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.

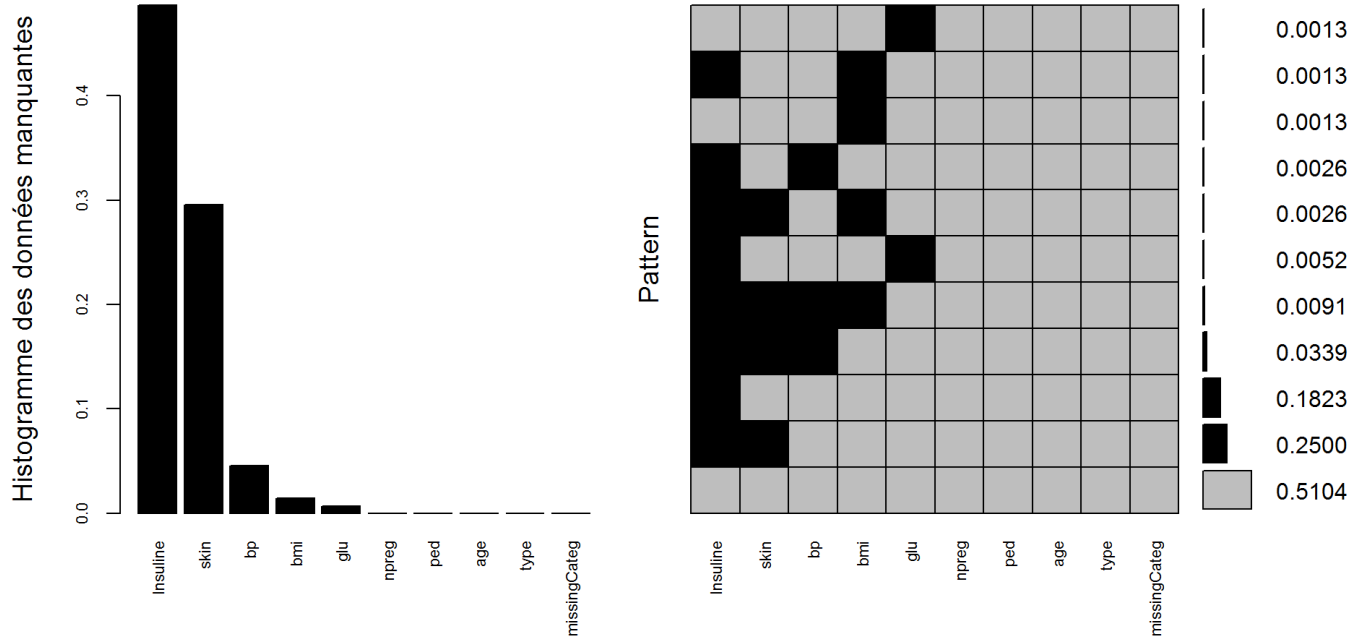
Recours aux librairies MICE, MASS, VIM pour visualiser les patterns du jeu de données grâce aux fonctions fluxplot() ou aggr_plot().

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

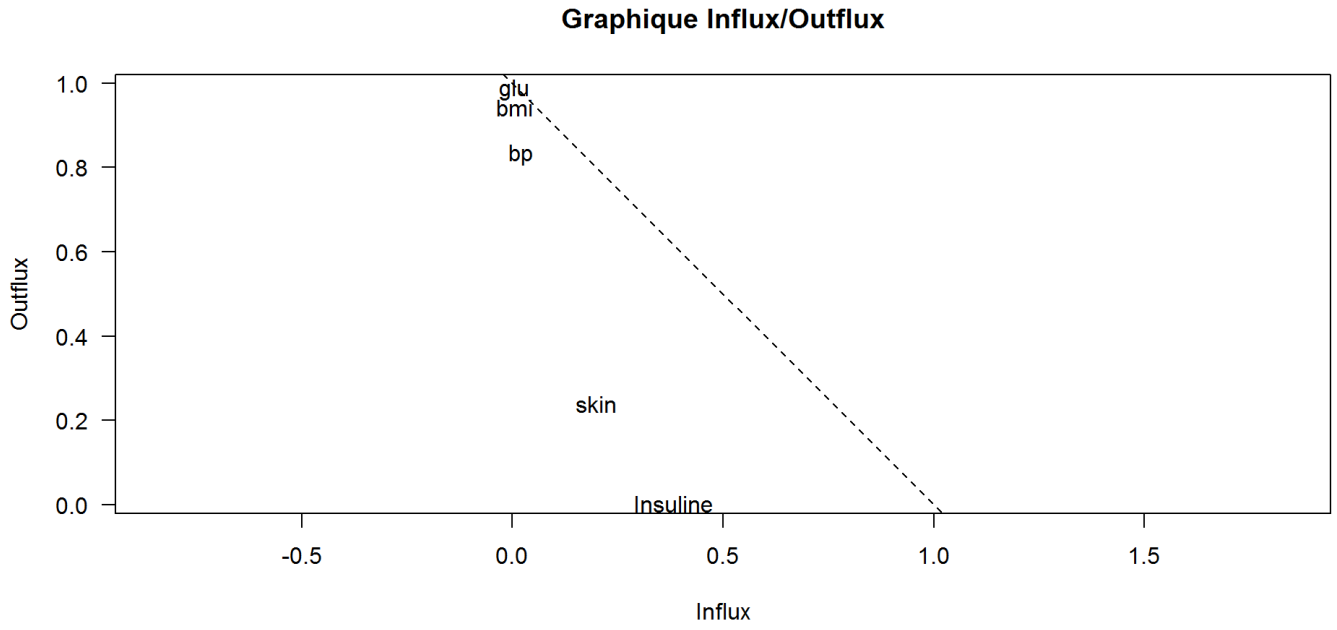
A elle seule, la variable Insuline, quand elle est manquante, regroupe 8 patterns. La variable skin concerne 4 patterns.

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
```



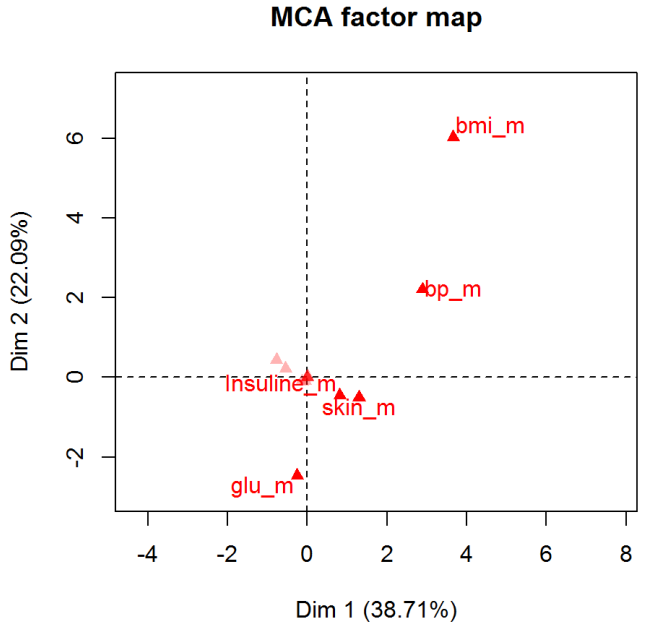
	influx <dbl>	outflux <dbl>
glu	0.005018821	0.986196319
bp	0.020388959	0.831288344
skin	0.200439147	0.239263804
Insuline	0.382685069	0.003067485
bmi	0.005646173	0.943251534

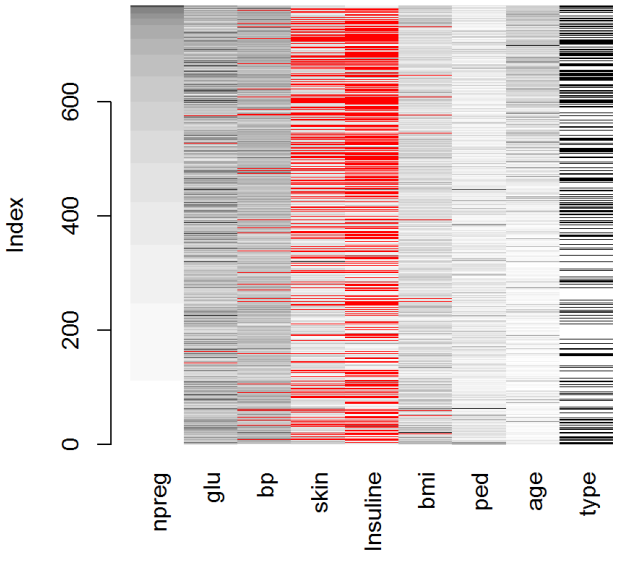
5 rows

```
## [1] "Nombre de patterns si Insuline manquant :8"
```

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



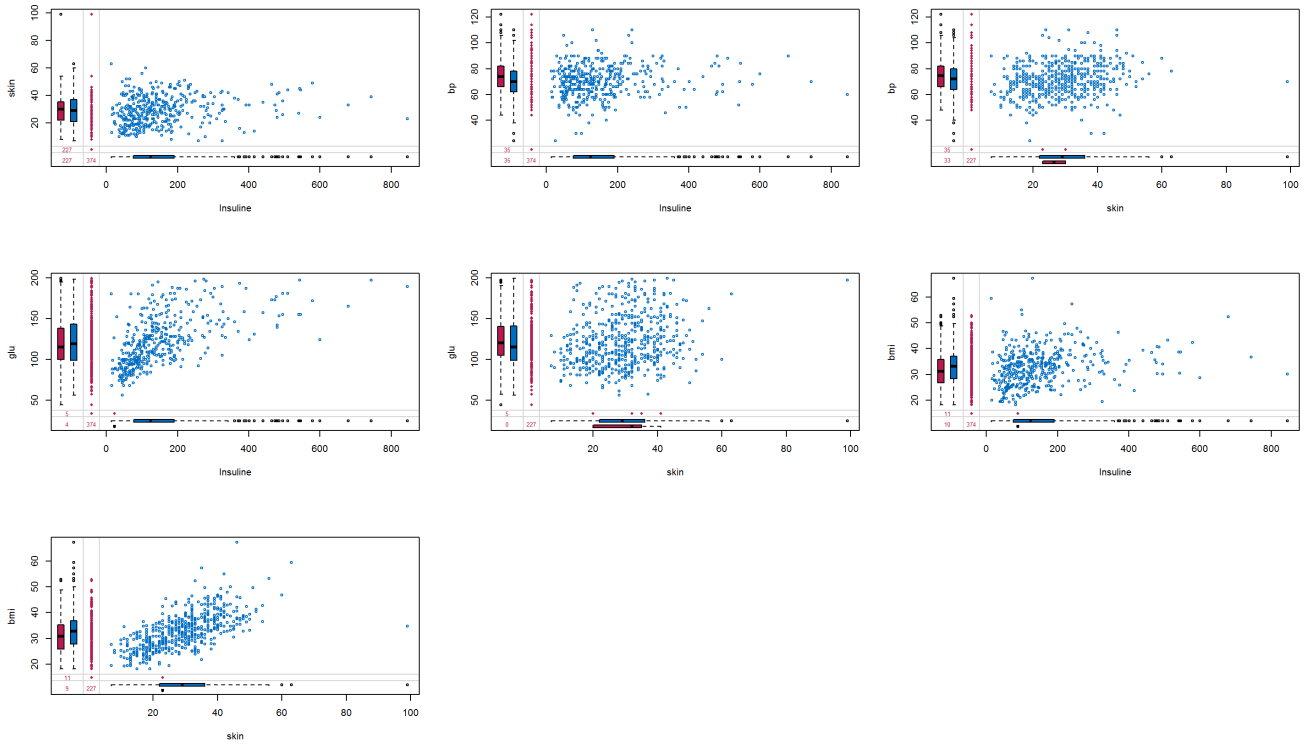


On ne remarque pas de groupes de variables bien séparés. Peut-être une liaison skin-insuline

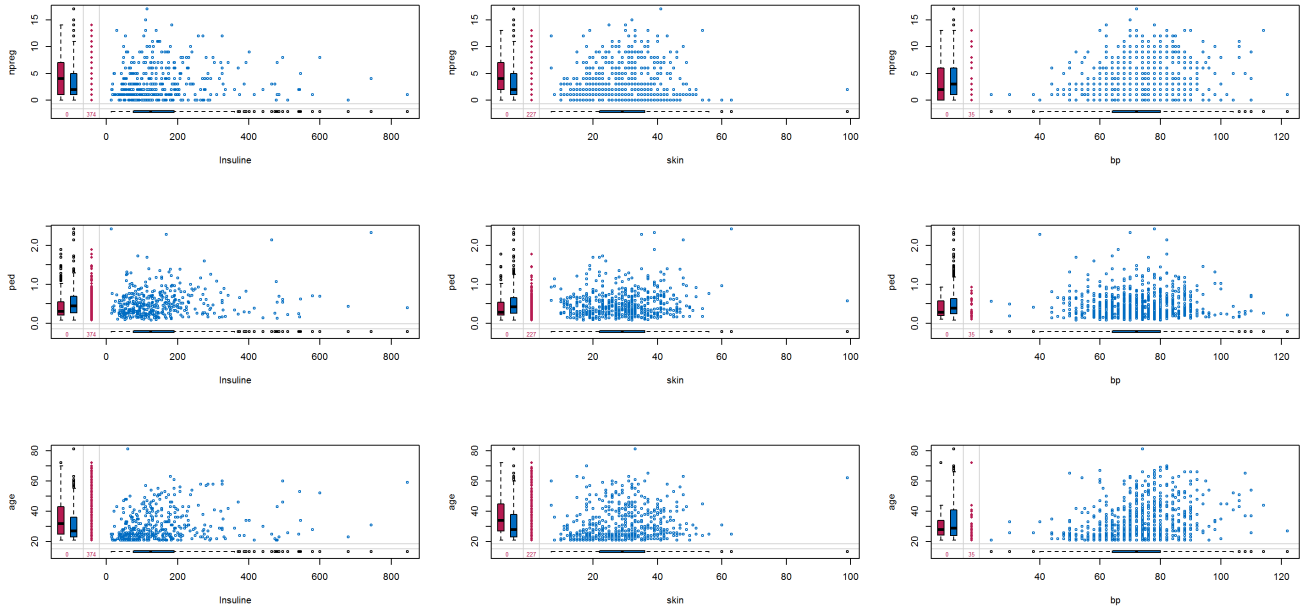
- L’hypothèse MCAR semble infirmée au regard des distributions des variables Insuline et Skin comparées aux variables complètes, npreg, age et ped. A ce stade, l’hypothèse d’ignorabilité du mécanisme peut par ailleurs être maintenue. L’analyse de sensibilité nous permettra de la valider. Nous pouvons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

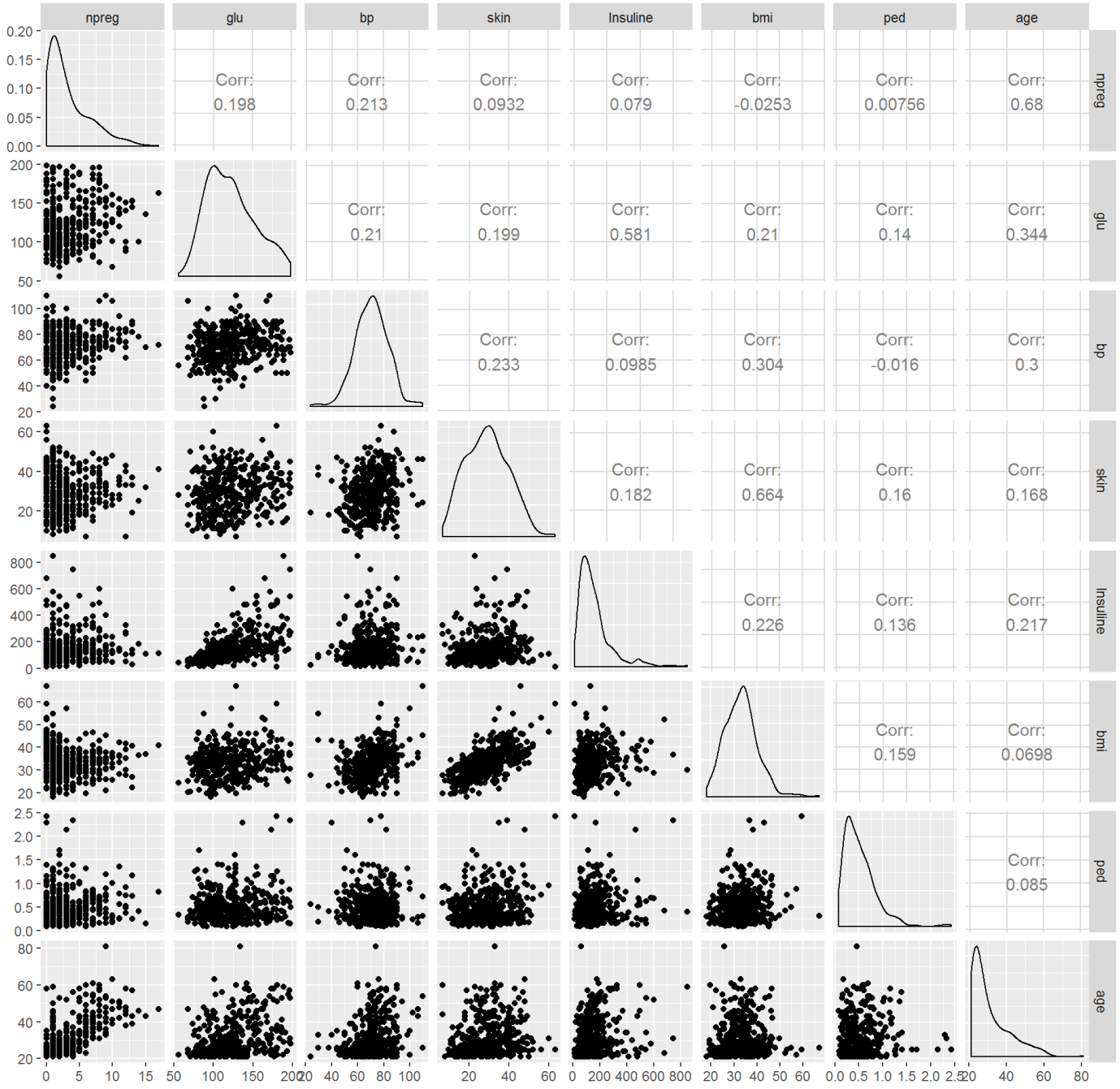
A nouveau, certaines imputations ne sont pas plausibles : avec le modèle joint les graphiques montrent des valeurs aberrantes pour l'épaisseur de peau et le taux d'insuline. S'agissant de la méthode FCS, les distributions des variables glu et bmi présentent des profils très divergents pour chacune des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

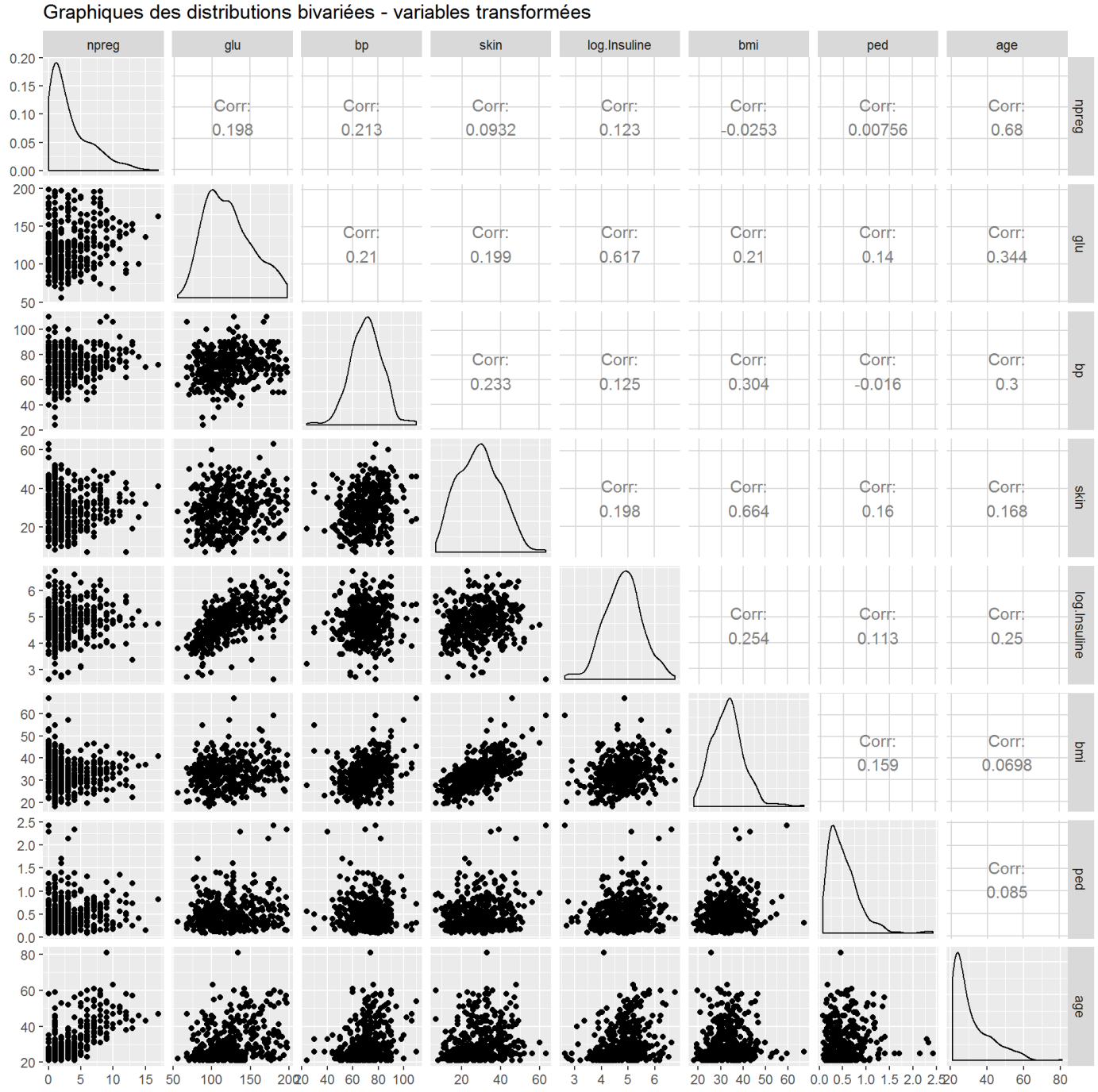
3.2 Imputation multiple par les méthodes d'analyse factorielle avec MIPCA

- 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.

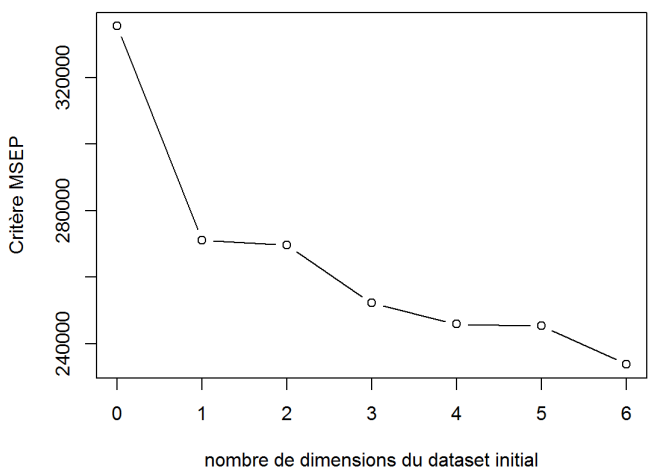
Graphiques des distributions bivariées

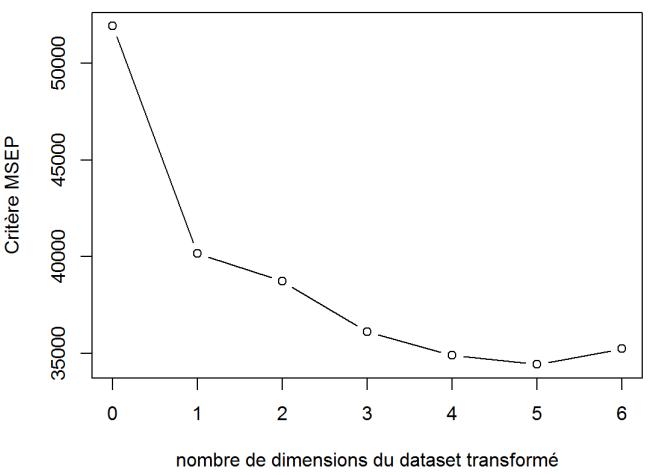




- Détermination de la dimension de l'espace de projection

Avec la méthode de validation croisée "kfold", le nombre de dimensions optimale est de 6 ou 5 selon le critère "MSEP".





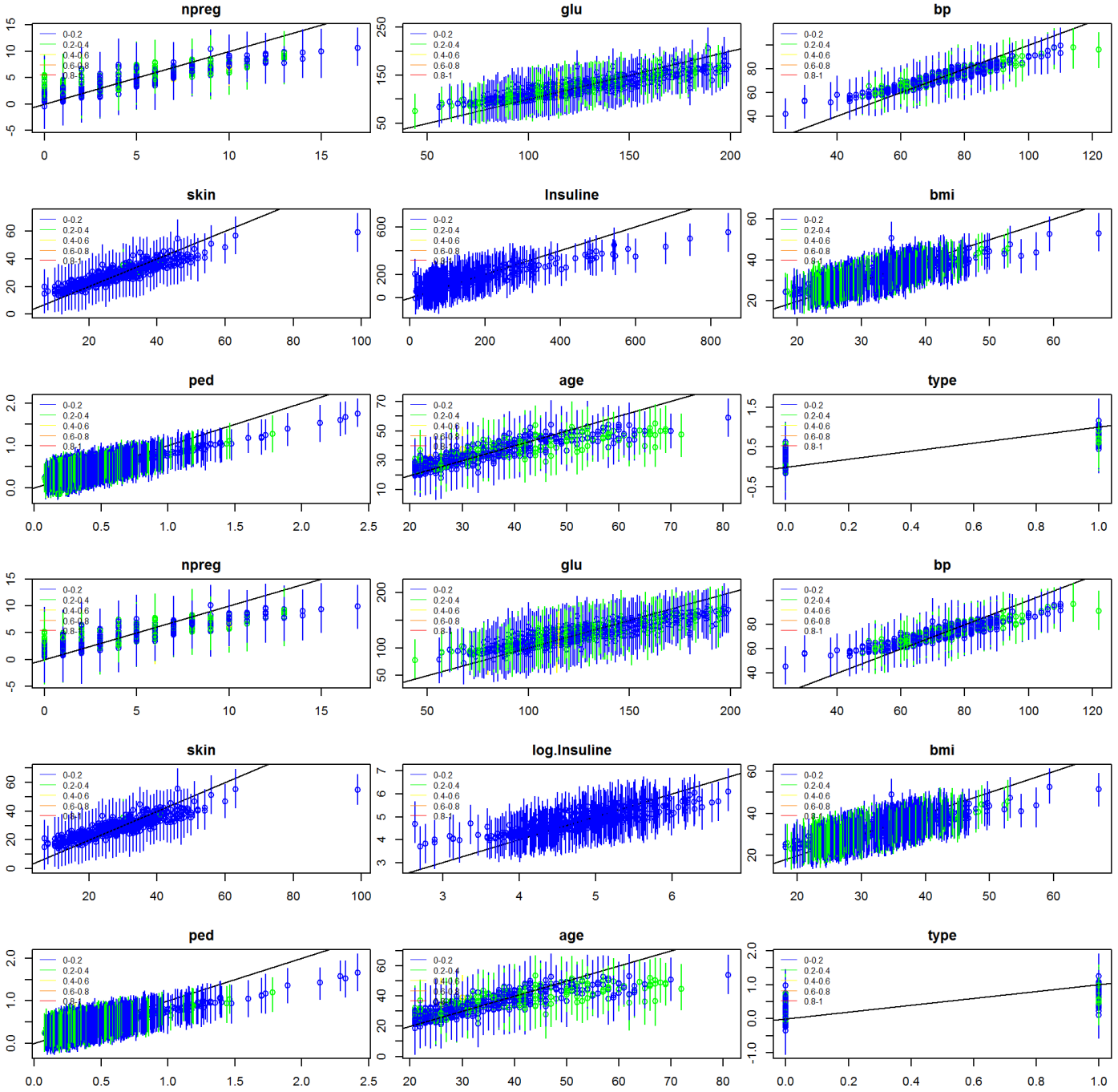
```
## [1] "Le nombre de dimensions retenu pour le jeu de données initial est de :6"
```

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

3.2.1 Imputation multiple - méthode bayésienne

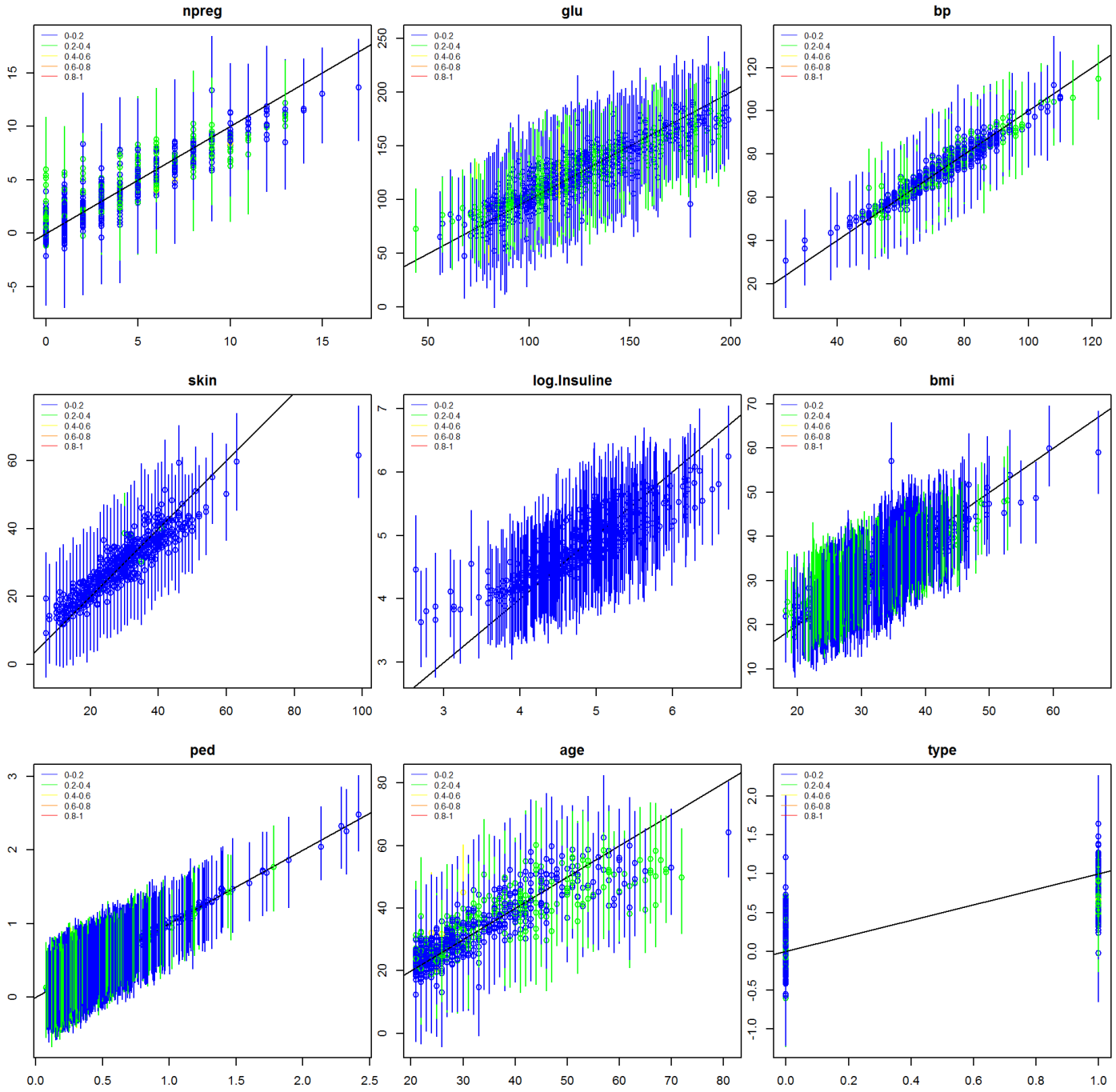
- Diagnostics pour l'imputation multiple - méthode bayésienne

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.



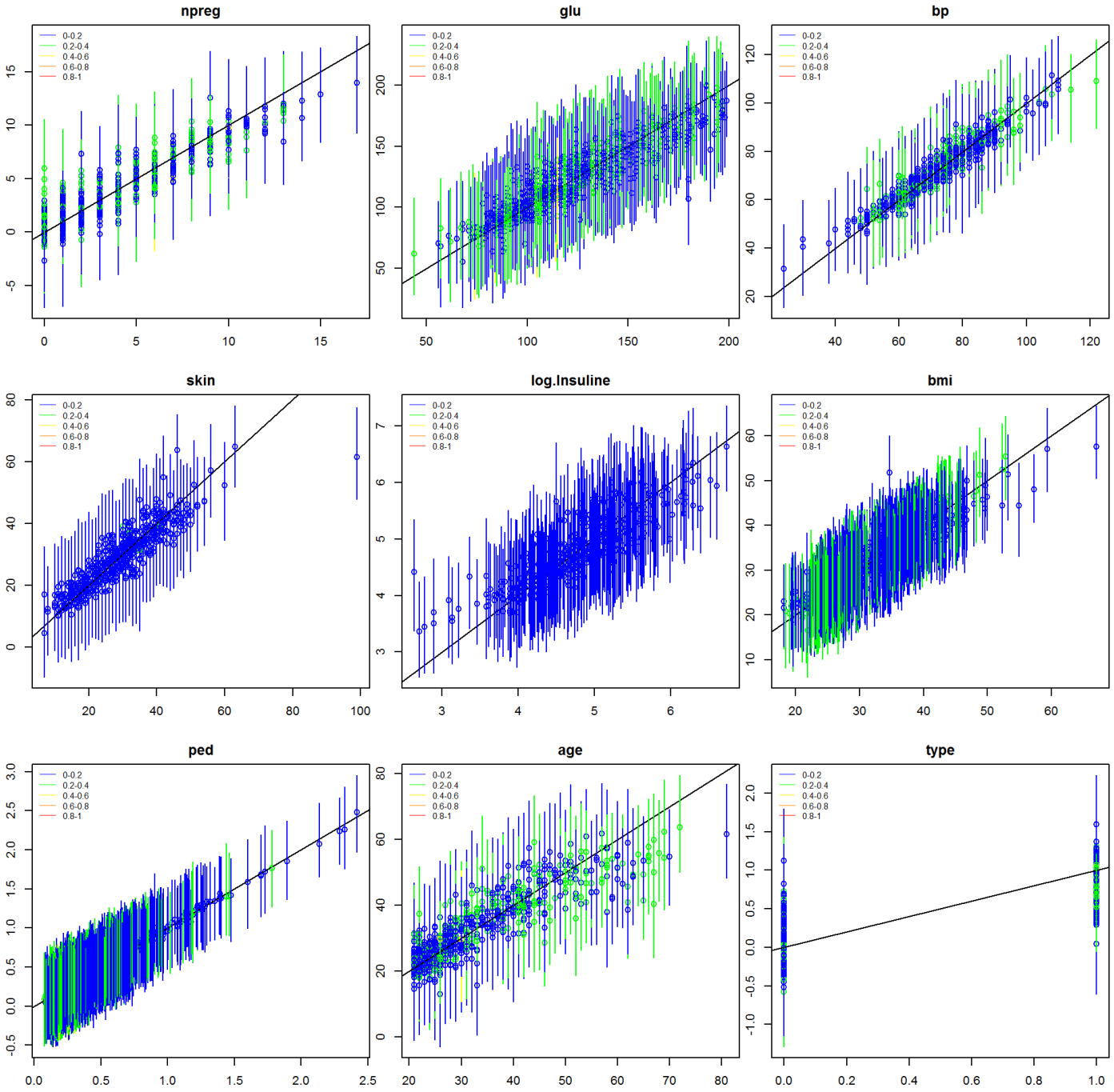
3.2.2 Multiple Imputation avec multiple imputation par boostrap

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



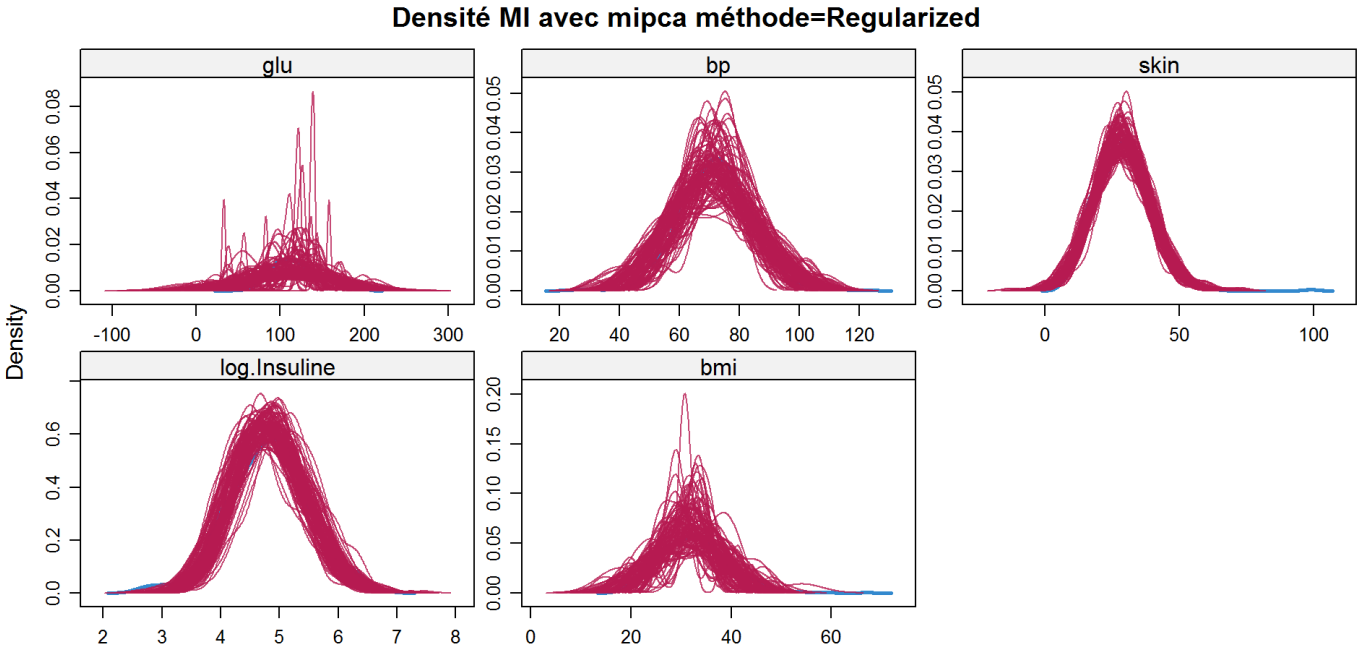
3.2.3 Multiple Imputation EM et bootstrap

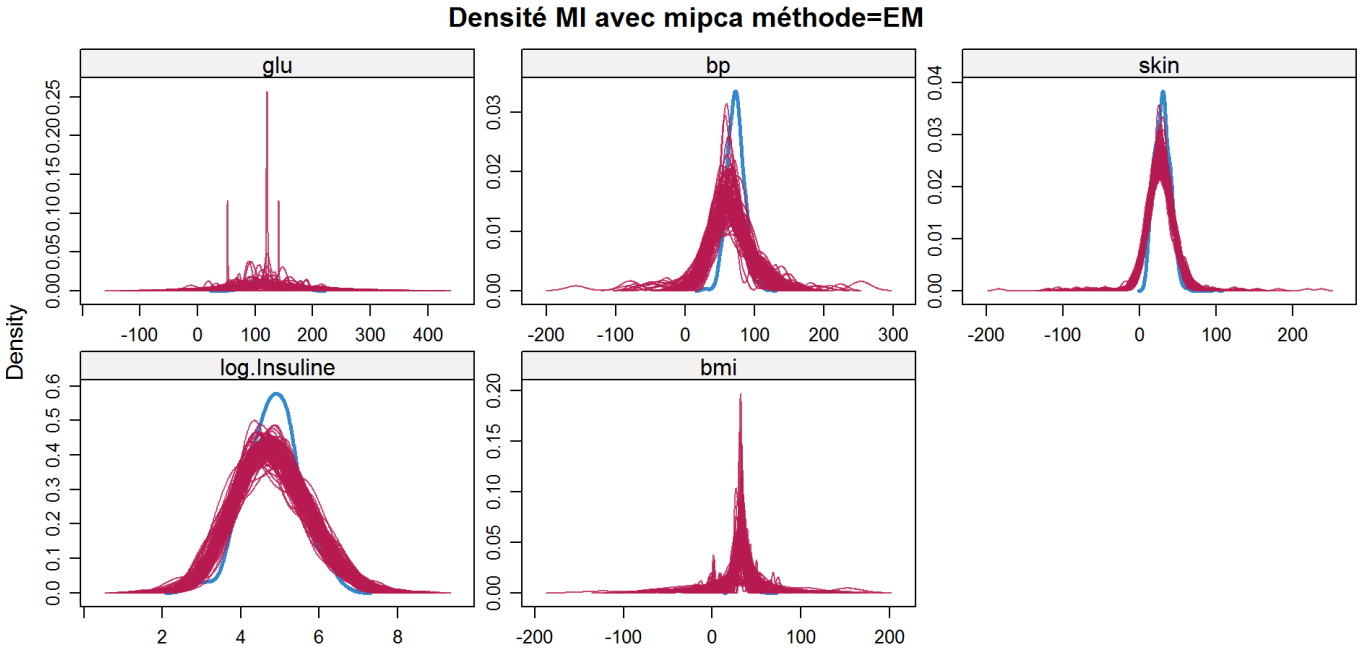
Cette dernière imputation semble *in fine* la plus appropriée. Plus de 90% des réimputations se situent dans l'intervalle de confiance.



- 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.

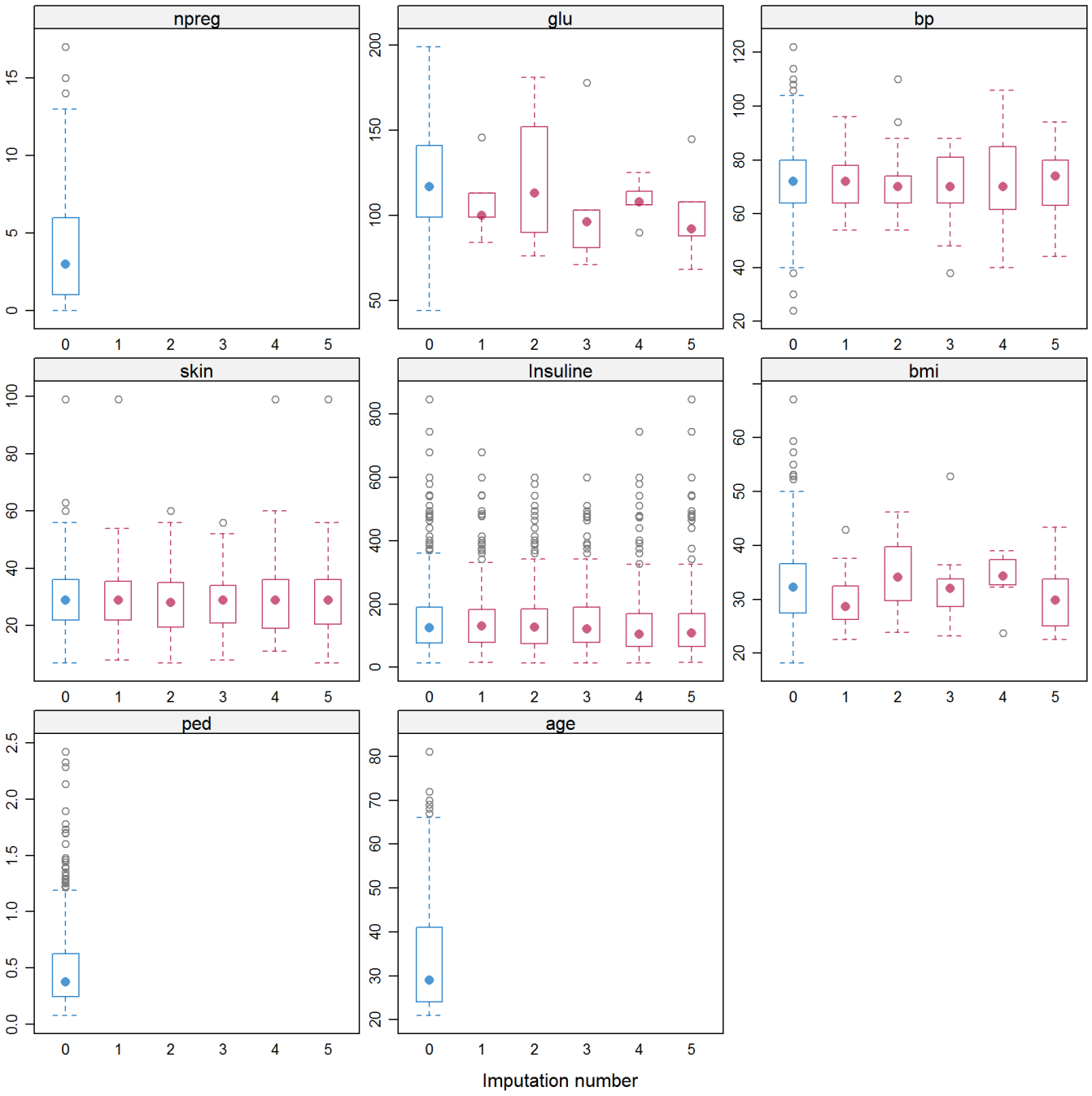




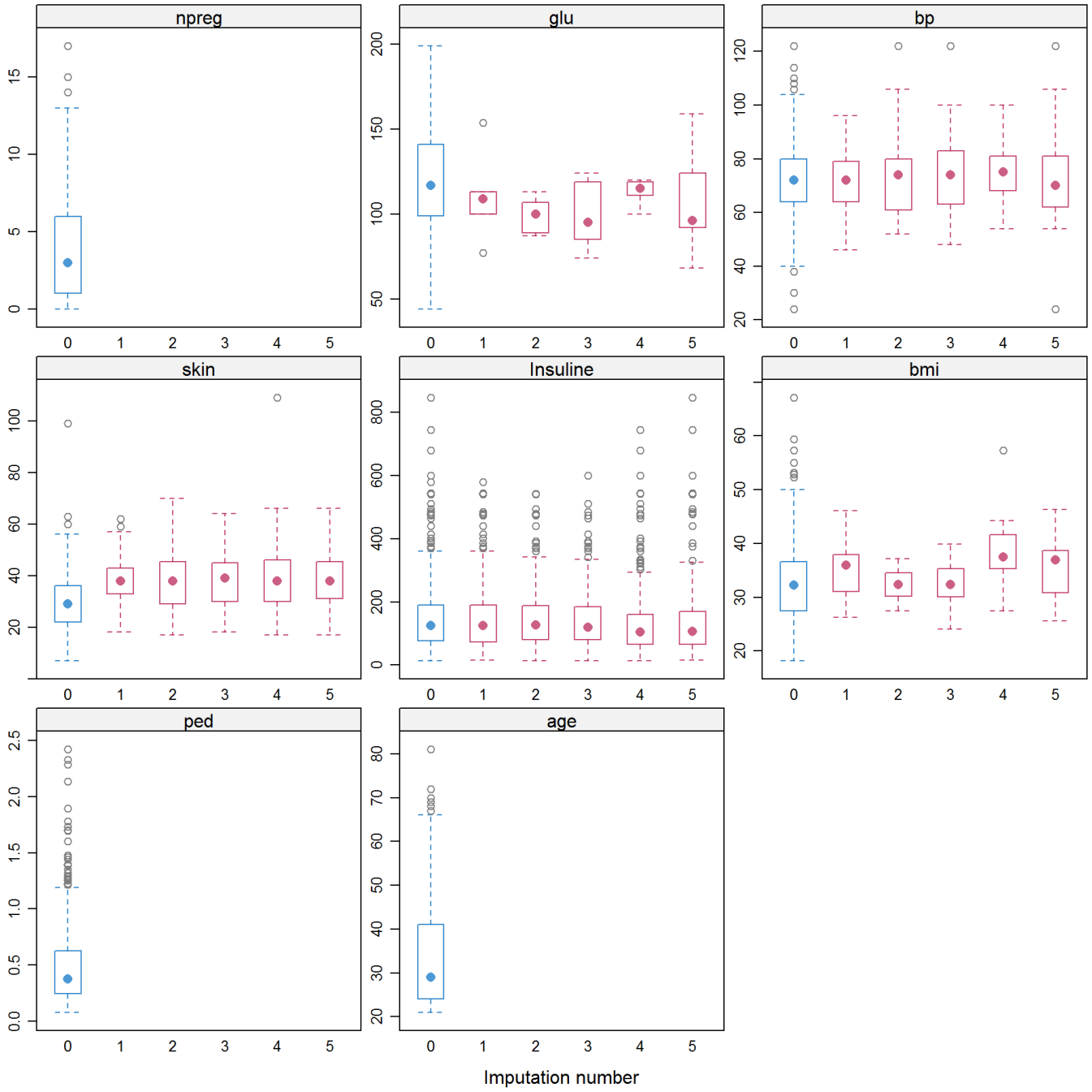
4. Analyse de sensibilité

L'analyse de sensibilité ne concerne que le modèle d'imputation multiple Avec pour hyptohèse 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).

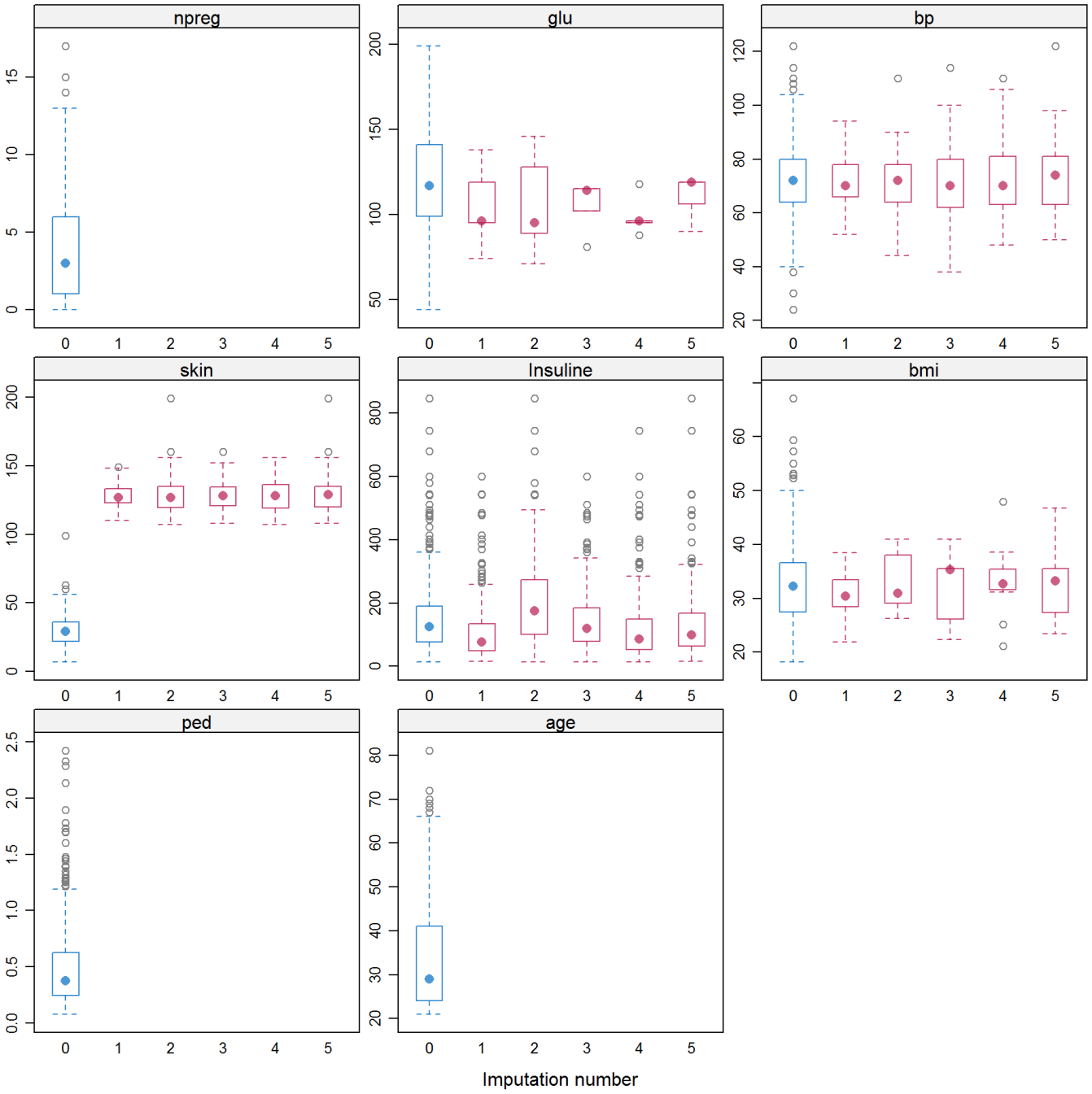
```
delta <- c(0,10,100)
imp.all <- vector("list", length(delta))
post <- mice(data[,1:8], maxit = 0)$post
for (i in 1:length(delta)){
  d <- delta[i]
  cmd <- paste("imp[[j]][,i] <- imp[[j]][,i] +", d)
  post["skin"] <- cmd
  imp <- mice(data[,1:8], post = post, maxit = 5, m=5, seed=111119, print = F)
  imp.all[[i]] <- imp
}
bwplot(imp.all[[1]])
```




```
bwplot(imp.all[[2]])
```




```
bwplot(imp.all[[3]])
```


npreg
glu
bp
skin
Insuline
bmi
ped
age
Imputation number

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.glm <- with(data=conv.imp.mipca.EM.boot.new ,exp=glm(type ~ npreg + glu + bp + skin + log.Insul
ine + 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)	-11.473132329	1.233102686	-9.3042797	415.2989	0.000000e+00
npreg	0.123452231	0.032913760	3.7507787	736.3944	1.901296e-04
glu	0.028974981	0.004255850	6.8082703	618.3012	2.342304e-11
bp	-0.008024487	0.008892683	-0.9023696	679.2862	3.671804e-01
skin	0.022192758	0.014056824	1.5787889	387.5162	1.152004e-01
log.Insuline	0.745989623	0.248061862	3.0072725	252.7903	2.901862e-03
bmi	0.065465844	0.019980012	3.2765668	591.8697	1.112249e-03
ped	0.829602597	0.302049957	2.7465741	734.1867	6.169650e-03
age	0.011515063	0.009782116	1.1771545	736.8243	2.395139e-01

9 rows

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

	estimate <dbl>	std.error <dbl>	statistic <dbl>	df <dbl>	p.value <dbl>
(Intercept)	-11.23018923	1.105783434	-10.155867	462.1078	0.000000e+00
npreg	0.13782643	0.035162928	3.919652	689.5783	9.752483e-05
glu	0.02558764	0.004503809	5.681334	607.5362	2.074468e-08
bp	-0.01738994	0.008754171	-1.986475	546.7153	4.747945e-02
skin	0.03298898	0.013471581	2.448783	317.7843	1.487394e-02
log.Insuline	0.85753460	0.205198638	4.179046	300.9873	3.840785e-05
bmi	0.05511323	0.021524661	2.560469	496.5623	1.074777e-02
ped	0.92504374	0.311661067	2.968108	695.9368	3.099262e-03
age	0.01834827	0.010456473	1.754728	696.6964	7.974538e-02

9 rows

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

	estimate <dbl>	std.error <dbl>	statistic <dbl>	df <dbl>	p.value <dbl>
(Intercept)	-8.62575377	0.853757289	-10.103286	510.2477	0.000000000
npreg	0.10756478	0.033270937	3.232995	735.1897	0.001279681
glu	0.03765235	0.003680887	10.229150	740.5630	0.000000000
bp	-0.01743161	0.008505779	-2.049384	546.1637	0.040901688
skin	0.02861586	0.013553199	2.111373	278.3677	0.035631588
bmi	0.06473609	0.020938079	3.091787	512.7718	0.002097811
ped	0.88886273	0.301582959	2.947324	748.5984	0.003305001
age	0.01994912	0.010016087	1.991708	738.1824	0.046771090

8 rows

On remarque que les variables skin (pour mice), age et bp ont une p-value assez importante.

- 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.glm, diabete.noinsuline.glm, method = "likelihood")$pvalue
```

```
## [1] 2.396047e-05
```

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

```
##    test statistic df1      df2 df.com      p.value       riv
##  1 ~~ 2  17.46443   1 374.3012    759 3.647477e-05 0.5507066
```

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 : bp et age

```
## [1] "Test du rapport de vraisemblance modele complet vs modele sans la variable age : 0.0763371217407416"
```

```
## [1] "Test du rapport de vraisemblance modele complet vs modele sans la variable skin : 0.017665142862979"
```

```
## [1] "Test du rapport de vraisemblance modele complet vs modele sans la variable insuline : 2.39604727777509e-
05"
```

```
## [1] "Test du rapport de vraisemblance modele complet vs modele sans la variable bp : 0.0277648269328263"
```

```
## [1] "Test du rapport de vraisemblance modele complet vs modele sans la variable bmi : 0.0114624223656384"
```

```
## [1] "Test du rapport de vraisemblance modele complet vs modele sans la variable glu  : 1.738609256563e-13"
```

```
## [1] "Test du rapport de vraisemblance modele complet vs modele sans la variable npreg : 3.1724938396982e-05"
```

```
## [1] "Test du rapport de vraisemblance modele complet vs modele sans la variable ped : 0.00323902343173565"
```

Toutes les variables semblent significatives, c'est le modèle complet qui a le meilleur AIC. A partir de ce AIC moyen on pourrait essayer de supprimer les variables : bmi, age et bp.

```
## [1] "AIC du modele complet : 687.872740691428"
```

```
## [1] "AIC du modele sans la variable age : 689.177761785113"
```

```
## [1] "AIC du modele sans la variable skin : 695.753913386109"
```

```
## [1] "AIC du modele sans la variable insuline : 715.424610330563"
```

```
## [1] "AIC du modele sans la variable bp : 690.883997273214"
```

```
## [1] "AIC du modele sans la variable bmi : 694.39336092397"
```

```
## [1] "AIC du modele sans la variable npreg : 703.017985200401"
```

```
## [1] "AIC du modele sans la variable glu : 725.753952742626"
```

```
## [1] "AIC du modele sans la variable ped : 695.574620108385"
```

On regarde maintenant et l'on compare certains modèles emboîtés en supprimant successivement les variables : age, bmi et bp. Le meilleur modèle qui minimise le critère BIC est le modèle diabete.insuline.glm sans les variables : bmi, age et bp Soit en ne conservant que les 5 variables: npreg, ped, glu, skin et log.Insuline C'est aussi confirmé par les différents MLE réalisés sur les modèles emboîtés. Pour argumenter ce choix on peut se référer à l'annexe 2 - choix de modèle

```
## [1] "Test du rapport de vraisemblance modele complet vs modele sans la variable age : 0.0763371217407394"
```

```
## [1] "Test du rapport de vraisemblance modele sans la variable age vs modele sans la variable age et bp: 0.086
0802995887788"
```

```
## [1] "Test du rapport de vraisemblance modele sans la variable age et bp vs modele sans la variable age, bp et
bmi: 0.0424498515374343"
```

```
## [1] "AIC du modele complet : 687.872740691428"
```

```
## [1] "AIC du modele sans la variable age : 689.177761785113"
```

```
## [1] "AIC du modele sans la variable age et bp : 690.502897697801"
```

```
## [1] "AIC du modele sans la variable age, bp et bmi: 694.564436934921"
```

```
## [1] "BIC du modele complet : 729.666848289757"
```

```
## [1] "BIC du modele sans la variable age : 726.328079650294"
```

```
## [1] "BIC du modele sans la variable age et bp : 723.009425829835"
```

```
## [1] "BIC du modele sans la variable age, bp et bmi: 722.427175333807"
```

	estimate <dbl>	**std.error** <dbl>	**statistic** <dbl>	**df** <dbl>	**p.value** <dbl>
(Intercept)	-11.23018923	1.105783434	-10.155867	462.1078	0.000000e+00

	estimate <dbl>	std.error <dbl>	statistic <dbl>	df <dbl>	p.value <dbl>
npreg	0.13782643	0.035162928	3.919652	689.5783	9.752483e-05
glu	0.02558764	0.004503809	5.681334	607.5362	2.074468e-08
bp	-0.01738994	0.008754171	-1.986475	546.7153	4.747945e-02
skin	0.03298898	0.013471581	2.448783	317.7843	1.487394e-02
log.Insuline	0.85753460	0.205198638	4.179046	300.9873	3.840785e-05
bmi	0.05511323	0.021524661	2.560469	496.5623	1.074777e-02
ped	0.92504374	0.311661067	2.968108	695.9368	3.099262e-03
age	0.01834827	0.010456473	1.754728	696.6964	7.974538e-02

9 rows

	estimate <dbl>	std.error <dbl>	statistic <dbl>	df <dbl>	p.value <dbl>
(Intercept)	-11.04067910	1.090456536	-10.124823	467.8522	0.000000e+00
npreg	0.16910218	0.030586114	5.528724	690.6388	4.577096e-08
glu	0.02647900	0.004472863	5.919922	610.5123	5.374825e-09
skin	0.03144380	0.013391854	2.347979	317.5859	1.948877e-02
log.Insuline	0.86650192	0.204314137	4.241028	303.5846	2.957352e-05
ped	0.93848498	0.309775160	3.029568	696.7771	2.539624e-03
bp	-0.01339933	0.008542912	-1.568473	510.8368	1.173900e-01
bmi	0.05238066	0.021531251	2.432773	486.6697	1.534295e-02

8 rows

	estimate <dbl>	std.error <dbl>	statistic <dbl>	df <dbl>	p.value <dbl>
(Intercept)	-11.59024584	1.033259015	-11.217174	479.5791	0.000000e+00
npreg	0.15991672	0.029842422	5.358704	693.3701	1.142104e-07
glu	0.02553309	0.004413431	5.785316	602.0179	1.164245e-08
skin	0.03153331	0.013563746	2.324823	297.9213	2.075330e-02
log.Insuline	0.86247299	0.205980484	4.187159	287.8886	3.758143e-05
ped	0.94423622	0.306896187	3.076728	697.6892	2.174730e-03
bmi	0.04481442	0.021371407	2.096934	450.7490	3.655608e-02

7 rows

	estimate <dbl>	std.error <dbl>	statistic <dbl>	df <dbl>	p.value <dbl>
(Intercept)	-10.83676467	0.981339759	-11.042826	414.2328	0.000000e+00
npreg	0.15135850	0.029276323	5.169997	690.5159	3.069309e-07
glu	0.02617796	0.004387507	5.966478	602.0101	4.138492e-09
skin	0.04841036	0.010630370	4.553967	275.8753	7.907462e-06
log.Insuline	0.89585879	0.202254123	4.429372	295.8230	1.331910e-05
ped	0.99219004	0.303056415	3.273945	686.1696	1.113808e-03

6 rows

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 variablilité 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.

```
Class: mipo    m = 100
                 estimate         ubar            b            t dfcom       df        riv     lambda        fmi
(Intercept)  -10.88502375 7.231029e-01 3.239346e-01 1.050277e+00   762 345.8641 0.45245827 0.31151206 0.31545909
npreg          0.15107510 8.032095e-04 5.978041e-05 8.635877e-04   762 683.0321 0.07517120 0.06991556 0.07262705
glu            0.02602838 1.685869e-05 2.779750e-06 1.966624e-05   762 574.4618 0.16653414 0.14275977 0.14572876
skin           0.04820190 7.002272e-05 3.368601e-05 1.040456e-04   762 329.4700 0.48588328 0.32699963 0.33104812
log.Insuline   0.90936111 2.642584e-02 1.944355e-02 4.606383e-02   762 242.1640 0.74313588 0.42632126 0.43100122
ped            1.00577056 8.573742e-02 5.498982e-03 9.129139e-02   762 695.2185 0.06477885 0.06083784 0.06352801
```

Au final, les estimations des coefficients sont du même ordre de grandeur en comparaison de la méthode des cas concrets (= listwise deletion). Mais de manière général sont plus faibles. Et en particulier biensûr pour les variables à fortes données manquantes: insuline et skin.

```
##
## Call:
## glm(formula = type ~ npreg + glu + skin + I(log(Insuline)) +
##     ped, family = binomial(link = "logit"), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0404  -0.6731  -0.3701   0.6458   2.5513
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -8.835302   1.183463  -7.466 8.29e-14 ***
## npreg             0.133916   0.039773   3.367  0.00076 ***
## glu               0.036197   0.005641   6.417 1.39e-10 ***
## skin              0.040632   0.013161   3.087  0.00202 **
## I(log(Insuline))  0.250019   0.250644   0.998  0.31852
## ped               1.091993   0.400249   2.728  0.00637 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.90  on 392  degrees of freedom
## Residual deviance: 354.25  on 387  degrees of freedom
##   (375 observations deleted due to missingness)
## AIC: 366.25
##
## Number of Fisher Scoring iterations: 5
```

En comparaison de la méthode des cas concrets (= listwise deletion), avec un modèle automatique de sélection de variables nous ne conserverions pas la variable insuline ni skin au risque de détériorer le caractère prédictif du modèle.

```
diabete.listwisedeletion.glm1 <- stepAIC(glm(type ~ npreg + glu + bp + skin + I(log(Insuline)) + bmi+ ped + age,
data=data[ok,], family = binomial(link="logit")), direction="both", trace = F)
summary(diabete.listwisedeletion.glm1)
```

```
##
## 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
```

6. Conclusion

Dans notre cas, nous avons réussi à démontrer la pertinence de l'imputation avec pour objectif l'implémentation d'un modèle de classification.

L'analyse comparative des données complétés par imputation multiple avec les packages mice et missMDA donne des structures similaires, lorsque l'on fait une analyse PCA (voir Annexe 3). De même, pour la structure des corrélations 2 à 2 entre les variables (cf. Annexe 4) on retrouve les mêmes formes générales.

Les méthodes sont nombreuses et nous n'avons pas pu toutes les expérimentés. Notament les méthodes bayésiennes de mipCA en ce qui concerne la détermination des paramètres optimaux. Ainsi que le package AMELIA couplé avec Zelig qui n'ont pas été utilisé.

Une question est apparue concernant l'utilisation pratique des données imputés. De nombresue méthodes R nécessitent en input un dataframe ou équivalent. Ici les méthode utilisent du boostraping, on a pour résultat une famille de structures comparables à un dataFrame. Que faut-il faire pour se ramener à un dataFrame: utiliser la moyenne? ou par exemple dans le cas de la méthode mipca utiliser la variable $res.ImputPCA? 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.

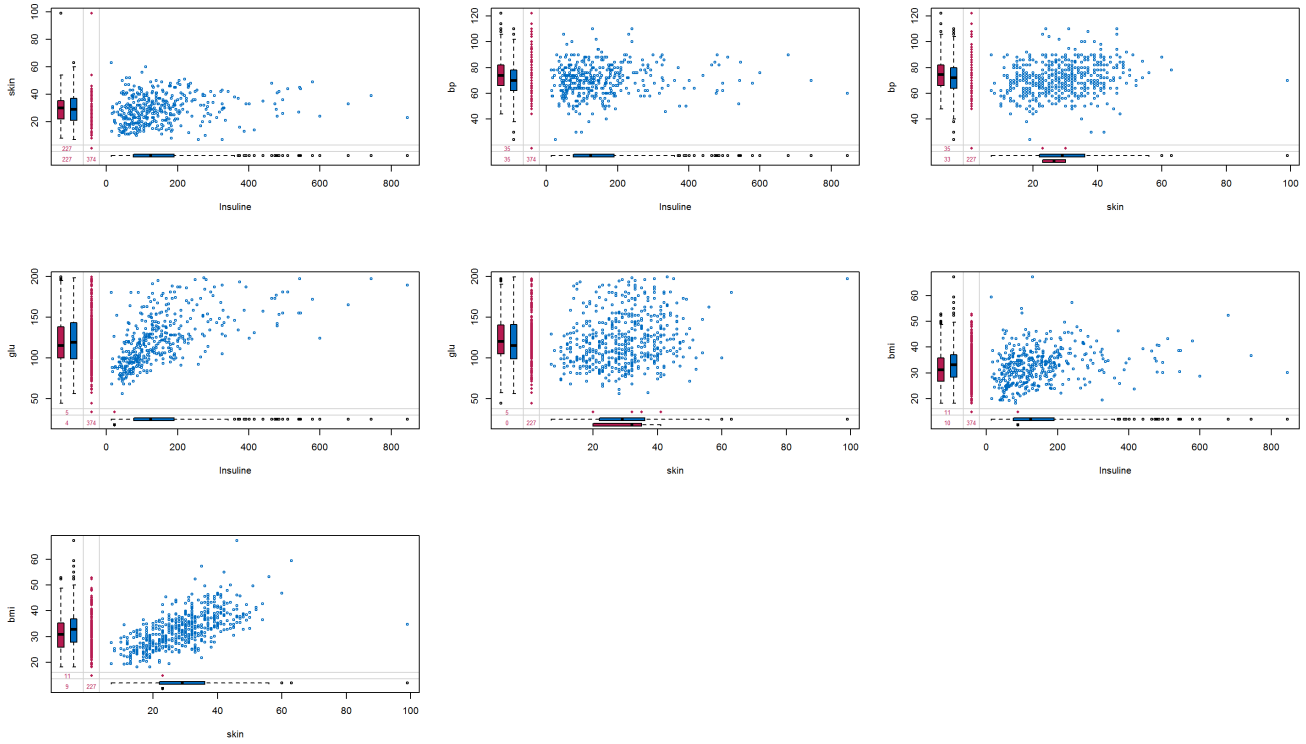
Les données Pima sont très utilisées (cf. Kaggle). Pour autant il ne semble pas qu'un modèle d'inputation ait été utilisé. Bien souvent les données manquantes sont supprimées (analyse des cas concrets ou listwise deletion) ou bien une imputation simple par moyenne est mise en oeuvre.

Annexes

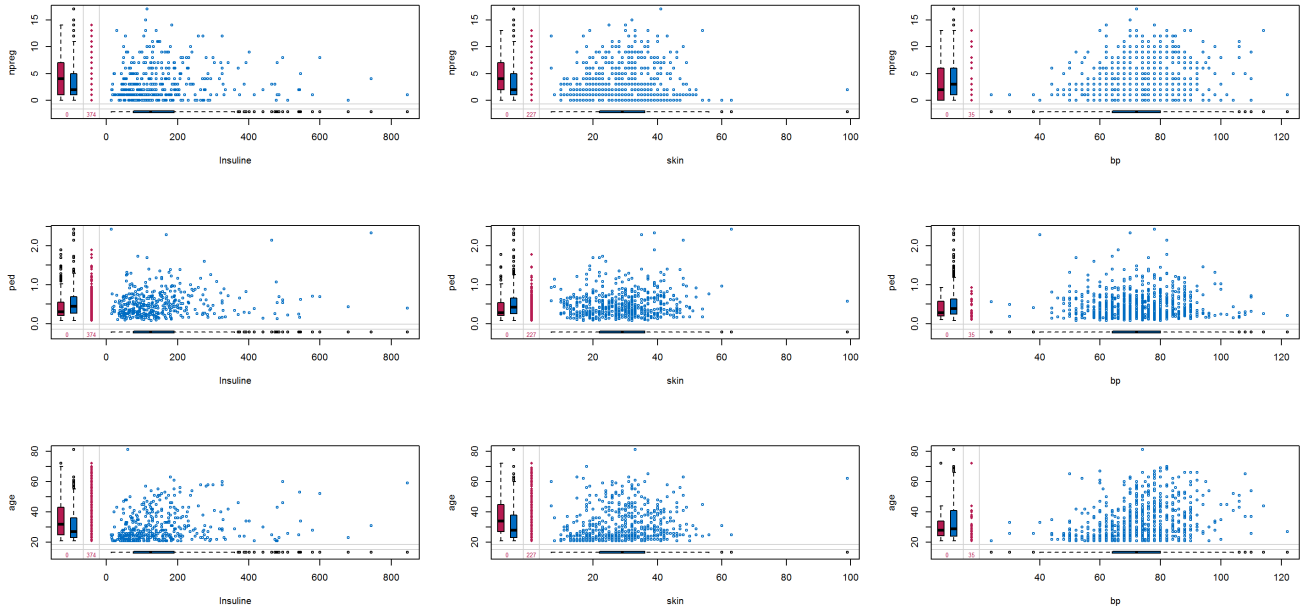
Annexe 1 : Graphiques

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

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



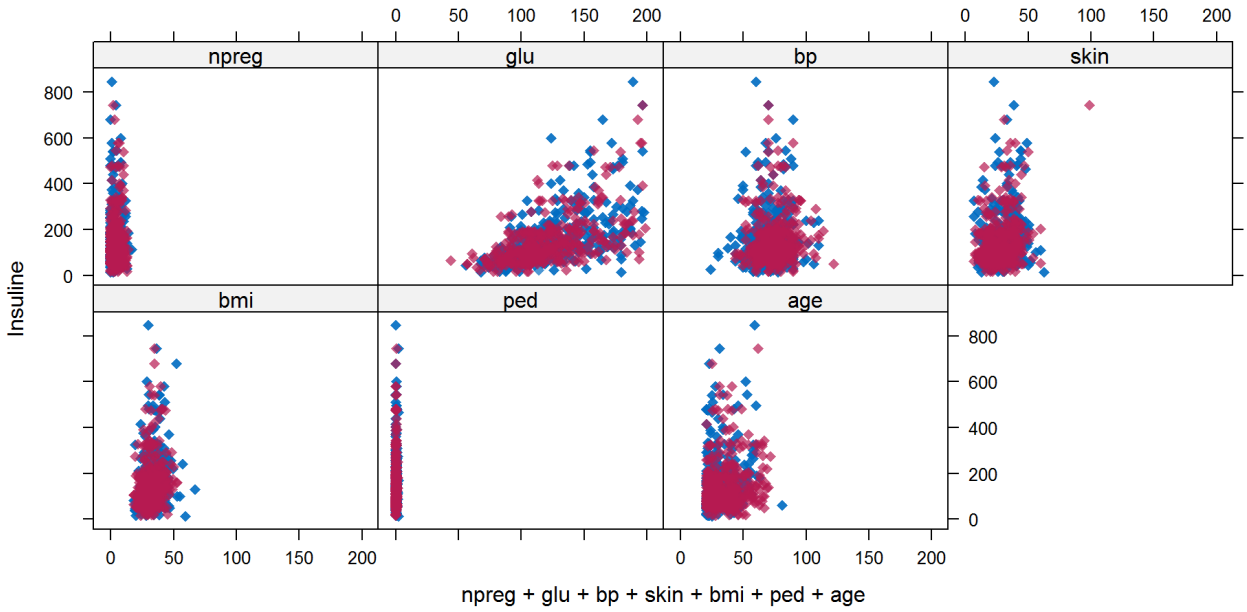
- Distributions marginales des mêmes variables avec les variables complètes

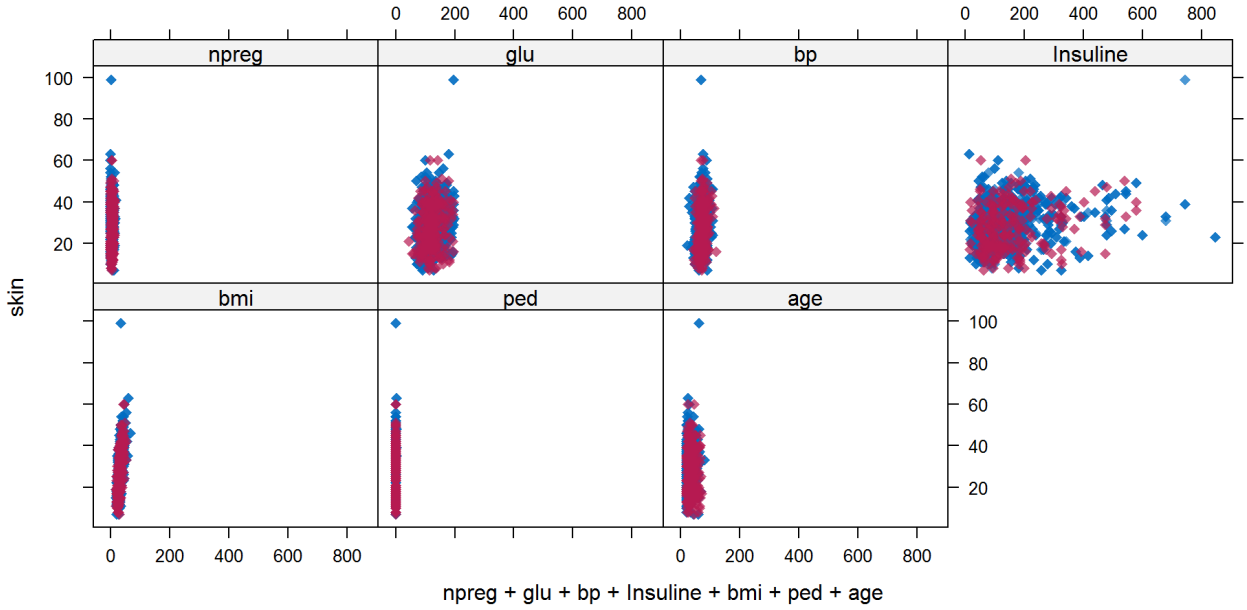


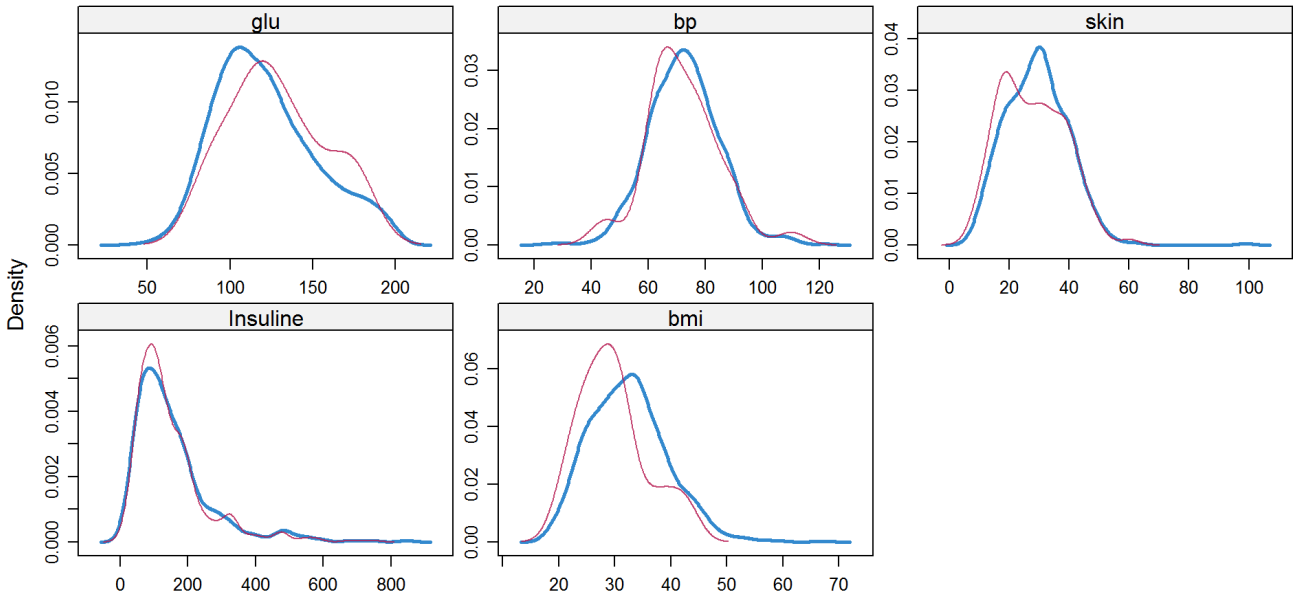
graphiques N°2 - Analyses des distributions pour les Imputations simples (§-3.1.3)

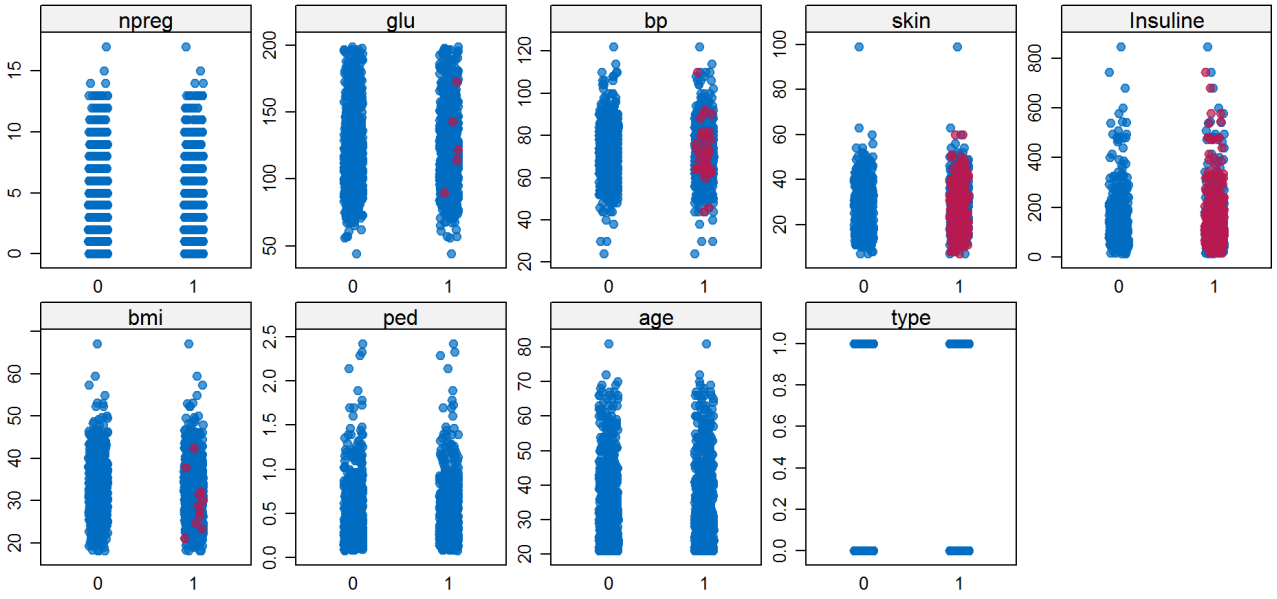
Les graphiques montrent que les imputations par régression linéaire stochastique comportent des valeurs aberrantes, en l'espèce, des valeurs négatives pour l'épaisseur de peau et le taux d'insuline. Les distributions des valeurs observées et imputées sont proches, ce qui n'infirme ni ne confirme le mécanisme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

- Imputations simples - PMM

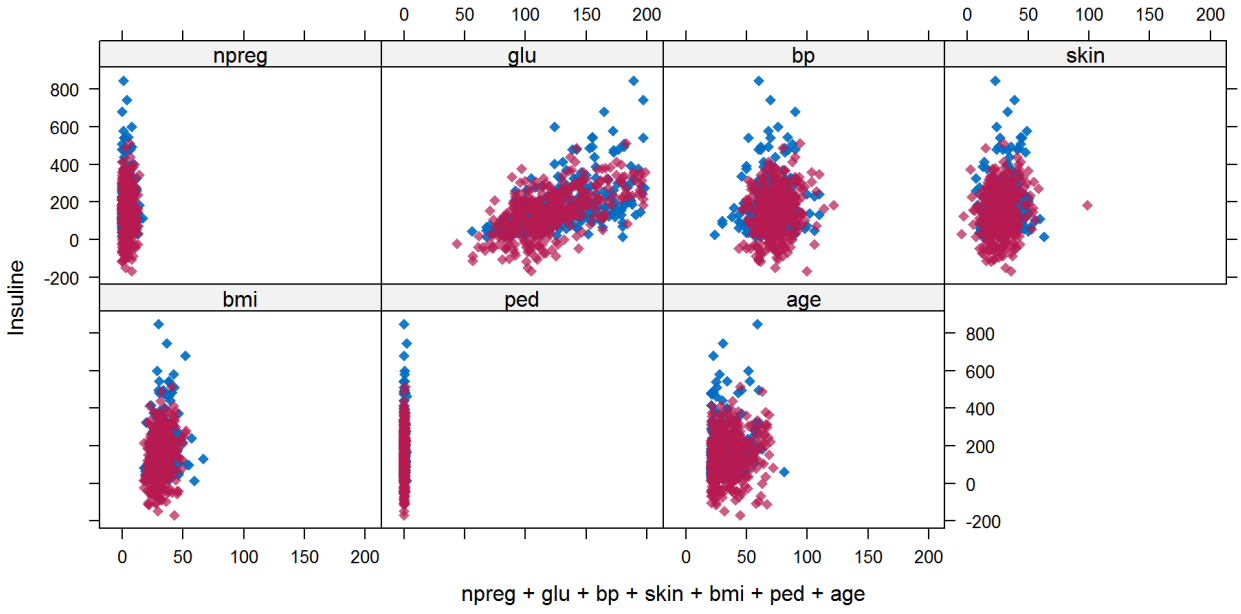


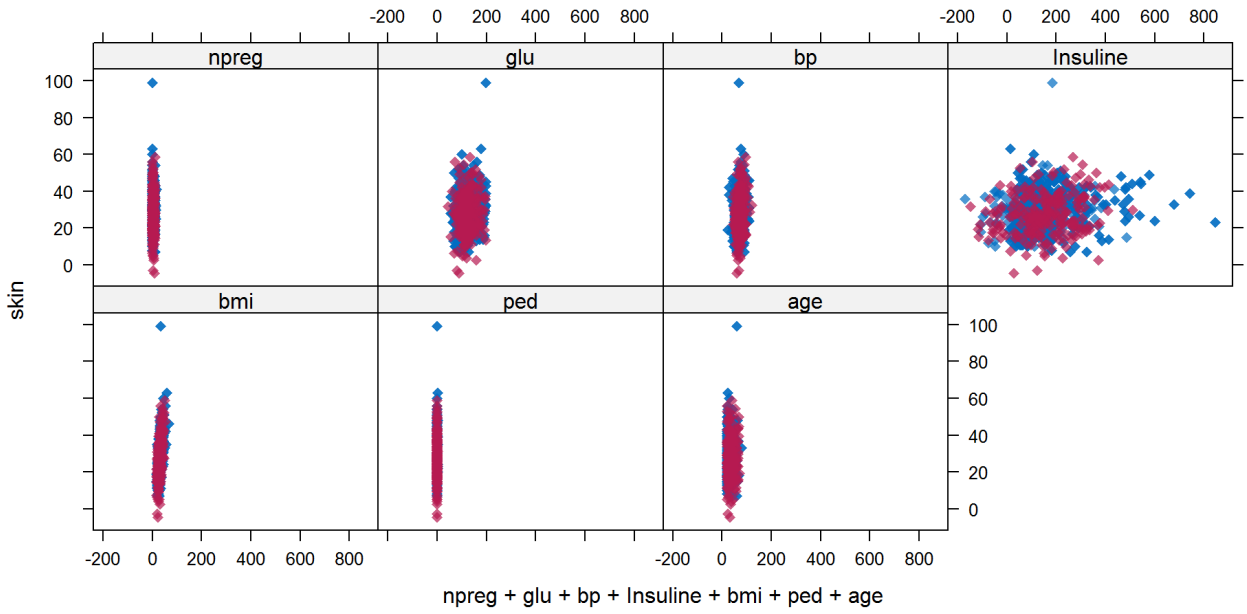


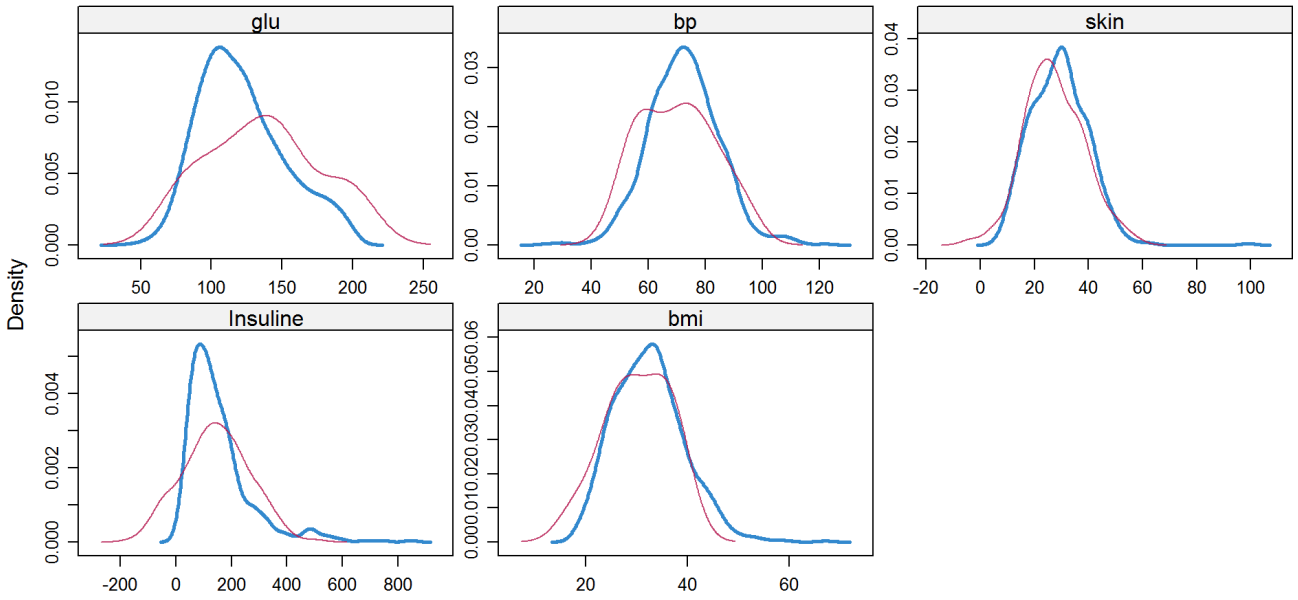


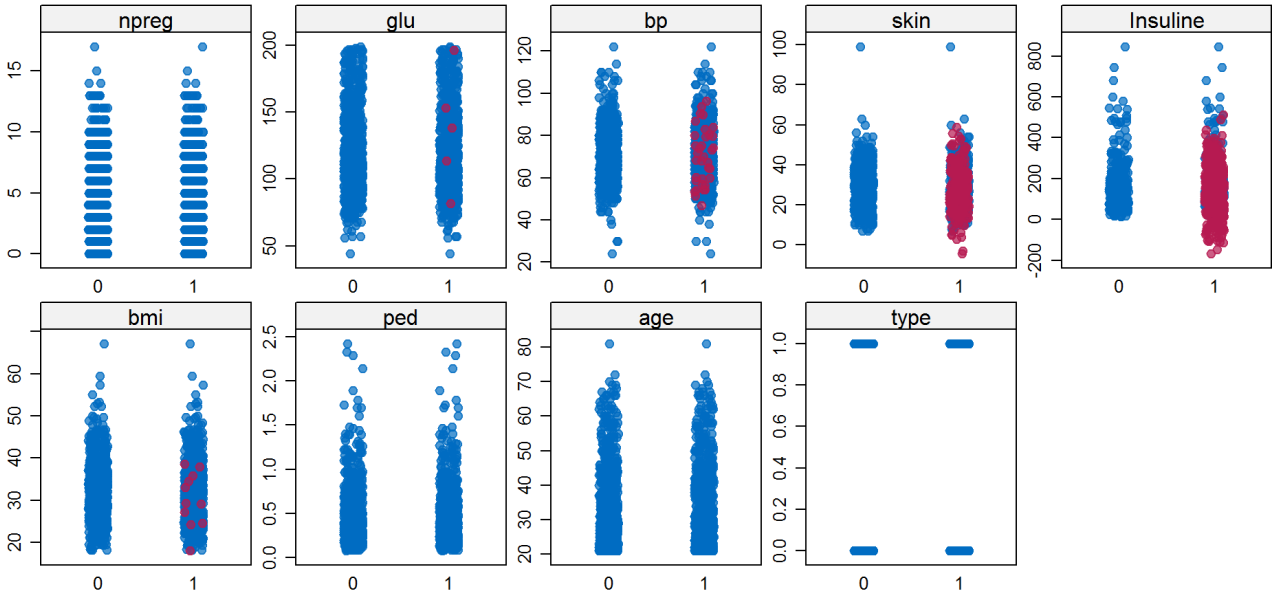


- Imputations simples - Normales




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

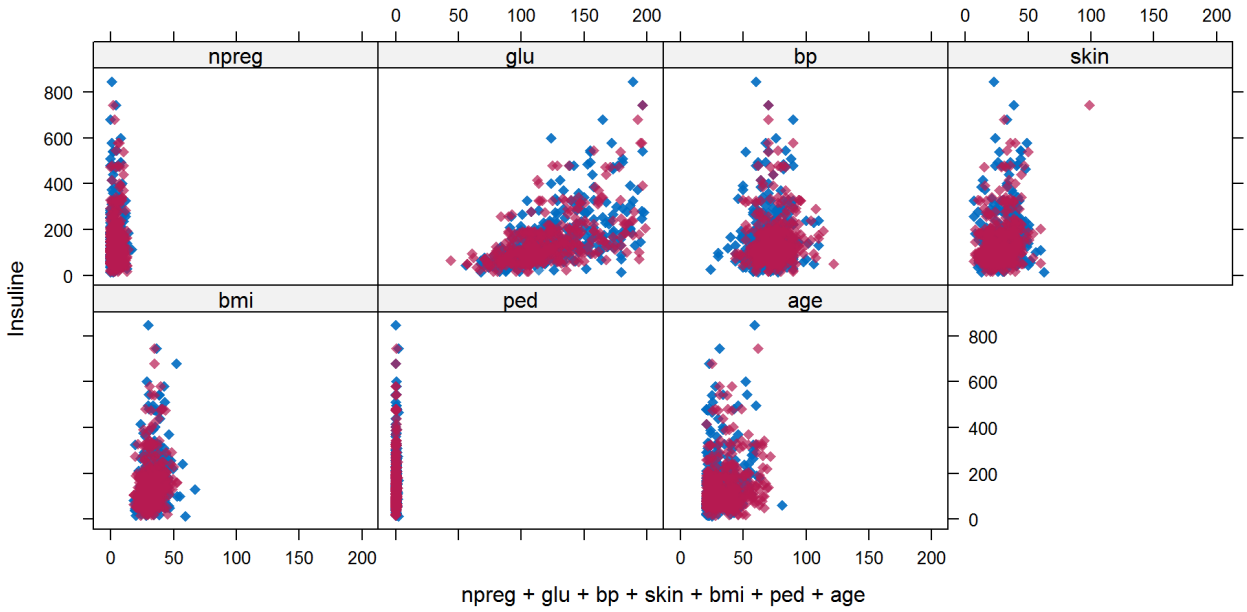

glu
bp
skin
Insuline
bmi
Density

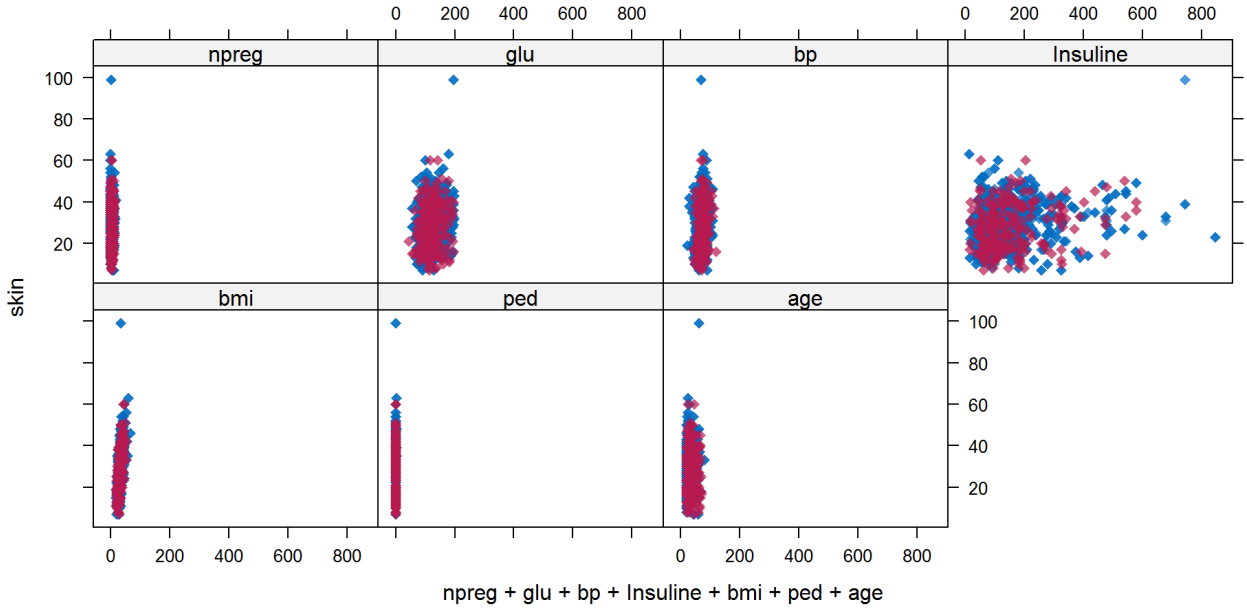

npreg
glu
bp
skin
Insuline
bmi
ped
age
type

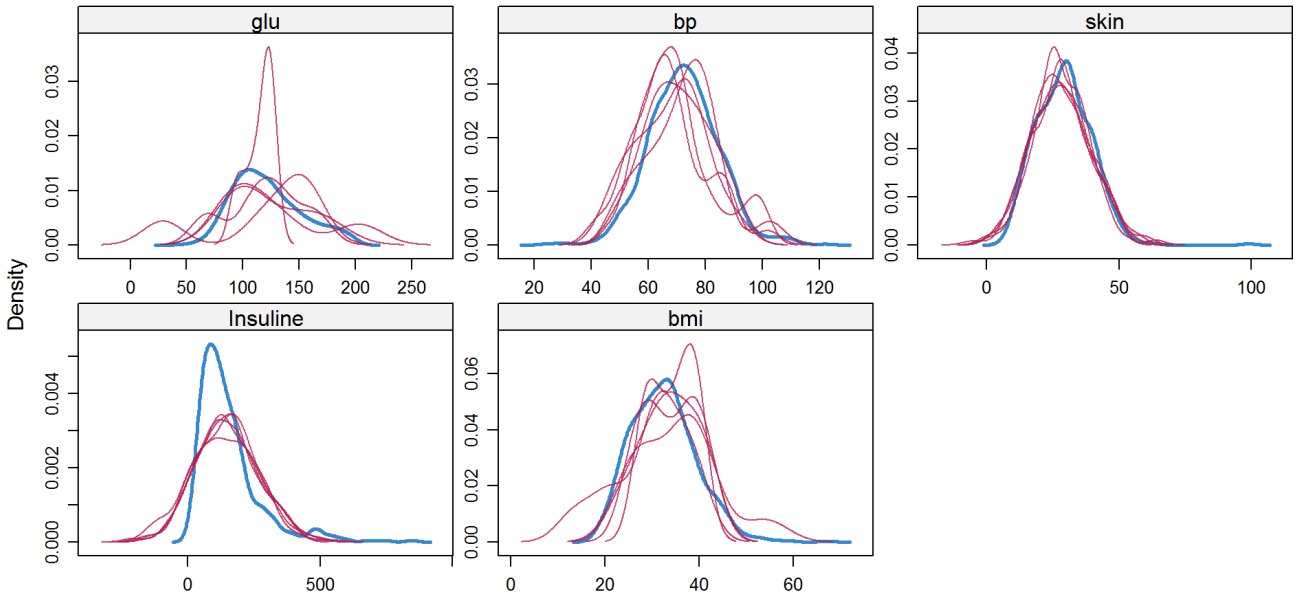
graphiques N°3 - Analyses des distributions pour les Imputations multiples (§-3.1.3)

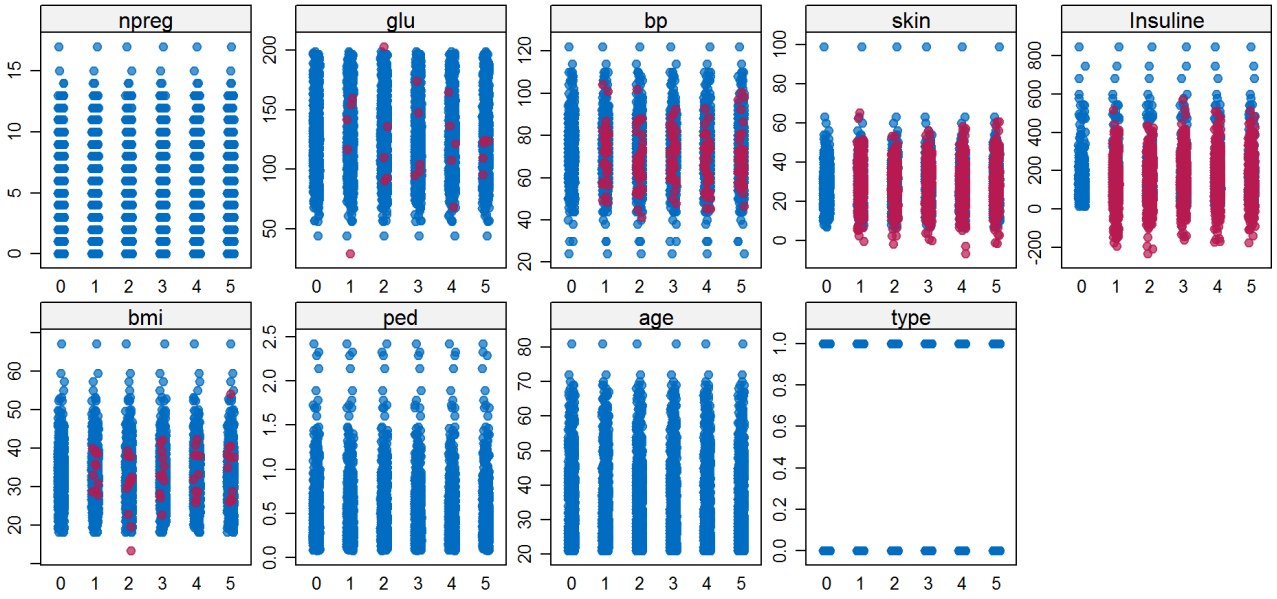
A nouveau, certaines imputations ne sont pas plausibles : avec le modèle joint les graphiques montrent des valeurs aberrantes pour l'épaisseur de peau et le taux d'insuline. S'agissant de la méthode FCS, les distributions des variables glu et bmi présentent des profils très divergents pour chacune des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

- Imputations multiples - JM

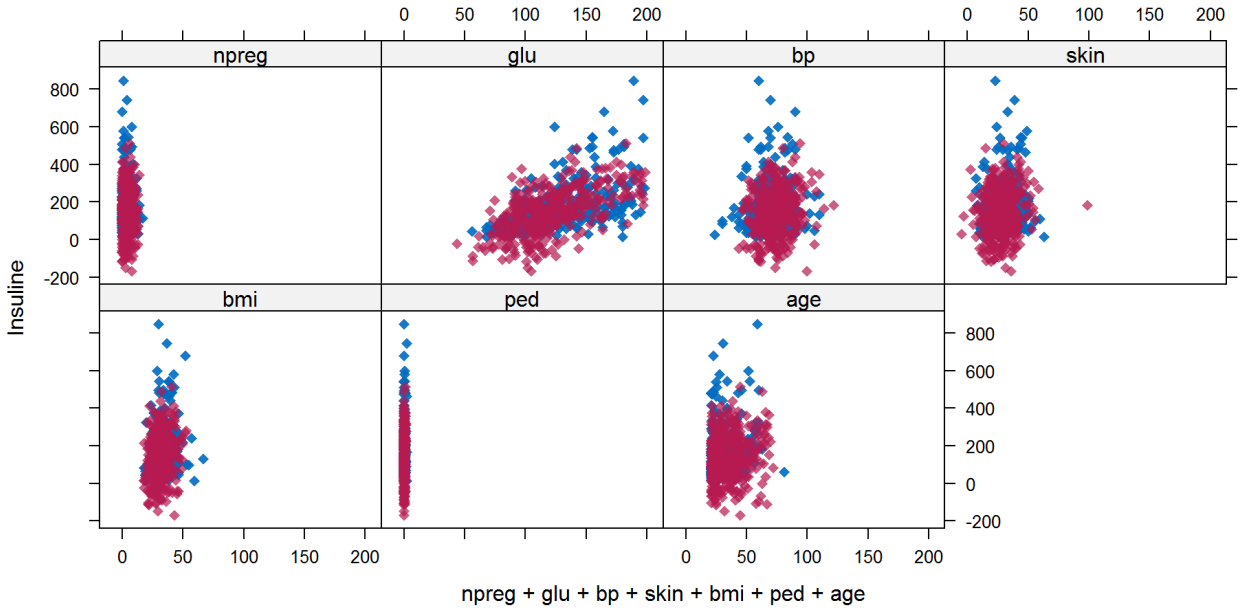


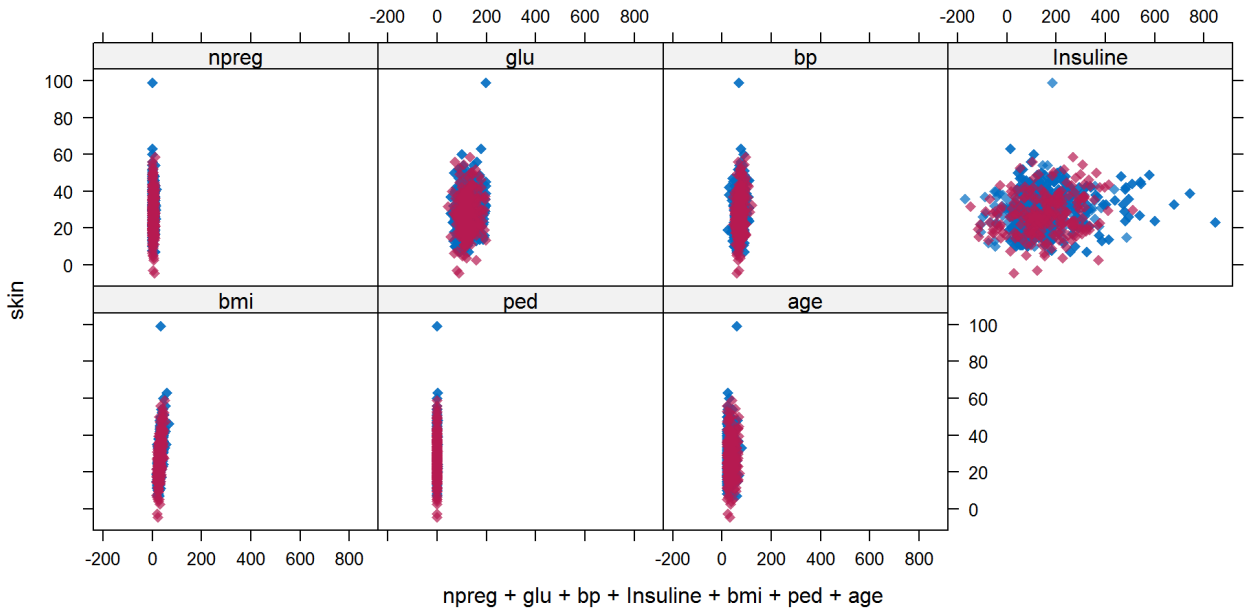


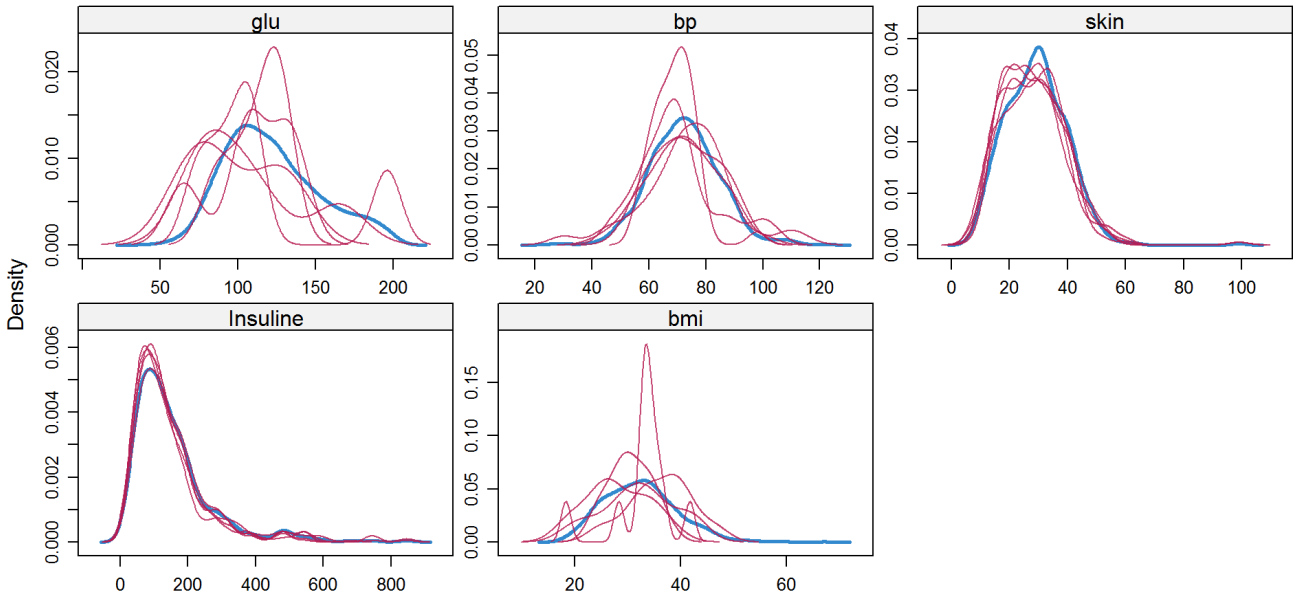


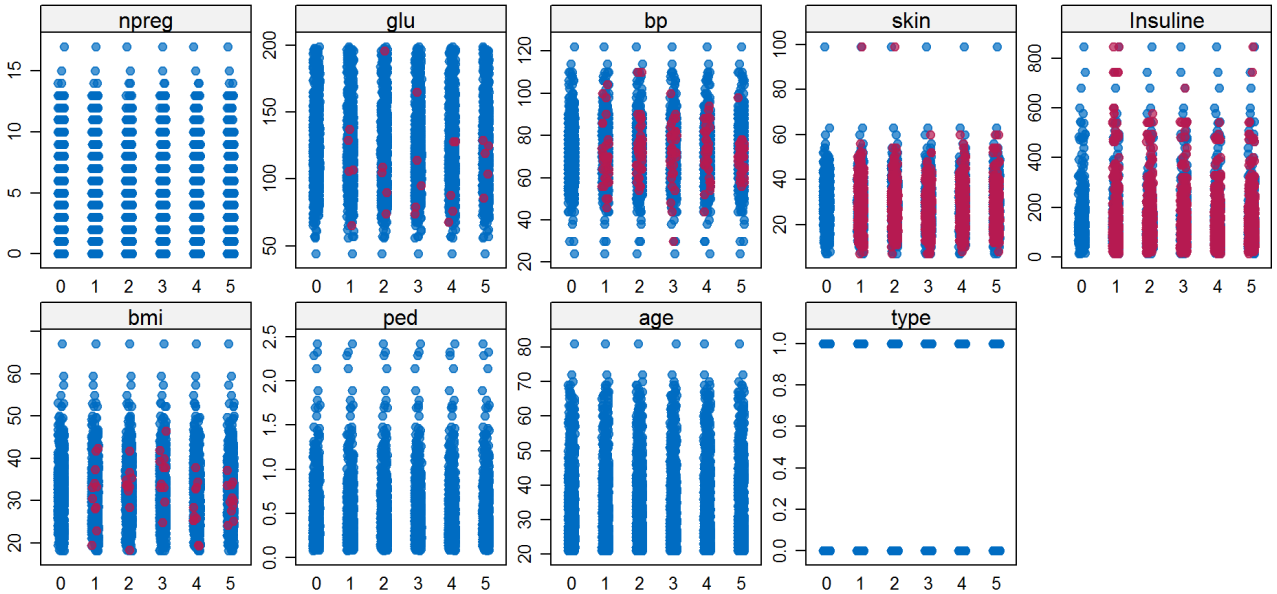


- Imputations multiples - FCS




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

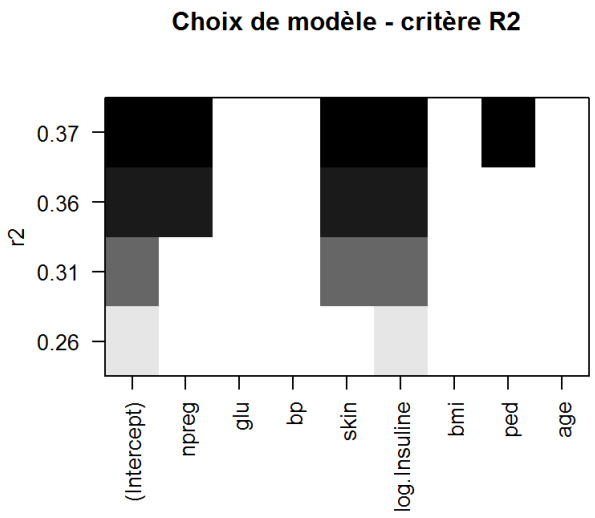

glu
bp
skin
Insuline
bmi
Density

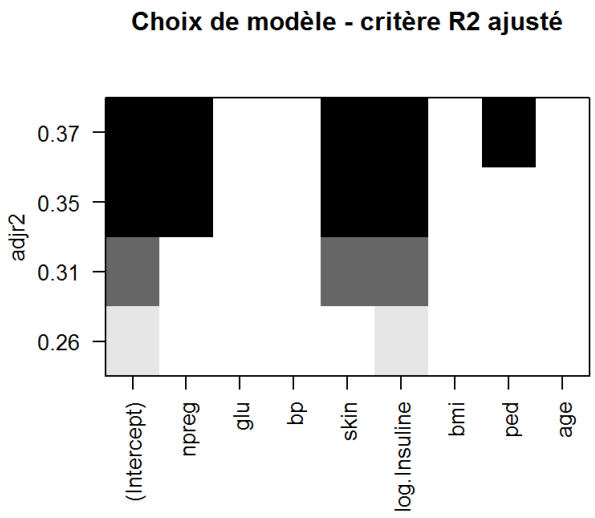

npreg
glu
bp
skin
Insuline
bmi
ped
age
type

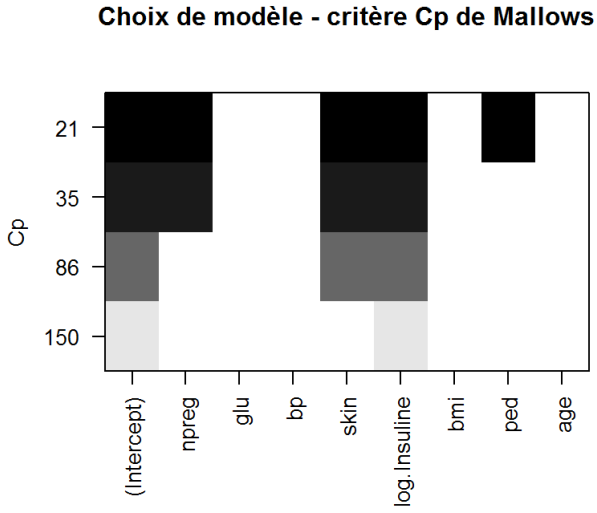
Annexe 2 - choix de modèle

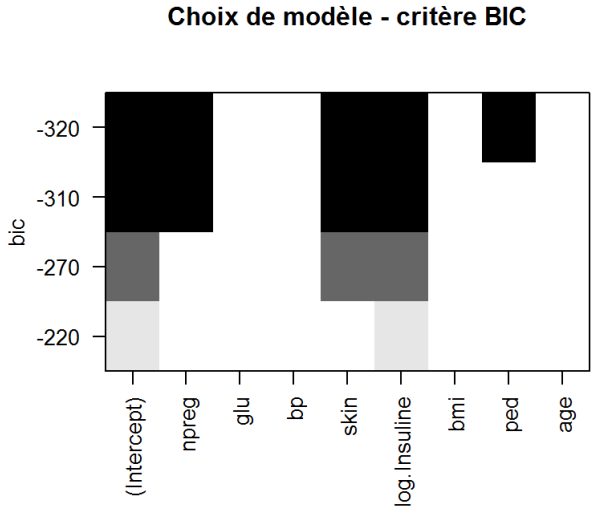
Dans cette annexe on applique les techniques regsubset et step pour le choix automatique des variables. Les données utilsées sont les données obtenu à partir de la variable $res.inputePCA On obtient à nouveau ce résultat en utilsant comme données d'input les moyennes des diffrérents jeux de données issues 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 ~ .  ,  family = binomial(link="logit"),  data = res.imputePCA[,1:maxDim])
summary(m_sature)
```

```
##
## Call:
## glm(formula = type ~ ., family = binomial(link = "logit"), data = res.imputePCA[,
##     1:maxDim])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3533  -0.6207  -0.3008   0.6030   3.0587
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -13.582666   1.151711 -11.793  < 2e-16 ***
## npreg          0.151973   0.035413   4.291 1.78e-05 ***
## glu            0.014893   0.004842   3.076  0.00210 **
## bp            -0.018386   0.008602  -2.137  0.03256 *
## skin           0.056456   0.014345   3.936 8.30e-05 ***
## log.Insuline   1.623134   0.227638   7.130 1.00e-12 ***
## bmi            0.028698   0.022177   1.294  0.19565
## ped            0.983273   0.304624   3.228  0.00125 **
## age            0.021588   0.010604   2.036  0.04175 *
## ---
## 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: 633.74  on 759  degrees of freedom
## AIC: 651.74
##
## Number of Fisher Scoring iterations: 5
```

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

- Critère AIC
- Critère BIC

Choix du modèle - Modèle obtenu à partir des données imputés

```
summary(modele_step_BwdFwd_AIC)
```

```
##
## Call:
## glm(formula = type ~ npreg + glu + bp + skin + log.Insuline +
##     ped + age, family = binomial(link = "logit"), data = res.imputePCA[,
##     1:maxDim])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3924  -0.6149  -0.3075   0.6151   3.1102
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -13.304451   1.124008 -11.837  < 2e-16 ***
## npreg          0.145590   0.034766   4.188 2.82e-05 ***
## glu            0.015047   0.004842   3.107  0.00189 **
## bp            -0.015841   0.008386  -1.889  0.05888 .
## skin           0.069347   0.010370   6.687 2.27e-11 ***
## log.Insuline   1.646173   0.226420   7.270 3.58e-13 ***
## ped            1.018496   0.302159   3.371  0.00075 ***
## age            0.021184   0.010588   2.001  0.04542 *
## ---
## 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: 635.41  on 760  degrees of freedom
## AIC: 651.41
##
## Number of Fisher Scoring iterations: 5
```

```
AIC(modele_step_BwdFwd_AIC)
```

```
## [1] 651.4134
```

```
BIC(modele_step_BwdFwd_AIC)
```

```
## [1] 688.5637
```

```
summary(m_StepBwdFwd_BIC)
```

```
##
## Call:
## glm(formula = type ~ npreg + glu + skin + log.Insuline + ped,
##     family = binomial(link = "logit"), data = res.imputePCA[,
##         1:maxDim])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3949  -0.6142  -0.3125   0.6117   3.0078
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -13.699976   1.040001 -13.173  < 2e-16 ***
## npreg          0.175578   0.029819   5.888 3.91e-09 ***
## glu            0.015125   0.004724   3.202 0.001364 **
## skin           0.062768   0.009899   6.341 2.28e-10 ***
## log.Insuline   1.654175   0.225200   7.345 2.05e-13 ***
## ped            1.027049   0.298013   3.446 0.000568 ***
## ---
## 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: 641.26  on 762  degrees of freedom
## AIC: 653.26
##
## Number of Fisher Scoring iterations: 5
```

```
AIC(m_StepBwdFwd_BIC)
```

```
## [1] 653.2596
```

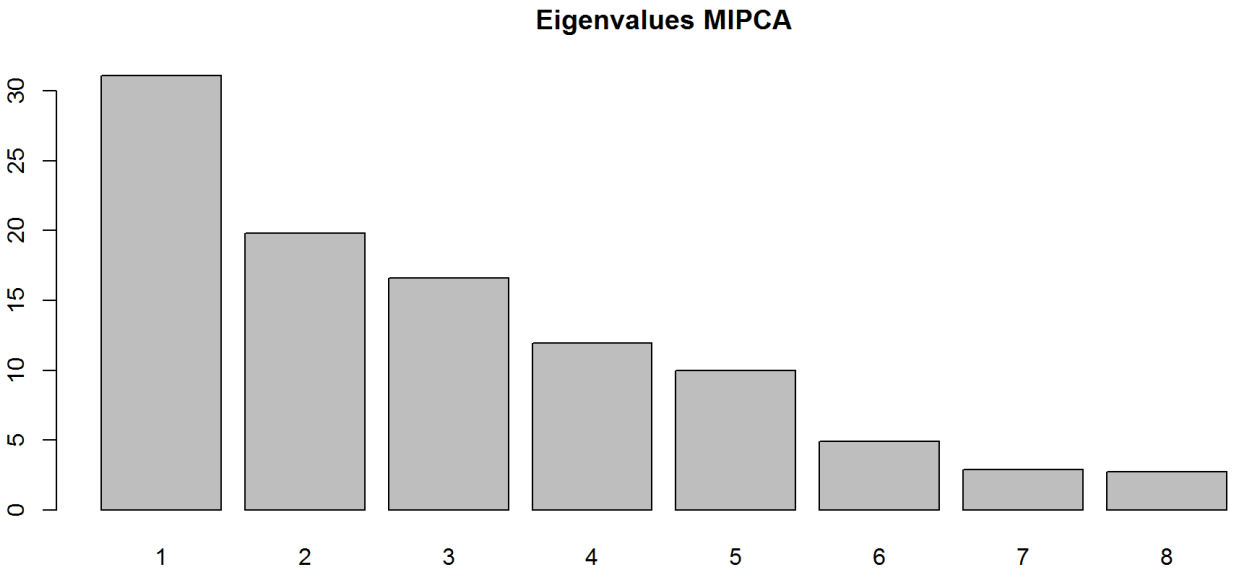
```
BIC(m_StepBwdFwd_BIC)
```

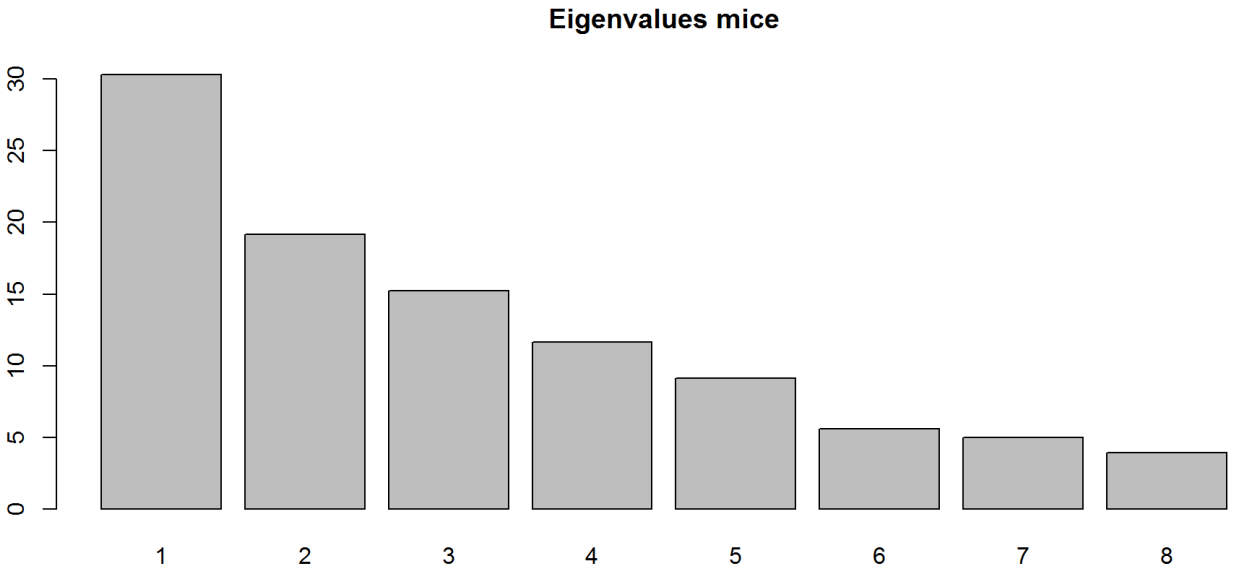
```
## [1] 681.1223
```

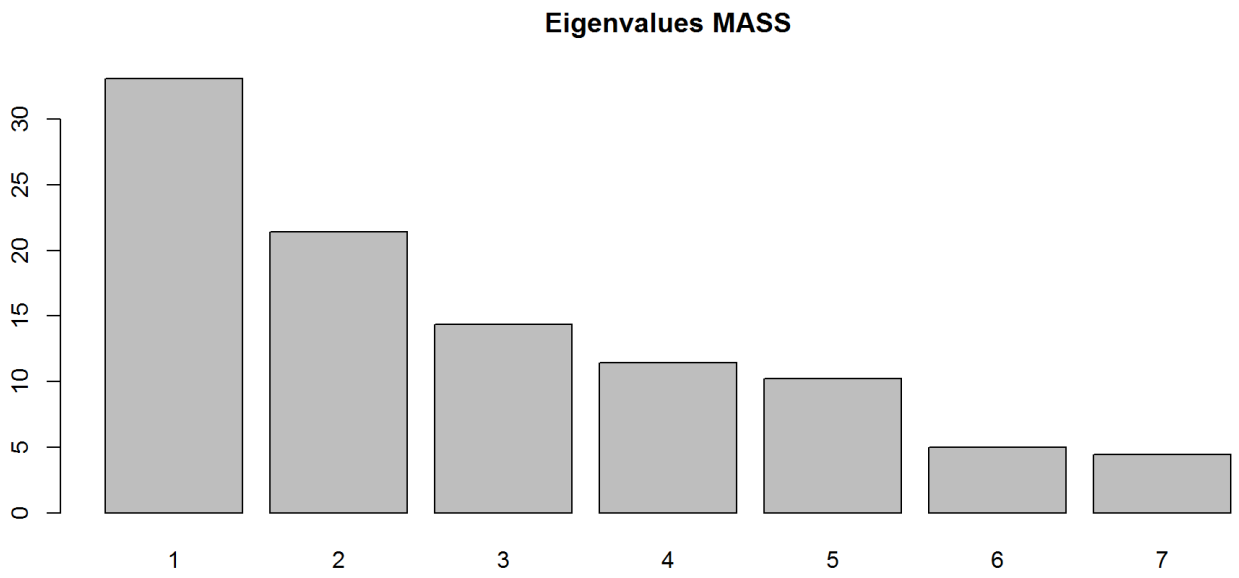
Annexe 3 : Analyse PCA - comparaison aux données non traitées

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

Choix du nombre d’axes

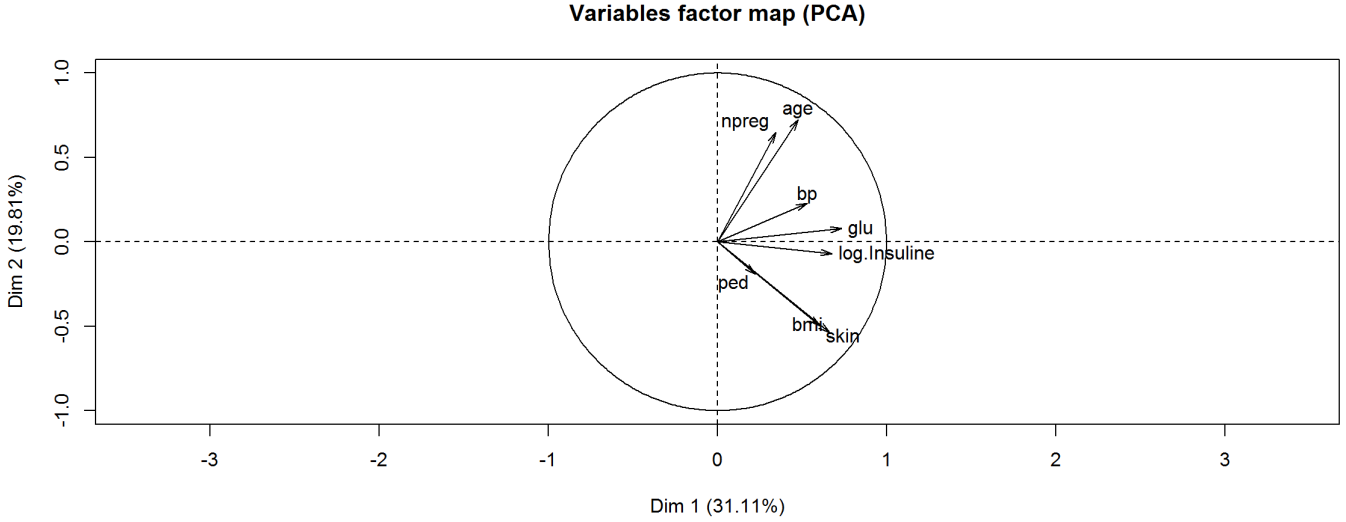


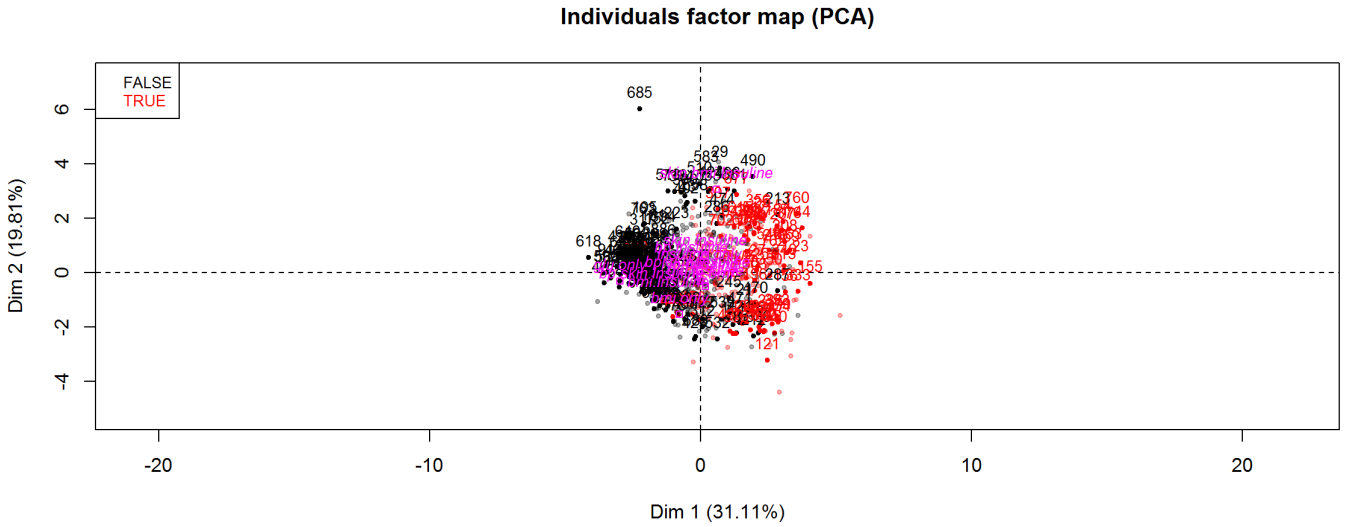


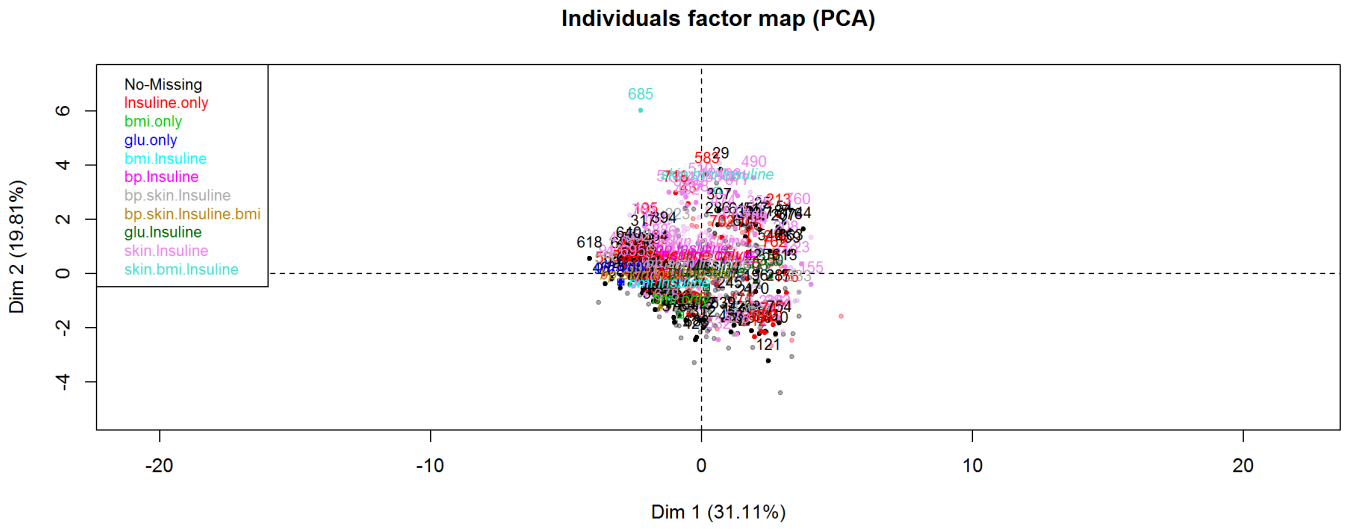


Graphiques : nuage des individus et cercle de qualité des projections

- Données Pima complétées avec la méthode MIPCA du package MissMDA

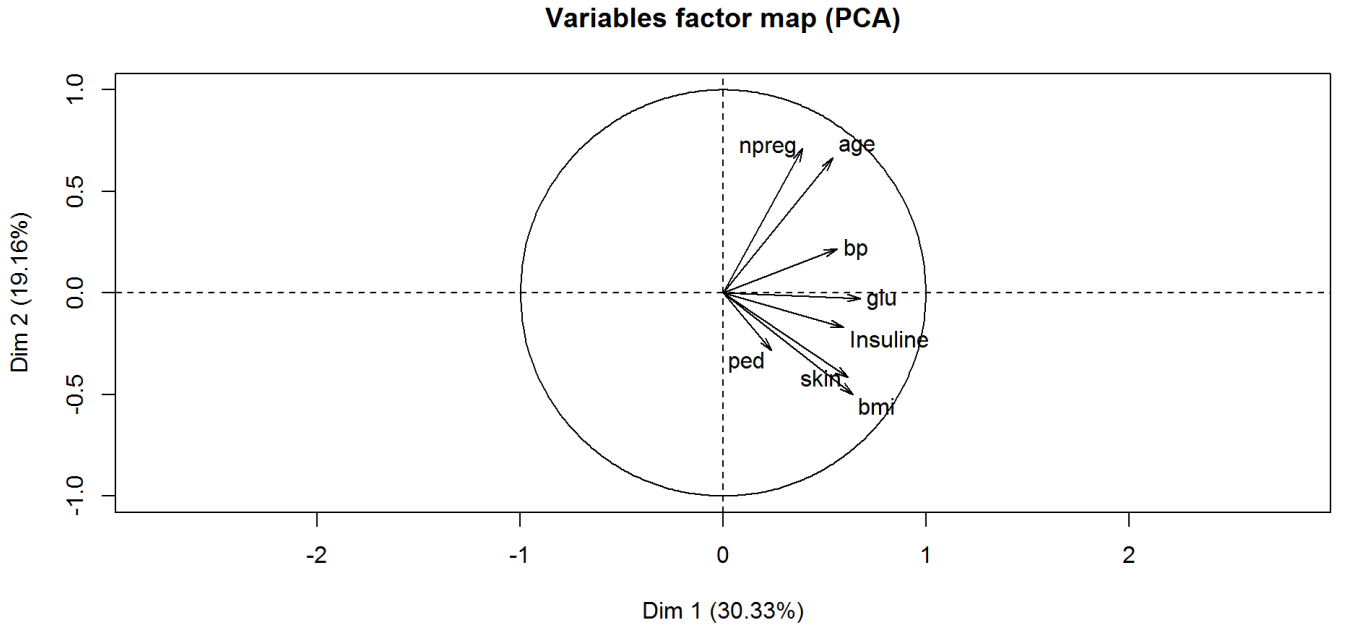


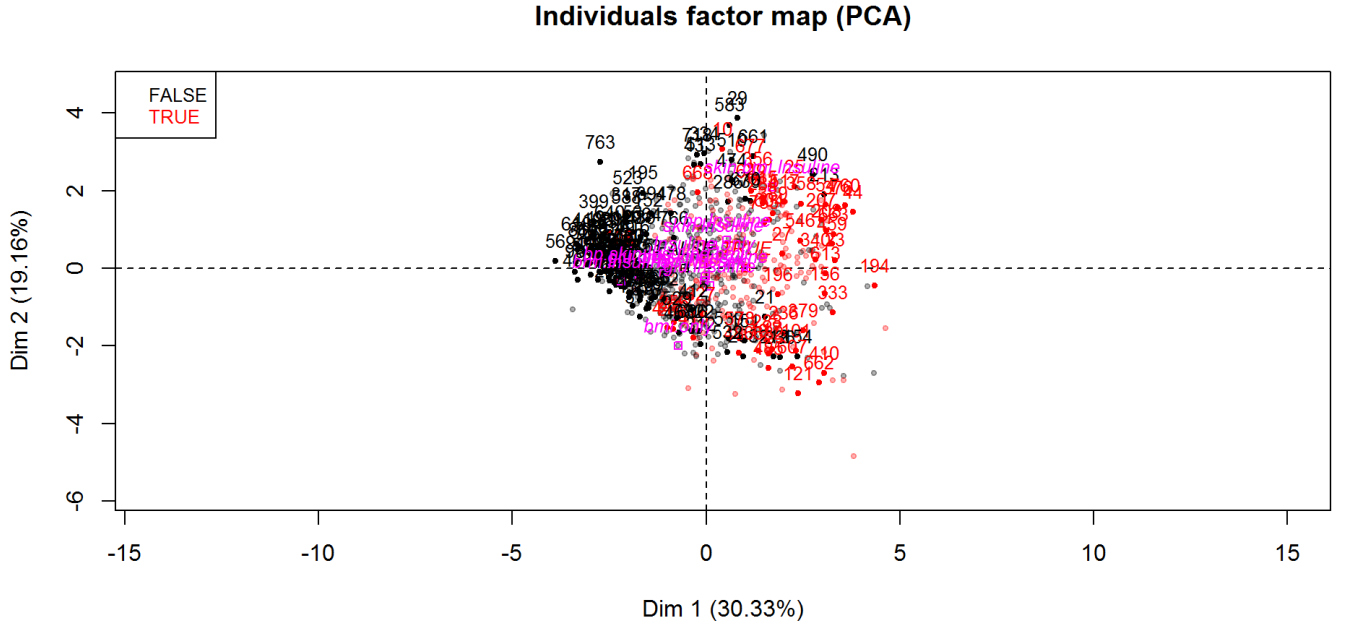


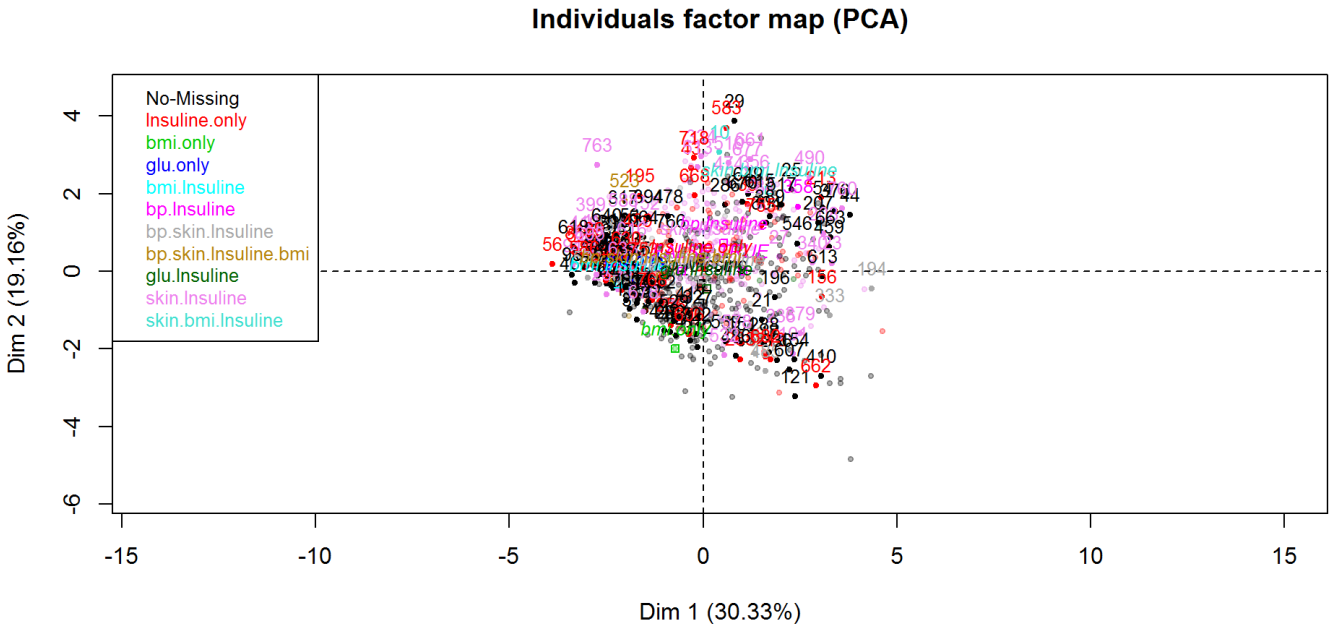


Le nuage des individus est bien centré

- Données Pima complétées avec la méthode mice du package mice

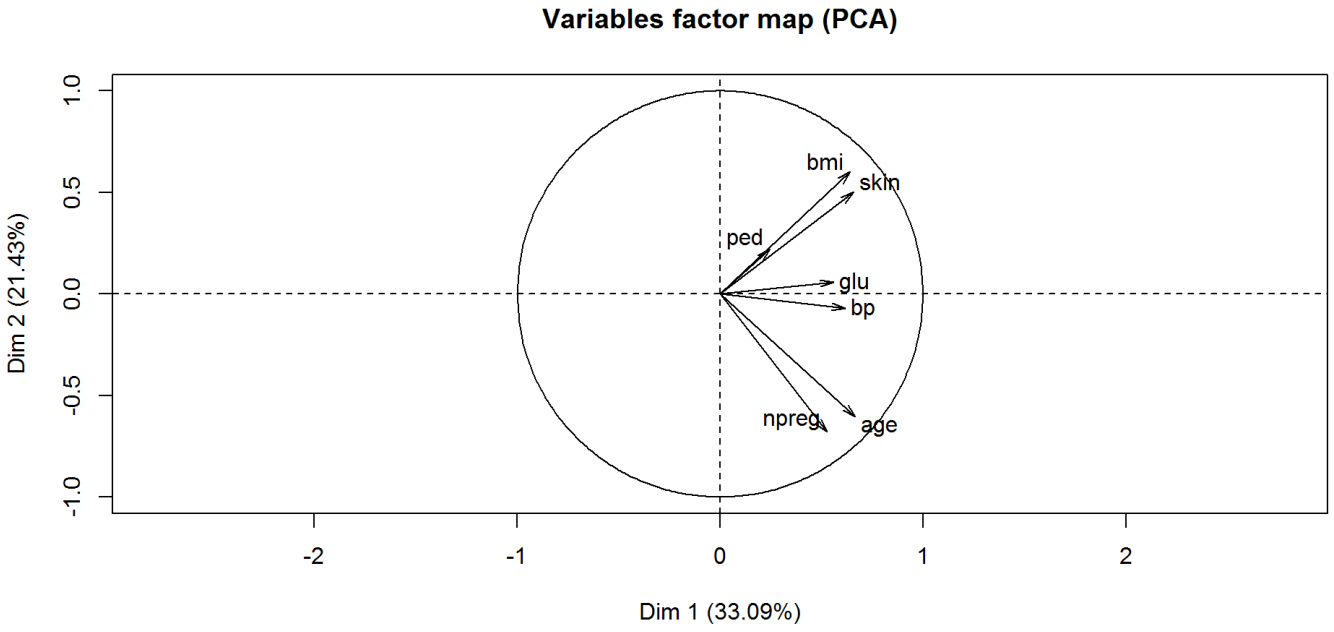


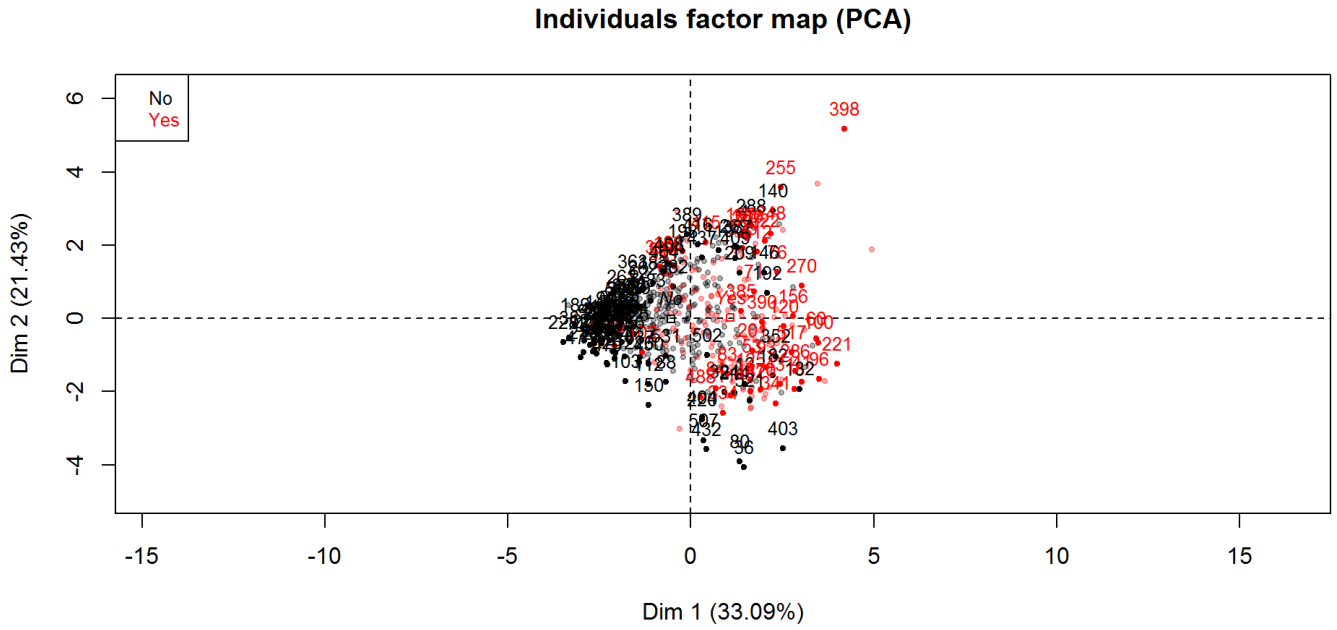




Dans le cas de mice toutes données sont relativement centrées.

- Données Pima de la librairie MASS

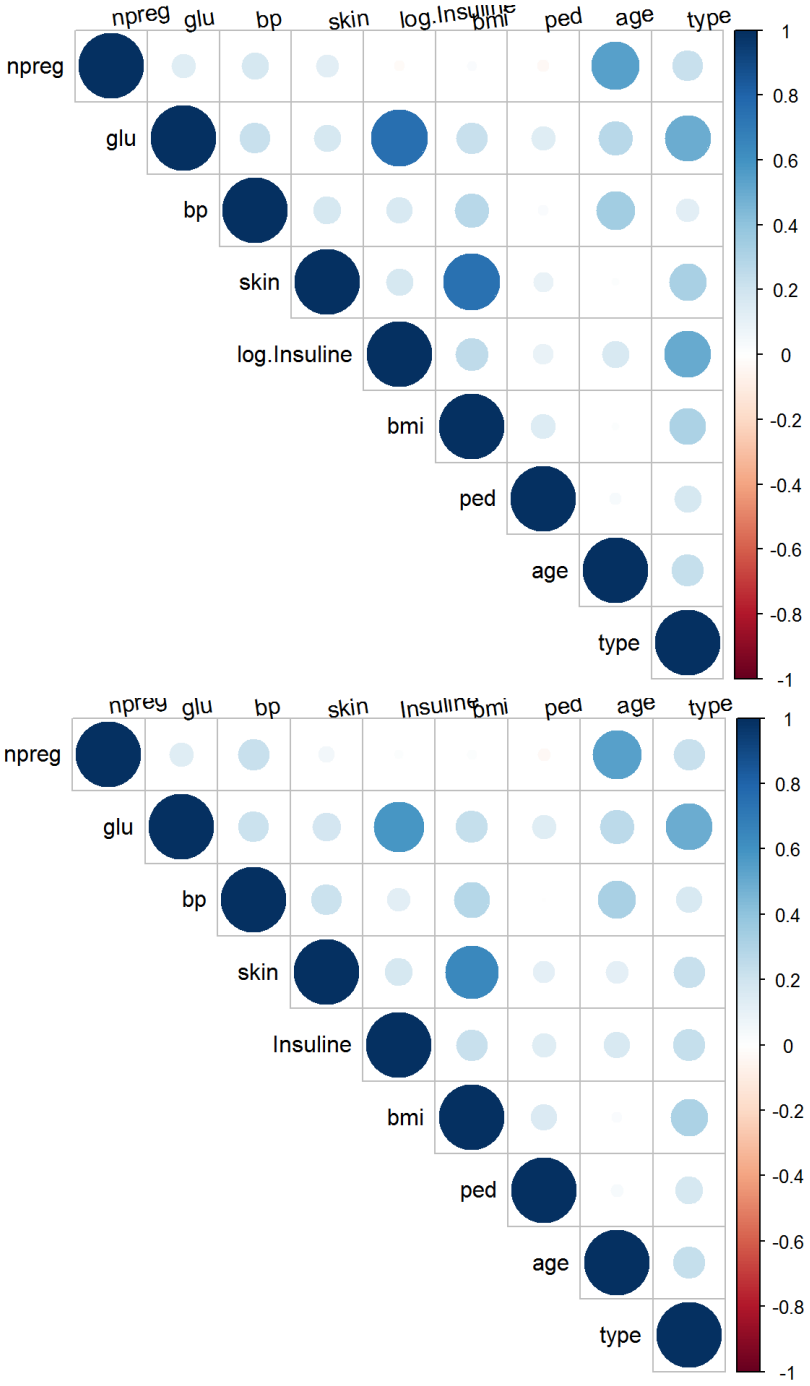


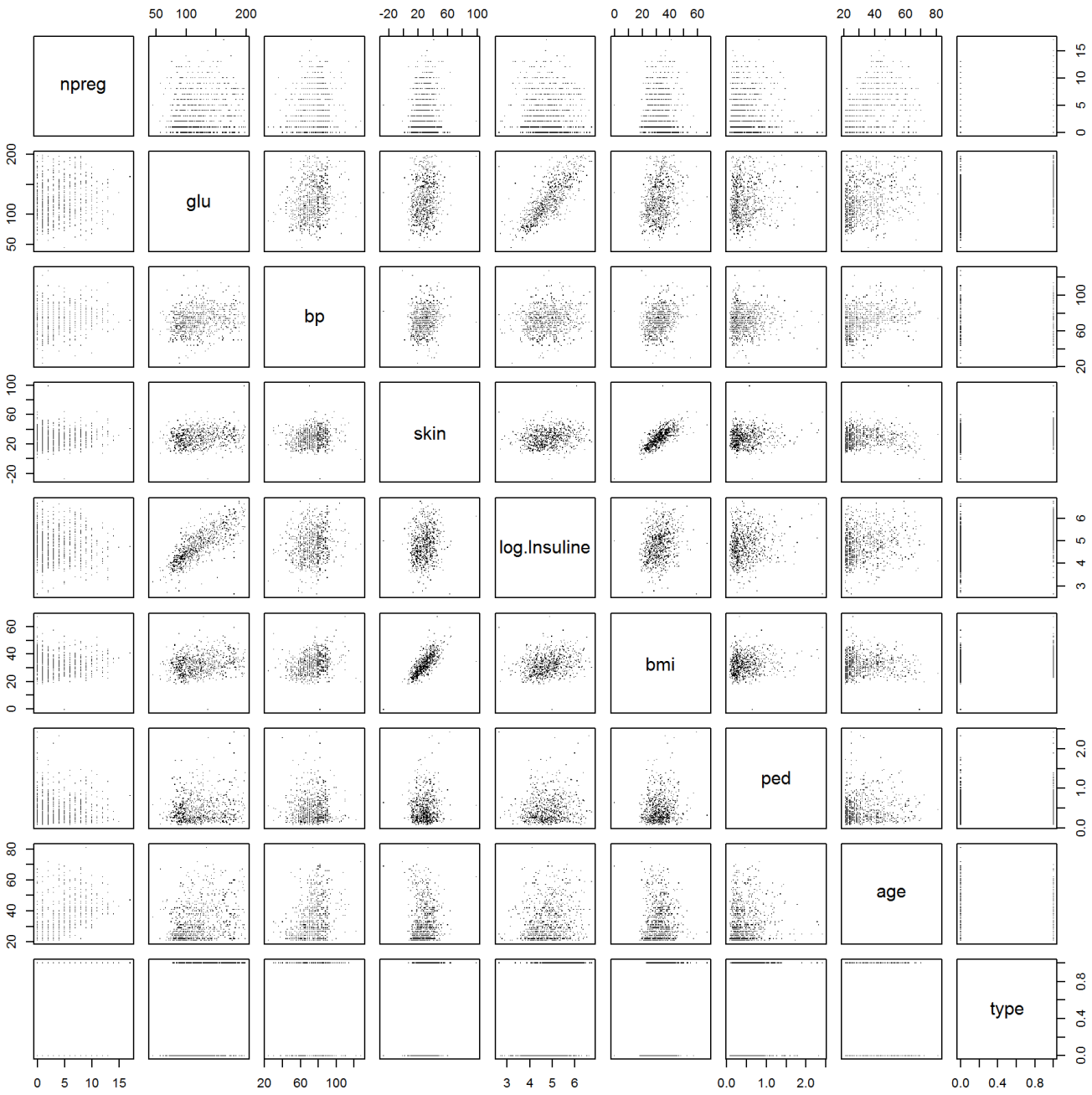


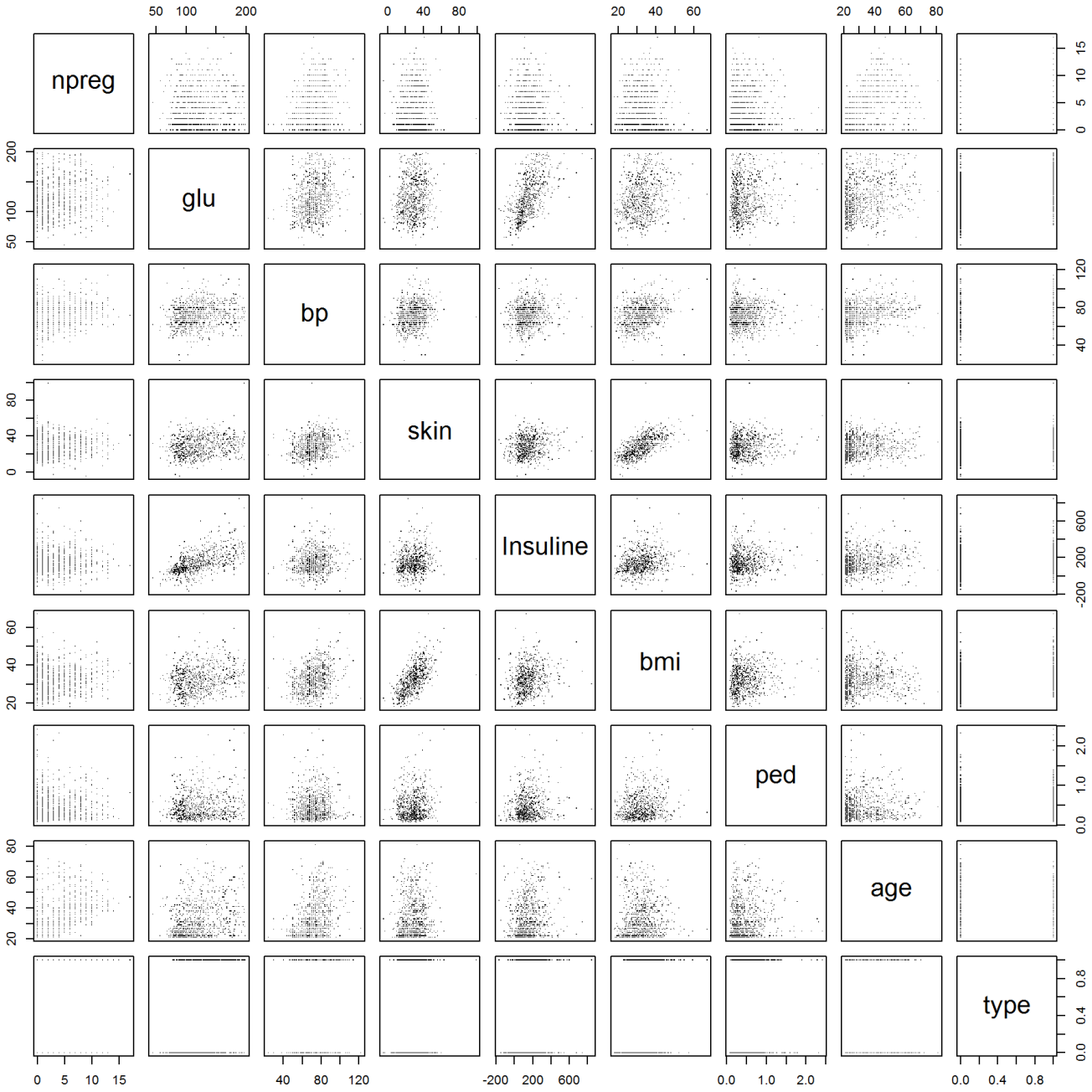
Remarque: Les données complétés par MICE et MIPCA ont une PCA similaire. Par contre pour les données de MASS on a une transformation par rapport aux axes. Même forme mais symétrique / à l'axe principal.

Annexe 4 : Comparaison des corrélations des données complétés obtenues avec mice et missMA

A la vu des graphiques, la structure globale des corrélations est très voisine.




npreg
glu
bp
skin
log.Insuline
bmi
ped
age
type


npreg
glu
bp
skin
Insuline
bmi
ped
age
type