

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 diabète, 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 incluses dans le jeu de données. Dans l'extraction, les patients sont des femmes et issues de la communauté 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 sucré génétique
- 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'agit de compléter les données manquantes par imputation.

1.4 Chargement des données - 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 identifiants. 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 manquantes: 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.

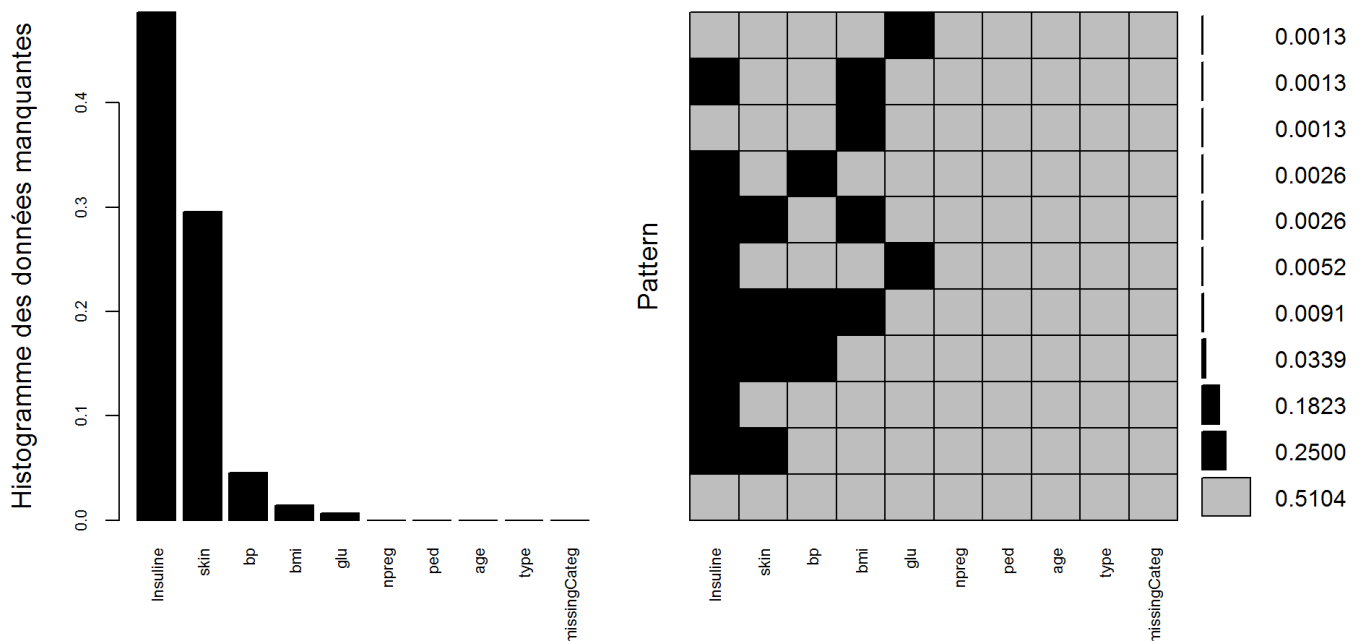
Recours aux bibliothèques MICE, MASS, VIM pour visualiser les patterns du jeu de données grâce aux fonctions `fluxplot()` ou `aggr_plot()`.

Le pourcentage de données manquantes est élevé, les informations sont exhaustives pour seulement 51% des individus, ce qui justifie 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.

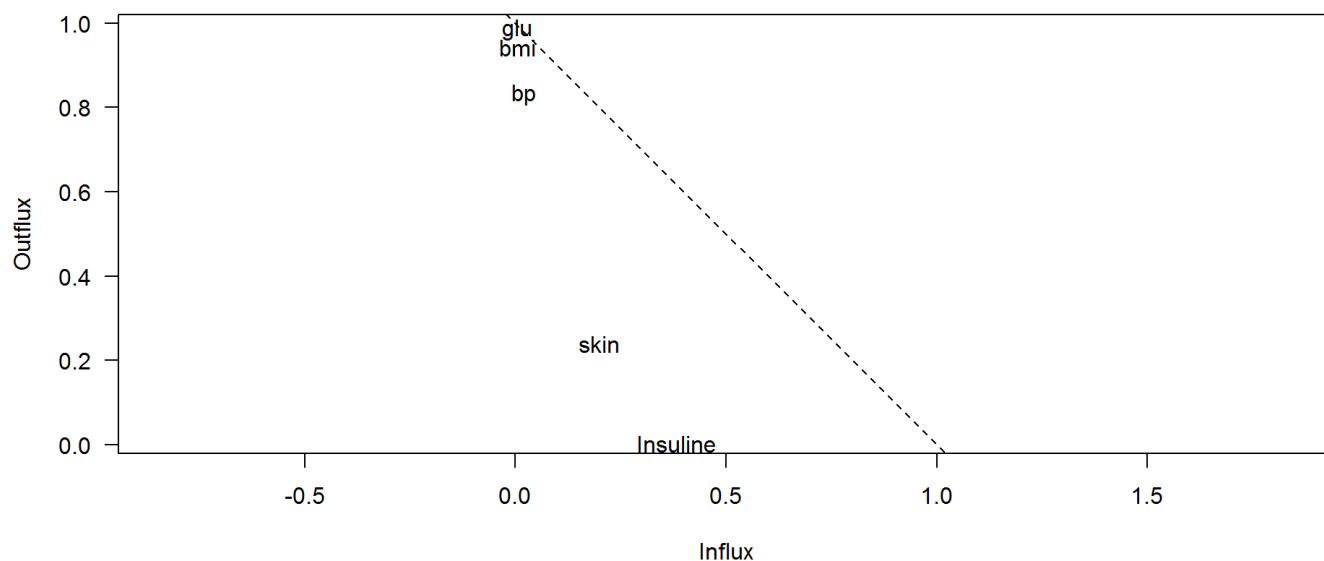
Au final, il ressort que le mécanisme des données manquantes, qui concernent 5 variables du dataframe "Pima", est non-monotone, ce qui justifiera ultérieurement le recours à l'imputation multiple (joint modeling, fully conditional specification, ACP).

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 connexion 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
```

Graphique Influx/Outflux



```
##          influx    outflux
## 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
```

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

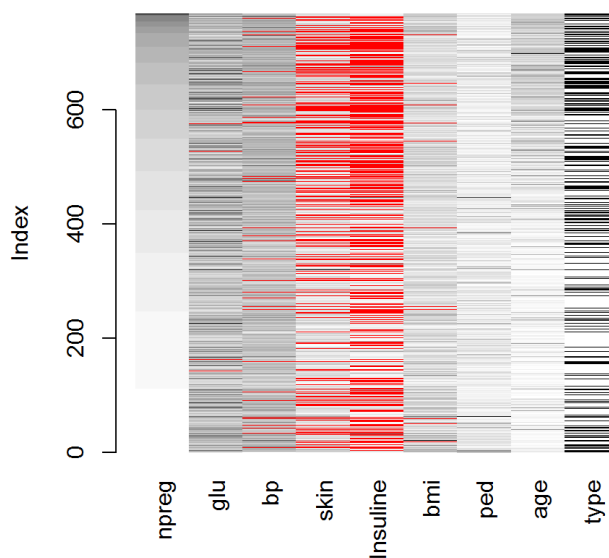
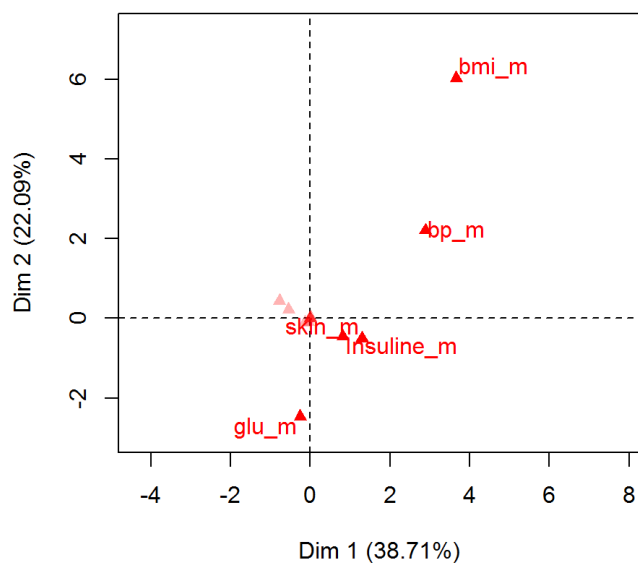
```
## [1] 8
```

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

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

On reprend la méthode d'analyse PCA des données manquantes vu en cours (chap 1) et de http://factominer.free.fr/missMDA/appendix_These_Audigier.pdf (http://factominer.free.fr/missMDA/appendix_These_Audigier.pdf)

MCA factor map

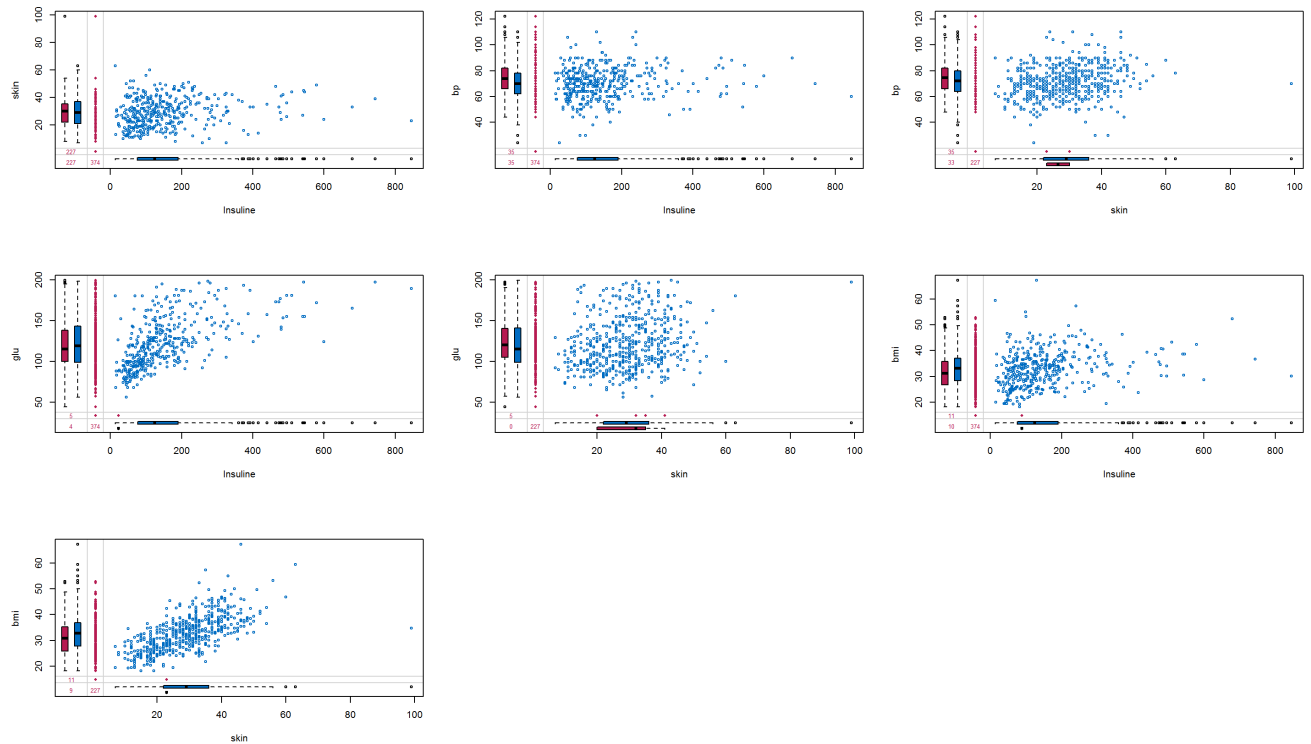


On ne remarque pas de groupes de variables bien séparés. Peut-être une liaison skin-insuline

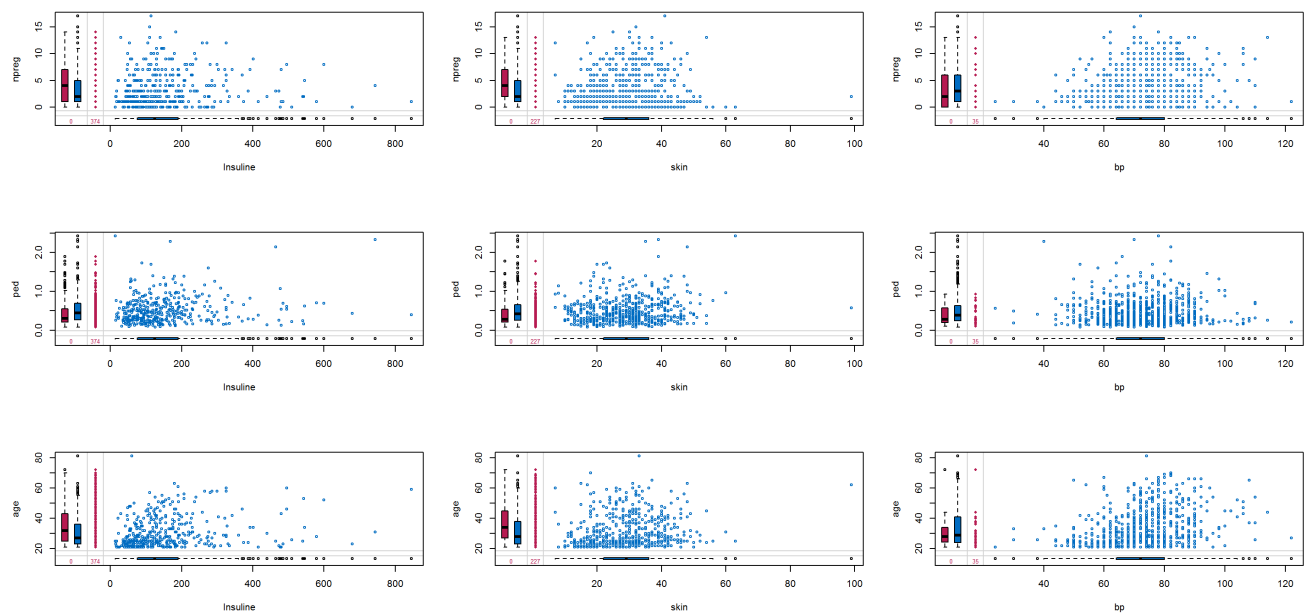
- L'hypothèse MCAR semble infirmée au regard des distributions des variables Insuline et Skin comparées aux variables complètes, npreg, age et ped. A ce stade, l'hypothèse d'ignorabilité du mécanisme peut par ailleurs être maintenue. L'analyse de sensibilité nous permettra de la valider. Nous pouvons 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 Conditional 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. Pour les graphiques cf. [annexe N°1 - graphiques N°2 - Analyses des distributions pour les Imputations simples].

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

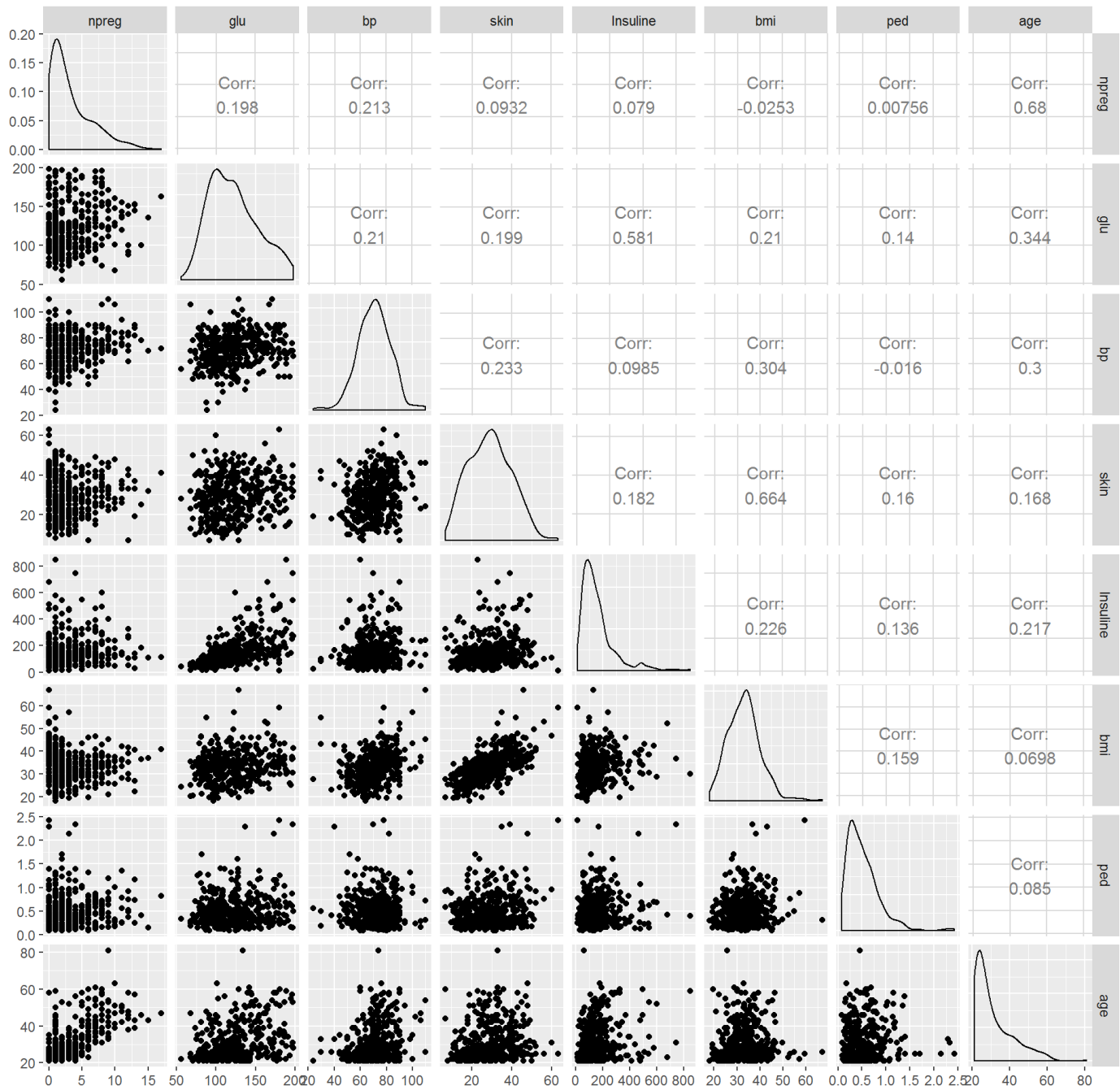
A nouveau, certaines imputations ne sont pas plausibles : avec le modèle joint les graphiques montrent des valeurs aberrantes pour l'épaisseur de peau et le taux d'insuline. S'agissant de la méthode FCS, les distributions des variables glu et bmi présentent des profils très divergents pour chacune des imputations, en raison du nombre peu élevé de données manquantes. Pour les graphiques cf. annexe N°1 - graphiques N°3 - Analyses des distributions pour les Imputations multiples.

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

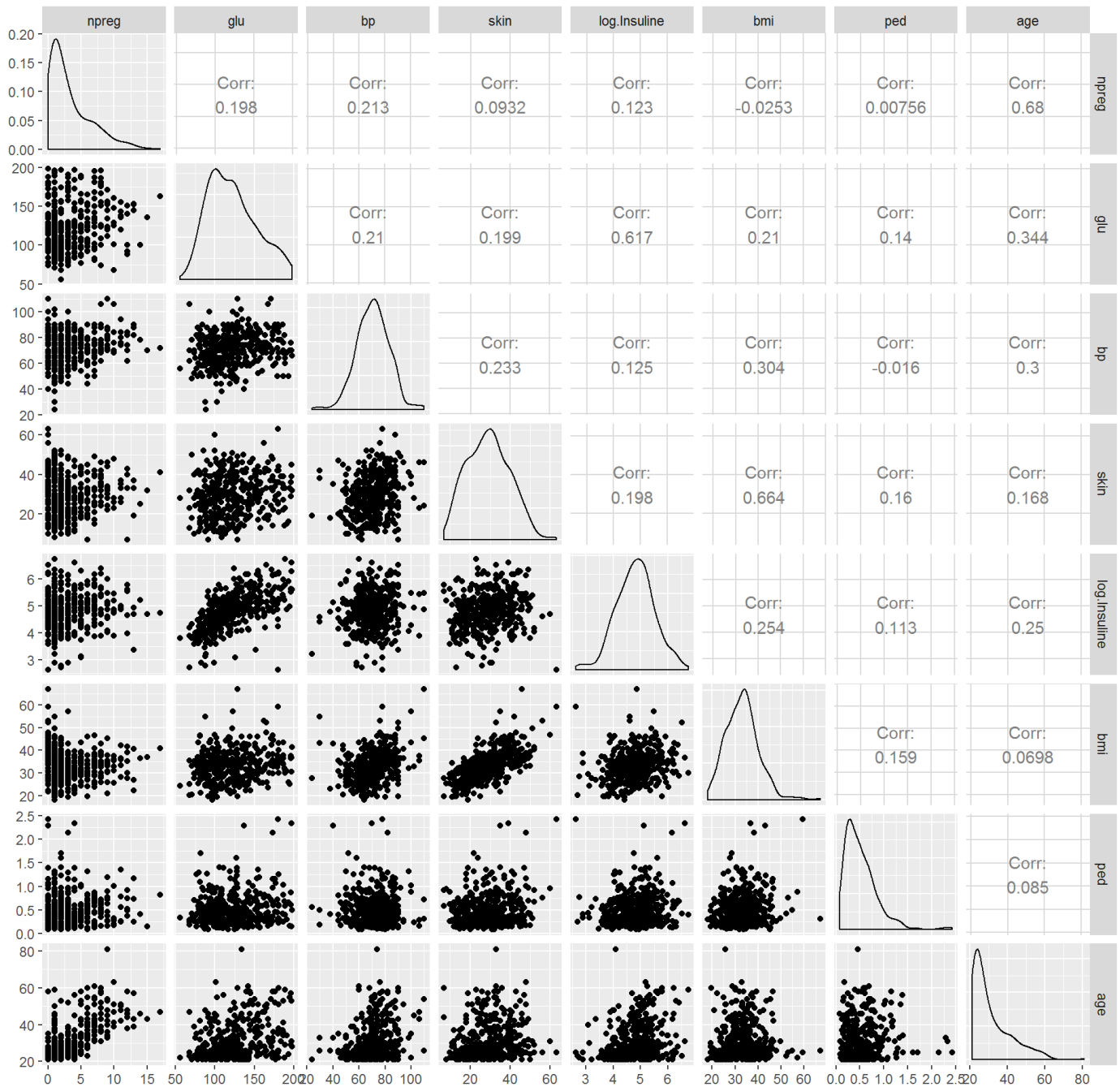
- Transformation des données

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

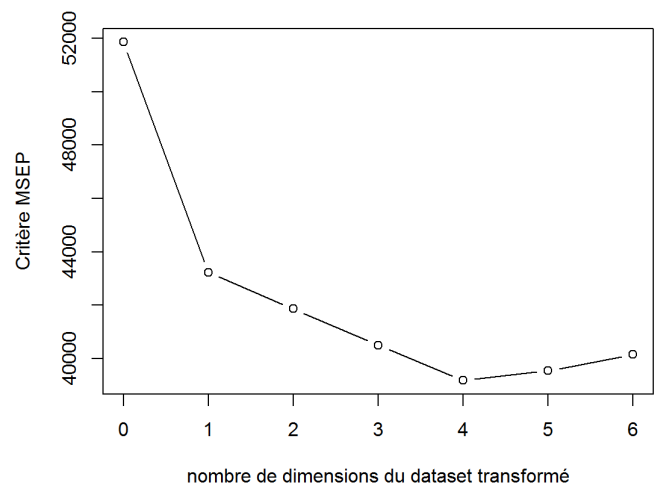
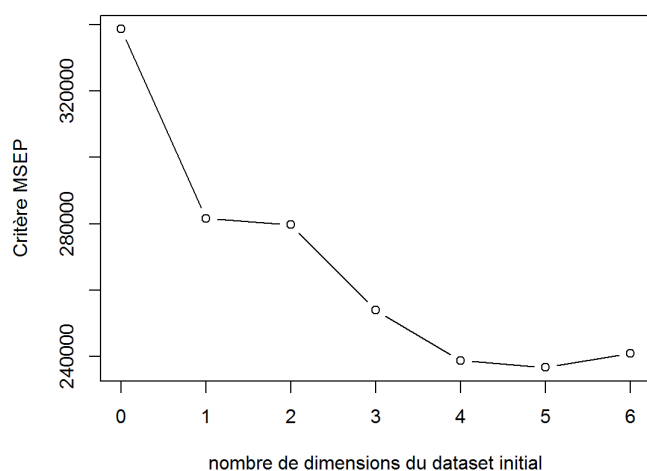


Graphiques des distributions bivariées - variables transformées



- 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".



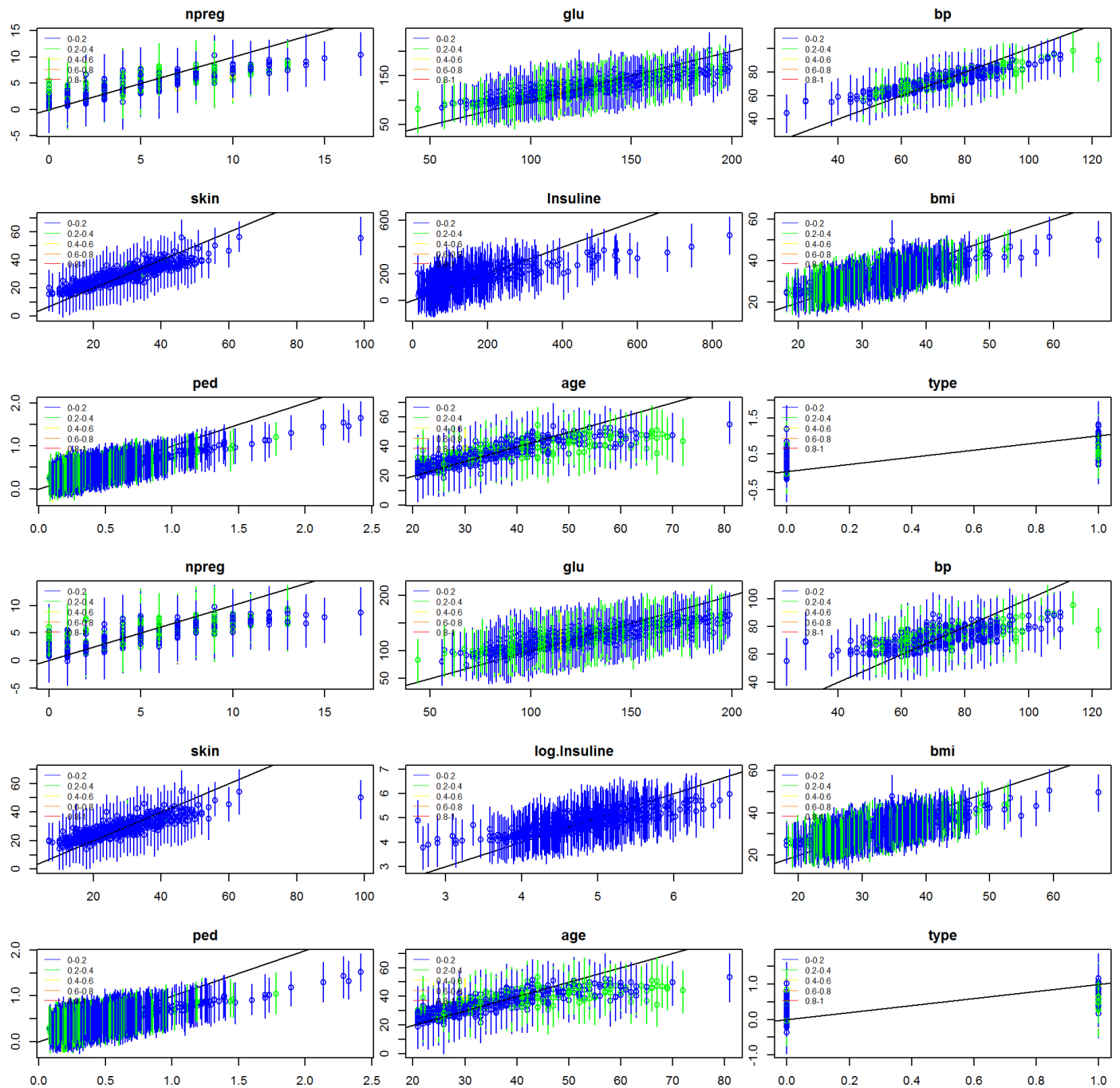

```
## [1] "Le nombre de dimensions retenu pour le jeu de données initial est de :5"
```

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

3.2.1 Imputation multiple - méthode bayésienne

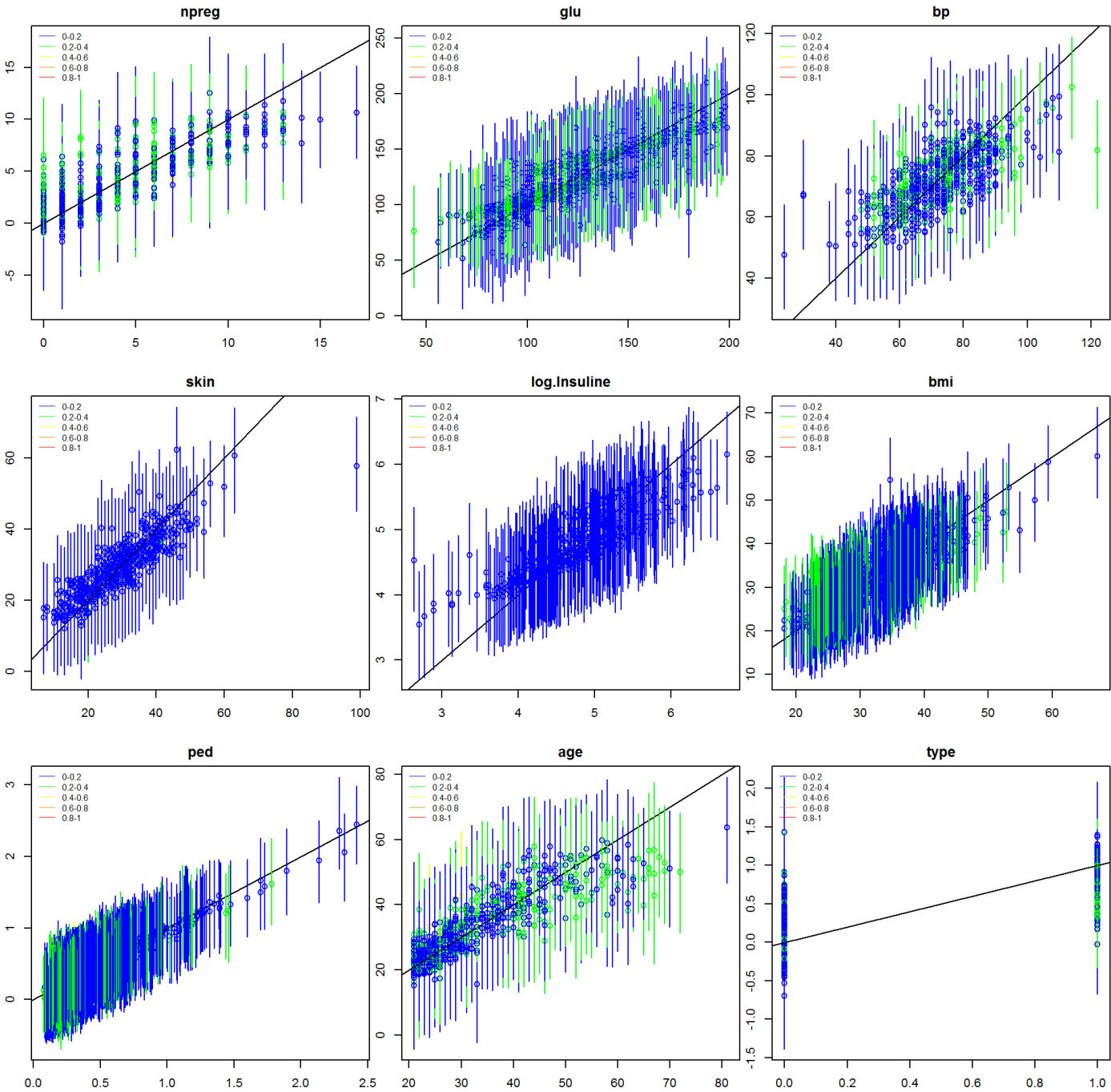
- Diagnostics pour l'imputation multiple - méthode bayésienne

Visuellement, l'imputation de la variable insuline semble de meilleure qualité, une fois celle-ci ayant fait l'objet d'une transformation en logarithmique. Le nombre d'itération pour la période de burn-in Lstart et le paramètre L n'ont pas été déterminé par faute de temps en suivant la méthode proposé dans http://factominer.free.fr/missMDA/appendix_These_Audigier.pdf (http://factominer.free.fr/missMDA/appendix_These_Audigier.pdf) Or la détermination de ces paramètres joue un rôle important dans la qualité de l'algorithme Bayésien. Ceci explique peut-être le résultat moyen obtenu avec les paramètres par défaut.



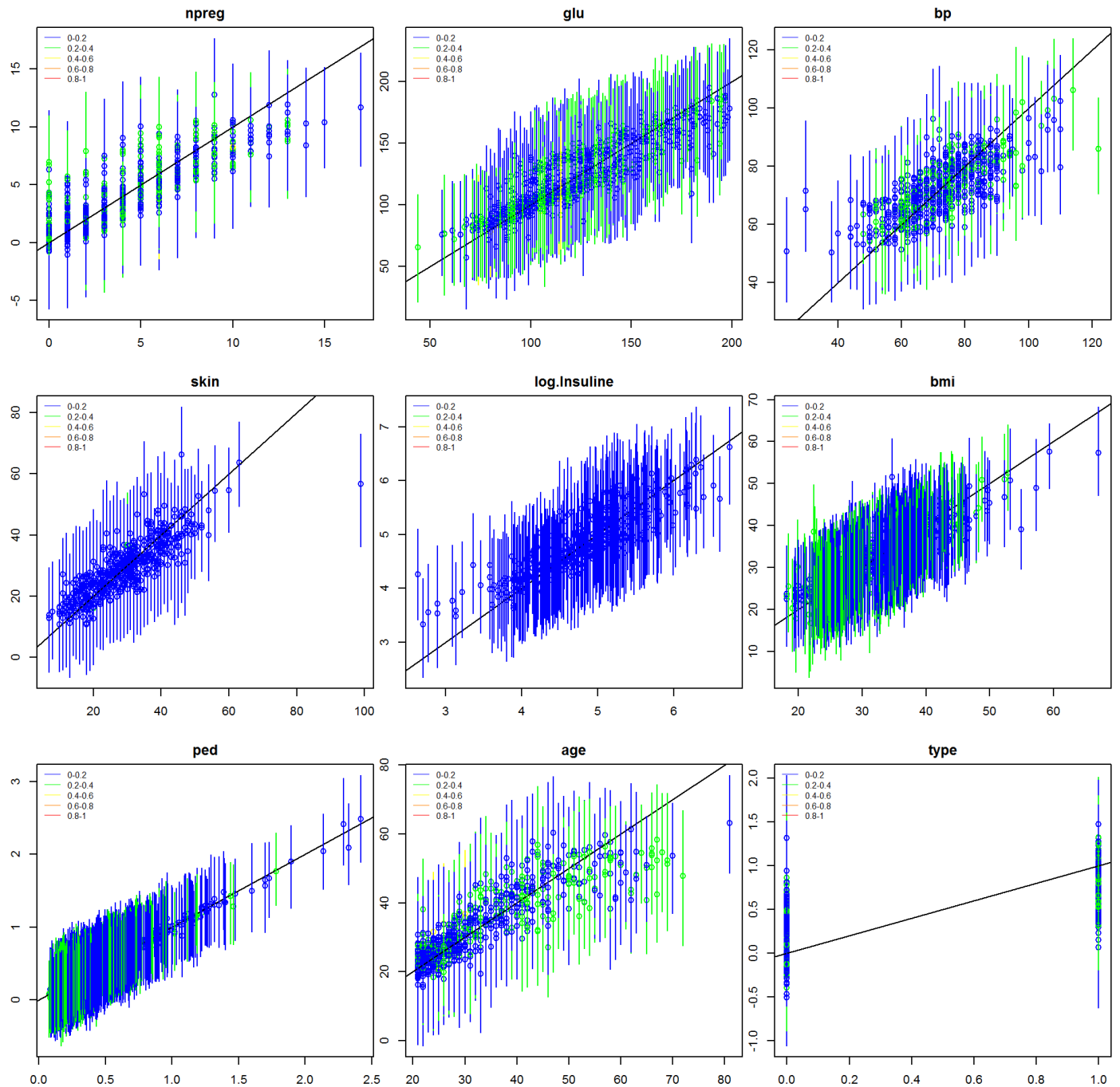
3.2.2 Multiple Imputation avec multiple imputation par bootstrap

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



3.2.3 Multiple Imputation EM et bootstrap

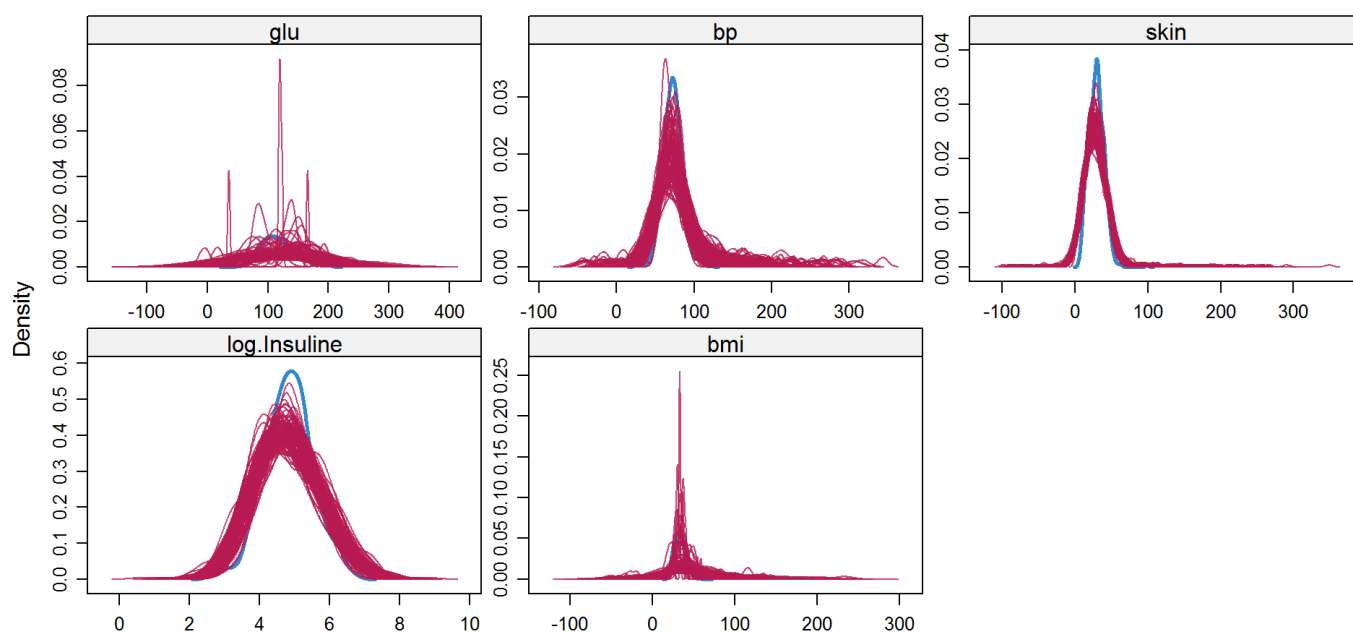
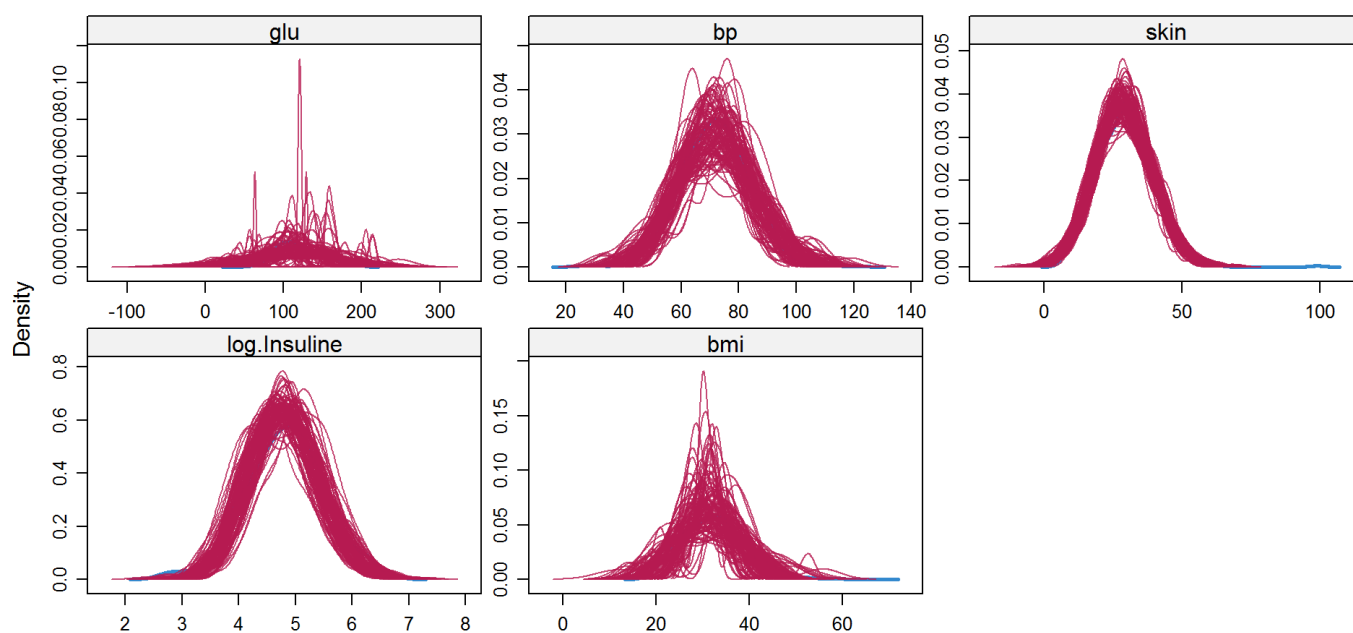
Cette dernière imputation semble *in fine* la plus appropriée. Plus de 90% des réimputations se situent dans l'intervalle de confiance.



L'analyse du résultat donné par le `plot(over.imp.mipca.EM.boot.new)` est bonne, voir graphiques N°4 - en Annexe N°1
(`plot(over.imp.mipca.EM.boot.new)`)

- Analyse des distributions

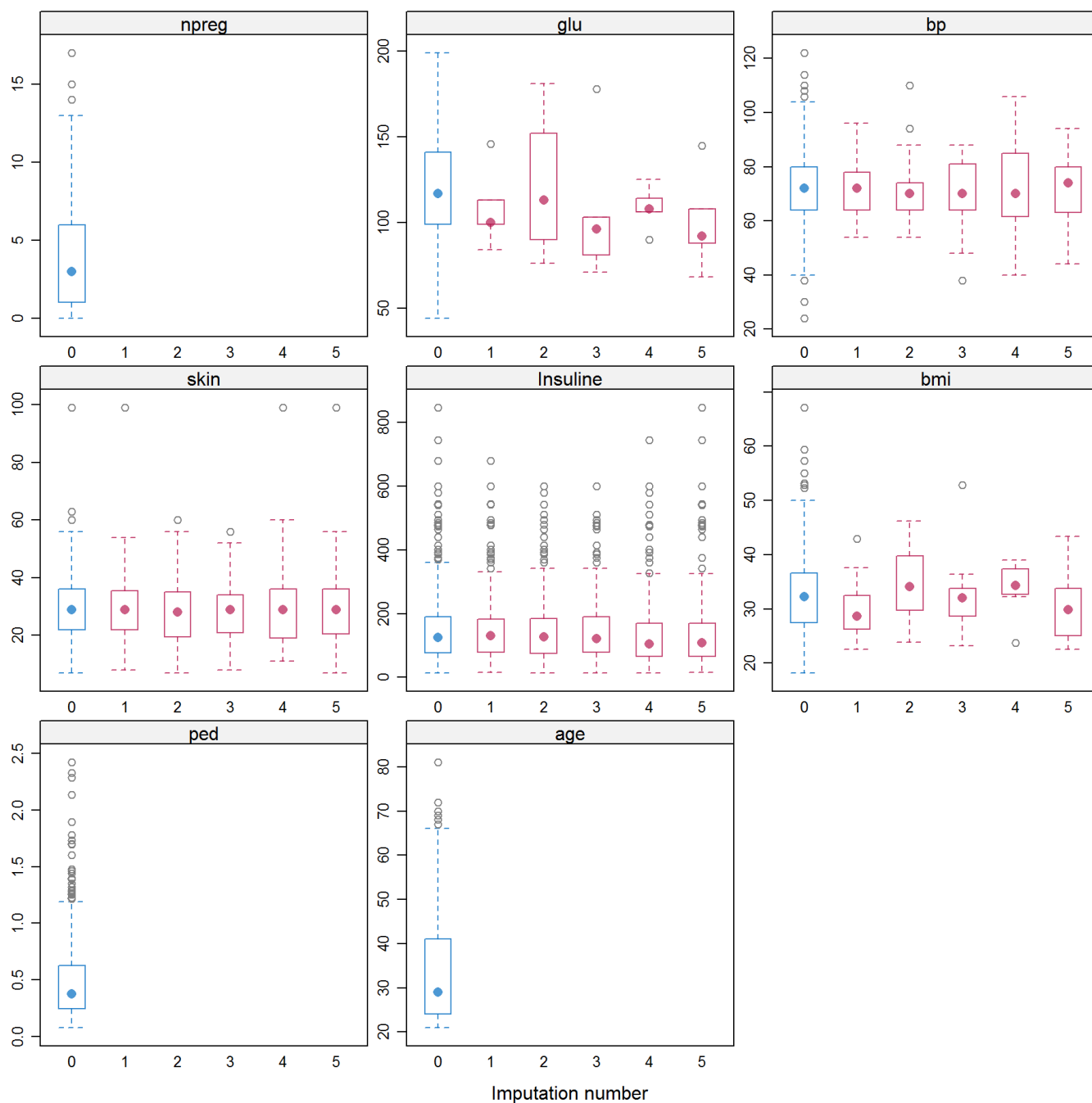
L'analyse des distributions des résultats est mise en oeuvre avec les méthodes du package "MICE". L'analyse requiert une conversion préalable des objets avant manipulation des fonctions et méthodes du package. *In fine*, ces dernières sorties semblent confirmer le diagnostic précédent. Le meilleur modèle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



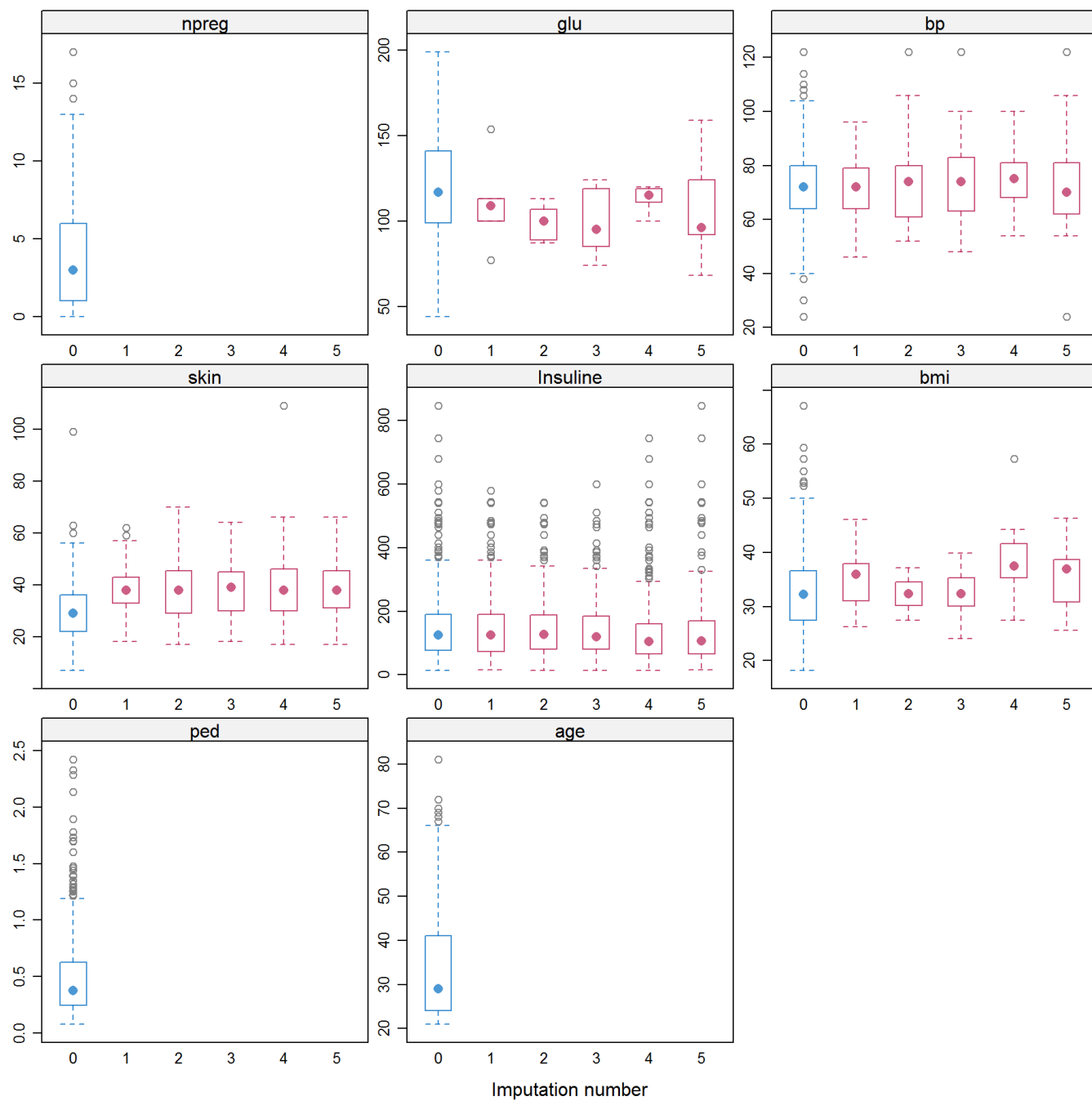
4. Analyse de sensibilité

L'analyse de sensibilité ne concerne que le modèle d'imputation multiple Avec pour hypothèse un mécanisme MAR, les imputations sont robustes à une modification des valeurs de "skin" (idem pour "Insuline"). Seulement possible avec MICE (post traitement à la volée).

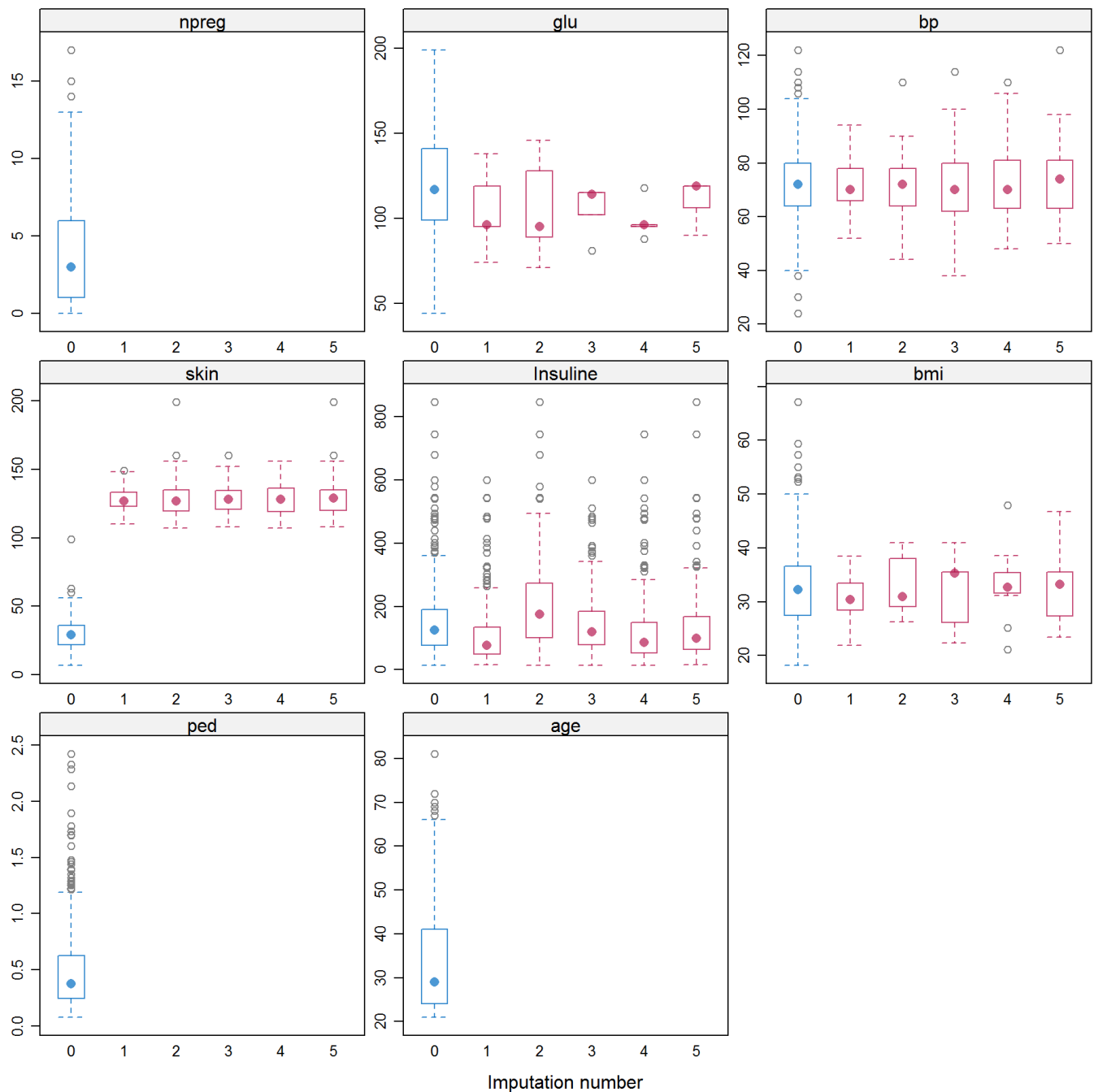
```
delta <- c(0,10,100)
imp.all <- vector("list", length(delta))
post <- mice(data[,1:8], maxit = 0)$post
for (i in 1:length(delta)){
  d <- delta[i]
  cmd <- paste("imp[[j]][,i] <- imp[[j]][,i] +", d)
  post["skin"] <- cmd
  imp <- mice(data[,1:8], post = post, maxit = 5, m=5, seed=111119, print = F)
  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.Insuline +
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.Insuline +
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	std.error	statistic	df	p.value
## (Intercept)	-11.400011546	1.187090424	-9.6033220	497.9434	0.000000e+00
## npreg	0.124011130	0.032727176	3.7892402	741.4134	1.633686e-04
## glu	0.029496982	0.004136649	7.1306467	659.7653	2.632117e-12
## bp	-0.006598732	0.008884396	-0.7427328	692.4153	4.578954e-01
## skin	0.011838423	0.013886438	0.8525169	372.0046	3.944758e-01
## log.Insuline	0.704670064	0.228798239	3.0798754	321.4320	2.249570e-03
## bmi	0.074139556	0.019743145	3.7552050	590.2862	1.903538e-04
## ped	0.828422502	0.301069750	2.7515966	735.8316	6.076469e-03
## age	0.011327311	0.009760453	1.1605312	731.1827	2.462114e-01

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

##	estimate	std.error	statistic	df	p.value
## (Intercept)	-12.141289302	1.189505730	-10.2070036	428.8726	0.000000e+00
## npreg	0.140332598	0.034593919	4.0565684	700.0002	5.540094e-05
## glu	0.023454826	0.004352253	5.3891223	624.2180	1.004849e-07
## bp	-0.008253068	0.009439432	-0.8743183	548.3962	3.823277e-01
## skin	0.009650793	0.012304159	0.7843521	417.5849	4.332782e-01
## log.Insuline	1.041124980	0.198733597	5.2387970	329.5332	2.889540e-07
## bmi	0.064975977	0.020883203	3.1113990	446.0616	1.981560e-03
## ped	1.019659026	0.315405391	3.2328522	654.1950	1.287183e-03
## age	0.015479336	0.010424886	1.4848446	646.7768	1.380722e-01

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

##	estimate	std.error	statistic	df	p.value
## (Intercept)	-8.459868105	0.869975039	-9.7242653	444.9883	0.0000000000
## npreg	0.120123947	0.032441664	3.7027677	750.7235	0.0002288809
## glu	0.037369472	0.003629265	10.2967055	735.9130	0.0000000000
## bp	-0.013625841	0.009035287	-1.5080695	536.0114	0.1321258969
## skin	0.007388408	0.012061235	0.6125747	375.3826	0.5405284843
## bmi	0.075228893	0.020525581	3.6651285	411.4270	0.0002795734
## ped	0.889040495	0.297423532	2.9891397	750.2212	0.0028890333
## age	0.014807704	0.009557484	1.5493307	744.2688	0.1217273294

On remarque que les variables 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] 1.996417e-07
```

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

##	test statistic	df1	df2	df.com	p.value	riv
## 1 ~ 2	27.44499	1	406.1332	759	2.601565e-07	0.4839518

Au regard des différentes comparaisons et du test final, on conserve, tant pour la méthode d'IM FCS (package MICE) que celle utilisant l'algorithme EM et le bootstrap (package missMDA), des modèles excluant les variables : skin, bp et age

```
## [1] "Test du rapport de vraisemblance modele complet vs modele sans la variable age : 0.124473399132631"
```


[1] "Test du rapport de vraisemblance modele complet vs modele sans la variable skin : 0.455756953252064"

[1] "Test du rapport de vraisemblance modele complet vs modele sans la variable insuline : 1.99641684295315e-07"

[1] "Test du rapport de vraisemblance modele complet vs modele sans la variable bp : 0.333629538015072"

[1] "Test du rapport de vraisemblance modele complet vs modele sans la variable bmi : 0.00446421960975407"

[1] "Test du rapport de vraisemblance modele complet vs modele sans la variable glu : 1.16584519815888e-12"

[1] "Test du rapport de vraisemblance modele complet vs modele sans la variable npreg : 4.49075084600903e-05"

[1] "Test du rapport de vraisemblance modele complet vs modele sans la variable ped : 0.000531896526746678"

[1] "AIC du modele complet : 686.575215017162"

[1] "AIC du modele sans la variable age : 687.090830762315"

[1] "AIC du modele sans la variable skin : 685.724748799093"

[1] "AIC du modele sans la variable insuline : 730.287720998959"

[1] "AIC du modele sans la variable bp : 685.690965915181"

[1] "AIC du modele sans la variable bmi : 697.79003435603"

[1] "AIC du modele sans la variable npreg : 702.820152174833"

[1] "AIC du modele sans la variable glu : 719.671491942039"

[1] "AIC du modele sans la variable ped : 696.526861899278"

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

[1] 685.691

[1] 685.6757

[1] 684.555

[1] 729.3403

[1] "Test du rapport de vraisemblance modele sans bp vs modele sans les variables bp et age : 0.18949780227073"

[1] "Test du rapport de vraisemblance modele sans bp et age vs modele sans bp, age et skin : 0.555176590089067"

[1] "Test du rapport de vraisemblance modele sans bp, age et skin vs modele sans bp, age, Insuline et skin : 8.75602027550215e-08"

	estimate	std.error	statistic	df	p.value
## (Intercept)	-12.479248405	1.106235162	-11.2808278	453.7552	0.000000e+00
## age	0.013263044	0.010132034	1.3090209	641.9648	1.909957e-01
## bmi	0.060863847	0.021292570	2.8584547	364.4331	4.501764e-03
## npreg	0.138274791	0.034439597	4.0149944	700.1340	6.586607e-05
## ped	1.042752647	0.314650927	3.3139983	640.2016	9.715804e-04
## glu	0.022993400	0.004288803	5.3612627	634.2586	1.158738e-07
## skin	0.007630877	0.012166366	0.6272109	400.0445	5.308788e-01
## log.Insuline	1.053970263	0.198096237	5.3204961	326.2683	1.926899e-07

	estimate	std.error	statistic	df	p.value
## (Intercept)	-12.168695446	1.080641634	-11.2606206	428.4867	0.000000e+00
## bmi	0.059180105	0.021159313	2.7968822	366.1879	5.432202e-03
## npreg	0.162162430	0.030006541	5.4042360	673.5206	9.041200e-08
## ped	1.052235683	0.314040127	3.3506409	641.1928	8.535679e-04
## glu	0.024051566	0.004237509	5.6758733	625.7047	2.112149e-08
## skin	0.007644363	0.012077137	0.6329615	405.8462	5.271153e-01
## log.Insuline	1.047312343	0.199241984	5.2564842	317.0080	2.703495e-07

	estimate	std.error	statistic	df	p.value
## (Intercept)	-12.20632757	1.078406739	-11.318853	427.1825	0.000000e+00
## bmi	0.06773580	0.017120111	3.956505	271.1527	9.717394e-05
## npreg	0.16331569	0.029951166	5.452732	675.6848	6.962027e-08
## ped	1.05694573	0.313494598	3.371496	644.8432	7.922767e-04
## glu	0.02410426	0.004226212	5.703514	629.8761	1.805863e-08
## log.Insuline	1.04096781	0.198196196	5.252209	320.0277	2.746890e-07

	estimate	std.error	statistic	df	p.value
## (Intercept)	-8.74941433	0.754614557	-11.594548	477.0087	0.000000e+00
## bmi	0.07191867	0.017895939	4.018714	225.0154	7.982653e-05
## npreg	0.13747715	0.027484392	5.002008	752.9991	7.065789e-07
## ped	0.93140018	0.294664145	3.160887	750.6905	1.636076e-03
## glu	0.03772112	0.003510608	10.744899	747.0039	0.000000e+00

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

```
pool.glm5<-pool(diabete.insuline.glm5)
summary(pool.glm5)
```

	estimate	std.error	statistic	df	p.value
## (Intercept)	-12.20632757	1.078406739	-11.318853	427.1825	0.000000e+00
## bmi	0.06773580	0.017120111	3.956505	271.1527	9.717394e-05
## npreg	0.16331569	0.029951166	5.452732	675.6848	6.962027e-08
## ped	1.05694573	0.313494598	3.371496	644.8432	7.922767e-04
## glu	0.02410426	0.004226212	5.703514	629.8761	1.805863e-08
## log.Insuline	1.04096781	0.198196196	5.252209	320.0277	2.746890e-07

```
pool.glm5
```

```
## Class: mipo      m = 100
##               estimate      ubar      b      t dfcom
## (Intercept) -12.20632757 8.793431e-01 2.808099e-01 1.162961e+00 762
## bmi          0.06773580 1.789290e-04 1.130388e-04 2.930982e-04 762
## npreg        0.16331569 8.295860e-04 6.681816e-05 8.970723e-04 762
## ped          1.05694573 8.878415e-02 9.400705e-03 9.827886e-02 762
## glu          0.02410426 1.595663e-05 1.885382e-06 1.786087e-05 762
## log.Insuline 1.04096781 2.607192e-02 1.307902e-02 3.928173e-02 762
##               df      riv      lambda      fmi
## (Intercept) 427.1825 0.32253401 0.24387578 0.24739114
## bmi          271.1527 0.63806966 0.38952535 0.39397889
## npreg        675.6848 0.08134942 0.07522954 0.07795472
## ped          644.8432 0.10694152 0.09660991 0.09939882
## glu          629.8761 0.11933818 0.10661494 0.10943820
## log.Insuline 320.0277 0.50666827 0.33628389 0.34039324
```

En reprenant http://factominer.free.fr/missMDA/appendix_These_Audigier.pdf (http://factominer.free.fr/missMDA/appendix_These_Audigier.pdf) (page 20) à propos de la sortie de la fonction pool. La colonne (fmi) (fraction of missing information) peut être interprété comme la part de variabilité due aux valeurs manquantes. De grande valeurs (>5) indique que le résultat est sensible à la méthode MI utilisée. Ici ce n'est pas le cas.

Au final, les estimations des coefficients sont assez proches en comparaison de la méthode des cas concrets (= listwise deletion). Mais de manière général (sauf pour glu) sont plus faibles.

While the point estimates are close to the ones obtained by MI since the mechanism is MCAR?

```
##
## Call:
## glm(formula = type ~ bmi + npreg + ped + glu + I(log(Insuline)),
##      family = binomial(link = "logit"), data = data)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -3.0230  -0.6611  -0.3869   0.6425   2.5525
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -9.773714   1.264774  -7.728 1.10e-14 ***
## bmi           0.076625   0.020993   3.650 0.000262 ***
## npreg         0.155902   0.041178   3.786 0.000153 ***
## ped           1.218329   0.423205   2.879 0.003992 **
## glu           0.036956   0.005743   6.435 1.23e-10 ***
## I(log(Insuline)) 0.117806   0.256832   0.459 0.646459
## ---
## 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: 348.51  on 386  degrees of freedom
## (376 observations deleted due to missingness)
## AIC: 360.51
##
## Number of Fisher Scoring iterations: 5
```

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
```

6. Conclusion

Intérêt : Les données Pima sont très utilisées (cf. Kaggle). Pour autant il ne semble pas qu'un modèle d'imputation 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. Les méthodes sont nombreuses et nous n'avons pas pu toutes les expérimentés. Notamment les méthodes bayésiennes de mipCA en ce qui concerne la détermination des paramètres. Ainsi que le package AMELIA couplé avec Zelig qui n'ont pas été utilisés. Une question est apparue aussi, concernant l'utilisation pratique des données imputés. De nombreuses méthodes R nécessitent en input un dataFrame ou équivalent. Ici les méthodes utilisant du bootstrapping, on a pour

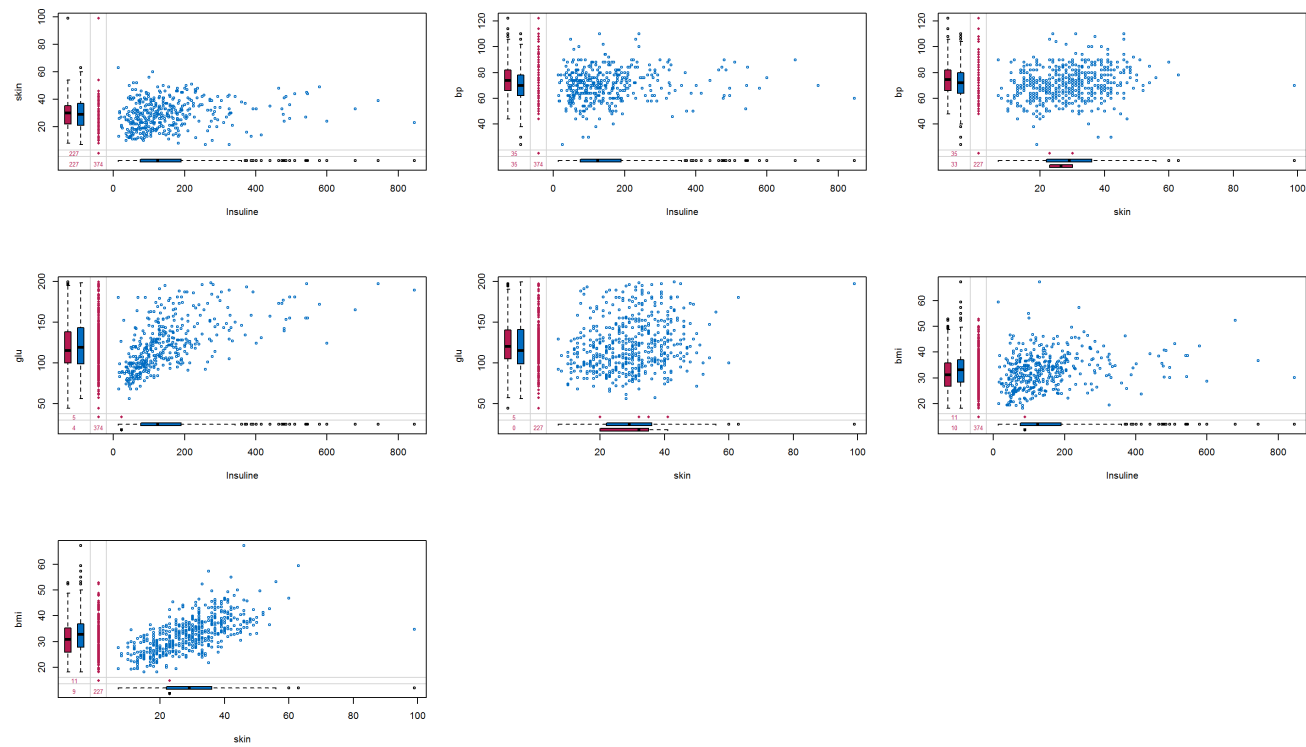
résultat une famille de structure comparable à un `dataFrame`. Que faut-il faire pour se ramener à un `dataFrame`: utiliser la moyenne? Ou par exemple dans le cas la méthode `mipca` utiliser la variable `$res.InputPCA` c'est ce qui est fait en Annexe 2 - pour utiliser les méthodes de choix de modèles et en annexe §3 pour une analyse PCA des données complétées.

Annexes

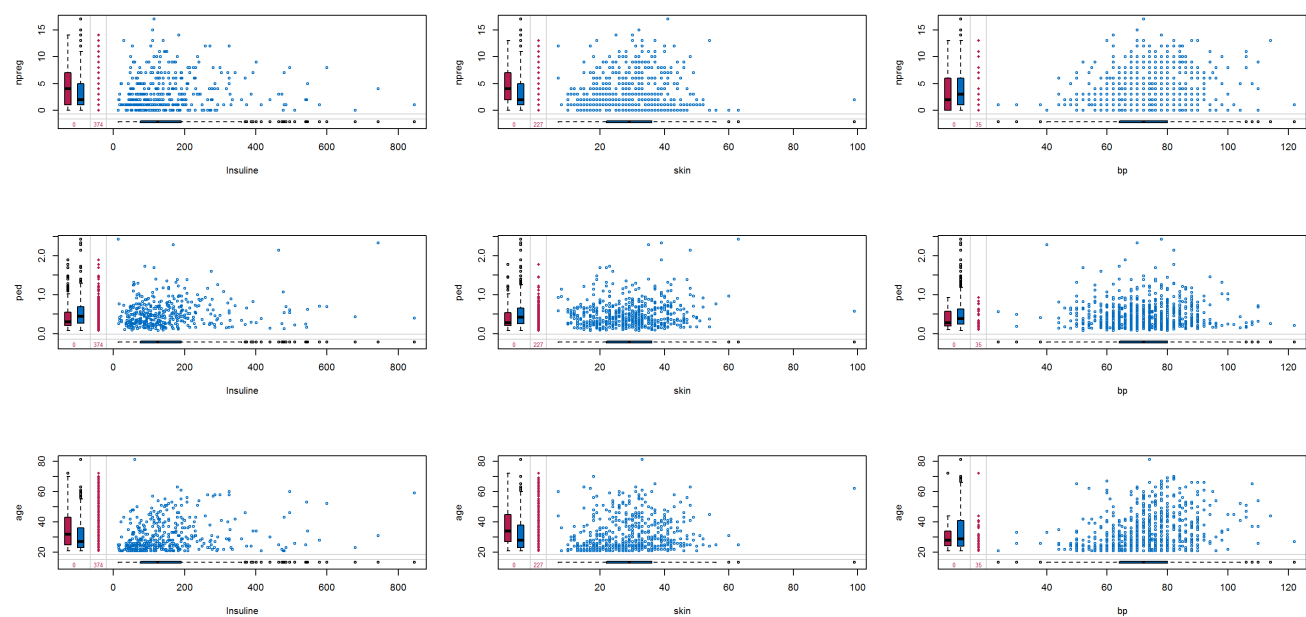
Annexe 1 : Graphiques

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

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



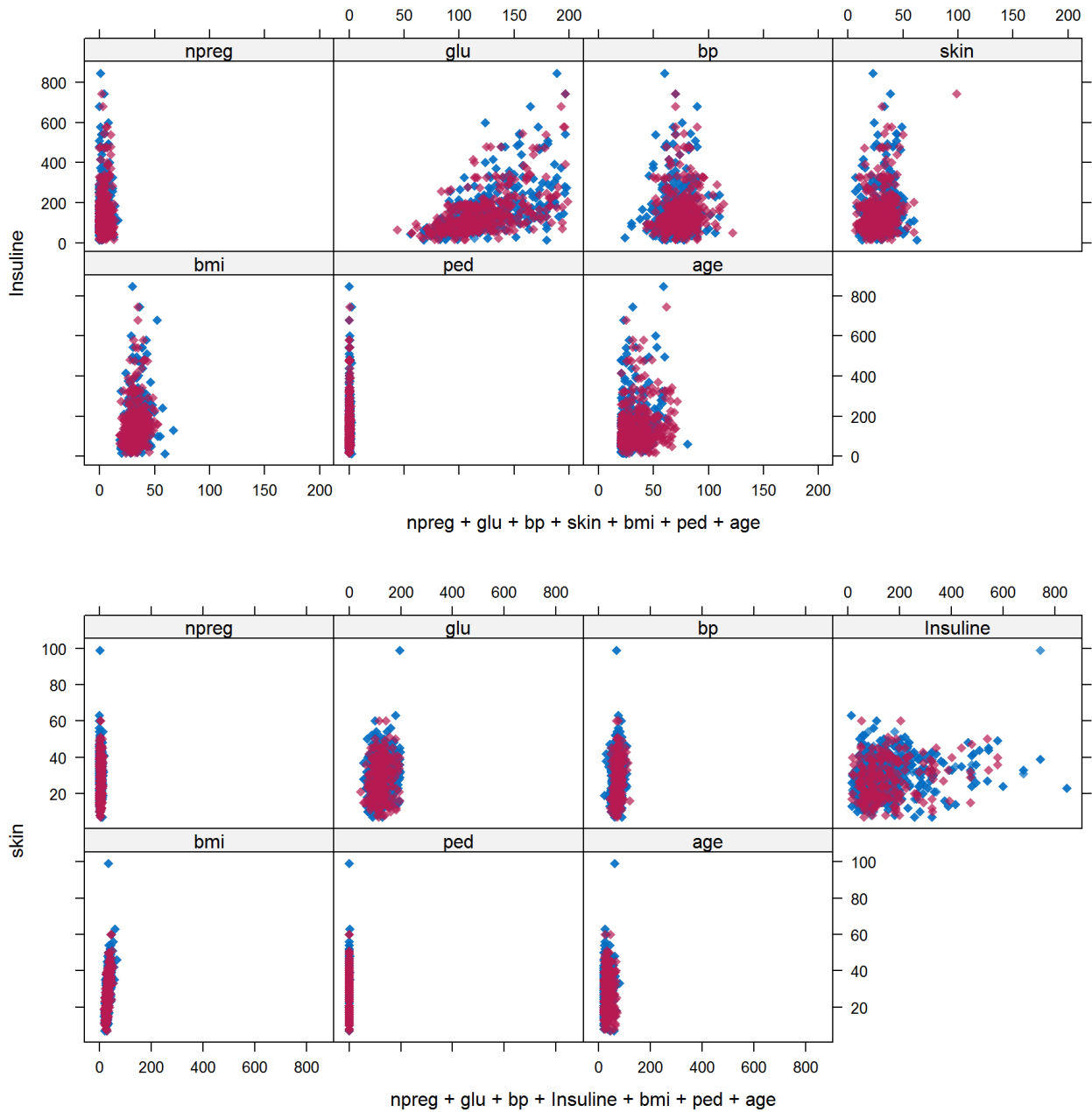
- Distributions marginales des mêmes variables avec les variables complètes

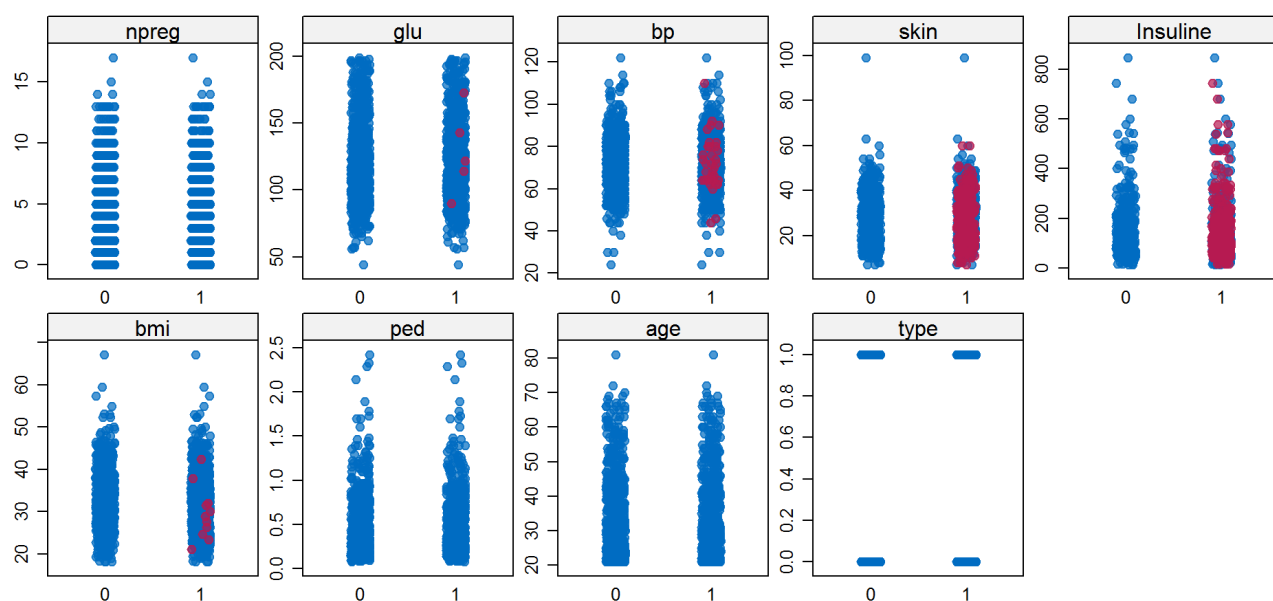
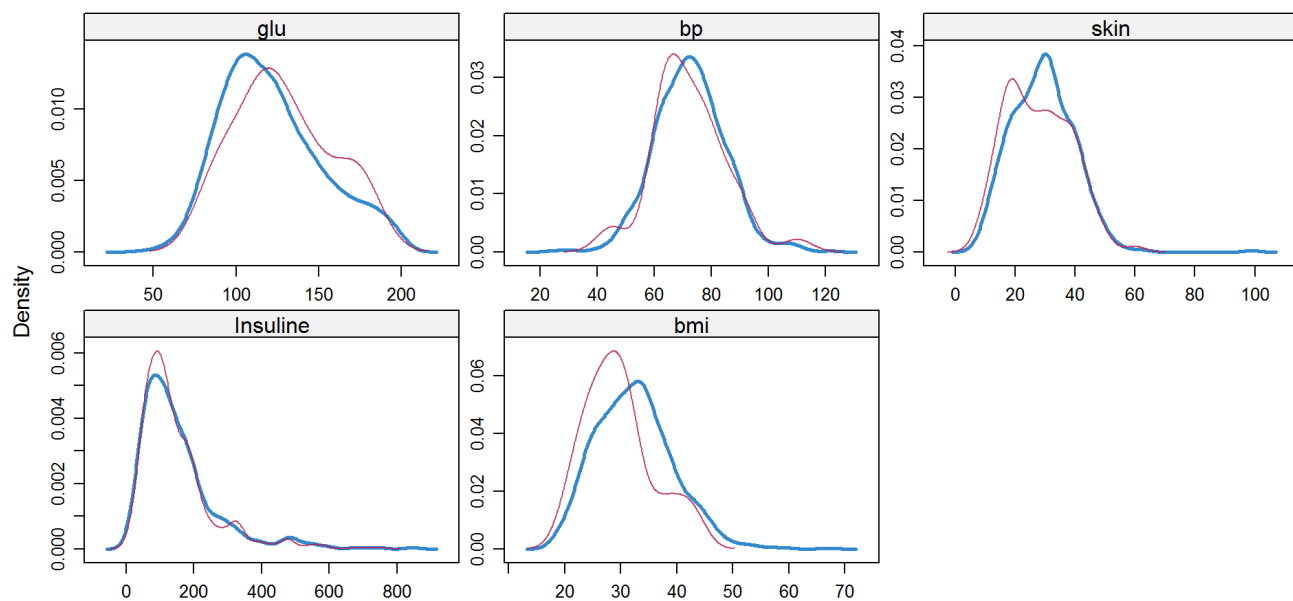


graphiques N°2 - Analyses des distributions pour les Imputations simples (§-3.1.3)

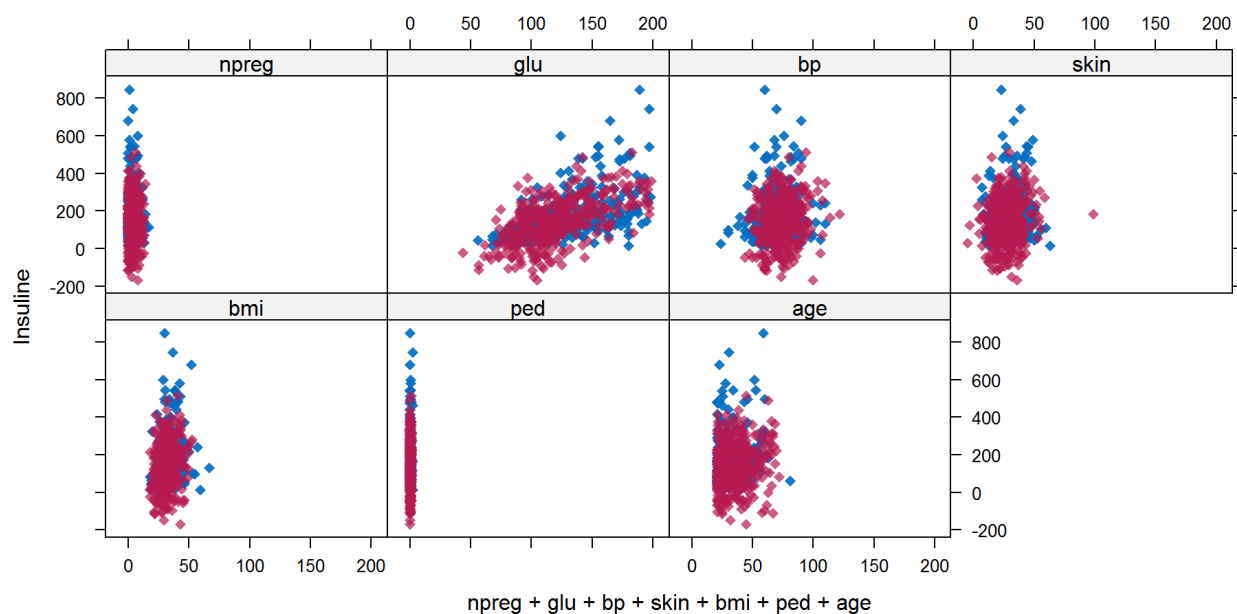
Les graphiques montrent que les imputations par régression linéaire stochastique comportent des valeurs aberrantes, en l'espèce, des valeurs négatives pour l'épaisseur de peau et le taux d'insuline. Les distributions des valeurs observées et imputées sont proches, ce qui n'infirme ni ne confirme le mécanisme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

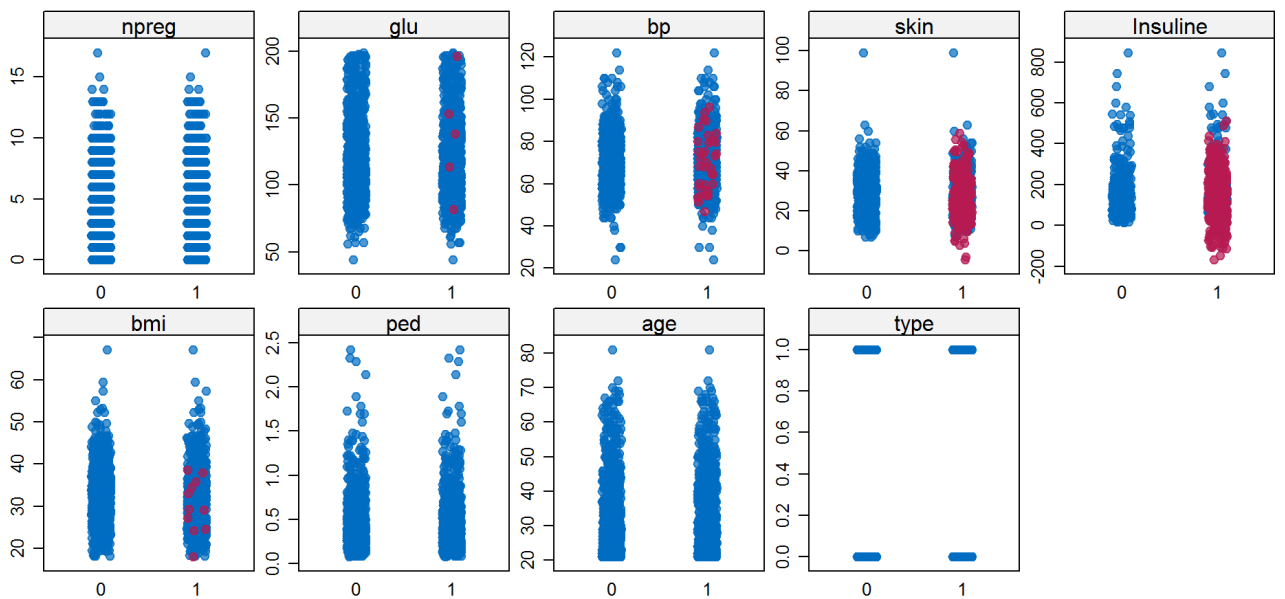
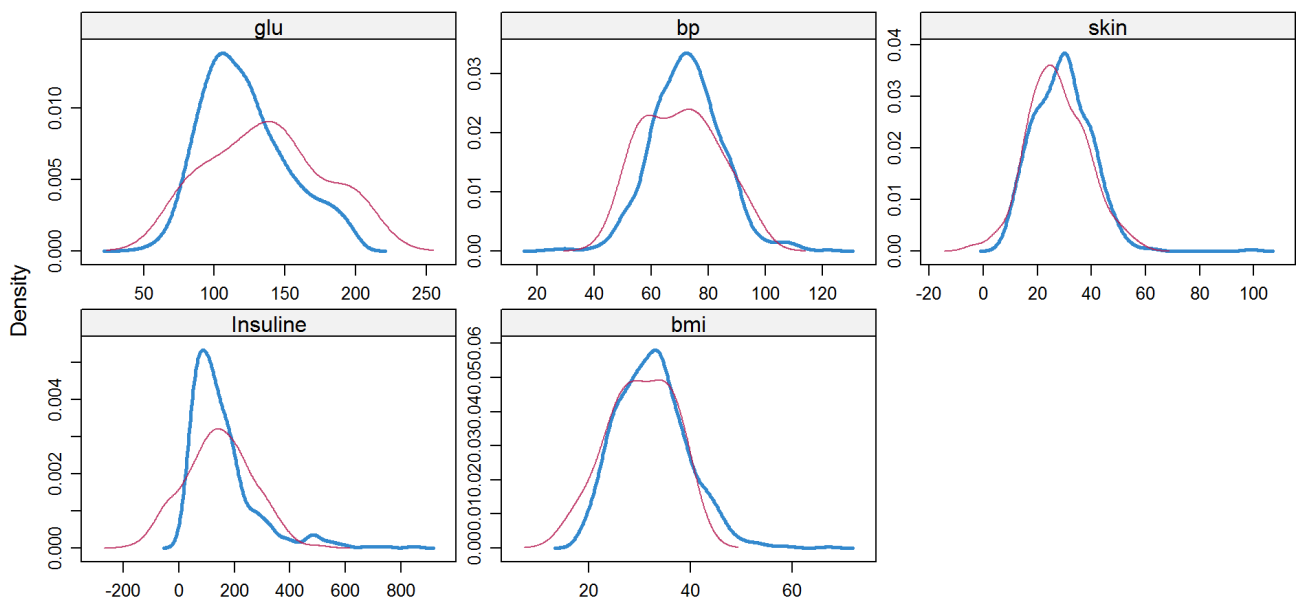
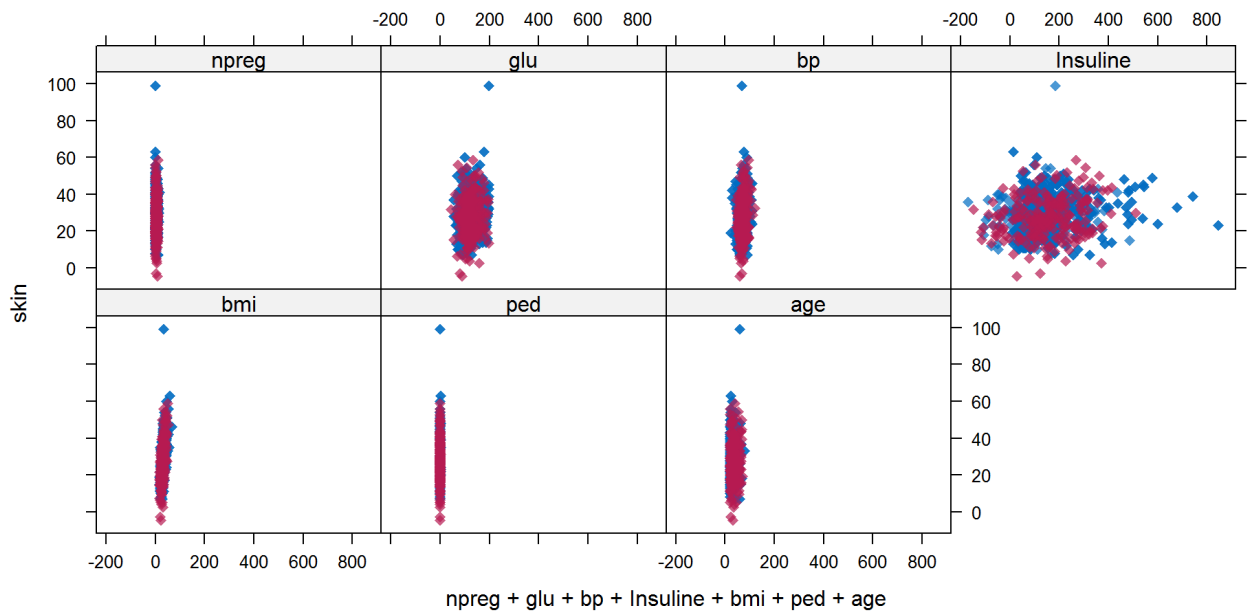
- Imputations simples - PMM





- Imputations simples - Normales

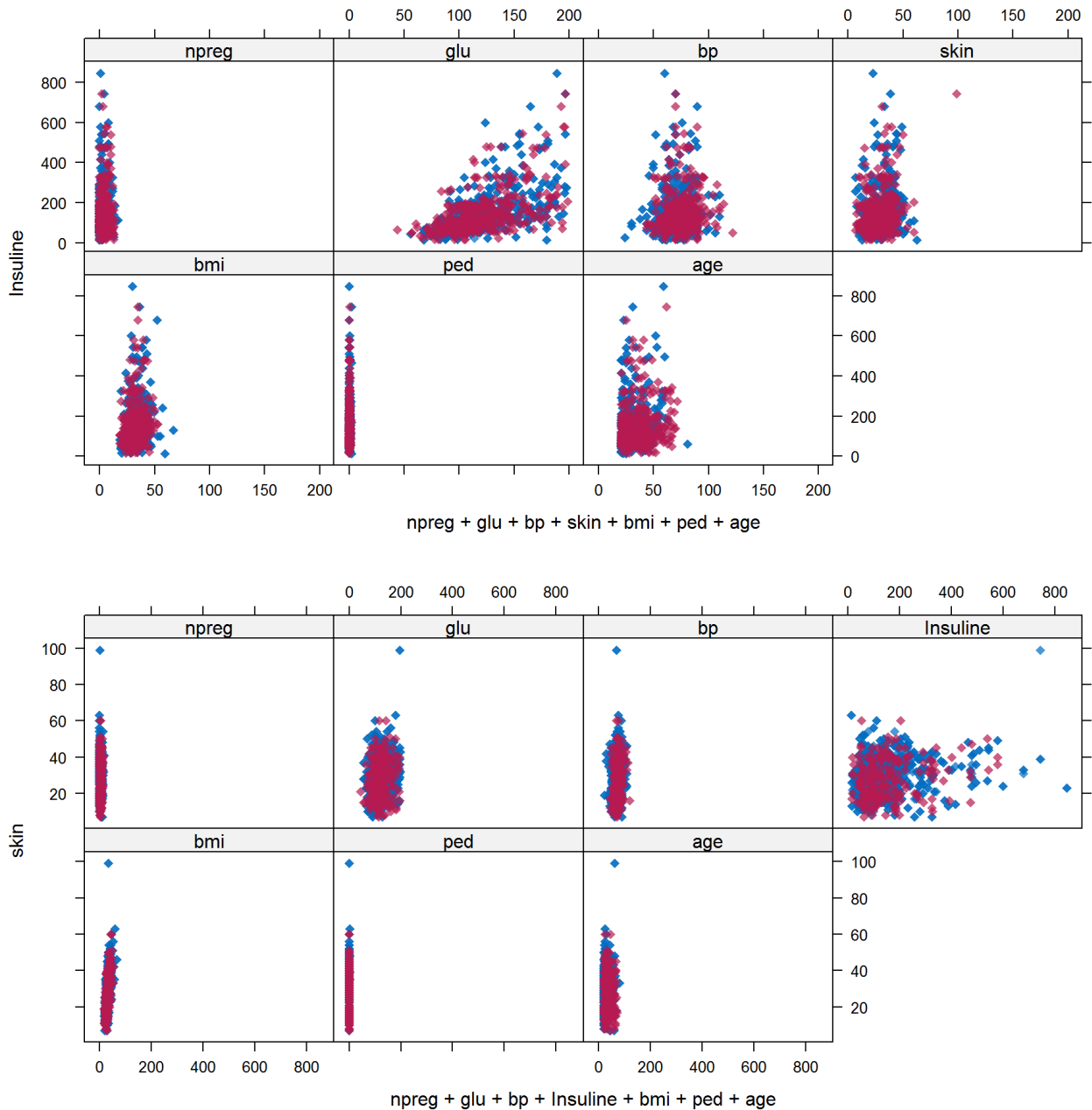


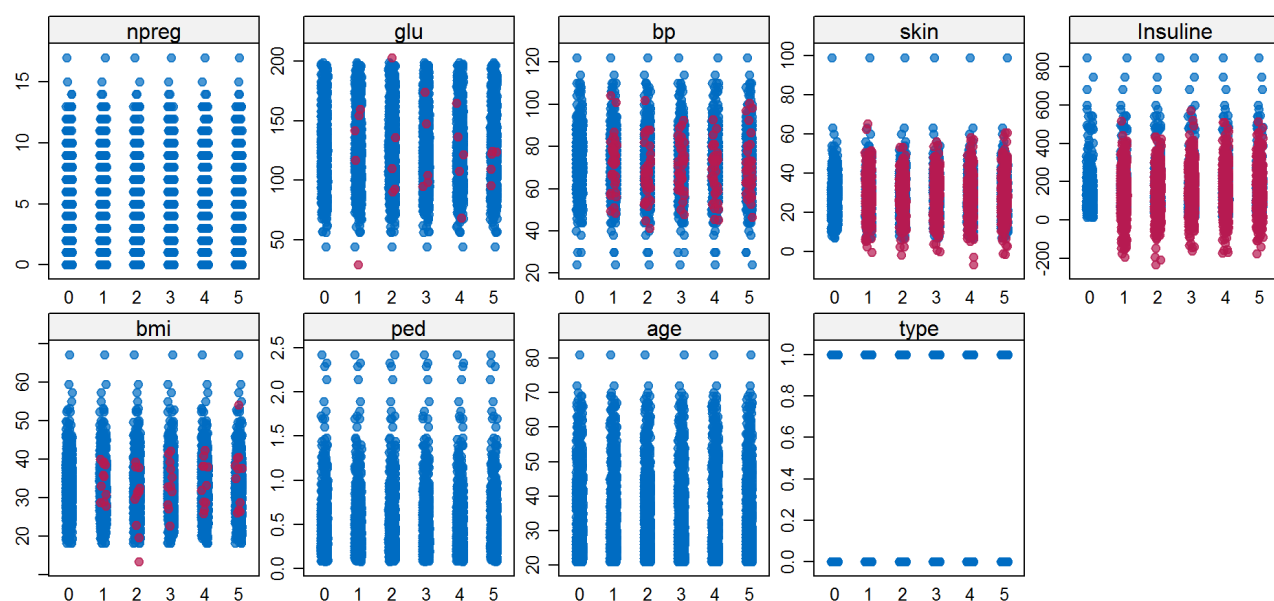
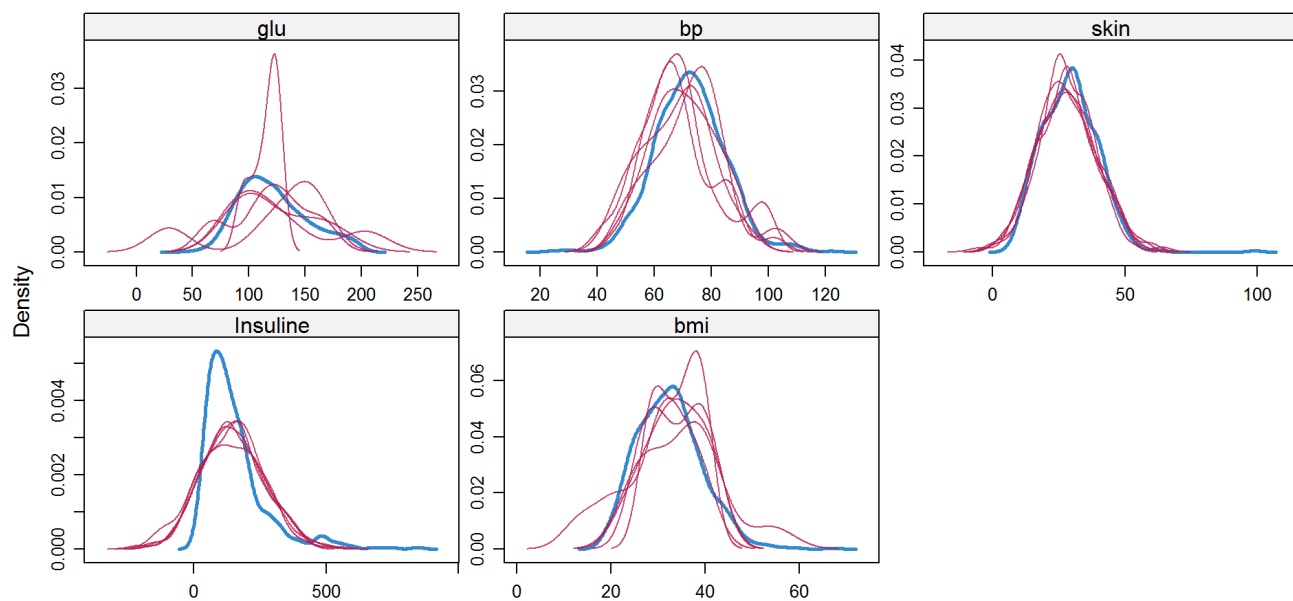


graphiques N°3 - Analyses des distributions pour les Imputations multiples (§-3.1.3)

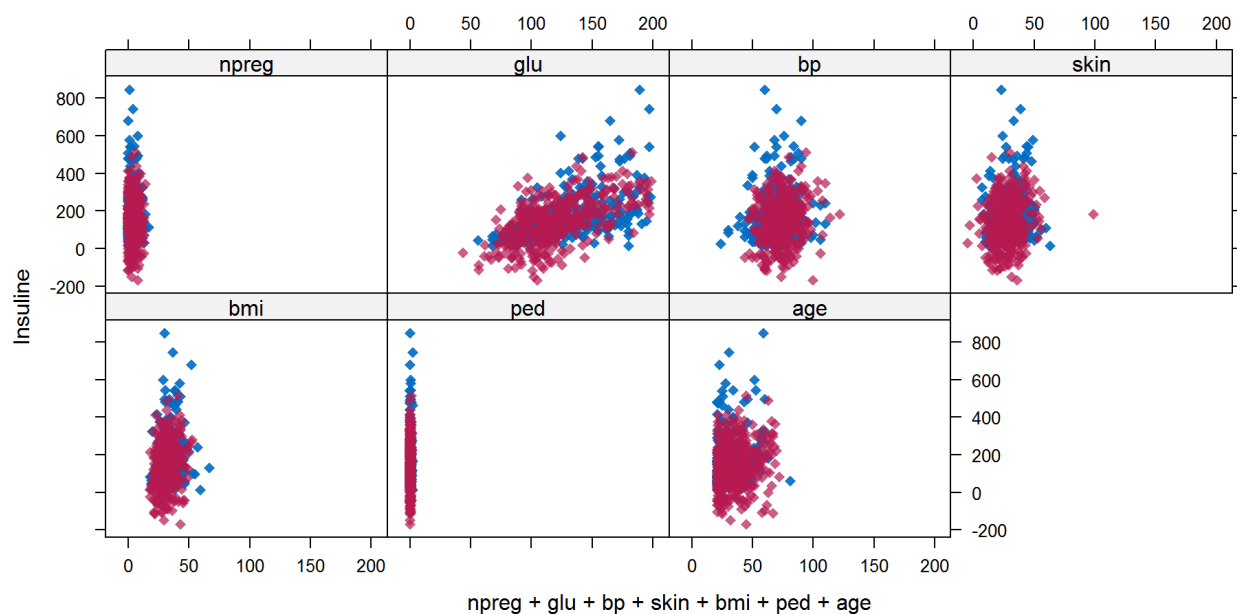
A nouveau, certaines imputations ne sont pas plausibles : avec le modèle joint les graphiques montrent des valeurs aberrantes pour l'épaisseur de peau et le taux d'insuline. S'agissant de la méthode FCS, les distributions des variables glu et bmi présentent des profils très divergents pour chacune des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

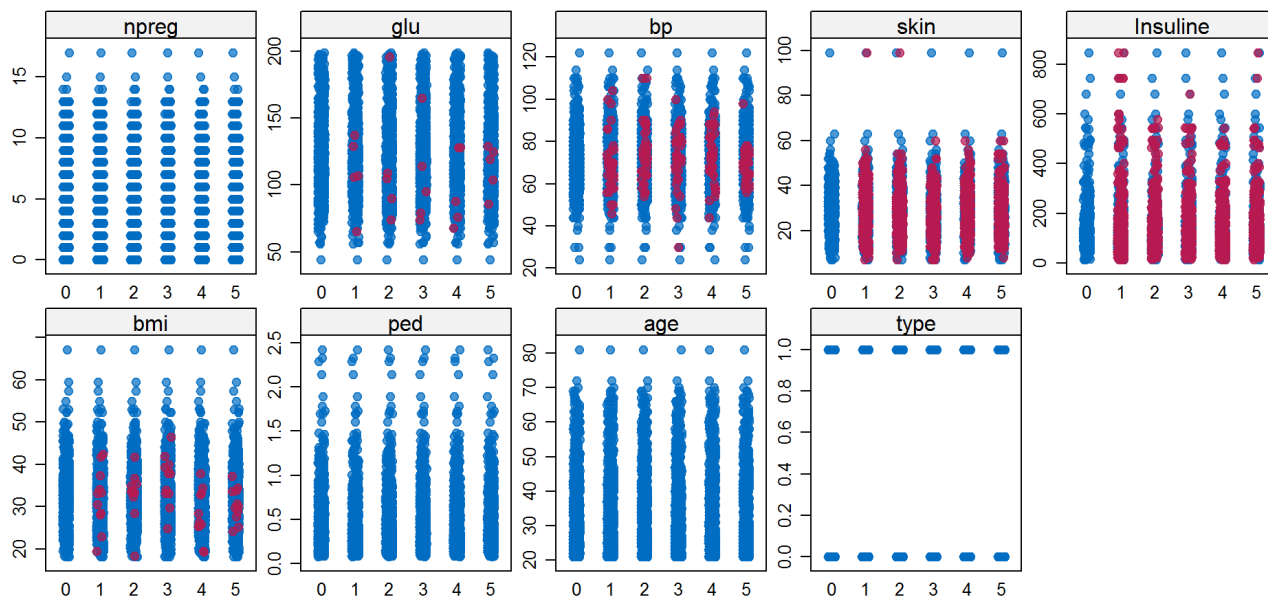
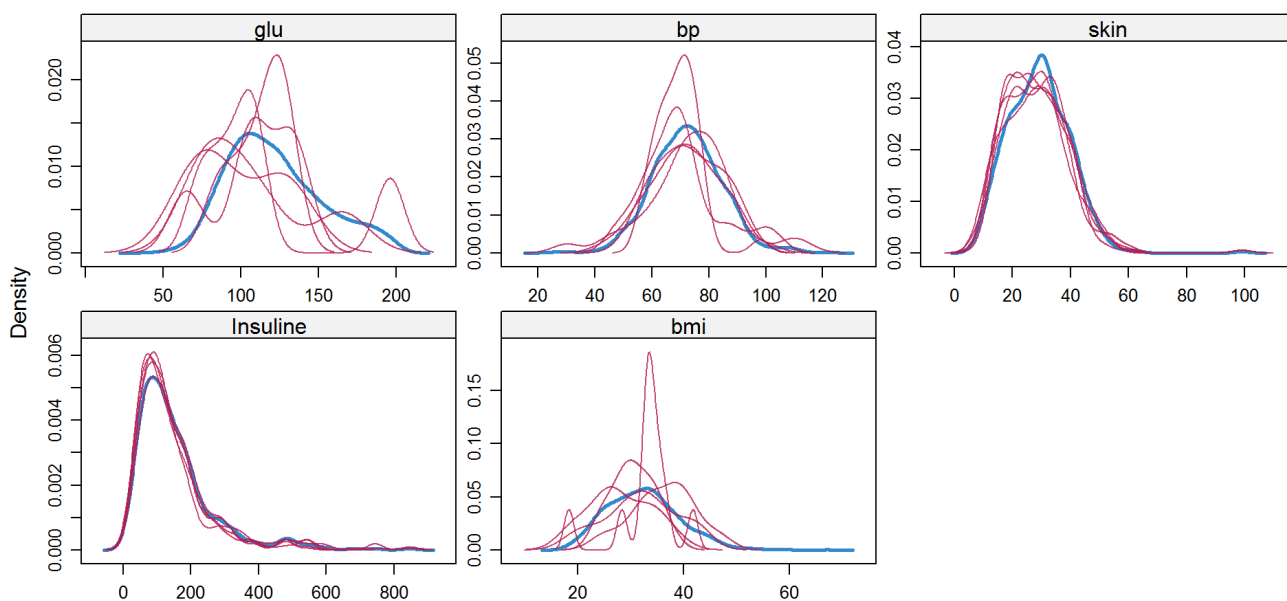
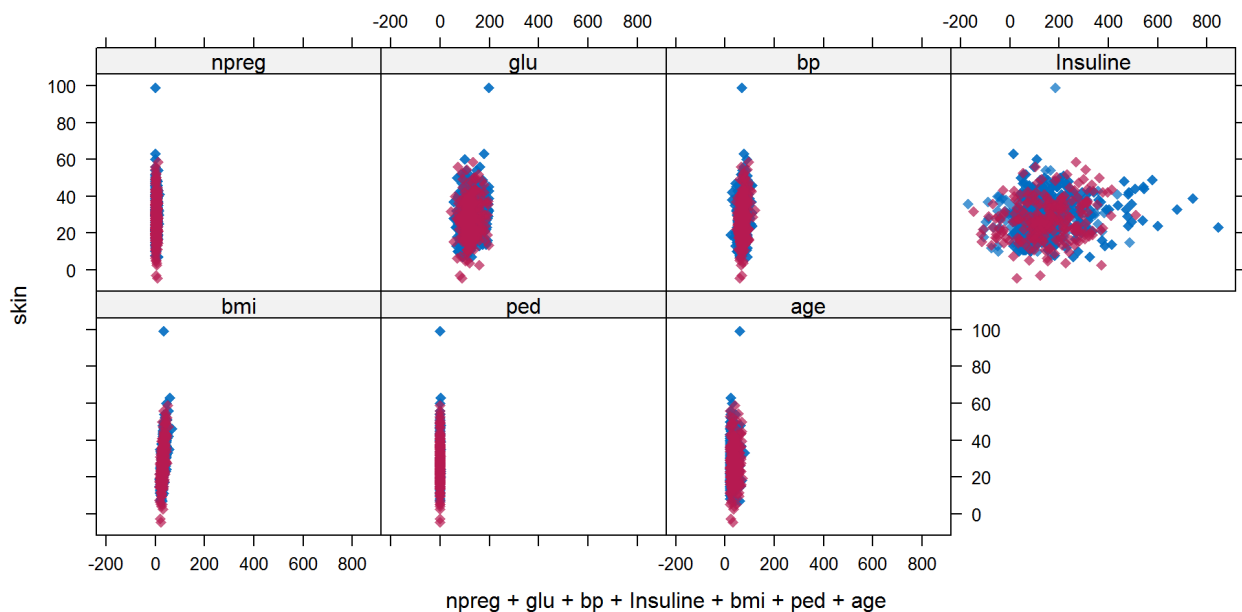
- Imputations multiples - JM





- Imputations multiples - FCS

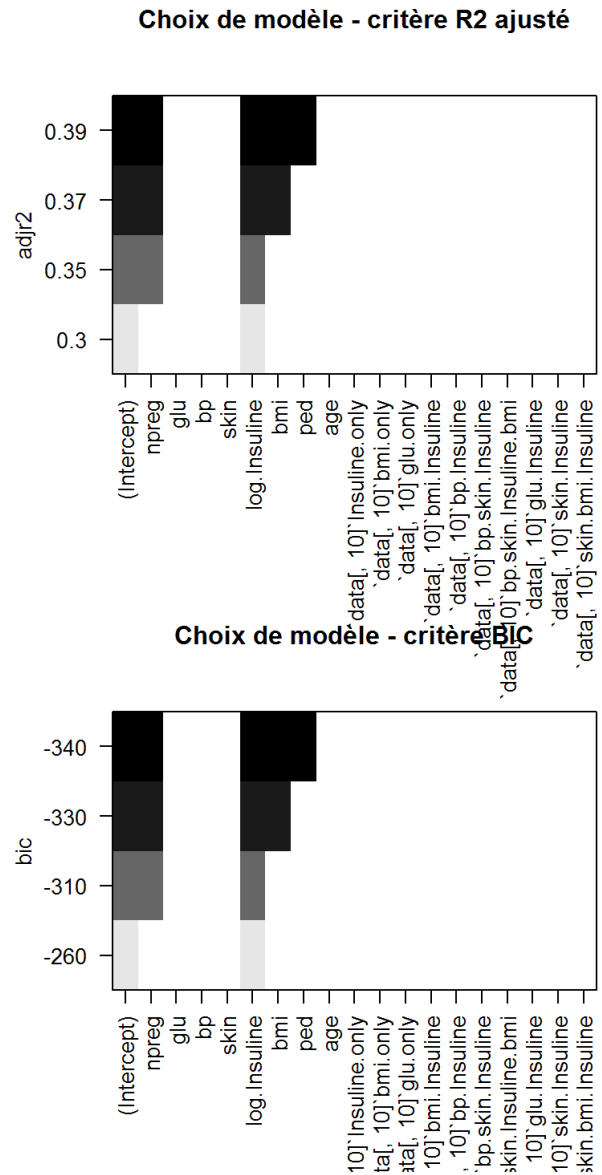
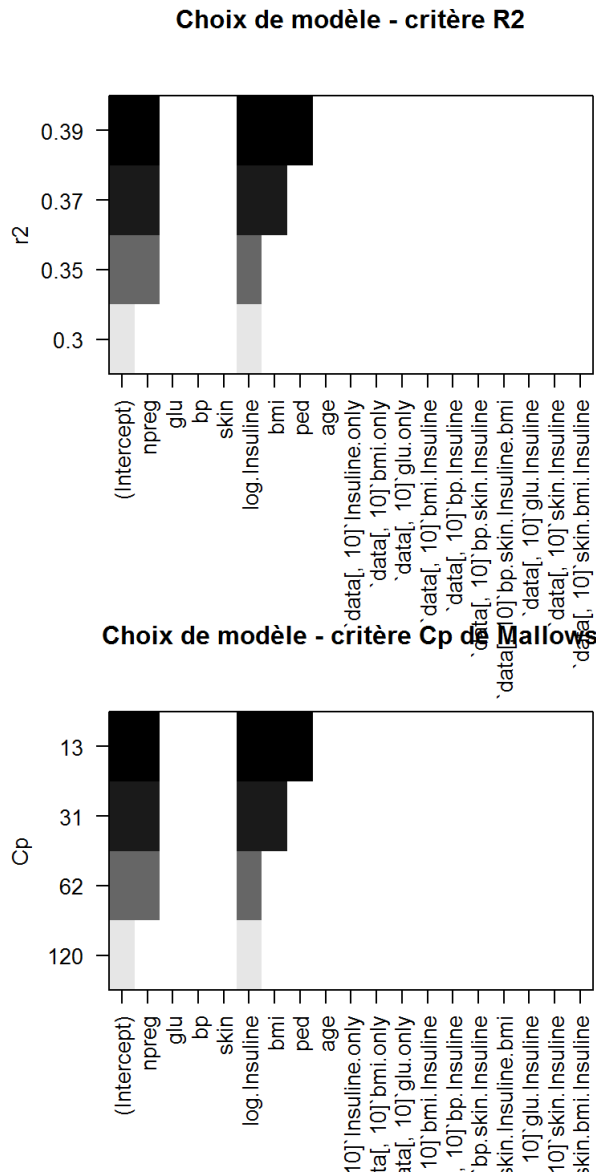




Annexe 2 - choix de modèle

Dans cette annexe on applique les techniques regsubsets et step pour le choix automatique des variables. Les données utilisées sont les données obtenu à partir de la variable \$res.imputePCA. On obtient à nouveau ce résultat en utilisant comme données d'input les moyennes des différents jeux de données issues du bootstrap 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 = binomial, 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 = dataFramePima)
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
```

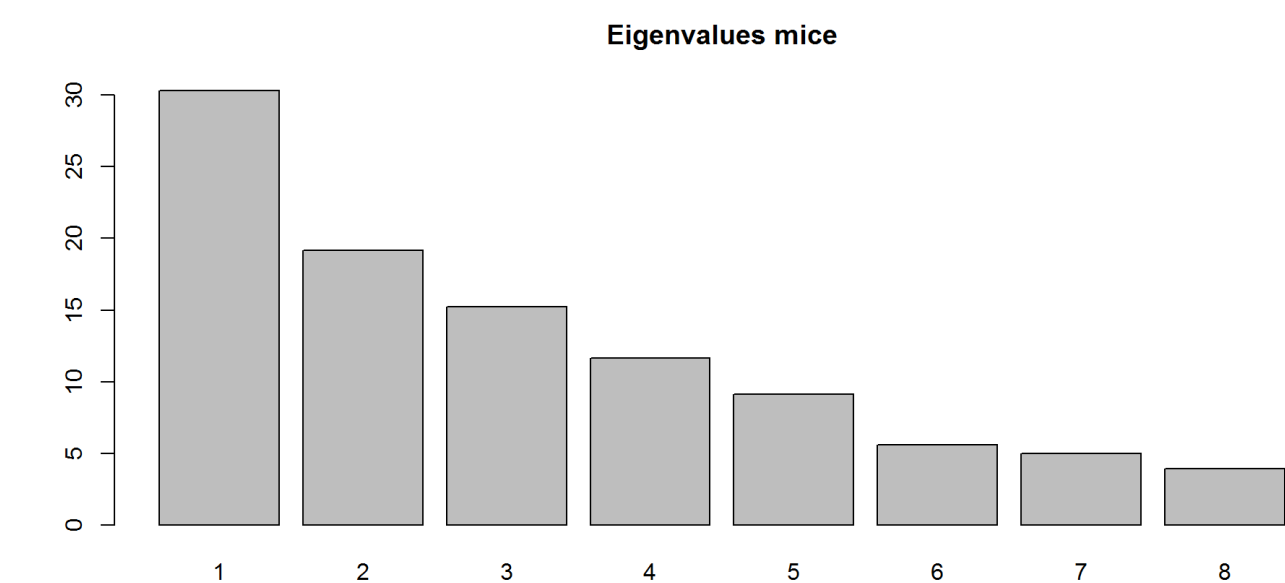
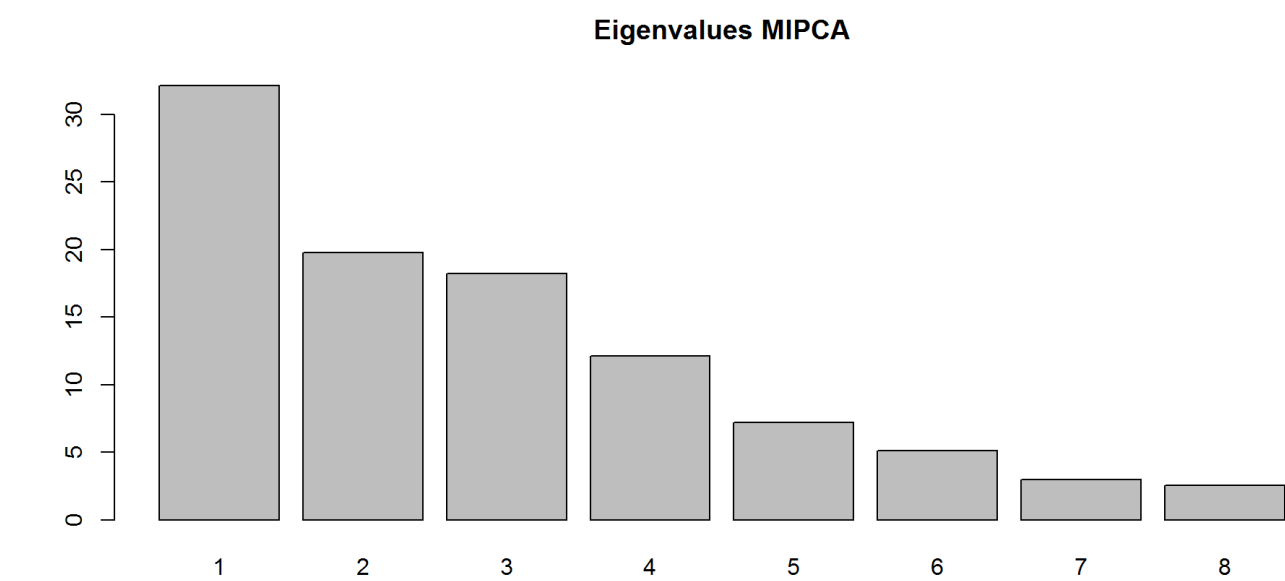
```
BIC(m_StepBwdFwd_BIC)
```

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

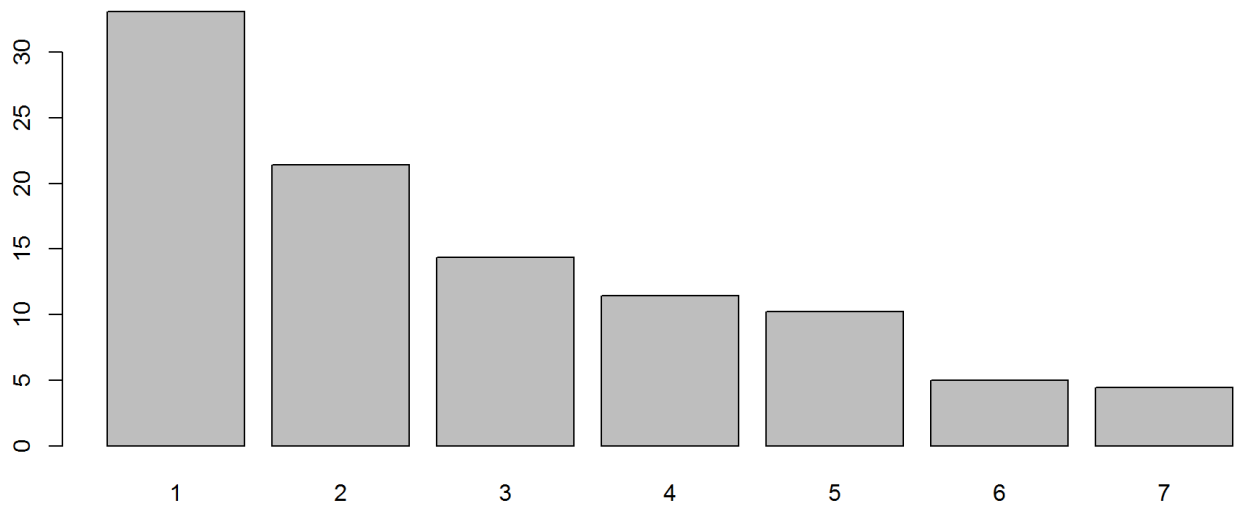

Annexe 3 : Analyse PCA - comparaison aux données non traitées

- 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



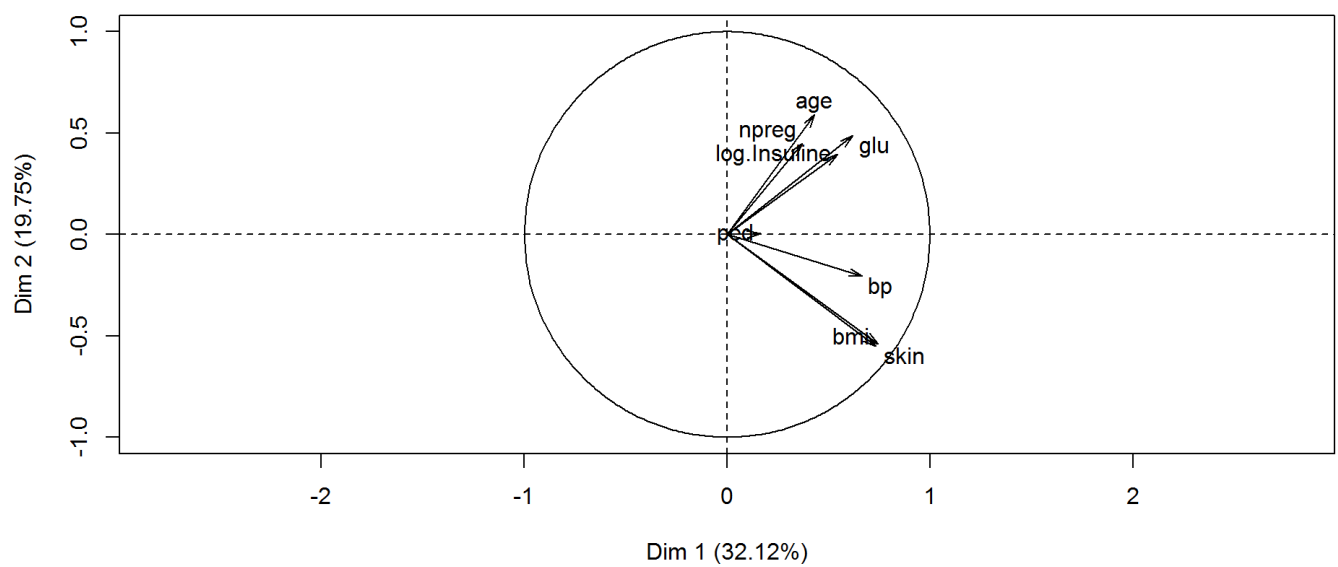
Eigenvalues MASS



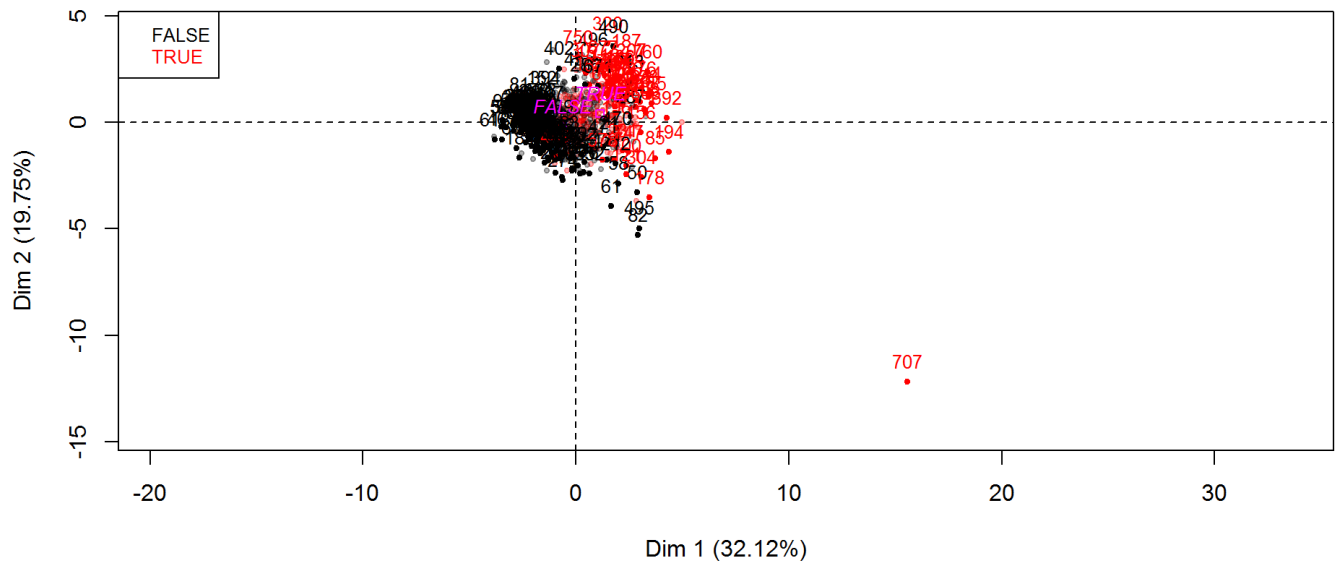
2. Graphiques : nuage des individus et cercle de qualité des projections

- Données Pima complétées avec la méthode MIPCA du package MissMDA

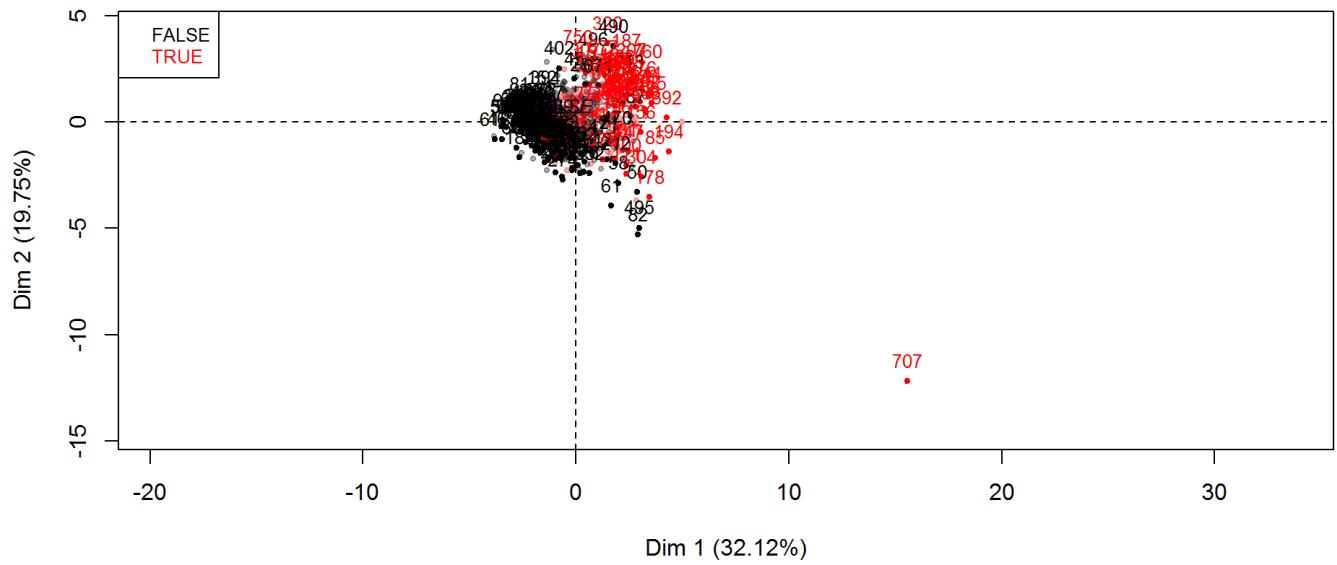
Variables factor map (PCA)



Individuals factor map (PCA)



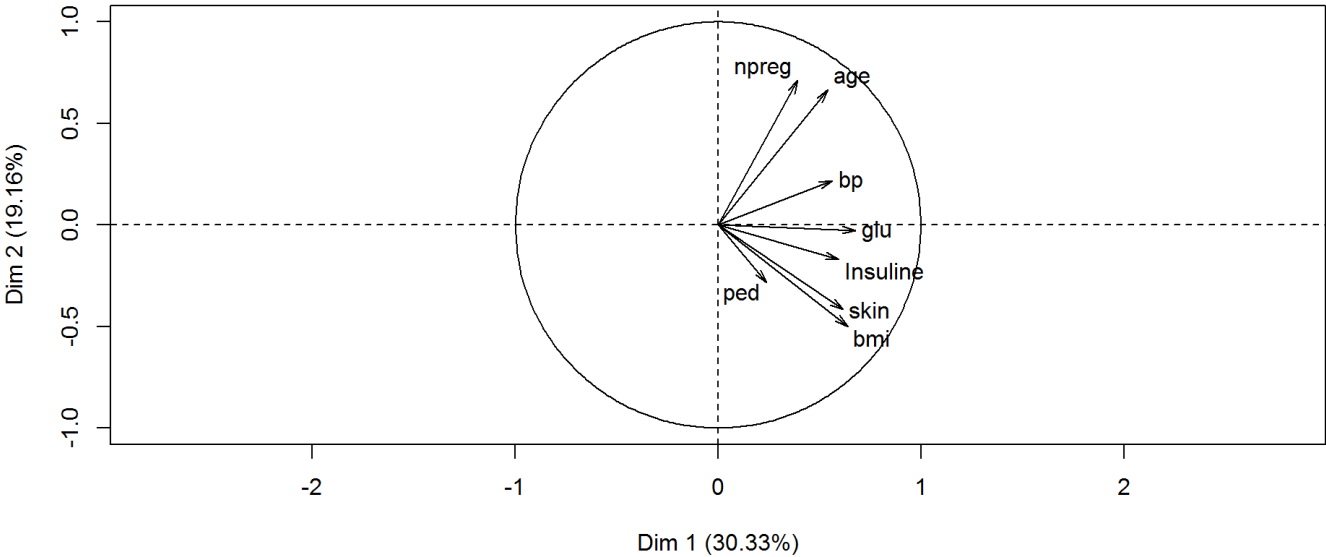
Individuals factor map (PCA)



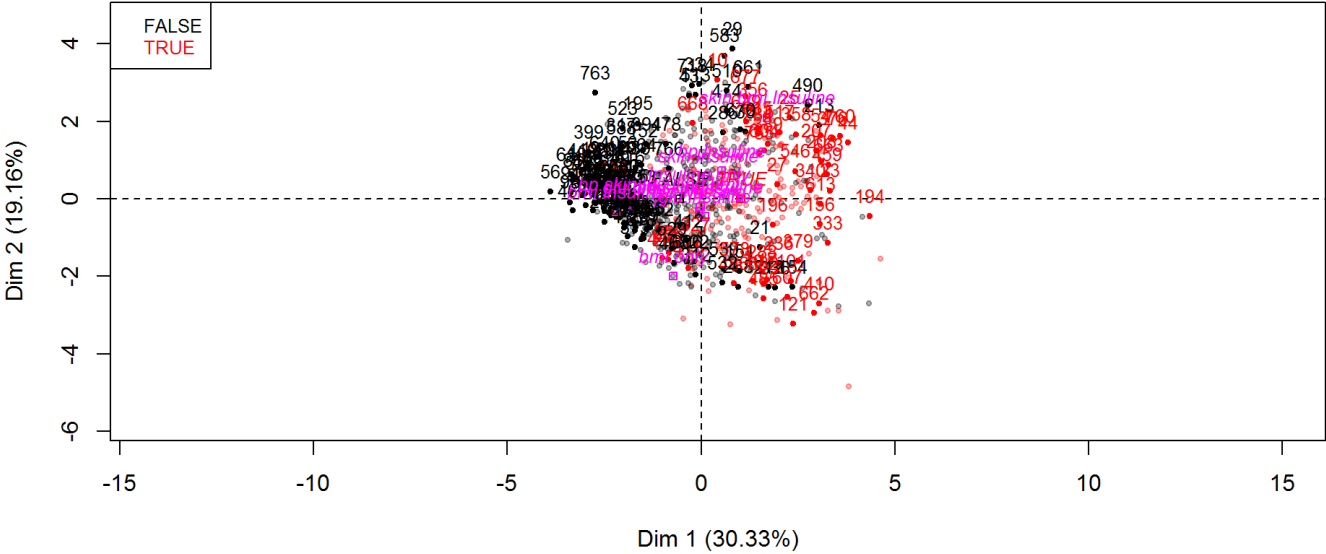
Le nuage des individus est bien centré

- Données Pima complétées avec la méthode mice du package mice

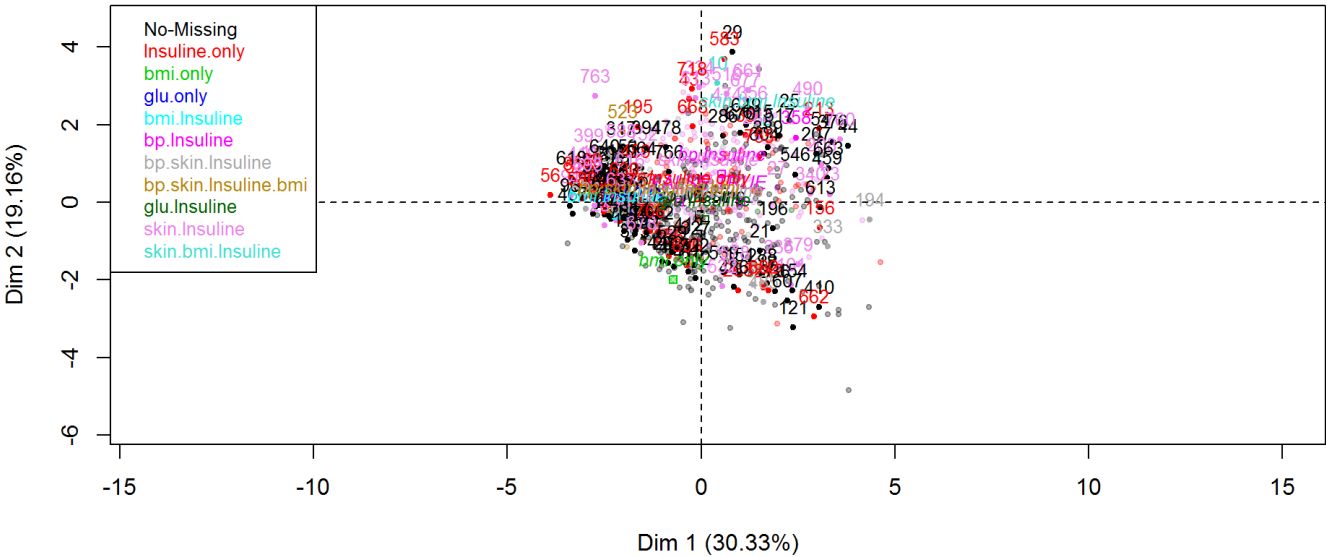
Variables factor map (PCA)



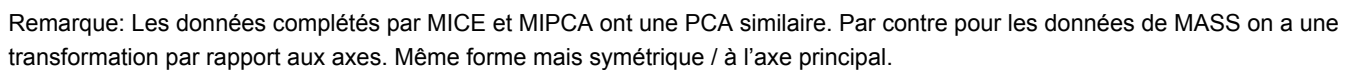
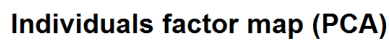
Individuals factor map (PCA)



Individuals factor map (PCA)



- Données Pima de la librairie MASS



3. coefficients de corrélation entre chacune 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

```

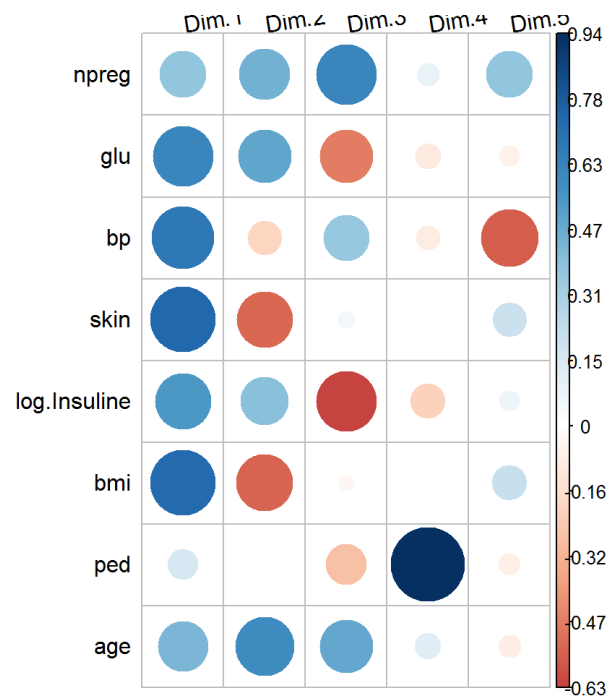
```
## typeMissing=TRUE -0.276669 9.264595e-10
## type=TRUE -0.276669 9.264595e-10
```

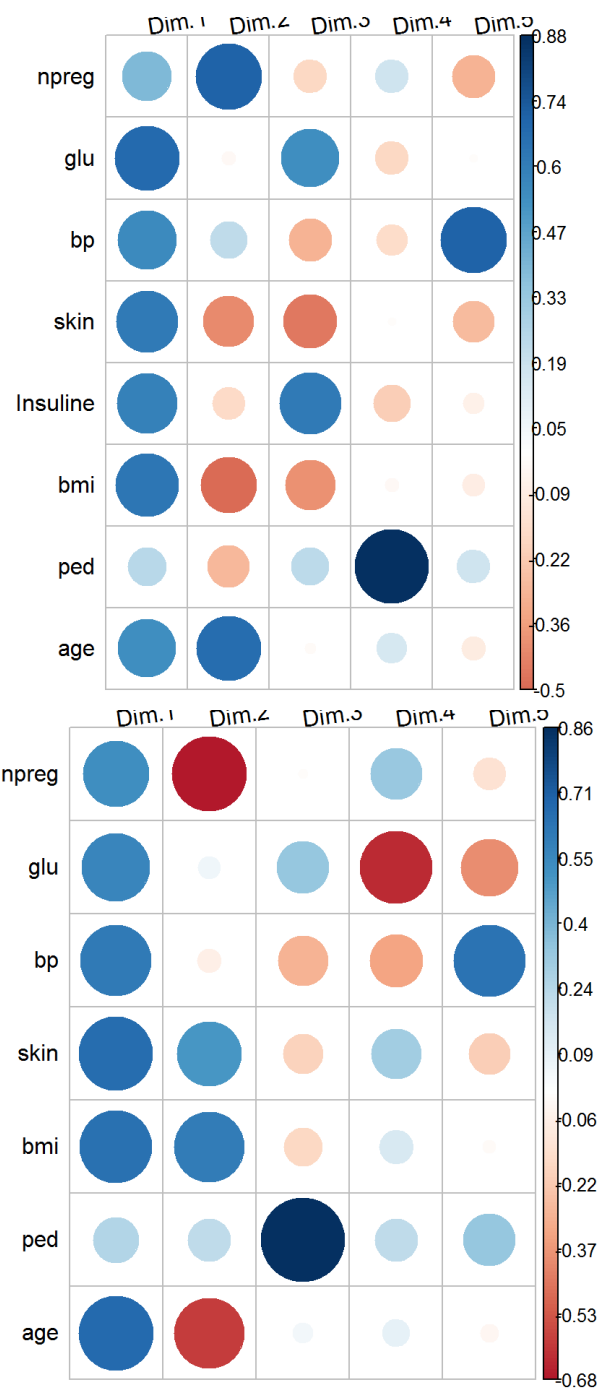
Graphique de corrélation avec corrplot

```
##          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
```



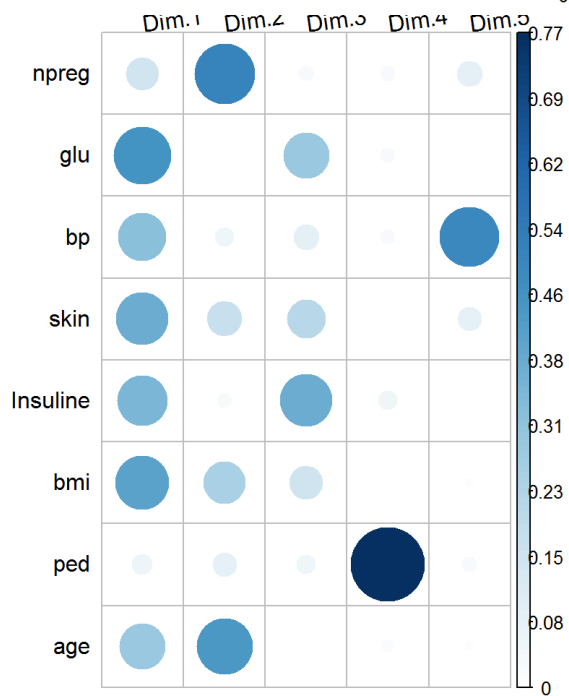
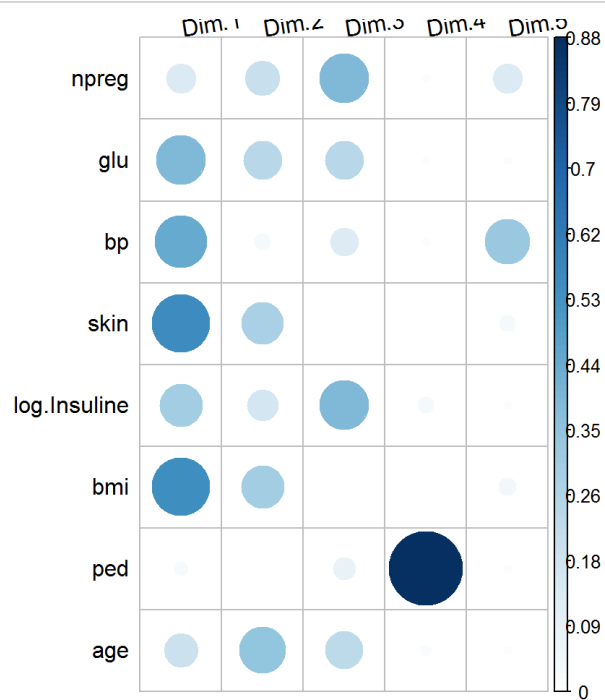


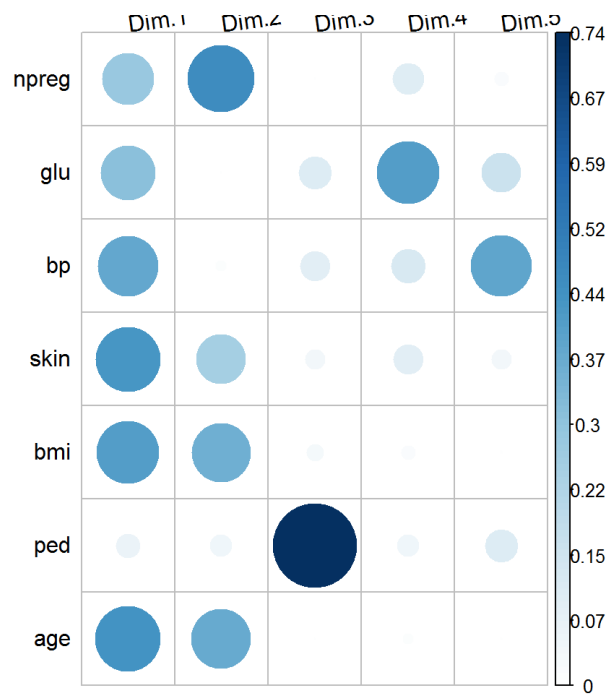
4. Indice de qualité de la représentation cos2

```
##          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
```

