Non-Standard Work, Unionization and Innovation Capacity

Anna Snaidero, Caterina Manicardi, Stefania Scrofani
May 2022

Interim presentation
Statistical Learning and Large Data

PhD Economics, Sant'Anna School of Advanced Studies

1. Motivation and background
2. Data cleaning (with imputation of the values for total turnover through Loess and cross-validation)

3. Principal Component Analysis on variables related to firm's workforce and subsequent k-means clusterization
4. K-means clusterization with variables related to innovation capacity
5. Mapping between the two clusterization
6. Linear Discriminant Analysis (and Quadratic Discriminant Analysis)

Single firms as observations (2746 obs.)

7. Canonical Correlation Analysis
8. Comparison between LS, Ridge, Lasso and Best Subset Selection as methods for model selection

Ateco codes as observations (1554 obs.)

9. Further improvements

Motivation and background

From the end of the 70s the economics debate focused on labour market policy, with the new orthodoxy that views flexibilisation as key to economic success.

However, due to the productivity slowdown in recent decades, and particularly after the 2008 global crisis, economic theories opposed to the orthodox consensus have regained importance within the debate around the pro-flexibility policy. New empirical and theoretical research works highlight the link between labour institutions and firms' innovativiness.

This contribution proposes new evidence on the link between non-standard work, unionization and innovation capacity. We employ the "*Rilevazione Imprese Lavoro*" carried out by INAPP in 2015 to investigate the links within Italian firms between their propensity to innovate and workforce characteristics.

Data cleaning

Original dataset:
Rilevazione Imprese Lavoro 2015
30091 observations of 219 variables

Data cleaning:
- Transform variables related workforce, training expenditure and investments in percentage unit
- Imputation of firm's turnover through Loess regression and cross-validation

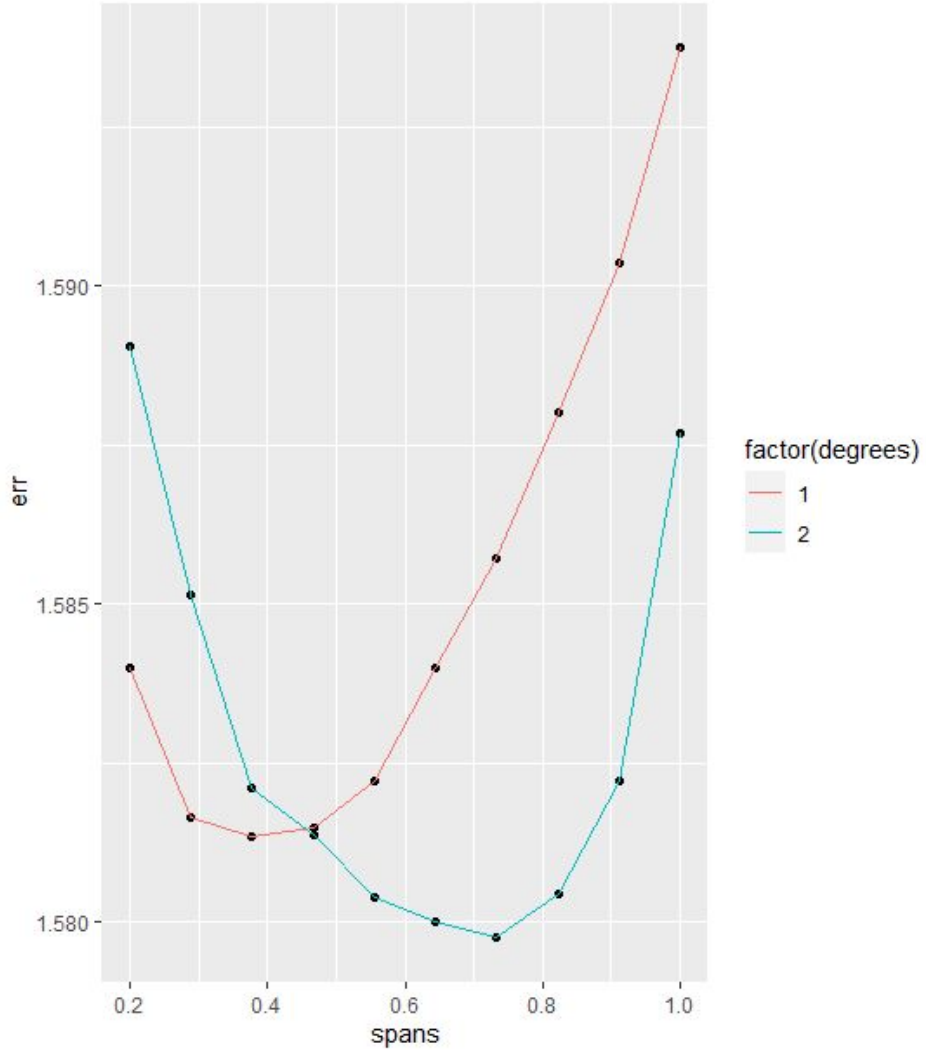
Dataset after cleaning and subset of variables of interest:
2746 observations of 19 variables (5 binary variables for innovation capacity and 14 variables about workforce composition)

Data cleaning

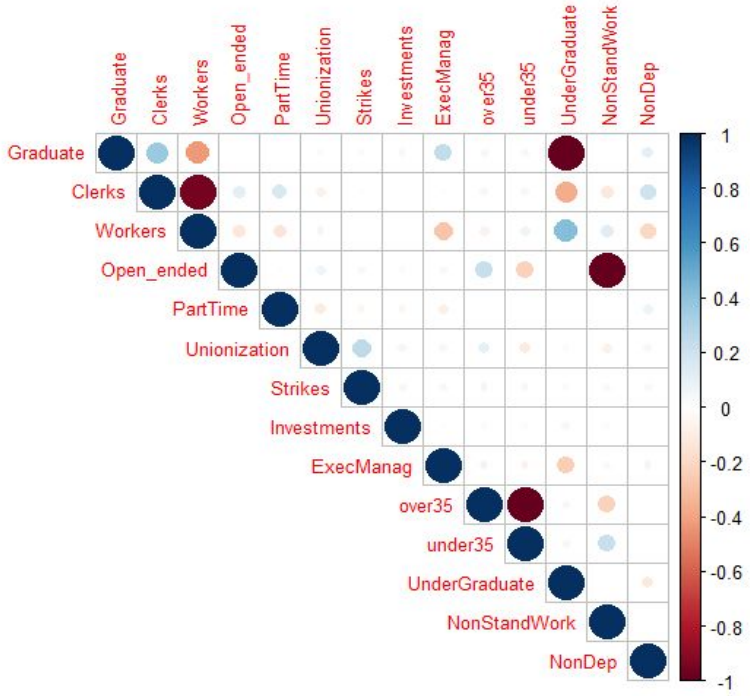
Variable description	Unit	Source
Share of managers and executives over total number of employees	%	RIL
Share of graduated employees over total number of employees	%	RIL
Share of non-standard workers [a] over total number of employees	%	RIL
Share of part-time workers [b] over total number of employees	%	RIL
Share of unionized workers over total number of employees	%	RIL
Share of hours not worked in 2014 because of strikes and labour conflicts in general over the average number of hours worked in Italy per worker [c]	%	RIL, OECD
Amount of training expenditures over total turnover	%	RIL
Amount of investments [d] over total turnover	%	RIL

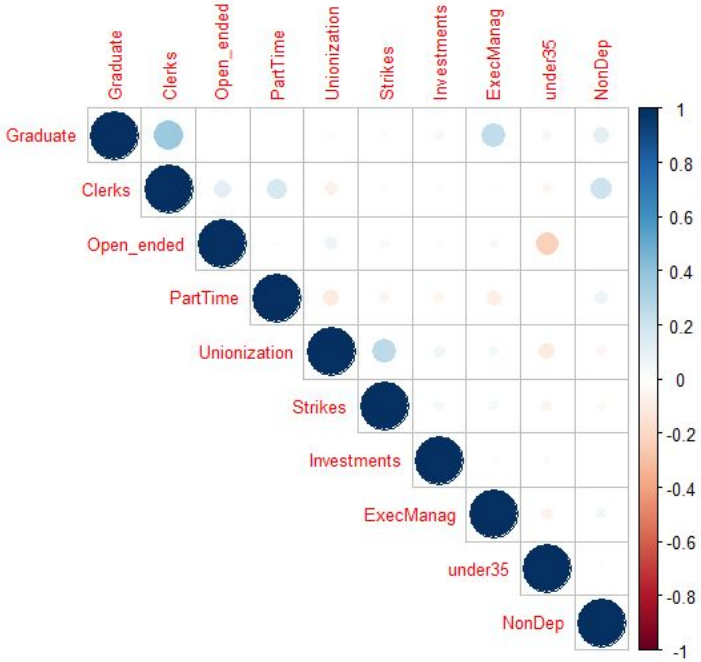
[a] The NSW are defined as those employees with a fixed-term contract, an apprenticeship contract or an on-call contract.
[b] Both open-ended and fixed-term part-time contracts.
[c] The average number of hours worked in Italy per worker in 2014 is fixed to 1716 according to the OECD (https://data.oecd.org/emp/hours-worked.htm).
[d] The amount of investments includes all the investments in marketing and advertisement, research and development, patents, licences, brands and software, land and buildings, industrial plants, machinery and equipment, IT equipment and other.

Turnover imputation: Loess and cross validation



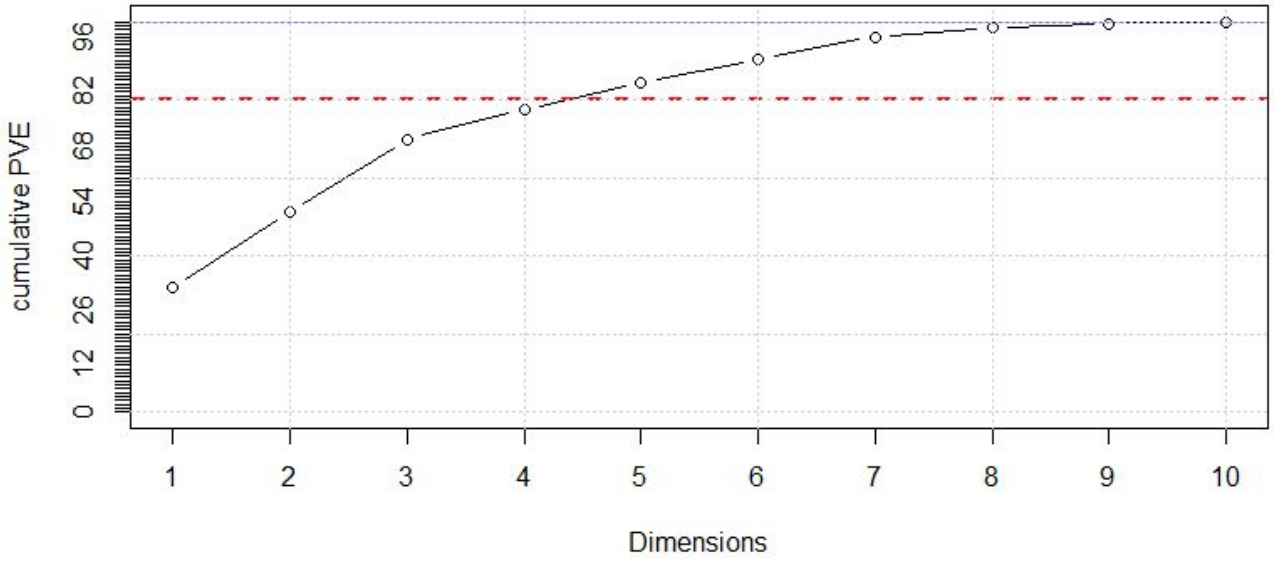
Find correlation (cutoff=0.8) prior to PCA

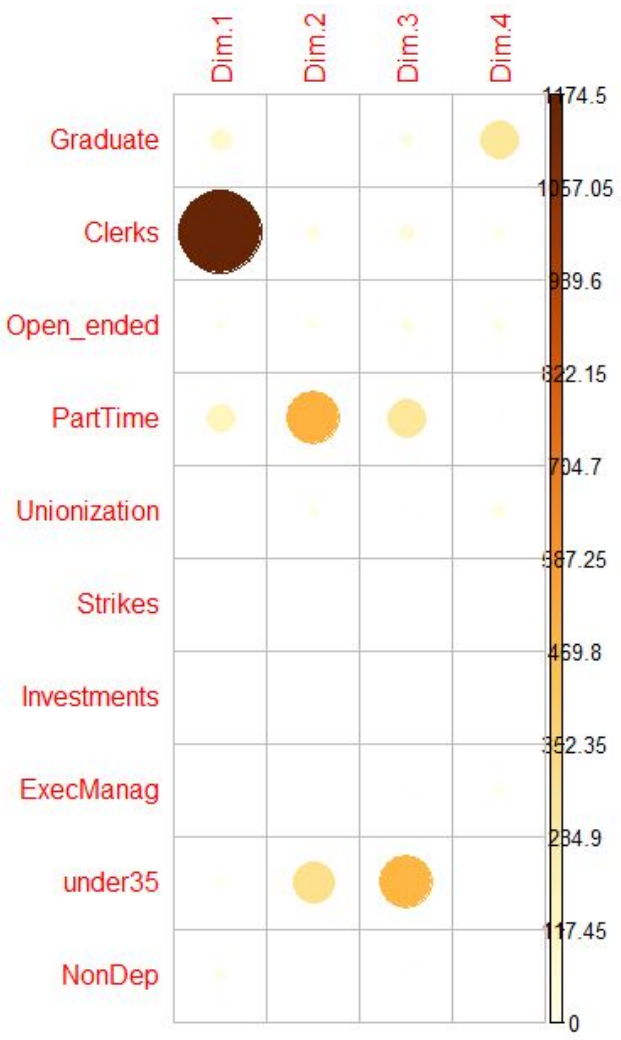




PCA on workforce composition

Loadings
PC1: -0.90 Clerks, -0.31 PartTime, -0.23 Graduates
PC2: -0.75 PartTime, -0.58 under35, -0.19 Clerks
PC3: 0.76 under35, -0.55 PartTime, 0.24 Clerks
PC4: 0.86 Graduates, -0.24 Open_ended, 0.24 Unionization

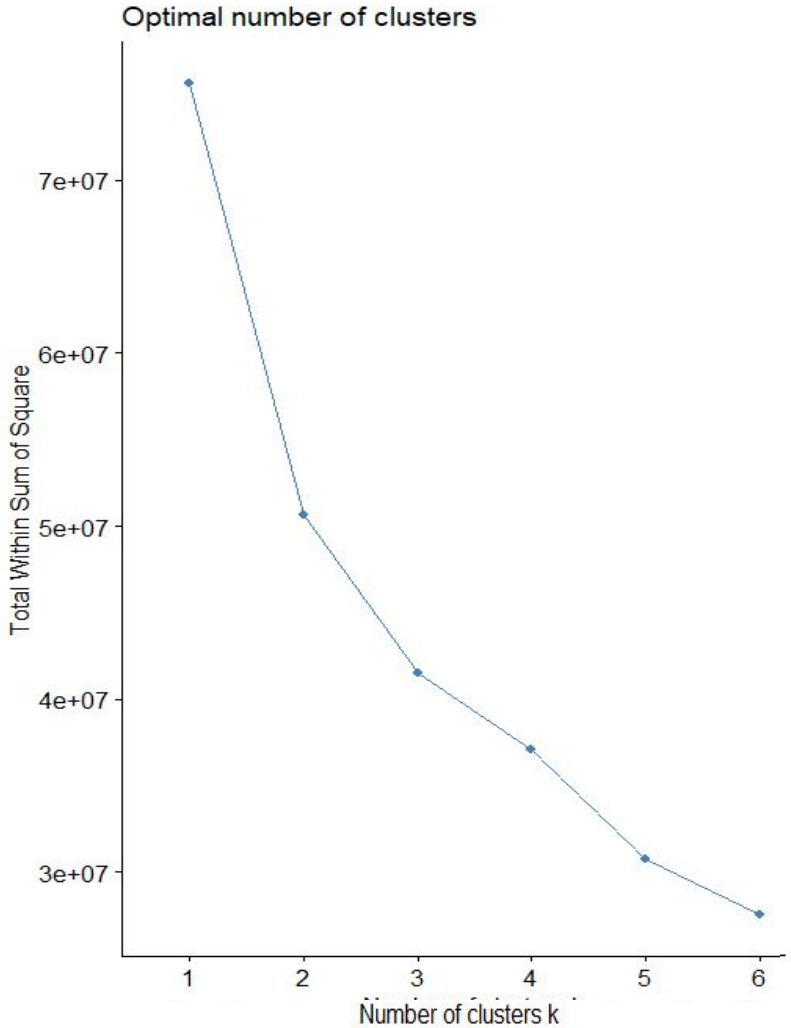




K-means clustering

	Cluster 1	Cluster 2	Cluster 3
Graduate	9.10	5.09	24.54
Clerks	46.95	18.44	86.67
Open_ended	91.98	90.09	92.59
PartTime	87.30	8.09	14.09
NonDep	10.34	6.06	13.33
Unionization	2.25	7.71	4.85
Strikes	0.12	0.79	0.55
Investments	1.44	2.46	2.23
ExecManag	1.64	3.36	5.31
under35	28.84	28.71	29.93

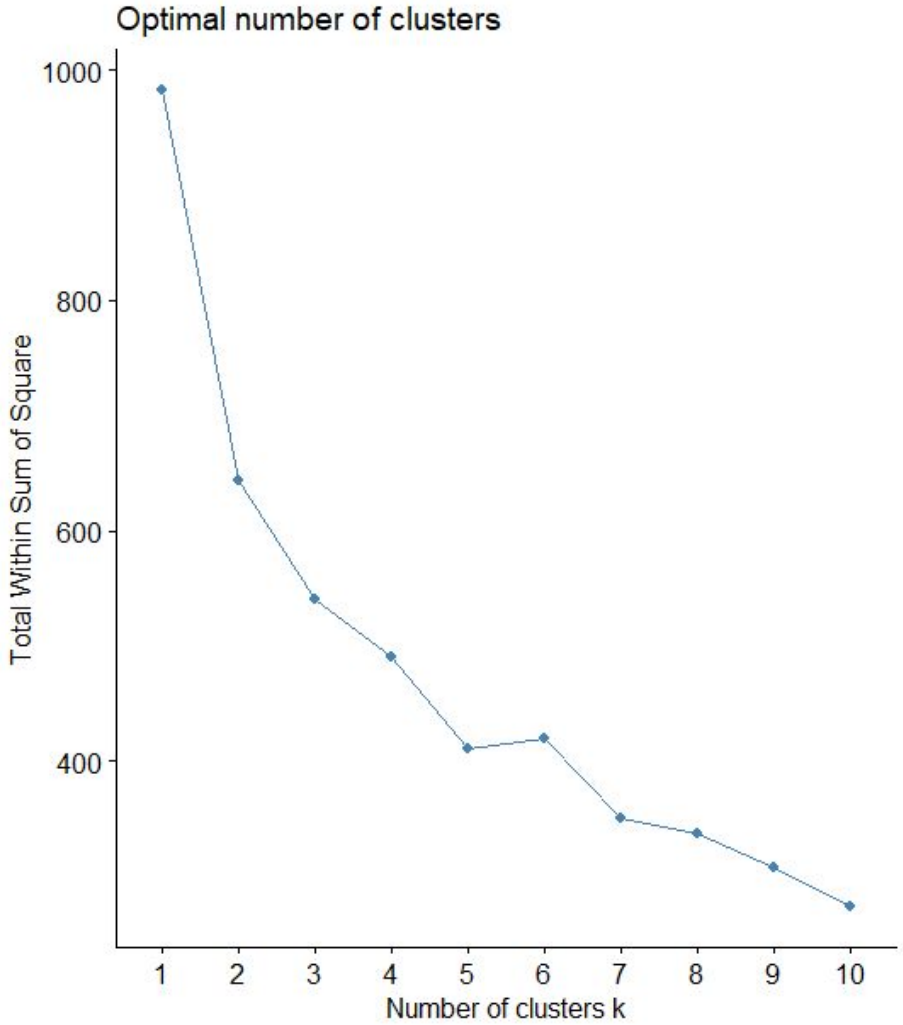
Cluster 1: high *PartTime*, low *ExecManag* and *Investments*.
Cluster 2: low *Clerks*, *PartTime* and *NonDep*, high *Unionization*.
Cluster 3: high *Graduate*, *Clerks*, *ExecManagers*.



Technological clustering

	Cluster 1	Cluster 2	Cluster 3
Product Innovation	0.24	0.33	0.83
Process Innovation	0.16	0.22	0.81
R&D	0.04	0.26	0.31
Plants	0.18	0.63	0.60
Tech	0.09	0.74	0.37

Cluster 1: lagged firms
Cluster 2: "supply dominated" (Pavitt, 1984)
Cluster 3: innovators



Mapping between the two clusterization

	ClusterWorkForce 1	ClusterWorkForce 2	ClusterWorkForce 3
ClusterTech 1	0.57	0.33	0.28
ClusterTech 2	0.24	0.29	0.46
ClusterTech 3	0.19	0.39	0.27

- The first workforce cluster, characterized by a higher level of non dependent workers and part time workers, is the one with the highest percentage of “laggard firms”.
- The second cluster, with the lowest level of part time and non dependent workers and the highest unionization level, has instead a quite homogeneous composition, with a particularly high level of “innovators”.
- The third workforce cluster, with the highest share of graduates and intermediate-upper figures (clerks, executives and managers) is instead highly characterized by supply dominated firms.

Linear (and Quadratic) Discriminant Analysis

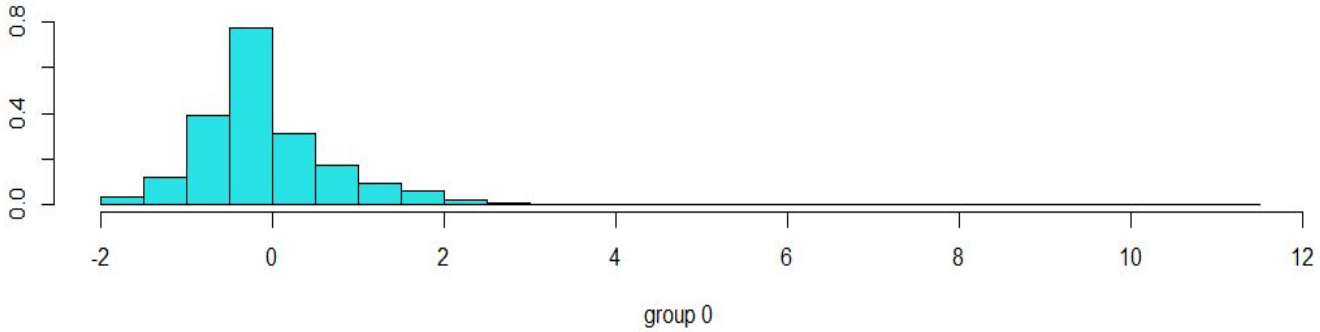
To classify those firms that pursue R&D and those who do not.

Dataset: 21746 observations of 15 variables (workforce+R&D)

1. Find correlation (cutoff=0.8) → From 15 to 11 variables
2. Split the dataset in train and test data (60-40 ratio)
3. Logit regression to perform VIF and check for multicollinearity (*RD*~.)
4. Check whether classes (*RD*=0, *RD*=1) have identical covariance matrices
5. Linear Discriminant Analysis (find a discriminant function linear in the predictors)
6. Quadratic Discriminant Analysis (find a discriminant function quadratic in the predictors)

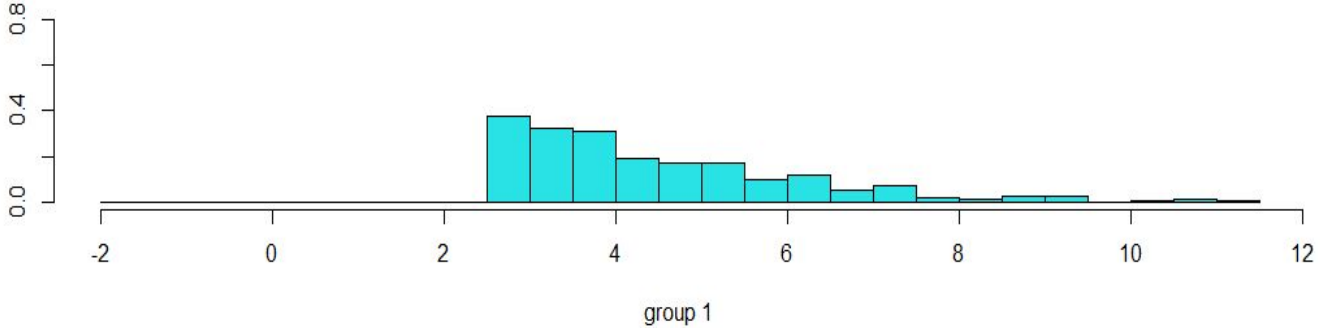
Group means:

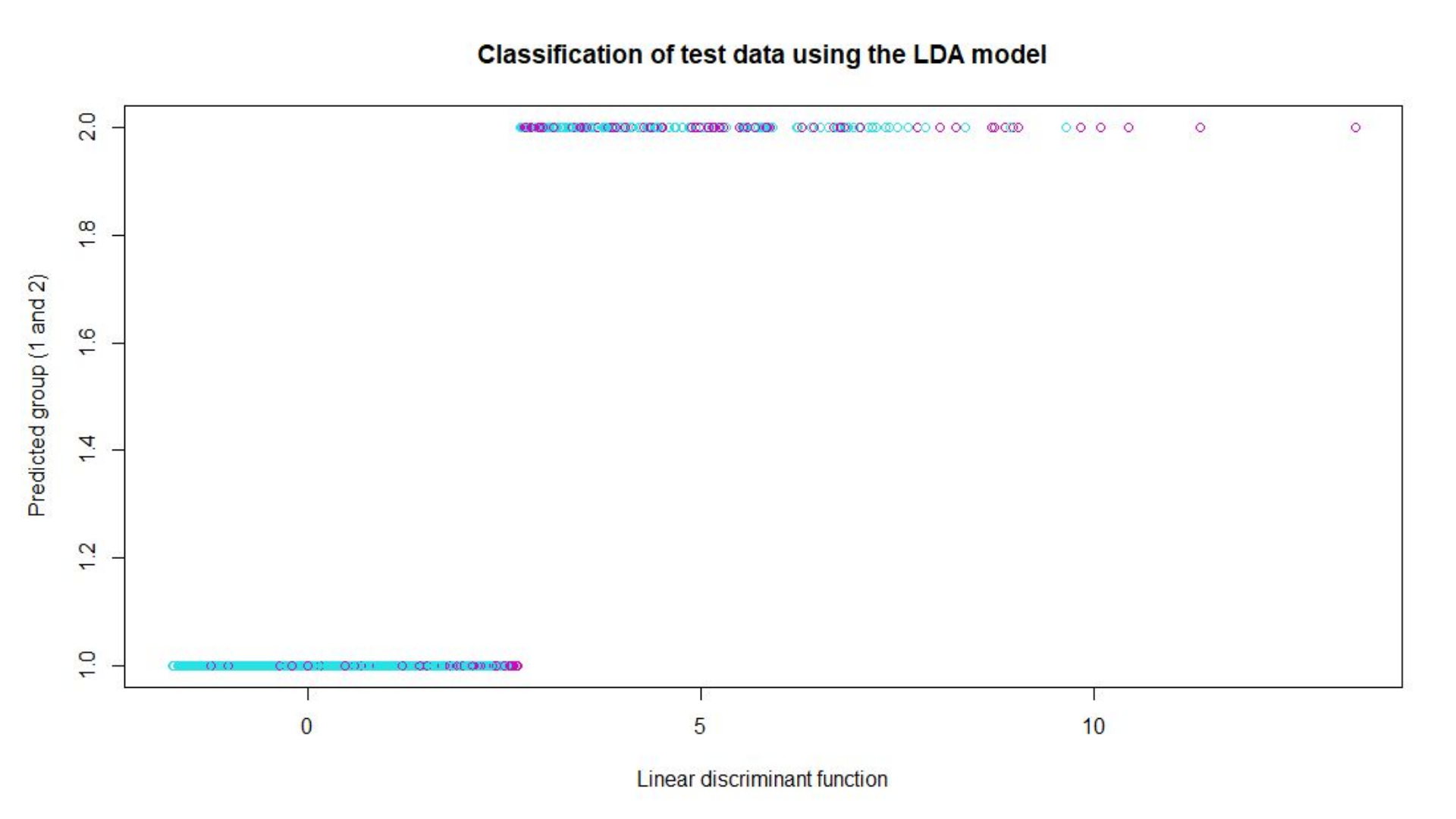
	0	1
Graduate	9.43	19.21
Clerks	38.16	44.74
PartTime	20.85	10.15
NonDep	8.09	9.68
Unionization	5.57	13.29
Strikes	0.37	3.13
Investments	1.93	6.04
ExecManag	3.44	6.35
under35	29.32	26.58
NonStandWork	8.93	8.46

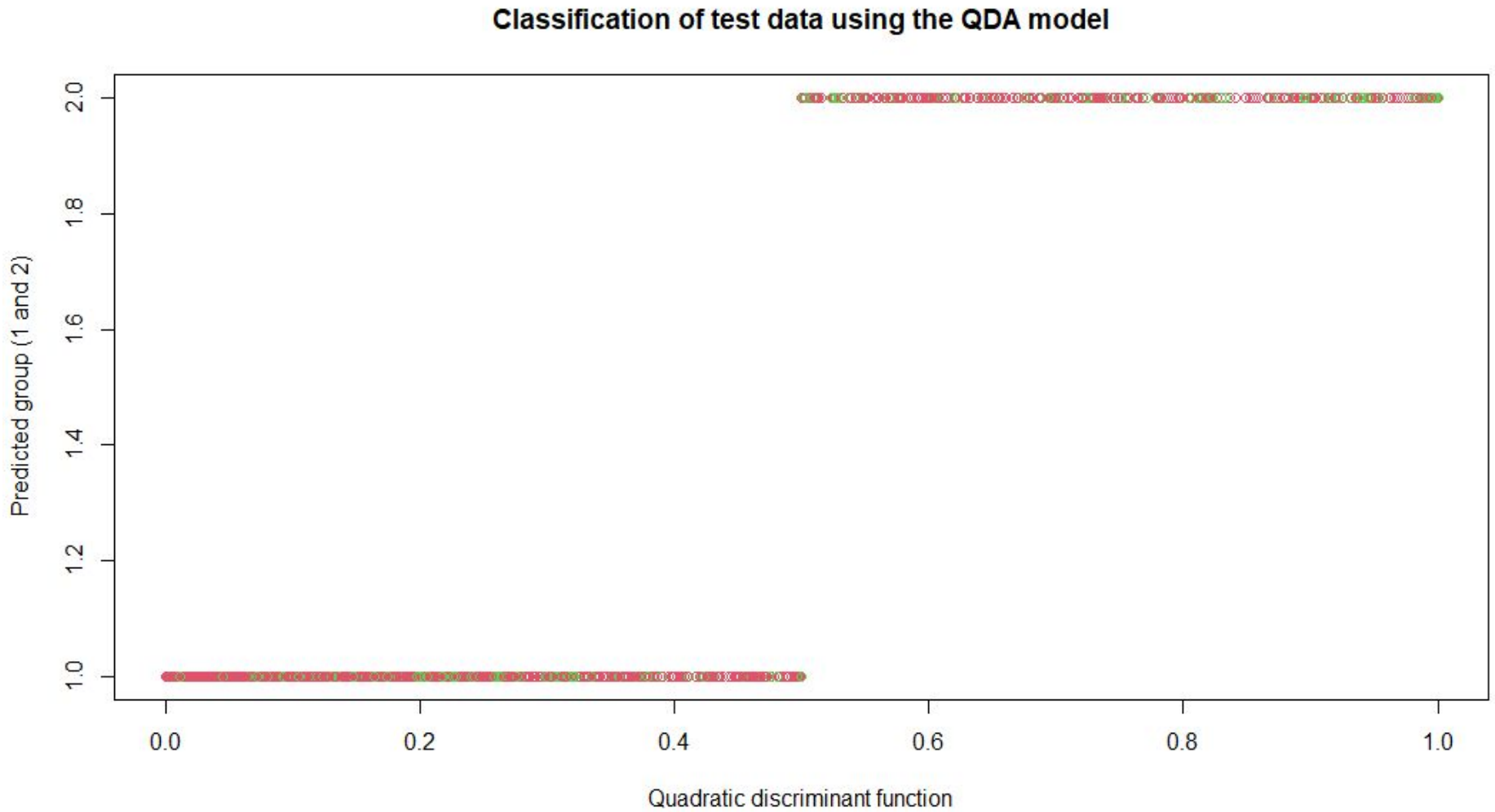


Coefficients of the linear discriminant:

	LD1
Graduate	0.0201
Clerks	0.0034
PartTime	-0.0103
NonDep	0.0051
Unionization	0.0171
Strikes	0.1083
Investments	0.0677
ExecManag	0.0085
under35	-0.0021
NonStandWork	0.0016




Classification of test data using the LDA model
Predicted group (1 and 2)
1.0
1.2
1.4
1.6
1.8
2.0
0
5
10
Linear discriminant function


Classification of test data using the QDA model
Predicted group (1 and 2)
1.0
1.2
1.4
1.6
1.8
2.0
0.0
0.2
0.4
0.6
0.8
1.0
Quadratic discriminant function

Canonical Correlation Analysis

Data reduction technique in a multiple output setting.

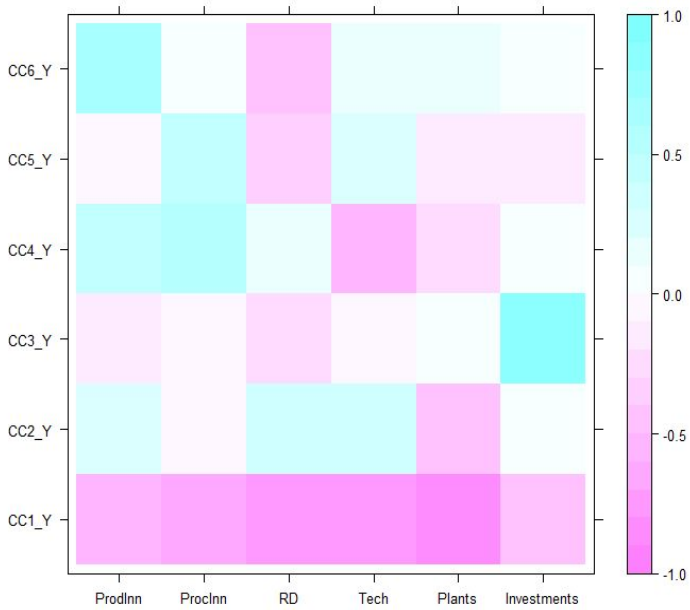
Dataset: 1554 observations of 34 variables (6 variables innovation-related and 28 variables for the workforce)

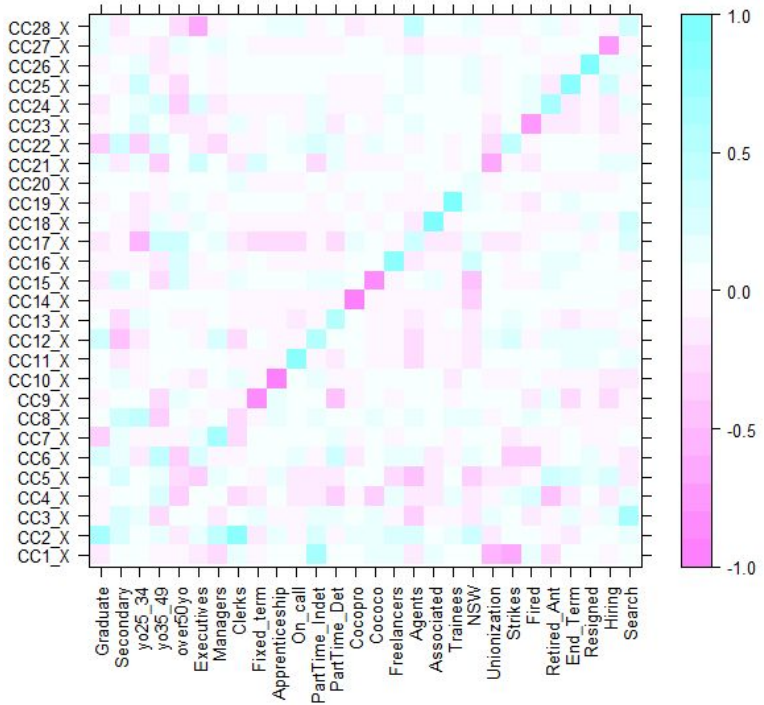
Define U as the set of linear combinations from the set of variables X (workforce)
Define V as the set of linear combinations from the set of variables Y (innovation)
Each member of U is paired with a member of V to give the i^{th} canonical component (U_i, V_i) and their coefficients are those that maximize the canonical correlation between them.

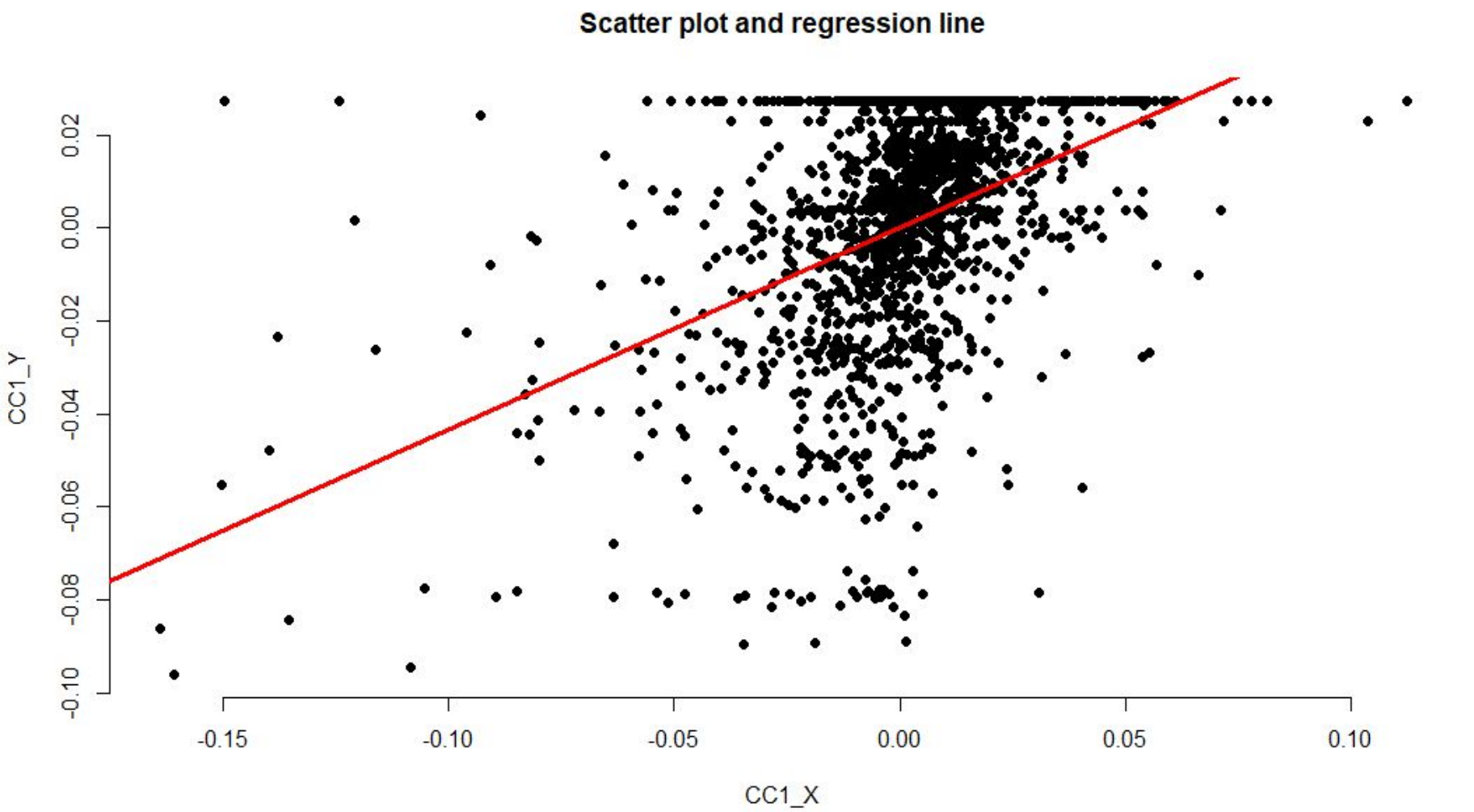
$$U_1 = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p$$

$$V_1 = b_{11}Y_1 + b_{12}Y_2 + ... + b_{1k}Y_k$$

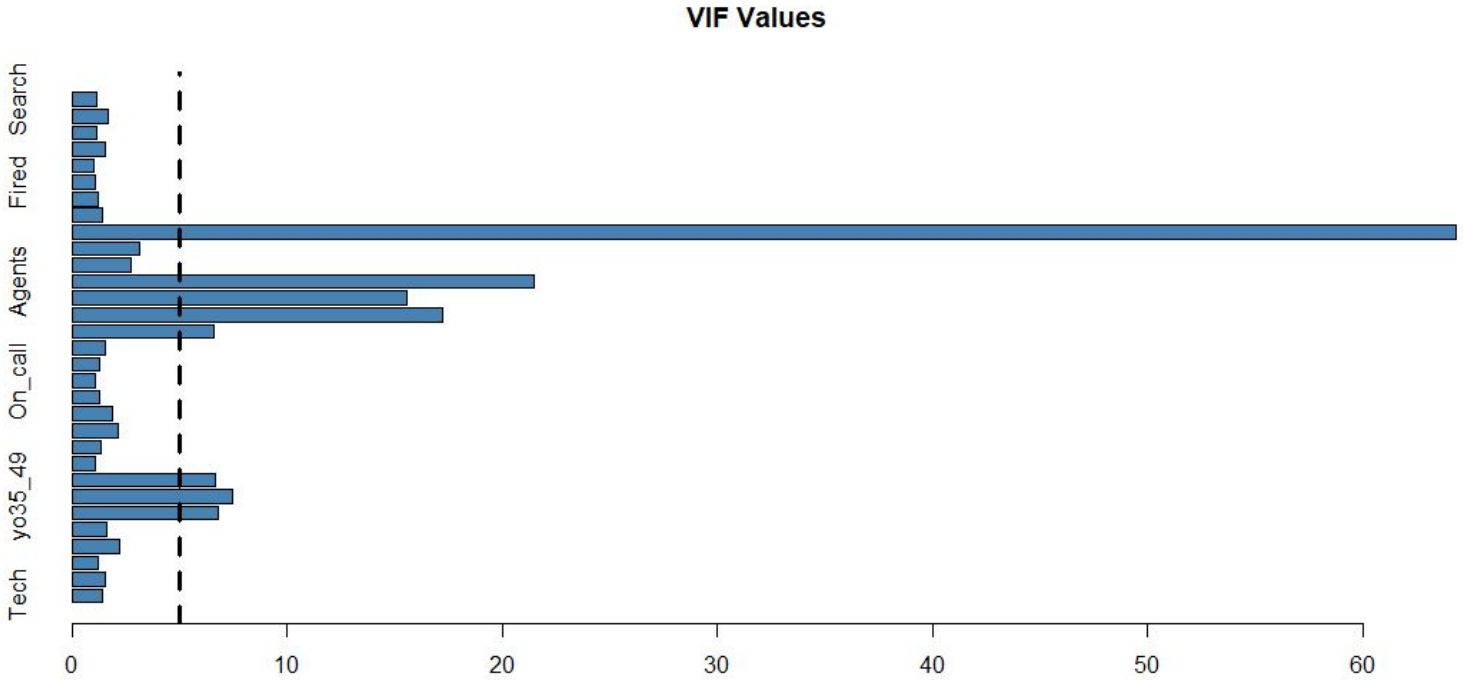
	Canonical Correlation	Squared Canonical Correlation
CC1	0.4337	0.1882
CC2	0.3352	0.1123
CC3	0.2630	0.0692
CC4	0.1976	0.0391
CC5	0.1777	0.0316
CC6	0.1141	0.0130


CC6_Y
CC5_Y
CC4_Y
CC3_Y
CC2_Y
CC1_Y
ProdInn
ProcInn
RD
Tech
Plants
Investments
1.0
0.5
0.0
-0.5
-1.0

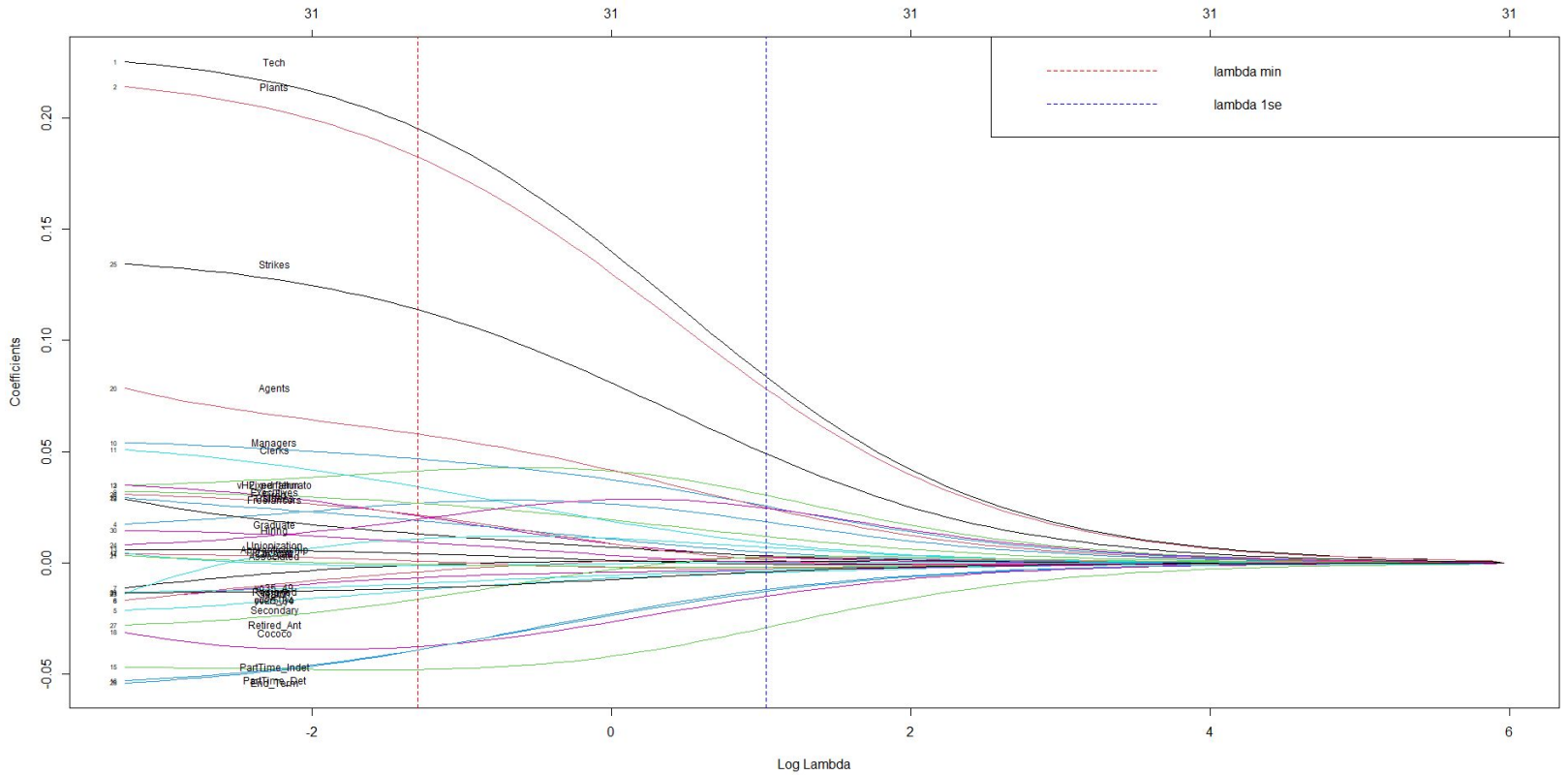

CC28_X
CC27_X
CC26_X
CC25_X
CC24_X
CC23_X
CC22_X
CC21_X
CC20_X
CC19_X
CC18_X
CC17_X
CC16_X
CC15_X
CC14_X
CC13_X
CC12_X
CC11_X
CC10_X
CC9_X
CC8_X
CC7_X
CC6_X
CC5_X
CC4_X
CC3_X
CC2_X
CC1_X
Graduate
Secondary
yo25_34
yo35_49
ower50yo
Executives
Managers
Clerks
Fixed_term
Apprenticeship
On_call
PartTime_Indet
PartTime_Det
Cocopro
Cococo
Freelancers
Agents
Associated
Trainees
NSW
Unionization
Strikes
Fired
Retired_Ant
End_Term
Resigned
Hiring
Search
1.0
0.5
0.0
-0.5
-1.0


Scatter plot and regression line
CC1_Y
0.02
0.00
-0.02
-0.04
-0.06
-0.08
-0.10
-0.15
-0.10
-0.05
0.00
0.05
0.10
CC1_X

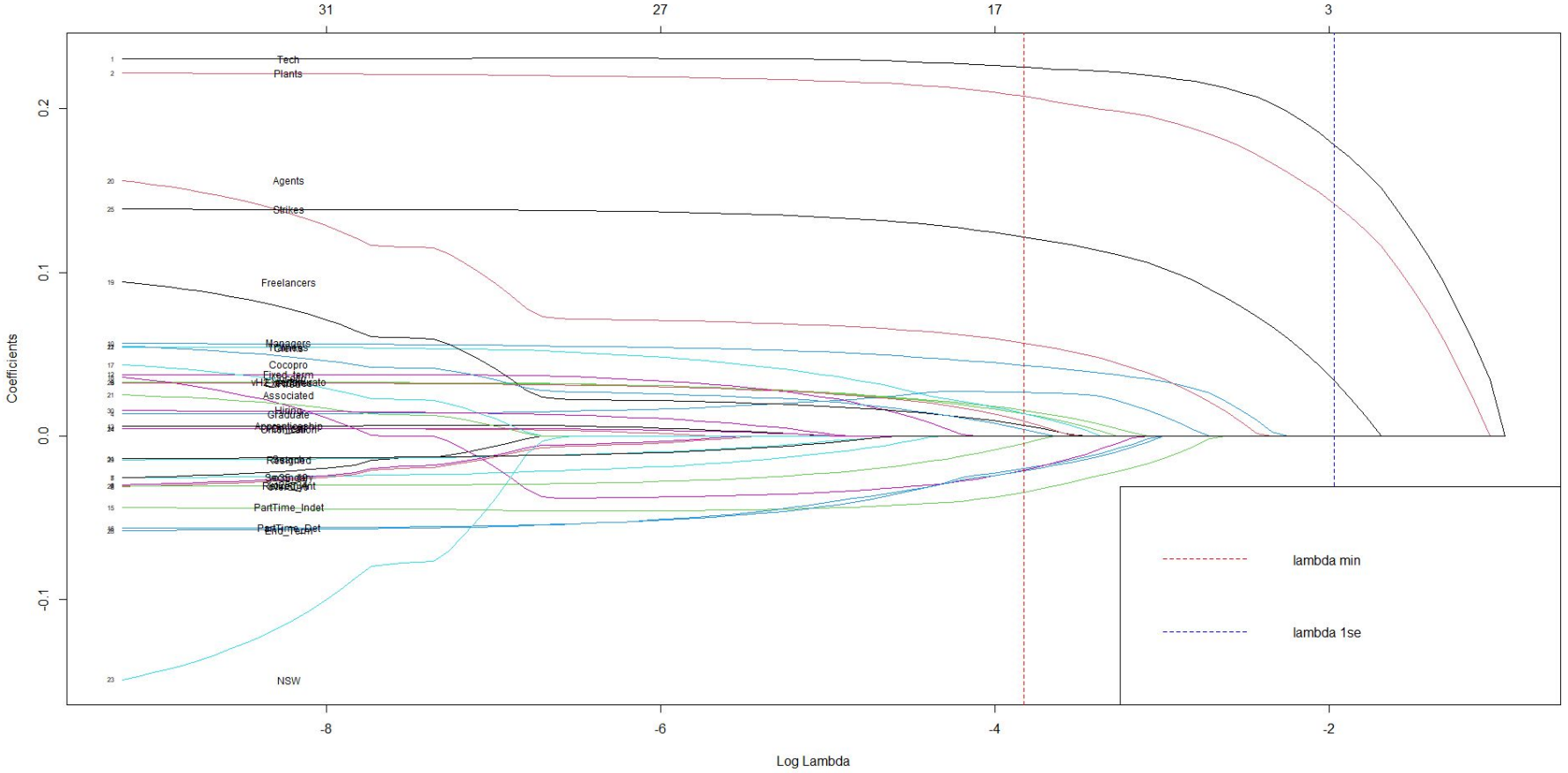
Model selection: preliminary step



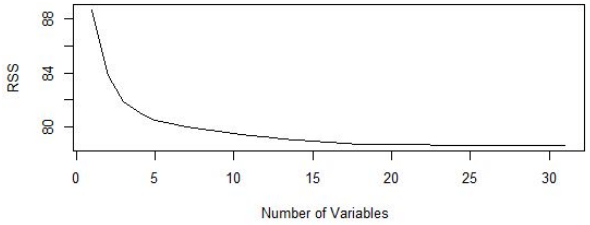
Ridge

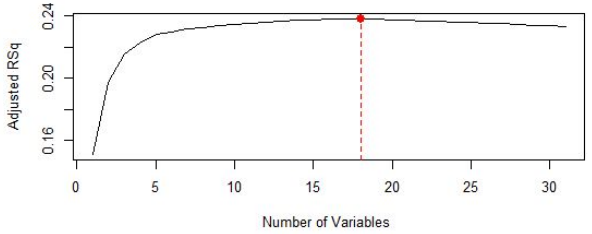

31
31
31
31
31
lambda min
lambda 1se
Tech
Plants
Strikes
Agents
Managers
Clerks
Graduate
Secondary
Retired_Ant
Cococo
PartTime_Indet
PartTime_Ret
Coefficients
0.20
0.15
0.10
0.05
0.00
-0.05
-2
0
2
4
6
Log Lambda

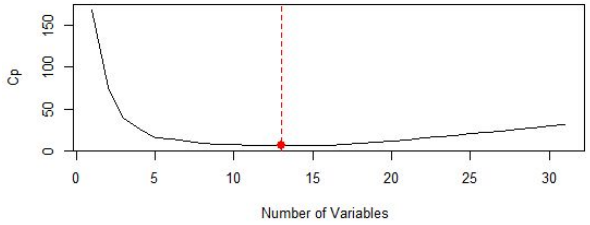
Lasso

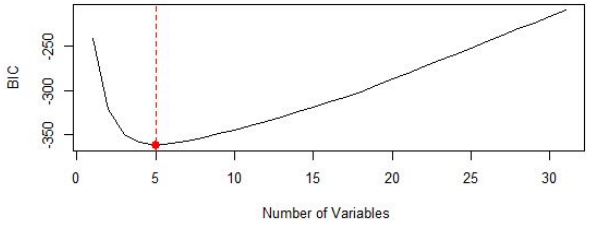

31
27
17
3
Tech
Plants
Agents
Strikes
Freelancers
Managers
Cocopro
Fixed_term
Associated
Hiring
Graduate
Apprenticeship
PartTime_Indet
PartTime_Det
NSW
0.2
0.1
0.0
-0.1
Coefficients
-8
-6
-4
-2
Log Lambda
lambda min
lambda 1se

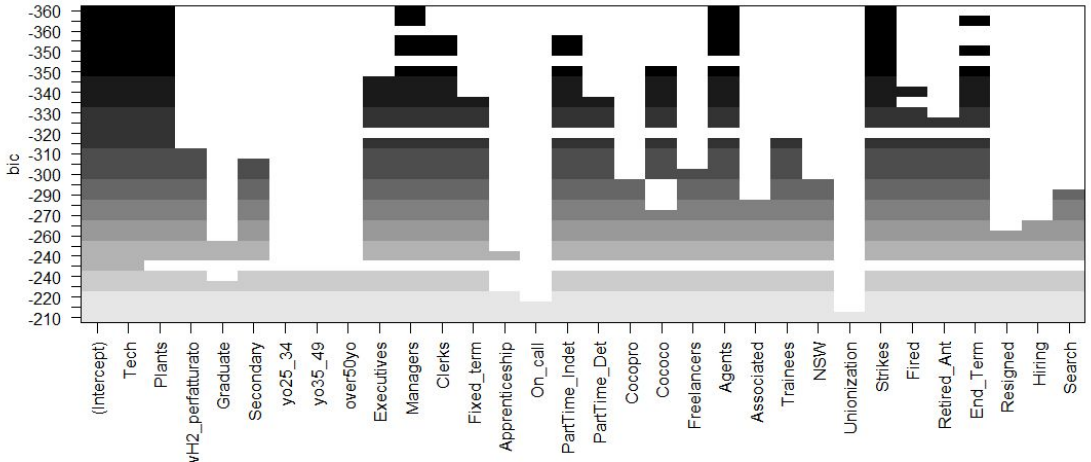
Best subset selection











	OLS	Ridge	Lasso	Best Subset Selection
Intercept	5.345e-02	6.011310e-17	0.0051481	0.008598220
Tech	1.882e-01	2.005099e-01	0.1894545	0.196269672
Plants	1.604e-01	1.877347e-01	0.1607161	0.162103744
Investment	1.410e-03	4.061088e-02	0.0012978	-
Graduate	1.900e-04	2.596943e-02	0.0003989	-
Secondary	-2.804e-04	-1.322601e-02	-	-
Yo25-34	-5.060e-04	-7.144631e-03	-	-
Yo35-49	-3.993e-04	-1.342709e-03	-	-
Over50yo	-5.162e-04	-4.924630e-03	-	-
Executives	1.505e-03	2.777630e-02	0.0014293	-
Managers	2.216e-03	4.779854e-02	0.0020202	0.003039718
Clerks	4.879e-04	3.636433e-02	0.0003370	-
Fixed term	9.790e-04	2.309893e-02	-	-
Apprenticeship	2.989e-04	4.624033e-03	-	-
On call	2.586e-04	1.416454e-03	-	-
Part time Indet	-4.989e-04	-4.793632e-02	-0.0006035	-
Part time Det	-2.844e-03	-4.105039e-02	-0.0017854	-
Cocopro	3.635e-03	-1.031898e-03	-0.0016019	-
Cococo	2.142e-03	-3.791958e-02	-	-
Freelancers	4.568e-03	1.441669e-02	0.0009239	-
Agents	6.068e-03	6.002363e-02	0.0024846	0.002744875
Associated	3.788e-03	-2.654105e-04	-	-
Trainees	6.609e-03	2.031376e-02	0.0028862	-
Non Dep over Dep	-3.604e-03	9.792917e-03	-	-
Unionization	8.319e-05	1.803342e-02	0.0000621	-
Strikes	9.083e-03	1.171805e-01	0.0089805	0.009321010
Fired	1.334e-03	2.311097e-02	0.0014677	-
Retired Ant	-1.800e-02	-1.809149e-02	-0.0179947	-
End term	-1.612e-03	-4.122521e-02	-0.0010656	-
Resigned	-6.160e-04	-9.738312e-03	-	-
Hiring	2.719e-04	1.028398e-02	-	-
Search	-7.230e-04	-1.168418e-02	-	-
				-
RMSE		0.8703748	0.9056538	0.230217

Further improvements

Economic motivation

- Add the spatial (province) dimension plus wage share
- Why agents are important for innovation capacity?

Statistical motivation

- Assess and compare predictive capacity of each model through cross-validation

References

Ada, Ayşen Altun. "*The relationship between "innovation, science and technology" and "economic development": canonical correlation analysis for the EU countries*." International Journal of Social Sciences 78.1 (2019).

Costa, Stefano, et al. "*From organizational capabilities to corporate performances: at the roots of productivity slowdown*." No. 2021/21. LEM Working Paper Series, 2021.

Gareth, James, et al. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013.

Pavitt, Keith."S*ectoral patterns of technical change: towards a taxonomy and a theory.*" Research policy 13.6 (1984): 343-373.