2013 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 22–25, 2013, SOUTHAMPTON, UK

BOUNDED GAUSSIAN PROCESS REGRESSION

Bjørn Sand Jensen, Jens Brehm Nielsen and Jan Larsen

Department of Applied Mathematics and Computer Science,
Technical University of Denmark,
Matematiktorvet Building 303B, 2800 Kongens Lyngby, Denmark
{bjje,jenb,janla}@dtu.dk

ABSTRACT

We extend the Gaussian process (GP) framework for *bounded* regression by introducing two bounded likelihood functions that model the noise on the dependent variable explicitly. This is fundamentally different from the implicit noise assumption in the previously suggested warped GP framework. We approximate the intractable posterior distributions by the Laplace approximation and expectation propagation and show the properties of the models on an artificial example. We finally consider two real-world data sets originating from perceptual rating experiments which indicate a significant gain obtained with the proposed explicit noise-model extension.

1. INTRODUCTION

Regression is typically defined as learning a mapping from a possible multi-dimensional input to an effectively unbounded one-dimensional observational space, i.e., the space of the dependent variable. However, in many regression problems the observational space is clearly bounded. Examples of such problems include prediction of betting odds, data compression ratios and ratings from perceptual experiments. When the observational space is bounded, modeling the observations with a distribution having infinite support such as the Gaussian distribution, is clearly incorrect from a probabilistic point of view. In this work we will extend the GP framework to allow for principle modeling of such observations.

Gaussian processes (GPs) are currently considered a state-of-the-art Bayesian regression method due to its flexible and non-parametric nature. However, *bounded* regression with GPs has only indirectly been addressed by mapping or *warping* the bounded observations onto a latent unbounded space in which the observational noise can be assumed to be Gaussian [1]. Hereby, the observational model is only modeled implicitly through the warping function. In contrast, we consider observational models or likelihood functions that make assumptions about the noise directly in the observational space, and thus, model the observational noise explicitly.

Possibly, the simplest way to derive a bounded likelihood function is to use a truncated distribution. A natural choice is to use the truncated version of the Gaussian distribution considered in this work. Alternatively, a bounded likelihood function could be derived from a distribution that only has finite support. Of this type, we will consider the beta distribution and derive a bounded likelihood function based on a re-parameterization. For both models we perform inference and predictions based on the Laplace approximation and expectation propagation (EP).

Employing a toy example, we compare the predictive distributions of warped GPs with regression based on the bounded likelihood functions mentioned above. We show that, as expected, the model with the correct noise assumption provides the best expected predictive negative log likelihood (or, alternatively, generalization error). Two examples are used to justify the models in real-world regression scenarios and they show that the two likelihood models provide better model fits compared to the warped GP.

2. GAUSSIAN PROCESS REGRESSION

A Gaussian process (GP) is a stochastic process defined as a collection of random variables, any finite subset of which must have a joint Gaussian distribution. In effect, we may place the GP as a prior over any finite set of functional values $\mathbf{f} = [f_1, f_2, ..., f_n]^\top$, where $f_i = f(\mathbf{x}_i)$, resulting in a finite multivariate (zero-mean) Gaussian distribution over the set as $p(\mathbf{f}|\mathcal{X}, \boldsymbol{\theta}_c) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K})$, where each element of the covariance matrix $[\mathbf{K}]_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)_{\boldsymbol{\theta}_c}$ is given by a covariance function $k(\cdot,\cdot)_{\boldsymbol{\theta}_c}$ with parameters $\boldsymbol{\theta}_c$, and where $\mathcal{X} = \{\mathbf{x}_i | i = 1, ..., n\}$ denotes the set of inputs. The GP is effectively used as a prior over functions in non-parametric Bayesian regression frameworks where either the outputs or a likelihood can be parameterized by a smooth and continuous function $f(\cdot)$. In the simplest case the set of observations, $\mathcal{Y} = \{y_i | i = 1, ..., n\}$, consists of the functional val-

This work was supported in part by the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation under the CoSound project, case number 11-115328.

978-1-4799-1180-6/13/$31.00 ©2013 IEEE

ues themselves with added i.i.d Gaussian noise with variance σ_n^2. Hereby, the likelihood function is a standard Gaussian likelihood function parameterized by $f(\cdot)$ defining the mean. Hence, $p(y_i|f_i, \boldsymbol{\theta}_{\mathcal{L}}) = \mathcal{N}(y_i|f_i, \sigma^2)$.

Bayes formula gives us—regardless of the likelihood function—the posterior distribution,

$$p(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta}) = \frac{p(\mathcal{Y}|\mathbf{f}, \boldsymbol{\theta}_{\mathcal{L}})p(\mathbf{f}|\mathcal{X}, \boldsymbol{\theta}_c)}{p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})},$$

where it is typically assumed that the likelihood factorizes over instances such that $p(\mathcal{Y}|\mathbf{f}, \boldsymbol{\theta}_{\mathcal{L}}) = \prod_{i=1}^{n} p(y_i|f_i, \boldsymbol{\theta}_{\mathcal{L}})$. The denominator, $p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})$, is called the *marginal likelihood* or *evidence* given as $p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}) = \int p(\mathcal{Y}|\mathbf{f}, \boldsymbol{\theta}_{\mathcal{L}})$ $p(\mathbf{f}|\mathcal{X}, \boldsymbol{\theta}_c)d\mathbf{f}$. In empirical Bayesian methods the evidence is used to learn point estimates of both likelihood function and prior parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_c, \boldsymbol{\theta}_{\mathcal{L}}\}$.

Provided that the likelihood is Gaussian, both the posterior and predictive distribution will be Gaussian (processes) available in closed form [2, Chapter 2]. However, not all real-world problems actually justify the observations to be Gaussian distributed. As mentioned, we consider *bounded* observations, meaning that they in contrast to Gaussian distributed observations do not have infinite support.

3. BOUNDED LIKELIHOOD FUNCTIONS

We consider a set $\mathcal{Y} = \{y_i|i = 1, ..., n\}$ of bounded responses $y_i \in \,]a, b[$ to an input $\mathbf{x}_i$. In the following we will present three different observational models for this type of response. The first is the warped GP [1], where the likelihood describes warped observations rather than the bounded responses directly. Following this, we propose two different likelihood functions that directly model the bounded responses in a principle probabilistic fashion by assuming particular distributions of the observations defining the noise in the original bounded domain.

3.1. Warping

Snelson *et. al* [1] learn a warping, that transforms the original data $\mathcal{Y}$ into a form where the data is modeled by a traditional GP with a Gaussian noise model. Here, we will not consider how to learn the correct warping, but instead use a fixed warping that transforms the bounded responses y_i into unbounded versions z_i. Several warping functions would apply, but to allow for direct comparison of all the models we use the inverse cumulative Gaussian (probit) $\Phi^{-1}(\cdot)$—with zero mean and unity variance—such that $z_i = \Phi^{-1}(y_i)$. The resulting model will be referred to as GP-WA.

3.2. Truncated Distributions

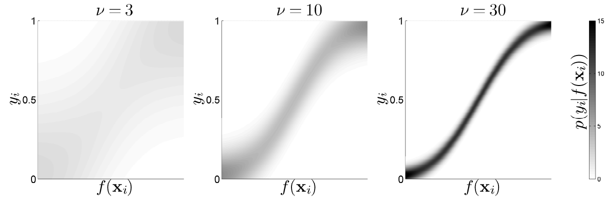

Fig. 1. Illustration of the proposed TG likelihood function with $p(y_i|f_i)$ shown as a gray-scale level. Left: $\nu = 3$, Middle: $\nu = 10$ and Right: $\nu = 30$.

The simplest route to a bounded likelihood function is to use distributions with infinite support and truncate them to the bounded domain. There are a number of relevant distributions including the truncated student-t and of course the truncated Gaussian (TG) distribution, see e.g. [3, 4]. As a representative for this type of bounding approach, we consider the TG and define the corresponding likelihood function as

$$\begin{aligned}\mathcal{L}_{TG} &\equiv p\left(y_i|f_i, \boldsymbol{\theta}_{\mathcal{L}}\right) \\ &= \frac{\nu \mathcal{N}\left(\nu\left(y_i - \mathrm{M}\left(f_i\right)\right)\right)}{\Phi\left(\nu\left(b - \mathrm{M}\left(f_i\right)\right)\right) - \Phi\left(\nu\left(a - \mathrm{M}\left(f_i\right)\right)\right)},\end{aligned} \tag{1}$$

where the distribution is parameterized by the mode $\mathrm{M}(f_i)$ and the domain limits a and b which we assume to be 0 and 1, respectably[1]. The mean of the TG distribution is given by

$$\begin{aligned}\mu\left(f_i\right) &= \mathrm{M}\left(f_i\right) \\ &+ \frac{1}{\nu} \frac{\mathcal{N}\left(\nu\left(a - \mathrm{M}\left(f_i\right)\right)\right) - \mathcal{N}\left(\nu\left(b - \mathrm{M}\left(f_i\right)\right)\right)}{\Phi\left(\nu\left(b - \mathrm{M}\left(f_i\right)\right)\right) - \Phi\left(\nu\left(a - \mathrm{M}\left(f_i\right)\right)\right)}.\end{aligned} \tag{2}$$

Eq. 2 in effect leaves two parametrization options in the sense that we may select the non-parametric function, $f(\cdot)$, to parameterize either the mode or the mean function. Both options are valid from a modeling perspective, but the easiest parametrization is by far the mode, $\mathrm{M}(f_i)$. For prediction speed it may be beneficial to indirectly parameterize the mean, but then the (unique) solution to the mode given the mean must be found numerically or approximately. The numerical approach will severely limit the effectiveness of the posterior approximation and in this work we will therefore focus on the mode parametrization for the TG. Thus, the likelihood function in Eq. 1 is parameterized by the mode as follows $\mathrm{M}(f_i) = \Phi(f_i)$ and the resulting model depicted in Fig. 1 will be referred to as GP-TG

3.3. Beta

A distribution that imposes bounded responses in a completely natural manner is the beta distribution which has also been applied in standard parametric settings [5, 6]. The beta distribution is therefore an obvious distribution for the bounded observations and we select a parametrization which

[1] We note that the truncated student-t has the same form as the TG and can easily be realized using the methods and implementations presented in this work.

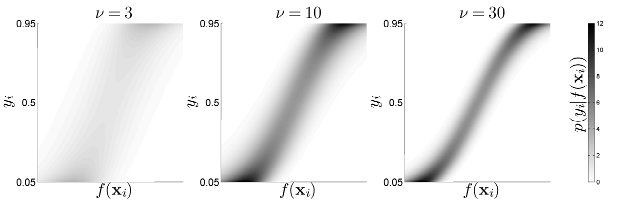

Fig. 2. Illustration of the proposed beta likelihood function with $p(y_i|f_i)$ shown as a gray-scale level. Left: $\nu = 3$, Middle: $\nu = 10$ and Right: $\nu = 30$.

expresses the shape parameters, α, β, of the beta distribution, Beta(α, β), in terms of the mean μ such that

$$\alpha = \nu\mu, \qquad \beta = \nu\left(1 - \mu\right).$$

We then parameterize the mean μ of the beta distribution by the cumulative Gaussian, such that $\mu(f_i) = \Phi(f_i)$. The reparameterized beta likelihood depicted in Fig. 2 is thereby given by

$$\mathcal{L}_{\text{BE}} \equiv p(y_i|f_i, \boldsymbol{\theta}_{\mathcal{L}}) = \text{Beta}(y_i|\nu\Phi(f_i), \nu\left(1 - \Phi(f_i)\right)),$$

and will be referred to as the GP-BE model. Note, that the ν parameter is an (inverse) dispersion parameter.

4. APPROXIMATE INFERENCE AND PREDICTION

For the GP-WA model the likelihood is effectively Gaussian, hence, inference is analytical tractable [1]. However, neither the GP-TG model nor the GP-BE model have analytical tractable posterior distributions. Instead, we must resort to approximations. We consider two different approximate inference schemes—the Laplace approximation and expectation propagation (EP). Both methods approximate the posterior distribution $p(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta})$ with a single Gaussian $q(\mathbf{f})$. In the following we briefly give an overview of the two approximate inference schemes in relations to the bounded likelihood functions. For more details on the approximation schemes see for instance [2].

4.1. Laplace Approximation

Possibly, the simplest inference method is the Laplace approximation in which a multivariate Gaussian distribution is used to approximate the posterior, such that $p(\mathbf{f}|\mathcal{X}, \mathcal{Y}, \theta) \approx q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\hat{\mathbf{f}}, \mathbf{A}^{-1})$, where $\hat{\mathbf{f}}$ is the mode of the posterior and $\mathbf{A}$ is the Hessian of the negative log posterior at the mode. The mode is found as $\hat{\mathbf{f}} = \arg\max_{\mathbf{f}}\, p\left(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta}\right) = \arg\max_{\mathbf{f}}\, p\left(\mathcal{Y}|\mathbf{f}, \boldsymbol{\theta}_{\mathcal{L}}\right) p\left(\mathbf{f}, \mathcal{X}, \boldsymbol{\theta}_c\right)$. The general solution to the problem can be found by considering the un-normalized log posterior and the resulting cost function which is to be maximized, is given by

$$\psi\left(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta}\right) = \log p\left(\mathcal{Y}|\mathbf{f}, \mathcal{X}, \boldsymbol{\theta}_{\mathcal{L}}\right) - \frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f} - \frac{1}{2}\log|\mathbf{K}| - \frac{N}{2}\log 2\pi,$$

where $\mathbf{K}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)_{\boldsymbol{\theta}_c}$. The maximization can be solved with a standard Newton-step algorithm given by

$$\hat{\mathbf{f}}^{new} = \left(\mathbf{K}^{-1} + \mathbf{W}\right)^{-1} \cdot \left[\mathbf{W}\hat{\mathbf{f}} + \nabla \log p(\mathcal{Y}|\mathbf{f}, \mathcal{X}, \boldsymbol{\theta}_{\mathcal{L}})\right],$$

where the Hessian $\mathbf{W} = -\nabla\nabla_{\mathbf{f}} \log p(\mathcal{Y}|\mathbf{f})$ is diagonal with elements defined by the second derivative of the log-likelihood function $[\mathbf{W}]_{i,i} = -\frac{\partial^2 \log p(y_i|f_i)}{\partial f_i^2}$. When converged, the resulting approximation is

$$p\left(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta}\right) \approx \mathcal{N}\left(\mathbf{f}|\hat{\mathbf{f}}, \boldsymbol{\Sigma}\right),$$
$$\text{where } \boldsymbol{\Sigma} = \left(\mathbf{W} + \mathbf{K}^{-1}\right)^{-1}.$$

Approximating the posterior of $\mathbf{f}$ by the Laplace approximation requires the first two derivatives of the log likelihood. For the TG we will report the general derivatives applicable for any truncated likelihood function based on symmetric densities for which the truncated density can be written as the TG, i.e. in the form

$$p\left(y_i|f_i\right) = \frac{r\left(g\left(y_i|f_i\right)\right)}{s\left(g\left(b|f_i\right)\right) - s\left(g\left(a|f_i\right)\right)}, \tag{3}$$

where we for the TG model defines $g\left(c|f_i\right) = \nu\left(c - \mathrm{M}\left(f_i\right)\right)$. The resulting derivatives for the TG likelihood requires the following partial derivatives

$$\frac{\partial r\left(\cdot\right)}{\partial f_i} = \nu^2 g\left(y_i\right)\mathcal{N}\left(g\left(y_i\right)\right)\mathcal{N}\left(f_i\right),$$
$$\frac{\partial^2 r(\cdot)}{\partial^2 f_i} = \nu^2\mathcal{N}\left(g\left(y_i\right)\right)\mathcal{N}\left(f_i\right) \left[-\nu\mathcal{N}\left(f_i\right) + g\left(y_i\right)\left(\nu g\left(y_i\right)\mathcal{N}\left(f_i\right) - f_i\right)\right],$$
$$\frac{\partial s\left(\cdot\right)}{\partial f_i} = -\nu\mathcal{N}\left(g\left(b\right)\right)\mathcal{N}\left(f_i\right) \quad \text{and}$$
$$\frac{\partial^2 s(\cdot)}{\partial^2 f_i} = -\nu\mathcal{N}\left(g\left(b\right)\right)\mathcal{N}\left(f_i\right)\left[\nu g\left(b\right)\mathcal{N}\left(f_i\right) - f_i\right],$$

which enter into the derivatives of Eq. 3. The two required partial derivatives for the beta distribution are given by

$$\frac{\partial \log \text{Beta}(y_i|\cdot)}{\partial f_i} = \nu \cdot \mathcal{N}\left(f_i\right) \cdot \left[\log(y_i) - \log(1 - y_i) - \psi(\alpha) + \psi(\beta)\right] \quad \text{and}$$
$$\frac{\partial^2 \log \text{Beta}(y_i|\cdot)}{\partial f_i^2} = - \nu^2 \cdot \mathcal{N}\left(f_i\right) \cdot \Big[\mathcal{N}\left(f_i\right) \cdot \left(\psi^{(1)}(\alpha) + \psi^{(1)}(\beta)\right) + \frac{f_i}{\nu} \cdot \left(\log(y_i) - \log(1 - y_i) - \psi(\alpha) + \psi(\beta)\right)\Big],$$

where $\psi(\cdot)$ and $\psi^{(1)}(\cdot)$ are the digamma function of zero'th and first order, respectively.

4.2. Expectation Propagation

EP also approximates the posterior distribution with a single multivariate Gaussian distribution $q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ by factorizing the likelihood by n Gaussian factors $t_i(f_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\Sigma}_i) = \tilde{Z}_i \mathcal{N}(f_i|\tilde{\mu}_i, \tilde{\Sigma}_i)$, where $i = 1, ..., n$. The EP approximation to the full posterior is thus given by

$$\begin{aligned} p(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta}) \approx q(\mathbf{f}) &= \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= p(\mathbf{f}, \mathcal{X}) \mathcal{N}(\mathbf{f}|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}) \prod_{i=1}^{n} \tilde{Z}_i, \end{aligned}$$

where the means $\tilde{\mu}_i$ and variances $\tilde{\Sigma}_i$ have been collected into the vector $\tilde{\boldsymbol{\mu}}$ and diagonal matrix $\tilde{\boldsymbol{\Sigma}}$, respectively. The mean and covariance of the approximation are given by

$$\boldsymbol{\mu} = \boldsymbol{\Sigma} \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\mu}}, \qquad \boldsymbol{\Sigma} = \left(\mathbf{K}^{-1} + \tilde{\boldsymbol{\Sigma}}^{-1} \right)^{-1}.$$

EP updates each factor t_i in turn by first removing the factor to yield what is called the *cavity distribution* $q_{-i}(f_i) = \mathcal{N}(f_i|\mu_{-i}, \Sigma_{-i})$, where $\mu_{-i} = \Sigma_{-i}([\boldsymbol{\Sigma}]_{i,i}^{-1}\mu_i - \tilde{\Sigma}_i^{-1}\tilde{\mu}_i)$ and $\Sigma_{-i} = ([\boldsymbol{\Sigma}]_{i,i}^{-1} - \tilde{\Sigma}_i^{-1})^{-1}$. Secondly, the factor t_i is updated by projecting the cavity distribution multiplied with the true likelihood term onto a univariate Gaussian. The projection is effectively done by solving the following three integrals

$$\begin{aligned} Z_i &= \int p(y_i|f_i) \mathcal{N}(f_i|\mu_{-i}, \Sigma_{-i}) df_i, && (4) \\ \frac{dZ_i}{d\mu_{-i}} &= \frac{d}{d\mu_{-i}} \int p(y_i|f_i) \mathcal{N}(f_i|\mu_{-i}, \Sigma_{-i}) df_i \\ &= \int p(y_i|f_i) \frac{d}{d\mu_{-i}} \left\{ \mathcal{N}(f_i|\mu_{-i}, \Sigma_{-i}) \right\} df_i, && (5) \\ \frac{d^2 Z_i}{d\mu_{-i}^2} &= \frac{d^2}{d\mu_{-i}^2} \int p(y_i|f_i) \mathcal{N}(f_i|\mu_{-i}, \Sigma_{-i}) df_i \\ &= \int p(y_i|f_i) \frac{d^2}{d\mu_{-i}^2} \left\{ \mathcal{N}(f_i|\mu_{-i}, \Sigma_{-i}) \right\} df_i. && (6) \end{aligned}$$

Neither the beta likelihood nor the TG likelihood yield analytical tractable solutions for these three integrals, but the one-dimensional integrals can be solved numerically for the EP inference.

4.3. Predictive Distributions

Naturally, we want to predict future values of both the latent functional value f^* and data label y^*. For all models the posterior distribution over $\mathbf{f}$ is effectively Gaussian[2]. Hence, the predictive distribution $p(f^*|\mathcal{Y}, \mathcal{X}, \mathbf{x}^*) = \mathcal{N}(f^*|\mu^*, \sigma_*^2)$ of latent functional values is Gaussian and is derived just as in the standard cases in a straight forward manner (see e.g. [2, Chapter 2-3]).

The predictive distribution of future targets $p(y^*|\mathcal{Y}, \mathcal{X}, \mathbf{x}^*)$ involves computing the integral

$$p(y^*|\mathcal{Y}, \mathcal{X}, \mathbf{x}^*) = \int p(y^*|f^*) \mathcal{N}(f^*|\mu^*, \sigma_*^2) df^*.$$

For the GP-WA, the predictive distribution has a closed-form solution [1]

$$p_{\text{GP-WA}}(y^*|\mathcal{Y}, \mathcal{X}, \mathbf{x}^*) = \frac{\mathcal{N}(\Phi^{-1}(y^*)|\mu^*, \sigma_*^2)}{\Phi(\Phi^{-1}(y^*))}.$$

In case of the GP-BE and GP-TG the predictive distribution is not given in closed form. Instead, the integral must be computed using numerical methods. Predictions of the mean, $\mathbb{E}(y) \in \,]\,0;1\,[$, are in the bounded case given by

$$\begin{aligned} \mathbb{E}_{p(y^*|\cdot)}\{y^*\} &= \int_0^1 y^* p(y^*|\mathcal{Y}, \mathcal{X}, \mathbf{x}^*) dy^* && (7) \\ &= \int \mathcal{N}(f^*|\mu^*, \sigma_*^2) \int_0^1 y^* p(y^*|f^*) dy^* df^* \\ &= \int \mathcal{N}(f^*|\mu^*, \sigma_*^2) \mathbb{E}_{p(y^*|f^*)}\{y^*\} df^*. && (8) \end{aligned}$$

Given the cumulative Gaussian warping, Eq. 7 can be solved analytically for the GP-WA model. In Eq. 8 the mean of the likelihood occurs, which in the beta case is parameterized by a cumulative Gaussian and given the specific choice of warping this results in a closed form solution expressed by[3]

$$\mathbb{E}_{\text{GP-WA}}\{y^*\} = \mathbb{E}_{\text{GP-BE}}\{y^*\} = \Phi\left(\frac{\mu^*}{\sqrt{1 + (\sigma^*)^2}} \right).$$

In case of the GP-TG model, Eq. 7 has no analytical form and must be solved by one-dimensional numerical approximation.

5. SIMULATION EXAMPLE

In order to illustrate the difference between the warped and bounded likelihood approaches we consider an artificial example with added noise. It is generated by drawing a one-dimensional function from a zero-mean Gaussian process with a squared exponential (SE) kernel with length scale, $\sigma_l = 1$, and noise variance $\sigma_f = \exp(1)$. Three different types of noise are then added: The first type (WA) is i.i.d Gaussian noise added directly on f and transformed through $\Phi(\cdot)$ which corresponds to the noise assumption in the warped

[2] For the warped GP the posterior is exactly Gaussian, whereas we for the two other models have approximated—either by Laplace or EP—the posterior with a Gaussian.

[3] Keep in mind that although there is an equal sign between the predictive mean of the cumulative-warped and the beta model, the means will in general be different due to difference in the *latent* predictive distributions of the GP.

	Squared Exponential ($\sigma_f^2 = 2$, $\ell = 1$)		
	WA	TG	BE
GP-WA	**-129.8** (6.4)	-82.0 (7.3)	-165.3 (31.1)
GP-TG	-91.0 (19.6)	**-96.8 (4.5)**	-81.8 (14.5)
GP-BE	-119.8 (7.7)	-91.2 (6.3)	**-195.2 (24.6)**
	Periodic ($\sigma_f^2 = 3$, $\ell = 0.8$, $\lambda = 5$)		
	WA	TG	BE
GP-WA	**-93.2 (6.9)**	-80.6 (11.0)	-70.8 (10.4)
GP-TG	-76.2 (10.3)	**-91.6 (9.4)**	-66.0 (12.9)
GP-BE	-88.5 (3.8)	-84.5 (7.8)	**-99.8 (15.5)**

Table 1. Expected predictive negative log likelihood (and standard deviation) for each of the three models (GP-WP, GP-TG, GP-BE)) evaluated on a specific function with additive noise from ten random realizations of the noise for each corresponding noise types: WA, BE and TG. The noise free function is drawn from a GP prior with the indicated covariance functions and parameter values (defined in [7])

GP. In the second case (TG), f is transformed through $\Phi(\cdot)$ before adding noise based on the mode-parameterized TG distribution, thus corresponding to the noise assumption of the TG likelihood. In the third case (BE), we add noise based on the mean-parameterized beta distribution.

In order to visualize the special nature of bounded responses and the difference between the models, we have illustrated the WA noise case in Fig. 3, where all three bounded models are evaluated. Both the Laplace approximation and EP have been used for inference for the beta and TG model. The hyper-parameters are in all cases optimized using evidence maximization. The main difference of the three models occurs at the domain boundaries, where the GP-WA model concentrates the entire mass almost at the boundary. The predictive distribution of the GP-TG model generally has a similar shape over the entire domain with its mean always spaced significantly far from the boundary, whereas the GP-BE can also have its mean very close to the boundary as for the GP-WA model, but still retain mass away from the boundary. No significant differences between the two inference schemes are evident. Since the EP scheme requires numerical solutions to the integrals in Eq. 4-6, the Laplace approximation will be used in the reminder of this article.

We evaluate the ability of the models to model different noise distributions by comparing the predictive log likelihood for the previously mentioned dataset based on the Laplace approximation. A second example is added in which the function is drawn from a GP with a periodic covariance function. The predictive log likelihood for both examples is reported in Tab. 5 and is the average over ten realizations of the noise. As expected, we see that the model corresponding to the added noise type always results in the lowest negative likelihood, indicating a better model fit.

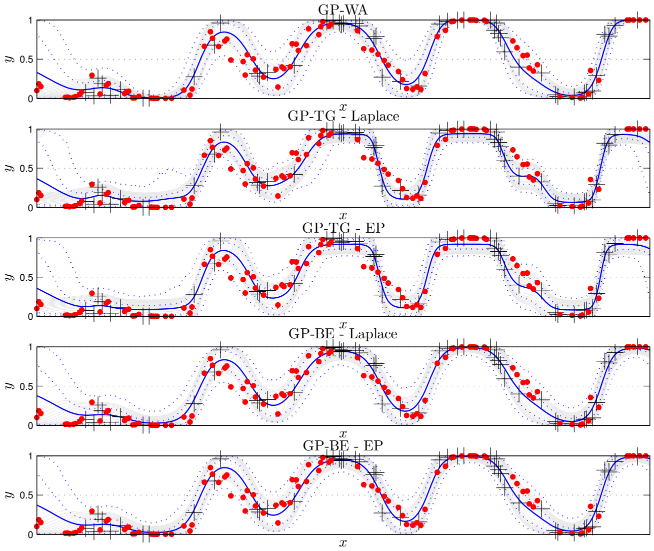

Fig. 3. Predictive distributions for the three models: GP-WA, GP-TG and GP-BE. For GP-TG and GP-BE both Laplace and EP inference are shown, where training data: $+$, test examples: $\cdot$, predictive mean: — and 68% and 95% percentiles: $\cdots$. Also, contours of the predictive distribution are shown in gray, where the intensity reflects probability mass concentration.

6. PERCEPTUAL AUDIO EVALUATIONS

In order to demonstrate the difference between the three considered models in a real-world scenario, we have tested the three models on two data sets consisting of subjective ratings performed while listening to audio through a hearing aid (HA) compressor with different settings.

The first dataset [8], HA-I, contains six compression ratio settings (including one without compression) and three release-time settings. This results in sixteen non-trivial combinations of settings that are rated three times by each of the seven test subjects while listening to a speech signal. The dataset also contains the audiogram of the hearing impaired test subjects. The audio signal resulting from each compressor setting is represented by standard audio features, namely thirty Mel frequency cepstral coefficients (MFCC). Thus, for one setting, s, each test subject, ts, rated the audio signal, a. This results in a collection of inputs for this specific rating which we collect in $\mathbf{x} = \{\mathbf{x}^{ts}, \mathbf{x}^{a}, \mathbf{x}^{s}\}$. We use the multi-task kernel formulation [9] and define the covariance function as $k\left(\mathbf{x}_i, \mathbf{x}_j\right) = k^u_{\text{ARD}}\left(\mathbf{x}^u_i, \mathbf{x}^u_j\right)\left(k^a\left(\mathbf{x}^a_i, \mathbf{x}^a_j\right) + k^s\left(\mathbf{x}^s_i, \mathbf{x}^s_j\right)\right)$ where all covariance functions are squared exponentials, the first one with automatic relevance determination (ARD).

The second dataset [10], HA-II, contains three input parameters related to the compression ratio, attack time and release time of a HA dynamic range compressor. Four subjects have rated 50 combinations of inputs in relation to general preference while listening to a speech-in-background-noise

		GP-WA	GP-TG	GP-BE
HA-I	-log $p(y^*)$	-66.1	-96.1	**-101.2**
	MSE	0.013	0.001	0.010
HA-II	-log $p(y^*)$	-7.7	-9.3	**-14.1**
	MSE	0.031	0.030	0.035

Table 2. **HA-I** Mean square error (MSE) and expected predictive negative log likelihood over 10 random sets. We find a significant difference in log likelihood at the 5% level between GP-TG and the two other models but not between GP-TG and GP-BE. For MSE the only significant difference is between GP-TG and GP-BE. **HA-II** Mean square error (MSE) and negative log likelihood over 10 folds. Considering the negative log likelihood only the GP-BE is significantly better than the GP-WA in a paired t-test. There is no significant difference between GP-TG and the other models. The GP-BE is significantly different in terms of MSE than the two others.

signal. The dataset does not contain any data describing the subjects, hence we use only one squared exponential covariance function.

We initialize the hyper-parameters in the (common) covariance function to the same value for all models, but initialize the likelihood noise parameter with multiple values in a grid pattern after which all the hyper-parameters are optimized using evidence maximization. We then report the performance of the model which yields the largest evidence after maximization. For the purpose of comparing the three models, we will simply consider the Laplace approximation and a retest scenario in which we train on a random repetition and test on another repetition for each setting. We repeat this three times and evaluate the resulting predictive likelihood and mean square error (MSE). The results are listed in Tab. 6. We note from the negative predictive log likelihood that the beta distribution provides a better fit to the noise compared to the other two models given the two real-world datasets presented here.

7. DISCUSSION AND CONCLUSION

In the present work, we outlined two bounded likelihood functions for bounded Gaussian process regression which in contrast to previous work make explicit assumptions about the noise in the bounded observation space. In the two considered examples we found the beta model to be better than the two other models in terms of the predictive log likelihood. These results together with the artificial examples support the application of all three models in the non-parametric Gaussian process framework. However, the optimal model obviously depends on the actual noise distribution in a given application. We therefore foresee addition and inclusion of other noise models based on other distribution with finite support.

Likelihood-model implementations are available [11] for use in the `gpml` toolbox [7] and can easily be extended to support more advanced link functions [12], which will make the models (both the bounded and the warped) even more flexible. In particular, we suggest to use a mixture of cumulative Gaussian link functions which do not complicate predictions significantly. Furthermore, we suggest to evaluate the performance of the deterministic approximations by the use of MCMC-sampling methods.

In conclusion, we have extended the Gaussian process framework to include bounded likelihood functions allowing for explicit specification of the likelihood model in applications where bounded observations are present and support an explicit noise model.

8. REFERENCES

[1] E. Snelson, C. E. Rasmussen, and Z. Ghahramani, “Warped Gaussian Processes,” in *Advances in Neural Information Processing Systems*, vol. 16. MIT Press, 2004.

[2] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[3] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, vol. 1, Wiley, 2nd edition, 1994.

[4] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, vol. 2, Wiley, 2nd edition, 1995.

[5] S. Ferrari and F. Cribari-Neto, “Beta Regression for Modelling Rates and Proportions,” *Journal of Applied Statistics*, vol. 31, no. 7, pp. 799–815, Aug. 2004.

[6] M Smithsen and J Verkuilen, “A Better Lemonsqueezer? Maximum Likelihood Regression with Beta-distributed Dependent Variables.,” *Australian Journal Psychology*, vol. 57, pp. 98–98, 2005.

[7] C. E. Rasmussen and H. Nickisch, “Matlab gpml toolbox,” 2010.

[8] E. Schmidt, *Hearing Aid Processing of Loud Speech and Noise Signals: Consequences for Loudness Perception and Listening Comfort.*, Ph.D. thesis, Technical University of Denmark, 2006.

[9] E. Bonilla, K. M. Chai, and C. Williams, “Multi-task gaussian process prediction,” in *Advances in Neural Information Processing Systems*, vol. 20, pp. 153–160. MIT Press, 2008.

[10] J. B. Nielsen, “M.sc. thesis, preference based personalization of hearing aids,” 2010.

[11] J. B. Nielsen and B. S. Jensen, “Bounded Gaussian Process Regression - Supplementary Material,” 2013.

[12] T. C. Martins Dias and C. A. R. Diniz, “The use of Several Link Functions on a Beta Regression Model: a Bayesian Approach.,” *AIP Conference Proceedings*, vol. 1073, no. 1, pp. 144, 2008.