Quelques *bons plans* construits par optimisation en régression polynômiale

18 avril 2003

Résumé

On considère un modèle de régression paramétrique. On rapelle tout d'abord les résultats classiques concernant l'estimateur des moindres carrés. On définit ensuite plusieurs critères d'optimalité du plan d'expérience. Ces critères sont bâtis sur la matrice de variance covariance de l'estimateur des moindres carrés. Nous examinons ensuite le cas particulier de la régression polynômiale unidimensionnelle dans lequel nous exhibons les plans optimaux. Une ouverture sur la régression polynômiale multidimensionnelle est donnée en conclusion. Cet exposé reprend pour l'essentiel le chapitre 5 du livre de H. Dette et al [DS97].

1 Régression paramétrique

Soit U un sous-ensemble compact de $\mathbb{R}^d$ ($d \geq 1$), et f une application de $\mathbb{R}^d$ dans $\mathbb{R}^k$ ($k \geq 1$). On considère le modèle de régression :

$$Y(x) = \langle f(x), \theta^* \rangle + \varepsilon(x) \quad (x \in U). \tag{1}$$

$(\varepsilon(x))_{x \in U}$ est un *bruit blanc* et $\langle \cdot, \cdot \rangle$ désigne le produit scalaire habituel sur $\mathbb{R}^k$. Le modèle (1) est observé sur une *la grille* de points $x_1, \ldots, x_n$ (ces points ne sont pas nécessairement tous disctints). Le modèle d'observation est alors :

$$Y = F\theta^* + \varepsilon. \tag{2}$$

Où $F = (f_j(x_i))_{i=1,\ldots n; j=1\ldots k}$, $(n > k)$ est la matrice du plan d'expèrience et $\varepsilon = (\varepsilon(x_i))_{i=1,\ldots n}$ est le vecteur des erreurs. On fait l'hypothèse que le modèle d'observation n'est pas dégénéré, c'est-à-dire que la matrice F est de rang plein (k). On appelle $\widehat{\theta}_n$ l'estimateur des moindres carrés de θ^* (c'est l'estimateur qui minimise en $\theta \in \mathbb{R}^k$ l'erreur d'ajustement $\|Y - F\theta\|_2$). On a alors le résultat classique suivant qui s'obtient de façon élémentaire en utilisant le théorème de projection et de l'algèbre linéaire de base.

Théorème 1

1) $\widehat{\theta}_n$ existe et est unique. Par ailleurs on a $\widehat{\theta}_n = (F^T F)^{-1} F^T Y$.

2) Si on suppose de plus que ε est un vecteur de carré intégrable centré et de matrice de variance covariance $\sigma_*^2 I_n$ (bruit décorellé), alors

$$\mathbb{E}(\widehat{\theta}_n) = \theta^* \quad \textit{(sans biais), et} \quad \textit{Var}\,\widehat{\theta}_n = \sigma_*^2 (F^T F)^{-1} \ \textit{(matrice de variance-covariance).} \quad (3)$$

Rappelons que si A et B sont des matrices carrés symétriques et positives, on dit que $A \leq B$ (ordre de Loewner) lorsque $A - B$ est positive. Cela définit un ordre partiel sur l'ensemble des matrices symétriques et positives. Le théorème de Gauss-Markov suivant confère à $\widehat{\theta}_n$ une propriété d'optimalité relativement à l'ordre de Loewner.

Proposition 1 *Soit B une matrice $l \times k$ ($l \leq k$), de rang plein.*

1) Sous les hypothèses du point 2) du Théorème 1, $B\widehat{\theta}_n$ est l'estimateur qui possède la matrice de variance-covariance minimale (pour l'ordre de Loewner) parmi tous les estimateurs linéaires (c'est-à-dire s'écrivant sous la forme AY) et sans biais de $B\theta^*$.

2) Sous les hypothèses précédentes on suppose de plus que ε est un vecteur gaussien. Alors, $B\widehat{\theta}_n$ est l'estimateur qui possède la matrice de variance-covariance minimale parmi tous les estimateurs sans biais de $B\theta^*$.

A la lecture de la proposition précédente on pourrait se sentir optimiste et essayer de chercher à optimiser la matrice du plan d'expèrience en minimisant en $x_1, \ldots, x_n$ la matrice de variance-covariance (pour l'ordre de Loewner) de $B\widehat{\theta}_n$ (pour un B donné). Malheureusement, comme nous avons à faire avec un ordre partiel, il n'y a pas de solution. Pour construire un plan d'expérience par minimisation d'un critère, il faut bâtir une fonction de coût sur les matrices symétriques positives. C'est l'objet du paragraphe suivant.

2 Construction de plans d'expérience

A partir de maintenant on suppose que les hypothèses du point 2) du Théorème 1 sont vérifiées (sur le bruit). Soit B une matrice comme dans la Proposition 1. Remarquons tout d'abord que la matrice de variance covariance de $B\widehat{\theta}_n$ est :

$$\text{Var}\, B\widehat{\theta}_n = \sigma_*^2 B (F^T F)^{-1} B^T.$$

Typiquement, on peut s'intéresser seulement à une partie du paramètre. Par exemple, si on s'intéresse au j-ème paramètre $1 \leq j \leq k$, B est le j-ème vecteur de la base canonique duale de $\mathbb{R}^k$.

On *résume* les points de la grille dans une mesure de probabilité :

$$\mu_n(dx) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(dx).$$

Il est alors immédiat que $n^{-1}(F^T F)$ n'est rien d'autre que la matrice des moments d'ordre 2 du vecteur aléatoire $f(X_n)$ quand X_n a la loi μ_n :

$$\frac{1}{n}(F^T F)_{i,j} = \int_U f_i(x) f_j(x) \mu_n(dx) = \frac{1}{n} \sum_{m=1}^{n} f_i(x_m) f_j(x_m) \ \ (i, j = 1, \ldots k). \quad (4)$$

Plaçon nous dans un cadre asymptotique. Puisque U est compact il est naturel de supposer que la suite μ_n converge en loi vers une mesure de probabilité μ concentrée sur U. Soit X de loi μ sans perte de généralité on peut supposer que la matrice des moments d'ordre 2 de $f(X)$ est non dégénérée (sinon on réduit la dimension). Notons $\Delta(\mu)$ cette matrice :

$$(\Delta(\mu))_{i,j} = \int_U f_i(x) f_j(x) \mu(dx) \;\; (i,j = 1, \dots k). \tag{5}$$

Soit Σ_U un ensemble de mesure de probabilité sur U. Pour construire un plan d'expérience on procède alors en deux temps :

1) On se place en asymptotique. On maximise un critère bâti sur la matrice de variance-covariance asymptotique renormalisée de $B\widehat{\theta}_n$: $B\Delta^{-1}(\mu)B^T$. Soit μ^* une mesure maximisante.
2) On approxime (dans un sens à préciser), la mesure μ^* par une mesure portée par un nombre fini de points.

Définissons maintenant les critères matricielles que nous allons considérer.

Définition 1 *Soit A une matrice symétrique $l \times l$ définie positive ($l \in \mathbb{N}^*$). Soit $\lambda_1 \geq \lambda_2 \geq \dots \lambda_l > 0$ les valeurs propres de A. On considère les fonctions de coût suivantes :*

$$D(A) = \det(A^{-1}) = \prod_{j=1}^{l} \frac{1}{\lambda_j},$$
$$E(A) = \frac{1}{\sup_{u \in \mathbb{R}^l} u^T A u} = \frac{1}{\lambda_1}.$$

La procédure de construction de plans d'expérience peut alors être formalisée de la façon suivante :

Définition 2 *Soit Σ_U un ensemble de mesure de probabilités sur U et B une matrice $l \times k$ ($l \leq k$) de rang plein.*

1) On dit que la mesure $\mu^* \in \Sigma_U$ est D-optimale (resp. E-optimale) lorsqu'elle maximise $D(B\Delta^{-1}(\mu)B^T)$ (resp. $E(B\Delta^{-1}(\mu)B^T)$) sur Σ_U.

2) Une suite de plans d'expériences (μ_n^*) est D-optimale (resp. E-optimale) lorsqu'elle converge en loi vers μ^* qui satisfait 1).

Interprétation

- ***D*-optimalité**. Dans le cas où le bruit ε est supposé gaussien, l'ellipsoïde de confiance, au niveau $\alpha \in]0,1[$, pour $B\theta^*$ est donné par :

$$\Omega_\alpha = \left\{ \theta \in \mathbb{R}^k : \; (\theta - \widehat{\theta}_n)^T [B(F^T F)^{-1} B^T]^{-1} (\theta - \widehat{\theta}_n) \leq l S_n^2 f_{\alpha,l,n-k} \right\}. \tag{6}$$

Où

$$S_n^2 = \frac{\|Y - F\widehat{\theta}_n\|^2}{n-k}$$

est l'estimateur sans biais de la variance σ_*^2 et $f_{\alpha,l,n-k}$ est le quantile supérieur de niveau α pour la loi de Fisher $F(l, n-k)$. Le volume de Ω_α est proportionel à $\det B(F^TF)^{-1}B^T$. Un plan D-optimal permet donc d'obtenir asymptotiquement un volume minimal pour l'ellipsoïde de confiance de $B\theta^*$.

- **E-optimalité**. Il s'agit ici de minimiser la plus grande valeur propre de la matrice de variance covariance de $B\widehat{\theta}_n$. C'est donc un critère minimax. On cherche le plan d'expérience qui minimise asymptotiquement la plus grande des variances des combinaisons linéaires construites sur $B\widehat{\theta}_n$.

Soit $\mathbb{P}(U)$ l'ensemble des mesures de probabilités sur U. A partir de maintenant, on travaile avec

$$\Sigma_U^* = \{\mu \in \mathbb{P}(U) : \text{ Card}(\text{Supp}\mu) < +\infty \text{ et } \Delta(\mu) \text{ est inversible}\}.$$

Si $\mu \in \Sigma_U^*$ est portée par m points de U et vaut $\mu = \sum_{j=1}^m q_j \delta_{x_j}$ on l'écrit

$$\mu = \begin{pmatrix} x_1 & x_2 & \dots & x_m \\ q_1 & q_2 & \dots & q_m \end{pmatrix}$$

On a la caractérisation suivante :

Théorème 2 *On suppose (pour simplifier) que $B = I$. Il existe un plan D-optimal. De plus, un plan μ_* est D-optimal si, et seulement si,*

$$\forall x \in U,\ f^T(x)\Delta^{-1}(\mu_*)f(x) \leq k.$$

Remarquons que l'on a toujours, pour $\mu \in \Sigma_U^*$,

$$k = \text{Tr}(\Delta(\mu)\Delta(\mu)^{-1}) = \sum_{j=1}^m q_j f^T(x_j)\Delta^{-1}(\mu)f(x_j) \leq \|f^T\Delta^{-1}(\mu)f\|_\infty.$$

Posons,

$$G(\mu) = \|f^T\Delta^{-1}(\mu)f\|_\infty = \sup_{x\in U} f^T(x)\Delta^{-1}(\mu)f(x).$$

D'après le Théorème précédent, si μ_* est un plan D-optimal on a $G(\mu_*) = k$.

Définition 3 *Un plan est dit G-optimal si il minimise la fonction G sur Σ_U^*.*

Il faut noter que $f^T(x)\Delta^{-1}(\mu)f(x)$ est la variance de $f^T(x)\widehat{\theta}_n$. Ainsi, le plan G-optimal a la propriété minimax de minimiser la plus grande variance de cette collections d'estimateurs quand x décrit U. Le Corollaire suivant s'obtient en utilisant les remarques précédentes ou par des arguments de perturbation sur la fonction $\Phi(\mu) = \log\det(\Delta(\mu)),\ (\mu \in \Sigma_U^*)$.

Corollaire 1 (Kiefer-Wolfowitz) *On suppose que $B = I$. μ_* est G-optimal si, et seulement si, elle est D-optimal.*

3 Cas de la régression polynômiale

3.1 Quelques systèmes de polynômes orthogonaux

Sur $[-1,1]$ on considère la loi béta de paramètres $\alpha, \beta > -1$ (notée $\beta(\alpha,\beta)$), dont la densité par rapport à la mesure de Lebesgue est :

$$w_{\alpha,\beta}(x) = \frac{2^{-1-\alpha-\beta}}{B(\alpha+1,\beta+1)}(1-x)^{\alpha}(1+x)^{\beta},$$

$B(\alpha,\beta)$ est la constante de normalisation. Pour $n \in \mathbb{N}$ et $x \in [-1,1]$, le polynôme de Jacobi d'ordre n est

$$\begin{aligned} P_n^{\alpha,\beta}(x) &= \frac{(-1)^n}{2^n n!}(1-x)^{-\alpha}(1+x)^{-\beta}\frac{d^n}{dx^n}\left[(1-x)^{n+\alpha}(1+x)^{n+\beta}\right] \\ &= 2^{-n}\sum_{j=0}^{n}\left(\begin{array}{c} n+\alpha \\ j \end{array}\right)\left(\begin{array}{c} n+\beta \\ n-j \end{array}\right)(x-1)^{n-j}(x+1)^{j}. \end{aligned}$$

Les polynômes de Jacobi sont orthogonaux pour $w_{\alpha,\beta}$ (produit scalaire L^2 sur $[-1,1]$). Les polynômes ultrasphériques $(C_n^{\lambda})_{n\in\mathbb{N}}$ sont obtenus dans le cas $\alpha = \beta = \lambda - 1/2$ ($\lambda > -1/2$) avec une autre normalisation :

$$C_n^{\lambda}(x) = \begin{cases} \dfrac{\Gamma(\lambda+\frac{1}{2})\Gamma(2\lambda+n)}{\Gamma(2\lambda)\Gamma(\lambda+n+\frac{1}{2})}P_n^{\lambda-1/2,\lambda-1/2}(x) & x \in [-1,1],\ \lambda \neq 0, \\ \dfrac{2}{n}\kappa_n P_n^{-1/2,-1/2}(x) & x \in [-1,1],\ \lambda = 0, \end{cases}$$

où pour $n \in \mathbb{N}^*$,

$$\kappa_n = \frac{2\cdot 4\cdots 2n}{1\cdot 3\cdots(2n-1)} \quad \text{et} \quad \kappa_0 = 1.$$

Dans le cas $\alpha = \beta = 0$ ($\lambda = 1/2$), on obtient les polynômes de Legendre orthogonaux pour la loi uniforme sur $[-1,1]$. On note

$$P_n(x) = P_n^{0,0}(x) = C_n^{1/2}(x) \quad (x \in [-1,1],\ n \in \mathbb{N}).$$

3.2 Régression polynômiale univariée

On se place maintenant dans le modèle de régression polynômiale :

$$Y(x) = \sum_{j=0}^{r}\theta_j^* x^j + \varepsilon(x),\ (x \in [-1,1]). \tag{7}$$

On a bien sûr $k = r+1$. Sans perte de généralité nous nous sommes placés sur $[-1,1]$. C'est en effet toujours possible par reparamétrisation du problème. On a alors le Théorème suivant :

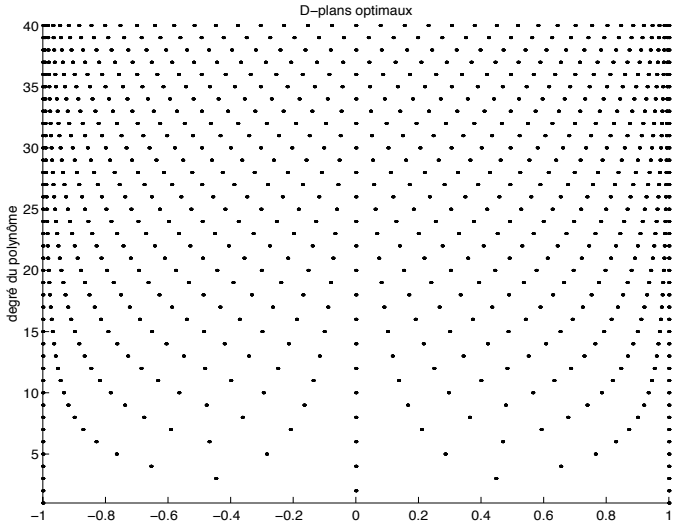

FIG. 1 – D-plans optimaux pour diverses valeurs de r

Théorème 3 *On suppose $B = I$. Le plan D-optimal dans le modèle de régression univarieé (7) est la mesure uniforme sur les zéros du polynôme de degré $r+1$:*

$$(x^2-1)P_r'(x).$$

La preuve de ce Théorème utilise des techniques de moments canoniques à travers la représentation des déterminants de Hankel sur ces moments. Lorsque le degré du polynôme augmente, le plan optimal converge en loi vers la loi de l'arcsinus (de densité $w_{-1/2,-1/2}$). La figure 1 est la représentation graphique des plans D-optimaux pour diverses valeurs de r.

3.3 Régression polynômiale multivariée

On s'intéresse maintenant au problème de régression multivariée :

$$\begin{aligned} Y(x) &= \theta_0^* + \sum_{i=1}^{d} \theta_i^* x_i + \sum_{1\le i_1\le i_2\le d} \theta_{i_1,i_2}^* x_{i_1}x_{i_2} + \cdots \\ &+ \sum_{1\le i_1\le\cdots\le i_d\le r} \theta_{i_1,\ldots,i_d}^* \prod_{j=1}^{r} x_{i_j} + \varepsilon(x),\ x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \in [-1,1]^d. \end{aligned} \tag{8}$$

On a ici $k = \begin{pmatrix} r+d \\ r \end{pmatrix}$. L'ensemble des mesures considéré ici n'est pas $\Sigma^*_{[-1,1]^d}$ mais son sous ensemble constitué de mesures produits :

$$\Sigma^{**}_{[-1,1]^d} = \left\{ \bigotimes_{i=1}^{d} \mu_i :\ \mu_i \in \Sigma^*_{[-1,1]},\ i = 1\ldots d \right\}. \tag{9}$$

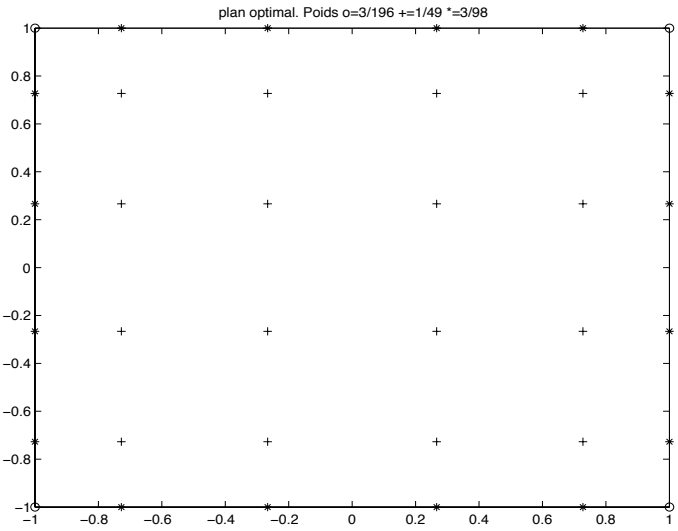

FIG. 2 – D-plans optimaux $d = 2$ $r = 5$

On a alors le Théorème suivant :

Théorème 4 *Le plan D-optimal dans le modèle de régression (8) est*

$$\mu_{**}(dx_1, \ldots, dx_d) = \bigotimes_{i=1}^{d} \mu_*(dx_i),$$

où μ_* est portée par les zéros du $(r-1)$-ème polynôme ultrasphérique $C_{r-1}^{d/2+1}$ et $\{-1, 1\}$. Elle donne le poids $1/(d+r)$ à ces zéros et $(d+1)/(2(d+r))$ à -1 et 1.

Ce Théorème se montre aussi en utilisant des méthodes de moments canoniques. Dans le cas d'un modèle de régression partielle, c'est-à-dire si l'on suppose que seulement une partie des monômes est présente dans (8), le plan D-optimal peut-être construit sous certaines hypothèses structurelles. La figure 2 donne le plan D-optimal dans le cas de la dimension 2 pour un polynôme de degé 5.

Références

[AD92] A. C. Atkinson and A. N. Donev. *Optimum experimental designs*. Oxford Science Publications, 1992.

[Co80] J. Coursol. *Technique statistique des modèles linéaires*. Les cours du CIMPA 1980.

[DS97] H. Dette and W. J. Studden. *The theory of canonical moments with applications in statistics, probability, and analysis*. John Wiley & Sons Inc., New York, 1997. A Wiley-Interscience Publication.

[Sc59] H. Scheffe. *The analysis of variance*. John Wiley & Sons Inc., New York, 1959.