Transformation Forests

Torsten Hothorn

Joint work with Achim Zeileis, Lisa Schlosser and Heidi Seibold

Linköpings Universitet, 2018-05-23

Machine Learning

Machine Learning methods give

> *computers the ability to learn without being explicitly programmed.*

(Arthur Samuel, 1959)

Actually: Fit statistical models to data by clever optimisation of appropriate target functions

Machine Learning

Source: `https://xkcd.com/1838/`

Statistical Learning

An oxymoron, like "Statistical Science"

Either you learn, or you estimate

Statistical Modelling

Too dull a term to attract any grant money

However: Explicitly acknowledges the underlying probabilistic theory

Statistical Models

What is a statistical model?

$$Y \sim \mathbb{P}_Y$$

What is a regression model?

$$Y \mid \mathbf{X} = \mathbf{x} \sim \mathbb{P}_{Y \mid \mathbf{X} = \mathbf{x}}$$

Random Forest

What is a random forest (in general, not only B&C)?

Classical:

$$\mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}) = f(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X}$$

Here:

$$\mathbb{P}(Y \leq y \mid \mathbf{X} = \mathbf{x}) = \mathbb{P}_{Y \mid \mathbf{X} = \mathbf{x}}(y) = f(y \mid \mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X}$$

Parametric (!) Setup

Unconditional model for response

$$\mathbb{P}_{Y,\Theta} = \{\mathbb{P}_{Y,\boldsymbol{\vartheta}} \mid \boldsymbol{\vartheta} \in \Theta\}$$

Assumption: Regression model belongs to this family:

$$\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}} = \mathbb{P}_{Y,\boldsymbol{\vartheta}(\mathbf{x})}$$

Task: Estimate $\boldsymbol{\vartheta}$ function

Likelihood Contributions

"Learning" data $(y_i, \mathbf{x}_i), i = 1, \dots, N$ plus family $\mathbb{P}_{Y,\Theta}$ defines likelihood function

$$\ell_i : \Theta \rightarrow \mathbb{R}$$

$\ell_i(\boldsymbol{\vartheta}(\mathbf{x}_i))$ gives the likelihood for observation i with candidate parameters $\boldsymbol{\vartheta}(\mathbf{x}_i)$

Handle censoring and truncation appropriately here

Adaptive Local Likelihood Estimators

$$\hat{\vartheta}^N(\mathbf{x}) := \operatorname*{arg\,max}_{\vartheta \in \Theta} \sum_{i=1}^{N} w_i^N(\mathbf{x}) \ell_i(\vartheta)$$

Conditioning works via weight functions $w_i^N(\mathbf{x})$ only

Unconditional Maximum Likelihood

$$\hat{\boldsymbol{\vartheta}}_{\text{ML}}^{N} := \operatorname*{arg\,max}_{\boldsymbol{\vartheta} \in \Theta} \sum_{i=1}^{N} \ell_i(\boldsymbol{\vartheta})$$

Trees

$$\mathcal{X} = \dot{\bigcup_{b=1,\dots,B}} \mathcal{B}_b$$

$$\begin{aligned} w^N_{\text{Tree},i}(\mathbf{x}) &:= \sum_{b=1}^{B} I(\mathbf{x} \in \mathcal{B}_b \wedge \mathbf{x}_i \in \mathcal{B}_b) \\ \hat{\boldsymbol{\vartheta}}^N_{\text{Tree}}(\mathbf{x}) &:= \operatorname*{arg\,max}_{\boldsymbol{\vartheta} \in \Theta} \sum_{i=1}^{N} w^N_{\text{Tree},i}(\mathbf{x}) \ell_i(\boldsymbol{\vartheta}) \end{aligned}$$

Forests

$$\mathcal{X} = \dot{\bigcup_{b=1,\dots,B_t}} \mathcal{B}_{tb} \text{ for } t = 1, \dots, T \text{ trees}$$

$$\begin{aligned} w^N_{\text{Forest},i}(\mathbf{x}) &:= \sum_{t=1}^{T} \sum_{b=1}^{B_t} I(\mathbf{x} \in \mathcal{B}_{tb} \wedge \mathbf{x}_i \in \mathcal{B}_{tb}) \\ \hat{\boldsymbol{\vartheta}}^N_{\text{Forest}}(\mathbf{x}) &:= \operatorname*{arg\,max}_{\boldsymbol{\vartheta} \in \Theta} \sum_{i=1}^{N} w^N_{\text{Forest},i}(\mathbf{x}) \ell_i(\boldsymbol{\vartheta}) \end{aligned}$$

OK, Done! Really?

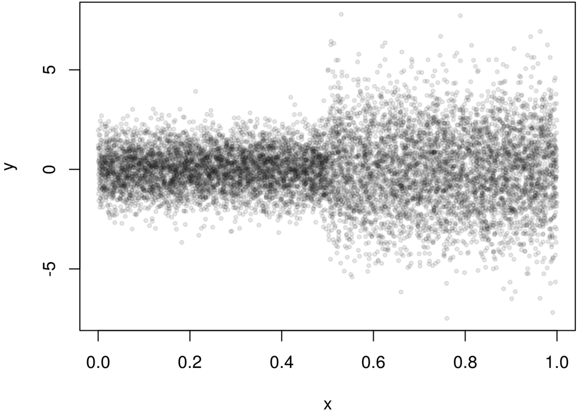
These “nearest neighbor weights” have been used before, first in

- “bagging survival trees” (2004), in
- “conditional inference forests” (**party(kit)**, since 2005) and in
- “quantile regression forests” (**quantregForest**, since 2006)

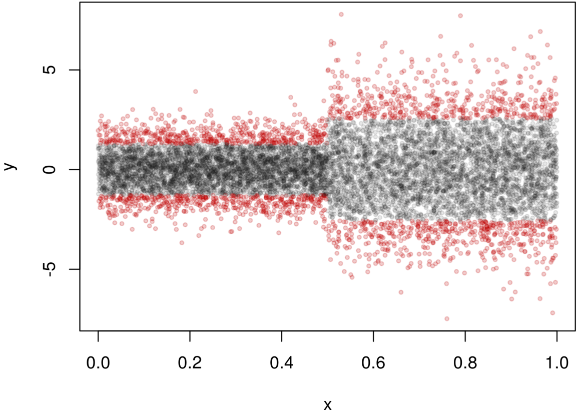
with *standard* trees (CART- or CTree-like).

Unfortunately, there is a catch.

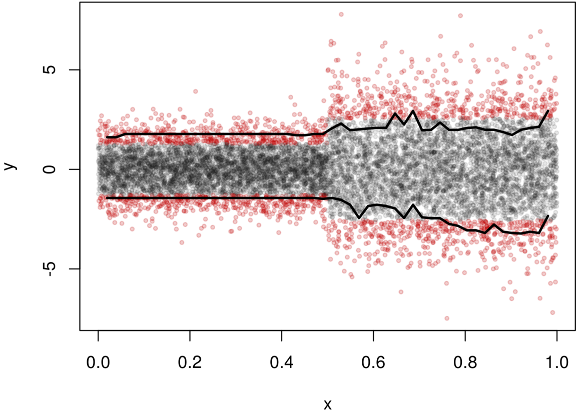
The Problem



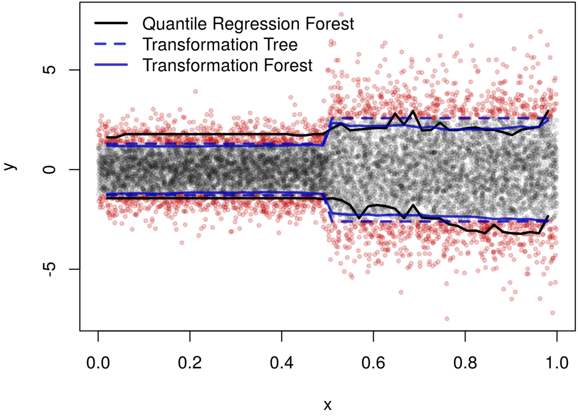
The Problem



The Problem



The Problem


Quantile Regression Forest
Transformation Tree
Transformation Forest
y
5
0
-5
0.0
0.2
0.4
0.6
0.8
1.0
x

The Solution

We need splits sensitive to *distributional* and not just *mean* changes.

Generic approach ("Distribution trees and forests"):

$$\mathbb{P}(Y \leq y \mid \mathbf{X} = \mathbf{x}) = \mathbb{P}_{Y,\boldsymbol{\vartheta}(\mathbf{x})}(y)$$

Here: Use transformation model

$$\mathbb{P}(Y \leq y \mid \mathbf{X} = \mathbf{x}) = F_Z(\mathbf{a}(y)^\top \boldsymbol{\vartheta}(\mathbf{x}))$$

Why Transformation Models?

With

$$\mathbb{P}(Y \leq y) = \mathbb{P}(h(Y) \leq h(y)) = F_Z(h(y))$$

we can generate *all* distributions $\mathbb{P}_Y$ from some F_Z and a corresponding h.

Suitable parameterisations of $h(y) = \mathbf{a}(y)^\top \boldsymbol{\vartheta}$ preserve much of this generality.

Why Transformation Models?

As we *always* observe intervals $(\underline{y}, \bar{y}]$ the exact likelihood is

$$\mathcal{L}(\boldsymbol{\vartheta} | Y \in (\underline{y}, \bar{y}]) := F_Z(\mathbf{a}(\bar{y})^\top \boldsymbol{\vartheta}) - F_Z(\mathbf{a}(\underline{y})^\top \boldsymbol{\vartheta})$$

- Always defined, always a probability (Lindsey, 1999, JRSS-D)
- Applicable to discrete responses
- Covers all types of random censoring and truncation
- For a precise datum y of some continuous Y, the likelihood can be *approximated* by the density

$$f_Y(y) = f_Z(\mathbf{a}(y)^\top \boldsymbol{\vartheta}) \mathbf{a}'(y)^\top \boldsymbol{\vartheta}$$

Why Transformation Models?

Three ways to look at a normal linear model:

1.

$$\begin{aligned} Y &= \alpha + \tilde{\mathbf{x}}^\top \boldsymbol{\beta} + \sigma\varepsilon, \quad \varepsilon \sim \mathsf{N}(0,1) \\ \mathbb{E}(Y - \alpha | \mathbf{X} = \mathbf{x}) &= \tilde{\mathbf{x}}^\top \boldsymbol{\beta} \end{aligned}$$

2.

$$\mathbb{P}(Y \le y | \mathbf{X} = \mathbf{x}) = \Phi\left(\frac{y - \alpha - \tilde{\mathbf{x}}^\top \boldsymbol{\beta}}{\sigma}\right)$$

3.

$$\begin{aligned} \mathbb{P}(Y \le y | \mathbf{X} = \mathbf{x}) &= \Phi(\tilde{\alpha}_1 + \tilde{\alpha}_2 y - \tilde{\mathbf{x}}^\top \tilde{\boldsymbol{\beta}}) \\ \mathbb{E}(\tilde{\alpha}_1 + \tilde{\alpha}_2 Y | \mathbf{X} = \mathbf{x}) &= \tilde{\mathbf{x}}^\top \tilde{\boldsymbol{\beta}} \end{aligned}$$

with $\tilde{\alpha}_1 = -\alpha/\sigma, \tilde{\alpha}_2 = 1/\sigma > 0$ and $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}/\sigma$.

Why Transformation Models?

View (3) allows us to see that the normal linear model is of the form

$$\begin{aligned}\mathbb{P}(Y \le y | \mathbf{X} = \mathbf{x}) &= F_Z(h_Y(y) - \tilde{\mathbf{x}}^\top \tilde{\boldsymbol{\beta}}) \\ \mathbb{E}(h_Y(Y) | \mathbf{X} = \mathbf{x}) &= \tilde{\mathbf{x}}^\top \tilde{\boldsymbol{\beta}}\end{aligned}$$

with F_Z a cdf of an absolutely continuous rv Z and h_Y a monotone "baseline transformation function".

With $F_Z(z) = 1 - \exp(-\exp(z))$ and "unspecified" h_Y we get the continuous proportional hazards, or Cox, model.

Other choices of F_Z and h_Y generate *all* linear transformation models.

Why Transformation Models?

"Linear" transformation models

$$\mathbb{P}(Y \leq y \mid \mathbf{X} = \mathbf{x}) = F_Z(\mathbf{a}(y)^\top \boldsymbol{\vartheta} - \tilde{\mathbf{x}}^\top \boldsymbol{\beta})$$

"Non-linear" transformation models

$$\mathbb{P}(Y \leq y \mid \mathbf{X} = \mathbf{x}) = F_Z(\mathbf{a}(y)^\top \boldsymbol{\vartheta} - \beta(\mathbf{x}))$$

Conditional transformation models

$$\mathbb{P}(Y \leq y \mid \mathbf{X} = \mathbf{x}) = F_Z(\mathbf{a}(y)^\top \boldsymbol{\vartheta}(\mathbf{x}))$$

with additive structure of $\boldsymbol{\vartheta}(\mathbf{x})$
Transformation trees/forests

$$\mathbb{P}(Y \leq y \mid \mathbf{X} = \mathbf{x}) = F_Z(\mathbf{a}(y)^\top \boldsymbol{\vartheta}(\mathbf{x}))$$

with non-linear structure of $\boldsymbol{\vartheta}(\mathbf{x})$

Parameterisation

Transformation trees and forests based on parameterisation

$$\mathbb{P}(Y \leq y \mid \mathbf{X} = \mathbf{x}) = F_Z(\mathbf{a}_{\text{Bs},d}(y)^\top \boldsymbol{\vartheta}(\mathbf{x}))$$

- $\mathbf{a}_{\text{Bs},d}(y)^\top \boldsymbol{\vartheta}(\mathbf{x})$ is a smooth, monotonic Bernstein polynomial of degree d
- $d = 1$ with $F_Z = \Phi$ means $\mathbb{P}_{Y \mid \mathbf{X} = \mathbf{x}} = \mathcal{N}(\mu(\mathbf{x}), \sigma^2(\mathbf{x}))$
- $d = 5$ is surprisingly flexible

Model-based Recursive Partitioning (MOB)

Core idea

- Fit parameters $\hat{\boldsymbol{\vartheta}}_{\text{ML}}$ in *unconditional* model $\mathbb{P}_{Y,\boldsymbol{\vartheta}}$
- Compute individual gradient contributions (“scores”)

$$\mathbf{s}_i = \left. \frac{\partial \ell_i(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} \right|_{\boldsymbol{\vartheta} = \hat{\boldsymbol{\vartheta}}_{\text{ML}}}$$

- Select predictor from **x** with strongest parameter instability as indicated by highest association to $\mathbf{s}_i, i = 1, \dots, N$
- Find “best” binary split; repeat recursively

Implemented for many models, including (G)LM(M)s, parametric survival, β-regression, spatial lag, Bradley-Terry-Luce, various Item Response Theory models, subgroup analyses, etc.

Transformation Trees (TTree)

- Start with $\hat{\boldsymbol{\vartheta}}_{\text{ML}}^N$
- Search for parameter instabilities in $\hat{\boldsymbol{\vartheta}}_{\text{ML}}^N$ as a function of $\mathbf{x}$ using (a beefed-up version) of MOB
- Potentially find changes in the mean AND higher moments
- Forests: Aggregate these trees via adaptive local likelihood estimation

Transformation Forests (TForest)

$$\hat{\mathbb{P}}(Y \le y \mid \mathbf{X} = \mathbf{x}) = \Phi(\mathbf{a}_{\text{Bs},d}(y)^\top \hat{\boldsymbol{\vartheta}}^N_{\text{Forest}}(\mathbf{x}))$$

makes the forest “parametric” (one model for each $\mathbf{x}$) with

- Forest likelihood
- Prediction intervals
- Likelihood-based variable importance
- Parametric bootstrap
- ...

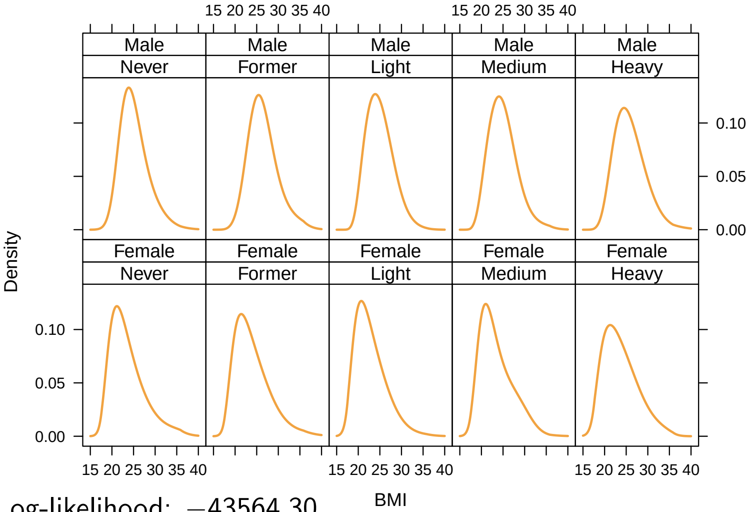
and applicable to censored and truncated data.

Swiss Body Mass Index Distributions

2012 survey ($N = 16427$) in Switzerland
Explain conditional distribution of BMI given

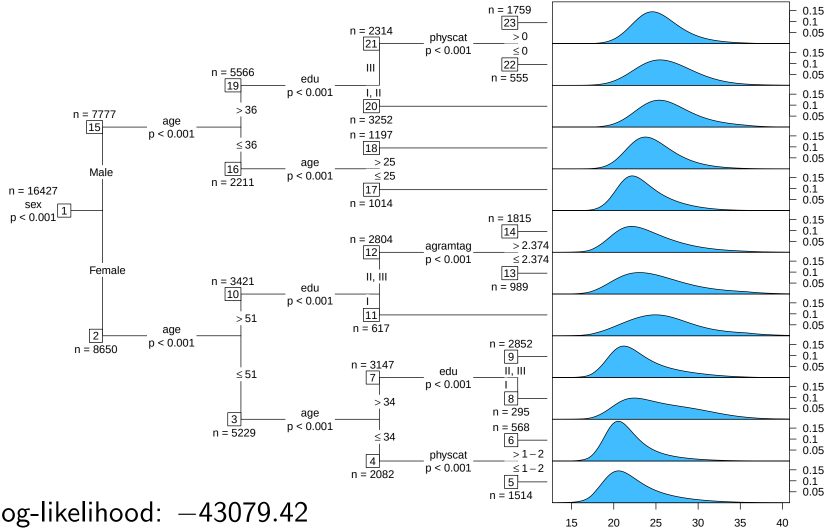
- Sex,
- Smoking status,
- Age,
- Education,
- Physical activity,
- Alcohol intake,
- Fruit and vegetable consumption,
- Region, and
- Nationality.

BMI by Sex and Smoking



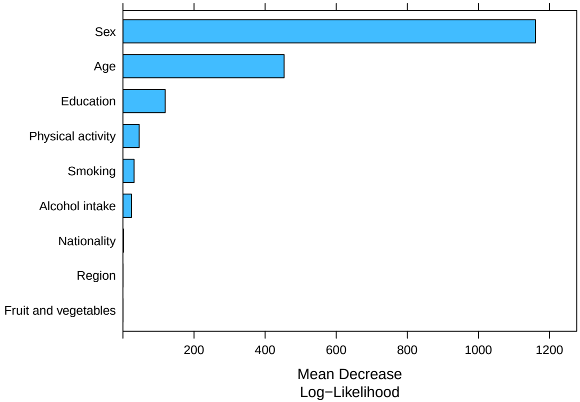
Log-likelihood: -43564.30

Transformation Tree



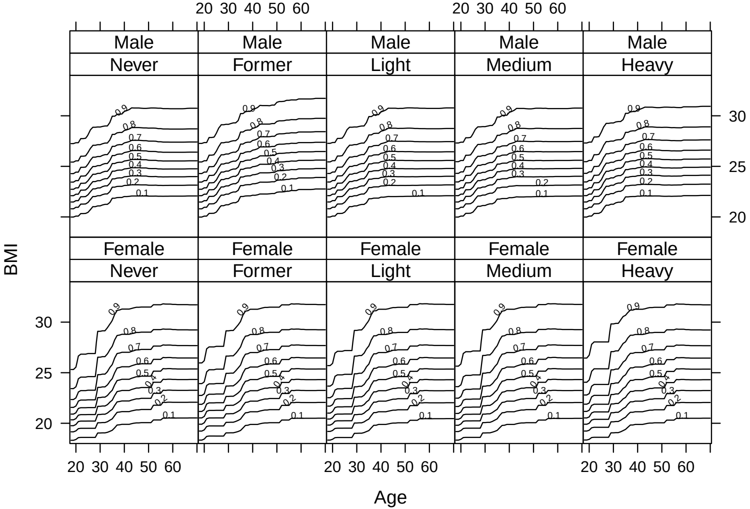
Log-likelihood: -43079.42

Transformation Forest: Variable Importance



Log-likelihood: -42520.18

Transformation Forest: Partial Deciles



More Complex Models

For example: Subgroup analysis, stratified / personalised medicine, ...

Conditional transformation model

$$\mathbb{P}(Y \le y \mid \text{treatment}, \mathbf{X} = \mathbf{x}) = F_Z(\mathbf{a}_{\text{Bs},d}(y)^\top \boldsymbol{\vartheta}(\mathbf{x}) - \beta(\mathbf{x}) I(\text{treated}))$$

- Both the “intercept function” $\mathbf{a}_{\text{Bs},d}(y)^\top \boldsymbol{\vartheta}(\mathbf{x})$ and
- the treatment effect $\beta(\mathbf{x})$ may depend on $\mathbf{x}$
- $F_Z() = 1 - \exp(-\exp())$ makes β a log-hazard ratio
- Include $\hat{\beta}$ in search for parameter instabilities

Stratified Medicine

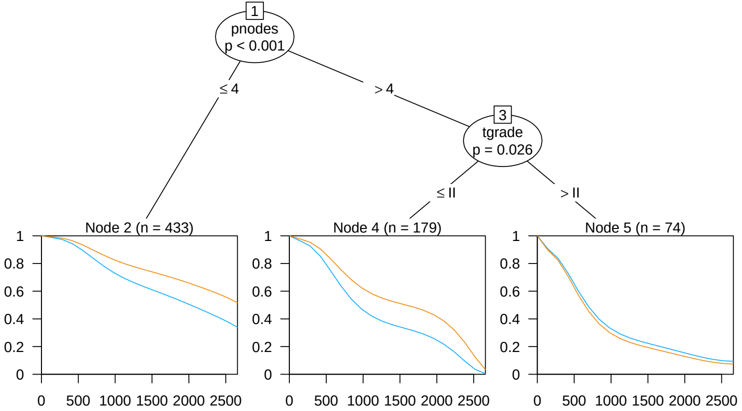
Partition log-hazard ratio β from a fully parametric Cox model

$$\mathbb{P}(T > t \mid \text{treatment}) = \exp(-\exp(\mathbf{a}_{\text{Bs},d}(t)^\top \boldsymbol{\vartheta} - \beta I(\text{treated}))$$

for a randomised controlled clinical trial on hormonal treatment of breast-cancer patients

```
> library("tram")
> cmod <- Coxph(ctime ~ horTh, data = GBSG2)
> library("trtf")
> tmod <- trafotree(cmod,
+                   formula = ctime ~ horTh | .,
+                   data = GBSG2)
```

Stratified Medicine



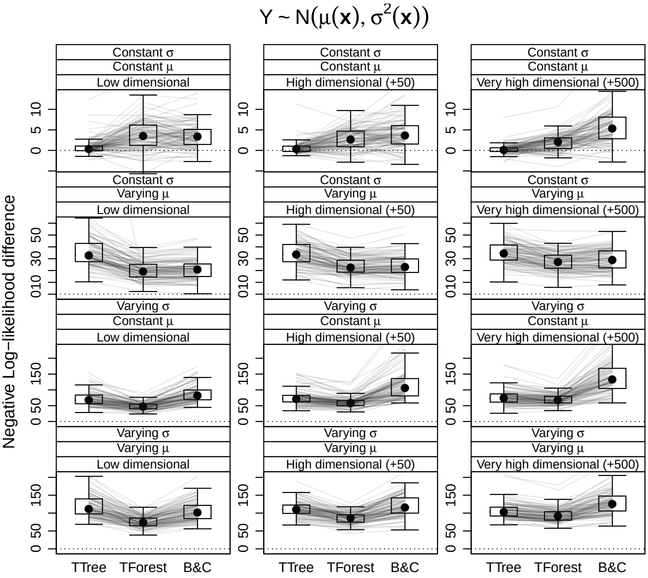
Discussion

- The "two cultures" of statistical modelling come closer
- With $Y =$ BMI, rain, house prices, survival time etc.

$$\hat{\mathbb{E}}(Y|\mathbf{X} = \mathbf{x}) = \hat{f}(\mathbf{x}) = \mathbf{x}^\top \hat{\boldsymbol{\beta}}$$

  not interesting (or even harmful)
- $\mathbb{P}_{Y,\hat{\boldsymbol{\vartheta}}(\mathbf{x})}$ more informative
- Flexibility (non-linear interactions) of B&C random forests preserved
- Simplicity of B&C random forests preserved
- Large sample behaviour?
- High dimensional?

Low and High

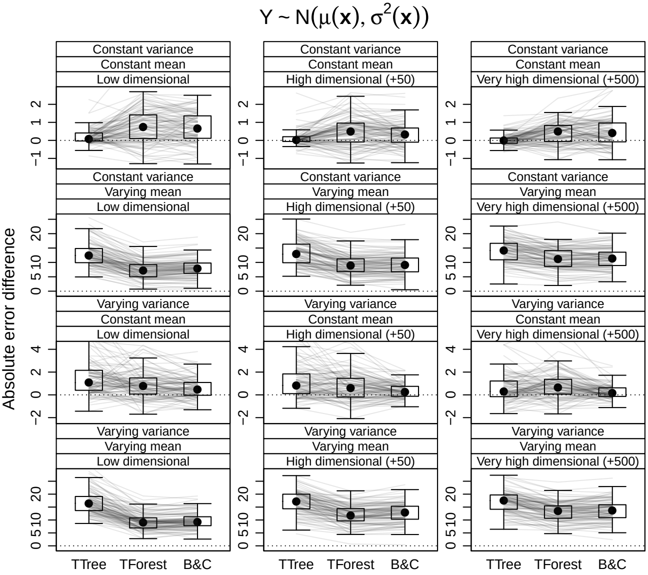


$Y \sim N(\mu(\mathbf{x}), \sigma^2(\mathbf{x}))$
Negative Log−likelihood difference
Constant σ
Constant μ
Low dimensional
Constant σ
Constant μ
High dimensional (+50)
Constant σ
Constant μ
Very high dimensional (+500)
Constant σ
Varying μ
Low dimensional
Constant σ
Varying μ
High dimensional (+50)
Constant σ
Varying μ
Very high dimensional (+500)
Varying σ
Constant μ
Low dimensional
Varying σ
Constant μ
High dimensional (+50)
Varying σ
Constant μ
Very high dimensional (+500)
Varying σ
Varying μ
Low dimensional
Varying σ
Varying μ
High dimensional (+50)
Varying σ
Varying μ
Very high dimensional (+500)
0
5
10
0
10
30
50
0
50
150
TTree
TForest
B&C

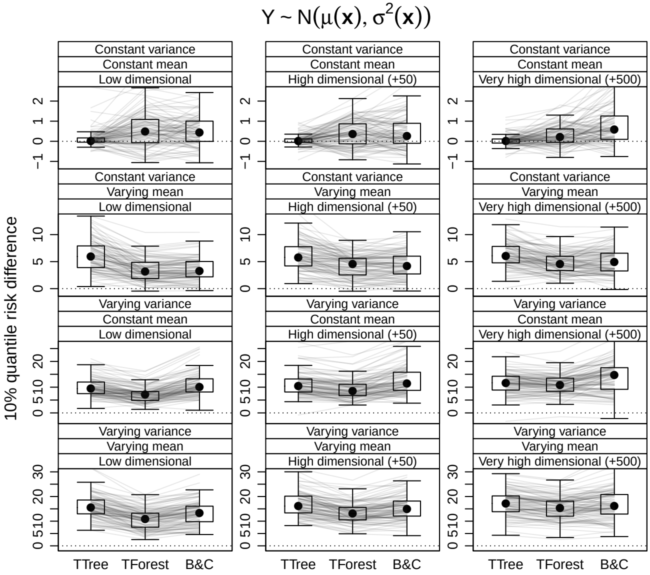
Resources

- "Transformation Forests", **trtf**,
  https://arxiv.org/abs/1701.02110,
- "Top-Down Transformation Choice" (with BMI example), SM, **trtf**, http://arxiv.org/abs/1706.08269
- "Most Likely Transformations", SJoS, **mlt**, **tram**,
  http://dx.doi.org/10.1111/sjos.12291
- "Conditional Transformation Models", JRSS-B,
  http://dx.doi.org/10.1111/rssb.12017
- "Model-based Recursive Partitioning", JCGS, **partykit**
  http://dx.doi.org/10.1198/106186008X319331,
- "Model-based Recursive Partitioning for Subgroup Analyses", IJB, **model4you**
  http://dx.doi.org/10.1515/ijb-2015-0032
- "Model-based Forests", SMMR, **model4you**,
  http://dx.doi.org/10.1177/0962280217693034

Low and High: Median



Low and High: 10% Quantile



Low and High: 90% Quantile