An outline of my thesis topic

M. Remedios Sillero-Denamiel
(Post-doctoral researcher working with Prof. Simon Wilson)

February 9, 2022

Insight

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Computational Methods for the Analysis of Complex Data

AUTHOR:
M. Remedios Sillero Denamiel

SUPERVISED BY:
Prof. Dr. Rafael Blanquero Bravo
Prof. Dr. Emilio Carrizosa Priego
Prof. Dr. Pepa Ramírez Cobo

Institute of Mathematics of the University of Seville (IMUS) and
Department of Statistics and Operations Research of the University of Seville

July 7, 2021

Computational Methods for the Analysis of Complex Data

AUTHOR:
M. Remedios Sillero Denamiel

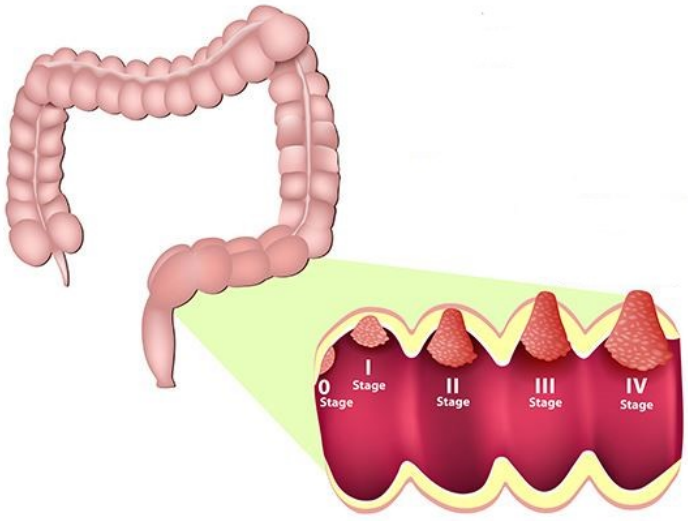
SUPERVISED BY:
Prof. Dr. Rafael Blanquero Bravo
Prof. Dr. Emilio Carrizosa Priego
Prof. Dr. Pepa Ramírez Cobo

Institute of Mathematics of the University of Seville (IMUS) and
Department of Statistics and Operations Research of the University of Seville

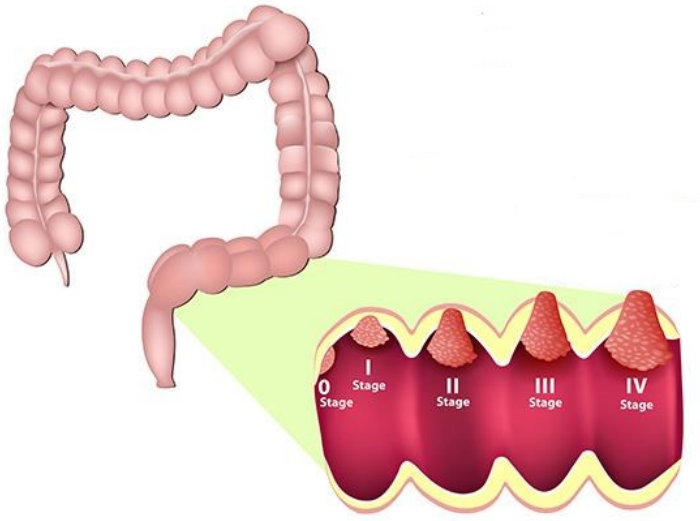
July 7, 2021

The **complexity of the raw data** in addition to **new requests posed by practitioners** entail a challenge.

Background and context: a transfer project

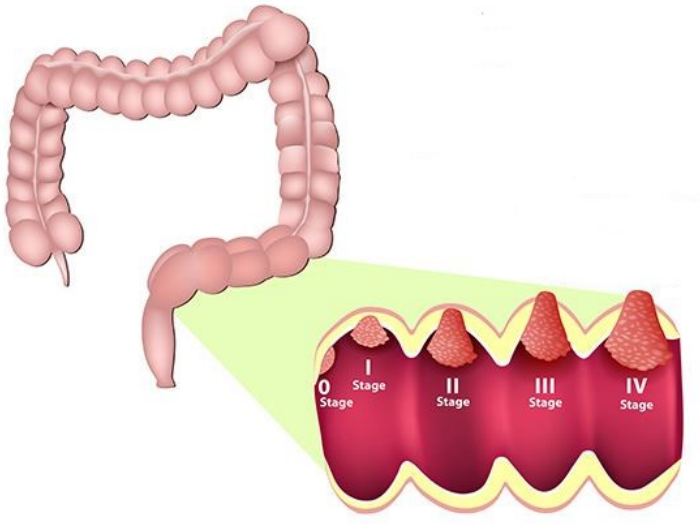


Background and context: a transfer project



- **Heterogeneous data.**
  Blood samples are collected from distinct sources (biobanks in this case), as it is standard in many biomedical contexts.

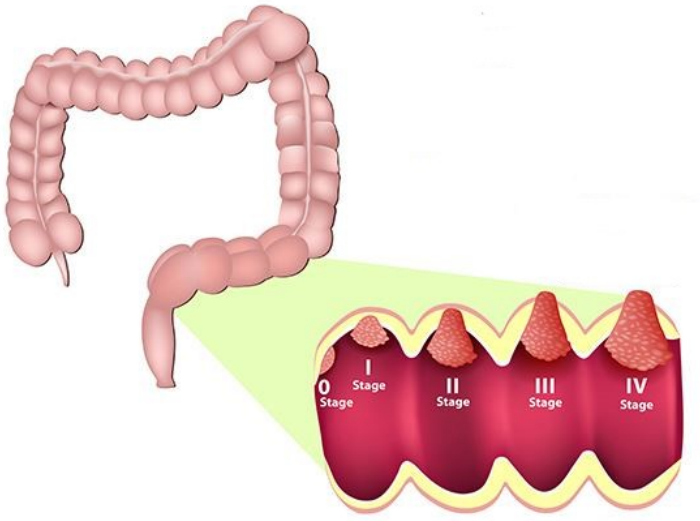
Background and context: a transfer project



- **Heterogeneous data.**

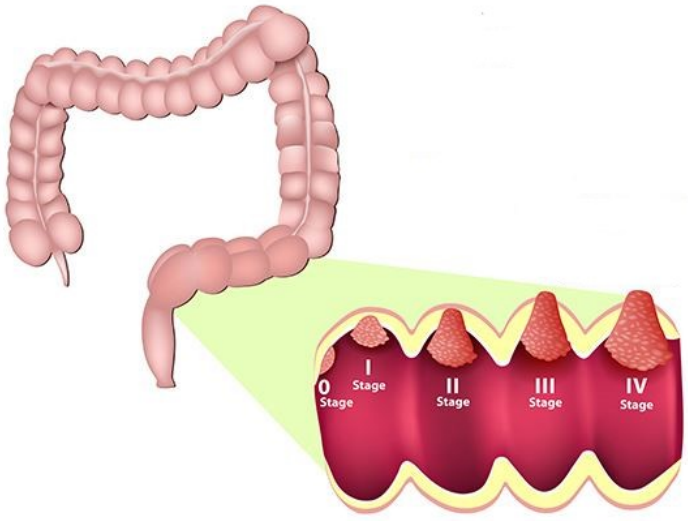
- **Cost-sensitive learning.**
  It has not the same impact to classify a stage IV cancer as a control patient (it can result in death) than a control patient as a stage IV cancer (it only implies further examination).

Background and context: a transfer project



- **Heterogeneous data.**
- **Cost-sensitive learning.**
- **Cost-sensitive feature selection.**
  In this context, millions of features (biomarkers) are obtained from the blood serum. Only a few of them may result relevant.

Background and context: a transfer project



- **Heterogeneous data.**
- **Cost-sensitive learning.**
- **Cost-sensitive feature selection.**

Regular Article | Published: 12 March 2020

A cost-sensitive constrained Lasso

Rafael Blanquero, Emilio Carrizosa, Pepa Ramírez-Cobo & M. Remedios Sillero-Denamiel

Advances in Data Analysis and Classification **15**, 121–158 (2021) | Cite this article

682 Accesses | **5** Citations | **8** Altmetric | Metrics

Abstract

The Lasso has become a benchmark data analysis procedure, and numerous variants have been proposed in the literature. Although the Lasso formulations are stated so that overall prediction error is optimized, no full control over the accuracy prediction on certain individuals of interest is allowed. In this work we propose a novel version of the Lasso in which quadratic performance constraints are added to Lasso-based objective functions, in such a way that threshold values are set to bound the prediction errors in the different groups of interest (not necessarily disjoint). As a result, a constrained sparse regression model is defined by a nonlinear optimization problem. This cost-sensitive constrained Lasso has a direct application in heterogeneous samples where data are collected from distinct sources, as it is standard in many biomedical contexts. Both theoretical properties and empirical studies concerning the new method are explored in this paper. In addition, two illustrations of the method on biomedical and sociological contexts are considered.

Blanquero, R. and Carrizosa, E. and Ramírez-Cobo, P. and Sillero-Denamiel, M.R. A cost-sensitive constrained Lasso. **Advances in Data Analysis and Classification** 15, 121-158 (2021)

Linear regression setting

- Consider p predictors.
- $\mathbf{y}$ the observed response vector, which is predicted by $\hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}_1 + \ldots + \hat{\beta}_p \mathbf{x}_p$
- $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ is estimated by the fitting procedure Ordinary Least Squares (OLS).
- The OLS solution is not sparse. For that reason, the Lasso problem and its different versions appear in the literature:

$$\hat{\boldsymbol{\beta}}^{Lasso}(\lambda) = \underset{\boldsymbol{\beta}}{argmin} \quad \frac{1}{n}\|\mathbf{y} - \mathcal{X}\boldsymbol{\beta}\|^2 + \lambda\|(\beta_1, \ldots, \beta_p)\|_1$$

Linear regression setting

- Consider p predictors.
- $\mathbf{y}$ the observed response vector, which is predicted by $\hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}_1 + \ldots + \hat{\beta}_p \mathbf{x}_p$
- $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ is estimated by the fitting procedure Ordinary Least Squares (OLS).
- The OLS solution is not sparse. For that reason, the Lasso problem and its different versions appear in the literature:

$$\hat{\boldsymbol{\beta}}^{Lasso}(\lambda) = \underset{\boldsymbol{\beta}}{argmin} \quad \frac{1}{n}\|\mathbf{y} - \mathcal{X}\boldsymbol{\beta}\|^2 + \lambda\|(\beta_1, \ldots, \beta_p)\|_1$$

It is known that if $\lambda \to \infty$, then sparsity of the solution increases, and the accuracy becomes less important.

Controlling the performance: The CSCLasso

Main idea

We shall demand that performance measures for groups of interest attain certain threshold values

Lasso with L performance constraints

$$\min_{\boldsymbol{\beta}} \frac{1}{n_0}\|\mathbf{y}_0 - \mathcal{X}_0\boldsymbol{\beta}\|^2 + \lambda\|(\beta_1, \ldots, \beta_p)\|_1$$

s.t.

$$\frac{1}{n_1}\|\mathbf{y}_1 - \mathcal{X}_1\boldsymbol{\beta}\|^2 - f_1 \leq 0$$

$$\vdots$$

$$\frac{1}{n_L}\|\mathbf{y}_L - \mathcal{X}_L\boldsymbol{\beta}\|^2 - f_L \leq 0,$$

where $f = (f_1, \ldots, f_L)$ positive threshold values which we impose.

Controlling the performance: The CSCLasso

Main idea

We shall demand that performance measures for groups of interest attain certain threshold values

Lasso with L performance constraints

$$\min_{\boldsymbol{\beta}} \frac{1}{n_0}\|\mathbf{y}_0 - \mathcal{X}_0\boldsymbol{\beta}\|^2 + \lambda\|(\beta_1, \ldots, \beta_p)\|_1$$

s.t.

$$\frac{1}{n_1}\|\mathbf{y}_1 - \mathcal{X}_1\boldsymbol{\beta}\|^2 - (1+\tau)\mathsf{MSE}_1(\hat{\boldsymbol{\beta}}^{\mathsf{ols}}) \leq 0$$

$$\vdots$$

$$\frac{1}{n_L}\|\mathbf{y}_L - \mathcal{X}_L\boldsymbol{\beta}\|^2 - (1+\tau)\mathsf{MSE}_L(\hat{\boldsymbol{\beta}}^{\mathsf{ols}}) \leq 0,$$

where $\tau \geq \tau_{min}$.

Controlling the performance: The CSCLasso

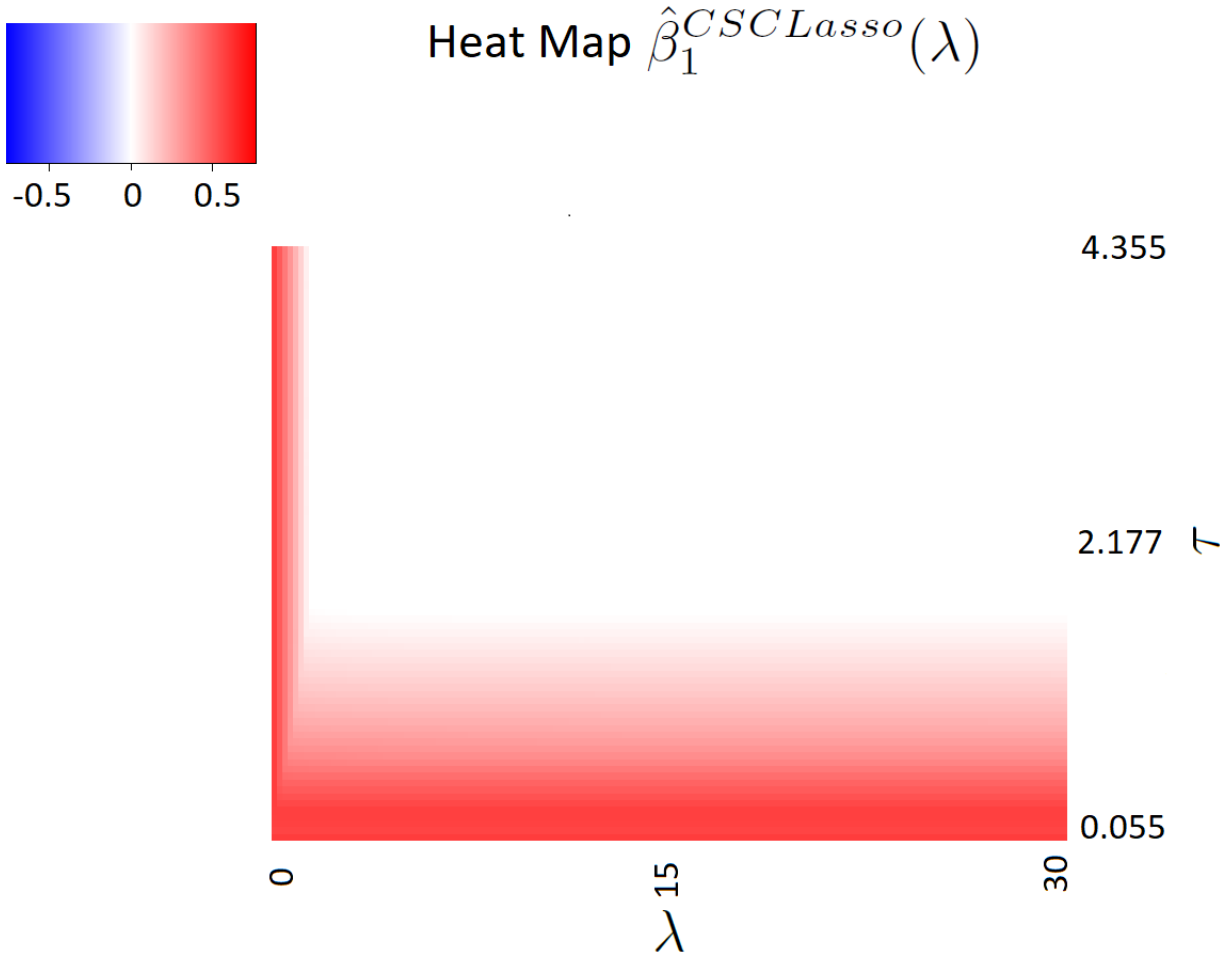


Figure: Heat map of $\hat{\beta}_1^{CSCLasso}(\lambda)$ using `prostate` dataset

Numerical illustration
Communities and Crime dataset[1]

- Response variable: the number of murders per 100K population
- **Group 1: communities from Midwest**
- Group 2: the rest of communities

	f_1	*Overall* MSE	MSE_1	MSE_2	NZ
Lasso	-	0.488	**0.433**	0.453	21.57
Improv. 5%	0.411	0.488	**0.422**	0.453	25.49
Improv. 7%	0.403	0.487	**0.420**	0.453	28.43
Improv. 10%	0.390	0.488	**0.416**	0.453	26.47
Improv. 15%	0.368	0.486	**0.403**	0.459	34.31

Table: Median errors over testing set for communities and crime dataset. Constraints imposed over *Group 1*

[1] Redmond, M. and Baveja, A. A data-driven software tool for enabling cooperative information sharing among police departments. European Journal of Operational Research 141(3), 660-678 (2002)

Research Lines

- **Statistics and Optimization**
  1. **Cost-sensitive Regression**

     Blanquero, R., Carrizosa, E., Ramírez-Cobo, P. and Sillero-Denamiel, M. Remedios
     "A cost-sensitive constrained Lasso", Advances in Data Analysis and Classification, 2021.
  2. **Cost-sensitive Classification**

     Blanquero, R., Carrizosa, E., Ramírez-Cobo, P. and Sillero-Denamiel, M. Remedios
     "Constrained Naïve Bayes with application to unbalanced data classification", Central European Journal of Operations Research, 2021.

Research Lines

- **Statistics and Optimization**
  1. **Cost-sensitive Regression**

     Blanquero, R., Carrizosa, E., Ramírez-Cobo, P. and Sillero-Denamiel, M. Remedios
     "A cost-sensitive constrained Lasso", Advances in Data Analysis and Classification, 2021.
  2. **Cost-sensitive Classification**

     Blanquero, R., Carrizosa, E., Ramírez-Cobo, P. and Sillero-Denamiel, M. Remedios
     "Constrained Naïve Bayes with application to unbalanced data classification", Central European Journal of Operations Research, 2021.
  3. **Cost-sensitive Feature Selection**

     Blanquero, R., Carrizosa, E. Ramírez-Cobo, P. and Sillero-Denamiel, M. Remedios
     "Variable selection for Naïve Bayes classification", Computers & Operations Research , 2021.

Thank you!

sillerom@tcd.ie

An outline of my thesis topic

M. Remedios Sillero-Denamiel
(Post-doctoral researcher working with Prof. Simon Wilson)

February 9, 2022

Insight

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin