Comment bien choisir des points dans un espace à beaucoup de dimensions?

Jean-Marc AZAÏS

Modélisation **A**léatoire à **F**inalité **I**ndustrielle et **A**ppliquée, LSP, IMT

Outline

La théorie classique des plans d'expériences donne principalement deux types de plans sur des domaines continus : hyper-rectangles, hyper-sphères etc...

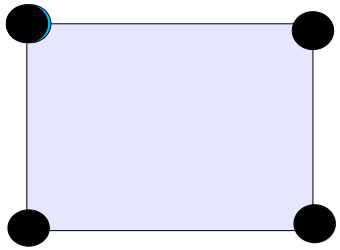
Il s'agit des plans **Isovariants** et des plans **D-optimaux**.

Nous allons d'abord présenter ces méthodes et montrer comment on peut les adapter aux expériences numériques.

Plans isovariants

Une variable Y est influencée par un certain nombre de variables quantitatives

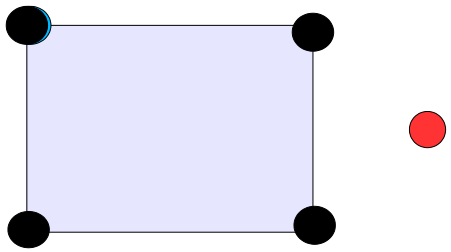
Par exemple : la pollution d'un moteur en fonction des réglages du calculateur embarqué. On suppose en général un modèle quadratique et on fait un certain nombre de mesures



Plans isovariants

Une variable Y est influencée par un certain nombre de variables quantitatives

Par exemple : la pollution d'un moteur en fonction des réglages du calculateur embarqué. On suppose en général un modèle quadratique et on fait un certain nombre de mesures



Ensuite, on veut estimer la réponse en un nouveau point et on veut que la variance de l'estimation ne dépende pas de la direction.

L'isovariance est souvent associée à des conditions qui font que la variance ne dépend pas beaucoup de la distance. **C'est donc une condition minimax** qui garantit que la pire variance est faible.
Si $Z^{(1)}, ..., Z^{(m)}$ sont les m régresseurs et si on définit les différents moments associés au plan par :

$$[\delta_1, \cdots, \delta_m] := \frac{1}{n} \sum_{i=1}^{n} (Z_i^{(1)})^{\delta_1} \times \cdots \times (Z_i^{(m)})^{\delta_m},$$

On a alors (Box et Hunter 1957) Le plan est isovariant si et seulement si

1. Tous les moments ayant un coef δ impair sont nuls
2. Les moments de type $[4, 0, \ldots, 0]$ sont tous égaux et sont le triple des moments du type $[2, 2, 0, \ldots, 0]$

Il existe plusieurs classes de solutions :

- les plans composites de Box et Wilson
- les plans de Box et Behnken
- les plans basés sur des polygones réguliers
- des plans hybrides.

Il existe plusieurs classes de solutions :

- les plans composites de Box et Wilson
- les plans de Box et Behnken
- les plans basés sur des polygones réguliers
- des plans hybrides.

Il existe plusieurs classes de solutions :

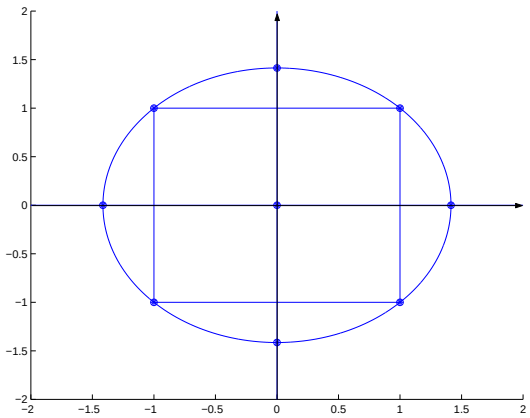
- les plans composites de Box et Wilson
- les plans de Box et Behnken
- les plans basés sur des polygones réguliers
- des plans hybrides.

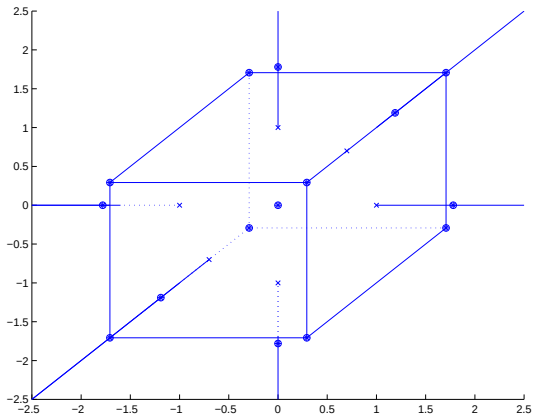
Il existe plusieurs classes de solutions :

- les plans composites de Box et Wilson
- les plans de Box et Behnken
- les plans basés sur des polygones réguliers
- des plans hybrides.

Plans de Box et Wilson

Il comprennent **une partie factorielle**, **une partie axiale**, **une partie centrale**





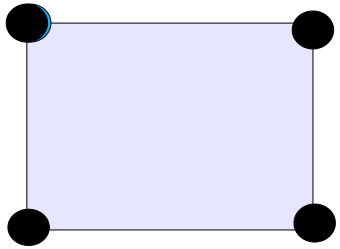
Pub

PUB PUB PUB PUB

PUB PUB PUB PUB

Plans D-optimaux

On considère un domaine



et un modèle qui sera le plus souvent un modèle polynomial, quadratique ou cubique mais qui peut être également un modèle de régression sur une base de Fourier ou sur une base d'ondelettes.

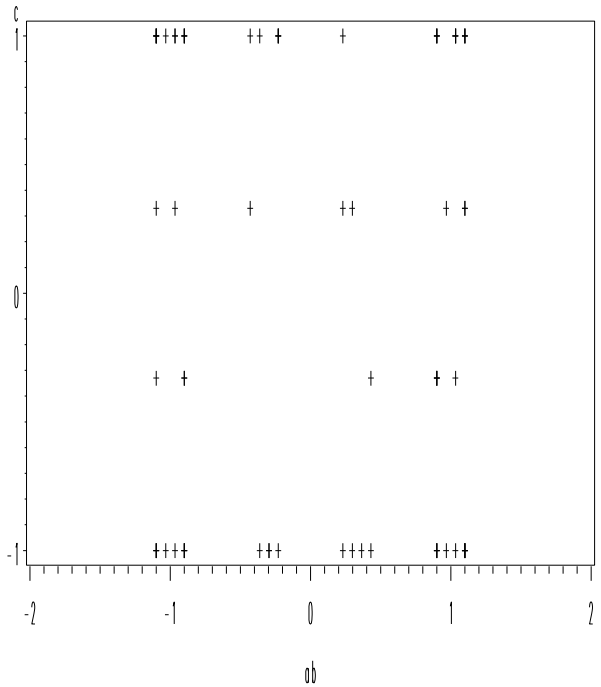
Ce modèle donne une matrice **X** qui est **la design matrix**. La matrice de Variance-covariance des estimateurs des coef de la régression vaut

$$(X'X)^{-1}$$

On ne peut pas minimiser une matrice, on doit donc en faire un résumé scalaire.
Une façon commode est de prendre le déterminant
On minimise $\det((X'X)^{-1})$ ou ce qui revient au même **on maximise** $\det(X'X)$.

Exemple de plan D-optimal

On représente un plan pour cinq variables en perspective cavalière (sur trois dimensions).



Le source

```
proc plan; factors a=4 b=4 c=4 d=4 e=4 ;
output out=table33 a nvals=(-1 -0.33 0.33 1) ..... e nvals=(-1
-0.33 0.33 1); run;quit;
proc optex data=table33; model a b c d e a*a a*b a*c a*d a*e
b*b b*c b*d b*e c*c c*d c*e d*d d*e e*e a*a*a ..... e*e*e ;
generate iter=15; output out=cubique1; run;quit;
```

Un aperçu de la sortie

Design Number	D-efficiency	A-efficiency	G-efficiency	Average Prediction Standard Error
1	18.2758	4.1033	58.7749	1.1321
2	18.1228	3.5925	56.5411	1.1613
3	18.0057	3.7093	53.2260	1.1718
4	17.9982	4.1690	52.8823	1.1233
5	17.9885	3.8112	56.3177	1.1356
6	17.9877	4.2374	58.4384	1.1089
7	17.9746	3.8078	58.8699	1.1499
8	17.9638	3.8835	56.7563	1.1396
9	17.9508	3.7556	55.1246	1.1577
10	17.8814	4.4257	58.4321	1.1118

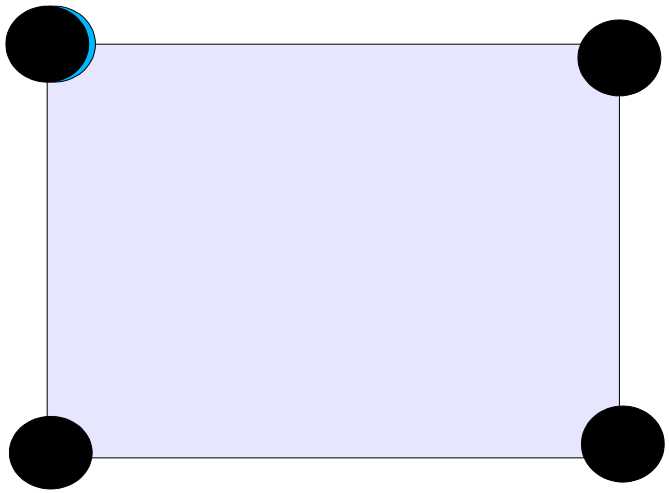
Conclusion

Les plans isovariants et D-optimaux sont vraiment conçus pour un environnement aléatoire. Ils ne sont pas exactement adaptés à notre problème.
De plus ils ont tendance à se concentrer "aux frontières de l'empire ".

On ne sait même pas si c'est bien ou mal!!

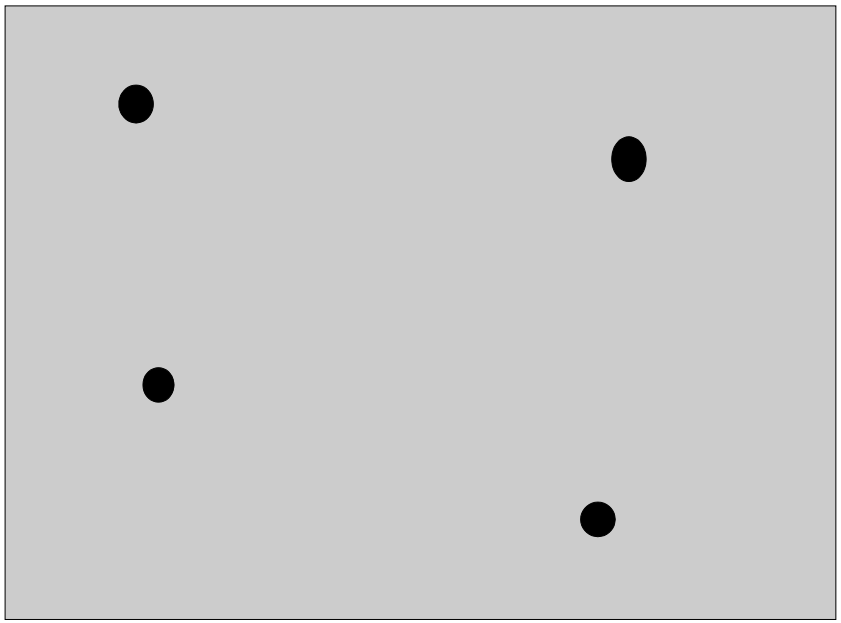
Le principe

En absence de toute information sur un modèle possible, on va juste chercher à répartir "le mieux possible" des points sur un domaine



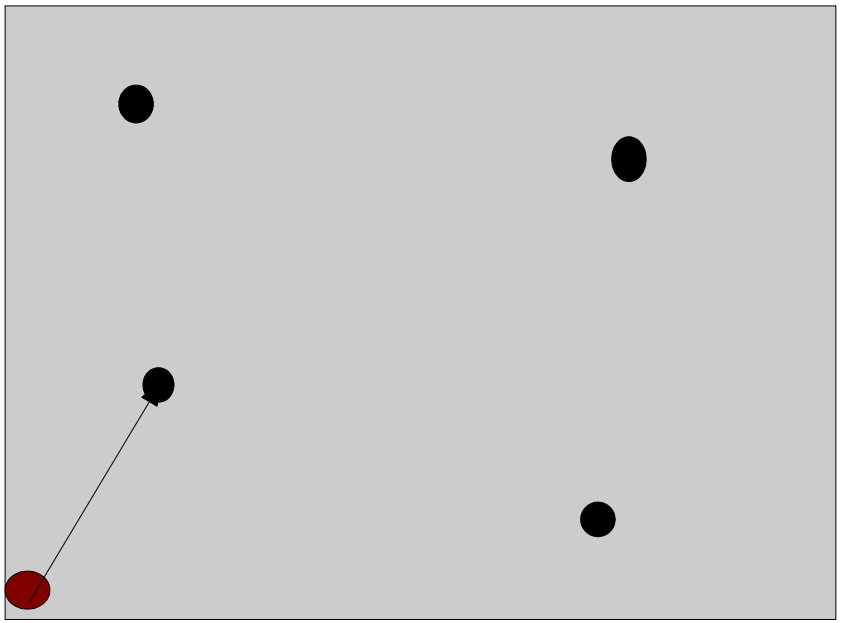
Critère minimax

On minimise la distance maximale d'un point quelconque du domaine au point le plus proche du plan d'expérience.



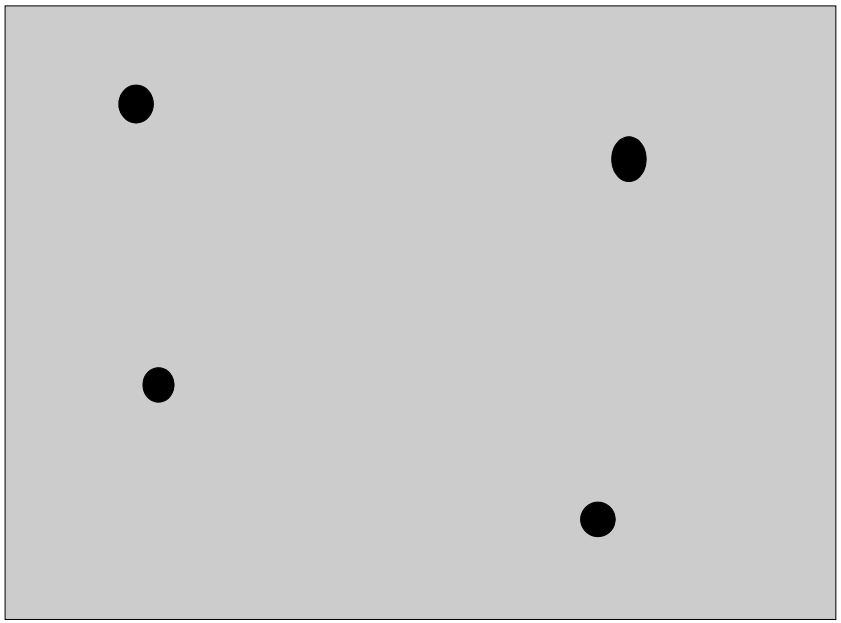
Critère minimax

On minimise la distance maximale d'un point quelconque du domaine au point le plus proche du plan d'expérience.



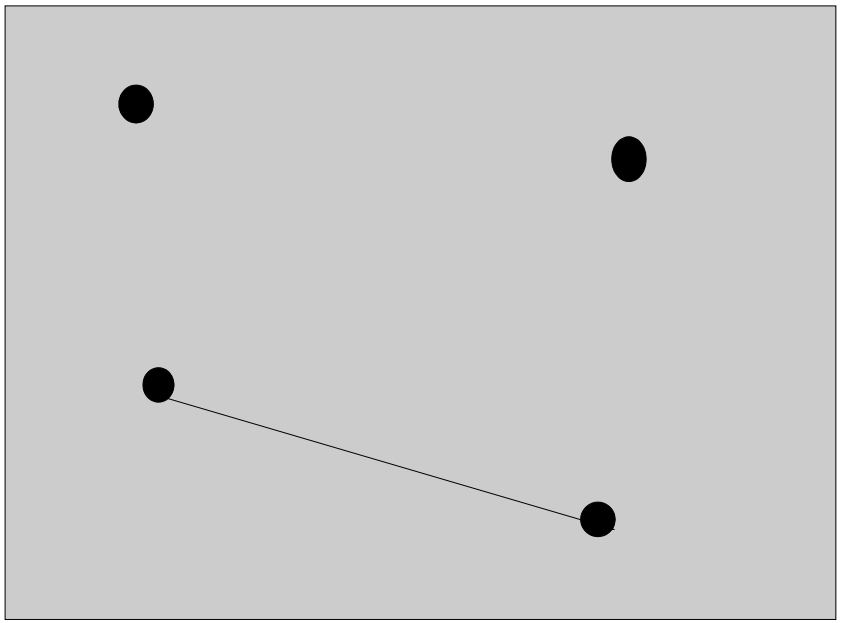
Critère maximin

On maximise la distance minimale entre deux points du plan d'expérience.



Critère maximin

On maximise la distance minimale entre deux points du plan d'expérience.



Solutions connues

De manière théorique on ne sait pas résoudre ce problème en général

- En dimension **un** il faut équirépartir les points.
- En dimension **deux** il faut plutôt les placer en quinconce.
- En dimension **trois** il faut plutôt les placer comme les centres des oranges chez le marchand de fruits.

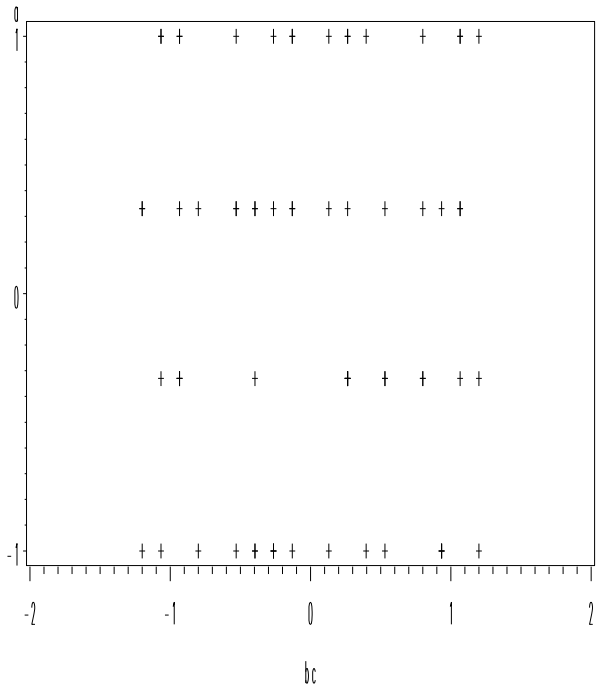

Minneola
Oranges
Honeydew

En général on ne sait pas. Mais il existe des solutions numériques :

- Proc optex de SAS
- Le design browser de la calibration tool-box de MATLAB
- et certainement beaucoup d'autres.

Plan space filling sur réseau

On représente une projection en dimension 3 en perspective cavalière d'un plan space-filling en dimension 5.



Les hypercubes latins

Une technique est donnée par les hypercubes latins (McKay, 79, Stein, 87)

n points dans l'hyper-cube $[0,1]^m$. Les variables sont $Z^{(1)}, ..., Z^{(m)}$

On découpe $[0,1]^m$ en n^m "petits cubes" $n = 4, m = 3$

• **Étape 1** On choisit $n = 4$ cubes parmi les 64.

Pour cela on choisit 3 permutations

- la première π_1 peut être prise égale à l'identité.
- la seconde et la troisième sont "tirées au hasard" parmi les permutations de 4 éléments. Par exemple :

$$\pi_2 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix}$$

$$\pi_3 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 2 & 4 \end{pmatrix}$$

On obtient donc le tableau

i	$\pi_1(i)$	$\pi_2(i)$	$\pi_3(i)$
1	1	4	3
2	2	3	1
3	3	2	4
4	4	1	2

- **Étape 2** On tire les points d'observation "au hasard" dans chaque "petit cube".

La méthode des hypercubes latins équilibre donc la marginale d'ordre 1 : pour chaque variable Il y a exactement une unité et une seule qui tombe dans chacun des n sous intervalles de $[0, 1]$.
On peut généraliser ce résultat en utilisant pour le choix des petits cubes parmi les n^m **un plan fractionnaire de résolution** r dont on peut montrer qu'il équilibre les marginales d'ordre $r - 1$.
cela implique que n soit une puissance de premier ce qui est une toute petite restriction :

2 3 4 5 6 **7 8 9** 10 **11**

Conclusion

Ce problème se rapproche du problème de la quantification ou plus généralement de toutes les méthodes de calcul d'intégrales multiples. Il existe une littérature sur les suites à discrépance faible.
Un lien est encore à faire entre ces méthodes.

Quantification de la gaussienne bi-dimensionnelle