BART as a Gaussian process

FAST ACCURATE LEGIBLE

Giacomo Petrillo <giacomo.petrillo@unifi.it>
University of Florence (UNIFI)
Department of Statistics, Computer Science, Applications (DISIA)

Summary

BART is a state-of-the-art Bayesian nonparametric regression method. In causal inference, you use it to impute the missing potential outcomes. It computes the posterior running a MCMC over an ensemble of trees. This work develops a completely different implementation using Gaussian processes.

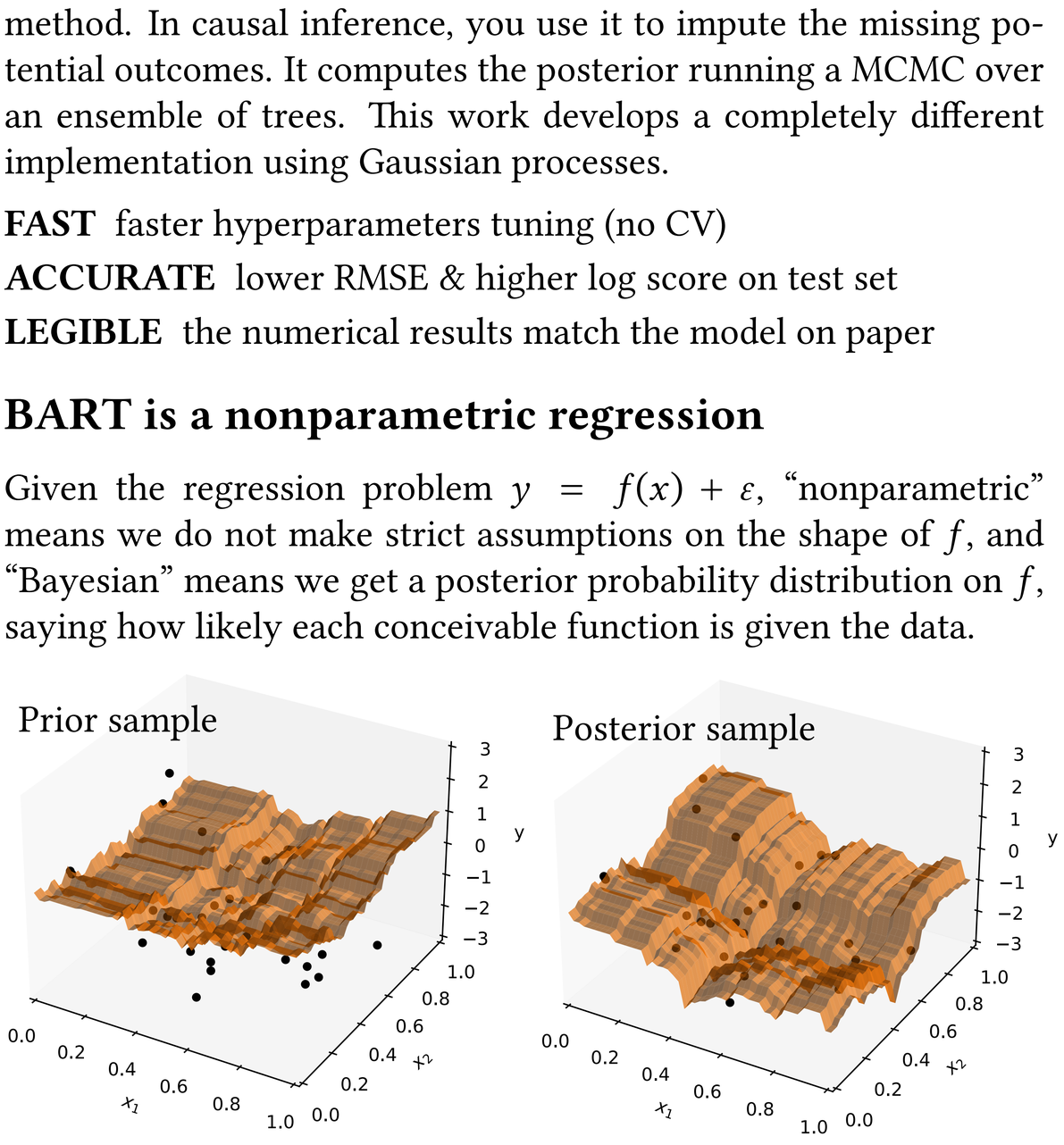
FAST faster hyperparameters tuning (no CV)

ACCURATE lower RMSE & higher log score on test set

LEGIBLE the numerical results match the model on paper

BART is a nonparametric regression

Given the regression problem $y = f(x) + \varepsilon$, "nonparametric" means we do not make strict assumptions on the shape of f, and "Bayesian" means we get a posterior probability distribution on f, saying how likely each conceivable function is given the data.



BART with ∞ trees is a Gaussian process

BART represents f as a sum of many regression trees: $f(x) = \sum_{i=1}^{m} T_i(x)$. The prior distribution is specified over the tree properties (depth, divisions, children).

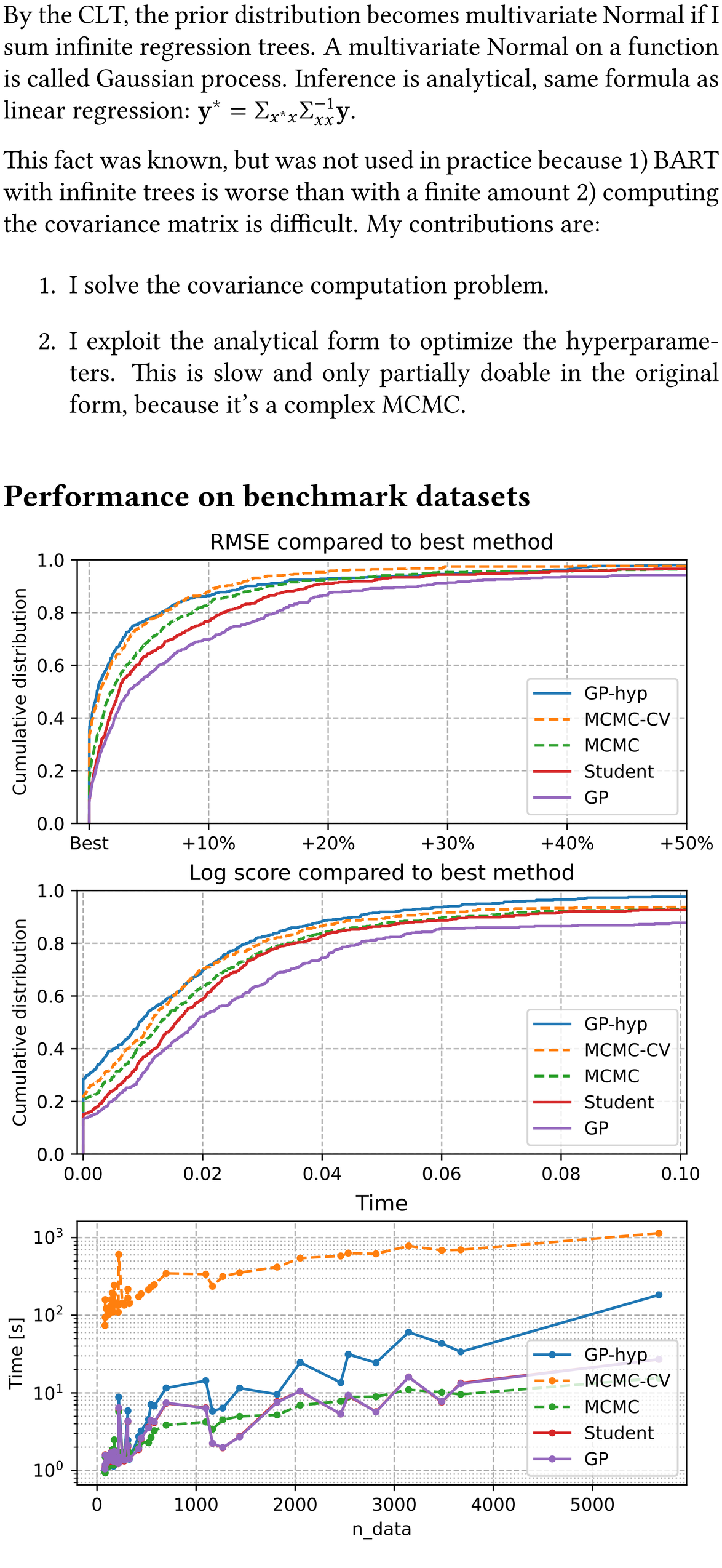
By the CLT, the prior distribution becomes multivariate Normal if I sum infinite regression trees. A multivariate Normal on a function is called Gaussian process. Inference is analytical, same formula as linear regression: $\mathbf{y}^* = \Sigma_{x^*x}\Sigma_{xx}^{-1}\mathbf{y}$.

This fact was known, but was not used in practice because 1) BART with infinite trees is worse than with a finite amount 2) computing the covariance matrix is difficult. My contributions are:

1. I solve the covariance computation problem.

2. I exploit the analytical form to optimize the hyperparameters. This is slow and only partially doable in the original form, because it's a complex MCMC.

Performance on benchmark datasets



BART is SoTA in causal inference

Results of the ACIC 2022 data challenge: