REVUE DE STATISTIQUE APPLIQUÉE

A. JOURDAN
D. COLLOMBIER

Régression trigonométrique et plans d'échantillonnage pour expériences simulées

Revue de statistique appliquée, tome 49, n° 2 (2001), p. 5-25
<http://www.numdam.org/item?id=RSA_2001__49_2_5_0>

Rev. Statistique Appliquée, 2001, **XLIX** (2), 5-25

RÉGRESSION TRIGONOMÉTRIQUE ET PLANS D'ÉCHANTILLONNAGE POUR EXPÉRIENCES SIMULÉES

A. Jourdan* et D. Collombier**

* *Université de Pau et des Pays de l'Adour,*
laboratoire de Mathématiques Appliquées, 64000 Pau
astrid.jourdan@univ-pau.fr

** *Université Louis Pasteur, IRMA, 67084 Strasbourg*
collombier@math.u-strasbg.fr

RÉSUMÉ

Un grand nombre de phénomènes physiques sont modélisés puis étudiés à l'aide de simulateurs. Ces derniers sont souvent très coûteux et donc difficilement exploitables. L'objectif de la planification des expériences simulées est d'ajuster à moindre coût un prédicteur pour la réponse du simulateur. On se limite ici aux simulateurs déterministes (la réponse du code ne dépend que des paramètres d'entrée et est donc reproductible).

Pour résumer la réponse du simulateur, on utilise ici un modèle de type géostatistique qui permet d'introduire une corrélation spatiale entre les réponses du simulateur. Les deux particularités de ce travail sont : l'utilisation d'une régression linéaire trigonométrique (ou de Fourier) – l'adaptation de plans d'expériences classiques tels que les tableaux orthogonaux linéaires pour l'échantillonnage d'expériences simulées et la construction de plans optimaux avec validation numérique.

Mots-clés : *Plans d'échantillonnage, Régression trigonométrique, Expériences simulées, Krigeage, Tableau orthogonal linéaire.*

ABSTRACT

Many scientific phenomena are now investigated by complex computer models or codes. A computer experiments is a number of runs of the code with various inputs. In this work the output of computer experiments is deterministic (rerunning the code with the same inputs give identical observation). The aim of computer experiments is to fit a predictor of the output to the data.

In order to resume the response of the code, we adopt here a geo-statistic model which permit to introduce spatial correlation between the responses of the code. The two particularities of this work are : the use of a trigonometric (or Fourier) regression – the adaptation of basic experimental designs to this kind of model and the construction of optimal designs with numerical validation.

Keywords : *Sampling designs, Trigonometric regression, Computer experiments, Kriging, Linear orthogonal arrays.*

1. Introduction

L'expérimentation de nombreux phénomènes physiques est souvent très onéreuse ou, dans certains cas, impossible à mettre en place. En revanche, certains de ces phénomènes peuvent être modélisés par des équations mathématiques. Il est alors possible de trouver une solution numérique au problème à l'aide d'un simulateur. Malgré les performances sans cesse croissantes des ordinateurs, ces simulateurs restent souvent difficiles à exploiter car très coûteux en temps. Cela est essentiellement dû au fait qu'ils prennent en compte un grand nombre de paramètres d'entrée. L'utilisateur souhaite bien souvent disposer d'un modèle simple et rapide pour résumer la réponse du code de calcul. Dans cet article le type de simulateur étudié est déterministe, c'est-à-dire que la réponse du simulateur est reproductible, *i.e* est entièrement déterminée par les paramètres de simulation. Par opposition à l'expérimentation physique, une expérience menée à l'aide d'un code de calcul est dite simulée.

La deuxième partie de l'article est consacrée à la mise en place d'un modèle de type géostatistique adapté à l'analyse les expériences simulées. La réponse déterministe du simulateur, $Y(x)$, est considérée comme la réalisation d'une fonction aléatoire formée d'une régression linéaire trigonométrique (ou de Fourier) et d'un processus gaussien permettant d'introduire une corrélation spatiale entre les réponses du simulateur.

L'originalité de ce travail réside dans la troisième partie qui consiste à adapter des outils spécifiques aux plans d'expériences classiques dans le contexte des expériences simulées.

La quatrième partie étudie de manière empirique à partir d'un exemple simple, l'efficacité et la robustesse des plans d'échantillonnage définis dans la partie 3.

2. Modélisation et Prédiction

2.1. Le modèle

Sacks *et al.* (1989) proposent d'utiliser un modèle qui traite la réponse déterministe du simulateur au point $x \in [0,1]^d$, comme la réalisation d'une fonction aléatoire, $Y(x)$, somme d'une régression linéaire et d'un processus gaussien,

$$Y(x) = X(x)\beta + \Gamma(x). \tag{1}$$

2.1.1. Le processus résiduel

$\Gamma(x)$ est un processus stochastique appelé processus résiduel, qui est supposé stationnaire gaussien d'espérance nulle et de matrice de covariance

$$\mathrm{cov}(\Gamma(x), \Gamma(w)) = \sigma^2 R(x,w),$$

où σ^2 est la variance du processus et $R(x, w)$ sa fonction de corrélation. Cette fonction est choisie stationnaire de la forme la plus classique,

$$R(x, w) = \exp(-\theta||x - w||^2),$$

où θ est le paramètre de corrélation (plus il est grand, moins il y a de corrélation et *vice-versa*), et $||\cdot||$ est la norme euclidienne de $\mathbb{R}^d$. La fonction de corrélation dépend de la distance entre les points, ce qui permet d'introduire une forte corrélation entre les réponses du simulateur si les simulations sont effectuées en des points très proches et *a contrario* une quasi-indépendance pour des points éloignés.

2.1.2. La régression linéaire trigonométrique

Reprenons les notations de Bates *et al.* (1996) et Riccomagno *et al.* (1997) pour définir la partie déterministe du modèle, $\forall x \in [0, 1]^d$,

$$E[Y(x)] = f(x) = \beta_0 + \sqrt{2} \sum_{h \in A^+} \beta_h \sin\big[2\pi(h|x)\big] + \phi_h \cos\big[2\pi(h|x)\big],$$

où $(\cdot|\cdot)$ le produit scalaire usuel de $\mathbb{R}^d$.

Il s'agit donc d'une régression trigonométrique (ou de Fourier) définie par un ensemble A^+ de vecteurs non nuls d'entiers tel que $h \in A^+ \Longrightarrow -h \notin A^+$, appelé ensemble des fréquences du modèle. Nous utilisons ici la terminologie suivante,

Effets simples : $A^+ \subseteq \{(h_1, 0, ..., 0), ..., (0, ..., 0, h_d)\}$

Interactions de deux facteurs :

$A^+ \subset \{(h_1, h_2, 0, \ldots, 0), \ldots, (0, ..., h_k, \ldots, h_j, \ldots, 0)\}$, etc..., où $h_i \neq 0$, représente l'ordre du $i^{\text{ème}}$ facteur.

Exemple 1. — Un modèle à 3 facteurs prenant en compte des effets simples des 2 premiers facteurs et une interaction du premier facteur avec le troisième, définis par les fréquences

$$A^+ = \big\{(1, 0, 0), (0, 1, 0), (1, 0, 1)\big\},$$

s'écrit, $\forall x = (x_1, x_2, x_3) \in [0, 1]^3$

$$f(x) = \beta_0 + \sqrt{2}\Big(\beta_{(1,0,0)}\sin(2\pi x_1) + \beta_{(0,1,0)}\sin(2\pi x_2) + \beta_{(1,0,1)}\sin\big(2\pi(x_1 + x_3)\big) + \phi_{(1,0,0)}\cos(2\pi x_1) + \phi_{(0,1,0)}\cos(2\pi x_2) + \phi_{(1,0,1)}\cos\big(2\pi(x_1 + x_3)\big)\Big).$$

La régression trigonométrique peut encore s'écrire sous une forme complexe

$$f(x) = \alpha_0 + \sum_{h \in A^+} \alpha_h \exp\big[2\pi i(h|x)\big] + \overline{\alpha}_h \exp\big[-2\pi i(h|x)\big] \qquad (2)$$

Les formes complexes et réelles des vecteurs du modèle et des paramètres,

$$\begin{cases} X(x) = \left[1, \left\{\sqrt{2}\sin(2\pi(x|h)\right\}_{h\in A^+}, \left\{\sqrt{2}\cos(2\pi(x|h)\right\}_{h\in A^+}\right] \\ \beta = {}^t[\beta_0, \{\beta_h\}_{h\in A^+}, \{\phi_h\}_{h\in A^+}] \end{cases}$$

$$\begin{cases} Z(x) = \left[1, \left\{\exp(2\pi i(x|h)\right\}_{h\in A^+}, \left\{\exp(-2\pi i(x|h)\right\}_{h\in A^+}\right] \\ \alpha = {}^t\left[\alpha_0, \{\alpha_h\}_{h\in A^+}, \{\overline{\alpha}_h\}_{h\in A^+}\right] \end{cases}$$

sont liées par les relations $Z(x) = X(x)H^{-1}$ et $\alpha = H\beta$ (d'où l'hypothèse $h \in A^+ \Longrightarrow -h \notin A^+$) où H est une matrice unitaire ($H^*H = I$ où $H^* = {}^t\overline{H}$)

$$H = \left(\begin{array}{c|c|c} 1 & 0 & 0 \\ \hline 0 & -iH' & H' \\ \hline 0 & iH' & H' \end{array}\right) \text{ avec } H' = \frac{1}{\sqrt{2}}\, I.$$

La régression s'écrit indifféremment sous l'une des deux formes

$$f(x) = Z(x)\alpha = X(x)\beta,$$

ce qui permet de travailler de façon théorique avec la forme complexe du modèle de régression afin de pouvoir utiliser des notions algébriques (caractères des représentations irréductibles des groupes finis (Kobilinsky (1990)) puis de passer à la forme trigonométrique pour les tests numériques.

2.2. La prédiction

La méthode utilisée classiquement pour analyser un tel modèle est connue en géostatistique sous le nom de *krigeage*. Elle a été introduite par Matheron (1963) et parallèlement et indépendamment par Henderson (1950).

Soient un plan de N points de $[0,1]^d$, $D = \{x_1, \ldots, x_N\}$, et le vecteur des observations aux points du plan, $Y_D = {}^t[Y(x_1), \ldots, Y(x_N)]$. On utilise ici comme prédicteur de la réponse $Y(x)$

$$\widehat{Y}(x) = {}^t c(x) Y_D.$$

C'est une combinaison linéaire spatiale des observations qui permet de prendre en compte le fait que plus le point x où l'on cherche à prédire la réponse du simulateur est loin des simulations effectuées, plus l'erreur de prédiction sera grande. Plus précisément, on utilise le prédicteur sans biais qui minimise l'erreur quadratique moyenne (BLUP) (*cf.* Sacks *et al.* (1989), Christensen (1990) et Robinson (1991))

$$\widehat{Y}(x) = X(x)\widehat{\beta} + {}^t r(x) R^{-1}[Y_D - X\widehat{\beta}],$$

où $\widehat{\beta} = ({}^t X R^{-1} X)^{-1}\, {}^t X R^{-1} Y_D$ est l'estimateur de Gauss-Markov de β, avec

• $X = {}^t[X(x_1), \ldots, X(X_N)]$ matrice d'ordre $N \times m$ du modèle aux points du plan,

• $R = \big(R(x_i, x_j)\big)_{i,j}$ matrice d'ordre $N \times N$ de corrélation,

• $r(x) = {}^t\big[R(x_1, x), \ldots, R(x_N, x)\big]$ vecteur des corrélations entre les points du plan et x.

Il peut encore s'écrire avec la notation complexe en utilisant la relation $X = ZH$,

$$\widehat{Y}(x) = Z(x)\widehat{\alpha} + {}^t r(x) R^{-1}[Y_D - Z\widehat{\alpha}] \tag{3}$$

où $\widehat{\alpha} = (Z^* R^{-1} Z)^{-1} Z^* R^{-1} Y_D$.

La *surface de réponse* ainsi prédite interpole les aléas observés, c'est-à-dire qu'en tout point du plan

$$\widehat{Y}(x_i) = Y(x_i).$$

Deux questions se posent alors : le choix du paramètre θ de la fonction de corrélation (*cf.* 2.1.1) et le choix du plan D.

3. Tableaux orthogonaux linéaires

Dans cette partie, seule la partie régression linéaire du modèle est prise en compte. Soit un modèle défini par son ensemble de fréquences A^+.

Le choix des points du plan d'échantillonnage se fait de la façon suivante. Chaque arrête du domaine expérimental $[0,1]^d$ est découpée en un nombre premier p de segments de même longueur que l'on numérote de 0 à $p-1$. Nous obtenons ainsi une partition du cube unité en p^d cellules (Fig. 1).

On munit l'ensemble des entiers $\{0, \ldots, p-1\}$ de la loi d'addition modulo p. Le domaine d'échantillonnage s'identifie alors à $GF(p)^d$, où $GF(p)$ est le corps de Galois d'ordre p, en associant à tout vecteur g de $GF(p)^d$, un point d'échantillonnage $x \in [0,1]^d$ tel que $x = g/p$. A tout point d'observation construit de la sorte, la régression linéaire trigonométrique (2) est de la forme

$$f(x) = \alpha_0 + \sum_{h \in A^+} \alpha_h \exp\left[i\,\frac{2\pi}{p}(h|g)\right] + \overline{\alpha}_h \exp\left[-i\,\frac{2\pi}{p}(h|g)\right], \tag{4}$$

où $(h|g)$ est le produit scalaire sur $GF(p)^d$.

Les plans d'échantillonnage utilisés dans cet article correspondent à des sous-ensembles de $GF(p)^d$ particuliers appelés tableaux orthogonaux linéaires.

Définition 2. — *Soient T un tableau d'ordre $N \times d$ à éléments dans $GF(p)$, et t un entier tel que $1 \leqslant t \leqslant d$.*

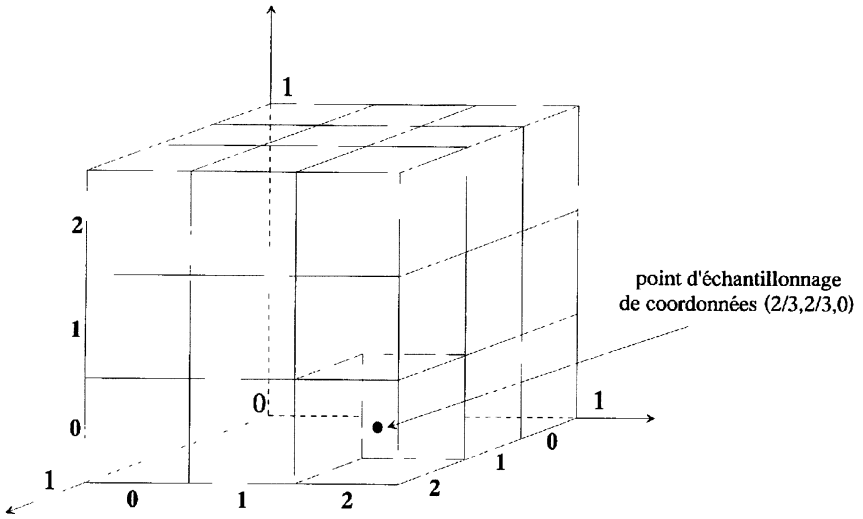


FIGURE 1
Schéma du cube unité avec GF(3)^3

T est un tableau orthogonal de force t et d'index λ, si et seulement si dans tout bloc formé de t colonnes de T, tous les éléments de $GF(p)^t$ figurent un même nombre de fois λ chacun.

T est un tableau orthogonal linéaire si de plus ses lignes sont toutes distinctes et constituent un sous-espace vectoriel de $GF(p)^d$.

Remarque 3 : Le coût de simulation N est égal à λp^t, il est donc préférable d'utiliser des tableaux d'index $\lambda = 1$.

L'ensemble S constitué des lignes de T est par définition un sous-espace vectoriel de $GF(p)^d$. Il est donc entièrement déterminé,

• soit par une matrice B d'ordre $t \times d$ formée de t vecteurs linéairement indépendants générateurs de S,

• soit par P une matrice d'ordre $(d-t) \times d$, de rang $d-t$, génératrice du dual de S, *i.e* l'ensemble $S^{\perp} = \{h \in GF(p)^d, (s|h) = 0 \; \forall s \in S\}$.

La matrice P étant de rang $d-t$, elle peut toujours s'écrire sous la forme $P = (I_{d-t}|Q)$ où Q est une matrice d'ordre $(d-t) \in t$. On a alors $B = (-{}^tQ|I_t)$.

Exemple 4. — Soit un domaine d'échantillonnage à 3 facteurs partitionné en 3^3 cellules. Soit T le tableau orthogonal linéaire à éléments dans $GF(3)$ engendré

par la matrice

$$B = \left(\begin{array}{c|cc} 2 & 0 & 1 \\ 2 & 1 & 0 \end{array}\right) \Longrightarrow T = {}^t\begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 2 \\ 0 & 1 & 2 & 1 & 0 & 2 & 2 & 0 & 1 \\ 0 & 2 & 1 & 1 & 2 & 0 & 2 & 1 & 0 \end{pmatrix}$$

Si on note $S^{\perp}$ le dual de l'ensemble S constitué des lignes de T et P la matrice génératrice de $S^{\perp}$, alors

$$P = (1 \mid 1\ 1) \quad \text{et} \quad S^{\perp} = \big\{(0,0,0),(1,1,1),(2,2,2)\big\}.$$

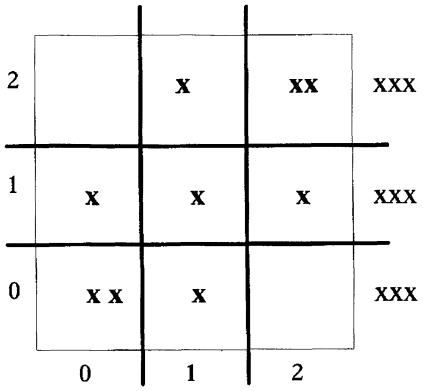
Les tableaux orthogonaux linéaires utilisés pour les tests numériques (§ 4) ont été construits à partir de la matrice génératrice du dual P, car celle-ci présente deux propriétés essentielles comme nous le verrons dans le deux paragraphes suivants. Elle permet, premièrement de s'assurer que les points d'échantillonnage représentent bien toutes les parties du cube unité, et deuxièmement de vérifier que les plans définis par P permettent de construire le BLUP.

3.1. Répartition des points d'échantillonnage dans le cube unité

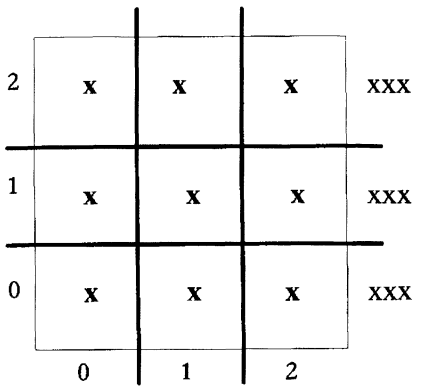
Les plans habituellement utilisés pour ajuster le modèle (1) sont construits de façon à optimiser des critères de qualité (IMSE, entropie, distance maximin,... *cf.* Sack *et al.* (1989)). Ils dépendent donc entièrement du modèle choisi, notamment du paramètre de corrélation θ. Pour éviter cet inconvénient, nous utilisons ici des plans construits indépendamment du modèle qui visent à ce que les points d'échantillonnage « remplissent » bien le cube unité (space filling designs), afin que toutes les parties de $[0,1]^d$ soient testées par les simulations. Les tableaux orthogonaux linéaires de force t entrent tout à fait dans ce cadre puisque par définition ils possèdent une bonne répartition de leurs points lorsqu'on les projette en dimension t (restreints à t facteurs). Prenons l'exemple de deux tableaux d'ordre 9×3 (3 facteurs et 9 points). Ces tableaux sont représentés sur la figure 2 projetés sur 2 facteurs puis sur 1. Celui de droite est de force 2 d'index 1 et celui de gauche est de force 1 d'index 3.

Si on replace les points des deux tableaux en dimension trois, on peut facilement imaginer que le tableau de force 2 représente uniformément le domaine d'échantillonnage alors que le tableau de force 1 laisse certaines parties du cube unité non représentées, donc non testées par les simulations. On peut supposer que plus la force du tableau est élevée, mieux les points d'échantillonnage sont répartis dans le domaine, et ainsi améliorer la qualité de la prédiction. Cette hypothèse sera confirmée avec les tests numériques (§ 4). Les résultats suivants permettent de vérifier rapidement à partir de la matrice P, si un tableau orthogonal linéaire est de force t (*cf.* Bailey (1985)).

Définition 5. — *On appelle poids d'un vecteur à éléments dans $GF(p)$, le nombre de coordonnées non nulles et poids d'une matrice, le plus petit poids des combinaisons linéaires de ses lignes.*



Tableaux orthogonal de force 1
sur une face du cube unité



Tableaux orthogonal de force 2
sur une face du cube unité

$$T = {}^{t}\begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 2 & 2 & 1 & 2 \\ 0 & 0 & 1 & 0 & 1 & 1 & 2 & 2 & 2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 \end{pmatrix} \quad T = {}^{t}\begin{pmatrix} 0 & 0 & 1 & 1 & 1 & 2 & 2 & 0 & 2 \\ 0 & 1 & 0 & 1 & 2 & 1 & 0 & 2 & 2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 2 & 2 & 1 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

FIGURE 2

Proposition 6. — *Un tableau orthogonal linéaire de matrice génératrice du dual P est de force $t-1$ si et seulement si P est de poids supérieur ou égal à t.*

Lemme 7. — *La matrice $P = (I_{d-t}|Q)$ ci-dessus est de poids supérieur ou égal à t si et seulement si $\forall k = 1, \ldots, \min(d-t, t)$ toute combinaison linéaire de k lignes de Q est de poids supérieur ou égal à $t-k$.*

Exemple 4 (suite). — Dans l'exemple 4 le tableau orthogonal linéaire T à éléments dans $GF(3)$ est défini par sa matrice génératrice du dual $P = (1|1\ 1)$. D'après la proposition 6 et le lemme 7, il est clair que le tableau T est de force 2 puisque P est de poids 3.

Pour p et la force t fixés, les tableaux orthogonaux linéaires utilisés dans les tests numériques ont été construits à l'aide de ces résultats, notamment du lemme 7 qui permet d'établir un algorithme de construction simple et moins coûteux.

3.2. Estimabilité et optimalité

Le meilleur prédicteur linéaire sans biais (3) n'existe pas pour tous les plans d'échantillonnage. En effet, il faut (et il suffit) que les colonnes de la matrice du modèle Z soient indépendantes (condition d'estimabilité de α par $\widehat{\alpha}$). Nous allons maintenant voir que dans le cas d'une régression linéaire trigonométrique et d'un tableau orthogonal linéaire la condition d'existence du prédicteur (3) s'exprime ici aussi à l'aide de la matrice P, et plus précisément grâce au dual $S^{\perp}$ engendré par P.

Les propriétés suivantes sont des résultats classiques que l'on peut retrouver par exemple dans les ouvrages de Collombier (1996) ou Christensen (1990).

Proposition 8. — *Pour un tableau orthogonal linéaire et une régression linéaire trigonométrique, les colonnes de la matrice du modèle sont, soit 2 à 2 identiques, soit 2 à 2 orthogonales (pour le produit scalaire usuel de $\mathbb{C}^d$).*

Le prédicteur linéaire sans biais de $Y(x)$ existe $\forall x$, si et seulement si les colonnes de la matrice du modèle sont 2 à 2 orthogonales.

La matrice d'information Z^*Z représente les produits scalaires de toutes les colonnes de la matrice du modèle. Pour que le prédicteur linéaire sans biais existe, il faut donc qu'elle soit multiple de la matrice identité. Le plan est alors universellement optimal pour le modèle en espérance (*i.e.* en l'absence de corrélation). Le lemme suivant est établi à partir de la régression trigonométrique (4), et donne la forme de la matrice d'information.

Lemme 9. — *Soient un modèle défini par son ensemble de fréquences A^+, et $S = \{s_1, \ldots, s_N\}$ l'ensemble constitué des lignes de T un tableau orthogonal linéaire à éléments dans $GF(p)$ de taille N. La matrice d'information s'écrit*

$$Z^*Z = \left(\begin{array}{c|c|c} N & a & a \\ \hline a^* & B & C \\ \hline {}^t a & {}^t C & {}^t B \end{array}\right)$$

où pour $h \in A^+$ et $n \in A^+$

$$a_h = \sum_{k=1}^{N} \exp\left[i\,\frac{2\pi}{p}(s_k|h)\right], \quad B_{hn} = \sum_{k=1}^{N} \exp\left[i\,\frac{2\pi}{p}(s_k|n-h)\right],$$
$$C_{hn} = \sum_{k=1}^{N} \exp\left[i\,\frac{2\pi}{p}\,(s_k|n+h)\right].$$

Donc le prédicteur de $Y(x)$ existe pour tout x si et seulement si $a_h = 0$, $B_{hn} = 0$ $h \neq n$ et $C_{hn} = 0$. Le résultat qui suit s'obtient facilement en utilisant les caractères de représentations linéaires des groupes finis, et permet de vérifier si ces conditions sont satisfaites.

Lemme 10. — *Soit $S = \{s_1, \ldots, s_N\}$ l'ensemble constitué des lignes de T un tableau orthogonal linéaire à éléments dans $GF(p)$ de taille N. Alors $\forall h \in GF(p)^d$,*

$$\sum_{k=1}^{N} \exp\left[i\,\frac{2\pi}{p}(s_k|h)\right] = \begin{cases} N & \text{si } h \in S^{\perp} \\ 0 & \text{sinon} \end{cases}$$

Ces deux lemmes permettent d'établir une condition nécessaire et suffisante pour l'existence du prédicteur, qui lie le dual du tableau $S^{\perp}$ et l'ensemble A^{+} des fréquences du modèle.

Définition 11. — *Soient h et m deux fréquences non nulles du modèle. On dit que les fréquences sont mutuellement orthogonales par rapport à S, si*

$$\begin{cases} h \notin S^{\perp},\ m \notin S^{\perp} \\ (h+m) \notin S^{\perp} \\ (h-m) \notin S^{\perp} \text{ pour } h \neq m \end{cases}$$

Remarques 12 :

1) $h \notin S^{\perp}$ est équivalent à $(-h) \notin S^{\perp}$ car $S^{\perp}$ est un sous-espace vectoriel de $GF(p)^d$.

2) Si $p > 2$ alors la condition $(h+m) \notin S^{\perp}$ n'est à vérifier que pour $m \neq h$.

Théorème 13. — *Soient un modèle défini par son ensemble de fréquences A^{+} et S un ensemble formé des lignes d'un tableau orthogonal linéaire. Le prédicteur linéaire sans biais de $Y(x)$ existe pour tout x, si et seulement si toutes les fréquences de A^{+} sont mutuellement orthogonales par rapport S.*

Remarque 14 : Ce résultat peut s'étendre à une classe plus large de plans d'échantillonnage, notamment aux fractions régulières (*cf.* Collombier (1996) pour la définition) qui permettent d'avoir un découpage du cube unité asymétrique, et avec un nombre de segments non plus premier mais puissance d'un nombre premier.

Exemple 4 (suite). — Dans l'exemple 4 le tableau orthogonal linéaire T à éléments dans $GF(3)$ est défini par sa matrice génératrice du dual $P = (1|1\ 1)$ et son dual est égal à $S^{\perp} = \{(0,0,0),(1,1,1),(2,2,2)\}$. On peut alors vérifier facilement qu'il est possible d'analyser avec ce tableau les effets simples des 3 facteurs avec les fréquences $A^{+} = \{(1,0,0),(0,1,0),(0,0,1)\}$. En revanche, il n'est pas possible d'ajouter l'interaction (0,1,1) des 2 derniers facteurs car alors les fréquences (1,0,0) et (0,1,1) ne sont pas mutuellement orthogonales. Si on veut introduire une interaction des 2 derniers facteurs dans le modèle, il faut utiliser la fréquence (0,1,2) ou (0,2,1), mais cela ne fait pas intervenir les deux facteurs de la même façon.

Là encore on remarque que la condition d'existence du prédicteur qui lie la régression linéaire trigonométrique et le plan d'échantillonnage se fait au travers la matrice génératrice du dual P. A partir du théorème 13, de la proposition 6 et du lemme 7, un programme Matlab a été conçu afin de construire tous les tableaux orthogonaux linéaires existants pour des paramètres p et t fixés, et de sélectionner ceux qui vont permettre d'ajuster une régression trigonométrique définie par son ensemble de fréquences A^{+}. La question qui se pose maintenant est le choix du ou des meilleurs plans parmi ceux sélectionnés et suivant quel critère. Nous allons tenter d'apporter des éléments de réponse dans la partie suivante.

4. Tests numériques

A partir d'un exemple simple, nous allons étudier l'efficacité des plans d'échantillonnage. On se fixe une fonction connue,

$$\frac{\sum_{h\in A^+} \cos(h|x) + \sin(h|x)}{1 + \sum_{h\in A^+} (h|x)}, \quad \forall x \in [0,1]^d,$$

et une régression trigonométrique définie par l'ensemble des fréquences, $A^+ = \{(1,0,\ldots,0),\ldots,(0,\ldots,0,1)\}$.

Ce choix a été fait de façon à s'assurer que le modèle en espérance est approprié à la prédiction de f puisque là n'est pas notre propos.

Afin de mesurer la différence entre la fonction et le prédicteur, nous allons utiliser l'erreur quadratique moyenne intégrée empirique (EISE) sur une grille de 10^d points de $[0,1]^d$ (Sacks *et al.* (1989)) pour un paramètre de corrélation θ fixé,

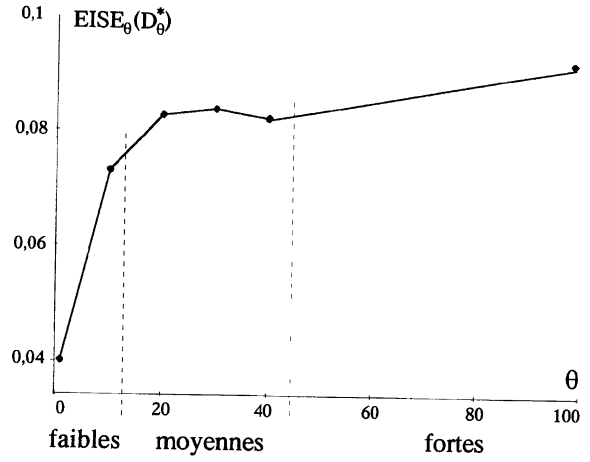
$$\text{EISE}_\theta = \frac{1}{10^d} \sum_{i=1}^{10^d} \left(\widehat{Y}(x_i) - f(x_i)\right)^2$$

Pour p et t fixés, le ou les meilleurs plans sont alors ceux qui minimisent cette erreur dans la classe des tableaux orthogonaux linéaires.

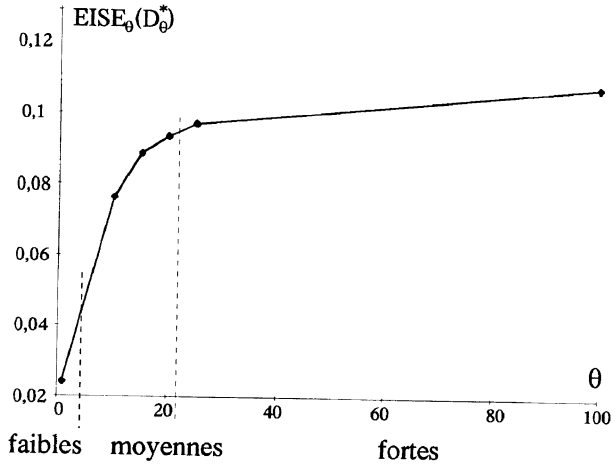
4.1. Robustesse des meilleurs plans par rapport aux variations de θ

Des premiers tests ont montré que, sauf exception, il n'existe pas de plan qui minimisent l'erreur quel que soit θ. En fait, on constate qu'il existe 3 types de plans, ceux qui minimisent l'erreur pour les faibles, les moyennes et les fortes corrélations (Fig. 3). La force de la corrélation correspond à des valeurs de θ qui dépendent de $N = p^t$ et d. On note D^*_θ le plan qui minimise EISE_θ où θ représente l'une des 3 catégories, faibles, moyennes ou fortes corrélations. Pour certains paramètres p et t, on a seulement détecté que deux catégories de corrélation faibles et fortes, voire même dans certains cas une seule catégorie, ce qui signifie alors que le meilleur plan D^*_θ minimise l'erreur quel que soit θ.

Notre but n'est pas de déterminer quel paramètre de corrélation choisir, il existe des techniques d'estimation pour cela (Mardia et Marshall (1984), Christensen (1990)). On peut toutefois constater sur cet exemple (Fig. 3), que l'erreur diminue lorsque les corrélations augmentent (θ diminue), avec un seuil inférieur pour θ (en général 1) en dessous duquel la matrice R est numériquement instable, et un seuil supérieur (en général 100) au-dessus duquel R devient proche de la matrice identité, donc plus de corrélation. Nous avons choisi ici d'étudier la robustesse des meilleurs plans des 3 catégories lorsque θ varie. Ceci est une propriété non négligeable



$N = 25$ et $d = 3$



$N = 49$ et $d = 5$

FIGURE 3
Évolution de l'erreur en fonction de θ

étant donné les problèmes qu'engendre l'estimation du paramètre de corrélation en pratique. Pour cela on calcule le coefficient d'efficacité (Sacks *et al.* (1989)) (Tab. 1),

$$\mathrm{EISE}_\theta(D^*_\theta)/\mathrm{EISE}_\theta(D^*_{\theta_i}), \tag{5}$$

où θ_i appartient à une des 2 autres catégories que θ.

TABLEAU 1

Tableaux d'efficacité des meilleurs plans lorsque θ_i varie

$N = 25$ et $d = 3$

θ / θ_i	Fortes	Moyennes	Faibles
Fortes	100 %	82 %	80 %
Moyennes	95 %	100 %	80 %
Fortes	7 %	53 %	100 %

$N = 49$ et $d = 4$

θ / θ_i	Fortes	Moyennes	Faibles
Fortes	100 %	100 %	72 %
Moyennes	95 %	100 %	80 %
Fortes	22 %	61 %	100 %

On remarque alors que $\boldsymbol{D_\theta^*}$, **le meilleur plan pour une forte ou moyenne corrélation, est très robuste aux changements de** θ. *A contrario*, l'efficacité de D_θ^*, le meilleur plan pour une faible corrélation, devient catastrophique si on utilise une petite valeur de θ. Donc quand on ne connaît pas la valeur de θ *a priori*, il est préférable d'utiliser les meilleurs plans des fortes et des moyennes corrélations. Les matrices génératrices de ces tableaux orthogonaux linéaires sont données dans les tableaux 2, 3 et 4 pour $d = 3$,4 et 5. Il suffit alors de multiplier cette matrice par la matrice d'ordre $p^d \times d$ formée des vecteurs de $GF(p)^d$ pour obtenir le tableau orthogonal linéaire.

Dans les tableaux 2 à 4, $C_r = N/(2d+1)$ représente le coût relatif, *i.e* la taille du plan divisée par le nombre de paramètres à estimer dans la régression trigonométrique.

TABLEAU 2

Matrices génératrices des meilleurs plans à 3 facteurs pour les fortes et moyennes corrélations

Caractéristiques				*Fortes corrélations*		*Moyennes corrélations*	
t	N	p	Cr	Efficacité	D_θ^*	Efficacité	D_θ^*
2	9	3	1.3		$\begin{pmatrix}2&1&0\\2&0&1\end{pmatrix}$ (*)		
2	25	5	3.6	100% / Moy 72% / Faibles	$\begin{pmatrix}1&1&0\\2&0&1\end{pmatrix}$ $\begin{pmatrix}1&1&0\\2&0&1\end{pmatrix}$ $\begin{pmatrix}2&1&0\\3&0&1\end{pmatrix}$	95% / Fortes 80% / Faibles	$\begin{pmatrix}3&1&0\\3&0&1\end{pmatrix}$ $\begin{pmatrix}2&1&0\\4&0&1\end{pmatrix}$
2	49	7	7	81% / Faibles	$\begin{pmatrix}3&1&0\\5&0&1\end{pmatrix}$	Pas de moyenne corrélation	

(*) Pour ces paramètres on ne détecte qu'une seule catégorie, *i.e* que D_θ^* minimise l'erreur quel que soit θ. On ne peut donc pas calculer le coefficient d'efficacité (5).

TABLEAU 3

Matrices génératrices des meilleurs plans à 4 facteurs pour les fortes et moyennes corrélations

Caractéristiques				*Fortes corrélations*		*Moyennes corrélations*	
t	N	p	Cr	Efficacité	D_0^*	Efficacité	D_θ^*
2	25	5	2.8	99% / Moy 93% / Faibles	$\begin{pmatrix}2&3&1&0\\3&4&0&1\end{pmatrix}$	80% / Fortes 96% / Faibles	$\begin{pmatrix}2&3&1&0\\3&4&0&1\end{pmatrix}\begin{pmatrix}1&3&1&0\\2&4&0&1\end{pmatrix}\begin{pmatrix}1&2&1&0\\3&4&0&1\end{pmatrix}$ $\begin{pmatrix}2&2&1&0\\2&4&0&1\end{pmatrix}\begin{pmatrix}1&3&1&0\\3&3&0&1\end{pmatrix}\begin{pmatrix}1&2&1&0\\2&3&0&1\end{pmatrix}$
3	27	3	3	100% / Moy 87% / Faibles	$\begin{pmatrix}1&1&0&0\\1&0&1&0\\2&0&0&1\end{pmatrix}$	96% / Fortes 90% / Faibles	$\begin{pmatrix}1&1&0&0\\1&0&1&0\\1&0&0&1\end{pmatrix}\begin{pmatrix}1&1&0&0\\2&0&1&0\\2&0&0&1\end{pmatrix}$
2	49	7	5.4	82% / Moy 80% / Faibles	$\begin{pmatrix}3&2&1&0\\5&6&0&1\end{pmatrix}\begin{pmatrix}4&3&1&0\\4&5&0&1\end{pmatrix}$	95% / Fortes 80% / Faibles	$\begin{pmatrix}4&2&1&0\\4&6&0&1\end{pmatrix}$
3	125	5	13.9	98% / Moy 76% / Faibles	$\begin{pmatrix}1&1&0&0\\2&0&1&0\\3&0&0&1\end{pmatrix}$	61% / Fortes 82% / Faibles	$\begin{pmatrix}3&1&0&0\\3&0&1&0\\4&0&0&1\end{pmatrix}\begin{pmatrix}2&1&0&0\\2&0&1&0\\4&0&0&1\end{pmatrix}$

TABLEAU 4

Matrices génératrices des meilleurs plans à 5 facteurs pour les fortes et moyennes corrélations

Caractéristiques				*Fortes corrélations*		*Moyennes corrélations*	
t	N	p	Cr	Efficacité	D_0^*	Efficacité	D_θ^*
2	25	5	2.3	87% / Faibles	$\begin{pmatrix}3&2&3&1&0\\3&4&4&0&1\end{pmatrix}$	Pas de moyenne corrélation	
2	27	3	2.4	99% / Faibles	$\begin{pmatrix}1&2&1&0&0\\1&2&0&1&0\\2&0&0&0&1\end{pmatrix}$	Pas de moyenne corrélation	
2	49	7	4.4	95% / Moy. 90% / Faibles	$\begin{pmatrix}2&1&3&1&0\\3&4&6&0&1\end{pmatrix}$	80% / Fortes 89% / Faibles	$\begin{pmatrix}2&1&2&1&0\\3&4&6&0&1\end{pmatrix}\begin{pmatrix}2&1&3&1&0\\2&3&5&0&1\end{pmatrix}$
4	81	3	7.3	91% / Faibles	$\begin{pmatrix}1&1&0&0&0\\1&0&1&0&0\\2&0&0&1&0\\2&0&0&0&1\end{pmatrix}\begin{pmatrix}1&1&0&0&0\\1&0&1&0&0\\1&0&0&1&0\\2&0&0&0&1\end{pmatrix}$	Pas de moyenne corrélation	
3	125	5	11.3	82% / moy 82% / Faibles	$\begin{pmatrix}1&1&1&0&0\\3&1&0&1&0\\3&4&0&0&1\end{pmatrix}\begin{pmatrix}1&1&1&0&0\\2&1&0&1&0\\2&3&0&0&1\end{pmatrix}$	82% / Fortes 95% / Faibles	$\begin{pmatrix}1&2&1&0&0\\2&3&0&1&0\\3&4&0&0&1\end{pmatrix}\begin{pmatrix}1&2&1&0&0\\3&2&0&1&0\\3&4&0&0&1\end{pmatrix}$

4.2. Influence du paramètre p et de la force du tableau sur l'erreur

A l'issue de ces tests, on remarque que globalement l'erreur diminue lorsque le nombre de points du plan augmente, mais surtout que cette diminution dépend de la valeur de p. En effet, **plus le découpage du domaine expérimental est fin (p grand), plus l'erreur est petite et ceci à moindre coût**. Prenons par exemple le cas des plans à 5 facteurs (Fig. 4).

Pour $N = 25$, 27 et 81, nous avons les plus petites erreurs que l'on puisse trouver quel que soit θ, ce qui veut dire qu'avec un plan de 27 points et $p = 3$, on commet une erreur 6,3 fois plus grande que pour un plan pratiquement de même taille, $N = 25$, mais avec $p = 5$. Pour $N = 49$ il se peut que l'erreur soit inférieure à 0,0243 avec un θ plus petit, mais cela ne change rien au fait qu'il est préférable d'utiliser ce plan que celui 2 fois plus grand, de taille 81 avec $p = 3$.

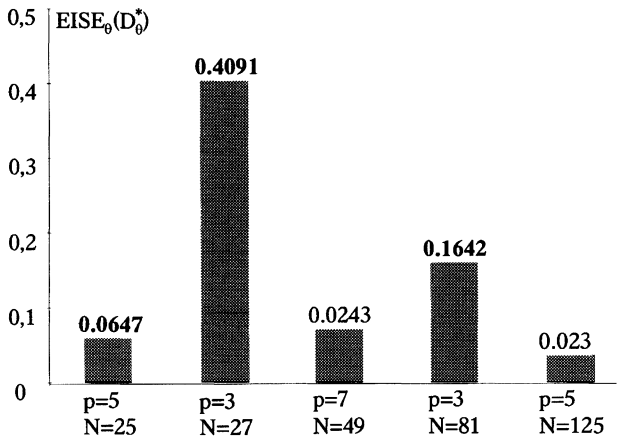


FIGURE 4
Évolution de l'erreur en fonction de p pour les plans à 5 facteurs.

Une fois que p est fixé, il reste à choisir t, la force du tableau orthogonal linéaire. On remarque alors que, **pour une taille fixée, N, les plans qui minimisent l'erreur sont toujours de force maximale et ceci quel que soit** θ. Attention toutefois, un tableau de force maximale n'est pas nécessairement d'erreur minimale. Ces résultats confirment l'intuition que nous avions eu dans le paragraphe § 3.2, c'est-à-dire qu'une bonne répartition marginale des points du plan est un critère de qualité.

La taille du plan est égale à $N = p^t$, il faut donc trouver un bon compromis entre les deux paramètres de façon à garder un coût de simulation raisonnable. Nous conseillons alors **de privilégier la finesse de découpage du cube unité (p grand) à la force du tableau orthogonal t**. En effet, comme on peut le voir sur la figure 4 et le tableau 4, il est préférable d'utiliser le tableau de taille 49, de force 2, au tableau de taille 81 et de force 4.

4.3. Test des critères de sélection des plans

En pratique il est impossible de sélectionner un plan à l'aide de la EISE puisque l'on ne connaît pas la réponse du simulateur aux points de la grille. Même avec l'exemple que nous traitons, calculer l'erreur sur 10^d points devient très long quand $d > 5$. Dans la bibliographie on trouve des critères pour sélectionner les plans bien plus simples à calculer, et surtout qui ne dépendent pas du modèle en espérance.

Le critère d'entropie (Schewry et Wynn (1987), Currin *et al.* (1991)) qui permet de mesurer la quantité d'information fournie par la simulation. De façon simplifiée, ce critère se réduit à la recherche du plus grand déterminant de la matrice de corrélation.

$$\max : \det(R)$$

Le critère de distance est basé sur la distance entre les points du plan (maximin distance) (Johnson *et al.* (1990)) de façon à trouver la meilleure répartition spatiale. Soit $||x||$ la norme euclidienne.

$$\max : \min_{1 \leqslant i < j \leqslant N} ||x_i - x_j||$$

Nous avons testé l'efficacité et la particularité de ces critères avec les tableaux orthogonaux linéaires à 3, 4 et 5 facteurs.

On remarque 2 choses.

Premièrement, il existe un lien évident entre la force du tableau et les critères. **Pour les plans de taille fixée, plus t est important et plus la distance ou le déterminant est grand** (Fig. 5, 6, 7 et 8).

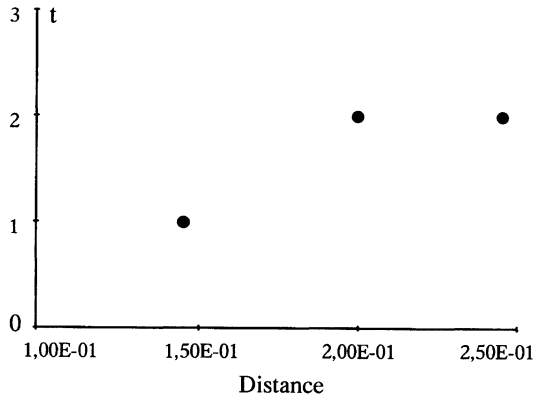


FIGURE 5
Force en fonction de la distance pour les plans de taille 49 à 3 facteurs

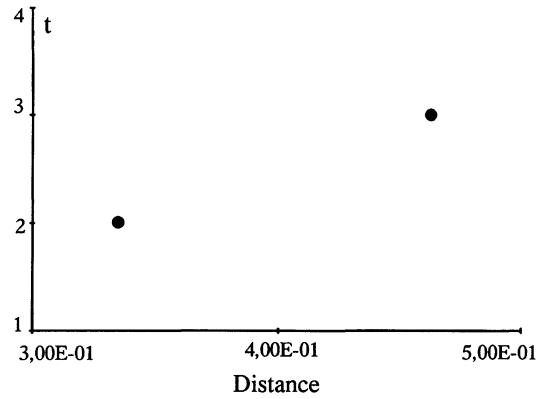


FIGURE 6
Force en fonction de la distance pour les plans de taille 27 à 4 facteurs

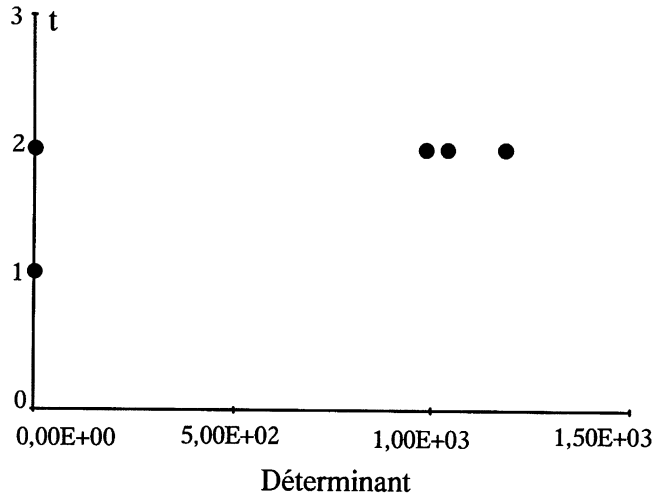


FIGURE 7
Force en fonction du déterminant pour les plans de taille 25 à 3 facteurs

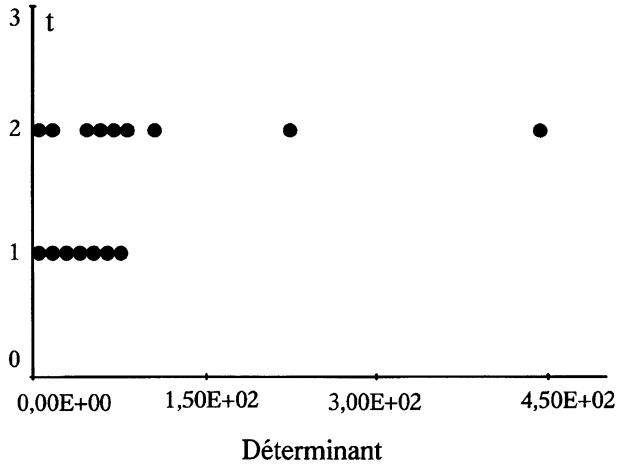


FIGURE 8
Force en fonction du déterminant pour les plans de taille 25 à 5 facteurs

De plus, pour les tableaux de même taille et de force maximale, un plan de déterminant maximal est toujours de distance maximale (Fig. 9 et 10).

Deuxièmement, les plans sélectionnés par ces critères ne minimisent pas l'erreur EISE. On est donc en droit de se demander s'ils sont tout de même efficaces par

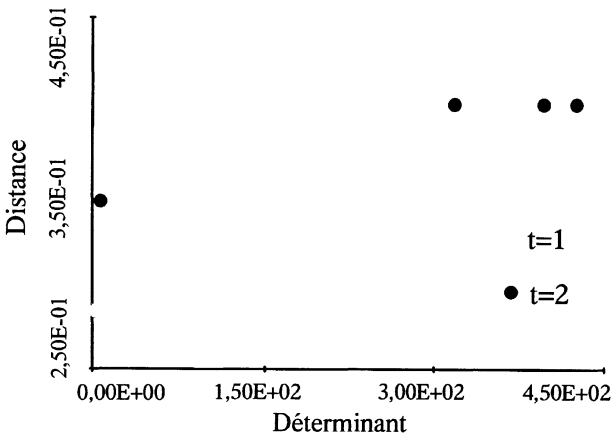


FIGURE 9
Distance en fonction du déterminant pour les plans de taille 25 à 4 facteurs

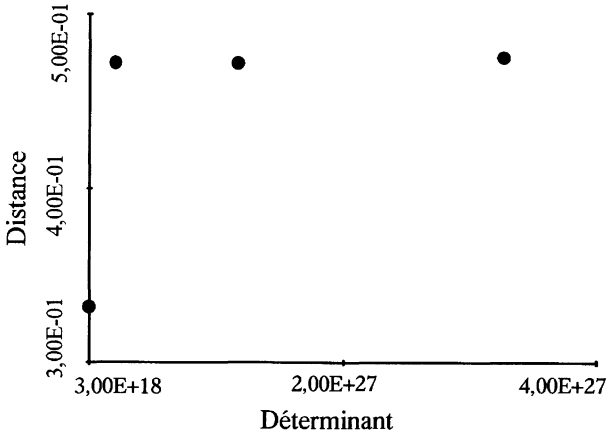


FIGURE 10
Distance en fonction du déterminant pour les plans de taille 27 à 5 facteurs

rapport aux meilleurs plans. Les tableaux 5, 6 et 7 donnent les matrices génératrices des tableaux orthogonaux linéaires de force maximale et de déterminant maximal, notés D^*_{det}, ainsi que leurs coefficients d'efficacité par rapport aux meilleurs plans des 3 catégories.

TABLEAU 5

Tableau d'efficacité des plans à 3 facteurs de déterminant maximal

N	D_{det}^*	Fortes corrélations	Faibles corrélations
9	$\begin{pmatrix}2&1&0\\2&0&1\end{pmatrix}$	100% (**)	
25	$\begin{pmatrix}2&1&0\\4&0&1\end{pmatrix}\begin{pmatrix}3&1&0\\3&0&1\end{pmatrix}$	95%	80%
49	$\begin{pmatrix}3&1&0\\5&0&1\end{pmatrix}$	100%	81%

(**) Il n'y a qu'une seule catégorie de corrélation pour $N = 9$ et $d = 3$ (*cf.* Tab. 2)

TABLEAU 6

Tableau d'efficacité des plans à 4 facteurs de déterminant maximal

N	D_{det}^*	Fortes corrélations	Faibles corrélations
25	$\begin{pmatrix}3&2&1&0\\3&4&0&1\end{pmatrix}$	100%	93%
27	Tous les plans ont même déterminant		
49	$\begin{pmatrix}1&4&1&0\\3&6&0&1\end{pmatrix}\begin{pmatrix}2&2&1&0\\2&5&0&1\end{pmatrix}$	59%	86%
125	$\begin{pmatrix}3&1&0&0\\3&0&1&0\\4&0&0&1\end{pmatrix}\begin{pmatrix}2&1&0&0\\2&0&1&0\\4&0&0&1\end{pmatrix}$	61%	82%

TABLEAU 7

Tableau d'efficacité des plans à 5 facteurs de déterminant maximal

N	D_{det}^*	Fortes corrélations	Faibles corrélations
25	$\begin{pmatrix}3&2&3&1&0\\3&4&4&0&1\end{pmatrix}$	100%	89%
27	$\begin{pmatrix}1&2&1&0&0\\2&2&0&1&0\\2&0&0&0&1\end{pmatrix}\begin{pmatrix}1&1&1&0&0\\1&2&0&1&0\\1&0&0&0&1\end{pmatrix}$	94%	100%
49	$\begin{pmatrix}4&2&4&1&0\\4&6&6&0&1\end{pmatrix}\begin{pmatrix}1&4&2&1&0\\2&4&6&0&1\end{pmatrix}$	91%	91%
81	Tous les plans ont même déterminant		
125	$\begin{pmatrix}1&2&1&0&0\\2&3&0&1&0\\3&4&0&0&1\end{pmatrix}\begin{pmatrix}1&2&1&0&0\\3&2&0&1&0\\3&4&0&0&1\end{pmatrix}$	62%	93%

On constate que même si le critère d'entropie ne donne pas les meilleurs plans, il est tout de même intéressant puisque les plans sélectionnés sont souvent très efficaces, et surtout, ils ne dépendent pas de la régression trigonométrique choisie. On peut donc espérer qu'ils seront aussi efficaces pour n'importe quel modèle en espérance

En conclusion, nous avons mis à jour une catégorie de plans, les tableaux orthogonaux linéaires, très intéressante, car très robustes aux variations de θ. Nous avons vu de plus que la répartition marginale des points du plan est un critère de qualité de ces tableaux. Il nous reste à étendre cette étude à une classe plus large de

plans telle que les fractions régulières, de façon à permettre un découpage du cube unité asymétrique et en un nombre de segments, non plus premier, mais puissance d'un nombre premier.

La stratégie préconisée est donc de coupler le critère de répartition marginale et d'entropie, ce qui permet de sélectionner (parmi les tableaux de même taille) des plans de bonne qualité et efficaces quel que soit le modèle.

Références

BAILEY R.A. (1985). Factorial design and abelian groups. *Linear Algebra Appl.* **70**, 349-368.

BATES R.A., BUCK R.J., RICCOMAGNO E., WYNN H.P. (1996). Experimental design and observation for large systems. *J. Roy. Statist. Soc. B* **58**, 77-94.

BATES R.A., RICCOMAGNO E., SCHWABE R., WYNN H.P. (1998). Lattice and dual lattice in optimal experimental design for Fourier models. *Comp. Statist. Data and Analysis* **28** N°3, 283-296.

CHRISTENSEN R. (1990). *Linear models for multivariate, time series, and spatial data.* Springer-Verlag.

COLLOMBIER D. (1996). *Plans d'expérience factoriels.* Springer-Verlag, Berlin.

CURRIN C.T., MITCHELL M., MORRIS M., YLVISAKER D. (1991). Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *J. Amer. Statist. Assoc.* **86**, 953-963.

HENDERSON C.R. (1950). Estimation of genetic parameters. *Ann. Math. Statist.* **21**, 309-310 (abstract).

JOHNSON M.E., MOORE L.M, YLVISAKER D. (1990). Minimax and maximin distance designs. *J. Statist. Planning and Inference* **26**, 131-148.

KOBILINSKY A. (1990). Complex linear models and cyclic designs. *Linear Algebra Appl.*, **127**, 227-282.

KOEHLER J. R., OWEN A.B. (1996). Computer experiments. Dans Ghosh S. et Rao C.R. : *Handbook of statistics, 13 : Design and analysis of experiments.* North-Holland, 261-308.

MARDIA K.V., MARSHALL R.J. (1984). Maximum likelihood estimation of models for reidual covariance in spatial regression. *Biometrika* **71**, 135-146.

MATHERON G. (1963). Principles of geostatistics. *Economic Geology* **58**, 1246-1266.

PARK J.-S. (1994). Optimal latin hypercube designs for computer experimnts. *J. Statist. Planning and Inference* **39**, 95-111.

RICCOMAGNO E., SCHWABE R., WYNN H.P. (1997). Lattice-based D-optimum design for Fourier regression. *Annals of Statist.* **24**, 2313-2327.

ROBINSON G.K. (1991). That BLUP is a good thing : the estimation of random effects. *Statist. Science.* **6**, 15-51.

SACKS J., SCHILLER S.B., WELCH W.J. (1989). Designs for computer experiments. *Technometrics* **31**, 41-47.

SACKS J., WELCH W.J., MITCHELL T.J., WYNN H.P. (1989). Design and analysis of computer experiments. *Statist. Science* **4**, 409-435.

SERRES J.P. (1967). *Représentations linéaires des groupes finis.* Herman, Paris.

SHEWRY M.C., WYNN H.P. (1987). Maximum entropy sampling. *J. Appl. Statist.* **14**, 165-170.

TANG B. (1994). A theorem for selecting OA-based latin hypercubes using a distance criterion. *Comm. Statist. – Theory Meth.* **27**, 2047-2058.