Spatial Inference with R-package BRISC

CUGOS, 2023 Spring Fling

Arkajyoti Saha, April 21, 2023
University of Washington
Department of Statistics

Outline

- What problem does BRISC solve?
- What can you do with BRISC?
- Applications of BRISC.

Outline

- **What problem does BRISC solve?**
- What can you do with BRISC?
- Applications of BRISC.

Geospatial/point-referenced data

Data: $\{(Y_i, X_i, s_i)\ : i = 1,\ldots,n\}$

- $\mathbf{S} = \left(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\right)$: locations
- $\mathbf{Y} = \left(y\left(\mathbf{s}_1\right), y\left(\mathbf{s}_2\right), \ldots, y\left(\mathbf{s}_n\right)\right)$: observed response
- $\mathbf{X} = \left(\mathbf{x}\left(\mathbf{s}_1\right), \mathbf{x}\left(\mathbf{s}_2\right), \ldots, \mathbf{x}\left(\mathbf{s}_n\right)\right)$: explanatory variables

Geospatial/point-referenced data

Data: $\{(Y_i, X_i, s_i)\ :\ i = 1, \ldots, n\}$

- $\mathbf{S} = \left(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\right)$: locations
- $\mathbf{Y} = \left(y\left(\mathbf{s}_1\right), y\left(\mathbf{s}_2\right), \ldots, y\left(\mathbf{s}_n\right)\right)$: observed response
- $\mathbf{X} = \left(\mathbf{x}\left(\mathbf{s}_1\right), \mathbf{x}\left(\mathbf{s}_2\right), \ldots, \mathbf{x}\left(\mathbf{s}_n\right)\right)$: explanatory variables

Objectives:

- Understand relationship between X and Y.
- Inference on spatial structure.
- Predict at a new location s_0.

How do we currently model this?

How do we currently model this?

Classical solution: **Ordinary Least Square Regression (OLS)**

$$Y(s) = X(s)\boldsymbol{\beta} + \epsilon(s)$$

How do we currently model this?

Classical solution: **Ordinary Least Square Regression (OLS)**

$$Y(s) = X(s)\boldsymbol{\beta} + \epsilon(s)$$

Linear Covariate effect

How do we currently model this?

Classical solution: **Ordinary Least Square Regression (OLS)**

$$Y(s) = X(s)\boldsymbol{\beta} + \epsilon(s)$$

Linear Covariate effect

White noise

How do we currently model this?

Classical solution: **Ordinary Least Square Regression (OLS)**

$$Y(s) = X(s)\boldsymbol{\beta} + \epsilon(s)$$

Linear Covariate effect

White noise

Usually modeled with $N\left(0, \tau^2\right)$

How do we currently model this?

Classical solution: **Ordinary Least Square Regression (OLS)**

$$Y(s) = X(s)\boldsymbol{\beta} + \epsilon(s)$$

Linear Covariate effect

White noise

Usually modeled with $N\left(0,\tau^2\right)$

Doesn't account for spatial effect.

How do we currently model this?

Account for spatial error: **Linear Mixed Model (LMM)**

$$Y(s) = X(s)\boldsymbol{\beta} + \epsilon(s) + W(s)$$

Linear Covariate effect

White noise

How do we currently model this?

Account for spatial error: **Linear Mixed Model (LMM)**

$$Y(s) = X(s)\boldsymbol{\beta} + \epsilon(s) + W(s)$$

Spatial random effect

Linear Covariate effect

White noise

How do we currently model this?

Account for spatial error: **Linear Mixed Model (LMM) with GP**

$$Y(s) = X(s)\boldsymbol{\beta} + \epsilon(s) + W(s)$$

Linear Covariate effect

White noise

Spatial random effect

Usually modeled with Gaussian Process (GP)

How do we estimate this?

How do we estimate this?

Maximum Likelihood Estimation

$$\text{Likelihood}\left(\mathbf{y}\right) \propto \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{ \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^{\top} \boldsymbol{\Sigma}^{-1} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) \right\}$$

How do we estimate this?

Maximum Likelihood Estimation

$$\text{Likelihood}\left(\mathbf{y}\right) \propto \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^{\top} \boldsymbol{\Sigma}^{-1} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)\right\}$$

How do we estimate this?

Maximum Likelihood Estimation

$$\text{Likelihood}(\mathbf{y}) \propto |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\top} \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\}$$

$\mathbf{C} + \tau^2\mathbf{I}$

Spatial error

White noise

What can go wrong?

Maximum Likelihood Estimation

$$\text{Likelihood}\left(\mathbf{y}\right) \propto \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{ \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^{\top} \boldsymbol{\Sigma}^{-1} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) \right\}$$

What can go wrong?

Maximum Likelihood Estimation

$$\text{Likelihood}\left(\mathbf{y}\right) \propto \left|\mathbf{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{ \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^{\top} \mathbf{\Sigma}^{-1} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) \right\}$$

$n \times n$

Dense

What can go wrong?

Maximum Likelihood Estimation

$$\text{Likelihood}\left(\mathbf{y}\right) \propto \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^{\top}\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\right\}$$

$n \times n$

Dense

$O\left(n^3\right)$. Infeasible in large data!!!

How do we propose to solve this?

Maximum Likelihood Estimation

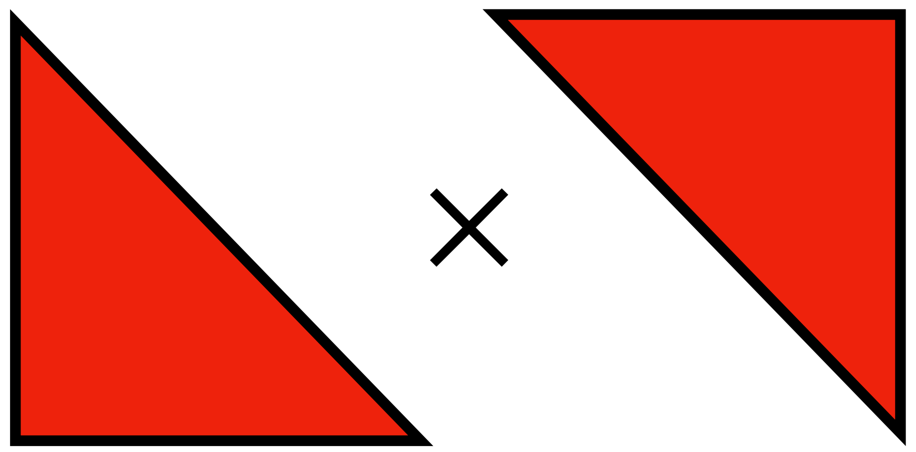
$$\text{Likelihood}\left(\mathbf{y}\right) \propto \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^{\top}\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\right\}$$

How do we propose to solve this?

Maximum Likelihood Estimation

$$\text{Likelihood}\left(\mathbf{y}\right) \propto \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{ \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^{\top} \boldsymbol{\Sigma}^{-1} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) \right\}$$

$$\boldsymbol{\Sigma}^{-1} \overset{Cholesky}{=}$$

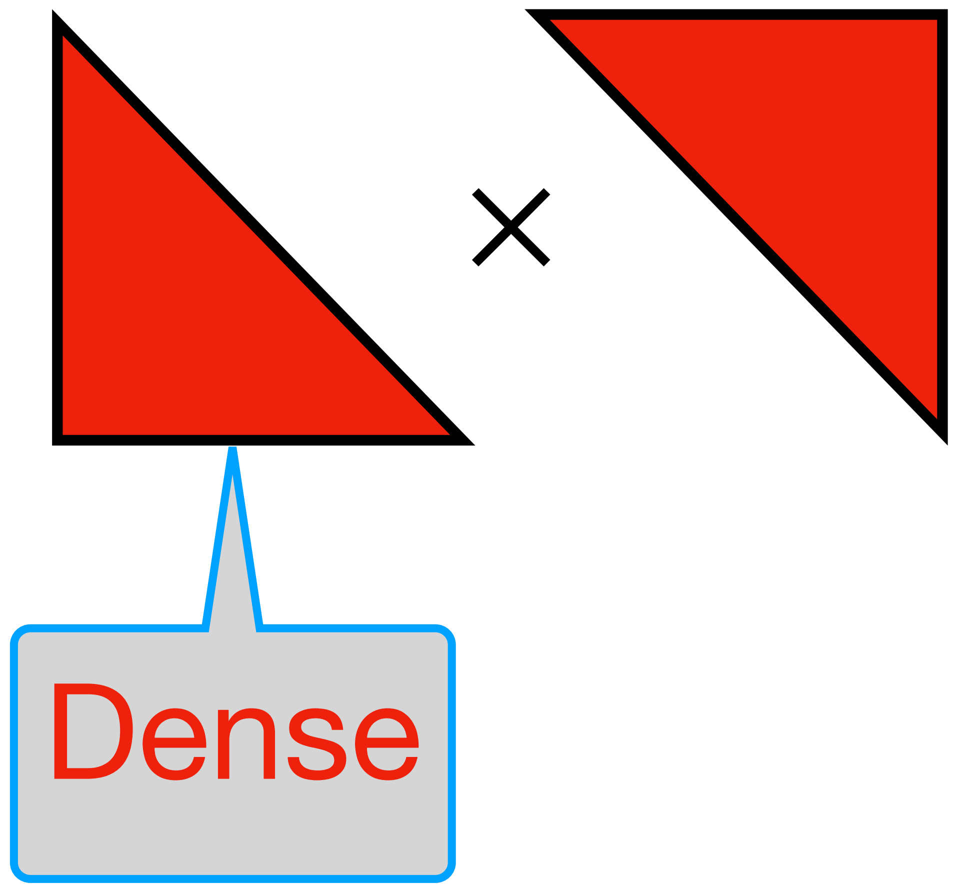


How do we propose to solve this?

Maximum Likelihood Estimation

$$\text{Likelihood}\left(\mathbf{y}\right) \propto \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{ \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^{\top} \boldsymbol{\Sigma}^{-1} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) \right\}$$

$$\boldsymbol{\Sigma}^{-1} \overset{Cholesky}{=}$$



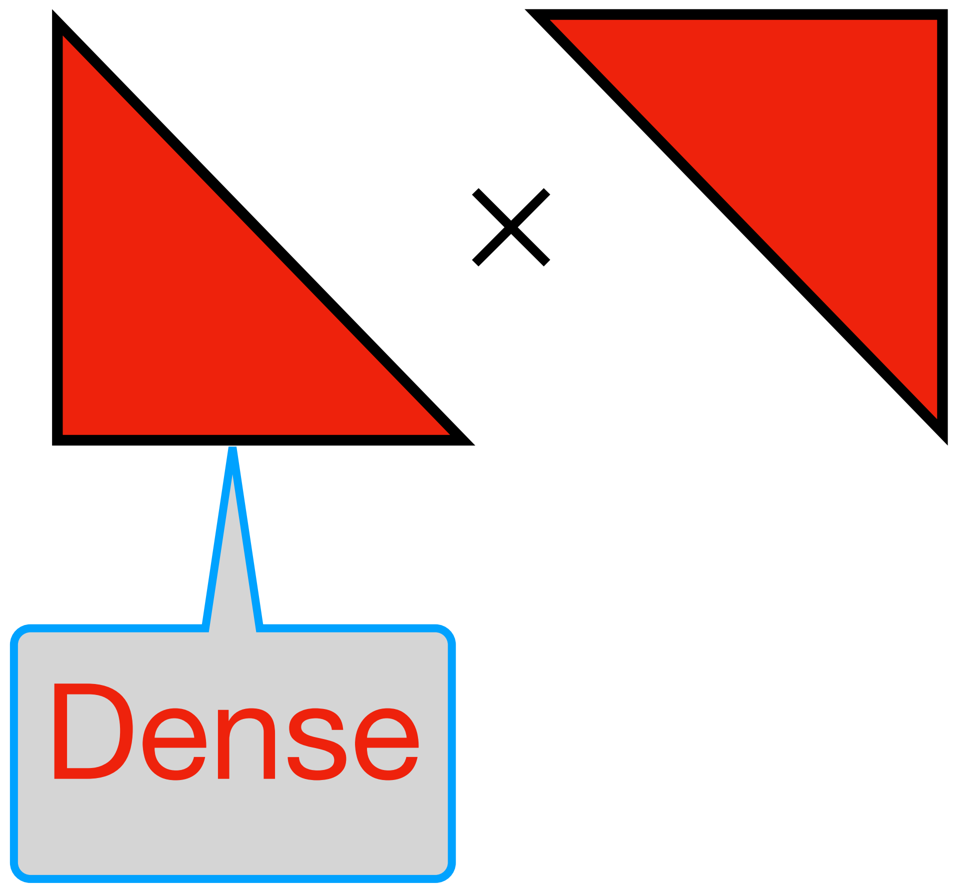
How do we propose to solve this?

Maximum Likelihood Estimation

$$\text{Likelihood}\left(\mathbf{y}\right) \propto \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{ \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^{\top} \boldsymbol{\Sigma}^{-1} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) \right\}$$

$$\boldsymbol{\Sigma}^{-1} \overset{\text{Cholesky}}{=}$$

$O\left(n^3\right)$

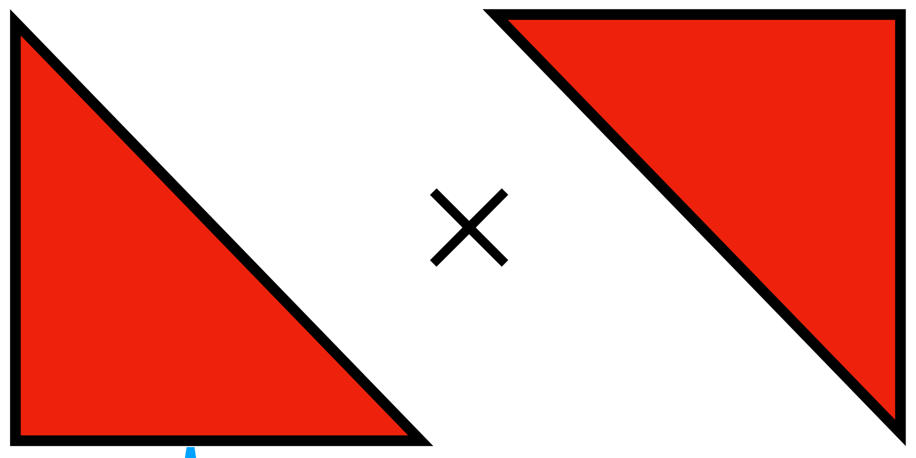


How do we propose to solve this?

Maximum Likelihood Estimation

$$\text{Likelihood}\left(\mathbf{y}\right) \propto \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{ \left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^{\top} \boldsymbol{\Sigma}^{-1} \left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right) \right\}$$

$$\boldsymbol{\Sigma}^{-1} \overset{Cholesky}{=}$$



$O\left(n^3\right)$

Dense

"everything is related to everything else, but near things are more related than distant things."

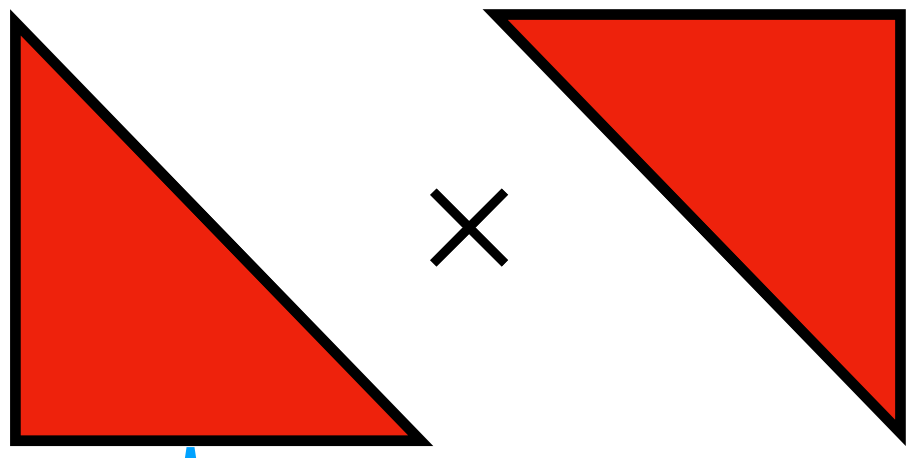
For any location, only consider its correlation with its m nearest neighbors!!

How do we propose to solve this?

Maximum Likelihood Estimation

$$\text{Likelihood}\left(\mathbf{y}\right) \propto \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{ \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^{\top} \boldsymbol{\Sigma}^{-1} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) \right\}$$

$$\boldsymbol{\Sigma}^{-1} \overset{Cholesky}{=}$$



$O\left(n^3\right)$

Dense

"everything is related to everything else, but near things are more related than distant things."

For any location, only consider its correlation with its m nearest neighbors!!

NN + GP

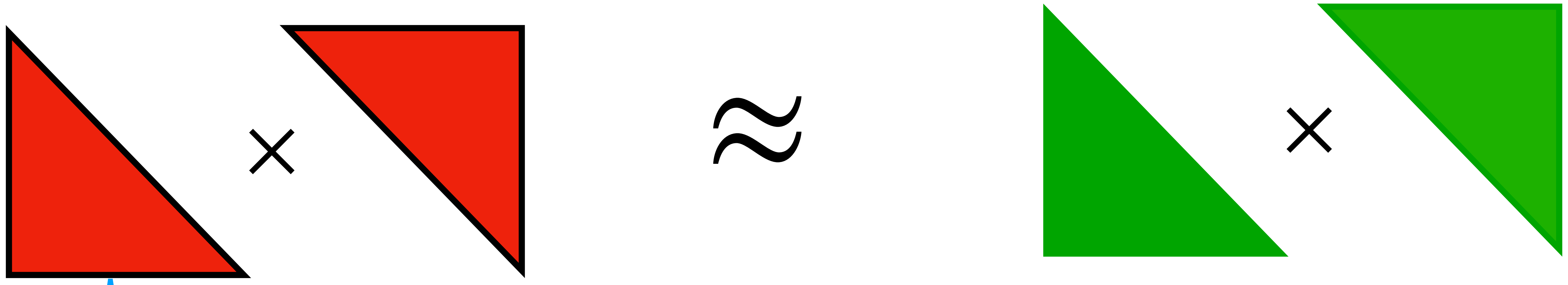
Datta A et al. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. JASA. 2016.

How do we propose to solve this?

Maximum Likelihood Estimation

$$\text{Likelihood}\left(\mathbf{y}\right) \propto \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{ \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^{\top} \boldsymbol{\Sigma}^{-1} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) \right\}$$

$$\boldsymbol{\Sigma}^{-1} \overset{Cholesky}{=}$$



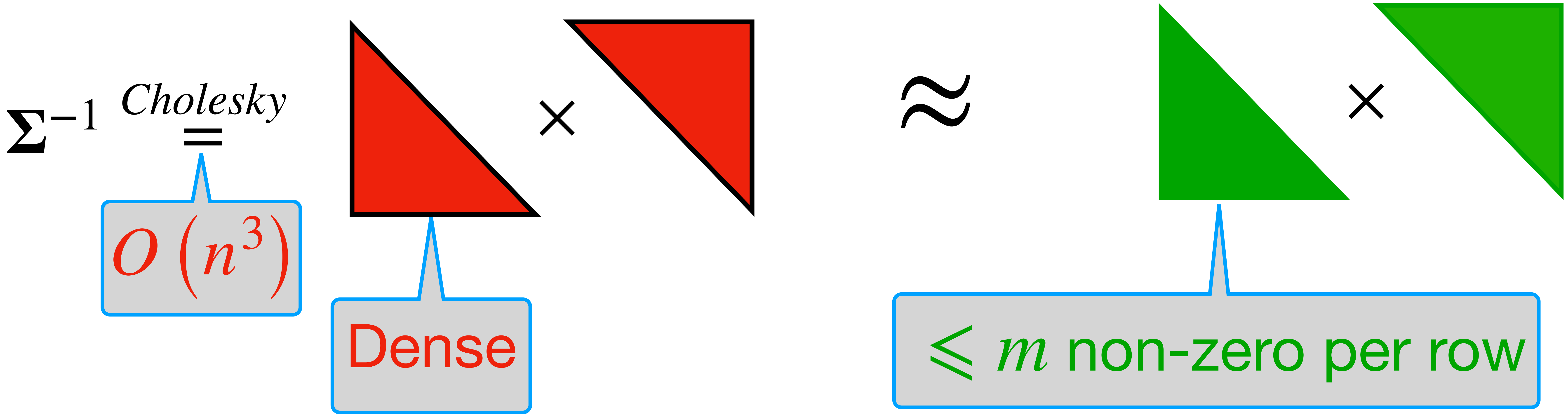
$O\left(n^3\right)$

Dense

How do we propose to solve this?

Maximum Likelihood Estimation

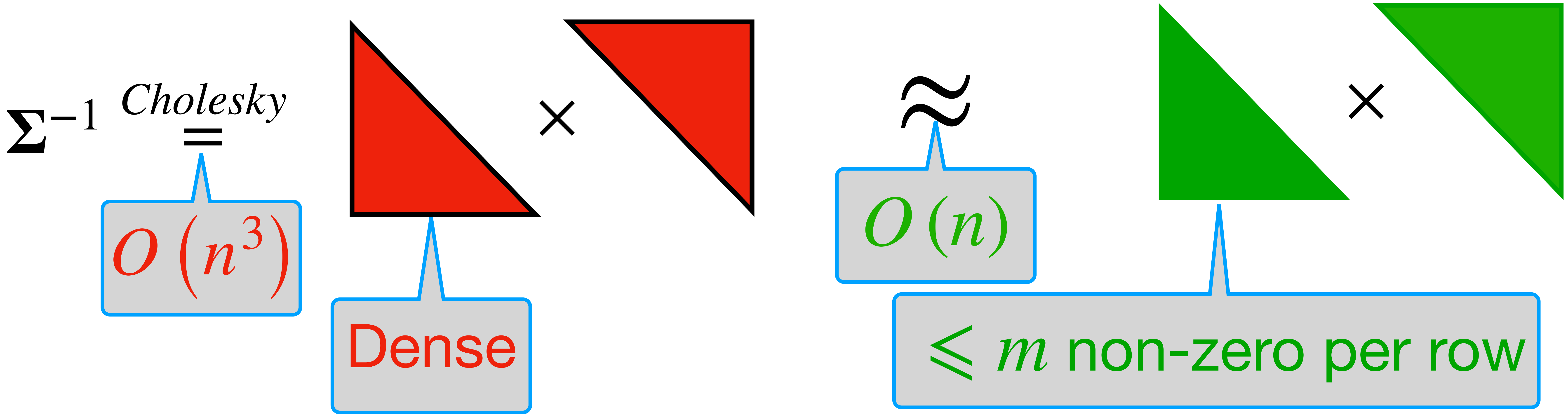
$$\text{Likelihood}\left(\mathbf{y}\right) \propto \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^{\top}\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\right\}$$



How do we propose to solve this?

Maximum Likelihood Estimation

$$\text{Likelihood}\left(\mathbf{y}\right) \propto \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left\{ \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^{\top} \boldsymbol{\Sigma}^{-1} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) \right\}$$



BRISC

- BRSIC implements this in R, a wrapper around C++ code.
- Embarrassingly parallel computation!

Package ‘BRISC’

October 12, 2022

Type Package

Title Fast Inference for Large Spatial Datasets using BRISC

Version 1.0.5

Maintainer Arkajyoti Saha <arkajyotisaha93@gmail.com>

Author Arkajyoti Saha [aut, cre],
Abhirup Datta [aut],
Jorge Nocedal [ctb],
Naoaki Okazaki [ctb],
Lukas M. Weber [ctb]

Depends R (>= 3.3.0), RANN, parallel, stats, rdist, matrixStats, pbapply, graphics

Description Fits bootstrap with univariate spatial regression models using Bootstrap for Rapid Inference on Spatial Covariances (BRISC) for large datasets using nearest neighbor Gaussian processes detailed in Saha and Datta (2018) <doi:10.1002/sta4.184>.

Outline

- What problem does BRISC solve?
- **What can you do with BRISC?**
- Applications of BRISC.

Inference

Significant improvement over state-of-the-art algorithms.

Training data ~ 105K

Test data ~ 45K

Inference

Significant improvement over state-of-the-art algorithms.

Training data ~ 105K

Test data ~ 45K

Classical methods do not work!!

Inference

Significant improvement over state-of-the-art algorithms.

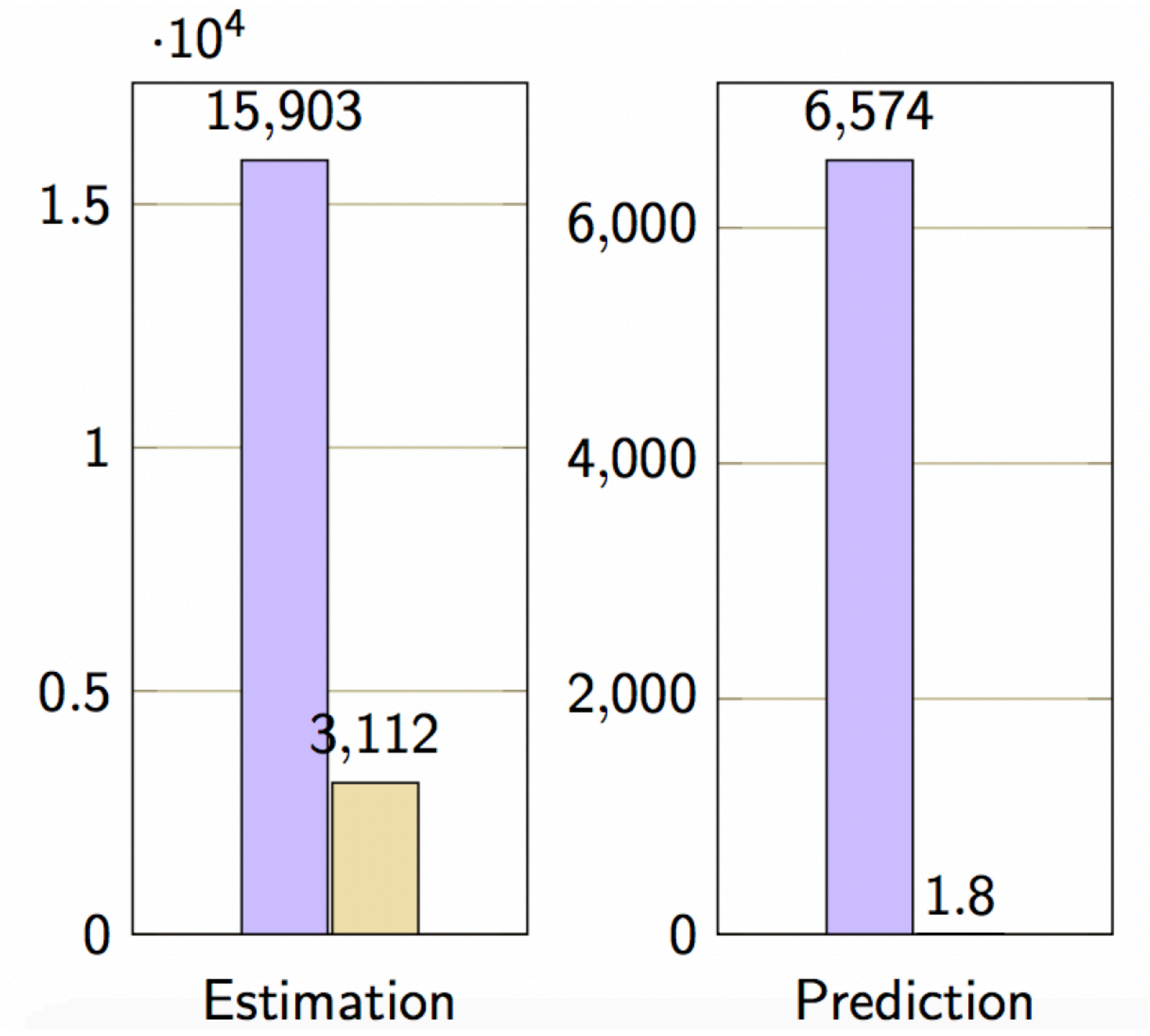
Training data ~ 105K

Test data ~ 45K

Classical methods do not work!!

NNGP with Bayesian (Datta et al.)

NNGP with BRISC



Inference

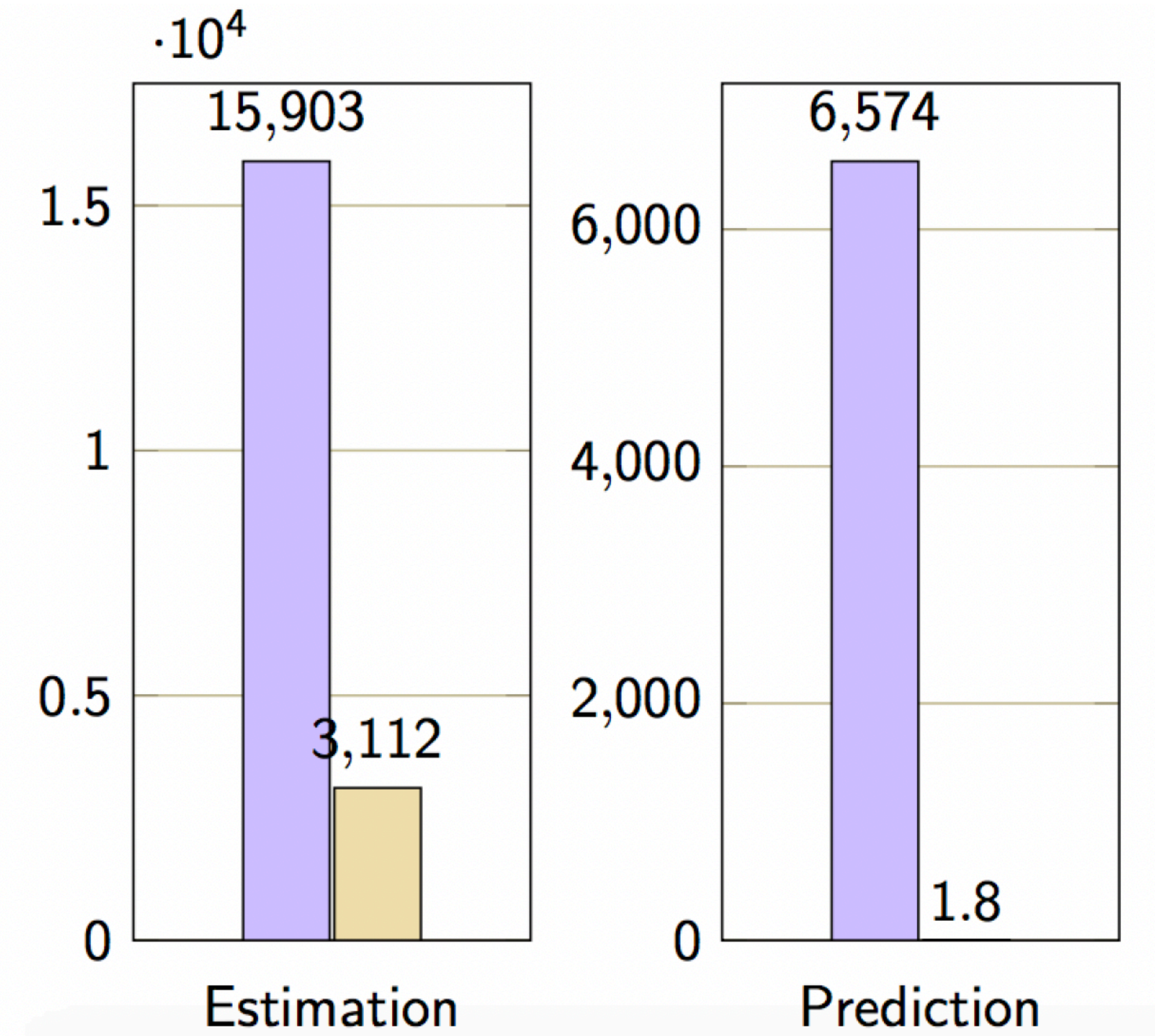
Significant improvement over state-of-the-art algorithms.

Training data ~ 105K

Test data ~ 45K

Classical methods do not work!!

- NNGP with Bayesian (Datta et al.)
- NNGP with BRISC



Saha A, Datta A. BRISC: Bootstrap for rapid inference on spatial covariances. Stat. 2018;7(1):e184.

Inference

- Estimation: `estimation_result <- BRISC_estimation(coords, y, x)`

Inference

- Estimation: `estimation_result <- BRISC_estimation(coords, y, x)`
- Uncertainty (via bootstrap): `BRISC_bootstrap(estimation_result)`

Inference

- Estimation: `estimation_result <- BRISC_estimation(coords, y, x)`
- Uncertainty (via bootstrap): `BRISC_bootstrap(estimation_result)`
- Prediction: `BRISC_prediction(estimation_result, coords_pred, x_pred)`

Inference

- Estimation: `estimation_result <- BRISC_estimation(coords, y, x)`
- Uncertainty (via bootstrap): `BRISC_bootstrap(estimation_result)`
- Prediction: `BRISC_prediction(estimation_result, coords_pred, x_pred)`

Prediction Location

Prediction X

Simulating from Gaussian Process

Simulating LARGE data from Gaussian Process is computationally challenging.

Simulating from Gaussian Process

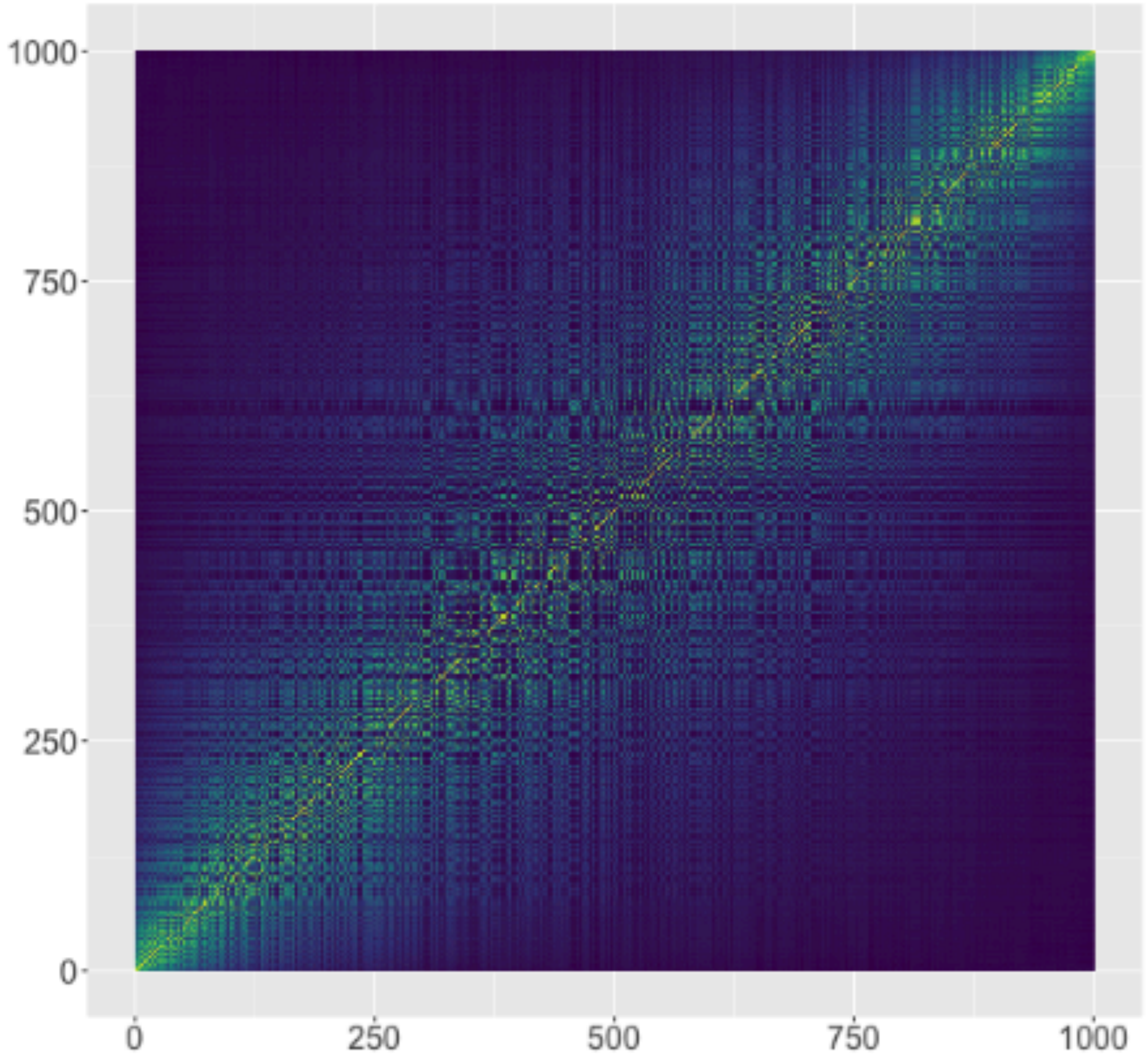
Simulating LARGE data from Gaussian Process is computationally challenging.

Simulate from NNGP with BRISC: `BRISC_simulation(coords)`

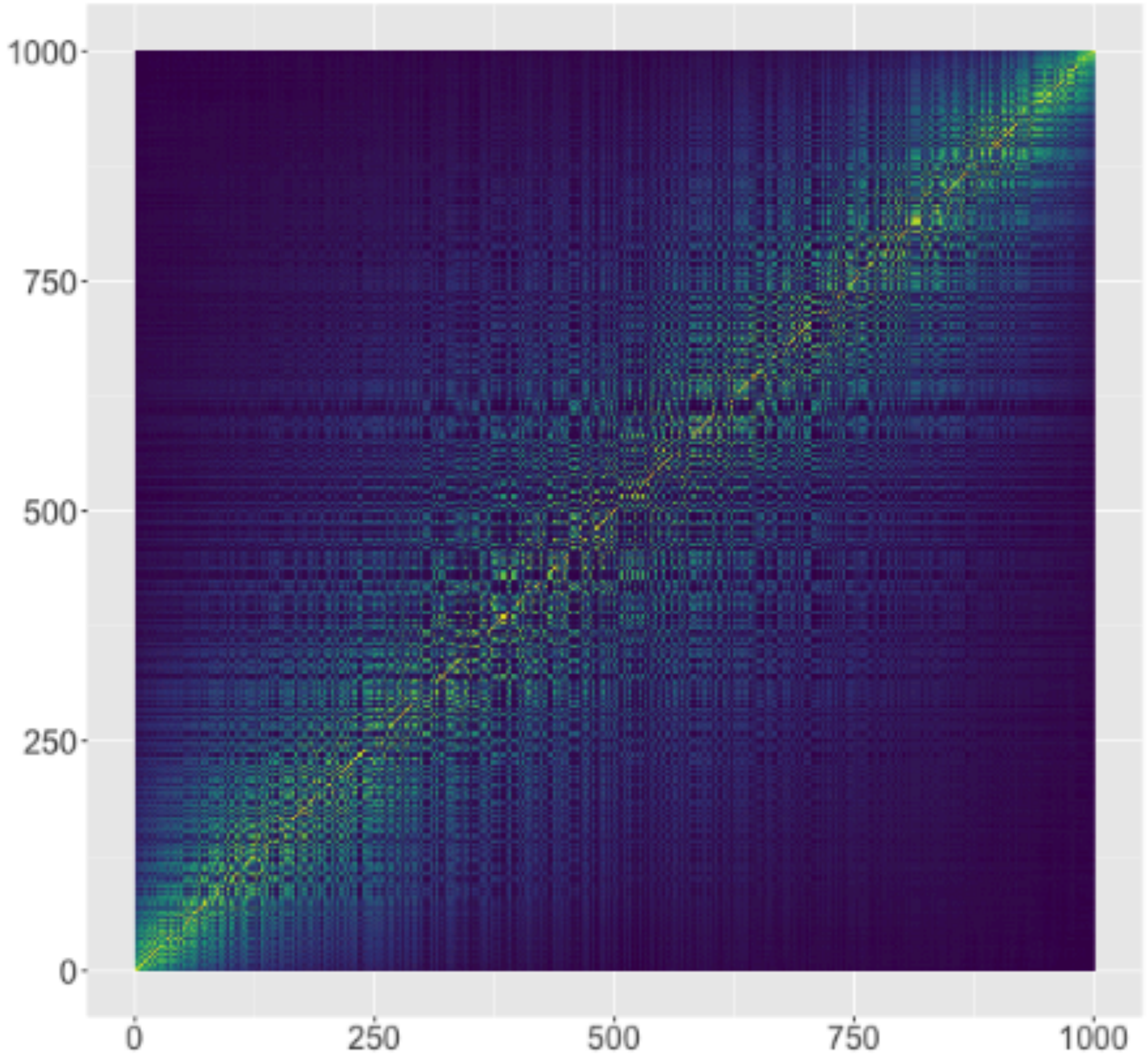
Simulating from Gaussian Process

Simulating LARGE data from Gaussian Process is computationally challenging.

Simulate from NNGP with BRISC: `BRISC_simulation(coords)`



Full GP

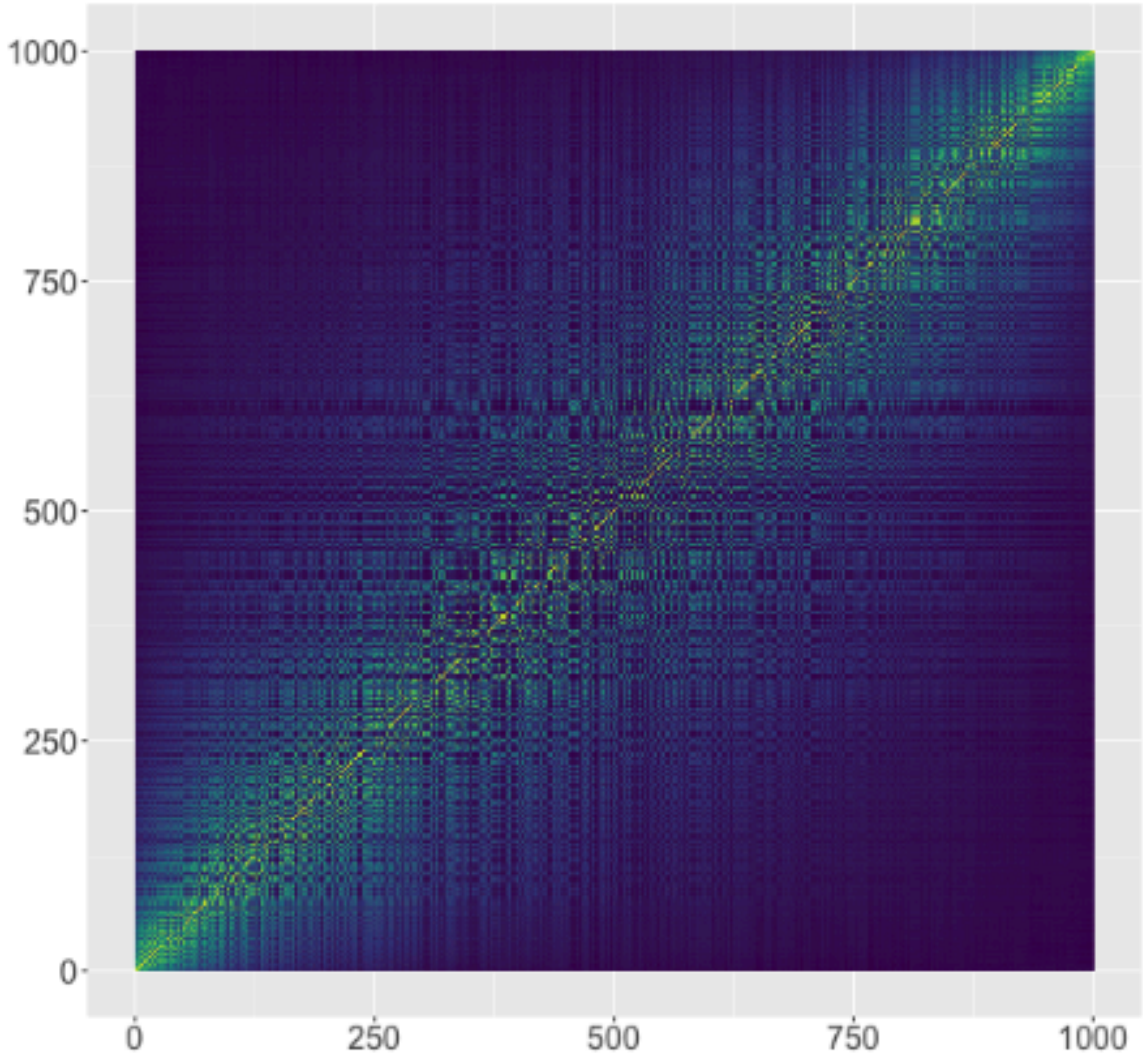


NNGP

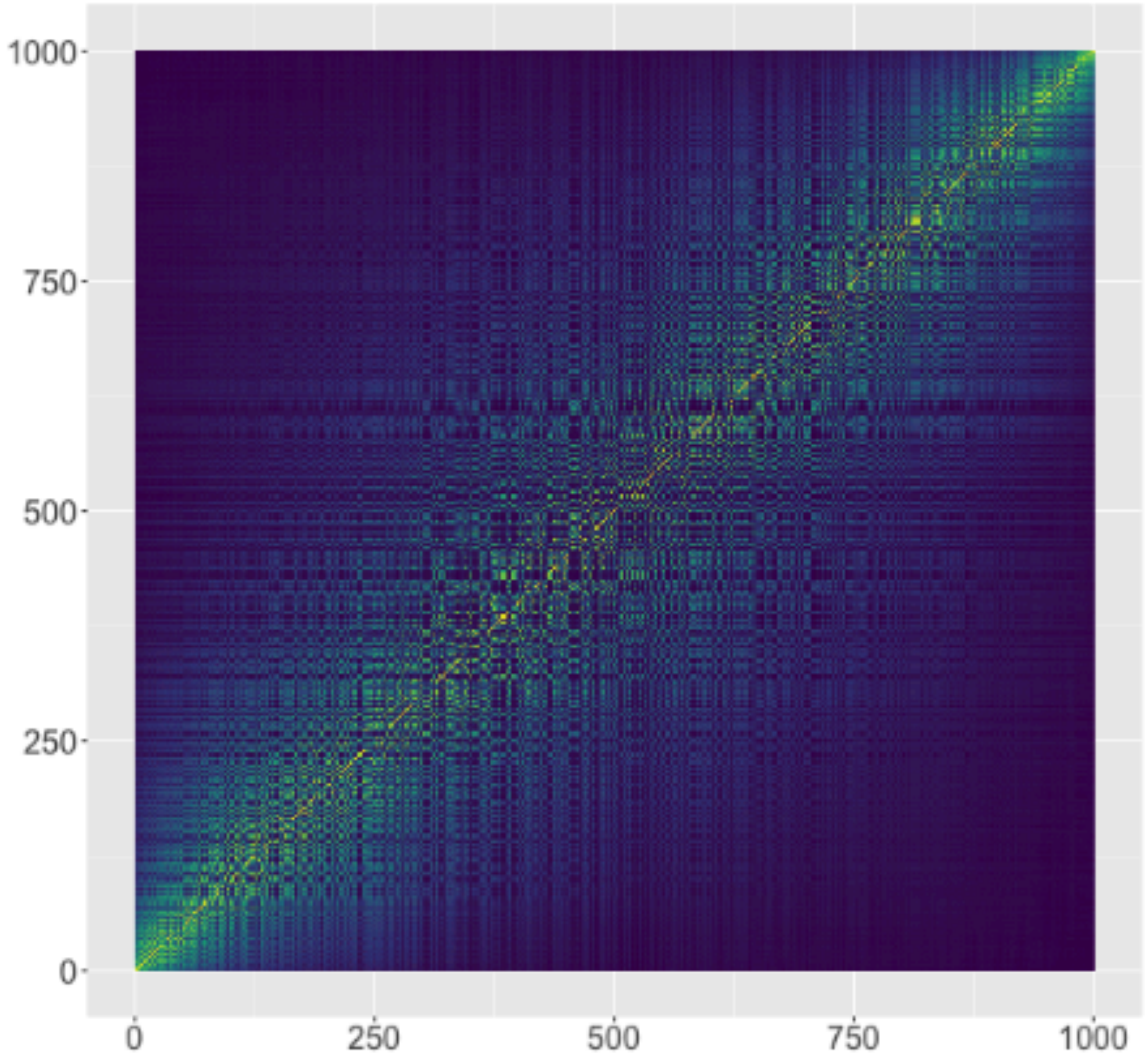
Simulating from Gaussian Process

Simulating LARGE data from Gaussian Process is computationally challenging.

Simulate from NNGP with BRISC: `BRISC_simulation(coords)`



Full GP



NNGP

Sample size	NNGP	full GP
1000	0.7 (0.04)	2.6 (0.08)
2500	1.6 (0.29)	31.8 (2.02)
5000	3.3 (0.25)	262.3 (9.33)
10000	8.3 (0.23)	NA
100000	121.5 (9.53)	NA

Probit Model

Estimates parameters by maximizing likelihood with a grid search.

Probit Model

Estimates parameters by maximizing likelihood with a grid search.

Methods	$n = 15^2$	$n = 25^2$	$n = 50^2$	$n = 100^2$
(a) Grid search for **one** parameter combination				
probit-NNGP	0.065	0.5	9	166
TLR	0.57	2.9	28	187
(b) Prediction at **one** out-of-sample location following estimation				
probit-NNGP	< 0.01	< 0.01	< 0.01	0.025
TLR	1.2	5.8	40	271

TLR = Low rank approximation of covariance matrix.

Probit Model

Estimates parameters by maximizing likelihood with a grid search.

Methods	$n = 15^2$	$n = 25^2$	$n = 50^2$	$n = 100^2$
(a) Grid search for **one** parameter combination				
probit-NNGP	0.065	0.5	9	166
TLR	0.57	2.9	28	187
(b) Prediction at **one** out-of-sample location following estimation				
probit-NNGP	< 0.01	< 0.01	< 0.01	0.025
TLR	1.2	5.8	40	271

TLR = Low rank approximation of covariance matrix.

Saha A et al. Scalable Predictions for Spatial Probit Linear Mixed Models Using Nearest Neighbor Gaussian Processes. Journal of Data Science. 2022.

Probit Model

Estimates parameters by maximizing likelihood with a grid search.

Likelihood evaluation for estimation: `llk_binary <- Binary_estimation(coords, y)`

Probit Model

Estimates parameters by maximizing likelihood with a grid search.

Likelihood evaluation for estimation: `llk_binary <- Binary_estimation(coords, y)`

Prediction: `Binary_prediction(llk_binary, coords_pred)`

Probit Model

Estimates parameters by maximizing likelihood with a grid search.

Likelihood evaluation for estimation: `llk_binary <- Binary_estimation(coords, y)`

Prediction: `Binary_prediction(llk_binary, coords_pred)`

Tutorial: https://github.com/ArkajyotiSaha/probit-NNGP-code

Outline

- What problem does BRISC solve?
- What can you do with BRISC?
- **Applications of BRISC.**

Applications of BRISC

- nn-SVG : Identifies of spatially variable genes transcriptomics data with BRISC.

Applications of BRISC

- nn-SVG : Identifies of spatially variable genes transcriptomics data with BRISC.
- RF-GLS : $Y(s) = X(s)\boldsymbol{\beta} + \epsilon(s) + W(s)$

Applications of BRISC

- nn-SVG : Identifies of spatially variable genes transcriptomics data with BRISC.

- RF-GLS : $Y(s) = \underbrace{\cancel{X(s)\boldsymbol{\beta}}}_{f(X)} + \epsilon(s) + W(s)$

Applications of BRISC

- nn-SVG : Identifies of spatially variable genes transcriptomics data with BRISC.
- RF-GLS : $Y(s) = \underbrace{\cancel{X(s)\boldsymbol{\beta}}}_{f(X)} + \epsilon(s) + W(s)$

  Fits Random Forest in spatially dependent data.

Applications of BRISC

- nn-SVG : Identifies of spatially variable genes transcriptomics data with BRISC.

- RF-GLS : $Y(s) = \underbrace{\cancel{X(s)\boldsymbol{\beta}}}_{f(X)} + \epsilon(s) + W(s)$

  Fits Random Forest in spatially dependent data.

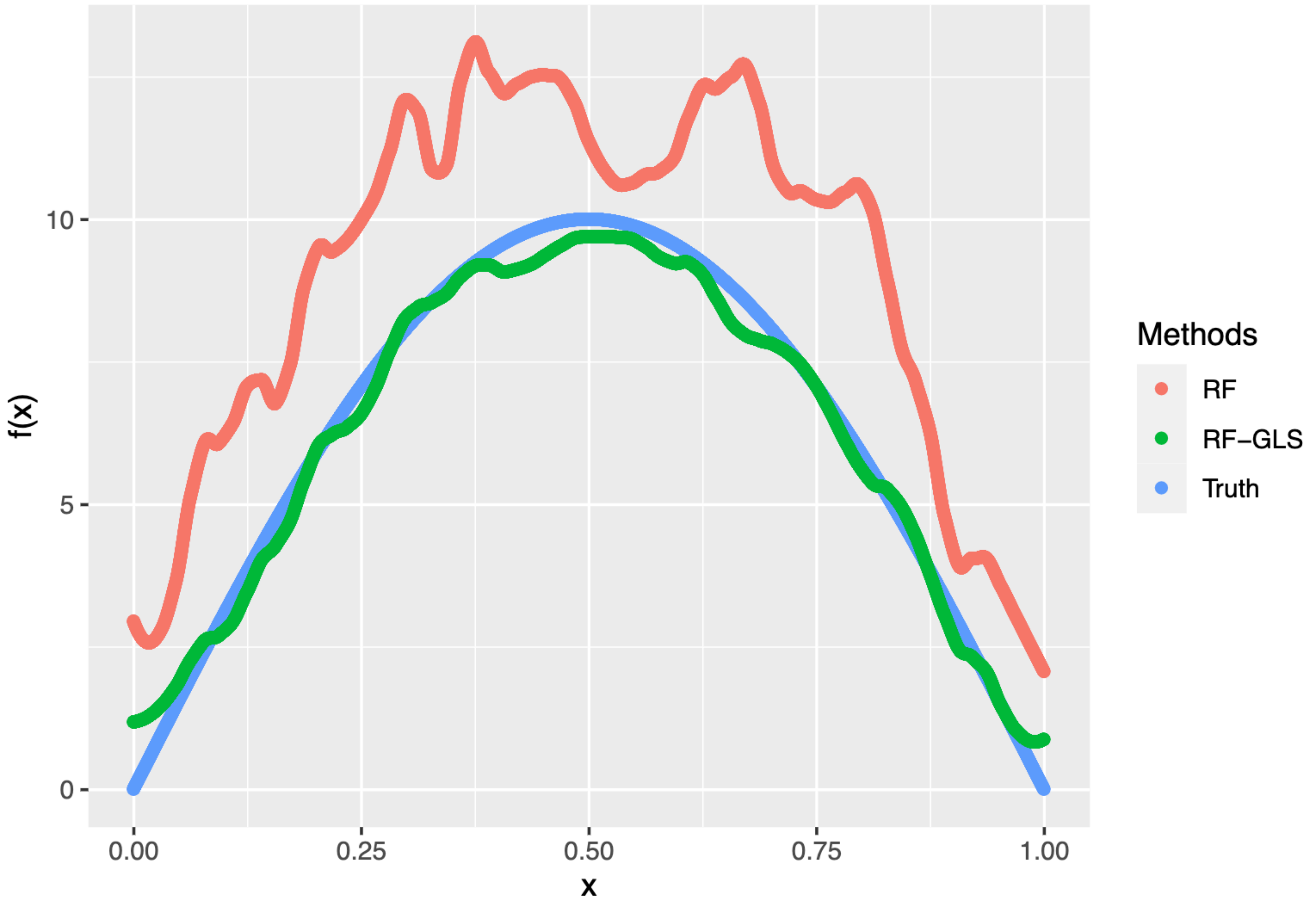
  Rewrite the split criteria as a Generalized Least Square (GLS) Loss.

Applications of BRISC

- nn-SVG : Identifies of spatially variable genes transcriptomics data with BRISC.
- RF-GLS : $Y(s) = \underset{f(X)}{\cancel{X(s)\boldsymbol{\beta}}} + \epsilon(s) + W(s)$

  Fits Random Forest in spatially dependent data.

  Rewrite the split criteria as a Generalized Least Square (GLS) Loss.



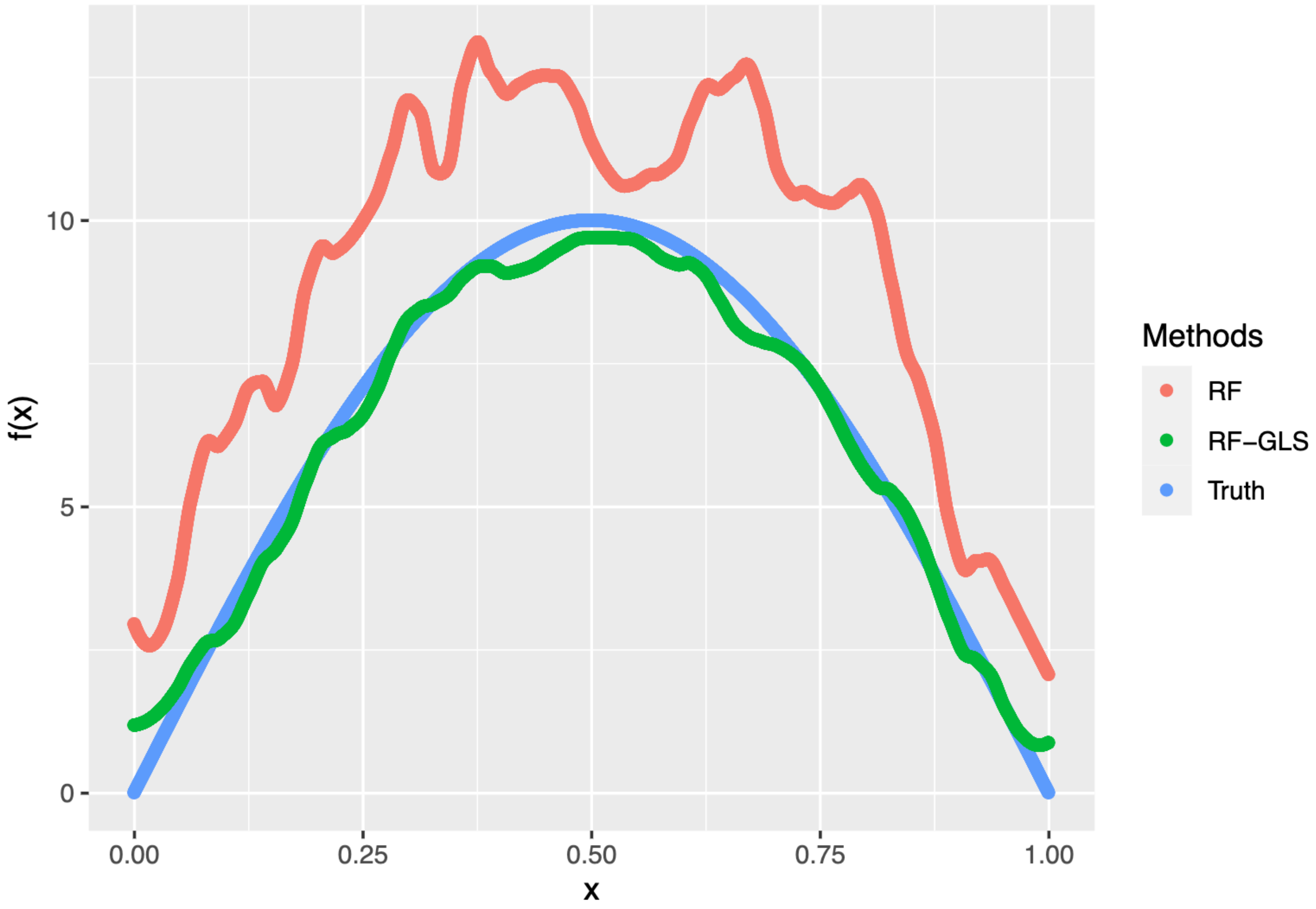
Applications of BRISC

- nn-SVG : Identifies of spatially variable genes transcriptomics data with BRISC.

- RF-GLS : $Y(s) = \underbrace{\cancel{X(s)\beta}}_{f(X)} + \epsilon(s) + W(s)$

  Fits Random Forest in spatially dependent data.

  Rewrite the split criteria as a Generalized Least Square (GLS) Loss.



<u>Continuous Outcome:</u> Saha A et al. Random forests for spatially dependent data. JASA. 2023.

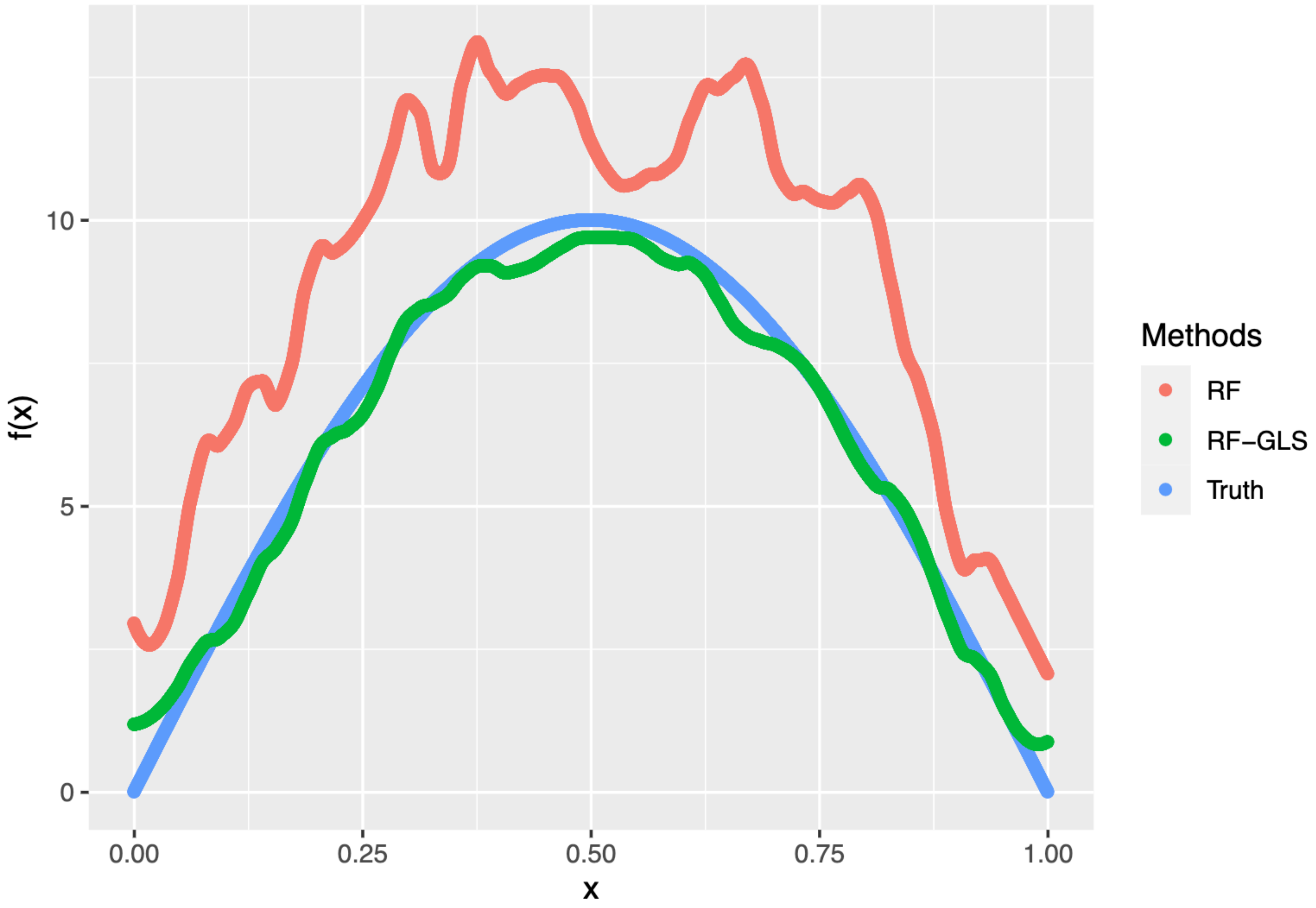
Applications of BRISC

- nn-SVG : Identifies of spatially variable genes transcriptomics data with BRISC.

- RF-GLS : $Y(s) = \underbrace{\cancel{X(s)\beta}}_{f(X)} + \epsilon(s) + W(s)$

  Fits Random Forest in spatially dependent data.

  Rewrite the split criteria as a Generalized Least Square (GLS) Loss.



<u>Continuous Outcome:</u> Saha A et al. Random forests for spatially dependent data. JASA. 2023.

<u>Binary Outcome:</u> Saha A, Datta A. Random forests for binary geospatial data. arXiv preprint arXiv:2302.13828. 2023

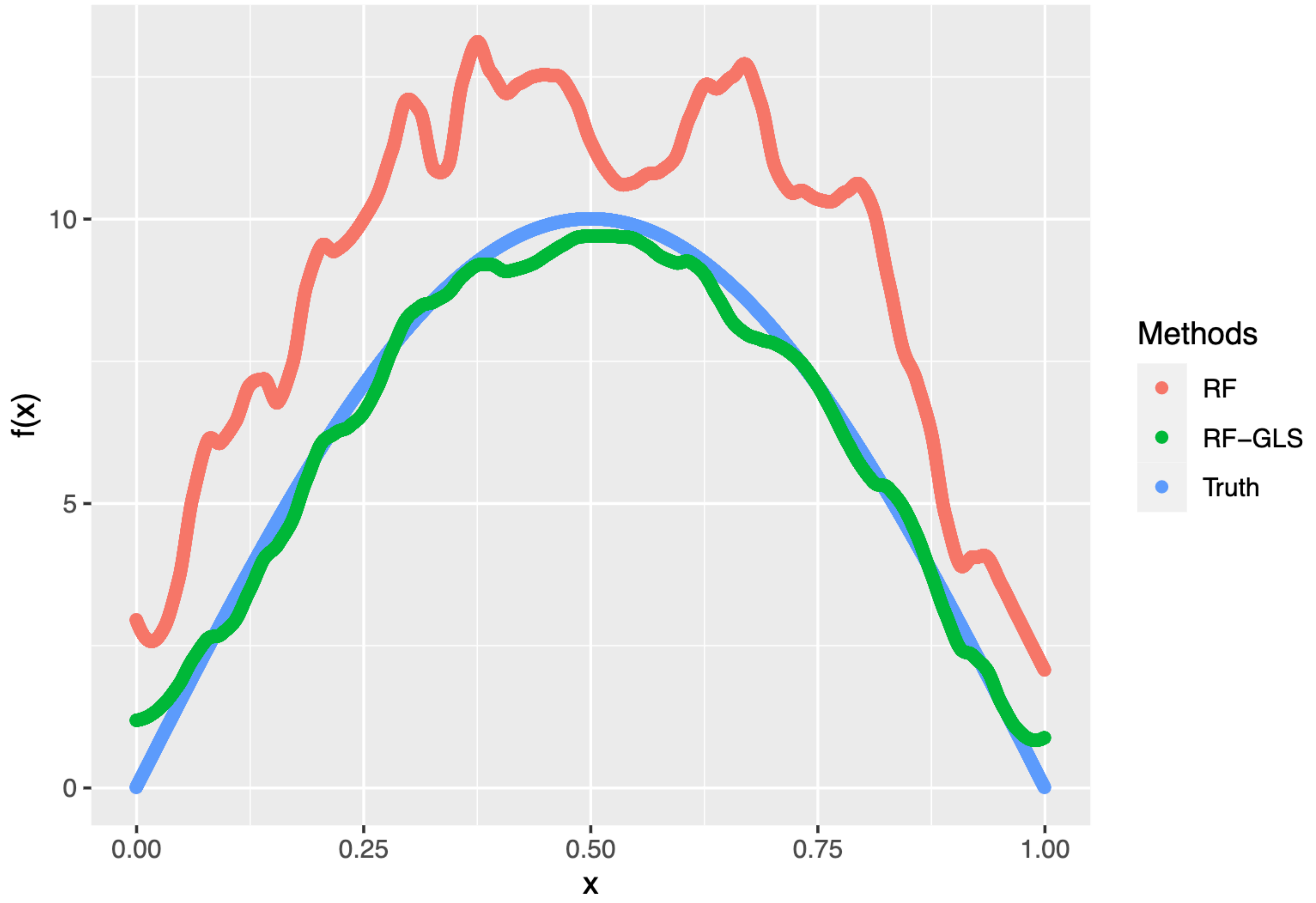
Applications of BRISC

- nn-SVG : Identifies of spatially variable genes transcriptomics data with BRISC.

- RF-GLS : $Y(s) = \underset{f(X)}{\cancel{X(s)\beta}} + \epsilon(s) + W(s)$

  Fits Random Forest in spatially dependent data.

  Rewrite the split criteria as a Generalized Least Square (GLS) Loss.



<u>Continuous Outcome:</u> Saha A et al. Random forests for spatially dependent data. JASA. 2023.

<u>Binary Outcome:</u> Saha A, Datta A. Random forests for binary geospatial data. arXiv preprint arXiv:2302.13828. 2023

<u>Package:</u> Saha A et al. RandomForestsGLS: An R package for Random Forests for dependent data. Journal of Open Source Software. 2022

Acknowledgements

Abhirup Datta
Biostatistics, BSPH, JHU

Sumanta Basu
Statistics & Data Science, Cornell