METHODES DE REGRESSION SEMIPARAMETRIQUE DE TYPES « SLICING » OU « POOLED SLICING » : MISE EN ŒUVRE SOUS LE LOGICIEL SAS® SOUS FORME DE MACRO-COMMANDES ET APPLICATION SUR DES JEUX DE DONNEES

Ali Gannoun[1,2], Christiane Guinot[3] et Jérôme Saracco[1]

[1] *Laboratoire de Probabilités et Statistique*
Université Montpellier II
Place Eugène Bataillon, 34 095 Montpellier Cedex 5, France
(Email : {gannoun,saracco}@stat.math.univ-montp2.fr)

[2] *Statistical Genetics and Bioinformatics unit*
National Human Genome Center
Howard University
Washington D.C. 200059, U.S.A.
(Email : agannoun@howard.edu

[3] *CE.R.I.E.S.*
20, rue Victor Noir, 92 521 Neuilly sur Seine Cedex, France
(Email : christiane.guinot@ceries-lab.com)

1. Introduction

De nombreuses méthodes de régression et de prédiction sont disponibles quand la variable à expliquer $Y \in \mathbf{R}$ est réelle et lorsque la variable explicative X est un vecteur de $\mathbf{R}^p$. Ces méthodes reposent soit sur une approche paramétrique (comme la régression linéaire multiple ou polynomiale), soit sur des approches non-paramétriques (comme la régression à noyau ou les splines de lissage) ou semi-paramétriques (telles que celles que nous allons présenter ici). Toutes ces approches et méthodes sont caractérisées par des avantages et des défauts spécifiques.

Ainsi, par exemple, dans les modèles paramétriques, la structure de la dépendance entre Y et X est complètement fixée. Pour l'interprétation du modèle, ceci est un point très positif. Cependant, dans certains cas, le modèle paramétrique n'est pas en adéquation avec le vrai modèle sous-jacent ; la méthode estime alors la « meilleure » approximation dans l'espace de la classe choisie de fonctions paramétriques. Cette approximation peut être parfois très « éloignée » du vrai modèle, et les conclusions découlant de l'estimation peuvent alors être erronées.

Afin de surmonter ce problème et en suivant le principe « laissons parler les données »[1], des modèles non-paramétriques ont été développés. Ces modèles plus généraux ne reposent souvent que sur des hypothèses de continuité et de dérivabilité de la fonction de lien entre X et Y. Ils sont alors d'une grande flexibilité. Cependant leur estimation nécessite la spécification des paramètres de lissage (tels que la largeur de fenêtre pour l'estimateur à noyau). Ces paramètres doivent être déterminés à partir des données, le plus souvent au moyen de procédures de calculs intensifs (comme pour le critère de validation croisée par exemple), avant que les estimations via le modèle puissent être obtenues. Néanmoins un important problème pratique se pose dès que la dimension p de X est trop grande ($p \geq 3$) : ceci est connu sous le terme « fléau (ou malédiction) de la dimension ». Plus précisément, les données vont être de plus en plus dispersées au fur et à mesure que la dimension de l'espace de x augmente. De ce fait, dans le voisinage d'un point d'intérêt x_0, pour lequel on veut estimer $f(x_0)$, il va être difficile d'avoir suffisamment de données afin d'obtenir une estimation de bonne qualité.

Pour contourner cette difficulté, une partie paramétrique peut alors être incluse dans le modèle non-paramétrique : ce dernier devient ainsi semi-paramétrique. Cela ajoute non seulement une certaine structure au modèle, mais cela permet aussi une interprétation plus aisée du rôle joué par les variables explicatives. A ce propos, Li (1991) a introduit un modèle semi-paramétrique de régression fondé sur la réduction de dimension de la variable explicative, sans perte d'information, et préservant la flexibilité du modèle non-paramétrique. La méthode d'estimation de la partie paramétrique du modèle qu'il propose, s'appelle la méthode SIR, pour « *Sliced Inverse Regression* » (que l'on peut traduire par « Régression inverse par tranches (ou par tranchage) »). Cette méthode sera notée SIR-I dans la suite.

La deuxième partie de cet article est consacrée à une présentation du modèle semi-paramétrique considéré, ainsi que des méthodes d'estimation correspondantes. Nous présentons tout d'abord, à la section 2.1, le modèle semi-paramétrique de régression lorsque la variable à expliquer Y est unidimensionnelle. Ensuite, à la section 2.2, nous donnons un rapide panorama des méthodes d'estimation de sa partie paramétrique : méthodes de type « Slicing » (basées sur un tranchage) et méthodes de type « Pooled Slicing » (basées sur une combinaison de tranchages). A la section 2.3, nous exposons le cas d'un modèle où la variable à expliquer est elle-même multidimensionnelle, i.e. $Y \in \mathbf{R}^q$ ($q>1$), et nous décrivons brièvement deux méthodes d'estimation adaptées à ce cadre. La troisième partie concerne la mise en œuvre sous le logiciel SAS® des méthodes décrites précédemment. Les diverses macro-commandes SAS qui ont été implémentées[2] y sont commentées en détails. La quatrième et dernière partie illustre l'utilisation de ces macro-commandes : tout d'abord avec des jeux de données simulées à la section 4.1, puis avec l'étude d'un jeu de données réelles décrivant des propriétés biophysiques de la peau de femmes françaises à la section 4.2.

[1] Traduction libre de "letting the data speak for themselves".

[2] Toutes les macro-commandes peuvent être récupérées auprès de Jérôme Saracco.

2. Présentation du modèle et des méthodes

2.1. *Présentation du modèle*

Nous considérons le modèle de régression semi-paramétrique suivant (introduit par Li (1991)) :

$$Y = f(X'\beta_1,...,X'\beta_K,\varepsilon) .$$

La variable à expliquer Y est reliée à la variable explicative p-dimensionnelle X par un vecteur explicatif de dimension réduite : $(X'\beta_1,...,X'\beta_K)$, $K<p$. Aucune hypothèse n'est faite ni sur le paramètre fonctionnel f, ni sur la distribution de l'erreur ε.

Nous nous intéressons à la partie paramétrique du modèle, c'est-à-dire que l'on désire connaître l'espace engendré par les K vecteurs de paramètres inconnus $\beta_1,...,\beta_K$. Cet espace est appelé l'espace EDR (comme « *Effective Dimension Reduction*). Remarquons que les vecteurs β_k ne sont pas identifiables individuellement (car le paramètre fonctionnel f est inconnu), ainsi seul l'espace EDR est globalement identifiable. On appelle direction EDR toute direction appartenant à l'espace EDR. Les méthodes SIR de type « Slicing » ou « Pooled Slicing » permettent d'estimer une base de cet espace EDR sans avoir à estimer f. Une fois qu'on a réduit la dimension de p à K en estimant cet espace EDR, il ne reste qu'à utiliser une méthode non-paramétrique afin d'estimer le paramètre fonctionnel f. Cette méthode peut être du type noyau ou splines de lissage. Cette étape d'estimation non-paramétrique sera d'autant plus aisée que le réduction de dimension est efficace, i.e. que K est petit.

Remarque : Dans certaines problématiques, il arrive qu'on n'ait pas à estimer la fonction de lien. C'est par exemple le cas de l'étude des courbes de références (très utilisées dans le milieu biomédical) qui sont fondées sur l'estimation des quantiles conditionnels. Plus précisément, les « courbes » de références contenant 100(2α-1)% des sujets de référence sont les ensembles de points $\{(x,q_{1-\alpha}(x))\}$ et $\{(x,q_\alpha(x))\}$ lorsque x varie, où $q_\alpha(x)$ est le quantile conditionnel d'ordre $\alpha \in]0,5;1[$ de la variable y sachant que $X=x$. Ces quantiles conditionnels peuvent être estimés non paramétriquement sans avoir à estimer au préalable la fonction de lien (voir par exemple Gannoun *et al* (2002)). Dans le cadre du modèle de réduction de dimension que l'on considère dans cet article, on peut montrer que $q_\alpha(x)=q_\alpha(x'\beta_1,...x'\beta_K)$, voir Gannoun *et al* (2001).

2.2. Méthodes de type « Slicing » ou « Pooled Slicing »

Les différentes méthodes qui vont être décrites succinctement dans la suite, reposent sur deux aspects spécifiques qui ont donnés le nom à ces méthodes : « régression inverse par tranches (ou par tranchage) ».

- Le premier aspect concerne l'utilisation de propriétés des moments conditionnels de X sachant Y (d'où le nom de « régression inverse » puisque généralement lorsque l'on cherche à modéliser la relation de dépendance de Y sur X, on utilise les moments conditionnels de Y sachant X).
- Le second aspect intervient au moment de l'estimation de ces moments conditionnels inverses : la technique repose sur un ou plusieurs découpages en « tranches » du support de y. Cette idée permet de construire des estimateurs très faciles à évaluer d'un point de vue numérique.

Nous détaillons maintenant ces deux aspects en présentant les philosophies sous-jacentes aux différentes méthodes de type « slicing » et leurs homologues de type « pooled slicing ».

● *Moments conditionnels de x sachant y et espace EDR*

L'idée la plus importante de ces méthodes est de considérer la régression « inverse » (c'est à dire d'inverser les rôles de Y et de X) et de trouver une connexion entre l'espace EDR et les « courbes » des deux premiers moments inverses $E[X|Y]$ et $V(X|Y)$. Ainsi, l'approche SIR-I repose exclusivement sur une propriété de la courbe de régression inverse $E[X|Y]$. L'utilisation de cette courbe représente un certain nombre d'avantages comme la simplicité de l'idée sous-jacente ainsi que celle de l'implémentation informatique. En effet, d'un point de vue purement calculatoire, le fléau de la dimension disparaît vu que p régressions unidimensionnelles sont maintenant nécessaires à l'estimation de $E[X|Y]$ au lieu d'une unique régression p-dimensionnelle pour estimer $E[Y|X]$. Cette approche SIR-I souffre cependant d'un problème pathologique connu : elle est en effet « aveugle » dans le cas où le modèle présente une « dépendance symétrique » (la définition de ce type de modèle ne sera pas précisée ici, un exemple sera donné ultérieurement pour illustrer cette pathologie). Dans ce cas, la méthode SIR-I peut ne pas trouver toutes les directions EDR. Pour résoudre ce problème pathologique et pour retrouver toutes les directions EDR, une extension naturelle est de considérer des moments conditionnels de x sachant y d'ordre supérieur. Dans ce but, la méthode SIR-II utilise des propriétés de la « courbe » de variance conditionnelle $V(X|Y)$. Afin d'utiliser les informations provenant des deux premiers moments conditionnels, Li (1991) propose de combiner les deux approches SIR-I et SIR-II au travers la méthode SIRα, le paramètre α permet de pondérer l'influence de chacune des méthodes.

● *Tranchage (« slicing ») ou combinaison de tranchages (« pooled slicing »)*

Dans la procédure d'estimation de la méthode SIR-I, Li (1991) propose d'estimer la courbe de régression inverse par une simple fonction étagée basée sur un tranchage, c'est à dire une partition du support de Y en H tranches. Cette idée est aussi utilisée pour l'estimation de la « courbe » de variance conditionnelle inverse dans le cas de la méthode SIR-II. Les arguments pour le choix d'une étape de tranchage sont principalement la simplicité de l'écriture des estimateurs (voir la section 2.3) et la rapidité des temps de calcul qui va en découler. D'un point de vue théorique, les méthodes de type « slicing » ne sont pas sensibles au choix du tranchage. Cependant d'un point de vue pratique, l'utilisateur doit choisir un tranchage particulier, et l'estimation des directions EDR peut être sensible au nombre et à la position des tranches choisies lorsque la taille de l'échantillon est relativement petite (inférieure à 50 individus). Pour venir à bout de ce défaut, Aragon et Saracco (1997) introduisent l'idée du « pooled slicing » qui consiste à combiner des résultats provenant de différents tranchages du support de y afin d'absorber les variations dues au choix d'un tranchage particulier. Sur la base de cet argument, toutes les méthodes de type « slicing » ont alors une version de type « pooled slicing » (voir Saracco (2001)). Dans la suite, ces méthodes seront appelées PSIR-I, PSIR-II et PSIRα.

Pour plus de détails théoriques et pratiques concernant ces deux aspects des différentes méthodes, le lecteur pourra consulter l'Annexe 2.

2.3. *Variable à expliquer multidimensionnelle*

On considère maintenant un modèle semi-paramétrique dans lequel la variable à expliquer Y est de dimension q :

$$Y_j = f_j(X'\beta_1,...,X'\beta_K,\varepsilon) \text{ pour } j=1,...,q.$$

Un des objectifs principaux dans ce modèle est aussi l'estimation de l'espace EDR par des méthodes de type « Slicing »[3]. On peut ensuite être intéressé par l'estimation non paramétrique des fonctions de lien f_j. Les méthodes de type « Slicing » que nous allons présenter ici et qui ont été implémentées sous le logiciel SAS®, reposent sur deux philosophies légèrement différentes.

- La première méthode, celle du « Complete Slicing », est fondée sur une utilisation directe de la propriété de la courbe de régression inverse, propriété déjà utilisée dans le cas d'un y unidimensionnel. Ici, le tranchage correspond à un découpage de l'espace des y_i en un nombre H de pavés de $\mathbf{R}^q$ (qui sont ici les « tranches »[4]). Il ne reste alors qu'à appliquer la méthode SIR-I comme dans le cas univarié.
- La seconde méthode, appelée « Pooled Marginal Slicing », est une méthode qui agrège diverses informations provenant des q composantes de Y : les informations concernent des propriétés des l'espérance conditionnelle de X sachant Y_j, la phase d'estimation nécessite le tranchage (marginal) de chaque composante de Y. Quelques précisions sur cette méthode sont données en Annexe 2.

3. Mise en œuvre sous le logiciel SAS®

Les différentes méthodes présentées dans la section précédente ont été implémentées dans le logiciel SAS® sous forme de macro-commandes SAS. Les principales étapes de calculs matriciels ont été réalisées au moyen de la procédure IML (du module SAS IML®), les divers graphiques disponibles appellent des procédures graphiques usuelles telles que GPLOT, G3D ou G3GRID (du module SAS GRAPH®).

Dans la suite de cette partie, nous décrivons tout d'abord les différentes macro-commandes SAS développées pour les méthodes de type « Slicing » ou « Pooled Slicing » dans le cadre où Y est unidimensionnel, puis lorsque Y est multidimensionnel.

3.1. Macro-commandes SAS pour les méthodes SIR et PSIR univariées

● Macro-commande `%PSIRuniv`

Elle permet de faire les calculs des différents éléments nécessaires à l'estimation de l'espace EDR (en particulier le calcul des directions EDR et des valeurs propres associées) par une des six méthodes de type « Slicing » ou de type « Pooled Slicing ». La liste des

[3] Pour un exposé introductif sur ces méthodes SIR multivariées, on peut consulter Aragon (1997).

[4] Le découpage en tranches est fait de manière récursive : on commence par découper selon la première composante de y, puis selon la seconde en tenant compte du découpage de la composante précédente, et ainsi de suite pour chacune des composantes. Chaque découpage est fait de manière à ce que chaque tranche contienne à peu près le même nombre d'individus. Attention, le nombre de tranches augmente très rapidement avec la dimension de y.

valeurs propres calculées est affichée dans la fenêtre Output. De plus, des tables SAS ont été créées et seront utilisées dans les autres macro-commandes. La syntaxe de cette macro-commande est la suivante :

```
%PSIRuniv(table,methode=1,alpha=0.5,NbMinTr=10,NbMinInd=10) ;
```

où les différents paramètres sont :

- `table` = nom de la table SAS contenant les données,
- `methode` = numéro de la méthode à utiliser pour l'estimation :

  1 pour la méthode SIR-I, 4 pour la méthode PSIR-I,

  2 pour la méthode SIR-II, 5 pour la méthode PSIR-II,

  3 pour la méthode SIR_{α},, 6 pour la méthode $PSIR_{\alpha}$.

Par défaut, la méthode SIR-I est utilisée.

- `alpha` = valeur du paramètre α à choisir entre 0 et 1 (utile uniquement pour les méthodes SIR_{α} et $PSIR_{\alpha}$). Pour α=0 (resp. 1), on retrouve les méthodes SIR-I ou PSIR-I (resp. SIR-II ou PSIR-II).

Par défaut, le paramètre α prend la valeur 0.5 pour les méthodes SIR_{α} et $PSIR_{\alpha}$.

- `NbMinTr` = nombre minimum de tranches à considérer dans les différents tranchages pris en compte dans les méthodes de type « Pooled Slicing ».

Par défaut, le nombre minimum de tranches à considérer est fixé à 10 tranches.

- `NbMinInd` = nombre minimum d'individus par tranches à considérer dans les différents tranchages pris en compte dans les méthodes de type « Pooled Slicing ».

Par défaut, le nombre minimum d'individus par tranches à considérer est fixé à 10 individus.

● Macro-commande `%gPSIRe`

Elle permet d'afficher dans la fenêtre graphique l'éboulis des valeurs propres (« Screeplot of eigenvalue »), ce graphique est utile pour choisir graphiquement le nombre de directions EDR à retenir pour la suite de l'étude. Cette macro-commande ne nécessite aucun paramètre particulier car elle utilise des tables SAS créées par la macro-commande `%PSIRuniv` , sa syntaxe est donc tout simplement :

```
%gPSIRe ;
```

● Macro-commande `%dirEDR`

Elle permet l'affichage dans la fenêtre Output de la (ou des) direction(s) EDR retenue(s) à l'étape précédente. Sa syntaxe est :

```
%dirEDR(EDR=1) ;
```

où le paramètre `EDR` correspond au nombre de directions EDR retenues au vu des valeurs propres (par défaut, une seule direction EDR est affichée).

● Macro-commande `%gPSIRp1`

Elle permet d'afficher dans la fenêtre graphique le nuage de points $\{(y_i,\ x_i'\hat{\beta}_k),\ i=1,...,n\}$ où $\hat{\beta}_k$ est la $k^{ème}$ direction EDR estimée. Une estimation de la fonction de lien par la méthode des splines de lissage est automatiquement superposée à ce nuage de points. La syntaxe est :

```
%gPSIRp1(indice=1) ;
```

où le paramètre `indice` correspond au numéro k de l'indice $x'\hat{\beta}_k$ représenté en abscisse (par défaut, il s'agit de l'indice calculé avec la première direction EDR estimée).

● Macro-commande `%gPSIRp2`

Elle permet d'afficher dans la fenêtre graphique le nuage de points en trois dimensions $\{(y_i,\ x_i'\hat{\beta}_k,\ x_i'\hat{\beta}_j),\ i=1,...,n\}$ où $\hat{\beta}_k$ (resp. $\hat{\beta}_j$) est la $k^{ème}$ (resp. $j^{ème}$) direction EDR estimée. La syntaxe est :

```
%gPSIRp2(indice1=1,indice2=2) ;
```

où les paramètres `indice1` et. `indice2` correspondent respectivement aux numéros k et j des indices. $x'\hat{\beta}_k$ et $x'\hat{\beta}_j$ représentés sur les deux axes du plan horizontal du graphique 3D (par défaut, il s'agit des indices calculés avec les deux premières directions EDR estimées).

● Macro-commande `%gPSIRp2s`

Elle permet d'obtenir dans la fenêtre graphique un graphique en trois dimensions représentant la surface de réponse de y en fonction de deux indices $x'\hat{\beta}_k$ et $x'\hat{\beta}_j$, cette surface de réponse est calculée par interpolation linéaire sur une grille. La syntaxe est :

```
%gPSIRp2s(indice1=1,indice2=2, rotate=70, tilt=70) ;
```

où les paramètres `indice1` et `indice2` sont définis de manière analogue à la macro-commande précédente. Les deux autres paramètres `rotate` et `tilt` permettent de faire tourner les axes afin d'obtenir une meilleure représentation de la surface de réponse : le paramètre `rotate` spécifie un ou plusieurs angles selon lesquels le plan horizontal va pivoter autour de l'axe vertical ; le paramètre `tilt` spécifie un ou plusieurs angles selon lesquels on désire incliner le graphique vers nous (par défaut, ces deux paramètres prennent la valeur 70 degrés).

3.2. *Macro-commandes SAS pour les méthodes SIR multivariées*

● Macro-commande `%SIRmulti`

Elle permet de faire les calculs des différents éléments nécessaires à l'estimation de l'espace EDR (en particulier le calcul des directions EDR et des valeurs propres associées) par une des deux méthodes multivariées disponibles. La liste des valeurs propres calculées est affichée dans la fenêtre Output. Des tables SAS ont aussi été créées et seront utilisées dans les autres macro-commandes.

La syntaxe de cette macro-commande est la suivante :

```
%SIRmulti(table,NbY,methode=1) ;
```

où les différents paramètres sont :

- `table` = nom de la table SAS contenant les données (les premières variables de cette table SAS doivent être les variables à expliquer, les suivantes sont les variables explicatives).
- `NbY =` nombre de variables à expliquer. Ces variables à expliquer doivent être les premières variables de la table SAS utilisée.
- `methode =` numéro de la méthode à utiliser pour l'estimation :

  1 pour la méthode du Complete Slicing,

  2 pour la méthode du Pooled Marginal Slicing ;

Par défaut, la méthode du Complete Slicing est utilisée.

● Macro-commandes `%gSIRe` et `%dirEDR`

Elles fonctionnent exactement de la même manière que dans le cadre univarié (elles sont décrites à la section précédente).

● Macro-commande `%gMSIRp1`

Elle permet d'afficher dans la fenêtre graphique les nuages de points $\{(y_{j,i}\ ,\ x_i'\hat{\beta}_k),\ i=1,...,n\}$ où y_j est la $j^{\text{ème}}$ variable à expliquer et où $\hat{\beta}_k$ est la $k^{\text{ème}}$ direction EDR estimée. Une estimation de la fonction de lien par la méthode des splines de lissage est automatiquement superposée à ces nuages de points. La syntaxe est :

```
%gMSIRp1(indice=1,NbY=3) ;
```

Il est nécessaire de préciser en paramètre le nombre `NbY` de variables à expliquer.

● Macro-commande `%gMSIRp2`

Elle permet d'afficher dans la fenêtre graphique le nuage de points en trois dimensions $\{(y_{j,i}\ ,\ x_i'\hat{\beta}_k\ ,\ x_i'\hat{\beta}_l\ ),\ i=1,...,n\}$ où y_j est la $j^{\text{ème}}$ variable à expliquer et où $\hat{\beta}_k$ (resp. $\hat{\beta}_l$) est la $k^{\text{ème}}$ (resp. $l^{\text{ème}}$) direction EDR estimée. La syntaxe de cette fonction est :

```
%gMSIRp2(indice1=1,indice2=2,varY=1,NbY=3) ;
```

où les paramètres `indice1` et. `indice2` correspondent respectivement aux numéros k et j des indices. $x'\hat{\beta}_k$ et $x'\hat{\beta}_l$ représentés sur les deux axes du plan horizontal du graphique 3D (par défaut, il s'agit des indices calculés avec les deux premières directions EDR estimées). Deux autres paramètres doivent être renseignés :

- `varY =` numéro j de la variable à expliquer qui sera représentée sur l'axe vertical du graphique 3D. Par défaut, la première variable à expliquer sera prise en compte.
- `NbY =` nombre de variables à expliquer (par défaut 3).

● Macro-commande `%gMSIRp2s`

Elle permet d'obtenir dans la fenêtre graphique un graphique en trois dimensions représentant la surface de réponse de y_j en fonction de deux indices $x'\hat{\beta}_k$ et $x'\hat{\beta}_l$, cette

surface de réponse est calculée par interpolation linéaire sur une grille. La syntaxe de cette fonction est :

```
%gMSIRp2s(indice1=1,indice2=2,varY=1,NbY=3,rotate=35,tilt=70);
```

où les paramètres `indice1` et `indice2`, `rotate` et `tilt` sont définis de manière analogue au cas univarié. Comme pour la macro-commande précédente, les deux autres paramètres devant être renseignés sont `varY` et `NbY`.

4. Application diverses

Dans cette partie, nous illustrons et commentons l'utilisation des macro-commandes décrites précédemment sur des jeux de données simulées, puis sur des données réelles.

4.1. *Utilisation des macros-commandes avec des jeux de données simulées*

4.1.1. Présentation des jeux de données simulées

Pour illustrer l'utilisation de ces macros-commandes et pour décrire les sorties numériques ou graphiques qui leur sont associées, nous avons simulé quatre jeux de données.

- Jeux de données n°1. Nous considérons un modèle univarié à un seul indice sans « dépendance symétrique » :

$$Y=(X'\beta)^3-(X'\beta)^2+\varepsilon,$$

où X suit une loi multinormale centrée réduite de dimension 5, l'erreur ε suit une loi normale centrée réduite et est indépendante de X, et β=(1,1,-1,-1,0)'. Selon ce modèle, nous avons simulé un échantillon de taille n=200 que nous avons « stocké » dans la table SAS **m1**, cette table SAS contient 6 variables notées Y, X1, …,X5. Avec ce jeu de données, les méthodes SIR-I ou PSIR-I doivent fournir une bonne estimation de la direction EDR.

- Jeux de données n°2. Nous considérons un modèle univarié à un seul indice avec « dépendance symétrique » :

$$Y=(X'\beta)^2+\varepsilon,$$

où X, ε, β sont définies comme précédemment. La table SAS **m2** contient l'échantillon correspondant de taille n=200. Ici, on est dans le cas pathologique des méthodes SIR-I et PSIR-I. Par contre, les méthodes SIR-II et PSIR-II doivent fournir une bonne estimation de la direction EDR. Les méthodes SIRα et PSIRα doivent elles aussi fonctionner correctement si le paramètre α donne suffisamment de poids à la méthode SIR-II ou PSIR-II.

- Jeux de données n°3. Nous considérons un modèle univarié à deux indices :

$$Y=(X'\beta_1)^2+(X'\beta_2)^2+\varepsilon,$$

où X, ε sont définies comme précédemment, β_1=(1,1,-1,0,0)' et β_2=(0,0,0,1,-2)'. La table SAS **m3** contient l'échantillon correspondant de taille n=1000. Ce modèle présente une « dépendance symétrique », ainsi la méthode SIR-II doit permettre d'obtenir une bonne estimation de l'espace EDR.

- Jeux de données n°4. Nous considérons un modèle mutivarié à un seul indice :

$$Y=\begin{cases} Y_1=X'\beta+\varepsilon_1 \\ Y_2=\exp(X'\beta)+\varepsilon_2 \\ Y_3=(X'\beta)^2+(X'\beta)\varepsilon_3 \end{cases}$$

où X, β sont définies comme précédemment pour les jeux de données 1 et 2. Les erreurs ε_1, ε_2 et ε_3 suivent des lois normales centrées réduites et sont indépendantes entre elles et de X. La table SAS **m4** contient l'échantillon correspondant de taille n=200, les variables sont Y1, Y2, Y3, X1, …, X5. Les méthodes du « Complete Slicing » et du « Pooled Marginal Slicing » doivent permettre d'estimer correctement la direction EDR.

4.1.2. Programmes SAS et commentaires des sorties numériques et graphiques

Pour chacun de ces quatre jeux de données simulées, nous donnons ci-après le programme SAS, des commentaires de quelques sorties numériques et graphiques obtenues avec les macro-commandes. Les extraits des sorties numériques sont disponibles en Annexe 3. Tous les graphiques ont été regroupés en Annexe 1.

● Etude de la table SAS **m1**

Programme SAS

```
%PSIRuniv(m1,methode=1);    /* méthode SIR-I */
%gPSIRe;
%dirEDR(EDR=1);
%gPSIRp1;

%PSIRuniv(m1,methode=4,alpha=0.5,NbMinTr=10,NbMinInd=10); /* méth PSIR-I */
%gPSIRe;
%dirEDR(EDR=1);
%gPSIRp1;
```

Commentaires des sorties numériques et graphiques

Pour la méthode SIR-I, le début des sorties numériques nous renseigne sur le nombre d'individus, de variables explicatives et de tranches utilisées. La répartition des individus par tranches est ensuite fournie. Au de la liste des valeurs propres ou bien de la **figure 1(a1)** représentant l'éboulis des valeurs propres, il apparaît clairement qu'il faut retenir une seule direction EDR. La direction EDR estimée est ensuite affichée à l'écran. (Notons que l'estimation est d'excellente qualité, le cosinus carré de l'angle entre la vraie direction et cette direction estimée est égal à 0,999.) **La figure 1(a2)** représente le nuage de points croisant la variable à expliquer et l'indice estimé, auquel est superposée une estimation de la fonction de lien.

Pour la méthode PSIR-I, les premières lignes de sorties nous informent sur les différents paramètres du « Pooling » (combinaison de tranches) : nombre minimum de tranches, nombre minimum d'individu par tranche et nombre de tranchages considérés. Ici, encore la liste des valeurs propres fait clairement ressortir qu'une seule direction EDR doit être retenue. La direction EDR estimée par la méthode PSIR-I est très similaire à celle obtenue par SIR-I (le cosinus carré de l'angle entre cette direction et le vecteur β est aussi égal à 0,999). Ainsi, nous

n'avons pas jugé nécessaire de fournir les deux graphiques produits par les macro-commandes SAS correspondantes.

● Etude de la table SAS **m2**

Programmes SAS

```
%PSIRuniv(m2,methode=1);     /* méthode SIR-I */
%gPSIRe;
%dirEDR(EDR=1);
%gPSIRp1;

%PSIRuniv(m2,methode=2);      /* méthode SIR-II */
%gPSIRe;
%dirEDR(EDR=1);
%gPSIRp1;

%PSIRuniv(m2,methode=6,alpha=0.5,NbMinTr=10,NbMinInd=10); /* PSIRalpha */
%gPSIRe;
%dirEDR(EDR=1);
%gPSIRp1;
```

Commentaires des sorties numériques et graphiques

Le modèle présentant une dépendance symétrique, on s'attend à ce que la méthode SIR-I ne fournisse pas une bonne estimation. Au vu des sorties numérique et graphique (voir la **figure 1(b1)**) portant sur les valeurs propres, on observe une décroissance sans rupture des valeurs propres, ce qui signifie que la méthode n'arrive pas à trouver des directions de réduction de la dimension. Cela est confirmé par le nuage de points de la **figure 1(b2)** qui ne présente aucune structure modélisable. (Notons que la qualité d'estimation mesurée par le cosinus carré est ici de 0,171.)

Par contre, les méthodes SIR-II et PSIRα (avec α=0,5) vont elles permettre d'estimer convenablement la direction de β.

Les listes des valeurs propres mettent en évidence une rupture entre la première et la deuxième valeur (ceci est bien visible sur l'éboulis des valeurs propres de la méthode SIR-II à **la figure (c2)**). Les directions EDR estimées sont d'excellente qualité (leur cosinus carré est respectivement égal à 0,974 pour SIR-II et 0,986 pour PSIRα). On peut noter que les deux vecteurs estimés sont de sens contraire. Le nuage de points croisant la variable à expliquer et l'indice estimé est fourni seulement pour la méthode SIR-II (voir **la figure 1(c2)**).

● Etude de la table SAS **m3**

Programme SAS

```
%PSIRuniv(m3,methode=2);     /* méthode SIR-II */
%gPSIRe;
```

%dirEDR(EDR=2);

%gPSIRp2;

%gPSIRp2s;

%gPSIRp2s(rotate= 0 20 40 60 80 100 180);

%gPSIRp2s(tilt= 0 20 40 60 80 100 180);

Commentaires des sorties numériques et graphiques

Remarquons ici que c'est bien l'espace EDR (engendré par les deux directions β_1 et β_2) qui a été estimé et non pas ces deux directions qui ne sont pas individuellement identifiables. Le modèle présentant une dépendance symétrique, la méthode d'estimation mise en œuvre est la méthode SIR-II. Au vu des sorties numériques portant sur les valeurs propres et de l'éboulis des valeurs propres (voir la **figure 1(d1)**), on va retenir deux directions EDR. Les directions EDR estimées fournissent une qualité d'estimation de l'espace EDR très bonne (cosinus carré égal à 0,994). La **figure 1(d2)** représente le nuage de points en 2D croisant le premier indice estimé et la variable à expliquer (le graphique utilisant le second indice estimé étant très similaire au graphique précédent, nous n'avons pas jugé nécessaire de l'inclure dans ces sorties). Les **figures 1(d3)** et **1(d4)** donnent les deux types de graphiques en 3D disponibles : le nuage de points et la surface de réponse estimée. Notons aussi que les graphiques qui sont produits par les deux dernières lignes du programme SAS n'ont pas non plus été fournis, ces deux lignes permettent de voir comment utiliser les paramètres rotate et tilt permettant de faire tourner les axes afin d'obtenir (éventuellement) une meilleure vision de la surface de réponse.

● Etude de la table SAS **m4**

Programme SAS

%SIRmulti(m4,methode=1,NbY=3); /* méthode du Complete Slicing */

%gSIRe;

%dirEDR(EDR=1);

%gMSIRp1;

%SIRmulti(m4,methode=2,NbY=3); /* méthode du Pooled Marginal Slicing */

%gSIRe;

%dirEDR(EDR=1);

%gMSIRp1;

Commentaires des sorties numériques et graphiques

Pour les deux méthodes, la liste des valeurs propres nous indique qu'une seule direction est nécessaire pour réduire la dimension. Les directions EDR estimées sont d'excellente qualité (avec des cosinus carrés respectivement égaux à 0,997 et 0,994). Les sorties graphiques ne concernent que la méthode du Complete Slicing : la **figure 1(e1)** donne l'éboulis des valeurs propres et les **figures 1(e2)**, **1(e3)** et **1(e4)** les nuages de points croisant le premier indice estimé et les trois variables à expliquer Y1, Y2 et Y3, les fonctions de lien estimées étant superposées à chacun des nuages de points.

4.2. *Utilisation des macro-commandes dans un problème de statistique appliquée*

4.2.1. Présentation des données

Les données étudiées portent sur des propriétés biophysiques de la peau de femmes françaises. L'étude dont sont issues les données, a été conduite à Paris du mois de novembre 1998 au mois de mars 1999 par le CE.R.I.E.S, centre de recherche sur la peau humaine financé par Chanel. Elle concerne des femmes françaises, âgées de 20 à 80 ans, présentant une peau apparemment saine (c'est-à-dire sans aucun signe de dermatose en cours ou de maladie générale avec manifestations cutanées avérées) et ayant respecté des consignes cosmétiques strictes. Cette étude s'est déroulée en atmosphère contrôlée (température de 23°C et humidité relative de 50±5%). Elle comportait des questionnaires sur les habitudes de vie, un interrogatoire et un examen médical cutané, en plus d'une évaluation des propriétés biophysiques cutanées. L'évaluation des paramètres biophysiques a été effectuée sur deux zones du visage (le front et la joue) et sur la face antérieure de l'avant-bras gauche. Les propriétés biophysiques de la peau incluaient pour chaque zone :

- la température cutanée mesurée en °C (variables TFRONT, TJOUE et TBRAS) ;
- la perte insensible en eau mesurée en g/m^2 (variables FRONT1, JOUE1 et BRAS1) ;
- le pH cutané (variables PFRONT, PJOUE et PBRAS) ;
- l'hydratation de la peau estimée par la capacitance (variables C2FRONT, C2JOUE et C2BRAS) et par la conductance de la peau mesurée en µS (variables KFRONT, KJOUE et KBRAS) ;
- le taux de sécrétion de sébum (taux instantané de lipides, mesuré en $\mu g/cm^2$) : variables SFRONT et SJOUE (ce taux a été mesuré uniquement sur les deux zones du visage) ;
- et la couleur de la peau. La couleur a été exprimée à l'aide des trois paramètres du système L*a*b* CIE 1976, où L* exprime la luminosité, a* les coordonnées de chromacité rouge/vert et b* les coordonnées de chromacité jaune/bleu. Les variables d'intensité rouge ou jaune et de luminosité correspondantes sont : AFRONT (axe A), BFRONT (axe B), LFRONT (axe L) pour la zone « front », AJOUE, BJOUE et LJOUE pour la zone « joue », et ABRAS, BBRAS et LBRAS pour la zone « bras ». Deux autres paramètres biophysiques de couleur sont aussi utilisés : la saturation de la peau, mesurée en C (variables C_FRONT, C_JOUE et C_BRAS); l'angle de teinte, exprimé en degrés (variables H_FRONT, H_JOUE et H_BRAS).

Les autres variables disponibles sont l'âge des volontaires (variable AGE en années), la température (variable TEMP en °C) et l'hygrométrie (variable HYGRO en %) de la pièce.

Dans cette étude, les variables d'intérêt sont les mesures de conductance (KFRONT, KJOUE et KBRAS). Le travail sera fait zone par zone, de manière indépendante. Les covariables pour chaque zone sont les paramètres biophysiques correspondants ainsi que l'âge, la température et l'hygrométrie de la salle.

Un des objectifs de cette étude est de rechercher, pour chaque zone, une ou des directions de réduction de dimension pour les variables d'intérêt afin de construire des courbes de référence à 90% pour ces paramètres biophysiques d'intérêt.

La démarche générale effectuée pour chaque zone est la suivante :

Etape 1 : On applique la méthode SIR-I en utilisant la variable d'intérêt et l'ensemble des covariables disponibles pour la zone considérée.

A partir du graphique de l'éboulis des valeurs propres, on détermine le nombre de directions EDR à conserver : on cherche un saut visible dans cet éboulis et K correspond au nombre de valeurs propres avant le saut. Remarquons que si aucun saut n'est détecté, aucune réduction de dimension n'est possible.

On visualise alors la structure du nuage de points formé par la variable d'intérêt et le ou les indices calculés à partir de la ou des K directions EDR estimées. On vérifie graphiquement qu'il n'y a pas de structure pour le nuage de points formé par la variable d'intérêt et les indices calculés à partir du $(K+1)^{\text{ème}}$ vecteur propre estimé (qui n'est pas une direction EDR).

Etape 2 : L'objectif est ici de simplifier le ou les indices afin de faciliter leur interprétation. A cette fin, pour chaque indice estimé, on fait une sélection automatique ascendante de régresseurs dans le modèle de régression linéaire multiple de cet indice estimé sur l'ensemble des covariables disponibles (on utilise ici la procédure REG (module SAS STAT®) et l'option *selection=forward*). Le sous-ensemble final des covariables retenues est celui obtenu en faisant l'union des sous-ensembles de régresseurs sélectionnés pour chaque indice.

On applique ensuite à nouveau la méthode SIR-I avec cet ensemble restreint de covariables et on obtient alors la ou les directions EDR estimées correspondantes. Finalement, on vérifie graphiquement que chaque nuage de points, croisant les indices calculés avec la totalité des covariables et les indices calculés avec le nombre restreint de covariables, présente une structure linéaire. Cela permet de vérifier que l'on n'a pas trop perdu d'information en supprimant les covariables les moins significatives dans l'explication des indices.

Etape 3 : On est maintenant en mesure de construire les courbes de référence pour la variable d'intérêt en fonction des indices estimés. Pour simplifier, plaçons-nous dans le cas où une seule direction EDR $\hat{\beta}$ a été retenue. On peut alors construire les indices estimés correspondants : $\left\{U_i = X_i\hat{\beta}, i=1,...,n\right\}$. Les courbes de référence sont alors obtenues à partir des données $\left\{(Y_i,U_i), i=1,...,n\right\}$ en utilisant, par exemple, la méthode du noyau pour estimer la fonction de répartition conditionnelle $\hat{F}(y \mid u)$. L'estimation du quantile conditionnel d'ordre α découlent naturellement de la résolution en y de l'équation $\alpha=\hat{F}(y \mid u)$. Pour avoir de plus amples détails concernant cette démarche de construction de courbes de référence, le lecteur pourra se référer à Gannoun *et al* (2001).

Cette troisième étape d'estimation n'a pas été faite ici. Les sorties décrites dans la suite ne concernent donc que les étapes 1 et 2.

4.2.2. Présentation des résultats

Nous présentons maintenant quelques extraits des résultats obtenus avec les étapes 1 et 2 pour chacune des trois zones.

● *Etude de la zone de l'avant- bras*

Etape 1 : On applique la méthode SIR-I à l'ensemble des covariables de cette zone.

	P
Le nombre de variables explicatives est :	12

Les valeurs propres estimées des directions EDR sont :

VALPREDR
0.3689291
0.1284043
0.0815379
0.0644682
0.0329833
0.0158766
0.0110742
0.0044383
0.0018435
8.568E-19
-2.71E-18
-2.11E-17

Direction(s) EDR estimée(s) :

	EDR1
AGE	-0.12555
HYGRO	0.24186
TEMP	0.41638
LBRAS	0.61243
ABRAS	0.69057
BBRAS	-0.11828
C2BRAS	0.02637
PBRAS	-0.15059
BRAS1	0.04342
TBRAS	0.02763
C_BRAS	-0.60865
H_BRAS	0.04246

Au vu de l'éboulis des valeurs propres (voir **figure 2(a1)**), on décide de ne retenir qu'une seule direction EDR. Ceci est confirmé par les graphiques des **figures 3(a1)** et **3(a2)** : le premier montrant une certaine structure entre la variable d'intérêt et le premier indice, le second ne montrant aucune structure entre la variable d'intérêt et le second indice.

Etape 2 : La procédure de sélection automatique ascendante (forward selection procedure) dont nous n'avons pas fourni ici les sorties, nous indique qu'en ne retenant que les 5 variables LBRAS, C_BRAS, C2BRAS, HYGRO et AGE (citées par ordre d'entrée dans le modèle), le R^2 est de plus de 81%. On applique alors à nouveau la méthode SIR-I avec ces 5 variables et on obtient :

	P
Le nombre de variables explicatives est :	5

Les valeurs propres estimées des directions EDR sont :

VALPREDR
0.3594497
0.0831848
0.0498697
0.0132621
0.0085565

Direction(s) EDR estimée(s) :

	EDR1
AGE	-0.13647
HYGRO	0.38632
LBRAS	-0.11962
C2BRAS	0.02762
C_BRAS	0.24308

Les graphiques des **figures 2(a2)**, **4(a1)** et **4(a2)** confirment le fait que la simplification de l'unique indice n'a pas fait perdre beaucoup d'information pour notre modélisation de la variable KBRAS.

Pour la peau de l'avant-bras, le modèle final qui a été construit, relie la conductance (KBRAS) *positivement* à l'humidité relative de la salle (HYGRO), à la capacitance (C2BRAS) qui est une autre mesure d'hydratation de la peau et à la saturation de la peau (C_BRAS) qui est une mesure de couleur de la peau, et *négativement* à l'âge et à un autre paramètre de couleur de la peau (LBRAS).

● *Etude de la zone du front*

Etape 1 : On applique la méthode SIR-I à l'ensemble des covariables de cette zone.

	P
Le nombre de variables explicatives est :	13

Les valeurs propres estimées des directions EDR sont :

VALPREDR
0.4396504
0.1163487
0.073608
0.0618144
0.0542457
0.018912
0.0164791
0.0091381
0.0024344
7.426E-18
7.1E-18
-2.49E-18

Direction(s) EDR estimée(s) :

	EDR1
AGE	-0.06135
HYGRO	0.01395
TEMP	0.35819
LFRONT	0.57197
AFRONT	1.40338
BFRONT	-0.01485
C2FRONT	0.00076
SFRONT	0.00562
PFRONT	-0.23933
FRONT1	0.07123
TFRONT	0.57539
C_FRONT	-1.11650
H_FRONT	-0.15001

Au vu de l'éboulis des valeurs propres (voir **figure 2(b1)**), on décide de ne retenir qu'une seule direction EDR. Ceci est confirmé par les graphiques des **figures 3(b1)** et **3(b2)** : le premier montrant une certaine structure entre la variable d'intérêt et le premier indice, le second ne montrant aucune structure entre la variable d'intérêt et le second indice.

Etape 2 : La procédure de sélection automatique ascendante (forward selection procedure) dont nous n'avons pas fourni ici les sorties, nous indique qu'en ne retenant que les 6 variables AGE, TFRONT, LFRONT, C_FRONT, SFRONT et FRONT1 (citées par ordre d'entrée dans le modèle), le R^2 est de plus de 89%. On applique alors à nouveau la méthode SIR-I avec ces 6 variables et on obtient :

	P
Le nombre de variables explicatives est :	6

Les valeurs propres estimées des directions EDR sont :

VALPREDR
0.3934356
0.0981441
0.0396528
0.0262265
0.015092
0.0071785

Direction(s) EDR estimée(s) :

	EDR1
AGE	0.37168
LFRONT	-0.00025
SFRONT	0.00680
FRONT1	0.08150
TFRONT	0.61821
C_FRONT	0.32407

Les graphiques des **figures 2(b2)**, **4(b1)** et **4(b2)** confirment le fait que la simplification de l'unique indice n'a pas fait perdre beaucoup d'information pour notre modélisation de la variable KFRONT.

Pour la peau du front, le modèle final qui a été construit, relie la conductance (KFRONT) *positivement* à l'âge, au taux de sécrétion de sébum (SFRONT), à la perte insensible en eau (FRONT1), à la température cutanée (TFRONT) et à la saturation de la peau (C_FRONT) qui est une mesure de couleur de la peau, et *négativement* à un autre paramètre de couleur de la peau (LBRAS).

● *Etude de la zone de la joue*

Etape 1 : On applique la méthode SIR-I à l'ensemble des covariables de cette zone.

	P
Le nombre de variables explicatives est :	13
	N

Les valeurs propres estimées des directions EDR sont :

VALPREDR
0.4893763
0.169949
0.0670676
0.052193
0.029233
0.0206825
0.0147654
0.00794
0.0040203
1.506E-17
3.691E-18
-2.61E-18

Direction(s) EDR estimée(s) :

	EDR1
AGE	-0.06036
HYGRO	-0.15447
TEMP	0.57636
LJOUE	0.41563
AJOUE	0.15192
BJOUE	-0.02082
C2JOUE	0.00018
SJOUE	0.00491
PJOUE	-0.29857
JOUE1	0.07759
TJOUE	-0.01757
C_JOUE	0.02494
H_JOUE	-0.00240

Au vu de l'éboulis des valeurs propres (voir **figure 2(c1)**), on décide de ne retenir qu'une seule direction EDR. Ceci est confirmé par les graphiques des figures **3(c1)** et **3(c2)** : le premier montrant une certaine structure entre la variable d'intérêt et le premier indice, le second ne montrant aucune structure entre la variable d'intérêt et le second indice.

Etape 2 : La procédure de sélection automatique ascendante (forward selection procedure) dont nous n'avons pas fourni ici les sorties, nous indique qu'en ne retenant que les 6 variables AGE, LJOUE, AJOUE, JOUE1, BJOUE et SJOUE (citées par ordre d'entrée dans le modèle), le R^2 est près de 95%. On applique alors à nouveau la méthode SIR-I avec ces 6 variables et on obtient :

	P
Le nombre de variables explicatives est :	6

Les valeurs propres estimées des directions EDR sont :

VALPREDR
0.4684183
0.1313528
0.041937
0.0143943
0.0081886
0.0038224

Direction(s) EDR estimée(s) :

	EDR1
AGE	0.56841
LJOUE	0.42401
AJOUE	0.16407
BJOUE	-0.00140
SJOUE	0.00537
JOUE1	0.08569

Les graphiques des **figures 2(c2)**, **4(c1)** et **4(c2)** confirment le fait que la simplification de l'unique indice n'a pas fait perdre beaucoup d'information pour notre modélisation de la variable KJOUE.

Pour la peau de la joue, le modèle final qui a été construit, relie la conductance (KJOUE) *positivement* à l'âge, à deux paramètres de couleur de la peau (LJOUE et AJOUE), au taux de sécrétion de sébum (SFRONT), à la perte insensible en eau (FRONT1), et *négativement* à un autre paramètre de couleur de la peau (BJOUE).

Commentaires sur ces trois modèles : Pour chacune des zones considérées, un certain nombre de variables (4 ou 5) entrent dans le modèle en plus de la covariable AGE qui était attendue. Contrairement à l'opinion des experts qui redoutaient que les conditions expérimentales (mesurées par les variables HYGRO et TEMP) aient une forte influence sur les paramètres biophysiques de la peau quelle que soit la zone, seule l'humidité relative de la salle (HYGRO) intervient pour la conductance réalisée au niveau de l'avant-bras. Les autres covariables retenues diffèrent selon les zones. La plupart sont reliées cliniquement à l'hydratation de la peau telles que la capacitance, la perte insensible en eau et le taux de sécrétion de sébum (ces deux dernières étant présentes dans les modèles sur les deux zones du visage). Enfin pour l'ensemble des zones, il est intéressant de noter qu'une ou plusieurs variables mesurant la couleur de la peau entrent dans les modèles.

Remerciements

Les auteurs remercient le Pr. E. Tschachler pour ses encouragements ainsi que toute l'équipe du CE.R.I.E.S. pour leur contribution aux données sur les propriétés biophysiques de la peau, en particulier J. Latreille, I. Le Fur et G. Heuvin.

RÉFÉRENCES

Aragon, Y. (1997) A Gauss implementation of Multivariate Sliced Inverse Regression. *Computational Statistics* **12** 355-372.

Aragon, Y., Saracco, J. (1997) Sliced Inverse Regression: an appraisal of small sample alternatives to slicing. *Computational Statistics* **12** 109-130.

Ferré, L. (1998). Determining the dimension in SIR and related methods. *Journal of the American Statistical Association* **441** 132-140.

Gannoun, A., Girard, S., Guinot, C., Saracco, J. (2001). Dimension reduction in reference curves estimations. *Rapport de Recherche, Unité de Biométrie, ENSAM-INRA-UMII, n°01-06*.

Gannoun, A., Girard, S., Guinot, C., Saracco, J. (2002). Trois méthodes non paramétriques pour l'estimation de courbes de référence – Application à l'analyse de propriétés biophysiques de la peau. *A paraître dans la Revue de Statistique Appliquée*.

Gannoun, A., Saracco, J. (2001). Cross Validation criteria for SIRα and PSIRα methods in view of prediction. *Proceeding of the 10th Symposium on.Applied Stochastic Models and Data Analysis* 443-448.

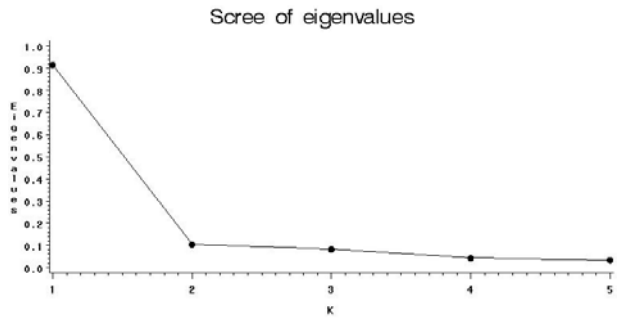
Li, K. C. (1991). Sliced Inverse Regression for dimension reduction, with discussions. *Journal of the American Statistical Association* **86** 316-382.

Saracco, J., Larramendy, I., Aragon, Y. (1999). La regression inverse par tranches ou méthode SIR : presentation générale. *La revue de Modulad* **22** 21-39

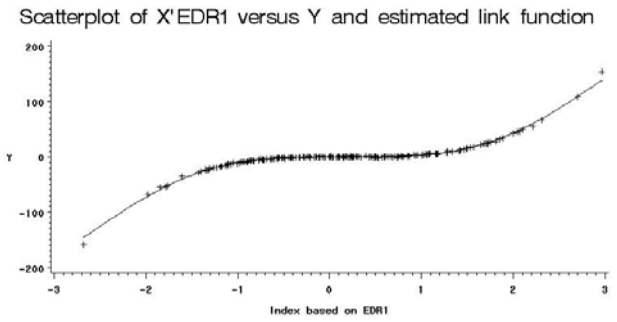
Saracco, J. (2001). Pooled Slicing methods versus Slicing methods. *Communications in Statistic – Simulation and Computation* **30** 489-511.

Schott, J. R. (1994). Determining the dimensionality in Sliced Inverse Regression. *Journal of the American Statistical Association* **89** 141-148.

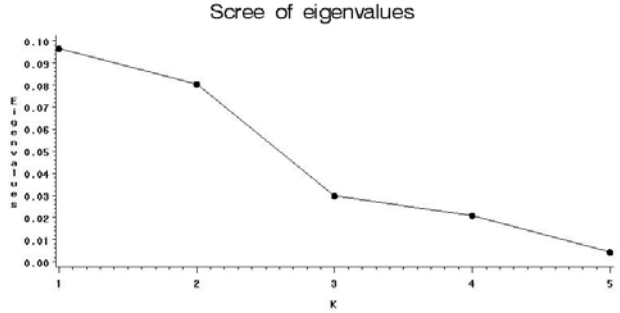
ANNEXE 1



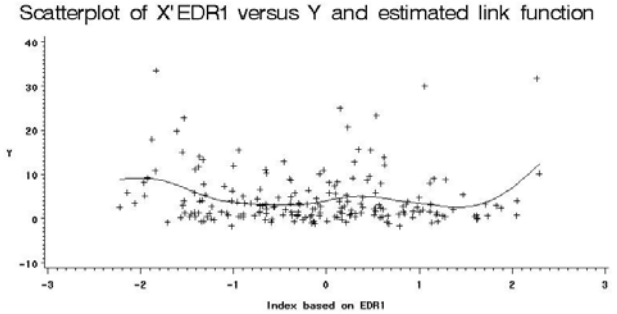
(a1) Méthode SIR-I avec la table SAS m1



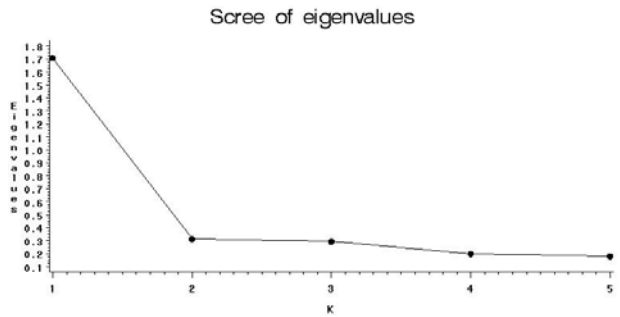
(a2) Méthode SIR-I avec la table SAS m1



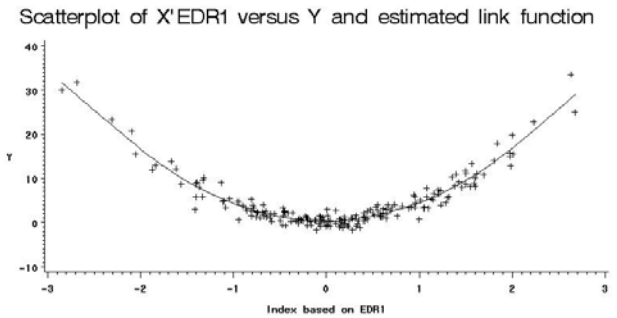
(b1) Méthode SIR-I avec la table SAS m2



(b2) Méthode SIR-I avec la table SAS m2

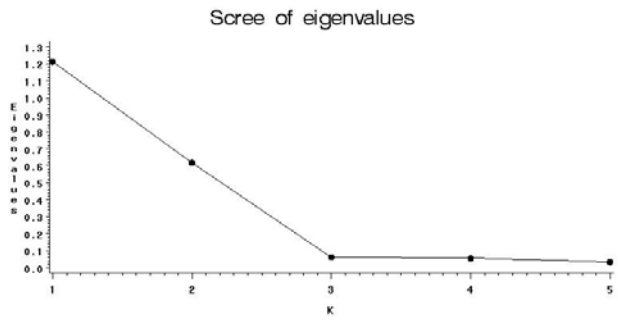


(c1) Méthode SIR-II avec la table SAS m2

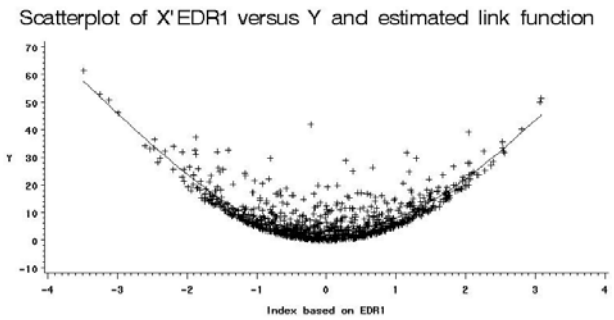


(c2) Méthode SIR-II avec la table SAS m2

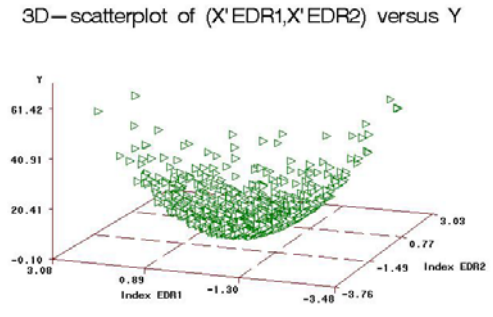
Figure 1 : Graphiques des éboulis des valeurs (à gauche) et graphiques des nuages du point croisant la variable à expliquer Y et le premier indice estimé (à droite).



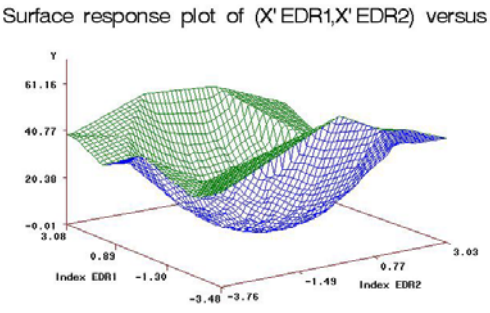
(d1) Méthode SIR-II avec la table SAS m3



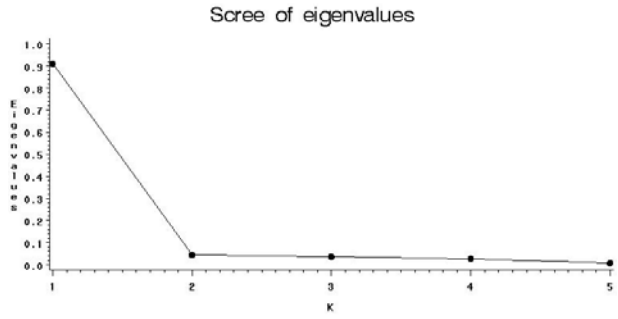
(d2) Méthode SIR-II avec la table SAS m3



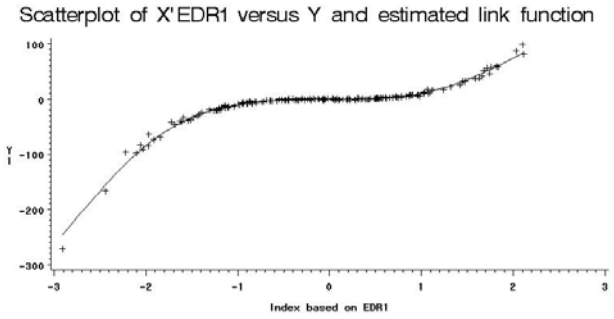
(d3) Méthode SIR-II avec la table SAS m3



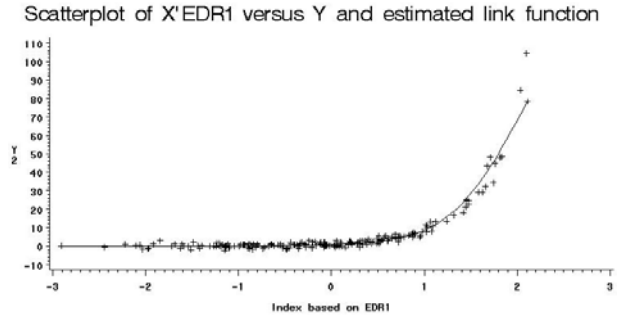
(d4) Méthode SIR-II avec la table SAS m3



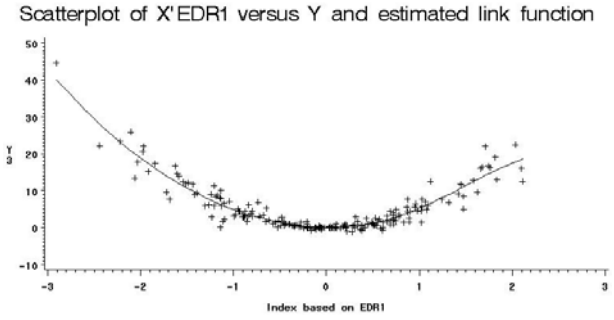
(e1) Complete Slicing avec la table SAS m4



(e2) Complete Slicing avec m4 (variable Y1)

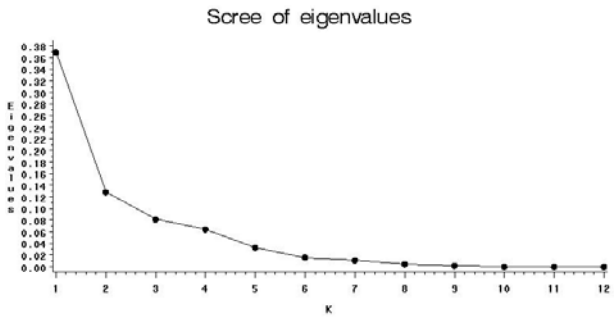


(e3) Complete Slicing avec m4 (variable Y2)

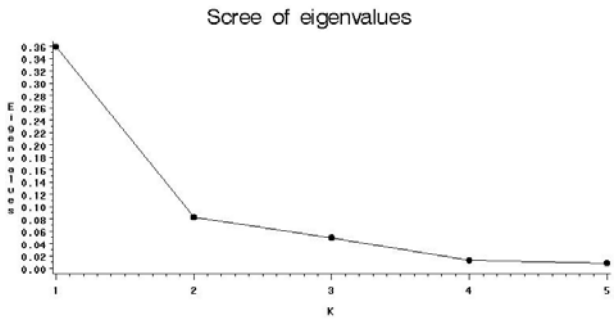


(e4) Complete Slicing avec m4 (variable Y3)

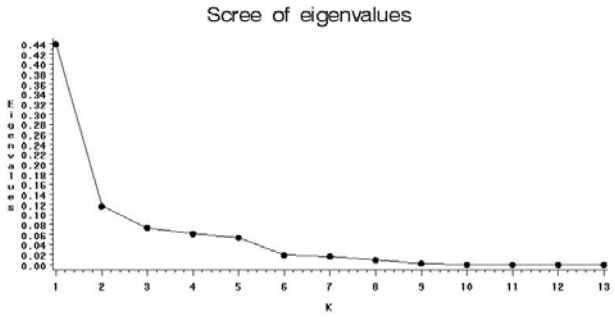
Figure 1 (suite) : Graphiques des éboulis des valeurs et graphiques des nuages du point croisant la (ou les) variable(s) à expliquer et le (ou les) premier(s) indice(s) estimé(s).



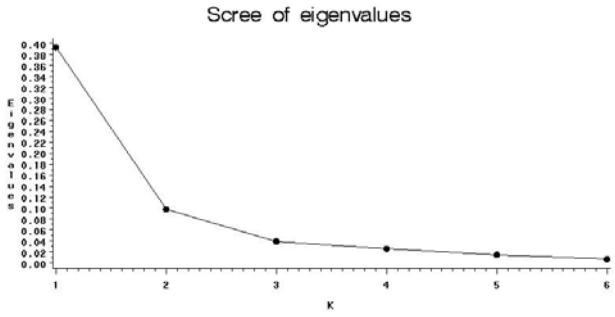
(a1) Zone de l'avant-bras (étape 1)



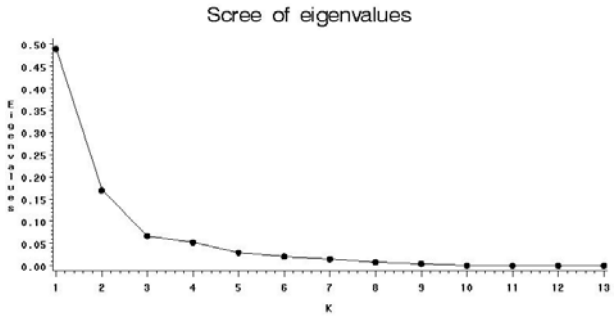
(a2) Zone de l'avant-bras (étape 2)



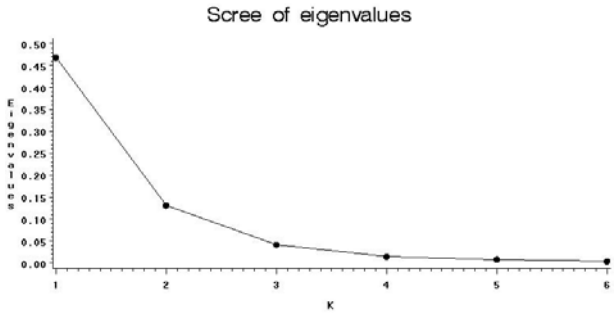
(b1) Zone du front (étape 1)



(b2) Zone du front (étape 2)



(c1) Zone de la joue (étape 1)



(c2) Zone de la joue (étape 2)

Figure 2 : Graphiques des éboulis des valeurs propres par zone pour les étapes 1 (avant simplification de l'index) et 2 (après simplification)

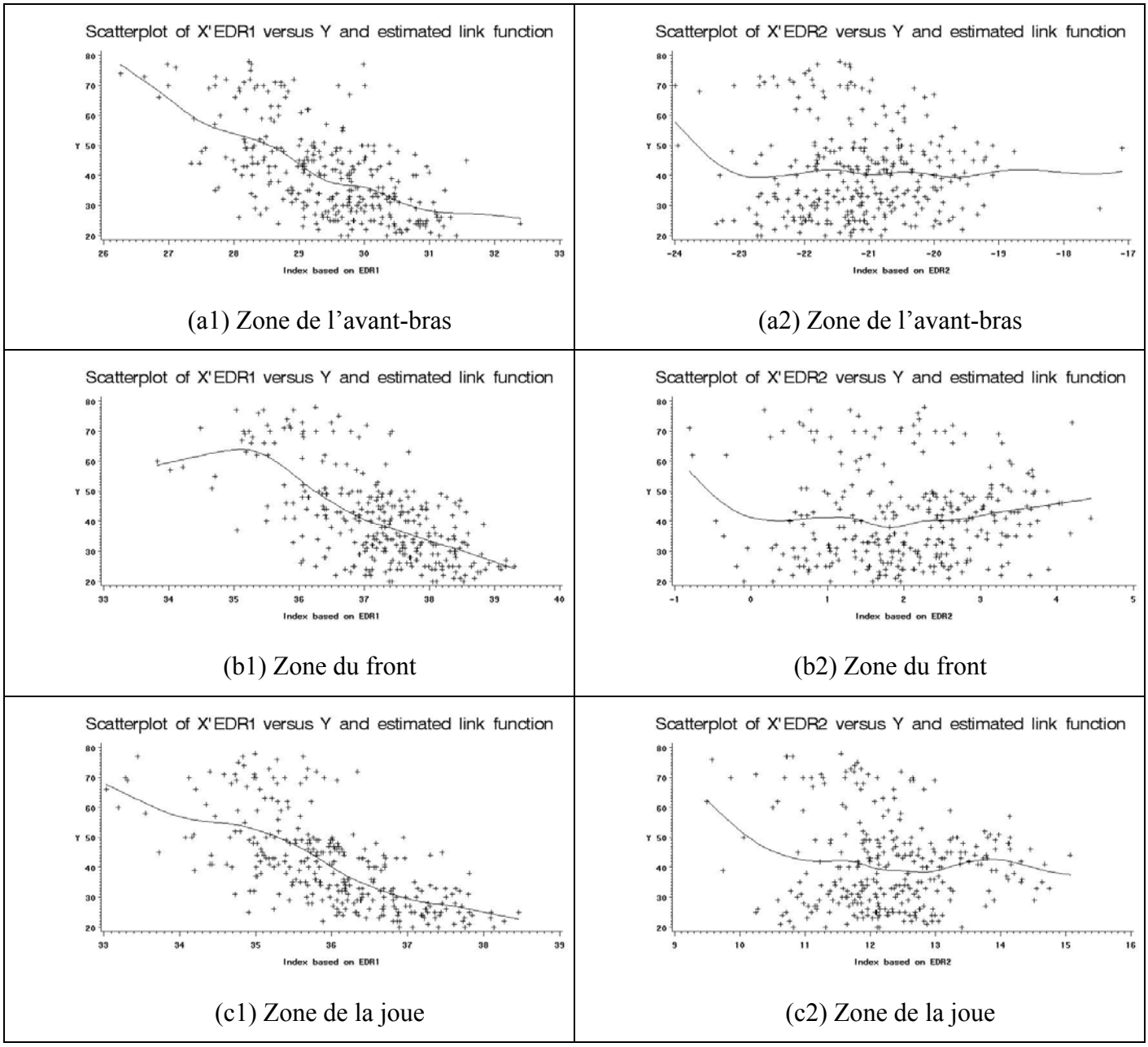
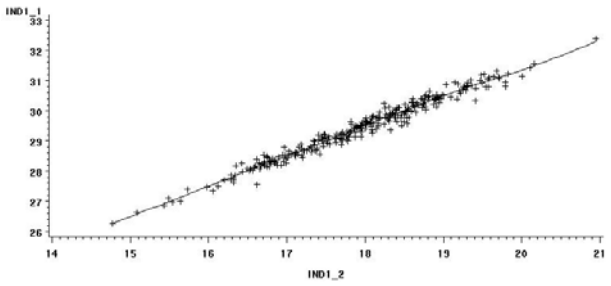
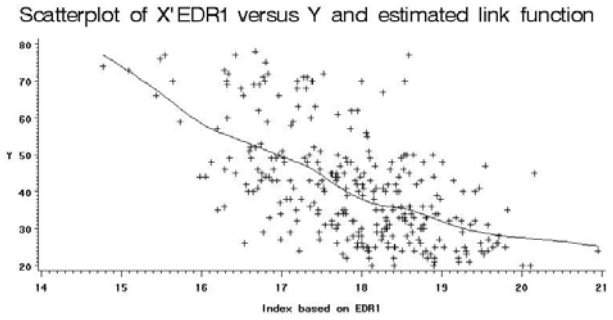


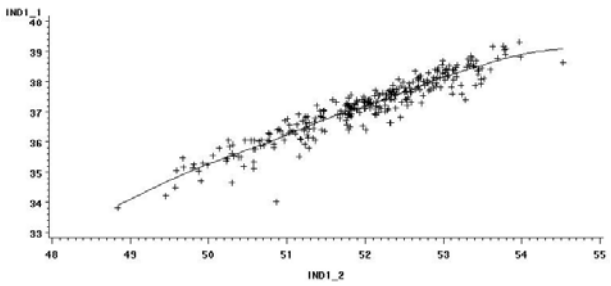
Figure 3 : Graphiques des nuages de points croisant la variable d'intérêt (la conductance de la peau) et les deux premiers indices pour chaque zone.



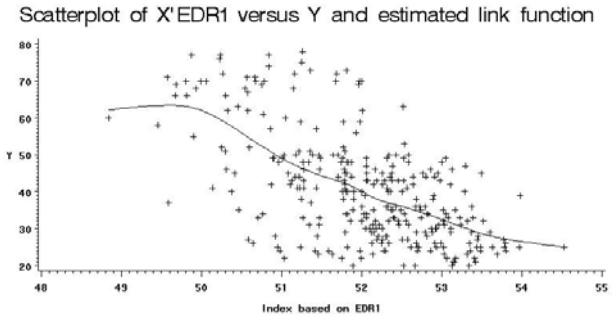
(a1) Zone de l'avant-bras



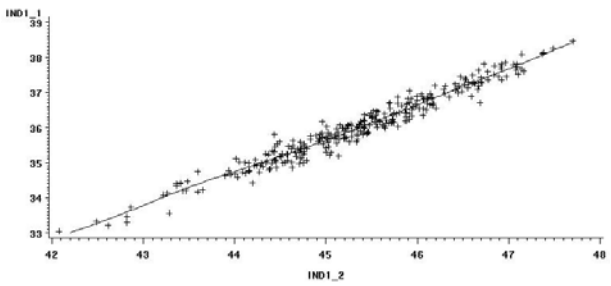
(a2) Zone de l'avant-bras (étape 2)



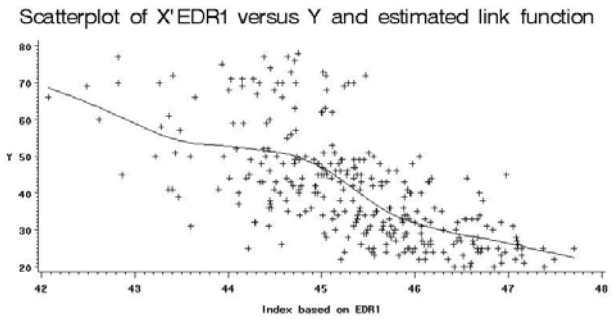
(b1) Zone du front



(b2) Zone du front (étape 2)



(c1) Zone de la joue



(c2) Zone de la joue (étape 2)

Figure 4 : Pour chaque zone, graphiques croisant les indices calculés avant et après leur simplification à l'étape 2 (à gauche), graphiques des nuages de points croisant la variable d'intérêt (la conductance de la peau) et l'indice calculés après simplification.

ANNEXE 2 : Quelques détails sur les méthodes de types « Slicing » et « Pooled Slicing »

Pour simplifier les notations, définissons z la version standardisée de x par $Z=\Sigma^{-1/2}(X-\mu)$, où μ est l'espérance de X et Σ sa variance. On appelle direction EDR standardisée associée à la $k^{\text{ème}}$ direction EDR β_k, le vecteur $\eta_k=\Sigma^{-1/2}\beta_k$. Ainsi, connaissant la $k^{\text{ème}}$ direction EDR standardisée η_k, on peut en déduire la $k^{\text{ème}}$ direction EDR : $\beta_k=\Sigma^{+1/2}\eta_k$.

● Dans le cadre de la théorie de la méthode **SIR-I** (respectivement **SIR-II** et **SIRα**), il a été démontré que les K vecteurs propres associés aux K plus grandes valeurs propres de la matrice M_I (resp. M_{II} et M_α) sont des directions EDR standardisées, où :

$$M_I=\text{var}(E[Z|T(Y)]),\ M_{II}=E\{\text{var}(Z|T(Y))-E[\text{var}(Z|T(Y))]\}^2 \text{ et } M_\alpha=(1-\alpha)M_I^2+\alpha M_{II},$$

la fonction T étant une transformation monotone de Y.

Considérons maintenant un échantillon $\{(Y_i,X_i), i=1,...,n\}$ et découpons l'étendue des y_i en H « tranches » $s_1,...,s_H$. Pour l'estimation des différentes matrices introduites ci-dessus, Li (1991) d'utiliser la fonction « tranchage » T définie comme suit :

$$T(Y_i)=\sum_{h=1}^{H} h\,1[Y_i\in s_h],\forall i=1,...,n.$$

Les matrices M_I et M_{II} s'écrivent alors :

$$M_I=\sum_{h=1}^{H} p_h m_h m_h' \text{ et } M_{II}=\sum_{h=1}^{H} p_h (V_h-\overline{V})(V_h-\overline{V})',$$

où $p_h=P(Y\in s_h)$, $m_h=E[Z|Y\in s_h]$, $V_h=V(Z|Y\in s_h)$ et $\overline{V}=\sum_{h=1}^{H} p_h V_h$.

En substituant les moments empiriques aux moments théoriques correspondants, on en déduit directement les matrices estimées $\hat{M}_I$, $\hat{M}_{II}$ et $\hat{M}_\alpha$ dont les vecteurs propres associés aux K plus grandes valeurs propres sont des directions EDR standardisées estimées.

● Pour les versions « Pooled Slicing » de ces méthodes, il faut introduire D tranchages différents notés T_d, $d=1,...,D$. Les matrices utilisées pour les méthodes **PSIR-I**, **PSIR-II** et **PSIRα** sont, respectivement :

$$M_I^P=\frac{1}{D}\sum_{d=1}^{D} M_I^d,\ M_{II}^P=\frac{1}{D}\sum_{d=1}^{D} M_{II}^d \text{ et } M_\alpha^P=(1-\alpha)(M_I^P)^2+\alpha M_{II}^P,$$

les matrices M_I^d et M_{II}^d étant définies de manière analogue aux matrices M_I et M_{II} pour le tranchage T_d. Aragon et Saracco (1997) et Saracco (2001) ont démontré que les K vecteurs propres associés aux K plus grandes valeurs propres de la matrice M_I^P (resp. M_{II}^P et M_α^P) sont des directions EDR standardisées L'estimation de ces trois matrices M_I^P, M_{II}^P et M_α^P est alors obtenue en estimant les matrices M_I^d et M_{II}^d comme précédemment mais en utilisant le tranchage T_d.

Remarque sur le choix des tranchages : Pour l'ensemble des méthodes, les tranchages sont construits de façon à ce que chaque tranche contienne à peu près le même nombre

d'individus (c'est à dire que les poids p_h sont à peu près égaux). Pour les méthodes de type « Slicing », le nombre H de tranches est choisi automatiquement en fonction de la taille n de l'échantillon. En ce qui concerne les méthodes de type « Pooled Slicing », le choix des différents tranchages T_d est géré par deux paramètres permettant de déterminer (en fonction de n) le nombre D de tranches utilisées dans l'estimation. Ces deux paramètres sont :

- le nombre minimum de tranches pour un tranchage : H_{min},
- le nombre minimum d'individus par tranche à prendre pour un tranchage : n_{min}.

Par exemple, si on dispose d'un échantillon de taille n=100, et que l'utilisateur choisi les valeurs H_{min}=8 et n_{min}=10, les trois tranchages suivants seront utilisés pour l'estimation des méthodes de type « Pooled Slicing » :

tranchage 1 : H_{min}=8 tranches dont 4 de 12 individus et 4 de 13 individus,

tranchage 2 : 9 tranches dont 8 tranches de 11 individus et une de 8,

tranchage 3 : 10 tranches de n_{min}=10 individus.

Remarque sur le choix du nombre *K* de directions EDR à retenir : Des tests ont été écrits pour déterminer le nombre de directions EDR nécessaires (voir Ferré (1998) ou Schott (1994)), mais ces tests n'ont pas encore été implémentées sous le logiciel SAS®. Ainsi, on prendra le critère empirique suivant pour déterminer K : on recherche dans la liste des valeurs propres (ou sur le graphique correspondant) une rupture entre celles qui semblent être significativement supérieures aux autres.

Remarque sur le choix du paramètre α : la valeur de α doit être choisie de manière adaptative. Deux méthodes de choix automatique pour ce paramètre ont été développées : l'une d'elles est basée sur des procédures de tests (voir Saracco (2001)), l'autre repose sur un critère de validation croisée (voir Gannoun et Saracco (2001)). Ces deux techniques n'ont pas encore été implémentées sous le logiciel SAS®.

Pour de plus détails et de références bibliographiques, le lecteur pourra se reporter à Saracco *et al* (1999) où une présentation générale des méthodes SIR est donnée.

● Lorsque la variable à expliquer Y est multidimensionnelle, la méthode **Pooled Marginal Slicing**, agrège les diverses informations provenant des q composantes de Y. Considérons les q transformations (« tranchages ») T_j de chaque composante Y_j de Y. Notons $M_j = \mathrm{var}(E[Z|T_j(Y)])$. Aragon (1997) montre que les K vecteurs propres associés aux K plus grandes valeurs propres de la matrice $M^P = \frac{1}{q}\sum_{j=1}^{q} M_j$ sont des directions EDR standardisées. L'estimation des matrices M_j se fait de manière analogue au cas unidimensionnel pour la variable à expliquer Y_j.

D'autres techniques ont été développées dans ce cadre d'un Y multidimensionnel, il s'agit des méthodes du **Marginal Slicing** et de l'**Alternating SIR.** Ces deux méthodes n'ont pas encore été implémentées sous forme de macro-commandes SAS. Le lecteur pourra trouver des renseignements et des références sur ces méthodes dans Saracco *et al* (1999).

ANNEXE 3 : Extraits des sorties numériques obtenues (données simulées)

● **Etude de la table SAS m1**

==========

Méthodes de Régression Inverse par Tranches (SIR univariée)

==========

Méthode SIR-I

==========

	P
Le nombre de variables explicatives est :	5
	N
Le nombre total d'individus est :	200
	NBTR
Le nombre de tranches utilisées est :	10

La répartition des individus par tranches est la suivante :

1	20
2	20
3	20
4	20
5	20
6	20
7	20
8	20
9	20
10	20

Les valeurs propres estimées des directions EDR sont :

VALPREDR
0.894304
0.071418
0.0517187
0.0158516
0.0100052

Direction(s) EDR estimée(s) :

	EDR1
X1	0.52125
X2	0.51237
X3	-0.53626
X4	-0.52064
X5	0.00118

Méthode PSIR-I

==========

NBMINTR

Le nombre minimum de tranches choisi pour le Pooling est : 10

NBMININD

Le nombre minimum d'individus par tranche choisi pour le Pooling est : 10

NBVECTR

Le nombre total de tranchages considérés pour le Pooling est : 11

P

Le nombre de variables explicatives est : 5

On obtient ensuite le détail de chacun des 11 tranchages (on ne fournit pas ici les sorties correspondantes).

Les valeurs propres estimées des directions EDR sont :

VALPREDR
0.9147043
0.1032708
0.0816372
0.0419382
0.0335431

Direction(s) EDR estimée(s) :

	EDR1
X1	0.53052
X2	0.50277
X3	-0.52354
X4	-0.53608
X5	-0.00496

● **Etude de la table SAS m2**

Méthode SIR-I

==========

(On ne donne pas ici la répartition des individus par tranches qui est identique à celle de l'exemple précédent.)

Les valeurs propres estimées des directions EDR sont :

VALPREDR
0.0964533
0.0804692
0.0297549
0.0208476
0.0042104

Direction(s) EDR estimée(s) :

	EDR1
X1	0.22646
X2	-0.27374
X3	0.60204
X4	0.18507
X5	0.70294

Méthode SIR-II

==========

(On ne fournit ici les valeurs propres et la direction EDR estimée.)

Les valeurs propres estimées des directions EDR sont :

VALPREDR
1.7067041
0.3132924
0.2930843
0.1994578
0.1811488

Direction(s) EDR estimée(s) :

	EDR1
X1	0.58730
X2	0.45192
X3	-0.37502
X4	-0.50534
X5	-0.01513

Méthode PSIR-Alpha

==========

NBMINTR

Le nombre minimum de tranches choisi pour le Pooling est : 10

NBMININD

Le nombre minimum d'individus par tranche choisi pour le Pooling est : 10

NBVECTR

Le nombre total de tranchages considérés pour le Pooling est : 11

ALPHA

La valeur choisie pour Alpha est : 0.5

(On ne fournit ici que les valeurs propres et la direction EDR estimée.)

Les valeurs propres estimées des directions EDR sont :

VALPREDR

0.9210995

0.2088512

0.1822934

0.1707416

0.1318815

Direction(s) EDR estimée(s) :

	EDR1
X1	-0.55810
X2	-0.42727
X3	0.42918
X4	0.51615
X5	0.02990

● **Etude de la table SAS m3**

Méthode SIR-II

==========

(On ne fournit ici que les valeurs propres et la direction EDR estimée.)

Les valeurs propres estimées des directions EDR sont :

VALPREDR
1.2133132
0.6179877
0.062553
0.0549464
0.0338443

Direction(s) EDR estimée(s) :

	EDR1	EDR2
X1	-0.09146	0.53322
X2	-0.06984	0.55973
X3	0.07132	-0.62736
X4	-0.46843	-0.05891
X5	0.86883	0.11063

● **Etude de la table SAS m4**

==========

Méthodes de Régression Inverse par Tranches (SIR multivariée)

==========

Complete Slicing

==========

(On ne fournit ici que les valeurs propres et la direction EDR estimée.)

Les valeurs propres des directions EDR sont :

VALEDR
0.9107777
0.0447011
0.0357902
0.0272404
0.0075109

Direction(s) EDR estimée(s) :

	EDR1
X1	0.49452
X2	0.44385
X3	-0.44974
X4	-0.46489
X5	0.03090

Pooled Marginal Slicing

==========

(On ne fournit ici que les valeurs propres et la direction EDR estimée.)

Les valeurs propres des directions EDR sont :

VALEDR
0.5588059
0.0606308
0.0554157
0.0333434
0.0206498

Direction(s) EDR estimée(s)

	EDR1
X1	0.51323
X2	0.45670
X3	-0.44510
X4	-0.42798
X5	0.03613