Prediction of Clinical Scores from Neuroimaging Data with Censored Likelihood Gaussian Processes

Anil Rao*†, Joao Monteiro *†, and Janaina Mourao-Miranda*†
*Department of Computer Science, University College London, London, U.K. Email: a.rao@ucl.ac.uk
†Max Planck University College London Centre for Computational Psychiatry and Ageing Research, University College London, London, UK

***Abstract*—In this paper, we explore the use of Censored Likelihoods in Gaussian Process Regression when predicting bounded clinical scores from neuroimaging data. The standard approach, which uses a Gaussian Likelihood, does not respect the fact that the clinical scores are bounded, and so may produce suboptimal models. Conversely, Censored Likelihoods explicitly model the restricted range of such clinical scores and carry this property through inference. We apply both the standard approach and the Censored Likelihood approach to the prediction of the MMSE score from structural MRI. Overall, we find small improvements in mean squared error when using the Censored Likelihood and in addition, the censored models are more favoured from a Bayesian perspective. We also discuss the qualitative nature of the predictions of the two approaches.**

***Keywords*—gaussian processes, clinical scores**

I. Introduction

There has been substantial interest in recent years in using multivariate regression models to predict clinical and psychometric scales from neuroimaging MRI [1].

One attractive framework for learning the predictive models for such measures is Gaussian Process Regression (GPR). Gaussian processes are flexible Bayesian methods for model estimation that have recently gained popularity for building predictive neuroimaging models for regression and classification [2], [3], [4]. The standard Gaussian Likelihood used for GPR , however, does not respect the fact that the clinical score y is often bounded below $y \in [a, \infty]$, above $y \in [\infty, b]$ or both $y \in [a, b]$, where $a, b \in \mathbb{R}$. This can mean that some of the attractive properties of Gaussian Processes, such as automatic hyperparameter estimation, become compromised resulting in sub-optimal models and poorer predictions.

In this work we explore the use of GPR with Censored Likelihoods when the clinical score we aim to predict is bounded. The use of Censored Likelihoods enables the modelling procedure to explicitly take into account the restricted range of the the clinical score during model building and inference. Using imaging data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, we show that the use of GPR with Censored Likelihoods can result in better models for the data from a Bayesian perspective, whilst also improving prediction accuracy.

II. Materials

The dataset consisted of the MP-RAGE images of 592 unique subjects from the ADNI database (adni.loni.usc.edu). The data was preprocessed using SPM12 (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/) by performing grey matter segmentation and group-wise registration with Dartel to a study-specific template. The aligned images were transformed to the 2mm MNI template and smoothed with a Gaussian kernel of 2mm FWHM. A mask was applied to select voxels that had a probability of being grey matter above 0.025, giving a set of images that provide the 157026 image features $\mathbf{x}$ in the matrix $\mathbf{X}$. In our experiments, the grey matter image features are used to predict the 'Mini-Mental State Examination' (MMSE) which is thus the target variable y. The MMSE is commonly used to diagnose and assess dementia and tests subject performance in areas such as arithmetic, comprehension, and basic motor skills. Subjects can achieve a minimum of zero and a maximum of thirty for the MMSE score, ie. $y \in [0, 30]$. Figure 1 shows the distribution of MMSE score over the 592 subjects used in our analysis, and we can see that a large proportion of the scores 'pile up' at the maximum possible score of 30.

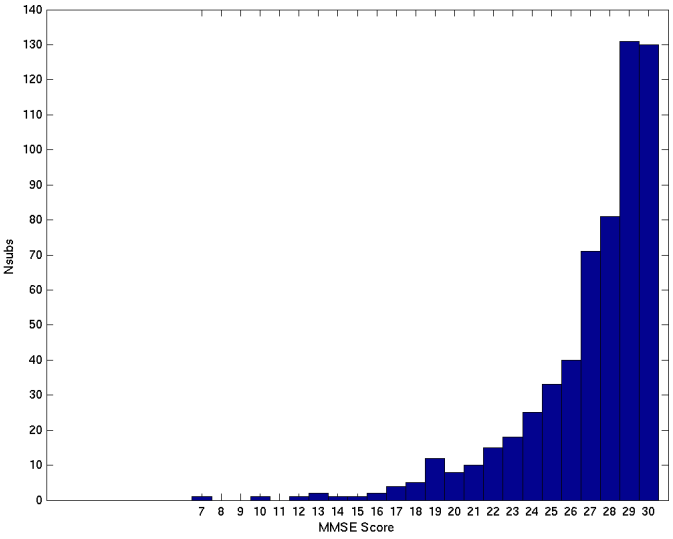

Fig. 1. This figure shows the distribution of the MMSE scores over the 592 subjects used from the ADNI database

III. Methods

A. Gaussian Process Regression

Given a set of n observations $\mathbf{x}_i \in \mathbb{R}^p$ with associated target variables y_i, Gaussian processes impose a multivariate Gaussian prior on a set of latent variables f_i, where the mean

and covariance of the prior are functions of the inputs $\mathbf{x}_i$ [5]:

$$\begin{aligned} E(f_i) &= m(\mathbf{x}_i) \\ Cov(f_i, f_j) &= k(\mathbf{x}_i, \mathbf{x}_j) \end{aligned} \tag{1}$$

We assume a zero mean function $m(\mathbf{x}_i) \equiv 0$ throughout this work. The covariance function $k(\mathbf{x}_i, \mathbf{x}_j)$, also referred to as the kernel function, describes how the values of the latent variables covary across the input space, and it will have a set of associated kernel parameters $\boldsymbol{\theta}$. We employ a linear covariance function with a bias term in our analysis:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i {\mathbf{x}_j}^T}{l^2} + b^2 \tag{2}$$

The posterior for the latent variables f_i is given by:

$$P(\mathbf{f} \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\sigma}) = \frac{1}{P(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\sigma})} \prod_{i=1}^{n} P(y_i \mid f_i, \boldsymbol{\sigma}) P(f_i \mid x_i) \tag{3}$$

where the likelihood function $P(y_i \mid f_i, \boldsymbol{\sigma})$ relates the observed targets to the latent variables, and has hyperparameters $\boldsymbol{\sigma}$. In order to perform inference, we firstly need to estimate the complete set of kernel and likelihood hyperparameters $\{\boldsymbol{\sigma}, \boldsymbol{\theta}\}$, and this is performed by maximizing the marginal likelihood $P(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\sigma})$ which is the normalizing factor in equation (3). Once we have the optimised hyperparameters $\{\boldsymbol{\theta}, \boldsymbol{\sigma}\}$, we can then perform inference for a set of test points $\mathbf{X}_*$ to derive the predictive distribution for the test targets $\mathbf{y}_*$. We now describe the Gaussian and Censored Likelihoods used for modelling in this paper and how inference proceeds in each case.

1) Gaussian Likelihood: The model underlying the Gaussian Likelihood is

$$y_i = f_i + \epsilon \tag{4}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $\sigma > 0$ is the standard deviation of the noise. This gives rise to the Gaussian Likelihood

$$P(y_i \mid f_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - f_i)^2}{2\sigma^2}} \quad , \forall y_i \in \mathbb{R} \tag{5}$$

For a Gaussian Likelihood, the predictive distribution of the target y_* of a test point $\mathbf{x}_*$, given the training data, has the closed form

$$\begin{aligned} y_* \mid \mathbf{X}, \mathbf{y}, \mathbf{x}_* &\sim \mathcal{N}(\bar{y}_*, Var(y_*)) \text{ where} \\ \bar{y}_* &= \mathbf{k}_*(K + \sigma^2 I)^{-1}\mathbf{y} \\ Var(y_*) &= k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*(K + \sigma^2 I)^{-1}\mathbf{k}_*^T + \sigma^2 \end{aligned} \tag{6}$$

in which K is the $n \times n$ matrix of training set covariances and $\mathbf{k}_*$ is the n-dimensional row vector of test-training covariances. Typically, the mean of the distribution, $\bar{y}_*$, is taken to be the prediction of the target at test point $\mathbf{x}_*$.

Note that the Gaussian Likelihood in equation (5) gives non-zero probabilities for every possible value of the target variable, given the latent variable. This means that even if the target variable y is bounded within $[a, b]$, this is not taken into account during inference. One consequence of this is that the mean predictions $\bar{y}_*$ may lie outside $[a, b]$. In this work, we ensure that all predictions using the Gaussian Likelihood lie within the appropriate range by taking the final predictions to be the closest value within $[a, b]$ to the mean predictions.

2) Left and Right Censored Likelihood: Here the targets y_i are explicitly bounded below ('right censoring') and above ('left censoring'), $y \in [a, b]$. The underlying model is

$$y_i = \begin{cases} a & \text{if } f_i + \epsilon < a \\ b & \text{if } f_i + \epsilon > b \\ f_i + \epsilon & \text{otherwise} \end{cases} \tag{7}$$

with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The likelihood for y_i is then

$$P(y_i \mid f_i) = \begin{cases} \Phi(\frac{a - f_i}{\sigma}) & \text{if } y_i = a \\ \Phi(\frac{f_i - b}{\sigma}) & \text{if } y_i = b \\ \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - f_i)^2}{2\sigma^2}} & \text{if } y_i \in (a, b) \end{cases} \tag{8}$$

where Φ is the cumulative distribution function for the normal distribution, $\Phi(z) = \int_{-\infty}^{z} \mathcal{N}(0, 1) dz$.

Comparing this with the Gaussian Likelihood in equation (5), we can see that the Censored Likelihood is identical to the Gaussian Likelihood for values of the target $y_i \in (a, b)$, but differs at the bounds a, b. The resulting likelihood also gives zero probabilities for values of the target y_i that lie outside $[a, b]$, ensuring that the range of the targets is explicitly included in the model and is carried through inference. Note that our main motivation for using a Censored Likelihood when predicting clinical scores is to respect their bounded nature: We are not using it to directly address any possible skewness in the distribution of target variables, such as that seen in figure 1, although skewness may naturally arise with bounded targets. Indeed, a Gaussian Likelihood may still be appropriate for a skewed distribution of non-bounded target variables.

Inference with a Censored Likelihood cannot be performed analytically as with the Gaussian Likelihood, and so approximations must be used. In this work we test both the Laplacian approximation for Censored Likelihoods, described in [6], and the Expectation Propogation (EP) approach from [7]. These methods are essentially different ways of approximating the posterior distribution of the latent variables given in equation (3), enabling the marginal likelihood to be optimized, and inference to be performed. In each case, this results in gaussian predictive distributions for the latent variables of the test points:

$$f_* \mid \mathbf{X}, x_*, y \sim \mathcal{N}(\mu_{f_*}, \sigma^2_{f_*}) \tag{9}$$

giving rise to the following predictive distribution for y_*:

$$P(y_* \mid \mathbf{X}, \mathbf{y}, \mathbf{x}_*) = \begin{cases} \Phi(\frac{a - \mu_{f_*}}{\sqrt{\sigma^2 + \sigma^2_{f_*}}}) & \text{if } y_* = a \\ \Phi(\frac{\mu_{f_*} - b}{\sqrt{\sigma^2 + \sigma^2_{f_*}}}) & \text{if } y_* = b \\ \frac{1}{\sqrt{2\pi(\sigma^2 + \sigma^2_{f_*})}} e^{-\frac{(y_* - \mu_{f_*})^2}{2(\sigma^2 + \sigma^2_{f_*})}} & \text{if } y_* \in (a, b) \end{cases} \tag{10}$$

The mean prediction $\bar{y}_*$ using the Censored Likelihood is then given by

$$\begin{aligned} \bar{y}_* &= a\Phi(z_1) + b\Phi(-z_2) \\ &+ (\Phi(z_2) - \Phi(z_1)) \left(\mu_{f_*} - \sigma_* \left(\frac{\phi(z_2) - \phi(z_1)}{\Phi(z_2) - \Phi(z_1)} \right) \right) \end{aligned} \tag{11}$$

in which $z_1 = \frac{a - u_{f_*}}{\sqrt{\sigma^2 + \sigma^2_{f_*}}}$, $z_2 = -\frac{u_{f_*} - b}{\sqrt{\sigma^2 + \sigma^2_{f_*}}}$, $\sigma_* = \sqrt{\sigma^2 + \sigma^2_{f_*}}$, and ϕ is the probability density function of the standard normal distribution.

B. Model Evaluation

We evaluate the different GPR models described in section III-A when predicting the MMSE score using the imaging features as predictors. In addition to those models, we also perform predictions with naive models that consist solely of a constant bias term, using both a Gaussian Likelihood and Censored Likelihood with EP inference. These models do not use the image features during inference and serve as reference models for the other approaches. Model performance was determined using 5-fold cross validation, repeated over 10 different splits of the data. In each case all features were standardised to mean zero and unit variance using the training folds only, and the predictions were rounded to the nearest whole number. We calculate the mean-squared error (MSE), the mean-absolute error (MAE), and Kendall's coefficient τ of each model for each of the 10 splits of the data, and then average them to give summary measures of predictive performance. In addition to the above metrics, which are calculated using only point estimates of the predictions, we determine extra measures that are often used when assessing Bayesian modelling approaches. Firstly, we calculate the mean negative log predictive density (MNLPD) for each model. The log predictive density (LPD) essentially describes how probable the real targets, ie. the MMSE scores, are, according to the predictive distribution given by a particular model [5] ie.

$$LPD = \log P(y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) \tag{12}$$

Unlike the MSE and MAE, the LPD takes into account the uncertainty in the predictive distribution. We take the negative log predictive density to turn this into a loss function, and average over each observation and split to give the summary measure MNLPD of how well the model has captured the predictive distribution of the data. Lower values of MNLPD will then correspond to better fits of the predictive distribution. Lastly, we determine the optimised Log Marginal Likelihood, $\log \mathcal{Z}$, using the whole dataset for training. Since $\log \mathcal{Z}$ is the optimized value we correct for the number of model hyperparameters to enable model comparison using the Bayesian Information Criterion:

$$\log \mathcal{Z}_{\text{BIC}} = \log \mathcal{Z} - \frac{p}{2}\log(n) \tag{13}$$

Here n is the size of the dataset, and p is the number of model hyperparameters (equal to the number of kernel parameters plus 1 likelihood parameter).

IV. Results and Discussion

TABLE I. Prediction of MMSE for the different models.

Model	MSE	MAE	τ	MNLPD	$\log \mathcal{Z}_{\text{BIC}}$
Bias Only (Gauss Lik)	11.86	2.51	-	2.66	-1578.4
Bias Only (Cens Lik)	11.86	2.51	-	2.43	-1443.7
Gauss Lik	7.06	1.88	0.48	2.42	-1457.5
Cens. Lik (Laplace)	6.94	1.88	0.48	2.20	-1328.8
Cens. Lik (EP)	6.88	1.86	0.48	2.20	-1327.4

Table I summarizes the predictive accuracies for the different models. Firstly, we can see that all the models that use the image features outperform the corresponding naive models that utilize a covariance function containing only a bias term. We can also see that the Censored Likelihood model with EP inference gives slightly more accurate predictions than the model using the standard Gaussian Likelihood, according to both the MSE and MAE. The Censored Likelihood model using Laplace inference gives a smaller improvement in MSE than the one using EP inference. The Kendall correlation coefficient appears similar for the different approaches. It is possible that the use of the more appropriate Censored Likelihood produces the improvements in MSE and MAE compared to the Gaussian likelihood model. We found that the censored models improved the MSE for all 10 splits of the data, providing further evidence that the slight improvements in accuracy were due to the modelling approach. As the Censored Likelihood with EP inference outperformed that with Laplace inference, we mostly focus on that approach in the following discussion.

Figure 2 shows the predicted MMSE scores against the true MMSE scores when using the Gaussian Likelihood, and the Censored Likelihood with EP inference, averaged over the 10 splits. The dashed red lines indicate predictions for a model that has zero test error. The first thing we can notice is that the predictions of subjects with low MMSE scores tend to be too high for all the models, as shown by the large number of points to the left of the red line. However, we can also observe subtle differences in the qualitative behaviour of the different approaches over the range of the targets. The principal apparent difference is that the predictions of the censored approach (bottom) do not appear to push predictions of high MMSE scores $\geq$ 27 as strongly to the maximal score of 30 as those with the Gaussian Likelihood (top). This can be further seen in table II, where we show the average errors over the 10 splits using the Gaussian Likelihood and Censored Likelihood (EP) within each quartile of the targets: $Q_1 : y \leq 26, (n = 179); Q_2 : 27 \leq y \leq 28, (n = 152); Q_3 : y = 29, (n = 131); Q_4 : y = 30, (n = 130)$. Here we can see that the 'softening' of predictions for subjects with high MMSE scores using the Censored Likelihood seems to have mostly affected the predictive accuracy of subjects with the highest MMSE score, as the Censored Likelihood gives smaller MSE and MAE for each quartile apart from Q_4.

TABLE II. Errors within each quartile for the different models.

	Q_1		Q_2		Q_3		Q_4	
Model	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Gauss Lik	14.09	2.64	3.51	1.47	3.62	1.45	4.99	1.75
Cens. Lik (EP)	13.67	2.63	3.20	1.34	3.27	1.28	5.48	1.99

In terms of the overall predictive distribution, the censored models perform better than the Gaussian Likelihood as shown by the lower MNLPD values for these approaches. Calculation of the Log Bayes Factors $\log K$ of pairs of models, given by the difference between the corresponding values of $\log \mathcal{Z}_{\text{BIC}}$, gives very strong support for both of the censored models over the Gaussian Likelihood model ($\log K = 129, 130$). Given that the Censored Likelihood explicitly restricts the range of the target variable to $[0, 30]$, it is perhaps not suprising that these models are so strongly favoured over those using the Gaussian Likelihood, according to the Bayes Factors and MNLPD.

For illustration, figure 3 shows the (unnormalized) weight images for the different approaches when training with the complete dataset. A positive weight at a voxel v indicates an increase in the prediction of the latent function (but not the predicted target variable for the Censored Likelihood models) as

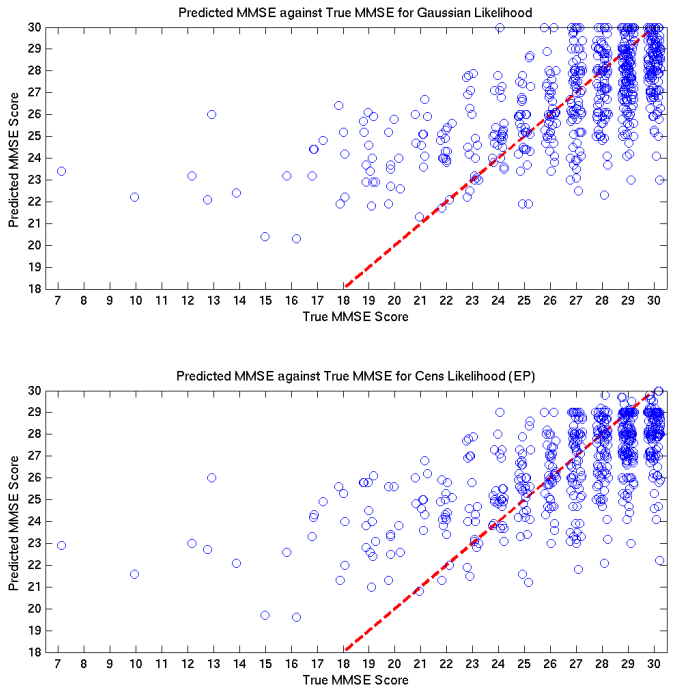

Fig. 2. This figure shows the predicted MMSE scores against the true MMSE scores for different models, averaged over 10 splits. Some jittering has been added to the true scores for visualisation purposes. The top shows the predictions using the Gaussian Likelihood, while in the bottom we see the predictions using the Censored Likelihood EP approach. The dashed red lines indicate predictions for a model that has zero test error.

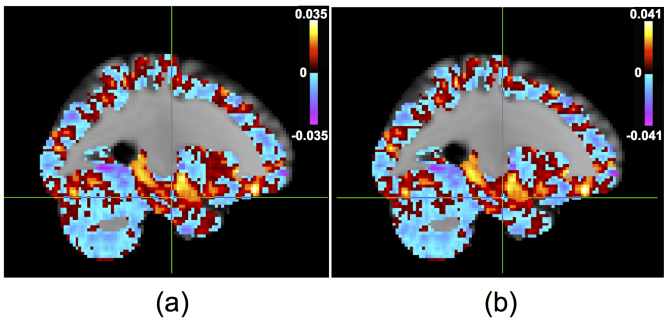

Fig. 3. This figure shows the weight images for each model when training using all 592 subjects. Positive weights are indicated with a hot colour and negative weights with cool colours. The Gaussian Likelihood model is shown on the right in (a), while (b) shows the model for the Censored Likelihood using EP inference.

the value of the image feature at v increases, holding the values of all other features constant. Visually, the weight vectors for the approaches are not substantially different, and all indicate large positive weights in the left hippocampus/amygdala which are proximal to the crosshair, positioned at (-26,-14,-22) in MNI space. However, the Censored Likelihood has a larger range of values in the weight vector, corresponding to a larger increase in the prediction of the latent function as the image feature values change.

V. Conclusion

In this work we have explored using Gaussian Process Regression with a Censored Likelihood when predicting clinical scores from neuroimaging data. We compared this approach to Gaussian Process Regression with a Gaussian Likelihood, which is the standard model used when predicting scores that have a large range. We found that the use of the Censored Likelihood with both Laplace and EP inference gave small improvements in the MSE compared to the standard model, although they did not appear to improve Kendalls τ correlation coefficient. The use of a Censored Likelihood also appeared to improve the overall predictive distribution for those models, and the Bayes factors gave strong supporting evidence that they were more appropriate for the data.

Whilst we saw modest improvements in prediction accuracy using the Censored Likelihood, it is possible that further improvement may be obtained by adding complexity to the model so that non-linear relationships within the range of target values $[a, b]$ can be captured, while simultaneously enforcing censoring of the likelihood. It would also be interesting to investigate performance using different kernels rather than the single linear kernel utilized in this work. Further experiments, using datasets with different properties and dimensionality and comparison with methods such as Gaussian Process Ordinal Regression [3] will be performed to investigate the behaviour of these different approaches.

Acknowledgment

Anil Rao and Janaina Mourao Miranda were supported by the Wellcome Trust under grant number WT102845/Z/13/Z. Joao M. Monteiro was supported by a PhD scholarship awarded by Fundacao para a Ciencia e a Tecnologia (SFRH/BD/88345/2012).

References

[1] C. M. Stonnington, C. Chu, S. Klöppel, C. R. Jack, J. Ashburner, and R. S. J. Frackowiak, "Predicting clinical scores from magnetic resonance scans in Alzheimer's disease," *NeuroImage*, vol. 51, no. 4, pp. 1405–1413, 2010.

[2] A. Marquand, M. Howard, M. Brammer, C. Chu, S. Coen, and J. Mourão Miranda, "Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes." *NeuroImage*, vol. 49, no. 3, pp. 2178–89, Mar. 2010.

[3] O. M. Doyle, J. Ashburner, F. O. Zelaya, S. C. R. Williams, M. a. Mehta, and a. F. Marquand, "Multivariate decoding of brain images using ordinal regression." *NeuroImage*, vol. 81, pp. 347–57, Nov. 2013.

[4] J. Young, M. Modat, M. J. Cardoso, A. Mendelson, D. Cash, and S. Ourselin, "Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment," *NeuroImage: Clinical*, vol. 2, pp. 735–745, 2013.

[5] C. E. Rasmussen, *Gaussian processes for machine learning*. MIT Press, 2006.

[6] E. Ertin, "Gaussian process models for censored sensor readings." *2007 IEEE/SP 14th Workshop on Statistical Signal Processing, SSP 07*, pp. 665–669, 2007.

[7] P. Groot and P. Lucas, "Gaussian Process Regression with Censored Data Using Expectation Propagation," *Sixth European Workshop on Probabilistic Graphical Models*, pp. 115–122, 2012.