# STATISTIQUE ET LOGICIELS

## Implémentation en Splus des méthodes SIR univariées et multivariées

Saracco Jérôme

Laboratoire de Probabilités et Statistique,
Département de mathématiques, case courrier 051,
Université Montpellier II,
Place Eugène Bataillon,
34 095 Montpellier Cédex 5.

Email : saracco@stat.math.univ-montp2.fr

Nous donnons dans cet article le mode d'emploi de l'implémentation en SPLUS des méthodes SIR (Sliced Inverse Regression) univariées ou multivariées. Nous ne fournissons ici aucun rappel théorique sur les méthodes d'estimation ; les différentes références bibliographiques sont précisées pour chacune des méthodes. Ce document est une description des différentes possibilités qu'offre le programme principal. Ce dernier permet à l'utilisateur, ayant une matrice de données à analyser,
(*i*) de choisir, parmi les variables du fichier, la (ou les) variable(s) dépendante(s) et les variables explicatives retenues pour l'étude,
(*ii*) de sélectionner dans le menu proposé (en fonction de son choix de variables), une des méthodes SIR univariées ou multivariées disponibles,
(*iii*) d'obtenir des estimations des directions EDR, de la fonction de lien, des graphiques, ...

## 1 Introduction

La régression inverse par tranches (Sliced Inverse Regression (SIR)) est une méthode de régression semiparamétrique qui contrairement aux autres méthodes de régression semiparamétrique, ne requiert que des temps de calcul informatique très courts. Ce document fait suite à l'article *"La régression inverse par tranches ou méthode SIR : présentation générale"* de Saracco, Larramendy et Aragon (1998) dans lequel les différentes méthodes SIR univariées et multivariées ont été présentées. Nous rappelons que les méthodes SIR univariées (resp. multivariées) font référence au fait que la variable à expliquer $y$ est unidimensionnelle (resp. multidimensionnelle). Pour une meilleure compréhension de cette mise en oeuvre informatique, la lecture de l'article mentionné précédemment apparaît nécessaire. Nous donnons ici une illustration et un descriptif de l'implémentation informatique des méthodes SIR réalisée en SPLUS. Les différentes procédures et fonctions sont disponibles auprès de l'auteur.

Nous rappelons brièvement les modèles de régression semiparamétrique de référence considérés dans les méthodes SIR.

Le modèle univarié de SIR est le suivant :

$$\begin{cases} y & = & f(x'\beta_1, \ldots, x'\beta_K, \epsilon) \\ \epsilon & \perp & x \end{cases} \tag{1}$$

où $y$ est une variable á expliquer univariée, $x$ une variable explicative $p$-dimensionnelle avec $p > 1$ (afin que l'introduction des vecteurs de paramètres de réduction de dimension $\beta_1, \ldots, \beta_K$ avec $K < p$ ait un intérêt) et $\epsilon$ un terme aléatoire d'erreur de loi inconnue et arbitraire. $f$ est appelée *fonction de lien*, elle est inconnue et arbitraire. Les vecteurs $\beta_1, \ldots, \beta_K$ de $I\!R^p$ sont linéairement indépendants et inconnus.
Le modèle multivarié est une extension du modèle (1) à un modèle à $q$ équations :

$$\begin{cases} y_1 & = & f_1(x'\beta_1, \ldots, x'\beta_K, \epsilon_1) \\ & \vdots & \\ y_q & = & f_q(x'\beta_1, \ldots, x'\beta_K, \epsilon_q) \\ \epsilon_j & \perp & x \quad \text{pour } j = 1, \ldots, q \end{cases} \tag{2}$$

Dans ce système, les fonctions de lien $f_1, \ldots, f_q$ et les vecteurs de paramètres $\beta_1, \ldots, \beta_K$ sont inconnus. La variable $y$ à expliquer est donc ici de dimension $q$.

Dans ces deux modèles, seul l'espace des paramètres $E = \text{Vect}(\beta_1, \ldots, \beta_K)$ (i.e. le sous-espace linéaire de dimension $K$ de $I\!R^p$ engendré par les vecteurs $\beta_1, \ldots, \beta_K$) est identifiable, chaque $\beta_k$ n'étant en effet pas individuellement et totalement identifiable. $E$ est appelé l'*espace EDR* (comme Effective Dimension Reduction). Toute direction appartenant au sous-espace EDR est appelée *direction EDR*, i.e. toute combinaison linéaire des $\beta_k$ est une direction EDR. Les termes $x'\beta_k$ sont eux appelés les indices du modèle.

Les différentes méthodes SIR permettent d'*estimer une base $\hat{\Sigma}$-orthonormales de l'espace EDR $E$*, , où $\hat{\Sigma}$ est la matrice de covariance empirique des variables explicatives.
L'estimation du paramètre fonctionnel $f$ dans le modèle (1) peut être obtenue par les méthodes de régression non paramétrique usuelles (méthodes des noyaux ou des splines, voir Eubank (1988)), ceci d'autant plus facilement que le paramètre $K$ de réduction de dimension du modèle est petit. Il en est de même pour l'estimation nonparamétrique des fonctions de liens $f_1, \ldots, f_q$ dans le modèle (2).

Une petite bibliographie sur SIR est donnée à la fin de ce document. Précisons tout de même que la méthode SIR a été introduite par Duan et Li (1991) dans le cadre de modèle univarié à un seul indice et par Li (1991) dans le cadre de modèle univarié à $K$ indices. Dans le cadre de ces modèles, diverses "méthodes SIR alternatives" ont été décrite par Aragon et Saracco (1997), Hsing et Carroll (1992) ... La méthode SIR a été ensuite étendue au cas où la variable explicative est multidimensionnelle par Aragon, Li et Thomas-Agnan (1995) et Li, Aragon et Thomas-Agnan (1997). Il est aussi possible de trouver un panorama sur toutes les méthodes SIR dans Saracco (1996 a).

## 1.1 Introduction à l'implémentation en Splus des méthodes SIR

L'implémentation réalisée comprend un programme principal permettant à l'utilisateur ayant une matrice de données à analyser, de sélectionner, en fonction du choix de la (ou

des) variable(s) á expliquer et des variables explicatives qu'il a fait, une des méthodes de régression inverse par tranches disponibles. Il est à noter que ce programme principal permet d'obtenir des sorties graphiques, il est donc nécessaire à l'utilisateur d'ouvrir la fenêtre graphique de SPLUS avant de lancer le programme principal.
Les données à étudier doivent être mises sous la forme d'une matrice SPLUS avec les individus en lignes et les variables en colonnes. Nous donnons, à la sous-section 1.3, trois modèles que nous avons utilisés pour simuler des matrices de données qui nous permettent d'illustrer le fonctionnement du programme et des méthodes.

**Lancement du programme principal.** Après avoir installé toutes les procédures dans Splus (voir la liste des procédures à la section 5), il suffit à l'utilisateur de taper

```
resultatSIR <- ProgPrincipal.prog(matrice)
```

où "matrice" est la matrice des données à étudier. L'utilisateur n'aura ensuite qu'à suivre les instructions qui s'afficheront à l'écran en fonction de ces choix au fur et à mesure du déroulement du programme. Des résultats numériques (directions EDR estimées, estimation nonparamétrique des $y_i$, ...) seront affectés dans l'objet Splus "resultatSIR".

## 1.2 Sélection des variables et méthodes SIR disponibles

L'étape de sélection des variables est décrite à la sous-section suivante. Les méthodes SIR univariées et multivariées disponibles sont présentées à la sous-section 1.2.2 ainsi que quelques références bibliographiques. Leur utilisation est détaillée à la section 2 (resp. 3) pour les méthodes univariées (resp. multivariées).

### 1.2.1 Sélection des variables à expliquer ou explicatives

Il est demandé tout d'abord à l'utilisateur s'il désire ou non afficher les listes des noms des individus et des variables de la matrice de données. L'utilisateur doit ensuite sélectionner le type de chacune des variables de la matrice de données : variables à expliquer (variable dépendante) ou bien variable explicative ou bien encore variable ne rentrant pas dans l'étude. Dans le cas où le choix ne serait pas judicieux dans le cadre de SIR : absence de variable á expliquer, absence de variable explicative, une seule variable explicative (dans ce cas-là, il n'y a en effet pas besoin de faire de la réduction de dimension), ..., des messages d'erreur s'affichent à l'écran et il est demandé de saisir à nouveau le choix des types de variables.

```
================================================
================================================
Methodes de Regression Inverse par Tranches (SIR)
        unidimensionnelle ou multidimensionnelle
et Methodes alternatives
================================================
================================================

================================================
Choix des variables dans la matrice des donnees.
================================================
Nombre d'individus =  200
Nombre de variables =  6
```

```
Affichage des noms des individus (1=oui/2=non) ? 1: 1

  [1] "ind1"   "ind2"   "ind3"   "ind4"   "ind5"   "ind6"   "ind7"   "ind8"
  [9] "ind9"   "ind10"  "ind11"  "ind12"  "ind13"  "ind14"  "ind15"  "ind16"
 [17] "ind17"  "ind18"  "ind19"  "ind20"  "ind21"  "ind22"  "ind23"  "ind24"
 [25] "ind25"  "ind26"  "ind27"  "ind28"  "ind29"  "ind30"  "ind31"  "ind32"
 [33] "ind33"  "ind34"  "ind35"  "ind36"  "ind37"  "ind38"  "ind39"  "ind40"
 [41] "ind41"  "ind42"  "ind43"  "ind44"  "ind45"  "ind46"  "ind47"  "ind48"
 ....
[185] "ind185" "ind186" "ind187" "ind188" "ind189" "ind190" "ind191" "ind192"
[193] "ind193" "ind194" "ind195" "ind196" "ind197" "ind198" "ind199" "ind200"
Affichage des noms des variables (1=oui/2=non) ? 1: 1

[1] "Y"  "X1" "X2" "X3" "X4" "X5"
Choix des variables dependantes et independantes :
1=var. dependante, 2=var. independante, 3=var. a supprimer
Y  - 1: 1
X1  - 1: 2
X2  - 1: 2
X3  - 1: 2
X4  - 1: 2
X5  - 1: 2
Var. dependantes choisies :
Y
Var. independantes choisies :
X1   X2   X3   X4   X5
-------------------------------------------
 - Nombre de variables dependantes =  1
 - Nombre de variables explicatives =  5
-------------------------------------------
```

### 1.2.2 Méthodes SIR disponibles

En fonction du choix des variables, le programme propose alors de choisir une méthode d'estimation parmi diverses propositions. Pour cela, un menu apparaît à l'écran. Nous donnons ci-dessous la liste des méthodes univariées (menu 1) et multivariées (menu 2) disponibles et les références bibliographiques les concernant.

- S'il n'y a qu'une seule variable à expliquer, le menu est le suivant :

*Menu 1 : méthodes SIR univariées disponibles.*

```
================================================
Choix des methodes disponibles
================================================

1: SIR classique (Duan et Li)
2: Pooled Slicing (PSIR)
3: SIR Hsing-Carroll (tranches distinctes)
4: SIR Hsing-Carroll (tranches glissantes)
5: Kernel Smoothing Inverse Regression
Selection: 1
```

On distingue la méthode SIR "classique" (initialement proposée par Duan et Li(1991) et par Li(1991)) avec plusieurs choix possibles pour l'étape de tranchage (découpage automatique en tranches, ou choix par l'utilisateur du découpage : tranches d'égale amplitude, tranches de poids égal,...) et les méthodes SIR "alternatives" : la méthode du Pooled Slicing (Aragon et Saracco (1997)), les méthodes SIR de Hsing et Carroll (1992) avec des tranches distinctes ou glissantes de 2 individus et la méthode de Régression Inverve par

Lissage (Aragon et Saracco (1997)) appelée dans ce menu "Kernel Smoothing Inverse Regression".

• Dans le cas où plusieurs variables á expliquer ont été sélectionnées, le menu proposé est alors :

*Menu 2 : méthodes SIR multivariées disponibles.*

```
================================================
Choix des methodes disponibles.
================================================

1: Complete Slicing
2: Marginal Slicing
3: Pooled Marginal Slicing
4: Alternating SIR
Selection: 3
```

Des résultats théoriques sur ces méthodes peuvent être trouvés dans Aragon, Li et Thomas-Agnan (1995) pour les méthodes du Complete Slicing, du Marginal Slicing et du Pooled Marginal Slicing, et dans Li, Aragon et Thomas-Agnan (1997) pour la méthode d'Alternating SIR.

**Une remarque sur le nombre de directions EDR à retenir.** Quelle que soit la méthode SIR sélectionnée, le nombre de directions EDR à retenir est choisi par l'utilisateur (lors de l'exécution du programme), en fonction des valeurs propres de la matrice diagonalisée (notée $\hat{\Gamma}_T$ dans Saracco, Larramendy et Aragon (1998)). On rappelle que les directions EDR correspondent à des valeurs propres non nulles de la matrice $\Gamma_T$. Des procédures permettant d'estimer cette dimension $K$ ont été developpées par Ferré (1996, 1998).

## 1.3 Modèles et matrices de données simulés

Nous avons simulé 5 matrices de données provenant de 4 modèles différents. Ces matrices vont nous permettre de parcourir l'ensemble des méthodes disponibles.

Les deux premières matrices simulées, **m1a** et **m1b**, proviennent du modèle univarié à un seul indice suivant :

$$\text{Modèle A :} \quad y = (x'\beta)^3 - (x'\beta)^2 - 3x'\beta + \varepsilon$$

où $x \sim N_5(0_5, I_5)$, $\varepsilon \sim N_1(0,1)$ indépendante de $x$, et $\beta = \frac{1}{2}(1,1,1,1,0)'$. La taille de ces deux échantillons simulés est $n = 200$.

La troisième matrice simulée **m2** provient du modèle univarié à un seul indice suivant :

$$\text{Modèle B :} \quad y = (x'\beta)^3 + \varepsilon$$

où $x \sim N_5(0_5, I_5)$, $\varepsilon \sim N_1(0,1)$ indépendante de $x$, et $\beta = \frac{1}{2}(1,1,1,1,0)'$. La taille de cet échantillon simulé est $n = 50$.

La matrice **m3**, de taille $n = 500$, a été simulée à partir d'un modèle univarié à deux indices :

$$\text{Modèle C :} \quad y = (x'\beta_1)^3 + 5x'\beta_2 + \varepsilon$$

avec $x \sim N_5(0_5, I_5)$, $\varepsilon \sim N_1(0,1)$ indépendante de $x$, et les vecteurs de paramètres $\beta_1 = (1,1,1,0,0)'$, $\beta_2 = (0,0,0,1,1)'$.

Enfin, **m4**, la dernière matrice de données simulées, provient d'un modèle de régression multivarié à un seul indice :

$$\text{Modèle D :} \quad y = \begin{cases} y_1 &= x'\beta + \varepsilon_1 \\ y_2 &= (x'\beta + \varepsilon_2)^2 \\ y_3 &= (x'\beta + \varepsilon_3)^3 \end{cases}$$

avec $x \sim N_5(0_5, \Sigma)$ où la matrice $\Sigma$ est telle que l'élément $(i,j)$ est égal à $\min(i,j)$. Les erreurs $\varepsilon_j \sim N_1(0,1)$, $j = 1, \ldots, 3$ sont indépendantes entre elles et indépendantes de $x$, et $\beta = (1,1,1,1,0)'$.

Ainsi, les matrices de données **m1a, m1b, m2, m3** contiennent, en colonnes, les valeurs des variables $y$, $x_1$, ..., $x_5$ mesurées sur les $n$ individus, et la matrice **m4** les valeurs des variables $y_1$, $y_2$, $y_3$, $x_1$, ..., $x_5$. Dans chacune de ces matrices, les individus sont codés : ind1, ind2, ind3, ...

Les matrices **m1a, m1b** et **m3** vont servir à illustrer la méthode SIR "classique" (par opposition aux méthodes SIR "alternatives"). La matrice **m2** de taille plus modeste va nous permettre de regarder le fonctionnement des méthodes SIR "alternatives". La matrice **m4** sera utilisée pour faire tourner les méthodes de SIR multivariées.

# 2 Méthodes SIR univariées

Elles correspondent au choix d'une des méthodes dans le Menu 1.

Nous appelons ici que les méthodes alternatives (à un tranchage particulier et arbitraire, fixé par l'utilisateur) ont surtout un intérêt dans le cadre d'échantillon de petite taille. La méthode SIR classique est en effet très peu sensible au choix des tranches lorsque la taille de l'échantillon est importante. Il faut cependant noter que le nombre $H$ de tranches doit être supérieur strictement au nombre $K$ de directions à retenir, afin de ne pas avoir une réduction de dimension artificielle. De plus, $H$ doit aussi être inférieur à la partie entière de $n/2$, où $n$ est la taille de l'échantillon, afin qu'il y ait au moins deux individus par tranche.

## 2.1 Méthode SIR classique

Il s'agit ici du choix 1 dans le menu 1. Nous illustrons cette méthode avec les matrices de données **m1a, m1b** et **m3**. Nous présentons tout d'abord les différents choix de type de tranchage puis les sorties numériques et graphiques fournies par le programme.

### 2.1.1 Choix du type de tranchage

Il est demandé à l'utilisateur s'il désire choisir personnellement le type de tranchage ou s'il préfère laisser le programme gérer automatiquement le découpage en tranches.
Dans le cas où l'utilisateur opte pour un tranchage particulier, il lui est alors proposé un nouveau menu permettant de sélectionner tel ou tel type de découpage en tranches :

```
==============================
Methode : SIR classique
==============================

Voulez-vous choisir le type de tranchage ?
   1=Automatique / 2=Choix personnel
1: 2
------------------------
Choix du type de tranchage

1: Choix du nombre de tranches (tranches de largeur constante)
2: Choix de la largeur (constante) des tranches
3: Choix du nombre de tranches (tranches contenant le meme nb d'individus)
4: Choix du nombre (constant) d'individus par tranche
5: Tranchage automatique
Selection: 5
```

**Etape préliminaire automatique de tranchage.** Avant de décrire les différents types de tranchage, il est intéressant de préciser qu'une *étape préliminaire automatique de tranchage* a lieu avant un éventuel découpage spécifique en tranches. Cette étape permet de prendre en compte des échantillons pour lesquels la variable à expliquer $y$ prend des valeurs particulières. Chaque type de tranchage tient compte de cette spécificité. Ce traitement initial des données permet de repérer s'il y a des $y_i$ ayant des valeurs identiques.
Par exemple, lorsque la variable dépendante $y$ est discrète, le programme construit autant de tranches qu'il y a de niveaux différents de $y$ et chaque tranche contient alors les individus de l'échantillon ayant cette modalité pour la variable à expliquer. Dans cet exemple particulier, le tranchage a donc déjà été réalisé à cette étape préliminaire. Un message apparaît alors à l'écran et précise cet état de fait.
Un autre exemple met en exergue la nécessité de cette étape préliminaire : il peut arriver que l'on dispose d'un échantillon dans lequel certaines valeurs de la variable dépendante $y$ soient censurées ou manquantes (c'est le cas des modèles de sélection). Le découpage préliminaire fabrique alors ici une tranche contenant les individus dont la valeur de $y$ est manquante (ou bien une tranche par niveau de censure) et regroupe les individus restants (pour lesquels la valeur de $y$ n'est ni manquante ni censurée) dans une tranche particulière. Un message apparaît alors à l'écran et précise que certains individus ont déjà été affectés dans une tranche, ainsi que le nombre d'individus restant à affecter. C'est sur ces individus restants que l'étape suivante de découpage en tranches (pour laquelle il faut préciser le type de tranchage désiré) va porter. Dans le cas où les valeurs de la variable à expliquer n'ont pas une des particularités mentionnées ci-dessus, l'étape de tranchage préliminaire n'a pas d'incidence, et le tranchage désiré se fait alors sur la totalité de l'échantillon. L'utilisateur doit préciser le type de tranchage qu'il souhaite effectuer sur les données. Nous donnons ci-dessous quelques explications concernant ces divers découpages en tranches.

**Tranchage de type 1.** L'utilisateur doit choisir le nombre $H$ de tranches et le programme construit automatiquement des tranches de largeur égale recouvrant l'étendue des valeurs prises par la variable à expliquer $y$, ceci après avoir demandé à l'utilisateur s'il désire ou non choisir les bornes des tranches extrêmes. (Ce choix des bornes extrêmes est contrôlé afin de s'assurer que tous les individus de l'échantillon seront pris en compte : ainsi, la borne inférieure de la première tranche doit être inférieure ou égale à la plus petite valeur de $y$ de l'échantillon, et la borne supérieure de la dernière tranche doit être supérieure à la plus grandeur valeur de la variable dépendante. Si le choix des bornes ne satisfait pas ces conditions, un message d'erreur est affiché et il faut alors ressaisir les bornes.)

```
------------------------
Tranchage de type 1 :
choix du nombre de tranches, et tranches de largeur constante
------------------------
Nombre d'individus pris en compte =  200
Valeur minimale de Y =  -10.2487597347281
Valeur maximale de Y =  5.2422306933793
Etendue de Y =  15.4909904281074
Entrer le nombre de tranches choisi - 1: 10
------------------------
Choix des bornes de tranches extremes :

1: Automatique(Borne Inf=min(Y), Borne Sup=max(Y))
2: Manuel
Selection: 1
------------------------
Les bornes de tranches calculees sont :
 [1] -10.2487597  -8.6996607 -7.1505616 -5.6014626 -4.0523636 -2.5032645
 [7]  -0.9541655   0.5949336  2.1440326  3.6931317  5.2422307
```

**Tranchage de type 2.** Il est demandé à l'utilisateur de choisir la largeur de tranche désirée ainsi que les bornes des tranches extrêmes (voir le tranchage de type 1 pour quelques commentaires à ce sujet). Le programme découpe alors l'étendue des valeurs de $y$ en un nombre de tranches de largeur égale et le plus proche possible de la largeur désirée par l'utilisateur.

```
------------------------
Tranchage de type 2 :
choix de la largeur constante des tranches
------------------------
Nombre d'individus pris en compte =  200
Valeur minimale de Y =  -10.2487597347281
Valeur maximale de Y =  5.2422306933793
Etendue de Y =  15.4909904281074
Entrer la largeur de tranches desiree -1: 2
------------------------
Choix des bornes de tranches extremes :

1: Automatique(Borne Inf=min(Y), Borne Sup=max(Y))
2: Manuel
Selection: 2
Entrer la borne inferieure du tranchage - 1: -10.5
Entrer la borne superieure du tranchage - 1: 5.5
Le tranchage est le suivant :
 -  5  tranches de largeur  1.7
------------------------
Les bornes de tranches calculees sont :
[1] -10.5  -8.5  -6.5  -4.5  -2.5 -0.5  1.5  3.5  5.5
------------------------
```

**Tranchage de type 3.** Il est demandé à l'utilisateur de choisir le nombre $H$ de tranches qu'il désire. Le programme construit alors un découpage en $H$ tranches contenant à peu près le même nombre d'individus. (Si la taille d'échantillon n'est pas divisible par le nombre de tranches, le nombre d'individus restants est réparti aléatoirement parmi les $H$ tranches de manière à ce que les poids des tranches ne diffèrent au plus que d'un seul individu.)

```
-----------------------
Tranchage de type 3 :
choix du nombre de tranches, et tranches contenant
a peu pres le meme nombre d'individus
-----------------------
Entrer le nombre de tranches desire pour le nombre  200
d'individus - 1: 15
-----------------------
Le tranchage effectue est le suivant :
5  tranches contenant a peu pres  13  individus,
les  5 individus restants etant repartis aleatoirement parmi les tranches.
-----------------------
```

**Tranchage de type 4.** L'utilisateur doit choisir le nombre désiré d'individus par tranches. Le programme calcule alors automatiquement un découpage en tranches de poids égal ou à peu près. (Il y aura au plus un individu de différence entre les tranches, voir le tranchage de type 3 pour quelques commentaires à ce sujet.)

```
-------------------------
Tranchage de type 4 :
choix du nombre constant d'individus par tranche
-------------------------
Entrer le nombre d'individu par tranche desire pour le nombre  200
d'individus - 1: 25
Le tranchage effectue est le suivant :
8 tranches  contenant 25 individus
-----------------------------------------------------------------
```

**Tranchage de type 5.** Enfin, si l'utilisateur opte pour un tranchage automatique, le programme calcule automatiquement le nombre $H$ de tranches et construit alors des tranches de poids à peu près égal.

```
---------------------
Tranchage de type 5 : Automatique (tranches de poids constant)
---------------------
```

### 2.1.2 Sorties numériques et graphiques

Après avoir choisi le type de tranchage, la liste des valeurs propres des directions s'affiche à l'écran. Il est alors demandé à l'utilisateur de choisir le nombre de directions EDR qu'il désire retenir pour la suite. Les directions EDR apparaissent ensuite à l'écran. (On retrouve ci-dessous les estimations obtenues pour la matrice de données **mla** avec le tranchage automatique.)

```
-----------------------------------------------------------------
Nombre de variables explicatives =  5
Les valeurs propres des directions EDR sont :
[1] 0.271742991 0.040834031 0.034069730 0.010141107 0.006465826
```

```
Entrer le nombre de directions EDR :
1: 1
------------------------------------------------------------
Les directions EDR calculees sont :
      "X1"       "X2"       "X3"       "X4"       "X5"
 -0.6179961 -0.4116174 -0.6661495 -0.4160846 0.02980023
============================================================
```

Avec l'échantillon **m1b**, les valeurs propres calculées sont du même ordre de grandeur et la direction EDR obtenue est :

```
------------------------------------------------------------
Les directions EDR calculees sont :
      "X1"       "X2"       "X3"       "X4"       "X5"
 0.4290617 0.5637701 0.5227722 0.4333068 0.01147877
============================================================
```

En fonction du nombre de direction(s) retenue(s), des sorties graphiques sont disponibles.

**Cas où une seule direction EDR a été retenue.** Le programme permet d'estimer la fonction de lien. Pour cela, deux méthodes sont disponibles : la méthode des noyaux avec un estimateur du type Nadaraya-Watson ou la méthode des splines de lissage.

```
------------------------------------------------------
Choix de la methode pour l'estimation de la
                        fonction de lien :

1: Methode des noyaux
2: Splines de lissage
Selection: 2
------------------------------------------------------
```

Sur la fenêtre graphique apparaît alors le nuage des points $\{(y_i, x_i'\hat{\beta}),\ i = 1, \ldots, n\}$ ainsi que la courbe de la régression nonparamétrique des $y_i$ sur les $x_i'\beta$. Dans le cas de l'estimation par la méthode des noyaux, afin de contrôler le choix de la fenêtre optimale, le graphe du critère de validation croisée (CV) en fonction de la largeur de fenêtre (bandwidth) s'affiche dans la partie de droite de la fenêtre graphique.
La figure 1 (resp. 2) illustre l'estimation de la fonction de lien $f$ par la méthode des splines de lissage avec l'échantillon **m1a** (resp. par la méthode des noyaux avec l'échantillon **m1b**).
En sortie du programme, on obtient la liste suivante : les directions EDR retenues, le vecteur des $\hat{y}_i$ (estimation des $y_i$ par une des deux méthodes nonparamétriques), ce vecteur prenant la valeur "NULL" dans le cas d'un modèle à plusieurs indices ou d'un modèle multivarié.

**Cas où plusieurs directions EDR sont retenues.** Pour la matrice de données **m3** (voir la sortie numérique ci-dessous, pour la méthode SIR classique avec un tranchage automatique), la sortie graphique est alors interactive (utilisation de la fonction graphique "brush()"). Cette fonction permet de visualiser un nuage de points en trois dimensions, de faire tourner les axes pour avoir la meilleure vue possible des données et repérer les formes particulières du nuage. Le choix des axes se fait ici parmi $y$, $x'\hat{\beta}_1$, ..., $x'\hat{\beta}_K$. Il n'est malheureusement pas possible pour le moment d'obtenir des sorties sur papier de cette fonction graphique.

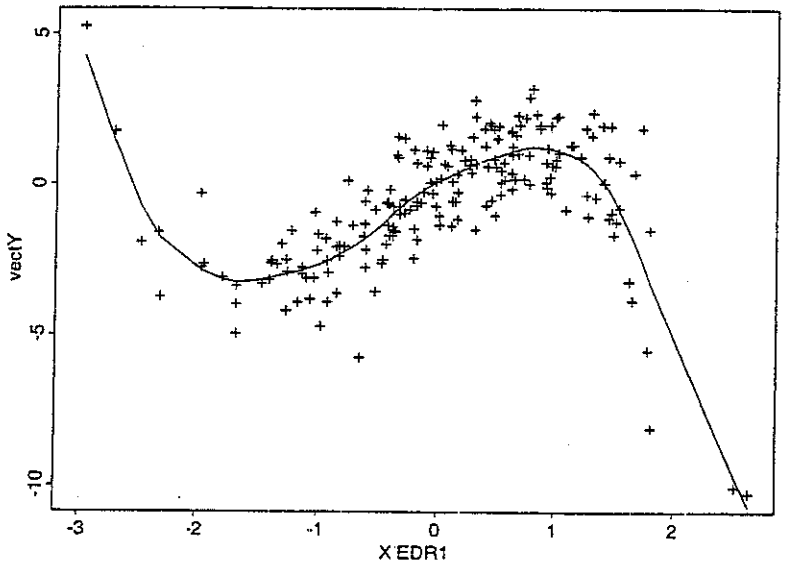


FIG. 1 - *Estimation par la méthode des splines de lissage de la fonction de lien pour la matrice des données* **m1a** *avec la direction EDR estimée par la méthode SIR classique avec un découpage en tranches automatique.*

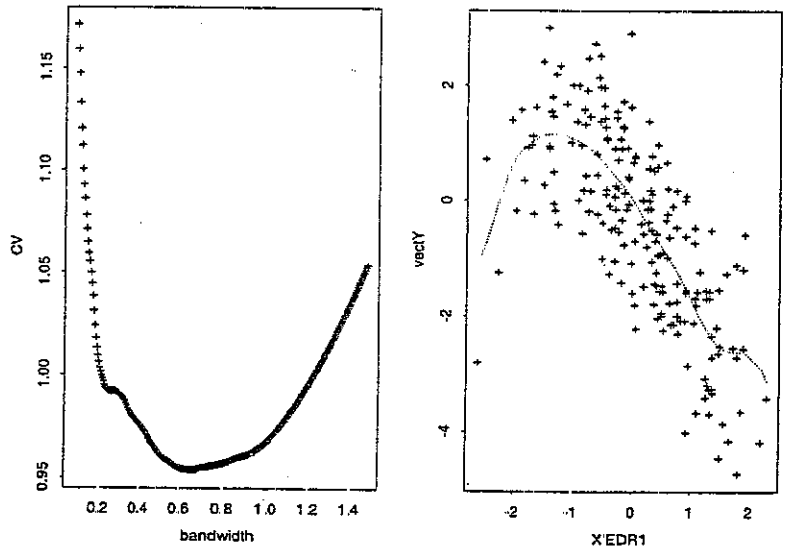


FIG. 2 - *Estimation par la méthode des noyaux de la fonction de lien pour la matrice des données* **m1b** *avec la direction EDR estimée par la méthode SIR classique avec un découpage en tranches automatique.*

```
------------------------------------------------------------
Nombre de variables explicatives =  5
Les valeurs propres des directions EDR sont :
[1] 0.874813127 0.131886825 0 011841643 0 009967834 0.006610356
Entrer le nombre de directions EDR :
1: 2
------------------------------------------------------------
Les directions EDR calculees sont :
            edr1        edr2
"X1"    0.4408820  -0.2503422
"X2"    0.4387593  -0.4391758
"X3"    0.4633077  -0.5240221
"X4"    0.4895389   0.5009442
"X5"    0.4967377   0.4540365
============================================================
```

## 2.2 Méthodes alternatives

Elles correspondent aux choix 2 à 5 dans le menu 1 donné à la page 4. Nous avons vu que ces méthodes alternatives sont intéressantes à utiliser lorsque la taille des échantillons à traiter est relativement faible ($n \leq 50$). Nous travaillons ici avec la matrice de données **m2** pour laquelle $n = 50$ afin d'illustrer ces quatre méthodes alternatives.
Pour les méthodes du Pooled Slicing (Aragon et Saracco (1997)) et de Hsing et Carroll (1992), il n'est demandé à l'utilisateur que de choisir le nombre de directions EDR qu'il désire retenir et ces directions s'affichent alors à l'écran. Des sorties graphiques identiques à celles présentées pour la méthode 1 sont disponibles. La figure 3 donne par exemple une estimation de la fonction de lien par la méthode des splines de lissage avec une seule direction EDR retenue et estimée par la méthode du Pooled Slicing. Nous donnons ci-dessous les sorties (affichés à l'écran) des valeurs propres et des directions EDR estimées par chacune de ces méthodes.

```
============================
Methode : Pooled Slicing
============================
Entrer le nombre minimum d'individus par tranches :
1: 4
Entrer le nombre minimum de tranches :
1: 4
------------------------------------------------------------
Nombre de variables explicatives =  5
Les valeurs propres des directions EDR sont :
[1] 0.77383860 0.12911400 0 11107708 0.07712931 0.03947388
Entrer le nombre de directions EDR :
1: 1
------------------------------------------------------------
Les directions EDR calculees sont :
    "X1"       "X2"       "X3"       "X4"       "X5"
-0.4177116 -0.3714004 -0.5404946 -0.4143466 0.0437689
============================================================

======================================================
Methode : SIR Hsing et Carroll (tranches distinctes)
======================================================
------------------------------------------------------------
Nombre de variables explicatives =  5
Les valeurs propres des directions EDR sont :
[1]  0.7998950  0.2986170  -0.0264591 -0.0814835 -0.1574808
Entrer le nombre de directions EDR :
1: 1
------------------------------------------------------------
```

```
Les directions EDR calculees sont :
    "X1"       "X2"       "X3"       "X4"        "X5"
 -0.3645908 -0.4044934 -0.5465454 -0.4626052 -0.04354753
============================================================

============================================================
Methode : SIR Hsing et Carroll (tranches glissantes)
============================================================
------------------------------------------------------------
Nombre de variables explicatives =  5
Les valeurs propres des directions EDR sont :
[1]  0.81549960  0.27141492  0.08397168 -0.09775105 -0.218893482
Entrer le nombre de directions EDR :
1: 1
------------------------------------------------------------
Les directions EDR calculees sont :
    "X1"       "X2"       "X3"       "X4"        "X5"
 -0.382748 -0.4103982 -0.5666585 -0.405805 -0.395872
============================================================
```

**Remarque.** Les estimateurs de la matrice de covariances, utilisés dans les méthodes de Hsing et Carroll (1992) (tranches distinctes ou glissantes) convergent en théorie vers une matrice semi-définie positive. Cependant, il semble que les conditions nécessaires à ce résultat théorique, conditions qui ne peuvent être vérifiées en pratique, soient relativement fortes. Ainsi, les matrices de covariances estimées par ces méthodes peuvent ne pas être semi-définies positives, ce qui explique la présence de valeurs propres négatives, voir Aragon et Saracco (1997) pour une étude (sur des simulations) de la qualité de l'estimation de l'espace EDR obtenue avec ces deux estimateurs.

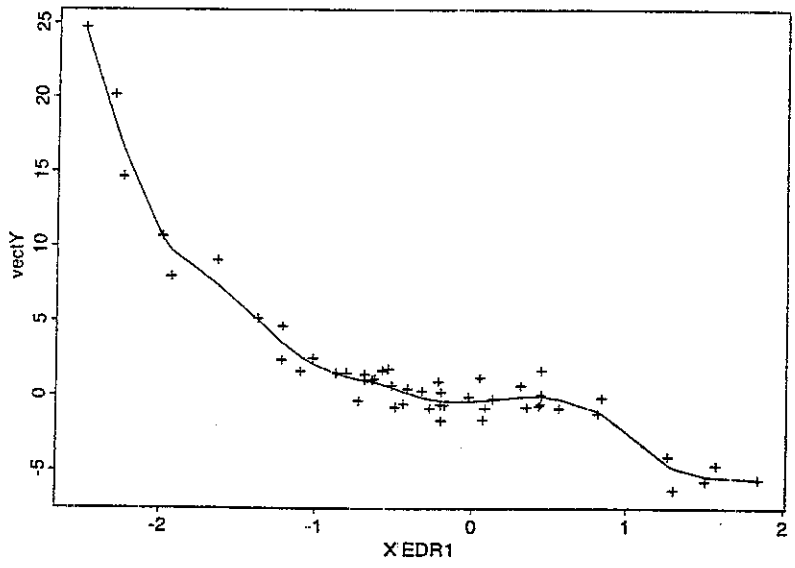


FIG. 3 - *Estimation par la méthode des splines de lissage de la fonction de lien pour la matrice des données* **m2** *avec la direction EDR estimée par la méthode SIR classique avec un découpage en tranches automatique.*

Pour la méthode de régression inverse par lissage avec la méthode des noyaux (méthode 5), l'utilisateur doit en plus choisir entre un estimateur de type "somme" ou un estimateur

de type "intégrale" pour estimer les composantes de la matrice de covariances à diagonaliser (voir Saracco, Larramendy et Aragon (1998) pour une description rapide de ces estimateurs). Le déroulement du programme est ensuite identique à celui des méthodes précédentes.

```
=============================================
Methode : Kernel Smoothing Inverse Regression
=============================================
Choix du type d'estimateur :

1: Estimateur de type somme
2: Estimateur de type integrale
Selection: 1
-------------------------------------------------------------
Nombre de variables explicatives =  5
Les valeurs propres des directions EDR sont :
[1] 0.61944381 0.13004612 0.06394399 0.05993324 0.02452732
Entrer le nombre de directions EDR :
1: 1
-------------------------------------------------------------
Les directions EDR calculees sont :
      "X1"        "X2"        "X3"        "X4"        "X5"
 -0.46134878 -0.43052669 -0.46633636 -0.4897622 0.02775195
=============================================================
```

# 3 Méthodes SIR multivariées

Elles correspondent au choix d'une des méthodes dans le menu 2 donné à la page 5 On s'intéresse ici à la matrice des données **m4** provenant d'un modèle dans lequel nous avons choisi 3 variables dépendantes.
Quatre méthodes de SIR multivariée sont disponibles. Il n'y a pas de sorties graphiques préprogrammées pour ces méthodes. En sortie, on récupère une liste contenant en particulier les directions EDR retenues pour les méthodes du Complete Slicing, du Marginal Slicing et du Pooled Marginal Slicing, et les directions EDR et MP pour la méthode d'Alternating SIR.

## 3.1 Complete Slicing

Si l'utilisateur opte pour la méthode du Complete Slicing, il lui est alors demandé de choisir les nombres désirés de tranches par composante de $y$. S'il préfère un choix automatique, le programme gère le découpage de chaque composante de la variable dépendante multidimensionnelle. Dans la sortie ci-dessous, c'est l'utilisateur qui a fourni le nombre de tranches par composante. De même que pour les méthodes univariées, l'utilisateur doit ensuite choisir le nombre de directions EDR à retenir, ces dernières s'affichent alors à l'écran.

```
==============================
Methode : Complete Slicing
==============================
Choix du nombre de tranches par composantes :
1: Automatique
2: Manuel
Selection: 2
----------------------
```

```
Entrer votre choix du nombre de tranches par composantes.
Nombre de tranches pour la composante  1  :
1: 3
Nombre de tranches pour la composante  2  :
1: 2
Nombre de tranches pour la composante  3  :
1: 2
----------------------
Les valeurs propres des directions EDR sont :
[1] 0.449200650 0.027774211 0.017908379 0.007980775 0.004979282
Entrer le nombre de directions EDR :
1: 1
------------------------------------------------------------
Les directions EDR calculees sont :
     "X1"       "X2"       "X3"       "X4"       "X5"
 0.2133111 0.2245927 0.1710295 0.1653583 0.02446524
============================================================
```

## 3.2 Marginal Slicing

Si la méthode du Marginal Slicing a été sélectionnée, un menu apparaît à l'écran et propose à l'utilisateur de choisir parmi trois méthodes permettant de réduire la dimension des variables dépendantes :

- une Analyse en Composantes Principales centrée réduite de ces variables, suivie ou bien (*i*) d'un tranchage automatique du même type que celui utilisé dans la méthode SIR classique lorsqu'une seule composante principale a été retenue, ou bien (*ii*) d'un tranchage automatique du même type que celui utilisé dans la méthode Complete Slicing quand au moins deux composantes principales sont retenues ;
- une méthode de classification non hiérarchique (méthode des k-means, voir par exemple Hartigan et Wong (1979)) pour laquelle l'utilisateur doit fixer le nombre de classes désirées (ces classes jouent en fait le rôle des tranches) ;
- une méthode de classification hiérarchique (voir par exemple Saporta (1990)) pour laquelle l'utilisateur doit aussi préciser le nombre de classes désirées.

L'utilisateur n'a ensuite qu'à choisir le nombre de directions EDR à retenir, et ces dernières sont alors affichées à l'écran.

```
============================
Methode : Marginal Slicing.
============================
1: ACP centree reduite des variables a expliquer
2: Classification non hierarchique (methode k-means).
3: Classification hierarchique.
Selection: 1

Inertie totale:  3
Pourcentages de variance expliquee par les composantes principales :
       f1       f2       f3
 59.42485 33.32929 7.245857
Entrer le nombre de composantes a retenir :
1: 2
Choix du nombre de tranches par composantes :

1: Automatique
2: Manuel
Selection: 1
Les valeurs propres des directions EDR sont :
[1] 0.449051953 0.035099214 0.019792912 0.005574957 0.004472711
```

```
Entrer le nombre de directions EDR :
1: 1
------------------------------------------------------------
Les directions EDR calculees sont :
    "X1"      "X2"       "X3"       "X4"        "X5"
 0.2013437 0.207196 0.1918523 0.1636334 0.02477621
============================================================
```

## 3.3 Pooled Marginal Slicing

Si l'utilisateur choisit la méthode du Pooled Marginal Slicing, un menu s'affiche et demande de choisir le type de pondération (poids égaux, ou poids proportionnels à la plus grande valeur propre). Ce choix ayant été effectué, le tranchage de chaque composante de $y$ se fait automatiquement, et l'utilisateur n'a plus qu'à préciser le nombre de directions EDR désirées qui seront alors affichées à l'écran.

```
====================================
Methode : Pooled Marginal Slicing
====================================

1: Poids egaux
2: Poids proportionnels a la plus grande valeur propre
Selection: 1

Tranchage de type 5 : Automatique (tranches de poids constants)
Tranchage de type 5 : Automatique (tranches de poids constants)
Tranchage de type 5 : Automatique (tranches de poids constants)

Les valeurs propres des directions EDR sont :
             [,1]
[1,]  0.2775690144
[2,]  0.0542072285
[3,]  0.0123884439
[4,]  0.0101352941
[5,]  0.0003410243
Entrer le nombre de directions EDR :
1: 1
------------------------------------------------------------
Les directions EDR calculees sont :
    "X1"      "X2"       "X3"       "X4"        "X5"
 0.2427267 0.1791123 0.1997344 0.1712326 0.01355517
============================================================
```

## 3.4 Alternating SIR

La méthode Alternating SIR est une méthode interactive par excellence. L'étape 0 de recherche des directions MP est une analyse des corrélations canoniques entres les variables dépendantes et les variables explicatives.

A chaque étape comprenant l'estimation des directions MP (Most Predictable) et l'estimation des directions EDR, l'utilisateur doit préciser le nombre de directions à retenir pour les calculs de l'étape suivante. De plus, il doit dire s'il désire ou non faire une itération supplémentaire. Si une seule direction (MP ou EDR) est retenue, le programme effectue un tranchage automatique de type 5 (voir méthode SIR classique univariée). Si plusieurs directions sont retenues, le programme effectue encore un tranchage automatique et du même type que le tranchage de la méthode du Complete Slicing. On obtient en sortie les directions EDR et MP estimées retenues à la dernière itération.

```
==============================
Methode : Alternating SIR
==============================
------------
Etape 0 :
------------
**********************************
** Canonical Correlation Analysis **
**********************************
Les correlations canoniques sont :
[1] 0.98071946 0.19142692 0.08519238
Entrer le nombre de dimensions a garder :
1: 1

********************************
** Recherche des directions EDR **
********************************
==============================
Methode : SIR classique
==============================
Voulez-vous choisir le type de tranchage ?
   1=Automatique / 2=Choix personnel
1: 1
Tranchage de type 5 : Automatique (tranches de poids constants)
Valeurs propres des directions EDR :
[1] 0.307466774 0.019755986 0.012970958 0.010279250 0.001083786
Entrer le nombre de directions EDR a retenir :
1: 1
Directions EDR retenues :
    "var4"   "var5"     "var6"    "var7"    "var8"
 0.2229941 0.183826 0.1929685 0.1846985 0.0107868
==========================================
1: Iteration supplementaire.
2: Arret
Selection: 1
==========================================
------------
Etape  1  :
------------
********************************
** Recherche des directions MP **
********************************
==============================
Methode : SIR classique
==============================
Voulez-vous choisir le type de tranchage ?
   1=Automatique / 2=Choix personnel
1: 1
Tranchage de type 5 : Automatique (tranches de poids constants)
Valeurs propres des directions MP :
[1] 0.3088026 0.2228016 0.1097790
Entrer le nombre de directions MP a retenir :
1: 1
Directions MP retenues :
     "Y1"       "Y2"        "Y3"
 -0.2136945 0.0006439235 0.0002790695

********************************
** Recherche des directions EDR **
********************************
==============================
Methode : SIR classique
==============================
Voulez-vous choisir le type de tranchage ?
   1=Automatique / 2=Choix personnel
1: 1
Tranchage de type 5 : Automatique (tranches de poids constants)
Valeurs propres des directions EDR :
```

```
[1] 0.301794198 0.017313709 0.013036378 0.006796130 0.003098046
Entrer le nombre de directions EDR a retenir :
1: 1
Directions EDR retenues :
     "X1"       "X2"       "X3"       "X4"       "X5"
 0.2399322 0.1638369 0.2171287 0.1560878 0.02473748
==========================================
1: Iteration supplementaire
2: Arret
Selection: 2
==========================================

==================================================
Directions EDR finales retenues :
     "X1"       "X2"       "X3"       "X4"       "X5"
 0.2399322 0.1638369 0.2171287 0.1560878 0.02473748
Directions MP finales retenues :
     "Y1"        "Y2"        "Y3"
 -0.2136945 0.0006439235 0.0002790695
==================================================
```

# 4 Quelques remarques sur l'implémentation

• Nous n'avons utilisé ici qu'un seul programme principal gérant les différentes méthodes SIR univariées ou multivariées à partir d'une matrice de données. Il est évidemment possible de lancer indépendamment les procédures correspondant à chacune de ces méthodes. Dans ce cas, il convient de noter que la matrice des données ne doit contenir que les variables entrant dans l'étude. Ces dernières sont ordonnées de la manière suivante : variables dépendantes, puis variables explicatives. Il faut de plus préciser pour les méthodes multivariées le nombre de variables dépendantes.

Le document Saracco (1996 b) donne les listages commentés du programme principal ainsi que tous les listings commentés des différentes procédures et fonctions appelées dans ce programme principal. Ainsi, pour une bonne utilisation d'une fonction ou procédure particulière (i.e. en dehors du programme principal), il est intéressant et indispensable de regarder les commentaires présents en début de listing. Le tableau des fonctions et procédures Splus developpées pour les méthodes SIR est donné à la section suivante.

• Le programme principal et les différentes procédures et fonctions appelées sont disponibles auprès de Jérôme Saracco (Email : saracco@stat.math.univ-montp2.fr).

• Certaines des procédures de régression nonparamétrique par la méthode des noyaux, utilisées dans certains des programmes, ont été développées par Haerdle (1991). C'est par exemple le cas de la procédure "G.WarpingReg" qui est utilisée dans les procédures "GraphiqueSIR.prog" ou "KerSmoIR.prog".

• Une implémentation des méthodes SIR univariées et multivariées a été développée en GAUSS, voir Aragon (1997).

# 5 Tableau des fonctions et procédures Splus pour SIR

Dans Saracco (1996 b), nous donnons les listings **commentés** de toutes les procédures et fonctions nécessaires aux diverses méthodes de Régression Inverse par Tranchage ou Lissage, univariées et multivariées, ainsi que quelques programmes de simulation de matrices de données.

Un programme proposant une méthode d'estimation pour les modèles de sélection est aussi fourni : estimation des vecteurs de paramètres des équations d'observation et de sélection, estimation nonparamétrique des fonctions de lien d'observation et de sélection. La théorie de cette méthode peut être trouvée dans Saracco (1997 b). Il s'agit d'une méthode en 2 étapes : une étape de type SIR puis une étape de type Analyse Canonique.

Les tableaux 1 et 2 ci-après donnent la liste de ces procédures Splus.

*Tableau 1 : Tableau des procédures et fonctions Splus présentées dans Saracco (1996 b)*

| Nom du programme | Description | Page |
|---|---|---|
| | PROGRAMME PRINCIPAL | |
| ProgPrincipal.prog | Programme principal gérant toutes les méthodes de type SIR | 4 |
| donnees.prog | Fonction permettant de créer la matrice des données à analyser | 7 |
| | MÉTHODES SIR UNIVARIÉES | |
| SIRglobal.prog | Programme gérant toutes les méthodes SIR et méthodes alternatives dans le cadre unidimensionnel | 9 |
| SIR.prog | Méthode SIR "classique" | 11 |
| Tranchage.prog | Fonction gérant tous les types de tranchage | 13 |
| tranche1.prog | Tranchage de type 1 | 14 |
| tranche2.prog | Tranchage de type 2 | 18 |
| tranche3.prog | Tranchage de type 3 | 22 |
| tranche4.prog | Tranchage de type 4 | 25 |
| tranche5.prog | Tranchage de type 5 (automatique) | 28 |
| ConstuctionGeneraleV.prog | Procédure de calcul de $\hat{\Gamma}_I$ | 30 |
| | MÉTHODES ALTERNATIVES UNIVARIÉES | |
| PSIR.prog | Méthode alternative du Pooled Slicing | 31 |
| SIRHCdist.prog | Méthode SIR alternative de Hsing et Carroll avec des tranches distinctes | 33 |
| constV.HC1.prog | Procédure de calcul de $\hat{\Gamma}_I$ dans le cadre de la méthode de Hsing et Carroll avec des tranches distinctes | 34 |
| SIRHCglis.prog | Méthode SIR alternative de Hsing et Carroll avec des tranches glissantes | 35 |
| constV.HC2.prog | Procédure de calcul de $\hat{\Gamma}_I$ dans le cadre de la méthode de Hsing et Carroll avec des tranches glissantes | 36 |
| KerSmoIR.prog | Méthodes de régression inverse par lissage (méthode des noyaux) | 37 |
| GraphiqueSIR.prog | Procédure permettant d'obtenir des graphiques dans le cadre univarié | 40 |

*Tableau 2 : Suite du Tableau des procédures et des fonctions Splus présentées dans Saracco (1996 b)*

| Nom du programme | Description | Page |
|---|---|---|
| | MÉTHODES SIR MULTIVARIÉES | |
| CompleteSlicing.prog | Méthode du Complete Slicing | 42 |
| CompleteTr.prog | Procédure de tranchage dans le cadre du Complete Slicing | 44 |
| MarginalSlicing.prog | Méthode du Marginal Slicing | 46 |
| MarginalTr.prog | Procédure de tranchage dans le cadre du Marginal Slicing | 48 |
| PooledMarginalSlicing.prog | Méthode du Pooled Marginal Slicing | 49 |
| AlternatingSIR.prog | Méthode de l'Alternating SIR | 50 |
| | ESTIMATION DE MODÈLES SEMIPARAMÈTRIQUE DE SÉLECTION | |
| SIRmodsel.prog | Méthode d'estimation dans les modèles de sélection | 53 |
| | FONCTIONS LOCALES | |
| ACP.prog | Analyse en Composantes Principales | 58 |
| Centrage.prog | Centrage d'une matrice | 59 |
| Reduction.prog | Calcul de la matrice de covariances à la puissance (-1/2) | 59 |
| Standardisation.prog | Standardisation d'une matrice | 59 |
| numobs.prog | Ajout d'une colonne contenant le numéro des observations | 59 |
| ordremat.prog | Tri d'une matrice selon une de ses colonnes | 60 |
| | SIMULATIONS DE VARIABLES EXPLICATIVES | |
| rmulnorm | Simulation d'une loi normale multivariée | 60 |
| GenereXellipt1.prog | Simulation d'une loi normale contaminée dans la variance | 60 |
| GenereXellipt2.prog | Simulation d'une distribution de Pearson de type II | 61 |
| GenereXellipt3.prog | Simulation d'une distribution de Pearson de type VII | 62 |
| | SIMULATION DE DONNÉES | |
| SimulationMatYX.prog | Simulation d'un modèle semiparamétrique (de type S.I.R) | 63 |
| SimulModSel.prog | Simulation d'un modèle semiparamétrique de sélection | 64 |

# 6 Références bibliographiques

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

ARAGON, Y., LI, K. C. and THOMAS-AGNAN, CH. (1995). Modelling income distributions using Multivariate Sliced Inverse Regression. *Rapport Technique du GREMAQ de l'Université des Sciences Sociales de Toulouse*.

ARAGON, Y., BARTHE, PH., CASSADOU, S. and THOMAS-AGNAN, CH. (1996). Analysing ambulatory blood pressure monitoring data with Multivariarate Sliced Inverse Regression. *Rapport Technique du GREMAQ de l'Université des Sciences Sociales de Toulouse*.

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

DUAN, N. and LI, K. C. (1991). Slicing regression: a link-free regression method. *The Annals of Statistics* **19** 505-530.

EUBANK, R. (1988). *Spline smoothing and nonparametric regression*. New York: Marcel Dekker Inc.

FERRÉ, L. (1996). Choix de dimension en régression inverse par tranches (SIR). *C. R. Acad. Sci. Paris*, **t. 323**, *série I*, *n° 4*, 403-406.

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

HAERDLE, W. (1991). *Smoothing techniques with implementation in S*. Springer-Verlag New-York Inc.

HARTIGAN, J. A. and WONG, M. A. (1979). A k-means clustering algorithm. *Applied Statistics* **28** 100-108.

HSING, T. and CARROLL, R. J. (1992). An asympotic theory for Sliced Inverse regression. *The Annals of Statistics* **20** 1040-1061.

LI, K.C. (1991). Sliced inverse regression for dimension reduction, with discussions. *Journal of the American Statistical Association* **86** 316-342.

LI, K. C., ARAGON, Y. and THOMAS-AGNAN, CH. (1997). Analysis of Multivariate outcome data: SIR and a nonlinear theory of Hotelling's most predictable variates. *Soumis à Journal of the American Statistical Association*.

SAPORTA, G. (1990). *Probabilités, Analyse des Données et Statistique*. Editions Technip.

SARACCO, J. (1996 a). Contributions à la Régression Inverse par Tranchage (Sliced Inverse Regression - SIR). *Thèse de troisième cycle - Université Paul Sabatier (Toulouse III) France*.

SARACCO, J. (1996 b). Méthodes de Régression Inverse par Tranchage (S.I.R), méthodes alternatives et reliées : implémentation en SPLUS (programmes et commentaires). *Rapport Technique du GREMAQ de l'Université des Sciences Sociales de Toulouse*.

SARACCO, J. (1997 a). An asymptotic theory for Sliced Inverse Regression (SIR). *Communications in statistics - Theory and methods* **26** 2141-2171.

SARACCO, J. (1997 b). Distribution-free and link-free fast estimation for Sample Selection Models. *Cahier du GREMAQ n° 97.10453, Université des Sciences Sociales de Toulouse.*

SARACCO, J., LARRAMENDY, I. et ARAGON, Y. (1998). La régression inverse par tranches ou méthode SIR : présentation générale. *La revue de Modulad (même numéro).*