Extraction séquentielle de conditions expérimentales optimales

Luc Pronzato

Laboratoire I3S (CNRS - UNSA)
Les Algorithmes - Bât Euclide B
2000 route des Lucioles – B.P. 121
F-06903 Sophia Antipolis Cedex
pronzato@i3s.unice.fr

RÉSUMÉ. *Nous considérons le problème posé par la sélection séquentielle de conditions expérimentales en vue de la construction d'un modèle paramétrique. Ceci couvre en particulier deux aspects. (i) Fouille de données : les données, c'est-à-dire les observations ainsi que les conditions expérimentales qui ont été utilisées pour les obtenir, sont déjà disponibles en masse et il s'agit d'explorer* en ligne *une base de données afin d'en sélectionner* séquentiellement *la fraction α la plus pertinente. (ii) Entreposage de données : il s'agit de faire croître au mieux une base de données, en n'y incorporant à l'avenir que la fraction α la plus pertinente des données qui sont proposées. La «qualité» des données est quantifiée à partir d'un critère de planification d'expérience Φ, fonction de la matrice d'information de Fisher. Nous donnons une règle de décision asymptotiquement optimale, qui conduit à une expérience Φ-optimale (sous la contrainte que seulement α% des données sont sélectionnées).*

ABSTRACT. *We consider the problem of sequential selection of experimental conditions for parameter estimation. This may concern two particular situations: (i) the data, including the observations and the associated experimental conditions, are already available in a massive data base, to be explored on-line in order to sequentially select the most informative α% of the data; (ii) in order to enlarge the data base at best in the future, only the most informative part of the new data will have to be included. The «quality» of data is measured by an experimental design criterion Φ, function of the Fisher information matrix. A decision rule is constructed which is asymptotically optimal and yields a Φ-optimum experiment (under the constraint that only α% of the data are selected).*

MOTS-CLÉS : *planification d'expériences, planification séquentielle, échantillonnage, protocole expérimental, fouille de données, extraction d'information, entreposage de données.*

KEYWORDS: *sequential design, sampling, constrained design measure, data mining, extraction of information.*

1. Présentation du problème

Nous souhaitons construire un modèle paramétrique, et, dans ce but, nous examinons séquentiellement des conditions expérimentales caractérisées par des grandeurs $X_k \in \mathcal{X} \subseteq \mathbb{R}^q$, $k = 1, 2, \ldots$, que nous supposerons former une suite de variables aléatoires indépendantes identiquement distribuées (avec une mesure de probabilité μ).

Dès qu'une nouvelle variable X_k est examinée, nous devons décider «en ligne» si elle doit être acceptée ou pas. Il peut s'agir de variables déjà présentes en masse dans une grande base de donnée, trop grande pour être utilisée dans sa totalité afin de construire le modèle qui nous intéresse : dans ce cas, nous allons tenter d'extraire séquentiellement de cette masse la partie la plus informative pour notre construction de modèle (*fouille de données*). Il peut s'agir de nouvelles variables recueillies «dans la nature» : on souhaite alors faire vivre une base de données en ne lui incorporant que les grandeurs les plus informatives, toujours en terme de construction de modèle (*entreposage de données*). Dans les deux cas, nous devons être sélectif, et la démarche adoptée ici consiste à n'accepter qu'une fraction α des variables expérimentales proposées, $\alpha \in (0, 1)$.

Comme de coutume, la qualité des conditions expérimentales est quantifiée par une fonction Φ de la matrice d'information de Fisher. On se place dans le cas où, pour des variables expérimentales $X_1, X_2, \ldots, X_n$, cette matrice est de la forme

$$\mathbf{M}_n = \sum_{k=1}^{n} \mathbf{f}(X_k)\mathbf{f}^\top(X_k)\,,$$

où $\mathbf{f}(\cdot)$ dépend du modèle paramétrique considéré, avec $\dim(\mathbf{f}) = \dim(\theta)$ et θ les paramètres d'intérêt. Nous supposons que $\mathbf{f}(\cdot)$ est une fonction continue de X. Ceci recouvre aussi bien le cas de la modélisation de la probabilité de succès dans une loi de Bernouilli (modèle logistique par exemple), que celui plus classique d'un modèle de régression, linéaire ou non. Notons que si le modèle est non linéaire en ses paramètres θ, $\mathbf{f}(\cdot)$ dépend de leur valeur. Nous supposerons alors qu'une valeur nominale $\hat{\theta}^0$ est disponible, laquelle est utilisée pour calculer $\mathbf{f}(\cdot)$.

Le critère $\Phi(\cdot)$ (à maximiser) est choisi concave, croissant et linéairement différentiable sur l'ensemble des matrices symétriques définies positives ; $\Phi(\mathbf{M}) = \log\det(\mathbf{M})$ et $\Phi(\mathbf{M}) = -\text{trace}(\mathbf{M}^{-1})$ sont des choix classiques, voir par exemple [SIL 80]. Nous noterons $\mathcal{F}_\Phi(\mathbf{M}_1, \mathbf{M}_2)$ la dérivée directionnelle $\lim_{\epsilon\to 0^+}\{\Phi[(1-\epsilon)\mathbf{M}_1 + \epsilon\mathbf{M}_2] - \Phi(\mathbf{M}_1)\}/\epsilon$.

Soit $(u_k)_k$ la suite des décisions prises, avec $u_k = 1$ si X_k est accepté, $u_k = 0$ sinon (X_k est connu quand la décision est prise). Si la suite des variables X_k proposées est de longueur N, et que seulement un nombre $n < N$ peut en être accepté, le problème à résoudre correspond à

$$\text{maximiser } \mathrm{E}\{\Phi(\mathbf{M}_{N,n}/n)\} \qquad [1]$$

par rapport à $(u_k)_k$ satisfaisant les contraintes $u_k \in \mathcal{U}_k \subseteq \{0,1\}$, $k = 1, \ldots, N$, $\sum_{k=1}^{N} u_k = n$, avec $\mathbf{M}_{N,n} = \sum_{k=1}^{N} u_k \mathbf{f}(X_k)\mathbf{f}^\top(X_k)$. L'espérance mathématique $\mathtt{E}\{\cdot\}$ dans (1) porte sur $X_1, \ldots, X_N$ et concerne donc la mesure produit $\mu^{\otimes N}$. On peut donner de ce problème une formulation du type contrôle optimal stochastique.

A l'étape k, $1 \leq k \leq N$, quand il s'agit d'accepter ou rejeter X_k, notons a_k le nombre de X_i déjà retenus : $a_k = \sum_{i=1}^{k-1} u_i$, avec $a_1 = 0$. On peut alors considérer $(a_k, \mathbf{M}_{k-1,a_k}, X_k)$ comme l'état et u_k comme la commande du «système» au «temps» k. Une stratégie $S_{N,n}$ est définie par une fonction $(k, a, \mathbf{M}, X) \mapsto u \in \{0,1\}$. Pour tout $k \in \{1, \ldots, N\}$, la décision optimale est obtenue par la résolution du problème suivant :

$$\max_{u_k \in \mathcal{U}_k} [\mathtt{E}_{X_{k+1}}\{ \max_{u_{k+1} \in \mathcal{U}_{k+1}} [\mathtt{E}_{X_{k+2}}\{ \max_{u_{k+2} \in \mathcal{U}_{k+1}} [\ldots$$

$$\mathtt{E}_{X_{N-1}}\{ \max_{u_{N-1} \in \mathcal{U}_{N-1}} [\mathtt{E}_{X_N}\{ \max_{u_N \in \mathcal{U}_N} [\Phi(\sum_{i=1}^{N} u_i \mathbf{f}(X_i)\mathbf{f}^\top(X_i))]\}]\} \ldots]\}]\}] \,,$$

où $\mathtt{E}_{X_j}\{.\}$ désigne l'espérance mathématique sur X_j, distribué avec la mesure de probabilité μ, et

$$\mathcal{U}_j = \mathcal{U}_j(a_j) = \begin{cases} \{0\} & \text{si } a_j = n \,, \\ \{1\} & \text{si } a_j + N - j + 1 \leq n \,, \\ \{0,1\} & \text{sinon.} \end{cases}$$

2. Résultats

Le cas $\dim(\theta) = 1$ est considéré dans [PRO 01a]. La stratégie optimale (en boucle fermée) est obtenue par programmation dynamique, et une règle de décision «boucle ouverte» est proposée, asymptotiquement optimale pour n fixé et N tendant vers l'infini (sa construction repose sur la distribution des valeurs extrêmes pour μ, et l'optimalité asymptotique est obtenue pour une fonction de distribution des X_k de type Von Mises).

La programmation dynamique ne conduit plus aussi simplement à la solution optimale dans le cas multidimensionnel ($\dim(\theta) > 1$). C'est ce cas qui nous intéresse ici. Des stratégies sous-optimales sont proposées dans [PRO 99] (boucle ouverte avec retour d'information) et [PRO 01b] («optimale à un pas», c'est-à-dire sur un horizon glissant de longueur 1). La première s'avère fortement sous-optimale.

La démarche proposée ici repose sur la construction d'un protocole expérimental approximatif (*approximative design*). Soit Ξ l'ensemble des mesures de probabilité sur $\mathcal{X}$. Pour tout $\xi \in \Xi$, nous noterons $\mathbf{M}(\xi) = \int_{\mathcal{X}} \mathbf{f}(x)\mathbf{f}^\top(x)\xi(dx)$ et $\phi(\xi) = \Phi[\mathbf{M}(\xi)]$. Nous supposons naturellement que $\mathbf{M}(\mu)$ existe et que $-\infty < \phi(\mu) < \infty$, avec μ la mesure de probabilité pour chaque X_k. Nous supposons de plus que μ ne possède pas d'atomes, c'est-à-dire que pour tout ensemble $\Delta\mathcal{X}$ il existe $\Delta\mathcal{X}' \subset \Delta\mathcal{X}$ tel que $\int_{\Delta\mathcal{X}'} \mu(dx) < \int_{\Delta\mathcal{X}} \mu(dx)$ (avec les mesures absolument continues par rapport à la

mesure de Lebesgue comme cas particulier). Puisque $\Phi(\cdot)$ est linéairement différentiable, nous avons $F_\Phi(\xi_1;\xi_2) = \mathcal{F}_\Phi[\mathbf{M}(\xi_1),\mathbf{M}(\xi_2)] = \int_\mathcal{X} F_\Phi(\xi_1,x)\xi_2(dx)$ pour tout ξ_1,ξ_2 dans Ξ avec $\phi(\xi_1) > -\infty$, où nous notons $F_\Phi(\xi,x) = F_\Phi(\xi;\delta_x)$ et δ_x la mesure de Dirac en x. Par exemple, le cas de la D-optimalité où $\Phi(\cdot) = \log\det(\cdot)$ donne $F_\Phi(\xi,x) = f^\top(x)\mathbf{M}^{-1}(\xi)f(x) - p$, avec $p = \dim(\theta)$.

Nous considérons le comportement asymptotique pour $n = \lfloor \alpha N \rfloor$, avec $\alpha \in (0,1)$ et $N \to \infty$. Nous montrons tout d'abord que la détermination d'une mesure ξ_α^* optimale pour $\phi(\cdot)$ sous la contrainte $\xi_\alpha^* \leq \mu/\alpha$, voir par exemple [WYN 82, FED 89], permet la construction d'une solution asymptotiquement optimale pour le problème (1). Nous montrons ensuite que la construction de ξ_α^* n'est pas un préalable nécessaire, et proposons une stratégie séquentielle $S_\alpha(\mu)$, définie par

$$S_\alpha(\mu) : \left\{ \begin{array}{l} \text{accepter } X_k \text{ si } P_k = P_k(X_k) = \mu\{x \,/\, F_\Phi(\xi_k,x) > F_\Phi(\xi_k,X_k)\} < \alpha\,, \\ \text{rejeter } X_k \text{ sinon}\,. \end{array} \right. \qquad [2]$$

Le résultat principal, obtenu sous certaines hypothèses techniques sur μ et $\Phi(\cdot)$, s'énonce ainsi.

THÉORÈME 1 *La mesure empirique ξ_k générée par les points acceptés par la stratégie $S_\alpha(\mu)$ définie par (2) satisfait $\lim_{k\to\infty}\phi(\xi_k) = \phi(\xi_\alpha^*)$, μ-a.s., avec ξ_α^* une mesure bornée Φ-optimale, $\xi_\alpha^* \leq \mu/\alpha$.*

Autrement dit, la stratégie définie par (2) échantillonne asymptotiquement suivant la mesure bornée Φ-optimale ξ_α^*, et est asymptotiquement optimale pour le problème (1) quand $n = \lfloor \alpha N \rfloor$ avec $\alpha \in (0,1)$ et $N \to \infty$.

3. Bibliographie

[FED 89] FEDOROV V., « Optimal design with bounded density : optimization algorithms of the exchange type », *J. Statist. Planning and Inference*, vol. 22, 1989, p. 1–13.

[PRO 99] PRONZATO L., « Sequential selection of observations in randomly generated experiments », *Proc. ProbaStat'98, Smolenice (Slovaquie), Feb. 98,* Tatra Mountains Mathematical Publications, vol. 17, 1999, p. 167–175.

[PRO 01a] PRONZATO L., « Optimal and asymptotically optimal decision rules for sequential screening and resource allocation », *IEEE Transactions on Automatic Control*, vol. 46, n° 5, 2001, p. 687–697.

[PRO 01b] PRONZATO L., « Sequential construction of an experiment design from and i.i.d. sequence of experiments without replacement », ATKINSON A., BOGACKA B., ZHIGLJAVSKY A., Eds., *Optimum Design 2000*, chapitre 11, p. 113–122, Kluwer, Dordrecht, 2001.

[SIL 80] SILVEY S., *Optimal Design*, Chapman & Hall, London, 1980.

[WYN 82] WYNN H., « Optimum submeasures with applications to finite population sampling », GUPTA S., BERGER J., Eds., *Statistical Decision Theory and Related Topics III. Proc. 3rd Perdue Symp., vol. 2*, p. 485-495, Academic Press, New York, 1982.