Gaussian Process Regression with Censored Data Using Expectation Propagation

Perry Groot, Peter Lucas

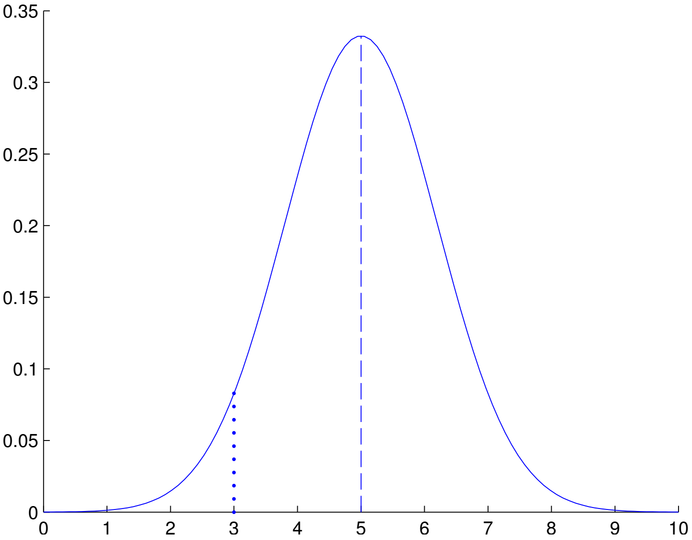
Radboud University Nijmegen

{perry,peterl}@cs.ru.nl

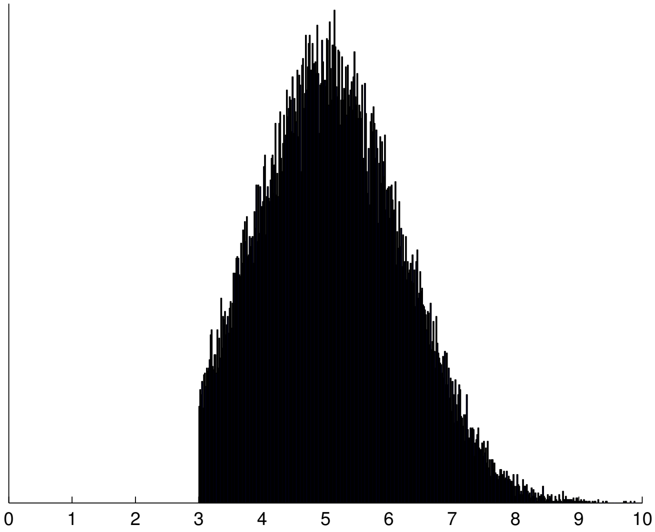
6^{th} European Workshop on Probabilistic Graphical Models

PGM 2012

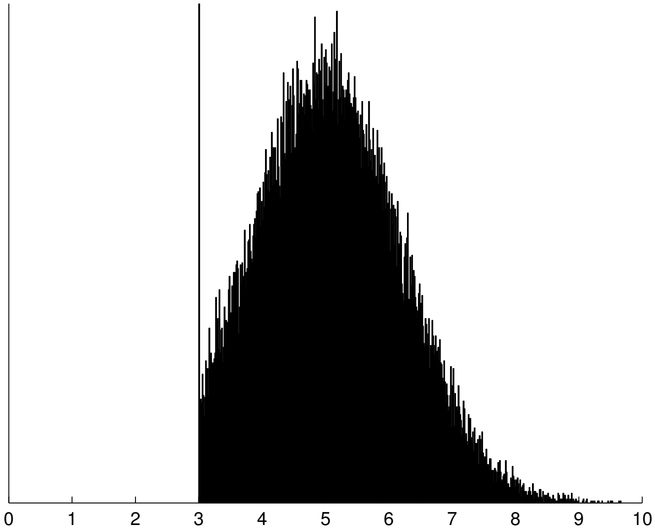
Introduction



Introduction - Truncation



Introduction - Censoring



Introduction

Both represent a limitation.

- Truncation: the population from which data is drawn
- Censoring: the variable of interest

Key difference is in the explanatory variable

- Truncation: missing
- Censoring: fully observable

Examples: Survival analysis, reliability testing

Problem Setting

Learn a function

$$f : \mathbb{R}^D \to \mathbb{R}$$

given a set of observations

$$\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$$

where y is a censored version of y^*:

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

We are interested in the posterior

$$p(f|\mathcal{D}) = \frac{p(f)p(\mathcal{D}|f)}{p(\mathcal{D})}$$

Gaussian Processes

A Gaussian process (GP) is collection of random variables $\{f_i\}$ with the property that the joint distribution of any finite subset has a joint Gaussian distribution.

A GP specifies a probability distribution over functions $f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$ and is fully specified by its mean function $m(\boldsymbol{x})$ and covariance (or kernel) function $k(\boldsymbol{x}, \boldsymbol{x}')$.

Typically $m(\boldsymbol{x}) = \mathbf{0}$, which gives

$$\{f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_l)\} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{K}) \text{ with } K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

Gaussian Processes - Posterior process

A priori, given data $\mathcal{D} = \{\boldsymbol{X}, \boldsymbol{y}\}$ with $\boldsymbol{y} = f(\boldsymbol{X})$ and test points $\boldsymbol{X}_*$ we have

$$\begin{bmatrix} f(\boldsymbol{X}) \\ f(\boldsymbol{X}_*) \end{bmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} K(\boldsymbol{X}, \boldsymbol{X}) & K(\boldsymbol{X}, \boldsymbol{X}_*) \\ k(\boldsymbol{X}_*, \boldsymbol{X}) & K(\boldsymbol{X}_*, \boldsymbol{X}_*) \end{bmatrix}\right)$$

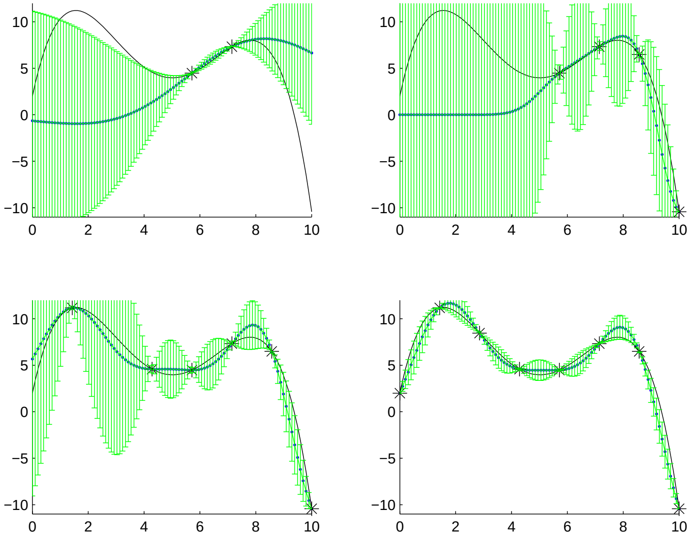
and after conditioning

$$f(\boldsymbol{X}_*)|\boldsymbol{X}_*, \boldsymbol{X}, \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$$

with

$$\begin{aligned} \boldsymbol{\mu} &= K(\boldsymbol{X}_*, \boldsymbol{X}) K(\boldsymbol{X}, \boldsymbol{X})^{-1} \boldsymbol{y} \\ \Sigma &= K(\boldsymbol{X}_*, \boldsymbol{X}_*) - K(\boldsymbol{X}_*, \boldsymbol{X}) \underbrace{K(\boldsymbol{X}, \boldsymbol{X})^{-1}}_{\mathcal{O}(n^3)} K(\boldsymbol{X}, \boldsymbol{X}_*) \end{aligned}$$

Gaussian Processes - 1D demo



Likelihood

Assume that latent function values are contaminated with Gaussian noise with zero mean and unknown variance.

Likelihood becomes a mixture of Gaussian and probit likelihood terms:

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

which is well-known as the Tobit likelihood.

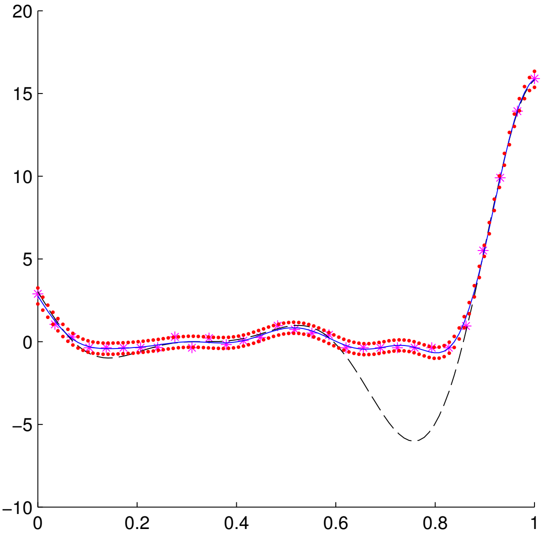
Expectation Propagation

- The posterior $p(f|\mathcal{D}) = \frac{p(f)p(\mathcal{D}|f)}{p(\mathcal{D})}$ is intractable
- EP approximates the likelihood by a Gaussian distribution making the posterior tractable
- Local likelihood approximations

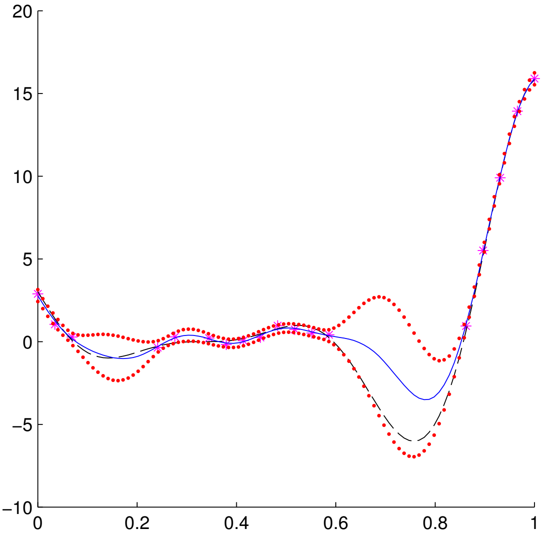
$$p(y_i|f_i) \simeq t_i(f_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2) = \tilde{Z}_i \mathcal{N}(f_i|\tilde{\mu}_i, \tilde{\sigma}_i^2)$$

- Approximation is iteratively updated
- In the Gaussian case the update step turns out to be the same as moment matching
- The zeroth, first, and second moments of the Tobit likelihood can be computed analytically

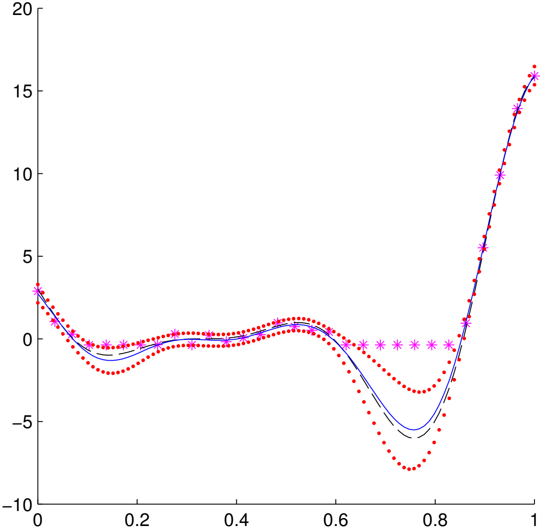
Experiments



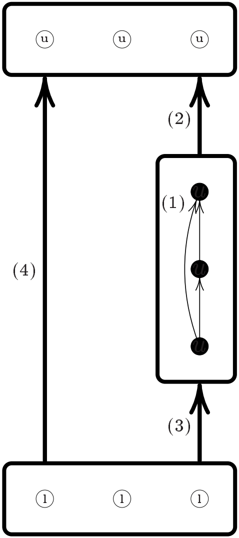
Experiments



Experiments



Experiments



Concordance index:

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

where $\mathcal{G} = (X, \mathcal{E})$ order graph with edges $\mathcal{E}$ according to (1)–(4)

Fraction of all pairs of inputs whose predicted values are correctly ordered among all inputs that can be ordered

Experiments

Housing data:

- 506 observations on 14 real-valued variables
- median value greater than $50.000 appear as $50.000
- 16 observations were censored (3.2% of the data)

Table: 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

Conclusions

- GPs provide a flexible, non-parametric Bayesian framework that can be extended to censored observations
- The intractable posterior in case of a Tobit likelihood can be approximated with EP using analytic update steps leading to a stable algorithm