http://jmlr.org/papers/v23/19-597.html

$\mathcal{X}$ $\mathcal{O}$

x

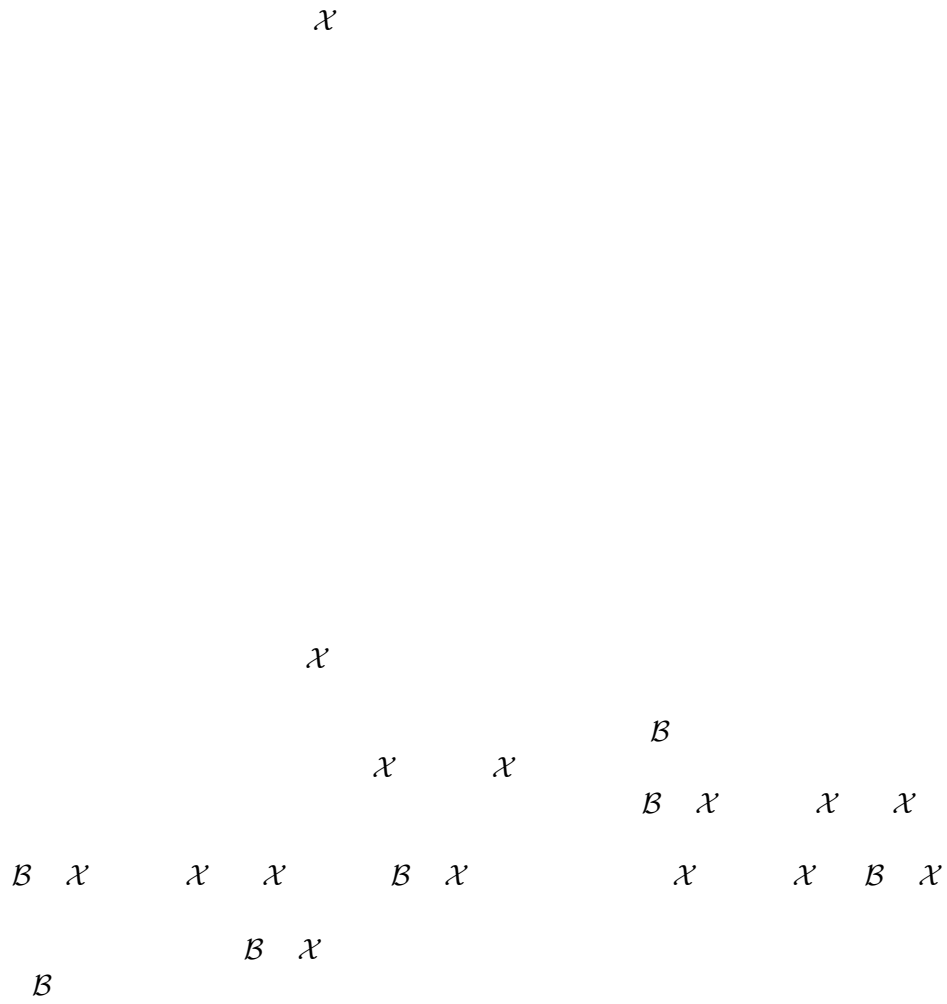
x

x x

x x

x x

x



$\mathcal{L}$ $\mathcal{L}$

$\mathcal{L}$

$\mathcal{L}$ $\mathcal{L}$

$\mathcal{L}$

$\mathcal{X}$

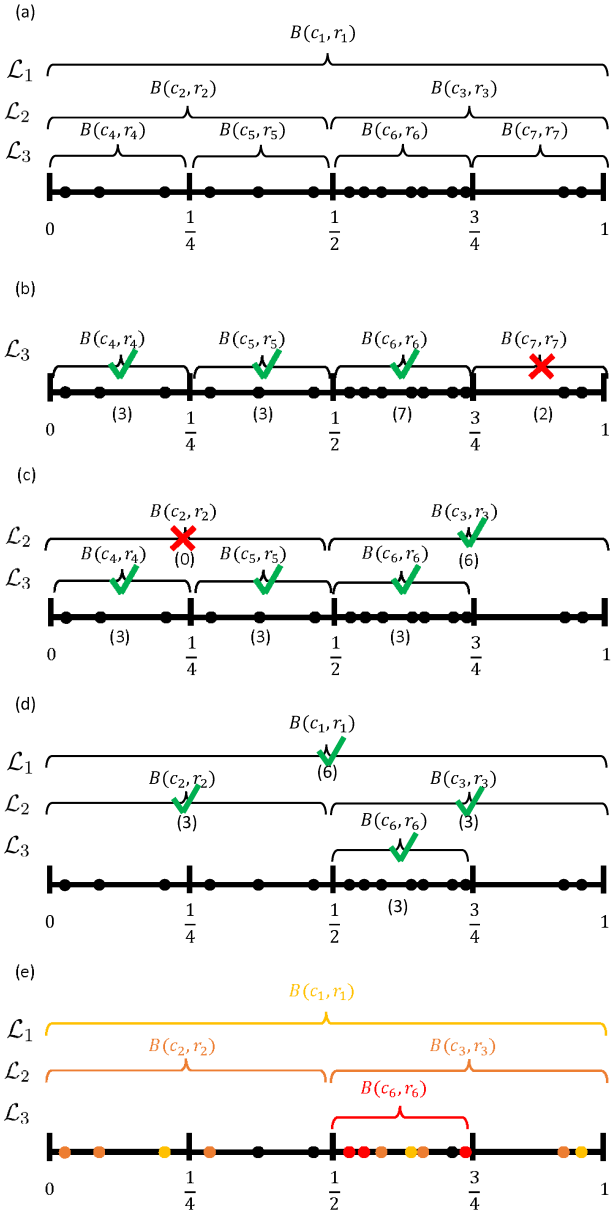
$\mathcal{L}$ $\mathcal{L}$ $\mathcal{L}$

$\mathcal{B}$

$\mathcal{X}$

$\mathcal{X}$

$\mathcal{X}$



The initial complete RP scheme as a binary tree with 3-layers on $[0, 1]$.

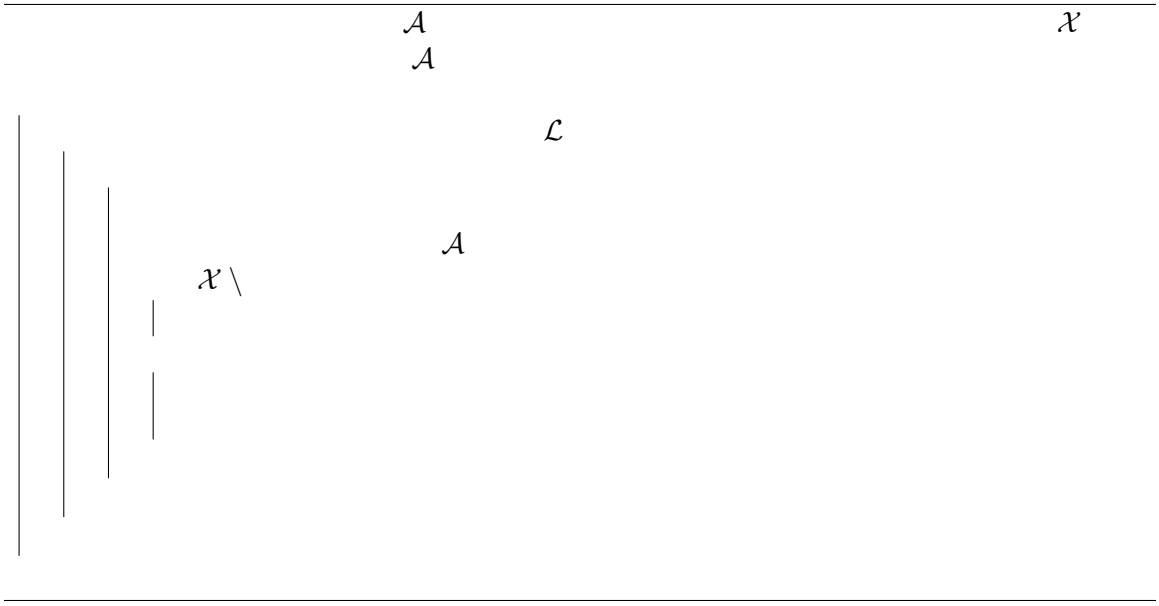
Starting from layer 3, we prune $B(\boldsymbol{c}_7, \boldsymbol{w}_7)$ since there are only $2 < 3$ observations available. We keep $B(\boldsymbol{c}_6, \boldsymbol{w}_6), B(\boldsymbol{c}_5, \boldsymbol{w}_5), B(\boldsymbol{c}_4, \boldsymbol{w}_4)$ as they all contain at least $m = 3$ observations.

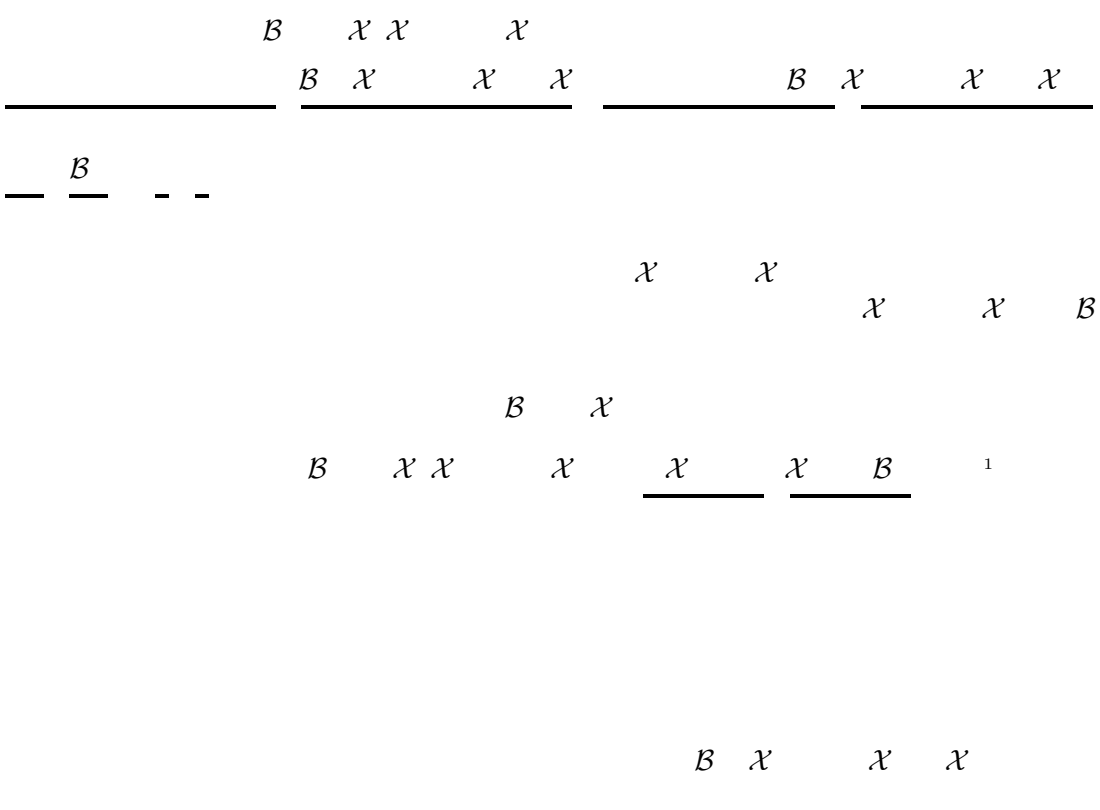
Moving to layer 2, $B(\boldsymbol{c}_3, \boldsymbol{w}_3)$ has the 6 observations required by itself and its child $B(\boldsymbol{c}_6, \boldsymbol{w}_6)$. Checking $B(\boldsymbol{c}_2, \boldsymbol{w}_2)$, it contains only 6 observations so we prune its children $B(\boldsymbol{c}_4, \boldsymbol{w}_4), B(\boldsymbol{c}_5, \boldsymbol{w}_5)$.

Moving to layer 1, $B(\boldsymbol{c}_1, \boldsymbol{w}_1)$ contains 15 observations, therefore there are sufficient observations for $B(\boldsymbol{c}_1, \boldsymbol{w}_1)$ and its children $B(\boldsymbol{c}_2, \boldsymbol{w}_2)$, $B(\boldsymbol{c}_3, \boldsymbol{w}_3)$ and $B(\boldsymbol{c}_6, \boldsymbol{w}_6)$. We keep $B(\boldsymbol{c}_1, \boldsymbol{w}_1)$ and its children, completing the partitioning.

Given the final RP scheme $B(\boldsymbol{c}_1, \boldsymbol{w}_1)$, $B(\boldsymbol{c}_2, \boldsymbol{w}_2)$, $B(\boldsymbol{c}_3, \boldsymbol{w}_3)$ and $B(\boldsymbol{c}_6, \boldsymbol{w}_6)$, one possible random selection of the pseudo-inputs $\bar{\mathcal{X}}$ for the $j = 1, \ldots, N$ different additive components (here $N = 4$) conditional on the RP scheme is shown as colored dots. Points with the same color belong to blocks on the same layer.

$\mathcal{X}$





—

$\mathcal{X}$ $\mathcal{X}$

$\mathcal{B}$ $\mathcal{X}$ $\mathcal{X}$

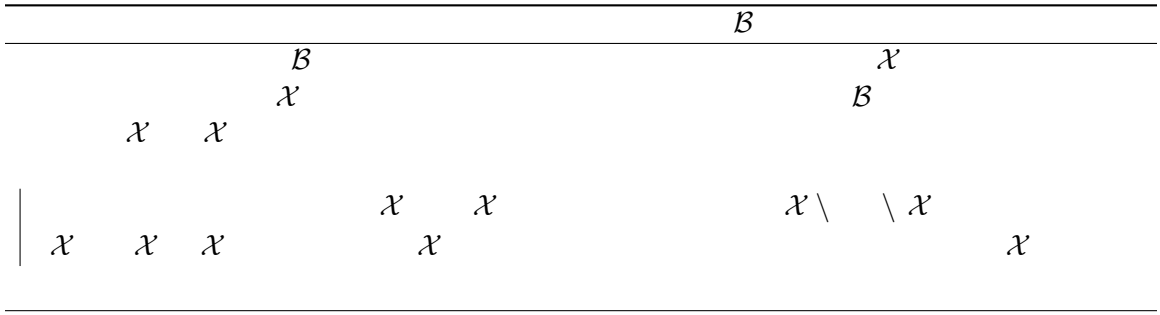
$\mathcal{B}$ $\mathcal{X}$ $\mathcal{X}$ $\mathcal{X}$

$\mathcal{B}$ $\mathcal{X}$ $\mathcal{X}$ $\mathcal{X}$ $\mathcal{B}$ $\mathcal{X}$ $\mathcal{X}$

$\mathcal{X}$ $\mathcal{X}$ $\mathcal{B}$ $\mathcal{X}$

$\mathcal{B}$

$\mathcal{I}$



$\mathcal{I}$

$\mathcal{C}$ $_{\mathcal{I}}\mathcal{X}$

$\mathcal{X}$ $\mathcal{X}$ $\mathcal{B}$ $\mathcal{X}$ $\mathcal{C}$

$_{\mathcal{L}}$

$\mathcal{X}$ $\mathcal{C}$ $\mathcal{C}$

$\mathcal{C}$ $\mathcal{X}$ $\mathcal{C}$

$\mathcal{B}$

$\mathcal{L}$

$\mathcal{X}$ $\mathcal{B}$ $\mathcal{X}$ $\mathcal{X}$

$\mathcal{B}$ $\mathcal{X}$ $\mathcal{X}$

$\mathcal{X}$ $\mathcal{B}$ $\mathcal{X}$

$\mathcal{X}$ $\mathcal{B}$ $\mathcal{X}$ $\mathcal{X}$

– –

$\mathcal{X}$

$\mathcal{O}$

$\mathcal{X}$ $\mathcal{O}$

$\mathcal{O}$

■

$\mathcal{X}$

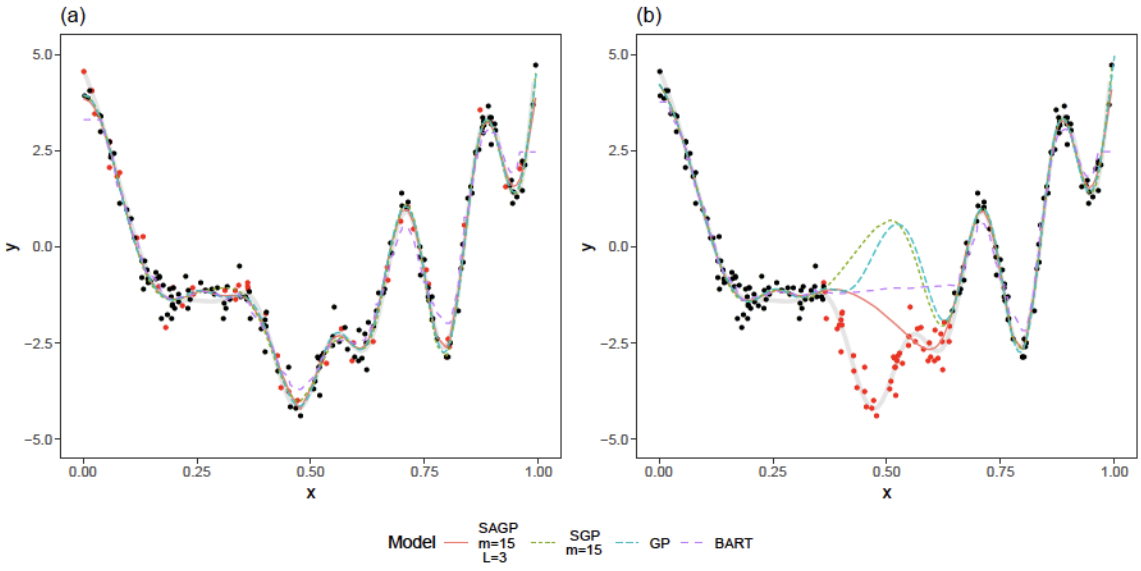


Figure 2: Example of data generated in the simulation study. The gray, bold, curve represents the true mean function $f(x)$. Training and testing sets are represented as black and red points. Panels (a) and (b) show the scenarios where the 50 data points of the testing set are chosen at random or as the input location that is closest to a randomly chosen point (0.5 in the example), respectively. The posterior predictive functions of four models, fit on the training portion of the data, are provided in both panels.

For each generated dataset, the models are fit on the training data and used to predict the response on the testing data. For each point in the testing set, we compute the estimated mean function $\hat{y}(x_i)$ (see Section 3.2.6) and the 95% prediction interval (PI) for y_i. The performance of the estimators of the mean function is evaluated in terms of root mean squared error (RMSE). To assess the appropriateness of the uncertainty quantification, we compute the coverage of the PIs and compare it to the nominal prediction level. Finally, we compare the methods in terms of the average value of interval scores, which is a summary measure to assess the quality of prediction intervals (Gneiting and Raftery, 2007). Given a $(1-\alpha)100\%$ PI for y_i with extremes (l_i, u_i), the interval score at y_i is defined as

$$s_\alpha(l_i, u_i; y_i) = (u_i - l_i) + \frac{2}{\alpha}(l_i - y_i)\mathbf{1}(y_i < l_i) + \frac{2}{\alpha}(y_i - u_i)\mathbf{1}(y_i > u_i).$$

We choose this metric to jointly evaluate a family of intervals in terms of precision (i.e. the width of the intervals) and accuracy (i.e., the coverage of the true value). Notably, low values of the score indicate good performance.

4.2 Results

Figure 3 summarizes the resulting RMSEs, PI coverages and averages of the interval scores across the 1000 generated datasets for the two scenarios.

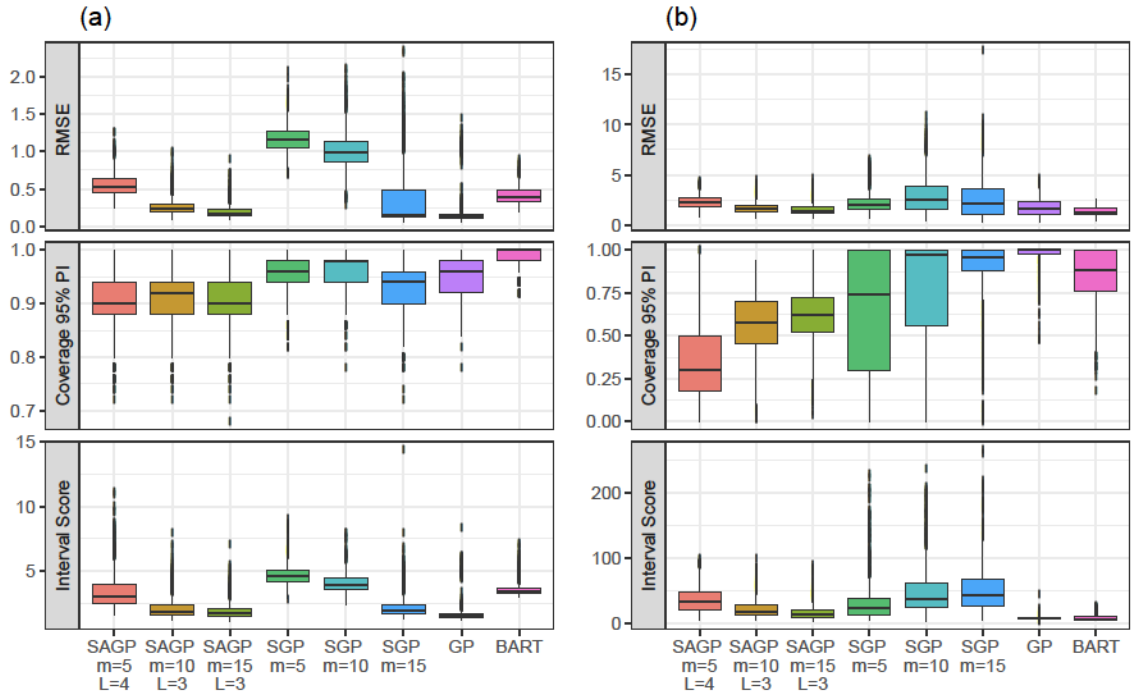


Figure 3: RMSE (top panels), coverage (central panels) and interval score (bottom panels) over the 1000 simulated datasets. Panel (a) shows the results in the case where the testing set is chosen at random over $[0, 1]$. Panel (b) shows the results in the case where the testing set is chosen in a random interval with center uniformly selected from $[0.25, 0.75]$.

Panel (a) provides the results in the scenario where the testing set is selected at random. In terms of RMSE, both the SAGP and SGP models perform better with larger values of m. As expected, the full GP model attains the smallest RMSEs. The SAGP models with $m = 5$ and 10 perform better than the SGP models with the same number of pseudo-inputs. For $m = 15$, the median RMSEs in the SAGP and SGP models are similar, but the performance of the SAGP model is more consistent across simulations (the upper quartile of SAGP with $m = 15$ is considerably smaller than the one of SGP with $m = 15$). With the considered configuration of the parameters, the BART model performs slightly better than the SAGP model with $m = 5$, but worse than the SAGP model with $m = 10$ and $m = 15$. The coverage of the 95% PIs is close to the nominal level for all the methods except for BART. The PIs of the SAGP model appear to be slightly too narrow, as most of the coverages are a little lower than .95. SGP and GP models show coverages perfectly matching the nominal value.

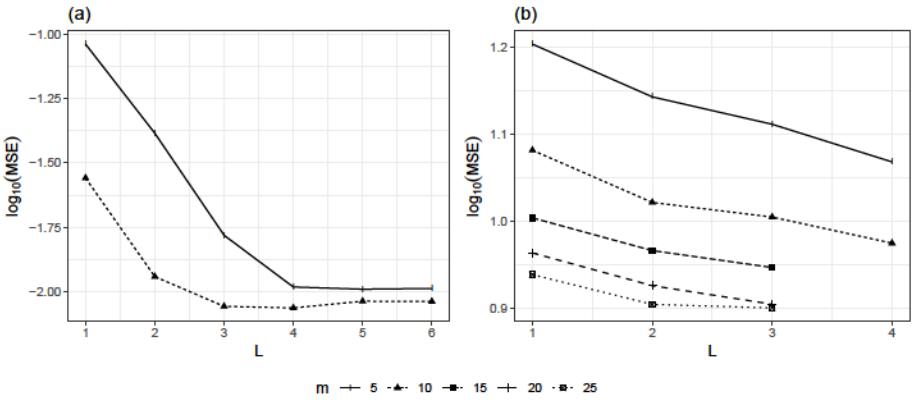


Figure 4: (a) Out-of-sample MSE (on $\log_{10}$ scale) attained by different models fitted on 1 dimensional heart-rate dataset with $m = 5, 10$ and $L = 1, \ldots, 6$.
(b) Out-of-sample MSE (on $\log_{10}$ scale) attained by different models fitted on 2 dimensional temperature dataset with $m = 5, 10, 15, 20, 25$ and $L = 1, \ldots, 4$. For any $L > 4$, our pruning algorithm 1 will reduce it to $L = 4$; for $m = 25$ our pruning algorithm will reduce SAGP model to $L = 3$

`www.image.ucar.edu/Data/US.monthly.met/USmonthlyMet.shtml`, US precipitation and temperature (1895-1997) dataset).

Qualitative comparisons of GP, BART and SAGP are shown in Figure 6. For GP regression we used MLE estimates with the Matern(5/2) kernel. For BART we use the default settings (Chipman et al., 2010). For SAGP, we choose $L = 3, m = 25$ and calibrate the α, β of the noise prior in SAGP and the noise estimate in BART according to MLE of noise estimate from GP. The GP model shows reasonable predictions, however, the prediction comes with high predictive variance in locations away from the observations and especially near the boundary (not shown). The predictive mean of BART shows it has a slight grid-like artifact due to its decision tree construction. In addition, the shape of the response around the mode is noticeably more rectangular than suggested by the other models.

This dataset provides us a 2-dimensional example where the data is limited, which is actually a disadvantage for SAGP since the sparsification does not cut down the computational cost significantly yet some information is lost in the procedure. Nonetheless, the SAGP method captures the major trends and even some of the extremal temperatures close to 40 degrees centigrade. Compared to BART and GP, the SAGP model behaves "in-between" these two methods and provides us with very competitive performance.

5.3 Ice Sheet Data

The Ice Sheet data is a larger 2-dimensional dataset but this time with noticeably uneven sampling as discussed in Park and Apley (2018).The response is ice sheet thickness in meters collected over a region of west Antarctica (Blankenship et al., 2004).We used the data from 1991, first converting the longitude and latitude into 2-dimensional Euclidean coordinates and standardizing the dataset to $[0, 1]^2$.

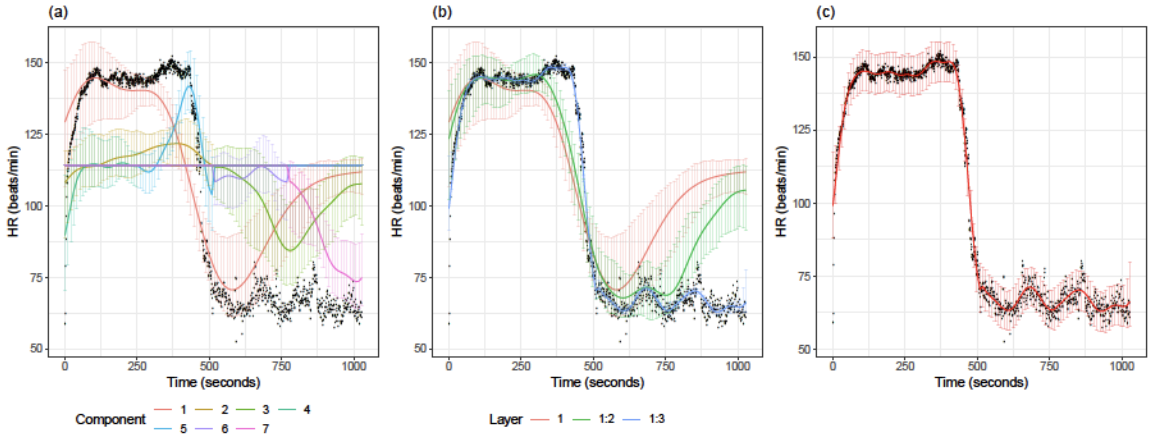
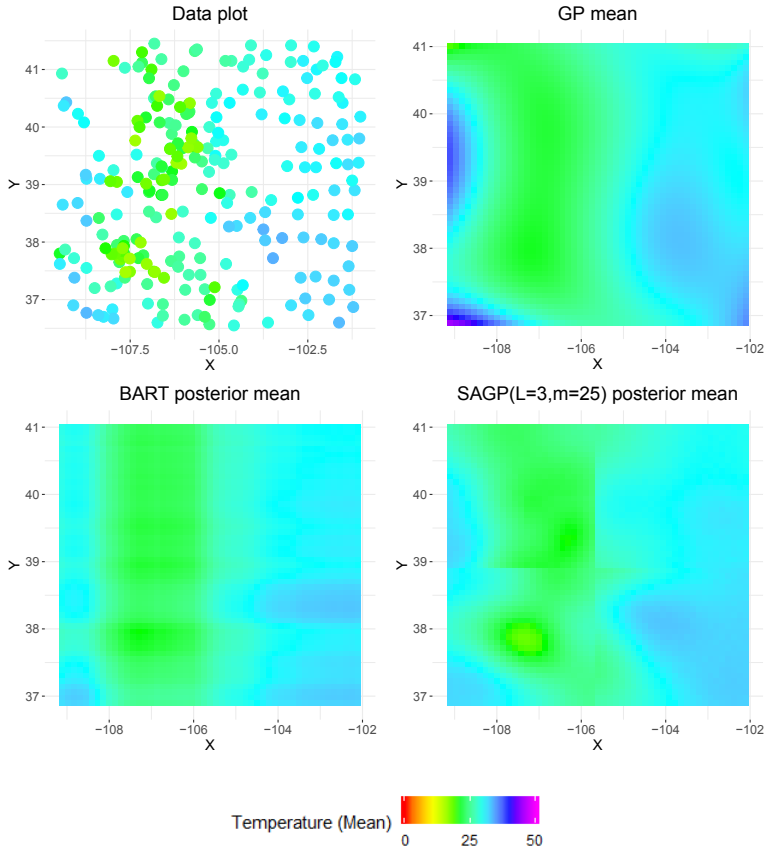
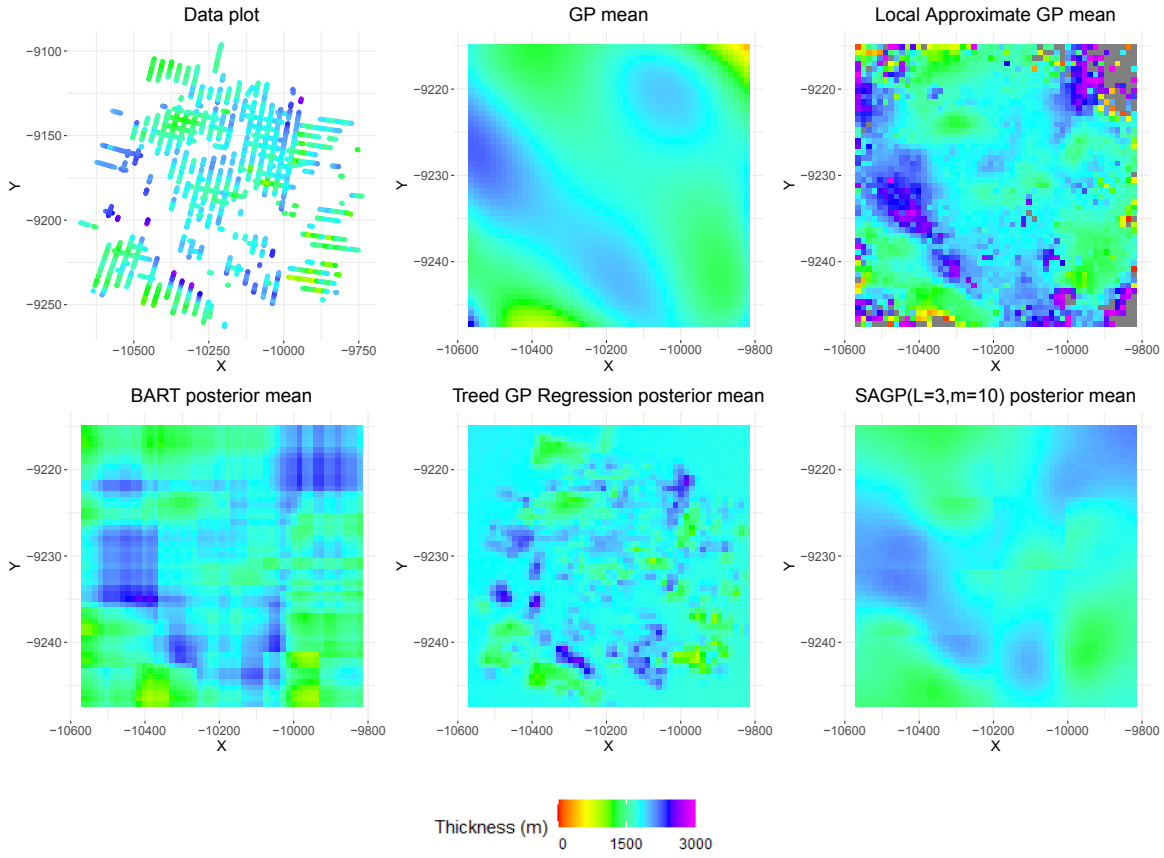


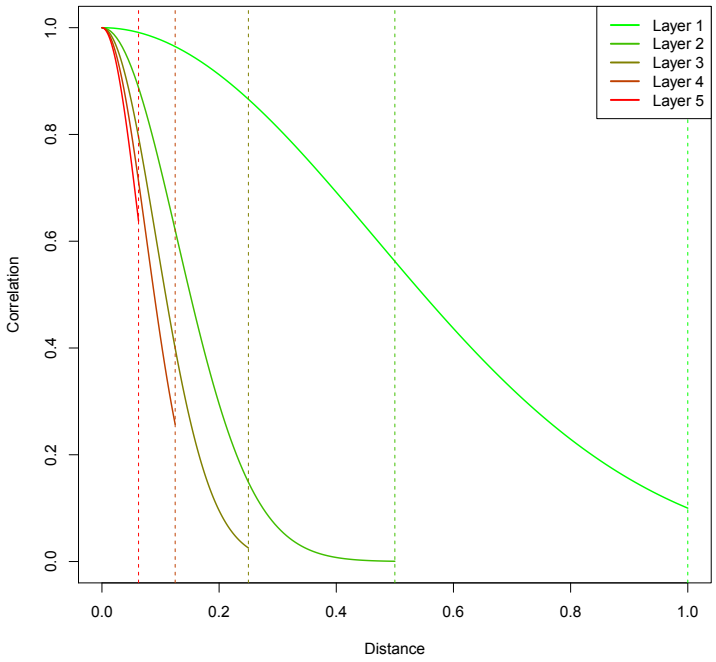
Figure 5: The panels show the observed HR values over time as black dots and results about the fit of the SAGP model with $m = 10$ and $L = 3$. Panel (a) shows the posterior means and the 95% CIs of the 7 additive components of the SAGP model on 100 equispaced locations on the support of the data. Panel (b) shows the posterior means and the 95% CIs of the sole component in layer 1 (red), of the components belonging to layer 1 and 2 (green) and of the complete model, including components from layer 1, 2 and 3. Panel (c) provides the predictive mean and the corresponding 95% prediction intervals.

A plot of the data and predictive fits for the GP (exponential correlation), laGP, BART, treed GP (TGP; (Gramacy et al., 2007)) and SAGP models are shown in Figure 7. We included TGP in this plot as we thought it may be helpful with the unevenly sampled data but did not end up including it in our overall quantitative results below. For the SAGP model, we show the fit obtained with $L = 3, m = 10$.

The fits obtained among these models show quite different behaviors. The full GP fit possess extreme boundary behavior due to the lack of data near the boundary. The BART model shows more noticeable grid-like artifacts in this dataset, but does not suffer from the boundary effects seen with the GP. The TGP regression also does not exhibit boundary effects but has much higher variability of the mean response in the data-rich region which does not agree with the other models. The dynamic partitioning of TGP also introduces considerable computational cost. The laGP model with its default settings and MSPE criterion exhibits some degree of variability in the fitted mean response, particularly near the boundaries, however, it is the most computationally efficient method.


Data plot
GP mean
BART posterior mean
SAGP(L=3,m=25) posterior mean
Y
X
41
40
39
38
37
−107.5
−105.0
−102.5
−108
−106
−104
−102
Temperature (Mean)
0
25
50


Data plot
GP mean
Local Approximate GP mean
BART posterior mean
Treed GP Regression posterior mean
SAGP(L=3,m=10) posterior mean
X
Y
Thickness (m)
0
1500
3000



$\mathcal{L}$

$\mathcal{X}$

$\mathcal{O}$ $\mathcal{X}$

$\mathcal{O}$ $_{\mathcal{L}}$ $\mathcal{X}$

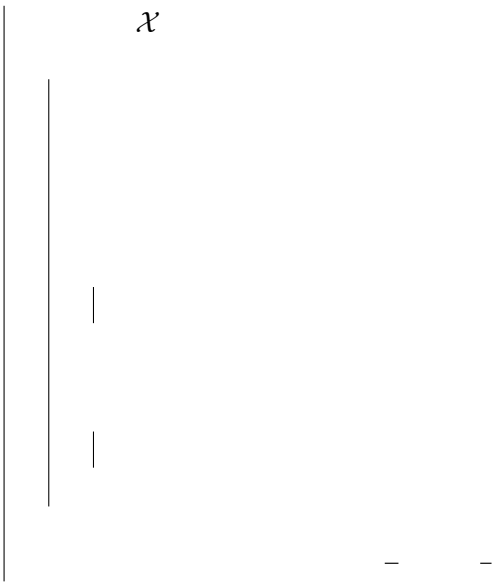
$\mathcal{O}$ $\mathcal{L}$

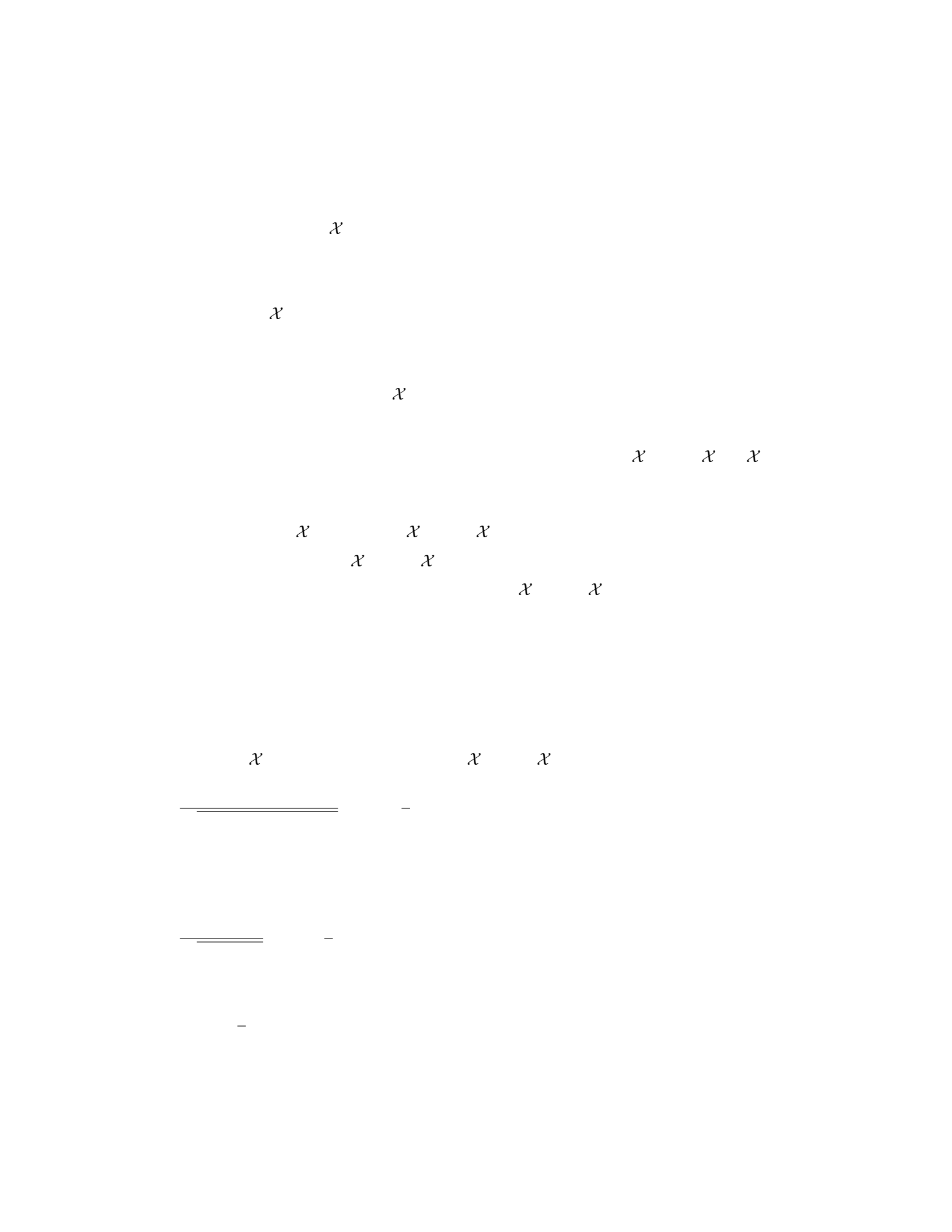
$\mathcal{O}$ $\mathcal{L}$ $\mathcal{O}$

$\mathcal{X}$

$\mathcal{X}$

$\mathcal{X}$





Appendix E. Diagnostic Statistics for the SAGP Model on 1000 Batches of Simulated Dataset

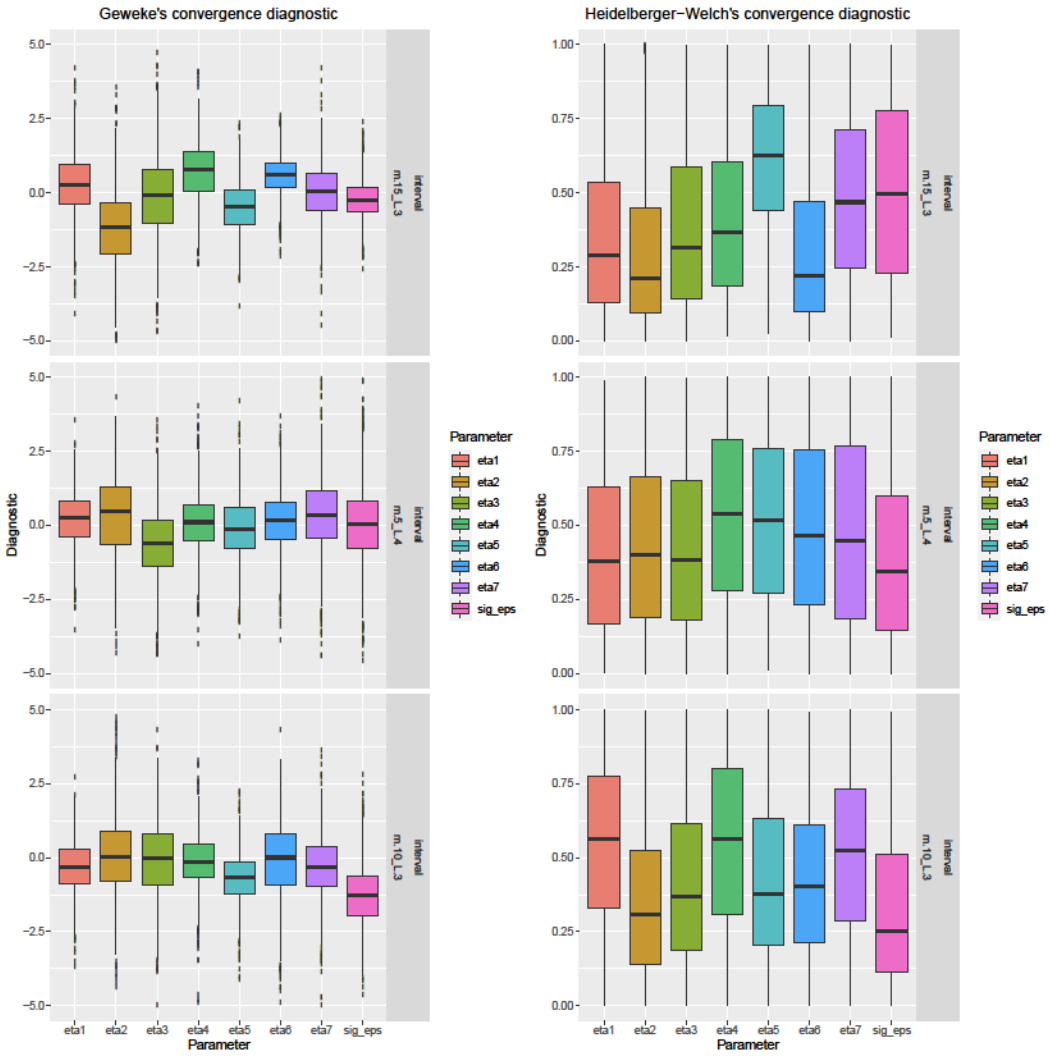


Figure 9: The panels of box plots show the Geweke's convergence diagnostic (Geweke et al., 1991) and Heidelberger-Welch's convergence diagnostic (Heidelberger and Welch, 1983) based on the MCMC sample of SAGP model, for parameter $\eta^{(j)}$ and σ_ϵ^2, calculated from the 1000 batches of simulated dataset from formula (11) with the testing set is random or interval.

Appendix F. Heart Rate Dataset Analyzed by SAGP Model Fitted with $m = 5$ and $L = 4$ (Figure 10)

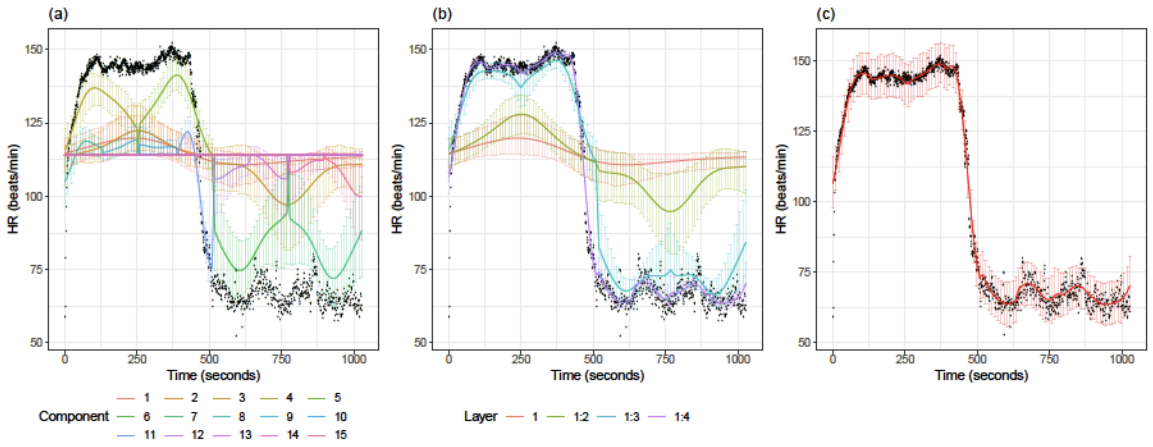


Figure 10: The panels show the observed HR values over time as black dots and results about the fit of the SAGP model with $m = 5$ and $L = 4$. Panel (a) shows the posterior means and the 95% CIs of the 15 additive components of the SAGP model on 100 equispaced locations on the support of the data. Panel (b) shows the posterior means and the 95% CIs of the sole component in layer 1 (red), of the components belonging to layer 1 and 2 (green) , of the components belonging to layer 1, 2, 3 and of the complete model, including components from layer 1, 2, 3 and 4. Panel (c) provides the predictive mean and the corresponding 95% prediction intervals.

References

Anjishnu Banerjee, David B. Dunson, and Surya T. Tokdar. Efficient gaussian process regression for large datasets. *Biometrika*, 100.1:75–89, 2012.

D. D. Blankenship et al. Ice thickness and surface elevation, southeastern ross embayment, west antarctica. *U.S. Antarctic Program (USAP) Data Center*, 2004. doi: doi:10.7265/N5WW7FKC. URL `https://www.usap-dc.org/view/dataset/609099`.

Richard Blundell, Alan Duncan, and Krishna Pendakur. Semiparametric estimation and consumer demand. *Journal of applied econometrics*, 13(5):435–461, 1998.

Leo Breiman. *Classification and Regression Trees*. Belmont, California: Wadsworth, 1984.

Thang D. Bui and Richard E. Turner. Tree-structured gaussian process approximations. In *Advances in Neural Information Processing Systems*, 2014.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bayesian cart model search. *Journal of the American Statistical Association*, 93.443:935–948, 1998.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4.1:266–298, 2010.