Using the LASSO and Ridge Regression in Case-control Studies with Zero-inflated Predictors

Maria Kohl[1,2]

maria.kohl@meduniwien.ac.at

Joint work with Georg Heinze[2], Hiddo Lambers Heerspink[3] and Joachim Jankowski[4]

[1] Universitätsklinikum Erlangen, Germany
[2] Section for Clinical Biometrics, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Austria
[3] Department of Clinical Pharmacology, University Medical Center Groningen, Netherlands
[4] Medical Clinic IV, Charité, Berlin, Germany

Contents

Background

- Case-control studies for detection of plasma-proteomic signature
- What are plasma-proteomics?

  Determine molecules (peptides) in blood plasma samples by mass spectrometry

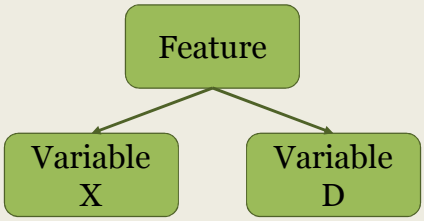
  → Estimation of peptide abundance ('intensity')

  Many peptides are found only in part of the samples

  → frequent occurrence of **zero intensities**

Background

- Identify features (biomarkers) to predict health status ('case' vs. 'control')



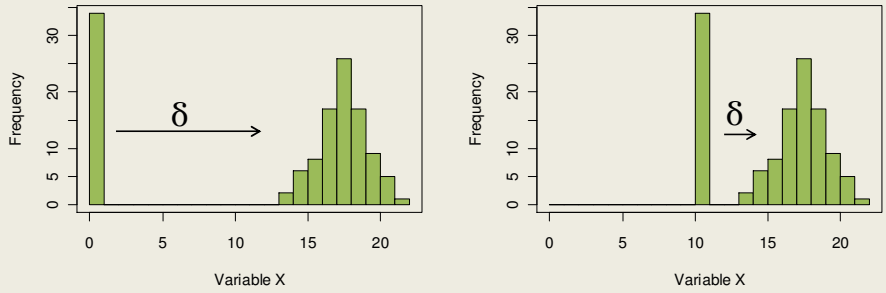
Subject	Feature_1	D_1	X_1
1	0	0	?
2	32768	1	15
3	1048576	0	20
4	0	0	?
...	...	...	...
170	4096	1	12
171	0	0	?
172	16384	1	14

- D ... dichotomous variable (1 if non-zero intensity, 0 if zero intensity)
- X ... continuous log2-transformed intensities

  $\log_2(0) = -\infty$ → What should we do?

Where should the zeros be placed?

At distance δ left to the minimum of the non-zero values.



If D=0, we set X to a feature-specific $\min(X_{D=1})$- δ
→ δ is a global tuning parameter
δ ϵ {log(2^i), i=1,...,8}

Two step model building process

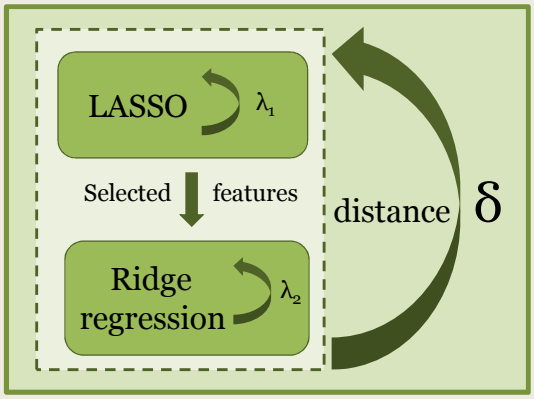
How to make use of both parts, X and D, for selecting features?

(1) **LASSO** for feature selection: $\ell(\beta) - \lambda_1 \sum |\beta_j|$
Using X only

(2) **Ridge regression** for re-estimation: $\ell(\beta) - \lambda_2 \sum \beta_j^2$
Using X and D of each selected feature

Lambda parameters of the LASSO (λ_1) and ridge regression (λ_2) are optimized using cross-validated deviance.

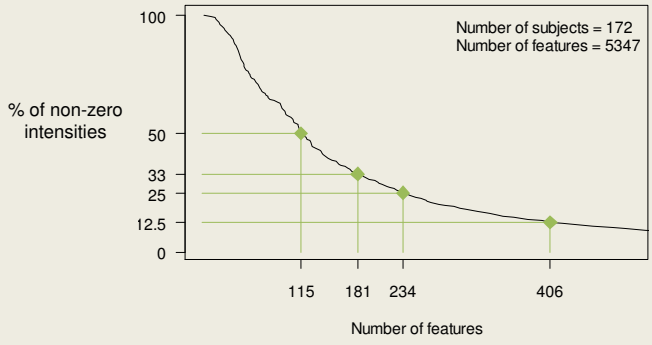
Process



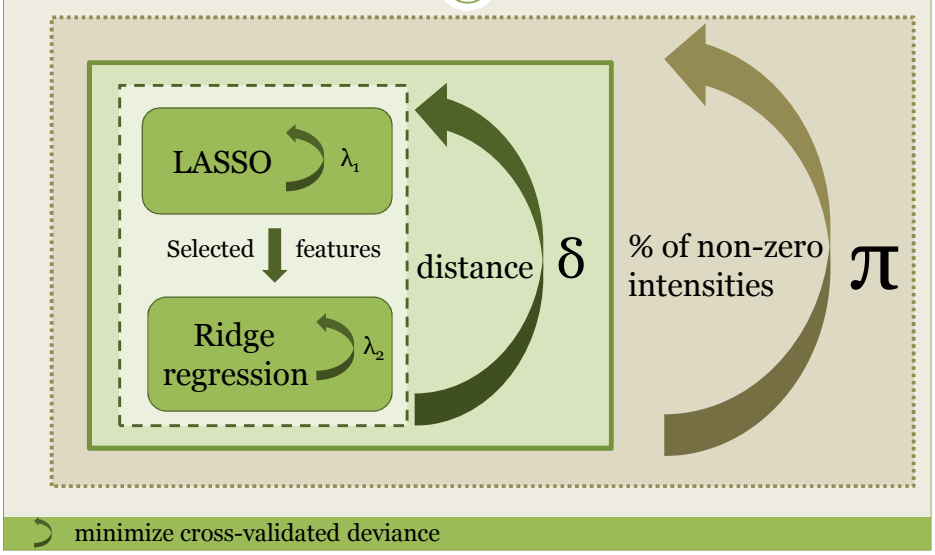
minimize cross-validated deviance

Non-informative filtering (π)

- Filtering out features with excess frequencies of zero intensities
- $\pi \in \{12.5\%, 25\%, 33\%, 50\%\}$



Process



Second-line models

- Investigators are most interested in obtaining a greedy *list of features*
- Some peptides are biologically *not identifiable*

'Second-line' model:

(1) Remove selected features from the 'first line' model of the pool of candidates

(2) Repeat model building to select features waiting in 'second line'

Approaches

Strategy	LASSO step	Ridge step
X_XD	using X only	using X and D of the selected features

Approaches

Strategy	LASSO step	Ridge step
X_XD	using X only	using X and D of the selected features
D_D	using D only	using only D of the selected features
X_X	using X only	using only X of the selected features
XD_XD	using X and D independently	using the selected X and D

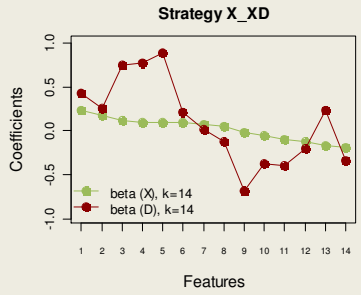
Model building: Results

Number of selected features
(% of columnwise set union of selected features)

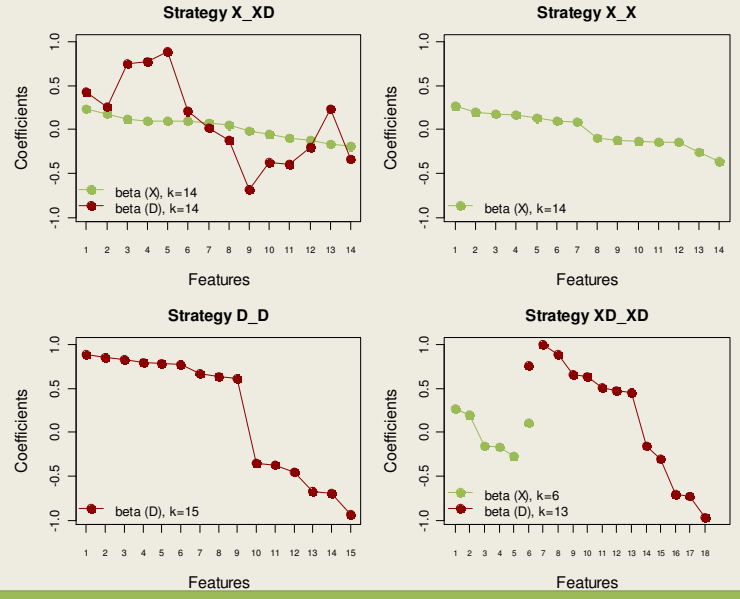
	% of non-zeros (π)							
Strategy	**12.5%**		**25%**		**33%**		**50%**	
X_XD	14	(70%)	**14**	**(82.4%)**	12	(80.0%)	9	(90%)
D_D	**14**	**(70%)**	14	(82.4%)	12	(80.0%)	9	(90%)
X_X	**15**	**(75%)**	12	(70.6%)	10	(66.7%)	6	(60%)
XD_XD	**18**	**(90%)**	13	(76.5%)	11	(73.3%)	9	(90%)
	20	(100%)	17	(100%)	15	(100%)	10	(100%)

Optimal π

Model building: Results



Model building: Results



Evaluating model performance: measures

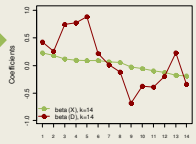
Models are compared by the following performance measures:

- Concordance index (C-index)
- Predictive accuracy: $n^{-1}\sum|y_i - \hat{p}_i|$

  Mean absolute difference of y and estimated predictor
- Misclassification rate (for ‘cases‘ and ‘controls‘)

Evaluating model performance: Another (outer) cross validation loop

‚Leave-two-out' cross validation

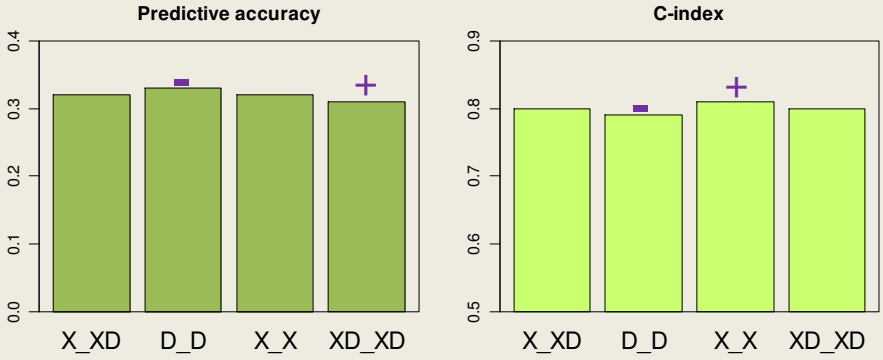


Subject	Feature_1	Feature_2
1	0	0
2	32768	0
4	0	0
6	0	8086
...	...	...
171	4096	16384
172	0	0

3	0	7062
5	8922	13965

Subject	CV
1	
2	
3	0.2
4	
5	0.7
6	
...	...
171	
172	

Results: Performance



Results: Performance

Conclusions

Analysing several data sets we conclude:

- X_X, X_XD and XD_XD performed similarly D_D was uniformly worse
- Including D does not improve performance
- Marginal impact of the value of distance δ
- Choice of the required minimum proportion of non-zeros (π) is crucial

Thanks for your attention!

References

1. R. Tibshirani. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society 1996.
2. A.E. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 1970.
3. M. Schemper. Predictive accuracy and explained variation. Statistics in Medicine 2003.
4. J. Friedman, T. Hastie and R. Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 2010.

Supported by European Union's 7 framework programme (SysKid, a Collaborative FP7 Research Project to Fight Chronic Kidney Disease; grant agreement number HEALTH–F2–2009–241544)