

Plans d'expériences optimaux et régression PLS pour l'analyse de sensibilité globale

GdR MASCOT NUM 19 mars 2010

J.-P. Gauchi¹, S. Lehuta², S. Mahévas²

¹*INRA, MIA(UR341), Domaine de Vilvert, Jouy-en-Josas, 78352, France*

²*IFREMER, Dpt Ecologie et Modèles pour l'Halieutique, Nantes, 44311, France*

Analyse de sensibilité globale sous trois contraintes particulières:

- 1 Un temps de **calcul trop long** pour réaliser toutes les simulations demandées lors de l'utilisation de méthodes usuelles d'AS (Sobol, Saltelli,...).
- 2 L'existence d'entrées **non indépendantes**:
 - corrélations structurelles (observées ou supposées)
 - relations fonctionnelles
- 3 La présence d'**entrées qualitatives** (discrètes à quelques modalités) à côté d'entrées quantitatives (continues).

- Introduction
- Les 5 étapes de la méthodologie proposée
 - ① Construction d'un réseau candidat de simulations
 - ② Postulation du modèle d'approximation et construction de sa matrice
 - ③ Calcul du déterminant maximum de la matrice d'information du modèle
 - ④ Calcul du plan optimal de simulation
 - ⑤ Analyse des résultats et calcul des indices de sensibilité *SI-VIP*
- Application à un problème halieutique de l'IFREMER
- Conclusion et perspectives

- C'est une tentative d'une nouvelle démarche en cinq étapes, alliant deux idées-force :
 - Un **plan de simulation D-optimal** pour trouver un plan de simulation de faible taille
 - L'utilisation de la **régression PLS**, conduisant à des **nouveaux indices de sensibilité** (différents de ceux de Sobol), pour prendre en compte la multicolinéarité et les liaisons fonctionnelles.
- Résultats fructueux sur une application réelle de l'IFREMER (modèle ISIS-FISH).

Etape 1 : Construction d'un réseau candidat de simulations

- On construit un plan usuel d'exploration de l'espace des k entrées continues (LHS, ...):
⇒ réseau R_1 de dimensions $(N_0 \times k)$.
- Les modalités de chacune des q entrées discrètes sont répétées et randomisées:
⇒ réseau R_2 de dimensions $(N_0 \times q)$.
- Les q colonnes dans R_2 sont remplacées par leurs indicatrices:
⇒ réseau R'_2 de dimensions $(N_0 \times q')$.
- On concatène R_1 et R'_2
⇒ réseau $R_3 = R_1 || R'_2$ de dimensions $(N_0 \times p)$, avec $p = (k + q')$.
- On applique les contraintes fonctionnelles et on corrèle les r ($\leq k$) entrées (Iman & Conover, 1982) sur R_3 :
⇒ réseau R_4 de dimensions $(N \times p)$, avec $N < N_0$.
- A partir de R_4 , en supprimant les colonnes qui créent les combinaisons linéaires, on forme le réseau candidat: ⇒réseau R_C de dimension $(N \times p^*)$.

Etape 2 : Postulation du modèle d'approximation et construction de sa matrice

- On postule un modèle statistique:

$$y = f(\text{entrées}) + \varepsilon$$

où:

- f est une approximation du modèle numérique M (inconnu) construite avec **les composantes PLS**, elles-mêmes formées par les coefficients d'**un polynôme Q de degré deux** (P termes: effets linéaires, effets des indicatrices, effets quadratiques, et interactions doubles).
- ε est l'erreur de modèle, on supposera ici $\varepsilon \sim_{iid} \mathcal{N}(0, \sigma^2)$, σ^2 inconnue.
- On construit la matrice X de dimensions $(N \times P)$ du modèle Q à partir de R_C .

- Un plan d'expériences à **mesure continue** est un **support continu** χ associé à **une densité de probabilité** $\pi \in \mathcal{D}$, définie sur ce support (design measure).
- La matrice d'information de Fisher de la mesure π s'écrit (avec $f(x)^T$ vecteur fonction de régression ($1 \times P$)):

$$M_F(\pi) = \int_{\chi} f(x)f(x)^T \pi(dx)$$

- La **mesure D-optimale** s'écrit:

$$\pi_D = \text{Arg} \left\{ \max_{\pi \in \mathcal{D}} \left[\det \left(\int_{\chi} f(x)f(x)^T \pi(dx) \right) \right] \right\}$$

- Propriété importante: discrétisation de π_D sur N_S points de support de χ , avec $N_S < P(P+1)/2$.

- Soit $M_F^n = X_n^T X_n$ la matrice d'information du modèle Q calculée pour une matrice X_n ($n \times P$) dont les n lignes $\in N$ lignes de X .
- Soit ζ_n le plan d'expériences = matrice formée par n lignes de R_C
- Soit le plan D-optimal **discret** à n ($n < N$) points de support:

$$\zeta_n^D = \text{Arg} \left\{ \max_{\zeta_n} \left[\det \left(X_n^T X_n \right) \right] \right\}$$

- Alors la D -efficacité d'un plan discret est:

$$D - \text{efficacité} = 100 \times \left[\frac{\det(M_F^n) / n^P}{\det(M_F(\pi_D))} \right]^{1/P}$$

Etape 3 : Calcul du déterminant maximum de la matrice d'information du modèle

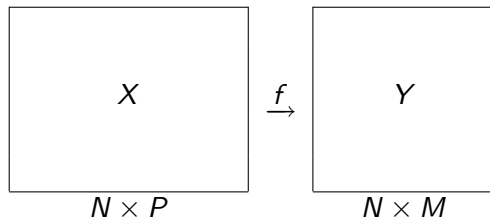
- On approxime le support continu χ par la discrétisation R_C .
 - On calcule **une approximation $\widetilde{\det}$ du maximum (global) de $\det(M_F(\pi_D))$.**
 - Algorithme utilisé : à partir des masses initiales (uniformes) en chacun des points de R_C on charge itérativement les points où la fonction de variance $f(x)^T (X_N^T X_N)^{-1} f(x)$ est maximum (Torsney, 1988).
- la qualité de cette approximation dépend du degré de discrétisation de R_C .
- **Objectif : obtenir avec $\widetilde{\det}$ une approximation de la D-efficacité de tout plan D-optimal discret ζ_n^D calculé sur R_C .**

Etape 4 : Calcul du plan optimal de simulation

- Le plan D-optimal (discret) de simulation cherché ξ_n^D est une matrice $(n \times p^*)$ avec $p^* \leq n \leq N$, sous-ensemble extrait de R_C qui sera optimal au sens du critère de D-optimalité, c'est-à-dire que, parmi tous les plans de taille n , ce sera celui qui présentera **la plus grande D-efficacité**.
- En général n est **très faible** par rapport à N et c'est ce qui fait l'un des points forts de la méthode.
- Plusieurs algorithmes discrets (pas de preuve de convergence) mais des algorithmes efficaces (par exemple l'algorithme à échanges doubles de Fedorov, 1971).

Regression PLS: rappels

- Soit le modèle de régression multivarié: $Y = f(X)$



où X est une matrice $N \times P$ constituée de M variables explicatives X_j et Y est une matrice $N \times M$ constituée de M réponses Y_k , observées sur N individus.

- On cherche à relier X à Y au moyen d'un ensemble de P fonctions linéaires (polynomiales) en les paramètres β_j sur les X_j .

Soit:

- E_0 : matrice des variables X_j centrées-réduites
- F_0 : matrice des variables Y_k centrées-réduites
- E_h : matrice des résidus de la décomposition de E_0 en utilisant h composantes PLS
- F_h : matrice des résidus de la décomposition de F_0 en utilisant h composantes PLS
- F_{hk} : $k - ième$ colonne de F_h
- H : nombre de composantes PLS t_h retenues

Regression PLS: rappels

- C'est une méthode de régression itérative

- **Etape 1 :**

- On construit une combinaison linéaire $u_1 = F_0 c_1$ des colonnes de F_0 et une combinaison linéaire $t_1 = E_0 w_1$ des colonnes de E_0 par la maximisation de

$$\text{cov}(u_1, t_1) = \text{cor}(u_1, t_1) \sqrt{\text{var}(u_1) \text{var}(t_1)}$$

sous les contraintes :

$$\|w_1\|_2 = \|c_1\|_2 = 1$$

Donc il faut maximiser simultanément la variance expliquée par t_1 , la variance expliquée par u_1 , et la corrélation entre les deux ce qui revient à maximiser le produit scalaire

$$\langle t_1, u_1 \rangle = \|t_1\| \cdot \|t_2\| \cdot \text{cor}(t_1, u_1)$$

$\implies u_1$ et t_1 sont aussi corrélées que possible et résumant au mieux les tableaux E_0 et F_0 .

- On construit ensuite les régressions: $E_0 = t_1 p_1^T + E_1$ et $F_0 = t_1 r_1^T + F_1$

• Etape 2 :

- $E_0 \longrightarrow E_1$; $F_0 \longrightarrow F_1 \implies$ deux nouvelles composantes:
 u_2 combinaison linéaire des colonnes de F_1 et t_2 combinaison linéaire des colonnes de E_1
- On construit ensuite les régressions:

$$E_0 = t_1 p_1^T + t_2 p_2^T + E_2$$

$$F_0 = t_1 r_1^T + t_1 r_2^T + F_2$$

• Etapes suivantes :

On itère la procédure jusqu'à ce que les composantes t_1, \dots, t_H expliquent suffisamment F_0

\implies Les composantes PLS t_h sont des combinaisons linéaires orthogonales des colonnes de E_0

\implies De la décomposition $F_0 = t_1 r_1^T + \dots + t_h r_h^T + F_h$ on peut déduire les équations de régression PLS (les $\hat{\beta}_j$ sont biaisés) :

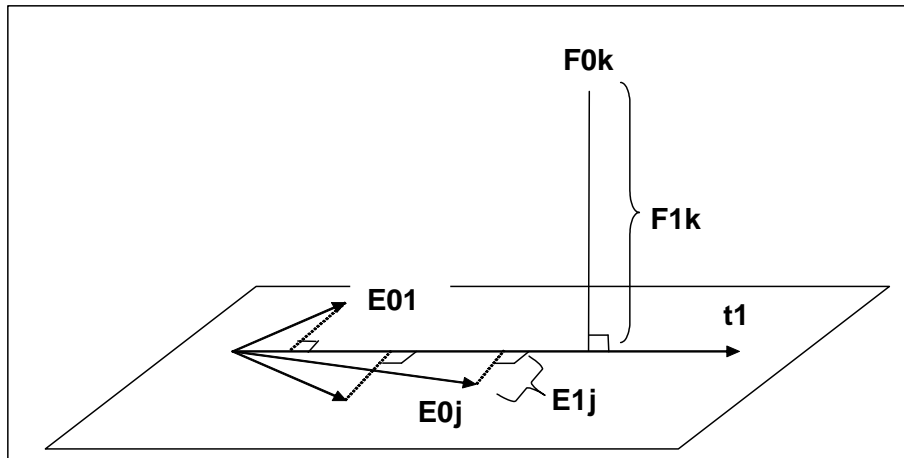
$$Y_k = \hat{\beta}_{k_0} + \hat{\beta}_{k_1} X_1 + \dots + \hat{\beta}_{k_M} X_P + F_{hk}$$

Solution:

- w_1 est vect propre de $E_0^T F_0 F_0^T E_0$ associé à la + grande val propre θ_1^2
- c_1 est vect propre de $F_0^T E_0 E_0^T F_0$ associé à la + grande val propre θ_1^2
- algorithme utilisé : Nipals

Regression PLS: rappels

Visualisation de $t_1 = E_0 w_1$



- Indice de sensibilité de Sobol au premier ordre:

$$SI_j = V(E(Y|X_j)) / V(Y) \underset{\text{if linear relation}}{=} \text{cor}^2(Y, X_j)$$

- On a:

$$y_0 = r_1 t_1 + \dots + r_H t_H + y_H = \sum_{h=1}^H \frac{\text{sign}(r_h) SI_h^{1/2}}{\sigma_h} t_h + y_H$$

$$y_0 = \sum_{h=1}^H r_h \left(\sum_{j=1}^P w_{hj}^* E_{0j} \right) + y_H = \sum_{j=1}^P \left(\sum_{h=1}^H r_h w_{hj}^* \right) E_{0j} + y_H$$

$$y_0 = \hat{\beta}_1^{PLS} E_{01} + \dots + \hat{\beta}_P^{PLS} E_{0P} + y_H$$

- Les coefficients $\hat{\beta}_j^{PLS}$ peuvent être vus comme des nouveaux indices (signés) SI_j :

$$SI_j = 100 \times \frac{\hat{\beta}_j^{PLS}}{\sum_{j=1}^P |\hat{\beta}_j^{PLS}|}$$

Propriété:

$$\det \left(\text{Varcov} \left(\hat{\beta}^{PLS} \right) \right) \ll \det \left(\text{Varcov} \left(\hat{\beta}^{OLS} \right) \right)$$

Définition des indices SI-VIP

- Soit la redondance

$$Rd(Y, t_1, \dots, t_H) = \sum_{h=1}^H cor^2(Y, t_h)$$

- Soit le *VIP*, "Variable Importance in the Projection" (Wold, 1992) pouvoir explicatif d'un facteur X_j sur la sortie Y , donné par:

$$VIP_{Hj} = \left[\frac{P}{Rd(Y, t_1, \dots, t_H)} \sum_{l=1}^H Rd(Y, t_l) w_{lj}^2 \right]^{1/2}$$

- Comme $\sum_{j=1}^P VIP_{Hj}^2 = P$
On propose (Ellouze & al., 2010) pour mesurer la sensibilité de Y par rapport au facteur X_j :

$$SI - VIP_j \text{ (en \%)} = 100 \times \frac{VIP_{Hj}^2}{P}$$

Etape 5 : Analyse des résultats et calcul des SI-VIP

- Comme de nombreux coefficients PLS ne sont pas significatifs, on utilise une méthode de sélection (méthode PLS-BQ , Gauchi & Chagnon, 2001, ou PLS-Forward, Bastien & al., 2005) pour sélectionner les P' termes significatifs vis-à-vis de la variation du Q_{cum}^2 de Wold ou du $Q2G$ (Lazraq & al, 2003)
⇒ En général $P' \ll P$ dans le modèle polynomial final Q' .
- On calcule les $SI - VIP$.

Application à un problème halieutique de l'IFREMER

Introduction

- Pêche ANCHOIS du Golfe de Gascogne.
- Modèle numérique : le modèle ISIS-Fish.
- L'anchois réalise des migrations annuelles (2 classes d'âge) vers 5 zones de ponte: assez grande variabilité de la répartition dans les différentes zones.
- Impact important de la distribution spatiale dans ces zones sur l'efficacité des mesures de gestion.
- **OBJECTIF: On souhaite tester finement l'impact des différents patrons de répartition par AS.**

Application à un problème halieutique de l'IFREMER

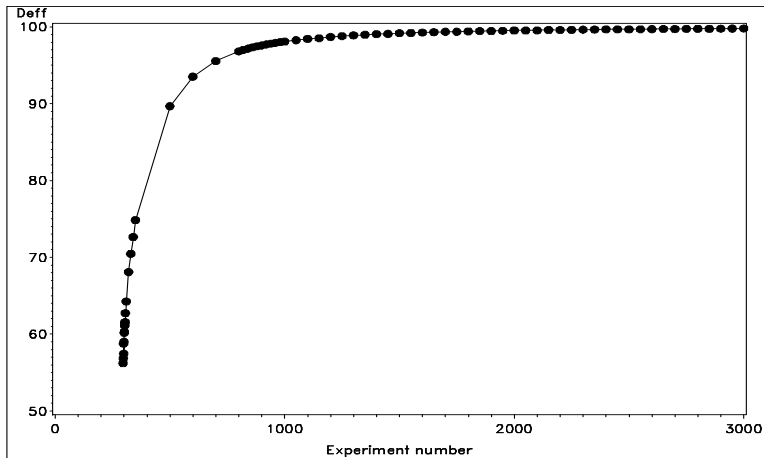
Les 20 facteurs (entrées)

- Les **10 coefficients de migration** (continus et bornés) depuis une unique zone de départ vers les cinq zones d'arrivée: ils somment à un à l'intérieur d'une classe d'âge et sur les séries historiques des corrélations non négligeables apparaissent entre certains d'entre eux.
- Date de migration (discret à 2 modalités), fécondité (continu), mortalité naturelle (discret, 3 modalités), effort total fleet1, ..., effort total fleet5, TAC (discret à 2 modalités), AMPR (discret à 13 modalités).
- Ici on traite **un jeu de données d'essai**, jeu final en cours de traitement.

Application à un problème halieutique de l'IFREMER

Le plan de simulation

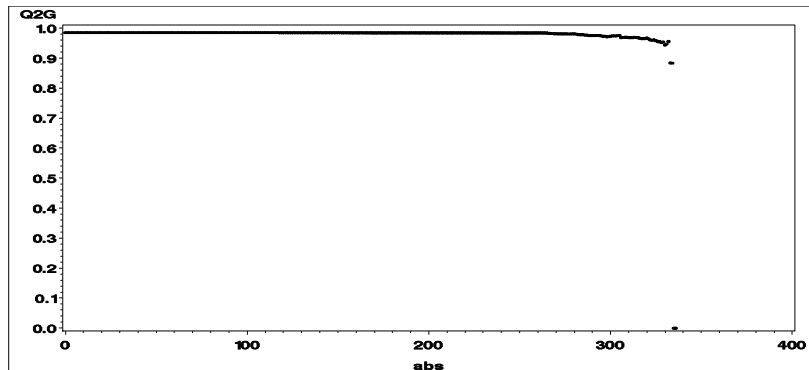
- Réseau candidat sous contraintes: 6854 lignes.
- Plan de simulation D-optimal à 790 lignes (790 calculs du modèle à lancer)



Application à un problème halieutique de l'IFREMER

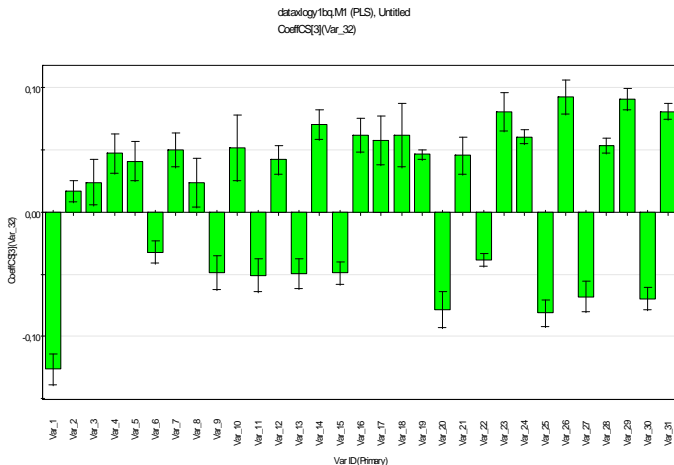
Analyse de la réponse "BIOFINALE" (1)

- Etape 1: Régression PLS sur 336 termes $\implies R^2(\%) = 98.47$;
Q2G cumulé = 0.98 (critère compris entre 0 et 1).
- Etape 2: Elimination des termes ne contribuant pas ou presque pas à ce critère, deux méthodes au choix
 - avec la méthode PLS-BQ (Gauchi & Chagnon, 2001):



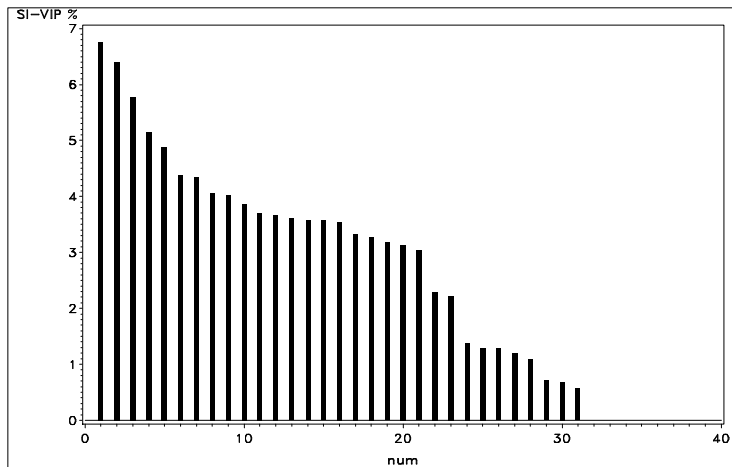
Application à un problème halieutique de l'IFREMER

Analyse de la réponse "BIOFINALE" (2): les coefficients de régression PLS des 31 termes retenus



Application à un problème halieutique de l'IFREMER

Analyse de la réponse "BIOFINALE" (3): les SI-VIP



Application à un problème halieutique de l'IFREMER

Analyse de la réponse "BIOFINALE" (4): un extrait des intitulés des SI-VIP

6.77 → morn1; 6.40 → A1G*morn1; 5.77 → morn3;
5.14 → A1G*morn3 ; 4.89 → A2G*morn1 ; 4.39 → A1N*morn1;
4.35 → A2N*morn1 ; 4.06 → A2LC*morn3; 4.03 → A2N*morn3;
3.87 → A1LL*morn1 ; 3.70 → A2LL*morn1 ; 3.67 → A1N*morn3;
3.61 → A2R*morn1

- L'application IFREMER a été traitée avec succès.
- Avantages:
 - prise en compte des corrélations et des relations fonctionnelles (point fort de PLS : pas d'inversion de matrice),
 - plan de simulation de taille réduite,
 - on peut choisir le degré du polynome Q et les termes d'interaction d'intérêt.
- Inconvénient: un simple polynome peut s'avérer insuffisant pour construire les composantes PLS si présence de fortes non-linéarité
- Suite en cours avec une approche PLS non linéaire