The multiple faces of shrinkage

Georg Heinze
Center for Medical Statistics, Informatics and Intelligent Systems
Section for Clinical Biometrics

Partly supported by Austrian Science Fund FWF, Project I2276-N33

The multiple faces of shrinkage

- Introduction

- Post-estimation shrinkage methods:
  Dunkler, Sauerbrei and Heinze, JStatSoft 2016

- From bias reduction to shrinkage and beyond
  Puhr, Heinze, Nold, Lusa and Geroldinger, StatMed 2017

Historical outline

- Gauss (early 1800s): Least Squares: unbiasedness as a paradigm
- James&Stein 1961: Biased but better
- Hoerl&Kennard 1970: Ridge regression
- Efron 1975: The two statisticians applying for a job
- Copas 1983: Variable selection, bias and shrinkage estimators
- Van Houwelingen&leCessie 1990: Jackknife-type global shrinkage factor
- Tibshirani 1996: Lasso
- Greenland 2000: The sharp-shooter
- Van Houwelingen 2001: Shrinkage and penalization review

Purposes of shrinkage estimators

- Sacrifice unbiasedness to reduce MSE of statistics (predictions, effect estimates)
- Correct miscalibration
- Reduce over-optimism
- Variable selection

Post-estimation shrinkage methods

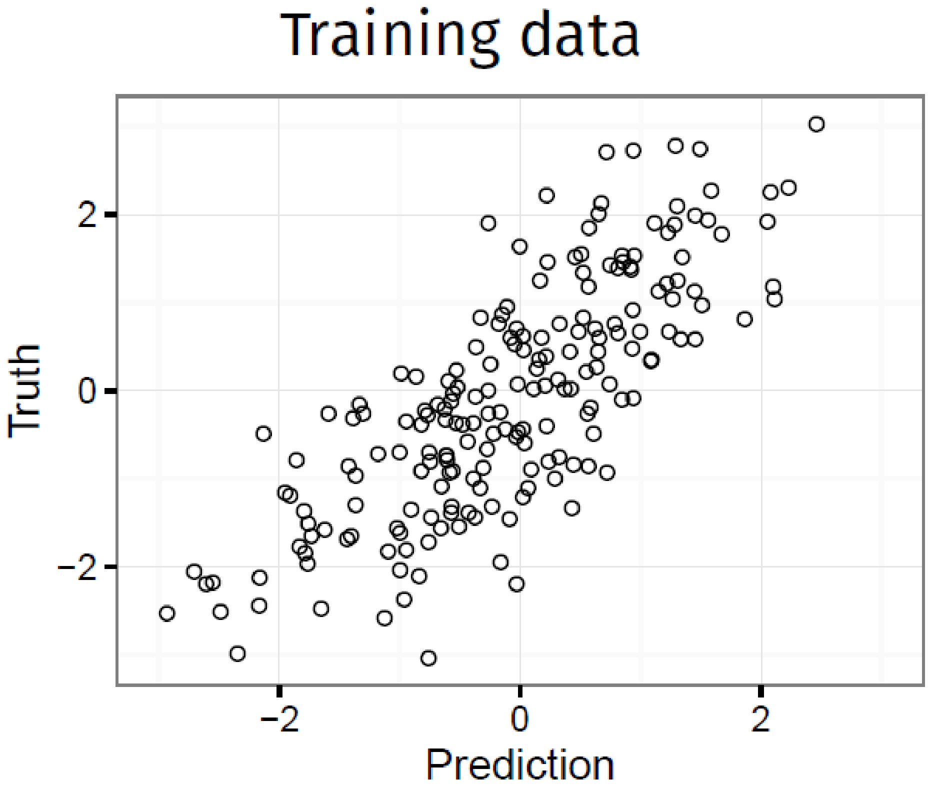
Joint work with Michael Kammer, Daniela Dunkler, Willi Sauerbrei

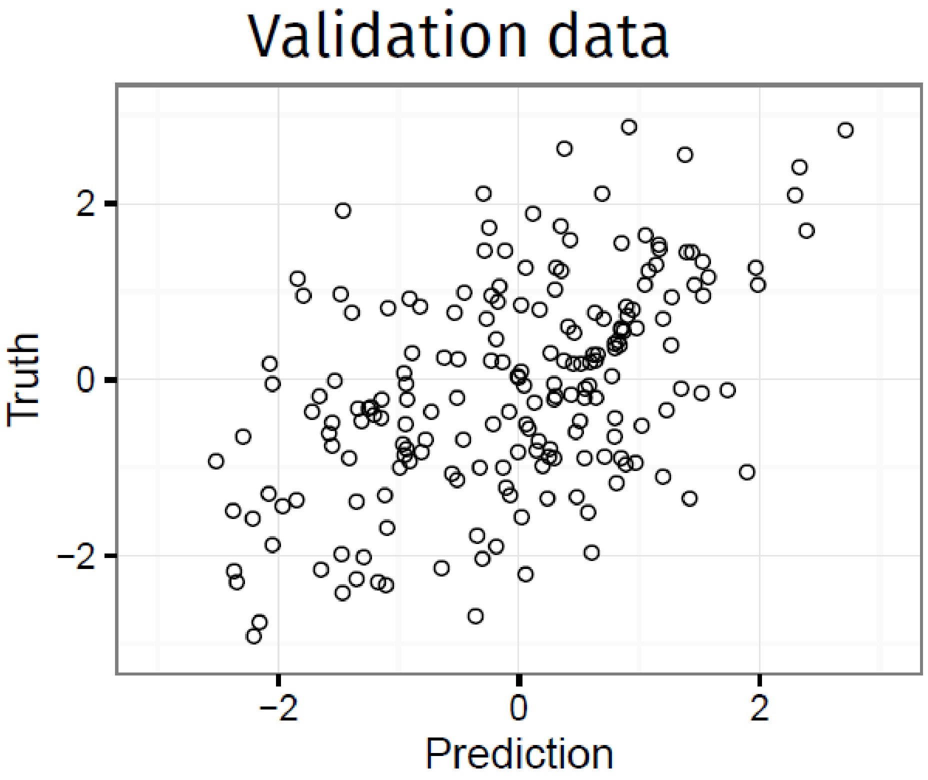
Shrinkage as a problem

- Predictions for new data are too extreme
- Consequence of regression to the mean (Copas 1997)
- Problematic: low level of information per variable or poor model fit

Shrinkage as a problem

- Predictions for new data are too extreme
- Consequence of regression to the mean (Copas 1997)
- Problematic: low level of information per variable or poor model fit

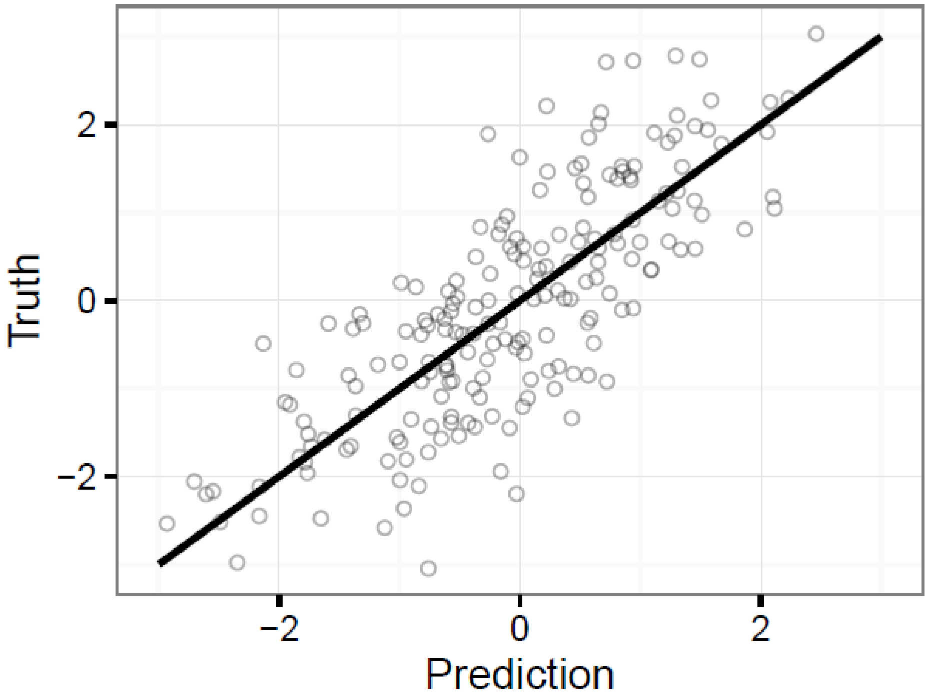




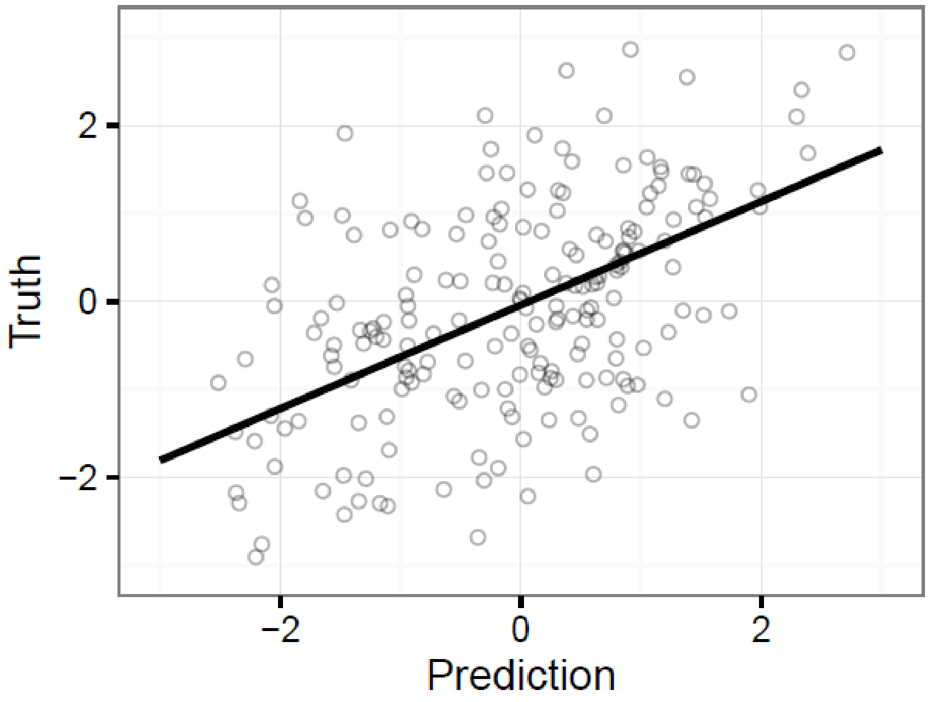
Shrinkage as a problem

- Predictions for new data are too extreme
- Consequence of regression to the mean (Copas 1997)
- Problematic: low level of information per variable or poor model fit

Training data



Validation data



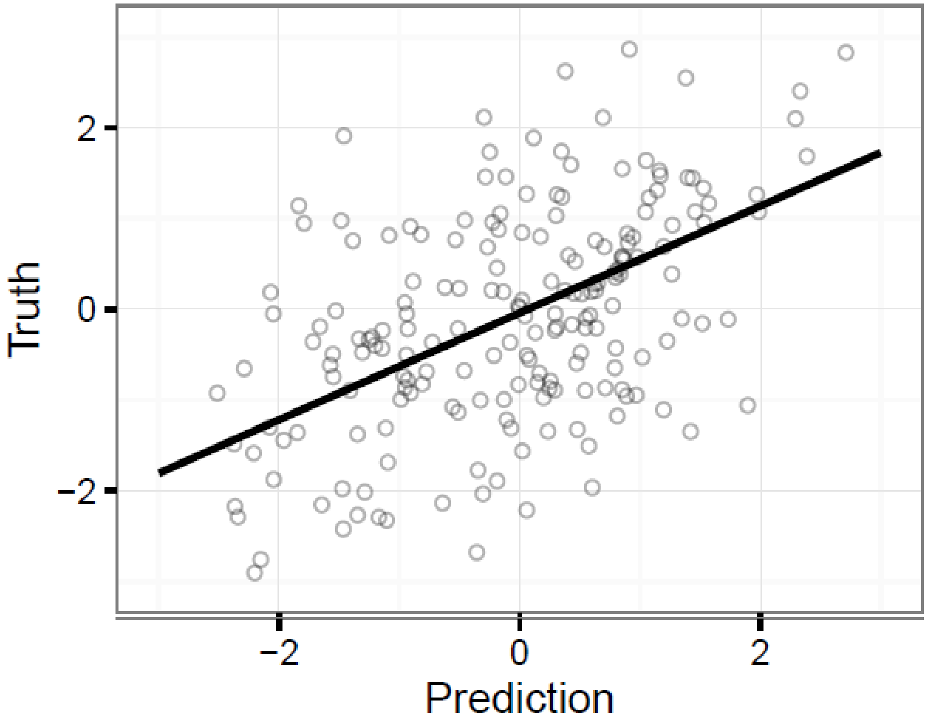
Anticipate and correct for shrinkage

Minimize overestimation by shrinkage of coefficients towards origin

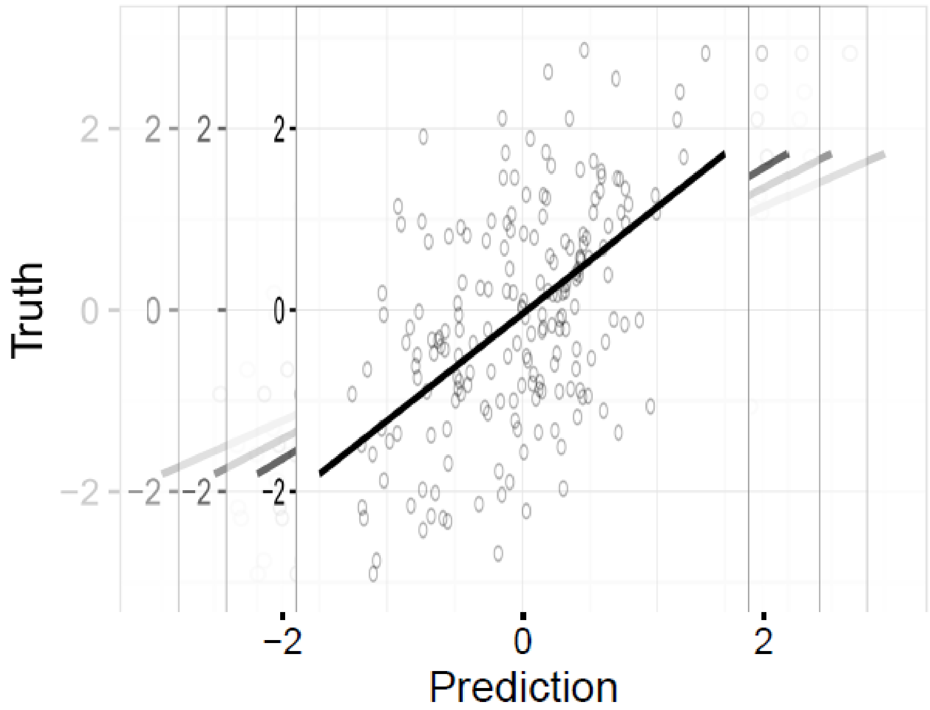
$\Rightarrow$ introduces (small) bias ...

$\Rightarrow$ ...but often leads to reduction in mean squared error

Poor calibration



Improved calibration



Post-estimation shrinkage methods

- Build model
  - Estimate vector of regression coefficients $\hat{\beta}$ using maximum likelihood (ML)
  - Use routine modeling strategies
  - Select variables if needed and re-estimate

- Estimate global shrinkage factor b
  - Leave-one-out resampling of $\hat{\beta}$: $\hat{\beta}^{(-i)}$
  - Perform ML regression of outcomes on jackknifed linear predictors $\eta_i = \sum_j x_{ij}\hat{\beta}_j^{(-i)}$
  - Shrinkage factor b = regression coefficient of this analysis (PESg)

Use of the shrinkage factors

- Shrinkage factor = 1 … no shrinkage
- Shrinkage factor = 0 … maximum shrinkage
- Shrinkage factor >0.8 … OK, application of shrinkage factor could improve predictions
- Shrinkage factor 0.5<0.8 … Modeling OK?
- Shrinkage factor 0<0.5 … bad fit, unnecessary information in the model

- Once estimated, the shrinkage factor is then used to multiply all regression coefficients:

$$\hat{y}^{new} = \hat{\beta}_0 + b\left(x_i^{new}\hat{\beta}\right)$$

- (Analogously for logistic or Cox regression)

Sauerbrei‘s (1999) ‚parameterwise shrinkage factors‘

- Leave-one-out resampling of $\hat{\beta}$: $\hat{\beta}^{(-i)}$
- Perform ML regression of outcomes on

  jackknifed *partial* linear predictors $\eta_{ij} = x_{ij}\hat{\beta}_j^{(-i)}$
- Regression coefficients b_j are used as parameterwise shrinkage factors (PESp)

- Recommended for models obtained by stepwise variable selection

Dunkler‘s (2016) extension of parameterwise shrinkage

- Dunkler et al (2016) investigated the parameterwise shrinkage factors b_j
- Provide a rough standard error estimate of the shrinkage factor
  - The closer to 1, the lower the standard error

- Joint shrinkage factor: hybrid between global and parameterwise shrinkage,
  - combine shrinkage factor estimation for groups of semantically related variables, or groups of regression coefficients which are uninterpretable alone
  - Given G such groups, compute $\eta_{ig} = \sum_{j \in J_g} x_{ij} \hat{\beta}_j^{(-i)}$, for $g = 1, \dots, G$
  - Use η_{ig} in second step and estimate SF‘s $b_g, g = 1, \dots, G$

- DFBETA method: considerable computational gain by approximating

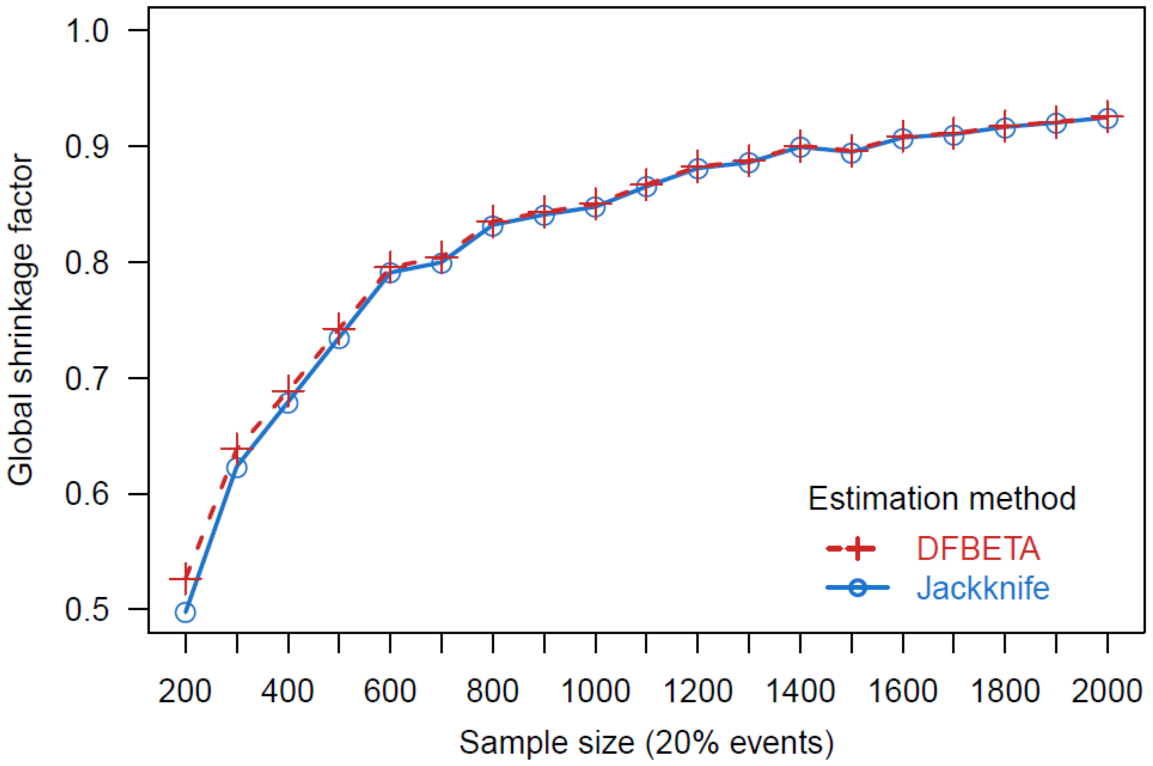
$$\hat{\beta}^{(-i)} \approx \hat{\beta} - DFBETA_i$$

Example: deep vein thrombosis study

```
              coef exp(coef) se(coef)     z      p
log2ddim     0.219     1.245   0.0854  2.56 0.0100
sex.male     0.491     1.634   0.1847  2.66 0.0079
loc.distal  -0.922     0.398   0.3101 -2.97 0.0029
loc.proximal -0.205    0.815   0.1787 -1.15 0.2500

Likelihood ratio test=24.5 on 4 df, p=6.37e-05, n=929, number of events=147
```

Explanatory variable	Jackknife	DFBETA	Relative difference
Global shrinkage	0.8076	0.8123	0.6%
Parameterwise shrinkage			
log2ddim	0.7321	0.7385	0.9%
sex.male	0.8351	0.8373	0.3%
loc.distal	0.8394	0.8449	0.7%
loc.proximal	0.1321	0.1470	11.2%
Joint shrinkage			
log2ddim	0.7806	0.7864	0.7%
sex.male	0.8364	0.8386	0.3%
loc	0.8055	0.8111	0.7%
Computing time	3.03	0.02	-99.3%



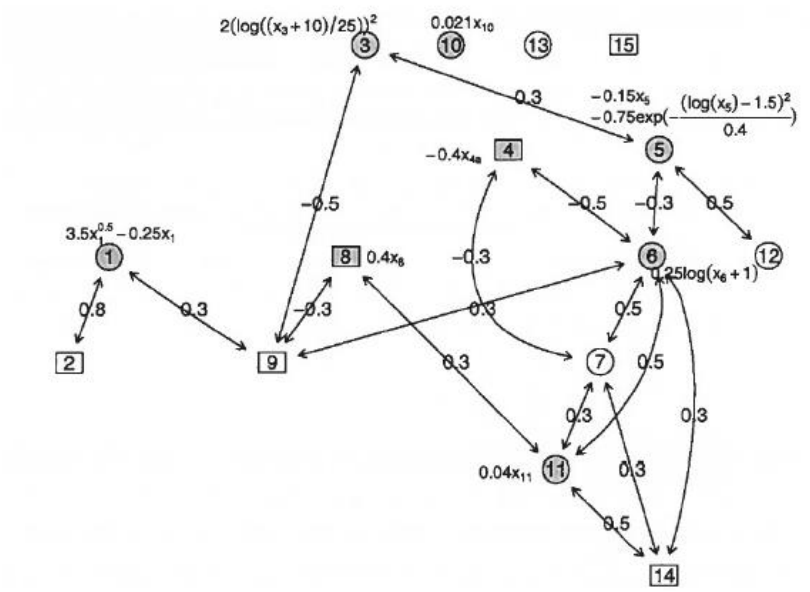
How do shrinkage effects of different methods compare?

- We evaluated shrinkage effects of
  - Post-estimation Shrinkage, parameterwise (PESp)
    applied after AIC-guided backward elimination of effects
  - Post-estimation Shrinkage, global (PESg)
    applied to full model
  - Ridge regression with optimized tuning parameter λ
  - Lasso regression with optimized tuning parameter λ

- Realistic scenario resembling a typical observational study
- Survival outcome, censoring

Simulations: realistic setup

- Correlation structure resembles medical data (Binder et al. 2011)
- Binary, ordinal and continuous variables
- Baseline hazard follows Weibull distribution
- Uniform censoring
- 1000 simulations



- Eventrate 50%
- Censoring 50%

Selected scenario (EPV 10): Shrinkage of coefficients

Compare coefficient for method M with estimation by ML: shrinkage($\hat{\beta}_i^M$) $= \hat{\beta}_i^M / \hat{\beta}_i^{ML}$

	Type	r_{mult} +	β_s *
c1	continuous	0.47	0.76
c2	continuous	0.40	0.59
c3	continuous	0.26	0.57
c4	continuous	0.38	0.53
c5	continuous	0	0.52
c6	continuous	0.42	0.52
b7	binary	0.13	0.48
b8	binary	0.15	0.32
b9	binary	0.38	0
b10	binary	0.10	0
c11	continuous	0.21	0
o12	ordinal	0.20	0
o13	ordinal	0.18	0
c14	continuous	0.27	0
c15	continuous	0	0
b16	binary	0.18	0
b17	binary	0	0

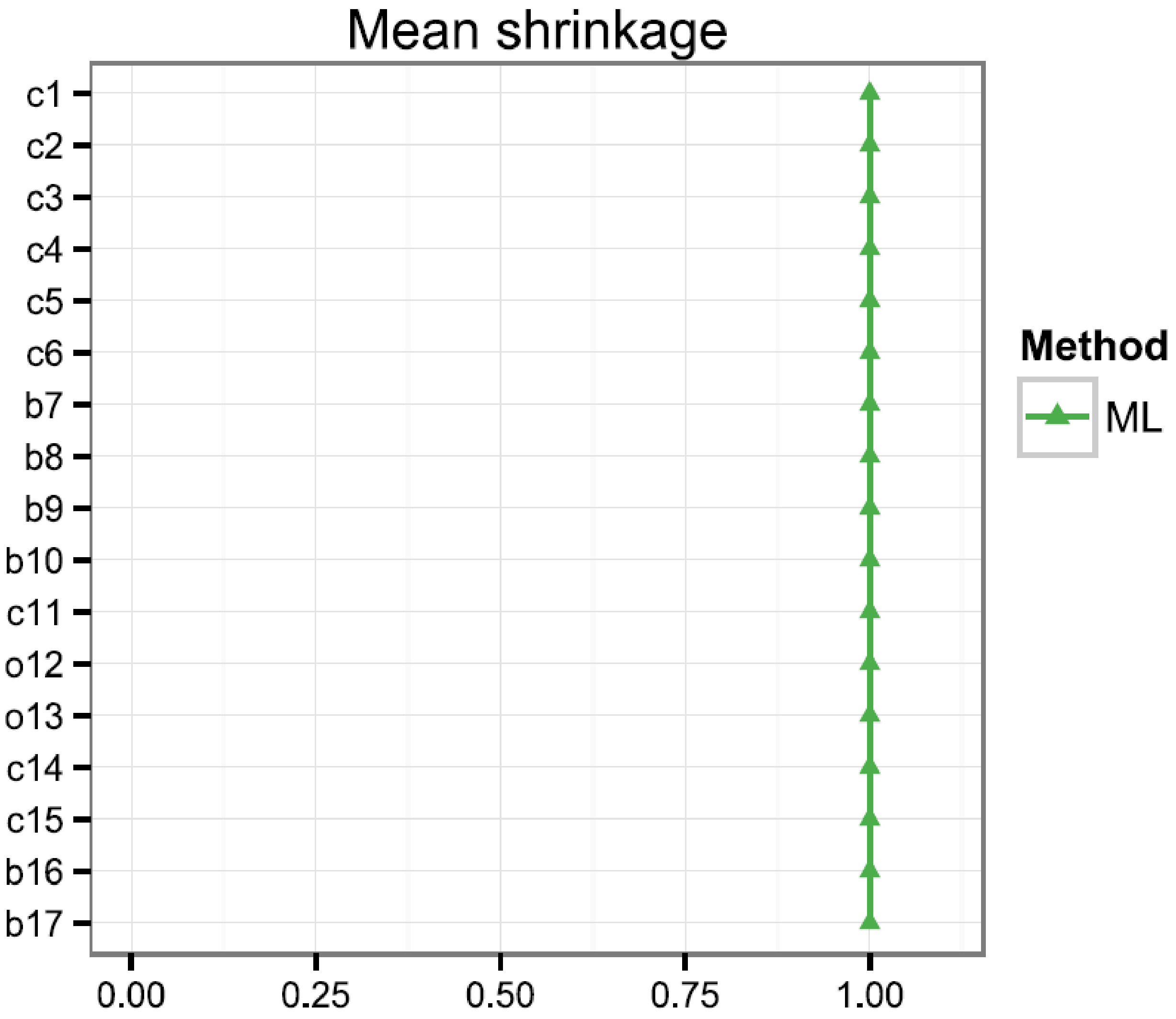
+ multiple correlation, * standardized coefficient

Selected scenario (EPV 10): Shrinkage of coefficients

Compare coefficient for method M with estimation by ML: $\text{shrinkage}(\hat{\beta}_i^M) = \hat{\beta}_i^M / \hat{\beta}_i^{ML}$

	Type	r_{mult} +	β_s *
c1	continuous	0.47	0.76
c2	continuous	0.40	0.59
c3	continuous	0.26	0.57
c4	continuous	0.38	0.53
c5	continuous	0	0.52
c6	continuous	0.42	0.52
b7	binary	0.13	0.48
b8	binary	0.15	0.32
b9	binary	0.38	0
b10	binary	0.10	0
c11	continuous	0.21	0
o12	ordinal	0.20	0
o13	ordinal	0.18	0
c14	continuous	0.27	0
c15	continuous	0	0
b16	binary	0.18	0
b17	binary	0	0

+ multiple correlation, * standardized coefficient

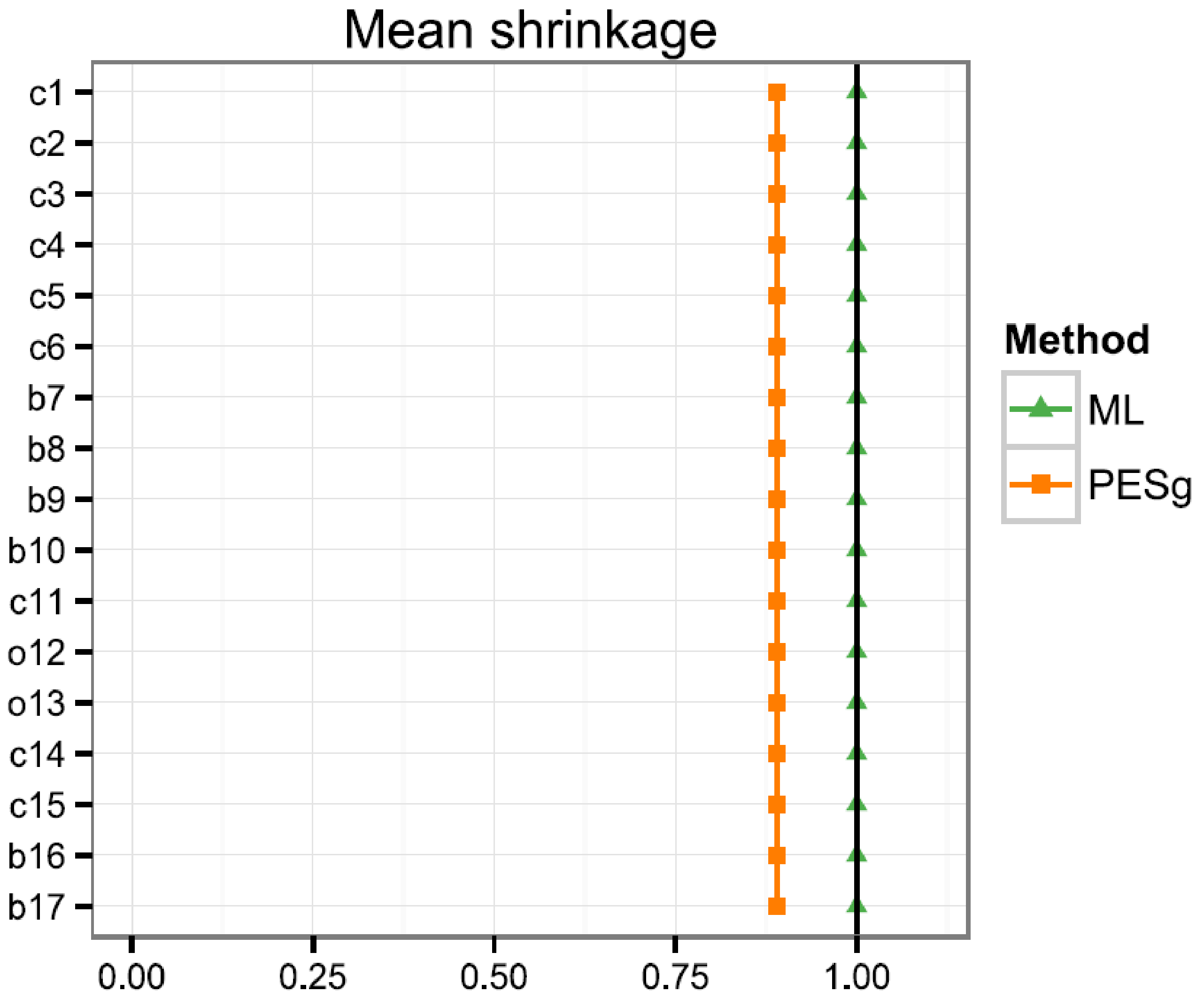


Selected scenario (EPV 10): Shrinkage of coefficients

Compare coefficient for method M with estimation by ML: shrinkage$(\hat{\beta}_i^M) = \hat{\beta}_i^M / \hat{\beta}_i^{ML}$

	Type	r_{mult} +	β_s *
c1	continuous	0.47	0.76
c2	continuous	0.40	0.59
c3	continuous	0.26	0.57
c4	continuous	0.38	0.53
c5	continuous	0	0.52
c6	continuous	0.42	0.52
b7	binary	0.13	0.48
b8	binary	0.15	0.32
b9	binary	0.38	0
b10	binary	0.10	0
c11	continuous	0.21	0
o12	ordinal	0.20	0
o13	ordinal	0.18	0
c14	continuous	0.27	0
c15	continuous	0	0
b16	binary	0.18	0
b17	binary	0	0

+ multiple correlation, * standardized coefficient

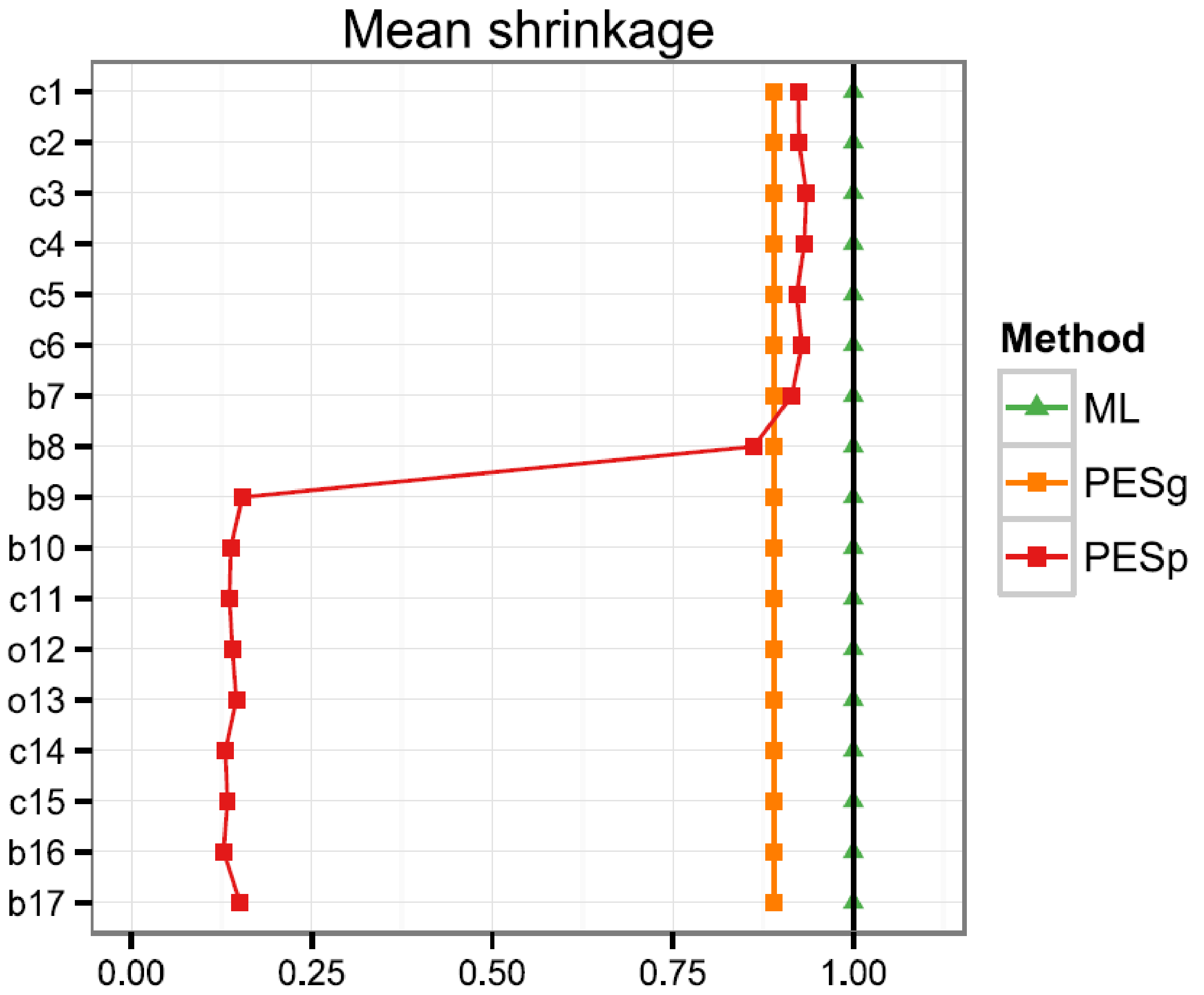


Selected scenario (EPV 10): Shrinkage of coefficients

Compare coefficient for method M with estimation by ML: $\text{shrinkage}(\hat{\beta}_i^M) = \hat{\beta}_i^M / \hat{\beta}_i^{ML}$

	Type	$r_{mult}{}^{+}$	$\beta_s{}^{*}$
c1	continuous	0.47	0.76
c2	continuous	0.40	0.59
c3	continuous	0.26	0.57
c4	continuous	0.38	0.53
c5	continuous	0	0.52
c6	continuous	0.42	0.52
b7	binary	0.13	0.48
b8	binary	0.15	0.32
b9	binary	0.38	0
b10	binary	0.10	0
c11	continuous	0.21	0
o12	ordinal	0.20	0
o13	ordinal	0.18	0
c14	continuous	0.27	0
c15	continuous	0	0
b16	binary	0.18	0
b17	binary	0	0

$^{+}$ multiple correlation, * standardized coefficient

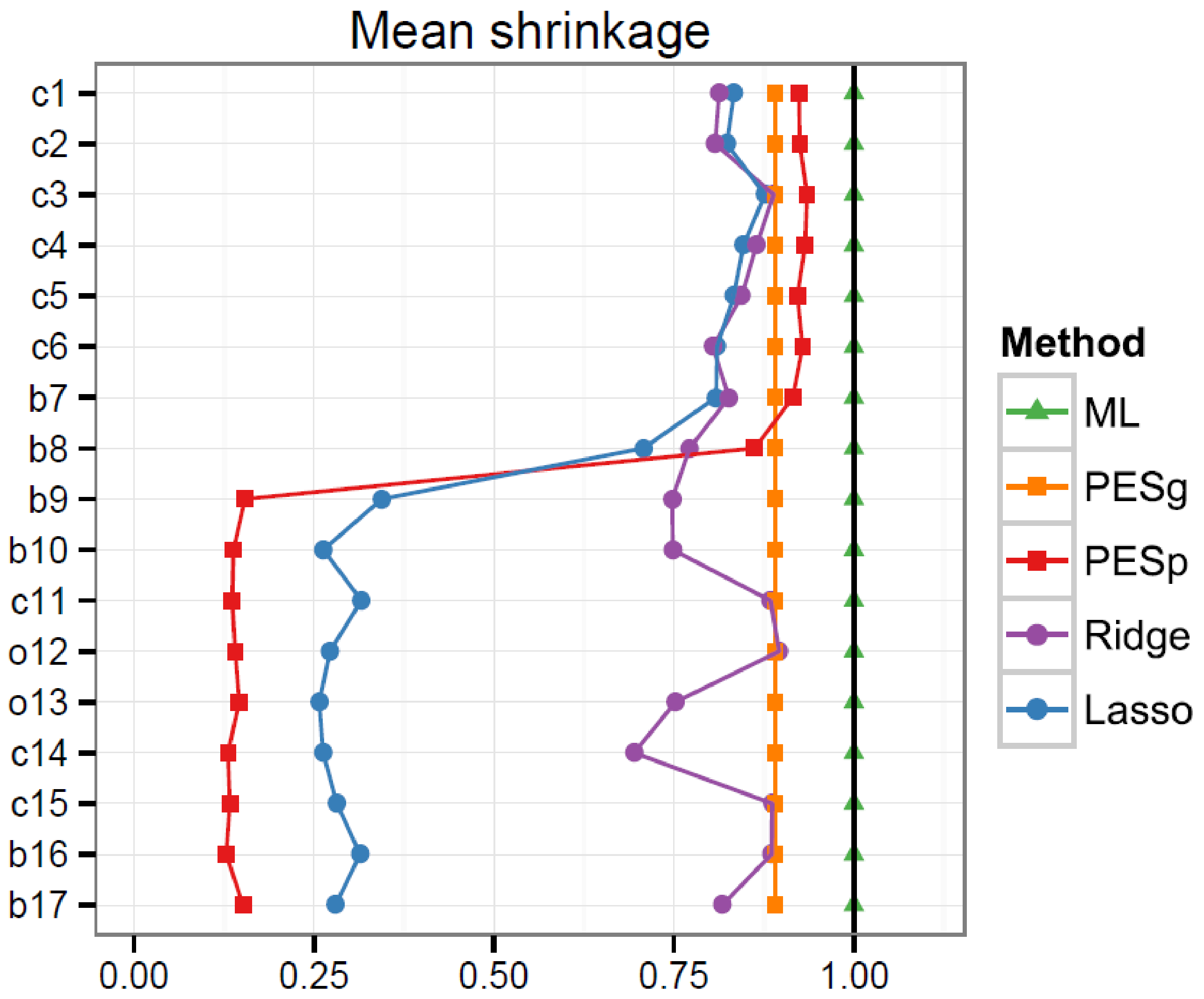


Selected scenario (EPV 10): Shrinkage of coefficients

Compare coefficient for method M with estimation by ML: $\text{shrinkage}(\hat{\beta}_i^M) = \hat{\beta}_i^M / \hat{\beta}_i^{ML}$

	Type	$r_{mult}{}^{+}$	$\beta_s{}^{*}$
c1	continuous	0.47	0.76
c2	continuous	0.40	0.59
c3	continuous	0.26	0.57
c4	continuous	0.38	0.53
c5	continuous	0	0.52
c6	continuous	0.42	0.52
b7	binary	0.13	0.48
b8	binary	0.15	0.32
b9	binary	0.38	0
b10	binary	0.10	0
c11	continuous	0.21	0
o12	ordinal	0.20	0
o13	ordinal	0.18	0
c14	continuous	0.27	0
c15	continuous	0	0
b16	binary	0.18	0
b17	binary	0	0

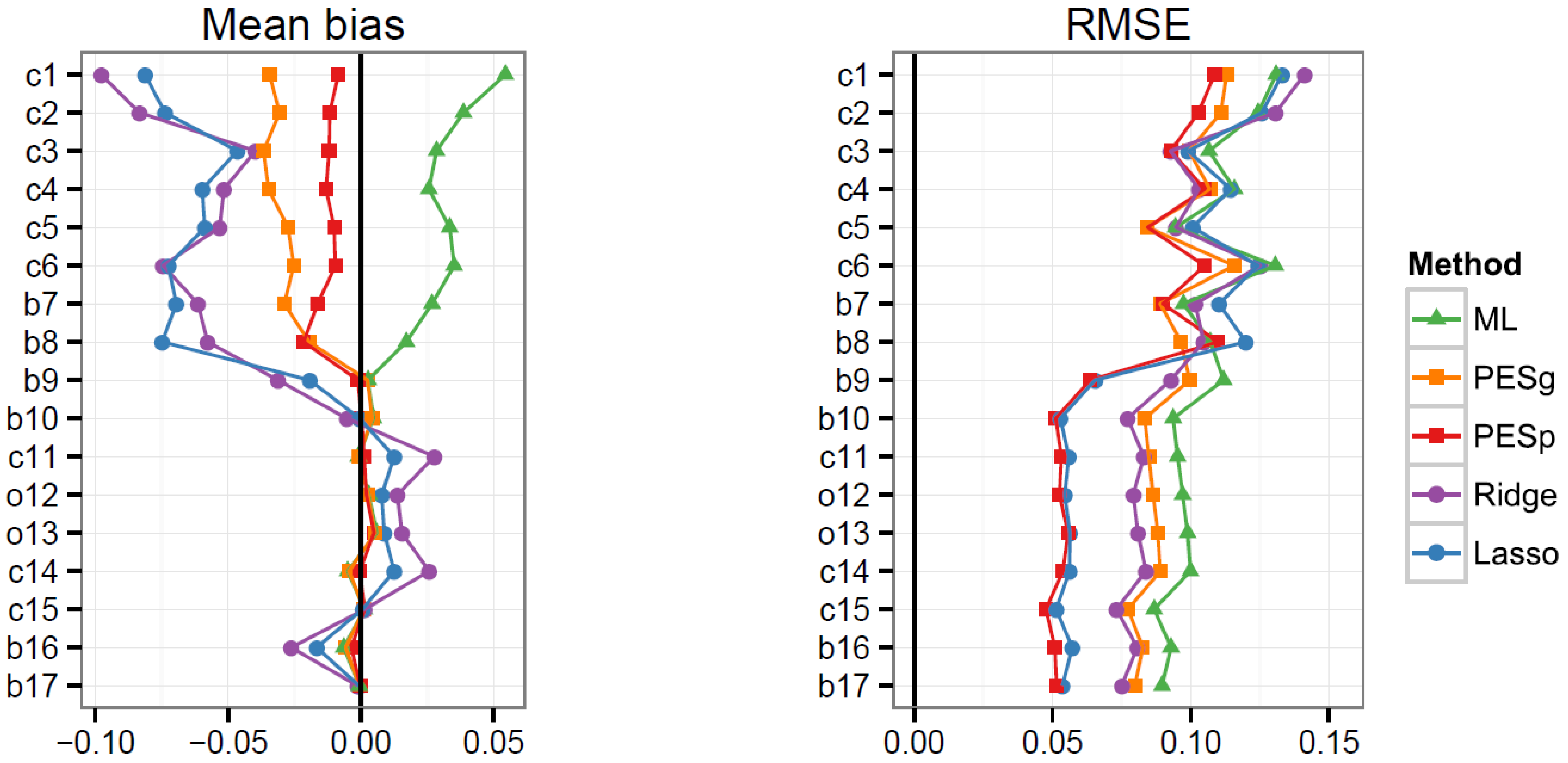
$^{+}$ multiple correlation, * standardized coefficient



Selected scenario (EPV 10): Bias and RMSE of coefficients

$$\text{bias}(\hat{\beta}_i^M) = \beta_i - \hat{\beta}_i^M$$

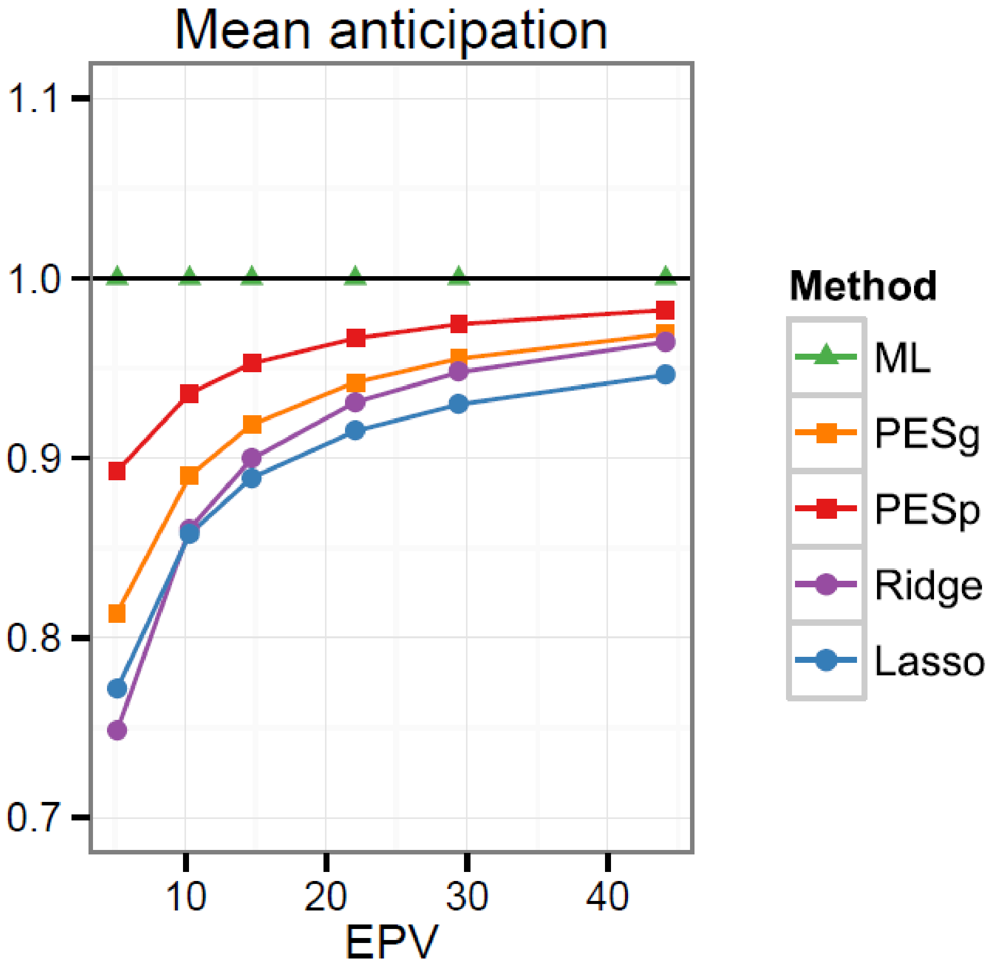
$$\text{RMSE}(\hat{\beta}_i^M) = \sqrt{\frac{1}{n_{sim}} \sum (\beta_i - \hat{\beta}_i^M)^2}$$



Selected scenario: Shrinkage correction over EPV

Shrinkage measured in terms of calibration of linear predictor: $h(t) = h_0(t)\exp(bX\hat{\beta})$

Anticipated shrinkage: $1/b^{\mathrm{M}}_{\mathrm{train}}$

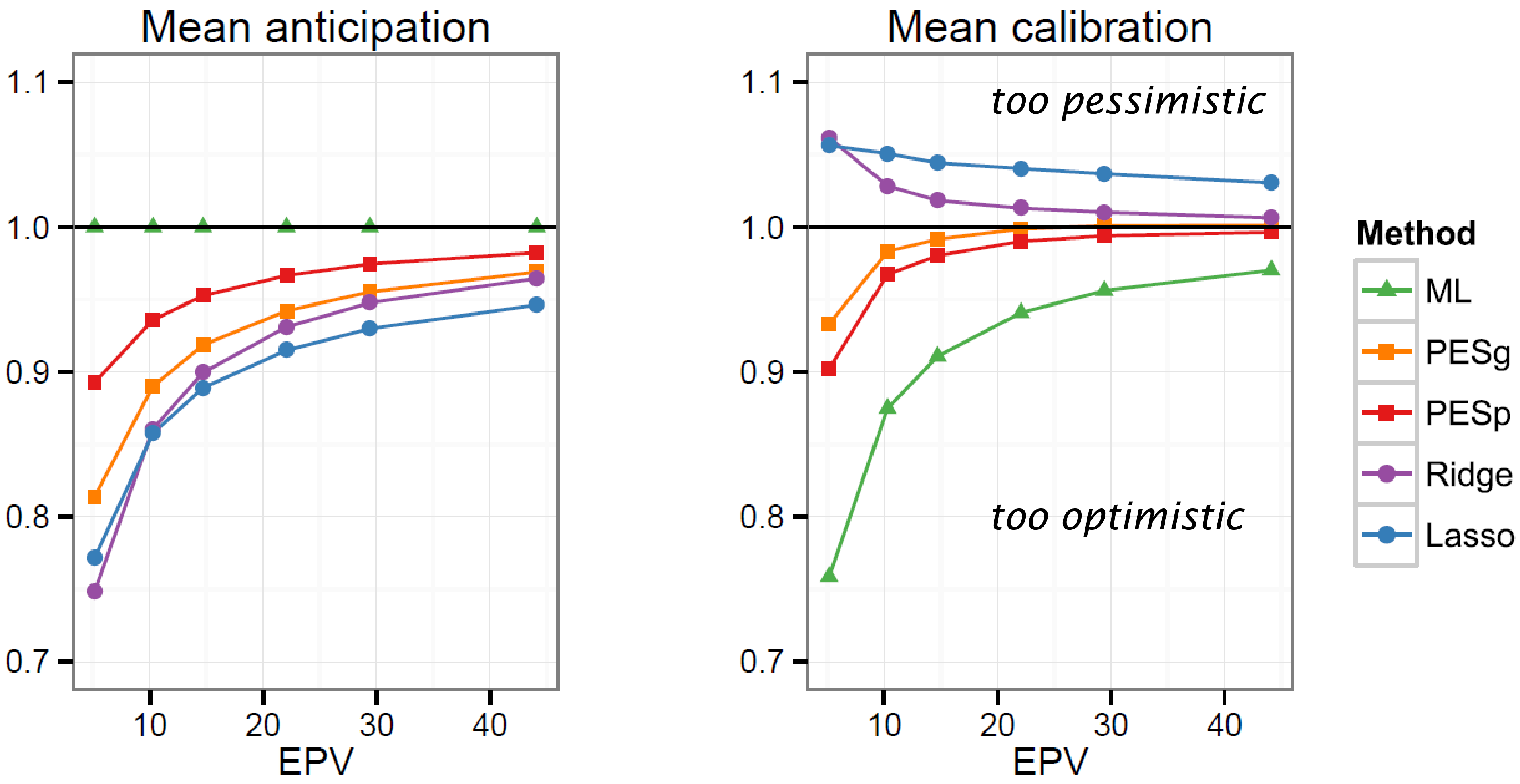


Selected scenario: Shrinkage correction over EPV

Shrinkage measured in terms of calibration of linear predictor: $h(t) = h_0(t)\exp(bX\hat{\beta})$

Anticipated shrinkage: $1/b^{M}_{\text{train}}$

Achieved calibration: $b^{M}_{\text{validation}}$



Conclusions: performance

- Penalized methods anticipate more shrinkage than parameterwise shrinkage
- Penalized methods tend to be too pessimistic, post-estimation methods too optimistic
- Penalized methods more "stable" even at very low EPV
- Backward selection usually leads to smaller models, which are often preferred by practicioners

From bias reduction to shrinkage and beyond

Joint work with Rainer Puhr, Angelika Geroldinger, Sander Greenland

Setting the scene

Logistic regression
(with fixed set of covariates,
rare events/critical events per variable ratio)

ML: small sample bias ($\hat{\beta}$)

Reduce bias: Firth correction → Biased predictions ($\hat{\pi}$)

↓

MSE reduction ($\hat{\beta}$)!

Correcting bias in $\hat{\pi}$

↓

FLAC (Puhr 2017)

↓

Bayesian methods with
weakly informative priors → Further MSE reduction ($\hat{\beta}$ and $\hat{\pi}$)!

Firth‘s penalization for logistic regression

In exponential family models with canonical parametrization the **Firth-type penalized likelihood** is given by

$$L^*(\beta) = L(\beta)\det(I(\beta))^{1/2},$$

where $I(\beta)$ is the Fisher information matrix and $L(\beta)$ is the likelihood.

Firth-type penalization

- **removes the first-order bias** of the ML-estimates of β,
- is **bias-preventive** rather than corrective,
- is available in **Software** packages such as SAS, R, Stata…

Firth's penalization for logistic regression

In logistic regression, the penalized likelihood is given by

$$L^*(\beta) = L(\beta)\,\det(X^t W X)^{1/2}, \text{ with}$$

$$W = \mathrm{diag}(\mathrm{expit}(\mathrm{X_i}\beta)(1 - \mathrm{expit}(\mathrm{X_i}\beta)))$$
$$= \mathrm{diag}(\pi_i(1 - \pi_i))\ .$$

- Firth-type estimates always exist.

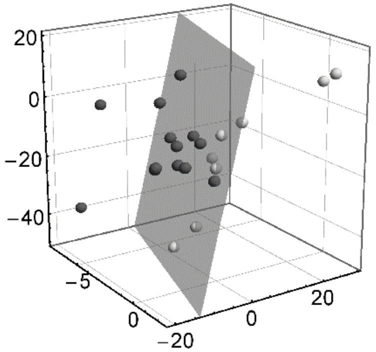
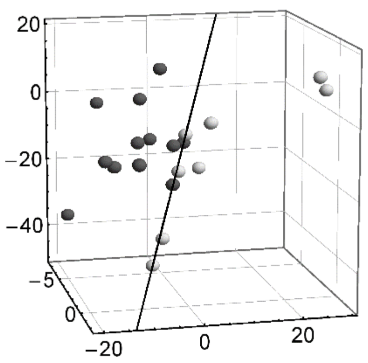
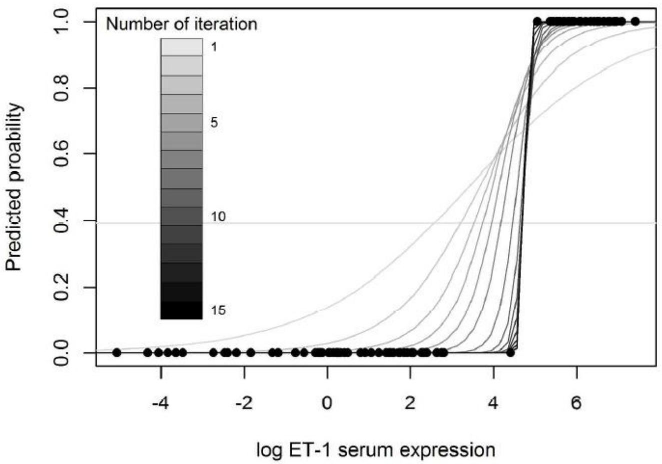
W is maximised at $\pi_i = \frac{1}{2}$, i.e. at $\beta = 0$, thus

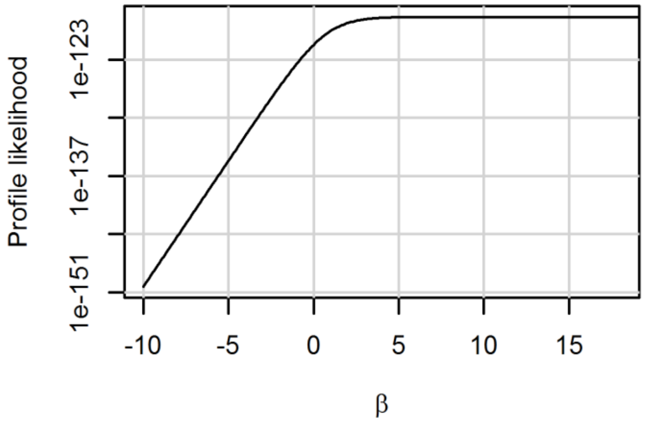
- predictions are usually pulled towards $\frac{1}{2}$,
- coefficients towards zero.

Shrinkage!

Firth‘s penalization for logistic regression

- Separation of outcome classes by covariate values (Figs. from Mansournia, 2017)



- Firth‘s bias reduction method was proposed as solution to the problem of separation in logistic regression (Heinze and Schemper, 2002
- Penalized likelihood has a unique mode
- It prevents infinite coefficients to occur

Firth‘s penalization for logistic regression

Bias reduction also leads to reduction in MSE:

- Rainey, 2017: Simulation study of LogReg for political science
  ‚Firth‘s methods dominates ML in bias and MSE‘

However, the predictions get biased…

- Elgmati et al, 2015

… and anti-shrinkage could occasionally arise:

- Greenland and Mansournia, 2015

Firth‘s Logistic regression

For logistic regression with one binary regressor*,
Firth’s bias correction amounts to adding $1/2$ to each cell:

original

	A	B
Y=0	44	4
Y=1	1	1

Firth-type penalization →

augmented

	A	B
0	44.5	4.5
1	1.5	1.5

$\text{event rate} = \frac{2}{50} = 0.04$

$OR_{BvsA} = 11$

$\text{event rate} = \frac{3}{52} \sim 0.058$

$OR_{BvsA} = 9.89$

$\text{av. pred. prob.} = 0.054$

* Generally: for saturated models

Example of Greenland 2010

original

	A	B	
Y=0	315	5	*320*
Y=1	31	1	*32*
	346	*6*	*352*

augmented

	A	B	
Y=0	315.5	5.5	*321*
Y=1	31.5	1.5	*33*
	346.5	*6.5*	*354*

$$\text{event rate} = \frac{32}{352} = 0.091$$

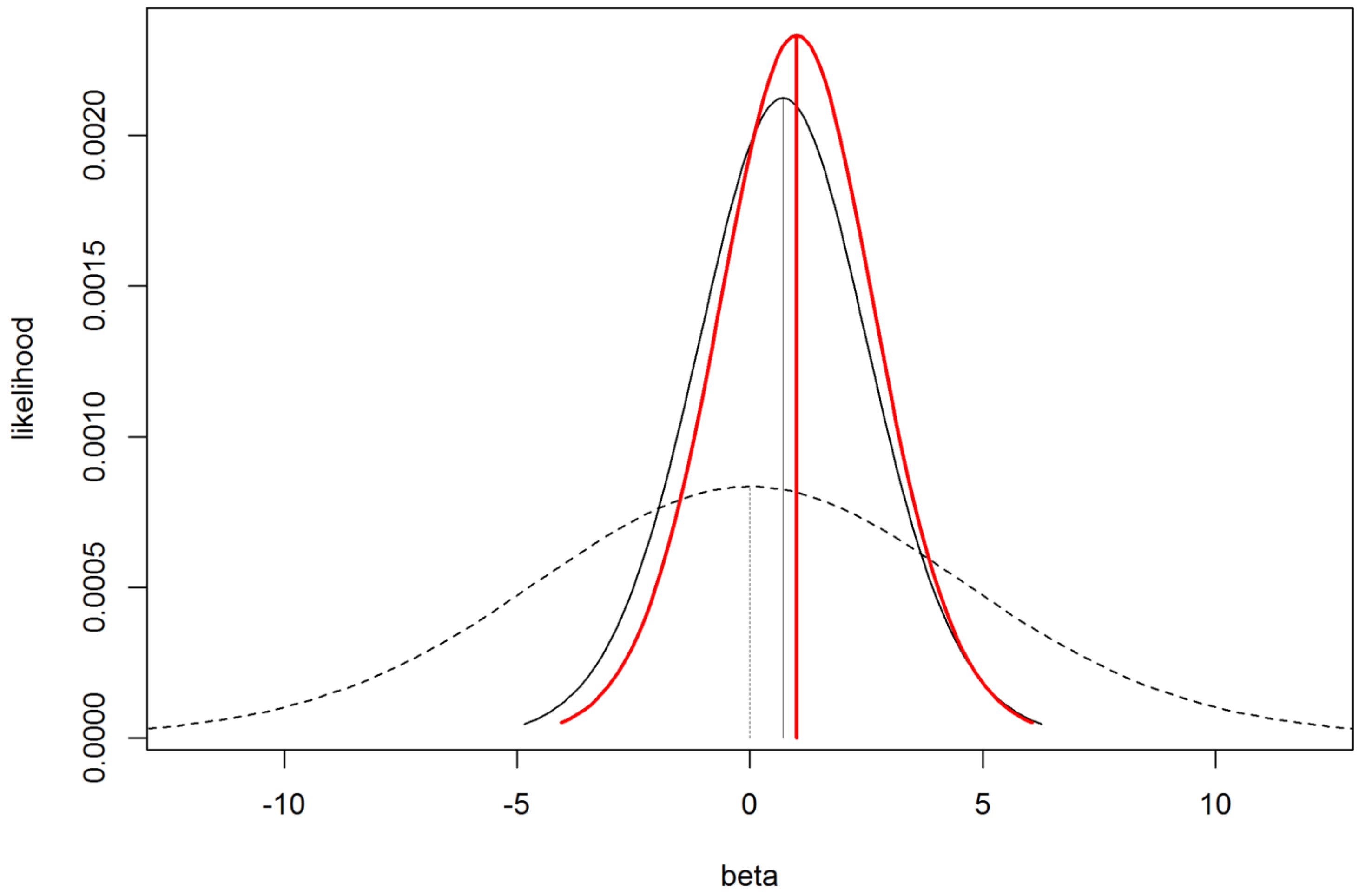
$$OR_{BvsA} = 2.03$$

$$\text{event rate} = \frac{33}{354} = 0.093$$

$$OR_{BvsA} = 2.73$$

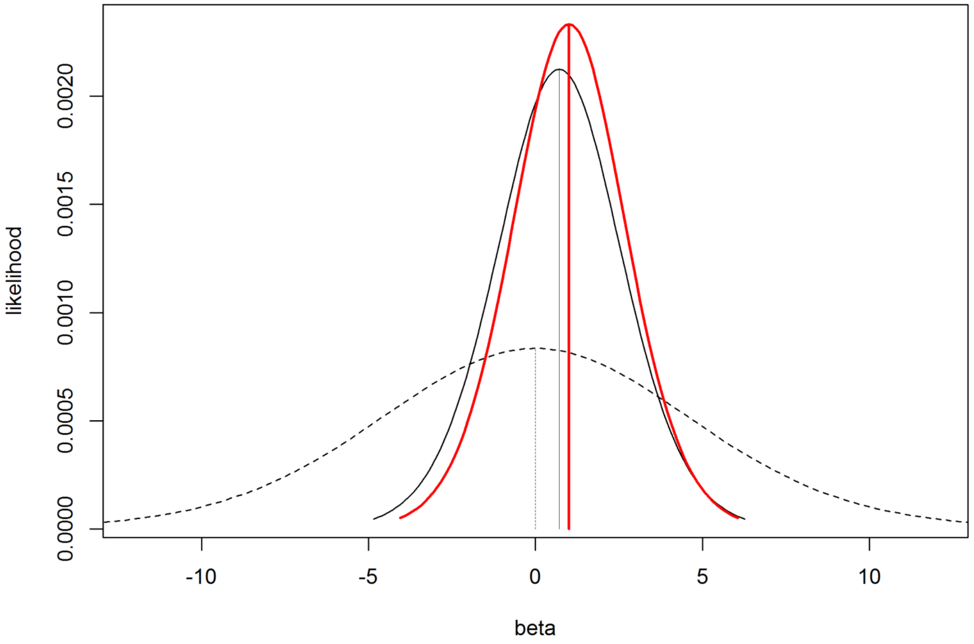
Greenland, AmStat 2010

Greenland example: likelihood, prior, posterior



Bayesian non-collapsibility: anti-shrinkage from penalization

- Prior and likelihood modes do not ‚collapse‘: posterior mode exceeds both
- The ‚shrunken‘ estimate is larger than ML estimate
- How can that happen???



An even more extreme example from Greenland 2010

- 2x2 table

	X=0	X=1	
Y=0	25	5	*30*
Y=1	5	1	*6*
	30	*6*	*36*

- Here we immediately see that the odds ratio = 1 ($\beta_1 = 0$)

- But the estimate from augmented data: odds ratio = 1.26 (try it out!)

Greenland, AmStat 2010

Simulating the example of Greenland

- We should distinguish BNC in a single data set from a systematic increase in bias of a method (in simulations)

	X=0	X=1	
Y=0	315	5	*320*
Y=1	31	1	*32*
	346	*6*	*352*

- Simulation of the example:
- Fixed groups x=0 and x=1, P(Y=1|X) as observed in example
- True log OR=0.709

Simulating the example of Greenland

- True value: log OR = 0.709

Parameter	ML	Jeffreys-Firth	
Bias β_1	*	+18%	
RMSE β_1	*	0.86	
Bayesian non-collapsibility $\boldsymbol{\beta_1}$		**63.7%**	

* Separation causes β_1 to be undefined ($-\infty$) in 31.7% of the cases

Simulating the example of Greenland

- To overcome Bayesian non-collapsibility, Greenland and Mansournia (2015) proposed not to impose a prior on the intercept
- They suggest a log-F(1,1) prior for all other regression coefficients
- The method can be used with conventional frequentist software because it uses a data-augmentation prior

Greenland and Mansournia, StatMed 2015

logF(1,1) prior (Greenland and Mansournia, 2015)

Penalizing by log-F(1,1) prior gives $L(\beta)^* = L(\beta) \cdot \prod \frac{e^{\frac{\beta_j}{2}}}{1+e^{\beta_j}}$.

This amounts to the following modification of the data set:

	x1	x2	y
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*

→

	x1	x2	y
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*
1	*	*	*
0	1	0	0
0	1	0	1
0	0	1	0
0	0	1	1

each assigned a weight of 1

each assigned a weight of ½

- No shrinkage for the intercept, no rescaling of the variables

Simulating the example of Greenland

- Re-running the simulation with the log-F(1,1) method yields:

Parameter	ML	Jeffreys-Firth	logF(1,1)
Bias β_1	*	+18%	
RMSE β_1	*	0.86	
Bayesian non-collapsibility $\boldsymbol{\beta_1}$		**63.7%**	**0%**

* Separation causes β_1 be undefined ($-\infty$) in 31.7% of the cases

Simulating the example of Greenland

- Re-running the simulation with the log-F(1,1) method yields:

Parameter	ML	Jeffreys-Firth	logF(1,1)
Bias β_1	*	+18%	-52%
RMSE β_1	*	0.86	1.05
Bayesian non-collapsibility $\boldsymbol{\beta_1}$		**63.7%**	**0%**

* Separation causes β_1 be undefined $(-\infty)$ in 31.7% of the cases

Other, more subtle occurrences of Bayesian non-collapsibility

- Ridge regression: normal prior around 0
- usually implies bias towards zero,
- But:
- With correlated predictors with different effect sizes, for some predictors the bias can be away from zero

Simulation of bivariable log reg models

- $X_1, X_2 \sim \mathrm{Bin}(0.5)$ with correlation $r = 0.8, n = 50$
- $\beta_1 = 1.5$, $\beta_2 = 0.1$, ridge parameter λ optimized by cross-validation

Parameter	ML	Ridge (CV λ)	Log-F(1,1)	Jeffreys-Firth
Bias β_1	+40% (+9%*)	-26%	-2.5%	+1.2%
RMSE β_1	3.04 (1.02*)	1.01	0.73	0.79
Bias β_2	-451% (+16%*)	+48%	+77%	+16%
RMSE β_2	2.95 (0.81*)	0.73	0.68	0.76
Bayesian non-collapsibility $\boldsymbol{\beta_2}$		**25%**	**28%**	**23%**

*excluding 2.7% separated samples

Anti-shrinkage from penalization?

Bayesian non-collapsibility/anti-shrinkage

- can be avoided in univariable models,
  but no general rule to avoid it in multivariable models
- Likelihood penalization can often decrease RMSE
  (even *with* occasional anti-shrinkage)
- **Likelihood penalization ≠ guaranteed shrinkage**

Reason for anti-shrinkage

- We look at the association of X and Y

- We could treat the source of data as a ‚ghost factor' G
- G=0 for original table
- G=1 for pseudo data

- We ignore that the conditional association of X and Y is confounded by G

Example of Greenland 2010 revisited

original

	A	B	
Y=0	315	5	*320*
Y=1	31	1	*32*
	346	*6*	*352*

augmented

	A	B	
Y=0	315.5	5.5	*321*
Y=1	31.5	1.5	*33*
	347	*7*	*352*

To overcome both the overestimation and anti-shrinkage problems:

- We propose to adjust for the confounding by including the ‚ghost factor' G in a logistic regression model

FLAC: **F**irth's **L**ogistic regression with **A**dded **C**ovariate

Split the augmented data into the original and pseudo data:

augmented

	A	B
0	315.5	5.5
1	31.5	1.5

original
G=0

	A	B
0	315	5
1	31	1

$OR_{BvsA} = 2.03$

+

pseudo
G=1 ← Ghost factor

	A	B
0	0.5	0.5
1	0.5	0.5

Define **F**irth type **L**ogistic regression with **A**dditional **C**ovariate as an analysis including the ghost factor as added covariate:

$OR_{BvsA} = 1.84$

FLAC: Firth‘s Logistic regression with Added Covariate

Beyond 2x2 tables:

Firth-type penalization can be obtained by solving modified score equations:

$$\sum_{i=1}^{N}(y_i - \pi_i)x_{ir} + h_i\left(\frac{1}{2} - \pi_i\right)x_{ir} = 0; \quad r = 0, \dots, p$$

where the h_i’s are the diagonal elements of the hat matrix $H = W^{\frac{1}{2}}X(X'WX)^{-1}XW^{1/2}$

They are equivalent to:

$$\sum_{i=1}^{N}(y_i - \pi_i)x_{ir} + \sum_{i}^{N} h_i\left(\frac{1}{2} - \pi_i\right)x_{ir} =$$

$$= \sum_{i=1}^{N}(y_i - \pi_i)x_{ir} + \sum_{i=1}^{N}\frac{h_i}{2}(y_i - \pi_i) + \sum_{i=1}^{N}\frac{h_i}{2}(1 - y_i - \pi_i) = 0$$

FLAC: **F**irth‘s **L**ogistic regression with **A**dded **C**ovariate

- A closer inspection yields:

$$\sum_{i=1}^{N}\underbrace{(y_i-\pi_i)x_{ir}}_{\text{The original data}}+\underbrace{\sum_{i=1}^{N}\underbrace{\frac{h_i}{2}(y_i-\pi_i)x_{ir}}_{\text{Original data, weighted by } h_i/2}+\sum_{i=1}^{N}\underbrace{\frac{h_i}{2}(1-y_i-\pi_i)x_{ir}}_{\text{Data with reversed outcome, weighted by } h_i/2}}_{\text{Pseudo data}}=0$$

FLAC: **F**irth‘s **L**ogistic regression with **A**dded **C**ovariate

- A closer inspection yields:

$$\sum_{i=1}^{N}\underbrace{(y_i-\pi_i)x_{ir}}_{\text{The original data}}+\underbrace{\sum_{i=1}^{N}\frac{h_i}{2}\underbrace{(y_i-\pi_i)x_{ir}}_{\text{Original data, weighted by } h_i/2}+\sum_{i=1}^{N}\frac{h_i}{2}\underbrace{(1-y_i-\pi_i)x_{ir}}_{\text{data with reversed outcome, weighted by } h_i/2}}_{\text{Pseudo data}}=0$$

Ghost factor: G=0 G=1
(‚Added covariate‘)

FLAC: Firth's Logistic regression with Added Covariate

FLAC estimates can be obtained by the following steps:

1) Define an indicator variable discriminating between original and pseudo data.
2) Apply ML on the augmented data including the indicator.

✓ unbiased pred. probabilities

FLIC

Firth type **L**ogistic regression with **I**ntercept **C**orrection:

1. Fit a Firth logistic regression model
2. Modify the intercept in Firth-type estimates such that the average pred. prob. becomes equal to the observed proportion of events.

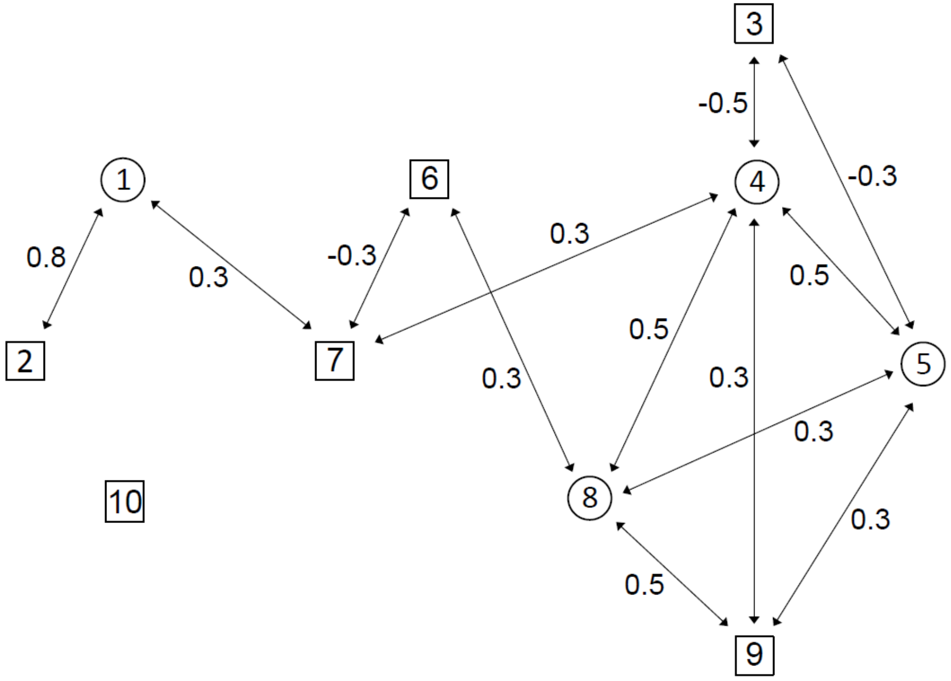
✓ unbiased pred. probabilities

effect estimates are the same as in Firth type logistic regression

Simulation study: the set-up

We investigated the performance of FLIC and FLAC, simulating 1000 data sets for 45 scenarios with:

- 500, 1000 or 1400 observations,
- event rates of 1%, 2%, 5% or 10%
- 10 covariables (6 cat., 4 cont.),
  see Binder et al., 2011
- none, moderate and strong effects
  of positive and mixed signs



Main evaluation criteria:

bias and RMSE of predictions and effect estimates

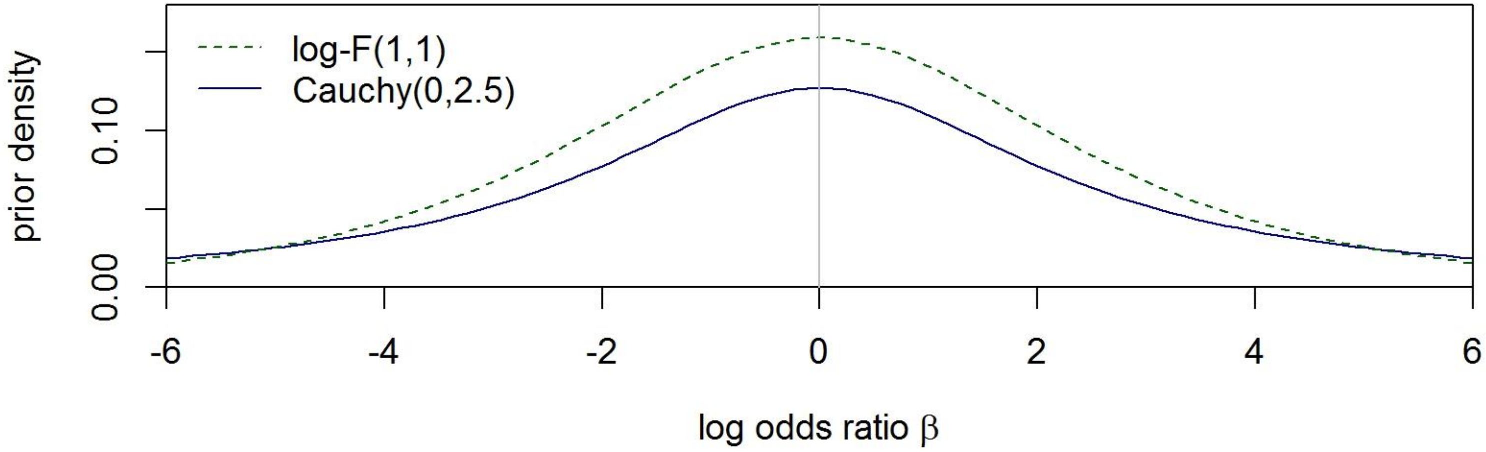
Other methods for accurate prediction

In our simulation study, we compared FLIC and FLAC to the following methods:

- weakened Firth-type penalization (Elgmati 2015),
  with $L(\beta)^* = L(\beta)\det(X^t W X)^\tau$, $\tau = 0.1$, WF
- ridge regression, RR
- penalization by log-F(1,1) priors, LF
- penalization by Cauchy priors with scale parameter=2.5. CP

Cauchy priors (CP)

Cauchy priors (scale=2.5) have heavier tails than log-F(1,1)-priors:



We follow Gelman 2008:

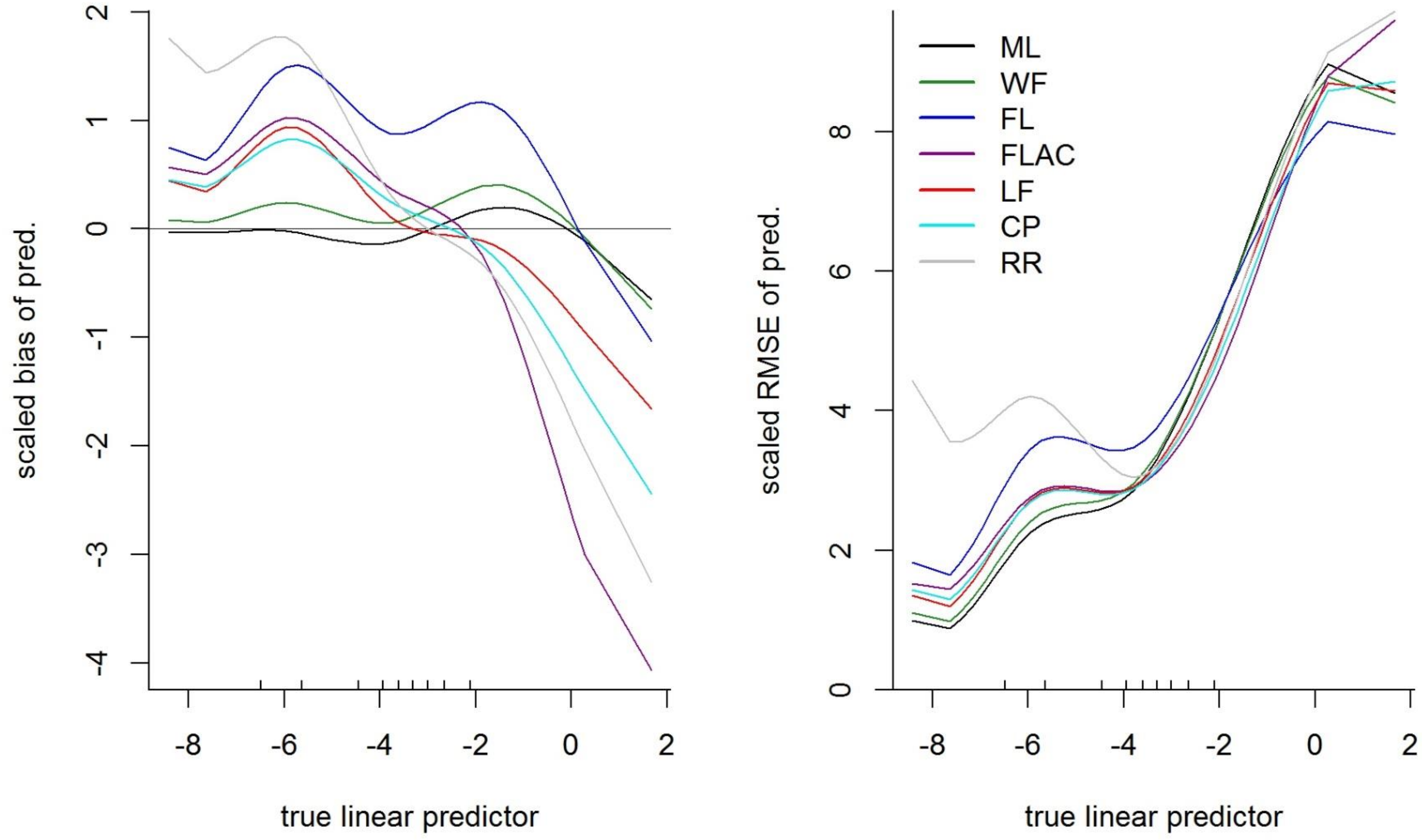
- all variables are centered,
- binary variables are coded to have a range of 1,
- all other variables are scaled to have standard deviation 0.5,
- the intercept is penalized by Cauchy(0,10).

This is implemented in the function `bayesglm` in the R-package `arm`.

Simulation results

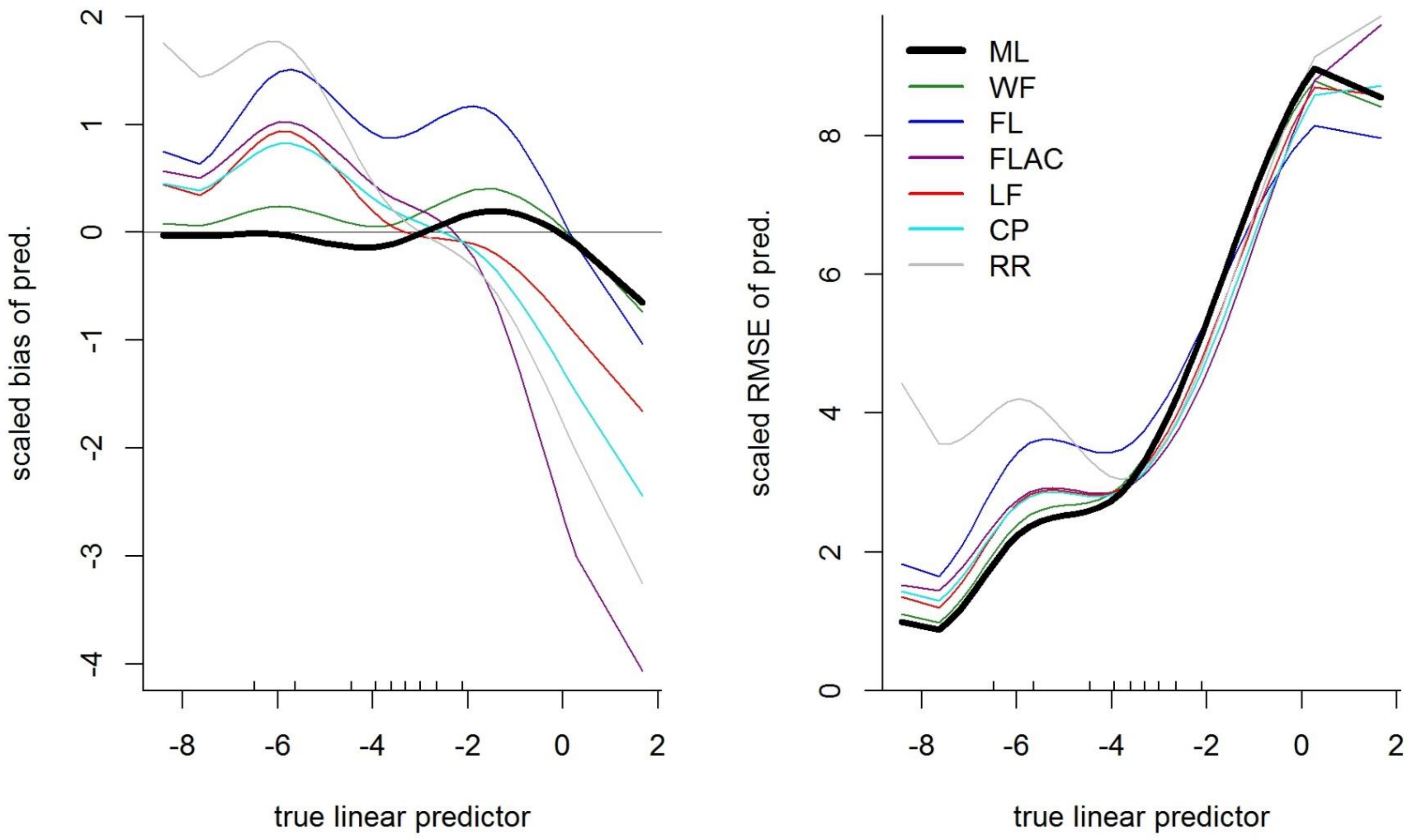
- Bias of $\hat{\beta}$: clear winner is Firth method
  FLAC, logF, CP: slight bias towards 0
- RMSE of $\hat{\beta}$:
  equal effect sizes: ridge the winner
  unequal effect sizes: very good performance of FLAC and CP closely followed by logF(1,1)
- Calibration: often FLAC the winner; considerable instability of ridge
- Bias and RMSE of $\hat{\pi}$: see following slides

Predictions: bias RMSE



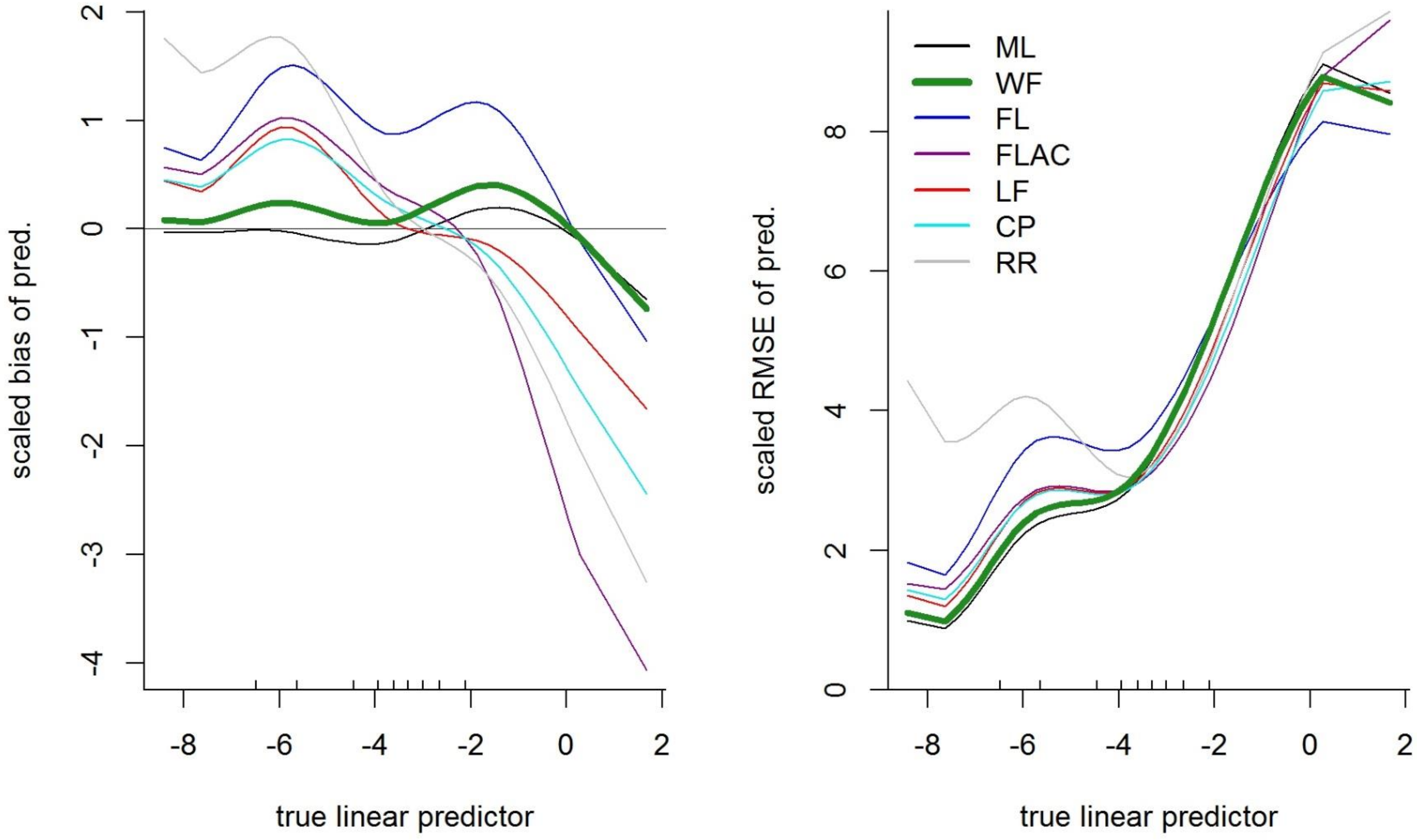
N=500, a=1, ybar=0.05, b.sign=-1

Predictions: bias RMSE



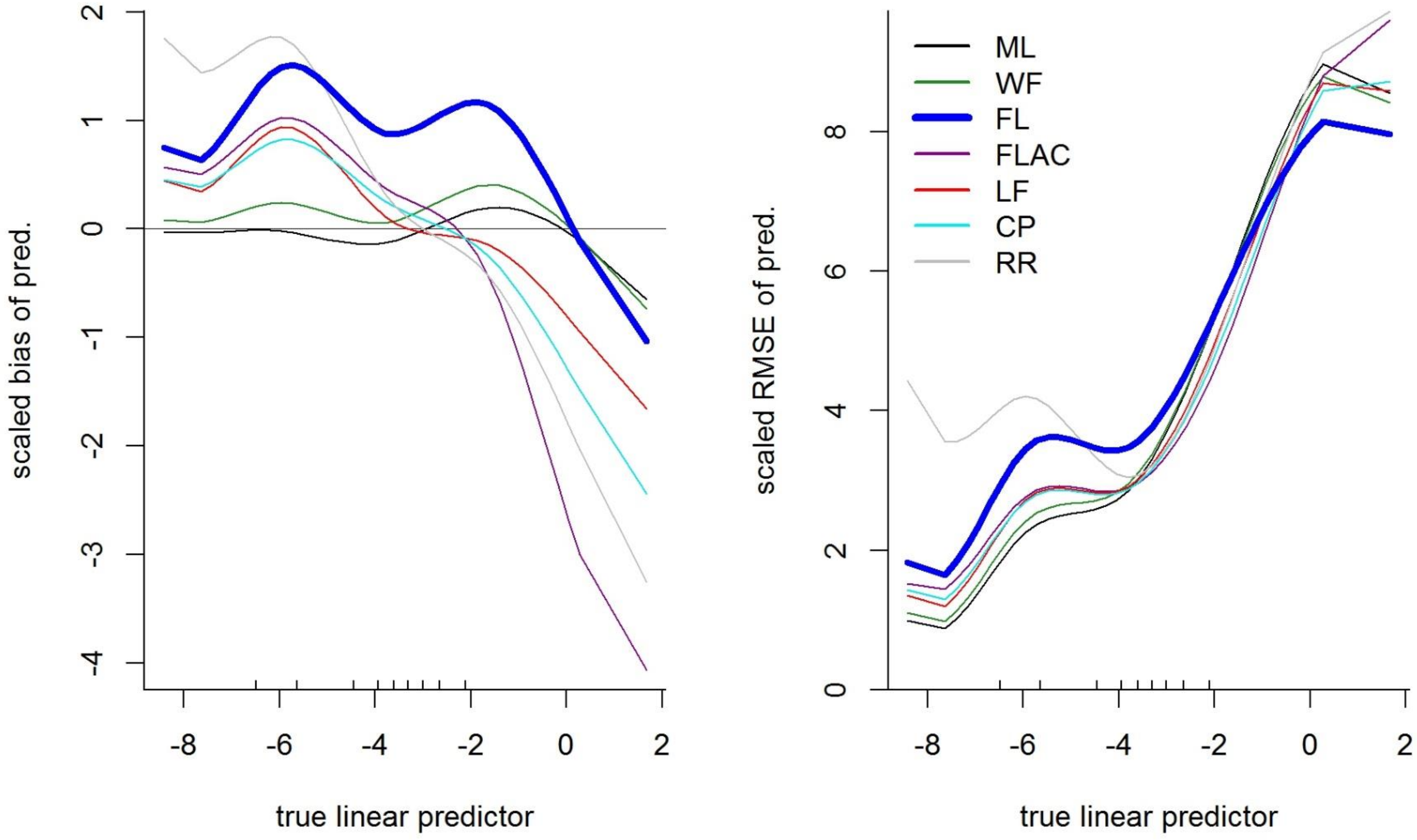
N=500, a=1, ybar=0.05, b.sign=-1

Predictions: bias RMSE



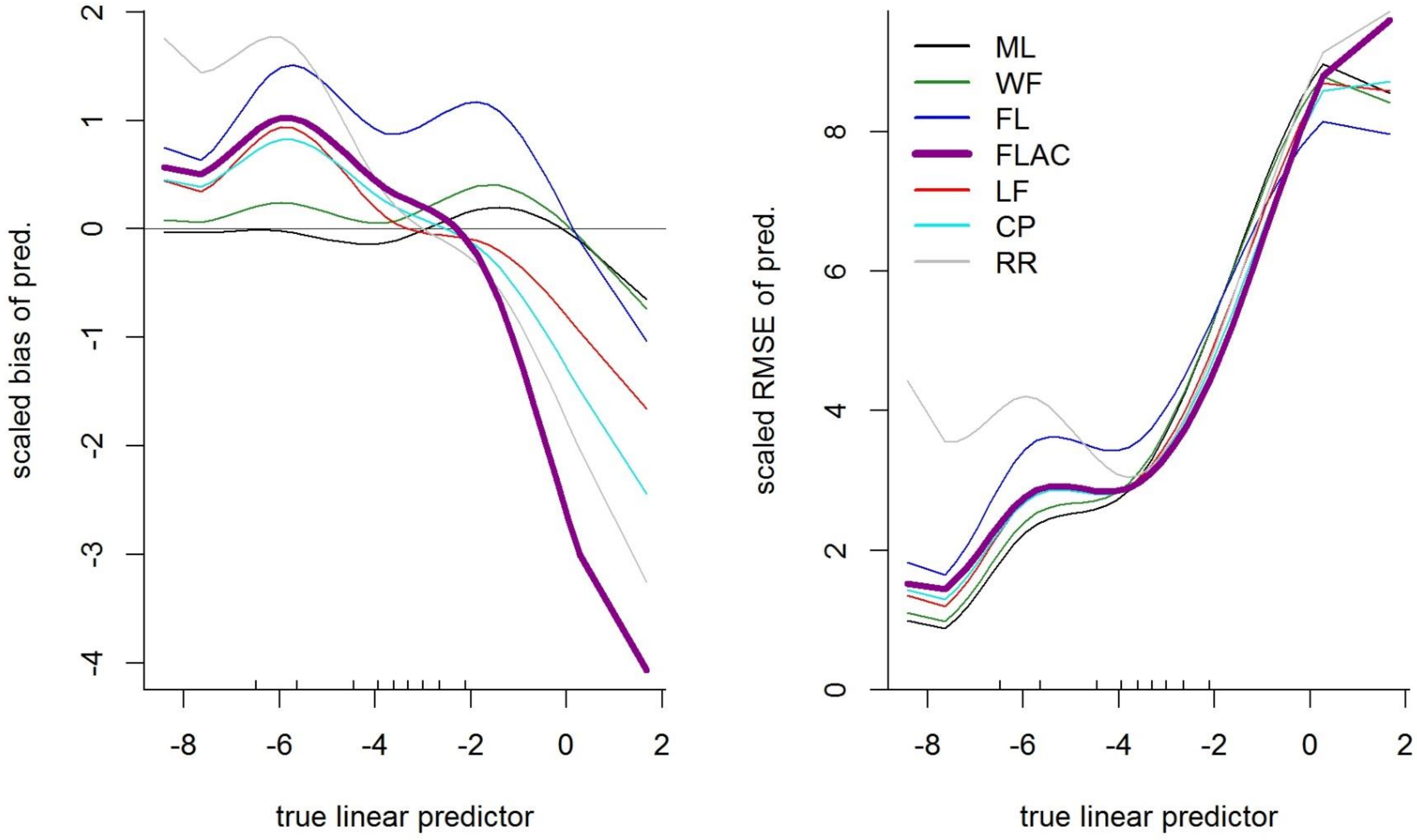
N=500, a=1, ybar=0.05, b.sign=-1

Predictions: bias RMSE



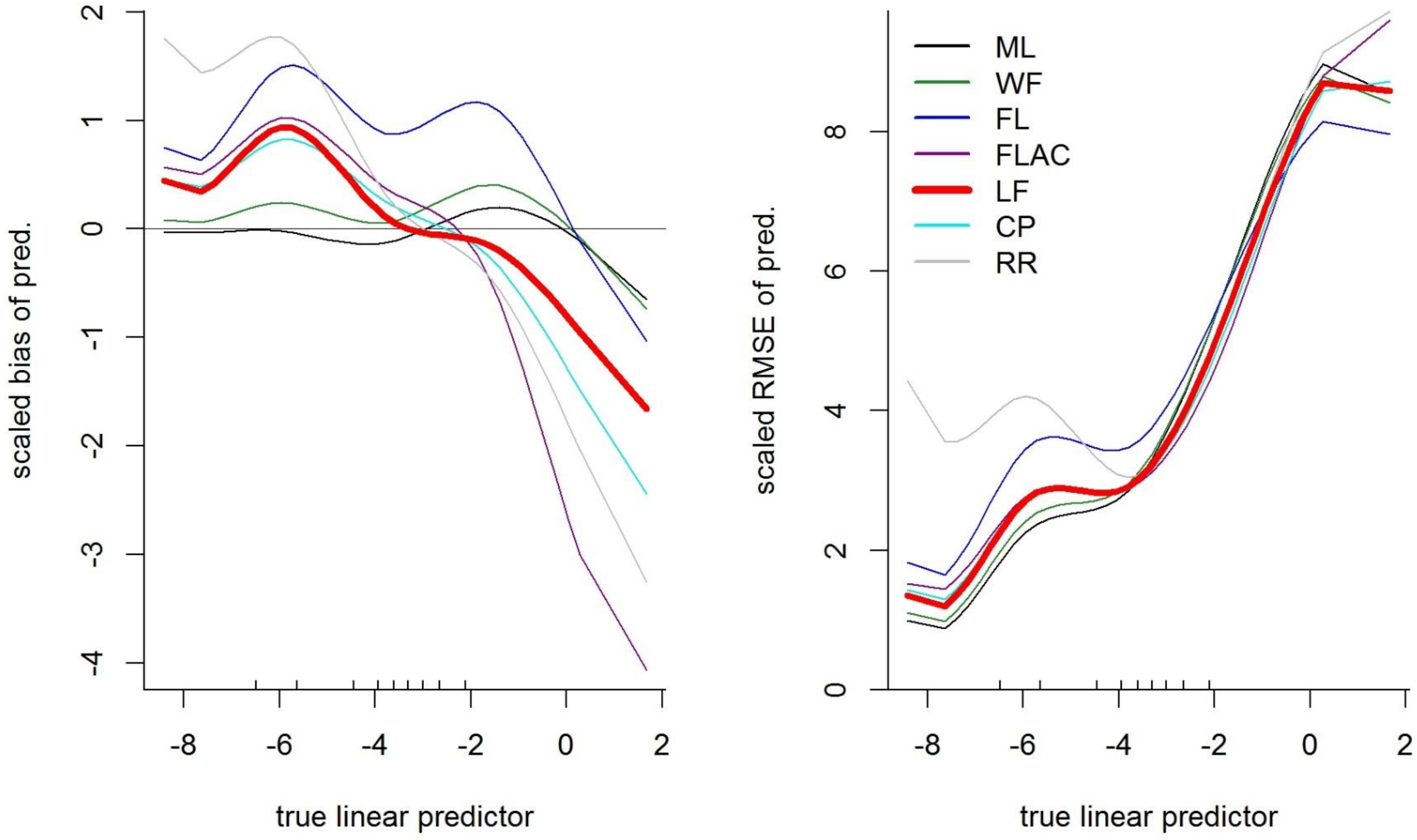
N=500, a=1, ybar=0.05, b.sign=-1

Predictions: bias RMSE



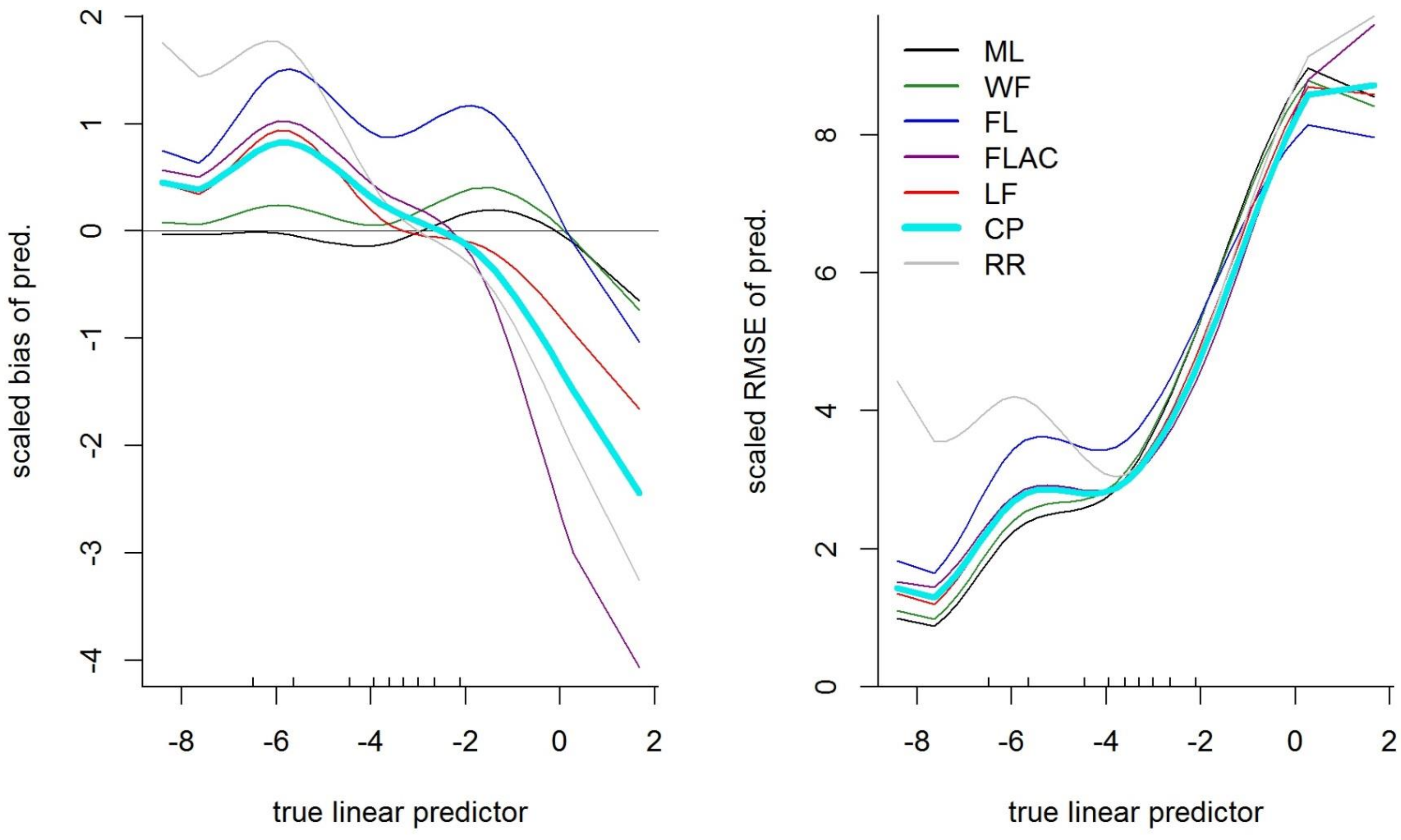
N=500, a=1, ybar=0.05, b.sign=-1

Predictions: bias RMSE



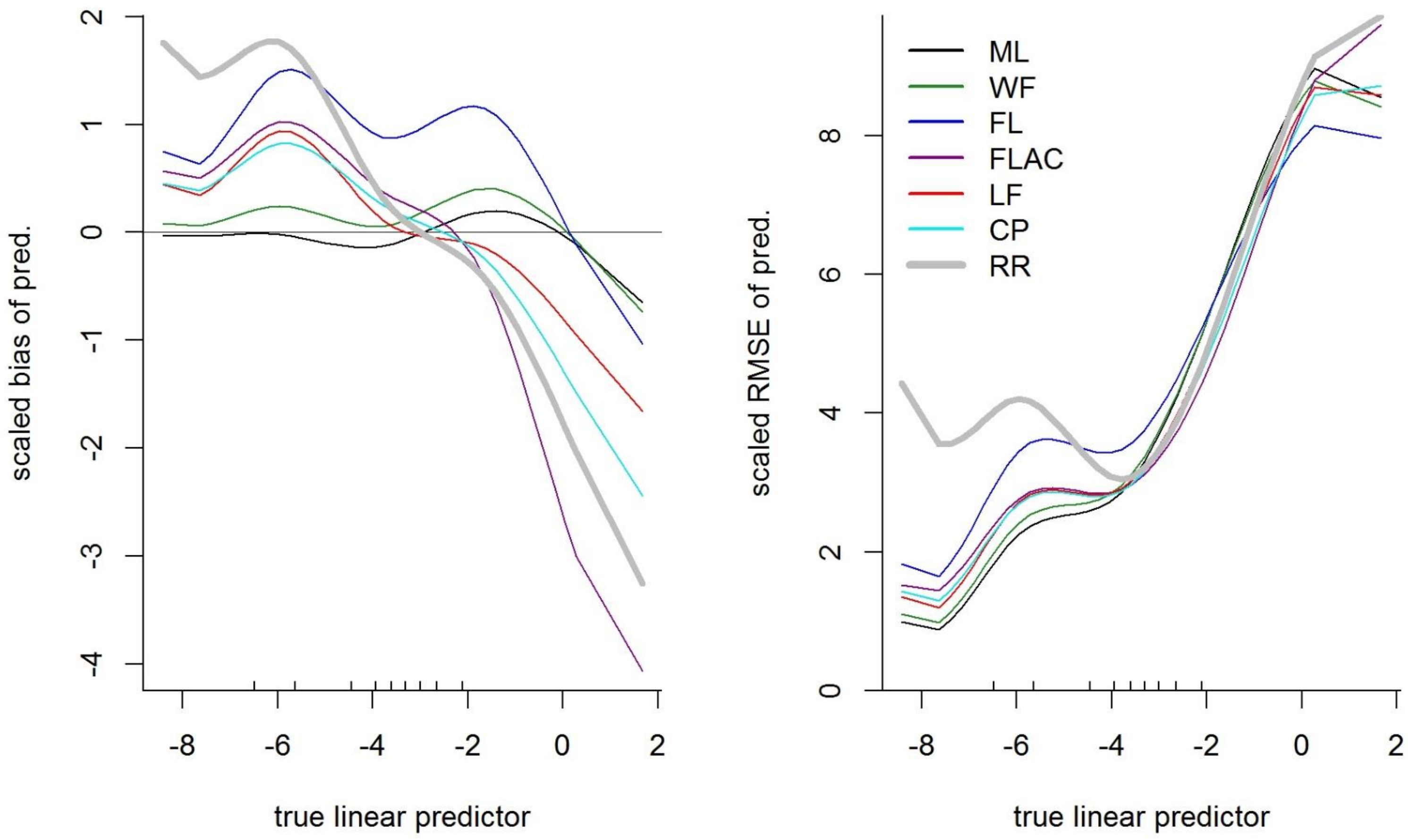
N=500, a=1, ybar=0.05, b.sign=-1

Predictions: bias RMSE



N=500, a=1, ybar=0.05, b.sign=-1

Predictions: bias RMSE



N=500, a=1, ybar=0.05, b.sign=-1

Comparison

FLAC

- No tuning parameter
- Transformation-invariant
- Often best MSE, calibration

Ridge

- Standardization is standard
- Tuning parameter
  – no confidence intervals
- Not transformation-invariant
- Performance decreases
  if effects are very different

Bayesian methods (CP, logF)

- CP: in-built standardization,
  no tuning parameter
- logF(m,m): choose m by '95% prior region' for parameter of interest
  m=1 for wide prior, m=2 less vague
- (in principle, m could be tuned as in ridge)
- logF: easily implemented
- CP and logF are not transformation-invariant

Confidence intervals

It is important to note that:

- With penalized (=shrinkage) methods one cannot achieve nominal coverage over all possible parameter values
- But one can achieve nominal coverage averaging over the implicit prior

- Prior – penalty correspondence can be *a-priori* established if there is no tuning parameter
- Important to use profile penalized likelihood method
- Wald method ($\hat{\beta} \pm 1.96\,SE$) depends on unbiasedness of estimate

Gustafson&Greenland, StatScience 2009

Conclusion

Part 1:
Prediction under model uncertainty

- Variable selection should be accompanied by shrinkage factor estimation
- BW–PESp unless EPV ratio very low
- PESp can also reveal modeling problems

Part 2:
Prediction under sparsity (fixed model)

- We recommend FLAC for:
- Good performance
- Invariance to transformations or coding
- Cannot be 'outsmarted' by creative coding

References

* Dunkler D, Sauerbrei W, Heinze G. Global, parameterwise and joint shrinkage factor estimation. Journal of Statistical Software 2016, 69(8).

* Puhr R, Heinze G, Nold M, Lusa L, Geroldinger A. Firth‘s logistic regression with rare events – accurate effect estimates and predictions? Statistics in Medicine 2017, early view.

Please cf. the reference lists therein for all other citations of this presentation.

Further references:

- Gustafson P, Greenland S. Interval estimation for messy observational data. Statistical Science 2009, 24:328-342.
- Mansournia M, Geroldinger A, Greenland S, Heinze G. Separation in logistic regression – causes, consequences and control. Submitted, 2017.
- Rainey C. Estimating logit models with small samples. www.carlislerainey.com/papers/small.pdf (27 March 2017)