ISV51: Programmation sous R

Introduction

L3 GBI – Université d"Evry

semestre d'automne 2015

`http://julien.cremeriefamily.info/teachings_L3BI_ISV51.html`

Laboratoire de Mathématiques et Modélisation d'Évry

Intervenant

Équipe « Statistique & Génome », LaMME

Laboratoire de Mathématiques et Modélisation d'Évry
LaMME

`http://www.math-evry.cnrs.fr/`

Julien Chiquet – maître de Conférences, statistiques

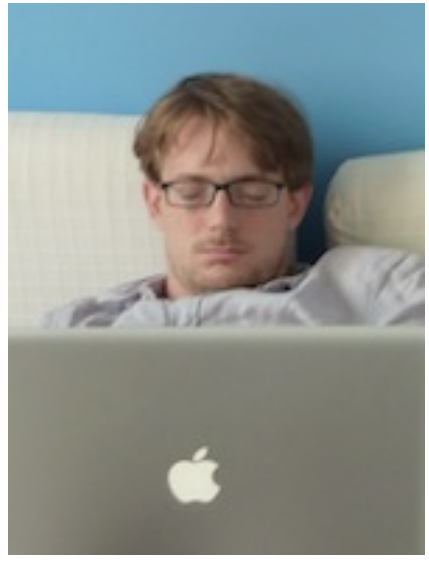
`julien.chiquet@genopole.cnrs.fr,@gmail.com`

Intervenant

Équipe « Statistique & Génome », LaMME

Laboratoire de Mathématiques et Modélisation d'Évry
LaMME

`http://www.math-evry.cnrs.fr/`



Julien Chiquet – maître de Conférences, statistiques

`julien.chiquet@genopole.cnrs.fr,@gmail.com`

Agenda (théorique) du semestre

1. Les Structures de données leur manipulation
2. Analyse de données élémentaire
3. Programmation sous `R` : du niveau débutant à intermédiaire
4. Vers le niveau avancé : (parallélisation, `R/C++`, `ggplot2`)

+ utilisation de R-studio (développement, publier un rapport).

+ évaluation sur machine.

Agenda (théorique) du semestre

1. Les Structures de données leur manipulation
2. Analyse de données élémentaire
3. Programmation sous `R` : du niveau débutant à intermédiaire
4. Vers le niveau avancé : (parallélisation, `R/C++`, `ggplot2`)

+ utilisation de R-studio (développement, publier un rapport).

+ évaluation sur machine.

Agenda (théorique) du semestre

1. Les Structures de données leur manipulation
2. Analyse de données élémentaire
3. Programmation sous `R` : du niveau débutant à intermédiaire
4. Vers le niveau avancé : (parallélisation, `R/C++`, `ggplot2`)

+ utilisation de R-studio (développement, publier un rapport).

+ évaluation sur machine.

Plan

Qu'est-ce que R ?

En deux mots,

R est un logiciel de développement scientifique spécialisé dans le calcul et l'analyse statistique.

R est aussi

- un langage,
- un environnement,
- un projet open source (projet GNU),
- un logiciel multi-plateforme (Linux, Mac, Windows),

Qu'est-ce que R ?

En deux mots,

R est un logiciel de développement scientifique spécialisé dans le calcul et l'analyse statistique.

R est aussi

- un langage,
- un environnement,
- un projet open source (projet GNU),
- un logiciel multi-plateforme (Linux, Mac, Windows),

Principales fonctionnalités

1. Gestionnaire de données
   - Lecture, manipulation, stockage.
2. Algèbre linéaire
   - Opérations classiques sur vecteurs, tableaux et matrices
3. Statistiques et analyse de données
   - Dispose d'un *grand* nombre de méthodes d'analyse de données (des plus anciennes et aux plus récentes)
4. Moteur de sorties graphiques
   - Sorties écran ou fichier
5. Système de modules
   - Alimenté par la communauté
6. Interface « facile » avec `C/C++`, `Fortran`,...

Principales fonctionnalités

1. Gestionnaire de données
   - Lecture, manipulation, stockage.
2. Algèbre linéaire
   - Opérations classiques sur vecteurs, tableaux et matrices
3. Statistiques et analyse de données
   - Dispose d'un *grand* nombre de méthodes d'analyse de données (des plus anciennes et aux plus récentes)
4. Moteur de sorties graphiques
   - Sorties écran ou fichier
5. Système de modules
   - Alimenté par la communauté
6. Interface « facile » avec `C/C++`, `Fortran`,...

Historique

Approche chronologique

- 1970s développement de S au Bell labs.
- 1980s développement de S-PLUS au AT&T. Lab
- 1993 développement de R sur le modèle de S par Robert Gentleman et Ross Ihaka au département de statistique de l'université d'Auckland.
- 1995 dépôts des codes sources sous licence GNU/GPL
- 1997 élargissement du groupe
- 2002 la fondation R dépose ses statuts sous la présidence de Gentleman et Ihaka
- 2007 création de revolution analytics
- 2011 première version public de R-studio

Mode de diffusion

Développement ~~entièrement~~ bénévole

- « R development core team » (20aine de personnes)
- Participation de *nombreux* chercheurs (>7000 packages en Août 2015)
- Avec l'essor du « Big Data », apparition d'outils payants « autour de R » (revolution, rstudio entreprise, etc.) version payante

Qualités et défauts de R

Plus ☺

1. Libre et gratuit,
2. Richesse des modules,
3. Souplesse,
4. Prise en main rapide (Syntaxe intuitive et compact),
5. Développement rapide (langage de scripts),
6. Nombreuses possibilités graphiques.

Moins ☹

1. Aide intégrée succincte,
2. ~~debugger un peu sec~~,
3. Code parfois illisible (compacité),
4. Facile de « mal » coder,
5. Lent par rapport à C/C++,
6. Personnalisation des graphiques un peu lourde.

Qualités et défauts de R

Plus ☺

1. Libre et gratuit,
2. Richesse des modules,
3. Souplesse,
4. Prise en main rapide (Syntaxe intuitive et compact),
5. Développement rapide (langage de scripts),
6. Nombreuses possibilités graphiques.

Moins ☹

1. Aide intégrée succincte,
2. ~~debugger un peu sec~~,
3. Code parfois illisible (compacité),
4. Facile de « mal » coder,
5. Lent par rapport à C/C++,
6. Personnalisation des graphiques un peu lourde.

Les concurrents plus ou moins directs

Les logiciels de développement scientifique sont spécialisés en

1. algèbre linéaire
   - `Matlab` (Mathworks), la référence,
   - `Scilab` (INRIA), l'alternative libre,
   - `Octave` (GNU), l'alternative open source ☺,
2. statistiques
   - `SAS` (SAS Inc.), la référence,
   - `S-PLUS` (TIBCO), le concurrent,
   - `R` (GNU), l'alternative open source ☺,
3. calcul symbolique
   - `Mathematica` (Wolfram), la référence,
   - `Maple` (Maplesoft), la référence aussi,
   - `Maxima` (GNU), l'alternative open source ☺.

+ `Python` + `SciPy` et `Julia`

Les concurrents plus ou moins directs

Les logiciels de développement scientifique sont spécialisés en

1. algèbre linéaire
   - `Matlab` (Mathworks), la référence,
   - `Scilab` (INRIA), l'alternative libre,
   - `Octave` (GNU), l'alternative open source ☺,
2. statistiques
   - `SAS` (SAS Inc.), la référence,
   - `S-PLUS` (TIBCO), le concurrent,
   - `R` (GNU), l'alternative open source ☺,
3. calcul symbolique
   - `Mathematica` (Wolfram), la référence,
   - `Maple` (Maplesoft), la référence aussi,
   - `Maxima` (GNU), l'alternative open source ☺.

+ `Python` + `SciPy` et `Julia`

Les concurrents plus ou moins directs

Les logiciels de développement scientifique sont spécialisés en

1. algèbre linéaire
   - `Matlab` (Mathworks), la référence,
   - `Scilab` (INRIA), l'alternative libre,
   - `Octave` (GNU), l'alternative open source ☺,
2. statistiques
   - `SAS` (SAS Inc.), la référence,
   - `S-PLUS` (TIBCO), le concurrent,
   - `R` (GNU), l'alternative open source ☺,
3. calcul symbolique
   - `Mathematica` (Wolfram), la référence,
   - `Maple` (Maplesoft), la référence aussi,
   - `Maxima` (GNU), l'alternative open source ☺.

+ Python + SciPy et Julia

Les concurrents plus ou moins directs

Les logiciels de développement scientifique sont spécialisés en

1. algèbre linéaire
   - `Matlab` (Mathworks), la référence,
   - `Scilab` (INRIA), l'alternative libre,
   - `Octave` (GNU), l'alternative open source ☺,
2. statistiques
   - `SAS` (SAS Inc.), la référence,
   - `S-PLUS` (TIBCO), le concurrent,
   - `R` (GNU), l'alternative open source ☺,
3. calcul symbolique
   - `Mathematica` (Wolfram), la référence,
   - `Maple` (Maplesoft), la référence aussi,
   - `Maxima` (GNU), l'alternative open source ☺.

+ `Python` + `SciPy` et `Julia`

Matlab versus R

- Obtenir de l'aide

```
help -i                 help.start ()
help                    help(help)
help sort               help(sort) _or_ ?sort
```

- Séquence de vecteurs

```
1:10                    1:10  _or_ seq(10)
1:3:10                  seq(1,10,by=3)
10:-1:1                 10:1
linspace(1,10,7)        seq(1,10,length=7)
```

- Manipulation de vecteurs

```
a=[2 7 8 5]              a <- c(2,7,8,5)
a=a[3:4                  a <- a[c(3,4)]
adash=[2 3 4 5]'         adash <- t(c(2,3,4,5))
```

Se tenir informé (« à l'ancienne »)

1. La page web de la fondation `R`
   - les statuts, des liens, des références.
   - `http://www.r-project.org/`

2. La page web du CRAN (Comprehensive R Arxiv Network)
   - binaires d'installation, packages, documentations, ...
   - `http://cran.r-project.org/`

3. La conférence des utilisateurs de `R` :
   - annuelle, prochaine édition au Danemark
   - `http://user2015.math.aau.dk/`

4. *The `R` journal* propose des articles sur
   - de nouvelles extensions, des applications, des actualités.
   - `http://journal.r-project.org/`

Se tenir informé (« à l'ancienne »)

1. La page web de la fondation R
   - les statuts, des liens, des références.
   - `http://www.r-project.org/`

2. La page web du CRAN (Comprehensive R Arxiv Network)
   - binaires d'installation, packages, documentations, ...
   - `http://cran.r-project.org/`

3. La conférence des utilisateurs de R :
   - annuelle, prochaine édition au Danemark
   - `http://user2015.math.aau.dk/`

4. *The R journal* propose des articles sur
   - de nouvelles extensions, des applications, des actualités.
   - `http://journal.r-project.org/`

Se tenir informé (« à l'ancienne »)

1. La page web de la fondation R
   - les statuts, des liens, des références.
   - `http://www.r-project.org/`

2. La page web du CRAN (Comprehensive R Arxiv Network)
   - binaires d'installation, packages, documentations, ...
   - `http://cran.r-project.org/`

3. La conférence des utilisateurs de R : useR!
   - annuelle, prochaine édition au Danemark
   - `http://user2015.math.aau.dk/`

4. *The R journal* propose des articles sur
   - de nouvelles extensions, des applications, des actualités.
   - `http://journal.r-project.org/`

Se tenir informé (« à l'ancienne »)

1. La page web de la fondation R
   - les statuts, des liens, des références.
   - `http://www.r-project.org/`

2. La page web du CRAN (Comprehensive R Arxiv Network)
   - binaires d'installation, packages, documentations, ...
   - `http://cran.r-project.org/`

3. La conférence des utilisateurs de R : useR!
   - annuelle, prochaine édition au Danemark
   - `http://user2015.math.aau.dk/`

4. *The R journal* propose des articles sur
   - de nouvelles extensions, des applications, des actualités.
   - `http://journal.r-project.org/`

Se tenir informé (comme les jeunes)

1. RSTUDIO,
   - Interface de développement multiplateforme pour R
   - binaires d'installation, packages, documentations, blog, etc.
   - `https://www.rstudio.com/`

2. REVOLUTION ANALYTICS, version « entreprise » de R
   - Support technique, développement spécifiques
   - passage à l'échelle/bigData orienté
   - `http://www.revolutionanalytics.com`

3. Datacamp, une plate-forme de cours en ligne
   - MOOC ne nécessitant pas d'installation préalable de R
   - `https://www.datacamp.com/`
   - voir aussi `http://tryr.codeschool.com/`

4. Blogs et plateformes alimentés par la communauté
   - `http://www.inside-r.org/`
   - `http://www.r-statistics.com/`
   - `http://www.r-bloggers.com/`
   - ...

Se tenir informé (comme les jeunes)

1. RSTUDIO,
   - Interface de développement multiplateforme pour R
   - binaires d'installation, packages, documentations, blog, etc.
   - `https://www.rstudio.com/`

2. REVOLUTION ANALYTICS, version « entreprise » de R
   - Support technique, développement spécifiques
   - passage à l'échelle/bigData orienté
   - `http://www.revolutionanalytics.com`

3. Datacamp, une plate-forme de cours en ligne
   - MOOC ne nécessitant pas d'installation préalable de R
   - `https://www.datacamp.com/`
   - voir aussi `http://tryr.codeschool.com/`

4. Blogs et plateformes alimentés par la communauté
   - `http://www.inside-r.org/`
   - `http://www.r-statistics.com/`
   - `http://www.r-bloggers.com/`
   - ...

Se tenir informé (comme les jeunes)

1. RSTUDIO,
   - Interface de développement multiplateforme pour `R`
   - binaires d'installation, packages, documentations, blog, etc.
   - `https://www.rstudio.com/`

2. REVOLUTION ANALYTICS, version « entreprise » de `R`
   - Support technique, développement spécifiques
   - passage à l'échelle/bigData orienté
   - `http://www.revolutionanalytics.com`

3. `Datacamp`, une plate-forme de cours en ligne
   - MOOC ne nécessitant pas d'installation préalable de `R`
   - `https://www.datacamp.com/`
   - voir aussi `http://tryr.codeschool.com/`

4. Blogs et plateformes alimentés par la communauté
   - `http://www.inside-r.org/`
   - `http://www.r-statistics.com/`
   - `http://www.r-bloggers.com/`
   - `...`

Se tenir informé (comme les jeunes)

1. RSTUDIO,
   - Interface de développement multiplateforme pour R
   - binaires d'installation, packages, documentations, blog, etc.
   - `https://www.rstudio.com/`

2. REVOLUTION ANALYTICS, version « entreprise » de R
   - Support technique, développement spécifiques
   - passage à l'échelle/bigData orienté
   - `http://www.revolutionanalytics.com`

3. `Datacamp`, une plate-forme de cours en ligne
   - MOOC ne nécessitant pas d'installation préalable de R
   - `https://www.datacamp.com/`
   - voir aussi `http://tryr.codeschool.com/`

4. Blogs et plateformes alimentés par la communauté
   - `http://www.inside-r.org/`
   - `http://www.r-statistics.com/`
   - `http://www.r-bloggers.com/`
   - `...`

Plan

Avant de démarrer

Installation et premiers contacts

Une session exemple

Installation

Rendez-vous sur la page du CRAN `http://cran.r-project.org/`

Mac

Télécharger `R-3.2.2.pkg`, cliquer.

Windows

Télécharger `R-3.2.2-win32.exe`.

Linux

Systèmes supportants `apt` (Debian, Ubuntu, . . .)

```
$ sudo apt-get update
$ sudo apt-get install r-base
```

Lancer R

Dans un terminal, taper 'R'

Premiers pas (mode console

```
$ R
R version 3.2.2 (2015-08-14) -- "Fire Safety"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)
[...]
Tapez 'demo()' pour des démonstrations, 'help()' pour l'aide
en ligne ou 'help.start()' pour obtenir l'aide au format HTML.
Tapez 'q()' pour quitter R.
> 1+1
[1] 2
```

Sortez moi de là !

```
> q()
Save workspace image? [y/n/c]:y
```

$\rightsquigarrow$ Sauve l'environnement et le réouvre la prochaine fois

Organiser un projet R

Solution sans R-studio

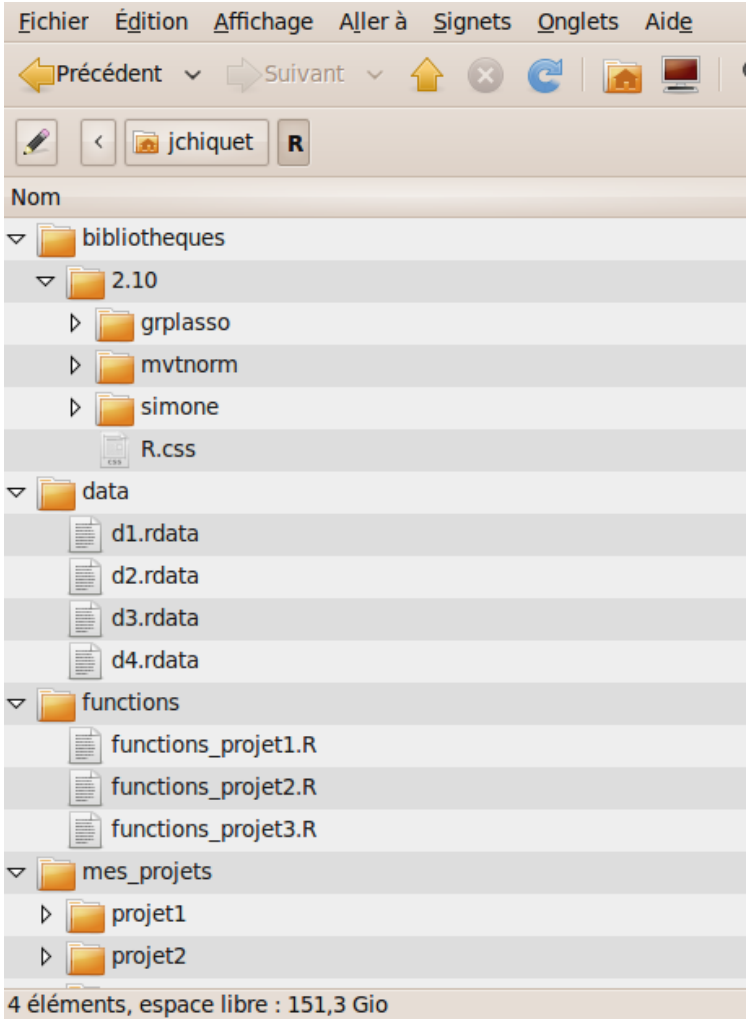


FIGURE: Arborescence type

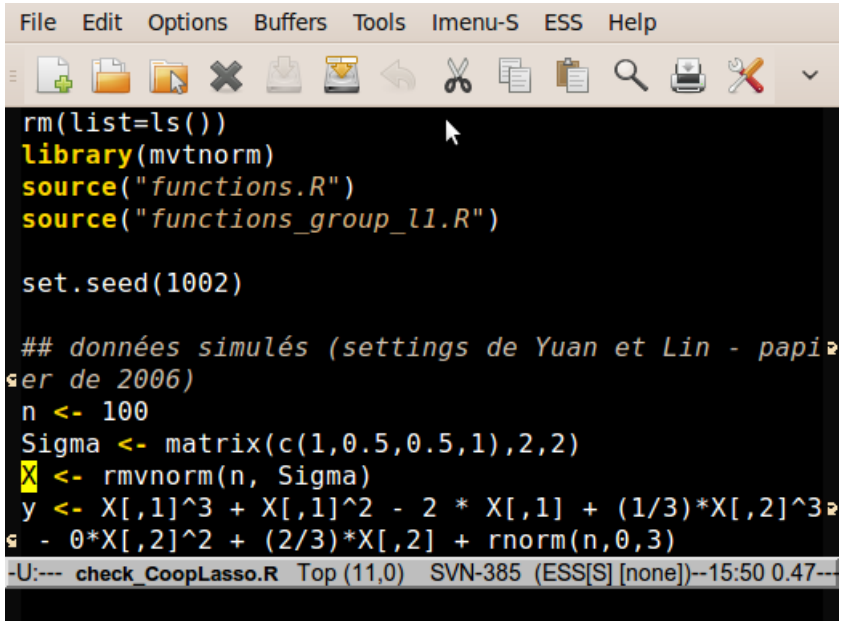
- Dans un répertoire `R`, placer
  - un répertoire `data`
  - un répertoire `mes_projets`
  - un répertoire `functions`
- Créer un répertoire par projet
  - sauvegarde des données

    ```
    save.image(file = "f.RData")
    ```
  - sauvegarde des instructions

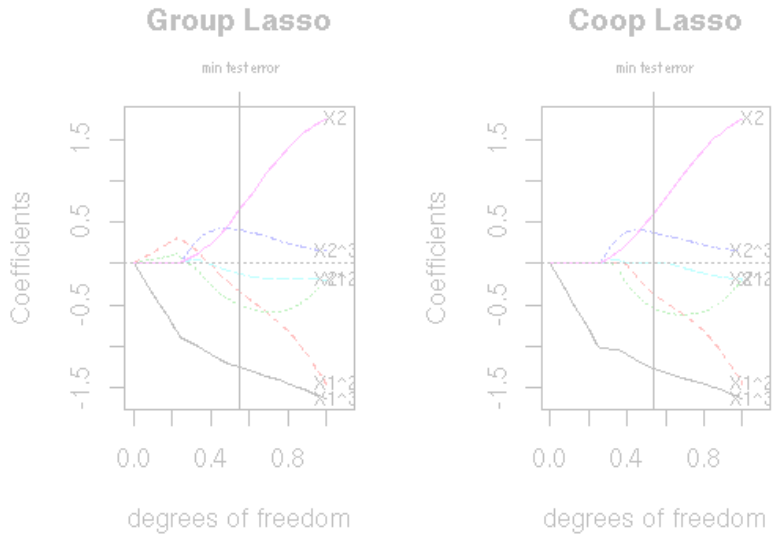
    ```
    savehistory(file = "f.Rhistory")
    ```
- `bibliotheques` contient les extensions installées.

Environnement de travail sous Linux

Un bureau de développement avec R(solution sans R-studio

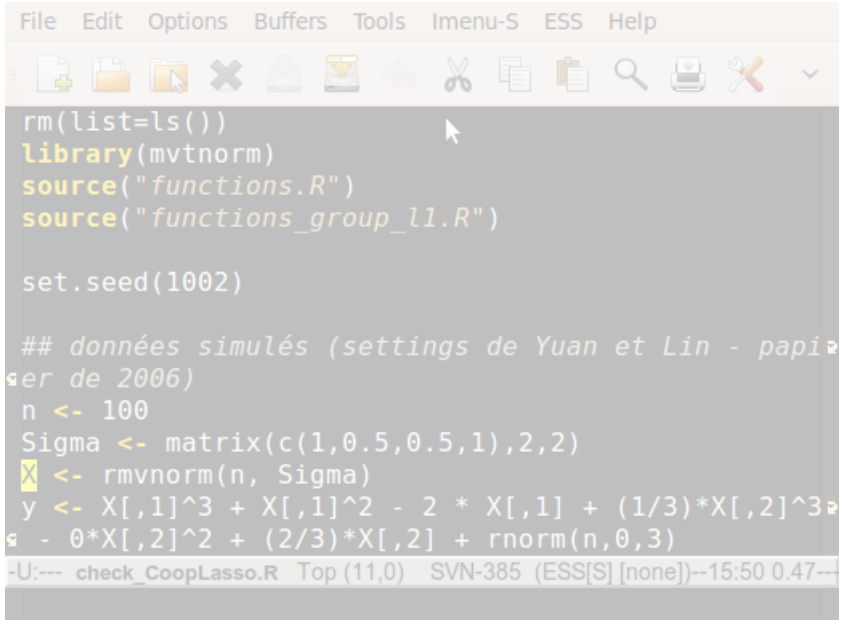


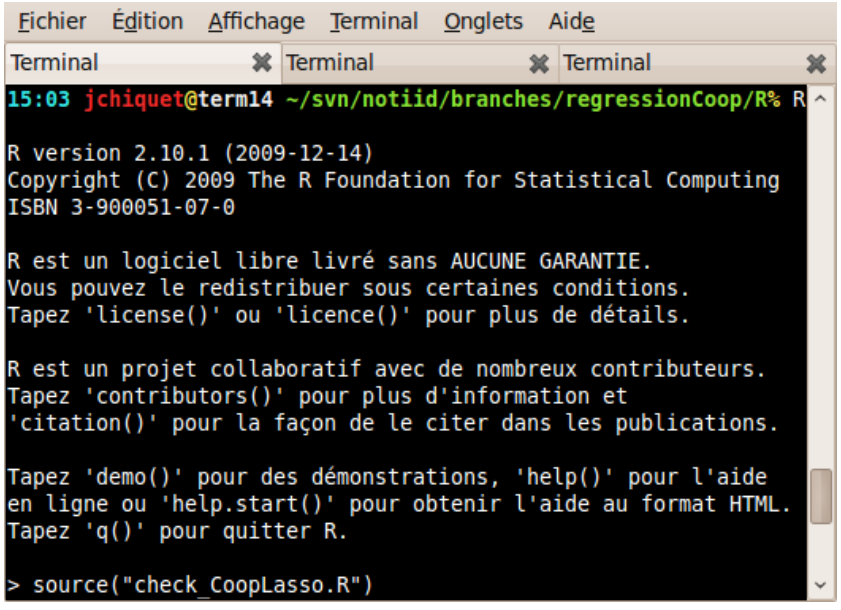
1. un éditeur de texte
2. un terminal avec R
3. des sorties graphiques



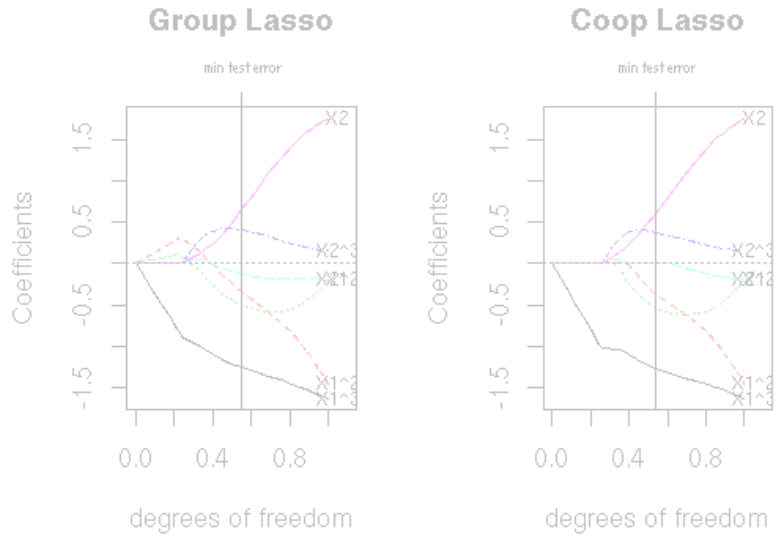
Environnement de travail sous Linux

Un bureau de développement avec R(solution sans R-studio





1. un éditeur de texte
2. un terminal avec R
3. des sorties graphiques



Environnement de travail sous Linux

Un bureau de développement avec R(solution sans R-studio

File Edit Options Buffers Tools Imenu-S ESS Help

```
rm(list=ls())
library(mvtnorm)
source("functions.R")
source("functions_group_l1.R")

set.seed(1002)

## données simulés (settings de Yuan et Lin - papier de 2006)
n <- 100
Sigma <- matrix(c(1,0.5,0.5,1),2,2)
X <- rmvnorm(n, Sigma)
y <- X[,1]^3 + X[,1]^2 - 2 * X[,1] + (1/3)*X[,2]^3 - 0*X[,2]^2 + (2/3)*X[,2] + rnorm(n,0,3)
```

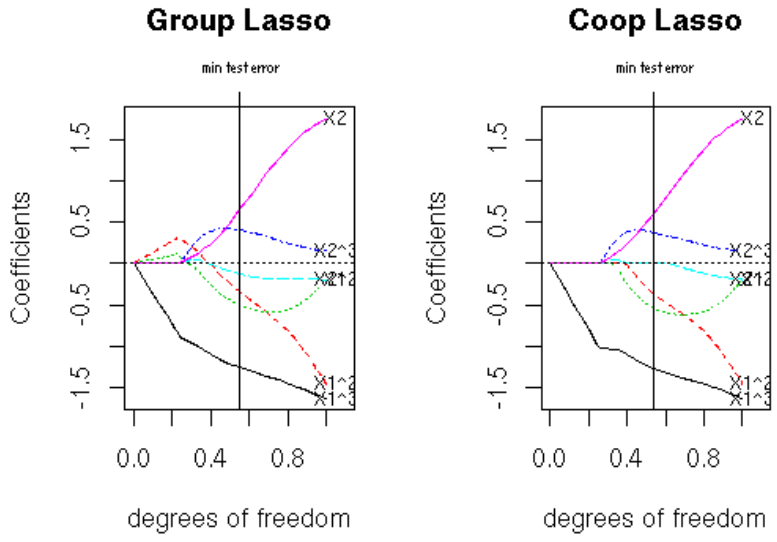
-U:--- check_CoopLasso.R Top (11,0) SVN-385 (ESS[S] [none])--15:50 0.47---

Fichier Édition Affichage Terminal Onglets Aide

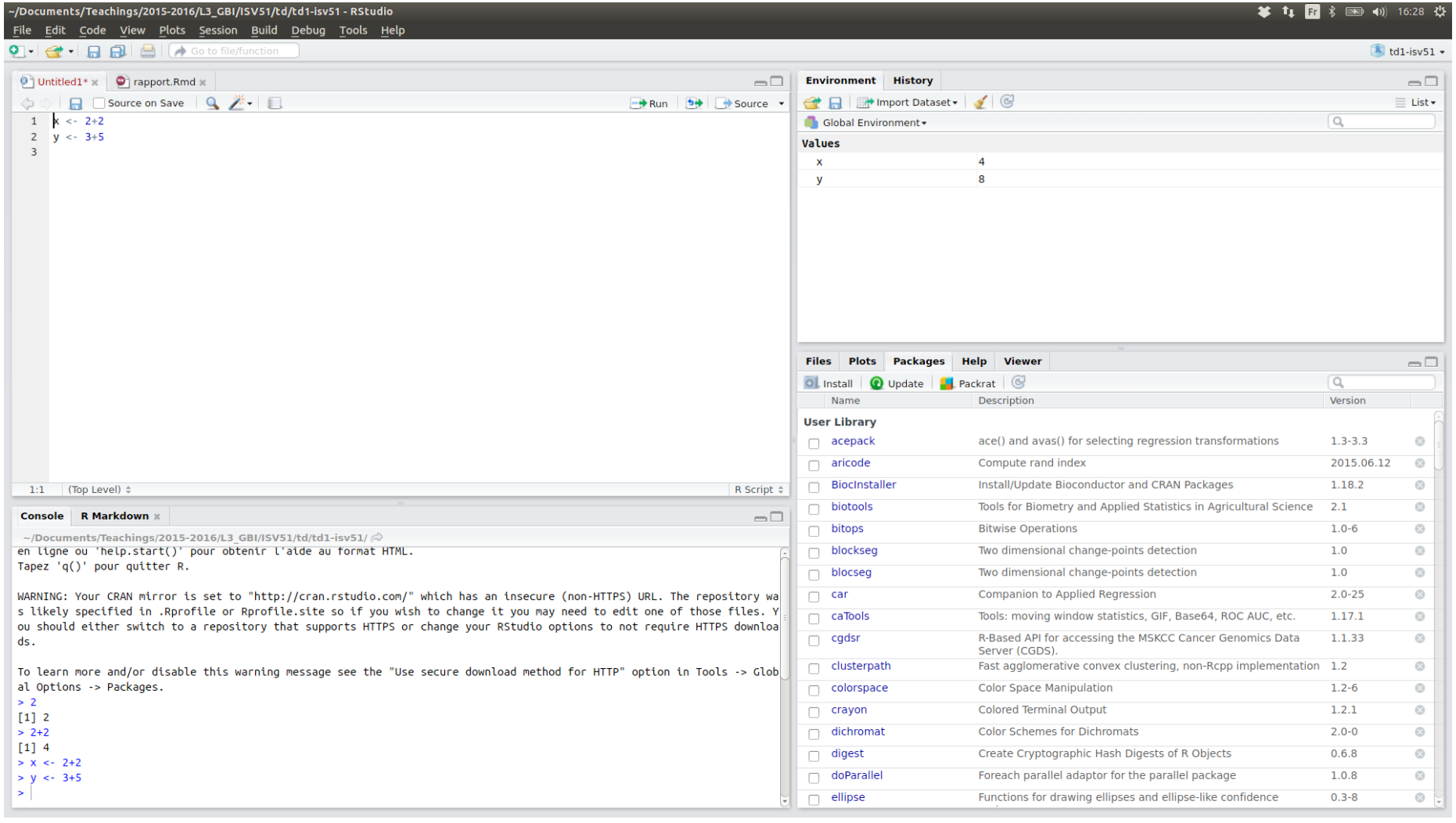
Terminal Terminal Terminal

```
15:03 jchiquet@term14 ~/svn/notiid/branches/regressionCoop/R% R

R version 2.10.1 (2009-12-14)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R est un logiciel libre livré sans AUCUNE GARANTIE.
Vous pouvez le redistribuer sous certaines conditions.
Tapez 'license()' ou 'licence()' pour plus de détails.

R est un projet collaboratif avec de nombreux contributeurs.
Tapez 'contributors()' pour plus d'information et
'citation()' pour la façon de le citer dans les publications.

Tapez 'demo()' pour des démonstrations, 'help()' pour l'aide
en ligne ou 'help.start()' pour obtenir l'aide au format HTML.
Tapez 'q()' pour quitter R.

> source("check_CoopLasso.R")
```

1. un éditeur de texte
2. un terminal avec R
3. des sorties graphiques



R-studio, environnement de travail intégré



Trouver de l'aide

Depuis R

- `help(str)` : lance l'aide associée à la commande `str`,
- `help.search("factorial")` : cherche les commandes contenant le mot-clé `factorial`,
- `help.start()` : lance l'aide HTML.

Sur le Web

En utilisant les media mentionné plus haut...

À tout moment

- la liste des commandes usuelles,
- le prof (pas infaillible mais rapide d'accès).

Plan

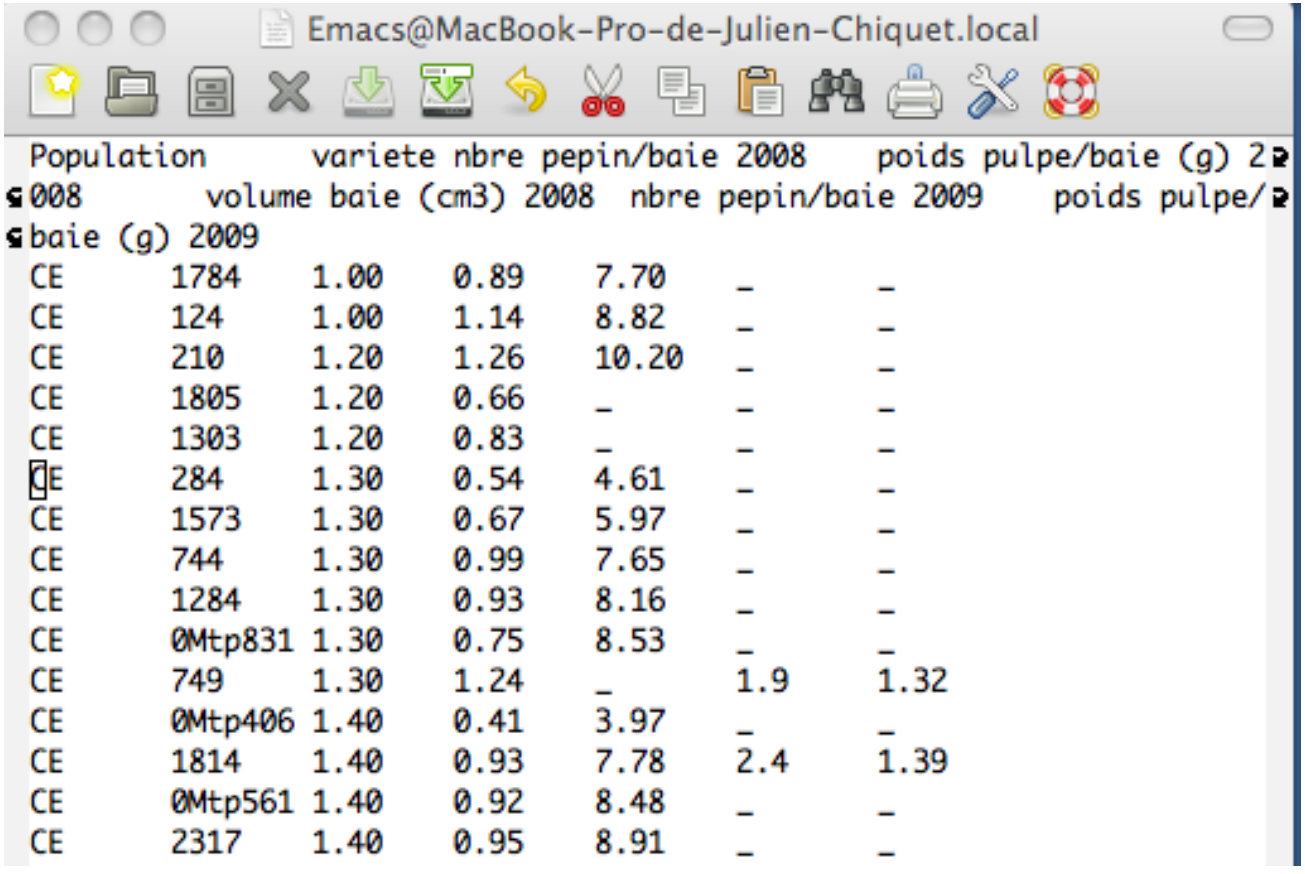
Avant de démarrer

Installation et premiers contacts

Une session exemple

Analyse élémentaire d'un jeu de données

Quelle tête ont les données ? On ouvre avec Emacs :



Population	variete	nbre pepin/baie 2008	poids pulpe/baie (g) 2008	volume baie (cm3) 2008	nbre pepin/baie 2009	poids pulpe/baie (g) 2009
CE	1784	1.00	0.89	7.70	_	_
CE	124	1.00	1.14	8.82	_	_
CE	210	1.20	1.26	10.20	_	_
CE	1805	1.20	0.66	_	_	_
CE	1303	1.20	0.83	_	_	_
CE	284	1.30	0.54	4.61	_	_
CE	1573	1.30	0.67	5.97	_	_
CE	744	1.30	0.99	7.65	_	_
CE	1284	1.30	0.93	8.16	_	_
CE	0Mtp831	1.30	0.75	8.53	_	_
CE	749	1.30	1.24	_	1.9	1.32
CE	0Mtp406	1.40	0.41	3.97	_	_
CE	1814	1.40	0.93	7.78	2.4	1.39
CE	0Mtp561	1.40	0.92	8.48	_	_
CE	2317	1.40	0.95	8.91	_	_

FIGURE: données baies de vignes 2008/2009

Je remplace tous les _ par du vide (R le comprendra mieux) !
(cette opération peut être faite directement dans R !)

Importation des données I

`getwd()` et `setwd()` gèrent le répertoire de travail :

```
setwd("~/Documents/Teachings/2015-2016/L3_GBI/ISV51/slides/0-introduction")
getwd()

## [1] "/home/jchiquet/Documents/Teachings/2015-2016/L3_GBI/ISV51/slides/0-introduc
```

Qu'est-ce qu'y se trouve dans ce répertoire ?

```
dir()

##  [1] "figures"
##  [2] "main.aux"
##  [3] "main.log"
##  [4] "main.nav"
##  [5] "main.out"
##  [6] "main.pdf"
##  [7] "main.Rnw"
##  [8] "main.Rnw~"
##  [9] "main.snm"
## [10] "main.tex"
## [11] "main.toc"
## [12] "mesures_baie_raisin_2008-2009.txt"
```

Importation des données II

Chargeons les données (délimitation par des tabulations)

```
donnees <- read.delim("mesures_baie_raisin_2008-2009.txt")
```

Qu'est-ce qui se trouve dorénavant dans mon itinéraire de recherche ?

```
ls()
```

```
##  [1] "C"          "cov.plot"   "donnees"    "invalid"    "matannot"
##  [6] "matgeno"    "matpheno2"  "mat.plot"   "my.heatmap" "nsnp"
## [11] "pheno"      "variable"
```

```
objects()
```

```
##  [1] "C"          "cov.plot"   "donnees"    "invalid"    "matannot"
##  [6] "matgeno"    "matpheno2"  "mat.plot"   "my.heatmap" "nsnp"
## [11] "pheno"      "variable"
```

Importation des données III

Quelle « tête » (au sens propre !) ont mes données ?

```
head(donnees)
```

```
##   Population variete nbre.pepin.baie.2008 poids.pulpe.baie..g..2008
## 1         CE    1784                   1.0                      0.89
## 2         CE     124                   1.0                      1.14
## 3         CE     210                   1.2                      1.26
## 4         CE    1805                   1.2                      0.66
## 5         CE    1303                   1.2                      0.83
## 6         CE     284                   1.3                      0.54
##   volume.baie..cm3..2008 nbre.pepin.baie.2009 poids.pulpe.baie..g..2009
## 1                   7.70                   NA                       NA
## 2                   8.82                   NA                       NA
## 3                  10.20                   NA                       NA
## 4                     NA                   NA                       NA
## 5                     NA                   NA                       NA
## 6                   4.61                   NA                       NA
```

Importation des données IV

Quelles sont ses attributs ?

```
str(donnees)

## 'data.frame': 245 obs. of  7 variables:
##  $ Population               : Factor w/ 3 levels "CE","CO","TE": 1 1 1 1 1 1 1 1
##  $ variete                  : Factor w/ 245 levels "0Mtp1004","0Mtp1005",..: 113
##  $ nbre.pepin.baie.2008     : num  1 1 1.2 1.2 1.2 1.3 1.3 1.3 1.3 1.3 ...
##  $ poids.pulpe.baie..g..2008: num  0.89 1.14 1.26 0.66 0.83 0.54 0.67 0.99 0.93
##  $ volume.baie..cm3..2008   : num  7.7 8.82 10.2 NA NA 4.61 5.97 7.65 8.16 8.53
##  $ nbre.pepin.baie.2009     : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ poids.pulpe.baie..g..2009: num  NA NA NA NA NA NA NA NA NA NA ...
```

Analyse statistique I

Le « nécessaire » résumé statistique :

```
summary(donnees)

##  Population     variete    nbre.pepin.baie.2008 poids.pulpe.baie..g..2008
##  CE:84      OMtp1004:  1   Min.   :0.000        Min.   :0.390
##  CO:89      OMtp1005:  1   1st Qu.:1.400        1st Qu.:0.820
##  TE:72      OMtp1033:  1   Median :1.800        Median :1.060
##             OMtp1068:  1   Mean   :1.858        Mean   :1.212
##             OMtp1072:  1   3rd Qu.:2.400        3rd Qu.:1.360
##             OMtp1073:  1   Max.   :3.200        Max.   :3.750
##             (Other) :239   NA's   :27           NA's   :28
##  volume.baie..cm3..2008 nbre.pepin.baie.2009 poids.pulpe.baie..g..2009
##  Min.   : 3.44          Min.   :1.000        Min.   :0.560
##  1st Qu.: 7.14          1st Qu.:1.500        1st Qu.:0.875
##  Median : 8.91          Median :2.000        Median :1.230
##  Mean   :10.47          Mean   :2.082        Mean   :1.513
##  3rd Qu.:11.90          3rd Qu.:2.600        3rd Qu.:1.607
##  Max.   :33.80          Max.   :3.600        Max.   :4.340
##  NA's   :44             NA's   :196          NA's   :203
```

Analyse statistique II

Et si je veux le nombre de baies moyen en 2008 pour chaque population ?

```
tapply(nbre.pepin.baie.2008,Population,mean,na.rm=TRUE)

##       CE       CO       TE
## 1.850649 2.034667 1.666667
```

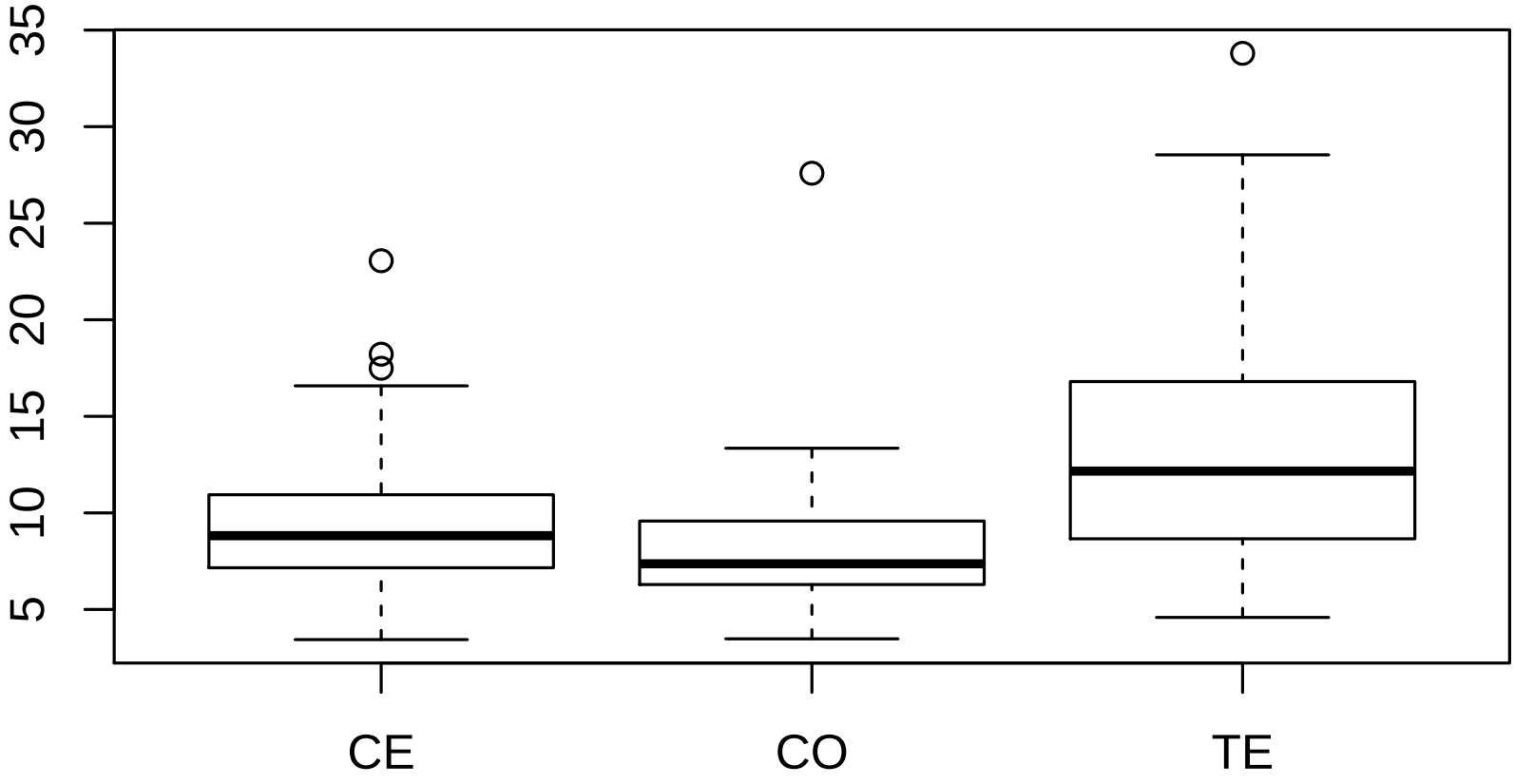
Et le volume moyen des baies ?

```
tapply(volume.baie..cm3..2008,Population,mean,na.rm=TRUE)

##        CE        CO        TE
##  9.667971  8.003000 14.132097
```

Ça a l'air disparate. Qu'est-ce que ça donne graphiquement ?

```
boxplot(volume.baie..cm3..2008 ~ Population)
```

Analyse statistique III

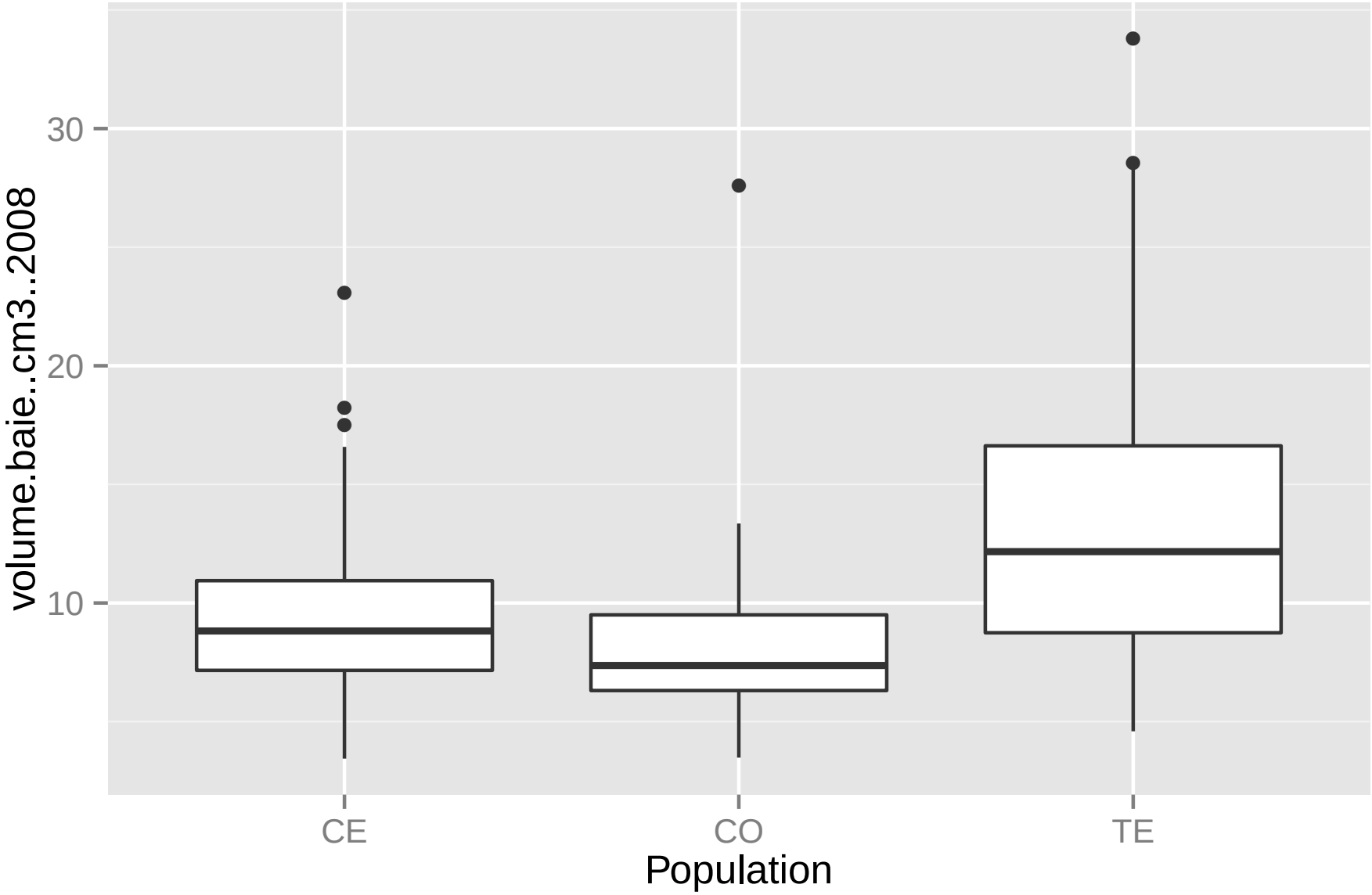


Analyse statistique IV

Voyons cela en plus joli/moderne...

```
library(ggplot2)
qplot(Population,volume.baie..cm3..2008, geom="boxplot")

## Warning in loop_apply(n, do.ply): Removed 44 rows containing non-finite
values (stat_boxplot).
```

Analyse statistique V



Analyse de la variance

Finalement, y-a-t-il un effet « population » pour le volume des baies ?

```
anova(lm(volume.baie..cm3..2008 ~ Population))

## Analysis of Variance Table
##
## Response: volume.baie..cm3..2008
##             Df Sum Sq Mean Sq F value    Pr(>F)
## Population   2 1301.9  650.94   29.45 6.357e-12 ***
## Residuals  198 4376.5   22.10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

etc. *beaucoup* d'autres choses sont possibles...

Un autre exemple : SNP et HIV I

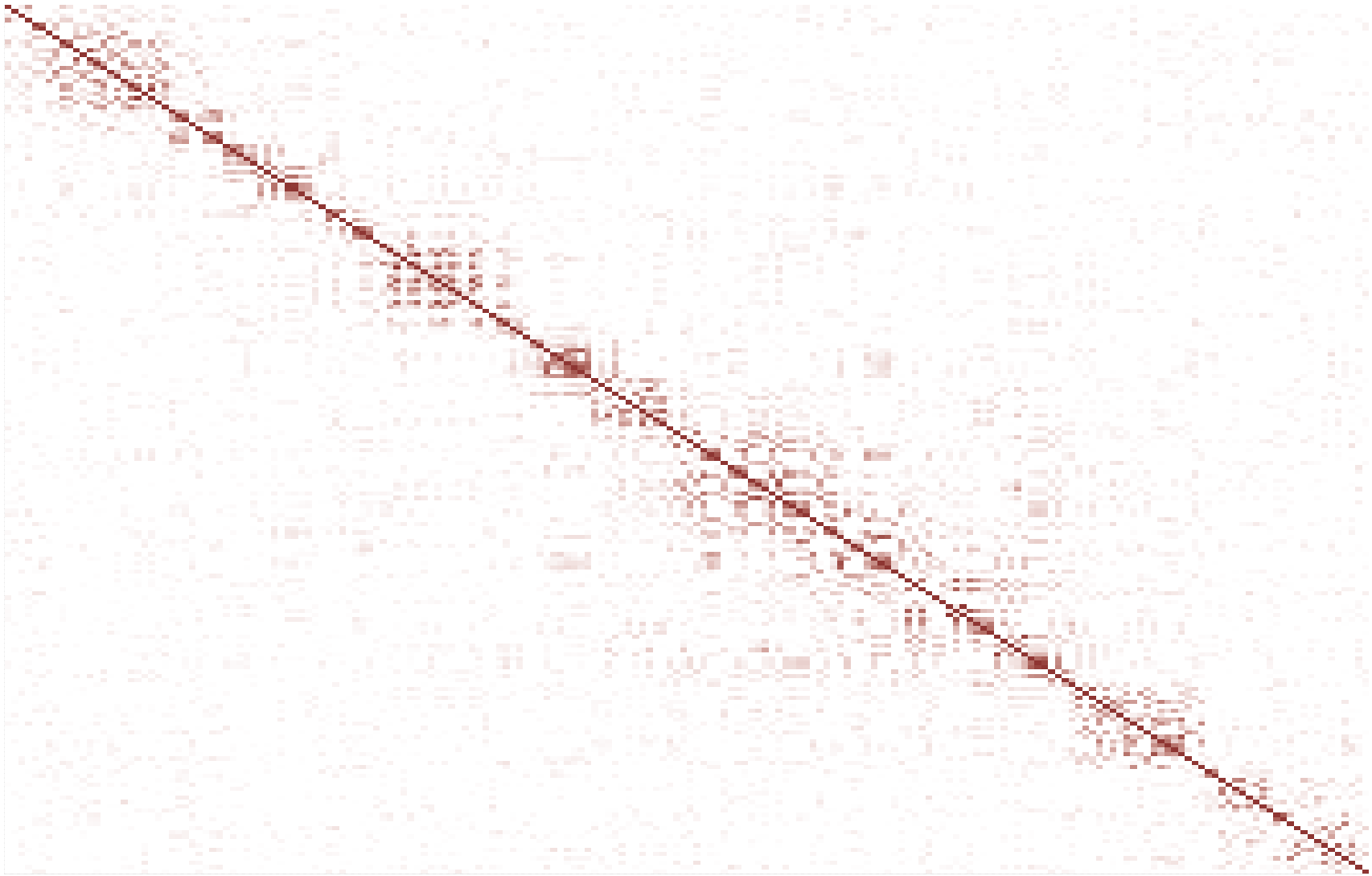
```
setwd("~/svn/hdr/code/examples/example_markers_selection")
source("../plot_func.R") #a couple of personal plotting function
library(Matrix) # better matrix manipulation
```

```
## LOAD PREPROCESSED SNP DATA
load("data/geno.Rdata")      # SNP data
load("data/matannot.RData") # annotation
load("data/pheno.Rdata")     # phenotype data
nsnp <- 200 # only consider the first nsnp SNP for plotting
```

Un autre exemple : SNP et HIV II

```
## Linkage desuiquilibrium / correlation between SNP
C <- cov2cor(var(t(matgeno[1:nsnp, ]), na.rm=TRUE))
mat.plot(C)

## Using row as id variables
```

Un autre exemple : SNP et HIV III



Un autre exemple : SNP et HIV IV

```
## Phenotype distributions
pheno <- melt(matpheno2[, 4:5])

## No id variables; using all as measure variables

variable <- rep("HIV-RNA level", nrow(pheno))
variable[pheno$variable == "DNAinc"] <- "HIV-DNA level"
pheno$variable <- factor(variable)
print(ggplot(pheno, aes(x=variable,y=value,fill=variable)) + geom_violin() + geom_j
      theme(legend.position="none",
            text=element_text(size=24),
            axis.title=element_blank()) +xlab("") +ylab(""))

## Warning in loop_apply(n, do.ply): Removed 34 rows containing non-finite
values (stat_ydensity).
## Warning in loop_apply(n, do.ply): Removed 34 rows containing missing values
(geom_point).
```

Un autre exemple : SNP et HIV V