Contributions à l'analyse de sensibilité et à l'analyse discriminante généralisée

Julien Jacques

THÈSE

présentée par

Julien JACQUES

pour obtenir le grade de **DOCTEUR**
de l'Université Joseph Fourier - Grenoble 1

Spécialité : **Mathématiques Appliquées**

CONTRIBUTIONS
À
L'ANALYSE DE SENSIBILITÉ
ET À
L'ANALYSE DISCRIMINANTE GÉNÉRALISÉE

Soutenue le 5 décembre 2005

Composition du jury :

Président :	Anatoli IOUDITSKI
Directeur de thèse :	Christian LAVERGNE
Rapporteurs :	Alain BACCINI
	Jérôme SARACCO
Examinateurs :	Christophe BIERNACKI
	Nicolas DEVICTOR

Thèse préparée à l'INRIA Rhône-Alpes, co-financée par le CEA.

à Carole.

Remerciements

J’exprime toute ma reconnaissance à mon directeur de thèse, Christian LAVERGNE, qui a été présent tout au long de ces quatre années. Je le remercie de ses conseils avisés ainsi que pour la rigueur qu’il a su m’inculquer.

Je remercie très chaleureusement Nicolas DEVICTOR, responsable du Laboratoire de Conduite et Fiabilité des Réacteurs du CEA de Cadarache, qui a été à l’initiative de cette thèse, et qui a toujours été disponible pour moi. Je lui suis reconnaissant pour son soutien scientifique et moral.

Je suis reconnaissant à Christophe BIERNACKI, Professeur à l’Université des Sciences et Technologies de Lille, de m’avoir transmis le goût de la recherche lors de mon DEA, et de m’avoir encadré très efficacement lors de ma dernière année de thèse. Je le remercie également de ses conseils quant à l’orientation de ma carrière de chercheur.

Je remercie Anatoli IOUDITSKI, Professeur à l’Université Joseph Fourier de Grenoble, pour avoir accepté de présider mon jury de thèse.

Je suis reconnaissant à Jérôme SARACCO, Professeur à l’Université de Bourgogne, pour l’enthousiasme qu’il a porté à mes travaux, et pour avoir accepté d’être rapporteur de ma thèse. Je le remercie également d’avoir pris le temps de me rencontrer pour me faire part de ses remarques.

J’exprime ma gratitude à Alain BACCINI, Professeur à l’Université Paul Sabatier de Toulouse, pour avoir accepté de rapporter mon travail de thèse, et pour le temps précieux qu’il a su consacrer à cette tache.

Merci également à Anestis ANTONIADIS, Professeur à l’Université Joseph Fourier de Grenoble, qui s’est investi dans ce travail, ainsi qu’à Gérard d’AUBIGNY, Professeur à l’Université Pierre Mendès France de Grenoble, qui m’a accueilli au sein du Laboratoire de Statistique et Analyse des Données.

Je remercie Gilles CELEUX et Florence FORBES pour m’avoir accueilli au sein des projets IS2 et Mistis de l’Inria Rhône-Alpes. J’en profite pour remercier toutes les personnes que j’ai pu cotoyer au sein de l’Inria, qui ont contribué au bon déroulement de ma thèse : Charles (pour son expertise en informatique et sa maîtrise de Linux), Gérard, Greg, Guillaume, Henri, Jean-Baptise, JéjéE et JéjéM, Matthieu, Myriam, Olivier et Stéphane. Je remercie également Chantal, Elodie et Françoise.

Un grand merci également à Franck CORSET pour son aide à différents niveaux lors de mes postes d’ATER au sein du département informatique de l’IUT II de Grenoble, et à Eric FONTENAS pour la confiance qu’il m’a témoigné.

Je remercie Bertrand IOOSS et Michel MARQUÈS, du Laboratoire de Conduite et Fiabilité des Réacteurs du CEA, pour les discussions fructueuses que j'ai pues avoir avec eux lors de mes séjours à Cadarache. Je suis reconnaissant à Bertrand d'avoir relu cette thèse.

Je tiens à remercier Carole LANGLOIS et Arlette DANZON, du Registre des tumeurs du Doubs, pour leur investissement dans notre collaboration.

Je remercie Christian LE MERDY pour m'avoir permis d'effectuer une partie de ma thèse au sein du Laboratoire de Mathématique de l'Université de Franche-Comté, ainsi que Stéphane CHRETIEN pour sa bonne humeur et son expertise scientifique. Merci également à Cédric qui a relu quelques parties de cette thèse, et à Sam, qui est venu me supporter lors de ma soutenance.

J'ai une pensée pour mon père à qui je dois beaucoup, pour ma mère, mes frères et soeur, ainsi que pour ma belle famille.

Enfin, je remercie ma future épouse Carole, qui a toujours cru en moi, qui a été présente tout au long de ma thèse et particulièrement lors des moments difficiles. Je lui dédie cette thèse.

Table des matières

Notations

m, M	modèle mathématique
X, Y	variables aléatoires
x	réalisation de X
$\mathbf{X}, \vec{X}$	vecteur aléatoire
$\mathbf{x}, \vec{x}$	réalisation du vecteur aléatoire $\mathbf{X}, \vec{X}$
X_i	i-ème composante de $\mathbf{X}$
$\mathbf{X}_{\sim i}$	vecteur $\mathbf{X}$ privé de sa i-ème composante
$\hat{T}$	estimation ou approximation de la quantité T
$\mathrm{E}[X]$	espérance de X
$\mathrm{V}(X)$	variance de X
$\mathrm{Cov}(X, Y)$	covariance de X et Y
$\rho_{X,Y}$	coefficient de corrélation entre X et Y
$p(A)$	probabilité de l'événement A
S_i	indice de sensibilité à la variable X_i
$S_{i_1,\ldots,i_s}$	indice de sensibilité à l'interaction entre les variables $X_{i_1}, \ldots, X_{i_s}$
S_{T_i}	indice de sensibilité total à la variable X_i
$\sum_{k \# j} S_{i_1,\ldots,i_s}$	somme sur tous les ensembles tel que $j \in \{i_1, \ldots, i_s\}$
$\mathcal{B}(\alpha)$	loi de Bernoulli de paramètre α
$\mathcal{N}(\mu, \Sigma)$	loi normale de moyenne μ et de matrice de variance Σ
$\mathcal{M}(1, p_1, \ldots, p_K)$	loi multinomiale d'ordre 1 et de paramètres $p_1, \ldots, p_K$
$L(.)$	vraisemblance
$l(.)$	log-vraisemblance
$l_c(., \mathbf{z})$	log-vraisemblance complétée par les données manquantes $\mathbf{z}$

Introduction

Deux sujets sont abordés dans cette thèse, l'analyse de sensibilité et l'analyse discriminante généralisée. Ces deux thèmes étant totalement distincts, nous les présentons séparément.

Analyse de sensibilité

Contexte

Dans de nombreux domaines comme la fiabilité des structures mécaniques, le comportement de systèmes thermohydrauliques ou l'analyse d'accidents graves, des modèles mathématiques sont utilisés, à la fois dans un but de simulation, lorsque les expérimentations sont trop chères ou même impossibles, mais aussi de prédiction. Ces modèles, destinés à mimer un certain phénomène étudié, font correspondre à un ensemble de données d'entrée (variables physiques, paramètres du modèle) un ensemble de sorties ou réponses, et ce de manière déterministe. Dans la réalité, certaines entrées du modèle ne sont pas connues de façon exacte, et sont entachées d'une incertitude. Une démarche probabiliste est alors adoptée, en considérant ces entrées comme des variables aléatoires. Les sorties sont alors elles aussi des variables aléatoires.

Un modèle mathématique est donc une représentation du phénomène étudié, dont la qualité dépend essentiellement de la connaissance de ce phénomène et des moyens dont on dispose pour construire le modèle. La connaissance étant souvent imparfaite (limitation de la compréhension du phénomène, du nombre de données et des expériences) et les moyens limités (scientifiques et numériques), à la majorité des modèles mathématiques sont associés différentes sources d'incertitudes.

Le contexte dans lequel nous nous intéressons à cette incertitude, est celui de l'analyse de sensibilité, qui consiste à déterminer, quantifier et analyser comment réagissent les sorties d'un modèle à des perturbations sur ses variables d'entrée. L'analyse de sensibilité informe sur la façon dont se répercutent les incertitudes d'entrée sur les variables de sortie. Comme ces informations sont utilisées pour prendre des décisions sur le phénomène étudié, il est important d'avoir à l'esprit que des incertitudes sont associées au modèle utilisé.

Problématique

L'objectif de cette thèse est d'étudier l'influence d'une incertitude de modèle sur une analyse de sensibilité, et de voir s'il est possible de prendre en compte cette incertitude lors de l'analyse, ou encore lors de l'interprétation des résultats. Comme nous le verrons à travers l'étude de la bibliographie, la notion d'incertitude de modèle est très large, et nous décidons de n'en étudier que deux aspects, nés de situations concrètes rencontrées au sein du Laboratoire de Conduite et Fiabilité des Réacteurs du CEA[1]/Cadarache. Ces deux aspects ou sources d'incertitudes sont :

[1]Commissariat à l'Energie Atomique

- l'incertitude due à un changement du modèle étudié,
- l'incertitude due à l'utilisation d'un modèle simplifié à la place du modèle étudié.

Détaillons ces deux sources d'incertitude, qui constituent les deux motivations applicatives de ce travail. Considérons un modèle mathématique de référence, M_1, dont les réponses sont réduites à une unique variable de sortie Y.
Nous supposons qu'une analyse de sensibilité a été réalisée sur ce modèle M_1. De nouvelles informations (nouvelles données ou informations complémentaires) nous conduisent à modifier le modèle M_1, c'est-à-dire à le muter en un nouveau modèle M_2. La variable de sortie Y n'aura pas la même valeur qu'elle soit calculée à partir de M_1 ou à partir de M_2. Etant donnée la nature de la mutation, est-il nécessaire de réaliser une nouvelle analyse de sensibilité sur le modèle M_2, souvent chère numériquement, ou alors l'analyse réalisée sur M_1 peut-elle être utilisée pour en déduire à moindre coût des résultats sur celle de M_2 ?
Nous supposons maintenant que le modèle M_1 nécessite un temps important pour calculer la valeur de la sortie correspondant à un jeu de réalisations des variables d'entrée (calcul 3D par exemple, qui peut prendre plusieurs heures de calcul). L'analyse de sensibilité est alors trop coûteuse pour pouvoir être réalisée sur M_1. L'alternative utilisée est de construire un modèle simplifié M_2, qui permet de réaliser le nombre de simulations nécessaires. Il est donc possible de faire l'analyse de sensibilité sur M_2 et de l'utiliser pour approximer celle de M_1. Le second problème auquel tente de répondre cette thèse est de déterminer l'impact de l'erreur d'approximation sur l'analyse de sensibilité de M_1.
Tandis que la problématique de l'impact d'une mutation de modèle sur une analyse de sensibilité n'est pas abordée dans la bibliographie, celle de l'utilisation d'un modèle simplifié l'est, mais dans un contexte de calculs fiabilistes.

La réalisation d'analyse de sensibilité sur des applications réelles, ainsi que la définition des possibilités de mutations d'un modèle, ont fait émerger un problème d'actualité en analyse de sensibilité. Ce problème est celui des modèles à entrées non indépendantes. Nous montrons dans cette thèse que l'utilisation d'une analyse de sensibilité classique en présence de dépendance entre les variables aléatoires d'entrée peut engendrer des erreurs d'interprétation, et proposons alors une méthode basée sur une approche multidimensionnelle.

Organisation de la thèse

La première partie de cette thèse, dédiée à l'analyse de sensibilité, est divisée en trois chapitres.

Le premier chapitre présente un état de l'art de l'analyse de sensibilité, en introduisant les différents indices de sensibilité, ainsi que leur méthode d'estimation. Une étude comparative est réalisée pour déterminer les points forts et points faibles de chaque méthode. Ensuite, nous discutons la notion d'incertitude de modèle, en présentant la bibliographie correspondante.

Le deuxième chapitre apporte des solutions à la double problématique de la thèse, à partir d'une étude théorique de différentes mutations de modèle envisageables. Pour chacune de ces mutations, la possibilité de relier formellement les indices de sensibilité du modèle avant mutation et après mutation est étudiée. Cette relation formelle est utilisée pour déduire les indices de sensibilité du modèle muté à moindre coût. La problématique de la simplification de modèle est alors abordée en caractérisant la simplification comme une mutation du modèle initial. Ces travaux sont illustrés par une application sur un logiciel utilisé en sûreté nucléaire pour des études d'impact sur l'environnement.

Le troisième chapitre apporte une solution au problème de l'analyse de sensibilité pour modèles à entrées non indépendantes. En constatant qu'il n'y a pas de sens à exprimer la sensibilité d'un modèle à une variable d'entrée si celle-ci est corrélée avec d'autres, nous adoptons une démarche multidimensionnelle consistant à exprimer la sensibilité du modèle à des groupes de variables corrélées. La méthode proposée est alors appliquée sur deux codes de calcul utilisés en dosimétrie dans le domaine de l'ingénierie nucléaire.

Analyse discriminante généralisée

Contexte

L'analyse discriminante, aussi appelée classification supervisée ou apprentissage statistique, consiste à classer des individus en groupes définis *a priori*, et ce en utilisant l'information contenue dans un échantillon d'apprentissage, pour lequel les appartenances des individus aux groupes sont connues.

L'analyse discriminante est utilisée dans de nombreux domaines, comme la biologie où l'on cherche par exemple à discriminer une population d'animaux en fonction de leur sexe, ou encore en médecine où l'on tente de différencier les patients malades des patients sains. D'autres applications existent aussi dans le domaine de la finance où l'analyse discriminante est appelée plus communément *scoring*, et où l'objectif consiste à discriminer des clients en fonction de leur rentabilité financière.

Problématique

Afin de pouvoir utiliser l'information contenue dans un échantillon d'apprentissage pour classer des individus issus d'un échantillon test, l'hypothèse fondamentale qui est faite en analyse discriminante est que les deux échantillons d'apprentissage et de test sont issus d'une même population.

Dans un certain nombre d'applications, cette hypothèse n'est pas respectée. L'analyse discriminante généralisée a donc été introduite pour discriminer les individus d'un échantillon issu d'une population « légérement » différente de celle dont est issu l'échantillon d'apprentissage. L'application qui a motivé le développement de la discrimination généralisée consiste à discriminer des oiseaux en fonction de leur sexe, en utilisant un échantillon d'apprentissage constitué d'oiseaux de la même espèce mais d'origine géographique différente.

L'objectif de cette thèse est d'étendre l'analyse discriminante généralisée, définie dans un cadre gaussien, au cas des données binaires. Le principal enjeu est alors de définir un lien entre les populations d'apprentissage et de test, ou autrement dit une relation stochastique qui transforme les variables binaires de la population d'apprentissage en celles de la population test. Pour ce faire, nous supposons que les variables binaires sont issues de la discrétisation de variables continues sous-jacentes gaussiennes. En utilisant alors les travaux précurseurs sur la discrimination généralisée dans un cadre gaussien sur ces variables sous-jacentes, il est possible d'obtenir une relation entre les variables binaires de ces deux populations.

Organisation de la thèse

La seconde partie de cette thèse, constituée d'un unique chapitre, est consacrée à l'analyse discriminante généralisée, et plus particulièrement au cas des données binaires. En définissant un lien entre les populations binaires d'apprentissage et de test, nous développons un certain nombre de modèles de discrimination généralisée, correspondant à des hypothèses naturelles sur la transformation entre les deux populations d'apprentissage et de test. Après avoir défini des techniques d'estimation, basées sur la maximisation de la vraisemblance, et testé ces modèles sur simulations, nous les utilisons pour discriminer des oiseaux en fonction de leur sexe, puis pour prédire parmi une population d'individus ayant eu un premier cancer, les individus ayant un risque important de survenue d'un deuxième cancer.

Première partie.
Analyse de sensibilité et incertitude de modèle

Chapitre 1

État de l'art

Ce premier chapitre contient un état de l'art de l'analyse de sensibilité, ainsi qu'une discussion sur la notion d'incertitude de modèle. Dans la première section, nous présentons les différents indices de sensibilité existants ainsi que les méthodes d'estimation correspondantes, et réalisons une étude comparative de ces différentes approches. Nous abordons ensuite la notion d'incertitude de modèle dans la seconde section en nous appuyant sur une étude bibliographique.

1.1. Analyse de sensibilité

Considérons un modèle mathématique, formé d'un ensemble de variables d'entrée aléatoires, d'une fonction déterministe, et d'un ensemble de variables de sortie (ou réponses) aléatoires.
Nous écrivons ce modèle sous la forme suivante :

$$\begin{array}{rcl} f : \mathbb{R}^p & \rightarrow & \mathbb{R} \\ \mathbf{X} & \mapsto & Y = f(\mathbf{X}) \end{array} \tag{1.1}$$

La fonction f du modèle peut être très complexe (système d'équations différentielles), et est en pratique évaluée à l'aide d'un code informatique, plus ou moins onéreux en temps de calcul. L'ensemble des variables d'entrée $\mathbf{X} = (X_1, ..., X_p)$ regroupe toutes les entités considérées comme aléatoires dans le modèle. Nous supposons dans ce premier chapitre les variables d'entrée indépendantes. L'ensemble des variables de sortie est quant à lui supposé réduit à une unique variable Y.
L'analyse de sensibilité étudie comment des perturbations sur les entrées du modèle engendrent des perturbations sur la réponse. Il est possible de grouper les méthodes d'analyse de sensibilité en trois classes : les méthodes de *screening*, l'analyse de sensibilité locale et l'analyse de sensibilité globale.
Les méthodes de *screening*, présentées par Saltelli et al. dans [48], analysent qualitativement l'importance des variables d'entrée sur la variabilité de la réponse du modèle. Elles permettent d'établir une hiérarchie au sein des variables d'entrée en fonction de leur influence sur la variabilité de la réponse. L'analyse de sensibilité locale, tout comme l'analyse globale, sont des méthodes d'analyse quantitative, qui permettent en plus d'établir une hiérarchie au sein des variables d'entrée, de donner un ordre de grandeur des écarts au sein de cette hiérarchie. L'analyse de sensibilité locale étudie comment de petites perturbations autour d'une valeur $\mathbf{x}^0 = (x_1^0, ..., x_p^0)$ des entrées se répercutent sur la valeur de la sortie. La méthode d'analyse locale la plus classique est l'approche OAT (*One factor At Time*), qui consiste à calculer ou estimer les indices de sensibilité définis par

$$S_i = \frac{\partial y}{\partial x_i}(x_1^0, ..., x_p^0),$$

exprimant l'effet sur la valeur de la variable aléatoire Y de perturber les valeurs des variables X_i autour d'une valeur nominale x_i^0. Pour une revue de ces méthodes, le lecteur pourra se référer à Turanyi [63].
L'analyse de sensibilité globale s'intéresse quant à elle à la variabilité de la sortie du modèle dans son domaine de variation. Elle étudie comment la variabilité des entrées se répercute sur celle de la sortie, en déterminant quelle part de variance de la sortie est due à telles entrées ou tel ensemble d'entrées. Il est possible de distinguer l'analyse locale de l'analyse globale de la sorte : l'analyse locale s'intéresse à la valeur de la réponse, tandis que l'analyse globale s'intéresse à sa variabilité.
Cette thèse traite exclusivement de l'analyse de sensibilité globale, nous nous permettons donc parfois d'omettre cet adjectif «globale».

1.1.1. Les ambitions de l'analyse de sensibilité

Au cours de l'élaboration, de la construction ou de l'utilisation d'un modèle mathématique, l'analyse de sensibilité peut s'avérer être un outil précieux. En effet, en étudiant comment la réponse du modèle réagit aux variations de ses variables d'entrée, l'analyse de sensibilité permet de déterminer :

(i) Si le modèle est bien fidèle au processus qu'il modélise. En effet, si l'analyse exhibe une importance forte d'une variable d'entrée qui en réalité est connue comme non influente, le modèle ne reflétera pas correctement le processus. Il sera alors nécessaire de modifier le modèle.

(ii) Quelles sont les variables qui contribuent le plus à la variabilité de la réponse du modèle ? Il sera alors possible, si besoin est, d'améliorer la qualité de la réponse du modèle. Connaissant les variables d'entrée les plus influentes, les erreurs sur la sortie du modèle pourront être diminuées, soit, lorsque cela est possible, en diminuant les erreurs sur les entrées les plus influentes, soit en adaptant la structure du modèle pour réduire l'effet des erreurs sur ces entrées.

(iii) Quelles sont au contraire les variables les moins influentes. Il sera possible de les considérer comme des paramètres déterministes, en les fixant par exemple à leur espérance, et ainsi d'obtenir un modèle plus *léger* avec moins de variables d'entrée. Dans le cas d'un code informatique, il sera possible de supprimer des parties de codes qui n'ont aucune influence sur la valeur et la variabilité de la réponse.

(iv) Quelles variables, ou quels groupes de variables, interagissent avec quelles (ou quels) autres : l'analyse de sensibilité peut permettre de mieux appréhender et comprendre le phénomène modélisé, en éclairant les relations entre les variables d'entrée et la variable de sortie.

Bon nombre de publications sur le sujet explicitent et illustrent ces objectifs. On pourra se référer notamment aux travaux de Saltelli et al. : [49],[50] et [52].
Les enjeux de l'analyse de sensibilité sont alors multiples : validation d'une méthode ou d'un code de calcul, orientation des efforts de recherche et développement, ou encore justification en terme de sûreté d'un dimensionnement ou d'une modification d'un système. De nombreux domaines d'applications sont alors interessés. Saltelli dans [48] en regroupe un certain nombre : l'ingénierie nucléaire, la chimie, l'écologie. On rencontre aussi d'autres applications en médecine [8] ou en économie.

1.1.2. Estimateurs statistiques de la sensibilité

L'objectif de l'analyse de sensibilité globale est de répartir l'incertitude sur la réponse d'un modèle entre les variables d'entrée. Elle consiste à déterminer quelle part d'incertitude sur la réponse est due à l'incertitude sur chaque variable d'entrée (ou groupe de variables d'entrée). Saltelli présente l'ensemble des méthodes existantes dans [48], en les regroupant de la sorte : les méthodes fiabilistes de type FORM et SORM traitant d'analyse de sensibilité pour l'analyse de risques, les méthodes bayesienne, les méthodes graphiques et enfin les méthodes basées sur l'étude de la variance. Nous nous intéressons dans ce mémoire exclusivement à cette dernière classe de méthode, qui consiste à déterminer quelle part de variance de la réponse est due à la variance de chaque variable d'entrée (ou groupe de variables d'entrée).
Nous présentons dans un premier temps les méthodes d'analyse de sensibilité globale (basées sur la variance)

pour modèles linéaires, puis monotones, et enfin une méthode plus générale, basée sur la décomposition de la variance du modèle. Chaque méthode définit des indices de sensibilité, exprimant la sensibilité du modèle aux variables ou groupes de variables d'entrée. Nous analysons ensuite différentes méthodes d'estimation de ces indices, que nous comparons sur un exemple test.

1.1.2.1. Indices de sensibilité pour modèles linéaires et/ou monotones

La première hypothèse faite sur le modèle, est celle de linéarité, sous laquelle nous définissons des indicateurs intuitifs de la sensibilité de la sortie du modèle aux variables d'entrée. Puis nous généralisons ces indicateurs au cas moins restrictif des modèles monotones.

Indice SRC et coefficient de corrélation

Supposons que le modèle étudié soit linéaire, et qu'il s'écrive sous la forme suivante :

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i. \tag{1.2}$$

Comme les variables X_i sont supposées indépendantes, la variance de Y s'écrit alors :

$$\mathrm{V}(Y) = \sum_{i=1}^{p} \beta_i^2 \mathrm{V}(X_i),$$

où $\beta_i^2 \mathrm{V}(X_i)$ est la part de variance due à la variable X_i. Ainsi, nous quantifions la sensibilité de Y à X_i par le rapport de la part de variance due à X_i sur la variance totale.

Définition 1.1.1. *L'indicateur ainsi construit est l'indice de sensibilité SRC (Standardized Regression Coefficient), défini par :*

$$SRC_i = \frac{\beta_i^2 V(X_i)}{V(Y)}. \tag{1.3}$$

Il exprime la part de variance de la réponse Y due à la variance de la variable X_i.
Cet indice SRC est équivalent, au carré près, au coefficient de corrélation linéaire entre la réponse du modèle et ses variables d'entrée $\rho_{X_i,Y}$, qui est parfois utilisé comme indicateur de sensibilité. En effet, pour le modèle linéaire (1.2), il est simple de vérifier que :

$$\mathrm{Cov}(X_i, Y) = \beta_i \mathrm{V}(X_i),$$

d'où

$$\rho_{X_i,Y} = \frac{\mathrm{Cov}(X_i, Y)}{\sqrt{\mathrm{V}(X_i)\mathrm{V}(Y)}} = \beta_i \sqrt{\frac{\mathrm{V}(X_i)}{\mathrm{V}(Y)}}$$

et donc, $\rho^2_{X_i,Y} = SRC_i$.
Contrairement au coefficient de corrélation linéaire, l'indice SRC est toujours positif ($SRC \in [0, 1]$). On le préférera alors au coefficient de corrélation linéaire, puisque l'on raisonne en terme de part de variance, et donc le signe du coefficient de corrélation linéaire n'a pas d'importance.

Indice PCC (*Partial Correlation Coefficient*)

Néanmoins, il est parfois difficile d'apprécier la sensibilité de Y à une variable d'entrée X_i, si les simulations successives du modèle sont faites pour des valeurs différentes de toutes les variables d'entrée. En effet, la corrélation entre Y et X_i peut être due à une tierce variable. On rencontre parfois en pratique des cas où

une corrélation entre deux variables est observée, alors qu'elle n'est en fait due qu'à une corrélation avec une troisième variable. Saporta illustre ceci dans [55], en citant une étude, qui concluait que le nombre de maladies mentales au sein d'une certaine population était corrélé au nombre de postes de radio que possédait cette population. Cette corrélation s'avérait en fait nulle après avoir fixé la variable temps.
Pour contrer cet effet, l'indice de corrélation partielle PCC a été proposé. Il permet d'évaluer la sensibilité de Y à X_i en éliminant l'effet des autres variables, toujours donc sous l'hypothèse de linéarité du modèle.

Définition 1.1.2. *L'indice de corrélation partielle de Y et de X_i, exprimant la sensibilité de Y à X_i, est donné par :*

$$PCC_i = \rho_{Y,X_i|\mathbf{X}_{\sim i}} = \frac{Cov(Y, X_i|\mathbf{X}_{\sim i})}{\sqrt{V(Y|\mathbf{X}_{\sim i})V(X_i|\mathbf{X}_{\sim i})}},$$

où $\mathbf{X}_{\sim i}$ est le vecteur $\mathbf{X}$ privé de sa i-ème composante.

La notion de corrélation partielle est développée dans [55]. Nous la présentons dans le contexte qui nous intéresse, c'est-à-dire pour définir la corrélation partielle de la sortie Y du modèle avec une variable d'entrée X_i, en annexe A.1.
Les deux indices de sensibilité PCC et SRC ne sont pas égaux,. Néanmoins, s'ils sont utilisés pour classer les variables d'entrée en fonction de leur importance sur la variable de sortie, les classements obtenus sont exactement les mêmes, le classement pour l'indice PCC se réalisant à partir de la valeur absolue de ce dernier, puisqu'il peut être négatif.

Indices basés sur la transformation des rangs $SRRC$ et $PRCC$

Les indices de sensibilité SRC et PCC trouvent leur légitimité dans la linéarité du modèle. Si cette hypothèse n'est pas vérifiée, mais si le modèle est monotone[1], il est possible de contourner le problème de la non linéarité, en se basant sur la transformation des rangs. En effet, l'indice SRC, basé sur la linéarité, n'est autre que le carré du coefficient de corrélation. Si le modèle n'est pas linéaire mais monotone, un estimateur non paramétrique du coefficient de corrélation est un coefficient de corrélation basé sur les rangs.
Considérons une matrice de N simulations du modèle étudié $Y = f(X_1, ..., X_p)$:

$$\begin{bmatrix} y^1 & x_1^1 & x_2^1 & \dots & x_p^1 \\ y^2 & x_1^2 & x_2^2 & \dots & x_p^2 \\ & & \vdots & & \\ y^N & x_1^N & x_2^N & \dots & x_p^N \end{bmatrix}.$$

À chaque simulation i de 1 à N, on associe son rang selon une variable. Le rang 1 sera affecté à la simulation qui a la plus petite valeur, et le rang N à celle qui a la plus grande valeur. Ainsi, étudier la liaison entre deux variables reviendra à comparer les classements issus de ces deux variables. Dans notre cas, nous créons le vecteur r_Y des rangs selon Y, et pour chaque variable d'entrée les vecteurs r_i, pour $i = 1, N$, des rangs selon X_i.
Nous définissons alors deux indices, l'indice $SRRC$ (*Standardized Regression Rank Coefficient*), qui est l'analogue de l'indice SRC mais calculé à partir des vecteurs de rangs, et l'indice $PRCC$ (*Partial Rank Correlation Coefficient*) pour l'indice PCC.

1.1.2.2. Indices de sensibilité sans hypothèse sur le modèle

Après avoir présenté des indices de sensibilité pour modèles linéaires, puis monotones, nous présentons une approche plus générale qui ne fait aucune hypothèse sur le modèle.

[1] monotone par rapport à chacune de ses variables lorsque les autres sont fixées

Considérons le modèle

$$Y = f(X_1, ..., X_p), \tag{1.4}$$

où les variables d'entrée sont indépendantes.
Pour apprécier l'importance d'une variable d'entrée X_i sur la variance de la sortie Y, nous étudions de combien la variance de Y décroît si on fixe la variable X_i à une valeur x_i^* :

$$\mathrm{V}(Y|X_i = x_i^*).$$

Cette quantité est la variance conditionnelle de Y sachant $X_i = x_i^*$. Le problème de cet indicateur est le choix de la valeur x_i^* de X_i, que l'on résout en considèrant l'espérance de cette quantité pour toutes les valeurs possibles de x_i^*, notée :

$$\mathrm{E}[\mathrm{V}(Y|X_i)].$$

Ainsi, plus la variable X_i sera importante vis-à-vis de la variance de Y, plus cette quantité sera petite.

Théorème. *Théorème de la variance totale.*
Soit un couple (X, Y) de variables aléatoires, où Y prend ses valeurs dans $\mathbb{R}$ et X_i dans un ensemble fini ou dénombrable, ou dans $\mathbb{R}$ ou $\mathbb{R}^p$. Si la variance de Y est finie, alors :

$$V(Y) = V(E[Y|X_i]) + E[V(Y|X_i)].$$

Etant donné le théorème de la variance totale, un indicateur de la sensibilité de Y à X_i sera la variance de l'espérance de Y conditionnellement à X_i :

$$\mathrm{V}(\mathrm{E}[Y|X_i]).$$

Plus la variable X_i sera importante, plus cette quantité sera grande. Afin d'utiliser un indicateur normalisé, nous définissons finalement l'indice de sensibilité (1.5).

Définition 1.1.3. *L'indice de sensibilité exprimant la sensibilité de Y à X_i est défini par :*

$$S_i = \frac{V(E[Y|X_i])}{V(Y)}. \tag{1.5}$$

Cet indice est appelé indice de sensibilité de premier ordre par Sobol [60], *correlation ratio* par McKay [34], ou encore *importance measure*. Il quantifie la sensibilité de la sortie Y à la variable d'entrée X_i, ou encore la part de variance de Y due à la variable X_i.

Remarque. *Dans le cas du modèle linéaire (1.2), cet indice de sensibilité est égal à l'indice SRC, puisque :*

$$V(E[Y|X_i]) = V(E[\beta_0 + \sum_{i=1}^{p} \beta_i X_i|X_i]) = V(\beta_i X_i) = {\beta_i}^2 V(X_i).$$

Sobol [60] introduit cet indice de sensibilité en décomposant la fonction f du modèle en somme de fonctions de dimensions croissantes. Considérons la fonction du modèle (1.4) : $f(x_1, ..., x_p)$, où nous supposons les variables $(x_1, ..., x_p)$ réelles appartenant $[0, 1]^p$.
Sobol montre la proposition suivante :

Proposition 1.1.1. *Décomposition de Sobol.*
Si f est intégrable sur $[0, 1]^p$, elle admet une unique décomposition :

$$f(x_1, ..., x_p) = f_0 + \sum_{i=1}^{p} f_i(x_i) + \sum_{1 \le i < j \le p} f_{i,j}(x_i, x_j) + \ldots + f_{1,2,...,p}(x_1, ..., x_p), \tag{1.6}$$

où f_0 est une constante et où les fonctions de la décomposition vérifient les conditions :

$$\int_0^1 f_{i_1,...,i_s}(x_{i_1},...,x_{i_s})dx_{i_k} = 0 \qquad \forall k = 1,...,s, \quad \forall\{i_1,...,i_s\} \subseteq \{1,...,p\}. \tag{1.7}$$

Démonstration. L'existence et l'unicité de cette décomposition (1.6) sont assurées par les conditions (1.7). On pourra se référer à [60] pour la démonstration. □

Ces conditions (1.7) signifient que l'intégrale de chaque fonction de la décomposition par rapport à une de ses variables est nulle.
Il vient alors naturellement que les fonctions $f_{i_1,...,i_s}$ sont orthogonales, *i.e.*

$$\int_0^1 f_{i_1,...,i_s}(x_{i_1},...,x_{i_s}) f_{j_1,...,j_t}(x_{j_1},...,x_{j_t})dx = 0, \tag{1.8}$$

si au moins un des indices $i_1,...,i_s,j_1,...,j_t$ n'est pas répété.
En utilisant les conditions (1.7), on obtient en intégrant (1.4) par rapport :

- à toutes les variables :

$$\int_0^1 f(x)dx = f_0,$$

- à toutes les variables sauf x_i :

$$\int_0^1 f(x)dx_{\sim i} = f_0 + f_i(x_i),$$

  où $x_{\sim i}$ représente le vecteur $(x_1,...,x_p)$ privé de sa i-ème composante.
- à toutes les variables sauf x_i et x_j :

$$\int_0^1 f(x)dx_{\sim ij} = f_0 + f_i(x_i) + f_j(x_j) + f_{i,j}(x_i,x_j),$$

- à toutes les variables sauf x_i, x_j et x_k :

$$\begin{aligned}\int_0^1 f(x)dx_{\sim ijk} &= f_0 + f_i(x_i) + f_j(x_j) + f_k(x_k) \\ &\quad + f_{i,j}(x_i,x_j) + f_{i,k}(x_i,x_k) + f_{j,k}(x_j,x_k) + f_{i,j,k}(x_i,x_j,x_k),\end{aligned}$$

  et ainsi de suite.

On en déduit alors les fonctions de la décomposition (1.6) :

$$\begin{aligned}f_0 &= \int_0^1 f(x)dx, \\ f_i(x_i) &= -f_0 + \int_0^1 f(x)dx_{\sim i}, \\ f_{i,j}(x_i,x_j) &= -f_0 - f_i(x_i) - f_j(x_j) + \int_0^1 f(x)dx_{\sim ij}, \\ f_{i,j,k}(x_i,x_j,x_k) &= -f_0 - f_i(x_i) - f_j(x_j) - f_k(x_k) - f_{i,j}(x_i,x_j) - f_{i,k}(x_i,x_k) - f_{j,k}(x_j,x_k) \\ &\quad + \int_0^1 f(x)dx_{\sim ijk},\end{aligned}$$

et ainsi de suite jusqu'au dernier terme :

$$f_{1,...,p}(x_1,...,x_p) = f(x) - f_0 - \sum_{i=1}^{p} f_i(x_i) - \ldots - \sum_{1\leq i_1<\ldots<i_{p-1}\leq p} f_{i_1,...,i_{p-1}}(x_{i_1},...,x_{i_{p-1}}).$$

Ce dernier terme $f_{1,...,p}$ est défini par $f(x)$ moins tous les autres termes, de sorte à ce que la décomposition soit vérifiée.
La décomposition ci-dessus est présentée pour une fonction de variables réelles à valeurs dans $[0,1]^p$. Revenons au modèle (1.4) : $Y = f(X_1,...,X_p)$, où les variables X_i sont aléatoires. Nous les supposons indépendantes et uniformes sur $[0,1]$. La décomposition (1.6) reste vraie :

$$Y = f(X_1,...,X_p) = f_0 + \sum_{i=1}^{p} f_i(X_i) + \sum_{1\leq i<j\leq p} f_{ij}(X_i,X_j) + \ldots + f_{1,...p}(X_1,...,X_p). \tag{1.9}$$

Il est possible d'interpréter les fonctions de la décomposition en terme d'espérance et de variance. En prenant l'espérance de Y suivant son expression (1.9), on obtient :

$$\mathrm{E}[Y] = f_0 + 0,$$

d'après la propriété (1.7). De même, en prenant l'espérance de Y conditionellement à X_i, puis à X_i et X_j, et ainsi de suite, on obtient :

$$\begin{aligned}
f_i(X_i) &= \mathrm{E}[Y|X_i] - f_0, \\
f_{i,j}(X_i,X_j) &= \mathrm{E}[Y|X_i,X_j] - f_i(X_i) - f_j(X_j) - f_0, \\
f_{i,j,k}(X_i,X_j,X_k) &= \mathrm{E}[Y|X_i,X_j,X_k] - f_{i,j}(X_i,X_j) - f_{i,k}(X_i,X_k) - f_{j,k}(X_j,X_k) \\
&\quad - f_i(X_i) - f_j(X_j) - f_k(X_k) - f_0, \\
&\ldots
\end{aligned} \tag{1.10}$$

ou encore :

$$\begin{aligned}
f_0 &= \mathrm{E}[Y], \\
f_i(X_i) &= \mathrm{E}[Y|X_i] - \mathrm{E}[Y], \\
f_{i,j}(X_i,X_j) &= \mathrm{E}[Y|X_i,X_j] - \mathrm{E}[Y|X_i] - \mathrm{E}[Y|X_j] + \mathrm{E}[Y], \\
f_{i,j,k}(X_i,X_j,X_k) &= \mathrm{E}[Y|X_i,X_j,X_k] - \mathrm{E}[Y|X_i,X_j] - \mathrm{E}[Y|X_i,X_k] - \mathrm{E}[Y|X_j,X_k] \\
&\quad + \mathrm{E}[Y|X_i] + \mathrm{E}[Y|X_j] + \mathrm{E}[Y|X_k] - \mathrm{E}[Y], \\
&\ldots
\end{aligned}$$

Remarque. *Cette décomposition est analogue à la décomposition réalisée en analyse de variance, connue sous le nom anglais «ANOVA decomposition», et présentée notamment par Efron et Stein dans [19].*

La variance de Y, V, peut se décomposer selon le théorème suivant.

Théorème 1.1.1. *Décomposition de Sobol de la variance.*
La variance du modèle à entrées indépendantes (1.4) se décompose en :

$$V = \sum_{i=1}^{p} V_i + \sum_{1\leq i<j\leq p} V_{ij} + \ldots + V_{1...p}, \tag{1.11}$$

où

$$\begin{aligned} V_i &= V(E[Y|X_i]), \\ V_{ij} &= V(E[Y|X_i, X_j]) - V_i - V_j, \\ V_{ijk} &= V(E[Y|X_i, X_j, X_k]) - V_{ij} - V_{ik} - V_{jk} - V_i - V_j - V_k, \\ &\dots \\ V_{1\dots p} &= V - \sum_{i=1}^{p} V_i - \sum_{1\le i<j\le p} V_{ij} - \dots - \sum_{1\le i_1<\dots<i_{p-1}\le p} V_{i_1\dots i_{p-1}} \end{aligned}$$

Démonstration. La démonstration est immédiate dès lors que l'on constate que le dernier terme de la décomposition $V_{1\dots p}$ n'est autre que la différence entre la variance de Y et toutes les parts de variance d'ordre inférieur. □

Nous montrons en annexe A.2.1 que les parts de variance de la décomposition (1.11) sont les variances des fonctions de la décomposition de Sobol (1.9), que l'on note $\tilde{V}_{i_1,\dots,i_s}$:

$$V_{i_1,\dots,i_s} = \mathrm{V}(f_{i_1,\dots,i_s}(X_{i_1}, \dots, X_{i_s})) = \tilde{V}_{i_1,\dots,i_s} \qquad \forall \{i_1, \dots, i_s\} \subseteq \{1, \dots, p\}.$$

Ainsi, la séparation des effets des différentes variables d'entrée faite dans la décomposition de Sobol de la fonction du modèle est bien transmise dans la décomposition de la variance de Y.

Remarque. *L'hypothèse d'uniformité ne sert en fait qu'à écrire les égalités (1.10) sous forme d'intégrales. Cette hypothèse n'est pas nécessaire à la décomposition.*

Définition 1.1.4. *Ainsi, on peut définir des indices de sensibilité de premier ordre :*

$$S_i = \frac{V_i}{V} = \frac{V(E[Y|X_i])}{V}, \tag{1.12}$$

les indices de sensibilité d'ordre deux :

$$S_{ij} = \frac{V_{ij}}{V},$$

qui expriment la sensibilité de la variance de Y à l'interaction des variables X_i et X_j, c'est-à-dire la sensibilité de Y aux variables X_i et X_j qui n'est pas prise en compte dans l'effet des variables seules.
On définit encore les indices de sensibilité d'ordre trois :

$$S_{ijk} = \frac{V_{ijk}}{V},$$

qui expriment la sensibilité de la variance de Y aux variables X_i, X_j et X_k qui n'est pas prise en compte dans l'effet des variables seules et des interactions deux à deux.
Et ainsi de suite jusqu'à l'ordre p.

Remarque. *Cette définition des indices de sensibilité de premier ordre obtenue à partir de la décomposition de Sobol est la même que celle donnée précédemment en (1.5).*

L'interprétation de ces indices est facile, puisque grâce à (1.11), leur somme est égale à 1, et étant tous positifs, plus l'indice sera grand (proche de 1), plus la variable aura d'importance.
Le nombre d'indices de sensibilité ainsi construit, de l'ordre 1 à l'ordre p, est égale à $2^p - 1$. Lorsque le nombre de variables d'entrée p est trop important, le nombre d'indices de sensibilité explose. L'estimation

et l'interprétation de tous ces indices deviennent vite impossibles.
Homma et Saltelli [27] ont alors introduit des indices de sensibilité totaux, qui expriment la sensibilité totale de la variance Y à une variable, c'est-à-dire la sensibilité à cette variable sous toutes ses formes (sensibilité à la variable seule et sensibilité aux interactions de cette variable avec d'autres variables).

Définition 1.1.5. *L'indice de sensibilité total S_{T_i} à la variable X_i est défini comme la somme de tous les indices de sensibilité relatifs à la variable X_i :*

$$S_{T_i} = \sum_{k \# i} S_k. \tag{1.13}$$

où $\#i$ représente tous les ensembles d'indices contenant l'indice i.

Par exemple, pour un modèle à trois variables d'entrée, nous avons :

$$S_{T_1} = S_1 + S_{12} + S_{13} + S_{123}.$$

Remarque. *De la même façon que l'on a l'équation suivante en conditionnant par rapport à X_i :*

$$1 = \underbrace{\frac{V(E[Y|X_i])}{V(Y)}}_{S_i} + \frac{E[V(Y|X_i)]}{V(Y)},$$

en conditionnant par rapport à $\mathbf{X}_{\sim i}$, on obtient :

$$1 = \frac{V(E[Y|\mathbf{X}_{\sim i}])}{V(Y)} + \underbrace{\frac{E[V(Y|\mathbf{X}_{\sim i})]}{V(Y)}}_{S_{T_i}}. \tag{1.14}$$

Ces indices de sensibilité ont l'avantage de ne faire aucune hypothèse sur la forme du modèle, mais nécessitent une hypothèse d'indépendance des variables d'entrée. Ces indices de sensibilité basés sur la décomposition de la variance seront ceux utilisés dans la suite de cette thèse.

1.1.3. Estimation des indices de sensibilité

Les indices de sensibilité qui viennent d'être présentés peuvent parfois être calculés formellement, lorsque la forme analytique de la fonction f du modèle est connue et relativement simple. Nous avons émis l'hypothèse que cette fonction pouvait être très complexe et non connue analytiquement (résultat d'un code informatique). Ne pouvant calculer ces indices de sensibilité, il est alors nécessaire de les estimer. Nous présentons dans un premier paragraphe l'estimation des indices de sensibilité SRC, PCC ainsi que leur version basée sur les rangs. Puis, dans un second paragraphe, nous présenterons l'estimation des indices de sensibilité basés sur la décomposition de la variance.

1.1.3.1. Indices SRC, PCC, $SRRC$ et $PRCC$

Les indices de sensibilité SRC et PCC nécessitent la linéarité du modèle :

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i.$$

En pratique, le modèle n'est généralement pas exactement linéaire, les β_i ne sont pas connus et il est nécessaire de les estimer. Une régression linéaire multiple permet alors d'estimer le modèle :

$$\hat{Y} = \hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_i X_i,$$

à partir d'un N-échantillon de simulations du modèle $(y_k, x_{k1}, \ldots, x_{kp})_{k=1..N}$. Nous définissons par simulation du modèle le vecteur $(y, x_1, \ldots, x_p)$, formé par le résultat d'une réalisation aléatoire des p variables d'entrée : $x_1, \ldots, x_p$ et du calcul de la valeur correspondante de la variable de sortie : y. Rappelons que la forme analytique du modèle étudié n'est pas nécessairement connue, mais que nous avons supposé savoir simuler ce modèle par un code informatique.

Indice SRC : les indices SRC sont alors estimés par (1.3) :

$$SRC_i = \frac{\hat{\beta}_i^2 \hat{\mathrm{V}}(X_i)}{\hat{\mathrm{V}}(Y)}.$$

Indice PCC : une méthode de calcul de cet indice PCC est présentée dans [48]. Elle est composée de deux étapes. Pour estimer l'indice PCC_i relatif à la variable X_i, on construit dans un premier temps les deux régressions linéaires multiples suivantes :

$$\hat{Y}^{(\sim i)} = \hat{b}_0 + \sum_{j \neq i} \hat{b}_j X_j \quad \text{et} \quad \hat{X}_i^{(\sim i)} = \hat{c}_0 + \sum_{j \neq i} \hat{c}_j X_j.$$

L'indice PCC_i n'est alors rien d'autre que la valeur absolue du coefficient de corrélation entre $X_i - \hat{X}_i^{(\sim i)}$ et $Y - \hat{Y}^{(\sim i)}$:

$$PCC_i = \frac{\hat{\mathrm{Cov}}(X_i - \hat{X}_i^{(\sim i)}, Y - \hat{Y}^{(\sim i)})}{\sqrt{\hat{\mathrm{V}}(X_i - \hat{X}_i^{(\sim i)}) \hat{\mathrm{V}}(Y - \hat{Y}^{(\sim i)})}}.$$

Indices $SRRC$ et $PRCC$: par définition, ces indices sont les versions basées sur les rangs des indices SRC et PCC. Pour les estimer, il faut transformer les simulations en vecteurs rangs (paragraphe 1.1.2.1), puis estimer l'indice $SRRC$ ($PRCC$) comme l'indice SRC (PCC) en utilisant ces vecteurs rangs.

1.1.3.2. Estimation de Monte Carlo (ou méthode de Sobol)

Rappel sur les méthodes de Monte Carlo [46]

Dans beaucoup de problèmes scientifiques, on est amené à calculer une intégrale du type

$$I = \int_D f(\mathbf{x}) d\mathbf{x}, \tag{1.15}$$

où D est un espace de plus ou moins grande dimension, et f une fonction (intégrable).
Soit $x_1, \ldots, x_N$ un N-échantillon d'une variable aléatoire uniforme sur D. Nous supposons cet échantillon pris de manière totalement aléatoire (échantillonnage aléatoire). Une approximation de I par la méthode de Monte Carlo est faite par :

$$\hat{I}_N = \frac{1}{N} \sum_{i=1}^{N} f(\mathbf{x}_i).$$

La loi forte des grands nombres assure que la moyenne d'une suite de variables aléatoires indépendantes de même espérance et de variances finies converge presque sûrement vers l'espérance, d'où :

$$\lim_{N \to +\infty} \hat{I}_N = I, \quad \text{avec probabilité 1.}$$

Etant données les hypothèses de la loi forte des grands nombres (même espérance et variances finies), toute espérance mathématique d'une variable aléatoire $f(X)$ de densité de probabilité μ :

$$\mathrm{E}[f(X)] = \int f(x)\mu(x)dx,$$

peut être estimée par :

$$\hat{\mathrm{E}}[f(X)] = \frac{1}{N}\sum_{i=1}^{N} f(x_i),$$

où $(x_i)_{i=1..N}$ est un N-échantillon de réalisations de la variable aléatoire X.
Le taux de convergence d'une méthode de Monte Carlo est estimé à l'aide du théorème centrale-limite, qui assure que :

$$\frac{\sqrt{N}}{\sigma}(\hat{I}_N - I) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

si $f(X)$ est de variance finie σ^2. La convergence est donc en $\mathcal{O}(N^{-\frac{1}{2}})$.

Accélération de la convergence

La méthode de Monte Carlo avec échantillonnage aléatoire est la méthode de base pour calculer l'intégrale (1.15). Bon nombre de méthodes alternatives ont été proposées pour améliorer la convergence, parmi lesquelles les méthodes de simulation pseudo-probabilistes[2], comme l'échantillonnage stratifié ou par hyper cube latin (*LHS*) [35], les méthodes de Quasi-Monte Carlo [37], ou encore les méthodes de Quasi-Monte Carlo Randomisé [40].
L'échantillonnage stratifié consiste à découper l'espace des variables d'entrée en petits espaces disjoints, puis à échantillonner au sein de chacun de ces sous espaces. L'échantillonnage *LHS* est basé sur le même principe, en s'assurant que le découpage a défini des espaces équiprobables, et que chaque espace est bien échantillonné ; le quadrillage se fait dans le cube unité, pour un tirage aléatoire d'échantillon uniforme, puis ces échantillons sont transformés via la fonction de répartition inverse.
Les méthodes de Quasi-Monte Carlo sont des versions déterministes des méthodes de Monte Carlo. Ces méthodes définissent des séquences d'échantillons déterministes qui ont une discrépance plus faibles que les séquences aléatoires, c'est-à-dire qu'elles ont une meilleure répartition uniforme dans l'espace des variables d'entrée. Ces méthodes de quasi-Monte Carlo permettent d'obtenir une convergence plus rapide en $\mathcal{O}(N^{-1}(logN)^{p-1})$ (sous des conditions relativement faibles de régularité de f). Parmi les séquences utilisées, celles de Halton [23], LP_τ*-sequences* de Sobol [59] ou encore de Faure [20] peuvent être citées.
Les méthodes de Quasi-Monte Carlo Randomisé, sous certaines conditions peu restrictives sur f, ont un taux de convergence en $\mathcal{O}(N^{-\frac{3}{2}}(logN)^{-\frac{p-1}{2}})$, et permettent une approximation de l'erreur d'estimation. Owen [39] présente ces méthodes comme une ré-randomisation des séquences utilisées dans les méthodes de quasi-Monte Carlo : on prend les séquences déterministes a_i de ces dernières, et on les transforme en variables aléatoire x_i. Cette transformation se fait par exemple par :

$$x_i = a_i + U \mod 1,$$

[2] «pseudo» puisqu'elle consiste en un échantillonnage non totalement aléatoire

où $U \sim U[0,1]^p$. Il n'existe pas de résultats connus concernant l'utilisation de ces méthodes pour l'analyse de sensibilité.
Homma et Saltelli [26] ont comparé la méthode d'échantillonnage par hypercube latin et la méthode de quasi-Monte Carlo basée sur les séquences LP_τ de Sobol aux méthodes classiques de Monte-Carlo (échantillonnage aléatoire), pour l'estimation d'indices de sensibilité. La figure 1.1 illustre ces trois méthodes d'échantillonnage.

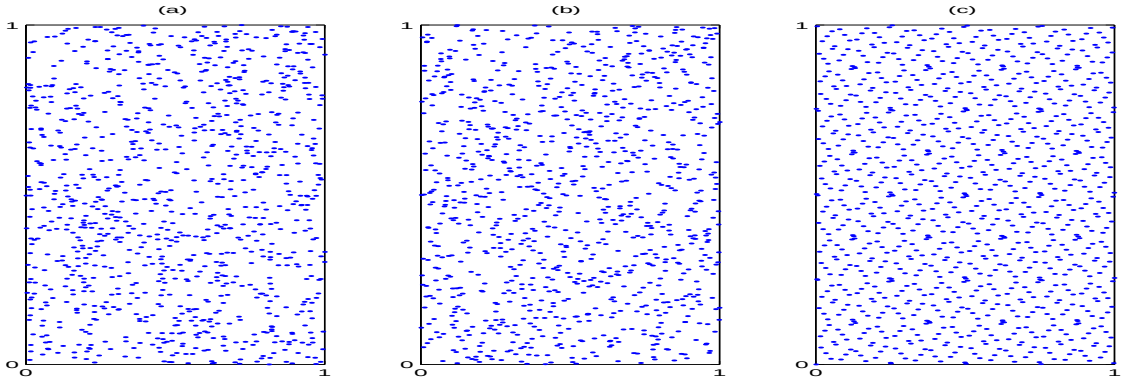


FIG. 1.1. : 1000-échantillon de variables uniformes sur $[0,1]^2$, par échantillonnage aléatoire (a), par hypercube latin (b) et par quasi-Monte Carlo (séquence LP_τ de Sobol) (c)

Homma et Saltelli ont montré que l'utilisation des séquences LP_τ de Sobol permettait une convergence plus rapide que les deux autres méthodes.
Néanmoins, nous utiliserons dans cette thèse la méthode de Monte Carlo basique (échantillonnage aléatoire simple) pour les estimations, qui est la plus simple à mettre en oeuvre et la plus couramment utilisée en analyse de sensibilité. L'utilisation d'autres méthodes d'échantillonnage est une des perspectives de ce travail de thèse.

Estimation des indices de sensibilité par Monte Carlo

Considérons un N-échantillon de réalisations des variables d'entrée $(X_1, \ldots, X_p)$:

$$\tilde{X}_{(N)} = (x_{k1}, \ldots, x_{kp})_{k=1..N}$$

L'espérance de Y, $\mathrm{E}[Y] = f_0$, et sa variance, $\mathrm{V}(Y) = V$, sont estimées par :

$$\hat{f}_0 = \frac{1}{N}\sum_{k=1}^{N} f(x_{k1}, \ldots, x_{kp}), \tag{1.16}$$

$$\hat{V} = \frac{1}{N}\sum_{k=1}^{N} f^2(x_{k1}, \ldots, x_{kp}) - \hat{f}_0{}^2. \tag{1.17}$$

L'estimation des indices de sensibilité nécessite l'estimation d'espérance de variance conditionnelle. Nous présentons une technique d'estimation due à Sobol [60].
L'estimation des indices de sensibilité de premier ordre (1.12) consiste à estimer la quantité :

$$V_i = \mathrm{V}(\mathrm{E}[Y|X_i]) = \underbrace{\mathrm{E}[\mathrm{E}[Y|X_i]^2]}_{U_i} - \mathrm{E}[\mathrm{E}[Y|X_i]]^2 = U_i - \mathrm{E}[Y]^2,$$

la variance de Y étant estimée classiquement par (1.17). Sobol propose d'estimer la quantité U_i, c'est-à-dire l'espérance du carré de l'espérance de Y conditionnellement à X_i, comme une espérance classique, mais

en tenant compte du conditionement à X_i en faisant varier dans les deux appels à la fonction f toutes les variables sauf la variable X_i. Ceci nécessite deux échantillons de réalisations des variables d'entrée, que nous notons $\tilde{X}^{(1)}_{(N)}$ et $\tilde{X}^{(2)}_{(N)}$:

$$\begin{aligned}\hat{U}_i = \frac{1}{N}\sum_{k=1}^{N} & \quad f\left(x^{(1)}_{k1},\ldots,x^{(1)}_{k(i-1)},x^{(1)}_{ki},x^{(1)}_{k(i+1)},\ldots,x^{(1)}_{kp}\right)\\ & \times \quad f\left(x^{(2)}_{k1},\ldots,x^{(2)}_{k(i-1)},x^{(1)}_{ki},x^{(2)}_{k(i+1)}\ldots,x^{(2)}_{kp},\right).\end{aligned}$$

Les indices de sensibilité de premier ordre sont alors estimés par :

$$\hat{S}_i = \frac{\hat{V}_i}{\hat{V}} = \frac{\hat{U}_i - \hat{f}_0{}^2}{\hat{V}}.$$

Pour les indices de sensibilité de second ordre $S_{ij} = \frac{V_{ij}}{V}$, où :

$$V_{ij} = \mathrm{V}(\mathrm{E}[Y|X_i,X_j]) - V_i - V_j = U_{ij} - \mathrm{E}[Y]^2 - V_i - V_j,$$

nous estimons les quantités $U_{ij} = \mathrm{E}[\mathrm{E}[Y|X_i,X_j]^2]$ de la même manière, en faisant varier toutes les variables sauf X_i et X_j :

$$\begin{aligned}\hat{U}_{ij} = \frac{1}{N}\sum_{k=1}^{N} & \quad f\left(x^{(1)}_{k1},\ldots,x^{(1)}_{k(i-1)},x^{(1)}_{ki},x^{(1)}_{k(i+1)},\ldots,x^{(1)}_{k(j-1)},x^{(1)}_{kj},x^{(1)}_{k(j+1)},\ldots,x^{(1)}_{kp}\right)\\ & \times \quad f\left(x^{(2)}_{k1},\ldots,x^{(2)}_{k(i-1)},x^{(1)}_{ki},x^{(2)}_{k(i+1)},\ldots,x^{(2)}_{k(j-1)},x^{(1)}_{kj},x^{(2)}_{k(j+1)},\ldots,x^{(2)}_{kp}\right).\end{aligned}$$

L'indice S_{ij} est alors estimé par :

$$\hat{S}_{ij} = \frac{\hat{U}_{ij} - \hat{f}_0{}^2 - \hat{V}_i - \hat{V}_j}{\hat{V}}.$$

Et ainsi de suite pour les indices de sensibilité d'ordre supérieur.

Remarque. *L'estimation des indices de sensibilité d'ordre i, ($1 < i \leq p$), nécessite l'estimation des indices de sensibilité d'ordre 1 à $i-1$.*

Par contre, les indices de sensibilité totaux peuvent être estimés directement. En effet, la propriété (1.14) permet d'écrire que :

$$S_{T_i} = 1 - \frac{\mathrm{V}(\mathrm{E}[Y|X_{\sim i}])}{\mathrm{V}(Y)} = 1 - \frac{V_{\sim i}}{V}.$$

où $V_{\sim i}$ est la variance de l'espérance de Y conditionnellement à toutes les variables sauf X_i. $V_{\sim i}$ est alors estimée comme V_i, sauf qu'au lieu de faire varier toutes les variables sauf X_i, nous ne faisons varier uniquement X_i.
Ainsi, pour estimer $V_{\sim i} = \mathrm{E}[\mathrm{E}[Y|X_{\sim i}]^2] - \mathrm{E}[\mathrm{E}[Y|X_{\sim i}]]^2 = U_{\sim i} - \mathrm{E}[Y]^2$, on estime $U_{\sim i}$ par :

$$\begin{aligned}\hat{U}_{\sim i} = \frac{1}{N}\sum_{k=1}^{N} & \quad f\left(x^{(1)}_{k1},\ldots,x^{(1)}_{k(i-1)},x^{(1)}_{ki},x^{(1)}_{k(i+1)},\ldots,x^{(1)}_{kp}\right)\\ & \times \quad f\left(x^{(1)}_{k1},\ldots,x^{(1)}_{k(i-1)},x^{(2)}_{ki},x^{(1)}_{k(i+1)}\ldots,x^{(1)}_{kp},\right),\end{aligned}$$

et donc

$$\hat{S}_{T_i} = 1 - \frac{\hat{U}_{\sim i} - \hat{f}_0^{\,2}}{\hat{V}}.$$

Remarque. *En utilisant une taille d'échantillon de Monte Carlo de N, le nombre réel de simulations des variables d'entrée nécessaires à l'estimation des indices de sensibilité est $2N$, puisque cette estimation nécessite deux jeux de simulations. Le nombre d'appels à la fonction du modèle est alors $N \times (k+1)$, où k est le nombre d'indices estimés. Pour un modèle à p variables d'entrée, l'estimation de tous les indices de sensibilité nécessite $N \times (2^p)$ appels à la fonction. En revanche, n'estimer que les indices de premier ordre et les indices totaux ne demande que $N \times (2p+1)$ appels.*

1.1.3.3. La méthode FAST

La méthode FAST (*Fourier Amplitude Sensitivity Test*) a été développée par Cukier et al. [13], [15], [14], ainsi que Schaibly et Shuler [56].
Considérons une fonction

$$f(\mathbf{x}) = f(x_1, \ldots, x_p),$$

où $\mathbf{x} \in [0,1]^p$, et le modèle à variables aléatoires $Y = f(X_1, \ldots, X_p)$ associé.
Cukier *et al.* montrent qu'il est possible d'obtenir une décomposition de la variance de Y, semblable à la décomposition (1.9) de Sobol, en utilisant la transformée de Fourier multi-dimensionnelle de f. Le calcul d'une telle décomposition multi-dimensionnelle étant trop complexe pour être réalisé en pratique, l'idée de la méthode FAST est de remplacer les décompositions multi-dimensionnelles par des décompositions uni-dimensionnelles le long d'une courbe parcourant l'espace $[0,1]^p$.
Cette courbe est définie par un ensemble d'équations paramétriques :

$$x_i(s) = g_i(\sin(\omega_i s)) \qquad \text{pour } i = 1, ..., p,$$

où g_i sont des fonctions à déterminer, permettant un recouvrement uniforme de $[0,1]^p$ (nous en présentons un exemple dans le paragraphe suivant), et où $(\omega_1, ..., \omega_p) \in \mathbb{N}^{*p}$ est un ensemble de fréquences entières linéairement indépendantes (aucune n'est combinaison linéaire des autres).
Ainsi, lorsque s varie dans $\mathbb{R}$, le vecteur $(x_1(s), \ldots, x_p(s))$ décrit une courbe qui parcours $[0,1]^p$.
Cukier et al. montrent que l'on a alors :

$$f_0 = \int_{[0,1]^p} f(\mathbf{x}) dx = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} f(\mathbf{x}(s)) ds.$$

Les fréquences $(\omega_1, ..., \omega_p)$ étant entières, la courbe ne remplie pas l'espace $[0,1]^p$ mais est périodique de période 2π, d'où :

$$f_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\mathbf{x}(s)) ds.$$

Si on applique ces idées au calcul de la variance V d'un modèle :

$$Y = f(X_1, \ldots, X_p),$$

en notant $f_0 = \mathrm{E}[Y]$, on obtient

$$\begin{aligned} V &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f^2(\mathbf{x}(s))ds - f_0^2 \\ &\simeq \sum_{j=-\infty}^{\infty}(A_j^2 + B_j^2) - (A_0^2 + B_0^2) \\ &\simeq 2\sum_{j=1}^{\infty}(A_j^2 + B_j^2), \end{aligned} \tag{1.18}$$

où A_j et B_j sont les coefficients de Fourier définis par :

$$A_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\mathbf{x}(s))\cos(js)ds,$$

$$B_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\mathbf{x}(s))\sin(js)ds.$$

Cukier et al. expliquent que la part de la variance (1.18) due à une variable X_i est la somme des carrés des coefficients de Fourier A_j et B_j attribués à la fréquence ω_i relative à X_i et à ses harmoniques (multiplication par un entier) :

$$V_i = 2\sum_{p=1}^{\infty}(A_{p\omega_i}^2 + B_{p\omega_i}^2).$$

L'indice de sensibilité S_i est alors défini par :

$$S_i = \frac{\sum_{p=1}^{\infty}(A_{p\omega_i}^2 + B_{p\omega_i}^2)}{\sum_{j=1}^{\infty}(A_j^2 + B_j^2)}.$$

Saltelli et al. [54] ont introduit la méthode *Extended FAST*, qui est une extension de cette méthode aux indices de sensibilité totaux, en évaluant la part de variance due à toutes les variables sauf X_i comme la somme des carrés des coefficients de Fourier A_j et B_j attribués à toutes les fréquences $\omega_{\sim i}$ autre que ω_i et ses harmoniques :

$$V_{\sim i} = 2\sum_{p=1}^{\infty}(A_{p\omega_{\sim i}}^2 + B_{p\omega_{\sim i}}^2).$$

L'indice de sensibilité total S_{T_i}, comme pour la méthode de Sobol, est donné par :

$$S_{T_i} = 1 - \frac{\sum_{p=1}^{\infty}(A_{p\omega_{\sim i}}^2 + B_{p\omega_{\sim i}}^2)}{\sum_{j=1}^{\infty}(A_j^2 + B_j^2)}.$$

Remarque. *Saltelli et Bolado [47] montrent, en s'appuyant sur un certain nombre d'exemples tests, que les indices de sensibilité définis par la méthode FAST sont équivalents à ceux définis précédement et estimés par Sobol : (1.12) et (1.13).*

Estimation FAST des indices de sensibilité

L'estimation des indices de sensibilité nécessite de définir les fonctions g_i et les fréquences ω_i utilisées. Une borne M pour l'évaluation des sommes doit aussi être choisie, les sommes infinies n'étant numériquement pas évaluables. En effet les indices de sensibilité de premier ordre et totaux seront estimés de la façon suivante :

$$\hat{S}_i = \frac{2\sum_{p=1}^{M}(A^2_{p\omega_i} + B^2_{p\omega_i})}{2\sum_{j=1}^{M}(A^2_j + B^2_j)} \quad \text{et} \quad \hat{S}_{T_i} = 1 - \frac{2\sum_{p=1}^{M}(A^2_{p\omega_{\sim i}} + B^2_{p\omega_{\sim i}})}{2\sum_{j=1}^{M}(A^2_j + B^2_j)},$$

où les intégrales A_j et B_j sont estimées par une méthode de Monte-Carlo classique, et où M, harmonique maximum considéré, est évalué en fonction des propriétés suivantes :

- plus M est grand, mieux les indices reflètent l'effet des variables,
- mais aussi plus M est grand, plus le nombre de simulations (nombre de points pris sur la courbe qui explore le cube unité) sera élevé.

Le choix de M revient à faire un compromis entre la qualité des indices et le coût de leur estimation. Cukier et al. ont déterminé de façon empirique que le meilleur compromis pour M était 4 ou 6, et ce quelque soit la dimension du modèle. Nous optons pour le choix classique $M = 4$ pour nos expérimentations.
Pour le choix des fonctions g_i, plusieurs possibilités ont été proposées. Nous retenons celles de Saltelli et al. [54], définies pour $s \in [-\pi, \pi]$ par :

$$x_i(s) = g_i(\sin(\omega_i s)) = \frac{1}{2} + \frac{1}{\pi}\arcsin(\sin(\omega_i s)), \tag{1.19}$$

qui sont celles qui recouvrent le mieux l'espace $[0,1]^p$ en respectant une distribution uniforme des échantillons. La figure 1.2 représente l'allure de la courbe $g_1(\sin(\omega_1 s))$, pour $s \in [-\pi, \pi]$ et $\omega_1 = 11$.

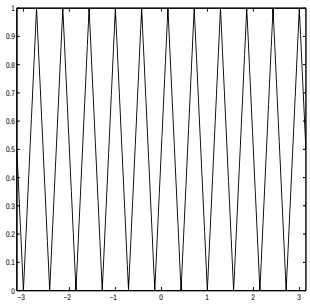


FIG. 1.2. : Fonction $g_1(\sin(\omega_1 s))$ pour $s \in [-\pi, \pi]$ et $\omega_1 = 11$.

Sur la figure 1.3, les graphiques (a) et (b) représentent un échantillonnage selon la transformation (1.19), respectivement dans le cas de deux ($\omega_1 = 11$ et $\omega_2 = 21$) et trois ($\omega_1 = 11$, $\omega_2 = 21$ et $\omega_3 = 27$) variables d'entrées.
Définissons enfin le choix des fréquences ω_i. En pratique, il arrive souvent que l'on utilise l'alternative simple suivante : on choisit une fréquence élevée pour ω_i (20 par exemple) et des fréquences plus petites pour les autres $\omega_{\sim i}$ (1 par exemple). Saltelli et al. [54] montrent que ce choix tend à surestimer les V_i, et propose un algorithme automatique pour choisir ces fréquences :

(i) choix d'une fréquence ω_i, sachant que le nombre de simulations minimum nécessaire (N donné en (1.20)) est fonction de ce choix,

(ii) détermination de la valeur maximale des autres fréquences $\max(\omega_{\sim i}) = \frac{\omega_i}{2M}$,

(iii) les autres fréquences varient de 1 à $\max(\omega_{\sim i})$, avec les deux conditions suivantes :

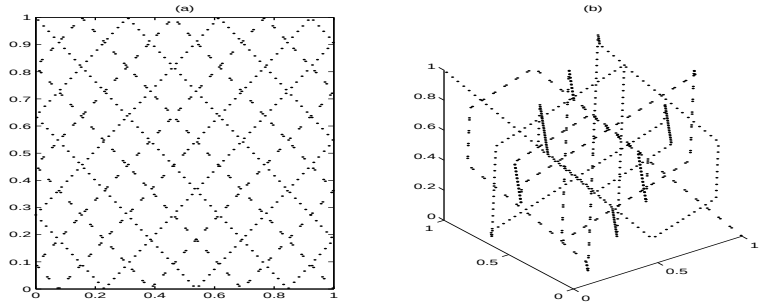


FIG. 1.3. : Echantillonages selon la transformation $g_i(\sin(\omega_i s))$ en dimension deux (a) et trois (b).

a) le pas entre chaque fréquence doit être aussi grand que possible,

b) le nombre de variables X_i avec la même fréquence doit être aussi faible que possible.

Par exemple, dans le cas d'un modèle à huit variables, où l'on veut estimer S_{T_4}, et où $\omega_4 = 32$, l'algorithme précédent conduit à choisir :

ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
1	2	3	32	1	2	3	4

Enfin, comme évoqué précédemment, Saltelli et al. [54] détermine une taille d'échantillon minimale pour l'évaluation de ces indices, donnée par :

$$N = 2M \max_i(\omega_i) + 1. \tag{1.20}$$

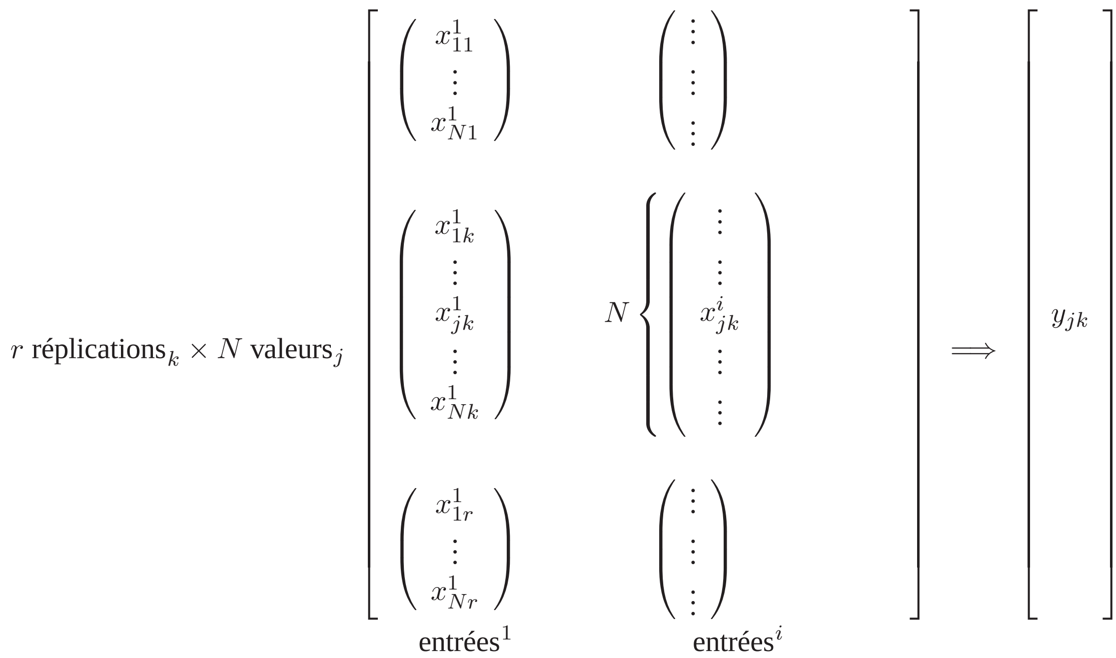
Un des avantages de cette méthode est que les indices de sensibilité peuvent être calculés indépendamment les uns des autres, à partir d'un même échantillon de simulations, ce que ne permet pas la méthode de Sobol qui nécessite deux échantillons. Par contre, la méthode de Sobol étant stochastique, elle permet d'obtenir un intervalle de confiance sur les estimations d'indices, ce que ne permet pas FAST, puisque pour une série de fréquences donnée, les estimations d'indices sont déterministes. Saltelli et Bolado [47] ont comparé FAST et Sobol sur un certain nombre de modèle. Ils ont conclu que FAST était, d'un point de vue complexité de calcul, plus avantageuse que Sobol. Néanmoins, FAST est parfois sujette à un biais (que Saltelli et Bolado assigne au choix des fréquences), tandis que Sobol converge toujours vers la vraie valeur des indices de sensibilité.

Remarque. *Entrées non uniformes*.
Comme pour tout système d'échantillonnage (cf 1.1.3.2), celui de FAST est défini de façon uniforme sur le cube unité. Lorsque les variables du modèle ne sont pas uniformes, une transformation sur les points échantillonnés sera appliquée pour les rendre de loi voulue. Cette transformation se fera en général par l'inverse de la fonction de répartition, et pourra parfois être approchée lorsque l'inverse n'existe pas (cas gaussien).

Les méthodes de Sobol et FAST sont les deux méthodes les plus couramment utilisées en analyse de sensibilité pour estimer les indices de sensibilité. Toutefois, une troisième méthode, antérieure à celles-ci, a été proposée par McKay.

1.1.3.4. La méthode de McKay

La méthode d'estimation des indices de sensibilité proposée par McKay, [34], se base sur l'échantillonnage par hypercube latin répliqué (*r-LHSampling*). À partir d'un N-échantillon créé selon le plan d'échantillonnage par hypercube latin, décrit au paragraphe 1.1.3.2 (N premières lignes de la matrice ci-dessous), on crée r réplications (paquet de N lignes) en permutant indépendamment et aléatoirement les N valeurs de chaque variable (i.e. colonne). La réunion de ces r réplications donnera $N \times r$ échantillons pour chaque variable.
Ce schéma d'échantillonnage par hypercube latin répliqué peut être représenté par la figure 1.4.



$1 \leq j \leq N$: N valeurs des variables d'entrée (prises dans des intervalles équiprobables),
$1 \leq k \leq r$: r permutations des N-vecteurs de simulations des entrées,
$1 \leq i \leq p$: p paramètres.

FIG. 1.4. : Échantillonnage par hypercube latin répliqué.

Les moyennes suivantes sont alors définies :

$$\overline{y}_{j.} = \frac{1}{r}\sum_{k=1}^{r} y_{jk} \qquad \overline{y} = \frac{1}{N}\sum_{j=1}^{N} \overline{y}_{j.},$$

où $\overline{y_{j.}}$ est la moyenne *inter* réplications et $\overline{y}$ est la moyenne sur toutes les valeurs de y.

L'estimation de l'indice de sensibilité de premier ordre de la variable X_i, défini par (1.12) nécessite l'estimation des quantités :

$$\mathrm{V}(\mathrm{E}[Y|X_i]) \quad \text{et} \quad \mathrm{V}(Y).$$

La variance totale $\mathrm{V}(Y)$ peut être estimée par :

$$\widehat{\mathrm{V}^{(*)}(Y)} = \frac{1}{r}\sum_{k=1}^{r}\underbrace{\frac{1}{N}\sum_{j=1}^{N}(y_{jk}-\overline{y}_{.k})^2}_{\widehat{\mathrm{V}_{k(Y)}}}, \tag{1.21}$$

où $\overline{y}_{.k} = \dfrac{1}{N}\sum_{j=1}^{N} y_{jk}$ et $\widehat{\mathrm{V}_k(Y)}$ sont les estimations de la moyenne et de la variance de Y au sein de la réplication k (*intra* réplications).
En utilisant la formule classique de l'analyse de la variance, pour une somme de carrés *intra* et *inter* réplications, qui s'écrit :

$$\begin{aligned}\sum_{k=1}^{r}\sum_{j=1}^{N}(y_{jk}-\overline{y})^2 &= \underbrace{\sum_{k=1}^{r}\sum_{j=1}^{N}(\overline{y}_{.k}-\overline{y})^2}_{\textit{inter}} + \underbrace{\sum_{k=1}^{r}\sum_{j=1}^{N}(y_{jk}-\overline{y}_{.k})^2}_{\textit{intra}} \\ &= N\sum_{k=1}^{r}(\overline{y}_{.k}-\overline{y})^2 + \sum_{k=1}^{r}\sum_{j=1}^{N}(y_{jk}-\overline{y}_{.k})^2,\end{aligned}$$

on a :

$$\widehat{\mathrm{V}^{(*)}(Y)} = \frac{1}{Nr}\sum_{k=1}^{r}\sum_{j=1}^{N}(y_{jk}-\overline{y})^2 - \frac{1}{r}\sum_{k=1}^{r}(\overline{y}_{.k}-\overline{y})^2.$$

Or, pour un échantillonnage *LHS*, comme $\mathrm{E}[(\overline{y}_{.k}-\overline{y})^2]$ est en $\frac{1}{N}$, le dernier terme de cette égalité peut être considéré comme négligeable pour une taille d'échantillon N suffisamment grande. McKay propose alors l'estimation de la variance totale suivante :

$$\widehat{\mathrm{V}(Y)} = \frac{1}{Nr}\sum_{j=1}^{N}\sum_{k=1}^{r}(y_{jk}-\overline{y})^2.$$

Soient $\overline{Y}_{j.}$ et $\overline{Y}$ les variables aléatoires dont $\overline{y}_{j.}$ et $\overline{y}$ sont les réalisations sur notre matrice d'échantillonnage. Comme :

$$\begin{aligned}\mathrm{E}\left[(\overline{Y}_{j.}-\overline{Y})^2\right] &\simeq \mathrm{V}(\overline{Y}_{j.}) \\ &= \mathrm{V}(\mathrm{E}[\overline{Y}_{j.}|X_i]) + \mathrm{E}[\mathrm{V}(\overline{Y}_{j.}|X_i)] \\ &= \mathrm{V}(\mathrm{E}[Y|X_i]) + \frac{1}{r}\mathrm{E}[\mathrm{V}(Y|X_i)],\end{aligned}$$

le terme $\mathrm{V}(E[Y|X_i])$ est estimé par :

$$\frac{1}{N}\sum_{j=1}^{N}(\overline{y}_{j.}-\overline{y})^2 - \frac{1}{r}\frac{1}{Nr}\sum_{j=1}^{N}\sum_{k=1}^{r}(y_{jk}^{(i)}-\overline{y}_{j.})^2,$$

où $\dfrac{1}{Nr}\sum_{j=1}^{N}\sum_{k=1}^{r}(y_{jk}^{(i)}-\overline{y}_{j.})^2$ est l'estimateur de $\mathrm{E}[\mathrm{V}(Y|X_i)]$, avec $y_{jk}^{(i)}$ obtenu en fixant, pour la variable X_i, les r réplications, (x_{jk}^i constant sur k, c'est-à-dire $x_{j1}^i = x_{j2}^i = ... = x_{jr}^i$ pour tout $1 \leq j \leq N$).

L'indice de sensibilité de premier ordre de la variable X_i, défini par (1.12) est alors estimé par :

$$S_i = \frac{r\sum_{j=1}^{N}(\overline{y}_{j.} - \overline{y})^2 - \frac{1}{r}\sum_{j=1}^{N}\sum_{k=1}^{r}(y_{jk}^{(i)} - \overline{y}_{j.})^2}{\sum_{j=1}^{N}\sum_{k=1}^{r}(y_{jk} - \overline{y})^2}.$$

La notion de sensibilité totale étant introduite postérieurement à la méthode d'estimation de McKay, cette dernière ne propose pas d'estimation pour les indices de sensibilité totaux.

1.1.4. Une approche non paramétrique : les modèles additifs

Une approche du problème de l'analyse de sensibilité basée sur les modèles additifs (*AM*) [24] est possible. Un modèle additif (*AM*) permet naturellement, *via* sa construction, de séparer l'effet de chaque variable d'entrée. Il est alors possible d'obtenir une décomposition de la variance en fonction des variables d'entrée, à partir de laquelle il est possible de déduire des estimations d'indices de sensibilité.
Un modèle additif est défini par :

$$Y = \alpha + \sum_{i=1}^{p} f_i(X_i), \tag{1.22}$$

où les f_i sont des fonctions d'une variable.
Comme les variables d'entrée sont supposées indépendantes, la variance d'un tel modèle s'écrit :

$$\mathrm{V}(Y) = \sum_{i=1}^{p} \mathrm{V}(f_i(X_i)).$$

Comme dans le cas d'un modèle linéaire (1.1.2.1), la part de variance due à la variable X_i peut être extraite, et est ici égale à $\mathrm{V}(f_i(X_i))$. On peut donc naturellement définir les indices de sensibilité de premier ordre par :

$$S_i = \frac{\mathrm{V}(f_i(X_i))}{\mathrm{V}(Y)}.$$

On a alors la proposition suivante :

Proposition 1.1.2. *Les indices de sensibilité $S_i = \frac{V(f_i(X_i))}{V(Y)}$ sont équivalents aux indices de sensibilité classiques (1.12).*

Démonstration. La preuve est immédiate dès l'instant où l'on remarque que l'espérance d'un modèle additif (1.23) conditionnellement à une variable X_i s'écrit :

$$\mathrm{E}[Y|X_i] = \alpha + \sum_{\substack{j=1 \\ j\neq i}}^{p} \mathrm{E}[f_j(X_j)] + f_i(X_i),$$

d'où

$$\mathrm{V}(\mathrm{E}[Y|X_i]) = \mathrm{V}(f_i(X_i)).$$

□

1.1.4.1. Estimation des indices de sensibilité basés sur les modèles additifs

Comme pour les modèles linéaires, le modèle étudié n'est généralement pas additif, et les fonctions f_i ne sont généralement pas connues. Il est donc nécessaire d'estimer ce modèle additif par :

$$\hat{Y} = \hat{\alpha} + \sum_{i=1}^{p} \hat{f}_i(X_i), \tag{1.23}$$

à partir d'un échantillon de simulations $(y, x_{1j}, \ldots, x_{pj})_{j=1..N}$. Les fonctions f_i sont estimées par exemple par des fonctions splines, que nous choisissons cubiques. Les fonctions splines cubiques sont des fonctions polynomiales par morceaux continues et ayant leurs deux premières dérivées continues aux noeuds [64].
L'algorithme généralement utilisé pour estimer un modèle additif est appelé *backfitting algorithm*, et résulte en un ajustement non paramétrique des fonctions f_i, c'est-à-dire les valeurs des fonctions aux points de l'échantillon : $(f_i(x_{ij}))_{j=1..N}$. À partir de ces valeurs, les variances $\mathrm{V}(f_i(X_i))$ sont estimées par une méthode de Monte-Carlo classique.
Il est possible d'apporter une première critique à cette méthode : si le modèle étudié est fortement non additif, l'ajustement d'un modèle additif ne sera que peu vraisemblable, et l'estimation des indices de sensibilité sera faussée. La qualité de l'ajustement devra être vérifiée avant toute interprétation des valeurs des indices de sensibilité. Par contre, pour un modèle additif, l'efficacité du *backfitting algorithm* devrait permettre une estimation des indices de sensibilité peu coûteuse. C'est ce que nous testons dans le paragraphe suivant.

1.1.4.2. Tests et applications des modèles additifs pour l'analyse de sensibilité

Considérons le modèle :

$$Y = \frac{{X_2}^4}{{X_1}^2}, \tag{1.24}$$

où $X_i \sim \mathcal{U}[0.5, 1.5]$ pour $i = 1, 2$.
La figure 1.5 représente Y en fonction de X_1 et X_2.
Les indices de sensibilité de ce modèle peuvent se calculer analytiquement, et sont égaux à :

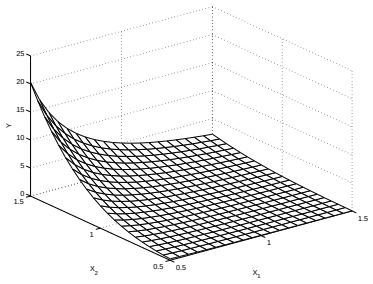


FIG. 1.5. : Représentation 3D de $Y = \frac{{X_2}^4}{{X_1}^2}$.

$$S_1 \simeq 0.262 \qquad S_2 \simeq 0.511.$$

Tous les résultats numériques de calcul d'indices de sensibilité présentés dans cette thèse ont été obtenus à l'aide de l'interface Matlab présentée en annexe A.4. Le premier résultat à analyser avant le calcul des indices de sensibilité, est la qualité de l'ajustement du modèle (1.24) par un modèle additif. La figure 1.6 présente ainsi le pourcentage de variance expliquée par le modèle additif, pour différentes tailles d'échantillon de simulations.

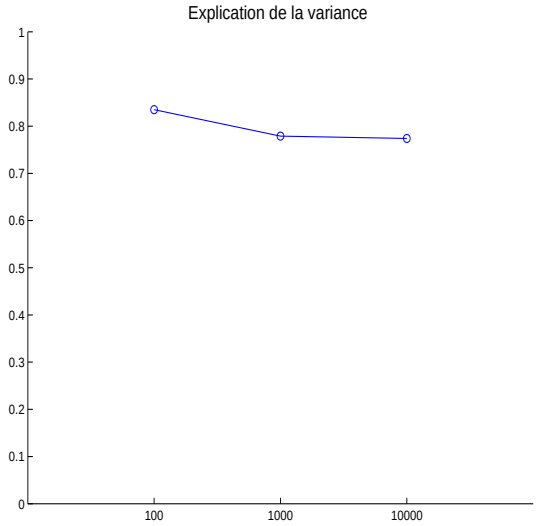


FIG. 1.6. : Part de variance expliquée pour l'ajustement de $Y = \frac{{X_2}^4}{{X_1}^2}$ par un modèle additif, en fonction du nombre de simulations du modèle.

Le modèle est relativement bien expliqué (80%), il est donc possible de se fier aux calculs d'indices. La figure 1.7 illustre la convergence des résultats de vingt calculs d'indices avec un échantillon de simulations de taille 100, 1000, 10000 et enfin 100000 (le point représente la moyenne et le trait l'étendue à plus ou moins 2 écarts-types). Le trait continu représente la vraie valeur des indices de sensibilité, calculée de façon analytique.

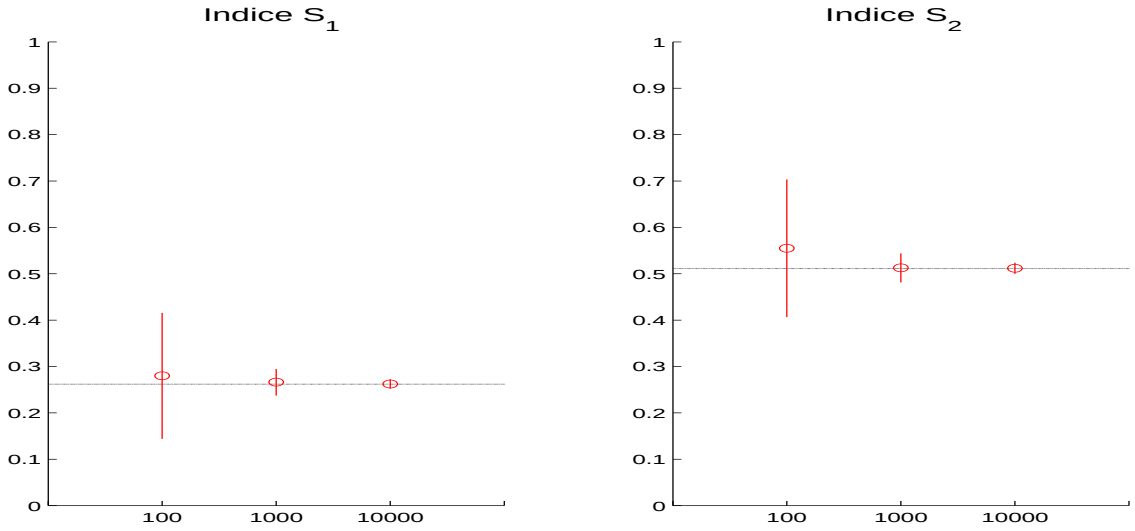


FIG. 1.7. : Indices de sensibilité de premier ordre de $Y = \frac{{X_2}^4}{{X_1}^2}$ en fonction du nombre de simulations du modèle.

Pour cet exemple, les variables aléatoires sont uniformes, ce qui confère une grande importance aux variables seules. Ainsi, l'estimation obtenue est correcte (relativement proche de la valeur réelle des indices), avec une variabilité relativement faible même pour un échantillon de taille 100.
Par contre, pour un modèle sensible seulement à l'interaction entre variables, le modèle additif que nous avons défini n'arrive pas à ajuster correctement le modèle. On obtient alors des résultats de sensibilité erro-

nés.
Illustrons ce propos à l'aide du modèle :

$$Y = X_1 X_2 X_3,$$

où $X_i \sim \mathcal{N}(0,1)$ pour $i = 1,2,3$. Les indices de sensibilité de ce modèle se calculent analytiquement, et sont tous nuls sauf $S_{123} = 1$.
L'explication de ce modèle par modèle additif est de l'ordre de 50% avec 100 points, mais chute rapidement à moins de 15% avec 1000 points. Comme ce modèle s'ajuste mal par un modèle additif, plus le nombre de points est important, plus la part de variance expliquée est faible. Les valeurs estimées des indices de sensibilité n'ont donc aucune signification et sont totalement erronées (non nulles), nous ne les présentons pas.
La méthode d'analyse de sensibilité par modèles additifs présentée ci-dessus souffre logiquement de sa construction, qui ne permet d'estimer les indices de sensibilité que pour des modèles qui s'expliquent relativement bien par des modèles additifs. Comme les indices SRC sont utilisables pour les modèles relativement linéaires, les indices définis ci-dessus sont utilisables pour les modèles relativement additifs. Une des possibilité d'extension de cette méthode serait, lorsque la part de variance expliquée n'est pas assez importante, d'introduire dans le modèle additif des fonctions de deux variables, puis de trois variables, et ainsi de suite, à la manière de la décomposition de Sobol.

1.1.5. Applications numériques des méthodes de McKay, Sobol, FAST et AM (modèles additifs)

Nous avons présenté trois méthodes d'estimation d'indices de sensibilité : celles due à McKay et à Sobol ainsi que la méthode FAST. Ces trois méthodes nous intéressent particulièrement car elles ne font aucune hypothèse sur la nature du modèle, contrairement aux indices SRC, PCC et leur version sur les rangs. Néanmoins, ces derniers indices doivent être utilisés lorsque les hypothèses nécessaires sont satisfaites, puisqu'ils sont généralement peu coûteux à estimer. Nous présentons une courte analyse comparative des trois méthodes précédemment citées, ainsi que de la méthode proposée s'appuyant sur les modèles additifs (AM).
Considérons le modèle d'Ishigami [48] :

$$Y = \sin(X_1) + 7\sin^2(X_2) + \frac{X_3^4}{10}\sin(X_1),$$

où $X_i \sim \mathcal{U}[-\pi, \pi]$ pour $i = 1,2,3$.
La figure 1.8 représente Y en fonction des trois variables X_1, X_2 et X_3.

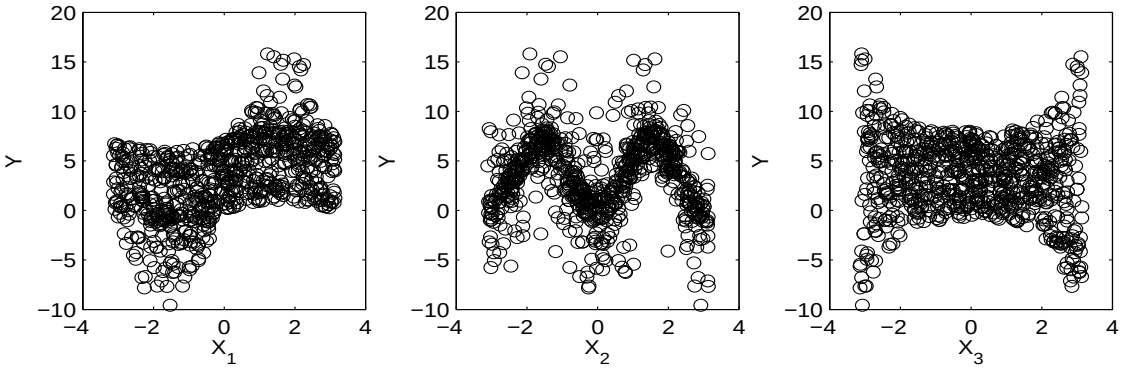


FIG. 1.8. : $Y = \sin(X_1) + 7\sin^2(X_2) + \frac{X_3^4}{10}\sin(X_1)$ en fonction de X_1, X_2 et X_3.

Les calculs numériques d'indices de sensibilité ont été réalisés avec l'interface Matlab présentée en annexe A.4. Pour chacune des méthodes, les indices ont été calculés vingt fois (sauf pour FAST qui donne un résultat déterministe pour une série de fréquence choisie), afin d'obtenir un intervalle de confiance sur la valeur des estimations, et ce pour une taille d'échantillon de 100, 1000 et 10000. Pour la méthode d'estimation de McKay, le calcul n'a été fait que pour un échantillon de taille 100, qui suffit à une estimation correcte (proche de la valeur réelle avec une variance faible). Nous comparons les résultats obtenus sur les indices de sensibilité de premier ordre.
Les valeurs moyennes des indices de sensibilité de premier ordre pour chacune des quatre méthodes sont présentées dans le tableau 1.1.

	N=100			N=1000			N=10000		
méthode	S_1	S_2	S_3	S_1	S_2	S_3	S_1	S_2	S_3
analytique	0.314	0.442	0	0.314	0.442	0	0.314	0.442	0
Sobol	0.318	0.456	$\simeq 0$	0.316	0.449	$\simeq 0$	0.310	0.441	$\simeq 0$
FAST	0.189	0.078	0.002	0.318	0.386	0.005	0.314	0.464	0.0001
AM	0.296	0.476	0.052	0.313	0.451	0.008	0.312	0.445	0.0008
McKay	0.313	0.444	$\simeq 0$						

TAB. 1.1.: Indices de sensibilité de premier ordre du modèle d'Ishigami, par les méthodes de Sobol, FAST, AM et McKay, en fonction de trois tailles d'échantillon.

Pour cette application, toutes les méthodes sauf FAST donnent des estimations correctes à partir d'une taille d'échantillon de 100. Il est normal que FAST échoue, puisque le nombre de simulations minimal pour FAST, donné par (1.20), est ici égal à $N = 2 \times 4 \times 39 + 1 = 313$, où 39 est la plus grande fréquence utilisée par l'interface Matlab pour un modèle à trois variables d'entrée. Remarquons que le pourcentage de variance expliquée par le modèle additif est de l'ordre de 80%, ce qui permet donc une bonne estimation des indices.

Remarque. *Seule la méthode de Sobol permet de calculer les indices de sensibilité d'ordre supérieur. On trouve alors $S_{13} \simeq 0.25$, ce qui permet de conclure que la sortie du modèle d'Ishigami est sensible à X_2 (environ 44%), à X_1 (environ 31%) et à l'interaction entre X_1 et X_3 (environ 25%). Si on calcule par la méthode AM l'indice de sensibilité au produit X_1X_3, nous n'arrivons pas à retrouver cette sensibilité à l'interaction entre X_1 et X_3, car la liaison entre ces deux variables dans le modèle n'est pas un simple produit. La méthode AM nécessiterait, pour calculer les indices d'ordre supérieur, de connaître la forme analytique exacte du modèle, ce qui n'est pas envisageable en pratique.*

Le tableau 1.2 et la figure 1.9 permettent de comparer les trois méthodes (Sobol, McKay et AM) du point de vue de la variabilité des estimateurs.

	N=100			N=1000			N=10000		
méthode	S_1	S_2	S_3	S_1	S_2	S_3	S_1	S_2	S_3
Sobol	0.157	0.122	0.150	0.050	0.057	0.073	0.011	0.014	0.018
AM	0.051	0.064	0.035	0.014	0.023	0.004	0.004	0.007	0.0004
McKay	0.012	0.20	0.003						

TAB. 1.2.: Variation de l'écart-type des estimations des indices de sensibilité du modèle d'Ishigami, pour les méthodes de Sobol, AM et McKay, pour 3 tailles d'échantillon.

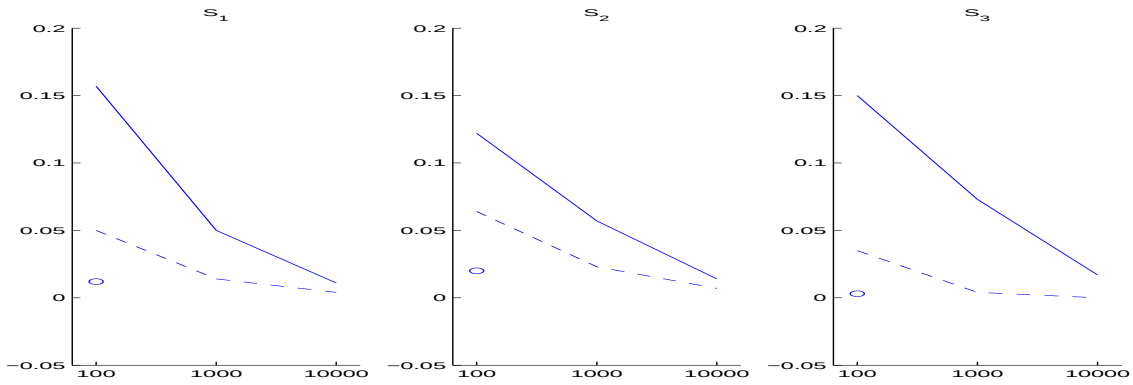


FIG. 1.9. : Variation de l'écart-type des estimations des indices de sensibilité du modèle d'Ishigami, pour les méthodes de Sobol(trait continu), de McKay(rond) et AM(trait discontinu), pour 3 tailles d'échantillon.

La méthode de McKay, donne un écart-type petit avec une taille d'échantillon de 100, alors que la méthode de Sobol nécessite une taille de 10000 pour obtenir la même précision. Mais cette méthode de McKay est très lourde en temps de calcul, même pour une petite taille d'échantillon. Les estimations par modèles additifs sont quant à elles toujours d'écarts-types plus faibles que ceux obtenus par la méthode de Sobol.
Nous concluons par un comparatif de ces quatre méthodes testées ici, fondé sur plusieurs applications (dont celle présentée ci-dessus) et sur une analyse bibliographique.

1.1.5.1. Comparatif des méthodes de McKay, AM (modèles additifs), Sobol et FAST

La méthode de McKay ne permet d'estimer que les indices de premier ordre, et est relativement lourde en temps de calcul, même pour un échantillon de petite taille.
La méthode AM basée sur les modèles additifs, permet d'estimer de façon efficace les indices de sensibilité de premier ordre, pour les modèles qui s'expliquent correctement par un modèle additif. Pour les modèles n'ayant pas cette propriété, la méthode AM ne permet pas d'estimer les indices de sensibilité.
La méthode FAST permet d'estimer les indices de sensibilité de premier ordre et totaux pour tous types de modèles. Bien que nous ne l'ayons pas montré sur l'application étudiée, elle est plus légère numériquement que la méthode de Sobol mais peut introduire un biais [47]. Néanmoins, la méthode n'étant pas exacte, on aurait apprécié d'avoir un intervalle de confiance sur les estimations, ce qui n'est pas possible puisque les estimations fournies par la méthode FAST sont déterministes pour une série de fréquences fixée.
La méthode de Sobol est alors la plus complète. Elle permet d'estimer un intervalle de confiance pour tous les indices de sensibilité (premier ordre, ordre intermédiaire et indices totaux). De plus, cette méthode est simple à mettre en oeuvre et à comprendre. Toutefois la méthode de Sobol est un peu gourmande en temps de calcul, mais l'utilisation de méthodes de quasi-Monte Carlo permet de limiter cet effet.

Nous utilisons donc dans cette thèse la méthode de Sobol pour estimer les indices de sensibilité basés sur la décomposition de la variance. Le système d'échantillonnage de Monte Carlo choisi est l'échantillonnage aléatoire simple, les applications étudiées permettant l'utilisation d'un tel système.

1.1.6. Conclusion

L'analyse de sensibilité (globale) étudie l'impact de la variabilité des entrées d'un modèle sur la variabilité de sa sortie. Elle consiste à évaluer des indices de sensibilité qui quantifient combien une variable ou un groupe de variables contribue à la variance de la sortie. Parmi les méthodes d'analyse de sensibilité nous avons montré que la plus complète est celle due à Sobol, qui définit des indices de sensibilité pour tous les sous-ensembles de variables d'entrée, ainsi que des indices de sensibilité totaux. L'estimation de Sobol

basées sur une méthode de Monte Carlo semble être la plus robuste, bien que relativement gourmande en temps de calcul.

1.2. Incertitude de modèle

Cette section a pour objectif de discuter la notion d'incertitude de modèle en s'appuyant sur la bibliographie correspondante. Les avancées de ces dernières décennies dans le domaine des statistiques et de l'informatique ont permis le développement d'outils de modélisation de plus en plus puissants, et donc de plus en plus utilisés dans de nombreux domaines. Dès lors, une attention particulière a été portée sur le problème de la relation entre les hypothèses de modélisation et les résultats obtenus. Plusieurs groupes de travail ont été créés pour débattre de ce sujet de l'incertitude de modèle, avec notamment en octobre 1993 à Anapolis (Maryland, USA) : *Model Uncertainty : its Characterization and Quantification*, et en 1995 à Bath (Angleterre) : *Model Uncertainty and Model Robustness*. Parmi les investigateurs de ces workshops nous pouvons citer Abramson, Apostolakis, Evans, Laskey, McKay ou encore Theofanous.
Les travaux que nous avons recensés jusqu'à présent peuvent être classés selon trois axes de recherches. Les premiers concernent la détermination et la définition des différentes sources d'incertitude qui interviennent lors de l'élaboration d'un modèle. Un deuxième axe traite de la sélection de modèle lorsque l'on est en présence de plusieurs modèles concurrents pour représenter un même phénomène. Enfin, quelques travaux moins nombreux traitent de l'incertitude de modèle issue de l'utilisation d'un modèle simplifié, qui est une des problématiques qui nous intéresse particulièrement dans ce travail de thèse.
Nous reprenons dans les paragraphes suivants ces principaux travaux en conservant une organisation en trois parties correspondant aux trois axes de recherches explorés sous le thème de l'incertitude de modèle.

1.2.1. Incertitude liée à l'élaboration d'un modèle

Ce sujet a été abordé à de nombreuses reprises dans la littérature, dans laquelle nous citons les articles dont nous nous sommes inspirés pour présenter cette synthèse : [1], [38], [43], [65] ou encore [67]. La construction ou l'élaboration de tout modèle mathématique est sujette à deux sources d'incertitudes : l'incertitude épistémique, due notamment à l'impossibilité de connaître parfaitement le problème étudié, et l'incertitude aléatoire, due au fait que certaines quantités ou paramètres déterministes du modèle sont des estimations de moments de variables aléatoires.

1.2.1.1. Incertitude épistémique

L'incertitude épistémique naît du passage du phénomène réel au modèle mathématique. Nous détaillons ici les différentes sources d'incertitude qui existent à chaque étape de cette construction.

- Du phénomène réel au modèle théorique. La définition du modèle théorique génère une incertitude due à la nature du phénomène considéré, qui peut ne jamais être connue ni comprise parfaitement par l'homme. On souffre souvent d'un manque de connaissance sur le phénomène, qui, de plus, peut évoluer avec le temps. En outre, la physique utilisée (mécanique des fluides, réaction chimique...) n'est pas nécessairement en adéquation parfaite avec le phénomène étudié, de même que les hypothèses nécessaires à l'application de théorèmes et postulats ne sont pas nécessairement exactement vérifiées. Enfin, les conditions initiales ou de bords utilisées sont elles aussi incertaines.
- Du modèle théorique au modèle mathématique. Ce passage entraîne des approximations, sources d'incertitude : il n'est pas toujours possible de prendre en compte la globalité du phénomène physique. Citons pour exemple le logiciel GASCON, étudié en section 2.3, qui simule le transfert de radionucléides d'une installation nucléaire jusqu'à l'homme, et qui par soucis d'exploitabilité, ne tient pas compte des turbulences atmosphériques. De plus, des erreurs de mesures ou un manque de données peuvent aussi apporter des incertitudes sur les valeurs données aux différents paramètres déterministes du modèle.
- Du modèle mathématique au modèle numérique informatisé. L'utilisation du modèle mathématique nécessite généralement qu'il soit discrétisé afin de pouvoir en tirer une solution numérique. Cette discrétisation engendre des approximations qui sont une source d'incertitude supplémentaire.

– Du modèle numérique informatisé à la valeur numérique de la sortie du modèle. Comme elle est calculée de façon informatique, cette valeur est sujette aux incertitudes dues à la précision finie de l'ordinateur utilisé.

Toutes ces incertitudes issues de l'élaboration du modèle sont définies comme des incertitudes épistémiques. Ces incertitudes peuvent être réduite en améliorant la connaissance du phénomène (plus de données par exemple).
La figure 1.10 illustre ces différentes sources d'incertitudes.

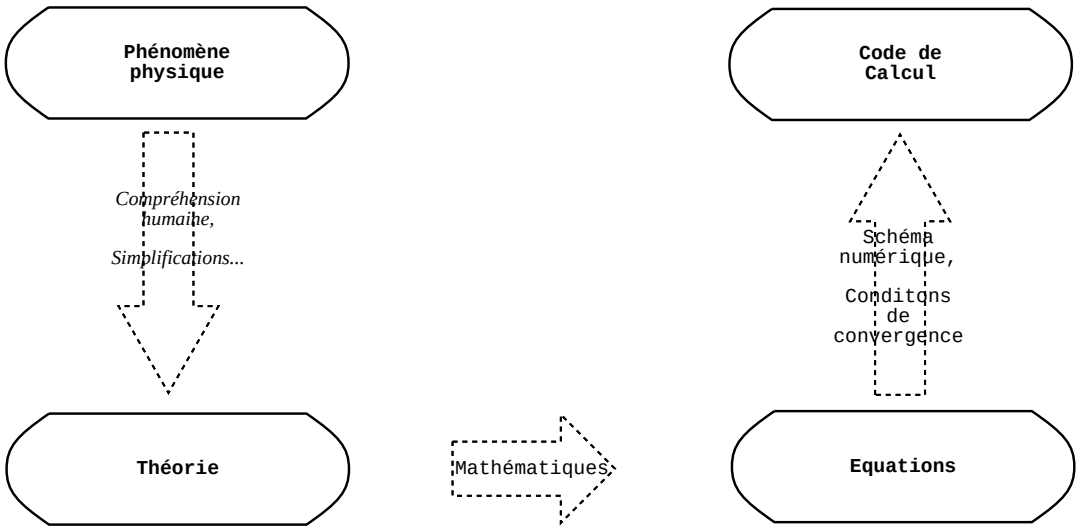


FIG. 1.10. : Différentes sources d'incertitude épistémiques présentes lors de l'élaboration d'un modèle.

1.2.1.2. Incertitude aléatoire

La seconde source d'incertitude est dite aléatoire et apparaît lors de l'estimation des paramètres (déterministes) du modèle. Cette incertitude est due à la variabilité naturelle de toute quantité physique mesurée. On l'explique parfois comme l'incertitude responsable de l'obtention de résultats différents lorsque l'on répète plusieurs fois dans des conditions identiques une expérience. Cette incertitude est inhérente à tout phénomène ou processus physique, et il est impossible de la supprimer.

1.2.2. Le problème des modèles concurrents

Nous présentons maintenant l'axe de recherche le plus abordé sur le thème de l'incertitude de modèle. Supposons que pour calculer une valeur d'une variable aléatoire d'intérêt Y, plusieurs modèles $(M_i)_{1\leq i\leq k}$ sont disponibles.
Les stratégies classiques de sélection de modèles, que nous ne détaillons pas ici, permettent de choisir le meilleur modèle parmi les différents candidats, suivant différents critères. Comme le remarque Raftery [44], les approches qui visent à sélectionner un unique modèle et à l'utiliser pour l'inférence statistique, ignorent l'incertitude issue du processus de sélection, et sous-estime donc l'incertitude totale existante sur Y.
Une idée naturelle est alors d'utiliser un modèle défini comme une moyenne pondérée des différents modèles possibles :

$$Y = \sum_{i=1}^{k} p(M_i) f(\mathbf{X}|M_i), \tag{1.25}$$

où $f(.|M_i)$ est la fonction du modèle M_i. Cette démarche peut être apparentée à celle des modèles de mélange [36].
Le principal problème est alors la définition et l'estimation des probabilités $p(M_i)$.

Winkler, [66], propose trois interprétations ou définitions possibles de cette probabilité :
- $p(M_i)$ est la probabilité que le modèle M_i soit juste. Mais dans ce cas, nous aurons théoriquement tous les $p(M_i) = 0$, puisque aucun modèle ne peut être exact (cf. Buslik [11]).
- $p(M_i)$ est la probabilité que le modèle M_i soit le meilleur parmi les modèles possibles. Mais en quel sens définir le terme «meilleur» ?
- $p(M_i)$ est la probabilité que le modèle M_i soit assez bon. Mais là encore, comment définir ceci ?

Apostolakis [6] et Laskey [32] optent pour la première définition, et propose d'estimer $p(M_i)$ par jugement d'expert. Cette approche, qui souffre de subjectivité, a l'avantage de ne nécessiter aucune donnée.
Lorsqu'au contraire on dispose de données, Bier [9] et Winkler proposent une méthode d'estimation de ces $p(M_i)$ basée sur la vraisemblance des modèles M_i. Il semble néanmoins que le coût d'estimation engendré soit alors important.
Clyde [12], Buckland *et al.* [10] ainsi que Kass et Raftery [30] s'intéressent quant à eux à une pondération utilisant les critères BIC (*Bayesian Information Criteria* [57]) et AIC (*Akaike Information Criteria* [2]). Le lecteur pourra se référer à [21] pour plus d'information.
L'ensemble de ces travaux ont été présentés au *Workshop* intitulé *Model Uncertainty : its Characterization and Quantification*, préalablement cité.
À noter la remarque de Abramson [1] qui met en garde sur l'utilisation d'une moyenne pondérée des différents modèles possibles, puisqu'elle n'a pas de sens physique, et peut masquer certaines aspérités des modèles.

En parallèle à ce *Workshop*, une approche bayésienne du problème des modèles concurrents émerge, initiée notamment par Raftery [25], sous le nom de *Bayesian Model Averaging*.

Bayesian Model Averaging

En adoptant un formalisme bayésien, l'objectif d'un modèle est de déterminer la probabilité que la quantité d'intérêt Y prenne une valeur y, conditionnellement à une certaine réalisation $\mathbf{x}_0$ des variables d'entrée $\mathbf{X}$. On note cette probabilité $p(Y = y|\mathbf{X} = \mathbf{x}_0)$ ou encore $p(Y = y|\mathcal{D})$, où $\mathcal{D}$ représente les «données» d'entrée $\mathbf{X} = \mathbf{x}_0$.
Lorsqu'un nombre fini k de modèles sont candidats, la méthode *Bayesian Model Averaging* consiste, sur le même principe que précédemment, à utiliser une pondération de tous les modèles possibles :

$$p(Y = y|\mathcal{D}) = \sum_{i=1}^{k} p(Y|M_i, \mathcal{D})p(M_i|\mathcal{D}), \tag{1.26}$$

où

$$p(M_i|\mathcal{D}) = \frac{p(\mathcal{D}|M_i)p(M_i)}{\sum_{j=1}^{k} p(\mathcal{D}|M_j)p(M_j)},$$

avec $p(M_j)$ la probabilité a priori que le modèle M_j soit vrai, et où

$$p(\mathcal{D}|M_i) = \int p(\mathcal{D}|\theta_i, M_i)p(\theta_i|M_i)d\theta_i, \tag{1.27}$$

avec θ_i les paramètres du modèle M_i, $p(\mathcal{D}|\theta_i, M_i)$ est la vraisemblance de M_i sur les données $\mathcal{D}$ et $p(\theta_i|M_i)$ la densité a priori de θ_i sous M_i.

Madigan et Raftery [33] montrent que l'utilisation de cette moyenne de tous les modèles possibles donne de meilleurs prédictions que l'utilisation de n'importe quel modèle seul (appartenant à la classe de modèle considérée).
Néanmoins, un certain nombre de difficultés existent :

- le nombre de termes à évaluer peut parfois devenir trop important,
- les intégrales ne sont pas calculables,
- comment évaluer $p(M_i)$, probabilité a priori que le modèle M_j soit vrai ?

Pour diminuer le nombre de terme à évaluer, la méthode *Occam's window* supprime dans l'ensemble des modèles possibles ceux qui sont trop mauvais par rapport au meilleur modèle, au sens de la probabilité $p(M_i|\mathcal{D})$, ainsi que ceux dont on dispose d'une version simplifiée tout aussi bonne (toujours au sens de $p(M_i|\mathcal{D})$).
Madigan et Raftery montrent comment utiliser les méthodes de Monte Carlo par chaînes de Markov pour l'estimation des intégrales implicites dans (1.26). D'autres solutions particulières ont été proposés suivant la forme du modèle (régression linéaire, modèle linéaire généralisé, modèle de survie et modèle graphique).
L'évaluation des probabilités a priori $p(M_i)$ est le problème le plus difficile. On retrouve alors les mêmes solutions que précédemment (jugement d'expert, BIC, AIC, etc.). Madigan et Raftery proposent de les considérer toutes égales, ce qui à l'avantage d'être un choix totalement neutre. Ils explicitent aussi quelques solutions particulières pour les modèles de régression linéaire, les modèles de Cox et les modèles graphiques.

1.2.3. Modèle simplifié et modèle de référence

Enfin, le dernier aspect abordé dans la bibliographie au sujet de l'incertitude de modèle, est lorsque l'incertitude est due à l'utilisation d'un modèle simplifié, alors qu'un modèle de référence existe.
Concernant cette problématique, des solutions n'ont été apportées, à notre connaissance, que dans un contexte de fiabilité des structures mécaniques, par Pendola dans sa thèse [41], en s'appuyant sur les travaux de Ditlevsen et Arnbjerg-Nielsen [18]. Il étudie l'impact de l'utilisation d'un modèle simplifié en fiabilité. Un problème de fiabilité d'une structure mécanique consiste à évaluer la probabilité qu'une réalisation y d'une certaine quantité d'intérêt aléatoire Y appartienne à une région de défaillance $\mathcal{F}$. Cette quantité d'intérêt est calculée à partir d'un modèle :

$$Y = f(\mathbf{X}),$$

où $\mathbf{X} \in \mathbb{R}^p$ est un ensemble de variables d'entrée (forces, moments, contraintes ...). Souvent, ce modèle est le résultat d'un calcul par éléments finis relativement coûteux en temps de calcul. Un modèle simplifié est alors utilisé :

$$Y = f_s(\mathbf{X}),$$

pour l'évaluation de la probabilité de défaillance. Cette évaluation consiste à résoudre une certaine équation, appelée équation d'état limite.
Ditlevsen montre que l'équation d'état limite du modèle de référence, intervenant dans le calcul de la probabilité de défaillance, est équivalente à celle du modèle simplifié à une correction près des variables d'entrée. Cette correction est fonction du rapport entre les solutions de l'équation d'état limite de référence et de l'équation d'état limite simplifiée. Considérant le modèle simplifié comme relativement proche du modèle de référence, la méthode du facteur de correction de modèle (*FCM*) consiste alors à supposer que dans un voisinage du point de conception x^* (estimé par exemple par une méthode de calcul de fiabilité de type *FORM, First Order Reliability Method*, ou *SORM, Second Order Reliability Method* [16]), le rapport des solutions des deux état limites peut être localement remplacé par une constante.
Cette méthode *FCM* n'est pas applicable à notre contexte d'analyse de sensibilité, puisqu'elle nécessite des approximations locales, et que l'analyse de sensibilité *globale* s'intéresse à la variabilité des variables du modèle sur l'ensemble de leur domaine de variation.
Dans un contexte fiabiliste, Pendola montre que cette méthode atteint rapidement ses limites lorsque le modèle de référence est lourd en temps de calcul (le calcul du point x^* nécessitant plusieurs appels à ce dernier).

De plus, elle suppose que le modèle de référence et le modèle simplifié sont fonction des mêmes variables. Pendola introduit alors la méthode d'approximation fonctionnelle de l'écart entre modèle (*AFE*).
Soit le modèle simplifié :

$$Y_s = f_s(\mathbf{X}_s),$$

avec $\mathbf{X}_s \in \mathbb{R}^q$, où cette fois le nombre de variables q peut être plus petit que celui du modèle de référence p.
La méthode *AFE* consiste à définir un ensemble de variables $\mathbf{X}_\Delta$ destinées à modéliser l'écart entre le modèle de référence et le modèle simplifié. Cet ensemble sera constitué au minimum des variables de $\mathbf{X}_s$ et au maximum de celle de $\mathbf{X}$:

$$\mathbf{X}_s \subseteq \mathbf{X}_\Delta \subseteq \mathbf{X}.$$

La nature physique des variables peut être utilisée pour les sélectionner.
La fonction écart est alors définie comme l'écart relatif entre le modèle simplifié et le modèle de référence :

$$\Delta(\mathbf{X}) = \frac{f_s(\mathbf{X}_s) - f(\mathbf{X})}{f(\mathbf{X})}.$$

Cette fonction écart est ensuite modélisée par une surface de réponse afin de pouvoir être utilisée dans les calculs fiabilistes. Pour ceci, une base de données de réalisations de Δ est construite à partir d'un échantillon de réalisations du modèle de référence et du modèle simplifié. Une surface de réponse $\hat{\Delta}(\mathbf{X}_\Delta)$ fonction uniquement des variables $\mathbf{X}_\Delta$ précédemment sélectionnées, est ajustée sur cette base de données. On considère alors que $\hat{\Delta}(\mathbf{X}_\Delta)$ est une bonne approximation de la fonction écart :

$$\hat{\Delta}(\mathbf{X}_\Delta) \simeq \Delta(\mathbf{X}),$$

avec $\mathbf{X} = \{\mathbf{X}_\Delta, \mathbf{X}_{\sim\Delta}\}$ où $\mathbf{X}_{\sim\Delta}$ sont les variables de $\mathbf{X}$ n'appartenant pas à $\mathbf{X}_\Delta$.
Pendola utilise comme surface de réponse des polynômes du second degré.
À partir de cette approximation de la fonction écart, pour les différents calculs fiabilistes, Pendola montre qu'il est intéressant d'apporter une correction au modèle simplifié sous la forme suivante :

$$\frac{f_s(\mathbf{X}_s)}{1 + \hat{\Delta}(\mathbf{X}_\Delta)}.$$

1.2.4. Conclusion

L'incertitude de modèle a été abordée sous trois aspects. Le premier consiste à déterminer les différentes sources d'incertitude existantes. Ce sujet a été étudié à de nombreuses reprises, ce qui a permis de définir une liste des sources d'incertitudes. Le deuxième aspect traite des modèles concurrents. Les solutions proposées dans la littérature consistent à utiliser une moyenne pondérée des différents modèles possibles, l'enjeu étant alors de définir les pondérations. Enfin, le dernier aspect propose une solution au problème de l'utilisation d'un modèle simplifié alors qu'un modèle de référence existe.
Ce dernier aspect est particulièrement intéressant vis-à-vis de notre problématique, mais la solution proposée par Pendola est spécifiquement adaptée au contexte de fiabilité des structures mécaniques. Nous proposons alors une démarche relativement équivalente, mais adaptée au contexte de l'analyse de sensibilité (cf. section 2.2).
L'incertitude due à un changement du modèle étudié, qui constitue notre seconde problématique, n'est pas abordée dans la littérature. Nous apportons des éléments de réponse dans le second chapitre de ce mémoire.

Chapitre 2

Analyse de sensibilité et incertitude de modèle

Ce chapitre traite de l'impact sur une étude de sensibilité de deux sources d'incertitude : l'incertitude due à une mutation dans le modèle étudié, et celle due à l'utilisation d'un modèle simplifié, alors qu'un modèle de référence existe. Ces deux sources d'incertitude, décrites en introduction de cette thèse, sont traitées dans les deux premières sections de ce chapitre.
La première section montre, pour un certain nombre de mutations types, comment les indices de sensibilité du modèle après mutation peuvent être déduits de ceux du modèle initial.
La deuxième section donne des éléments de réponse au problème de l'impact de l'utilisation d'un modèle simplifié, en utilisant notamment les résultats issus de la section précédente sur les mutations de modèle.
Enfin, une application sur le logiciel GASCON illustre ces travaux dans une troisième section.

2.1. Mutations de modèles et analyse de sensibilité

Considérons un modèle mathématique de référence M_1 caractérisé par :

$$Y = f_1(\mathbf{X}),$$

où Y est la variable aléatoire de sortie, $\mathbf{X} = (X_1, ..., X_p)$ est le vecteur des variables aléatoires d'entrée indépendantes, et f_1 la fonction déterministe qui associe Y à $\mathbf{X}$. Nous supposons qu'une analyse de sensibilité a été réalisée sur ce modèle M_1. De nouvelles informations (nouvelles données ou informations complémentaires) nous conduisent à modifier le modèle M_1, c'est-à-dire à le muter en un nouveau modèle M_2, caractérisé par :

$$Y = f_2(\mathbf{X}, \mathbf{Z}),$$

où $\mathbf{Z}$ est un vecteur d'éventuelles variables d'entrée complémentaires (corrélées ou non avec les composantes de $\mathbf{X}$).
La mutation est entiérement déterministe et opère sur la fonction déterministe du modèle.
Nous distinguons quatre types de mutations.

Le premier type de mutation consiste à ajouter ou à supprimer des variables aléatoires du modèle. Cette mutation est rencontrée en pratique lorsque l'on décide de considérer une variable aléatoire comme déterministe, ou inversement. Citons quelques applications éventuelles :

- Un doute quant à l'estimation d'un paramètre apparaît. Nous décidons alors de le considérer comme aléatoire, plutôt que de lui donner une valeur déterministe fausse.
- Le cas contraire, où une nouvelle information nous permet de donner une estimation déterministe cohérente à une variable que l'on considérait aléatoire.
- La sensibilité est due essentiellement à une variable (95% par exemple). Les valeurs des indices de

sensibilité relatifs aux autres variables sont très petites, ce qui ne nous permet pas, compte-tenu des incertitudes d'estimation, de distinguer une hiérarchie au sein de ces dernières. Une méthode envisageable est alors d'étudier la sensibilité du modèle en fixant la valeur de la variable la plus importante.

Le deuxième type de mutation consiste à ajouter une certaine fonction g au modèle initial. Le modèle muté peut alors s'écrire $Y = f_2(\mathbf{X}, \mathbf{Z}) = f_1(\mathbf{X}) + g(\mathbf{X}, \mathbf{Z})$. Ce type de mutation peut être rencontré sous une forme différente : deux analyses de sensibilité ont été réalisées sur deux modèles, et l'intérêt est finalement porté sur la somme de ces deux modèles. La situation inverse peut aussi être envisagée : si un modèle se sépare additivement en deux sous modèles, il peut être plus simple de faire l'analyse séparément sur ces sous modèles. Enfin, une dernière application peut être celle où deux modèles sont en concurrence pour simuler un phénomène et que, ne sachant lequel choisir, on décide d'utiliser une moyenne ou une combinaison linéaire de ces deux modèles.

Le troisième type de mutation est analogue au deuxième, en considérant une multiplication à la place d'une addition. Le modèle muté s'écrit $Y = f_2(\mathbf{X}, \mathbf{Z}) = f_1(\mathbf{X}) \times g(\mathbf{X}, \mathbf{Z})$. Afin d'illustrer cette mutation par une application possible, considérons le logiciel GASCON, présenté en section 2.3, qui étudie l'impact sur l'homme d'un rejet gazeux chronique d'une installation nucléaire, à partir notamment d'une quantité de rejet unitaire. L'impact sur l'homme est linéaire en fonction de la quantité de rejet. Supposons que l'on dispose d'un autre code de calcul permettant pour une installation donnée de définir la quantité de rejet. Il serait alors intéressant d'étudier la sensibilité du produit de ces deux codes, qui modéliserait le rejet de l'installation à l'homme.

Le dernier type de mutation à considérer est la composition du modèle par une certaine fonction. Le modèle muté, qui s'écrit $Y = f_2(\mathbf{X}, \mathbf{Z}) = g(f_1(\mathbf{X}), \mathbf{Z})$, semble avoir moins d'applications pratiques, mais complète la liste des mutations envisagées.

L'approche que nous proposons consiste à étudier, pour chaque mutation, les relations formelles entre les indices de sensibilité avant mutation et après mutation. Nous examinerons tout d'abord le premier type de mutation, qui intervient au niveau des variables d'entrées du modèle : introduire une nouvelle variable (transformer un paramètre déterministe en une variable aléatoire), fixer une variable aléatoire afin de la rendre déterministe ou encore transformer une variable (changement de loi, d'échelle...). Nous nous intéresserons ensuite aux trois autres types de mutations : après avoir constaté que le cas général de la composition par une fonction n'était pas exploitable formellement, nous nous concentrons sur l'addition et la multiplication d'une fonction au modèle initial, que nous interpréterons pratiquement comme l'ajout ou la multiplication de deux modèles.

2.1.1. Mutations des variables d'entrée du modèle

2.1.1.1. Transformation d'une variable d'entrée

Supposons qu'une des variables d'entrée du modèle soit transformée. Par exemple, on décide de changer la nature ou les paramètres de la loi de cette variable.
Le modèle $Y = f(X_1, \ldots, X_i, \ldots, X_p)$ devient $Y' = f(X_1, \ldots, h(X_i), \ldots, X_p)$. Nous supposons les variables X_i de loi μ_i à densité ρ_i sur $\mathbb{R}$ (pour $i = 1, \ldots, p$). Etant donné la définition des indices de sensibilité, étudier les relations entre indices avant mutation et après mutation nécessite d'étudier les espérances conditionnelles de Y et Y'. Nous étudions dans un premier temps le cas des indices de premier ordre.

Indices de premier ordre S_j avec $j \neq i$
L'espérance de $Y^{'}$ conditionnellement à X_j s'écrit :

$$\begin{aligned} \mathrm{E}[Y^{'}|X_j] &= \int_{\mathbb{R}^p} f(x_1,...,X_j,...,h(x_i),...,x_p)d\mu_1...d\mu_i...d\mu_p \\ &= \int_{\mathbb{R}^p} f(x_1,...,X_j,...,h(x_i),...,x_p)\rho_1(x_1)...\rho_i(x_i)...\rho_p(x_p)dx_1...dx_i...dx_p, \end{aligned}$$

d'où par changement de variable :

$$\mathrm{E}[Y^{'}|X_j] = \int_{\mathbb{R}^p} f(x_1,...,X_j,...,x_i,...,x_p)\rho_1(x_1)...\rho_i(h^{-1}(x_i))(h^{-1})^{'}(x_i)...\rho_p(x_p)dx_1...dx_i...dx_p.$$

Cette espérance peut être mise en relation avec $\mathrm{E}[Y|X_j]$, définie par :

$$\mathrm{E}[Y|X_j] = \int_{\mathbb{R}^p} f(x_1,...,X_j,...,x_i,...,x_p)\rho_1(x_1)...\rho_i(x_i)...\rho_p(x_p)dx_1...dx_i...dx_p,$$

uniquement sous la condition suivante :

$$\rho_i(h^{-1}(x_i))(h^{-1})^{'}(x_i) \propto \rho_i(x_i), \tag{2.1}$$

ou autrement dit si le rapport des deux est constant.

– En se plaçant dans un cadre simple (variables gaussiennes, transformation affine $h(t) = at + b$), est-il possible de vérifier cette relation (2.1) ? La densité d'une variable gaussienne de moyenne m et de variance σ^2 ($\rho_i(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-m)^2}{2\sigma^2}}$) n'étant jamais nulle, nous pouvons étudier le rapport des deux parties de (2.1), étant donné que $(h^{-1})^{'}(t) = \frac{1}{a}$ est constant :

$$\frac{\rho_i(h^{-1}(x_i))}{\rho_i(x_i)} = e^{\frac{(x_i-m)^2-(h^{-1}(x_i)-m)^2}{2\sigma^2}} = e^{\frac{(x_i+h^{-1}(x_i)-2m)(x_i-h^{-1}(x_i))}{2\sigma^2}},$$

qui n'est constant que pour $x_i = h^{-1}(x_i)$ (transformation identité $h(t) = t$), ou pour $x_i = -h^{-1}(x_i) + 2m$ (transformation $h(t) = -t + 2m$). Sous ces conditions, $\rho_i(h^{-1}(x_i))(h^{-1})^{'}(x_i) = \rho_i(x_i)$, et donc $\mathrm{E}[Y^{'}|X_j] = \mathrm{E}[Y|X_j]$, les indices de sensibilité du modèle après mutation peuvent donc être obtenus à partir de ceux du modèle initial par :

$$S_j^{'} = \frac{\mathrm{V}(\mathrm{E}[Y^{'}|X_j])}{\mathrm{V}(Y^{'})} = \frac{\mathrm{V}(\mathrm{E}[Y|X_j])}{\mathrm{V}(Y)} \frac{\mathrm{V}(Y)}{\mathrm{V}(Y^{'})} = S_j \times \frac{\mathrm{V}(Y)}{\mathrm{V}(Y^{'})}. \tag{2.2}$$

Cette relation est aussi valable pour la transformation $h(t) = -t + 2m$ avec des variables d'entrée de densité de probabilité symétriques par rapport à m (Student avec $m = 0$, Cauchy avec $m = 0$, uniforme $\mathcal{U}[a, b]$ avec $m = \frac{a+b}{2}$). Seulement, d'un point de vue pratique, cette transformation n'a pas d'application immédiate, et risque de ne pas être rencontrée souvent.

– Considérons maintenant le cas particulier où la variable X_i apparaît de façon séparable dans le modèle, c'est-à-dire que le modèle s'écrit :

$$f(X_1,...,X_p) = f_1(X_i) + f_2(X_1,...,X_{i-1},X_{i+1},...,X_p).$$

On montre facilement que l'on a alors $\mathrm{E}[Y^{'}|X_j] = \mathrm{E}[Y|X_j] + (\mathrm{E}[f_1(X_i)] - \mathrm{E}[f_1(h(X_i))])$, pour

tout $j \neq i$, d'où $\mathrm{V}(\mathrm{E}[Y^{'}|X_j]) = \mathrm{V}(\mathrm{E}[Y|X_j])$. La variable X_i n'a pas d'influence sur les variances conditionnelles. Il est alors possible d'obtenir les indices de sensibilité du modèle après mutation à partir de ceux du modèle initial par (2.2), et ce sans hypothèse sur la transformation h ni sur la loi des variables d'entrée.

Indices de premier ordre S_i
Dans ce cas :

$$\mathrm{E}[Y^{'}|X_i = x_i] = \mathrm{E}[Y|X_i = h(x_i)],$$

qui sont des fonctions de x_i. Soit $e(t) = \mathrm{E}[Y|X_i = t]$. Pour exprimer les indices de sensibilité du modèle après mutation à partir de ceux du modèle initial, il est nécessaire de pouvoir exprimer formellement $\mathrm{V}(e(h(X_i)))$ en fonction de $\mathrm{V}(e(X_i))$. Les seuls cas de figure adéquats sont ceux où les fonctions e et h sont affines, à cause de la linéarité de l'intégrale.

– Supposons donc que le modèle s'écrive :

$$f(X_1, ..., X_p) = f_1(X_i) f_2(X_1, ..., X_{i-1}, X_{i+1}, ..., X_p) + f_3(X_1, ..., X_{i-1}, X_{i+1}, ..., X_p).$$

Dans ce cas, $e(t) = \mathrm{E}[Y|X_i = t] = at + b$, où $a = \mathrm{E}[f_2(X_1, ..., X_p)]$ et $b = \mathrm{E}[f_3(X_1, ..., X_p])$. Si on suppose que $h(t) = ct + d$, on a alors $\mathrm{V}(e(h(X_i))) = a^2c^2\mathrm{V}(X_i) = c^2\mathrm{V}(e(X_i))$, et donc :

$$S_i^{'} = S_i \times c^2 \frac{\mathrm{V}(Y)}{\mathrm{V}(Y^{'})}.$$

Indices d'ordre supérieur
Les indices d'ordre supérieur et totaux se comportent de manière identique aux indices de premier ordre, selon qu'ils expriment une sensibilité à la variable X_i ou non.

2.1.1.2. Une variable aléatoire d'entrée devient déterministe

Nous supposons qu'une des variables d'entrée du modèle initial $Y = f(X_1, \ldots, X_i, \ldots, X_p)$ soit finalement considérée comme déterministe. Le modèle devient alors $Y^{'} = f(X_1, \ldots, \alpha, \ldots, X_p)$. Définir une relation entre les indices de sensibilité du modèle avant mutation et ceux après mutation nécessite des hypothèse sur la forme de f, concernant la séparabilité linéaire de la variable X_i par rapport aux autres variables.

Cas où le modèle est $Y = \sum_{k=1}^{K} f_k^1(X_i) f_k^2(\boldsymbol{X}_{\sim i})$
C'est le modèle le plus général que nous présenterons. Les deux modèles suivants sont des cas particuliers de celui-ci. Les indices de sensibilité relatifs à la variable X_i n'existent plus dans le nouveau modèle $Y^{'} = \sum_{k=1}^{K} f_k^1(\alpha) f_k^2(\boldsymbol{X}_{\sim i})$.
Indices de premier ordre S_j
En écrivant l'espérance de Y et celle de $Y^{'}$ conditionnellement à X_j :

$$\begin{aligned} \mathrm{E}[Y|X_j] &= \sum_{k=1}^{K} \mathrm{E}[f_k^1(X_i)]\mathrm{E}[f_k^2(\boldsymbol{X}_{\sim i})|X_j], \\ \mathrm{E}[Y^{'}|X_j] &= \sum_{k=1}^{K} \mathrm{E}[f_k^1(\alpha)]\mathrm{E}[f_k^2(\boldsymbol{X}_{\sim i})|X_j], \end{aligned}$$

on obtient la relation suivante entre ces espérances conditionnelles :

$$\mathrm{E}[Y^{'}|X_j] = \mathrm{E}[Y|X_j] + \sum_{k=1}^{K}(\mathrm{E}[f_k^1(\alpha)] - \mathrm{E}[f_k^1(X_i)])\mathrm{E}[f_k^2(\boldsymbol{X}_{\sim i})|X_j].$$

– Le passage à la variance donne une relation entre les indices de sensibilité avant et après mutation, sous les conditions $\mathrm{E}[f_k^1(\alpha)] = \mathrm{E}[f_k^1(X_i)], \quad \forall k \in \{1,..,K\}$, c'est-à-dire lorsque :

$$\begin{cases} f_k^1 \text{ affine } : t \mapsto at + b \\ \mathrm{E}[X_i] = \alpha \end{cases} \tag{2.3}$$

Ainsi, sous ces hypothèses, les indices de sensibilité du modèle après mutation peuvent être obtenus à partir de ceux du modèle initial, par une simple remise à l'échelle, en les multipliant par le rapport des deux variances :

$$S_j^{'} = S_j \times \frac{\mathrm{V}(Y)}{\mathrm{V}(Y^{'})} \qquad \forall j \in \{1,\ldots,p\} \setminus \{i\}.$$

Indices d'ordre supérieur $S_{i_1,\ldots,i_s}$

Sous les mêmes conditions (2.3), il est possible, pour les mêmes raisons, d'obtenir les indices de sensibilité à tout ordre de $Y^{'}$ en fonction de ceux de Y par :

$$S_{i_1,\ldots,i_s}^{'} = S_{i_1,\ldots,i_s} \times \frac{\mathrm{V}(Y)}{\mathrm{V}(Y^{'})}, \qquad \forall \{i_1,...,i_s\} \subset \{1,\ldots,p\} \setminus \{i\}.$$

Indices totaux S_{T_j}

Nous sommes confrontés à deux cas de figure : soit tous les indices de sensibilité de Y ont été estimés, soit seuls les indices totaux l'ont été. Dans le second cas, nous ne pouvons tirer aucune information des indices de sensibilité totaux de Y pour estimer ceux de $Y^{'}$, puisqu'il est impossible d'extraire des indices totaux de Y la part due aux interactions avec la variable X_i. Par contre, si on est dans le premier cas, nous pouvons alors, sous les conditions (2.3), estimer les indices totaux par $S_{T_j}^{'} = \sum_{\#j} S_r^{'} = (\sum_{\#j} S_r) \times \frac{\mathrm{V}(Y)}{\mathrm{V}(Y^{'})}$, où $\sum_{\#j} S_r$ est la somme de tous les indices de sensibilité ayant j comme indice. En effet, sachant exprimer les indices de sensibilité de $Y^{'}$ à tout ordre en fonction de ceux de Y, il est possible d'en faire autant pour leur somme.

Cas où le modèle est $Y = f_1(X_i) + f_2(\boldsymbol{X}_{\sim i})$

Ce modèle est un cas particulier du précédent qui permet d'omettre les conditions (2.3). Etant donné les espérance conditionnelle $\mathrm{E}[Y|X_j] = \mathrm{E}[f_1(X_i)] + \mathrm{E}[f_2(\boldsymbol{X}_{\sim i})|X_j]$ et $\mathrm{E}[Y^{'}|X_j] = f_1(\alpha) + \mathrm{E}[f_2(\boldsymbol{X}_{\sim i})|X_j]$, on a naturellement que $\mathrm{V}(\mathrm{E}[Y^{'}|X_j]) = \mathrm{V}(\mathrm{E}[Y|X_j])$.
Ainsi, les indices de sensibilité de premier ordre S_j et d'ordre supérieur $S_{i_1,\ldots,i_s}$ s'obtiennent par

$$\begin{aligned} S_j^{'} &= S_j \times \frac{\mathrm{V}(Y)}{\mathrm{V}(Y^{'})}, \quad \text{pour tout } j \in \{1,...,p\} \setminus \{i\}, \\ S_{i_1,\ldots,i_s}^{'} &= S_{i_1,\ldots,i_s} \times \frac{\mathrm{V}(Y)}{\mathrm{V}(Y^{'})}, \quad \text{pour tout } \{i_1,...,i_s\} \subset \{1,...,p\} \setminus \{i\}. \end{aligned}$$

Comme la variable X_i est indépendante des autres variables, et séparée additivement de ces dernières, les indices d'interaction au sein du modèle initial entre X_i et d'autre variables sont tous nuls. En effet, les indices d'interaction $S_{i,j}$, pour tout $j \in \{1,\ldots,p\} \setminus \{i\}$, égaux à :

$$S_{i,j} = \frac{\mathrm{V}(\mathrm{E}[Y|X_i,X_j] - \mathrm{E}[Y|X_i] - \mathrm{E}[Y|X_j])}{\mathrm{V}(Y)}$$

sont nuls puisque le terme :

$$\begin{aligned} \mathrm{E}[Y|X_i,X_j]-\mathrm{E}[Y|X_i]-\mathrm{E}[Y|X_j] &= \mathrm{E}[f_1(X_i)|X_i]+\mathrm{E}[f_2(\boldsymbol{X}_{\sim i})|X_j] \\ &\quad -\mathrm{E}[f_1(X_i)]-\mathrm{E}[f_2(\boldsymbol{X}_{\sim i})|X_j] \\ &\quad -\mathrm{E}[f_1(X_i)|X_i]-\mathrm{E}[f_2(\boldsymbol{X}_{\sim i})] \\ &= -\mathrm{E}[f_1(X_i)]-\mathrm{E}[f_2(\boldsymbol{X}_{\sim i})], \end{aligned} \tag{2.4}$$

est constant donc de variance nulle.
Ainsi, il est possible d'obtenir les indices de sensibilité totaux du modèle après mutation par :

$$S'_{T_j} = S_{T_j} \times \frac{\mathrm{V}(Y)}{\mathrm{V}(Y')} \qquad \forall j \in \{1,\dots,p\} \setminus \{i\}.$$

Cas où le modèle est $Y = f_1(X_i) f_2(\boldsymbol{X}_{\sim i})$
Ce second cas particulier permet aussi d'omettre les conditions (2.3).
Comme $\mathrm{E}[Y|X_j] = \mathrm{E}[f_1(X_i)]\mathrm{E}[f_2(\boldsymbol{X}_{\sim i})|X_j]$ et $\mathrm{E}[Y'|X_j] = f_1(\alpha)\mathrm{E}[f_2(\boldsymbol{X}_{\sim i})|X_j]$, on obtient après passage à la variance $\mathrm{V}(\mathrm{E}[Y'|X_j]) = \frac{f_1(\alpha)^2}{\mathrm{E}[f_1(X_i)]^2}\mathrm{V}(\mathrm{E}[Y|X_j])$. Ainsi les indices de sensibilité de premier ordre et d'ordre supérieur du modèle après mutation Y' s'obtiennent à partir de ceux de Y par :

$$\begin{aligned} S'_j &= S_j \times \frac{f_1(\alpha)^2}{\mathrm{E}[f_1(X_i)]^2}\frac{\mathrm{V}(Y)}{\mathrm{V}(Y')} \qquad \forall j \in \{1,\dots,p\} \setminus \{i\}, \\ S'_{i_1,\dots,i_s} &= S_{i_1,\dots,i_s} \times \frac{f_1(\alpha)^2}{\mathrm{E}[f_1(X_i)]^2}\frac{\mathrm{V}(Y)}{\mathrm{V}(Y')} \quad \text{pour tout } \{i_1,...,i_s\} \subset \{1,\dots,p\} \setminus \{i\}. \end{aligned}$$

Les indices de sensibilité totaux s'obtiennent par $S'_{T_j} = (\sum_{\#j} S_r) \times \frac{f_1(\alpha)^2}{\mathrm{E}[f_1(X_i)]^2}\frac{\mathrm{V}(Y)}{\mathrm{V}(Y')}$, uniquement si les indices de sensibilité de Y sont connus à tout ordre.

2.1.1.3. Un paramètre déterministe devient une variable aléatoire

Nous considérons cette fois la mutation inverse de la précédente.
La mutation consiste à considérer un paramètre déterministe α du modèle comme aléatoire, et à le représenter par une nouvelle variable X_{p+1}, que l'on suppose indépendante des autres. Le modèle devient alors $Y' = f(X_1,..,X_p,X_{p+1})$.
Comme pour la mutation inverse, nous sommes dans l'obligation de faire des hypothèses sur la forme du modèle, afin de pouvoir exprimer les indices de sensibilité après mutation en fonction de ceux avant mutation. Naturellement, il est nécessaire d'estimer tous les indices relatifs à la nouvelle variable, ainsi que tous les indices totaux S_{T_1} à S_{T_p}, afin qu'ils prennent en compte les interactions éventuelles avec la nouvelle variable.

Cas où le modèle est $Y = \sum_{k=1}^{K} f_k^1(\alpha) f_k^2(X_1, ..., X_p)$

Indices de premier ordre S_j

En écrivant les espérance de Y et de $Y^{'}$ conditionnellement à X_j :

$$\begin{aligned} \mathrm{E}[Y|X_j] &= \sum_{k=1}^{p} f_k^1(\alpha)\mathrm{E}[f_k^2(X_1, ..., X_p)|X_j], \\ \mathrm{E}[Y^{'}|X_j] &= \sum_{k=1}^{p} \mathrm{E}[f_k^1(X_{p+1})]\mathrm{E}[f_k^2(X_1, ..., X_p)|X_j], \end{aligned}$$

on constate que $\mathrm{E}[Y^{'}|X_j] = \mathrm{E}[Y|X_j] + \sum_{k=1}^{K}(\mathrm{E}[f_k^1(X_{p+1}) - f_k^1(\alpha)])\mathrm{E}[f_k^2(X_1, ..., X_p)|X_j]$. Pour passer à la variance et exprimer les indices de sensibilité de $Y^{'}$ en fonction de ceux de Y, il est intéressant que $\mathrm{E}[f_k^1(X_{p+1})] = f_k^1(\alpha)$ pour tout $k \in \{1, \ldots, K\}$.

– Ainsi, sous les conditions suivantes :

$$\begin{cases} f_k^1 \text{ affine } : t \mapsto at + b \\ \mathrm{E}[X_{p+1}] = \alpha, \end{cases} \tag{2.5}$$

les indices de sensibilité de premier ordre de $Y^{'}$, ainsi que ceux d'ordre supérieur (non relatif aux interactions avec la nouvelle variable X_{p+1}), s'obtiennent par :

$$\begin{aligned} S_j^{'} &= S_j \times \frac{\mathrm{V}(Y)}{\mathrm{V}(Y^{'})} \qquad \forall j \in \{1, \ldots, p\}, \\ S_{i_1,...,i_s}^{'} &= S_{i_1,...,i_s} \times \frac{\mathrm{V}(Y)}{\mathrm{V}(Y^{'})}, \qquad \forall \{i_1, ..., i_s\} \subset \{1, \ldots, p\}. \end{aligned}$$

L'indice de sensibilité à la nouvelle variable X_{n+1} est défini par :

$$S_{p+1}^{'} = \sum_{k=1}^{K} \mathrm{E}[f_k^2(X_1, ..., X_p)]^2 \times \frac{\mathrm{V}(f_k^1(X_{p+1}))}{\mathrm{V}(Y^{'})}.$$

Cette définition n'est pas exploitable en pratique, puisqu'il serait nécessaire d'évaluer séparément $\mathrm{E}[f_k^2(X_1, ..., X_p)]^2$ pour chaque k, ce qui, d'un point de vue calcul, est aussi coûteux qu'une nouvelle analyse. Il est donc préférable d'estimer $S_{p+1}^{'}$ par une méthode classique (Sobol ou autre) comme celle utilisée pour l'analyse sur Y. De même pour les indices d'interaction avec la nouvelle variable X_{p+1}, puisque pour tout $1 \leq j \leq p$:

$$\mathrm{E}[Y^{'}|X_j, X_{p+1}] = \sum_{k}^{p} f_k^1(X_{p+1})\mathrm{E}[f_k^2(X_1, ..., X_p)|X_j],$$

ce qui n'est pas exploitable de façon bénéfique en pratique.

Cas où le modèle est $Y = f_1(\alpha) + f_2(X_1, \ldots, X_p)$ Ce cas particulier du précédent ne demande pas la vérification des hypothèses (2.5). En effet, les espérances de Y et de $Y^{'}$ conditionnellement à X_j :

$$\mathrm{E}[Y|X_j] = f^1(\alpha) + \mathrm{E}[f^2(X_1, ..., X_p)|X_j]$$

et

$$\mathrm{E}[Y^{'}|X_j] = \mathrm{E}[f^1(X_{n+1})] + \mathrm{E}[f_k^2(X_1, ..., X_p)|X_j]$$

satisfont la relation $\mathrm{E}[Y^{'}|X_j] = \mathrm{E}[Y|X_j] - f^1(\alpha) + \mathrm{E}[f^1(X_{p+1})]$.
Ainsi, le passage à la variance donne $\mathrm{V}(\mathrm{E}[Y^{'}|X_j]) = \mathrm{V}(\mathrm{E}[Y|X_j])$. Les indices de sensibilité de premier ordre et d'ordre supérieur (n'incluant pas les interactions avec X_{p+1}), peuvent alors être obtenus par :

$$\begin{aligned} S_j^{'} &= S_j \times \frac{\mathrm{V}(Y)}{\mathrm{V}(Y^{'})} \qquad \forall j \in \{1, \ldots, p\}, \\ S_{i_1,\ldots,i_s}^{'} &= S_{i_1,\ldots,i_s} \times \frac{\mathrm{V}(Y)}{\mathrm{V}(Y^{'})} \qquad \forall \{i_1, ..., i_s\} \subset \{1, \ldots, p\}. \end{aligned}$$

Il est important de noter que tous les indices d'interaction entre les variables initiales $X_1, ..., X_p$ et la nouvelle variable X_{p+1} sont nuls, puisque ces variables sont indépendantes et apparaîssent additivement dans le modèle (se reporter à (2.4)).
Introduction d'un bruit additif. L'introduction dans le modèle d'un bruit aléatoire additif est une application de ce type de mutation, pour lequel nous pourrons déterminer aussi l'indice de sensibilité relatif à la nouvelle variable représentant ce bruit. Soit $Y^{'} = \epsilon + f(X_1, \ldots, X_p)$ le modèle après mutation (introduction d'un bruit ϵ), où ϵ est supposé gaussien $\mathcal{N}(0, \sigma^2)$. Outre les résultats précédents concernant les indices de sensibilité de $Y^{'}$ relatifs aux variables $(X_1, \ldots, X_p)$, il est aussi possible de déterminer l'indice de sensibilité relatif à ce bruit, qui est $S_\epsilon^{'} = \frac{\sigma^2}{\mathrm{V}(Y^{'})}$.

Si le modèle est $Y = f_1(\alpha) f_2(X_1, \ldots, X_p)$:
Ce cas particulier permet lui aussi d'omettre les hypothèses (2.5). Les espérances de Y et de $Y^{'}$ conditionnellement à X_j, égales respectivement à

$$\mathrm{E}[Y|X_j] = f^1(\alpha)\mathrm{E}[f^2(X_1, ..., X_p)|X_j],$$

et

$$\mathrm{E}[Y^{'}|X_j] = \mathrm{E}[f^1(X_{p+1})]\mathrm{E}[f_k^2(X_1, ..., X_p)|X_j],$$

sont proportionnelles :

$$\mathrm{E}[Y^{'}|X_j] = \mathrm{E}[Y|X_j]\frac{\mathrm{E}[f^1(X_{p+1})]}{f^1(\alpha)}.$$

On en déduit que les indices de sensibilité de premier ordre et d'ordre supérieur (n'incluant pas les interactions avec X_{p+1}), peuvent être obtenus par :

$$\begin{aligned} S_j^{'} &= S_j \times \frac{\mathrm{V}(Y)}{\mathrm{V}(Y^{'})} \times (\frac{\mathrm{E}[f^1(X_{p+1})]}{f^1(\alpha)})^2 \qquad \forall j \in \{1, \ldots, p\}, \\ S_{i_1,\ldots,i_s}^{'} &= S_{i_1,\ldots,i_s} \times \frac{\mathrm{V}(Y)}{\mathrm{V}(Y^{'})} \times (\frac{\mathrm{E}[f^1(X_{p+1})]}{f^1(\alpha)})^2, \qquad \forall \{i_1, ..., i_s\} \subset \{1, .., p\}. \end{aligned}$$

Introduction d'un bruit multiplicatif. L'introduction dans le modèle d'un bruit aléatoire multiplicatif est une application de ce type de mutation, pour laquelle il est possible d'obtenir aussi les indices d'interaction avec la nouvelle variable X_{p+1}. Supposons qu'un bruit $\epsilon \sim \mathcal{N}(1, \sigma^2)$ soit introduit multiplicativement dans le modèle, qui s'écrit alors $Y^{'} = \epsilon f(X_1, .., X_p)$. Les indices de sensibilité de ce nouveau modèle peuvent être déduits de ceux de Y en utilisant les résultats ci-dessus. Mais il est aussi possible, sans nouvelle analyse de sensibilité, d'obtenir les indices de sensibilité relatifs à la nouvelle variable qu'est ϵ.

L'indice de premier ordre relatif à la nouvelle variable ϵ est donné par :

$$S'_\epsilon = \mathrm{E}[Y]^2 \frac{\sigma^2}{\mathrm{V}(Y')},$$

puisque $\mathrm{V}(\mathrm{E}[Y'|\epsilon]) = \mathrm{V}(\epsilon \mathrm{E}[Y]) = \mathrm{E}[Y]^2 \mathrm{V}(\epsilon)$.
Quant aux indices d'interaction entre les variables $(X_1, .., X_p)$ et ϵ, un calcul élémentaire donne :

$$S'_{j,\epsilon} = \sigma^2 S_j \frac{\mathrm{V}(Y)}{\mathrm{V}(Y')} \qquad \forall j \in \{1, \ldots, p\}.$$

Ce résultat est aussi vrai pour les indices d'ordre supérieur à deux, et peut être démontré simplement en utlisant les deux propritétés suivantes.

Propriété 2.1.1. *Si (A, B) est un couple de variables indépendantes, alors :*

$$\begin{aligned} V(AB) &= E[V(AB|B)] + V(E[AB|B]) \\ &= E[B^2]V(A) + E[A]^2 V(B) \\ &= E[B]^2 V(A) + V(A)V(B) + E[A]^2 V(B), \end{aligned}$$

Propriété 2.1.2. *Si (A, B) est un couple de variables indépendantes, alors :*

$$Cov(A, AB) = E[B]V(A).$$

2.1.2. Composition de plusieurs analyses de sensibilité

Nous envisageons désormais des mutations que l'on peut qualifier d'externe au modèle. Nous considérerons le cas général d'un modèle $Y = f(X_1, .., X_p)$ muté en un certain modèle Y' défini par $Y' = g \circ f(X_1, .., X_p)$, puis les cas particuliers où cette composition est une multiplication ou une addition. Ces deux derniers cas peuvent être interprétés comme la multiplication ou l'addition de deux modèles, sur chacun desquels une analyse de sensibilité préalable a été réalisée.

Supposons donc que le modèle $Y = f(X_1, .., X_p)$ mute en $Y' = g \circ f(X_1, .., X_p)$. Comme nous l'avons déjà vu précédemment, relier formellement les indices de sensibilité de ces deux modèles revient à écrire une relation mathématique entre $\mathrm{E}[Y|X_i]$ et $\mathrm{E}[Y'|X_i]$, pour tout $1 \leq i \leq p$, et ce en utilisant les propriétés de linéarité du calcul intégral ($\mathrm{E}[aZ + b] = a\mathrm{E}[Z] + b$). Il apparaît donc indispensable pour relier les indices de Y et de Y', qu'ils satisfassent une relation linéaire, autrement dit que g soit affine ($g(t) = at + b$).
Cela revient alors à considérer les deux cas particuliers de la composition, que sont la multiplication par une fonction (ou multiplication de deux modèles), et l'addition d'une fonction à Y (ou addition de deux modèles).

2.1.2.1. Multiplication de deux modèles

Nous considérons deux modèles Y_1 et Y_2, sur lesquels deux analyses de sensibilité ont été réalisées. Les indices de sensibilité de ces derniers peuvent-ils être utilisés pour obtenir les indices de sensibilité du modèle produit $Y = Y_1 Y_2$? Nous différencions les cas où les variables des modèles Y_1 et Y_2 sont les mêmes, différentes, ou encore certaines sont en commun et d'autres non.
Nous notons S^1 et S^2 les indices de sensibilité de Y_1 et Y_2.

Modèles à variables différentes

Soient les modèles $Y_1 = f_1(X_1, ..., X_p)$ et $Y_2 = f_2(X_{p+1}, ..., X_{p+q})$. Comme les variables de chacun des

deux modèles sont différentes :

$$\begin{aligned}
\mathrm{E}[Y|X_j] &= \mathrm{E}[Y_1Y_2|X_j] \\
&= \mathrm{E}[f_1(X_1,...,X_p)f_2(X_{p+1},...,X_{p+q})|X_j] \\
&= \begin{cases} \mathrm{E}[Y_1|X_j] \times \mathrm{E}[Y_2] & \text{si} \quad 1 \leq j \leq p \\ \mathrm{E}[Y_1] \times \mathrm{E}[Y_2|X_j] & \text{si} \quad p+1 \leq j \leq p+q \end{cases}
\end{aligned}$$

Ainsi, les indices de sensibilité du modèle Y, de premier ordre et d'ordre supérieur (relatif à l'interaction de variable propre à Y_1 ou à Y_2), peuvent être obtenus par :

$$S_j = \begin{cases} \mathrm{E}[Y_2]^2 S_j^1 \times \frac{\mathrm{V}(Y_1)}{\mathrm{V}(Y)} & \text{si} \quad 1 \leq j \leq n \\ \mathrm{E}[Y_1]^2 S_j^2 \times \frac{\mathrm{V}(Y_2)}{\mathrm{V}(Y)} & \text{si} \quad n+1 \leq j \leq n+p \end{cases} \tag{2.6}$$

Les indices de sensibilité relatifs à l'interaction entre variables de Y_1 et variables de Y_2 sont quant à eux obtenus par :

$$S_{j,k} = \mathrm{E}[Y_2]^2 S_j^1 \frac{\mathrm{V}(Y_1)}{\mathrm{V}(Y)} + \mathrm{E}[Y_1]^2 S_k^2 \frac{\mathrm{V}(Y_2)}{\mathrm{V}(Y)} + S_j^1 S_k^2 \frac{\mathrm{V}(Y_1)\mathrm{V}(Y_2)}{\mathrm{V}(Y)^2}, \tag{2.7}$$

où $1 \leq j \leq p$ et $p+1 \leq k \leq p+q$.
En effet, en appliquant la propriété 2.1.1 :

$$\begin{aligned}
\mathrm{V}(\mathrm{E}[Y|X_j,X_k]) &= \mathrm{V}(\mathrm{E}[Y_1|X_j]\mathrm{E}[Y_2|X_k]) \\
&= \mathrm{E}[\mathrm{E}[Y_2|X_k]]^2\mathrm{V}(\mathrm{E}[Y_1|X_j]) + \mathrm{V}(\mathrm{E}[Y_1|X_j])\mathrm{V}(\mathrm{E}[Y_2|X_k]) \\
&\quad +\mathrm{E}[\mathrm{E}[Y_1|X_j]]^2\mathrm{V}(\mathrm{E}[Y_2|X_k]) \\
&= \mathrm{E}[Y_2]^2\mathrm{V}(\mathrm{E}[Y_1|X_j]) + \mathrm{V}(\mathrm{E}[Y_1|X_j])\mathrm{V}(\mathrm{E}[Y_2|X_k]) + \mathrm{E}[Y_1]^2\mathrm{V}(\mathrm{E}[Y_2|X_k])
\end{aligned}$$

Ce résultat est encore valable pour les indices de sensibilité liés à l'interaction entre un groupe de variable de Y_1 et un groupe de variable de Y_2. Il suffit alors de remplacer dans (2.7) l'indice j par les indices $j_1, \ldots, j_s$ relatifs au groupe de variables de Y_1, et k par $k_1, \ldots, k_{s'}$ relatifs au groupe de variables de Y_2.
Concernant les indices de sensibilité totaux, on est en présence de deux cas de figure. Soit tous les indices de sensibilité à tout ordre ont été estimés pour Y_1 et Y_2, auquel cas il est possible de tous les estimer à nouveau grâce à (2.6) et (2.7), et ainsi en les sommant de retrouver les indices de sensibilité totaux de Y. Par contre, si les indices de sensibilité totaux de Y_1 et Y_2 ont été estimés directement, il sera alors impossible d'obtenir une estimation de ceux de Y sans nouvelle analyse de sensibilité.

Modèles à variables identiques

Nous considérons désormais les deux modèles $Y_1 = f_1(X_1,...,X_p)$ et $Y_2 = f_2(X_1,...,X_p)$. Comme l'espérance d'un produit de variables ne peut pas être exprimée de façon simple en fonction des espérances de ces variables lorsque ces dernières ne sont pas indépendantes, nous ne pouvons obtenir les indices de sensibilité de Y à partir de ceux de Y_1 et de Y_2.
Citons tout de même un cas particulier lorsqu'un des deux modèles ne contient qu'une seule variable. Supposons par exemple que $Y = f_1(X_1,..,X_p) \times f_2(X_j)$ avec $1 \leq j \leq p$.
Comme $\mathrm{E}[f_1(X_1,..,X_p) \times f_2(X_j)|X_j] = \mathrm{E}[f_2(X_j)]\mathrm{E}[f_1(X_1,..,X_p)|X_j]$, il est possible d'obtenir l'indice de sensibilité relatif à la variable X_j, qui est :

$$S_j = \mathrm{E}[Y_2]^2 \times S_j^1 \frac{\mathrm{V}(Y_1)}{\mathrm{V}(Y)}.$$

Modèles à variables identiques et différentes

Nous sommes dans le cas où $Y_1 = f_1(X_1, .., X_p)$ et $Y_2 = f_2(X_1, .., X_{p+q})$. L'espérance de $Y = Y_1 Y_2$ conditionnellement à une variable X_j est égale à l'intégrale du produit des fonctions f_1 et f_2 par rapport à toutes les variables sauf X_j. Ainsi, pour les mêmes raisons que précédemment, il ne sera pas possible d'exprimer cette intégrale (ou espérance) en fonction séparément des intégrales (ou espérances) de f_1 et de f_2 (de Y_1 et de Y_2). Nous serons donc dans l'obligation de refaire un nouvelle analyse de sensibilité pour obtenir les indices de sensibilité du modèle Y.

2.1.2.2. Addition de deux modèles

Comme pour les cas de la multiplication de deux modèles, nous différencions les cas où les variables des modèles sont identiques, différentes, ou un mélange des deux.

Modèles à variables différentes

Soit les deux modèles $Y_1 = f_1(X_1, ..., X_p)$ et $Y_2 = f_2(X_{p+1}, ..., X_{p+q})$. Nous nous intéressons au modèle somme $Y = Y_1 + Y_2$.
Etant donné que les variables des deux modèles sont différentes (et indépendantes), la variance de Y est $\mathrm{V}(Y) = \mathrm{V}(Y_1 + Y_2) = \mathrm{V}(Y_1) + \mathrm{V}(Y_2)$. L'espérance de Y conditionnellement à une variable X_j, est quant à elle égale à $\mathrm{E}[Y_1|X_j] + \mathrm{E}[Y_2]$ si $1 \leq j \leq p$, ou à $\mathrm{E}[Y_2|X_j] + \mathrm{E}[Y_1]$ si $p+1 \leq j \leq p+q$. Comme les termes $\mathrm{E}[Y_2]$ et $\mathrm{E}[Y_1]$ sont constants et n'interviennent pas dans la variance, les indices de sensibilité de premier ordre, et ceux d'ordre supérieur relatifs à l'interaction entre variables propre à l'un ou à l'autre des deux modèles, sont donnés par :

$$S_j = \begin{cases} S_j^1 \times \frac{\mathrm{V}(Y_1)}{\mathrm{V}(Y_1)+\mathrm{V}(Y_2)} & \text{si} \quad 1 \leq j \leq p \\ S_j^2 \times \frac{\mathrm{V}(Y_2)}{\mathrm{V}(Y_1)+\mathrm{V}(Y_2)} & \text{si} \quad p+1 \leq j \leq p+q, \end{cases} \tag{2.8}$$

et

$$S_{i_1,...,i_s} = \begin{cases} S^1_{i_1,...,i_s} \times \frac{\mathrm{V}(Y_1)}{\mathrm{V}(Y_1)+\mathrm{V}(Y_2)} & \text{si} \quad \{i_1, ..., i_s\} \subset \{1, ..., p\} \\ S^2_{i_1,...,i_s} \times \frac{\mathrm{V}(Y_2)}{\mathrm{V}(Y_1)+\mathrm{V}(Y_2)} & \text{si} \quad \{i_1, ..., i_s\} \subset \{p+1, ..., p+q\}, \end{cases} \tag{2.9}$$

Les indices de sensibilité liés à l'interaction entre variables de Y_1 et de Y_2 sont nuls puisque les variables sont toutes indépendantes les unes des autres, et que Y est la somme de ces deux modèles.
Comme cette mutation n'entraîne aucune interaction supplémentaire entre les variables de Y_1 et de Y_2, les indices de sensibilité totaux de Y sont :

$$S_{T_i} = \begin{cases} S^1_{T_i} \times \frac{\mathrm{V}(Y_1)}{\mathrm{V}(Y_1)+\mathrm{V}(Y_2)} & \text{si} \quad 1 \leq i \leq p \\ S^2_{T_i} \times \frac{\mathrm{V}(Y_2)}{\mathrm{V}(Y_1)+\mathrm{V}(Y_2)} & \text{si} \quad p+1 \leq i \leq p+q. \end{cases}$$

Modèles à variables identiques

Supposons que les deux modèles $Y_1 = f_1(X_1, ..., X_p)$ et $Y_2 = f_2(X_1, ..., X_p)$ aient les mêmes variables d'entrée. L'espérance d'une somme étant égale à la somme des espérances, l'espérance de Y conditionnellement à une variable X_j est égale à la somme de l'espérance de Y_1 conditionnellement à X_j et de celle de Y_2 conditionnellement à X_j.
Ainsi, par passage à la variance, sachant que la variance d'une somme est égale à la somme des variances plus deux fois la covariance, on obtient les indices de sensibilité de Y par :

$$S_j = S_j^1 \times \frac{\mathrm{V}(Y_1)}{\mathrm{V}(Y)} + S_j^2 \times \frac{\mathrm{V}(Y_2)}{\mathrm{V}(Y)} + \frac{2\mathrm{Cov}(\mathrm{E}[Y_1|X_j], \mathrm{E}[Y_2|X_j])}{\mathrm{V}(Y)} \qquad \forall j \in \{1, \ldots, p\} \tag{2.10}$$

Outre la variance de Y, il est nécessaire d'estimer la covariance de $\mathrm{E}[Y_1|X_j]$ et $\mathrm{E}[Y_2|X_j]$.
Cette estimation est généralement aussi coûteuse que les estimations nécessaires à une nouvelle analyse, et utiliser les indices préalablement estimé n'a pas d'intérêt. Il existe néanmoins des cas particuliers pour lesquels on peut penser que ces covariances s'estiment facilement. C'est le cas des modèles linéaires et additifs.

Cas particulier des modèles linéaires et additifs.
Supposons que $Y_1 = \sum_{i=1}^{p} \alpha_i X_i$ et $Y_2 = \sum_{i=1}^{p} \beta_i X_i$. Dans ce cas $\mathrm{Cov}(\mathrm{E}[Y_1|X_j], \mathrm{E}[Y_2|X_j]) = \alpha_j \beta_j \mathrm{V}(X_j)$.
L'estimation de cette covariance ne coûte pas cher, mais vu la simplicité du modèle, l'estimation des indices de sensibilité n'est pas plus coûteuse. Ce travail n'apporte dans ce cas aucune économie intéressante dans le processus d'estimation.
Si l'on suppose cette fois que les modèles s'écrivent $Y_1 = \sum_{i=1}^{p} \alpha_i f_i(X_i)$ et $Y_2 = \sum_{i=1}^{p} \beta_i g_i(X_i)$, on obtient $\mathrm{Cov}(\mathrm{E}[Y_1|X_j], \mathrm{E}[Y_2|X_j]) = \alpha_j \beta_j \mathrm{Cov}(f_j(X_j), g_j(X_j))$. Là encore le gain numérique est nul, puisque l'estimation de cette dernière covariance est du même coût que l'estimation de la covariance initiale de l'équation (2.10).

Les indices de sensibilité d'ordre supérieur et totaux souffrent eux aussi du même problème, que nous ne savons donc résoudre que dans de très simples cas (additifs) pour lesquels l'économie d'une nouvelle analyse n'est pas nécessaire.

Modèles à variables identiques et différentes

Nous envisageons maintenant un mélange des deux cas précédent : les deux modèles Y_1 et Y_2 ont des variables en communs et d'autres propres à chacun. L'espérance du modèle $Y = Y_1 + Y_2$ conditionnellement à une variable X_i est :

$$\mathrm{E}[Y|X_i] = \begin{cases} \mathrm{E}[Y_1 + Y_2|X_i] = \mathrm{E}[Y_1|X_i] + \mathrm{E}[Y_2|X_i] & \text{si} \quad 1 \leq i \leq p \\ \mathrm{E}[Y_2|X_i] \quad \text{si} \quad p+1 \leq i \leq p+q. \end{cases} \tag{2.11}$$

Par analogie aux deux cas précédent (modèles à variables identiques et modèles à variables différentes), il est possible d'obtenir tous les indices de sensibilité relatifs aux variables qui n'appartiennent qu'à un seul des deux modèles, en l'occurrence Y_2, puisque les autres souffrent du même problème que celui rencontré avec la covariance lorsque les modèles sont à variables identiques.
Ainsi, pour $p+1 \leq j \leq p+q$ et $\{i_1, ..., i_s\} \subset \{p+1, ..., p+q\}$, les indices de sensibilité du modèle Y sont donnés part :

$$\begin{aligned} S_j &= S_j^2 \times \frac{\mathrm{V}(Y_2)}{\mathrm{V}(Y)}, \\ S_{i_1,...,i_s} &= S_{i_1,...,i_s}^2 \times \frac{\mathrm{V}(Y_2)}{\mathrm{V}(Y)}, \\ S_{T_j} &= S_{T_j}^2 \times \frac{\mathrm{V}(Y_2)}{\mathrm{V}(Y)}. \end{aligned}$$

Les autres indices de sensibilité doivent être estimés par une nouvelle analyse de sensibilité.

2.1.3. Bilan

À partir d'un modèle mathématique sur lequel une analyse de sensibilité a été menée, nous avons envisagé un certain nombre de mutations auxquelles il peut être sujet. Pour certaines d'entre elles, il n'est pas nécessaire de recommencer l'analyse de sensibilité pour obtenir les indices de sensibilité du modèle après

mutation. En effet, nous avons montré comment il était possible de les estimer à partir de ceux du modèle initial, et ce à moindre coût.
Nous citons les principales mutations observant de telles propriétés.

- mutations des variables d'entrée du modèle :
  - transformation d'une variable d'entrée.
    Le modèle de départ $Y = f(X_1, .., X_i, ..., X_p)$ est muté en $Y^{'} = f(X_1, .., h(X_i), ..., X_p)$.
    Il est possible d'obtenir les indices de sensibilité de $Y^{'}$ (non relatif à X_i) à partir de ceux de Y si $Y = f_1(X_i) + f_2(\boldsymbol{X}_{\sim i})$, ou sinon sous l'hypothèse (très contraignante) que $h(t) = -t + 2m$ et que la loi des variables d'entrée soit symétrique par rapport à m. Sous des hypothèses de linéarité de h et du modèle en X_i, les indices de sensibilité relatifs à la variable X_i peuvent être obtenus à partir de ceux avant mutation.
  - Une variable d'entrée devient déterministe (on décide de fixer une variable aléatoire du modèle).
    On différencie alors trois formes particulières du modèle pour lesquelles il est possible de déduire les indices du modèle après mutation :
    - $Y = \sum_{k=1}^{K} f_k^1(\boldsymbol{X}_{\sim i}) f_k^2(X_i)$, sous les conditions suivantes :

$$\begin{cases} {f_k}^2(t) = at + b, & \forall 1 \leq k \leq K \\ \mathrm{E}[X_i] = \alpha & \text{où } \alpha \in \mathbb{R} \end{cases}$$

    - $Y = f_1(X_i) + f_2(\boldsymbol{X}_{\sim i})$ sans condition,
    - $Y = f_1(X_i) f_2(\boldsymbol{X}_{\sim i})$ sans condition.
  - Un paramètre déterministe devient une variable aléatoire (on décide de considérer un paramètre déterministe du modèle comme aléatoire).
    Le paramètre déterministe en question est considéré comme une variable aléatoire supplémentaire X_{p+1}. Quelques cas particuliers sur la forme du modèle permettent de déduire les indices de sensibilité après mutation (sauf pour les indices totaux, et pour ceux relatifs à la nouvelle variable) :
    - $Y = \sum_{k=1}^{K} f_k^1(\alpha) f_k^2(X_1, .., X_p)$, sous les conditions suivantes :

$$\begin{cases} f_k^1(t) = at + b, & \forall 1 \leq k \leq K \\ \mathrm{E}[X_{n+1}] = \alpha & \text{où } \alpha \in \mathbb{R} \end{cases}$$

    et sans condition pour :
    - $Y = f^1(\alpha) + f_2(X_1, .., X_p)$,
    - $Y = f^1(\alpha) f_2(X_1, .., X_p)$,
    - introduction d'un bruit blanc additif,
    - introduction d'un bruit blanc multiplicatif.
- Composition de plusieurs analyses de sensibilité.
  L'addition de deux modèles permet d'obtenir les indices de sensibilité du modèle somme, lorsque ces modèles sont à entrées différentes. Lorsque parmi leurs entrées, certaines sont identiques et d'autres sont différentes, nous ne pouvons obtenir les indices de sensibilité que pour les variables propres à un seul des deux modèles. La multiplication de deux modèles permet quant à elle d'obtenir les indices de sensibilité du produit que si les variables d'entrée des deux modèles sont différentes.

2.1.4. Conclusion

Pour chaque mutation décrite précédemment, nous avons relié formellement, lorsque cela était possible, les indices de sensibilité du modèle avant mutation avec ceux du modèle après mutation. Ces relations mathématiques permettent, connaissant les indices de sensibilité avant mutation, de calculer ceux après mutation à

moindre coût (sans refaire une analyse de sensibilité complète), en estimant par exemple une simple variance ou espérance. En présence d'une certaine mutation de modèle, il suffira donc d'appliquer les résultats correspondants. La liste des mutations étudiées n'étant pas exhaustive, il peut arriver en pratique de rencontrer une mutation non étudiée. Il sera alors possible d'appliquer une démarche analogue à celle employée pour les autres mutations, et ainsi de compléter la liste des mutations étudiées. Un exemple illustratif sur le logiciel GASCON est présenté dans la section 2.3.

2.2. Utilisation d'un modèle simplifié

Considérons un modèle mathématique de référence M_1 caractérisé par :

$$Y = f_1(\mathbf{X}),$$

où Y est la variable de sortie, $\mathbf{X} = (X_1, ..., X_p)$ est le vecteur des variables d'entrée que nous supposons indépendantes, et f_1 la fonction qui associe Y à $\mathbf{X}$.
Nous supposons que ce modèle M_1 nécessite un temps de calcul trop important pour pouvoir réaliser une analyse de sensibilité. En effet, il arrive parfois de rencontrer des modèles dont l'exécution peut prendre plusieurs heures. C'est le cas par exemple des modèles de calcul thermohydraulique du coeur d'un réacteur nucléaire, ou de simulation des phénomènes apparaissant lors d'un scénario d'accident grave.
Nous considérons alors une version simplifiée M_2 du modèle de référence. Ce modèle M_2 est caractérisé par :

$$\hat{Y} = f_2(\mathbf{X}),$$

où $\hat{Y}$ est une approximation de Y. L'idée est alors d'utiliser l'analyse de sensibilité de M_2 pour approximer celle de M_1.
La qualité de l'approximation dépend de la nature du modèle simplifié. Si le modèle M_2 est trop simplifié, il risque de ne plus être très proche de M_1, et la qualité de l'approximation de l'analyse de sensibilité de M_1 pourrait alors être médiocre.
Nous discernons dès lors deux situations. Dans la première, le modèle simplifié est une surface de réponse (cf. définition 2.2.1), construite sur un jeu de données de réalisations du modèle de référence. Dans la seconde situation, nous considérons que la simplification a été effectuée à partir de considérations physiques.

2.2.1. Première situation : surface de réponse

Nous supposons pouvoir constituer un jeu de données simulées à partir du modèle de référence M_1, dont la taille dépend directement des connaissances a priori, des moyens dont on dispose et du temps de calcul de M_1.
La construction d'un modèle simplifié, que nous envisageons ici, consiste en l'ajustement d'une surface de réponse sur ces données. Nous définissons ci-dessous la notion de surface de réponse.

Définition 2.2.1. *Surface de réponse.*
Une surface de réponse est une fonction utilisée pour modéliser un certain processus ou phénomène, à partir d'une base de données de réalisations de ce dernier. Cette fonction doit avoir les propriétés suivantes :

- *une bonne approximation de la base de données d'apprentissage,*
- *une bonne capacité de prédiction dans le domaine d'étude,*
- *un temps de calcul associé à une réalisation négligeable afin de permettre des simulations intensives.*

Les familles de surface de réponse répondant à ces critères sont alors nombreuses [17] : modèles linéaires multiples, modèles additifs, modèles mixtes, splines, ondelettes ou encore réseaux de neurones, *etc*. Chacune de ces familles a des avantages et des inconvénients. Ce n'est pas l'objet de ce document de les évaluer.
Nous écrivons le problème de l'utilisation d'une surface de réponse sous la forme suivante :

$$Y = f_1(\mathbf{X}) = f_2(\mathbf{X}) + \Delta, \tag{2.12}$$

où Δ est l'erreur due à l'approximation. Nous distinguons trois situations.
Dans la première situation, le modèle M_2 a déjà été construit à partir d'une base de données de simulations de M_1. On dispose donc d'une base de données de résidus issus de cet ajustement. Nous proposons de modéliser ces résidus, par une fonction écart $\hat{\Delta}$. Suivant la nature de cette modélisation, nous discuterons au paragraphe 2.2.1.1 la possibilité d'utiliser cette fonction écart pour améliorer l'approximation de l'analyse

de sensibilité de M_1.
Dans la seconde situation, on suppose que la surface de réponse M_2 n'a pas encore été construite, mais que la base de données est imposée. Nous avons donc le choix du type de modèle simplifié à construire, ou de la famille de surface de réponse à utiliser. Ce choix, discuté au paragraphe 2.2.1.2, se fera de sorte à obtenir la meilleure approximation possible de l'analyse de sensibilité de M_1.
Dans la dernière situation, nous supposons avoir une liberté totale sur le choix de la base de données, ainsi que sur celui de la surface de réponse à utiliser.

2.2.1.1. Utilisation des résidus issus de l'ajustement de la surface de réponse

Nous supposons disposer d'une base de données $(y_i, \mathbf{x}_i)_{i=1,\dots,N}$ de N simulations de ce modèle de référence. Cette base de données a servi de support à la construction de la surface de réponse M_2 :

$$\hat{Y} = f_2(\mathbf{X}),$$

qui est une approximation de M_1, dont les propriétés permettent de faire des simulations intensives. La surface de réponse M_2 n'est pas nécessairement fonction des mêmes variables que le modèle de référence M_1, il est possible que certaines variables aient été omises (volontairement ou non) au cours de la simplification. L'erreur d'approximation, introduite en (2.12), est définie comme l'écart entre le modèle de référence et la surface de réponse :

$$\Delta = f_1(\mathbf{X}) - f_2(\mathbf{X}).$$

Soient δ_i les résidus issus de l'ajustement de M_2 sur la base de données de M_1 :

$$\delta_i = y_i - f_2(\mathbf{x}_i) \qquad \forall i = 1, \dots, N.$$

Ces résidus sont des réalisations de la variable aléatoire Δ. Nous proposons de modéliser l'écart Δ par une fonction écart $\hat{\Delta}$ (une autre surface de réponse), en utilisant ces résidus $(\delta_i)_{i=1,\dots,N}$. Comme $\hat{\Delta}$ est une surface de réponse, il est possible de réaliser une analyse de sensibilité sur cette fonction sans engendrer un coût trop important.

Remarque. *Une approche analogue, spécifiquement adaptée au contexte de la fiabilité des structures mécaniques, a été proposée par Pendola [41]. Comme nous l'avons vu dans la section 1.2, elle consiste à modéliser l'écart de simplification par une surface de réponse polynomiale du second degré.*

La modélisation de Δ par $\hat{\Delta}$ nous permet de proposer une nouvelle approximation du modèle de référence :

$$Y = f_1(\mathbf{X}) = f_2(\mathbf{X}) + \hat{\Delta} + \epsilon,$$

où ϵ est la nouvelle erreur faite en modélisant Δ par $\hat{\Delta}$.
Nous supposons que l'ajustement de cette fonction écart permet d'obtenir une meilleure approximation du modèle de référence, c'est-à-dire que le modèle $M_2 + \hat{\Delta}$ est une meilleure approximation de M_1 que M_2 seule. Il existe différents critères statistiques permettant de comparer deux modèles, parmi lesquels les critères BIC (*Bayesian Information Criteria* [57]) et AIC (*Akaike Information Criteria* [1, 2]). Si ce n'est pas le cas, le processus de modélisation de Δ par $\hat{\Delta}$ devra être itéré.
L'analyse de sensibilité de M_1, qui a été approximée par celle de M_2, peut alors être améliorée en l'approximant par celle de $M_2 + \hat{\Delta}$. Le coût engendré sera celui d'une nouvelle analyse de sensibilité de $M_2 + \hat{\Delta}$, que nous considérons comme négligeable, puisque $M_2 + \hat{\Delta}$ est une surface de réponse.
Il reste néanmoins encore une erreur d'approximation, que l'on note cette fois ϵ. Si cette erreur est de variance négligeable devant celles de M_1 et de $M_2 + \hat{\Delta}$ (par exemple, ϵ peut être un biais), l'approximation sera alors correcte. Sinon, il sera possible d'itérer ce processus jusqu'à l'obtention d'une erreur d'approximation

de variance négligeable, qui pourra alors être considérée comme un bruit.

Etant données ces conclusions, il aurait été plus judicieux, si possible, de construire M_2 de sorte que l'approximation faite soit la meilleure possible, ou tout au moins que l'on ait une idée de sa qualité. C'est ce que nous discutons dans le paragraphe suivant.

2.2.1.2. Choix du modèle simplifié à partir d'une base de données existante

Nous supposons que la base de données de simulations du modèle est imposée. Il n'est pas possible de faire de nouvelles simulations. Nous donnons ici quelques conseils pour la construction de cette surface de réponse.
L'approximation de M_1 par M_2 entraîne un certain écart Δ, défini en (2.12) par :

$$Y = f_1(\mathbf{X}) = f_2(\mathbf{X}) + \Delta.$$

La surface de réponse M_2 doit être construite de sorte que son analyse de sensibilité soit la plus proche possible de celle de M_1. Or, l'analyse de sensibilité consiste à expliquer la variance de Y en fonction des variables d'entrée. Cette variance s'écrit :

$$\mathrm{V}(Y) = \mathrm{V}(f_1(\mathbf{X})) = \mathrm{V}(f_2(\mathbf{X})) + \mathrm{V}(\Delta) + 2\mathrm{Cov}(f_2(\mathbf{X}), \Delta).$$

La situation idéale est donc celle où Δ est de variance nulle (d'où covariance avec f_2 nulle), ce qui impliquerait l'égalité des variances de f_1 et f_2, et donc des analyses de sensibilité de M_1 et de M_2. La surface de réponse M_2 devra être construite de façon à s'approcher au maximum de cette situation utopique (il est possible de réitérer la construction de M_2 pour intégrer Δ si sa variance n'est pas négligeable). Si toutefois il est impossible d'obtenir un écart Δ de variance négligeable, l'approximation de l'analyse de sensibilité de Y par celle de $\hat{Y}$ ne sera pas correcte, puisqu'une part de la variance de Y ne sera pas expliquée. Comme pour la problématique précédente (paragraphe 2.2.1.1), il sera alors nécessaire de définir une meilleure approximation de Y en construisant une nouvelle surface de réponse sur les résidus Δ, et ainsi de suite.
Outre les conditions classiques assurant que la surface de réponse M_2 est une bonne approximation de M_1 (comparaison des premiers moments statistiques, étude des résidus Δ), les conditions spécifiques à notre contexte de l'approximation pour l'analyse de sensibilité sont : variance de Δ minimale, orthogonalité «maximale» de M_2 et Δ (au sens covariance minimale).

Antoniadis [5] montre qu'en choisissant une base orthonormée adéquate, il est possible de construire un modèle simplifié en contrôlant l'erreur Δ commise. Ses travaux, décrits en annexe A.3, introduisent aussi une nouvelle approche de l'analyse de sensibilité, utilisant l'analyse de la variance fonctionnelle.

2.2.1.3. Choix du modèle simplifié et de la base de données

Supposons que l'on soit dans l'une des deux situations suivantes :

- Une base de données de simulations du modèle existe, mais il est possible de l'enrichir par un certain nombre de simulations supplémentaires. Afin de conserver un sens à l'utilisation d'une surface de réponse, nous supposons ce nombre limité.
- Aucune donnée n'a encore été générée. La construction de la base de données est totalement libre.

Il est important de choisir judicieusement la façon dont sont générées les simulations, en vue de l'ajustement de la surface de réponse. La théorie des plans d'expériences [22] peut être utilisée afin d'optimiser le nombre de données requises en fonction de la nature de la surface de réponse utilisée.
Ce thème constitue un axe de recherche intéressant que nous n'avons pas développé dans cette thèse.

2.2.2. Seconde situation : modèle simplifié à partir de considérations physiques

Nous considérons maintenant que le modèle simplifié M_2 :

$$\hat{Y} = f_2(\mathbf{X}),$$

a été élaboré à partir de considérations physiques. Le modèle M_2 pourra ou non être une fonction des mêmes variables $\mathbf{X}$ que M_1 (il se peut que certaines variables de M_1 n'entre pas en compte dans M_2). Nous distinguons alors deux cas de figure concernant l'élaboration de M_2.

Dans le premier nous considérons que le modèle M_2 utilise une autre approche de modélisation plus simple que celle du modèle de référence. Par exemple, M_2 modélise le problème en dimension 1 alors que M_1 le considère en dimension 3, ou encore on remplace le modèle éléments finis par une règle simplifiée pour M_2.

Dans le second cas de figure, le modèle simplifié suppose non influents certains phénomènes pris en compte dans le modèle de référence. Nous citons trois exemples illustrant ce cas de figure :

- Le modèle de référence est un modèle de comportement mécanique, intégrant à la fois rupture fragile et rupture ductile. Si l'on suppose que dans la situation étudiée la rupture ductile est très peu probable, il est possible d'utiliser un modèle simplifié n'incluant que la rupture fragile.
- Le modèle de référence est cette fois un modèle de comportement des fluides dans la cuve d'un réacteur intégrant les situations flux laminaires et flux turbulents. Un modèle simplifié peut alors être de ne considérer qu'une seule de ces deux situations.
- Enfin, le modèle de référence est un modèle de relâchement de gaz de fusion d'un combustible nucléaire. En jouant sur les paramètres du modèle de référence, il est possible de construire un modèle simplifié inhibant certains phénomènes.

Pour chacune de ces situations, étant donnés les domaines de variations considérés des variables d'entrée, et selon le contexte de l'étude, l'analyse du phénomène physique permet de montrer que certaines parties du modèle de référence peuvent être négligées afin de construire un modèle plus simple.

Dans chacun de ces deux cas de figure, la simplification a généré une erreur d'approximation du modèle de référence par le modèle simplifié :

$$\Delta = f_1(\mathbf{X}) - f_2(\mathbf{X}).$$

Si cette erreur Δ est négligeable d'un point de vue variance, l'approximation de l'analyse de sensibilité de M_1 par celle de M_2 est correcte.
Si Δ n'est pas négligeable, nous proposons alors, comme dans le paragraphe 2.2.1 de la modéliser par une surface de réponse $\hat{\Delta}$:

$$\Delta = \hat{\Delta} + \epsilon.$$

Pour cela, nous devons construire une base de données de réalisations de cette erreur Δ, autrement dit, sur un certain nombre de simulations des variables aléatoires d'entrée, nous calculons la réponse du modèle de référence M_1 et celle du modèle simplifié M_2. Les différences entre ces deux réponses forment un base de réalisations de Δ, que l'ont note $(\delta_i)_{i=1\ldots N}$.
Ainsi, une nouvelle approximation du modèle de référence est obtenue :

$$Y = f_1(\mathbf{X}) = f_2(\mathbf{X}) + \hat{\Delta} + \epsilon.$$

Comme dans le cas des surfaces de réponse, une nouvelle erreur d'approximation ϵ apparaît. Si cette erreur n'est pas négligeable d'un point de vue variance, il sera nécessaire d'itérer le processus. Si au contraire elle est négligeable, l'analyse de sensibilité de $f_2(\mathbf{X}) + \hat{\Delta}$ est alors bonne approximation de l'analyse de M_1 à la condition que le modèle $f_2(\mathbf{X}) + \hat{\Delta}$ soit une meilleure approximation du modèle M_1 que celle faite en

n'utilisant que f_2 (à partir par exemple de l'utilisation des critères statistiques de type BIC ou AIC). Si cette condition n'est pas remplie, le processus de construction de la surface de réponse $\hat{\Delta}$ devra être itéré.
Le coût engendré par la nouvelle analyse de $f_2(\mathbf{X}) + \hat{\Delta}$ est bien moindre que celui de l'analyse de M_1, puisque $\hat{\Delta}$ est une surface de réponse et M_2 un modèle simplifié construit pour cela.
Dans le cas général où M_2 et $\hat{\Delta}$ ont des variables communes, nous pouvons refaire cette analyse de sensibilité, et ainsi obtenir une bonne approximation de l'analyse de sensibilité de M_1 puisque l'erreur est considérée négligeable.
Néanmoins, si M_2 et $\hat{\Delta}$ n'ont pas de variables communes, il peut être intéressant d'utiliser les résultats de la section 2.1, qui nous permettent d'obtenir les indices de sensibilité de $f_2(\mathbf{X}) + \hat{\Delta}$ à moindre coût. En effet, la nouvelle approximation du modèle de référence peut être vue comme une mutation consistant à ajouter les deux modèles $f_2(\mathbf{X})$ et $\hat{\Delta}$. Comme il n'est pas vraisemblable d'un point de vue applicatif que les variables de la fonction écart soient de nouvelles variables, nous supposons que ce sont des variables du modèle de référence qui ne sont pas exprimées dans le modèle simplifié.
Le problème s'écrit dans ce cas :

$$Y \simeq f_2(X_1, \ldots, X_k) + \hat{\Delta}(X_{k+1}, \ldots, X_p) \quad \text{avec } 1 \leq k < p.$$

Les indices de sensibilité du modèle de référence sont alors approximés (cf. paragraphe 2.1.2.2) par :

$$\begin{aligned} S_j^1 &\simeq S_j^2 \times \frac{\mathrm{V}(f_2(X_1, \ldots, X_k))}{\mathrm{V}(f_2(f_2(X_1, \ldots, X_k)) + \mathrm{V}(\hat{\Delta}(X_{k+1}, \ldots, X_p))} \quad \text{si } 1 \leq j \leq k, \\ S_j^1 &\simeq S_j^{\hat{\Delta}} \times \frac{\mathrm{V}(\hat{\Delta}(X_{k+1}, \ldots, X_p))}{\mathrm{V}(f_2(f_2(X_1, \ldots, X_k)) + \mathrm{V}(\hat{\Delta}(X_{k+1}, \ldots, X_p))} \quad \text{si } k+1 \leq j \leq p. \end{aligned}$$

L'ensemble des différentes situations de ce paragraphe sont résumées (schématiquement) par la figure 2.1.

Remarque (Application aux modèles concurrents)**.** *Dans les différentes situations étudiées précédemment, nous avons défini une fonction écart entre deux modèles (modèle de référence et modèle simplifié ou surface de réponse). Cette fonction écart peut avoir d'autres applications.*
Dans de nombreuses situations (mécanique, stockage profond...), des modèles concurrents pour représenter un même phénomène sont développés (par des équipes différentes parfois et sur des outils différents). Par exemple, l'utilisation de modèles 3D peut parfois être en concurrence avec des modèles 1D ou 2D. Ce problème de l'existence de modèles concurrents est souvent étudié dans la littérature relative à l'incertitude de modèle, et a été décrit au paragraphe 1.2.2.
À partir de la fonction écart entre deux modèles concurrents, il peut être possible sous certaines conditions de déduire les indices de sensibilité d'un modèle à partir de ceux de l'autre, comme cela a été fait précédemment entre le modèle de référence et la surface de réponse (ou modèle de référence et modèle simplifié). Nous ne développons pas ici cette application.

2.2.3. Conclusion

Le problème de l'utilisation d'un modèle simplifié pour une analyse de sensibilité alors qu'un modèle de référence existe a été étudié en modélisant l'écart entre le modèle et sa simplification.
Si le modèle simplifié est une surface de réponse et si cette dernière peut être améliorée, une nouvelle analyse de sensibilité (non coûteuse en temps de calcul) donnera de meilleures approximations des indices de sensibilité de M_1. En outre, nous avons aussi discuté les conditions à remplir par une surface de réponse en vue d'obtenir la meilleure approximation possible de l'analyse de sensibilité d'un modèle de référence.
Si le modèle simplifié est élaboré à partir de considérations physiques sur le phénomène étudié, et s'il est possible et utile de l'améliorer, une nouvelle analyse donnera elle aussi de meilleures approximation des

indices de M_1. Cette analyse sera d'un coût moindre par rapport à celui d'une analyse de sensibilité de M_1. Néanmoins, lorsque le modèle simplifié et la surface de réponse construite sur l'écart avec le modèle de référence sont fonction de variables différentes, les travaux sur les mutations de modèle peuvent être appliqués pour obtenir les indices de sensibilité de la nouvelle approximation de M_1 à moindre coût. L'utilité de cette démarche, en terme de gain numérique, est discutée au paragraphe 2.3.6.

Dans la section suivante, le logiciel GASCON est présenté. L'analyse de sensibilité d'une partie de ce code nécessite l'utilisation d'un modèle simplifié. Nous étudions alors la possibilité de modéliser l'écart entre ce modèle simplifié et le modèle de référence.

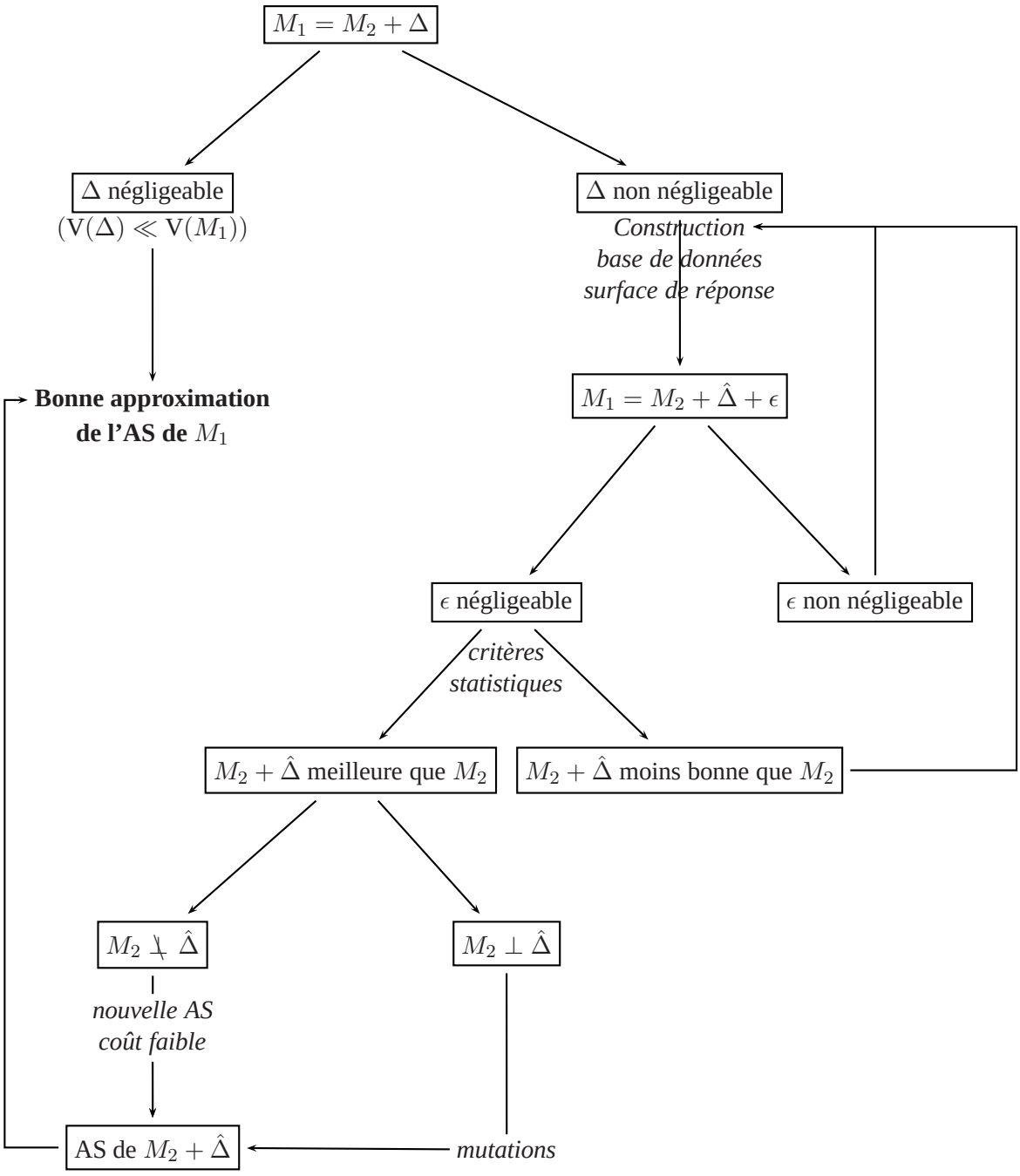


FIG. 2.1. : Approximation de l'analyse de sensibilité (AS) de M_1

2.3. Applications au logiciel GASCON

Cette section a pour objectif d'illustrer sur une application réelle l'applicabilité des deux sections 2.1 et 2.2. Nous présentons dans un premier temps le logiciel GASCON, la variable de sortie étudiée, ainsi que la surface de réponse utilisée pour l'analyse de sensibilité (GASCON étant trop lourd en temps de calcul). Nous analysons ensuite l'impact de l'utilisation de la surface de réponse à la place du logiciel GASCON lui-même. Puis nous introduisons un certain nombre de mutations de cette surface de réponse. Pour chacune d'elle, les résultats de la section 2.1 sont utilisés pour déduire les nouveaux indices de sensibilité après mutation. En extrapolant ce travail de traitement des mutations, nous donnons finalement des conseils sur la réalisation d'une analyse de sensibilité lorsque le modèle étudié peut se diviser en sous modèles.

2.3.1. Le logiciel GASCON

Le code GASCON [4] est un code utilisé pour quantifier les transferts dans l'environnement et son impact sur l'homme suite à un rejet atmosphérique continu de radionucléides. Il permet d'évaluer les doses reçues par une population (groupe de référence) à partir du calcul des concentrations de radionucléides dans les différentes étapes des chaînes de transport incluant l'air, la couche superficielle du sol, les végétaux, les animaux, etc. Les résultats fournis par GASCON sont sous la forme de doses efficaces annuelles par voie de transfert et de doses totales reçues par le groupe de référence. Ce groupe est divisé en trois tranches d'âge : adulte, enfant (environ dix ans) et nouveau-né.

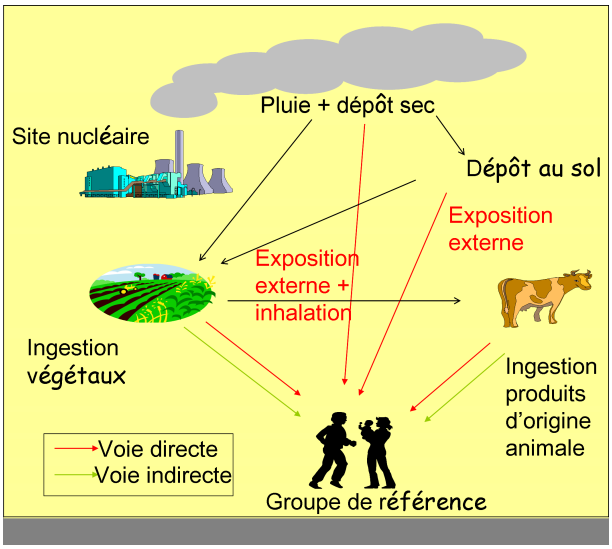


FIG. 2.2. : Voies de transfert prises en compte par GASCON.

Les principales voies d'exposition ou d'atteinte prises en compte dans GASCON, représentées par la figure 2.2, sont :

- l'exposition externe au nuage,
- l'inhalation,
- l'exposition externe au dépôt au sol,
- l'ingestion de végétaux contaminés par voie directe et indirecte,
- l'ingestion de productions animales par voie directe et indirecte.

Parmi les variables de sorties de GASCON, nous étudions dans cette section la dose efficace annuelle (Sv/an) en iode 129 (^{129}I), reçue par un adulte à la suite d'un rejet sur une année d'une cheminée d'une installation de Cadarache. Ce rejet est fixé à une certaine valeur de référence sans signification physique (1 Becquerel

par an). GASCON permet plus généralement l'analyse de l'ensemble des radionucléides, en fonction de différentes durées de fonctionnement. Nous considérons aussi que l'adulte habite au voisinage de Cadarache (France, Bouches-du-Rhône). La réponse étudiée est notée Ad_I_1.

Les variables d'entrée considérées pour l'étude de sensibilité peuvent être regroupées sous les classes suivantes :

- facteurs de dose pour la tranche d'âge «adulte»,
- facteurs de transfert aux productions animales,
- facteurs de transfert sol-plante séparés par famille de végétaux„
- facteurs de translocation (caractérisant le transfert interne au végétal entre la surface follaire et les organes consommés),
- coefficients de sorption ou de partage (coefficient de rétention de particules selon le sol),
- vitesse de dépôt sec,
- rations alimentaires du groupe de référence pour l'adulte,
- rations alimentaires des animaux liés aux produits consommés par le groupe de référence,
- pluviométrie.

GASCON est un code complexe programmé en Visual Basic sous Excel. Sa complexité n'est pas due aux formulations mathématiques utilisées (relativement simples), mais à l'importance du nombre de phénomènes pris en compte. Son exécution est relativement longue (une trentaine de secondes par calcul sur un Pentium IV), et ne nous permet de constituer qu'un échantillon de taille 1000. Cet échantillon ne sera pas de taille suffisante pour effectuer l'analyse de sensibilité par la méthode de Sobol. La méthode employée dans [29] pour réaliser une analyse de sensibilité sur GASCON consiste à construire des surfaces de réponse pour les différentes variables de sorties. Les surfaces de réponse employées sont des polynômes de degré deux, obtenus par régression polynomiale.

La surface de réponse construite pour la variable de sortie Ad_I_1 est le modèle sur lequel nous réaliserons l'analyse de sensibilité.

2.3.2. Surface de réponse pour $Y = Ad\_I\_1$

Parmi les différentes régressions multiples testées dans [29], le meilleur ajustement pour Ad_I_1 est obtenu lorsque les régressions linéaires sont faites en fonction de deux chaînes alimentaires (pour la situation étudiée ici) : celle du lait de chèvre et celle du lait de brebis. En effet, les dépôts de radionucléides se font essentiellement dans un voisinage très proche de l'installation de Cadarache, dans lequel évoluent des chèvres et des brebis. Et parmi les principaux aliments régionaux que consomment les individus du groupe de référence étudié, les produits laitiers sont ceux qui fixent le plus les radionucléides.
Le modèle d'ajustement de $Y = Ad\_I\_1$ qui en découle est :

$$\begin{aligned} Y - \alpha_0 &= \alpha_1 X_1 X_2 X_3 X_4 X_5 + \alpha_2 X_1 X_2 X_3 X_4 X_5^2 + \alpha_3 X_1 X_2 X_3 X_6 X_5 + \alpha_4 X_1 X_2 X_3 X_6 X_5^2 \\ &\quad + \alpha_5 X_1 X_7 X_8 X_9 X_5 + \alpha_6 X_1 X_7 X_8 X_9 X_5^2 + \alpha_7 X_1 X_7 X_8 X_{10} X_5, \end{aligned} \tag{2.13}$$

où les X_i sont les variables d'entrée :

$$\begin{aligned}
X_1 &: \text{ingestion efficace de l'adulte},\\
X_2 &: \text{facteur de transfert au lait de chèvre de } {}^{129}I,\\
X_3 &: \text{ration alimentaire de l'adulte en lait de chèvre},\\
X_4 &: \text{ration alimentaire en herbe de la chèvre},\\
X_5 &: \text{vitesse de dépôt sec de } {}^{129}I,\\
X_6 &: \text{ration alimentaire en foin de la chèvre},\\
X_7 &: \text{facteur de transfert au lait de brebis de } {}^{129}I,\\
X_8 &: \text{ration alimentaire de l'adulte en lait de brebis},\\
X_9 &: \text{ration alimentaire en herbe du mouton},\\
X_{10} &: \text{ration alimentaire en foin du mouton},
\end{aligned}$$

et les α_i les coefficients de la régression :

coeff.	valeur	coeff.	valeur
α_0	1.0303×10^{-14}	α_4	-3.5456×10^{-10}
α_1	5.3945×10^{-11}	α_5	6.2968×10^{-11}
α_2	-8.1069×10^{-10}	α_6	-1.0495×10^{-9}
α_3	2.2303×10^{-11}	α_7	1.0339×10^{-11}

TAB. 2.1.: Coefficients du modèle de régression multiple.

Les investigateurs du code GASCON ont supposé que chaque variable d'entrée était de loi «bi-uniforme» sur son intervalle de variation $[min, max]$ autour d'une valeur nominale (valeur choisie comme étant la plus vraisemblable), c'est-à-dire uniforme sur $[min, nominale]$ avec la probabilité $\frac{1}{2}$ et uniforme sur $[nominale, max]$ avec la probabilité $\frac{1}{2}$. La densité de probabilité d'une telle variable aléatoire s'écrit :

$$\rho(x) = \begin{cases} \frac{1}{nominale-min}\mathbb{1}_{[min,nominale]}(x) & \text{avec la probabilité } p=\frac{1}{2} \\ \frac{1}{max-nominale}\mathbb{1}_{[nominale,max]}(x) & \text{avec la probabilité } p=\frac{1}{2} \end{cases}$$

L'espérance et la variance de telles variables aléatoires sont données par :

$$\begin{aligned}
\mathrm{E}[X_i] &= \frac{max + 2nominale + min}{4},\\
\mathrm{V}(X_i) &= \frac{max^2 + 2nominale^2 + min^2}{4} - \mathrm{E}[X_i]^2.
\end{aligned}$$

Les plages de variations des 10 variables d'entrée, correspondant à Ad_I_1, pour le site étudié, sont données par le tableau 2.2.

2.3.3. Impact de l'utilisation d'un modèle simplifié

La surface de réponse (2.13) a été construite dans l'étude [29] à partir de 666 points de la base de données (choisi aléatoirement). Les autres points ayant servi à la validation de cette surface de réponse.
La figure 2.3 représentent les points ajustés par la surface de réponse en fonction des points de la base de données. La répartition selon la droite $y = x$ indique une bonne approximation de GASCON par la surface

variable	min	nominale	max	espérance	variance
X_1	1.1×10^{-8}	1.1×10^{-7}	1.1×10^{-6}	3.3275×10^{-7}	1.9786×10^{-13}
X_2	0.06	0.43	0.65	0.3925	0.0449
X_3	11	110	1100	332.75	1.9786×10^{5}
X_4	255.7	2557	25570	7734.925	1.0691×10^{8}
X_5	0.0005	0.005	0.05	0.01512	4.088×10^{-4}
X_6	109.6	1096	10960	3315.4	1.9642×10^{7}
X_7	0.08	0.49	0.94	0.5	0.0925
X_8	1	10	100	30.25	1.6352×10^{3}
X_9	255.7	2557	25570	7734.925	1.0691×10^{8}
X_{10}	109.6	1096	10960	3315.4	1.9642×10^{7}

TAB. 2.2.: Plage de variations des variables d'entrée.

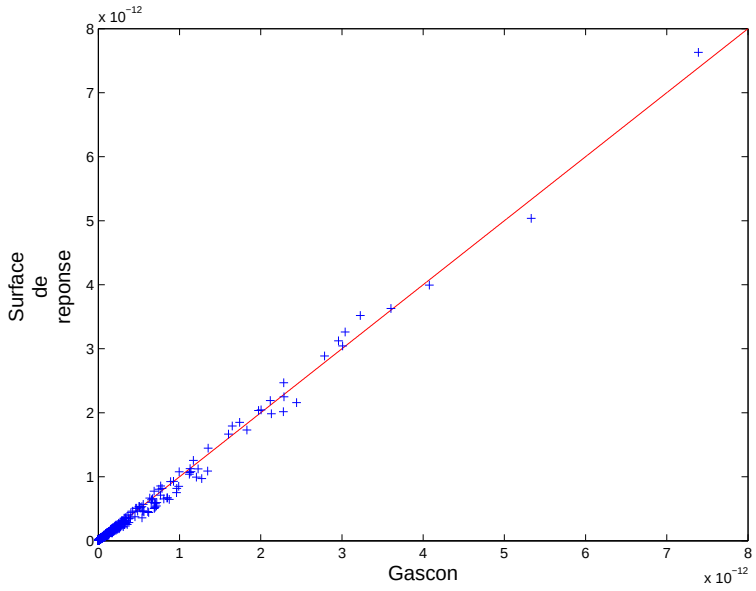


FIG. 2.3. : Surface de réponse versus base de données de construction de GASCON

de réponse, et ce sur toute l'étendue des valeurs (de 0 à 8×10^{-12}). Les moyennes, écarts-types, minimums et maximums des points de la base de données (y_i^{Gascon}), de leur ajustement par la surface de réponse (y_i^{SR}), et des résidus ($e_i = y_i^{Gascon} - y_i^{SR}$) sont présentés dans le tableau 2.3.

Les moyennes et écart-type de Gascon et de la surface de réponse sont très proche. Cette surface de réponse a été validée dans l'étude [29], et nous confirmons cette validité à l'aide d'une analyse de variance de cette régression, en décomposant classiquement la variance totale en somme de la variance résiduelle et de la

	minimum	moyenne	maximum	écart-type
GASCON	8.053×10^{-17}	$1.67633789 \times 10^{-13}$	7.390×10^{-12}	5.096×10^{-13}
Surface de réponse	1.113×10^{-14}	$1.67633774 \times 10^{-13}$	7.629×10^{-12}	5.078×10^{-13}
résidus	-2.935×10^{-13}	1.569×10^{-20}	3.021×10^{-13}	4.251×10^{-14}

TAB. 2.3.: Statistiques sur les bases de données de construction, d'ajustement et de résidus.

variance expliquée par la régression :

$$\underbrace{\frac{1}{666}\sum_{i=1}^{666}(y_i^{Gascon} - \overline{y^{Gascon}})^2}_{\text{Variance totale}} = \underbrace{\frac{1}{666}\sum_{i=1}^{666}(y_i^{Gascon} - y_i^{SR})^2}_{\text{Variance résiduelle}} + \underbrace{\frac{1}{666}\sum_{i=1}^{666}(y_i^{SR} - \overline{y^{Gascon}})^2}_{\text{Variance expliquée}}.$$

Le coefficient de détermination R^2, défini par le rapport de la variance expliquée sur la variance totale, est ici égal à :

$$R^2 = \frac{\sum_{i=1}^{666}(y_i^{Gascon} - \overline{y^{Gascon}})^2 - \sum_{i=1}^{666}(y_i^{Gascon} - y_i^{SR})^2}{\sum_{i=1}^{666}(y_i^{Gascon} - \overline{y^{Gascon}})^2} \simeq \frac{2.2514 \times 10^{-22}}{2.2672 \times 10^{-22}} \simeq 0.9930.$$

Il indique en étant très proche de 1 que l'ajustement est de très bonne qualité.
Un test classique de la qualité de l'ajustement peut être réalisé à partir de ce coefficient de détermination, sachant que sous l'hypothèse que la régression est non significative ou non valide (coefficients de la régression tous nuls) :

$$\frac{R^2}{1-R^2}\frac{n-p-1}{p} \sim F(p, n-p-1),$$

où n est le nombre de points de la base de données ($n = 666$), p est le nombre de variables de la régression ($p = 10$), et $F(p, n-p-1)$ est une loi de Fisher-Sndecor de paramètres p et $n-p-1$.
En se fixant un risque de première espèce entre 5% et 1%, l'hypothèse que notre régression n'est pas valide est toujours rejetée.
On confirme donc la validation de cette surface de réponse faite dans l'étude [29].

Ce qui nous intéresse plus particulièrement ici est l'étude des résidus issus de cette régression. La figure 2.4 présente les résidus studentisés, c'est-à-dire divisés par leur écart-type estimé sur l'échantillon des 666 résidus. Nous pouvons remarquer que 95% de ces résidus studentisés sont compris entre -2 et 2, ce qui traduit une bonne qualité de la régression.

En outre, nous savons que les résidus sont d'espérance nulle. Nous avons vu aussi que la variance résiduelle était petite devant la variance totale. De plus, en calculant la matrice de covariance des points estimés par la surface de réponse (y_i^{SR}) et des résidus (e_i), on peut conclure à une indépendance des résidus vis-à-vis de la surface de réponse, puisque la covariance est très petite devant les variances :

$$\Sigma_{(y^{SR},e)} \simeq \begin{bmatrix} 2.5790 \times 10^{-25} & 4.3899 \times 10^{-32} \\ 4.3899 \times 10^{-32} & 1.8074 \times 10^{-27} \end{bmatrix}$$

Les résidus peuvent donc être modélisés par une variable aléatoire indépendante de celles de la surface de réponse, d'espérance nulle et de variance petite devant la variance de la surface de réponse.

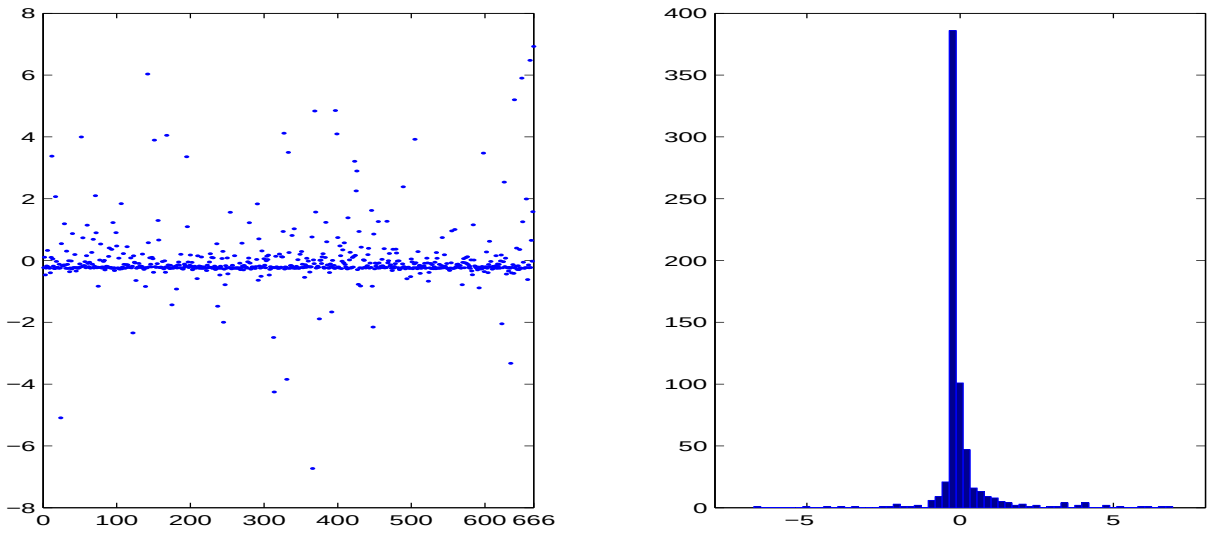


FIG. 2.4. : Résidus studentisés issus de l'ajustement de la surface de réponse sur la base de données de construction de GASCON (valeurs et histogramme)

Comme nous l'avons vu en section 2.2, il n'est donc pas utile de corriger l'estimation des indices de sensibilité en prenant en compte ces résidus. L'estimation des indices faite en utilisant la surface de réponse est tout à fait correcte.

Remarque. *L'étude des résidus présentée ci-dessus a été faite à partir de la base de données d'apprentissage de 666 points, ce qui peut conduire à sous-estimer ces résidus. Il est important de réaliser cette étude en parallèle sur la base de données de test (334 points). Nous ne le présentons pas ici, mais cela a été fait dans [29]. Les résultats sont légérement moins bons que pour la surface d'apprentissage (variance résiduelle légèrement supérieure et donc R^2 légérement inférieur :* 0.988*), mais conduisent néanmoins à accepter la validité de la régression.*

2.3.4. Analyse d'incertitude et de sensibilité de $Y = Ad\_I\_1$

À partir de la surface de réponse (2.13), une analyse d'incertitude est menée sur $Y = Ad\_I\_1$. L'estimation des deux premiers moments de Y est faite par Monte-Carlo avec 10000 simulations, cette estimation étant répétée 200 fois. Le choix de 10000 suffit à obtenir une estimation de ces moments avec une précision satisfaisante, comme le montre le tableau 2.4 (rapport écart-type sur moyenne inférieur à 10%).

moment	moyenne	écart-type	$\frac{\text{écart-type}}{\text{moyenne}}$
$\mathrm{E}[Y]$	2.0265×10^{-13}	6.7420×10^{-15}	0.0333
$\mathrm{V}(Y)$	3.9544×10^{-25}	3.8850×10^{-26}	0.0982

TAB. 2.4.: Estimation des deux premiers moments de Y.

Une analyse de sensibilité de $Y = Ad\_I\_1$ est réalisée par la méthode de Sobol. Comme la sensibilité d'une variable aléatoire est invariante par l'addition d'une constante, les analyses de sensibilité de Y et $Y - \alpha_0$ sont les mêmes. Ainsi, nous réaliserons par la suite les analyses de sensibilité de Y sur $Y - \alpha_0$, ce qui

permet d'omettre la constante α_0 dans (2.13). De la même manière que pour l'analyse d'incertitude, la taille des échantillons de Monte Carlo utilisé est de 10000. En répétant 200 fois cette analyse de sensibilité, on vérifie que cette taille de 10000 est suffisante à une estimation correcte des indices de sensibilité (écart-type jugé suffisamment petit devant la moyenne, inférieur à 10% pour les valeurs supérieure à 0.2). Les indices de sensibilité totaux et de premier ordre sont présentés dans la figure 2.5, et leur valeur numérique tableau 2.5. Les valeurs données correspondent aux moyennes et écart-types des estimations sur les 200 analyses de sensibilité. Dans le tableau 2.5, les valeurs sont considérées comme quasiment nulles ($\simeq 0$) lorsque la valeur moyenne est proche de 0 et plus petite que l'écart-type.
Toutes les analyses de sensibilité présentées dans cette section seront réalisées suivant ce protocole (200 répétitions, échantillons de Monte Carlo de taille 10000).

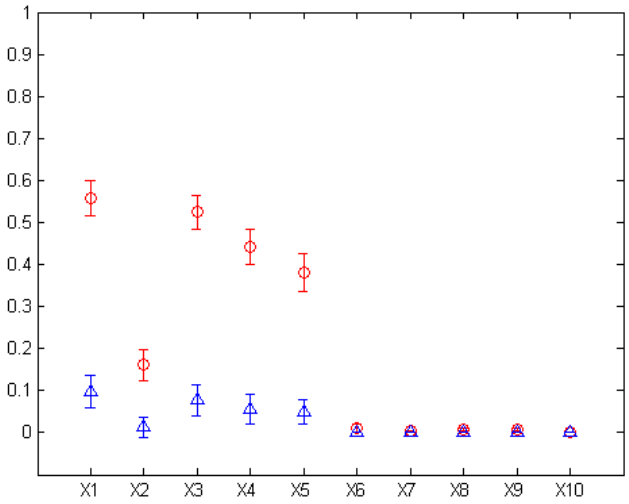


FIG. 2.5. : Indices de sensibilité totaux (○) et de premier ordre (△) du modèle (2.13) avec intervalles de dispersion à plus ou moins deux écarts-types.

Indices	moyenne	écart-type	Indices	moyenne	écart-type
S_{T_1}	0.558	0.042	S_1	0.097	0.019
S_{T_3}	0.524	0.041	S_3	0.077	0.018
S_{T_4}	0.442	0.043	S_4	0.056	0.017
S_{T_5}	0.381	0.044	S_5	0.050	0.015
S_{T_2}	0.161	0.036	S_2	0.014	0.012
S_{T_6}	0.012	0.010	S_6	$\simeq 0$	
S_{T_8}	0.008	0.004	S_8	$\simeq 0$	
S_{T_9}	0.007	0.003	S_9	$\simeq 0$	
S_{T_7}	0.003	0.002	S_7	$\simeq 0$	
$S_{T_{10}}$	$\simeq 0$		S_{10}	$\simeq 0$	

TAB. 2.5.: Indices de sensibilité du modèle (2.13).

L'analyse de sensibilité montre que la variance de la variable Ad_I_1 est due essentiellement aux variances de cinq variables, qui sont, par ordre d'importance, l'ingestion efficace de l'adulte (X_1), la ration alimen-

taire de l'adulte en lait de chèvre (X_3) et de la chèvre en herbe (X_4), la vitesse de dépôt sec en ^{129}I (X_5), et enfin, avec une importance plus faible, le facteur de transfert du lait de chèvre de ^{129}I (X_2). On remarque que toutes les variables importantes sont celles relatives à la chaîne alimentaire du lait de chèvre. Celles spécifiques à la chaîne alimentaire du lait de brebis n'ont pas d'importance sur la variance de Ad_I_1. Mais elles interviennent dans le calcul de sa valeur (cf. équation 2.13).
L'importance de la chaîne alimentaire du lait de chèvre peut être décelée en estimant séparément la variance de la partie du modèle (2.13) relative à cette chaîne alimentaire, et celle de la partie relative à la chaîne alimentaire du lait de brebis. En effet, nous verrons au paragraphe 2.3.5.2 que le modèle (2.13) peut être décomposé en deux sous modèles relatifs à ces deux chaînes alimentaires, et que la variance du sous modèle «chèvre» est près de 100 fois plus grande que celle du sous modèle «brebis».
Comme en témoignent les indices de premier ordre, l'influence des variables seules sur la variance de Ad_I_1 est relativement faible (de près de 10% pour X_1 à 1% pour X_2). Par contre, en prenant en compte les interactions entre variables avec les indices de sensibilité totaux, on constate que l'importance est plus forte (de 56% pour X_1 à 16% pour X_2). Les interactions entre les variables de la chaîne alimentaire du lait de chèvre jouent donc un rôle très important sur la variance de Ad_I_1. À noter que la hiérarchie sur les variables seules est la même que celle en prenant en compte les interactions. Cette conservation de la hiérarchie ainsi que l'importance des interactions sont logiques, puisque dans le modèle (2.13) les variables importantes apparaîssent toutes sous la forme de produit.

2.3.5. Mutations de la surface de réponse de Ad_I_1

Nous considérons que le modèle de calcul de la variable de sortie Ad_I_1 est la surface de réponse (2.13). Nous supposons que ce modèle subit un certain nombre de mutations, que nous décrivons ci-dessous. Ces mutations ont été choisies en fonction de deux paramètres : leur intérêt pratique et leur aptitude à illustrer les résultats théoriques de la section précédente.
Toutes les variables aléatoires du modèle avant mutation restent les mêmes après mutation sauf indications contraires.

Mutations transformant une ou plusieurs variables d'entrée.

- *Mutation 1.1* : nous supposons que la valeur de l'ingestion efficace chez l'adulte est connue avec assez de précision pour ne pas avoir à la considérer comme aléatoire. Nous décidons donc de fixer la variable X_1 à sa valeur nominale.
- *Mutation 1.2* : nous décidons de considérer que la ration de foin mangée par les moutons est fixe. En effet, les paysans donnent une quantité de foin précise pour chaque mouton, mais la quantité d'herbe mangée reste aléatoire. La variable X_{10} du modèle est donc fixée non pas à sa valeur nominale mais à son espérance (nous avons vu en section 2.1, comme ici X_{10} n'est pas multiplicatif de toutes les autres variables du modèle, contrairement à X_1, qu'il était nécessaire pour pouvoir traiter ce type de mutation que la valeur de la variable soit fixée à son espérance).
- *Mutation 1.3* : nous considérons comme modèle initial le modèle (2.13) où seules trois variables sont aléatoires : l'ingestion efficace de l'adulte (X_1), la ration alimentaire de l'adulte en lait de chèvre (X_3), et la vitesse de dépôt sec de ^{129}I (X_5). Ceci permet de calculer les indices de sensibilité à tout ordre. Nous envisageons alors la mutation qui consiste, comme pour la mutation 1.1, à fixer l'ingestion efficace de l'adulte.

Ajoût de deux modèles. La deuxième catégorie de mutation étudiée consiste à créer un nouveau modèle en sommant deux modèles existants.

Soient les deux modèles :

$$\begin{aligned} Y_1 &= \alpha_1 X_1 X_2 X_3 X_4 X_5 + \alpha_2 X_1 X_2 X_3 X_4 X_5^2 + \alpha_3 X_1 X_2 X_3 X_6 X_5 + \alpha_4 X_1 X_2 X_3 X_6 X_5^2, \\ \text{et} \\ Y_2 &= \alpha_5 X_1 X_7 X_8 X_9 X_5 + \alpha_6 X_1 X_7 X_8 X_9 X_5^2 + \alpha_7 X_1 X_7 X_8 X_{10} X_5. \end{aligned}$$

Le premier modèle correspond à la chaîne alimentaire du lait de chèvre, et le second à la chaîne alimentaire du lait de brebis. Nous supposons qu'une analyse de sensibilité a été menée sur chacun de ces deux modèles. La variable de sortie de GASCON qui nous intéresse, $Y = Ad\_I\_1$, est définie, à la constante α_0 près, comme la somme de ces deux modèles : $Y - \alpha_0 = Y_1 + Y_2$.
Nous distinguons dans un premier temps le cas où les variables X_1 et X_5 sont déterministes, ce qui rend Y_1 et Y_2 indépendantes (*Mutation 2.1*). Nous étudions ensuite le cas plus général où X_1 et X_5 sont aléatoires, c'est-à-dire où Y_1 et Y_2 sont deux modèles ayant deux variables communes (*Mutation 2.2*).

Multiplication de deux modèles. La dernière catégorie de mutation que nous étudions consiste encore à créer un nouveau modèle à partir de deux modèles existants, mais cette fois en les multipliant.
Soient les deux modèles :

$$\begin{aligned} Z_1 &= \alpha_1 X_2 X_4 X_5 + \alpha_2 X_2 X_4 X_5^2 + \alpha_3 X_2 X_6 X_5 + \alpha_4 X_2 X_6 X_5^2, \quad (2.14) \\ \text{et} \\ Z_2 &= X_1 X_3. \quad (2.15) \end{aligned}$$

Le modèle Z_1 est fonction des variables relatives aux chèvres et à l'Iode 129. Le modèle Z_2 est relatif à l'homme adulte du groupe de référence étudié. Le produit de ces deux modèles Z_1 et Z_2 forme le modèle $Y_1 = Z_1 Z_2$ précédemment introduit, qui correspond à la chaîne alimentaire du lait de chèvre.

Pour chacune de ces mutations, nous voulons obtenir les indices de sensibilité du modèle formé après mutation, sans avoir à refaire une analyse de sensibilité classique, c'est-à-dire en utilisant les résultats démontrés section 2.1.

2.3.5.1. Mutation transformant une variable d'entrée

Mutation 1.1
Le modèle (2.13), dont l'analyse de sensibilité est présentée au paragraphe 2.3.4 mute en :

$$\begin{aligned} Y' - \alpha_0 &= 1.1 \times 10^{-7} \alpha_1 X_2 X_3 X_4 X_5 + 1.1 \times 10^{-7} \alpha_2 X_2 X_3 X_4 X_5^2 + 1.1 \times 10^{-7} \alpha_3 X_2 X_3 X_6 X_5 \\ &\quad + 1.1 \times 10^{-7} \alpha_4 X_2 X_3 X_6 X_5^2 + 1.1 \times 10^{-7} \alpha_5 X_7 X_8 X_9 X_5 + 1.1 \times 10^{-7} \alpha_6 X_7 X_8 X_9 X_5^2 \\ &\quad + 1.1 \times 10^{-7} \alpha_7 X_7 X_8 X_{10} X_5, \quad (2.16) \end{aligned}$$

Nous sommes dans le cas de figure développé au paragraphe 2.1.1.2. Tous les indices de sensibilité, relatifs à la variable X_1 disparaissent. Les indices de sensibilité de premier ordre de (2.16) peuvent être obtenus en multipliant directement ceux de (2.13) par :

$$\frac{\mathrm{V}(Y)}{\mathrm{V}(Y')} \left(\frac{1.1 \times 10^{-7}}{\mathrm{E}[X_1]} \right)^2 .$$

Connaissant des estimations de $\mathrm{V}(Y)$ et $\mathrm{E}[X_1]$ (cf. tableau 2.4), il nous suffit donc d'estimer la variance du nouveau modèle Y' pour obtenir ses indices de sensibilité. Par le même protocole que celui utilisé pour

estimer $\mathrm{V}(Y)$ (200 calculs de Monte-Carlo à 10000 points), on obtient :

$$\hat{\mathrm{V}}(Y^{'}) = 1.9152 \times 10^{-26}.$$

Il nous faut donc multiplier les indices de (2.13) par :

$$\frac{3.9544 \times 10^{-25}}{1.9152 \times 10^{-26}} \left(\frac{1.1 \times 10^{-7}}{3.3275 \times 10^{-7}} \right)^2 \simeq 2.2564. \tag{2.17}$$

Cette valeur n'a pas d'interprétation pratique.
Les indices de sensibilité du modèle initial ont été estimés avec une certaine incertitude (écarts-types des estimations). L'incertitude est alors elle aussi multipliée par cette constante $\gamma = 2.2564$.
En effet, si on note $(S_i^1, \ldots, S_i^{200})$ les 200 estimations d'un indice de sensibilité S_i, l'écart-type des estimations S_i^k multiplié par γ (γS_i^k) est égal à γ fois celui des estimations S_i^k :

$$\sigma_{\gamma S_i^k} = \sqrt{\frac{\sum_{k=1}^{200} (\gamma S_i^k - \overline{\gamma S_i^k})^2}{200}} = \gamma \sqrt{\frac{\sum_{k=1}^{200} (S_i^k - \overline{S_i^k})^2}{200}} = \gamma \sigma_{S_i^k}.$$

Nous présentons les indices de sensibilité de (2.16) obtenus par cette méthode ($\times$ rouge), ainsi que ceux obtenus par une nouvelle analyse de sensibilité ($+$ bleu), dans la figure 2.6.

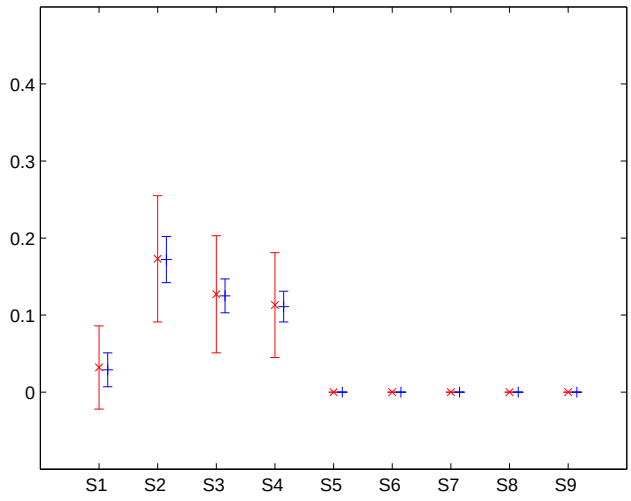


FIG. 2.6. : Indices de sensibilité de premier ordre du modèle (2.17) par mutation ($\times$ rouge) et par une nouvelle analyse ($+$ bleu).

Les deux résultats coïncident d'un point de vue de leur valeur moyenne. Les indices de sensibilité après mutation sont plus grands que ceux avant mutation (coefficient multiplicatif γ supérieur à 1). Même si l'incertitude relative sur l'estimation des indices est la même avant et après mutation, l'incertitude absolue devient elle plus grande après mutation car elle est multipliée par γ. Comme l'incertitude d'une nouvelle analyse est équivalente à celle de l'analyse initiale, l'incertitude issue de l'estimation par la méthode des mutations est plus importante que celle d'une nouvelle analyse. Le comportement inverse aurait été observé si les indices de sensibilité après mutation avaient été plus petits que ceux avant mutations (cas d'un coeffi-

cient multiplicatif γ inférieur à 1, cf. mutation inverse considérée en remarque à la fin de ce paragraphe).
L'analyse de sensibilité du modèle muté (2.16) exhibe la même hiérarchie d'importance au sein des variables d'entrée que celle du modèle initial (en supprimant la variable X_1). Effectivement, en multipliant tous les indices par une constante plus grande que 1, seules les valeurs des indices de sensibilité augmentent mais pas les différences relatives entre indices.
Les indices de sensibilité de premier ordre du modèle muté (2.16) ont donc été obtenus à moindre coût (estimation d'une variance), en utilisant l'analyse faite sur la version initiale du modèle. Par contre, pour les indices de sensibilité totaux, comme ils ont été estimés directement dans l'analyse de (2.13), nous ne pouvons les obtenir que par une nouvelle analyse de sensibilité, qui est présentée tableau 2.6.

Indices	moyenne	écart-type
S_{T_3}	0.577	0.028
S_{T_4}	0.484	0.025
S_{T_5}	0.427	0.028
S_{T_2}	0.176	0.026
S_{T_6}	0.013	0.006
S_{T_8}	0.008	0.003
S_{T_9}	0.007	0.002
S_{T_7}	0.003	0.002
$S_{T_{10}}$	$\simeq 0$	

TAB. 2.6.: Indices de sensibilité totaux du modèle (2.16).

On remarque que la hiérarchie des variables d'entrée est encore conservée pour les indices totaux. Alors que les indices de premier ordre du modèle muté sont égaux à 2.2564 fois ceux du modèle initial, les indices totaux sont eux égaux approximativement à 1.1 fois ceux du modèle initial. Fixer la variable X_1 augmente donc proportionnellement plus la sensibilité de premier ordre aux autres variables que la sensibilité aux interactions.

Remarque. *Mutation inverse*.
On considère désormais la mutation inverse de la précédente. Le modèle (2.16) où X_1 est fixé mute en le modèle initial (2.13).
Pour cette mutation, traitée section 2.1.1.3, il est possible d'obtenir les indices de sensibilité de premier ordre de Y (2.13) à partir de ceux de $Y^{'}$ (2.16), en multipliant ces derniers par :

$$\frac{\hat{V}(Y')}{\hat{V}(Y)} \left(\frac{E[X_1]}{1.1 \times 10^{-7}} \right)^2 \simeq 0.4432,$$

qui est naturellement l'inverse du coefficient (2.17) utilisé pour la mutation précédente. Nous retomberons donc naturellement sur les indices de (2.13) présentés dans le tableau 2.5, à l'exception des indices relatifs à X_1 qu'il sera nécessaire d'estimer par une nouvelle analyse. Les indices de sensibilité totaux devront être estimés par une nouvelle analyse puisqu'ils ont été estimés directement et non comme la somme des indices à tout ordre.

Mutation 1.2

Dans la mutation précédente, c'est la variable la plus importante, X_1, qui a été fixée à sa valeur nominale. Nous examinons ce qu'il se passe si on fixe une variable qui n'a pas d'importance sur la variance de la variable de sortie $Ad\_I\_1 : X_{10}$ (vitesse de dépôt sec de l'iode 129).

Le modèle (2.13) mute donc en :

$$\begin{aligned} Y^{'} - \alpha_0 &= \alpha_1 X_1 X_2 X_3 X_4 X_5 + \alpha_2 X_1 X_2 X_3 X_4 X_5^2 + \alpha_3 X_1 X_2 X_3 X_6 X_5 + \alpha_4 X_1 X_2 X_3 X_6 X_5^2 \\ &+ \alpha_5 X_1 X_7 X_8 X_9 X_5 + \alpha_6 X_1 X_7 X_8 X_9 X_5^2 + 3315.4 \alpha_7 X_1 X_7 X_8 X_5, \end{aligned} \quad (2.18)$$

où $3315.4 = \mathrm{E}[X_{10}]$.
L'analyse de sensibilité de (2.13) a montré que la variable X_{10} n'était d'aucune influence sur la variance de Y (indice total quasi-nul). On peut alors inférer que les indices de sensibilité du nouveau modèle ne changeront pas. En effet, outre les indices de sensibilité relatifs à la variable X_{10} qui disparaissent, les autres indices de sensibilité de premier ordre sont obtenus, comme pour la mutation précédente, en multipliant ceux de (2.13) par :

$$\frac{\hat{\mathrm{V}}(Y)}{\hat{\mathrm{V}}(Y')} = \frac{3.9785 \times 10^{-25}}{3.9699 \times 10^{-25}} \simeq 1.0022,$$

où $\mathrm{V}(Y^{'})$ est estimé toujours par le même protocole. Comme prévu, le coefficient multiplicatif étant quasiment égal à 1, les indices de sensibilité du nouveau modèle seront approximativement les mêmes que ceux du modèle original.
Les indices de sensibilité totaux devront être estimés par une nouvelle analyse, puisque nous ne disposons pas des indices de sensibilité du modèle initial à tout ordre. Mais comme pour les indices de premier ordre, l'indice total de sensibilité à X_{10} étant quasi nul, les indices de sensibilité totaux du modèle (2.18) devraient être approximativement les mêmes que ceux de (2.13).

Remarque. *Mutation inverse*.
Considérons la mutation inverse consistant à muter le modèle (2.16) (où X_{10} est fixé) en le modèle initial (2.13). Le comportement est le même que pour la mutation précédente. Les indices de sensibilité de premier ordre du modèle (2.13) peuvent être obtenus en multipliant ceux du modèle (2.18) par :

$$\frac{\hat{V}(Y')}{\hat{V}(Y)} = \frac{3.9699 \times 10^{-25}}{3.9785 \times 10^{-25}} \simeq 0.9978,$$

ce qui ne changera évidement pas significativement les valeurs des indices, puisque ce coefficient est très proche de 1. Le nouvel indice S_{10} devra quant à lui être estimé par une nouvelle analyse, ainsi que tous les indices totaux, afin de prendre en compte les interactions avec X_{10}.

Mutation 1.3

Soit le modèle (2.13), au sein duquel nous supposons que seules les variables X_1, X_3 et X_5 sont aléatoires (les autres variables sont supposées déterministes et fixées à leur espérance). Ce modèle peut s'écrire :

$$\begin{aligned} Y - \alpha_0 &= \alpha_1^{'} X_1 X_3 X_5 + \alpha_2^{'} X_1 X_3 X_5^2 + \alpha_3^{'} X_1 X_3 X_5 + \alpha_4^{'} X_1 X_3 X_5^2 \\ &+ \alpha_5^{'} X_1 X_5 + \alpha_6^{'} X_1 X_5^2 + \alpha_7^{'} X_1 X_5. \end{aligned} \quad (2.19)$$

Les coefficients $\alpha_i^{'}$ s'expriment en fonction de ceux du modèle (2.13) (*c.f.* tableau 2.7) :

$\alpha_0^{'} = \alpha_0$	$\alpha_4^{'} = \alpha_4 \times \mathrm{E}[X_2]\mathrm{E}[X_6]$
$\alpha_1^{'} = \alpha_1 \times \mathrm{E}[X_2]\mathrm{E}[X_4]$	$\alpha_5^{'} = \alpha_5 \times \mathrm{E}[X_7]\mathrm{E}[X_8]\mathrm{E}[X_9]$
$\alpha_2^{'} = \alpha_2 \times \mathrm{E}[X_2]\mathrm{E}[X_4]$	$\alpha_6^{'} = \alpha_6 \times \mathrm{E}[X_7]\mathrm{E}[X_8]\mathrm{E}[X_9]$
$\alpha_3^{'} = \alpha_3 \times \mathrm{E}[X_2]\mathrm{E}[X_6]$	$\alpha_7^{'} = \alpha_7 \times \mathrm{E}[X_7]\mathrm{E}[X_8]\mathrm{E}[X_{10}]$

TAB. 2.7.: Coefficients du modèle 2.19.

Comme ce modèle ne comporte que trois variables d'entrée, il est possible de calculer les indices de sensibilité à tout ordre. Les valeurs de ces indices sont présentées dans le tableau 2.8.

Indices	moyenne	écart-type	Indices totaux	moyenne	écart-type
S_1	0.220	0.015	S_{T_1}	0.616	0.025
S_3	0.177	0.016	S_{T_3}	0.551	0.018
S_5	0.115	0.015	S_{T_5}	0.427	0.025
S_{13}	0.176	0.020			
S_{15}	0.115	0.013			
S_{35}	0.092	0.019			
S_{135}	0.105	0.018			
somme	1.000				

TAB. 2.8.: Indices de sensibilité de (2.19)

Nous décidons de fixer la variable X_1 (ingestion efficace chez l'adulte) à 1.1×10^{-7}, sa valeur nominale. Nous formons ainsi le modèle :

$$\begin{aligned} Y' - \alpha_0 = 1.1 \times 10^{-7} \quad (\quad & \alpha_1' X_3 X_5 + \alpha_2' X_3 X_5^2 + \alpha_3' X_3 X_5 + \alpha_4' X_3 X_5^2 \\ + \quad & \alpha_5' X_5 + \alpha_6' X_5^2 + \alpha_7' X_5 \quad). \end{aligned} \tag{2.20}$$

Cette mutation, analogue à celles 1.1 et 1.2, est abordée dans le paragraphe 2.1.1.2. Les indices de sensibilité de Y' de premier ordre et d'ordre supérieur s'obtiennent en multipliant ceux de Y par :

$$\frac{\mathrm{V}(Y)}{\mathrm{V}(Y')} \left(\frac{1.1 \times 10^{-7}}{\mathrm{E}[X_1]} \right)^2 .$$

Il suffit donc d'estimer la variance du nouveau modèle :

$$\hat{\mathrm{V}}(Y') = 7.3580 \times 10^{-27},$$

(protocole habituel), et de multiplier les indices de sensibilité de Y par :

$$\frac{1.7657 \times 10^{-25}}{7.3580 \times 10^{-27}} \left(\frac{1.1 \times 10^{-7}}{3.3275 \times 10^{-7}} \right)^2 \simeq 2.6226.$$

Dans ce calcul, nous disposons des indices à tout ordre, il est donc possible d'estimer les indices totaux de Y' en sommant ses indices de sensibilité à tout ordre.

Les indices de sensibilité obtenus ainsi que ceux obtenus par une nouvelle analyse, présentés par la figure 2.7, coïncident d'un point de vue de leur valeur moyenne. Comme pour la mutation 1.1, les incertitudes sur les estimations par mutation sont plus grandes que celles par une nouvelle analyse, puisqu'elles correspondent à 2.6226 fois les incertitudes issues de l'analyse initiale. Pour les indices de sensibilité totaux, il n'est pas possible d'estimer l'incertitude de leur estimation de façon correcte, puisqu'ils sont estimés comme la somme de deux estimations non indépendantes (indices d'ordre un et deux). Effectivement les estimations des indices de premier ordre et de second ordre ne sont pas indépendantes, puisque l'indice de deuxième ordre S_{23} est défini comme la part de variance due au couple (X_2, X_3) moins la part de variance due à X_2 et à X_3, sur la variance totale.
La variable X_1 était la plus importante d'un point de vue variance dans le modèle initial. La fixer augmente les indices de sensibilité de toutes les autres variables, sans modifier leur hiérarchie. Si l'on avait fixé les autres variables, X_3 ou X_5, un comportement similaire aurait été observé, c'est-à-dire une augmentation

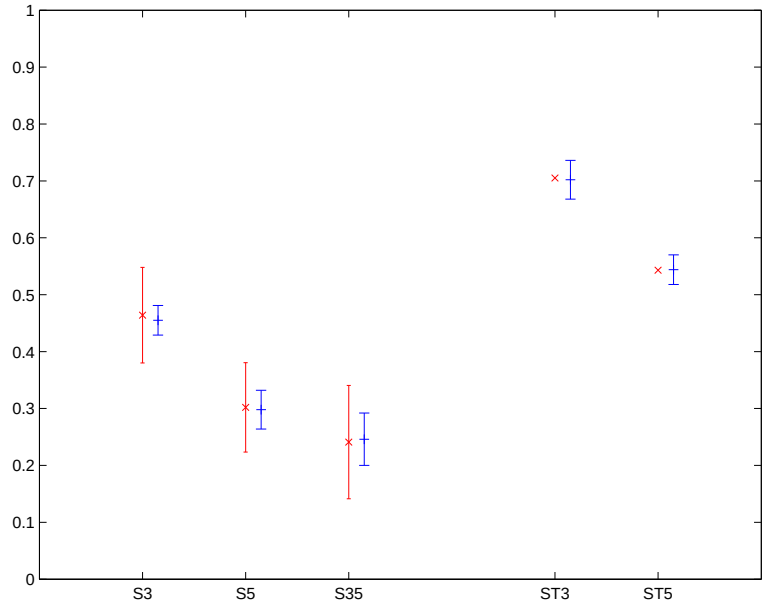


FIG. 2.7. : Indices de sensibilité de premier ordre du modèle (2.17) par mutation ($\times$ rouge) et par une nouvelle analyse ($+$ bleu).

des valeurs des indices de sensibilité et une conservation de la hiérarchie, mais dans une moindre mesure puisque leur importance était plus faible que celle de X_1.

Remarque. *Mutation inverse.*
Considérons finalement la mutation inverse de celle étudiée ci-dessus. Le modèle (2.20), où X_1 est déterministe, mute en (2.19), où X_1 est aléatoire. Cette mutation, étudiée au paragraphe 2.1.1.3, permet d'obtenir les indices de sensibilité de (2.19) en multipliant ceux de (2.20) par :

$$\frac{\hat{V}(Y')}{\hat{V}(Y)}\left(\frac{E[X_1]}{1.1\times 10^{-7}}\right)^2 \simeq 0.3813,$$

sauf pour les indices relatifs à X_1 qui devront être estimés par une nouvelle analyse de sensibilité. Les indices totaux seront alors estimés en sommant les indices à tout ordre ainsi obtenus. On retrouve alors de façon évidente les résultats de sensibilité du modèle initial (2.19) présenté par le tableau 2.8.

2.3.5.2. Ajoût de deux modèles

Soient les deux modèles :

$$\begin{aligned} Y_1 &= \alpha_1 X_1X_2X_3X_4X_5 + \alpha_2 X_1X_2X_3X_4X_5^2 + \alpha_3 X_1X_2X_3X_6X_5 \\ &\quad +\alpha_4 X_1X_2X_3X_6X_5^2, \end{aligned} \tag{2.21}$$

et

$$Y_2 = \alpha_5 X_1X_7X_8X_9X_5 + \alpha_6 X_1X_7X_8X_9X_5^2 + \alpha_7 X_1X_7X_8X_{10}X_5. \tag{2.22}$$

Le modèle (2.21) est défini par la chaîne alimentaire du lait de chèvre, et (2.22) par celle du lait de brebis. Une analyse de sensibilité a été faite sur chacun de ces modèles.
Le modèle (2.13) relatif à la variable aléatoire $Y = Ad\_I\_1$ qui nous intéresse est égal à la somme de ces

deux modèles : $Y - \alpha_0 = Y_1 + Y_2$. Comme ils ont des variables communes, on commence par les fixer afin d'avoir deux modèles à variables distinctes.

Mutation 2.1 : Y_1 et Y_2 n'ont pas de variables communes

Nous considérons donc les variables X_1 et X_5 déterministes (fixées à leurs valeur nominales).
Les tableaux 2.9 et 2.10 présentent les indices de sensibilité de Y_1 et Y_2.

Indices	moyenne	écart-type	Indices	moyenne	écart-type
S_2	0.055	0.008	S_{T_2}	0.203	0.011
S_3	0.316	0.013	S_{T_3}	0.664	0.013
S_4	0.229	0.011	S_{T_4}	0.551	0.014
S_6	0.007	0.007	S_{T_6}	0.017	0.003

TAB. 2.9.: Indices de sensibilité de Y_1.

Indices	moyenne	écart-type	Indices	moyenne	écart-type
S_7	0.065	0.01	S_{T_7}	0.251	0.01
S_8	0.274	0.016	S_{T_8}	0.641	0.015
S_9	0.233	0.012	S_{T_9}	0.587	0.018
S_{10}	$\simeq 0$		$S_{T_{10}}$	0.005	0.002

TAB. 2.10.: Indices de sensibilité de Y_2.

Pour cette mutation, étudiée au paragraphe 2.1.2.2, il est possible d'obtenir les indices de sensibilité (premier ordre et totaux) du modèle somme par :

$$S'_j = \begin{cases} S_j \times \frac{\mathrm{V}(Y_1)}{\mathrm{V}(Y_1)+\mathrm{V}(Y_2)} & \text{si } j = 2,4,6 \\ S_j \times \frac{\mathrm{V}(Y_2)}{\mathrm{V}(Y_1)+\mathrm{V}(Y_2)} & \text{si } j = 7,8,9,10. \end{cases}$$

Nous n'avons donc besoin que de la variance de Y_1 et de Y_2, qui ont pu être estimées grâce aux analyses de sensibilité de Y_1 et Y_2 :

$$\hat{\mathrm{V}}(Y_1) = 2.2569 \times 10^{-26}, \qquad \hat{\mathrm{V}}(Y_2) = 3.814 \times 10^{-28},$$

d'où

$$\frac{\hat{\mathrm{V}}(Y_1)}{\hat{\mathrm{V}}(Y_1) + \hat{\mathrm{V}}(Y_2)} = 0.9834, \qquad \frac{\hat{\mathrm{V}}(Y_2)}{\hat{\mathrm{V}}(Y_1) + \hat{\mathrm{V}}(Y_2)} = 0.0166.$$

Ainsi, il est donc possible à partir de l'analyse sur les deux parties indépendantes d'un modèle de déduire les indices de sensibilité de leur somme, et ce sans calculs supplémentaires.
Les résultats obtenus par mutation ($\times$ rouge) et par une nouvelle analyse ($+$ bleu) sont présentés aux figures 2.8 et 2.9.

Etant donné que la variance du modèle Y_2 est 100 fois plus petite que celle de Y_1, la variance de leur somme est essentiellement due à celle de Y_1. Ainsi, seules les variables du modèle Y_1 ont un rôle non négligeable sur la variance de Y. Cela correspond à la conclusion de l'analyse initiale du modèle (2.13) qui révélait que seule la chaîne alimentaire du lait de chèvre (Y_1) jouait un rôle sur la variance de Y.

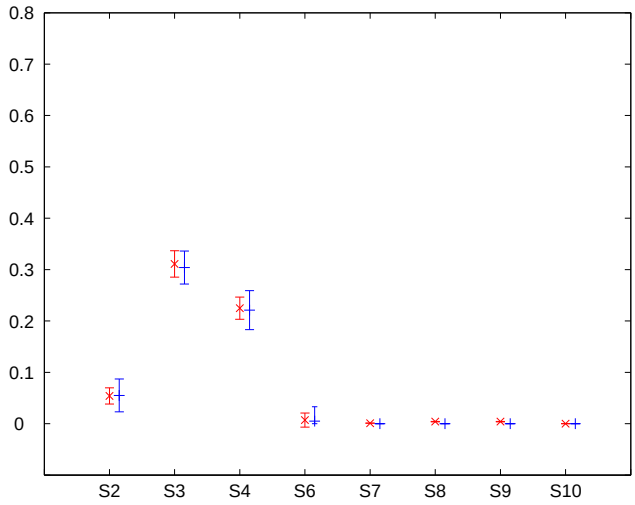


FIG. 2.8. : Indices de sensibilité de premier ordre du modèle «somme» par mutation ($\times$ rouge) et par une nouvelle analyse ($+$ bleu).

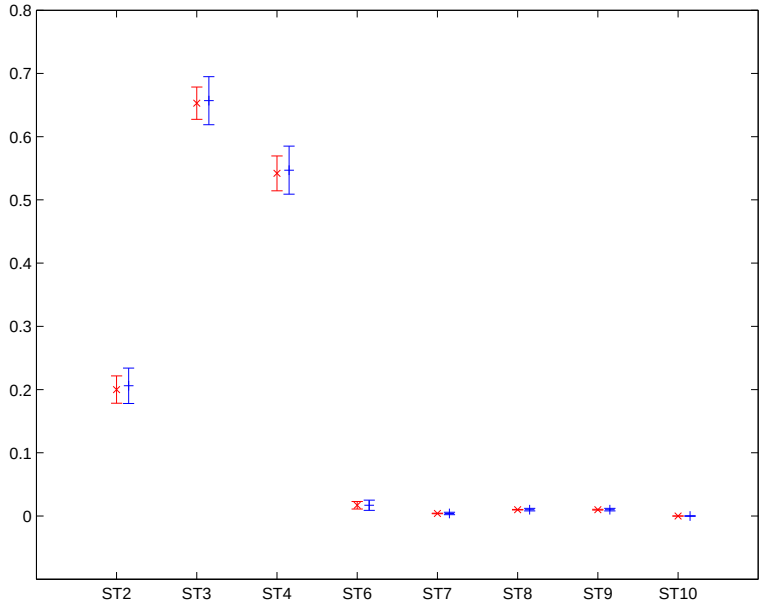


FIG. 2.9. : Indices de sensibilité totaux du modèle «somme» par mutation ($\times$ rouge) et par une nouvelle analyse ($+$ bleu).

L'incertitude d'estimation est approximativement identique par la méthode des mutations et par la nouvelle analyse. Ceci est dû au fait que le coefficient multiplicatif des indices non nuls (donc de Y_1) est quasiment égal à 1.

Mutation 2.2 : Y_1 et Y_2 ont deux variables communes

Nous ne supposons plus que les variables X_1 et X_5 sont déterministes. Les deux modèles Y_1 et Y_2 ont donc deux variables communes.

Les analyses de sensibilité de Y_1 et Y_2 donnent les résultats des tables 2.11 et 2.12.

Indices	moyenne	écart-type	Indices	moyenne	écart-type
S_1	0.084	0.018	S_{T_1}	0.551	0.044
S_2	0.016	0.013	S_{T_2}	0.172	0.040
S_3	0.083	0.018	S_{T_3}	0.546	0.045
S_4	0.062	0.018	S_{T_4}	0.460	0.042
S_5	0.045	0.015	S_{T_5}	0.379	0.043
S_6	$\simeq 0$		S_{T_6}	0.012	0.009

TAB. 2.11.: Indices de sensibilité de Y_1.

Indices	moyenne	écart-type	Indices	moyenne	écart-type
S_1	0.079	0.019	S_{T_1}	0.548	0.048
S_5	0.040	0.016	S_{T_5}	0.367	0.048
S_7	0.018	0.013	S_{T_7}	0.210	0.048
S_8	0.079	0.019	S_{T_8}	0.552	0.046
S_9	0.060	0.019	S_{T_9}	0.462	0.045
S_{10}	$\simeq 0$		$S_{T_{10}}$	0.013	0.010

TAB. 2.12.: Indices de sensibilité de Y_2.

Nous avons vu en étudiant ce type de mutation au paragraphe 2.1.2.2, qu'il n'est possible d'obtenir que les indices de sensibilité (premier ordre et totaux) relatifs aux variables qui n'appartiennent qu'à un seul des deux modèles, et ce par :

$$S_j' = \begin{cases} S_j \times \frac{\mathrm{V}(Y_1)}{\mathrm{V}(Y)} & \text{si } j = 2, 4, 6 \\ S_j \times \frac{\mathrm{V}(Y_2)}{\mathrm{V}(Y)} & \text{si } j = 7, 8, 9, 10. \end{cases}$$

Les deux analyses de sensibilité sur Y_1 et Y_2 ont permis entre autre d'estimer les variances de Y_1 et Y_2 :

$$\hat{\mathrm{V}}(Y_1) = 3.7950 \times 10^{-25}, \qquad \hat{\mathrm{V}}(Y_2) = 5.7048 \times 10^{-27},$$

d'où

$$\frac{\hat{\mathrm{V}}(Y_1)}{\hat{\mathrm{V}}(Y)} = 0.9597, \qquad \frac{\hat{\mathrm{V}}(Y_2)}{\hat{\mathrm{V}}(Y)} = 0.0144.$$

Nous pouvons donc obtenir les indices de sensibilité de premier ordre et totaux pour toutes les variables sauf X_1 et X_5, pour lesquels il est nécessaire d'estimer leurs indices par une nouvelle analyse. Les résultats sont présentés aux figures 2.10 et 2.11.

Comme dans le cas précédent où ces modèles étaient supposés sans variables communes, la variance de Y_1 est près de 100 fois plus grande que celle de Y_2. Ce sont donc encore les variables de Y_1 (X_1 à X_6) qui contribuent le plus à la variance de Y.

Remarque. *Estimation des indices relatifs aux variables X_1 et X_5.*
Une remarque intéressante à noter concerne les valeurs des indices de sensibilité aux variables X_1 et X_5

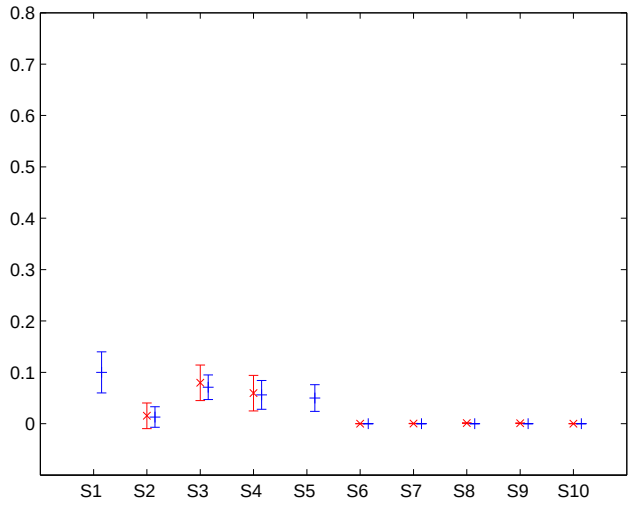


FIG. 2.10. : Indices de sensibilité de premier ordre du modèle somme par mutation ($\times$ rouge) et par une nouvelle analyse ($+$ bleu).

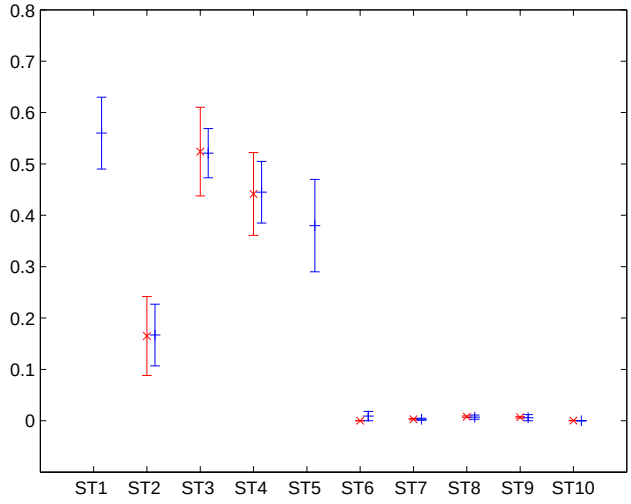


FIG. 2.11. : Indices de sensibilité totaux du modèle somme par mutation ($\times$ rouge) et par une nouvelle analyse ($+$ bleu).

des modèles Y_1, Y_2 et leur somme (tableau 2.13).
On constate que les indices de sensibilité de Y sont plus grands que ceux de Y_1 et de Y_2, et ce pour les deux variables X_1 et X_5, au premier ordre et à l'ordre total. Ceci peut être expliqué par la propriété démontrée au paragraphe 2.1.2.2 :

$$S_1 = S_1^{Y_1} \times \frac{V(Y_1)}{V(Y)} + S_1^{Y_2} \times \frac{V(Y_2)}{V(Y)} + \frac{2Cov(E[Y_1|X_j], E[Y_2|X_j])}{V(Y)}.$$

Ainsi, il n'est pas surprenant que S_1 puisse être supérieur à la fois à $S_1^{Y_1}$ et $S_1^{Y_2}$.

Indices	moyenne	écart-type	Indices total	moyenne	écart-type
S_1	0.097	0.020	S_{T_1}	0.555	0.045
$S_1^{Y_1}$	0.084	0.018	$S_{T_1}^{Y_1}$	0.551	0.044
$S_1^{Y_2}$	0.079	0.019	$S_{T_1}^{Y_2}$	0.548	0.048
S_5	0.052	0.016	S_{T_5}	0.387	0.043
$S_5^{Y_1}$	0.045	0.016	$S_{T_5}^{Y_1}$	0.379	0.043
$S_5^{Y_2}$	0.040	0.016	$S_{T_5}^{Y_2}$	0.367	0.048

TAB. 2.13.: Indices de sensibilité du modèle somme Y, de Y_1 et de Y_2 aux variables X_1 et X_5.

2.3.5.3. Multiplication de deux modèles

Considérons finalement les deux modèles suivants :

$$Z_1 = \alpha_1 X_2 X_4 X_5 + \alpha_2 X_2 X_4 X_5^2 + \alpha_3 X_2 X_6 X_5 + \alpha_4 X_2 X_6 X_5^2, \tag{2.23}$$

$$Z_2 = X_1 X_3, \tag{2.24}$$

où Z_1 est fonction des variables relatives à la chaîne alimentaire de la chèvre et à l'Iode 129 (facteur de transfert au lait de chèvres, rations en herbe et en foin de la chèvre et vitesse de dépôt sec de l'iode 129), et Z_2 de celles relatives à l'homme adulte du groupe de référence étudié (ingestion efficace et ration alimentaire en lait de chèvre).
Le produit de ces deux modèles Z_1 et Z_2 forme le modèle $Y_1 = Z_1 Z_2$ qui correspond à la chaîne alimentaire du lait de chèvre.

Les analyses de sensibilité des modèles Z_1 et Z_2 sont présentées par les tableaux 2.14 et 2.15.

Indices	moyenne	écart-type	Indices	moyenne	écart-type
S_2	0.083	0.008	S_{T_2}	0.214	0.018
S_4	0.353	0.024	S_{T_4}	0.633	0.010
S_5	0.236	0.017	S_{T_5}	0.490	0.013
S_6	$\simeq 0$		S_{T_6}	0.014	0.001

TAB. 2.14.: Indices de sensibilité de Z_1.

Indices	moyenne	écart-type	Indices	moyenne	écart-type
S_1	0.344	0.018	S_{T_1}	0.677	0.013
S_3	0.322	0.013	S_{T_3}	0.656	0.018

TAB. 2.15.: Indices de sensibilité de Z_2.

Ce type de mutation, étudié au paragraphe 2.1.2.1, permet d'obtenir les indices de sensibilité du modèle produit Y_1, par :

$$S_j = \begin{cases} \mathrm{E}[Z_2]^2 S_j \times \dfrac{\mathrm{V}(Z_1)}{\mathrm{V}(Y_1)} & \text{si} \quad j = 2, 4, 5 \text{ ou } 6 \\ \mathrm{E}[Z_1]^2 S_j \times \dfrac{\mathrm{V}(Z_2)}{\mathrm{V}(Y_1)} & \text{si} \quad j = 1 \text{ ou } 3. \end{cases} \tag{2.25}$$

L'espérance et la variance de Z_1 et Z_2 sont estimées lors de leur analyse de sensibilité présentée ci-dessus :

$$\begin{aligned} \hat{\mathrm{E}}[Z_1] &= 1.5501 \times 10^{-9}, \qquad \hat{\mathrm{E}}[Z_2] = 1.1041 \times 10^{-4}, \\ \hat{\mathrm{V}}(Z_1) &= 5.5010 \times 10^{-18}, \qquad \hat{\mathrm{V}}(Z_2) = 3.8644 \times 10^{-8}, \end{aligned}$$

tandis que la variance de Y_1 a été estimée lors des mutations précédente :

$$\hat{\mathrm{V}}(Y_1) = 3.7950 \times 10^{-25}.$$

En multipliant donc les indices de sensibilité de premier ordre de Z_1 par :

$$\hat{\mathrm{E}}[Z_2]^2 \times \frac{\hat{\mathrm{V}}(Z_1)}{\hat{\mathrm{V}}(Y_1)} = 0.1767,$$

et ceux de Z_2 par :

$$\hat{\mathrm{E}}[Z_1]^2 \times \frac{\hat{\mathrm{V}}(Z_2)}{\hat{\mathrm{V}}(Y_1)} = 0.2447,$$

on obtient les indices du modèle Y_1.
La figure 2.12 présente les indices de premier ordre ainsi obtenus, et les compare avec ceux d'une nouvelle analyse de sensibilité.

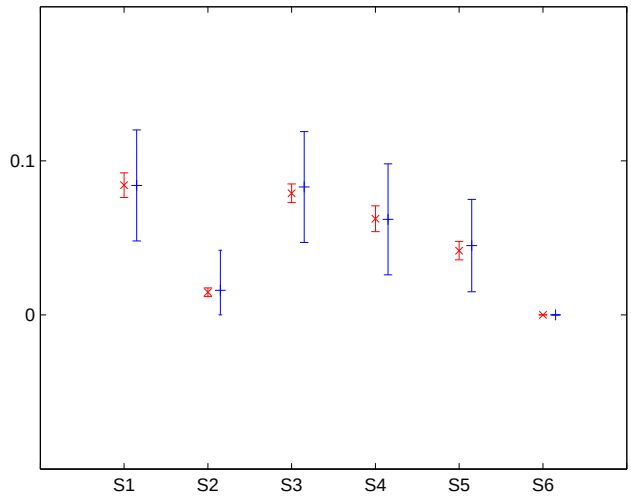


FIG. 2.12. : Indices de sensibilité de premier ordre du modèle produit, par mutation ($\times$ rouge) et par une nouvelle analyse ($+$ bleu).

Cette fois, l'incertitude est plus petite par mutation, puisque les valeurs des indices issues des analyses de Y_1 et Y_2 sont multipliées par des coefficients plus petits que 1.
Quant aux indices de sensibilité totaux, ils ne peuvent pas être obtenus par mutation, car ils ont été estimés directement. Si tous les indices d'ordre intermédiaire avaient été estimés, il serait alors possible d'estimer par cette méthode des mutations les indices de sensibilité à tout ordre du modèle produit (même ceux relatifs aux interactions entre variables des deux sous modèles), et ainsi en les sommant il serait possible d'obtenir une estimation des indices de sensibilité totaux.

2.3.6. Intérêts pratiques de diviser un modèle en sous-modèles

Les deux applications précédentes consistent, à partir d'analyses de sensibilité sur deux modèles, à déduire l'analyse du modèle «somme» ou du modèle «produit». Nous avons démontré en section 2.1 sous quelles conditions cela était possible.
Abordons ce problème sous un autre aspect : supposons qu'une analyse de sensibilité sur un certain modèle Y doit être réalisée. Supposons qu'il soit possible de décomposer ce modèle Y en une somme ou un produit de deux sous modèles orthogonaux Y_1 et Y_2, c'est-à-dire fonction de variables différentes. Dans ce cas, sachant que le nombre de simulations de Monte Carlo nécessaires (et le temps de calcul associé) aux estimations des indices de sensibilité dépend de la complexité du modèle, il peut être intéressant de faire l'analyse sur les deux sous modèles séparément, qui sont naturellement moins complexes.

Exemple : considérons l'exemple suivant où le modèle Y peut se séparer additivement en deux sous modèles Y_1 et Y_2 :

$$Y = \sum_{i=1}^{20} X_i = \underbrace{\sum_{i=1}^{10} X_i}_{Y_1} + \underbrace{\sum_{i=11}^{20} X_i}_{Y_2} .$$

L'estimation des indices de sensibilité de premier ordre de Y par la méthode de Sobol, utilisant des échantillons de Monte Carlo de taille 100000, demande 4.8 secondes (Pentium IV à 3GHz). Nous la réalisons 200 fois afin d'obtenir une idée de la précision de cette estimation. Les indices estimés ont une valeur moyenne approximativement égale à 0.05, et l'écart-type sur les 200 répétitions vaut pour chaque indice approximativement 0.033.
Pour estimer les indices de premier ordre des modèles Y_1 et Y_2 avec la même précision, une étude expérimentale sur différentes tailles d'échantillon a permis de montrer qu'une taille de seulement 20000 suffisait. En effet, comme le nombre de variables de chacun de ces deux modèles n'est que de 10, l'estimation des indices nécessite moins de simulations que pour Y qui en compte 20. Cette taille de 20000 entraîne un temps de calcul des indices de premier ordre de 0.3 seconde pour chacun des deux sous modèles. Pour déduire de ces derniers les indices de premier ordre de Y, il suffit de multiplier les indices de Y_1 par $\frac{\mathrm{V}(Y_1)}{\mathrm{V}(Y)}$ et ceux de Y_2 par $\frac{\mathrm{V}(Y_2)}{\mathrm{V}(Y)}$. Ici ces deux coefficients sont trivialement égaux à $\frac{1}{2}$, mais dans un cas plus général, leur estimation ne demande pas de calcul supplémentaire, puisque les variances de Y_1 et de Y_2 ont déjà été évaluées au cours de leur analyse de sensibilité.

Ainsi, pour cet exemple, l'analyse des deux sous modèles séparément apporte en 0.6 seconde la même information que l'analyse du modèle complet en 4.8 secondes. Le gain est important, en partie grâce au fait que les deux sous modèles ont deux fois moins de variables d'entrée que le modèle somme. Si Y_1 comptait par exemple 3 variables d'entrée et Y_2 en comptait 17, le gain d'estimation aurait été moindre, car l'estimation de Y_2 aurait été quasiment aussi coûteuse que l'estimation de Y.
Dans le cas où le modèle se sépare additivement en deux sous modèles indépendants, il peut donc être intéressant d'un point de vue coût de calcul de réaliser l'analyse de sensibilité sur chacun des deux sous modèles, lorsque la différence de complexité entre modèle et sous modèle est importante. La même conclusion est valable pour un modèle qui se sépare multiplicativement en deux sous modèles. À noter dans ce dernier cas, que des interactions apparaîssent suite à la multiplication des deux sous modèles, mais il est encore possible d'estimer les indices de sensibilité à ces interactions à partir des indices de sensibilité des sous modèles (cf. paragraphe 2.1.2.1).
Pour les applications sur GASCON présentées précédemment, la différence de complexité est trop faible pour constater une différence significative entre les incertitudes d'estimation des indices des sous modèles et du modèle somme pour une même taille d'échantillon de Monte-Carlo. En effet, pour les mutations 2.1

et 2.2, le nombre de variables du modèle somme est de 10 (ou 8 lorsque X_1 et X_5 sont fixées), et celui des deux sous modèles est 5 (ou 4), d'où une différence de 5 variables (ou 4). Cette différence ne permet pas d'observer une différence au niveau de l'incertitude d'estimation, contrairement à l'exemple présenté ci-dessus, où les sous modèles ont 10 variables et le modèle somme en a 20.

Lorsqu'un modèle se décompose en somme ou produit de deux sous modèles de complexité bien inférieure mais cette fois non totalement indépendants (variables communes et différentes), il sera possible (et plus efficace) d'estimer les indices de sensibilité des variables qui n'appartiennent qu'à un seul des deux sous modèles par l'analyse de sensibilité du sous modèle. Les indices de sensibilité des variables communes devront être estimés par l'analyse de sensibilité du modèle complet.

Remarque. *Un autre avantage de réaliser l'analyse de sensibilité séparément sur les deux sous modèles, est que l'on obtient une information plus précise sur l'analyse de chaque sous modèle. Mais il faut être très vigilant quant à l'interprétation de ces résultats, car l'objectif est toujours d'analyser la sensibilité du modèle complet aux variables d'entrée. Par exemple, lors de la mutation 2.1, la partie du modèle relative à la chaîne alimentaire du lait de brebis n'a aucune influence sur la variance de la variable* Ad_I_1. *Certaines variables qui contribuent beaucoup à la variance de cette partie du modèle, n'ont néanmoins aucune influence sur la variance de* Ad_I_1.

Remarque. *Concernant l'application du paragraphe 2.3.5.2, l'estimation de la variance des deux sous modèles Y_1 et Y_2 ne demande que peu de calcul (une centaine de simulation de ces modèles), alors que leur analyse de sensibilité par la méthode de Sobol est beaucoup plus coûteuse. En constatant que la variance de Y_1 est près de 100 fois plus grande que celle de Y_2, il est raisonnable de concentrer l'analyse uniquement sur Y_1, et ainsi de réaliser l'analyse de sensibilité sur un modèle plus petit.*

2.3.7. Conclusions et applicabilité

Nous avons présenté le logiciel GASCON ainsi que la variable de sortie étudiée : la dose efficace annuelle en ^{129}I reçue par un adulte d'un groupe de référence proche de Cadarache (France, Bouches-du-Rhône) au bout d'un an. Le logiciel GASCON étant trop lourd en temps de calcul pour une analyse de sensibilité par la méthode de Sobol, une surface de réponse est utilisée. Cette surface de réponse est considérée comme le modèle de référence de calcul de la variable de sortie étudiée. Nous avons montré par l'étude des résidus de l'ajustement de la surface de réponse sur GASCON, que les estimations des indices de sensibilité de GASCON faites en utilisant la surface de réponse étaient correctes.
Nous avons ensuite illustré, à partir de cette surface de réponse, l'applicabilité des travaux de la section 2.1, en définissant un certain nombre de mutations (fixer une variable aléatoire, découper le modèle en somme ou produit de sous modèles). Les indices de sensibilité ont pu être estimés, sans calcul supplémentaire important. Pour ces mutations, le coût d'une nouvelle analyse de sensibilité a donc pu être diminué.
Ce travail nous a en outre permis de définir des conseils sur la démarche à suivre pour réaliser une analyse de sensibilité, lorsque le modèle étudié peut se séparer en sous modèles (de complexité significativement plus faible).

Chapitre 3

Analyse de sensibilité et modèles à entrées dépendantes

L'analyse de sensibilité, présentée en section 1.1, définit des indices de sensibilité sous l'hypothèse d'indépendance des variables d'entrée du modèle. Ces indices expriment la sensibilité de la variance de la sortie du modèle aux variables d'entrée, ainsi qu'aux interactions éventuelles entre ces entrées. En pratique, l'hypothèse d'indépendance des variables d'entrée peut s'avérer ne pas être réaliste.
Dans un premier temps, nous illustrons le problème de l'analyse de sensibilité en présence de variables non indépendantes à travers un exemple simple. Nous présentons ensuite une synthèse bibliographique sur le sujet et nous montrons que de nouveaux termes apparaissent dans la décomposition de la variance du modèle en fonction des variables d'entrée lorsque ces dernières sont corrélées. Les indices de sensibilité définis dans le cas indépendant à partir de cette décomposition, n'ont alors plus de signification et il est impossible d'interpréter les nouveaux termes de la décomposition de la variance en terme de sensibilité. Nous introduisons alors des indices de sensibilité multidimensionnels.
Enfin, nous illustrons ces travaux par une application dans le domaine de l'ingénierie nucléaire.

3.1. Indices de sensibilité multidimensionnels

3.1.1. Un exemple d'analyse de sensibilité classique

Considérons le modèle :

$$Y = X_1 + X_2,$$

où X_1 et X_2 sont deux variables indépendantes et de loi uniforme sur [0,1]. Les indices de sensibilité sont trivialement égaux à $\frac{1}{2}$ pour S_1 et S_2 et à 0 pour l'interaction S_{12}. L'interprétation est alors simple : la variance du modèle est sensible autant à X_1 qu'à X_2, *i.e.* la variance de Y est expliquée à moitié par la variance de X_1 et à moitié par celle de X_2.
Supposons maintenant les variables X_1 et X_2 dépendantes, de loi conjointe :

$$\mu_{X_1,X_2}(x_1,x_2) = \begin{cases} 2 & \text{si} \quad 0 \leq x_1, x_2 \leq 0.5 \quad \text{ou} \quad 0.5 \leq x_1, x_2 \leq 1, \\ 0 & \text{sinon.} \end{cases}$$

La figure 3.1 présente la répartition d'un échantillon de 5000 simulations du couple (X_1, X_2) sur le carré unité $[0,1]^2$.

Les indices de sensibilité classiques, définis en section 1.1, peuvent être calculés formellement, sur ce modèle. Ils sont alors égaux à :

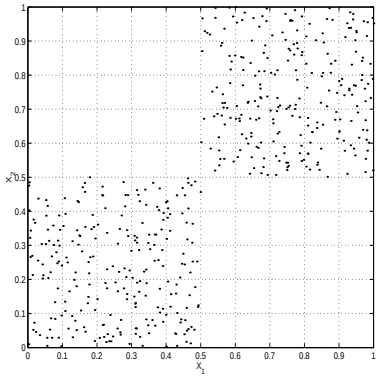


FIG. 3.1. : Répartition d'un échantillon de réalisations du couple (X_1, X_2) sur le carré unité.

S_1	S_2	S_{12}
$\frac{13}{14} \simeq 0.93$	$\frac{13}{14} \simeq 0.93$	$-\frac{6}{7} \simeq -0.86$

L'indice de sensibilité S_{12} est égal à la part de la variance de Y qui n'est pas expliquée par X_1 et X_2 seules. Or, comme ces deux dernières sont dépendantes, la part de variance due à X_1 seule prend en compte une partie de la variance due à X_2, de même pour la part de variance due à X_2 seule. La somme des parts de variance due à X_1 seule et à X_2 seule dépasse donc la variance totale de Y. C'est pourquoi l'indice de sensibilité S_{12} est négatif.
Cet exemple illustre bien qu'en absence d'indépendance, l'indice de sensibilité de premier ordre n'exprime pas la part de la variance de la sortie due uniquement à une entrée, l'indice d'ordre deux n'exprime pas non plus la part due uniquement à l'interaction (sinon il devrait être nul, étant donné qu'il n'y a pas de termes d'interaction dans le modèle, ce qui n'est pas le cas), et ainsi de suite.

Nous présentons en premier lieu un état de l'art de l'analyse de sensibilité pour modèles à entrées corrélées ou dépendantes.

3.1.2. Synthèse bibliographique

Le problème des modèles à entrées non indépendantes en analyse de sensibilité est sujet à de nombreuses études actuellement, dont nous présentons une synthèse chronologique.
Le premier auteur à aborder le sujet est McKay [34] en 1995, qui explique comment utiliser sa méthode d'estimation des indices de premier ordre (décrite en section 1.1) pour des variables non indépendantes. Sa méthode d'estimation se basant sur un échantillonnage par hypercube latin répliqué, il utilise les méthodes de Iman et Conover [28], ou de Stein [61], pour garder la dépendance des variables d'entrées lors des réplications (définies comme des permutations d'un échantillon initial). Il ne remet pas en cause l'utilisation des indices de sensibilité de premier ordre comme ils sont définis pour les modèles à entrées indépendantes (Définition 1.1.4) : les indices de premier ordre sont toujours supposés exprimer la sensibilité de la variance de la sortie du modèle à la variance d'une variable d'entrée.
Kraan et Cooke montrent en 1997 [31], par l'étude de deux cas pratiques, que la corrélation des variables d'entrée ne doit pas être négligée lors de l'étude de la distribution de sortie d'un modèle. En effet, la corrélation fait partie intégrante du modèle, la négliger revient à changer de modèle.
En 1998, Bedford [7] s'appuye sur l'idée que lorsque les modèles sont à entrées dépendantes, les fonctions de la décomposition de Sobol ne sont plus orthogonales. Il propose donc d'orthogonaliser cette famille de fonction, à l'aide d'un procédé de type Gramm-Schmidt. Le premier problème de cette méthode est que la valeur des indices de sensibilité calculée dépend du procédé d'orthogonalisation choisi, plus précisément de

l'ordre dans lequel les fonctions sont orthogonalisées. Le second problème est que la sensibilité est évaluée sur les variables orthogonalisées qui ne correspondent plus exactement aux variables initiales du modèle. Or si ces dernières ont une signification physique réelle, les variables orthogonalisées n'en ont pas forcément une et on n'exprime plus alors la sensibilité sur une variable physique mais sur une transformation de variables, pouvant parfois mêler plusieurs variables.
Toujours en 1998, RamaRao *et al.* [45] utilise les coefficients SRC et PCC dans le cas de modèles à entrées dépendantes. Il affirme (sans démonstration) que l'indice SRC ne peut pas être utilisé comme un indicateur de sensibilité pour de tels modèles. En se plaçant dans un cadre de régression multiple, il montre que le coefficient PCC peut être relié à l'augmentation du coefficient de détermination lorsqu'une variable est ajoutée au modèle de régression. Cette liaison, qui permet d'expliquer que l'indice PCC est un indicateur de la sensibilité du modèle à une variable (en terme de part de variance), est toujours vraie si les variables d'entrée sont corrélées. Ainsi, pour un modèle linéaire, le coefficient PCC exprime toujours la sensibilité du modèle à une de ses variables d'entrée même si ces dernières sont corrélées.
Les derniers travaux en date sur le problème des modèles à entrées corrélées sont dus à Tarantola et Saltelli ([62] et [51], résumés dans [53]). Les différentes questions auxquelles l'analyse de sensibilité répond sont classées en catégories (*settings*) :

- *Factors Prioritisation (FP)* : trouver quelle est la variable d'entrée du modèle qui, fixée, apporte la plus grande réduction de la variance de la sortie.
- *Factors Fixing (FF)* : trouver les variables ou groupes de variables auxquels le modèle n'est pas sensible, au sens de la contribution à la variance.
- *Variance Cutting (VC)* : quel est l'ensemble minimal de variables à fixer pour obtenir une certaine réduction de la variance ?
- *Factors Mapping (FM)* : quelles sont les variables qui sont les plus responsables de réalisations de la sortie dans une région donnée (valeur de la sortie dans tel ou tel intervalle) ?

Si l'analyse de sensibilité est utilisée avec comme objectif *FP*, Saltelli et Tarantola expliquent que les indices de sensibilité de premier ordre classiques sont toujours utilisables. La méthode d'estimation qu'ils utilisent alors est celle de McKay. Si l'objectif est *VC*, ils proposent une heuristique permettant de trouver l'ensemble minimal qui apporterait une réduction donnée de la variance. Cette heuristique nécessite de calculer tous les indices de sensibilité totaux. Si la réduction de la variance obtenue en fixant la variable qui a le plus grand indice total est suffisante, on s'arrête. Sinon, on calcule la réduction de variance obtenue en fixant les deux variables ayant les deux indices de sensibiltié totaux les plus important. Et ainsi de suite jusqu'à obtention de la réduction désirée.

Lorsque les variables d'entrée sont non indépendantes, l'utilisation de l'analyse de sensibilité est plus restreinte (*(FP),(VC)*). Les indices de sensibilité de premier ordre peuvent être estimés par la méthode de McKay (ou par la méthode d'estimation de l'indice PCC si le modèle est relativement linéaire). Afin d'élargir le champ d'utilisation de l'analyse de sensibilité en présence de corrélation, nous étudions dans le paragraphe suivant, comme l'a fait Sobol pour le cas indépendant, comment décomposer la variance de la sortie du modèle en fonction des variables d'entrée, mais cette fois sans l'hypothèse d'indépendance de ces dernières.

3.1.3. Décomposition de la variance d'une fonction de variables non indépendantes

Dans le cas indépendant, les indices de sensibilité sont définis par décomposition de la variance de la sortie du modèle en part de variance de premier ordre (due à une variable), de deuxième ordre (due à l'interaction entre deux variables), et ainsi de suite. Nous examinons ce que devient cette décomposition lorsque les variables d'entrée ne sont plus indépendantes.

Considérons le modèle :

$$Y = f(X_1, ..., X_p),$$

où les variables d'entrée peuvent être corrélées ou dépendantes. La décomposition de Sobol (1.9) reste toujours vraie sous la forme, puisque la dépendance n'influe pas sur les calculs d'espérance :

$$f(X_1, ..., X_p) = f_0 + \sum_{i=1}^{p} f_i(X_i) + \sum_{1 \le i < j \le p} f_{i,j}(X_i, X_j) + \ldots + f_{1,...,p}(X_1, ..., X_p),$$

avec

$$\begin{aligned}
f_0 &= \mathrm{E}[Y], \\
f_i(X_i) &= \mathrm{E}[Y|X_i] - f_0, \\
f_{i,j}(X_i, X_j) &= \mathrm{E}[Y|X_i, X_j] - f_i(X_i) - f_j(X_j) - f_0, \\
f_{i,j,k}(X_i, X_j, X_k) &= \mathrm{E}[Y|X_i, X_j, X_k] - f_{i,j}(X_i, X_j) - f_{i,k}(X_i, X_k) - f_{j,k}(X_j, X_k) \\
&\quad - f_i(X_i) - f_j(X_j) - f_k(X_k) - f_0, \\
&\ldots
\end{aligned}$$

puisque la dernière fonction $f_{1,...,p}(X_1, ..., X_p)$ est définie de sorte à vérifier cette égalité (égal à f moins toutes les fonctions d'ordre 1 à $p-1$).
La décomposition de Sobol de la variance est encore vérifiée :

$$\begin{aligned}
\mathrm{V}(Y) &= \sum_{i=1}^{p} V_i + \sum_{1 \le i < j \le p} V_{ij} + \ldots + V_{1...p}, \\
&\text{avec} \\
V_i &= \mathrm{V}(\mathrm{E}[Y|X_i]), \\
V_{ij} &= \mathrm{V}(\mathrm{E}[Y|X_i, X_j]) - V_i - V_j, \\
V_{ijk} &= \mathrm{V}(\mathrm{E}[Y|X_i, X_j, X_k]) - V_{ij} - V_{ik} - V_{jk} - V_i - V_j - V_k, \\
&\ldots \\
V_{1...p} &= V - \sum_{i=1}^{p} V_i - \sum_{1 \le i < j \le p} V_{ij} - \ldots - \sum_{1 \le i_1 < ... < i_{p-1} \le p} V_{i_1...i_{p-1}}
\end{aligned}$$

puisque le dernier terme $V_{1...p}$ de la décomposition est défini comme la différence entre la variance de Y et toutes les autres parts de variance d'ordre inférieur.
Par contre, contrairement au cas indépendant, les parts de variance $V_{i_1...i_s}$ ne sont pas les variances des fonctions de la décomposition de Sobol $\tilde{V}_{i_1...i_s} = \mathrm{V}(f_{i_1,...,i_s}(X_{i_1}, ..., X_{i_s}))$. La séparation des effets des différentes variables d'entrée faite dans la décomposition de Sobol de la fonction du modèle n'est plus transmise dans la décomposition de la variance de Y. En effet, si $f_{i_1,...,i_s}(X_{i_1}, ..., X_{i_s})$ traduit l'effet des s variables $X_{i_1}, ..., X_{i_s}$ sur Y qui n'est pas compris dans les effets des sous-ensembles stricts de $\{X_{i_1}, ..., X_{i_s}\}$, et donc si $\tilde{V}_{i_1,...,i_s}$ traduit la part de variance de Y due à l'interaction entre ces s variables $X_{i_1}, ..., X_{i_s}$, ce n'est plus le cas de $V_{i_1,...,i_s}$. Or, c'est à partir de $V_{i_1,...,i_s}$ que l'indice de sensibilité de Y à l'interaction entre $X_{i_1}, ..., X_{i_s}$ est défini.

Nous montrons par la suite, dans le cas d'un modèle à deux variables d'entrée non indépendantes, qu'il est possible de définir la différence entre $V_{i_1,...,i_s}$ et $\tilde{V}_{i_1,...,i_s}$.

Le cas de deux variables dépendantes

Proposition 3.1.1.
Soit le modèle $Y = f(X_1, ..., X_p)$ à p variables d'entrée. On suppose $(X_3, ..., X_p)$ indépendantes deux à deux, et X_1 (respectivement X_2) indépendante de X_j pour tout $j = 3, \ldots, p$. Les deux variables X_1 et X_2 peuvent être dépendantes.
La variance de Y se décompose alors en :

$$V(Y) = \sum_{i=1}^{p} \tilde{V}_i + \sum_{1 \leq i < j \leq p} \tilde{V}_{ij} + \ldots + \tilde{V}_{1\ldots p} + V_{12}^{*(p)}, \tag{3.1}$$

où :

$$\begin{aligned}
\tilde{V}_i &= V(f_i(X_i)), \\
\tilde{V}_{ij} &= V(f_{i,j}(X_i, X_j)), \\
\tilde{V}_{ijk} &= V(f_{i,j,k}(X_i, X_j, X_k)), \\
&\ldots \\
V_{12}^{*(p)} &= 2 \sum_{s \in \mathcal{S}} (E[f_{1,s} f_{2,s}] + E[f_{1,s} f_{1,2,s}] + E[f_{2,s} f_{1,2,s}]),
\end{aligned}$$

avec $\mathcal{S} = \{$sous-ensemble de $\{\emptyset\} \cup \{3, ..., p\}\}$, la notation $f_{i_1,\ldots,i_r}$ désignant la fonction $f_{i_1,\ldots,i_r}(X_{i_1}, \ldots, X_{i_r})$, définie suivant le modèle :

$$\begin{aligned}
f_0 &= E[Y], \\
f_i(X_i) &= E[Y|X_i] - f_0, \\
f_{i,j}(X_i, X_j) &= E[Y|X_i, X_j] - f_i(X_i) - f_j(X_j) - f_0, \\
f_{i,j,k}(X_i, X_j, X_k) &= E[Y|X_i, X_j, X_k] - f_{i,j}(X_i, X_j) - f_{i,k}(X_i, X_k) - f_{j,k}(X_j, X_k) \\
&\quad - f_i(X_i) - f_j(X_j) - f_k(X_k) - f_0, \\
&\ldots
\end{aligned}$$

Démonstration. Pour vérifier la proposition précédente, considérons le cas où le nombre de variables du modèle est $p = 2$, puis $p = 3$. La généralisation pour un p quelconque sera alors naturelle.
Soit un modèle à deux variables d'entrée :

$$Y = f(X_1, X_2),$$

où X_1 et X_2 sont supposées non indépendantes. La décomposition de Sobol de ce modèle s'écrit :

$$Y = f_0 + f_1(X_1) + f_2(X_2) + f_{1,2}(X_1, X_2),$$

où les fonctions de la décomposition sont d'espérance nulle et définies par :

$$\begin{aligned}
f_0 &= \mathrm{E}[Y], \\
f_1(X_1) &= \mathrm{E}[Y|X_1] - f_0, \\
f_2(X_2) &= \mathrm{E}[Y|X_2] - f_0, \\
f_{1,2}(X_1, X_2) &= \mathrm{E}[Y|X_1, X_2] - f_0 - f_1 - f_2.
\end{aligned}$$

La variance de Y s'écrit alors :

$$\begin{aligned}
\mathrm{V}(Y) &= \mathrm{V}(f_0 + f_1(X_1) + f_2(X_2) + f_{1,2}(X_1, X_2)) \\
&= \underbrace{\mathrm{V}(f_1(X_1))}_{\tilde{V}_1} + \underbrace{\mathrm{V}(f_2(X_2))}_{\tilde{V}_2} + \underbrace{\mathrm{V}(f_{1,2}(X_1, X_2))}_{\tilde{V}_{12}} \\
&\quad +2\mathrm{Cov}(f_1(X_1), f_2(X_2)) + 2\mathrm{Cov}(f_1(X_1), f_{1,2}(X_1, X_2)) + 2\mathrm{Cov}(f_2(X_2), f_{1,2}(X_1, X_2)) \\
&= \tilde{V}_1 + \tilde{V}_2 + \tilde{V}_{12} + 2\mathrm{Cov}(\mathrm{E}[Y|X_1], \mathrm{E}[Y|X_2]) \\
&\quad +2\mathrm{Cov}(\mathrm{E}[Y|X_1], \mathrm{E}[Y|X_1, X_2] - \mathrm{E}[Y|X_1] - \mathrm{E}[Y|X_2]) \\
&\quad +2\mathrm{Cov}(\mathrm{E}[Y|X_2], \mathrm{E}[Y|X_1, X_2] - \mathrm{E}[Y|X_1] - \mathrm{E}[Y|X_2]) \\
&= \tilde{V}_1 + \tilde{V}_2 + \tilde{V}_{12} + 2\mathrm{Cov}(\mathrm{E}[Y|X_1], \mathrm{E}[Y|X_2]) \\
&\quad +2\mathrm{V}(\mathrm{E}[Y|X_1]) - 2\mathrm{V}(\mathrm{E}[Y|X_1]) - 2\mathrm{Cov}(\mathrm{E}[Y|X_1], \mathrm{E}[Y|X_2]) \\
&\quad +2\mathrm{V}(\mathrm{E}[Y|X_2]) - 2\mathrm{V}(\mathrm{E}[Y|X_2]) - 2\mathrm{Cov}(\mathrm{E}[Y|X_1], \mathrm{E}[Y|X_2]) \\
&= \tilde{V}_1 + \tilde{V}_2 + \tilde{V}_{12} - 2\mathrm{Cov}(\mathrm{E}[Y|X_1], \mathrm{E}[Y|X_2])
\end{aligned}$$

en utilisant la propriété $\mathrm{Cov}(\mathrm{E}[Y|X_1], \mathrm{E}[Y|X_1, X_2]) = \mathrm{V}(\mathrm{E}[Y|X_1])$ démontrée en annexe A.2.1.
Un terme de covariance apparaît donc dans la décomposition de la variance de Y.

Afin de généraliser au cas d'un modèle à p variables d'entrée, nous montrons que cette covariance peut s'écrire sous la forme de la Proposition 3.1.1.
En écrivant la variance de Y comme :

$$\begin{aligned}
\mathrm{V}(Y) &= \mathrm{E}[Y^2] - f_0^2 \\
&= \mathrm{E}[f_1^2] + \mathrm{E}[f_2^2] + \mathrm{E}[f_{12}^2] + 2\mathrm{E}[f_1 f_2] + 2\mathrm{E}[f_1 f_{12}] + 2\mathrm{E}[f_2 f_{12}],
\end{aligned}$$

on obtient :

$$\mathrm{V}(Y) = \tilde{V}_1 + \tilde{V}_2 + \tilde{V}_{12} + V_{12}^{*(2)}.$$

avec :

$$\begin{aligned}
V_{12}^{*(2)} &= 2(\mathrm{E}[f_1 f_2] + \mathrm{E}[f_1 f_{12}] + \mathrm{E}[f_2 f_{12}]) \\
&= -2\mathrm{Cov}(\mathrm{E}[Y|X_1], \mathrm{E}[Y|X_2]),
\end{aligned}$$

Remarque.
- *Lorsqu'il y a indépendance il vient naturellement que $V_{12}^{*(2)} = 0$.*
- *Le terme $V_{12}^{*(2)}$ étant une covariance, il est important de noter qu'il peut aussi bien être négatif que positif.*

Considérons maintenant un modèle à trois variables d'entrée :

$$Y = f(X_1, X_2, X_3),$$

où les deux variables X_1 et X_2 sont non indépendantes. La décomposition de Sobol de Y est :

$$\begin{aligned}
Y &= f_0 + f_1(X_1) + f_2(X_2) + f_3(X_3) \\
&\quad +f_{1,2}(X_1, X_2) + f_{1,3}(X_1, X_3) + f_{2,3}(X_2, X_3) \\
&\quad +f_{1,2,3}(X_1, X_2, X_3).
\end{aligned}$$

La variance de Y s'écrit :

$$\begin{aligned}\mathrm{V}(Y) &= \mathrm{E}[Y^2] - f_0^2 \\ &= \mathrm{E}[f_1^2] + \mathrm{E}[f_2^2] + \mathrm{E}[f_3^2] + \mathrm{E}[f_{1,2}^2] + \mathrm{E}[f_{1,3}^2] + \mathrm{E}[f_{2,3}^2] + \mathrm{E}[f_{1,2,3}^2] \\ &\quad +2\mathrm{E}[f_1 f_2] + 2\mathrm{E}[f_1 f_{1,2}] + 2\mathrm{E}[f_2 f_{1,2}] + 2\mathrm{E}[f_{1,3} f_{2,3}] \\ &\quad +2\mathrm{E}[f_{1,3} f_{1,2,3}] + 2\mathrm{E}[f_{2,3} f_{1,2,3}],\end{aligned}$$

en utilisant le fait que la variable X_3 est indépendante des deux autres (les calculs sont détaillés en annexe A.2.2) .
Le terme supplémentaire est alors :

$$V_{12}^{*(3)} = 2(\mathrm{E}[f_1 f_2] + \mathrm{E}[f_1 f_{1,2}] + \mathrm{E}[f_2 f_{1,2}] + \mathrm{E}[f_{1,3} f_{2,3}] + \mathrm{E}[f_{1,3} f_{1,2,3}] + \mathrm{E}[f_{2,3} f_{1,2,3}]),$$

et ainsi :

$$\mathrm{V}(Y) = V_1 + V_2 + V_3 + V_{12} + V_{13} + V_{23} + V_{123} + V_{12}^{*(3)}.$$

La proposition est alors obtenue par généralisation de cette décomposition pour un modèle à p variables d'entrée. □

Un nouvel indice de sensibilité

Proposons un nouvel indice de sensibilité, défini par :

$$S_{12}^{*} = \frac{V_{12}^{*(p)}}{\mathrm{V}(Y)}.$$

Cet indice, qui est nul lorsque les variables d'entrée sont indépendantes, semble exprimer la sensibilité du modèle à la corrélation des variables d'entrée. Cependant, nous montrons à l'aide d'un exemple test, qu'il n'en est rien.

Remarque. *Si les variables d'entrée sont indépendantes, leur covariance est nulle, ce qui implique la nullité de l'indice S_{12}^{*}. Par contre, le fait que la covariance et S_{12}^{*} soient nuls n'implique pas l'indépendance des variables d'entrée.*

Exemple

Soit le modèle :

$$Y = X_1 + X_2 + X_3, \tag{3.2}$$

où

$$(X_1, X_2, X_3) \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & 0 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix} \right).$$

La variance de Y est :

$$\mathrm{V}(Y) = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + 2\rho\sigma_1\sigma_2,$$

et se décompose selon (3.1) en :

$$\mathrm{V}(Y) \quad = \quad V_1 + V_2 + V_3 + V_{12} + V_{13} + V_{23} + V_{123} + V_{12}^{*(2)},$$

où :

$$\begin{aligned} V_1 &= (\sigma_1 + \rho\sigma_2)^2, \\ V_2 &= (\sigma_2 + \rho\sigma_1)^2, \\ V_3 &= \sigma_3^2, \\ V_{12} &= \rho^2(\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2), \\ V_{13} &= V_{23} = V_{123} = 0, \\ V_{12}^* &= -2(\rho\sigma_1\sigma_2 + \rho^2(\sigma_1 + \sigma_2) + \rho^3\sigma_1\sigma_2). \end{aligned}$$

La figure 3.2 illustre les valeurs des indices de sensibilité en fonction de la corrélation ρ, pour des variables X_i centrées réduites ($\mu_i = 0$ et $\sigma_i^2 = 1$ pour $i = 1, 2, 3$).

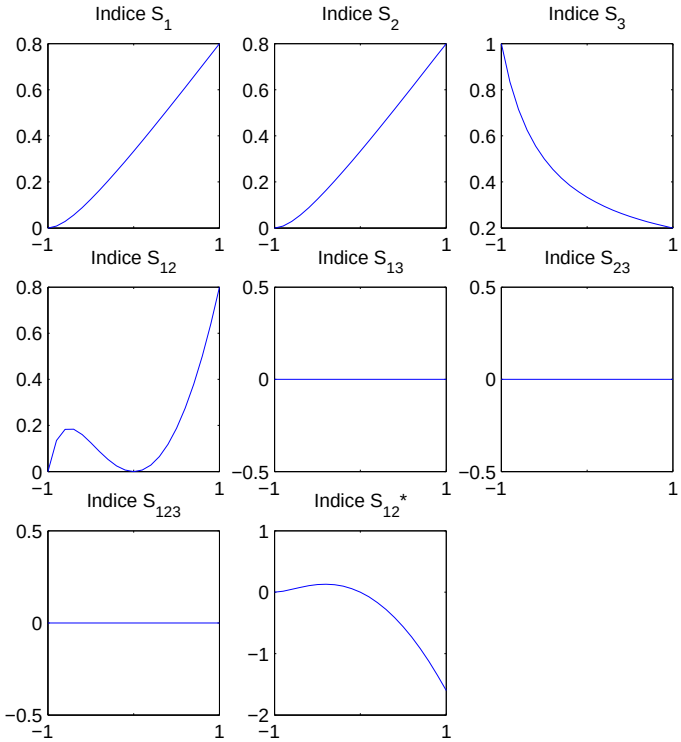


FIG. 3.2. : Indice de sensibilité du modèle $Y = X_1 + X_2 + X_3$, en fonction de la corrélation de X_1 et X_2

- Lorsque la corrélation est nulle, seuls les indices de sensibilité de premier ordre sont égaux à $\frac{1}{3}$.
- Lorsqu'elle est égale à -1, le terme $X_1 + X_2$ du modèle annule l'effet des variables X_1 et X_2, puisque la corrélation parfaite implique $X_1 = -X_2$. Le modèle est alors assimilable au modèle $Y = X_3$, d'où le seul indice non nul et égal à 1 est S_3.
- Lorsque la corrélation est cette fois égale à 1 ($X_1 = X_2$), le modèle est $Y = 2X_1 + X_3$ ou $Y = 2X_2 + X_3$, les variances de X_1 et X_2 ont donc quatre (2^2) fois plus d'importance que celle de X_3.

D'une façon générale, lorsque la corrélation varie entre -1 et 1, cet exemple illustre que la corrélation influe sur les valeurs de la plupart des indices de sensibilité, et non seulement sur S_{12}^* comme on aurait pu l'espérer. En effet, le coefficient de corrélation ρ apparaît dans les équations analytiques des variances partielles V_1,

V_2, V_{12} et V_{12}^*. Ceci est toutefois logique, puisque la corrélation fait partie entière du modèle, et modifier cette corrélation change le modèle lui même, et donc la sensibilité du modèle aux différentes variables. Le comportement propre de l'indice S_{12}^* n'est pas aisément compréhensible, puisqu'il est quasiment nul lorsque la corrélation varie négativement de -1 à 0, et décroît lorsque la corrélation croît positivement.
La décomposition (3.1) ne permet donc pas d'isoler l'effet de la corrélation sur la variance du modèle. Définir un nouvel indice de sensibilité à partir des termes supplémentaires qui apparaîssent dans la décomposition de la variance d'un modèle en présence de dépendance ne semble pas avoir d'intérêt pratique, cet indice n'ayant pas d'interprétation naturelle en terme de sensibilité.
Si l'on observe le comportement de l'indice S_{12}, on constate qu'il croît avec la corrélation. On est alors tenté de dire que plus X_1 et X_2 sont corrélées (positivement), plus leur interaction est responsable d'une part importante de la variance de Y. Or il n'en est rien puisqu'il n'y a pas d'interaction au sein du modèle, ces deux variables étant séparées additivement. Il s'avère donc dangereux d'interpréter séparément les indices S_1, S_2, S_{12} et S_{12}^*. Un autre raisonnement menant à la même conclusion est de considérer la partie du modèle (3.2) constituée par la somme de X_1 et X_2. Comme ces deux variables sont corrélées, il existe une infinité de façon d'écrire leur somme $X_1 + X_2$, c'est-à-dire une infinité de couples (W_1, W_2) tels que $X_1 + X_2 = W_1 + W_2$, d'où une infinité de valeurs pour les indices S_1, S_2, S_{12} et S_{12}^*. Il n'est donc pas possible de donner un sens à la sensibilité du modèle à la variable X_1 seule, ou à la variable X_2 seule.
Ainsi, en quantifiant la sensibilité à une variable par une analyse de sensibilité classique, on prend aussi en compte, au moins partiellement, la sensibilité aux autres variables avec lesquelles elle est dépendante. Il est impossible d'isoler la part de variance due uniquement à une variable puisqu'il y a naturellement une partie de la variance qui est due à la dépendance avec ces autres variables. Nous proposons donc d'exprimer la sensibilité pour des groupes de variables dépendantes, et d'introduire ainsi des indices de sensibilité multidimensionnels.
Comme nous l'a fait remarquer Saltelli, la possibilité de calculer des indices de sensibilité sur des groupes de variables est connue (cf. travaux de Sobol [60]). L'originalité de notre travail consiste en l'utilisation de cette sensibilité «par groupe» pour l'analyse de sensibilité de modèles à entrées non indépendantes.

3.1.4. Indices de sensibilité multidimensionnels

Soit le modèle :

$$Y = f(X_1, \ldots, X_p),$$

où on suppose qu'il y a des groupes de variables corrélées. Les variables d'un même groupe sont corrélées, mais les variables de groupes différents sont indépendantes. Ces groupes sont réduits à une unique variable si cette dernière est indépendante de toutes les autres variables.
Pour une représentation plus aisée, nous choisissons d'écrire le vecteur des variables d'entrée de la façon suivante :

$$(\underbrace{X_1}_{\vec{X}_1}, ..., \underbrace{X_I}_{\vec{X}_I}, \underbrace{X_{I+1}, ..., X_{I+k_1}}_{\vec{X}_{I+1}}, \underbrace{X_{I+k_1+1}, ..., X_{I+k_2}}_{\vec{X}_{I+2}}, ..., \underbrace{X_{I+k_{L-1}+1}, ..., X_p}_{\vec{X}_{I+L}}), \tag{3.3}$$

où les p variables unidimensionnelles corrélées $X_1, \ldots, X_p$ sont regroupées en $I + L$ groupes, formant les variables multidimensionnelles indépendantes : $\vec{X}_1, \ldots, \vec{X}_{I+L}$. Parmi ces $I + L$ groupes, nous supposons que les I premiers sont constitués d'une seule variable.
Nous avons mis en évidence précédemment la difficulté d'exprimer la sensibilité du modèle à chacune des variables $X_1, \ldots, X_p$, lorsqu'elles sont corrélées. Intéressons nous alors à la sensibilité du modèle aux variables indépendantes $\vec{X}_1, \ldots, \vec{X}_{I+L}$.
Par généralisation des indices de sensibilité basés sur la décomposition de la variance dans le cas unidimensionnel, nous introduisons les indices de sensibilité multidimensionnels de premier ordre.

Définition 3.1.1. *La sensibilité de la réponse Y du modèle à la variable $\vec{X}_j$ est quantifiée par l'indice de sensibilité multidimensionnel de premier ordre à la variable défini par :*

$$S_j = \frac{V(E[Y|\vec{X}_j])}{V(Y)} \quad \forall j \in \{1, ..., I+L\}. \tag{3.4}$$

Afin de relier cet indice aux indices classiques unidimensionnels, envisageons le cas où $\vec{X}_j$ est de dimension 1 et de dimension supérieure.
Si $\vec{X}_j$ est de dimension 1, c'est-à-dire $j \in \{1, ..., I\}$, l'indice de sensibilité multidimensionnel est strictement égal à l'indice classique unidimensionnel :

$$S_j = \frac{\mathrm{V}(\mathrm{E}[Y|\vec{X}_j])}{\mathrm{V}(Y)} = \frac{\mathrm{V}(\mathrm{E}[Y|X_j])}{\mathrm{V}(Y)}.$$

Si $\vec{X}_j$ est de dimension plus grande que 1 ($j \in \{I+1, ..., I+L\}$), par exemple $j = I+2$, l'indice de sensibilité multidimensionnel, noté $S_{\{i+k_1+1,...,i+k_2\}}$, est alors égal à :

$$S_j = S_{\{i+k_1+1,...,i+k_2\}} = \frac{\mathrm{V}(\mathrm{E}[Y|\vec{X}_j])}{\mathrm{V}(Y)} = \frac{\mathrm{V}(\mathrm{E}[Y|X_{i+k_1+1}, ..., X_{i+k_2}])}{\mathrm{V}(Y)}.$$

Comme dans le cas classique, il est possible de définir une décomposition de la variance du modèle, en fonction des groupes de variables corrélées. Cette décomposition permet de définir les différents indices de sensibilité multidimensionnels.
La Proposition 3.1.2 est la version multidimensionnelle de la Proposition 1.1.1.

Proposition 3.1.2. *Décomposition multidimensionnelle de la variance.*
Soit la fonction $Y = f(\vec{X}_1, \ldots, \vec{X}_{I+L})$ où $\vec{X}_j$, pour tout $1 \leq j \leq I+L$, sont $I+L$ variables aléatoires multidimensionnelles indépendantes. Alors la décomposition de $V(Y) = V$ est :

$$V = \sum_{j=1}^{I+L} V_j + \sum_{1 \leq j < k \leq I+L} V_{jk} + \ldots + V_{1\ldots I+L} \tag{3.5}$$

où

$$\begin{aligned} V_j &= V(E[Y|\vec{X}_j]), \\ V_{jk} &= V(E[Y|\vec{X}_j, \vec{X}_k]) - V_j - V_k, \\ V_{jkm} &= V(E[Y|\vec{X}_j, \vec{X}_k, \vec{X}_m]) - V_{jk} - V_{jm} - V_{km} - V_j - V_k - V_m, \\ &\ldots \end{aligned}$$

et où le dernier terme de la décomposition $V_{1\ldots I+L}$ s'écrit :

$$V_{1\ldots I+L} = \underbrace{V(E[Y|\vec{X}_1, \ldots, \vec{X}_{I+L}])}_{V} - \sum_{1 \leq j_1 < \ldots < I+L-1 \leq I+L} V_{j_1 \ldots j_{I+L-1}} - \ldots - \sum_{1 \leq j < k \leq I+L} V_{jk} - \sum_{j=1}^{I+L} V_j.$$

Il est défini de sorte à vérifier l'égalité (3.5).

Définition 3.1.2. *Ainsi, on peut définir :*
les indices de sensibilité (multidimensionnels) de premier ordre :

$$S_j = \frac{V_j}{V} = \frac{V(E[Y|\vec{X}_j])}{V} \qquad \forall 1 \leq j \leq I+L, \tag{3.6}$$

les indices de sensibilité d'ordre deux :

$$S_{jk} = \frac{V_{jk}}{V} \qquad \forall 1 \leq j < k \leq I + L,$$

qui expriment la sensibilité de la variance de Y à l'interaction entre les variables multidimensionnelles $\vec{X}_j$ et $\vec{X}_k$.
On définit encore les indices de sensibilité d'ordre trois :

$$S_{jkm} = \frac{V_{jkl}}{V} \qquad \forall 1 \leq j < k < l \leq I + L,$$

qui expriment la sensibilité de la variance de Y aux variables multidimensionnelles $\vec{X}_j$, $\vec{X}_k$ et $\vec{X}_l$ qui n'est pas prise en compte dans l'effet des variables seules et des interactions deux à deux.
Et ainsi de suite jusqu'à l'ordre p.

La notion de sensibilité totale à une variable multidimensionnelle est aussi introduite.

Définition 3.1.3. *La sensibilité totale de Y à la variable $\vec{X}_j$ est quantifiée par :*

$$S_{T_j} = \sum_{k \# j} S_k,$$

où $\# j$ représente tous les sous ensemble de $\{1, ..., I + L\}$ qui incluent j.

Il est important de noter que si les variables d'entrée du modèle sont toutes indépendantes, ces indices de sensibilité sont égaux aux indices de sensibilité unidimensionnels classiques (Définition 1.1.4). Les indices de sensibilité multidimensionnels sont donc une généralisation des indices de sensibilité unidimensionnels.
Les valeurs des indices de sensibilité multidimensionnels s'interprètent de la même façon que celles des indices de sensibilité unidimensionnels, puisqu'ils sont positifs et que la somme des indices de sensibilité multidimensionnels à tout ordre est égale à 1, grâce à la décomposition multidimensionnelle de la variance (Proposition 3.1.2).

Remarque. *Revenons à l'exemple du modèle (3.2). Nous considérons donc la variable bidimensionnelle (X_1, X_2) et la variable X_3. La variance de Y ($V(Y) = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + 2\rho\sigma_1\sigma_2$) se décompose en la part de variance de premier ordre due au couple (X_1, X_2) ($\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2$), et la part de variance due à X_3 (σ_3^2). Si on considère les trois variables d'entrée centrées réduites ($\mu_i = 0$ et $\sigma_i^2 = 1$ pour $i = 1, 2, 3$), les indices de sensibilité sont :*

$$S_{\{1,2\}} = \frac{2 + 2\rho}{3 + 2\rho} \qquad \text{et} \qquad S_3 = \frac{1}{3 + 2\rho}.$$

Cette fois il est possible d'interpréter la sensibilité de Y au couple (X_1, X_2), qui croît lorsque ρ croît vers 1, pour être jusqu'à quatre fois plus grande que la sensibilité à X_3, et décroît jusqu'à 0 lorsque ρ décroît vers -1. Cette définition élimine les risques d'erreur d'interprétation que l'on pouvait faire en interprétant séparément S_1, S_2 et S_{12}.

Nous illustrons ces indices de sensibilité multidimensionnels sur un exemple théorique.

3.1.5. Un exemple théorique

Considérons le modèle mathématique :

$$Y = aX_1X_2 + bX_3X_4 + cX_5X_6, \tag{3.7}$$

où les coefficients a, b et c sont réels, et les variables d'entrée sont gaussiennes :

$$(X_1, X_2, X_3, X_4, X_5, X_6) \sim \mathcal{N}(\vec{0}, \Sigma), \quad \text{avec} \quad \Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \rho_1 & 0 & 0 \\ 0 & 0 & \rho_1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \rho_2 \\ 0 & 0 & 0 & 0 & \rho_2 & 1 \end{bmatrix}.$$

Conformément au schéma (3.3), les variables X_3 et X_4 corrélées sont considérées comme une seule variable bidimensionnelle (X_3, X_4). L'indice de sensibilité multidimensionnel de premier ordre relatif à cette variable bidimensionnelle est noté $S_{\{3,4\}}$. De même, l'indice $S_{\{5,6\}}$ quantifie la sensibilité de la sortie du modèle Y au couple (X_5, X_6).
Ainsi, nous écrivons les six variables d'entrée unidimensionnelle non indépendantes comme quatre variables multidimensionnelles indépendantes :

$$(\underbrace{X_1}_{X_1}, \underbrace{X_2}_{X_2}, \underbrace{X_3, X_4}_{(X_3,X_4)}, \underbrace{X_5, X_6}_{(X_5,X_6)}).$$

Une analyse de sensibilité complète de ce modèle consiste à calculer les quatre indices de sensibilité multidimensionnels de premier ordre, les six indices d'ordre deux, les quatre indices d'ordre trois, et l'indice d'ordre quatre. Les indices de sensibilité multidimensionnels totaux sont obtenus en sommant les indices de l'ordre un à quatre. Dans cet exemple, les indices sont calculés formellement, puisque la forme analytique de la fonction est connue et simple (voir les détails des calculs en annexe A.2.3). Hormis les indices totaux, les seuls indices de sensibilité non nuls sont :

$$\begin{aligned} S_{12} &= \frac{a^2}{a^2 + b^2(1+{\rho_1}^2) + c^2(1+{\rho_2}^2)}, \\ S_{\{3,4\}} &= \frac{b^2(1+{\rho_1}^2)}{a^2 + b^2(1+{\rho_1}^2) + c^2(1+{\rho_2}^2)}, \\ S_{\{5,6\}} &= \frac{c^2(1+{\rho_2}^2)}{a^2 + b^2(1+{\rho_1}^2) + c^2(1+{\rho_2}^2)}. \end{aligned}$$

Le dénominateur de tous ces indices est la variance de Y. La valeur du numérateur de l'indice de sensibilité à l'interaction entre X_1 et X_2, S_{12}, est fonction du coefficient a. Les valeurs des numérateurs des autres indices non nuls $S_{\{3,4\}}$ et $S_{\{5,6\}}$ sont eux fonction des coefficients b et c mais aussi des coefficients de corrélation ρ_1 et ρ_2. Le tableau 3.1 illustre ceci en présentant les valeurs des indices de sensibilité pour différentes valeurs des coefficients a, b et c et des coefficients de corrélation ρ_1 et ρ_2. Comme les coefficients de corrélation sont élevés au carré dans les expressions des indices, leur signe n'a pas d'importance sur la valeur de ces derniers. Nous ne considérons donc que des coefficients de corrélation positifs.

Il faut avant tout souligner qu'étant donné l'indépendance de X_1 et X_2, les indices de sensibilité S_1, S_2, et S_{12} sont les indices classiques (ceux pour variables indépendantes), ils peuvent être calculés sans utiliser la forme multidimensionnelle.

- Dans la situation (i), une analyse de sensibilité classique permet donc de calculer S_1, S_2, et S_{12}, et de conclure que la variance de Y est due essentiellement (73%) à l'interaction entre X_1 et X_2.
- Par contre, dans les autres situations, comme l'effet des variables X_1 et X_2 est moins important, nous avons besoin de calculer les indices de sensibilité multidimensionnels pour montrer que les couples (X_3, X_4) et (X_5, X_6) ont la même importance dans la situation (ii), et que (X_5, X_6) est le plus important dans la situation (iii).

situation	a	b	c	ρ_1	ρ_2	S_{12}	$S_{\{3,4\}}$	$S_{\{5,6\}}$
(i)	3	1	1	0.8	0.8	0.7329	0.1336	0.1336
(ii)	1	1	1	0.8	0.8	0.2336	0.3832	0.3832
(iii)	1	1	3	0.8	0.8	0.0575	0.0943	0.8483
(iv)	1	1	1	0.8	0.3	0.2881	0.4397	0.2922
(v)	1	1	3	0.8	0.3	0.0803	0.1317	0.7880
(vi)	1	1	3	0.3	0.8	0.0593	0.0647	0.8760

TAB. 3.1.: Valeur des indices de sensibilité multidimensionnels en fonction des coefficients a, b et c du modèle (3.7) et des coefficients de corrélation des entrées.

- Effectivement, dans la situation (ii), les couples (X_3, X_4) et (X_5, X_6) sont symétriques dans le modèle, et donc ils ont la même importance. Dans la situation (iii) le produit X_5X_6 est multiplié par un coefficient 3, ce qui entraîne une importance prédominante du couple (X_5, X_6).
- Les situations (iv), (v) et (vi) illustrent la dépendance des indices $S_{\{3,4\}}$ et $S_{\{5,6\}}$ avec les coefficients de corrélation (S_{12} est aussi fonction de la corrélation, mais ceci est dû au dénominateur (la variance de Y) qui dépend de la corrélation des entrées). Comme les couples (X_3, X_4) et (X_5, X_6) sont sous la forme de produit dans le modèle, plus la corrélation est grande (positivement), plus le couple est important, et donc plus l'indice de sensibilité à une grande valeur.
- Dans la situation (iv) la corrélation du couple (X_3, X_4) est plus grande que celle de (X_5, X_6), et donc l'indice de sensibilité $S_{\{3,4\}}$ a une plus grande valeur que $S_{\{5,6\}}$.
- Dans la situation (v), le coefficient multiplicatif $c = 3$ a plus d'importance que la différence de corrélation entre (X_3, X_4) ($\rho_1 = 0.8$) et (X_5, X_6) ($\rho_2 = 0.3$), ce qui confère à $S_{\{5,6\}}$ une plus grande valeur qu'à $S_{\{3,4\}}$.
- Enfin, dans la situation (vi), les coefficients c et ρ_2 contribuent tous les deux à l'importance du couple (X_5, X_6) sur la variance de Y.

Cette simple application illustre bien l'utilité des indices de sensibilité multidimensionnels, par rapport à une fausse interprétation que l'on aurait pu faire en calculant les indices de sensibilité classiques sur ce modèle à entrées corrélées. En effet, si on calcule par exemple l'indice de sensibilité à la variable unidimensionnelle X_3 (ou X_4), on obtient :

$$S_3 = S_4 = \frac{2b^2{\rho_1}^2}{a^2 + b^2(1 + {\rho_1}^2) + c^2(1 + {\rho_2}^2)}.$$

Cet indice étant différent de 0 lorsque la corrélation n'est pas nulle, on serait tenté de conclure à une sensibilité de la variable X_3 seule, *i.e.* de premier ordre, (ou X_4), alors qu'il n'en est rien : les variables X_3 et X_4 n'apparaîssent dans le modèle que sous la forme d'un produit. Leur influence n'est pas due à chaque variable seule, mais aux interactions et à la corrélation entre X_3 et X_4.

Comme dans le cas unidimensionnel, les indices de sensibilité multidimensionnels ne peuvent pas toujours être calculés formellement, notamment lorsque la fonction du modèle n'est pas explicitement connue, ou alors trop complexe. En effet, dès que la fonction n'est plus linéaire, le calcul formel devient vite très compliqué.
Nous introduisons donc une méthode d'estimation de ces indices.

3.1.6. Estimation numérique des indices de sensibilité multidimensionnels

La méthode que nous proposons s'inspire de celle de Sobol en dimension 1, qui utilise des estimations de Monte Carlo.

Soit un N-échantillon de réalisations des variables d'entrée multidimensionnelles $(\vec{X}_1, \ldots, \vec{X}_{I+L})$:

$$\tilde{X}_{(N)} = (\vec{x}_{k1}, \ldots, \vec{x}_{k(I+L)})_{k=1..N}$$

L'espérance de la sortie Y du modèle, $\mathrm{E}[Y] = f_0$, et sa variance, $\mathrm{V}(Y) = V$, sont estimées par :

$$\hat{f}_0 = \frac{1}{N}\sum_{k=1}^{N} f(\vec{x}_{k1}, \ldots, \vec{x}_{k(I+L)}), \tag{3.8}$$

$$\hat{V} = \frac{1}{N}\sum_{k=1}^{N} f^2(\vec{x}_{k1}, \ldots, \vec{x}_{k(I+L)}) - \hat{f}_0{}^2. \tag{3.9}$$

L'estimation des indices de sensibilité de premier ordre (3.6) consiste à estimer la quantité :

$$V_j = \mathrm{V}(\mathrm{E}[Y|\vec{X}_j]) = \underbrace{\mathrm{E}[\mathrm{E}[Y|\vec{X}_j]^2]}_{U_j} - \mathrm{E}[\mathrm{E}[Y|\vec{X}_j]]^2 = U_j - \mathrm{E}[Y]^2,$$

la variance de Y étant estimée classiquement par (3.9). À l'image de la méthode de Sobol en dimension 1, nous utilisons deux échantillons de réalisations des variables d'entrée, que nous notons $\tilde{X}^{(1)}_{(N)}$ et $\tilde{X}^{(2)}_{(N)}$, pour estimer la quantité U_j :

$$\begin{aligned}\hat{U}_j = \frac{1}{N}\sum_{k=1}^{N} & \quad f\left(\vec{x}^{(1)}_{k1}, \ldots, \vec{x}^{(1)}_{k(j-1)}, \vec{x}^{(1)}_{kj}, \vec{x}^{(1)}_{k(j+1)}, \ldots, \vec{x}^{(1)}_{k(I+L)}\right) \\ \times & \quad f\left(\vec{x}^{(2)}_{k1}, \ldots, \vec{x}^{(2)}_{k(j-1)}, \vec{x}^{(1)}_{kj}, \vec{x}^{(2)}_{k(j+1)} \ldots, \vec{x}^{(2)}_{k(I+L)},\right).\end{aligned}$$

Les indices de sensibilité de premier ordre sont alors estimés par :

$$\hat{S}_j = \frac{\hat{U}_j - \hat{f}_0{}^2}{\hat{V}}.$$

Pour les indices de sensibilité de second ordre S_{jm}, où :

$$V_{jm} = \mathrm{V}(\mathrm{E}[Y|\vec{X}_j, \vec{X}_m]) - V_j - V_m = U_{jm} - \mathrm{E}[Y]^2 - V_j - V_m,$$

nous estimons les quantités $U_{jm} = \mathrm{E}[\mathrm{E}[Y|\vec{X}_j, \vec{X}_m]^2]$ de la même manière, en faisant varier toutes les variables sauf $\vec{X}_j$ et $\vec{X}_m$:

$$\begin{aligned}\hat{U}_{jm} = \frac{1}{N}\sum_{k=1}^{N} & \quad f\left(\vec{x}^{(1)}_{k1}, \ldots, \vec{x}^{(1)}_{k(j-1)}, \vec{x}^{(1)}_{kj}, \vec{x}^{(1)}_{k(j+1)}, \ldots, \vec{x}^{(1)}_{k(m-1)}, \vec{x}^{(1)}_{km}, \vec{x}^{(1)}_{k(m+1)}, \ldots, \vec{x}^{(1)}_{k(I+L)}\right) \\ \times & \quad f\left(\vec{x}^{(2)}_{k1}, \ldots, \vec{x}^{(2)}_{k(j-1)}, \vec{x}^{(1)}_{kj}, \vec{x}^{(2)}_{k(j+1)}, \ldots, \vec{x}^{(2)}_{k(m-1)}, \vec{x}^{(1)}_{km}, \vec{x}^{(2)}_{k(m+1)}, \ldots, \vec{x}^{(2)}_{k(I+L)}\right).\end{aligned}$$

L'indice S_{jm} est alors estimé par :

$$\hat{S}_{jm} = \frac{\hat{U}_{jm} - \hat{f}_0{}^2 - \hat{V}_j - \hat{V}_m}{\hat{V}}.$$

Et ainsi de suite pour les indices de sensibilité d'ordre supérieur.
Les indices de sensibilité totaux peuvent être estimés directement, en appliquant la même méthode qu'en

dimension 1 :

$$S_{T_j} = 1 - \frac{\mathrm{V}(\mathrm{E}[Y|\vec{X}_{\sim j}])}{\mathrm{V}(Y)} = 1 - \frac{V_{\sim j}}{V}.$$

Pour estimer $V_{\sim j} = \mathrm{E}[\mathrm{E}[Y|\vec{X}_{\sim j}]^2] - \mathrm{E}[\mathrm{E}[Y|\vec{X}_{\sim j}]]^2 = U_{\sim j} - \mathrm{E}[Y]^2$, on estime $U_{\sim j}$ par :

$$\begin{aligned}\hat{U}_{\sim j} = \frac{1}{N}\sum_{k=1}^{N} & \quad f\left(\vec{x}_{k1}^{(1)}, \ldots, \vec{x}_{k(j-1)}^{(1)}, \vec{x}_{kj}^{(1)}, \vec{x}_{k(j+1)}^{(1)}, \ldots, \vec{x}_{k(I+L)}^{(1)}\right) \\ \times & \quad f\left(\vec{x}_{k1}^{(1)}, \ldots, \vec{x}_{k(j-1)}^{(1)}, \vec{x}_{kj}^{(2)}, \vec{x}_{k(j+1)}^{(1)} \ldots, \vec{x}_{k(I+L)}^{(1)}, \right),\end{aligned}$$

et donc

$$\hat{S}_{T_j} = 1 - \frac{\hat{U}_{\sim j} - \hat{f}_0^{\,2}}{\hat{V}}.$$

3.1.7. Conclusion

L'analyse de sensibilité classique ne permet pas d'expliquer la sensibilité d'un modèle à ses variables d'entrée si celles-ci ne sont pas indépendantes. Nous proposons de regrouper les variables non indépendantes et d'exprimer la sensibilité à ces groupes de variables, et avons défini une méthode d'estimation de ces indices de sensibilité multidimensionnels, basée sur des simulations de Monte-Carlo.
Nous utilisons cette méthode pour l'analyse de sensibilité de modèles à entrées corrélées dans le domaine de l'ingénierie nucléaire dans la section suivante.

3.2. Applications liées à la dosimétrie neutronique pour des irradiations d'acier

L'évaluation de l'évolution des propriétés mécaniques d'aciers soumis à irradiation est basée sur des tests mécaniques dont l'interprétation utilise un indicateur de l'irradiation reçue, la fluence neutronique. La dosimétrie a pour objectif de déterminer cette fluence neutronique, définie comme le nombre de réactions neutroniques qui ont eu lieu au cours de l'irradiation d'un échantillon dans des conditions données (spectre neutronique, durée de l'irradiation, puissance du coeur...) à partir de l'étude d'un ou plusieurs dosimètres irradiés dans les mêmes conditions. Depuis 1989 un programme de recherche et de développement vise au CEA à améliorer la méthode de détermination de cette fluence et de l'incertitude associée. Ce programme est important vis-à-vis de l'évaluation de la durée de vie résiduelle des structures soumises à l'irradiation, et donc de la disponibilité de l'installation nucléaire.
Des échantillons de métal représentatifs sont alors irradiés conjointement avec des dosimètres par activation ou fissiles (nous en considérons 5 dans cette étude, nommés Det1, Det2, Det3, Det4 et Det5) destinés à enregistrer l'irradiation reçue. Après la fin de l'irradiation, les activités des dosimètres sont mesurées par un laboratoire du Département d'Études des Réacteurs du CEA. La fluence neutronique est ensuite évaluée à partir de ces mesures d'activités, des spectres neutroniques calculés par un code de calcul, des bibliothèques de sections efficaces et de l'historique d'irradiation selon un processus en quatre étapes : mesures d'activités d'un dosimètre, estimation de la fluence neutronique pour chacun des dosimètres, calcul du taux moyen de réactions en position de référence et estimation de la fluence neutronique reçue par le dispositif.
Nous nous intéressons dans un premier temps au modèle de calcul de l'indice épithermique, utilisé au cours de la deuxième étape. Puis, nous réalisons une analyse de sensibilité sur le logiciel utilisé pour la quatrième étape : Stay'SL [42], qui a été développé par l'ORNL[1] (USA) pour l'ajustement du spectre neutronique de référence.

3.2.1. Indice Epithermique

Une réaction neutronique entre un isotope et un neutron dans un spectre de neutrons peut être schématisée par l'équation suivante :

$$\tau = N \times \int_0^\infty e\ \sigma(e)\ n(e)\ de,$$

où :

$$\begin{array}{rcl} \tau & : & \text{taux de réaction par seconde,} \\ N & : & \text{nombre d'atomes de l'isotope considéré,} \\ \sigma(.) & : & \text{section efficace de l'isotope } (cm^2), \\ n(.) & : & \text{spectre de densité de neutrons } (n/cm^2 s). \end{array}$$

Cette équation de réaction est généralement utilisée sous la forme suivante :

$$\tau = N \times (\phi_{rap}\sigma_{rap}) + \phi_0 \times (I\rho + \sigma_0),$$

[1] *Oak Ridge National Laboratory*

où :

$\phi_{rap} = \int_1^\infty e\, n(e)\, de$: flux de neutrons «rapide» (énergie supérieure à 1MeV),

$\sigma_{rap} = \frac{\int_{E_c}^\infty e\, \sigma(e)\, n(e)\, de}{\int_1^\infty e\, n(e)\, de}$: section efficace effective «rapide», E_c étant un certain niveau d'énergie fixé,

ϕ_0 : flux thermique conventionnel $(n/cm^2 s)$(flux de neutrons ramené à une énergie de $E_0 = 0.0253eV$ correspondant à une vitesse de $V_0 = 2200m/s,$ vitesse la plus probable dans le spectre maxwellien à 20˚C),

I : intégrale de résonance,

$\rho = \frac{\phi_{Uepi}}{\phi_0}$: **indice épithermique**, où ϕ_{Uepi} est le flux épithermique (intégrale du spectre, au dessus de la coupure du cadmium) par unité de léthargie $(n/cm^2 s)$,

σ_0 : section efficace thermique (cm^2).

Pour calculer l'indice épithermique, nous utilisons un modèle (décrit ci-dessous), qui comporte quatre variables d'entrée aléatoires. L'indice épithermique issu de ce calcul est donc une variable aléatoire, et chaque calcul de cet indice est une réalisation de cette variable. L'objectif de cette analyse de sensibilité est d'améliorer la connaissance de cet indice, via la détermination des variables d'entrée qui contribuent le plus à sa variance. Cette étude pourra servir à choisir des actions pour réduire l'incertitude sur cet indice épithermique.

3.2.1.1. Le modèle

Un des modèles de calcul de l'indice épithermique utilisé par le CEA possède quatre variables d'entrée aléatoires supposées gaussiennes, dont deux sont corrélées.
Les variables d'entrée sont donc supposées :

$$\begin{aligned}
\text{intégrale de résonance du Co59 (Cobalt 59)} &: X_1 \sim \mathcal{N}(72, 7.2^2),\\
\text{facteur Fcd} &: X_2 \sim \mathcal{N}(\log(1.01989), 0.0147051^2),\\
\text{activité du dosimètre Co59 "nu"} &: X_3 \sim \mathcal{N}(4.703 \times 10^7, 1147732^2),\\
\text{activité du dosimètre Co59 sous Cadmium} &: X_4 \sim \mathcal{N}(2.522 \times 10^7, 615368^2),\\
\text{avec} &\quad \rho_{X_3 X_4} = 0.85.
\end{aligned}$$

Les écarts-types des variables d'entrée sont en réalité connus comme un pourcentage d'incertitude par rapport à leur espérance :

variable	X_1	X_2	X_3	X_4
incertitude	10%	1.44%	2.44%	2.44%

L'indice épithermique, Y, est alors calculé à partir des variables d'entrée par :

$$Y = \frac{(1.008843 - 0.02114316X_1 + 9.858080.10^{-5}{X_1}^2 + 1.931988.10^{-8}X_4)\exp(X_2)}{(1 - \exp(X_2)\frac{X_4}{X_3})(-0.00575077 + 3.73935.10^{-8}X_3)}. \quad (3.10)$$

Une analyse d'incertitude permet d'obtenir une estimation par Monte Carlo de l'espérance et de la variance de l'indice épithermique défini par ce modèle (3.10). Le nombre de simulations de Monte Carlo utilisé est de 10000. Nous répétons 30 fois cette estimation pour obtenir une idée de la variabilité des estimations. Le choix de 10000 simulations de Monte Carlo suffit à obtenir une variabilité fine des estimations des premiers

moments de l'indice épithermique :

$$\begin{aligned} \hat{\mathrm{E}}[Y] &= 0.6303 \quad (0.0008), \\ \hat{\mathrm{V}}(Y) &= 0.0050 \quad (0.0001), \\ \hat{\sigma}_Y &= 0.0707 \quad (0.0007), \end{aligned}$$

où figure entre parenthèse l'écart-type (absolu) des 30 estimations.

Nous présentons maintenant une analyse de sensibilité du modèle (3.10), qui permet de déterminer quelles sont les variables d'entrée qui contribuent le plus à la variance de l'indice épithermique.

3.2.1.2. Analyse de sensibilité sur le modèle de calcul de l'indice épithermique

Les variables d'entrée n'étant pas indépendantes, on considère pour cette analyse X_1, X_2 et la variable bidimensionnelle (X_3, X_4).
Nous réalisons dans un premier temps une étude sur le nombre de simulations de Monte Carlo nécessaires à une estimation des indices de sensibilité que l'on jugera correcte. Puis nous présentons les résultats de l'analyse de sensibilité, que l'on utilise ensuite pour améliorer la connaissance de l'indice épithermique, c'est-à-dire pour diminuer sa variance.

Etude du nombre de simulations de Monte Carlo nécessaires

Le nombre de simulations nécessaires aux estimations des indices de sensibilité est propre à chaque modèle et à la précision désirée, sachant qu'en règle générale, plus le modèle sera complexe (nombre de variables d'entrée et forme du modèle) plus ce nombre devra être important. Néanmoins, il arrive souvent en pratique, pour les modèles à temps de calcul important, que ce choix du nombre de simulations de Monte Carlo à effectuer soit dicté par les moyens dont on dispose (temps, machines...).
La précision des estimations des indices de sensibilité doit permettre de distinguer une hiérarchie au sein des variables d'entrée en fonction de leur importance sur la variance de la sortie. Une incertitude de 10% pour des indices de sensibilité assez grands sera acceptable si elle permet néanmoins cette distinction.
L'analyse de sensibilité du modèle (3.10) étant peu coûteuse en temps de calcul, nous pouvons nous permettre de réaliser 100 fois l'analyse (calculs des indices), pour différentes tailles d'échantillon de Monte Carlo, variant de 5000 à 80000, afin de déterminer la taille adéquate.
Le tableau 3.2 présente les moyennes (m) et écarts-types (σ) des 100 évaluations d'indices. Les écarts-types permettent de définir un pourcentage d'incertitude sur la valeur estimée. Les valeurs estimées sont arrondies à zéro ($\simeq 0$) lorsqu'elles sont proches de 0 et que leur écart-type est plus grand que leur moyenne.
Une taille d'échantillon de 5000 permet d'obtenir une bonne estimation de l'indice S_1 (incertitude inférieure à 10%). Néanmoins, cette taille n'est pas suffisante, puisque la somme des indices de premier ordre est alors de 0.838, il reste donc plus de 16% de la variance de l'indice épithermique à expliquer. Pour les indices totaux, l'estimation est moins bonne que pour S_1, puisque l'on obtient 20% d'incertitude sur les indices à grande valeur, et plus de 50% d'incertitude sur les petits indices inférieurs à 0.1.
Une taille d'échantillon de 40000 permet de faire apparaître toutes les parts de variance de l'indice épithermique (S_1, S_2, $S_{\{3,4\}}$), puisque la somme des indices de sensibilité est alors quasiment égale à 1. Comme seuls les indices de premier ordre sont non nuls et que leur somme est égale à 1, il est alors possible d'en déduire que les interactions n'ont aucun effet. L'analyse de sensibilité peut donc se ramener à la seule estimation des indices de premier ordre. Nous continuerons néanmoins à estimer tous les indices, n'ayant aucun problème de temps de calcul.
Finalement, une taille de 80000 permet d'obtenir tous les indices de premier ordre et totaux avec une précision acceptable (suffisante pour hiérarchiser les variables d'entrée). Cette taille peut paraître importante, mais cela est en partie due à la variance de Y qui est très proche de zéro ($\simeq 0.005$). Les imprécisions numériques sont alors accrues lors des estimations d'indices puisqu'elles nécessitent une division par cette

N	5000		10000		20000		40000		80000	
indice	m	σ	m	σ	m	σ	m	σ	m	σ
S_1	0.838	0.073	0.848	0.059	0.844	0.038	0.848	0.023	0.846	0.017
S_2	$\simeq 0$		$\simeq 0$		0.084	0.084	0.090	0.056	0.084	0.039
$S_{\{3,4\}}$	$\simeq 0$		$\simeq 0$		$\simeq 0$		0.074	0.059	0.070	0.059
S_{12}	$\simeq 0$		$\simeq 0$		$\simeq 0$		$\simeq 0$		$\simeq 0$	
$S_{1\{3,4\}}$	$\simeq 0$		$\simeq 0$		$\simeq 0$		$\simeq 0$		$\simeq 0$	
$S_{2\{3,4\}}$	$\simeq 0$		$\simeq 0$		$\simeq 0$		$\simeq 0$		$\simeq 0$	
$S_{12\{3,4\}}$	$\simeq 0$		$\simeq 0$		$\simeq 0$		$\simeq 0$		$\simeq 0$	
S_{T_1}	0.857	0.169	0.843	0.111	0.844	0.081	0.842	0.054	0.847	0.038
S_{T_2}	0.088	0.048	0.084	0.041	0.083	0.032	0.084	0.020	0.084	0.015
$S_{T_{\{3,4\}}}$	0.073	0.050	0.069	0.036	0.073	0.022	0.069	0.018	0.070	0.011

TAB. 3.2.: Évolution des indices de sensibilité du modèle de calcul de l'indice épithermique en fonction du nombre N de simulation de Monte Carlo utilisée

variance.

Remarque. *Pour une taille d'échantillon de Monte Carlo de N, le nombre réel de simulations des variables d'entrée nécessaires à une analyse de sensibilité est $2N$, puisque l'estimation des indices nécessite deux jeux de simulations du modèle. Le nombre d'appels à la fonction du modèle pour estimer k indices de sensibilité est $N\times(k+1)$. Pour un modèle à p variables d'entrée, l'estimation de tous ses indices de sensibilité demande donc $N \times (2^p)$ appels à sa fonction, ce qui représente ici $8N$ appels à la fonction (3.10). L'estimation des indices de premier ordre et totaux ne demande que $N \times (6 + 1) = 7N$ appels à la fonction.*

Comme le modèle (3.10) ne pose aucun problème de temps de calcul, nous choisissons pour toutes les analyses de sensibilité suivantes une taille d'échantillon de 100000. Chaque estimation sera répétée 30 fois, afin de vérifier leur précision. Ce protocole demande un temps de calcul d'environ 13 secondes pour une programmation Matlab sur un ordinateur Pentium IV à 3GHz.

Analyse de sensibilité de référence

L'analyse de sensibilité du modèle (3.10) donne les indices de sensibilité présentés Table 3.3 et figure 3.3 (les barres d'erreur sont étendues à plus ou moins deux écarts-types). Les indices d'interaction ne figurent pas dans cette table, car leur valeur estimée est nulle.

	S_1	S_2	$S_{\{3,4\}}$	S_{T_1}	S_{T_2}	$S_{T_{\{3,4\}}}$
moyenne	0.846	0.083	0.071	0.848	0.083	0.071
écart-type	0.018	0.038	0.046	0.040	0.013	0.014

TAB. 3.3.: Indices de sensibilité du modèle de calcul de l'indice épithermique

Une remarque intéressante, qui sera confirmée et que nous développerons avec l'application suivante sur le code Stay'SL, est que l'estimation des indices de sensibilité totaux est moins bonne (variabilité plus grande) que celle des indices de premier ordre, lorsque les indices ont une grande valeur, et meilleure lorsqu'ils ont une petite valeur.
L'intégrale de résonance du Co59 (X_1) est de loin la variable la plus influente sur la variance de l'indice épithermique. En effet, la part de variance de ce dernier due à X_1 est de près de 85% de la variance totale. Les autres variables ont alors un effet bien moins important.

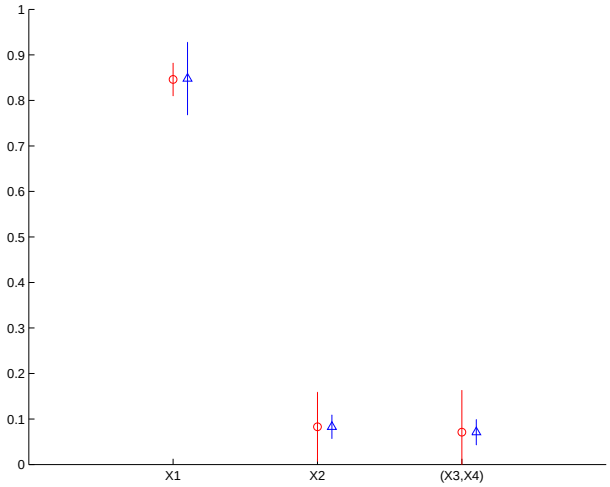


FIG. 3.3. : Indices de sensibilité de premier ordre (○) et totaux (△) du modèle de calcul de l'indice épithermique.

Si l'on veut mieux connaître la valeur de l'indice épithermique, c'est-à-dire réduire la variance issue de son calcul, il est nécessaire d'essayer de réduire l'incertitude sur X_1, qui est actuellement de 10% (écart-type égal à 10% de la valeur moyenne). Une amélioration de la détermination du facteur Fcd ou des activités des dosimètres du Cobalt ne permettrait de réduire la valeur de la variance sur Y que d'au plus 10%.

Illustration de l'impact d'une réduction de l'incertitude sur les variables d'entrée

Nous supposons pouvoir réduire l'incertitude sur l'intégrale de résonance du Co59 à 5%, puis à 1%.
Sans toucher aux autres variables, et en réduisant l'incertitude sur X_1 à 5% (réduction de moitié), on obtient une variance et un écart-type de Y de :

$$\hat{\mathrm{V}}(Y) = 0.0018, \qquad \hat{\sigma}_Y = 0.0423, \qquad \hat{\mathrm{E}}[Y] = 0.6252,$$

d'où une réduction de l'écart-type, et donc de l'incertitude sur l'indice épithermique de 41% (l'écart-type initial étant 0.0707). L'impact sur l'espérance de Y est alors une légère diminution, ce qui s'explique par la présence du terme ${X_1}^2$ dans le modèle (3.10) (l'espérance de Y n'est donc pas fonction que de l'espérance de X_1, mais aussi de sa variance). L'analyse de sensibilité donne les résultats suivants :

	S_1	S_2	$S_{\{3,4\}}$	S_{T_1}	S_{T_2}	$S_{T_{\{3,4\}}}$
moyenne	0.578	0.238	0.182	0.579	0.240	0.183
écart-type	0.043	0.061	0.055	0.041	0.028	0.028

La variable la plus importante est toujours l'intégrale de résonance du Co59, c'est encore sur elle qu'il faudrait porter les efforts pour encore diminuer l'incertitude sur l'indice épithermique.

En réduisant l'incertitude sur l'intégrale de résonance à 1%, on obtient une variance et un écart-type de Y, ainsi qu'une moyenne de :

$$\hat{\mathrm{V}}(Y) = 0.0008, \qquad \hat{\sigma}_Y = 0.0282, \qquad \hat{\mathrm{E}}[Y] = 0.6237,$$

d'où une diminution de 60% par rapport à l'écart-type d'origine. L'espérance de Y décroît encore légèrement. Les indices de sensibilité alors obtenus sont :

	S_1	S_2	$S_{\{3,4\}}$	S_{T_1}	S_{T_2}	$S_{T_{\{3,4\}}}$
moyenne	0.038	0.507	0.425	0.052	0.522	0.441
écart-type	0.099	0.062	0.077	0.026	0.075	0.054

La variable X_1 n'a cette fois quasiment plus d'influence sur la variance de Y. Toutefois, on remarque que la somme des indices de sensibilité est plus petite que 1, et qu'il reste encore près de 3 à 4% de la variance de l'indice épithermique à expliquer. Cela peut être dû soit aux imprécisions d'estimation, soit alors à l'apparition de l'influence d'une interaction entre certaines variables. Une nouvelle analyse de sensibilité, avec cette fois des échantillons de Monte Carlo de taille 1 million, nous montre que cela est dû aux imprécisions numériques, puisque l'on obtient cette fois une somme des indices de premier ordre quasiment égale à 1. Les estimations, répétées 30 fois, sont les suivantes :

	S_1	S_2	$S_{\{3,4\}}$	S_{T_1}	S_{T_2}	$S_{T_{\{3,4\}}}$
moyenne	0.053	0.520	0.433	0.052	0.514	0.432
écart-type	0.033	0.019	0.026	0.008	0.024	0.018

Le temps de calcul associé à cette nouvelle taille d'échantillon de 1 million est 2 minutes et 20 secondes (Matlab, Pentium IV, 3GHz).
Si on veut continuer à diminuer la variance de Y, il faut maintenant porter les efforts sur X_2 et (X_3, X_4). Et ainsi de suite.

Afin de comparer aux réductions obtenues en travaillant sur X_1, on estime la variance que l'on aurait obtenue en divisant par 10 l'incertitude sur toutes les autres variables (X_2, X_3 et X_4), c'est-à-dire pour les incertitudes suivantes :

variable	X_1	X_2	X_3	X_4
incertitude	10%	0.144%	0.244%	0.244%

On obtient alors :

$$\hat{\mathrm{V}}(Y) = 0.0042, \qquad \hat{\sigma}_Y = 0.0648, \qquad \hat{\mathrm{E}}[Y] = 0.6292,$$

soit une diminution de 8% de l'écart-type de l'indice épithermique. L'espérance de Y est légèrement plus faible que l'espérance d'origine, mais cette différence n'est pas statistiquement significative, compte-tenu de l'incertitude d'estimation de l'espérance d'origine.
Quant aux indices de sensibilité, on obtient logiquement $S_1 \simeq S_{T_1} \simeq 1$.

En divisant par deux l'incertitude sur l'intégrale de résonance du Co59, on diminue l'incertitude sur l'indice épithermique de 41%, alors qu'en divisant par dix l'incertitude sur les trois autres variables, on n'obtient une réduction de seulement 8%. Ceci illustre l'intérêt de travailler sur les variables les plus importantes (d'un point de vue variance).

3.2.1.3. Conclusion de l'étude

Cette analyse de sensibilité montre clairement que la variable qui contribue le plus à l'incertitude sur l'indice épithermique, est l'intégrale de résonance du Co59. Une diminution de moitié de son incertitude (écart-type) entraîne une diminution de près de 41% de celle sur l'indice épithermique. Si l'incertitude sur l'intégrale de résonance du Co59 est irréductible (impossibilité de faire de nouvelles mesures...), il est toujours possible de diminuer celles sur les autres variables (en priorité sur le facteur Fcd), mais le gain sera beaucoup moins important.
Une autre conclusion intéressante concerne la valeur de l'indice épithermique calculé. On remarque qu'en diminuant sa variance, sa valeur a aussi tendance à décroître légèrement. Une trop grande incertitude sur cet indice a donc tendance à surestimer sa valeur.
Si cette étude de sensibilité avait été menée avec une analyse de sensibilité classique, les résultats obtenus auraient été semblables. Cela aurait été une coïncidence due au fait que la variable la plus importante X_1 est indépendante des autres. Si le poids des variables corrélées avait été plus important, l'analyse classique aurait pû conduire à de fausses conclusions, comme cela a été illustré par les applications théoriques de la section

précédente. Nous présentons une autre étude de sensibilité d'un modèle plus complexe, où les variables les plus importantes sont corrélées.

3.2.2. Code Stay'SL

L'application que nous présentons sur le code Stay'SL permet de tester et prouver l'applicabilité du calcul des indices multidimensionnels sur un problème réel complexe (185 variables d'entrée, structure de corrélation au sein de ces entrées...).

3.2.2.1. Le modèle

Le code Stay'SL a pour objectif d'estimer un spectre de neutron, c'est-à-dire le nombre de neutrons par groupe d'énergie, à partir des éléments suivants :

- les taux de réactions mesurés, c'est-à-dire le nombre de réactions nucléaires par seconde, pour chaque détecteur,
- le spectre de référence lié au réacteur nucléaire,
- les sections efficaces pour chaque combinaison de réaction et de groupe d'énergie, issue de librairies internationales de données nucléaires.

Pour cette étude, les taux de réactions mesurés et le spectre d'origine ont été normalisés, par soucis de confidentialité. Les librairies de sections efficaces sont du domaine publique.
Le processus d'ajustement du spectre de neutron repose sur une méthode de type moindres carrés, opérant la minimisation de la norme de la différence entre les données d'entrées et leur contrepartie ajustée (dont seul le spectre est effectivement calculé par Stay'SL). La méthode d'ajustement est décrite dans [42].

Les variables d'entrée

Le code Stay'SL comporte 185 variables d'entrée, supposées dans cette étude toutes gaussiennes, que nous nommons X_1 à X_{185} :

- $X_1, ..., X_{150}$: 150 sections efficaces (pour chaque combinaison des 5 dosimètres fissiles (Det1, Det2, Det3, Det4 et Det5) et des 30 groupes d'énergie). Elles sont données sous la forme d'une matrice de valeur moyenne et d'une matrice de variance. La matrice de variance est une matrice 150×150 par bloc de 30×30.
- $X_{151}, ..., X_{180}$: flux de neutrons pour chacun des 30 groupes d'énergie. Les flux sont donnés par une matrice de valeur moyenne et une matrice de variance absolue. Ces données sont normalisées.
- $X_{181}, ..., X_{185}$: les 5 taux de réactions mesurés : τ_{Det1}, τ_{Det2}, τ_{Det3}, τ_{Det4}, τ_{Det5}. On dispose de leur valeur moyenne ainsi que des corrélations entre τ_{Det1} et τ_{Det2}, ainsi qu'entre τ_{Det3} et τ_{Det4}. Ces données sont normalisées.

La figure 3.4 représente les matrices de variance des variables d'entrée.

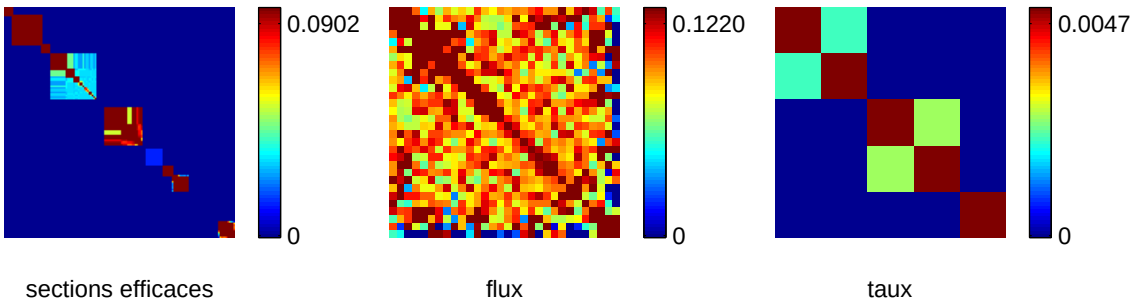


FIG. 3.4. : Matrices de variance pour les sections efficaces, les flux et les taux mesurés.

Les variables de sortie

Le code Stay'SL calcule 21 variables de sortie :

- $Y_1, ..., Y_5$: taux de réactions ajustés de Det1, Det2, Det3, Det4, Det5,
- $Y_6, ..., Y_{10}$: flux de neutrons ajustés pour Det1, Det2, Det3, Det4, Det5,
- Y_{11} : somme des flux de neutrons pondérés qui ont une énergie supérieure à 1 MeV,
- Y_{12} : somme des flux de neutrons pondérés qui ont une énergie supérieure à 0.1 MeV,
- Y_{13} : χ^2 issu de l'ajustement,
- Y_{14} : flux de neutrons qui ont une énergie supérieure à 1 MeV,
- Y_{15} : flux de neutrons qui ont une énergie supérieure à 0.1 MeV,
- Y_{16} : rapport des deux flux de neutrons précédents (Y_{14} et Y_{15}),
- $Y_{17}, ..., Y_{21}$: sections efficaces ajustées pour les dosimètres Det1, Det2, Det3, Det4, Det5.

Parmi ces variables de sortie, nous nous intéressons à neuf d'entre elles : Y_{11}, Y_{12}, Y_{14}, Y_{16} puis Y_1 à Y_5.
La version du code Stay'SL utilisée est programmée en Matlab, et prend un temps de calcul inférieur à la seconde.

3.2.2.2. Analyse d'incertitude

Une analyse d'incertitude préliminaire nous permet d'estimer l'espérance et la variance des différentes variables de sortie de Stay'SL. Ces estimations sont faites par Monte Carlo, en utilisant des échantillons de simulations du modèle de taille 10000. Nous réalisons 20 fois ces estimations, et présentons la moyenne ainsi que l'écart-type de ces 20 estimations dans le tableau 3.4.
La taille de 10000 suffit pour estimer les moments des vartiables de sortie avec assez de précision (écart-type

	espérance		variance	
	m	σ	m	σ
Y_1	1.091×10^{-1}	1.202×10^{-4}	1.090×10^{-4}	1.805×10^{-6}
Y_2	1.32×10^{-2}	3.219×10^{-6}	1.695×10^{-7}	2.120×10^{-9}
Y_3	2.097×10^{-3}	8.331×10^{-7}	5.245×10^{-9}	7.757×10^{-11}
Y_4	3.087×10^{-3}	1.654×10^{-6}	3.869×10^{-8}	5.183×10^{-10}
Y_5	1.861×10^{-5}	1.186×10^{-8}	$< 10^{-12}$	$< 10^{-14}$
Y_{11}	4.012×10^{-2}	1.737×10^{-5}	2.977×10^{-6}	4.590×10^{-8}
Y_{12}	1.379×10^{-1}	6.542×10^{-5}	3.340×10^{-5}	5.425×10^{-7}
Y_{14}	4.110×10^{-1}	8.091×10^{-6}	9.374×10^{-7}	1.220×10^{-8}
Y_{15}	1.413×10^{-1}	3.212×10^{-5}	9.426×10^{-6}	1.262×10^{-8}
Y_{16}	3.438	5.210×10^{-4}	3.219×10^{-3}	5.522×10^{-7}

TAB. 3.4.: Analyse d'incertitude des différentes variables de sortie du code Stay'SL.

de l'ordre de 10^{-2} à 10^{-3} fois la moyenne).

3.2.2.3. Analyse de sensibilité

Comme les variables d'entrée ne sont pas indépendantes, nous réalisons une analyse de sensibilité multidimensionnelle. Nous regroupons les 185 variables d'entrée non indépendantes en 9 groupes de variables indépendantes :

- $\vec{X}_1 = (X_1, ..., X_{30})$: sections efficaces sur 30 groupes d'énergie pour le dosimètre Det1,

- $\vec{X}_2 = (X_{31}, ..., X_{60})$: sections efficaces sur 30 groupes d'énergie pour le dosimètre Det2,
- $\vec{X}_3 = (X_{61}, ..., X_{90})$: sections efficaces sur 30 groupes d'énergie pour le dosimètre Det3,
- $\vec{X}_4 = (X_{91}, ..., X_{120})$: sections efficaces sur 30 groupes d'énergie pour le dosimètre Det4,
- $\vec{X}_5 = (X_{121}, ..., X_{150})$: sections efficaces sur 30 groupes d'énergie pour le dosimètre Det5,
- $\vec{X}_6 = (X_{151}, ..., X_{180})$: flux de neutrons pour chacun des 30 groupes d'énergie,
- $\vec{X}_7 = (X_{181}, X_{182})$: taux de réactions (τ_{Det1},τ_{Det2}),
- $\vec{X}_8 = (X_{183}, X_{184})$: taux de réactions (τ_{Det3},τ_{Det4}),
- $\vec{X}_9 = (X_{185})$: taux de réactions (τ_{Det5}).

Une analyse de sensibilité est réalisée pour chacune des variables de sortie qui nous intéressent : Y_{11}, Y_{12}, Y_{14}, Y_{16}, puis Y_1 à Y_5. Comme le nombre de variables d'entrée est important, la taille des échantillons de Monte Carlo utilisés doit aussi être relativement importante (nous la choisirons de 100000), ce qui pose alors des problèmes de temps de calcul. Nous nous restreindrons au calcul des indices de sensibilité de premier ordre et totaux. Le nombre d'appels au code Stay'SL sera pour une analyse égal à 1.9×10^6 (100000 fois le nombre d'indices estimés plus un). Une telle analyse demande environ 12 heures de calcul, sur un ordinateur Pentium IV à 3GHz.

Précisons que l'on ne s'intéresse pas à la valeur exacte des indices de sensibilité mais à leur ordre de grandeur, l'objectif étant de déterminer une hiérarchie au sein des variables d'entrée en fonction de leur influence sur la variance de la sortie. Ainsi, nous verrons avec les premières applications que la taille d'échantillon choisie de 100000 est suffisante.

Analyse de sensibilité de la somme des flux de neutrons pondérés qui ont une énergie supérieure à 1 MeV (Y_{11}) et à 0.1 MeV (Y_{12})

Afin d'obtenir une idée de la précision des estimations des indices de sensibilité, nous réalisons les premières analyses de sensibilité sur Y_{11} et Y_{12} respectivement 7 et 4 fois. Les figures 3.5 et 3.6 et les Tables 3.5 et 3.6 présentent les résultats de ces analyses, avec, pour chaque variable multidimensionnelle, les moyennes des indices de sensibilité de premier ordre et totaux sur les 7 et 4 calculs. Les barres d'erreur des figures représentent l'étendue entre l'estimation minimale et maximale sur les 7 (respectivement 4) répétitions.

	sections efficaces du dosimètre					flux	taux de réactions		
	Det1	Det2	Det3	Det4	Det5		(τ_{Det1},τ_{Det2})	(τ_{Det3},τ_{Det4})	τ_{Det5}
indice S_j	**0.501**	0.039	0.066	**0.182**	0.029	**0.112**	**0.192**	0.069	0.026
indice S_{T_j}	**0.514**	0.007	0.015	**0.154**	0.006	**0.109**	**0.159**	0.052	0.001

TAB. 3.5.: Indices de sensibilité de premier ordre et totaux pour la sortie Y_{11} du code Stay'SL.

	sections efficaces du dosimètre					flux	taux de réactions		
	Det1	Det2	Det3	Det4	Det5		(τ_{Det1},τ_{Det2})	(τ_{Det3},τ_{Det4})	τ_{Det5}
indice S_j	**0.522**	0.035	0.044	**0.219**	0.041	**0.126**	**0.190**	0.060	0.034
indice S_{T_j}	**0.490**	$\simeq 0$	0.011	**0.184**	0.007	**0.091**	**0.156**	0.026	$\simeq 0$

TAB. 3.6.: Indices de sensibilité de premier ordre et totaux pour la sortie Y_{12} du code Stay'SL.

Ces résultats montrent que la variable de sortie Y_{11} est sensible aux mêmes variables d'entrée que Y_{12}, et ce dans les mêmes proportions. La variable qui contribue à la plus grande part de la variance des sorties Y_{11} et Y_{12}, est la variable $\vec{X}_1$, les sections efficaces du dosimètre Det1. La variance de ces sections efficaces est responsable de près de la moitié de la variance totale. Ensuite, les deuxièmes variables les plus importantes sont $\vec{X}_4$ et $\vec{X}_7$ (sections efficaces du dosimètre Det4 et taux de réactions de Det1 et de Det2), avec des parts de la variance totale d'un peu moins de 20%. Enfin vient la variable $\vec{X}_6$ (flux de neutrons), qui contribue à

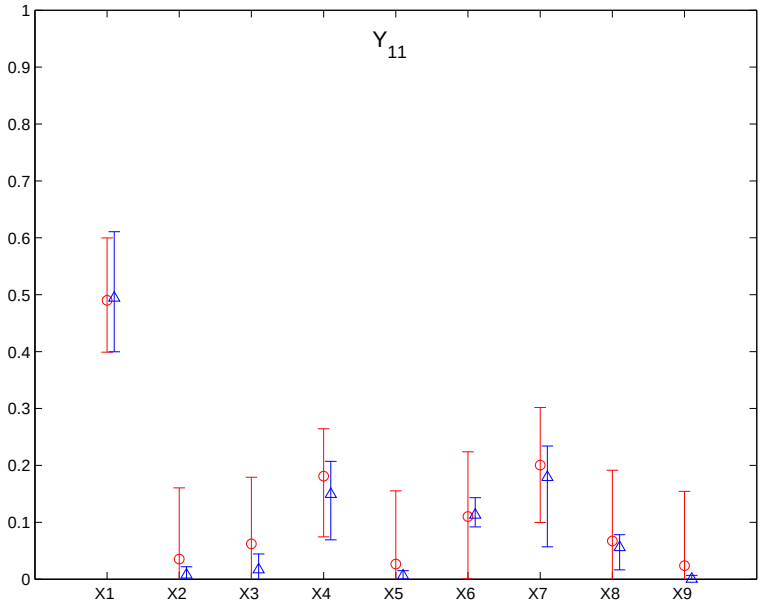


FIG. 3.5. : Indices de sensibilité de premier ordre (○) et totaux (△) pour la sortie Y_{11} du code Stay'SL.

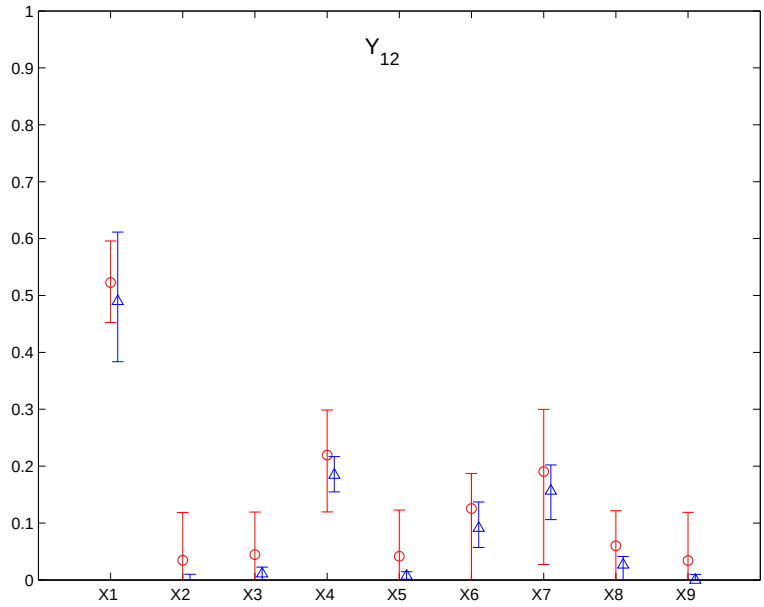


FIG. 3.6. : Indices de sensibilité de premier ordre (○) et totaux (△) pour la sortie Y_{12} du code Stay'SL.

un peu moins de 10% de la variance totale, et la variable $\vec{X}_8$ (taux de réactions de Det3 et de Det4), qui a une importance légèrement non nulle.

La première constatation importante à faire, et qui sera vraie pour toutes les autres variables de sortie, est que les indices totaux sont équivalents d'un point de vue valeur estimée, compte-tenu des incertitudes d'estimation, aux indices de premier ordre. Cela signifie que les interactions n'ont qu'une très faible influence sur la variance des différentes sorties étudiées (ici Y_{11} et Y_{12}).

La seconde concerne la qualité d'estimation des indices de sensibilité de premier ordre et totaux. Les estimations des indices totaux sont meilleures (variabilité plus fine) que celles des indices de premier ordre,

lorsque les indices estimés ont une petite valeur. Par contre, lorsque les indices à estimer ont une valeur importante, ce sont les indices de premier ordre qui sont les mieux estimés. Ce comportement avait déjà été remarqué de façon encore plus flagrante sur l'application concernant l'indice épithermique.
Expliquons le en considérant S_1 et S_{T_1} ainsi que S_4 et S_{T_4}. Les indices S_1 et S_{T_1} ont des valeurs estimées proches (environ 0.5), mais la variabilité des estimations est plus importante pour S_{T_1} que pour S_1. Les indices S_4 et S_{T_4}, qui ont une plus petite valeur (environ 0.2), ont un comportement inverse, puisque c'est la variabilité de l'indice de premier ordre qui cette fois est plus importante que celle de l'indice total.
L'estimation de S_{T_1} nécessite l'estimation de la variance de l'espérance de la sortie du modèle conditionnellement à toutes les variables sauf la variable $\vec{X}_1$, autrement dit l'estimation de la variance d'une fonction de $\vec{X}_1$. L'estimation de S_1 nécessite quant à elle l'estimation de la variance de l'espérance de la sortie du modèle conditionnellement à la variable $\vec{X}_1$, c'est-à-dire l'estimation de la variance d'une fonction des variables $\vec{X}_2, \ldots, \vec{X}_9$. De même pour les indices relatifs à $\vec{X}_4$.
La variabilité issue de l'estimation de la variance d'une fonction d'une variable qui contribue beaucoup à la variance totale ($\vec{X}_1$), s'avère plus grande que la variabilité issue de l'estimation de la variance d'une fonction de plusieurs variables qui contribuent peu à la variance totale ($\vec{X}_2, \ldots, \vec{X}_9$). C'est la raison qui rend la variabilité de S_{T_1} plus grande que celle de S_1. En revanche, la variabilité issue de l'estimation de la variance d'une fonction d'une variable qui contribue peu à la variance totale ($\vec{X}_4$) est plus fine que celle issue de l'estimation de la variance d'une fonction de plusieurs variables ($\vec{X}_1, \ldots, \vec{X}_3, \vec{X}_5, \ldots, \vec{X}_9$) dont une ($\vec{X}_1$) contribue beaucoup à la variance totale. C'est pourquoi la variabilité de S_{T_4} est plus petite que celle de S_4.

Sur ces premières applications, l'incertitude absolue sur les estimations des indices de sensibilité, est de l'ordre de 0.1 à 0.2. Nous supposons que les analyses de sensibilité des autres variables de sortie sont sujettes à une incertitude équivalente. Il sera important, lors de l'interprétation des résultats, de tenir compte de cette incertitude. De plus, les imprécisions numériques d'estimation engendrent parfois des indices estimés légèrement négatif ou légèrement plus grand que 1. Nous arrondissons alors ces valeurs respectivement à 0 et à 1.
Toutes les analyses réalisées sur les autres variables de sortie utilisent des échantillons de Monte Carlo de taille 100000.

Analyse de sensibilité du flux de neutrons qui ont une énergie supérieure à 1 MeV (Y_{14})

Le tableau 3.7 présente les valeurs des indices de sensibilité. Les indices ayant les plus grandes valeurs sont en caractères gras.

	sections efficaces du dosimètre					flux	taux de réactions		
	Det1	Det2	Det3	Det4	Det5		(τ_{Det1},τ_{Det2})	(τ_{Det3},τ_{Det4})	τ_{Det5}
indice S_j	0.080	0.092	0.094	0.080	0.070	**0.771**	**0.237**	0.105	0.085
indice S_{T_j}	0.003	0.013	0.016	0.003	$\simeq 0$	**0.694**	**0.161**	0.028	0.009

TAB. 3.7.: Indices de sensibilité de premier ordre et totaux pour la sortie Y_{14} du code Stay'SL.

La variance de la variable Y_{14} (flux de neutrons qui ont une énergie supérieure à 1 MeV), est due majoritairement au flux de neutrons, ce qui est somme toute logique. Le reste de la variance de Y_{14} est dû essentiellement aux taux de réactions de Det1 et de Det2.

Analyse de sensibilité du rapport du flux de neutrons qui ont une énergie supérieure à 1 MeV et du flux de neutrons qui ont une énergie supérieure à 0.1 MeV (Y_{16})
Les résultats sont présentés dans la Table 3.8.

	sections efficaces du dosimètre					flux	taux de réactions		
	Det1	Det2	Det3	Det4	Det5		(τ_{Det1},τ_{Det2})	(τ_{Det3},τ_{Det4})	τ_{Det5}
indice S_j	0.074	0.048	0.053	0.077	0.067	**0.986**	0.157	0.061	0.076
indiceS_{T_j}	0.001	$\simeq 0$	0.010	0.006	0.002	$\simeq$**1**	0.081	$\simeq 0$	0.004

TAB. 3.8.: Indices de sensibilité de premier ordre et totaux pour la sortie Y_{16} du code Stay'SL.

La variance de Y_{16} n'est due qu'à une variable, qui représente le flux de neutrons, ce qui ici encore est cohérent, puisque la variable Y_{16} est le rapport du flux de neutrons qui ont une énergie supérieure à 1 MeV (Y_{14}) et du flux de ceux qui ont une énergie supérieure à 0.1 MeV (Y_{15}). Les sections efficaces et les taux de réactions n'ont pas d'influence sur la variance du rapport des flux. Par rapport à l'analyse de Y_{14}, les flux de neutrons ont ici encore plus d'importance, et les taux de réactions de Det1 et de Det2 n'en ont quasiment plus. Ceci semble indiquer que Y_{15} est aussi très sensible aux flux de neutrons. En effet, une analyse de sensibilité de Y_{15} donne des résultats semblables à celles de Y_{14} (Table 3.9), avec une importance encore un peu plus forte du flux de neutrons.

	sections efficaces du dosimètre					flux	taux de réactions		
	Det1	Det2	Det3	Det4	Det5		(τ_{Det1},τ_{Det2})	(τ_{Det3},τ_{Det4})	τ_{Det5}
indice S_j	0.080	0.092	0.094	0.080	0.070	**0.823**	**0.156**	0.105	0.085
indice S_{T_j}	0.003	0.013	0.016	0.003	$\simeq 0$	**0.722**	**0.141**	0.028	0.009

TAB. 3.9.: Indices de sensibilité de premier ordre et totaux pour la sortie Y_{15} du code Stay'SL.

Analyse de sensibilité des taux de réactions ajustés de Det1, Det2, Det3, Det4, Det5
Les résultats sont présentés dans la Table 3.10.

Entrées / Sorties	sections efficaces du dosimètre					flux	taux de réactions		
	Det1	Det2	Det3	Det4	Det5		(τ_{Det1},τ_{Det2})	(τ_{Det3},τ_{Det4})	τ_{Det5}
taux ajusté	**0.945**	$\simeq 0$	$\simeq 0$	$\simeq 0$	$\simeq 0$	$\simeq 0$	$\simeq 0$	$\simeq 0$	$\simeq 0$
Det1	$\simeq$**1**	$\simeq 0$	0.005	0.004	0.004	0.038	0.004	$\simeq 0$	0.001
taux ajusté	0.056	**0.459**	0.067	0.014	0.008	**0.233**	**0.163**	0.036	0.008
Det2	0.052	**0.451**	0.065	0.011	0.004	**0.221**	**0.147**	0.033	0.003
taux ajusté	0.030	0.046	**0.721**	0.075	0.067	**0.147**	0.083	0.062	0.031
Det3	$\simeq 0$	0.011	**0.686**	0.041	0.031	**0.112**	0.047	0.030	0.015
taux ajusté	$\simeq 0$	$\simeq 0$	0.022	**0.886**	0.004	0.056	0.021	$\simeq 0$	$\simeq 0$
Det4	$\simeq 0$	$\simeq 0$	0.026	**0.889**	0.008	0.060	0.025	$\simeq 0$	0.002
taux ajusté	0.040	0.043	0.051	0.041	**0.616**	**0.367**	0.050	0.052	0.053
Det5	$\simeq 0$	0.002	0.010	$\simeq 0$	**0.581**	**0.330**	0.010	0.011	0.022

TAB. 3.10.: Indices de sensibilité de premier ordre (ligne supérieure) et totaux (ligne inférieure) pour les sortie Y_1 à Y_5 du code Stay'SL.

Logiquement, on constate que les variances des taux de réactions ajustés sont sensibles essentiellement aux sections efficaces des dosimètres correspondants : par exemple, la variance du taux de réactions ajusté de Det1 est sensible exclusivement aux sections efficaces du dosimètre Det1. De même, la variance du taux de réactions ajusté de Det4 est sensible en grande partie aux sections efficaces du dosimètre Det4.
La variance du taux de réactions ajusté de Det2, est quant à elle sensible pour moitié aux sections efficaces du dosimètre de Det2, et pour l'autre moitié au flux de neutrons et aux taux de réactions $(\tau_{Det1}, \tau_{Det2})$.

Quant à la variance du taux de réactions ajusté de Det3, elle est elle aussi sensible au flux de neutrons, mais minoritairement par rapport à sa sensibilité aux sections efficaces du dosimètre Det3.
Finalement, la variance du taux de réactions ajusté de Det5 est sensible, dans cet ordre d'importance, aux sections efficaces du dosimètre Det5 et au flux de neutrons.

3.2.2.4. Conclusion de l'étude

Le code de calcul Stay'SL est un code complexe utilisé dans une étape d'un processus lié à l'évaluation de la durée de vie de certaines structures irradiées. Diminuer l'incertitude sur les différentes variables de sortie peut contribuer à mieux estimer ces durées de vie, par des intervalles de confiance plus fin.
L'enjeu d'une analyse de sensibilité de ce code est donc primordial, et il n'est pas possible de prendre le risque de tirer de fausses conclusions en utilisant une analyse classique. Une analyse de sensibilité multi-dimensionnelle est nécessaire et utile puisqu'elle permet de tirer des conclusions justes sur le modèle en prenant en compte les corrélation entre variables d'entrée. Ainsi, nous avons mis en évidence, pour chacune des différentes variables de sortie étudiées, quels étaient les groupes de variables d'entrée qui contribuent le plus à leur variance.

Annexe A

Annexes de la partie I

A.1. Corrélation partielle dans un cadre gaussien

Soit le modèle défini par :

$$Y = f(X_1, ..., X_p).$$

Nous nous plaçons dans un formalisme gaussien, *i.e.* le vecteur M est gaussien.
Considérons le vecteur M tel que :

$$M = \begin{bmatrix} Y \\ X_i \\ X_1 \\ \vdots \\ X_{i-1} \\ X_{i+1} \\ \vdots \\ X_p \end{bmatrix}, \text{ que l'on divise en } M_1 = \begin{bmatrix} Y \\ X_i \end{bmatrix} \text{ et } M_2 = \begin{bmatrix} X_1 \\ \vdots \\ X_{i-1} \\ X_{i+1} \\ \vdots \\ X_p \end{bmatrix} = \mathbf{X}_{\sim i} \text{ de sorte que } M = \begin{bmatrix} M_1 \\ M_2 \end{bmatrix}.$$

La matrice de variance de M peut se partitionner en 4 blocs : $\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$, où Σ_{11} est la matrice de variance de M_1, Σ_{12} est la matrice de variance de M_1 et de M_2, *etc*.
La loi de M_1 conditionnellement à M_2 est normale de matrice de variance : $\Sigma_{11|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Posons :

$$\Sigma_{11|2} = \begin{bmatrix} \sigma_{11|2} & \sigma_{12|2} \\ \sigma_{21|2} & \sigma_{22|2} \end{bmatrix}.$$

Les termes $\sigma_{ij|2}$ extra diagonaux ($i \neq j$) de la matrice $\Sigma_{11|2}$ sont appelés les covariances partielles.
La covariance partielle de Y et de X_i sachant les autres variables d'entrée est :

$$\sigma_{12|2} = \text{Cov}(Y, X_i|\mathbf{X}_{\sim i}),$$

de laquelle on déduit la corrélation partielle de Y et de X_i :

$$\rho_{Y,X_i|X_{\sim i}} = \frac{\text{Cov}(Y, X_i|X_{\sim i})}{\sqrt{\text{V}(Y|X_{\sim i})\text{V}(X_i|X_{\sim i})}} = \frac{\sigma_{12|2}}{\sqrt{\sigma_{11|2}\sigma_{22|2}}}.$$

A.2. Décomposition de la variance

Cet annexe contient les différents calculs analytiques de décomposition de variance. Le dernier paragraphe contient les calculs analytiques des indices de sensibiltié d'un modèle théorique test à entrées corrélées.

A.2.1. Décomposition de la variance d'un modèle à entrées indépendantes

Soit le modèle $Y = f(X_1, ..., X_p)$ que l'on écrit sous la forme de sa décpmposition de Sobol :

$$Y = f_0 + \sum_{i=1}^{p} f_i(X_i) + \sum_{1 \leq i < j \leq p} f_{i,j}(X_i, X_j) + \ldots + f_{1,\ldots,p}(X_1, ..., X_p),$$

où toutes les variables d'entrée sont indépendantes, et où :

$$\begin{aligned}
f_0 &= \mathrm{E}[Y], \\
f_i(X_i) &= \mathrm{E}[Y|X_i] - \mathrm{E}[Y], \\
f_{i,j}(X_i, X_j) &= \mathrm{E}[Y|X_i, X_j] - \mathrm{E}[Y|X_i] - \mathrm{E}[Y|X_j] + \mathrm{E}[Y], \\
f_{i,j,k}(X_i, X_j, X_k) &= \mathrm{E}[Y|X_i, X_j, X_k] - \mathrm{E}[Y|X_i, X_j] - \mathrm{E}[Y|X_i, X_k] - \mathrm{E}[Y|X_j, X_k] \\
&\quad + \mathrm{E}[Y|X_i] + \mathrm{E}[Y|X_j] + \mathrm{E}[Y|X_k] - \mathrm{E}[Y], \\
&\ldots
\end{aligned}$$

D'après le théorème 1.1.1, la variance de Y peut se décomposer sous la forme suivante :

$$V = \sum_{i=1}^{p} V_i + \sum_{1 \leq i < j \leq p} V_{ij} + \ldots + V_{1\ldots p},$$

avec :

$$\begin{aligned}
V_i &= \mathrm{V}(\mathrm{E}[Y|X_i]), \\
V_{ij} &= \mathrm{V}(\mathrm{E}[Y|X_i, X_j]) - V_i - V_j, \\
V_{ijk} &= \mathrm{V}(\mathrm{E}[Y|X_i, X_j, X_k]) - V_{ij} - V_{ik} - V_{jk} - V_i - V_j - V_k, \\
&\ldots
\end{aligned}$$

Nous montrons ici que les parts de variance de cette décomposition sont les variances des fonctions de la décomposition de Sobol (1.9), que l'on note $\tilde{V}_{i_1,\ldots,i_s}$:

$$V_{i_1,\ldots,i_s} = \mathrm{V}(f_{i_1,\ldots,i_s}(X_{i_1}, ..., X_{i_s})) = \tilde{V}_{i_1,\ldots,i_s} \qquad \forall \{i_1, ..., i_s\} \subseteq \{1, ..., p\}.$$

Démonstration. La variance de Y s'écrit :

$$\begin{aligned}
V &= \mathrm{V}\left(f_0 + \sum_{i=1}^{p} f_i(X_i) + \sum_{1 \leq i < j \leq p} f_{ij}(X_i, X_j) + \ldots + f_{1,\ldots,p}(X_1, ..., X_p) \right) \\
&= \sum_{i=1}^{p} \underbrace{\mathrm{V}(f_i(X_i))}_{\tilde{V}_i} + \sum_{1 \leq i < j \leq p} \underbrace{\mathrm{V}(f_{ij}(X_i, X_j))}_{\tilde{V}_{ij}} + \ldots + \underbrace{\mathrm{V}(f_{1,\ldots,p}(X_1, ..., X_p))}_{\tilde{V}_{1\ldots p}},
\end{aligned}$$

puisque les covariances entre toutes ces fonctions sont nulles. En effet, la covariance de deux fonctions de variables différentes est nulle :

$$\mathrm{Cov}(f_i(X_i), f_j(X_j)) = 0,$$

puisque les variables sont indépendantes. Celle de deux fonctions ayant des variables communes est aussi nulle :

$$\mathrm{Cov}(f_i(X_i), f_{ij}(X_i, X_j)) = \mathrm{E}[f_i(X_i) f_{ij}(X_i, X_j)] - \mathrm{E}[f_i(X_i)]\mathrm{E}[f_{ij}(X_i, X_j)] = 0,$$

puisque $\mathrm{E}[f_i(X_i) f_{ij}(X_i, X_j)]$ est nulle d'après la propriété de la décomposition de Sobol (1.8), et $\mathrm{E}[f_i(X_i)] =$

$\mathrm{E}[f_{ij}(X_i, X_j)] = 0$ d'après la propriété (1.7).
Les parts de variances de premier ordre sont :

$$\tilde{V}_i \quad = \quad \mathrm{V}(f_i(X_i)) = \mathrm{V}(\mathrm{E}[Y|X_i] - \mathrm{E}[Y]) = \mathrm{V}(\mathrm{E}[Y|X_i]) = V_i,$$

puisque $\mathrm{E}[Y]$ est une constante.
Les développements suivants nécessitent quatres propriétés, qui seront démontrées à la fin de ce paragraphe.

Propriété A.2.1. *Cov*$(E[Y|X_i, X_j], E[Y|X_i]) =$ *V*$(E[Y|X_i])$.

Propriété A.2.2. *Cov*$(E[Y|X_i, X_j, X_k], E[Y|X_i]) =$ *V*$(E[Y|X_i])$.

Propriété A.2.3. *Cov*$(E[Y|X_i, X_j, X_k], E[Y|X_i, X_j]) =$ *V*$(E[Y|X_i, X_j])$.

Propriété A.2.4. *Cov*$(E[Y|X_i, X_j], E[Y|X_i, X_k]) =$ *V*$(E[Y|X_i])$.

Les parts de variances de deuxième ordre sont :

$$\begin{aligned}
\tilde{V}_{ij} \quad &= \quad \mathrm{V}(f_{i,j}(X_i, X_j)) \\
&= \quad \mathrm{V}(\mathrm{E}[Y|X_i, X_j] - \mathrm{E}[Y|X_i] - \mathrm{E}[Y|X_j] + \mathrm{E}[Y]) \\
&\overset{X_i \perp X_j}{=} \quad \mathrm{V}(\mathrm{E}[Y|X_i, X_j]) + \mathrm{V}(\mathrm{E}[Y|X_i]) + \mathrm{V}(\mathrm{E}[Y|X_j]) - 2\mathrm{Cov}(\mathrm{E}[Y|X_i, X_j], \mathrm{E}[Y|X_i]) \\
&\quad\quad -2\mathrm{Cov}(\mathrm{E}[Y|X_i, X_j], \mathrm{E}[Y|X_j]) - 2\underbrace{\mathrm{Cov}(\mathrm{E}[Y|X_i], \mathrm{E}[Y|X_j])}_{=0 \text{ car } X_i \perp X_j} \\
&\overset{\text{Ppté A.2.1}}{=} \quad \mathrm{V}(\mathrm{E}[Y|X_i, X_j]) - \mathrm{V}(\mathrm{E}[Y|X_i]) - \mathrm{V}(\mathrm{E}[Y|X_j]) \\
&= \quad \mathrm{V}(\mathrm{E}[Y|X_i, X_j]) - V_i - V_j \\
&= \quad V_{ij}.
\end{aligned}$$

Les parts de variances de troisième ordre sont :

$$\begin{aligned}
\tilde{V}_{ijk} \quad &= \quad \mathrm{V}(f_{i,j,k}(X_i, X_j, X_k)) \\
&= \quad \mathrm{V}(\mathrm{E}[Y|X_i, X_j, X_k] - \mathrm{E}[Y|X_i, X_j] - \mathrm{E}[Y|X_i, X_k] - \mathrm{E}[Y|X_j, X_k] \\
&\quad\quad +\mathrm{E}[Y|X_i] + \mathrm{E}[Y|X_j] + \mathrm{E}[Y|X_k] - \mathrm{E}[Y]).
\end{aligned}$$

Afin de simplifier les calculs, introduisons les notations suivantes :

$$\begin{aligned}
E_i \quad &= \quad \mathrm{E}[Y|X_i] \\
E_{ij} \quad &= \quad \mathrm{E}[Y|X_i, X_j] \\
E_{ijk} \quad &= \quad \mathrm{E}[Y|X_i, X_j, X_k] \\
&\quad\quad ... \\
D_i \quad &= \quad \mathrm{V}(\mathrm{E}[Y|X_i]) \\
D_{ij} \quad &= \quad \mathrm{V}(\mathrm{E}[Y|X_i, X_j]) \\
D_{ijk} \quad &= \quad \mathrm{V}(\mathrm{E}[Y|X_i, X_j, X_k]) \\
&\quad\quad ...
\end{aligned}$$

Nous avons donc :

$$\begin{aligned}
\tilde{V}_{ijk} &= \mathrm{V}(E_{ijk} - E_{ij} - E_{ik} - E_{jk} + E_i + E_j + E_k) \\
&= D_{ijk} + D_{ij} + D_{ik} + D_{jk} + D_i + D_j + D_k \\
&\quad -2\underbrace{\mathrm{Cov}(E_{ijk}, E_{ij})}_{D_{ij}} -2\underbrace{\mathrm{Cov}(E_{ijk}, E_{ik})}_{D_{ik}} -2\underbrace{\mathrm{Cov}(E_{ijk}, E_{jk})}_{D_{jk}} \\
&\quad +2\underbrace{\mathrm{Cov}(E_{ijk}, E_i)}_{D_i} +2\underbrace{\mathrm{Cov}(E_{ijk}, E_j)}_{D_j} +2\underbrace{\mathrm{Cov}(E_{ijk}, E_k)}_{D_k} \\
&\quad +2\underbrace{\mathrm{Cov}(E_{ij}, E_{ik})}_{D_i} +2\underbrace{\mathrm{Cov}(E_{ij}, E_{jk})}_{D_j} -2\underbrace{\mathrm{Cov}(E_{ij}, E_i)}_{D_i} -2\underbrace{\mathrm{Cov}(E_{ij}, E_j)}_{D_j} -2\underbrace{\mathrm{Cov}(E_{ij}, E_k)}_{0} \\
&\quad +2\underbrace{\mathrm{Cov}(E_{ik}, E_{jk})}_{D_k} -2\underbrace{\mathrm{Cov}(E_{ik}, E_i)}_{D_i} -2\underbrace{\mathrm{Cov}(E_{ik}, E_j)}_{0} -2\underbrace{\mathrm{Cov}(E_{ik}, E_k)}_{D_k} \\
&\quad -2\underbrace{\mathrm{Cov}(E_{jk}, E_i)}_{0} -2\underbrace{\mathrm{Cov}(E_{jk}, E_j)}_{D_j} -2\underbrace{\mathrm{Cov}(E_{jk}, E_k)}_{D_k} \\
&\quad +2\underbrace{\mathrm{Cov}(E_i, E_j)}_{0} +2\underbrace{\mathrm{Cov}(E_i, E_k)}_{0} +2\underbrace{\mathrm{Cov}(E_j, E_k)}_{0}
\end{aligned}$$

d'où :

$$\tilde{V}_{ijk} = D_{ijk} - D_{ij} - D_{ik} - D_{jk} + D_i + D_j + D_k$$

En écrivant cette équation avec les termes V_i, V_{ij}, ..., on obtient finalement après quelques calculs :

$$\tilde{V}_{ijk} = \mathrm{V}(\mathrm{E}[Y|X_i, X_j, X_k]) - V_{ij} - V_{ik} - V_{jk} - V_i - V_j - V_k = V_{ijk}$$

□

A.2.1.1. Démonstrations des propriétés A.2.1 à A.2.4

Démonstration de la propriété A.2.1.

$$\begin{aligned}
\mathrm{Cov}(\mathrm{E}[Y|X_i, X_j], \mathrm{E}[Y|X_i]) &= \mathrm{E}[\mathrm{E}[Y|X_i, X_j]\mathrm{E}[Y|X_i]] - \underbrace{\mathrm{E}[\mathrm{E}[Y|X_i, X_j]]}_{\mathrm{E}[Y]}\underbrace{\mathrm{E}[\mathrm{E}[Y|X_i]]}_{\mathrm{E}[Y]} \\
&= \int\int \mathrm{E}[Y|X_i, X_j]\mathrm{E}[Y|X_i]dx_i dx_j - \mathrm{E}[Y]^2 \\
&= \int \mathrm{E}[Y|X_i] \underbrace{\left(\int \mathrm{E}[Y|X_i, X_j]dx_j\right)}_{\mathrm{E}[Y|X_i]} dx_i - \mathrm{E}[Y]^2 \\
&= \int \mathrm{E}[Y|X_i]^2 dx_i - \mathrm{E}[Y]^2 \\
&= \mathrm{E}[\mathrm{E}[Y|X_i]^2] - \mathrm{E}[Y]^2,
\end{aligned}$$

d'où

$$\begin{aligned}
\mathrm{Cov}(\mathrm{E}[Y|X_i, X_j], \mathrm{E}[Y|X_i]) &= \mathrm{E}[\mathrm{E}[Y|X_i]^2] - \mathrm{E}[\mathrm{E}[Y|X_i]]^2 \\
&= \mathrm{V}(\mathrm{E}[Y|X_i]).
\end{aligned}$$

□

Démonstration de la propriété A.2.2.

$$\begin{aligned}
\text{Cov}(\text{E}[Y|X_i,X_j,X_k],\text{E}[Y|X_i]) &= \text{E}[\text{E}[Y|X_i,X_j,X_k]\text{E}[Y|X_i]] - \text{E}[\text{E}[Y|X_i,X_j,X_k]]\text{E}[\text{E}[Y|X_i]] \\
&= \int_i\int_j\int_k \text{E}[Y|X_i,X_j,X_k]\text{E}[Y|X_i]dx_i dx_j dx_k - \text{E}[Y]^2 \\
&= \int_i \text{E}[Y|X_i] \underbrace{\left(\int_j\int_k \text{E}[Y|X_i,X_j,X_k]dx_j dx_k\right)}_{\text{E}[Y|X_i]} dx_i - \text{E}[Y]^2 \\
&= \text{E}[\text{E}[Y|X_i]^2] - \text{E}[Y]^2 \\
&= \text{V}(\text{E}[Y|X_i]).
\end{aligned}$$

□

Démonstration de la propriété A.2.3.

$$\begin{aligned}
\text{Cov}(\text{E}[Y|X_i,X_j,X_k],\text{E}[Y|X_i,X_j]) &= \text{E}[\text{E}[Y|X_i,X_j,X_k]\text{E}[Y|X_i,X_j]] - \text{E}[\text{E}[Y|X_i,X_j,X_k]]\text{E}[\text{E}[Y|X_i,X_j]] \\
&= \int_i\int_j\int_k \text{E}[Y|X_i,X_j,X_k]\text{E}[Y|X_i,X_j]dx_i dx_j dx_k - \text{E}[Y]^2 \\
&= \int_i\int_j \text{E}[Y|X_i,X_j] \underbrace{\left(\int_k \text{E}[Y|X_i,X_j,X_k]dx_k\right)}_{\text{E}[Y|X_i,X_j]} dx_i dx_j - \text{E}[Y]^2 \\
&= \text{E}[\text{E}[Y|X_i,X_j]^2] - \text{E}[Y]^2 \\
&= \text{V}(\text{E}[Y|X_i,X_j]).
\end{aligned}$$

□

Démonstration de la propriété A.2.4.

$$\begin{aligned}
\text{Cov}(\text{E}[Y|X_i,X_j],\text{E}[Y|X_i,X_k]) &= \text{E}[\text{E}[Y|X_i,X_j]\text{E}[Y|X_i,X_k]] - \text{E}[\text{E}[Y|X_i,X_j]]\text{E}[\text{E}[Y|X_i,X_k]] \\
&= \int_i\int_j\int_k \text{E}[Y|X_i,X_j]\text{E}[Y|X_i,X_k]dx_i dx_j dx_k - \text{E}[Y]^2 \\
&= \int_i \underbrace{\left(\int_j \text{E}[Y|X_i,X_j]dx_j\right)}_{\text{E}[Y|X_i]} \underbrace{\left(\int_k \text{E}[Y|X_i,X_k]dx_k\right)}_{\text{E}[Y|X_i]} dx_i - \text{E}[Y]^2 \\
&= \text{E}[\text{E}[Y|X_i]^2] - \text{E}[Y]^2 \\
&= \text{V}(\text{E}[Y|X_i]).
\end{aligned}$$

□

A.2.2. Décomposition de la variance d'un modèle à trois variables d'entrée dont deux sont corrélées

Considérons un modèle à trois variables d'entrée :

$$Y = f(X_1, X_2, X_3),$$

où les deux variables X_1 et X_2 sont non indépendantes. La décomposition de Sobol de Y est :

$$\begin{aligned} Y = f(X_1, X_2) &= f_0 + f_1(X_1) + f_2(X_2) + f_3(X_3) \\ &\quad + f_{1,2}(X_1, X_2) + f_{1,3}(X_1, X_3) + f_{2,3}(X_2, X_3) \\ &\quad + f_{1,2,3}(X_1, X_2, X_3). \end{aligned}$$

La variance de Y s'écrit, en notant f_i à la place de $f_i(X_i)$ pour alléger la notation :

$$\begin{aligned} \mathrm{V}(Y) &= \mathrm{E}[Y^2] - \mathrm{E}[Y]^2 = \mathrm{E}[Y^2] - f_0^2 \\ &= \mathrm{E}[f_1^2] + \mathrm{E}[f_2^2] + \mathrm{E}[f_3^2] + \mathrm{E}[f_{1,2}^2] + \mathrm{E}[f_{1,3}^2] + \mathrm{E}[f_{2,3}^2] + \mathrm{E}[f_{1,2,3}^2] \\ &\quad + 2\mathrm{E}[f_1 f_2] + 2\mathrm{E}[f_1 f_3] + 2\mathrm{E}[f_1 f_{1,2}] + 2\mathrm{E}[f_1 f_{1,3}] + 2\mathrm{E}[f_1 f_{2,3}] + 2\mathrm{E}[f_1 f_{1,2,3}] \\ &\quad + 2\mathrm{E}[f_2 f_3] + 2\mathrm{E}[f_2 f_{1,2}] + 2\mathrm{E}[f_2 f_{1,3}] + 2\mathrm{E}[f_2 f_{2,3}] + 2\mathrm{E}[f_2 f_{1,2,3}] \\ &\quad + 2\mathrm{E}[f_3 f_{1,2}] + 2\mathrm{E}[f_3 f_{1,3}] + 2\mathrm{E}[f_3 f_{2,3}] + 2\mathrm{E}[f_3 f_{1,2,3}] \\ &\quad + 2\mathrm{E}[f_{1,2} f_{1,3}] + 2\mathrm{E}[f_{1,2} f_{2,3}] + 2\mathrm{E}[f_{1,2} f_{1,2,3}] \\ &\quad + 2\mathrm{E}[f_{1,3} f_{2,3}] + 2\mathrm{E}[f_{1,3} f_{1,2,3}] \\ &\quad + 2\mathrm{E}[f_{2,3} f_{1,2,3}]. \end{aligned}$$

Or, en notant E_i pour l'espérance par rapport à la loi de la variable X_i, on a :

$$\begin{aligned} \mathrm{E}[f_1 f_3] &= \mathrm{E}_{12}[f_1 \underbrace{\mathrm{E}_3[f_3]}_{=0}] = 0, \qquad \mathrm{E}[f_2 f_3] = \mathrm{E}_{12}[f_2 \underbrace{\mathrm{E}_3[f_3]}_{=0}] = 0, \\ \mathrm{E}[f_1 f_{1,3}] &= \mathrm{E}_{12}[f_1 \underbrace{\mathrm{E}_3[f_{1,3}]}_{=0}] = 0, \qquad \mathrm{E}[f_1 f_{2,3}] = \mathrm{E}_{12}[f_1 \underbrace{\mathrm{E}_3[f_{2,3}]}_{=0}] = 0, \\ \mathrm{E}[f_2 f_{1,3}] &= \mathrm{E}_{12}[f_2 \underbrace{\mathrm{E}_3[f_{1,3}]}_{=0}] = 0, \qquad \mathrm{E}[f_2 f_{2,3}] = \mathrm{E}_{12}[f_2 \underbrace{\mathrm{E}_3[f_{2,3}]}_{=0}] = 0, \\ \mathrm{E}[f_3 f_{1,2}] &= \mathrm{E}_3[f_3 \underbrace{\mathrm{E}_{12}[f_{1,2}]}_{=0}] = 0, \qquad \mathrm{E}[f_3 f_{1,3}] = \mathrm{E}_3[f_3 \underbrace{\mathrm{E}_{12}[f_{1,3}]}_{=0}] = 0, \\ \mathrm{E}[f_3 f_{2,3}] &= \mathrm{E}_3[f_3 \underbrace{\mathrm{E}_{12}[f_{2,3}]}_{=0}] = 0, \qquad \mathrm{E}[f_{1,2} f_{1,3}] = \mathrm{E}_{12}[f_{1,2} \underbrace{\mathrm{E}_3[f_{1,3}]}_{=0}] = 0, \\ \mathrm{E}[f_{1,2} f_{2,3}] &= \mathrm{E}_{12}[f_{1,2} \underbrace{\mathrm{E}_3[f_{2,3}]}_{=0}] = 0, \qquad \mathrm{E}[f_{1,2} f_{1,2,3}] = \mathrm{E}_{12}[f_{1,2} \underbrace{\mathrm{E}_3[f_{1,2,3}]}_{=0}] = 0. \end{aligned}$$

On obtient donc finalement la décomposition suivante de la variance :

$$\begin{aligned} \mathrm{V}(Y) &= \mathrm{E}[f_1^2] + \mathrm{E}[f_2^2] + \mathrm{E}[f_3^2] + \mathrm{E}[f_{1,2}^2] + \mathrm{E}[f_{1,3}^2] + \mathrm{E}[f_{2,3}^2] + \mathrm{E}[f_{1,2,3}^2] \\ &\quad + 2\mathrm{E}[f_1 f_2] + 2\mathrm{E}[f_1 f_{1,2}] + 2\mathrm{E}[f_2 f_{1,2}] + 2\mathrm{E}[f_{1,3} f_{2,3}] \\ &\quad + 2\mathrm{E}[f_{1,3} f_{1,2,3}] + 2\mathrm{E}[f_{2,3} f_{1,2,3}], \end{aligned}$$

d'où

$$\mathrm{V}(Y) = V_1 + V_2 + V_3 + V_{12} + V_{13} + V_{23} + V_{123} + V_{12}^{*(3)}$$

avec $V_{12}^{*(3)} = 2\mathrm{E}[f_1 f_2] + 2\mathrm{E}[f_1 f_{1,2}] + 2\mathrm{E}[f_2 f_{1,2}] + 2\mathrm{E}[f_{1,3} f_{2,3}] + 2\mathrm{E}[f_{1,3} f_{1,2,3}] + 2\mathrm{E}[f_{2,3} f_{1,2,3}].$

A.2.3. Calcul des indices de sensibilité du modèle $Y = aX_1X_2 + bX_3X_4 + cX_5X_6$

Considérons le modèle mathématique :

$$Y = aX_1X_2 + bX_3X_4 + cX_5X_6, \tag{A.1}$$

où les coefficients a, b et c sont réels, et les variables d'entrée sont gaussiennes :

$$(X_1, X_2, X_3, X_4, X_5, X_6) \sim \mathcal{N}(\vec{0}, \Sigma), \quad \text{avec} \quad \Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \rho_1 & 0 & 0 \\ 0 & 0 & \rho_1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \rho_2 \\ 0 & 0 & 0 & 0 & \rho_2 & 1 \end{bmatrix}.$$

Pour calculer les indices de sensibilité, nous utilisons les quatre propriétés suivantes.

Propriété A.2.5. *Soient les variables* $(X_1, X_2) \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} {\sigma_1}^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & {\sigma_2}^2 \end{bmatrix}\right)$*. Alors* X_1 *suit conditionnellement à* X_2 *une loi gaussienne :*

$$X_1|X_2 \sim \mathcal{N}(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(X_2 - \mu_2), {\sigma_1}^2(1-\rho^2)).$$

Cette propriété est issue de [55], page 88.

Propriété A.2.6. *Soit* $X \sim \mathcal{N}(\mu, \sigma^2)$*. Alors, pour tout* $k \in \mathbb{Z}$*,*

$$\begin{aligned} E[(X-\mu)^{2k+1}] &= 0, \\ E[(X-\mu)^{2k}] &= \frac{(2k)!}{2^k k!}. \end{aligned}$$

Pour la démonstration se référer à [55], page 45.

Propriété A.2.7. *L'espérance d'un produit de deux variables aléatoires* X *et* Y *peut s'écrire :*

$$E[XY] = E[XE[Y|X]].$$

Démonstration. D'après le théorème de l'espérance totale, $\mathrm{E}[XY] = \mathrm{E}[\mathrm{E}[XY|X]]$, d'où $\mathrm{E}[XY] = \mathrm{E}[X\mathrm{E}[Y|X]]$. □

Propriété A.2.8. *Soient les variables* $(X_1, X_2) \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$*. On a alors :*

$$V(X_1X_2) = 1 + \rho^2.$$

Démonstration. La variance du produit des variables X_1 et X_2 s'écrit :

$$\mathrm{V}(X_1X_2) = \mathrm{E}[{X_1}^2{X_2}^2] - \mathrm{E}[X_1X_2]^2.$$

Or, $\mathrm{E}[X_1X_2] = \mathrm{Cov}(X_1, X_2) + \mathrm{E}[X_1]\mathrm{E}[X_2] = \rho + 0 = \rho$, et :

$$\begin{aligned} \mathrm{E}[{X_1}^2{X_2}^2] &= \mathrm{E}[{X_1}^2\mathrm{E}[{X_2}^2|X_1]] \qquad \text{(propriété A.2.7)} \\ &= \mathrm{E}[{X_1}^2((1-\rho^2) + \rho^2X_1^2)] \qquad \text{(propriété A.2.5)} \\ &= (1-\rho^2)\mathrm{E}[{X_1}^2] + \rho^2\mathrm{E}[{X_1}^4] \\ &= (1-\rho^2) + 3\rho^2 \qquad \text{(propriété A.2.6)} \\ &= 1 + 2\rho^2, \end{aligned}$$

d'où :

$$\mathrm{V}(X_1X_2) = 1 + 2\rho^2 - \rho^2 = 1 + \rho^2.$$

□

Étant donné la structure de corrélation du modèle (A.1), nous considérons les quatre variables multidimensionnelles suivantes : X_1, X_2, (X_3, X_4) et (X_5, X_6). Nous devons donc calculer 4 indices de sensibilité de premier ordre, 6 de second ordre, 4 indices d'ordre trois et 1 indice d'ordre quatre.
La variance de Y est :

$$\begin{array}{rcl} \mathrm{V}(Y) & = & \mathrm{V}(aX_1X_2 + bX_3X_4 + cX_5X_6) \\ & = & a^2\mathrm{V}(X_1X_2) + b^2\mathrm{V}(X_3X_4) + c^2V(X_5X_6) \\ & \overset{\text{(propriété A.2.8)}}{=} & a^2 + b^2(1+{\rho_1}^2) + c^2(1+{\rho_2}^2) \end{array}$$

Indices de premier ordre

Les quatre indices de premier ordre sont :

$$S_1 = \frac{\mathrm{V}(\mathrm{E}[Y|X_1])}{\mathrm{V}(Y)}, \quad S_2 = \frac{\mathrm{V}(\mathrm{E}[Y|X_2])}{\mathrm{V}(Y)}, \quad S_{\{3,4\}} = \frac{\mathrm{V}(\mathrm{E}[Y|X_3,X_4])}{\mathrm{V}(Y)}, \quad S_{\{5,6\}} = \frac{\mathrm{V}(\mathrm{E}[Y|X_5,X_6])}{\mathrm{V}(Y)}.$$

Or,

$$\begin{array}{rcl} \mathrm{V}(\mathrm{E}[Y|X_1]) & = & \mathrm{V}(aX_1\underbrace{\mathrm{E}[X_2]}_{=0} + b\mathrm{E}[X_3X_4] + c\mathrm{E}[X_5X_6]) = 0, \\ \mathrm{V}(\mathrm{E}[Y|X_3,X_4]) & = & \mathrm{V}(a\mathrm{E}[X_1X_2] + bX_3X_4 + c\mathrm{E}[X_5X_6]) \\ & = & b^2\mathrm{V}(X_3X_4) \\ & \overset{\text{(propriété A.2.8)}}{=} & b^2(1+{\rho_1}^2). \end{array}$$

Pour les mêmes raisons :

$$\begin{array}{rcl} \mathrm{V}(\mathrm{E}[Y|X_2]) & = & 0, \\ \mathrm{V}(\mathrm{E}[Y|X_5,X_6]) & = & c^2(1+{\rho_2}^2). \end{array}$$

Les indices de premier ordre sont donc :

$$\begin{array}{rcl} S_1 & = & S_2 = 0, \\ S_{\{3,4\}} & = & \dfrac{b^2(1+{\rho_1}^2)}{a^2 + b^2(1+{\rho_1}^2) + c^2(1+{\rho_2}^2)}, \\ S_{\{5,6\}} & = & \dfrac{c^2(1+{\rho_2}^2)}{a^2 + b^2(1+{\rho_1}^2) + c^2(1+{\rho_2}^2)}. \end{array}$$

Indices d'ordre deux

L'indice relatif à l'interaction entre X_1 et X_2 est :

$$S_{12} = \frac{\mathrm{V}(\mathrm{E}[Y|X_1,X_2]) - \mathrm{V}(\mathrm{E}[Y|X_1]) - \mathrm{V}(\mathrm{E}[Y|X_2])}{\mathrm{V}(Y)}.$$

Comme :

$$\mathrm{V}(\mathrm{E}[Y|X_1,X_2]) - \mathrm{V}(\mathrm{E}[Y|X_1]) - \mathrm{V}(\mathrm{E}[Y|X_2]) = a^2\mathrm{V}(X_1X_2) - 0 - 0 = a^2,$$

l'indice est égal à :

$$S_{12} = \frac{a^2}{a^2 + b^2(1+{\rho_1}^2) + c^2(1+{\rho_2}^2)}.$$

Par définition, toutes les parts de variance (dues aux variables seules et aux différentes intercations) sont des variances et donc sont positives, puisque nos quatre variables X_1, X_2, (X_3, X_4) et (X_5, X_6) sont indépendantes. Or comme la somme des parts de variance dues au couple (X_3, X_4), au couple (X_5, X_6) et à l'interaction entre X_1 et X_2 est égal à la variance totale de Y, ce qui peut se traduire par :

$$S_{\{3,4\}} + S_{\{5,6\}} + S_{12} = 1,$$

on en déduit que toutes les parts de variance dues aux autres interactions (d'ordre deux et plus) sont nulles, et donc que tous les autres indices de sensibilité sont nuls.

Ainsi, pour ce modèle seuls trois indices de sensibilité sont non nuls : $S_{\{3,4\}}$, $S_{\{5,6\}}$ et S_{12}.

A.3. Analyse de la variance fonctionnelle pour l'analyse de sensibilité

Ce travail est dû à Antoniadis [5]. La méthode qu'il propose consiste à écrire le modèle étudié dans une certaine base orthonormée. L'erreur faite en estimant les indices de sensibilité est alors contrôlable.
Nous définissons dans un premier temps une base orthonormée. Puis, en écrivant le modèle étudié dans cette base on extrait les indices de sensibilité, qui peuvent être estimé par régression. Enfin, nous montrons comment il est possible de contrôler les erreurs d'estimation et d'approximation faites lors de l'évaluation des indices de sensibilité.

Définition et construction d'une base orthonormée

Considérons un ensemble de fonctions $\phi_0, \phi_1, \phi_2, \ldots$ définies sur $[0,1]$, telle que :

$$\begin{aligned} \phi_0(x) &= 1 \quad \forall x \in [0,1], \\ \int_0^1 \phi_j(x)dx &= 0 \quad \forall j \geq 1, \\ \int_0^1 \phi_j(x)\phi_i(x)dx &= \delta_{i,j}. \end{aligned}$$

Ces fonctions peuvent être par exemple des polynômes orthogonaux, des fonctions sinusoïdales ou des ondelettes.
Considérons le cas p-dimensionnel, avec :

$$\mathbf{x} = (x_1, x_2, \ldots, x_p) \quad \in \quad [0,1]^p,$$

et définissons les fonctions :

$$\psi_{\mathbf{r}}(\mathbf{x}) = \prod_{j=1}^{p} \phi_{r(j)}(x_j). \tag{A.2}$$

où

$$\mathbf{r} = (r(1), r(2), \ldots, r(p)) \quad \in \quad \{0,1,2,\ldots\}^p.$$

Les fonctions $\psi_{\mathbf{r}}(\mathbf{x})$ ont alors les propriétés suivantes :

$$\begin{aligned} \psi_{0,0,\ldots,0}(\mathbf{x}) &= 1, \\ \int_{[0,1]^p} \psi_{\mathbf{r}}(\mathbf{x})\psi_{\mathbf{s}}(\mathbf{x})d\mathbf{x} &= \delta_{\mathbf{r},\mathbf{s}}, \end{aligned}$$

et définissent une base orthonormée de $L^2([0,1]^p)$.

Analyse de la variance fonctionnelle pour l'analyse de sensibilité.

Soit le modèle mathématique :

$$Y = f(\mathbf{X}),$$

où $\mathbf{X} \in [0,1]^p$.
Il s'écrit dans la base définie précédemment :

$$Y = f(\mathbf{X}) = \sum_{\mathbf{r} \in \mathcal{U}} \beta_{\mathbf{r}} \psi_{\mathbf{r}}(\mathbf{X}),$$

avec $\beta_r \in \mathbb{R}$ et où $\mathcal{U}$ est l'ensemble des arrangements de p éléments parmi $\{0,1,2,\ldots,p-1\}$. Le cardinal de $\mathcal{U}$ étant p^p, l'écriture du modèle dans cette base devient vite impossible lorsque la dimension augmente.
Nous définissons alors une approximation du modèle :

$$Y = f(\mathbf{X}) = \sum_{\mathbf{r} \in \mathcal{R}} \beta_{\mathbf{r}} \psi_{\mathbf{r}}(\mathbf{X}) + \eta(\mathbf{X}),$$

où $\mathcal{R} \subset \mathcal{U}$, et $\eta(\mathbf{X})$ est l'erreur d'approximation. Cette erreur d'approximation est déterministe puisqu'elle ne dépend que de l'ensemble $\mathcal{R}$ choisi. Il est donc possible de la connaître et de la contrôler.
Comme les $\psi_{\mathbf{r}}(\mathbf{x})$ forment une base orthonormée, les coefficients $\beta_{\mathbf{r}}$ sont égaux à :

$$\beta_{\mathbf{r}} = \int_{[0,1]^p} f(\mathbf{x})\psi_{\mathbf{r}}(\mathbf{x})d\mathbf{x}. \tag{A.3}$$

La variance de Y est donnée par :

$$\mathrm{V}(Y) = \sum_{\mathbf{r}\in\mathcal{R},\mathbf{r}\neq 0} \beta_{\mathbf{r}}^2 + \int_{[0,1]^p} \eta(\mathbf{X})^2 d\mathbf{x}.$$

Il est alors possible d'exprimer la part de variance de Y due à un sous ensemble $\mathcal{S}$ de variables d'entrée par :

$$\mathrm{V}_{\mathcal{S}} = \sum_{\mathbf{r}\in\mathcal{S}} \beta_{\mathbf{r}}^2,$$

d'où l'indicateur de sensibilité normalisé :

$$\frac{\mathrm{V}_{\mathcal{S}}}{\mathrm{V}(Y)}.$$

Cet indicateur est défini à partir des coefficients inconnus β_r qu'il faut donc estimer.

Estimation par régression et régression approchée.

En écrivant les fonctions de la base orthonormée :

$$Z(\mathbf{x}) = (\psi_0(\mathbf{x}), \ldots, \psi_{p-1}(\mathbf{x}))',$$

le vecteur β des coefficients (A.3) s'écrit :

$$\beta = \int Z(\mathbf{x}) f(\mathbf{x}) d\mathbf{x},$$

et peut être estimé par régression :

$$\hat{\beta} = (Z'Z)^{-1}Z'Y,$$

où Y est calculé à partir d'un N-échantillon $(\mathbf{X}_i)_{i=1..N}$ de variables aléatoires uniformes indépendantes et identiquement distribuées (i.i.d.) sur $[0,1]^p$.
Antoniadis propose aussi une approche par régression approchée numériquement plus économique :

$$\hat{\hat{\beta}} = \frac{1}{n}Z'Y.$$

Contrôle des erreurs d'approximation et d'estimation.

L'estimation des indices de sensibilité par cette méthode est sujette à deux sources d'erreur : l'erreur due à la troncature $\mathcal{R}$, et l'erreur due à l'estimation des coefficients β.
La troncature $\mathcal{R} \subset \mathcal{U}$ peut être vue comme une simplification de modèle, analogue à celle rencontrée lorsque l'analyse est réalisée sur une surface de réponse plutôt que sur la fonction elle-même. L'erreur commise peut alors être contrôlée puisque l'on connait théoriquement la troncature choisie.
L'erreur globale commise en estimant les indices est l'erreur faite en estimant l'approximation de f par :

$$\hat{f}(\mathbf{X})^{(N)} = \sum_{\mathbf{r}\in\mathcal{R}} \hat{\beta}_{\mathbf{r}}^{(N)} \psi_{\mathbf{r}}(\mathbf{X}),$$

où N est la taille d'échantillon utilisée pour la régression (approchée).
La qualité des résultats d'approximation peut être mesurée par :

$$\int (f(\mathbf{x}) - \hat{f}(\mathbf{x})^{(N)})^2 d\mathbf{x},$$

qui s'évalue en pratique sur les m derniers tirages d'échantillon par :

$$\frac{1}{m} \sum_{i=N-m+1}^{N} (f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)^{(i-1)})^2,$$

où le choix de m se fait par un compromis biais-variance donnant $m \sim N^{\frac{2}{3}}$.
En pratique, on utilise la mesure de manque d'adéquation *Lack Of Fit* :

$$LOF(N) = \frac{\int (f(\mathbf{x}) - \hat{f}(\mathbf{x})^{(N)})^2 d\mathbf{x}}{\mathrm{V}(Y)},$$

pour évaluer la qualité de l'estimation. Cette mesure est positive, et plus elle est petite plus l'estimation est bonne. Cet indicateur *LOF* renseigne sur la qualité des estimations des indices de sensibilité.

A.4. Une interface Matlab d'analyse de sensibilité

La section 1.1 présente les différents indices de sensibilité, ainsi que les différentes méthodes d'estimation de ces indices. Nous présentons ici une interface graphique, développée sous Matlab, permettant de calculer ces indices de sensibilité pour un modèle défini.
Cette interface a plusieurs objectifs :

- réaliser des analyses de sensibilité sur des cas d'école simples dont le code peut être saisi directement par l'utilisateur sous l'interface,
- réaliser des analyses de sensibilité de modèles plus complexes dont le code est contenu dans un fichier Matlab,
- mettre à disposition des utilisateurs une bibliothèque de fonctions Matlab des différentes méthodes d'analyse de sensibiltié. L'utilisateur pourra alors utiliser ces fonctions en les intégrant à ses programmes.

Les différentes méthodes d'analyse de sensibilité disponibles sont :

- le calcul des indices de sensibilité :
  - SRC,
  - PCC,
  - $SRRC$,
  - $PRCC$,
- le calcul des indices de sensibilité basés sur la décomposition de la variance (définition 1.1.4) par les méthodes :
  - de Sobol (indices à tout ordre),
  - FAST (indices de premier ordre et d'ordre total),
  - de McKay (indices de premier ordre),
  - des modèles GAM (indices de premier ordre).

A.4.1. Utiliser l'interface Matlab

A.4.1.1. Architecture du répertoire `/SAinterface/`

Le répertoire `/SAinterface/` contient les fichiers et dossiers suivants :

- `myfunction.m` qui est un exemple de fichier contenant le code d'un modèle,
- `Resultats/` qui contient les fichiers résultats, avec les extension :
  - `*.sobol` pour les indices de premier ordre estimés par la méthode de Sobol,
  - `*.sobolT` pour les indices totaux estimés par la méthode de Sobol,
  - `*.sobolTout` pour les indices à tout ordre estimés par la méthode de Sobol,
  - `*.fast` pour les indices de premier ordre estimés par FAST,
  - `*.fastT` pour les indices totaux estimés par FAST,
  - `*.mckay` pour les indices de premier ordre estimés par la méthode de McKay,
  - `*.gam` pour les indices de premier ordre estimés par GAM,
  - et enfin `*.src`, `*.pcc`,`*.srrc`,`*.prcc`, pour les indices du même nom,
- `SAinterface.fig` qui est le fichier graphique de l'interface,
- `SAinterface.m` qui contient le code de l'interface,
- `SAToolBox` qui contient les différentes fonctions Matlab relatives aux différentes méthodes d'estimation des indices de sensibilité :
  - `Fast.m`,
  - `Gam.m`,
  - `/GAMToolBox/` qui contient les différentes fonctions utilisées pour l'estimation d'un modèle GAM,
  - `Mckay.m`,
  - `Pcc.m`,
  - `Prcc.m`,
  - `simule.m` qui permet de simuler des variables de différentes lois de probabilité (uniforme, normale, exponentielle, Weibull, gamma, Student, beta, χ^2, et log-normale),
  - `Sobol.m`,
  - `Src.m`,
  - `Srrc.m`,
  - `unif2loi.m` qui transforme une variable de loi uniforme en une variable d'autre loi de probabilité (citées ci-dessus).

A.4.1.2. Chargement de l'Interface

Pour utiliser l'interface Matlab, il suffit de se placer dans le répertoire `/SAinterface/` et de lancer `SAinterface`. La fenêtre graphique représentée par la figure A.1 apparaît alors.

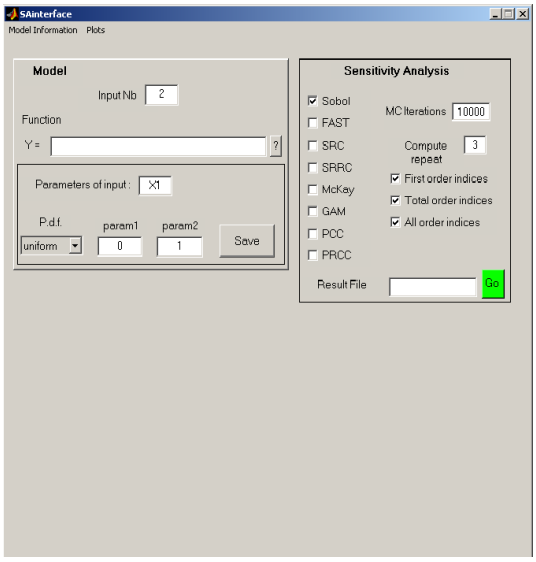


FIG. A.1. : SAinterface

Cette fenêtre est composée de deux parties : «Model», dans laquelle est paramétré le modèle et «Sensitivity Analysis» dans laquelle on choisit les indices à calculer et la méthode d'estimation. Nous détaillons ces deux étapes dans les paragraphes suivants.

A.4.1.3. Paramétrer le modèle

Après avoir saisi le nombre de variables d'entrée du modèle («Input Nb»), il faut :

- soit saisir spécifiquement le code de la fonction en langage Matlab, en nommant les variables d'entrée $X1$, $X2$, ..., sans oublier que ces dernières sont considérés comme des vecteurs :
  par exemple, pour 3 variables d'entrée : $X1 + \exp(X2) .* X3$,
- soit saisir le nom de fichier qui contient la fonction :
  par exemple : $myfunction(x)$.

Il faut ensuite spécifier dans l'espace «Parameter of input» la loi des variables d'entrée du modèle, celles-ci étant par défaut toutes uniformes sur l'intervalle $[0, 1]$. Il n'est pas possible à l'heure actuelle de spécifier des variables d'entrée corélées, ceci étant prévu pour une future version. La nature des lois des variables d'entrée peut être visualisée à l'aide du menu «Model Information/Input Distributions».

A.4.1.4. Choix de la méthode

Il faut dans un premier temps choisir la méthode à utiliser. Il est possible de spécifier plusieurs méthodes, l'estimation des indices sera alors faite indépendamment par chaque méthode.
Il faut ensuite préciser la taille des échantillons de Monte Carlo utilisée («MC iterations»), et le nombre de fois que l'on veut répéter l'estimation (pour obtenir un ordre d'idée de la précision des résultats). Pour la méthode FAST, le calcul ne sera toujours fait qu'une seule fois puisqu'il est déterminsite en fonction de la taille des échantillons choisie (cf. méthode FAST, section 1.1).
Enfin, il faut préciser le nom du fichier de sauvegarde.

Les calculs démarrent en actionnant le bouton «Go».

A.4.1.5. Résultats

Nous présentons les résultats obtenus pour l'exemple cité ci-dessus : $X1+\exp(X2)\times X3$ où les variables d'entrée sont uniformes sur $[0,1]$. La taille des échantillons de Monte Carlo est fixée à 10000, et nous demandons 30 répétitions du calcul des indices de sensibilité à tout ordre, par la méthode de Sobol.
Les résultats sont sauvegardés dans les fichiers de sauvegarde, et apparaîssent dans la fenêtre de commande Matlab. Comme nous avons demandé 30 calculs, les moyennes et écarts-types des 30 estimations sont présentées.
Il est possible de visualiser graphiquement ces résultats à l'aide de l'option «Plots/Draw Plots» du menu. L'interface graphique est alors représentée par la figure A.2.
Les graphiques présentent les indices de premier ordre et totaux du modèle, avec une barre d'erreur étendue à plus ou

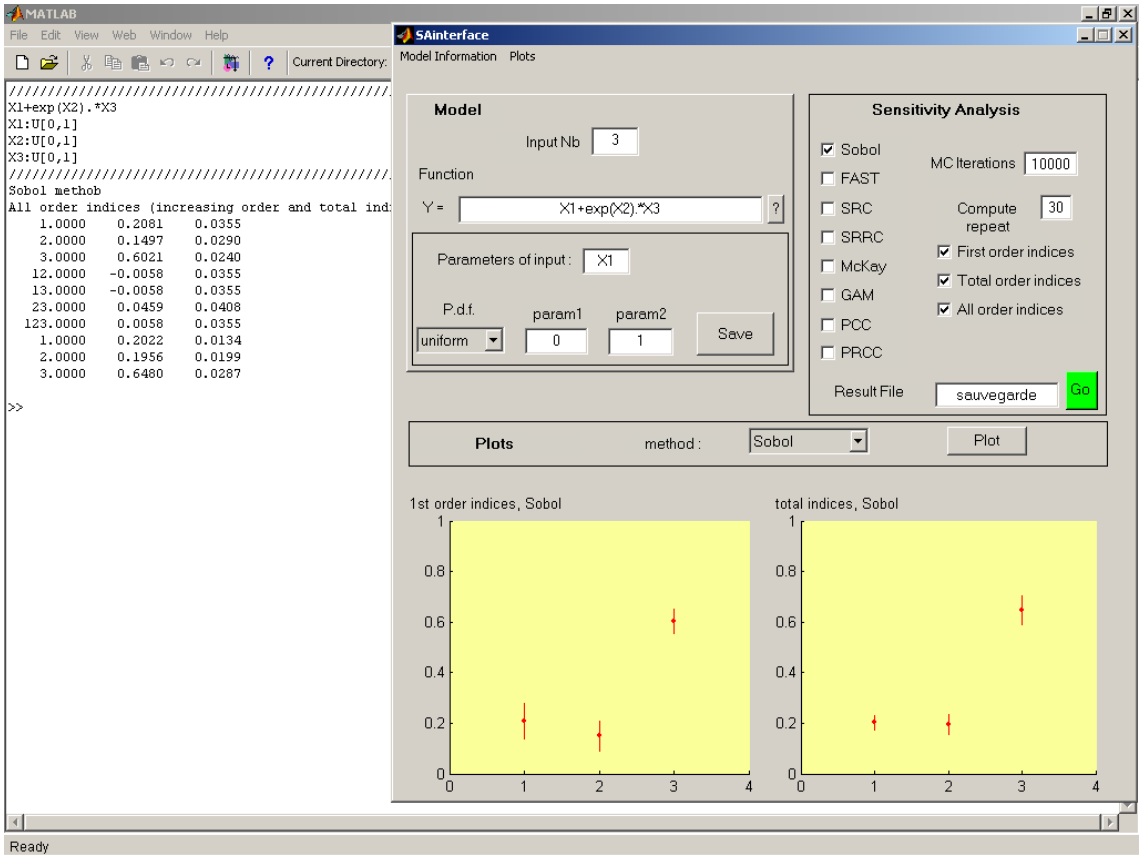


FIG. A.2. : SAinterface

moins deux écarts-types.

A.4.2. Les autres logiciels d'analyse de sensibilité

Peu de logiciels destinés à l'analyse de sensibilité existent. Nous n'en citerons qu'un, développé par l'équipe de Saltelli et Tarantola : SIMLAB. Ce logiciel, régulièrement amélioré à travers la sortie de nouvelles versions, est libre [58]. Il est beaucoup plus complet que l'interface que nous venons de présenter, et propose tout un panel de méthode d'analyse de sensibilité et d'incertitude.

A.4.3. Développements futurs

Cette interface Matlab sera prochainement enrichie du calcul des indices de sensibilité multidimensionnels (section 3.1) dans le cas de modèles à entrées non indépendantes.

Bibliographie

[1] L.R Abramson. Model uncertainty from a regulatory point of view. In *Workshop "Model Uncertainty : its Characterization and Quantification"*, Anapolis (Maryland, USA), 1993.

[2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19 :716–723, 1974.

[3] N. Akaike. Information theory as an extension of the maximum likelihood principle. In *B.Petrov and F. Csaki, editors, Second International Symposium on Information Theory*, pages 267–281, Budapest, Akademiai Kiado, 1973.

[4] A. Ameur (Ligeron S.A.). Code gascon - dossier de définition. CEA/DSNQ, 1999.

[5] A. Antoniadis. Analyse de la variance fonctionnelle pour l'analyse de sensibilité globale. In *Journées scientifiques du GDR Momas*, Luminy (France), 2003.

[6] G. Apostolakis. The concept of probability in safety assessments of technological systems. *Science*, 250 :1359–1364, 1990.

[7] Bedford. Sensitivity indices for (tree)-dependant variables. In *International Symposium on Sensitivity Analysis of Model Output*, Venise (Italie), 1998.

[8] C. Bielza, S. Rios-Insua, M. Gomez, and J.A. Fernandez del Pozo. Sensitivity analysis in ictneo. In *International Symposium on Sensitivity Analysis of Model Output*, Madrid (Espagne), 2001.

[9] M. Bier. Some illustrative examples of model uncertainty. In *Workshop "Model Uncertainty : its Characterization and Quantification"*, Anapolis (Maryland, USA), 1993.

[10] S.T. Buckland, K.P. Burnham, and N.H. Augustin. Model selection : an integral part of inference. *Biometrics*, 53 :603–618, 1997.

[11] A. Buslik. Bayesian approach to model uncertainty. In *Workshop "Model Uncertainty : its Characterization and Quantification"*, Anapolis (Maryland, USA), 1993.

[12] M. Clyde. Comment on "bayesian model averaging : a tutorial". *Statistical Science*, 14(4) :382–417, 1999.

[13] R.I. Cukier, C.M. Fortuin, K.E. Shuler, A.G. Petschek, and J.H. Schaibly. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients - theory. *Journal Chemical Physics*, 59 :3873–3878, 1973.

[14] R.I. Cukier, R.I. Levine, and K.E. Shuler. Nonlinear sensitivity analysis of multiparameter model systems. *Journal Computational Physics*, 26 :1–42, 1978.

[15] R.I. Cukier, K.E. Shuler, and J.H. Schaibly. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients - analysis of the approximations. *Journal Chemical Physics*, 63 :1140–1149, 1975.

[16] N. Devictor. *Fiabilité mécanique : méthodes FORM/SORM et couplages avec des codes éléments finis par des surfaces de réponse adaptatives*. PhD thesis, Université Blaise Pascal, 1996.

[17] N. Devictor, M. Marques, and M. Lemaire. Adaptative use of response surfaces in the reliability computations of mechanical components. *In : Advance in Safety and Reliability*, pages 1269–1277, 1997.

[18] O. Ditlevsen and T. Arnbjerg-Nielsen. Model-correction-factor method in structural reliability. *Journal of Engineering Mechanics*, 120(1), 1994.

[19] b. Efron and C. Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9(3) :586–596, 1981.

[20] H. Faure. Discrépance de suites associées à un système de numération (en dimension s). *Acta Arith.*, 41 :337–351, 1982.

[21] J.H. Friedman, T. Hastie, and R. Tibshirani, editors. *Element of statistical learning*. Springer, 2001.

[22] J. Goupy, editor. *La méthode des plans d'expériences*. Paris, Editions Dunod, 1996.

[23] J.H. Halton. On the efficiency of certain quasi-random sequences of point sequence. *Numerical Mathematics*, 2 :84–90, 1960.

[24] T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1990.

[25] J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volinsky. Bayesian model averaging : a tutorial. *Statistical Science*, 14(4) :382–417, 1999.

[26] T. Homma and A. Saltelli. Use of sobol's quasirandom sequence generator for integration of modified uncertainty importance measure. *Journal of Nuclear Science and Technology*, 32(11) :1164–1173, 1995.

[27] T. Homma and A. Saltelli. Importance measures in global sensitivity analysis of non linear models. *Reliability Engineering and System Safety*, 52 :1–17, 1996.

[28] R.L. Iman and W.J. Conover. A distribution-free approach to including rank correlation among input variables. *Communications in Statistics*, B11(3) :311–334, 1982.

[29] B. Iooss, F. Van Dorpe, and N. Devictor. Responses surfaces and sensitivity analyses for an environmental model of dose calculations. Reliability Engineering and System Safety, to appear.

[30] R.E. Kass and A.E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90 :773–795, 1995.

[31] B. Kraan and R. Cooke. The effect of correlations in uncertainty analysis : two cases. Technical Report In R. Cooke, ed. Technical Committee Uncertainty Modeling : Report on the Benchmark Workshop Uncertainty, European Safety and Reliability Association, Delft, Netherlands, 1997.

[32] K.B. Laskey. Applications of model uncertainty for the practice of risk assessment. In *Workshop "Model Uncertainty : its Characterization and Quantification"*, Anapolis (Maryland, USA), 1993.

[33] D. Madigan and A.E. Raftery. Model selection and accounting for model uncertainty in graphical model using occam's window. *Journal of the American Statistical Association*, 89 :1335–1346, 1994.

[34] M.D. McKay. Evaluating prediciton uncertainty. Technical Report NUREG/CR-6311, US Nuclear Regulatory Commission and Los Alamos National Laboratory, 1995.

[35] M.D. McKay, R. Beckman, and W. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2) :239–245, 1979.

[36] G.J. McLachlan and D. Peel, editors. *Finite Mixture Model*. New York : Wiley, 2000.

[37] H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. Philadelphia : SIAM, 1992.

[38] W.L. Oberkampf, S.M. DeLand, B.M. Rutherford, K.V. Diegert, and K.F. Alvin. Error uncertainty in modeling and simulation. *Reliability Engineering and System Safety*, 2002.

[39] A. Owen. *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, chapter Randomly Permuted (t,m,s)-Nets and (t,s)-Sequences. New York : Springer-Verlag, Niederreiter,H. and Shiue,P.J.-S. (Eds), 1995.

[40] A. Owen. Monte carlo extension of quasi-monte carlo. In *1998 Winter Simulation Conference*, Washington (DC, USA), 1998.

[41] M. Pendola. *Fiabilité des structures en contextes d'incertitudes statistique et de modélisation*. PhD thesis, Université Blaise Pascal, 2000.

[42] F.G. Perey. Least-squares dosimetry unfolding : the program stay'sl. Technical Report ORNL/TM-6062 ENDF-254, Oak Ridge National Laboratory, 1977.

[43] U. Pulkkinen and T. Huovinen. Model uncertainty in safety assessment. Technical Report STUK-YTO-TR 95, Finnish Center for Radiation and Nulcear Safety, 1996.

[44] A.E. Raftery. *Sociological Methodology 1995*, chapter Bayesian model selection in social research (with discussion by Andrew Gelman, Donald B. Rubin and Robert M. Hauser), pages 111–196. Oxford, U.K. : Blackwells, peter v. marsden edition, 1995.

[45] B.S. RamaRao, S. Mishra, S.D. Sevougian, and R.W. Andrews. Uncertainty importance of correlated variables in the probabilistic performance assessment of a nuclear waste repository. In *Second International Symposium on Sensitivity Analysis of Model Output*, pages 215–218, Venise (Italie), 1998.

[46] C. Robert and Casella G. *Monte Carlo Statistical Methods*. Springer, 1999.

[47] A. Saltelli and R. Bolado. An alternative way to compute fourier amplitude sensitivity test (fast). *Computational Statistics Data Analysis*, 26 :445–460, 1998.

[48] A. Saltelli, K. Chan, and E.M. Scott, editors. *Sensitivity Analysis*. Wiley, 2000.

[49] A. Saltelli and E.M. Scott. Guest editorial : The role of sensitivity analysis in the corroboration of models and its link to model structural and parametric uncertainty. *Reliability Engineering and System Safety*, 1997.

[50] A. Saltelli and S. Tarantola. Sensitivity analysis : a prerequisite in model building ? *Foresight and Precaution*, 2000.

[51] A. Saltelli and S.. Tarantola. On the relative importance of input factors in mathematical models : safety assessment for nuclear waste disposal. *Journal of the American Statistical Association*, 97(459) :702–709, 2002.

[52] A. Saltelli, S. Tarantola, and F. Campolongo. Sensitivity analysis as an ingredient of modeling. *Statistical Science*, 15(4) :377–395, 2000.

[53] A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto, editors. *Sensitivity Analysis in Practice*. Wiley, 2004.

[54] A. Saltelli, S. Tarantola, and K.P.-S. Chan. A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics*, 41(1) :39–56, 1999.

[55] G. Saporta, editor. *Probabilités, Analyse Des Données et Statistique*. Technip, 1990.

[56] J.H. Schaibly and K.E. Shuler. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficeints. ii applications. *Journal Chemical Physics*, 59 :3879–3888, 1973.

[57] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6 :461–464, 1978.

[58] SIMLAB. Uncertainty and sensitivity analysis software. *"http ://www.jrc.cec.eu.int/uasa/prj-sa-soft.asp"*, Joint Research Center European Commission.

[59] I.M. Sobol. On the distribution of points in a cube and the approximate evaluation of integrals. *U.S.S.R Comput. Math. and Math. Phys.*, 7 :86–112, 1967.

[60] I.M. Sobol. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, 1 :407–414, 1993.

[61] M. Stein. Large sample properties of simulations using latin hypercube sampling. *Technometrics*, 29(2), 1987.

[62] S. Tarantola. Quantifying uncertainty importance when inputs are correlated. In *Foresight and Precaution Conference (ESREL)*, Edinburgh (UK), 2000.

[63] T. Turanyi. Sensitivity analysis of complex kinetic system, tools and applications. *Journal of Mathematical Chemistry*, 5 :203–248, 1990.

[64] G. Wabba, editor. *Spline methods for observational data*. SIAM. Philadephia, 1990.

[65] R.W. Walters and L. Huyse. Uncertainty analysis for fluid mechanics with applications. Technical Report NASA/CR-2002-211449 ICASE Report No. 2002-1, NASA, 2002.

[66] L. Winkler. Model uncertainty : Probabilities for models ? In *Workshop "Model Uncertainty : its Characterization and Quantification"*, Anapolis (Maryland, USA), 1993.

[67] E. Zio and G. Apostolakis. Two methods for the structured assessment of model uncertainty by experts in performance assessments of radioactive waste repositories. *Reliability Engineering and System Safety*, 54 :225–241, 1996.

Seconde partie.
Analyse discriminante généralisée :
cas des données binaires
avec modèles des classes latentes

Chapitre 4

Analyse discriminante généralisée : cas des données binaires avec modèles des classes latentes

4.1. Introduction

4.1.1. L'analyse discriminante

Considérons un ensemble d'individus décrits par d variables explicatives $X^1, \ldots, X^d$, et une partition de ces individus en K groupes (ou classes) $G_1, \ldots, G_K$ définis *a priori*. L'analyse discriminante, aussi appelée classification supervisée ou « *scoring* », consiste à prédire l'appartenance d'un individu aux groupes *a priori* [13].
Nous disposons pour cela d'un ensemble d'apprentissage $\mathcal{D}$ constitué de N individus

$$\mathcal{D} = ((\mathbf{x}_1, \mathbf{z}_1), \ldots, (\mathbf{x}_N, \mathbf{z}_N)),$$

pour lesquels nous connaissons les valeurs des variables explicatives $\mathbf{x}_i = (x_i^1, \ldots, x_i^d) \in \mathcal{X}$, où $\mathcal{X}$ correspond à $\mathbb{R}^d$ dans le cas continu et à $\{0,1\}^d$ dans le cas binaire, ainsi que les appartenances $\mathbf{z}_i \in \{0,1\}^K$ des individus aux groupes $G_1, \ldots, G_K$, où $z_i^k = 1$ si l'individu i appartient au k-ème groupe et $z_i^k = 0$ sinon.

4.1.2. Les évolutions de l'analyse discriminante

L'analyse discriminante a été sujette à de nombreuses évolutions au cours du vingtième siècle. C'est en 1936 que Fisher [17] proposa une des premières techniques de discrimination linéaire entre deux classes, utilisant la distance de Mahalanobis. Ses travaux ont été par la suite étendus au cas de plus de deux classes par Rao en 1948 [27]. Plus récemment, l'utilisation des mélanges gaussiens multivariés a permis d'obtenir des règles de discrimination quadratiques (cf. par exemple [34]). Banfield et Raftery en 1993 [5] puis Celeux et Govaert en 1995 [12] ont alors étudié un certain nombre de modèles évoluant entre les règles de discrimination linéaire et quadratique.
Anderson proposa quant à lui en 1972 une méthode que l'on peut qualifier de semi-paramétrique : la discrimination logistique [3].
Des méthodes non paramétriques ont aussi été définies, parmi lesquelles celle due à Fix et Hodges en 1951 [18], basée sur la méthode des k plus proches voisins. Citons aussi les travaux de Friedman et Stuetzle en 1981 [19], basés sur la notion de directions révélatrices (*projection pursuit*), ainsi que ceux de Hand en 1982 [20] ou encore de Silverman en 1986 [31], basés sur la méthode des noyaux.

4.1.3. Une évolution transversale : la discrimination généralisée

Toutes les procédures classiques d'analyse discriminante citées précédemment font l'hypothèse que l'échantillon d'apprentissage est issu de la même population que l'échantillon à classer. Il arrive néanmoins en pratique que cette hypothèse ne soit pas vérifiée.
Van Franeker et Ter Brack [35] illustrent ce problème par une application biologique concrète consistant à déterminer le sexe d'oiseaux sur lesquels un certain nombre de variables biométriques ont été mesurées. L'échantillon d'apprentissage est constitué d'oiseaux originaires des Pays-Bas, pour lesquels le sexe est connu par dissection. L'échantillon test est quant à lui constitué d'oiseaux d'origine géographique différente, de même espèce mais de taille différente, pour lesquels le sexe n'est pas connu. En raison de la différence de taille entre les oiseaux de ces deux populations, il est évident que la règle de classification issue de l'échantillon d'apprentissage ne peut être appliquée directement sur l'échantillon test. Sous l'hypothèse de normalité des variables biométriques mesurées, Van Franeker et Ter Brack proposent alors une des premières techniques de discrimination généralisée, construite de façon relativement empirique.
S'appuyant sur une application en biologie semblable à celle décrite précédemment, Biernacki et al. [9] développent dans leur article de référence sur la discrimination généralisée un certain nombre de modèles dans un cadre gaussien. Nous présentons succinctement leur méthode dans le paragraphe suivant.

Remarque. *Un contexte similaire, dans lequel l'échantillon d'apprentissage et l'échantillon test ne sont pas issus d'une même population est abordé par Baccini et al. en 2001 [4]. En effet, une dérive des conditions expérimentales, entre le moment où l'échantillon d'apprentissage est mesuré et celui où l'échantillon à classer est mesuré rend inopérant une analyse discriminante classique. La méthode d'analyse discriminante conditionnelle qu'ils proposent consiste à mesurer, en parallèle des variables explicatives pour chaque échantillon, une ou plusieurs variables supplémentaires étant supposées corrélées avec cette dérive temporelle des conditions expérimentale. Cette ou ces variables sont alors prise en compte dans la définition de la règle de classement. Les résultats de classement obtenus sont bons, mais l'hypothèse de l'existence et de la mesurabilité de variables corrélées à la dérive temporelle est très forte et restreint donc les possibilités d'applications.*

4.1.3.1. Discrimination généralisée dans un cadre gaussien

Soit P la population d'apprentissage, pour laquelle l'appartenance des individus aux classes est connue, et soit P^* la population test dont sont issus les individus à classer. Lorsque la loi conditionnelle aux groupes est une gaussienne multivariée, ces deux populations sont modélisées par un mélange de lois normales. Au contraire de la discrimination classique, les paramètres de ces mélanges peuvent être différents pour les deux populations P et P^* en discrimination généralisée. Nous supposons néanmoins que l'espace des données est identique pour ces deux populations, c'est-à-dire que le nombre de classes K est le même, que chaque classe est de même nature, et que les variables mesurées sont également de même nature.
L'approche proposée par Biernacki et al. consiste alors à exhiber une transformation en loi liant les vecteurs aléatoires d'une même classe k mais de populations différentes. Sous des hypothèses raisonnables (transformation composante par composante et de classe $\mathcal{C}^1$), cette transformation est nécessairement affine et s'écrit :

$$Y^*_{|Z^{*k}=1} \sim A_k Y_{|Z^k=1} + b_k,$$

où A_k est une matrice diagonale de $\mathbb{R}^{d\times d}$, et b_k un vecteur de $\mathbb{R}^d$.
Sous certaines hypothèses restrictives sur la nature du lien entre les populations P et P^*, plusieurs modèles de liaisons ont alors été proposés. L'application de ces modèles au problème biologique des sexes d'oiseaux donne de très bons résultats, meilleurs que ceux obtenus par une approche de discrimination classique ou de classification automatique.

4.2. Analyse discriminante généralisée pour données binaires

Dans de nombreux domaines d'application, comme la médecine ou la finance (cf. paragraphe 4.3), les variables explicatives ne sont pas continues mais binaires. L'objectif de nos travaux est d'étendre l'analyse discriminante généralisée, définie par Biernacki et al. [9] dans un cadre gaussien, au cas de données binaires. Pour ceci, nous adoptons une modélisation des variables binaires connus sous le nom de modèle des classes latentes, qui définit la loi des variables binaires conditionnellement à l'appartenance à une classe (latente). Puis nous verrons que sous l'hypothèse que ces variables binaires sont issues d'une discrétisation de variables gaussiennes sous-jacentes, il est possible de définir un lien entre les variables binaires de la population d'apprentissage et celles de la population test. Nous présentons ensuite les techniques d'estimation nécessaires aux différentes étapes de la discrimination généralisée. Enfin, nous montrons à l'aide d'un ensemble de tests sur simulations la robustesse de la discrimination généralisée ainsi que son intérêt vis-à-vis de la discrimination classique et de la classification automatique.

4.2.1. Le modèle des classes latentes

Les modèles à variables latentes supposent l'existence de variables inobservables directement, telle l'intelligence, mais dont il est possible de mesurer les effets, comme les résultats à certains tests. La notion de structure latente a été introduite par Lazarsfeld et Henry [22] dans les années 1950, dans le cadre d'étude socio-psychologique. Depuis, de nombreux travaux ont permis de développer un certain nombre de techniques d'analyse de structure latente, que l'on peut regrouper en quatre catégories, selon que les variables latentes et observées sont discrètes ou continues (tableau 4.1).

		variables latentes	
		discrètes	continues
variables observées	discrètes	modèle des classes latentes	modèle des traits latents
	continues	modèles des profils latents	modèle d'analyse factorielle

TAB. 4.1.: Modèles d'analyse de structures latentes.

Pour plus de détail, le lecteur pourra se référer aux ouvrages d'Everitt [15] ou encore de Bartholomew et Knott [6].
Dans notre cas où toutes les variables sont discrètes, les variables observées sont les variables explicatives binaires X^j et les variables latentes sont les appartenances aux classes Z. Ainsi un individu possède un certain caractère j si $X^j = 1$ tandis qu'il ne le possède pas si $X^j = 0$. De même, cet individu appartient au groupe G_k si $Z^k = 1$ et n'appartient pas à ce groupe si $Z^k = 0$. Le modèle des classes latentes (cf. [11] et [15]) suppose que les variables $X^j_{|Z^k=1}$ sont conditionnellement indépendantes (indépendance des variables observées conditionnellement aux variables latentes) et qu'elles suivent une loi de Bernoulli :

$$X^j_{|Z^k=1} \quad \sim \quad \mathcal{B}(\alpha_{kj}) \quad \forall j = 1, \ldots, d, \quad \forall k = 1, \ldots, K.$$

En d'autres termes, la probabilité qu'un individu du groupe G_k possède le caractère j est égale à α_{kj}.
Ainsi, la distribution de probabilité de la variable $\mathbf{X} = (X^1, \ldots, X^d)$ peut s'écrire, conditionnellement à l'appartenance à la classe k :

$$f_k(x^1, \ldots, x^d) = \prod_{j=1}^{d} {\alpha_{kj}}^{x^j} (1 - \alpha_{kj})^{1-x^j}. \tag{4.1}$$

De façon similaire, la loi de $\mathbf{Z}$ est donnée par

$$\mathbf{Z} \sim \mathcal{M}(1, p_1, \ldots, p_K), \tag{4.2}$$

où $\mathcal{M}(1, p_1, \ldots, p_K)$ représente la loi multinomiale d'ordre 1 de paramètres $p_1, \ldots, p_K$, avec p_k représentant la probabilité *a priori* du groupe G_k dans la population P.

4.2.2. Les données

Les données consistent en deux échantillons : un étiqueté, S, issu d'une population P, et un non étiqueté, S^*, issu d'une population P^*. Les deux populations P et P^* peuvent être différentes.
L'échantillon d'apprentissage S est composé de N couples $(\mathbf{x}_1, \mathbf{z}_1), \ldots, (\mathbf{x}_N, \mathbf{z}_N)$, réalisations indépendantes du couple aléatoire $(\mathbf{X}, \mathbf{Z})$ de distribution :

$$X^j_{|Z^k=1} \sim \mathcal{B}(\alpha_{kj}) \quad \forall j = 1, \ldots, d \quad \text{et} \quad \mathbf{Z} \sim \mathcal{M}(1, p_1, \ldots, p_K). \tag{4.3}$$

De plus, l'hypothèse d'indépendance conditionnelle des variables explicatives X^j $(j = 1, \ldots, d)$ permet d'écrire la distribution de probabilité de la variable $\mathbf{X}$ conditionnellement à l'appartenance à la classe k selon l'équation (4.1). On en déduit l'expression suivante de la distribution de probabilité de $\mathbf{X}$

$$f(x^1, \ldots, x^d) = \sum_{k=1}^{K} p_k \prod_{j=1}^{d} {\alpha_{kj}}^{x^j} (1 - \alpha_{kj})^{1-x^j}. \tag{4.4}$$

L'échantillon test S^* est quant à lui composé de N^* individus pour lesquels seules les variables explicatives $\mathbf{x}^*_1, \ldots, \mathbf{x}^*_{N^*}$ sont connues (les variables sont les mêmes que pour l'échantillon d'apprentissage). Les appartenances aux classes $\mathbf{z}^*_1, \ldots, \mathbf{z}^*_{N^*}$ sont inconnues. Nous considérons alors les couples $(\mathbf{x}^*_i, \mathbf{z}^*_i)$ $(i = 1, \ldots, N^*)$ comme des réalisations indépendantes du couple aléatoire $(\mathbf{X}^*, \mathbf{Z}^*)$ de distribution :

$${X^*}^j_{|Z^{*k}=1} \sim \mathcal{B}(\alpha^*_{kj}) \quad \forall j = 1, \ldots, d \quad \text{et} \quad \mathbf{Z}^* \sim \mathcal{M}(1, p^*_1, \ldots, p^*_K). \tag{4.5}$$

La distribution du couple $(\mathbf{X}^*, \mathbf{Z}^*)$ diffère donc de celle de $(\mathbf{X}, \mathbf{Z})$ par la valeur des paramètres α_{kj} et p_k.
Nous supposons ici encore l'indépendance conditionnelle des variables explicatives ${X^*}^1, \ldots, {X^*}^d$, ce qui nous permet d'écrire la distribution de probabilité de $\mathbf{X}^*$ de façon analogue à celle de $\mathbf{X}$:

$$f^*(x^1, \ldots, x^d) = \sum_{k=1}^{K} p^*_k \prod_{j=1}^{d} {\alpha^*_{kj}}^{x^j} (1 - \alpha^*_{kj})^{1-x^j}. \tag{4.6}$$

L'objectif de la discrimination généralisée est alors d'estimer les N^* étiquettes $\mathbf{z}^*_1, \ldots, \mathbf{z}^*_N$ inconnues en utilisant l'information contenue à la fois dans S et dans S^*.

4.2.3. Modélisation du lien entre populations

Étant donné que les populations P et P^* peuvent être différentes, nous ne pouvons pas appliquer directement sur la population P^* une règle de classement apprise sur P. Nous cherchons donc à définir une relation entre les deux populations P et P^*, ou autrement dit une relation qui transforme les variables binaires de la population P en les variables binaires de la population P^*. Au contraire du cas gaussien où la transformation affine est non seulement justifiée mais aussi relativement intuitive, il n'est pas évident de définir directement une relation entre variables binaires.
Nous supposons alors que les variables binaires observées sont une discrétisation de variables continues gaussiennes sous-jacentes. Par exemple, lors de l'achat d'un produit par un client, nous supposons que ce client donne une note au produit en fonction de différents critères (qualité, esthétique, prix...), et qu'il décide de ne l'acheter que si cette note dépasse un certain seuil. Seule la décision (acte d'achat) est observable, la note ayant conduit à cette décision n'étant pas observée.
L'hypothèse qu'une variable binaire est une discrétisation d'une variable sous-jacente continue n'est pas

nouvelle dans le domaine des statistiques. Thurstone [33] est le premier à utiliser cette notion en 1927, dans son modèle de jugement comparatif, qui consiste à choisir entre deux stimuli celui dont la variable continue associée est la plus grande. Puis, au milieu du vingtième siècle apparaissent les modèles des traits latents, qui supposent que les variables binaires observées sont une observation discrète de variables continues latentes. Ces modèles ont été introduits précédemment dans la sous-section 4.2.1. Enfin, en 1987, Everitt [16] propose un algorithme de classification pour données mélangées (*i.e.* de plusieurs types : continues, ordinales et catégorielles), en s'appuyant sur l'hypothèse que les données ordinales et catégorielles sont une discrétisation observable de données continues inobservables.

La démarche que nous proposons consiste dans un premier temps à modéliser les variables binaires $\mathbf{X}$ et $\mathbf{X}^*$ par des variables continues gaussiennes sous-jacentes $\mathbf{Y}$ et $\mathbf{Y}^*$, puis, à l'image de la discrimination généralisée dans un cadre gaussien, à définir une relation entre les populations P et P^* en s'appuyant sur ces variables continues $\mathbf{Y}$ et $\mathbf{Y}^*$. Ensuite, à partir de la relation entre variables gaussiennes, nous montrons qu'il est possible de déduire une relation explicite entre les variables binaires.

4.2.3.1. Utilisation de variables continues sous-jacentes

Nous considérons donc les variables explicatives $X^j_{|Z^k=1}$, de loi de Bernoulli $\mathcal{B}(\alpha_{kj})$, comme issues d'une discrétisation de variables continues $Y^j_{|Z^k=1}$ de loi normale $\mathcal{N}(\mu_{kj}, \sigma^2_{kj})$.
Nous définissons cette discrétisation de la façon suivante :

$$X^j_{|Z^k=1} = \begin{cases} 0 & \text{si } \lambda_j Y^j_{|Z^k=1} < 0 \\ 1 & \text{si } \lambda_j Y^j_{|Z^k=1} \geq 0 \end{cases} \qquad \text{pour } j = 1, \ldots, d, \tag{4.7}$$

où $\lambda_j \in \{-1, 1\}$ est introduit afin de s'affranchir du choix d'un ordre au sein des variables explicatives, et ainsi de ne pas avoir à décider à quelle modalité de X^j faire correspondre un Y^j positif. Ce λ_j devient donc un paramètre du modèle, qui peut être connu s'il existe un ordre naturel au sein des variables X_j.

Remarque. *Le choix du découpage en 0 pour la discrétisation est arbitraire mais n'a pas d'incidence sur la méthode.*

On en déduit naturellement la relation entre α_{kj} et les paramètres λ_j, μ_{kj} et σ_{kj} :

$$\alpha_{kj} = p(X^j_{|Z^k=1} = 1) = p(\lambda_j Y^j_{|Z^k=1} \geq 0) = \begin{cases} \Phi\left(\frac{\mu_{kj}}{\sigma_{kj}}\right) & \text{si } \lambda_j = 1 \\ 1 - \Phi\left(\frac{\mu_{kj}}{\sigma_{kj}}\right) & \text{si } \lambda_j = -1 \end{cases} \tag{4.8}$$

où Φ désigne la fonction de répartition de la loi normale centrée réduite.
Nous supposons de plus que l'hypothèse d'indépendance conditionnelle des variables binaires faite précédemment peut être étendue aux variables continues Y^j. Cette hypothèse entraîne de grandes simplifications de calcul en permettant d'exprimer α_{kj} uniquement en fonction des paramètres μ_{kj}, σ_{kj} et λ_j (pour k et j fixés).

Remarque. *L'indépendance conditionnelle des variables binaires $X^j_{|Z^k=1}$ ne suffit pas à induire celle des variables continues sous-jacentes, comme le montre le contre-exemple suivant. Soient les deux variables aléatoires indépendantes Y_1 et T, et soit la variable $Y_2 = T\,Y_1$. Les variables binaires X_1 et X_2 engendrées par discrétisation de Y_1 et Y_2 sont clairement indépendantes, puisque la positivité ou la négativité de Y_1 et de Y_2 sont indépendantes (du fait de la variable T). Or, les variables continues Y_1 et Y_2 ne sont pas nécessairement indépendantes.*

Remarque. *Sans cette hypothèse d'indépendance conditionnelle des variables Y^j, on a si $\lambda_j = 1$:*

$$\begin{aligned}\alpha_{kj} &= p(Y^j_{|Z^k=1} \geq 0) \\ &= (2\pi)^{-\frac{d}{2}}|\Sigma_k|^{-\frac{1}{2}} \int_{\mathbb{R}\times\ldots\times\mathbb{R}\times[0,+\infty[\times\mathbb{R}\times\ldots\times\mathbb{R}} \exp(-\frac{1}{2}(y-\mu_k)'|\Sigma_k|^{-1}(y-\mu_k))dy.\end{aligned}$$

À supposer que l'on puisse identifier les structures de matrice de variance Σ_k correspondant à des variables binaires conditionnellement indépendantes, nous serions bloqués par les intégrations multidimensionnelles induites par le calcul des α_{kj} ci-dessus. Elles sont d'une complexité suffisamment importante pour pouvoir poser des problèmes d'évaluation en pratique.

Ainsi, sous cette hypothèse d'indépendance conditionnellement à l'appartenance à la classe k, les variables continues sont gaussiennes indépendantes :

$$\mathbf{Y}_{|Z^k=1} \sim \mathcal{N}_d(\mu_k, \Sigma_k),$$

avec Σ_k une matrice de variance diagonale, d'éléments diagonaux $(\sigma^2_{kj})_{1\leq j\leq d}$.
Si on note $\phi_k(.)$ la densité de probabilité d'une telle loi, la variable $\mathbf{Y}_{|Z^k=1}$ sous-jacente à la variable $\mathbf{X}_{|Z^k=1}$ est de densité de probabilité :

$$f(y) = \sum_{k=1}^{K} p_k \phi_k(y). \tag{4.9}$$

Enfin, connaissant y, la variable $\mathbf{Z}|y$ a pour distribution de probabilité :

$$p(Z^k = 1|y) = \frac{p_k\phi_k(y)}{f(y)} = t_k(y) \qquad \text{pour } k = 1, \ldots, K. \tag{4.10}$$

De la même façon, nous supposons que les variables binaires $\mathbf{X}^*_{|Z^{*k}=1}$ de la population P^* sont aussi issues de la discrétisation de variables sous-jacentes gaussiennes $\mathbf{Y}^*_{|Z^{*k}=1}$ de paramètres μ^*_k et Σ^*_k et de proportions $p^*_1, \ldots, p^*_K$.

4.2.3.2. Modèle de liaison entre les deux populations

L'objectif de l'analyse discriminante généralisée, qui consiste à établir un lien entre les populations P et P^*, revient donc à déterminer une relation entre les paramètres α_{kj} et α^*_{kj} des variables explicatives binaires. Dans le cas gaussien, nous avons rappelé que sous des hypothèses raisonnables (transformation composante par composante et de classe $\mathcal{C}^1$), la relation entre les variables des populations P et P^* est de la forme :

$$Y^*_{|Z^{*k}=1} \sim A_k Y_{|Z^k=1} + b_k, \tag{4.11}$$

où A_k est une matrice diagonale de $\mathbb{R}^{d\times d}$, et b_k un vecteur de $\mathbb{R}^d$.
On en déduit les relations suivantes entre les paramètres des variables gaussiennes $Y^*_{|Z^{*k}=1}$ et $Y_{|Z^k=1}$:

$$\begin{aligned}\mu^*_{kj} &= a_{kj}\mu_{kj} + b_{kj}, \\ \sigma^*_{kj} &= |a_{kj}|\,\sigma_{kj}.\end{aligned}$$

De plus, la relation (4.8) entre les paramètres des variables binaires et ceux des variables continues sous-jacentes au sein de la population P peut aussi s'écrire pour la population P^*. On a donc :

$$\alpha^*_{kj} = \begin{cases} \Phi\big(\frac{\mu^*_{kj}}{\sigma^*_{kj}}\big) & \text{si } \lambda_j = 1 \\ 1 - \Phi\big(\frac{\mu^*_{kj}}{\sigma^*_{kj}}\big) = \Phi\big(-\frac{\mu^*_{kj}}{\sigma^*_{kj}}\big) & \text{si } \lambda_j = -1. \end{cases}$$

Remarque. *Comme $\lambda^*_j = \lambda_j$, il est inutile d'introduire λ^*_j dans les équations.*

On en déduit finalement la relation suivante entre les paramètres α^*_{kj} et α_{kj} :

$$\alpha^*_{kj} = \Phi\Big(\underbrace{\text{sgn}(a_{kj})}_{\delta_{kj}}\ \Phi^{-1}(\alpha_{kj}) + \lambda_j \underbrace{\frac{b_{kj}}{|a_{kj}|\ \sigma_{kj}}}_{\gamma_{kj}}\Big), \tag{4.12}$$

où $\text{sgn}(t)$ désigne le signe du réel t.
Sous l'hypothèse que les variables binaires sont issues de variables gaussiennes sous-jacentes, nous avons donc déterminé une relation entre les paramètres des variables binaires des populations P et P^*.

Remarque. *Pour estimer cette relation, il n'est nul besoin d'estimer les paramètres a_{kj}, b_{kj} et σ_{kj} qui sont les paramètres des variables gaussiennes et de la transformation entre ces variables gaussiennes. De plus, l'estimation de ces paramètres entraînerait un certain nombre de problèmes d'identifiabilité qui nécessiterait l'introduction de contraintes supplémentaires.*

En écrivant (4.12) sous la forme suivante :

$$\alpha^*_{kj} = \Phi\Big(\delta_{kj}\ \Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma_{kj}\Big), \tag{4.13}$$

on constate en effet que le lien entre les populations P et P^* est entièrement déterminé par les paramètres δ_{kj}, γ_{kj} et λ_j.

Le nombre de paramètres définissant le lien entre P et P^* est alors relativement important, Kd paramètres continus (les γ_{kj}) et $Kd + d$ paramètres discrets (δ_{kj} et λ_j), et ce pour estimer les Kd paramètres continus α^*_{kj}. Nous définissons alors quatre sous-modèles particuliers, en émettant certaines hypothèses sur les paramètres A_k et b_k de la transformation entre variables gaussiennes, ainsi que sur les matrices de variance Σ_k des variables gaussiennes.

Modèle M_1 : σ_{kj} libre, A_k libre et $b_k = 0$ (transformation réduite à une homothétie). Le modèle est alors :

$$\alpha^*_{kj} \quad = \quad \Phi\Big(\delta_{kj}\ \Phi^{-1}(\alpha_{kj})\Big) \qquad \text{avec} \qquad \delta_{kj} \in \{-1, 1\}.$$

Modèle M_2 : pour le deuxième modèle M_2, nous supposons l'homoscédasticité de P, *i.e.* $\sigma_{kj} = \sigma$, et une transformation entre P et P^* constante, *i.e.* indépendante de la classe et de la dimension, de sorte que $A_k = aI_d$ et $b_k = \beta e$ où $e = (1, \ldots, 1)'$ de dimension d. Le modèle s'écrit alors $\alpha^*_{kj} = \Phi\Big(\delta\ \Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma\Big)$ avec $\lambda_j \in \{-1, 1\}$, $\delta \in \{-1, 1\}$ et $\gamma \in \mathbb{R}$.
Or, ce modèle n'est pas identifiable, pour deux raisons :

- le produit $\lambda_j\gamma$ n'est pas identifiable (on peut intervertir les signes des deux membres du produit en gardant la même valeur du produit). On introduit alors le paramètre $\lambda'_j = \lambda_j\ \text{sgn}(\gamma)$, qui est donc le signe du produit $\lambda_j\gamma$.
- la transformation (4.11) n'est pas identifiable, car il existe plusieurs couples (A_k, b_k) conduisant à la même transformation. Ce problème n'existe pas pour M_1 car b_k est supposé nul pour ce dernier.

Par exemple, considérons la transformation qui fait correspondre à la gaussienne de centre $(-1,-1)$ et de matrice de variance identité la gaussienne de centre $(-2,-2)$ et de matrice de variance égale à quatre fois l'identité. On a alors deux solutions pour cette transformation : soit $A = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ et $b = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, soit $A = \begin{pmatrix} -2 & 0 \\ 0 & -2 \end{pmatrix}$ et $b = \begin{pmatrix} -4 \\ -4 \end{pmatrix}$.
Pour remédier à ceci, dès que le paramètre de translation b_k n'est pas nul, on impose une homothétie positive ($a_{kj} > 0$ pour tout $1 \leq k \leq K$ et $1 \leq j \leq d$). Le modèle est alors identifiable, et cette hypothèse n'entraîne aucune restriction puisque la translation reste libre. Cette hypothèse sera donc faite pour tous les modèles à venir.

Sous les hypothèses $\sigma_{kj} = \sigma$, $A_k = aI_d$ avec $a > 0$ et $b_k = \beta e$ (transformation constante pour la dimension et la classe), le modèle M_2 est :

$$\alpha^*_{kj} = \Phi\Big(\Phi^{-1}(\alpha_{kj}) + \lambda'_j|\gamma|\Big) \qquad \text{avec} \qquad \lambda'_j = \lambda_j \ \text{sgn}(\gamma) \in \{-1,1\} \quad \text{et} \quad |\gamma| \in \mathbb{R}^+.$$

Modèle M_3 : $\sigma_{kj} = \sigma_k$, $A_k = a_k I_d$ (avec $a_k > 0$ pour tout $1 \leq k \leq K$) et $b_k = \beta_k e$ (transformation dépendant uniquement de la classe). Le modèle est :

$$\alpha^*_{kj} = \Phi\Big(\Phi^{-1}(\alpha_{kj}) + \lambda'_{kj}|\gamma_k|\Big) \qquad \text{avec} \qquad \lambda'_{kj} = \lambda_j \ \text{sgn}(\gamma_k) \in \{-1,1\} \quad \text{et} \quad |\gamma_k| \in \mathbb{R}^+.$$

Modèle M_4 : $\sigma_{kj} = \sigma_j$, $A_k = A$ (avec $a_{kj} > 0$ pour tout $1 \leq k \leq K$ et $1 \leq j \leq d$) et $b_k = \beta$ (transformation dépendant uniquement de la dimension). Le modèle est alors :

$$\alpha^*_{kj} = \Phi\Big(\Phi^{-1}(\alpha_{kj}) + \gamma_j\Big) \qquad \text{avec} \qquad \gamma_j \in \mathbb{R}.$$

Notons que dans ce cas, comme le paramètre γ_j est libre, il intègre le paramètre λ_j.

À chacun de ces quatre modèles M_i ($i = 1, \ldots, 4$), nous pouvons associer une hypothèse supplémentaire sur les proportions : les proportions des groupes sont conservées ou non de P vers P^*. Nous notons alors M_i le modèle avec proportions conservées, et pM_i avec proportions différentes, ce qui conduit finalement à 8 modèles.

Remarque. *Emboîtement de modèles.*
Il est important de préciser les recoupements qui existent entre les huit modèles que l'on vient de définir. Le modèle M_2 (pM_2) est un sous-modèle des modèles M_3 (pM_3) et M_4 (pM_4), qui eux sont disjoints. Le modèle M_1 (pM_1) n'est compris dans aucun des autres modèles. En effet, si on modifie ses hypothèses en obligeant A_k à être positif, pour être en adéquation avec les autres modèles, on devrait alors transférer la liberté de A_k vers b_k. Ce dernier serait alors entièrement libre, c'est-à-dire vis-à-vis de la dimension mais aussi de la classe, ce qui n'est considéré dans aucun des trois autres modèles. Il existe néanmoins quelques cas particuliers (suivants que les centres des gaussiennes sous-jacentes et que la matrice d'homothétie A_k ne dépendent pas respectivement de la classe, de la dimension, ou d'aucun des deux) pour lesquels le modèle M_1 (pM_1) est un sous-modèle des autres modèles (respectivement M_4 (pM_4), M_3 (pM_3) et M_2 (pM_2)). Notons enfin que lorsque les paramètres δ_{kj} du modèle M_1 sont tous égaux à 1, le modèle de discrimination généralisée ainsi défini correspond au modèle de discrimination classique.

Remarque. *Nombre de paramètres à estimer.*
Les Kd paramètres α^*_{kj} sont obtenus en estimant les paramètres du modèle choisi, comme le montre le Tableau 4.2. On note que le nombre de paramètres continus au sein des modèles est alors moins grand que Kd.

	M_1	M_2	M_3	M_4	pM_1	pM_2	pM_3	pM_4
continus	0	1	K	d	$K-1$	K	$2K-1$	$d+K-1$
discrets	Kd	d	Kd	0	$K(d+1)-1$	$K+d-1$	$K(d+1)-1$	$d+K-1$

TAB. 4.2.: Nombre de paramètres continus et discrets à estimer pour les différents modèles.

4.2.4. Estimation des paramètres

L'analyse discriminante généralisée pour données binaires que nous proposons est une démarche composée de trois étapes, chacune donnant lieu à des estimations.
La première étape consiste à modéliser l'échantillon S de données binaires issu de la population P par un mélange de lois de Bernoulli, décrit en (4.3). Pour cet échantillon d'apprentissage, les appartenances aux classes des individus étant connues, l'estimation des paramètres α_{kj} (pour $1 \leq k \leq K$ et $1 \leq j \leq d$) et des proportions $p_1, \ldots, p_K$ du mélange est immédiate :

$$\hat{p}_k = \frac{\sum_{i=1}^{N} z_i^k}{N} \qquad \text{et} \qquad \hat{\alpha}_{kj} = \frac{\sum_{i=1}^{N} x_i^j z_i^k}{N}.$$

La deuxième étape consiste à modéliser les lois des variables binaires (mélange de lois de Bernoulli) de la population P^*. Comme nous l'avons vu précédemment, les paramètres α_{kj}^* de ces lois sont exprimés en fonction des paramètres α_{kj} de la population P ainsi que des paramètres λ_j, δ_{kj} et γ_{kj} caractérisant la transformation entre les deux populations P et P^*. Les α_{kj} étant déjà estimés, il suffira d'estimer à cette étape les paramètres de la transformation, ainsi que les proportions des mélanges si elles sont différentes entre P et P^*. Nous détaillerons ci-après les différentes techniques nécessaires à l'estimation de ces paramètres.
Enfin, une fois la population P^* modélisée, la troisième et dernière étape consiste à estimer les appartenances aux classes $z_1^*, \ldots, z_N^*$ des individus de S^* par maximum *a posteriori*.

Nous décrivons les techniques d'estimation nécessaires à la deuxième étape. À cette étape, nous supposons disposer d'une estimation des paramètres α_{kj} (pour $1 \leq k \leq K$ et $1 \leq j \leq d$) et des proportions $p_1, \ldots, p_K$ modélisant la population P. Nous présentons la situation où les proportions des mélanges sont différentes pour P et P^* et qu'il est nécessaire de les estimer dans P^*. Le cas contraire est immédiat.
Soit θ l'ensemble des paramètres à estimer, constitué des proportions $p_1^*, \ldots, p_K^*$, ainsi que des paramètres δ_{kj}, λ_j et γ_{kj} (pour $1 \leq k \leq K$ et $1 \leq j \leq d$) de la transformation entre P et P^*. Soit Θ l'espace dans lequel évolue θ.
En utilisant l'expression de la distribution de probabilité de $\mathbf{X}^*$, donnée en (4.6), la vraisemblance du modèle s'écrit :

$$L(\theta) = \prod_{i=1}^{N^*} \sum_{k=1}^{K} p_k^* \prod_{j=1}^{d} {\alpha_{kj}^*}^{x_i^{*j}} (1 - \alpha_{kj}^*)^{1 - x_i^{*j}}.$$

Étant en présence de données manquantes que sont les appartenances z_i^* des individus x_i^* aux classes, la maximisation de la vraisemblance peut se faire avec l'algorithme EM [14] qui est bien adapté à ce cas.
La log-vraisemblance complétée s'écrit :

$$l_c(\theta; z_1^*, \ldots, z_{N^*}^*) = \sum_{i=1}^{N^*} \sum_{k=1}^{K} z_i^{*k} \log \left(p_k^* \prod_{j=1}^{d} {\alpha_{kj}^*}^{x_i^{*j}} (1 - \alpha_{kj}^*)^{(1 - x_i^{*j})} \right).$$

Partant d'une valeur initiale du paramètre à estimer, l'algorithme EM est un algorithme itératif composé de deux étapes successives, l'étape E (pour *Expectation*) et l'étape M (pour *Maximization*). L'étape E consiste à calculer l'espérance de la log-vraisemblance complétée par rapport à la loi des données manquantes (ici les appartenances aux classes) conditionnellement aux données (ici les valeurs des variables explicatives) et à la valeur courante du paramètre à estimer. L'étape M consiste alors à choisir comme nouvelle estimation du paramètre celle qui maximise l'espérance de la log-vraisemblance complétée calculée à l'étape E. Ces deux étapes sont répétées jusqu'à ce que la log-vraisemblance devienne stationnaire.
Nous détaillons les deux étapes de l'algorithme EM.

L'étape E de l'algorithme EM consiste, à partir d'une estimation courante $\theta^{(q)}$, à calculer l'espérance de la log-vraisemblance complétée par rapport à la loi $p(\mathbf{z}^*|\mathbf{x}^*;\theta^{(q)})$ des appartenances aux classes $\mathbf{z}^* = (z_1^*,\ldots,z_{N^*}^*)$ conditionnellement aux valeurs des variables explicatives $\mathbf{x}^* = (x_1^*,\ldots,x_{N^*}^*)$ et à la valeur courante $\theta^{(q)}$ du paramètre θ. Après quelques calculs, cette espérance s'écrit :

$$\begin{aligned}\mathcal{Q}(\theta;\theta^{(q)}) &= E_{\theta^{(q)}}[l_c(\theta;Z_1^*,\ldots,Z_{N^*}^*)|x_1^*,\ldots,x_{N^*}^*] \\ &= \sum_{i=1}^{N^*}\sum_{k=1}^{K} t_{ik}^{(q)}\Big\{\log(p_k^*)+\log\Big(\prod_{j=1}^{d}{\alpha_{kj}^*}^{x_i^{*j}}(1-\alpha_{kj}^*)^{1-x_i^{*j}}\Big)\Big\}\end{aligned}$$

où

$$t_{ik}^{(q)} = p(Z_i^{*k}=1|x_1^*,\ldots,x_{N^*}^*;\theta^{(q)}) = \frac{p_k^{*(q)}\prod_{j=1}^{d}(\alpha_{kj}^{*}{}^{(q)})^{x_i^{*j}}(1-\alpha_{kj}^{*}{}^{(q)})^{(1-x_i^{*j})}}{\sum_{\kappa=1}^{K}p_\kappa^{*(q)}\prod_{j=1}^{d}(\alpha_{\kappa j}^{*}{}^{(q)})^{x_i^{*j}}(1-\alpha_{\kappa j}^{*}{}^{(q)})^{(1-x_i^{*j})}},$$

est la probabilité conditionnelle que l'individu i appartienne à la classe k.

L'étape M de l'algorithme EM consiste alors à choisir $\theta^{(q+1)}$ dans Θ qui maximise cette espérance conditionnelle :

$$\theta^{(q+1)} = \underset{\theta\in\Theta}{\operatorname{argmax}}\,\mathcal{Q}(\theta;\theta^{(q)}). \tag{4.14}$$

Cette maximisation conduit au nouvel estimateur des proportions p_k^* suivant :

$$p_k^{*(q+1)} = \frac{\sum_{i=1}^{N^*} t_{ik}^{(q)}}{N^*}.$$

Cette expression est obtenue par résolution du problème d'optimisation (4.14), sous la contrainte que la somme des proportions soit égale à 1.
Pour les paramètres λ_j, δ_{kj} et γ_{kj}, il n'est pas évident d'obtenir une formulation explicite de leur estimateur obtenu à l'étape M. Néanmoins la maximisation de $\mathcal{Q}(\theta;\theta^{(q)})$ en fonction des paramètres λ_j et δ_{kj} ne pose aucun problème (éventuellement des problèmes combinatoires comme nous le verrons plus loin) puisqu'ils prennent leurs valeurs dans un ensemble fini ($\{-1,1\}$). Ce qui résout l'estimation du modèle pM_1.
Pour les autres modèles, on doit maximiser $\mathcal{Q}(\theta;\theta^{(q)})$ en fonction des paramètres (λ_j,γ_{kj}). Pour cela, nous maximisons la fonctionnelle $\mathcal{Q}$ en fonction de γ_{kj} pour chaque valeur possible des paramètres λ_j, et nous choisissons ensuite parmi tous les (λ_j,γ_{kj}) obtenus celui qui rend $\mathcal{Q}$ maximum. Notons que dans les modèles 2 et 3 ce sont les valeurs absolues des paramètres γ et γ_k qui interviennent, mais que nous estimons

néanmoins leur valeur algébrique, puisque nous avons besoin de leur signe pour déduire des estimations de λ celle de $\lambda_j^{'}$ et $\lambda_{kj}^{'}$.

Remarque. *En pratique, le bouclage sur les valeurs possibles des paramètres discrets ne se fait pas au sein de l'étape M, mais en dehors de l'algorithme EM : nous estimons les paramètres p_k^* et γ_{kj} à l'aide de l'algorithme EM pour chaque valeur possible des paramètres discrets δ_{kj} et λ_j, puis nous choisissons la solution de vraisemblance maximale.*

La maximisation de la fonctionnelle $\mathcal{Q}$ en fonction du paramètre continu γ_{kj} est un problème d'optimisation classique. Soient $\mathcal{Q}_2$, $\mathcal{Q}_3$ et $\mathcal{Q}_4$ les fonctionnelles à maximiser en fonction de γ, γ_k et γ_j pour les modèles M_2 (pM_2), M_3 (pM_3) et M_4 (pM_4), les paramètres λ_j et λ_{kj} étant fixés :

$$\begin{aligned}\mathcal{Q}_2(\gamma) = \sum_{i=1}^{N^*}\sum_{k=1}^{K} t_{ik}\Big\{ \log(p_k^*) \quad &+ \quad \sum_{j=1}^{d} x_i^{*j} \log\Big(\Phi(\Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma)\Big) \\ &+ \quad \sum_{j=1}^{d}(1 - x_i^{*j}) \log\Big(1 - \Phi(\Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma)\Big)\Big\}.\end{aligned}$$

$$\begin{aligned}\mathcal{Q}_3(\gamma_1,\ldots,\gamma_K) = \sum_{i=1}^{N^*}\sum_{k=1}^{K} t_{ik}\Big\{ \log(p_k^*) \quad &+ \quad \sum_{j=1}^{d} x_i^{*j} \log\Big(\Phi(\Phi^{-1}(\alpha_{kj}) + \lambda_{kj}\gamma_k)\Big) \\ &+ \quad \sum_{j=1}^{d}(1 - x_i^{*j}) \log\Big(1 - \Phi(\Phi^{-1}(\alpha_{kj}) + \lambda_{kj}\gamma_k)\Big)\Big\},\end{aligned}$$

$$\begin{aligned}\mathcal{Q}_4(\gamma_1,\ldots,\gamma_d) = \sum_{i=1}^{N^*}\sum_{k=1}^{K} t_{ik}\Big\{ \log(p_k^*) \quad &+ \quad \sum_{j=1}^{d} x_i^{*j} \log\Big(\Phi(\Phi^{-1}(\alpha_{kj}) + \gamma_j)\Big) \\ &+ \quad \sum_{j=1}^{d}(1 - x_i^{*j}) \log\Big(1 - \Phi(\Phi^{-1}(\alpha_{kj}) + \gamma_j)\Big)\Big\},\end{aligned}$$

Comme Φ est une fonction de répartition, il vient que

$$\lim_{|\gamma|\to\infty} \mathcal{Q}_2(\gamma) = \lim_{\|\gamma_1,\ldots,\gamma_K\|\to\infty} \mathcal{Q}_3(\gamma_1,\ldots,\gamma_K) = \lim_{\|\gamma_1,\ldots,\gamma_d\|\to\infty} \mathcal{Q}_4(\gamma_1,\ldots,\gamma_d) = -\infty$$

De plus, les fonctionnelles $\mathcal{Q}_2$, $\mathcal{Q}_3$ et $\mathcal{Q}_4$ sont des fonctions de γ, $(\gamma_1,\ldots,\gamma_K)$ et $(\gamma_1,\ldots,\gamma_d)$ strictement concaves (cf. annexe B.1). Il est donc possible d'appliquer un algorithme de type Newton pour rechercher leur maximum [10].
Ce qui résout les estimations des modèles M_2 (pM_2), M_3 (pM_3) et M_4 (pM_4).

4.2.4.1. Estimation en grande dimension

L'augmentation de la dimension pose des problèmes numériques. En effet, à l'étape M de l'algorithme EM, pour les modèles M_1 et pM_1, il faut rechercher parmi toutes les valeurs possibles de δ_{kj} celle qui maximise une certaine quantité. Or en dimension 10, avec $K = 2$, le nombre de δ_{kj} possibles est 2^{20}, soit plus d'un million de possibilités. Le même type de problème existe pour les modèles M_2 (pM_2) et M_3 (pM_3), avec les paramètres λ_j et λ_{kj}. Précisons néanmoins que pour ces derniers modèles, lorsqu'un ordre naturel est connu au sein de variables explicatives, celui-ci impose le choix de la discrétisation (4.7), et les paramètres λ_j et λ_{kj} sont donc connus. Dans ce paragraphe, nous traitons l'estimation de δ_{kj}, mais les mêmes

méthodes sont utilisables pour l'estimation des autres paramètres discrets λ_j et λ_{kj}.
Il devient vite impossible lorsque la dimension d dépasse 10 (situation fréquente dans le cas de données binaires) de parcourir toutes les valeurs de ces paramètres discrets.
Nous proposons alors deux méthodes d'optimisation stochastique : une méthode de relaxation et la méthode du recuit simulé. Ces méthodes sont des méthodes approchées qui nous permettent de trouver une estimation de la valeur optimale du paramètre sans avoir à parcourir exhaustivement l'espace dans lequel évolue ce paramètre.
Nous nous plaçons dans le cas du modèle M_1 où les proportions sont connues. Le seul paramètre de la log-vraisemblance est alors δ_{kj}. Nous verrons plus loin comment faire lorsque l'on doit estimer d'autres paramètres conjointement à ce paramètre δ_{kj}.

La méthode de relaxation consiste à relâcher l'hypothèse que le paramètre δ_{kj} appartient à l'espace discret $\{-1, 1\}^{Kd}$, et à supposer qu'il appartient à l'espace continu $[-1, 1]^{Kd}$. De la même façon que nous venons de le faire pour l'optimisation en fonction des paramètres continus γ_{kj} à l'étape M de l'algorithme EM (cf. annexe B.1), on peut montrer que les fonctions à optimiser par rapport aux paramètres continus $\tilde{\delta}_{kj}$ sont strictement concaves et tendent vers $-\infty$ sur les bords du domaine (de même pour les autres paramètres discrets λ_j et λ_{kj}). Ainsi, une méthode d'optimisation classique de type Newton permet de résoudre le problème d'optimisation. Nous obtenons ainsi une solution continue $\tilde{\delta}_{kj}$ que nous discrétisons pour obtenir une solution binaire par $\delta_{kj} = \text{sgn}(\tilde{\delta}_{kj})$. Le principal problème de ce genre de technique est qu'il n'existe aucun résultat théorique prouvant la convergence vers la valeur optimale du paramètre.

La seconde méthode est celle du recuit simulé qui a été introduite par Metropolis *et al.* en 1953 [23] pour minimiser un critère sur un ensemble fini de très grande taille. On en trouve une présentation claire dans [28].
Le principe du recuit simulé consiste à se déplacer plus rapidement sur la surface de la fonction à optimiser, afin d'éviter de rester au voisinage d'un mimimum local, et ce en effectuant des changements d'échelle.
L'algorithme du recuit simulé pour la recherche du paramètre δ_{kj} qui maximise la log-vraisemblance est le suivant.
Soit l'estimation courante $\delta_{kj}^{(i)}$.

(i) Simuler ξ_{kj} dans un voisinage proche de $\delta_{kj}^{(i)}$.

(ii) Accepter $\delta_{kj}^{(i+1)} = \xi_{kj}$ avec la probabilité $p_i = \min(\exp(\frac{\Delta E_i}{T_i}), 1)$,
prendre $\delta_{kj}^{(i+1)} = \delta_{kj}^{(i)}$ sinon.

(iii) Augmenter T_i en T_{i+1}.

La première étape consiste à simuler une réalisation ξ_{kj} de δ_{kj} qui soit dans un voisinage de $\delta_{kj}^{(i)}$. Cette réalisation peut être vue comme une perturbation de la valeur courante de δ_{kj}. Dans le cas continu on simulera par exemple une variable uniforme dans un voisinage de $\delta_{kj}^{(i)}$. Dans notre cas discret, on change à chaque perturbation une seule des Kd composantes de $\delta_{kj}^{(i)}$ (de 1 en -1 ou vice-versa), cette composante étant choisie aléatoirement.
La deuxième étape consiste à évaluer le gain en log-vraisemblance obtenu avec ce ξ_{kj}. Si ce gain $\Delta E_i = l(\xi_{kj}) - l(\delta_{kj}^{(i)})$ est positif, $l(.)$ étant la log-vraisemblance, on choisi alors ξ_{kj} comme nouvelle réalisation de δ_{kj}. Sinon, on se donne une probabilité non nulle de le choisir néanmoins comme nouvelle réalisation de δ_{kj}. C'est cette probabilité qui nous permet d'échapper à l'attraction de $\delta_{kj}^{(i)}$ si jamais ce dernier est un maximum local. Cette probabilité p_i est définie par $p_i = \min(\exp(\frac{\Delta E_i}{T_i}), 1)$. Ainsi, si le gain en log-vraisemblance ΔE_i est positif, $\exp(\frac{\Delta E_i}{T_i})$ est supérieur à 1, d'où $p_i = 1$. Si le gain est négatif $p_i = \exp(\frac{\Delta E_i}{T_i})$.
Le paramètre T_i est un paramètre de « température » que l'on fait croître au cours des simulations de réalisations de δ_{kj}, afin de favoriser de plus en plus les solutions de grande vraisemblance à mesure que l'on avance dans les simulations. Un T_i proche de 1 en début de simulation permet d'explorer au maximum l'espace dans lequel évolue la log-vraisemblance, afin de visiter tous les maxima locaux, et lorsque ce T_i croît,

on se resserre de plus en plus vers le maximum global.
La troisième étape consiste donc à faire croître ce paramètre T_i, comme nous venons de l'annoncer. Comme il l'est explicité dans [28], la convergence théorique de l'algorithme du recuit simulé est établie lorsque la croissance de T_i est logarithmique. Mais en pratique, cette croissance est trop faible, et on utilise généralement une croissance linéaire, du type $T_{i+1} = \vartheta T_i$, où $\vartheta > 1$.
Contrairement à la méthode de relaxation, nous savons donc que si l'on attendait un temps infini, la valeur optimale du paramètre δ_{kj} serait atteinte. Mais la principale difficulté de cette méthode est le choix du paramètre ϑ, qui est spécifique à chaque problème.

Remarque. *L'algorithme présenté ci-dessus est en fait l'algorithme de Metropolis pour simuler la densité* $exp(\frac{1}{T}l(\delta_{kj}))$*, qui est utilisé dans les méthodes de Monte Carlo par Chaîne de Markov [28]. Nous créons ainsi une chaîne de Markov dont chaque élément est une réalisation du paramètre* δ_{kj}*.*

Estimation de plusieurs paramètres

Si les proportions au sein de la population test sont inconnues, ou si le modèle étudié contient les paramètres γ ou γ_k (modèles M_2, M_3, pM_2 et pM_3), les paramètres discrets ne sont plus les seuls à devoir être estimés. On procède alors de la façon suivante :

- pour la méthode de relaxation, on utilise l'algorithme EM défini précédemment, en incluant à chaque étape M la recherche du paramètre discret « relaxé » qui maximise la log-vraisemblance.
- pour la méthode du recuit simulé, l'algorithme EM n'est plus utilisé. À chaque génération d'un nouvel élément de la chaîne de Markov, on recherche les autres paramètres optimaux, soit par une méthode de type Newton (pour les paramètres γ et γ_k), soit en utilisant l'expression explicite de l'estimateur optimal obtenue à l'étape M de l'algorithme EM (pour les proportions).

Test des méthodes de relaxation et du recuit simulé

Nous testons les deux méthodes du recuit simulé et de relaxation pour l'estimation du paramètre δ_{kj} du modèle M_1, dans le cas où les proportions sont les mêmes dans les populations d'apprentissage et de test. Les tests sont réalisés en dimension 5 et 10.
Le tableau 4.3 présente les pourcentages moyens (sur 30 répétitions) de δ_{kj} bien estimés par la méthode de relaxation.

		dimension	
méthode	N	$d = 5$	$d = 10$
relaxation	100	93	80
relaxation	1000	100	100

TAB. 4.3.: Pourcentages de δ_{kj} bien estimés par la méthode de relaxation.

Remarque. *Les tests ont été effectués avec une transformation entre* P *et* P^* *respectant les hypothèses du modèle* M_1*. Les paramètres utilisés (loi des données et transformation) sont les mêmes que ceux utilisés pour les tests sur simulation présentés ci-après.*

La méthode de relaxation donne de très bons résultats pour l'estimation du paramètre δ_{kj}, et semble converger lorsque la taille de l'échantillon augmente.
Quant à la méthode du recuit simulé, elle souffre du choix du paramètre ϑ, qui influe beaucoup sur les résultats d'estimations. La figure 4.1 illustre l'évolution de la log-vraisemblance du modèle M_1, en dimension 10, pour la méthode du recuit simulé pour deux mauvais choix du paramètre ϑ. En effet, dans la figure de gauche, ϑ est trop petit ($\vartheta = 1.0000000001$) ce qui implique que la surface de la log-vraisemblance n'est

pas assez explorée. On risque alors de rester sur un maximum local. Au contraire, dans la figure de droite, ϑ est trop grand ($\vartheta = 1.0001$), et on n'arrive pas à se focaliser sur le maximum global.

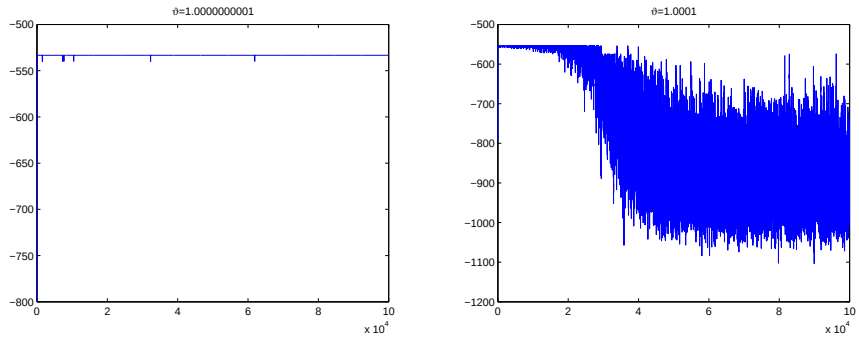


FIG. 4.1. : Évolution au cours des simulations (en abscisse) de la log-vraisemblance (en ordonnée) du modèle M_1 pour la méthode du recuit simulé avec un ϑ trop petit (gauche) et un ϑ trop grand (droite).

La figure 4.2 illustrent l'évolution de la log-vraisemblance du modèle M_1 pour un choix plus convenable du paramètre ϑ ($\vartheta = 1.000001$). Mais là encore, les fluctuations d'échantillonnage influent sur l'exploration de la log-vraisemblance. En effet, avec le même paramètre $\vartheta = 1.000001$, on explore plus (figures 4.2 du haut) ou moins (figure 4.2 du bas) la surface de la log-vraisemblance. Tout l'enjeu étant de l'explorer assez pour ne manquer aucun maximum local, mais de pouvoir se concentrer sur le maximum global au fur et à mesure que l'on avance dans les simulations.

Le pourcentage de δ_{kj} bien estimés avec ce paramètre $\vartheta = 1.000001$ est de 65% en dimension 10. Si l'on augmente ϑ, ce pourcentage diminue car plus on avance dans les simulations plus les sauts à chaque pas sont importants et plus on s'éloigne de la solution otpimale. Si l'on diminue ϑ, on augmente au contraire ce pourcentage, car on converge très rapidement vers la solution optimale. Mais malheureusement il est possible que la solution obtenue ne soit qu'un maximum local, ce qui ne semble pas être le cas ici mais qui pourrait très bien l'être.

Enfin, concernant le temps de calcul, la méthode de relaxation est plus rapide, puisque pour la méthode du recuit simulé on est obligé d'effectuer un nombre important de simulations (100000 par exemple), alors que pour l'algorithme EM la convergence est souvent observée bien avant ce nombre d'itérations.

En raison des difficultés de réglage qui existent pour la méthode du recuit simulé, et sachant que ces réglages sont propres à chaque problème, nous utilisons pour les tests de la discrimination généralisée en dimension 10 la méthode de relaxation.

Le logiciel Matlab est utilisé pour implémenter l'estimation des huit modèles de discrimination généralisée.

4.2.5. Tests sur simulations

Un ensemble de tests sur simulations a été réalisé, afin de comparer l'analyse discriminante généralisée à deux autres démarches consistant à : (1) utiliser directement sur l'échantillon test la règle de classement construite à partir de l'échantillon d'apprentissage (discrimination classique), (2) oublier que l'on a un échantillon d'apprentissage et effectuer une classification automatique (ou *clustering*) sur l'échantillon test.
La discrimination généralisée propose huit modèles différents : M_1 à M_4 et pM_1 à pM_4. Afin de choisir le

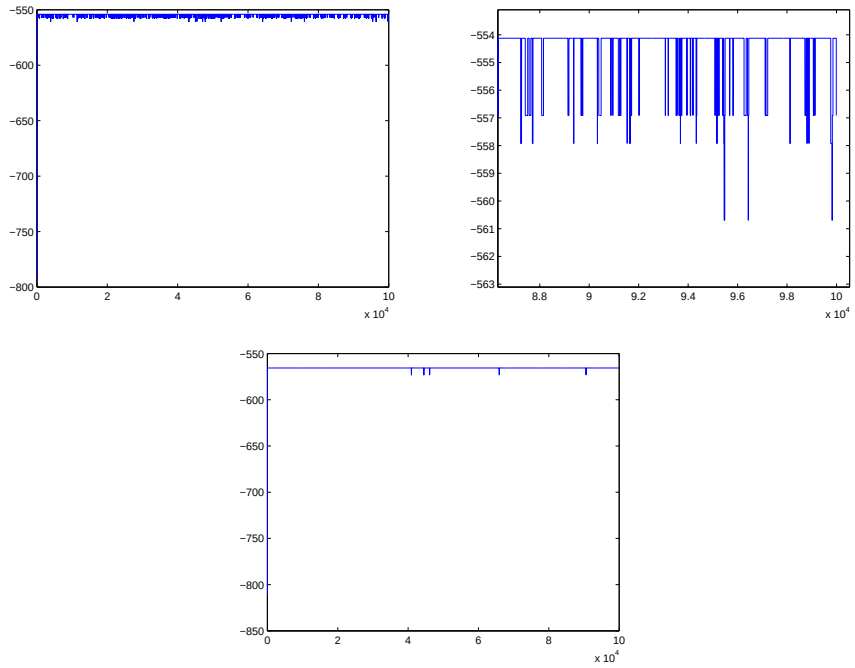


FIG. 4.2. : Évolution au cours des simulations (en abscisse) de la log-vraisemblance (en ordonnée) du modèle M_1 pour la méthode du recuit simulé avec $\vartheta = 1.000001$. Deux essais différents sont présentés (premier essai en haut, évolution sur les 100000 simulations à gauche puis sur les 3000 dernières à droite, et second essai en bas).

modèle le plus approprié aux données testées, nous les comparons à l'aide d'un critère fréquemment utilisé en choix de modèle : le critère BIC (*Bayesian Information Criterion*, [57]). L'idée de cette approche bayésienne de choix de modèle consiste à choisir le modèle m de plus grande probabilité *a posteriori* $p(m|\mathcal{D})$. Cette probabilité peut être exprimée par : $p(m|\mathcal{D}) \propto p(\mathcal{D}|m)p(m)$, où $p(\mathcal{D}|m)$ est appelée vraisemblance intégrée ou vraisemblance marginale et $p(m)$ est la probabilité *a priori* du modèle m. L'évaluation de cette vraisemblance intégrée étant souvent difficile, on utilise l'approximation consistant à approcher la log-vraisemblance intégrée par le maximum de la log-vraisemblance, pénalisée par une fonction du nombre ν de paramètres continus du modèle et de la taille N de l'échantillon. Cette approximation conduit alors au critère BIC, défini par :

$$BIC = -2l(\hat{\theta}) + \nu \log(N),$$

où $l(\hat{\theta})$ est le maximum de la log-vraisemblance.
Ce critère est préféré à des critères issus d'une approche fréquentiste comme le critère NIC (*Network Information Criterion*, [24, 25, 26]) ou sa version simplifiée AIC (*An Information Criterion*, [**?**, 2]), en raison de l'inconsistence de ces derniers qui les mène à choisir un modèle trop complexe avec une probabilité non nulle [8].

Les comparaisons de l'analyse discriminante généralisée, de la discrimination classique et de la classification automatique sont faites en terme d'erreur de classement, évaluées sur l'échantillon test simulé.

4.2.5.1. Protocole

Nous appliquons les trois méthodes de classification (discrimination généralisée, discrimination classique et classification automatique) sur des échantillons d'apprentissage et de test simulés à partir d'une discrétisation de variables gaussiennes. Huit transformations entre les populations d'apprentissage et de test sont testées, chacune d'entres elles correspondant aux hypothèses spécifiques des huit modèles M_1 à M_4 et pM_1 à pM_4. Ces tests sont naturellement favorables à la discrimination généralisée, puisqu'ils respectent les hypothèses de cette dernière :

- les variables binaires sont issues d'une discrétisation de variables gaussiennes,
- l'indépendance conditionnelle des variables gaussiennes Y^j est respectée,
- la transformation entre P et P^* est $\mathcal{C}^1$ et de matrice d'homothétie A_k diagonale.

Nous testons ensuite quatre autres types de simulations mettant en défaut les hypothèses de la discrimination généralisée précédemment citées, afin de tester sa robustesse :

- Les variables binaires sont simulées par discrétisation de variables gaussiennes bruitées par une certaine proportion p de variables uniformes, centrées sur chaque gaussienne du mélange et étendues à plus ou moins deux écart-types (bruit 1).
- Les variables binaires, simulées par discrétisation de variables gaussiennes, sont bruitées par une proportion p de variables binaires réparties uniformément sur $\{0,1\}^d$ (bruit 2).
- Les variables binaires sont simulées par discrétisation de variables gaussiennes non conditionnellement indépendantes (bruit 3).
- La matrice d'homothétie A_k de la transformation entre P et P^* n'est pas diagonale (bruit 4).

Notons que l'hypothèse d'une transformation $\mathcal{C}^1$ entre P et P^* est naturelle, et nous ne réalisons donc pas de test la mettant en défaut.
Comme pour la simulation sans bruit, nous testons pour chaque type de bruit huit transformations correspondant aux huit modèles M_1 à M_4 et pM_1 à pM_4.

Pour chaque test, les estimations sont répétées trente fois, afin d'avoir une estimation moyenne :

- des taux d'erreur de classement obtenus par chaque méthode de discrimination,
- des valeurs des critères BIC pour les modèles de discrimination généralisée et de discrimination classique,
- des pourcentages de choix de chaque modèle de discrimination généralisée par le critère BIC.

Nous choisissons un espace des variables explicatives de dimension $d = 5$ ($\mathcal{X} = \mathbb{R}^5$) puis $d = 10$ ($\mathcal{X} = \mathbb{R}^{10}$), les données étant réparties en deux classes ($K = 2$). Pour la dimension 10, quelques restrictions sur le nombre de tests effectués seront appliquées. Deux tailles d'échantillons sont utilisées : $N = 100$ puis $N = 1000$. Le détail des autres paramètres utilisés (génération des données, transformation), figure en annexe B.2.
Pour l'algorithme EM, nous avons fixé le seuil de convergence de la log-vraisemblance à 10^{-6}, en se fixant un nombre maximum d'itérations à 200.

4.2.5.2. Résultats en dimension 5

Les tests sur des échantillons de données binaires simulées par des données gaussiennes non bruitées, dont les résultats sont présentés en annexe B.2.1 par les tableaux B.1 et B.2, permettent de tirer les conclusions suivantes :

- le critère BIC permet de choisir généralement le modèle correspondant à la transformation entre P et P^* utilisée. Néanmoins, pour une taille d'échantillon de 100, l'emboîtement des modèles 2, 3 et 4 rend difficile le choix entre ces trois modèles (le bon modèle est choisi dans 60% des cas). Mais ce problème est quasiment résolu lorsque l'on passe à une taille d'échantillon de 1000, puisque le bon modèle est alors choisi dans plus de 98% des cas.

– Les taux d'erreurs obtenus par discrimination généralisée pour les modèles choisis par BIC sont significativement meilleurs que ceux obtenus par discrimination classique (même lorsque BIC ne choisit pas le modèle correspondant à la vraie transformation), et sont très proches des taux d'erreur optimaux pour une taille d'échantillon de 1000. La classification automatique est quant à elle moins bonne que les deux autres méthodes.

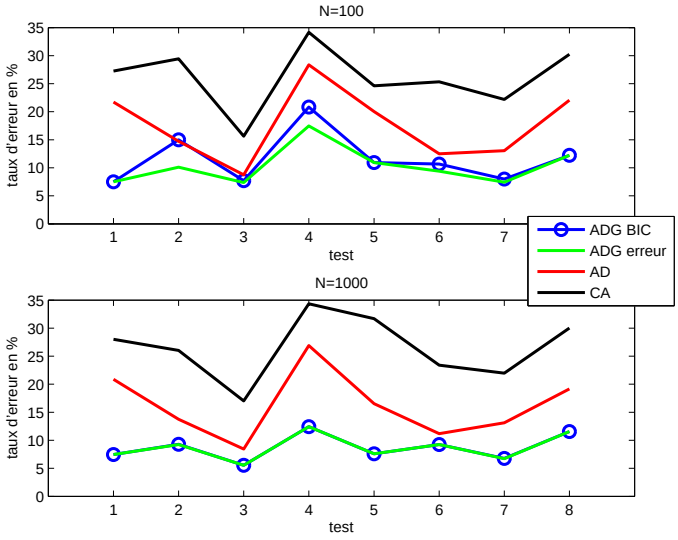


FIG. 4.3. : Tests sans bruits.

La figure 4.3 illustre ces résultats, en présentant les taux de mauvais classements obtenus par le meilleur modèle de discrimination généralisée selon le critère BIC (ADG BIC) puis selon l'erreur de classement (ADG erreur), par discrimination classique (AD) et enfin par classification automatique (CA). L'axe des abscisses représente les huit tests effectués.

Pour le premier type de bruit (bruit uniforme sur les variables gaussiennes), les résultats sont présentés en annexe B.2.2 dans les tableaux B.3 et B.4 pour une proportion de bruit de 10%, et dans les tableaux B.5 et B.6 pour une proportion de bruit de 30%. Pour le deuxième type de bruit (bruit uniforme sur les variables binaires), les résultats sont présentés en annexe B.2.3 dans les tableaux B.7 et B.8 pour 10% de bruit, et dans les tableaux B.9 et B.10 pour 30% de bruit. Les figures 4.4 et 4.5 illustrent ces résultats, avec les mêmes notations que pour la figure 4.3.

Pour ces deux types de bruits, les conclusions sont semblables :

– avec une proportion raisonnable de bruit (10%), le critère BIC a plus de mal à choisir le bon modèle correspondant à la transformation que lorsqu'il n'y a pas de bruit, et le problème dû à l'emboîtement des modèles 2, 3 et 4 est amplifié. Il arrive néanmoins à choisir le bon modèle dans 73% des cas avec une taille d'échantillon de 1000.
– Du point de vue des erreurs de classements, les modèles de discrimination généralisée choisis par le critère BIC, qu'ils soient ou non le bon modèle vis-à-vis de la transformation, donnent de meilleurs résultats (ou semblables dans certains cas rares) que la discrimination classique ou que la classification automatique, qui cette fois se rapproche des deux autres méthodes d'un point de vue erreur de classement.
– Pour une proportion de bruit plus importante (30%), le choix des modèles devient impossible puisque les données sont tellement bruitées qu'elles ne correspondent plus à aucun des modèles. Les erreurs de classements par discrimination généralisée et discrimination classique deviennent très importantes,

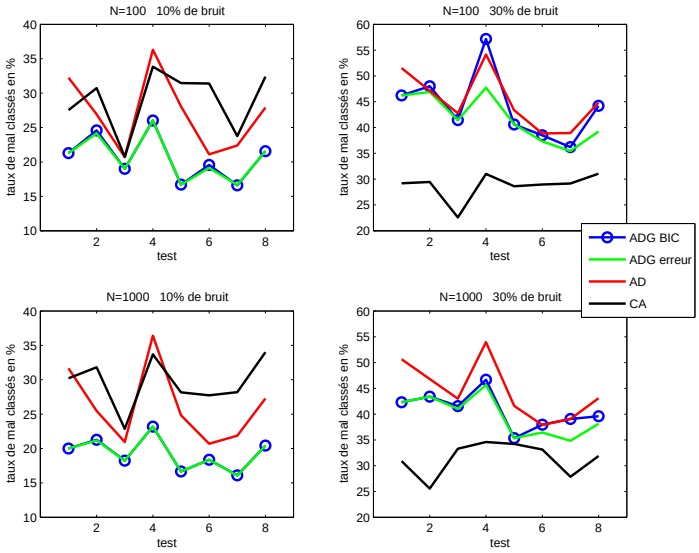


FIG. 4.4. : Tests avec bruit 1.

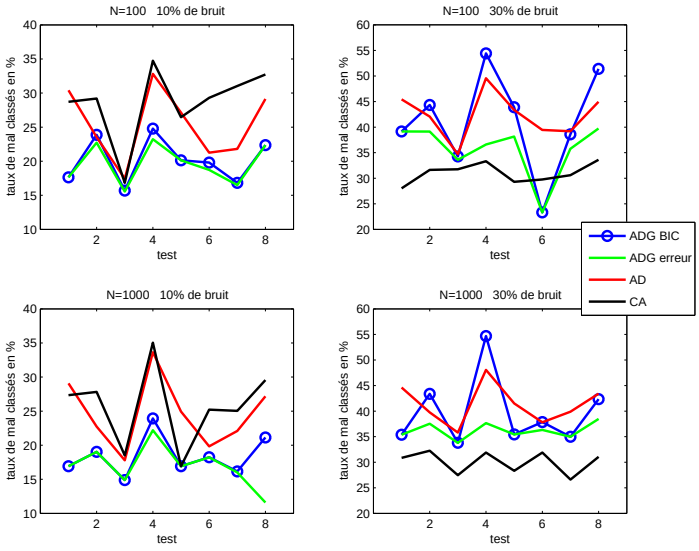


FIG. 4.5. : Tests avec bruit 2.

c'est alors la classification automatique qui est la meilleure.

Pour le troisième type de bruit, où l'hypothèse d'indépendance conditionnelle des variables gaussiennes est mise à défaut, les résultats sont présentés en annexe B.2.3 dans le tableau B.11. Les conclusions sont les suivantes :

- le critère BIC choisit toujours le modèle de discrimination généralisée pM_3, qui n'est pas forcément celui qui donne le taux d'erreur de classement le plus faible.
- Les taux d'erreurs obtenus par discrimination généralisée avec le modèle pM_3 sont meilleurs que ceux obtenus par discrimination classique, et meilleurs dans 75% des cas que ceux obtenus par classifica-

tion automatique.

Enfin, pour le quatrième type de bruit, qui propose une transformation avec matrice d'homothétie non diagonale, les résultats sont présentés en annexe B.2.4 dans le tableau B.12. Les conclusions sont les suivantes :

- le critère BIC hésite entre les modèles M_3 et pM_3, sans trop de cohérence : il ne converge pas forcément vers un choix lorsque la taille de l'échantillon augmente, et lorsqu'il converge ce n'est pas forcément vers le bon modèle d'un point de vue conservation des proportions.
- Néanmoins, les taux d'erreurs obtenus par discrimination généralisée avec les modèles choisis par BIC sont meilleurs que ceux obtenus par discrimination classique ou par classification automatique.

Remarque. *Choix de modèle en présence de bruit important.*
On peut remarquer que pour les bruits 3 et 4, les meilleurs modèles sont les modèles M_3 et pM_3. Nous pensons que ces modèles sont ceux qui permettent de s'adapter le mieux à ce type de bruit. Quant au bruit 1, le choix à l'aide du critère BIC semble se porter vers le modèle pM_4 lorsque la taille de l'échantillon est assez grande, bien que ce ne soit pas ce modèle qui donne les plus petites erreurs de classement.

4.2.5.3. Résultats en dimension 10

Pour les essais numériques que nous présentons ici, nous nous restreignons à l'étude d'un seul type de bruit, le bruit 1 qui consiste en un bruit uniforme sur les variables gaussiennes qui génèrent les variables binaires par discrétisation.
L'optimisation des paramètres discrets des modèles 2, 3 et 4 est faite à l'aide de la méthode de relaxation présentée précédemment.
La figure 4.6 illustre les résultats obtenus.

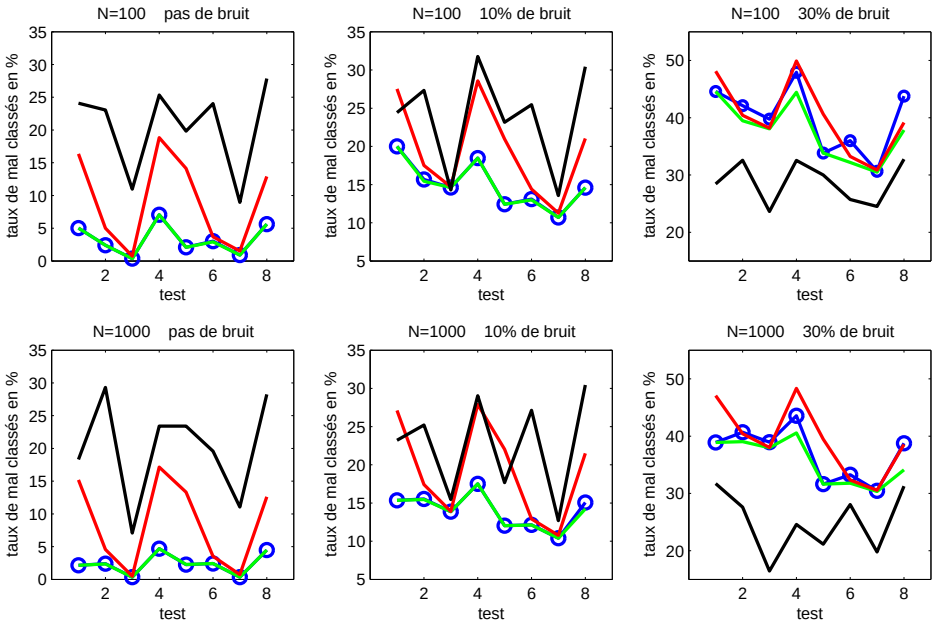


FIG. 4.6. : Tests sans bruit et avec bruit 1.

Les résultats obtenus sont relativement analogues à ceux obtenus en dimension 5. D'un point de vue erreur

de classement, les modèles de discrimination généralisée choisis par BIC sont meilleurs que la discrimination classique et que la classification automatique, pour des simulations non bruitées et avec 10% de bruit. Comme en dimension 5 lorsque le bruit est trop important c'est la classification automatique qui est la plus performante.

4.2.5.4. Conclusions

Dans un cadre de discrimination sur variables binaires avec populations d'apprentissage et de test différentes, et lorsque les variables binaires sont dues à une discrétisation de variables gaussiennes sous-jacentes, l'analyse discriminante généralisée donne des erreurs de classement plus faibles que l'analyse discriminante classique ou que la classification automatique.
De plus, les différents tests effectués ont permis de montrer la robustesse de la discrimination généralisée :

- à l'hypothèse de normalité sous-jacente,
- à l'hypothèse des classes latentes (indépendance conditionnelle des variables gaussiennes sous-jacentes),
- à l'hypothèse qui contraint la transformation entre P et P^* à être « variable par variable » (matrice d'homothétie A_k diagonales).

Notons néanmoins que lorsque l'on s'éloigne « trop » de la normalité sous-jacente, les discriminations classique et généralisée deviennent mauvaises, et c'est alors la classification automatique qui est la meilleure.

4.3. Applications à des données réelles

Ce paragraphe contient deux applications de l'analyse discriminante généralisée pour variables binaires. La première application est issue des travaux précurseurs sur l'analyse discriminante généralisée ([9], [35]), et consiste à discriminer une population d'oiseaux en fonction de leur sexe, à partir de données biométriques que nous avons discrétisées pour le besoin de l'étude. La seconde application, dans un contexte de santé publique, consiste à discriminer une population de patients ayant eu un premier cancer en fonction du risque de survenue d'un deuxième cancer.
Nous introduirons finalement une troisième possibilité d'application de l'analyse discriminante généralisée dans le domaine des assurances.

4.3.1. Sexes d'oiseaux

La discrimination généralisée sur variables gaussiennes donne de très bons résultats pour la problématique des sexes d'oiseaux présentée en introduction de ce chapitre (cf. Biernacki et *al*. [9]), qui consiste à discriminer une population d'oiseaux à partir de variables biométriques, avec des populations d'apprentissage et de test d'origines géographiques différentes.
L'espèce d'oiseau de mer considérée, *Calanectris diomedea*, est présente en Méditerranée et dans l'Atlantique Nord, ces conditions océanographiques contrastées étant présumées à l'origine de l'existence de différences notables en taille, en couleur et en comportement entre ces différentes sous-espèces. Nous disposons d'un échantillon de trois sous-espèces différentes, pour lesquelles cinq variables morphologiques ont été mesurées sur chaque individu (cf. Thibault et *al*. [32]). Ces variables sont les suivantes :

- profondeur du bec,
- longueur du bec,
- longueur du tarse (os de la patte),
- longueur des ailes,
- longueur de la queue.

Chaque échantillon comporte une partie de mâles et une partie de femelles, et chaque individu est de sexe connu.

Nous proposons une première illustration de nos travaux à partir de cette application, en discrétisant les

variables morphologiques continues en variables binaires, tout en sachant que l'information perdue est alors importante. La figure 4.7 illustre les différences entre deux variables morphologiques (tailles des ailes et de la queue), pour les deux espèces *diomedea* et *borealis*.

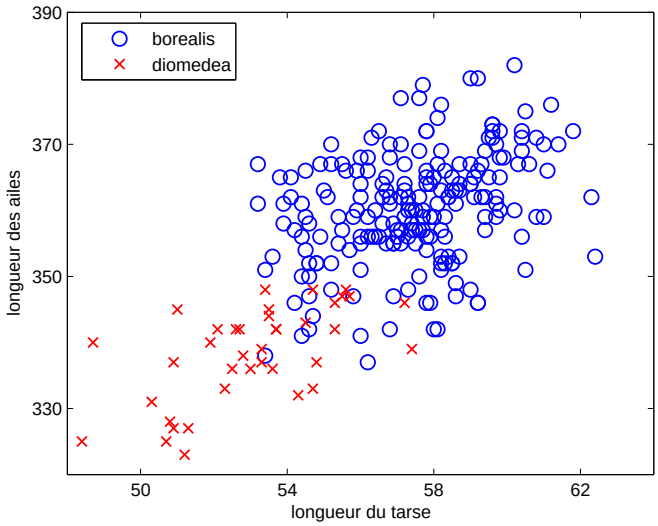


FIG. 4.7. : Longueur des ailes et de la queue pour les deux espèces *diomedea* et *borealis*.

Comme on peut le constater sur cette figure, la discrétisation doit être réalisée avec précaution. Effectivement, si l'on discrétise par exemple par rapport aux moyennes des variables biométriques pour une espèce, toutes les valeurs pour l'autre espèce se trouvent alors du même coté de cette discrétisation, du fait des différences importantes de taille entre ces espèces.
Ainsi, nous choisissons de ne travailler que sur les deux espèces les plus proches, *diomedea* et *borealis*, et nous choisissons les seuils de discrétisation de sorte à ce que l'on ait un maximum d'individus de chaque coté de ce seuil pour les deux espèces.
La tableau 4.4 présente les valeurs minimales et maximales des cinq variables morphologiques pour les deux espèces considérés (en mm).

		prof. bec		long. bec		long. tarse		long. ailes		long. queue	
		min	max	min	max	min	max	min	max	min	max
diomedea	mâle	12.7	15.3	47.5	55.2	48.4	57.4	325	348	122	134
	femelle	11.7	15.5	46.1	55	48.7	55.3	323	345	117	140
borealis	mâle	14.3	18.3	51.3	61.2	53.9	62.4	337	382	128	153
	femelle	13	16.4	47.8	59.1	53.2	61	338	376	127	153

TAB. 4.4.: Minimums et maximums des différentes variables morphologiques pour les deux espèces *diomedea* et *borealis*.

Les seuils de discrétisation optimaux, choisis en explorant tous les seuils possibles, sont alors donnés par le tableau 4.5.

Nous choisissons pour population d'apprentissage l'espèce *borealis*, dont nous disposons d'un échantillon de 206 individus (45% de femelles), et pour population test l'espèce *diomedea*, dont nous disposons d'un échantillon de 38 individus (58% de femelles).

prof. bec	long. bec	long. tarse	long. ailes	long. queue
14.35	51.50	54.45	342.50	128.40

TAB. 4.5.: Seuils de discrétisation des variables morphologiques pour les deux espèces *diomedea* et *borealis*.

Les trois méthodes de classification, discrimination généralisée, discrimination classique et classification automatique sont testées. Les résultats sont présentés dans le Tableau 4.6. Pour la classification automatique, la procédure a été répétée 100 fois, et nous présentons le taux de mauvais classement moyen, ainsi que l'écart-type entre parenthèse.

modèle	pM_1	pM_2	pM_3	pM_4	M_1	M_2	M_3	M_4	D.	C.A.
taux	57.9	21	21	21	57.9	23.7	15.8	18.4	42.1	29.5 (9.4)
nombre	22	8	8	8	22	9	6	7	16	11.2 (9.45)

TAB. 4.6.: Taux (en $\%$) et nombre (sur 38 individus) de mauvais classements obtenus sur la population test *diomedea*, avec apprentissage sur la population *borealis*.

L'utilisation du critère BIC pour choisir entre les huit modèles de discrimination généralisée (tableau 4.7) conduit à sélectionner le modèle pM_3, qui donne une erreur de classement de 21.05%. Cette erreur est plus faible que celles obtenues par discrimination classique (42.11%) et par classification automatique (29.50%). Notons que les modèles M_3 et M_4 donnent des erreurs de classements encore plus faibles (respectivement deux et un mauvais classements en moins), mais le critère BIC ne les choisit pas. Néanmoins, la différence de valeur de BIC entre M_3 et pM_3 est peu significative.

pM_1	pM_2	pM_3	pM_4	M_1	M_2	M_3	M_4
269.75	222.50	**220.55**	237.02	267.35	221.59	221.47	233.62

TAB. 4.7.: Valeurs du critère BIC pour les huit modèles de discrimination généralisée.

Cette application illustre l'intérêt de la discrimination généralisée vis-à-vis des procédures classiques de discrimination linéaire ou de classification automatique. En effet, en adaptant la règle de classement issue de l'échantillon d'apprentissage à l'échantillon test en fonction des différences entre les deux populations d'apprentissage et de test, la discrimination généralisée permet de classer les individus de la population test de façon plus efficace qu'en appliquant directement la règle de classement issue de l'échantillon d'apprentissage, ou encore en oubliant l'échantillon d'apprentissage et en effectuant une classification automatique directement sur l'échantillon test.
Remarquons aussi que l'hypothèse gaussienne conditionnellement au sexe était relativement acceptable sur les variables biométriques sous-jacentes. Néanmoins, il y avait une forte corrélation entre ces différents caractères, violant alors les hypothèses d'indépendance conditionnelle de nos modèles.

Nous présentons maintenant une seconde application où certaines variables sont naturellement binaires.

4.3.2. Risque de deuxième cancer

Le Registre des tumeurs du Doubs est un registre général de population dont les missions sont :
– la production de statistiques pour la surveillance épidémiologique des cancers par l'enregistrement continu et exhaustif des nouveaux cas de cancers résidant au moment du diagnostic dans le département du Doubs,

– la réalisation d'études et de recherches relatives notamment à l'épidémiologie, l'évaluation du dépistage des cancers, les prises en charge diagnostiques et thérapeutiques, la qualité des soins et la qualité de vie.

Un des axes de recherche actuel au sein de ce registre porte sur le risque d'apparition d'un deuxième cancer chez des patients ayant été atteints par un premier cancer du testicule. Dans le cadre de cette étude, nous disposons d'un ensemble de données constitué de 299 patients ayant eu ce premier cancer entre 1978 et 2002, la date de point (date de fin de suivi des patients) étant fixée au 31 décembre 2002.
L'étude que nous proposons consiste à prédire parmi les individus ayant eu un premier cancer du testicule récemment, ceux qui ont un risque important de survenue d'un quelconque deuxième cancer. Pour effectuer cette discrimination, nous nous appuyons sur les données relatives aux patients ayant eu un premier cancer antérieurement, pour lesquels nous avons assez de recul pour constater s'ils ont eu ou non un deuxième cancer.

La population d'apprentissage représente donc les individus ayant eu un premier cancer du testicule entre 1978 et 1992 et habitant dans le département du Doubs au moment du diagnostic, pour lesquels nous savons s'ils ont eu un deuxième cancer ou non. Nous disposons ainsi d'un échantillon de 149 patients, parmi lesquels :

– 19 patients ont eu un deuxième cancer,
– 20 patients sont décédés sans avoir eu de deuxième cancer.

Pour les 110 patients restants, nous supposons qu'ils n'ont pas de deuxième cancer. En effet, leur premier cancer ayant été contracté avant 1992, on dispose d'une durée de suivi d'au moins 10 années (on a vérifié qu'ils habitaient toujours dans le Doubs en 2002), que nous supposons suffisante à la survenue d'un deuxième cancer. Cette hypothèse semble raisonnable et est discutée à la fin du paragraphe 4.3.2.2.
La population test à classer est composée d'individus du Doubs ayant eu un premier cancer du testicule entre 1993 et 2002. Nous disposons d'un échantillon de 150 patients, parmi lesquels 14 sont étiquetés :

– 1 patient a eu un deuxième cancer,
– 13 patients sont décédés sans avoir eu de deuxième cancer.

Puisque le temps de suivi de ces patients est inférieur à 10 années, notre objectif est de prédire quels sont parmi ces 150 patients ceux qui ont un risque important de survenue d'un deuxième cancer, sachant que l'on dispose de 14 individus pour tester notre prévision.

Étant donné la différence temporelle entre nos deux populations d'apprentissage et de test, il est fort probable qu'une discrimination classique donne de mauvais résultats. En effet, l'évolution de certains paramètres comme les traitements médicaux ou le comportement des patients vis-à-vis de ces traitements entre les périodes 1978-1992 et 1993-2002 a certainement un impact sur l'évolution du risque d'apparition d'un deuxième cancer.

Pour les deux populations d'apprentissage et de test, nous disposons pour chaque individu de cinq variables dites explicatives ou descriptives. Ces variables sont soit binaires par nature, soit discrétisées pour le besoin de notre étude :

– l'état du patient à la date de dernière nouvelle (1 : décédé ou 0 : vivant),
– la nature du cancer du testicule (1 : tumeurs germinales séminomateuses ou 0 : non séminomateuses),
– l'âge du patient (1 : supérieur à l'âge médian, 34 ans, calculé sur les deux échantillons d'apprentissage et de test ou 0 : inférieur à l'âge médian),
– la commune de résidence à la date de dernière nouvelle (1 : ville et agglomération (éloignée d'au maximum 7km) ou 0 : campagne),
– la commune de résidence à la date de diagnostic du premier cancer (même discrétisation).

La date de dernière nouvelle étant la date de point, *i.e.* le 31 décembre 2002.

Remarque 4.3.1. *L'utilisation de la variable état pour la prédiction du risque de survenue d'un second cancer peut être criticable, puisque les patients décédés ne pourront plus avoir de nouveau cancer. Nous la conservons néanmoins afin de ne pas trop réduire le nombre de variables explicatives, et nous nous assurerons dans les conclusions de notre étude que les patients détectés comme étant à risque sont bien encore vivant à la date de dernière nouvelle.*

Avant toute discrimination, nous devons choisir parmi les cinq variables explicatives celles qui sont (les plus) discriminantes, à partir de l'information contenue dans l'échantillon d'apprentissage. Nous appliquons ensuite l'analyse discriminante généralisée ainsi que l'analyse discriminante classique et la classification automatique sur l'échantillon test. À l'aide des 14 données tests de ce dernier, nous choisissons le meilleur modèle pour finalement prédire quels sont les individus ayant un risque important de survenue d'un deuxième cancer.

4.3.2.1. Étude préliminaire : choix des variables discriminantes

Les cinq variables dont on dispose ne sont pas forcément toutes discriminantes vis-à-vis du risque de deuxième cancer. Nous étudions dans un premier temps la liaison de chacune de ces cinq variables avec l'apparition d'un deuxième cancer. Les tableaux de contingence correspondant aux données de l'échantillon d'apprentissage sont présentés par le tableau 4.8.

		état		
		0	1	
deuxième cancer	0	110	20	130
	1	8	11	19
		118	31	149

		nature		
		0	1	
deuxième cancer	0	68	62	130
	1	4	15	19
		72	77	149

		âge		
		0	1	
deuxième cancer	0	79	51	130
	1	3	16	19
		82	67	149

		com. résid.		
		0	1	
deuxième cancer	0	53	77	130
	1	11	8	19
		64	85	149

		com. diag.		
		0	1	
deuxième cancer	0	53	77	130
	1	10	9	19
		63	86	149

TAB. 4.8.: Tableaux de contingence des variables explicatives et de l'apparition d'un deuxième cancer (0 : non, 1 : oui).

Remarque. *Une question importante que se posent les praticiens hospitaliers est de savoir si la nature de la tumeur du cancer du testicule a une influence sur l'apparition d'un deuxième cancer. La lecture du tableau de contingence correspondant à la nature du cancer du testicule nous apporte un premier élément de réponse : parmi les 149 individus de la population d'apprentissage, la répartition de cancer du testicule à tumeurs germinales séminomateuses et non séminomateuses est approximativement égale (77 contre 72 patients sur 149), tandis que si l'on se restreint aux patients ayant eu un deuxième cancer, on constate que la proportion d'individus ayant eu un premier cancer à tumeurs germinales séminomateuses est beaucoup*

plus importante (15 sur 19 pour les tumeurs germinales séminomateuses contre 4 sur 19 pour les tumeurs germinales non séminomateuses). Il semble donc que les cancers du testicule à tumeurs germinales séminomateuses favorisent l'apparition d'un deuxième cancer.

La liaison entre les différentes variables explicatives et l'apparition d'un deuxième cancer peut être évaluée à l'aide du coefficient de contingence de Pearson (voir par exemple [29]). Les valeurs de ces coefficients sont présentés dans le tableau 4.9.

Coeff. Conting.	état	nature	âge	commune diag.	commune résid.
deuxième cancer	0.33	0.2	0.29	0.08	0.11

TAB. 4.9.: Coefficients de contingence de Pearson des variables explicatives et de l'apparition d'un deuxième cancer.

On constate que les variables relatives aux communes de résidence à la date du diagnostic et à la date de dernière nouvelle semblent avoir une moins forte corrélation avec la variable d'apparition de deuxième cancer que les trois autres variables. Ceci ne suffit néanmoins pas à choisir les variables discriminantes.
Pour ce faire, nous effectuons plusieurs analyses discriminantes sur la population d'apprentissage, en utilisant les cinq variables explicatives, puis toutes les combinaisons possibles de quatre variables, et enfin de trois variables. Pour chaque discrimination, nous découpons l'échantillon d'apprentissage en deux sous-échantillons servant de base d'apprentissage et de test. Nous répétons 300 fois ce découpage et évaluons les taux moyens de mauvais classements obtenus. Nous choisissons le plus petit ensemble de variables conduisant au plus petit taux d'erreur de classement.
Les résultats, présentés en annexe B.4.1, confirment ce que nous pressentions suite à l'examen des coefficients de contingence, à savoir que les trois variables discriminantes sont l'âge du patient au moment du diagnostic, la nature de la tumeur et l'état du patient à la date de dernière nouvelle.

4.3.2.2. Prédiction du risque de deuxième cancer

Ayant choisi trois variables discriminantes, nous appliquons l'analyse discriminante généralisée afin de classer les 150 patients ayant eu un premier cancer entre 1993 et 2002.
Nous rappelons que l'échantillon d'apprentissage est constitué des 149 patients ayant eu un premier cancer entre 1978 et 1992, pour lesquels nous supposons connaître s'ils ont eu ou non un deuxième cancer.

Remarque. *Il peut être envisagé de se restreindre à n'utiliser que les patients pour lesquels nous savons avec certitude s'ils ont eu ou non un deuxième cancer (patients ayant eu un deuxième cancer et patients décédés sans deuxième cancer). On obtient alors un échantillon d'apprentissage de 39 patients, qui est trop petit pour notre étude. En effet, il conduit à classer tous les patients (ou au moins une grande majorité) comme patients à risque, et ce quelle que soit la méthode utilisée.*

Afin de choisir le meilleur modèle pour classer les individus (discrimination généralisée, discrimination classique ou classification automatique), nous évaluons les taux de mauvais classements obtenus sur les 14 patients pour lesquels nous savons s'ils ont eu un deuxième cancer (1 patient) ou non (13 patients décédés sans deuxième cancer). Les résultats sont présentés par le tableau 4.10. La première ligne de ce tableau correspond aux taux de mauvais classements obtenus pour les 13 patients décédés sans deuxième cancer, la deuxième ligne correspond aux taux de mauvais classements pour le patient ayant eu un deuxième cancer, et la troisième ligne aux taux de mauvais classements globaux.

Le meilleur modèle est le modèle de discrimination généralisée pM_3, qui classe correctement les 14 patients étiquetés de l'échantillon test. Par contre, le critère BIC choisit le modèle pM_4, qui est celui qui donne le

deuxième cancer	modèle									
	pM_1	pM_2	pM_3	pM_4	M_1	M_2	M_3	M_4	D.	C.A.
oui	38.46	38.46	**0**	61.54	38.46	38.46	0	38.46	38.46	92.31
non	100	100	**0**	100	100	100	100	100	100	100
total	42.86	42.86	**0**	64.29	42.86	42.86	7.14	42.86	42.86	92.86

TAB. 4.10.: Taux de mauvais classements sur les 14 patients étiquetés de l'échantillon test.

plus grand taux d'erreur (tableau 4.11).

modèle							
pM_1	pM_2	pM_3	pM_4	M_1	M_2	M_3	M_4
513.34	524.79	529.80	506.22	508.65	523.56	526.42	509.05

TAB. 4.11.: Critères BIC pour les huit modèles de discrimination généralisée.

Nous choisissons néanmoins d'utiliser le modèle pM_3 pour prédire les individus à risque de l'échantillon d'apprentissage. Ainsi, la discrimination généralisée détecte 56 individus ayant un risque important de survenue de deuxième cancer parmi les 150 patients ayant eu un premier cancer du testicule entre 1993 et 2002. Après examen précis de ces 56 patients, il s'avère qu'ils correspondent tous au même type de patients : patient ayant eu son cancer du testicule de type séminomateux à un âge supérieur à 34 ans (âge médian), et étant encore vivant à la date de dernière nouvelle (ce qui est logique conformément à la remarque 4.3.1).

Remarque. *Il n'est pas surprenant que ces 56 patients correspondent à la même modalité de nos trois variables explicatives, étant donné que le nombre de modalités possibles pour ces trois variables binaires est restreint :* $2^3 = 8$.

Remarque. *Nous avons supposé qu'un suivi de 10 années suffisait à détecter la survenue d'un deuxième cancer. Cette hypothèse qui pouvait sembler raisonnable à l'examen des données (nous le confirmons ci-dessous), et qui permettait d'avoir un échantillon d'apprentissage de taille confortable, est peut-être trop forte. Nous l'avons donc supposé dans un second temps de 15 ans. Ainsi, nous avons enlevé de l'échantillon d'apprentissage les individus ayant eu un premier cancer entre 1988 et 1992, et nous les avons intégré à l'échantillon test. Nous présentons en annexe B.4.2 cette étude, qui considère donc comme population d'apprentissage les individus ayant eu un premier cancer entre 1978 et 1987 et comme population test les individus ayant eu un premier cancer entre 1988 et 2002. Néanmoins, en examinant les 69 patients que l'on enlève de l'échantillon d'apprentissage par rapport à l'étude précédente (patients ayant eu un premier cancer entre 1988 et 1992), on constate que 7 ont déjà eu avant 2002 un deuxième cancer, soit une proportion d'environ* 10.14%*, ce qui n'est pas très éloigné des* 12.90% *de deuxième cancer pour les individus ayant eu un premier cancer entre 1978 et 1987. Il semble donc raisonnable de les conserver dans l'échantillon d'apprentissage, conformément à l'étude présentée dans ce paragraphe.*

4.3.2.3. Conclusion

La discrimination généralisée permet de détecter relativement efficacement les patients ayant un risque important de survenue de deuxième cancer (aucune erreur de classement avec le modèle pM_3 sur les 14 patients étiquetés de l'échantillon test).
De plus, les prévisions obtenues semblent être pessimistes au vue des proportions de deuxième cancer estimées (environ 30% de patients à fort risque de deuxième cancer au sein de la population test alors qu'au

sein de la population d'apprentissage la proportion était de 13%). Il est préférable d'être pessimiste plutôt qu'optimiste, puisqu'il est plus intéressant de se tromper en classant à risque un patient qui n'aura pas de deuxième cancer plutôt que le contraire.
Enfin, les 56 patients détectés comme patients à risque sont les patients ayant eu leur cancer du testicule de type séminomateux à un âge supérieur à 34 ans (âge médian) et étant encore vivant à la date de dernière nouvelle (31 décembre 2002).

Ces résultats sont transmis au Registre des tumeurs du Doubs et sont actuellement à l'étude par les médecins et statisticiens de ce registre. Une utilisation de ces résultats peut être dans un premier temps d'accentuer la surveillance de ces 56 patients, afin de détecter au plus tôt la survenue d'un éventuel deuxième cancer, et ainsi d'augmenter les chances de guérison.
Ces travaux donneront suite à une autre étude plus précise au sein du Registre des tumeurs de Doubs, qui consistera à rechercher au sein des dossiers des patients des informations supplémentaires : taille de la tumeur, présence ou non de métastase, traitement par radiothérapie ou chimiothérapie.

4.3.3. Une perspective d'application dans le domaine des assurances

Un des domaines d'application où l'on rencontre souvent des données binaires est le monde des assurances. Nous présentons ici un exemple d'application de l'analyse discriminante généralisée dans ce domaine, sachant que nous n'avons pas de données concrètes à l'heure actuelle.
Considérons qu'une compagnie d'assurance, basée à Paris, désire s'implanter dans une certaine région de province. Son expérience parisienne lui a permis d'acquérir un certain nombre d'informations sur ses clients parisiens, et entre autre de construire des règles de discrimination entre différents groupes de clients (par exemple, clients à risque et clients sûrs). Les caractéristiques de la population de province n'étant pas les mêmes que celles de la population parisienne (salaire, santé...), les règles de discrimination construites à partir de son expérience parisienne ne peuvent pas être appliquées directement sur les clients de province. La discrimination généralisée s'avère donc être un outil très intéressant pour la compagnie d'assurance en vue d'utiliser son expérience parisienne pour s'implanter dans une autre région.
Une partie de ces travaux a en outre été présentée au colloque : *Data Mining et Apprentissage Statistique. Applications en Assurance*, 12-13 mai 2005, Niort, France [21].

Annexe B

Annexes de la partie II

B.1. Concavité stricte des fonctionnelles $\mathcal{Q}_2$, $\mathcal{Q}_3$ et $\mathcal{Q}_4$

B.1.1. Concavité stricte de $\mathcal{Q}_2$

Il faut montrer que la fonction suivante est une fonction strictement concave de γ :

$$\mathcal{Q}_2(\gamma) = \sum_{i=1}^{N^*}\sum_{k=1}^{K} t_{ik}\Big\{ \log(p_k^*) \quad + \quad \sum_{j=1}^{d} x_i^{*j} \log\Big(\Phi(\Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma)\Big) \\ + \quad \sum_{j=1}^{d}(1 - x_i^{*j}) \log\Big(1 - \Phi(\Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma)\Big)\Big\}.$$

On a :

$$\frac{\partial \mathcal{Q}_2(\gamma)}{\partial \gamma} = \sum_{i=1}^{N^*}\sum_{k=1}^{K} t_{ik} \sum_{j=1}^{d} \Big\{ x_i^{*j}\lambda_j \frac{\phi(\Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma)}{\Phi(\Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma)} + (1 - x_i^{*j})\lambda_j \frac{-\phi(\Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma)}{1 - \Phi(\Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma)}\Big\},$$

où ϕ est la densité de probabilité de la loi normale centrée réduite.
En utilisant la propriété $\frac{\partial\phi(bx+a)}{\partial x} = -b(x+a)\phi(x+a)$, et en posant $\zeta_{kj} = \Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma$ on obtient :

$$\frac{\partial^2 \mathcal{Q}_2(\gamma)}{\partial \gamma} = \sum_{i=1}^{N^*}\sum_{k=1}^{K} t_{ik} \sum_{j=1}^{d} \lambda_j \Big\{ \qquad x_i^{*j}\Big(\frac{-\lambda_j\zeta_{kj}\phi(\zeta_{kj})\Phi(\zeta_{kj}) - \lambda_j\phi(\zeta_{kj})^2}{[\Phi(\zeta_{kj})]^2}\Big) \\ + \quad (1 - x_i^{*j})\Big(\frac{-\lambda_j\zeta_{kj}\phi(\zeta_{kj})(1 - \phi(\zeta_{kj})) - \lambda_j\phi(\zeta_{kj})^2}{[1 - \Phi(\zeta_{kj})]^2}\Big)\Big\},$$

ou encore

$$\frac{\partial^2 \mathcal{Q}_2(\gamma)}{\partial \gamma} = \sum_{i=1}^{N^*}\sum_{k=1}^{K} t_{ik} \sum_{j=1}^{d} \lambda_j^2 \Big\{ \qquad \frac{x_i^{*j}}{[\Phi(\zeta_{kj})]^2} \overbrace{(-\zeta_{kj}\phi(\zeta_{kj})\Phi(\zeta_{kj}) - \phi(\zeta_{kj})^2)}^{-g_1(\zeta_{kj})} \\ + \quad \frac{(1 - x_i^{*j})}{[1 - \Phi(\zeta_{kj})]^2} \overbrace{(-\zeta_{kj}\phi(\zeta_{kj})(1 - \phi(\zeta_{kj})) - \phi(\zeta_{kj})^2)}^{-g_2(\zeta_{kj})}\Big\}.$$

Pour montrer que $\mathcal{Q}_2$ est concave, il suffit donc de montrer que les deux numérateurs des fractions précédentes sont strictement négatifs, pour toute valeur ζ_{kj} réelle, ou autrement dit que les fonctions g_1 et g_2 sont strictement positives :

- $\forall x \in \mathbb{R} : g_1(x) = x\Phi(x) + \phi(x) > 0$,
  puisque $g_1'(x) = \Phi(x) + x\phi(x) - x\phi(x) = \Phi(x) > 0$ donc g_1 est strictement croissante et $\lim_{x\to-\infty} g_1(x) = 0$,

– $\forall x \in \mathbb{R} : g_2(x) = -x + x\Phi(x) + \phi(x) > 0$,
puisque $g_2'(x) = -1 + \Phi(x) + x\phi(x) - x\phi(x) = \Phi(x) - 1 < 0$ donc g_2 est strictement décroissante et $\lim_{x\to+\infty} g_2(x) = 0$.

Ainsi $\frac{\partial^2 \mathcal{Q}_2(\gamma)^2}{\partial \gamma} < 0$ et donc $\mathcal{Q}_2$ est strictement concave. □

B.1.2. Concavité stricte de $\mathcal{Q}_3$

Il faut montrer que la fonction suivante est une fonction strictement concave de $\gamma_1, \ldots, \gamma_K$:

$$\mathcal{Q}_3(\gamma_1, \ldots, \gamma_K) = \sum_{i=1}^{N^*}\sum_{k=1}^{K} t_{ik}\Big\{ \log(p_k^*) \quad + \quad \sum_{j=1}^{d} x_i^{*j} \log\Big(\Phi(\Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma_k)\Big) + \quad \sum_{j=1}^{d}(1 - x_i^{*j}) \log\Big(1 - \Phi(\Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma_k)\Big)\Big\},$$

ce qui revient à montrer la stricte concavité de $q_3^{(k)}(\gamma_k)$:

$$q_3^{(k)}(\gamma_k) = \sum_{i=1}^{N^*} t_{ik}\Big\{ \log(p_k^*) \quad + \quad \sum_{j=1}^{d} x_i^{*j} \log\Big(\Phi(\Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma_k)\Big) + \quad \sum_{j=1}^{d}(1 - x_i^{*j}) \log\Big(1 - \Phi(\Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma_k)\Big)\Big\}.$$

On a :

$$\frac{\partial q_3^{(k)}(\gamma_k)}{\partial \gamma_k} = \sum_{i=1}^{N^*} t_{ik} \sum_{j=1}^{d} \Big\{ x_i^{*j}\lambda_j \frac{\phi(\Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma_k)}{\Phi(\Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma_k)} + (1 - x_i^{*j})\lambda_j \frac{-\phi(\Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma_k)}{1 - \Phi(\Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma_k)} \Big\}.$$

En notant $\zeta_{kj} = \Phi^{-1}(\alpha_{kj}) + \lambda_j\gamma_k$ on obtient comme précédemment :

$$\frac{\partial^2 q_3^{(k)}(\gamma)}{\partial \gamma^2} = \sum_{i=1}^{N^*} t_{ik} \sum_{j=1}^{d} \lambda_j^2 \Big\{ \qquad x_i^{*j}\Big(\frac{-\zeta_{kj}\phi(\zeta_{kj})\Phi(\zeta_{kj}) - \phi(\zeta_{kj})^2}{[\Phi(\zeta_{kj})]^2}\Big) + \quad (1 - x_i^{*j})\Big(\frac{-\zeta_{kj}\phi(\zeta_{kj})(1 - \phi(\zeta_{kj})) - \phi(\zeta_{kj})^2}{[1 - \Phi(\zeta_{kj})]^2}\Big)\Big\}.$$

Or, nous avons montré précédemment que les deux numérateurs des fractions précédentes sont strictement négatifs, d'où $q_3^{(k)}$ est strictement concave. □

B.1.3. Concavité stricte de $\mathcal{Q}_4$

Il faut montrer que la fonction suivante est une fonction strictement concave de $\gamma_1, \ldots, \gamma_d$:

$$\mathcal{Q}_4(\gamma_1, \ldots, \gamma_d) = \sum_{i=1}^{N^*}\sum_{k=1}^{K} t_{ik}\Big\{ \log(p_k^*) \quad + \quad \sum_{j=1}^{d} x_i^{*j} \log\Big(\Phi(\Phi^{-1}(\alpha_{kj}) + \gamma_j)\Big) + \quad \sum_{j=1}^{d}(1 - x_i^{*j}) \log\Big(1 - \Phi(\Phi^{-1}(\alpha_{kj}) + \gamma_j)\Big)\Big\},$$

ce qui revient à montrer la stricte concavité de $q_4^{(k)}(\gamma_j)$:

$$q_4^{(j)}(\gamma_j) = \sum_{i=1}^{N^*}\sum_{k=1}^{K} t_{ik}\Big\{ \log(p_k^*) \quad + \quad x_i^{*j}\log\Big(\Phi(\Phi^{-1}(\alpha_{kj})+\gamma_j)\Big) \\ + \quad (1-x_i^{*j})\log\Big(1-\Phi(\Phi^{-1}(\alpha_{kj})+\gamma_j)\Big)\Big\}.$$

On a :

$$\frac{\partial q_4^{(j)}(\gamma_j)}{\partial \gamma_j} = \sum_{i=1}^{N^*}\sum_{k=1}^{K} t_{ik}\Big\{ x_i^{*j}\frac{\phi(\Phi^{-1}(\alpha_{kj})+\gamma_j)}{\Phi(\Phi^{-1}(\alpha_{kj})+\gamma_j)} + (1-x_i^{*j})\frac{-\phi(\Phi^{-1}(\alpha_{kj})+\gamma_j)}{1-\Phi(\Phi^{-1}(\alpha_{kj})+\gamma_j)}\Big\}.$$

En notant $\zeta_{kj} = \Phi^{-1}(\alpha_{kj})+\gamma_j$ on obtient :

$$\frac{\partial^2 q_4^{(j)}(\gamma_j)}{\partial \gamma_j^2} = \sum_{i=1}^{N^*}\sum_{k=1}^{K} t_{ik}\Big\{ \qquad x_i^{*j}\Big(\frac{-\zeta_{kj}\phi(\zeta_{kj})\Phi(\zeta_{kj})-\phi(\zeta_{kj})^2}{[\Phi(\zeta_{kj})]^2}\Big) \\ + \quad (1-x_i^{*j})\Big(\frac{-\zeta_{kj}\phi(\zeta_{kj})(1-\phi(\zeta_{kj}))-\phi(\zeta_{kj})^2}{[1-\Phi(\zeta_{kj})]^2}\Big)\Big\}.$$

Comme nous avons déjà montré que les deux numérateurs des fractions précédentes sont strictement négatifs, $q_4^{(j)}$ est strictement concave. □

B.2. Résultats des tests sur simulations en dimension 5

Les valeurs des paramètres pour tous les tests sur simulations ont été choisies arbitrairement.
Les données binaires sont simulées à partir d'une discrétisation d'un mélange de deux gaussiennes de centres et de matrice de variance :

$$\mu_1 = \begin{pmatrix} -2 \\ -1 \\ -1.5 \\ 1.4 \\ -1.2 \end{pmatrix} \mu_2 = \begin{pmatrix} 1 \\ 1.5 \\ 2 \\ 2.1 \\ 0.9 \end{pmatrix} \qquad \Sigma_1 = \Sigma_2 = \begin{pmatrix} 3 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix}.$$

Les proportions de ce mélange dans la population P sont $p_1 = \frac{1}{2}$ et $p_2 = \frac{1}{2}$. Dans la population P^* elles sont soit les mêmes (proportions inchangées), soit $p_1 = \frac{3}{10}$ et $p_2 = \frac{7}{10}$ (proportions différentes).

Nous présentons les résultats de classification (taux d'erreur, valeur du critère BIC et pourcentage de choix de chaque modèle de discrimination généralisée) obtenus par discrimination généralisée, discrimination classique et classification automatique, pour des données simulées sans bruit, puis avec les quatre type de bruit définis au paragraphe 4.2.5.1.

Pour les tests avec simulations sans bruit, avec le bruit 1 et le bruit 2, les paramètres utilisés pour les transformations entre P et P^* sont :

– paramètres de la transformation 1 (elle respecte les hypothèse du modèle M_1) :

$$A_1 = \begin{pmatrix} 0.5 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix} A_2 = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & -3 \end{pmatrix} b_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} b_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

– paramètres de la transformation 2 (elle respecte les hypothèse du modèle M_2) :

$$A_1 = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix} A_2 = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix} b_1 = \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \\ -1 \end{pmatrix} b_2 = \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \\ -1 \end{pmatrix}$$

– paramètres de la transformation 3 (elle respecte les hypothèse du modèle M_3) :

$$A_1 = \begin{pmatrix} 3 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix} A_2 = \begin{pmatrix} 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.5 \end{pmatrix} b_1 = \begin{pmatrix} 2 \\ 2 \\ 2 \\ 2 \\ 2 \end{pmatrix} b_2 = \begin{pmatrix} 0.6 \\ 0.6 \\ 0.6 \\ 0.6 \\ 0.6 \end{pmatrix}$$

– paramètres de la transformation 4 (elle respecte les hypothèse du modèle M_4) :

$$A_1 = \begin{pmatrix} 0.5 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix} A_2 = \begin{pmatrix} 0.5 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix} b_1 = \begin{pmatrix} -1 \\ -2 \\ -2 \\ -3 \\ -2 \end{pmatrix} b_2 = \begin{pmatrix} -1 \\ -2 \\ -2 \\ -3 \\ -2 \end{pmatrix}$$

Pour les tests avec les bruits 3 et 4, les paramètres seront spécifiés plus loin.

B.2.1. Pas de bruit

		transformations avec proportions							
		non conservées				conservées			
modèle	critère	1	2	3	4	1	2	3	4
	erreur	7.50	14.73	24.63	25.13	**10.93**	14.86	18.20	30.13
M_1	BIC	588.62	629.02	497.03	662.62	599.79	620.28	534.44	636.28
	choix	3.33	13.33	0	13.33	**100**	0	0	3.33
	erreur	33.86	**15.00**	10.43	17.66	32.73	**10.66**	7.83	13.30
M_2	BIC	635.11	626.14	443.53	661.17	671.56	610.18	511.88	617.18
	choix	0	**40**	13.33	6.66	0	**63.33**	46.66	26.66
	erreur	25.56	15.93	12.00	22.70	23.13	10.20	**8.00**	13.96
M_3	BIC	630.12	626.90	443.85	660.96	666.63	610.37	515.15	619.53
	choix	3.33	10	10	3.33	0	23.33	**50**	0
	erreur	20.70	14.33	11.83	17.46	18.8	9.40	7.43	**12.23**
M_4	BIC	633.80	641.57	455.81	653.24	663.08	621.46	523.01	608.00
	choix	3.33	0	0	20	0	0	0	**56.66**
	erreur	**7.50**	14.73	16.66	27.26	10.93	14.86	18.16	28.96
PM_1	BIC	581.24	631.45	468.52	665.00	603.98	620.17	531.11	632.60
	choix	**90**	3.33	0	0	0	0	0	3.33
	erreur	30.10	14.43	**7.7**	18.60	35.90	11.63	8.03	15
PM_2	BIC	635.40	626.32	440.76	653.84	655.58	612.37	515.32	614.34
	choix	0	30	**46.66**	20	0	6.66	3.33	6.66
	erreur	24.83	13.63	7.36	23.80	25.13	11.33	8.06	18.26
PM_3	BIC	630.24	627.95	440.79	655.30	654.47	612.95	515.52	617.19
	choix	0	3.33	30	3.33	0	6.66	0	0
	erreur	22.00	10.10	7.50	**20.83**	28.23	10.93	7.50	15.16
PM_4	BIC	623.11	639.13	454.39	651.61	655.53	625.02	527.18	611.47
	choix	0	0	0	**33.33**	0	0	0	3.33
Discrim.	erreur	21.73	14.73	8.76	28.36	20.03	12.5	13.06	22.06
	BIC	697.27	631.85	512.70	685.15	711.04	592.63	532.00	668.40
Cluster.	erreur	27.26	29.43	15.63	34.16	24.63	25.33	22.20	30.23
erreur optimale		7.74	9.23	6.30	10.51	7.19	8.79	6.56	10.27

TAB. B.1.: N=100, pas de bruit.

		transformations avec proportions							
		non conservées				conservées			
modèle	critère	1	2	3	4	1	2	3	4
	erreur	7.45	13.74	10.69	22.98	**7.58**	11.18	13.13	17.93
M_1	BIC	5863.56	6278.92	4994.59	6565.57	5928.46	6090.47	5375.51	6310.57
	choix	0	0	0	0	**100**	0	0	0
	erreur	39.80	13.68	10.26	15.33	31.85	**9.24**	6.77	12.04
M_2	BIC	6323.18	6251.71	4391.99	6555.22	6634.37	6001.75	5115.39	6108.84
	choix	0	0	0	0	0	**90**	0	0
	erreur	25.69	14.81	10.53	22.10	19.18	9.24	**6.77**	12.10
M_3	BIC	6292.58	6266.54	4373.12	6538.45	6591.59	6006.97	5083.94	6113.17
	choix	0	0	0	0	0	3.33	**100**	0
	erreur	20.27	13.68	11.33	16.71	19.63	9.28	6.72	**11.56**
M_4	BIC	6229.07	6264.48	4400.29	6316.26	6448.58	6020.23	5134.15	5913.85
	choix	0	0	0	0	0	3.33	0	**100**
	erreur	**7.45**	13.74	8.44	22.98	7.58	11.18	13.13	17.74
pM_1	BIC	5749.60	6266.54	4625.29	6570.89	5934.95	6062.53	5297.93	6250.36
	choix	**100**	0	0	0	0	0	0	0
	erreur	39.92	**9.29**	5.97	16.52	32.56	9.29	6.79	13.10
pM_2	BIC	6310.21	6191.49	4319.49	6471.25	5484.27	6007.70	5120.87	6097.11
	choix	0	**96.66**	0	0	0	3.33	0	0
	erreur	28.87	9.52	**5.55**	21.75	26.11	9.26	6.77	13.93
pM_3	BIC	6286.99	6197.32	4293.18	6472.54	6454.44	6012.87	5090.27	6101.49
	choix	0	0	**100**	0	0	0	0	0
	erreur	16.95	9.25	5.98	**12.43**	29.31	9.26	6.72	11.81
pM_4	BIC	6079.45	6209.99	4340.77	6272.39	6388.84	6026.43	5139.50	5919.96
	choix	0	3.33	0	**100**	0	0	0	0
Discrim.	erreur	20.88	13.74	8.44	26.89	16.54	11.18	13.13	19.16
	BIC	7305.54	6299.81	5068.45	6954.34	7294.42	6023.81	5307.96	6652.46
Cluster.	erreur	28.02	26.02	17.03	34.35	31.7	23.39	21.98	30.02
erreur optimale		7.54	9.13	6.69	11.34	7.66	9.12	6.75	11.01

TAB. B.2.: N=1000, pas de bruit.

B.2.2. bruit 1

B.2.2.1. 10% de bruit

		transformations avec proportions							
		non conservées				conservées			
modèle	critère	1	2	3	4	1	2	3	4
	erreur	21.28	34.42	44.61	39.13	**16.70**	23.90	33.20	35.50
M_1	BIC	653.33	682.12	561.87	705.99	626.16	643.58	576.12	659.39
	choix	3.33	3.33	0	13.33	**100**	0	0	0
	erreur	42.01	27.50	19.87	27.98	36.16	**19.56**	**16.66**	21.96
M_2	BIC	692.14	673.05	526.85	704.84	693.04	633.38	560.85	540.07
	choix	0	26.66	20	0	0	**50**	**70**	20
	erreur	34.23	28.17	**20.22**	33.14	26.10	20.33	17.36	22.66
M_3	BIC	689.05	672.72	525.92	704.30	688.99	633.24	562.63	641.37
	choix	0	23.33	**30**	0	0	30	16.66	3.33
	erreur	31.08	26.69	21.98	27.82	27.43	19.13	16.76	**21.56**
M_4	BIC	693.93	685.92	532.82	697.64	683.33	644.87	568.85	632.96
	choix	0	0	3.33	10	0	0	3.33	**50**
	erreur	**21.28**	38.94	30.86	44.32	16.70	21.36	31.53	30.73
pM_1	BIC	645.80	684.61	541.07	708.40	630.11	644.63	573.96	641.37
	choix	**90**	0	3.33	0	0	0	3.33	0
	erreur	36.92	**24.58**	**19.00**	33.42	42.26	21.30	17.96	23.36
pM_2	BIC	691.04	671.94	525.27	694.99	672.48	634.61	563.68	638.05
	choix	0	**33.33**	**30**	30	0	13.33	6.66	10
	erreur	34.80	26.69	19.26	35.25	39.13	21.06	18.40	25.90
pM_3	BIC	688.98	673.93	524.60	697.50	671.94	634.92	565.30	639.85
	choix	0	6.66	13.33	3.33	0	6.66	0	6.66
	erreur	35.67	24.16	19.64	**26.02**	36.56	22.00	17.36	24.33
pM_4	BIC	673.28	681.09	534.92	690.97	668.88	646.25	572.92	635.19
	choix	6.66	6.66	0	**43.33**	0	0	0	10
Discrim.	erreur	32.24	26.92	20.70	36.31	28.06	21.10	22.40	27.90
	BIC	761.44	705.84	575.12	823.49	715.49	633.06	570.13	704.17
Cluster.	erreur	27.53	30.73	20.73	33.84	31.46	31.40	23.76	32.40
erreur optimale		7.29	8.75	6.81	10.57	7.78	9.32	5.91	10.28

TAB. B.3.: N=100, bruit 1, 10% de bruit.

		transformations avec proportions							
		non conservées				conservées			
modèle	critère	1	2	3	4	1	2	3	4
	erreur	20.00	25.46	20.95	33.01	**16.66**	20.69	21.86	26.41
M_1	BIC	6421.50	6771.28	5645.42	7049.06	6098.48	6324.37	5657.25	6492.26
	choix	0	0	0	0	**100**	0	0	0
	erreur	40.83	26.06	18.93	25.13	35.40	**18.37**	16.16	20.50
M_2	BIC	6838.17	6723.93	5250.39	7034.62	6756.43	6200.72	5471.90	6256.94
	choix	0	0	0	0	0	**56.66**	10	0
	erreur	36.08	26.68	19.16	28.36	27.32	18.37	**16.10**	20.52
M_3	BIC	6805.53	6700.22	5237.56	7009.98	6721.93	6203.19	5460.36	6260.29
	choix	0	0	0	0	0	16.66	**73.33**	0
	erreur	30.87	25.24	22.36	27.58	27.78	18.40	16.11	**20.42**
M_4	BIC	6789.57	6718.53	5197.03	6806.18	6566.43	6213.13	5478.02	6109.34
	choix	0	0	0	0	0	3.33	13.33	**56.66**
	erreur	**20.00**	25.46	20.95	33.01	16.66	20.69	21.86	26.67
pM_1	BIC	6344.25	6763.84	5349.72	7054.14	6104.63	6291.98	5597.36	6435.69
	choix	**100**	0	0	0	0	0	0	0
	erreur	39.51	**21.27**	18.25	23.86	33.45	18.53	16.17	21.86
pM_2	BIC	6823.23	6642.87	5196.57	6875.46	6592.94	6202.79	5478.16	6232.26
	choix	0	**60**	3.33	0	0	16.66	0	0
	erreur	35.59	21.59	**18.24**	23.89	37.97	18.47	16.04	21.90
pM_3	BIC	6794.84	6646.46	5185.78	6877.12	6560.93	6206.02	5465.76	6237.19
	choix	0	16.66	**36.66**	0	0	0	3.33	0
	erreur	26.73	21.18	18.81	**23.18**	37.97	18.59	16.11	20.79
pM_4	BIC	6590.38	6647.05	5179.79	6717.29	6479.20	6216.92	5487.55	6237.19
	choix	0	23.33	60	**100**	0	6.66	0	43.33
Discrim.	erreur	31.66	25.46	20.95	36.40	24.83	20.69	21.86	27.27
	BIC	7611.01	6757.52	5704.32	7502.93	7482.82	6415.07	5707.42	6744.98
Cluster.	erreur	30.21	31.81	22.86	33.69	28.16	27.73	28.19	34.00
erreur optimale		7.72	9.22	6.56	11.28	7.65	9.22	6.66	11.28

TAB. B.4.: N=1000, bruit 1, 10% de bruit.

B.2.2.2. 30% de bruit

		transformations avec proportions non conservées				conservées			
modèle	critère	1	2	3	4	1	2	3	4
	erreur	46.22	52.08	52.32	63.42	40.66	40.56	46.93	53.73
M_1	BIC	764.45	788.89	694.35	799.10	666.12	679.81	631.20	690.08
	choix	**33.33**	3.33	0	3.33	**70**	6.66	3.33	0
	erreur	56.78	48.77	41.45	48.57	47.9	38.50	36.26	39.23
M_2	BIC	797.73	775.81	670.56	795.42	721.42	668.27	618.56	681.78
	choix	0	13.3	**33.33**	3.33	0	**26.66**	**60**	6.66
	erreur	53.98	48.00	41.69	56.22	46.43	38.56	36.96	39.53
M_3	BIC	782.67	764.13	669.99	785.08	705.35	665.08	620.04	678.32
	choix	0	**26.66**	13.33	0	0	23.33	16.66	6.66
	erreur	52.26	46.90	42.41	47.73	42.83	37.33	35.43	39.76
M_4	BIC	795.72	780.50	669.55	787.52	698.20	674.80	625.92	673.53
	choix	0	0	**33.33**	0	0	3.33	0	20
	erreur	49.64	54.61	50.41	65.20	40.66	42.36	46.36	52.56
pM_1	BIC	761.58	790.34	683.04	799.23	668.95	680.08	632.29	687.81
	choix	**33.33**	0	3.33	0	6.66	3.33	0	3.33
	erreur	53.42	49.52	42.55	57.17	49.23	41.00	36.86	44.23
pM_2	BIC	780.49	765.92	669.70	771.59	703.55	665.91	621.20	672.11
	choix	3.33	16.66	10	**40**	0	16.66	6.66	**30**
	erreur	52.32	48.89	42.44	56.22	49.73	39.36	37.56	44.60
pM_3	BIC	776.53	762.04	669.60	770.29	689.05	664.38	622.19	670.33
	choix	0	16.66	6.66	16.66	3.33	13.33	6.66	10
	erreur	60.11	45.59	44.13	57.85	49.90	40.73	35.93	42.83
pM_4	BIC	765.16	765.32	672.12	771.34	681.83	671.21	626.33	669.03
	choix	30	23.33	0	36.66	20	6.66	6.66	23.33
Discrim.	erreur	51.51	47.14	42.76	54.16	43.36	38.86	38.93	44.83
	BIC	902.11	802.52	739.18	754.88	778.12	667.11	641.42	721.25
Cluster.	erreur	**29.19**	**29.46**	**22.58**	**31.01**	**28.63**	**28.96**	**29.16**	**31.06**
erreur optimale		7.65	8.63	5.99	10.90	7.29	8.49	6.55	10.83

TAB. B.5.: N=100, bruit 1, 30% de bruit.

		transformations avec proportions							
		non conservées				conservées			
modèle	critère	1	2	3	4	1	2	3	4
	erreur	42.31	46.81	43.00	51.84	35.35	37.90	39.06	44.17
M_1	BIC	7489.63	7794.07	6943.85	7960.42	6554.45	6697.53	6224.37	6839.19
	choix	0	0	0	0	**100**	0	0	0
	erreur	54.75	50.23	41.15	45.65	44.34	36.45	34.99	38.19
M_2	BIC	7840.14	7678.26	6604.55	7902.26	7053.93	6565.87	6082.03	6628.10
	choix	0	0	0	0	0	0	3.33	0
	erreur	53.70	49.80	41.01	54.60	42.60	36.55	35.37	38.20
M_3	BIC	7804.61	7794.38	6592.16	7809.90	7030.33	6540.17	6079.00	6619.12
	choix	0	0	0	0	0	0	3.33	0
	erreur	50.22	46.58	41.55	47.66	44.14	36.43	34.84	38.14
M_4	BIC	7762.31	7611.51	6547.69	7706.48	6836.45	6542.61	6061.86	6497.91
	choix	0	0	**66.66**	0	0	0	0	10
	erreur	42.31	46.81	43.00	51.84	35.35	37.90	39.06	42.69
pM_1	BIC	7445.68	7794.38	6782.64	7965.35	6559.56	679.42	6192.01	6803.46
	choix	**56.66**	0	0	0	0	0	0	0
	erreur	46.02	43.87	41.02	48.83	46.91	37.82	35.13	40.02
pM_2	BIC	7657.15	7514.65	6600.69	7608.39	6778.75	6531.35	6085.83	6535.36
	choix	0	6.66	0	0	0	10	3.33	0
	erreur	50.82	45.37	41.07	48.93	54.09	37.19	35.31	40.15
pM_3	BIC	7638.8	7506.30	6586.85	7610.35	6749.04	6521.36	6083.47	6540.43
	choix	0	6.66	0	0	0	40	0	0
	erreur	54.96	43.38	42.58	46.67	54.22	37.95	39.06	39.60
pM_4	BIC	7445.21	7480.73	6548.95	7511.64	6690.06	6517.78	6045.53	6455.62
	choix	43.33	**86.66**	33.33	**100**	0	**50**	**90**	**90**
Discrim.	erreur	50.65	46.81	43.00	53.96	41.60	37.90	39.06	43.09
	BIC	8379.20	7829.20	7038.95	8205.20	7789.55	6685.10	6200.43	6965.80
Cluster.	erreur	**30.90**	**25.57**	**33.30**	**34.60**	**34.22**	**33.14**	**27.85**	**31.89**
erreur optimale		7.64	9.15	6.49	11.28	7.60	9.00	6.61	11.01

TAB. B.6.: N=1000, bruit 1, 30% de bruit.

B.2.3. bruit 2

B.2.3.1. 10% de bruit

		transformations avec proportions non conservées				conservées			
modèle	critère	1	2	3	4	1	2	3	4
	erreur	**17.64**	23.53	25.08	33.70	**20.13**	21.20	26.23	39.46
M_1	BIC	587.37	645.45	532.24	658.73	608.21	636.94	569.77	628.76
	choix	**56.66**	3.33	0	20	**96.66**	0	0	6.66
	erreur	41.78	**23.87**	15.82	23.23	39.93	**19.80**	17.03	22.53
M_2	BIC	638.59	639.38	507.30	661.76	676.20	622.92	558.47	619.33
	choix	0	**40**	6.66	0	0	**63.33**	26.66	3.33
	erreur	33.06	24.74	15.75	29.22	31.50	19.73	**16.83**	22.96
M_3	BIC	632.81	636.15	498.74	662.02	672.06	628.68	553.41	620.52
	choix	0	36.66	20	6.66	0	10	**63.33**	3.33
	erreur	28.01	22.99	16.32	**24.78**	27.4	18.73	16.50	**22.36**
M_4	BIC	636.57	648.96	513.84	651.43	667.52	632.92	568.30	611.03
	choix	0	6.66	0	**23.33**	0	3.33	6.66	**33.33**
	erreur	20.16	23.53	22.65	34.34	22.80	25.36	26.33	30.66
pM_1	BIC	587.59	649.33	519.26	661.18	611.60	633.74	571.34	622.99
	choix	40	0	3.33	3.33	3.33	0	0	16.66
	erreur	42.15	23.36	15.55	**29.46**	43.46	21.53	17.3	25.8
pM_2	BIC	635.01	641.92	505.13	654.19	658.17	625.76	557.02	616.30
	choix	0	3.33	10	**23.33**	0	10	0	20
	erreur	35.62	25.52	**15.69**	32.52	34.60	21.03	17.40	27.90
pM_3	BIC	630.62	638.91	496.01	655.66	655.96	625.76	557.02	616.30
	choix	0	10	**60**	3.33	0	13.33	3.33	6.66
	erreur	37.03	22.72	15.58	31.01	37.16	20.50	16.73	26.03
pM_4	BIC	630.47	650.19	514.31	650.79	655.96	635.64	572.38	612.72
	choix	3.33	0	0	20	0	0	0	10
Discrim.	erreur	30.40	23.53	17.44	32.82	27.13	21.26	21.80	29.13
	BIC	718.69	653.32	537.01	704.94	770.81	644.84	568.51	744.92
Cluster.	erreur	28.72	29.19	16.83	34.74	26.46	29.30	31.03	32.73
erreur optimale		8.03	9.29	6.34	10.29	7.68	8.80	6.85	9.68

TAB. B.7.: N=100, bruit 2, 10% de bruit.

		transformations avec proportions							
		non conservées				conservées			
modèle	critère	1	2	3	4	1	2	3	4
	erreur	16.91	22.73	17.78	30.99	16.91	19.83	22.09	26.19
M_1	BIC	5914.78	6445.93	5298.70	6563.72	5949.62	6278.94	5660.81	6301.71
	choix	0	0	0	0	26.66	0	0	0
	erreur	43.48	21.98	15.06	22.19	35.22	**18.23**	16.28	11.59
M_2	BIC	6416.03	6388.53	4994.28	6645.46	6716.16	6131.43	5547.67	6131.63
	choix	0	0	0	0	0	**36.66**	0	0
	erreur	33.22	23.89	15.06	26.21	27.31	18.22	16.03	20.62
M_3	BIC	6379.80	6380.79	4884.52	6628.77	6687.02	6208.04	5487.55	6130.78
	choix	0	23.33	0	0	0	0	96.66	0
	erreur	28.41	21.37	15.69	24.70	26.29	18.12	**16.15**	20.33
M_4	BIC	6346.30	6376.47	4957.69	6394.04	6477.84	6144.19	5541.20	5974.73
	choix	0	16.66	0	23.33	0	0	**96.66**	50
	erreur	**16.91**	22.73	17.78	30.99	**16.91**	19.83	22.09	26.19
pM_1	BIC	5872.43	6452.22	5120.29	6559.77	5947.32	6208.04	5644.13	6190.78
	choix	**100**	0	0	0	**73.33**	0	0	0
	erreur	43.25	**19.02**	14.90	24.48	40.39	18.54	16.29	22.46
pM_2	BIC	6363.40	6372.32	4942.31	6532.63	6483.58	6136.18	5492.00	6125.43
	choix	0	**43.33**	0	0	0	26.66	0	0
	erreur	33.15	21.16	**14.88**	27.67	33.67	18.89	16.01	21.68
pM_3	BIC	6341.17	6373.69	4825.87	6536.92	6483.58	6136.18	5492.00	6125.43
	choix	0	0	**100**	0	0	3.33	3.33	0
	erreur	52.20	19.70	15.01	**23.93**	26.66	18.70	16.12	**21.14**
pM_4	BIC	6262.47	6374.51	4926.57	6381.31	6373.90	6132.61	5546.39	5973.64
	choix	0	16.66	0	**76.66**	0	33.33	0	**50**
Discrim.	erreur	29.08	22.73	17.78	33.63	24.94	19.83	22.09	27.18
	BIC	7194.76	6388.11	5315.73	7108.69	7599.17	6325.08	5693.29	6758.33
Cluster.	erreur	27.34	27.82	18.50	35.04	16.91	25.22	25.03	29.56
erreur optimale		7.66	9.21	6.53	11.37	7.71	9.06	6.59	11.59

TAB. B.8.: N=1000, bruit 2, 10% de bruit.

B.2.3.2. 30% de bruit

		transformations avec proportions non conservées				conservées			
modèle	critère	1	2	3	4	1	2	3	4
	erreur	39.16	43.90	43.13	55.46	38.16	51.60	51.70	60.90
M_1	BIC	610.25	674.24	582.28	661.50	610.94	669.82	625.66	633.05
	choix	**93.33**	13.33	16.66	6.66	16.66	0	10	3.33
	erreur	54.53	40.46	33.76	41.03	51.13	38.30	36.60	41.13
M_2	BIC	679.94	673.07	581.10	680.96	679.13	647.35	621.13	666.20
	choix	0	13.33	13.33	0	0	3.33	3.33	0
	erreur	46.26	40.40	34.26	43.03	50.16	38.13	38.60	43.43
M_3	BIC	670.60	670.38	565.06	676.12	671.60	647.51	612.50	642.27
	choix	0	6.66	**70**	0	0	3.33	**46.66**	6.66
	erreur	42.73	39.13	33.60	40.80	44.40	37.50	36.70	39.73
M_4	BIC	667.50	675.72	586.91	665.51	665.24	651.12	628.92	644.34
	choix	0	6.66	0	0	0	0	3.33	6.66
	erreur	39.16	42.43	43.26	54.76	43.90	44.06	51.70	51.40
pM_1	BIC	613.40	670.38	584.46	655.60	607.29	657.36	628.77	620.91
	choix	3.33	0	0	26.66	**80**	0	0	**46.66**
	erreur	53.86	44.33	33.83	50.93	53.33	23.33	36.50	43.70
pM_2	BIC	666.21	666.84	585.11	660.86	648.46	636.60	619.75	650.74
	choix	3.33	**36.66**	0	23.33	3.33	**43.33**	23.33	6.66
	erreur	49.40	43.70	34.40	36.66	49.53	39.46	37.90	46.66
pM_3	BIC	656.81	665.40	568.74	660.95	650.65	635.88	613.41	631.64
	choix	0	16.66	0	6.66	0	23.33	10	10
	erreur	55.96	45.76	33.76	54.43	48.66	40.46	35.76	45.23
pM_4	BIC	650.63	672.10	590.90	653.44	644.31	638.69	627.51	627.08
	choix	0	6.66	0	**36.66**	0	26.66	3.33	20
Discrim.	erreur	45.43	42.06	34.73	49.56	43.36	39.46	39.20	44.96
	BIC	724.08	688.11	587.08	679.33	764.94	703.31	657.29	713.49
Cluster.	erreur	**28.03**	**31.63**	**31.76**	**33.33**	**29.30**	**29.76**	**30.60**	**33.60**
erreur optimale		6.85	9.32	6.55	9.89	8.19	8.47	6.28	10.40

TAB. B.9.: N=100, bruit 2, 30% de bruit.

		transformations avec proportions							
		non conservées				conservées			
modèle	critère	1	2	3	4	1	2	3	4
	erreur	35.38	38.61	37.98	46.34	35.45	37.39	42.55	56.97
M_1	BIC	5998.33	6605.39	5721.27	6501.16	6027.84	6592.72	6170.90	6318.82
	choix	**96.66**	0	0	0	0	0	0	0
	erreur	53.42	37.54	33.79	37.65	47.05	37.22	35.22	40.16
M_2	BIC	6609.89	6633.92	5684.90	6699.98	6741.62	6394.23	6083.91	6287.09
	choix	0	0	3.33	0	0	0	0	0
	erreur	49.28	39.07	33.81	37.84	44.66	37.84	38.21	38.83
M_3	BIC	6578.37	6616.92	5604.74	6623.08	6657.24	6374.23	6057.60	6212.81
	choix	0	0	**93.33**	0	0	0	3.33	0
	erreur	44.78	37.79	33.80	40.71	42.34	36.32	36.02	38.47
M_4	BIC	6486.57	6518.48	5688.67	6469.13	6505.29	6341.16	6073.83	6147.32
	choix	0	0	0	0	0	0	0	0
	erreur	35.38	38.61	37.98	46.34	35.45	37.69	41.49	42.79
pM_1	BIC	5992.62	6586.57	5704.33	6398.46	5455.12	6441.97	6173.24	6081.18
	choix	3.33	0	3.33	23.33	**100**	0	0	3.33
	erreur	54.43	45.31	33.78	44.48	51.99	38.86	34.98	43.01
pM_2	BIC	6422.10	6576.22	5690.83	6526.47	6398.22	6328.03	6037.85	6194.38
	choix	0	0	0	0	0	0	**33.33**	0
	erreur	47.36	42.52	33.83	43.54	49.77	38.32	35.29	40.58
pM_3	BIC	6413.72	6548.08	5610.57	6517.56	6402.52	6312.09	6024.90	6168.84
	choix	0	0	0	0	0	0	**33.33**	0
	erreur	58.58	43.38	33.80	54.68	45.75	37.85	35.25	42.34
pM_4	BIC	6306.76	6474.13	5694.71	6352.46	6254.52	6210.94	6035.12	5999.89
	choix	0	**100**	0	**76.66**	0	**100**	30	**96.66**
Discrim.	erreur	44.61	39.76	35.81	48.07	41.50	37.82	39.89	43.42
	BIC	7493.45	6714.67	5968.43	7400.03	7629.71	6673.24	6178.03	7182.57
Cluster.	erreur	**30.85**	**32.25**	**27.46**	**31.90**	**28.33**	**31.90**	**26.61**	**31.05**
erreur optimale		7.54	9.18	6.71	11.01	7.57	9.08	6.66	11.45

TAB. B.10.: N=1000, bruit 2, 30% de bruit.

B.2.4. bruit 3

Le mélange de variables gaussiennes à l'origine de la simulation des variables binaires a pour centres et matrices de variance :

$$\mu_1 = \begin{pmatrix} -2 \\ -1 \\ -1.5 \\ 1.4 \\ -1.2 \end{pmatrix}, \mu_2 = \begin{pmatrix} 1 \\ 1.5 \\ 2 \\ 2.1 \\ 0.9 \end{pmatrix}, \qquad \Sigma_1 = \Sigma_2 = \begin{pmatrix} 3 & 2.5 & 2.5 & 2.5 & 2.5 \\ 2.5 & 3 & 2.5 & 2.5 & 2.5 \\ 2.5 & 2.5 & 3 & 2.5 & 2.5 \\ 2.5 & 2.5 & 2.5 & 3 & 2.5 \\ 2.5 & 2.5 & 2.5 & 2.5 & 3 \end{pmatrix}.$$

Les paramètres de la transformation entre P et P^* sont :

$$A_1 = A_2 = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}, \qquad b_1 = b_2 = \begin{pmatrix} -4 \\ -4 \\ -4 \\ -4 \\ -4 \end{pmatrix}.$$

Ce jeu de paramètres permet de ne favoriser aucun des six modèles M_2, M_3 et M_4 et leur version avec proportions différentes.
Les résultats sont présentés dans le tableau B.11.

Remarque. *Si nous avions choisi des vecteurs de translation b_1 et b_2 nuls, les modèles M_1 et pM_1 ne serait pas exclus comme c'est le cas ici. Mais alors, la matrice de translation étant proportionnelle à l'identité et positive, la transformation entre les populations P et P^* serait nulle d'un point de vue variables binaires ($P = P^*$). L'analyse discriminante fonctionnerait très bien et la discrimination généralisée n'aurait donc pas d'intérêt.*

		transformations avec proportions non conservées		conservées	
modèle	critère	N=100	N=1000	N=100	N=1000
	erreur	47.00	44.20	40.26	33.23
M_1	BIC	500.55	5263.57	493.73	5025.59
	choix	0	0	0	0
	erreur	37.00	33.40	25.63	25.25
M_2	BIC	476.95	4932.27	444.46	4431.26
	choix	0	0	0	0
	erreur	39.00	38.30	25.93	25.44
M_3	BIC	414.88	4469.81	411.63	4035.35
	choix	0	0	0	0
	erreur	46.00	45.00	32.93	31.72
M_4	BIC	486.72	48925.73	457.96	4430.38
	choix	0	0	0	0
	erreur	47.00	44.20	38.53	32.04
pM_1	BIC	479.80	5045.35	469.74	4658.26
	choix	0	0	0	0
	erreur	37.00	34.50	26.00	25.49
pM_2	BIC	473.62	4873.22	435.18	4302.28
	choix	0	0	0	0
	erreur	**39.00**	38.30	**27.93**	**28.06**
pM_3	BIC	409.68	4371.85	394.09	3830.81
	choix	**100**	**100**	**100**	**100**
	erreur	42.00	38.60	27.33	26.39
pM_4	BIC	476.75	4788.3	443.68	4250.23
	choix	0	0	0	0
Discrim.	erreur	46.00	45.00	32.93	31.72
	BIC	626.96	5992.74	652.51	6030.81
Cluster.	erreur	46.00	**34.00**	33.40	33.55
erreur optimale		12.04	13.79	12.98	13.39

TAB. B.11.: bruit 3.

B.2.5. bruit 4

Le mélange de variables gaussiennes à l'origine de la simulation des variables binaires a pour centres et matrices de variance :

$$\mu_1 = \begin{pmatrix} -2 \\ -1 \\ -1.5 \\ 1.4 \\ -1.2 \end{pmatrix}, \mu_2 = \begin{pmatrix} 1 \\ 1.5 \\ 2 \\ 2.1 \\ 0.9 \end{pmatrix}, \qquad \Sigma_1 = \Sigma_2 = \begin{pmatrix} 3 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix}.$$

Les paramètres de la transformation entre P et P^* sont :

$$A_1 = A_2 = \begin{pmatrix} 2 & 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 & 2 \end{pmatrix}, \qquad b_1 = b_2 = \begin{pmatrix} -6 \\ -6 \\ -6 \\ -6 \\ -6 \end{pmatrix}.$$

Comme pour le bruit 3, ce jeu de paramètres permet de ne favoriser aucun des six modèles M_2, M_3 et M_4 et leur version avec proportions différentes.
Les résultats sont présentés dans le tableau B.12.

		transformations avec proportions			
		non conservées		conservées	
modèle	critère	N=100	N=1000	N=100	N=1000
	erreur	20.96	18.70	37.20	13.19
M_1	BIC	450.27	4459.05	431.52	4282.70
	choix	0	0	0	0
	erreur	18.90	18.52	12.93	11.14
M_2	BIC	531.09	5250.77	512.76	5108.30
	choix	0	0	0	0
	erreur	**16.50**	**16.36**	**11.16**	10.58
M_3	BIC	395.84	3771.86	338.89	3233.19
	choix	**96.66**	**46.66**	**50**	0
	erreur	18.93	18.47	12.26	12.07
M_4	BIC	518.03	5018.72	484.43	4709.34
	choix	0	0	0	0
	erreur	20.96	18.70	34.80	13.19
pM_1	BIC	454.46	4463.79	429.27	4224.73
	choix	0	0	0	0
	erreur	18.90	18.52	14.50	11.24
pM_2	BIC	535.57	5256.46	512.88	5078.39
	choix	0	0	0	0
	erreur	16.13	**16.15**	**11.60**	**10.60**
pM_3	BIC	399.17	3769.89	338.77	3203.85
	choix	3.33	**53.33**	**50**	**100**
	erreur	18.93	18.44	12.53	12.10
pM_4	BIC	522.41	5024.60	483.57	4666.88
	choix	0	0	0	0
Discrim.	erreur	19.26	18.87	14.00	13.28
	BIC	608.25	5531.24	583.08	5758.43
Cluster.	erreur	20.66	18.72	14.63	13.37
erreur optimale		22.49	23.68	23.66	21.97

TAB. B.12.: bruit 4.

B.3. Résultats des tests sur simulations en dimension 10

Les données binaires sont simulées à partir d'une discrétisation d'un mélange de deux gaussiennes de centres et de matrice de variance :

$$\begin{aligned}\mu_1 &= (-2,-1,-1.5,1.4,-1.2,-1.8,1.6,0.5,-1.4,-1.8)',\\ \mu_2 &= (1,1.5,2,2.1,0.9,1.6,-2.3,0.8,0.5,2)',\\ \Sigma_1 &= \Sigma_2 = 3I_{10},\end{aligned}$$

où I_{10} est la matrice identité en dimension 10.
Les proportions de ce mélange dans la population P sont $p_1 = \frac{1}{2}$ et $p_2 = \frac{1}{2}$. Dans la population P^* elles sont soit les mêmes (proportions inchangées), soit $p_1 = \frac{3}{10}$ et $p_2 = \frac{7}{10}$ (proportions différentes).

Nous présentons les résultats de classification (taux d'erreur, valeur du critère BIC et pourcentage de choix de chaque modèle de discrimination généralisée) obtenus par discrimination généralisée, discrimination classique et classification automatique, pour des données simulées sans bruit, puis avec le bruit 1 défini au paragraphe 4.2.5.1.

Pour les tests avec simulations sans bruit et avec le bruit 1, les paramètres utilisés pour les transformations entre P et P^* sont :

- paramètres de la transformation 1 :

$$\begin{aligned}A_1 &= \text{diag}\left(\frac{1}{2},2,3,-2,3,-\frac{1}{2},3,2,-2,\frac{1}{2}\right),\\ A_2 &= \text{diag}\left(2,-2,2,3,-3,-2,3,-\frac{1}{2},3,2\right),\\ b_1 &= b_2 = (0,0,0,0,0,0,0,0,0,0)',\end{aligned}$$

- paramètres de la transformation 2 :

$$\begin{aligned}A_1 &= A_2 = 2I_{10},\\ b_1 &= b_2 = (-1,-1,-1,-1,-1,-1,-1,-1,-1,-1)',\end{aligned}$$

- paramètres de la transformation 3 :

$$\begin{aligned}A_1 &= 2I_{10} \qquad A_2 = \frac{1}{2}I_{10},\\ b_1 &= \left(-\frac{1}{2},-\frac{1}{2},-\frac{1}{2},-\frac{1}{2},-\frac{1}{2},-\frac{1}{2},-\frac{1}{2},-\frac{1}{2},-\frac{1}{2},-\frac{1}{2}\right)',\\ b_2 &= \left(\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2}\right)',\end{aligned}$$

- paramètres de la transformation 4 :

$$\begin{aligned}A_1 &= A_2 = \text{diag}\left(\frac{1}{2},2,3,2,4,\frac{1}{2},3,2,4,2\right),\\ b_1 &= b_2 = (-1,1,2,1,2,-2,3,2,1,-4)'.\end{aligned}$$

B.3.1. Pas de bruit

		transformations avec proportions non conservées				conservées			
modèle	critère	1	2	3	4	1	2	3	4
	erreur	**5.03**	4.90	0.73	18.86	2.13	4.70	2.30	8.23
M_1	BIC	1148.10	1213.08	957.38	1104.52	1135.42	1190.10	1045.36	1077.79
	choix	**96.66**	0	0	0	0	0	0	0
	erreur	14.43	**2.40**	1.00	24.43	16.30	2.93	1.23	18.16
M_2	BIC	1491.60	1185.49	921.94	1298.40	1579.36	1159.74	1026.77	1218.30
	choix	0	**86.66**	0	0	0	0	0	0
	erreur	12.06	2.56	**0.36**	22.60	13.50	3.03	0.90	14.46
M_3	BIC	1423.72	1188.92	879.00	1252.74	1479.21	1161.55	974.53	1191.40
	choix	0	6.66	**100**	0	0	0	0	0
	erreur	38.93	2.40	0.60	**7.06**	44.50	3.00	1.30	6.20
M_4	BIC	1312.74	1205.13	946.43	1017.45	1358.31	1179.42	1046.73	982.96
	choix	0	3.33	0	**93.33**	0	3.33	3.33	3.33
	erreur	6.06	5.06	0.73	19.20	**2.10**	4.70	2.33	8.30
pM_1	BIC	1148.69	1218.03	970.23	1108.78	1130.89	1188.66	1042.34	1075.73
	choix	3.33	3.33	0	0	**100**	10	0	0
	erreur	12.33	2.70	1.00	21.63	14.60	**3.03**	1.23	14.70
pM_2	BIC	1516.81	1195.47	932.64	1300.41	1580.02	1155.39	1022.39	1230.94
	choix	0	0	0	0	0	**63.33**	0	0
	erreur	12.20	2.80	0.46	22.06	13.26	3.10	**0.90**	13.50
pM_3	BIC	1433.47	1198.79	890.83	1250.87	1476.78	1157.20	970.13	1199.32
	choix	0	0	0	0	0	16.66	**96.66**	0
	erreur	35.46	2.96	0.66	6.80	33.16	2.96	1.30	**5.63**
pM_4	BIC	1392.13	1215.14	957.46	1023.84	1436.89	1175.22	1042.45	980.42
	choix	0	0	0	6.66	0	6.66	0	**96.66**
Discrim.	erreur	16.36	5.00	0.73	18.83	14.13	3.70	1.53	12.90
	BIC	1423.03	1290.04	954.58	1341.10	1470.49	1195.52	1029.76	1273.90
Cluster.	erreur	24.10	23.06	10.96	25.33	19.83	24.03	8.96	27.86
erreur optimale		1.84	1.98	2.57	5.10	1.95	2.11	2.30	5.11

TAB. B.13.: N=100, pas de bruit.

		transformations avec proportions non conservées				conservées			
modèle	critère	1	2	3	4	1	2	3	4
	erreur	**2.15**	4.43	0.50	15.57	2.24	3.58	0.73	5.89
M_1	BIC	10966	11938	9468	10965	11185	11630	10021	10666
	choix	**100**	0	0	0	0	0	0	0
	erreur	13.02	**2.40**	0.71	19.68	13.87	2.43	0.55	16.98
M_2	BIC	14600	11593	9061	12843	15579	11307	9900	11928
	choix	0	**93.33**	0	0	0	0	0	0
	erreur	11.42	2.37	**0.35**	19.68	13.21	2.45	0.35	13.63
M_3	BIC	13953	11597	8608	12325	14582	11312	9374	11652
	choix	0	6.66	**100**	0	0	0	0	0
	erreur	30.18	2.42	0.66	**4.69**	45.13	2.46	0.56	4.66
M_4	BIC	12683	11633	9099	9672	13194	11351	9936	9511
	choix	0	0	0	**100**	0	0	0	0
	erreur	2.36	4.55	0.49	16.01	**2.26**	3.59	0.73	5.96
pM_1	BIC	11112	12033	9635	10977	11178	11628	10014	10666
	choix	0	0	0	0	**100**	0	0	0
	erreur	11.10	2.50	0.73	19.09	13.16	**2.43**	0.55	14.52
pM_2	BIC	14892	11738	9206	12844	15583	11300	9894	12054
	choix	0	0	0	0	0	**96.66**	0	0
	erreur	11.65	2.50	0.36	18.98	13.15	2.45	**0.35**	13.03
pM_3	BIC	14093	11743	8763	12321	14580	11305	9367	11751
	choix	0	0	0	0	0	3.33	**100**	0
	erreur	30.87	2.75	0.72	5.71	31.62	2.43	0.57	**4.48**
pM_4	BIC	13502	11773	9245	9787	14048	11344	9930	9505
	choix	0	0	0	0	0	0	0	**100**
Discrim.	erreur	15.19	4.56	0.49	17.14	13.32	3.52	0.73	12.61
	BIC	15343.04	11999	9837	12940	15902	11715	10063	12087
Cluster.	erreur	18.32	29.31	7.08	23.40	23.41	19.59	11.05	28.25
erreur optimale		2.17	2.36	2.66	5.87	2.20	2.37	2.73	5.91

TAB. B.14.: N=1000, pas de bruit.

B.3.2. 10% de bruit 1

		transformations avec proportions non conservées				conservées			
modèle	critère	1	2	3	4	1	2	3	4
	erreur	**20.00**	17.33	19.26	25.76	12.46	15.40	13.80	19.03
M_1	BIC	1259	1322	1111	1226	1196	1261	1131	1142
	choix	**93.33**	0	0	0	0	0	0	0
	erreur	26.28	**15.67**	14.71	32.75	24.96	13.23	11.06	23.96
M_2	BIC	1589	1295	1092	1398	1582	1226	1092	1259
	choix	0	**70**	0	0	0	0	0	0
	erreur	25.03	15.76	**14.61**	31.76	21.40	13.13	10.70	22.30
M_3	BIC	1519	1297	1064	1349	1494	1227	1063	1236
	choix	0	16.66	**93.33**	0	0	0	0	0
	erreur	48.17	15.86	14.77	**18.46**	51.26	13.63	11.03	15.36
M_4	BIC	1390	1308	1102	1168	1366	1241	1109	1083
	choix	0	13.33	0	**96.66**	0	6.66	0	6.66
	erreur	20.96	17.40	19.26	26.12	**12.43**	15.30	15.26	20.36
pM_1	BIC	1258	1327	1121	1226	1192	1259	1113	1140
	choix	6.66	0	0	0	**100**	0	0	0
	erreur	25.12	15.64	14.71	29.96	22.26	**13.10**	11.06	21.73
pM_2	BIC	1613	1307	1096	1399	1583	1222	1088	1266
	choix	0	0	0	0	0	**60**	0	0
	erreur	25.32	15.67	14.61	31.69	21.13	13.00	**10.70**	21.63
pM_3	BIC	1530	1308	1069	1346	1491	1223	1059	1241
	choix	0	0	0	0	0	23.33	**96.66**	0
	erreur	45.38	15.44	15.03	18.07	37.23	13.03	10.96	**14.60**
pM_4	BIC	1477	1323	1106	1185	1442	1238	1105	1080
	choix	0	0	6.66	3.33	0	10	3.33	**93.33**
Discrim.	erreur	27.53	17.50	14.71	28.58	20.96	14.40	11.26	21.03
	BIC	1694	1296	1143	1395	1646	1234	1023	1169
Cluster.	erreur	24.42	27.33	14.29	31.76	23.16	25.46	13.56	30.43
erreur optimale		1.78	2.14	2.56	5.67	2.06	1.84	2.48	5.49

TAB. B.15.: N=100, bruit 1, 10% de bruit.

		transformations avec proportions non conservées				conservées			
modèle	critère	1	2	3	4	1	2	3	4
	erreur	**15.34**	17.33	13.91	22.63	12.06	12.88	10.73	15.33
M_1	BIC	12188	13089	10874	12219	11734	12065	10673	11333
	choix	**100**	0	0	0	0	0	0	0
	erreur	25.53	15.45	14.08	29.04	23.51	12.13	10.62	25.53
M_2	BIC	15463	12737	10597	13815	15688	11784	10595	12367
	choix	0	10	0	0	0	0	0	0
	erreur	23.56	**15.53**	**13.86**	29.45	21.89	12.17	10.40	22.21
M_3	BIC	14716	12724	10329	13242	15688	11784	10273	12047
	choix	0	**76.66**	**100**	0	0	0	0	0
	erreur	38.79	15.47	14.14	**17.50**	50.89	12.25	10.57	**15.07**
M_4	BIC	13510	12744	10589	11347	13361	11815	10610	10528
	choix	0	13.33	0	**100**	0	0	0	**86.66**
	erreur	15.45	17.47	13.91	23.28	**12.03**	12.88	10.77	15.38
pM_1	BIC	12298	13160	10997	12244	11728	12063	10666	11325
	choix	0	0	0	0	**100**	0	0	0
	erreur	23.73	15.64	14.09	28.90	22.74	**12.14**	10.62	22.86
pM_2	BIC	15757	12880	10681	13814	15703	11777	10589	12479
	choix	0	0	0	0	0	**60**	0	0
	erreur	23.73	15.66	13.87	29.12	21.87	12.18	**10.39**	21.81
pM_3	BIC	14832	12859	10413	13238	14677	11778	10268	12130
	choix	0	0	0	0	0	40	**100**	0
	erreur	40.61	15.64	14.31	17.67	42.27	12.18	10.58	14.23
pM_4	BIC	14451	12903	10666	11562	14272	11810	10604	10533
	choix	0	0	0	0	0	0	0	13.33
Discrim.	erreur	27.12	17.42	13.91	27.89	22.04	12.91	10.77	21.49
	BIC	16546	13185	11140	13939	15988	12186	10611	12387
Cluster.	erreur	23.20	25.21	15.45	29.03	17.67	27.13	12.70	30.45
erreur optimale		2.22	2.36	2.73	6.01	2.17	2.41	2.74	5.95

TAB. B.16.: N=1000, bruit 1, 10% de bruit.

B.3.3. 30% de bruit 1

		transformations avec proportions non conservées				conservées			
modèle	critère	1	2	3	4	1	2	3	4
	erreur	44.58	41.54	39.88	48.21	33.83	36.06	31.86	38.03
M_1	BIC	1503	1586	1392	1503	1326	1371	1236	1298
	choix	**66.66**	0	6.66	0	0	0	0	0
	erreur	47.20	39.46	38.12	51.22	41.63	32.43	30.63	40.90
M_2	BIC	1785	1556	1397	1621	1625	1333	1225	1382
	choix	0	0	3.33	0	0	0	0	0
	erreur	44.94	40.02	38.09	51.54	38.73	32.20	30.50	40.43
M_3	BIC	1707	1537	1396	1559	1529	1326	1222	1332
	choix	0	13.33	0	0	0	0	0	0
	erreur	63.51	42.11	40.32	47.94	61.96	36.00	30.86	43.76
M_4	BIC	1552	1512	1370	1434	1389	1318	1222	1243
	choix	3.33	**83.33**	20	**100**	6.66	**33.33**	16.66	**86.66**
	erreur	44.67	41.81	39.82	48.45	33.83	36.66	31.86	37.93
pM_1	BIC	1508	1589	1399	1506	1327	136	1234	1301
	choix	30	0	0	0	**93.33**	10	23.33	10
	erreur	45.77	39.55	38.09	49.82	40.13	32.30	30.66	39.43
pM_2	BIC	1812	1564	1398	1623	1627	1329	1221	1389
	choix	0	0	6.66	0	0	16.66	10	0
	erreur	45.05	40.26	38.06	51.39	38.66	32.20	30.50	40.13
pM_3	BIC	1712	1543	1397	1558	1527	1323	1219	1335
	choix	0	3.33	3.33	0	0	23.33	20	0
	erreur	61.01	40.23	39.76	44.43	55.30	33.30	30.66	37.86
pM_4	BIC	1656	1545	1368	1481	1481	1321	1220	1264
	choix	0	0	**60**	0	0	16.66	**30**	3.33
Discrim.	erreur	48.12	40.41	38.18	49.91	40.63	33.23	30.86	39.16
	BIC	1879	1575	1438	1610	1674	1423	1195	1386
Cluster.	erreur	**28.45**	**32.58**	**23.63**	**32.55**	**30.00**	**25.73**	**24.53**	**32.76**
erreur optimale		1.68	2.01	2.13	5.34	1.85	1.84	2.41	4.59

TAB. B.17.: N=100, bruit 1, 30% de bruit.

		transformations avec proportions non conservées				conservées			
modèle	critère	1	2	3	4	1	2	3	4
	erreur	38.90	40.39	38.03	42.68	31.62	32.36	30.57	34.11
M_1	BIC	14660	15302	13676	14801	12790	13067	11961	12540
	choix	**100**	0	0	0	0	0	0	0
	erreur	45.72	39.04	38.06	47.43	40.50	31.78	30.43	40.99
M_2	BIC	17229	15006	13576	15946	15874	12822	11920	13281
	choix	0	0	0	0	0	0	0	0
	erreur	44.81	39.33	37.98	48.70	39.37	31.88	30.34	39.16
M_3	BIC	16286	14887	13551	15208	14867	12771	11856	12809
	choix	0	0	0	0	0	6.66	6.66	0
	erreur	59.17	40.67	39.18	43.56	61.93	33.27	30.47	38.76
M_4	BIC	15084	14691	13311	13993	13574	12713	11800	12051
	choix	0	**100**	0	**100**	0	**86.66**	**83.33**	**100**
	erreur	38.98	40.46	38.01	43.01	31.63	32.36	30.56	34.19
pM_1	BIC	14713	15334	13740	14818	12784	13065	11955	12543
	choix	0	0	0	0	**100**	0	0	0
	erreur	44.91	39.25	38.06	47.61	39.94	31.81	30.42	38.85
pM_2	BIC	17482	15133	13600	15953	15896	12822	11917	13351
	choix	0	0	0	0	0	0	0	0
	erreur	44.85	39.54	37.98	48.68	39.30	31.87	30.35	39.03
pM_3	BIC	16353	14977	13570	15204	14866	12768	11857	12850
	choix	0	0	0	0	0	6.66	3.33	0
	erreur	54.85	39.12	38.95	40.57	56.10	31.99	30.42	34.29
pM_4	BIC	16183	15000	13305	14472	14530	12750	11808	12171
	choix	0	0	**100**	0	0	0	6.66	0
Discrim.	erreur	47.07	40.43	38.01	48.34	39.48	32.34	30.55	38.53
	BIC	17811	15228	13676	15929	16147	12971	11971	13368
Cluster.	erreur	**31.72**	**27.62**	**16.46**	**24.60**	**21.15**	**28.07**	**19.78**	**31.28**
erreur optimale		2.17	2.44	2.69	6.05	2.15	2.33	2.61	5.93

TAB. B.18.: N=1000, bruit 1, 30% de bruit.

B.4. Applications sur données réelles : risque de second cancer

B.4.1. Choix des variables discriminantes

Les analyses discriminantes et classifications automatiques sont réalisées en découpant l'échantillon d'apprentissage de 149 individus en deux sous-échantillons d'apprentissage et de test.
Le tableau B.19 présente les taux moyens de mauvais classements (sur 300 découpages) évalués sur le sous-échantillon test, ainsi que sur la restriction de ce dernier aux patients à risque. Effectivement, certaines classifications ont tendance à classer quasiment tous les individus comme patients non à risque, sans que cela n'affecte trop le taux de mauvais classement global puisque les individus réellement à risque sont en faible proportion (19 sur 149).
Les variables explicatives sont les suivantes :

- X_1 : état du patient à la date de dernière nouvelle,
- X_2 : nature de la tumeur,
- X_3 : âge du patient au moment du diagnostic,
- X_4 : commune de résidence à la date de dernière nouvelle.
- X_5 : commune de résidence au moment du diagnostic.

	Taux de mauvais classements global (en $\%$)		Taux de mauvais classements sur population à risque (en $\%$)	
Variables	discrimination	classif. auto.	discrimination	classif. auto.
$X_1X_2X_3X_4X_5$	12.82	35.93	64.98	50.17
$X_1X_2X_3X_4$	12.35	32.61	63.31	51.10
$X_1X_2X_3X_5$	12.67	31.72	65.35	51.53
$X_1X_2X_4X_5$	11.94	36.46	77.43	50.26
$X_1X_3X_4X_5$	12.30	34.50	68.45	51.62
$X_2X_3X_4X_5$	16.51	36.98	83.17	53.20
$X_1X_2X_3$	**12.94**	**32.84**	**62.39**	**49.39**
$X_1X_2X_4$	11.91	33.46	80.03	49.47
$X_1X_2X_5$	12.60	31.87	81.27	51.42
$X_1X_3X_4$	13.10	31.39	67.33	48.79
$X_1X_4X_5$	12.59	34.88	84.15	49.84
$X_2X_3X_4$	14.75	37.59	92.03	53.40
$X_2X_3X_5$	14.78	38.07	93.55	48.17
$X_2X_4X_5$	13.61	39.37	98.42	52.37
$X_3X_4X_5$	14.51	38.74	95.85	46.39

TAB. B.19.: Sélection des variables discriminantes.

Le choix des variables discriminantes est porté sur les trois variables X_1, X_2 et X_3, qui sont l'état du patient à la date de dernière nouvelle, la nature de la tumeur et l'âge du patient au moment du diagnostic.

B.4.2. Augmentation du délai d'apparition d'un deuxième cancer

Nous présentons ici les résultats obtenus en augmentant à 15 ans le délai d'apparition d'un second cancer, précédemment supposé à 10 ans.
L'échantillon d'apprentissage est constitué de 93 patients ayant eu un premier cancer du testicule entre 1978 et 1987, parmi lesquels 17 sont décédés sans avoir eu de deuxième cancer, et 12 ont eu un deuxième cancer. L'échantillon test est composé de 206 patients ayant eu un premier cancer entre 1988 et 2002, parmi lesquels 16 sont décédés sans avoir eu de deuxième cancer, et 8 ont eu un deuxième cancer.
Les résultats obtenus pour les différentes méthodes de discrimination sont présentés dans le tableau B.20.

Ici encore, le meilleur modèle est le modèle de discrimination généralisée pM_3, qui classe les 24 patients étiquetés de l'échantillon test avec 29% d'erreur. Le modèle pM_4 donne un taux d'erreur plus faible, mais est trop mauvais pour

deuxième cancer	modèle									
	pM_1	pM_2	pM_3	pM_4	M_1	M_2	M_3	M_4	D.	C.A.
oui	0	0	**37.5**	6.25	0	0	6.25	0	0	56.25
non	100	100	**12.5**	62.5	100	100	62.5	100	100	87.5
total	33.33	33.33	**29.17**	25	33.33	33.33	25	33.33	33.33	66.67

TAB. B.20.: Taux de mauvais classements sur les 24 patients étiquetés de l'échantillon test.

être accepter si on se restreint à la détection des deuxièmes cancers.
De plus, le critère BIC conduit à choisir ce modèle pM_3 (tableau B.21).

modèle							
pM_1	pM_2	pM_3	pM_4	M_1	M_2	M_3	M_4
837.21	841.67	**832.28**	849.35	853.57	855.30	856.52	855.41

TAB. B.21.: Critères BIC pour les huit modèles de discrimination généralisée.

Les résultats de classification sont moins bons que ceux obtenus avec un délai d'apparition fixé à 10 ans, ce qui peut être expliqué par deux causes :

- l'échantillon d'apprentissage est plus petit (93 contre 149 patients),
- les patients étiquetés de l'échantillon test sur lesquels sont évalués les taux d'erreur sont plus nombreux.

En utilisant le modèle pM_3 de discrimination généralisée, nous détectons parmi les 206 patients de l'échantillon test 82 patients à fort risque de deuxième cancer.
Après analyse de ces 82 patients, on constate qu'ils ont eu un premier cancer du testicule à tumeurs germinales séminomateuses à un âge avancé. Ceci coïncide avec les conclusions obtenues avec un délai d'apparition de deuxième cancer de 10 ans. Par contre, on rencontre les deux modalités de la variable état dans ces 86 patients, ce qui est impossible (cf. remarque 4.3.1) et remet en cause l'utilisation de cette variable. En omettant cette variable état, la conclusion est donc semblable à celle obtenue avec un délai d'apparition de deuxième cancer de 10 ans. Précisons tout de même que nous avions mis en doute l'utilisation de la variable état dans la remarque 4.3.1, mais que nous avions décider de la conserver néanmoins puisque les conclusions obtenues au paragraphe 4.3.2.2 étaient cohérentes.
En conclusion, se restreindre à un échantillon d'apprentissage constitué des patients ayant eu un premier cancer entre 1978 et 1987 conduit à la même conclusion qu'en incluant les patients ayant eu un premier cancer entre 1988 et 1992, si on prend le soin de ne plus considérer la variable état.

Bibliographie

[1] Akaike, N. (1973). *Information Theory as an Extension of the Maximum Likelihood Principle.* In B.Petrov and F. Csaki, editors, Second International Symposium on Information Theory, pages 267-281, Budapest, Akademiai Kiado.

[2] Akaike, N. (1974). *A new look at the statistical model identification.* IEEE Transactions on AUtomatic Control, AC-19, 716-723.

[3] Anderson, J.A. (1972). *Separate sample logistic discrimination.* Biometrika, 59, 19-35.

[4] Baccini, A., Caussinus, H. and Ruiz-Gazen, A. (2001). *Apprentissage progressif en analyse discriminante.* Revue de Statistique Appliquée, XLIX(4), 87-99.

[5] Banfield, J.D. and Raftery, A.E. (1993). *Model-based Gaussian and non-Gaussian clustering.* Biometrics, 49, 803-821.

[6] Bartholomew, D.J. and Knott, M. (1999). *Latent Variable Models and Factor Analysis.* Second edition, Arnold.

[7] Biernacki, C. and Beninel,F. *Apprentissage sur une sous population et prédiction sur une autre : une extension à (et de) la discrimination logistique.* Colloque Data Mining et Apprentissage Statistique. Applications en Assurance, du 12 au 13 mai 2005, Niort, France.

[8] Biernacki, C. (2004). *Contributions à l'étude des mélanges de lois et à leurs applications.* Mémoire d'Habilitation à Diriger des Recherches, Université de Franche-Comté.

[9] Biernacki, C., Beninel, F. and Bretagnolle, V. (2002). *A generalized discriminant rule when training population and test population differ on their descriptive parameters.* Biometrics, 58, 2, 387-397.

[10] Bonnans, J.F., Gilbert, J.C., Lemarechal, C. and Sagastizabal, C.A. (2003). *Numerical Optimization : Theoretical and Practical Aspects.* Springer.

[11] Celeux, G. and Govaert, G. (1991). *Clustering criteria for discrete data and latent class models.* Journal of classification, 8, 157-176.

[12] Celeux, G. and Govaert, G. (1995). *Gaussian parcimonious models.* Pattern Recognition, 28 (5), 781-793.

[13] Celeux, G. (2003). *Analyse discriminante.* Dans *L'analyse des données,* G. Govaert éditeur, Hermes Science, Paris, France.

[14] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). *Maximum likelihood from incomplete data (with discussion).* Journal of the Royal Statistical Society, Series B 39, 1-38.

[15] Everitt, B.S. (1984). *An introduction to latent variables models*. London : Chapman & Hall.

[16] Everitt, B.S. (1987). *A Finite Mixture Model for the Clustering of Mixed-Mode Data*. Statistics and Probability Letters 6, 305-309.

[17] Fisher, R.A. (1936). *The use of multiple measurements in taxonomic problems.* Annals of Eugenics, 7, 179-188, Pt. II.

[18] Fix, A. and Hodges, J.L. (1951). *Discriminatory analysis - non parametric discrimination : Consistency properties.* Technical Report, report of the U.S.A.F. School of Aviation Medicine, Agrawala (1977).

[19] Friedman, J.H. and Stuetzle, W. (1981). *Projection pursuit regression.* Journal of the American Statistical Association, 76, 817-823.

[20] Hand, D.J. (1982). *Kernel discriminant analysis.* Research studies press, Wiley, New-York.

[21] Jacques, J. and Biernacki, C. *Analyse discriminante généralisée : cas des données binaires avec modèle des classes latentes.* Colloque Data Mining et Apprentissage Statistique. Applications en Assurance, du 12 au 13 mai 2005, Niort, France.

[22] Lazarsfeld, P.F. and Henry, N.W. (1968). *Latent structure analysis*. Houghton Mifflin, Boston.

[23] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953) *Equations of state calculations by fast computing machines.* J. Chem. Phys. 21, 1087-1092.

[24] Murata, N., Yoshizawa, S. and Amari, S. (1991). *A criterion for determining the number of parameters in an artificial neural network model.* In T.Hohonen, K.Mäkisara, O.Simula and J.Hangas, editors, Artificial Neual Networks. Proceedings of ICANN-91, volume 1, pages 9-14, Amsterdam : North Holland.

[25] Murata, N., Yoshizawa, S. and Amari, S. (1993). *Learning curves, model selection and complexity of neural networks.* In NIPS5, pages 607-614.

[26] Murata, N., Yoshizawa, S. and Amari, S. (1994). *Network Information Criterion - determining the number of hidden units for artificial neural networks models.* IEEE Transactions on Neural Networks, 5, 865-872.

[27] Rao, C.R. (1948). *The utilization of multiple measurements in problems of biological classification (with discussion)*. Journal of the Royal Statistical Society, Series B, 10, 159-203.

[28] Robert, C. (1996). *Méthodes de Monte Carlo par Chaînes de Markov.* Economica, Paris.

[29] Saporta, G. (1990). *Probabilités, analyse des données et statistique.* Technip, Paris.

[30] Schwarz, G. (1978). *Estimating the dimension of a model*. Annals of Statistics, 6, 461-464.

[31] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. London : Chapman & Hall.

[32] Thibault, J-.C., Bretagnolle, V. and Rabouam, C. (1997). Cory's shearwater calonectris diomedia. Birds of Western Paleartic Update, 1, 75-98.

[33] Thurstone, L.L. (1927). *A law of comparative judgement* Amer. J. Psychol., 38, 368-389.

[34] Tomassone, R., Danzard, M., Daudin, J.J. and Masson, J.P. (1988). *Discrimination et classement*. Masson, Paris.

[35] Van Franeker, J.A. and Ter Brack, C.J.F. (1993). *A generalized discriminant for sexing fulmarine petrels from external measurements.* The Auk, 110(3), 492-502.

Conclusions et Perspectives

Nous concluons cette thèse comme nous l'avons commencée, en séparant les deux sujets abordés que sont l'analyse de sensibilité et l'analyse discriminante généralisée.

Analyse de sensibilité

Les méthodes d'analyse de sensibilité sont de plus en plus utilisées de nos jours dans tout processus de modélisation, que ce soit pour améliorer la prédiction du modèle, pour alléger le modèle ou encore pour mieux comprendre le phénomène étudié en analysant les différentes interactions entre variables.
Dans ce contexte, cette thèse a eu pour objectif d'étudier l'impact de deux sources d'incertitudes particulières, liées à la problématique générale de l'incertitude de modèle, en répondant aux questions suivantes :

- Quelles conséquences une mutation du modèle étudié a-t-elle sur l'analyse de sensibilité de ce dernier ?
- Quelle est l'influence sur les résultats d'une analyse de sensibilité de l'utilisation d'un modèle simplifié à la place d'un modèle de référence trop complexe ?

L'analyse de ces deux problématiques a fait naître un troisième objectif :

- étudier le problème de l'analyse de sensibilité pour des modèles à variables d'entrée non indépendantes.

Le premier travail effectué fût de synthétiser les méthodes et travaux existants en analyse de sensibilité. Nous avons comparé et testé les différentes approches de la sensibilité, et sélectionné une définition des indices de sensibilité ne nécessitant aucune hypothèse sur le modèle (indices définis par décomposition de la variance, introduits par Sobol et Saltelli). Une approche alternative utilisant des modèles additifs a aussi été proposée et testée, mais il faut être conscient de ses limites lors des applications (comme pour toutes les autres méthodes). La méthode d'estimation des indices de sensibilité que nous avons choisie est celle de Sobol, basée sur l'estimation d'intégrales par simulations de Monte-Carlo. Le second travail fût d'analyser la problématique de l'incertitude de modèle, en réalisant notamment une étude bibliographique sur le sujet. Nous n'avons pas trouvé de travaux se rapportant à notre double problématique. Tout ceci a été présenté dans la première partie de ce mémoire.
La méthodologie employée pour étudier l'impact sur une analyse de sensibilité, d'une mutation du modèle étudié, a consisté en une étude cas par cas. Nous avons déterminé une liste des mutations de modèles possibles, pour chacune desquelles nous avons établi les relations formelles entre les indices de sensibilité avant mutation et ceux après mutation. Ainsi, sous certaines conditions, nous avons montré pour quels cas il était possible de déduire les valeurs des indices de sensibilité du modèle après mutation, à partir de ceux avant mutation, avec un minimum de calculs supplémentaires. Nous nous sommes ensuite servis de ces résultats pour apporter quelques éléments de réponses au problème de l'impact de l'utilisation d'un modèle simplifié. En caractérisant une fonction écart entre le modèle simplifié et le modèle de référence, nous avons déterminé sous quelles conditions l'approximation faite en utilisant le modèle simplifié était correcte, et dans le cas contraire comment cette approximation pouvait être améliorée à moindre coût. Ces travaux, ainsi

que quelques applications illustratives sur un logiciel d'impact environnemental de rejets gazeux chroniques d'une installation nucléaire, ont été présentés dans la deuxième partie de ce mémoire.
Enfin, nous nous sommes aperçus au cours de cette thèse, de l'importance du problème des modèles à variables d'entrée non indépendantes. Les recherches menées ont abouti à la conclusion qu'il n'y a pas de sens de considérer la notion de sensibilité de la sortie d'un modèle à une variable d'entrée indépendamment des autres, si celle-ci est corrélée ou dépendante d'autres entrées. Nous avons donc proposé une approche multidimensionnelle, en exprimant la sensibilité pour des groupes de variables corrélées ou dépendantes. La méthode d'estimation de Sobol peut être généralisée à l'estimation d'indices multidimensionnels. Des applications dans le domaine de l'ingénierie nucléaire ont montré l'intérêt de cette méthode. Ceci a constitué finalement la troisième partie de ce mémoire.

L'estimation des indices de sensibilité, dans le cas unidimensionnel classique ainsi que dans le cas multidimensionnel, demande pour des modèles complexes un temps de calcul important. Les plans d'échantillonnage utilisés peuvent permettre une accélération de la convergence de ces estimations. Jusqu'à présent, les méthodes de Quasi-Monte Carlo sont les plus efficaces. Les dernières innovations dans ce domaine de type Quasi-Monte Carlo Randomisé, devraient apporter à l'analyse de sensibilité un bienfait novateur.
L'approche fonctionnelle de l'analyse de sensibilité, proposée par Antoniadis, est un axe de recherche porteur d'espoir. En effet, les atouts de l'analyse de la variance fonctionnelle devraient permettre une estimation de la sensibilité du modèle nécessitant un nombre de simulations de ce dernier moins important que dans le cas classique. En outre, l'écriture du modèle dans une base de fonction orthogonale, apporte un cadre idéal à la problématique des mutations de modèles, et plus généralement à celle de l'incertitude de modèle.

Analyse discriminante généralisée

L'analyse discriminante généralisée est une extension de l'analyse discriminante classique qui permet de traiter le cas où les échantillons d'apprentissage et de test ne sont pas issus d'une même population. Notre contribution étend l'analyse discriminante généralisée, définie initialement dans un cadre gaussien, au cas des données binaires. Le principal défi fût d'établir une relation entre les variables binaires des deux populations d'apprentissage et de test. Ainsi, nous avons obtenu une règle de discrimination avec un nombre réduit de paramètres à estimer.
Un ensemble de tests sur simulations a permis de montrer la supériorité de notre approche vis-à-vis de la discrimination classique et de la classification automatique, dans le cas où les populations d'apprentissage et de test sont différentes. Enfin, deux applications ont été traitées, un première dans un contexte biologique puis une seconde en santé publique, où notre méthode a permis de détecter parmi un ensemble de patients ayant eu un premier cancer ceux pour lesquels le risque de survenue de deuxième cancer est important.

Les perspectives de ces travaux sont nombreuses. Tout d'abord, nous avons vu que la relation entre les variables binaires des deux populations d'apprentissage et de test était définie à partir de la fonction de répartition de la loi normale centrée réduite. Il peut être intéressant de tester d'autres fonctions de répartition. En effet, nous avons vu qu'en s'éloigant trop de l'hypothèse que les données binaires étaient dues à une discrétisation de variables gaussiennes, la discrimination généralisée pouvait être en difficulté. En changeant la fonction de répartition utilisée, on peut espérer s'adapter à ce type de données. Il faudra bien sûr être vigilant sur la justification théorique de cette approche car la transformation linéaire stochastique était garantie par l'hypothèse normale.
Nos travaux ont permis d'élargir le domaine d'application de l'analyse discriminante généralisée du cas gaussien au cas binaire. Afin de pouvoir traiter un maximum de cas pratiques, il serait très intéressant de pouvoir traiter les variables catégorielles, *i.e.* à plus de deux modalités. Pour ce faire, une solution peut être d'utiliser le même artifice en supposant que les variables catégorielles sont issues d'une discrétisation de variables continues. Par contre, contrairement au cas binaire où le choix du seuil de discrétisation n'a pas

d'influence sur les modèles de discrimination généralisée, ce choix est problématique dans le cas catégoriel, puisque l'on a alors plusieurs seuils par variable à définir. Les travaux d'Everitt [16] qui définissent un algorithme de classification pour données mélangées (binaires, catégorielles et continues) peuvent être d'une aide intéressante pour traiter ce problème.
S'il est possible d'adapter la discrimination généralisée au cas catégoriel, l'objectif suivant sera alors de traiter les problèmes où l'on rencontre, comme dans les travaux d'Everitt, les trois sortes de variables : binaires, catégorielles et continues.
Toujours dans la même optique de s'adapter à un maximum de cas pratiques différents, on peut se poser la question de comment réaliser une classification lorsque les individus des populations d'apprentissage et de test ne sont décrits par le même type de variables ? Par exemple, on peut imaginer que les variables de la population d'apprentissage sont continues et que celles de la population test sont binaires, tout en exprimant la même caractéristique.
Enfin, la généralisation au cas de populations d'apprentissage et de test différentes n'a été étudiée que dans le cadre de l'analyse discriminante multi-normale. Elle peut néanmoins être envisagée pour la discrimination logistique [7] ou encore pour des méthodes de discrimination non paramétriques.

CONTRIBUTIONS À L'ANALYSE DE SENSIBILITÉ ET À L'ANALYSE DISCRIMINANTE GÉNÉRALISÉE

Résumé :
Deux thèmes sont abordés dans cette thèse : l'analyse de sensibilité et l'analyse discriminante généralisée.
L'**analyse de sensibilité** globale d'un modèle mathématique étudie comment les variables de sortie de ce dernier réagissent à des perturbations de ses entrées. Les méthodes basées sur l'étude de la variance quantifient les parts de variance de la réponse du modèle dues à chaque variable d'entrée et chaque sous-ensemble de variables d'entrée. Le premier problème abordé est l'impact d'une incertitude de modèle sur les résultats d'une analyse de sensibilité. Deux formes particulières d'incertitude sont étudiées : celle due à une mutation du modèle de référence, et celle due à l'utilisation d'un modèle simplifié à la place du modèle de référence. Un second problème relatif à l'analyse de sensibilité a été étudié au cours de cette thèse, celui des modèles à entrées corrélées. En effet, les indices de sensibilité classiques n'ayant pas de signification (d'un point de vue interprétation) en présence de corrélation des entrées, nous proposons une approche multidimensionnelle consistant à exprimer la sensibilité de la sortie du modèle à des groupes de variables corrélées. Des applications dans le domaine de l'ingénierie nucléaire illustrent ces travaux.
L'**analyse discriminante généralisée** consiste à classer les individus d'un échantillon test en groupes, en utilisant l'information contenue dans un échantillon d'apprentissage, lorsque ces deux échantillons ne sont pas issus d'une même population. Ce travail étend les méthodes existantes dans un cadre gaussien au cas des données binaires. Une application en santé publique illustre l'utilité des modèles de discrimination généralisée ainsi définis.
Mots Clés :
Analyse de sensibilité globale, incertitude de modèle, mutation de modèle, modèle à entrées statistiquement dépendantes, indice de sensibilité multidimensionnel. Analyse discriminante généralisée, données binaires, populations d'apprentissage et de test différentes, algorithme EM.

CONTRIBUTIONS TO SENSITIVITY ANALYSIS AND TO GENERALIZED DISCRIMINANT ANALYSIS

Abstract :
Two topics are studied in this thesis : sensitivity analysis and generalized discriminant analysis.
Global **sensitivity analysis** of a mathematical model studies how the output variables of this last react to variations of its inputs. The methods based on the study of the variance quantify the part of variance of the response of the model due to each input variable and each subset of input variables. The first subject of this thesis is the impact of a model uncertainty on results of a sensitivity analysis. Two particular forms of uncertainty are studied : that due to a change of the model of reference, and that due to the use of a simplified model with the place of the model of reference. A second problem was studied during this thesis, that of models with correlated inputs. Indeed, classical sensitivity indices not having significance (from an interpretation point of view) in the presence of correlation of the inputs, we propose a multidimensional approach consisting in expressing the sensitivity of the output of the model to groups of correlated variables. Applications in the field of nuclear engineering illustrate this work.
Generalized discriminant analysis consists in classifying the individuals of a test sample in groups, by using information contained in a training sample, when these two samples do not come from the same population. This work extends existing methods in a Gaussian context to the case of binary data. An application in public health illustrates the utility of generalized discrimination models thus defined.
Keywords :
Global sensitivity analysis, model uncertainty, model mutation, model with statistically dependent inputs, multidimensional sensitivity index. Generalized discriminant analysis, binary data, different training and test populations, EM algorithm.