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

GdR MASCOT NUM 19 mars 2010

J.-P. Gauchi[1], S. Lehuta[2], S. Mahévas[2]

[1] *INRA, MIA(UR341), Domaine de Vilvert, Jouy-en-Josas, 78352, France*
[2] *IFREMER, Dpt Ecologie et Modèles pour l'Halieutique, Nantes, 44311, France*

Le contexte

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).

Plan

- 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

Introduction

- 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, ...):
  $\Longrightarrow$ réseau R_1 de dimensions $(N_0 \times k)$.
- Les modalités de chacune des q entrées discrètes sont répétées et randomisées:
  $\Longrightarrow$ réseau R_2 de dimensions $(N_0 \times q)$.
- Les q colonnes dans R_2 sont remplacées par leurs indicatrices:
  $\Longrightarrow$ réseau R_2' de dimensions $(N_0 \times q')$.
- On concatène R_1 et R_2'
  $\Longrightarrow$ 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:
  $\Longrightarrow$ 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: $\Longrightarrow$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:

$$y = f\left(entrées\right) + \varepsilon$$

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 $\varepsilon \sim_{iid} \mathcal{N}\left(0, \sigma^2\right)$, σ^2 inconnue.

- On contruit la matrice X de dimensions $(N \times P)$ du modèle Q à partir de R_C.

Plan D-optimal: rappels

- 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:

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

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

D-efficacité: rappels

- 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 ξ_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:

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

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

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

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** $\det\left(M_F\left(\pi_D\right)\right)$.
  - 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 $f\left(x\right)^T \left(X_N^T X_N\right)^{-1} f\left(x\right)$ est maximum (Torsney, 1988).

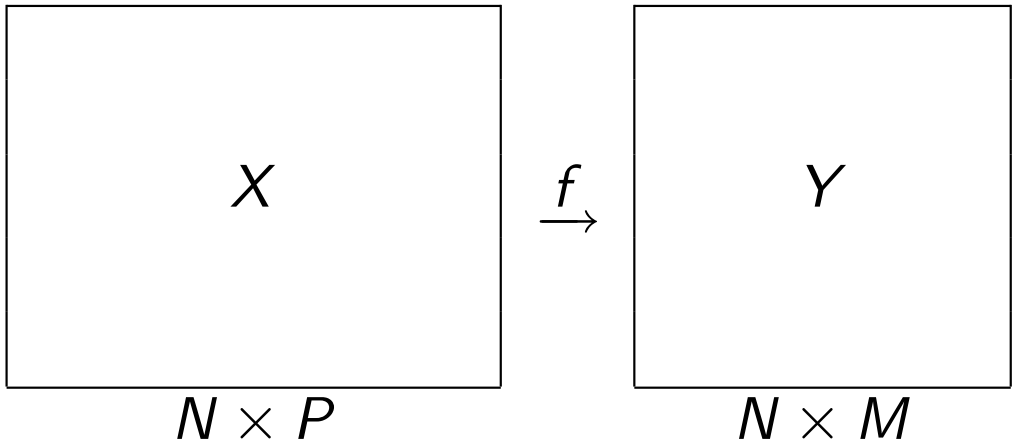
  $\longrightarrow$ la qualité de cette approximation dépend du degré de discrétisation de R_C.
- **Objectif : obtenir avec** $\widetilde{\det}$ **une approximation de la D-efficacité de tout plan D-optimal discret** ξ_n^D **calculé sur** R_C.

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.

Regression PLS: rappels

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

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

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

$$\langle t_1, u_1 \rangle = \|t_1\| \cdot \|t_2\| \cdot cor\left(t_1, u_1\right)$$

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

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

Regression PLS: rappels

- **Etape 2 :**
  - $E_0 \longrightarrow E_1$; $F_0 \longrightarrow F_1$ $\Longrightarrow$ 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:

$$\begin{aligned} E_0 &= t_1 p_1^T + t_2 p_2^T + E_2 \\ F_0 &= t_1 r_1^T + t_1 r_2^T + F_2 \end{aligned}$$

- **Etapes suivantes :**
  On itère la procédure jusqu'à ce que les composantes $t_1, \ldots, t_H$ expliquent suffisamment F_0
  $\Longrightarrow$ Les composantes PLS t_h sont des combinaisons linéaires orthogonales des colonnes de E_0
  $\Longrightarrow$ De la décomposition $F_0 = t_1 r_1^T + \ldots + t_h r_h^T + F_h$ on peut déduire les équations de régression PLS (les $\hat{\beta}_j$ sont biaisés) :

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

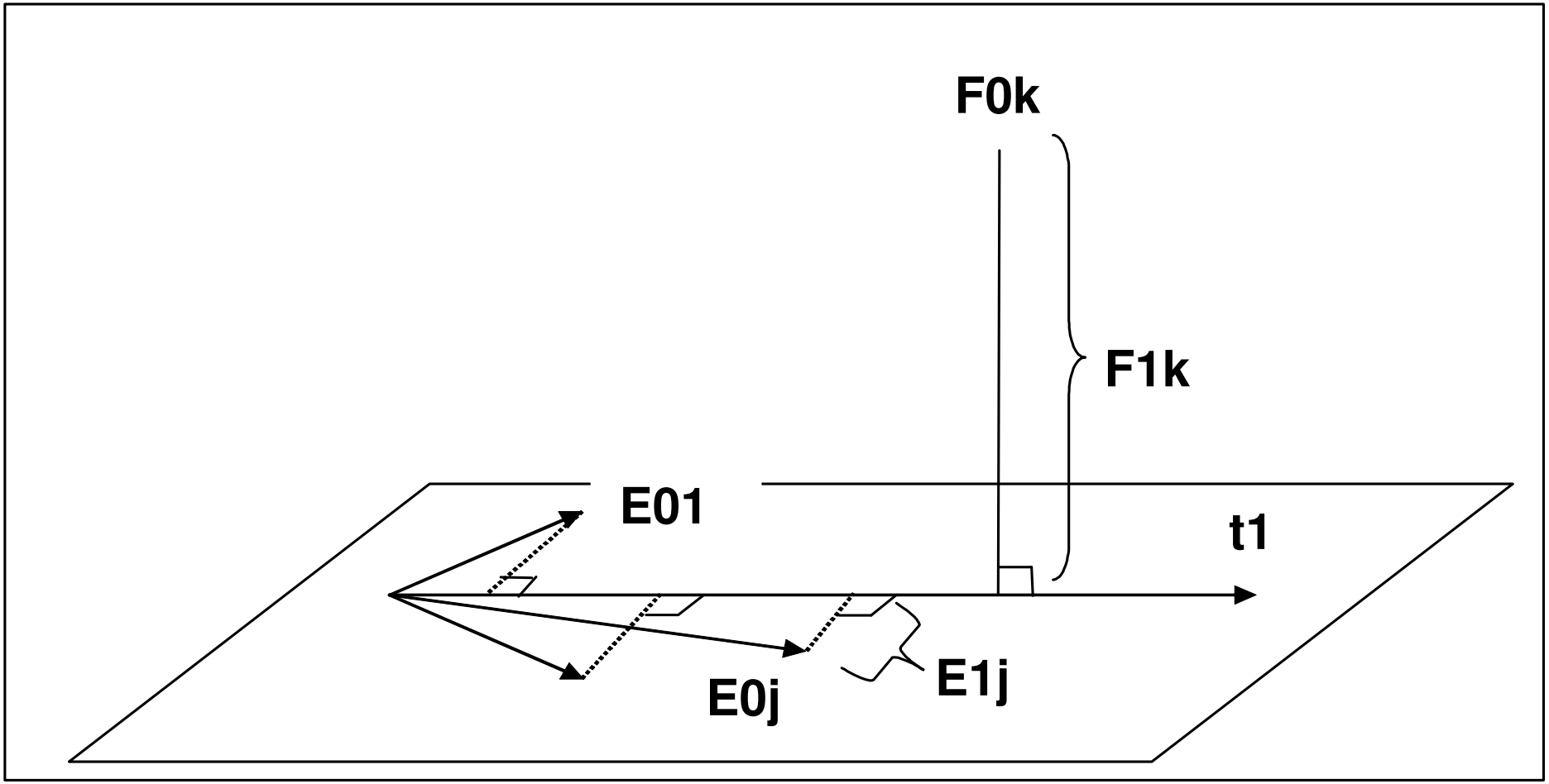
Regression PLS: rappels

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 $t_1 = E_0 w_1$



Lien avec l'analyse de sensibilité

- Indice de sensibilité de Sobol au premier ordre:

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

- On a:

$$\begin{aligned} y_0 &= r_1 t_1 + \ldots + r_H t_H + y_H = \sum_{h=1}^{H} \frac{sign(r_h) SI_h^{1/2}}{\sigma_h} t_h + y_H \\ y_0 &= \sum_{h=1}^{H} r_h \left(\sum_{j=1}^{P} w_{hj}^* E_{0j} \right) + y_H = \sum_{j=1}^{P} \left(\sum_{h=1}^{H} r_h w_{hj}^* \right) E_{0j} + y_H \\ y_0 &= \hat{\beta}_1^{PLS} E_{01} + \ldots + \hat{\beta}_P^{PLS} E_{0P} + y_H \end{aligned}$$

Lien avec l'analyse de sensibilité

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

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

Propriété:

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

Définition des indices SI-VIP

- Soit la redondance

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

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

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

- 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:

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

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^2_{cum} de Wold ou du $Q2G$ (Lazraq & al, 2003)
  $\Longrightarrow$ En général $P' << P$ dans le modèle polynomial final Q'.
- On calcule les $SI - VIP$.

- Pêcherie 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

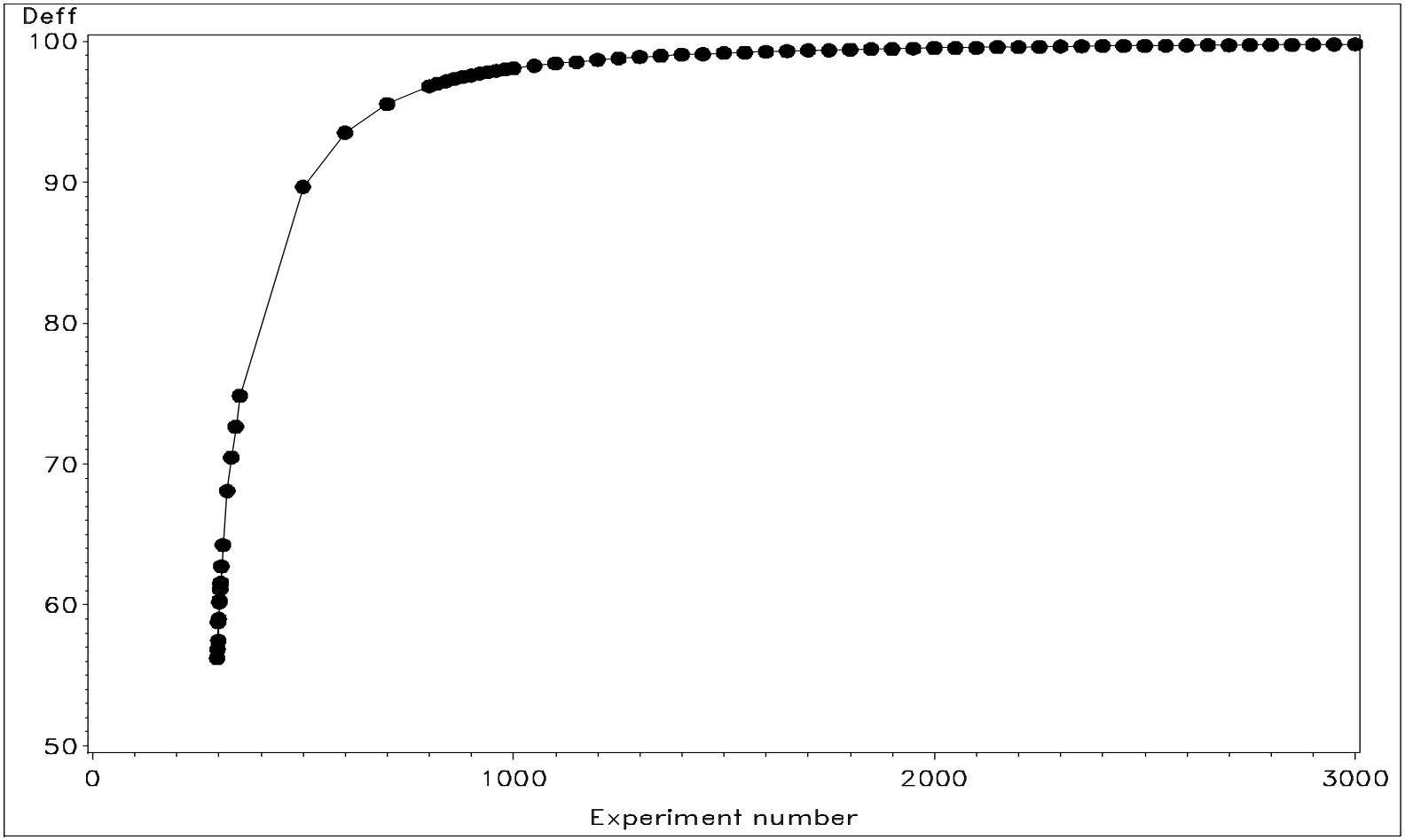
Les 20 facteurs (entrées)

- 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

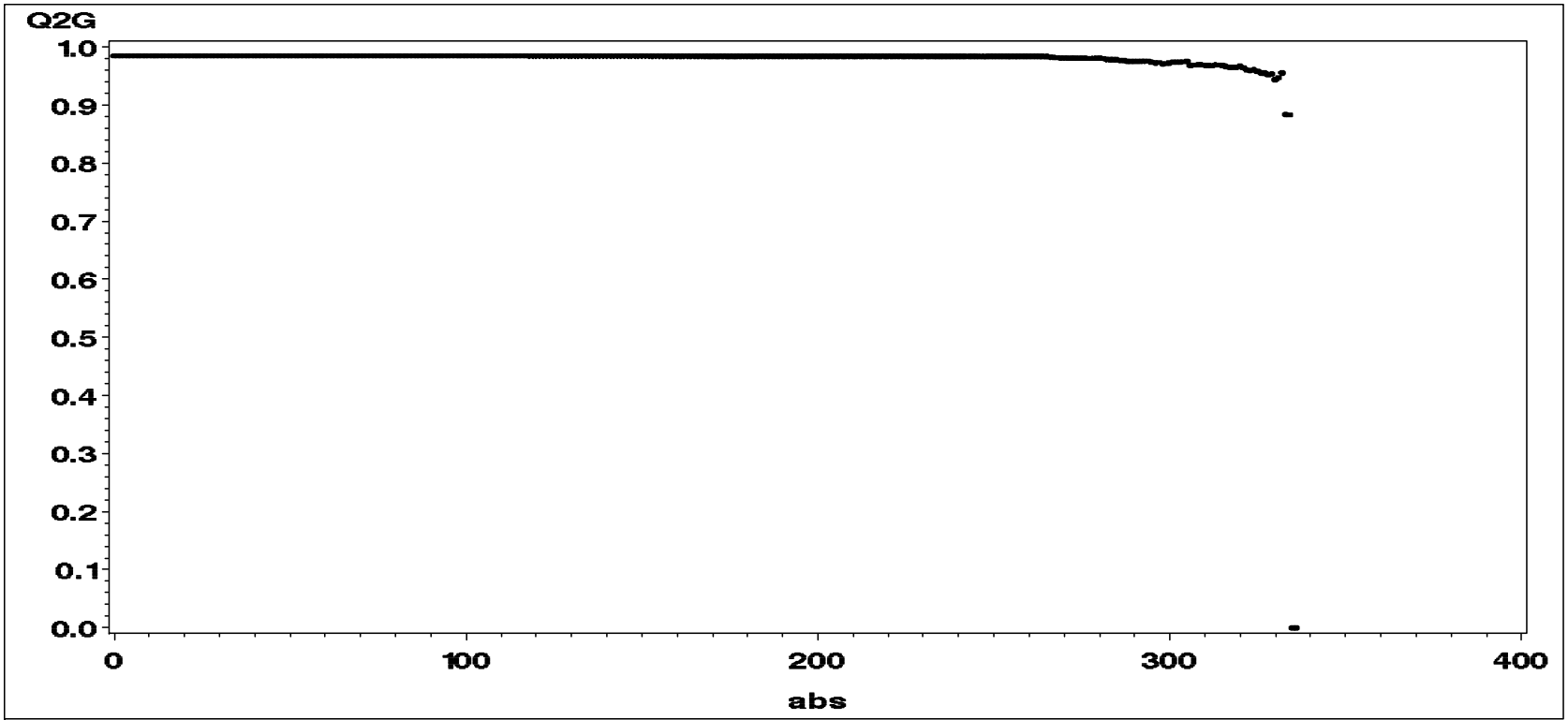
- Réseau candidat sous contraintes: 6854 lignes.
- Plan de simulation D-optimal à 790 lignes (790 calculs du modèle à lancer)



Application à un problème halieutique de l'IFREMER

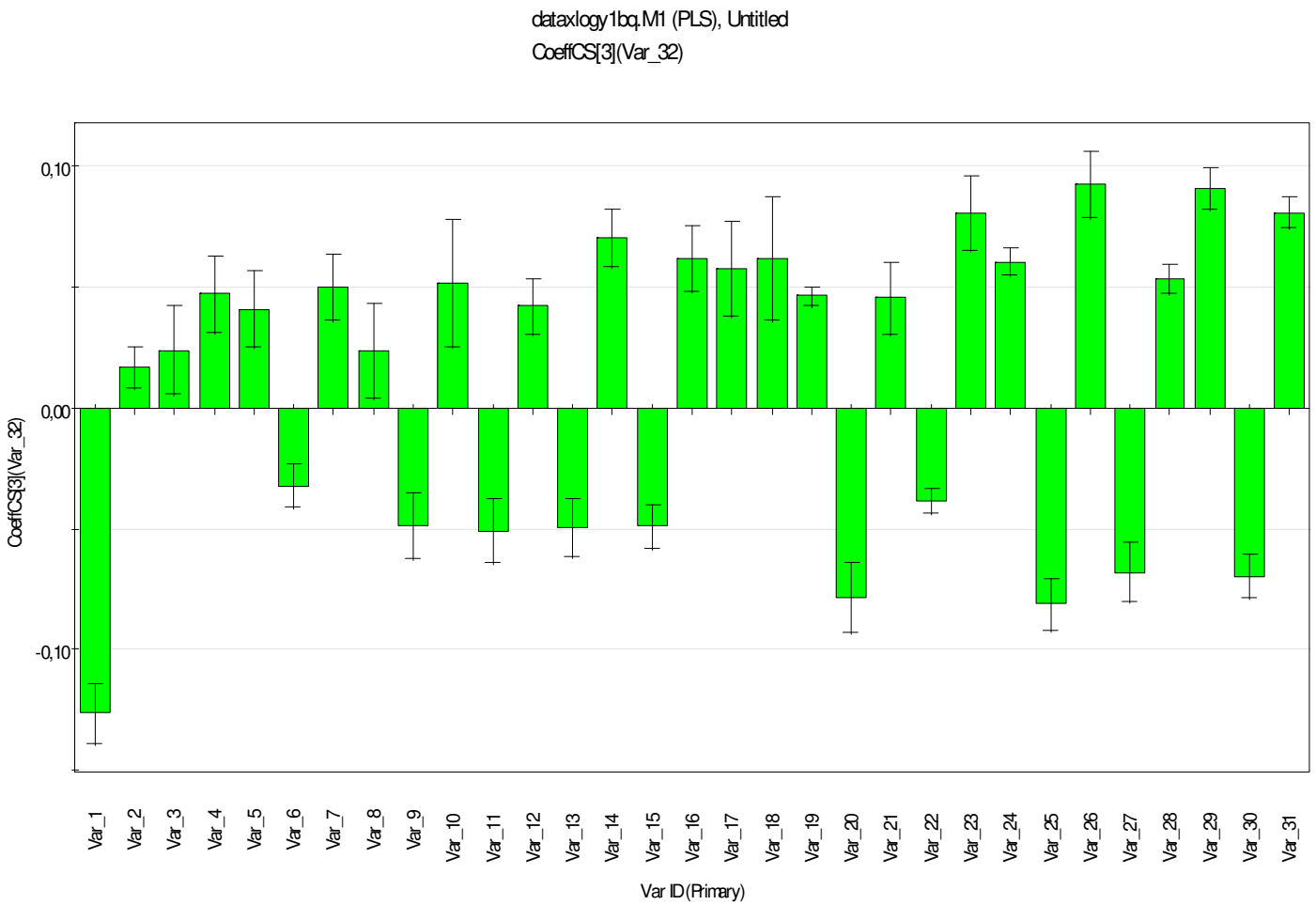
Analyse de la réponse "BIOFINALE" (1)

- Etape 1: Régression PLS sur 336 termes $\Longrightarrow R^2(\%) = 98.47$; $Q2G$ 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):



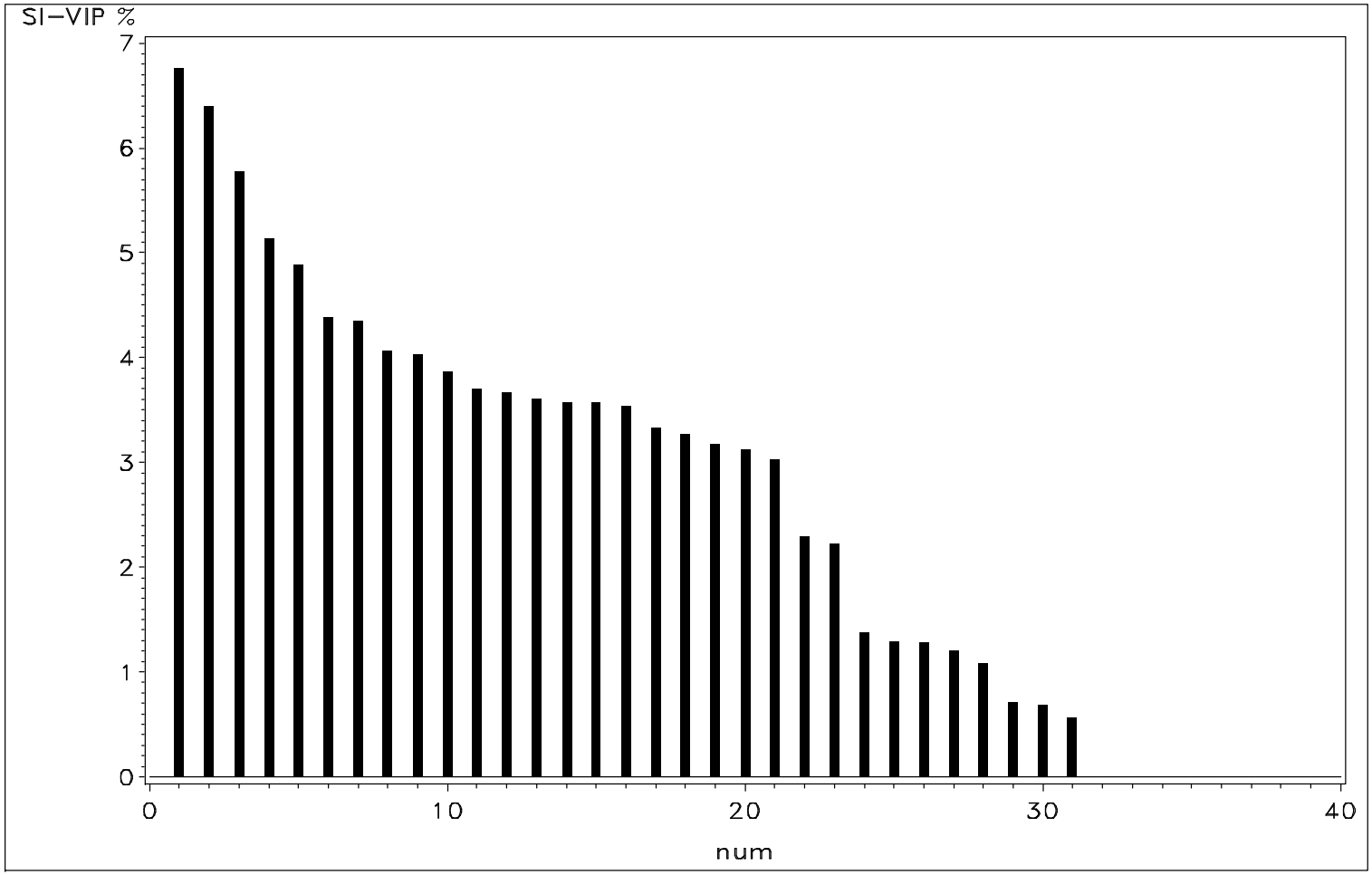
Application à un problème halieutique de l'IFREMER

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

Conclusion et perspectives

- 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