Outils statistiques pour l'analyse de sensibilité : Analyse de la variance et quasi-régression

Anestis Antoniadis (UJF)
e-mail:antonia@imag.fr

Toulouse, 2 et 3 Février 2006

Plan de l'exposé

- Généralités sur l'analyse de sensibilité globale et indices de sensibilité.
- L'analyse de la variance fonctionnelle comme outil pour l' ASG.
- Régression et quasi-régression.
- Sélection des entrées, pénalisation et seuillage.
- Extensions possibles et problèmes ouverts.

Introduction

Nous considérerons un modèle (code numérique) représenté de manière générique par une fonction f définie sur un domaine de $\mathbb{R}^p$ et à valeurs dans $\mathbb{R}^m$:

$$\mathbf{Y} = \mathbf{Y}(\mathbf{X}) = \mathbf{f}(\mathbf{X}),$$

avec $\mathbf{X} \in \mathbb{R}^p$ les “entrées” et $\mathbf{Y} \in \mathbb{R}^m$ les “sorties”.

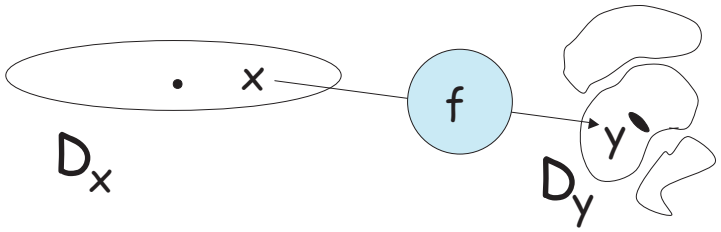
L’*analyse de sensibilité globale* (ASG) tente de déterminer l’importance et le poids des incertitudes des composantes de $\mathbf{X}$ sur la variabilité de la réponse $\mathbf{Y}(\mathbf{X})$.

Objectifs

On cherche généralement à déterminer :

- les paramètres et les entrées qui contribuent le plus à la variabilité des sorties
- Les paramètres qui ne sont pas significatifs
- si (et lesquels) certains facteurs interagissent entre eux.

Analyse globale



- Des valeurs plausibles de $\mathbf{x}$ dans D_x génèrent un domaine D_y de $\mathbf{y}$.
- Des petites perturbations $\bullet$ de $\mathbf{x}$ donnent des perturbations $\bullet$ de $\mathbf{y}$

GLOBALE : $D_x \rightarrow D_y$ **LOCALE :** $\bullet \rightarrow \bullet$

- Une densité de probabilité f_X modélise l'incertitude sur les entrées et l'action du modèle définit l'incertitude sur les sorties par une densité f_Y.

Pour simplifier l'exposé on considère dans la suite des sorties uni-dimensionnelles ($m = 1$).

Quantifier l'incertitude

Il s'agit de localiser au mieux les valeurs de Y en caractérisant sa loi (la loi prédictive). Il est habituel d'utiliser sa moyenne

$$\mu_Y = \mathbb{E}(Y) = \int_{D_y} y f_Y(y) dy$$

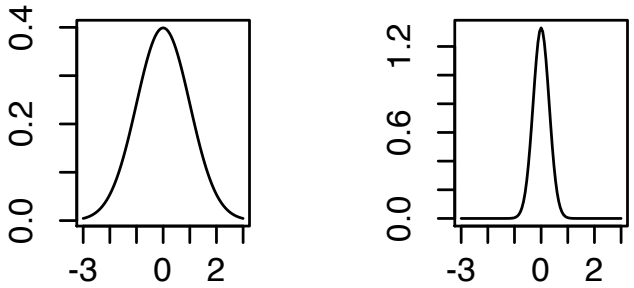
et sa variance

$$\sigma_Y^2 = \mathsf{Var}(Y) = \int_{D_y} (y - \mu_Y)^2 f_Y(y) dy$$

ou, encore mieux, sa fonction de répartition $F_Y(y) = \mathbb{P}(Y \leq y)$. Ces quantités sont en général estimées par leur analogues empiriques.

Incertitudes dues aux composantes des entrées

Pour un sous ensemble $s \subset \{1, 2, \ldots, p\}$ considérons la partition du vecteur des entrées $\mathbf{X} = \mathbf{X}^s \cup \mathbf{X}^{-s}$, où $\mathbf{X}^s$ désigne le vecteur dont les composantes sont indexées par s et $\mathbf{X}^{-s}$ sont complémentaire.



Evaluer l'incertitude due aux entrées $\mathbf{X}^s$ consiste alors à étudier comment la connaissance de $\mathbf{X}^s$ réduit l'incertitude sur la variable de sortie Y.

(suite)

Losque $\mathbf{X}^s$ est fixé à une valeur donnée $\mathbf{x}^s$, ce sont les entrées de $\mathbf{X}^{-s}$ qui causent l'incertitude sur Y. Cette incertitude locale est donc caractérisée par la densité conditionnelle $f^{\mathbf{x}^s}_{Y|X^s}(y)$ et l'incertitude globale, caractérisée par la densité de Y est alors vue comme une moyenne des densités conditionnelles

$$f_Y(y) = \int f^{\mathbf{x}^s}_{Y|X^s}(y) f_{X^s}(\mathbf{x}^s) d\mathbf{x}^s.$$

Intuitivement, dire que $\mathbf{X}^s$ est important signifie que l'incertitude sur Y se modifie beaucoup en fonction des valeurs de $\mathbf{x}^s$. Par contre si cette dernière reste constante cela signifie que Y et $\mathbf{X}^s$ sont indépendantes.

Variance prédictive

Considérons maintenant la qualité des prédictions lorsque l'on ne se sert que de $\mathbf{X}^s$.

Soit $\hat{Y}(\mathbf{x}^s)$ la prédiction lorsque l'on ne se sert que de $\mathbf{x}^s$. Comparons la à $\hat{Y}(\mathbf{x}) = Y(\mathbf{x})$, prédiction obtenue à partir du modèle complet, avec une fonction de perte quadratique :

$$\mathcal{L} = \mathbb{E}_X \left([\hat{Y}(\mathbf{x}) - \hat{Y}(\mathbf{x}^s)]^2 \right) = \int \left[\hat{Y}(\mathbf{x}) - \hat{Y}(\mathbf{x}^s) \right]^2 f_X(\mathbf{x}) d\mathbf{x}.$$

Quel est alors un bon choix de $\hat{Y}(\mathbf{x}^s)$?

Solution

On peut montrer que parmi toutes les fonctions $m(\mathbf{x}^s)$ de $\mathbf{x}^s$ de carré intégrable, la fonction qui minimise

$$\mathbb{E}\{(Y - m(\mathbf{X}^s)\}$$

est l'espérance conditionnelle de Y à $\mathbf{X}^s$, c'est à dire, que dans notre cas la fonction $\hat{y}(\mathbf{x}^s)$ qui minimise $\mathcal{L}$ est

$$\hat{y}(\mathbf{x}^s) = \mathbb{E}^{\mathbf{x}^s}(Y|\mathbf{X}^s) = \int y(\mathbf{x}) f^{\mathbf{x}^s}_{X^{-s}|X^s}(\mathbf{x}^{-s}) d\mathbf{x}^{-s}$$

où $f^{\mathbf{x}^s}_{X^{-s}|X^s}(\mathbf{x}^{-s})$ est la densité de $\mathbf{X}^{-s}$ conditionnelle à $\mathbf{X}^s = \mathbf{x}^s$.

Valeur minimale de $\mathcal{L}$

Par substitution la valeur minimale de $\mathcal{L}$ est alors

$$\begin{aligned}
\min \mathcal{L} &= \mathbb{E}\left[Y - \mathbb{E}[Y|X^s]\right]^2 \\
&= V(Y) - V(\mathbb{E}(Y|X^s)) \\
&= V(Y)\left[1 - \frac{V(\mathbb{E}(Y|X^s))}{V(Y)}\right] \\
&= V(Y)(1 - h^2)
\end{aligned}$$

et le coefficient h^2 indique donc la fraction de la variance de Y expliquée par $\mathbf{X}^s$. Ce n'est autre que le rapport des corrélations entre Y et sa prédiction $\mathbb{E}(Y|\mathbf{X}^s)$.

Analyse de sensibilité globale

Les calculs précédents expliquent la raison pour laquelle la variance et les techniques d'analyse de la variance sont des notions essentielles pour l'analyse de sensibilité globale. Les indices de sensibilité du premier ordre sont définis par

$$S_i = \frac{\mathrm{Var}(\mathbb{E}(Y|X_i))}{\mathrm{Var}(Y)}$$

alors que les indices d'ordre supérieur par

$$S_{i_1,i_2,\ldots,i_k} = \frac{\mathrm{Var}(\mathbb{E}(Y|\{X_{i_1},\ldots,X_{i_k}\}))}{\mathrm{Var}(Y)}$$

La plupart des méthodes d'estimation de ces indices de sensibilité repose sur une décomposition de la réponse $Y = f(\mathbf{X})$ de type analyse de la variance.

Les étapes d'implémentation du AS globale empirique

Etape 1 Définition du modèle, des facteurs et des réponses.

Etape 2 Assigner des lois de probabilités aux paramètres/facteurs (les entrées) et si nécessaire une structure de dépendance (i.e. covariance) entre ceux-ci. DIFFICILE.

Etape 3 Simuler des réalisations à partir des lois des entrées pour produire un ensemble de sorties à l'aide du modèle étudié.

Etape 4 Procéder à l'analyse des données ainsi obtenues en estimant les quantités voulues par leur analogues empiriques.

Choix de la méthode de simulation

Echantillonnage aléatoire simple ou stratifié (SRS). Chaque entrée est échantillonnée indépendamment un grand nombre de fois à partir des lois marginales pour créer l'ensemble des entrées pour l'analyse (ou échantillonnage du vecteur des entrées à partir de la loi conjointe). C'est une méthode coûteuse en temps de calcul surtout si il y a beaucoup de facteurs, et peut éventuellement ne pas parcourir l'ensemble des valeurs possibles des facteurs.

Echantillonnage aléatoire par hypercube latin (LHS). L'étendue de chacun des facteurs est divisée en N classes de probabilités égales, et une observation pour chacun des facteur est réalisée dans chacune des classes.

Analyse empirique

A la fin des simulations les donnés obtenues sont de la forme $(y_i, \mathbf{x}_i)$, où $\mathbf{x}_1, \ldots, \mathbf{x}_n$ sont les réalisations du vecteurs des entrées.

L'analyse comporte alors

- Analyse de la régression (sur donnés brutes ou sur les rangs)
- Inférence statistique sur les lois (Tests)
- Analyse de la variance

Des “nouvelles” méthodes d’analyse

Une grande variété de problèmes et de modèles font usage de fonctions de type “boîte noire”, c’est à dire de fonctions de structure compliquée dépendant de plusieurs variables et ceci de manière peu compréhensible. Etant donné une telle fonction, il est alors important de pouvoir examiner quelles variables, s’il y en a, dominent son comportement.

La classes des méthodes que nous allons maintenant décrire on été développées par Sobol, Saltelli et ses collaborateurs et reposent sur la notion d’analyse de la variance fonctionnelle (FANOVA) pour des fonctions de carré intégrable, définies sur le cube $[0,1]^p$.

ANOVA fonctionnelle

Une des applications importantes de l'ASG est de pouvoir contrôler de manière la plus précise possible la variabilité de la "sortie" au vu de l'ensemble des variables d' "entrée" tout en éliminant parmi elles celles dont l'influence est négligeable.

Il est donc important de disposer d'une approximation universelle de la surface de régression ne nécessitant pas de trop d'hypothèses sur la nature de cette dernière et pouvant être raisonnablement estimée avec un nombre limité d'observations.

Dans ce contexte l'analyse de la variance fonctionnelle joue un rôle important.

Décomposition (ANOVA) de $L^2([0,1]^p)$

Hoeffding, Antoniadis, Efron & Morris, Stone, Owen, Sobol, Roosen, Saltelli.

On décompose une fonction de carré intégrable et supposée être évaluable en tout point de $[0,1]^p$ en effets principaux et intéractions selon :

$$f(\mathbf{x}) = \sum_{u \subseteq \{1,2,\ldots,p\}} f_u(\mathbf{x}),$$

- f_u ne dépend que des composantes de $\mathbf{x}$ dont l'indice est dans u
- $\int f_u(\mathbf{x}) f_v(\mathbf{x}) d\mathbf{x} = 0$, $u \neq v$, $\quad f_\emptyset = \int f(\mathbf{x}) d\mathbf{x}$ est la moyenne globale.
- $\sigma^2(f) = \sum_{u \neq \emptyset} \int f_u(\mathbf{x})^2 d\mathbf{x}$.

On a $\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) = \sum_u \frac{1}{n}\sum_{i=1}^{n} f_u(\mathbf{x}_i)$. La décomposition fonctionnelle précédente (avec les conditions d'orthogonalité) est unique. Il est facile de montrer d'ailleurs que

$$f_u(\mathbf{x}) = \int \left(f(\mathbf{x}) - \sum_{v \subset u} f_v(\mathbf{x}) \right) d\mathbf{x}^{-u} = \int f(\mathbf{x}) d\mathbf{x}^{-u} - \sum_{v \subset u} f_v(\mathbf{x}).$$

Si l'on suppose que la distribution de $\mathbf{X}$ est uniforme sur $[0,1]^p$ (mesure de Lebesgue), on obtient, en intégrand la décomposition ANOVA :

$$f_u(\mathbf{X}) = \mathbb{E}(Y|\mathbf{X}^u) - \sum_{v \subset u} f_v(\mathbf{X})$$

Lorsque $u = \{j\}$, f_u sont les effets principaux, alors que dans le cas général f_u est une intéraction d'ordre le cardinal de u.

On retiendra la liste des propriétés les plus importantes de l'ANOVA fonctionnelle :

- **Moyennes centrées** $\int f_u(\mathbf{x}) d\mathbf{x}^u = 0$, pour $u \neq \emptyset$.
- **Orthogonalité** $\int f_u(\mathbf{x}) f_v(\mathbf{x}) d\mathbf{x} = 0$, $u \neq v$.
- **Décomposition de la variance** $\sigma^2(f) = \sum_{u \neq \emptyset} \sigma_u^2$.

Une fonction sera *additive* en u, si $f(\mathbf{x}) = \sum_{v \subset u} f_v(\mathbf{x})$. Dans ce cas $f_u(\mathbf{x}) = 0$.

Importance des variables : les indices de Sobol

Notons la variance des composantes u (avec la convention $\sigma^2_\emptyset = 0$) par $\sigma^2_u = \int f^2_u(\mathbf{x})d\mathbf{x}$.

L'orthogonalité des f_u et l'indépendance des composantes de $\mathbf{X}$ impliquent alors que la sensibilité de chacune des composantes (sensibilité d'ordre 1) est donnée par

$$S_i = \sigma^2_{\{i\}}/\sigma^2(f),$$

et

$$S_u = (\sum_{v \subseteq u} \sigma^2_v)/\sigma^2(f)$$

Les indices de sensibilité globales de Sobol (1993) sont définis par

$$\underline{\tau}^2_u = S_u \quad \text{et} \quad \bar{\tau}^2_u = (\sum_{v \cap u \neq \emptyset} \sigma^2_v)/\sigma^2(f).$$

Clairement on a

$$\underline{\tau}_u^2 \leq \bar{\tau}_u^2$$

On voit donc, qu'une grande valeur de $\underline{\tau}_u^2$ indique un sous-ensemble de variables qui agissant ensemble peuvent fortement affecter Y.

Au contraire des petites valeurs de $\bar{\tau}_u^2$ indiquent un groupe de variables qui a peu d'effet sur Y, même en présence d'autres variables.

Remarquons que l'on a :

$$\underline{\tau}_u^2 + \bar{\tau}_{-u}^2 = 1.$$

L'équation $\underline{\tau}_u^2 = (\sum_{v \subseteq u} \sigma_v^2)/\sigma^2(f)$ exprime les 2^p valeurs de $\underline{\tau}_u^2$ comme combinaison linéaire des 2^p valeurs de σ_u^2.

La relation inverse est donnée par

$$\sigma_u^2/\sigma^2(f) = \sum_{v \subseteq u} (-1)^{|u-v|} \underline{\tau}_v^2$$

Pour calculer σ_u^2 à partir de $\bar{\tau}_u^2$, on peut utiliser la relation précédente et le fait que $\underline{\tau}_v^2 + \bar{\tau}_{-v}^2 = 1$.

Sans perdre de généralité, en supposant que $\sigma^2(f) = 1$, Sobol (1993) donne les identités

$$\underline{\tau}_u^2 = \int f(x^u, x^{-u}) f(x^u, z^{-u}) dx dz^{-u} - f_\emptyset^2$$

et

$$\bar{\tau}_u^2 = \frac{1}{2} \int (f(x^u, x^{-u}) - f(z^u, x^{-u}))^2 dx dz^u$$

Ces intégrales sont respectivement calculées sur les cubes unité de dimension $2p - |u|$ et $p + |u|$ et sont à la base des calculs des indices de sensibilité par la méthode de Sobol fondée sur des intégrations de type Monte Carlo, faisable dans le cas où p est relativement petit et implémentés dans le logiciel SIMLAB.

Autres méthodes

Le calcul des divers indices de sensibilité globale se ramène finalement au calcul numérique d'intégrales multidimensionnelles du type $I = \int_{]0,1]^p} g(\mathbf{x}) d\mathbf{x}$. Losrque la dimension du domaine d'intégration augmente la plupart des méthodes numériques deviennent problématiques et on a recours à des procédures de nature plus statistique. Les principales raisons sont :

- l'échantillonnage devient inévitablement "creux"
- des erreurs d'approximation proviennent du fait que certaines régions sont non ou mal échantillonnées.

Saltelli, Chan et Scott (2000) donnent un aperçu complet des méthodes existantes de calcul des indices de sensibilité et seront abordées par Saltelli dans sa conférence.

Remarque

Les principales méthodes “stochastiques” pour calculer les intégrales précedentes sont

- Monte Carlo : $n^{-1/2}$.
- Quasi-Monte Carlo : $n^{-1}(\log n)^{d-1}$, mais sans une approximation de l’erreur de l’estimateur.
- Quasi-Monte Carlo Randomisé : Une approximation de l’erreur de l’estimation fondée sur des répétitions et un taux d’approximation de l’ordre $n^{-3/2}(\log n)^{(d-1)/2}$.

Les taux sont asymptotiques et valables sous des conditions peu restrictives sur f.

Régression et Monte Carlo

Jusqu'à présent nous nous sommes intéressés à une interprétation des σ_u^2 et à leur estimation.

Souvent, une estimation des composantes de la décomposition fonctionnelle de la variance est également d'un grand intérêt (par exemple pour des objectifs d'optimisation (surfaces de réponses) ou de visualisation.

La dimension du domaine de variation des entrées étant souvent importante, et l'absence d'une forme analytique de f requièrent encore une fois des méthodes de Monte Carlo pour approcher la fonction f.

Approche statistique pour l'approximation

L'approche statistique pour l'approximation commence par

$$f(\mathbf{x}) = \sum_{j=0}^{d-1} \beta_j \psi_j(\mathbf{x}) + \eta(\mathbf{x}).$$

Les fonctions $\psi_j(\mathbf{x})$ sont des fonctions de base de $L^2([0,1]^p)$ choisies de sorte que :

- $\psi_0(\mathbf{x}) = 1, \quad \forall \mathbf{x} \in [0,1]^p$
- $\int \psi_j^2(\mathbf{x}) d\mathbf{x} = 1, \ \forall j \in \{0, \ldots, d-1\}$
- $\int \psi_j(\mathbf{x}) \psi_k(\mathbf{x}) d\mathbf{x} = 0, \quad j \neq k$

Dans l'expression précédente, les coefficients β_j sont des scalaires et $\eta(\mathbf{x})$ est une erreur d'approximation.

Dans la suite, inspirés des travaux de Stone nous prendrons pour fonctions ψ des produits tensoriels de fonctions univariées (sinusoïdes, B-splines orthogonales, polynômes orthogonaux, ondelettes).

Dans l'approche par régression ou quasi-régression, l'erreur d'approximation η est considérée être déterministe, alors que dans l'approche du Kriging issue de la géostatistique, $\eta(\mathbf{x})$ est supposée être la réalisation d'un processus stationnaire gaussien centré de noyau de corrélation paramétrisé par un paramètre de dimension finie.

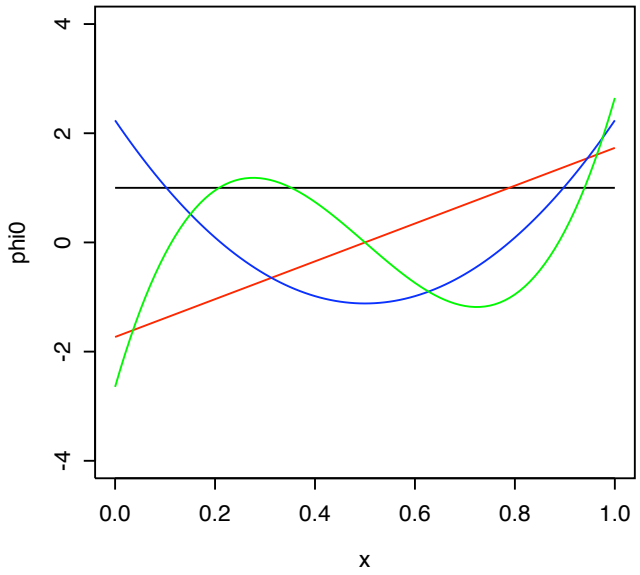
Bases orthonormées

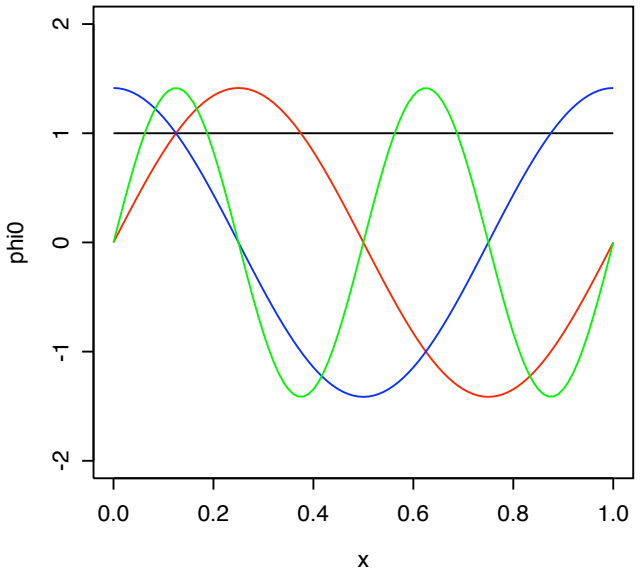
Commençons par le cas scalaire, en considérant une suite de fonctions définies sur $[0,1]$:

$$\phi_0, \phi_1, \phi_2, \ldots$$

avec

$$\begin{aligned} \phi_0(x) &= 1, \quad 0 \leq x \leq 1, \\ \int_0^1 \phi_j(x)dx &= 0, \quad j \geq 1, \\ \int_0^1 \phi_j(x)\phi_k(x)dx &= \delta_{j,k}. \end{aligned}$$


phi0
4
2
0
-2
-4
0.0
0.2
0.4
0.6
0.8
1.0
x


phi0
2
1
0
-1
-2
0.0
0.2
0.4
0.6
0.8
1.0
x

Produits tensoriels

$$\mathbf{x} = (x_1, x_2, \ldots, x_p) \in [0,1]^p$$
$$\mathbf{r} = (r(1), r(2), \ldots, r(p)) \in \{0, 1, 2, \ldots\}^p$$
$$\psi_{\mathbf{r}}(\mathbf{x}) = \prod_{j=1}^{p} \phi_{r(j)}(x_j)$$

Il est facile de voir que

$$\psi_{0,0,\ldots,0}(\mathbf{x}) = \psi_0(\mathbf{x}) = 1$$

et que

$$\int_{[0,1]^p} \psi_{\mathbf{r}}(\mathbf{x})\psi_{\mathbf{s}}(\mathbf{x})d\mathbf{x} = \delta_{\mathbf{r},\mathbf{s}}.$$

On obtient ainsi une base orthonormée de $L^2([0,1]^p)$.

Approximation

Posons : $y = f(\mathbf{x}) = \sum_{\mathbf{r} \in \mathcal{U}} \beta_r \psi_{\mathbf{r}}(\mathbf{x}) = \sum_{\mathbf{r} \in \mathcal{R}} \beta_{\mathbf{r}} \psi_{\mathbf{r}}(\mathbf{x}) + \eta(\mathbf{x})$

$\eta(\mathbf{x})$ erreur d'approximation déterministe

Estimer $\beta_{\mathbf{r}}$ à partir de valeurs $f(\mathbf{x}_i)$ où $\mathbf{x}_i \sim U(0,1)^p$ pour obtenir

$$\tilde{f}(\mathbf{x}) = \sum_{\mathbf{r} \in \mathcal{R}} \tilde{\beta}_{\mathbf{r}} \psi_{\mathbf{r}}(\mathbf{x})$$

Faire alors tout le travail avec $\tilde{f}(\mathbf{x})$

$$\beta_{\mathbf{r}} = \int_{[0,1]^d} f(\mathbf{x}) \psi_{\mathbf{r}}(\mathbf{x}) d\mathbf{x}$$

- La variance de f est donnée par

$$\sigma^2(f) = \sum_{\mathbf{r}\in\mathcal{R},\mathbf{r}\neq 0} \beta_{\mathbf{r}}^2 + \int \eta(\mathbf{x})^2 d\mathbf{x}$$

- L'importance d'un sous ensemble $\mathcal{S}$ de coefficients par

$$\sigma_{\mathcal{S}}^2 = \sum_{\mathbf{r}\in\mathcal{S}} \beta_{\mathbf{r}}^2$$

- Version normalisée

$$\sigma_{\mathcal{S}}^2/\sigma^2(f)$$

Les sous-ensembles d'intérêt diffèrent selon les applications :

- On regroupe ensemble les fonctions de base $\psi_{\mathbf{r}}$ qui ne dépendent que d'un sous-ensemble des entrées.
- On regroupe ensemble les fonctions de bases dont l'ordre en $\mathbf{r}$ est faible
- ...

Plusieurs manières de définir l'ordre.

$$\|\mathbf{r}\|_0 = \sum_{j=1}^{p} 1_{r_j>0} \text{ rang}, \quad \|\mathbf{r}\|_1 = \sum_{j=1}^{p} r_j, \text{ degré} \quad \|\mathbf{r}\|_\infty = \max_{1\leq j\leq p} r_j \text{ ordre}$$

Exemples

$\{\mathbf{r} \mid r(1) > 0\}$	dépend de x_1
$\{\mathbf{r} \mid r(1) = 0\}$	ne dépend pas de x_1
$\{\mathbf{r} \mid \|\mathbf{r}\|_0 = 1\}$	partie additive à effets simples
$\{\mathbf{r} \mid 0 < \|\mathbf{r}\|_0 \leq k\}$	intéractions jusqu'à l'ordre k
$\{\mathbf{r} \mid 0 < \|\mathbf{r}\|_1 \leq k\}$	de degré au plus k
$\{\mathbf{r} \mid r(j) = 0, j > 3\}$	fonction que de x_1, x_2, x_3

En général nous construirons $\mathcal{R}$ sous la forme

$$\mathcal{R} = \mathcal{R}_{B_0,B_1,B_\infty} = \{\mathbf{r} \mid \|\mathbf{r}\|_0 \leq B_0, \|\mathbf{r}\|_1 \leq B_1, \|\mathbf{r}\|_\infty \leq B_\infty\}$$

Nous noterons désormais $d = \text{Cardinal}(\mathcal{R})$.

Implémentation

Posons : $Z(\mathbf{x}) = (\psi_0(\mathbf{x}), \ldots, \psi_{d-1}(\mathbf{x}))^T$

Le vecteur optimal des coefficients $\boldsymbol{\beta}$ est alors

$$\begin{aligned} \boldsymbol{\beta}^* &= \text{argmin}_{\boldsymbol{\beta}} \int (f(\mathbf{x}) - Z(\mathbf{x})^T \boldsymbol{\beta})^2 d\mathbf{x} \\ &= \left(\int Z(\mathbf{x}) Z(\mathbf{x})^T d\mathbf{x} \right)^{-1} \int Z(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \end{aligned}$$

L'erreur quadratique intégrée est aussi donnée par

$$ISE = \int (f(\mathbf{x}) - Z(\mathbf{x})^T \boldsymbol{\beta})^2 d\mathbf{x}.$$

L'orthogonalité implique

$$\begin{aligned}\boldsymbol{\beta}^* &= \left(\int Z(\mathbf{x})Z(\mathbf{x})^T d\mathbf{x}\right)^{-1} \int Z(\mathbf{x})f(\mathbf{x})d\mathbf{x} \\ &= \int Z(\mathbf{x})f(\mathbf{x})d\mathbf{x}.\end{aligned}$$

Observations

$$\mathbf{x}_i \sim U([0,1]^p),\ 1 \leq i \leq n,\ i.i.d.$$

Régression

$$\hat{\boldsymbol{\beta}} = (\mathcal{Z}^T\mathcal{Z})^{-1}\mathcal{Z}^T\mathcal{Y}, \quad \mathcal{Z}_{n\times d},\ \mathcal{Y}_{n\times 1} \text{ et } \hat{f}(\mathbf{x}) = Z(\mathbf{x})^T\hat{\boldsymbol{\beta}}.$$

Quasi-régression ou régression approchée

La quasi-régression exploite le fait que d'une part pour une suite orthonormée de $L^2([0,1]^p)$ la matrice $\left(\int Z(\mathbf{x})Z(\mathbf{x})^T d\mathbf{x}\right) = I$ et que d'autre part, le plan d'expérience étant issu d'une loi uniforme sur $[0,1]^p$, la loi des grands nombre permet d'approcher la matrice empirique $\mathcal{Z}^T\mathcal{Z}$ par l'identité. On obtient ainsi pour estimation des coefficients de régression l'expression :

$$\tilde{\boldsymbol{\beta}} = \frac{1}{n}\mathcal{Z}^T\mathcal{Y}.$$

Ce type d'approximation a été utilisé par Efromovich (1992) pour le lissage par séries de fonctions orthogonales ainsi que par Owen (1992, 1998) pour l'élaboration de méthodes de quasi-MonteCarlo.

Comparaison

Nous sommes en présence de deux estimateurs

$$\hat{\boldsymbol{\beta}} = (\mathcal{Z}^T\mathcal{Z})^{-1}\mathcal{Z}^T\mathcal{Y} \quad \text{et} \quad \tilde{\boldsymbol{\beta}} = \frac{1}{n}\mathcal{Z}^T\mathcal{Y}.$$

Pour les analyser posons

$$\frac{1}{n}\mathcal{Z}^T\mathcal{Z} = I + A \quad \text{et} \quad \frac{1}{n}\mathcal{Z}^T\boldsymbol{\eta} = \boldsymbol{\delta}.$$

Notons que la matrice A et le vecteur $\boldsymbol{\delta}$ sont tous deux centrés et de termes de variance proportionnelle à $1/n$. On a alors

$$\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} = \boldsymbol{\delta} + A\boldsymbol{\beta} \quad \text{et} \quad \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (I + A)^{-1}\boldsymbol{\delta}.$$

Comparaison (suite)

Considérons le cas où d est fixé et $n \to \infty$. Dans ce cas on sait que les valeurs propres de $I + A$, disons λ_k, convergent presque sûrement vers 1 à la vitesse $\log n/n$ (Anderson (1984)) et donc les valeurs propres $\nu_k = \lambda_k - 1$ de A convergent vers 0 à la même vitesse. Pour n grand, on obtient ainsi

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (I + A)^{-1}\boldsymbol{\delta} = (I - A + A^2 - A^3 + \dots)\boldsymbol{\delta} \approx \boldsymbol{\delta} - A\boldsymbol{\delta}.$$

Les erreurs des deux estimateurs dépendent toutes les deux de $\boldsymbol{\delta}$ avec un terme en $A\boldsymbol{\beta}$ pour la quasi-régression et un terme en $A\boldsymbol{\delta}$ pour la régression. Comme $\boldsymbol{\delta}$ et que $A\boldsymbol{\beta}$ sont tous deux de l'ordre $O_p(n^{-1/2})$, les deux estimateurs ont le même taux de convergence mais comme $\boldsymbol{\beta}$ est constant, la régression est plus efficace. Par contre on peut montrer que pour d important la quasi-régression peut être plus efficace.

Si l'on pose

$$\tilde{\sigma}_d^2 = \sum_{j=1}^{d} \tilde{\beta}_j^2$$

et

$$\sigma_d^2 = \sum_{j=1}^{d} \beta_j^2$$

et si $n = \alpha d$ avec $\alpha > 0$, on peut montrer que

$$\lim_{d\to\infty, n=[\alpha d]} \mathbb{E}\left((\tilde{\sigma}_d^2 - \sigma_d^2)^2\right) = \frac{\mathbb{E}(f^2(\mathbf{X}))}{\alpha^2} + O(n^{-1/2}),$$

et donc l'erreur d'estimation des indices de sensibilité est affectée d'un biais. Cela justifie donc une correction du biais en quasi-régression.

Régression Approchée

$$\tilde{\boldsymbol{\beta}} = \frac{1}{n}\mathcal{Z}^T\mathcal{Y}$$

Complexité

	Temps	Espace
Régression	$\mathcal{O}(nd^2 + d^3)$	$\mathcal{O}(nd)$
Régression A.	$\mathcal{O}(nd)$	$\mathcal{O}(d)$

Régression approchée et Monte Carlo

On simule séquentiellement des tirages $\mathbf{x}_1, \ldots, \mathbf{x}_n$ selon la loi $U([0,1]^p)$

$$\begin{aligned} \tilde{\beta}_{\mathbf{r}}^{(n)} &= \frac{1}{n}\sum_{i=1}^{n} \psi_{\mathbf{r}}(\mathbf{x}_i) f(\mathbf{x}_i), \\ S_{\mathbf{r}}^{(n)} = \hat{\text{var}}(\tilde{\beta}_{\mathbf{r}}^{(n)}) &= \frac{1}{n(n-1)}\sum_{i=1}^{n} \left(\psi_{\mathbf{r}}(\mathbf{x}_i) f(\mathbf{x}_i) - \tilde{\beta}_{\mathbf{r}}^{(n)}\right)^2 \end{aligned}$$

Ces deux quantités peuvent être mises à jour séquentiellement de façon stable (Chan, Golub et LeVeque (1983)) :

$$\begin{aligned} \tilde{\beta}_{\mathbf{r}}^{(n)} &= \frac{1}{n}\left((n-1)\tilde{\beta}_{\mathbf{r}}^{(n-1)} + \psi_{\mathbf{r}}(\mathbf{x}_n) f(\mathbf{x}_n)\right) \\ S_{\mathbf{r}}^{(n)} = &= \frac{n-2}{n} S_{\mathbf{r}}^{(n-1)} + \frac{1}{n^2}\left(\psi_{\mathbf{r}}(\mathbf{x}_n) f(\mathbf{x}_n) - \tilde{\beta}_{\mathbf{r}}^{(n-1)}\right)^2 S_{\mathbf{r}}^{(n)} \end{aligned}$$

Estimation des indices

Pour un ensemble $\mathcal{R}$ de coefficients un estimateur sans biais de

$$S_{\mathcal{R}} = \sum_{\mathbf{r}\in\mathcal{R}} \beta_{\mathbf{r}}^2$$

est obtenu par

$$\sum_{\mathbf{r}\in\mathcal{R}} \left(\tilde{\beta}_{\mathbf{r}}^2 - S_{\mathbf{r}}^{(n)} \right)$$

puisque

$$\mathbb{E}\left(\tilde{\beta}_{\mathbf{r}}^2 - S_{\mathbf{r}}^{(n)} \right) = \beta_{\mathbf{r}}^2 - \mathrm{var}(\tilde{\beta}_{\mathbf{r}}^{(n)}) + \mathrm{var}(\tilde{\beta}_{\mathbf{r}}^{(n)}).$$

La qualité des résultats d'approximation dépend d'une part de la troncature $\mathcal{R}$ choisie et des erreurs d'estimation $(\tilde{\beta}_{\mathbf{r}}^{(n)} - \beta_{\mathbf{r}})$ pour $\mathbf{r} \in \mathcal{R}$.

Elle est mesurée par

$$ISE(n) = \int (f(\mathbf{x}) - \tilde{f}^{(n)}(\mathbf{x}))^2 d\mathbf{x}.$$

Cette dernière peut être estimée à l'aide de l'erreur de prédiction empirique $(f(\mathbf{x}_{n+1}) - \tilde{f}^{(n)}(\mathbf{x}_{n+1}))^2$ et afin de réduire la variance de ce dernier estimateur on moyenne sur les m tirages les plus récents

$$I\hat{S}E(n) = \frac{1}{m} \sum_{i=n-m+1}^{n} (f(\mathbf{x}_i) - \tilde{f}^{(i-1)}(\mathbf{x}_i))^2.$$

Le choix de m se fait par un équilibre entre biais^2 et variance de $I\hat{S}E(n)$ ce qui donne $m \sim n^{2/3}$.

Qualité de l'approximation

Comme mesure de diagnostic de la qualité de l'approximation on utilise alors comme mesure de manque d'adéquation le LOF défini par

$$LOF(n) = ISE(n)/\sigma^2(f), \quad \hat{LOF}(n) = \hat{ISE}(n)/\hat{\sigma^2(f)}$$

Le LOF décrit la fraction de la variance de f non expliquée par la régression approchée. Si LOF est grand et que $\sum_r S_{\mathbf{r}}^{(n)}$ est petit, on a besoin d'une troncature moins sévère.

Régularisation

Nous avons vu que la qualité de l'approximation pour la quasi-régression dépend fortement de la norme du vecteur $\boldsymbol{\beta}$ des coefficients. Il semble alors naturel de régulariser l'estimateur des moindres carrés à l'aide d'une pénalisation et ceci d'une part pour limiter la variance des estimateurs des indices de sensibilité mais aussi afin de simplifier le modèle pour une meilleure interprétation. Une première méthode de régularisation, conduisant à des estimateurs biaisés des coefficients mais de moindre variance est la régularisation de type ridge :

$$\tilde{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \|\mathbf{Y} - Z\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2$$

Dans ce contexte et lorsque la régression est orthogonale on obtient les estimateurs à rétrécissement (voir Beran (2003)).

Régression approchée et rétrécissement

Efromovich, Donoho, Johnstone, Antoniadis

Il s'agit d'améliorer l'algorithme pour obtenir des estimateurs dont la variance de prédiction est plus faible.

$$\tilde{f}_{\boldsymbol{\gamma}}^{(n)}(\mathbf{x}) = \sum_{\mathbf{r}} \gamma_{\mathbf{r}}^{(n)} \tilde{\beta}_{\mathbf{r}}^{(n)} \psi_{\mathbf{r}}(\mathbf{x}), \quad \gamma_{\mathbf{r}}^{(n)} \in [0,1]$$

Optimalement

$$\gamma_{\mathbf{r}}^{(n)} = \frac{\beta_{\mathbf{r}}^2}{\beta_{\mathbf{r}}^2 + \mathsf{var}(\tilde{\beta}_{\mathbf{r}}^{(n)})}$$

On utilise les données pour estimer les coefficients de rétrécissement $\gamma_{\mathbf{r}}^{(n)}$

$$\hat{\gamma}_{\mathbf{r}}^{(n)} = \frac{\tilde{\beta}_{\mathbf{r}}^{(n-1)^2}}{\tilde{\beta}_{\mathbf{r}}^{(n-1)^2} + S_{\mathbf{r}}^{(n-1)}}$$

Seuillage

Lorsque la dimension d est importante, il est aussi souhaitable de rechercher parmi tous les modèles de quasi-régression, le plus parsimonieux ou creux et cela est réalisé comme dans le cadre des ondelettes par un critère de pénalisation de type “Lasso” :

$$\tilde{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \|\mathbf{Y} - Z\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1$$

qui conduit à des estimateurs de type seuillage doux :

$$\tilde{f}^{(n)}_{\boldsymbol{\gamma}}(\mathbf{x}) = \sum_{\mathbf{r}} \gamma^{(n)}_{\mathbf{r}} \tilde{\beta}^{(n)}_{\mathbf{r}} \psi_{\mathbf{r}}(\mathbf{x}), \quad \gamma^{(n)}_{\mathbf{r}} \in [0,1]$$

avec

$$\gamma^{(n)}_{\mathbf{r}} = \frac{(|\tilde{\beta}^{(n)}_{\mathbf{r}}| - t_n)_+}{|\tilde{\beta}^{(n)}_{\mathbf{r}}|}$$

Et pour un plan non orthonormé?

Tous les développements précédents sont fondés sur une analyse de la variance fonctionnelle dérivée sous l'hypothèse d'une loi uniforme sur l'hypercube unité et sont généralisables au cas de n'importe quelle mesure produit sur $\mathbb{R}^p$.

Dans la plupart des cas pratiques, une telle représentation produit n'est pas raisonnable pour la loi des entrées. Ces dernières présentent souvent des dépendances non linéaires avec un support ne couvrant pas l'hypercube, et l'intégration avec une mesure uniforme place trop de poids sur des régions vides ou peu probables de l'espace des variables, entraînant des biais et des distorsions.

Illustration

Considérons la fonction à deux entrées $f(x_1, x_2) = x1 + x_2^2$ et supposons que la loi conjointe du couple des entrées est une loi uniforme sur le carré $[0, 2]^{\times 2}$ privé de son quadrant supérieur droit.

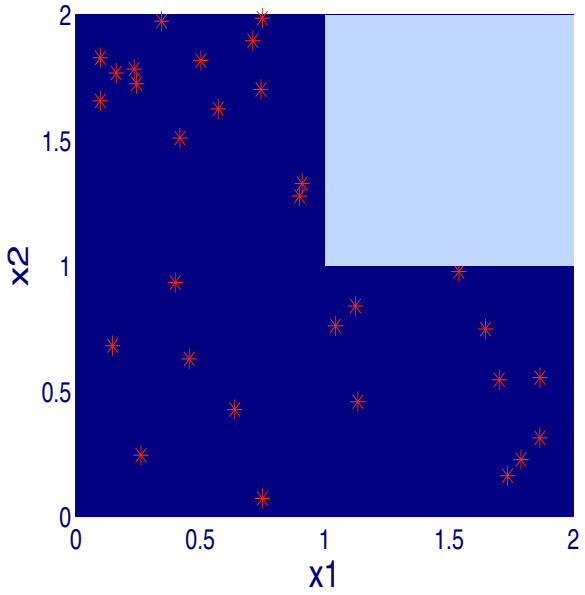
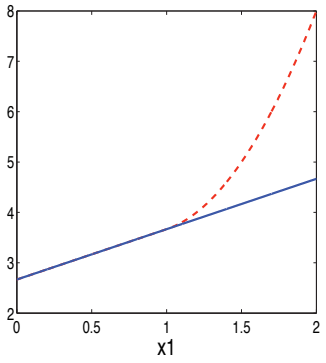
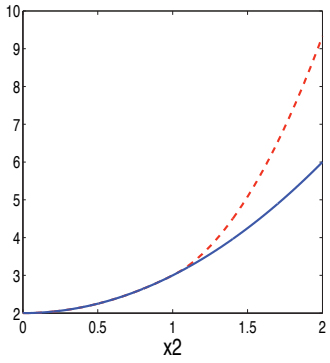


Illustration (suite)

Ajoutons à f un terme supplémentaire $g(x_1, x_2) = 10(x_1 - 1)_+^2 (x_2 - 1)_+^2$, dont le support est contenu dans le quadrant vide. L'utilisation d'une loi uniforme sur le carré $[0, 2]^{\times 2}$ produira une estimation de $F = f + g$ alors qu'en réalité sur le vrai support de la loi du couple (X_1, X_2), f et F sont indistinguables. Cette extrapolation en dehors du support produit des distorsions sur l'estimation des effets des "entrées" $\int f(x_1, x_2)dx_2$ et $\int F(x_1, x_2)dx_2$ (idem pour x_2).





FANOVA sur L^2 pondéré

On remplace d'abord la mesure uniforme sur le cube unité par une mesure admettant une densité w par rapport à la mesure de Lebesgue et on considère l'espace $L^2([0,1]^p, w)$ dont le produit scalaire est défini par $\langle f, g \rangle_w = \int f(\mathbf{x}) g(\mathbf{x}) w(\mathbf{x}) d\mathbf{x}$. On peut alors définir l'ensemble des effets $\{f_u, u \subset \{1, \ldots, p\}\}$ par

$$\{f_u\}_{u \subset p} = \text{argmin}_{\{g_u \in L^2(\mathbb{R}^u), u \in \{1,\ldots,p\}\}} \int \left(\sum_{u \subset p} g_u(\mathbf{x}) - f(\mathbf{x}) \right)^2 w(\mathbf{x}) d\mathbf{x},$$

sous les contraintes d'orthogonalité hiérarchiques

$$\forall v \subset u \ : \int f_u(\mathbf{x}) f_v(\mathbf{x}) w(\mathbf{x}) d\mathbf{x} = 0.$$

Problèmes ouverts

- Définition et algorithmes de calcul des sensibilités pour le schéma pondéré.
- Choix des bases de manière adaptative.
- Comparaison de f et $\tilde{f}$ sur des observations d'apprentissage.
- Outils pour distinguer la structure de f de celle de $\tilde{f}$
- Choix de fonctions de bases (utilisation de la régression sans approximation à l'aide de tels choix)
- Contrôle du bruit dans les mesures.
- Poser les mêmes problèmes dans un cadre bayésien.

Un exemple de boîte noire

Venables et Ripley (2000) décrivent un problème de performance d'un ordinateur. Pour 209 ordinateurs, les variables suivantes ont été mesurées :

nom	Fabricant et modèle
syct	temps d'un cycle en ns
mmin	mémoire principale minimum en KB
mmax	mémoire principale maximale en KB
cach	taille du cache en KB
chmin	nombre minimum de chaînes
chmax	nombre maximum de chaînes
perf	performance publiée par rapport à un IBM 370/158-3
estperf	performance estimée

Un ajustement par réseau de neurones à 3 couches

Après avoir supprimé nom et estperf et avoir remplacé chmin et chmax par chmax-chmin Ils ont ajusté à ce jeu de données un réseau de neurones à 3 couches pour expliquer la variable $\log$(perf) en fonction de 6 variables $(X_1, X_2, \ldots, X_6) \in [0,1]^6$ par un modèle

$$b_0 + \sum_{i=1}^{6} w_i X_i + \sum_{h=1}^{3} w_{h0} \ell(b_h + \sum_{i=1}^{6} w_{ih} X_i)$$

avec $\ell(z) = (1 + \exp(-z))^{-1}$. Cela donne lieu à une fonction dépendant de 31 paramètres que nous considérons comme notre boîte noire dans cet exemple.

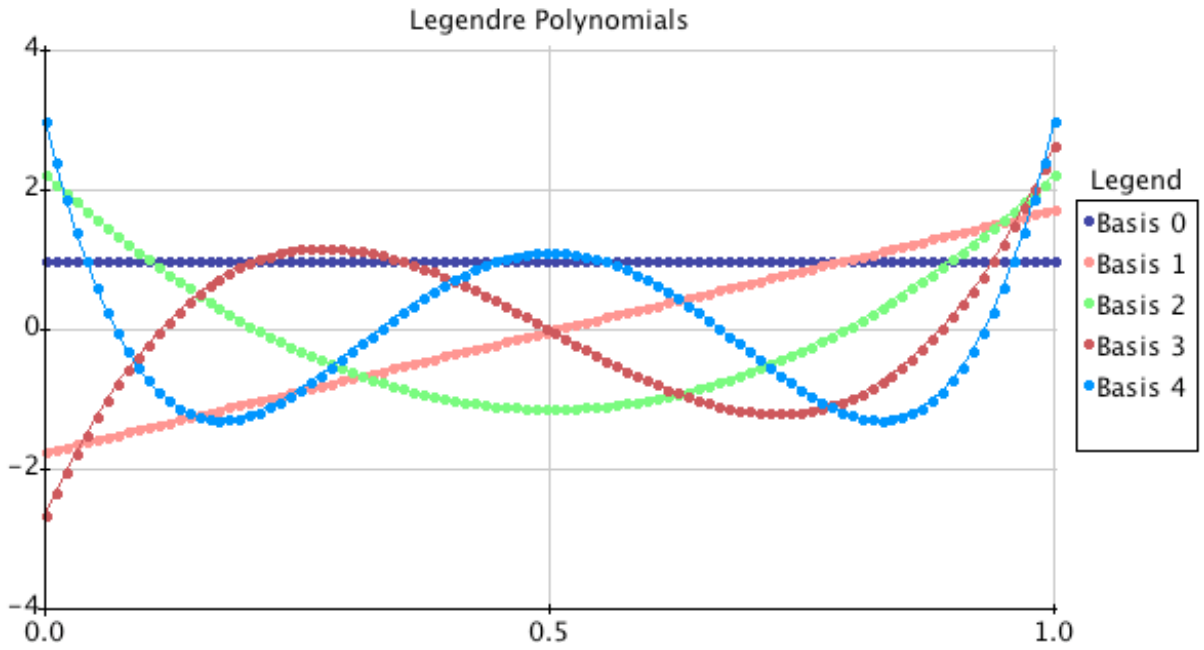
Modèle de quasi-régression

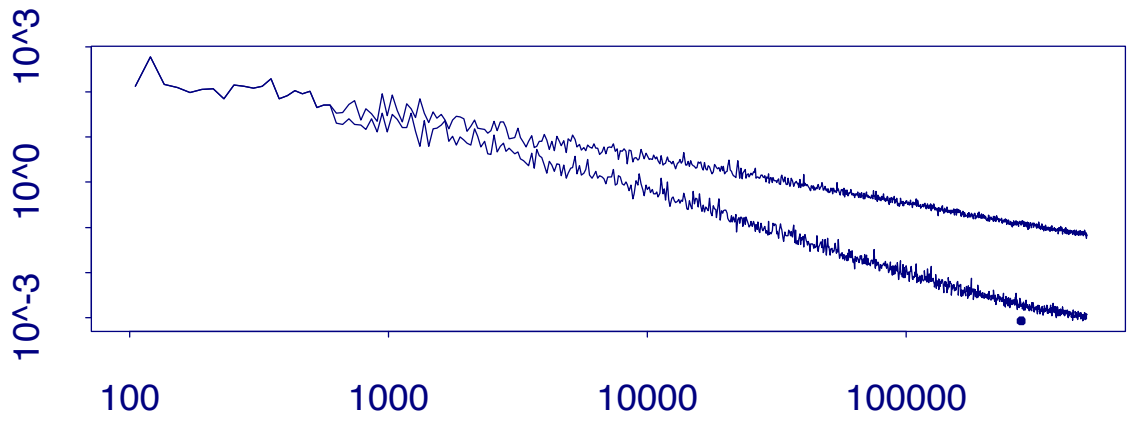
Nous avons ajusté un tel modèle en choisissant pour fonction de base les polynômes de Legendre et choisissant

$$\mathcal{R} = \mathcal{R}_{3,8,4} = \{\mathbf{r} \mid \|\mathbf{r}\|_0 \leq 3, \|\mathbf{r}\|_1 \leq 8, \|\mathbf{r}\|_\infty \leq 4\}$$

L'erreur du l'approximation de $f(\mathbf{x})$ par $\tilde{f}(\mathbf{x})$ est estimée en fonction du nombre de simulations N.

Nous avons normalisé le ISE par la variance de f. Ainsi une valeur du ISE$/\sigma^2 = 1$ signifie que $\tilde{f}$ fait aussi bien que la moyenne de f, alors qu'une valeur ISE$/\sigma^2 = 0.01$ signifie que $\tilde{f}$ explique 99% de la variance de f.


Legendre Polynomials
4
2
0
−2
−4
0.0
0.5
1.0
Legend
Basis 0
Basis 1
Basis 2
Basis 3
Basis 4


10^3
10^0
10^-3
100
1000
10000
100000

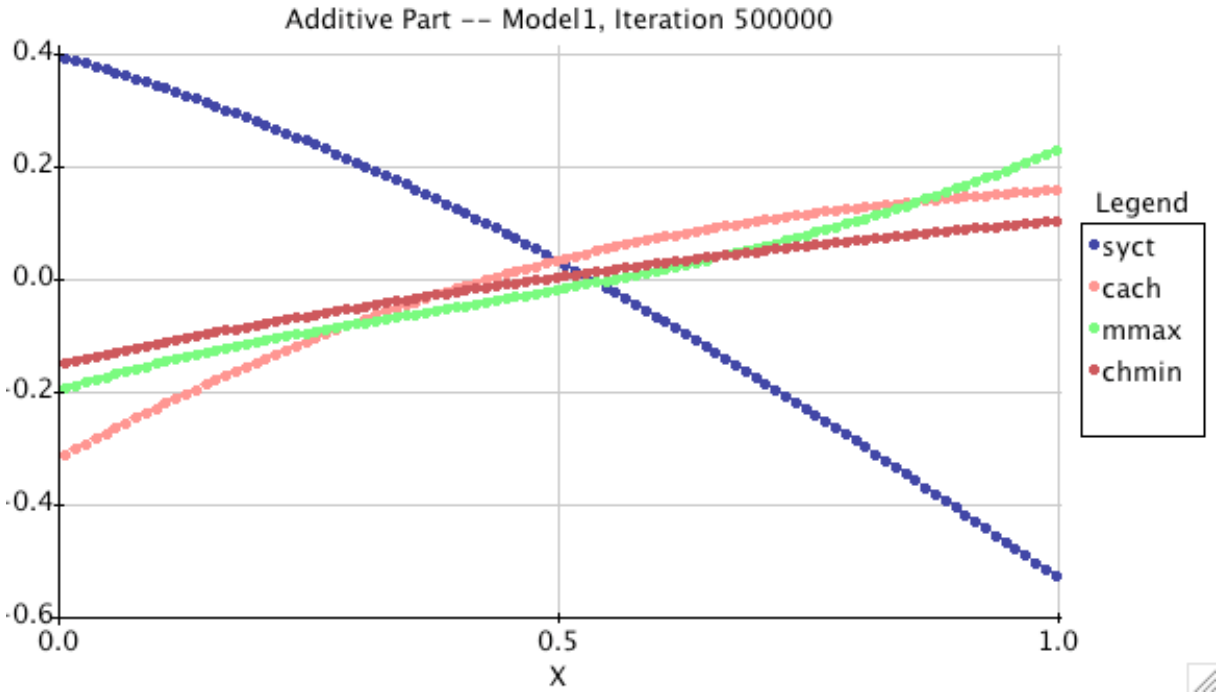
Résultats

A l'issue des 500000 itérations, on explique 98.2 % de la variance par le modèle de régression.

```
Number of bases is 1145

######  ANOVA at Iteration 500000  ######
1-R^2 (LOF) is 0.0012087 at iteration 499500
Beta[0] (constant factor) is 2.0720

Variances on one and two variables / sample variance
syct        mmin        mmax        cach        chmin       deltach
0.52184
8.6385E-4   0.010800
0.0089844   0.026059    0.087686
0.055498    0.0063268   0.054925    0.13143
0.011096    5.9349E-4   0.0084105   0.010159    0.036990
2.4175E-4   4.9982E-4   2.6319E-4   0.0014536   2.2085E-4   0.0093188
```


Additive Part -- Model1, Iteration 500000
0.4
0.2
0.0
-0.2
-0.4
-0.6
0.0
0.5
1.0
X
Legend
syct
cach
mmax
chmin

Plus de la moitié de la variation est expliquée par l'effet du seul facteur syct. Le figure précédente montre les estimateurs des effets individuels des facteurs. On voit que seul des fonctions linéaires ou quadratiques (dues au fait que $\|\mathbf{r}\|_\infty \leq 4$) sont présentes. Parmi les interactions on notera uniquement les interactions entre mmax et cach et entre syct et cach qui pou chacune prennent compte respectivement 5.5 % de la variance totale.

Références Bibliographiques

An, J. and A. B. Owen (2001). Quasi-regression. Journal of Complexity 17 (4), 588–607.

Anderson, T. W. (1984), An Introduction to Multivariate Statistical Analysis (Second Edition), Wiley.

Antoniadis, A. (1983) Analysis of variance on function spaces, *Math. Oper. Forsch. und Statist.*, series Statistics, vol **15**, No 1, 59–71, 1984.

Chan, T. F., G. H. Golub, and R. J. LeVeque (1983). Algorithms for computing the sample variance: Analysis and recommendations. The American Statistician 37, 242–247.

Efromovich, S. (1992), On orthogonal series estimators for random design nonparametric regression, in 'Computing Science and Statistics. Proceedings of the 24rd Symposium on the Interface', pp. 375– 379.

Efron, B. and C. Stein (1981). The jackknife estimate of variance. Annals of Statistics 9, 586–596.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. Annals of Mathematical Statistics 19, 293–325.

Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional anova models. Annals of Statistics 26, 242–272.

Owen, A. B. (1992), 'A central limit theorem for Latin hypercube sampling', Journal of the Royal Statistical Society, Series B 54, 541–551.

Owen, A. B. (1992). Orthogonal arrays for computer experiments, integration and visualization. Statistica Sinica 2, 439–452.

Ripley, B. (1996). Pattern Recognition and Neural Networks. Cambridge University Press.

Roosen, C. B. (1995). Visualization and exploration of high-dimensional functions using the functional ANOVA decomposition. Ph. D. thesis, Stanford University,

Department of Statistics.

Saltelli, A., K. Chan, and E. M. Scott (2000). *Sensitivity Analysis*. Chichester: Wiley.

Sobol', I. M. (1993). Sensitivity estimates for nonlinear mathematical models. Mathematical Modeling and Computational Experiment 1, 407–414.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. Annals of Statistics 22, 118–171.

Venables, W. and B. Ripley (1999). *Modern Applied Statistics with S-Plus*, 3rd Edition. New York: Springer.