Transformation Boosting Machines: Empirical Evaluation of the tbm Package

Torsten Hothorn
Universität Zürich

Abstract

This document discusses technical details on the empirical evaluation of the reference implementation of transformation boosting machines in the **tbm** package. Model estimation, interpretation, and criticism is illustrated for eight life science applications. Setup and detailed results of a simulation study based on artificial data generating processes are presented.

Keywords: Conditional Transformation Model, Shift Transformation Model.

1. Introduction

Transformation boosting machines estimate conditional transformation models (CTMs) and shift transformation models (STMs) by optimising the corresponding log-likelihood. In a nutshell, transformation models are fully parameterised by appropriate basis functions those parameters are (in total or partially) linked to predictor variables. The method is applicable to a broad range of problems because (1) the likelihood allows to handle discrete and continuous responses under all forms or random censoring and truncation and (2) simple linear, more complex nonlinear additive, or completely unstructured (tree-based) conditional paramater functions can be estimated by using appropriate basis functions for the predictor variables.

This document describes technical details of the empirical evaluation of transformation boosting machines by means of eight applications (Section 2) and artificial data generating processes (Section 3). For each of the eight applications, the exact model setup, the best performing model, and the corresponding model interpretation and model criticism is presented. The source code for these benchmark analysis is available from the directory

```
R> system.file("applications", package = "tbm")
```

Simulation experiments based on the data generating processes presented in Section 3 can be reproduced using the source code in

```
R> system.file("simulations", package = "tbm")
```

This document also contains graphical representations of the out-of-sample log-likelihoods for all settings in the simulation experiments (Section 3.2).
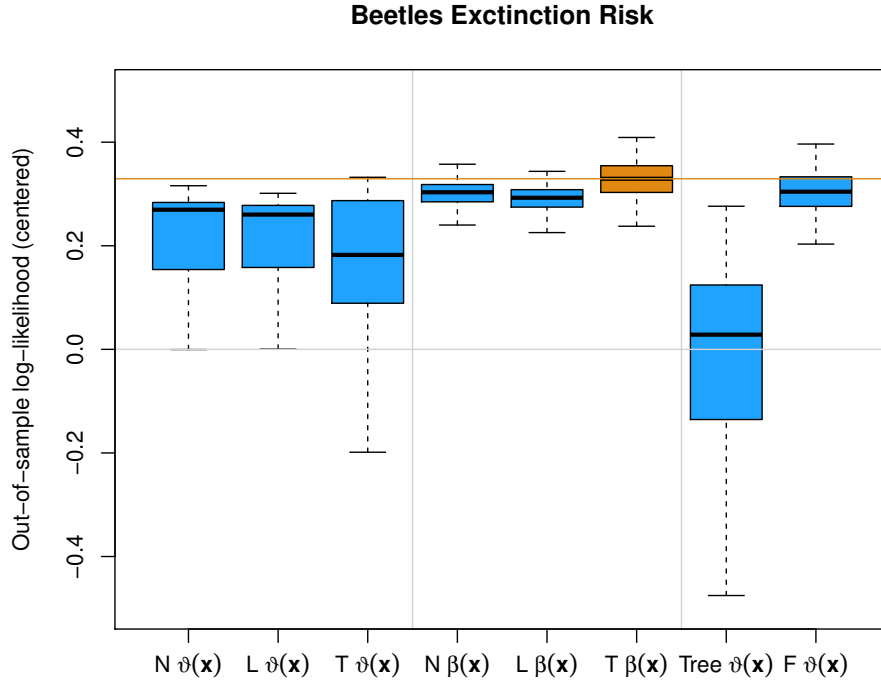
Beetles Exctinction Risk

Figure 1: Beetles Exctinction Risk. Out-of-sample log-likelihood (centered by out-of-sample log-likelihood of unconditional model).

2. Applications

For eight prediction problems, the out-of-sample log-likelihood was evaluated based on 100 subsamples ($3/4N$ as learning sample and $1/4N$ as test sample). For problems with categorical responses or right-censored responses, subsamples were stratified with respect to the response class or censoring status (such that the distribution of the response or the the censoring rate were the same in learning and test samples). The out-of-sample log-likelihood was centered by the out-of-sample log-likelihood of the uninformative model which was estimated and tested on the very same folds.

Transformation boosting machines were estimated using package **tbm** (Hothorn 2019b). Transformation trees and forests as implemented in package **trtf** (Hothorn 2019a) served as main competitors.

Boosting CTM Likelihoods (parameter $\boldsymbol{\vartheta}(\boldsymbol{x})$) with nonlinear (N, B-spline) basis functions, linear (L) basis functions, and tree-based (T, depth two) basis functions. Boosting STM Likelihoods (parameter $\beta(\boldsymbol{x})$) with nonlinear (N, B-spline) basis functions, linear (L) basis functions, and tree-based (T, depth two) basis functions.

2.1. Beetles Exctinction Risk

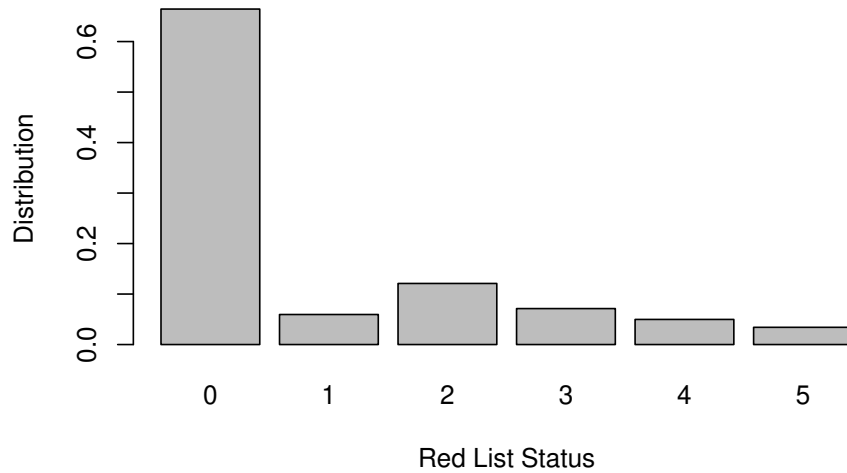This problem aims at the prediction of the exctinction risk of beetles (Seibold *et al.* 2015).

Figure 2: Beetles Exctinction Risk. Unconditional response distribution.

The response is the Red List status ranging from 0 (least concern) to 5 (regionally extinct) of $N = 1025$ saproxylic beetles, where for each species 10 (7 numeric and 3 categorical) predictors are available.

We start with an unconditional proportional odds model (using F_Z = expit and one parameter for each but the highest Red List category):

```
R> m_mlt <- Polr(RL ~ 1, data = ldata)
R> logLik(m_mlt)

'log Lik.' -1176.571 (df=5)
```

The corresponding unconditional distribution is depicted in Figure 2.

Model evaluation (Figure 1) indicated that a tree-structured model for $\beta(\boldsymbol{x})$ had the best performance; this model was fitted using

```
R> fm_tree

RL ~ mean_body_size + mean_elev + flowers + niche_decay + niche_diam +
    niche_canopy + distribution + tree + feeding + habitat

R> bm <- stmboost(m_mlt, formula = fm_tree, data = ldata,
+                 method = quote(mboost::blackboost))[626]
```

(the default tree was grown to a depth of two). The in-sample log-likelihood is shown in Figure 3.

The model consists of trees with two-way interactions, basically all variables seemed to play a role in the model (the numbers giving the absolute number of splits in each of these variables):
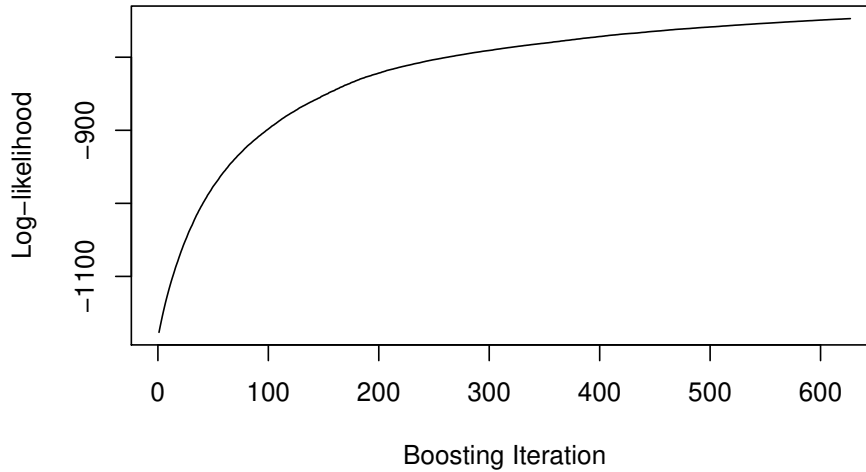
Figure 3: Beetles Exctinction Risk. In-sample log-likelihood maximisation.

```
mean_body_size        mean_elev         flowers      niche_decay      niche_diam
           215              100              66               60             252
  niche_canopy     distribution            tree          feeding         habitat
           141              483             307              195              59
```

For six selected beetle species, the conditional distribution of extinction as predicted from the tree-based model is given in Figure 4.

In this special case, the algorithm is identical to boosting for proportional odds models as implemented in the `PropOdds()` family (Schmid *et al.* 2011); and thus the in-sample risks are essentially identical:

```
R> po <- blackboost(fm_tree, data = ldata, family = PropOdds())[626]
R> max(abs(risk(bm) - risk(po)))

[1] 0.488843
```

2.2. Birth Weight Prediction

Models shall be used to improve birth weight prediction in small fetuses (weighting ≤ 1600 g at birth) based on measurements obtained using three-dimensional (3D) sonography (Schild *et al.* 2008).

The response is the birth weight in gram of $N = 150$ singleton fetuses, where for each fetus 5 numeric predictors are available.

We start with an unconditional model (using $F_Z = \Phi$ and an Bernstein polynomial of order six):
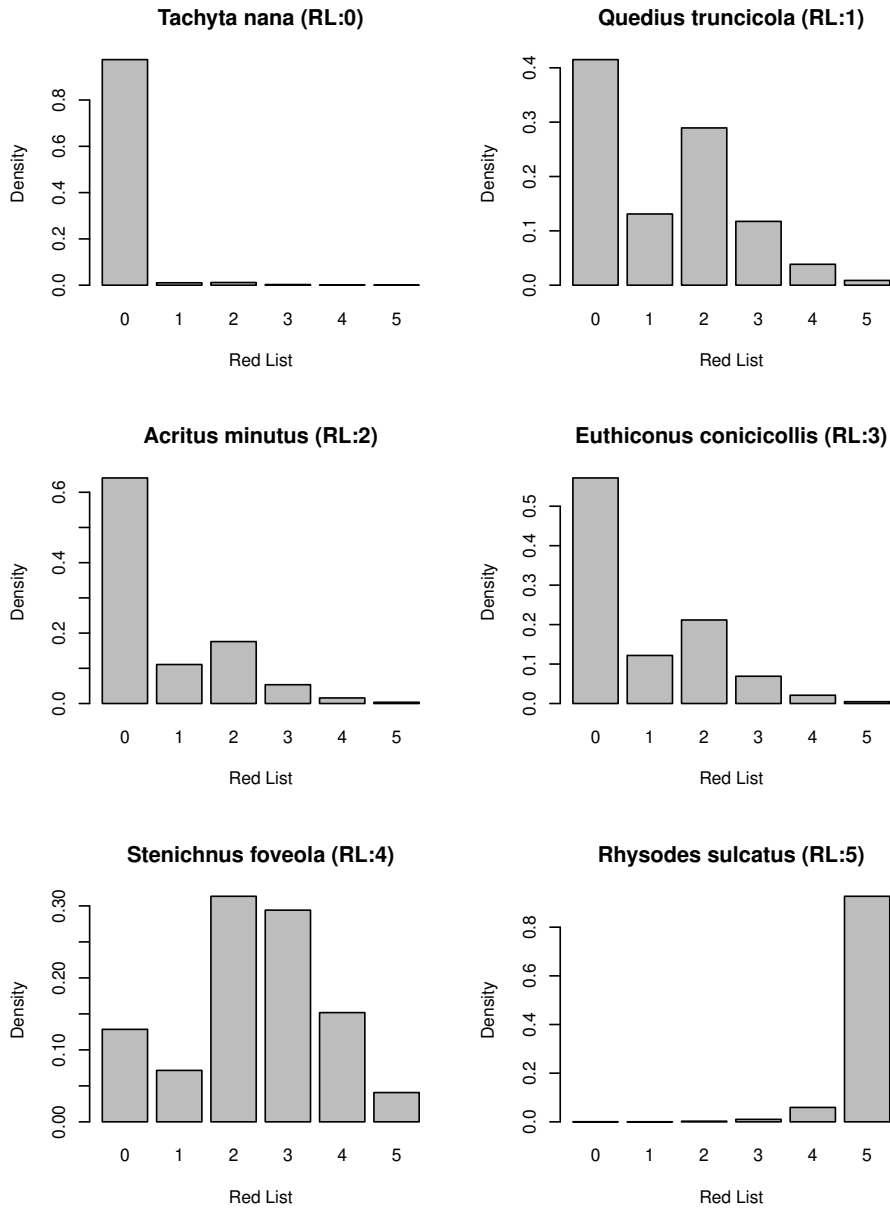
Figure 4: Beetles Exctinction Risk. Conditional response distribution.

Birth Weight Prediction

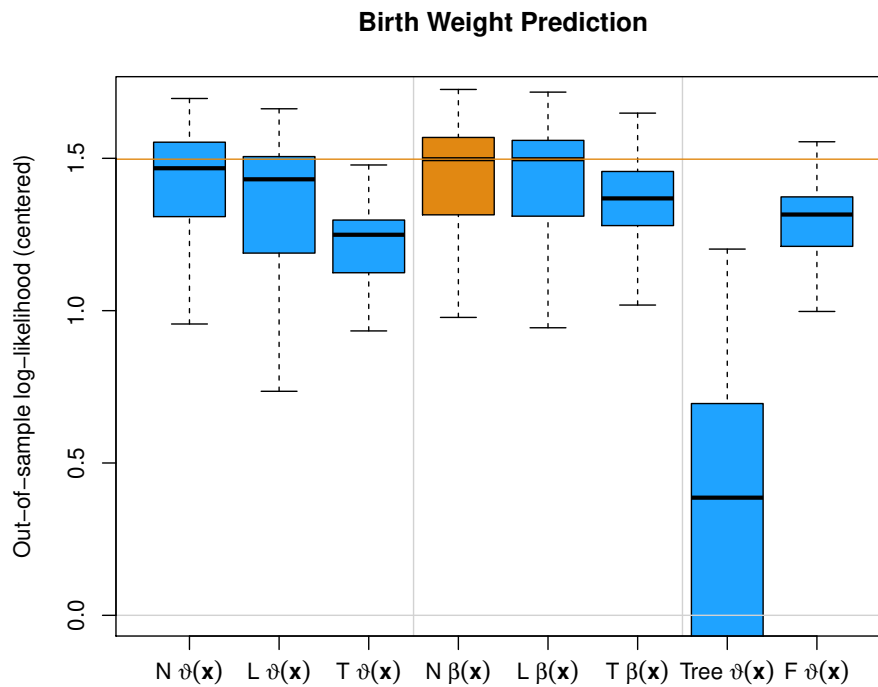Figure 5: Birth Weight Prediction. Out-of-sample log-likelihood (centered by out-of-sample log-likelihood of unconditional model).

```
R> m_mlt <- BoxCox(birthweight ~ 1, data = ldata, extrapolate = TRUE)
R> logLik(m_mlt)

'log Lik.' -1082.233 (df=7)
```

The corresponding unconditional distribution is depicted in Figure 6.

Model evaluation (Figure 5) indicated that an additive smooth model for $\beta(\boldsymbol{x})$ had the best performance; this model was fitted using

```
R> fm_gam[["stm"]]

birthweight ~ bols(volabdo, intercept = FALSE) + bols(hccalc,
    intercept = FALSE) + bols(volos, intercept = FALSE) + bols(fe,
    intercept = FALSE) + bols(bip, intercept = FALSE) + bbs(volabdo,
    center = TRUE, df = 1) + bbs(hccalc, center = TRUE, df = 1) +
    bbs(volos, center = TRUE, df = 1) + bbs(fe, center = TRUE,
    df = 1) + bbs(bip, center = TRUE, df = 1)

R> bm <- stmboost(m_mlt, formula = fm_gam[["stm"]], data = ldata,
+                 method = quote(mboost::mboost))[253]
```
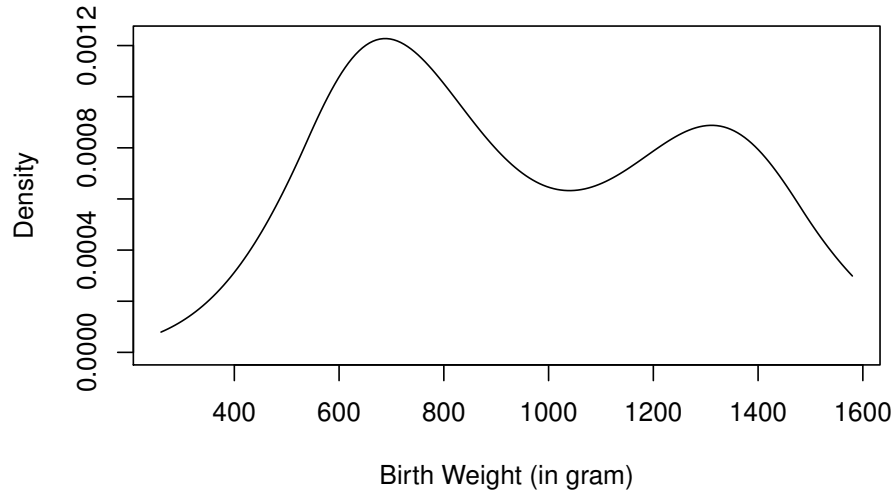
Figure 6: Birth Weight Prediction. Unconditional response distribution.

The in-sample log-likelihood is shown in Figure 7 and a QQ-plot based on the probability-integral transform in Figure 8. It is important to note that the term $\beta(\boldsymbol{x})$ must not contain an intercept. The model was therefore composed of linear model terms (without intercept) and smooth deviations from linear functions (all with the same degree of freedom to facilitate unbiasedness). Thus, the model selects between linear and smooth additive functions.

The selected terms were

```
bols(volabdo, intercept = FALSE)    66
bols(hccalc, intercept = FALSE)     58
bols(volos, intercept = FALSE)      38
bols(fe, intercept = FALSE)         40
bols(bip, intercept = FALSE)         0
bbs(volabdo, df = 1, center = TRUE)  0
bbs(hccalc, df = 1, center = TRUE)   0
bbs(volos, df = 1, center = TRUE)    0
bbs(fe, df = 1, center = TRUE)       0
bbs(bip, df = 1, center = TRUE)     51
```

and the corresponding partial contributions are plotted in Figure 9. Only minor deviations from linearity can be observed (and the out-of-sample risk of the linear model was almost the same). In addition, the baseline transformation (Figure 10) is almost linear, indicating that a normal linear model (reported by Schild *et al.* 2008) might be a good compromise between prediction accuracy and interpretability for this problem.

For five selected observations, the conditional distribution of birth weight is shown in Figure 11. The blue dots indicate the actual observations. Except for the right-most density, the
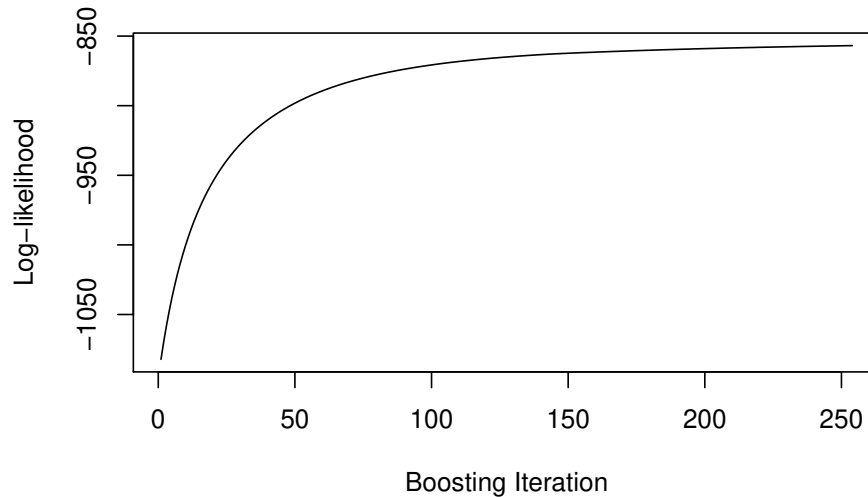
Figure 7: Birth Weight Prediction. In-sample log-likelihood maximisation.

predicted conditional densities are almost perfectly symmetric (as a consequence of the linear baseline transformation).

2.3. Body Fat Mass

The response is the body fat mass (in kilogram) of $N = 71$ study participants, where for each subject 9 numeric predictors are available (Garcia *et al.* 2005).

We start with an unconditional model (using $F_Z = \Phi$ and an Bernstein polynomial of order six):

```
R> m_mlt <- BoxCox(DEXfat ~ 1, data = ldata, prob = c(.1, .99))
R> logLik(m_mlt)

'log Lik.' -267.0084 (df=7)
```

The corresponding unconditional distribution is depicted in Figure 13.

Model evaluation (Figure 12) indicated that an additive smooth model for $\boldsymbol{\vartheta}(\boldsymbol{x})$ had the best performance; this model was fitted using

```
R> fm_gam[["ctm"]]

DEXfat ~ bbs(age) + bbs(waistcirc) + bbs(hipcirc) + bbs(elbowbreadth) +
    bbs(kneebreadth) + bbs(anthro3a) + bbs(anthro3b) + bbs(anthro3c) +
    bbs(anthro4)

R> bm <- ctmboost(m_mlt, formula = fm_gam[["ctm"]], data = ldata,
+                 method = quote(mboost::mboost))[1000]
```
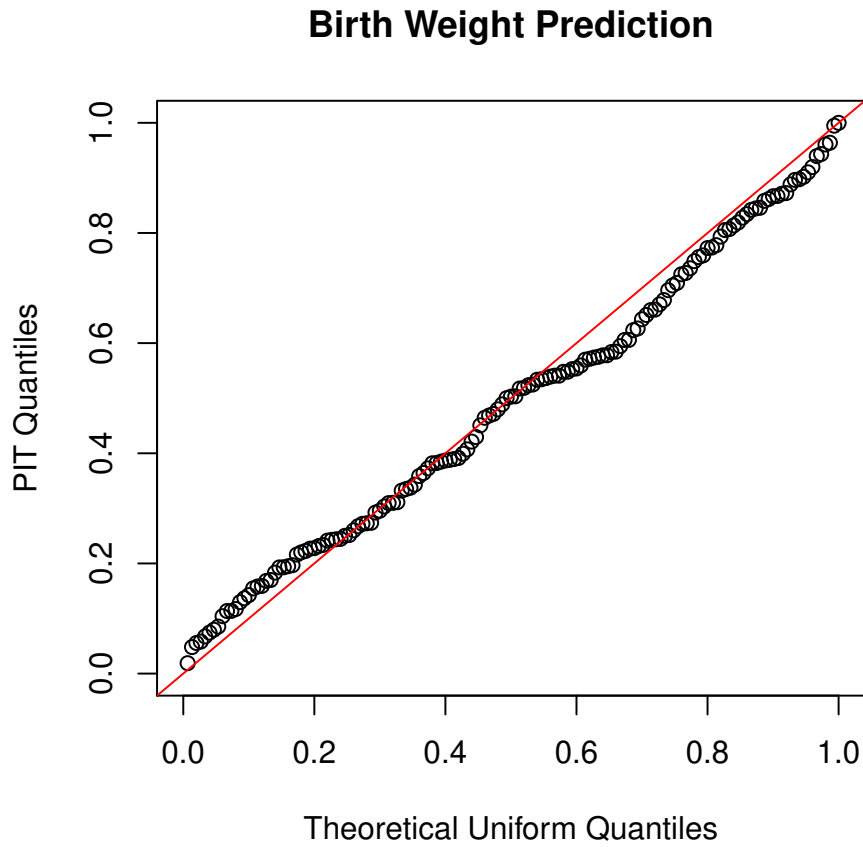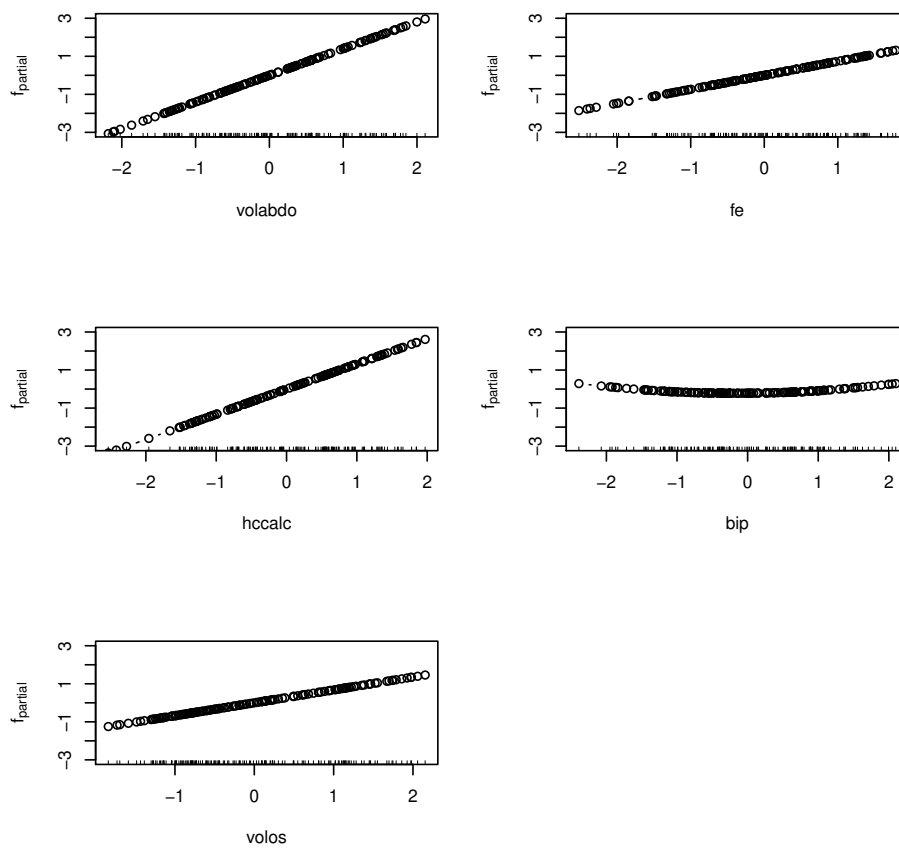
Birth Weight Prediction

Figure 8: Birth Weight Prediction. Probability-integral transform QQ-plot.

The in-sample log-likelihood is shown in Figure 14 and a QQ-plot based on the probability-integral transform in Figure 15.

The model is additive in the sense that $h(y \mid \boldsymbol{x}) = h_0(y) + \sum_{j=1}^{J} h_j(y \mid x_j)$. The partial transformation functions h_j are plotted in Figure 16.

```
bbs(age)          101
bbs(waistcirc)    178
bbs(hipcirc)      145
bbs(elbowbreadth)  72
bbs(kneebreadth)  156
bbs(anthro3a)       4
bbs(anthro3b)     117
bbs(anthro3c)     187
bbs(anthro4)       40
```

For five selected observations, the conditional distribution of body fat mass is shown in Figure 17. The blue dots indicate the actual observations.

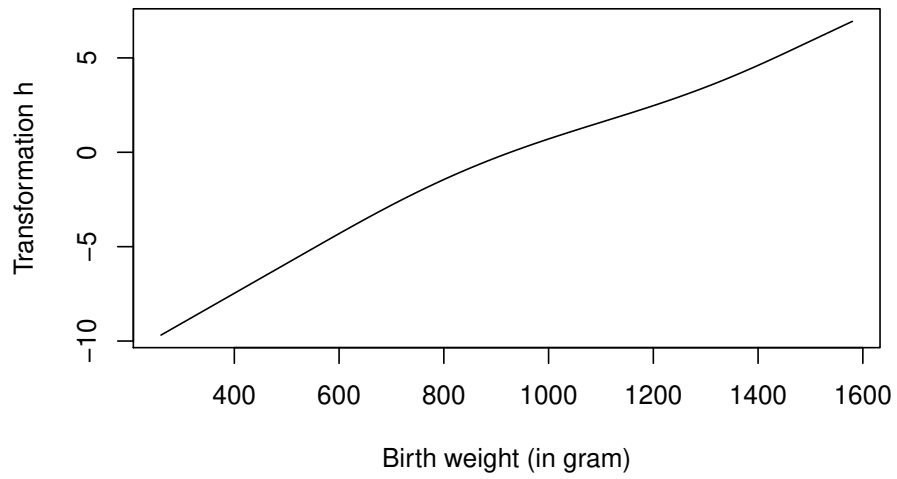Figure 9: Birth Weight Prediction. Partial contributions to additive predictor $\beta(\boldsymbol{x})$.

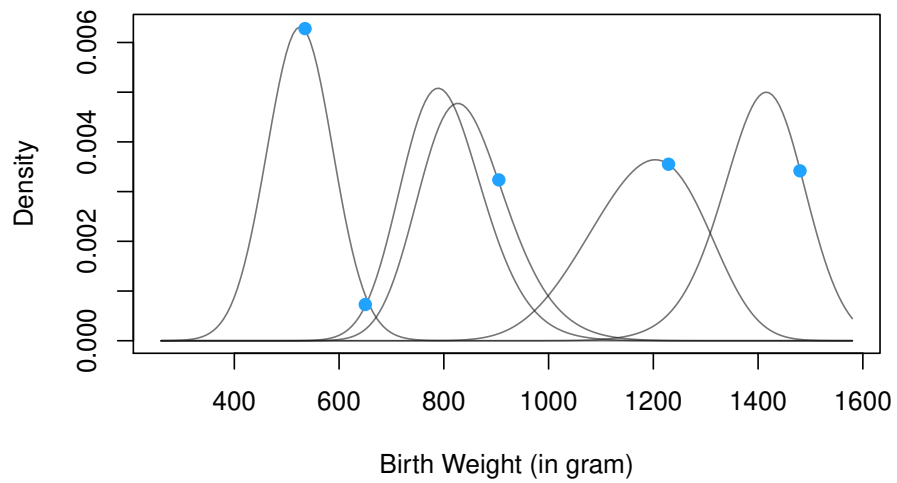Figure 10: Birth Weight Prediction. Baseline transformation.

Figure 11: Birth Weight Prediction. Conditional response distribution.
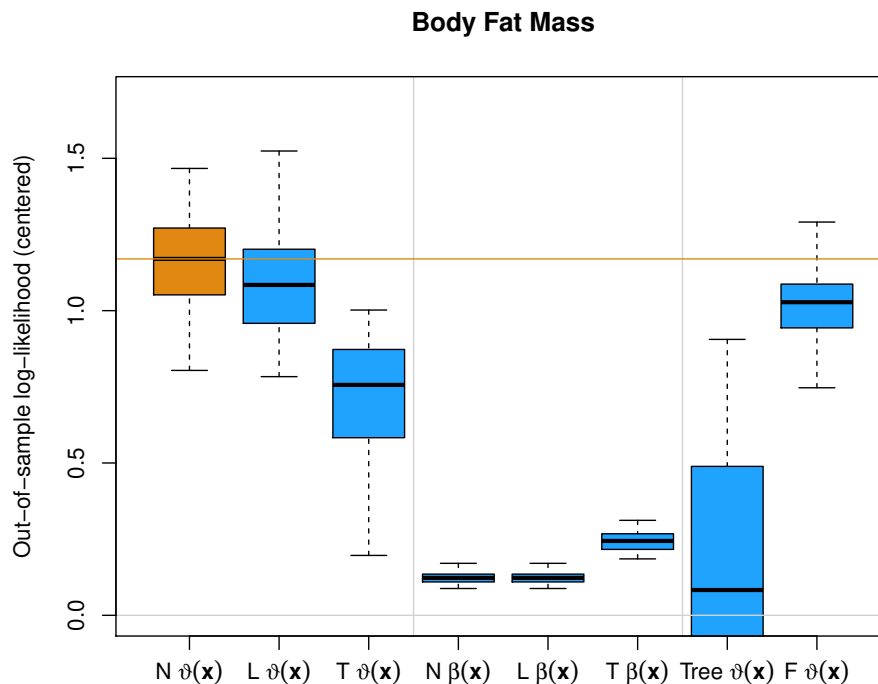
Body Fat Mass

Figure 12: Body Fat Mass. Out-of-sample log-likelihood (centered by out-of-sample log-likelihood of unconditional model).

2.4. CAO/ARO/AIO-04 DFS

The response is the disease-free survival time of rectal cancer patients from the CAO/ARO/AIO-04 randomised controlled clinical trial (Rödel *et al.* 2015). For $N = 1153$ patients, 18 (3 numeric and 15 categorical) baseline predictors are available.

We start with an unconditional Cox model ($F_Z = 1 - \exp(-\exp())$ and a Bernstein polynomial of order six):

```
R> m_mlt <- Coxph(DFS ~ 1, data = ldata, prob = c(0, .9))
R> logLik(m_mlt)
```

```
'log Lik.' -3095.702 (df=7)
```

The corresponding unconditional distribution is depicted in Figure 19.

Model evaluation (Figure 18) indicated that a linear model for $\beta(\boldsymbol{x})$ had the best performance, this model was fitted using

```
R> fm_glm[["stm"]]
R> bm <- stmboost(m_mlt, formula = fm_glm[["stm"]], data = ldata,
+               method = quote(mboost::mboost))[963]
```
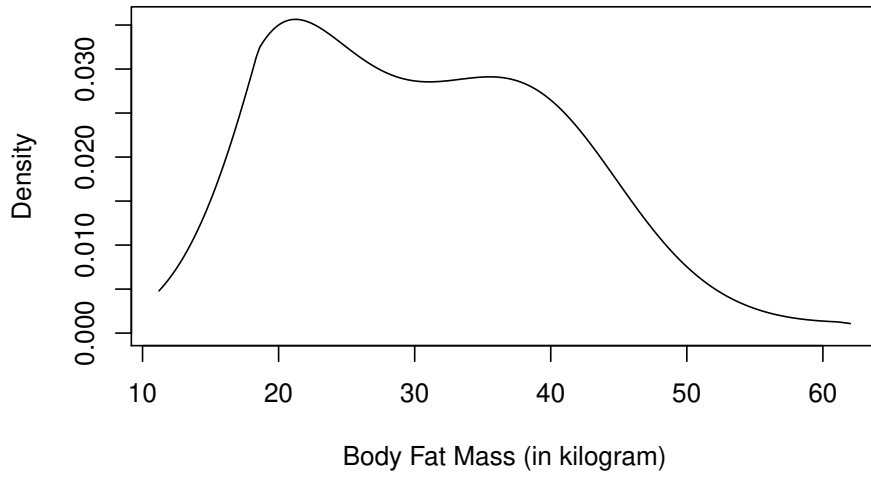
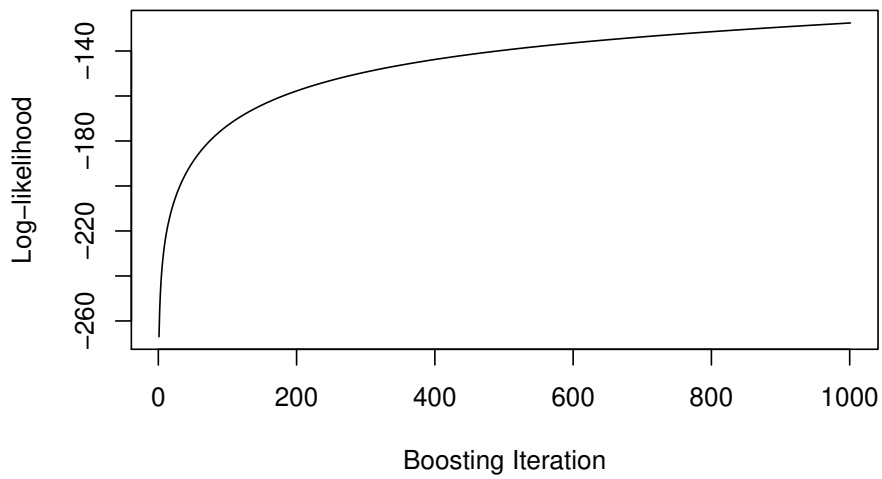Figure 13: Body Fat Mass. Unconditional response distribution.

Figure 14: Body Fat Mass. In-sample log-likelihood maximisation.
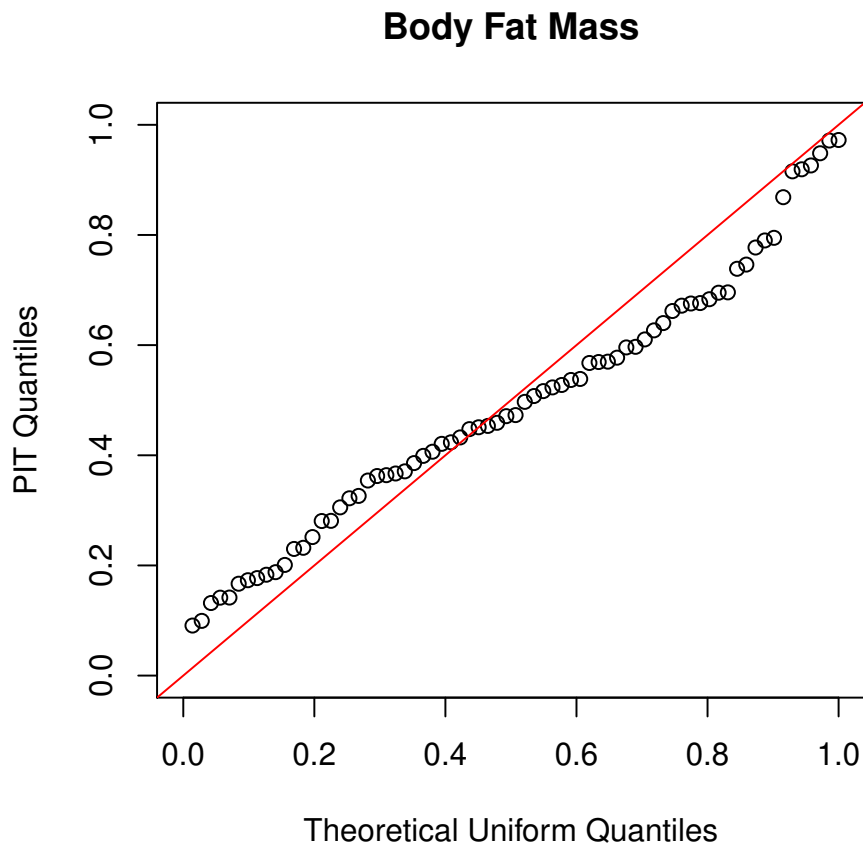
Body Fat Mass

Figure 15: Body Fat Mass. Probability-integral transform QQ-plot.

This is, in fact, a linear Cox model which could have also been fitted using

```
R> bmCoxPH <- mboost(fm_glm[["stm"]], data = ldata, family = CoxPH())[963]
```

2.5. Childhood Malnutrition

The aim is to predict childhood malnutrition, here measured as stunting, that is, insufficient height for age. The data for $N = 24166$ children from India were compiled by Fenske *et al.* (2011) and include 20 (6 numeric and 14 categorical) predictors.

We start with an unconditional model (using $F_Z = \Phi$ and an Bernstein polynomial of order six):

```
R> m_mlt <- BoxCox(stunting ~ 1, data = ldata, prob = c(.05, .975), extrapolate = TRUE)
R> logLik(m_mlt)
```
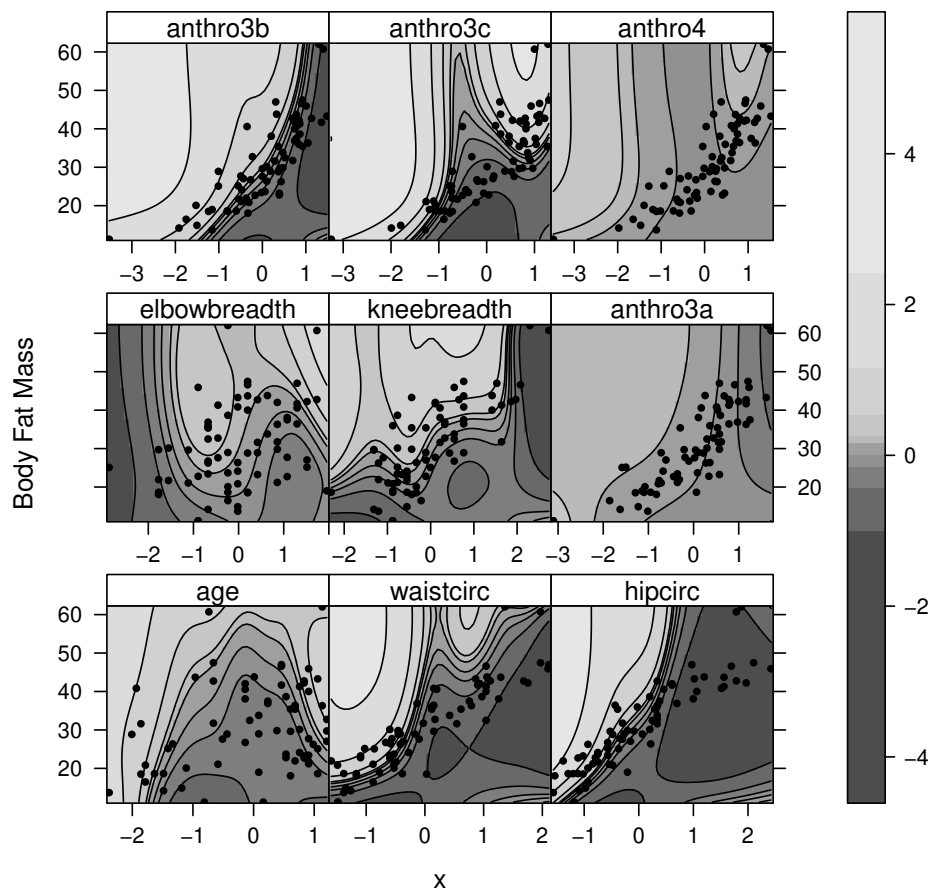
```
'log Lik.' -157843.8 (df=7)
```

Figure 16: Body Fat Mass. Scatterplots of all predictor variables and the response, overlaid with partial transformation functions.
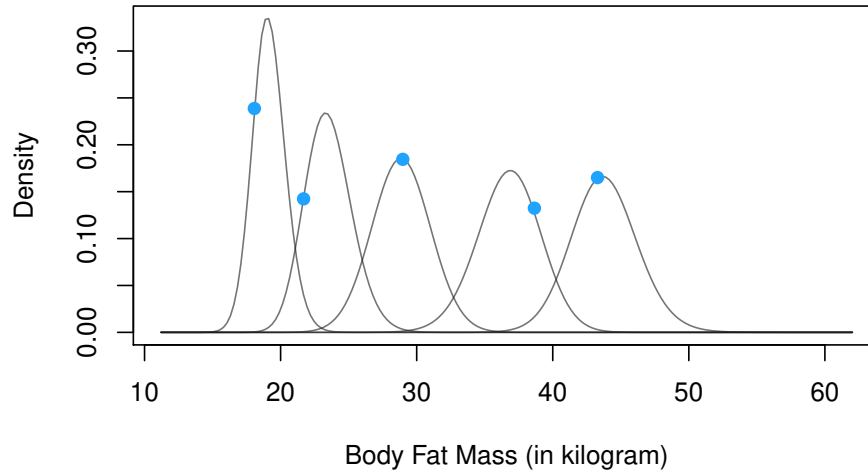
Figure 17: Body Fat Mass. Conditional response distribution.

The corresponding unconditional distribution is depicted in Figure 21.

Model evaluation indicated that a tree-based model for $\boldsymbol{\vartheta}(\boldsymbol{x})$ had the best performance; this model was fitted using

```
R> fm_tree
R> bm <- ctmboost(m_mlt, formula = fm_tree, data = ldata,
+                 method = quote(mboost::blackboost))[515]
```

2.6. Head Circumference

The Fourth Dutch Growth Study (Fredriks *et al.* 2000) is a cross-sectional study that measures growth and development of the Dutch population between the ages of 0 and 22 years. The study measured, among other variables, head circumference (HC) and age of 7482 males and 7018 females. Stasinopoulos and Rigby (2007) analysed the head circumference of 7040 males with explanatory variable age using a GAMLSS model with a Box-Cox t distribution describing the first four moments of head circumference conditionally on age. The models show evidence of kurtosis, especially for older boys.

We start with an unconditional model (using $F_Z = \Phi$ and an Bernstein polynomial of order six):

```
R> m_mlt <- BoxCox(head ~ 1, data = ldata)
R> logLik(m_mlt)

'log Lik.' -21182.17 (df=7)
```

Model evaluation (Figure 22) indicated that an additive smooth model for $\boldsymbol{\vartheta}(\boldsymbol{x})$ had the best performance; this model was fitted using
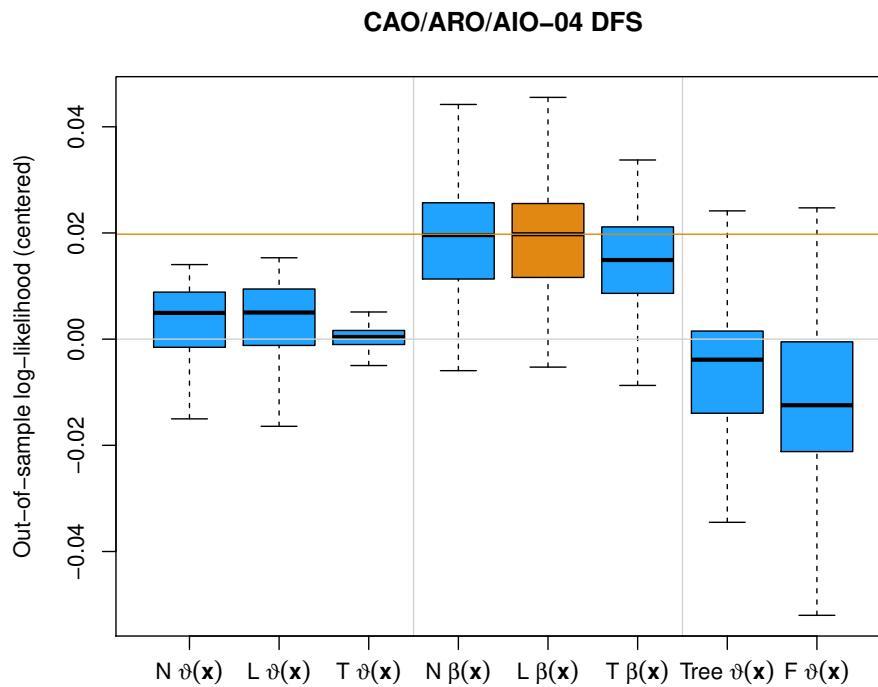
CAO/ARO/AIO–04 DFS

Figure 18: CAO/ARO/AIO-04 DFS. Out-of-sample log-likelihood (centered by out-of-sample log-likelihood of unconditional model).

```
R> fm_gam[["ctm"]]

head ~ bbs(lage)

R> bm <- ctmboost(m_mlt, formula = fm_gam[["ctm"]], data = ldata,
+                 method = quote(mboost::mboost))[1000]
```

The in-sample log-likelihood is shown in Figure 23 and a QQ-plot based on the probability-integral transform in Figure 24.

Figure 25 shows the model as a growth curve model. The observations are overlaid with quantile curves obtained via inversion of the estimated conditional distributions. The figure can be directly compared with Figure 16 of Stasinopoulos and Rigby (2007) (the quantile curves in these two plots are essentially equivalent) and also indicates a certain asymmetry towards older boys.

2.7. PRO-ACT ALSFRS

This data origines from the PRO-ACT Prize4Life challenge 2012 (http://prize4life.org.il/en/prediction-prize/), comprising clinical trials data from $N = 1013$ patients suffering Amypthropic Lateral Sclerosis (ALS). The response is the ALS-Functional Rating Scale
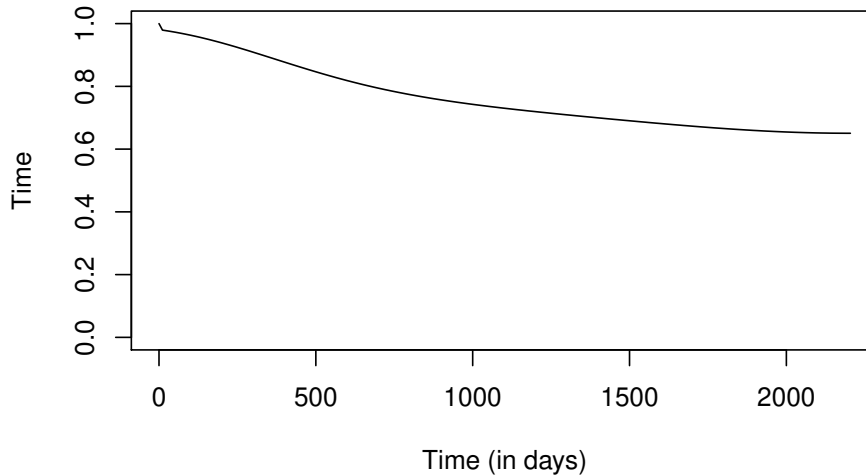
Figure 19: CAO/ARO/AIO-04 DFS. Unconditional response distribution.

(ALSFRS) six months after diagnosis. For each patient, 59 (43 numeric and 16 categorical) predictors are available. An overview on the challenge and winning algorithms is available from Küffner *et al.* (2015).

The response is an ordinal variable, ranging from 0 to 40. We start with an unconditional proportional odds model featuring a continuous basis transformation of the response ($F_Z =$ expit and a Bernstein polynomial of order six):

```
R> m_mlt <- Colr(ALSFRS.halfYearAfter ~ 1, data = ldata, prob = c(.05, .9), extrapolate =
R> logLik(m_mlt)


'log Lik.' -3416.895 (df=7)
```

The corresponding unconditional distribution is depicted in Figure 27.

Model evaluation (Figure 26) indicated that a linear model for $\beta(\boldsymbol{x})$ (also known as "distribution regression") had the best performance; this model was fitted using

```
R> bm <- ctmboost(m_mlt, formula = fm_glm[["ctm"]], data = ldata,
+                 method = quote(mboost::mboost))[613]
```

The in-sample log-likelihood is shown in Figure 28.

Time since onset and ALSFRS at time of diagnosis are the two most important variables (roughly assessed by the selection frequency):

```
          bols(time_onset_treatment, df = 2)
                                        119
               bols(ALSFRS.Start, df = 2)
```

Childhood Malnutrition

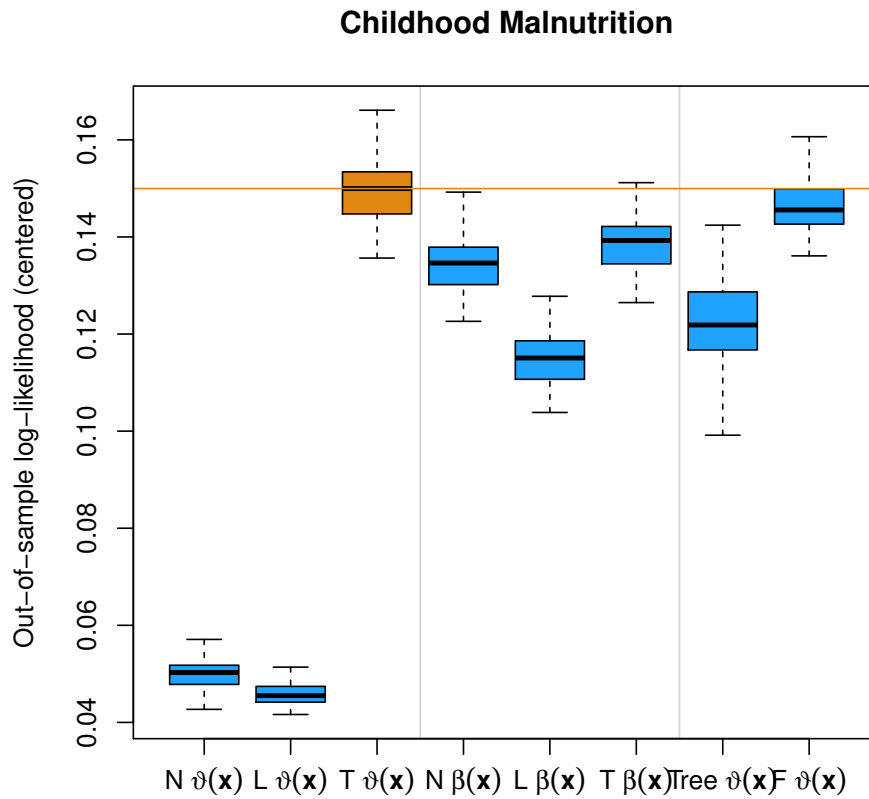Figure 20: Childhood Malnutrition. Out-of-sample log-likelihood (centered by out-of-sample log-likelihood of unconditional model).

```
                                    192
                     bols(sex, df = 2)
                                      5
                     bols(age, df = 2)
                                      7
                  bols(height, df = 2)
                                     20
                 bols(atrophy, df = 2)
                                      5
          bols(fasciculations, df = 2)
                                      5
                  bols(speech, df = 2)
                                     15
               bols(stiffness, df = 2)
                                      1
              bols(swallowing, df = 2)
```
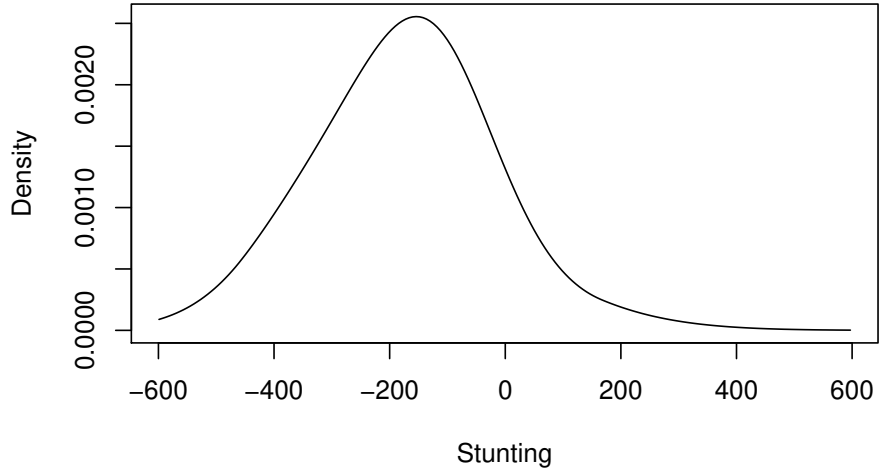
Figure 21: Childhood Malnutrition. Unconditional response distribution.

```
                                              8
          bols(family_history_same, df = 2)
                                             12
                  bols(bp_systolic, df = 2)
                                              7
             bols(subjectliters_fvc, df = 2)
                                             33
               bols(value_chloride, df = 2)
                                              9
                     bols(value_ck, df = 2)
                                             16
    bols(value_white_blood_cell_wbc, df = 2)
                                             19
                bols(value_glucose, df = 2)
                                              7
    bols(value_alkaline_phosphatase, df = 2)
                                              9
                bols(value_calcium, df = 2)
                                              2
                 bols(value_sodium, df = 2)
                                             11
              bols(value_potassium, df = 2)
                                             11
          bols(value_bilirubin_total, df = 2)
                                              1
             bols(value_lymphocytes, df = 2)
```

Head Circumference

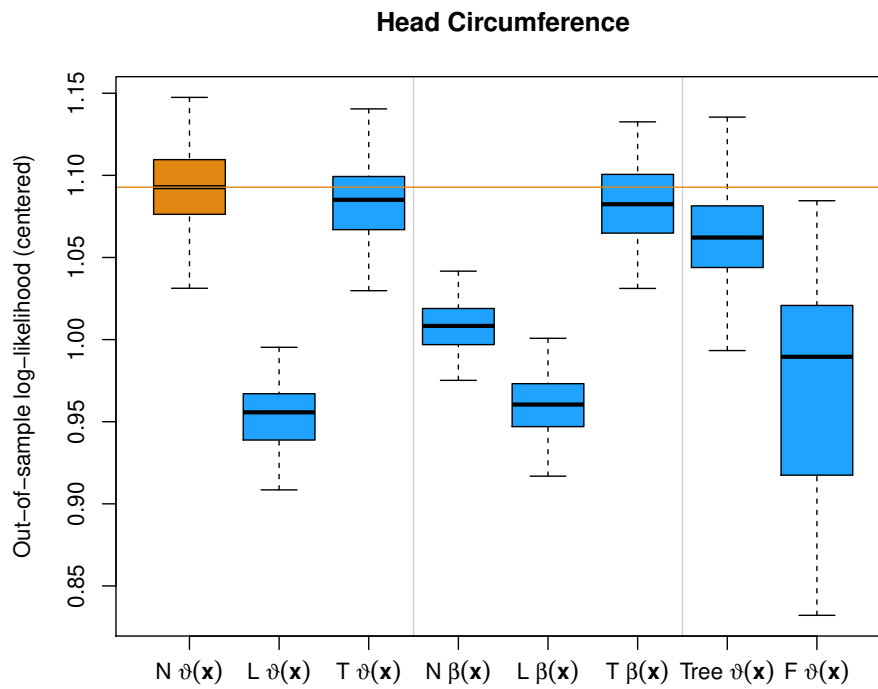Figure 22: Head Circumference. Out-of-sample log-likelihood (centered by out-of-sample log-likelihood of unconditional model).

```
                                                         6
            bols(value_red_blood_cells_rbc, df = 2)
                                                        16
                      bols(value_phosphorus, df = 2)
                                                        15
                         bols(value_altsgpt, df = 2)
                                                         4
      bols(value_absolute_eosinophil_count, df = 2)
                                                         2
                      bols(value_creatinine, df = 2)
                                                        11
                        bols(value_urine_ph, df = 2)
                                                         6
      bols(value_hba1c_glycated_hemoglobin, df = 2)
                                                        21
        bols(value_absolute_monocyte_count, df = 2)
                                                         2
      bols(value_absolute_lymphocyte_count, df = 2)
                                                         7
               bols(value_total_cholesterol, df = 2)
```
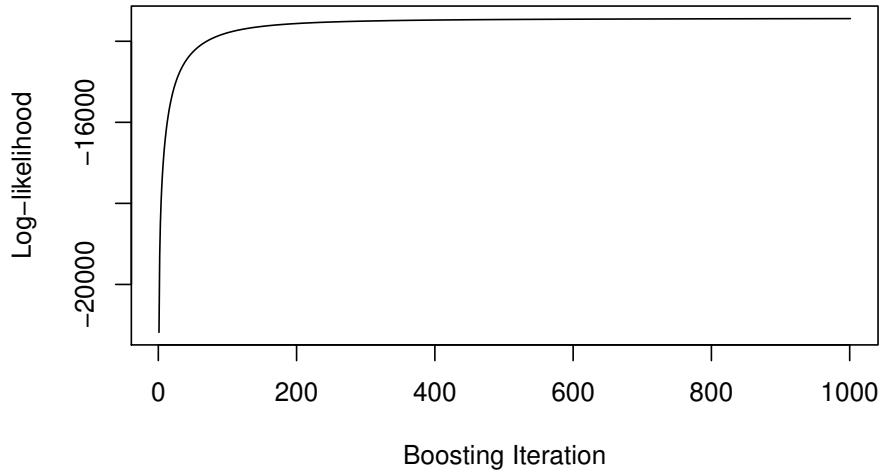
Figure 23: Head Circumference. In-sample log-likelihood maximisation.

```
                             6
       bols(value_triglycerides, df = 2)
                             3
```

The response-varying coefficients for these two variables are given in Figure 29.

2.8. PRO-ACT OS

Overall survival data from ALS patients were compiled from the PRO-ACT database by Seibold *et al.* (2017). For each of $N = 2711$ patients, 19 (3 numeric and 16 categorical) predictors are available.

We start with an unconditional Cox model ($F_Z = 1-\exp(-\exp())$ and a Bernstein polynomial of order six):

```
R> m_mlt <- Coxph(y ~ 1, data = ldata)
R> logLik(m_mlt)


'log Lik.' -6827.902 (df=7)
```

The corresponding unconditional distribution is depicted in Figure 31.

Model evaluation indicated that tree-based model for $\beta(\boldsymbol{x})$ had the best performance; this model was fitted using

```
R> fm_tree
R> bm <- stmboost(m_mlt, formula = fm_tree, data = ldata,
+                 control = boost_control(nu = 0.01),
+                 method = quote(mboost::blackboost))[451]
```

Head Circumference
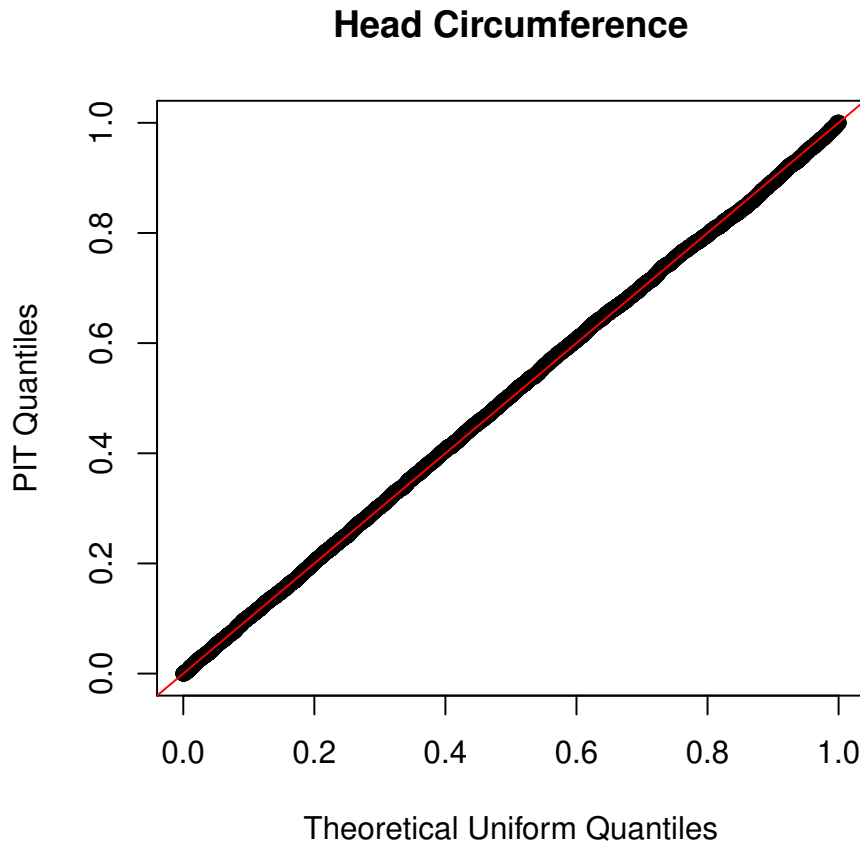
Figure 24: Head Circumference. Probability-integral transform QQ-plot.

This model is roughly equivalent to a boosted Cox-model using trees as baselearners:

```
R> blackboost(fm_tree, data = ldata, family = CoxPH())
```

3. Artificial Data Generating Processes

3.1. Simulation Models

Data were generated based on two groups and one numeric predictor variable $x \in [0, 1]$ based on transformation models of the form

$$\mathbb{P}(Y \leq y \mid \text{Group}, x) = \Phi(h(y \mid \text{Group}, x))$$

where the conditional transformation function $h(y \mid \text{Group}, x) = \Phi^{-1}(\mathbb{P}(Y \leq y \mid \text{Group}, x))$ for four data generating processes is given in Table 1.

General parameters of the simulation were defined as
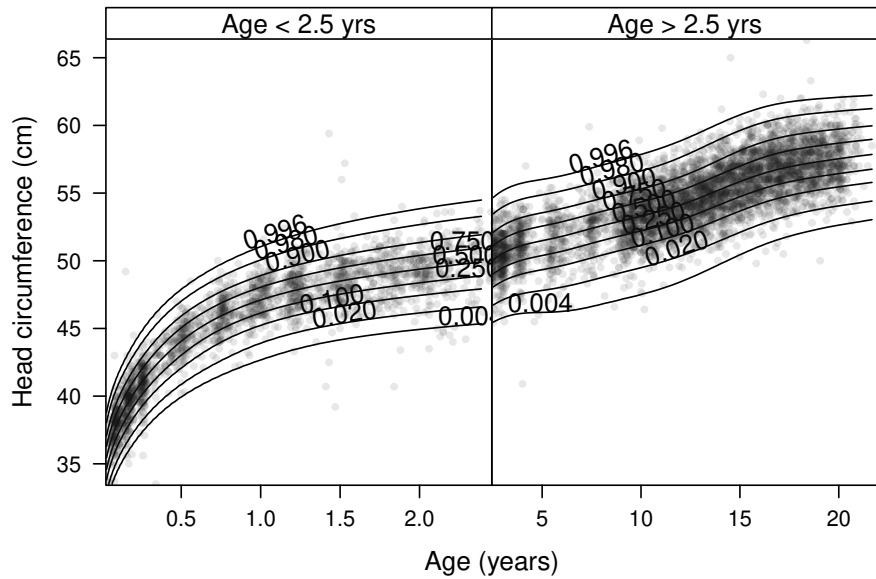
Figure 25: Head Circumference. Observed head circumference and age for 7040 boys with estimated quantile curves for $\tau = 0.04, 0.02, 0.1, 0.25, 0.5, 0.75, 0.9, 0.98, 0.996$.

DGP	$\Phi^{-1}(\mathbb{P}(Y \leq y \mid \text{Group } 1, x))$	$\Phi^{-1}(\mathbb{P}(Y \leq y \mid \text{Group } 2, x))$
Linear $\beta(\boldsymbol{x})$	$h_Y(y) - 2x$	$h_Y(y) + 2 - x$
Nonlinear $\beta(\boldsymbol{x})$	$h_Y(y) - 2g(x)$	$h_Y(y) + 2 - g(x)$
Linear $\boldsymbol{\vartheta}(\boldsymbol{x})$	$h_Y(y) - \beta_1(y) - \beta_2(y)x$	$h_Y(y) + \beta_1(y) - (\beta_2(y) + \beta_3(y))x$
Nonlinear $\boldsymbol{\vartheta}(\boldsymbol{x})$	$h_Y(y) - \beta_1(y) - \beta_2(y)g(x)$	$h_Y(y) + \beta_1(y) - (\beta_2(y) + \beta_3(y))g(x)$

Table 1: Artificial Data Generating Processes (DGPs). Description of four simulation models. $g(x) = \sin(2\pi x)(1 + x)$, $h_Y, \beta_1, \beta_2, \beta_3$ are Bernstein polynomials of order six.

```
R> library("tram")
R> ### set-up RNG
R> set.seed(27031105)
R> ### order of Bernstein polynomials
R> order <- 6
R> ### support of distributions
R> sup <- c(-4, 6)
R> ### bounds (essentially for plots)
R> bds <- c(-8, 8)
R> ### shift effects: main effect of grp, x, and interaction effect
R> beta <- c(-2, 2, -1)
```

Group information and x along with $g(x)$ was generated via

```
R> ### two groups
R> grp <- gl(2, 1)
R> ### uniform predictor
```
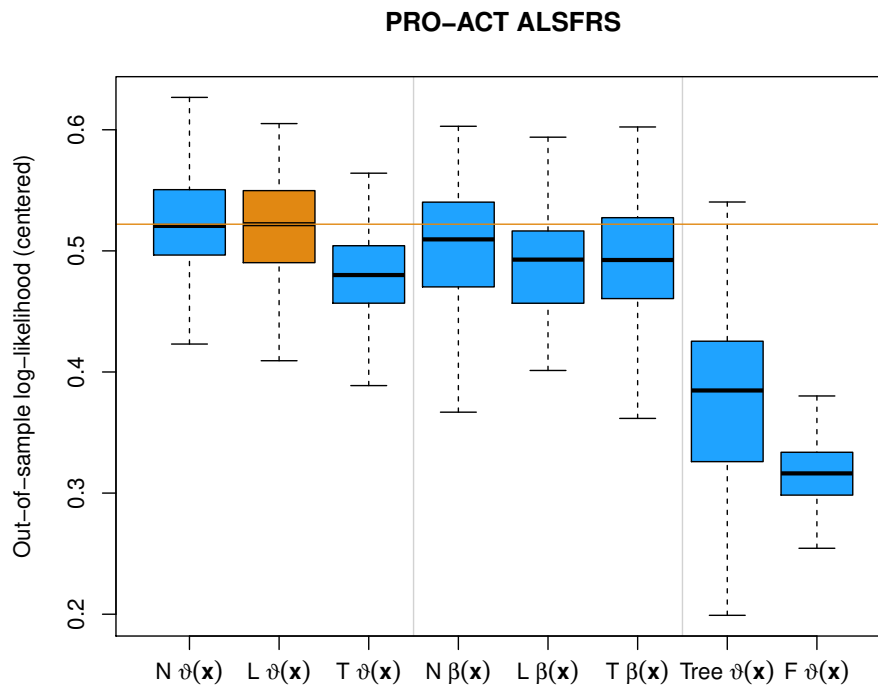
PRO–ACT ALSFRS

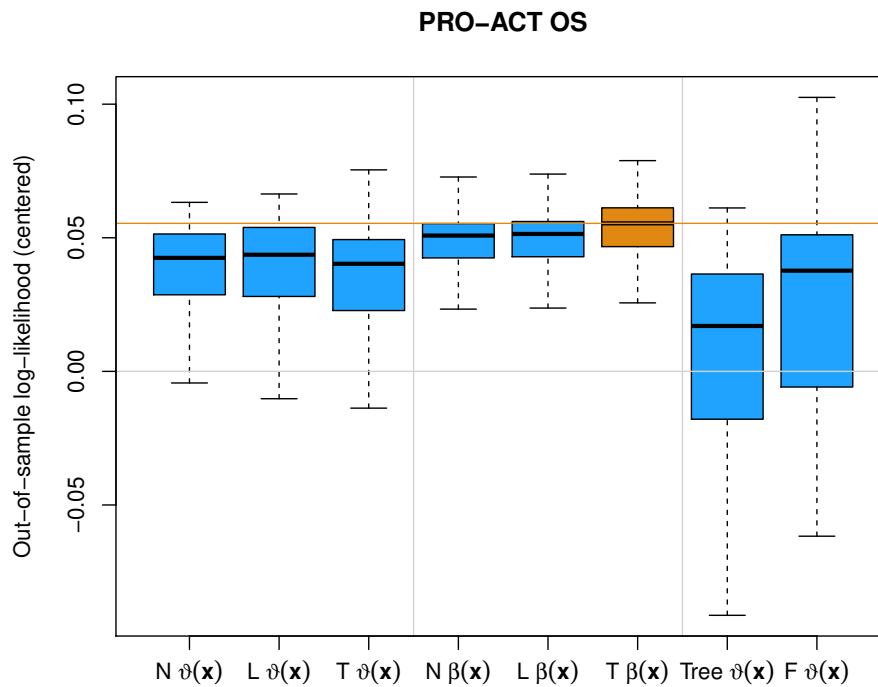Figure 26: PRO-ACT ALSFRS. Out-of-sample log-likelihood (centered by out-of-sample log-likelihood of unconditional model).

```
R> x <- seq(from = 0, to = 1, length.out = 1000)
R> d <- expand.grid(grp = grp, x = x)
R> ### transformation of x: sin(2 pi x) (1 + x)
R> d$g <- with(d, sin(x * 2 * pi) * (1 + x))
R> ### generate some response (this is NOT the
R> ### response used for the simulations)
R> X <- model.matrix(~ grp * x, data = d)[,-1]
R> d$y <- rnorm(nrow(d), mean = X %*% beta)
```

For the first model ("Linear $\beta(x)$"), the baseline transformation h_Y is a nonlinear function, parameterised as a Bernstein polynomial with the following coefficients

```
R> ### h_Y
R> (cf0 <- seq(from = sup[1], to = sup[2], length = order + 1) +
+          sin(seq(from = 0, to = 2*pi, length = order + 1)) *
+          (seq(from = 0, to = 2*pi, length = order + 1) <= pi) * 2)

[1] -4.0000000 -0.6012825  1.0653841  1.0000000  2.6666667  4.3333333
[7]  6.0000000

R> m1 <- BoxCox(y ~ grp * x, data = d, order = order,
```
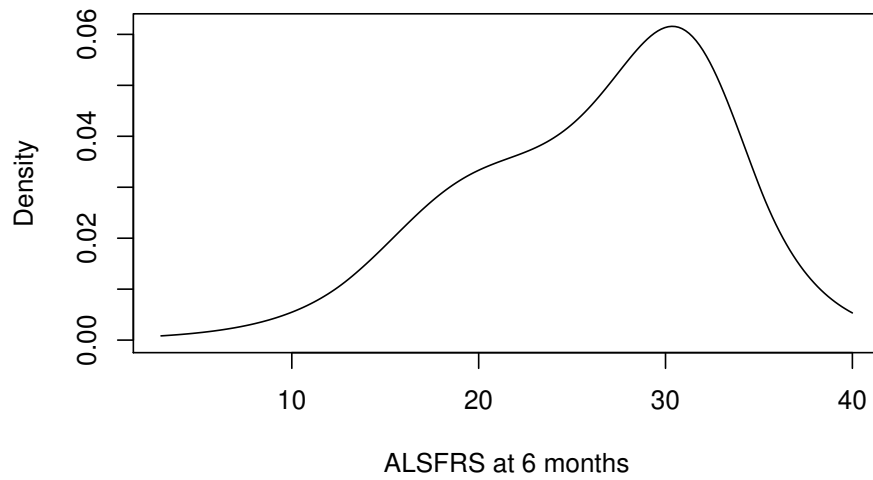
Figure 27: PRO-ACT ALSFRS. Unconditional response distribution.
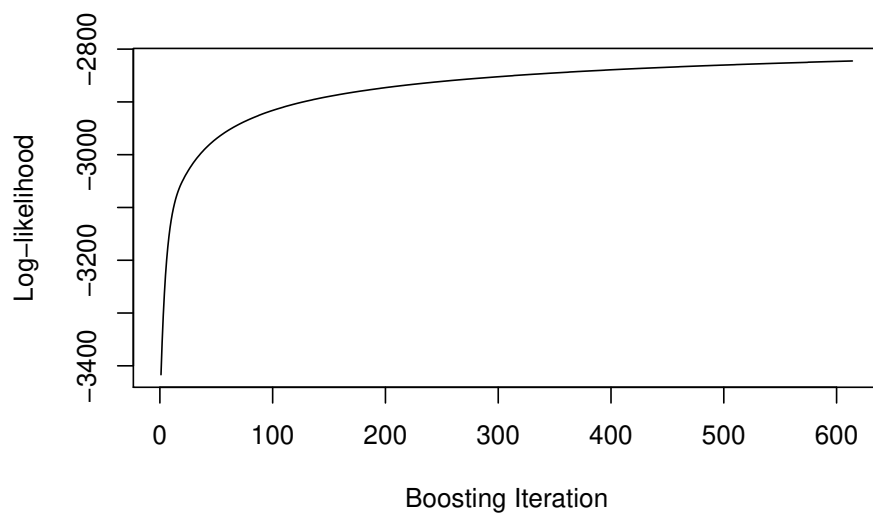
Figure 28: PRO-ACT ALSFRS. In-sample log-likelihood maximisation.
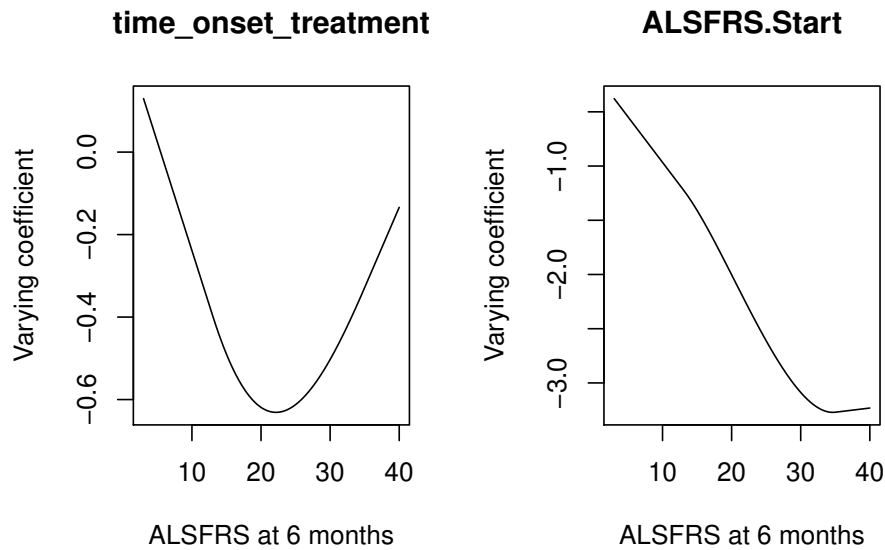
Figure 29: PRO-ACT ALSFRS. Response-varying coefficients.

```
+                model_only = TRUE, support = sup, bounds = bds)
R> cf <- coef(m1)
R> (cf[] <- c(cf0, beta))


 [1] -4.0000000 -0.6012825  1.0653841  1.0000000  2.6666667  4.3333333
 [7]  6.0000000 -2.0000000  2.0000000 -1.0000000


R> coef(m1) <- cf
```

The linear part simply consists of a main effect of group, a main effect of x and an interaction effect. This is a linear transformation model with nonnormal response.

The second model ("Nonlinear $\beta(\boldsymbol{x})$") is based on the same coefficients, however, after the transformation $g(x)$

```
R> m2 <- BoxCox(y ~ grp * g, data = d, order = order,
+                model_only = TRUE, support = sup, bounds = bds)
R> cf <- coef(m2)
R> cf[] <- c(cf0, beta)
R> coef(m2) <- cf
```

The third model ("Linear $\boldsymbol{\vartheta}(\boldsymbol{x})$") is a distribution regression model featuring response-varying coefficients:

```
R> m3 <- BoxCox(y | grp * x ~ 1, data = d, order = order,
+                model_only = TRUE, support = sup, bounds = bds)
R> cf <- coef(m3)
```

PRO–ACT OS

Figure 30: PRO-ACT OS. Out-of-sample log-likelihood (centered by out-of-sample log-likelihood of unconditional model).

```
R>
R> ### beta_1
R> (cf1 <- sin(seq(from = 0, to = pi / 2, length.out = order + 1)) * beta[1])

[1]  0.0000000 -0.5176381 -1.0000000 -1.4142136 -1.7320508 -1.9318517
[7] -2.0000000

R> ### beta_2
R> (cf2 <- sqrt(seq(from = 0, to = 2, length.out = order + 1)) / sqrt(2) * beta[2])

[1] 0.0000000 0.8164966 1.1547005 1.4142136 1.6329932 1.8257419 2.0000000

R> ### beta_3
R> (cf21 <- sin(seq(from = 0, to = pi / 2, length.out = order + 1)) * beta[3])

[1]  0.0000000 -0.2588190 -0.5000000 -0.7071068 -0.8660254 -0.9659258
[7] -1.0000000

R> cf[] <- c(cf0, cf1, cf2, cf21)
R> coef(m3) <- cf
```
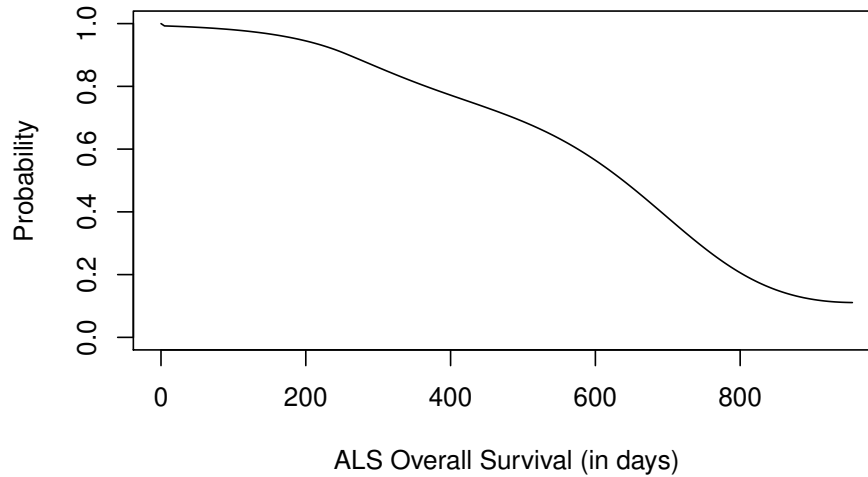
Figure 31: PRO-ACT OS. Unconditional response distribution.

In the last model ("Nonlinear $\vartheta(\boldsymbol{x})$"), the same response-varying coefficients are used after the transformation $g(x)$:

```
R> m4 <- BoxCox(y | grp * g ~ 1, data = d, order = order,
+               model_only = TRUE, support = sup, bounds = bds)
R> cf <- coef(m4)
R>
R> cf[] <- c(cf0, cf1, cf2, cf21)
R> coef(m4) <- cf
```

The conditional transformation functions are depicted in Figure 32.

Samples from the conditional distributions described by these four models were drawn as follows:

```
R> y1 <- simulate(m1, newdata = d, n = 100)
R> y2 <- simulate(m2, newdata = d, n = 100)
R> y3 <- simulate(m3, newdata = d, n = 100)
R> y4 <- simulate(m4, newdata = d, n = 100)
```

For each combination of data generating process, sample size ($N = 75, 150, 300$), number of additional uninformative predictor variables ($J^+ = 0, 5, 25$), two choices of F_Z (standard normal and standard logistic in the specification of the boosting procedure, the data were always generated using $F_Z = \Phi$), and two choices of the order of the Bernstein polynomials (six and 12, also only for the specification of the boosting procedures), the out-of-sample log-likelihood was estimated for 100 simulation runs. Very extreme outliers (centered log-likelihoods smaller than -10^5) were not drawn.
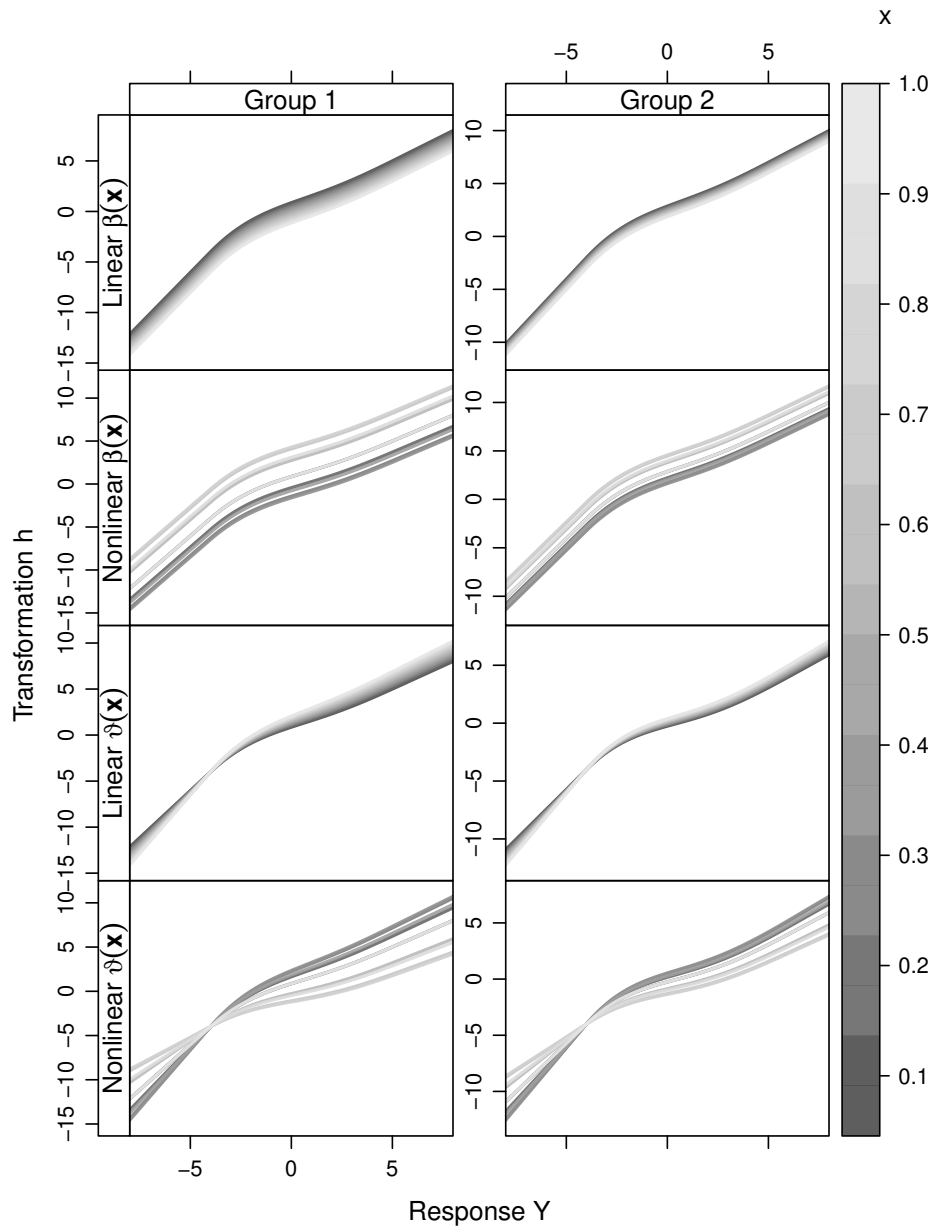
Figure 32: Artificial Data Generating Processes. Conditional transformation functions $h(y \mid \text{Group}, x)$ given two groups (left and right panel) and $x \in [0, 1]$ (grey color coding) for four different types of transformation models.

3.2. Results

The following figures depict the out-of-sample log-likelihoods (centered by the out-of-sample log-likelihood of the true model) for each combination of distribution function F_Z, number of additional noninformative uniform predictors J^+, and order M of the Bernstein polynomial. The panels correspond to the different DGPs (columns) and sample sizes (rows). Abbreviations of the boosting procedures and basis functions used are given at the x-axis. The boxplot of the best performing model is highlighted in yellow.
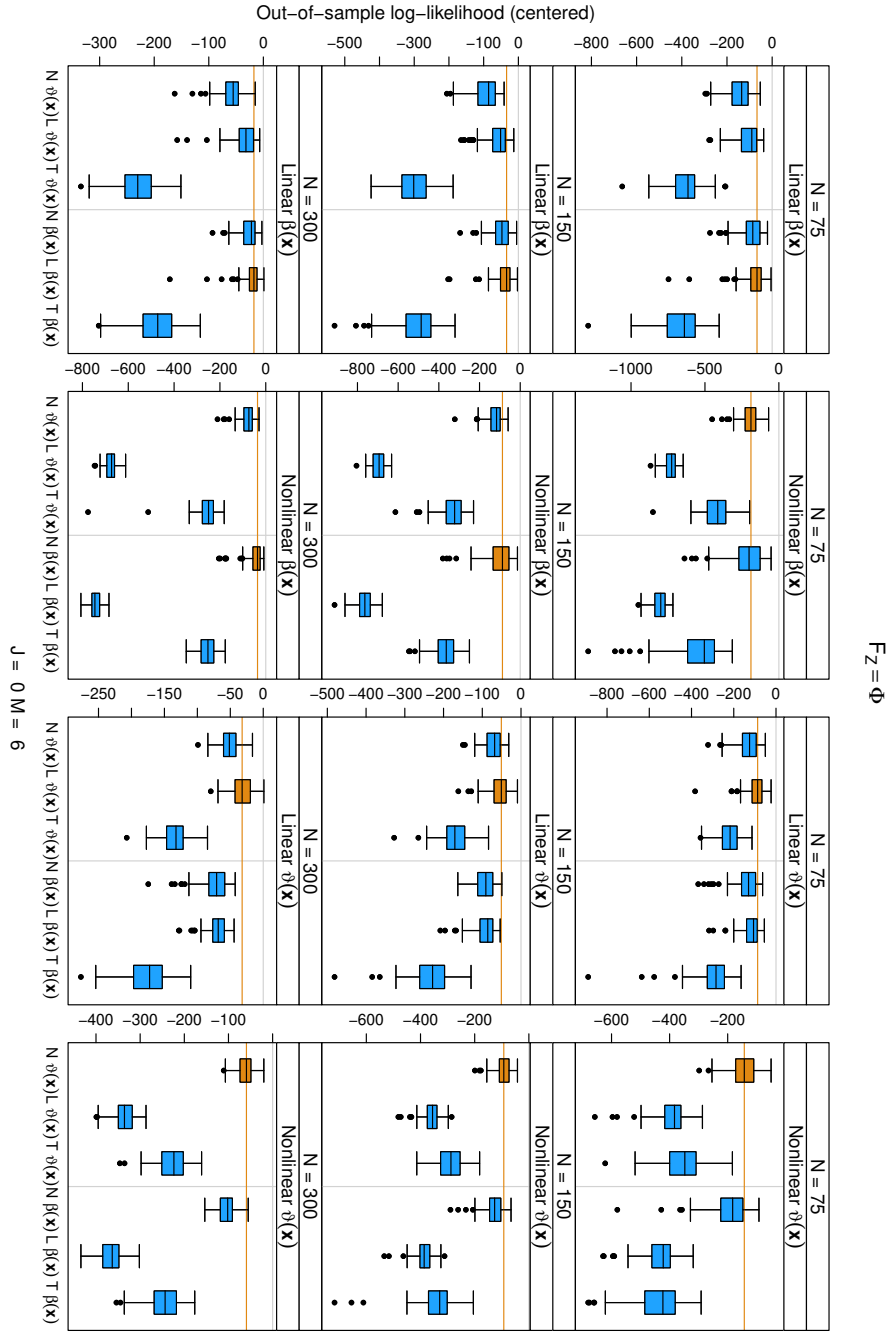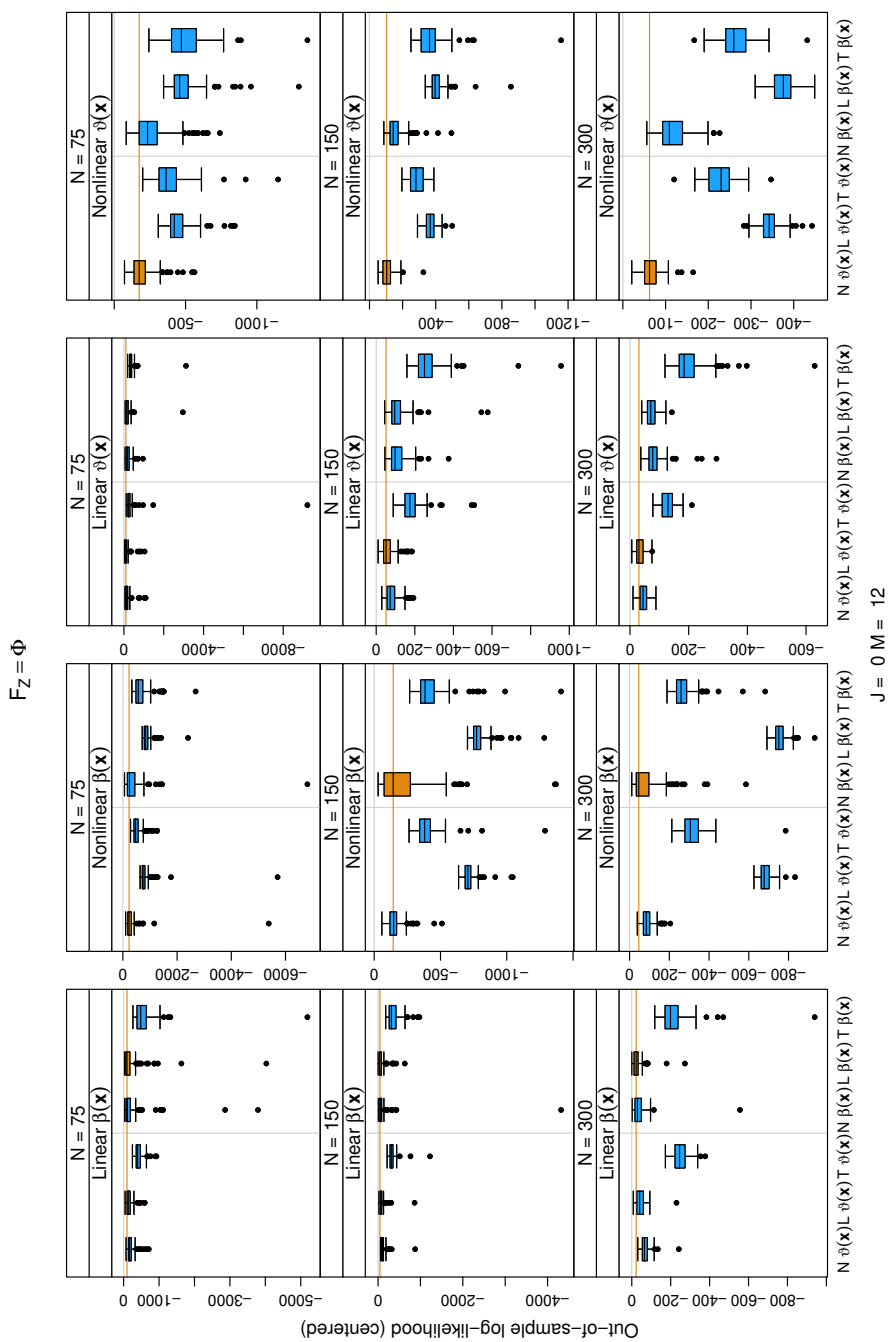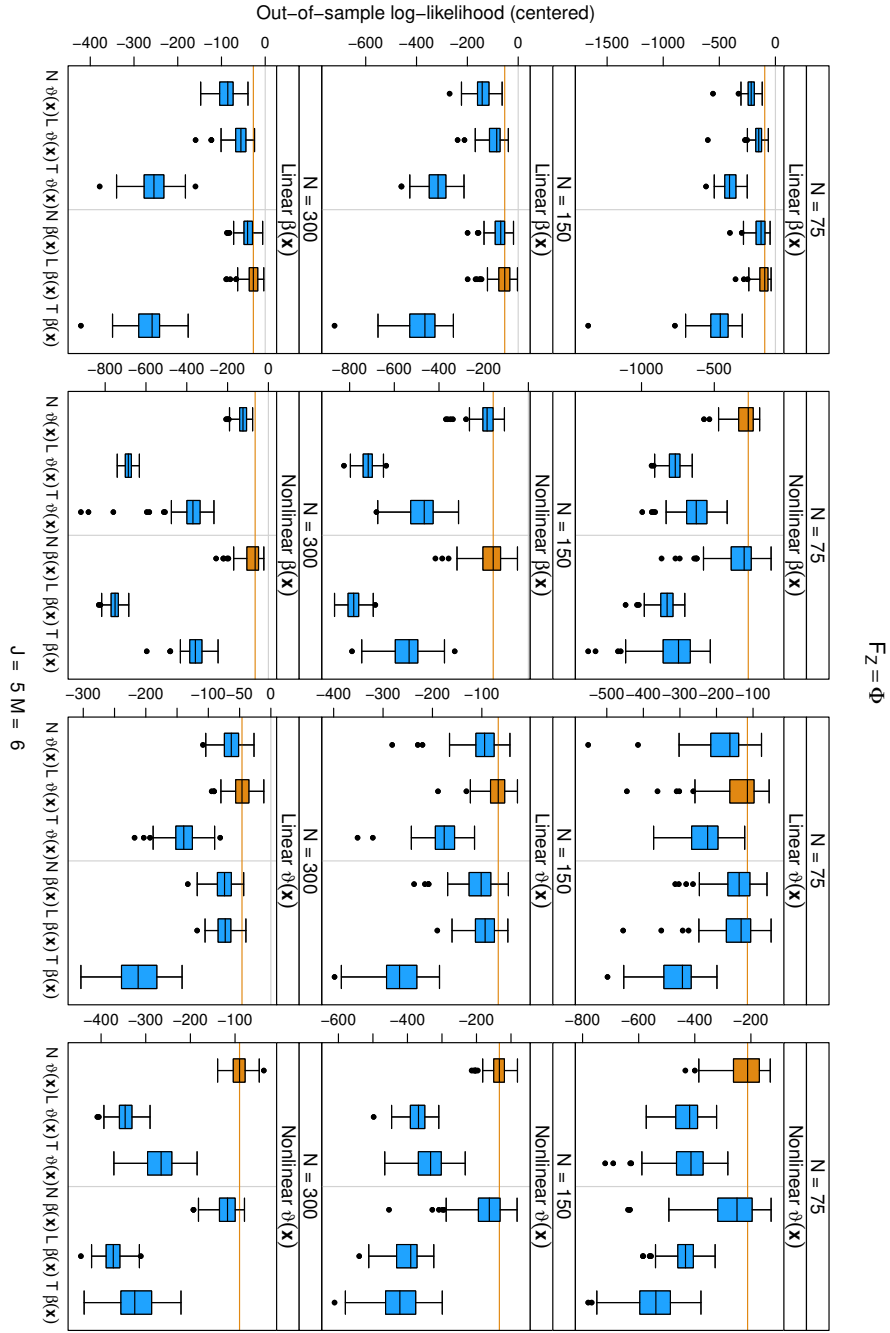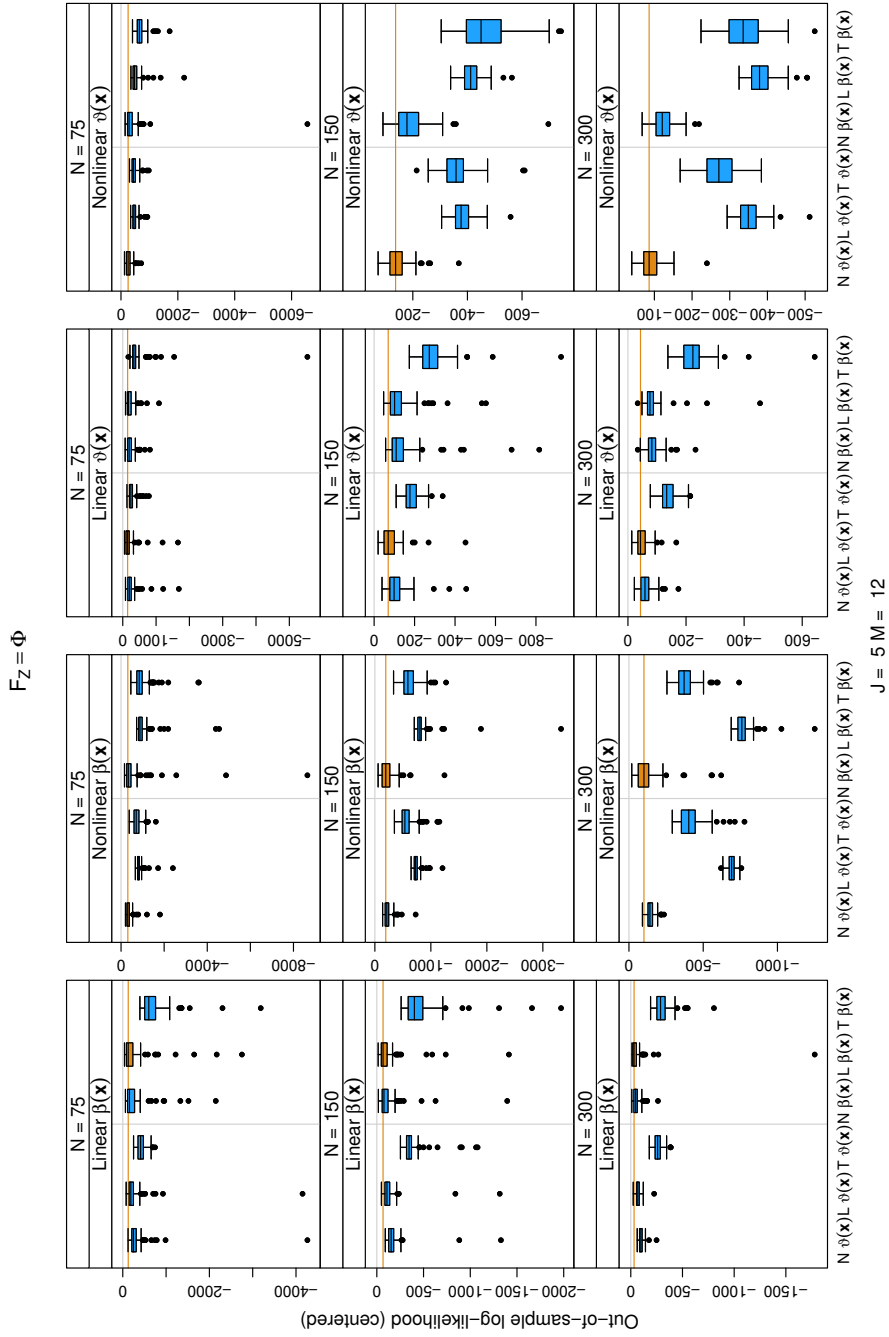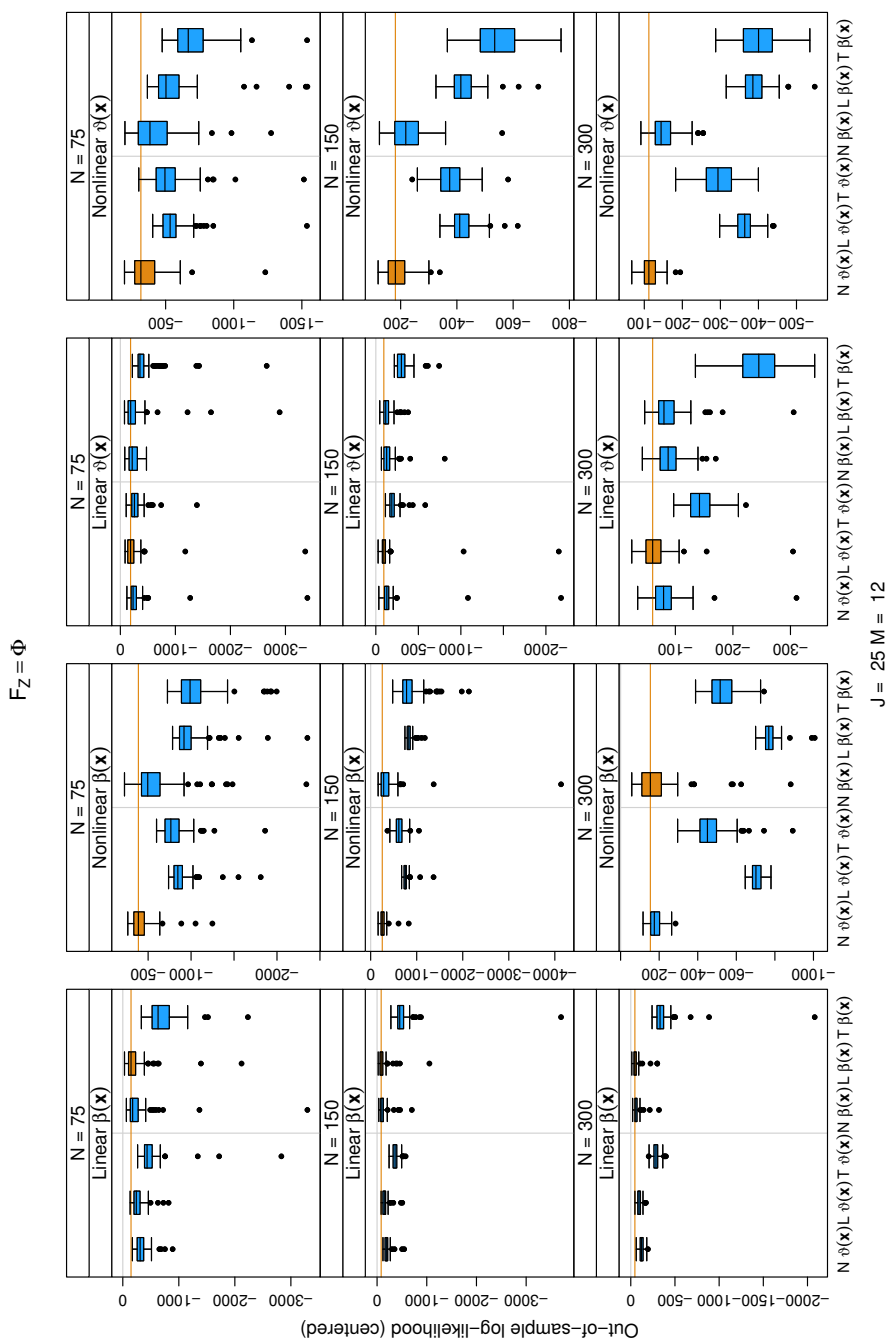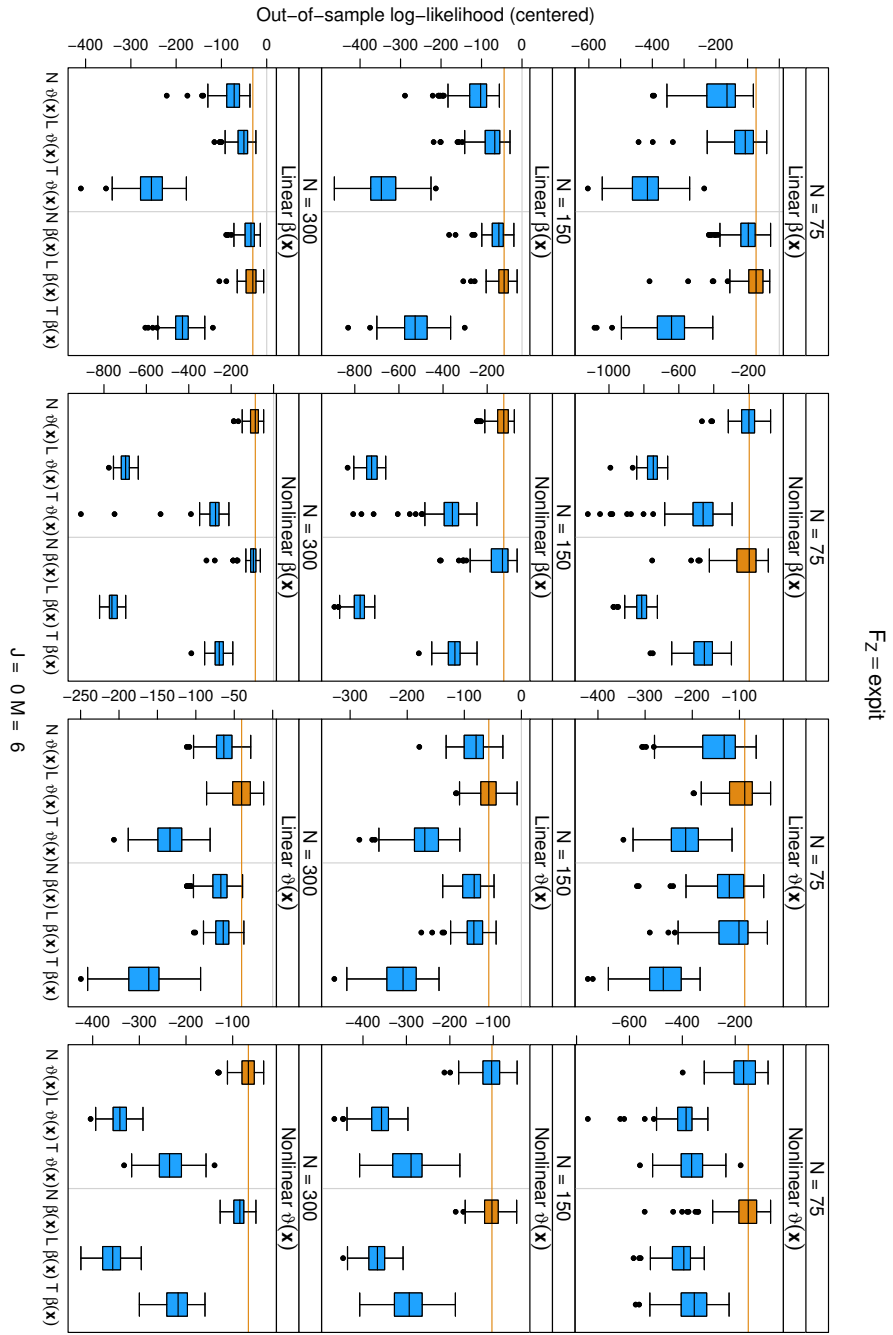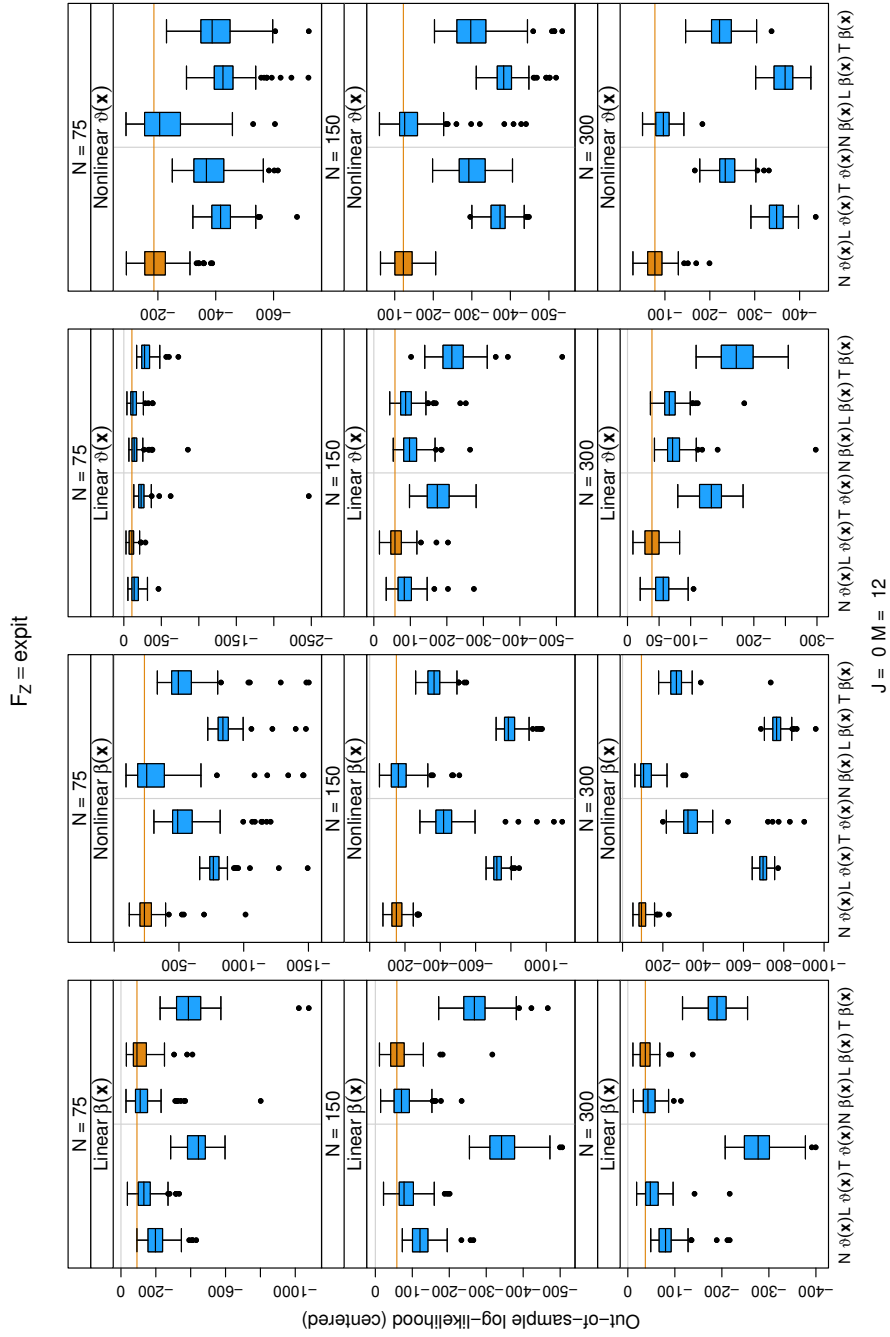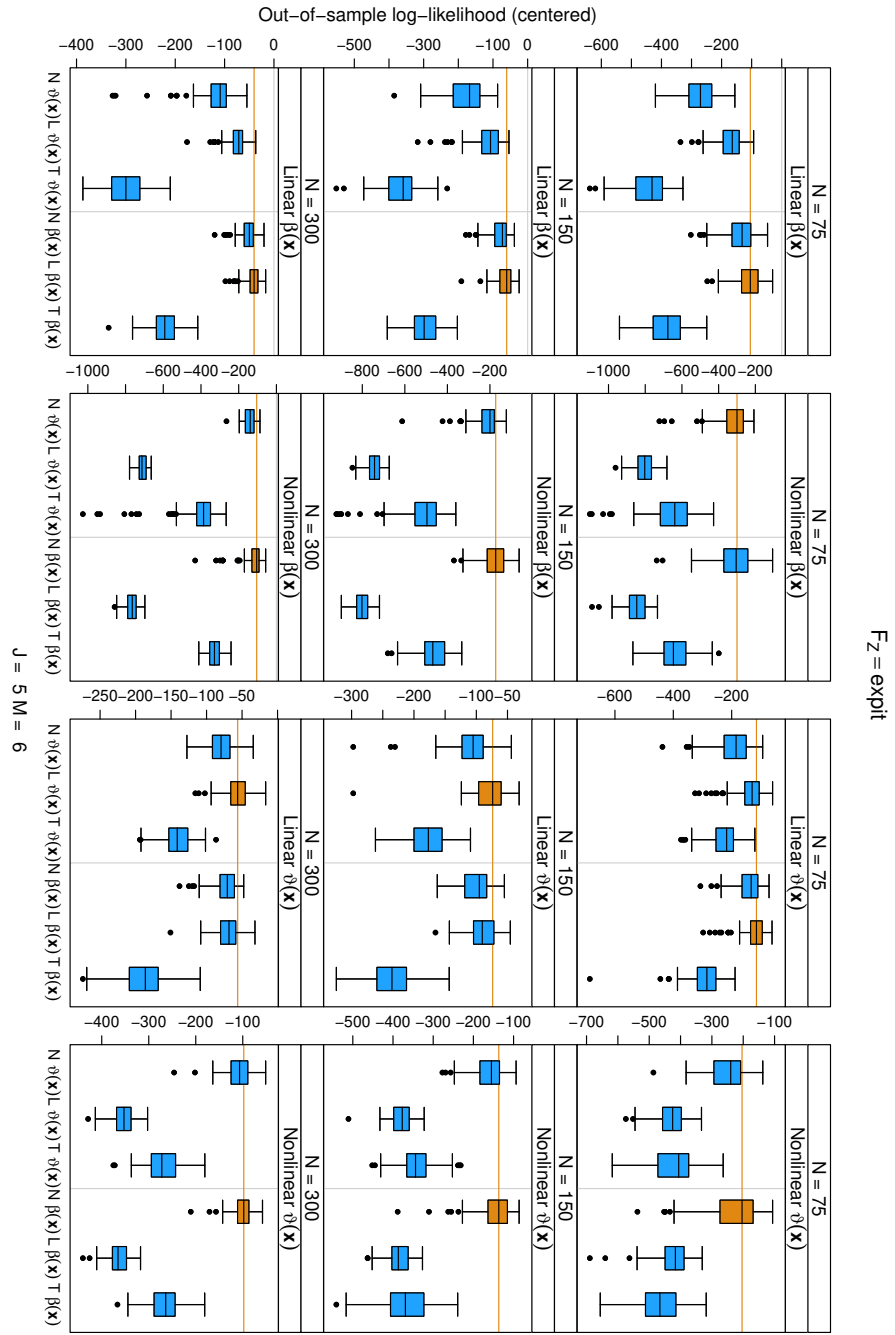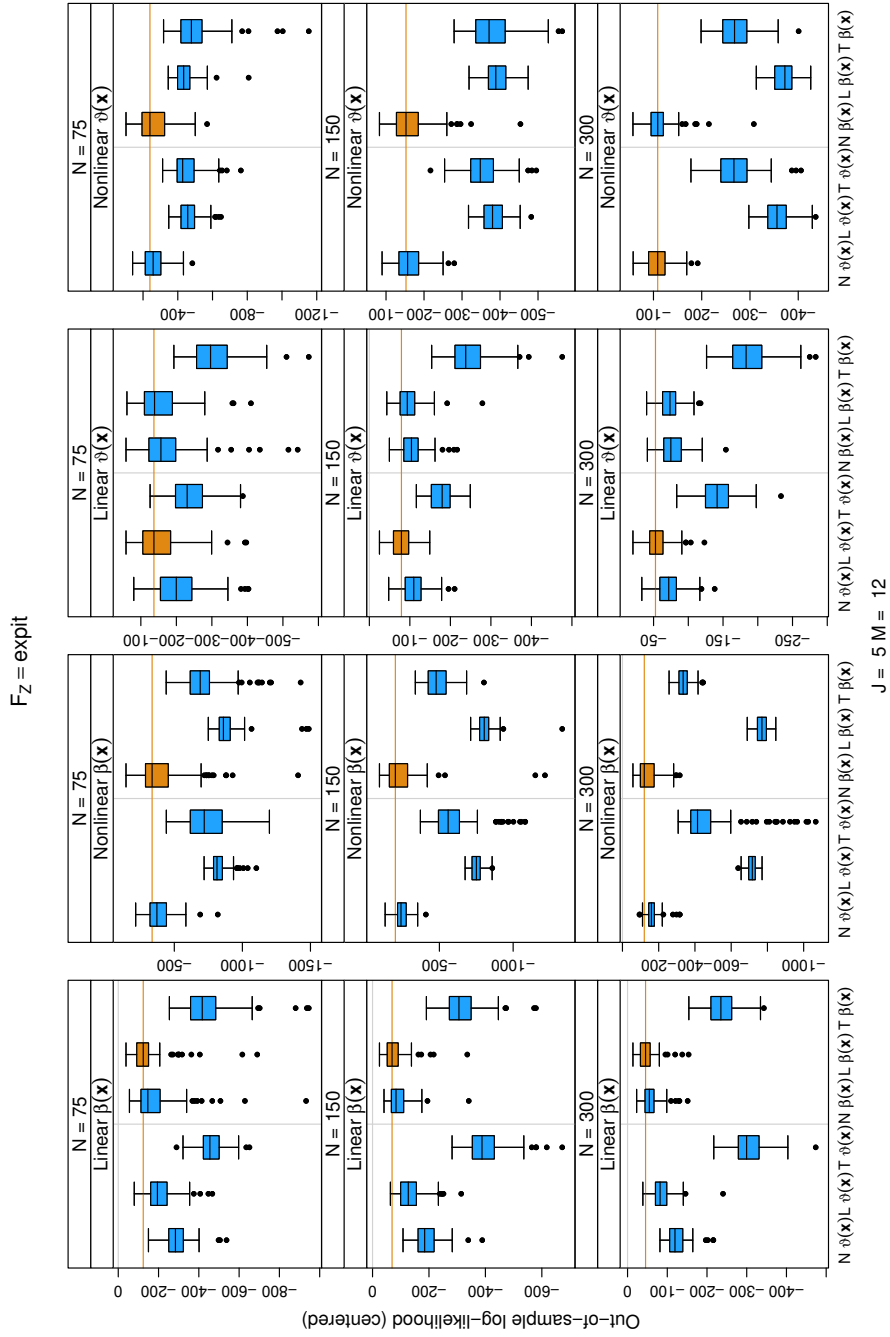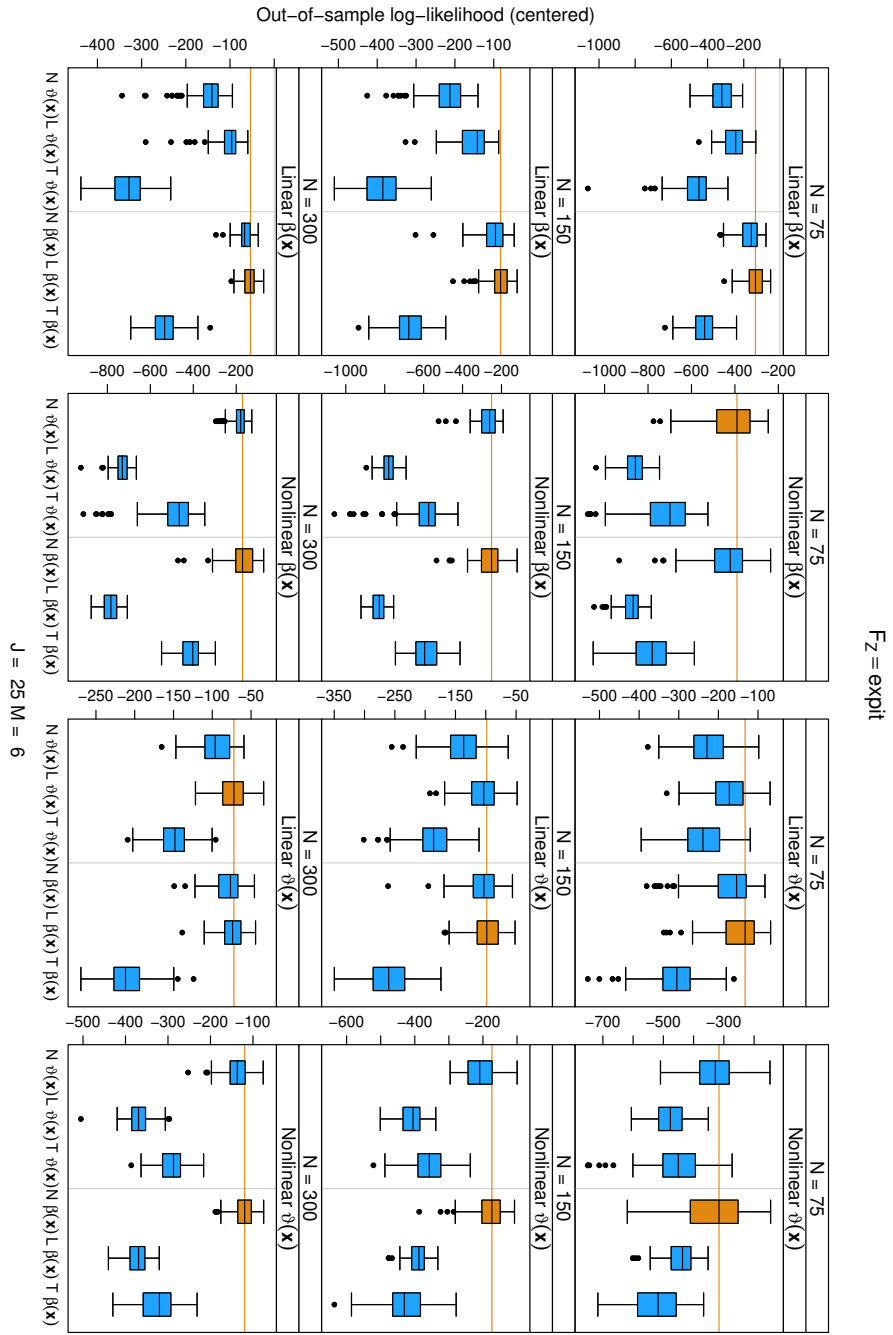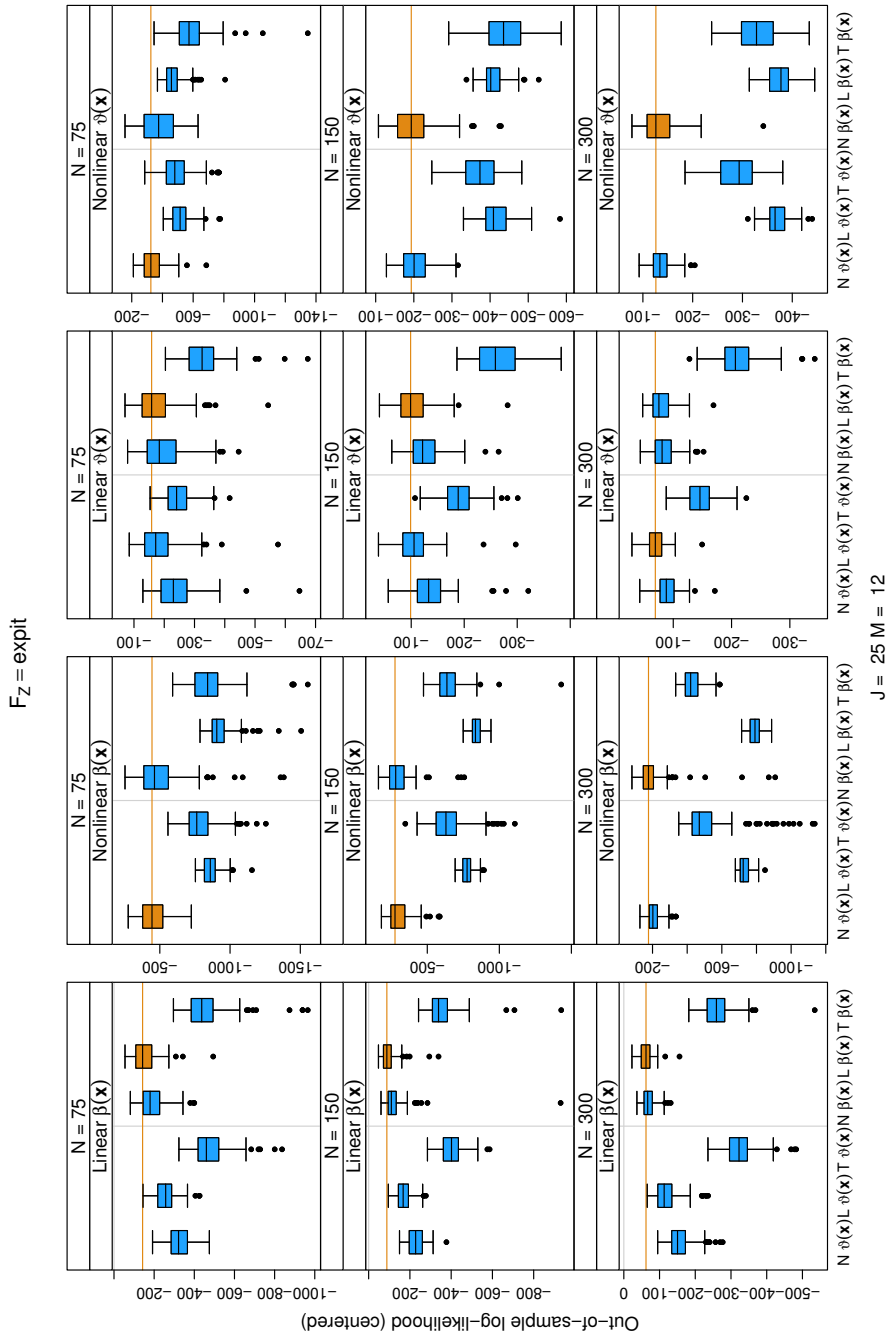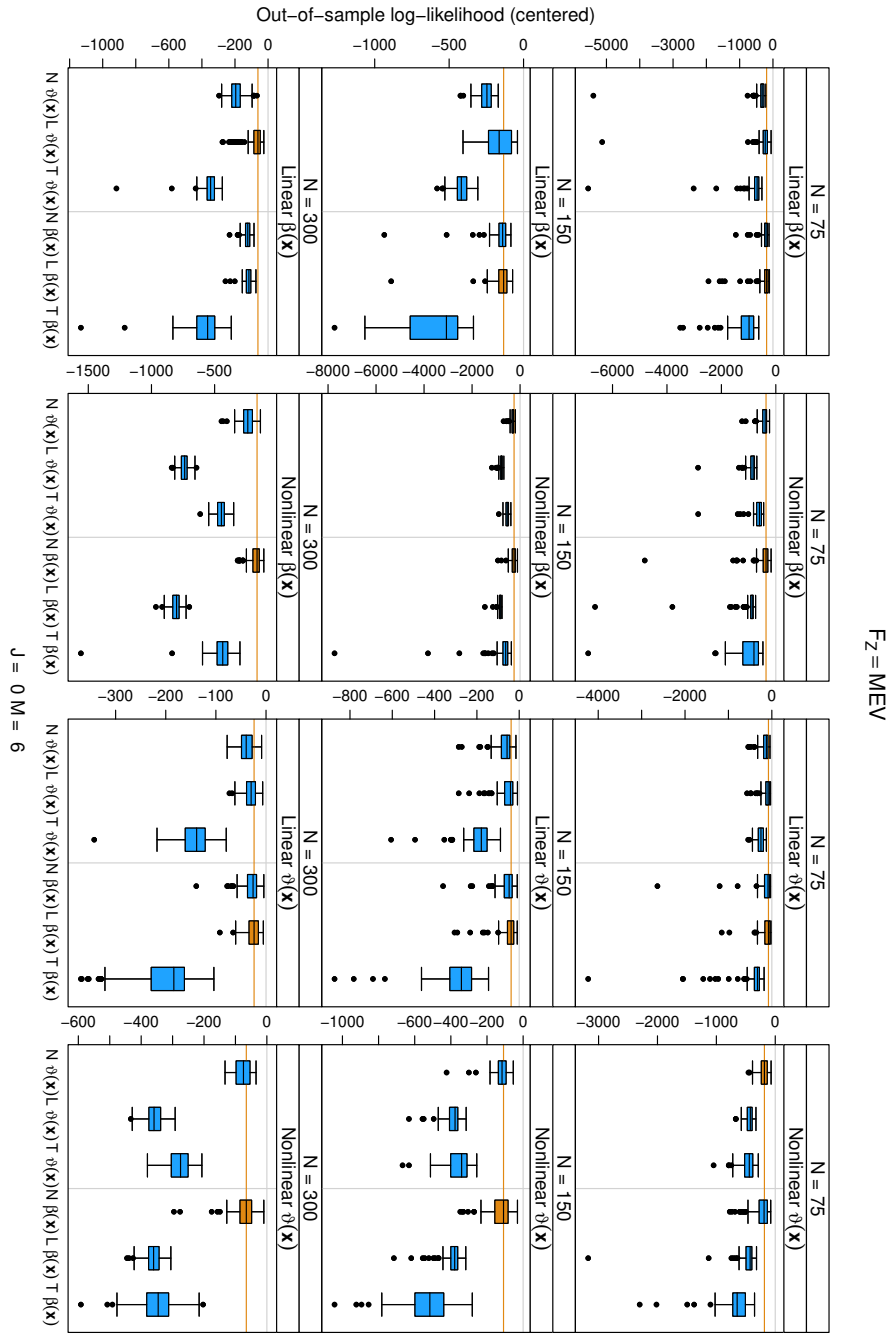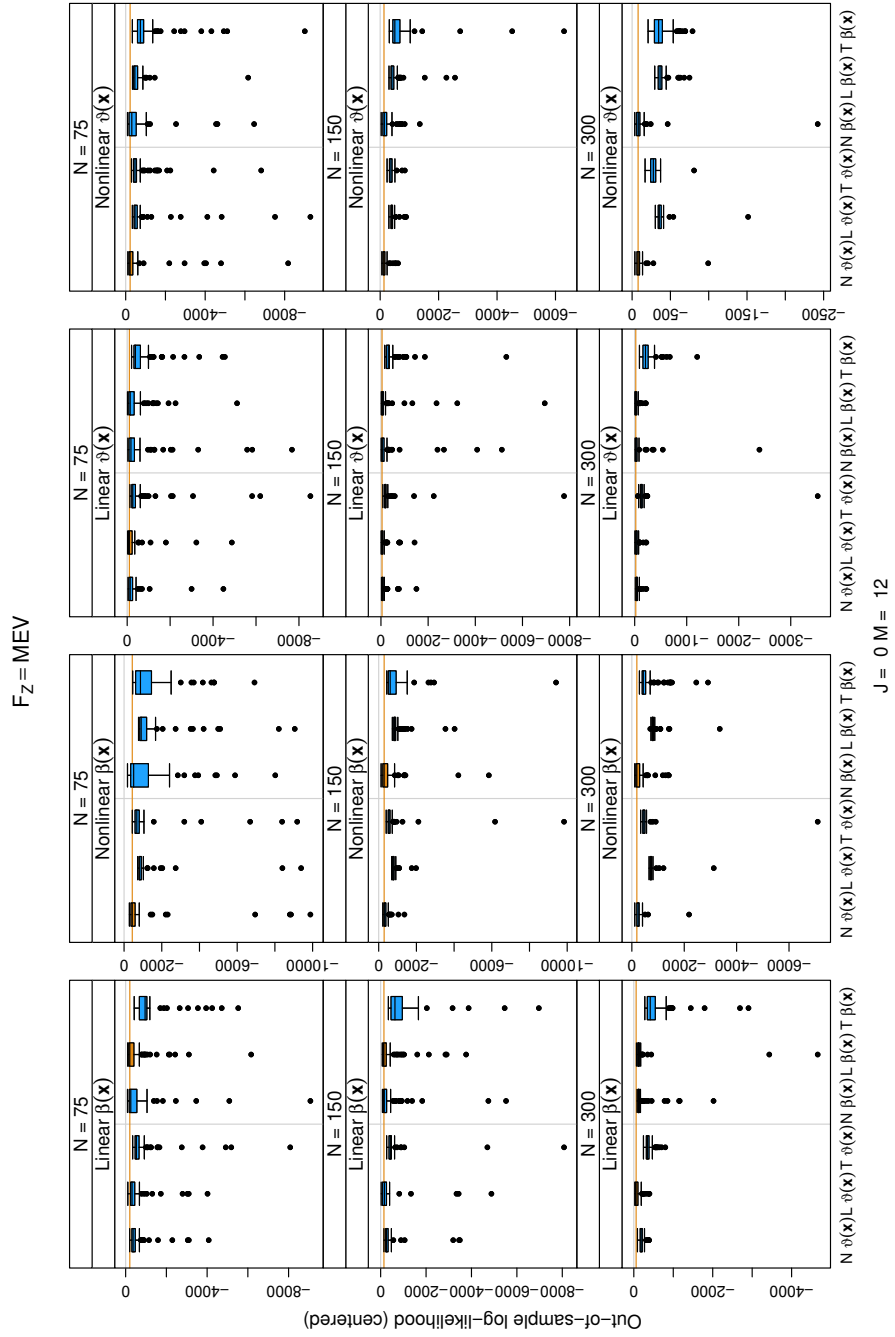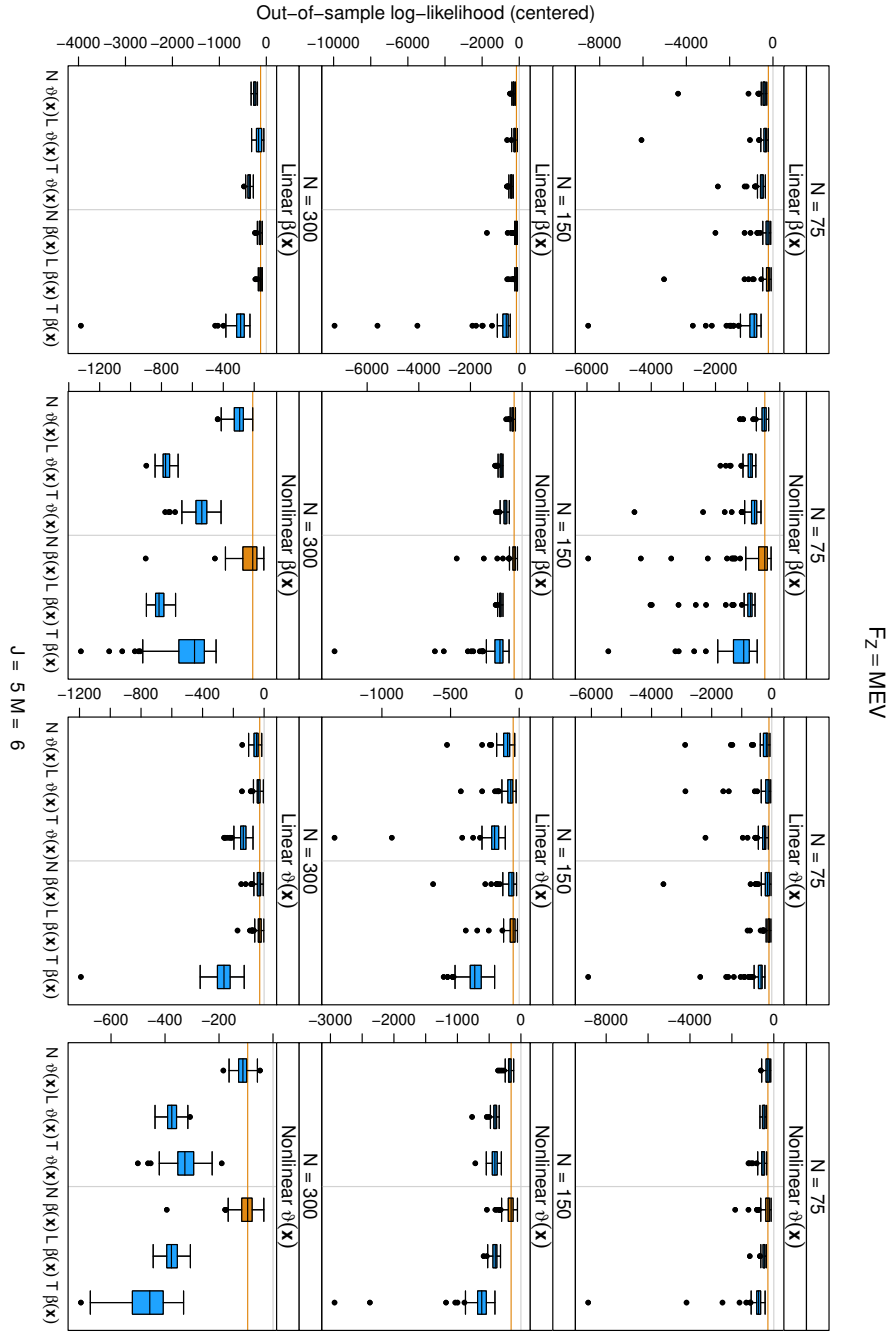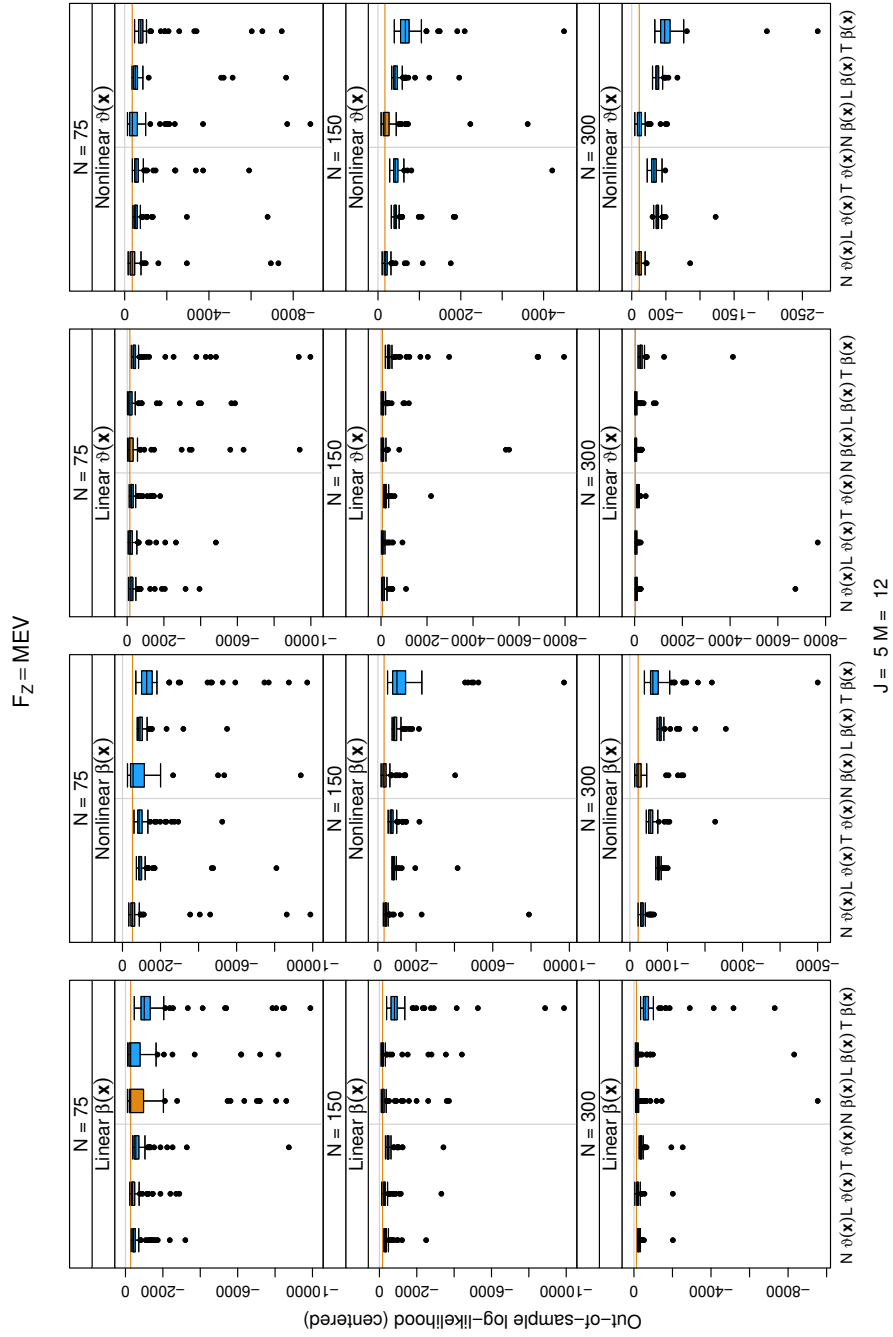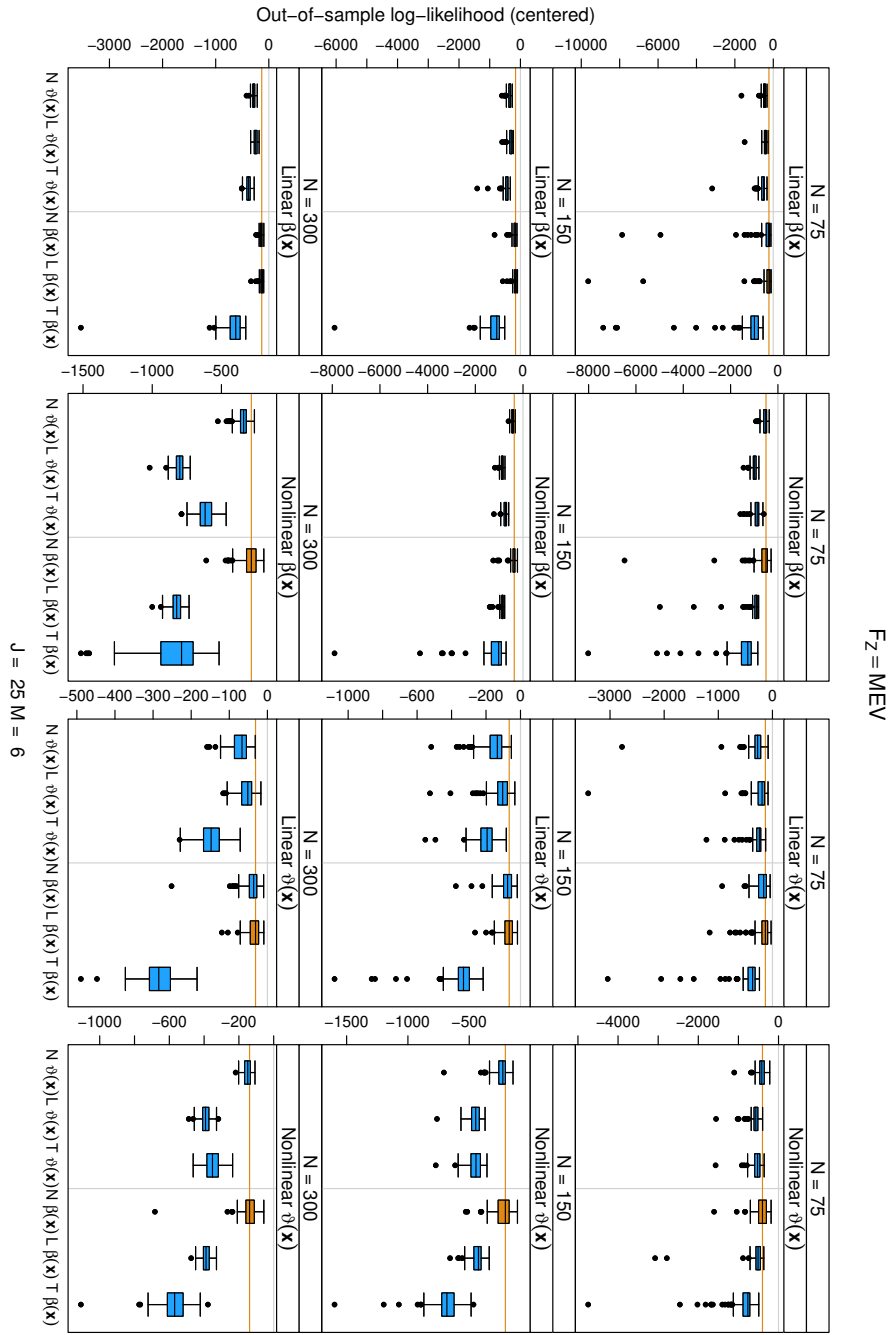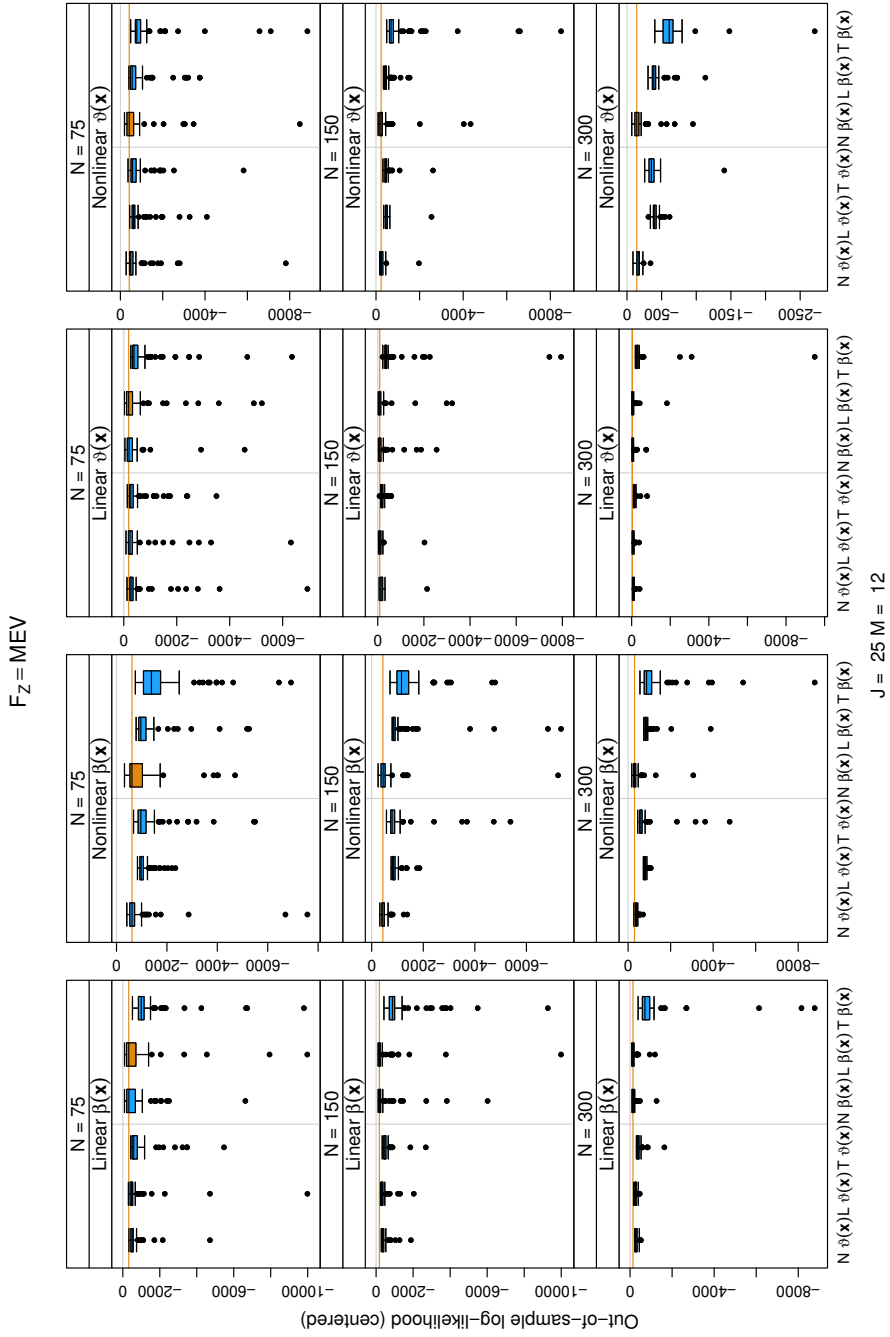
References

Fenske N, Kneib T, Hothorn T (2011). "Identifying Risk Factors for Severe Childhood Malnutrition by Boosting Additive Quantile Regression." *Journal of the American Statistical Association*, **106**(494), 494–510. `doi:10.1198/jasa.2011.ap09272`.

Fredriks AM, van Buuren S, Burgmeijer RJF, Meulmeester JF, Beuker RJ, Brugman E, Roede MJ, Verloove-Vanhorick SP, Wit J (2000). "Continuing Positive Secular Growth Change in The Netherlands 1955–1997." *Pediatric Research*, **47**(3), 316–323.

Garcia AL, Wagner K, Hothorn T, Koebnick C, Zunft HJF, Trippo U (2005). "Improved Prediction of Body Fat by Measuring Skinfold Thickness, Circumferences, and Bone Breadths." *Obesity*, **13**(3), 626–634. `doi:10.1038/oby.2005.67`.

Hothorn T (2019a). ***trtf**: Transformation Trees and Forests.* R package version 0.3-6, URL `https://CRAN.R-project.org/package=trtf`.

Hothorn T (2019b). *tbm: Transformation Boosting Machines.* R package version 0.3-2.1, URL `https://r-forge.r-project.org/projects/ctm/`.

Küffner R, Zach N, Norel R, Hawe J, Schoenfeld D, Wang L, Li G, Fang L, Mackey L, Hardiman O, Cudkowicz M, Sherman A, Ertaylan G, Grosse-Wentrup M, Hothorn T, van Ligtenberg J, Macke JH, Meyer T, Schölkopf B, Tran L, Vaughan R, Stolovitzky G, Leitner ML (2015). "Crowdsourced Analysis of Clinical Trial Data to Predict Amyotrophic Lateral Sclerosis Progression." *Nature Biotechnology*, **33**, 51–57. `doi:10.1038/nbt.3051`.

Rödel C, Graeven U, Fietkau R, Hohenberger W, Hothorn T, Arnold D, Hofheinz RD, Ghadimi M, Wolff HA, Lang-Welzenbach M, Raab HR, Wittekind C, Ströbel P, Staib L, Wilhelm M, Grabenbauer GG, Hoffmanns H, Lindemann F, Schlenska-Lange A, Folprecht G, Sauer R, Torsten Liersch on behalf of the German Rectal Cancer Study Group (2015). "Oxaliplatin Added to Fluorouracil-based Preoperative Chemoradiotherapy and Postoperative Chemotherapy of Locally Advanced Rectal Cancer (the German CAO/ARO/AIO-04 study): Final Results of the Multicentre, Open-label, Randomised, Phase 3 Trial." *The Lancet Oncology*, **16**(8), 979–989. `doi:10.1016/S1470-2045(15)00159-X`.

Schild RL, Maringa M, Siemer J, Meurer B, Hart N, Goecke TW, Schmid M, Hothorn T, Hansmann ME (2008). "Weight Estimation by Three-Dimensional Ultrasound in the Small Fetus." *Ultrasound in Obstetrics & Gynecology*, **32**(2), 168–175. `doi:10.1002/uog.6111`.

Schmid M, Hothorn T, Maloney KO, Weller DE, Potapov S (2011). "Geoadditive Regression Modeling of Stream Biological Condition." *Environmental and Ecological Statistics*, **18**(4), 709–733. `doi:10.1007/s10651-010-0158-4`.

Seibold H, Zeileis A, Hothorn T (2017). "Individual Treatment Effect Prediction for ALS Patients." *Statistical Methods in Medical Research*. `doi:10.1177/0962280217693034`.

Seibold S, Brandl R, Schmidl J, Busse J, Thorn S, Hothorn T, Müller J (2015). "Extinction Risk Status of Saproxylic Beetles Reflects the Ecological Degradation of Forests in Europe." *Conservation Biology*, **29**(2), 382–390. `doi:10.1111/cobi.12427`.

Stasinopoulos DM, Rigby RA (2007). "Generalized Additive Models for Location Scale and Shape (GAMLSS) in R." *Journal of Statistical Software*, **23**(7), 1–46. URL http://www.jstatsoft.org/v23/i07.

Affiliation:

Torsten Hothorn
Institut für Epidemiologie, Biostatistik und Prävention
Universität Zürich
Hirschengraben 84, CH-8001 Zürich, Switzerland
Torsten.Hothorn@uzh.ch
http://tiny.uzh.ch/bh