Gaussian Process Regression with Censored Data Using Expectation Propagation

Perry Groot, Peter Lucas
Radboud University Nijmegen, the Netherlands
{perry,peterl}cs.ru.nl

Abstract

Censoring is a typical problem of data gathering and recording. Specialized techniques are needed to deal with censored (regression) data. Gaussian processes are Bayesian non-parametric models that provide state-of-the-art performance in regression tasks. In this paper we propose an extension of Gaussian process regression models to data in which some observations are subject to censoring. Since the model is not analytically tractable we use Expectation propagation to perform approximate inference on it.

1 Introduction

The method of data gathering and recording is a central issue in data analysis. A typical practical problem is censoring, which occurs when the value of a measurement or observation is only partially known. For example, censoring occurs when monitoring university freshmen in order to predict the drop out after the first semester. Another example of censoring occurs when a value occurs outside the range of a measuring instrument like a scale with a fixed limit. Censoring is also typical in clinical studies when individuals may withdraw from the study prematurely. Censored regression models were first suggested in Econometrics by Tobin (1958) and have since been a topic of active research (DeMaris, 2004; Greene, 2012; Wooldridge, 2002).

Censored regression models come in a variety of different forms: parametric, semi-parametric, and non-parametric. The earliest models were mostly parametric whereas nowadays interest is in the development of semi- and non-parametric models. The parametric modeling strategy is often not applicable since econometric data sets often have non-linear relationships and the underlying data generating mechanism is not precisely known. Gaussian processes (GPs) (Rasmussen and Williams, 2006) are Bayesian non-parametric graphical models that provide state-of-the-art performance in regression tasks. Gaussian processes are an extension of multivariate Gaussian random variables to infinite index sets that define a probability distribution over functions and inference takes place directly in the space of functions. An interesting property of Gaussian processes is that they provide a full predictive distribution, i.e., a mean estimate and an uncertainty estimate for the mean.

The contribution of this paper is an extension of Gaussian process regression models to data in which some observations are subject to censoring. Since the model is not analytically tractable we use an approximation for computing the posterior. In (Ertin, 2007) a Laplace approximation was considered. Since the censored Gaussian process model is, however, closely related to the probit model for which Expectation propagation was found to be the method of choice (Nickisch and Rasmussen, 2008) we use Expectation propagation to perform approximate inference on the model. As the model satisfies integrability with respect to Gaussian measures all necessary derivations for EP can be done analytically leading to a numerically stable algorithm. The method is validated on a benchmark data set. The rest of the paper is structured as follows. Section 2 describes the Bayesian framework. Section 3 describes experimental results. Section 4 describes related work. Section 5 gives conclusions.

2 Bayesian framework

In the following we denote vectors $\boldsymbol{x}$ and matrices $\boldsymbol{K}$ with bold-face type and their components with regular type, i.e., x_i , K_{ij}. With $\boldsymbol{x}^T$ we denote the transpose of the vector $\boldsymbol{x}$. Let $\boldsymbol{x} \in \mathbb{R}^D$ be a vector of explanatory variables, $y^* \in \mathbb{R}$ a response variable. We consider the problem of learning a function $f : \mathbb{R}^D \to \mathbb{R}$ from a set $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ of n observations, where y is a censored version of y^*:

$$y = \begin{cases} l & \text{if } y^* \leq l \\ y^* & \text{if } l < y^* < u \\ u & \text{if } y^* \geq u \end{cases} \tag{1}$$

where $l, u \in \mathbb{R}$ with $l < u$, are lower and upper thresholds, respectively. Hence, values in a certain range are transformed to (or reported as) a single value. Censoring represents a limitation in the response variable y^* of interest, but the explanatory variable x is assumed to be fully observable.[1] When $y = l$ we only know a lower bound on the target value $y^* \in [l, \infty)$ which is referred to as being *right censored.* When $y = u$ we only know an upper bound on the target value $y^* \in (-\infty, u]$ which is referred to as being *left censored.*

2.1 Prior

We assume that the latent values $y^* = f(\boldsymbol{x})$ are the realization of a zero-mean Gaussian process (Rasmussen and Williams, 2006), which can then be fully specified by the covariance matrix. The covariance matrix can be specified in terms of a kernel function $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. An often used kernel function is the Gaussian kernel

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D} \frac{(x_d - x'_d)^2}{\ell_d^2}\right) \tag{2}$$

where x_d denotes the d-th element of $\boldsymbol{x}$, and $\boldsymbol{\theta} = \{\sigma_f^2, \ell_1, \ldots, \ell_D\}$ hyperparameters with σ_f^2 specifying the signal variance and ℓ_d specifying how much the function can vary in the d-th element of the explanatory variables. For brevity we often omit the dependence on $\boldsymbol{\theta}$ in formulas. The kernel function leads to a multivariate Gaussian prior of latent function values $\{f(\boldsymbol{x}_i)\}$

$$p(\boldsymbol{f}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{K}|^{\frac{D}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{f}^T \boldsymbol{K}^{-1} \boldsymbol{f}\right) \tag{3}$$

where $\boldsymbol{K}$ is the $n \times n$ covariance matrix with $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

2.2 Likelihood

To capture the relations in (1) we can define an ideal likelihood for noise-free cases:

$$p_{id}(y|f(x)) = \begin{cases} 1 & \text{if } y = l, f(x) \leq l, \\ & \text{or } l < y = f(x) < u, \\ & \text{or } y = u, f(x) \geq u \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

To tolerate noise in the explanatory variables or response variables we assume that the unobserved latent function values are contaminated with noise and that these noisy values are then censored. A common choice is Gaussian noise with zero mean and unknown variance σ^2.[2] We use $\mathcal{N}(\delta|\mu, \sigma^2)$ to denote a Gaussian random variable δ with mean μ and variance σ^2 and use $\phi(\cdot)$, $\Phi(\cdot)$ to denote the standard normal density and cdf, respectively. Under these assumptions the following equations hold

$$\begin{aligned} p(y = l|f) &= \int p_{id}(y = l|f + \delta)\mathcal{N}(\delta|0, \sigma^2)\, \mathrm{d}\delta \\ &= \Phi\left(\frac{l - f}{\sigma}\right) = 1 - \Phi\left(\frac{f - l}{\sigma}\right) \\ p(y = y^*|f) &= \mathcal{N}(y|f, \sigma^2) = \frac{1}{\sigma}\phi\left(\frac{y - f}{\sigma}\right) \\ p(y = u|f) &= \Phi\left(\frac{f - u}{\sigma}\right) \end{aligned} \tag{5}$$

[1] In contrast, if the explanatory variable is not observable, i.e., missing, the data is said to be truncated.

[2] In principle, one could assume any distribution for the noise on the latent functions. Another often used assumption is heteroscedastic Gaussian noise.

The likelihood therefore becomes a mixture of Gaussian and probit likelihood terms:

$$L = \prod_{i=1}^{n} p(y_i|f_i) = \prod_{y_i=l} \left[1 - \Phi\left(\frac{f_i - l}{\sigma}\right)\right] \prod_{l<y_i<u} \left[\frac{1}{\sigma}\phi\left(\frac{y_i - f_i}{\sigma}\right)\right] \prod_{y_i=u} \left[\Phi\left(\frac{f_i - u}{\sigma}\right)\right] \tag{6}$$

This likelihood is well known in the literature as a Tobit likelihood (DeMaris, 2004; Greene, 2012), or a type I Tobit model according to the taxonomy of (Amemiya, 1984). The model can nowadays be estimated by many software packages, however, these are typically limited by a *parametric, linear form.*

2.3 Posterior

The posterior distribution over the latent variables is given by Bayes' theorem as a product of a normalization term, the prior, and the likelihood

$$p(\boldsymbol{f}|X, y) = \frac{1}{Z} p(\boldsymbol{f}|X) \prod_{i=1}^{n} p(y_i|f_i) \tag{7}$$

where the normalization term is called the evidence or marginal likelihood

$$Z = p(y|X) = \int p(\boldsymbol{f}|X) \prod_{i=1}^{n} p(y_i|f_i)\, \mathrm{d}\boldsymbol{f} \tag{8}$$

The Tobit likelihood in (6) makes the posterior in (7) analytically intractable and one has to resort to approximation techniques for computing the posterior. Below we describe the Expectation propagation method.

2.3.1 Expectation Propagation

Expectation propagation (EP) (Minka, 2001) is a deterministic approximate inference method that tackles the non-analytic nature of the posterior in (7) by approximating the likelihood by a local likelihood approximation using an unnormalized Gaussian function in the latent variable f_i

$$\begin{aligned} p(y_i|f_i) &\simeq t_i(f_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2) \\ &= \tilde{Z}_i \mathcal{N}(f_i|\tilde{\mu}_i, \tilde{\sigma}_i^2) \end{aligned} \tag{9}$$

where t_i is called the i-th *site* with site parameters $\tilde{Z}_i$, $\tilde{\mu}_i$, and $\tilde{\sigma}_i^2$, which we often abbreviate as $t_i(f_i)$ or simply t_i.

The product of the n likelihood approximations t_i can be expressed as

$$\prod_{i=1}^{n} t_i(f_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2) = \mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}) \prod_{i=1}^{n} \tilde{Z}_i \tag{10}$$

where $\tilde{\boldsymbol{\mu}}$ is the vector of $\tilde{\mu}_i$ and $\tilde{\boldsymbol{\Sigma}}$ is diagonal with $\tilde{\Sigma}_{ii} = \tilde{\sigma}_i^2$. The posterior $p(\boldsymbol{f}|\mathcal{D})$ is approximated by $q(\boldsymbol{f}|\mathcal{D})$ where

$$\begin{aligned} q(\boldsymbol{f}|\mathcal{D}) &= \frac{1}{Z_{EP}} p(\boldsymbol{f}) \prod_{i=1}^{n} t_i(f_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2) \\ &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \boldsymbol{\mu} &= \boldsymbol{\Sigma}\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}} \\ \boldsymbol{\Sigma} &= (\boldsymbol{K}^{-1} + \tilde{\boldsymbol{\Sigma}}^{-1})^{-1} \end{aligned} \tag{11}$$

using the product rule of Gaussian distributions and where Z_{EP} is the approximation of the EP algorithm to the marginal likelihood in (8).

The EP algorithm iteratively updates the t_i approximations sequentially. This is done by removing the i-th term from the approximate posterior, resulting in a *cavity distribution*, which is then combined with the i-th exact likelihood term. Finally a Gaussian approximation to the non-Gaussian marginal is computed, which is then used to update t_i.

EP defines the cavity distribution $q_{\setminus i}$, which is a Gaussian distribution, as

$$\begin{aligned} q_{\setminus i} &\propto \int p(\boldsymbol{f}) \prod_{j \neq i} t_j(f_j) = \mathcal{N}(\mu_{\setminus i}, \sigma_{\setminus i}^2) \\ \mu_{\setminus i} &= \sigma_{\setminus i}^2(\sigma_i^{-2}\mu_i - \tilde{\sigma}_i^{-2}\tilde{\mu}_i) \\ \sigma_{\setminus i}^2 &= (\sigma_i^{-2} - \tilde{\sigma}_i^{-2})^{-1} \end{aligned} \tag{12}$$

We then continue to find the new unnormalized Gaussian marginal which best approximates the product of the cavity distribution and the exact likelihood

$$\hat{q}(f_i) = \hat{Z}_i \mathcal{N}(\hat{\mu}_i, \hat{\sigma}_i^2) \simeq q_{\setminus i} p(y_i|f_i) \tag{13}$$

When minimizing $KL(p(x)||q(x))$ in the case of a Gaussian distribution $q(x)$, the minimum is obtained by matching the zeroth, first, and second moment. The parameters of the local likelihood approximation t_i can be computed using

$$
\begin{aligned}
\tilde{\mu}_i &= \tilde{\sigma}_i^2(\hat{\sigma}_i^{-2}\hat{\mu}_i - \sigma_{\setminus i}^{-2}\mu_{\setminus i}) \\
\tilde{\sigma}_i^2 &= (\hat{\sigma}_i^{-2} - \sigma_{\setminus i}^{-2})^{-1} \\
\tilde{Z}_i &= \hat{Z}_i\sqrt{2\pi(\sigma_{\setminus i}^2 + \tilde{\sigma}_i^2)}\exp\left(\frac{1}{2}\frac{(\mu_{\setminus i} - \tilde{\mu}_i)^2}{(\sigma_{\setminus i}^2 + \tilde{\sigma}_i^2)}\right)
\end{aligned}
$$

where $\hat{Z}_i$, $\hat{\mu}_i$, and $\hat{\sigma}_i^2$ are the moments that minimize the KL divergence which are described in the next section for the Tobit likelihood.

2.3.2 Tobit Moments

The minimum of the KL divergence is obtained when the moments of the two Gaussian distributions match (Minka, 2001). We thus need to compute the zeroth, first, and second order moments

$$
\begin{aligned}
\hat{Z}_i &= \int q_{\setminus i} t_i \,\mathrm{d}f_i \\
\hat{\mu}_i &= \mathbf{E}_q[f_i] \\
\hat{\sigma}_i^2 &= \mathbf{E}_q[(f_i - \mathbf{E}_q[f_i])^2]
\end{aligned}
\tag{14}
$$

where $q = \hat{Z}_i^{-1} q_{\setminus i} t_i$, which in the case of the Tobit likelihood in (6) can be computed analytically (cf. (Rasmussen and Williams, 2006)). The moments depend on whether the observation is left censored, right censored, or non-censored. Let $z_i^m = \frac{\mu_{\setminus i} - m}{\sqrt{\sigma^2 + \sigma_{\setminus i}^2}}$ for $m \in \{l, u\}$. If $y_i = l$ then

$$
\begin{aligned}
\hat{Z}_i &= 1 - \Phi\left(z_i^l\right) \\
\hat{\mu}_i &= \mu_{\setminus i} + \frac{\sigma_{\setminus i}^2 \mathcal{N}(z_i^l)}{(1 - \Phi(z_i^l))\sqrt{\sigma^2 + \sigma_{\setminus i}^2}} \\
\hat{\sigma}_i^2 &= \hat{Z}_i^{-1}[(\sigma_{\setminus i}^2 + \mu_{\setminus i}^2) - \hat{Z}_{i_{(p)}}(\hat{\sigma}_{i_{(p)}}^2 + \hat{\mu}_{i_{(p)}}^2)] - \hat{\mu}_i^2
\end{aligned}
$$

where $\hat{Z}_{i_{(p)}}$, $\hat{\sigma}_{i_{(p)}}^2$, $\hat{\mu}_{i_{(p)}}^2$ are the zeroth, first, and second order moments of a probit likelihood $\Phi(\frac{f-l}{\sigma})$. If $l < y_i < u$ then

$$
\begin{aligned}
\hat{Z}_i &= \frac{1}{\sqrt{2\pi(\sigma^2 + \sigma_{\setminus i}^2)}}\exp\left(-\frac{1}{2}\frac{(y_i - \mu_{\setminus i})^2}{(\sigma^2 + \sigma_{\setminus i}^2)}\right) \\
\hat{\mu}_i &= \mu_{\setminus i} + \sigma_{\setminus i}^2\left(\frac{y_i - \mu_{\setminus i}}{\sigma^2 + \sigma_{\setminus i}^2}\right) \\
\hat{\sigma}_i^2 &= \sigma_{\setminus i}^2 - \sigma_{\setminus i}^4\frac{1}{(\sigma^2 + \sigma_{\setminus i}^2)}
\end{aligned}
$$

If $y_i = u$ then

$$
\begin{aligned}
\hat{Z}_i &= \Phi\left(z_i^u\right) \\
\hat{\mu}_i &= \mu_{\setminus i} + \frac{\sigma_{\setminus i}^2 \mathcal{N}(z_i^u)}{\Phi(z_i^u)\sqrt{\sigma^2 + \sigma_{\setminus i}^2}} \\
\hat{\sigma}_i^2 &= \sigma_{\setminus i}^2 - \frac{\sigma_{\setminus i}^4 \mathcal{N}(z_i^u)}{(\sigma^2 + \sigma_{\setminus i}^2)\Phi(z_i^u)}\left(z_i^u + \frac{\mathcal{N}(z_i^u)}{\Phi(z_i^u)}\right)
\end{aligned}
$$

2.4 Prediction

The approximate predictive distribution $q(f_*)$ given an input $\boldsymbol{x}_*$ results from the approximation to the posterior distribution given by EP (11) and can be written as

$$
\begin{aligned}
q(f_*) &= \int p(f_* | \boldsymbol{x}_*, \mathcal{D}, \boldsymbol{f}) q(\boldsymbol{f}) \,\mathrm{d}\boldsymbol{f} \\
&= \mathcal{N}(\mu_*, \sigma_*^2) \\
\mu_* &= \boldsymbol{k}_*^T(\boldsymbol{K} + \tilde{\boldsymbol{\Sigma}})^{-1}\tilde{\boldsymbol{\mu}} \\
\sigma_*^2 &= k(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_*^T(\boldsymbol{K} + \tilde{\boldsymbol{\Sigma}})^{-1}\boldsymbol{k}_*
\end{aligned}
\tag{15}
$$

where $\boldsymbol{k}_* = [k(\boldsymbol{x}_1, \boldsymbol{x}_*), \ldots, k(\boldsymbol{x}_n, \boldsymbol{x}_*)]^T$.

2.5 Marginal Likelihood

The approximation of the EP algorithm to the marginal likelihood in (8) is given by

$$
\begin{aligned}
p(\mathcal{D}|\boldsymbol{\theta}) &= \int p(\boldsymbol{f}|\boldsymbol{\theta}_1)\prod_{i=1}^{n} p(y_i|f_i, \boldsymbol{\theta}_2) \,\mathrm{d}\boldsymbol{f} \\
&\approx \int p(\boldsymbol{f}|\boldsymbol{\theta}_1)\prod_{i=1}^{n} t_i(f_i) \,\mathrm{d}\boldsymbol{f}
\end{aligned}
\tag{16}
$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$ are the hyperparameters with $\boldsymbol{\theta}_1$ the covariance hyperparameters and $\boldsymbol{\theta}_2$ the likelihood hyperparameters. The hyperparameters can be optimized by minimizing the negative log marginal likelihood using gradient descent where

$$\frac{\partial}{\partial \boldsymbol{\theta}_1} \log p(\mathcal{D}|\boldsymbol{\theta}) = -\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\theta}_1}\left(\log|\boldsymbol{\Sigma}+\boldsymbol{K}| + \boldsymbol{\mu}^T(\boldsymbol{\Sigma}+\boldsymbol{K})^{-1}\boldsymbol{\mu}\right)$$

$$\frac{\partial}{\partial \boldsymbol{\theta}_2} \log p(\mathcal{D}|\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}_2}\sum_{i=1}^{n} \log \hat{Z}_i$$

3 Experiments

In this section we compare the standard Gaussian process model with the Tobit Gaussian process model, which was implemented in the publicly available GPstuff toolbox (Vanhatalo et al., 2011) on both artificial data and real data. In order to compare the predictive performance of the models we cast the analysis as a ranking problem, which is an elegant way of dealing with the censoring of the data. Two predictions can be ordered if (1) both values are non-censored, (2) if the non-censored value of one is smaller than the left censored value of the other, (3) if the non-censored value of one is larger than the right censored value of the other, or (4) if the right censored value of one is larger than the left censored value of the other. These four possible ordering relations are illustrated in Figure 1.

For these reasons the *concordance index* or *c-index* is often used as a performance measure for comparing models with censored data (Harrell Jr., 2001). The c-index can be interpreted as the fraction of all pairs of inputs whose predicted values are correctly ordered among all inputs that can be ordered. The c-index can be written as

$$c(\mathcal{D}, \mathcal{G}, f) = \frac{1}{|\mathcal{E}|}\sum_{\mathcal{E}_{ij}} \mathbf{1}_{f(x_i)<f(x_j)} \qquad (17)$$

where $\mathcal{G} = (X, \mathcal{E})$ is the order graph with edges according to the four criteria given above, $|\mathcal{E}|$ is the number of edges in $\mathcal{E}$, and $\mathbf{1}_{x<y} = 1$ if $x < y$, 0 otherwise, is an indicator function.

The concordance index is a generalization of the Wilcoxon-Mann-Whitney statistics (Wilcoxon, 1945; Mann and Whitney, 1947) for which $c = 1$ indicates a perfect predictive model and $c = 0.5$ indicates a random predictive model and is thus an AUC-like metric for the scenario of censored data.

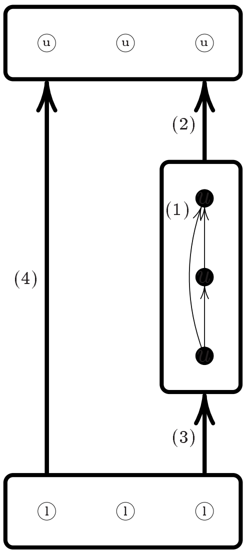

Figure 1: Four possible ordering relations when data is censored. The white/open circles represent censored data points and are labeled u when it is left censored or l when right censored. Black/closed dots represents non-censored data points. Arrows between (groups of) points represents which data points can possibly be ordered where the number refers to the enumeration used in the article.

3.1 Artificial data

To illustrate the behaviour of the standard GP model and the Tobit GP model, we first illustrate their behaviour on a simple 1-dimensional function. We created a data set of 30 equally spaced inputs in $[0, 1]$ and outputs using

$$f(x) = (6x - 2)^2 \sin(2(6x - 2)) \qquad (18)$$

which were contaminated with Gaussian noise with zero mean and variance 0.1. To censor 40% of the observations we calculated the 40th percentile of the sample distribution and used the corresponding value of $l = -0.2265$ as lower threshold. We trained a standard GP model and a Tobit GP model on the full data set of 30 samples including censored observations, as

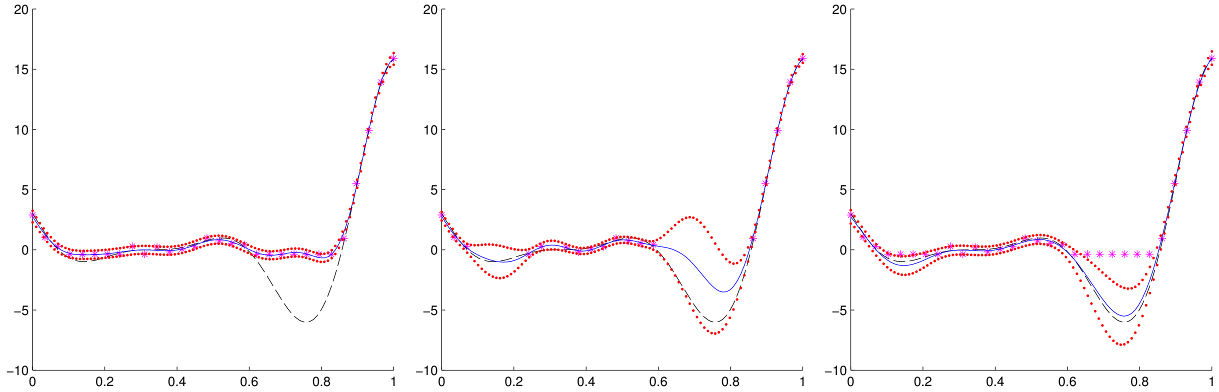

Figure 2: Regression with censored data. Left panel: standard GP model on data with censored samples. Middle panel: standard GP model on data where censored samples were removed. Right panel: Tobit GP model on data with censored samples. The $*$ denotes observations (possibly noisy and/or censored). Dashed line denotes the underlying true function. Solid line denotes the mean model prediction. Dotted line denotes two standard deviations from the mean.

well as a standard GP model on the data set of 18 samples of non-censored observations.

The results are shown in Figure 2. The left panel shows a standard GP model on the data set of 30 observations which clearly shows a bad predictive model of the underlying true function as it tries to match the mean prediction with the censored observations and also severely undervalues the model uncertainty in the censored areas. The middle panel shows a standard GP model on the data of 18 non-censored samples, which performs a bit better than the standard GP model on the full data set both in terms of the mean prediction and the uncertainty of the mean prediction. The model, however, throws out information and is not able to predict that the underlying true function lies strictly below the censoring threshold. The right panel shows the Tobit GP model which shows the best mean prediction, has a lower predictive uncertainty, and in the censored areas gives predictions that lie strictly below the censoring threshold.

We report the results of 10-fold crossvalidation of both the standard GP and Tobit GP in Table 1. As the number of censored observations is quite large, i.e., 40% of the data, we observe a large increase in the c-index for the Tobit GP model. Since we know the true underlying latent function we also report the root mean squared error

$$RMSE(t, f) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - f_i)^2} \qquad (19)$$

and the mean absolute error

$$MAE(t, f) = \frac{1}{n}\sum_{i=1}^{n}|t_i - f_i| \qquad (20)$$

in Table 1 for both the standard GP and Tobit GP models.

Table 1: Concordance results and error measures artificial data.

	GP	**Tobit GP**
c-index	0.73	0.95
RMSE	2.21	0.72
MAE	1.42	0.60

3.2 Housing data

We validated the Gaussian process Tobit model on the ‘housing’ benchmark data set obtained from the UCI machine learning repository. Originally the housing data set was used to investigate methodological issues related to the use of housing data to estimate the demand for clean air (Harrison and Rubinfeld, 1978). The data consists of 506 observations on 14 real-valued variables. The objective is to predict the median value of owner-occupied homes from the remaining variables. In (Gilley and Pace, 1996), the data has been checked against the original census data and it was discovered that the Census Bureau censored tracts whose median value was over \$50.000. Hence, all tracts with a median value greater than \$50.000 appear as \$50.000. In total 16 observations were censored.

We report the results of 10-fold crossvalidation of both the standard GP and Tobit GP in Table 2 averaged over 10 runs. Although the number of censored observations is quite small, i.e., only 3.2% of the data, we observe an increase in the c-index for the Tobit GP model. We also see an improvement in predictive performance for the Tobit GP model when using the Expectation algorithm (EP) instead of the Laplace approximation (LA) of (Ertin, 2007).

Table 2: Concordance results housing data (mean c-index and standard deviation).

method	**c-index**
GP	0.866 ± 0.003
Tobit-GP (LA)	0.879 ± 0.008
Tobit-GP (EP)	0.892 ± 0.007

4 Related work

In recent years researchers have striven to devise regression techniques that do not rely on the classical assumptions of parametric methods. In this paper we extended the Gaussian process framework, a principled non-parametric Bayesian framework for regressions tasks, to data subject to censoring. Since the model is no longer analytically tractable approximations are needed to compute the posterior. In this paper we considered the Expectation propagation algorithm, which was shown to outperform the Laplace approximation considered in (Ertin, 2007). In (Hutter et al., 2011) a different approach is followed to handle censored data. Hutter et al., uses a random forest, initialized using non-censored data only, which is iteratively improved using an EM algorithm. Their method samples data from the predictive distribution to fill in samples for censored observations. However, since the predictive distribution is not guaranteed to lie strictly below (or above) the censoring threshold (cf. Figure 2), additional countermeasures, like truncating the predictive distribution, are needed to ensure the data generated satisfies the censoring constraints.

In this paper we considered common lower and upper censoring thresholds for each data point, however, the model can easily be extended to allow for data points with individual lower and/or upper thresholds. This would allow the method to be used for the analysis of survival times (Shivaswamy et al., 2007) in which censoring is extremely common. The concordance index can still be used as a performance measure as given in Equation (17), but the situation depicted in Figure 1 changes. Since data points have individual censoring thresholds the ordering relations in Figure 1 do not longer hold and some may become incomparable (Shivaswamy et al., 2007).

In the likelihood model we assumed that noise was Gaussian distributed. If this assumption is violated this can have considerable impact on the predictive performance of the model (Greene, 2012). A more general approach would be to allow for heteroscedastic Gaussian distributed noise by using a second Gaussian process to model the noise process (Kersting et al., 2007). One could follow a similar EP approach as in (Muñoz-González et al., 2011), although with respect to this second Gaussian process the model would then no longer satisfy integrability with respect to Gaussian measures.

5 Conclusion

In this paper we have presented an extension to Gaussian process regression to handle data subject to censoring. To tackle the non-analytic nature of the posterior we used Expectation propagation to perform approximate inference on the model which was shown to outperform the Laplace approximation. Directions for further work include among others handling heteroscedastic noise, multi-variate Tobit models, and extending the model to other domains such as survival analysis.

The software was implemented in the publicly available GPstuff toolbox (Vanhatalo et al., 2011) and is freely available from the first authors website as well as some supplementary material describing implementation details.

References

T. Amemiya. 1984. Tobit models: A survey. *Journal of Econometrics*, 24(1-2):3–61.

A. DeMaris. 2004. *Regression With Social Data: Modeling Continuous and Limited Response Variables*. Wiley Series in Probability and Statistics. Wiley-Interscience.

E. Ertin. 2007. Gaussian process models for censored sensor readings. In *Proceedings of the 2007 IEEE/SP 14th Workshop on Statistical Signal Processing*, SSP '07, pages 665–669, Washington, DC, USA. IEEE Computer Society.

O. W. Gilley and R. K. Pace. 1996. On the Harrison and Rubinfeld data. *Journal of Environmental Economics and Management*, 31:403–405.

W. H. Greene. 2012. *Econometric Analysis*. Prentice Hall, 7th edition.

F. E. Harrell Jr. 2001. *Regression Modeling Strategies, With applications to linear models, logistic regression, and survival analysis*. Springer.

D. Harrison and D. L. Rubinfeld. 1978. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102.

F. Hutter, H. H. Hoos, and K. Leyton-Brown. 2011. Bayesian optimization with censored response data. In *NIPS workshop on Bayesian optimization, sequential experimental design, and bandits*.

K. Kersting, C. Plagemann, P. Pfaff, and W. Burgard. 2007. Most likely heteroscedastic gaussian process regression. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 393–400, New York, NY, USA. ACM.

H. B. Mann and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1):50–60.

T. Minka. 2001. *A family of approximation methods for approximate Bayesian inference*. Ph.D. thesis, MIT.

L. Muñoz-González, M. Lázaro-Gredilla, and A. R. Figueiras-Vidal. 2011. Heteroscedastic Gaussian process regression using expectation propagation. In *Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

H. Nickisch and C. E. Rasmussen. 2008. Approximations for Binary Gaussian Process Classification. *Journal of Machine Learning Research*, 9:2035–2078, October.

C. E. Rasmussen and C. K. I. Williams. 2006. *Gaussian processes for machine learning*. MIT Press, Cambridge, MA.

P. K. Shivaswamy, W. Chu, and M. Jansche. 2007. A support vector approach to censored targets. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pages 655–660, Washington, DC, USA. IEEE Computer Society.

J. Tobin. 1958. Estimation of relationships for limited dependent variables. *Econometrica*, 26:24–36.

J. Vanhatalo, J. Riihimäki, J. Hartikainen, and A. Vehtari. 2011. Bayesian modeling with Gaussian processes using the MATLAB toolbox GPstuff. submitted.

F. Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

J. M. Wooldridge. 2002. *Econometric Analysis of Cross Section and Panel Data*. Mit Press.