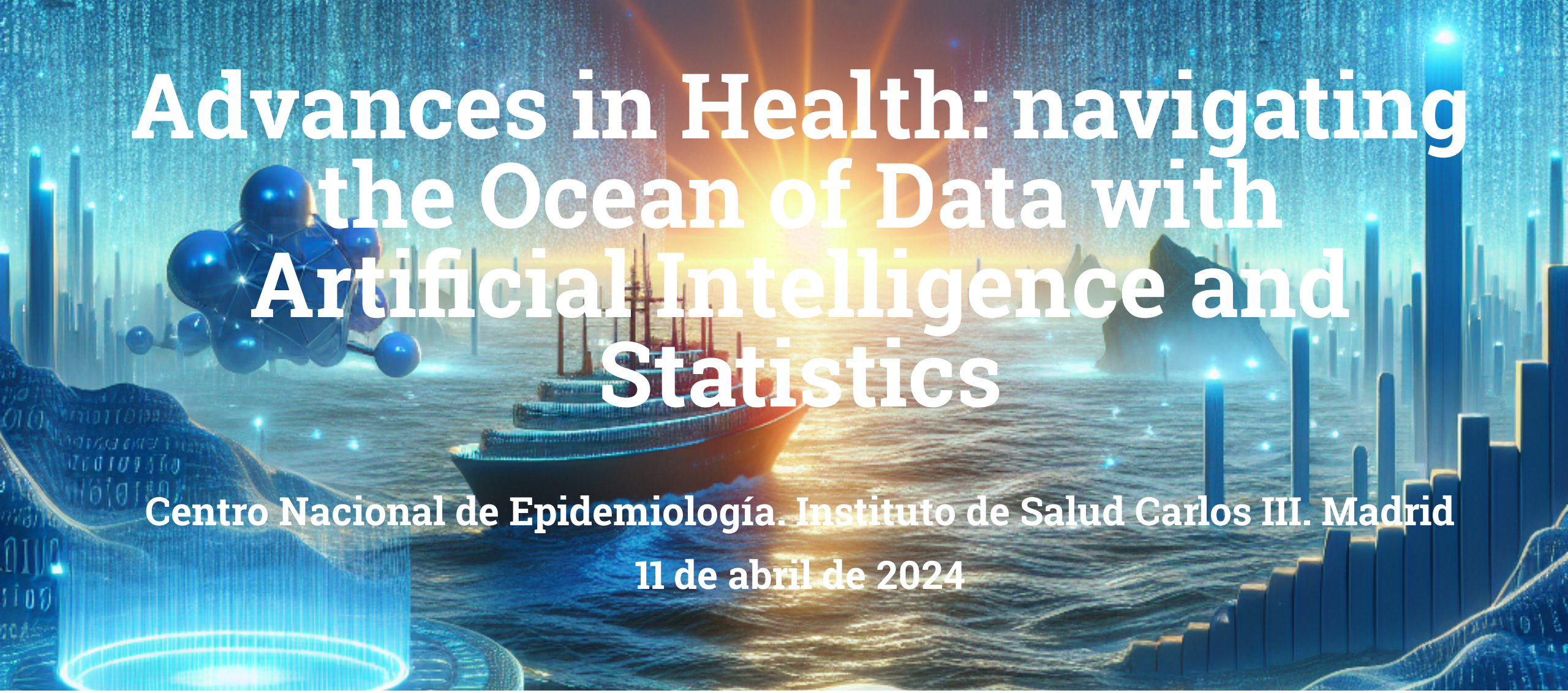


Advances in Health: navigating the Ocean of Data with Artificial Intelligence and Statistics

Centro Nacional de Epidemiología. Instituto de Salud Carlos III. Madrid

11 de abril de 2024

uc3m | Universidad Carlos III de Madrid

Rosa Elvira Lillo

IBiDat

IBiDat...About us

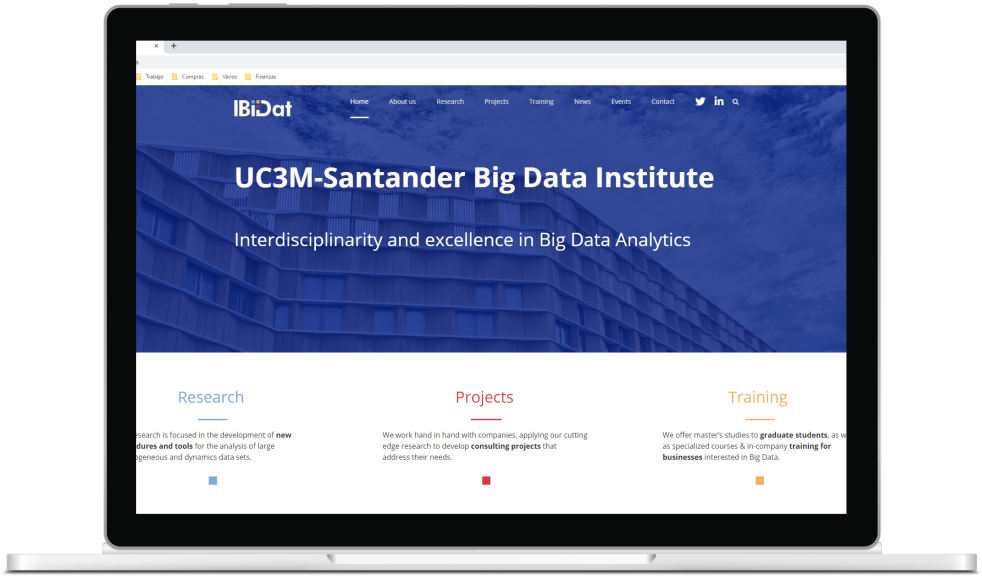
About us

uc3m-Santander Big Data Institute...and IA

Interdisciplinar team

Cutting-Edge Academic research

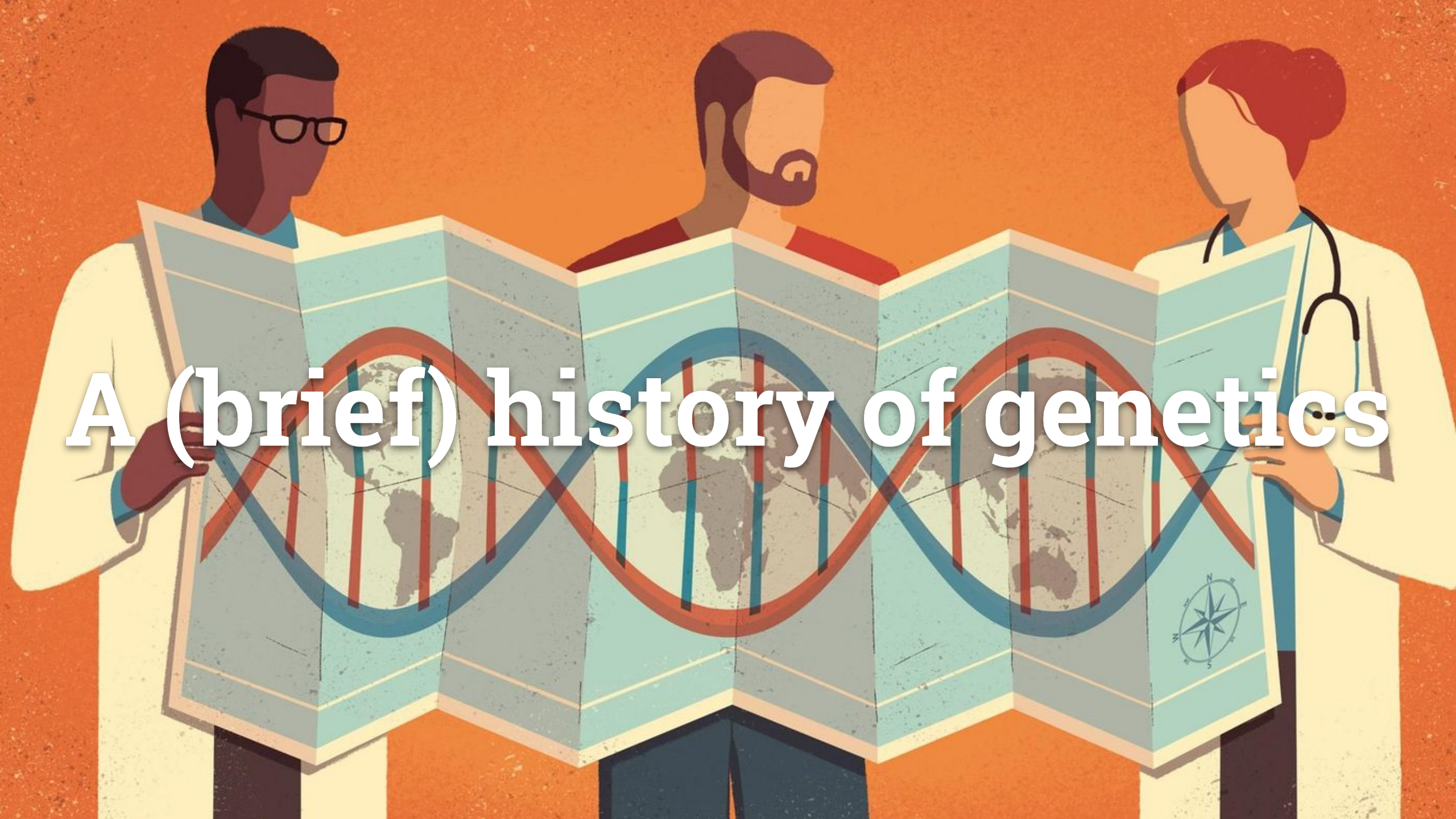
Boutique projects and specialized courses



ibidat.es

@BigData_uc3m

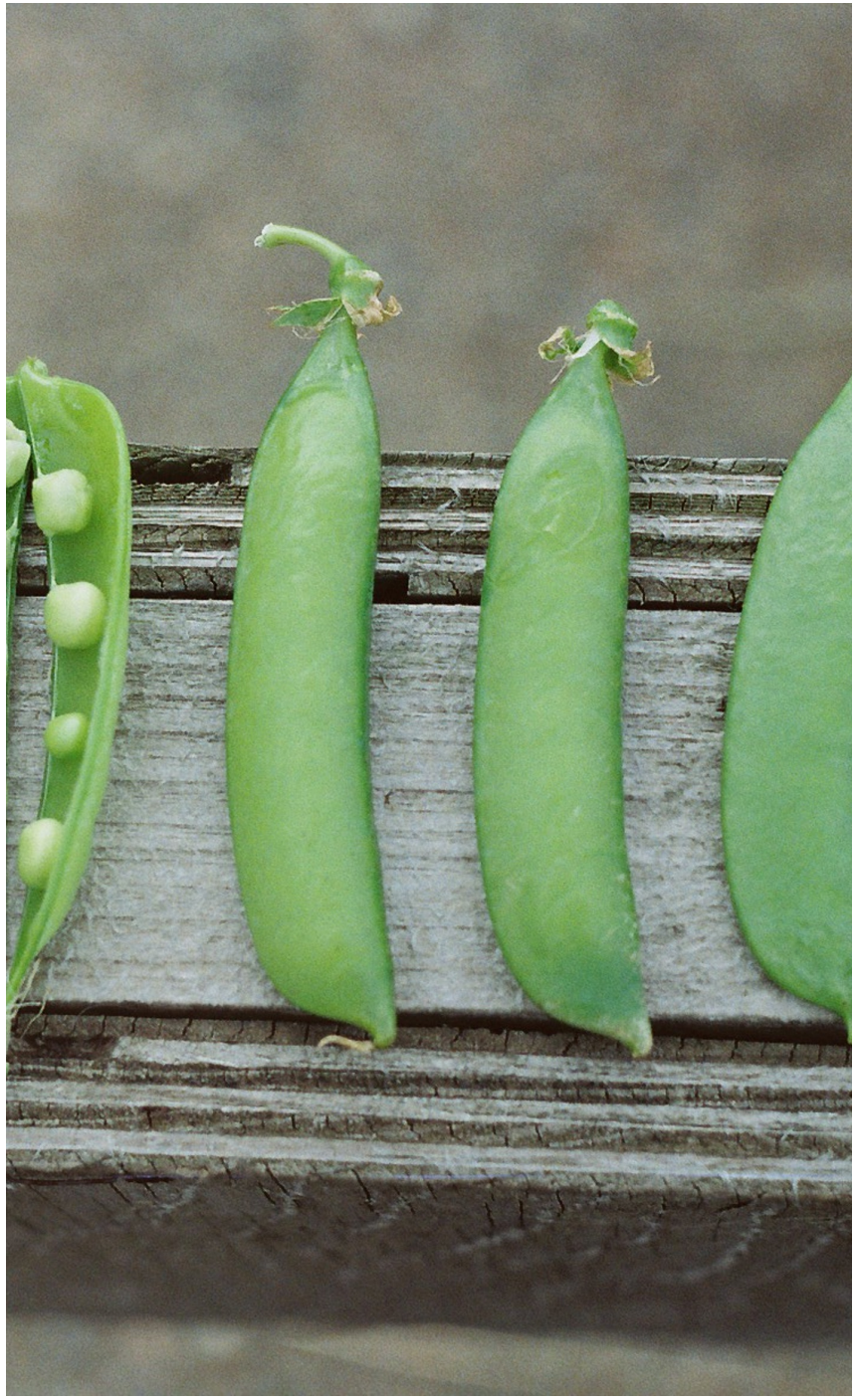
Instituto Big Data UC3M-Santander



A (brief) history of genetics

1859

Mendel defines principles of inheritance

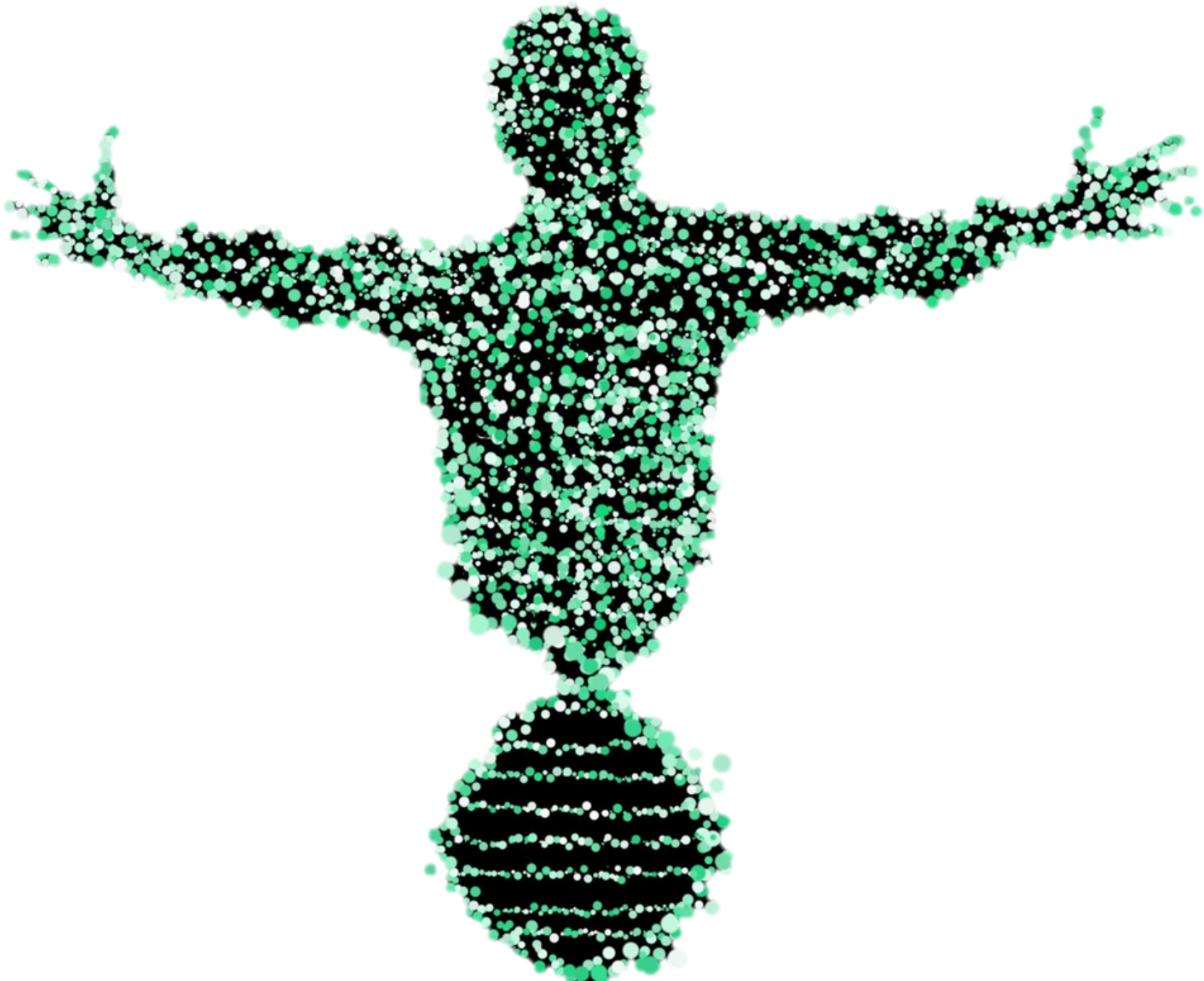


1944

DNA isolated as genetic material

1990

The Human Genome project starts



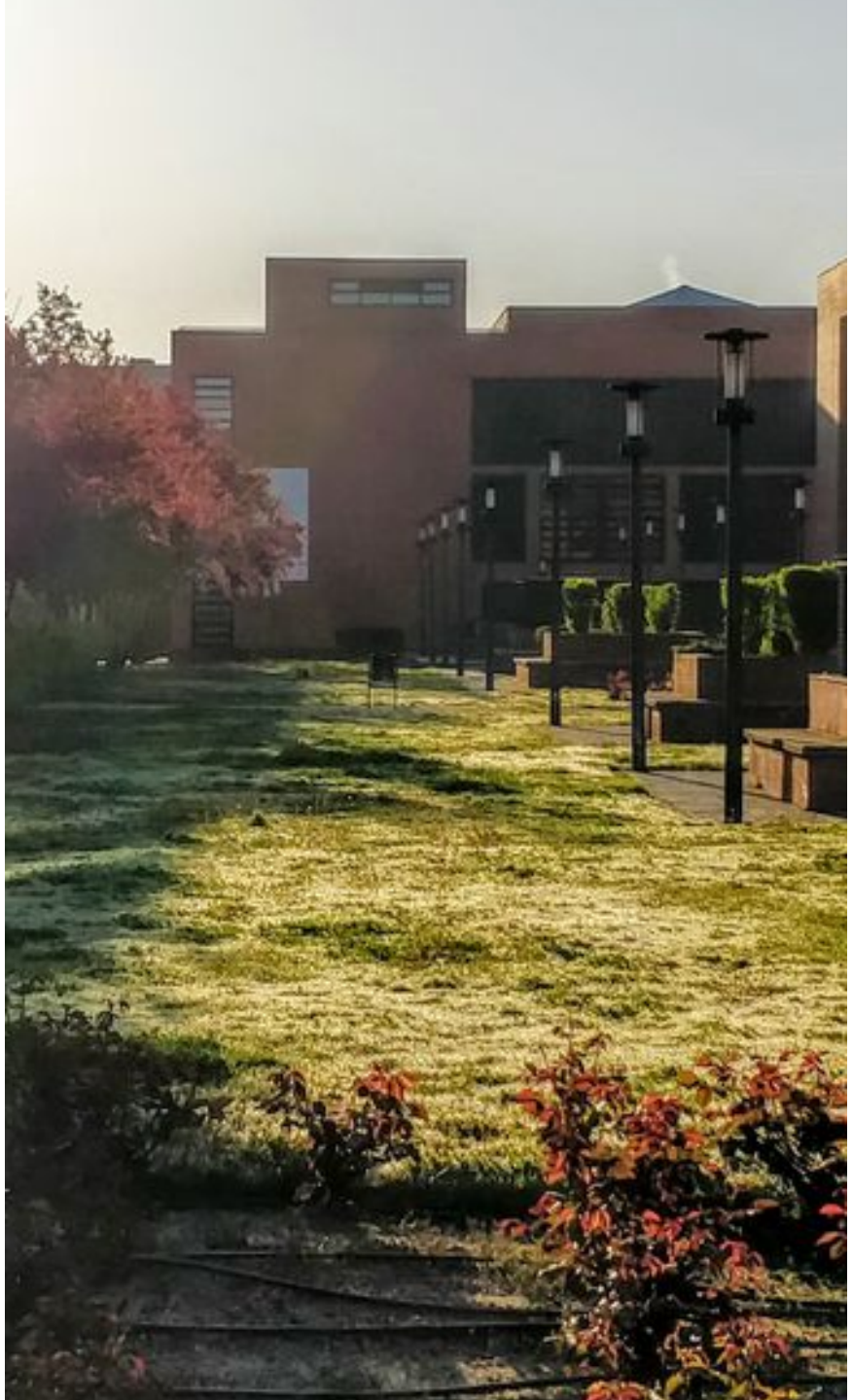

What now?

Biomedical data TODAY

- It becomes cheaper to obtain biomedical (genomic) data
- How to analyze it
- Storing the data is no longer a concern
- Understand and help the patient

And here, our story begins...

ADVANCED STATISTICAL TECHNIQUES FOR HIGH-DIMENSIONAL DATA



uc3m | Universidad Carlos III de Madrid

2016

Relation between breast cancer and genetic information.

Genetics problem

A breast cancer new treatment

Breast cancer patients

Triple-negative breast cancer patients

Genetics problem

A breast cancer new treatment

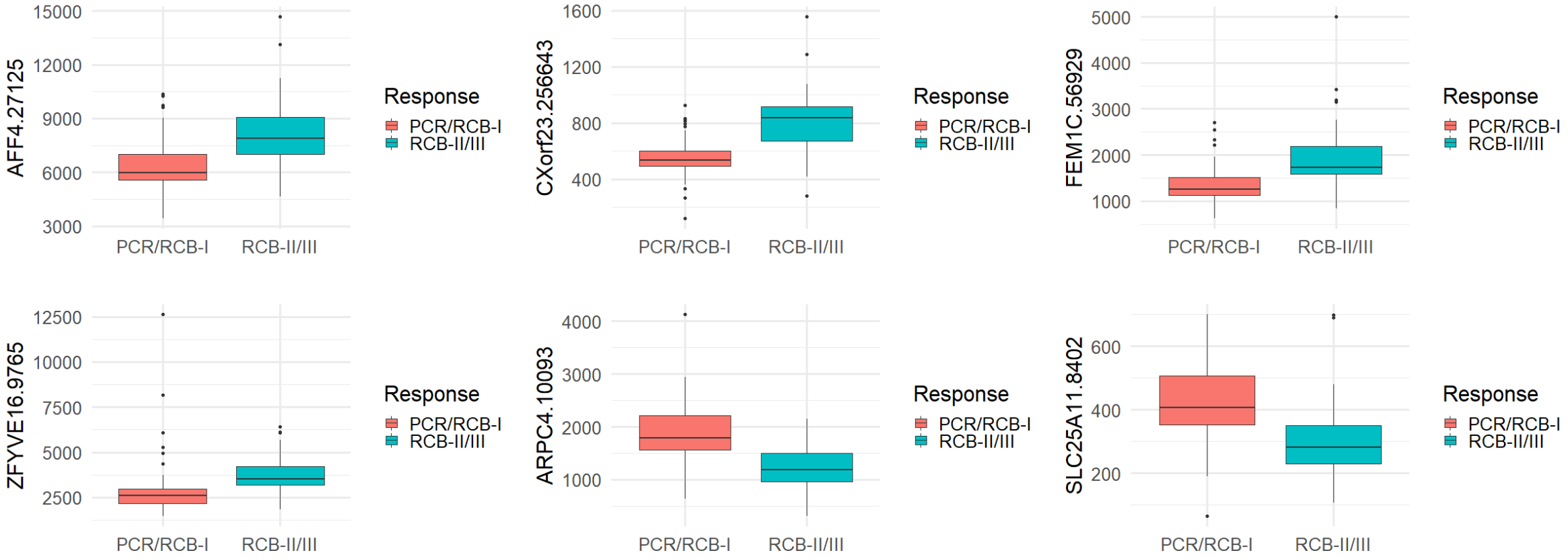
New treatment under research (Docetaxel and Carboplatino)

Is inter individual human genetic variation related to effectiveness?

95 TNBC Patients. RNA sequencing technologies retrieving 20531 gene values

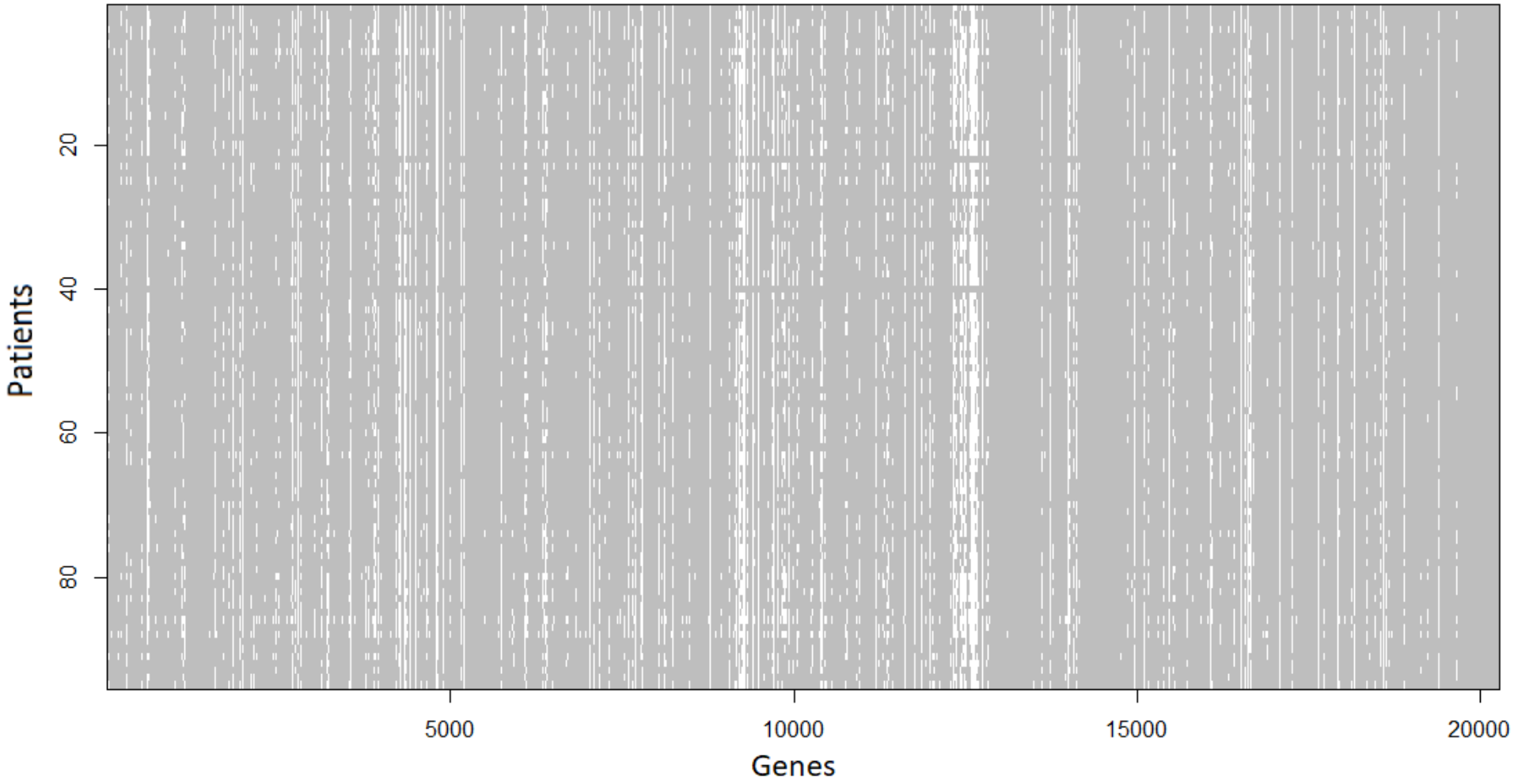
First approach: Univariate lab analysis

Difference in gene value




Data has a better idea

Matrix representation

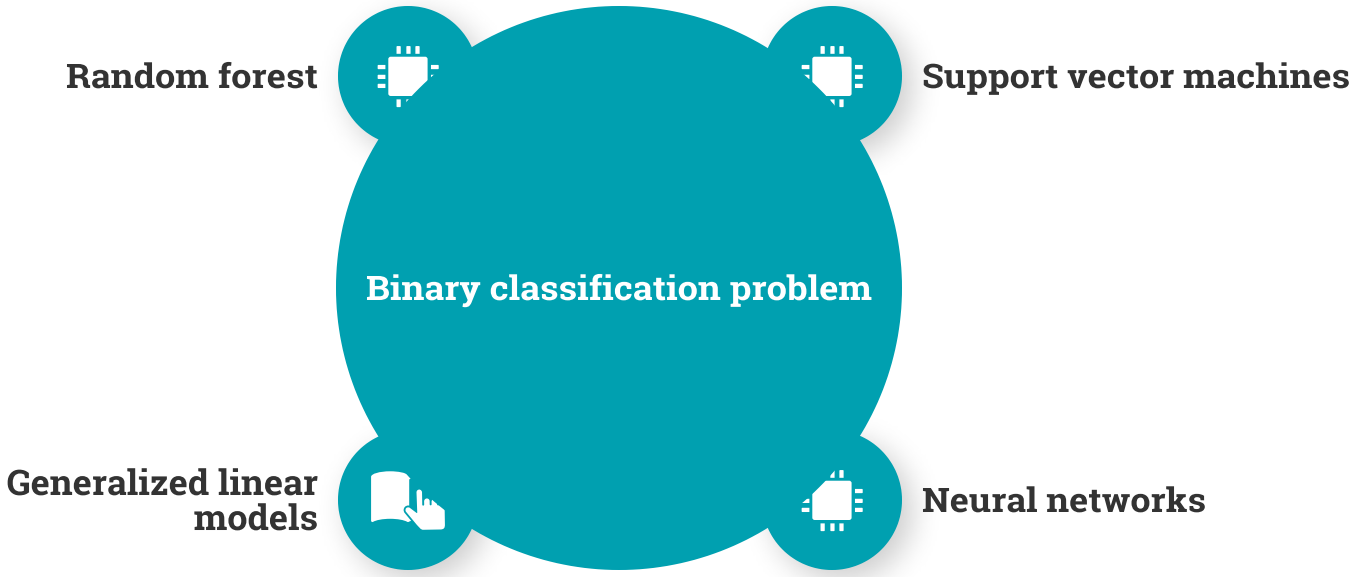


Matrix representation

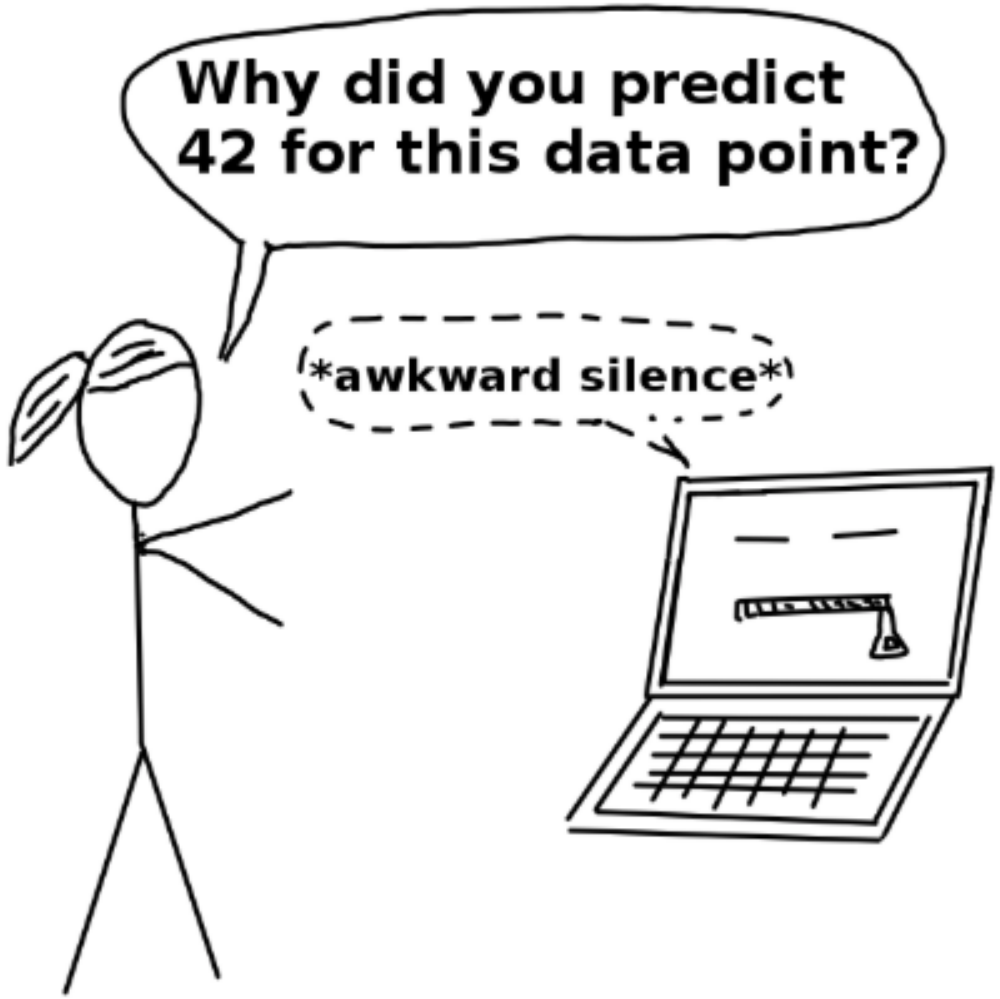
zooming in

$$X = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}; \; y = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

The statistical/data problem



- (Supervised) binary classification problem
- Final objective: “Genetic signature” able to predict effectiveness of the treatment


Why did you predict
42 for this data point?
awkward silence

The model

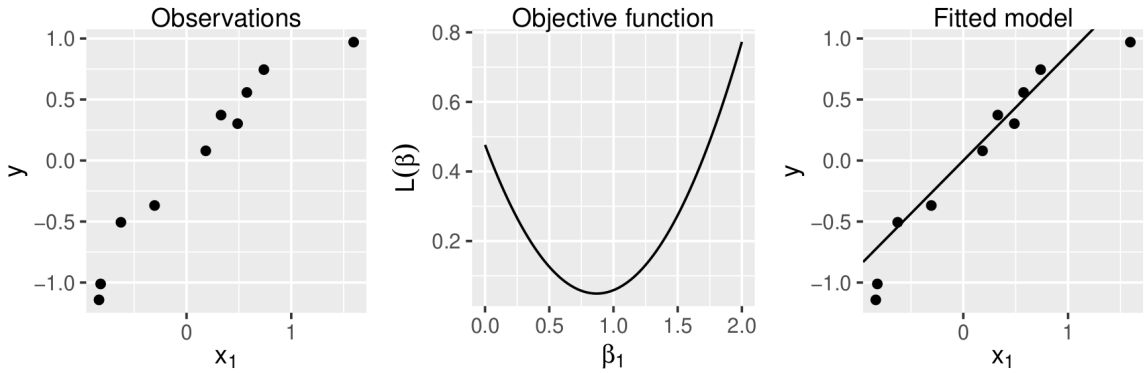
- Under the **generalized linear model** framework, we approximate $\boldsymbol{y}$ through $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$.
- The **quality** of approximations is measured using a **risk function** $L(\boldsymbol{\beta})$, that depends on the data $\mathbf{X}, \boldsymbol{y}$.
- This function $L(\boldsymbol{\beta})$ is often derived from the model's likelihood, written as a function of $\boldsymbol{\beta}$.

Basic problem

- To find an estimation of the optimal β, we **minimize** the empirical risk in the training data,

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \, L(\beta).$$

- The quality of an approximation can be evaluated using $L_{valid}(\hat{\beta})$, on a separate **validation** data set.



Common choices

- Linear regression

$$L(\beta) = \frac{1}{N} \|\boldsymbol{y} - \mathbf{X}\beta\|_2^2$$

- Logistic regression

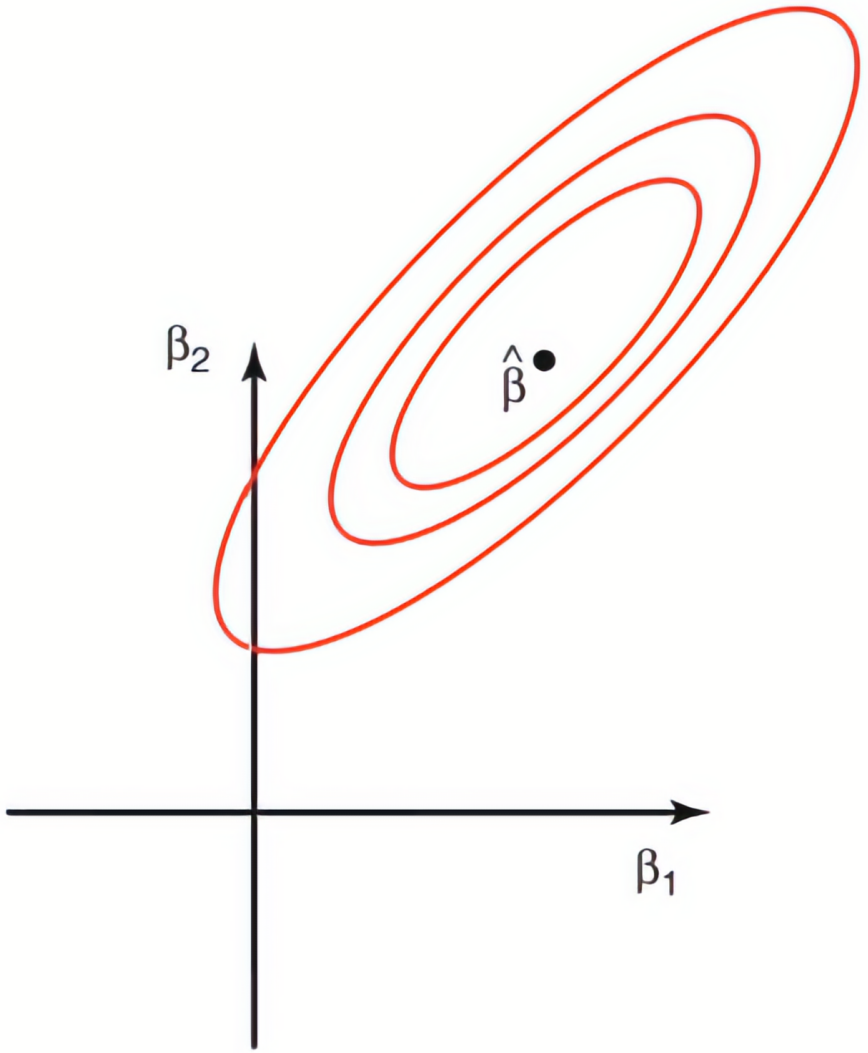
$$L(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left\{ \log \left[1 + \exp(\mathbf{x}_i^\top \beta) \right] - \boldsymbol{y}_i \mathbf{x}_i^\top \beta \right\}$$

- Cox regression

$$L(\beta) = \sum_{i \in D} \mathbf{x}_i^\top \beta - \sum_{i \in D} \log \left(\sum_{k \in R_i} \exp(\mathbf{x}_k^\top \beta) \right)$$

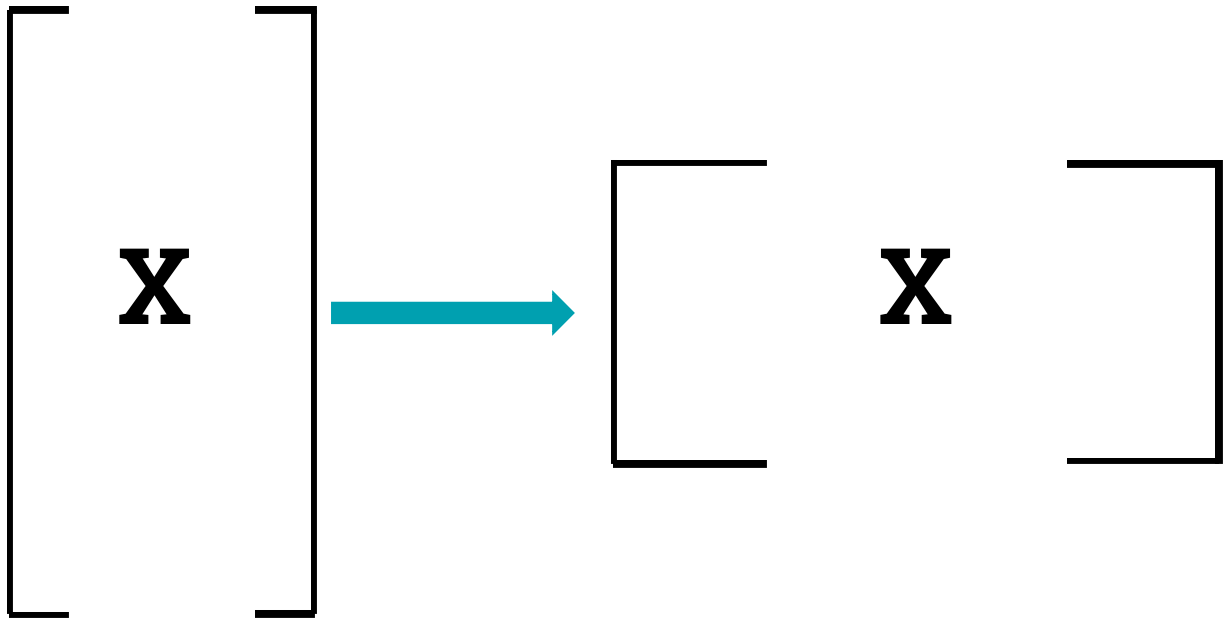
Functions that are **convex** and **differentiable**.

Maths behind regression



- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
- $\hat{\beta} = \arg\min \left\{ \|y - X\beta\|_2^2 \right\}$
- $\hat{\beta} = (X^t X)^{-1} X^t y$

High dimensional regression



$$\hat{\beta} = (X^t X)^{-1} X^t y$$

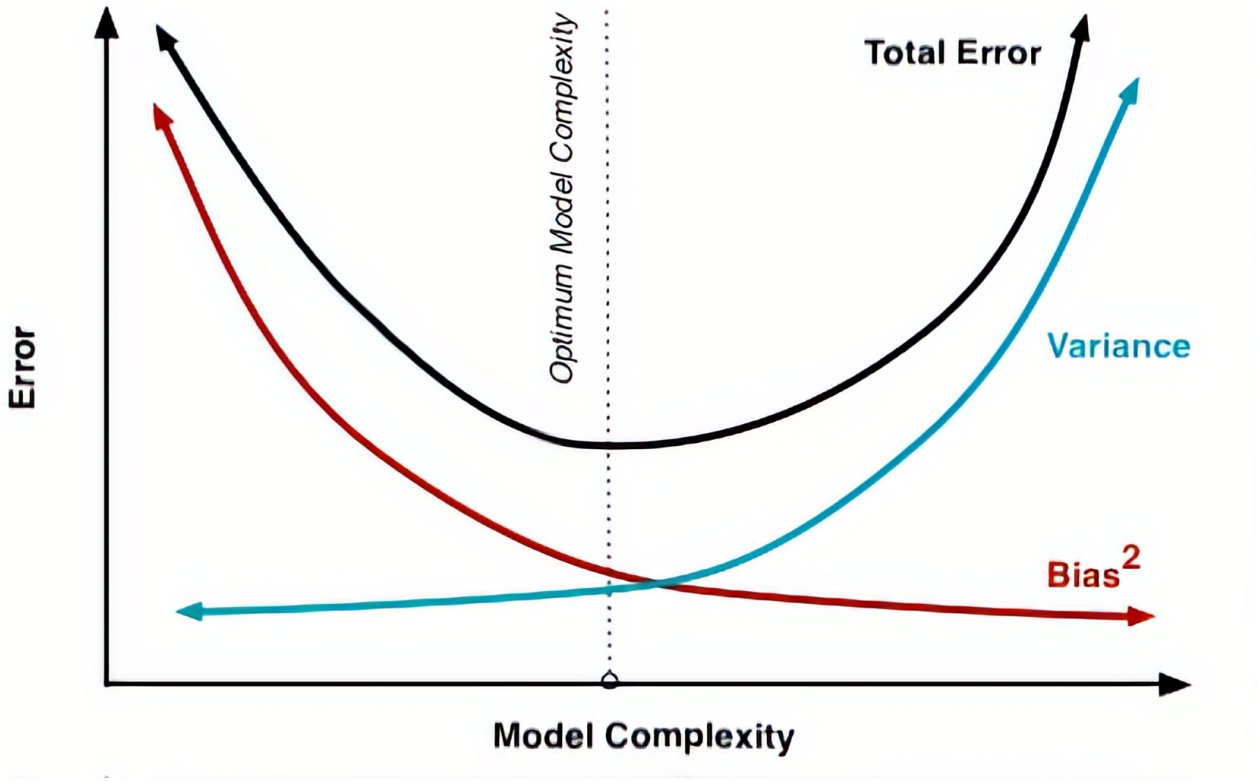
Il-posed problems

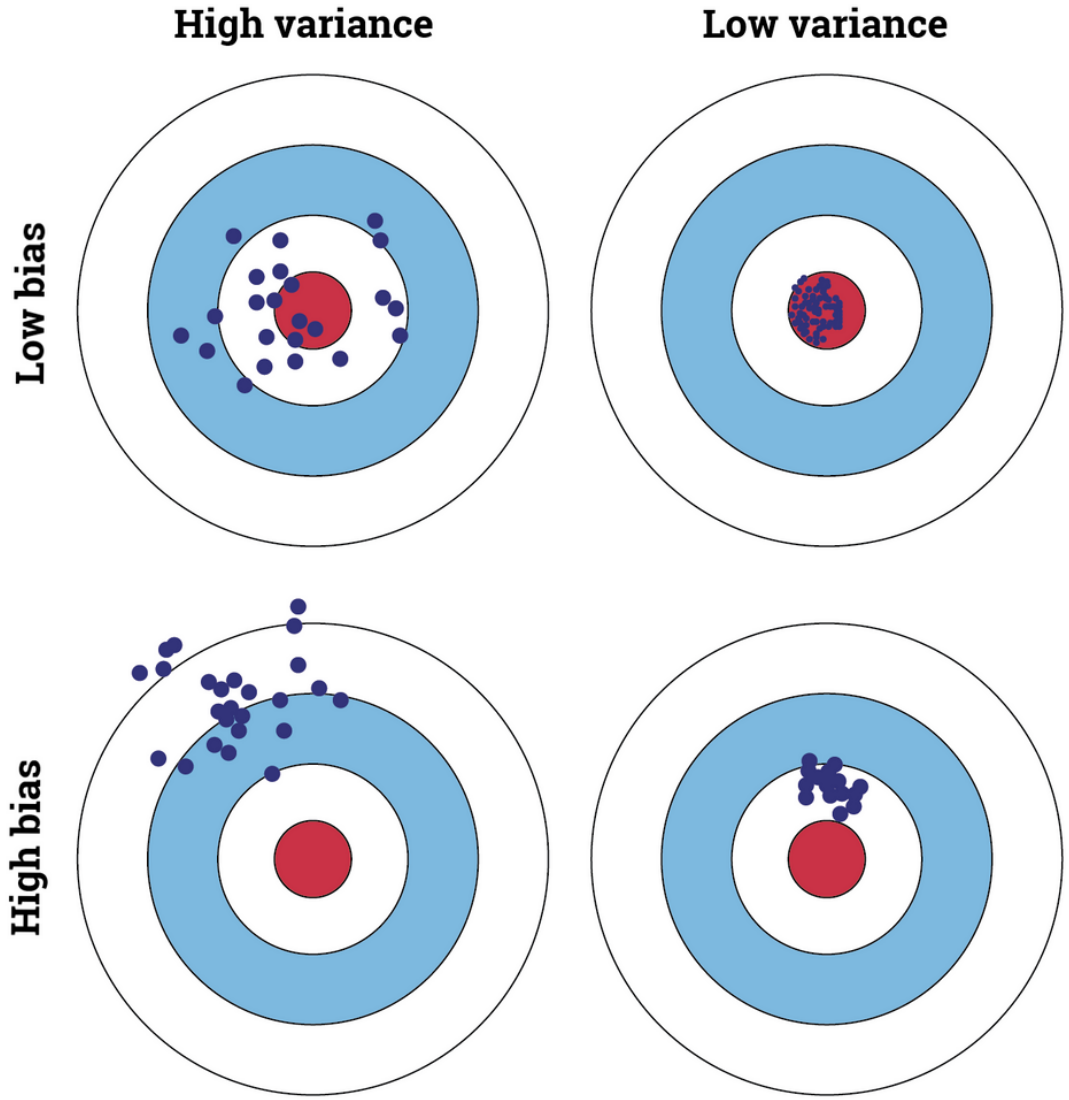
Low dimensional data → **High dimensional data**

So lets penalize the models

Variance – bias tradeoff

Penalized regression





Penalization approach

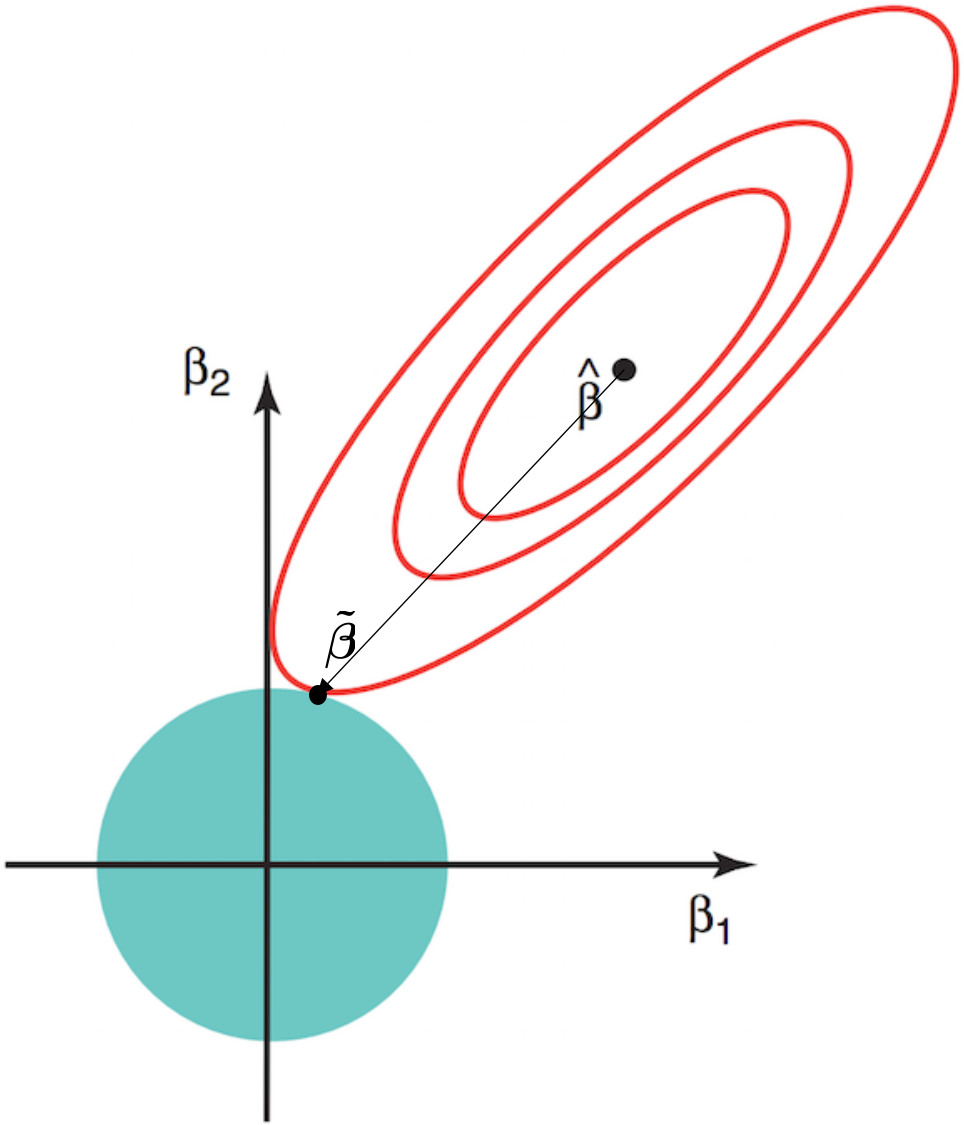
- In the case of **high-dimensionality** ($p > n$), it becomes difficult to work with enough data samples to achieve a high density of points. The **penalization** approach provides a formalism for controlling the **complexity** of the approximating functions, to fit available finite data.

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ L(\boldsymbol{\beta}) + \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) \right\}, \quad \text{where } \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) \geq 0.$$

- If $\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = \boldsymbol{\lambda}\phi(\boldsymbol{\beta})$, this problem is equivalent to,

$$\hat{\boldsymbol{\beta}}(t) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} L(\boldsymbol{\beta}), \quad \text{subject to } \phi(\boldsymbol{\beta}) \leq t, \quad t \geq 0.$$

Ridge penalization

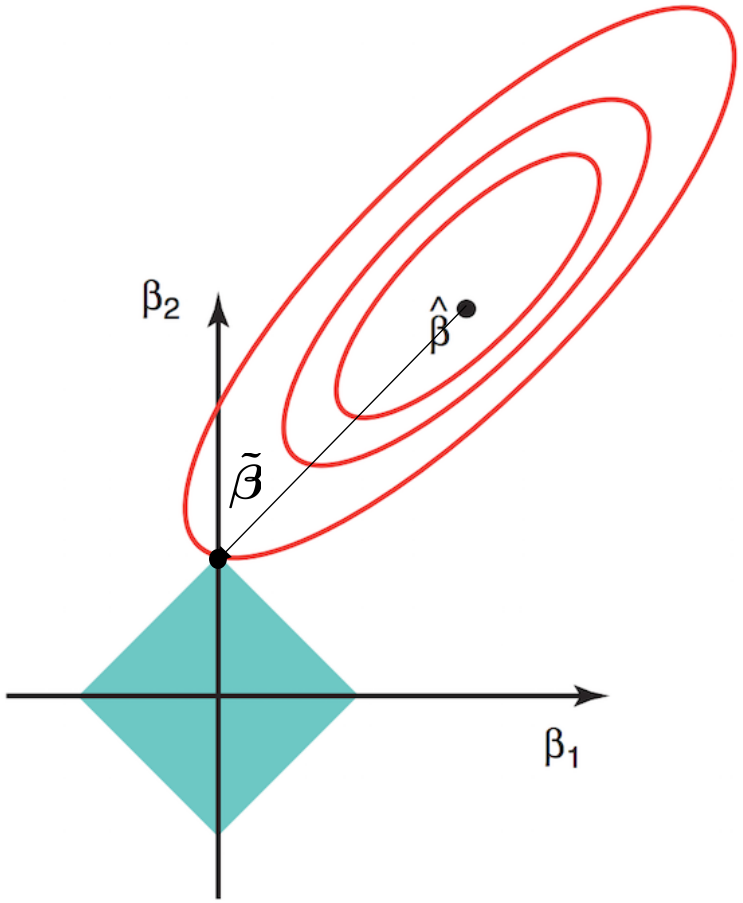


- $\min_{\vec{\beta}} \left\{ \frac{1}{2} \left\| \vec{y} - X\vec{\beta} \right\|_2^2 + \lambda \|\vec{\beta}\|_2 \right\}$

All the parameters are shrunk in the same way.

We are not selecting variables.

lasso penalization



Tibshirani R (1996)
Regression Shrinkage and Selection via the Lasso
Journal of the Royal Statistical Society. Series B (Methodological) (1996) 58(1) 267-288

- $\min_{\vec{\beta}} \left\{ \frac{1}{2} \left\| \vec{y} - X\vec{\beta} \right\|_2^2 + \lambda \|\vec{\beta}\|_1 \right\}$

Some parameters are shrunk to 0.

Regularization= variable selection.

Biological interpretation

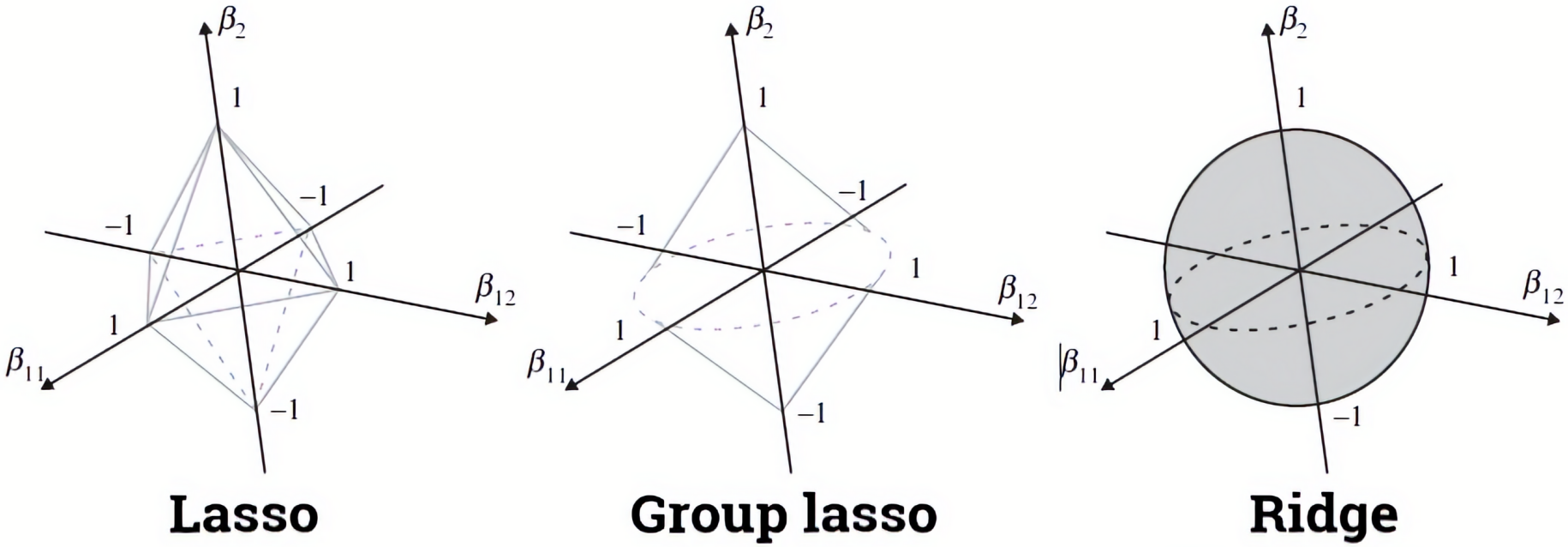
We require to select variables

Genes work in groups

Group lasso penalization

- $$\min_{\vec{\beta}} \left\{ \frac{1}{2} \left\| \vec{y} - X\vec{\beta} \right\|_2^2 + \lambda \sum_{k=1}^{K} \sqrt{p_k} \left\| \vec{\beta}^k \right\|_2 \right\}$$

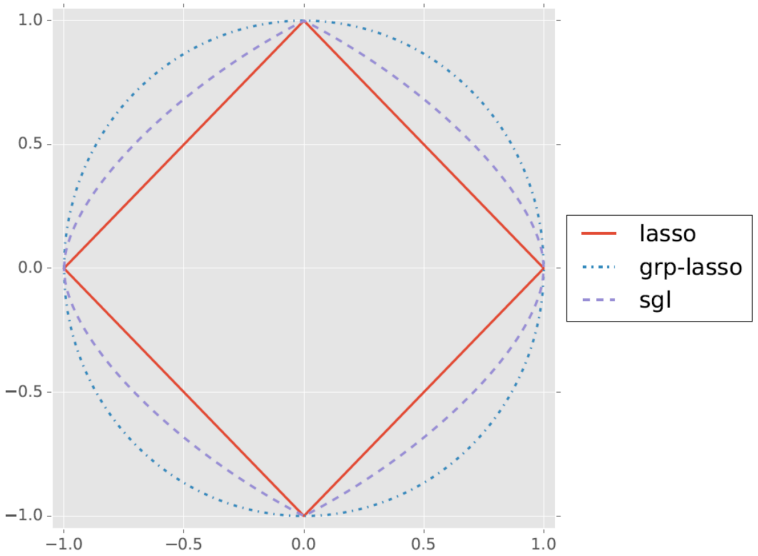
- **Lasso between groups**
- **Ridge within groups**



Yuan M, Lin Y (2006)
Model selection and estimation in regression with grouped variables
Journal of the Royal Statistical Society. Series B (Methodological) (2006) 68(1) 49-67

Sparse group lasso penalization

- $$\min_{\vec{\beta}} \left\{ R(\vec{\beta}) + \alpha\lambda \left\| \vec{\beta} \right\|_1 + (1-\alpha)\lambda \sum_k^K \sqrt{p_k} \left\| \vec{\beta^k} \right\|_2 \right\}$$



Simon N, Friedman J, Tibshirani R (2013)

A sparse-group lasso
Journal of Computational and Graphical Statistics (2013) 22(2) 231-245

Linear combination of lasso and group lasso.

Sparsity both between and within groups of variables.


BUT we need to create the groups
Or an unsupervised classification problem


Our contributions start!

Common penalty functions for variable selection

$\boldsymbol{\lambda}$	$\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta})$	**Regularization**
λ	$\lambda\|\boldsymbol{\beta}\|_1$	Lasso
λ, α	$(1-\alpha)\lambda\|\boldsymbol{\beta}\|_2^2 + \alpha\lambda\|\boldsymbol{\beta}\|_1$	Elastic-net
λ	$\lambda \sum_{j=1}^{J} \sqrt{p_j}\|\boldsymbol{\beta}^{(j)}\|_2$	Group-lasso
λ, α	$(1-\alpha)\lambda \sum_{j=1}^{J} \sqrt{p_j}\|\boldsymbol{\beta}^{(j)}\|_2 + \alpha\lambda\|\boldsymbol{\beta}\|_1$	Sparse-group lasso
$\lambda_1, \lambda_2, \boldsymbol{\gamma}$	$\lambda_2 \sum_{j=1}^{J} \gamma_j\|\boldsymbol{\beta}^{(j)}\|_2 + \lambda_1\|\boldsymbol{\beta}\|_1$	Sparse-group lasso

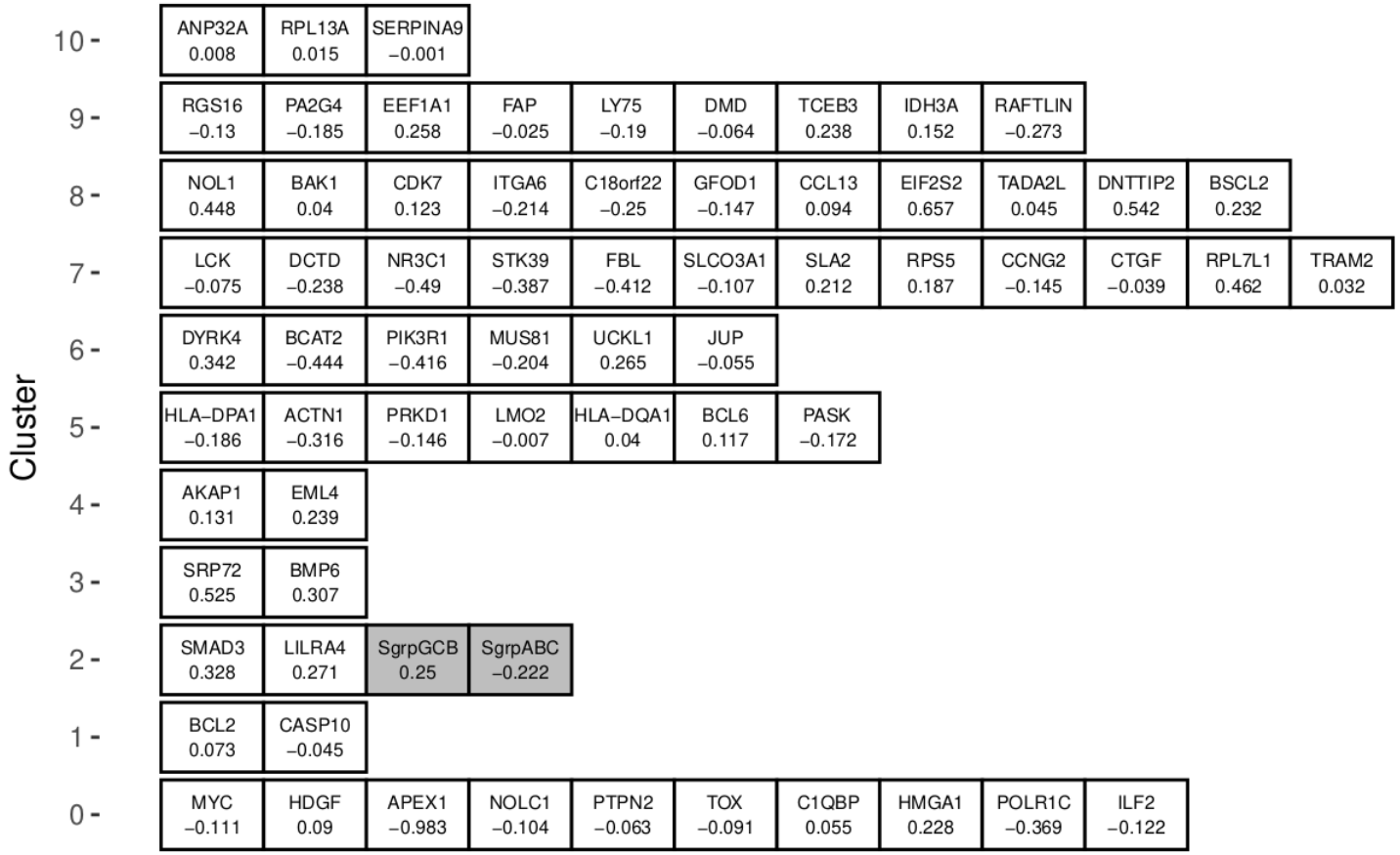
Sparse Group Lasso formulation

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left[L(\boldsymbol{\beta}) + \lambda_2 \overbrace{\sum_{j=1}^{J} \gamma_j \|\boldsymbol{\beta}^{(j)}\|_2}^{\text{Group Lasso}} + \lambda_1 \underbrace{\|\boldsymbol{\beta}\|_1}_{\text{Lasso}} \right],$$

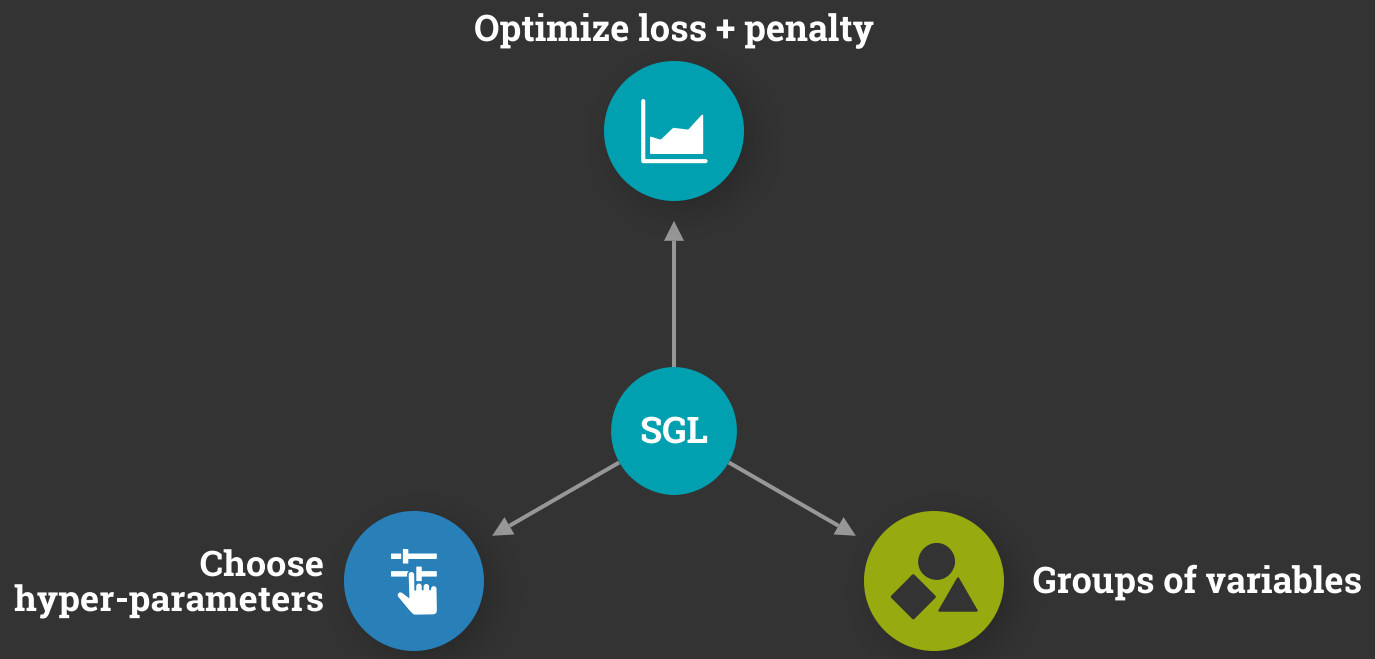
where

- Variables are **grouped** (in J predefined groups),
- $\boldsymbol{\beta}$ is the coefficient vector, with $\boldsymbol{\beta}^{(j)}$ relative to variables in j–th group,
- $L(\boldsymbol{\beta})$ is the empirical **risk** function in the training data (mean squared error, binary cross-entropy),
- $\lambda_1, \lambda_2, \gamma_j \geq 0$ are $(J+2)$ regularization **hyperparameters**.

Variables are grouped in predefined clusters

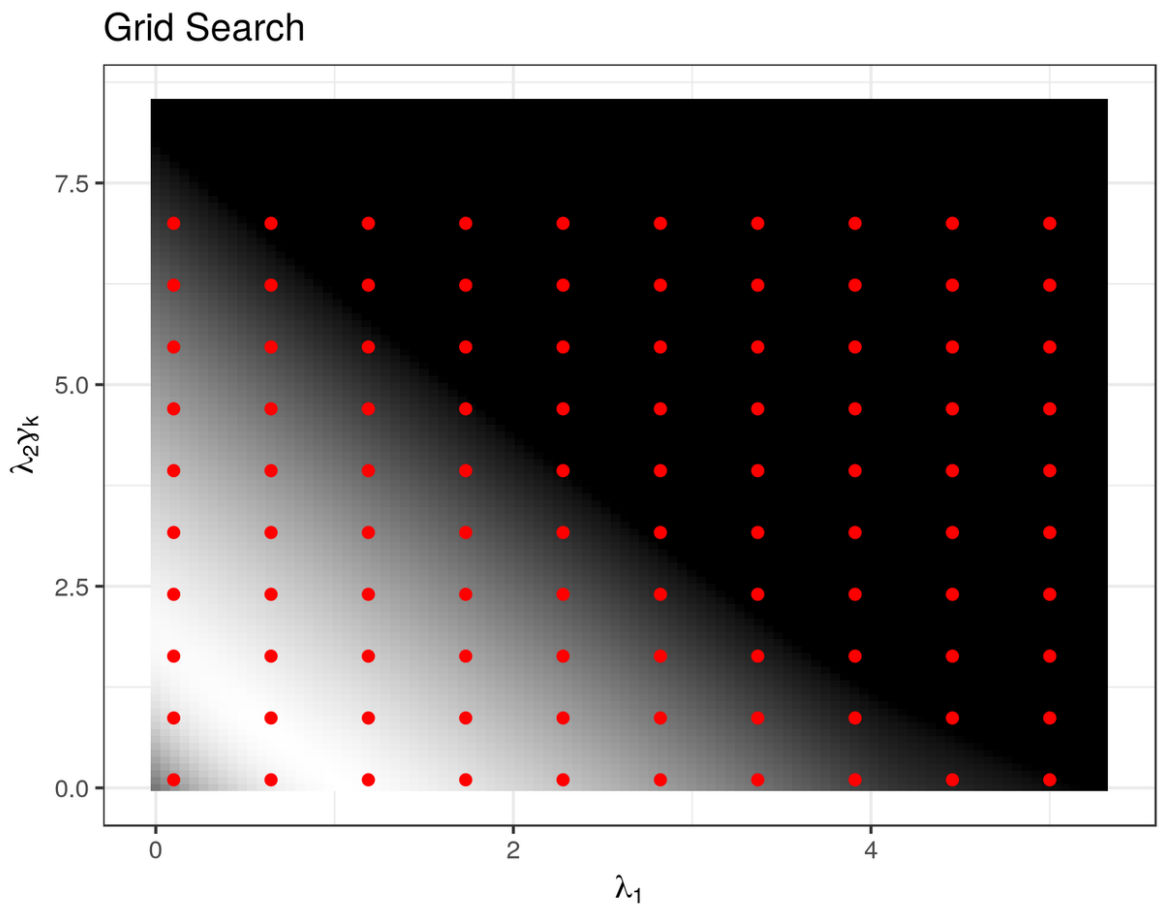


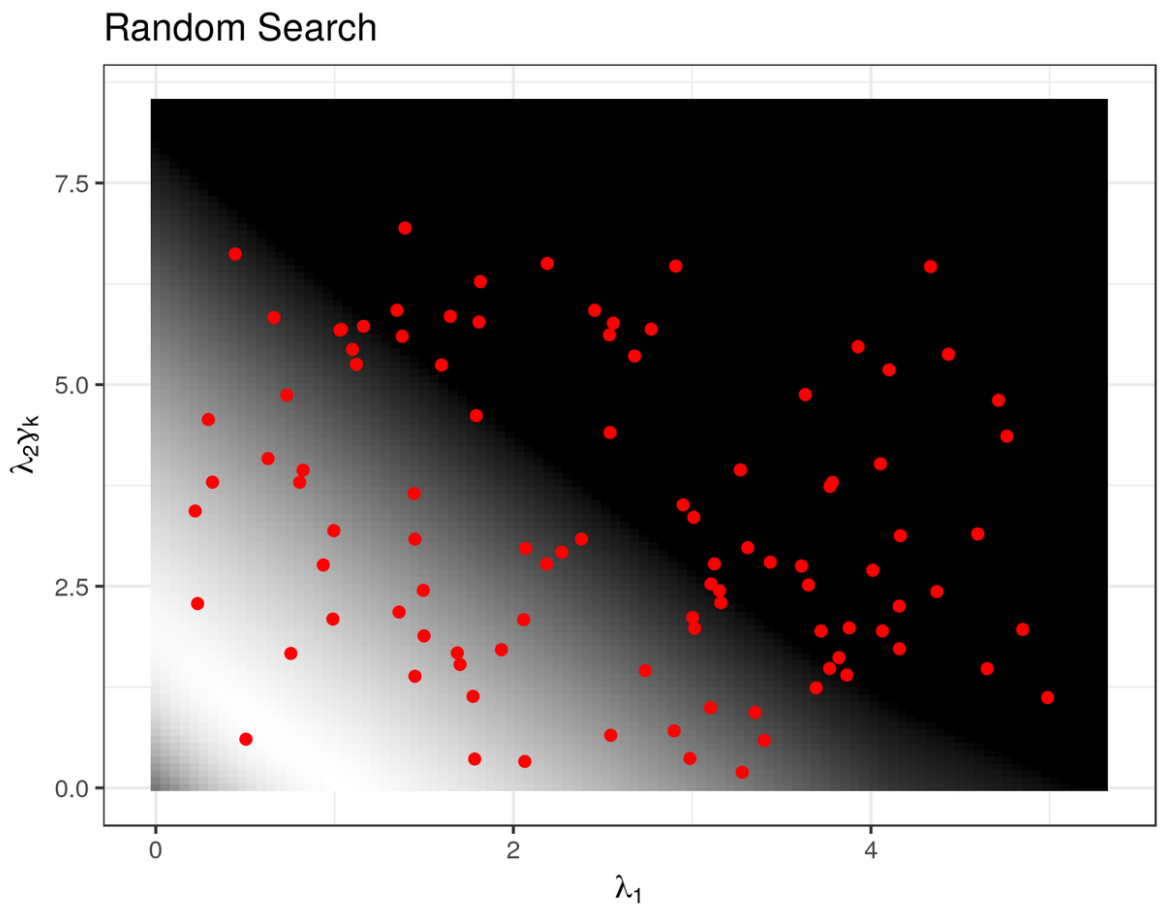
Sub-problems



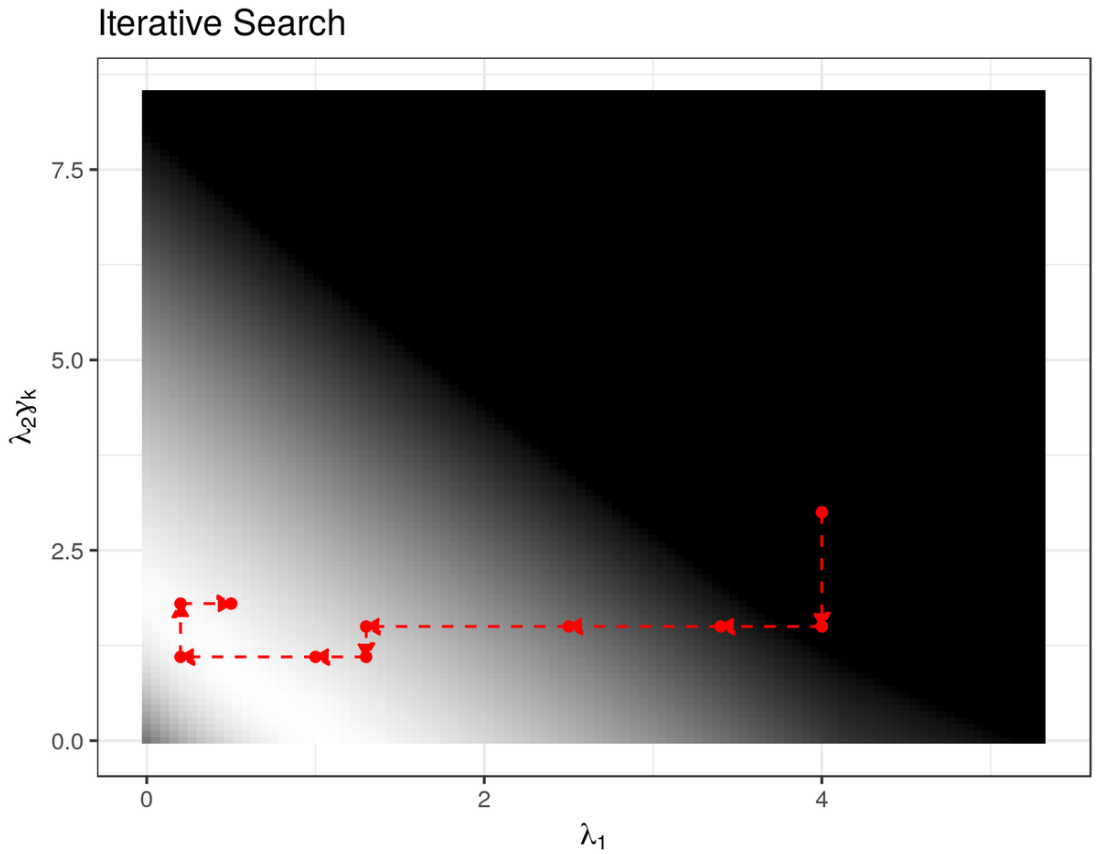
An iterative Sparse Group Lasso

Grid Search vs Random Search





The iterative Sparse Group Lasso



Juan C. Laria, M. Carmen Aguilera-Morillo & Rosa E. Lillo (2019) An Iterative Sparse-Group Lasso, Journal of Computational and Graphical Statistics. 28(3), Pages 722-731

Initial regularization hyper-parameters

Proposition 1

Consider the Sparse Group Lasso problem with coefficients fixed except those in k-th group.

$$F(\beta) = R(\beta) + \lambda_2 \gamma_k \|\beta\|_2 + \lambda_1 \|\beta\|_1.$$

$\beta^* = \mathbf{0}$ minimizes $F(\beta)$ if

$$\left\| S(\nabla R(\mathbf{0}), \lambda_1) \right\|_2 \leq \lambda_2 \gamma_k.$$

Initial regularization hyper-parameters

Proposition 2

Consider the Sparse Group Lasso problem with coefficients fixed except those in k-th group.

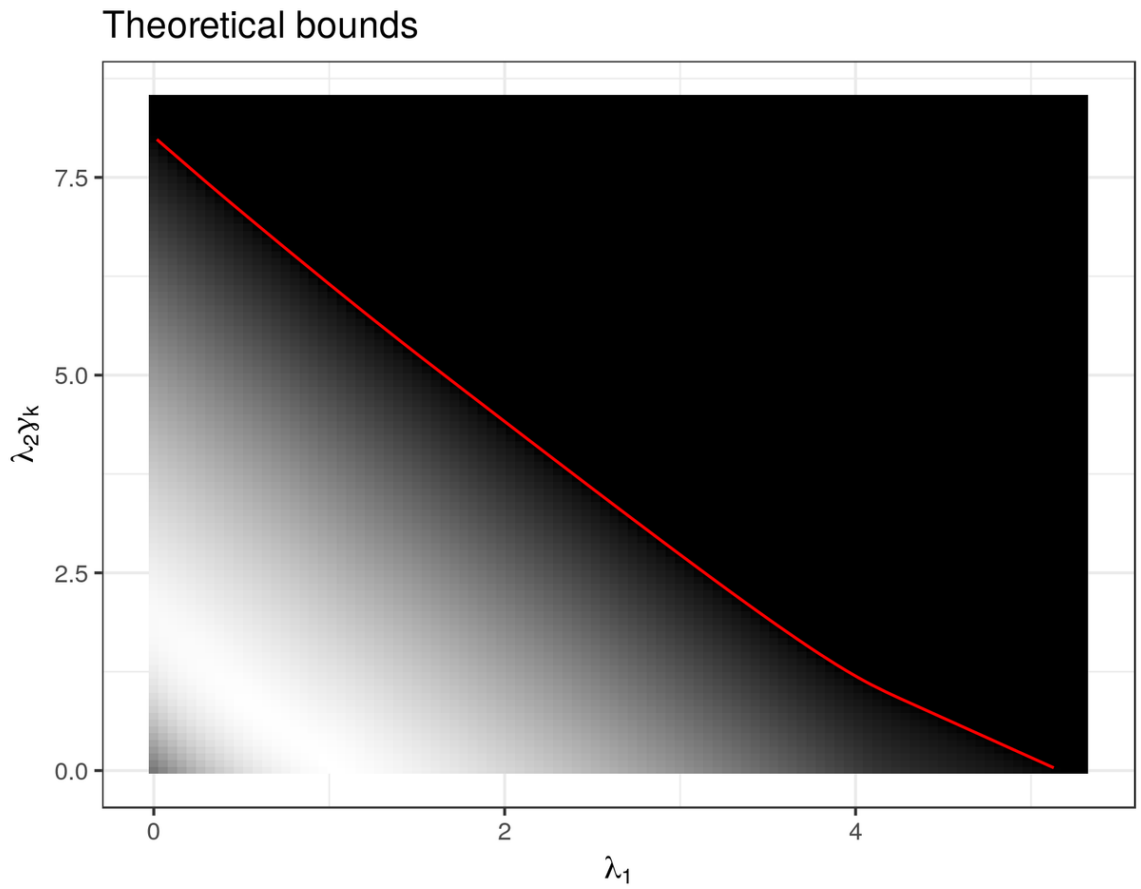
$$F(\beta) = R(\beta) + \lambda_2 \gamma_k \|\beta\|_2 + \lambda_1 \|\beta\|_1.$$

$\beta^* = \mathbf{0}$ minimizes $F(\beta)$ if

$$\max_{1 \leq i \leq p} \left| \nabla R(\mathbf{0})_i \right| \leq \lambda_1.$$

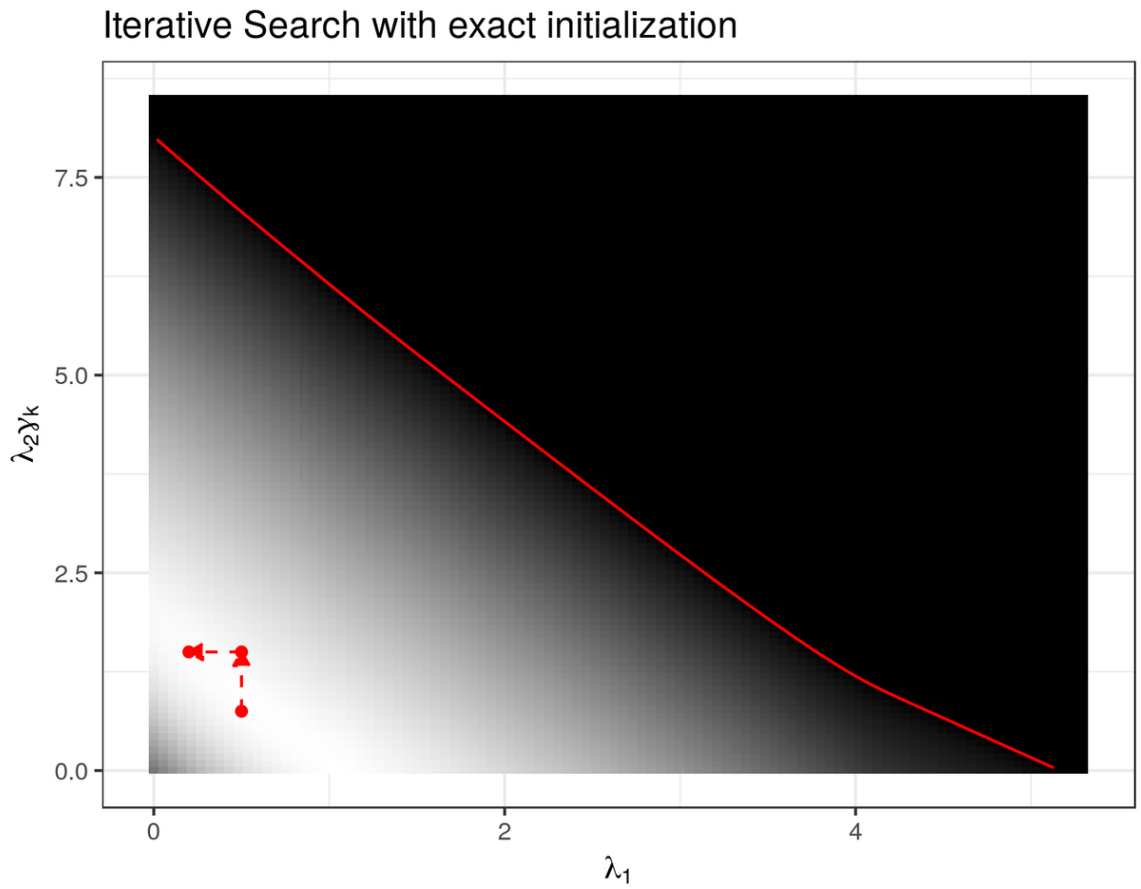
Exact method to find the boundary

Propositions 1 and 2



Juan C. Laria, M. Carmen Aguilera-Morillo & Rosa E. Lillo (2019) An Iterative Sparse-Group Lasso, Journal of Computational and Graphical Statistics. Volume 28(3), Pages 722-731

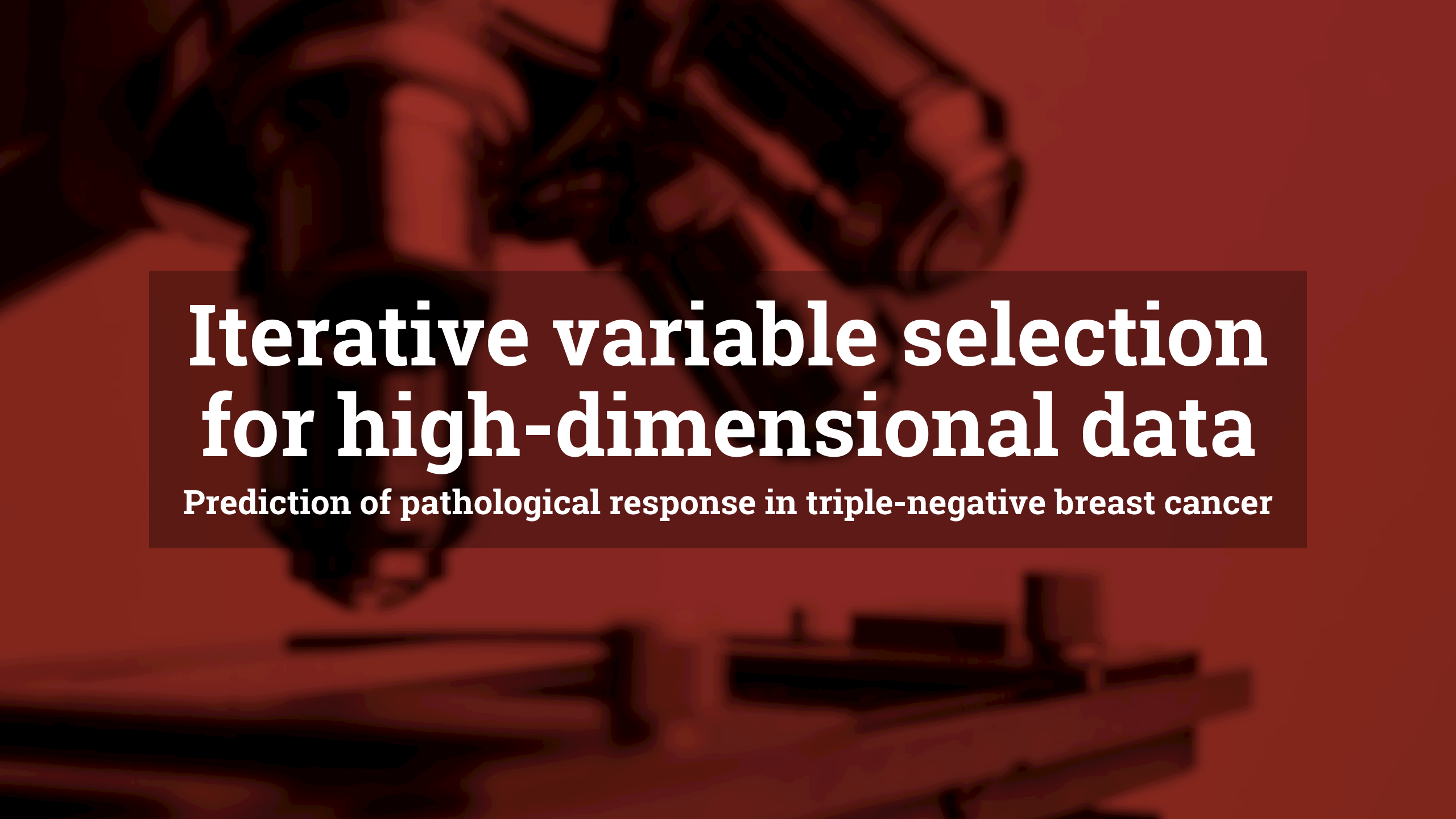
Initial regularization parameters



Juan C. Laria, M. Carmen Aguilera-Morillo & Rosa E. Lillo (2019) An Iterative Sparse-Group Lasso, Journal of Computational and Graphical Statistics. Volume 28(3), Pages 722-731

Simulations

	2 generating groups (10 non-zero coefficients)				
	$\#\lambda$	R_{valid}	R_{test}	$\|\|\hat{\beta} - \beta\|\|_2^2$	runtime(min.)
GD	201	**28.27**	48.65	20.49	466.8
GD_0	2	59.12	63.8	34.69	44.8
iSGL	202	**28.96**	**43.21**	**15.15**	27.1
$iSGL_0$	2	57.97	61.35	32.14	**1.3**
NM	2	58.63	63.38	34.21	82
GS	2	53.58	59.21	30.35	16.6
RS	2	53.57	59.51	30.69	36.7
	3 generating groups (15 non-zero coefficients)				
	$\#\lambda$	R_{valid}	R_{test}	$\|\|\hat{\beta} - \beta\|\|_2^2$	runtime(min.)
GD	201	57.09	84.79	45.33	445.3
GD_0	2	132.39	128.24	88.48	43.5
iSGL	202	**49.83**	**77.3**	**36.15**	38.2
$iSGL_0$	2	127.54	121.92	82.04	**1.2**
NM	2	131.22	126.87	87.34	75.8
GS	2	120.18	119.49	79.1	22.6
RS	2	120.27	118.92	78.36	46.1



Iterative variable selection for high-dimensional data

Prediction of pathological response in triple-negative breast cancer

Biomedical application

Hospital General Universitario Gregorio Marañón

SaludMadrid

Triple negative breast cancer study

RNA-Seq analysis

p > 20K genes and N < 100 samples

Upfront (neoadjuvant) chemotherapy has a 50% chance of success

Laria Juan C., Aguilera-Morillo M. Carmen, Álvarez Enrique,Lillo Rosa E., López-Taruella Sara, del Monte-Millán María,Picornell Antonio C., Martín Miguel and Romo Juan (2021).“Iterative variable selection high-dimensional data: Predictionof pathological response in triple-negative breast cancer”.• Mathematics, https://doi.org/10.3390/math9030222

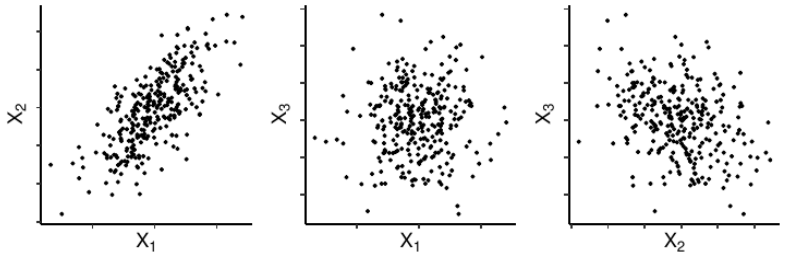
Methodology

Model estimation with iterative Sparse Group Lasso

Grouping variables using PCA

Selecting the best model with a variable importance index

Grouping variables using PCA



	PC1	PC2
X_1	**-0.67**	0.40
X_2	**-0.70**	-0.08
X_3	0.23	**0.91**

A variable importance index for classification

Algorithm 4:

/* sample data $\mathcal{Z}$, # of runs R */

Function `isgl`($\mathcal{Z}$, R)**:**

  for r *in* $1, 2 \dots R$ **do**

    $\mathcal{Z}_T, \mathcal{Z}_V \leftarrow$ random partition of $\mathcal{Z}$

    $\beta^{(r)} \leftarrow \text{ISGL}(Z_T, Z_V)$

    $ccr_V^{(r)} \leftarrow$ Correct classification rate of $\beta^{(r)}$ in $\mathcal{Z}_V$

  end

  return β, ccr_V

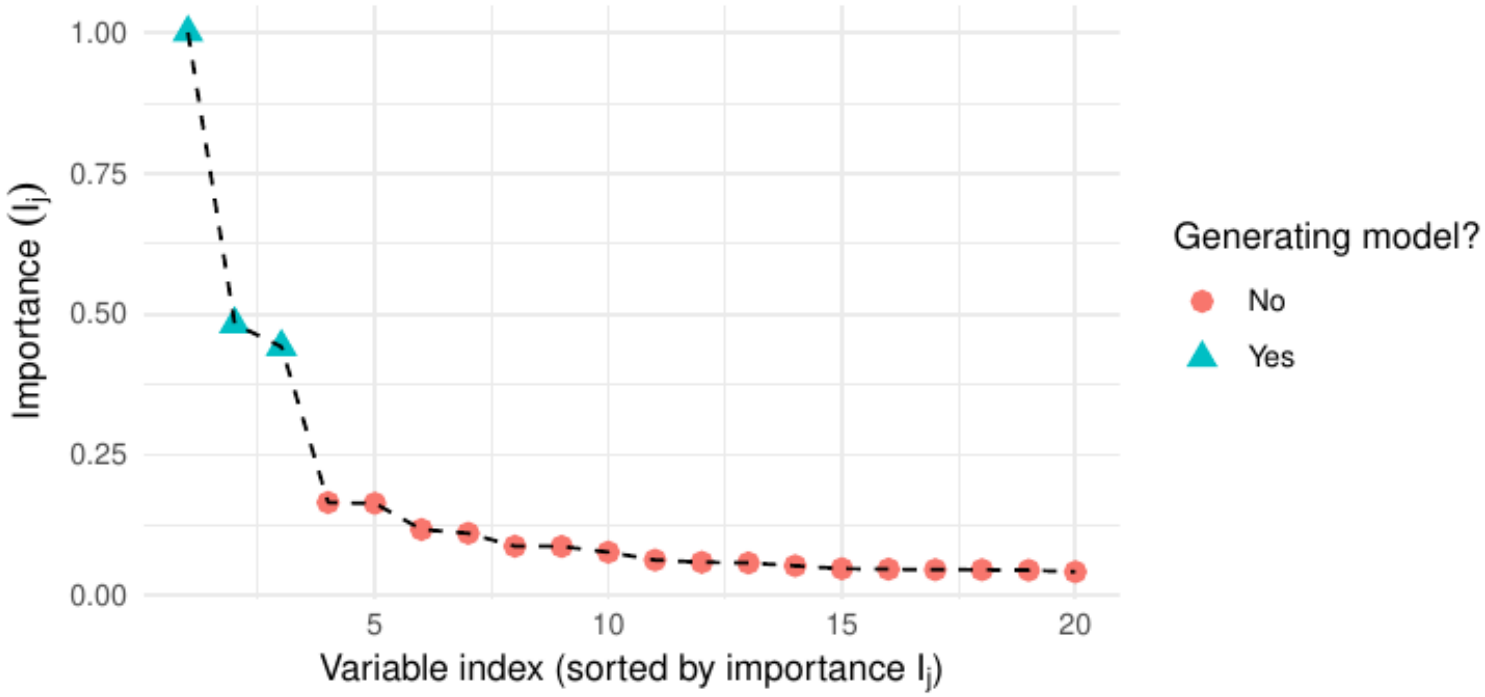
We consider the **importance index** I_j of variable X_j defined as,

$$I_j = \sum_{r=1}^{R} |\beta_j^{(r)}| \cdot (ccr_V^{(r)} - \delta) / \max_j \left\{ \sum_{r=1}^{R} |\beta_j^{(r)}| \cdot (ccr_V^{(r)} - \delta) \right\},$$

where

- $\beta^{(r)}$ and $ccr_V^{(r)}$ are those returned by Algorithm 4.
- δ is the null model's accuracy.

Why an importance index?



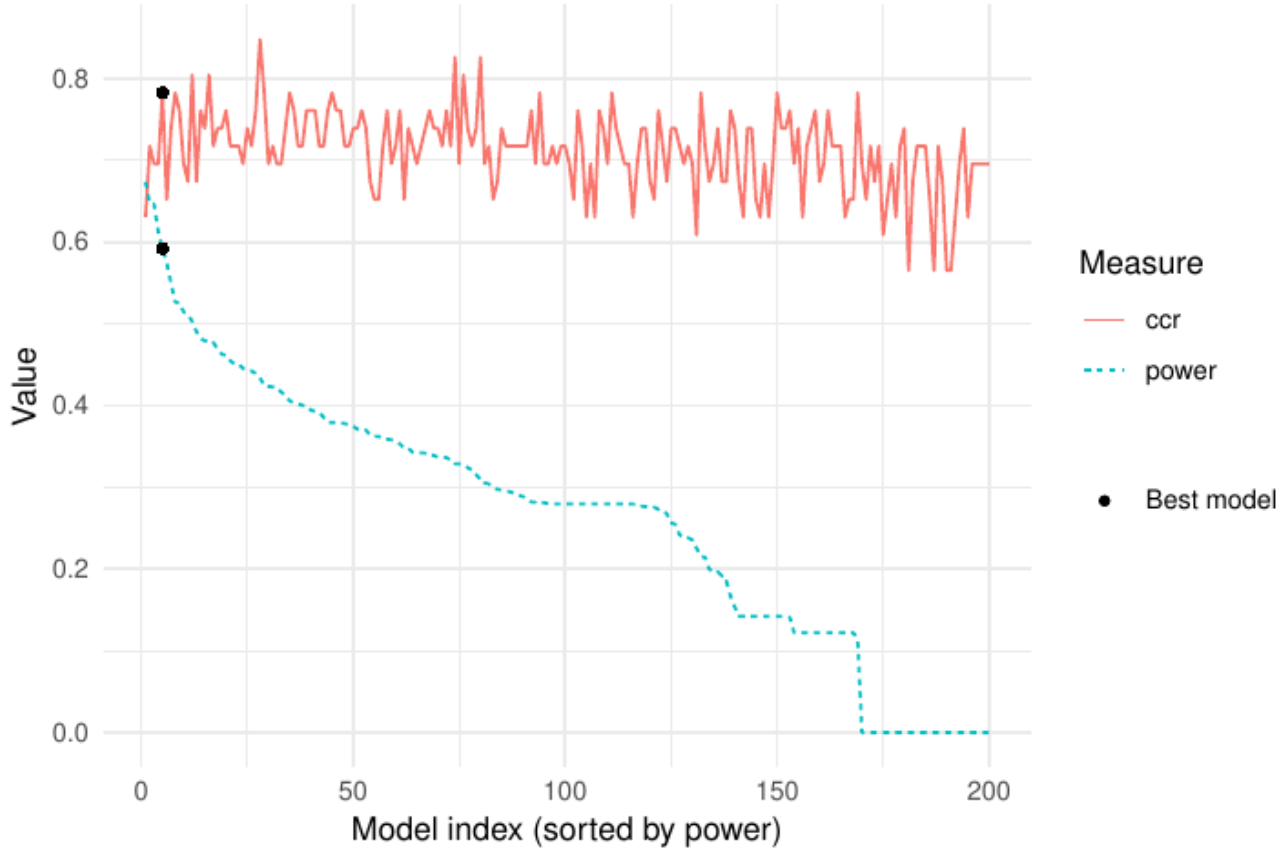
Power of a model

- Using the important index of the best K variables, we define the **power** of a model as,

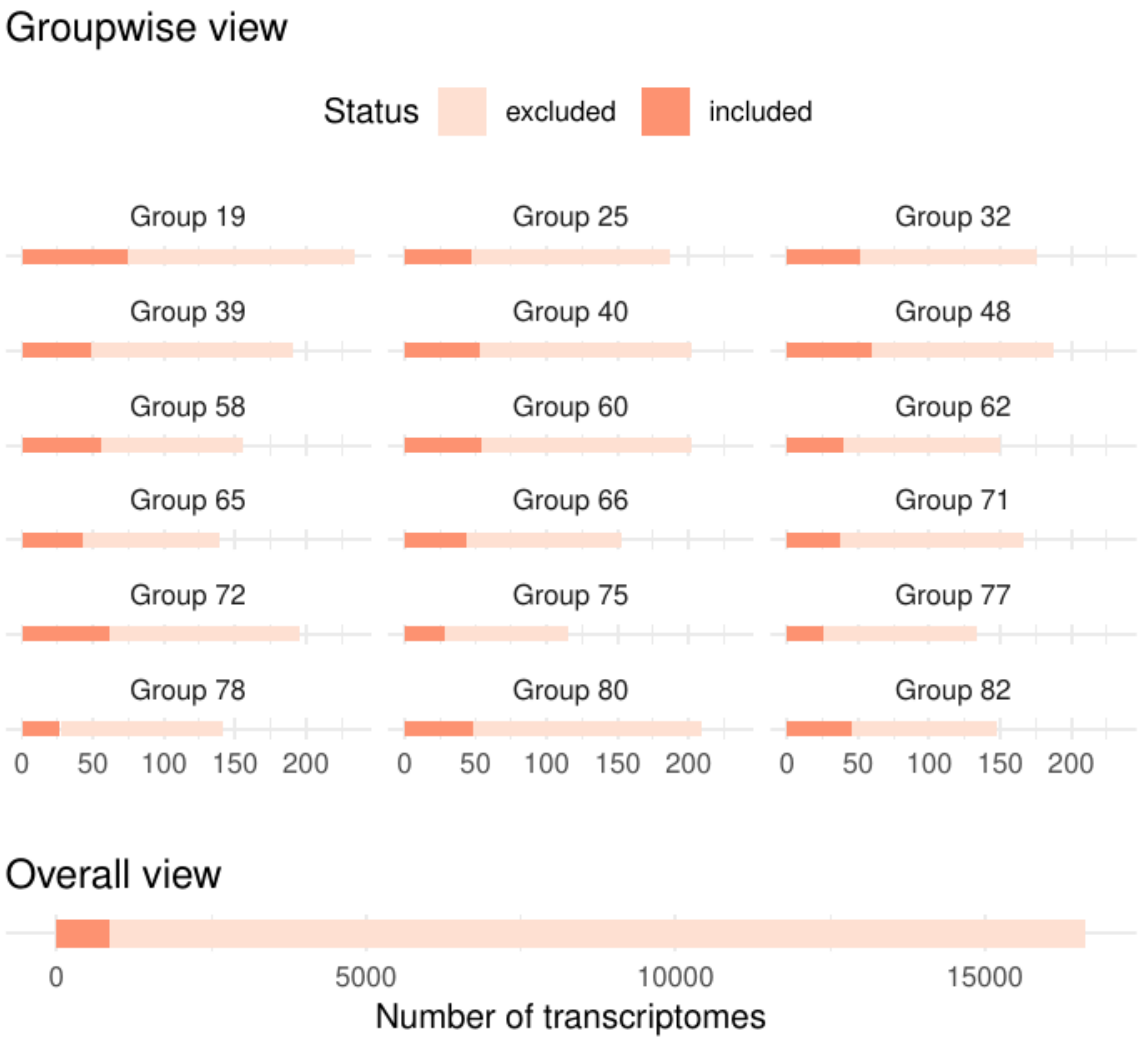
$$P_r = \frac{1}{\sum_{k=1}^{K} I_{(k)}} \sum_{j: I_j \geq I_{(K)}} I_j |\beta_j^{(r)}| / \|\boldsymbol{\beta}^{(r)}\|_1, \quad r = 1, 2 \ldots R,$$

  where $I_{(k)}$ denotes the $k-$th greatest **importance index**.
- The **power index** P weights each model, depending on the importance of its included variables.

Power of a model



Application to Biomedical Data

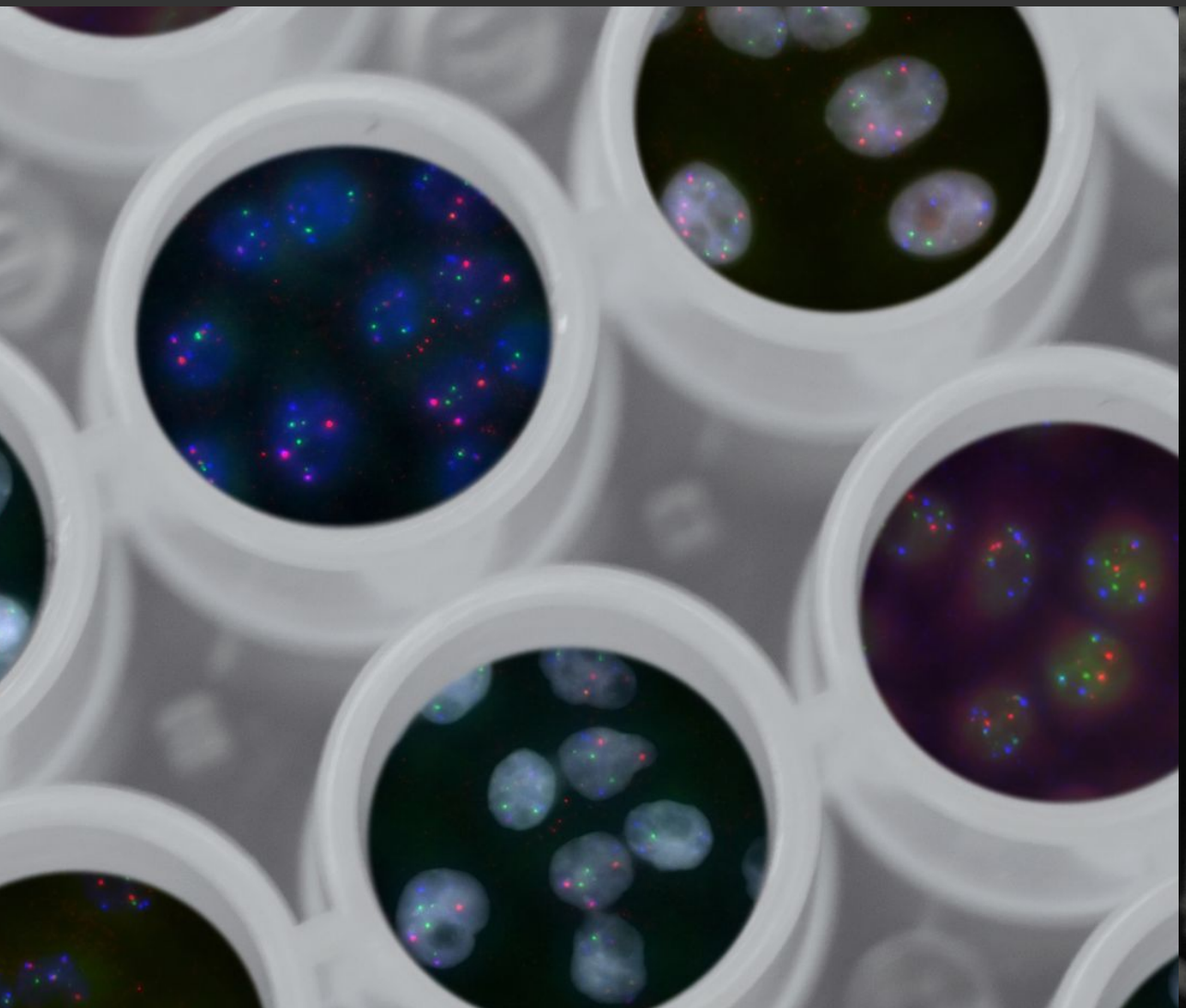


Group Linear Algorithm with Sparse Principal decomposition

A variable selection and clustering method for generalized linear models

Laria, J.C., Aguilera-Morillo, M.C. and Lillo, R.E. Group linear algorithm with sparse principal decomposition: a variable selection and clustering method for generalized linear models. (2023) Statistical Papers, 64, 227-253 https://doi.org/10.1007/s00362-022-01313-z

Motivation




EUR/USD - 1,35379 - 00:00:00 14 giu (EEST)
EUR/USD (Bid), Ticks, # 300 / 300
Gold, spot - 1.276,820 - 23:00:00 13 giu (CEST)
Gold, spot (Bid), 1 minute, # 159 / 300, Logarithmic, Heikin Ashi
Quote List [2]
World Markets

Assumptions

Sparsity

Clustering

Structure

Sparsity

There is a **small** number of columns of $\mathbf{X}$ that are actually related to $\boldsymbol{y}$, and therefore many components of β are **exactly** zero.

Clustering

There is a (possibly unknown) number K of **unknown** groups, or clusters, among the variables of $\mathbf{X}$.

Structure

- For every group, there is associated a **latent variable** that summarizes the information provided by all the variables in that cluster.
- In linear models, information is measured in terms of **linear predictors**.
- A variable $\mathbf{X}_j$ provides **information** to the model through $\mathbf{X}_j\beta_j$.
- Knowing those groups will improve the estimation of β, and knowing β will give us **insight** into the groups.

Solving the sparse **regression** problem and, at the same time, finding the **clusters** in the columns of $\boldsymbol{X}$, motivates the **GLASP** optimization problem,

$$\min_{\beta, \boldsymbol{W}, \boldsymbol{T}} \left\{ L(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \sum_{k=1}^{K} \|\boldsymbol{J}_k \beta\|_2 + \frac{\lambda_3}{2} \left\| \boldsymbol{X} \sum_{j=1}^{p} (e_j e_j^\top) \beta_j - \boldsymbol{T}\boldsymbol{W}^\top \right\|_F^2 \right\},$$

where

- $\|\cdot\|_F^2$ is the squared **Frobenius** norm, given by $\|\boldsymbol{M}\|_F^2 = Tr(\boldsymbol{M}\boldsymbol{M}^\top)$.
- $\boldsymbol{W} \in \mathbb{R}^{p \times K}$ is an **orthogonal** matrix with cluster information, $\boldsymbol{W}^\top \boldsymbol{W}$ diagonal.
- $\boldsymbol{T} \in \mathbb{R}^{N \times K}$ (latent groups) is a low-rank **unitary** representation of the linear predictors, $\boldsymbol{T}^\top \boldsymbol{T} = \boldsymbol{I}_K$.
- e_j is the j–th vector in the canonical basis of $\mathbb{R}^{p \times 1}$.
- $\boldsymbol{X} \sum_{j=1}^{p} (e_j e_j^\top) \beta_j \in \mathbb{R}^{N \times p}$ is the matrix of **linear predictors**.

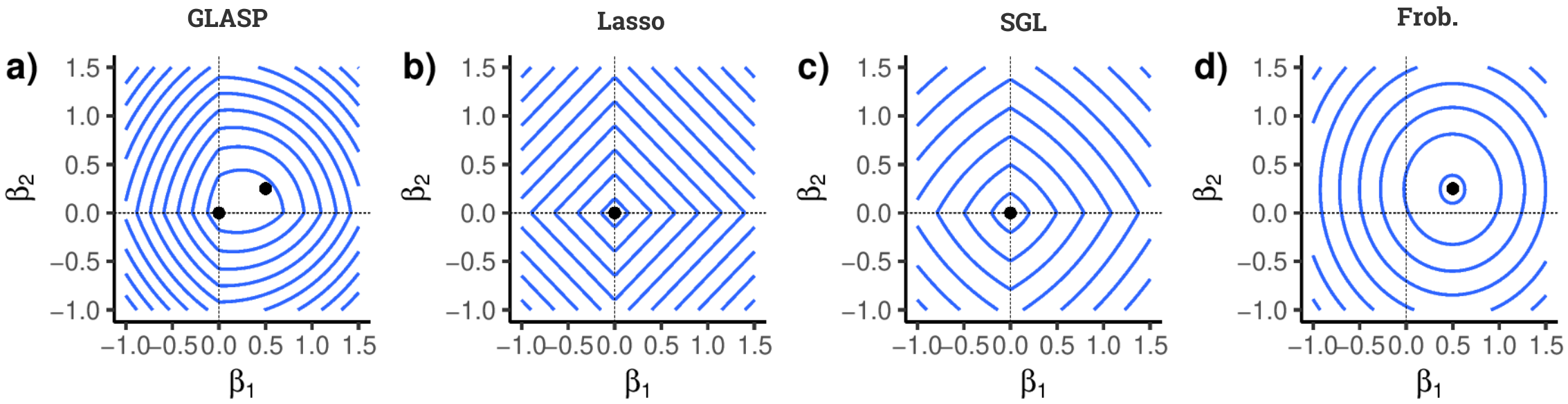
Solving the sparse **regression** problem and, at the same time, finding the **clusters** in the columns of $\boldsymbol{X}$, motivates the **GLASP** optimization problem,

$$\min_{\beta, \boldsymbol{W}, \boldsymbol{T}} \left\{ L(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \sum_{k=1}^{K} \|\boldsymbol{J}_k \beta\|_2 + \frac{\lambda_3}{2} \left\| \boldsymbol{X} \sum_{j=1}^{p} (e_j e_j^\top) \beta_j - \boldsymbol{T}\boldsymbol{W}^\top \right\|_F^2 \right\},$$

where

- $\boldsymbol{J}_k = \|\boldsymbol{W}_k\|_0 \sum_{j=1}^{p} (e_j e_j^\top) \mathbb{1}(W_{jk} \neq 0)$ is a **diagonal projection** matrix such that $\|\boldsymbol{J}_k \beta\|_2$ is the euclidean norm of the vector of coefficients associated with group k, penalized by the size of the group. Here $\|\boldsymbol{W}_k\|_0$ denotes the number of elements in column k–th of $\boldsymbol{W}$ that are non-zero, which is the size of group k.
- $\lambda_1, \lambda_2, \lambda_3$ are regularization hyperparameters.

GLASP penalty



Two-step iterative solution

GLASP can be separated in two optimization sub-problems.

$$\min_{\boldsymbol{\beta}} \left\{ L(\boldsymbol{\beta}) + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{k=1}^{K} \|\boldsymbol{J}_k \boldsymbol{\beta}\|_2 + \frac{\lambda_3}{2} \left\| \boldsymbol{X} \sum_{j=1}^{p} (e_j e_j^\top) \beta_j - \boldsymbol{T}\boldsymbol{W}^\top \right\|_F^2 \right\},$$

and

$$\min_{\boldsymbol{W},\boldsymbol{T}} \left\{ \frac{\lambda_3}{2} \left\| \boldsymbol{X} \sum_{j=1}^{p} (e_j e_j^\top) \beta_j - \boldsymbol{T}\boldsymbol{W}^\top \right\|_F^2 + \lambda_2 \sum_{k=1}^{K} \|\boldsymbol{J}_k \boldsymbol{\beta}\|_2 \right\}.$$

Step 1

If the groups are fixed (except group k), then

$$\min_{\boldsymbol{\beta}} \left\{ L(\boldsymbol{\beta}) + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sqrt{p_k} \|\boldsymbol{\beta}\|_2 + \frac{\lambda_3}{2} \sum_{j=1}^{p_k} \left\| \boldsymbol{X}_j \beta_j - \boldsymbol{T}\boldsymbol{W}_{j\cdot}^\top \right\|_2^2 \right\},$$

is a particular case of

$$\min_{\boldsymbol{\beta}} \{ F(\boldsymbol{\beta}) := R(\boldsymbol{\beta}) + \Phi(\boldsymbol{\beta}) \},$$

where

$$R(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) + \frac{\lambda_3}{2} \sum_{j=1}^{p_k} \left\| \boldsymbol{X}_j \beta_j - \boldsymbol{T}\boldsymbol{W}_{j\cdot}^\top \right\|_2^2,$$

and

$$\phi(\boldsymbol{\beta}) = \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sqrt{p_k} \|\boldsymbol{\beta}\|_2 .$$

To solve it, we will use the **fast iterative shrinkage-thresholding algorithm**.

Step 2

Given the optimization problem,

$$\min_{\boldsymbol{W},\boldsymbol{T}} \left\{ \frac{\lambda_3}{2} \left\| \boldsymbol{X} \sum_{j=1}^{p} (e_j e_j^\top) \beta_j - \boldsymbol{T}\boldsymbol{W}^\top \right\|_F^2 + \lambda_2 \sum_{k=1}^{K} \|\boldsymbol{J}_k \boldsymbol{\beta}\|_2 \right\}.$$

Consider the simpler problem,

$$\min_{\boldsymbol{u},\boldsymbol{v}} \left\{ \left\| \boldsymbol{M} - \boldsymbol{u}\boldsymbol{v}^\top \right\|_F^2 + \gamma \left(\sum_{j=1}^{p} \beta_j^2 \mathbb{1}(\boldsymbol{v}_j \neq 0) \|\boldsymbol{v}\|_0 \right)^{1/2} \right\}.$$

This is a special case of **one-rank regularized singular value decomposition**.
Although the regularization term is discontinuous, an iterative solution is possible.

Step 2

Given the optimization problem,

$$\min_{\boldsymbol{W},\boldsymbol{T}} \left\{ \frac{\lambda_3}{2} \left\| \boldsymbol{X} \sum_{j=1}^{p} (e_j e_j^\top) \beta_j - \boldsymbol{T}\boldsymbol{W}^\top \right\|_F^2 + \lambda_2 \sum_{k=1}^{K} \|\boldsymbol{J}_k \boldsymbol{\beta}\|_2 \right\}.$$

Consider the simpler problem,

$$\min_{\boldsymbol{u},\boldsymbol{v}} \left\{ \left\| \boldsymbol{M} - \boldsymbol{u}\boldsymbol{v}^\top \right\|_F^2 + \gamma \left(\sum_{j=1}^{p} \beta_j^2 \mathbb{1}(\boldsymbol{v}_j \neq 0) \|\boldsymbol{v}\|_0 \right)^{1/2} \right\}.$$

This is a special case of **one-rank regularized singular value decomposition**.
Although the regularization term is discontinuous, an iterative solution is possible.

Step 2

The optimal $\boldsymbol{v}$ is such that, for $l = 1, 2 \ldots p$,

$$\boldsymbol{v}_l = (\boldsymbol{M}^\top \boldsymbol{u})_l \mathbb{1}\left((\boldsymbol{M}^\top \boldsymbol{u})_l^2 > \gamma \left(C_{\beta,v}^{(-l)} + \beta_l^2 \right)^{1/2} \left(C_v^{(-l)} + 1 \right)^{1/2} - \gamma \left(C_{\beta,v}^{(-l)} C_v^{(-l)} \right)^{1/2} \right),$$

where

$$C_{\beta,v}^{(-l)} = \sum_{\substack{j=1 \\ j \neq l}}^{p} \beta_j^2 \mathbb{1}(\boldsymbol{v}_j \neq 0), \quad C_v^{(-l)} = \sum_{\substack{j=1 \\ j \neq l}}^{p} \mathbb{1}(\boldsymbol{v}_j \neq 0).$$

Step 2

Algorithm 5: One-rank regularized singular value decomposition (*1rSVD*).

Result: $\boldsymbol{u}, \boldsymbol{v}$ that minimize (4.15)
Input: $\boldsymbol{M}, \boldsymbol{\beta}$
Compute $\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}, s$ that minimize $\left\| \boldsymbol{M} - \hat{\boldsymbol{u}} s \hat{\boldsymbol{v}}^\top \right\|_F^2$ (one-rank SVD).
Initialize $\boldsymbol{u} \leftarrow \hat{\boldsymbol{u}};\ \boldsymbol{v} \leftarrow s\hat{\boldsymbol{v}}$
while $\boldsymbol{v}$ *not stationary* **do**
    Update $\boldsymbol{v}$ with (4.16), cyclically iterating component-wise until convergence.
    Update $\boldsymbol{u} \leftarrow \boldsymbol{M}\boldsymbol{v} / \|\boldsymbol{M}\boldsymbol{v}\|_2$
end

Algorithm 6: Regularized singular value decomposition.

Result: $\boldsymbol{W}, \boldsymbol{T}$ that minimize (4.4)
Input: $\boldsymbol{M}, \boldsymbol{\beta}, K$
for $k = 1 \ldots K$ **do**
    $\boldsymbol{u}, \boldsymbol{v} \leftarrow$ 1rSVD$(\boldsymbol{M}, \boldsymbol{\beta})$ (Solve the one-rank SVD problem)
    Set $\boldsymbol{T}_k \leftarrow \boldsymbol{u};\ \boldsymbol{W}_k \leftarrow \boldsymbol{v}$
    $\boldsymbol{M} \leftarrow \boldsymbol{M} - \boldsymbol{u}\boldsymbol{v}^\top$ (update $\boldsymbol{M}$ with the residuals)
end

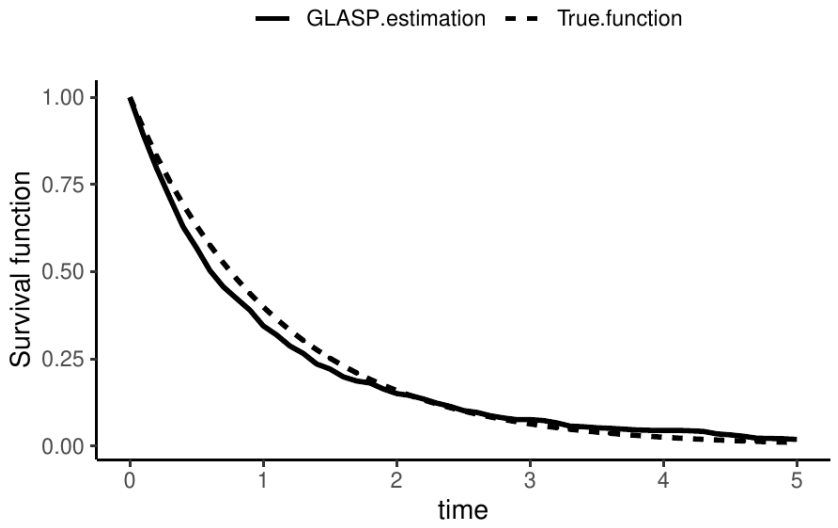
Simulations

Linear regression

	$\rho = 0.5$			
Method	RMSE	Correct Zeros	Num. Non-Zeros	RI
Lasso + Kmeans	64.632(1.583)	0.958(0.002)	91.933(2.505)	0.982(0.001)
Ridge + Kmeans	149.217(1.627)	0.05(0)	1000(0)	0.91(0)
EN + Kmeans	63.369(1.407)	0.946(0.003)	103.7(3.077)	0.984(0.001)
CEN	61.36(1.52)	0.789(0.049)	260.567(49.496)	0.988(0.001)
CEN Known Groups	52.998(1.119)	0.813(0.022)	236.667(22.255)	1(0)
Cluster Group Lasso	59.377(0.888)	0.2(0.062)	850(62.284)	0.906(0)
Group Lasso Known Groups	29.144(0.691)	0.905(0.053)	145(52.923)	1(0)
GLASP	58.516(1.757)	0.968(0.002)	82.3(1.675)	0.963(0.003)

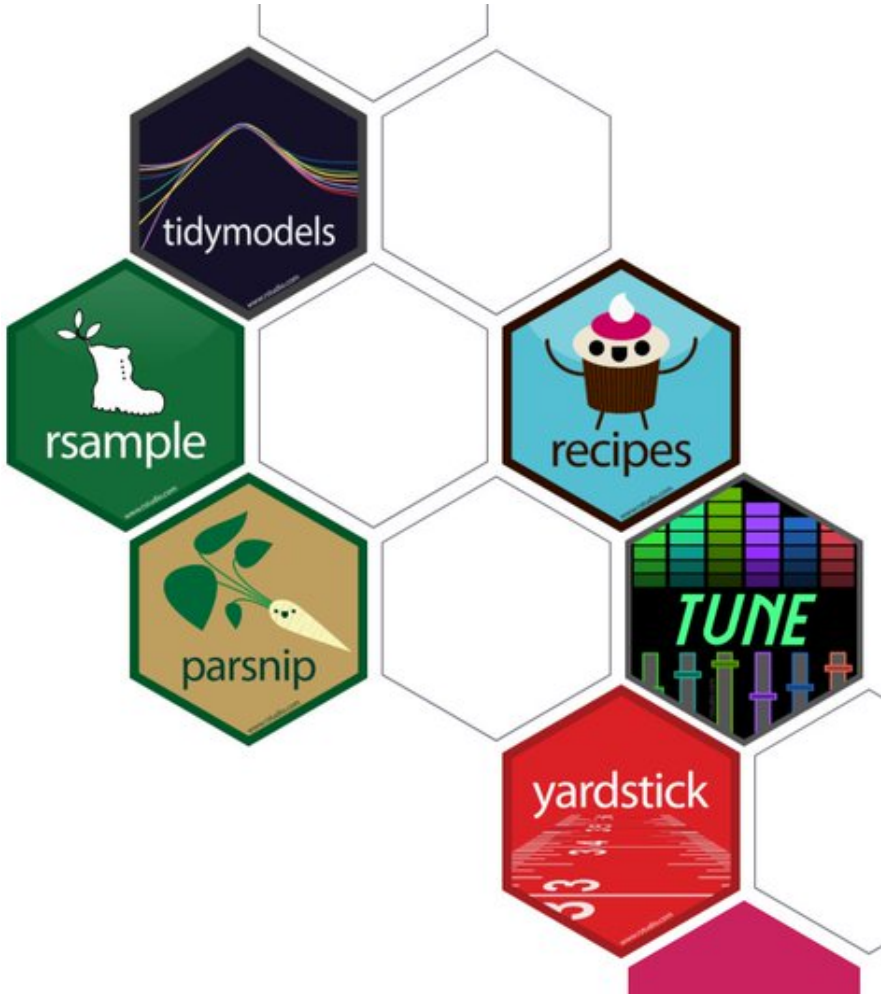
Simulations

Right-censored **survival** data



Method	β WMSE	Correct Zeros	β TPR	β TNR	Num. Non-Zeros	$S(t\|x)$ error
coxph	9.11 (1.54)	0.2 (0)	1 (0)	0 (0)	20 (0)	15.88 (0.35)
GLASP	3.27 (0.61)	0.37 (0.04)	0.95 (0.03)	0.23 (0.06)	16.03 (1.01)	13.86 (0.43)

Implementation details



R package glasp integrates with tidyverse packages

RcppArmadillo backend

Automatic hyper-parameter selection

- Grid Search
- Random Search
- Bayesian Optimization

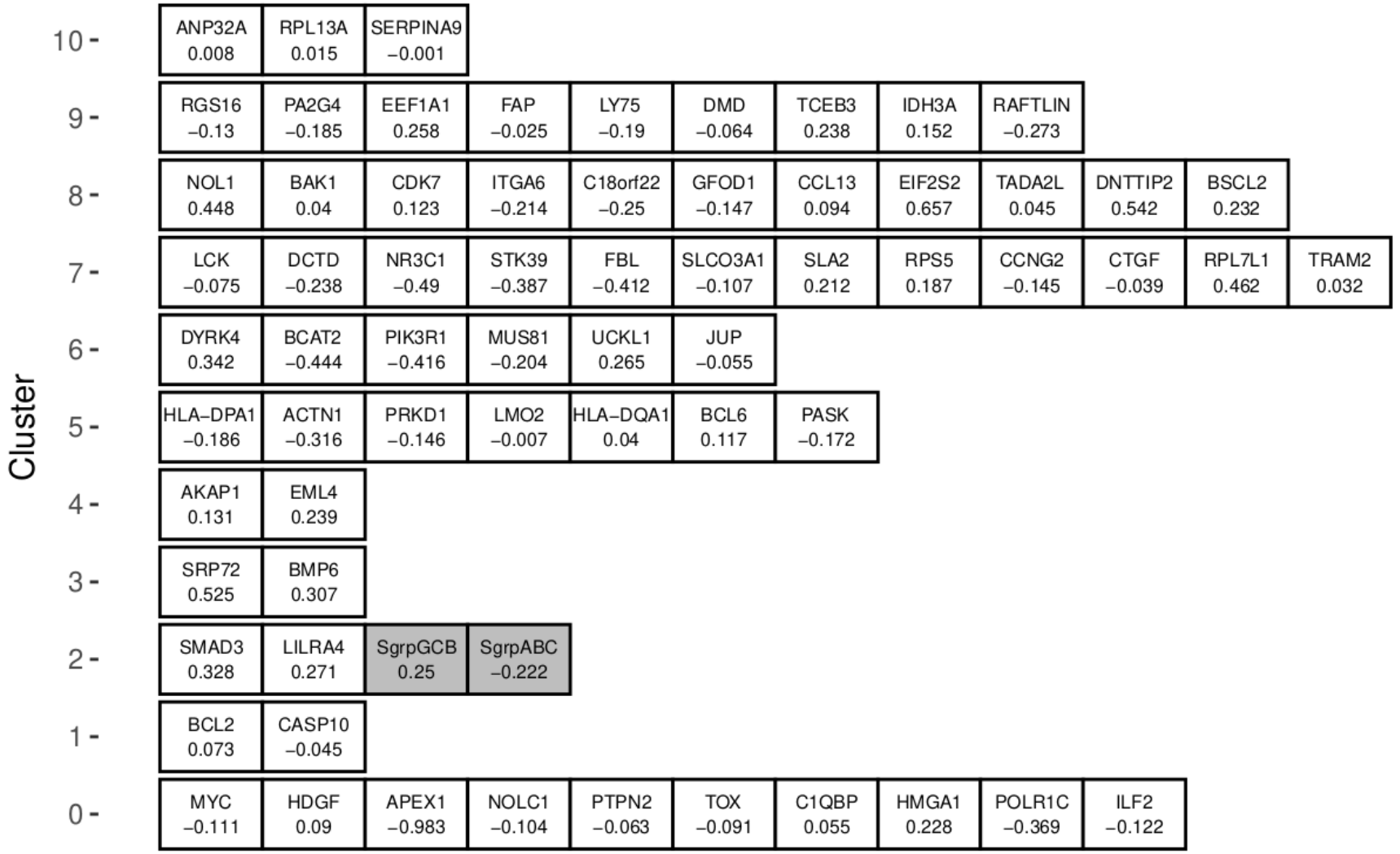
Flexible code

- Linear regression
- Logistic classification
- Cox regression

Docker image jlaria/glasp:0.0.1

Application to right-censored survival data

Gene-expression profiles (diffuse large-B-cell lymphoma)

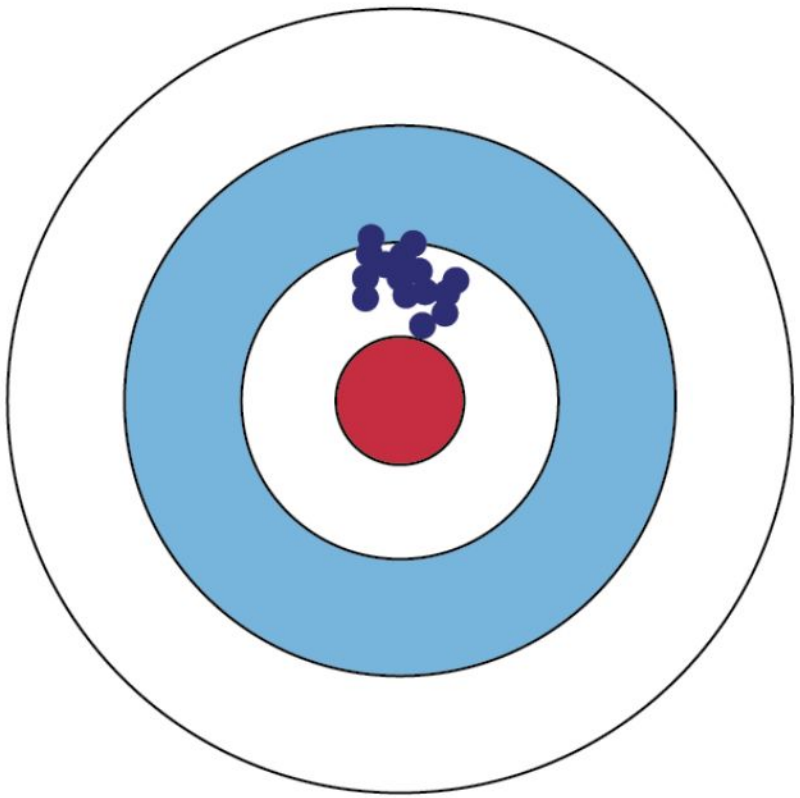



BUT

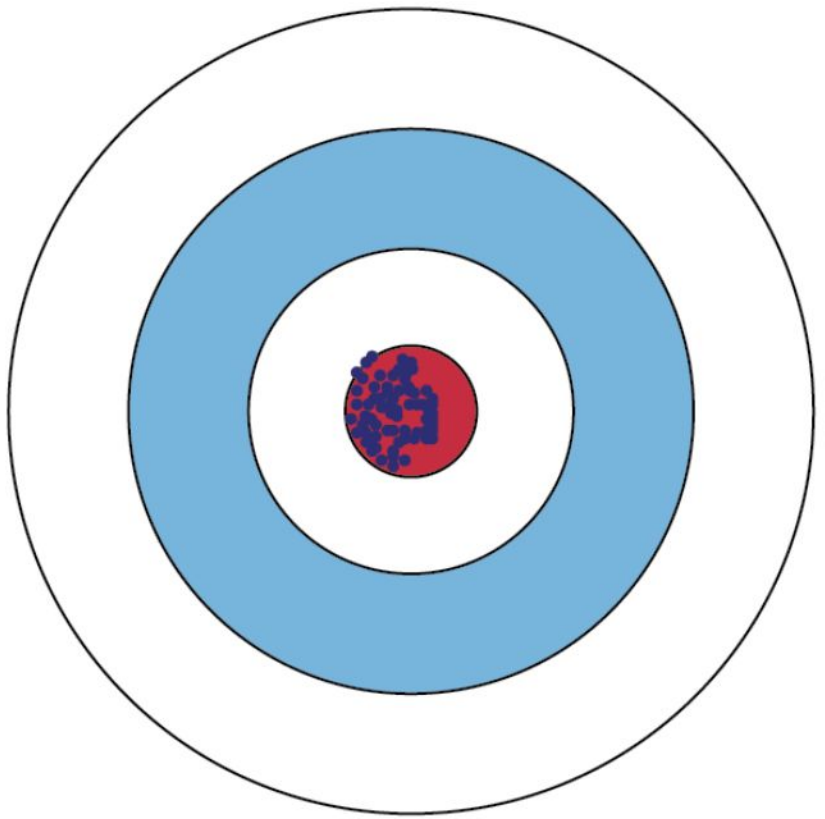
Penalized regression

- $$\min_{\vec{\beta}} \left\{ R(\vec{\beta}) + \alpha\lambda \left\| \vec{\beta} \right\|_1 + (1-\alpha)\lambda \sum_k^K \sqrt{p_k} \left\| \vec{\beta^k} \right\|_2 \right\}$$

Low variance, high bias



Low variance, low bias



Oracle estimators

Given a set of variables

$$\{\beta_1, \ldots, \beta_p\}$$

Define:

$\mathcal{A} = \{j : \beta_j \neq 0\}$ ➞ **Truly important variables**

$\hat{\mathcal{A}} = \{j : \hat{\beta}_j \neq 0\}$ ➞ **Variables selected by the model**

An oracle estimator...

Asymptotically recovers the right subset:

$$\hat{\mathcal{A}} = \mathcal{A}$$

Is normally distributed:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}) \rightarrow_d N(0, \Sigma)$$


The adaptive idea

The adaptive idea

Adaptive sparse group lasso

- $$\min_{\vec{\beta}} \left\{ R(\vec{\beta}) + \alpha\lambda \sum_{j=1}^{P} \tilde{w}_j |\beta_j| + (1-\alpha)\lambda \sum_{k=1}^{K} \sqrt{p_k} \tilde{v}_k \left\| \vec{\beta}^k \right\|_2 \right\}$$

w_j and v_k and are **pre-specified weight**

Intuition: if a variable is important, it should have a small weight

Zou H (2006)
The Adaptive Lasso and Its Oracle Properties
Journal of the American Statistical Association (2006) 101(476) 1418-1429

Mendez-Civieta A, Aguilera-Morillo M, Lillo R (2021)

Adaptive sparse group LASSO in quantile regression

Advances in Data Analysis and Classification15(3) 547-573

Weight calculation

1. Obtain a first estimation of the coefficients

2. Calculate the weights as:

$$\tilde{w} = \frac{1}{|\hat{\beta}|}$$

3. Plug in the weights and solve the adaptive model

$$\min_{\vec{\beta}} \left\{ R(\vec{\beta}) + \alpha\lambda \sum_{j=1}^{P} \tilde{w}_j |\beta_j| + (1-\alpha)\lambda \sum_{k=1}^{K} \sqrt{p_k}\tilde{v}_k \left\| \vec{\beta}^k \right\|_2 \right\}$$

Weight calculation

Obtain a first estimation of the $\hat{\beta}$ coefficients

- **Principal Component Analysis (PCA)**
- **Partial Least Squares (PLS)**

Mendez-Civieta A, Aguilera-Morillo M, Lillo R (2021)
Adaptive sparse group LASSO in quantile regression
Advances in Data Analysis and Classification15(3) 547-573

- **Sparse PCA**
- **Sparse PLS**
- **Sparse Group Lasso (SGL)**
- **Support Vector Regression (SVR)**
- **Xtreme Gradient Boosting (XGB)**

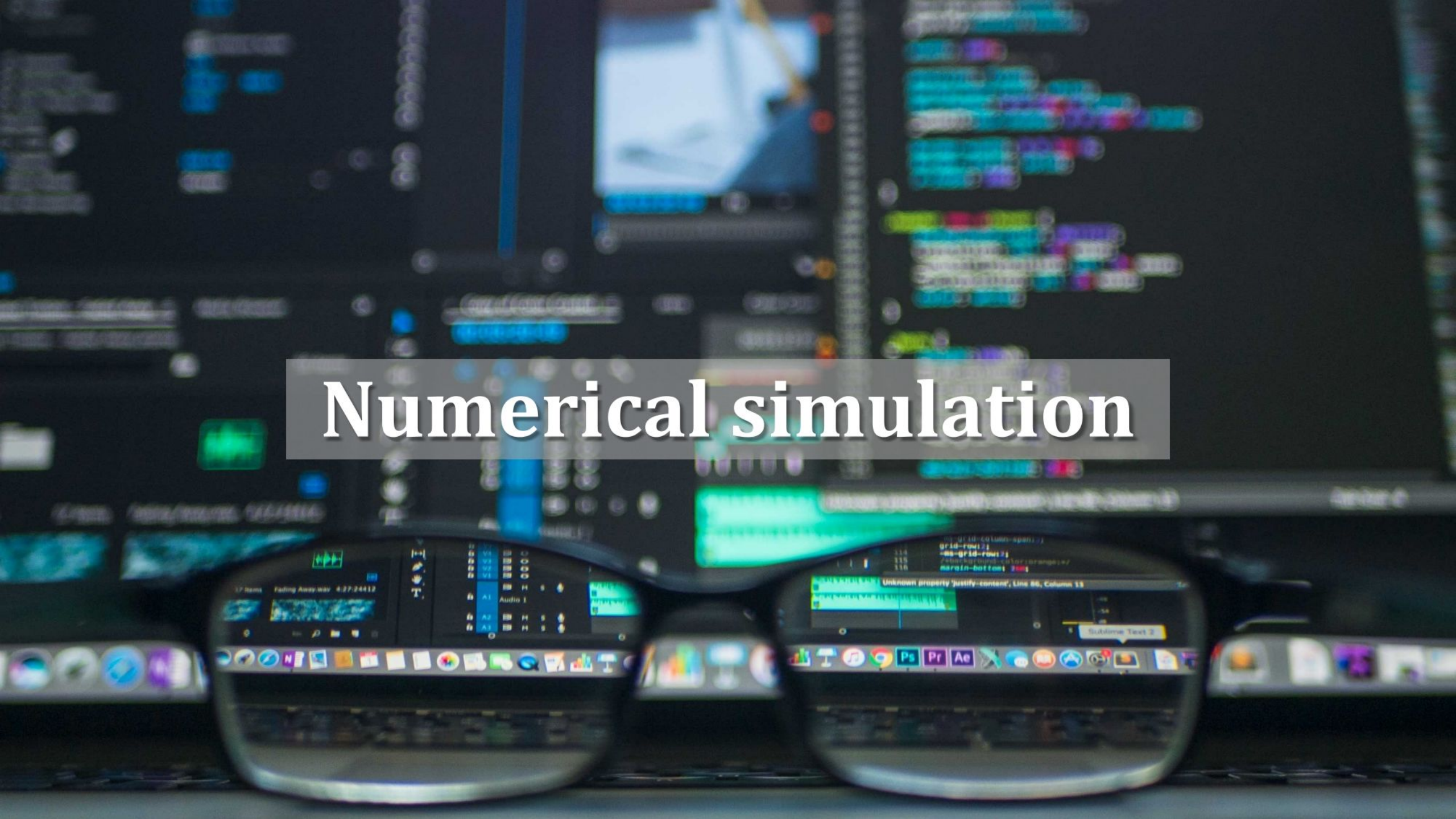
Master tesis
Sánchez Iglesias M. Cristina (2021);
Adaptive alternatives for variable selection in quantile regression


IT'S BACK AND BETTER THAN EVER!
CATS
Kinky Boots
BEST MUSICAL
THE FUNNIEST MUSICAL IN 400 YEARS!
LOL
ROTTEN!
CATS
SANDWICHES
Advertising

The asgl package

- **asgl** is an open-source Python package to solve penalized least squares and quantile regression
- Available at the pypi repository and github
  - https://pypi.org/project/asgl
  - https://github.com/alvaromc317/asgl

	Penalization	R	Matlab	Python	asgl
Linear regression	Lasso	✓	✓	✓	✓
	Group lasso	✓	×	✓	✓
	SGL	✓	×	×	✓
	Adaptive lasso	✓	×	×	✓
	Adaptive group lasso	×	×	×	✓
	ASGL	×	×	×	✓
Quantile regression	Lasso	✓	×	×	✓
	Group lasso	×	×	×	✓
	SGL	×	×	×	✓
	Adaptive lasso	×	×	×	✓
	Adaptive group lasso	×	×	×	✓
	ASGL	×	×	×	✓


Numerical simulation

Numeric simulation

- **Synthetic dataset containing 100 observations and 625 variables. There are 56 informative variables**

$$y = X\beta + \varepsilon, \varepsilon \sim t(3)$$

- **Non adaptive and adaptive models are solved;**

- **Train / validate / test split and repeat the process 100 times.**

Results

	TPR		TNR		CSR		$\|\|\hat{\beta} - \beta\|\|_2$		E_t	
	Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd
LASSO	0.76	0.05	0.90	0.00	0.89	0.01	22.9	3.10	7.60	1.08
GL	1.00	0.00	0.32	0.12	0.38	0.11	24.7	1.16	7.65	0.48
SGL	0.90	0.04	0.75	0.09	0.77	0.08	19.2	1.81	6.11	0.52
PCA_{pct}	0.92	0.04	0.82	0.05	0.82	0.04	15.73	1.84	4.84	0.54
PLS_{pct}	0.91	0.04	0.82	0.03	0.83	0.03	15.61	3.12	4.86	0.95
SPCA	0.92	0.03	0.8	0.06	0.81	0.05	16.17	3.43	5.13	1.14
SPLS	0.91	0.04	0.84	0.04	0.84	0.04	16.23	4.00	4.97	1.22
SGL	0.9	0.05	0.91	0.03	0.91	0.03	11.59	3.05	3.6	0.93
SVR	0.94	0.03	0.9	0.03	0.91	0.03	8.40	2.50	2.67	0.76
XGB	0.9	0.05	0.79	0.06	0.80	0.05	18.58	3.62	5.84	1.12

Non adaptive penalizations

Adaptive penalizations


Genetic data analysis

Research milestones

A genetic problem

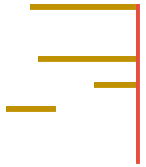
Master tesis, **now PhD**
González Barquero María del Pilar (2023)
Adaptive penalized survival analysis

Genetic data analysis

- Dataset contains expression values from 20531 different genes from 235 patients;

- Objective: Relation between survival time and genetic information;

- Cox models with lasso and adaptive lasso penalizations are solved;



- High level of right censoring (~80%)

Results

First repetition

Model	min(CVE)	λ	Nº variables	$\hat{C}_{130}$
Lasso	10.346	0.126	1	0.499
Ridge	10.214	1.491	44	0.544
PCA	9.792	4.665	17	0.412
Uni	8.596	2.581	44	0.598

Second repetition

Model	min(CVE)	λ	Nº variables	$\hat{C}_{130}$
Lasso	9.809	0.078	26	0.441
Ridge	8.587	1.572	39	0.479
PCA	9.256	2.046	32	0.615
Uni	9.232	4.283	25	0.481

Variable importance

Lasso

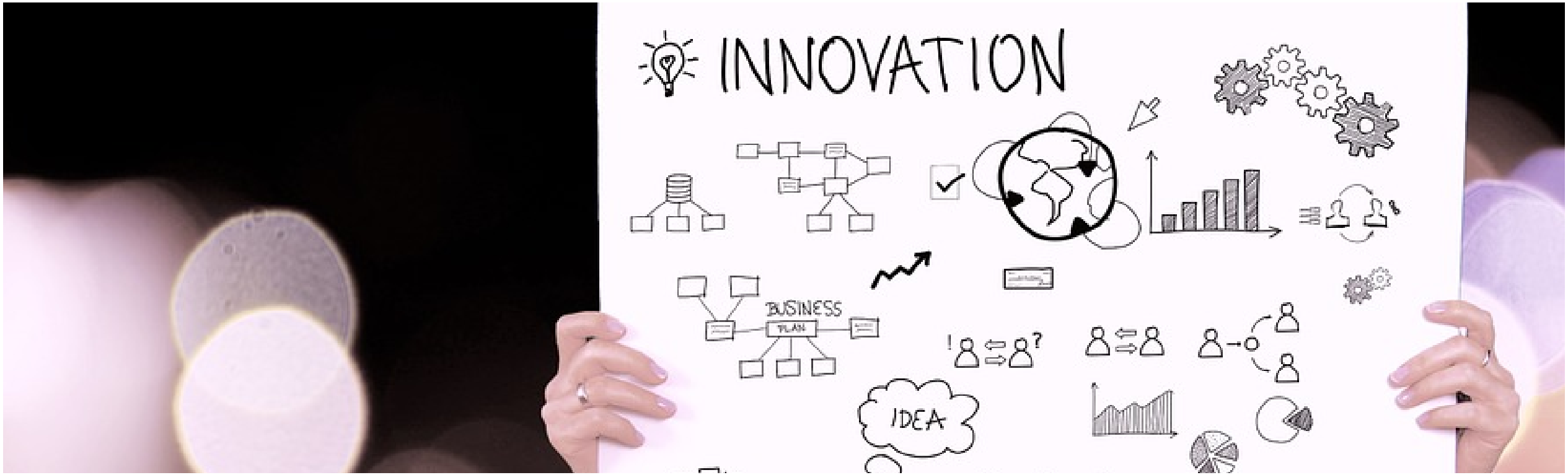
Variables	Importance index
n_cat	1
ENSG00000139880	0.47278
ENSG00000213397	0.2771
ENSG00000167889	0.26186
ENSG00000145780	0.22188
ENSG00000101974	0.20706
ENSG00000183336	0.16669
ENSG00000183155	0.16275
ENSG00000156017	0.15957
ENSG00000100811	0.13426
ENSG00000186350	0.11984

Adaptive

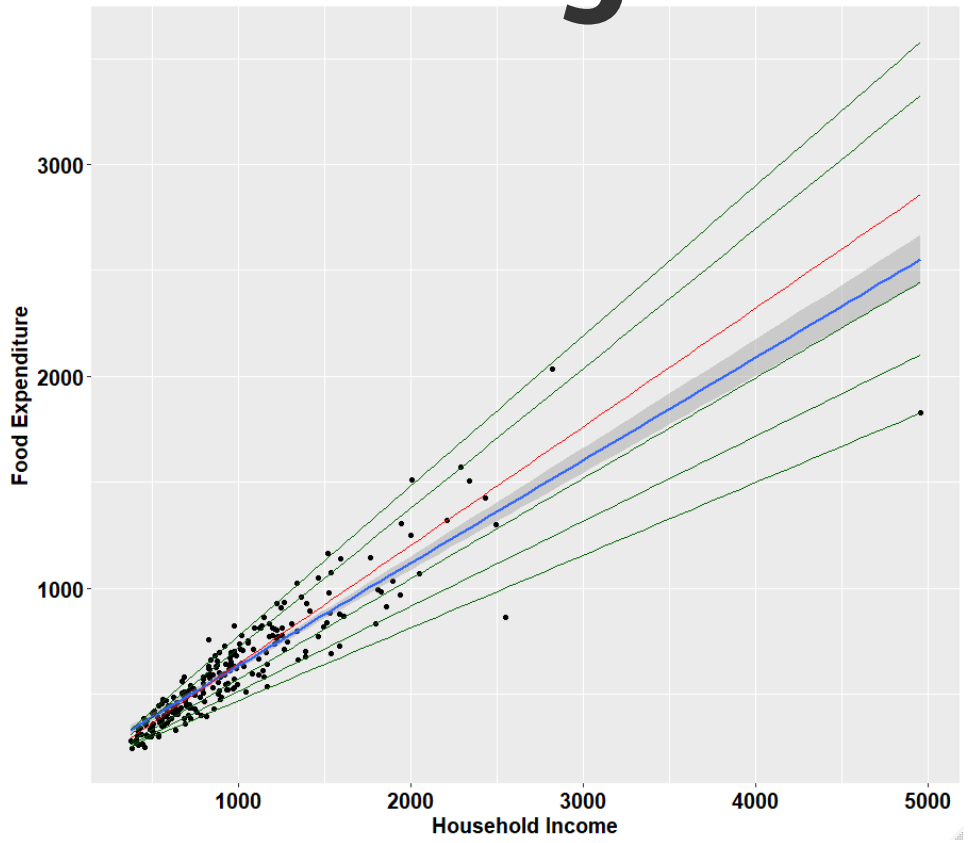
Variables	Importance index
ENSG00000110680	1
multicentric	0.44917
ENSG00000120008	0.34356
ENSG00000188778	0.33855
Instituto_Nacional_Enfermedades_Neoplasicas_Lima	0.27353
Hospital_Universitario_Gran_Canaria	0.25396
n_cat	0.17583
ENSG00000186191	0.16819
ENSG00000129862	0.15875
ENSG00000136352	0.14487
ENSG00000139880	0.10099

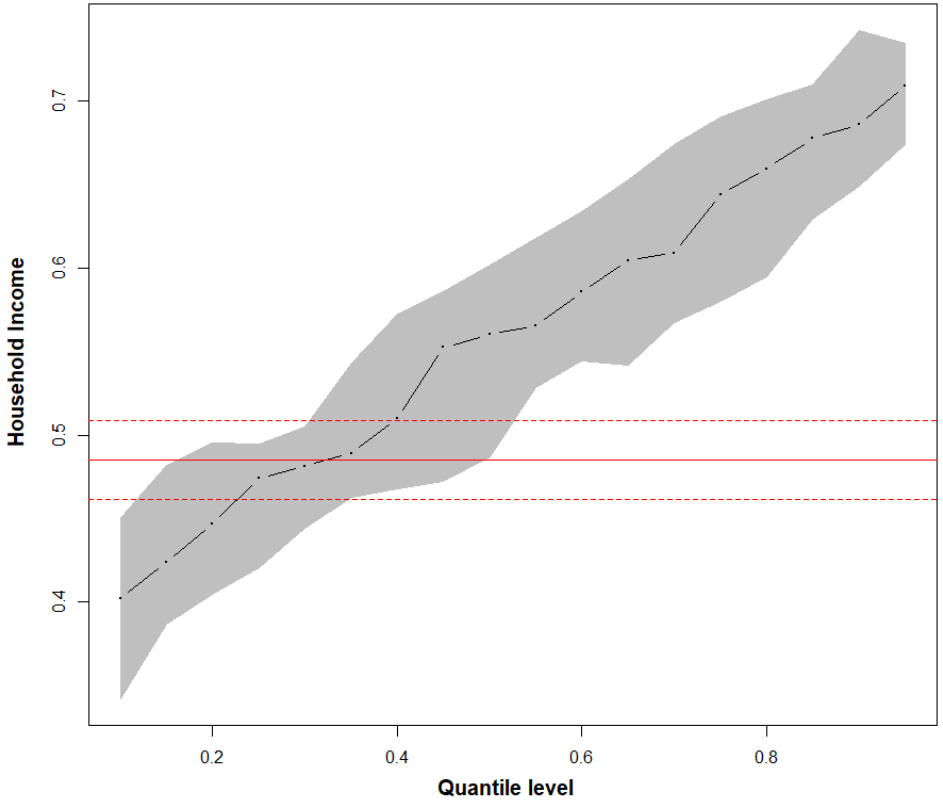
Laria J, et al., (2021)
Iterative variable selection for high-dimensional data: Prediction of pathological response in triple-negative breast cancer
Mathematics 9(3) 1-14

What else?



Quantile regression





Engel Food Expenditure Data used in [Koenker, 2005]. Income vs Food expenditure for Belgian working class households

OLS vs Quantile regression

OLS	QR
Provides conditional mean estimators;	Provides conditional quantile estimators;
Influenced by outliers;	Robust against outliers;
Fails on heteroscedastic data.	Works on heteroscedastic data.

The quantile regression model

- Given a sample of n observations structured as $\mathbb{D} = (y_i, \boldsymbol{x}_i),\ i = 1, \ldots, n$;
- Consider the loss check-function $\rho_\tau(u) = u(\tau - I(u < 0))$, where $I(\cdot)$ is the indicator function;
- The τ-th quantile of the response variable Y is estimated by solving,

$$R(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^t \boldsymbol{\beta}) \right\}$$

.

The sparse group LASSO in QR

First step: extend the sparse group LASSO (SGL) penalization to quantile regression.

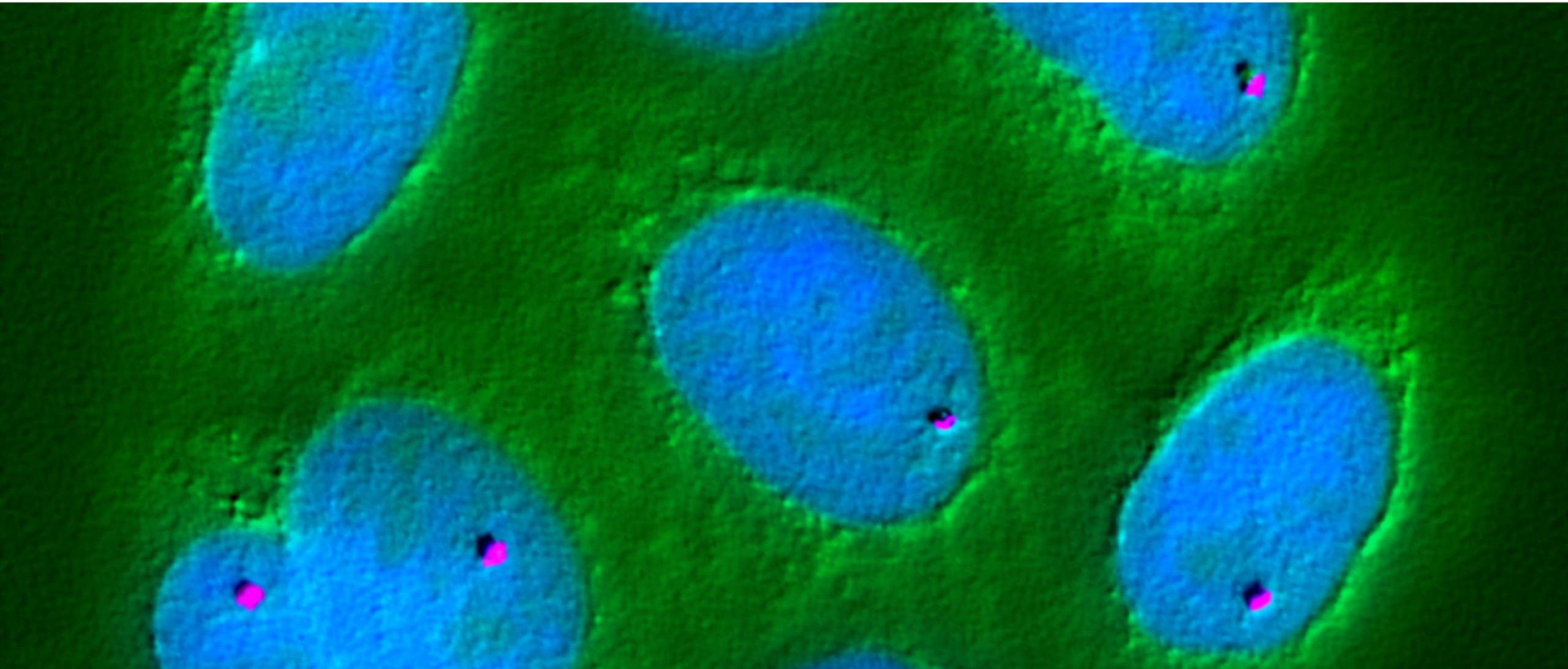
$$\min_{\vec{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - (\beta_0 + \vec{\beta}^t \vec{x}_i)) + \alpha\lambda \left\| \vec{\beta} \right\|_1 + (1-\alpha)\lambda \sum_{l=1}^{m} \sqrt{p_l} \left\| \vec{\beta^{(l)}} \right\|_2 \right\}$$

- Variables divided into m groups of size $\sqrt{p_l}$;
- Sparsity within and between groups of variables.

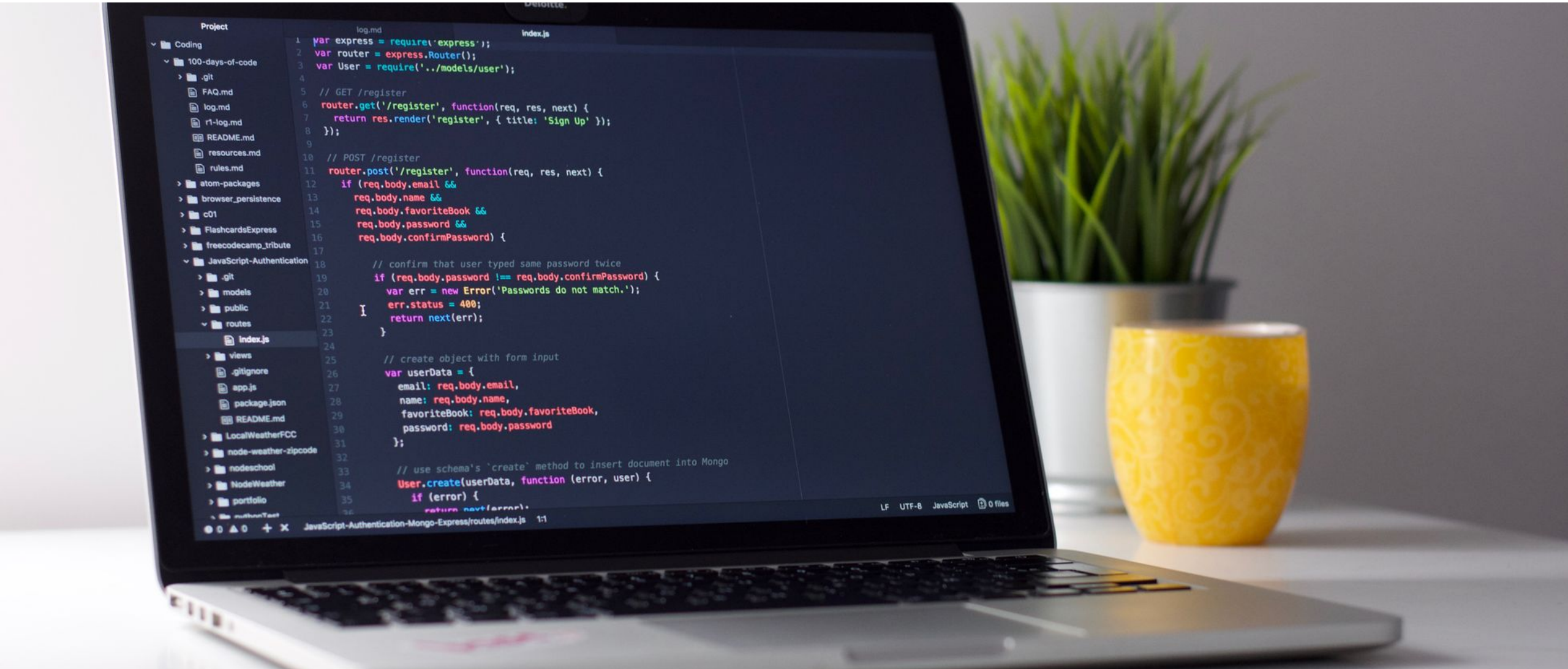

IDEA

Biological interpretation

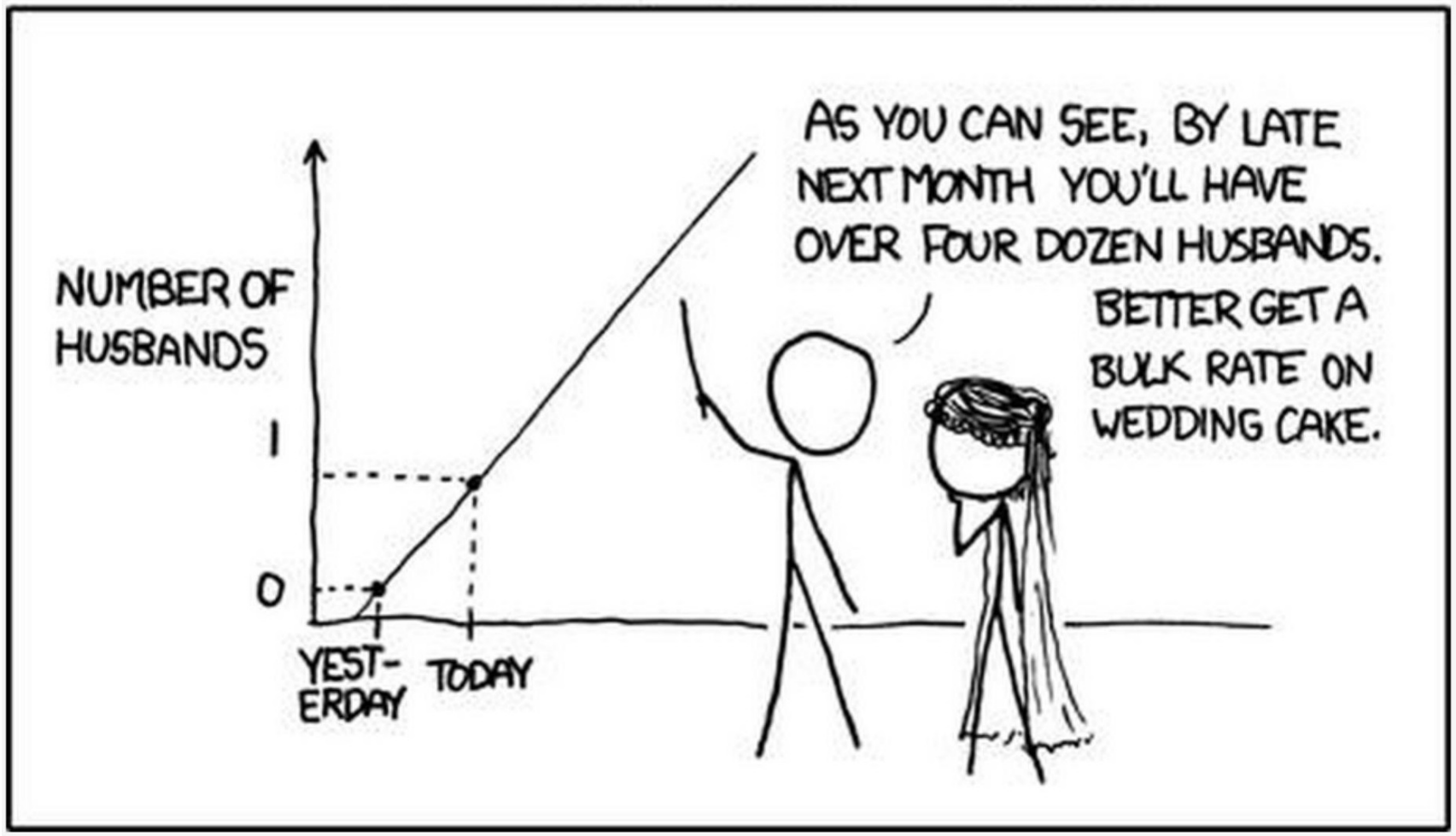
Gene grouping



Computational optimization

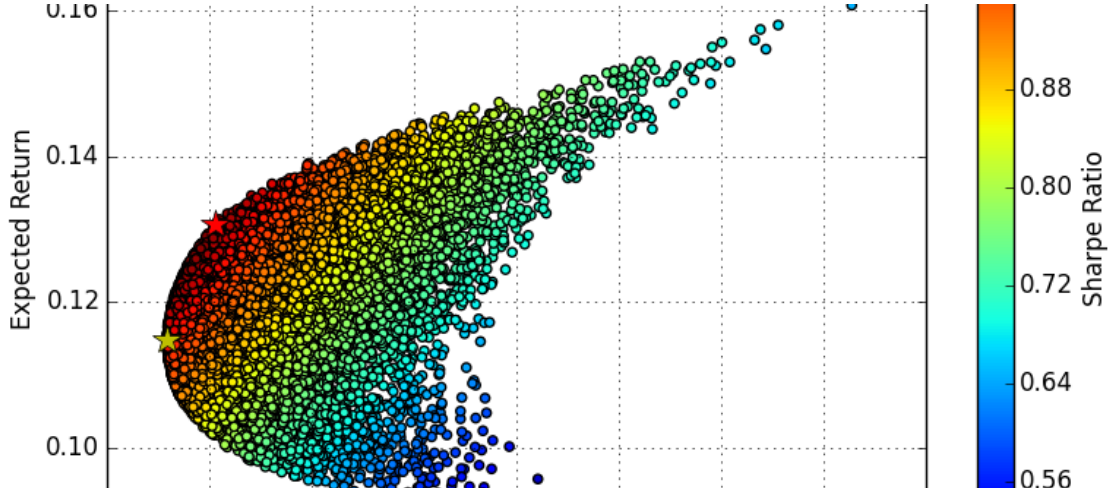


Statistical modelling



Efficient and interpretable solutions

...and more important



Challenge!!

To study the methodologies discussed in this talk when the response variable is multivariate

Important sites!!!

- **Iterative Sparse Group Lasso:** https://github.com/jlaria/sglfast
- **GLASP:** https://github.com/jlaria/glasp
- **Adaptive Sparse Group LASSO (Quantile):** https://pypi.org/project/asgl/ y https://github.com/alvaromc317/asgl


The “No free lunch” theorem

rosaelvira.lillo@uc3m.es