Leave-one-out crossvalidation favors inaccurate estimators

Angelika Geroldinger, Lara Lusa, Mariana Nold, Georg Heinze

8–12 May 2017, Thessaloniki

This work is supported by FWF under project number I 2276 (“PREMA”).

Motivation

Study on rare events in logistic regression,

compared 10 different model estimation methods

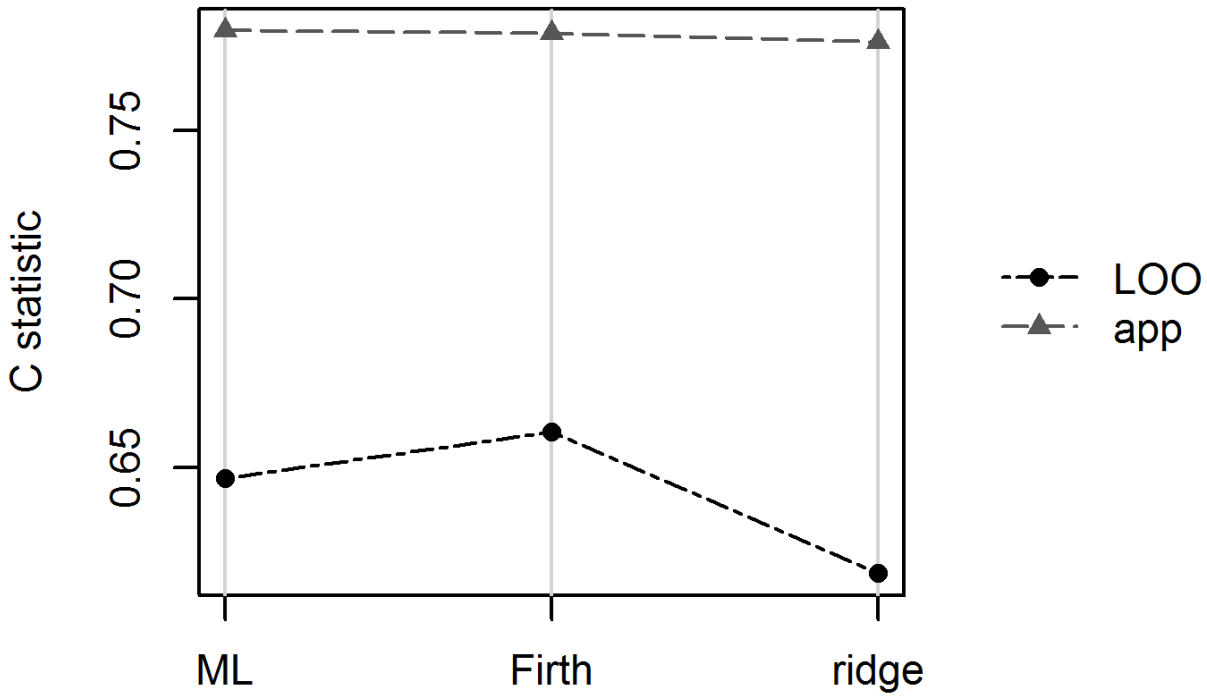
Real data example: arterial closure devices in minimally invasive cardiac surgery

		Type of surgical access	
		conven-tional	arterial closure device
Complication	no	82	342
	yes	8	8

Motivation

Simulations: all 10 model estimation methods similar with respect to discrimination ability (c statistic)

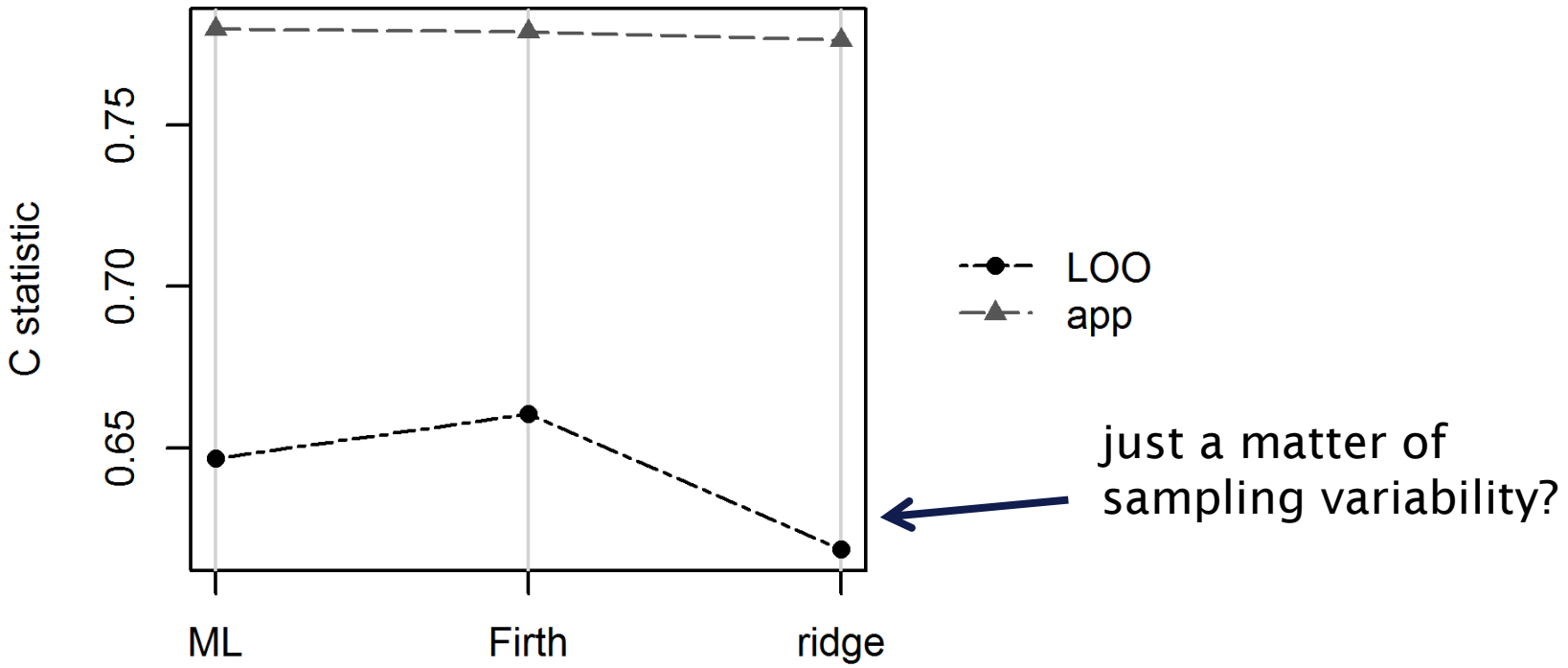
In the real data example leave-one-out crossvalidation (LOO CV) gave:



Motivation

Simulations: all 10 model estimation methods similar with respect to discrimination ability (c statistic)

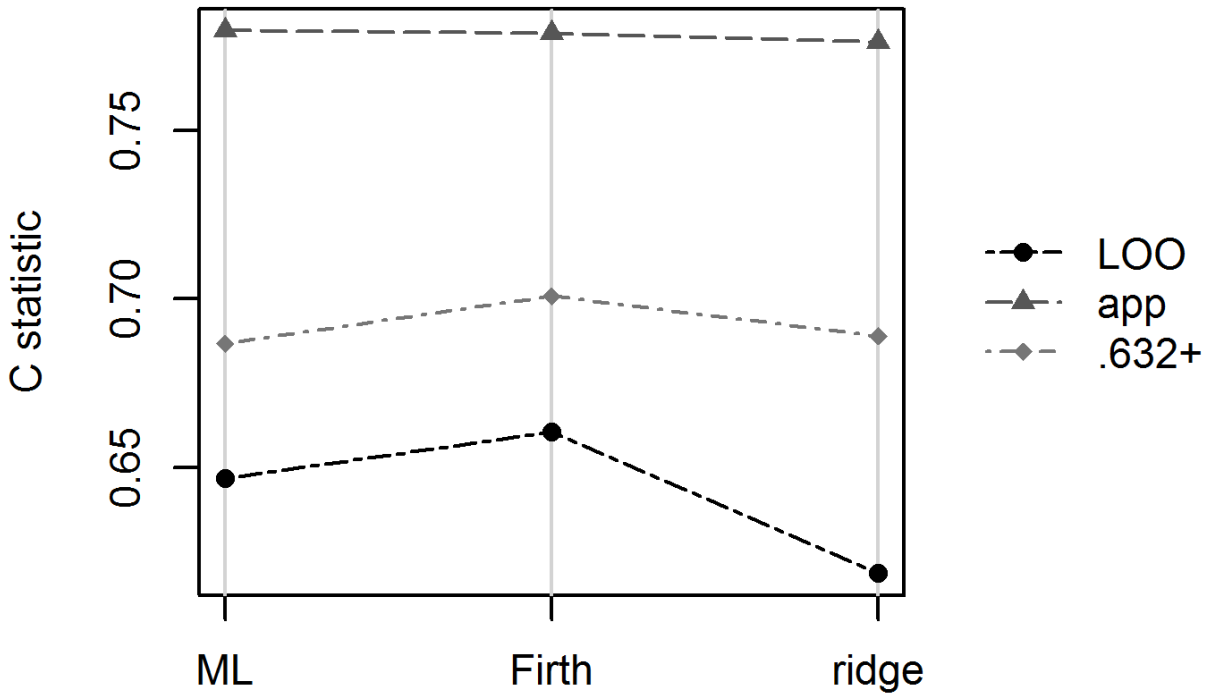
In the real data example leave-one-out crossvalidation (LOO CV) gave:



Motivation

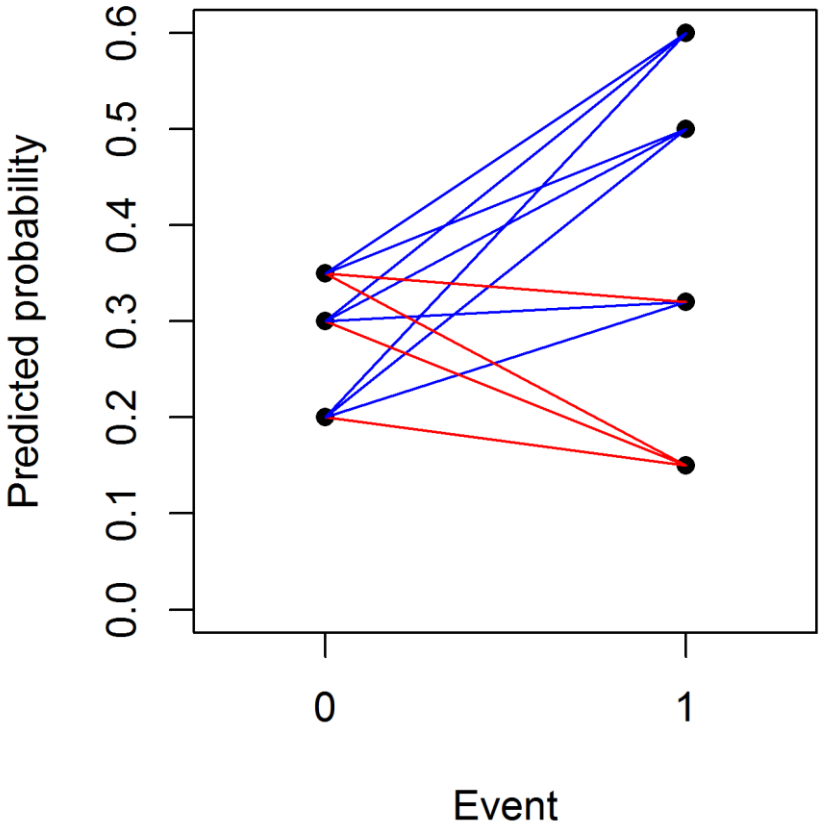
Simulations: all 10 model estimation methods similar with respect to discrimination ability (c statistic)

Using the .632+ bootstrap gives different results:



Definitions

C statistic: proportion of pairs with opposite outcomes, which are ranked correctly by the model (equal to AUC of ROC curves)

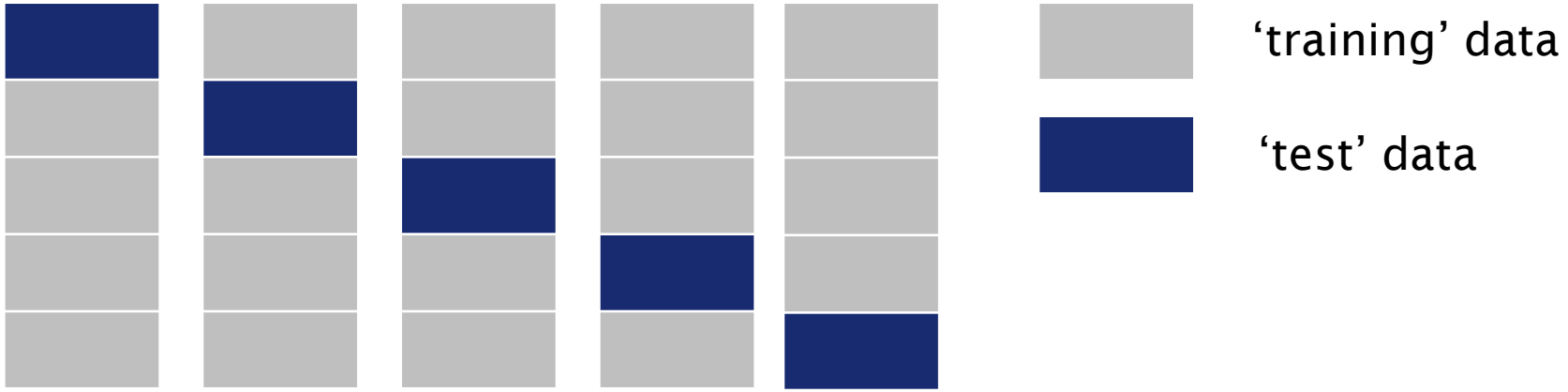


c statistic of $\frac{8}{4*3} = 0.667$

The c-statistic does not have to be ≥0.5!

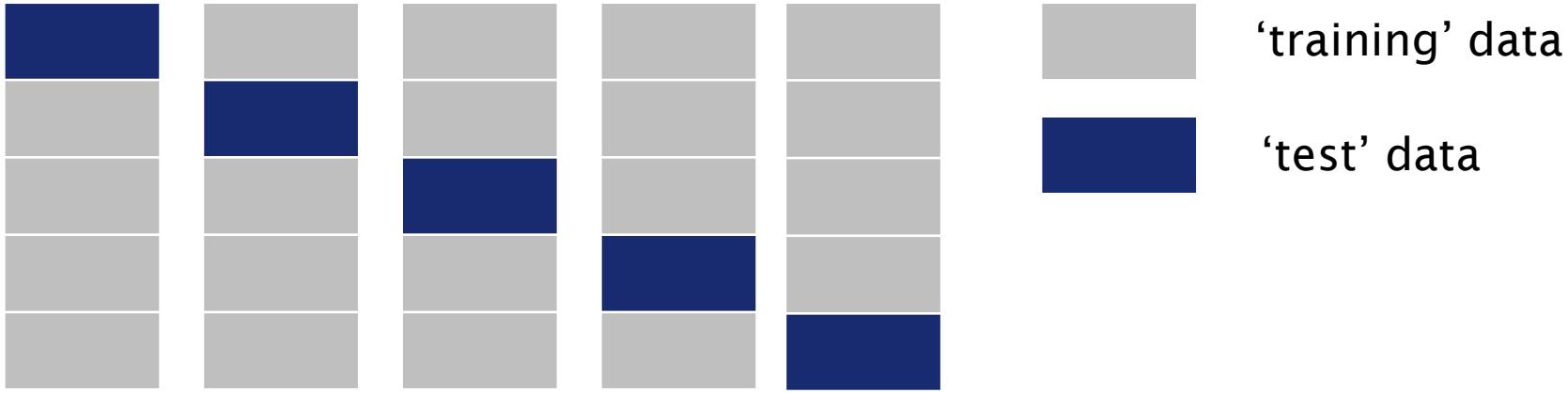
Definitions

C statistics with 5-fold CV:



Definitions

C statistics with 5-fold CV:



With LOO CV: only one observation per fold,

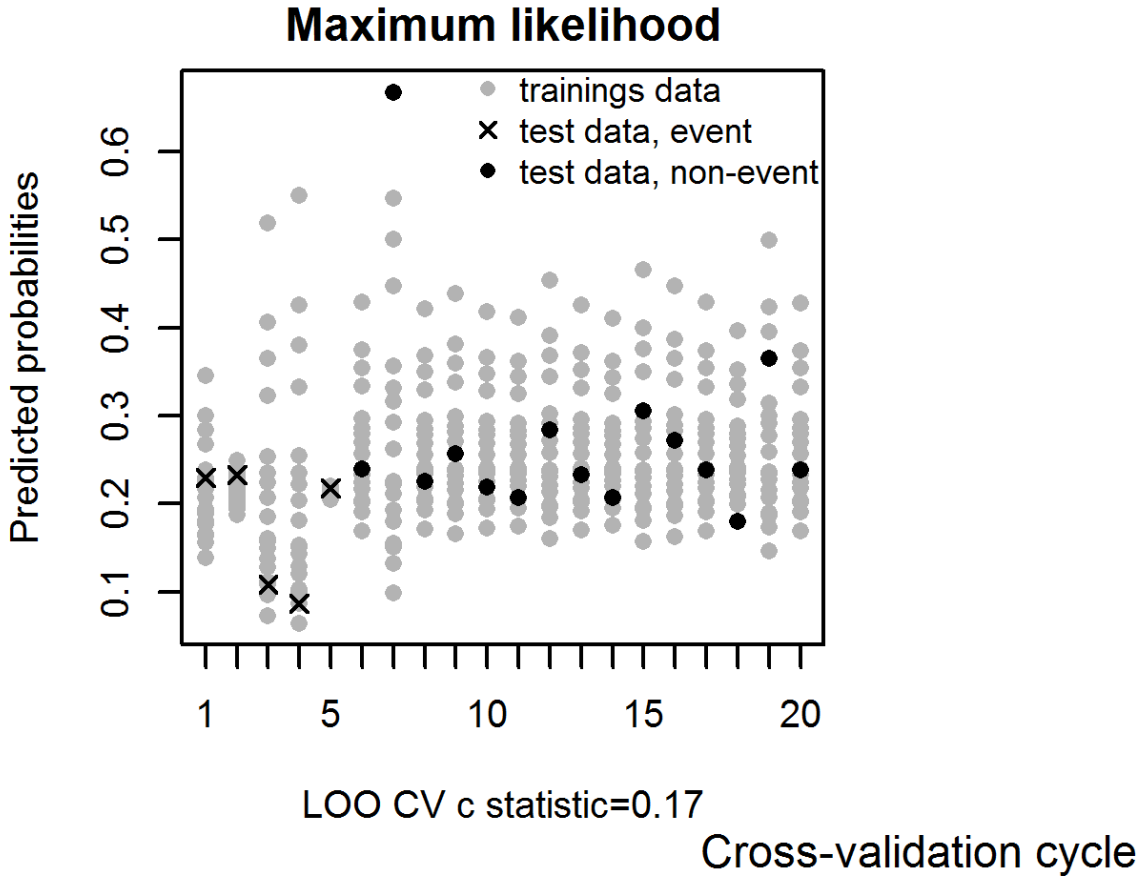
calculate pred. probability on test data,

calculate the c statistic for the pooled pred. probabilities.

The problem with LOO CV

20 observations of one explanatory variable (N(0,1)),

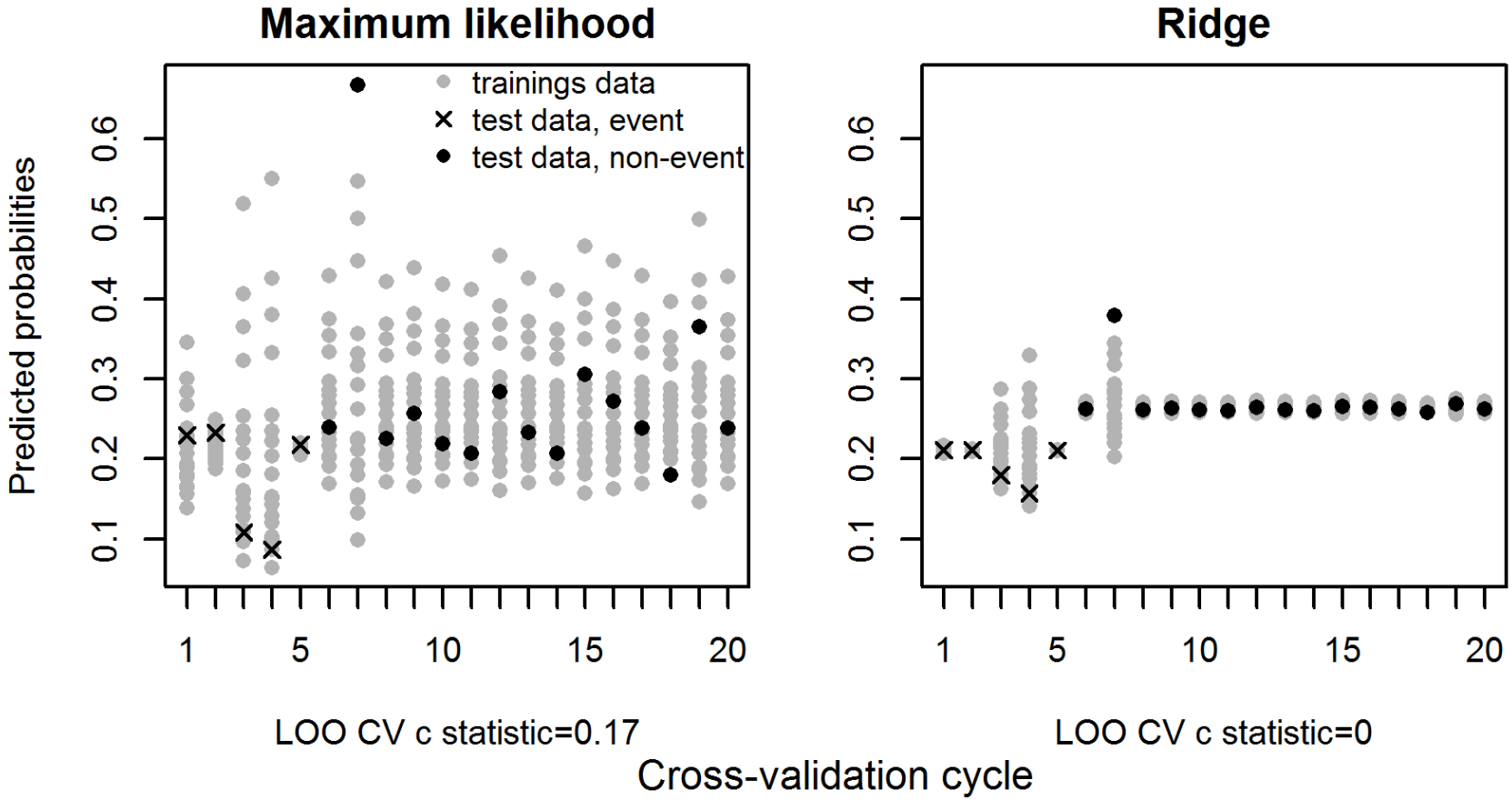
independent outcome with 5 events.



The problem with LOO CV

20 observations of one explanatory variable (N(0,1)),

independent outcome with 5 events.



Leave-pair-out crossvalidation (LPO CV)

Algorithm:

- Each pair of observations with opposite outcomes is used as test data once.
- Calculate the proportion of pairs, where the ranking is correct.

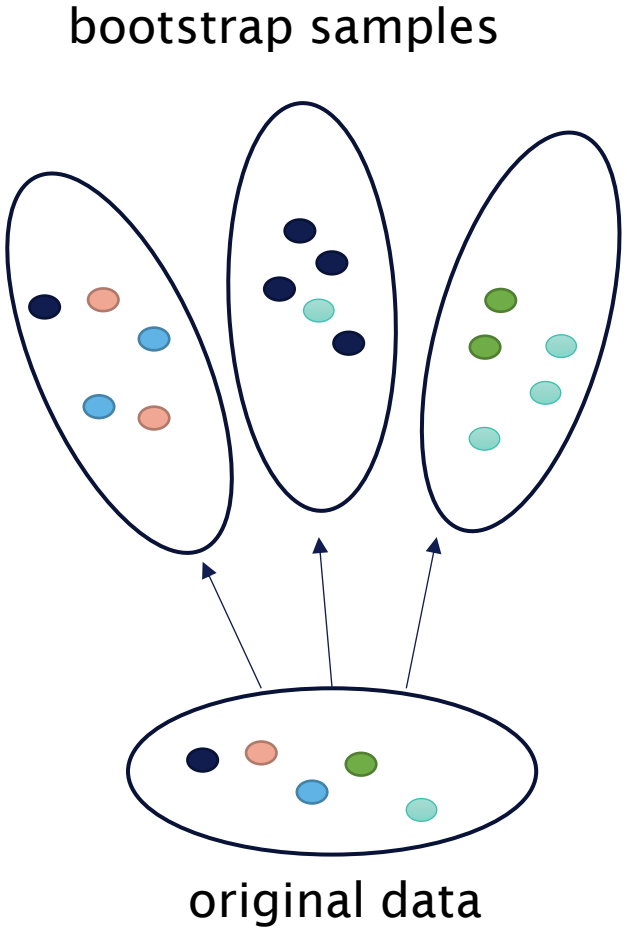
(Airola et al., 2011)

- tailored for the estimation of c statistics,
- high computational burden: (#events * #non-events) models to fit,
- not depending on random sampling.

Bootstrap

Enhanced bootstrap (Harrell, 2001):

1. Fit the model on bootstrap sample.
2. Calculate the c statistic using the model from 1. in the bootstrap sample and the original data.
3. The difference is the estimated "optimism".
4. Subtract the optimism averaged over multiple bootstrap samples from the apparent estimator.



Bootstrap

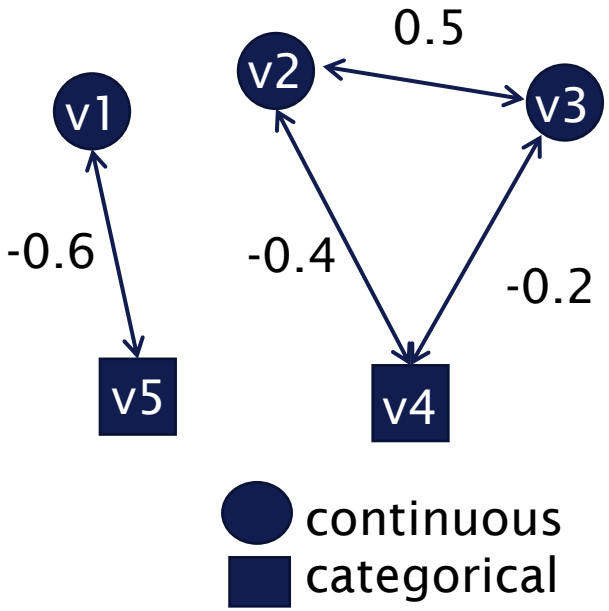
.632+ bootstrap (Efron, 1997):

- weighted average of the apparent c statistic and c statistic calculated from the omitted observations of the bootstrap sample,
- puts less weight on the apparent c statistic than the .632 bootstrap.

Simulation study: set up

We evaluated the performance of the resampling methods, simulating 1000 data sets for 12 scenarios with:

- 50 or 100 observations,
- event rates of 0.25 or 0.5,
- 5 covariables (2 cat., 3 cont.), see Binder et al., 2011,
- none, moderate and strong effects.



Main evaluation criteria:

mean difference and root mean squared difference to true value

Simulation study: set up

We consider the following resampling methods

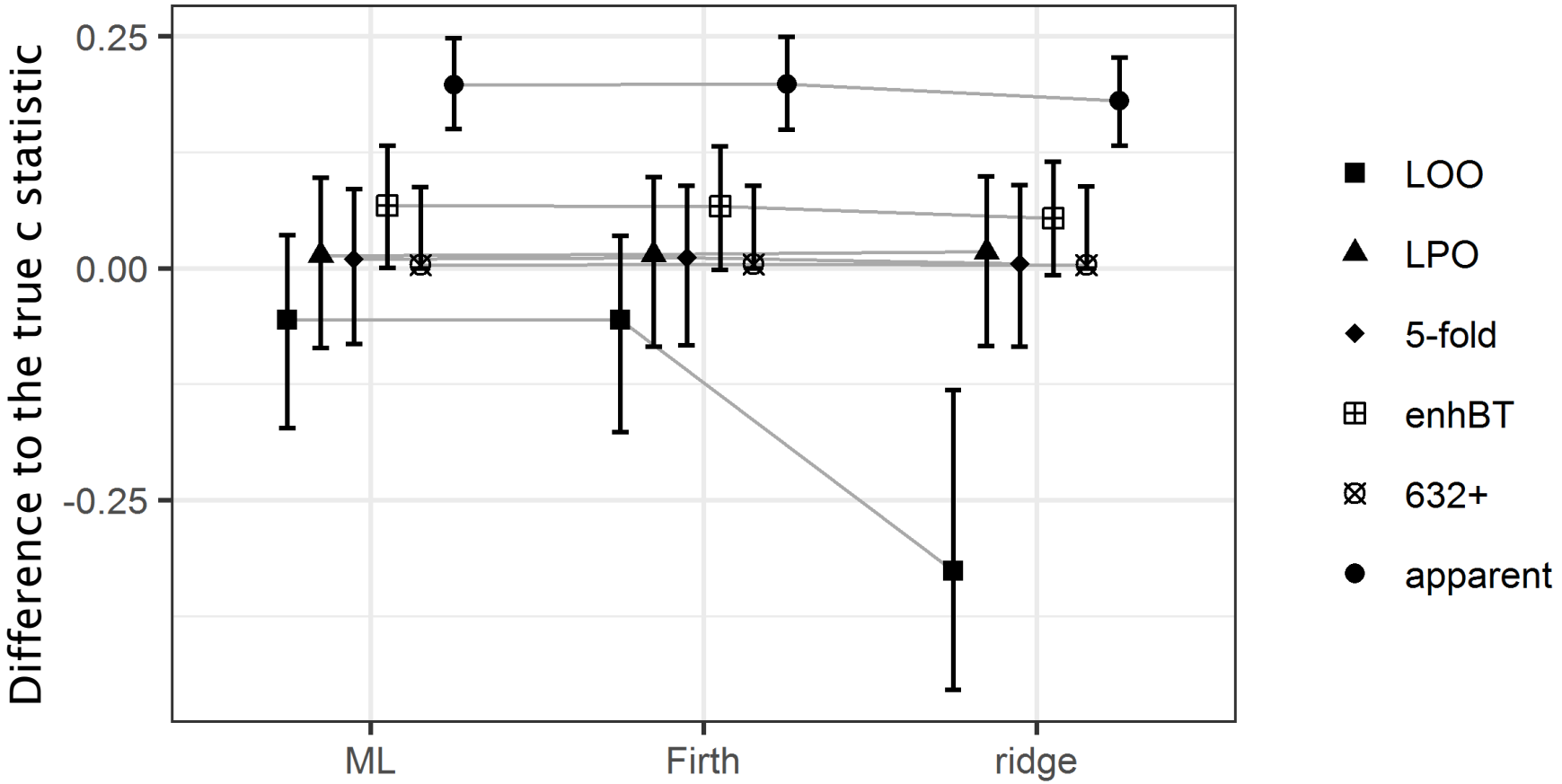
- 5-fold CV,
- LOO,
- LPO,
- enhanced bootstrap,
- .632+ bootstrap,

in combination with the following model estimation methods

- ML,
- Firth‘s penalization,
- ridge regression with AIC as tuning criterion.

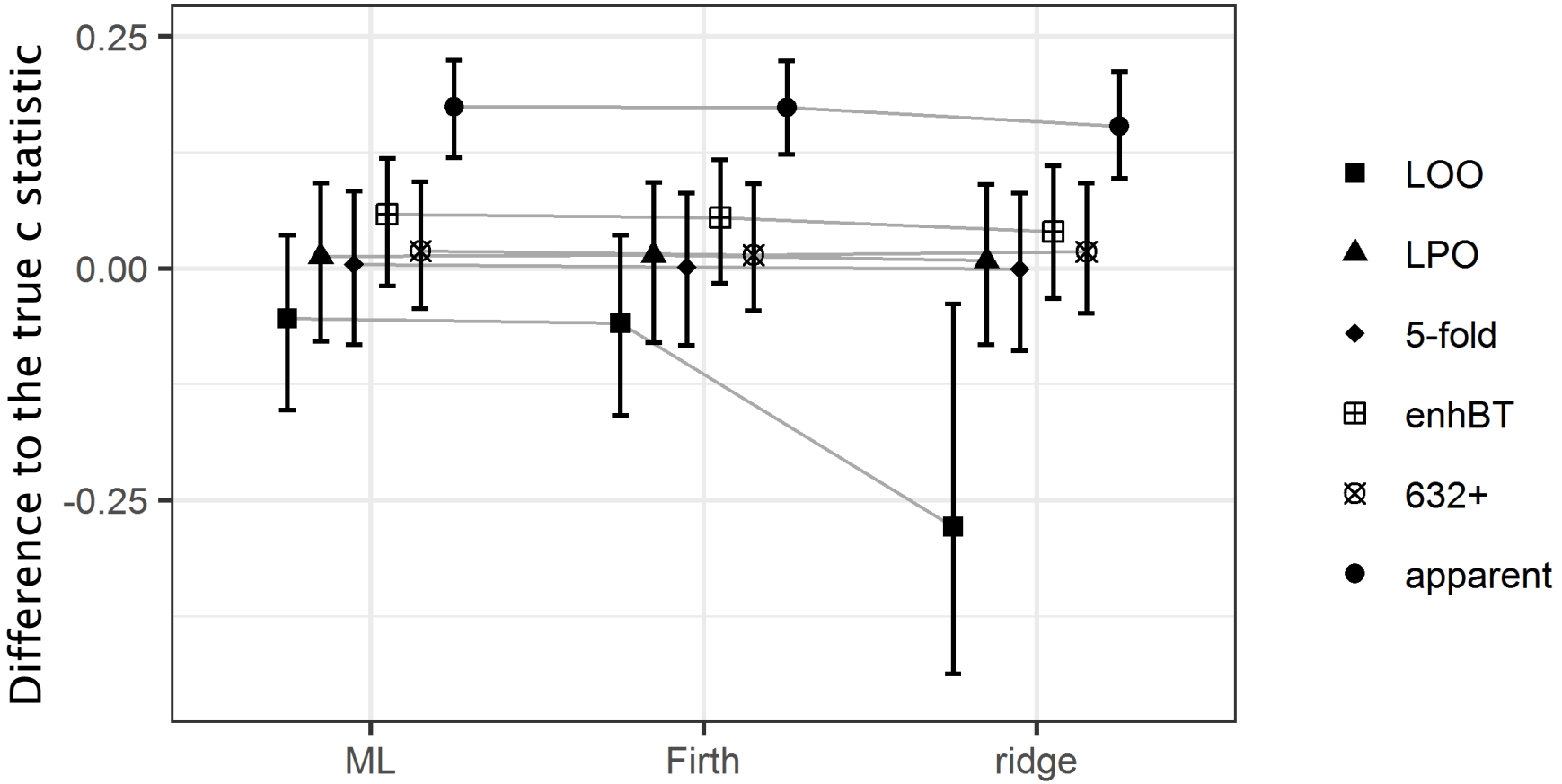
Simulation results

Median (IQR) difference to the true c statistic
N=50, event rate=0.25, no effect



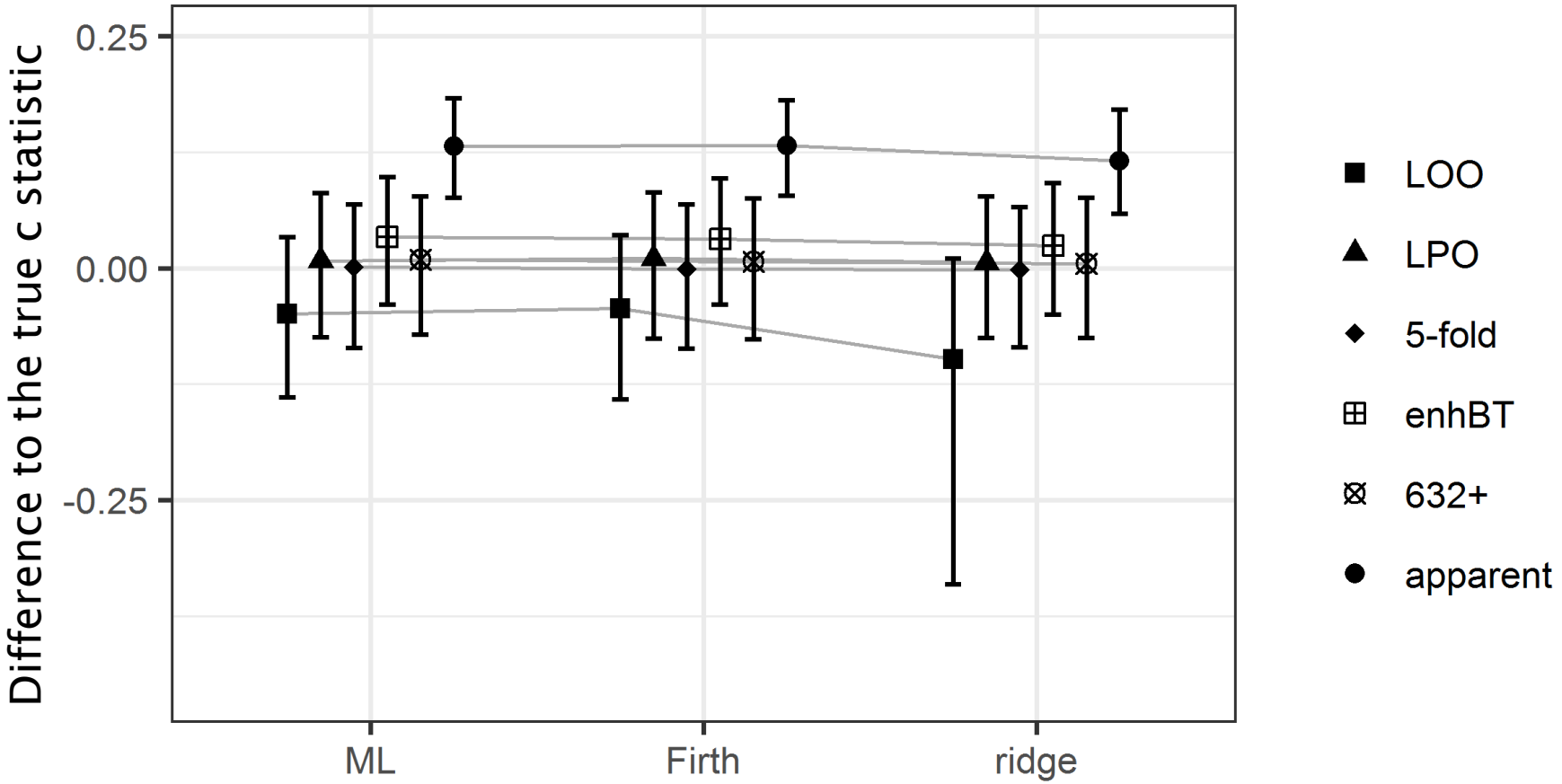
Simulation results

Median (IQR) difference to the true c statistic
N=50, event rate=0.25, small effect



Simulation results

Median (IQR) difference to the true c statistic
N=50, event rate=0.25, large effect



Discussion

- problem of non-estimability in subsamples: most frequent with bootstrap and 5-fold CV,
- different problems with estimation of the discrimination slope.

Conclusion

- LOO CV underestimates the c statistic,
- the bias in LOO CV is larger for methods with small variance,
- .632+ bootstrap, LPO CV and 5-fold CV were most accurate.

Literature

- Airola A, Pahikkala T, Waegeman W, De Baets B, Salakoski T. An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. Computational Statistics & Data Analysis. 2011;55(4):1828-44. doi: 10.1016/j.csda.2010.11.018.
- Binder H, Sauerbrei W and Royston P. Multivariable Model-Building with Continuous Covariates: Performance Measures and Simulation Design 2011. Technical Report FDM-Preprint 105, University of Freiburg, Germany.
- Efron B, Tibshirani R. Improvements on cross-validation: The .632+ bootstrap method. Journal of the American Statistical Association. 1997;92(438):548-60. doi: Doi 10.2307/2965703.
- Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis: Springer; 2001.
- Firth D. Bias reduction of maximum likelihood estimates. Biometrika 1993; 80(1): 27-38.
- Puhr R, Heinze G, Nold M, Lusa L, Geroldinger A. Firth's logistic regression with rare events: accurate effect estimates AND predictions? Stat Med In Press. 2017. doi: 10.1002/sim.7273.