Rapport - Données manquantes

Nom et Prénom des étudiants du groupe :

```
- Nom : Prénom : REAL Philippe
- Nom : Prénom : GUYONVARCH Alexis
```

1. Introduction

Présentation des données et objectifs de l'étude https://www.kaggle.com/uciml/pima-indians-diabetes-database (https://www.kaggle.com/uciml/pima-indians-diabetes-database)

1.1 Contexte

Le jeu de données provient de l'Institut national du diabete, des maladies digestives et rénales. Il rend possible la prédiction de la pathologie, en l'espèce le diabète, pour le patient à partir d'analyses inclues dans le jeu de données. Dans l'extraction, les patients sont des femmes et issues de la commmunauté des indiens Pima.

1.2 Description des colonnes

- npreg : Number of times pregnant - Nombre de grossesses
- glu : GlucosePlasma 2 hours in an oral glucose tolerance test - Concentration de glucose dans le sang
- bp : BloodPressureDiastolic - Pression sanguine (mm Hg)
- skin : SkinThicknessTriceps - Epaisseur de la peau (mm)
- Insuline : Insulin 2- Hour serum insuline - Taux d'insuline présent dans le sang (mu U/ml)
- bmi : BMIBody mass index - Indice de masse corporelle (poids en kg/(taille en m)élevée au carré)
- ped : Diabetes Pedigree Function Diabetes pedigree function - Antécédent de diabète sucrégénétique
- age : Age (years)
- type : Outcome Class variable - Variable réponse binaire (0/1)

1.3 Objectifs

Objectif : prédire si l'individu a ou non le diabète. Préalablement, il s'agi de compléter les données manquante par imputation.

1.4 Chargement des donnés - Résumé

	vars <int>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
npreg	1	768	3.8450521	3.3695781	3.0000	3.4610390	2.9652000	0.000	17.00
glu	2	763	121.6867628	30.5356411	117.0000	119.6579378	29.6520000	44.000	199.00
bp	3	733	72.4051842	12.3821582	72.0000	72.2896082	11.8608000	24.000	122.00
skin	4	541	29.1534196	10.4769824	29.0000	28.8752887	10.3782000	7.000	99.00
Insuline	5	394	155.5482234	118.7758552	125.0000	135.3196203	81.5430000	14.000	846.00
bmi	6	757	32.4574637	6.9249883	32.3000	32.1105437	6.8199600	18.200	67.10
ped	7	768	0.4718763	0.3313286	0.3725	0.4215536	0.2483355	0.078	2.42
age	8	768	33.2408854	11.7602315	29.0000	31.5438312	10.3782000	21.000	81.00
type	9	768	0.3489583	0.4769514	0.0000	0.3116883	0.0000000	0.000	1.00

9 rows | 1-10 of 14 columns

```
## [1] "Pourcentage de données manquantes par variable"
```

```
##    npreg      glu       bp     skin Insuline      bmi      ped      age
##      0.0      0.7      4.6     29.6     48.7      1.4      0.0      0.0
##     type
##      0.0
```

- Traitement des données

Suppression de la 1ère colonne comprenant les identifants. La colonne insuline est, dans l'immédiat, conservée en dépit de la part importante de données manquantes : 48.7 %.

	vars <int>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
npreg	1	768	3.8450521	3.3695781	3.0000	3.4610390	2.9652000	0.000	17.00
glu	2	763	121.6867628	30.5356411	117.0000	119.6579378	29.6520000	44.000	199.00
bp	3	733	72.4051842	12.3821582	72.0000	72.2896082	11.8608000	24.000	122.00
skin	4	541	29.1534196	10.4769824	29.0000	28.8752887	10.3782000	7.000	99.00
Insuline	5	394	155.5482234	118.7758552	125.0000	135.3196203	81.5430000	14.000	846.00
bmi	6	757	32.4574637	6.9249883	32.3000	32.1105437	6.8199600	18.200	67.10
ped	7	768	0.4718763	0.3313286	0.3725	0.4215536	0.2483355	0.078	2.42
age	8	768	33.2408854	11.7602315	29.0000	31.5438312	10.3782000	21.000	81.00

8 rows | 1-10 of 14 columns

- Les différentes catégories de données manquantes sont "taguées" en vue d'analyses ultérieures. La description deux à deux des patterns est effectuée avec avec la fonction md.pairs() du package mice qui nous permet déjà de mettre en lumière les 10 combinaisons du jeu de données parmi les 28 possibles.

```
## [1] "Combinaisons 2 à 2 des données manquantes"
```

```
##          glu bp skin Insuline bmi
## glu        5  0    0        4   0
## bp         0 35   33       35   7
## skin       0 33  227      227   9
## Insuline   4 35  227      374  10
## bmi        0  7    9       10  11
```

	npreg <int>	glu <int>	bp <int>	skin <int>	Insuline <int>	bmi <dbl>	ped <dbl>	age <int>	type <lgl>
1	6	148	72	35	*NA*	33.6	0.627	50	TRUE
2	1	85	66	29	*NA*	26.6	0.351	31	FALSE
3	8	183	64	*NA*	*NA*	23.3	0.672	32	TRUE
4	1	89	66	23	94	28.1	0.167	21	FALSE
5	0	137	40	35	168	43.1	2.288	33	TRUE
6	5	116	74	*NA*	*NA*	25.6	0.201	30	FALSE

6 rows | 1-10 of 11 columns

```
## [1] "Création des catégories de données manquantes"
```

```
##
##           No-Missing        Insuline.only              bmi.only
##                  392                  140                    1
##             glu.only         bmi.Insuline          bp.Insuline
##                    1                    1                    2
##     bp.skin.Insuline bp.skin.Insuline.bmi         glu.Insuline
##                   26                    7                    4
##        skin.Insuline    skin.bmi.Insuline
##                  192                    2
```

2. Exploration des données

2.1 Classification des données manqunates: MCAR/MNAR

Rapide classification de données manquantes :

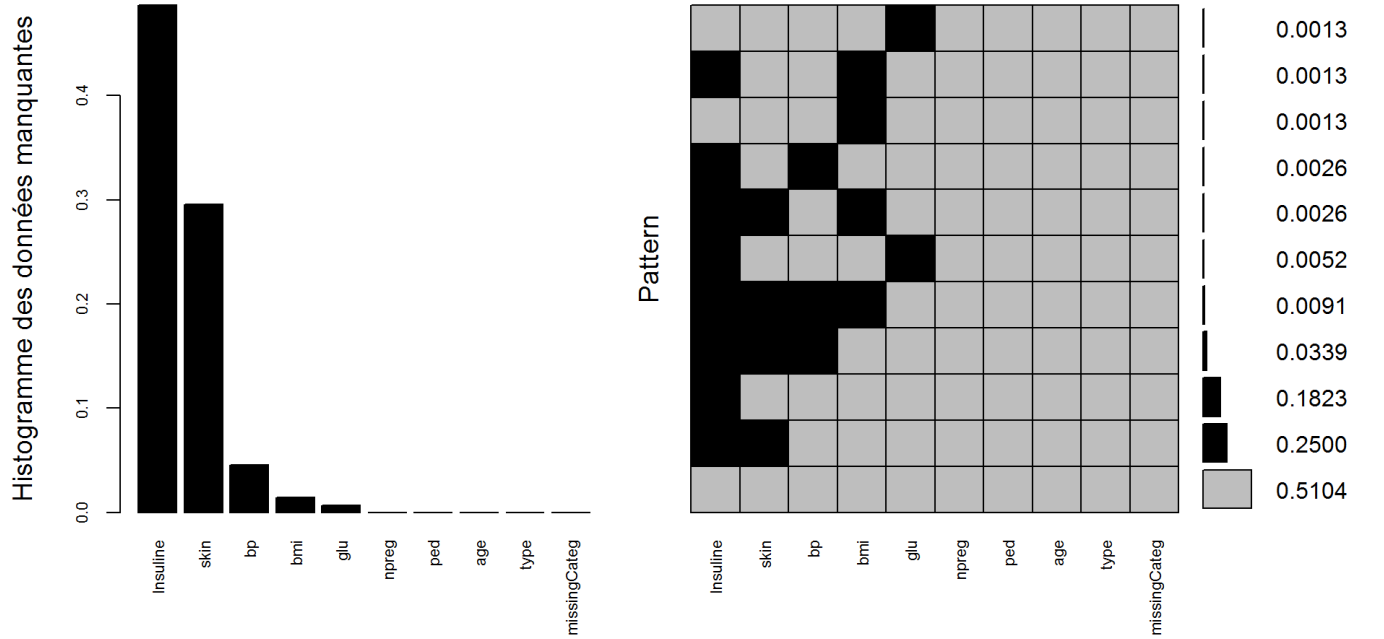
- MCAR (missing completely at random): Donnée manquante de façon complètement aléatoire => la probabilité d'absence est la même pour toutes les observations. et ne dépend donc que de paramètres exogènes indépendants de la variable.
- MAR (missing at random) : Survient lorsque les données ne manquent pas de façon complètement aléatoire; la probabilité d'absence est liée à une ou plusieurs autres variables observées.
- MNAR (missing not at random): La probabilité d'absence dépend de la variable en question.Les données MNAR induisent une perte de précision mais aussi un biais qui nécessite le recours à une analyse de sensibilité.

Recours aux librairies MICE, MASS, VIM pour visualiser les patterns du jeu de données grâce aux fonctions fluxplot() ou aggr_plot().

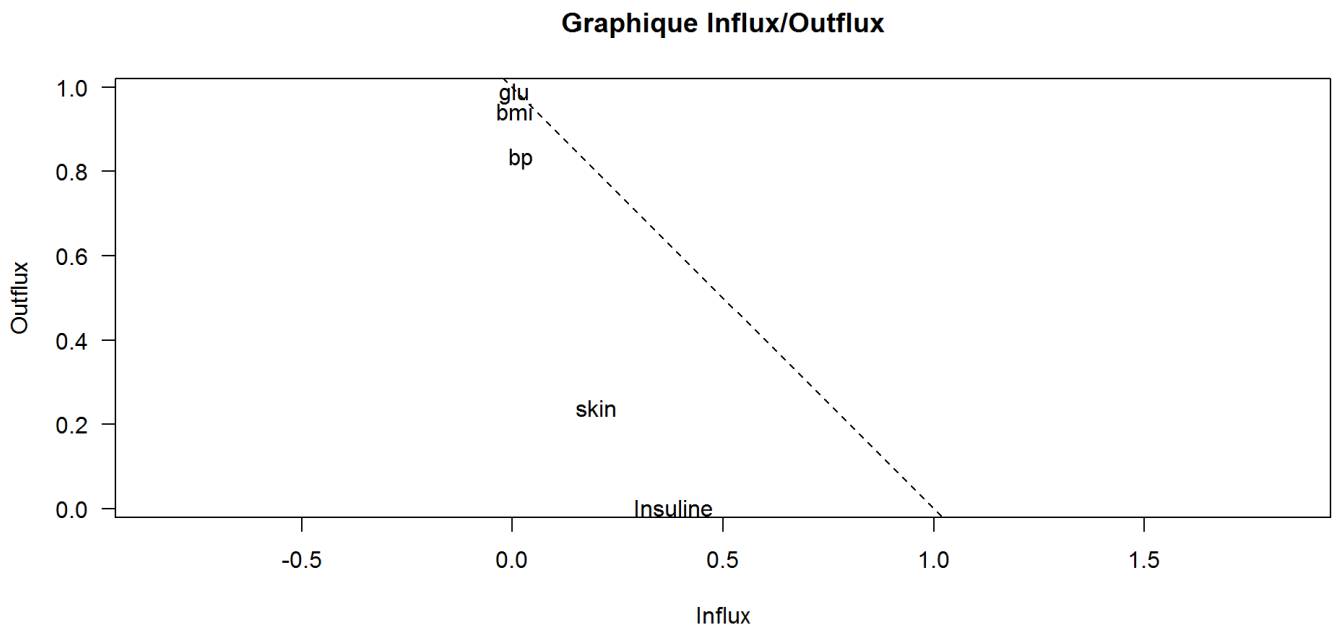
Le pourcentage de données manquantes est élevé, les informations sont exhaustives pour seulement 51% des individus, ce qui jusitifie le recours à l'imputation multiple. Le graphique des combinaisons confirme, ce qui avait déjà été mis en lumière ci-devant, à savoir la prédominance de plusieurs combinaisons : * Insuline + Skin * Insuline + Skin + BP A elle seule, la variable Insuline, quand elle est manquante, regroupe 8 patterns. La variable skin concerne 4 patterns.

Au final, il ressort que le mécanisme des données manquantes, qui concernent 5 variables du dataframe "Pima", est non-monotone, ce qui justifiera ultérieurement le recours à l'imputation multiple (joint modeling, fully conditionnal specification, ACP).

La valeur d' "influx" de la variable Insuline est plus élevée que la valeur d'"influx" de la variable Skin en dépit d'une proportion plus importante de données manquantes. Cela suggère une connection plus forte aux variables observées. Une valeur d'"outflux" très faible nous indique que la variable "Insuline", ainsi que la variable "Skin", quoique dans une moindre mesure, seront potentiellement moins utiles à l'imputation des autres variables.




```
##
##  Variables sorted by number of missings:
##      Variable       Count
##      Insuline 0.486979167
##          skin 0.295572917
##            bp 0.045572917
##           bmi 0.014322917
##           glu 0.006510417
##         npreg 0.000000000
##           ped 0.000000000
##           age 0.000000000
##          type 0.000000000
##  missingCateg 0.000000000
```



	influx <dbl>	outflux <dbl>
glu	0.005018821	0.986196319
bp	0.020388959	0.831288344
skin	0.200439147	0.239263804
Insuline	0.382685069	0.003067485
bmi	0.005646173	0.943251534

5 rows

```
## [1] "Nombre de patterns si Insuline manquant"
```

```
## [1] 8
```

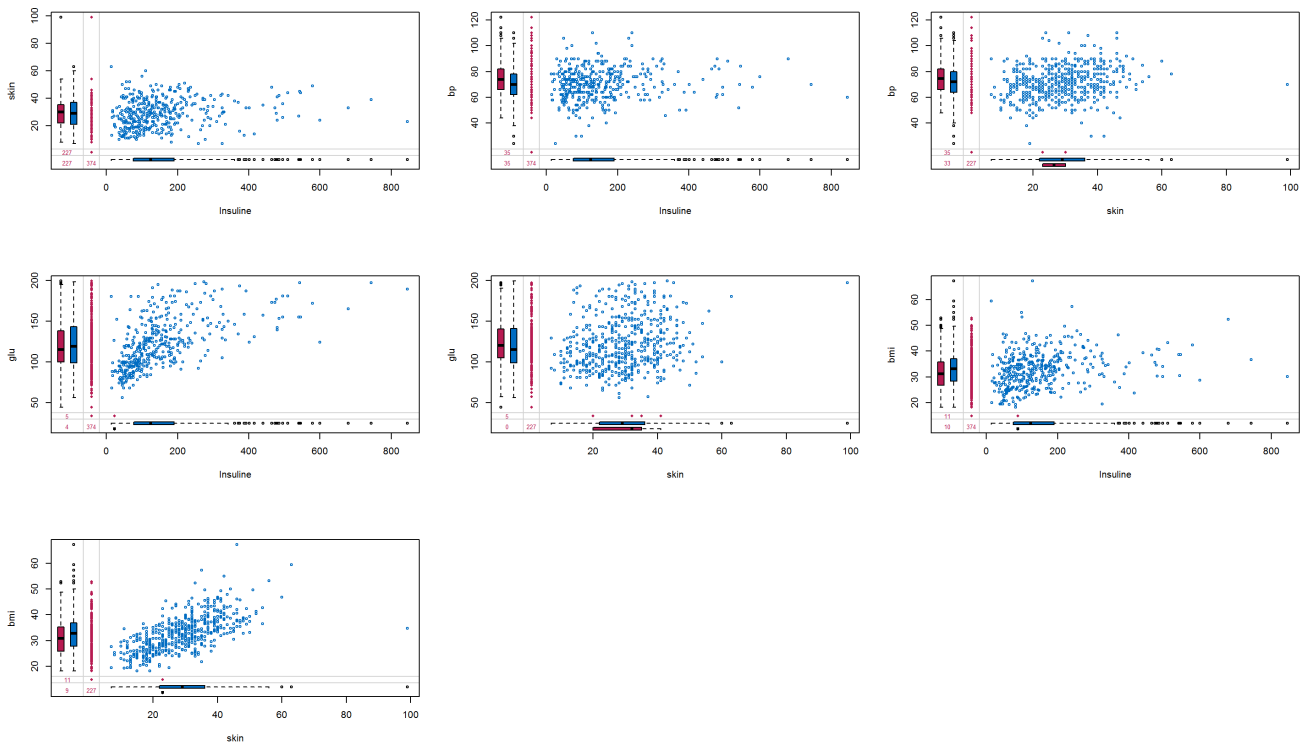
```
## [1] "Nombre de patterns si skin manquant"
```

```
## [1] 4
```

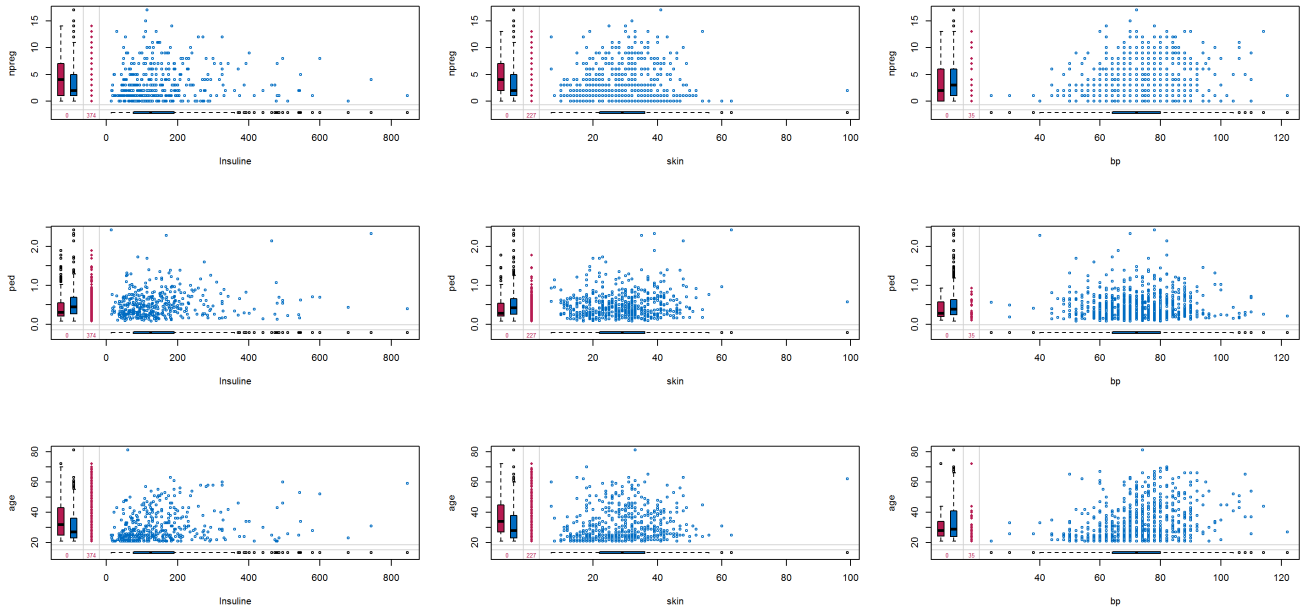
- L'hypothèse MCAR semble infirmée au regard des distributions des variables Insuline et Skin comparées aux variables complètes, npreg, age et ped. A ce stade, l'hypothèse d'ignorabilité du mécanisme peut par ailleurs être maintenue. L'analyse de sensibilité nous permettra de la valider. Nous pouvons enfin admettre que les mécanismes des variables "Insuline" et "skin" sont proches.

graphiques N°1 - Mécanismes des variables "Insuline" et "skin" (§2.1)

- Distributions marginales des variables Insuline, Skin et bp avec les autres variables incomplètes



- Distributions marginales des mêmes variables avec les variables complètes



3. Imputations

3.1 Imputations avec MICE

3.1.1 Imputations simples

Dans un premier temps on traite les valeurs manquantes par imputation simple avec le package MICE : * par le biais de la méthode PMM (predictive mean matching), * puis au moyen d'une régression linéaire - non bayésienne - stochastique.

A priori, ces imputations ne seront pas conservées en raison de la sensibilité à la spécification du modèle (pour la méthode paramétrique) et du biais souvent généré par la méthode PMM (semi paramétrique).

```
#PMM
imp.si.pmm <- mice(data, m=1, seed = 111119, print = F)

#régression stochastique avec bootstrap
imp.si.norm <- mice(data, method = "norm.nob", m = 1, maxit = 1, print = F)
```

3.1.2 Imputations multiples

https://datascienceplus.com/imputing-missing-data-with-r-mice-package/ (https://datascienceplus.com/imputing-missing-data-with-r-mice-package/)

Nous recourons, au cours d'une première étape, à l'imputation multiple au moyen des méthodes Joint Modeling puis Fully Conditionnal Specification.

```
#JM
imp.mi.jm <- mice(data, m=5, method='norm',seed=111119, print=F)
#FCS
imp.mi.fcs <- mice(data, m=5, seed=111119, print = F)
```

3.1.3 Analyses des distributions

- Analyse des distributions des variables observées et imputées pour les imputations simples

Les graphiques montrent que les imputations par régression linéaire stochastique comportent des valeurs aberrantes, en l'espèce, des valeurs négatives pour l'épaisseur de peau et le taux d'insuline. Les distributions des valeurs observées et imputées sont proches, ce qui n'infirme ni ne confirme le mécanisme MAR.

graphiques N°2 - Analyses des distributions pour les Imputations simples

Pour les graphiques cf. annexe N°1

- Analyse des distributions des variables observées et imputées pour les imputations multiples A nouveau, certaines imputations ne sont pas plausibles : avec le modèle joint les graphiques montrent des valeurs aberrantes pour l'épaisseur de peau et le taux d'insuline. S'agissant de la méthode FCS, les distributions des variables glu et bmi présentent des profils très divergents pour chacune des imputations, en raison du nombre peu élevé de données manquantes.

graphiques N°3 - Analyses des distributions pour les Imputations multiples

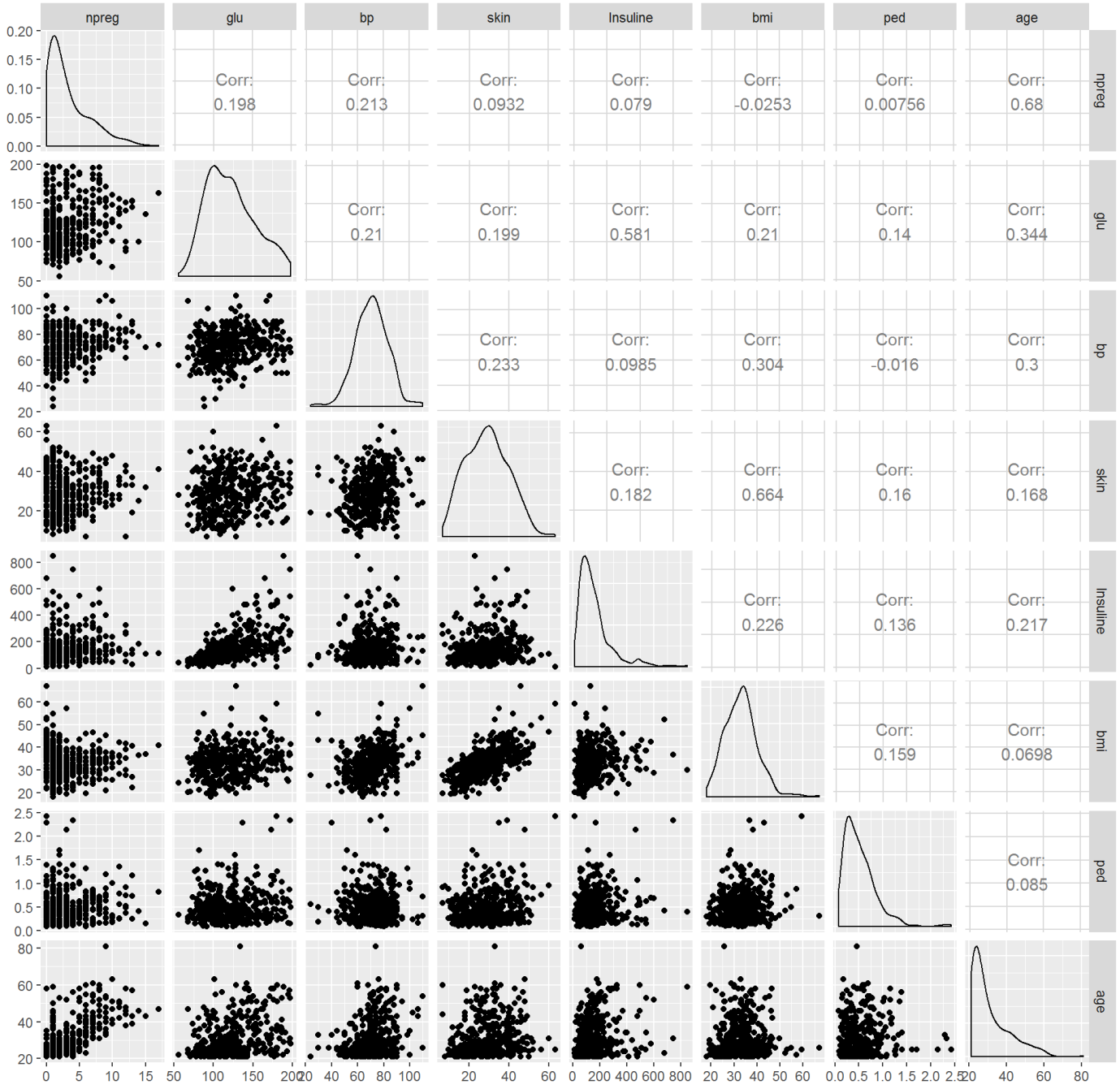
Pour les graphiques cf. annexe N°1

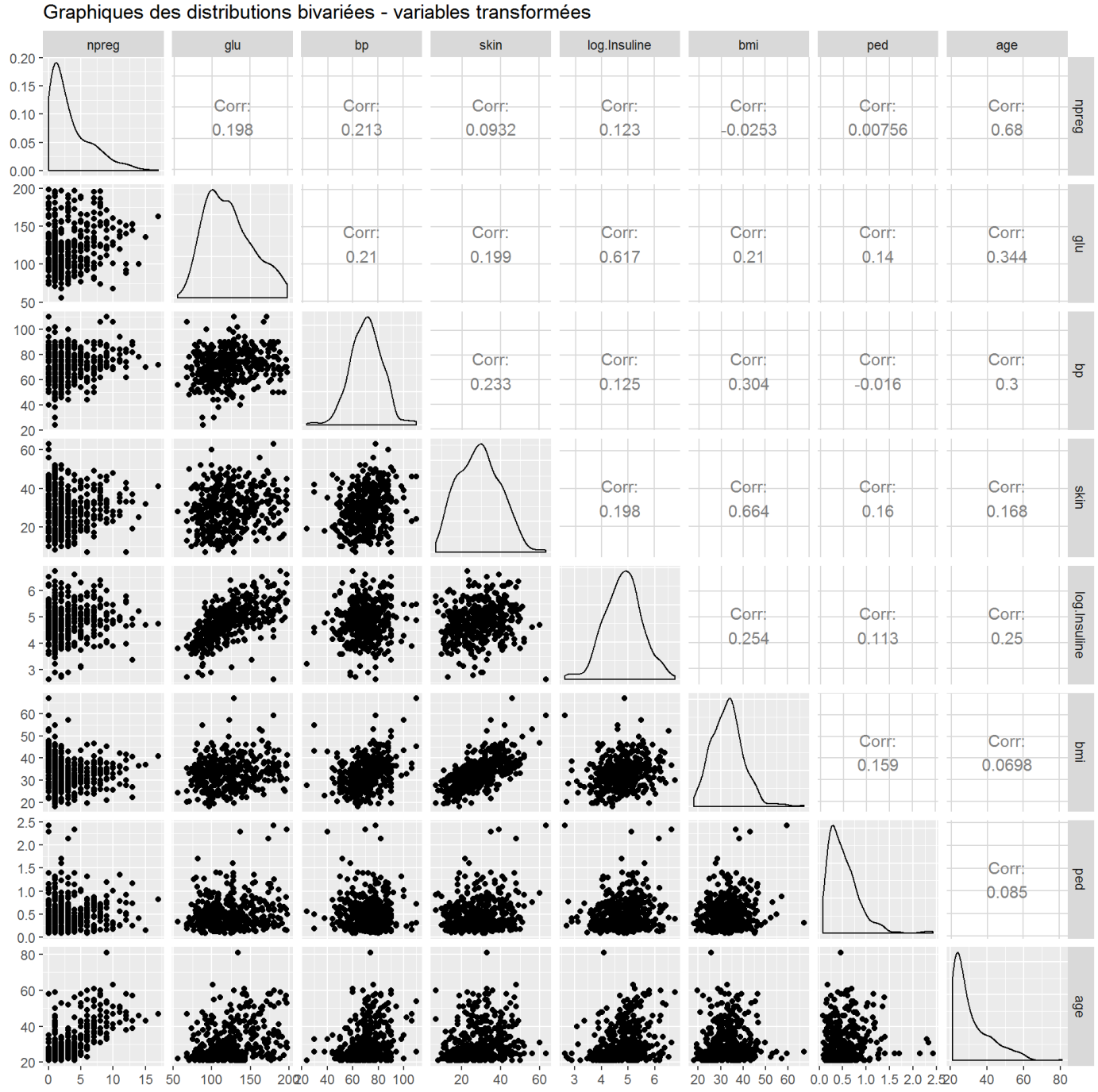
3.2 Imputation multiple par les méthodes d'analyse factorielle avec MIPCA

- Transformation des données

Préalablement à l'étape de réduction des dimensions, nous procédons à un essai de transformation des variables visant à renforcer la linéarité des liens entre elles. Finalement, après plusieurs tentatives (log, logistic et racine carrée), seule une transformation logarithmique de la variable insuline est conservée.

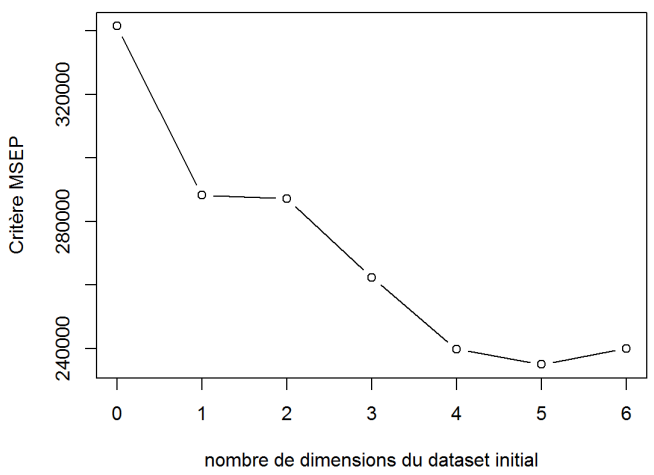
Graphiques des distributions bivariées

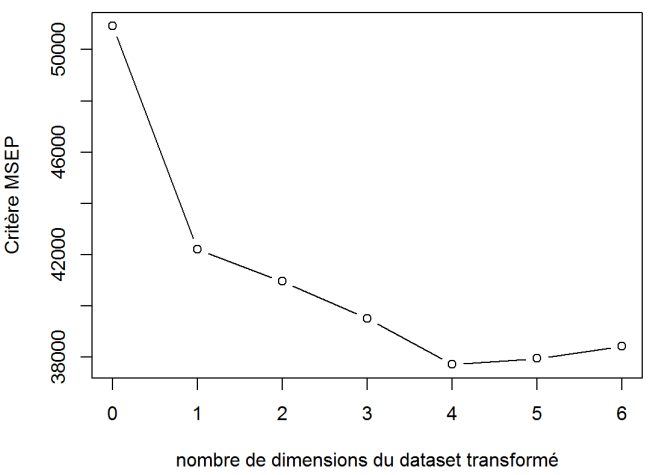




- Détermination de la dimension de l'espace de projection

Avec la méthode de validation croisée "kfold", le nombre de dimensions optimale est de 5 ou 4 selon le critère "MSEP".





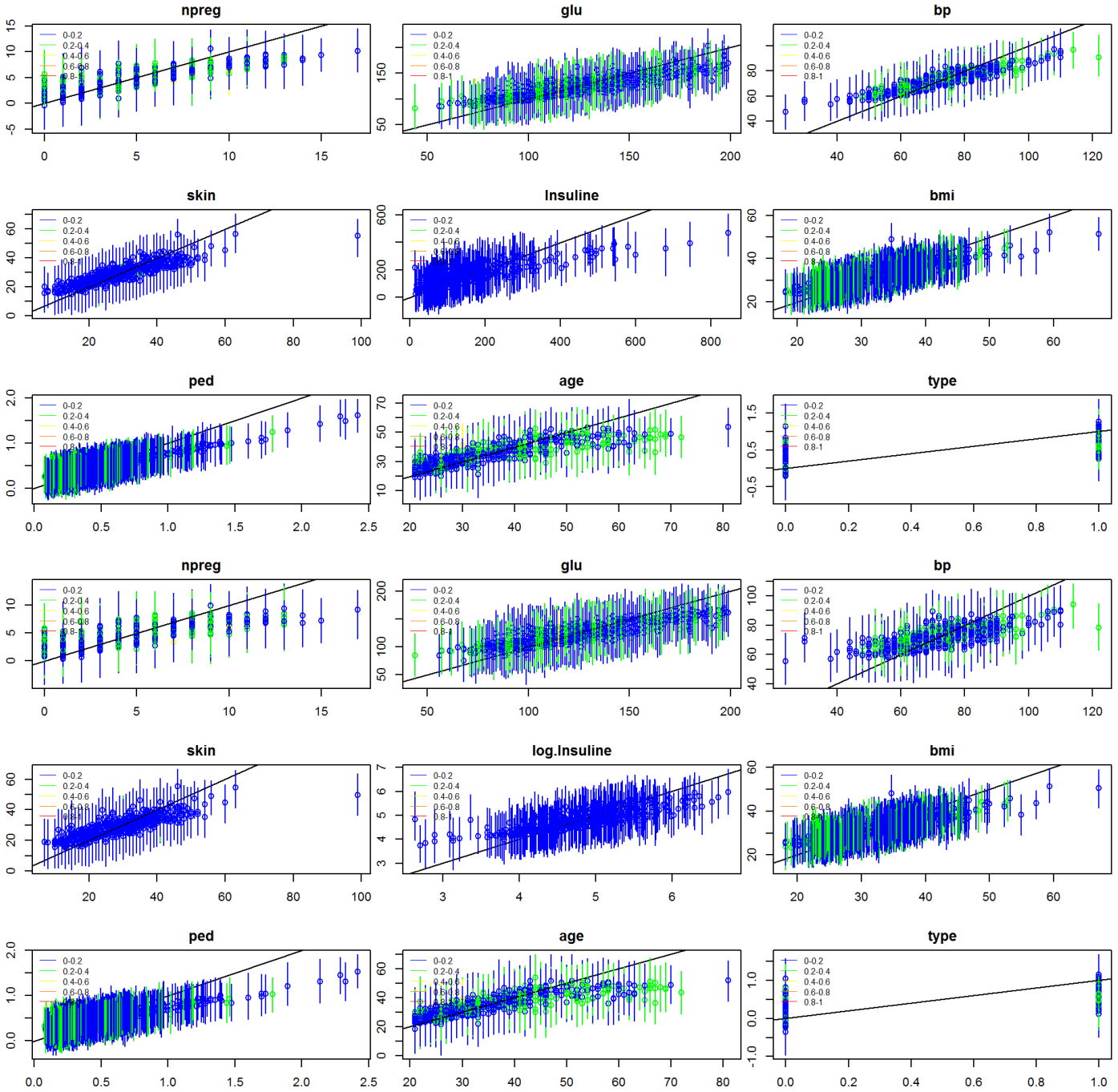
```
## [1] "Le nombre de dimensions retenu pour le jeu de données initial est de :5"
```

```
## [1] "Le nombre de dimensions retenu pour le jeu de données transformé est de :4"
```

3.2.1 Imputation multiple - méthode bayésienne

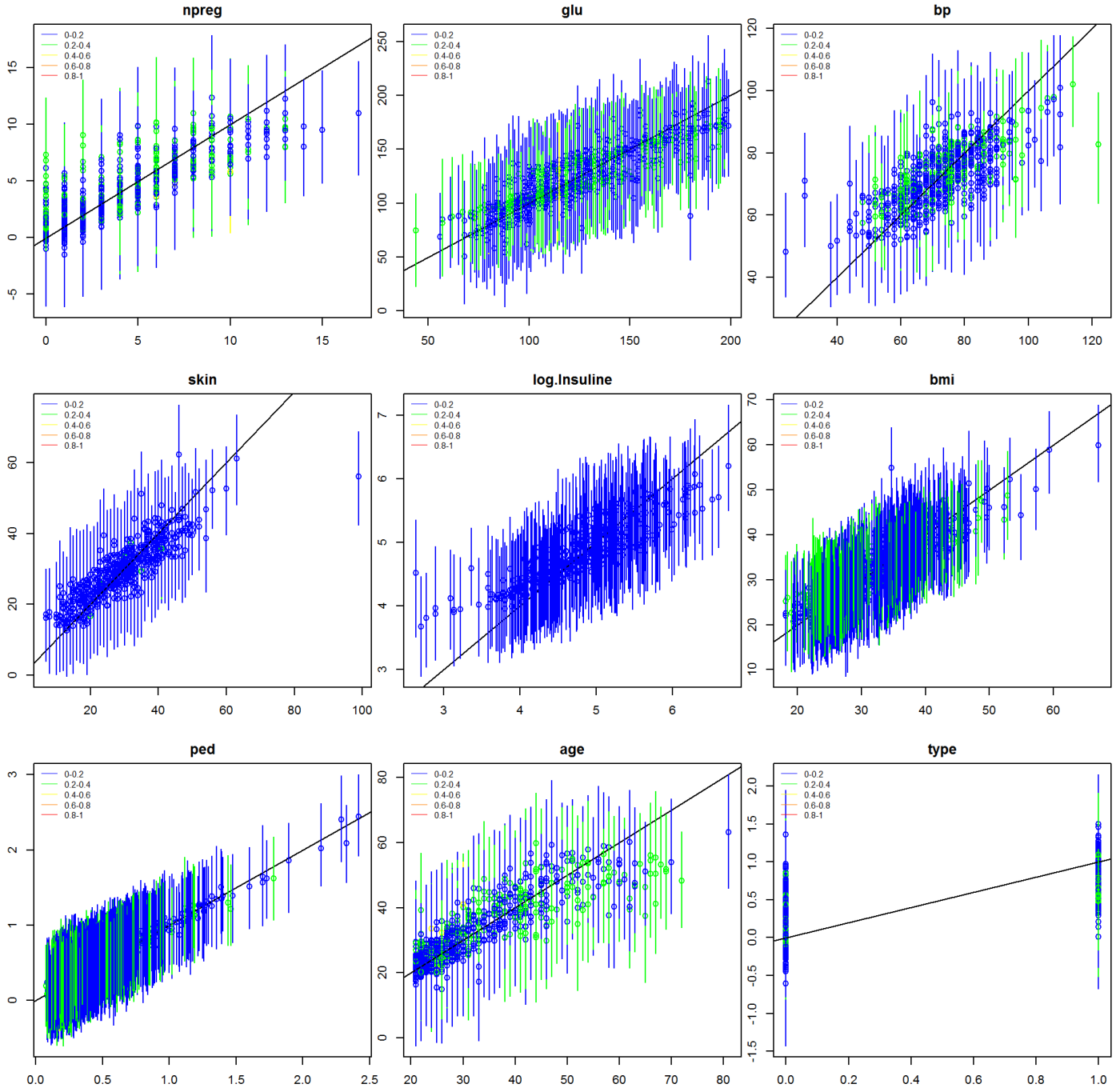
*Diagnostics pour l'imputation multiple - méthode bayésienne

Visuellement, l'imputation de la variable insuline semble de meilleure qualité, une fois celle-ci ayant fait l'objet d'une transformation en logarithmique.



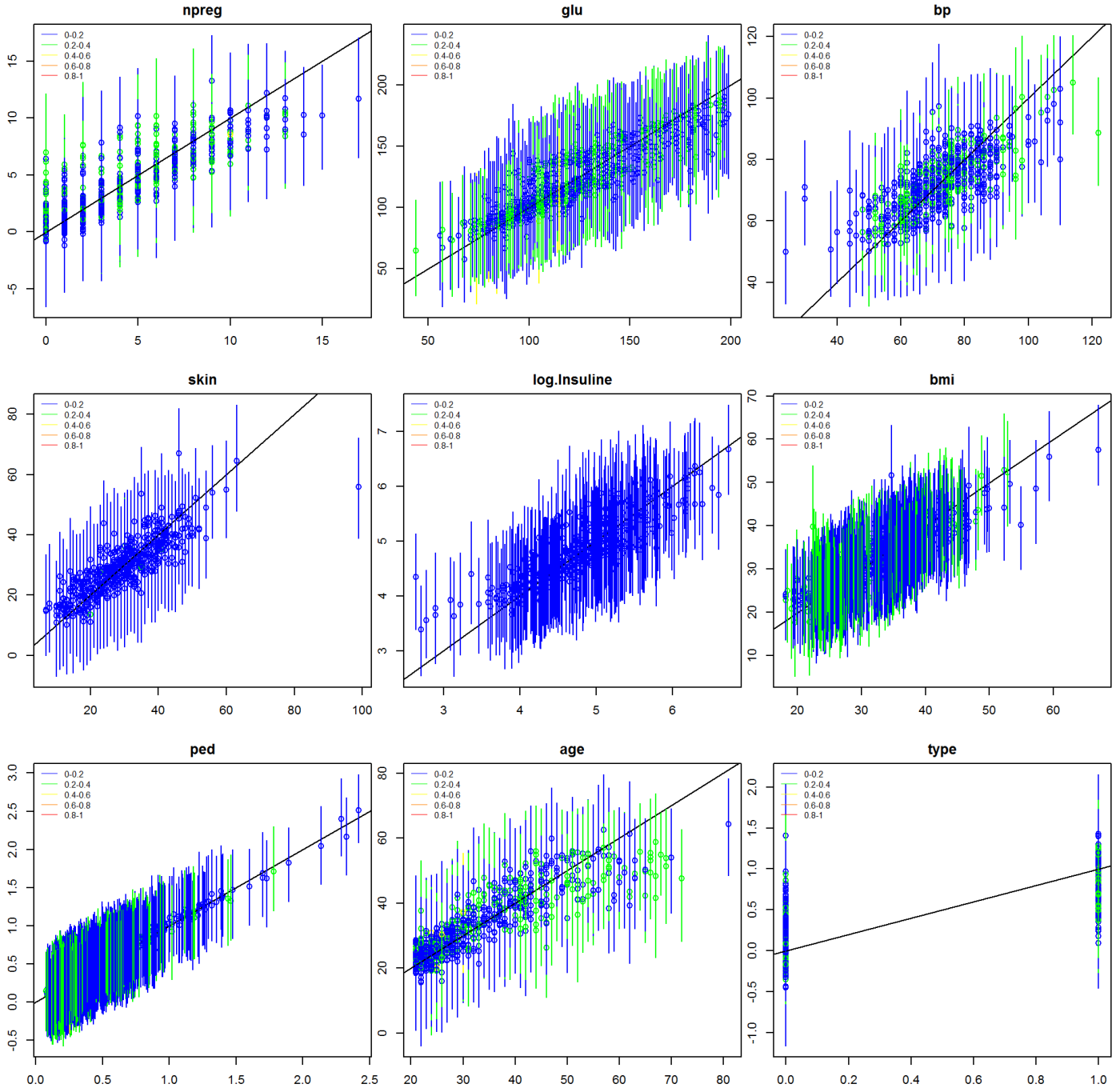
3.2.2 Multiple Imputation avec multiple imputation par boostrap

Visuellement, le modèle semble plus en adéquation avec l'imputation multiple par bootstrap.



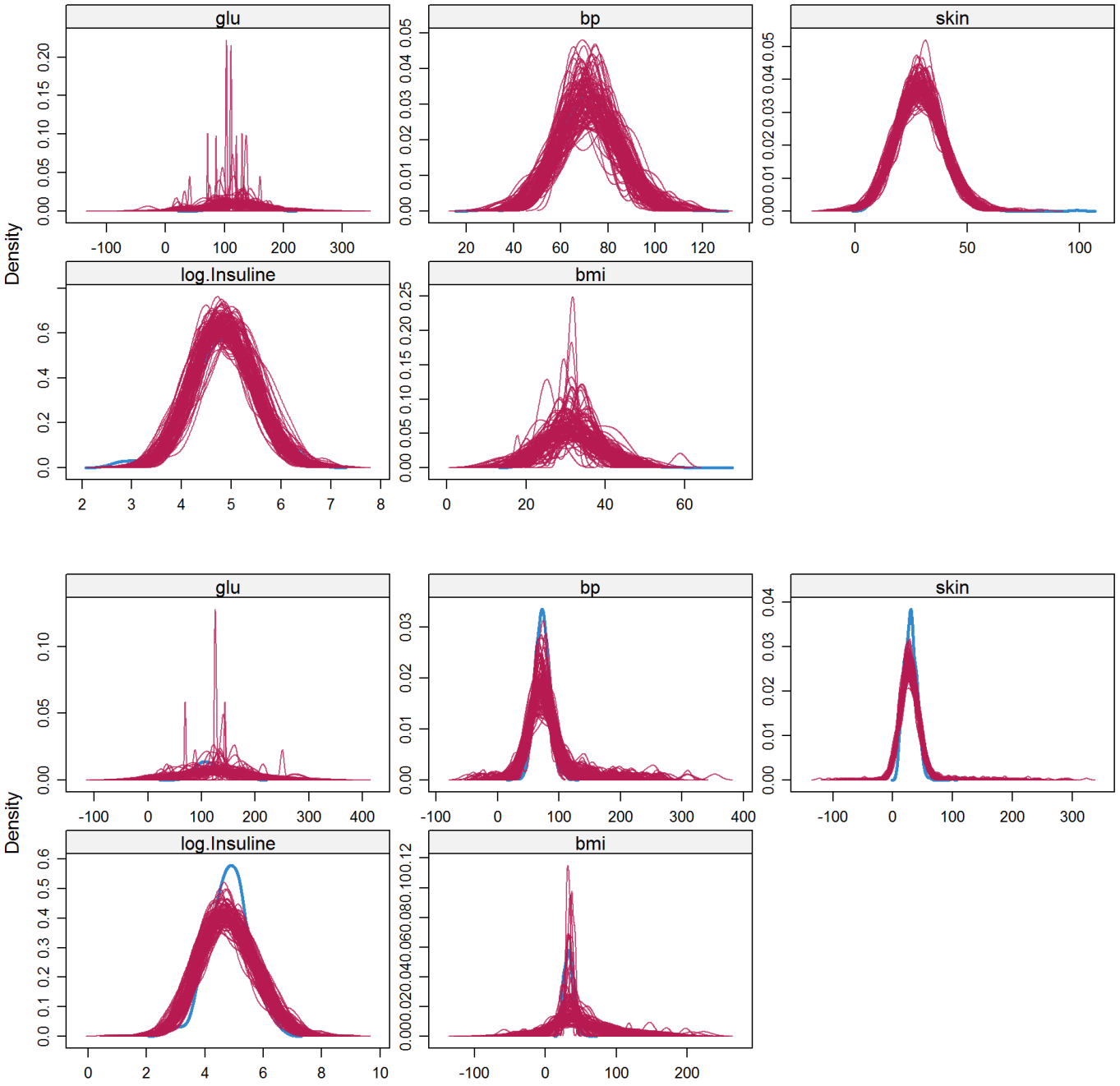
3.2.3 Multiple Imputation EM et bootstrap

Cette dernière imputation semble la plus appropriée.



* Analyse des distibutions

L'analyse des distributions des résultats est mise en peuvre avec les méthodes du package "MICE". L'analyse requiert une conversion préalable des objets avant manipulation des fonctions et méthodes du package. In fine, ces dernières sorties semblent confirmer le diagnostic précédent. Le meilleur modèle d'imputation semble être celui implémenté section 3.2.3.



4. Analyse de sensibilité

Ne fonctionne pas avec avec MIPCA (pas de post traitement à la volée).

```
# delta <- c(0,10,20)
# imp.all <- vector("list", length(delta))
#
# post <- mice(data, maxit = 0)$post
# for (i in 1:length(delta)){
#   d <- delta[i]
#   cmd <- paste("imp[[j]][,i] <- imp[[j]][,i] +", d)
#   post["skin"] <- cmd
#   imp <- prelim(MIPCA(new.data[,1:8],ncp=ncp.res.new,verbose=F, method = "EM", method.mi = "Boot", nboot = 1
0),new.data[,1:8])
#   imp.all[[i]] <- imp
# }
#
# bwplot(imp.all[[1]])
# bwplot(imp.all[[2]])
# bwplot(imp.all[[3]])
```

5. Modélisation de la variable réponse

A partir des données complétées, la variable réponse, “Individus diabétiques”, est modélisée avec un modèle GLM logit avec pour prédicteurs : -npreg -glu -bp -skin -Insuline -bmi -ped -age

```
#Avec imputation MICE FCS
diabete.insuline.glm1 <- with(data=imp.mi.fcs ,exp=glm(type ~ npreg + glu + bp + skin + Insuline + bmi +  ped +
 age,family=binomial(link="logit")))

#Avec imputations missMDA
diabete.insuline.glm2 <- with(data=conv.imp.mipca.boot.new ,exp=glm(type ~ npreg + glu + bp + skin + log.Insulin
e + bmi +  ped + age,family=binomial(link="logit")))

diabete.insuline.glm3 <- with(data=conv.imp.mipca.EM.boot.new ,exp=glm(type ~ npreg + glu + bp + skin + log.Insu
line + bmi +  ped + age,family=binomial(link="logit")))

#Avec imputations missMDA mais sans la variable Insuline
diabete.noinsuline.glm <- with(data=conv.imp.mipca.EM.boot.new ,exp=glm(type ~ npreg + glu + bp + skin + bmi +
 ped + age,family=binomial(link="logit")))
```

```
summary(pool(diabete.insuline.glm2))
```

	estimate <dbl>	std.error <dbl>	statistic <dbl>	df <dbl>	p.value <dbl>
(Intercept)	-9.073823458	1.050510687	-8.63753560	568.5832	0.000000e+00
npreg	0.125198612	0.032402578	3.86384725	755.3016	1.212004e-04
glu	0.036190271	0.004336223	8.34603472	619.4073	4.440892e-16
bp	-0.008118974	0.008631724	-0.94059706	711.6450	3.472305e-01
skin	0.003091763	0.013530404	0.22850486	383.5685	8.193756e-01
log.Insuline	0.018618293	0.223183478	0.08342147	317.3460	9.335690e-01
bmi	0.089698642	0.019655605	4.56351475	598.4031	6.106823e-06
ped	0.854648731	0.298357413	2.86451315	752.1412	4.292883e-03
age	0.012552526	0.009578159	1.31053641	747.8874	1.904167e-01

9 rows

```
summary(pool(diabete.insuline.glm3))
```

	estimate <dbl>	std.error <dbl>	statistic <dbl>	df <dbl>	p.value <dbl>
(Intercept)	-8.679808600	0.998183517	-8.6956040	404.4840	0.000000e+00
npreg	0.123864604	0.032914054	3.7632740	739.6652	1.809831e-04
glu	0.035638279	0.004687291	7.6031724	553.4942	1.245670e-13
bp	-0.009817762	0.008983007	-1.0929261	577.4229	2.748820e-01
skin	0.002568682	0.012735887	0.2016885	344.7643	8.402792e-01
log.Insuline	0.022761193	0.182204250	0.1249213	287.2606	9.006732e-01
bmi	0.082368968	0.022591864	3.6459571	327.9252	3.096904e-04
ped	0.873535102	0.296812783	2.9430508	747.0641	3.350638e-03
age	0.013564960	0.009587980	1.4147881	737.3803	1.575528e-01

9 rows

```
summary(pool(diabete.noinsuline.glm))
```

	estimate <dbl>	std.error <dbl>	statistic <dbl>	df <dbl>	p.value <dbl>
(Intercept)	-8.610548345	0.930142155	-9.2572391	361.3933	0.0000000000
npreg	0.123071043	0.032301980	3.8100155	750.7349	0.0001503275
glu	0.035988543	0.003669684	9.8069856	678.1816	0.0000000000
bp	-0.009825297	0.008978393	-1.0943269	575.5626	0.2742693685
skin	0.002477095	0.012704567	0.1949768	344.8976	0.8455259343
bmi	0.082483896	0.022253031	3.7066365	342.4423	0.0002448617
ped	0.873251472	0.296319782	2.9469901	748.1890	0.0033085867
age	0.013658141	0.009499184	1.4378225	745.0783	0.1509042681

8 rows

On remarque que les varables skin, age et bp ont une grande p-value.

- Comparaison des modèles https://stefvanbuuren.name/mice/reference/pool.html (https://stefvanbuuren.name/mice/reference/pool.html)

On accepte l'utilité de la variable "Insuline" et on conserve le modèle complet. p-value importante => on rejette l'influence de la variable insuline.

```
#Rapport de vraisemblance
pool.compare(diabete.insuline.glm3, diabete.noinsuline.glm, method = "likelihood")$pvalue
```

```
## [1] 0.9015207
```

```
#Anova
anova(diabete.insuline.glm3,diabete.noinsuline.glm)
```

```
##    test  statistic df1      df2 df.com   p.value       riv
##  1 ~~ 2 0.01560533   1 358.8463    759 0.9006558 0.5876564
```

Au regard des différentes comparaisons et du test ANOVA final, on conserve un modèle excluant les variables : bmi,glu,npreg <= NOK? On aurait tendance à rejetter les variables skin, bp et age.

```
## [1] "Test du rapport de vraisemblance modele complet vs modele sans la variable age : 1"
```

```
## [1] "Test du rapport de vraisemblance modele complet vs modele sans la variable skin : 1"
```

```
## [1] "Test du rapport de vraisemblance modele complet vs modele sans la variable insuline : 1"
```

```
## [1] "Test du rapport de vraisemblance modele complet vs modele sans la variable bp : 1"
```

```
## [1] "Test du rapport de vraisemblance modele complet vs modele sans la variable bmi : 1"
```

```
## [1] "Test du rapport de vraisemblance modele complet vs modele sans la variable glu  : 1"
```

```
## [1] "Test du rapport de vraisemblance modele complet vs modele sans la variable npreg : 1"
```

```
## [1] "Test du rapport de vraisemblance modele complet vs modele sans la variable ped : 1"
```

```
## [1] "AIC du modele complet : 737.565215113922"
```

```
## [1] "AIC du modele sans la variable age : 688.42701414307"
```

```
## [1] "AIC du modele sans la variable skin : 686.912022364828"
```

```
## [1] "AIC du modele sans la variable insuline : 730.945697138493"
```

```
## [1] "AIC du modele sans la variable bp : 686.77352700022"
```

```
## [1] "AIC du modele sans la variable bmi : 698.374069190742"
```

```
## [1] "AIC du modele sans la variable npreg : 703.804145730773"
```

```
## [1] "AIC du modele sans la variable glu : 721.406244095475"
```

```
## [1] "AIC du modele sans la variable ped : 697.710740945811"
```

Le AIC moyen tendrait à supprimer les variables: age, bp, skin.

```
## [1] 686.7735
```

```
## [1] 686.9861
```

```
## [1] 685.9922
```

```
## [1] 730.1093
```

```
## [1] "Test du rapport de vraisemblance modele sans bp vs modele sans les variables bp et age : 0.1563246664374
84"
```

```
## [1] "Test du rapport de vraisemblance modele sans bp et age vs modele sans bp, age et skin : 0.60293811978104
3"
```

```
## [1] "Test du rapport de vraisemblance modele sans bp, age et skin vs modele sans bp, age, Insuline et skin :
5.05620865176581e-08"
```

	estimate <dbl>	std.error <dbl>	statistic <dbl>	df <dbl>	p.value <dbl>
(Intercept)	-12.405428484	1.098190018	-11.2962495	458.9924	0.000000e+00
age	0.014117503	0.010000333	1.4117032	672.8819	1.584996e-01
bmi	0.059382526	0.020760972	2.8602961	396.0860	4.456552e-03
npreg	0.137604743	0.034300488	4.0117430	705.6599	6.670660e-05
ped	1.042897041	0.314970534	3.3110940	631.8189	9.822520e-04
glu	0.023181453	0.004250367	5.4539892	651.8361	6.998909e-08
skin	0.007251079	0.012711429	0.5704377	317.7945	5.687839e-01
log.Insuline	1.042178242	0.198250167	5.2568846	316.2518	2.701822e-07

8 rows

	estimate <dbl>	std.error <dbl>	statistic <dbl>	df <dbl>	p.value <dbl>
(Intercept)	-12.066872334	1.065782421	-11.3220786	441.1780	0.000000e+00
bmi	0.057465610	0.020645050	2.7835054	396.1766	5.635059e-03
npreg	0.162916386	0.030030135	5.4250967	672.8558	8.088187e-08
ped	1.052934793	0.314393739	3.3490959	632.1484	8.589503e-04
glu	0.024332608	0.004187943	5.8101567	648.5969	9.785025e-09
skin	0.007360835	0.012657256	0.5815507	317.3671	5.612822e-01
log.Insuline	1.033176320	0.198907292	5.1942606	310.3146	3.728642e-07

7 rows

	estimate <dbl>	std.error <dbl>	statistic <dbl>	df <dbl>	p.value <dbl>
(Intercept)	-12.10524395	1.061637662	-11.402425	442.8446	0.000000e+00
bmi	0.06562467	0.016812126	3.903413	277.0179	1.191724e-04
npreg	0.16414688	0.029970531	5.476943	673.6068	6.115410e-08
ped	1.05605752	0.312953079	3.374492	642.8348	7.839937e-04
glu	0.02439667	0.004176126	5.841938	652.6172	8.143884e-09
log.Insuline	1.02785016	0.197323566	5.208958	316.7345	3.427197e-07

6 rows

	estimate <dbl>	std.error <dbl>	statistic <dbl>	df <dbl>	p.value <dbl>
(Intercept)	-8.70638030	0.746811093	-11.658076	491.1198	0.000000e+00
bmi	0.07043438	0.017712957	3.976433	225.5763	9.423665e-05
npreg	0.13702218	0.027473997	4.987341	752.2558	7.606701e-07
ped	0.93701266	0.294394502	3.182847	750.6832	1.518464e-03
glu	0.03775943	0.003515003	10.742360	743.9005	0.000000e+00

5 rows

Le meilleur modèle est le modèle diabete.insuline.glm5 avec les variables : bmi, npreg, ped, glu et log.Insuline

Au final, les estimations des coefficients sont assez proches en comparaison de la méthode des cas concrets (= listwise deletion). En revanche, le pouvoir prédictif du modèle sort amélioré, toutes les variables étant significatives avec les jeux de données imputés.

```
##
## Call:
## glm(formula = type ~ npreg + glu + bmi + ped + age, family = binomial(link = "logit"),
##     data = data[ok, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8827  -0.6535  -0.3694   0.6521   2.5814
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.992080   1.086866  -9.193  < 2e-16 ***
## npreg        0.083953   0.055031   1.526 0.127117
## glu          0.036458   0.004978   7.324 2.41e-13 ***
## bmi          0.078139   0.020605   3.792 0.000149 ***
## ped          1.150913   0.424242   2.713 0.006670 **
## age          0.034360   0.017810   1.929 0.053692 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 344.89  on 386  degrees of freedom
## AIC: 356.89
##
## Number of Fisher Scoring iterations: 5
```

```
##
## Call:
## glm(formula = type ~ bp + skin + I(log(Insuline)) + ped + age,
##     family = binomial(link = "logit"), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5066  -0.7784  -0.4446   0.9042   2.4168
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -10.09698    1.30143  -7.758 8.60e-15 ***
## bp                 0.01354    0.01075   1.259  0.20799
## skin               0.03597    0.01259   2.858  0.00427 **
## I(log(Insuline))   1.01633    0.20295   5.008 5.51e-07 ***
## ped                1.01821    0.35280   2.886  0.00390 **
## age                0.05652    0.01280   4.417 1.00e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.70  on 393  degrees of freedom
## Residual deviance: 393.96  on 388  degrees of freedom
##   (374 observations deleted due to missingness)
## AIC: 405.96
##
## Number of Fisher Scoring iterations: 4
```

5. Conclusion

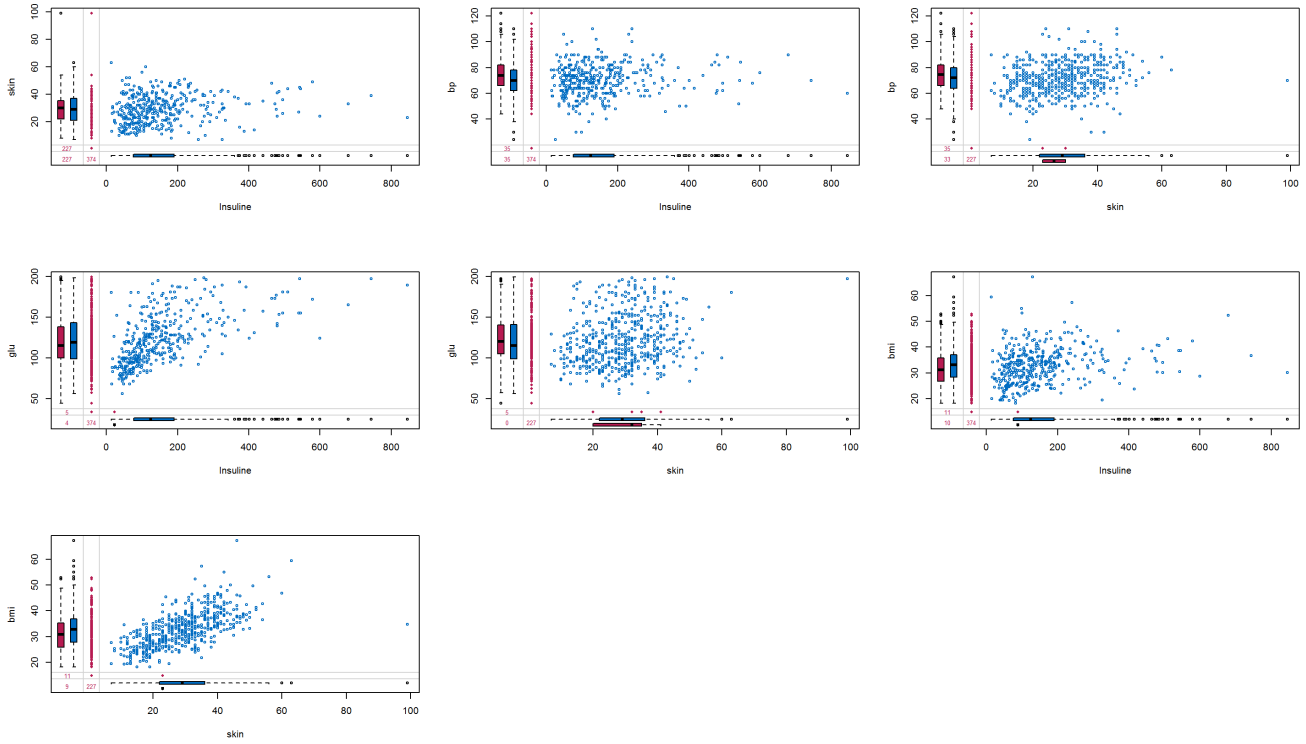
Intérêt : Les données Pima sont très utilisées (cf. Kaggle). Pour autant il ne semble pas qu'un modèle d'inputation ait été utilisé. Bien souvent les données manquantes sont supprimées (analyse des cas concrets ou listwise deletion) ou bien une imputation simple par moyenne est mise en oeuvre. Dans notre cas, nous avons réussi à démontrer la pertinence de l'imputation avec pour objectif l'implémentation d'un modèle de classification.

Annexes

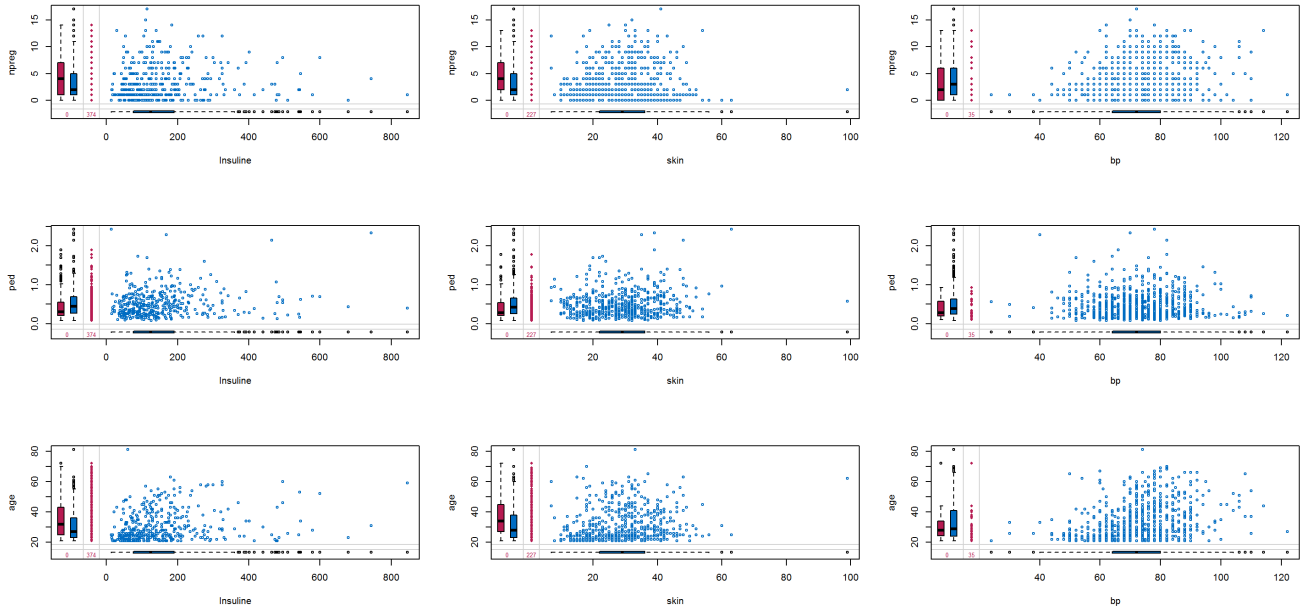
Annexe 1 : Graphiques

graphiques N°1 - Mécanismes des variables “Insuline” et “skin” (§2.1)

- Distributions marginales des variables Insuline, Skin et bp avec les autres variables incomplètes



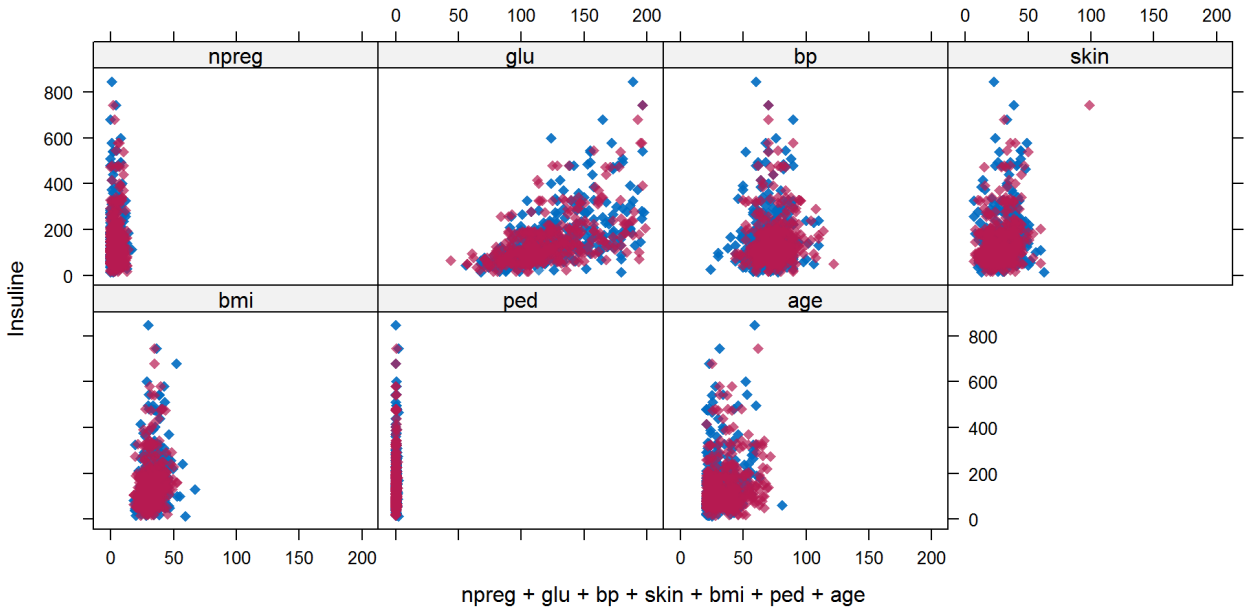
- Distributions marginales des mêmes variables avec les variables complètes

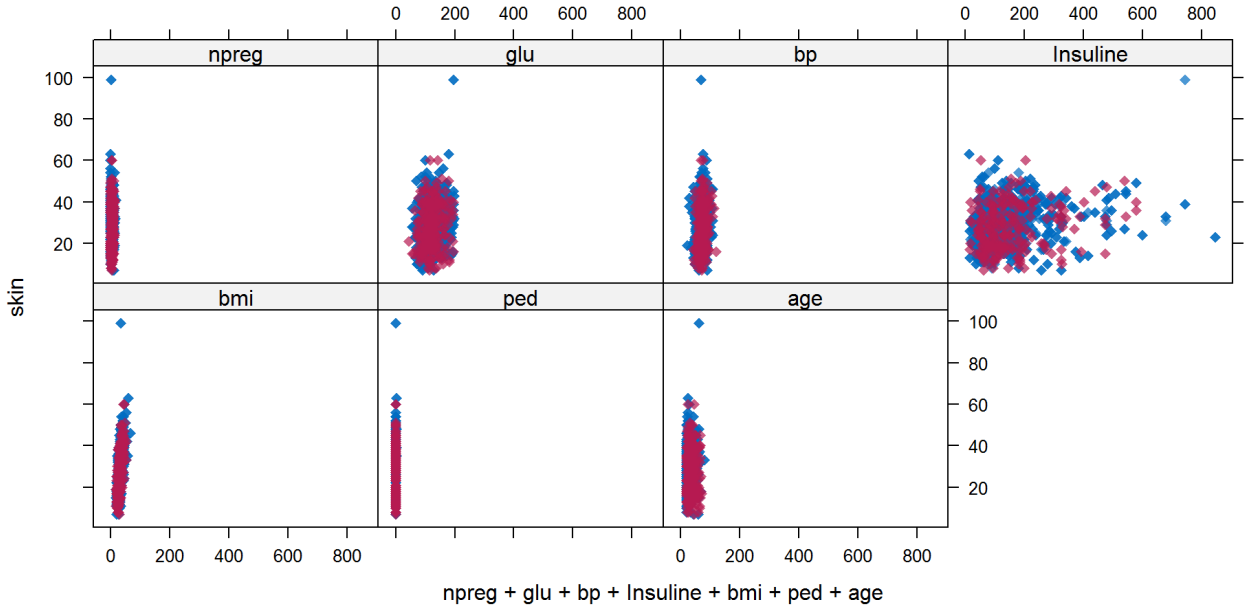


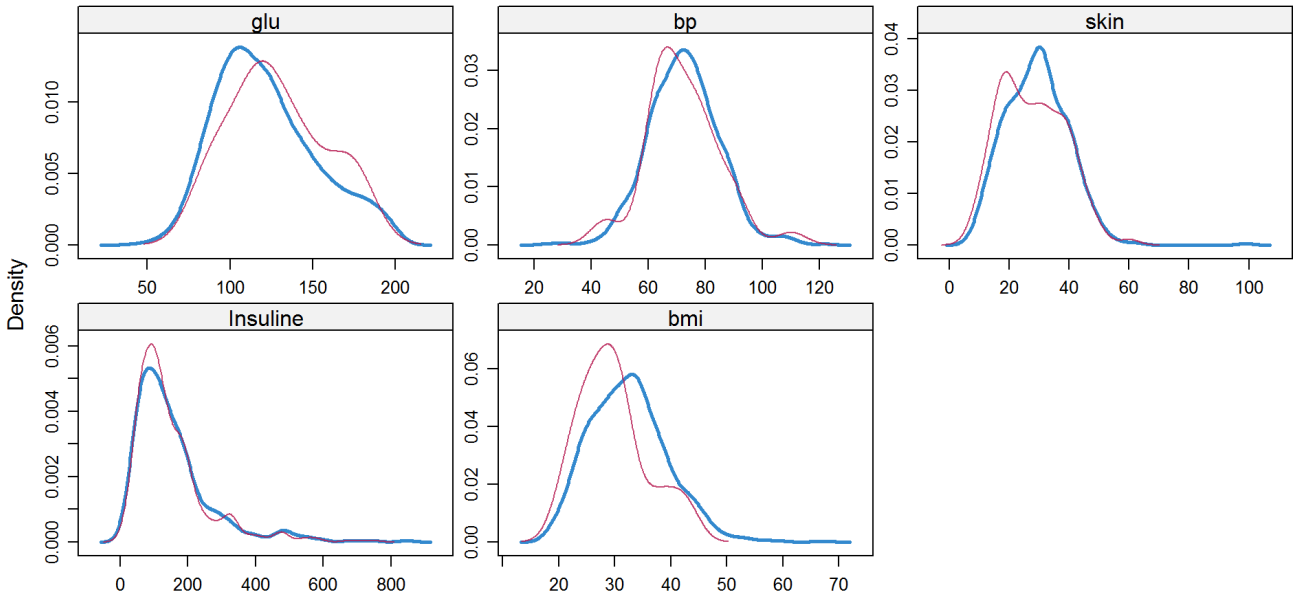
graphiques N°2 - Analyses des distributions pour les Imputations simples (§-3.1.3)

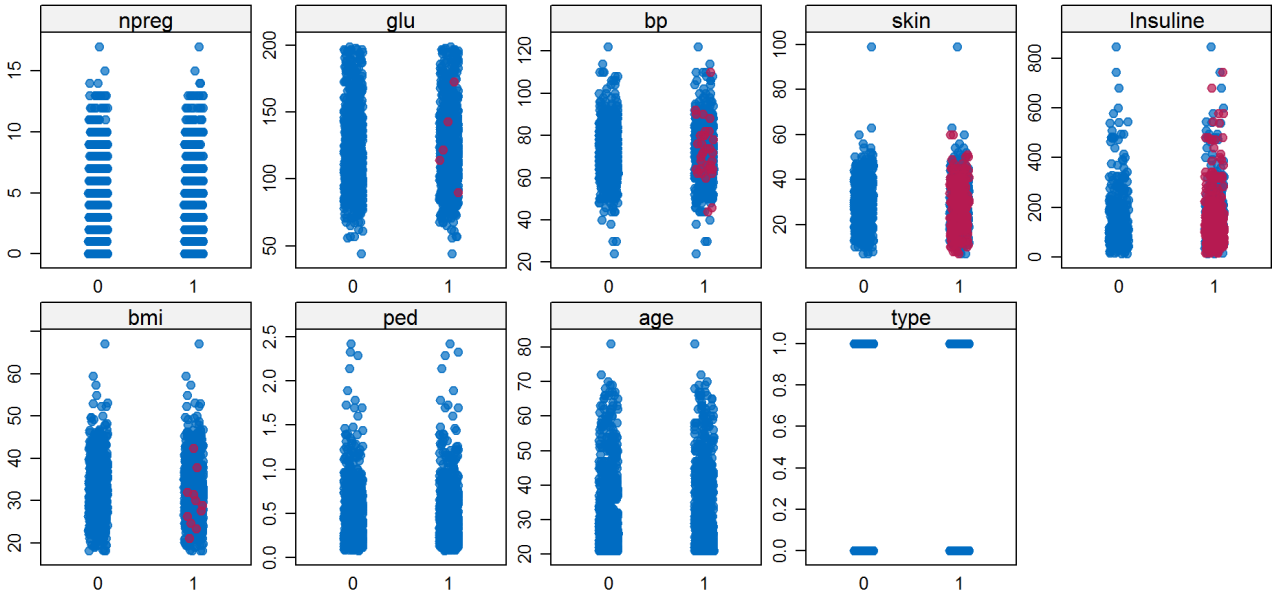
Les graphiques montrent que les imputations par régression linéaire stochastique comportent des valeurs aberrantes, en l'espèce, des valeurs négatives pour l'épaisseur de peau et le taux d'insuline. Les distributions des valeurs observées et imputées sont proches, ce qui n'infirme ni ne confirme le mécanisme MAR.

- Imputations simples - PMM

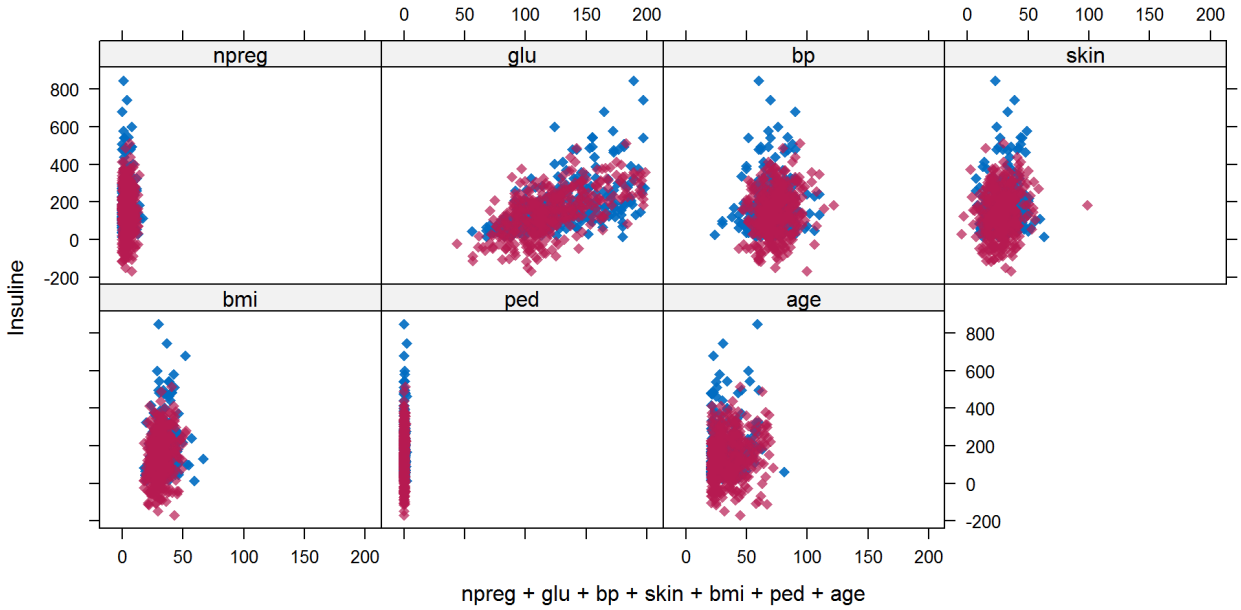


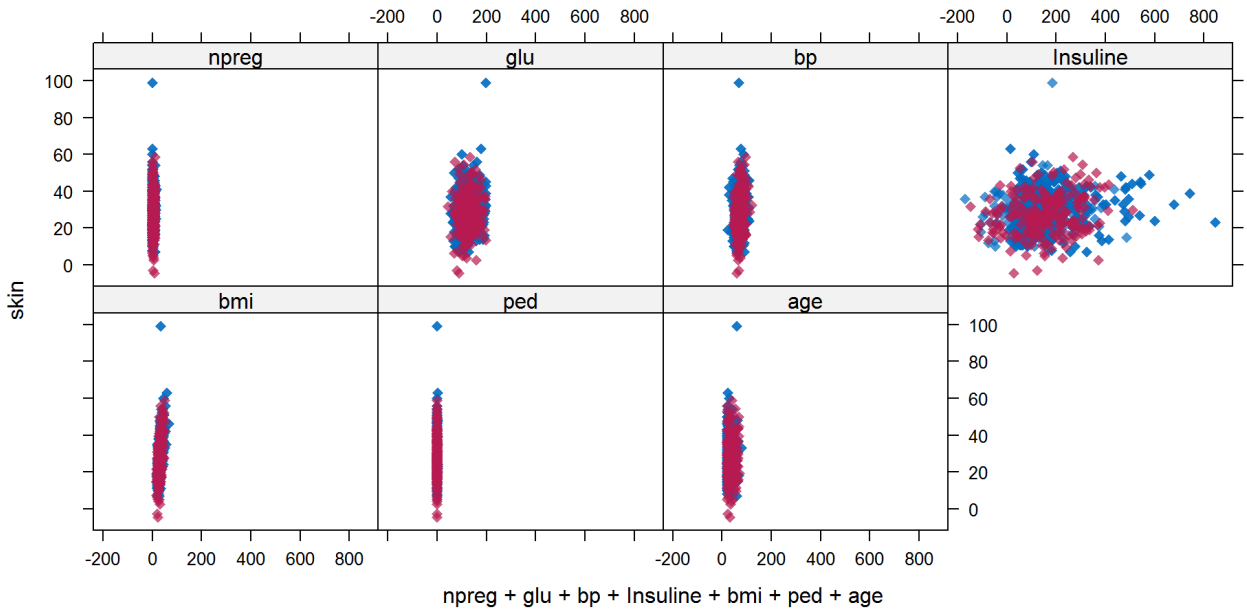


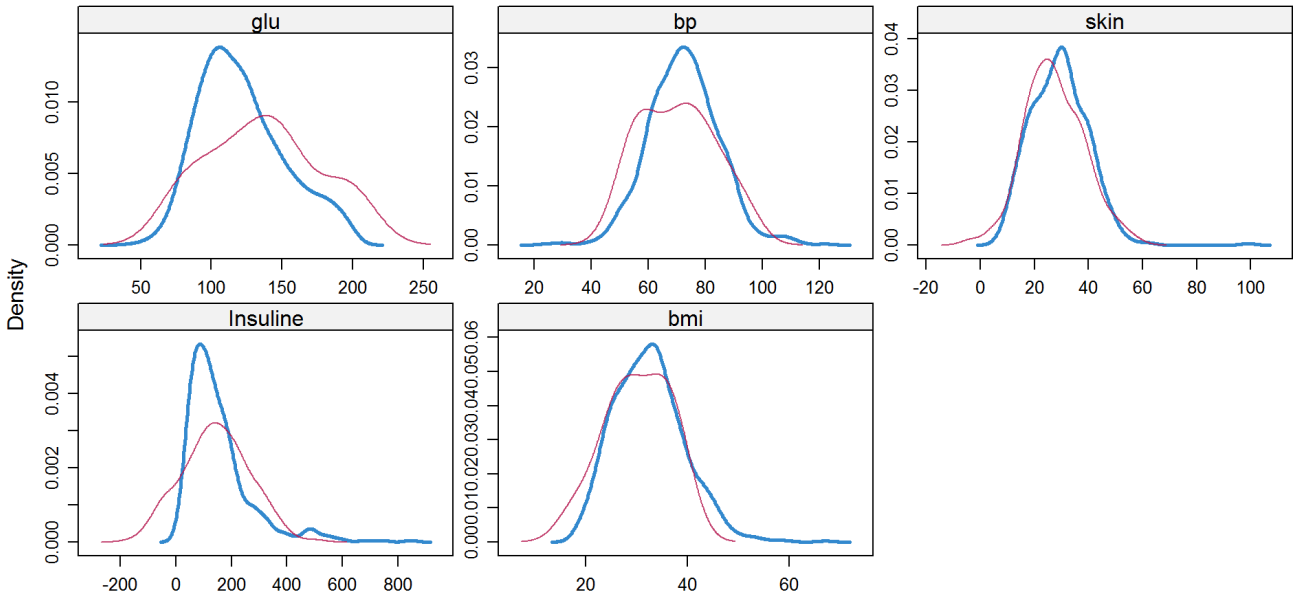


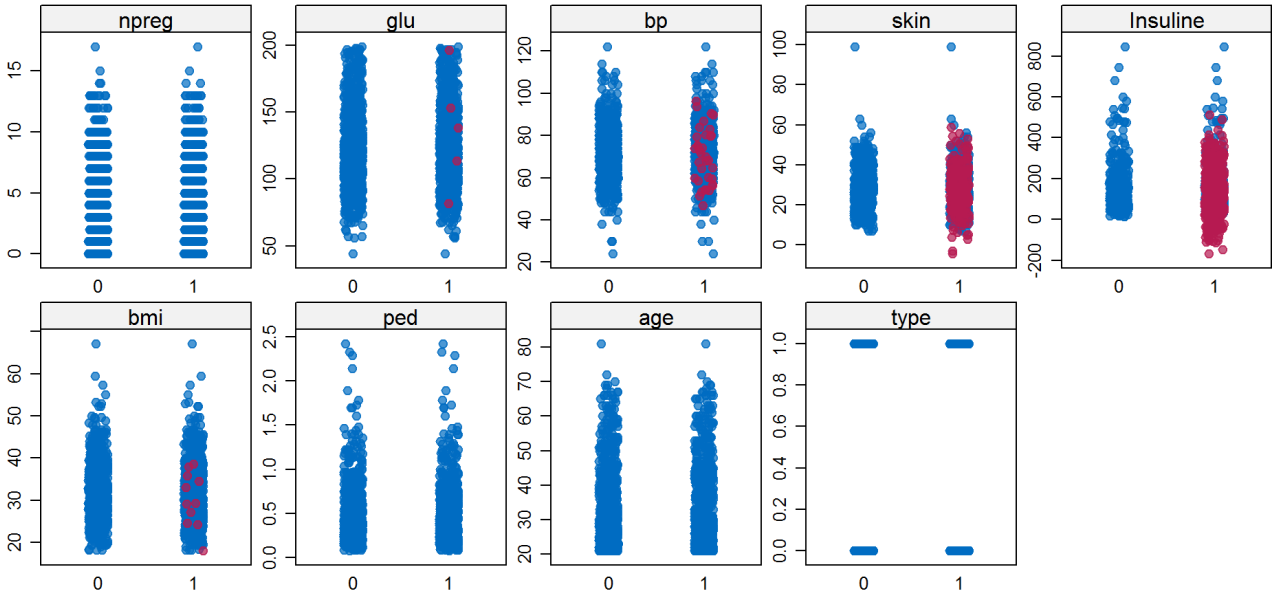


- Imputations simples - Normales




npreg
glu
bp
Insuline
bmi
ped
age
skin
npreg + glu + bp + Insuline + bmi + ped + age

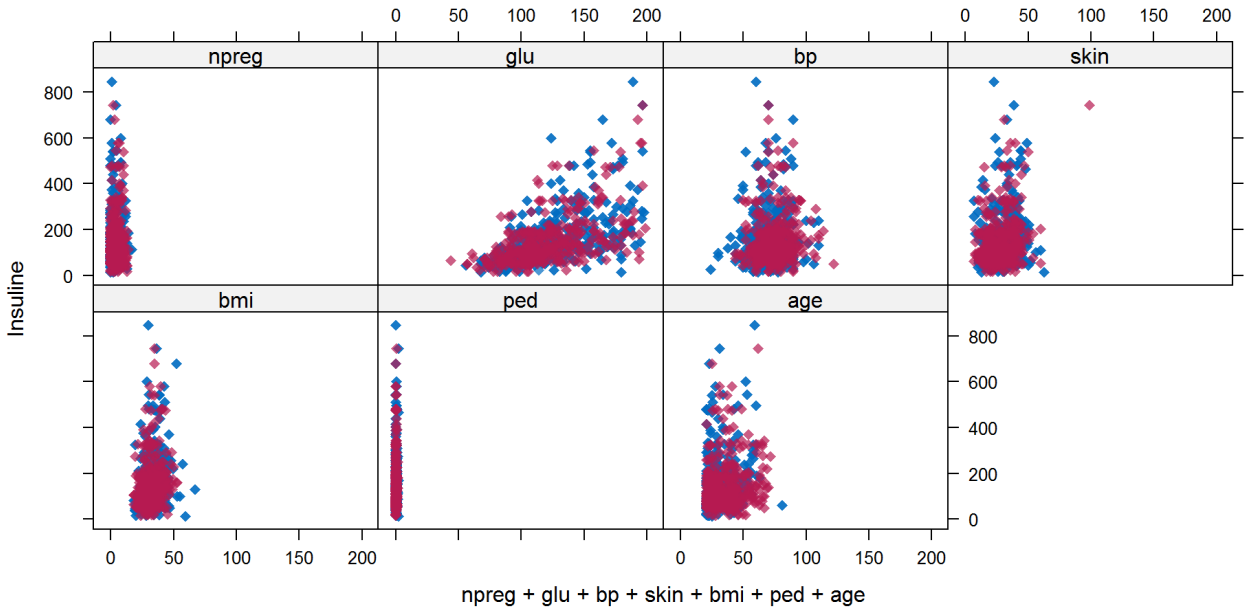

glu
bp
skin
Insuline
bmi
Density

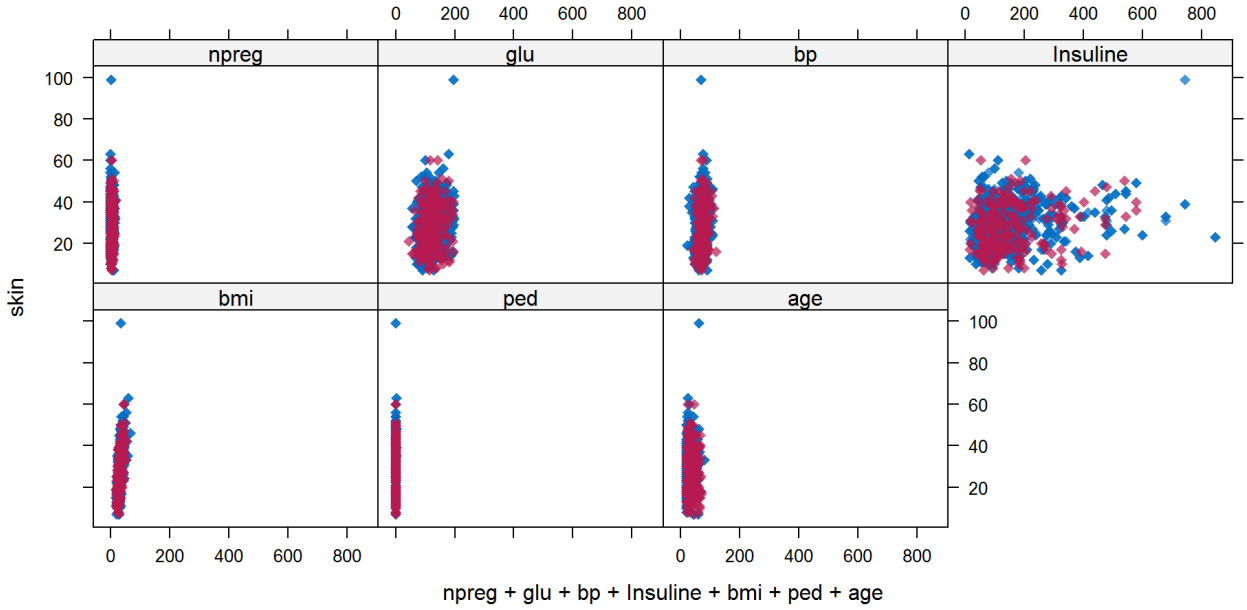

npreg
glu
bp
skin
Insuline
bmi
ped
age
type

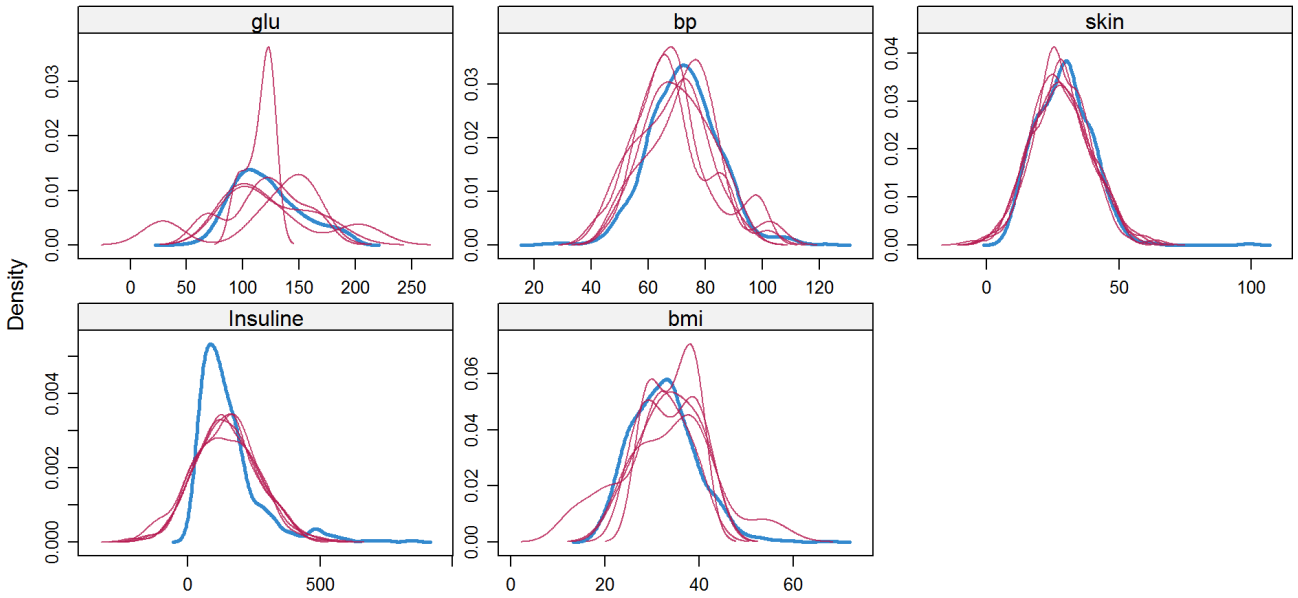
graphiques N°3 - Analyses des distributions pour les Imputations multiples (§-3.1.3)

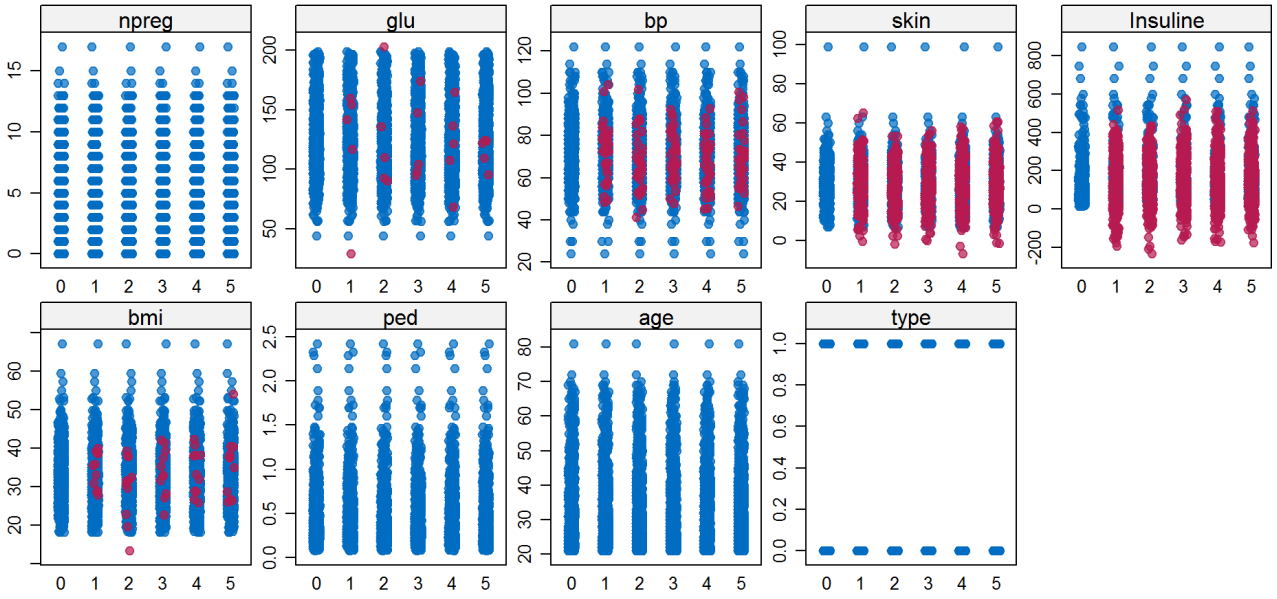
A nouveau, certaines imputations ne sont pas plausibles : avec le modèle joint les graphiques montrent des valeurs aberrantes pour l'épaisseur de peau et le taux d'insuline. S'agissant de la méthode FCS, les distributions des variables glu et bmi présentent des profils très divergents pour chacune des imputations, en raison du nombre peu élevé de données manquantes.

- Imputations multiples - JM

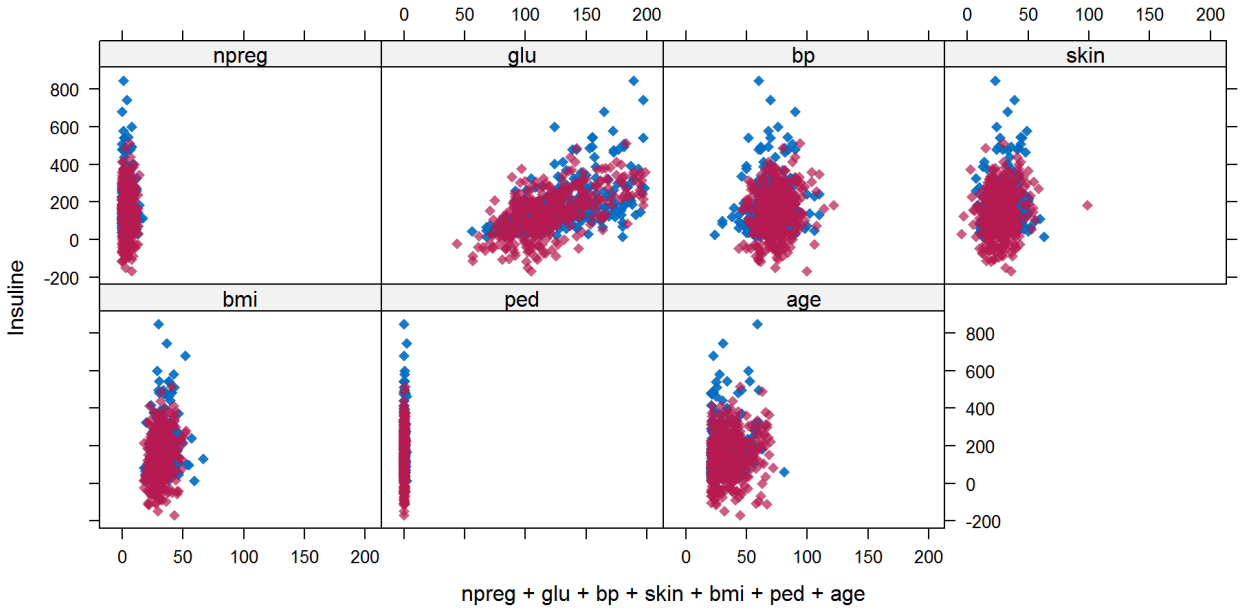


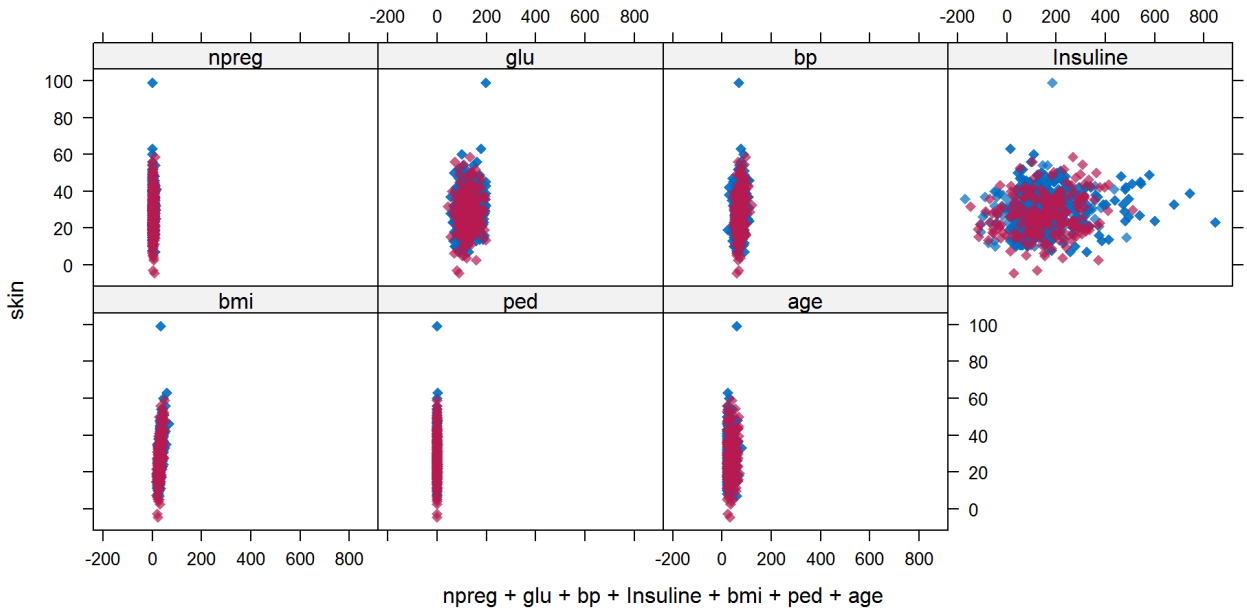


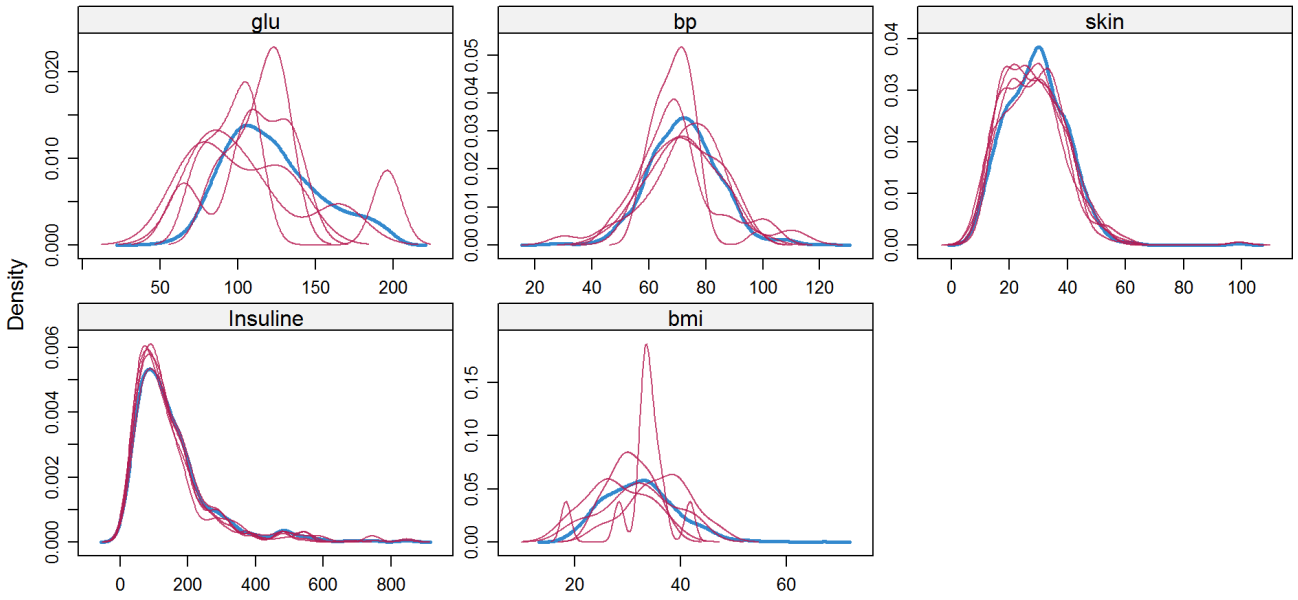


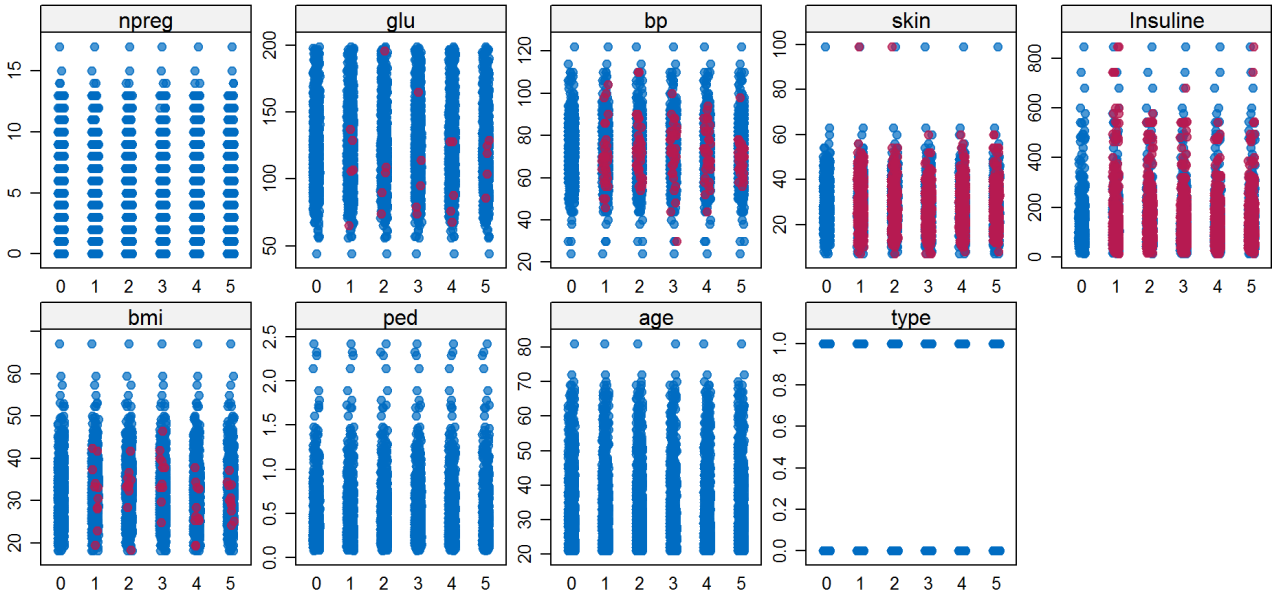


- Imputations multiples - FCS




npreg
glu
bp
Insuline
bmi
ped
age
skin
npreg + glu + bp + Insuline + bmi + ped + age

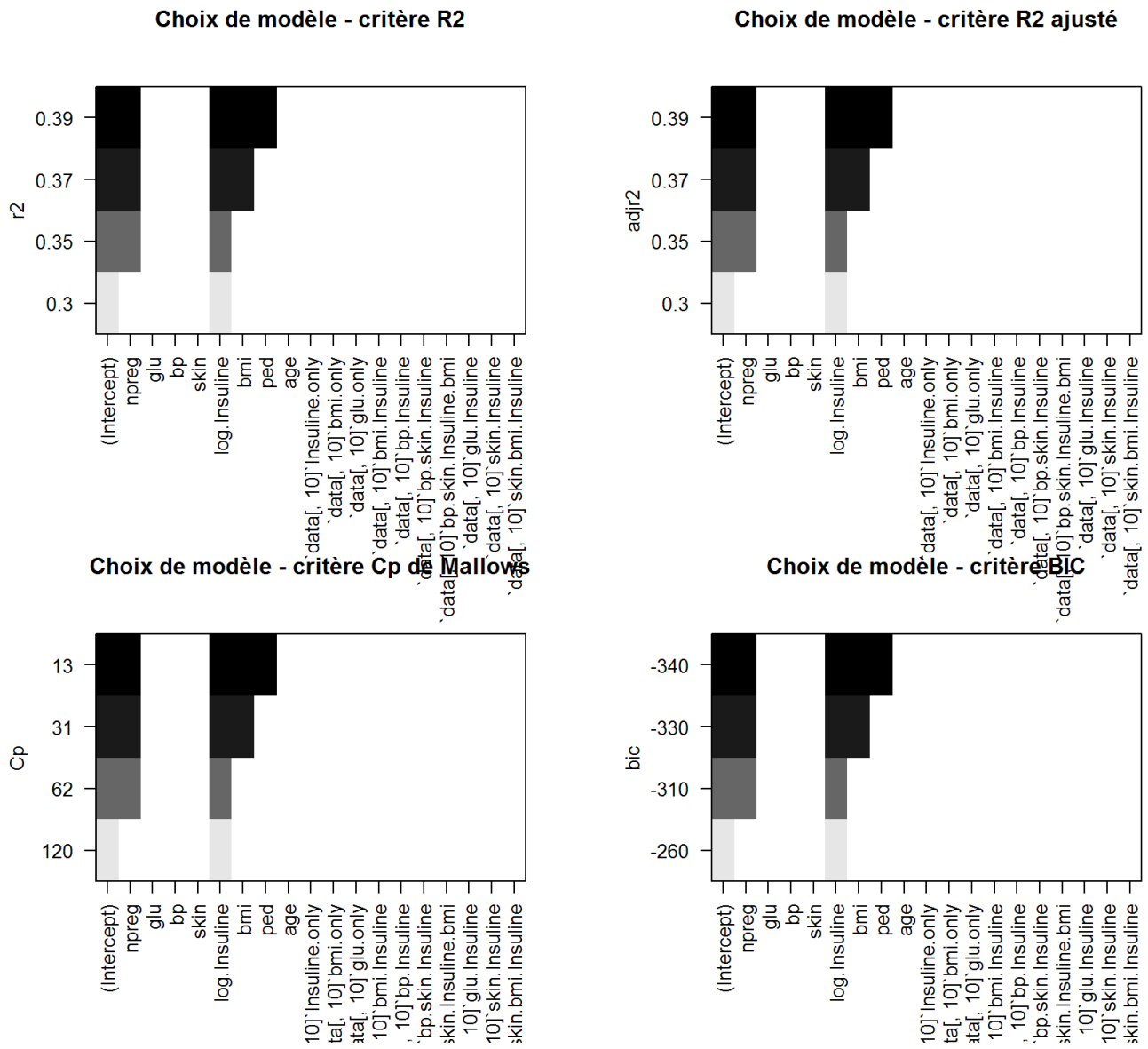

glu
bp
skin
Insuline
bmi
Density


npreg
glu
bp
skin
Insuline
bmi
ped
age
type

Annexe 2 - choix de modèle

Dans cette annexe on applique les techniques regsubset et step pour le choix automatique des variables. Les données utilsées sont les données obtenu à partir de la variable $res.inputePCA On obtient à nouveau ce résultat en utilsant comme données d'input les moyennes des diffrérents jeux de données issues du boostrap résultant de mipca.

Choix du modèle - Utilisation de regsubsets



La méthode step du package leaps nous fait choisir un modèle à 4 variables explicatives: npreg + glu + skin + ped

Choix du modèle - méthode step du package MASS à partir du modèle saturé

On reprend le modèle saturé obtenu au §2: m_sature

```
m_sature = glm(formula = type ~ . -typeMissing ,  family = binomial(link="logit"),  data = res.imputePCA)
summary(m_sature)
```

```
##
## Call:
## glm(formula = type ~ . - typeMissing, family = binomial(link = "logit"),
##     data = res.imputePCA)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4482  -0.5833  -0.2864   0.5372   3.1055
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -15.983280   1.291451 -12.376  < 2e-16
## npreg                            0.155352   0.035245   4.408 1.04e-05
## glu                              0.010871   0.004833   2.250 0.024476
## bp                              -0.004187   0.009825  -0.426 0.670016
## skin                             0.023228   0.014993   1.549 0.121326
## log.Insuline                     1.983972   0.234466   8.462  < 2e-16
## bmi                              0.057096   0.022174   2.575 0.010028
## ped                              1.231368   0.322506   3.818 0.000134
## age                              0.015108   0.010609   1.424 0.154438
## `data[, 10]`Insuline.only        0.531297   0.293752   1.809 0.070504
## `data[, 10]`bmi.only           -13.325297 882.743485  -0.015 0.987956
## `data[, 10]`glu.only            -8.092051 882.743507  -0.009 0.992686
## `data[, 10]`bmi.Insuline       -11.475459 882.743421  -0.013 0.989628
## `data[, 10]`bp.Insuline          1.184762   2.439781   0.486 0.627250
## `data[, 10]`bp.skin.Insuline     1.266642   0.564965   2.242 0.024963
## `data[, 10]`bp.skin.Insuline.bmi -3.465829   4.110241  -0.843 0.399106
## `data[, 10]`glu.Insuline         0.225128   2.110605   0.107 0.915055
## `data[, 10]`skin.Insuline        0.386300   0.257823   1.498 0.134051
## `data[, 10]`skin.bmi.Insuline    0.177155   1.630963   0.109 0.913504
##
## (Intercept)                    ***
## npreg                          ***
## glu                            *
## bp
## skin
## log.Insuline                   ***
## bmi                            *
## ped                            ***
## age
## `data[, 10]`Insuline.only      .
## `data[, 10]`bmi.only
## `data[, 10]`glu.only
## `data[, 10]`bmi.Insuline
## `data[, 10]`bp.Insuline
## `data[, 10]`bp.skin.Insuline   *
## `data[, 10]`bp.skin.Insuline.bmi
## `data[, 10]`glu.Insuline
## `data[, 10]`skin.Insuline
## `data[, 10]`skin.bmi.Insuline
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 993.48  on 767  degrees of freedom
## Residual deviance: 611.71  on 749  degrees of freedom
## AIC: 649.71
##
## Number of Fisher Scoring iterations: 13
```

Choix du modèle - Méthode progressive - step backward-forward à partir de m_sature

- Critère AIC
- Critère BIC

Choix du modèle - Modèle obtenu

```
#modele_step_BwdFwd_AIC<-glm(formula = type ~ (npreg + glu + bp + skin + Insuline + bmi + ped + age), family = b
inomial, data = dataFramePima)
summary(modele_step_BwdFwd_AIC)
```

```
##
## Call:
## glm(formula = type ~ npreg + glu + skin + log.Insuline + bmi +
##     ped + age, family = binomial(link = "logit"), data = res.imputePCA)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4264  -0.5889  -0.2909   0.5870   3.0197
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -15.747213   1.182951 -13.312  < 2e-16 ***
## npreg          0.154876   0.035028   4.422  9.8e-06 ***
## glu            0.009759   0.004732   2.063 0.039156 *
## skin           0.020873   0.014447   1.445 0.148521
## log.Insuline   2.003656   0.231134   8.669  < 2e-16 ***
## bmi            0.049770   0.021666   2.297 0.021610 *
## ped            1.128815   0.309756   3.644 0.000268 ***
## age            0.017885   0.010084   1.774 0.076142 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 993.48  on 767  degrees of freedom
## Residual deviance: 622.63  on 760  degrees of freedom
## AIC: 638.63
##
## Number of Fisher Scoring iterations: 5
```

```
AIC(modele_step_BwdFwd_AIC)
```

```
## [1] 638.6294
```

```
BIC(modele_step_BwdFwd_AIC)
```

```
## [1] 675.7797
```

```
#m_StepBwdFwd_BIC<-glm(formula = type ~ (npreg + glu + skin + Insuline + ped),family = binomial, data = dataFram
ePima)
summary(m_StepBwdFwd_BIC)
```

```
##
## Call:
## glm(formula = type ~ npreg + log.Insuline + bmi + ped, family = binomial(link = "logit"),
##     data = res.imputePCA)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4006  -0.5985  -0.3206   0.5928   2.9994
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -15.67803    1.12331 -13.957  < 2e-16 ***
## npreg          0.20231    0.03032   6.672 2.52e-11 ***
## log.Insuline   2.30853    0.18860  12.240  < 2e-16 ***
## bmi            0.07005    0.01502   4.664 3.11e-06 ***
## ped            1.20771    0.30968   3.900 9.62e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 993.48  on 767  degrees of freedom
## Residual deviance: 634.01  on 763  degrees of freedom
## AIC: 644.01
##
## Number of Fisher Scoring iterations: 5
```

```
AIC(m_StepBwdFwd_BIC)
```

```
## [1] 644.0087
```

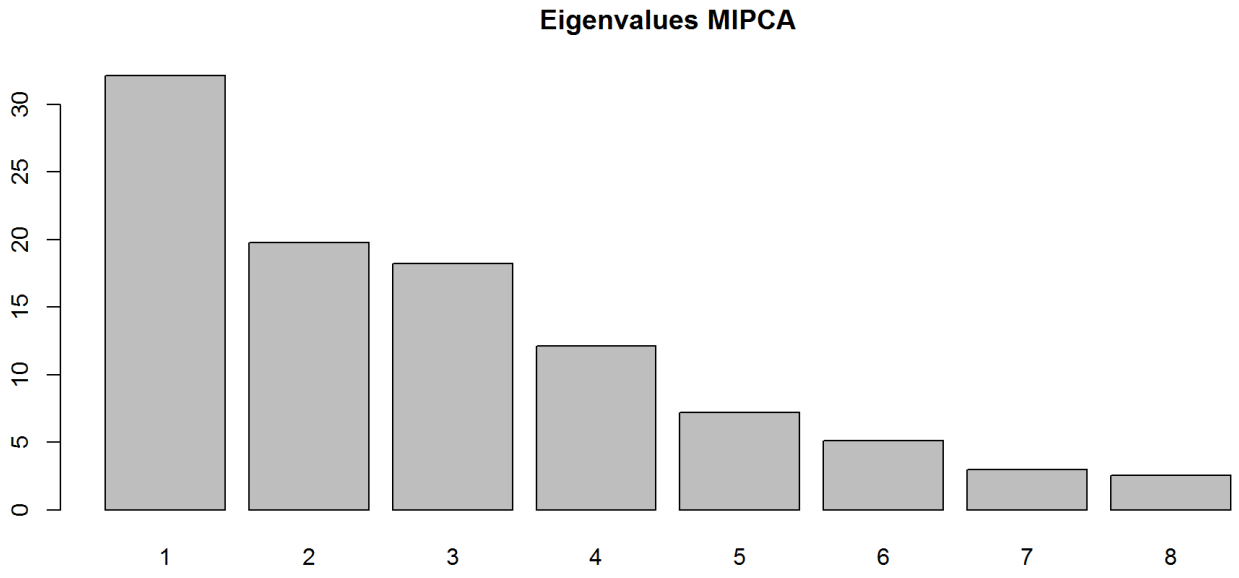
```
BIC(m_StepBwdFwd_BIC)
```

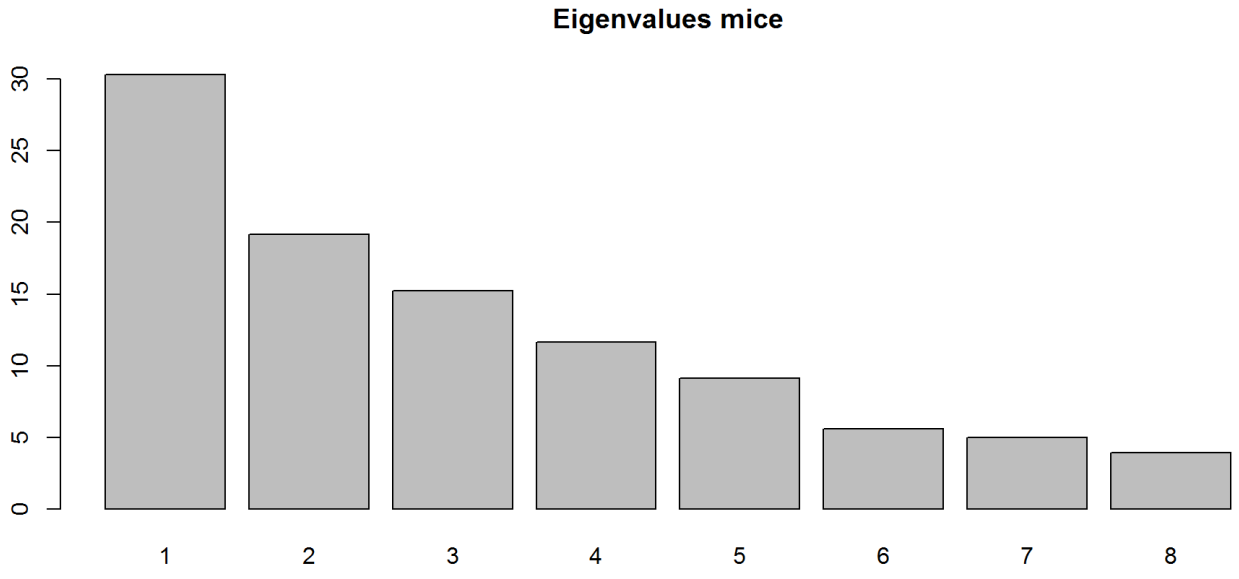
```
## [1] 667.2276
```

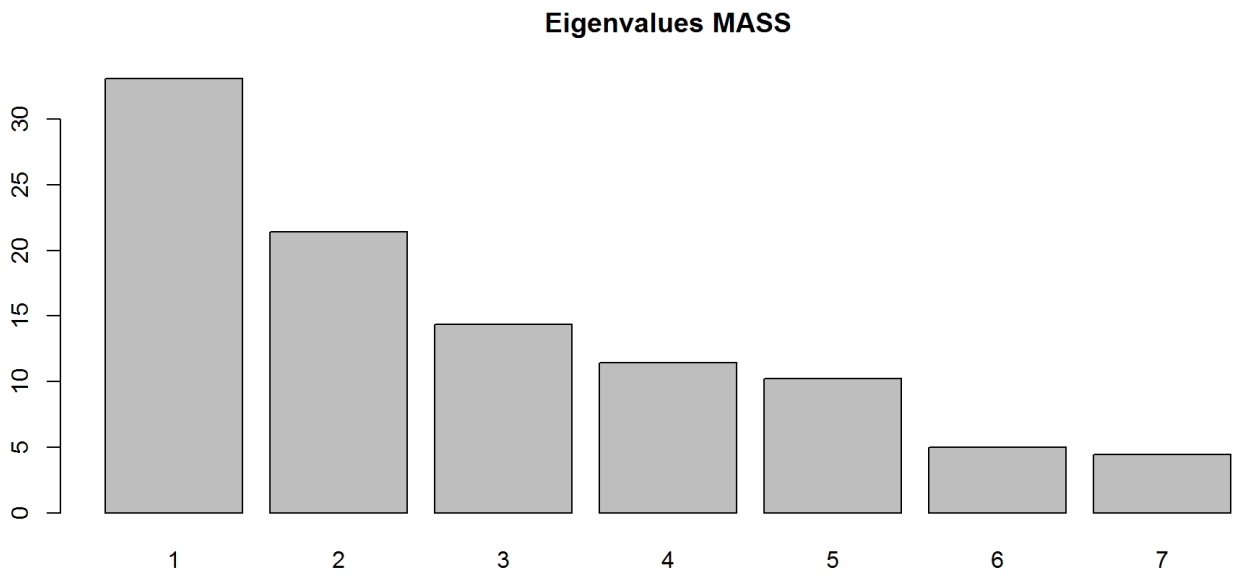
Annexe 3 : Analyse PCA - comparaison aux données non traitées

- PCA sur les données complétées par MIPCA
- PCA sur les données imputé par mice
- PCA sur les données Pima de MASS (non imputées)

1. Choix du nombre d’axes

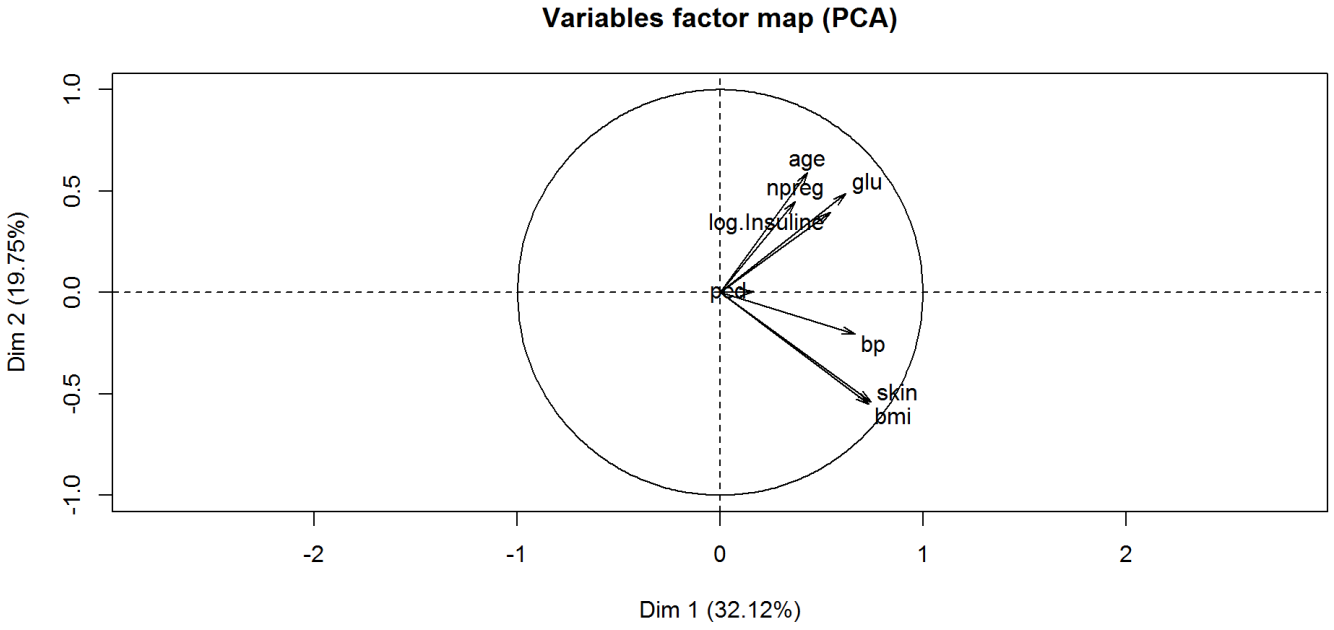


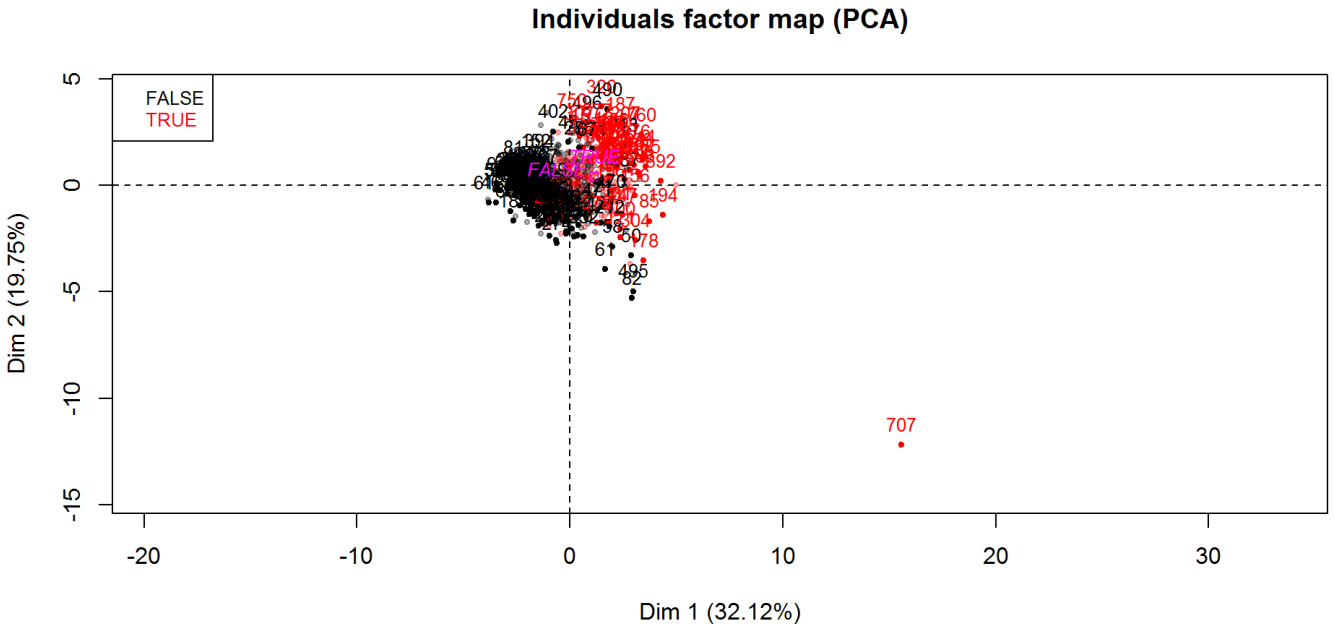


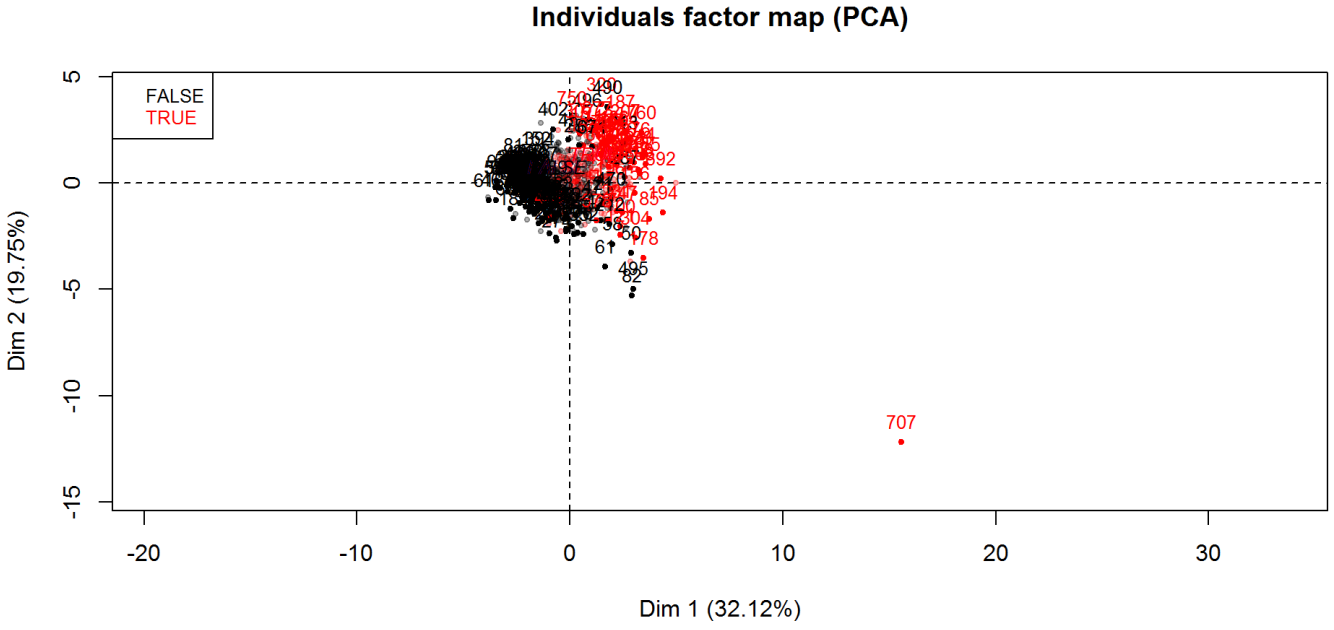


2. Graphiques : nuage des individus et cercle de qualité des projections

- Données Pima complétées avec la méthode MIPCA du package MissMDA

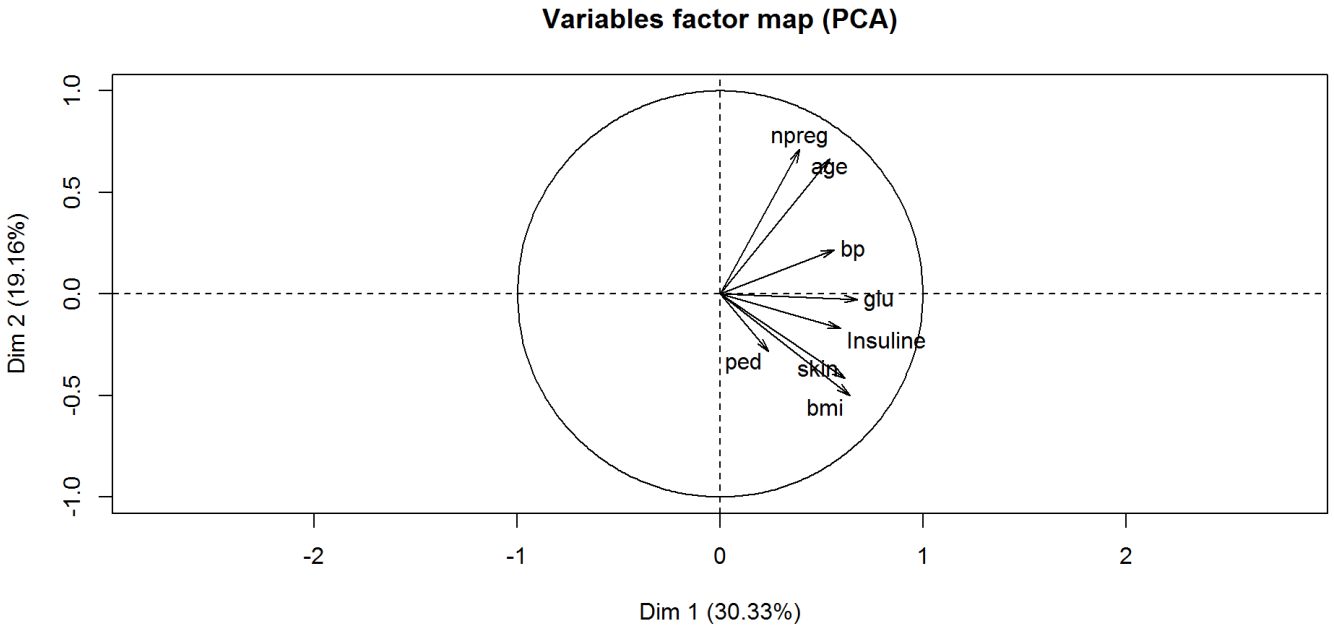


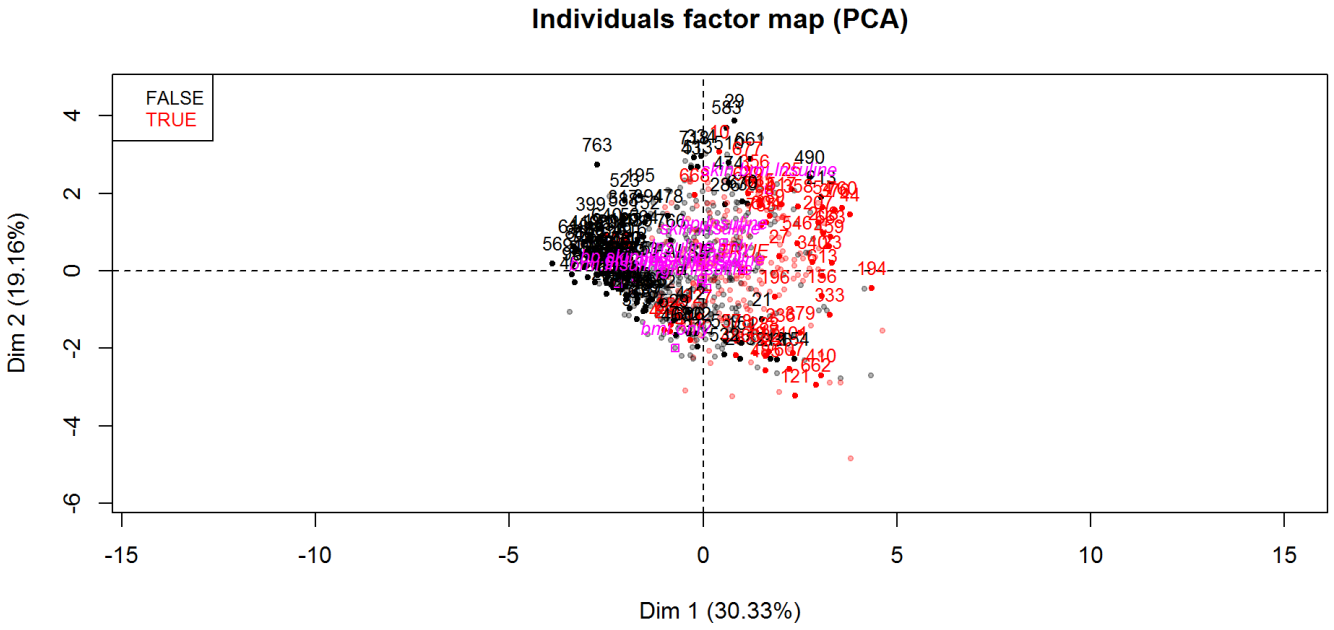


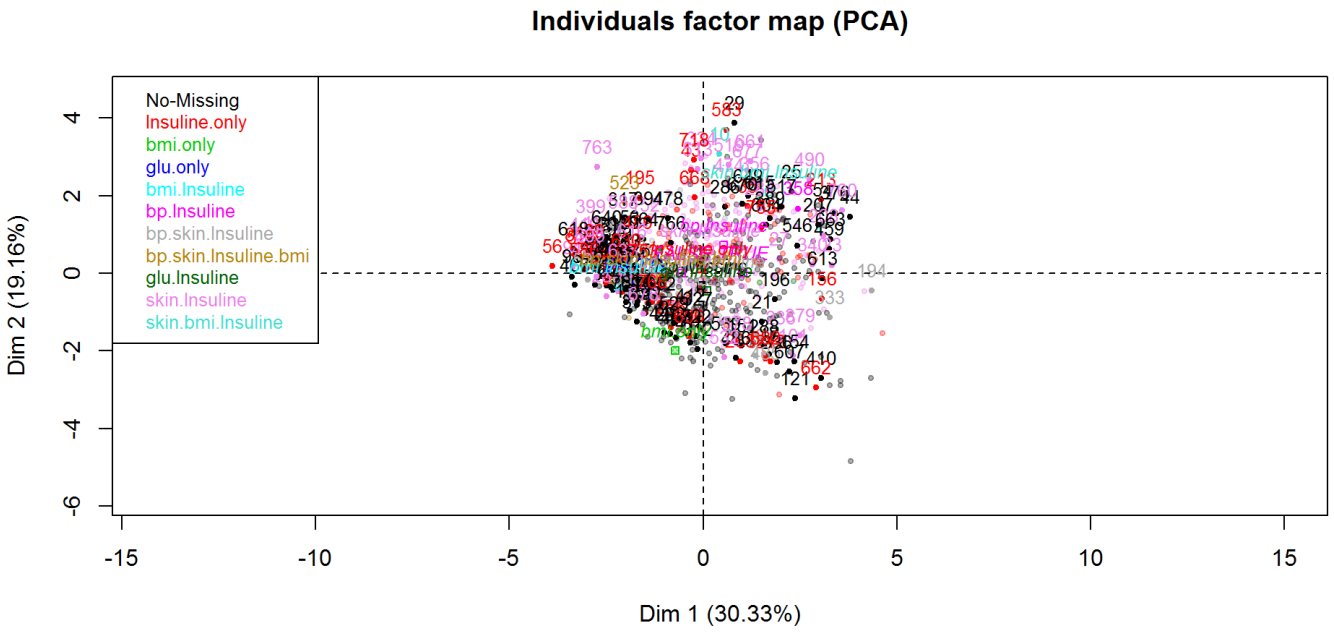


Le nuage des individus est bien centré

- Données Pima complétées avec la méthode mice du package mice

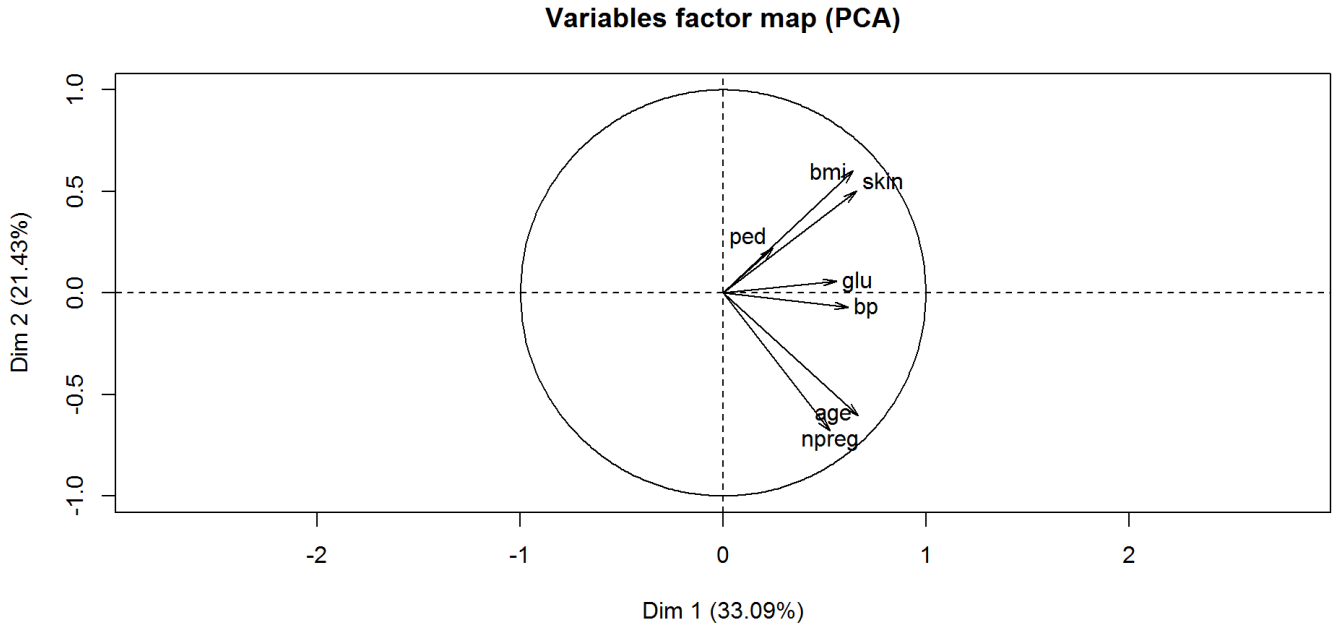

Variables factor map (PCA)
npreg
age
bp
glu
Insuline
ped
skin
bmi
Dim 2 (19.16%)
Dim 1 (30.33%)

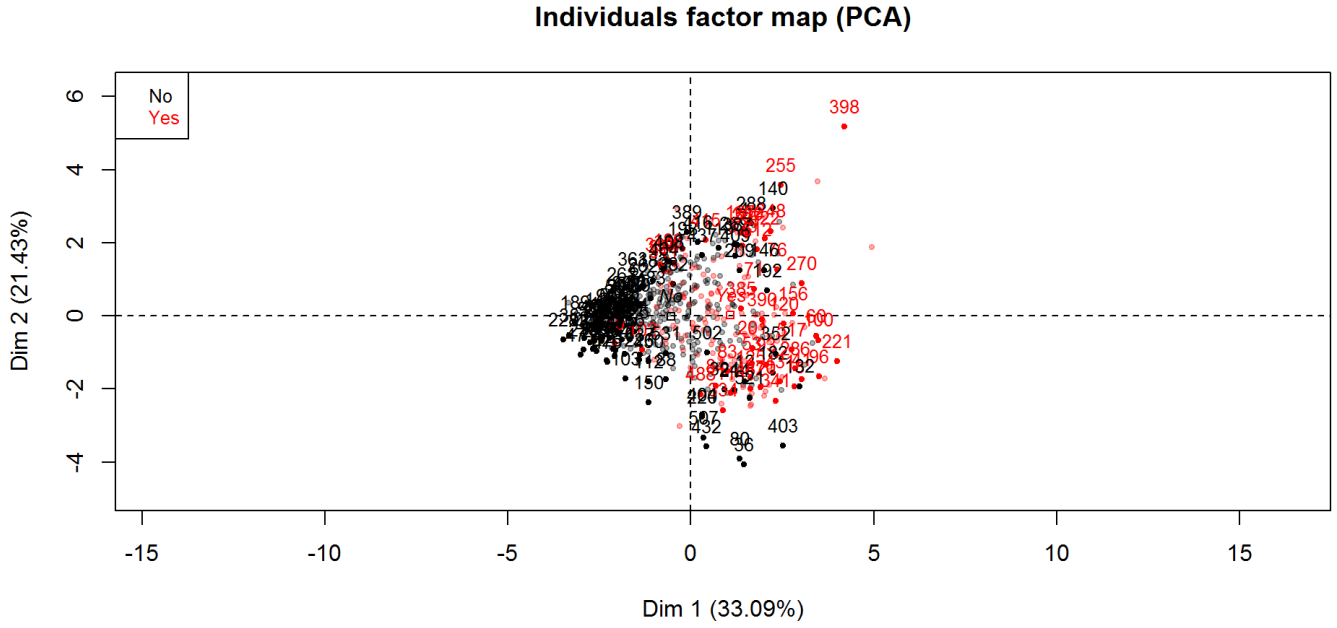

Individuals factor map (PCA)
FALSE
TRUE
Dim 2 (19.16%)
Dim 1 (30.33%)


Individuals factor map (PCA)
No-Missing
Insuline.only
bmi.only
glu.only
bmi.Insuline
bp.Insuline
bp.skin.Insuline
bp.skin.Insuline.bmi
glu.Insuline
skin.Insuline
skin.bmi.Insuline
Dim 2 (19.16%)
Dim 1 (30.33%)

Dans le cas de mice toutes données sont relativement centrées.

- Données Pima de la librairie MASS





Remarque: Les données complétés par MICE et MIPCA ont une PCA similaire. Par contre pour les données de MASS on a une transformation par rapport aux axes. Même forme mais symétrique / à l'axe principal.

3. coefficients de corrélation entre chacunes des variables et les 5 premières composantes principales

(ce qui correspond aux coordonnées des individus sur les 5 premiers axes) ##### A partir de la fonction dimdesc

```
dimdesc <- dimdesc(res.pca, proba=1e-5) #, nbelements = 16)
dimdesc
```

```
## $Dim.1
## $Dim.1$quanti
##              correlation       p.value
## skin           0.7432351 7.578981e-136
## bmi            0.7339801 7.754023e-131
## bp             0.6663183  1.011311e-99
## glu            0.6211169  3.908420e-83
## log.Insuline   0.5437096  2.710384e-60
## age            0.4328252  2.072657e-36
## npreg          0.3723705  1.140718e-26
## ped            0.1641726  4.800145e-06
##
## $Dim.1$quali
##                     R2      p.value
## type        0.2679501 7.298189e-54
## typeMissing 0.2679501 7.298189e-54
##
## $Dim.1$category
##                     Estimate      p.value
## typeMissing=TRUE   0.870416 7.298189e-54
## type=TRUE          0.870416 7.298189e-54
## typeMissing=FALSE -0.870416 7.298189e-54
## type=FALSE        -0.870416 7.298189e-54
##
##
## $Dim.2
## $Dim.2$quanti
##              correlation      p.value
## age            0.5905051 2.325475e-73
## glu            0.4876006 4.094351e-47
## npreg          0.4497885 1.610690e-39
## log.Insuline   0.3948880 4.590636e-30
## bp            -0.2049076 1.002973e-08
## skin          -0.5387835 4.914203e-59
## bmi           -0.5507535 3.953331e-62
##
## $Dim.2$quali
##                      R2      p.value
## type        0.05610182 2.968141e-11
## typeMissing 0.05610182 2.968141e-11
##
## $Dim.2$category
##                      Estimate      p.value
## typeMissing=TRUE   0.3123481 2.968141e-11
## type=TRUE          0.3123481 2.968141e-11
## typeMissing=FALSE -0.3123481 2.968141e-11
## type=FALSE        -0.3123481 2.968141e-11
##
##
## $Dim.3
## $Dim.3$quanti
##              correlation      p.value
## npreg          0.6246086 2.555349e-84
## age            0.4790428 2.601648e-45
## bp             0.3564601 1.996067e-24
## ped           -0.2808256 2.193051e-15
## glu           -0.4847836 1.626863e-46
## log.Insuline  -0.6254777 1.289097e-84
##
## $Dim.3$quali
##                      R2      p.value
## type        0.04777441 9.264595e-10
## typeMissing 0.04777441 9.264595e-10
##
## $Dim.3$category
##                     Estimate      p.value
## typeMissing=FALSE  0.276669 9.264595e-10
## type=FALSE         0.276669 9.264595e-10
```

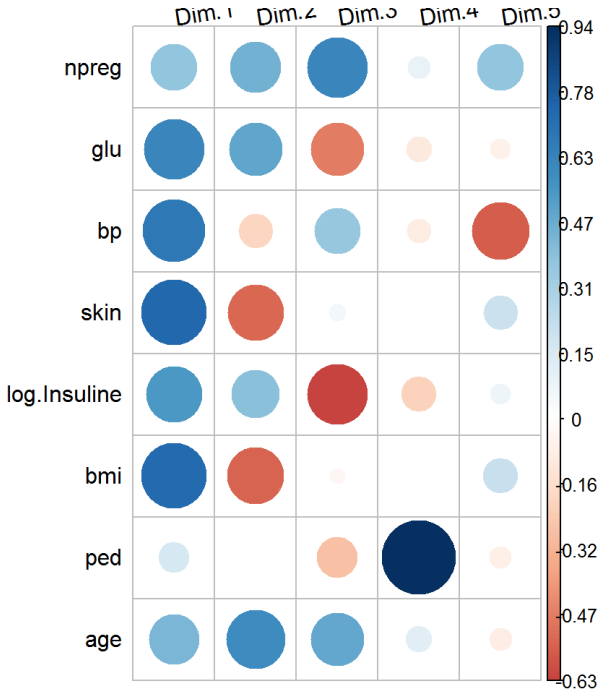
```
## typeMissing=TRUE  -0.276669 9.264595e-10
## type=TRUE         -0.276669 9.264595e-10
```

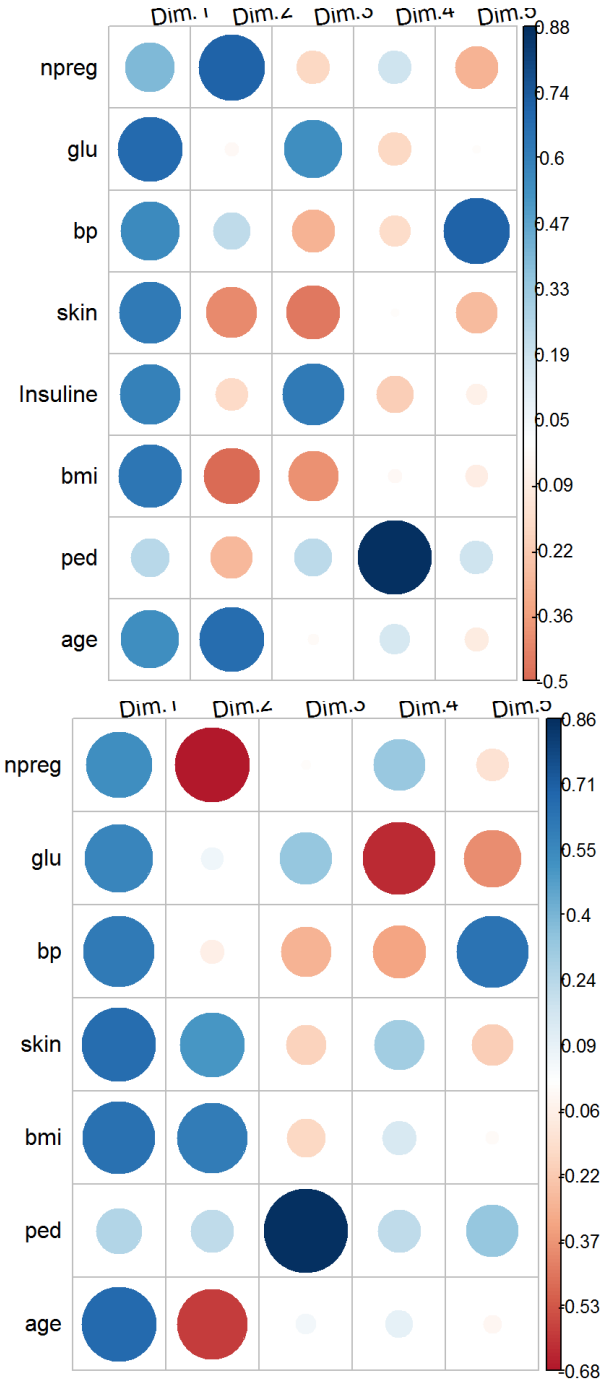
Graphique de corrélation avec corrplot

```
##              Dim.1 Dim.2 Dim.3 Dim.4 Dim.5
## npreg         0.37  0.45  0.62  0.09  0.37
## glu           0.62  0.49 -0.48 -0.11 -0.07
## bp            0.67 -0.20  0.36 -0.10 -0.57
## skin          0.74 -0.54  0.05  0.00  0.20
## log.Insuline  0.54  0.39 -0.63 -0.21  0.07
## bmi           0.73 -0.55 -0.04  0.00  0.21
## ped           0.16  0.00 -0.28  0.94 -0.08
## age           0.43  0.59  0.48  0.12 -0.09
```

```
##          Dim.1 Dim.2 Dim.3 Dim.4 Dim.5
## npreg     0.39  0.71 -0.18  0.18 -0.30
## glu       0.68 -0.03  0.54 -0.18 -0.01
## bp        0.56  0.22 -0.30 -0.16  0.71
## skin      0.62 -0.42 -0.46 -0.01 -0.28
## Insuline  0.59 -0.17  0.62 -0.22 -0.07
## bmi       0.64 -0.50 -0.40 -0.03 -0.08
## ped       0.24 -0.29  0.23  0.88  0.18
## age       0.54  0.67 -0.02  0.15 -0.09
```

```
##       Dim.1 Dim.2 Dim.3 Dim.4 Dim.5
## npreg  0.53 -0.68 -0.01  0.32 -0.13
## glu    0.56  0.06  0.33 -0.64 -0.40
## bp     0.61 -0.07 -0.30 -0.34  0.63
## skin   0.66  0.50 -0.19  0.30 -0.21
## bmi    0.64  0.60 -0.18  0.14 -0.02
## ped    0.25  0.22  0.86  0.22  0.33
## age    0.67 -0.60  0.05  0.09 -0.04
```



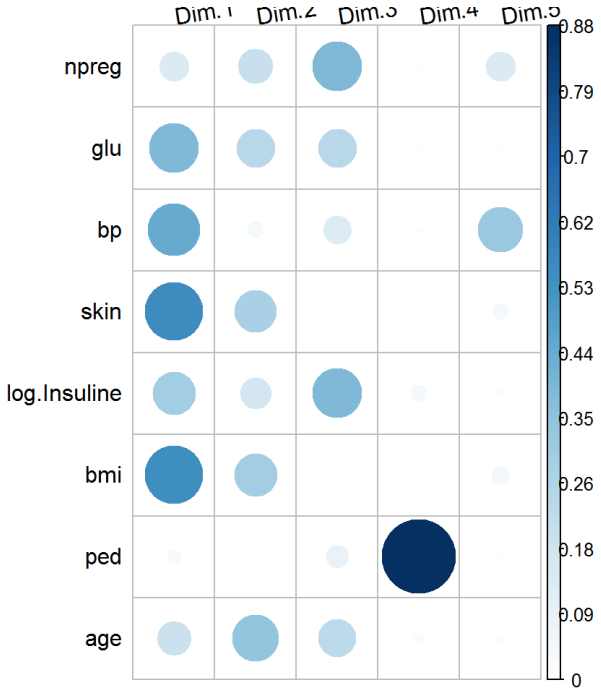


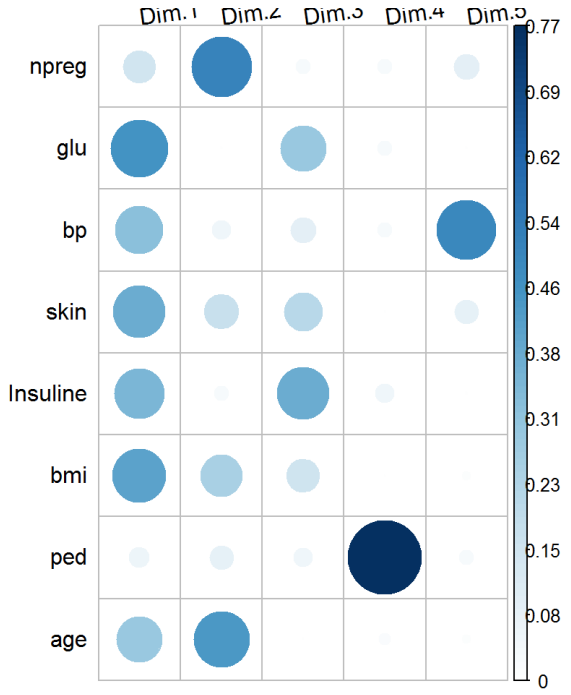
4. Indice de qualité de la représentation cos2

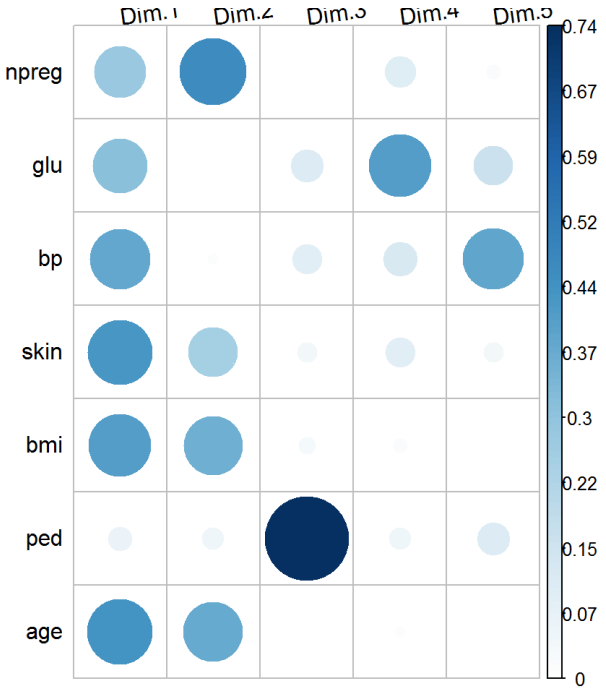
```
##              Dim.1 Dim.2 Dim.3 Dim.4 Dim.5
## npreg         0.14  0.20  0.39  0.01  0.14
## glu           0.39  0.24  0.24  0.01  0.01
## bp            0.44  0.04  0.13  0.01  0.33
## skin          0.55  0.29  0.00  0.00  0.04
## log.Insuline  0.30  0.16  0.39  0.04  0.01
## bmi           0.54  0.30  0.00  0.00  0.05
## ped           0.03  0.00  0.08  0.88  0.01
## age           0.19  0.35  0.23  0.02  0.01
```

```
##          Dim.1 Dim.2 Dim.3 Dim.4 Dim.5
## npreg     0.15  0.51  0.03  0.03  0.09
## glu       0.46  0.00  0.29  0.03  0.00
## bp        0.32  0.05  0.09  0.03  0.50
## skin      0.38  0.17  0.21  0.00  0.08
## Insuline  0.35  0.03  0.38  0.05  0.00
## bmi       0.41  0.25  0.16  0.00  0.01
## ped       0.06  0.08  0.05  0.77  0.03
## age       0.29  0.44  0.00  0.02  0.01
```

```
##        Dim.1 Dim.2 Dim.3 Dim.4 Dim.5
## npreg   0.28  0.46  0.00  0.10  0.02
## glu     0.31  0.00  0.11  0.41  0.16
## bp      0.38  0.01  0.09  0.12  0.39
## skin    0.43  0.25  0.04  0.09  0.04
## bmi     0.41  0.36  0.03  0.02  0.00
## ped     0.06  0.05  0.74  0.05  0.11
## age     0.44  0.37  0.00  0.01  0.00
```





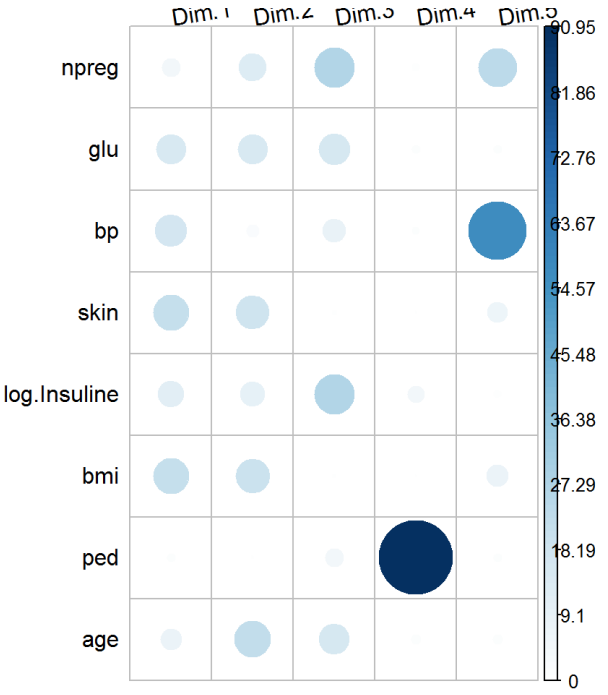


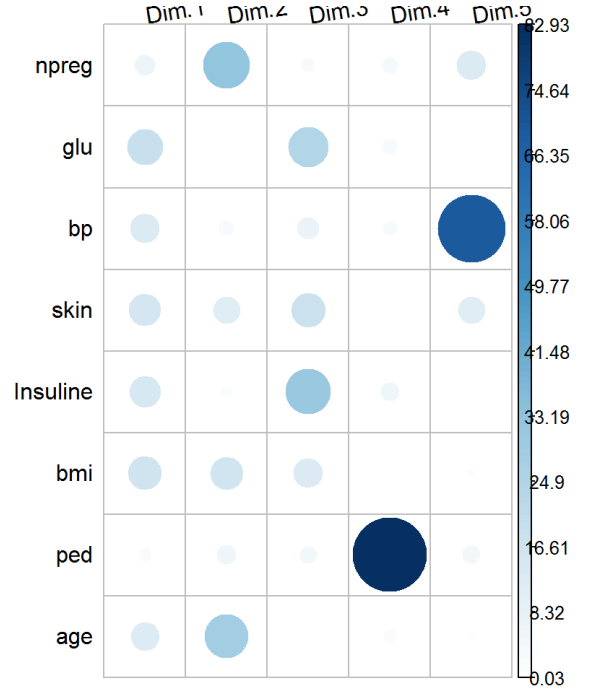
5. Contribution des variables à la construction des axes

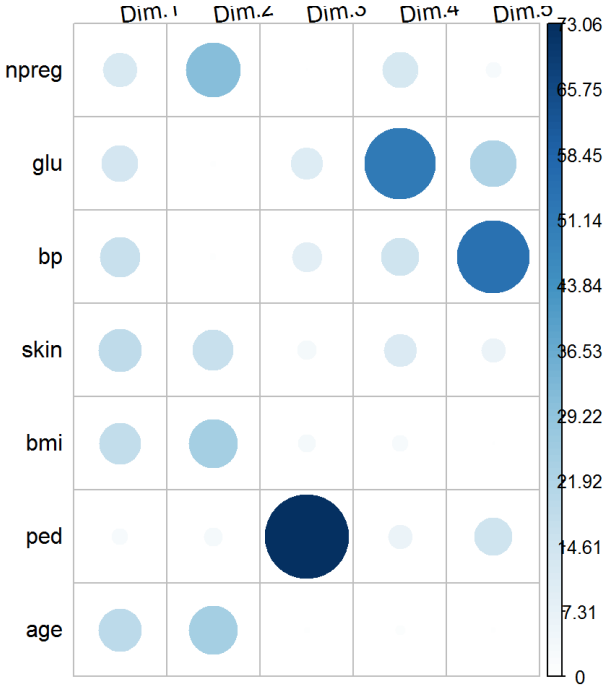
```
##              Dim.1 Dim.2 Dim.3 Dim.4 Dim.5
## npreg         5.40 12.80 26.79  0.83 24.32
## glu          15.01 15.04 16.14  1.15  0.97
## bp           17.28  2.66  8.73  0.95 56.84
## skin         21.50 18.37  0.19  0.00  6.69
## log.Insuline 11.51  9.87 26.87  4.55  0.89
## bmi          20.97 19.19  0.11  0.00  7.82
## ped           1.05  0.00  5.42 90.95  0.98
## age           7.29 22.07 15.76  1.57  1.48
```

```
##          Dim.1 Dim.2 Dim.3 Dim.4 Dim.5
## npreg     6.36 33.11  2.58  3.43 12.68
## glu      18.94  0.05 24.18  3.28  0.03
## bp       13.09  3.14  7.21  2.82 69.08
## skin     15.69 11.28 17.51  0.00 11.03
## Insuline 14.56  1.90 31.17  5.05  0.67
## bmi      16.86 16.31 13.04  0.10  0.90
## ped       2.38  5.30  4.27 82.93  4.58
## age      12.12 28.91  0.04  2.39  1.03
```

```
##       Dim.1 Dim.2 Dim.3 Dim.4 Dim.5
## npreg 11.95 30.81  0.00 13.04  2.41
## glu   13.46  0.24 10.71 51.52 22.05
## bp    16.31  0.34  9.15 14.83 54.60
## skin  18.74 16.79  3.56 11.20  5.95
## bmi   17.75 24.14  3.24  2.59  0.06
## ped    2.62  3.30 73.06  5.89 14.75
## age   19.17 24.39  0.28  0.92  0.19
```


Dim.1
Dim.2
Dim.3
Dim.4
Dim.5
npreg
glu
bp
skin
log.Insuline
bmi
ped
age
90.95
81.86
72.76
63.67
54.57
45.48
36.38
27.29
18.19
9.1
0


Dim.1
Dim.2
Dim.3
Dim.4
Dim.5
npreg
glu
bp
skin
Insuline
bmi
ped
age
82.93
74.64
66.35
58.06
49.77
41.48
33.19
24.9
16.61
8.32
0.03


Dim.1
Dim.2
Dim.3
Dim.4
Dim.5
npreg
glu
bp
skin
bmi
ped
age
73.06
65.75
58.45
51.14
43.84
36.53
29.22
21.92
14.61
7.31
0