Feature selection by replicate reproducibility and non-redundancy

Tümay Capraz
Huber Group EMBL Heidelberg

Feature selection

Unsupervised  **Supervised**

Filter methods

- evaluated on intrinsic properties of the data
- fast and simple
- e. g. selection of highly variable genes

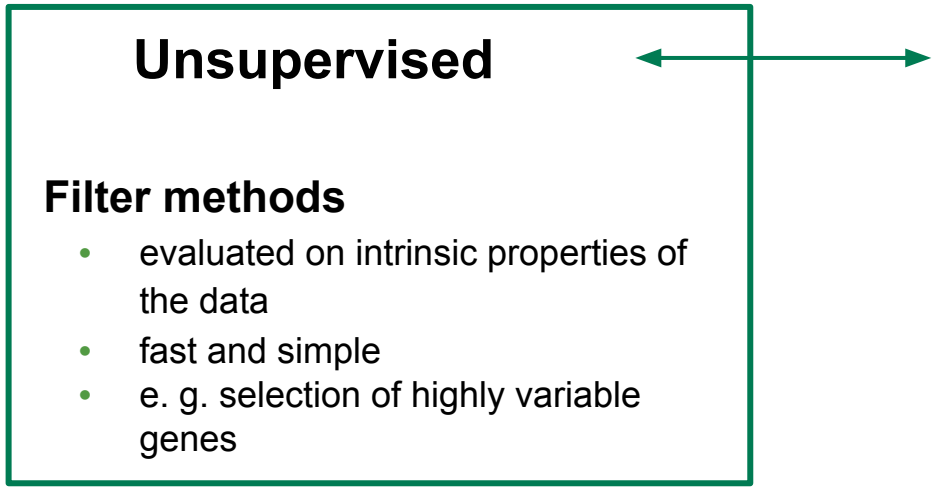
Wrapper methods

- feature subsets evaluated on performance of a model
- search algorithms combined with classifier
- computationally expensive

Embedded methods

- feature selection embedded in a model
- e. g. decision trees, lasso regression

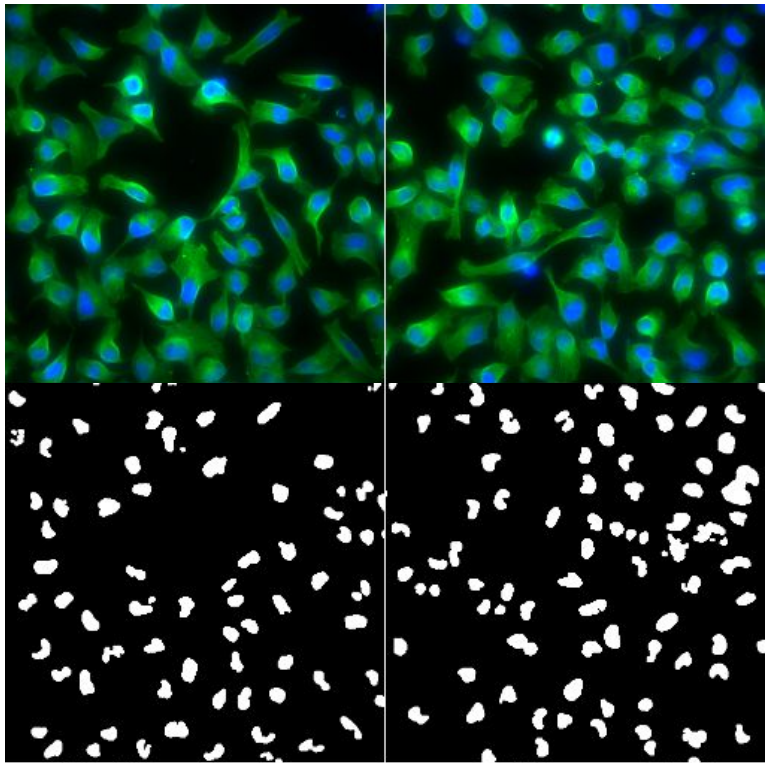
Feature selection



Embedded methods

- feature selection embedded in a model
- e. g. decision trees, lasso regression

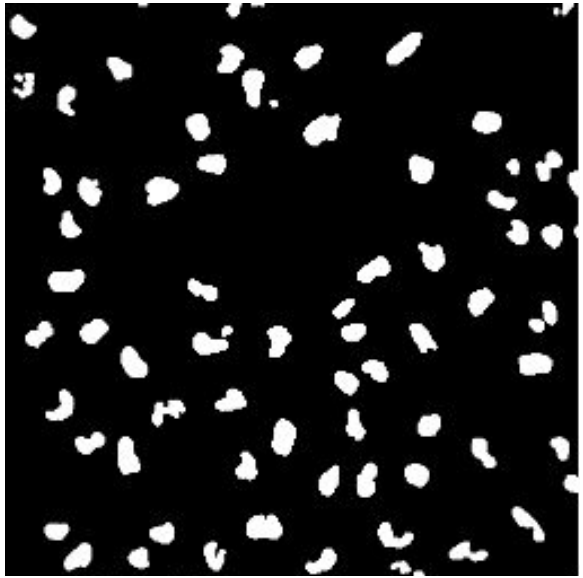
Feature selection applied to image analysis



Pau et al. 2010, EBImage—an R package for image processing with applications to cellular phenotypes

- Intrinsically high dimensional
- Analysis with neural networks
- Feature extraction (e.g. EBImage)
  - Highly correlated features
  - Noisy features

Feature selection applied to image analysis



Examples of extracted features:

- Cell count

Shape:

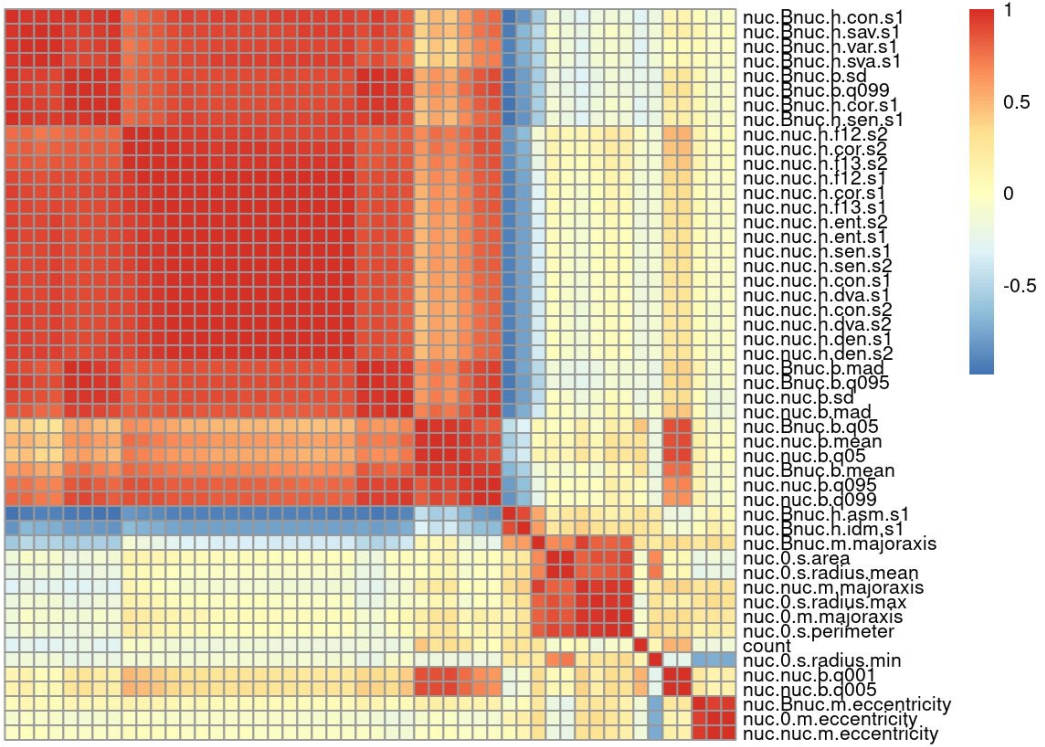
- Major axis length
- Eccentricity
- Area

Spatial independent:

- Quantile intensity
- Mean intensity

Pau et al. 2010, EBImage—an R package for image processing with applications to cellular phenotypes

Feature extraction produces highly correlated features



Data from: Laufer, C. et al. 2013. Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping.

Published: 07 April 2013

Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping

Christina Laufer, Bernd Fischer, Maximilian Billmann, Wolfgang Huber & Michael Boutros

Nature Methods **10**, 427–431(2013) | Cite this article

A map of directional genetic interactions in a metazoan cell

Bernd Fischer, Thomas Sandmann, Thomas Horn, Maximilian Billmann, Varun Chaudhary, Wolfgang Huber, Michael Boutros
European Molecular Biology Laboratory, Germany; German Cancer Research Center (DKFZ), Germany; Heidelberg University, Germany

Research Article · Mar 6, 2015

FeatSeekR: R package for unsupervised feature selection

Goal:

- Non-redundant feature subset with high replicate reproducibility

Method:

- Pre-select feature set
- Model each feature as a function of the selected features
- Project out dimension of the selected features
- Selection based on reproducibility of the signal across replicates
- Stop if there is only noise left

FeatSeekR: R package for unsupervised feature selection

Input:

- Data: $(X_{n\times p})^r$
- Selected features: $(S_{n\times q})^r$

Rank features according to reproducibility between replicates:

- Fit linear model:

$$x^r_{t,i} = S^r_t \beta^r_{t,i} + \epsilon^r_{t,i} \quad with \quad i = 1 \dots p$$

- Select feature i with highest reproducibility between replicates:

$$\max_i(g(\epsilon^1_i, \dots, \epsilon^r_i))$$

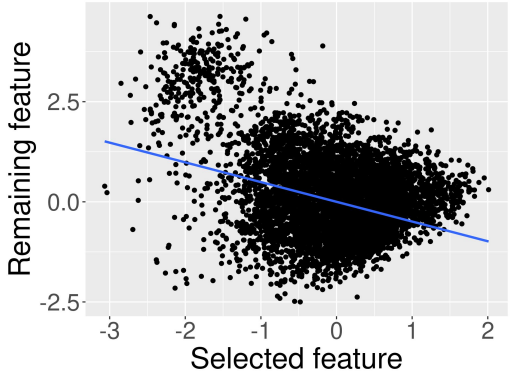
- Project out dimension spanned by previously selected features by setting:

$$x^r_{t+1,i} = \epsilon^r_{t,i}$$

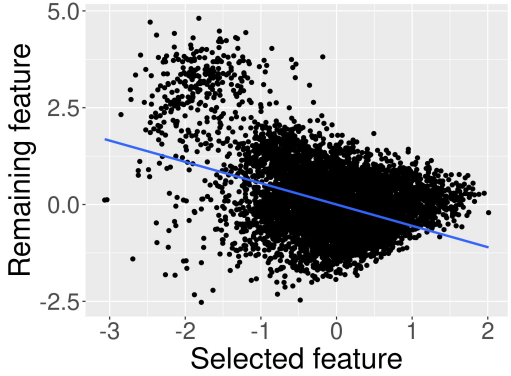
Iteration 1

1.) Fit linear model

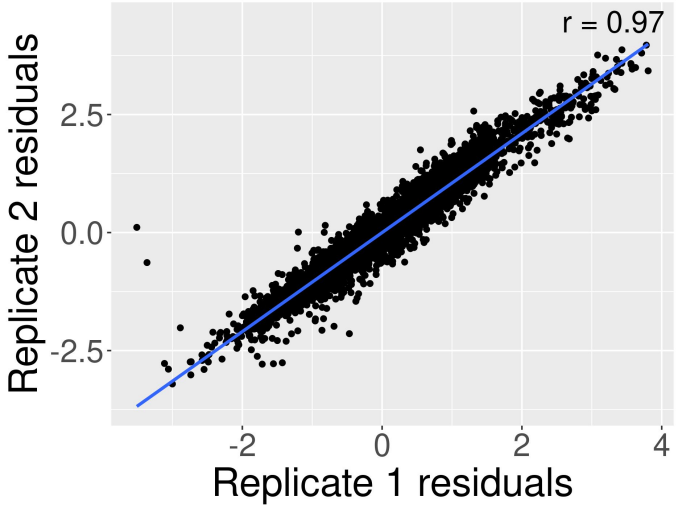
R1



R2



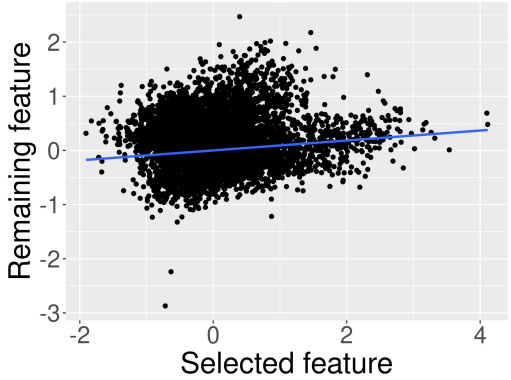
2.) Evaluate replicate reproducibility



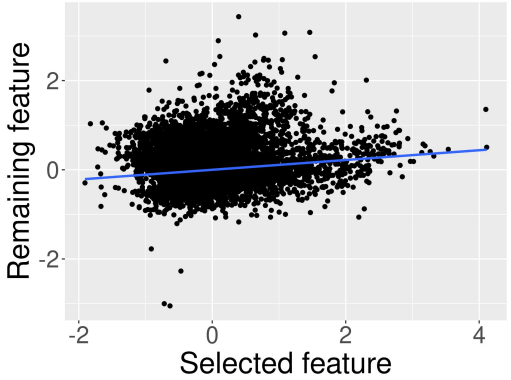
Iteration 5

1.) Fit linear model

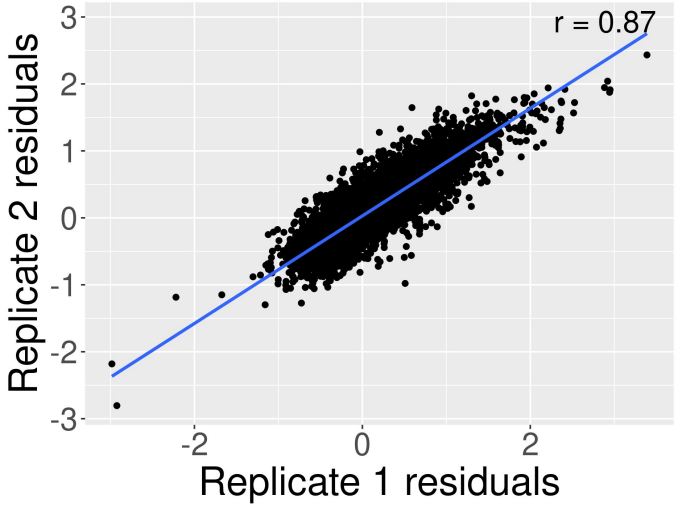
R1



R2



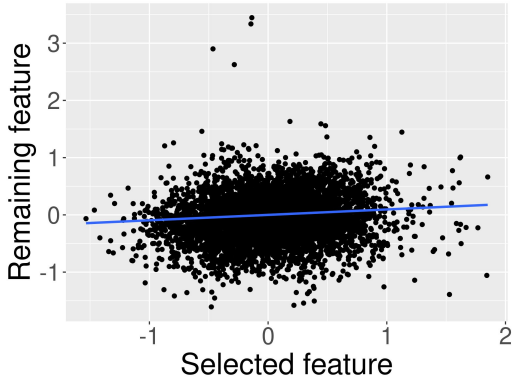
2.) Evaluate replicate reproducibility



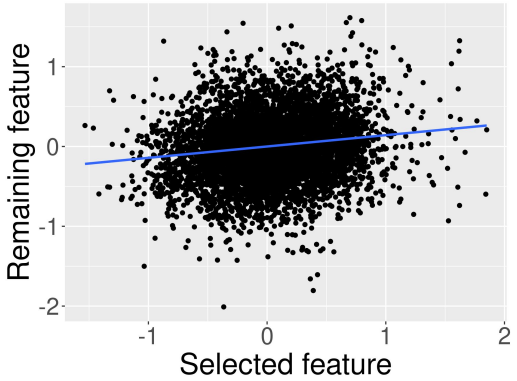
Iteration 10

1.) Fit linear model

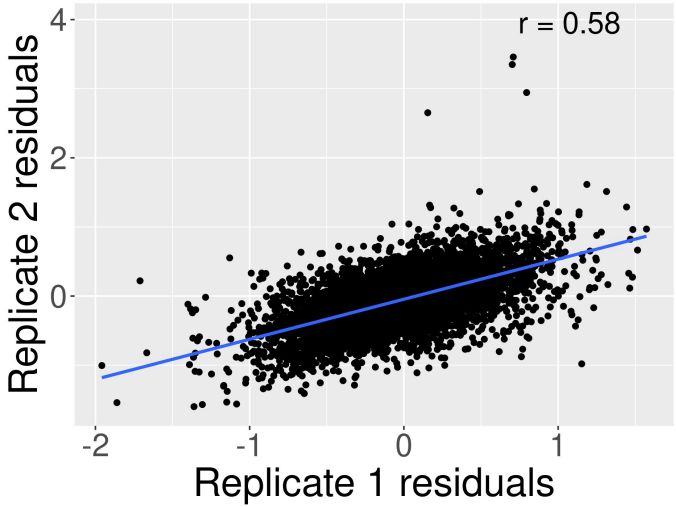
R1



R2



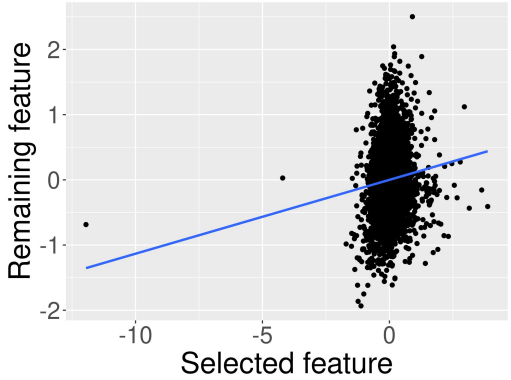
2.) Evaluate replicate reproducibility



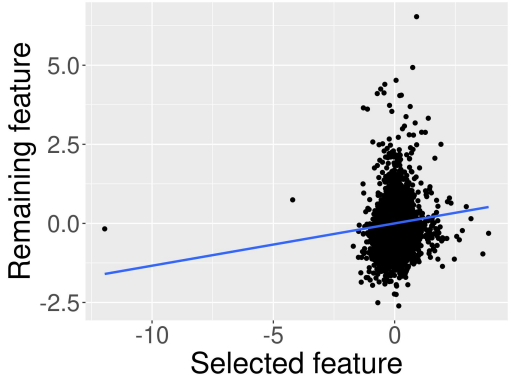
Iteration 18

1.) Fit linear model

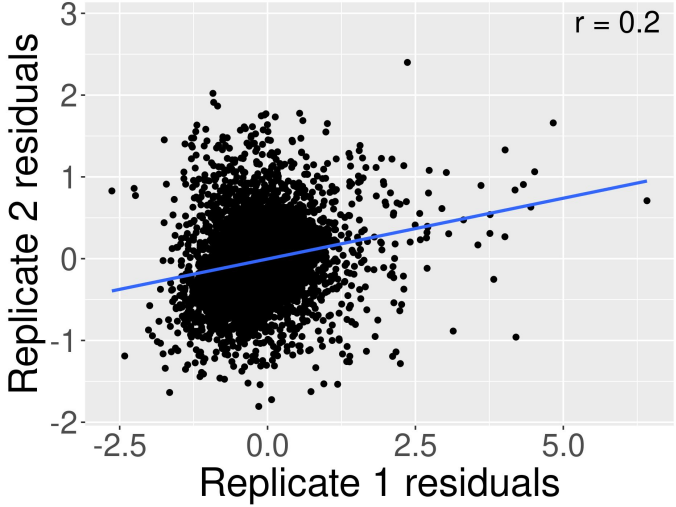
R1



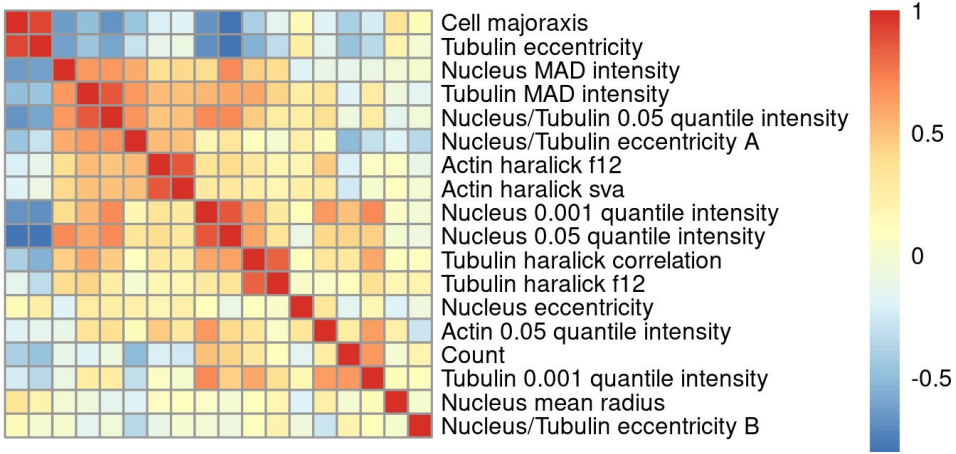
R2



2.) Evaluate replicate reproducibility



FeatSeekR identifies non-redundant and reproducible feature set



- 5820 samples
- 2 replicates
- 216 features

Data from: Laufer, C. et al. 2013. Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping.

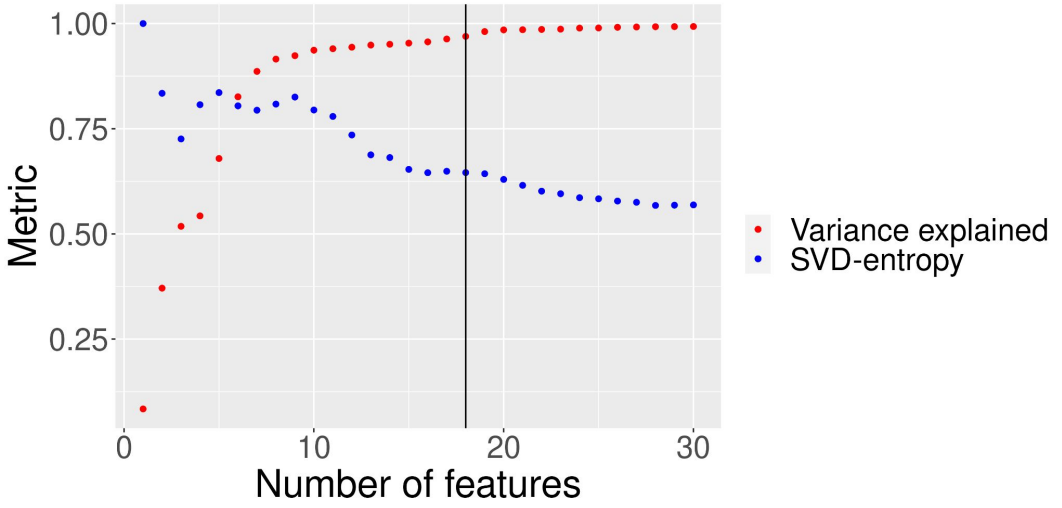
FeatSeekR identifies non-redundant and reproducible feature set

Variance explained:

$$\frac{\sum_{i=1}^{p} R^2_{adj,i}}{p}$$

Non-redundancy by SVD entropy:

$$E = -\frac{1}{log(q)} \sum_{i=1}^{q} V_i \, log(V_i)$$



Alter et al. 2000. Singular value decomposition for genome-wide expression data processing and modeling.

Data from: Laufer, C. et al. 2013. Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping.

Acknowledgement

Huber group

Wolfgang Huber
Simone Bell
Constantin Ahlmann-Eltze
Hosna Baniadam
Donnacha Fitzgerald
Holly Giles
Alexander Helmboldt
Nick Hirschmüller
Katharina Imkeller
Sarah Kaspar
Vladislav Kim
Junyan Lu
Dorothee Mersch
Tobias Roider
Thomas Naake
Thomas Schwarzl
Mike Smith
Harald Vöhringer
Julia Philipp

EMBL

erc
European Research Council
Established by the European Commission