain access to the Tobit-based (B)GP-LVM with uncertain model inputs, i.e., x in our framework, if we restrict the uncertain inputs to have Gaussian prior densities (Titsias and Lawrence 2010; Damianou et al. 2016). This would allow us to extend the (B)GP-LVM to the censored data regime which can prove very useful in many real-world applications where practitioners typically collect noise-corrupted observations for the model inputs and outputs (where the output data can be subjected to some censoring mechanism).

Note that the VT-GPR framework's biggest limitation arises from the proposed 'local' lower bounding strategy we introduced in Section 3.2. Recall that the local likelihood lower bound parameters b_i and c_i can be expressed in terms of a single freely optimizable parameter ζ_i. However, due to constraints imposed by asymptotics, the parameter a_i is restricted to the value $(-\frac{1}{2})$. Consequently, the local lower bound is unable to adjust its width (i.e., the function domain support) since the parameter a_i is fixed and not a function of ζ_i. This directly influences the VT-GPR framework's approximation performance (Nickisch and Rasmussen 2008). Despite the introduction of the additional variance parameters in an attempt to regulate the local lower bound support, the VT-GPR framework still tends to underestimate the predictive variance, relative to the competing frameworks, in the censored latent function regions (see the MNLL error measure scores for the various illustrative examples).

Furthermore, due to the imposed local lower bounding strategy and the limitation associated with parameter a_i, the optimal, and only free-form variational density, $q(\boldsymbol{u})$ must obey certain restrictions, i.e., the optimal density $q(\boldsymbol{u})$ must obey the restrictions associated with each local lower bound to ensure that we have a valid secondary variational lower bound, which can result in worse approximation/prediction performance (Nickisch and Rasmussen 2008).

Another limiting feature worth pointing out is the tightness of the secondary lower bound. Recall that we are free to choose the parameters ζ_i, which we do by finding the values of ζ_i that maximize our lower bound. The resulting secondary variational lower bound value then represents the tightest bound within the entire family of bounds that can be used as an approximation to $\ln p(\boldsymbol{y})$. However, the optimized bound will in general not be exact. Despite being able to exactly optimize the local lower bound for each Gaussian cdf factor, the required value for ζ_i depends on the value of f_i. Therefore, the local lower bound is tight for only one value of f_i (refer to Fig. 1). However, note that the quantity $\mathcal{F}^*(\boldsymbol{\theta})$ is obtained by integrating over all possible values of the latent vector $\boldsymbol{f}$, followed by integrating over all possible values of the inducing variable vector $\boldsymbol{u}$. Consequently, the values of ζ_i that maximize our secondary variational lower bound represent a compromise, as weighted by the variational posterior densities $p(\boldsymbol{f}|\boldsymbol{u})$ and $q(\boldsymbol{u})$, which directly influences the predictive performance of the proposed VT-GPR framework (Bishop 2009).

For future work, we would like to explore the idea of allowing each censored observation to have its own unique regulating variance parameter which, from a theoretical standpoint, should increase the VT-GPR's regulating capacity, resulting in improved approximation and prediction performance. This would be in line with the ideas initially proposed by Gammelli et al. (2020b).

Furthermore, since the values of ζ_i represent a compromise, as weighted by the posterior densities, we can consider using some form of regularizer to encourage better point estimates for ζ_i in an attempt to improve the approximation performance. Alternatively, we could dispense with the local lower bound approach introduced in Section 3.2 and follow the ideas outlined in Hensman et al. (2015, Section 4) where we use numerical integration to circumvent the intractabilities that arise from the presence of the Gaussian cdf terms in the likelihood function.

Acknowledgements This work was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) (EP/P028403/1) and the School of Data Science and Computational Thinking (Stellenbosch University).

Author Contributions Conceptualization, Methodology, Formal analysis and investigation, Writing - original draft preparation: Marno Basson; Writing - review and editing, Supervision: Tobias M. Louw and Theresa R. Smith; Funding acquisition: Tobias M. Louw.

Funding Open access funding provided by Stellenbosch University. This work was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) (EP/P028403/1) and the School of Data Science and Computational Thinking (Stellenbosch University).

Data Availability No novel data was produced during this study. The electrical conductivity (EC) data set, for the Vaal River at Groot Vadersbosch/Buffelsfontein, was obtained from the Department of Water and Sanitation, South Africa.

Code Availability No code has been made publicly available.

Declarations

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval Not Applicable.

Consent to participate Not Applicable.

Consent for publication All authors agreed with the content of this article and have given explicit consent for submission/publication.

Appendix A

In this Appendix, we provide supplementary information such that the reader can gain further insights into the proposed VT-GPR framework. We also provide additional mathematical steps to aid in the derivation of the VT-GPR lower bound.

A.1 Heteroskedasticity and Censored Regression Models

The primary motivation for the ideas discussed below can be found in the preprint of Gammelli et al. (2020b). Refer to Fig. 6 and Eq. (9) in the article.

When we, from a maximum likelihood perspective, consider the candidate mean values the left panel in Fig. 6 shows that all candidate mean values parameterize a Gaussian density that fits the uncensored observation, i.e., the observed value can be a potential sample from each Gaussian density. However, out of the 3 candidate mean values, the candidate mean value associated with the red Gaussian density is less likely to be a contender.

Refer to the right panel in Fig. 6. We now assume that the observed value (black cross) has been upper bound censored (magenta cross). We see that, from a maximum likelihood perspective, we favour the candidate mean value associated with the red Gaussian cdf. This, in turn, implies that we favour the red Gaussian density (left panel) as the most likely contender from which the observation was generated before censoring. However, based on our previous argument we know we are selecting the candidate mean value that corresponds to a Gaussian density that is less likely to have generated the uncensored observation (see the left panel). From this perspective, we can argue that we are overestimating the candidate mean value which translates to overestimating the latent function value of interest.

Since Eq. (9) in the article depends on a constant noise variance parameter, we would enforce the same amount of overestimation for all the observations (a similar observation is also made by Gammelli et al. 2020b). Gammelli et al. (2020b) suggest bypassing the overestimation phenomena by allowing the Tobit-based likelihood function to account for input-dependent noise, i.e., allowing for a heteroskedastic observation model. Refer to Fig. 7.

When we consider the right panel in Fig. 7, we observe that, from a maximum likelihood perspective, there are 3 candidate mean values that give rise to the same likelihood contribution (black dot, right panel) for the upper bound censored observation. These 3 candidate mean values correspond to the yellow, red, and magenta Gaussian densities (left panel), respectively.

Note that for the magenta Gaussian cdf to contribute the same likelihood value as the red Gaussian cdf, we require the magenta Gaussian density with higher variance (left panel) to have a higher candidate mean value. Similarly, for the yellow Gaussian cdf, we require the yellow Gaussian density with a smaller variance (left panel) to have a smaller candidate mean value.

Hence, we observe that the variance parameter directly controls the slope of the Gaussian cdf. In other words, if we regulate/adjust the variance parameter of the Gaussian cdf, the model can automatically tune the amount of overestimation associated with the candidate mean values. This translates to automatically tuning the amount of overestimation associated with the latent function value of interest resulting in better predictive performance. It is this observation that gave rise to the heteroskedastic-based variance parameterization that was suggested and implemented by Gammelli et al. (2020b). Note that a similar argument holds when we consider lower bound censored observations.

A.2 The Adjusted Mixed-Likelihood

Here we provide a qualitative understanding for our use of the adjusted mixed-likelihood which extends beyond the heteroskedastic-based motivation outlined above. At the end of Section 2, as well as in Section 3 of the article, we show that exact inference for the Tobit-based Gaussian process regression model is not possible due to the presence of the Gaussian cdf terms that arise as a result of censoring. Consequently, we need to resort to approximate inference techniques. In Section 3.1 of the article, we introduce and motivate the use of the variational sparse GP framework. However, the resulting variational lower bound which we obtain as a consequence of using the variational sparse GP framework remains intractable due to the presence of the Gaussian cdf terms. More specifically, the intractability arises due to the inability

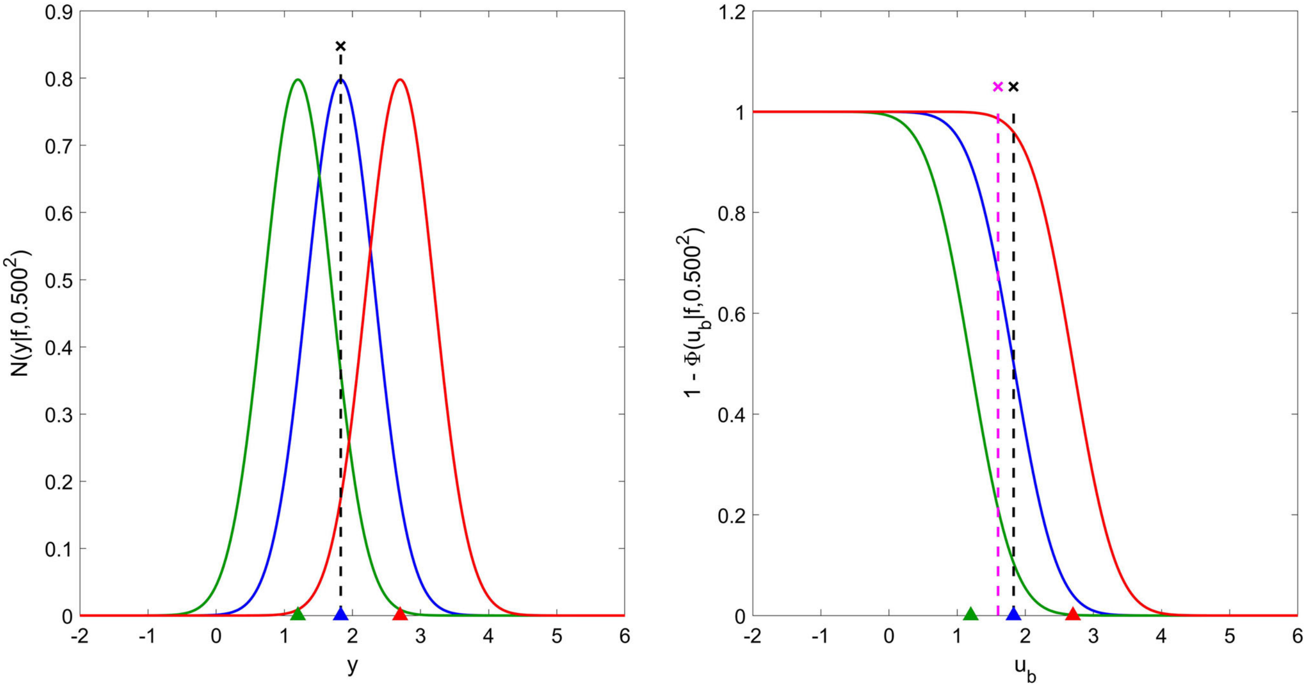

Fig. 6 Likelihood contribution, relative to the dashed lines, for an uncensored observation (left panel) and an upper bound censored observation (right panel) viewed as a function of u_b. The black cross denotes the assumed position for the uncensored observation and the magenta cross denotes the assumed position for the upper bound censored observation associated with the uncensored observation (black cross). The triangles correspond to 3 candidate mean values for f (latent function value of interest). Note that the Gaussian density variance associated with all candidate mean values have been fixed to 0.5^2. Figure 6 has been reproduced and adjusted from the preprint of Gammelli et al. (2020b). (Color figure online)

to analytically calculate the expectation of the log Gaussian cdf factor with respect to a Gaussian density function.
In Section 3.2 we induce tractability by exploiting 'local' variational methods (Jordan et al. 1999; Gibbs and MacKay 2000; Bishop 2009). The local variational-based approach allows us to locally lower bound each log Gaussian cdf factor with a quadratic function (Nickisch and Rasmussen 2008). This, in turn, allows us to lower bound the analytically intractable expectation with an expectation of a quadratic function under a Gaussian density (which is analytically tractable). However, within the context of our approach, we do pay a price when we use the proposed 'local' variational-based method.

Based on Eq. (9), and without loss of generality, we present the quadratic function that locally lower bounds the log Gaussian cdf factor for an upper bound censored observation (refer to Eqs. (56) to (57)).

$$\begin{aligned}
&\ln \Phi(f_i|u_b, \sigma_y^2) \\
&\quad \geq \frac{1}{\sigma_y^2}\left[a_i f_i^2 + b_i(f_i - u_b) + c_i\right] \\
&\therefore \Phi(f_i|u_b, \sigma_y^2) \\
&\quad \geq \exp\left\{\frac{1}{\sigma_y^2}\left[a_i f_i^2 + b_i(f_i - u_b) + c_i\right]\right\} \qquad (56)
\end{aligned}$$

$$\begin{aligned}
&\text{Asymptotic behaviour} \Rightarrow a_i = -\frac{1}{2} \\
&b_i = \zeta_i + (\sigma_y^2)\frac{\mathcal{N}(\zeta_i|u_b, \sigma_y^2)}{\Phi(\zeta_i|u_b, \sigma_y^2)} \\
&c_i = \frac{1}{2}\zeta_i^2 - b_i(\zeta_i - u_b) \\
&\quad +(\sigma_y^2)\ln \Phi(\zeta_i|u_b, \sigma_y^2) \qquad (57)
\end{aligned}$$

We observe that the local lower bound parameters b_i and c_i can be expressed in terms of a single freely optimizable variational parameters ζ_i. Details on how to calculate these parameters can be found in Section 3.2 of the article. However, due to constraints imposed by asymptotic behaviour, the local lower bound parameter a_i is restricted to the value $(-\frac{1}{2})$. Observe that the parameter a_i does not depend on the freely optimizable variational parameter ζ_i. Consequently, the local lower bound is unable to adjust its width (i.e., the function domain support) since the parameter a_i is fixed and not a function of ζ_i (Nickisch and Rasmussen 2008). This implies that the local lower bound will only be able to provide support for a small region of the log Gaussian cdf domain which, in turn, influences the quality of the expectation of the quadratic function with respect to the Gaussian density. However, we can leverage the heteroskedastic-based parameterization in Section A.1 to help us circumvent this limitation. Refer to Fig. 8.

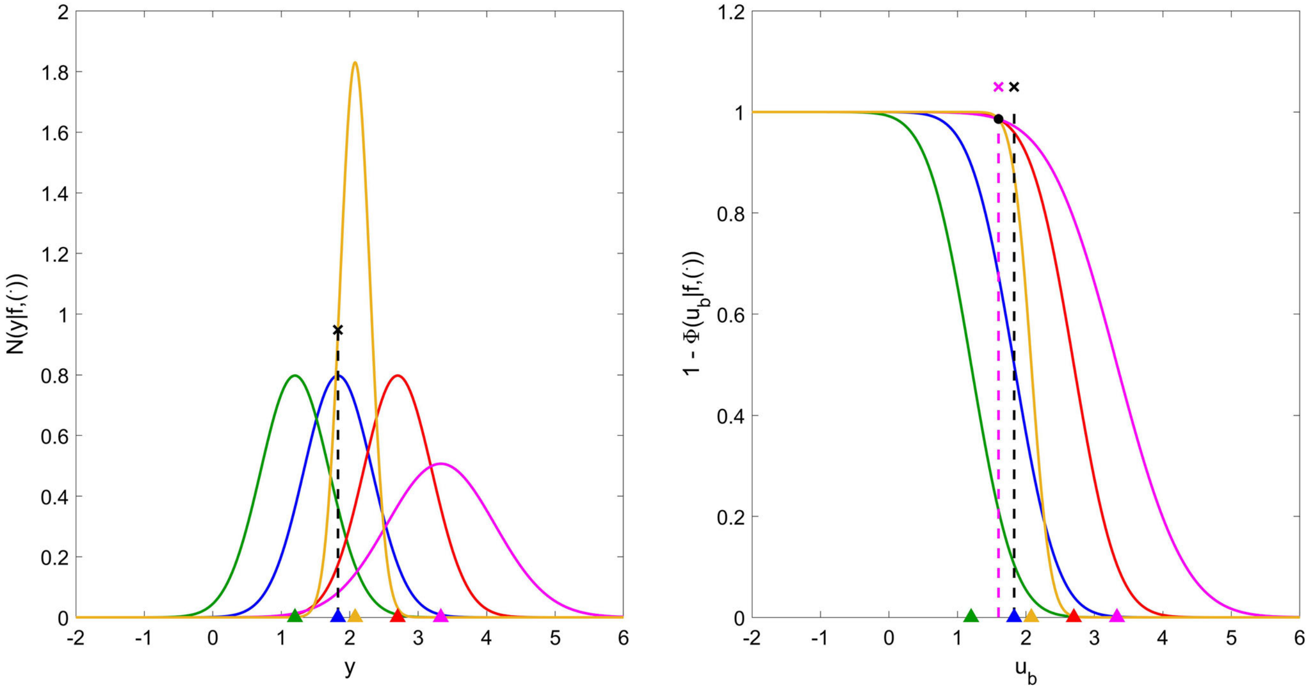

Fig. 7 Likelihood contribution, relative to the dashed lines, for an uncensored observation (left panel) and an upper bound censored observation (right panel), viewed as a function of u_b, with two additional proposal Gaussian densities (yellow and magenta). The yellow Gaussian density has a smaller (in numerical value) variance parameter, while the magenta Gaussian has a higher (in numerical value) variance parameters, when compared to the Gaussian densities in Fig. 6 (left panel). The black cross denotes the assumed position for the uncensored observation and the magenta cross denotes the assumed position for the upper bound censored observation associated with the uncensored observation (black cross). The black dot denotes a shared likelihood contribution value. The triangles correspond to 5 candidate mean values for f (latent function value of interest). Figure 7 has been reproduced and adjusted from the preprint of Gammelli et al. (2020b). (Color figure online)

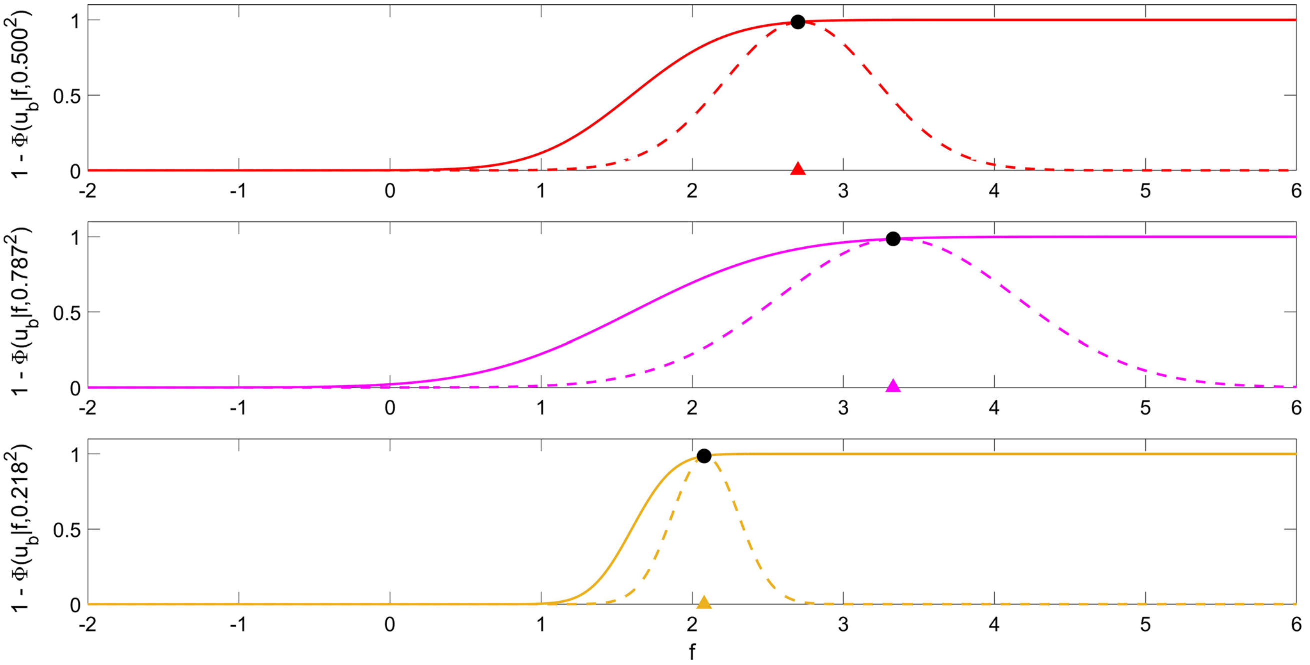

Fig. 8 Likelihood contribution for an upper bound censored observation (magenta cross, Fig. 7 - right panel) viewed as a function of f. The triangles correspond to the 3 candidate mean values (red, magenta and yellow) depicted in Fig. 7 (right panel) with a shared likelihood contribution value (black dot). The freely optimizable variational parameter for each local lower bound (dashed red, magenta and yellow curve) has been set to the corresponding candidate mean value to ensure the bound is tight, i.e., exact. (Color figure online)

Observe from Fig. 8 that for the same likelihood contribution value (black dot) the heteroskedastic parameterization associated with a Gaussian cdf factor with an increased variance corresponds to a local lower bound that provides support for a larger region of the Gaussian cdf domain (middle panel, Fig. 8). In other words, despite the asymptotic behaviour which limits the local lower bound support, we can use the increase in the variance parameter which arises from the heteroskedastic parameterization to regulate/tune the support of the local lower bound. This, in turn, affects the quality of the expectation of the quadratic function with respect to the Gaussian density. Note that one drawback of this approach stems from the fact that we are implicitly selecting the magenta Gaussian density (left panel, Fig. 7) as the most likely contender from which the observation was generated before censoring. However, the magenta density corresponds to a candidate mean value that can overestimate the true latent function value of interest (refer back to the arguments outlined in Sect. A.1).

Based on our previous discussion, we propose augmenting each Gaussian cdf factor in Eq. (9) with an additional variance parameter in an attempt to regulate/adjust the local lower bound support. Note that for training input locations associated with the upper detection limit u_b we assume a constant (with respect to the input x_i) heteroskedastic noise model with a total variance contribution which is the sum of the original mixed-likelihood variance σ_y^2 in Eq. (9) and a regulating variance parameter $\sigma_{u_b}^2$.

A.3 Deriving the Secondary Variational Lower Bound

Here we provide further details on how to derive the optimal 'collapsed' secondary variational lower bound that can be used for Bayesian model training and inference. Recall Eq. (43) in the article which is given by

$$\mathcal{F}^*(\boldsymbol{\theta}) = \ln \int\limits_{\boldsymbol{u}} p(\boldsymbol{u}) \exp\{\Psi(\boldsymbol{u})\} d\boldsymbol{u} \tag{58}$$

We start with our definition of $\Psi(\boldsymbol{u})$ by noting that

$$\Psi(\boldsymbol{u}) = \int\limits_{\boldsymbol{f}} p(\boldsymbol{f}|\boldsymbol{u}) \ln p_l(\boldsymbol{y}|\boldsymbol{f}) d\boldsymbol{f} \tag{59}$$

Next, we recall Eq. (36) and observe from Eq. (59) that

$$\begin{aligned} \Psi(\boldsymbol{u}) &= \int\limits_{\boldsymbol{f}} p(\boldsymbol{f}|\boldsymbol{u}) \ln p_l(\boldsymbol{y}|\boldsymbol{f}) d\boldsymbol{f} \\ p_l(\boldsymbol{y}|\boldsymbol{f}) &= g(\boldsymbol{f}_{l_b}|\cdots) \\ &\quad \times \left[\prod_{l_b < y_i < u_b} \mathcal{N}(y_i | f_i, \sigma_y^2) \right] g(\boldsymbol{f}_{u_b}|\cdots) \end{aligned} \tag{60}$$

Note that the definitions for functions $g(\boldsymbol{f}_{l_b}|\cdots)$ and $g(\boldsymbol{f}_{u_b}|\cdots)$ are given by Eqs. (33) and (34), respectively. Using properties of the natural logarithm, as well as the marginalization property of the multivariate Gaussian density function, we can show that Eq. (60) decomposes into the sum of the following expectations

$$\begin{aligned} \Psi(\boldsymbol{u}) &= \int\limits_{\boldsymbol{f}_{l_b}} p(\boldsymbol{f}_{l_b}|\boldsymbol{u}) \ln g(\boldsymbol{f}_{l_b}|\cdots) d\boldsymbol{f}_{l_b} \\ &\quad + \int\limits_{\boldsymbol{f}_{un}} p(\boldsymbol{f}_{u_n}|\boldsymbol{u}) \ln \mathcal{N}(\boldsymbol{y}_o|\boldsymbol{f}_{u_n} \cdots) d\boldsymbol{f}_{u_n} \\ &\quad + \int\limits_{\boldsymbol{f}_{u_b}} p(\boldsymbol{f}_{u_b}|\boldsymbol{u}) \ln g(\boldsymbol{f}_{u_b}|\cdots) d\boldsymbol{f}_{u_b} \\ \mathcal{N}(\boldsymbol{y}_o|\boldsymbol{f}_{u_n} \cdots) &= \prod_{l_b < y_i < u_b} \mathcal{N}(y_i | f_i, \sigma_y^2) \end{aligned}$$

We can analytically evaluate $\Psi(\boldsymbol{u})$ and show that $\exp\{\Psi(\boldsymbol{u})\}$ corresponds to

$$\begin{aligned} &\exp\{\Psi(\boldsymbol{u})\} \\ &\quad = \exp\left\{-\frac{1}{2}\boldsymbol{\rho}_l^T \boldsymbol{\Sigma}_{y_l}^{-1} \boldsymbol{\rho}_l\right\} \\ &\qquad \times \exp\left\{\frac{1}{2}\boldsymbol{b}_{l_b}^T \boldsymbol{\Sigma}_{l_b}^{-1} \boldsymbol{b}_{l_b} + \frac{1}{2}\boldsymbol{b}_{u_b}^T \boldsymbol{\Sigma}_{u_b}^{-1} \boldsymbol{b}_{u_b} + B\right\} \\ &\boldsymbol{\rho}_l = \boldsymbol{y}_l - \boldsymbol{K}_{NM}^l \boldsymbol{K}_{MM}^{-1} \boldsymbol{u} \end{aligned}$$

The definitions for $\boldsymbol{y}_l$, $\boldsymbol{K}_{NM}^l$, $\boldsymbol{\Sigma}_{y_l}$, $\boldsymbol{b}_{l_b}$ and $\boldsymbol{b}_{u_b}$ are given in the article. We define $\boldsymbol{\Sigma}_{l_b}$, $\boldsymbol{\Sigma}_{u_b}$ and B as follows

$$\begin{aligned} \boldsymbol{\Sigma}_{l_b} &= \left(\sigma_y^2 + \sigma_{l_b}^2\right) \boldsymbol{I}_{N_{l_b} N_{l_b}} \\ \boldsymbol{\Sigma}_{u_b} &= \left(\sigma_y^2 + \sigma_{u_b}^2\right) \boldsymbol{I}_{N_{u_b} N_{u_b}} \\ B &= -\frac{N_{y_o}}{2} \ln(2\pi\sigma_y^2) \\ &\quad -\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}_{l_b}^{-1}\left[\boldsymbol{K}_{N_{l_b} N_{l_b}} - \boldsymbol{K}_{N_{l_b} M} \boldsymbol{K}_{MM}^{-1} \boldsymbol{K}_{M N_{l_b}}\right]\right\} \\ &\quad +\boldsymbol{c}_{l_b}^T \boldsymbol{\Sigma}_{l_b}^{-1} \boldsymbol{1}_{l_b} - \boldsymbol{b}_{l_b}^T \boldsymbol{\Sigma}_{l_b}^{-1} \left((l_b) \times \boldsymbol{1}_{l_b}\right) \\ &\quad -\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}_{y_o}^{-1}\left[\boldsymbol{K}_{N_{y_o} N_{y_o}} - \boldsymbol{K}_{N_{y_o} M} \boldsymbol{K}_{MM}^{-1} \boldsymbol{K}_{M N_{y_o}}\right]\right\} \\ &\quad + \boldsymbol{c}_{u_b}^T \boldsymbol{\Sigma}_{u_b}^{-1} \boldsymbol{1}_{u_b} - \boldsymbol{b}_{u_b}^T \boldsymbol{\Sigma}_{u_b}^{-1} \left((u_b) \times \boldsymbol{1}_{u_b}\right) \\ &\quad - \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}_{u_b}^{-1}\left[\boldsymbol{K}_{N_{u_b} N_{u_b}} - \boldsymbol{K}_{N_{u_b} M} \boldsymbol{K}_{MM}^{-1} \boldsymbol{K}_{M N_{u_b}}\right]\right\} \end{aligned}$$

We denote with the symbol $\mathrm{tr}(\cdot)$ the matrix trace operator. We define $\boldsymbol{\Sigma}_{y_o}$ as follows

$$\boldsymbol{\Sigma}_{y_o} = \sigma_y^2 \boldsymbol{I}_{N_{y_o} N_{y_o}}$$

Next, we require evaluating the term $p(\boldsymbol{u}) \exp\{\Psi(\boldsymbol{u})\}$. The multivariate normal density $p(\boldsymbol{u})$ takes the following form (for more details, see Titsias 2008, 2009)

$$p(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u}|\boldsymbol{0}, \boldsymbol{K}_{MM})$$

We then have that

$$p(\boldsymbol{u}) \exp\{\Psi(\boldsymbol{u})\} = c \exp\left\{-\frac{1}{2}\left[\boldsymbol{\rho}_l^T \boldsymbol{\Sigma}_{y_l}^{-1} \boldsymbol{\rho}_l + \boldsymbol{u}^T \boldsymbol{K}_{MM}^{-1} \boldsymbol{u}\right]\right\} \tag{61}$$

We define the constant c associated with Eq. (61) as follows

$$\begin{aligned} c &= \frac{1}{(2\pi)^{\frac{M}{2}}} \frac{1}{|\boldsymbol{K}_{MM}|^{\frac{1}{2}}} \\ &\times \exp\left\{\frac{1}{2}\boldsymbol{b}_{l_b}^T \boldsymbol{\Sigma}_{l_b}^{-1} \boldsymbol{b}_{l_b} + \frac{1}{2}\boldsymbol{b}_{u_b}^T \boldsymbol{\Sigma}_{u_b}^{-1} \boldsymbol{b}_{u_b} + B\right\} \end{aligned}$$

We denote with the symbol $|\cdot|$ the matrix determinant. After completing the multivariate square of Eq. (61) with respect to $\boldsymbol{u}$ we can show that

$$p(\boldsymbol{u}) \exp\{\Psi(\boldsymbol{u})\} = d \exp\left\{-\frac{1}{2}\left(\boldsymbol{u} - \boldsymbol{\mu}_u\right)^T \boldsymbol{S}_u^{-1} \left(\boldsymbol{u} - \boldsymbol{\mu}_u\right)\right\} \tag{62}$$

The definitions for vector $\boldsymbol{\mu}_u$ and matrix $\boldsymbol{S}_u$ are given in the main text of the article. We define the constant d as follows

$$d = c \exp\left\{-\frac{1}{2}\left[-\boldsymbol{\mu}_u^T \boldsymbol{S}_u^{-1} \boldsymbol{\mu}_u + \boldsymbol{y}_l^T \boldsymbol{\Sigma}_{y_l}^{-1} \boldsymbol{y}_l\right]\right\}$$

Next, we substitute Eq. (62) into Eq. (58) to obtain

$$\begin{aligned} \mathcal{F}^*(\boldsymbol{\theta}) &= \ln \int_{\boldsymbol{u}} d \exp\left\{-\frac{1}{2}\mathcal{A}_u\right\} d\boldsymbol{u} \\ \mathcal{A}_u &= \left(\boldsymbol{u} - \boldsymbol{\mu}_u\right)^T \boldsymbol{S}_u^{-1} \left(\boldsymbol{u} - \boldsymbol{\mu}_u\right) \\ \mathcal{F}^*(\boldsymbol{\theta}) &= \ln\left[d \int_{\boldsymbol{u}} \exp\left\{-\frac{1}{2}\mathcal{A}_u\right\} d\boldsymbol{u}\right] \end{aligned} \tag{63}$$

We recognize the integral in Equation (63) as the normalization constant for a multivariate Gaussian density. Hence, we have that

$$\mathcal{F}^*(\boldsymbol{\theta}) = \ln\left[d (2\pi)^{\frac{M}{2}} |\boldsymbol{S}_u|^{\frac{1}{2}}\right] \tag{64}$$

We expand Eq. (64) by noting that

$$\begin{aligned} \mathcal{F}^*(\boldsymbol{\theta}) &= \ln d + \ln (2\pi)^{\frac{M}{2}} + \ln |\boldsymbol{S}_u|^{\frac{1}{2}} \\ \mathcal{F}^*(\boldsymbol{\theta}) &= \ln (2\pi)^{\frac{M}{2}} + \ln |\boldsymbol{S}_u|^{\frac{1}{2}} - \ln (2\pi)^{\frac{M}{2}} \\ &- \ln |\boldsymbol{K}_{MM}|^{\frac{1}{2}} + \frac{1}{2}\boldsymbol{b}_{l_b}^T \boldsymbol{\Sigma}_{l_b}^{-1} \boldsymbol{b}_{l_b} + \frac{1}{2}\boldsymbol{b}_{u_b}^T \boldsymbol{\Sigma}_{u_b}^{-1} \boldsymbol{b}_{u_b} \\ &- \frac{1}{2}\left[\boldsymbol{y}_l^T \boldsymbol{\Sigma}_{y_l}^{-1} \boldsymbol{y}_l - \boldsymbol{y}_l^T \boldsymbol{\Sigma}_{y_l}^{-1} \boldsymbol{K}_{NM}^l \boldsymbol{Q}^{-1} \boldsymbol{K}_{MN}^l \boldsymbol{\Sigma}_{y_l}^{-1} \boldsymbol{y}_l\right] \\ &- \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}_{l_b}^{-1}\left[\boldsymbol{K}_{N_{l_b} N_{l_b}} - \boldsymbol{K}_{N_{l_b} M} \boldsymbol{K}_{MM}^{-1} \boldsymbol{K}_{M N_{l_b}}\right]\right\} \\ &- \frac{N_{y_o}}{2} \ln (2\pi\sigma_y^2) + \boldsymbol{c}_{l_b}^T \boldsymbol{\Sigma}_{l_b}^{-1} \boldsymbol{1}_{l_b} - \boldsymbol{b}_{l_b}^T \boldsymbol{\Sigma}_{l_b}^{-1} \left((l_b) \times \boldsymbol{1}_{l_b}\right) \\ &+ \boldsymbol{c}_{u_b}^T \boldsymbol{\Sigma}_{u_b}^{-1} \boldsymbol{1}_{u_b} - \boldsymbol{b}_{u_b}^T \boldsymbol{\Sigma}_{u_b}^{-1} \left((u_b) \times \boldsymbol{1}_{u_b}\right) \\ &- \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}_{y_o}^{-1}\left[\boldsymbol{K}_{N_{y_o} N_{y_o}} - \boldsymbol{K}_{N_{y_o} M} \boldsymbol{K}_{MM}^{-1} \boldsymbol{K}_{M N_{y_o}}\right]\right\} \\ &- \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}_{u_b}^{-1}\left[\boldsymbol{K}_{N_{u_b} N_{u_b}} - \boldsymbol{K}_{N_{u_b} M} \boldsymbol{K}_{MM}^{-1} \boldsymbol{K}_{M N_{u_b}}\right]\right\} \end{aligned} \tag{65}$$

Note that to arrive at Eq. (65) we used the fact that

$$\boldsymbol{\mu}_u^T \boldsymbol{S}_u^{-1} \boldsymbol{\mu}_u = \boldsymbol{y}_l^T \boldsymbol{\Sigma}_{y_l}^{-1} \boldsymbol{K}_{NM}^l \boldsymbol{Q}^{-1} \boldsymbol{K}_{MN}^l \boldsymbol{\Sigma}_{y_l}^{-1} \boldsymbol{y}_l$$

Furthermore, the definition for $\boldsymbol{Q}$ is given in the article. Using the definition for $\boldsymbol{S}_u$ in the article and properties of the matrix determinant, we can simplify Eq. (65) to obtain

$$\begin{aligned} \mathcal{F}^*(\boldsymbol{\theta}) &= \ln |\boldsymbol{K}_{MM}|^{\frac{1}{2}} - \ln |\boldsymbol{Q}|^{\frac{1}{2}} \\ &- \frac{N_{y_o}}{2} \ln (2\pi) - \frac{N_{y_o}}{2} \ln (\sigma_y^2) - \frac{1}{2}\boldsymbol{y}_l^T \boldsymbol{A} \boldsymbol{y}_l \\ &+ \left[\frac{1}{2}\boldsymbol{b}_{l_b}^T \boldsymbol{\Sigma}_{l_b}^{-1} \boldsymbol{b}_{l_b} + \boldsymbol{c}_{l_b}^T \boldsymbol{\Sigma}_{l_b}^{-1} \boldsymbol{1}_{l_b} - \boldsymbol{b}_{l_b}^T \boldsymbol{\Sigma}_{l_b}^{-1} \left((l_b) \times \boldsymbol{1}_{l_b}\right)\right] \\ &+ \left[\frac{1}{2}\boldsymbol{b}_{u_b}^T \boldsymbol{\Sigma}_{u_b}^{-1} \boldsymbol{b}_{u_b} + \boldsymbol{c}_{u_b}^T \boldsymbol{\Sigma}_{u_b}^{-1} \boldsymbol{1}_{u_b} - \boldsymbol{b}_{u_b}^T \boldsymbol{\Sigma}_{u_b}^{-1} \left((u_b) \times \boldsymbol{1}_{u_b}\right)\right] \\ &- \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}_{l_b}^{-1}\left[\boldsymbol{K}_{N_{l_b} N_{l_b}} - \boldsymbol{K}_{N_{l_b} M} \boldsymbol{K}_{MM}^{-1} \boldsymbol{K}_{M N_{l_b}}\right]\right\} \\ &- \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}_{y_o}^{-1}\left[\boldsymbol{K}_{N_{y_o} N_{y_o}} - \boldsymbol{K}_{N_{y_o} M} \boldsymbol{K}_{MM}^{-1} \boldsymbol{K}_{M N_{y_o}}\right]\right\} \\ &- \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}_{u_b}^{-1}\left[\boldsymbol{K}_{N_{u_b} N_{u_b}} - \boldsymbol{K}_{N_{u_b} M} \boldsymbol{K}_{MM}^{-1} \boldsymbol{K}_{M N_{u_b}}\right]\right\} \end{aligned} \tag{66}$$

The definition for $\boldsymbol{A}$ is given in the article. We can further simplify Eq. (66) by rewriting the local likelihood lower bound contributions, as well as the trace term contributions, into a compact format such that

$$\begin{aligned} \mathcal{F}^*(\boldsymbol{\theta}) &= \ln |\boldsymbol{K}_{MM}|^{\frac{1}{2}} - \ln |\boldsymbol{Q}|^{\frac{1}{2}} \\ &- \frac{N_{y_o}}{2} \ln (2\pi) - \frac{N_{y_o}}{2} \ln (\sigma_y^2) - \frac{1}{2}\boldsymbol{y}_l^T \boldsymbol{A} \boldsymbol{y}_l \\ &+ \frac{1}{2}\boldsymbol{b}^T \boldsymbol{\Sigma}_c^{-1} \boldsymbol{b} + \boldsymbol{c}^T \boldsymbol{\Sigma}_c^{-1} \boldsymbol{1}^* - \boldsymbol{b}^T \boldsymbol{\Sigma}_c^{-1} \boldsymbol{d} \\ &- \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}_{y_l}^{-1}\left[\boldsymbol{K}_{NN}^l - \boldsymbol{K}_{NM}^l \boldsymbol{K}_{MM}^{-1} \boldsymbol{K}_{MN}^l\right]\right\} \end{aligned} \tag{67}$$

The definitions for vectors $\boldsymbol{b}$, $\boldsymbol{c}$, $\mathbf{1}^*$ and $\boldsymbol{d}$, as well as for the matrices $\boldsymbol{\Sigma}_c$ and $\boldsymbol{K}^l_{NN}$, are given in the article. We arrive at the optimal ‘collapsed’ secondary variational lower bound (refer to Eq. (46) in the article) by rewriting Equation (67) to obtain

$$
\begin{aligned}
&\mathcal{F}^*(\boldsymbol{\theta})\\
&\quad = \ln\left\{\frac{|\boldsymbol{K}_{MM}|^{\frac{1}{2}}}{(2\pi)^{\frac{N_{y_o}}{2}}(\sigma_y^2)^{\frac{N_{y_o}}{2}}|\boldsymbol{Q}|^{\frac{1}{2}}}\exp\{\mathcal{A}_{\mathcal{F}^*}\}\right\}\\
&\qquad -\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}_{y_l}^{-1}\left[\boldsymbol{K}^l_{NN}-\boldsymbol{K}^l_{NM}\boldsymbol{K}^{-1}_{MM}\boldsymbol{K}^l_{MN}\right]\right\}\\
&\mathcal{A}_{\mathcal{F}^*} = -\frac{1}{2}\boldsymbol{y}_l{}^T\boldsymbol{A}\boldsymbol{y}_l+\frac{1}{2}\boldsymbol{b}^T\boldsymbol{\Sigma}_c^{-1}\boldsymbol{b}\\
&\qquad +\boldsymbol{c}^T\boldsymbol{\Sigma}_c^{-1}\mathbf{1}^*-\boldsymbol{b}^T\boldsymbol{\Sigma}_c^{-1}\boldsymbol{d}
\end{aligned}
$$

Appendix B

Here we provide supplementary simulation-based results for Section 5.

B.1 Selecting the Number of Inducing Variables: Example 1

In this subsection, we give some insight into how we selected the number of inducing variables for our VT-GPR implementation in Example 1. For our randomly generated data set, we selected 7 inducing variables followed by running our gradient-based optimizer to find point estimates for $\boldsymbol{\theta}$. From an implementation perspective, we minimized the negative secondary variational lower bound (see Eq. (46) in the article) using fmincon, in conjunction with the MultiStart algorithm, in MATLAB.

The MultiStart algorithm allows users to explore multiple starting points for $\boldsymbol{\theta}$ and we arbitrarily selected 1000 starting points. The algorithm returns multiple point estimates for $\boldsymbol{\theta}$, each associated with a local minimum, ranked according to the objective function value. We recorded the lowest objective function value which corresponds to the best local minimum found by the optimizer. The same procedure is then repeated but we incrementally increase the number of inducing variables until we reach 20 inducing variables.

Refer to Fig. 9 for a plot of the objective function value against the number of inducing variables. We see that as the number of inducing variables increases, the objective function value decreases until it stabilizes around (roughly) 15 inducing variables. This indicates that the secondary variational lower bound has reached a point where it is sufficiently tight (for more details, see Titsias, 2008, 2009) and we would gain no further benefit from increasing the number of inducing variables. We will point out that our proposed approach for selecting the number of inducing variables is not necessarily practical, especially in an online setting or when limited computational resources and time are available. The interested reader is referred to the work of Galy-Fajou and Opper (2021) and Uhrenholt et al. (2021) for recent approaches on selecting the number of inducing variables.

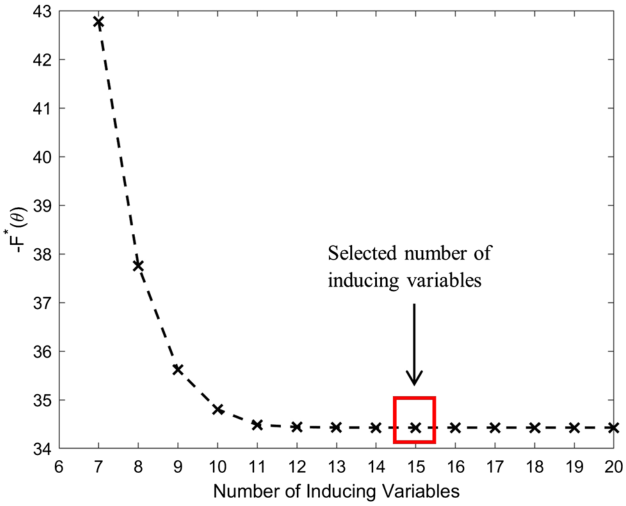

Fig. 9 Plot of the negative secondary variational lower bound (see Eq. (46) in the article) as the number of inducing variables are incrementally increased from 7 to 20. (Color figure online)

B.2 Deteriorated MNLL Performance: Example 1

In order to investigate the worse MNLL performance, we simulated three additional independently generated data sets which highlight some key features of the proposed VT-GPR framework. From the right panels in Fig. 10 we observe that the VT-GPR framework tends to overestimate the true latent function in the lower bound censored regions (see Sections. A.1 and A.2), relative to the mean model prediction, and also produces less conservative credibility intervals, i.e., smaller credibility intervals, when compared to the LA and EP-based frameworks.

When looking at the functional form of the MNLL error measure (see Eq. (55) in the article) we observe that the overestimating mean model prediction and the less conservative credibility intervals inflate the MNLL performance measure, especially in the lower bound censored latent function regions (see, for example, the stratified results in Table 1 of the article).

We attribute this behaviour to the adjusted mixed-likelihood which has a single regulating variance parameter that must regulate/tune {1} the amount of overestimation associated with the latent function value (see Section A.1) and {2} regulate/adjust the local likelihood lower bound support (see Section A.2) for all of the lower bound censored observations. We postulate that, despite having a single regulating

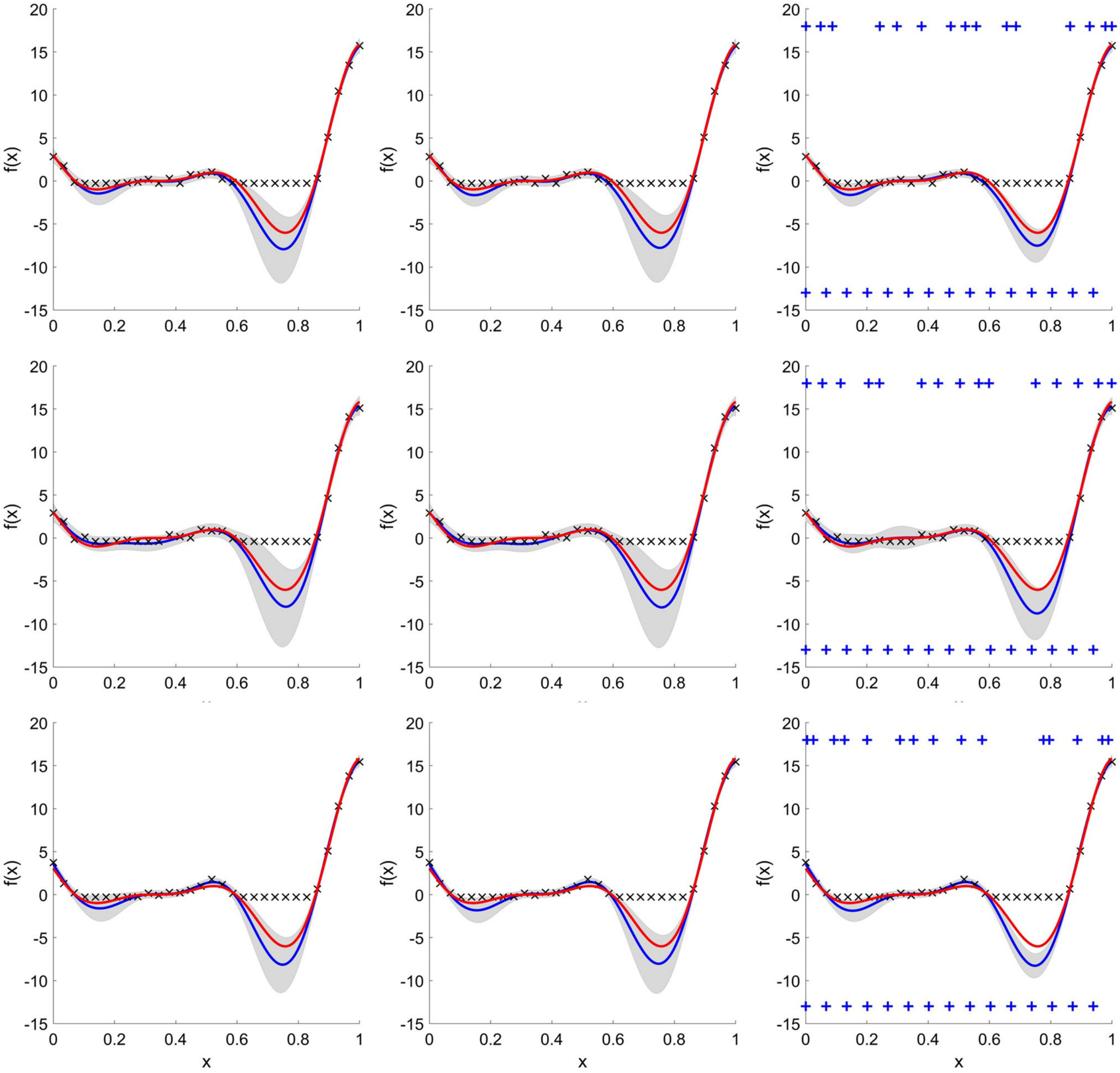

Fig. 10 Tobit Gaussian process regression results obtained from 3 independently generated data sets with l_b set to the 40th percentile of the uncensored observational data. Left Panels: T-GPR latent function predictive results using the Laplace approximation. Middle Panels: T-GPR latent function predictive results using the expectation propagation framework. Right Panels: VT-GPR latent function predictive results. Additional Information: The black '×'-sign denotes the observational data (noisy and/or censored), the red line denotes the underlying latent function (see Eq. (52) in the article) while the blue curve denotes the mean model prediction (model MAP estimate). The corresponding grey shaded area depicts the 99% point-wise credibility interval. Furthermore, the blue '+'-sign at the bottom of the right panels depict the initial inducing input locations while the optimized inducing input locations are depicted at the top of the right panels. We arbitrarily selected 15 as the number of inducing variables for our VT-GPR implementation (see Section B.1 for more details). (Color figure online)

variance parameter, the VT-GPR framework does not have enough regulating capacity for all the lower bound censored observations, i.e., the single regulating variance parameter does not provide sufficient regulating/tuning capacity. This limitation can potentially be circumvented by allowing each lower bound censored observation to have a unique regulating variance parameter.

B.3 Additional Figures for Example 2

Plots of the latent function predictive results (Fig. 11) from the single simulation, as well as box plots (Fig. 12) comparing the performance metrics across 1000 independently generated data sets, for the various T-GPR frameworks from Example 2. For the data in Fig. 11, we calculate that the

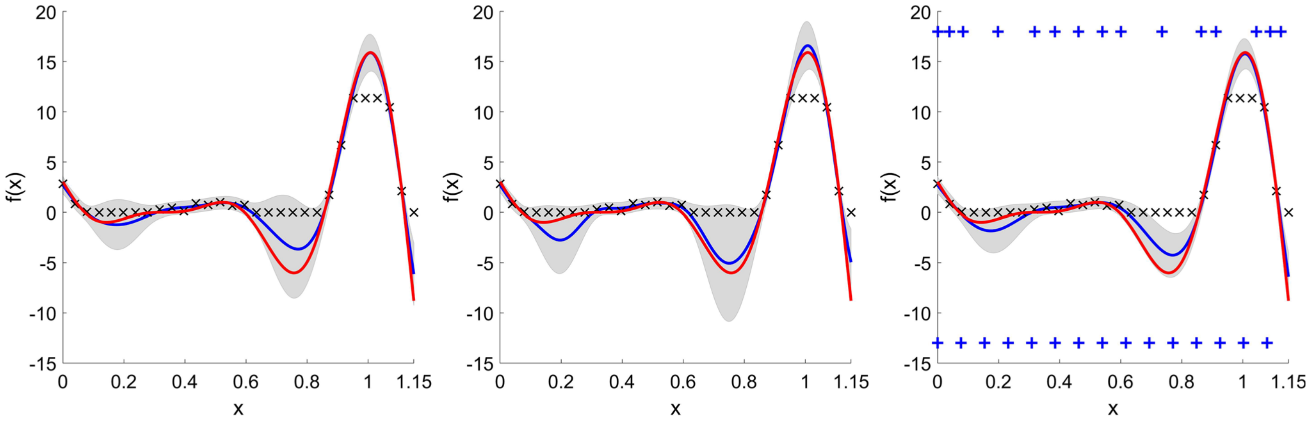

Fig. 11 Tobit Gaussian process regression results with l_b and u_b set to the 40th and 90th percentile of the uncensored observational data, respectively. Left Panel: T-GPR latent function predictive results using the Laplace approximation. Middle Panel: T-GPR latent function predictive results using the expectation propagation framework. Right Panel: VT-GPR latent function predictive results. Additional Information: The black '×'-sign denotes the observational data (noisy and/or censored), the red line denotes the underlying latent function (see Eq. (52)) while the blue curve denotes the mean model prediction (model MAP estimate). The corresponding grey shaded area depicts the 99% point-wise credibility interval. Furthermore, similar to Fig. 2, the blue '+'-sign denotes the inducing input locations. (Color figure online)

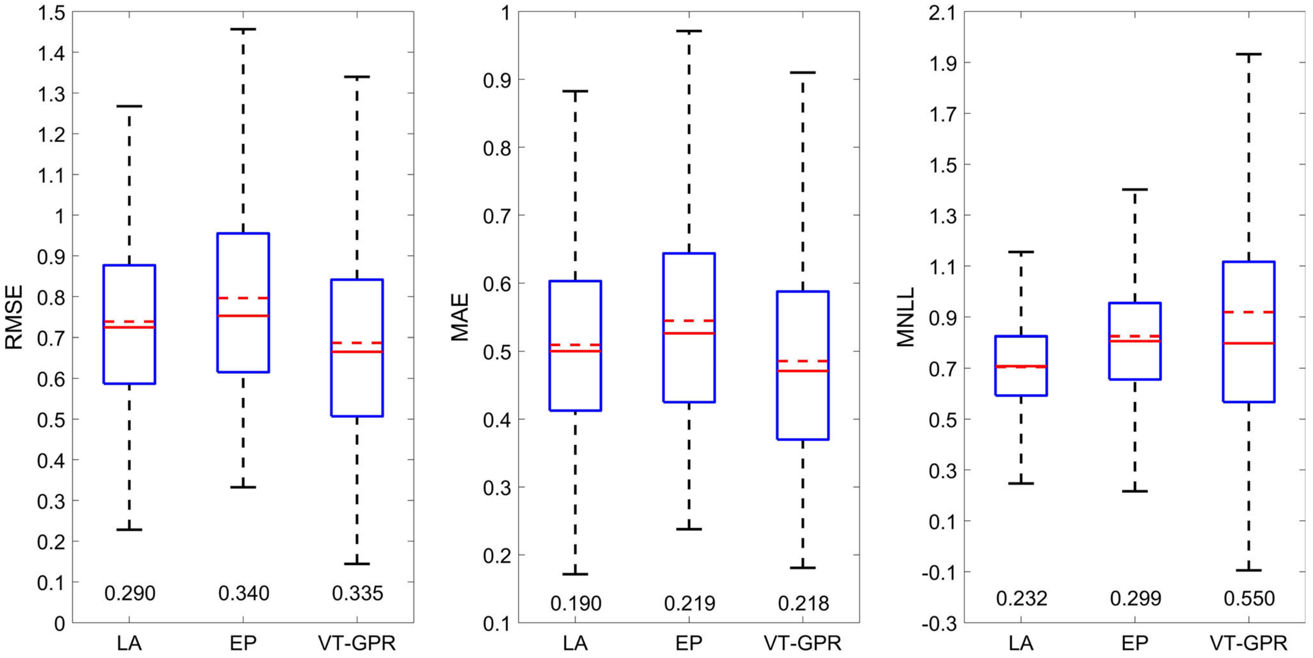

Fig. 12 Box plot visualization for the generated RMSE (left panel), MAE (middle panel) and MNLL (right panel) results, respectively, for each T-GPR framework. The dashed red line depicts the mean value for each quantitative performance measure across the 1000 additional independently generated data sets. The interquartile range is denoted at the bottom whisker of each box plot. (Color figure online)

lower and upper limits of detection are $l_b = -0.0239$ and $u_b = 11.3555$, respectively.

References

Alaa, A.M., van der Schaar, M.: Deep multi-task Gaussian processes for survival analysis with competing risks. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 2326–2334 (2017)

Allik, B., Miller, C., Piovoso, M.J., et al.: The Tobit Kalman filter: an estimator for censored measurements. IEEE Trans. Control Syst. Technol. **24**(1), 365–371 (2016). https://doi.org/10.1109/TCST.2015.2432155

Amemiya, T.: Tobit models: a survey. J. Econom. **24**(1), 3–61 (1984). https://doi.org/10.1016/0304-4076(84)90074-5

Barrett, J.E., Coolen, A.C.C.: Covariate dimension reduction for survival data via the Gaussian process latent variable model. Stat. Med. **35**(8), 1340–1353 (2016). https://doi.org/10.1002/sim.6784

Bishop, C.M.: Pattern Recognition and Machine Learning, Information science and statistics. Springer, New York (2009)

Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: a review for statisticians. J. Am. Stat. Assoc. **112**(518), 859–877 (2017). https://doi.org/10.1080/01621459.2017.1285773

Bui, T., Turner, R.: On the paper: Variational Learning of Inducing Variables in Sparse Gaussian Processes (Titsias, 2009). https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.716.942&rep=rep1&type=pdf (2014)

Campbell, K.R., Yau, C.: Bayesian Gaussian process latent variable models for pseudotime inference in single-cell RNA-seq data. bioRxiv **2015**, 1–6 (2015). https://doi.org/10.1101/026872

Chen, N., Qian, Z., Nabney, I.T., et al.: Wind power forecasts using Gaussian processes and numerical weather prediction. IEEE Trans. Power Syst. **29**(2), 656–665 (2013). https://doi.org/10.1109/TPWRS.2013.2282366

Csató, L., Opper, M.: Sparse on-line Gaussian processes. Neural Comput. **14**(3), 641–668 (2002). https://doi.org/10.1162/089976602317250933

Damianou, A., Lawrence, N.D.: Deep Gaussian processes. In: Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, pp. 207–215 (2013)

Damianou, A.C., Titsias, M.K., Lawrence, N.D.: Variational inference for latent variables and uncertain inputs in Gaussian processes. J. Mach. Learn. Res. **17**(42), 1–62 (2016)

DWS: National Water Management System data extracted on 2019-07-17. Department of Water and Sanitation. www.dwa.gov.za/iwqs/wms/data/C_reg_WMS_nobor.htm (2019)

EPA: Indicators: Conductivity. United States Environmental Protection Agency. www.epa.gov/national-aquatic-resource-surveys/indicators-conductivity (2022)

Ertin, E.: Gaussian process models for censored sensor readings. In: 2007 IEEE/SP 14th Workshop on Statistical Signal Processing, pp. 665–669 (2007) https://doi.org/10.1109/SSP.2007.4301342

Galy-Fajou, T., Opper, M.: Adaptive inducing points selection For Gaussian processes. arXiv preprint arXiv:2107.10066v1 https://doi.org/10.48550/arXiv.2107.10066 (2021)

Gammelli, D., Peled, I., Rodrigues, F., et al.: Estimating latent demand of shared mobility through censored Gaussian Processes. Transp. Res. Part C Emerg. Technol. **120**, 102775 (2020). https://doi.org/10.1016/J.TRC.2020.102775

Gammelli, D., Rolsted, K.P., Pacino, D., et al.: Generalized Multi-Output Gaussian Process Censored Regression. arXiv preprint arXiv:2009.04822v2 https://doi.org/10.48550/arXiv.2009.04822 (2020b)

Gibbs, M.N., MacKay, D.J.: Variational Gaussian process classifiers. IEEE Trans. Neural Netw. **11**(6), 1458–1464 (2000). https://doi.org/10.1109/72.883477

Groot, P., Lucas, P.: Gaussian Process Regression with Censored Data Using Expectation Propagation. In: Proceedings of the 6th European Workshop on Probabilistic Graphical Models, pp. 115–122 (2012)

Hensman, J., Fusi, N., Lawrence, N.D.: Gaussian processes for big data. In: Uncertainty in Artificial Intelligence—Proceedings of the 29th Conference, UAI 2013, pp. 282–290 (2013) https://doi.org/10.48550/arXiv.1309.6835

Hensman, J., Matthews, A., Ghahramani, Z.: Scalable variational Gaussian process classification. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, pp. 351–360 (2015) http://proceedings.mlr.press/v38/hensman15.pdf

Hoffman, M.D., Blei, D.M., Wang, C., et al.: Stochastic variational inference. J. Mach. Learn. Res. **14**, 1303–1347 (2013). https://doi.org/10.48550/arXiv.1206.7051

Hutter, F., Hoos, H., Leyton-Brown, K.: Bayesian Optimization With Censored Response Data. arXiv preprint arXiv:1310.1947v1 (2013) https://doi.org/10.48550/arXiv.1310.1947

Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., et al.: An introduction to variational methods for graphical models. Mach. Learn. **37**(2), 183–233 (1999). https://doi.org/10.1023/A:1007665907178

Lawrence, N.D.: Gaussian process latent variable models for visualisation of high dimensional data. In: Advances in Neural Information Processing Systems 16 (2004)

Lawrence, N.D.: Learning for Larger Datasets with the Gaussian Process Latent Variable Model. In: Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, pp. 243–250 (2007)

Lázaro-Gredilla, M.: Bayesian Warped Gaussian Processes. In: Advances in Neural Information Processing Systems (2012)

Lázaro-Gredilla, M., Quiñnero-Candela, J., Rasmussen, C.E., et al.: Sparse Spectrum Gaussian Process Regression. J. Mach. Learn. Res. **11**(63), 1865–1881 (2010)

Li, A.H., Bradic, J.: Censored quantile regression forest. In: International Conference on Artificial Intelligence and Statistics, pp. 2109–2119 (2020)

Logan, J.D.: Applied Mathematics, 3rd edn. Wiley, New York (2006)

MacKay, D.J.C.: Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge (2004)

Minka, T.P.: Expectation Propagation for approximate Bayesian inference. In: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, pp. 362–369 (2001a) https://doi.org/10.48550/arXiv.1301.2294

Minka, T.P.: A family of algorithms for approximate Bayesian inference. PhD thesis, Massachusetts Institute of Technology (2001b)

Nickisch, H., Rasmussen, C.E.: Approximations for binary gaussian process classification. J. Mach. Learn. Res. **9**(67), 2035–2078 (2008)

Pishro-Nik, H.: Introduction to Probability, Statistics, and Random Processes. Kappa Research LLC, London (2014)

Quiñonero-Candela, J., Rasmussen, C.E.: A unifying view of sparse approximate gaussian process regression. J. Mach. Learn. Res. **6**(65), 1939–1959 (2005)

Rao, A., Monteiro, J., Mourao-Miranda, J.: Prediction of clinical scores from neuroimaging data with censored likelihood Gaussian processes. In: 2016 International Workshop on Pattern Recognition in Neuroimaging (PRNI), pp. 1–4 (2016). https://doi.org/10.1109/PRNI.2016.7552358

Rasmussen, C.E., Williams, C.K.: Gaussian Processes for Machine Learning, Adaptive Computation and Machine Learning. MIT Press, Cambridge (2006)

Saul, A.D., Hensman, J., Vehtari, A., et al.: Chained Gaussian Processes. In: Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, pp. 1431–1440 (2016). https://doi.org/10.48550/arXiv.1604.05263

Seeger, M., Williams, C. K.I., Lawrence, N.D.: Fast Forward Selection to Speed Up Sparse Gaussian Process Regression. In: Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics (2003)

Snelson, E., Ghahramani, Z.: Sparse Gaussian processes using pseudo-inputs. In: Advances in Neural Information Processing Systems vol. 18 (2005)

Snelson, E., Rasmussen, C.E., Ghahramani, Z.: Warped Gaussian processes. Adv. Neural. Inf. Process. Syst. **16**, 337–344 (2004)

Titsias, M.: Variational Model Selection for Sparse Gaussian Process Regression. Tech. rep., University of Manchester, UK https://www2.aueb.gr/users/mtitsias/papers/sparseGPv2.pdf (2008)

Titsias, M.: Variational Learning of Inducing Variables in Sparse Gaussian Processes. In: Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, pp. 567–574 (2009)

Titsias, M., Lawrence, N.D.: Bayesian Gaussian process latent variable model. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 844–851 (2010)

Tobin, J.: Estimation of relationships for limited dependent variables. Econometrica **26**(1), 24–36 (1958). https://doi.org/10.2307/1907382

Uhrenholt, A.K., Charvet, V., Jensen, B.S.: Probabilistic selection of inducing points in sparse Gaussian processes. In: Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, pp. 1035–1044 (2021). https://doi.org/10.48550/arXiv.2010.09370

Urtasun, R., Fleet, D.J., Fua, P.: 3D people tracking with Gaussian process dynamical models. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 238–245 (2006). https://doi.org/10.1109/CVPR.2006.15

Vanhatalo, J., Riihimäki, J., Hartikainen, J., et al.: GPstuff: Bayesian Modeling with Gaussian Processes. J. Mach. Learn. Res. **14**, 1175–1179 (2013)

Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian process dynamical models for human motion. IEEE Trans. Pattern Anal. Mach. Intell. **30**(2), 283–298 (2008). https://doi.org/10.1109/TPAMI.2007.1167

Wu, W. C.-H., Yeh, M.-Y., Chen, M.-S.: Deep Censored Learning of the Winning Price in the Real Time Bidding. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining pp. 2526–2535 (2018). https://doi.org/10.1145/3219819.3220066

Zhang, J., Zhu, Z., Zou, J.: Supervised Gaussian process latent variable model based on Gaussian mixture model. In: 2017 International Conference on Security, Pattern Analysis, and Cybernetics, pp. 124–129 (2017). https://doi.org/10.1109/SPAC.2017.8304262