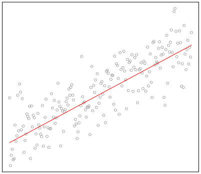
Distributional Trees and Forests

Lisa Schlosser, Torsten Hothorn, Achim Zeileis

`https://R-Forge.R-project.org/projects/partykit/`

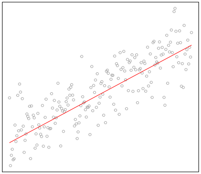
Motivation



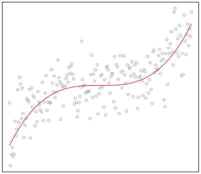
LM, GLM

`lm`
`glm`

Motivation



$\longrightarrow$



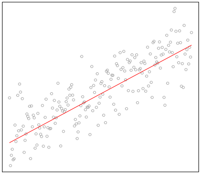
LM, GLM

GAM

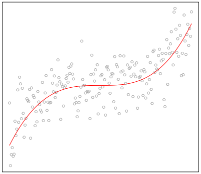
`lm`
`glm`

`mgcv`
`VGAM`
...

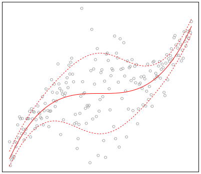
Motivation



→



→



LM, GLM

`lm`
`glm`

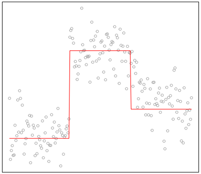
GAM

`mgcv`
`VGAM`
...

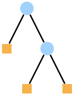
GAMLSS

`gamlss`
`mgcv`
`VGAM`
`gamboostLSS`
...

Motivation

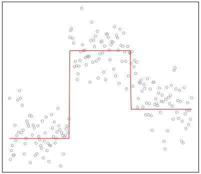


Regression Tree

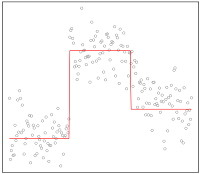


`rpart`
`party(kit)`

Motivation

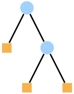


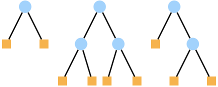
$\longrightarrow$



Regression Tree

Random Forest

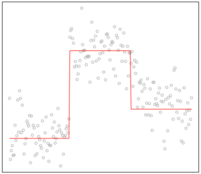




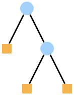
`rpart`
`party(kit)`

`randomForest`
`ranger`
`party(kit)`
...

Motivation

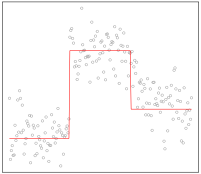


Regression Tree

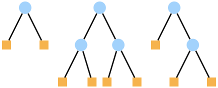


`rpart`
`party(kit)`

→

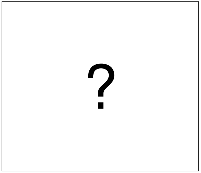


Random Forest



`randomForest`
`ranger`
`party(kit)`
...

→



Distributional Trees and Forests

`disttree`
based on
`partykit`

Goals

Tree:

- Specify the complete distribution in each subgroup.
  (location, scale and shape)
- Automatic detection of steps and abrupt changes.
- Capture non-linear and non-additive effects and interactions.

Forest:

- Smoother effects.
- Stabelization of the model.

Building Distributional Trees and Forests

Tree:

1. Specify a distribution with log-likelihood function $\ell(\theta; y)$.
2. Estimate $\hat{\theta}$ via maximum likelihood.
3. Test for associations or instabilities of the scores $\frac{\partial \ell}{\partial \theta}(\hat{\theta}; y_i)$ and each partitioning variable x_i.
4. Split the sample along the partitioning variable with the strongest association or instability. Choose breakpoint with highest improvement in log-likelihood.
5. Repeat steps 2–4 recursively in the subgroups until some stopping criterion is met.

Forest: Ensemble of trees.

- Bootstrap or subsamples.
- Random input variable sampling.

Prediction

Tree:

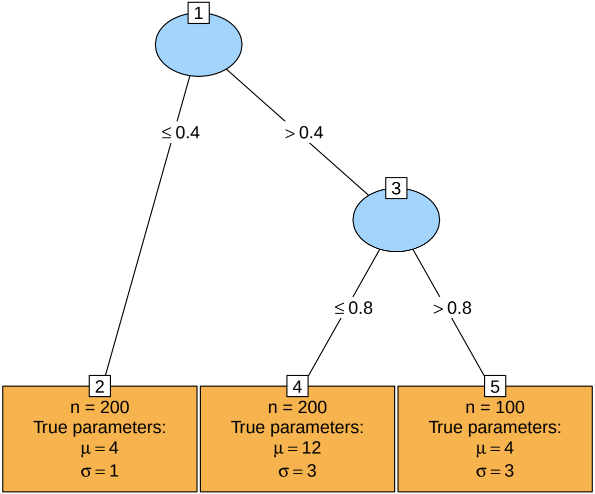
Estimate $\hat{\theta}$ on the subsample of the learning data which ends up in the same terminal node as the new observation.

Forest:

Estimate $\hat{\theta}$ on the whole learning data but weighted by the number of trees in which a learning observation ends up in the same terminal node as the new observation.

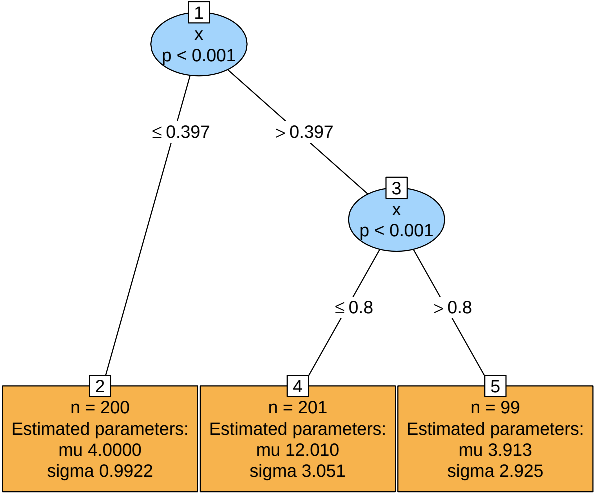
Fitting a Tree

DGP: $Y \sim \mathcal{N}(\mu(X), \sigma(X))$



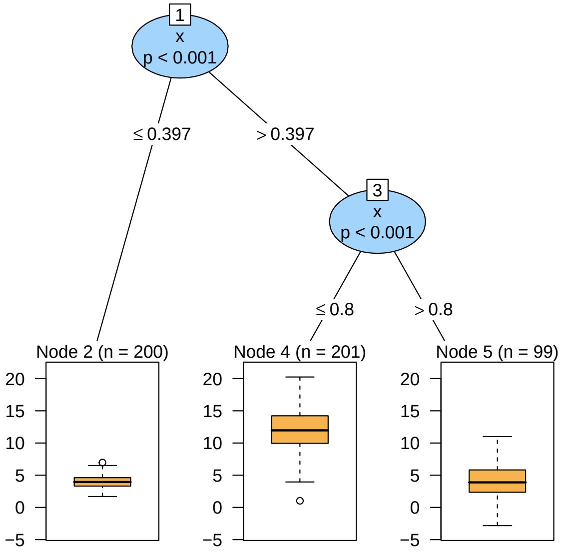
Fitting a Tree

Model: `disttree(y~x)`

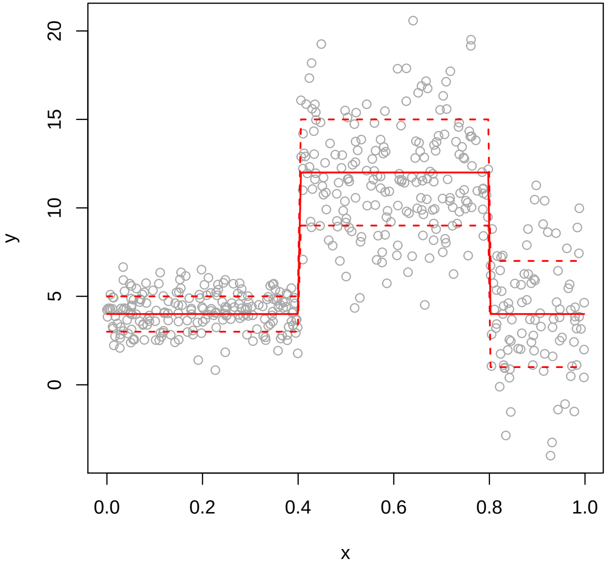


Fitting a Tree

Model: `disttree(y~x)`

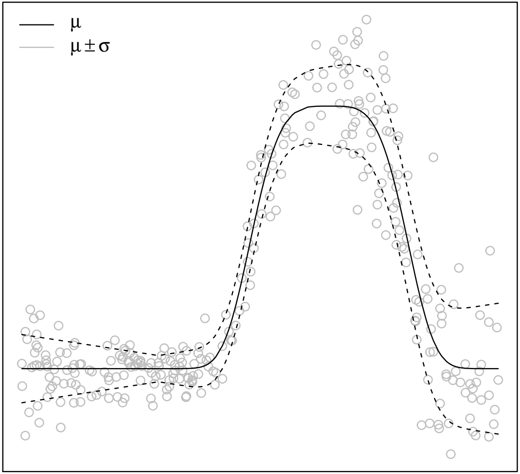


Fitting a Tree

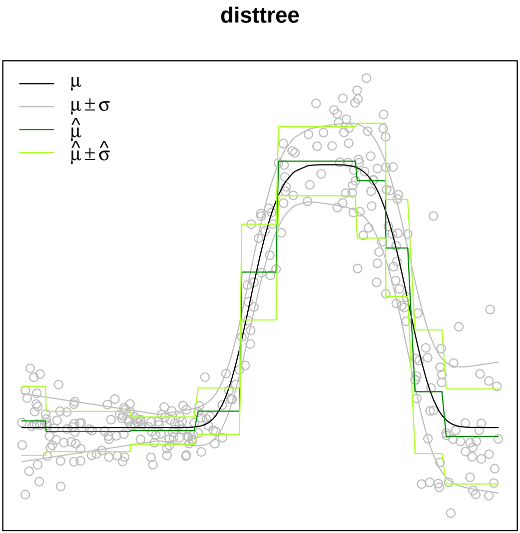


Simulation

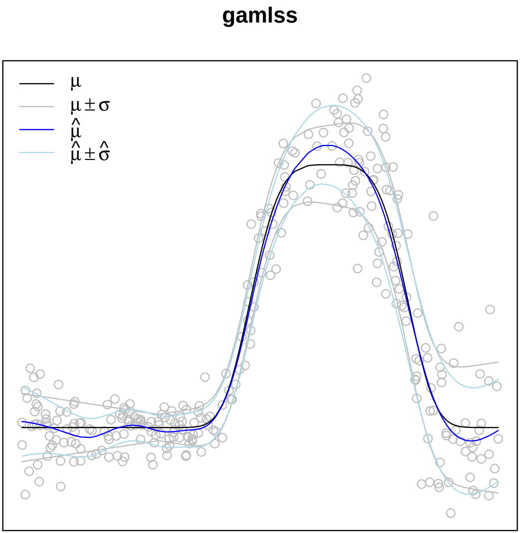
true parameters



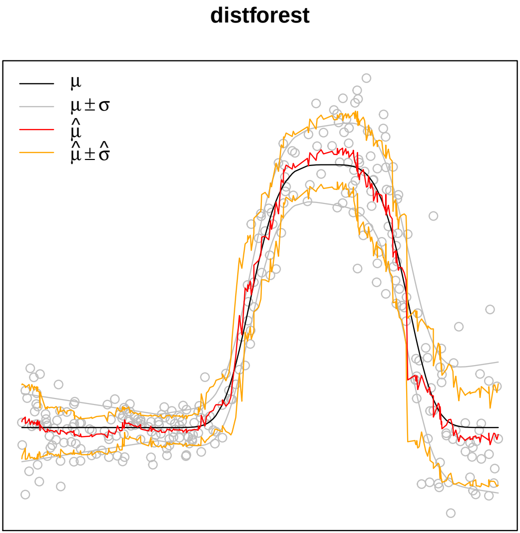
Simulation



Simulation

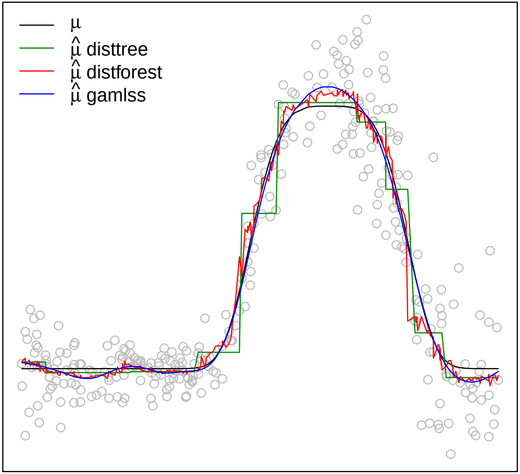


Simulation

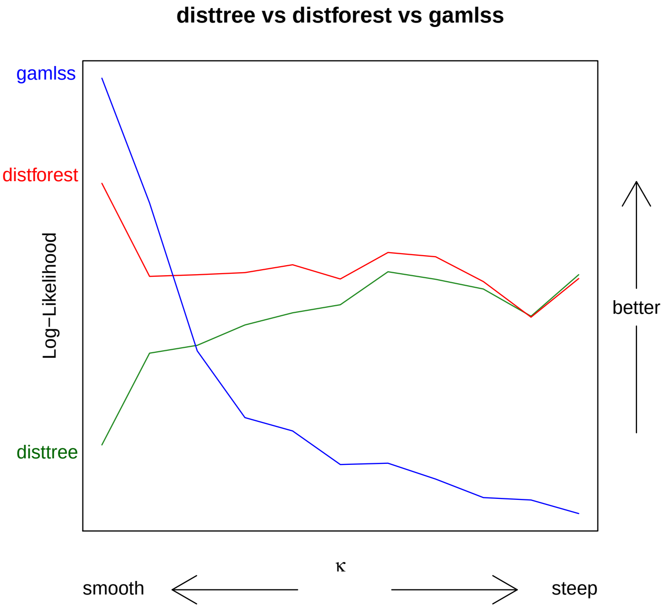


Simulation

disttree vs distforest vs gamlss



Simulation


disttree vs distforest vs gamlss
gamlss
distforest
disttree
Log−Likelihood
better
κ
smooth
steep

Software

R-package **disttree** available on R-Forge:

`https://R-Forge.R-project.org/projects/partykit/` Main

functions:

- `distfit()`
- `disttree()`
- `distforest()`

References

Hothorn T, Hornik K, Zeileis A (2006).
"Unbiased Recursive Partitioning: A Conditional Inference Framework."
Journal of Computational and Graphical Statistics, **15**(3), 651–674.

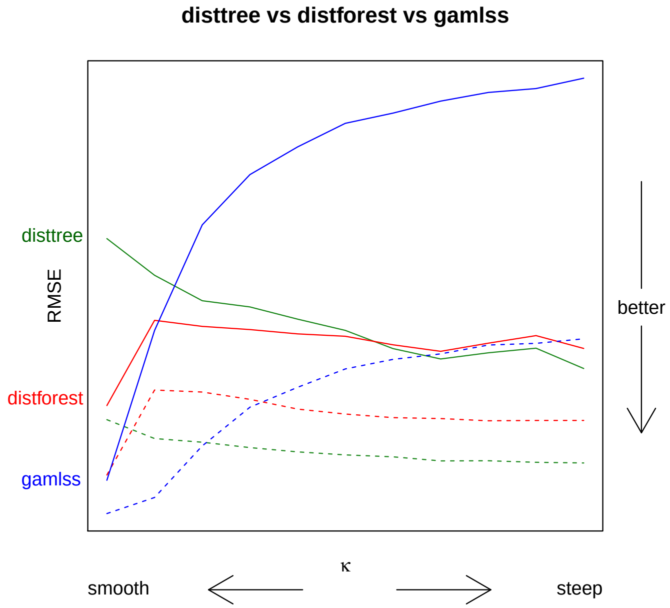
Hothorn T, Zeileis A (2015).
"Partykit: A Modular Toolkit for Recursive Partytioning in R."
Journal of Machine Learning Research, **16**, 3905–3909.

Stasinopoulos DM, Rigby RA, et al. (2007).
"Generalized Additive Models for Location Scale and Shape (GAMLSS) in R."
Journal of Statistical Software, **23**(7), 1–46.

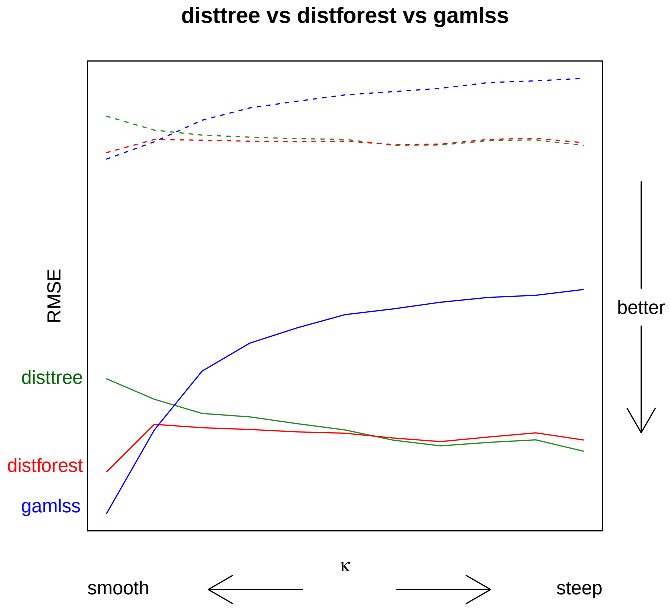
Zeileis A, Hothorn T, Hornik K (2008).
"Model-Based Recursive Partitioning."
Journal of Computational and Graphical Statistics, **17**(2), 492–514.

Seibold H, Zeileis A, Hothorn T (2016).
"Model-Based Recursive Partitioning for Subgroup Analyses."
The international journal of biostatistics, **12**(1), 45–63.

Simulation


disttree vs distforest vs gamlss
disttree
distforest
gamlss
RMSE
better
κ
smooth
steep

Simulation



disttree vs distforest vs gamlss
RMSE
disttree
distforest
gamlss
better
κ
smooth
steep