Abstract

The tgp package is a tool for fully Bayesian semiparametric and nonstationary regression by treed Gaussian processes with jumps to the limiting linear model. Special cases also implemented include the Bayesian linear model, linear CART, and the stationary separable or isotropic Gaussian process (GP).

In addition to inference and posterior prediction, the package supports the (sequential) design of experiments under these models paired with several objective criteria.

1D and 2D plotting, with higher dimension projection and slice capabilities, and tree drawing functions are provided.

3. Nonstationary Predictive Uncertainty

- Predictive uncertainty of stationary GPs is always largest where sampling is lowest. Partitioning (e.g., treed GP) allows for more reasonable estimates.

![plot](sin.btlm, layout="surf")

- A key feature of the treed GP is its ability to learn about input-dependent noise. A classic example is the Motorcycle data, in R.

library(MASS); X <- scyclists[1]; Z <- scyclists[2]

3. Nonstationary Predictive Uncertainty

- Predictive uncertainty of stationary GPs is always largest where sampling is lowest. Partitioning (e.g., treed GP) allows for more reasonable estimates.

4. Semiparametric Modeling

- The Friedman data set has 10 covariates, but the response $E(Z(x)) = \mu = 10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$ depends only on the first 5, combining nonlinear, linear, and irrelevant effects.

   - The following R code compares Bayesian linear CART to the (non-treed) GP LLM in terms of RMSE

   ```r
   f <- Friedman 1 data(200); X <- fr[1:10]; Z <- f
   ff <- Friedman 1 data(1000); X <- fr[1:10]
   fr.bglm <- bgp(X, Z); fr.bglm2 <- bgp(X, Z, variances=200)
   fr.bglm3 <- bgp(X, Z, s=c(1,1,1))
   exp.bglm <- bgp(X, Z, s=c(1,1,1))
   exp.bglm2 <- bgp(X, Z, s=(1,1))
   exp.bglm3 <- bgp(X, Z, s=(1,1,1))
   sqrt(mean((fr.btlm$Ytrue - fr.btlm$Ypred)^2)) #> 0.4157481
   sqrt(mean((fr.btlm$Ytrue - fr.btlm$Ypred)^2)) #> 2.191923
   ```

5. Sequential Design

Consider sequentially designing an experiment for the exp data (left), by subsampling from candidates:

- Start by finding the MAP tree $T$
- Then, obtain a sequential D-optimal design, i.e., for each GP in $T$.

![plot](exp.btlm, layout="as")

- Use the treed GP model again, and gather “adaptive sampling” statistics:
  - ALM: pred var $\sigma^2(k)$
  - ALC: expected reduction in pred var $\Delta \sigma^2(k) = \int \sigma^2(\tilde{x}) - \sigma^2(x) f(\tilde{x} | x) d\tilde{x}$
  - EGO: expected global optimization (min) $E[\max_{x \in \Omega} f(x)]$

6. Concluding Thoughts

- tgp is licensed under the LGPL, and available from CRAN
- Parallelization via Pthreads can be enabled at compile time
- For a tutorial, from R simply enter vignette("tgp")