THÈSE DE DOCTORAT DE
L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité : **Mathématiques**

Option : **Statistique**

Présentée par :

Vincent FEUILLARD

sujet de la thèse :

ANALYSE D'UNE BASE DE DONNÉES POUR LA CALIBRATION D'UN CODE DE CALCUL

Soutenue le 21 mai 2007 devant le jury composé de :

Directeur de Thèse	M. Paul Deheuvels	Université Paris VI
Rapporteurs	M. Jean-Jacques Droesbeke	Université Libre de Bruxelles
	M. Stéphane Lallich	Université Lyon II
Examinateurs	M. Laurent Carraro	École Nationale des Mines de Saint Étienne
	M. Nicolas Devictor	Responsable C.E.A.
	M. Roger Phan-Tan-Luu	Université Aix-Marseille III

Thèse réalisée dans le cadre d'un Contrat de Formation par la Recherche au Commissariat à l'Energie Atomique

A mes parents,
ma petite soeur,
mes frères,
Maud.

REMERCIEMENTS

A Monsieur le Professeur Paul Deheuvels pour l'excellence de son soutien scientifique et ses nombreux conseils.

A Nicolas Devictor, pour la qualité de son encadrement et sa bienveillance. Cette thèse n'aurait pu voir le jour sans la grande confiance qu'il a bien voulu m'accorder.

A Bertrand Iooss, pour sa disponibilité, son écoute, ses conseils avisés et ses qualités humaines. Nos nombreuses discussions m'ont grandement aidé et réconforté lors de mes recherches.

A Monsieur le Professeur Roger Phan-Tan-Luu pour l'orientation de mon travail.

A Messieurs les Professeurs, Jean-Jacques Droesbeke et Eric Lallich, rapporteurs, pour leur lecture attentive et leurs remarques pertinentes.

Aux membres du jury, pour avoir eu l'amabilité de participer à la soutenance.

Au Commissariat à l'Energie Atomique pour le financement de ma thèse.

A l'équipe du LCFR qui m'a permis d'évoluer dans un environnement convivial et agréable.

A toutes celles et ceux que j'ai rencontrés lors de mes trois années passées au C.E.A. de Cadarache, particulièrement à mes collocataires Stouf et Toinou pour leur soutien amical, à Nicolas, Elie, Amandine, Mila, Léopoldine, Mo, Pierre, Mathieu, Damien, Patrick, Eric, Guilhem, Laure, Claire, et à mes amis de toujours, Dan, Atsa, Aurélie et Aurore.

Table des matières

Introduction **15**

1 Critères déterministes **25**
1.1 Contexte . . . 25
1.2 Les critères . . . 26
1.2.1 Notations . . . 26
1.2.2 Critères associés aux distances . . . 27
1.2.2.1 Distances entre points de la BDDE . . . 27
1.2.2.2 Distances entre points de la BDDE et points de l'espace . 30
1.2.3 Critères associés aux volumes des cellules de Voronoi . . . 35
1.2.4 Récapitulatif . . . 36
1.3 Discrépance . . . 38
1.3.1 Notations . . . 38
1.3.2 Définitions . . . 43
1.3.3 Propriétés . . . 47
1.3.3.1 Inégalité . . . 47
1.3.3.2 Propriétés des suites aléatoires de loi uniforme sur $[0,1]^d$. 51
1.3.3.3 Expressions, Discussion . . . 55
1.3.4 Récapitulatif . . . 58
1.4 Utilisation, méthodologie . . . 60
1.4.1 Etude d'une base de données d'entrée . . . 60
1.4.2 Sélection de points . . . 65
1.4.2.1 Méthode 1 . . . 65
1.4.2.2 Méthode 2 . . . 67
1.4.2.3 Méthode 3 . . . 68
1.4.3 Application des méthodes de sélection . . . 69
1.4.4 Spécification de points . . . 73
1.4.4.1 Méthode 1 . . . 73
1.4.4.2 Méthode 2 . . . 74
1.4.4.3 Méthode 3 . . . 75
1.4.5 Application des méthodes de spécification . . . 75
1.4.6 Discussion . . . 78

2 Liens entre discrépance et estimation non-paramétrique 81
2.1 Introduction . . . 81
2.2 Inégalité de Koksma-Hlwaka généralisée . . . 83
2.2.1 Notations et hypothèse . . . 83
2.2.2 Inégalité généralisée de Koksma-Hlwaka . . . 83
2.2.3 Considération du processus empirique uniforme . . . 86
2.3 Majoration de critères . . . 87
2.3.1 Introduction . . . 87
2.3.2 Majoration de l'$\mathbb{IMSE}$ . . . 88
2.3.3 Majoration de la $\mathbb{MSE}$ . . . 94
2.3.4 Interprétation . . . 96
2.4 Cadre méthodologique . . . 96
2.5 Application . . . 99
2.5.1 Présentation de l'exemple . . . 99
2.5.2 Analyse initiale des points disponibles . . . 100
2.5.3 Sélection d'un sous-ensemble de points . . . 101
2.5.4 Estimation et Validation . . . 105
2.6 Discussion . . . 107

3 Critères probabilistes 111
3.1 Notion de Test Statistique . . . 113
3.2 Partition du pavé unité . . . 115
3.3 Test sur le vecteur de paramètres d'une loi multinomiale . . . 118
3.3.1 Espérance et Variance de $S_{f_{n,k}}(Y_1, \ldots, Y_k)$ . . . 119
3.3.2 Test de Pearson . . . 121
3.3.3 Test du rapport de vraisemblance . . . 123
3.3.4 « *ϕ-divergence family* » . . . 125
3.3.5 Discussion . . . 129
3.4 « *Sparse case* » . . . 129
3.4.1 Application du théorème de Holst (1972) . . . 130
3.4.2 Recherche du « plus grand pavé vide » . . . 133
3.5 « *Scan Statistics* » . . . 135
3.5.1 Approximation par une loi de Poisson conditionnelle . . . 135
3.5.2 Lois des « *conditional two-dimensional discrete scan statistic* » . . . 138
3.5.3 Cas continu . . . 142
3.6 Utilisation . . . 144
3.7 Discussion . . . 148

Conclusion 151

Annexe 157
4.1 « Connaissance » de la fonction de code . . . 157
4.2 Absence de « Connaissance » de la fonction de code . . . 159
4.2.1 Différence entre réponses du code et réponses expérimentales . . . 160

4.2.2 Méthode GLUE 161
4.2.3 Approche de Kennedy et O'Hagan (2001) 162
4.2.4 Méthodes de « calibration multivariée » 167
4.2.4.1 Approche indirecte 167
4.2.4.2 Approche directe 168
4.3 Discussion 170

Table des figures

1.1 Grille de Sukharev (25 points) . . . 31
1.2 Approximation de la dispersion . . . 32
1.3 Régions de Voronoi d'un réseau . . . 34
1.4 Pavé « ancré à l'origine » associée à$z = (0.6, 0.7)$ . . . 47
1.5 Pavé « centré » associé à$z = (0.6, 0.7)$ . . . 47
1.6 Union de pavé « pairs » associée à$z = (0.5, 0.7)$ . . . 47
1.7 Densité de $n \times \mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2$ . . . 53
1.8 Base de données d'entrée étudiée (400 points) . . . 61
1.9 Exemple d'un ensemble de points redondants . . . 67
1.10 Évolution de la discrépance par application de l'algorithme A_1 . . . 71
1.11 Évolution de la discrépance par application de l'algorithme A_2 . . . 71
1.12 Évolution de la discrépance par application de l'algorithme A_3 . . . 71
1.13 Évolution de la discrépance en fonction de ε (Algorithme A_3) . . . 71
1.14 Évolution de la discrépance par application de l'algorithme B_1 . . . 76
1.15 Évolution de la discrépance par application de l'algorithme B_2 . . . 76
1.16 Évolution de la discrépance par application de l'algorithme B_3 . . . 76
1.17 Évolution de la discrépance L2 . . . 79
1.18 « Pathologie » d'une suite de Halton (100 points) . . . 79

2.1 Ensemble $\boldsymbol{x}(100)$ initial . . . 99
2.2 Densité de $n \times \mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2$ . . . 101
2.3 Points sélectionnés par l'algorithme A_1 . . . 102
2.4 Évolution de Q par l'algorithme A_1 . . . 102
2.5 Points sélectionnés par l'algorithme A_2 . . . 103
2.6 Évolution de Q par l'algorithme A_2 . . . 103
2.7 Points sélectionnés par l'algorithme A_3 . . . 104
2.8 Évolution de Q par l'algorithme A_3 . . . 104
2.9 Évolution de Q en fonction de la distance ε par l'algorithme A_3 . . . 104
2.10 Validation de l'estimation $\hat{f}_{ini}(\cdot)$ . . . 106
2.11 Validation de l'estimation $\hat{f}_{A_1}(\cdot)$ . . . 106
2.12 Validation de l'estimation $\hat{f}_{A_3}(\cdot)$ . . . 106
2.13 Validation de l'estimation $\hat{f}_{A_2}(\cdot)$ . . . 106

3.1 Ensemble de $n = 150$ points dans $\mathcal{X} = [0,1)^2$ 112
3.2 Représentation du plus grand pavé vide 134
3.3 Pavé constitué de 2×2 cellules 139
3.4 Ensemble de 150 points dans $\mathcal{X} = [0,1)^2$ 145
3.5 Localisation d'un groupe de points 147

Introduction

CONTEXTE GÉNÉRAL

Au cours de ces dernières décennies, le développement général des mathématiques appliquées a entrepris de représenter systématiquement les phénomènes expérimentaux à l'aide de *modèles*. Nous nous placerons ici, plus particulièrement, dans le cadre de la statistique et de l'informatique, pour lequel un *modèle* désignera un *code de calcul*. Celui-ci sera la version numérique d'un ensemble d'équations mathématiques, issues, par exemple, d'applications dans des domaines tels que la physique ou la chimie. Un tel *modèle* dépend naturellement d'un ensemble de paramètres. Dans le cadre de l'utilisation d'un *code de calcul*, les *réponses du modèle* constituent des *expériences simulées*, qui pour être pleinement exploitables doivent être *validées*. Il faut au préalable vérifier que *les réponses du modèle « représentent de façon acceptable »* des réponses observées. Cette étape est généralement connue sous l'appellation de la ***calibration du code de calcul***.

La calibration consiste à déterminer le ou les paramètres de façon à ce que le modèle s'approche au mieux de la réalité.

Cette opération se pratique dans un large éventail de disciplines scientifiques. Donnons quelques exemples.

- la calibration, le « calage », des paramètres d'un code simulant les énergies reçues par une cible dans différentes conditions expérimentales (Sancandi (2006)),
- l'estimation du taux d'émission de radionucléides (ou radioisotopes) pour des modèles d'accidents nucléaires (Kennedy et O'Hagan (2001a)),
- en hydrologie, pour estimer des paramètres liés à la transmissivité hydraulique dans des modèles de pluie (Romanowicz (2006)).

Dans beaucoup d'applications de la calibration, il apparaît que la base de données d'entrée (BDDE) du modèle a une grande influence sur la précision de l'estimation des paramètres (dans un sens qui sera défini et illustré plus loin). Il est donc important de disposer de critères permettant d'apprécier la qualité de la BDDE. Lorsqu'on ne dispose pas de connaissances précises sur le modèle (ou code de calcul), ces critères doivent être les plus généraux possibles afin d'être utilisés quel que soit le domaine d'application et quelle que soit la méthode d'estimation. Leur objectif est de vérifier que les variables

de la BDDE ont une répartition aussi « uniforme » que possible dans leur domaine de variation. Pour ce faire, il existe deux grandes familles de critères : les critères dits « déterministes », et les critères dits « probabilistes ».

Ainsi cette thèse traite, au sein de la problématique de la calibration, la qualité d'une base de données au sens de sa répartition uniforme. Avant d'en présenter le plan, nous allons préciser le problème de la calibration et la méthodologie générale que nous avons considérée pour la résoudre.

PROBLÉMATIQUE DE LA CALIBRATION

Afin de formaliser la présentation du problème de calibration, précisons les notations de base ainsi que les différentes variables qui seront considérées par la suite.

√ Nous notons $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ *une base de données d'entrée* (BDDE) avec $x_i = (x_i^{(1)}, \ldots, x_i^{(d)})^t \in \mathcal{X}$. En pratique, à cause de problèmes liés à la précision des observations, il est naturel d'observer de nombreux « collages » (« *ties* » en anglais) dans les bases de données, faisant donc place à des observations multiples. Le caractère distinct des observations sera, ici, admis, pour développer leur analyse théorique. Les données $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ seront donc supposées distinctes.

Elles correspondent aux « points » d'un espace $\mathcal{X}$ dans lequel les observations prennent leurs valeurs. En général $\mathcal{X}$ pourra être choisi soit comme une partie fermée d'un espace de Hilbert, soit, en dimension finie, comme une partie fermée bornée de $\mathbb{R}^d$. Cependant la plupart des applications présentées seront obtenues lorsque $\mathcal{X}$ est une partie compacte de $\mathbb{R}^d$, ce qui sera supposé être le cas implicitement par la suite. En physique ou en chimie, ces points peuvent correspondre à des conditions expérimentales. Pour chaque réponse i, nous avons alors d conditions expérimentales scalaires à valeurs dans un ensemble $\mathcal{X}$.

√ Nous notons $\boldsymbol{y}(n) = \{y_1 = y(x_1), \ldots, y_n = y(x_n)\}$ *l'ensemble des valeurs de réponse observées*, elles-mêmes fonctions des données de la BDDE.

En pratique, nous supposons que $y_i \in \mathcal{Y}, \forall i \in \{1, \ldots, n\}$, où $\mathcal{Y}$ est une partie (en général un intervalle) spécifiée de $\mathbb{R}$. En physique ou en chimie ce sont les résultats expérimentaux obtenus pour chaque condition x_i.

√ Nous notons $\theta \in \Theta$, *le vecteur des paramètres* intervenant dans le *modèle*.

Il s'agit du *paramètre* (vectoriel) de calibration. C'est celui que nous cherchons à estimer pour que le *modèle* s'approche au mieux de la réalité (selon des critères énoncés plus loin). Le plus souvent nous supposerons que l'espace des paramètres Θ est l'adhérence d'une partie ouverte de $\mathbb{R}^p$.

√ Nous désignons par $\mathcal{M}(\theta)$, le *modèle*. Celui-ci représente la règle de calcul permettant d'obtenir les réponses à partir des entrées x_i.

Plus concrètement, le modèle $\mathcal{M}(\theta)$ représente une valeur approchée du vecteur

d'observations

$$\boldsymbol{y}(n) = \{y_1 = y(x_1), \ldots, y_n = y(x_n)\},$$

par une approximation de la forme

$$\boldsymbol{z}(n) = \{z_1 = f(x_1, \theta) \ldots, z_n = f(x_n, \theta)\}.$$

Le modèle est donc résumé par la fonction $f(x, \theta)$. La fonction f est appelée *code* ou *fonction de code*.

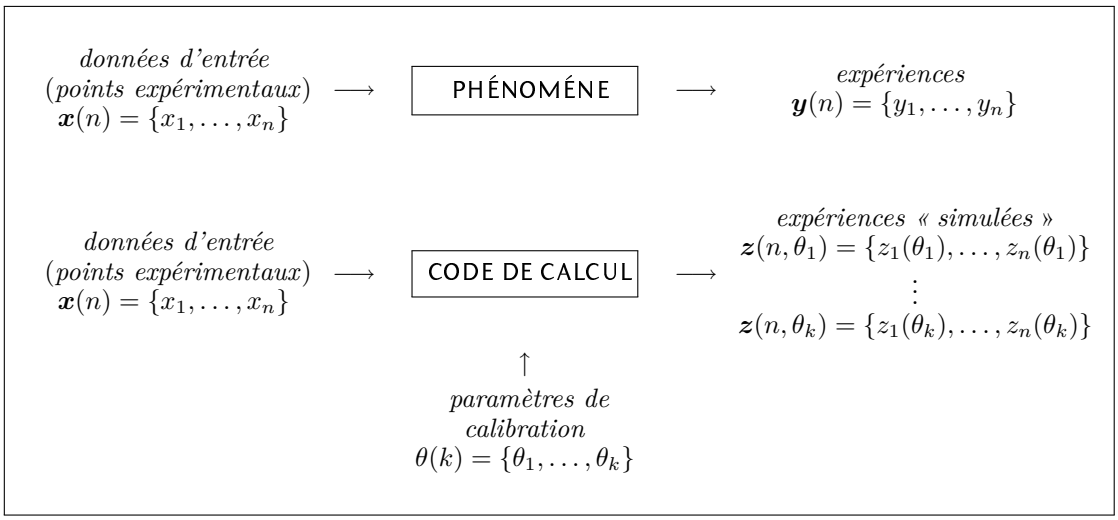
√ Nous notons $\boldsymbol{z}(n) = \{z_1 = f(x_1, \theta), \ldots, z_n = f(x_n, \theta)\}$, *l'ensemble des réponses du modèle*, en fonction des $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ de la BDDE.

Ce sont les résultats du code en chaque point de la BDDE comme mentionné ci-dessus. Selon les configurations possibles, le vecteur de paramètres $\theta \in \Theta$ peut être éventuellement choisi librement parmi plusieurs valeurs possibles $\theta_j \in \Theta$, $j = 1, \ldots, k$. Dans ce cas, nous disons qu'il y a un *contrôle du paramètre*. Nous parlerons de la base de données des paramètres pour désigner $\theta(k) = \{\theta_1, \ldots, \theta_k\}$. Lorsque nous avons un contrôle du paramètre, nous disposons de plusieurs ensembles de réponses

$$\boldsymbol{z}(n, \theta_j) = \{z_1(\theta_j) = f(x_1, \theta_j), \ldots, z_n(\theta_j) = f(x_n, \theta_j)\},$$

où $\theta_j \in \Theta$, $j = 1, \ldots, k$, est connu. Dans certaines situations, les valeurs de $\theta \in \Theta$ seront considérées comme fixées à l'avance. Cela signifie que les ensembles $\boldsymbol{z}(n, \theta_j)$, $j = 1, \ldots, k$, de réponses du code ont déjà été obtenues en différents $\theta_j \in \Theta$, et qu'il n'est pas possible d'obtenir de nouvelles réponses du code en d'autres valeurs du paramètre.

L'objectif fondamental de la calibration est d'ajuster le paramètre vectoriel θ de manière à ce que $\boldsymbol{z}(n, \theta) = \{f(x_1, \theta), \ldots, f(x_n, \theta)\}$ s'approche au mieux (selon des critères d'ajustement précisés ultérieurement) de $\boldsymbol{y}(n) = \{y(x_1), \ldots, y(x_n)\}$.



MÉTHODOLOGIE

Pour résoudre le problème de calibration de façon *générique*, c'est-à-dire *quel que soit le contexte d'application de l'étude considérée*, nous proposons la méthodologie suivante :

***Etape* 1** Collecte d'informations ;

***Etape* 2** Analyse de la base de données d'entrée ;

***Etape* 3** Application de la démarche de calibration ;

***Etape* 4** Validation du (ou des) paramètres estimé(s).

• **L'étape 1** consiste à prendre connaissance de l'ensemble des données à notre disposition. La méthode de résolution du problème de calibration dépendra de ces informations. Les données de la BDDE : $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$, $x_i \in \mathcal{X}$, et des réponses expérimentales $\boldsymbol{y}(n) = \{y_1, \ldots, y_n\}$, $y_i \in \mathcal{Y}$, pour $i = 1, \ldots, n$, seront toujours supposées disponibles. Plusieurs cas sont possibles : la fonction de code est connue (cas 1), partiellement connue ou inconnue (cas 2). Suivant les cas considérés, la calibration peut ou non nécessiter des réponses de code : $\boldsymbol{z}(n, \theta_j) = \{z_1 = f(x_1, \theta_j) \ldots, z_n = f(x_n, \theta_j)\}$ en un ensemble de paramètres $\theta(k) = \{\theta_1, \ldots, \theta_k\}$. Plus précisément, nous distinguerons les cas suivants :

Cas 1 **« Connaissance de la fonction de code »**

Nous dirons qu'il y a connaissance de la fonction de code f lorsqu'elle est connue de façon analytique et que son expression est « simple ». Dans ce contexte, les données de la BDDE : $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$, $x_i \in \mathcal{X}$, et des réponses expérimentales $\boldsymbol{y}(n) = \{y_1, \ldots, y_n\}$, $y_i \in \mathcal{Y}$, pour $i = 1, \ldots, n$, seront suffisantes pour effectuer la calibration.

Cas 2 **« Absence de connaissance de la fonction de code »**

Nous dirons qu'il y a absence de connaissance de la fonction de code f lorsque nous considérons qu'elle n'est pas connue de façon analytique (expression analytique inconnue ou trop « complexe » pour l'application des méthodes classiques de régression). Pour pouvoir estimer le paramètre de calibration θ, nous aurons alors besoin de réponses de code : $\boldsymbol{z}(n, \theta_j) = \{z_1 = f(x_1, \theta_j) \ldots, z_n = f(x_n, \theta_j)\}$ en un ensemble de paramètres $\theta(k) = \{\theta_1, \ldots, \theta_k\}$. Les techniques d'estimation de θ seront différentes lorsqu'il y a :

a) possibilité d'obtenir de nouvelles réponses du code en des $\theta_{j'} \in \Theta$ supplémentaires, on dit alors qu'il y a contrôle du paramètre θ ;

b) impossibilité d'obtenir de nouvelles réponses du code.

Les méthodes de calibration associées à ces différents cas seront discutées lors de la présentation de l'étape 3.

• **L'étape 2** de la méthodologie, à savoir l'*analyse de la base de données*, consiste à vérifier que les données vont permettre d'obtenir une estimation « correcte » du paramètre de calibration (par exemple, avec une bonne précision, intervalle de confiance réduit, et/ou robuste, peu sensible aux variations des entrées du code). Cette analyse concernera donc la BDDE $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$, avec $x_i \in \mathcal{X}$, et éventuellement $\theta(k) = \{\theta_1, \ldots, \theta_k\}$, avec $\theta_j \in \Theta$, lorsque cet ensemble de paramètres existe. Selon la méthode d'estimation du paramètre retenue, les critères de qualité des données peuvent être différents.

Dans le cas où il y a *connaissance de la fonction* de code (cas 1 de l'étape 1), il nous faudra analyser uniquement la BDDE $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$. On a recours aux critères bien connus de la théorie des *plans d'expériences*. Les critères alors considérés font intervenir l'expression analytique de la fonction f et permettent le plus souvent de réduire l'intervalle de confiance (ou une région de confiance) du paramètre estimé. Ces critères ayant fait l'objet de nombreux travaux au cours des dernières décennies, ils ne seront pas discutés ici en détail. Nous renvoyons entre autres à Chernoff (1953), Kiefer (1959, 1961, 1974), Fedorov (1972, 1980), Wynn (1970), Pronzato (1986), ainsi qu'à l'ouvrage de Droesbeke *et al.* (1997) pour des exposés généraux et des applications. Lorsque la fonction de code f est non linéaire par rapport aux paramètres, les techniques correspondantes sont parfois délicates à mettre en oeuvre, voir Gauchi et Pázman (2006). Aussi, nous traiterons parfois ce cas de la même façon que lorsqu'il y a absence d'informations sur f.

Lorsque la fonction de code est *inconnue* ou *considérée comme telle* (dans le cas 2 de l'étape 1), l'utilisation des critères de plan d'expériences est délicate. Cependant, certaines méthodes de calibration impliquent une substitution de la fonction de code par une approximation de celle-ci appelée *métamodèle* ou *surface de réponse*. Des techniques de *régression linéaire* ou *non linéaire* peuvent être employées. Le cadre théorique alors posé permet l'utilisation des critères de plan d'expériences. Toutefois il s'agit d'un cas particulier, car pour de nombreux *métamodèles*, il n'existe pas de tels critères (réseau de neurones, krigeage, méthode des fonctions orthogonales, par exemple). De façon à ce que la méthodologie proposée puisse être appliquée dans un contexte général, nous chercherons à vérifier que les données ont une « répartition uniforme » au sens de critères « déterministes » ou « probabilistes » (cf. chapitres I, II, II).

Les premiers sont liés à la notion de « remplissage de l'espace » ou« space filling » (cf. chapitre I). L'objectif est d'évaluer la répartition « uniforme » des données, de vérifier qu'elles « recouvrent » (« remplissent ») au mieux l'espace dans lequel elles sont définies. Il s'agit d'une approche heuristique puisqu'il n'est pas rigoureusement démontré que des points de « répartition uniforme » (au sens précédemment explicité) permettent une « meilleure » estimation de paramètres, et ceci quelle que soit la méthode d'estimation. Certains des critères étudiés, issus de la notion de *discrépance*, peuvent cependant trouver une justification théorique. En effet, ils interviennent dans *l'inégalité de Koksma-Hlwaka* qui permet de majorer l'erreur d'estimation de l'intégrale d'une fonction par sa

moyenne. Or, certains types d'estimateurs font intervenir des moyennes, approximations d'intégrale (par exemple, ceux de la *méthode des moments*, procédé qui bien que rendu obsolète par les techniques du maximum de vraisemblance demeure néanmoins utilisé par nombre d'expérimentateurs). Ainsi, l'inégalité de Koksma-Hlwaka (cf. chapitre II) montre que la diminution de critères de discrépance d'un ensemble de points en entrée du modèle réduira, en conséquence, les erreurs d'estimations. Plus particulièrement, nous détaillerons et développerons le lien qui existe entre ces critères particuliers et une méthode d'estimation d'un paramètre fonctionnel (à l'aide des travaux de Hickernell (1999) et Rafajlowicz et Schwabe (2005)). Nous en déduirons qu'un critère de *discrépance* faible permet d'obtenir une « meilleure » estimation du paramètre fonctionnel (au sens de la *mean square error*, $\mathbb{MSE}$ et de l'*integrated mean square error*, $\mathbb{IMSE}$, critères définis dans le chapitre II).

Les critères de répartition uniforme au sens « probabiliste » sont liés à la notion de *tests statistiques* (cf. chapitre III). Par cette approche, les données sont considérées comme des variables aléatoires. Il s'agit alors d'effectuer des *tests de l'hypothèse d'indépendance et d'une distribution uniforme* de ces variables. De nombreux tests existent dans ce cadre. Leur objectif est de tester la qualité de générateurs de nombres aléatoires (essentiellement en une dimension) utilisés pour les méthodes de Monte Carlo (cf. chapitre III). Cela constitue aussi une approche heuristique dans le sens où la considération de variables aléatoires uniformes ne permet pas forcément une « meilleure » estimation de paramètres (dans un sens qui serait à définir), ceci quelle que soit la méthode d'estimation utilisée. Remarquons cependant que si les données ne correspondent pas à des variables aléatoires uniformes et qu'elles sont considérées comme telles, les estimateurs des moments seront bien entendu biaisés.

• **L'étape 3**, appelée *application de la démarche de calibration*, est l'estimation du paramètre recherché. Comme précisé à l'étape 1, nous distinguons les deux cas :

Cas 1 « Connaissance de la fonction de code » :

Les techniques bien connues de régressions linéaire et non linéaire permettront d'obtenir une estimation du paramètre θ (nous nous référerons entre autres à Antoniadis *et al.* (1992) et Walter et Pronzato (1997)).

Cas 2 « Absence de connaissance de la fonction de code » :

a) Possibilité d'obtenir de nouvelles réponses du code en des $\theta_{j'} \in \Theta$ supplémentaires,
on dit alors qu'il y a contrôle du paramètre θ, différents appels de la fonction de code f sont possibles pour l'estimation du paramètre de calibration, les techniques d'optimisation multiobjectif (voir Collette et Siarry (2002)), ou la méthode GLUE (*Global Likelihood Uncertainty Estimation*, voir Beven et Binley (1992)) peuvent alors être appliquées ;

b) Impossibilité d'obtenir de nouvelles réponses du code,
lorsqu'une étude des résultats pré-existants de la fonction de code f permet de savoir que celle-ci est linéaire par rapport au paramètre de calibration $\theta \in \Theta$, nous utiliserons les techniques dites de *calibration multivariée* ;
lorsque f est « complexe » (essentiellement non linéaire par rapport à $\theta \in \Theta$), il faudra construire un *métamodèle* (ou surface de réponse) permettant d'obtenir une estimation de la fonction de code (voir Santner *et al.* (2003), par exemple) ; l'estimation du paramètre de calibration se fera alors en substituant la fonction de code par le *métamodèle* (voir Kennedy et O'Hagan (2001a)).

• **L'étape 4** est la validation du (ou des) paramètres estimé(s). La (ou les) méthode(s) de calibration adaptée(s) au phénomène étudié sera (seront) réalisée(s) à l'aide de données convenablement sélectionnées parmi la BDDE $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ (en utilisant une technique définie au chapitre I, par exemple). Les données non-sélectionnées serviront à la validation. Ceci implique qu'il est possible d'effectuer une expérience simulée (appel à la fonction de code) en de nouveaux paramètres (notamment le paramètre estimé).

ORGANISATION DE LA THESE

L'étape 2, dans le contexte où il y a « Absence de connaissance de la fonction de code », sera principalement développée et fera l'objet des chapitres I, II, III. L'étape 3 sera détaillée en annexe

De nombreux critères d'uniformité, souvent utilisés dans le contexte de l'intégration numérique, seront présentés au chapitre I (inspirés, entre autres des travaux de Gunzburger et Burkardt (2004) et Hickernell (1998)). L'objectif est de vérifier qu'un espace est « bien recouvert », « bien rempli » par un ensemble de points (au sens défini chapitre I). Ainsi, la notion d'uniformité est différente de celle de la théorie statistique. Nous ne considérerons pas les données comme des variables aléatoires et nous parlerons d'approche « déterministe ». Nous nous focaliserons sur la notion de *discrépance* (essentiellement celle définie par Hickernell (1998)). Il s'agit de critères permettant d'effectuer des comparaisons entre le nombre de points compris dans un pavé (produit d'intervalles) et le volume (ou mesure de Lebesgue) de ce pavé. Ils sont donc, par leur définition même, tout à fait adaptés à l'objectif fixé. Nous les utiliserons pour développer des techniques d'« extraction » et de « spécification » de points de façon à constituer un ensemble qui recouvre uniformément l'espace dans lequel ils sont définis.

Le chapitre II est l'étude de liens entre la discrépance et une méthode de régression non paramètrique. La discrépance intervient dans l'inégalité de Koksma-Hlwaka (voir Hlwaka (1961) et Hickernell (1998)) qui fournit une borne de l'erreur de l'estimation d'une intégrale. De façon générale, pour une fonction f de carré intégrable, pour une suite de points $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ dans un espace $\mathcal{X}$, l'inégalité de Koksma-Hlwaka

généralisée peut s'écrire comme suit (cf. Niederreiter et Spanier (Eds) (1998)) :

$$|I(f) - \hat{I}_{\boldsymbol{x}(n)}(f)| \leq V(f)\mathrm{D}(\boldsymbol{x}(n)), \tag{1}$$

où :

i) $I(f) = \int_{\mathcal{X}} f(x)dx$,

ii) $\hat{I}_{\boldsymbol{x}(n)}(f) = \frac{1}{n}\sum_{x_i \in \boldsymbol{x}(n)} f(x_i)$ avec $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$,

iii) $V(f)$ est une *variation* de f,

iv) et $\mathrm{D}(\boldsymbol{x}(n))$ est un terme qui dépend des points $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ et correspond à la notion de *discrépance*.

Nous renvoyons au chapitre II pour des définitions précises de ces différentes quantités. A l'aide de cette inégalité, et en considérant les travaux de Hickernell (1999) et de Rafajlowicz et Schwabe (2005), nous proposerons une majoration de critères de qualité d'estimation d'un paramètre fonctionnel obtenu par la méthode des fonctions orthogonales (introduite par Cencov (1962), voir aussi Földes et Révész (1974), Sansone (1977), Devroye et Györfi (1985), Bosq et Lecoutre (1987), Hardle (1989), par exemple). Ce chapitre souligne l'importance de la qualité d'une base de données (au sens de sa répartition uniforme) pour l'estimation de paramètres et justifie de façon théorique les techniques de sélection et de spécification de points présentées au chapitre I. Bien qu'il s'agisse d'une méthode d'estimation particulière, nous pouvons raisonnablement penser qu'en l'absence d'information sur la fonction f, un ensemble de points ayant une discrépance faible permettra, en général, d'obtenir une estimation de paramètres « convenable ». Les estimateurs de paramètres, mais aussi les estimateurs de leurs espérances et de leurs variances, font parfois intervenir des moyennes, approximations d'intégrale (par exemple, ceux obtenus par la méthode de Monte Carlo). Ainsi, l'inégalité (1) montre que la diminution de la discrépance d'un ensemble de points réduira, en conséquence, les erreurs d'estimation.

Dans le chapitre III, les données seront considérées comme aléatoires. Nous effectuerons donc des *tests statistiques d'uniformité multidimensionnelle*. Plusieurs tests seront proposés, issus de différentes théories, « *sparse-serial tests* », « *scan statistics* ». Nous nous référerons respectivement à L'Ecuyer *et al.* (2002), Glaz *et al.* (2001), et ainsi qu'à leurs propres références. Pour la plupart de ces tests, une partition en cellules disjointes de l'espace sera effectuée, et nous définirons des statistiques à l'aide du nombre de points compris dans ces cellules.

Chapitre 1

Critères déterministes

1.1 Contexte

Nous ferons ici référence au vocabulaire, aux notations et à la méthodologie définis en introduction.

Le présent chapitre étudie principalement l'étape 2 de la méthodologie dans un contexte où la structure de la fonction du code (propriété de continuité, de linéarité, etc.), est inconnue. Nous allons donc nous intéresser à l'étude de la BDDE $\boldsymbol{x}(n)$. En particulier, cette étude sera menée dans le cas où il existe différents ensembles $\boldsymbol{z}(n, \theta_j)$ de réponses du code en différentes valeurs $\theta_j \in \Theta$, $j = 1, \ldots, k$, de la base de données des paramètres : $\boldsymbol{\theta}(k) = \{\theta_1, \ldots, \theta_k\}$. Par la suite, nous parlerons de l'analyse de la BDDE $\boldsymbol{x}(n)$, mais toutes les méthodes que nous allons définir doivent aussi s'appliquer à $\boldsymbol{\theta}(k)$ lorsque cette base existe.

Ne connaissant pas la structure de modèle que nous prenons en compte pour la calibration, notre analyse consiste à vérifier que les points de $\boldsymbol{x}(n)$ ont une répartition uniforme dans leur domaine de variation. Pour ce faire, nous allons définir, d'un point de vue formel, différents critères. Comme le titre de ce chapitre l'indique, ceux-ci sont essentiellement dictés par des considérations déterministes. Il s'agit d'une approche différente de celle des tests statistiques d'uniformité faisant l'hypothèse d'une répartition aléatoire des points dans l'espace. Par exemple, lorsque nous vérifions que l'espacement entre les points d'une BDDE est *régulier*, ce qui est par exemple le cas pour un *réseau*, cette propriété relève a priori d'une autre modélisation que celle qui consiste à étudier des répartitions aléatoires basées sur des observations indépendantes. La répartition « uniforme » des points n'est donc pas à comprendre ici au sens *probabiliste* (i.e., des points dont la distribution correspond à une loi uniforme, ou autre, sur le domaine $\mathcal{X}$). [1]

Ce chapitre est composé de trois parties. Tout d'abord, nous définirons des critères

[1] Les tests statistiques d'uniformité feront l'objet du chapitre III.

permettant de vérifier que les points sont régulièrement espacés, et des critères permettant de vérifier qu'il n'existe pas de « trous » dans la BDDE. Les premiers critères feront intervenir les distances entre les points de la BDDE (afin de vérifier qu'il y a un espacement convenable entre ceux-ci), les seconds prendront en compte les distances entre les points de la BDDE et les points de l'espace $\mathcal{X}$, il s'agira donc de considérations basées sur des distances convenablement choisies. Nous présenterons ensuite des critères permettant d'apprécier le *recouvrement*, le « remplissage », de l'espace, en particulier à l'aide d'objets géométriques spécifiques de $\mathcal{X}$ (hypercubes, cellules de Voronoi).

Dans une seconde partie nous étudierons la notion de *discrépance*. Il s'agit de critères permettant d'effectuer des comparaisons entre le nombre de points compris dans un pavé (produit d'intervalles) et le volume (ou mesure de Lebesgue) de ce pavé. Nous détaillerons précisément ces notions qui se trouvent, par leur définition même, constituer des critères tout à fait adaptés pour l'évaluation du *recouvrement uniforme* d'une BDDE $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ de n points dans un espace $\mathcal{X} = [0,1]^d$.

Dans la troisième partie de ce chapitre nous expliquerons comment utiliser et interpréter ces différents critères. Nous proposerons enfin une méthodologie d'étude de la qualité d'une base de données, que nous illustrerons à l'aide d'exemples.

1.2 Les critères

1.2.1 Notations

Commençons par introduire quelques notations générales que nous utiliserons pour l'analyse de la BDDE :

(1) $\mathcal{X} = \bar{\mathcal{O}}$, où $\mathcal{O}$ est un ouvert borné de $\mathbb{R}^d$, désigne l'espace des valeurs possibles des éléments de la BDDE, $x_i \in \mathcal{X}$, pour $i = 1, \ldots, n$. L'espace de la BDDE $\mathcal{X}$ est donc une partie compacte de $\mathbb{R}^d$.

(2) $\boldsymbol{x}(n)$ la suite de points de la BDDE : $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ avec $x_i \in \mathcal{X}$,

(3) $\rho_p(.,.)$ la distance ℓ^p sur $\mathcal{X}$ définie par :

$$\rho_p(w,v) := \left[\sum_{j=1}^{d} |w_j - v_j|^p\right]^{1/p} \quad \text{avec } p \geq 1 \text{ et } w, v \in \mathcal{X}, \tag{1.1}$$

$$\rho_\infty(w,v) := \max_{j=1,\ldots,d} |w_j - v_j| \quad w, v \in \mathcal{X}. \tag{1.2}$$

Pour $p = 2$, ρ_2 est la distance euclidienne classique dans $\mathbb{R}^d$.

(4) $\gamma_i := \min_{\ell \neq i} \rho_p(x_i, x_\ell)$ désigne la distance minimale entre points x_i, x_ℓ, $i \neq \ell$.

(5) $\bar{\gamma}$ désigne la moyenne des γ_i :

$$\bar{\gamma} := \frac{1}{n}\sum_{i=1}^{n}\gamma_i.$$

1.2.2 Critères associés aux distances

1.2.2.1 Distances entre points de la BDDE

Les critères que nous allons définir ici permettent de vérifier que les points ne sont pas trop « proches » les uns des autres et régulièrement espacés. Ils sont essentiellement inspirés par Gunzburger et Burkardt (2004). Notons que la notion de proximité dépend bien entendu des caractéristiques du problème étudié. Par exemple, une distance entre deux points pourra être considérée comme « faible » lors de l'étude d'un phénomène linéaire, et « importante » si le phénomène est fortement non linéaire. Pour les applications il est donc nécessaire d'avoir recours à l'avis d'un expert du domaine étudié afin de pouvoir apprécier les distances les plus appropriées au phénomène et à l'étude.

- **Distances minimales**

Pour apprécier la proximité mutuelle des points de la BDDE, nous pouvons, dans un premier temps, faire usage de la distance minimale entre deux points de la base de données d'entrée. Celle-ci est définie par :

$$\boxed{\mathrm{dmin}_p(\boldsymbol{x}(n)) := \min_{x_i \neq x_j \in \boldsymbol{x}(n)} \rho_p(x_i, x_j).} \tag{1.3}$$

S'il existe des points trop « proches » les uns des autres, vis-à-vis du problème étudié, la quantité dmin_p sera faible. Lorsque nous souhaitons que tous les points soient au moins séparés les uns des autres d'une distance d_v, nous pouvons définir le critère :

$$\boxed{\min_{x_i \neq x_j \in \boldsymbol{x}(n)} \rho_p(x_i, x_j) > d_v.} \tag{1.4}$$

Dans le cas où la répartition des points $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ composant la BDDE peut être planifiée (conformément à un procédé de planification), nous définissons le plan de répartition suivant :

Définition 1.2.1
Le plan $\boldsymbol{x}^*(n) = \{x_1^*, \ldots, x_n^*\}$, permettant d'assurer une distance minimale satisfaisante entre chaque point, au sens de l'égalité :

$$\min_{x_i \neq x_j \in \boldsymbol{x}^*(n)} \rho_p(x_i, x_j) = \max_{\boldsymbol{x}(n) \in \mathcal{X}^n} \min_{x_i \neq x_j \in \boldsymbol{x}(n)} \rho_p(x_i, x_j), \tag{1.5}$$

est dit maximin.

Partant du constat que, pour un maillage où les points sont régulièrement espacés les uns des autres, nous avons $\gamma_1 = \ldots = \gamma_n$ où γ_i désigne la distance entre x_i et son plus proche voisin dans la BDDE, nous pouvons aussi définir le critère de qualité suivant :

$$\boxed{\gamma(\boldsymbol{x}(n)) := \frac{\max_{i=1,\ldots,n} \gamma_i}{\min_{i=1,\ldots,n} \gamma_i}.} \tag{1.6}$$

Comme dans un cas de parfaite régularité, ce critère vaut 1, une valeur voisine de 1 exprimera un critère de qualité de la régularité de la répartition des points. Remarquons que le critère basé sur $\gamma(\boldsymbol{x}(n))$ est plus global que dmin_p (voir (1.3)).

• **Distances moyennes**

Une autre approche pour apprécier la régularité des espacements entre les points de la BDDE est de considérer des distances « moyennes ».

Après avoir éventuellement normalisé l'espace $\mathcal{X} \subset \mathbb{R}^d$ associé à notre BDDE de façon à avoir $\mathcal{X} = [0,1]^d$, nous considérons un critère de distance moyenne correspondant à l'inverse de la moyenne harmonique des distances (nous avons $n(n-1)/2$ distances) entre les points de la BDDE, rapportées au coefficient $d^{1/p}$. Cette expression est définie par

$$m_p(\boldsymbol{x}(n)) := \frac{2}{n(n-1)} \sum_{x_i \neq x_j \in \boldsymbol{x}(n)} \left[\frac{d^{1/p}}{\rho_p(x_i, x_j)} \right]. \tag{1.7}$$

Plus généralement, voir Santner *et al.* (2003), nous utilisons le critère suivant, portant sur la moyenne des distances :

$$\boxed{m_{p,\lambda}(\boldsymbol{x}(n)) := \left(\frac{2}{n(n-1)} \sum_{x_i \neq x_j \in \boldsymbol{x}(n)} \left[\frac{d^{1/p}}{\rho_p(x_i, x_j)} \right]^{\lambda} \right)^{1/\lambda}.} \tag{1.8}$$

où $\lambda \geq 1$ désigne un paramètre réel convenablement choisi.

Notons que nous avons les inégalités :

$$0 < \rho_p(x_1, x_2) \leq d^{1/p}. \tag{1.9}$$

Ici $d^{1/p}$ n'est rien d'autre que la valeur de la norme de la « diagonale » de l'hypercube unité. En effet, pour cette diagonale, $x_i = (0, \ldots, 0)^t$, et $x_j = (1, \ldots, 1)^t$, de sorte que $\rho_p(x_i, x_j) = d^{1/p}$.

L'inégalité (1.9) implique encore $m_{p,\lambda}(\boldsymbol{x}(n)) \geq 1$. Plus cette quantité sera faible, moins il y aura de redondance, i.e., de « groupement(s) » de points trop proches les uns des autres dans $\boldsymbol{x}(n)$.

Lorsque nous imposons à la BDDE $\boldsymbol{x}(n)$ de conserver une distance minimale d_v entre deux points, nous devons aussi vérifier l'inégalité :

$$m_{p,\lambda}(\boldsymbol{x}(n)) > d^{1/p}/d_v. \tag{1.10}$$

Ce critère n'assure pas cependant une absence « totale » de *redondance*, certains points pouvant être « proches » les uns des autres, mais assure que la distance minimale entre points distincts de la BDDE est « en moyenne » supérieure à d_v, et que nous n'avons donc pas de groupes trop importants de points proches les uns des autres.

Dans le cas où la représentation des points $\boldsymbol{x}(n)$ peut être *planifiée* en $\boldsymbol{x}^*(n)$, nous adoptons la nomenclature suivante faisant référence à l'égalité (1.8).

Définition 1.2.2
Les plans $x^*(n)$ sont dits de critère $m_{p,\lambda}$ optimaux si :

$$m_{p,\lambda}(\boldsymbol{x}^*(n)) = \min_{\boldsymbol{x}(n)\in\mathcal{X}^n}(m_{p,\lambda}(\boldsymbol{x}^*(n))). \tag{1.11}$$

Notons que lorsque $\lambda \to \infty$ dans (1.8) nous avons :

$$\begin{aligned} m_{p,\infty}(\boldsymbol{x}(n)) &= \lim_{\lambda\to\infty} m_{p,\lambda}(\boldsymbol{x}(n)) \\ &= \max_{x_i\neq x_j\in\boldsymbol{x}(n)} \frac{d^{1/p}}{\rho_p(x_i,x_j)}. \end{aligned}$$

Par conséquent, un plan *optimal* pour le critère $m_{p,\infty}(\boldsymbol{x}(n))$ est aussi un plan *maximin* (voir 1.2.1).

Avec les notations (4) et (5) du paragraphe (1.2.1), nous définissons un autre critère utilisant une moyenne de distances donné par :

$$\boxed{\Lambda(\boldsymbol{x}(n)) := \frac{1}{\bar{\gamma}}\left(\frac{1}{n}\sum_{i=1}^{n}(\gamma_i-\bar{\gamma})^2\right)^{1/2}.} \tag{1.12}$$

Notons que ce critère demeure défini pour un espace $\mathcal{X}$ non restreint à $[0,1]^d$. Il s'agit, pour ce critère, de vérifier que la variabilité (l'écart-type) des distances des points les plus proches (les γ_i) n'est pas trop importante. Lorsque nous considérons un maillage régulier de l'espace $\mathcal{X}$, ce critère est nul, puisque $\gamma_1 = \ldots = \gamma_n$. Ainsi, plus ce critère sera faible, plus la répartition des points pourra être considérée comme régulière.

1.2.2.2 Distances entre points de la BDDE et points de l'espace

Nous allons introduire ici des critères qui permettent de vérifier qu'il n'existe pas de « trous » dans la BDDE, et, par conséquent, de vérifier que l'ensemble du domaine $\mathcal{X}$ est recouvert de façon « acceptable » (voir Feuillard *et al.* (2005)). Nous ne considérons donc plus uniquement des distances entre points de la BDDE, mais des distances entre points de la BDDE et des points convenables de l'espace $\mathcal{X}$.

- **Dispersion**

Un premier critère permettant de quantifier les « trous » de la BDDE est celui de la *dispersion* au sens de la définition suivante :

Définition 1.2.3
Nous définissons la dispersion de la BDDE $\boldsymbol{x}(n) \in \mathcal{X}^n$ *par :*

$$d_p(\boldsymbol{x}(n)) := \sup_{w \in \mathcal{X}} \left\{ \min_{x_i \in \boldsymbol{x}(n)} \rho_p(w, x_i) \right\}. \tag{1.13}$$

Dans le cas particulier $p = 2$ et $p = \infty$, nous obtenons :

$$\begin{aligned} d_2(\boldsymbol{x}(n)) &:= \sup_{w \in \mathcal{X}} \left\{ \min_{x_i \in \boldsymbol{x}(n)} \|w - x_i\| \right\}, \\ d_\infty(\boldsymbol{x}(n)) &:= \sup_{w \in \mathcal{X}} \left(\min_{1 \le i \le n} \left\{ \max_{j=1,\dots,d} |w_j - x_j| \right\} \right). \end{aligned}$$

Intuitivement, la *dispersion* au sens de la définition (1.2.3) s'interprète comme le rayon de la plus grande boule ne contenant aucun point de la BDDE dans l'espace $\mathcal{X}$. Cette quantité peut aussi être vue comme l'infimum de tous les rayons r tels que les boules $B(x_1, r), \dots, B(x_n, r)$ recouvrent $\mathcal{X}$ (rappelons que lorsque $\mathcal{X}$ est compact, il existe un nombre fini de boules de rayon r recouvrant $\mathcal{X}$). Par conséquent, lorsque la *dispersion* (au sens de la définition (1.2.3)) est élevée, la suite de points comporte des « trous » dans le domaine $\mathcal{X}$, et une dispersion faible assure une bonne répartition des points dans $\mathcal{X}$, un *recouvrement* de l'espace sans « trou ».

Précisons quelques propriétés intéressantes. Nous avons les inégalités

$$\frac{1}{2\lfloor n^{1/d} \rfloor} \le d_\infty(\boldsymbol{x}(n)) \le d_2(\boldsymbol{x}(n)) \quad \le d^{1/2} d_\infty(\boldsymbol{x}(n)), \tag{1.14}$$

où $\lfloor n^{1/d} \rfloor$ désigne la partie entière de $n^{1/d}$. Cette dernière inégalité est notamment atteinte pour les grilles de Sukharev (voir l'illustration graphique (1.1)). Pour davantage de détails sur cette notion, nous nous référerons à Sukharev (1971), Niederreiter (1988), Niederreiter (1992), Niederreiter et Wills (1975).

Une façon de calculer la dispersion est de considérer une dispersion « relative » à une suite dont nous savons qu'elle *recouvre* bien tout l'espace (nous pourrons considérer des

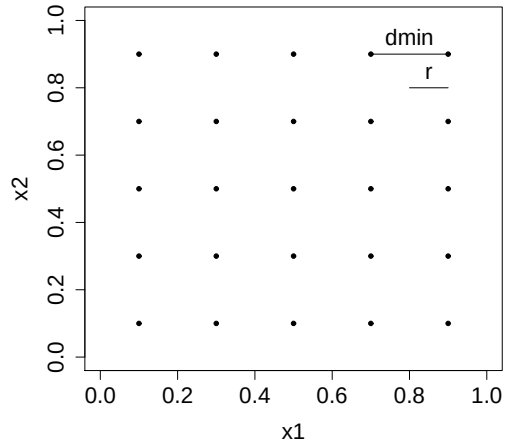


FIG. 1.1 – Grille de Sukharev (25 points)

suites à discrépances faibles dans $[0,1]^d$, définies au paragraphe (1.3)). Nous supposons disposer d'une *suite auxiliaire* BDDREF $\boldsymbol{x}_f(N) = \{x_{f,1}, \ldots, x_{f,N}\}$ de points de $\mathcal{X}$ dont nous considérons les propriétés d'uniformité « acceptables ». Ici, N correspond à un nombre important (en pratique, largement supérieur à n) de points (de façon à bien recouvrir tout l'espace). Nous calculons :

$$\mathrm{Disp}(\boldsymbol{x}(n), \boldsymbol{x}_f(N)) \quad := \quad \max_{x_{f_i} \in \boldsymbol{x}_f(N)} \left\{ \min_{x_j \in \boldsymbol{x}(n)} \rho_p(x_{f_i}, x_j) \right\}. \tag{1.15}$$

Ceci est illustré par le graphique (1.2).

Nous faisons donc, en utilisant (1.15), une approximation de la dispersion au sens de l'égalité (1.13). Ne pouvant calculer le supremum relativement à $w \in \mathcal{X}$ dans la définition (1.2.3) sur tout le domaine $\mathcal{X}$, nous prenons, dans (1.15), le maximum relativement aux $x_{f_i} \in \boldsymbol{x}_f(n)$, parmi un ensemble $\boldsymbol{x}_f(n)$ de points qui *recouvre* $\mathcal{X}$ de façon jugée satisfaisante. Nous pouvons aussi interpréter le critère (1.15) comme une comparaison entre les points de la BDDE et des points d'une BDDREF ayant une bonne répartition uniforme, c'est pourquoi nous avons choisi l'appellation dispersion « relative » pour (1.15). Précisons que cette notion que nous introduisons ici n'a pas été rencontrée dans la littérature, et est à notre connaissance, nouvelle.

Signalons dès à présent que certaines des suites $\boldsymbol{x}_f$ que nous choisirons comme BDDREF seront les *suites à discrépance faible* (au sens de la définition (1.3.7) du paragraphe (1.3)). Celles-ci ont aussi l'avantage d'être à *dispersion faible* (au sens de l'égalité (1.30) du paragraphe (1.3)). Ainsi, si la suite de points de la BDDE est « proche » (dispersion

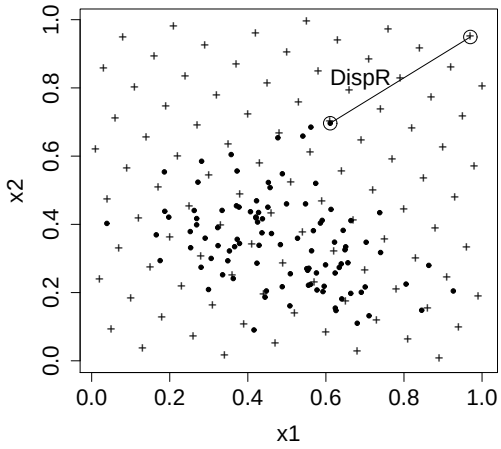


FIG. 1.2 – Approximation de la dispersion

relative faible) d'une *suite à discrépance faible*, nous pouvons penser que la discrépance, et par conséquent la dispersion, sera faible. Nous calculerons donc pour la BDDE $\boldsymbol{x}(n)$ différentes dispersions « relatives » à l'aide de différentes BDDREF, *suites à discrépance faible*.

Au lieu de prendre comme BDDREF une suite « déterministe » (comme le sont les suites à discrépance faible), nous pouvons aussi considérer des suites de variables aléatoires indépendantes et de loi uniforme dans $\mathcal{X} = [0,1]^d$. Nous calculerons le critère (1.15) pour ces différentes suites, et garderons la plus grande valeur comme approximation de la dispersion. En pratique, le nombre de points de ces suites peut ne pas être trop élevé (de l'ordre du nombre de points de la BDDE étudiée n, par exemple) de façon à effectuer de nombreux calculs de la dispersion « relative ».

Lorsque les différentes approximations ont le même ordre de grandeur, nous pourrons considérer une approximation correcte de la dispersion théorique.

Pour apprécier la valeur de l'approximation obtenue, nous pouvons la comparer avec les différents critères de distance que nous aurons précédemment calculés, comparant ainsi les distances entre points de la BDDE et le rayon de la plus grande boule vide dans l'espace $\mathcal{X}$. Il est, ici aussi, utile d'avoir recours à l'avis d'un expert du domaine étudié qui soit capable d'apprécier la pertinence des différentes distances utilisées dans les critères d'uniformité. Plus la dispersion (définition (1.2.3) et approximation par la formule (1.15)) sera faible et plus la BDDE *recouvrira* (« remplira ») l'espace, i.e. avec des « petits » domaines (boules de rayon faible) de l'espace sans point de la BDDE.

Nous pourrons aussi comparer la dispersion (obtenue par l'approximation de la formule (1.15)) de la BDDE $\boldsymbol{x}(n)$ avec celles de suites $(\boldsymbol{x}^*(n))$ planifiées ayant le même nombre de points n que celui de la BDDE et dont nous savons que leur dispersion est faible, i.e., considérer des ratios du type :

$$\text{Rdisp}(\boldsymbol{x}(n)) \quad := \quad \frac{d_p(\boldsymbol{x}(n), \mathcal{X})}{d_p(\boldsymbol{x}^*(n)), \mathcal{X}}. \tag{1.16}$$

Plus $\text{Rdisp}(\boldsymbol{x}(n))$ sera proche de 1, et plus la BDDE *recouvrira* l'espace, au sens des critères précédents.

Remarquons aussi, que lorsque nous remplaçons $\boldsymbol{x}_f(N)$ par $\boldsymbol{x}(n)$ dans le critère de dispersion « relative » (1.15), nous obtenons le critère *maximin* de la suite $\boldsymbol{x}(n)$ défini par le terme de droite de l'égalité (1.5) et rappelé ci-dessous :

$$\max_{\boldsymbol{x}(n)\in\mathcal{X}^n} \min_{x_i\neq x_j\in\boldsymbol{x}(n)} \rho_p(x_i, x_j),$$

- **Utilisation des cellules de Voronoi**

Il s'agit d'une approche comparable à la précédente. Dans ce qui suit la valeur de $1 \leq p \leq \infty$ sera fixée et ρ_p définie comme au paragraphe (1.2.1). Nous ne calculons plus ici le rayon de la plus grande boule vide dans $\mathcal{X}$, mais la distance caractérisant la plus grande cellule de Voronoi associée à la BDDE. Commençons par préciser la notion de *région de Voronoi* :

Définition 1.2.4
Soit $x \in \boldsymbol{x}(n)$. La cellule de Voronoi $V(x) = V(x, \mathcal{X})$ associée au point x dans $\mathcal{X}$ est l'ensemble des points w de l'espace $\mathcal{X}$ tels que la distance $\rho_p(x, w)$ soit inférieure à la distance $\rho_p(x', w)$ pour n'importe quel autre point x' de $\boldsymbol{x}(n)$:

$$V(x) := \{\ w \in \mathcal{X} : \rho_p(x, w) \leq \rho_p(x', w)\ \forall x' \in \boldsymbol{x}(n)\}.$$

(Pour illustrer cette définition, l'ensemble des régions de Voronoi d'un réseau plan est représenté dans la figure (1.3) ci-dessous).

La construction des cellules de Voronoi et le calcul les différents critères définis à partir de celles-ci présentent une forte complexité algorithmique. Nous utilisons à cet effet les programmes de Gunzburger et Burkardt (2004)[2].

[2] voir `http://www.csit.fsu.edu/~burkardt/pdf/ptmeas.pdf`

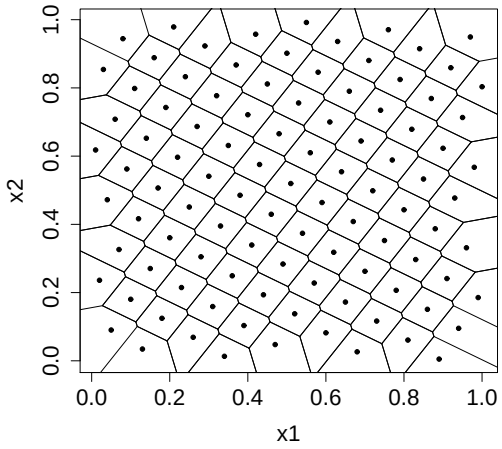


FIG. 1.3 – Régions de Voronoi d'un réseau
3

Un critère que nous pouvons définir à l'aide de cette notion est, en notant $V(x)$ la cellule de Voronoi associée à x dans $\mathcal{X}$,

$$\boxed{h(\boldsymbol{x}(n)) := \max_{i=1,\dots,n} h_i \quad \text{où} \quad h_i := \max_{w \in V(x_i)} \rho_p(x_i, w).} \tag{1.17}$$

Par définition, le critère (1.17) permet d'apprécier le maximum des distances entre un point x_i de la BDDE et un point w de la cellule de Voronoi qui lui est associée. Ce critère permet donc de vérifier qu'il n'existe pas de cellule de Voronoi trop importante. En d'autres termes, un domaine important de l'espace $\mathcal{X}$ de la BDDE peut contenir un seul point x de la BDDE. Plus la valeur de ce critère sera faible, meilleur sera le *recouvrement* de l'espace. Il est ici aussi nécessaire d'avoir recours à l'avis d'un expert pour apprécier la pertinence de ce critère (puisqu'il fait intervenir le choix d'une distance appropriée dans l'espace $\mathcal{X}$).

Nous pouvons aussi définir, à partir des $h_i = \max_{w \in V(x_i)} \rho_p(x_i, w)$, $i = 1, \dots, n$, le critère :

$$\boxed{\mu(\boldsymbol{x}(n)) := \frac{\max_{i=1,\dots,n} h_i}{\min_{i=1,\dots,n} h_i}.} \tag{1.18}$$

Pour un maillage parfaitement régulier de l'espace, $\mu = 1$. Ainsi, plus la valeur du critère (1.18) sera proche de 1, plus le *recouvrement* de l'espace de la BDDE sera satisfaisant.

Nous pouvons également définir le critère χ suivant :

$$\boxed{\chi(\boldsymbol{x}(n)) := \max_{i=1,\dots,n} \chi_i \quad \text{où} \quad \chi_i := \frac{2h_i}{\gamma_i},} \tag{1.19}$$

où $\gamma_i = \min_{\ell \neq i} \rho_p(x_i, x_\ell)$ et $h_i = \max_{w \in V(x_i)} \rho_p(x_i, w)$, pour $i = 1 \dots, n$. Ce critère permet non seulement d'apprécier la qualité de *recouvrement* des points de la BDDE $\boldsymbol{x}(n) = \{x_1, \dots, x_n\}$ de l'espace $\mathcal{X}$, mais aussi d'apprécier les propriétés liées aux espacements entre points de la base de données d'entrée $\boldsymbol{x}(n)$. Dans le cas d'une BDDE de répartition « uniforme » (i.e. régulièrement répartie et qui *recouvre* l'espace $\mathcal{X}$ de façon « acceptable »), $\chi_i(\boldsymbol{x}(n))$ est idéalement constant lorsque i varie de 1 à n. Lorsque nous nous éloignerons d'une répartition uniforme, le rapport $\chi(\boldsymbol{x}(n))$ aura tendance à augmenter. Ainsi plus cette quantité sera faible, meilleur sera le *recouvrement* de $\mathcal{X}$ par les points de $\boldsymbol{x}(n)$ et meilleure sera la régularité de leur disposition dans l'espace.

RECOMMANDATIONS :
Nous rappelons que, pour pouvoir interpréter l'utilité et la pertinence des distances entre les points de la BDDE, il est nécessaire d'avoir recours à un avis d'expert. Par conséquent, l'utilisation des différents critères définis précédemment doit se faire en collaboration avec les spécialistes du domaine étudié.

1.2.3 Critères associés aux volumes des cellules de Voronoi

Dans ce paragraphe, nous allons définir des critères faisant intervenir les volumes (au sens de la mesure de Lebesgue dans $\mathbb{R}^d$) des cellules de Voronoi. Comme les précédents critères, ceux que nous développons ici sont inspirés de Gunzburger et Burkardt (2004).

Soit $V_i = V(x_i)$, $i = 1, \dots, n$, les cellules de Voronoi associées aux points x_i, $i = 1, \dots, n$ dans $\mathcal{X}$. Nous notons $|V_i|$ leur volume. Le critère suivant peut alors être défini. Nous posons

$$\boxed{\nu(\boldsymbol{x}(n)) := \frac{\max_{i=1,\dots,n} |V_i|}{\min_{i=1,\dots,n} |V_i|}.} \tag{1.20}$$

Pour un maillage régulier de l'espace $\mathcal{X}$, nous avons une valeur idéale : $\nu(\boldsymbol{x}(n)) = 1$. Ainsi, une valeur proche de l'unité donnera une information exprimant que les volumes des cellules de Voronoi seront presque identiques, et par conséquent que la BDDE considérée sera régulièrement répartie dans l'espace $\mathcal{X}$ au sens de ce critère.

Nous pouvons encore définir d'autres critères faisant intervenir des moments convenablement choisis associés aux cellules de Voronoi. Notons

$$\bar{x_i} := \frac{1}{|V_i|} \int_{V_i} x dx,$$

le premier moment (ou barycentre, ou centre de gravité) de V_i ($\bar{x}_i$ un vecteur de $\mathbb{R}^d$), et posons

$$M_i := \frac{1}{|V_i|}\int_{V_i}(x-\bar{x}_i)(x-\bar{x}_i)'dx = \frac{1}{|V_i|}\int_{V_i} xx'dx - \bar{x}_i\bar{x}_i',$$

le moment d'ordre 2 associé au centre de gravité $\bar{x}_i$ de V_i, appelé aussi matrice de covariance ou moment d'inertie. Dans cette définition, M_i est une matrice $d \times d$ symétrique positive. Nous notons $T_i = \mathrm{tr}(M_i)$, la trace de la matrice M_i, nous posons $\bar{M}_i = T_i/d$, et nous désignons par $\Delta_i := \det(M_i - \bar{M}_i.I)$, le déterminant de la matrice de déviation de M_i relativement à une matrice diagonale (« deviatoric matrix », voir Gunzburger et Burkardt (2004)).

Pour un maillage parfaitement régulier, nous aurions idéalement :

$$T_1 = \cdots = T_n = \bar{T}, \quad \text{où} \quad \bar{T} := \sum_{i=1}^{n} T_i\,,$$
$$\text{et} \quad \Delta_1 := \cdots = \Delta_n = 0.$$

Nous définissons donc les critères :

$$\boxed{\tau(\boldsymbol{x}(n)) = \max_{i=1,\ldots,n} |T_i - \bar{T}|,} \tag{1.21}$$

$$\boxed{\Delta(\boldsymbol{x}(n)) = \max_{i=1,\ldots,n} |\Delta_i|.} \tag{1.22}$$

Ainsi, plus les critères (1.21) et (1.22) seront proches de 0, plus la répartition des points sera uniforme, i.e., plus les points recouvriront bien tout l'espace $\mathcal{X}$ au sens de ce critères.

1.2.4 Récapitulatif

Dans cette partie, nous avons défini différents critères permettant de vérifier la régularité de la répartition des points de la BDDE $\boldsymbol{x}(n)$ (« l'équirépartition »), et le bon *recouvrement* de l'espace $\mathcal{X}$ par $\boldsymbol{x}(n)$. Avant de définir une dernière notion dans la partie suivante, nous récapitulons l'ensemble des critères vus jusqu'à présent dans le Tableau (1.1). Les séparations horizontales du tableau indiquent les 3 types de critères que nous avons définis :

- les critères associés aux distances entre les points de la BDDE qui permettent d'apprécier la régularité des espacements des points.
- les critères associés aux distances entre les points de la BDDE et les points de l'espace $\mathcal{X}$, qui permettent d'apprécier le *recouvrement* (« le remplissage ») de l'espace $\mathcal{X}$.
- les critères associés à des volumes qui permettent, eux aussi, d'apprécier le *recouvrement* de l'espace $\mathcal{X}$.

Les *valeurs de référence* indiquées dans la colonne de droite représentent les valeurs des critères que nous aurions si la BDDE était régulièrement répartie. Lorsque celles-ci n'existent pas, les flèches $\downarrow$ signifient que nous souhaitons une valeur la plus faible possible ($\downarrow$ signifie « à minimiser »).

CRITERE	VALEUR DE REFERENCE
$\mathrm{dmin}_\infty(\boldsymbol{x}(n))$	$\mathrm{Disp}_\infty(\boldsymbol{x}(n))/2$
$\gamma(\boldsymbol{x}(n))$	1
$m_{2,1}(\boldsymbol{x}(n))$	$\downarrow$
$\Lambda(\boldsymbol{x}(n))$	0
$\mathrm{Disp}_\infty(\boldsymbol{x}(n), \boldsymbol{x}_{Ha}(N))$	$2.\mathrm{dmin}_\infty(\boldsymbol{x}(n))$
$\mathrm{Disp}_\infty(\boldsymbol{x}(n), \boldsymbol{x}_{Ham}(N))$	$2.\mathrm{dmin}_\infty(\boldsymbol{x}(n))$
$\mathrm{Disp}_\infty(\boldsymbol{x}(n), \boldsymbol{x}_{Fa}(N))$	$2.\mathrm{dmin}_\infty(\boldsymbol{x}(n))$
$\mathrm{Disp}_\infty(\boldsymbol{x}(n), \boldsymbol{x}_{Res}(N))$	$2.\mathrm{dmin}_\infty$
$h(\boldsymbol{x}(n))$	$\downarrow$
$\mu(\boldsymbol{x}(n))$	1
$chi(\boldsymbol{x}(n))$	$\downarrow$
$\nu(\boldsymbol{x}(n))$	1
$\tau(\boldsymbol{x}(n))$	0
$\Delta(\boldsymbol{x}(n))$	0

TAB. 1.1 – Différents critères et valeurs de référence

1.3 Discrépance

Nous allons à présent évoquer la notion de *discrépance*. Celle-ci est le plus souvent rencontrée dans le domaine de l'intégration numérique multi-dimensionnelle, car elle permet de définir des bornes pour l'erreur d'intégration. D'un point de vue général, *la discrépance est une différence entre le nombre de points d'une suite* $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ *contenus dans certains pavés* P *(produit d'intervalles) d'un espace* $\mathcal{X} = [0,1]^d$*, et les volumes (mesures de Lebesgue) de ces pavés notés* $\boldsymbol{\lambda}(P)$*, pour* $P \subset \mathcal{X}$. Comme il est possible de choisir différentes catégories de pavés (hypercubes, hyperrectangles, etc.), ainsi que d'autres familles d'ensembles, et différentes distances (normes), pour définir ces quantités, il existe plusieurs définitions possibles de la discrépance. Pour mieux comprendre les origines et les différentes propriétés des discrépances que nous allons définir, nous renvoyons au paragraphe 2.2 du chapitre II, ainsi qu'à Thiémard (2000), Hickernell (1998), et aux références bibliographiques contenues dans ces articles.

1.3.1 Notations

Précisons les notations que nous utilisons pour les définitions qui suivent. Nous notons :

(I.1) $\mathcal{I} := [0,1)^d$;

(I.2) $\mathcal{X} := \bar{\mathcal{I}} = [0,1]^d$;

(I.3) z un élément de $\mathcal{X} = [0,1]^d$, $z := (z^{(1)}, \ldots, z^{(d)})' \in \mathcal{X}$;

(I.4) $\boldsymbol{x}(n) := \{x_1, \ldots, x_n\}$ une suite de points dans $\mathcal{X}$; $x_i = (x_i^{(1)}, \ldots, x_i^{(d)})' \in \mathcal{X}$ pour $i = 1, \ldots, n$ et $\boldsymbol{x}(n) \in \mathcal{X}^n$;

(I.5) $P := J(\alpha, \beta)$ un pavé (produit d'intervalles) dans $\mathcal{I} = [0,1)^d$; pour

$$\alpha = (\alpha^{(1)}, \ldots, \alpha^{(d)})' \in \mathcal{X} = [0,1]^d, \quad \text{et}$$
$$\beta = (\beta^{(1)}, \ldots, \beta^{(d)})' \in \mathcal{X}, \qquad \text{avec} \quad 0 \le \alpha^{(j)} \le \beta^{(j)} \le 1, \forall j \in \{1, \ldots, d\},$$

$$P := J(\alpha, \beta) = \prod_{j=1}^{d} [\alpha^{(j)}, \beta^{(j)});$$

(I.6) $\mathcal{P}$ l'ensemble des pavés P contenus dans $\mathcal{I} = [0,1)^d$,

$$\mathcal{P} = \left\{ P := \prod_{j=1}^{d} [\alpha^{(j)}, \beta^{(j)}) : 0 \le \alpha^{(j)} \le \beta^{(j)} \le 1 \right\};$$

(I.7) $P^*(z)$ un pavé *ancré à l'origine* (produit d'intervalles ayant tous pour extrémité inférieure 0) dans $\mathcal{I} = [0,1)^d$; pour $z = (z^{(1)}, \ldots, z^{(d)})' \in \mathcal{X} = [0,1]^d$ avec $0 \leq z^{(j)} \leq 1$,

$$P^*(z) := \prod_{j=1}^{d} [0, z^{(j)});$$

(I.8) $\mathcal{P}^*$ l'ensemble des pavés *ancrés à l'origine* contenus dans $\mathcal{I} = [0,1)^d$,

$$\mathcal{P}^* := \left\{ P^* = \prod_{j=1}^{d} [0, z^{(j)}) : 0 \leq z^{(j)} \leq 1 \right\};$$

(I.9) $\boldsymbol{\lambda}(P)$ le volume (mesure de Lebesgue) d'un pavé P dans $\mathcal{X}$;

pour $\alpha = (\alpha^{(1)}, \ldots, \alpha^{(d)})' \in \mathcal{X} = [0,1]^d$,
et $\beta = (\beta^{(1)}, \ldots, \beta^{(d)})' \in \mathcal{X}$, avec $0 \leq \alpha^{(j)} \leq \beta^{(j)} \leq 1, \forall j \in \{1, \ldots, d\}$,
et $P = \prod_{j=1}^{d} [\alpha^{(j)}, \beta^{(j)})$,

$$\boldsymbol{\lambda}(P) := \prod_{j=1}^{d} (\beta^{(j)} - \alpha^{(j)});$$

(I.10) $\#\{E \cap \boldsymbol{x}(n)\}$ le nombre d'éléments d'une suite $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ appartenant à l'ensemble $E \subset \mathcal{X} = [0,1]^d$;

(I.11) A l'ensemble des 2^d sommets du cube unité dans $\mathbb{R}^d$,

$$A := \left\{ a \in \mathcal{X} : a^{(j)} = 0 \text{ ou } 1, \forall j \in \{1, \ldots, d\} \right\};$$

(I.12) $J(a,z)$ le pavé dont les extrémités inférieures et supérieures sont définies par les points $z \in \mathcal{X}$ et $a \in A$; $J(a,x)$ est défini de la façon suivante, pour un sommet $a \in A$ et pour $z \in \mathcal{X}$,

$$J(a,z) := \left\{ z' \in \mathcal{I} = [0,1)^d : \min(a^{(j)}, z^{(j)}) \leq z'^{(j)} < \max(a^{(j)}, z^{(j)}),\ \forall j \in \{1,\dots,d\} \right\};$$

(I.13) $a(z)$ l'unique sommet de A (défini en (I.11)) qui est le plus proche du point $z \in \mathcal{X}$, c'est-à-dire, l'unique sommet de $[0,1]^d$ tel que $z \in J(a(z), (1/2,\dots,1/2)')$, où J est définie comme en (I.5) ;

(I.14) F_n la fonction de répartition empirique d'une suite de points $\boldsymbol{x}(n) = \{x_1,\dots,x_n\}$ dans $\mathcal{X}$; pour $z = (z^{(1)},\dots,z^{(d)})' \in \mathcal{X}$,

$$F_n(z) := \sum_{i=1}^{n} \mathbb{1}_{\{x_i^{(1)} \leq z^{(1)},\dots,x_i^{(d)} \leq z^{(d)}\}};$$

(I.15) U la fonction de répartition uniforme dans $\mathcal{X} = [0,1]^d$; pour $z = (z^1,\dots,z^{(n)})' \in \mathcal{X}$, où $\mathcal{X} = [0,1]^d$,

$$U(z) := \prod_{j=1}^{d} z^{(j)};$$

(I.16) $\|.\|_{L^p(\mathcal{I})}$ la norme L^p dans $L^p(\mathcal{I})$ (où, cf. (I.1) $\mathcal{I} = [0,1)^d$) ; pour $f \in L^p(\mathcal{I})$ avec $1 \leq p < \infty$,

$$\|f\|_{L^p(\mathcal{I})} := \left(\int_{\mathcal{I}} |f(z)|^p dz \right)^{1/p};$$

(I.17) $\|.\|_{L^\infty(\mathcal{I})}$ la norme L^∞ dans $L^\infty(\mathcal{I})$; pour $f \in L^\infty(\mathcal{I})$,

$$\|f\|_{L^\infty(\mathcal{I})} := \sup_{z \in \mathcal{I}} |f(z)|;$$

(I.18) $\boldsymbol{u}$ un ensemble non vide d'indices distincts de 1 à d, soit $\boldsymbol{u} = \{1 \leq u_1,\dots,u_\ell \leq d\}$;

(I.19) $\mathcal{I}^{(\boldsymbol{u})}$, le pavé unité dont les axes sont déterminés par l'ensemble non vide d'indices $\boldsymbol{u} = \{1 \leq u_1, \ldots, u_\ell \leq d\}$; pour $(1^{(1)}, \ldots, 1^{(d)})$,

$$\mathcal{I}^{(\boldsymbol{u})} := \prod_{j=1}^{\ell} [0, 1^{(\boldsymbol{u}_j)}), \quad \text{et} \quad \bar{\mathcal{I}}^{(\boldsymbol{u})} := \prod_{j=1}^{\ell} [0, 1^{(\boldsymbol{u}_j)}];$$

(I.20) $z^{(\boldsymbol{u})} \in [0,1]^\ell$ le point (vecteur) extrait de $z = (z^{(1)}, \ldots, z^{(d)})' \in \mathcal{X}$ dont les composantes sont indexées par les indices de l'ensemble non vide $\boldsymbol{u} = \{1 \leq u_1, \ldots, u_\ell \leq d\}$,

$$z^{(\boldsymbol{u})} := (z^{(u_1)}, \ldots, z^{(u_\ell)})';$$

(I.21) $\boldsymbol{x}^{(\boldsymbol{u})}(n) = \{x_1{}^{(\boldsymbol{u})}, \ldots, x_n{}^{(\boldsymbol{u})}\}$ la suite des points issus de $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ dont les composantes de chaque point $x_i{}^{(\boldsymbol{u})} \in \bar{\mathcal{I}}^{(\boldsymbol{u})}$ sont restreintes à celles indexées par $\boldsymbol{u} \subset \{1, \ldots, d\}$, ensemble d'indices non vide ; $x_i{}^{(\boldsymbol{u})} = (x_i{}^{(u_1)}, \ldots, x_i{}^{(u_\ell)})'$, pour $i = 1, \ldots, n$;

(I.22) $J(\alpha^{(\boldsymbol{u})}, \beta^{(\boldsymbol{u})})$ un pavé (produit d'intervalles) dans $\mathcal{I}^{(\boldsymbol{u})}$; pour $\alpha^{(\boldsymbol{u})} = (\alpha^{(u_1)}, \ldots, \alpha^{(u_\ell)})' \in \bar{\mathcal{I}}^{(\boldsymbol{u})}$ et $\beta^{(\boldsymbol{u})} = (\beta^{(u_1)}, \ldots, \beta^{(u_l)})' \in \bar{\mathcal{I}}^{(\boldsymbol{u})}$ avec $0 \leq \alpha^{(u_j)} \leq \beta^{(u_j)} \leq 1$, $\forall j \in \{1, \ldots, \ell\}$,

$$J(\alpha^{(\boldsymbol{u})}, \beta^{(\boldsymbol{u})}) := \prod_{j=1}^{\ell} [\alpha^{(u_j)}, \beta^{(u_j)});$$

(I.23) $P^*(z^{(\boldsymbol{u})})$ un pavé *ancré à l'origine* ayant pour extrémités supérieures les composantes de $z^{(\boldsymbol{u})}$, où $z^{(\boldsymbol{u})}$ est un vecteur extrait de $z = (z^{(1)}, \ldots, z^{(d)})' \in \mathcal{X}$ dont les composantes sont indexées par les indices de l'ensemble non vide $\boldsymbol{u} = \{1 \leq u_1, \ldots, u_\ell \leq d\}$; pour $z^{(\boldsymbol{u})} = (z^{(u_1)}, \ldots, z^{(u_\ell)})$,

$$P^*(z^{(\boldsymbol{u})}) := \prod_{j=1}^{\ell} [0, z^{(u_j)});$$

(I.24) $\boldsymbol{\lambda}^{(\boldsymbol{u})}\left(J(\alpha^{(\boldsymbol{u})}, \beta^{(\boldsymbol{u})})\right)$ le volume (mesure de Lebesgue dans $\mathcal{I}^{(\boldsymbol{u})}$ voir (I.19)) d'un pavé $J(\alpha^{(\boldsymbol{u})}, \beta^{(\boldsymbol{u})})$ (voir (I.5)) ; pour $\alpha^{(\boldsymbol{u})} = (\alpha^{(u_1)}, \ldots, \alpha^{(u_\ell)})' \in \mathcal{I}^{(\bar{\boldsymbol{u}})}$ et $\beta = (\beta^{(u_1)}, \ldots, \beta^{(u_l)})' \in \mathcal{I}^{(\bar{\boldsymbol{u}})}$ avec $0 \leq \alpha^{(u_j)} \leq \beta^{(j)} \leq 1$, $\forall j \in \{1, \ldots, \ell\}$,

$$\boldsymbol{\lambda}^{(\boldsymbol{u})}\left(J(\alpha^{(\boldsymbol{u})},\beta^{(\boldsymbol{u})})\right) := \prod_{j=1}^{\ell}(\alpha^{(u_j)}-\beta^{(u_j)});$$

(I.25) $A^{(\boldsymbol{u})}$ les 2^ℓ sommets du cube $\bar{\mathcal{I}}^{(\boldsymbol{u})}$ (voir (I.19)) dont les composantes sont indexées par
$\boldsymbol{u} = \{1 \le u_1, \ldots, u_\ell \le d\}$ (ℓ est le nombre d'éléments de l'ensemble (non vide) $\boldsymbol{u}$, $\ell = \#(\boldsymbol{u})$),

$$A^{(\boldsymbol{u})} := \left\{a^{(\boldsymbol{u})} \in \bar{\mathcal{I}}^{(\boldsymbol{u})} : a^{(u_j)} = 0 \text{ ou } 1,\ \forall j \in \{1,\ldots,\ell\}\right\};$$

(I.26) $J(a^{(\boldsymbol{u})}, z^{(\boldsymbol{u})})$ le pavé dont les extrémités inférieures et supérieures sont définies par les points $z^{(\boldsymbol{u})} \in \mathcal{I}^{(\boldsymbol{u})}$ et $a^{(\boldsymbol{u})} \in A^{(\boldsymbol{u})}$, voir (I.19) et (I.25) ; pour $x^{(\boldsymbol{u})} \in \mathcal{I}^{(\boldsymbol{u})}$ et $a^{(\boldsymbol{u})} \in A^{(\boldsymbol{u})}$, $J(a^{(\boldsymbol{u})}, z^{(\boldsymbol{u})})$ est défini comme suit,

$$J(a^{(\boldsymbol{u})}, z^{(\boldsymbol{u})}) :=$$
$$\left\{z'^{(\boldsymbol{u})} \in \mathcal{I}^{(\boldsymbol{u})} : \min(a^{(u_j)}, z^{(u_j)}) \le z'^{(u_j)} < \max(a^{(u_j)}, z^{(u_j)}),\ \forall j \in \{1,\ldots,\ell\}\right\};$$

(I.27) $a^{(\boldsymbol{u})}(z)$ l'unique sommet de $A^{(\boldsymbol{u})}$ (voir (I.25)) qui est le plus proche du point $z^{(\boldsymbol{u})} \in \mathcal{I}^{(\boldsymbol{u})}$ (voir (I.19)) où $z^{(\boldsymbol{u})} = \{z^{(u_1)}, \ldots, z^{(u_\ell)}\}$ est un point (vecteur) extrait de $z = (z^{(1)}, \ldots, z^{(d)}) \in \mathcal{X}$, pour un ensemble non vide d'indices
$\boldsymbol{u} = \{1 \le u_1, \ldots, u_\ell \le d\}$;

(I.28) $\boldsymbol{\sigma}$ une somme définie pour les sommets $a^{(\boldsymbol{u})} \in A^{(\boldsymbol{u})}$ (voir (I.25)) dont les composantes sont celles correspondant aux indices d'un ensemble non vide $\boldsymbol{u} = \{1 \le u_1, \ldots, u_\ell \le d\}$,

$$\boldsymbol{\sigma}(a^{(\boldsymbol{u})}) := \sum_{j \in \boldsymbol{u} = \{1 \le u_1, \ldots, u_\ell \le d\}} a^{(u_j)} \quad (\mod 2),$$

cette somme est égale à 0 quand la somme des composantes du sommet $a^{(\boldsymbol{u})}$ est paire et est égale à 1 sinon ; lorsque la somme $\boldsymbol{\sigma}(a^{(\boldsymbol{u})})$ est égale à 0, nous dirons que le sommet $a^{(\boldsymbol{u})}$ est *pair*, et que le pavé $J(a^{(\boldsymbol{u})}, z^{(\boldsymbol{u})})$ est *pair* ;

(I.29) $J_e(a^{(\boldsymbol{u})}, z^{(\boldsymbol{u})})$ l'union de pavés *pairs* ; pour les sommets *pairs* (i.e. tels que $a^{(\boldsymbol{u})} \in A^{(\boldsymbol{u})}$ et $\boldsymbol{\sigma}(a^{(\boldsymbol{u})}) = 0$) et $z^{(\boldsymbol{u})}$ un point extrait de $z = (z^{(1)}, \ldots, z^{(d)})' \in \mathcal{X}$ dont les composantes sont indexées par l'ensemble non vide $\boldsymbol{u} = \{1 \leq u_1, \ldots, u_\ell \leq d\}$,

$$J_e(a^{(\boldsymbol{u})}, z^{(\boldsymbol{u})}) := \bigcup_{\boldsymbol{\sigma}(a^{(\boldsymbol{u})})=0} J(a^{(\boldsymbol{u})}, z^{(\boldsymbol{u})}).$$

(I.30) $F_n^{(\boldsymbol{u})}$ une fonction de répartition d'une suite de points $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ restreinte aux composantes indexées par l'ensemble non vide $\boldsymbol{u} = \{1 \leq u_1, \ldots, u_\ell \leq d\}$ des éléments de $\boldsymbol{x}(n)$; pour $\boldsymbol{u} = \{1 \leq u_1 \leq \cdots \leq u_\ell\}$ et $z = (z^{(1)}, \ldots, z^{(d)}) \in \mathcal{X}$,

$$F_n^{(\boldsymbol{u})}(z) := \sum_{i=1}^{n} \mathbb{1}_{\{x_i^{(u_1)} \leq z^{(u_1)}, \ldots, x_i^{(u_\ell)} \leq z^{(u_\ell)}\}};$$

sur $\bar{\mathcal{I}}^{(\boldsymbol{u})}$ (voir (I.19)), il s'agit de la fonction de répartition de la suite $\boldsymbol{x}^{(\boldsymbol{u})}(n) = \{x_1{}^{(\boldsymbol{u})}, \ldots, x_n{}^{(\boldsymbol{u})}\}$;

(I.31) $U^{(u)}$ la fonction de répartition uniforme dans $\bar{\mathcal{I}}^{(\boldsymbol{u})}$ (voir (I.19)) ;
pour $z^{(\boldsymbol{u})} = (z^{(u_1)}, \ldots, z^{(u_\ell)}) \in \bar{\mathcal{I}}^{(\boldsymbol{u})}$,

$$U(z^{(\boldsymbol{u})}) := \prod_{j=1}^{\ell} z^{(u_j)};$$

(I.32) $\|.\|_{L^p(\mathcal{I}^{(\boldsymbol{u})})}$ la norme L^p dans $L^p(\mathcal{I}^{(\boldsymbol{u})})$; pour $f \in L^p(\mathcal{I}^{(\boldsymbol{u})})$ avec $1 \leq p < \infty$,

$$\|f\|_{L^p(\mathcal{I}^{(\boldsymbol{u})})} := \left(\int_{\mathcal{I}^{(\boldsymbol{u})}} |f(z)|^p dz \right)^{1/p}.$$

1.3.2 Définitions

Nous exposons ci-dessous différentes définitions, correspondant à différentes versions de la *discrépance* (voir aussi chapitre II).

– La discrépance extrême

Définition 1.3.1
Soit $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ une suite de n points dans $\mathcal{X}$, la discrépance extrême de $\boldsymbol{x}(n)$ est définie par :

$$\boxed{\mathrm{D}(\boldsymbol{x}(n)) := \sup_{P \in \mathcal{P}} \left| \frac{\#\{P \cap \boldsymbol{x}(n)\}}{n} - \boldsymbol{\lambda}(P) \right|.} \tag{1.23}$$

Rappelons, par (I.6), que $\mathcal{P}$ est l'ensemble des pavés contenus dans $\mathcal{I} = [0,1)^d$. Cette définition correspond bien à une comparaison entre le nombre de points contenus dans un pavé et le volume de ce pavé. Son calcul revient à chercher le pavé qui contient la densité de points la plus anormalement élevée comparativement à son volume.

– La discrépance à l'origine

Définition 1.3.2
Soit $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ une suite de n points dans $\mathcal{X}$, nous définissons la discrépance à l'origine par :

$$\boxed{\mathrm{D}^*(\boldsymbol{x}(n)) := \sup_{P \in \mathcal{P}^*} \left| \frac{\#\{P \cap \boldsymbol{x}(n)\}}{n} - \boldsymbol{\lambda}(P) \right|.} \tag{1.24}$$

Nous désignons ici des pavés *ancrés à l'origine*, par $\mathcal{P}^*$, voir (I.8). Remarquons que nous pouvons plus simplement écrire :

$$\mathrm{D}^*(\boldsymbol{x}(n)) := \|F_n - U\|_{L^\infty(\mathcal{I})},$$

où F_n désigne la fonction de répartition empirique de la suite $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ et U désigne la fonction de répartition uniforme sur $\mathcal{X} = [0,1]^d$. La discrépance à l'origine correspond donc à une distance entre F_n et U. C'est la norme L^∞ de la différence de ces deux fonctions.

Comme il existe différentes normes L^p correspondant à des choix différents de $1 \leq p \leq \infty$, nous pouvons aussi définir différentes discrépances L^p.

– La discrépance L^p

Définition 1.3.3
Soit $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ une suite de n points dans $\mathcal{X}$. Désignons par F_n et U, respectivement, la fonction de répartition empirique de $\boldsymbol{x}(n)$, et la fonction de répartition uniforme dans $\mathcal{X} = [0,1]^d$. Nous définissons la discrépance L^p par :

$$\mathrm{D}^{L^p}(\boldsymbol{x}(n)) := \|F_n - U\|_{L^p(\mathcal{I})}.$$

La plus étudiée parmi les discrépances L^p est, bien entendu, la discrépance L^2, soit :

$$\mathrm{D}^{L^2}(\boldsymbol{x}(n)) := \left\{ \int_{\mathcal{X}} |F_n(z) - U(z)|^2 dz \right\}^{1/2}. \tag{1.25}$$

– La discrépance généralisée

Nous allons définir d'autres variantes de la discrépance issues de la notion de *discrépance généralisée* introduite par Hickernell (1998). Cette dernière sera brièvement

rappelée et expliquée au chapitre II. Les discrépances modifiée, centrée, et symétrique, définies ci-dessous, en sont des cas particuliers. Les définitions qui sont ici données correspondent en fait plus à leur interprétation géométrique, qu'à leur définition théorique (cf chapitre II).

Pour les définir nous utilisons les notations (I.1) à (I.32) des notations introduites en début de cette partie.

- La discrépance modifiée

Nous utilisons ici les notations du paragraphe (1.3.1), en particulier : (I.10), (I.18), (I.20), (I.21), (I.23), (I.24) et (I.32).

Définition 1.3.4
Pour $1 \leq p < \infty$, $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$, *nous définissons la discrépance modifiée par :*

$$\boxed{\mathrm{DM}^{L^p}(\boldsymbol{x}(n)) := \left[\sum_{\boldsymbol{u} \subset \{1,\ldots,d\}} \left\| \frac{\#\left\{P^*(z^{(\boldsymbol{u})}) \cap \boldsymbol{x}^{(\boldsymbol{u})}(n)\right\}}{n} - \boldsymbol{\lambda}^{(\boldsymbol{u})}\left(P^*(z^{(\boldsymbol{u})})\right) \right\|^p_{L^p(\mathcal{I}^{(\boldsymbol{u})})} \right]^{1/p},}$$

que nous notons encore :

$$\mathrm{DM}^{L^p}(\boldsymbol{x}(n)) := \left\| \left(\frac{\#\left\{P^*(z^{(\boldsymbol{u})}) \cap \boldsymbol{x}^{(\boldsymbol{u})}(n)\right\}}{n} - \boldsymbol{\lambda}^{(\boldsymbol{u})}\left(P^*(z^{(\boldsymbol{u})})\right) \right)_{\boldsymbol{u} \neq \emptyset} \right\|_p .$$

Nous pouvons encore écrire cette discrépance sous la forme :

$$\mathrm{DM}^{L^p}(\boldsymbol{x}(n)) := \left[\sum_{\boldsymbol{u}} \| F_n^{(\boldsymbol{u})} - U^{(\boldsymbol{u})} \|^p_{L^p(\mathcal{I}^{\boldsymbol{u}})} \right]^{1/p} .$$

Conformément aux notations (I.30), (I.31) et (I.32) introduites au paragraphe (1.3.1), $F_n^{(\boldsymbol{u})}$ désigne ici la fonction de répartition empirique de $x^{(\boldsymbol{u})}(n)$, et $U^{(\boldsymbol{u})}$ la fonction de répartition uniforme dans le cube $\bar{\mathcal{I}}^{(\boldsymbol{u})}$, dont les composantes sont indexées par un ensemble non vide $\boldsymbol{u} \subset \{1, \ldots, d\}$. Nous notons encore :

$$\mathrm{DM}^{L^p}(\boldsymbol{x}(n)) = \| F_n - U \|_p .$$

Remarquons que la généralisation de DM^{L^p} pour le cas $p = \infty$ coïncide avec *la discrépance à l'origine* (voir définition (1.3.2)).

- La discrépance L^p centrée

Les définitions des discrépances L^p, et L^p modifiée, dépendent fortement du choix de l'origine, puisque nous considérons dans leur calcul, des pavés *ancrés à l'origine*. L'idée est de considérer l'ensemble A de tous les sommets du cube unité de façon à ce que la valeur de la discrépance soit la même lorsque nous effectuons une réflexion de la suite par rapport à n'importe quel plan $z^{(j)} = 1/2$. Pour cela, nous utilisons les notations introduites au paragraphe (1.3.1).

Définition 1.3.5
Pour $1 \leq p < \infty$, $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$, nous définissons la discrépance centrée par

$$\mathrm{DC}^{L^p}(\boldsymbol{x}(n)) := \left[\sum_{\boldsymbol{u} \subset \{1,\ldots,d\}} \left\| \frac{\#\left\{\boldsymbol{x}^{(\boldsymbol{u})}(n) \cap J\left(a^{(\boldsymbol{u})}(z), z^{(\boldsymbol{u})}\right)\right\}}{n} - \boldsymbol{\lambda}^{(\boldsymbol{u})}\left(J(a^{(\boldsymbol{u})}(z), z^{(\boldsymbol{u})})\right) \right\|^p_{L^p(\mathcal{I}^{(\boldsymbol{u})})} \right]^{1/p}.$$

Cette discrépance ne dépend plus uniquement des pavés *ancrés à l'origine* puisque elle fait intervenir les pavés $J\left(a^{(\boldsymbol{u})}(z), z^{(\boldsymbol{u})}\right)$ qui, pour un point $z \in \mathcal{X}$, sont définis par le point $z^{(\boldsymbol{u})} \in \mathcal{I}^{(\boldsymbol{u})}$ et son plus proche sommet $a^{(\boldsymbol{u})}(z) \in \mathcal{I}^{(\boldsymbol{u})}$, voir les notations (I.19), (I.20), (I.27) du paragraphe (1.3.1). Ceci est illustré par la Figure (1.5) où le point $z = (0.6, 0.7) \in [0,1)^2$, le pavé alors considéré est celui en rouge. Elle est invariante si nous remplaçons la suite $\boldsymbol{x}(n)$ par $1 - \boldsymbol{x}(n)$. Cette définition est valable pour $p < \infty$ et ne se généralise pas convenablement pour $p = \infty$ (voir Hickernell (1998)).

Signalons enfin une dernière définition de la discrépance : la discrépance L^p symétrique, que nous décrivons ci-dessous.

- La discrépance L^p symétrique

Nous utilisons les notations du paragraphe (1.3.1), en particulier les notations (I.28) et (I.29).

Définition 1.3.6
Pour $1 \leq p < \infty$, $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$, la discrépance L^p symétrique est définie par

$$\mathrm{DS}^{L^p}(\boldsymbol{x}(n)) := \left[\sum_{\boldsymbol{u} \subset \{1,\ldots,d\}} \left\| \frac{\#\left\{\boldsymbol{x}^{(\boldsymbol{u})}(n) \cap J_e\left(a^{(\boldsymbol{u})}, z^{(\boldsymbol{u})}\right)\right\}}{n} - \boldsymbol{\lambda}^{(\boldsymbol{u})}\left(J_e\left(a^{(\boldsymbol{u})}, z^{(\boldsymbol{u})}\right)\right) \right\|^p_{L^p(\mathcal{I}^{(\boldsymbol{u})})} \right]^{1/p}.$$

Cette définition de la discrépance ne fait pas intervenir uniquement les pavés *ancrés à l'origine*, puisque nous considérons les pavés *pairs* $J_e\left(a^{(\boldsymbol{u})}, z^{(\boldsymbol{u})}\right)$, voir en particulier la notation (I.29) du paragraphe 1.3.1. Une union de pavés « pairs » associée au point $z = (0.5, 0.7)$ est représentée Figure (1.6). Cette discrépance est invariante si nous remplaçons la suite $\boldsymbol{x}(n)$ par $1 - \boldsymbol{x}(n)$.

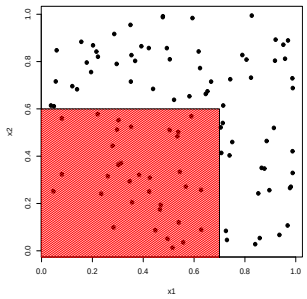


FIG. 1.4 – Pavé « ancré à l'origine » associée à $z = (0.6, 0.7)$

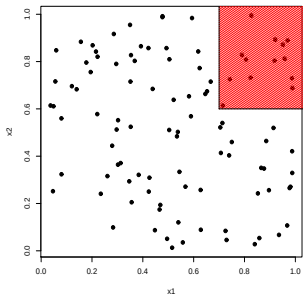


FIG. 1.5 – Pavé « centré » associé à $z = (0.6, 0.7)$

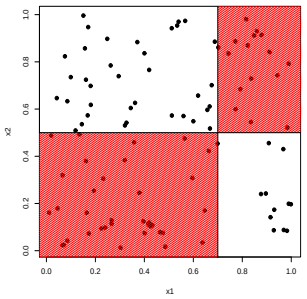


FIG. 1.6 – Union de pavé « pairs » associée à $z = (0.5, 0.7)$

Les Figures (1.4), (1.5), (1.6), représentent en rouge les pavés considérés pour les définitions des différentes discrépances définies ci-dessus dans l'espace $\mathcal{X} = [0, 1]^2$. Pour chaque discrépance, il s'agit de considérer l'ensemble de ces pavés et de comparer le nombre de points qu'il contiennent avec leur volume à l'aide d'une norme (c.f. définitions (1.3.4), (1.3.5), (1.3.6)).

1.3.3 Propriétés

Décrivons à présent différentes propriétés des discrépances préalablement définies.

1.3.3.1 Inégalité

Nous commençons par préciser les différentes inégalités entre les discrépances (voir Hickernell (1998) et Thiémard (2000)). Pour toute suite $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ de n points dans $\mathcal{X}$, en notant (conformément aux définitions du paragraphe précédent) D^{L^p}, la discrépance L^p, DM^{L^p}, la discrépance L^p modifiée, D^*, la discrépance à l'origine, D, la discrépance extrême, nous avons :

$$0 < \mathrm{D}^{L^p}(\boldsymbol{x}(n)) \leq \mathrm{DM}^{L^p}(\boldsymbol{x}(n)) \leq \mathrm{D}^*(\boldsymbol{x}(n)) \leq D(\boldsymbol{x}(n)) \leq 1\,. \tag{1.26}$$

Définissons à présent quelques bornes inférieures.

Théorème 1.3.1 (Roth 1954)
Pour toute suite $\boldsymbol{x}(n)$ contenant n points dans $\mathcal{X} = [0,1]^d$, il existe une constante $c_d^{L^2}$ ne dépendant que de la dimension d, telle que :

$$\mathrm{D}^{L^2}(\boldsymbol{x}(n)) \quad \geq \quad c_d^{L^2} \frac{(\log n)^{(d-1)/2}}{n}.$$

Démonstration :
Pour la démonstration de ce théorème, nous nous référons à Roth (1954).

□

Notons qu'il est théoriquement possible de construire des suites de points $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ dont l'ordre de la discrépance L^2 est en $O\left(n^{-1}(\log n)^{(d-1)/2}\right)$. Cet ordre est optimal, puisqu'il est le même pour les bornes inférieures et supérieures. La construction de telles suites (d'ordre optimal pour la discrépance L^2) n'est cependant pas connue sous forme explicite (à notre connaissance).

A l'aide des inégalités (1.26) et du théorème précédent, le lemme suivant est immédiat,

Lemme 1.3.1
Pour toute suite $\boldsymbol{x}(n)$ contenant n points dans $\mathcal{X} = [0,1]^d$, il existe une constante $c_d > 0$ ne dépendant que de la dimension d, telle que :

$$D(\boldsymbol{x}(n)) \geq \mathrm{D}^*(\boldsymbol{x}(n)) \geq \mathrm{DM}^{L^2}(\boldsymbol{x}(n)) \geq c_d \frac{(\log n)^{(d-1)/2}}{n}, \tag{1.27}$$

où D, D^* et DM^{L^2} désignent respectivement la discrépance modifiée L^2, la discrépance à l'origine et la discrépance extrême.

Pour le cas de la discrépance à l'origine D^*, il n'existe, à notre connaissance, aucune construction générale de suite permettant d'atteindre la borne inférieure $\frac{(\log n)^{(d-1)/2}}{n}$.

A titre d'indication, signalons une minoration plus précise de la discrépance à l'origine. Celle-ci est obtenue à l'aide du théorème suivant,

Théorème 1.3.2 (Baker 1999)
Pour toute suite $\boldsymbol{x}(n)$ contenant n points dans $\mathcal{X} = [0,1]^d$, il existe une constante $B_d > 0$ ne dépendant que de la dimension d, telle que :

$$\mathrm{D}^*(\boldsymbol{x}(n)) \geq B_d \frac{(\log n)^{(d-1)/2}}{n} \left(\frac{\log \log n}{\log \log \log n} \right)^{1/(2d-3)}. \tag{1.28}$$

Démonstration :
Pour la démonstration de ce théorème, nous nous référons à Baker (1999).

□

Plus généralement, nous ***conjecturons*** que pour une suite contenant une infinité de points $\boldsymbol{x} = \{x_1, \ldots, x_n, \ldots\}$ dans $\mathcal{X}$, l'ordre optimal de convergence vers 0 que nous pouvons obtenir pour la discrépance à l'origine D^* des n premiers points $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ de la suite $\boldsymbol{x}$ est, lorsque $n \to \infty$

$$\mathrm{D}^*(\boldsymbol{x}(n)) = O\left(\frac{(\log n)^d}{n} \right). \tag{1.29}$$

Aucune suite présentant un taux de décroissance vers 0 plus rapide n'a été construite à notre connaissance, et nous supposons donc qu'il n'en existe pas. C'est à l'aide de cette ***conjecture*** que sont définies les *suites à discrépance faible*, comme suit

Définition 1.3.7
Une suite contenant une infinité de points $\{x_1, \ldots, x_n, \ldots\}$ dans $\mathcal{X}$, et dont les n premiers termes $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ sont construits de façon à vérifier

$$\mathrm{D}^*(\boldsymbol{x}(n)) = O\left(\frac{(\log n)^d}{n} \right),$$

est appelée suite à discrépance faible.

Parmi ces suites, nous pouvons citer les suites de Halton, de Hammersley, de Faure, de Sobol, et aussi plus généralement les réseaux (voir Halton (1960), Hammersley (1960), Faure (1982)).

Lorsque nous souhaitons évaluer la qualité de répartition uniforme d'une BDDE $\boldsymbol{x}(n)$ constituée de n points dans $\mathcal{X}$, les *suites à discrépance faible* peuvent servir comme élément de comparaison. Nous pouvons par exemple comparer la valeur de la discrépance obtenue avec la BDDE $\boldsymbol{x}(n)$ avec celle d'une *suite à discrépance faible* $\boldsymbol{x}_f(n)$ comportant le même nombre de points n. La suite $\boldsymbol{x}_f(n)$ peut alors être considérée comme une suite « étalon ». Nous détaillerons cette approche dans le paragraphe (1.3.3.3).

Une autre inégalité extrêmement importante, faisant intervenir la dispersion,

$$d_\infty(\boldsymbol{x}(n)) \quad := \quad \sup_{w\in\mathcal{X}} \left(\min_{1\leq i\leq n} \left\{ \max_{j=1,\ldots,d} |w_j - x_j| \right\} \right), \tag{1.30}$$

et la discrépance à l'origine D^* (voir définition (1.3.2)) est donnée par le théorème ci-dessous.

Théorème 1.3.3 (Niederreiter 1988)
Pour toute suite $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ d'au moins n points dans $\mathcal{X} = [0,1]^d$, nous avons :

$$d_\infty(\boldsymbol{x}(n)) \leq 2.\left(\mathrm{D}^*(\boldsymbol{x}(n))\right)^{1/d}. \tag{1.31}$$

Démonstration :
Pour la démonstration de ce théorème, nous renvoyons à Niederreiter (1988).

□

Ainsi, toute *suite à discrépance faible* au sens de la définition (1.3.7) est aussi une suite à *dispersion faible* au sens où elle vérifie :

$$d_\infty(\boldsymbol{x}(n)) = O\left(\frac{\log n}{n^{1/d}}\right). \tag{1.32}$$

Nous faisons donc ici le lien avec le paragraphe (Dispersion) du paragraphe (1.2.2.2).

Précisons enfin une propriété concernant le nombre de points nécessaires pour que la discrépance à l'origine soit inférieure à un seuil ε, et la dimension d de $\mathcal{X}$.

Remarque
Pour tout choix d'un $\varepsilon \in (0, 1/2)$, le nombre d'éléments minimum $n(d,\varepsilon)$ de la plus courte suite $\boldsymbol{x}\left(n(d,\varepsilon)\right) = \{x_1, \ldots, x_{n(d,\varepsilon)}\}$ telle que $\mathrm{D}^*\left(\boldsymbol{x}(n(d,\varepsilon))\right) \leq \varepsilon$, croît de façon linéaire avec d, voir Henrich *et al.* (2001).

Autrement dit, si nous nous fixons un seuil ε de la discrépance à l'origine D^*, plus la dimension d est importante, plus le nombre minimal de points $n(d)$ pour que la discrépance à l'origine de $\boldsymbol{x}(n(d)) = \{x_1, \ldots, x_{n(d)}\}$ dans $\mathcal{X} = [0,1]^d$ soit inférieure à ε sera important. Ce nombre de points croît de façon linéaire avec la dimension. Cependant, la croissance du nombre de points nécessaires, $n(d)$, en fonction de la dimension d, des *suites à discrépance faible* $\boldsymbol{x} = \{x_1, \ldots, x_n, \ldots,\}$ construites à ce jour, pour que $\mathrm{D}^*(\boldsymbol{x}(n(d))) < \varepsilon$ n'est jamais linéaire, mais le plus souvent exponentielle.

1.3.3.2 Propriétés des suites aléatoires de loi uniforme sur $[0,1]^d$

Dans ce paragraphe, nous supposons que la suite $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ correspond à la réalisation de n variables aléatoires indépendantes et identiquement distribuées (v.a. i.i.d.) de loi uniforme dans $\mathcal{X} = [0,1]^d$. Ceci n'était pas le cas dans les paragraphes précédents. L'approche est donc différente, et nous parlons d'*approche probabiliste*. Cette approche fera l'objet d'un autre chapitre. Nous la considérons ici car, sous *l'hypothèse d'une distribution uniforme* des points de la BDDE, les discrépances que nous avons définies correspondent à des statistiques dont la loi est parfois connue.

□ Étude des discrépances à l'aide du processus empirique uniforme

Définition 1.3.8
Pour $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ une suite de variables aléatoires indépendantes et de loi uniforme dans $\mathcal{X} = [0,1]^d$, nous définissons le processus empirique uniforme par :

$$\alpha_n(x) \quad := \quad n^{1/2}(F_n(x) - U(x)), \quad \textit{pour } x \in \mathcal{X} \tag{1.33}$$

où F_n désigne la fonction de répartition empirique de la suite $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ dans $\mathcal{X}$, U la fonction de répartition uniforme dans $\mathcal{X}$.

Une façon de construire des statistiques à partir de ce processus est d'évaluer une fonctionnelle de celui-ci, par exemple une norme.

• Si nous utilisons la norme $L^\infty(\mathcal{I})$ (voir notation (I.32) du paragraphe (1.3.1)), par définition de la discrépance à l'origine D^*, (définition (1.3.2)), nous avons :

$$n^{1/2}\mathrm{D}^*(\boldsymbol{x}(n)) \quad := \quad \|\alpha_n\|_{L^\infty(\mathcal{I})}. \tag{1.34}$$

La statistique $\|\alpha_n\|_{L^\infty(\mathcal{I})}$ définie ci-dessus est connue sous le nom de *statistique de Kolmogorov-Smirnov*. Une majoration de sa loi est donnée par le théorème suivant,

Théorème 1.3.4 (Dvoretzky, Kiefer, Wolfowitz 1956)
Soit $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ une suite de variables aléatoires indépendantes et de loi uniforme dans $\mathcal{X} = [0,1]^d$, pour tout $\varepsilon \geq 0$, il existe une constante $C_{\varepsilon,d}$ telle que

$$P(n^{1/2}\mathrm{D}^*(\boldsymbol{x}(n)) \geq t) \quad \leq \quad C_{\varepsilon,d} \exp\left(-(2-\varepsilon)t^2\right). \tag{1.35}$$

Démonstration :
Pour la démonstration de ce théorème, nous renvoyons à Dvoretzky *et al.* (1956).

□

Précisons que dans le cas de la dimension $d = 1$, nous pouvons obtenir, (voir Massart (1990)) :

$$P(n^{1/2}\mathrm{D}^*(\boldsymbol{x}(n)) \geq t) \leq 2\exp(-2t^2).$$

Ainsi, il est possible d'effectuer des tests statistiques. Nous ferons alors le test d'hypothèse : les x_i, $i = 1, \ldots, n$ sont des réalisations de variables aléatoires indépendantes et identiquement distribuées de loi uniforme dans $\mathcal{X} = [0, 1]$.

Dans le cas où la dimension $d \geq 2$, le choix de $C_{\varepsilon,d} = 2$, optimal en dimension $d = 1$, ne s'applique plus. Le calcul de la loi de la statistique de Kolmogorov-Smirnov est alors délicat et une expression exacte de la loi limite n'est pas connue à ce jour. Le théorème de Dvoretzky-Kiefer-Wolfowicz est par conséquent difficilement exploitable dans notre contexte. Dans ce contexte, une stratégie de tests par projection des données sur divers sous-espaces est proposée par Franco *et al.* (2006).

• Si nous utilisons la norme $L^2(\mathcal{I})$, avec les notations (I.30), (I.31), (I.32) introduites au paragraphe (1.3.1), et par définition des discrépances L^2 et L^2 modifiée (définitions (1.3.3), (1.3.4)), nous avons :

$$n^{1/2}\mathrm{D}^{L^2}(\boldsymbol{x}(n)) := \|\alpha_n\|_{L^2(\mathcal{I})}, \tag{1.36}$$

$$n^{1/2}\mathrm{DM}^{L^2}(\boldsymbol{x}(n)) := \sum_{\boldsymbol{u}\subset 1,\ldots,d} \|\alpha_n^{(\boldsymbol{u})}\|_{L^2(\mathcal{I}^{(\boldsymbol{u})})}, \tag{1.37}$$

où $\alpha_n^{(\boldsymbol{u})} = n^{1/2}(F_n^{(\boldsymbol{u})}(x) - U^{(\boldsymbol{u})}(x))$ est le processus empirique associée à la projection des points $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ dans les sous espaces définis par les axes de $[0, 1]^d$ indexés par $\boldsymbol{u}$ (voir notations (I.19), (I.30), (I.31), paragraphe (1.3.1)).

La statistique définie par l'égalité (1.36) est appelée statistique de Cramer-Von-Mises. Celle-ci converge vers un pont brownien standard multivarié (voir Araujo et Giné (1980), par exemple). A l'aide d'un développement de Karhunen-Loève de ce pont brownien (voir Deheuvels *et al.* (2006)), il est possible d'obtenir :

$$\int_{[0,1]^{2d}} (\alpha_n(x))^2\, dx^{(1)} \ldots dx^{(d)} \xrightarrow{\mathcal{L}} \sum_{k_1,\ldots,k_d\geq 0} \lambda_{k_1\ldots k_d} Y^2_{k_1\ldots k_d}$$

où : $\lambda_{k_1\ldots k_d}$ sont des constantes positives croissantes, et $Y_{k_{u_1}\ldots k_d}$ des v.a. i.i.d. de loi $\mathcal{N}(0, 1)$. La statistique de Cramer-Von-Mises converge donc vers une somme pondérée de variables aléatoires de loi du χ^2.

La statistique définie par l'égalité (1.37), faisant intervenir la discrépance modifiée, est en fait la somme des statistiques de Cramer-Von-Mises associées à toutes les projections possibles des points de $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ selon les axes de $[0, 1]^d$. Elle converge

donc aussi vers une somme pondérée de variables aléatoires de loi du χ^2. Les cas des statistiques que l'on peut définir à l'aide des discrépances centrée et symétrique (définitions (1.3.5) et (1.3.6)) sont similaires.

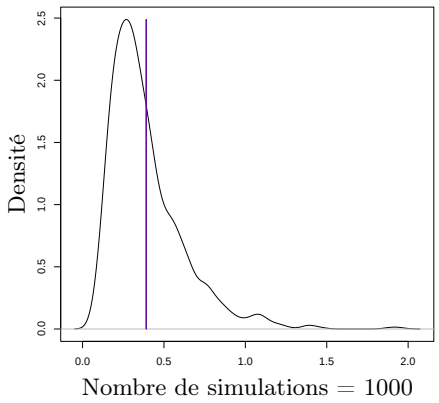


FIG. 1.7 – Densité de $n \times \mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2$

Cependant, les coefficients $\lambda_{k_1 \ldots k_d}$ ne sont pas connus de façon explicite. Pour pallier ce problème, des processus définis sur les marges du pavé unité sont le plus souvent utilisés. On se réfèrera à Deheuvels (1981) et Deheuvels *et al.* (2006). Il semble cependant délicat d'y avoir recours pour connaître la loi exacte des statistiques que l'on peut définir à l'aide des discrépances modifiée, centrée et symétrique. A notre connaissance, ces lois ne sont pas toutes connues de façon explicite pour le moment. Toutefois, il est possible de les tabuler par simulation. La densité obtenue par simulation de la statistique $n \times \mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2$ où $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ est une suite de v.a. i.i.d. dans $\mathcal{X} = [0,1]^2$ est représentée Figure 1.7 (la multiplication par n permet de ne pas faire dépendre la loi de la statistique du nombre de points).

□ Esperances des discrépances au carré

Concernant les moyennes des discrépances carrées nous avons la proposition suivante,

Proposition 1.3.1
Si $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ est une suite de variables aléatoires indépendantes et de loi uniforme dans $\mathcal{X}$ alors,

$$\mathbb{E}(\mathrm{D}^{L^2}(\boldsymbol{x}(n))^2) = \frac{(1/2)^d - (1/3)^d}{n}, \tag{1.38}$$

$$\mathbb{E}(\mathrm{DM}^{L^2}(\boldsymbol{x}(n))^2) = \left(\frac{4}{3}\right)^d \frac{(9/8)^d - 1}{n}, \tag{1.39}$$

$$\mathbb{E}(\mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2) = \frac{1}{n}\left[\left(\frac{13}{12} + \frac{1}{6}\right)^d - \left(\frac{13}{12}\right)^d\right], \tag{1.40}$$

$$\mathbb{E}(\mathrm{DS}^{L^2}(\boldsymbol{x}(n))^2) = \frac{1}{n}\left[\left(\frac{4}{3} + \frac{2}{6}\right)^d - \left(\frac{4}{3}\right)^d\right]. \tag{1.41}$$

Démonstration :
Les égalités (1.38), (1.39), (1.40), (1.41) sont directement obtenues en appliquant les formules dans Hickernell (1996a) et Hickernell (1998).

□

Insistons sur le fait que ces égalités sont valables lorsque la BDDE $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ est la réalisation d'une suite de v.a. i.i.d. de loi uniforme dans $\mathcal{X}$. Théoriquement, nous ne devrions donc pas les utiliser lorsque les n points $x_i \in \mathcal{X}$ ne sont pas définis comme des réalisations de *v.a. i.i.d.* de loi uniforme.

Cependant, elles sont parfois utilisées dans la littérature (essentiellement la discrépance L^2) pour apprécier la qualité de répartition uniforme de *suites à discrépance faible* qui ne sont pas par définition des suites de *v.a. i.i.d.* de loi uniforme. N'ayant aucune valeur de référence des discrépances de type L^2 pour affirmer que la BDDE $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ *recouvre* uniformément l'espace $\mathcal{X}$ de façon « acceptable », les valeurs de ces moyennes seront données à titre d'indication.

Remarquons que, pour un nombre de points n donné d'une suite $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$, contrairement aux moyennes des carrés de discrépance L^2 modifiée, centrée, et symétrique, celle de la discrépance L^2 exprimée par l'égalité (1.38) diminue lorsque la dimension d augmente (ceci peut être vérifié par une simple étude analytique en considérant $\mathbb{E}(\mathrm{D}^{L^2}(\boldsymbol{x}(n))^2)$ comme une fonction de la dimension d). Par conséquent, supposons que nous nous fixons une valeur seuil ε pour la discrépance L^2. Construisons une suite $\boldsymbol{x}(d_1, n)$, de n *v.a. i.i.d.* de loi uniforme dans $\mathcal{X}_{d_1} = [0,1]^{d_1}$, et une suite $\boldsymbol{x}(d_2, n)$, de n *v.a.*

i.i.d. de loi uniforme dans $\mathcal{X}_{d_2} = [0,1]^{d_2}$, avec $d_2 > d_1$. En moyenne la discrépance carrée L^2 de la suite $\boldsymbol{x}(d_2, n)$ dans $\mathcal{X}^{d_2}$ est plus faible que celle de de la suite $\boldsymbol{x}(d_1, n)$. Autrement dit, pour un nombre de points n donné d'une suite de *v.a. i.i.d.*, plus la dimension d de l'espace considéré est grande, plus la discrépance L^2 de la suite serait inférieure à ε. Cette propriété « étrange » de la discrépance L^2 va à l'encontre de la remarque faite en fin du paragraphe précédent et qui concernait la discrépance à l'origine D^*.

Nous pourrions donc penser que la discrépance L^2 n'est pas un critère adapté pour apprécier la qualité de *recouvrement* uniforme d'une BDDE $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$. En fait, cela signifie d'une part, qu'il est n'est pas souhaitable de comparer les valeurs des discrépances L^2 pour des suites qui sont définies dans des espaces ayant des dimensions différentes, et d'autre part, que la définition de la propriété de *suite à discrépance faible* (qui rappelons-le est définie à l'aide de la discrépance à l'origine) devrait être définie, dans le cadre de la discrépance L^2, à l'aide d'un critère qui diminue lorsque la dimension de l'espace de la suite considérée augmente. Ainsi, plus la dimension de l'espace serait importante, plus la discrépance L^2 de la suite devrait être faible (selon un critère à définir), et par conséquent, plus le nombre de points devrait être important. En fait cette propriété illustre le fait que la discrépance L^2 est très délicate à interpréter.

1.3.3.3 Expressions, Discussion

Nous allons donner à présent certaines formules analytiques concernant les discrépances. Nous nous intéressons au cas de la discrépance à l'origine et au cas des discrépances de type L^2.

– Discrépance à l'origine

Le calcul de la discrépance à l'origine est « abordable » dans le cas $d = 1$ et $d = 2$. Nous indiquons ci-dessous sous forme de proposition une façon de la calculer lorsque la BDDE $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ est constituée de n points dans $\mathcal{X} = [0,1]^d$, avec $d = 1$ ou $d = 2$.

En dimension $d = 1$:

Proposition 1.3.2
Pour une suite $\boldsymbol{x}(n)$ de n points dans $\mathcal{X} = [0,1]$ avec $0 \leq x_1 \leq x_2 \leq \cdots \leq x_n \leq 1$,

$$\mathrm{D}^*(\boldsymbol{x}(n)) \quad = \quad \frac{1}{n} + \max_{1 \leq i \leq n} \left(\frac{i}{n} - x_i \right) - \min_{1 \leq i \leq n} \left(\frac{i}{n} - x_i \right).$$

Démonstration :
Nous nous référons à Niederreiter (1992).

□

En dimension $d = 2$:
- Soit $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$, une suite de n points dans $\mathcal{X} = [0,1]^2$ triée dans l'ordre croissant de leur première composante, i.e., telle que $0 \leq x_1^{(1)} \leq \cdots \leq x_n^{(1)} \leq 1$.
- Nous posons $x_0 = (0,0)$ et $x_{n+1} = (1,1)$, et nous notons à présent $\boldsymbol{x}(n) = \{x_0, x_1, \ldots, x_n, x_{n+1}\}$ la BDDE à laquelle nous avons ajouté les points x_0 et x_{n+1}.
- Pour $i \in \{0, \ldots, n\}$ nous notons $\{\xi_{i,0}, \xi_{i,1}, \ldots, \xi_{i,i+1}\}$ la suite des $i+1$ valeurs dans $[0,1]$ obtenue en réordonnant les i deuxièmes composantes de la suite $\boldsymbol{x}(n) = \{x_0, x_1, \ldots, x_n, x_{n+1}\}$ à laquelle nous ajoutons $\xi_i^{i+1} = 1$. C'est la suite obtenue en réordonnant les deuxièmes composantes de $\{x_0, x_1, \ldots, x_i, x_{n+1}\}$ c'est-à-dire, selon nos notations, en réordonnant : $\{x_0^{(2)}, x_1^{(2)}, \ldots, x_i^{(2)}, x_{n+1}^{(2)}\}$. Nous avons donc :

$$0 = \xi_{i,0} \leq \xi_{i,1} \leq \cdots \leq \xi_{i,i} \leq \xi_{i,i+1}.$$

Proposition 1.3.3
L'expression de la discrépance (ou discrépance à l'origine) est alors, à l'aide des notations précédentes,

$$\mathrm{D}^*(\boldsymbol{x}(n)) = \max_{0 \leq i \leq n} \max_{0 \leq k \leq n} \max \left\{ \left| \frac{k}{n} - x_i^{(1)} \xi_{i,k} \right|, \left| \frac{k}{n} - x_{i+1}^{(1)} \xi_{i,k+1} \right| \right\}.$$

Démonstration :
Nous pouvons trouver cette proposition dans Thiémard (2000), et une démonstration de celle-ci dans Zhu (1993) (voir également Thiémard (2000)).

□

Pour $d > 2$ le calcul est possible (voir Niederreiter (1972), Thiémard (2000)) mais devient beaucoup plus complexe. Le coût des algorithmes connus permettant de réaliser cette opération semble augmenter de manière exponentielle avec la dimension. Nous nous contentons le plus souvent de bornes inférieures et supérieures, voir Thiémard (2000).

– Discrépance de type L^2

Pour les discrépance de type L^2, il est possible d'obtenir des formules simples permettant leur calcul. Nous donnons ici leurs expressions sous forme de proposition.

Proposition 1.3.4
Pour une suite de n points $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ *dans* $\mathcal{X} = [0,1]^d$*, les expressions des discrépances de type* L^2 *sont données par*

- ***Discrépance*** L^2

$$[\mathrm{D}^{L^2}(\boldsymbol{x}(n))]^2 = 3^{-d} - \frac{2^{1-d}}{n}\sum_{i=1}^{n}\prod_{k=1}^{d}(1-(x_i^{(k)})^2) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{l=1}^{n}\prod_{j=1}^{d}[1-\max(x_i^{(k)}, x_j^{(k)})]; \tag{1.42}$$

- ***Discrépance*** L^2 ***modifiée***

$$\mathrm{DM}_n^{L^2}(\boldsymbol{x}(n)) = \left(\frac{4}{3}\right)^d - \frac{2^{1-s}}{n}\sum_{i=1}^{n}\prod_{k=1}^{d}[3-(x_i^{(k)})^2] + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\prod_{k=1}^{d}[2-\max(x_i^{(k)}, x_j^{(k)})]; \tag{1.43}$$

- ***Discrépance*** L^2 ***centrée***

$$\begin{aligned}\mathrm{DC}_n^{L^2}(\boldsymbol{x}(n)) &= \left(\frac{13}{12}\right)^d - \frac{2}{n}\sum_{i=1}^{n}\prod_{k=1}^{d}\left(1+\frac{1}{2}|1+x_i^{(k)}-1/2| - \frac{1}{2}|1+x_i^{(k)}-1/2|^2\right)\\ &\quad + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\prod_{k=1}^{d}\left(1+\frac{1}{2}|1+x_i^{(k)}-1/2| + \frac{1}{2}|1+x_j^{(k)}-1/2|\right.\\ &\qquad \left. - \frac{1}{2}|x_i^{(k)}-x_j^{(k)}|\right);\end{aligned} \tag{1.44}$$

- ***Discrépance*** L^2 ***symétrique***

$$\begin{aligned}\mathrm{DS}_n^{L^2}(\boldsymbol{x}(n)) &= \left(\frac{4}{3}\right)^d - \frac{2}{n}\sum_{i=1}^{n}\prod_{k=1}^{d}[1-2x_i^{(k)}-2(x_i^{(k)})^2]\\ &\quad + \frac{2^d}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\prod_{k=1}^{d}[1-|x_i^{(k)}-x_j^{(k)}|].\end{aligned} \tag{1.45}$$

Démonstration :
L'expression (1.42) a été obtenue par Warnock (1972), les expressions (1.43), (1.44) et (1.45) ont été obtenues par Hickernell (1998).

□

Essentiellement par pragmatisme, nous avons choisi les discrépances de type L^2 pour apprécier la qualité de *recouvrement* uniforme d'une BDDE $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ dans $\mathcal{X}$, puisque celles-ci sont facilement calculables. De plus les discrépances L^2 centrée et L^2 symétrique ne souffrent pas d'une dépendance à l'origine trop forte, inconvénient souvent formulé à l'encontre de la discrépance L^2 (voir Hickernell (1998), Thiémard (2000), Morokoff et Caflisch (1994)).

Une des questions que nous nous posons alors est de savoir quand ces différentes quantités nous indiquent que la BDDE est de « qualité acceptable », au sens où elle *recouvre* « uniformément » l'espace $\mathcal{X}$ d'une façon satisfaisante.

Définir une valeur « acceptable » de la discrépance pour une suite de points initiaux $\boldsymbol{x}(n)$ quelconque dans $\mathcal{X}$ semble irréaliste, puisque celle-ci dépend de la dimension de l'espace $\mathcal{X}$ et du nombre de points de la suite.

Pour répondre à cette question, nous allons comparer les différentes valeurs des discrépances de type L^2 de la BDDE que nous trouvons avec celles de *suites à discrépance faible*. Pour effectuer une telle comparaison, nous utiliserons des ratios de la forme :

$$\mathrm{RD}(\boldsymbol{x}(n)) := \frac{\text{Discrépance de type } L^2(\boldsymbol{x}(n))}{\text{Discrépance de type } L^2(\boldsymbol{x}_f(n))}\,, \tag{1.46}$$

où $\boldsymbol{x}_f(n)$ désigne une *suite à discrépance faible* (ayant même nombre de points n que la BDDE $\boldsymbol{x}(n)$), et « Discrépance de type L^2 » désigne les différentes discrépances de type L^2, D^{L^2}, DM^{L^2}, DC^{L^2} et DS^{L^2}. Pour une base de donnée $\boldsymbol{x}(n)$, nous aurons donc 4 ratios : $\mathrm{RD}^{\mathrm{L}^2}$, $\mathrm{RDM}^{\mathrm{L}^2}$, $\mathrm{RDC}^{\mathrm{L}^2}$, $\mathrm{RDS}^{\mathrm{L}^2}$. Une valeur proche de 1 signifiera donc que la discrépance de type L^2 considérée de la BDDE est comparable à celle d'une suite dont le *recouvrement* uniforme de l'espace est « acceptable ». Par conséquent, plus ces ratios seront proches de 1, plus la BDDE pourra être considérée comme uniformément répartie dans l'espace $\mathcal{X}$.

Nous comparerons aussi les valeurs calculées avec les discrépances carrées moyennes définies par les égalités (1.38), (1.39), (1.40) et (1.41). Bien que ces moyennes soient obtenues lorsque nous considérons une suite de *v.a. i.i.d.* de loi uniforme elles permettent d'avoir une idée sur la qualité du *recouvrement* uniforme de la BDDE dans l'espace $\mathcal{X}$.

1.3.4 Récapitulatif

Nous rappelons ici l'ensemble des critères que nous avons étudiés jusqu'à présent à l'aide du Tableau (1.2). Nous avons simplement ajouté ici aux critères du Tableau (1.1) de la partie 1.2 les critères associés à la notion de discrépance. N'ayant pas *de valeurs de références* pour les discrépances de types L^2, nous avons inscrit dans la colonne consacrée, une flèche vers le bas $\downarrow$ qui signifie que nous souhaitons que ces valeurs soient les plus faibles possibles, ainsi que les expressions des différentes moyennes des discrépances carrées (égalités (1.38), (1.39), (1.40) et (1.41)) dans le cas où la BDDE est considérée comme une suite de *v.a. i.i.d.* de loi uniforme.

CRITERE	VALEUR DE REFERENCE
$\mathrm{dmin}_\infty(\boldsymbol{x}(n))$	$\mathrm{Disp}_\infty(\boldsymbol{x}(n))/2$
$\gamma(\boldsymbol{x}(n))$	1
$m_{2,1}(\boldsymbol{x}(n))$	$\downarrow$
$\Lambda(\boldsymbol{x}(n))$	0
$\mathrm{Disp}_\infty(\boldsymbol{x}(n), \boldsymbol{x}_{Ha}(N))$	$2.\mathrm{dmin}_\infty(\boldsymbol{x}(n))$
$\mathrm{Disp}_\infty(\boldsymbol{x}(n), \boldsymbol{x}_{Ham}(N))$	$2.\mathrm{dmin}_\infty(\boldsymbol{x}(n))$
$\mathrm{Disp}_\infty(\boldsymbol{x}(n), \boldsymbol{x}_{Fa}(N))$	$2.\mathrm{dmin}_\infty(\boldsymbol{x}(n))$
$\mathrm{Disp}_\infty(\boldsymbol{x}(n), \boldsymbol{x}_{Res}(N))$	$2.\mathrm{dmin}_\infty$
$h(\boldsymbol{x}(n))$	$\downarrow$
$\mu(\boldsymbol{x}(n))$	1
$\chi(\boldsymbol{x}(n))$	$\downarrow$
$\nu(\boldsymbol{x}(n))$	1
$\tau(\boldsymbol{x}(n))$	0
$\Delta(\boldsymbol{x}(n))$	0
$(\mathrm{D}^{L^2}(\boldsymbol{x}(n)))^2$	$\downarrow$ et $\frac{(1/2)^d-(1/3)^d}{n}$
$\mathrm{RDiscL}^2(\boldsymbol{x}(n))$	1
$(\mathrm{DM}^{L^2}(\boldsymbol{x}(n)))^2$	$\downarrow$ et $\left(\frac{4}{3}\right)^d \frac{(9/8)^d-1}{n}$
$\mathrm{RDiscL}^2\mathrm{M}(\boldsymbol{x}(n))$	1
$(\mathrm{DC}^{L^2}(\boldsymbol{x}(n)))^2$	$\downarrow$ et $\frac{1}{n}\left[\left(\frac{13}{12}+\frac{1}{6}\right)^d - \left(\frac{13}{12}\right)^d\right]$
$\mathrm{RDiscL}^2\mathrm{C}(\boldsymbol{x}(n))$	1
$(\mathrm{DS}^{L^2}(\boldsymbol{x}(n)))^2$	$\downarrow$ et $\frac{1}{n}\left[\left(\frac{4}{3}+\frac{2}{6}\right)^d - \left(\frac{4}{3}\right)^d\right]$
$\mathrm{RDiscL}^2\mathrm{S}(\boldsymbol{x}(\mathrm{n}))$	1

TAB. 1.2 – Différents critères et valeurs de référence

Précisons que les critères que nous utilisons n'ont pas pour objectif la validation d'uniformité de suites mais l'appréciation de la répartition uniforme d'une BDDE ; celle-ci est en général quelconque. Si l'objectif est de valider l'uniformité de suites ayant de très bonnes propriétés, ces critères peuvent être insuffisants ; il en existe alors d'autres faisant parfois directement intervenir leurs propriétés de construction, voir Lemieux et L'Ecuyer (2001).

1.4 Utilisation, méthodologie

Les différents critères définis dans les parties précédentes vont être utilisés pour définir une méthodologie d'étude d'une base de données d'entrée (voir aussi Feuillard *et al.* (2005)). Celle-ci comporte 3 étapes :

1. Tout d'abord, évaluer la qualité de la base de données d'entrées. Notre objectif est de vérifier que la base de données d'entrées *recouvre* tout l'espace de façon « acceptable » (espacements réguliers entre les points, absence de trous). Il s'agit donc ici d'interpréter les valeurs des différents critères que nous avons définis.

2. Sélectionner certains points de la BDDE de façon à conserver un maximum d'information. L'information que nous considérons ici est la qualité de répartition uniforme de la BDDE. Il s'agit donc d'extraire de la BDDE, constituée initialement de la suite de points $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ dans $\mathcal{X} = [0,1]^d$, un sous-ensemble de points, $\boldsymbol{x}_1(n_1) = \{x_{i_1}, \ldots, x_{n_1}\} \subset \boldsymbol{x}(n)$, qui permette de recouvrir au mieux l'espace $\mathcal{X}$. Le critère d'uniformité que nous utiliserons pour cette sélection est la discrépance. Par sa définition même, celle-ci permet aussi d'apprécier la qualité des autres critères que nous avons définis.
3. Spécifier de nouveaux points de la BDDE. Après avoir sélectionné au mieux certains points de la BDDE, nous en spécifierons des nouveaux à l'aide *de suites à discrépance faible*, de façon à obtenir une BDDE de « qualité de répartition uniforme acceptable », si nécessaire.

Pour illustrer nos propos, nous considérons une base de données d'entrée comportant 400 points dans $\mathcal{X} = [0,1]^3$ et appliquons la méthodologie ci-dessus.

1.4.1 Etude d'une base de données d'entrée

La base de données d'entrée (400 points) considérée est représentée dans le graphique (1.8).

La première étape consiste à calculer les critères du Tableau (1.2). Malheureusement, certains d'entre eux n'ont pas de valeurs de référence permettant une appréciation relative

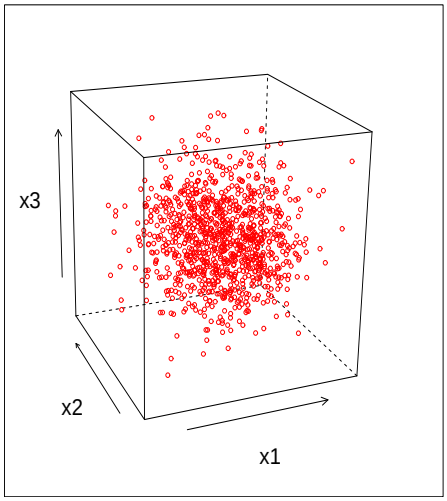


FIG. 1.8 – Base de données d'entrée étudiée (400 points)

des critères obtenus. Pour interpréter les résultats, nous allons aussi calculer ces critères pour des suites dont la « qualité de répartition uniforme » (au sens déterministe) est considérée comme « acceptable ». Ces suites $\boldsymbol{x}(n)$ « étalon » sont les *suites à discrépance faible* de Halton, $\boldsymbol{x}_{Ha}(n)$, de Hammersley, $\boldsymbol{x}_{Ham}(n)$, de Faure, $\boldsymbol{x}_{Fa}(n)$, et un réseau (*lattice*) de rang 1 de type $n\alpha$, $\boldsymbol{x}_{Res}(n)$ (un réseau de rang 1 est aussi un hypercube-latin), (voir Halton (1960), Hammersley (1960), Faure (1982), respectivement). Elles comportent le même nombre de points, $n = 400$. Les résultats sont présentés dans le Tableau (1.3).

CRITERE	$\boldsymbol{x}_{Ha}(n)$	$\boldsymbol{x}_{Ham}(n)$	$\boldsymbol{x}_{Fa}(n)$	$\boldsymbol{x}_{Res}(n)$	VALEUR DE REFERENCE
$\mathrm{dmin}_2(\boldsymbol{x}(n))$	0.0476	0.0415	0.0342	0.0931	
$\mathrm{dmin}_\infty(\boldsymbol{x}(n))$	0.0384	0.03	0.0304	0.0718	$\mathrm{Disp}_\infty(\boldsymbol{x}(\mathrm{n}))/2$
$\gamma(\boldsymbol{x}(n))$	3.72	3.72	4.86	1.88	1
$m_{2,1}(\boldsymbol{x}(n))$	3.200	3.192	3.212	3.195	
$\Lambda(\boldsymbol{x}(n))$	0.261	0.166	0.236	0.0620	0
$\mathrm{Disp}_2(\boldsymbol{x}(n)) \approx$	0.21	0.17	0.19	0.18	
$\mathrm{Disp}_\infty(\boldsymbol{x}(n)) \approx$	0.15	0.15	0.15	0.15	$2.\mathrm{dmin}_\infty$
$h(\boldsymbol{x}(n))$	0.203	0.169	0.204	0.183	$\downarrow$
$\mu(\boldsymbol{x}(n))$	2.164	1.737	2.248	1.883	1
$\chi(\boldsymbol{x}(n))$	6.117	6.186	8.422	3.939	$\downarrow$
$\nu(\boldsymbol{x}(n))$	9.367	2.853	5.990	6.564	1
$\tau(\boldsymbol{x}(n))$	0.00236	0.00210	0.003301	0.00222	0
$\Delta(\boldsymbol{x}(n))$	1.042E-8	1.340E-9	5.144E-9	2.992 E-9	0
$\mathrm{D}^{L^2}(\boldsymbol{x}(n))^2$	1.76E-5	8.19E-6	1.44E-5	1.24E-5	$\downarrow \quad \mathbb{E}(\mathrm{D}^{L^2}(\boldsymbol{x}(n))^2) = 0.00022$
$\mathrm{DM}^{L^2}(\boldsymbol{x}(n))^2$	1.19E-4	5.76E-5	5.80E-5	4.56E-5	$\downarrow \quad \mathbb{E}(\mathrm{DM}^{L^2}(\boldsymbol{x}(n))^2) = 0.0025$
$\mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2$	6.85E-5	4.25E-4	4.39E-5	3.62E-5	$\downarrow \quad \mathbb{E}(\mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2) = 0.0017$
$\mathrm{DS}^{L^2}(\boldsymbol{x}(n))^2$	7.49E-4	4.5E-4	8.65E-4	7.1E-4	$\downarrow \quad \mathbb{E}(\mathrm{DS}^{L^2}(\boldsymbol{x}(n))^2) = 0.014$

TAB. 1.3 – Etude de *suites à discrépance faible* en dimension 3 (400 points, $n = 400$)

Les critères du réseau faisant intervenir les distances entre les points de la suite considérée ($\mathrm{dmin}_\infty(\boldsymbol{x}_{Res}(n))$, $\gamma(\boldsymbol{x}_{Res}(n))$, $m_{2,1}(\boldsymbol{x}_{Res}(n))$, $\Lambda(\boldsymbol{x}_{Res}(n))$) sont les meilleurs. En effet ces critères sont les plus proches des valeurs de référence : $\mathrm{dmin}_\infty(\boldsymbol{x}_{Res}(n))$ est supérieure à celle des autres suites, ce qui signifie que les points les plus proches du réseau sont plus éloignés que ceux des autres suites (donc moins de « redondance de points »), $\gamma(\boldsymbol{x}_{Res}(n))$ est plus proche de 1, les espacements entre points sont donc plus réguliers, $\Lambda(\boldsymbol{x}_{Res}(n))$ est plus proche de 0, la variabilité des espacements est beaucoup plus faible. La régularité de la répartition des points du réseau est donc meilleure que celle des autres suites (c'est une propriété qui est bien entendu liée à sa définition). Si nous établissons un classement par ordre de préférence en considérant ces critères, nous avons : 1 Réseau, 2 Hammersley, 3 Faure, 4 Halton.

Considérons les critères faisant intervenir les points de la suite et les points de l'espace. Pour chaque suite, nous avons plusieurs approximations de la dispersion. Dans le cas d'une grille de répartition parfaitement uniforme (au sens déterministe) la dispersion est le double de la distance minimale entre deux points, voir Niederreiter (1992) et graphique (1.1). L'approximation de la dispersion du réseau semble vérifier cette propriété. Le rayon de la plus grande boule vide est donc comparable à la distance minimale entre deux points du réseau. Le critère $\chi(\boldsymbol{x}_{Res}(n))$, plus proche de 1, montre que le maximum des rayons des cellules de Voronoi est aussi comparable à la distance minimale entre deux points. Ces critères montrent que les points du réseau occupent l'espace de la meilleure façon.

Les critères faisant intervenir les volumes des régions de Voronoi donnent de meilleurs résultats avec la suite de Hammersley. Le critère $\mu(\boldsymbol{x}_{Ham}(n))$ de la suite de Hammersley montrait déjà que les rayons de ces régions étaient comparables ($\max h_i / \min h_i$ proche de 1). Les différentes discrépances indiquent quant à elles que le réseau semble de meilleure qualité. En général, dans un pavé de $[0,1]^3$, le nombre de points comparativement à son volume, semble meilleur pour le réseau. Remarquons que l'interprétation de ce critère correspond à la réalisation d'un compromis entre les critères des distances entre points et les critères des distances entre points et espace. En effet, une distance (moyenne ou minimale) faible entre deux points et une région vide de l'espace trop importante (dispersion) impliquent une mauvaise relation entre le nombre de points et les volumes d'intervalles du cube unité et par conséquent une dispersion élevée.

Les différents critères montrent que le réseau a une meilleure répartition uniforme dans l'espace.

Ayant quelques valeurs de références, nous allons maintenant nous intéresser à la base de données d'entrée expérimentale fournie à l'appui de notre étude. Les résultats sont présentés Tableau (1.4).

Par comparaison aux *suites à discrépance faible* et aux valeurs de référence, cette

CRITERE	VALEUR	VALEUR DE REFERENCE
$\mathrm{dmin}_2(\boldsymbol{x}(n))$	0.00338	
$\mathrm{dmin}_\infty(\boldsymbol{x}(n))$	0.00241	$\mathrm{Disp}_\infty(\boldsymbol{x}(\mathrm{n}))/2$
$\gamma(\boldsymbol{x}(n))$	21.236	1
$m_{2,1}(\boldsymbol{x}(n))$	6.112	
$\Lambda(\boldsymbol{x}(n))$	0.553	0
$\mathrm{Disp}_\infty(\boldsymbol{x}(n), \boldsymbol{x}_{Ha}(1000))$	0.576	$2.\mathrm{dmin}_\infty(\boldsymbol{x}(n))$
$\mathrm{Disp}_\infty(\boldsymbol{x}(n), \boldsymbol{x}_{Ham}(1000))$	0.0.564	$2.\mathrm{dmin}_\infty(\boldsymbol{x}(n))$
$\mathrm{Disp}_\infty(\boldsymbol{x}(n), \boldsymbol{x}_{Fa}(1000))$	0.587	$2.\mathrm{dmin}_\infty(\boldsymbol{x}(n))$
$\mathrm{Disp}_\infty(\boldsymbol{x}(n), \boldsymbol{x}_{Res}(1000))$	0.546	$2.\mathrm{dmin}_\infty(\boldsymbol{x}(n))$
$h(\boldsymbol{x}(n))$	0.574308	$\downarrow$
$\mu(\boldsymbol{x}(n))$	17.58	1
$\chi(\boldsymbol{x}(n))$	36.08	$\downarrow$
$\nu(\boldsymbol{x}(n))$	1670.67	1
$\tau(\boldsymbol{x}(n))$	0.0447	0
$\Delta(\boldsymbol{x}(n))$	6.95E-7	0
$\mathrm{D}^{L^2}(\boldsymbol{x}(n))^2$	0.018	$\downarrow \quad \mathbb{E}(\mathrm{D}^{L^2}(\boldsymbol{x}(n))^2) = 0.00022$
$\mathrm{RD}^{\mathrm{L}^2}(\boldsymbol{x}(\mathrm{n}))$	40→46	1
$\mathrm{DM}^{L^2}(\boldsymbol{x}(n))^2$	0.18	$\downarrow \quad \mathbb{E}(\mathrm{DM}^{L^2}(\boldsymbol{x}(n))^2) = 0.0025$
$\mathrm{RDM}^{\mathrm{L}^2}(\boldsymbol{x}(\mathrm{n}))$	39→62	1
$\mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2$	0.094	$\downarrow \quad \mathbb{E}(\mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2) = 0.0017$
$\mathrm{RDC}^{\mathrm{L}^2}(\boldsymbol{x}(\mathrm{n}))$	37→51	1
$\mathrm{DS}^{L^2}(\boldsymbol{x}(n))^2$	0.91	$\downarrow \quad \mathbb{E}(\mathrm{DS}^{L^2}(\boldsymbol{x}(n))^2) = 0.014$
$\mathrm{RDS}^{\mathrm{L}^2}(\boldsymbol{x}(\mathrm{n}))$	31→47	1

TAB. 1.4 – Etude d'une base de données d'entrée à 3 dimensions (400 points, $n = 400$)

base de données d'entrée n'a pas une « bonne répartition uniforme ». Aucun critère n'a de valeur pouvant être qualifiée « d'acceptable ».

Nous allons à présent sélectionner certains points de façon à construire une base ayant de meilleures propriétés d'uniformité (selon nos critères).

1.4.2 Sélection de points

L'objectif est d'extraire parmi les points de la BDDE initiale, $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ avec $x_i \in \mathcal{X} = [0,1]^d$, un sous-ensemble $\boldsymbol{x}_1(n_1) = \{x_{i_1}, \ldots, x_{n_1}\} \subset \boldsymbol{x}(n)$ dont la répartition est uniforme. Nous allons donc sélectionner des points de façon à réduire la valeur d'un critère d'uniformité. Ce critère est la discrépance L^2 centrée, DC^{L^2} (voir définition (1.3.5)). Par sa définition même, il est adapté à l'objectif fixé puisqu'il correspond à une comparaison entre le nombre de points compris dans certains pavés de l'espace et le volume de ces pavés. Nous prenons donc en compte le nombre de points de la suite. La qualité de répartition uniforme est évaluée comparativement aux nombres de points de la suite. De plus, c'est un critère simple à calculer (voir la formule (1.44)) et donc d'utilisation très aisée.

Pour qualifier un sous-ensemble $\boldsymbol{x}_1(n_1)$ de points de la BDDE de qualité « acceptable », nous comparerons la discrépance carrée centrée, $\mathrm{DC}^{L^2}(\boldsymbol{x}_1(n_1))^2$, de cet ensemble à son espérance, $\mathbb{E}\left[\left(\mathrm{DC}^{L^2}(\boldsymbol{x}_1(n_1))\right)^2\right]$, donnée par la formule (1.40). Bien que nous ne considérons pas notre BDDE comme une suite de v.a i.i.d. de loi uniforme, cette valeur constituera une valeur seuil de référence. Lorsqu'un sous-ensemble de points de la BDDE, $\boldsymbol{x}_1(n_1)$, aura une discrépance carrée inférieure à cette valeur seuil :

$$\left(\mathrm{DC}^{L^2}(\boldsymbol{x}_1(n_1))\right)^2 < \mathbb{E}\left[\left(\mathrm{DC}^{L^2}(\boldsymbol{x}_1(n_1))\right)^2\right], \tag{1.47}$$

il sera jugé de qualité « acceptable ».

1.4.2.1 Méthode 1

La première méthode consiste tout simplement à trouver un sous-ensemble $\boldsymbol{x}_1(n_1)$ de $\boldsymbol{x}(n)$ dont $\mathrm{DC}^{L^2}(\boldsymbol{x}(n_1))$ est minimale. L'algorithme A_1 utilisé est le suivant :

Algorithme A_1

Étape 1 : calcul de $\mathrm{D}_{\mathrm{iter}_0} = \mathrm{DC}^{L^2}(\boldsymbol{x}(n))$;

Étape 2 : Pour $i = 1, \ldots, n$,
- Calcul de $\mathrm{D}_{\mathrm{iter}_1,\mathrm{i}} = \mathrm{DC}^{L^2}(\boldsymbol{x}_{-i}(n-1))$ où $\boldsymbol{x}_{-i}(n-1) = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$,
- Calcul de $\mathrm{diff}_\mathrm{i} = \mathrm{D}_{\mathrm{iter}_1,\mathrm{i}} - \mathrm{D}_{\mathrm{iter}_0}$;

Étape 3 : Si $\mathrm{diff}_{\mathrm{i}^*} = \{\min_{i=1,\ldots,n} \mathrm{diff}_\mathrm{i}\} < 0$,
- sélection de l'ensemble $\boldsymbol{x}_{-i^*}(n-1)$,
- itération (retour à l'Étape 1) en posant : $\boldsymbol{x}(n) \leftarrow \boldsymbol{x}_{-i^*}(n-1))$, $D_{iter_0} \leftarrow D_{iter_1,i}$;

Étape 4 : Arrêt lorsque $D_{iter_1,i^*} = \min_{i=1,\ldots,n} D_{iter_1,i} > 0$.

Cet algorithme ne permet pas forcément de trouver le sous-ensemble optimal, puisqu'on itére en considérant uniquement la différence $\mathrm{diff}_\mathrm{i} = \mathrm{D}_{\mathrm{iter}_1,\mathrm{i}} - \mathrm{D}_{\mathrm{iter}_0}$. Cependant, ce sous-ensemble peut être qualifié « $\mathrm{DC}^{L^2}(\boldsymbol{x}(n))$ - irréductible », puisqu'il y a arrêt de l'algorithme lorsqu'aucune suppression de points ne permet de réduire cette quantité.

Remarquons que, lors de l'itération l'ensemble de points $\boldsymbol{x}_r(n_r) = \{x_i : \mathrm{diff}_\mathrm{i} < 0\} \subset \boldsymbol{x}(n)$ constitue un ensemble de points qui ne contribue pas à une répartition uniforme de l'ensemble $\boldsymbol{x}(n)$. En effet la suppression d'un de ces points permet de diminuer le critère DC^{L^2}, et donc la discrépance, critère de répartition uniforme. L'ensemble $\boldsymbol{x}_r(n_r) = \{x_i : \mathrm{diff}_\mathrm{i} < 0\} \subset \boldsymbol{x}(n)$ peut donc être qualifié d'ensemble de points redondants. Un exemple de points redondants d'une suite de 100 points dans un espace $\mathcal{X} = [0,1]^2$ est illustré Figure (1.9). L'ensemble $\boldsymbol{x}_r(n_r) = \{x_i : \mathrm{diff}_\mathrm{i} < 0\} \subset \boldsymbol{x}(n)$ n'est pas pour autant un ensemble redondant, car seule la suppression d'un unique point de l'ensemble permet de réduire la discrépance (les points de l'ensemble peuvent être considérés comme redondants, mais pas l'ensemble). Il n'y a aucune garantie pour que la suppression de l'ensemble $\boldsymbol{x}_r(n_r)$ permette de réduire DC^{L^2}. Pour ce faire, il faudrait vérifier que :

$$\mathrm{diff}_{\mathrm{i}_1,\ldots,\mathrm{i}_{\mathrm{n_r}}} = \mathrm{DC}^{L^2}(\boldsymbol{x}(n)) - \mathrm{DC}^{L^2}(\boldsymbol{x}(n) \backslash \boldsymbol{x}_r(n_r)) < 0.$$

L'algorithme A_1 est réalisable car le nombre de points ($n = 400$) et la dimension de l'espace ($d = 3$) considérés pour l'application sont relativement faibles. Le temps de calcul pour obtenir ces résultats est ici de quelques minutes (sur un PC[4]). Lorsque le nombre de points et la dimension sont plus élevés, l'exécution de cet algorithme est plus coûteuse en temps de calcul. Il est donc nécessaire de le modifier quelque peu. Par exemple, à l'étape 2, une méthode consiste à remplacer $\mathrm{D}_{\mathrm{iter}_1,\mathrm{i}}$ par $\mathrm{D}_{\mathrm{iter}_1,\boldsymbol{i}_\mathrm{k}} = \mathrm{DC}^{L^2}(\boldsymbol{x}_{-\boldsymbol{i}_k}(n-k))$ où $\boldsymbol{x}_{-\boldsymbol{i}_k}(n-k) = \{x_1, \ldots, x_n\} \backslash \{x_{i_1}, \ldots, x_{i_k}\}$, pour un ensemble d'indices $\boldsymbol{i}_k = \{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$, puis à arrêter la boucle dès que $\mathrm{diff}_{\boldsymbol{i}_\mathrm{k}} = \mathrm{D}_{\mathrm{iter}_1,\boldsymbol{i}_\mathrm{k}} - \mathrm{D}_{\mathrm{iter}_0} < 0$.

[4]Précisons que ce programme a été réalisé avec le langage interprété R. L'exécution serait encore plus rapide avec un langage bas niveau.

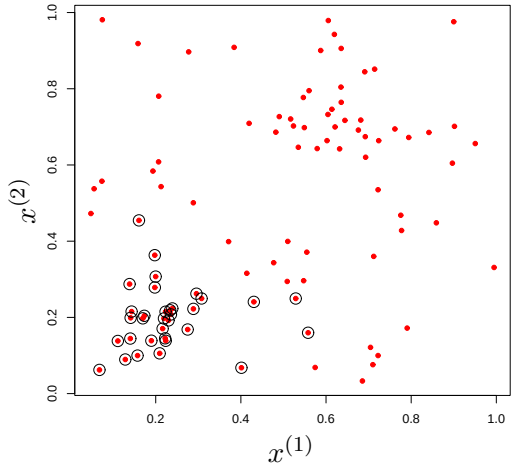


FIG. 1.9 – Exemple d'un ensemble de points redondants (points encerclés) pour une suite de 100 points dans $[0,1]^2$

1.4.2.2 Méthode 2

La seconde méthode consiste simplement à sélectionner des points proches de ceux d'une suite à discrépance faible. Ces suites « déterministes » sont essentiellement utilisées pour les méthodes de quasi-Monte Carlo et sont construites dans l'objectif de réduire au mieux la discrépance à l'origine (voir définition (1.3.7)). L'algorithme A_2 est le suivant :

Algorithme A_2

Étape 1 : Construction d'une suite à discrépance faible dans $\mathcal{X}$ comportant n_f points, $\boldsymbol{x}_f(n_f) = \{x_{f_1}, \ldots, x_{f_{n_f}}\}$.

Étape 2 : Pour $i = 1, \ldots, n_f$,
- sélection de $x_{i_1} \in \boldsymbol{x}(n)$ le plus proche du point $x_{f_i} \in \boldsymbol{x}_f(n_f)$,
- $\boldsymbol{x}(n-1) = \boldsymbol{x}(n) \setminus x_{i_1}$,
- $\boldsymbol{x}_f(n_f - 1) = \boldsymbol{x}_f(n_f) \setminus x_{f_i}$.

Soit $\boldsymbol{x}(n_f) \subset \boldsymbol{x}(n)$ la suite obtenue.
Calcul de $\mathrm{DC}^{L^2}(\boldsymbol{x}(n_f))$.

Étape 3 : Construction d'une suite à discrépance faible comportant $n_{f'}$ points, $\boldsymbol{x}_{f'}(n_{f'}) = \{x_{f'_1}, \ldots, x_{f'_{n_{f'}}}\}$ avec $n_{f'} < n_f$.
Application de l'Étape 2.

L'avantage de cette méthode simple est qu'elle permet d'obtenir immédiatement un ensemble ayant un nombre de points souhaité (à l'aide des étapes 1 et 2 de l'algorithme) sans réaliser de calculs répétés de la discrépance.

1.4.2.3 Méthode 3

L'objectif de la troisième méthode est d'extraire un sous-ensemble de points régulièrement répartis dans $\mathcal{X} = [0,1]^3$. L'algorithme A_3 est le suivant :

Algorithme A_3

Étape 1 : Soit dist $= \{\varepsilon_1, \ldots, \varepsilon_{end}\}$ un ensemble de valeurs croissantes dans $(0,1)$. Sélection du point $x_i^* \in \boldsymbol{x}(n)$ le plus proche (au sens de la norme euclidienne) du « centre » du domaine (i.e. du point $(0.5, 0.5)$).
- $\boldsymbol{x}_{iter_1}(1) = \{x_i^*\}$,
- $\boldsymbol{x}_{iter_2}(n-1) = \boldsymbol{x}(n) \setminus \{x_i^*\}$.

Étape 2 : Soient $\boldsymbol{x}_r(n_r) = \{x_{i_1}, \ldots, x_{i_{n_r}}\} \subset \boldsymbol{x}_{iter_2}(n-1)$ les points inclus dans la boule de centre x_i^* et de rayon ε_1 : $\boldsymbol{x}_r(n_r) = \boldsymbol{x}_{iter_2}(n-1) \cap B(x_i^*, \varepsilon_1)$.
- $\boldsymbol{x}_{iter_2}(n-n_r-1) = \boldsymbol{x}_{iter_2}(n-1) \setminus \boldsymbol{x}_r(n_r)$.

Étape 3 : Soit $x_{i'} \in \boldsymbol{x}_{iter_2}(n-n_r-1)$ le point le plus proche de $\boldsymbol{x}_{iter_1}(1)$ (réalisant le minimum des distances entre les points de $\boldsymbol{x}_2(n-1-n_r)$ et $\boldsymbol{x}_{iter}(1)$).
- $\boldsymbol{x}_{iter_1}(2) = \boldsymbol{x}_{iter}(1) \cup \{x_{i'}\}$
- $\boldsymbol{x}_{iter_2}(n-n_r-2) = \boldsymbol{x}_{iter_2}(n-n_r-1) \setminus \{x_{i'}\}$.

Étape 4 : Soient $\boldsymbol{x}_{r_2}(n_{r_2}) \in \boldsymbol{x}_{iter_2}(n-n_r-2)$ les points inclus dans la boule de centre $x_{i'}$ et de rayon ε_1 : $\boldsymbol{x}_{r_2}(n_{r_2}) = \boldsymbol{x}_{iter_2}(n-n_r-2) \cap B(x_{i'}, \varepsilon_1)$.
- $\boldsymbol{x}_{iter_2}(n-n_r-2) = \boldsymbol{x}_{iter_2}(n-n_r-n_{r_2}-2) \setminus \boldsymbol{x}_{r_2}(n_{r_2})\}$

Étape 5 : Itération des Étapes 3 et 4 tant que $\boldsymbol{x}_{iter_2} \neq \emptyset$.
On note $\boldsymbol{x}_{\varepsilon_1}(n_{\varepsilon_1})$ l'ensemble $\boldsymbol{x}_{iter_1}(n_1)$ obtenu.

Étape 6 : Calcul de $\mathrm{DC}^{L^2}(\boldsymbol{x}_{\varepsilon_1}(n_{\varepsilon_1}))$. Retour à l'Étape 1 avec $\varepsilon_2 \in$ dist .

Étape 7 : Sélection de $\boldsymbol{x}_1(n_1) = \boldsymbol{x}_\varepsilon(n_\varepsilon) = \arg_{\varepsilon_i \in \mathrm{dist}} \left\{ \min \mathrm{DC}^{L^2}(\boldsymbol{x}_{\varepsilon_i}(n_{\varepsilon_i})) \right\}$.

Comme la discrépance est une comparaison entre le nombre de points de $\boldsymbol{x}(n)$ compris dans certains pavés de $\mathcal{X} = [0,1]^3$ et le volume de ces pavés, nous pouvons raisonnablement penser qu'un ensemble de points régulièrement espacés permettra d'avoir une discrépance faible. C'est par exemple le cas des réseaux. L'objectif de l'algorithme A_3 est

donc de sélectionner des points régulièrement espacés, tous distants d'au moins $\varepsilon > 0$, et qui permettent de recouvrir au mieux l'espace $\mathcal{X} = [0,1]^3$ puisque nous cherchons à diminuer le critère DC^{L^2}.

Notons que cet algorithme peut aussi être modifié en remplaçant à l'étape 3 : $x_{i'} \in \boldsymbol{x}_{iter_2}(n-n_r-1)$ le point le plus proche de $\boldsymbol{x}_{iter_1}(1)$, par $x_{i'} \in \boldsymbol{x}_{iter_2}(n-n_r-1)$ le point le plus *éloigné* de $\boldsymbol{x}_{iter_1}(1)$ (réalisant le *maximum* des distances euclidiennes entre les points de $\boldsymbol{x}_2(n-1-n_r)$ et $\boldsymbol{x}_{iter}(1)$). L'ensemble de points alors retenu à l'aide de l'algorithme A_3 modifié comporte des caractéristiques similaires à celui retenu par l'algorithme A_3 initial.

En général cet algorithme permet bien de trouver un sous-ensemble de points de qualité acceptable. Le critère DC^{L^2} diminue cependant moins rapidement et moins fortement que lors de l'application de l'algorithme A_1 .

1.4.3 Application des méthodes de sélection

Lors de l'application des algorithmes A_1, A_2, A_3, l'évolution de la discrépance L^2 centrée carrée en fonction du nombre de points est représentée Figure (1.10), (1.11), et (1.12). La courbe verte représente l'évolution de l'espérance de la discrépance L^2 centrée carrée d'une suite de points ayant une loi de probabilité uniforme dans $\mathcal{X} = [0,1]^3$. Comme attendu (de par la construction des algorithmes), la méthode 1 est la plus efficace pour la sélection d'un ensemble de points de discrépance faible. L'algorithme A_3 converge plus difficilement. Rappelons que ce dernier n'a pas pour objectif la sélection d'un ensemble de points de discrépance minimale, mais la sélection de points régulièrement espacés. Les points sélectionnés doivent tous être distants d'au moins une distance ε. L'évolution de la discrépance en fonction de cette distance est représentée Figure (1.13). La distance ε retenue est $\varepsilon = 0.285$.

Les suites de points sélectionnées, $x_{A_1}(n_{A_1})$, $x_{A_2}(n_{A_2})$, $x_{A_3}(n_{A_3})$, comportent respectivement $n_{A_1} = 20$, $n_{A_2} = 25$, et $n_{A_3} = 18$ points. Le fait que le nombre de points de ces suites est faible par rapport au nombre de points de la suite initiale ($n = 400$) indique que beaucoup de points de la suite initiale étaient « redondants » dans le sens où il ne permettaient pas de contribuer à une meilleure répartition uniforme.

L'ensemble des valeurs des critères du Tableau (1.2). appliqués à $x_{A_1}(n_{A_1})$, $x_{A_2}(n_{A_2})$, $x_{A_3}(n_{A_3})$, sont présentées Tableau (1.5).

Les critères de distance entre les points de la suite considérée (γ, $m_{2,1}$, Λ) montrent que les points de $x_{A_3}(n_{A_3})$ sont répartis de façon plus régulière que ceux de $x_{A_2}(n_{A_2})$ et $x_{A_3}(n_{A_3})$ (c'est l'objectif poursuivi par A_3).

Les critères de distance entre les points de la suite considérée et les points de l'espace $\mathcal{X} = [0,1]^3$, montrent que les différentes suites $x_{A_1}(n_{A_1})$, $x_{A_2}(n_{A_2})$, $x_{A_3}(n_{A_3})$ recouvrent

l'espace $\mathcal{X}$ de façon comparable. Le critère χ, plus important pour $x_{A_1}(n_{A_1})$, nous indique tout simplement qu'il existe dans cette suite des points relativement proches (en comparaison avec le rayon de la cellule de Voronoi associé à l'un de ces points, voir équation (1.19)).

Les critères faisant intervenir des volumes sont aussi semblables. Le critère ν (rapport du maximum par le minimum des cellules de Voroni, voir équation (1.20)) indique que les espacements des points de $x_{A_1}(n_{A_1})$ sont moins réguliers que ceux $x_{A_2}(n_{A_2})$ et $x_{A_3}(n_{A_3})$. les valeurs des discrépances semblent montrer que la répartition uniforme de $x_{A_1}(n_{A_1})$ est la meilleure.

On vérifie bien que les discrépances carrées des suites sélectionnées sont inférieures à leurs espérances (lorsqu'il s'agit de suite ayant une loi de probabilité uniforme dans $\mathcal{X} = [0,1]^3$). Ainsi, les suites de points sélectionnées ont toute une répartition uniforme « acceptable » selon ce critère.

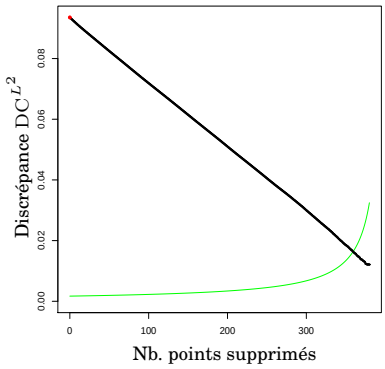


FIG. 1.10 – Évolution de la discrépance carrée centrée par application de l'algorithme A_1

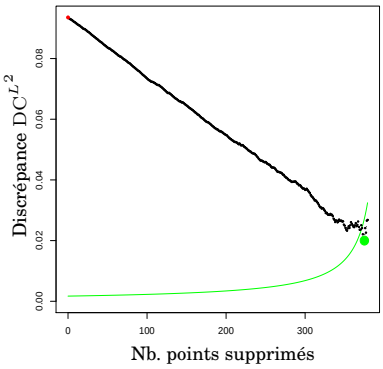


FIG. 1.11 – Évolution de la discrépance carrée centrée par application de l'algorithme A_2

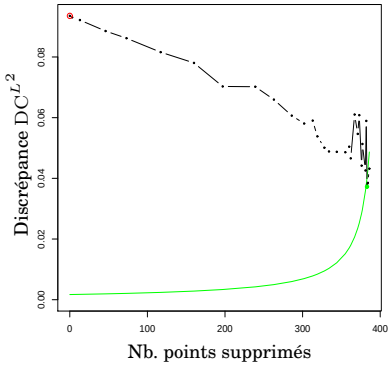


FIG. 1.12 – Évolution de la discrépance carrée centrée par application de l'algorithme A_3

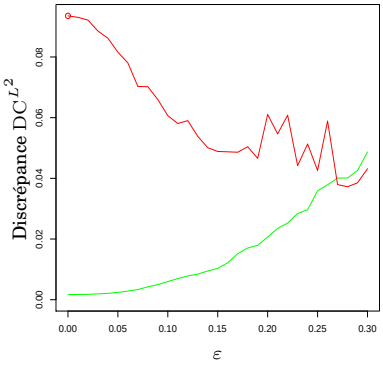


FIG. 1.13 – Évolution de la discrépance carrée centrée par application de l'algorithme A_3 en fonction de ε

	$\boldsymbol{x}_{ini}(n_{ini})$	$\boldsymbol{x}_{A_1}(n_{A_1})$	$\boldsymbol{x}_{A_2}(n_{A_2})$	$\boldsymbol{x}_{A_3}(n_{A_3})$	$\boldsymbol{x}_{B_1}(n_{B_1})$	$\boldsymbol{x}_{B_2}(n_{B_2})$	$\boldsymbol{x}_{B_3}(n_{B_3})$
n	400	20	25	18	43	40	43
$\mathrm{dmin}_2(\boldsymbol{x}(n))$	0.00338	0.072	0.141	0.285	0.072	0.12	0.285
$\mathrm{dmin}_\infty(\boldsymbol{x}(n))$	0.002	0.006	0.11	0.17	0.06	0.09	0.17
$\gamma(\boldsymbol{x}(n))$	21.23	6.04	2.08	1.22	4.40	2.88	1.15
$m_{2,1}(\boldsymbol{x}(n))$	12.22	6.71	6.88	7.15	6.03	6.44	5.52
$\Lambda(\boldsymbol{x}(n))$	0.553	0.472	0.226	0.061	0.3113	0.289	0.029
$\mathrm{Disp}_2(\boldsymbol{x}(n))$	0.59	0.59	0.60	0.61	0.39	0.50	0.36
$\mathrm{Disp}_\infty(\boldsymbol{x}(n))$	0.45	0.47	0.45	0.46	0.32	0.43	0.29
$h(\boldsymbol{x}(n))$	0.574	0.578	0.574	0.603	0.361	0.487	0.318
$\mu(\boldsymbol{x}(n))$	17.58	1.74	1.89	2.18	1.55	2.09	1.43
$\chi(\boldsymbol{x}(n))$	36.1	15.1	7.0	4.14	8.30	6.27	2.22
$\nu(\boldsymbol{x}(n))$	1670	5.7	3.36	3.95	3.83	4.48	3.58
$\tau(\boldsymbol{x}(n))$	0.047	0.024	0.016	0.022	0.0092	0.015	0.0091
$\Delta(\boldsymbol{x}(n))$	7E-7	8.5E-07	9.7e-07	9.42e-07	1.88e-07	8.32e-07	1.40e-07
$\mathrm{D}^{L^2}(\boldsymbol{x}(n))^2$	0.018	0.0018	0.0037	0.0059	0.00052	0.00089	0.00063
$\mathrm{DM}^{L^2}(\boldsymbol{x}(n))^2$	0.18	0.015	0.032	0.060	0.002	0.010	0.0075
$\mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2$	0.094	0.012	0.020	0.035	0.0016	0.0079	0.0063
$\mathrm{DS}^{L^2}(\boldsymbol{x}(n))^2$	0.91	0.14	0.14	0.28	0.025	0.058	0.053

TAB. 1.5 – Étude des suites obtenues à l'aide des algorithmes A_1, A_2, A_3, B_1, B_2, B_3

1.4.4 Spécification de points

Selon la méthodologie que nous avons définie au paragraphe (1.4), lorsque la base de données d'entrée, $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ dans $\mathcal{X} = [0,1]^d$, n'est pas jugée de « qualité acceptable » et lorsqu'il est possible d'obtenir des résultats (expériences et/ou simulations) supplémentaires, nous spécifierons de nouveaux points. Nous supposerons alors disposer d'une suite à discrépance faible $\boldsymbol{x}_f(N_f) = \{x_{f_1}, \ldots, x_{f_{Nf}}\}$ dont la répartition uniforme dans l'espace $\mathcal{X}$ est jugée de bonne qualité. Les points spécifiés seront convenablement choisis à l'aide de cette suite. Sur le modèle des algorithmes A1, A2, et A3, définis au paragraphe précédent, nous proposons différentes méthodes. L'objectif de ces méthodes sera, cette fois-ci, *d'augmenter* le nombre de points en cherchant à réduire la discrépance (critère de répartition uniforme, définition (1.3.5)).

Précisons que le nombre de points nécessaire pour qu'une suite à discrépance faible puisse être considérée comme telle croît de façon exponentielle avec la dimension (voir paragraphe (1.3.3)). Pour cette raison, ces méthodes deviennent délicates à appliquer (stockage des données, temps de calcul) lorsque la dimension d de $\mathcal{X} = [0,1]^d$ est supérieure à 9.

1.4.4.1 Méthode 1

Selon les notations introduites au paragraphe (1.3.1) et la définition (1.3.5), nous notons $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ la BDDE dans $\mathcal{X} = [0,1]^d$, et DC^{L^2}, la discrépance L^2 centrée. La nomenclature : $\#(\boldsymbol{x}(n))$ désigne le nombre d'éléments de la suite $\boldsymbol{x}(n)$, et $\boldsymbol{x}_f(N_f)$ une suite à discrépance faible comportant N_f points dans $\mathcal{X} = [0,1]^d$.

Nous proposons l'algorithme suivant :

Algorithme B_1

Étape 1 : Soit n_{sup} un nombre fixé, on pose : $\boldsymbol{x}_{ini}(n_{ini}) = \boldsymbol{x}(n)$

Étape 2 : Pour $i = 1, \ldots, N_f$,
- $\boldsymbol{x}(n+1) = \{x_1, \ldots, x_n\} \cup x_{f_i}$;
- calcul de $\mathrm{Dif}_i = \mathrm{DC}^{L^2}(\boldsymbol{x}(n+1)) - \mathrm{DC}^{L^2}(\boldsymbol{x}(n))$;

Étape 3 : sélection de i^* tel que $\mathrm{Dif}_{i^*} = \min_{i=1,\ldots,N_f} \mathrm{Dif}_i$,
- on pose $\boldsymbol{x}_{iter}(n_{iter}) = \{x_1, \ldots, x_n\} \cup x_{f_{i^*}}$;
- $\boldsymbol{x}_f(N_f - 1) = \{x_{f_1}, \ldots, x_{f_{Nf}}\} \backslash x_{f_{i^*}}$

Étape 4 : itération (retour à l'Étape 2) avec $\boldsymbol{x}(n) = \boldsymbol{x}(n_{iter})$ et arrêt lorsque : $\#(\boldsymbol{x}(n)) = n_{ini} + n_{supp}$

Cet algorithme consiste simplement à ajouter à la suite de points initiale, $\boldsymbol{x}(n)$, les points d'une suite à discrépance faible, $\boldsymbol{x}_f(N_f)$, de manière à ce que la différence entre la

discrépance L^2 centrée de la suite de points initiale, $\boldsymbol{x}(n)$, et la discrépance de la suite à laquelle on a ajouté ces points soit la plus faible possible. Ceci se fait ici point par point : on ajoute un unique point à chaque itération. Il est bien entendu possible de modifier l'algorithme B1 de façon à ajouter plusieurs points à chaque itération.

Le nombre N_f de points de la suite à discrépance faible $\boldsymbol{x}_f(N_f)$ dépend de la dimension et aussi du nombre de points n_{supp} que l'on veut spécifier. Plus la dimension d de l'espace $\mathcal{X} = [0,1]^d$ est grande, plus le nombre de points N_f de $\boldsymbol{x}_f(N_f)$ sera important. De même plus le nombre de points n_{supp} que l'on souhaite spécifier est important, plus N_f devra être grand.

1.4.4.2 Méthode 2

Nous reprenons ici les mêmes notations que précédemment (paragraphe (1.4.4.1)). Nous présentons une méthode analogue à celle présentée au paragraphe (1.4.2.2) (algorithme A_2). L'objectif de cette méthode est de « fusionner » directement la suite de points initiale avec une suite à discrépance faible ayant un nombre de points plus important.

L'algorithme est le suivant :

Algorithme B_2

Étape 1 : Soit n_{sup} un nombre fixé, on pose : $\boldsymbol{x}_{ini}(n_{ini}) = \boldsymbol{x}(n)$;

Étape 2 : On désigne par $\boldsymbol{x}_f(N_f)$ une suite à discrépance faible, avec $N_f = n + 1$;

Étape 3 : pour $i = 1, \ldots, n$,
- $x_{f_{i^*}} = \arg\min_{\{j=1,\ldots,N_f\}} \|x_{f_j} - x_i\|$;

Étape 4 : on pose :
- $\boldsymbol{x}_f(N_f - n_r) = \{x_{f_1}, \ldots, x_{f_{Nf}}\} \backslash \{x_{f_{i_1^*}}, \ldots, x_{f_{i_{n_r}^*}}\}$;
- $\boldsymbol{x}_{iter}(n_{iter}) = \boldsymbol{x}(n) \cup \boldsymbol{x}_f(N_f - n_r)$ et calcul de $\mathrm{DC}^{L^2}(\boldsymbol{x}_{iter}(n_{iter}))$;

Étape 5 : itération (retour à l'Étape 2) avec $\boldsymbol{x}(n) = \boldsymbol{x}_{iter}(n_{iter})$, ou $\boldsymbol{x}(n) = \boldsymbol{x}_{ini}(n_{ini})$, et arrêt lorsque : $\#(\boldsymbol{x}(n)) \geq n_{ini} + n_{supp}$.

Il s'agit donc dans un premier temps de considérer une suite à discrépance faible, $\boldsymbol{x}_f(N_f)$, de même cardinalité que la suite $\boldsymbol{x}(n)$ augmentée de 1, $\#(\boldsymbol{x}_f(N_f)) = n + 1$, et de remplacer les points de la suite $\boldsymbol{x}_f(N_f)$ les plus proches de la suite $\boldsymbol{x}(n)$ par ceux de $\boldsymbol{x}(n)$. Le nombre de points de la suite obtenue $\boldsymbol{x}_{iter}(n_{iter})$ n'est pas nécessairement égal à $n + 1$, puisqu'un point de la suite $\boldsymbol{x}_f(N_f)$ peut être à la fois le plus proche de plusieurs points de $\boldsymbol{x}(n)$. On itère ensuite en considérant une suite à discrépance faible ayant encore plus de points. Lors de cette itération, on peut, soit considérer la nouvelle

suite construite $\boldsymbol{x}_{iter}(n_{iter})$, soit la suite de points initiale.

Le nombre initial, N_f, de la suite à discrépance faible peut bien entendu être supérieur à $n+1$. Plus ce nombre sera important, plus le nombre de points de la suite obtenue par application de l'algorithme B_2 sera proche du nombre souhaité $n + n_{supp}$.

1.4.4.3 Méthode 3

Les notations utilisées sont celles précisées au paragraphe (1.4.4.1). L'objectif de cette méthode (analogue à celle du paragraphe (1.4.2.3)) est de spécifier des points régulièrement espacés. Elle s'appliquera donc particulièrement dans le cas où la suite $\boldsymbol{x}(n)$ est régulièrement espacée. Elle peut, par exemple, être utilisée après application de l'algorithme A_3 (paragraphe (1.4.2.3)).

Algorithme B_3

Étape 1 : Soit n_{sup} un nombre fixé, $\boldsymbol{x}_f(N_f)$, une suite à discrépance faible, et ε, une distance fixée ; on pose : $\boldsymbol{x}_{ini}(n_{ini}) = \boldsymbol{x}(n)$;

Étape 2 : pour $i = 1, \ldots, n$,
- $\mathrm{dmin}_i = \min_{j=1,\ldots,N_f} \|x_{f_j} - x_i\|$;

Étape 3 : soit $x_{f_{i^*}}$ tel que $\mathrm{dmin}_{i^*} = \arg\min_{i=1,\ldots,n}\{\mathrm{dmin}_i > \varepsilon\}$, on pose :
- $\boldsymbol{x}_{iter}(n_{iter}) = \boldsymbol{x}(n) \cup x_{f_{i^*}}$ et calcul de $\mathrm{DC}^{L^2}(n_{iter})$;

Étape 4 : itération (retour à l'Étape 2) avec $\boldsymbol{x}(n) = \boldsymbol{x}_{iter}(n_{iter})$, et arrêt lorsque $\#(\boldsymbol{x}(n)) = n_{ini} + n_{supp}$.

Le nombre de points n_{supp} à spécifier dépend bien entendu de la dimension de l'espace $\mathcal{X} = [0,1]^d$, mais aussi de la distance ε que l'on s'est fixée. Plus celle-ci sera faible, plus le nombre de points à spécifier devra être important, et plus elle sera importante, plus ce nombre sera faible. En effet, l'objectif poursuivi ici est de spécifier des points relativement proches de l'ensemble considéré (distants d'au moins ε). Ainsi, si le nombre de points spécifiés est faible, les points auront tendance à former des groupes, et ne seront donc pas répartis de façon uniforme. Lors de l'application de l'agorithme A_3 il est donc important de vérifier l'évolution de la discrépance.

1.4.5 Application des méthodes de spécification

Nous avons appliqué les algorithmes B_1, B_2, B_3, aux suites de points que nous avions sélectionnées à l'aides des algorithmes A_1, A_2, et A_3 (à partir de la suite initiale $\boldsymbol{x}(n)$ représentée Figure (1.8)). Précisons que les suites à discrépance faible que nous avons utilisées sont des suites de Hammersley.

Nous avons commencé par appliquer l'algorithme B_3 à la suite $\boldsymbol{x}_{A_3}(n_{A_3})$. Pour cette application, la valeur ε est celle que nous avions retenue lors de l'application de l'algorithme A_3 : $\varepsilon = 0.285$. Comme, par construction de l'algorithme B_3 (et A_3), les points doivent tous être distants d'au moins ε, il n'est pas possible d'augmenter indéfiniment le nombre de points à spécifier. Pour cet exemple, le nombre total de la suite de points obtenue est de 43 points. Nous avons donc appliqué les algorithmes B_1 et B_2 aux suites $\boldsymbol{x}_{A_1}(n_{A_1})$ et $\boldsymbol{x}_{A_2}(n_{A_2})$ en vue d'obtenir le même nombre de points.

Les évolutions de la discrépance centrée carrée en fonction du nombre de points lors de l'application des algorithmes B_1, B_2, et B_3 (ayant respectivement comme suite de points initiale, $\boldsymbol{x}_{A_1}(n_{A_1})$, $\boldsymbol{x}_{A_2}(n_{A_2})$, et $\boldsymbol{x}_{A_3}(n_{A_3})$) sont représentées Figure (1.14), (1.15), (1.16).

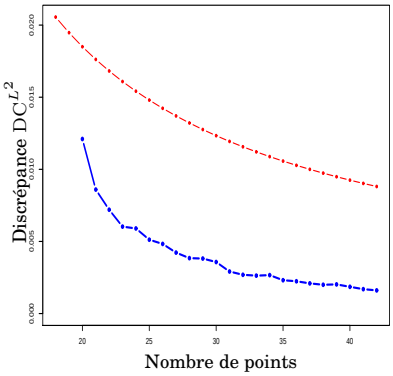


FIG. 1.14 – Évolution de la discrépance carrée centrée par application de l'algorithme B_1

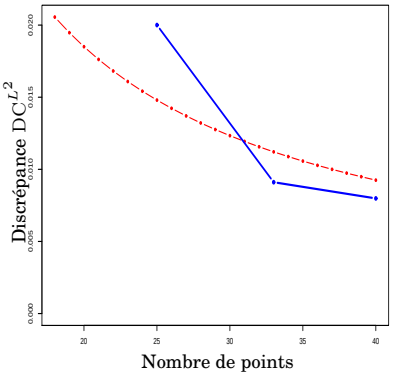


FIG. 1.15 – Évolution de la discrépance carrée centrée par application de l'algorithme B_2

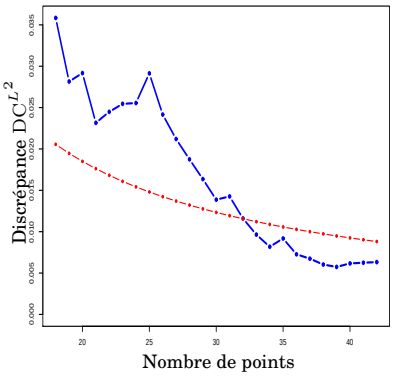


FIG. 1.16 – Évolution de la discrépance carrée centrée par application de l'algorithme B_3

Les suites initiales étaient toutes de qualité « acceptables », dans le sens où leur

discrépance carrée centrée était inférieure à leur espérance (voir équation (1.47)). Nous avons donc choisi de représenter une nouvelle valeur seuil. Elle correspond au quantile à 90% de la statistique définie par la discrépance carrée centrée d'une suite de variables aléatoires indépendantes uniformes dans $[0,1]^3$ (voir paragraphe (1.3.3.2)). Elle est représentée en rouge sur les différentes Figures (1.14), (1.15), (1.16). Il ne s'agit pas ici d'un critère d'arrêt puisque nous avons appliqué les algorithmes dans l'objectif d'avoir le même nombre de points.

Les suites obtenues suites sont notées $\boldsymbol{x}_{B_1}(n_{B_1})$, $\boldsymbol{x}_{B_2}(n_{B_2})$, et $\boldsymbol{x}_{B_3}(n_{B_3})$ et comportent $n_{B_1} = 43$, $n_{B_2} = 40$, $n_{B_3} = 43$ points. L'ensemble des critères du Tableau (1.2) appliquées à $\boldsymbol{x}_{B_1}(n_{B_1})$, $\boldsymbol{x}_{B_2}(n_{B_2})$, et $\boldsymbol{x}_{B_3}(n_{B_3})$ est présenté Tableau (1.5).

La Figure (1.16), représentant l'évolution de la discrépance carrée en fonction du nombre de points lors de l'application de l'algorithme B_3, montre bien que les points spécifiés n'ont pas pour objectif direct la réduction de la discrépance. En effet des points relativement proches de la suite initiale (ici $\boldsymbol{x}_{A_3}(n_{A_3})$) sont tout d'abord spécifiés formant ainsi des groupes de points. Puis, après quelques itérations, les points spécifiés permettent de recouvrir l'espace $\mathcal{X} = [0,1]^3$ de façon uniforme et donc de réduire la discrépance. La Figure (1.16) montre que l'algorithme B_1 semble le plus efficace pour la spécification de points ayant une discrépance faible.

Les critères du Tableau (1.5) montrent que les suites $\boldsymbol{x}_{B_1}(n_{B_1})$ et $\boldsymbol{x}_{B_3}(n_{B_3})$ « occupent » (« recouvrent », « remplissent ») l'espace $\mathcal{X}$ de façon comparable et de meilleure façon que $\boldsymbol{x}_{B_2}(n_{B_2})$. Comme attendu, par construction de l'algorithme B_3, la suite $\boldsymbol{x}_{B_3}(n_{B_3})$ est celle dont la répartition des points est la plus régulière.

Remarquons que certains critères des suites spécifiées sont plus sévères que ceux des suites sélectionnées. Par exemple, nous avons $\text{dmin}_2(\boldsymbol{x}_{B_2}(n_{B_2})) > \text{dmin}_2(\boldsymbol{x}_{A_2}(n_{A_2}))$, $\gamma(\boldsymbol{x}_{B_2}(n_{B_2})) > \gamma(x_{A_2}(n_{A_2}))$, $\Lambda(\boldsymbol{x}_{B_2}(n_{B_2})) > \Lambda(\boldsymbol{x}_{A_2}(n_{A_2}))$. Ces critères sont fonction des distances entre les points de la suite considérée. Par leur construction, les suites spécifiées sont de cardinalité plus élevée que les suites sélectionnées. Par conséquent, pour une suite spécifiée, il existera une distance entre deux points, inférieure ou égale à toutes les distances entre les points de la suite sélectionnée qui lui est associée (critère dmin_2). De plus, les algorithmes B_1 et B_2 n'ont pas pour objectif de spécifier des points régulièrement espacés. Ainsi, les critères exprimant la régularité des espacements de $\boldsymbol{x}_{B_1}(n_{B_1})$ et $\boldsymbol{x}_{B_2}(n_{B_2})$ ne sont pas nécessairement meilleurs que ceux de $\boldsymbol{x}_{B_1}(n_{B_1})$ et $\boldsymbol{x}_{A_2}(n_{A_2})$ (critère γ, Λ, par exemple). Cependant, les critères de discrépance assurent que ces suites « recouvrent », « occupent » mieux l'espace (puisqu'ils sont plus faibles).

Toutes les suites de points ici spécifiées sont de qualité *acceptable* selon notre critère : elles ont une discrépance centrée carrée inférieure à leur espérance.

1.4.6 Discussion

Les critères que nous avons utilisés sont appropriés pour l'évaluation de l'uniformité de la base de données d'entrée au sens déterministe. Rappelons que nous nous plaçons dans le contexte d'une base de données d'entrée pré-existante et que nous ignorons la façon dont elle a été obtenue. Le caractère aléatoire, et l'indépendance ne sont pas pris en compte par cette méthode. Pour l'algorithme A2, la sélection des points n'a pas été réalisée dans l'objectif de conserver l'aléatoire de la base de données d'entrée mais de conserver la structure uniforme au sens déterministe, i.e. avec espacement régulier et *recouvrement* de l'espace. Les tests d'uniformité au sens probabiliste feront l'objet du chapitre III.

Rappelons aussi qu'il ne s'agit pas de valider l'uniformité de la base mais d'apprécier sa qualité. Pour la validation de l'uniformité, il existe d'autres critères faisant parfois directement intervenir les propriétés de construction des suites, voir Morokoff et Caflisch (1994), Lemieux et L'Ecuyer (2001), Hickernell (1996b). Or nous nous plaçons dans le contexte où la méthode utilisée pour l'obtention des données (base de données d'entrée, réponse du code) est inconnue.

Le choix des *suites à discrépance faible* comme suites de référence ou suites permettant la spécification de points peut être délicat. En effet, la propriété de « *discrépance faible* » est asymptotique, et selon le nombre de points et la dimension certaines suites peuvent ne pas être adaptées, voir graphique (1.17) représentant l'évolution de la discrépance L^2 en fonction du nombre de points de suites à 2 dimensions. D'autre part, pour certaines dimensions, elles peuvent avoir certaines pathologies comme, par exemple, une suite de Halton en base $(17, 19)$ (qui correspond à la projection de la suite de Halton à 8 dimensions sur les axes x_7,x_8) représentée dans le graphique (1.18). Il est cependant possible de les modifier de façon à supprimer ces problèmes (voir Chi *et al.* (2005), Vandewoestyne et Cools (2004)).

Pour planifier des expériences à dimension élevée et lorsque le modèle de calibration est inconnu, nous pouvons, par exemple, utiliser des réseaux (*lattice*), voir L'Ecuyer (2004), Lemieux et L'Ecuyer (2001), Niederreiter et Wills (2005).

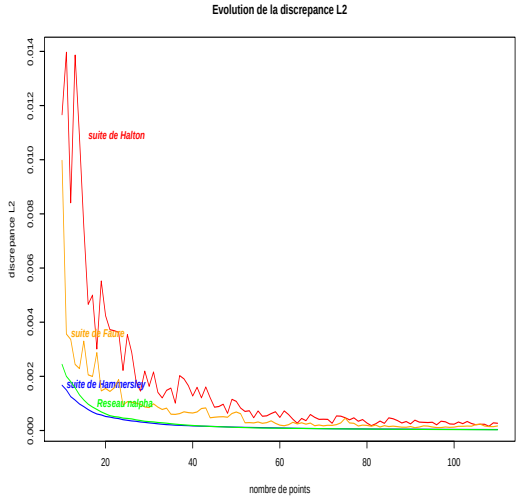


FIG. 1.17 – Évolution de la discrépance L2

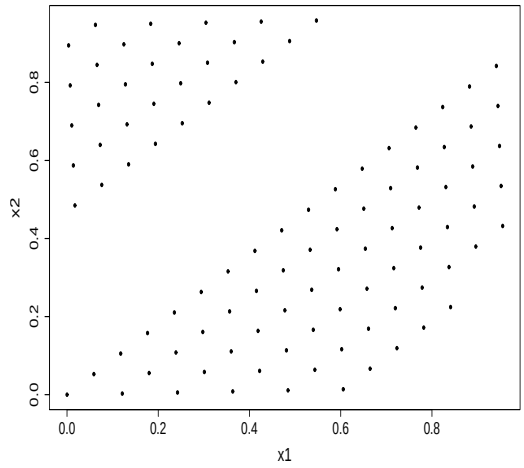


FIG. 1.18 – « Pathologie » d'une suite de Halton (100 points)

Chapitre 2

Liens entre discrépance et estimation non-paramétrique, méthodologie de sélection de points selon les données disponibles

2.1 Introduction

Ce chapitre reproduit, en l'adaptant au format du présent mémoire l'article de Feuillard (2006) soumis pour publication.

Nous nous intéressons ici aux liens qui existent entre une méthode d'estimation d'un paramètre fonctionnel et un critère d'uniformité d'un ensemble de points appelé discrépance. Nous complétons et développons ici l'article de Feuillard *et al.* (2006). Une méthodologie ayant pour objectif la construction ou l'amélioration d'une base de données en vue d'une meilleure estimation (au sens des critères d'$\mathbb{IMSE}$, *Integrated Mean Square Error* et de $\mathbb{MSE}$, *Mean Square Error*, définis plus loin) d'un paramètre fonctionnel en sera déduite.

Le modèle considéré est le suivant. On observe les y_i, $i = 1 \ldots, n$, où :

$$y_i \quad = \quad f(x_i) + \varepsilon_i, \quad \text{où,} \tag{2.1}$$

- $x_i \in \mathcal{X} = [0,1)^d$,
- $f(\cdot) \in L^2(\mathcal{X})$ est le paramètre fonctionnel à estimer,
- ε_i, $i = 1, \ldots, n$, est une suite de variables aléatoires, mutuellement indépendantes de moyenne nulle de variance σ^2, et indépendantes des x_i, $i = 1, \ldots, n$.

La méthode d'estimation fonctionnelle utilisée, introduite par Cencov (1962), consiste tout d'abord à projeter $f(\cdot)$ sur un sous-espace de dimension finie de l'espace de Hilbert

des fonctions de carré intégrable $(L^2(\mathcal{X}), \|.\|_{L^2})$. Ce sous-espace est défini par les premières composantes d'un système spécifié de fonctions orthonormales de $(L^2(\mathcal{X}), \|.\|_2)$. Une estimation de la projection de $f(\cdot)$ sur ce sous-espace est alors obtenue par une méthode de moments empiriques. Cette méthode, aujourd'hui classique, et dite *des fonctions orthogonales*, a fait l'objet de nombreuses études, notamment dans le cas de l'estimation d'une densité de probabilité. Nous renvoyons, par exemple, à Schwartz (1967), Kronmal et Tarter (1968), Bosq (1969), Watson (1969), Földes et Révész (1974), Szegö (1975), Sansone (1977), Sterbuchner (1980), Greblicki et Pawlak (1981), Prakasa Rao (1983), Eubank et Speckman (1990). Les ouvrages de Devroye et Györfi (1985), Bosq et Lecoutre (1987), Hardle (1989), Nadaraya (1989), Bosq (2005) peuvent aussi être consultés à ce sujet ainsi que leurs références bibliographiques.

La fonction $f(\cdot)$ étant supposée de carré intégrable sur $\mathcal{X}$, celle-ci peut s'écrire à partir d'un système de fonctions orthonormales $\{v_1(\cdot), \ldots, v_k(\cdot), \ldots\} \subset L^2(\mathcal{X})$, soit

$$f(x) = \sum_{k \geq 1} a_k v_k(x) \quad \text{où} \quad a_k := \int_{\mathcal{X}} f(x) v_k(x) dx. \tag{2.2}$$

L'estimation de $f(\cdot)$, par projection sur le sous-espace défini par les N premières fonctions du système orthonormal $\{v_1(\cdot), \ldots, v_k(\cdot), \ldots\}$, s'écrit alors sous la forme :

$$\hat{f}_n(x) := \sum_{k=1}^{N} \hat{a}_k v_k(x) \quad \text{où} \quad \hat{a}_k := \frac{1}{n} \sum_{i=1}^{n} y(x_i) v_k(x_i). \tag{2.3}$$

Les coefficients $\hat{a}_k$ sont des estimateurs empiriques sans biais des a_k. Pour le choix de la dimension du sous-espace N en fonction de la taille n de l'échantillon, nous renvoyons, par exemple, à Bosq et Bluez (1978), et Aubin (2005), ainsi qu'à leurs références bibliographiques. Il est classique de supposer que, lorsque $n \to \infty$, on a $N/n \to 0$ et $N \to \infty$ (cf. Bosq et Bluez (1978)).

Pour ce type d'estimation, Hickernell (1999) et Rafajlowicz et Schwabe (2005) ont montré qu'il existe un lien entre un critère d'uniformité d'un ensemble de points $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ dans l'espace $\mathcal{X} = [0,1)^d$, appelé discrépance, et des critères de qualité d'estimation de $f(\cdot)$. Nous retiendrons ici les critères de la $\mathbb{MSE}$ (Mean Square Error), et de l'$\mathbb{IMSE}$, *Integrated Mean Square Error* (voir l'équation (2.12) du paragraphe 2.2 et les équations 2.21 du §2.3 pour des définitions précises de ces quantités). Leur approche est ici développée.

Dans la suite de ce chapitre, nous rappellerons tout d'abord l'inégalité de Koksma-Hlwaka généralisée, faisant intervenir la discrépance généralisée. Cette inégalité sera ensuite utilisée pour fournir des majorations des critères d'$\mathbb{IMSE}$ et de $\mathbb{MSE}$. A partir de ces résultats, une méthodologie d'analyse, de sélection et de spécification de points

expérimentaux sera proposée. Enfin, une application à un exemple illustrera cette méthodologie.

2.2 Inégalité de Koksma-Hlwaka généralisée

2.2.1 Notations et hypothèse

Avant de définir l'inégalité de Koksma-Hlwaka généralisée, commençons par préciser quelques notations.

i) On pose $\mathcal{X} = [0,1)^d$;

ii) $\boldsymbol{u} = \{u_1, \ldots, u_\ell\}$ désigne un sous-ensemble non vide d'indices de $\{1, \ldots, d\}$;

iii) $|\boldsymbol{u}|$ désigne la cardinalité d'un ensemble non vide $\boldsymbol{u} \subset \{1, \ldots, d\}$, par exemple, pour $\boldsymbol{u} = \{u_1, \ldots, u_\ell\} \subset \{1, \ldots, d\}$, $|\boldsymbol{u}| = l$;

iv) Pour $x \in \mathcal{X}$ et $\boldsymbol{u} \subset \{1, \ldots, d\}$, nous désignons par $x^{(\boldsymbol{u})}$ le point (vecteur) extrait de $x = (x^{(1)}, \ldots, x^{(d)})' \in \mathcal{X}$ dont les composantes sont indexées par les indices de $\boldsymbol{u}$, $x^{(\boldsymbol{u})} = (x^{(u_1)}, \ldots, x^{(u_\ell)})'$.

Dans la notation (iv), x' désigne la transposée du vecteur x.

Pour définir l'inégalité de Koksma-Hlwaka généralisée, l'hypothèse suivante doit être vérifiée (voir Hickernell (1998)), pour une fonction $f(\cdot)$ donnée, définie sur $\mathcal{X} = [0,1)^d$.

Hypothèse 1

La fonction $f(\cdot)$ est continue et indéfiniment dérivable sur $\mathcal{X} = [0,1)^d$, et toute dérivée partielle croisée d'ordre inférieur ou égal à d de $f(\cdot)$ est intégrable d'ordre p :

$$f(\cdot) \in C(\mathcal{X}), \quad et \quad f(\cdot) \in \mathcal{W}_p(\mathcal{X}) \equiv \left\{ \begin{array}{l} f := \frac{\partial^{|\boldsymbol{u}|} f}{\partial x^{(\boldsymbol{u})}} = \frac{\partial^{|\boldsymbol{u}|} f}{\partial x^{(u_1)} \ldots \partial x^{(u_\ell)}} \in L^p([0,1]^{|\boldsymbol{u}|}), \\ \forall \boldsymbol{u} = \{u_1, \ldots, u_\ell\} \subseteq \{1, \ldots, d\} \end{array} \right\}, \quad \text{avec } p \text{ entier et } 1 \leq p < \infty. \tag{2.4}$$

2.2.2 Inégalité généralisée de Koksma-Hlwaka

De façon générale, sous l'hypothèse (1), pour une suite de points $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ dans $\mathcal{X}$, l'inégalité de Koksma-Hlawaka généralisée peut s'écrire comme suit (voir, par exemple, Niederreiter et Spanier (Eds) (1998)) :

$$|I(f) - \hat{I}_{\boldsymbol{x}(n)}(f)| = \left| \int_{\mathcal{X}} f(x) d(F_n(x) - U(x)) \right| \leq V(f) \mathrm{D}(\boldsymbol{x}(n)), \tag{2.5}$$

où :

i) $I(f) = \int_{\mathcal{X}} f(x)dx$, où $dx = dU(x)$ est la mesure de Lebesgue sur $\mathcal{X} = [0,1)^d$,

ii) $\hat{I}_{\boldsymbol{x}(n)}(f) = \frac{1}{n} \sum_{x_i \in \boldsymbol{x}(n)} f(x_i)$ avec $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$,

iii) F_n désigne la fonction de répartition empirique de l'échantillon $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ dans $[0,1)^d$,

iv) U désigne la fonction de répartition de la loi uniforme sur $[0,1)^d$,

v) $V(f)$ est une *variation* de $f(\cdot)$ (voir plus loin),

vi) et $\mathrm{D}(\boldsymbol{x}(n))$ est un terme qui dépend des points $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$, et correspond à la notion de *discrépance* (voir plus loin).

• Dans le cas de l'inégalité classique de Koksma-Hlwaka (voir Hlwaka (1961) et Niederreiter (1992)), $V(f)$ désigne *la variation totale au sens de Hardy et Krause de* $f(\cdot)$, et $\mathrm{D}(\boldsymbol{x}(n))$ la *discrépance à l'origine de* $\boldsymbol{x}(n)$, usuellement notée D^*, et définie par :

$$\mathrm{D}^*(\boldsymbol{x}(n)) := \|F_n - U\|_{L^\infty(\mathcal{X})}. \tag{2.6}$$

Dans le cas où la suite $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ est la réalisation d'une suite de variables aléatoires indépendantes identiquement distribuées, $\mathrm{D}^*(\boldsymbol{x}(n))$ correspond à la statistique de Kolmogorov-Smirnov (voir les inégalités de Dvoretzky *et al.* (1956), et de Massart (1990) lorsque $d = 1$, pour l'étude à distance finie de cette statistique).

• Dans le cas de l'inégalité de Koksma-Hlwaka généralisée, la définition de la discrépance $\mathrm{D}(\boldsymbol{x}(n))$ fait usage d'une norme convenable de l'application $(I - \hat{I}_{\boldsymbol{x}(n)})$ définie par :

$$f \to \mathrm{Err}(f, \boldsymbol{x}(n)) = (I - \hat{I}_{\boldsymbol{x}(n)})(f). \tag{2.7}$$

La définition de la variation $V(f)$ utilise une norme de la fonction $f(\cdot)$ dans un espace de Hilbert à noyau auto-reproduisant (en abrégé, RKHS pour *Reproducing Kernel Hilbert Space*, voir Hickernell (1998), on pourra aussi consulter Carraro (2007) pour une présentation de différents RKHS). Comme il est possible de définir différents noyaux et, par conséquent, différentes normes sur de tels espaces, on obtient de nombreuses variantes de ce résultat.

Soit un espace de Hilbert $\mathcal{H} \equiv \mathcal{W}_2(\mathcal{X})$ de fonctions sur $\mathcal{X}$ (selon la notation (2.4) de l'Hypothèse 1, prise pour $p = 2$) muni d'un noyau auto-reproduisant K, symétrique, de *type positif*, et de carré intégrable.

$$K(x,y) = K(y,x), \qquad \forall x, y \in [0,1)^d \tag{2.8}$$

$$\sum_{i,k} a_i\, a_k\, K(x_i, x_k) \geq 0, \qquad \forall a_i \in \mathbb{R}, \quad x_i \in [0,1)^d \tag{2.9}$$

$$\int_{[0,1)^d} K(x,x)dx < \infty. \tag{2.10}$$

Soit $f \in \mathcal{H}$ une fonction sur $\mathcal{X}$. Lorsque $f(\cdot)$ est constante, nous avons $\mathrm{Err}(f, \boldsymbol{x}(n)) = 0$ dans (2.7). Nous désignons par $f_\perp$ la projection de $f(\cdot)$ sur le sous-espace de $\mathcal{X}$ orthogonal

à la fonction 1 (le fait que $1 \in \mathcal{H}$ est conséquence de (2.10)). On notera $< ., . >_K$ le produit scalaire induit par le noyau K, et

$$\|.\|_K = (< ., . >_K)^{1/2}, \tag{2.11}$$

la norme correspondante. Nous avons $\mathrm{Err}(f, \boldsymbol{x}(n)) := \mathrm{Err}(f_\perp, \boldsymbol{x}(n))$, $\forall f(\cdot) \in \mathcal{H}$. Par le théorème de représentation de Riesz (voir Riesz et Nagy (1955)), il existe $\xi \in \mathcal{H}$ tel que : $\mathrm{Err}(f, \boldsymbol{x}(n)) =< \xi, f >_K$, $\forall f \in \mathcal{H}$. Ainsi :

$$\mathrm{Err}(f, \boldsymbol{x}(n)) = \mathrm{Err}(f_\perp, \boldsymbol{x}(n)) =< \xi, f_\perp >_K .$$

Par application de l'inégalité de Cauchy-Schwarz, on obtient :

$$\mathrm{Err}(f, \boldsymbol{x}(n)) \leq \|f_\perp(\cdot)\|_K \, \|\xi\|_K . \tag{2.12}$$

L'inégalité (2.12) est appelée *inégalité de Koksma-Hlwaka généralisée* (Niederreiter et Spanier (Eds) (1998)). La *discrépance généralisée* de $\boldsymbol{x}(n)$ est alors définie par $\mathrm{D}(\boldsymbol{x}(n)) = \mathrm{D}^{L^2}(\boldsymbol{x}(n), K) = \|\xi\|_K$ (voir Hickernell (1998)). En utilisant la forme explicite de ξ, la *discrépance généralisée* ci-dessus s'exprime comme suit :

$$\mathrm{D}^{L^2}(\boldsymbol{x}(n), K) := \left\{ \int_{\mathcal{X}^{2d}} K(x, y) d(F_n(x) - U(x)) d(F_n(y) - U(y)) \right\}. \tag{2.13}$$

Une forme du noyau K fréquemment utilisée (voir Hickernell (1998)) est la suivante. Pour $x = (x^{(1)}, \ldots, x^{(d)})' \in \mathcal{X}$ et $y = (y^{(1)}, \ldots, y^{(d)}) \in \mathcal{X}$, on a

$$\begin{aligned} K(x, y) &:= \prod_{j=1}^{d} K_1(x^{(j)}, y^{(j)}), \\ &\text{avec :} \\ K_1(x^{(1)}, y^{(1)}) &:= M + \beta^2 \left[\mu(x^{(1)}) + \mu(y^{(1)}) + \frac{1}{2} B_2(\{x^{(1)} - y^{(1)}\}) + B_1(x^{(1)}) B_1(y^{(1)}) \right]. \end{aligned} \tag{2.14}$$

Ici $\{x^{(1)} - x'^{(1)}\}$ désigne la partie fractionnelle de $(x^{(1)} - x'^{(1)})$, soit $\{u\} := u - \lfloor u \rfloor$, où $\lfloor u \rfloor$ est la partie entière de u, $\lfloor u \rfloor \leq u \leq \lfloor u \rfloor + 1$. Les fonctions $B_1(\cdot)$ et $B_2(\cdot)$ utilisées dans (2.14) sont :

$$B_1(x) := x - \frac{1}{2}, \qquad B_2(x) := x^2 - x + \frac{1}{6}.$$

Des choix possibles de β, M et μ dans (2.14) sont :

$$\beta := 1, \qquad M := \frac{4}{3}, \qquad \mu(x) := \frac{1}{6} - \frac{x^2}{2}, \tag{2.15}$$

$$\beta := 1, \qquad M := \frac{13}{12}, \qquad \mu(x) := \frac{1}{2} B_2(\{x - 1/2\}), \tag{2.16}$$

$$\beta := \frac{1}{2}, \qquad M := \frac{4}{3}, \qquad \mu(x) := -\frac{1}{2} B_2(x). \tag{2.17}$$

Les cas définis par les équations (2.15), (2.16), et (2.17) correspondent, respectivement, à la discrépance L^2 *modifiée* $\mathrm{DM}^{L^2}(\boldsymbol{x}(n))$, L^2 *symétrique*, $\mathrm{DS}^{L^2}(\boldsymbol{x}(n))$, et L^2 *centrée* $\mathrm{DC}^{L^2}(\boldsymbol{x}(n))$.

L'intérêt pratique des trois discrépances définies ci-dessus est qu'elles sont facilement calculables quelle que soit la dimension d de l'espace $\mathcal{X} = [0,1)^d$. De simples formules analytiques existent à cet effet (voir Hickernell (1998)). Ceci n'est pas le cas de la *discrépance à l'origine*, définie en (2.6), $\mathrm{D}^*(\boldsymbol{x}(n))$, dont le calcul est délicat en dimension $d > 2$ (voir Thiémard (2000)). De plus, les discrépances définies par (2.15), (2.16), (2.17), ont une interprétation géométrique simple. Ce sont des comparaisons entre la proportion de points de $\boldsymbol{x}(n)$ compris dans des pavés de $[0,1)^d$ (la mesure empirique) et le volume de ces pavés. Elles définissent donc bien des caractérisations de la disposition des points $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ dans l'espace $\mathcal{X} = [0,1)^d$.

Il est aussi possible de définir des inégalités de Koksma-Hlwaka analogues à (2.12) pour des espaces de Banach $\mathcal{W}_p(\mathcal{X})$ (voir équation (2.4) dans l'Hypothèse 1) Il suffit alors de remplacer l'inégalité (2.12) de Cauchy-Schwarz par une inégalité de Hölder :

$$\mathrm{Err}(f, \boldsymbol{x}(n)) \leq \left\| f_\perp^p(\cdot) \right\|_K^{1/p} \left\| \xi^q \right\|_K^{1/q}. \tag{2.18}$$

Ici, p et q sont tels que $1/p + 1/q = 1$, et $1 \leq p,\, q \leq \infty$, dans le cas des espaces de Hilbert à noyau défini par (2.15) ou (2.16). On supposera par contre que $1 < p,\, q < \infty$, pour le cas de l'espace de Hilbert muni d'un noyau défini par (2.17). L'inégalité de Hölder appliquée dans l'espace de Hilbert à noyau (2.15) pour $p = 1$, $q = \infty$ correspond à l'inégalité de Koksma-Hlwaka classique.

2.2.3 Considération du processus empirique uniforme

Lorsque $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ correspond à la réalisation d'une suite de variables aléatoires indépendantes et identiquement distribuées sur $\mathcal{X} = [0,1)^d$, la *discrépance généralisée* peut encore s'écrire sous la forme :

$$\mathrm{D}^{L^2}(\boldsymbol{x}(n), K) = \frac{1}{\sqrt{n}} \left\| \alpha_n(\cdot) \right\|_K, \tag{2.19}$$

où α_n est le processus empirique défini par $\alpha_n := n^{1/2}(F_n(x) - U(x))$ cf *iii)* du §2.2.1. Dans le cas particulier de la *discrépance de type L^2 modifiée*, on a

$$\left[\mathrm{DM}^{L^2}(\boldsymbol{x}(n))\right]^2 = \sum_{\boldsymbol{u} \subset \{1,\ldots,d\}} \left\| F_n^{(\boldsymbol{u})}(\cdot) - U^{(\boldsymbol{u})}(\cdot) \right\|_{L^2}, \tag{2.20}$$

où $\boldsymbol{u}$ parcourt tous les ensembles d'indices non vides de $\{1, \ldots, d\}$. Ici $F_n^{(\boldsymbol{u})}(\cdot)$, et $U^{(\boldsymbol{u})}(\cdot)$, désignent respectivement la fonction de répartition empirique et la fonction de répartition uniforme, de la projection des points $\boldsymbol{x}(n) = \{x_1, \ldots, x_d\}$ dans les sous espaces

définis par les axes de $[0,1)^d$ indexés par l'ensemble $\boldsymbol{u}$. Il s'agit donc de la somme des statistiques de Cramer-Von-Mises associées à toutes les projections possibles des points de $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ selon les axes de $[0,1)^d$. Or, pour un ensemble d'indices non vide $\boldsymbol{u} = \{u_1, \ldots, u_\ell\} \subset \{1, \ldots, d\}$, le processus empirique $\alpha_n^{(u)}$ converge vers un pont brownien standard multivarié (voir Araujo et Giné (1980), par exemple). A l'aide d'un développement de Karhunen-Loève de ce pont brownien (voir Deheuvels *et al.* (2006)), il est alors possible d'établir la convergence en loi

$$\int_{[0,1)^{2d}} \left[\alpha_n^{(\boldsymbol{u})}(x)\right]^2 dx^{(u_1)} \ldots dx^{(u_\ell)} \quad \stackrel{\mathcal{L}}{\to} \quad \sum_{k_{\boldsymbol{u}} \geq 0} \lambda_{k_{\boldsymbol{u}}} Y_{k_{\boldsymbol{u}}}^2,$$

où : $\lambda_{k_{\boldsymbol{u}}} = \lambda_{k_{u_1} \ldots k_{u_\ell}}$ désigne un tableau ordonné de constantes positives convenables, et $Y_{k_{\boldsymbol{u}}} = Y_{k_{u_1} \ldots k_{u_\ell}}$ un tableau de v.a. i.i.d. de loi $\mathcal{N}(0,1)$. Ainsi, compte tenu de (2.20), on a

$$n\left[\mathrm{DM}^{L^2}(\boldsymbol{x}(n))\right]^2 = \|\alpha_n(\cdot)\|_K \quad \stackrel{\mathcal{L}}{\to} \quad \sum_{\boldsymbol{u} \subset \{1,\ldots,d\}} \sum_{k_{\boldsymbol{u}} \geq 0} \lambda_{k_{\boldsymbol{u}}} Y_{k_{\boldsymbol{u}}}^2,$$

et la statistique $n\left[\mathrm{DM}^{L^2}(\boldsymbol{x}(n))\right]^2$ converge vers une somme pondérée de variables aléatoires de loi du χ_1^2. Les cas des statistiques définies par les discrépances centrée, DC^{L^2}, et symétrique, DS^{L^2}, sont similaires. Cependant, les coefficients $\lambda_{k_{\boldsymbol{u}}}$ ne sont pas toujours connus de façon explicite. Pour pallier ce problème, des processus définis sur les marges du pavé unité sont utilisés (voir Deheuvels (1981) et Deheuvels *et al.* (2006)). Il semble toutefois délicat d'y avoir recours pour connaître la loi exacte des statistiques : $\mathrm{DM}^{L^2}(\boldsymbol{x}(n))$, $\mathrm{DS}^{L^2}(\boldsymbol{x}(n))$, $\mathrm{DC}^{L^2}(\boldsymbol{x}(n))$. A notre connaissance, les lois de ces statistiques ne sont en général pas toutes connues de façon explicite. Mais, les espérances de ces statistiques peuvent être évaluées (voir Hickernell (1996b) et Hickernell (1998)) et une tabulation des lois demeure possible par simulation.

2.3 Majoration de critères

2.3.1 Introduction

Rappelons que nous nous plaçons dans le contexte de l'estimation $\hat{f}_n(\cdot)$ d'un paramètre fonctionnel $f(\cdot)$ par la méthode des fonctions orthogonales décrite par (4.1), (2.2) et (2.3). Dans ce qui suit, nous nous intéresserons à la majoration de critères exprimant la qualité de l'estimation $\hat{f}_n(\cdot)$ de $f(\cdot)$, l'$\mathbb{IMSE}$, et la $\mathbb{MSE}$, définis comme suit :

$$\mathbb{IMSE}(f, \hat{f}_n) := \int_{\mathcal{X}} \mathbb{E}(\hat{f}_n(x) - f(x))^2 dx, \quad \text{et} \quad \mathbb{MSE}(\hat{a}_1, a_1) := \mathbb{E}(\hat{a}_1 - a_1)^2. \quad (2.21)$$

Notons que lorsque la première fonction du système orthogonal $\{v_k(\cdot) : k \geq 1\}$ est égale à 1, $v_1 = 1$, $\mathbb{MSE}(\hat{a}_1, a_1)$ correspond à un critère de robustesse d'estimation de la moyenne $a_1 = \int_{\mathcal{X}} f(x)dx$ (voir (2.2) et (2.3)).

Par la suite, le nombre N des fonctions orthogonales $\{v_1(\cdot), \ldots, v_N(\cdot)\} \subset L^2(\mathcal{X})$ utilisées pour l'estimation de $f(\cdot)$ (cf. (2) et (3)) sera alors supposé spécifié (voir par exemple Bosq et Bluez (1978), Aubin (2005) ou Devroye et Györfi (1985)).

Les majorations de l'$\mathbb{IMSE}$ et de la $\mathbb{MSE}$ que nous proposerons seront obtenues à l'aide de l'inégalité, (2.12) de Koksma-Hlwaka généralisée, appliquée aux fonctions $f(\cdot)$ et $v_k(\cdot)$, $1 \leq k \leq N$. Aussi, nous supposerons que $f(\cdot)$ vérifie l'Hypothèse 1 présentée au §2.2.1 avec $p = 2$ dans (2.4), $f \in \mathcal{C}(\mathcal{X}) \cap \mathcal{W}_2(\mathcal{X})$. Pour la majoration de l'$\mathbb{IMSE}$, nous ferons usage de l'hypothèse supplémentaire suivante portant sur le choix de fonctions orthogonales $\{v_k(\cdot) : 1 \leq k \leq N\}$ de $L^2(\mathcal{X})$.

Hypothèse 2

Les fonctions $v_k(\cdot)$, $k \geq 1$, vérifient l'Hypothèse 1 avec $p = 2$ dans (2.4) et ont pour norme dans $L^2(\mathcal{X})$ la fonction constante 1.

$$v_k(\cdot) \in \mathcal{C}(\mathcal{X}) \cap \mathcal{W}_2(\mathcal{X}), \qquad et \qquad \int_0^1 v_k(x)^2 dx = 1. \tag{2.22}$$

L'application de l'inégalité (2.12) de Koksma-Hlwaka permettra de faire apparaître les discrépances généralisées de type L^2 (cf. (2.13), (2.14), (2.15), (2.16), (2.17)) dans les termes de majoration de l'$\mathbb{IMSE}$ et de la $\mathbb{MSE}$.

2.3.2 Majoration de l'$\mathbb{IMSE}$

A l'aide de l'*inégalité classique de Koksma-Hlwaka* (cf. (2.5) et (2.6)), il est possible de majorer l'$\mathbb{IMSE}$ par un terme dépendant de la *discrépance à l'origine* (on se référera aux arguments de Rafajlowicz et Schwabe (2005)). En reprenant cette approche, et en la généralisant au cas des discrépances généralisées de type L^2, nous avons obtenu la majoration suivante :

$$\begin{aligned} \mathbb{IMSE}(f, \hat{f}_n) \quad \leq \quad & \frac{N\sigma^2}{n} \left[1 + 2\, M_v\, C_{V(v)}\, \mathrm{D}^{L^2}\left(\boldsymbol{x}(n), K\right)\right] \\ & + N \left[C_{V(f)} M_v + M_f C_{V(v)}\right]^2 \mathrm{D}^{L^2}(\boldsymbol{x}(n), K)^2 \\ & + R(N, f). \end{aligned} \tag{2.23}$$

Ici, $C_{V(v)}$, et $C_{V(f)}$, sont des constantes qui dépendent, respectivement, des fonctions $v_k(\cdot)$, $k = 1, \ldots, N$, et $f(\cdot)$. M_v, et $M(f)$, sont des constantes qui majorent, respectivement, les normes uniformes, des $v_k(\cdot)$, pour $k = 1, \ldots, N$, et de la fonction $f(\cdot)$. Le terme $R(N, f)$ est un terme d'erreur. Enfin, $\mathrm{D}^{L^2}(\boldsymbol{x}(n), K)$ désigne *une discrépance généralisée de type L^2* (voir équation (2.13)).

Démonstration (de l'inégalité (2.23)) :
(inspirée de Rafajlowicz et Schwabe (2005))

A) Rappel des Hypothèses et notations supplémentaires

Nous renvoyons à (4.1) et (2.2) pour les définitions de $f(\cdot)$ et $v_k(\cdot)$, $k \geq 1$. Comme précisé au §2.3.1, nous supposons que $f(\cdot)$ vérifie l'Hypothèse 1, et $v_k(\cdot)$, $k \geq 1$, l'Hypothèse 2.

Par la suite, nous ne traiterons que le cas où l'on considère le RKHS $\mathcal{H} \equiv \mathcal{W}_2(\mathcal{X})$ muni du noyau défini par (2.14) et (2.15). Dans ce cadre théorique, l'application de l'inégalité (2.12) de Koksma-Hlwaka à une fonction $g \in \mathcal{H}$ en un ensemble de points $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ fera intervenir la *discrépance modifiée* de $\boldsymbol{x}(n)$ notée $\mathrm{DM}^{L^2}(\boldsymbol{x}(n)$ (cf. §2.2.2). Les cas des RKHS $\mathcal{H}$ de noyau défini par (2.14) et (2.16), et, (2.14) et (2.17) étant similaires, le détail des calculs ne sera pas précisé. Nous noterons :

i) $\partial^{|\boldsymbol{u}|} g / \partial x^{(\boldsymbol{u})}$, la dérivée partielle croisée d'ordre $|\boldsymbol{u}|$ par rapport à $x^{(\boldsymbol{u})} = (x^{(u_1)}, \ldots, x^{(u_\ell)})$, (voir (2.4) et les notations (i), (ii), (iii) (iv) du paragraphe 2.2) ;

ii) $x^{(\boldsymbol{u}^c)} = (1, \ldots, 1)$, l'ensemble des points $x = (x(1), \ldots, x^{(d)}) \in \mathcal{X} = [0,1)^d$ tels que $x^{(m)} = 1$, pour tout $m \notin \boldsymbol{u} = \{u_1, \ldots, u_\ell\} \subseteq \{1, \ldots, d\}$. De manière explicite,

$$\begin{aligned} &\left\{x^{(\boldsymbol{u}^c)} = (1, \ldots, 1)\right\} := \\ &\qquad \left\{x \,:\, x \in \mathcal{X} = [0,1)^d \text{ tel que : } x^{(j)} = 1,\ \forall j \in \{1, \ldots, d\}, \text{ et } j \notin \boldsymbol{u}\right\} ; \end{aligned}$$

iii) $g|_{\left\{x^{(\boldsymbol{u}^c)} = (1, \ldots, 1)\right\}}$, la restriction d'une fonction g sur $\left\{x^{(\boldsymbol{u}^c)} = (1, \ldots, 1)\right\}$;

iv) pour une fonction $g(\cdot) \in \mathcal{W}_2(\mathcal{X})$ et pour un ensemble non vide d'indices $\boldsymbol{u} \subset \{1, \ldots, d\}$,

$$VM^{(\boldsymbol{u})}(g) \;:=\; \int_{[0,1]^{|\boldsymbol{u}|}} \left[\left. \frac{\partial^{|\boldsymbol{u}|}(g(x))}{\partial x^{(\boldsymbol{u})}} \right|_{x^{(\boldsymbol{u}^c)} = (1, \ldots, 1)} \right] dx^{(\boldsymbol{u})}. \tag{2.24}$$

Nous désignerons la norme (cf. (2.11))

$$\mathcal{V}_{\mathcal{M}}(g) := \|g(\cdot)_{\perp}\|_K, \tag{2.25}$$

sous le nom de *variation modifiée*, par analogie avec la *discrépance modifiée* obtenue dans le contexte où le noyau K du RKHS $\mathcal{H}$ est défini par (2.14) et (2.15) (cf §2.2.2). Pour

$g \in \mathcal{H}$, il ressort de Hickernell (1998) que $\mathcal{V}_{\mathcal{M}}(g)$ peut s'écrire sous la forme

$$\mathcal{V}_{\mathcal{M}}(g) := \left[\sum_{\boldsymbol{u}\subset\{1,\ldots,d\}} \left\| \frac{\partial^{|\boldsymbol{u}|} g}{\partial x^{(\boldsymbol{u})}}\Big|_{\left\{x^{(\boldsymbol{u}^c)}=(1,\ldots,1)\right\}} \right\|^2_{L^2([0,1]^{|\boldsymbol{u}|})}\right]^{1/2}, \tag{2.26}$$

$$:= \left[\sum_{\boldsymbol{u}\subset\{1,\ldots,d\}} \left(VM^{(\boldsymbol{u})}(g)\right)^2\right]^{1/2}, \quad \text{selon nos notations.} \tag{2.27}$$

B) Décomposition de l'$\mathbb{IMSE}$

En utilisant le théorème de Fubini pour $f \in L^2(\mathcal{X})$, et $v_k(\cdot) \in L^2(\mathcal{X})$, nous écrivons,

$$\begin{aligned}
\mathbb{IMSE}(f,\hat{f}_n) &= \int_{\mathcal{X}} \mathbb{E}[\hat{f}_n(x) - f(x)]^2 dx \\
&= \int_{\mathcal{X}} \mathbb{E}\left[\sum_{k=1}^{N} \hat{a}_k v_k(x) - \sum_{k\geq 1} a_k v_k(x)\right]^2 dx \\
&= \mathbb{E}\left\{\int_{\mathcal{X}} \left[\sum_{k=1}^{N} (\hat{a}_k - a_k) v_k(x) + \sum_{k\geq N+1} a_k v_k(x)\right]^2 dx\right\}.
\end{aligned}$$

Par l'orthogonalité des $\{v_k(\cdot) : k \geq 1\}$, et une nouvelle application de Fubini, on en déduit que

$$\begin{aligned}
\mathbb{IMSE}(f,\hat{f}_n) &= \mathbb{E}\left\{\sum_{k=1}^{N} \int_{\mathcal{X}} [(\hat{a}_k - a_k)\, v_k(x)]^2 dx\right\} + \mathbb{E}\left\{\sum_{k\geq N+1} \int_{\mathcal{X}} [a_k v_k(x)]^2\right\} dx \\
&= \sum_{k=1}^{N} \int_{\mathcal{X}} \mathbb{E}\left[(\hat{a}_k - a_k)\, v_k(x)\right]^2 dx + \sum_{k\geq N+1} \int_{\mathcal{X}} \mathbb{E}[a_k v_k(x)]^2 dx\,.
\end{aligned}$$

Compte tenu du fait que les $\{v_k(\cdot), k \geq 1\}$ sont orthonormés dans $L^2(\mathcal{X})$ (Hypothèse 2), on conclut que

$$\begin{aligned}
\mathbb{IMSE}(f,\hat{f}_n) &= \sum_{k=1}^{N} \mathbb{E}(\hat{a}_k - a_k)^2 \int_{\mathcal{X}} v_k(x)^2 dx + \sum_{k\geq N+1} \mathbb{E}(a_k)^2 \int_{\mathcal{X}} v_k(x)^2 dx \\
&= \sum_{k=1}^{N} \mathbb{V}\mathrm{ar}(\hat{a}_k) + \sum_{k=1}^{N} [\mathbb{E}(\hat{a}_k - a_k)]^2 + \sum_{k\geq N+1} a_k^2\,.
\end{aligned}$$

Finalement l'$\mathbb{IMSE}$ se décompose en

$$\int_{\mathcal{X}} \mathbb{E}[\hat{f}_n(x) - f(x)]^2 dx = \mathrm{W}_n + \mathrm{B}_n^2 + \mathrm{R}(N, f),$$

où

$$\mathrm{W}_n := \sum_{k=1}^{N} \mathbb{V}\mathrm{ar}\{\hat{a}_k\}, \qquad \mathrm{B}_n^2 := \sum_{k=1}^{N} (\mathbb{E}\{\hat{a}_k - a_k\})^2, \qquad \mathrm{R}(N, f) := \sum_{k \geq N+1} a_k^2. \tag{2.28}$$

Nous allons, à présent, majorer le terme de variance, W_n, et le terme de biais, B_n^2, à l'aide de l'inégalité de Koksma-Hlwaka généralisée (voir §2.2).

C) Majoration du terme de Variance W_n

Considérons la composante $\mathbb{V}\mathrm{ar}\{\hat{a}_k\}$ de W_n dans (2.28). Par hypothèse d'indépendance des y_i (voir (4.1)), nous avons :

$$\begin{aligned} \mathbb{V}\mathrm{ar}\{\hat{a}_k\} &= \mathbb{V}\mathrm{ar}\left\{\frac{1}{n}\sum_{i=1}^{n} v_k(x_i) y_i\right\} \\ &= \frac{1}{n^2}\sum_{i=1}^{n} v_k(x_i)^2 \mathbb{V}\mathrm{ar}(y_i) \\ &= \frac{\sigma^2}{n}\left\{\frac{1}{n}\sum_{i=1}^{n} v_k(x_i)^2\right\}. \end{aligned} \tag{2.29}$$

• Par l'Hypothèse 2, nous pouvons appliquer au membre de droite de (2.29) l'inégalité (2.12) de Koksma-Hlwaka généralisée, en considérant une fonction du système orthonormal élevée au carré $v_k^2(\cdot)$. Nous obtenons

$$\left|\frac{1}{n}\sum_{i=1}^{n} v_k(x_i)^2 - \int_{\mathcal{X}} v_k(x)^2 dx\right| \leq \left\|v_k^2(\cdot)_{\perp}\right\|_K \mathrm{D}^{L^2}(\boldsymbol{x}(n), K), \tag{2.30}$$

où $\mathrm{D}^{L^2}(\boldsymbol{x}(n), K)$ désigne la discrépance généralisée (de type L^2) associée au noyau K (cf. (2.14)), et $\left\|v_k^2(\cdot)\right\|_K$, la norme de la fonction $v_k(\cdot)^2$ induite par ce noyau (cf. (2.11)). L'inégalité précédente implique que

$$\left|\frac{1}{n}\sum_{i=1}^{n} v_k(x_i)^2\right| \leq \left|\int_{\mathcal{X}} v_k(x)^2 dx\right| + \left\|v_k^2(\cdot)_{\perp}\right\|_K \mathrm{D}^{L^2}(\boldsymbol{x}(n), K). \tag{2.31}$$

Selon les considérations et les notations introduites en A) ci-dessus, le noyau K a la forme définie par (2.14) et (2.15), et l'inégalité (2.31) se réécrit

$$\left|\frac{1}{n}\sum_{i=1}^{n} v_k(x_i)^2\right| \leq \left|\int_{\mathcal{X}} v_k(x)^2 dx\right| + \mathcal{V}_{\mathcal{M}}(v_k^2)\, \mathrm{DM}^{L^2}(\boldsymbol{x}(n)). \tag{2.32}$$

Nous posons maintenant

$$\left(VM_2^{(\boldsymbol{u})}(v_k^2)\right)^2 := \int_{[0,1]^{|\boldsymbol{u}|}} \left[\left. \frac{\partial^{|\boldsymbol{u}|}(v_k^2(x))}{\partial x^{(\boldsymbol{u})}} \right|_{x^{(\boldsymbol{u}^c)}=(1,\ldots,1)} \right]^2 dx^{(\boldsymbol{u})}. \tag{2.33}$$

Par différentiation du terme $v_k^2(x)$ dans l'intégrale (2.33), nous obtenons :

$$\begin{aligned} \left(VM^{(\boldsymbol{u})}(v_k^2)\right)^2 &= \int_{[0,1]^{|\boldsymbol{u}|}} \left[\left(2\frac{\partial^{|\boldsymbol{u}|}(v_k(x))}{\partial x^{(\boldsymbol{u})}} v_k(x) \right) \Bigg|_{x^{(\boldsymbol{u}^c)}=(1,\ldots,1)} \right]^2 dx^{(\boldsymbol{u})} \\ &= 4 \int_{[0,1]^{|\boldsymbol{u}|}} \left[\left(\frac{\partial^{|\boldsymbol{u}|}(v_k(x))}{\partial x^{(\boldsymbol{u})}} v_k(x) \right) \Bigg|_{x^{(\boldsymbol{u}^c)}=(1,\ldots,1)} \right]^2 dx^{(\boldsymbol{u})}. \end{aligned} \tag{2.34}$$

Par la définition (2.2) des $v_k(.)$ et l'Hypothèse 2, nous avons

$$\frac{\partial^{|\boldsymbol{u}|}(v_k)}{\partial x^{(\boldsymbol{u})}} \in L^2([0,1]^{|\boldsymbol{u}|}), \qquad \text{et} \qquad v_k(\cdot) \in C(\mathcal{X}).$$

Nous pouvons donc appliquer l'inégalité de Hölder à (2.34) pour obtenir les inégalités

$$\begin{aligned} \left(VM^{(\boldsymbol{u})}(v_k^2)\right)^2 &\leq 4 \left\|v_k^2\right\|_{L^\infty([0,1]^{|\boldsymbol{u}|})} \left\| \left(\left. \frac{\partial^{|\boldsymbol{u}|}(v_k)}{\partial x^{(\boldsymbol{u})}} \right|_{x^{(\boldsymbol{u}^c)}=(1,\ldots,1)} \right)^2 \right\|_{L^1([0,1]^{|\boldsymbol{u}|})} \\ &\leq 4 \sup_{x\in\mathcal{X}} \left|v_k(x)^2\right| \int_{[0,1]^{|\boldsymbol{u}|}} \left(\left. \frac{\partial^{|\boldsymbol{u}|}(v_k)}{\partial x^{(\boldsymbol{u})}} \right|_{x^{(\boldsymbol{u}^c)}=(1,\ldots,1)} \right)^2 dx^{(u)} \\ &\leq 4 \sup_{x\in\mathcal{X}} \left|v_k(x)^2\right| \left(VM^{(\boldsymbol{u})}(v_k)\right)^2. \end{aligned}$$

La *variation modifiée* d'une fonction étant définie, comme en (2.26) et (2.27), en sommant les termes de gauche et droite de cette inégalité sur tous les sous-ensembles d'indices $\boldsymbol{u} = \{u_1, \ldots, u_\ell\}$ non vides possibles de $\{1, \ldots, d\}$, on a, par continuité des $v_k(\cdot)$ sous l'Hypothèse 2,

$$\begin{aligned} \left[\sum_{\boldsymbol{u}\subset\{1,\ldots,d\}} VM^{(\boldsymbol{u})}(v_k^2) \right]^{1/2} &= \mathcal{V}_{\mathcal{M}}(v_k^2) \\ &\leq 2 \left(\sup_{x\in\mathcal{X}} |v_k(x)^2| \right)^{1/2} \left(\sum_{\boldsymbol{u}\subset\{1,\ldots,d\}} VM^{(\boldsymbol{u})}(v_k) \right)^{1/2} \\ &\leq 2 \sup_{x\in\mathcal{X}} |v_k(x)|\, \mathcal{V}_{\mathcal{M}}(v_k). \end{aligned} \tag{2.35}$$

Dans le cadre de l'estimation par projection induite par les équations (4.1), (2.2) et (2.3), nous considérons les N premières fonctions $v_1(\cdot), \ldots, v_N(\cdot)$ dans $L^2(\mathcal{X})$, où $N \geq 1$ est un entier spécifié. Ainsi, pour N fixé, posons

$$C_{\mathcal{V}_{\mathcal{M}}(v_k),N} := \max\left\{\mathcal{V}_{\mathcal{M}}(v_1), \ldots, \mathcal{V}_{\mathcal{M}}(v_N)\right\}, \quad M_v := \max\left\{\sup_{x\in\mathcal{X}}|v_1(x)|, \ldots, \sup_{x\in\mathcal{X}}|v_N(x)|\right\}.$$

Avec ces notations, et en utilisant l'Hypothèse 2 impliquant l'orthonormalité des $v_k(\cdot)$, pour $k = 1, \ldots, N$, les inégalités (2.32) et (2.35) impliquent que

$$\left|\frac{1}{n}\sum_{i=1}^{n} v_k(x_i)^2\right| \quad \leq \quad 1 + 2\, M_v\, C_{\mathcal{V}_{\mathcal{M}}(v_k),N} \mathrm{DM}^{L^2}(\boldsymbol{x}(n))\,. \tag{2.36}$$

Les relations (2.29) et (2.36) impliquent que

$$\mathbb{V}\mathrm{ar}(\hat{a}_k) \quad \leq \quad \frac{\sigma^2}{n}\,\left[1 + 2\, M_v\, C_{\mathcal{V}_{\mathcal{M}}(v_k),N} \mathrm{DM}^{L^2}(\boldsymbol{x}(n))\right]. \tag{2.37}$$

Par sommation sur k de 1 à N, le terme de variance W_n dans (2.28) est majoré par

$$\mathrm{W}_n = \sum_{k=1}^{N} \mathbb{V}\mathrm{ar}(\hat{a}_k) \quad \leq \quad \frac{N\,\sigma^2}{n}\,\left[1 + 2\, M_v\, C_{\mathcal{V}_{\mathcal{M}}(v_k),N} \mathrm{DM}^{L^2}(\boldsymbol{x}(n))\right]. \tag{2.38}$$

• De la même façon, en considérant un RKHS muni d'un noyau défini par (2.14) et (2.16), ou par (2.14) et (2.17), il est possible d'obtenir les inégalités

$$\mathrm{W}_n \quad \leq \quad \frac{N\,\sigma^2}{n}\,\left[1 + 2\, M_v\, C_{\mathcal{V}_{\mathcal{C}}(v_k),N}\ \mathrm{DC}^{L^2}(v_k)\right], \tag{2.39}$$

$$\mathrm{W}_n \quad \leq \quad \frac{N\,\sigma^2}{n}\,\left[1 + 2\, M_v\, C_{\mathcal{V}_{\mathcal{S}}(v_k),N}\ \mathrm{DS}^{L^2}(v_k)\right], \tag{2.40}$$

où $\mathrm{DC}^{L^2}(\boldsymbol{x}(n))$ désigne la discrépance L^2 centrée, $\mathrm{DS}^{L^2}(\boldsymbol{x}(n))$, la discrépance L^2 symétrique, voir le §2.2.1, et $C_{\mathcal{V}_{\mathcal{C}}(v_k),N}$ et $C_{\mathcal{V}_{\mathcal{S}}(v_k),N}$, des constantes convenables.

D) Majoration du terme de biais

Rappelons l'expression du terme de biais, B_n^2, définie en (2.28) :

$$\mathrm{B}_n^2 := \sum_{k=1}^{N} (\mathbb{E}\{\hat{a}_k - a_k\})^2.$$

Par définition (voir les équations (4.1), (2.2) et (2.3)), nous avons, pour tout $k \in \{1 \ldots N\}$,

$$\mathbb{E}(\hat{a}_k) - a_k \quad = \quad \frac{1}{n}\sum_{i=1}^{n} f(x_i)v_k(x_i) - \int_{\mathcal{X}} f(x)v_k(x)dx. \tag{2.41}$$

Par l'Hypothèse 1 et l'Hypothèse 2, l'inégalité (2.12) de Koksma-Hlwaka s'applique. Conformément aux notations (2.13) et (2.11) du §2.2.2, nous déduisons de celle-ci et de (2.41) que, pour $k \in \{1, \ldots, N\}$,

$$|\mathbb{E}(\hat{a}_k) - a_k| \;\leq\; \|(f(\cdot)v_k(\cdot))_\perp\|_K \; \mathrm{D}^{L^2}(\boldsymbol{x}(n), K). \tag{2.42}$$

Nous effectuons un raisonnement analogue à celui utilisé plus haut lors de la majoration du terme de variance W_n (inégalités (2.38), (2.39), (2.40)). La modification consiste à remplacer les fonctions $v_k^2(\cdot)$ de (2.30) par les fonctions $f(\cdot) \times v_k(\cdot)$ de (2.42). Par application de (2.12) sous les Hypothèses (1) et (2), nous obtenons

$$\mathrm{B}_n^2 \;\leq\; N\left[C_{\mathcal{V}_{\mathcal{M}}(f)}M_v + M_f C_{\mathcal{V}_{\mathcal{M}}(v_k),N}\right]^2 \mathrm{DM}^{L^2}(\boldsymbol{x}(n))^2, \tag{2.43}$$

$$\mathrm{B}_n^2 \;\leq\; N\left[C_{\mathcal{V}_{\mathcal{C}}(f)}M_v + M_f C_{\mathcal{V}_{\mathcal{C}}(v_k),N}\right]^2 \mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2, \tag{2.44}$$

$$\mathrm{B}_n^2 \;\leq\; N\left[C_{\mathcal{V}_{\mathcal{S}}(f)}M_v + M_f C_{\mathcal{V}_{\mathcal{S}}(v_k),N}\right]^2 \mathrm{DS}^{L^2}(\boldsymbol{x}(n))^2. \tag{2.45}$$

E) Majoration de l'$\mathbb{IMSE}$

D'après (2.28), par sommation des expressions obtenues ci-dessus en (2.38), (2.39), (2.40), et, en (2.43), (2.44), (2.45), respectivement, comme majorations du terme de variance W_n, et de biais, B_n^2, nous aboutissons à l'inégalité (2.23), soit

$$\begin{aligned}\mathbb{IMSE}(f, \hat{f}_n) \;\leq\; & \frac{N\sigma^2}{n}\left[1 + 2\,M_v\,C_{V(v)}\;\mathrm{D}^{L^2}\left(\boldsymbol{x}(n), K\right)\right] \\ & + N\left[C_{V(f)}M_v + M_f C_{V(v)}\right]^2\;\mathrm{D}^{L^2}(\boldsymbol{x}(n), K)^2 \\ & + R(N, f).\end{aligned} \tag{2.46}$$

□

2.3.3 Majoration de la $\mathbb{MSE}$

Dans ce qui suit, nous considérerons que la première fonction du système de fonction $\{v_1(\cdot), \ldots, v_N(\cdot)\} \subset L^2(\mathcal{X})$ utilisé dans (2.2) et (2.3) est la fonction constante 1, $v_1 = 1$ (ce qui est rendu possible par (2.10)). Nous supposerons aussi que la fonction $f(\cdot)$ vérifie l'Hypothèse 1 avec $p = 2$ dans (2.4).

En faisant usage, soit d'une approche bayésienne, soit de l'utilisation de l'inégalité de Koksma-Hlwaka généralisée (voir Hickernell (1999)), il est possible de majorer le critère, dit de *robustesse*, $\mathbb{MSE}(\hat{a}_1, a_1)$ en (2.21), par

$$\mathbb{MSE}(\hat{a}_1, a_1) \;\leq\; \frac{\sigma^2}{n} + \left[\mathrm{D}^{L^2}\left(\boldsymbol{x}(n), K\right) V(f)\right]^2, \tag{2.47}$$

où $V(f)$ désigne une *variation de* $f(\cdot)$ et $\mathrm{D}^{L^2}(\boldsymbol{x}(n), K)$ une *discrépance généralisée* de type L^2 (voir §2.2.2 équation (2.13)). Nous détaillons ci-dessous les arguments correspondants, en suivant Hickernell (1999).

Ici, l'Hypothèse 2 portant sur le choix des $v_k(\cdot) \in L^2(\mathcal{X})$, $k = 1, \ldots, N$, n'est pas nécessaire.

Démonstration (de l'inégalité (2.47)) :

A) *Rappel des Hypothèses*

Pour les définitions de $f(\cdot)$ et $v_k(\cdot)$, $k \geq 1$, nous renvoyons à (4.1), (2.2) et (2.3). Nous supposerons que la fonction $f(\cdot)$ vérifie l'Hypothèse 1 avec $p = 2$ dans (2.4). Ainsi $f(\cdot)$ appartient à un RKHS $\mathcal{H}$ muni d'un noyau auto-reproduisant K. Nous nous limiterons aux noyaux K définis par (2.14) et (2.15), (2.14) et (2.16), (2.14) et (2.17). L'Hypothèse 1 permettra d'appliquer l'inégalité (2.12) de Koksma-Hlwaka généralisée à $f(\cdot)$.

La première fonction du système orthogonal $\{v_1(\cdot), \ldots, v_N(\cdot)\}$ de $L^2(\mathcal{X})$ sera supposée égale à la fonction constante 1, $v_1 = 1$. Ceci est rendu possible par (2.10).

B) *Décomposition de la* $\mathbb{MSE}$

L'erreur quadratique moyenne de l'estimateur $\hat{a}_1$ (voir (2.2) et (2.3)) se décompose classiquement en la somme d'un terme faisant intervenir son biais (au carré) et sa variance, sous la forme,

$$\mathbb{E}[\hat{a}_1 - a_1]^2 \;=\; \mathbb{E}\left[\hat{a}_1 - \mathbb{E}(\hat{a}_1)\right]^2 + \left[\mathbb{E}(\hat{a}_1) - a_1\right]^2 = \mathbb{V}\mathrm{ar}(\hat{a}_1) + \left[\mathbb{E}(\hat{a}_1) - a_1\right]^2 . \quad (2.48)$$

C) Étude du terme de variance

En utilisant l'hypothèse d'indépendance des y_i (voir (4.1)), et le fait que $v_1 = 1$, d'après (2.3) nous avons,

$$\begin{aligned} \mathbb{E}\left[\hat{a}_1 - \mathbb{E}(\hat{a}_1)\right]^2 &= \mathbb{V}\mathrm{ar}(\hat{a}_1) = \mathbb{V}\mathrm{ar}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}\mathrm{ar}(y_i) = \frac{\sigma^2}{n} . \end{aligned} \quad (2.49)$$

D) Majoration du terme de biais

Par hypothèse de modélisation (équations (4.1), (2.2) et (2.3)), nous avons $\mathbb{E}(y_i) = f(x_i)$, pour $i = 1, \ldots, n$. On a donc

$$\mathbb{E}(\hat{a}_1) = \frac{1}{n}\sum_{i=1}^{n} f(x_i), \qquad \text{et} \qquad a_1 = \int_{\mathcal{X}} f(x)dx\,.$$

Ainsi, le terme faisant intervenir le biais au carré de l'estimateur $\hat{a}_1$ dans (2.48) s'écrit :

$$[\mathbb{E}(\hat{a}_1) - a_1]^2 \;=\; \left[\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \int_{\mathcal{X}} f(x)dx\right]^2. \tag{2.50}$$

D'après l'Hypothèse 1, il est possible d'appliquer à (2.50) l'inégalité (2.12) de Koksma-Hlwaka généralisée, pour obtenir,

$$[\mathbb{E}(\hat{a}_1) - a_1]^2 \;=\; \left[\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \int_{\mathcal{X}} f(x)dx\right]^2 \leq \left[\mathrm{D}^{L^2}(\boldsymbol{x}(n), K)\ \|f(\cdot)_{\perp}\|_K\right]^2 \tag{2.51}$$

E) Majoration de la $\mathbb{MSE}$

D'après (2.48), par sommation des termes majorant, le biais élevé au carré (cf. (2.51)), et la variance de $\hat{a}_1$ (cf. (2.49)), nous aboutissons à (2.47).

□

2.3.4 Interprétation

Les inégalités (2.23) et (2.47) montrent que les termes dominant les critères d'$\mathbb{IMSE}$ et de *robustesse*, $\mathbb{MSE}$, peuvent être majorés par un terme dépendant de la discrépance généralisée. Ces majorations montrent que la caractérisation de la disposition des points dans l'espace $\mathcal{X} = [0,1)^d$, exprimée par les discrépances généralisées, influe sur la qualité (au sens de l'$\mathbb{IMSE}$ et de la $\mathbb{MSE}$) de l'estimation de $f(\cdot)$.

Pour pouvoir utiliser au mieux cette propriété selon les données disponibles, il est nécessaire de définir un cadre méthodologique, ce qui est fait dans le §2.4 ci-dessous.

2.4 Cadre méthodologique

Rappelons que nous nous plaçons dans le contexte de l'estimation d'un paramètre fonctionnel $f(\cdot)$ par la méthode des fonctions orthogonales décrite en (4.1), (2.2) et (2.3). L'objectif de ce paragraphe sera la *construction*, la *sélection*, ou la *spécification* d'un ensemble de points $\boldsymbol{x}(n) \in \mathcal{X}^n$ apte à fournir une estimation $\hat{f}_n(\cdot)$ satisfaisante de $f(\cdot)$. Les critères considérés pour apprécier la qualité de l'estimation $\hat{f}_n(\cdot)$ seront l'$\mathbb{IMSE}$ et la

$\mathbb{MSE}$ présentés en (2.21) du §2.3.

Les inégalités (2.23) et (2.47) montrent que, sous les Hypothèses 1 et 2, l'$\mathbb{IMSE}$ et la $\mathbb{MSE}$ sont majorés par un terme faisant intervenir la discrépance de $\boldsymbol{x}(n)$ noté $\mathrm{D}^{L^2}(\boldsymbol{x}(n), K)$ (voir la définition (2.13) §2.2.2). Plus précisément, leur étude permet de formuler les remarques suivantes.

Remarque 1
Lorsque n augmente et $\mathrm{D}^{L^2}(\boldsymbol{x}(n), K)$ diminue, l'$\mathbb{IMSE}$ et la $\mathbb{MSE}$ diminuent. Signalons de plus que lorsque $n \to \infty$, sous certaines hypothèses supplémentaires, sur la fonction $f(\cdot)$, les fonctions $v_k(\cdot)$ et l'ordre de troncature N en fonction de n, l'utilisation de *suites à discrépance faible* assure une vitesse de convergence optimale des estimations $\hat{f}_n(\cdot)$ vers $f(\cdot)$ au sens de l'$\mathbb{IMSE}$ (Stone (1982) et Rafajlowicz et Schwabe (2005)).

Remarque 2
Lorsque n est constant et que $\mathrm{D}^{L^2}(\boldsymbol{x}(n), K)$ diminue, l'$\mathbb{IMSE}$ et la $\mathbb{MSE}$ diminuent.

Remarque 3
Lorsque n diminue et $\mathrm{D}^{L^2}(\boldsymbol{x}(n), K)$ diminue, l'étude des inégalités (2.23) et (2.47) ne permet pas de connaître l'évolution de la $\mathbb{MSE}$ et de l'$\mathbb{IMSE}$. En effet, elles font toutes deux intervenir un terme en $1/n$ qui augmente dans ce contexte.

Le cadre méthodologique défini ci-dessous permet de prendre en compte ces remarques de façon pratique. Nous distinguerons différents cas de figure en fonction des données (« points d'observation », « observations ») disponibles, c'est-à-dire des (x_i, y_i), $i = 1, \ldots, n$ selon nos notations, cf. (4.1)).

Cas 1 *Absence initiale de la base de données $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ dans $[0,1)^d$*
Compte tenu de la Remarque 1, lorsqu'on peut maîtriser le choix des points $\boldsymbol{x}(n)$ dans $\mathcal{X} = [0,1)^d$, il est préférable de choisir une *suite à discrépance faible.*

Cas 2 *Choix d'un nombre « imposé » de points dans une base de données $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$*
Lorsqu'on doit choisir un ensemble $\boldsymbol{x}_1(n_1)$ de n_1 points dans un ensemble de points candidats $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$, on pourra commencer par étudier la qualité de la base $\boldsymbol{x}(n)$ à l'aide de critères faisant intervenir la discrépance (voir le §2.5.2). Compte tenu de la Remarque 2, on sélectionnera l'ensemble de points dont la quantité $D^{L^2}(\boldsymbol{x}_1(n_1))$ est la plus faible. Là encore, une étude de l'ensemble des points sélectionnés sera menée (cf. §2.5.2).

Cas 3 *Choix « libre » de points dans une base de données $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$*
Nous devons ici choisir un ensemble de points $\boldsymbol{x}_1(n_1)$ dans un ensemble $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ (n_1 n'est pas imposé). Compte tenu de la Remarque 3, ce

choix est délicat. Nous proposons la sélection de points $\boldsymbol{x}_1(n_1)$ parmi $\boldsymbol{x}(n)$ qui permettent de réduire la quantité $Q(\boldsymbol{x}(n)) := \mathrm{D}^{L^2}(\boldsymbol{x}(n), K)/n$. Le nombre n_1 de points sélectionnés sera plus important que si l'on cherche simplement à diminuer la quantité $\mathrm{D}^{L^2}(\boldsymbol{x}(n), K)$. Ainsi, on limitera l'augmentation du terme en $1/n$ dans les majorations de l'$\mathbb{IMSE}$ et de la $\mathbb{MSE}$ décrites par les inégalités (2.23) et (2.47). De plus, la diminution de $Q(\boldsymbol{x}(n))$ implique la diminution de $\mathrm{D}^{L^2}(\boldsymbol{x}(n), K)$. En effet, si

$$\frac{D^{L^2}(\boldsymbol{x}(n-1), K)}{n-1} \quad \leq \quad \frac{\mathrm{D}^{L^2}(\boldsymbol{x}(n), K)}{n},$$

alors,

$$D^{L^2}(\boldsymbol{x}(n-1), K) \quad \leq \quad \frac{n-1}{n}\, \mathrm{D}^{L^2}(\boldsymbol{x}(n), K),$$

ce qui implique,

$$D^{L^2}(\boldsymbol{x}(n-1), K) \quad \leq \quad \mathrm{D}^{L^2}(\boldsymbol{x}(n), K)^2. \tag{2.52}$$

Bien entendu, une étude préalable de la qualité de $\boldsymbol{x}(n)$, puis de $\boldsymbol{x}_1(n_1)$ à l'aide de critères faisant intervenir la discrépance sera menée (voir §2.5.2).

Cas 4 *Données* (x_i, y_i), $i = 1, \ldots, n$ *disponibles*

Il est d'usage d'appeler les couples (x_i, y_i), $i = 1, \ldots, n$ des données (point d'observation, observation). Tout d'abord une estimation $\hat{f}_n(\cdot)$ de $f(\cdot)$ à l'aide des données initiales (x_i, y_i), $i = 1, \ldots, n$, sera réalisée conformément à (2.3). Parmi les données (x_i, y_i), $i = 1, \ldots, n$, nous chercherons un sous-ensemble de nature à fournir une estimation satisfaisante $\hat{f}_{n_1}(\cdot)$ de $f(\cdot)$ (au sens de la $\mathbb{MSE}$ et de l'$\mathbb{IMSE}$). Dans ce contexte, nous sélectionnerons un sous-ensemble $\boldsymbol{x}_1(n_1) \subset \boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ de la même façon que celle décrite au cas 3. L'estimation $\hat{f}_{n_1}(\cdot)$ sera réalisée à l'aide des points de $\boldsymbol{x}_1(n)$ et des observations correspondantes (selon (2.3)). On comparera les estimations $\hat{f}_n(\cdot)$ et $\hat{f}_{n_1}(\cdot)$ de $f(\cdot)$. Ceci se fera à l'aide de critères comme par exemple celui de l'erreur quadratique entre les observations et les estimations aux points x_i, $i = 1, \ldots, n$, (voir le §2.5.4). L'estimation de $f(\cdot)$ qui est la meilleure, au sens de ces critères, sera retenue.

Lorsqu'une *spécification* de points dans $\mathcal{X}$ est possible (au sens de l'ajout contrôlé de nouvelles données (point d'observation, obervation)), compte tenu de la remarque 2, nous choisirons des points additionnels permettant de réduire la discrépance. Ces points *spécifiés* pourront, par exemple, être choisis parmi ceux d'une suite à discrépance faible.

2.5 Application

2.5.1 Présentation de l'exemple

Pour l'application, le cas 4 de la méthodologie est abordé ci-dessous. Des observations $y_i \in \mathbb{R}$ associées à des points $x_i \in \mathcal{X} = [0,1)^2$, $i = 1, \ldots, 100$ sont disponibles. L'ensemble $\boldsymbol{x}(100) = \{x_1, \ldots, x_{100}\}$ est représenté dans la Figure 2.1.

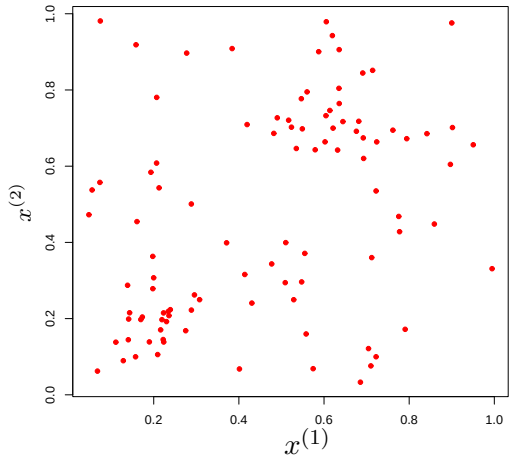


FIG. 2.1 – Ensemble $\boldsymbol{x}(100)$ initial

Les observations $y_1, \ldots, y_{100}$ corespondent à la modélisation :

$$y(x_i) \quad = \quad f(x_i) + \varepsilon_i, \quad \text{avec,} \tag{2.53}$$

- $x_i \in \mathcal{X} = [0,1)^2$,
- $f \in L^2(\mathcal{X})$, le paramètre fonctionnel à estimer,
- ε_i, $i = 1, \ldots, 100$, une suite de variables aléatoires indépendantes de moyenne nulle de variance σ^2.

La fonction $f(\cdot)$ considérée pour cette application est une somme de produits tensoriels de polynômes de Legendre et s'écrit comme suit,

$$\text{pour } x^{(j)} \in [0,1) : \qquad \phi_0(x^{(j)}) = 1, \qquad \phi_1(x^{(j)}) = 12\,(x^{(j)} - 1/2),$$

$$\begin{aligned} \text{pour } \; x = (x^{(1)}, x^{(2)}) \in \mathcal{X} = [0,1)^2 & \\ f(x^{(1)}, x^{(2)}) = a_{0,0}\,\phi_0(x^{(1)})\phi_0(x^{(2)}) &+ a_{0,1}\,\phi_0(x^{(1)})\phi_1(x^{(2)}) \\ + a_{1,0}\,\phi_1(x^{(1)})\phi_0(x^{(2)}) &+ a_{1,1}\,\phi_1(x^{(1)})\phi_1(x^{(2)}). \end{aligned}$$

Les coefficients $a_{0,0}$, $a_{0,1}$, $a_{1,0}$, et $a_{1,1}$, sont estimés par la méthode des moments introduite par la formule (2.3).

Les variables aléatoires ε_i, $i = 1, \ldots, 100$, utilisées pour la simulation sont indépendantes et de loi normale centrée $\mathcal{N}(0, \sigma^2)$, d'écart type $\sigma = 0.4$.

2.5.2 Analyse initiale des points disponibles

Selon la méthodologie définie au paragraphe 2.4, il est nécessaire de commencer par étudier la « qualité » de l'ensemble $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$. Pour ce faire la discrépance de type L^2 centrée (voir les équations (2.13), (2.14), (2.16)), notée DC^{L^2}, sera utilisée comme critère de qualité. Pour le calcul de cette quantité, la formule introduite par Hickernell (1998) est utilisée,

$$\begin{aligned}\left[\mathrm{DC}^{L^2}(\boldsymbol{x}(n))\right]^2 &= \left(\frac{13}{12}\right)^d - \frac{2}{n}\sum_{i=1}^{n}\prod_{k=1}^{d}\left(1+\frac{1}{2}|1+x_i^{(k)}-1/2| - \frac{1}{2}|1+x_i^{(k)}-1/2|^2\right) \\ &\quad + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\prod_{k=1}^{d}\left(1+\frac{1}{2}|1+x_i^{(k)}-1/2| + \frac{1}{2}|1+x_j^{(k)}-1/2| \right. \\ &\qquad \left. - \frac{1}{2}|x_i^{(k)}-x_j^{(k)}|\right),\end{aligned}$$

avec ici, $d = 2$, $n = 100$, et $x_i = (x_i^{(1)}, x_i^{(2)}) \in [0,1)^2$.

Par définition, la discrépance s'interprète comme une comparaison entre le nombre de points compris dans certains pavés de $\mathcal{X} = [0,1)^d$ et le volume de ces pavés. Il s'agit donc d'un critère de répartition uniforme de l'ensemble des points $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ dans $\mathcal{X} = [0,1)^2$. On parle encore de critère de « remplissage de l'espace » ou « *space filling* ».

Pour que l'ensemble $\boldsymbol{x}(n)$ soit jugé de qualité « acceptable », une première approche consisterait à effectuer des *tests statistiques*. L'hypothèse $H0$ à tester serait : « les x_i sont des réalisations de variables aléatoires indépendantes et de même loi uniforme sur $\mathcal{X} = [0,1)^d$ », et la statistique utilisée serait : $\mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2$. A notre connaissance, la loi de cette statistique n'est pas connue de façon exacte à ce jour. Il n'est donc pas aisé d'effectuer de tels tests en pratique. Cependant son espérance (Hickernell (1996a) et Hickernell (1998)) se calcule et vaut

$$\mathbb{E}\left[\mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2\right] = \frac{1}{n}\left[\left(\frac{13}{12}+\frac{1}{6}\right)^d - \left(\frac{13}{12}\right)^d\right]. \tag{2.54}$$

Puisque, par définition, plus la discrépance d'un ensemble de points $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ dans un espace $\mathcal{X} = [0,1)^d$ est faible, meilleure est la qualité du recouvrement des points

$\boldsymbol{x}(n)$ dans $\mathcal{X}$, l'espérance (2.54) constituera notre valeur seuil *supérieur* de référence. Nous jugerons de « qualité acceptable » un ensemble de points $\boldsymbol{x}(n)$ dont $\mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2 < E\left[\mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2\right]$. Notons que, selon nos simulations, lorsque $\boldsymbol{x}(n)$ correspond à un ensemble de v.a. i.i.d uniformes, $\mathbb{E}\left[\mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2\right]$ correspond au quantile à 61% de $\mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2$:

$$\mathbb{P}\left(\mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2 \leq \mathbb{E}\left[\mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2\right]\right) \approx 0.61.$$

La densité de la statistique $n \times \mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2$ obtenue par simulation est représentée dans la Figure 2.2 (la multiplication par n permet de ne pas faire dépendre la loi de la statistique du nombre de points).

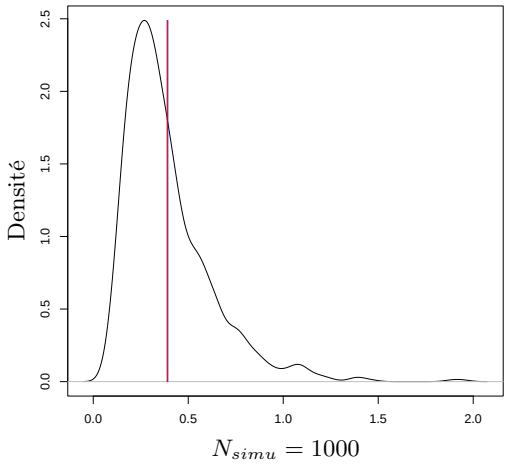


FIG. 2.2 – Densité de $n \times \mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2$, la moyenne est représentée par le trait vertical

Pour l'ensemble initial $\boldsymbol{x}(100) = \{x_1, \ldots, x_{100}\}$, nous avons

$$[\mathrm{DC}^{L^2}(\boldsymbol{x}(n))]^2 = 0.0100 \quad > \quad \mathbb{E}\left[\mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2\right] = 0.0038\,.$$

L'ensemble $\boldsymbol{x}(100) = \{x_1, \ldots, x_{100}\}$ n'est donc pas jugé de qualité acceptable. La Figure 2.10 montre aussi que l'ajustement de l'estimation de $f(\cdot)$ par la méthode de projection sur une base de fonctions orthonormales n'est pas satisfaisant lorsque les points de $\boldsymbol{x}(100) = \{x_1, \ldots, x_{100}\}$ sont pris en compte.

2.5.3 Sélection d'un sous-ensemble de points

Selon le cas 4 de la méthodologie introduite au paragraphe 2.4, nous proposons de sélectionner un sous-ensemble $\boldsymbol{x}_1(n_1) \subset \boldsymbol{x}(n) = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ de façon à réduire la

quantité

$$Q(\boldsymbol{x}(n)) \quad = \quad \frac{\mathrm{DC}^{L^2}(\boldsymbol{x}(n))}{n}\,. \tag{2.55}$$

Ici, $\mathrm{DC}^{L^2}(\boldsymbol{x}(n))$ désigne la discrépance de type L^2 centrée (voir les équations (2.13), (2.14), (2.16)). Pour effectuer cette sélection, plusieurs méthodes sont possibles et vont être illustrées.

Nous utiliserons les méthodes présentées au paragraphe 1.4.2.1, 1.4.2.2, 1.4.2.3 du chapitre I, en considérant ici la réduction du critère $Q(\boldsymbol{x}(n_1))$ (cf. (2.55)).

• **Application de l'Algorithme** A_1

La première méthode (cf. l'algorithme A_1 de §1.4.2.1) consiste à trouver un sous-ensemble $\boldsymbol{x}_1(n_1)$ de $\boldsymbol{x}(100)$ dont $Q(\boldsymbol{x}(n_1))$ est minimale (voir (2.55)).

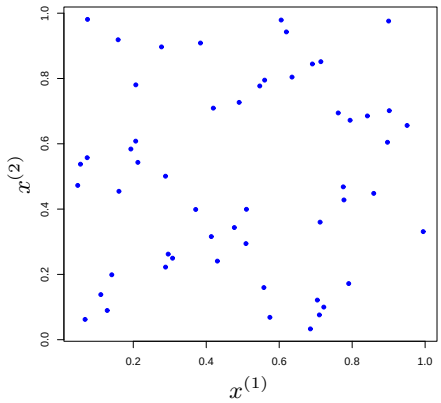


FIG. 2.3 – Points sélectionnés par l'algorithme A_1

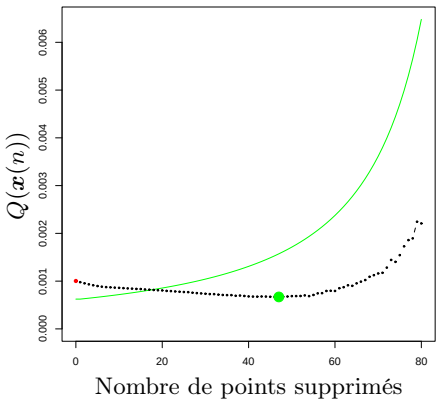


FIG. 2.4 – Évolution de Q en fonction du nombre de points supprimés par l'algorithme A_1

L'évolution de la quantité Q en fonction du nombre de points supprimés est représentée Figure 2.4. La courbe en vert représente $\sqrt{\mathbb{E}\left[\mathrm{DC}^{L^2}(\boldsymbol{x}(n))^2\right]}/n$ qui constitue une valeur seuil de référence (voir §2.5.2). L'ensemble de points retenu sera celui dont la quantité Q est la plus faible. Il comporte 54 points. On remarque toutefois, qu'à partir de 80 points il est possible de sélectionner un ensemble de points de qualité « acceptable ».

• Application de l'Algorithme A_2

La deuxième méthode (utilisation de l'algorithme A_2 du §1.4.2.2) consiste simplement à sélectionner des points proches de ceux d'une *suite à discrépance faible*. Ces suites « déterministes » sont essentiellement utilisées pour les méthodes de quasi-Monte Carlo dont l'objectif est de réduire au mieux la discrépance à l'origine (voir le §2.2). Pour l'étude de ces suites, nous renvoyons, par exemple, à Niederreiter (1992).

L'évolution de la discrépance en fonction du nombre de points supprimés est présentée dans la Figure 2.6. A partir de 80 points, il est possible d'obtenir un ensemble de qualité acceptable selon nos critères (voir le §2.5.2). L'ensemble de points retenu par cet algorithme comporte 67 points. Il est illustré dans la Figure 2.5. Entre 67 points et 45 points, la valeur de Q obtenue évolue relativement peu.

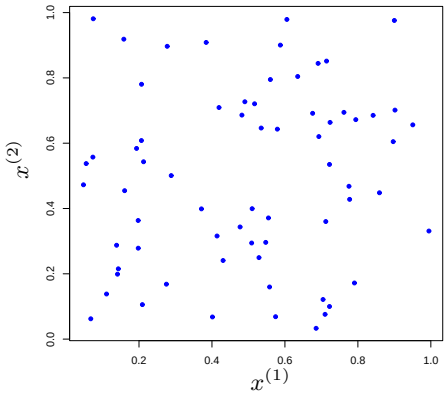


FIG. 2.5 – Points sélectionnés par l'algorithme A_2

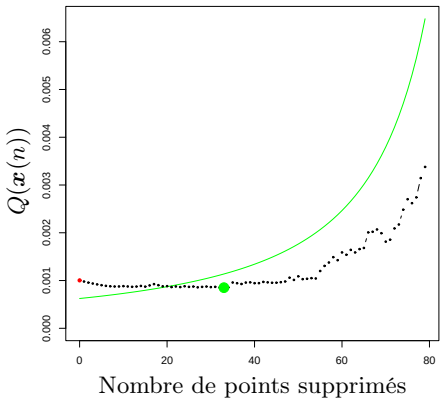


FIG. 2.6 – Évolution de Q en fonction du nombre de points supprimés par l'algorithme A_2

• Application de l'Algorithme A_3

L'objectif de la troisième méthode est d'extraire un sous-ensemble de points régulièrement répartis dans $\mathcal{X} = [0,1)^2$ (voir l'algorithme A_3 de §1.4.2.3 avec ici la réduction du critère $Q(\boldsymbol{x}(n_1))$, cf (2.55).).

Les Figures 2.8 et 2.9 représentent, l'évolution de la discrépance en fonction du nombre de points supprimés, et de la distance ε fixée. L'ensemble de points retenu comporte 48 points et est illustré Figure 2.7. Nous remarquons qu'à partir de 75 points il est possible d'obtenir un ensemble de points de qualité acceptable selon nos critères (voir §2.5.2).

En général, cet algorithme permet de trouver un sous-ensemble de points de qualité

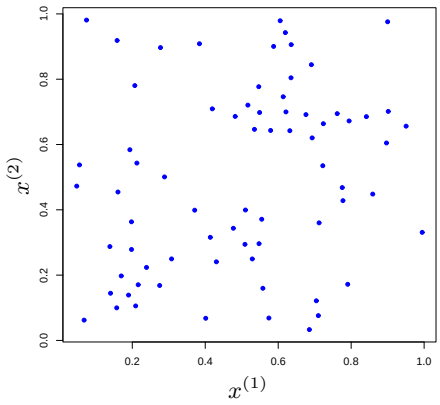


FIG. 2.7 – Points sélectionnés par l'algorithme A_3

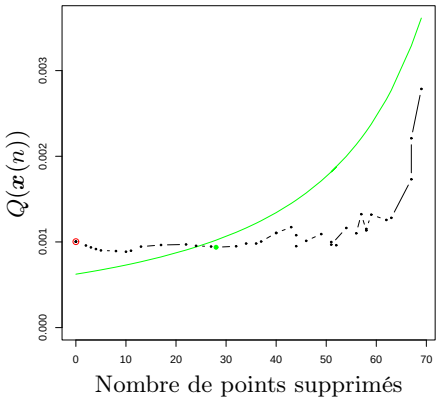


FIG. 2.8 – Évolution de Q en fonction du nombre de points supprimés par l'algorithme A_3

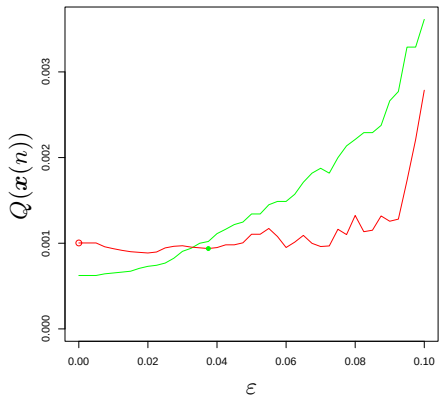


FIG. 2.9 – Évolution de Q en fonction de la distance ε par l'algorithme A_3

acceptable. Le critère Q diminue, cependant, moins rapidement et moins fortement que lors de l'application de l'algorithme A_1 (voir les Figures 2.4 et 2.8).

2.5.4 Estimation et Validation

Comparons à présent les différentes estimations de $f(\cdot)$, désignées par $\hat{f}_{ini}(\cdot)$, $\hat{f}_{A_1}(\cdot)$, $\hat{f}_{A_3}(\cdot)$, $\hat{f}_{A_2}(\cdot)$, obtenues, respectivement, avec l'ensemble $\boldsymbol{x}(n)$ de points initiaux, l'ensemble $\boldsymbol{x}_{A_1}(n_1)$ de points sélectionnés par l'algorithme A_1, l'ensemble $\boldsymbol{x}_{A_3}(n_3)$ de points sélectionnés par l'algorithme A_3, l'ensemble $\boldsymbol{x}_{A_2}(n_2)$ de points sélectionnés par l'algorithme A_2.

Pour la validation des estimations de $f(\cdot)$, nous avons introduit un nouvel ensemble $\boldsymbol{x}_{val}(100)$ de 100 points correspondant à des réalisations de variables aléatoires indépendantes uniformes dans $[0,1)^2$. Pour $i = 1, \ldots, 100$, des observations $y_{val_i} = f(x_i) + \varepsilon_i$, où $\varepsilon_i \sim \mathcal{N}(0, 0.4)$, ont été simulées aux points $x_{val_i} \in \boldsymbol{x}_{val}(100)$. Les Figures 2.10, 2.11, 2.12, 2.13, représentent, respectivement, les estimations $\hat{f}_{ini}(x_{val_i})$, $\hat{f}_{A_1}(x_{val_i})$, $\hat{f}_{A_3}(x_{val_i})$, $\hat{f}_{A_2}(x_{val_i})$, de $f(x_{val_i})$, rapportés aux observations « de validation », y_{val_i}, $i = 1, \ldots, 100$. Le Tableau 2.1 récapitule l'ensemble des résultats obtenus lors de l'étape de sélection de points, ainsi que l'erreur quadratique estimée par la formule

$$EQ_{A_j} = \frac{1}{100} \sum_{x_i \in \boldsymbol{x}_{val}(100)} \left(\hat{f}_{A_j}(x_i) - y_i \right)^2 \quad \text{pour } A_j = A_1, A_3, A_2.$$

Ensemble de points	Nb. points	DC^{L^2}	$\mathbb{E}(\mathrm{DC}^{L^2})$	Q	EQ
$\boldsymbol{x}_{A_1}(n_1)$	54	1.3E-3	7.2E-3	6.6E-4	0.22
$\boldsymbol{x}_{A_2}(n_2)$	67	3.0E-3	5.8E-3	8.1E-4	0.34
$\boldsymbol{x}_{A_3}(n_3)$	72	4.6E-3	5.4E-3	9.4E-4	0.66
$\boldsymbol{x}(n)$	100	10E-2	3.8E-3	10E-4	1.37

TAB. 2.1 – Tableau comparatif des résultats des différents algorithmes et des ajustements associés

Le Tableau 2.1 montre que l'algorithme A_1 donne de meilleurs résultats que les algorithmes A_3 et A_2. Ceci n'est pas surprenant car l'objectif de A_1 est la minimisation directe du critère Q.

Comme l'objectif principal de l'algorithme A_3 n'est pas la minimisation du critère Q mais la sélection d'un ensemble de points régulièrement répartis dans $\mathcal{X}$, celui-ci donne de moins bons résultats. De plus, le nombre de points sélectionnés par l'algorithme A_3 est plus important que ceux des algorithmes A_1 et A_2. Ainsi, l'ensemble $\boldsymbol{x}_{A_3}(n_3)$ conservera un nombre relativement important de points de l'ensemble initial $\boldsymbol{x}(n)$. Comme le

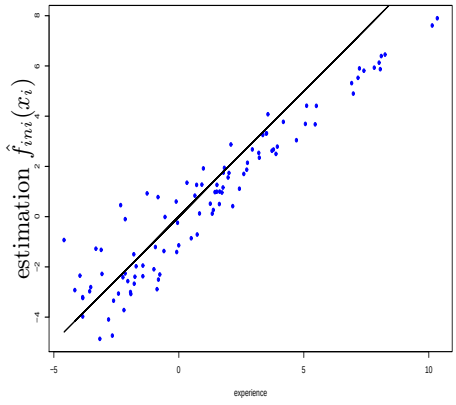


FIG. 2.10 – Validation de l'estimation $\hat{f}_{ini}(\cdot)$

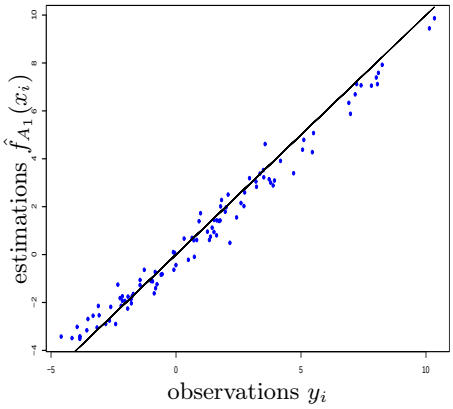


FIG. 2.11 – Validation de l'estimation $\hat{f}_{A_1}(\cdot)$

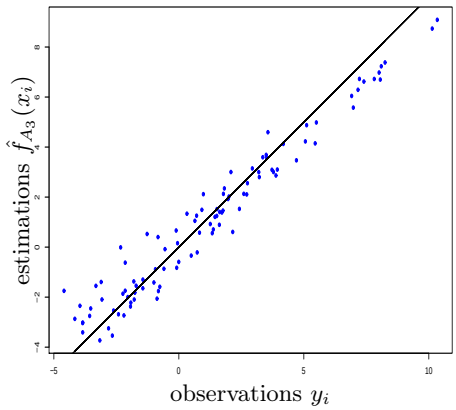


FIG. 2.12 – Validation de l'estimation $\hat{f}_{A_3}(\cdot)$

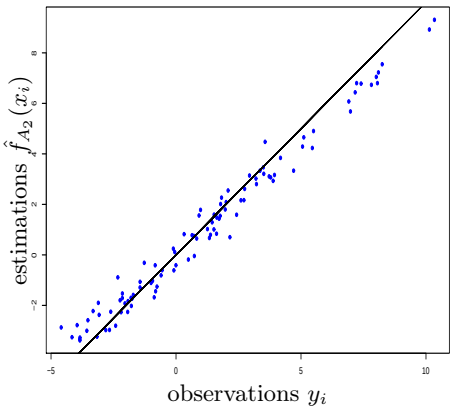


FIG. 2.13 – Validation de l'estimation $\hat{f}_{A_2}(\cdot)$

caractère de répartition uniforme de l'ensemble initial n'est pas jugé « acceptable » selon nos critères, le fait de sélectionner, parmi cet ensemble, une partie de cardinalité trop élevée risque de conserver ce manque d'uniformité. Ainsi $\boldsymbol{x}_{A_3}(n_3)$, risquera lui aussi de ne pas avoir une bonne « répartition ». Ceci est notamment confirmé par un critère sevère d'uniformité tel que celui de la discrépance L^2 centrée, voir le Tableau 2.1.

L'algorithme A_2 permet, en général, d'extraire un ensemble de points « proche » d'une *suite à discrépance faible* et donc de bénéficier parfois de caractéristiques semblables (recouvrement uniforme de l'espace). Il est également moins coûteux en temps de calcul, ce qui le rend particulièrement intéressant lorsque la dimension de l'espace $\mathcal{X}$ est élevée.

La sélection d'un ensemble de points par minimisation du critère $Q = \mathrm{DC}^{L^2}(\boldsymbol{x}(n))/n$ prend en compte l'inégalité (2.23). Comme précisé au §2.4, la division par n contraint la sélection à conserver un nombre de points relativement important. Bien entendu, il est aussi possible d'appliquer cette méthodologie en cherchant simplement à minimiser la discrépance centrée DC^{L^2}. Le nombre de points sélectionné sera alors inférieur à celui traité par le critère Q. Les résultats obtenus seront alors comparables.

2.6 Discussion

En nous appuyant sur les travaux de Hickernell (1998) et Rafajlowicz et Schwabe (2005) nous avons précisé les liens entre les critères d'$\mathbb{IMSE}$ et de $\mathbb{MSE}$ et la notion de discrépance généralisée, dans le cadre de l'estimation d'un paramètre fonctionnel basée sur sa décomposition à partir d'une somme de fonctions orthonormales. Les critères d'$\mathbb{IMSE}$ et de $\mathbb{MSE}$ peuvent tous deux être majorés par un terme faisant intervenir la discrépance généralisée.

De façon générale, une « discrépance » faible d'une suite de points $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ en lesquels sont observés $\{y_1, \ldots, y_n\}$ permet d'obtenir une estimation *robuste*, au sens de la $\mathbb{MSE}$, et de bonne qualité au sens de l'$\mathbb{IMSE}$. La méthodologie proposée au chapitre I et adaptée au contexte de l'estimation par la méthode des fonctions orthogonales trouve donc ici une justification théorique.

Le lien entre la discrépance et la méthode des fonctions orthogonales a été établi à l'aide de l'inégalité de Koksma-Hlwaka généralisée. D'origine récente (Hickernell (1998)), cette dernière permet de définir la discrépance à partir d'un noyau auto-reproduisant convenable d'un espace de Hilbert. Lorsqu'une modélisation par un processus aléatoire est utilisée, cela revient aussi à considérer des espaces de Hilbert où le noyau auto-reproduisant correspond à la fonction de covariance du processus. Il semble donc pertinent d'appliquer la notion de discrépance en adaptant la définition à la fonction de covariance comme mentionné par Hickernell (1999).

D'autre part, lorsqu'une méthode d'estimation fera intervenir des moyennes, approxi-

mations d'intégrale, l'inégalité de Koksma-Hlwaka généralisée montre que la diminution de la discrépance d'un ensemble de points permettra de réduire l'erreur de l'estimation. Cette propriété pourrait aussi être employée dans le contexte de l'*apprentissage statistique* où l'on considère une fonction de risque empirique, approximation d'une intégrale (ceci fait notamment l'objet de recherches récentes, voir Cervellera et Muselli (2004), Marry (2005)). Une perspective consisterait donc à utiliser les méthodes de sélection et de spécification de points du chapitre I dans la procédure d'apprentissage.

Chapitre 3

Critères probabilistes

Dans ce chapitre, nous approfondirons les méthodes destinées à l'étape 2 de la méthodologie définie en Introduction. Comme pour le chapitre I, l'objectif sera de vérifier qu'un ensemble de points est réparti de façon uniforme dans un espace $\mathcal{X}$. L'approche sera cependant différente de celle du chapitre I où les techniques utilisées étaient dictées par des considérations déterministes. Le caractère aléatoire éventuel des données de la BDDE (cf. Introduction) sera pris en compte. Ici, nous parlerons de répartition uniforme des points dans $\mathcal{X}$ lorsque ces points peuvent être considérés comme des *variables aléatoires indépendantes et de loi de probabilité uniforme* dans $\mathcal{X}$. Pour accepter ou rejeter cette dernière hypothèse nous aurons donc recours aux *tests statistiques*.

Nous nous intéresserons au cas où la dimension d de $\mathcal{X}$ est strictement supérieure à 1. Ainsi, certaines méthodes classiques pour $d = 1$ ne peuvent plus être employées. En effet, pour certains tests, les lois des *statistiques* (fonctions des variables considérées dans $\mathcal{X}$) prises en compte lorsque $d = 1$ ne sont pas connues de façon explicite en dimension supérieure (par exemple, les statistiques de Kolmogorov-Smirnov, de Cramer-Von-Mises). De plus, le calcul de ces statistiques est parfois délicat (comme la statistique de Kolmogorov-Smirnov qui correspond alors à l'évaluation de la discrépance à l'origine, cf. chapitre I).

Les statistiques que nous proposerons d'utiliser consistent à effectuer une partition en cellules disjointes de l'espace $\mathcal{X}$ (voir la Figure 3.1 et le §3.2), puis à définir des fonctions du nombre de points contenus dans chaque cellule (cf. équation (3.18)). Les tests réalisés permettront de vérifier si la proportion de points contenus dans ces cellules correspond ou non à celle d'une répartition de variables aléatoires indépendantes et uniformes.

Dans un premier temps, nous préciserons le formalisme. Nous présenterons la notion de test statistique et la partition de $\mathcal{X} = [0, 1)^d$ que nous utiliserons. Nous remarquerons notamment que la loi de probabilité caractérisant le nombre de points par cellule (de la partition de $\mathcal{X}$) est une loi multinomiale.

Nous rappellerons ensuite deux techniques classiques, faisant usage des statistiques

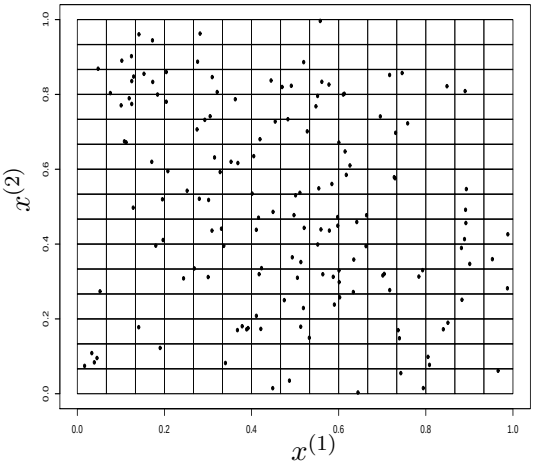


FIG. 3.1 – Ensemble de $n = 150$ points dans $\mathcal{X} = [0,1)^2$ avec une partition en $k = 25 \times 25$ cellules.

de Pearson, et du *rapport de vraisemblance*, utilisées ici pour tester l'uniformité dans le cadre d'une loi multinomiale. Nous remarquerons que ces statistiques relèvent de cas particuliers de la notion de *divergence*. Nous ferons dans une première étape l'hypothèse que le nombre n de points dans $\mathcal{X}$ tend vers l'infini et que le nombre k de cellules de la partition de l'espace $\mathcal{X}$ est une constante fixée (cf. théorèmes 3.3.1, 3.3.2, et 3.3.3). Les résultats que nous fournirons dans ces deux paragraphes seront donc exploitables lorsque le nombre de points dans $\mathcal{X}$ est important.

En second lieu, nous nous placerons dans le cas où le nombre n de points dans $\mathcal{X}$ et le nombre k_n de cellules de la partition de $\mathcal{X}$ sont dépendants et tendent vers l'infini. Nous supposerons, plus particulièrement, que

$$n \to \infty, \quad k_n \to \infty, \quad \lambda_n \to \lambda_\infty, \quad \text{avec,} \quad \lambda_n := \frac{n}{k_n}, \quad \text{et} \quad 0 < \lambda_\infty < \infty. \tag{3.1}$$

La quantité λ_n correspondra au nombre « moyen » de points par cellule. Dans le contexte d'une BDDE pré-existante (cf. Introduction), constituée de points dans un espace $\mathcal{X}$ de dimension relativement élevée, il sera délicat d'avoir, « en moyenne », plus d'un point par cellule (voir (3.15) du §3.2). Par conséquent, le cas où $0 < \lambda_\infty < 1$ dans (3.1) sera, pour nous, le plus important (nous parlerons de « *sparse case* », voir L'Ecuyer *et al.* (2002)).

– Dans un premier paragraphe, nous proposerons d'exploiter les résultats issus du théorème de Holst (1972) (cf. théorème 3.4.1). Au delà des statistiques de Pearson et du rapport de vraisemblance, nous pourrons considérer des statistiques caractérisant le nombre de points de l'espace $\mathcal{X}$ par cellule de la partition de cet espace.

Par exemple, nous étudierons le nombre de cellules « vides » (i.e. ne contenant aucun point), ou le nombre de cellules contenant au moins m points (où $m \geq 1$ est un entier).

– Nous proposerons ensuite l'utilisation de « statistiques par balayage de l'espace », ou « *scan statistics* » (cf. §3.5), qui s'interprètent comme le nombre maximal de points contenus dans un pavé constitué de plusieurs cellules. Ces dernières ont pour objectif de vérifier qu'il n'existe pas de groupe de points dans $\mathcal{X}$ de cardinalité élevée (en comparaison avec des variables aléatoires indépendantes et uniformément distribuées dans $\mathcal{X}$). Pour les utiliser, nous ferons des approximations que nous justifierons (cf. §3.5.1). Nous exploiterons essentiellement les résultats des « *scan statistics* » *discrètes conditionnelles.*

Un paragraphe précisant l'interprétation et l'utilisation des différentes statistiques introduites sera présenté à la fin du présent chapitre.

De façon à respecter les notations usuelles de la statistique nous désignerons les variables aléatoires par des lettres majuscules. Les points de la BDDE (cf. Introduction) dans $\mathcal{X}$ étant considérés comme tels, nous les noterons à présent $X_1, \ldots, X_n$.

3.1 Notion de Test Statistique

L'objectif de ce paragraphe est de fournir un exposé didactique et élémentaire des notions de base et du vocabulaire qui sera utile dans la suite de ce chapitre. Précisons que la notion de *test statistique* telle que nous la présenterons a été principalement introduite par J. Neyman et E.S. Pearson dans les années 1920-1930 (cf. Neyman et Pearson (1928)). Pour des informations supplémentaires, on pourra consulter, par exemple, Monfort (1997).

De façon générale, un problème de *test* consiste à établir une *règle de décision* permettant de choisir entre deux hypothèses,

$H0$: une *hypothèse nulle*,
$H1$: une *hypothèse alternative.*

Nous nous plaçons dans le contexte où nous observons des éléments aléatoires pour lesquels nous avons spécifié un *modèle statistique paramétrique* $(\Omega, \mathcal{A}, \mathbb{P}_p; p \in \mathcal{P})$ défini comme suit,

i) Ω est l'*espace des résultats* et correspond à l'espace dans lequel les observations prennent leurs valeurs ;

ii) $\mathcal{A}$ est une *tribu* de Ω, ainsi $(\Omega, \mathcal{A})$ est un *espace mesurable* ;

iii) $\mathcal{P} := \{\mathbb{P}_p\,; p \in \mathcal{P}\}$ est une *famille de probabilités* sur $\mathcal{A}$ dépendant d'un *paramètre* p, où p appartient à un espace Euclidien $\mathcal{P} \subset \mathbb{R}^q$.

Le problème de *test statistique* que nous étudierons reviendra à *accepter*, ou à *rejeter*, l'hypothèse selon laquelle la « *vraie* » valeur du paramètre, p, correspondant à la loi de probabilité $\mathbb{P}_p$ des observations, appartient, ou non, à un sous-ensemble non vide $\mathcal{P}_0$ de $\mathcal{P}$. Les hypothèses seront :

$$\begin{aligned} H0 : &\quad p \in \mathcal{P}_0 \subset \mathcal{P}, \\ H1 : &\quad p \in \mathcal{P}_0^c \quad \text{où} \quad \mathcal{P}_0^c := \mathcal{P} \setminus \mathcal{P}_0. \end{aligned} \tag{3.2}$$

Le test ainsi défini est une *fonction mesurable* :

$$\delta : \Omega \;\; \rightarrow \;\; \{d_0, d_1\}, \tag{3.3}$$

où *l'ensemble des décisions* $D := \{d_0, d_1\}$ est :

$$\begin{aligned} d_0 : &\quad \text{on choisit } H0, \\ d_1 : &\quad \text{on choisit } H1. \end{aligned}$$

Par définition, δ est une *règle de décision pure*. Le *test statistique* défini par les hypothèses (3.2) est donc appelé *test pur*.

Dans ce contexte, nous pouvons distinguer quatre configurations possibles présentées Tableau (3.1).

	$H0$ « vraie » : $p \in \mathcal{P}_0$	$H1$ « vraie » : $p \in \mathcal{P} \setminus \mathcal{P}_0$
d_0 : On choisit $H0$	« bonne décision »	*erreur de deuxième espèce*
d_1 : On choisit $H1$	*erreur de première espèce*	« bonne décision »

TAB. 3.1 – configurations possibles lors d'un *test statistique*

Nous appelons *région de rejet* ou *région critique* $\mathcal{R}$, l'ensemble des observations de Ω qui conduisent à refuser $H0$,

$$\mathcal{R} \subset \Omega \quad \text{et} \quad \mathcal{R} := \delta^{-1}(d_1). \tag{3.4}$$

Le complémentaire de $\mathcal{R}$ dans Ω est la *région d'acceptation*, $\mathcal{R}^c := \Omega \setminus W$.

Nous désignons par

$$a(p) \;\; = \;\; \mathbb{P}_p(\mathcal{R}), \quad \text{pour} \quad p \in \mathcal{P}, \tag{3.5}$$

la *fonction puissance du test*. Les restrictions de cette fonction sur $\mathcal{P}_0$, $\mathcal{P}_0^c$, sont, respectivement, le *risque de première espèce*, le *risque de deuxième espèce*. La *puissance du test* correspond à la quantité,

$$\sup_{p \in \mathcal{P}_0} \mathbb{P}_p(\mathcal{R}). \tag{3.6}$$

Par la suite, nous considérerons essentiellement le cas où l'hypothèse $H0$ ne contient qu'un seul élément, $\mathcal{P}_0 = \{p_0\}$, où $p_0 \in \mathcal{P} \subset \mathbb{R}^q$ dans (3.2). Il s'agit alors du test d'une *hypothèse simple* $H0$. Dans ce contexte, la *fonction de risque de première espèce* (cf. (3.5)) est une constante égale *au niveau de puissance* du test (cf. (3.6)).

En pratique, nous procéderons de la façon suivante.

1. Nous commencerons par définir l'ensemble $\mathcal{P}$ des paramètres. En général, $\mathcal{P}$ sera un sous-ensemble non vide fermé de $\mathbb{R}^q$.
2. Nous testerons une *hypothèse simple*, $H0 : p \in \mathcal{P}_0 = \{p_0\}$ où $p_0 \in \mathcal{P} \subset \mathbb{R}^q$, contre une *hypothèse alternative*, $H_1 : p \neq p_0$ et $p \in \mathcal{P}$.
3. Nous spécifierons la *puissance du test* $a \in (0, 1)$.
4. Nous déterminerons la *région critique* $\mathcal{R}_a$ de niveau a qui est définie par l'ensemble des observations conduisant à écarter *l'hypothèse simple* $H0$ à tort (i.e. alors que $H0$ est « vraie ») tel que

$$\mathcal{R}_a \subset \delta^{-1}(d_1) \quad \text{et,} \quad \mathbb{P}_{p_0}(\mathcal{R}_a) = a. \tag{3.7}$$

   $\mathcal{R}_a$ correspond à l'*erreur de première espèce* (cf. Tableau 3.1). Cette région $\mathcal{R}_a$ sera déterminée à l'aide d'une *statistique*, fonction des observations, dont la loi de probabilité est connue sous l'Hypothèse $H0$.
5. Nous *accepterons* l'hypothèse $H0$ si les observations n'appartiennent pas à la *région critique* $\mathcal{R}_a$, et la *rejetterons* dans le cas contraire.

Pour la méthode de test définie par les étapes 1-5 ci-dessus, nous ne considérons que l'erreur de première espèce a (cf. Tableau (3.1)). Ce type de test est appelé *test de premier ordre*.

3.2 Partition du pavé unité

Par la suite, nous serons amené à considérer une partition du pavé unité $\mathcal{X} = [0, 1)^d$. Celle-ci sera effectuée de la façon suivante :

i) on réalisera une division de $[0, 1]$ en s segments de même longueur, $h := 1/s$;

ii) pour un ensemble d'entiers $j_1, \ldots, j_d$, tels que $1 \leq j_1 \leq s, \ldots, 1 \leq j_d \leq s$, nous considérerons la *cellule* $A_{j_1,\ldots,j_d} \subset \mathcal{X}$ définie par :

$$A_{j_1,\ldots,j_d} := [(j_1 - 1)h, j_1 h) \times [(j_2 - 1)h, j_2 h) \times \ldots \times [(j_d - 1)h, j_d h)\,. \tag{3.8}$$

Pour $1 \leq j_1 \leq s, \ldots, 1 \leq j_d \leq s$, les s^d cellules $A_{j_1,\ldots,j_d}$ définies par (3.8), sont disjointes et forment une partition de $\mathcal{X} = [0, 1)^d$:

$$[0,1)^d = \bigcup_{j_1=1,\ldots,j_d=1}^{j_1=s,\ldots,j_d=s} A_{j_1,\ldots,j_d}. \tag{3.9}$$

Elles ont toutes le même volume (mesure de Lebesgue sur $\mathcal{X} = [0,1)^d$) : $\boldsymbol{\lambda}(A_j) = h^d$.

Dans un premier temps, nous ne nous intéresserons pas aux « emplacements » des cellules dans le pavé unité. Aussi, de façon à simplifier les notations, nous noterons A_j une cellule, en faisant varier l'indice j de 1 à $k := s^d$.

Pour $X_1, \ldots, X_n$ des variables aléatoires indépendantes indentiquement distribuées de loi uniforme sur $\mathcal{X} = [0,1)^d$, nous adopterons les notations suivantes.

- Soit $\mathbb{1}_{A_j}(X_i)$, la fonction indicatrice de la cellule A_j en X_i :

$$\mathbb{1}_{A_j}(X_i) = \begin{cases} 1 \text{ si } X_i \in A_j, \\ 0 \text{ si } X_i \notin A_j, \end{cases} \quad \text{avec } j = 1, \ldots, k \text{ et } i = 1, \ldots, n. \tag{3.10}$$

  Par définition, pour $j = 1, \ldots, k$ et $i = 1, \ldots, n$, les $\mathbb{1}_{A_j}(X_i)$ sont des variables aléatoires de *Bernouilli* ayant pour paramètre $p_{unif} := 1/k$, c'est-à-dire,

$$\begin{cases} \mathbb{P}(\mathbb{1}_{A_j}(X_i) = 1) = 1/k, \\ \mathbb{P}(\mathbb{1}_{A_j}(X_i) = 0) = 1 - 1/k. \end{cases} \tag{3.11}$$

  Pour un indice $j \in \{1, \ldots, k\}$ fixé, pour des indices distincts i_1 et i_2 tels que $1 \leq i_1 \leq n$ et $1 \leq i_2 \leq n$, les variables aléatoires $\mathbb{1}_{A_j}(X_{i_1})$ et $\mathbb{1}_{A_j}(X_{i_2})$ sont indépendantes.

- Pour un indice $j \in \{1, \ldots, k\}$ spécifié, nous désignerons par Y_j, le nombre de X_i, $i = 1, \ldots, n$, contenus dans la cellule $A_j \subset \mathcal{X}$. Ceci s'écrit formellement,

$$Y_j = \sum_{i=1}^{n} \mathbb{1}_{A_j}(X_i), \qquad j = 1, \ldots, k = s^d. \tag{3.12}$$

  Pour $j = 1, \ldots, k = s^d$, Y_j est la somme de n variables aléatoires indépendantes de Bernouilli (cf (3.11)) et suit donc une loi *binomiale* de paramètre $(n, p_{unif} = \frac{1}{k})$,

$$\begin{aligned} \mathbb{P}(Y_j = n_1) &= \binom{n}{n_1} p_{unif}^{n_1} (1 - p_{unif})^{n-n_1}, \\ \mathbb{P}(Y_j = n_1) &= \binom{n}{n_1} \left(\frac{1}{k}\right)^{n_1} \left(1 - \frac{1}{k}\right)^{n-n_1}, \end{aligned} \tag{3.13}$$

avec la notation usuelle :

$$\binom{n}{n_1} \;=\; \frac{n!}{n_1!(n-n_1)!}.$$

Ainsi, la loi jointe de $(Y_1,\ldots,Y_k)$ est une loi *multinomiale* de paramètres $\big(n, \big(p_1=\frac{1}{k},\ldots,p_k=\frac{1}{k}\big)\big)$,

$$\begin{aligned}
\mathbb{P}(Y_1=n_1,\ldots,Y_k=n_k) &= \binom{n}{n_1\ldots n_k}\left(\frac{1}{k}\right)^n && \text{si } n_1+\ldots+n_k=n,\\
\mathbb{P}(Y_1=n_1,\ldots,Y_k=n_k) &= \quad 0 && \text{si } n_1+\ldots+n_k\neq n. \quad (3.14)
\end{aligned}$$

- Nous noterons

$$\lambda \;:=\; \frac{n}{k}\,. \qquad (3.15)$$

Considérons les cellules A_j, $j=1,\ldots,k$ de la partition du pavé unité $\mathcal{X}=[0,1)^d$ (voir (3.8)). Le terme λ dans (3.15) représente le « nombre moyen » de variables $X_1,\ldots,X_n$ par cellule.

  - Lorsque $\lambda>1$, nous parlerons de « *dense case* » (en « moyenne », plus d'un point par cellule).
  - Lorsque $\lambda<1$, nous parlerons de « *sparse case* » (en « moyenne », moins d'un point par cellule).

Dans notre contexte,

a) la dimension d de $\mathcal{X}=[0,1)^d$ sera fixée et sera relativement élevée,

b) le nombre n de variables aléatoires $X_1,\ldots,X_n$ dans $\mathcal{X}$ sera fixé à l'avance et limité,

c) nous pourrons choisir le nombre $k=s^d$ de cellules comme une puissance d du nombre s de segments sur chaque axe de l'hypercube unité (d étant fixé on fera varier s, cf. *i*) ci-dessus).

Bien qu'ayant le choix de k, il sera souvent difficile de le désigner de façon à ce que $k << n$. En effet, $k=s^d$ devient très important lorsque la dimension d augmente. Le « *sparse case* » (i.e. $n<k$) nous concernera donc particulièrement. En pratique, plusieurs choix de k seront étudiés. En dimension élevée, ce nombre pourra être choisi le plus grand possible afin de prendre en compte un nombre important de sous parties de l'espace $\mathcal{X}$.

La partition du pavé unité décrite ci-dessus est souvent utilisée pour vérifier la qualité de générateurs de nombres aléatoires (voir, par exemple, L'Ecuyer *et al.* (2002)). On

génére $n \times d$ variables aléatoires uniformes sur $[0,1)$, $U_1, ..., U_{n\times d}$. On construit n vecteurs de la façon suivante $V_{d_i} = (U_{d_i}, ..., U_{d_i+d-1})$, pour $i = 1 \dots n$. On réalise ensuite des *tests statistiques* en considérant une partition de l'hypercube unité $[0,1)^d$ comme celle présentée ci-dessus. Cette méthode classique est aussi connue sous le nom de *test périodique*, « *serial test* ». Dans ce contexte, le nombre n de variables dans $\mathcal{X} = [0,1)^d$ peut être élevé. La plupart des tests périodiques existant concernent donc le « dense case ». Dans certains cas, le nombre n est aussi considéré comme une variable aléatoire de loi de *Poisson* de moyenne n. Les Y_j, $j = 1, \dots, k$, telles que nous les avons définies en (3.12) sont alors des variables aléatoires indépendantes et de loi de Poisson de paramètre n/k. La caractérisation des lois de statistiques définies à partir des Y_j, $j = 1, \dots, k$, est alors plus aisée du fait de leur indépendance.

Nous pouvons aussi remarquer des analogies entre cette approche par partition du pavé unité et des travaux réalisés dans le domaine du *« pavage »* et de la *dispersion spatiale* en *biostatistique*. Nous renvoyons entre autres à Greig et Smith (1952), Rogers (1974), Chessel (1978), et Cliff et Ord (1981).

3.3 Test sur le vecteur de paramètres d'une loi multinomiale

Dans ce qui suit, nous ferons référence à l'ensemble des considérations du §(3.2).

Nous désignerons par $H0_{Unif}$ l'hypothèse d'indépendance et d'uniformité des $X_1, \dots, X_n$ dans $\mathcal{X} = [0,1)^d$.

Comme exposé précédemment (voir le §(3.2)), sous l'hypothèse $H0_{Unif}$, le vecteur $(Y_1, \dots, Y_k)$ suit une loi multinomiale de paramètres $\left(n, p = \left(p_1 = \frac{1}{k}, \dots, p_k = \frac{1}{k}\right)\right)$. On en déduit qu'*accepter* ou *rejeter* l'hypothèse $H0_{Unif}$ définie ci-dessus revient à effectuer un *test d'une hypothèse simple* sur le vecteur de paramètres $p \in \mathcal{P}_{Multi}$ de la loi de $(Y_1, \dots, Y_k)$ (cf. (3.12) et (3.14)). Nous renvoyons entre autres à L'Ecuyer *et al.* (2002) ainsi qu'à leurs références bibliographiques. L'ensemble $\mathcal{P}_{Multi}$ alors considéré est défini par,

$$\mathcal{P}_{Multi} \quad := \quad \left\{ (p_1, \dots, p_k) \in (0,1)^k \ : \quad \sum_{j=1}^{k} p_j = 1 \right\}. \tag{3.16}$$

Les hypothèses du test seront désignées par :

$$\begin{aligned} H0_{Multi} : &\qquad p = (1/k, \dots, 1/k) \\ H1_{Multi} : &\qquad p \neq (1/k, \dots, 1/k) \ \text{ et } \ p \in \mathcal{P}_{Multi}, \end{aligned} \tag{3.17}$$

Pour effectuer ce test, nous allons étudier différentes statistiques, fonctions des Y_j, $j = 1, \dots k$ (cf. (3.12) et (3.14)). Pour une fonction $f_{n,k}$, de $\{1, \dots, n\} \subset \mathbb{N}$ dans $\mathbb{R}$, intégrable

par rapport à la mesure de comptage sur $\{1,\ldots,n\}$, les statistiques $S_{f_{n,k}}(Y_1,\ldots,Y_k)$ que nous prendrons en compte sont définies de la façon suivante,

$$S_{f_{n,k}}(Y_1,\ldots,Y_k) \;=\; \sum_{j=1}^{k}\left\{f_{n,k}\left(Y_j\right)\right\}, \tag{3.18}$$

où les Y_j sont définies en (3.12) et (3.14). Avant de proposer des exemples de fonctions f intervenant dans (3.18), nous présenterons ci-dessous le calcul de l'espérance et de la variance de $S_{f_{n,k}}(Y_1,\ldots,Y_k)$.

3.3.1 Espérance et Variance de $S_{f_{n,k}}(Y_1,\ldots,Y_k)$

Soit $(N_1,\ldots,N_k)$, un vecteur aléatoire de loi multinomiale de paramètres $(n,(p_1,\ldots,p_k))$. Pour $j=1,\ldots,k$, pour une fonction $f_{n,k}$, de $\{1,\ldots,n\}\subset\mathbb{N}$ dans $\mathbb{R}$, intégrable par rapport à la mesure de comptage sur $\{1,\ldots,n\}$, nous avons,

$$\mathbb{E}(f_{n,k}(N_j)) \;=\; \sum_{n_1=1}^{n}\binom{n}{n_1}p_j^{n_1}(1-p_j)^{n-n_1}f_{n,k}(n_1), \tag{3.19}$$

pour $n-n_1\geq n_2\geq 0$,

$$\mathbb{P}(N_1=n_1;N_2=n_2) \;=\; \frac{n!\,(p_1)^{n_1}(p_2)^{n_2}(1-p_1-p_2)^{n-n_1-n_2}}{n_1!n_2!(n-n_1-n_2)!}\,. \tag{3.20}$$

Ainsi, pour $(Y_1,\ldots,Y_k)$ défini en (3.12) et (3.14), de loi multinomiale de paramètres $(n,(p_1,\ldots,p_k))$ où $p_1=\ldots=p_k=1/k$, (3.19) implique

$$\mathbb{E}(f_{n,k}(Y_j)) \;=\; \sum_{n_1=1}^{n}\binom{n}{n_1}\frac{(k-1)^{n-n_1}}{k^n}\,f_{n,k}(n_1), \tag{3.21}$$

et, pour $n-n_1\geq n_2\geq 0$, (3.20) implique,

$$\begin{aligned}\mathbb{P}(Y_1=n_1;Y_2=n_2) \;&=\; \frac{n!}{n_1!(n-n_1)!}\frac{(n-n_1)!}{n_2!(n-n_1-n_2)!}\,\frac{(k-2)^{n-n_1-n_2}}{k^n}\\ &=\; \binom{n}{n_1}\binom{n-n_1}{n_2}\,\frac{(k-2)^{n-n_1-n_2}}{k^n}\,.\end{aligned} \tag{3.22}$$

Pour les Y_j, $j=1,\ldots,k$ définies par (3.12) et (3.14), nous posons à présent,

$$\begin{aligned}\mathbb{E}(f_{n,k}(Y_j)) \;&:=\; \mu,\\ \text{avec}\quad \mu \;&:=\; \sum_{n_1=1}^{n}\binom{n}{n_1}\left(\frac{(k-1)^{n-n_1}}{k^{n-1}}\right)f_{n,k}(n_1).\end{aligned} \tag{3.23}$$

On déduit de (3.22) que

$$\begin{aligned}
&\mathbb{E}\left[(f_{n,k}(Y_1)-\mu)(f_{n,k}(Y_2)-\mu)\right] \\
&= \sum_{n_1,n_2;n_1+n_2\le n}^{n} \binom{n}{n_1}\binom{n-n_1}{n_2} \frac{(k-2)^{n-n_1-n_2}}{k^n} (f_{n,k}(n_1)-\mu)(f_{n,k}(n_2)-\mu) \\
&= \sum_{n_1=n_2;\, n_1+n_2\le n}^{n} \binom{n}{n_1}\binom{n-n_1}{n_2} \frac{(k-2)^{n-n_1-n_2}}{k^n} (f_{n,k}(n_1)-\mu)(f_{n,k}(n_2)-\mu) \\
&+ \sum_{n_2<n-n_1,\, n_2\neq n_1}^{n} \binom{n}{n_1}\binom{n-n_1}{n_2} \frac{(k-2)^{n-n_1-n_2}}{k^n} (f_{n,k}(n_1)-\mu)(f_{n,k}(n_2)-\mu) \\
&= \sum_{n_1=1}^{\lfloor n/2\rfloor} \binom{n_1}{n} \frac{(k-2)^{n-2n_1}}{k^n} (f_{n,k}(n_1)-\mu)(f_{n,k}(n_2)-\mu) \\
&+ 2\sum_{n_1=1}^{n} \sum_{n_2=1}^{min(n-n_1,n_1-1)} \binom{n}{n_1}\binom{n-n_1}{n_2} \frac{(k-2)^{n-n_1-n_2}}{k^n} \\
&\qquad \times (f_{n,k}(n_1)-\mu)(f_{n,k}(n_2)-\mu)\,. \qquad (3.24)
\end{aligned}$$

Avec quelques calculs supplémentaires (voir aussi L'Ecuyer *et al.* (2002)), nous obtenons la proposition suivante concernant une statistique de la forme $S_{f_{n,k}}(Y_1,\ldots,Y_k)$ (cf. (3.18)).

Proposition 3.3.1
Soit $f_{n,k}$ une fonction de $\{1,\ldots,n\}\subset\mathbb{N}$ dans $\mathbb{R}$ intégrable par rapport à la mesure de comptage sur $\{1,\ldots,n\}$.

$$\mathbb{E}\left[S_{f_{n,k}}(Y_1,\ldots,Y_k)\right] = k\,\mu = \sum_{n_1=1}^{n} \binom{n}{n_1} \frac{(k-1)^{n-n_1}}{k^{n-1}} f_{n,k}(n_1), \qquad (3.25)$$

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[S_{f_{n,k}}(Y_1,\ldots,Y_k)\right] &= \sum_{n_1=0}^{n} \binom{n}{n_1} \frac{(k-1)^{n-n_1}}{k^{n-1}} (f_{n,k}(n_1)-\mu)^2 \\
&+ \sum_{n_1=0}^{[n/2]} \binom{n}{n_1}\binom{n-n_1}{n_1} \frac{(k-1)(k-2)^{n-2n_1}}{k^{n-1}} (f_{n,k}(n_1)-\mu)^2 \\
&+ 2\sum_{n_1=0}^{n} \sum_{n_2=0}^{\min(n-n_1,n_1-1)} \binom{n}{n_1}\binom{n-n_1}{n_1} \\
&\qquad \times \frac{(k-1)(k-2)^{n-n_1-n_2}}{k^{n-1}} \\
&\qquad \times (f_{n,k}(n_1)-\mu)\,(f_{n,k}(n_2)-\mu), \qquad (3.26)
\end{aligned}$$

où : n est le nombre de variables aléatoires, $X_1, \ldots, X_n$, indépendantes et de loi uniforme sur $\mathcal{X}$ utilisées pour construire les Y_j, $j = 1, \ldots, k$ (voir §(3.2) et (3.12)), $S_{f_{n,k}}(Y_1, \ldots, Y_k)$ est la statistique définie en (3.18), et $\mu := \mathbb{E}(Y_j)$, $j = 1, \ldots, k$ (cf. (3.23)).

Ci-dessous nous allons tout d'abord rappeler deux statistiques classiques pouvant s'écrire sous la forme $S_{f_{n,k}}(Y_1, \ldots, Y_k)$ décrite en (3.18), la *statistique de Pearson* et la *statistique du rapport de vraisemblance*.

Plus généralement, nous étudierons ensuite des familles de fonctions permettant de définir des statistiques, dont celles de *Pearson* et du *rapport de vraisemblance* sont des cas particuliers. Nous verrons que ces statistiques convergent (lorsque $n \to \infty$) vers la même loi. La loi sera différente selon que le nombre k de cellules A_j, $j \in \{1, \ldots, k\}$ (cf. (3.8)), dépend de n (k sera alors noté k_n) ou non.

Les premiers résultats que nous présenterons correspondront au cas où k est en entier fixé, et $n \to \infty$. Ils seront donc exploitables dans le « dense case » (i.e. $n > k$, voir (3.15)). Enfin, nous supposerons que $k \to \infty$ et $n \to \infty$ tels que $n/k \to \lambda_\infty$, où $\lambda_\infty \in \mathbb{R}$ est une constante strictement positive. Lorsque, $0 < \lambda_\infty < 1$, les méthodes présentées s'appliqueront dans le « *sparse case* ».

3.3.2 Test de Pearson

Une statistique fréquemment utilisée pour le test défini par (3.16) et (3.17) est la *statistique de Pearson*. Soit $p_1, \ldots, p_k$, tels que

$$0 < p_j < 1 \quad \text{et} \quad \sum_{j=1}^{k} p_j = 1 \,, \tag{3.27}$$

et $(N_1, \ldots, N_k)$ un vecteur aléatoire de loi multinomiale de paramètres $(n, p_1, \ldots, p_k)$.

Théorème 3.3.1 (Pearson)
Pour un vecteur aléatoire $(N_1, \ldots, N_k)$ de loi multinomiale de paramètres $(n, p_1, \ldots, p_k)$ où $(p_1, \ldots, p_k) \in \mathcal{P}_{multi}$, (cf. (3.16)), on a

$$\sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j} \quad \xrightarrow{\mathcal{L}} \quad \chi^2(k-1), \quad n \to \infty, \tag{3.28}$$

où $\chi^2(k-1)$ désigne la loi du χ^2 à $k-1$ degrés de liberté.

Ce théorème classique a été obtenu par Pearson (1900). Originellement, la statistique de Pearson est utilisée pour effectuer *des tests d'adéquation de lois* permettant de caractériser la qualité d'ajustement d'une distribution théorique à une distribution observée. Nous l'utiliserons ci-dessous pour effectuer un test d'hypothèse simple sur les paramètres d'un vecteur aléatoire de loi multinomiale (voir (3.16) et (3.17)).

Le vecteur aléatoire $(Y_1, \ldots, Y_k)$ défini en (3.12) suit une loi multinomiale de paramètres $(n, (p_1, \ldots, p_k))$ avec $p_1 = \ldots = p_k = 1/k$ (cf. (3.14)). Ainsi, d'après le théorème de Pearson, nous obtenons que

$$\chi^2_p := \sum_{j=1}^{k} \frac{k\,(Y_j - n/k)^2}{n} \quad \xrightarrow{\mathcal{L}} \quad \chi^2(k-1), \quad n \to \infty. \tag{3.29}$$

Pour un niveau de puissance a spécifié, on déduit de (3.29) une région critique $\mathcal{R}_{a\,\chi^2}$ du problème de test défini par (3.16) et (3.17),

$$\mathcal{R}_{a,\,\chi^2} \quad = \quad \left\{ \chi^2_p = \sum_{j=1}^{k} \frac{k\,(Y_j - n/k)^2}{n} \geq c_a \right\}, \tag{3.30}$$

où c_a correspond au a-quantile d'une loi du χ^2 à $k-1$ degrés de liberté.

La région $\mathcal{R}_{a\,,\chi^2}$ définie ci-dessus (cf. (3.30)) est une *région critique asymptotique* car elle est obtenue pour $n \to \infty$. En pratique, on admet que celle-ci est valable lorsque $n \times p_j > 5$, $j = 1, \ldots, k$.

Dans notre contexte, l'application du test de Pearson décrit ci-dessus est délicate. En effet, nous considérons le cas où $p_j = p_{unif} = 1/k$ (cf. (3.12) et (3.14)) où k est le nombre total de cellules de la partition de $\mathcal{X} = [0,1)^d$ ($k = s^d$, cf. le §(3.2)). Pour appliquer le test de Pearson, nous devons donc avoir $n > 5 \times s^d$. Lorsque la dimension d est élevée, cela suppose de disposer d'un nombre n, très important de variables $X_1, \ldots, X_n$. Or, nous supposons que ce nombre est spécifié à l'avance et limité.

Remarques

• Nous avons, pour la statistique χ^2_p définie en (3.29) :

$$\chi^2_p \quad = \quad \sum_{j=1}^{k} \frac{k\,(Y_j - n/k)^2}{n} \tag{3.31}$$

$$\quad = \quad \sum_{j=1}^{k} n \left[\frac{1}{k} \left(\frac{Y_j/n}{1/k} - 1 \right)^2 \right]. \tag{3.32}$$

Nous posons, pour $y > 0$,

$$f_{\chi^2\,n,k}(y) := n\,\frac{1}{k}\,\phi_{\chi^2}\left(\frac{y/n}{1/k}\right), \qquad \text{avec} \qquad \phi_{\chi^2}(t) := (t-1)^2 \quad \text{pour} \quad t \geq 0. \tag{3.33}$$

En substituant $f_{n,k}$ à $f_{{\chi^2}_{n,k}}$ dans $S_{f_{n,k}}(Y_1,\ldots,Y_k)$ définie en (3.18), nous obtenons que

$$S_{f_{{\chi^2}_{n,k}}}(Y_1,\ldots,Y_k) = \chi^2_p. \tag{3.34}$$

• Une autre façon d'écrire χ^2_p (voir (3.29)) est la suivante,

$$\begin{aligned}\chi^2_p &= \sum_{j=1}^{k} \frac{(Y_j - n/k)^2}{n/k} \\ &= \frac{\sum_{j=1}^{k} Y_j^2 - 2n/k \sum_{j=1}^{k} Y_j + n^2/k}{n/k}\,.\end{aligned}$$

Pour les $Y_1,\ldots,Y_k$, définies en (3.12), nous avons $\sum_{j=1}^{k} Y_j = n$, ce qui implique $n/k \sum_{j=1}^{k} Y_j = n^2/k$, ainsi χ^2_p se ré-écrit comme suit,

$$\chi^2_p = \sum_{j=1}^{k} n\,\frac{1}{k}\left\{\left(\frac{Y_j/n}{1/k}\right)^2 - \left(\frac{Y_j/n}{1/k}\right)\right\}. \tag{3.35}$$

Nous posons, pour $y > 0$,

$$f_{1n,k}(y) := 2\,n\left[\frac{1}{k}\,\phi_1\left(\frac{y/n}{1/k}\right)\right], \quad \text{avec} \quad \phi_1(t) := \frac{1}{2}(t^2 - t) \quad \text{pour} \quad t > 0. \tag{3.36}$$

En remplaçant $f_{n,k}$ par $f_{1n,k}(y)$ dans $S_{f_{n,k}}(Y_1,\ldots,Y_k)$ définie en (3.18), nous avons

$$S_{f_{1n,k}}(Y_1,\ldots,Y_k) = \chi^2_p. \tag{3.37}$$

Les remarques ci-dessus nous permettront de constater que χ^2_p est un cas particulier d'une famille de statistiques appelée *ϕ-divergence* (voir le §3.3.4).

3.3.3 Test du rapport de vraisemblance

Une autre technique classique utilisée pour résoudre le problème de test défini par (3.16) et (3.17) est l'utilisation du *rapport des vraisemblances maximales.*

Soit $\mathcal{P}$, un espace de paramètres, sous-ensemble non-vide fermé de $\mathbb{R}^q$, et $\mathcal{P}_0$ un sous-ensemble non vide de $\mathcal{P}$ de dimension ℓ. Soit $Z_1,\ldots,Z_k$, des variables aléatoires dont la *vraisemblance* $L_k(p\,; Z_1,\ldots,Z_k)$ (correspondant à la densité jointe de $Z_1,\ldots,Z_k$) est définie pour $p \in \mathcal{P}$ et $p \in \mathcal{P}_0$.

Le théorème suivant est issu du lemme de Neyman et Pearson (1933).

Théorème 3.3.2 (Rapport de vraisemblances)
Sous les considérations définies ci-dessus, on définit le « rapport des vraisemblances maximal » Λ_k par,

$$\Lambda_k \quad = \quad \frac{\sup_{p\in\mathcal{P}} L_k(p; Z_1, \ldots, Z_k)}{\sup_{p\in\mathcal{P}_0} L_k(p; Z_1, \ldots, Z_k)} \, . \tag{3.38}$$

Nous avons alors,

$$2\log(\Lambda_k) \quad \xrightarrow{\mathcal{L}} \quad \chi^2(q-\ell). \tag{3.39}$$

Ce théorème peut être appliqué au vecteur aléatoire $(Y_1, \ldots, Y_k)$ défini en (3.12) et (3.14) (où $dim(\mathcal{P}) = k-1$ et $dim(\mathcal{P}_0) = 0$). Il s'ensuit,

$$2\ln(\Lambda_k) = \sum_{j=1}^{k} Y_j \, \ln\left(\frac{k}{n}\, Y_j\right) \quad \xrightarrow{\mathcal{L}} \quad \chi^2(k-1). \tag{3.40}$$

Ainsi, pour un niveau de puissance a donné, on déduit de (3.40), une région critique $\mathcal{R}_{a\,\Lambda}$ pour le problème de test défini par (3.16) et (3.17),

$$\mathcal{R}_{a,\,\Lambda} \quad = \quad \left\{ 2\ln(\Lambda_k) = \sum_{j=1}^{k} 2\, Y_j \ln\left(\frac{k}{n}\, Y_j\right) \geq c_a \right\},$$

où c_a correspond au a-quantile d'une loi du χ^2 à $k-1$ degrés de liberté.

Comme $\mathcal{R}_{a,\,\chi^2}$ définie en (3.30), $\mathcal{R}_{a,\,\Lambda}$ est une *région critique asymptotique.*

Remarque
Nous avons,

$$2\ln(\Lambda_k) \quad = \quad \sum_{j=1}^{k} 2n \left[\frac{1}{k}\, \frac{Y_j/n}{1/k} \ln\left(\frac{Y_j/n}{1/k}\right)\right]. \tag{3.41}$$

Nous posons, pour $y > 0$,

$$f_{0n,k}(y) := 2\, n \left[\frac{1}{k}\, \phi_0\left(\frac{y/n}{1/k}\right)\right], \qquad \text{avec} \qquad \phi_0(t) := t\, \ln(t) \quad \text{pour} \quad t > 0. \tag{3.42}$$

En remplaçant $f_{n,k}$ par $f_{0n,k}(y)$ dans $S_{f_{n,k}}(Y_1, \ldots, Y_k)$ définie en (3.18), nous constatons que :

$$S_{f_{0n,k}}(Y_1, \ldots, Y_k) = 2\ln(\Lambda_k). \tag{3.43}$$

A l'aide de la remarque ci-dessus, nous verrons que $S_{f_{0n,k}}(Y_1, \ldots, Y_k)$ est un cas particulier d'une famille de statistiques appelée *ϕ-divergence* (voir le §3.3.4).

3.3.4 « *ϕ-divergence family* »

Dans ce paragraphe, nous exposerons tout d'abord le concept général de divergence. Nous présenterons ensuite des « familles » de statistiques ayant la forme $S_{f_{n,k}}(Y_1, \ldots, Y_k)$ définie en (3.18).

– Soient $n > 1$ et $k_n > 1$ des entiers, où k_n peut éventuellement être une suite dépendant de n non décroissante.

– Soient $q_1, \ldots, q_{k_n}$ tels que,

$$0 < q_j < 1 \quad \text{et} \quad \sum_{j=1}^{k_n} q_j = 1 \,, \tag{3.44}$$

– Soit $(N_1, \ldots, N_{k_n})$ un vecteur aléatoire de loi multinomiale de paramètres $(n, (p_1, \ldots, p_{k_n}))$.

– Nous notons

$$p := (p_1, \ldots, p_{k_n}), \tag{3.45}$$
$$q := (q_1, \ldots, q_{k_n}). \tag{3.46}$$

– Soit $\psi(u, v)$ une fonction, définie pour $u > 0$, $v > 0$, et à valeur dans $\mathbb{R}$ telle que,

a) $\psi(u, v)$ peut être prolongée par continuité en $(u, v) = (0, 0)$ en posant

$$\psi(0, 0) = 0, \tag{3.47}$$

b) pour $u > 0$, et $v > 0$,

$$\psi(0, v) = \lim_{t \to 0} \psi(t, v), \qquad \psi(u, 0) = \lim_{t \to 0} \psi(u, t), \tag{3.48}$$

où les limites dans (3.48) peuvent éventuellement être infinies.

– Nous posons :

$$D_\psi(q, p) = \sum_{j=1}^{k_n} \psi(q_j, p_j). \tag{3.49}$$

La quantité $D_\psi(q, p)$ dans (3.49) correspond à la notion de *divergence*. Elle est fréquemment employée pour comparer des mesures de probabilités $\{p(j), j \in \mathbb{E}\}$ et $\{q(j), j \in \mathbb{E}\}$ où $\mathbb{E} \subset \mathbb{N}$ est non vide. Dans notre contexte, nous l'utiliserons pour effectuer un test sur les paramètres d'une loi multinomiale d'un vecteur aléatoire $(N_1, \ldots, N_{k_n})$ (cf. (3.16) et (3.17)). Dans (3.49) les p_j, $j = 1, \ldots, k_n$, sont alors remplacés par les estimateurs du maximum de vraisemblance $\hat{p}_j := N_j/n$, $j = 1, \ldots, k_n$.

Plusieurs ensembles de fonctions ψ ont été considérés dans la littérature conduisant à différentes familles de divergences. Parmi ces familles, citons par exemple, les *divergences de Csiszár* (cf. Csiszár (1963, 1967)), de *Bergman* (cf. Bergman (1967)), de *Burbea-Rao* (cf. Burbea et Rao (1983), Pardo (1999)).

Dans ce qui suit, nous nous intéresserons essentiellement aux *divergences de Csiszár*, connues aussi sous le nom de *ϕ-divergences*. Nous les écrivons de la façon suivante,

$$D_\phi(q,p) \;=\; \sum_{j=1}^{k_n} p_j\, \phi\left(\frac{q_j}{p_j}\right), \tag{3.50}$$

où ϕ est une fonction telle que

i) ϕ est définie sur $(0,\infty)$ à valeurs dans $\mathbb{R}$,

ii) ϕ admet un prolongement par continuité en 0,

iii) ϕ est convexe,

iv) ϕ vérifie,

$$\begin{aligned}
&\text{pour } u=v=0, && v\,\phi\left(\frac{u}{v}\right) \equiv 0,\\
&\text{pour } u=0,\ v>0, && v\,\phi\left(\frac{u}{v}\right) \equiv v\phi(0),\\
&\text{pour } v=0,\ u>0, && v\,\phi\left(\frac{u}{v}\right) \equiv \lim_{t\to\infty}\frac{\phi(t)}{t}.
\end{aligned} \tag{3.51}$$

Nous supposerons de plus que

v) ϕ est localement différentiable d'ordre 2 au voisinage de 1, avec

$$\phi(1)=0 \qquad \text{et} \qquad \phi''(1)>0. \tag{3.52}$$

L'ensemble des fonctions vérifiant $i)-v)$ ci-dessus sera noté Φ^*.

Pour un vecteur aléatoire de loi multinomiale, $(N_1,\ldots,N_{k_n})$, de paramètres $(n, p=(p_1,\ldots,p_{k_n}))$, nous désignons par $\hat{p} := (\hat{p}_1,\ldots,\hat{p}_{k_n})$ l'estimateur du maximum de vraisemblance de p. On a, par calcul, $\hat{p}_j = N_j/n$ pour $j=1,\ldots,k_n$. Dans ce contexte, la loi asymptotique ($n\to\infty$) des ϕ-divergences $D_\phi(\hat{p},p)$, $\phi\in\Phi^*$, définies en (3.50), est fournie par le théorème suivant.

Théorème 3.3.3 (Lois asymptotiques des ϕ-*divergences*)
Dans le contexte précisé ci-dessus,

a) Si n, k_n, et p_j vérifient

$$\lim_{n\to\infty} k_n = k, \quad \text{et} \quad \liminf_{n\to\infty} \min_{1\le k_n} p_j > 0 \quad \text{où } k \ge 1 \text{ est une constante dans } \mathbb{N}$$

alors,

$$\frac{2n}{\phi''(1)} D_\phi(\hat{p}, q) \quad \xrightarrow{\mathcal{L}} \quad \chi^2(k-1). \tag{3.53}$$

b) Si la fonction $\phi \in \Phi^*$ (voir $i-v$ ci-dessus) est localement lipschitzienne dans un voisinage de 1, si n et k_n vérifient, pour $\beta \ge 1$

$$\lim_{n\to\infty} k_n = \infty, \qquad \text{et} \qquad \lim_{n\to\infty} \frac{k_n^{1+\beta}}{n} = 0, \tag{3.54}$$

et les p_j sont tels que

$$\liminf_{n\to\infty} k_n^\beta \times \left(\min_{1\le j\le k_n} p_j \right) \quad > \quad 0, \tag{3.55}$$

alors

$$\frac{2n}{\phi''(1)} \frac{D_\phi(\hat{p}, q) - \phi''(1)\, k_n}{\sqrt{2\, k_n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \tag{3.56}$$

Précisons que les convergences en loi des cas *a*) et *b*) existent également avec des hypothèses différentes sur n, k_n, et le choix des fonctions $\phi \in \Phi^*$ (voir Tumanyan (1954, 1975), Inglot *et al.* (1990)). Les résultats du théorème 3.3.3 sont, à notre connaissance, ceux dont les hypothèses sont les plus générales (cf. Györfi et Vajda (2002)).

Remarques

Les théorèmes 3.3.1 de Pearson, et 3.3.2 du rapport de vraisemblance, sont des cas particuliers du théorème 3.3.3.

• Pour la statistique de Pearson, nous avions remarqué (voir (3.33) et (3.34)) que

$$\chi^2_p = n \sum_{j=1}^{k} \frac{1}{k}\, \phi_{\chi^2}\left(\frac{y/n}{1/k} \right), \qquad \text{avec} \qquad \phi_{\chi^2}(t) = (t-1)^2, \quad t \ge 0\,.$$

En utilisant la notation (3.50), avec $p = (1/k, \ldots, 1/k)$ et $q = \hat{p} = (Y_1/n, \ldots, Y_k/n)$,

nous avons donc

$$\chi^2_p \;=\; nD_{\phi_{\chi^2}}(\hat{p},p). \tag{3.57}$$

Nous vérifions bien que $\phi_{\chi^2} \in \Phi^*$, et nous avons $\phi''(1) = 2$. Ainsi pour k fixé, nous pouvons appliquer le cas a) du théorème 3.3.3. Nous obtenons alors le même résultat qu'avec le théorème 3.3.1 de Pearson.

• Nous avions aussi constaté (cf. (3.36) et (3.37)) que

$$\chi^2_p = 2n\sum_{j=1}^{k}\frac{1}{k}\,\phi_1\left(\frac{y/n}{1/k}\right), \qquad \text{avec} \qquad \phi_1(t)=\frac{1}{2}(t^2-t), \quad t\geq 0\,. \tag{3.58}$$

Ainsi, χ^2_p peut s'écrire de la façon suivante,

$$\chi^2_p \;=\; 2nD_{\phi_1}(\hat{p},p),$$

avec $p = (1/k,\ldots,1/k)$ et $q = \hat{p} = (Y_1/n,\ldots,Y_k/n)$ dans (3.50). Nous vérifions que $\phi_1 \in \Phi^*$ et nous avons, $\phi_1''(t) = 1$. Pour k fixé, nous pouvons appliquer le cas a) du théorème 3.3.3 et obtenons le même résultat que le théorème 3.3.1 de Pearson.

• Concernant la statistique du rapport de vraisemblance, l'équation (3.43) implique que

$$2\ln(\Lambda_k) = 2\,n\sum_{j=1}^{k}\frac{1}{k}\,\phi_0\left(\frac{y/n}{1/k}\right), \qquad \text{avec} \qquad \phi_0(t)=t\ln(t), \quad t>0. \tag{3.59}$$

La fonction ϕ_0 admet un prolongement par continuité en 0 en posant $\phi_0(0) = 0$. On vérifie que $\phi_0 \in \Phi^*$. Nous avons $\phi_0''(1) = 1$. En posant $p = (1/k,\ldots,1/k)$ et $q = \hat{p} = (Y_1/n,\ldots,Y_k/n)$ dans (3.50), on a

$$2\ln(\Lambda_k) \;=\; 2nD_{\phi_0}(\hat{p},p).$$

Pour k fixé, le cas a) du théorème 3.3.3 s'applique et permet d'avoir le même résultat que le théorème 3.3.2 du rapport de vraisemblance.

• Plus généralement nous pouvons considérer la famille de fonctions définies par,

$$\phi_\delta(t) \;:=\; \frac{1}{\delta(\delta+1)}\,t(t^\delta-1), \quad \text{pour} \quad t\geq 0 \quad \text{et} \quad \delta>-1. \tag{3.60}$$

La fonction ϕ_1 (voir (3.58)) est obtenue pour $\delta = 1$ dans (3.60). La fonction ϕ_0 (voir (3.59)), correspond au choix de $\delta = 0$ (obtenue pour $t \geq 0$, en prolongeant $\phi_\delta(t)$ par continuité lorsque $\delta \to 0$). La famille de divergences définies à l'aide de (3.60) est appelée « power divergence ». Nous renvoyons à l'étude de Read et Cressie (1988). Les statistiques alors considérées peuvent s'écrire sous la forme de (3.18) c'est-à-dire,

$$S_{f_{\delta k,n}}(Y_1,\ldots,Y_k) = \sum_{j=1}^{k} f_{\delta k,n}(Y_j), \tag{3.61}$$

avec,

$$f_{\delta k,n}(y) = 2\,\frac{n}{k}\phi_\delta\left(\frac{y}{n/k}\right) \tag{3.62}$$

$$= \frac{2}{\delta(\delta+1)}\,y\left(\left(\frac{y}{n/k}\right)^{\delta}-1\right) \quad y\geq 0\,. \tag{3.63}$$

3.3.5 Discussion

Dans notre contexte, le nombre n de variables aléatoires $X_1,\ldots,X_n$ dans $\mathcal{X}=[0,1)^d$ (voir le §3.2) est fixé à l'avance. Le nombre $k=s^d$ de cellules de la partition du pavé unité (voir (3.8) du §3.2)) n'est pas imposé, et peut être choisi.

Pour appliquer le cas *a*) du théorème 3.3.3, nous devrons donc choisir k tel que $s^d << n$. Un tel choix semble donc délicat, car ceci implique de disposer d'un nombre très important de points lorsque la dimension d de $\mathcal{X}$ est relativement élevée.

L'exploitation du cas *b*) du théorème 3.3.3 implique que k_n (ici, k dépend de n) doit être choisi de façon à ce que $k_n^{1+\beta} < n$ (avec $\beta \geq 1$). Bien qu'ici k_n n puisse être une fonction strictement croissante de n avec $k_n \to \infty$, la condition (3.54) implique d'avoir un nombre important de points.

Que ce soit le cas *a*) ou le cas *b*) du théorème 3.3.3, celui-ci s'applique dans le « *dense case* », c'est-à-dire lorsque le nombre moyen de variables $X_1,\ldots,X_n$ par cellule de la partition de $\mathcal{X}$ considérée, noté λ (voir (3.15)), est strictement supérieur à 1. Pour revenir au « *sparse case* » ($\lambda < 1$), nous allons faire l'hypothèse suivante sur k_n et n, nous supposerons que

$$\text{pour} \quad n\to\infty,\ k_n\to\infty, \qquad \frac{n}{k_n}\to\lambda_\infty, \tag{3.64}$$

où $\lambda_\infty \in \mathbb{R}$ est une constante strictement positive. Nous verrons qu'il existe des résultats analogues à ceux du cas *b*) du théorème (3.3.3).

3.4 « *Sparse case* »

Dans ce qui suit, nous ferons usage des notations du §3.2 et considérerons des statistiques de la forme de (3.18) dont nous rappelons l'expression ci-dessous :

$$S_{f_{n,k_n}}(Y_1,\ldots,Y_{k_n}) = \sum_{j=1}^{k_n}\left\{f_{n,k_n}\left(Y_j\right)\right\}, \tag{3.65}$$

où ici, k_n est un entier qui dépend de n, et $(Y_1, \ldots, Y_{k_n})$ un vecteur aléatoire de loi multinomiale de paramètres $(n, 1/k_n, \ldots, 1/k_n)$ (voir (3.12) et (3.14)).

Nous supposerons que

$$n \to \infty, \qquad k_n \to \infty, \qquad \frac{n}{k_n} \to \lambda_\infty. \tag{3.66}$$

Le contexte est donc différent du §3.3 puisqu'ici, il est possible de considérer le cas où $0 < \lambda_n < 1$ (avec $\lambda_n = n/k_n$). Ceci signifie qu'en « moyenne », le nombre de points par cellule pourra être inférieur à 1 (voir l'équation (3.15) du §3.2).

Un premier paragraphe présentera des résultats issus du théorème de Holst (1972). Celui-ci permettra, entre autres, l'étude du nombre de cellules vides (i.e. ne contenant aucun point) de la partition de l'espace $\mathcal{X}$ (cf. §3.2). Nous parlons alors de « *sparse test* » (cf. L'Ecuyer *et al.* (2002)).

Nous proposerons ensuite une stratégie dont l'objectif est de déterminer le plus grand pavé (constitué de cellules) ne contenant aucun point.

3.4.1 Application du théorème de Holst (1972)

Nous désignons par $Z_1, \ldots, Z_{k_n}$ des variables aléatoires indépendantes et de même loi de Poisson de paramètre $\lambda := n/k_n$, et nous posons,

$$S^\star_{f_{n,k_n}}(Z_1, \ldots, Z_{k_n}) = \sum_{j=1}^{k_n} \{f_{n,k_n}(Z_j)\}, \tag{3.67}$$

$$\mu_{S^\star_{f_{n,k_n}}} = \sum_{j=1}^{k_n} \mathbb{E}\left[f_{n,k_n}(Z_j)\right], \tag{3.68}$$

$$\sigma^2_{S^\star_{f_{n,k_n}}} = \sum_{j=1}^{k_n} \mathbb{V}\mathrm{ar}\left[f_{n,k_n}(Z_j)\right] - \frac{1}{n}\left\{\sum_{j=1}^{k_n} \mathbb{C}\mathrm{ov}\left[Z_j, f_{n,k_n}(Z_j)\right]\right\}^2. \tag{3.69}$$

Nous avons le théorème suivant.

Théorème 3.4.1 (Holst (1972))
Nous reprenons les notations ci-dessus, et faisons les hypothèses suivantes.

i) Soit g_n une fonction mesurable de $[0, \infty) \times [0, 1]$ dans $\mathbb{R}$ telle que

$$f_{n,k_n}(Y_j) = g_n(Y_j, j/k_n).$$

ii) Soient $c_1 \in \mathbb{R}, c_2 \in \mathbb{R}$ des constantes, ne dépendant pas de n, telles que,

$$|g_n(u,v)| \quad \leq c_1 \quad \exp(c_2\, y).$$

iii) Soient n et k_n qui vérifient (3.66).

iv) Soit $\sigma^2_{S^\star_{f_{n,k_n}}}$ qui vérifie

$$0 < \liminf \frac{\sigma^2_{S^\star_{f_{n,k_n}}}}{n} \leq \limsup \frac{\sigma^2_{S^\star_{f_{n,k_n}}}}{n} < \infty.$$

Alors, on a

$$\frac{S_{f_{n,k_n}}(Y_1,\ldots,Y_k) - \mu_{S^\star_{f_{n,k_n}}}}{\sigma_{S^\star_{f_{n,k_n}}}} \quad \xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \tag{3.70}$$

• On vérifie que le théorème 3.4.1 s'applique, entre autres, pour les « *power divergence statistics* », définies en (3.50) et (3.60). Nous notons, pour $\delta \geq 0$ (voir aussi (3.59), (3.58) pour, $\delta = 0$, $\delta = 1$),

$$S_{f_{\delta k_n,n}}(Y_1,\ldots,Y_{k_n}) \quad = \quad \sum_{j=1}^{k_n} f_{\delta k_n,n}(Y_j), \tag{3.71}$$

avec,

$$f_{\delta k_n,n}(y) \quad = \quad \frac{2\,y}{\delta(\delta+1)}\left(\left(\frac{y}{n/k_n}\right)^{\delta} - 1\right) \quad \text{pour} \quad y \geq 0\,. \tag{3.72}$$

• Les hypothèses faites sur n et k_n en (3.66) permettent aussi d'introduire, pour $y \geq 0$, les fonctions indicatrices de la forme $f_{k_n,n}(y) = \mathbb{1}_{y=m}$ avec $m \geq 0$, et $f_{k_n,n}(y) = \mathbb{1}_{y\geq m}$ avec $m \geq 1$. En effet, celles-ci ne pouvaient pas être considérées lorsque k est fixé et $n \to \infty$. Dans ce cas, par la *loi forte des grands nombres*, les statistiques $S_{f_{k,n}}(Y_1,\ldots,Y_k) = \sum_{j=1}^{k} \mathbb{1}_{Y_j=m}$, et $S_{f_{k,n}}(Y_1,\ldots,Y_k) = \sum_{j=1}^{k} \mathbb{1}_{Y_j\geq m}$, convergent presque sûrement, vers 0, et vers k, respectivement. Nous posons,

$$S_{eq(m)_{k_n,n}}(Y_1,\ldots,Y_k) = \sum_{j=1}^{k_n} \mathbb{1}_{Y_j=m}, \qquad \text{pour } m \geq 0,$$

$$S_{up(m)_{k_n,n}}(Y_1,\ldots,Y_k) = \sum_{j=1}^{k_n} \mathbb{1}_{Y_j\geq m}, \qquad \text{pour } m \geq 1. \tag{3.73}$$

$S_{eq(m)}(Y_1,\ldots,Y_k)$ s'interprète comme le nombre de cellules A_j, $j=1,\ldots,k_n$, contenant exactement m variables parmi les $X_1,\ldots,X_n$ (cf. (3.8), (3.12)).

$S_{up(m)_{k_n,n}}(Y_1,\ldots,Y_k)$ s'interprète comme le nombre de cellules A_j, $j=1,\ldots,k_n$, contenant au moins m variables parmi les $X_1,\ldots,X_n$ (cf. (3.8), (3.12)).

• Pour effectuer le test défini par (3.16) et (3.17) nous utiliserons les statistiques définies par,

$$T_{f_{\delta n,k_n}} = \frac{S_{f_{\delta n,k_n}}(Y_1,\ldots,Y_{k_n}) - \mu_{S^\star_{f_{\delta n,k_n}}}}{\sigma^2_{S^\star_{f_{\delta n,k_n}}}}, \tag{3.74}$$

$$T_{eq(m)_{k_n,n}} = \frac{S_{eq(m)}(Y_1,\ldots,Y_{k_n}) - \mu_{S^\star_{eq(m)}}}{\sigma^2_{S^\star_{eq(m)}}}, \tag{3.75}$$

$$T_{up(m)_{k_n,n}} = \frac{S_{up(m)}(Y_1,\ldots,Y_{k_n}) - \mu_{S^\star_{up(m)}}}{\sigma^2_{S^\star_{up(m)}}}, \tag{3.76}$$

où $\mu_{S^\star_{f_{\delta n,k_n}}}$, $\sigma^2_{S^\star_{f_{\delta n,k_n}}}$, $\mu_{S^\star_{eq(m)}}$, $\sigma^2_{S^\star_{eq(m)}}$, $\mu_{S^\star_{up(m)}}$, $\sigma^2_{S^\star_{up(m)}}$, sont obtenues à l'aide de (3.68), (3.69), avec, respectivement, $f_{\delta n,k_n}$ (cf. (3.72)), $f_m(y)=\mathbb{1}_{y=m}$, $y\geq 0$, et $f_m(y)=\mathbb{1}_{y\geq m}$, $y\geq 0$.

Après calcul, nous obtenons,

$$\mu_{S^\star_{eq(m)}} = \frac{\lambda^m}{m!}\exp(-\lambda), \tag{3.77}$$

$$\sigma^2_{S^\star_{eq(m)}} = k\,\mu_{S^\star_{eq(m)}}\left(1-\mu_{S^\star_{eq(m)}}\right) - \frac{1}{n}\left[k\,\mu^\star_{S_{eq(m)}}(m-\lambda)\right]^2, \tag{3.78}$$

$$\mu_{S^\star_{up(m)}} = 1-\sum_{\ell=0}^{m-1}\frac{\lambda^\ell}{\ell!}\exp(-\lambda), \tag{3.79}$$

$$\sigma^2_{S^\star_{up(m)}} = k\,\mu_{S^\star_{up(m)}}\left(1-\mu_{S^\star_{up(m)}}\right) - \left[k\left(\lambda-\sum_{l=0}^{m-1}\frac{\lambda^\ell}{\ell!}\,\ell\,\exp(-\lambda)-\lambda\,\mu_{S^\star_{up(m)}}\right)\right]^2, \tag{3.80}$$

et, pour $\delta=1$,

$$\mu_{S^\star_{f_{1n,k_n}}} = k_n, \tag{3.81}$$

$$\sigma^2_{S^\star_{f_{1n,k_n}}} = k_n\left(\lambda+4+\frac{1}{\lambda}\right) - \frac{1}{n}\left[k_n(\lambda+1)\right]^2. \tag{3.82}$$

Par application du théorème 3.4.1 nous avons :

$$T_{f_{\delta n,k_n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \tag{3.83}$$

$$T_{eq(m)_{k_n,n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \tag{3.84}$$

$$T_{up(m)_{k_n,n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \tag{3.85}$$

• Pour un niveau de puissance a, on déduit de (3.83), (3.84) et (3.85) les régions critiques suivantes du problème de test défini par (3.16) et (3.17),

$$\mathcal{R}_{a,T_{f_{\delta n,k_n}}} = \left\{ \left| T_{f_{\delta n,k_n}} \right| \geq c_{a/2} \right\}, \tag{3.86}$$

$$\mathcal{R}_{a,T_{eq(m)_{k_n,n}}} = \left\{ \left| T_{eq(m)_{k_n,n}} \right| \geq c_{a/2} \right\}, \tag{3.87}$$

$$\mathcal{R}_{a,T_{up(m)_{k_n,n}}} = \left\{ \left| T_{up(m)_{k_n,n}} \right| \geq c_{a/2} \right\}, \tag{3.88}$$

où $c_{a/2}$ correspond au $a/2$ quantile supérieur d'une loi normale centrée réduite, $\mathcal{N}(0,1)$.

Lorsque $m = 0$, effectuer un test à l'aide de $T_{eq(0)_{k_n,n}}$ (voir (3.75)) revient à considérer le nombre de cellules « vides » (i.e. ne contenant aucune variable $X_1, \ldots, X_n$, cf. (3.12) du §3.2). La région critique (3.87) nous fournira alors les nombres minimum et maximum, (associés à un niveau de puissance a) de ces cellules vides. Ceci nous permettra de vérifier que les variables $X_1, \ldots, X_n$ « recouvrent » l'hypercube unité de façon satisfaisante (au sens de la puissance du test a que nous aurons spécifiée) sans qu'il existe un nombre trop faible ou trop important de « trous » (i.e. de cellules vides).

Lorsque $T_{eq(0)_{k_n,n}} \in \mathcal{R}_{a,T_{eq(0)_{k_n,n}}}$ (le test est rejeté), nous présentons ci-dessous une stratégie dont l'objectif est de déterminer le plus grand pavé constitué de cellules ne contenant aucun point.

3.4.2 Recherche du « plus grand pavé vide »

Conformément à la notation (3.8) du §3.2, nous notons, pour un ensemble d'entiers $j_1, \ldots, j_d$, tels que $1 \leq j_1 \leq s, \ldots, 1 \leq j_d \leq s$,

$$A_{j_1,\ldots,j_d} := [(j_1-1)h, j_1 h) \times [(j_2-1)h, j_2 h) \times \ldots \times [(j_d-1)h, j_d h), \tag{3.89}$$

une cellule de la partition de l'hypercube unité $\mathcal{X}$. Nous nous intéressons à présent à « l'emplacement » des cellules et notons,

$$Y_{j_1,\ldots,j_d} = \sum_{i=1}^{n} \mathbb{1}_{X_i \in A_{j_1,\ldots,j_d}}, \qquad j_1, \ldots, j_d \in \{1, \ldots, s\}. \tag{3.90}$$

Pour un entier $b \geq 1$, pour des indices, $s - b + 1 \geq j_1 \geq 1$, et $s - b + 1 \geq j_d \geq 1$ nous posons,

$$W_{b j_1,\ldots,j_d}(Y_{1,\ldots,1}, \ldots, Y_{s,\ldots,s}) = \sum_{r_1=j_1}^{j_1+b-1} \ldots \sum_{r_d=j_d}^{j_d+b-1} Y_{r_1,\ldots,r_d}. \tag{3.91}$$

Les $W_{bj_1,\ldots,j_d}$, définis ci-dessus, correspondent au nombre de variables $X_1, \ldots, X_n$, contenues dans des pavés « carrés », union de $b \times \ldots \times b$ cellules, dont les côtés sont parallèles aux axes de l'hypercube unité et ont tous la même longueur b/s. Nous considérerons

$$S_{minb}\left(Y_{1,\ldots,1}, \ldots, Y_{s,\ldots,s}\right) = \min_{\substack{1 \leq j_1 \leq s-b+1, \\ \ldots \\ 1 \leq j_d \leq s-b+1}} \left\{W_{bj_1,\ldots,j_d}\left(Y_{1,\ldots,1}, \ldots, Y_{s,\ldots,s}\right), \right\} \tag{3.92}$$

le pavé contenant le moins de variables. Nous ferons varier la taille des pavés à l'aide de diffréntes valeurs $b_1, \ldots, b_\ell$, et retiendrons le plus grand pavé ne contenant aucun point,

$$W_{bo}\left(Y_{1,\ldots,1}, \ldots, Y_{s,\ldots,s}\right) = \max_{b \in \{b_1,\ldots,b_\ell\}} \{S_{minb}\left(Y_{1,\ldots,1}, \ldots, Y_{s,\ldots,s}\right) : S_{minb}\left(Y_{1,\ldots,1}, \ldots, Y_{s,\ldots,s}\right) = 0\} . \tag{3.93}$$

Un tel pavé est illustré Figure 3.2 (en vert). Cette technique nous permettra de localiser les parties « vides » de l'hypercube unité (i.e. ne contenant pas de variables $X_1, \ldots, X_n$).

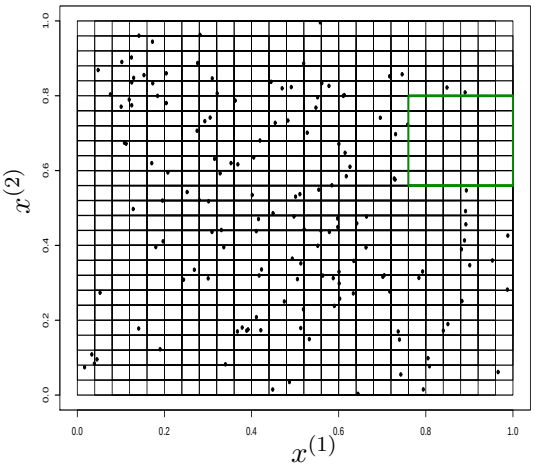


FIG. 3.2 – Représentation du plus grand pavé vide pour un ensemble de $n = 150$ points dans $\mathcal{X} = [0, 1)^2$ avec une partition en $k = 25 \times 25$ cellules.

De façon analogue, nous considérerons dans le paragraphe suivant des pavés « carrés » constitués de cellules. L'objectif sera de déterminer si ces pavés contiennent un nombre important de points en comparaison avec des variables aléatoires indépendantes et de loi uniforme dans $\mathcal{X}$.

3.5 « *Scan Statistics* »

Nous considérons l'ensemble des notations du §3.2 et étudierons ci-dessous le cas où $\mathcal{X} = [0,1)^2$ ($d = 2$).

Comme précédemment, nous testerons l'Hypothèse $H0_{unif}$ d'indépendance et d'uniformité des $X_1, \ldots, X_n$ dans $\mathcal{X}$. L'objectif poursuivi ici est de vérifier qu'il n'existe pas de parties de l'hypercube unité $\mathcal{X}$ contenant un nombre « trop important » (dans un sens précisé plus loin) de variables $X_1, \ldots, X_n$. Nous proposons d'utiliser des résultats issus de la théorie des « scan statistics ». Ces statistiques sont utilisées dans de nombreux domaines (santé publique, étude de séquences d'ADN, télécommunications, etc.), voir Glaz *et al.* (2001), et aussi, Deheuvels *et al.* (1988). Elles consistent à effectuer un « balayage » (« *scaning* ») du temps ou de l'espace à la recherche de groupes d'événements. Dans le contexte présenté au §3.2, nous prendrons en compte des groupes de variables $X_1, \ldots, X_n$ compris dans des pavés constitués de cellules A_{j_1,j_2}, $1 \leq j_1 \leq s$, $1 \leq j_2 \leq s$.

Pour $1 \leq j_1 \leq s$, $1 \leq j_2 \leq s$, nous désignons le nombre de variables $X_1, \ldots, X_n$ contenues dans un pavé « carré » constitués de $b \times b$ cellules par

$$W_{bj_1,j_2}(Y_{1,1}, \ldots, Y_{s,s}) = \sum_{r_1=j_1}^{j_1+b-1} \sum_{r_2=j_2}^{j_2+b-1} Y_{r_1,r_2}. \tag{3.94}$$

Nous renvoyons à (3.90) pour la définition de Y_{r_1,r_2}. Nous introduisons, pour un entier $b \geq 1$,

$$M_b(Y_{1,1}, \ldots, Y_{s,s}) = \max_{\substack{1 \leq j_1 \leq s-b+1, \\ 1 \leq j_2 \leq s-b+1}} \left\{ W_{bj_1,j_2}(Y_{1,1}, \ldots, Y_{s,s}) \right\}. \tag{3.95}$$

Cette statistique s'interprète comme le nombre maximum de variables $X_1, \ldots, X_n$ contenues dans des pavés « carrés » (union de $b \times b$ cellules).

Lorsque les variables Y_{j_1,j_2} (voir (3.12) et (3.90)), sont indépendantes et identiquement distribuées, il est possible de déterminer la loi de $M_b(Y_{1,1}, \ldots, Y_{s,s})$ (voir Glaz *et al.* (2001)). Cependant, dans le contexte décrit au §3.2, les Y_{j_1,j_2} ne vérifient pas cette hypothèse (leur somme vaut n).

3.5.1 Approximation par une loi de Poisson conditionnelle

Faisons l'hypothèse asymptotique (3.66), c'est-à-dire, supposons que

$$n \to \infty, \qquad k_n \to \infty, \qquad \frac{n}{k_n} \to \lambda_\infty. \tag{3.96}$$

En considérant $\lambda_n = n/k_n$ défini comme en (3.15), l'équation (3.13) caractérisant la

loi des Y_{j_1,j_2}, $1 \leq j_1 \leq s$, $1 \leq j_2 \leq s$, (cf. (3.12), (3.90)) se ré-écrit,

$$\begin{aligned}\mathbb{P}(Y_{j_1,j_2} = n_1) &= \frac{n!}{n_1!(n-n_1)!}\left(\frac{\lambda_n}{n}\right)^{n_1}\left(1-\frac{\lambda_n}{n}\right)^{n-n_1} \\ &= \frac{\lambda_n^{n_1}}{n_1!}\left(1-\frac{\lambda_n}{n}\right)^{n-n_1}\frac{n!}{n^{n_1}(n-n_1)!}\,. \end{aligned} \tag{3.97}$$

Pour n_1 un entier fixé, pour n et k_n qui vérifient (3.96), l'équation (3.97) implique que

$$\mathbb{P}(Y_{j_1,j_2} = n_1) \longrightarrow \frac{\lambda_\infty^{n_1}}{n_1!},\ \exp(-\lambda_\infty). \tag{3.98}$$

Par conséquent, pour $1 \leq j_1 \leq s$, $1 \leq j_2 \leq s$, la loi asymptotique (au sens de (3.96)) de Y_{j_1,j_2} est une loi de Poisson de paramètre λ_∞.

De la même façon, pour n_1 et n_2 des entiers fixés, remarquons que l'équation (3.20) se ré-écrit, pour $(j_{r_1}, j_{r_2}) \neq (j_{r_3}, j_{r_4})$,

$$\begin{aligned}\mathbb{P}(Y_{j_{r_1},j_{r_2}} = n_1; Y_{j_{r_3},j_{r_4}} = n_2) &= \frac{n!}{n_1!n_2!(n-n_1-n_2)!}\left(\frac{\lambda_n}{n}\right)^{n_1+n_2}\left(1-\frac{2\lambda_n}{n}\right)^{n-n_1-n_2} \\ &= \frac{\lambda_n^{n_1}}{n_1!}\,\frac{\lambda_n^{n_2}}{n_2!}\left(1-\frac{2\lambda_n}{n}\right)^{n} \\ &\quad \times\left(1-\frac{2\lambda_n}{n}\right)^{-n_1-n_2}\frac{n!}{n^{n_1+n_2}(n-n_1+n_2)!}\,. \end{aligned} \tag{3.99}$$

Nous obtenons, pour n et k_n qui vérifient (3.96),

$$\mathbb{P}(Y_{j_{r_1},j_{r_2}} = n_1; Y_{j_{r_3},j_{r_4}} = n_2) \longrightarrow \frac{\lambda_\infty^{n_1}}{n_1!}\exp(-\lambda_\infty)\,\frac{\lambda_\infty^{n_2}}{n_2!}\exp(-\lambda_\infty). \tag{3.100}$$

Les équations (3.98) et (3.100) impliquent l'indépendance asymptotique (au sens de (3.96)) de $Y_{j_{r_1},j_{r_2}}$ et $Y_{j_{r_3},j_{r_4}}$.

De la même façon il est possible de montrer l'indépendance asymptotique des Y_{j_1,j_2}, $1 \leq j_1 \leq s$, $1 \leq j_2 \leq s$, lorsque n et k_n vérifient (3.66).

Nous avons donc la proposition suivante.

Proposition 3.5.1
Soit $n \to \infty$ et $k_n \to \infty$, tels que

$$\frac{n}{k_n} \quad \to \quad \lambda_\infty, \tag{3.101}$$

où $\lambda_\infty \in \mathbb{R}$ est une constante strictement positive, les variables aléatoires Y_{j_1,j_2}, $1 \leq j_1 \leq s$, $1 \leq j_2 \leq s$ définies par (3.12) *sont asymptotiquement indépendantes et identiquement distribuées de loi de Poisson de paramètre λ_∞.*

Rappelons que nous nous plaçons dans le contexte où le nombre n de variables $X_1, \ldots, X_n$ dans $\mathcal{X} = [0,1)^d$ sera limité et le nombre de cellules $k = s^2$, que l'on spécifiera, pourra être important (cf. *a*) *b*) *c*) du §3.2). Par conséquent, pour $1 \leq j_1 \leq s$ $1 \leq j_2 \leq s$, les variables Y_{j_1,j_2}, correspondant aux nombres de variables $X_1, \ldots, X_n$, par cellule A_{j_1,j_2}, prendront des valeurs faibles (par rapport à n). Ainsi, pour de faibles valeurs n_1 et n_2 ($n_1 << n$ et $n_2 << n$) dans (3.97), et (3.99), les approximations données par les limites dans (3.98), et (3.100) seront justifiées.

Cependant, considérer les variables Y_{j_1,j_2}, $1 \leq j_1 \leq s$, $1 \leq j_2 \leq s$, comme indépendantes reviendrait à ne pas prendre en compte le fait que leur somme vaut n (cf. le §3.2). Remarquons de plus que le théorème 3.4.1 montre qu'il n'est pas possible d'appliquer le *théorème central limite* (qui s'applique lorsqu'il y a indépendance).

Pour considérer les variables Y_{j_1,j_2}, $1 \leq j_1 \leq s$, $1 \leq j_2 \leq s$, comme indépendantes et prendre en compte le fait que leur somme est égale à n, nous considérerons les variables conditionnelles,

$$\left(Y_{j_1,j_2} \quad \middle| \quad \sum_{j_1=1}^{s} \sum_{j_2=1}^{s} Y_{j_1,j_2} = n \right). \tag{3.102}$$

Les Y_{j_1,j_2}, $1 \leq j_1 \leq s$, $1 \leq j_2 \leq s$ étant considérées comme des variables aléatoires indépendantes de loi de Poisson de paramètre $\lambda = n/k$, la loi jointe des variables condi-

tionnelles définies en (3.102) s'écrit à l'aide de la *formule de Bayes*,

$$\begin{aligned}
&\mathbb{P}\left(Y_{1,1} = n_1, \ldots, Y_{s,s} = n_k \,\middle|\, \sum_{j_1=1}^{s} \sum_{j_2=1}^{s} Y_{j_1,j_2} = n \right) \\
&= \frac{\mathbb{P}\left(Y_{1,1} = n_1, \ldots, Y_{s,s} = n_k \ , \ \sum_{j_1=1}^{s} \sum_{j_2=1}^{s} Y_{j_1,j_2} = n \right)}{\mathbb{P}\left(\sum_{j_1=1}^{s} \sum_{j_2=1}^{s} Y_{j_1,j_2} = n \right)} \\
&= \frac{\exp(-k\lambda)\, \lambda^n}{n_1! \ldots n_k!} \, \frac{n!}{\exp(-k\lambda)(k\lambda)^n} \\
&= \frac{n!}{n_1! \ldots n_k!} \left(\frac{1}{k} \right)^n . \qquad (3.103)
\end{aligned}$$

Les variables conditionnelles (3.102) suivent donc une loi multinomiale de paramètres $(n, (1/k, \ldots, 1/k))$.

Ainsi, que l'on choisisse de considérer les variables représentant le nombre d'éléments contenus dans une cellule A_{j_1,j_2}, $1 \leq j_1 \leq s$, $1 \leq j_2 \leq s$, comme des variables aléatoires dépendantes de loi binomiale (cf. (3.13), (3.14)), ou comme des variables aléatoires conditionnelles (cf. (3.102)), la loi jointe de l'ensemble de ces variables est la même (loi multinomiale de paramètres $(n, (1/k, \ldots, 1/k))$).

Dans ce qui suit, nous ferons l'hypothèse que pour $1 \leq j_1 \leq s$, $1 \leq j_2 \leq s$, les variables Y_{j_1,j_2} sont indépendantes et équidistribuées de loi de Poisson de paramètre $\lambda = n/k$. A l'aide de la notation (3.95), pour un entier $b \geq 1$, nous introduisons la variable conditionnelle,

$$M_b^{(cond)}\left(Y_{1,1}, \ldots, Y_{s,s} \right) = \left(M_b\left(Y_{1,1}, \ldots, Y_{s,s} \right) \,\middle|\, \sum_{j_1=1}^{s} \sum_{j_2=1}^{s} Y_{j_1,j_2} = n \right). \quad (3.104)$$

Ainsi définie, $M_b^{(cond)}$ correspond à la « *conditional two-dimensional discrete scan statistic* » (voir Glaz *et al.* (2001)). Il est alors possible d'effectuer différentes approximations de la loi de cette statistique conditionnelle.

3.5.2 Lois des « *conditional two-dimensional discrete scan statistic* »

Nous présentons ci-dessous des approximations de la loi de $M_b^{(cond)}\left(Y_{1,1}, \ldots, Y_{s,s} \right)$ (cf. (3.104)). Pour une démonstration détaillée des expressions que nous fournirons, nous renvoyons à Glaz *et al.* (2001), Chen et Glaz (2002).

Nous utilisons la nomenclature suivante.

Pour un entier $b \geq 1$, pour un indice $1 \leq j_1 \leq s-b+1$, nous notons

$$E_{j_1}(m) \quad = \quad \left\{ \bigcap_{j_2=1}^{s-b+1} \left[W_{bj_1,j_2}\left(Y_{1,1},\ldots,Y_{s,s}\right) < m^2 \right] \right\}, \tag{3.105}$$

(où W_{bj_1,j_2} est définie en (3.91)). Nous appelons « colonne » du carré unité, une bande verticale de largeur b/s constituée de $b \times (sb)$ cellules (voir Figure 3.3). L'événement E_{j_1}, $1 \leq j_1 \leq s-m+1$, consiste à considérer que les cardinalités des pavés carrés de $b \times b$ cellules constituant une « colonne » (dont l'abscisse inférieure gauche est en $(j_1-1)/s$) sont toutes strictement inférieures à m^2.

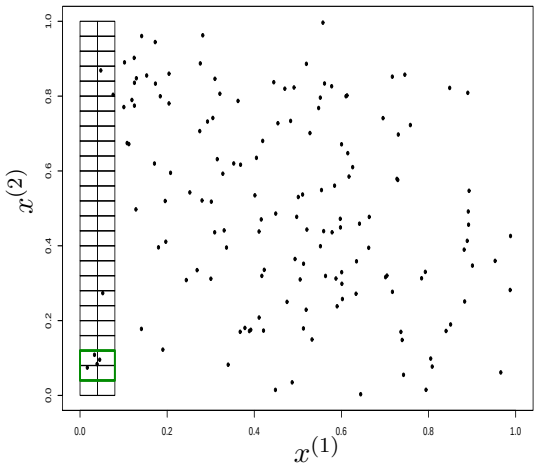


FIG. 3.3 – Pavé (en vert) constitué de 2×2 cellules contenu dans une « colonne d'abscisse inférieure gauche 0 » pour une partition de $[0,1)^2$ en $k = 25 \times 25$ cellules.

Nous notons l'événement,

$$B \quad = \quad \left(\sum_{j_1=1}^{s} \sum_{j_2=1}^{s} Y_{j_1,j_2} = n \right). \tag{3.106}$$

Les approximations de la loi de $M_b^{(cond)}$ feront intervenir les quantités $\mathbb{P}(E_1(m)|B)$, $\mathbb{P}(E_1(m) \cap E_2(m)|B)$, $\mathbb{P}(E_1(m) \cap E_3(m)|B)$ et $\mathbb{P}(E_1(m) \cap E_2(m) \cap E_3(m)|B)$. Pour calculer ces quantités, nous procéderons de la manière suivante.

Soit $(N_{1,1},\ldots,N_{s,s})$ un vecteur de loi multinomiale de paramètres $(n,(1/s^2,\ldots,1/s^2))$.

On déduit de (3.94) et (3.103) (où $k := s^2$), que, pour $1 \leq j_1 \leq s-b+1$, $1 \leq j_2 \leq s-b+1$,

$$\left(W_{bj_1,j_2} \leq m^2 \mid B\right) \stackrel{\mathcal{L}}{=} \left(\sum_{j_1=t}^{s+b-1} \sum_{j_2=t}^{s+b-1} N_{j_1,j_2} \leq m-1\right). \tag{3.107}$$

Ainsi, pour $t = 1, 2, 3$, on a

$$\mathbb{P}\left\{\bigcap_{r=1}^{t} E_r(m) \,|B\right\} = \mathbb{P}\left\{\bigcap_{r=1}^{t} \bigcap_{j_2=1}^{s-b+1} \left(\sum_{j_1=t}^{s+b-1} \sum_{j_2=t}^{s+b-1} N_{j_1,j_2} \leq m-1\right)\right\}, \tag{3.108}$$

$$\mathbb{P}\left\{E_1(m) \cap E_3(m) \mid B\right\} = \mathbb{P}\left\{\bigcap_{r=1,3} \bigcap_{j_2=1}^{s-b+1} \left(\sum_{j_1=t}^{s+b-1} \sum_{j_2=t}^{s+b-1} N_{j_1,j_2} \leq m-1\right)\right\}. \tag{3.109}$$

Pour estimer les quantités ci-dessus, nous ferons des simulations de vecteurs aléatoires $N_{1,1}, \ldots, N_{b+2,s}$ de loi

$$\begin{aligned} &\mathbb{P}\left(N_{1,1} = n_{1,1}, \ldots, N_{b+2,s} = n_{b+2,s}\right) \\ &= \binom{n}{n_{1,1} \ldots n_{b+2,s} \, (n-n^*)} \left(\frac{1}{s}\right)^n \left(1 - \frac{(b+2) \times s}{s^2}\right)^{n-n^*}, \end{aligned} \tag{3.110}$$

où $n^* = \sum_{j_1=1}^{b+2} \sum_{j_2=1}^{s} n_{j_1,j_2}$. Nous organiserons ces vecteurs sous forme de matrices de taille $(b+2) \times s$. Pour l'estimation de $\mathbb{P}(E_1(m) \,|\, B)$ nous considérerons les b premières lignes des ces matrices. Nous calculerons la proportion de ces matrices (de taille $b \times s$) telles que, les sommes des éléments de toutes les sous-matrices carrées $b \times b$ que l'on peut définir sont strictement inférieures à m.Les estimations de $\mathbb{P}(E_1(m) \cap E_2(m)|B)$, $\mathbb{P}(E_1(m) \cap E_3(m)|B)$ et $\mathbb{P}(E_1(m) \cap E_2(m) \cap E_3|B)$ sont analogues.

• Une approximation de la loi de $M_b^{(cond)}$ (cf. (3.104)) est donnée par

$$\begin{aligned} &\mathbb{P}\left(M_b^{(cond)}\left(Y_{1,1}, \ldots, Y_{s,s}\right) \geq m\right) \\ &\approx 1 - \frac{\mathbb{P}\left(E_1(m) \cap E_2(m) \cap E_3(m)|B\right)^{s-b-1}}{\mathbb{P}\left(E_1(m) \cap E_2(m)|B\right)^{s-b-2}}. \end{aligned} \tag{3.111}$$

Nous renvoyons à Chen et Glaz (2002) pour la démonstration de cette approximation.

• En utilisant des lois de Poisson, il est possible d'obtenir

$$\begin{aligned} &\mathbb{P}\left(M_b^{(cond)(Y_{1,1},\ldots,Y_{s,s})} \geq m\right) \\ &\approx 1 - \exp(-\beta(m)), \end{aligned} \tag{3.112}$$

avec,

$$\beta(m) = (s-b+1)(1 - \mathbb{P}(E_1(m)|B)).$$

• Par composition de lois de Poisson, nous avons (voir Ross (1993, 1994) et Glaz *et al.* (2001)),

$$\begin{aligned}\mathbb{P}\left(M_b^{(cond)}\left(Y_{1,1},\ldots,Y_{s,s}\right)\geq m\right)\\ \approx\ 1-\exp(-\beta_1(m)-\beta_2(m)-\beta_3(m)),\end{aligned}\tag{3.113}$$

avec,

$$\beta_i(m) \;=\; \frac{1}{i}\left[2\pi_{1,i}+(s-b-1)\pi_{2,i}\right],\quad i=1,2,3,$$

où,

$$\beta_1(m) \;=\; \frac{1}{i}\left[2\pi_{1,i}+(s-b-1)\pi_{2,i}\right]\quad i=1,2,3,$$

et,

$$\begin{aligned}
\pi_{1,1} &= \mathbb{P}(E_1(m)|B)-\mathbb{P}(E_1(m)\cap E_2(m)|B),\\
\pi_{1,2} &= 1-2\mathbb{P}(E_1(m)|B)+\mathbb{P}(E_1(m)\cap E_2(m)|B),\\
\pi_{2,1} &= \mathbb{P}(E_1(m)\cap E_3(m)|B)-\mathbb{P}(E_1(m)\cap E_2(m)\cap E_3(m)|B),\\
\pi_{2,2} &= 2\left[\mathbb{P}(E_1(m)B)-\mathbb{P}(E_1(m)\cap E_2(m)|B)-\mathbb{P}(E_1(m)\cap E_3(m)|B)\right.\\
&\quad \left.+\,\mathbb{P}(E_1(m)\cap E_2(m)\cap E_3(m)|B)\right],\\
\pi_{2,3} &= 1-3\mathbb{P}(E_1(m)B)+2\mathbb{P}(E_1(m)\cap E_2(m)|B)+\mathbb{P}(E_1(m)\cap E_3(m)|B)\\
&\quad -\mathbb{P}(E_1(m)\cap E_2(m)\cap E_3(m)|B).
\end{aligned}$$

• A l'aide d'inégalités à la *Bonferroni*, on montre que (voir Glaz *et al.* (2001))

$$\begin{aligned}\mathbb{P}\left(M_b^{(cond)}\left(Y_{1,1},\ldots,Y_{s,s}\right)\geq m\right) \;\leq\; & 1+(s-b-1)\,\mathbb{P}(E_1(m)|B)\\ & -(s-b)\,\mathbb{P}(E_1(m)\cap E_2(m)|B),\end{aligned}\tag{3.114}$$

et

$$\begin{aligned}\mathbb{P}\left(M_b^{(cond)}\left(Y_{1,1},\ldots,Y_{s,s}\right)\geq m\right) \;\leq\; & 1+(s-b-1)\,\mathbb{P}(E_1(m)\cap E_2(m)|B)\\ & -(s-b-1)\,\mathbb{P}(E_1(m)\cap E_2(m)|B).\end{aligned}\tag{3.115}$$

Le Tableau 3.2 présente le calcul des expressions (3.111), (3.112), (3.113), (3.114), pour différentes valeurs de n, k, b, et m (voir aussi Glaz). La colonne Simu présente des estimations obtenues par simulation dans le contexte du §3.2. On fait n_{simu} simulations de n variables aléatoires dans le carré unité. Pour chaque simulation, on considère le nombre maximal de points contenus dans un pavé de $b\times b$ cellules. La colonne Simu donc représente la proportion de ce nombre tel que celui-ci est supérieur à m. Ce tableau montre que, dans le contexte du §3.2 la considération de $M_b^{cond}\left(Y_{1,1},\ldots,Y_{s,s}\right)$ présentée

s	b	n	m	(3.111)	(3.112)	(3.113)	(3.114)	Simu
25	5	150	12	0.8952	....	....	....	0.945
			13	0.6784	....	....	....	0.722
			14	0.4006	....	....	....	0.415
			15	0.1895	....	....	....	0.21
			16	0.0794	....	....	....	0.085
			17	0.0278	....	....	....	0.038
			18	0.0100	....	....	....	0.014
			19	0.0027	....	....	....	0.006
25	5	300	20	0.9193	....	....	....	0.946
			21	0.7766	....	....	....	0.819
			22	0.5761	....	....	....	0.59
			23	0.3747	....	....	....	0.369
			24	0.2137	....	....	....	0.201
			25	0.1097	....	....	....	0.106
			26	0.0521	....	....	....	0.052
			27	0.0252	....	....	....	0.02

TAB. 3.2 – Approximations de $\mathbb{P}(M_b^{(cond)}(Y_{1,1},\ldots,Y_{s,s}) \geq m)$

ci-dessus fournit des résultats corrects.

Pour effectuer le test de répartition uniforme des variables $X_1,\ldots,X_n$ dans $[0,1)^2$, nous chercherons n_{bmax}, le nombre maximum de variables contenues dans un pavé consitué de $b \times b$ cellules. Nous rejetterons le test si

$$\mathbb{P}\left(M_b^{(cond)}(Y_{1,1},\ldots,Y_{s,s}) \geq n_{bmax}\right) \leq a, \tag{3.116}$$

où a est la puissance du test.

3.5.3 Cas continu

Il est aussi possible d'utiliser les « *scan-statistics* » dans le cas continu (i.e. sans effectuer de partition de l'espace $\mathcal{X}$ en cellules). Pour des valeurs, $0 < u_1 < 1$ et $0 < u_2 < 1$ fixées, on désigne par $W_{t_1,t_2}(u_1,u_2;Y_{1,1},\ldots,Y_{s,s})$ le nombre de variables $X_1,\ldots,X_n$, contenues dans un rectangle $[t_1,t_1+u_1)\times[t_2,t_2+u_2) \subset [0,1)^2$, où $0 \leq t_1 \leq 1-u_1$, $0 \leq t_2 \leq 1-u_2$, et on considère la statistique,

$$M_{u_1,u_2}(Y_{1,1},\ldots,Y_{s,s}) \;=\; \max_{\substack{0\leq t_1\leq 1-u_1\\ 0\leq t_2\leq 1-u_2}} \{W_{t_1,t_2}(u_1,u_2;Y_{1,1},\ldots,Y_{s,s})\}. \tag{3.117}$$

• D'après Naus (1965), des minorations et des majorations de la loi de probabilité de

cette variable convergent, pour de faibles valeurs de u_1 et u_2 vers,

$$\mathbb{P}(M_{u_1,u_2}(Y_{1,1},\ldots,Y_{s,s}) \geq m) \quad = \quad m^2 \binom{n}{s} (u_1u_2)^{m-1} + o\left([u_1u_2]^{k-1}\right). \quad (3.118)$$

• Lorsque $u_1 > 0$, $u_2 > 0$, et $\varepsilon > 0$, sont tels que $m = Nu_1u_2(1+\varepsilon)$ est un entier, pour de grandes valeurs de n, on a (voir Loader (1991)),

$$\begin{aligned}\mathbb{P}\left(M_{u_1,u_2}(Y_{1,1},\ldots,Y_{s,s}) \geq m\right) \quad &\approx \quad \frac{m^2u_1u_2(1-u_1)(1-u_2)\varepsilon^3}{(1-u_1u_2)^3(1+\varepsilon)} B(m,n,u_1u_2) \\ &+ \left[\frac{nu_2(1-u_1)\varepsilon}{1-u_1u_2} + \frac{nu_1(1-u_2)\varepsilon^2}{(1+\varepsilon)(1-u_1u_2)^2}\right. \\ &+ \quad \left.\frac{(1+\varepsilon)(1-u_1u_2)}{\varepsilon}\right] B(m,n,u_1u_2),\end{aligned} \quad (3.119)$$

avec

$$B(m,n,u_1u_2) \quad = \quad \binom{n}{m} (u_1u_2)^m (1-u_1u_2)^{n-m}. \quad (3.120)$$

• Nous considérons à présent l'hypercube unité $\mathcal{X} = [0,1]^d$. De façon analogue au cas $d = 2$, nous présentons la loi de $M_{u_1,\ldots,u_d}(Y_{1,\ldots,1},\ldots,Y_{s,\ldots,s})$, le nombre maximal de variables $X_1,\ldots,X_n \in \mathcal{X}$ contenues dans des hyperrectangles de côté $u_1,\ldots,u_d$ parallèles aux axes. Nous notons $w = u_1 \times \ldots \times u_d$, le volume de ces hyperrectangles. Tu (1997) (voir aussi Glaz *et al.* (2001)) a obtenu l'approximation suivante,

$$\begin{aligned}&\mathbb{P}\left(M_{u_1,\ldots,u_d}(Y_{1,\ldots,1},\ldots,Y_{s,\ldots,s}) \geq m\right) \\ &\quad \approx \quad \left(1 - \frac{nw}{m(1-w)}\right)^{2d-1} \frac{m^r}{w} B(m,n,w).\end{aligned} \quad (3.121)$$

Pour effectuer le test de répartition uniforme des variables $X_1,\ldots,X_n$ dans $[0,1)^d$, nous chercherons $n_{u_1,\ldots,u_d\,max}$, le nombre maximum de variables contenues dans un pavé de la forme $[r_1, r_1+u_1) \times \ldots \times [t_d, t_d+v_d) \subset \mathcal{X}$. Nous rejetterons le test si

$$\mathbb{P}(M_{u_1,\ldots,u_d}(Y_{1,\ldots,1},\ldots,Y_{s,\ldots,s}) \geq n_{u_1,\ldots,u_d\,max}) \leq a, \quad (3.122)$$

où a est la puissance du test.

En pratique, il est cependant délicat d'effectuer des tests à l'aide de $M_{u_1,\ldots,u_d}(Y_{1,\ldots,1}, \ldots, Y_{s,\ldots,s})$. Pour n fixé, la précision des formules présentées ci-dessus concernent les probabilités $\mathbb{P}\left(M_{u_1,\ldots,u_d}(Y_{1,\ldots,1},\ldots,Y_{s,\ldots,s}) \geq m\right)$ faibles (≤ 0.1 le plus souvent). La puissance du test a sera alors très faible, et un test sera systématiquement rejeté.

3.6 Utilisation

Nous nous plaçons dans le contexte décrit au §3.2 où,

a) la dimension d de $\mathcal{X} = [0,1)^d$ est fixée et est relativement élevée,

b) le nombre n de variables aléatoires $X_1, \ldots, X_n$ dans $\mathcal{X}$ est fixé à l'avance et limité,

c) nous pouvons choisir le nombre k de cellules comme une puissance d du nombre s de segments sur chaque axe de l'hypercube unité, $k = s^d$.

Pour $Y_{1,\ldots,1}, \ldots, Y_{s,\ldots,s}$ définies en (3.12), nous utiliserons les statistiques,

$$\begin{aligned} S_{f_1} &:= S_{f_1\, n,k}\left(Y_{1,\ldots,1}, \ldots, Y_{s,\ldots,s}\right), \\ S_{eq(m)} &:= S_{eq(m)}\left(Y_{1,\ldots,1}, \ldots, Y_{s,\ldots,s}\right), \\ S_{up(m)} &:= S_{up(m)}\left(Y_{1,\ldots,1}, \ldots, Y_{s,\ldots,s}\right), \\ M_b^{(cond)} &:= M_b^{(cond)}\left(Y_{1,\ldots,1}, \ldots, Y_{s,\ldots,s}\right). \end{aligned}$$

Nous renvoyons aux équations (3.72), (3.73), et (3.104) pour leurs définitions.

Comme précisé dans le §3.4 (« *Sparse case* »), ces statistiques permettent de tester l'hypothèse d'indépendance et de répartition uniforme de variables $X_1, \ldots, X_n \in \mathcal{X} = [0,1)^d$. Cependant, elles s'interprètent de façon différente et permettent d'étudier plusieurs caractéristiques de $X_1, \ldots, X_n$. Ci-dessous, nous précisons leurs interprétations, et présentons quelques applications des tests présentés au §3.4 et 3.5.

• Utilisation de S_{f_1}

Effectuer un test à l'aide de la statistique S_{f_1} revient à effectuer un *test d'adéquation (ou ajustement) de loi de probabilité*. Nous testons en fait l'hypothèse $p = (1/k, \ldots, 1/k)$ où p est la loi de probabilité des $\mathbb{1}_{A_j}(X_i)$, $j = 1, \ldots, k$ (voir le §3.2). La statistique S_{f_1} s'interprète comme une *« distance »* entre la loi de probabilité $p = (1/k, \ldots, 1/k)$ et la loi de probabilité empirique $\hat{p} = (Y_1/k, \ldots, Y_k/k)$ (cf. les §3.2 et §3.4).

Nous choisirons s, tel que le nombre de cellules k soit supérieur à n (i.e. $k = s^d > n$). Nous augmenterons ensuite s afin d'effectuer plusieurs tests. Ce test peut parfois être rejeté pour certaines valeurs de s et accepté pour d'autres. Par exemple, pour la suite de variables $X_1, \ldots, X_{150}$ représentée Figure 3.4 (150 points dans $\mathcal{X} = [0,1)^2$), pour un *niveau de puissance* $a = 0.1$, le test a été rejeté pour $k = 13^2, 14^2$ et accepté pour $k = 15^2$ ([1]). Pour comprendre ce phénomène, on effectuera, par exemple, différents tests à l'aide

[1] il est rejeté aussi pour $s = 16^2, 17^2, 18^2, 20^2$ et accepté pour $s = 19^2$

de $S_{eq(m)}$, ou $S_{up(m)}$. Précisons que pour $s = 15$, le calcul de la région d'acceptation de niveau $a = 0.1$ (complémentaire de la région critique, cf. le §3.1), obtenu à l'aide de (3.81), (3.82) et (3.83), aboutit à $\mathcal{R}_a^c(S_{f_1}) = [191, 259]$. Nous observons une valeur de la statistique $S_{f_1} = 249 \in [191, 259]$.

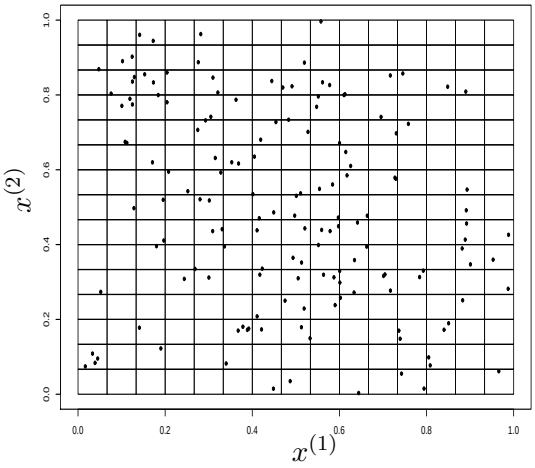


FIG. 3.4 – Ensemble de 150 points dans $\mathcal{X} = [0, 1)^2$

• Utilisation de $S_{eq(m)}$ et $S_{up(m)}$

La première de ces statistiques, $S_{eq(m)}$, s'interprète comme le nombre de cellules contenant exactement m points. Lorsqu'un test effectué à l'aide de S_{f_1} est rejeté, on pourra utiliser cette statistique afin de caractériser une des propriétés singulières de la suite de variables $X_1, \ldots, X_n$ étudiée. Par exemple lorsque $m = 0$, $S_{eq(m)}$ représente le nombre de cellules vides (ne contenant aucune variable $X_1, \ldots, X_n$), et peut donc s'interpréter comme le nombre de « trous » dans la partition de $\mathcal{X}$. Le rejet d'un test faisant usage de $S_{eq(m)}$ indique que le nombre de cellules vides est soit insuffisant, soit trop important (par comparaison au nombre obtenu lorsque $X_1, \ldots, X_n$ est une suite de v.a. i.i.d. de loi uniforme).

Pour $m > 0$, la statistique $S_{up(m)}$ correspond au nombre de cellules contenant au moins m points. Ainsi, nous pourrons évaluer si cette quantité est soit trop faible, soit trop élevée, en comparaison avec une suite de v.a i.i.d. uniformes.

En pratique, $k = s^d$ doit être relativement élevé, de façon à ce que puisse être étudiée la répartition des variables $X_1, \ldots, X_n$ dans de nombreuses sous-parties (cellules) de $\mathcal{X}$. Comme nous avons choisi k supérieur à n, le nombre m intervenant dans $S_{eq(m)}$ et $S_{up(m)}$ devra être faible. En effet, pour m grand, les probabilités $\mathbb{P}(S_{eq(m)} \geq m)$, $\mathbb{P}(S_{up(m)} \geq m)$,

m	$S_{eq(m)}$	$\mathcal{R}_a^c\left(S_{eq(m)}\right)$	$S_{up(m)}$	$\mathcal{R}_a^c\left(S_{up(m)}\right)$
0	122	$[109, 122]$		
1	65	$[66, 88]$	103	$[103, 116]$
2	31	$[20, 32]$	38	$[28, 37]$
3	5	$[3, 9]$	7	$[4, 10]$
4	2	$[0, 2]$	2	$[0, 2]$

TAB. 3.3 – Valeurs et régions d'acceptation de $S_{eq(m)}$ et $S_{up(m)}$ pour $X_1, \ldots, X_{150}$ (cf. Figure 3.5) ($k = 15^2$)

convergent vers 0.

Reprenons l'exemple de l'ensemble de 150 variables dans $\mathcal{X} = [0, 1)^2$ illustré par la Figure 3.4. Nous fixons $k = 15^2$ et essayons de comprendre pourquoi le test effectué avec S_{f_1} a été accepté. Le Tableau 3.3 reproduit les différentes valeurs observées pour $m = 0, 1, 2, 3, 4$ des statistiques $S_{eq(m)}$ et $S_{up(m)}$, ainsi que leurs *régions d'acceptation*, $\mathcal{R}_a^c\left(S_{eq(m)}\right)$, $\mathcal{R}_a^c\left(S_{up(m)}\right)$ (complémentaires des régions critiques voir le §3.1), pour un niveau de puissance $a = 0.1$.

Les valeurs en rouge correspondent aux statistiques qui permettent de rejeter le test. Par comparaison à une suite de variables $U_1, \ldots, U_{150}$ uniformes et indépendantes dans $\mathcal{X} = [0, 1)^2$, on déduit de $S_{eq(1)} = 65 < 66$ que la proportion de cellules contenant exactement une variable est trop faible, et de $S_{up(2)} = 38 > 37$ que le nombre de cellules contenant au moins deux variables est trop important. Nous remarquons aussi que les statistiques, $S_{eq(0)} = 122$, et $S_{up(1)} = 103$, atteignent, respectivement, les bornes, supérieure, et inférieure, de leurs régions d'acceptation.

• Utilisation de $M_b^{(cond)}$

Par sa définition, $M_b^{(cond)}$ permet de « localiser » un groupe de variables $X_1, \ldots, X_n$ contenant un nombre « important » de variables $X_1, \ldots, X_n$. La recherche de ce groupe se fait ici parmi tous les pavés carrés composés de $b \times b$ cellules (contrairement aux statistiques $S_{eq(m)}$ où l'on considère simplement chaque cellule). La statistique $M_b^{(cond)}$ correspond au nombre maximum de variables contenues dans l'un de ces pavés.

Pour l'exemple des variables $X_1, \ldots, X_{150}$ représentées Figure 3.4, pour $s = 25$, $b = 5$, nous obtenons que $M_b^{(cond)} = 18$. Le pavé constitué de 5×5 cellules correspondant est représenté à la Figure 3.5 par le carré rouge. Le carré vert représente le plus grand pavé carré ne contenant aucun point, pour $s = 25$, la longueur d'un coté de ce carré est $6 \times 1/s$. D'après le tableau 3.2, nous avons,

$$\mathbb{P}\left(M_b^{(cond)} \geq 18\right) \approx 0.1.$$

Ainsi, pour un niveau de puissance $a = 0.1$, la statistique $M_b^{(cond)}$ atteint sa borne supérieure de région d'acceptation.

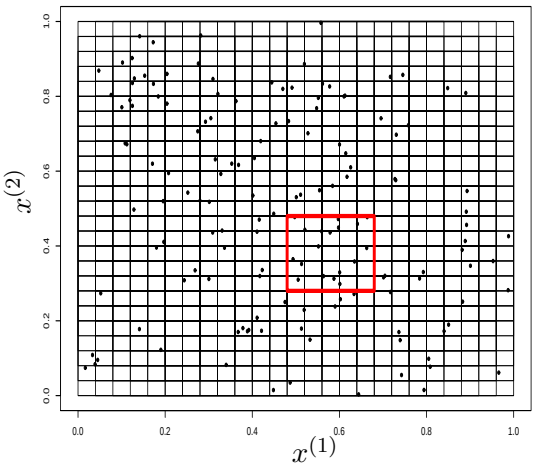


FIG. 3.5 – Localisation d'un groupe de points à l'aide de la « *scan statistic* » parmi l'ensemble illustré 3.4 ($k = 25^2$, $b = 5$)

• L'exemple présenté illustre qu'il est parfois délicat d'interpréter certains résultats de tests. Cependant, les statistiques que nous avons présentées permettent de caractériser quelques propriétés d'un ensemble de variables $X_1, \ldots, X_n$ dans $\mathcal{X} = [0,1)^d$. Par comparaison avec une suite de variables aléatoires indépendantes et uniformes, les résultats concernant l'ensemble illustré Figure 3.4 montrent que la proportion générale de variables dans les cellules semble convenable (test accepté avec S_{f_1}), mais que l'ensemble présente certaines pathologies,

– nombre de cellules vides élevé ,

$$S_{eq(0)} \quad = \quad 122,$$

– nombre de cellules contenant un unique point « anormalement » faible,

$$S_{eq(1)} = 65 < 66,$$

– nombre de cellules contenant plus de deux variables « anormalement » élevé,

$$S_{up(2)} = 38 > 37,$$

– existence d'un groupe de variables de cardinalité importante,

$$\mathbb{P}\left(M_b^{(cond)} \geq 18\right) \quad \approx \quad 0.1.$$

Par conséquent, nous rejetterons l'hypothèse d'indépendance et d'uniformité de la suite $X_1, \ldots, X_{150}$.

3.7 Discussion

Dans la pratique, il est nécessaire d'effectuer l'ensemble des tests présentés, et de faire varier le nombre de cellules de façon à avoir un aperçu global de la répartition de la suite de variables étudiée dans l'hypercube unité[2]. On pourra alors avoir recours au contrôle du risque multiple (voir, par exemple, Lallich *et al.* (2004)).

Pour l'étude d'un ensemble de points dans l'hypercube unité, ces statistiques nous permettront, entre autres, de localiser certaines parties « vides » de l'espace (i.e. ne contenant aucune variables) voir (3.93) du §3.5. Si une spécification de points est possible en vue d'améliorer la répartition uniforme de l'ensemble, celle-ci pourra se faire dans cette partie. La « localisation » de groupes de points redondants se fera à l'aide des « *scan statistics* ». Si une suppression de points est envisageable en vue d'améliorer la répartition uniforme de l'ensemble, nous choisirons un (ou des) point(s) de ce groupe. L'ensemble obtenu devra faire l'objet de nouveaux tests de façon à accepter l'hypothèse de répartition uniforme.

En dimension élevée la partition du pavé unité devient délicate, le nombre de cellules devient vite important puisqu'il s'agit d'une fonction exponentielle de la dimension. Les résultats concernant les lois des « *scan statistics* » discrètes en dimension supérieure à deux font actuellement l'objet de recherche (cf Glaz *et al.* (2001)).

[2]Ces tests ont été programmés avec le logiciel R et sont réalisables en dimension assez grande (8 ou 9) avec un nombre de cellules inférieur à 3^9 (temps de calcul de l'ordre d'une heure pour le cas de la dimension 9 avec 3^9 cellules).

Conclusion

L'objectif des travaux de recherche présentés dans ce mémoire est l'analyse d'une base de données en vue d'effectuer la *calibration d'un code de calcul*. Nous nous sommes placé dans le contexte général où la *fonction de code*, c'est-à-dire le modèle représentant un phénomène expérimental, est analytiquement inconnue. Nous avons aussi supposé que la méthode de calibration, i.e. d'estimation des paramètres, n'a pas été choisie. Ainsi, l'objectif de notre étude consistait à vérifier que les données en entrée de la fonction de code occupent au mieux leur domaine de variation (ce qui correspond à la notion de « *space filling* »). Pour atteindre cet objectif, nous avons distingué deux approches qui se différencient par la modélisation des données,

- ***l'approche déterministe*** où les données sont considérées comme des *variables « déterministes »* ,
- ***l'approche probabiliste*** où les données sont considérées comme des *variables aléatoires*.

• Concernant **l'approche déterministe**, nous avons introduit de nombreux critères utilisés parfois dans des contextes différents (comme l'intégration numérique). Dans ce cadre, ces critères sont *quantitatifs*.

Les premiers critères que nous avons considérés sont définis à l'aide de distances, distances entre les points des données considérées, ou distances entre les points des données et les points de l'espace dans lequel elles sont définies, comme la *dispersion* (cf. §1.2.2). Ils permettent de caractériser la disposition des points dans l'espace, par exemple, la régularité des espacements entre les données, la présence éventuelle de « trous », i.e. de parties de l'espace ne contenant aucun point. Des « défauts » de la répartition uniforme des données dans l'espace peuvent ainsi être identifiés qui risquent d'entraîner une « mauvaise » calibration de paramètres par la suite (faible précision, sensibilité aux variations des données prises en compte pour l'estimation des paramètres).

Nous avons ensuite étudié des critères plus généraux qui permettent de comparer le nombre de points contenus dans des pavés et le volume de ces pavés, comme les critères de *discrépance* (cf. §1.3). Nous avons fait le choix d'étudier la notion générale de discrépance définie à partir d'un noyau auto-reproduisant d'un espace de Hilbert convenable (RKHS pour « *Reproducing Kernel Hilbert Space* », cf. §2.2.2). S'agissant d'un critère général de répartition uniforme de points dans l'espace, nous l'avons utilisé pour proposer des

méthodes de sélection et de spécification d'un ensemble de points en vue d'améliorer sa « qualité de répartition uniforme » (au sens des critères introduits).

La considération de *l'inégalité de Koksma-Hlwaka généralisée* nous a permis de montrer l'existence de liens entre la discrépance et une méthode d'estimation d'un paramètre fonctionnel dite des *fonctions orthogonales. Une discrépance faible d'un ensemble de points permet d'obtenir une estimation robuste* ($\mathbb{MSE}$ *faible*) *et de bonne qualité* ($\mathbb{IMSE}$ *faible*). Parfois certaines méthodes de calibration impliquent une estimation d'un paramètre fonctionnel. Ce pourra être, par exemple, la fonction de code, ou une fonctionnelle faisant intervenir une différence entre la fonction de code et les résultats du phénomène observé. Nous avons donc formellement établi que l'utilisation de la méthode des fonctions orthogonales avec la méthodologie de sélection et de spécification des données présentée au chapitre I permettra d'obtenir une estimation robuste et de bonne qualité. Plus généralement, lorsqu'une méthode d'estimation fera intervenir des moyennes, approximations d'intégrale, l'inégalité de Koksma-Hlwaka généralisée montre que la diminution de la discrépance d'un ensemble de points permettra de réduire l'erreur de l'estimation. Cette propriété pourra aussi être utilisée dans le contexte de l'*apprentissage statistique* où l'on considère une fonction de risque empirique, approximation d'une intégrale (ceci fait notamment l'objet de recherches récentes, voir Cervellera et Muselli (2004), Marry (2005)). Ceci justifie donc les méthodes de sélection et de spécification de points proposées.

Lorsque la modélisation du phénomène fera intervenir un processus aléatoire, cela reviendra à considérer des espaces de Hilbert où le noyau auto-reproduisant correspond à la fonction de covariance du processus. Comme la discrépance est définie à partir d'un RKHS, il semble pertinent d'adapter la définition de la discrépance à la fonction de covariance comme mentionnée par Hickernell (1999) et de poursuivre la recherche de relations dans ce contexte.

• **L'approche probabiliste** nous a permis de prendre en compte le caractère aléatoire des données disponibles. Les critères de répartition uniforme des données correspondent alors à des résultats de *tests statistiques*. On *rejette* ou on *accepte* l'hypothèse d'indépendance et de loi de probabilité uniforme des données (considérées alors comme des *variables aléatoires*). Ici, les critères correspondent à des *règles de décision*.

Précisons que la propriété d'indépendance des données n'étaient pas prise en compte par l'approche déterministe. En effet, nous prenions en compte des critères qui permettaient de vérifier la régularité des espacements entre les données. Or si les données sont régulièrement réparties, celles-ci ne peuvent pas être indépendantes.

Une partition de l'espace des données en pavés disjoints (appelées aussi cellules) ayant été réalisée, les *statistiques* prises en compte pour ces tests permettent de comparer la proportion des données contenues dans ces pavés avec celle d'une suite de variables aléatoires indépendantes et de loi de probabilité uniforme.

Nous nous sommes particulièrement intéressé au cas où le « nombre moyen » de points par cellule est faible (« *sparse case* », cf. §3.4). En effet, en dimension relativement élevée, le nombre de cellules intervenant dans la partition de l'espace devient très important.

En pratique, il est alors délicat de disposer d'un nombre suffisant de points pour qu'il puisse y en avoir au moins un par cellule. Dans ce contexte, les lois de statistiques classiques utilisées pour effectuer des tests sont donc modifiées. Par exemple, la statistique de Pearson et la statistique du rapport de vraisemblance ont une loi gaussienne.

Nous avons aussi étudié les « statistiques par balayage de l'espace », ou « *scan statistics* ». L'objectif de ces statistiques est de vérifier qu'il n'existe pas de groupe de points de cardinalité trop importante parmi les données (par comparaison avec une suite de variables aléatoires indépendantes et uniformes). A l'aide d'approximations (cf. §3.5.1), nous nous sommes focalisé sur les *scan statistics discrètes conditionnelles*. Celles-ci semblent particulièrement adaptées au contexte de notre étude. L'étude des lois de ces statistiques en dimension élevée est récente et fait actuellement l'objet de recherches (voir Glaz *et al.* (2001)).

En pratique, il est nécessaire d'appliquer plusieurs tests à l'aide des différentes statistiques que nous avons introduites. Comme expliqué au Chapitre III, ces différentes statistiques caractérisent certaines propriétés des données. Par exemple, elles permettent d'affirmer qu'il existe de trop nombreuses sous-parties de l'espace sans point, ou qu'il existe des points redondants, i.e. un groupe de points de cardinalité élevée (tout ceci, en comparaison avec une suite de v.a i.i.d. de loi uniforme). Nous serons donc à même d'identifier certains défauts de la répartition uniforme des données, où ici, l'uniformité correspond à la réalisation de variables aléatoires indépendantes et de loi uniformes.

Nous proposons aussi de faire varier le nombre de cellules formant la partition de l'espace prises en compte pour la définition de ces statistiques. Le développement de techniques appropriées dans ce contexte fait actuellement l'objet de recherches (voir Glaz et Zhang (2006)).

Bien que la formalisation des données par l'approche déterministe et par l'approche probabiliste soit différente, certains critères utilisés comportent des « analogies ». Parmi les critères déterministes, la *discrépance* peut s'interpréter comme une comparaison entre le nombre de points contenus dans des pavés et le volume de ces pavés. Les *statistiques* utilisées pour l'approche probabiliste permettent de vérifier que la proportion de points contenus dans des pavés (formant une partition de l'espace des données) est « acceptable » (en comparaison avec celle de variables aléatoires indépendantes de loi de probabilité uniforme). Dans les deux cas, nous étudions donc les mêmes propriétés de la base de données à savoir, la répartition des points dans des sous-parties (des pavés) de leur domaine de variation. Le critère de discrépance tel que nous l'avons utilisé pour l'étude d'une base de données peut notamment être vu comme un critère probabiliste puisque nous considérons une valeur seuil qui correspond à une espérance ou à un quantile lorsque les données sont indépendantes et de loi de probabilité uniforme. La loi de la statistique alors définie n'est pas connue de façon explicite, et nous l'obtenons par simulation. Une perspective de recherche consisterait en l'étude de sa loi exacte.

Le travail réalisé permet à la fois de faire une synthèse des outils pouvant être utiles à l'analyse de la qualité d'une base de données au sens de sa répartition uniforme et de

proposer des techniques de sélection ou de spécification de points en vue de l'améliorer. Nous avons notamment justifié formellement l'emploi de ces techniques dans le cas de l'estimation d'un paramètre fonctionnel par la méthode des fonctions orthogonales. Dans le contexte de la calibration, les outils proposés permettront d'identifier des défauts parmi les données qui risquent d'entraîner une estimation peu robuste de paramètres. Notre approche est heuristique puisqu'il n'est pas formellement établi que des données uniformément réparties permettront d'obtenir une estimation de paramètres convenables quelle que soit la modélisation utilisée pour représenter le phénomène étudié. Dans ce contexte, des perspectives de recherche consisteraient à prendre en compte les réponses du phénomène et/ou du modèle dans l'analyse des données, et à poursuivre l'étude de liens théoriques entre certaines méthodes d'estimation utilisées pour la calibration et les critères que nous avons introduits. Ceux-ci pourraient être établis, par exemple, à l'aide de l'inégalité de Koksma-Hlwaka généralisée.

Annexe

Dans ce qui suit nous utiliserons le vocabulaire et les notations utilisés dans l'introduction du mémoire.

Cette annexe concerne les méthodes de calibration. Il s'agit de détailler quelques méthodes utilisées dans l'étape 3 de la méthodologie définie en introduction. Comme précisé lors de la présentation de l'étape 3, nous distinguerons les cas où il y a « Connaissance de la fonction de code » et ceux où il y a « Absence de connaissance de la fonction de code ».

Dans le premier cas, il s'agira d'un rappel succinct des méthodes bien connues de régression linéaire et non linéaire (voir Walter et Pronzato (1994), Antoniadis *et al.* (1992), et aussi de Crécy et Bazin (2004) par exemple).

Dans le second cas, nous présenterons des techniques générales qui prennent en compte une différence entre les expériences simulées (résultats de la fonction de code) et les observations, une technique mise au point dans le contexte d'application hyodrologique par Beven et Binley (1992), la méthode GLUE (*Global Likelihood Uncertainty Estimation*), une technique d'estimation bayésienne développée par Kennedy et O'Hagan (2001a), et des techniques communément utilisées dans des applications de la chimie, les méthodes dites de *calibration multivariée* ou *d'étalonnage multivarié* (pour *multivariate calibration*, voir Martens et Naes (1991) et Sundberg (1999) ainsi que leurs références).

4.1 « Connaissance » de la fonction de code

Ci-dessous, nous supposons que la « *fonction de code* », f, représentant le phénomène expérimental, est connue analytiquement. Dans ce contexte, les méthodes utilisées pour résoudre le problème de calibration correspondent essentiellement aux techniques classiques et bien connues de régression linéaire ou non linéaire.

Nous supposerons que la relation entre les résultats expérimentaux et la fonction de code peut s'exprimer de la façon suivante :

$$y(x) = f(x, \theta) + \varepsilon(x), \tag{4.1}$$

où,

• $y(x) \in \mathcal{Y} \subset \mathbb{R}$ est la réponse expérimentale en un point $x \in \mathcal{X}$ (cf. Introduction) ;

• $f(x, \theta)$ est le résultat de la fonction de code (ou modèle) en un point $x \in \boldsymbol{x}(n)$, et en $\theta \in \Theta \subset \mathbb{R}^p$, vecteur de paramètres de la calibration (cf. Introduction) ;

• $\varepsilon(x)$ est une variable aléatoire représentant l'erreur entre la réponse expérimentale et la réponse du modèle au point $x \in \mathcal{X}$, liée par exemple à des erreurs de mesure. Ces erreurs seront supposées de même loi de probabilité aux différents points $x \in \mathcal{X}$.

De façon à écrire la modélisation à l'aide de vecteurs, nous désignerons par,

• Y, le vecteur des réponses expérimentales :

$$Y \quad = \quad (y(x_1), \ldots, y(x_n))' \in \mathcal{Y}^n \subset \mathbb{R}^n; \tag{4.2}$$

• $M(\theta)$, le vecteur correspondant aux réalisations du modèle aux points $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$ ($x_i \in \mathcal{X}$, cf. Introduction) connus, et en $\theta \in \Theta$, inconnu, paramètre à calibrer :

$$M(\theta) \quad = \quad (f(x_1, \theta), \ldots, f(x_n, \theta))'; \tag{4.3}$$

• ε, le vecteur aléatoire représentant les erreurs d'observations aux points de la BDDE $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$:

$$\varepsilon \quad = \quad (\varepsilon(x_1), \ldots, \varepsilon(x_n))', \tag{4.4}$$

celui-ci sera supposé centré, et nous notons

$$\begin{array}{ll} V = \sigma^2 R & \text{sa matrice de covariance, où :} \\ R = (K(x_i, x_j))_{1 \leq i \leq n, 1 \leq j \leq n} & \text{pour } x_i, x_j \in \boldsymbol{x}(n). \end{array}$$

Ainsi, la modélisation (4.1) s'écrit sous forme vectorielle comme suit :

$$Y = M(\theta) + \varepsilon. \tag{4.5}$$

Nous estimerons θ par moindres carrés généralisés. C'est-à-dire, nous chercherons à minimiser le critère

$$\mathcal{L}_V(\theta) \quad := \quad \|Y - M(\theta)\|_{V^{-1}} \tag{4.6}$$

où

$$< ., . >_{V^{-1}}, \qquad \text{et} \qquad \|.\|_{V^{-1}}, \tag{4.7}$$

désignent respectivement, le produit scalaire, et la norme, induits par l'inverse de la matrice V (supposée symétrique définie positive).

• Lorsque f est linéaire en θ, lorsque la matrice M est de rang plein, l'estimateur des moindres carrés est donné par :

$$\hat{\theta}_{MC} = \left(M'R^{-1}M\right)^{-1} M'R^{-1}Y. \tag{4.8}$$

Cet estimateur est le meilleur estimateur linéaire sans biais de θ (BLUE). Lorsque les erreurs $\varepsilon(x_i)$, $i = 1, \ldots, n$ sont gaussiennes, il s'agit aussi de l'estimateur du maximum de vraisemblance, il est alors efficace, on a donc,

$$\hat{\theta}_{MC} \sim \mathcal{N}\left(\theta, \sigma^2 \left(M'R^{-1}M\right)^{-1}\right). \tag{4.9}$$

• Lorsque f est non linéaire en θ, lorsque l'opérateur M vérifie un certain nombre de propriétés (voir Antoniadis *et al.* (1992)), l'estimateur $\hat{\theta}_{MC}$ des moindres carrés est solution des équations dites normales :

$$(Y - M(\hat{\theta}_{MC}))'R^{-1}\dot{M}(\hat{\theta}_{MC}) \quad = \quad 0, \tag{4.10}$$

où $\dot{M}(\theta)$ est la matrice jacobienne de M en θ. L'estimateur $\hat{\theta}_{MC}$ est alors estimé par optimisation. Lorsque les erreurs sont gaussiennes, il s'agit aussi du maximum de vraisemblance et il est, sous certaines hypothèses, asymptotiquement efficace.

Nous ne détaillerons pas davantage cette partie et nous renvoyons à Pronzato (1986), Antoniadis *et al.* (1992) et de Crécy et Bazin (2004). Il existe d'autres méthodes d'estimation pouvant être employées pour l'estimation du paramètre. Celles-ci consistent à considérer d'autres critères (critère des moindres valeurs absolues, critère maximal, etc.). Le critère des moindres carrés a l'avantage d'être relativement simple à calculer. De plus, comme il s'agit aussi de l'estimateur du maximum de vraisemblance dans le cas d'erreurs gaussiennes, il dispose d'un ensemble de propriétés particulièrement intéressantes (notamment la propriété d'efficacité). Il est aussi possible d'effectuer une *approche bayésienne* en faisant une hypothèse a priori sur le paramètre θ.

Dans le contexte décrit ci-dessus, il existe de nombreux critères permettant d'analyser la qualité de données $\boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$. Ce sont les critères utilisés pour la construction de *plans d'expériences*. Leur objectif est de réduire une région de confiance de l'estimateur de θ (réduction du volume de cette région, par exemple). Nous renvoyons entre autres à Chernoff (1953), Kiefer (1959, 1961, 1974), Fedorov (1972, 1980), Wynn (1970), Pronzato (1986) et Droesbeke *et al.* (1997).

4.2 Absence de « Connaissance » de la fonction de code

Nous supposons que la fonction de code, f, est inconnue analytiquement. Les méthodes exposées précédemment ne peuvent donc pas être utilisées, puisque celles-ci font

directement intervenir la fonction f (au travers de l'utilisation de l'opérateur M, cf. (4.8) et (4.10)). Dans ce contexte, nous supposerons disponibles des résultats de la fonction de code en différents $x_i \in \mathcal{X}, i = 1, \dots n$ et différents vecteurs de paramètres $\theta_j \in \Theta, j = 1, \dots, k$. Nous les noterons :

$$\boldsymbol{z}(n, \theta_j) = \{z_1(\theta_j) = f(x_1, \theta_j), \dots, z_n(\theta_j) = f(x_n, \theta_j)\},$$

où $\theta_j \in \Theta$, $j = 1, \dots, k$, est connu.

Les premières méthodes que nous décrirons consistent simplement à considérer des différences entre les résultats de la fonction de code et les résultats expérimentaux.

Nous présenterons ensuite la méthode GLUE, qui peut être vue comme une méthode de pondération des vecteurs de paramètres $\theta(k) = \{\theta_1, \dots, \theta_k\}$ utilisés pour les réponses du code $\boldsymbol{z}(n, \theta_1), \dots, \boldsymbol{z}(n, \theta_k)$.

Nous décrirons une approche bayésienne, développée par Kennedy et O'Hagan (2001a), qui est essentiellement adaptée pour réaliser une prédiction de la réponse expérimentale.

Enfin, nous présenterons succinctement les méthodes dites de « calibration multivariée » utilisées essentiellement dans le domaine de la chimie.

4.2.1 Différence entre réponses du code et réponses expérimentales

Une première quantité (ou fonction objectif) que l'on peut considérer pour cette approche est,

$$\begin{aligned} \mathrm{Err}_2(\theta_j) &:= \|Y - Z(\theta_j)\|_2 \\ &:= \left(\sum_{i=1}^{n} |y_i - z_i(\theta_j)|^2 \right)^{1/2}, \end{aligned}$$

où, $\|.\|_2$ représente la norme euclidienne dans $\mathbb{R}^n$, $Y = (y_1, \dots, y_n)' \in \mathbb{R}^n$, le vecteur des réponses expérimentales, et $Z(\theta_j) = (z_1(\theta_j), \dots, z_n(\theta_j))'$, le vecteur des réponses du code en $\theta_j \in \Theta$, $j \in \{1, \dots, k\}$.

Une première façon d'estimer le vecteur de paramètres $\theta \in \Theta$ consiste à poser simplement,

$$\theta^* := \arg \min_{\theta_j \{\in \theta_1, \dots, \theta_k\}} \mathrm{Err}_2(\theta_j). \tag{4.1}$$

D'autres fonctions objectifs peuvent être prises en compte pour l'estimation du vecteur de paramètres, par exemple,

$$\begin{aligned} \mathrm{Err}_1(\theta_j) &:= \sum_{i=1}^{n} |y_i - z_i(\theta_j)|, \\ \mathrm{Err}_\infty(\theta_j) &:= \max_{i \in \{1, \dots, n\}} |y_i - z_i(\theta_j)|. \end{aligned}$$

Par cette approche, le vecteur solution appartient à l'ensemble $\theta(k) = \{\theta_1, \ldots, \theta_k\}$. Il paraît donc important que les paramètres de cet ensemble recouvrent au « mieux » leur domaine de variation. Pour vérifier cette propriété, les différentes techniques présentées aux Chapitres I et II pourront donc être utilisées. Remarquons aussi que l'utilisation de cette méthode implique un nombre élevé d'éléments de l'ensemble $\theta(k) = \{\theta_1, \ldots, \theta_k\}$, et, par conséquent, de disposer de nombreuses réponses de code $\boldsymbol{z}(\theta_j) = (z_1(\theta_j), \ldots, z_n(\theta_j))'$, $j \in \{1, \ldots, k\}$. Lorsque ceci n'est pas possible, on a alors recours à la construction d'un métamodèle (ou surface de réponse) de la fonction critère considérée pour l'estimation du vecteur de paramètres. Nous renvoyons, par exemple, aux études de Pérot (2005), Hervouet *et al.* (2006) et Jones *et al.* (1998). Ces métamodèles permettent de fournir une estimation rapide de la fonction objectif pour tout $\theta \in \Theta$. A l'aide de ces surfaces de réponses, il est alors possible d'utiliser des techniques d'optimisation multiobjectifs (par exemple à l'aide d'algorithmes génétiques, voir Collette et Siarry (2002)) de façon à prendre en compte différentes fonctions objectifs. On obtient alors un ensemble de solutions du vecteur de paramètres (appelé front de Pareto dans le cadre de l'utilisation d'algorithmes génétiques).

4.2.2 Méthode GLUE

La méthode de calibration appelée méthode GLUE pour *Generalized Likelihood Uncertainty Estimation* a été introduite par Beven et Binley (1992). Précisons que la « vraisemblance généralisée » (*generalized likelihood*) considérée par cette méthode ne correspond pas à la définition de la vraisemblance usuelle en statistique.

Conformément aux notations introduites précédemment et en Introduction, nous représentons un phénomène de la façon suivante, pour $x_i \in \boldsymbol{x}(n) = \{x_1, \ldots, x_n\}$, $\theta \in \Theta$,

$$y_i \quad = \quad z_i(\theta) + \varepsilon_i \; .$$

Formellement, le principe de la méthode GLUE consiste à considérer le vecteur des paramètres comme un vecteur aléatoire défini sur Θ, puis à approcher la loi de ce vecteur aléatoire par une loi discrète sur $\theta(k) = \{\theta_1, \ldots, \theta_k\}$. Cette méthode peut être résumée de la façon suivante.

1. On considère que les k vecteurs θ_j ont une loi de probabilité a priori $\pi_0(\theta)$ sur Θ.

2. On désigne par « vraisemblance généralisée » la quantité $L(\theta_j|\boldsymbol{y}) = \pi(Y|\theta_j)\pi_0(\theta_j)$, où $\pi(Y|\theta_i)$ désigne la densité de probabilité conditionnelle du vecteur d'observations $Y = (y_1, \ldots, y_n)'$, et $\pi_0(\theta_j)$ la densité de probabilité a priori en θ_j, $j \in \{1, \ldots, k\}$.

3. On calcule, pour $j \in \{1, \ldots, k\}$,

$$p_j \quad := \quad \frac{L(\theta_j|Y)}{\sum_{\ell=1}^{k} L(\theta_\ell|Y)}. \tag{4.2}$$

Cette méthode consiste simplement à effectuer des pondérations $p_1, \ldots, p_k$ des différents vecteurs de paramètres de $\theta_1, \ldots, \theta_k$. Les couples (θ_j, p_j) sont ensuite utilisés pour déterminer différentes caractéristiques du paramètre à calibrer θ. Une estimation du paramètre par cette méthode peut être donnée par :

$$\hat{\theta}_{GLUE} \quad := \quad \frac{1}{k} \sum_{j=1}^{k} \theta_j L(\theta_j | Y). \tag{4.3}$$

En général, de simples fonctions appelées, par abus de langage, « vraisemblances généralisées », sont utilisées pour le calcul des p_j, $j \in \{1, \ldots, k\}$. Un exemple d'une telle fonction consiste à prendre en compte l'erreur quadratique entre réponses observées et réponses du modèle en $\theta_j \in \theta(k) = \{\theta_1, \ldots, \theta_k\}$,

$$L(\theta_j | Y) := \qquad \left(\frac{1}{\sigma_j}\right)^k, \qquad \text{avec} \qquad \sigma_j^2 := \qquad \frac{1}{n} \sum_{i=1}^{n} (z_i(\theta_j) - y(x_i))^2.$$

A l'aide de cette fonction, pour $j \in \{1, \ldots, k\}$, le poids p_j de θ_j, sera d'autant plus important que le vecteur des réponses du code, $\{z_1(\theta_j), \ldots, z_n(\theta_j)\}$, sera proche du vecteur des réponses expérimentales $\{y_1, \ldots, y_n\}$. Nous renvoyons à Beven et Binley (1992), Ratto *et al.* (2001) et Romanowicz (2006) pour d'autres exemples de fonctions dites de « vraisemblances généralisées », ainsi qu'à leurs références.

Comme pour la méthode présentée au §(4.2.1), la technique GLUE implique de disposer de nombreuses réponses de code $\boldsymbol{z}(\theta_j) = (z_1(\theta_j), \ldots, z_n(\theta_j))'$, $j \in \{1, \ldots, k\}$. Lorsque ceci n'est pas possible, des métamodèles (dépendant de $\theta \in \Theta$) permettant de considérer une substitution de la fonction de code pourront être utilisées de façon à considérer un nombre plus important de paramètres $\theta_j \in \Theta$, $j = 1, \ldots, k$. Les différentes pondérations $p_1, \ldots, p_k$ alors obtenues (telles que $p_1 + \ldots + p_k = 1$) pourront alors être assimilées à des « probabilités », fournissant ainsi une loi de probabilité (a posteriori) discrète du vecteur de paramètre de calibration θ en $\{\theta_1, \ldots, \theta_k\}$. Il apparaît donc nécessaire de vérifier que les paramètres $\theta(k) = \{\theta_1, \ldots, \theta_k\}$ occupent « au mieux » leur domaine de variation Θ. Les outils présentés aux Chapitres I et III pourront être employés à cet effet.

4.2.3 Approche de Kennedy et O'Hagan (2001)

L'objectif de cette méthode est de fournir une prédiction de la réponse expérimentale. La modélisation utilisée par Kennedy et O'Hagan (2001a) est la suivante, pour $i = 1, \ldots, n$:

$$y_i \quad = \quad \zeta_i + \varepsilon_i \tag{4.4}$$

$$\quad = \quad \rho\, z_i(\theta) + \delta_i + \varepsilon_i, \tag{4.5}$$

où,

• y_i, $i = 1, \ldots, n$, représentent des résultats expérimentaux issus de l'observation d'un phénomène en $\{x_1, \ldots, x_n\} \in \mathcal{X}^n$ (cf. Introduction).

• ζ_i, $i = 1, \ldots, n$, représentent les « vrais » résultats du phénomène en $\{x_1, \ldots, x_n\} \in \mathcal{X}^n$, c'est-à-dire, les résultats du phénomène sans sources d'incertitude liées à la nature des observations.

• $z_i(\theta)$, $i = 1, \ldots, n$, représentent les réponses d'une fonction de code en $\{x_1, \ldots, x_n\} \in \mathcal{X}^n$, et en $\theta \in \Theta$ (cf. Introduction). Par la suite, on utilisera un métamodèle de la fonction de code qui interpolera les $\{z_1(\theta_j), \ldots, z_n(\theta_j)\}$, $j = 1, \ldots, k$, déjà réalisés (conformément au contexte décrit au début du §4.2). Le métamodèle sera construit par « krigeage » (voir Cressie (1993), Vazquez (2005), Marrel *et al.* (2006) par exemple). La réponse du code sera donc assimilée à la réalisation d'un processus gaussien stationnaire (isotropique). Nous supposerons que la moyenne de ce processus sera de la forme

$$m_1(x, \theta) \quad = \quad h_1(x, \theta)' \beta_1, \tag{4.6}$$

où $h_1(x, \theta)$ est un vecteur de fonctions en $x \in \mathcal{X}$ et $\theta \in \Theta$, et β_1 est un vecteur de paramètres. La fonction de variance-covariance du processus représentant la réponse de la fonction de code pourra s'écrire de la façon suivante,

$$c_1(u, t)(v, s) \quad = \quad \sigma_1^2 \exp\left[-(u - v)' \, \Omega_x (u - v) - (t - s)' \, \Omega_\theta (t - s)\right], \tag{4.7}$$

où Ω_x et Ω_θ sont des matrices diagonales. On notera le vecteur de paramètres composé de σ_1 et des éléments diagonaux de ces deux matrices ψ_1,

$$\psi_1 \quad := \quad \left(\sigma_1, \mathrm{diag}(\Omega_x)', \mathrm{diag}(\Omega_\theta)'\right)'. \tag{4.8}$$

• ρ, est un paramètre issu d'une régression.

• δ_i correspond à une erreur de « représentativité » de la fonction de code. Nous ferons l'hypothèse que les δ_i sont des variables aléatoires corrélées et indépendantes de la fonction de code. Nous les considérerons comme des réalisations d'un processus gaussien stationnaire (isotropique). La moyenne de ce processus sera de la forme,

$$m_2(x) \quad = \quad h_2(x)' \beta_2, \tag{4.9}$$

où $h_2(x)$ est un vecteur de fonctions en $x \in \mathcal{X}$, et β_1 est un vecteur de paramètres. La fonction de variance-covariance du processus s'écrira comme suit,

$$c_2(u, v) \quad = \quad \sigma_2^2 \exp\left[-(u - v)' \, \Omega_\delta (u - v)\right], \tag{4.10}$$

où Ω_δ est une matrice diagonale. On notera le vecteur des paramètres composé de σ_2 et des éléments diagonaux de cette matrice ψ_2,

$$\psi_2 \quad := \quad (\sigma_2, \mathrm{diag}(\Omega_\delta))'. \tag{4.11}$$

• ε_i, $i = 1, \ldots, n$, correspondent aux erreurs (incertitudes) liées à la nature des observations du phénomène. Nous les considérerons comme des variables aléatoires indépendantes de loi gaussienne, de moyenne nulle, et de variance λ^2,

$$\varepsilon_i \quad \sim \quad \mathcal{N}(0, \lambda). \tag{4.12}$$

Nous renvoyons à O'Hagan (1998) et Kennedy et O'Hagan (2001a) pour justifier l'utilisation de la modélisation décrite par les équations (4.4) et (4.5).

Pour fournir une prédiction de la réponse expérimentale y, nous devrons dans un premier temps estimer les hyperparamètres, ρ (cf. (4.5)), β_1 (cf. (4.6)), β_2 (cf. (4.9)), ψ_1 (cf. (4.8)), ψ_2 (cf. (4.11)), λ (cf. (4.12)). Ci-dessous, nous posons,

$$\beta \quad := \quad (\beta_1', \beta_2'), \tag{4.13}$$
$$\psi \quad := \quad (\psi_1', \psi_2'), \tag{4.14}$$
$$\phi \quad := \quad (\rho, \lambda, \psi). \tag{4.15}$$

Les estimations des hyperparamètres β et ϕ se feront par approche bayésienne. Nous commencerons donc par faire des hypothèses a priori concernant les paramètres introduits ci-dessus.

Hypothèses a priori

Nous supposons disposer de lois a priori des vecteurs de paramètres ϕ (voir (4.15)) et θ. Nous supposons que le vecteur de paramètres $\beta = (\beta_1', \beta_2')$ (voir (4.6), (4.9), (4.13)) suit une loi uniforme,

$$p(\beta_1, \beta_2) \quad \propto \quad 1. \tag{4.16}$$

Nous faisons l'hypothèse d'indépendance de θ et ϕ. Par conséquent, d'après (4.16), nous obtenons que,

$$p(\theta, \beta, \phi) \quad = \quad p(\theta)\, p(\phi). \tag{4.17}$$

Estimation des hyperparamètres

L'estimation des hyperparamètres introduits ci-dessus se fait en deux étapes.

Étape 1

Tout d'abord, nous estimons le vecteur de paramètres ψ_1 intervenant dans la fonction de covariance $c_1(.,.)(.,.)$ (cf. (4.8)) à l'aide des réponses du code $z_i(\theta_j)$, $i = 1, \ldots, n$, $j = 1, \ldots, k$. A cette étape, les θ_j, $j = 1, \ldots, k$, intervenant dans la fonction de code sont connus. Par conséquent, la variable θ ne correspond pas à un paramètre à estimer. Nous notons donc simplement Z, le vecteur aléatoire représentant l'ensemble des réponses du

code déjà réalisées. A l'aide d'une loi a priori sur le vecteur de paramètres ψ_1, notée $p(\psi_1)$, nous estimons le paramètre ψ_1, en considérant la quantité,

$$p(\beta_1, \psi_1|Z) \quad \propto \quad p(\beta_1, \psi_1)p(Z|\beta_1, \psi_1), \tag{4.18}$$

avec β_1 défini en (4.6), et, où par hypothèse, la loi de $Z|\beta_1, \psi_1$, désignée par $p(Z|\beta_1, \psi_1)$, est une loi normale de moyenne $h_1(x, \theta)$ (cf. (4.6)), et de fonction de variance-covariance $c_1(.,.)(.,.)$, (cf. (4.7)). Pour estimer ψ_1, on maximise la quantité (4.18) que l'on aura intégrée par rapport à β_1,

$$\hat{\psi}_1 \quad = \quad \arg\max_{\psi_1} \int p(\beta_1, \psi_1|Z)d\beta_1. \tag{4.19}$$

Étape 2

Nous estimons à présent les paramètres ρ, β_2, ψ_2 (voir (4.5), (4.9), (4.11), respectivement), à l'aide de l'ensemble des observations y_i et des réponses du code $z_i(\theta_j)$, $i = 1, \ldots, n$, $j = 1, \ldots, k$. Nous notons $d = (Z', Y')$, le vecteur constitué de l'ensemble des réponses du code, et de l'ensemble des réponses expérimentales. Nous cherchons à estimer les paramètres ρ, β_2, ψ_2 par maximisation de la quantité

$$p(\beta_2, \rho, \lambda, \psi_2|d, \psi_1). \tag{4.20}$$

Ici, selon la modélisation décrite par la formule (4.5), la réponse du code considérée est obtenue en un paramètre θ (paramètre à calibrer). Nous désignons donc ici par $Z(\theta)$ le vecteur des réponses du code obtenues en θ. Comme la réponse du code est indépendante des paramètres ρ, β_2, ψ_2, nous pouvons considérer, de façon équivalente à (4.20), la quantité,

$$p(\beta_2, \rho, \lambda, \psi_2|d, \psi_1) \quad \propto \quad p(\beta_2, \rho, \lambda, \psi_2)\, p(Y|Z(\theta), \beta_2, \phi), \tag{4.21}$$

où $\phi = (\rho, \lambda, \psi_1', \psi_2')$. Cependant, il n'est pas possible d'obtenir la loi de $Y|Z(\theta), \beta_2, \phi$ de façon analytique. Néanmoins, par hypothèse, nous savons que le vecteur aléatoire $Y|Z(\theta), \beta_2, \phi, \theta$ est normalement distribué. De plus, après calcul (voir Kennedy et O'Hagan (2001b)), il est possible d'obtenir la moyenne et la variance de ce vecteur aléatoire. Nous faisons ensuite l'approximation que $p(Y|Z(\theta), \beta_2, \phi, \theta)$ correspond aussi à la loi de $Y|Z(\theta), \beta_2, \phi$. Nous approchons donc la quantité (4.20) par

$$p(\beta_2, \rho, \lambda, \psi_2|d, \psi_1) \quad \propto \quad p(\beta_2, \rho, \lambda, \psi_2)p(Y|Z(\theta), \beta_2, \phi, \theta). \tag{4.22}$$

Nous estimons alors les paramètres $\beta_2, \rho, \lambda, \psi_2$ par maximisation de (4.22).

Nous désignerons par $\hat{\phi}$ le vecteur correspondant à l'estimation des hyperparamètres ϕ (cf. (4.15),(4.20), (4.19)).

Une fois obtenue $\hat{\phi}$, nous cherchons à approcher la loi a posteriori du vecteur des paramètres de calibration θ. Par hypothèse a priori, le paramètre θ est indépendant de ϕ (voir (4.17)), et β suit une loi uniforme (voir (4.16)), ce qui implique,

$$p(\theta, \beta, \phi) \quad = \quad p(\theta)\, p(\phi). \tag{4.23}$$

Nous approchons la loi a posteriori du vecteur de paramètres θ en considérant que

$$p(\theta|\phi = \hat{\phi}, d) \quad \propto \quad p(\theta)\, p(d|\theta, \phi = \hat{\phi}). \tag{4.24}$$

Nous considérons ensuite le processus $y(x)|\theta, \phi, d$ (où $d = (Z, Y)$). Celui-ci est gaussien par hypothèse, et, après calculs, il est possible d'obtenir son espérance et sa fonction de variance-covariance (voir Kennedy et O'Hagan (2001b)). En combinant la loi de $y(x)|\theta, \phi, d$ et la loi a posteriori de θ (cf. (4.24)), nous pouvons alors effectuer une prédiction d'une réponse expérimentale. Nous considérerons, par exemple, pour $\phi = \hat{\phi}$,

$$\mathbb{E}(y(x)|\phi, d) \quad = \quad \int_{\Theta} \mathbb{E}\left[y(x)\,|\theta, \phi, d\right] p(\theta|\phi, d) d\theta. \tag{4.25}$$

Rappelons que l'objectif de la méthode décrite ci-dessus est la prédiction de la réponse expérimentale. Elle pourra donc être appliquée, par exemple, dans le contexte où l'on dispose d'un nombre restreint de résultats expérimentaux, ou lorsque le code utilisé pour décrire le phénomène est difficilement réalisable (temps de calcul assez long) et qu'une prédiction rapide de la réponse expérimentale est souhaitée.

L'expression de la loi a posteriori du paramètre θ (cf. (4.24)) semble indiquer qu'il est possible d'obtenir une estimation de ce paramètre. Cependant, tel qu'il a été introduit ici, ce paramètre intervient dans le métamodèle utilisé pour représenter la fonction de code. L'estimation de θ à l'aide de (4.24) (par maximisation de cette quantité, par exemple) ne correspond donc pas à la calibration de la fonction de code, mais plutôt, à la « calibration du métamodèle » qui représente la fonction de code. L'utilisation de cette technique d'estimation du paramètre θ est donc délicate et à manier avec précaution.

Nous avons supposé ici que les réponses expérimentales et les réponses du code sont obtenues pour les mêmes données $x_i \in \mathcal{X}$, $i = 1, \ldots, n$. Ceci n'est en fait pas nécessaire. Nous pouvons donc avoir en entrée de la réponse expérimentale une base de données $\boldsymbol{u}(n) = \{u_1, \ldots, u_n\} \in \mathcal{X}^n$, et en entrée de la fonction de code une base de données $\boldsymbol{v}(m) = \{v_1, \ldots, v_m\} \in \mathcal{X}^m$ et un jeu de paramètres $\theta(k) = \{\theta_1, \ldots, \theta_k\} \in \Theta^k$.

La principale difficulté pour la mise en oeuvre de cette méthode est, à l'étape 2, l'inversion d'une matrice carrée de taille $(n + m) \times (n + m)$, où n est le nombre de données $\boldsymbol{u}(n) = \{u_1, \ldots, u_n\} \in \mathcal{X}^n$ en entrée des réponses expérimentales et m le nombre de données $\boldsymbol{v}(m) = \{v_1, \ldots, v_m\} \in \mathcal{X}^m$ en entrée de la fonction de code (voir

Kennedy et O'Hagan (2001a)). Des données bien choisies pourront permettre de considérer plusieurs matrices de taille bien inférieure à inverser (voir Kennedy et O'Hagan (2001b)). Une autre difficulté concerne l'intégration par rapport à θ dans (4.25). Une méthode envisageable consisterait alors à utiliser les méthodes de sélection et/ou de spécification des paramètres introduites au Chapitre I de ce mémoire, de façon à réduire au mieux l'erreur de l'approximation d'une intégrale par sa moyenne. En effet, ces méthodes ont pour objectif de réduire la discrépance qui intervient dans l'erreur de l'estimation d'une intégration par sa moyenne du fait de l'inégalité de Koksma-Hlwaka (voir Chapitre II).

4.2.4 Méthodes de « calibration multivariée »

Ces méthodes sont essentiellement appliquées dans le domaine de la chimie. Elles consistent à exprimer une relation entre un jeu de données d'entrées et de sorties. On distingue deux approches, l'approche indirecte, et l'approche directe (voir Sundberg (1999)). Dans le contexte décrit en introduction ces deux approches peuvent être vues de la façon suivante.

- Par l'approche indirecte,
  on considérera les paramètres $\theta_j, j = 1, \ldots, k$ de la fonction de code comme des données d'entrées et les résultats $\{\boldsymbol{z}(\theta_1), \ldots, \boldsymbol{z}(\theta_k)\}$ comme des sorties. Les techniques utilisées alors consisteront essentiellement à effectuer des régressions linéaires multivariées.

- Par l'approche directe,
  on considérera les résultats de la fonction de code $\{\boldsymbol{z}(\theta_1), \ldots, \boldsymbol{z}(\theta_k)\}$ comme des données d'entrée, et les différents paramètres, $\theta_j, j = 1, \ldots, k$, comme des données de sortie. On utilisera les techniques de régression PLS pour « prédire » une valeur d'un paramètre $\theta \in \Theta$.

Dans ce qui suit nous présentons succinctement ces deux approches. Certains principes des techniques utilisées pour chacune de ces approches seront décrits de façon simplifiée.

4.2.4.1 Approche indirecte

Les méthodes indirectes consistent à effectuer une régression de $\boldsymbol{z}(\theta) \in \mathbb{R}^n$ sur $\theta \in \Theta \subset \mathbb{R}^p$, puis à en déduire une estimation de θ. Une méthode fréquemment employée est la régression linéaire multiple. On souhaite alors obtenir une relation de la forme :

$$\boldsymbol{z}(\theta) \quad \approx \quad \alpha + B\theta. \tag{4.26}$$

Dans le contexte du paragraphe §4.2, le vecteur α et la matrice B sont des paramètres à estimer. Pour $i = 1, \ldots, n$, on effectue des régressions des vecteurs $(z(x_i, \theta_1), \ldots, z(x_i, \theta_k))'$

sur $\theta_1, \ldots, \theta_k$ pour trouver une relation du type,

$$\begin{pmatrix} z(x_i, \theta_1) \\ \vdots \\ z(x_i, \theta_k) \end{pmatrix} \approx \hat{\alpha}_i \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \begin{pmatrix} \theta_1' \\ \vdots \\ \theta_k' \end{pmatrix} \hat{\beta}_i. \tag{4.27}$$

Les paramètres $\hat{\alpha}_i$ et $\hat{\beta}_i$ issus de la régression de $(z(x_i, \theta_1), \ldots, z(x_i, \theta_k))'$ sur $\theta_1, \ldots, \theta_k$ correspondront respectivement aux composantes du vecteur α, et aux lignes de la matrice B dans (4.26). Sous l'hypothèse que la matrice B est de rang plein, on estime alors θ par moindres carrés à l'aide du vecteur des observations $Y = (y_1, \ldots, y_n)'$,

$$\hat{\theta}_{LS} = \left(\hat{B}\hat{B}'\right)^{-1} \hat{B}\,(Y - \hat{\alpha})\,. \tag{4.28}$$

On peut aussi utiliser l'estimation par moindres carrés généralisés en prenant en compte une matrice Σ de variance-covariance du vecteur des observations Y,

$$\hat{\theta}_{GLS} = \left(\hat{B}\Sigma^{-1}\hat{B}'\right)^{-1} \hat{B}\Sigma^{-1}\,(Y - \hat{\alpha})\,. \tag{4.29}$$

Pour obtenir une approximation de la variance de l'estimation de θ par cette approche, nous renvoyons entres autres à Sundberg (1996).

La technique de calibration décrite ci-dessus ne correspond pas exactement au contexte décrit en Introduction. En effet, cette technique ne prend pas en compte, dans la modélisation, la différence entre la nature des incertitudes liées aux réalisations de la fonction de code et la nature des incertitudes liées aux observations du phénomène. Certaines approches indirectes de calibration en tiennent compte, nous renvoyons par exemple à Brown (2002). L'objectif de ces méthodes est alors d'estimer un paramètre à l'aide d'observations « précises » d'un phénomène, et d'observations « moins précises » de ce même phénomène. Dans notre contexte (cf. Introduction), les observations « précises » correspondraient aux observations du phénomène, i.e. aux expériences, et les observations « moins précises » aux expériences simulées, résultats d'une fonction de code. Les techniques indirectes de calibration multivariée s'appliquent le plus souvent lorsque le phénomène étudié est linéaire par rapport au paramètre à estimer, θ. Ces méthodes pourront donc être utilisées dans ce cadre.

4.2.4.2 Approche directe

Cette approche consiste à établir une relation entre les différentes sorties du code $\{\boldsymbol{z}(\theta_1), \ldots, \boldsymbol{z}(\theta_k)\} \in \mathbb{R}^n \times \mathbb{R}^k$ et les paramètres $\{\theta_1, \ldots, \theta_k\} \in \Theta^k$, où $\Theta \in \mathbb{R}^p$, en effectuant directement une régression de θ sur $\boldsymbol{z}(\theta)$. On cherche donc à obtenir une relation de la forme

$$\theta \approx \sum_{i=1}^{n} \beta_i\, z(x_i, \theta), \tag{4.30}$$

où les β_i sont des coefficients à estimer. Une fois obtenue cette relation, nous pourrons considérer le vecteur des réponses expérimentales $y_1, \ldots, y_n$ (cf. Introduction), et estimer le paramètre de calibration par :

$$\hat{\theta} \quad = \quad \sum_{i=1}^{n} \beta_i \, y_i. \tag{4.31}$$

De façon à employer les notations usuelles en régression, nous posons à présent :

$$X \quad := \quad \begin{pmatrix} z(x_1, \theta_1) & \ldots & z(x_n, \theta_1) \\ \vdots & & \vdots \\ z(x_1, \theta_k) & \ldots & z(x_n, \theta_k) \end{pmatrix}, \tag{4.32}$$

$$Y \quad := \quad \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \end{pmatrix}. \tag{4.33}$$

Nous désignerons par $x_i \in \mathbb{R}^k$, $j = 1, \ldots, n$, une colonne de la matrice X, et $y_j \in \mathbb{R}^p$, $j = 1, \ldots, k$, une ligne de la matrice Y.

Dans ce contexte, une méthode de régression fréquemment employée pour expliquer Y à l'aide de X est la régression PLS. Celle-ci a été introduite par Wold *et al.* (1982) Nous présentons ci-dessous brièvement son principe dans le contexte où la variable à expliquer, y, est unidimensionnelle, $y \in \mathbb{R}$ (il s'agit alors de la régression PLS1, voir Tenenhaus (1998)). Le terme y_j, $j = 1, \ldots, k$ est alors un réel, et Y un vecteur, $Y \in \mathbb{R}^k$. Nous supposerons que les données sont centrées et réduites.

On considère une composante t_1,

$$t_1 \quad := \quad w_{1,i} x_1 + \ldots + w_{1,n} x_n, \tag{4.34}$$

où

$$w_{1,i} \quad := \quad \frac{\mathbb{C}\mathrm{ov}(x_i, y)}{\sqrt{\sum_{i=1}^{n} \mathbb{C}\mathrm{ov}(x_i, y)^2}}. \tag{4.35}$$

On réalise ensuite une régression simple de y sur t_1.

$$y \quad = \quad c_1 t_1 + y_1, \tag{4.36}$$

où c_1 est le coefficient de la régression, et y_1 le vecteur des résidus. Si le pouvoir explicatif de cette régression est trop faible on construit une nouvelle composante t_2. On considère les résidus $x_{1,i}$ des régressions des x_i sur t_1. La composante t_2 est construite de la façon suivante,

$$t_2 \quad := \quad w_{2,1} x_{1,1} + \ldots + w_{2,n} x_{1,n}, \tag{4.37}$$

où

$$w_{2,i} \quad := \quad \frac{\mathbb{C}\mathrm{ov}(x_{1,i}, y_1)}{\sqrt{\sum_{i=1}^{n} \mathbb{C}\mathrm{ov}(x_{1,i}, y_1)}} \, . \tag{4.38}$$

On réalise ensuite une régression de y sur t_1 et t_2,

$$y \quad = \quad c_1 t_1 + c_2 t_2 + y_2, \tag{4.39}$$

où c_2 est le coefficient de la régression, et y_2 le vecteur des résidus. On itère le procédé jusqu'à obtenir une régression ayant un pouvoir explicatif satisfaisant, celui-ci étant déterminé par validation croisée.

Cette méthode se généralise aussi lorsqu'il existe des données manquantes ou lorsque Y est une matrice (régression PLS2). Il est aussi parfois nécessaire d'effectuer un traitement des données X. Nous renvoyons à Tenenhaus (1998) pour une présentation détaillée des algorithmes de régression PLS1, PLS2, et de leurs propriétés mathématiques.

Cette méthode est extrêmement employée dans le domaine de la spectroscopie, notamment la spectroscopie proche infrarouge. Les spectres considérés représentent l'absorbance (mesure de la capacité d'un milieu à absorber la lumière) en fonction de la longueur d'onde. Les $x_1, \ldots, x_p$ correspondent à des valeurs des spectres en différentes longueurs d'ondes. La variable à expliquer est le plus souvent une concentration. Le contexte est donc quelque peu différent de celui fixé en Introduction. Précisons que d'autres méthodes bien différentes peuvent encore être employées, analyse de Fourier, réseaux de neurones. Nous renvoyons à Martens et Naes (1991) ou Naes *et al.* (2002).

Les méthodes présentées dans ce dernier paragraphe ne concernent pas exactement le contexte décrit en Introduction puisque le plus souvent, elles ne font pas intervenir une fonction de code. Elles semblent cependant être appropriées pour fournir une estimation du paramètre recherché. Précisons que, pour ces approches, il existe des méthodes de sélection de données appropriées à la technique de calibration que l'on utilisera. Nous renvoyons entre autres à Martens et Martens (2001), ainsi qu'à leurs propres références.

4.3 Discussion

Dans la pratique, la technique de calibration que l'on utilisera dépendra essentiellement des propriétés de la fonction de code.

- Lorsqu'il y a connaissance de la fonction de code, l'estimation du paramètre se fera à l'aide des méthodes classiques présentées au §4.1. L'estimation ne pose absolument aucune difficulté dans le cas où la fonction de code est linéaire. Elle peut parfois être délicate lorsque celle-ci est non linéaire. Il s'agit alors essentiellement d'un problème d'optimisation.

- Lorsqu'il y a absence de connaissance de la fonction de code, nous pouvons distinguer les cas suivants.

  – Si la fonction de code est « simple » à utiliser (temps de calcul négligeable) et qu'il est possible de réaliser un grand nombre d'appels, alors, les techniques des §4.2.1 et §4.2.2 sont les plus appropriées.

  – Si la fonction de code est complexe (coûteuse en temps de calcul) et que l'on dispose d'un nombre relativement important de résultats déjà réalisés, il est alors préférable de construire un métamodèle de la fonction de code et d'utiliser les techniques des §4.2.1 et §4.2.2 en substituant la fonction de code par la surface de réponse.

  – Si la fonction de code est complexe et que le nombre de résultats de cette fonction est restreint, on utilisera la technique présentée au §4.2.3 et, lorsque le code est linéaire par rapport au paramètre de calibration, on pourra aussi utiliser les techniques d'approche indirecte de calibration multivariée du §4.2.4.1.

Les différentes méthodes présentées illustrent la diversité des domaines d'application de la calibration. L'analyse des données présentée dans ce mémoire pourra être utile pour chacune d'entre elles, et plus particulièrement pour les méthodes présentées aux §4.2.1, §4.2.2 et §4.2.3. En effet, pour ces méthodes, il est semble important que les paramètres $\theta_1, \ldots, \theta_k$, en lesquels sont disponibles les résultats du code, $\boldsymbol{z}(\theta_1), \ldots, \boldsymbol{z}(\theta_k)$, « explorent » (remplissent) « au mieux » l'espace Θ. Ceci permettra de prendre en compte un jeu de paramètres « représentatifs » de l'espace Θ pour la calibration.

Bibliographie

A. Antoniadis, J. Berruyer et R. Carmona : *Régression non linéaire et applications.* Economica, 1992.

A. Araujo et E. Giné : *The central Limit Theorem for Real and Banach Valued Random Variables.* Wiley, 1980.

J. B. Aubin : *Estimation fonctionnelle par projection adaptative et applications.* Thèse de doctorat, Université Pierre et Marie Curie, Paris VI, 2005.

R. Baker : On irregularities of distribution ii. *London Mathematical Society*, 59: 50–64, 1999.

L. Bergman : The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3): 200–217, 1967.

K. Beven et A. Binley : The future of distributed models : model calibration and uncertainty prediction. *Hydrological processes*, 6: 279–298, 1992.

D. Bosq : Sur l'estimation d'une densité multivariée par une série de fonctions orthogonales. *Comptes Rendus de l'Académie des Sciences de Paris*, 268: 555–557, 1969.

D. Bosq : *Inférence et prévision en grandes dimensions.* Economica, 2005.

D. Bosq et J. Bluez : Etude d'une classe d'estimateurs non-paramétriques de la densité. *Annales de l'institut Henri Poincaré*, 14: 479–498, 1978.

D. Bosq et J. Lecoutre : *Théorie de l'estimation fonctionnelle.* Economica, 1987.

P. Brown : Inverse prediction. *Encyclopaedia of Environmetrics*, 2: 1075–1079, 2002.

J. Burbea et C. Rao : On the convexity of divergence measures based on entropy function. *IEEE Transactions on Information Theory*, 28(3): 489–495, 1983.

L. Carraro : Introduction au RKHS. Rencontre GDR MASCOTT, 2007. http ://www.lsp.ups-tlse.fr/Fp/Gamboa/GDR/RKHS_Mascot_2.pdf.

M. Cencov : Evaluation of an unknown distribution density from observations. *Soviet Mathematics*, 3: 1559–1562, 1962.

C. Cervellera et M. Muselli : Deterministic design for neural network learning : An approach based on discrepancy. *IEEE Transactions on Neural Networks*, 15: 533–544, 2004.

J. Chen et J. Glaz : Approximations for a conditional two-dimensional scan statistic. *Statistics and Probability Letters*, 58:287–296, 2002.

H. Chernoff : Locally optimum designs for estimating parameters. *Annals of Mathematical Statistics*, 24: 586–602, 1953.

D. Chessel : La description non paramétrique de la dispersion spatiale des individus d'une espèce. *Biométrie et Écologie*, 1: 45–135, 1978. http ://pbil.univ-lyon1.fr/R/perso/pagechessel.htm.

H. Chi, M. Mascagni et T. Warnock : On the optimal Halton sequence. *Mathematics and computers in simulation*, 70: 9–21, 2005.

A. D. Cliff et J. K. Ord : *Spatial Procsses, Models and Appliations.* Pion, 1981.

Y. Collette et P. Siarry : *Optimisation multiobjectif.* Eyrolles, 2002.

N. Cressie : *Statistics for Spatial Data.* Wiley, 1993.

I. Csiszár : Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der erdodizität von markoffschen ketten. *Publication of Maththematics Institute Hungarian Academy of Sciences*, 8: 85–108, 1963.

I. Csiszár : Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2: 299–318, 1967.

A. de Crécy et P. Bazin : Quantification of the uncertainties of the physical models of CATHARE 2. *In Best Estimates*, p. 111–117. OCDE NRC, 2004.

P. Deheuvels : An asymptotic decomposition for multivariate distribution-free tests of independence. *Journal of Multivariate Analysis*, 11: 102–113, 1981.

P. Deheuvels, J. Einmahl, D. Mason et F. H. Ruymagaart : The almost sure behaviour of maximal and minimal multivariate k_n-spacings. *Journal of Multivariate Analysis*, 24(1): 155–176, 1988.

P. Deheuvels, G. Peccati et M. Yor : On quadratic functionals of the brownian sheet and related processes. *Stochastic Processes and their Applications*, 116: 493–538, 2006.

L. Devroye et L. Györfi : *Nonparametric Density Estimation The L_1 view.* John Wiley & Sons,Inc., 1985.

J. Droesbeke, J. Fine et G. Saporta : *Plans d'expériences, Applications à l'entreprise.* Technip, 1997.

A. Dvoretzky, J. Kiefer et J. Wolfowitz : Asymptotic minimax character of the sample distribution function and of a classical multinomial estimator. *Annals of Mathematical Statistics*, 33: 642–669, 1956.

R. Eubank et P. Speckman : Curve fitting by polynomak-trigonometrics regression. *Biometrika*, 77: 1–9, 1990.

H. Faure : Discrépance de suites associées à un système de numération (en dimension s). *Acta Arithmetica*, 41(4): 337–351, 1982.

V. Fedorov : *Theory of Optimal Experiments*. Academic Press, New York, 1972.

V. Fedorov : Convex design theory. *Math. Operationsforsch. Statist., Ser. Statistics*, 11(3): 403–411, 1980.

V. Feuillard : Liens entre discrépance et estimation non-paramétrique, méthodologie de sélection de points selon les données disponibles. *Revue de Statistiques appliquées*, 2006. soumis.

V. Feuillard, N. Devictor et R. Phan-Tan-Luu : Liens entre discrépance et estimation non-paramétrique, méthodologie de sélection de points selon les données disponibles. $38^{\text{ème}}$ journées de Statistiques, SFDS, 2006.

V. Feuillard, R. Phan-Tan-Luu et N. Devictor : Methodology for evaluating the quality on an input database. *Chemiometrics and Intelligent Laboratory Systems*, 2005. soumis.

A. Földes et P. Révész : A general method for density estimation. *Studia Scientiarium Mathematicaum Hugarica*, 9: 81–92, 1974.

J. Franco, L. Carraro et O. Roustant : Un nouveau critère statistique en forme de radar pour la mesure de l'uniformité des plans d'expériences et de leurs projection sur les sous espaces. $38^{\text{ème}}$ journées de Statistiques, SFDS, 2006.

J. Gauchi et A. Pázman : Designs in nonlinear regression by stochastic minimization of functionals of the mean square error matrix. *Annals of Mathematical Statistics*, 136: 1135–1152, 2006.

J. Glaz, J. I. Naus et S. Wallenstein : *Scan Statistics*. Springer, 2001.

J. Glaz et Z. Zhang : Maximum scan score-type statistics. *Statistics and Probability Letters*, 76(13):1316–1322, 2006.

W. Greblicki et M. Pawlak : Asymptotical efficiency of classifying procedures using the Hermite series estimate of probability density functions. *IEEE Transactions on Systems, Mann and Cybernetics*, SMC-11: 364–366, 1981.

Greig et Smith : The use of random and contiguous quadrats in the study of the structue of plant. *Anals of Botany*, 16: 293–316, 1952.

M. Gunzburger et J. Burkardt : Uniformity measures for point samples in hypercubes. Rap. tech., 2004. http ://www.csit.fsu.edu/ burkardt/pdf/ptmeas.pdf.

I. Györfi et L. Vajda : Asymptotic distribution for goodness-of-fit statistics in a sequence of multinomial models. *Statistics and Probability Letters*, 56: 57–67, 2002.

J. Halton : On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik*, 33(2): 84–90, 1960.

J. Hammersley : Monte Carlo methods for solving multivariable problems. *Annals of the New York Academy of Science*, 86: 844–874, 1960.

W. Hardle : *Applied nonparametric regression.* Cambridge University Press, 1989.

S. Henrich, G.W.Wasilkowski et H. Wozniakowski : The inverse of star-dicrepancy depends lineary on the dimension. *Acta Arithmetica*, 96: 279–302, 2001.

J. M. Hervouet, J. M. Martinez, G. Arnaud et M. Dumas : Parameter estimation in shallow water equation based on surrogate neural network models and genetic algorithms. Proceedings of 7^{th} International Conference of Hydroinformatics HIC, 2006. Nice.

F. Hickernell : The mean square discrepancy of randomized nets. *Modeling and Computer Simulation*, 6(4): 274–296, 1996a.

F. Hickernell : Quadrature error bounds with applications to lattice rules. *SIAM Journal on Numerical Analysis*, 33(5): 1995–2016, 1996b.

F. Hickernell : A generalized discrepancy and quadrature error bound. *Mathematics of computation*, 67(221): 299–322, 1998.

F. Hickernell : Goodness-of-fit statistics, discrepancies and robust designs. *Statistics and Probability Letters*, 44: 73–78, 1999.

E. Hlwaka : Funktionen von beschränter variation in der theorie der gleichverteilung. *Annali di Matematica Pura ed Applicata*, 54: 325–333, 1961.

L. Holst : Asymptotic normality and efficiency for certain goodness-of-fit tests. *Biometrika*, 59: 137–145, 1972.

T. Inglot, T. Jurlewicz et T. Ledwina : Asymptotics for multinomial goodness of fit tests for a simple hypothesis. *Theory of Probability and its Applications*, 35(4):771–777, 1990.

D. Jones, M. Schonlau et W. Welch : Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, p. 455–492, 1998.

M. Kennedy et A. O'Hagan : A bayesian calibration of computer models. *Journal of the Royal Statistical Society*, 63: 425–464, 2001a.

M. Kennedy et A. O'Hagan : Supplementary details on bayesian calibration of computer. Rap. tech., University of Nottingham, Statistics Section, 2001b.

J. Kiefer : Optimum experimental designs. *Journal of the Royal Statistical Society B.*, 21: 272–319, 1959.

J. Kiefer : Optimum designs in regression problems ii. *Annals of Mathematical Statistics*, 31: 298–325, 1961.

J. Kiefer : General equivalence theory for optimum designs (approximate theory). *Annals of statistics*, 2: 849–879, 1974.

R. Kronmal et M. Tarter : The estimation of probability densities and cumulatives by fourier series method. *Journal of the American Statistical Association*, 63: 925–952, 1968.

S. Lallich, E. Prudhommes et O. Teytaud : Contrôle du risque multiple pour la sélection de règles d'assiociation significatives. *Revue Nouvelles Technologies de l'Information, RNTI-E-2*, 2: 305–316, 2004.

P. L'Ecuyer : Polynomial integration lattices, 2004. http ://www.iro.umontreal.ca/ lecuyer/papers.html.

P. L'Ecuyer, R. Simard et S. Wegenkittl : Sparse serial tests of uniformity for random number generators. *SIAM J. Sci. Comput.*, 24(2):652–668, 2002.

C. Lemieux et P. L'Ecuyer : On selection criteria for lattice rules and other Quasi-Monte carlo point sets. *Mathematics and Computers in Simulation*, 55:'139–148, 2001.

C. Loader : Large-deviation approximations to the distribution of scan statistics. *Advances in Applied Probability*, 23:751–771, 1991.

A. Marrel, B. Iooss, F. V. Dorpe et E. Volkova : An efficient methodology for modeling complex computer codes with gaussian processes. *Computational Statistics and Data Analysis*, 2006. soumis.

J. Marry : *Étude de l'Apprentissage Actif : Application à la conduite d'expérience.* Thèse de doctorat, Université Paris Sud, 2005.

H. Martens et M. Martens : *Multivariate Analysis of Quality.* Wiley, 2001.

H. Martens et T. Naes : *Multivariate Calibration.* Wiley, 1991.

P. Massart : The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Annals of Probability*, 18(3): 1269–1283, 1990.

A. Monfort : *Cours de Statistique mathématique.* Economica, 1997.

W. J. Morokoff et R. E. Caflisch : Quasi-random sequences and their discrepancies. *SIAM Journal on Scientific Computing*, 15(6): 1251–1279, 1994.

E. Nadaraya : *Nonparametric estimation of probability densities and regression curves.* Mathematics and its applications. Soviet series. Kluwer academic Publisher, 1989.

T. Naes, T. Isaksson, T. Fearn et T. Davies : *Multivariate Calibration and classification.* NIR, 2002.

J. Naus : A power comparaison of two tests of non random clustering. *Technometrics*, 8:493–517, 1965.

J. Neyman et E. S. Pearson : On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, Part I. 20A:1–66, 1928.

J. Neyman et E. S. Pearson : On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, Ser. A (1933) (231):289–337, 1933.

H. Niederreiter : Discrepancy and convex programming. *Annali di Matematica Pura ed Applicata*, IV(93): 89–97, 1972.

H. Niederreiter : Low-discrepancy and low-dispersion sequences. *Journal of Number Theory*, 30: 51–70, 1988.

H. Niederreiter : *Random number generation and quasi-Monte Carlo methods.* Society for Industrial and Applied Mathematics, 1992.

H. Niederreiter et J. Spanier (Eds) : *Monte Carlo and Quasi-Monte Carlo Methods.* Springer, 1998.

H. Niederreiter et J. Wills : Diskrepanz und distanz von maben bezüglich knovexer und jordanscher mengen. *Mathematische Zeitschrift*, 144: 125–134, 1975.

H. Niederreiter et J. Wills : Constructions of (t, m, s)-nets and (t, s)-sequences. *Finite Fields and their Applications*, 11: 578–600, 2005.

A. O'Hagan : A Markov property for covariance structures. Rap. tech., University of Nottingham, Statistics Section, 1998.

M. C. Pardo : On Burbea-Rao divergence based goodness-of-fit tests for multinomial models. *Journal of Multivariate Analysis*, 69(1): 65–87, 1999.

K. Pearson : On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Philosophical Magazine*, 50: 157–175, 1900.

N. Pérot : Présentation de la méthodologie envisagée pour la calibration du code PACTOLE. Rap. tech. CEA/DEN/CAD/DER/SESI/LCFR/NT DO 25 2005., C.E.A, 2005.

B. L. S. Prakasa Rao : *Nonparametric functional estimation*. Probability and Mathematical Statistics. Academic Press Inc.[Harcourt Brace Jovanovich Publishers], 1983.

L. Pronzato : *Synthèse d'expériences robustes pour modèles à paramètres incertains*. Thèse de doctorat, Thèse de l'Université d'Orsay, 1986.

E. Rafajlowicz et R. Schwabe : Halton and Hammersley sequences in multivariate nonparametric regression. *Statistics and probability letters*, Uncorrected proof, 2005.

M. Ratto, S. Tarantola et A. Saltelli : Sensitivity analysis in model calibration : Gsa-glue approach. *Computer Physics Communications*, 136: 212–224, 2001.

T. Read et N. Cressie : *Goodness-of-fit statistics for discrete multivariate data*. Springer-Verlag, 1988.

F. Riesz et B. S. Nagy : *Functional Analysis*. Ungar Publishing Co., 1955.

A. Rogers : *Statsistical Analysis of Spatial Dispersion : A Quadrat Method*. Pion, 1974.

K. B. R. Romanowicz : The future of distributed models : model calibration and uncertainty prediction. *Reliability Engineering and System Safety*, 91: 1315–1321, 2006.

M. Ross : Coumpound poisson approximations for the numbers of extreme spacings. *Advances in Applied Probability*, 25:847–874, 1993.

M. Ross : Stein's method for compound poisson approximation : the local approach. *Annals of Applied Probability*, 4:1177–1187, 1994.

K. F. Roth : On irregularities of distribution. *Mathematika*, 1: 73–79, 1954.

M. Sancandi : Utilisation des codes de calcul en présence d'incertitudes, application à Mirò. Rap. tech., Commissariat à l'Énergie Atomique/Direction des Applications Militaires, 2006.

G. Sansone : *Orthogonal Functions*. Krieger, 1977.

T. Santner, B. Williams et W. Notz : *The Design and Analysis of Computer Experiments*. Springer, 2003.

S. Schwartz : Estimation of a probability density by an orthogonal series. *Annal of Mathematics Statistics*, 38: 1262–1265, 1967.

H. Sterbuchner : On nonparametric multivariate density estimation. *Revue roumaine de Mathématiques Pures et Appliquées*, 25: 111–118, 1980.

C. Stone : Optimal global rate of convergence for nonparametric regression. *Annals of Statistics*, 10: 1040–1053, 1982.

A. Sukharev : Optimal strategies of the search for an extremum. *U.S.S.R. Computational Mathematics and Mathematical Physics*, 11: 910–924, 1971.

R. Sundberg : The precision of the estimated generalized least square estimator in multivariate calibration. *Scandinavian Journal of Statistics*, 23: 257–274, 1996.

R. Sundberg : Multivariate calibration direct and indirect regression methodology. *Scandinavian Journal of Statistics*, 26: 161–207, 1999.

G. Szegö : *Orthogonal Polynomials*. American Mathematical Society Colloquia Publications, 1975.

M. Tenenhaus : *La régression PLS*. Technip, 1998.

E. Thiémard : *Sur le calcul et la majoration de la discrépance à l'origine*. Thèse de doctorat, Ecole Polytechnique Fédérale de Lausanne, 2000.

I. Tu : *Theory and applications of scan statistics*. Thèse de doctorat, Stanford University, 1997.

S. Tumanyan : On the asymptotic distribution of the chi-square criterion. *Dokl. Akad. Nauk. SSSR*, 94:1011–1012, 1954.

S. Tumanyan : On the asymptotic distribution of the chi-square criterion. *Annals of Statistics*, 3(1):165–188, 1975.

B. Vandewoestyne et R. Cools : Good permutations for deterministic scrambled Halton sequences in terms of l_2 -discrepancy. *Journal of computational and applied mathematics*, 189:341–361, 2004.

E. Vazquez : *Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et applications*. Thèse de doctorat, Thèse de l'Université Paris XI, 2005.

E. Walter et L. Pronzato : *Identification de Modèles Paramétriques à partir de données expérimentales*. Masson, 1994.

E. Walter et L. Pronzato : *Identification of Parametric Models from Experimental Data*. Springer-Verlag, 1997.

T. Warnock : Computational investigations of low discrepancy points sets. *Application of Number Theory to Numerical Analysis*, p. 319–343, 1972.

G. Watson : Density estimation by orthogonal series. *Annal of Mathematical Statistics*, 40: 1496–1498, 1969.

S. Wold, H. Martens et H. Wold : The multivariate calibration problem in chemistry solved by the pls method. In Proc. Conf. Matrix Pencils, Ruhe A. & Kåstrøm B. (Eds), Lecture Notes in Mathematics, 1982.

H. Wynn : The sequential generation of d-optimum experimental designs. *Annals of Mathematical Statistics*, 41: 1655–1664, 1970.

Y. Zhu : A method for exact calculation of the discrepancy of low-dimensional finite point sets (ii). *Acta Mathematica Sinica*, 11(4): 422–435, 1993.

Résumé :

Cette recherche s'insère dans le contexte général de la calibration, en vue d'applications industrielles. Son objectif est d'évaluer la qualité d'une base de données, représentant la manière dont celle-ci occupe, au mieux des objectifs recherchés, son domaine de variation. Le travail réalisé ici fournit une synthèse des outils mathématiques et algorithmiques permettant de réaliser une telle opération. Nous proposons en outre des techniques de sélection ou d'importation de nouvelles observations permettant d'améliorer la qualité globale des bases de données. Les méthodes élaborées permettent entre autres d'identifier des défauts dans la structure des données. Leurs applications sont illustrées dans le cadre de l'évaluation de paramètres fonctionnels, dans un contexte d'estimation par fonctions orthogonales.

Mots clés : calibration, « space filling design », discrépance, estimation fonctionnelle, test d'uniformité.

Abstract :

This thesis takes place in the general context of the calibration for industrial application. It aims at evaluating the quality of a data base by checking that the data, with respect to our objectives, "best fill" the space. This work provides a synthesis of algorithmic and mathematic tools to achieve such a purpose. Extraction and importation techniques to improve the global quality of the data are proposed. These methods allow identifying some defaults of the data structure. An illustration of its application is exposed in the context of functional estimation with orthogonal functions.

Keywords : calibration, space filling design, discrepancy, functional estimation, uniformity test.