Web Supplement to

A Balanced Statistical Boosting Approach for GAMLSS via New Step Lengths

published in the Journal *Computational Statistics*

Alexandra Daub*, Andreas Mayr, Boyao Zhang, Elisabeth Bergherr

Appendix B Simulation Results

B.1 Gaussian location and scale model

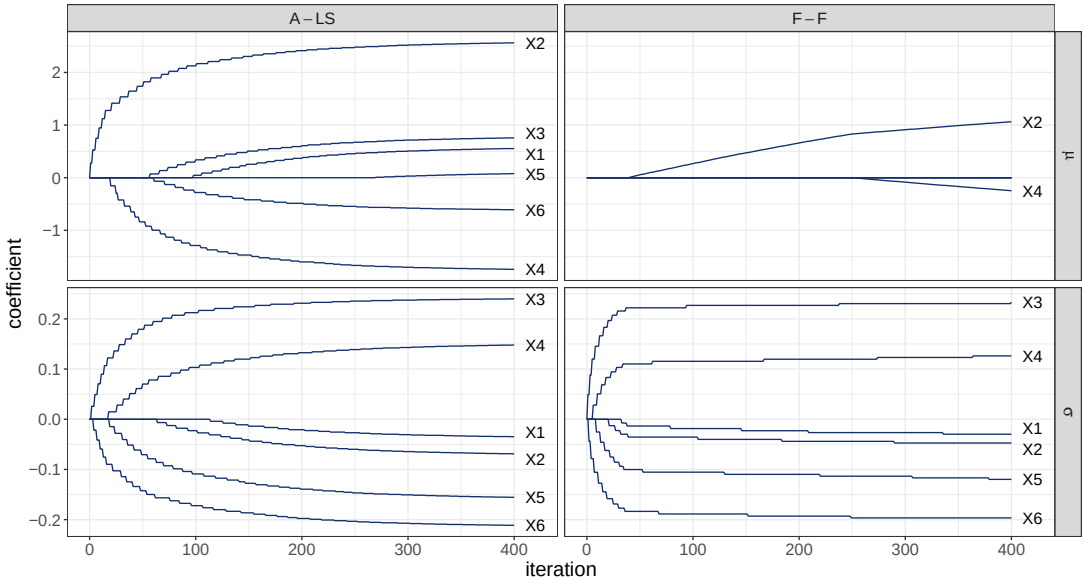

Fig. 9 Path plots for a boosting run in the Gaussian simulation setting (8) with shrunk optimal step lengths (left) and fixed step lengths (right). Further explanations on the abbreviations of the step length approaches can be found at the end of section 2

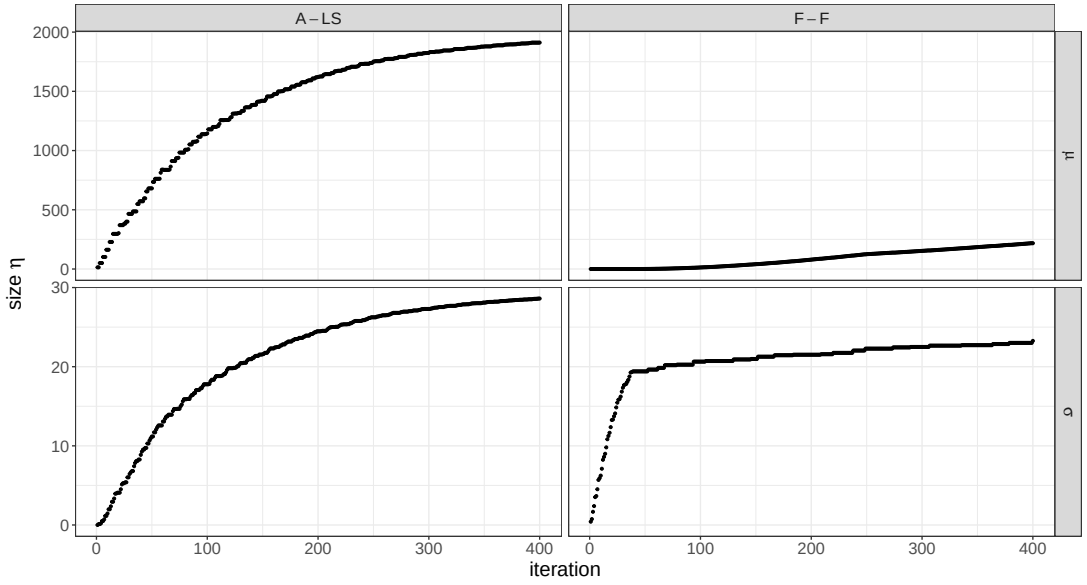

Fig. 10 Squared Euclidean norms of the predictors $\boldsymbol{\eta}_{\mu}^{[m]}$ and $\boldsymbol{\eta}_{\sigma}^{[m]}$ without the intercept for the same boosting run in the Gaussian simulation setting (8) with shrunk optimal step lengths (left) and fixed step lengths (right) as in Fig. 9. Further explanations on the abbreviations of the step length approaches can be found at the end of section 2

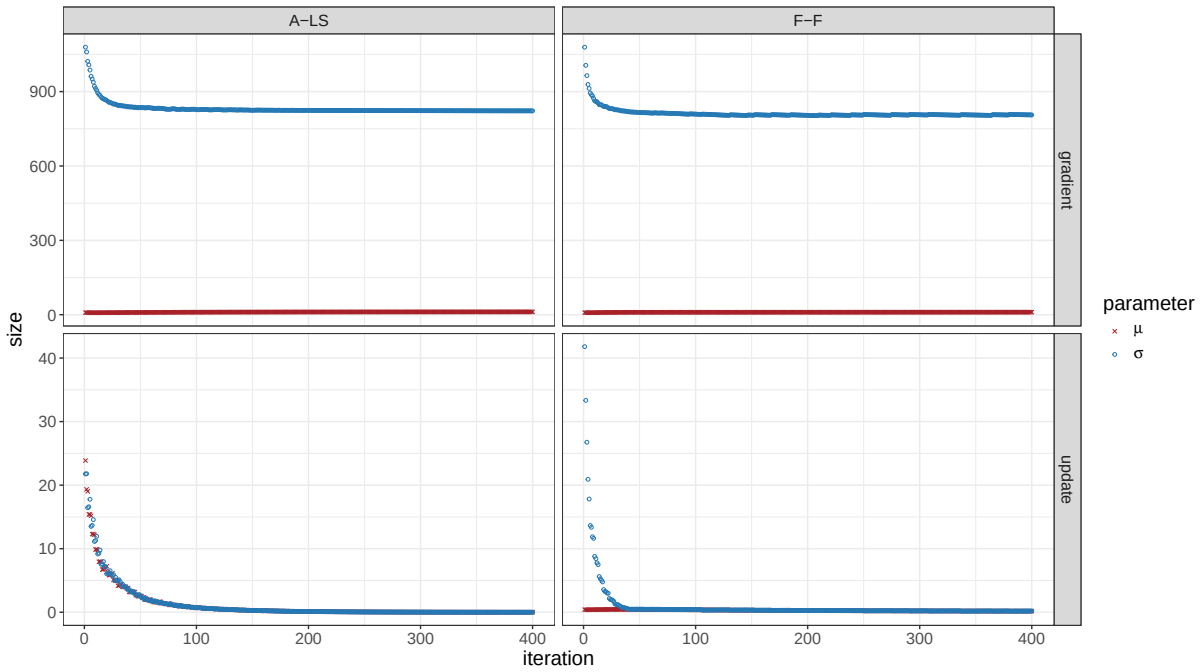

Fig. 11 Squared Euclidean norm of the negative gradients $\boldsymbol{u}_\mu^{[m]}$ and $\boldsymbol{u}_\sigma^{[m]}$ in iteration m evaluated at the current model (top) as well as update sizes $\zeta_\mu^{[m]}$ and $\zeta_\sigma^{[m]}$ as specified in (4) (bottom) for the same boosting run in the Gaussian simulation setting (8) with shrunk optimal step lengths (left) and fixed step lengths (right) as in Fig. 9. Further explanations on the abbreviations of the step length approaches can be found at the end of section 2

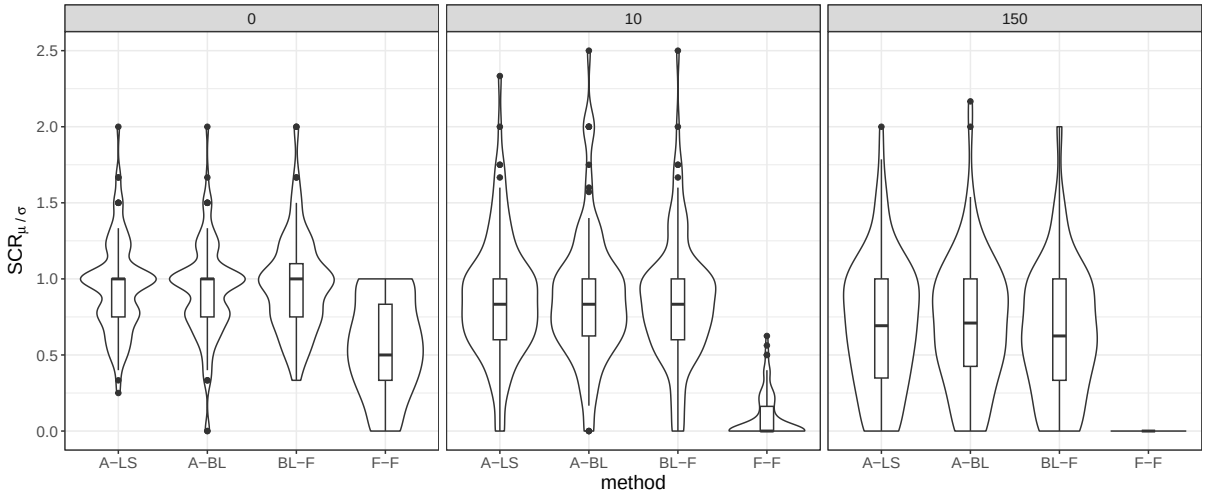

Fig. 12 Distribution of the ratio of numbers of covariates included in η_μ relative to η_σ ($\text{SCR}_{\mu/\sigma}$) in the Gaussian simulation setting (8) for a varying number of additional non-informative covariates (columns). 18 outliers that exceed an $\text{SCR}_{\mu/\sigma}$ of 2.5 are regarded in the distributions but not included in the plot for a better comprehensibility. The excluded outliers occured in 7 simulation runs, mainly with 150 additional non-informative covariates

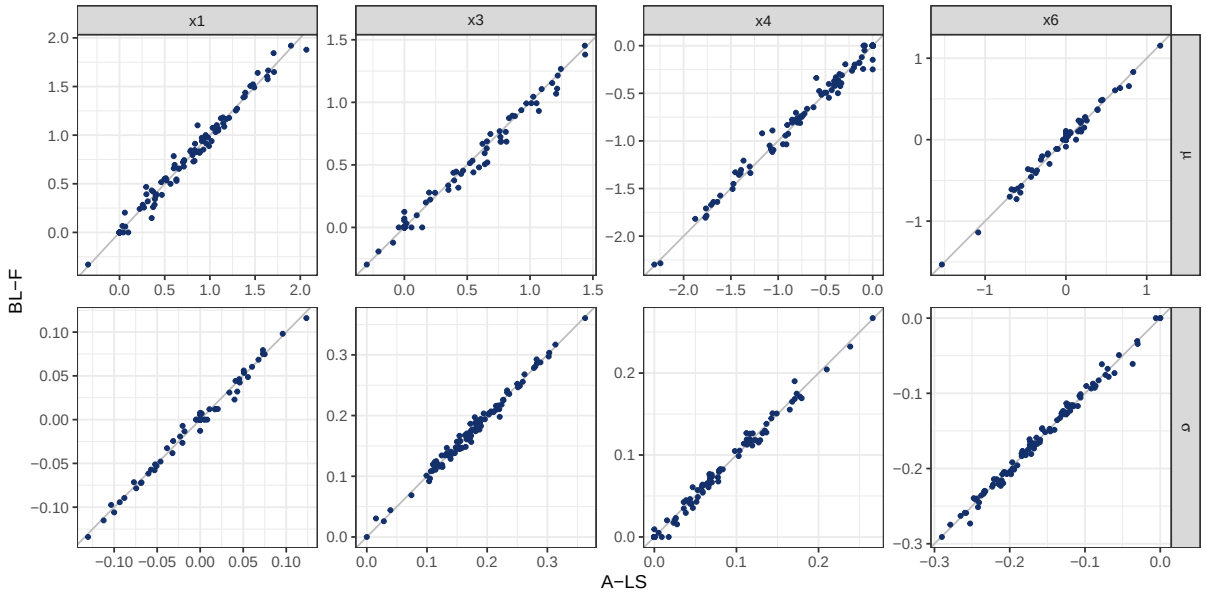

Fig. 13 Direct comparison of coefficient estimates of the two adaptive step length approaches A-LS and BL-F in the Gaussian simulation setting (8) without additional non-informative covariates

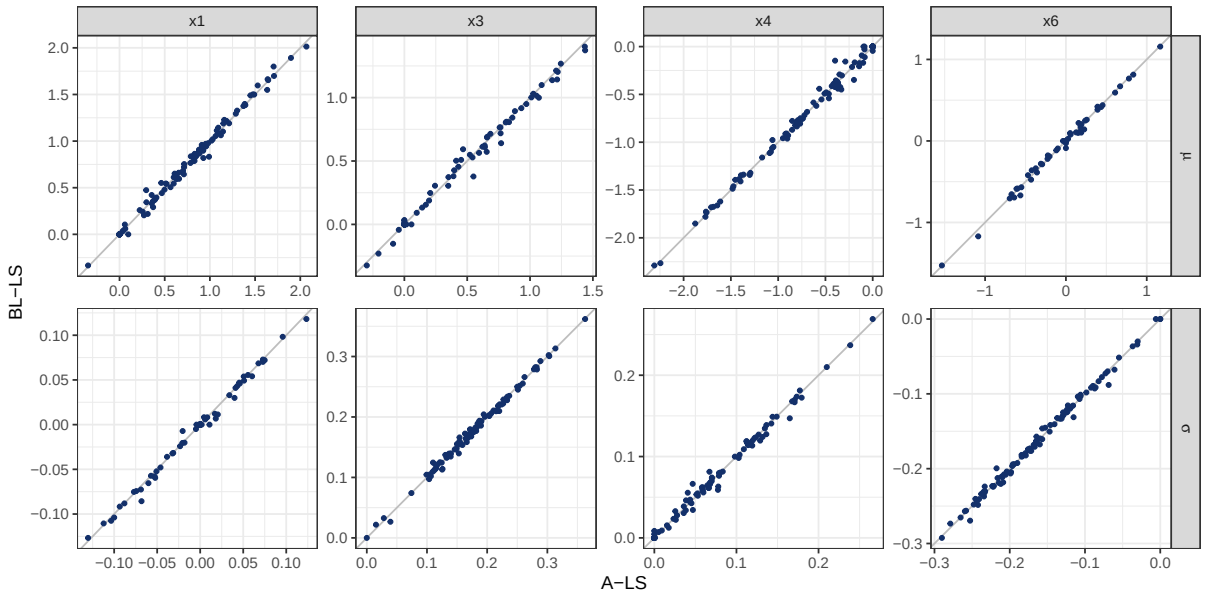

Fig. 14 Direct comparison of coefficient estimates of the two adaptive step length approaches A-LS and BL-LS in the Gaussian simulation setting (8) without additional non-informative covariates

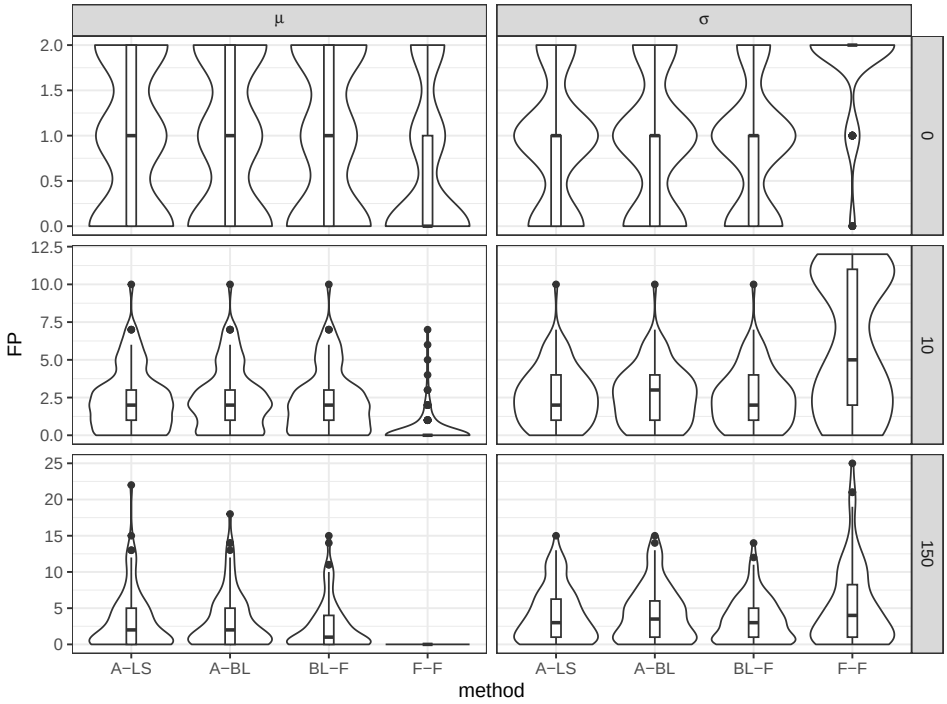

Fig. 15 Distribution of the number of false positives (FP) in the Gaussian simulation setting (8) for a varying number of additional non-informative covariates (rows)

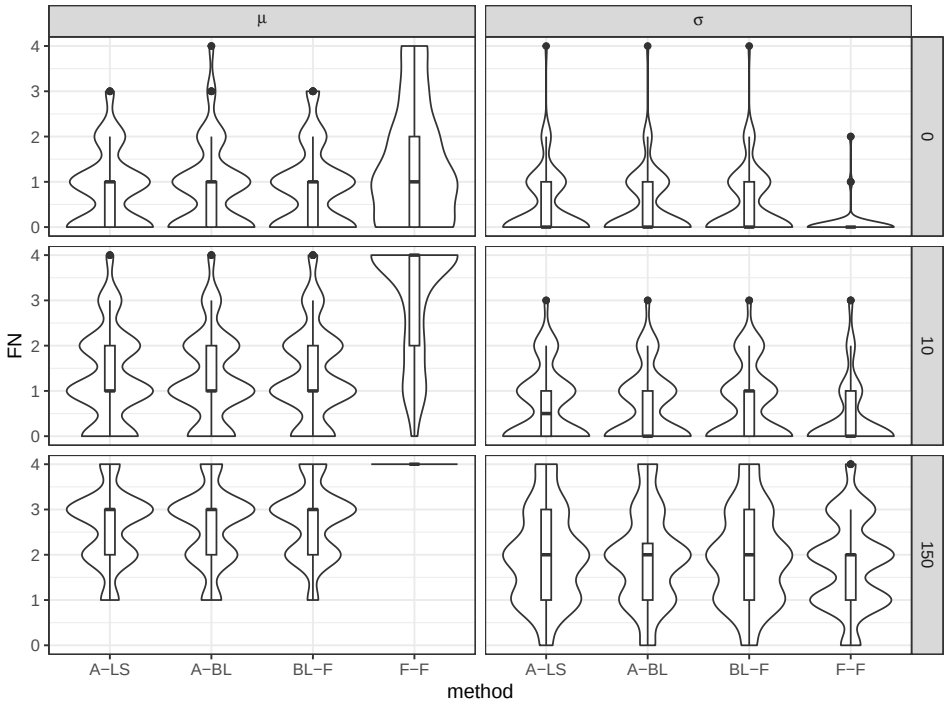

Fig. 16 Distribution of the number of false negatives (FN) in the Gaussian simulation setting (8) for a varying number of additional non-informative covariates (rows)

Table 5 Quartiles of the run times until m_{stop} without cross-validation in seconds and stopping iterations (in parenthesis) in the Gaussian simulation setting (8) with 10 (top) and 150 (bottom) additional non-informative covariates

non-inf		A-LS	A-BL	BL-F	F-F
10	1st Qu.	0.046 (48)	0.034 (51)	0.016 (22)	0.014 (22)
	Median	0.067 (72)	0.052 (79)	0.024 (38)	0.022 (38)
	3rd Qu.	0.092 (99)	0.069 (109)	0.033 (53)	0.425 (806)
150	1st Qu.	0.067 (11)	0.065 (11)	0.046 (6)	0.041 (4)
	Median	0.123 (26)	0.133 (30)	0.073 (13)	0.062 (10)
	3rd Qu.	0.196 (43)	0.196 (45)	0.105 (22)	0.093 (20)

B.2 Negative binomial location and scale model

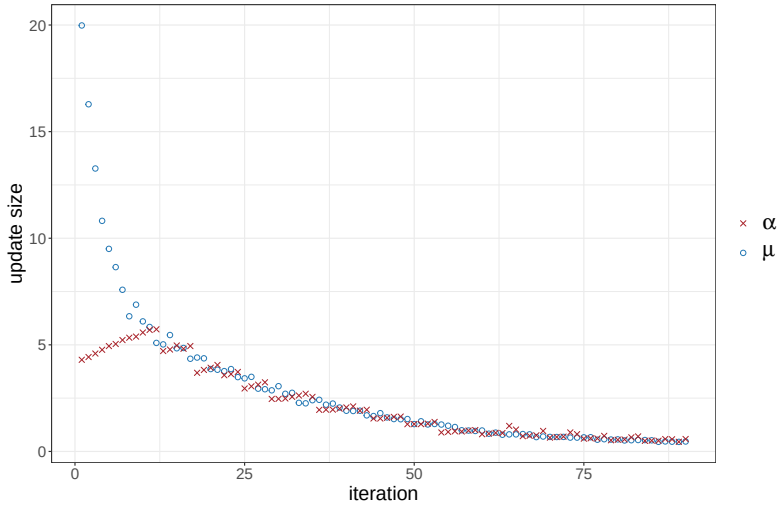

Fig. 17 Update sizes corresponding to numerically obtained shrunk optimal step lengths for an exemplary simulation run in the negative binomial simulation setting (9) without additional non-informative covariates

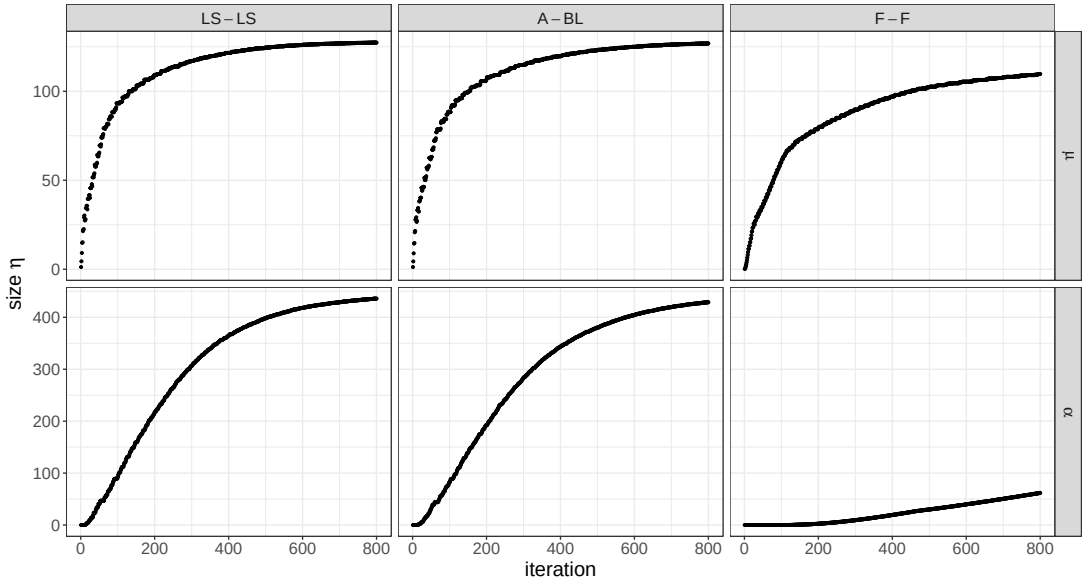

Fig. 18 Squared Euclidean norms of the predictors $\boldsymbol{\eta}_{\mu}^{[m]}$ and $\boldsymbol{\eta}_{\alpha}^{[m]}$ without the intercept for a boosting run in the negative binomial simulation setting (9) with shrunk optimal step lengths (left), base-learner based step lengths (middle) and fixed step lengths (right)

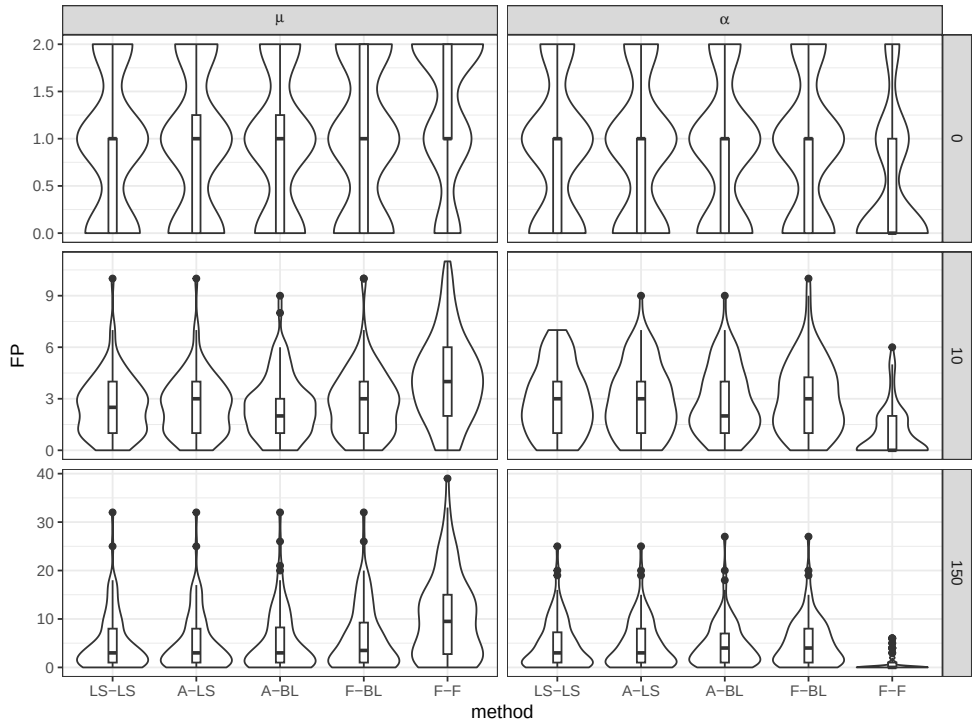

Fig. 19 Distribution of the number of false positives (FP) in the negative binomial simulation setting (9) for a varying number of additional non-informative covariates (rows)

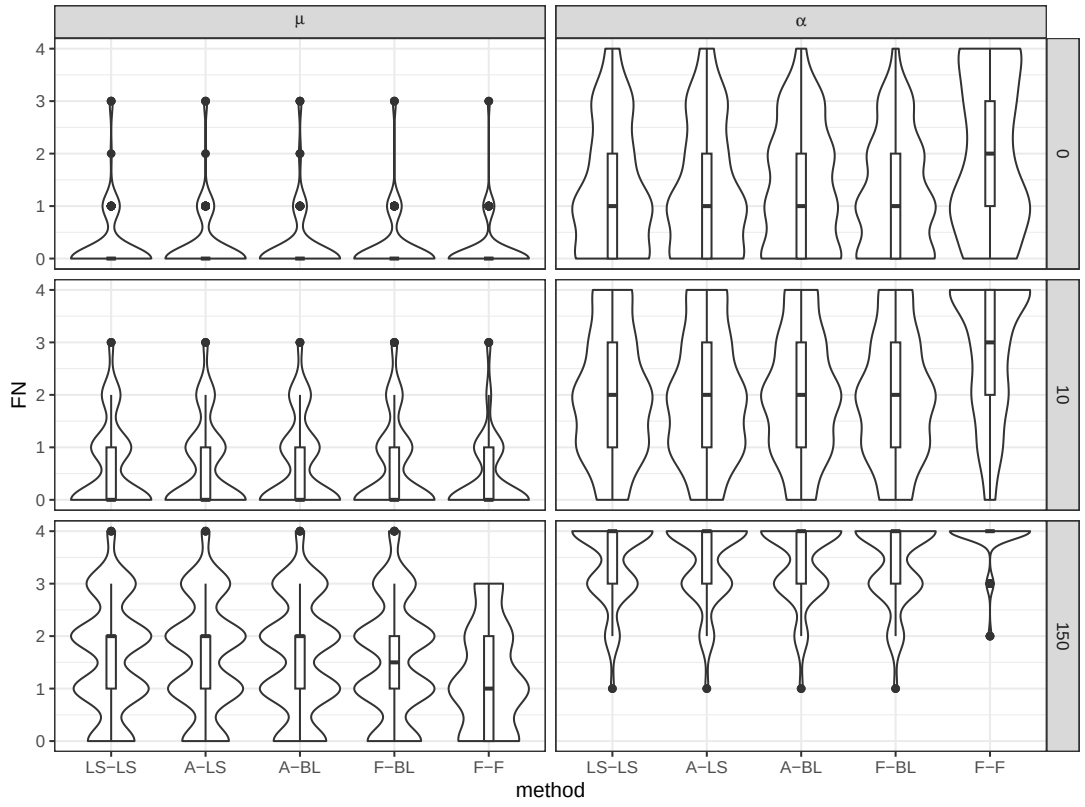

Fig. 20 Distribution of the number of false negatives (FN) in the negative binomial simulation setting (9) for a varying number of additional non-informative covariates (rows)

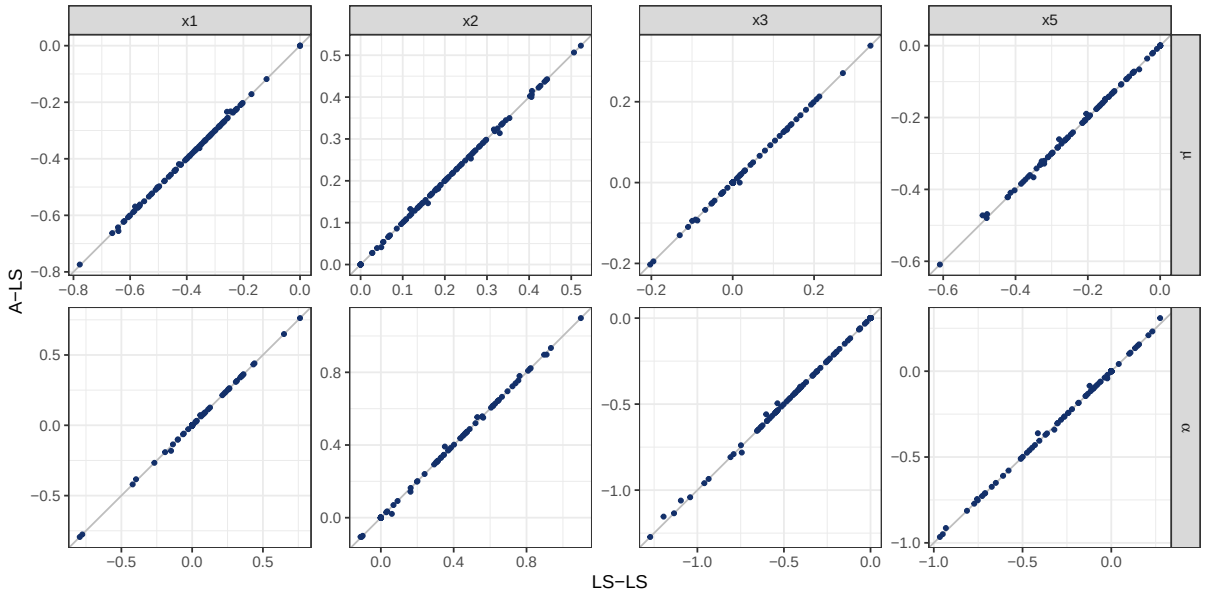

Fig. 21 Direct comparison of coefficient estimates of the two adaptive step length approaches LS-LS and A-LS in the negative binomial simulation setting (9) without additional non-informative covariates

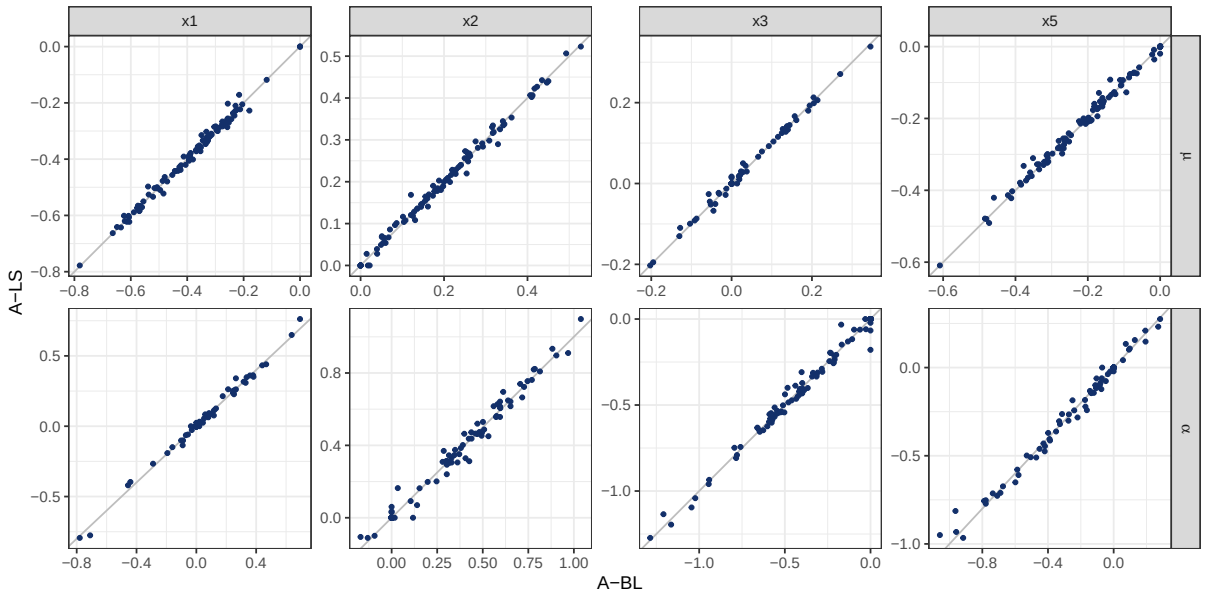

Fig. 22 Direct comparison of coefficient estimates of the two adaptive step length approaches A-LS and A-BL in the negative binomial simulation setting (9) without additional non-informative covariates

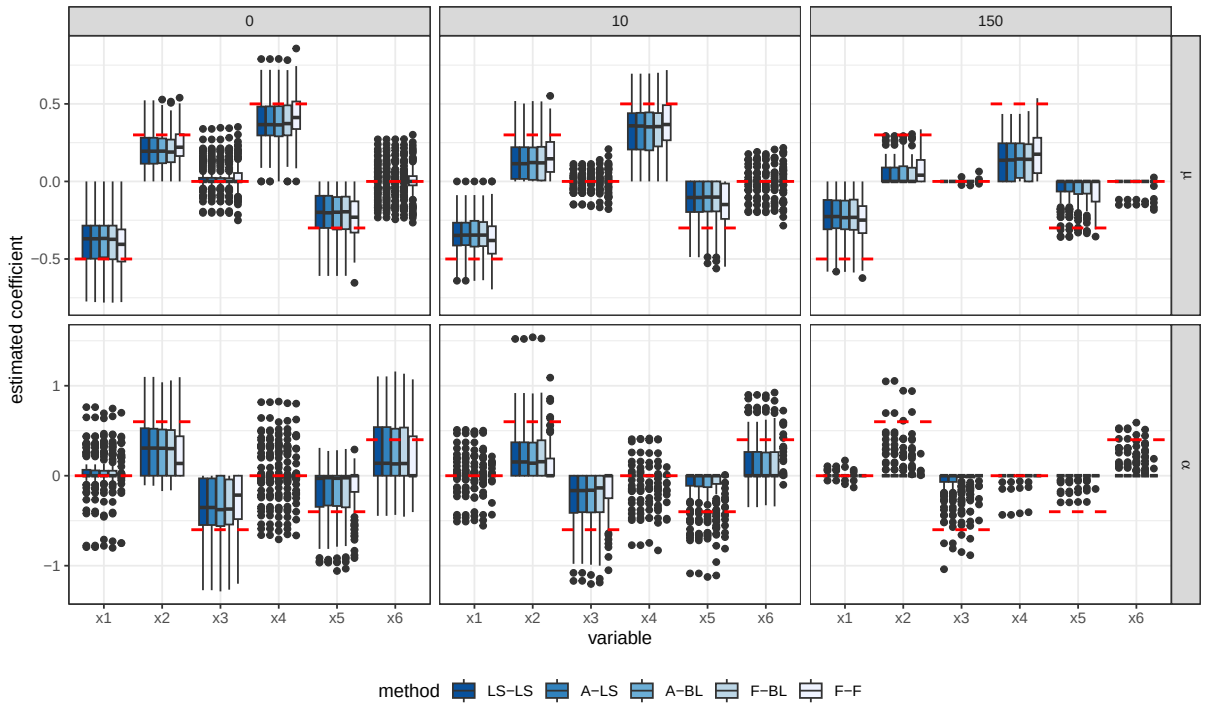

Fig. 23 Distribution of the coefficient estimates in the negative binomial simulation setting (9) for a varying number of additional non-informative covariates (columns). The red dashed lines represent the true coefficients

Table 6 Quartiles of the run times until m_{stop} without cross-validation in seconds and stopping iterations (in parenthesis) in the negative binomial simulation setting (9) with 0 (top), 10 (middle) and 150 (bottom) additional non-informative covariates

non-inf		LS-LS	A-LS	A-BL	F-BL	F-F
0	1st Qu.	0.166 (64)	0.117 (64)	0.061 (61)	0.144 (166)	0.183 (208)
	Median	0.263 (96)	0.171 (98)	0.102 (99)	0.239 (284)	0.438 (524)
	3rd Qu.	0.393 (134)	0.244 (133)	0.144 (142)	0.340 (398)	0.923 (1,042)
10	1st Qu.	0.130 (41)	0.099 (41)	0.049 (42)	0.116 (112)	0.142 (119)
	Median	0.212 (67)	0.162 (67)	0.088 (70)	0.243 (197)	0.288 (244)
	3rd Qu.	0.317 (93)	0.248 (93)	0.136 (102)	0.390 (308)	0.528 (436)
150	1st Qu.	0.105 (11)	0.105 (11)	0.118 (13)	0.182 (31)	0.219 (35)
	Median	0.231 (24)	0.234 (25)	0.189 (25)	0.482 (70)	0.464 (96)
	3rd Qu.	0.492 (62)	0.448 (62)	0.337 (57)	0.992 (176)	1.033 (208)

B.3 Weibull scale and shape model

We consider the following model for the Weibull response variable $y_i \sim \mathcal{WB}(\alpha_i, \lambda_i)$

$$\begin{aligned} \eta_{\lambda,i} &= \log(\lambda_i) = 0.6 + 0.15 \cdot x_{1i} - 0.2 \cdot x_{2i} + 0.4 \cdot x_{3i} - 0.25 \cdot x_{4i} \\ \eta_{\alpha,i} &= \log(\alpha_i) = -0.15 \cdot x_{3i} + 0.15 \cdot x_{4i} - 0.1 \cdot x_{5i} + 0.1 \cdot x_{6i}, \end{aligned} \tag{B.4}$$

where $x_1, ..., x_6$ are drawn independently from a uniform distribution on $[-1, 1]$. In this simulation setting, 0, 10, 150 additional non-informative covariates are considered.

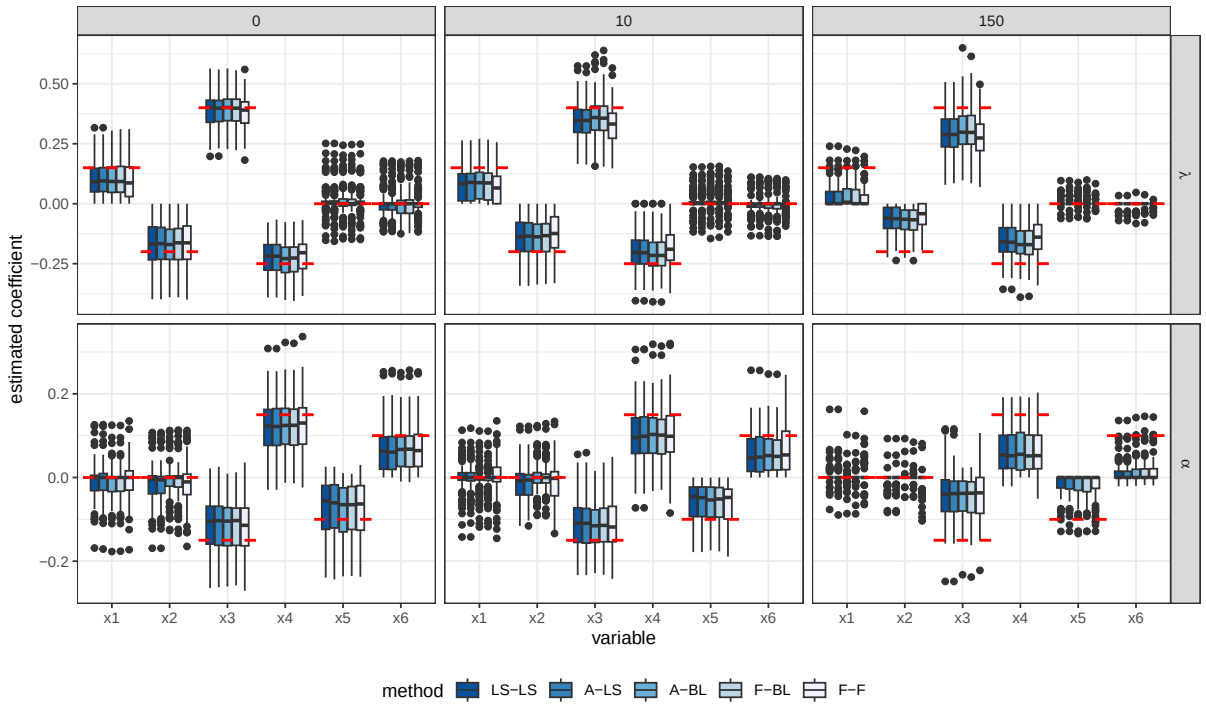

Fig. 24 Distribution of the coefficient estimates in the Weibull simulation setting (B.4) for a varying number of additional non-informative covariates (columns). The red dashed lines represent the true coefficients

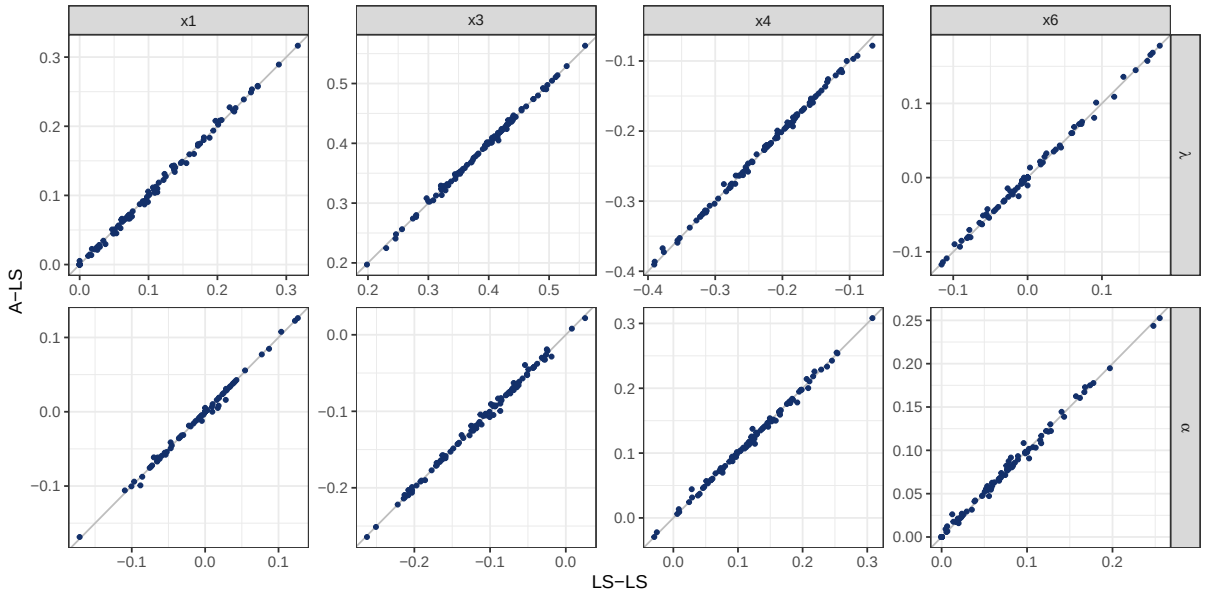

Fig. 25 Direct comparison of coefficient estimates of the two adaptive step length approaches LS-LS and A-LS in the Weibull simulation setting (B.4) without additional non-informative covariates

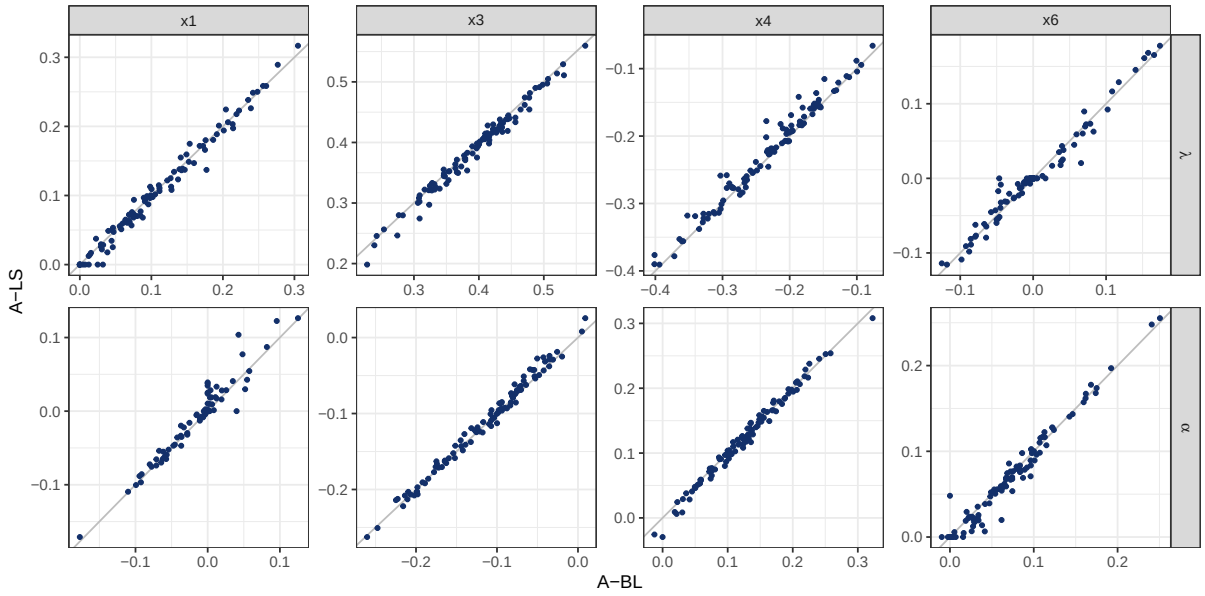

Fig. 26 Direct comparison of coefficient estimates of the two adaptive step length approaches A-LS and A-BL in the Weibull simulation setting (B.4) without additional non-informative covariates

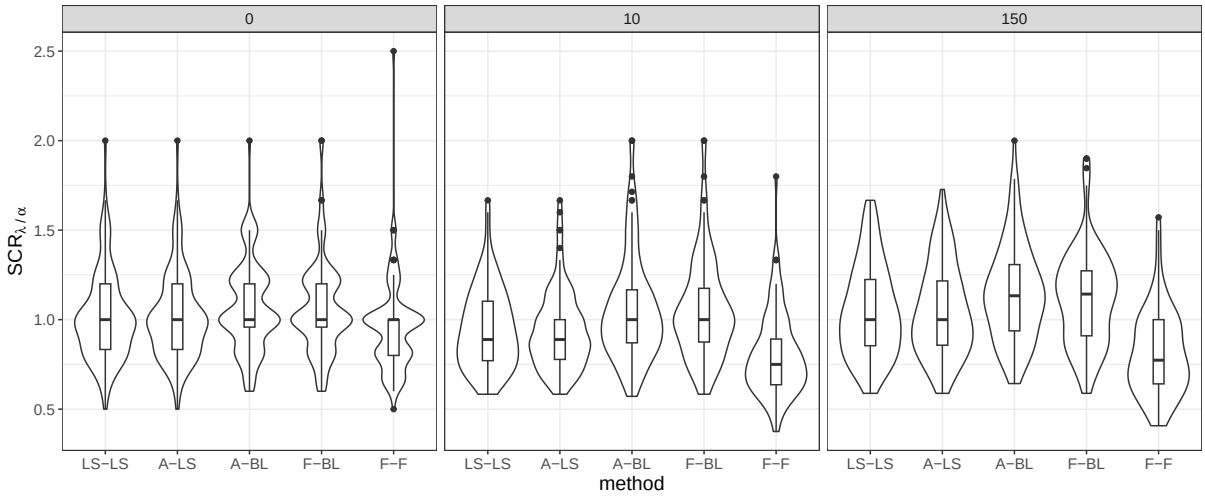

Fig. 27 Distribution of the ratio of numbers of covariates included in η_λ relative to η_α ($\text{SCR}_{\lambda/\alpha}$) in the Weibull simulation setting (B.4) for a varying number of additional non-informative covariates (columns)

Table 7 Quartiles of the run times until m_{stop} without cross-validation in seconds and stopping iterations (in parenthesis) in the Weibull simulation setting (B.4) with 0 (top), 10 (middle) and 150 (bottom) additional non-informative covariates

non-inf		LS-LS	A-LS	A-BL	F-BL	F-F
0	1st Qu.	0.173 (109)	0.117 (109)	0.105 (140)	0.062 (101)	0.059 (90)
	Median	0.224 (130)	0.160 (130)	0.133 (159)	0.085 (121)	0.075 (108)
	3rd Qu.	0.325 (180)	0.202 (180)	0.189 (196)	0.113 (151)	0.110 (147)
10	1st Qu.	0.165 (94)	0.136 (94)	0.213 (128)	0.081 (91)	0.065 (73)
	Median	0.218 (112)	0.192 (112)	0.282 (143)	0.108 (103)	0.091 (92)
	3rd Qu.	0.311 (141)	0.260 (141)	0.368 (163)	0.157 (130)	0.154 (114)
150	1st Qu.	0.455 (72)	0.424 (72)	0.577 (107)	0.388 (70)	0.292 (51)
	Median	0.555 (82)	0.556 (82)	0.752 (115)	0.506 (79)	0.382 (57)
	3rd Qu.	0.741 (93)	0.775 (93)	1.000 (126)	0.804 (88)	0.567 (68)

Appendix C Application Results

C.1 Australian Health Care Data

C.1.1 Information on the data

Table 8 Overview of the variables in the Australian health care data following Cameron et al. (1988) and Karlis and Ntzoufras (2005). For more a more detailed description of the data, see Cameron et al. (1988).

doctorco	Number of consultations with a doctor or specialist in the past 2 weeks.
sex	1 if female, 0 if male.
age	Age in years.
income	Monthly income in Australian dollars.
illness	Number of illnesses in the past 2 weeks, 0 to 5 or more.
actdays	Number of days of reduced activity in the past two weeks due to illness or injury.
hospadmi	Number of admissions to a hospital, psychiatric hospital, nursing or convalescent home in the past 12 months (up to 5 or more admissions).
hospdays	Number of nights in a hospital, etc. during most recent admission, 0 if no admission in the past 12 months.
medicine	Total number of prescribed and non-prescribed medications used in past 2 days.
prescrib	Total number of prescribed medications used in past 2 days.
nondocco	Number of consultations with non-doctor health professionals in the past 4 weeks.
levyplus	1 if covered by private health insurance fund for private patient in public hospital (with doctor of choice), 0 otherwise.
freepoor	1 if covered free by government because low income, recent immigrant, unemployed, 0 otherwise.
freepera	1 if covered free by government because of old age, disability pension, invalid veteran or family of deceased veteran, 0 otherwise.
hscore	General health questionnaire score using Goldberg's method. High score indicates bad health.
chcond1	1 if chronic condition(s) but not limited in activity, 0 otherwise.
chcond2	1 if chronic condition(s) and limited in activity, 0 otherwise.

Table 9 Overview of the variables and their summary statistics for the Australian health care data

Variable	Short description	Mean/frequency	standard dev.	Min/max	Frequency 0
doctorco	number of doctor consultations	0.302	0.798	0/9	79.8%
sex	1 = female	52.1%			
	0 = male	47.9%			
age	age in years	40.6	20.5	19/72	
income	monthly income in Austr. dollars	583	369	0/1,500	1.5%
illness	number of illnesses	1.432	1.384	0/5	29.9%
actdays	number of non-active days	0.862	2.888	0/14	85.8%
hospadmi	number of hospital admissions	0.174	0.508	0/5	86.5%
hospdays	number of nights in a hospital	1.334	6.120	0/80	86.5%
medicine	number of all medications	1.218	1.557	0/8	42.9%
prescrib	number of prescibed medications	0.863	1.415	0/8	59.4%
nondocco	number of non-doctor consultations	0.215	0.965	0/11	90.9%
levyplus	1 = private health insurance	44.3%			
	0 = no private health insurance	55.7%			
freepoor	1 = no insurance payment (e.g. due to low income)	4.3%			
	0 = otherwise	95.7%			
freepera	1 = no insurance payment (e.g. due to old age)	21.0%			
	0 = otherwise	79.0%			
hscore	health questionaire score	1.218	2.124	0/12	58.3%
chcond1	1 = chronic health condition, no limited activity	40.3%			
	0 = otherwise	59.7%			
chcond2	1 = chronic health condition, limited activity	11.7%			
	0 = otherwise	88.3%			

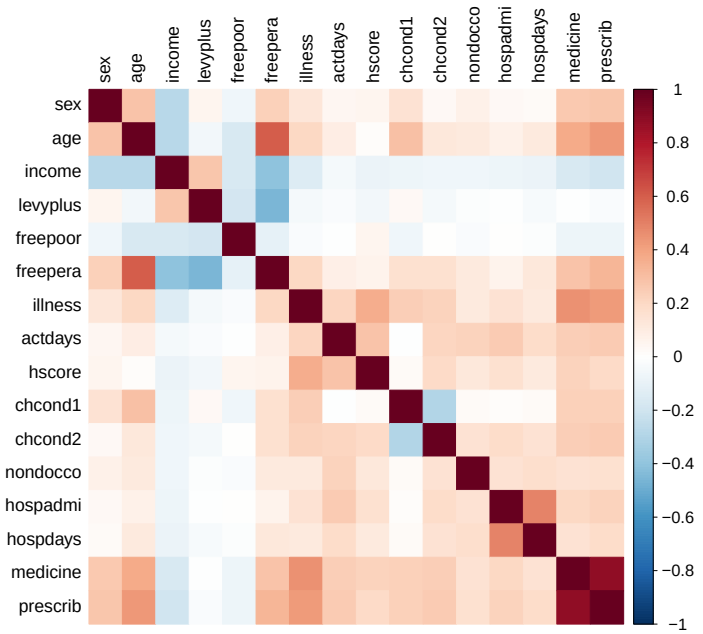

Fig. 28 Correlation plot of the Australia health care data

C.1.2 Further properties of the methods

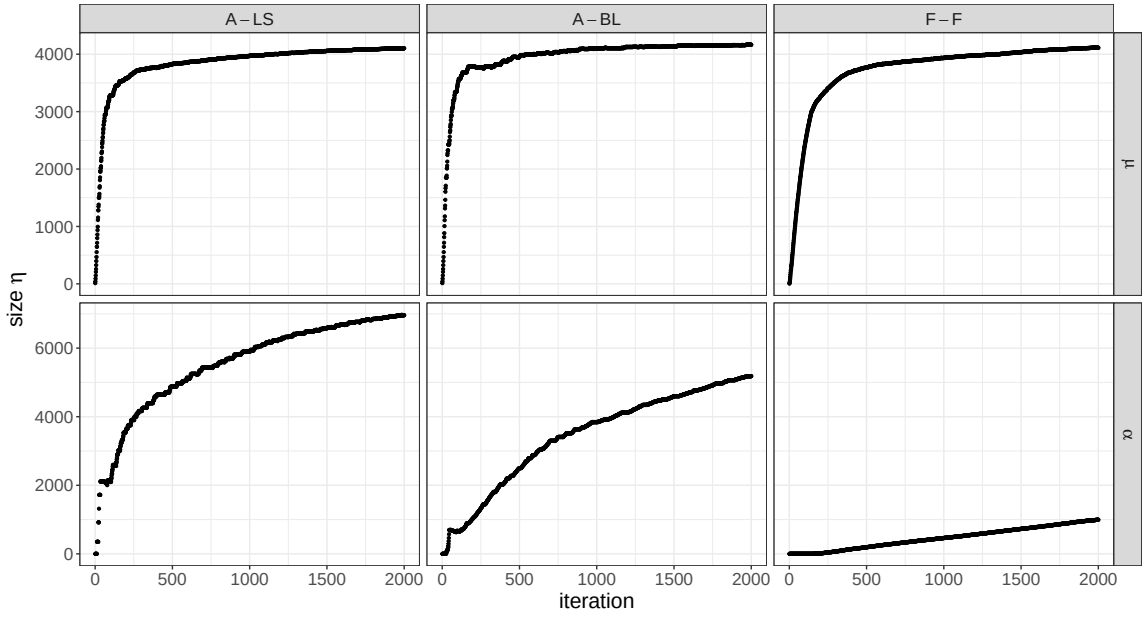

Fig. 29 Squared Euclidean norms of the predictors $\boldsymbol{\eta}_{\mu}^{[m]}$ and $\boldsymbol{\eta}_{\alpha}^{[m]}$ without the intercept for boosting runs of the Australian health care data with shrunk optimal (left), base-learner based (middle) and fixed step lengths (right)

Table 10 Quartiles of the run times until m_{stop} without cross-validation in seconds and stopping iterations (in parenthesis) for 100 cross-validation runs for the Australian health care data

	LS-LS	A-LS	A-BL	F-F
1st Qu.	2.62 (79)	1.93 (83)	1.36 (120)	12.05 (975)
Median	3.30 (102)	2.38 (103)	1.62 (142)	12.39 (1,084)
3rd Qu.	5.14 (146)	3.84 (151)	2.75 (220)	15.26 (1,241)

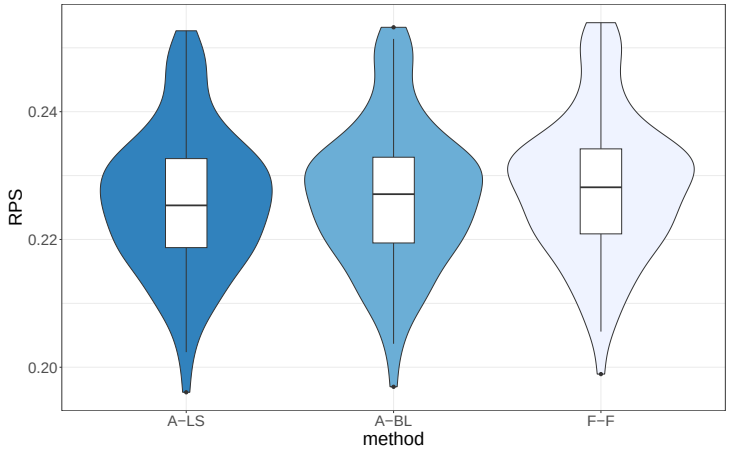

Fig. 30 Distribution of the ranked probability scores (RPS) for 100 random splits into training and validation set of the Australian health care data applying different step length approaches

C.1.3 Coefficients for LS-LS and F-BL step lengths

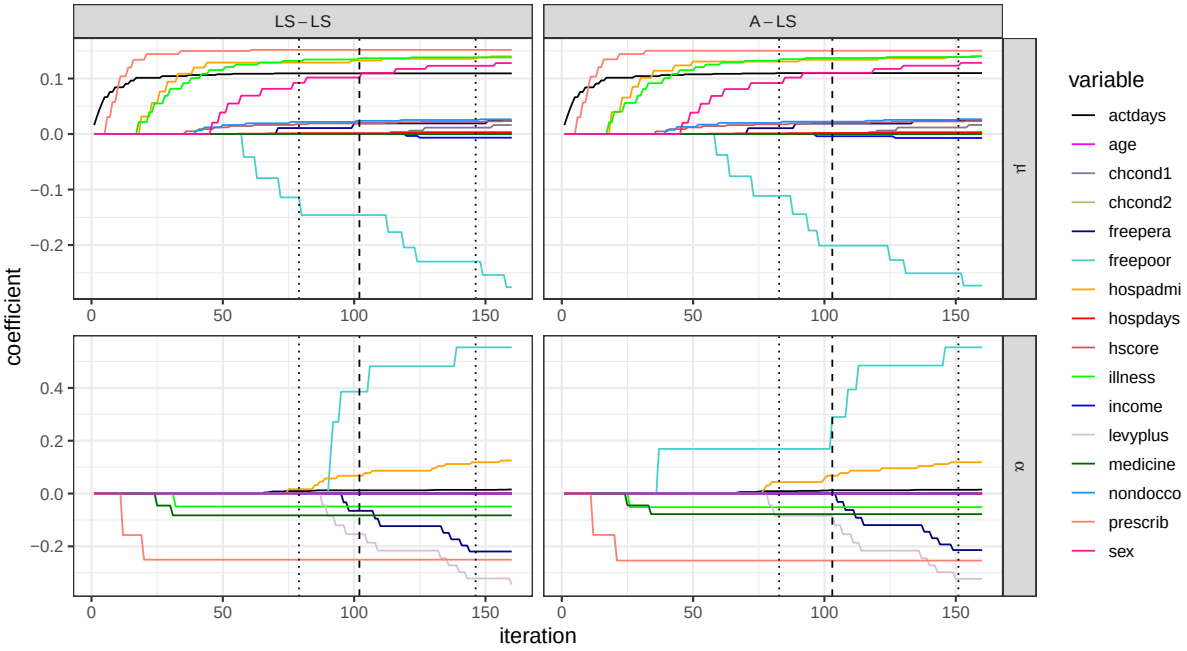

Fig. 31 Comparison of the coefficient paths for the Australian health care data using LS-LS and A-LS step lengths (columns). The vertical lines represent the median stopping iteration (dashed) as well as the 1st and 3rd quartile (dotted) of the stopping iteration obtained via cross-validation on 100 randomly drawn folds

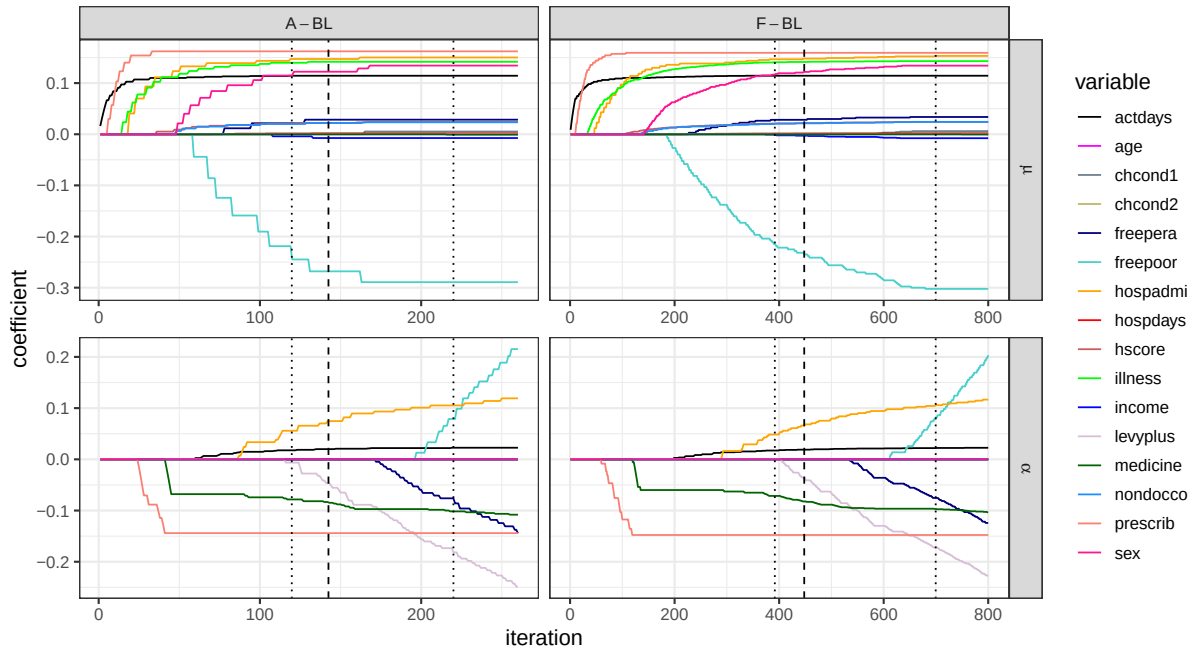

Fig. 32 Comparison of the coefficient paths for the Australian health care data using A-BL and F-BL step lengths (columns). The vertical lines represent the median stopping iteration (dashed) as well as the 1st and 3rd quartile (dotted) of the stopping iteration obtained via cross-validation on 100 randomly drawn folds. Please be aware of the differences in the x-axis scaling

Table 11 Coefficient estimates at the median stopping iteration for the Australian health care data applying different step length approaches

	η_μ					η_α				
	LS-LS	A-LS	A-BL	F-BL	F-F	LS-LS	A-LS	A-BL	F-BL	F-F
(Intercept)	-1.998	-2.009	-2.062	-2.054	-2.099	0.729	0.696	0.346	0.347	0.257
sex	0.102	0.110	0.122	0.122	0.142	0	0	0	0	0
age	0	0	0	0	0	0	0	0	0	0
income	0	-0.004	-0.007	-0.003	-0.012	0	0	0	0	0
illness	0.137	0.137	0.142	0.141	0.148	-0.049	-0.052	0	0	0
actdays	0.109	0.110	0.114	0.114	0.117	0.012	0.012	0.020	0.019	0.022
hospadmi	0.132	0.134	0.147	0.148	0.158	0.067	0.068	0.075	0.066	0.063
hospdays	0.002	0.002	0.002	0.002	0.002	0	0	0	0	0
medicine	0	0	0	0	-0.016	-0.083	-0.078	-0.083	-0.082	-0.100
prescrib	0.152	0.150	0.162	0.159	0.177	-0.250	-0.254	-0.144	-0.148	-0.089
nondocco	0.024	0.022	0.022	0.021	0.023	0	0	0	0	0
levyplus	0	0	0	0	0	-0.154	-0.084	-0.046	-0.036	0
freepoor	-0.146	-0.201	-0.268	-0.235	-0.347	0.386	0.290	0	0	0
freepera	0.019	0.019	0.028	0.028	0.034	-0.065	0	0	0	0
hscore	0.019	0.020	0.022	0.022	0.025	0	0	0	0	0
chcond1	0	0	0	0	0.008	0	0	0	0	0
chcond2	0	0	0	0	0	0	0	0	0	0

C.2 Diffuse Large-B-cell Lymphoma Data

Table 12 Coefficient estimates of η_λ for the DLBCL data applying different step length approaches (columns)

	LS-LS	A-LS	A-BL	F-BL	F-F
(Intercept)	1.056	1.054	1.054	1.047	0.857
X28641	0.061	0.061	0.073	0.052	0.088
X24683	0.031	0.031	0.057	0.074	0.019
X16799	0.043	0.044	0.040	0.050	0.064
X30240	-0.040	-0.042	-0.167	-0.224	0
X16545	0.036	0.037	0.066	0.063	0
X34688	-0.109	-0.110	-0.106	0	-0.099
X28673	-0.087	-0.087	0	0	-0.102
X30109	0.019	0.019	0	0	0.021
X26321	-0.072	-0.072	0	0	-0.071
X28910	-0.037	-0.037	0	0	0
X32118	0	0	-0.043	-0.083	0
X15937	0	0	-0.109	-0.143	0
X27593	0	0	0.020	0	0.020
X33186	0	0	-0.043	0	0
X16264	0	0	0	0.048	0
X24816	0	0	0	0	-0.051
X24412	0	0	0	0	-0.049
X27341	0	0	0	0	0.048
X33198	0	0	0	0	0.093

Table 13 Coefficient estimates of η_α for the DLBCL data applying different step length approaches (columns)

	LS-LS	A-LS	A-BL	F-BL	F-F
(Intercept)	0.306	0.306	0.509	0.493	0.091
X34729	0.046	0.046	0.059	0.074	0.067
X29181	0.035	0.035	0.048	0.085	0
X26250	0.054	0.054	0.013	0	0.164
X27129	-0.102	-0.102	0	0	-0.140
X17180	0.020	0.020	0	0	0
X27593	0.021	0.021	0	0	0
X24325	0	0	-0.005	-0.013	0
X34552	0	0	0.052	0.039	0
X26012	0	0	-0.027	0	0
X30727	0	0	0.004	0	0
X28178	0	0	0	0	0.036

Table 14 Quartiles of the run times until m_{stop} without cross-validation in seconds and stopping iterations (in parenthesis) for 100 cross-validation runs for the DLBCL data

	LS-LS	A-LS	A-BL	F-BL	F-F
1st Qu.	3.93 (19)	4.40 (19)	5.86 (29)	4.83 (9)	5.69 (21)
Median	4.28 (21)	5.02 (21)	7.40 (35)	5.50 (12)	6.57 (24)
3rd Qu.	4.74 (24)	5.60 (24)	8.65 (41)	6.08 (14)	7.07 (26)

References

Cameron, A. C., Trivedi, P. K., Milne, F., and Piggott, J. (1988). A microeconometric model of the demand for health care and health insurance in Australia. *The Review of economic studies*, 55(1):85–106.

Karlis, D. and Ntzoufras, I. (2005). Bivariate poisson and diagonal inflated bivariate poisson regression models in R. *Journal of Statistical Software*, 14(10):1–36.