A New Extension For Bayesian Additive Regression Trees Model Adopting Gaussian Processes

Mateus Maia, Keefe Murphy & Andrew Parnell

CASI 2022 - 16th of May, 2022

A joint work

(Yes, they are real people)

Andrew C. Parnell

Keefe Murphy

Bayesian Additive Regression Trees (BART)

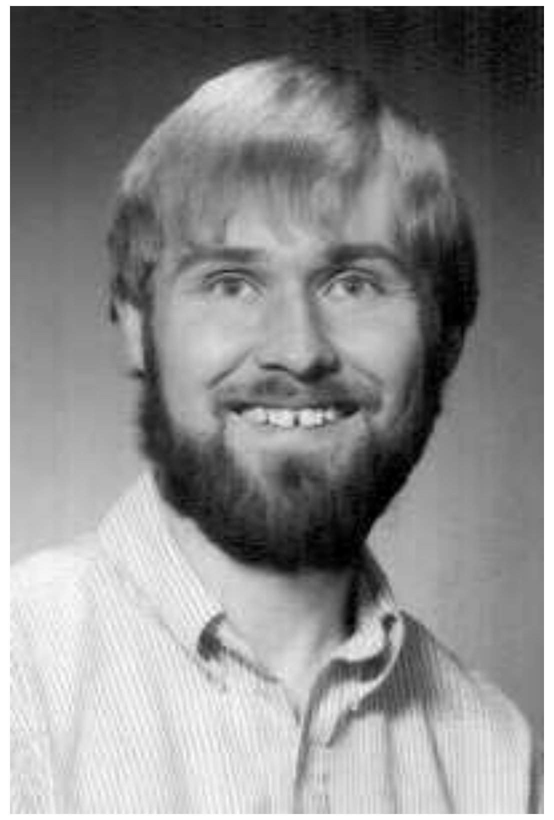
Bayesian Additive Regression Trees (BART)

March 2010

BART: Bayesian additive regression trees

Hugh A. Chipman, Edward I. George, Robert E. McCulloch

Ann. Appl. Stat. 4(1): 266-298 (March 2010). DOI: 10.1214/09-AOAS285



Huge A. Chipman

Edward I. George

Robert E. McCulloch

The model

Observed Output

Stochastic Error

$$y_i = f(\mathbf{x}_i) + \varepsilon \qquad \varepsilon \sim N(0, \tau^{-1})$$

Unknown Function

The model

Observed Output

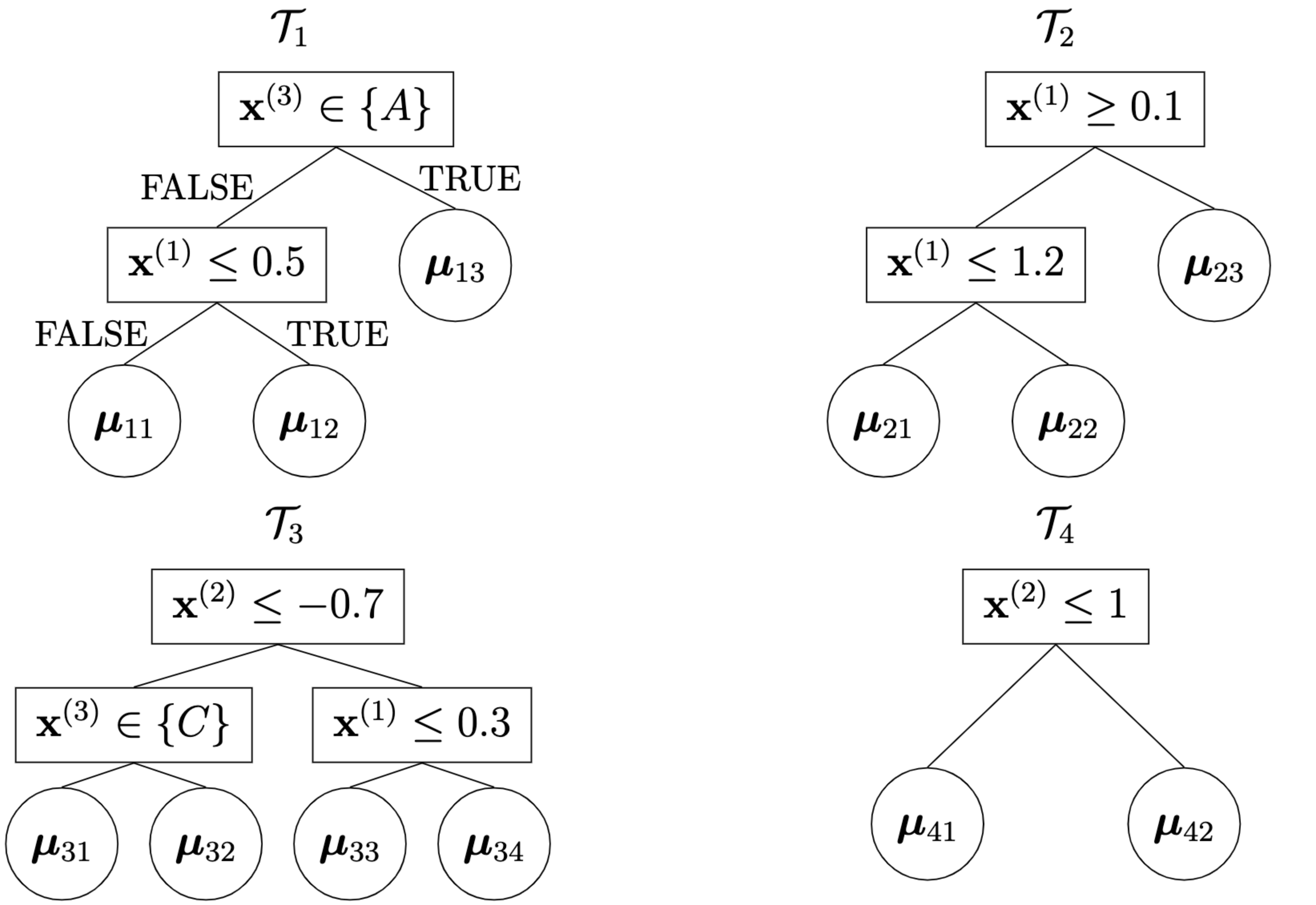
Stochastic Error

$$y_i = \sum_{t=1}^{T} g(\boldsymbol{x}_i, \mathcal{T}_t, M_t) + \varepsilon \qquad \varepsilon \sim N(0, \tau^{-1})$$

BART

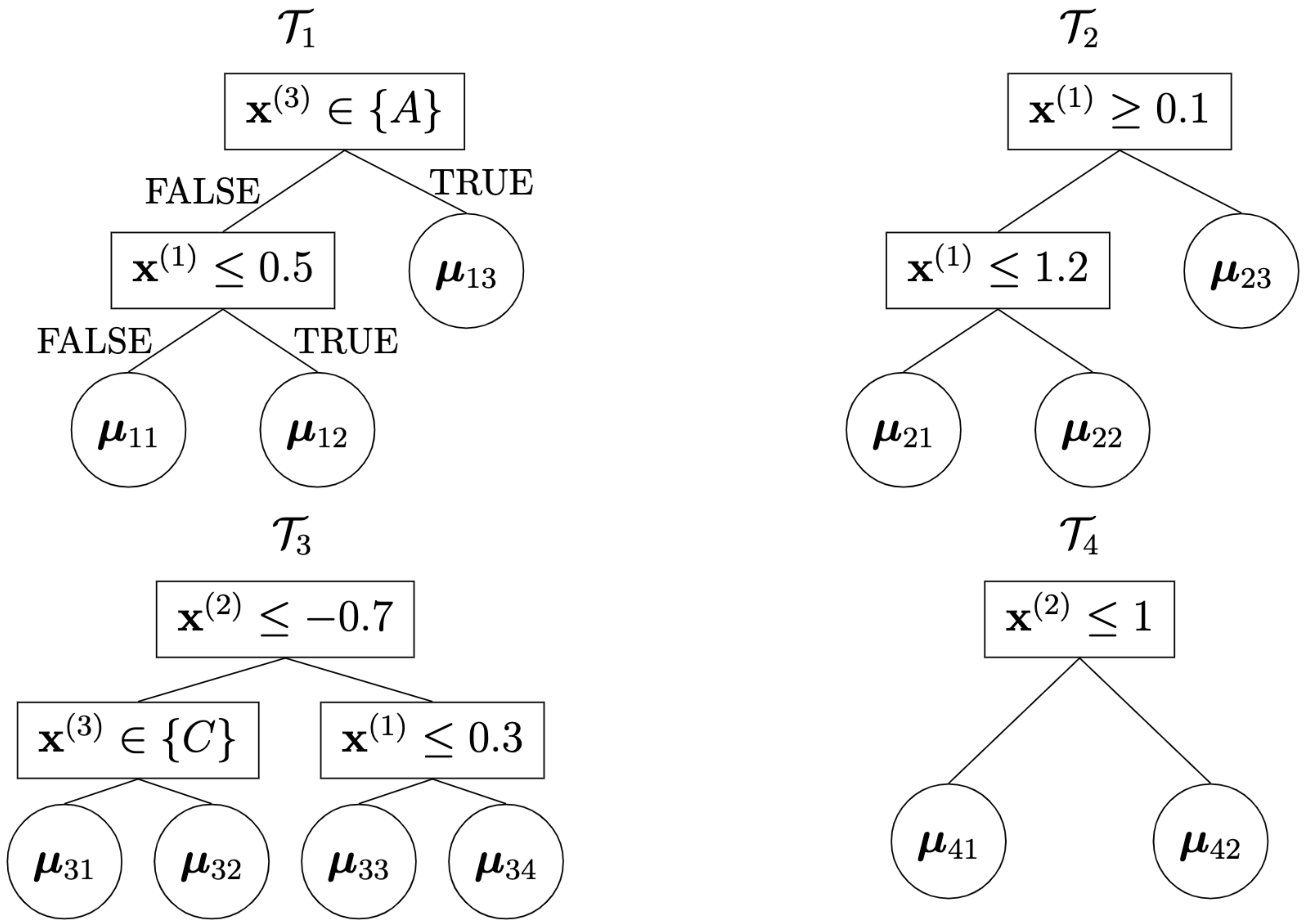
BART

$$y_i = \sum_{t=1}^{T} g(\boldsymbol{x}_i, \mathscr{T}_t, M_t) + \varepsilon$$



BART

$$y_i = \sum_{t=1}^{T} g(\boldsymbol{x}_i, \mathcal{T}_t, M_t) + \varepsilon$$



$\mu_{\ell t}$ priors:

$$\mu_{t\ell} \mid T_t, \tau \sim N(0, \tau_\mu^{-1})$$

$$\tau_\mu = \frac{k^2 m}{0.25}$$

Defining the prior and likelihood

$$y_i \sim N\left(\sum_{t=1}^{T} g(\boldsymbol{x}_i, M_t, \mathcal{T}_t), \tau^{-1}\right)$$

Posterior

$$p\left((M_1, T_1), \ldots, (M_T, \mathcal{T}_T), \tau \,|\, \boldsymbol{Y}, \boldsymbol{x}\right)$$

Back-Fitting MCMC algorithm

$$(T_t, M_t) \,|\, T_{-t}, M_{-t}, \tau, \mathbf{y}, \boldsymbol{X}$$

Only depends through

$$R_t \equiv \boldsymbol{y} - \sum_{k \neq t} g(\boldsymbol{X}, \mathcal{T}_t, M_t)$$

BART & Extensions

BART an illustrated example

$$y = sin(x) + N(0,\sigma^2)$$

BARTextensions

- **Piecewise-constant nature from the model ($\mu_{t\ell}$)**
  - **Soft-BART** - Linero, A. R., & Yang, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity Series B Statistical methodology.
  - **MOTR-BART** - Prado, E. B., Moral, R. A., & Parnell, A. C. (2021). Bayesian additive regression trees with model trees. Statistics and Computing, 31(3), 1-13.
- **Extensions to another applications**
  - **BAVART** - Huber, Florian, and Luca Rossini. "Inference in Bayesian additive vector autoregressive tree models." *The Annals of Applied Statistics* 16.1 (2022): 104-123.
  - **Survival BART** - Sparapani, Rodney A., et al. "Nonparametric survival analysis using Bayesian additive regression trees (BART)." *Statistics in medicine* 35.16 (2016): 2741-2753.

Spatial Data

"With the advancement of GPS and remote sensing technologies, large amounts of geospatial data are being collected from various domains."

"Zhe Jhiang, 2018"

In traditional statistical learning model, one of the main assumptions is that the samples are assumed to be independent and identically distributed (i.i.d).

$$y(\boldsymbol{s}) = f(\boldsymbol{x}(\boldsymbol{s})), \forall \boldsymbol{s}$$

Spatial Data

“With the advancement of GPS and remote sensing technologies, large amounts of geospatial data are being collected from various domains.”

“Zhe Jhiang, 2018”

In traditional statistical learning model, one of the main assumptions is that the samples are assumed to be independent and identically distributed (i.i.d).

Gaussian Processes (GP)

Gaussian Process

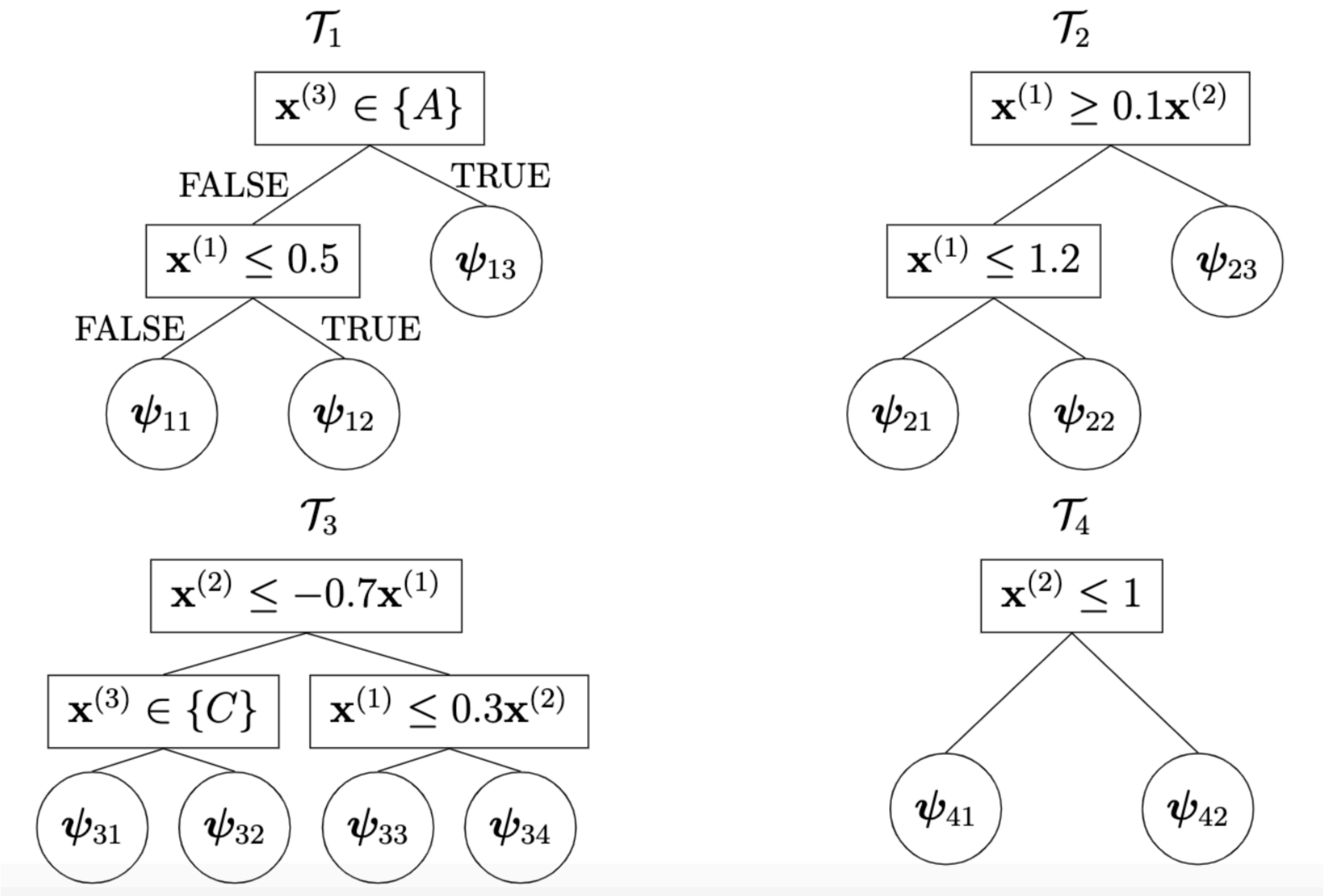
Bayesian
Additive Regression
Trees

(GP-BART)

GP-BART

$$y_i = \sum_{t=1}^{T} g\left(\mathbf{x}_i; \mathcal{T}_t, \mathbf{G}_t\right) + \varepsilon$$

$$\mathbf{G}_t = (\{\mu_{t1}, \phi_t, \nu\}, \ldots, \{\mu_{tb_t}, \phi_t, \nu\})$$



GP-BART

$$y_i = \sum_{t=1}^{T} g\left(\mathbf{x}_i; \mathcal{T}_t, \mathbf{G}_t\right) + \varepsilon$$

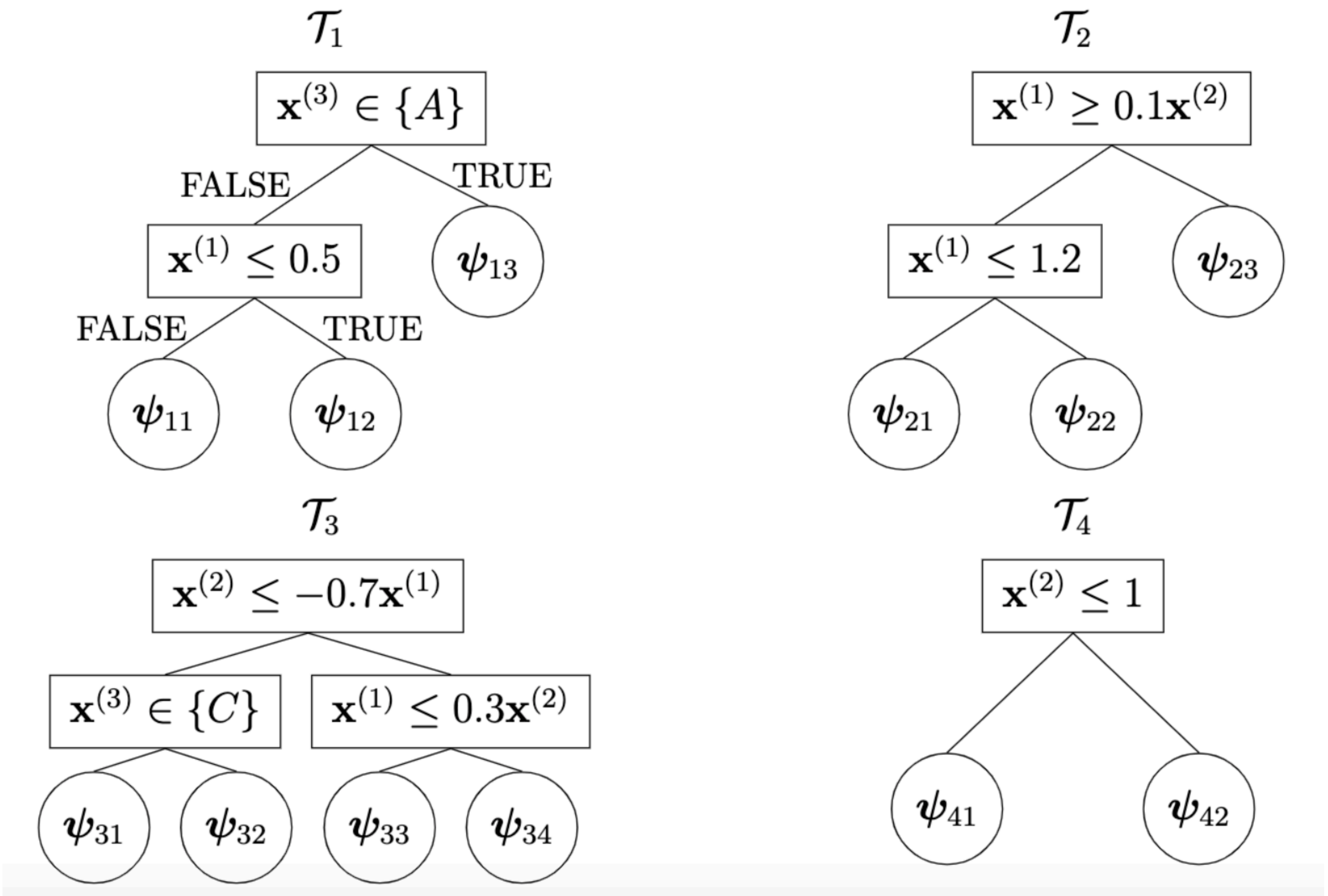
$$\mathbf{G}_t = (\{\mu_{t1}, \phi_t, \nu\}, \ldots, \{\mu_{tb_t}, \phi_t, \nu\})$$

$\psi_{t\ell}$ priors:

$$\boldsymbol{\psi}_{t\ell} \,|\, \mathcal{T}_t, \mu_{t\ell}, \phi_t, \nu, \kappa$$

$$\sim$$

$$\text{MVN}\left(\boldsymbol{\mu}_{t\ell} = \mu_{t\ell}\mathbf{1}_{n_{t\ell}}, (1-\kappa)\,\boldsymbol{\Omega}_{t\ell}\right)$$



GP-BART priors

GP-BART priors

GP-prior

$$\boldsymbol{\psi}_{t\ell} \mid \mathcal{T}_t, \mu_{t\ell}, \phi_t, \nu, \kappa$$

$$\sim$$

$$\text{MVN}\left(\boldsymbol{\mu}_{t\ell} = \mu_{t\ell}\mathbf{1}_{n_{t\ell}}, (1-\kappa)\,\boldsymbol{\Omega}_{t\ell}\right)$$

Covariance Matrix

$$\boldsymbol{\Omega}_{t\ell} = \nu^{-1}\exp\left\{-\frac{\|\mathbf{w}-\mathbf{w}'\|_2^2}{2\phi_t^2}\right\}$$

GP-BART priors

GP-prior

$$\boldsymbol{\psi}_{t\ell} \mid \mathcal{T}_t, \mu_{t\ell}, \phi_t, \nu, \kappa$$
$$\sim$$
$$\text{MVN}\left(\boldsymbol{\mu}_{t\ell} = \mu_{t\ell}\mathbf{1}_{n_{t\ell}}, (1-\kappa)\,\boldsymbol{\Omega}_{t\ell}\right)$$

Covariance Matrix

$$\boldsymbol{\Omega}_{t\ell} = \nu^{-1}\exp\left\{-\frac{\|\mathbf{w}-\mathbf{w}'\|_2^2}{2\phi_t^2}\right\}$$

$\mu_{t\ell}$ prior

$$\mu_{t\ell} \mid \mathcal{T}_t, \kappa \sim \text{N}\left(\mu_\mu, \kappa\tau_\mu^{-1}\right)$$

GP-BART priors

GP-prior

$$\boldsymbol{\psi}_{t\ell} \mid \mathcal{T}_t, \mu_{t\ell}, \phi_t, \nu, \kappa$$
$$\sim$$
$$\text{MVN}\left(\boldsymbol{\mu}_{t\ell} = \mu_{t\ell}\mathbf{1}_{n_{t\ell}}, (1-\kappa)\,\boldsymbol{\Omega}_{t\ell}\right)$$

Covariance Matrix

$$\boldsymbol{\Omega}_{t\ell} = \nu^{-1}\exp\left\{-\frac{\|\mathbf{w}-\mathbf{w}'\|_2^2}{2\phi_t^2}\right\}$$

$\mu_{t\ell}$ prior

$$\mu_{t\ell} \mid \mathcal{T}_t, \kappa \sim \text{N}\left(\mu_\mu, \kappa\tau_\mu^{-1}\right)$$

ϕ_t prior

$$\phi_t \sim \text{Unif}\left(\min\{D(\mathbf{X},\mathbf{X})\}, \max\{D(\mathbf{X},\mathbf{X})\}\right)$$

GP-BART priors

GP-prior

$$\boldsymbol{\psi}_{t\ell} \mid \mathcal{T}_t, \mu_{t\ell}, \phi_t, \nu, \kappa$$
$$\sim$$
$$\text{MVN}\left(\boldsymbol{\mu}_{t\ell} = \mu_{t\ell}\mathbf{1}_{n_{t\ell}}, (1-\kappa)\,\boldsymbol{\Omega}_{t\ell}\right)$$

Covariance Matrix

$$\boldsymbol{\Omega}_{t\ell} = \nu^{-1}\exp\left\{-\frac{\|\mathbf{w}-\mathbf{w}'\|_2^2}{2\phi_t^2}\right\}$$

$\mu_{t\ell}$ prior

$$\mu_{t\ell} \mid \mathcal{T}_t, \kappa \sim \text{N}\left(\mu_\mu, \kappa\tau_\mu^{-1}\right)$$

ϕ_t prior

$$\phi_t \sim \text{Unif}\left(\min\{D(\mathbf{X},\mathbf{X})\}, \max\{D(\mathbf{X},\mathbf{X})\}\right)$$

Shrinkage Parameters

$$\nu = \tau_\mu = 4k^2T$$

GP-BART sampling

Algorithm 1: GP-BART sampling algorithm

Input: $\mathbf{X}$, $\mathbf{y}$, T, N_{MCMC}, and all hyperparameters of the priors.

Initialise: T tree stumps s.t. $\mu_{t\ell} = 0$, $\phi_t = 0.1$, and $\tau = 1$, $\forall\,(t, \ell)$.

for *iterations* m *from* 1 *to* N_{MCMC} **do**

  for *trees* t *from* 1 *to* T **do**

    Calculate the partial residuals $\mathbf{R}_t$ via Equation (10);

    Propose a new tree $\mathcal{T}_t^\star$ by a grow, grow-project, prune, change, change-project, or swap move;

    Accept and update $\mathcal{T}_t = \mathcal{T}_t^\star$ with probability

$$\gamma^\star\left(\mathcal{T}_t, \mathcal{T}_t^\star\right) = \min\left\{1, \frac{\pi\left(\mathbf{R}_t \mid \mathcal{T}_t^\star, \phi_t, \tau, \nu, \kappa, \tau_\mu\right)\pi\left(\mathcal{T}_t^\star\right)}{\pi\left(\mathbf{R}_t \mid \mathcal{T}_t, \phi_t, \tau, \nu, \kappa, \tau_\mu\right)\pi\left(\mathcal{T}_t\right)}\right\}$$

    for *terminal nodes* ℓ *from* 1 *to* b_t **do**

      Update $\mu_{t\ell}$ via Equation (11);

      Update $\boldsymbol{\psi}_{t\ell}$ via Equation (12).

    end

    Update ϕ_t using MH.

  end

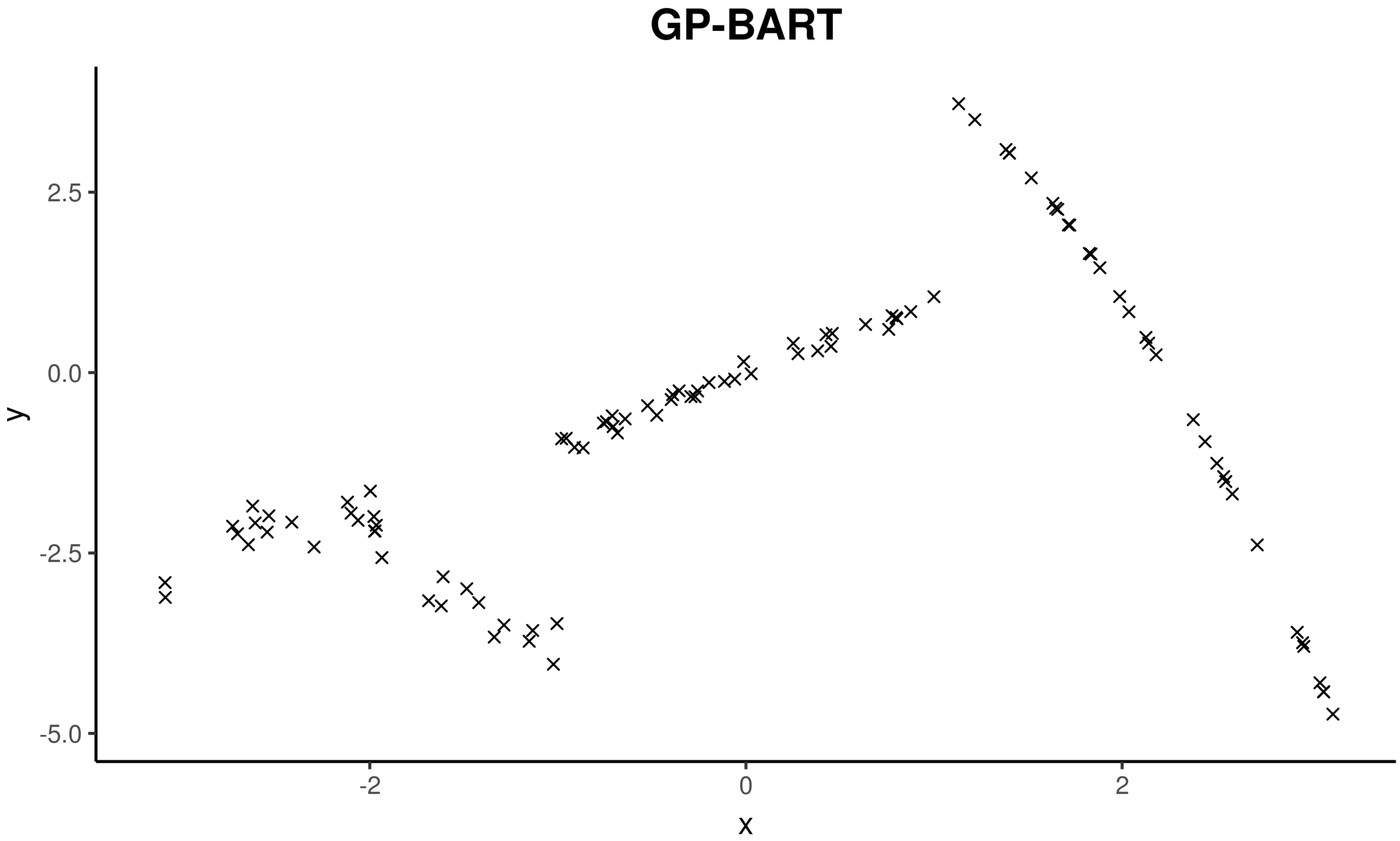
  Update τ via Equation (13).

end

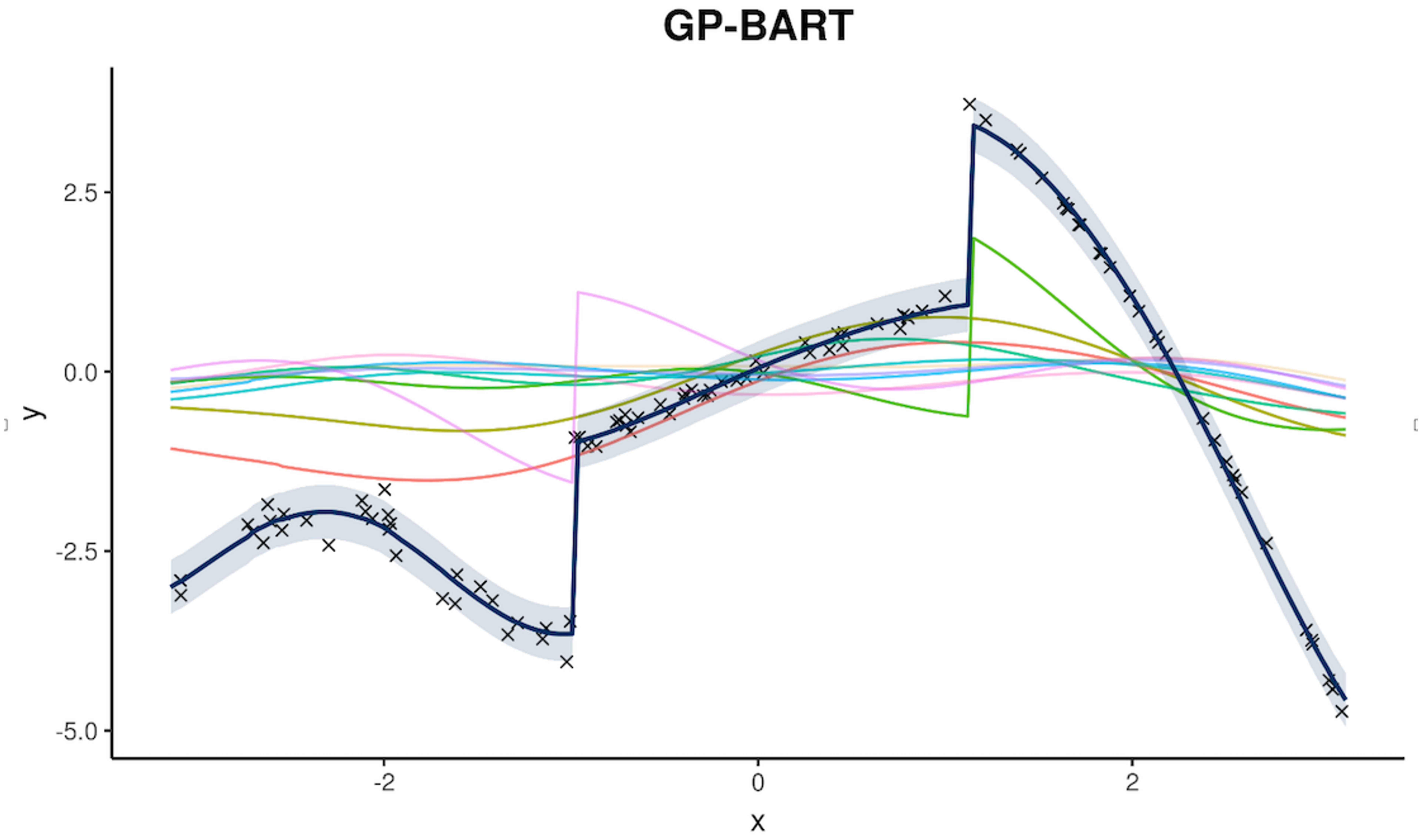
Output: Approx. samples from $\pi\left((\mathcal{T}_1, \mathbf{G}_1), \ldots, (\mathcal{T}_T, \mathbf{G}_T), \tau \mid \mathbf{y}\right)$.

GP-BART
In action

GP-BART an illustrated example



GP-BART an illustrated example



Simulations

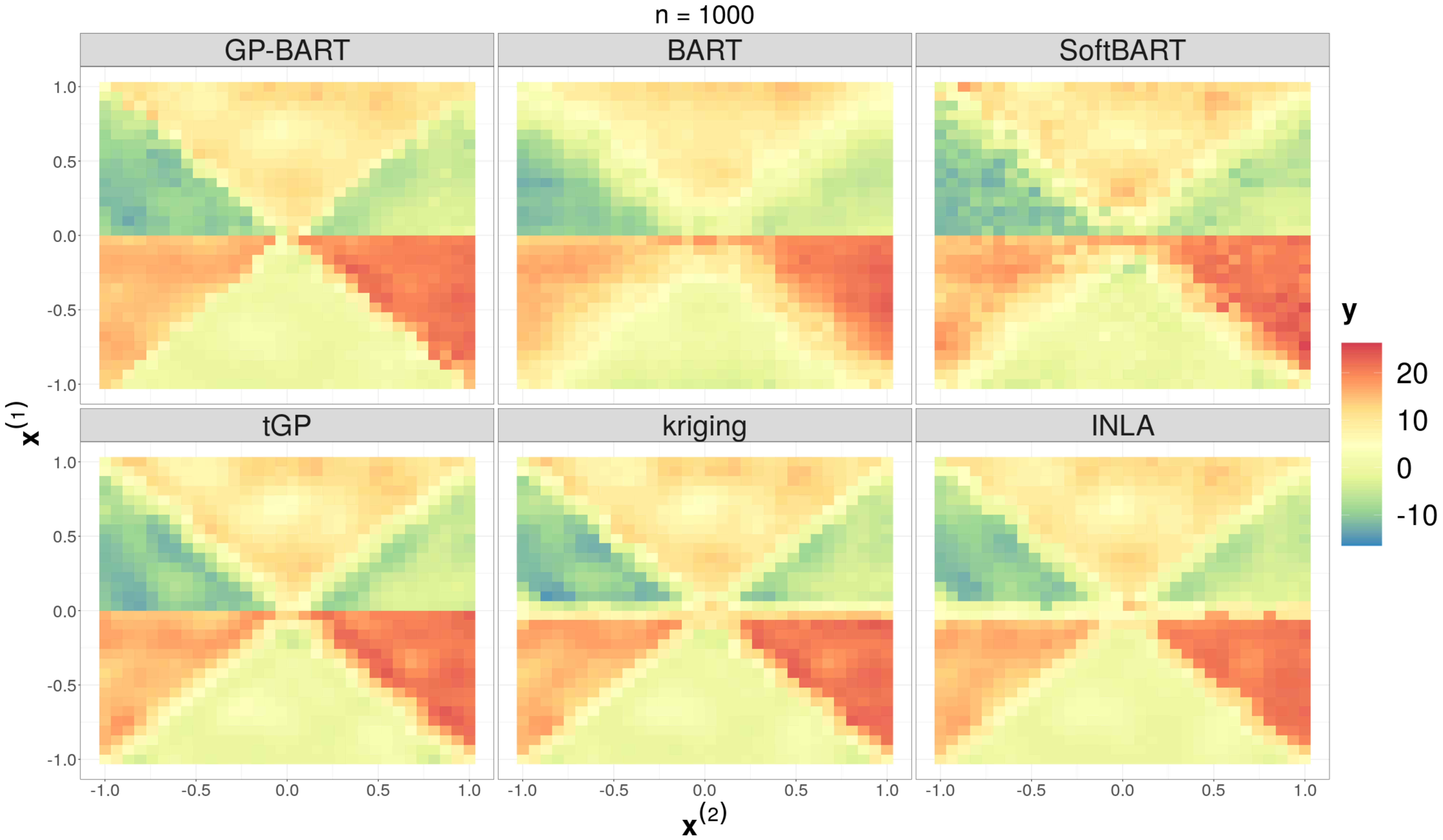
Comparison and Cross-validation

- **Compared with:**
  - BART (Chipman, *et. al* 2010)
  - SoftBART (Linero, 2018)
  - tGP (Gramacy, *et. al, 2008)*
  - Kriging
  - INLA (Rue, *et. al, 2015)*

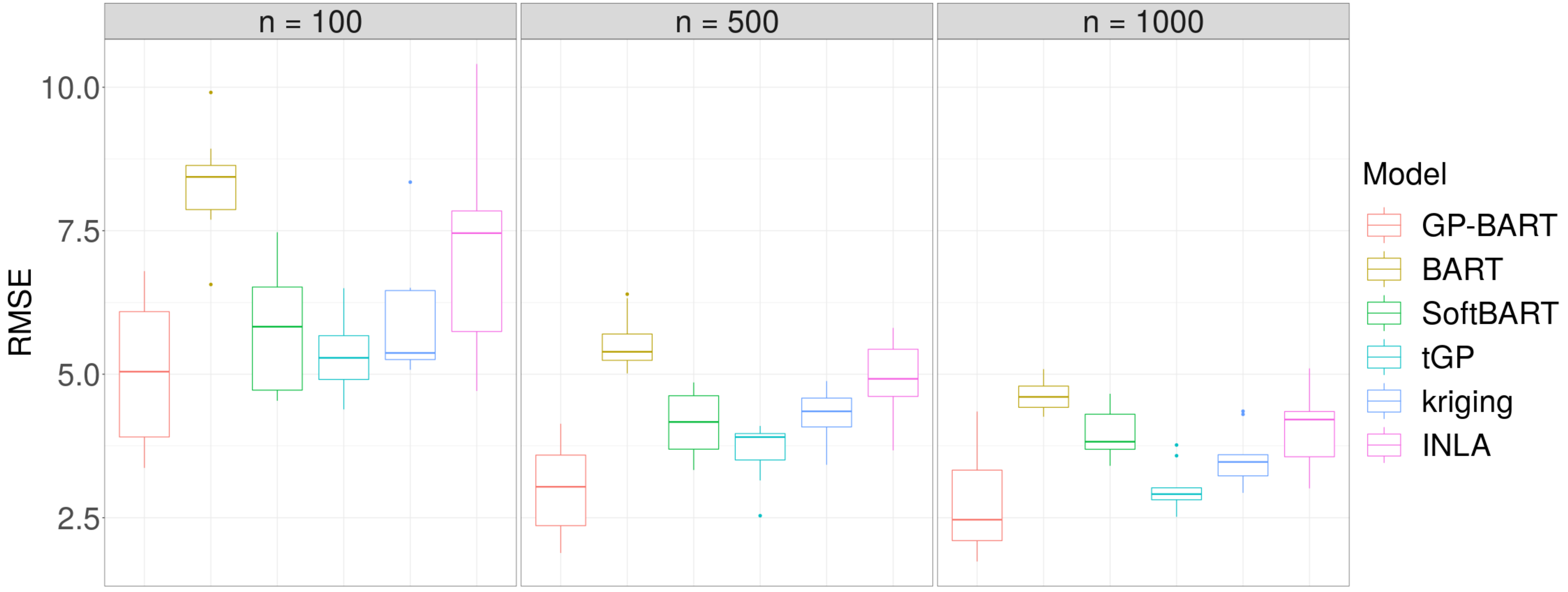
Evaluation:

- 10-fold
- Root Mean Squared Error (RMSE)
- Coverage in a 50% prediction interval.

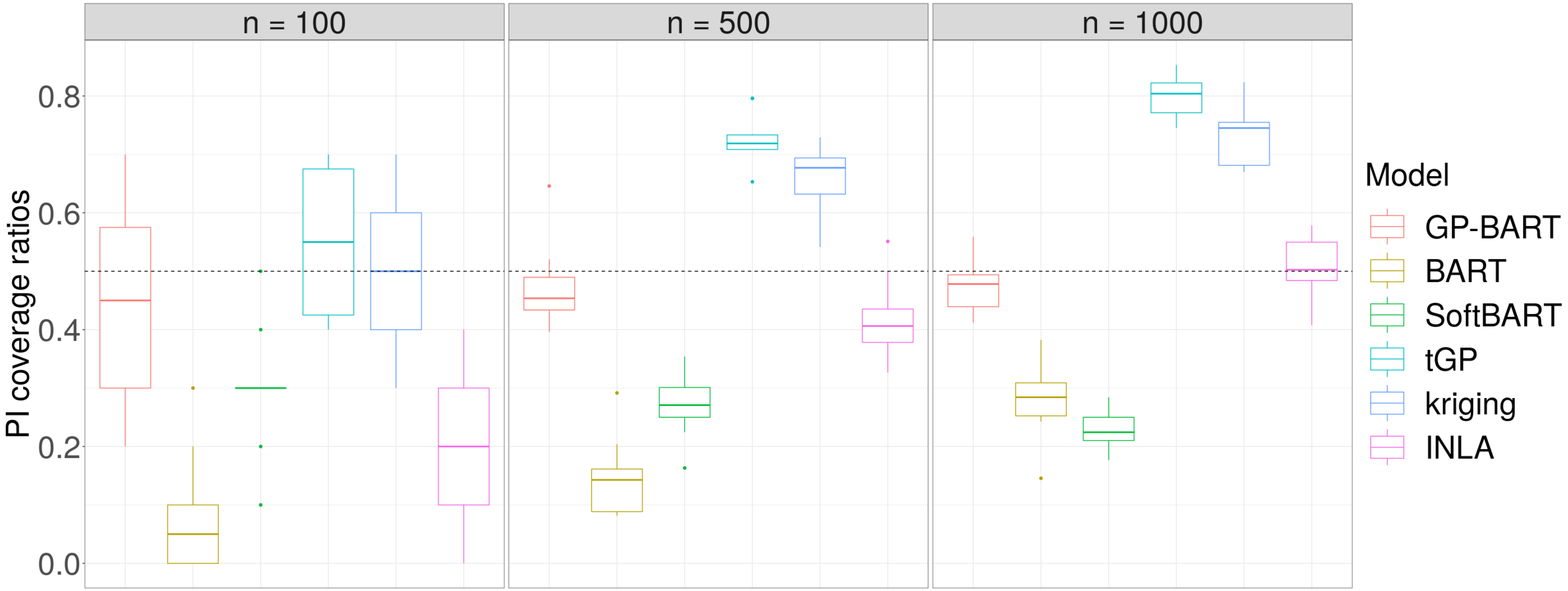
Simulations (n =1000)



Quantitative Evaluation



Quantitative Evaluation



GP-BART
In action

GP-BART

~~In action~~

Real action

Benchmarkings

- ***Auckland*** dataset dataset consists of 166 rows describing infant mortality in Auckland, with two spatial covariates and the target variable.
- ***Baltimore*** dataset comprises 221 observations of house sales prices, two spatial features, and 13 covariates characterising each property.
- ***Boston*** dataset contains 506 observations of the median values of owner-occupied suburban homes, two spatial features, and 13 related covariates.
- ***Swmud*** is a dataset of seabed mud content in the southwest Australia Exclusive Economic Zone with 177 observations of two sets of spatial coordinates and mud content as the target variable

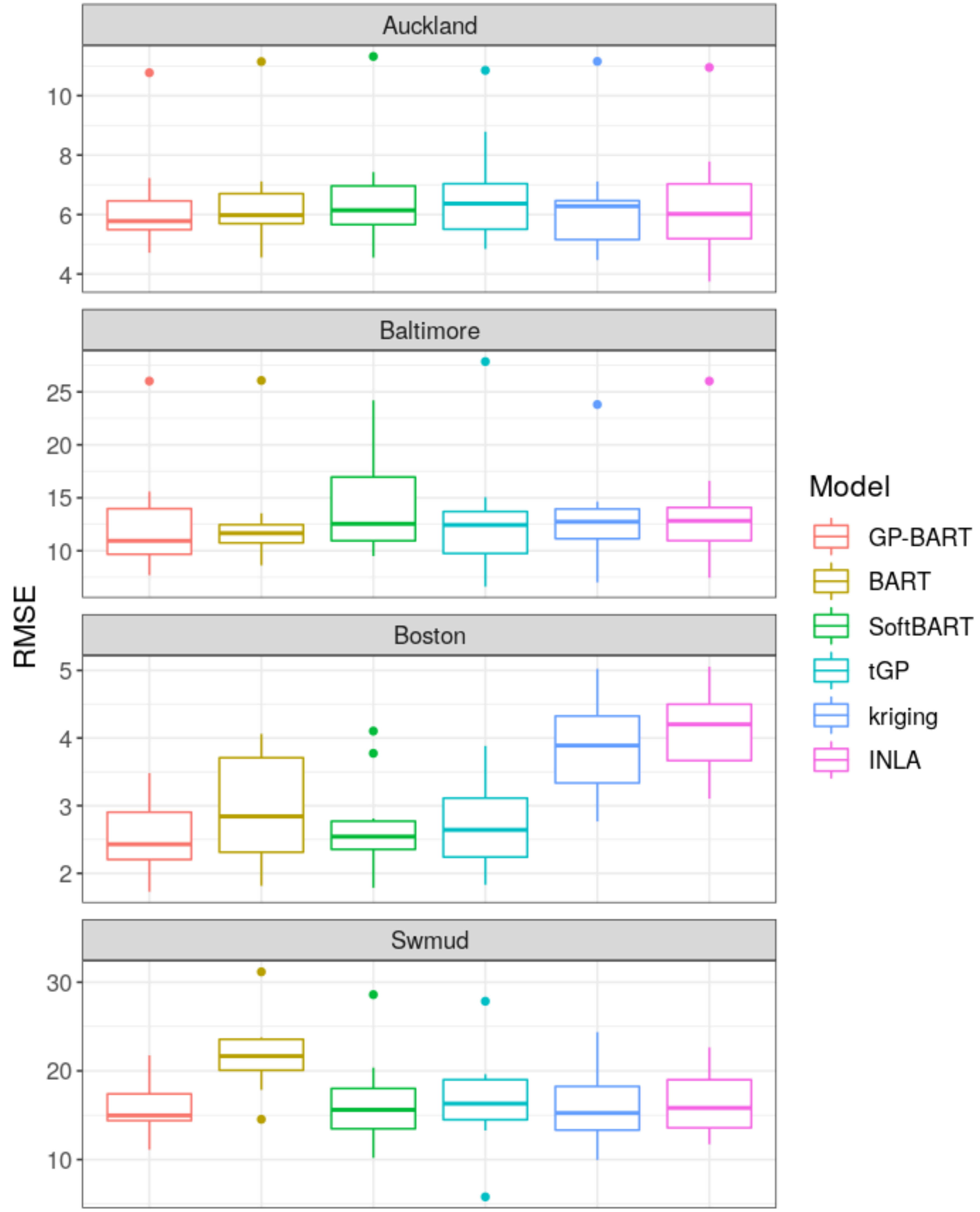
Benchmarkings

- ***Auckland*** dataset dataset consists of 166 rows describing infant mortality in Auckland, with two spatial covariates and the target variable.
- ***Baltimore*** dataset comprises 221 observations of house sales prices, two spatial features, and 13 covariates characterising each property.
- ***Boston*** dataset contains 506 observations of the median values of owner-occupied suburban homes, two spatial features, and 13 related covariates.
- ***Swmud*** is a dataset of seabed mud content in the southwest Australia Exclusive Economic Zone with 177 observations of two sets of spatial coordinates and mud content as the target variable

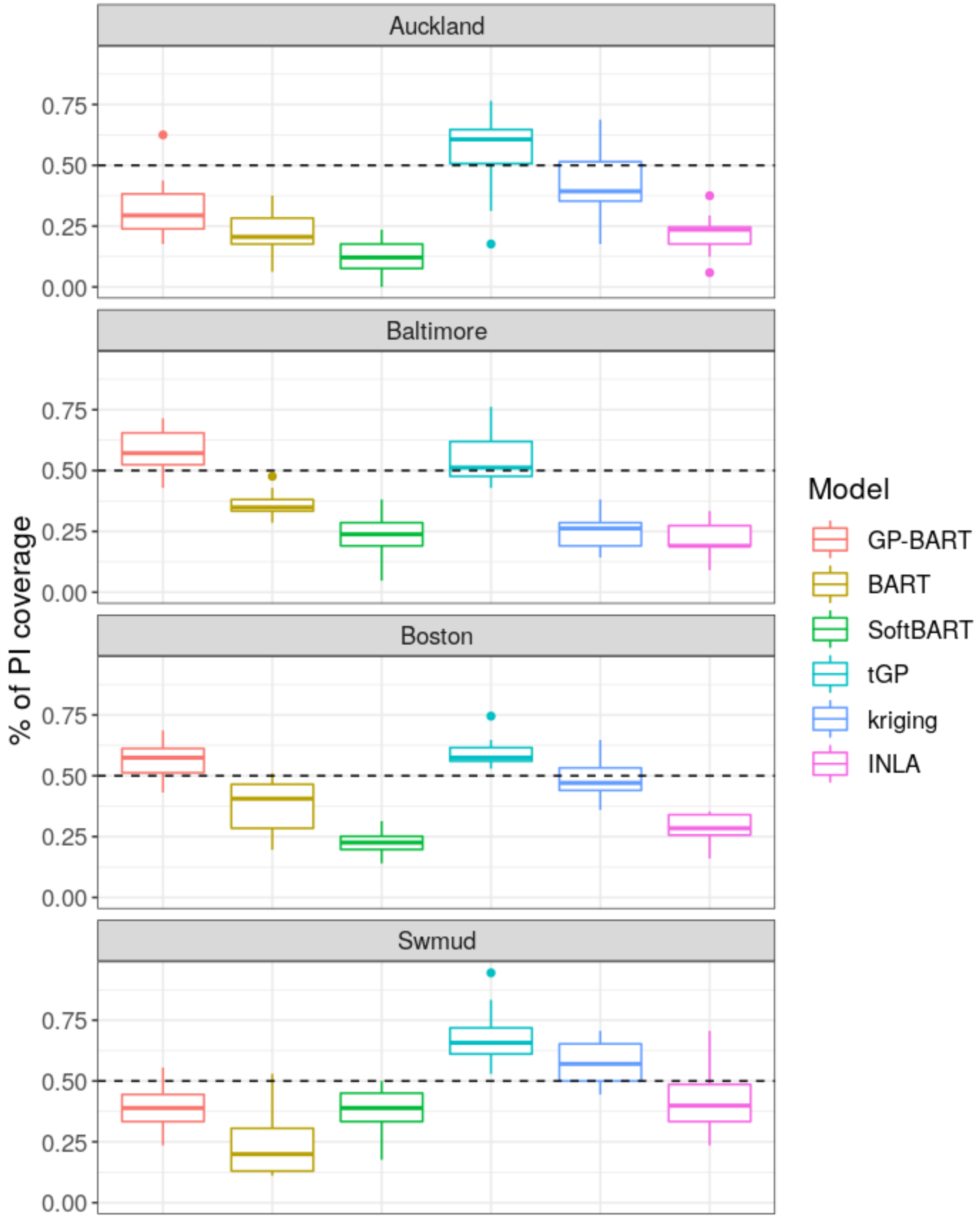
!

Cross-validation and evaluation settings are the same from simulations

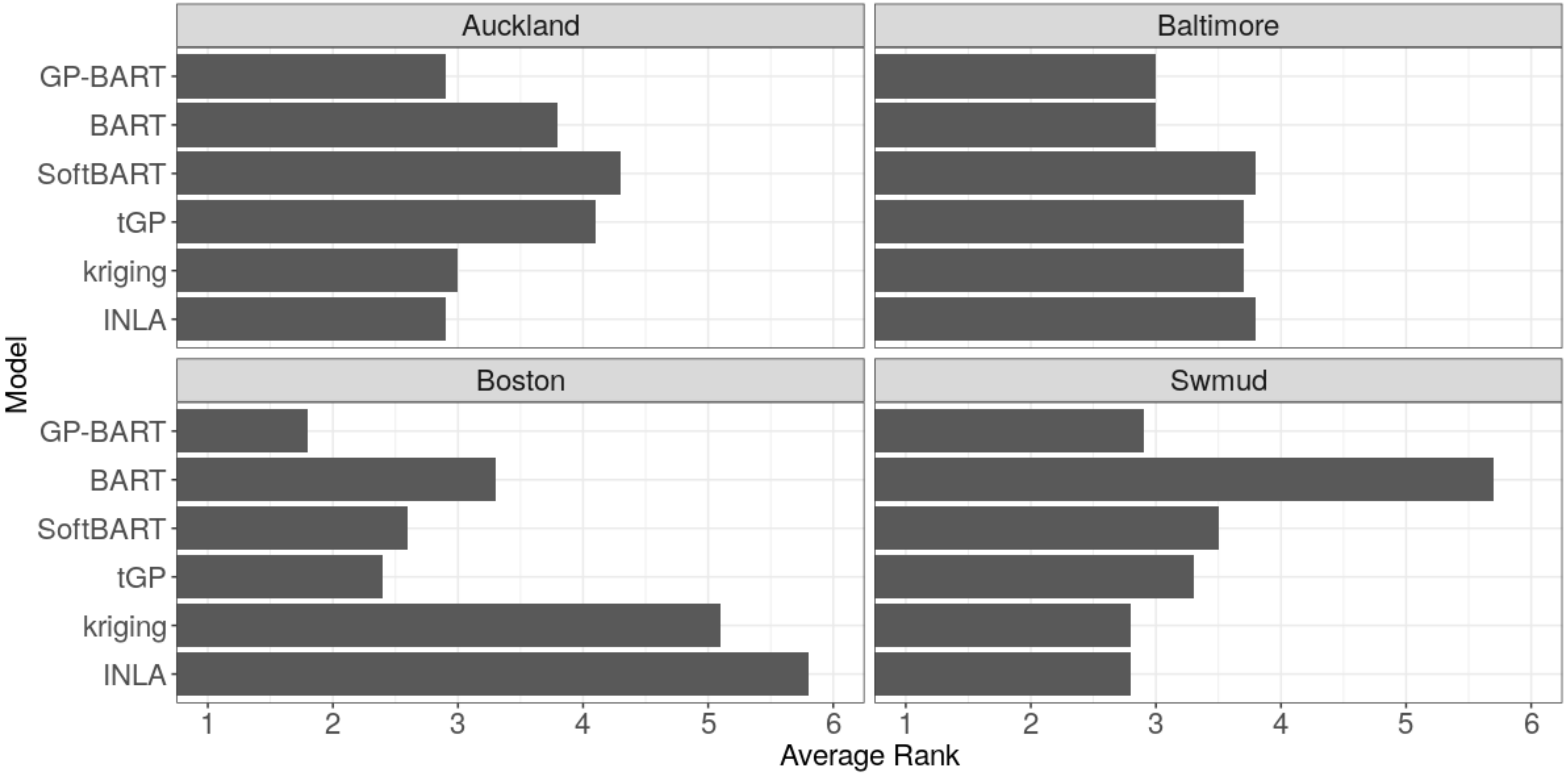
Quantitative Evaluation



Quantitative Evaluation



Quantitative Evaluation



Highlights
&
Challenges

Highlights

- GP-BART provides a model that is able to capture non-linear relations and spatial dependence
- Contains a theoretical justification of the prior structure of the Gaussian Processes applied to terminal nodes of the trees.
- Achieved excellent prediction and uncertainty estimations for simulated and real data
- Available to use by the **R** package **gpbart** at: https://github.com/MateusMaiaDS/gpbart.

Challenges

- The model can be computationally challenging to fit larger datasets
- A superior approach may introduce non-stationarity to the auto covariance structure.

References

- Maia, Mateus, Keefe Murphy, and Andrew C. Parnell. "GP-BART: a novel Bayesian additive regression trees approach using Gaussian processes." arXiv preprint arXiv:2204.02112 (2022).

- Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. "BART: Bayesian additive regression trees." The Annals of Applied Statistics 4.1 (2010): 266-298.

- Jiang, Zhe. "A survey on spatial prediction methods." IEEE Transactions on Knowledge and Data Engineering 31.9 (2018): 1645-1664.

- Linero, Antonio R., and Yun Yang. "Bayesian regression tree ensembles that adapt to smoothness and sparsity." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80.5 (2018): 1087-1110.

- Lindgren, Finn, and Håvard Rue. "Bayesian spatial modelling with R-INLA." Journal of statistical software 63 (2015): 1-25.

- Gramacy, Robert B., and Herbert K. H. Lee. "Bayesian treed Gaussian process models with an application to computer modeling." Journal of the American Statistical Association 103.483 (2008): 1119-1130.

Acknowledgements

- Mateus Maia's work was supported by Science Foundation Ireland Career Development Award grant number 17/CDA/4695 and SFI research centre award 12/RC/2289P2.

sfi
Fondúireacht Eolaíochta Éireann
Dá bhfuil romhainn
Science Foundation Ireland
For what's next