Supplementary Material for: Change-point detection in regression models for ordered data via the max-EM algorithm

Modibo Diabaté[1], Grégory Nuel[2] and Olivier Bouaziz[1]

[1]Université Paris Cité, CNRS, MAP5, F-75006 Paris, France
[2]LPSM (UMR CNRS 8001), Sorbonne Université, France

1 Comparison of the initialization methods in the regression models

In this section we provide some additional simulations for the regression settings of Section 5.2 of the main document. The same samples generated from the three regression models of Section 5.2 (linear, logistic and Weibull survival) are used but this time the algorithm is initialized using the Fused Lasso (FL) method instead of the Binary Segmentation (BS) method (see Section 3.4.2 of the main document). Table 1 contains the results in the one breakpoint situation. For ease of comparison, the results for BS initialization is also reported in the table. In Table 2, comparisons between BS and FL intializations for the two breakpoint situation with the Weibull survival model are presented. Detailed comments on those results can be found in Section 5.2 of the main document.

$n = 1,000$		Linear Model bp = 553	Logistic Model bp = 112	Survival Model bp = 666
One bp (BS)	MSE($\hat{\boldsymbol{\theta}}$)	0.86471	1.41481	0.10759
	BIAS^2($\hat{\boldsymbol{\theta}}$)	0.00280	0.01496	0.00157
	VAR($\hat{\boldsymbol{\theta}}$)	0.86191	1.39566	0.10602
	MAPE($\hat{\boldsymbol{\theta}}$)	4.37378	2.74543	0.26998
	ACCE(bp)	0.01367	0.01011	0.00160
One bp (FL)	MSE($\hat{\boldsymbol{\theta}}$)	0.85800	1.16945	0.10388
	BIAS^2($\hat{\boldsymbol{\theta}}$)	0.00256	0.02084	0.00075
	VAR($\hat{\boldsymbol{\theta}}$)	0.85543	1.14861	0.10312
	MAPE($\hat{\boldsymbol{\theta}}$)	4.36168	2.79238	0.26682
	ACCE(bp)	0.01347	0.01022	0.00257

Table 1: Comparison for the Max-EM algorithm between Binary Segmentation (BS) and Fused Lasso (FL) initializations in the one breakpoint regression model. The same three models as in Section 5.2 of the main paper are considered. The first model is a linear homoscedastic regression model with two covariates, the second model is a logistic model with intercept and one covariate and the third model is a Weibull survival regression model with two covariates (see Table 2 of the main paper for the values of the true parameters). The Mean Squared Error (MSE) of the estimated mean parameters, decomposed as the variance (VAR) plus squared bias (BIAS^2), along with the MAPE of the estimated parameters and the ACCE of the estimated breakpoints are provided. The results for the BS initialization are reported again in this table for ease of comparison.

$n = 1,000$		Survival Model bp $= (375, 689)$
Two bp (BS)	MSE($\hat{\boldsymbol{\theta}}$)	0.26253
	BIAS^2($\hat{\boldsymbol{\theta}}$)	0.00220
	VAR($\hat{\boldsymbol{\theta}}$)	0.26033
	MAPE($\hat{\boldsymbol{\theta}}$)	0.52188
	ACCE(bp)	0.01122
Two bp (FL)	MSE($\hat{\boldsymbol{\theta}}$)	0.26587
	BIAS^2($\hat{\boldsymbol{\theta}}$)	0.00182
	VAR($\hat{\boldsymbol{\theta}}$)	0.26405
	MAPE($\hat{\boldsymbol{\theta}}$)	0.52432
	ACCE(bp)	0.01164

Table 2: Comparison for the Max-EM algorithm between Binary Segmentation (BS) and Fused Lasso (FL) initializations in the two breakpoint regression model. The data were generated from a Weibull survival regression model with two covariates (see Table 2 of the main paper for the values of the true parameters). The Mean Squared Error (MSE) of the estimated mean parameters, decomposed as the variance (VAR) plus squared bias (BIAS^2), along with the MAPE of the estimated parameters and the ACCE of the estimated breakpoints are provided. The results for the BS initialization are reported again in this table for ease of comparison.

2 The bike sharing dataset

In this section we provide the estimated values of the intercept parameters and of the dates of the breakpoints in the bike sharing dataset. They supplement the analysis presented in Section 6.1 of the main paper. The max-EM algorithm was implemented with different number of breakpoints where the date was used as a covariate in a homoscedastic linear regression model. Detailed results on the analysis with the estimated values of the slopes and the figures displaying the piecewise linear estimation of the number of rental bikes with respect to the date can be found in the main document.

bp	Intercept values							Dates of change-point					
0	-83989.3243												
1	-113957.1407	562011.5949						2012-10-27					
2	-185647.7541	-215762.3792	562011.5949					2011-10-25	2012-10-27				
3	-243219.4985	90046.2795	-104970.0236	562011.5949				2011-04-22	2012-03-06	2012-10-27			
4	-243219.4985	53561.2934	-161676.0086	-98060.2346	562011.5949			2011-04-22	2011-11-15	2012-03-10	2012-10-27		
5	-212343.2410	-201228.2821	140884.0108	-401538.3899	-96473.5158	562011.5949		2011-04-16	2011-07-17	2011-12-22	2012-03-11	2012-10-27	
6	-212343.2410	-201228.2821	62873.8498	-191883.7781	-401538.3899	-96473.5158	562011.5949	2011-04-16	2011-07-17	2011-11-15	2011-12-22	2012-03-11	2012-10-27

Table 3: Estimated intercept values along with date of breakpoints, obtained from the max-EM algorithm in the different models ranging from 1 to 6 breakpoints.

3 The heart disease dataset

In this section, we compute all the pairwise correlations between the five continuous covariates (age, trestbps, chol, thalach, oldpeak) that were used to construct the proximal space for the heart disease dataset (see Section 6.2 of the main document for more details). The values are represented in Table 4. The scatterplots for all the pairwise combinations are also displayed in Figure 1. A discussion on those results can be found in the main document.

	Segment 1					Segment 2				
Variables	age	trestbps	chol	thalach	oldpeak	age	trestbps	chol	thalach	oldpeak
age	1					1				
trestbps	0.2090	1				0.3010	1			
chol	0.0754	0.0248	1			0.0420	0.0460	1		
thalach	-0.2530	-0.0055	0.1870	1		-0.4610	-0.0398	0.2280	1	
oldpeak	0.0236	0.2530	0.0961	-0.1530	1	0.0791	0.1510	-0.1340	-0.1780	1

Table 4: Pairwise Pearson correlations between all covariates used in the construction of the proximal space in each segment. The correlations with the variable oldpeak were calculated only on individuals that had an ST depression (in other words, the 99 individuals with value 0 for oldpeak were excluded in the calculations).

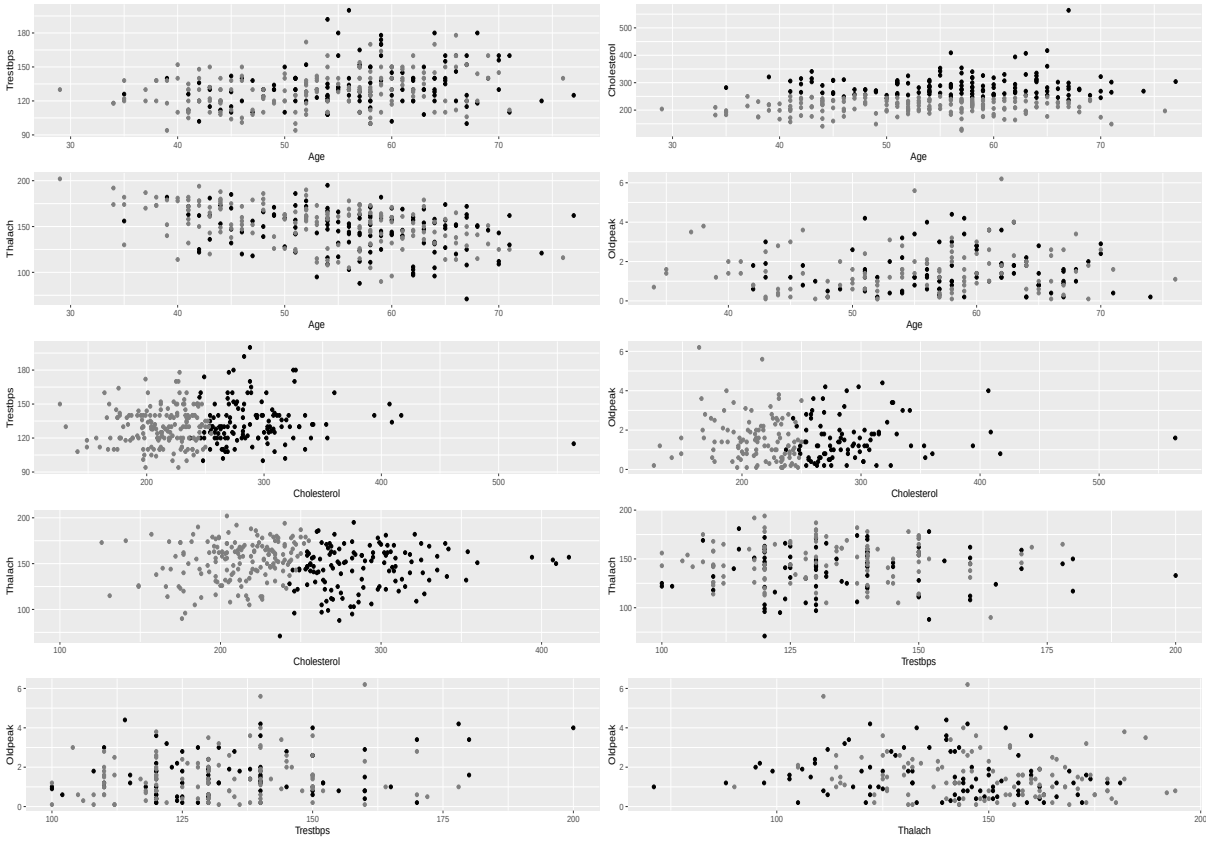

Figure 1: Scatterplots for all pair of covariates used in the construction of the proximal space in each segment. The scatterplots with the variable oldpeak were calculated only on individuals that had an ST depression (in other words, the 99 individuals with value 0 for oldpeak were excluded in those plots).