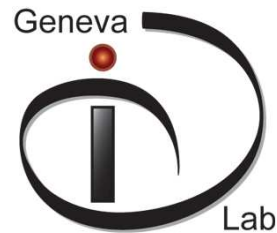


Feature selection in a
small-sample-high-dimensionality setting:
accuracy, stability, and learning curves.

Artificial Intelligence Group
Department of Computer Science
University of Geneva
Alexandros kalousis



The learning problem

Given:

- a set of mass spectra from urine samples of male and female, and
- the class label of each spectrum (i.e. male, female)

Produce:

- a good classification model (low predictive error).
- a minimal *set* of biomarkers discriminating female and male samples.



The *Male vs Female* Dataset

- Training set: 67 Male, 67 Female
- Blind Hold Out set : 92 instances
- Number of features, i.e. m/z masses : 5616
- Very large number of missing values, more than 3000 features had more than 80% missing values.
- Removing features with more than 80% missing values left us with 1524 features.



The Learning Experiments

- Two types of experiments:
 - Without feature selection, i.e. construct classification models with all the features,
 - With feature selection, i.e. construct classification models with a limited number of features.
- Performance estimation with 10-fold Cross Validation on the training set
- Two performance criteria: *Accuracy* and *Stability* of learned models
- Based on the performance estimates from the 10-fold CV choose:
 - which classification algorithm to use
 - which feature selection algorithm to use
 - and what is the best minimal cardinality feature setfor the construction of the final classification model.
- Estimate the performance of the final model on the blind hold-out set.



The Learning Algorithms

- Classification algorithms used:
 - Support Vector Machines, SMO, discovers linear decision surfaces
 - Decision Trees, J48, can discover complex decision surfaces provided there are enough training data,
 - Nearest Neighbors, IBL, classifies based on the distance from the most similar instance

the parameters of the classification algorithms were optimized by internal 10-CV.

- Feature Selection algorithms used:
 - *Univariate measures*: CHI-square, Information Gain, SYMmetrical uncertainty they determine the importance of each feature on its own
 - *Multivariate measures*: SVMONE and SVMRFE, both based on SMO, which determine the importance of features accounting for their interactions:

Feature set cardinalities examined: 50, 100, 150, 200, 300



Error Estimations

	Opt	NoFS	CHI	IG	Relief	SVMONE	SVMRFE	SYM
			<i>D_r</i> , feat sel 300					
SMO	5.97	5.97	12.69	12.69	13.43	6.72	7.46	12.69
J48	29.85	32.84	29.85	29.10	29.10	31.34	27.61	32.09
IBk	29.10	42.54	32.84	32.09	34.33	26.87	23.88	30.60
			<i>D_r</i> , feat sel 250					
SMO			14.93	14.93	9.70	6.72	9.70	13.43
J48			29.85	29.10	27.61	31.34	30.60	30.60
IBk			24.63	26.12	33.58	23.13	27.61	28.36
			<i>D_r</i> , feat sel 200					
SMO			15.67	14.93	10.45	7.46	9.70	14.18
J48			30.60	29.10	29.10	26.87	29.10	30.60
IBk			30.60	30.60	29.85	29.10	24.63	32.09
			<i>D_r</i> , feat sel 150					
SMO			19.40	19.40	11.94	10.45	11.94	17.91
J48			28.36	28.36	32.84	24.63	35.82	30.60
IBk			30.60	32.84	28.36	20.15	23.88	31.34
			<i>D_r</i> , feat sel 100					
SMO			19.40	18.66	11.19	11.19	10.45	19.40
J48			32.09	31.34	32.09	28.36	29.85	30.60
IBk			28.36	30.60	23.88	20.15	21.64	29.10
			<i>D_r</i> , feat sel 50					
SMO			14.18	14.18	19.40	18.66	16.42	14.18
J48			26.12	29.85	29.10	24.63	29.10	29.85
IBk			23.88	25.37	26.12	25.37	23.13	25.37

Which one of the **equally good** solutions to choose?

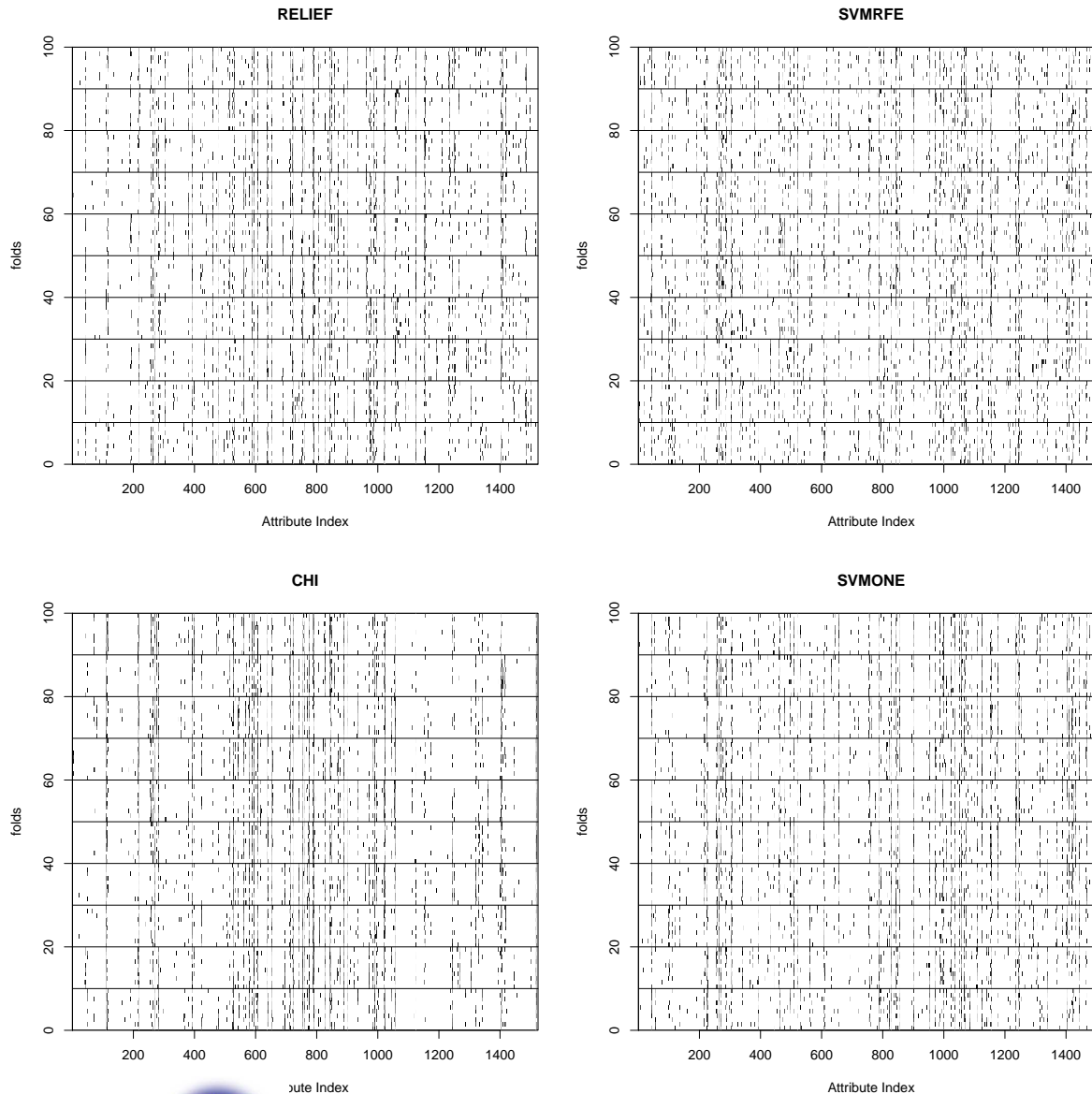


Controlling for stability of the learning tools

- Very simply idea: We do not want different subsets of the same dataset to produce different models, i.e. different biomarkers. Because we loose confidence of users on the models.
- How to control stability?
 - Take different subsets of your training data.
 - Apply the feature selection algorithms
 - Are the sets of selected features similar? Do they point out the same biomarkers?



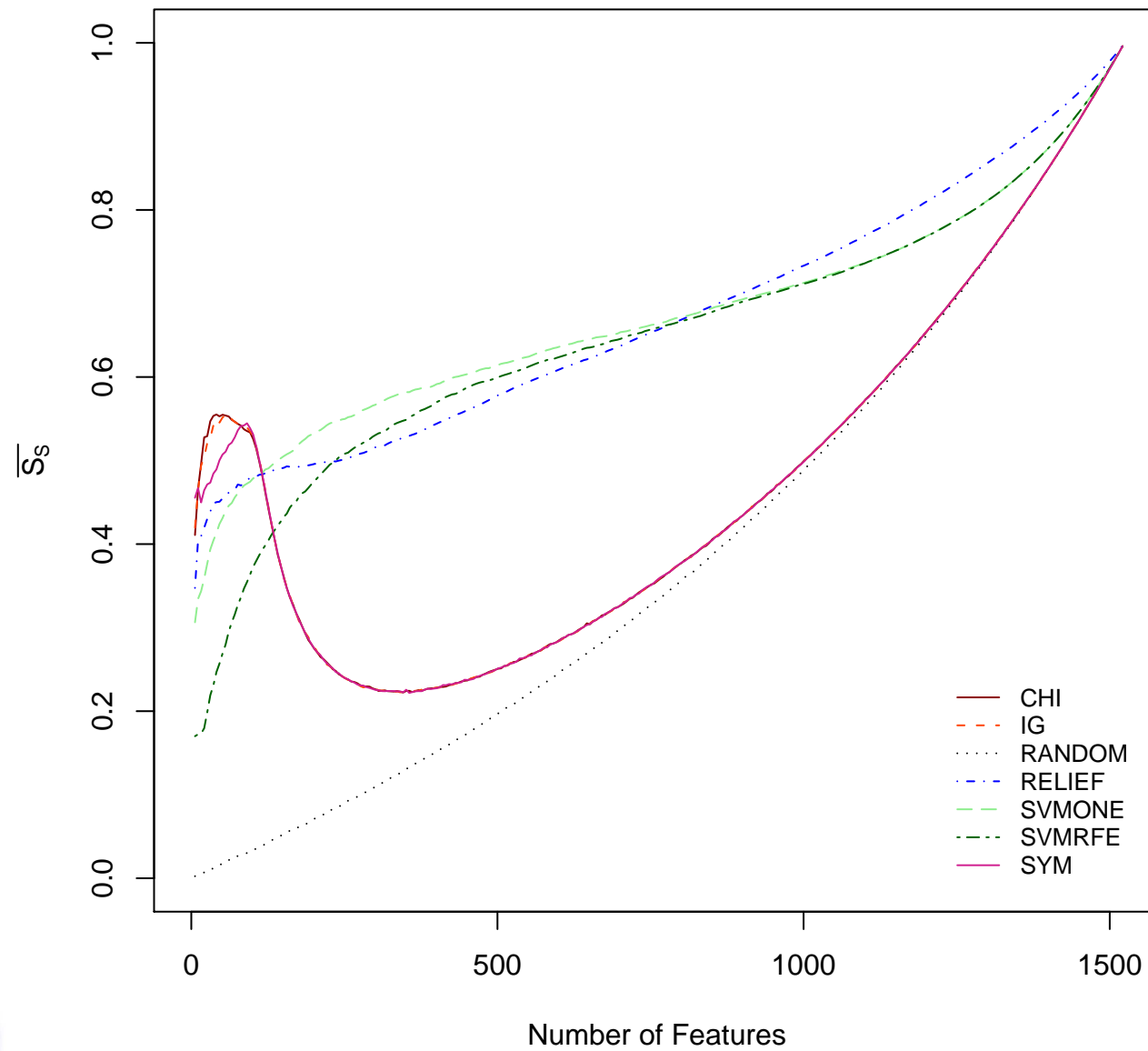
Visualizing Stability, Feature set cardinality=50



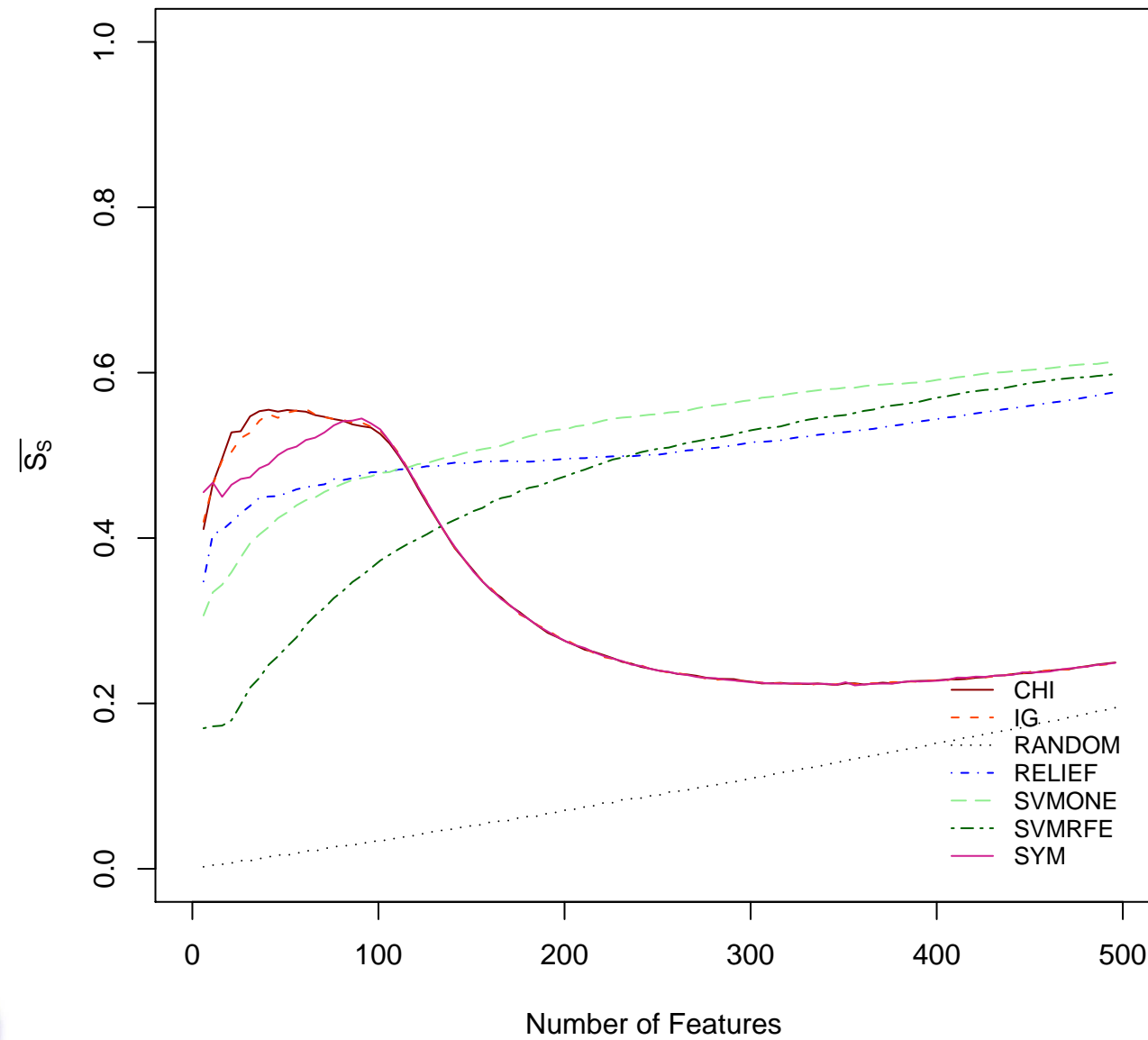
- X-axis=features, Y-axis=folds
- Two consecutive horizontal lines contain the ten inner cv-folds of a single outer cv-fold.
- A point indicates that the corresponding feature was selected.
- The more *complete* vertical lines (i.e. same feature selected among different folds) the more stable the algorithm.
- Clear picture of stability behavior *and also* of which features are important.



Visualizing Stability, for all possible feature set cardinalities



Visualizing Stability, for feature set cardinalities up to 500



A last comment on stability

- It is possible to have high instability with very good accuracy
- All the different sets of selected features have high predictive power
- This is an indication of redundancy on the original feature set



The final model

- the best algorithm combinations according to error were:

SMO with SVMONE or SVMRFE

- the best feature set cardinalities according to error were:

200, 250, 300 or the full training set.

- the most stable results for SVMONE and SVMRFE feature selection were observed around 300.
- So the winning combination based on error and stability was:

SMO+SVMONE @ 300 features

- and its error on the hold out set was:

10.84%



On the relation of classification error and sample size

- We construct learning curves at different feature set cardinalities.
- A learning curve shows how the classification error evolves when the number of training instances increases.
- Size of training sets used to construct the learning curves:

10%, 20%, ... 100% of the original training set

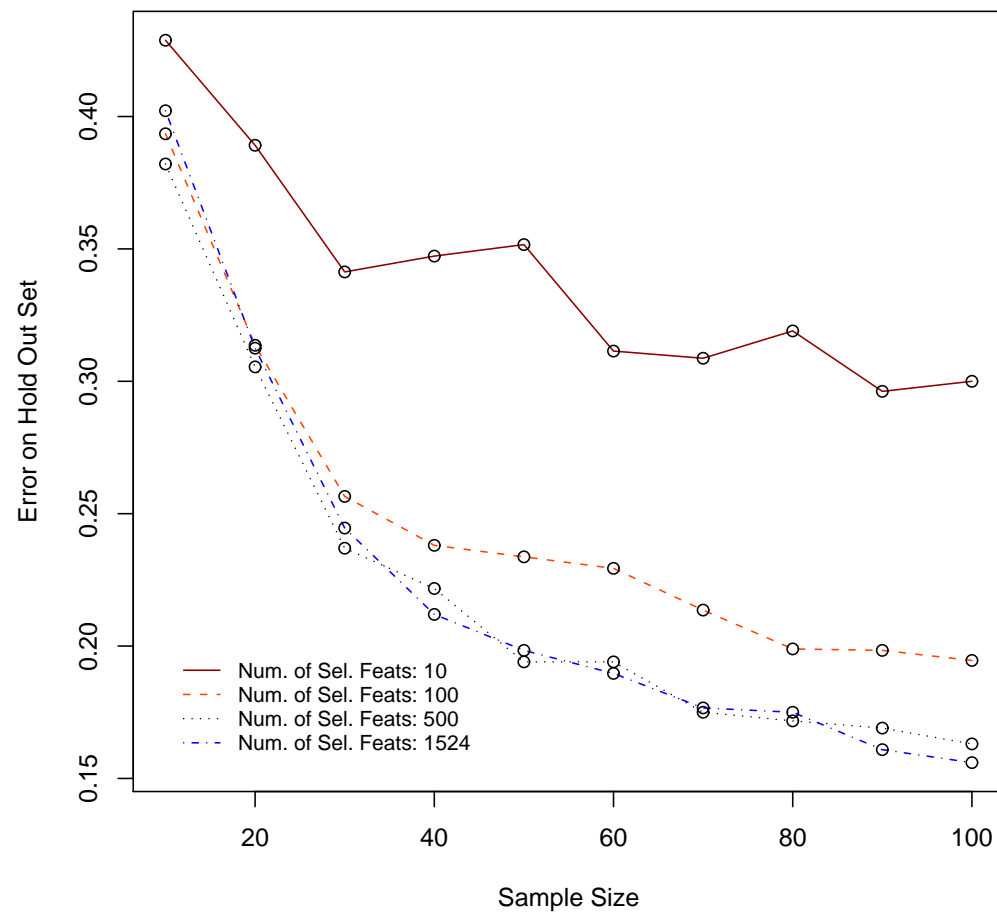
- Selected feature set cardinalities:

10,20,...,100, 150, 200, 250, 300, 350, 400, 450, 500, 1524

- Classification algorithm used to construct the curves was SMO.



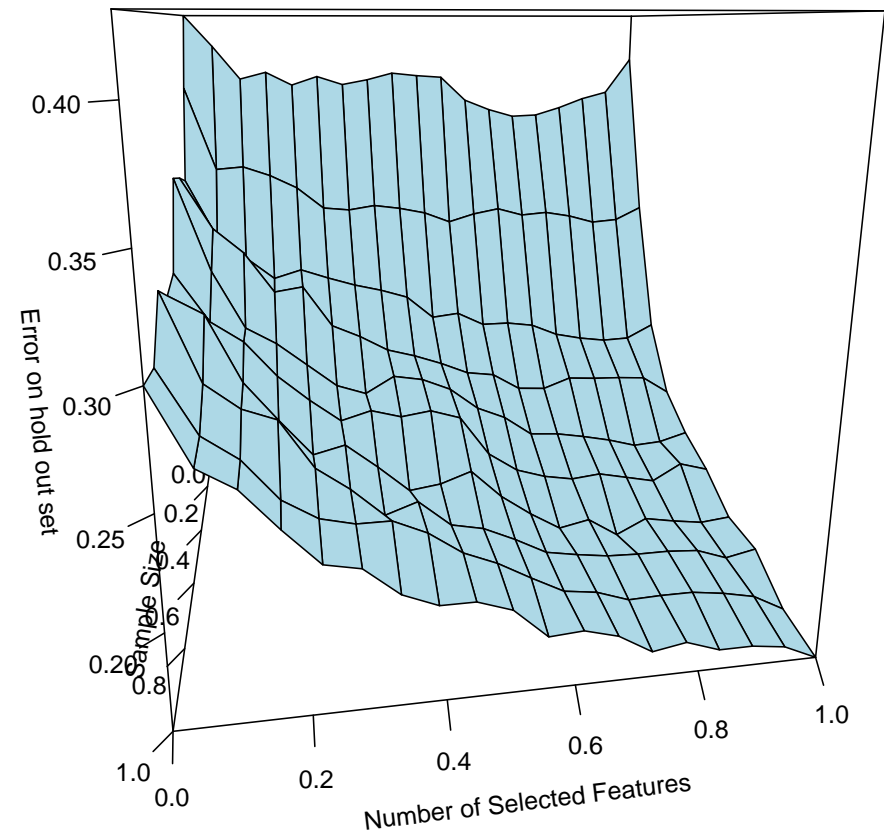
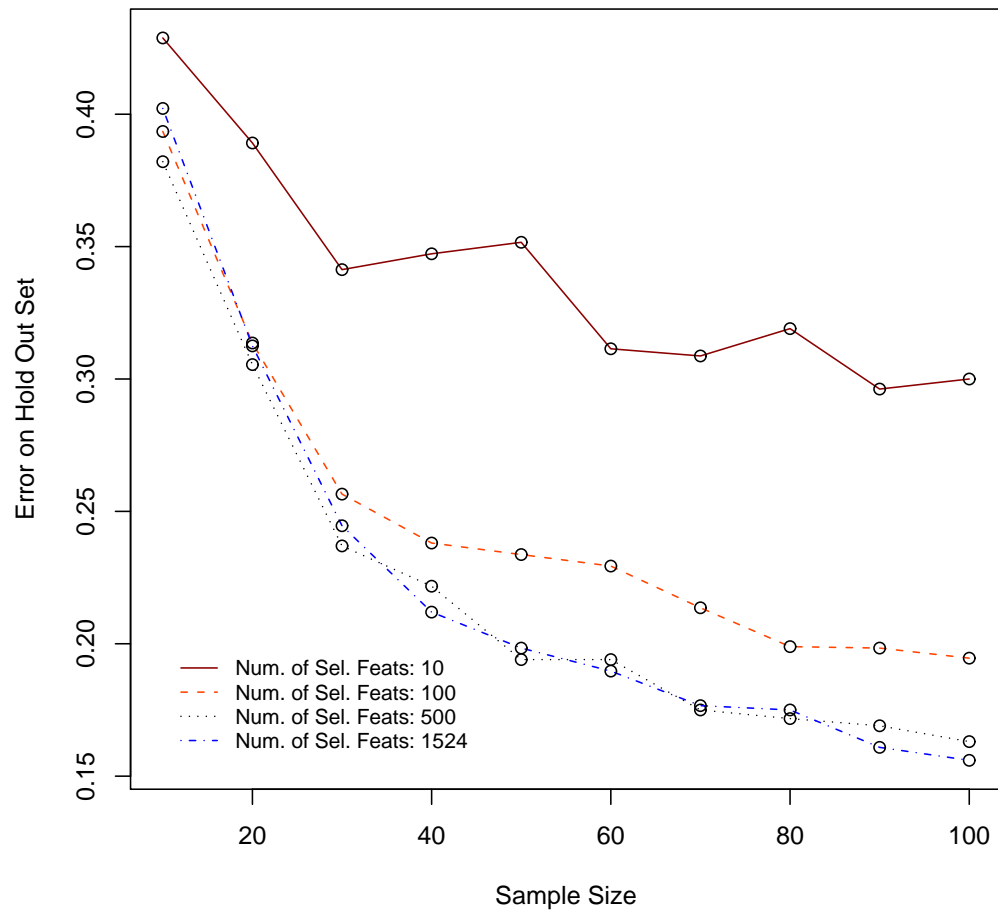
Learning curves at feature set cardinalities=10, 100, 500, 1524, SVMRFE



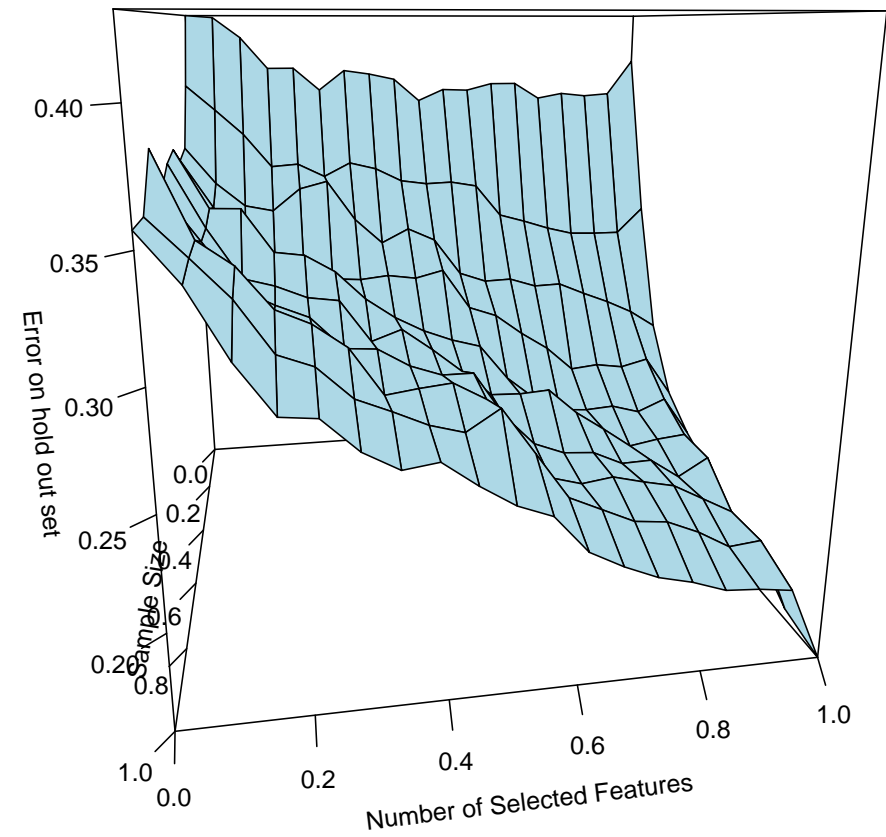
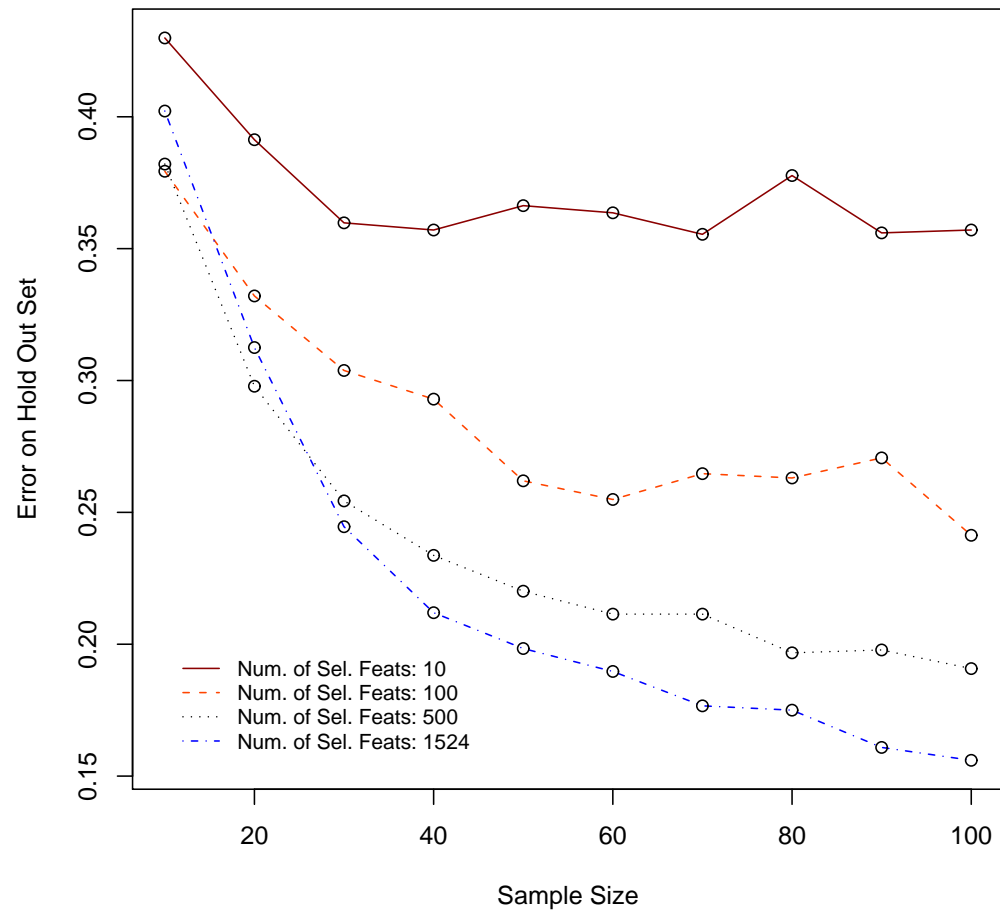
For the male vs female problem best performance with the *full* training set and *all features*.



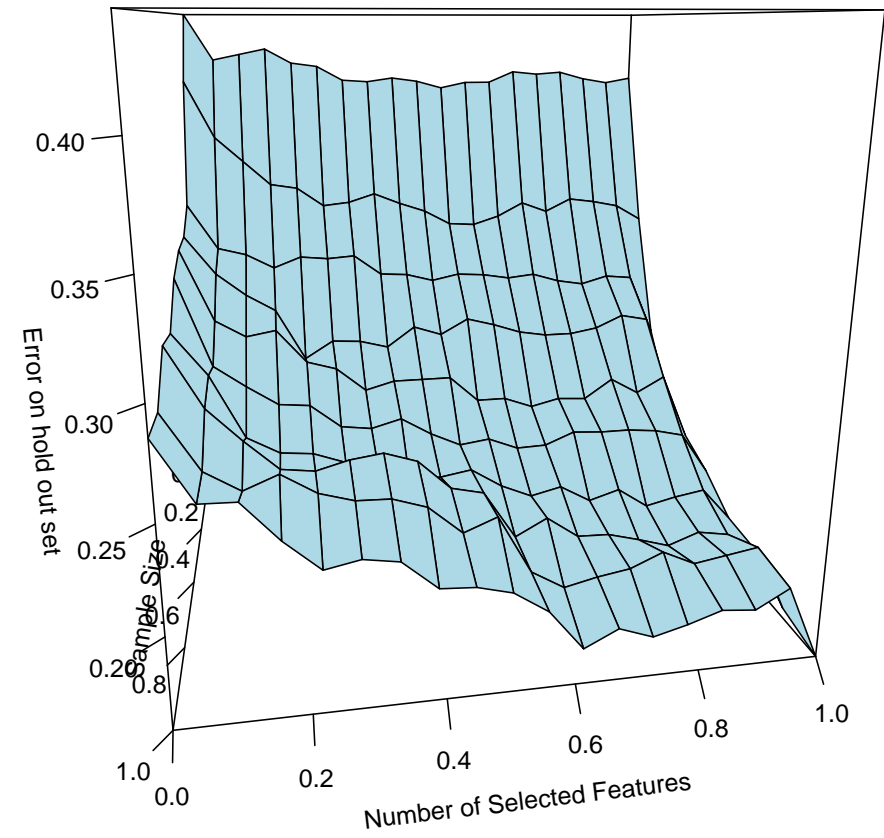
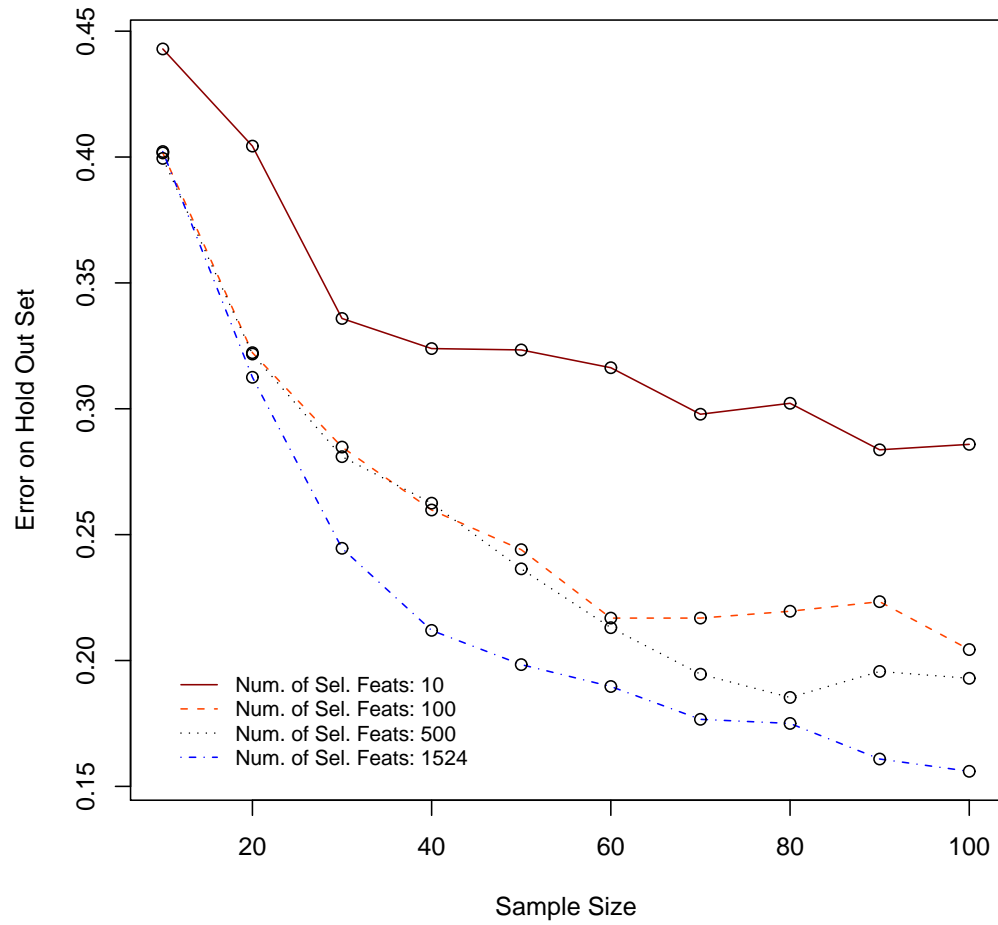
Learning curves, SVMRFE



Learning curves, RELIEF



Learning curves, IG



Conclusions

- Finding the best set of markers is a complex process that has to account for many factors.
- A very strict methodological setup to control for the validity of the selected markers must always be followed.

- This just a part of the things that



can do for



- For more...

- come on the next session WG.4: 18:00-19:30
- or talk to the e-lico people during the COST meeting :)
- or visit: www.e-lico.eu