Latent Gaussian Process for data-driven disease stratification with composite likelihoods

Siddharth Ramchandran[1]

Miika Koskinen[2*]

Harri Lähdesmäki[1*]

1 - Dept. Of Computer Science, Aalto University
2 - HUS / Helsinki Biobank

siddharth.ramchandran@aalto.fi

What are we trying to do?

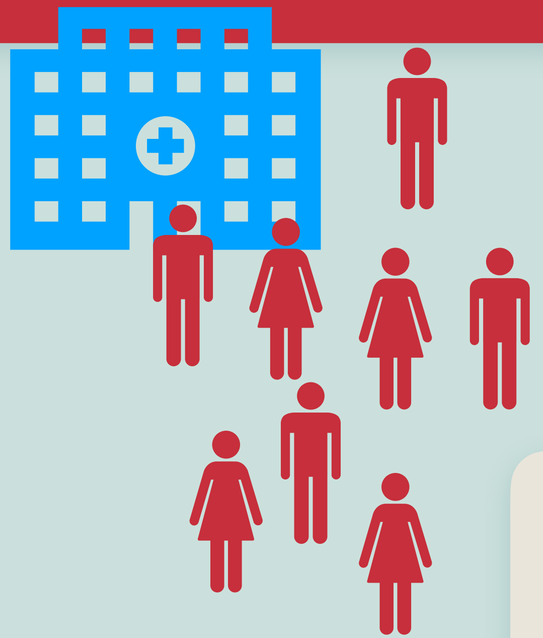
- Trying to understand patient records from several disparate sources (observation spaces/different likelihoods).
- Our method seeks to embed these observations in a low dimensional space while capturing the similarities between the observations.

CHALLENGE

How to handle different likelihoods and high dimensionality?

siddharth.ramchandran@aalto.fi

0, 1, 1, 0, 0, -0.5, 1.9, 0.23, 0.1, 0.2
1, 1, 1, 0, 0, 1.2, 2.3, 1.2, 0.8, 0.87
0, 0, 0, 0, 0, 1.3, -0.1, Null, 0.2, 0.3
....
1, 0, 0, 1, 1, 1.5, 2.4, 1.5, 0.9, 0.22



OUR METHOD

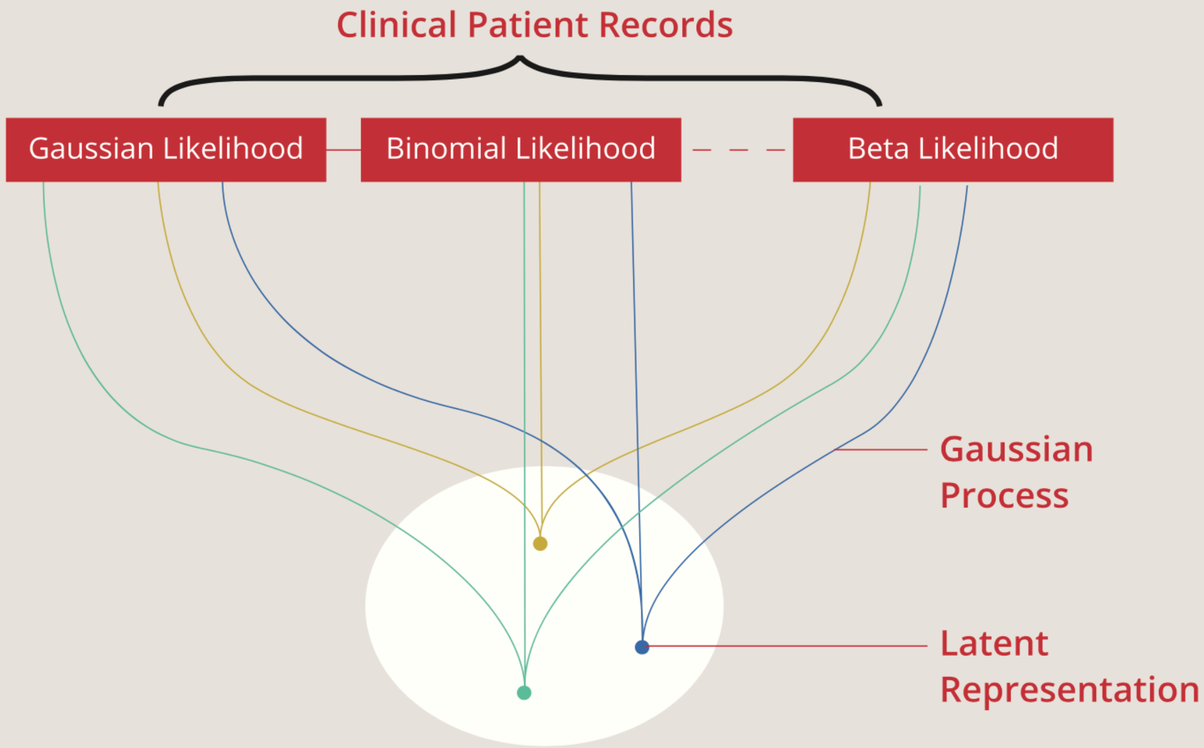
Patient records comprising of Binomial, Gaussian and Beta distributed data and missing values

Latent space with novel clustering among patients (can be of higher dimension)

siddharth.ramchandran@aalto.fi

How?

- Modify the Gaussian Process Latent Variable Model (GP-LVM) to learn a shared latent representation from the different observation spaces.
- Model the different observation spaces as generative models from a shared low dimensional latent representation.



siddharth.ramchandran@aalto.fi

What is a GP-LVM?

- A probabilistic and non-linear embedding of data in a lower dimensional space, where the latent variables are integrated out and the other hyper parameters are optimised.

IN A NUT SHELL

$$y = g(x) + \epsilon$$

Data

Gaussian Process

Latent variables

Noise

siddharth.ramchandran@aalto.fi

How to handle different likelihoods?

- Our model can handle data that comprises of different likelihoods like Binomial, Gaussian, Beta and Poisson (for now).

We learn the distribution parameters as follows:

$$\mathcal{F} \sim GP(K)$$

$$f = \mathcal{F}(x)$$

$$y \sim distribution(f)$$

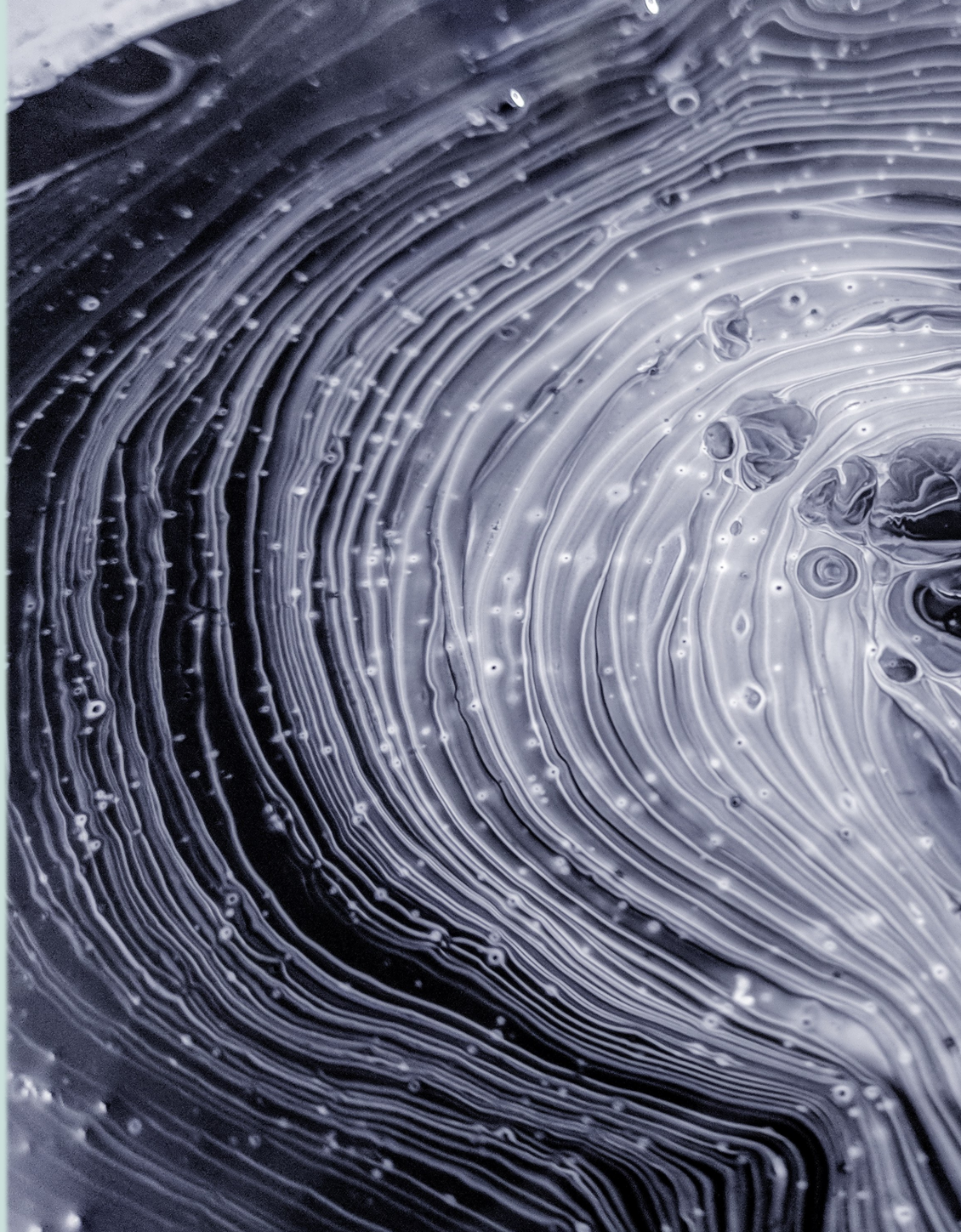
CHALLENGE

Intractable likelihood

siddharth.ramchandran@aalto.fi

Inference and Optimisation

- We make use of sampling-based variational inference to overcome the intractability.
- Obtain a lower bound on the log-evidence (ELBO).
- Compute gradients of the lower bound with Monte Carlo estimates.
- Use RMSProp optimiser to find an optimal variational distribution.



Demonstration on Clinical data of Parkinson’s disease

Data from

2200

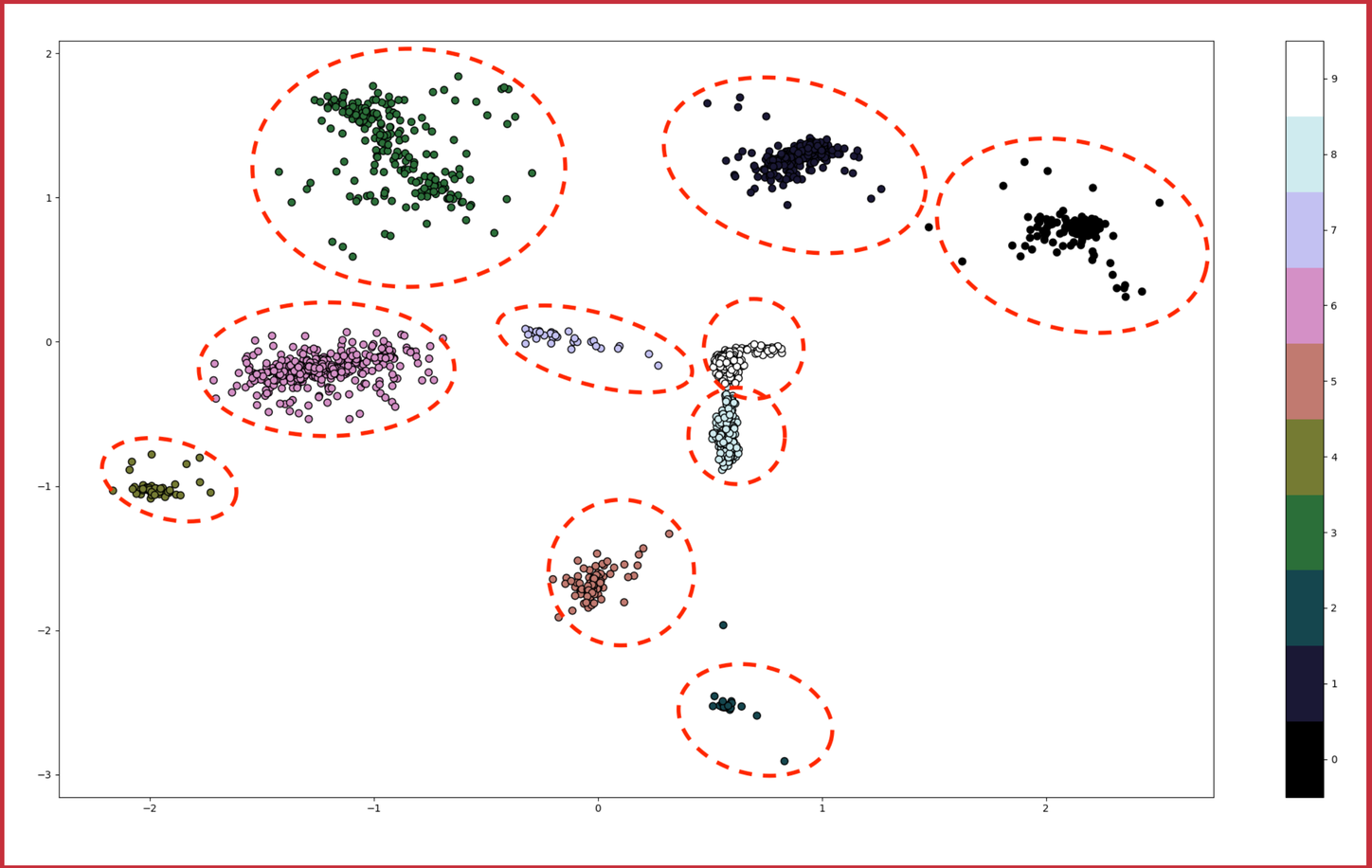
patients with

117 **covariates**

- **90 gaussian variables**
  laboratory data
- **26 binary variables**
  disease identification codes
- **1 binary variables**
  indicating gender

siddharth.ramchandran@aalto.fi

Results on Parkinson’s disease data



siddharth.ramchandran@aalto.fi

Thank You.

See you at the poster session

Aalto University
School of Science
and Technology

HELSINGIN BIOPANKKI
HELSINGFORS BIOBANK
HELSINKI BIOBANK

HUS

siddharth.ramchandran@aalto.fi