Comparing Pseudo-input Selection Methods in Sparse Gaussian Process Regression with Applications to Heart Rate Data

Hengrui Luo and Giovanni Nattino

The Ohio State University

STAT 8810 Special Topics in Uncertainty Quantification via Tree-based Models and Approximate Computations

Modeling Heart Rate Data

- Detecting early signs of HR deterioration in critical patients can improve patients' outcomes and decrease hospital costs.
- Several methods have been used with this goal, including GPs.

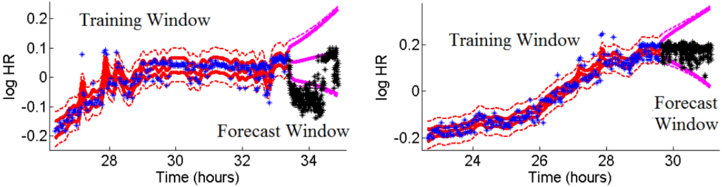


Figure: From Colopy et al. (2016).

Pros and Cons of GP Regression in HR modeling

Pros:

- Flexible models
- It is possible to account for generative physiology

Cons: high computational costs of GP regression on large datasets (as HR data!)

Possible solutions?

An Application to Real HR Data

We consider the data from the study by Zakynthinaki (2015):

- HR recordings in beats/minute of a single healthy subject;
- Eight datasets:
    - 4 recordings under physical exercise at different intensities (running at $v = 13.4$, 14.4, 15.7 and 17.0km/h);
    - 4 recordings in recovery phase after each exercise;
- About 400-500 records per dataset.

Gaussian Process Regression

Gaussian processes (GP) are widely used in a variety of fields [1]

- Engineering: Forecasting and prediction.
- Economy: Evaluation of the price stock and derivatives.
- Medicine: Epidemiology.

GP regression is proved to work well in scenes where the data set is relatively small, but there is a substantial obstruction in applying GP regression for general large data sets like spatial-temporal data sets and social networks [2].

The major computation cost occurs in the inversion of a large covariance matrix.

Existing Methods I

Due to the limitation of exact GP calculation, approximation methods are proposed to address the computational issue. [3] regards these sparse approximations as different methods of modifying priors on the GP regression model and categorized them as types below.

- Subset of regressors approximation (SoR).
  Take a subset of regressors by minimizing a certain criterion, approximate the joint likelihood directly.
- Deterministic training conditional approximation (DTC).
  Suggest a set of projected latent variables, approximate the joint likelihood directly.

Existing Methods II

- Fully independent training conditional approximation (FITC). Suggest a set of pseudo-inputs (inducing variables), does not give a deterministic relation between the inducing variables $\boldsymbol{f}$ and original regressors $\boldsymbol{u}$ but choose them by minimizing their training conditionals.
  - **The authors of [2] considered this problem on the basis of the sparse variational approximation proposed by [6, 9] with efficient algorithm.**
- Partially independent training conditional approximation (PITC). Same as FITC except that we now use another independent distribution to approximate training conditionals.

Existing Methods III

- Transduction and augumentation methods.
  They restrict the prediction set to a prespecified set of test cases, rather than trying to learn an entire function.
- Nystrom Approximation.
  This approximates the covariance function directly.

These approximation methods have their strengths and weak points, in our scenario of application. The problem we considered is to model and monitor the heart rate data. The heart rate data is collected in a continuous manner and usually of a large sample size, therefore such a task will face two problems that we hope to address using sparse GP methodologies.

1. Computational cost of fitting GP.
2. Predictive efficiency and accuracy from fitted GP.

Equi-spaced Grids and Simple Random Sampling (SRS)

The equi-spaced grids is simply divide the bounded location space into m equally spaced locations and used them as pseudo-inputs. This is more like DTC method but it serves as a standard method that we can used for comparative purpose.

The idea of simple random sampling (SRS) will choose the location of pseudo-inputs using a random sample of size m from a uniform distribution on the bounded location space. SRS is a naive FITC method because we are basically using uniform distribution as the training conditional (though it contains no information from the observations at all.). This will also be a referential method that we compare with more complicated FITC and PITC methods as shown below.

Preferential Sampling (PS)

The idea of preferential sampling (PS) method of choosing pseudo-inputs $\mathrm{f}_m = \mathrm{f}_m(X_m)$ is to select, among all those n observed locations $\mathrm{f} = \mathrm{f}(X)$ randomly with probabilities specified below. For each pair of observations $(y_i, \mathrm{f}(x_i))_{i=1}^n$ we calculate the quantity $\frac{\Delta y}{\Delta x} := \left|\frac{y_i - y_{i-1}}{x_i - x_{i-1}}\right|, i = 2, \cdots n$ and $\frac{y_1}{x_1}$ for the boundary observation. Then we randomly select m pseudo-inputs from all n locations with probabilities that are proportional to $\left\{\frac{\Delta y}{\Delta x}\right\}$ accordingly. Intuitively, this method will select those observed locations near which the output changes greatly with higher probability. By proceeding in this way we hope that we can capture more information from those greatly varying locations.

Maximizing the Marginal Likelihood

The idea of choosing pseudo-inputs $\mathrm{f}_m = \mathrm{f}_m(X_m)$ is to maximize the marginal likelihood of

$$\mathrm{f}_m = argmax_{\mathrm{f}_m} p(\boldsymbol{y} \mid \mathrm{f}_m)\, p(\mathrm{f}_m)$$

Or equivalently when the prior on pseudo-inputs is Gaussian with mean 0 and variance σ^2, we maximize the following

$$\mathrm{f}_m = argmax_{\mathrm{f}_m} \log\left[N\left(\boldsymbol{y} \mid 0, \sigma^2 \boldsymbol{I} + K_{nn}\right)\right]$$

where K_{nm}, K_{mm}, K_{mn} denoting the submatrices from the covariance matrix $K = \begin{bmatrix} K_{nn} & K_{nm} \\ K_{mn} & K_{mm} \end{bmatrix}$ of the joint distribution $p(\mathrm{f}, \mathrm{f}_m)$. This is a PITC method.

Minimizing the Lower Bound of KL divergence I

Theorem

[6, 9] To minimize the Kullback-Leibler divergence inbetween the full posterior $p(\mathrm{f},\mathrm{f}_m \mid \boldsymbol{y})$ with $dim\mathrm{f} = dim\boldsymbol{y} = n$ and the approximating posterior $q(\mathrm{f},\mathrm{f}_m) = p(\mathrm{f} \mid \mathrm{f}_m)\phi(\mathrm{f}_m)$, it suffices to maximize the following lower bound

$$\begin{aligned} F_V(\mathrm{f}_m,\phi) &= \int p(\mathrm{f} \mid \mathrm{f}_m)\phi(\mathrm{f}_m)\log\frac{p(\boldsymbol{y} \mid \mathrm{f})p(\mathrm{f}_m)}{\phi(\mathrm{f}_m)}d\mathrm{f}d\mathrm{f}_m \\ &= \log\left[N\left(\boldsymbol{y} \mid 0,\sigma^2\boldsymbol{I} + Q_{nn}\right)\right] - \frac{1}{2\sigma^2}trace\left(\tilde{\boldsymbol{K}}\right) \end{aligned}$$

where $Q_{nn} = K_{nm}K_{mm}^{-1}K_{mn}$, $\tilde{\boldsymbol{K}} = Cov(\mathrm{f},\mathrm{f}_m) = K_{nn} - K_{nm}K_{mm}^{-1}K_{mn}$ and K_{nm},K_{mm},K_{mn} denoting the submatrices from the covariance matrix $K = \begin{bmatrix} K_{nn} & K_{nm} \\ K_{mn} & K_{mm} \end{bmatrix}$ of the joint distribution $p(\mathrm{f},\mathrm{f}_m)$; $\phi(\mathrm{f}_m) = p(\mathrm{f} \mid \mathrm{f}_m)$

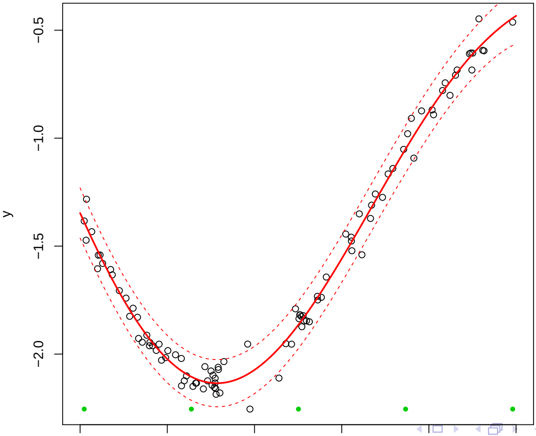
Minimizing the Lower Bound of KL divergence II

The idea of choosing pseudo-inputs is to minimize the KL-divergence between the exact posterior with all inputs $\mathrm{f} = \mathrm{f}(X)$ and the approximating posterior with pseudo-inputs $\mathrm{f}_m = \mathrm{f}_m(X_m)$. This major theoretic result greatly facilitates the variational optimization procedure and makes the idea of minimizing the KL divergence between two distributions full posterior $p(\mathrm{f}, \mathrm{f}_m \mid \boldsymbol{y})$ and the approximating posterior $q(\mathrm{f}, \mathrm{f}_m)$ feasible and relatively efficient. Due to its multimodal behavior, we use standard annealing optimization instead of SGD to make the result more stable and tractable.

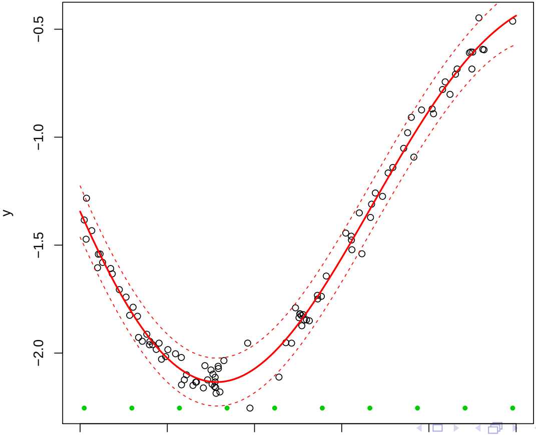
Simulation Analysis 1: Effect of m

- We assume $\boldsymbol{y} = \mathrm{f}(\boldsymbol{x}) + \boldsymbol{\varepsilon}$, with:
  - $\mathrm{f}(\boldsymbol{x}) \sim GP(0_n, \sigma_K^2 K)$, where $K(x_1, x_2) = \rho^{(x_1 - x_2)^2}$
  - $\boldsymbol{\varepsilon} \sim N\left(0_n, \sigma^2 \boldsymbol{I}_n\right)$
- We draw two GPs, one with $\rho = .1$ (“smooth”) and one with $\rho = 10^{-9}$ (“wiggly”). Other parameters:
  - $n = 100$
  - $\boldsymbol{x}$ sampled uniformly on $(0,1)$
  - $\sigma_K = 1$ and $\sigma = 0.05$
- We fit the five methods to both the realizations with $m = 5$, 7, 10, 12, 15, 20, 30, 40 and 50.
- We evaluate the reliability of the fitted GP with the average prediction error on the observed data.

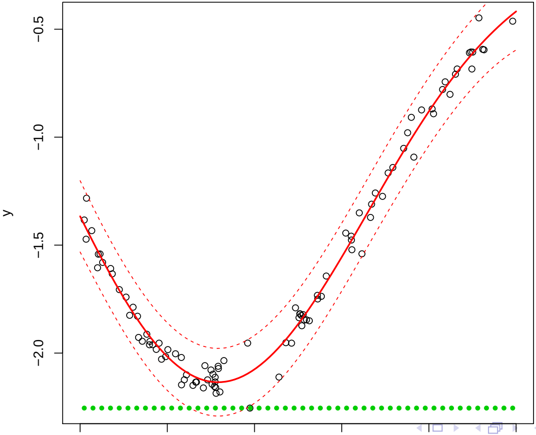
Grid - $m = 5$



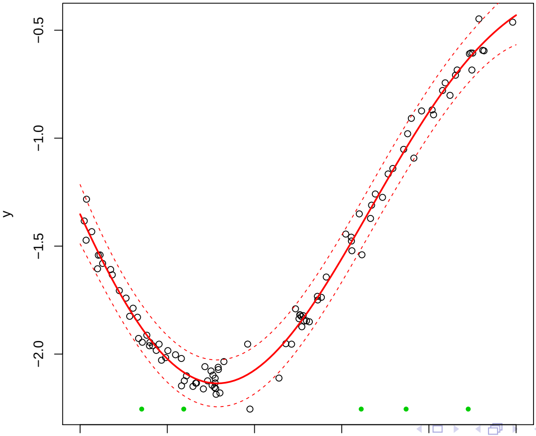
Grid - $m = 10$



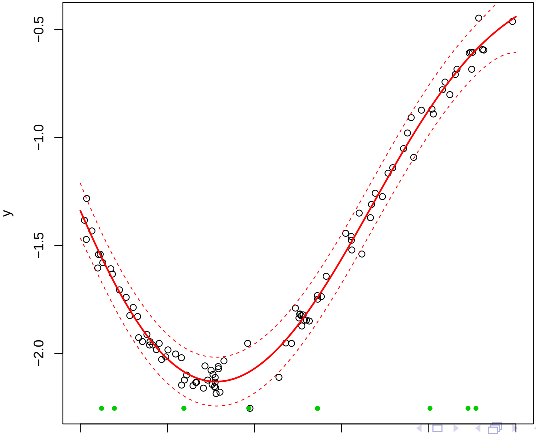
Grid - $m = 50$



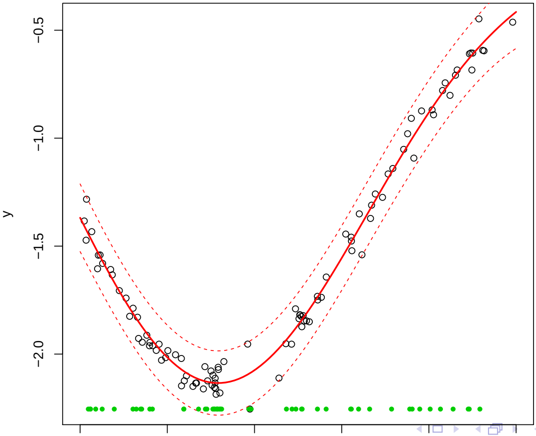
PS - $m = 5$



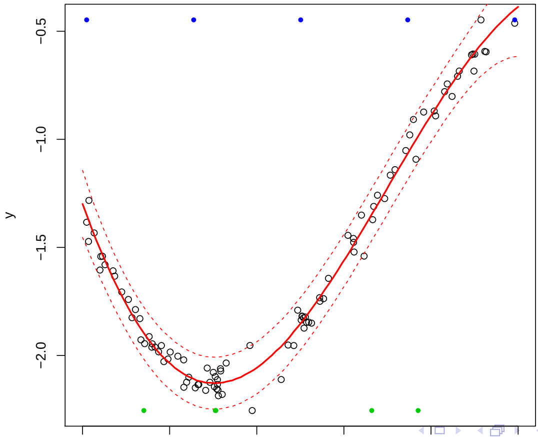
PS - $m = 10$



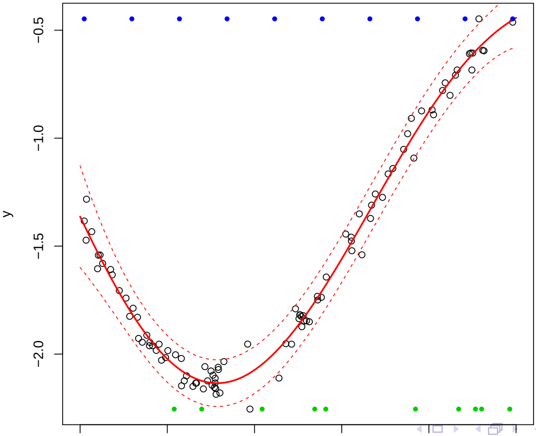
PS - $m = 50$



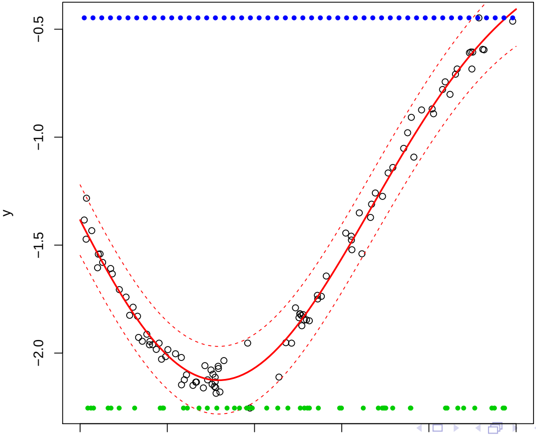
KL - $m = 5$



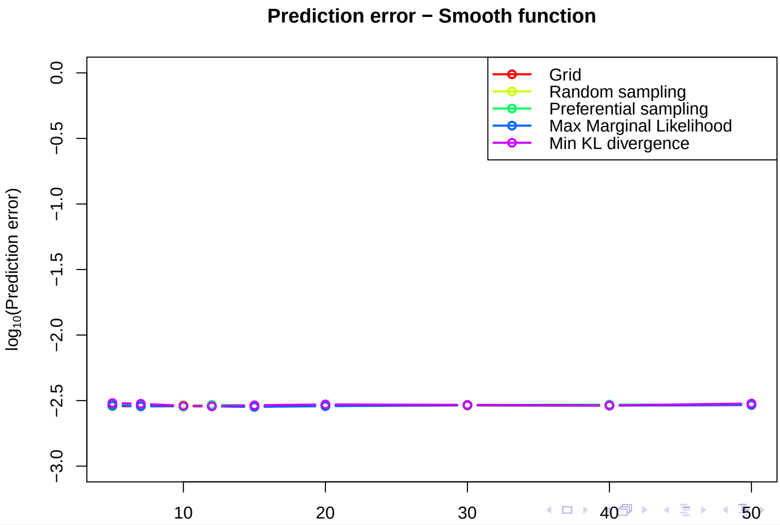
KL - $m = 10$



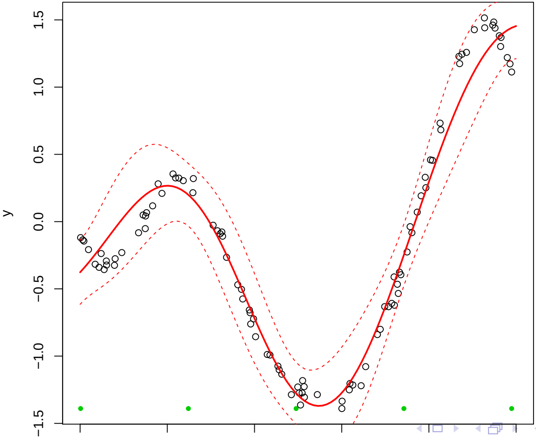
KL - $m = 50$



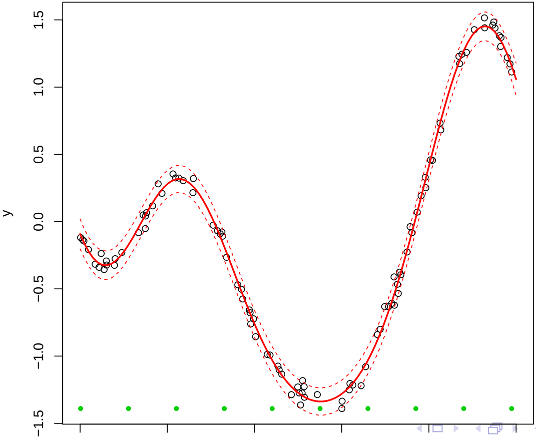
Simulation Analysis 1: Effect of m



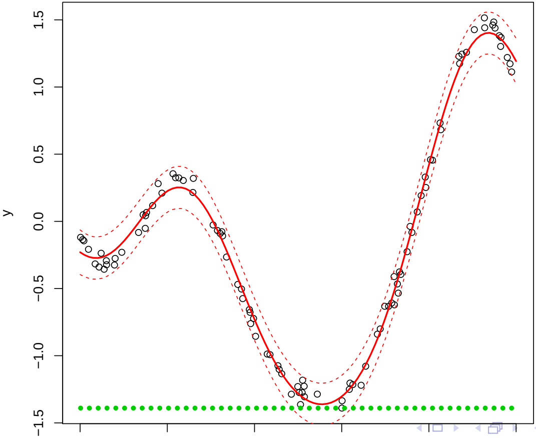
Grid - $m = 5$



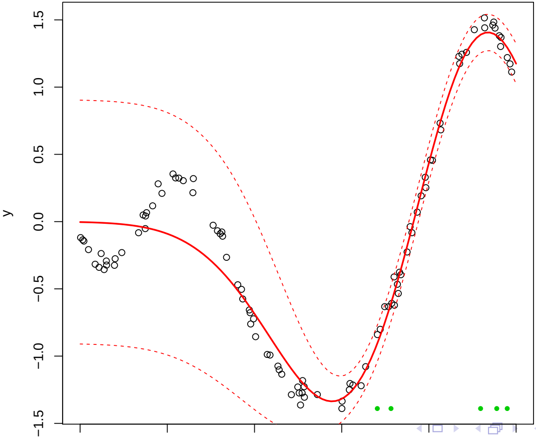
Grid - $m = 10$



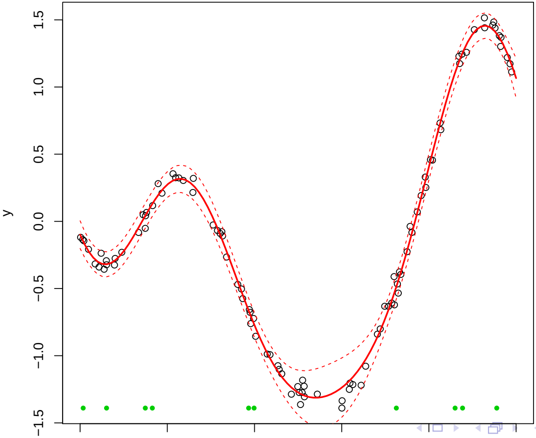
Grid - $m = 50$



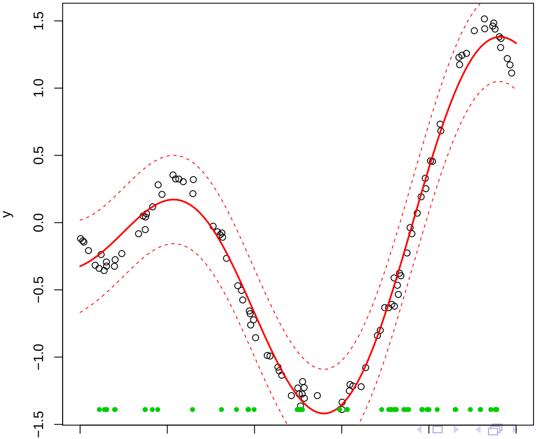
PS - $m = 5$



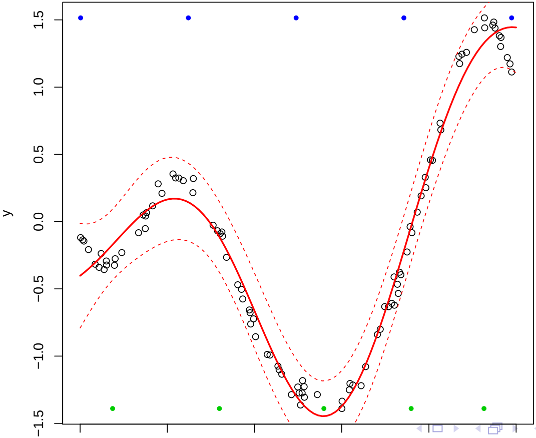
PS - $m = 10$



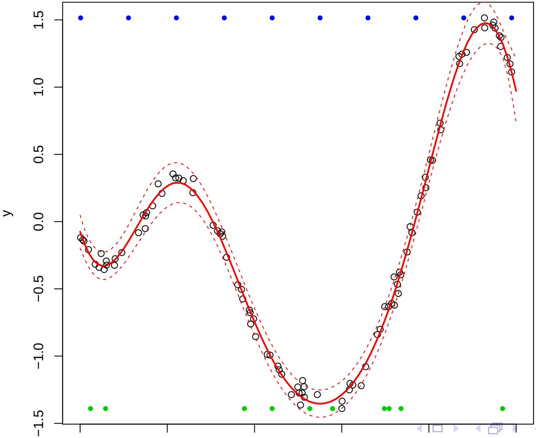
PS - $m = 50$



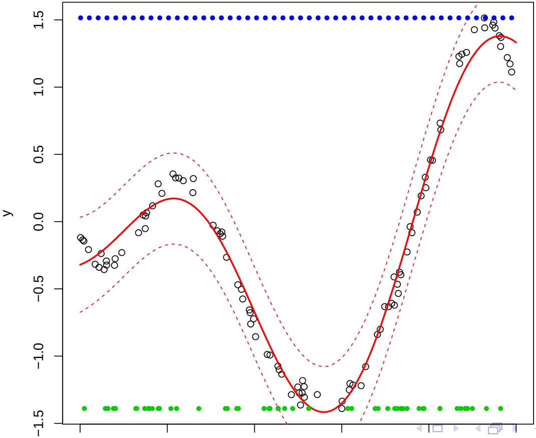
KL - $m = 5$



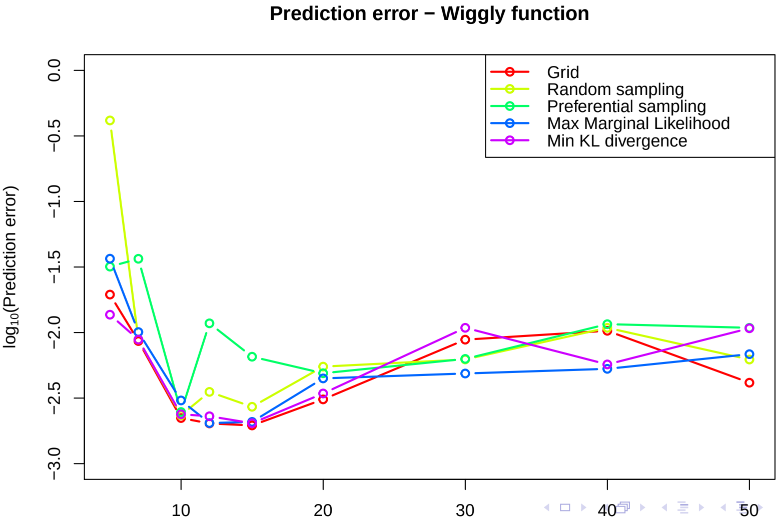
KL - $m = 10$



KL - $m = 50$



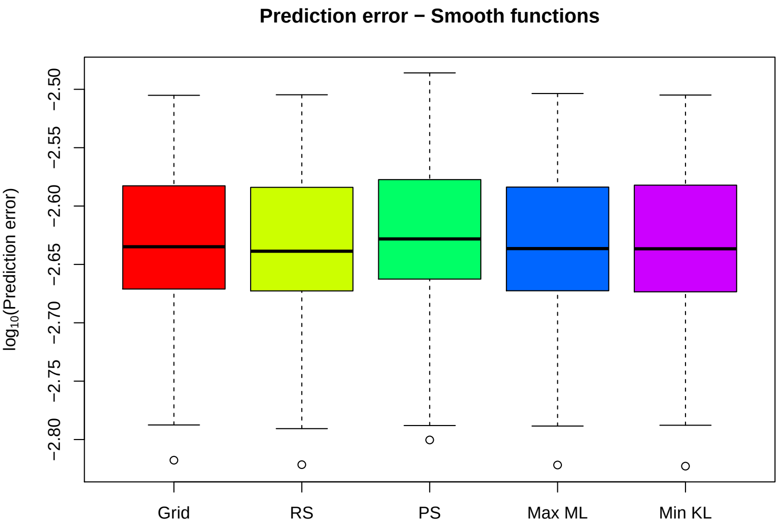
Simulation Analysis 1: Effect of *m*



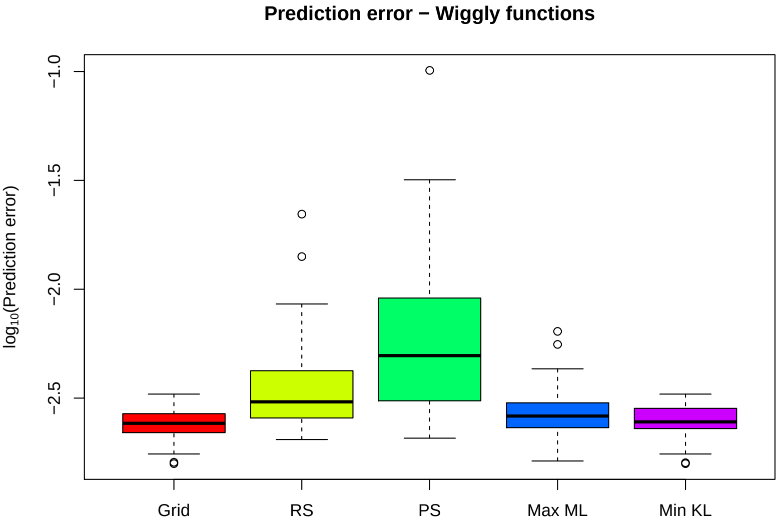
Simulation Analysis 2: Average Performance

- We assume $\boldsymbol{y} = \mathrm{f}(\boldsymbol{x}) + \boldsymbol{\varepsilon}$, with:
  - $\mathrm{f}(\boldsymbol{x}) \sim GP(0_n, \sigma_K^2 K)$, where $K(x_1, x_2) = \rho^{(x_1 - x_2)^2}$
  - $\boldsymbol{\varepsilon} \sim N\left(0_n, \sigma^2 \boldsymbol{I}_n\right)$
- We draw $N = 50$ realizations from the "smooth" ($\rho = .1$) and "wiggly" ($\rho = 10^{-9}$) family of GPs. Other parameters:
  - $n = 100$
  - $\boldsymbol{x}$ sampled uniformly on $(0, 1)$
  - $\sigma_K = 1$ and $\sigma = 0.05$
- We fit the five methods to both the families of simulations with $m = 10$.
- We evaluate the reliability of the fitted GP with the average prediction error on the observed data.

Simulation Analysis 2: Average Performance

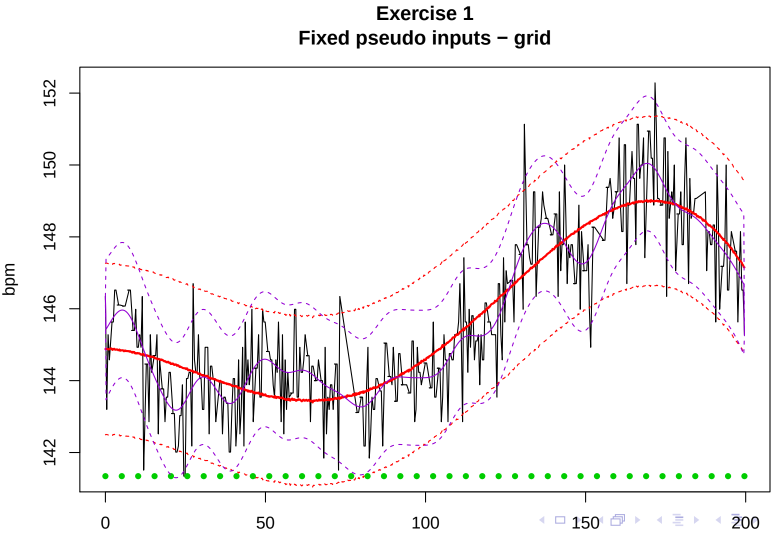


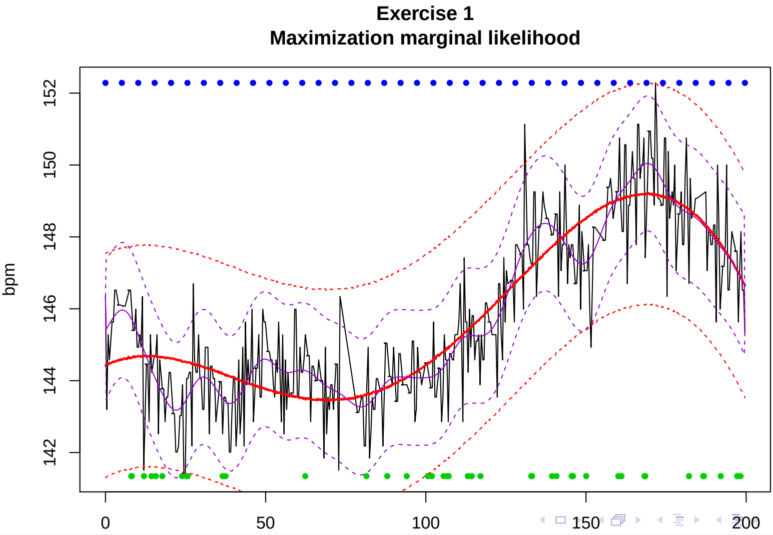
Simulation Analysis 2: Average Performance

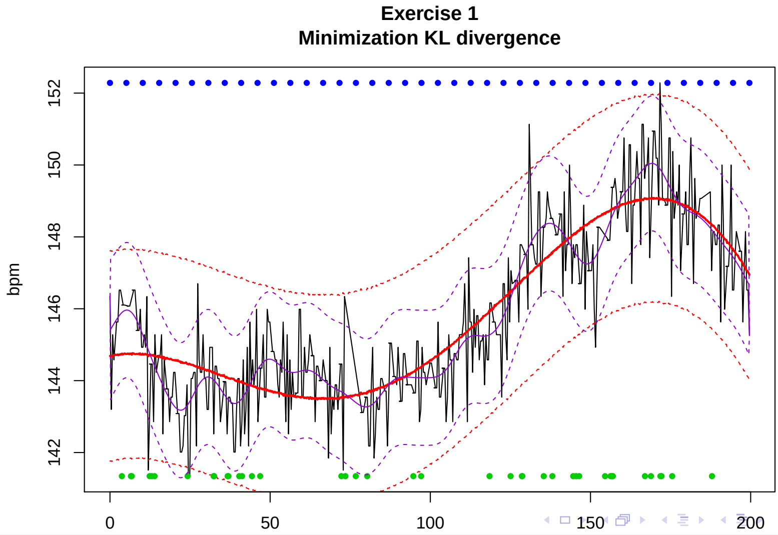


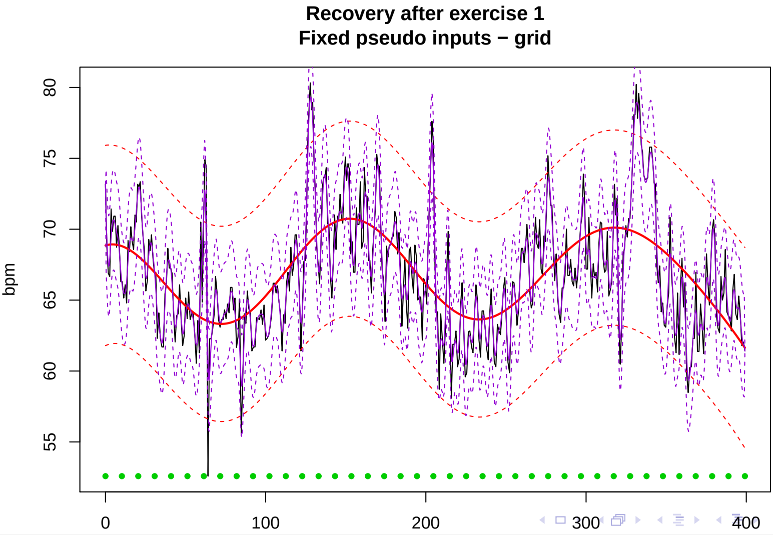
Application to HR Data

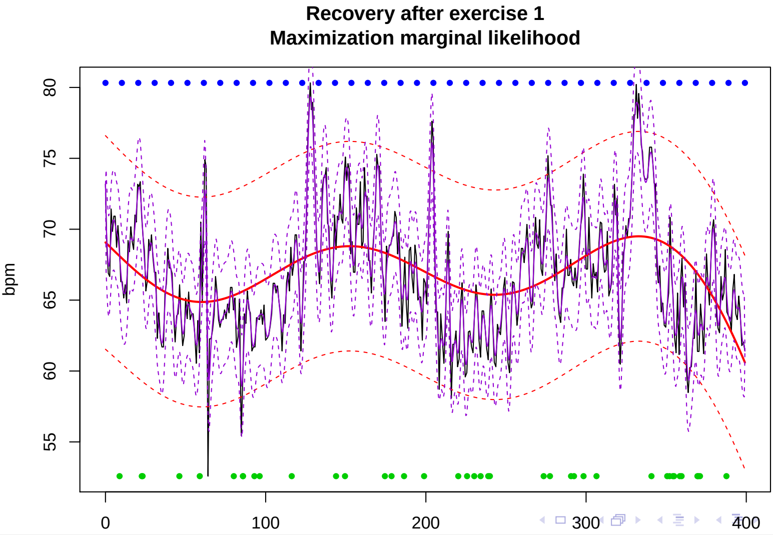
- Grid, Max ML and Min KL performed better than the other two approaches. We apply this three methods to fit GP to HR data.
- We selected the number of pseudo-inputs to be$\sim 10\%$ of the number of observations, i.e., $m = 40$.


Exercise 1
Fixed pseudo inputs − grid
bpm
152
150
148
146
144
142
0
50
100
150
200

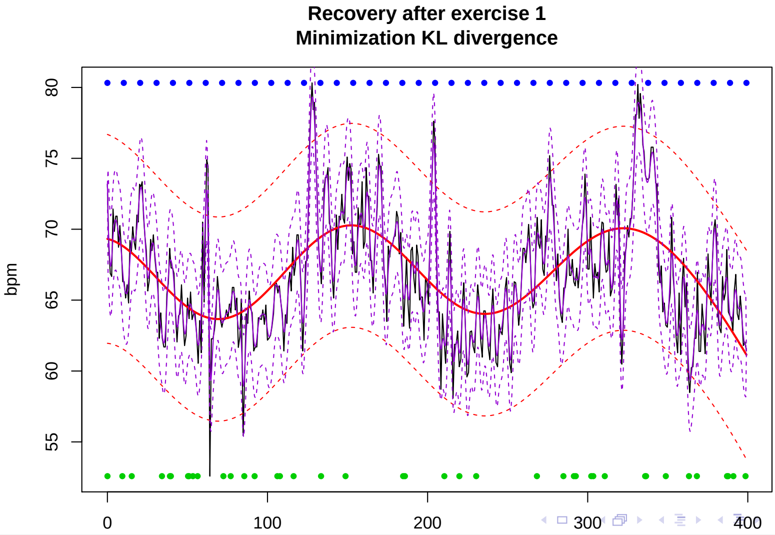

Exercise 1
Maximization marginal likelihood
bpm
152
150
148
146
144
142
0
50
100
150
200


Exercise 1
Minimization KL divergence
bpm
142
144
146
148
150
152
0
50
100
150
200


Recovery after exercise 1
Fixed pseudo inputs − grid
bpm
55
60
65
70
75
80
0
100
200
300
400


Recovery after exercise 1
Maximization marginal likelihood
bpm
80
75
70
65
60
55
0
100
200
300
400

Application to HR Data



Summary and Discussion

- Effect of number of pseudo-inputs *m*.
  - Elbow effect.
  - Correlation structure of GP data.
  - Optimization methods. Will the optimization methods be a factor that caused such difference?
- Empirical performance of the sparse GP.
  - Empirical prediction error.
  - Real data analysis. HR data.
  - Outlier detection on HR data.
  - Dimensionality and nonstationarity problem.

- Rasmussen, Carl Edward, and Christopher KI Williams. Gaussian processes for machine learning. Vol. 1. Cambridge: MIT press, 2006.
- Hensman, James, Nicolo Fusi, and Neil D. Lawrence. "Gaussian processes for big data." arXiv preprint arXiv:1309.6835 (2013).
- Quiñonero-Candela, Joaquin, and Carl Edward Rasmussen. "A unifying view of sparse approximate Gaussian process regression." Journal of Machine Learning Research 6.Dec (2005): 1939-1959.
- Damianou, Andreas, and Neil Lawrence. "Deep gaussian processes." Artificial Intelligence and Statistics. 2013.

- Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. "Variational inference: A review for statisticians." Journal of the American Statistical Association (2017).

- Titsias, Michalis K. "Variational learning of inducing variables in sparse Gaussian processes." International Conference on Artificial Intelligence and Statistics. 2009.

- Snelson, Edward, and Zoubin Ghahramani. "Sparse Gaussian processes using pseudo-inputs." Advances in neural information processing systems. 2006.

- Wainwright, Martin J., Tommi S. Jaakkola, and Alan S. Willsky. "A new class of upper bounds on the log partition function." IEEE Transactions on Information Theory 51.7 (2005): 2313-2335.

Titsias, Michalis K. Variational model selection for sparse Gaussian process regression. Technical report, School of Computer Science, University of Manchester, 2009.