Rev. Statistique Appliquée, 2005, **LIII** (4), 87-103

CONSTRUCTION D'UN CRITÈRE D'OPTIMALITÉ POUR PLANS D'EXPÉRIENCES NUMÉRIQUES DANS LE CADRE DE LA QUANTIFICATION D'INCERTITUDES

L. CARRARO[1], B. CORRE[1,2], C. HELBERT[1], O. ROUSTANT[1]

[1] *Département Méthodes et Modèles Mathématiques pour l'Industrie, École Nationale Supérieure des Mines de Saint-Etienne*

[2] *Direction Exploration Production, Division Techniques Géosciences Total*
courriel : nom@emse.fr

RÉSUMÉ

De nombreux phénomènes physiques sont étudiés à l'aide de simulateurs numériques coûteux, avec lesquels une variable d'intérêt – ou «réponse» – est une fonction déterministe des variables d'entrée (les facteurs). Cependant, on est souvent amené à évaluer la réponse sous forme d'incertitudes du fait de la méconnaissance du niveau des facteurs. Ainsi en Exploration/Production pétrolière, on s'intéresse par exemple à la distribution de la production d'huile d'un réservoir dans dix ans.

Dans cet article nous construisons un critère conçu pour planifier les simulations de sorte que la quantification des incertitudes sur la réponse soit la meilleure possible. Baptisé «MC-V optimalité», le critère obtenu est alors équivalent à un critère IMSE (Integrated Mean Squared Error) où l'intégration est effectuée selon la distribution des facteurs. La démarche sera illustrée par l'exposé du contexte de l'Exploration/Production pétrolière dont l'étude est à l'origine de ce critère.

Mots-clés : *planification d'expériences, surface de réponse, critères d'optimalité*

ABSTRACT

Many physical phenomena are studied using expensive numerical simulators, where a variable of interest-the response- is a deterministic function of the input variables or factors. However, one often needs to evaluate the uncertainties of the response, due to the lack of knowledge of factors levels. For example, in the field of oil Exploration/Production, one may be interested by the distribution of the oil production in ten years for a particular reservoir. In this paper we build a criterion suited to plan simulations so that the quantification of uncertainties of the response is the best possible. It is shown that this criterion, called MC-V, is equivalent to IMSE (Integrated Mean Squared Errors), where integration is carried out according to the distribution of the factors.

Keywords : *design of experiments, response surface, optimality criterions*

1. Introduction

De nombreux phénomènes physiques sont étudiés à l'aide de simulations numériques, ou *expériences numériques*, avec lesquels une variable d'intérêt – ou *réponse* – est une fonction déterministe des variables d'entrée – *les facteurs*. Cependant, on est souvent amené à évaluer la réponse sous forme d'incertitudes du fait de la méconnaissance des valeurs des facteurs. Partant du constat que les méthodes usuelles de construction de plans d'expériences ne sont pas adaptées pour optimiser la quantification des incertitudes, nous construisons un nouveau critère d'optimalité conçu spécifiquement pour répondre à cet objectif.

En Exploration/Production pétrolière par exemple, on étudie la capacité de production des réservoirs dans les années à venir au moyen de simulateurs d'écoulement. On dispose ainsi d'un simulateur qui fournit la production cumulée d'huile sur les dix prochaines années à des coordonnées données en fonction de champs de facteurs physiques *sub-surface* comme la porosité, la perméabilité, la perméabilité relative, ... en différents points du réservoir. Mais si la fonction donnant la production d'huile à dix ans est parfaitement déterministe (*via* le simulateur), le niveau des facteurs sub-surface est très mal connu. Le problème consiste alors à fournir des intervalles de confiance sur la production d'huile.

Le problème générique de la quantification des incertitudes sur la réponse peut être traité en deux étapes. Une première partie du problème consiste à traduire les incertitudes sur les facteurs en lois de probabilité. Cette étape est en général traitée au moins partiellement par les experts métier (dans l'exemple précédent, les géostatisticiens, les géologues, les géophysiciens). La seconde partie consiste alors à obtenir de façon approchée la loi de probabilité de la réponse par *propagation* des incertitudes des facteurs à travers le simulateur d'écoulement. Nous nous intéressons ici à cette seconde étape, qui est réalisée par le statisticien. Pour ce faire, on peut utiliser la méthode de Monte Carlo (MC) dont voici la forme la plus rudimentaire :

Pour $r = 1, \ldots, R$,

- Simuler $x = (x_1, \ldots, x_k)$ selon la loi des k facteurs
- Calculer $y_r = Y_{\text{sim}}(x)$ où Y_{sim} désigne le simulateur

Lorsque R est grand, la loi des grands nombres permet de justifier l'approximation de la fonction de répartition théorique de la réponse $Y = Y_{\text{sim}}(\tilde{x})$ [1] par la fonction de répartition empirique :

$$P(Y \leqslant y) \approx \frac{1}{R} \sum_{r=1}^{R} 1_{\{y_r \leqslant y\}}$$

où $1_{\{y_r \leqslant y\}}$ vaut 1 si $y_r \leqslant y$ et 0 sinon.

Cependant une difficulté importante provient du coût en temps de calcul souvent prohibitif d'une expérience. Ainsi dans l'exemple précédent, une simulation des écoulements au sein du réservoir demande plusieurs heures, voire plusieurs jours. Dans ces conditions, la propagation des incertitudes ne peut s'appliquer sur le simulateur lui-même. L'approche retenue est alors de modéliser simplement le

[1] $\tilde{x}$ est le vecteur aléatoire dont chaque simulation x est une réalisation.

simulateur afin de rendre réalisable en terme de coût la propagation des incertitudes. La difficulté principale est que le modèle simple doit être estimé avec un petit nombre d'expériences. Tout le problème consiste alors à bien choisir ces expériences afin d'assurer la meilleure estimation possible du simulateur d'écoulement.

Cette méthodologie est commune à un nombre croissant de problèmes faisant intervenir des codes de calcul complexes. Le simulateur est modélisé soit par approximation en utilisant la méthodologie des *surfaces de réponse*, usuelle en planification d'expériences (voir par exemple Box, Draper, 1987) et encore largement répandue dans le milieu industriel; soit par interpolation en utilisant des techniques issues du krigeage des géostatisticiens (Sacks, Welch, Mitchell, Wynn, 1989, et Sacks, Schiller, Welch, 1989). Ces dernières ont l'avantage de tenir compte du fait que les expériences sont réalisées avec un simulateur, mais ne semblent pas encore très employées (sans doute du fait de leur plus grande difficulté d'utilisation). Il existe des critères pour planifier de façon optimale les expériences (D-optimalité, G-optimalité, IMSE, entropie, distance maximin, distance minimax, etc.) mais, à notre connaissance, aucun d'entre eux n'est conçu pour répondre à l'objectif de quantification d'incertitudes. En outre, ces plans placent les points, dans les cas les plus simples, aux bords du domaine de variation des facteurs, ce qui pose problème dans la mesure où les points sont très souvent de probabilité nulle dans le cadre d'une simulation de Monte Carlo.

Dans ce contexte où il s'agit à la fois de tenir compte du fait que le résultat des expériences numériques est déterministe et qu'il s'agit de propager des incertitudes, nous avons choisi d'aborder ici le problème de propagation des incertitudes. L'article est structuré en 6 parties. Dans la première (section 2), on fixe le cadre probabiliste et on donne la modélisation du problème. La construction du critère est alors réalisée dans la seconde partie (section 3); différentes expressions sont données (section 4) qui permettent ensuite de faire le lien avec d'autres critères et de montrer l'équivalence avec l'IMSE (section 5). On donne ensuite des exemples de plans optimaux pour un cas issu d'une situation réelle (section 6). Enfin dans la section 7, une discussion est menée pour évoquer des améliorations possibles du critère et de sa généralisation à d'autres cadres.

2. Modélisation

2.1. Modèle d'approximation pour le simulateur

Dans les cas pratiques, la surface de réponse est souvent obtenue par approximation. Cela revient à considérer que le simulateur est donné par le modèle d'approximation suivant :

$$Y_{\text{sim}}(x) = X(x)\beta + \varepsilon(x) \tag{1}$$

avec :

- $x = (x_, \ldots, x_k)'$, vecteur $k \times 1$ définissant les valeurs des facteurs
- une partie «modèle » (appelée aussi «surface de réponse»), $X(x)\beta$, donnée sous la forme d'une combinaison linéaire de fonctions de base (par exemple polynomiales) :
  - $X(x) = [f_1(x), \ldots, f_p(x)]$, vecteur $1 \times p$ donnant les valeurs de fonctions de base en x
  - β, vecteur $p \times 1$ des coefficients inconnus, à estimer
- un écart au modèle $\varepsilon(x)$. On suppose que le processus $(\varepsilon(x))_x$ est un processus gaussien stationnaire, centré, de covariance $\gamma(x, y) = \sigma^2 R(||x - y||)$ où $||.||$ est la norme euclidienne et R représente une corrélation isotrope, telle que $R(\rho) = 0$ si $\rho > h$. On supposera R connue [2] ; le terme de variance σ^2, inconnu, est à estimer.

Comme dans (Santner *et al.*, 2003) et dans la littérature géostatique, il nous a paru plus simple de considérer les variables aléatoires comme des fonctions aléatoires de la variable d'espace x et d'utiliser une forme fonctionnelle pour la réponse du simulateur $Y_{\text{sim}}(x)$ et pour le terme d'erreur $\varepsilon(x)$, au lieu des notations de type indiciel $Y_{\text{sim},x}$ et ε_x que l'on rencontre pour les processus aléatoires. Par rapport à la référence précédente où la notation $Y(x)$ est employée, nous avons ajouté le suffixe «sim» de façon à pouvoir clairement différencier dans la suite le modèle pour le simulateur de son approximation (voir paragraphe suivant).

La nature de l'erreur $\varepsilon(x)$ est ici difficile à interpréter. S'agissant d'expériences simulées par ordinateur, son interprétation en tant qu'erreur de mesure, bien que théoriquement possible, n'est pas la plus naturelle. On pourrait en effet interpréter une expérience comme la réalisation d'une variable aléatoire, en considérant le fait que les nombreux «jeux» inéluctablement présents dans le simulateur (critères de convergence, discrétisation en temps, discrétisation en espace, orientation du maillage, etc.) influent sur le résultat obtenu en sortie. Cependant, cela ne permet pas de modéliser la parfaite reproductibilité d'une expérience réalisée avec les mêmes valeurs des facteurs.

On peut davantage parler d'erreur de modèle, car elle vise à expliquer la différence par rapport à des fonctions simples. Cependant, les hypothèses somme toute très minimales sur le terme d'erreur, montrent bien que l'intention est d'approcher le résultat du code par des fonctions de base connues : il s'agit avant tout d'un modèle d'approximation paramétrique. En général, on choisit une forme régulière pour les fonctions de base. Ainsi, dans les applications, la partie modèle sera souvent un polynôme de degré 2. Ceci sous-entend que le phénomène physique sous-jacent est relativement «lisse», ce qui peut parfois ne pas être le cas (voir la section «Discussion»).

[2] Il s'agit de proposer un processus à trajectoires continues dont le comportement s'approche autant que possible de celui d'un bruit blanc. Ainsi, R peut être par exemple donnée par un modèle sphérique, i.e. $R(\rho) = 1 + (1/2)(\rho/h)^3 - (3/2)(\rho/h)$.

2.2. Approximation du simulateur

Pour un plan d'expériences $x^{(1)}, \ldots, x^{(n)}$ donné [3], on approche la réponse du simulateur à partir du formalisme qui suit, conséquence de la formule (1). En notant

$X = [X(x^{(1)})', \ldots, X(x^{(n)})']'$, la matrice d'expériences $n \times p$, et

$Y = (Y_{\text{sim}}(x^{(1)}), \ldots, Y_{\text{sim}}(x^{(n)}))$, le vecteur $n \times 1$ des réponses aux points du plan,

on obtient le modèle :

$$Y = X\beta + E,$$

où le vecteur aléatoire E est de loi $\mathcal{N}(0, \sigma^2 \text{ Id})$.

En d'autres termes, le modèle obtenu est le modèle linéaire standard, et la réponse approchée est alors obtenue en deux étapes :

a) Estimation de β et σ par moindres carrés. Les estimateurs sont donnés par :

- $\widehat{\beta} = (X'X)^{-1}X'Y$ (estimateur de Gauss-Markov),
- $\widehat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - X(x^{(i)})\widehat{\beta})^2$, d'où la covariance estimée $\hat{\gamma}(x,y) = \hat{\sigma}^2 R(||x-y||)$

b) Approximation du simulateur, en remplaçant dans le modèle du simulateur et par leurs estimateurs :

$$Y_{\text{app}}(x) = X(x)\widehat{\beta} + \eta(x) \tag{2}$$

où $(\eta(x))_x$ est un processus gaussien stationnaire centré, de covariance $\hat{\gamma}$.

Remarques :

- Comme dans le modèle classique de la régression $\widehat{\sigma}^2$ est indépendant de $\widehat{\beta}$, et donc le processus $(\eta(x))_x$ est indépendant de $\widehat{\beta}$.
- On notera que dans ce contexte d'évaluation d'incertitude sur la prédiction, on ajoute – comme dans le cadre classique de la régression – le terme d'erreur estimée.

2.3. Propagation des incertitudes

La propagation des incertitudes des facteurs vers la réponse s'effectue par simulation de Monte Carlo sur l'approximation du simulateur [4]. Néanmoins, deux procédés sont possibles selon que l'on tient compte ou non de la variabilité des réponses aux points du plan d'expériences.

---

[3] On supposera que les points $x^{(i)}$ sont à distance supérieure à h les uns des autres, ce qui n'est pas une contrainte forte dans notre contexte.

[4] Nous négligerons dans cette partie introductive l'autocorrélation du processus $(\eta(x))_x$.

2.3.1. Propagation conditionnelle au jeu de réponses

Les étapes sont les suivantes :

1. Calculer les estimations $\widehat{\beta}$ et $\widehat{\sigma}$ de β et σ à partir du plan d'expériences et des réponses du simulateur correspondantes

   pour $r = 1, \ldots, R$, *faire* :
2. Simuler x^{*r} selon la loi de probabilité des facteurs
3. Simuler ε^{*r} de loi $N(0, \widehat{\sigma}^2)$, et calculer y^{*r} selon la formule $y^{*r} = X(x^{*r})\widehat{\beta} + \varepsilon^{*r}$

L'étape 2 permet de tenir compte de la source d'incertitude associée à chaque facteur, et l'étape 3 de l'erreur dans le modèle de régression.

Le défaut de ce procédé est de remplacer purement et simplement le simulateur par son approximation. En effet, cette approximation est réalisée à partir du jeu de réponses données par le simulateur aux points du plan d'expériences. Or le modèle de simulateur est un modèle d'approximation où l'on suppose que ce jeu de réponses est aléatoire (voir § 2.1). Il convient donc de tenir compte de cette source de variabilité dans la propagation des incertitudes. Formellement, les étapes 1-2-3 décrivent un procédé de propagation des incertitudes *conditionnel* à Y. Nous proposons maintenant un procédé non conditionnel.

2.3.2. Propagation non conditionnelle

Pour ce faire, on peut remarquer que la variabilité sur le jeu de réponses apparaît au niveau de l'estimation de β et σ. On peut donc reprendre les étapes 1-2-3 en remplaçant les valeurs fixées $\widehat{\beta}$ et $\widehat{\sigma}$ par des valeurs simulées dans la loi de probabilité des estimateurs correspondants.

Le nouveau procédé peut ainsi s'écrire :

1. Calculer les estimations $\widehat{\beta}$ et $\widehat{\sigma}$ de β et σ à partir du plan d'expériences et des réponses du simulateur correspondantes

   pour $r = 1, \ldots, R$, *faire* :
2. Simuler x^{*r} selon la loi de probabilité des facteurs

3a. Simuler β^{*r} dans la loi $N(\beta', (\sigma')^2(X'X)^{-1})$ et de façon indépendante $(\sigma^{*r})^2$ dans la loi $\frac{(\sigma')^2}{n-p-1}\chi^2_{n-p-1}$. Idéalement, β' et σ' devraient être égaux à β et σ. Comme ces derniers ne sont pas connus, on utilise donc leurs estimations : $\beta' = \widehat{\beta}$ et $\sigma' = \widehat{\sigma}$

3b. Simuler ε^{*r} de loi $N(0, (\widehat{\sigma}^{*r})^2)$, et calculer y^{*r} selon la formule $y^{*r} = X(x^{*r})\beta^{*r} + \varepsilon^{*r}$

Remarque. –

- L'étape 3a. est en fait équivalente à une simulation par ré-échantillonnage (ou bootstrap), et on pourrait écrire :
  3a. Simuler $\varepsilon^1, \ldots, \varepsilon^n$ indépendamment et de loi $N(0, \widehat{\sigma}^2)$, calculer $y^1, \ldots, y^n$

selon la formule $y^i = X(x^{(i)})\widehat{\beta} + \varepsilon^i$, et en déduire de nouvelles estimations β^{*r} et σ^{*r} de β et σ.
Le lecteur intéressé pourra consulter (Davinson, Hinkley, 1997).

- Les étapes 1, 3a et 3b peuvent être remplacées dans notre contexte de régression par le calcul d'intervalles de prévision standards.

3. Construction du critère

3.1. Motivation : un critère pour l'évaluation d'incertitudes

Notation. – Pour l'évaluation d'incertitudes, on s'intéresse aux réalisations de la réponse lorsque les valeurs des facteurs sont tirées aléatoirement suivant la loi des facteurs. Afin de faire apparaître cette considération plus clairement dans les notations, nous noterons $\widetilde{x}$ (x tilde) lorsque les facteurs sont vus comme des variables aléatoires, et x lorsqu'il s'agit d'une variable d'espace.

Par rapport à l'objectif annoncé de fournir des incertitudes sur des variables d'intérêt, il est naturel de rechercher un critère permettant de planifier les expériences $x^{(1)}, \ldots x^{(n)}$ de façon à estimer au mieux la loi de probabilité *non conditionnelle* de la réponse $Y_{\text{sim}}(\widetilde{x})$ du simulateur. Il s'agit donc de la distribution de la réponse du simulateur vu comme un modèle probabiliste dépendant des facteurs aléatoires $\widetilde{x}$. On notera cette distribution d_{sim}.

Or, cette distribution est connue seulement de façon approchée par l'intermédiaire de la distribution d_{app} de l'approximation $Y_{\text{app}}(\widetilde{x})$ donnée par la formule (2). Le critère proposé devra donc permettre de réduire l'écart entre les deux distributions d_{sim} et d_{app}.

3.2. Contrainte : méconnaissance des incertitudes des facteurs

Il serait alors naturel d'utiliser une des nombreuses définitions de «distance» pour les distributions de variable aléatoire, comme la distance de Kullback-Leibler, la distance du χ^2, la distance de Hellinger ... (voir par exemple (Borovkov, 1987)).

Cependant, ces distances qui ont été conçues pour mesurer avec précision l'écart entre des lois, en faisant souvent intervenir l'écart entre les densités, ne sont pas les mieux adaptées ici. En effet, dans notre contexte la loi μ des facteurs sub-surface $\widetilde{x}$ n'est pas connue avec précision, et l'incertitude sub-surface elle-même n'est donc connue que de manière approchée. Choisir une distance basée sur la densité des lois aurait donc l'inconvénient de trop faire dépendre le résultat de l'incertitude de spécification.

Il est donc plus indiqué ici d'utiliser un critère basé sur des éléments assez pauvres des distributions afin d'accroître la robustesse des résultats. C'est pourquoi, nous avons choisi de mesurer l'écart des deux distributions d_{sim} et d_{app} en utilisant uniquement leurs deux premiers moments. Dans le cas courant d'un modèle polynomial d'ordre 2, nous verrons que le critère ne fera alors intervenir la loi des facteurs sub-surface que par leurs 4 premiers moments.

3.3. Définition du critère

Pour définir le critère, il nous reste à remarquer que les deux distributions et ont le même premier moment :

$$E(d_{\text{sim}}) - E(d_{\text{app}}) = E(E(Y_{\text{sim}}(\widetilde{x})|\widetilde{x}) - E(Y_{\text{app}}(\widetilde{x})|\widetilde{x})) = E(X(\widetilde{x})\beta - X(\widetilde{x})\beta) = 0$$

ce qui nous amène à considérer uniquement les moments d'ordre 2, et à proposer la définition du critère C selon la formule qui suit :

$$\boxed{C(x^{(1)}, \ldots, x^{(n)}) = \text{var}(Y_{\text{app}}(\widetilde{x})) - \text{var}(Y_{\text{sim}}(\widetilde{x}))} \tag{3}$$

Il s'agit donc de l'écart de variance dû à l'estimation du modèle du simulateur.

Comme nous le verrons dans la section IV, cette expression est positive; c'est pourquoi nous n'avons pas utilisé de valeur absolue dans la définition. Ce résultat est assez intuitif, puisqu'il y a dans $Y_{\text{app}}(\widetilde{x})$ une source d'erreur supplémentaire provenant de l'estimation du modèle à partir du plan d'expériences.

Nous proposons de nommer ce critère MC-V optimalité, pour Monte Carlo – Variance. Ceci pour rappeler que le critère est conçu dans un objectif de propagation d'incertitudes (effectuée par une méthode de Monte Carlo) et basé sur la différence de Variance entre distribution théorique et distribution approchée.

Naturellement, les plans MC-V optimaux sont obtenus en minimisant ce critère.

4. Expressions du critère de MC-V optimalité

Dans toute la section, nous supposons que le plan d'expériences est connu, ce qui revient à raisonner conditionnellement à $x^{(1)}, \ldots, x^{(n)}$. Nous donnons alors deux expressions de la MC-V optimalité. La première est utile au plan théorique, et aboutira à une autre interprétation du critère (voir § V.1), en lien avec la G-optimalité, et à l'équivalence avec l'IMSE (§ V.2); La seconde est utile au plan pratique pour le calcul effectif du critère.

PROPOSITION. – *Notons* $X = [X(x^{(1)})', \ldots, X(x^{(n)})']'$ *la matrice d'expériences associée aux expériences* $x^{(1)}, \ldots, x^{(n)}$. *On a alors :*

- $C(x^{(1)}, \ldots, x^{(n)}) = \sigma^2\{E(X(\widetilde{x})(X'X)^{-1}X(\widetilde{x})')\}$
- $C(x^{(1)}, \ldots, x^{(n)}) = \sigma^2\{m_{X(\widetilde{x})}(X'X)^{-1}m'_{X(\widetilde{x})} + \text{Tr}((X'X)^{-1}\Gamma_{X(\widetilde{x})}))\}$, *où* $m_{X(\widetilde{x})}$ *et* $\Gamma_{X(\widetilde{x})}$ *sont respectivement l'espérance et la matrice de covariance du vecteur aléatoire* $X(\widetilde{x})$.

Remarques :

- La première expression confirme que le critère de MC-V optimalité est positif (par le fait que la matrice $X'X$ est définie positive).

- Les expressions du critère correspondent au cas où l'on a choisi pour $\tau^2 = \widehat{\sigma}^2$ l'estimateur sans biais de σ^2. Cependant, pour les autres estimateurs classiques (en normalisant par n au lieu de n-p par exemple), le biais se traduit par l'ajout d'une constante au critère (dépendant uniquement de σ^2). Le plan optimal correspondant est donc le même.
- Comme indiqué ci-avant, dans le cas courant d'un modèle polynomial de degré 2 (pour $X(x)$), le critère ne dépend que des 4 premiers moments des facteurs (x).

Démonstration. – D'après la formule de la variance conditionnelle,

$$\mathrm{var}(Y_{\mathrm{sim}}(\widetilde{x})) = E[\mathrm{var}(Y_{\mathrm{sim}}(\widetilde{x})|\widetilde{x})] + \mathrm{var}[(E(Y_{\mathrm{sim}}(\widetilde{x})|\widetilde{x})]$$

$$\mathrm{var}(Y_{\mathrm{app}}(\widetilde{x})) = E[\mathrm{var}(Y_{\mathrm{app}}(\widetilde{x})|\widetilde{x})] + \mathrm{var}[E(Y_{\mathrm{app}}(\widetilde{x})|\widetilde{x})]$$

On a déjà remarqué que les deux premiers moments conditionnels sont égaux :

$$E(Y_{\mathrm{sim}}(\widetilde{x})|\widetilde{x}) - E(Y_{\mathrm{app}}(\widetilde{x})|\widetilde{x}) = X(\widetilde{x})\beta - X(\widetilde{x})\beta = 0$$

Il en résulte :

$$C(x^{(1)}, \ldots, x^{(n)}) = E[\mathrm{var}(Y_{\mathrm{app}}(\widetilde{x})|\widetilde{x})] - E[\mathrm{var}(Y_{\mathrm{sim}}(\widetilde{x})|\widetilde{x})]$$

soit :

$$C(x^{(1)}, \ldots, x^{(n)}) = E[\mathrm{var}(Y_{\mathrm{app}}(\widetilde{x})|\widetilde{x})] - \sigma^2$$

Pour la suite du calcul, nous fixons la source d'aléa due à l'incertitude sur les facteurs, en raisonnant conditionnellement à $\widetilde{x}$. On a :

$$\begin{aligned} Y_{\mathrm{app}}(\widetilde{x}) - E(Y_{\mathrm{app}}(\widetilde{x})|\widetilde{x}) &= X(\widetilde{x})\widehat{\beta} + \eta(\widetilde{x}) - X(\widetilde{x})\beta \\ &= X(\widetilde{x})(\widehat{\beta} - \beta) + \eta(\widetilde{x}) \end{aligned}$$

Or par construction, $\eta(\widetilde{x})$ est indépendant de $\widehat{\beta}$ (voir la 1ère remarque, § 2.2). Il en résulte :

$$\mathrm{var}(Y_{\mathrm{app}}(\widetilde{x})|\widetilde{x}) = \mathrm{var}(X(\widetilde{x})(\widehat{\beta} - \beta)|\widetilde{x}) + \mathrm{var}(\eta(\widetilde{x})|\widetilde{x})$$

D'où, en notant $\mathrm{var}(\widehat{\beta}|\widetilde{x})$ la matrice de covariance de $\widehat{\beta}$ conditionnellement à $\widetilde{x}$:

$$\mathrm{var}(Y_{\mathrm{app}}(\widetilde{x})|\widetilde{x}) = X(\widetilde{x})\mathrm{var}(\widehat{\beta}|\widetilde{x})X(\widetilde{x})' + \widehat{\sigma}^2$$

Rendons maintenant à $\widetilde{x}$ son caractère aléatoire pour terminer l'évaluation du critère. Notons que d'après le théorème de Gauss-Markov, $\mathrm{var}(\widehat{\beta}|\widetilde{x}) = \sigma^2(X'X)^{-1}$, et $\widehat{\sigma}^2$ est

un estimateur non biaisé, c'est-à-dire ici : $E(\widehat{\sigma}^2|\widetilde{x}) = \sigma^2$, et *a fortiori* $E(\widehat{\sigma}^2) = \sigma^2$. Finalement,

$$E[\mathrm{var}(Y_{\mathrm{app}}(\widetilde{x})|\widetilde{x})] = \sigma^2 E(X(\widetilde{x})(X'X)^{-1}X(\widetilde{x})') + \sigma^2$$

d'où la première expression :

$$C(x^{(1)}, \ldots, x^{(n)}) = \sigma^2 E(X(\widetilde{x})(X'X)^{-1}X(\widetilde{x})')$$

Pour obtenir la seconde, il suffit de remarquer que, si Q est une matrice symétrique et u un vecteur aléatoire centré,

$$E(uQu') = \sum_{k,l} q_{k,l}E(u_k u_l) = \sum_{k,l} q_{k,l}\mathrm{cov}(u)_{k,l} = \mathrm{Tr}(Q\mathrm{cov}(u))$$

En centrant le vecteur $X(x)$, on obtient alors :

$$\begin{aligned} C(x^{(1)}, \ldots, x^{(n)}) =& \sigma^2\{E(m_{X(\widetilde{x})}(X'X)^{-1}m'_{X(\widetilde{x})}) \\ &+ E((X(\widetilde{x}) - m_{X(\widetilde{x})})(X'X)^{-1}(X(\widetilde{x}) - m_{X(\widetilde{x})})') + 2 \times 0\} \end{aligned}$$

d'où :

$$C(x^{(1)}, \ldots, x^{(n)}) = \sigma^2\{m_{X(\widetilde{x})}(X'X)^{-1}m'_{X(\widetilde{x})} + \mathrm{Tr}((X'X)^{-1}\Gamma_{X(\widetilde{x})})\}$$

Remarque. –

- Des expressions analogues du critère peuvent être obtenues en présence d'hétéroscédasticité, c'est-à-dire lorsque le terme d'erreur $\varepsilon(x)$ n'est plus de variance constante. Par exemple, si on suppose que $\varepsilon(x)$ est de loi $N(0, \sigma(x)^2)$, et si l'on note Γ_ε la matrice de variance-covariance du vecteur des erreurs aux points de données $(\varepsilon(x^{(1)}), \ldots, \varepsilon(x^{(n)}))'$,

$$\Gamma_\varepsilon = \begin{pmatrix} \sigma(x^{(1)})^2 & 0 & 0 \\ 0 & \ldots & 0 \\ 0 & 0 & \sigma(x^{(n)})^2 \end{pmatrix}$$

les expressions précédentes doivent être modifiées en remplaçant $\sigma^2(X'X)^{-1}$ par la forme plus générale de la matrice de covariance de $\widehat{\beta}$: $(X'\Gamma_\varepsilon^{-1}X)^{-1}$ (notons que les $\eta(x)$ restent indépendants de $\widehat{\beta}$ dans ce contexte).

5. Liens avec d'autres critères

5.1. Une autre interprétation du critère – Lien avec la G-optimalité

La première expression du critère de MC-V optimalité fait apparaître une interprétation en terme de prévision. En effet, dans le cadre qui est le nôtre,

$$Y_{\text{sim}}(x) = X(x)\beta + \varepsilon(x)$$

l'estimation du terme linéaire est donnée par l'estimateur :

$$\widehat{Y}(x) = X(x)\widehat{\beta}$$

Comme $\widehat{\beta}$ est de loi normale $N((\beta, \sigma^2(X'X)^{-1})$ (théorème de Gauss-Markov), on en déduit que $\widehat{Y}(x)$ est de loi normale d'espérance $X(x)\beta$ et de variance $\sigma^2 X(x)(X'X)^{-1}X(x)'$. On voit donc que l'on aboutit à une troisième expression pour la MC-V optimalité :

PROPOSITION. – *En notant $\widehat{Y}(x)$ l'estimation en x de $X(x)\beta$, $\widehat{Y}(x) = X(x)\widehat{\beta}$, on a :*

$$C(x^{(1)}, \ldots, x^{(n)}) = E[\text{var}(\widehat{Y}(\widetilde{x})|\widetilde{x})]$$

Il s'agit donc de la moyenne des carrés des erreurs d'estimation, moyenne réalisée selon la distribution des facteurs.

Naturellement, cette expression est à rapprocher du critère de G-optimalité :

$$G(x^{(1)}, \ldots, x^{(n)}) = \max_x(\text{var}(\widehat{Y}(x))$$

5.2. Équivalence avec l'IMSE

Le critère d'IMSE consiste à faire la moyenne des carrés des erreurs de prévision :

$$IMSE = \int_{x\in\Delta} E[(\widehat{Y}(x) - Y_{\text{sim}}(x))^2]\text{d}\mu(x)$$

où on a noté μ la loi de probabilité des facteurs x, et Δ le domaine expérimental.

En général, le choix de la mesure μ n'est pas aisé, et on prend souvent, faute de mieux, la mesure uniforme. Cependant, dans notre contexte, nous disposons souvent *a priori* d'une distribution des facteurs, celle proposée par les experts métier. Dans l'exemple de l'Exploration/Production pétrolière, il s'agit de la distribution μ des facteurs sub-surface. Nous allons voir qu'avec ce choix de distribution, l'IMSE est équivalent, au critère de MC-V optimalité lorsque le modèle du simulateur se

rapproche d'un bruit blanc, c'est-à-dire lorsque h tend vers 0 dans la structure de corrélation.

En effet, dans ce contexte, l'IMSE peut s'exprimer selon :

$$IMSE = E(E[(\widehat{Y}(\widetilde{x}) - Y_{\text{sim}}(\widetilde{x}))^2|\widetilde{x}])$$

soit, en remarquant que $E(\widehat{Y}(\widetilde{x})|\widetilde{x}) = X(\widetilde{x})\beta = E(Y_{\text{sim}}(\widetilde{x})|\widetilde{x})$:

$$IMSE = E[\text{var}(\widehat{Y}(\widetilde{x}) - Y_{\text{sim}}(\widetilde{x})|\widetilde{x})]$$

Or, en développant la variance conditionnelle, il vient :

$$\begin{aligned}\text{var}(\widehat{Y}(\widetilde{x}) - Y_{\text{sim}}(\widetilde{x})|\widetilde{x}) &= \text{var}(X(\widetilde{x})(\widehat{\beta} - \beta) - \varepsilon(\widetilde{x})|\widetilde{x}) \\ &= \text{var}(X(\widetilde{x})(\widehat{\beta} - \beta)|\widetilde{x}) + \text{var}(\varepsilon(\widetilde{x})|\widetilde{x}) \\ &\quad + 2\text{cov}(X(\tilde{x})(\hat{\beta} - \beta); \varepsilon(\tilde{x})|\tilde{x})\end{aligned}$$

Un calcul direct montre que le terme de covariance est égal à $X(x)(X'X)^{-1}X'\gamma_n$, avec $\gamma_n = (\gamma(x, x^{(1)}), \ldots, \gamma(x, x^{(n)}))'$. Faisons maintenant tendre h tend vers 0. Ce dernier terme tend alors vers 0 et on a :

$$\text{var}(\widehat{Y}(\widetilde{x}) - Y_{\text{sim}}(\widetilde{x})|\widetilde{x}) = \text{var}(\widehat{Y}(\widetilde{x})|\widetilde{x}) + \sigma^2$$

d'où finalement, dans le cas où μ est la distribution des facteurs connue *a priori* :

$$IMSE(x^{(1)}, \ldots, x^{(n)}) = C(x^{(1)}, \ldots, x^{(n)}) + \sigma^2$$

Remarques :

- L'équivalence avec l'IMSE justifie *a posteriori* l'emploi de ce dernier pour la spécificité du problème rencontré. En effet, l'emploi de l'IMSE avec la distribution des facteurs peut maintenant s'interpréter non seulement comme un simple cumul, mais *en fonction de l'objectif de quantification d'incertitudes.*
- On trouve également dans (Santner, Williams, Notz, 2003, § 6.2.2.) un lien entre les plans IMSE optimaux et les plans L-optimaux, ainsi que les plans A-optimaux.

6. Algorithme de calcul et exemples

6.1. Un algorithme de calcul des plans MC-V optimaux en petite dimension

Le calcul de plans IMSE-optimaux est un problème assez difficile en général. Les algorithmes de type quasi-Newton peuvent convenir à condition de prendre plusieurs points d'initialisation pour ne pas rester bloqué dans un minima local (Sacks,

Schiller, Welch, 1989; Santner, Williams, Notz, § 6.2.). En petite dimension, nous avons choisi de calculer les plans MC-V optimaux par un algorithme d'échange analogue à celui de Fedorov pour le calcul de plans D-optimaux. Ce dernier est présenté par exemple dans (Benoist, Tourbier, Germain-Tourbier, 1994), et nous en rappelons maintenant le principe.

L'algorithme repose sur une discrétisation de l'espace $[-1,1]^k$ en un ensemble fini S [5] . On initialise le plan d'expériences en choisissant au hasard uniformément et sans remise n points dans S. Le plan ainsi constitué est noté $X = \{x^{(1)}, \ldots, x^{(n)}\}$. On notera $S' = S \setminus X$, ensemble des points de S privés de ceux de X. On améliore alors de façon itérative le plan initial en procédant par échange : à chaque itération, on cherche à remplacer un point de X par un point de S' de façon à faire décroître le critère le plus possible. On choisit donc un point x de X et un point y de S' tels que le plan $V = (X \setminus \{x\}) \cup \{y\}$ (plan privé de l'expérience x auquel on a ajouté l'expérience y) soit de critère minimal. On met à jour X et S' en fonction du couple (x, y) trouvé (en prenant comme nouveau X le plan V ainsi obtenu). On arrête l'algorithme quand la diminution du critère est considérée comme non significative. Enfin, de façon à limiter la dépendance du résultat par rapport au plan initial, on réitère le procédé plusieurs fois.

Remarques :

- Chaque itération nécessite $n \times (N - n)$ évaluations du critère afin de trouver le meilleur couple (x, y), où N est le cardinal de S. Pour chaque plan initial, la recherche est exhaustive.
- Alors que dans le cas D-optimal le critère peut être mis à jour en tenant compte de la relation suivante :

$$\det({}^{x}M_{y}) = \det(M)[1+yM^{-1}y'+(xM^{-1}y')^2-xM^{-1}x'-(xM^{-1}x')(yM^{-1}y')]$$

avec $M = X'X$ et ${}^{x}M_{y} = V'V$ (avec $V = (X \setminus \{x\}) \cup \{y\}$), nous n'avons pas trouvé de mise à jour du critère en ce qui concerne la MC-V optimalité.

- Plus la discrétisation est fine et plus le plan optimal est bon. La qualité du plan dépend donc du temps de calcul que l'on s'autorise. Une question non étudiée ici est de trouver le bon compromis discrétisation/qualité.

6.2. Exemples

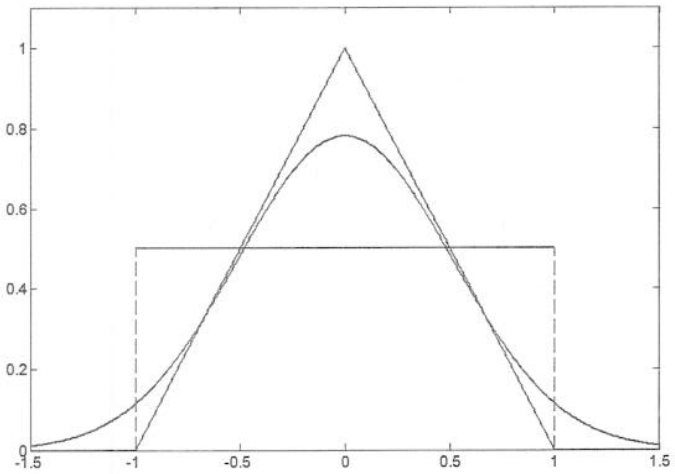
Afin d'apprécier le fonctionnement du critère de MC-V optimalité, on a calculé les plans MC-V optimaux pour 3 variables et 15 expériences, sur le domaine expérimental cubique $[-1,1]^3$. Trois cas ont été examinés, correspondant à trois lois de probabilité des facteurs : dans chaque cas, tous les facteurs sont supposés avoir la même loi, correspondant à l'une des lois suivantes :

- loi uniforme sur $[-1,1]$;
- loi normale, telle que $[-1,1]$ supporte 95 % de la masse ($N(0,\sigma^2)$ avec $\sigma = 1/1.96$);

[5] Ce qui assure au passage la condition de distance supérieure à h entre tous les points du plan.

– loi «triangulaire» sur $[-1, 1]$ (densité $1 - |x|$).

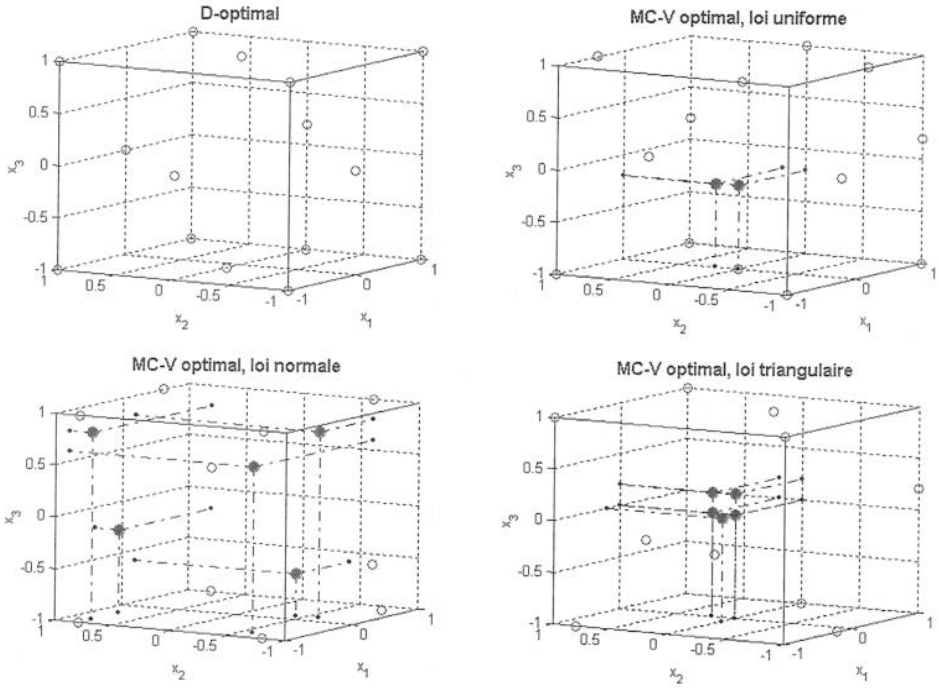
Leur densité respective est représentée ci-dessous.



La partie modèle est un polynôme de degré 2 complet. Par conséquent, le critère ne fait intervenir les lois des facteurs que par leurs 4 premiers moments, et même ici, par symétrie, seulement par leurs moments d'ordre 2 et 4 :

	Moment d'ordre 2	Moment d'ordre 4
Loi uniforme	$1/3 = 0.3333\ldots$	$1/5 = 0.2$
Loi normale	$\sigma^2 = 0.2603\ldots$	$3\sigma^4 = 0.2033\ldots$
Loi triangulaire	$1/6 = 0.1666\ldots$	$1/15 = 0.0666\ldots$

La discrétisation choisie pour l'algorithme est telle que chaque dimension de $[-1, 1]^3$ est discrétisée de façon uniforme en 11 niveaux. La figure ci-dessous représente les résultats obtenus.



Dans le plan D-optimal toutes les expériences sont situées au bord du domaine cubique; en particulier, tous les sommets sont occupés. Cela n'est pas surprenant car le critère D-optimal minimise l'erreur d'estimation.

Dans les plans MC-V optimaux, certains sommets sont vides et quelques expériences sont placées à l'intérieur du domaine. En fait le critère de MC-V optimalité réalise un compromis entre placer des points dans une zone de probabilité non nulle et assurer la qualité de l'estimation. La grappe de points que l'on observe dans le cas de lois triangulaires, pourrait ainsi s'expliquer par la «réaction» du critère au fait que la densité de probabilité d'observer une expérience est nulle au bord du domaine. Lorsque cette probabilité est plus importante, les points à l'intérieur sont toujours présents mais en plus petit nombre.

7. Discussion

Dans les sections précédentes, nous avons construit un critère adapté pour répondre à l'objectif de quantification d'incertitudes d'une variable d'intérêt, obtenue comme la sortie d'un code complexe. Ce critère est en fait équivalent à l'IMSE avec intégration selon la distribution des facteurs sub-surface. Pratiquement, l'optimisation de l'IMSE aboutit à des plans d'expériences originaux, avec davantage de points situés à l'intérieur du domaine expérimental que les plans optimaux classiques. En fait, l'optimisation réalise un compromis entre la nécessité de placer des points aux

bords pour bien contrôler l'estimation, et la nécessité de placer des points aux endroits où la probabilité d'observation est forte.

Nous évoquons maintenant deux problèmes pour l'amélioration du critère.

7.1. Définition du critère

Comme nous l'avons souligné au moment de la définition du critère, le fait que la MC-V optimalité soit basée sur des éléments assez pauvres des distributions lui confère une certaine robustesse relativement aux erreurs de spécification des incertitudes sub-surface. Cependant, le critère retenu, basé sur la variance, n'est pas parfaitement adapté à l'évaluation des incertitudes telle qu'elle est réalisée en pratique, sous la forme de quantiles à 10, 50 et 90 %.

7.2. Adaptation du critère aux modèles d'interpolation

Dans cet article, l'approche générale est basée sur une modélisation du bruit menant à un modèle de régression linéaire. Or, ce modèle ne prend pas en compte la spécificité des expériences numériques, à savoir leur parfaite reproductibilité. Pour pallier cet inconvénient, des modèles d'interpolation ont été développés. Les articles fondateurs sont ceux de Sacks, Welch, Mitchell, Wynn (1989) et Sacks, Schiller, Welch (1989). Voir également les présentations de Jones, Schonlau, Welch (1998), Jourdan, Collombier (2001) et Jourdan (2002). Dans ces approches de nature non paramétrique, l'interpolation peut s'interpréter en terme de fonctions splines, ce qui lui confère une plus grande souplesse que les bases de fonction polynomiales (voir Jones, Schonlau, Welch (1998) ou Wahba (1990)).

Ces modèles commencent à être employés de façon plus systématique pour l'approximation de codes complexes, et une adaptation du critère dans ce cadre est à prévoir.

Remerciements. – Nous tenons à remercier Xavier BAY, Eric TOUBOUL et André HAAS pour toutes leurs remarques, toujours constructives, qui nous ont été très utiles pour la rédaction de cet article. Nous sommes également reconnaissants aux deux rapporteurs pour leurs remarques constructives qui nous ont permis d'améliorer la forme de cet article, ainsi qu'au professeur Georges OPPENHEIM pour sa lecture approfondie ayant permis une amélioration substantielle du texte.

Références

BENOIST D., TOURBIER Y. et GERMAIN-TOURBIER S. (1994), *Plans d'expériences : construction et analyse*, Lavoisier.

BOROVKOV A. (1987), *Statistique mathématique*, Editions Mir Moscou.

BOX G.E.P., DRAPER N.R. (1987), *Empirical model-building and response surfaces*, Wiley.

DAVINSON A.C. and HINKLEY D.V. (1997), *Bootstrap Methods and their Applications*, Cambridge University Press.

JOBSON J.D. (1991), *Applied Multivariate Data Analysis, Volume I : Regression and Experimental Design*, Springer-Verlag.

JONES D.R., SCHONLAU M. and WELCH W.J. (1998), Efficient global optimisation of expensive black-box function, *Journal of Global Optimization*, **13**, 455-492.

JOURDAN A. (2002), Approche statistique des expériences simulées, *Revue de Statistique Appliquée*, L, N° 1.

JOURDAN A. et COLLOMBIER D. (2001), Régression trigonométrique et plans d'échantillonnage pour expériences simulées, *Revue de Statistique Appliquée*, XLIX, N° 2.

SACKS J., WELCH W.J., MITCHELL T.J. and WYNN H.P. (1989), Design and Analysis of Computer Experiments, *Statistical Science*, 4, N° 4, 409-435.

SACKS J., SCHILLER S.B. and WELCH W.J. (1989), Designs for Computer Experiments, *Technometrics*, **31**, N° 1, 41-47.

SANTNER T.J., WILLIAMS B.J. and NOTZ W.I. (2003), *The design and analysis of computer experiments*, Springer.

WAHBA G., (1990), *Spline models for observational data*, Society for industrial and applied mathematics.

BIBLIOGRAPHIE

HISTOIRE DES SONDAGES

par J. ANTOINE

1 vol, 286 pages, Editions ODILE JACOB (15 rue Soufflot, 75005, PARIS) 2005, 25,90 euros, ISBN 2 7381 1587 X

Jacques ANTOINE, premier dirigeant de la SOFRES, raconte dans cet ouvrage l'histoire des sondages des origines jusqu'à nos jours. Outre une introduction, une conclusion et un index, le livre comporte 7 chapitres.

Le premier est relatif à la préhistoire et à la genèse des sondages avec en particulier le succès des sondages préélectoraux lors de l'élection présidentielle américaine de 1936, l'apport des psychologues et sociologues, l'origine des études de marché, etc. Le second chapitre traite de la naissance des sondages en France avec la création de l'IFOP, de la SOFRES, et les premiers sondages de la statistique publique, tandis que le troisième est relatif à l'organisation de la profession, avec les problèmes de déontologie et de secret statistique, la loi informatique et libertés, la qualité des sondages, le statut des enquêteurs, etc. Le quatrième chapitre étudie les sondages politiques avec les erreurs dans les sondages préélectoraux, les sondages en France de 1997 (élections législatives) et 2002 (élections présidentielles), la réglementation de ces sondages avec la loi de 1977 créant en particulier la commission des sondages. Le cinquième chapitre traite des sondages marketing, avec les panels et études de marché, les études qualitatives et les champs d'application de ces sondages. Le sixième chapitre est relatif à l'audience des média (presse quotidienne, radio, télévision, etc.) tandis que le dernier chapitre étudie un autre champ d'application des sondages, celui de la recherche économique et sociale avec l'étude des conditions et dépenses de logement, des consommations alimentaires, des dépenses de santé, d'habillement, d'équipement et d'ameublement, de loisirs, etc

L'ouvrage se termine par une conclusion sur l'évolution de la société d'ici 2050 et ses répercussions sur les sondages.

Ce livre écrit par l'un de ceux qui a participé au développement des sondages en France est passionnant. Il s'adresse bien sûr aux statisticiens, mais de façon plus générale à tous ceux qui sont intéressés (hommes politiques, sociologues, économistes, etc) par les sondages.

OUVRAGE REÇU

J.M. LEGAY, A.M. SCHMID : Philosophie de l'interdisciplinarité. Correspondance (1999-2004) sur la recherche scientifique, la modélisation et les objets complexes, 1 Vol., 300 pages, Editions PETRA (12 rue de la Réunion, 75020, Paris) 2004, 25 euros, ISBN 2-84743-004-3

louisjeanimprimeur
59, Avenue Émile DIDIER - 05003 GAP Cedex - Tél. 04 92 53 17 00 • Dépôt légal : 643 – Novembre 2005
Imprimé en France