Universitat de Girona
Escola Politècnica Superior

Treball final de grau

Estudi: Grau en Enginyeria Informàtica

Títol: Connexió de CoDaPack amb R

Document: Resum

Alumne: Daniel Re Lartigue

Tutor: Santiago Thio Fernandez de Henestrosa
Departament: Informàtica, Matemàtica Aplicada i Estadística
Àrea: Estadística i Investigació Operativa

Convocatòria (mes/any) Juny/2020

1 Introducció

Durant els darrers anys, s'ha desenvolupat una nova visió metodològica per l'anàlisi estadística de dades composicionals, després del plantejament introduït a principis dels anys vuitanta per John Aitchison. El Grup de Recerca de Dades Composicionals de Girona són els seus deixebles i actualment líders en el món en aquest camp. Aquesta metodologia no és senzilla per utilitzar-la amb paquets estadístics estàndard.

És per això que va ser desenvolupat el CoDaPack (Copositional Data Package), un software gratuït, independent i multiplataforma que implementa en aquest moment el més elemental del mètodes estadístics anomenats anteriorment. Aquest software està orientat a usuaris sense una amplia experiència en l'ús de diversos paquets informàtics.

Hi ha llibreries CODA per a R però que no tenen connexió amb el CoDaPack, ja que està fet amb JAVA, d'aquí és on sorgeix la necessitat de recerca per tal de buscar la manera de connectar R amb JAVA i poder incloure les llibreries al CoDaPack.

Aquest treball per tant intenta resoldre aquesta necessitat i fusionar els dos recursos que tenim en un mateix d'una manera transparent i senzilla per l'usuari de l'aplicació.

2 Objectius

L'objectiu principal d'aquest treball és analitzar, (dissenyar) i implementar una llibreria amb JAVA que permeti utilitzar les funcionalitats de les llibreries que ja tenim de CODA fetes amb R mitjançant menús del CoDaPack i de forma totalment transparent per l'usuari. També està l'objectiu de crear una interfície genèrica per tal que usuaris de l'aplicació puguin crear les seves pròpies rutines i executar-les amb el CoDaPack.

Per tal d'aconseguir aquests objectius s'haurà d'aconseguir fer les següents fites:

- Estudiar com obrir R i carregar paquets de forma automàtica des de JAVA.
- Implementar rutines de JAVA específiques que facin crides a les rutines CODA d'R.
- Incorporar aquestes rutines al programa CoDaPack.
- Establir una interfície que permeti als usuaris programar les seves pròpies rutines amb R i que les puguin executar de forma transparent en el CoDaPack.
- Fer un instal·lador integrat.

3 Conceptes previs

3.1 Llibreries d'enllaç dinàmic

Una biblioteca d'enllaç dinàmic, es refereix als arxius de codi executable que es carreguen sota demanda d'un programa per part del sistema operatiu. Aquesta denominació és exclusiva dels sistemes operatius Windows però el terme existeix en la majoria de sistemes operatius actuals.

3.2 Maven

Maven és una eina per la gestió de projectes de software, es basa en el concepte de POM(Proyect ObjectModel). Amb Maven, podrem compilar, empaquetar, generar documentació, passar tests, preparar les builds, etc.

En el fitxer *pom.xml* és on descriurem el nostre projecte, una de les coses més importants que ens permet fer Maven és posar-hi en aquest fitxer dependències, això ens permet fer ús de llibreries i importar-les des de Maven si ho indiquem correctament en aquest fitxer.

3.3 Una petita introducció a R

R és un llenguatge de programació interpretat, de distribució lliure sota llicència GNU, i té un ambient per al còmput estadístic i gràfic. És un software multiplataforma.

R està dividit en dues parts conceptuals, la primera seria el seu sistema base que és el que ens podem descarregar a **CRAN**. L'altra part consisteix en els paquets modulars que el poden fer extensible.

Algunes de les tasques que podem fer per exemple utilitzant aquests paquets són les següents: mineria de textos, processament d'imatges, visualització interactiva de dades...

3.4 Variables d'entorn

Independent del sistema operatiu que estiguem utilitzant, les variables d'entorn són la manera més senzilla de poder passar informació d'una aplicació a una altra. Per tant en el cas que necessitem que una informació estigui disponible en alguna aplicació o eina del nostre sistema, simplement haurem de crear una variable d'entorn que la contingui.

Com el seu nom ens indica són variables per tant el seu valor pot variar, per tant poden ser alterades per un usuari, una aplicació o un script.

La importància de les variables d'entorn és que molta informació d'aplicacions o eines depèn d'elles per obtenir-ne informació i si alguna d'aquesta informació es trobés alterada llavors l'aplicació o eina deixaria de funcionar correctament.

4 Requisits funcionals

Els requisits funcionals defineixen una funció del sistema de software o dels seus components. Aquestes funcions són descrites com un conjunt d'entrades, uns comportaments i finalment unes sortides. En definitiva els requisits funcionals estableixen el comportament del software.

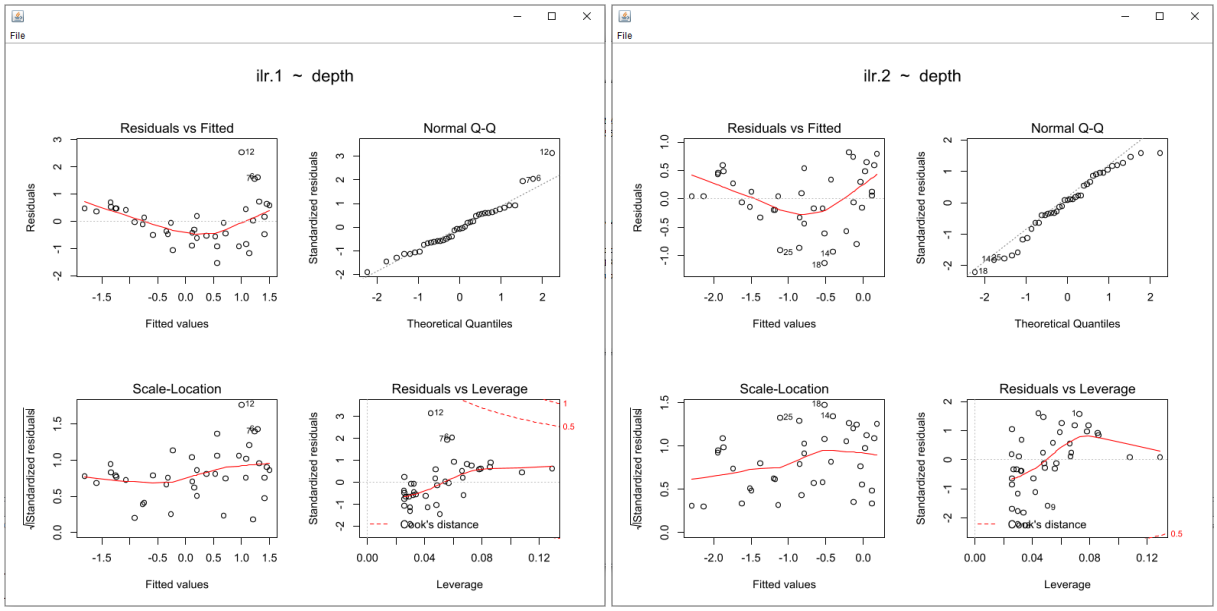
Els requisits funcionals dels usuaris finals que utilitzin l'aplicació són els següents:

- Esborrar totes les taules que hi hagin carregades en el sistema.
- Netejar la consola de missatges en el cas que així ho vulgui l'usuari.
- S'ha de poder definir un closure To per defecte en el menú de configuració.
- Mostrar o amagar els menús de desenvolupador en el menú de configuració.
- S'ha de poder canviar la ruta del directori on es busquen els scripts per defecte.
- Canviar els valors seleccionats d'una variable categòrica.
- Poder filtrar observacions amb el mateix valor en la variable categòrica seleccionada.
- Poder crear un data frame de zero (podent fer copy paste des de un excel), indicant el nombre de columnes i de files.
- Poder visualitzar els centres en els ternary/quaternary plot.

- Poder guardar els gràfics generats amb Java en format PDF.
- Visualitzar etiquetes en el balance dendogram.
- Poder guardar l'estat de la consola en sortir de l'aplicatiu.
- Guardar els gràfics generats amb R amb el format SVG.
- Poder canviar els noms de les variables directament en la taula.
- Poder visualitzar help en cada un dels menús.
- Discretitzar/Segmentar una variable continua en una variable categòrica (factor) segons el següent:
  - Method
  - # de levels
  - Si es vol afegir noms als grups o no
- Calcular una nova variable a partir d'altres amb una expressió d'R.
- Ordenar les dades seleccionades segons si es vol en ordre ascendent o descendent.
- Poder fer un subconjunt del conjunt de dades que compleixen una certa condició en les variables seleccionades.
- Executar la rutina Non-parametric Replacement esperant obtenir com a resultats unes noves variables en el data frame actual a partir dels següents paràmetres:
  - Selecció de variables
  - DL proportion
  - Closure result
  - Closure to
- Executar la rutina Logratio-EM zero Replacement esperant obtenir com a resultats unes noves variables en el data frame actual a partir dels següents paràmetres:
  - Selecció de variables
  - Rob Option
  - IniCov Option
  - DL proportion
- Executar la rutina Classical Univariate Noramlity test esperant obtenir com a resultat una sortida en la consola a partir dels següents paràmetres:
  - Selecció de variables
  - Normality Test
- Poder mostrar la següent informació amb R:
  - La versió d'R, el sistema operatiu i els paquets carregats
  - Obtenir els valors de les variables d'entorn
  - Informe sobre les característiques opcionals que han sigut compilades en la construcció d'R
- Poder executar rutines del CoDaPack fent crides a scripts d'R i passant paràmetres de Java a R un cop recollits en el menú interactiu del CoDaPack.

5 Resultats

Com que hi han molts resultats, s'ha decidit per resumir la feina feta mostrar les captures de l'execució d'un script d'R des del CoDaPack, on es recullen tots els possibles resultats (Regressió X real Y compositional): noves variables, gràfics, una nova taula i una sortida textual.



(a) Gràfic 1. (b) Gràfic 2.

Figura 1: Gràfics de la rutina Regressió X real Y compositional.

	Coefficients	ilr.1	ilr.2
1	intercept	2.76	1.04
2	depth	-0.07	-0.05

Figura 2: Nou data frame regressió X real Y compositional.

	num_sedim	sand	silt	clay	depth	ilr.1	ilr.2	ilr.1.r	ilr.2.r	ilr.1.f	ilr.2.f
1	S01	77.50	19.50	3.00	10.40	2.09	0.98	0.04	0.43	2.05	0.54
2	S02	71.90	24.90	3.20	11.70	2.11	0.75	0.14	0.27	1.97	0.48
3	S03	50.70	36.10	13.20	12.80	0.96	0.24	-0.93	-0.19	1.89	0.43
4	S04	52.36	41.02	6.62	13.00	1.59	0.17	-0.29	-0.25	1.88	0.42
5	S05	70.00	26.50	3.50	15.70	2.05	0.69	0.35	0.40	1.69	0.29
6	S06	66.50	32.20	1.30	16.30	2.92	0.51	1.26	0.25	1.65	0.26
7	S07	43.10	55.30	1.60	18.00	2.79	-0.18	1.25	-0.36	1.54	0.18
8	S08	53.40	36.80	9.80	18.70	1.23	0.26	-0.26	0.12	1.49	0.15
9	S09	15.50	54.40	30.10	20.70	-0.03	-0.89	-1.39	-0.94	1.36	0.05
10	S10	31.70	41.50	26.80	22.10	0.25	-0.19	-1.01	-0.17	1.26	-0.02
11	S11	65.70	27.80	6.50	22.40	1.54	0.61	0.30	0.64	1.24	-0.03
12	S12	70.40	29.00	0.60	24.40	3.53	0.63	2.42	0.76	1.11	-0.13
13	S13	17.40	53.60	29.00	25.80	0.04	-0.80	-0.97	-0.60	1.01	-0.20
14	S14	10.60	69.80	19.60	32.50	0.27	-1.33	-0.29	-0.81	0.56	-0.52
15	S15	38.20	43.10	18.70	33.60	0.63	-0.09	0.15	0.49	0.48	-0.57
16	S16	10.80	52.70	36.50	36.80	-0.35	-1.12	-0.61	-0.40	0.27	-0.72
17	S17	18.40	50.70	30.90	37.80	-0.01	-0.72	-0.21	0.06	0.20	-0.77
18	S18	4.60	47.40	48.00	36.90	-0.96	-1.68	-1.22	-0.92	0.26	-0.73
19	S19	15.60	50.40	34.00	42.20	-0.16	-0.83	-0.06	0.15	-0.10	-0.98
20	S20	31.90	45.10	23.00	47.00	0.41	-0.24	0.83	0.97	-0.42	-1.21
21	S21	9.50	53.50	37.00	47.10	-0.40	-1.22	0.03	-0.00	-0.43	-1.22
22	S22	17.10	48.00	34.90	48.40	-0.16	-0.73	0.36	0.55	-0.52	-1.28
23	S23	10.50	55.40	34.10	49.40	-0.28	-1.18	0.30	0.15	-0.59	-1.33
24	S24	4.78	54.43	40.80	49.50	-0.76	-1.72	-0.16	-0.39	-0.59	-1.33
25	S25	2.60	45.20	52.20	59.20	-1.28	-2.02	-0.03	-0.22	-1.25	-1.80

Figura 3: Noves variables Regressió X real Y compositional.

```
X real Y composition regression:

LINEAR REGRESSION

Response ilr.1 :

Call:
lm(formula = ilr.1 ~ as.matrix(X))

Residuals:
      Min        1Q    Median        3Q       Max
-1.38562 -0.28996 -0.03353  0.30368  2.42271

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.75750    0.40346   6.835 5.71e-07 ***
as.matrix(X) -0.06769    0.01212  -5.587 1.10e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8632 on 23 degrees of freedom
Multiple R-squared:  0.5757,    Adjusted R-squared:  0.5573
F-statistic: 31.21 on 1 and 23 DF,  p-value: 1.102e-05


Response ilr.2 :

Call:
lm(formula = ilr.2 ~ as.matrix(X))

Residuals:
      Min        1Q    Median        3Q       Max
-0.93682 -0.35501  0.05587  0.39758  0.96976

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.043614   0.245805   4.246 0.000305 ***
as.matrix(X) -0.048047   0.007382  -6.509 1.22e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5259 on 23 degrees of freedom
Multiple R-squared:  0.6481,    Adjusted R-squared:  0.6328
F-statistic: 42.36 on 1 and 23 DF,  p-value: 1.216e-06



r^2 =  [1] 59.81165
```

Figura 4: Sortida textual regressió X real Y compositional.

6 Conclusió

Els objectius del projecte inicialment consistien a analitzar, dissenyar i implementar una llibreria amb Java que permeti utilitzar llibreries de CODA desenvolupades amb R. Per tant també crear una interfície genèrica per tal que usuaris poguessin crear les seves pròpies rutines escrites amb R i executar-les en el CoDaPack.

Com que els objectius i funcionalitats definides al principi del projecte estan presents en l'aplicació desenvolupada fins ara, es considera que els objectius s'han assolit correctament. A més a més s'han desenvolupat diferents tipus de menús de desenvolupador, segons l'usuari vulgui utilitzar alguns paràmetres per la seva rutina o uns altres.

L'aplicació ha evolucionat molt favorablement i l'objectiu d'aconseguir ajuntar R amb Java s'ha assolit. Aquest projecte podríem dir que té moltes possibilitats i un futur al davant amb possibles noves millores i/o rutines noves.

L'aplicació actualment està disponible i operativa. Aquesta nova versió del CoDaPack ha permès ampliar considerablement les coses que es poden fer amb el CoDaPack gràcies al fet que ara té connexió amb R i per tant moltes coses que es poden fer amb R també ara es poden fer amb el CoDaPack. Aquestes millores i ampliacions els usuaris i la comunitat les agrairan bastant.

A continuació comentaré una sèrie de coneixements i experiència que he assolit al llarg del desenvolupament del projecte:

- Coneixements complementaris amb programació de Java, ja que ara sé treballar molt millor amb dependències i llibreries. També comentar que he après a incorporar llenguatges dins de Java, ja sigui HTML, R, JavaScript, CSS o Yaml.
- He après també a crear instal·ladors per als diferents sistemes operatius amb la creació de variables d'entorn sigui Windows o MacOS.
- Manipular llibreries d'R i saber-les utilitzar correctament en l'àmbit de Java i rJava.
- Treballar en un equip de recerca en un àmbit professional.

La meva conclusió personal final ha estat bona, ja que he après moltes coses i també a saber tractar certs problemes des de diferents punts de vista, també hem sento molt satisfet d'haver desenvolupat correctament el projecte i haver aconseguit els objectius marcats al principi del projecte.

Finalment vull comentar que des d'un principi he rebut una molt bona acollida en el departament i per qualsevol cosa sempre he pogut comptar amb qualsevol persona. Amb qui més contacte he tingut ha sigut amb en Marc Comas i en Santi Thió que han estat en tot moment al meu costat per ajudar-me per resoldre dubtes o simplement anar al meu costat durant el desenvolupament del projecte. Per tot això vull expressar el meu agraïment a tot el departament que m'han fet sentir un més del departament, en especial a en Santi Thió i en Marc Comas.