Les plans orientés modèle: régression

29 janvier 2015

Introduction

- Un plan d'expérience est généralement associé à un modèle (régression, krigeage)
- Principe : choisir les expériences pour maximiser la « qualité » du modèle
- Le modèle est choisi a priori
- Le plan va dépendre du modèle choisi

Cours de régression (A. Badea, Science des données)

- Données subies
- Ici : on choisit nos données !

Retour sur la régression (1/2)

Notations

- Base de fonctions : $f_1, \ldots, f_p$
- Plan d'expériences : $\mathbf{x}_1, \ldots, \mathbf{x}_n$
- $\mathbf{F} = \begin{bmatrix} f_1(\mathbf{x}_1) & \ldots & f_p(\mathbf{x}_1) \\ & \vdots & \\ f_1(\mathbf{x}_n) & \ldots & f_p(\mathbf{x}_n) \end{bmatrix} = \begin{bmatrix} \mathbf{f}(\mathbf{x}_1) \\ \vdots \\ \mathbf{f}(\mathbf{x}_n) \end{bmatrix} = \mathbf{f}(\mathbf{X})$

Hypothèse

$Y = \mathbf{f}(\mathbf{x}^*)\beta + \epsilon$ avec ϵ gaussien i.i.d $N(0, \sigma^2)$

Prédicteur en $\mathbf{x}^*$

$\hat{y} = \mathbf{f}(\mathbf{x}^*)\hat{\beta}$

La qualité de l'apprentissage des coefficients est liée à celle de la prédiction

Retour sur la régression (2/2)

Calcul de $\hat{\beta}$

$$\hat{\beta} = \left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{Y}$$

Propriétés de $\hat{\beta}$

- *Best linear unbiased estimate* : $E(\hat{\beta}) = \beta$
- Covariance : $\left(cov(\hat{\beta}_i, \hat{\beta}_j)\right)_{i,j} = \sigma^2\left(\mathbf{F}^T\mathbf{F}\right)^{-1}$

Variance de prédiction

$$var(\hat{y}) = \sigma^2\mathbf{f}(\mathbf{x}^*)\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{f}(\mathbf{x}^*)^T = \sigma^2 d(\mathbf{x}^*)$$

Critère d'optimalité : meilleure connaissance des coefficients

Optimisation d'un plan

Quantité critique : $\mathbf{F}^T\mathbf{F}$

- Pas de dépendance dans les observations
- σ^2 se factorise : pas de dépendance dans le niveau de bruit

Définition du problème d'optimisation

Choisir $\mathbf{x}_1, \ldots, \mathbf{x}_n$ tel que $\left(\mathbf{F}^T\mathbf{F}\right)^{-1}$ ait de "bonnes" propriétés

Exercice : optimisation d'un plan à un point

Modèle :

- Un seul degré de liberté
- $y(x) = \beta_0 x$

Expérience :

- un seul point x^1
- $x^1 \in [0, 1]$

Où placer l'observation de manière optimale ?

- pour minimiser l'erreur sur β_0
- pour minimiser la variance de prédiction

Critères d'optimalité pour la régression

Quantité critique : $\mathbf{F}^T\mathbf{F}$

D-optimalité

- $\min \det \left(\mathbf{F}^T\mathbf{F}\right)^{-1} = \max \det \left(\mathbf{F}^T\mathbf{F}\right)$
- Minimiser le volume de l'ellipsoïde de confiance

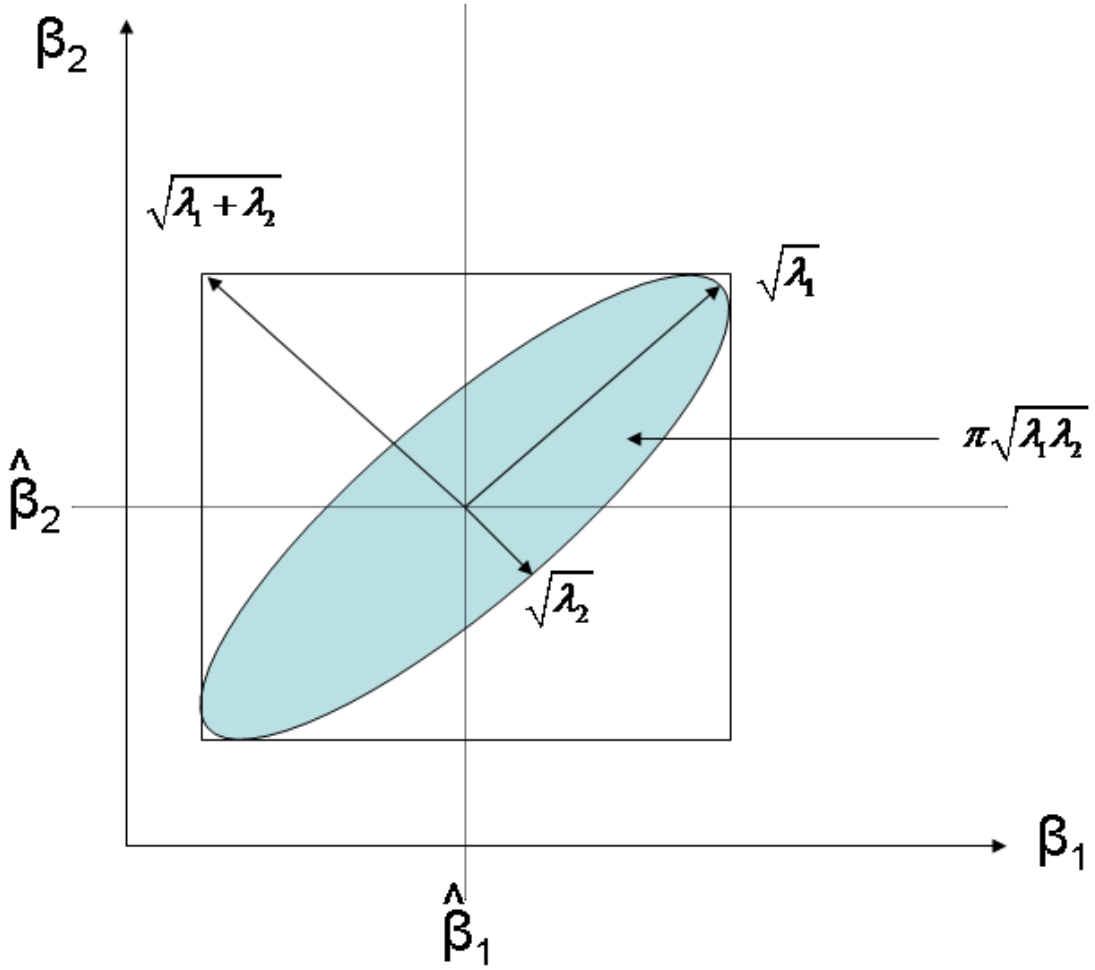
A-optimalité

- $\min tr \left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right]$
- Minimiser la somme des variances des coefficients

E-optimalité

- Minimiser la valeur propre maximale de $(\mathbf{F}^T\mathbf{F})^{-1}$:
- $\min \max_{1\leq i\leq n} \frac{1}{\lambda_i}$ (λ_i v.p. de $\mathbf{F}^T\mathbf{F}$)

Critères d'optimalité : interprétation graphique



Exercice : optimisation d'un plan à deux point

Modèle :

$$y(x) = \beta_0 + \beta_1 x$$

Expériences :

$$x_1, x_2 \in [0, 1]$$

Trouver le plan D-optimal

G-optimalité

Basé sur la variance de prédiction

- Minimiser le maximum de la variance de prédiction :
- $\min \max_{\mathbf{x}^*} \mathbf{f}(\mathbf{x}^*) \left(\mathbf{F}^T \mathbf{F}\right)^{-1} \mathbf{f}(\mathbf{x}^*)^T$
- Nécessite *a priori* une boucle d'optimisation imbriquée

Optimisation en pratique

Problème très complexe

- Nombre de variables : $n \times d$
- Problème très multimodal (invariances...)

En pratique : algorithmes d'échange

- On détermine le plus mauvais point du plan
- On cherche l'endroit du domaine le plus critique (par exemple, là où la variance est maximale)
- On supprime le mauvais point et on ajoute une observation au point critique

L'Algorithme à échange double de Fedorov (1/2)

Principe

On remplace une expérience $\mathbf{x}_i$ par $\mathbf{x}_j$

Equations

- $[\mathbf{F}^T\mathbf{F}]_{[t+1]} = [\mathbf{F}^T\mathbf{F}]_{[t]} - \mathbf{f}(\mathbf{x}_i)^T\mathbf{f}(\mathbf{x}_i) + \mathbf{f}(\mathbf{x}_j)^T\mathbf{f}(\mathbf{x}_j)$
- $det(\mathbf{F}^T\mathbf{F})_{[t+1]} = det(\mathbf{F}^T\mathbf{F})_{[t]} \times [1 + \delta_{ij}]$

avec :

- $\delta_{ij} = d(\mathbf{x}_j) - \left[d(\mathbf{x}_i)d(\mathbf{x}_j) - d(\mathbf{x}_i, \mathbf{x}_j)^2\right] - d(\mathbf{x}_i)$
- $d(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{f}(\mathbf{x}_i)\left(\mathbf{F}^T\mathbf{F}\right)^{-1}_{[t]}\mathbf{f}(\mathbf{x}_j)^T$

Tout se calcule sans inverser de matrice !

L'Algorithme à échange double de Fedorov (2/2)

Recherche du couple optimal

- En théorie : n optimisations en dimension d
- Historiquement : optimisation sur une grille
- Simplification 1 : on choisit $\mathbf{x}_i = argmin_{1 \leq k \leq n} d(\mathbf{x}_k)$
- Simplification 2 : on choisit $\mathbf{x}_j = argmax_{\mathbf{x} \in D} d(\mathbf{x})$
- Simplification 3 : on combine 1 et 2

Convergence

- Pas de preuve formelle
- Très efficace en pratique
- Nombreuses améliorations "modernes"

Pour aller plus loin : plans optimaux continus

Intérêt des répétitions

- On peut faire $N >> p$ expériences
- Vaut-il mieux faire N expériences distinctes ou des répétitions ?

Agrégation des répétitions

- Soit 2 observations $y_1(x_1)$ et $y_2(x_1)$, variance d'erreur σ^2
- On définit : $\bar{y} = \frac{1}{2}(y_1 + y_2)$, variance d'erreur $\sigma^2/2$
- Pas d'information perdue pour la régression !

Le modèle de régression généralisée

- Observations hétéroscédastiques : $var(\varepsilon_1) \neq var(\varepsilon_2) \ldots \neq var(\varepsilon_n)$
- Moindres carrés généralisés :

$$\beta^* = \left(\mathbf{F}^T \mathbf{\Gamma}^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^T \mathbf{\Gamma}^{-1} \mathbf{Y}$$

- avec : $\mathbf{\Gamma} = diag\left[var(\varepsilon_1), var(\varepsilon_2), ..., var(\varepsilon_n)\right]$

Gestion des répétitions

- en chaque point : $k_1, k_2, \ldots, k_n$ répétitions
- $\mathbf{\Gamma} = \sigma^2 diag\left[k_1^{-1}, ..., k_n^{-1}\right]$

Plans continus

Nouvelle définition d'un plan d'expériences

Pour un plan ξ à n points :

$$\xi = \left\{ \begin{array}{ccc} \mathbf{x}_1, & \dots & ,\mathbf{x}_n \\ k_1, & \dots & ,k_n \end{array} \right\} \qquad \text{avec : } \sum k_i = N$$

Plan continu normalisé

- En posant $\omega_i = k_i/N$, on peut factoriser la matrice de Fisher par $N\sigma^2$
- Si $N >> n$, alors les ω_i sont presque continus.
- Plan continu normalisé :

$$\xi = \left\{ \begin{array}{ccc} \mathbf{x}_1, & \dots & ,\mathbf{x}_n \\ \omega_1, & \dots & ,\omega_n \end{array} \right\} \qquad \text{avec : } \sum \omega_i = 1$$

Plans continus normalisés

Intérêt de cette définition

- On peut comparer des plans à différents nombres de points
- Le plan est défini par une mesure de probabilité (discrète)
- On a : $M(\xi) = \mathbf{F}^T\mathbf{F} = \int_D \mathbf{f}(\mathbf{x})^T\mathbf{f}(\mathbf{x})d\xi(\mathbf{x})$

Résultats fondamentaux

- Le support d'une mesure D-optimale est fini
- Théorème d'équivalence généralisé (TEG) : les plans D-optimaux et G-optimaux sont identiques
- Les valeurs optimales de D et G sont connues

Théorème d'équivalence de Kiefer et Wolfowitz (1960)

Les trois conditions sont équivalentes :

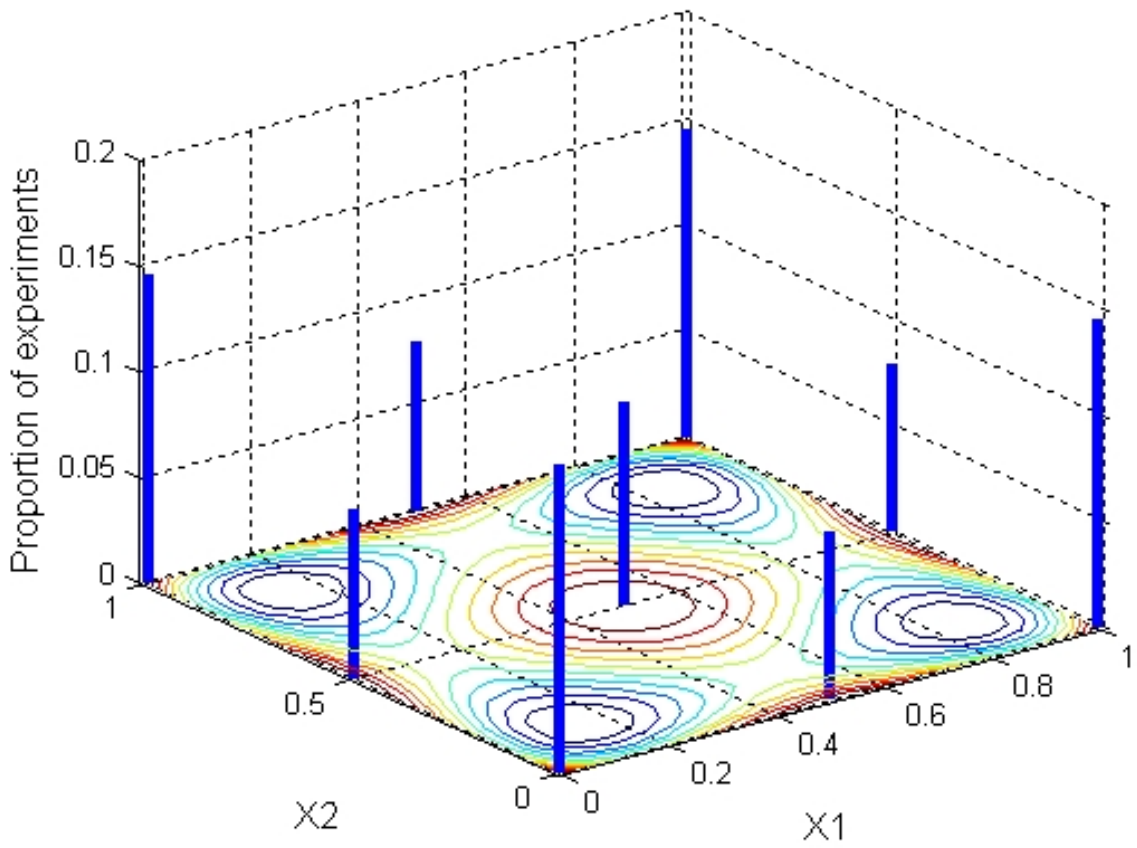
- Un plan est D-optimal
- Un plan est G-optimal
- $\max_{\mathbf{x}^*} \mathbf{f}(\mathbf{x}^*) \left(\mathbf{F}^T \mathbf{\Gamma}^{-1} \mathbf{F}\right)^{-1} \mathbf{f}(\mathbf{x}^*)^T = p$

Conséquences

- Maximiser le déterminant minimise la variance de prédiction maximale.
- On connait la plus petite valeur atteignable !

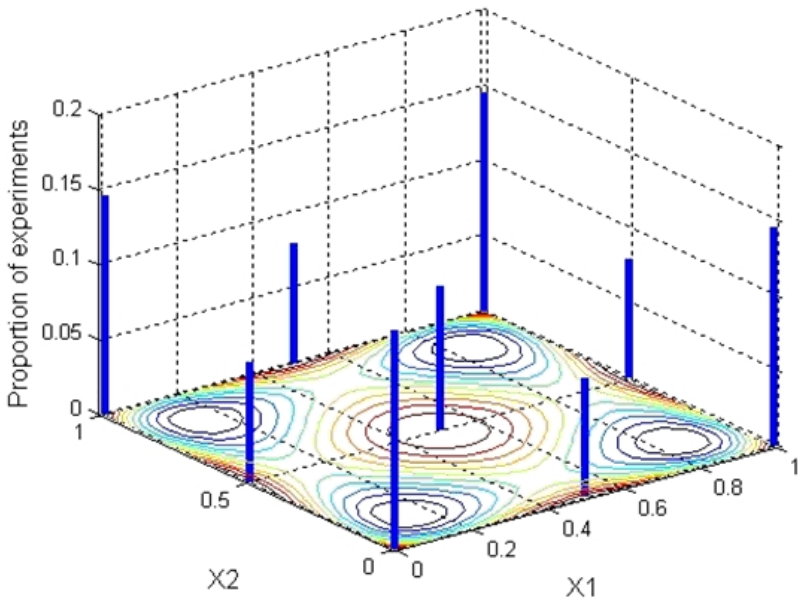
Un plan continu D-optimal (1/2)

- 14.6 % aux coins
- 8.0 % aux milieux des arêtes
- 9.6 % au centre

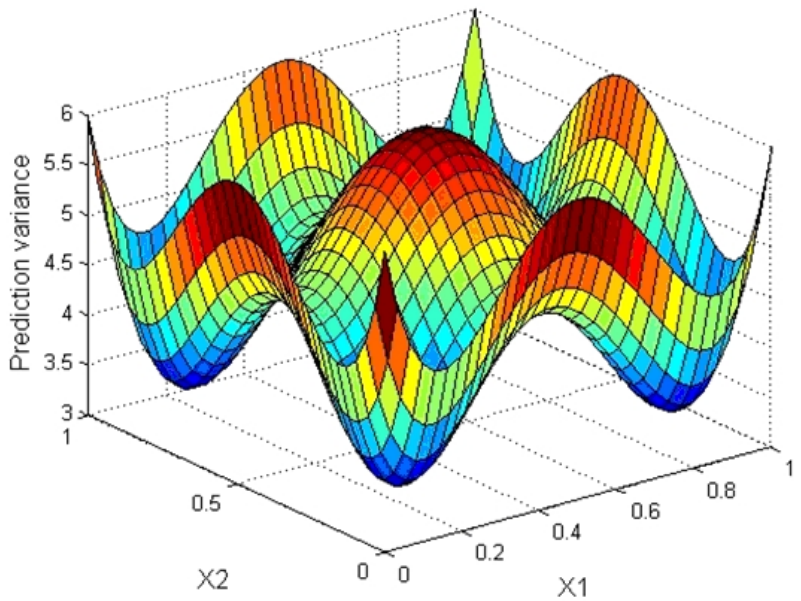


Un plan continu D-optimal (2/2)

Variance de prédiction du modèle :



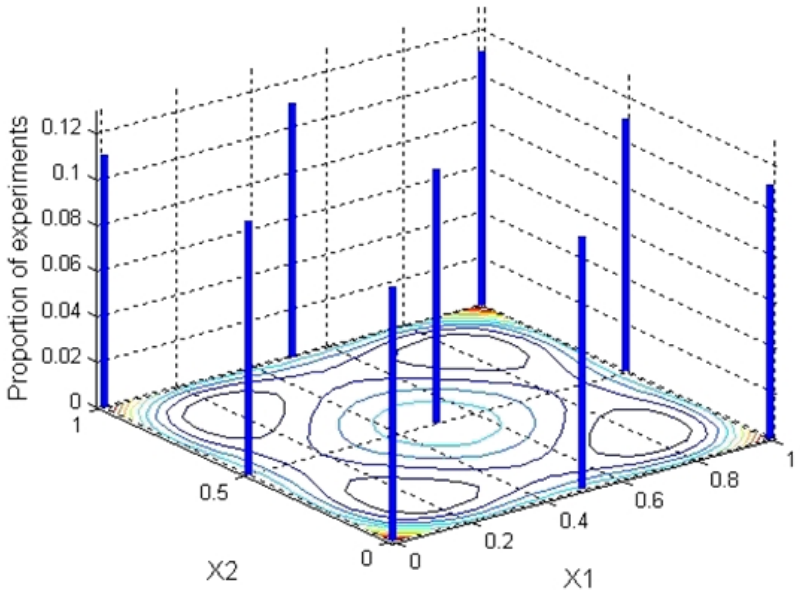
A



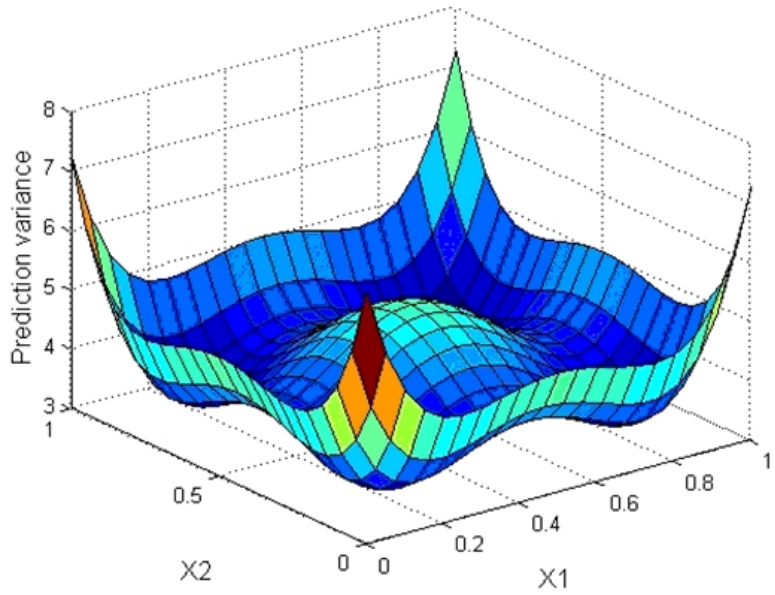
B

Le plan factoriel uniforme

Le plan n'est pas G-optimal :



A



B

D-efficacité, G-efficacité

D-efficacité

- $D_{eff} = 100 \times \frac{D(\xi)}{D(\xi^*)}$ (où ξ^* est la mesure optimale)
- Ecart d'un plan à l'optimalité
- $D(\xi^*)$ connu dans certains cas particuliers
- Très utilisé pour les plans simples

G-efficacité

- Connue analytiquement :

$$G_{eff} = 100 \times \sqrt{\frac{\max_{\mathbf{x}^*} \mathbf{f}(\mathbf{x}^*) \left(\mathbf{F}^T \mathbf{F}\right)^{-1} \mathbf{f}(\mathbf{x}^*)^T}{p}}$$

Nombre de points optimal

Retour sur l'exemple 1D

- $y(x) = \beta_0 + \beta_1 x$
- Trouver les poids optimaux pour $x_1 = 0$, $x_2 = 1$
- Calculer la variance maximale pour le plan à deux points D-optimal
- Calculer sa G-efficacité

Un plan D/G optimal contient au plus n_0 expériences distinctes :

$$n_0 = \frac{p(p+1)}{2} + 1$$

Construction de plans optimaux

Algorithme de Fedorov (1972)

0. On choisit une mesure initiale ξ_0
1. On cherche : $\mathbf{x}^* = argmax_D[d(\mathbf{x})]$ (variance maximum)
2. On calcule le poids attribué au nouveau point : $\omega^* = \frac{d(\mathbf{x}^*)-p}{p(d(\mathbf{x}^*)-1)}$
3. On met à jour la mesure : $\xi_{i+1} = (1-\omega^*)\xi_i + \omega^*\delta_{\mathbf{x}^*}$
4. On répète 1-2-3 jusqu'à ce que $max_D[d(\mathbf{x})] \approx p$

Robustesse à l'erreur de modèle

Si le modèle pour lequel le plan est optimisé est mal choisi

- Véritable modèle plus simple : OK
- Véritable modèle plus complexe : nouvelles expériences requises

Il existe des plans optimalement robustes à l'erreur de modèle !

Conclusion

Plans optimaux en régression

- Objectif : "meilleur" apprentissage des coefficients β_i
- De nombreux critères existent (basés sur la matrice de Fisher)

Utilité

- Pratique : études précises (modèles bien connus)
- Théorique : nombreux résultats fondamentaux (D/G optimalité, nombre d'observations optimal)