HAL
open science

Modélisation de la production d'hydrocarbures dans un bassin pétrolier

Bertrand Michel

To cite this version:

Bertrand Michel. Modélisation de la production d'hydrocarbures dans un bassin pétrolier. Mathématiques [math]. Université Paris Sud - Paris XI, 2008. Français. NNT : . tel-00345753

HAL Id: tel-00345753

https://theses.hal.science/tel-00345753

Submitted on 9 Dec 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° d'ordre: 9121

UNIVERSITE PARIS-SUD
FACULTE DES SCIENCES D'ORSAY

THESE

Présentée pour obtenir

LE GRADE DE DOCTEUR EN SCIENCES
DE L'UNIVERSITE PARIS XI

Spécialité : Mathématiques

par

Bertrand MICHEL

MODÉLISATION DE LA PRODUCTION D'HYDROCARBURES DANS UN BASSIN PÉTROLIER

Soutenue le 25 septembre 2008 devant la commission d'examen :

M.	Gérard Biau	(Rapporteur)
M.	Christophe Biernacki	(Rapporteur)
M.	Pascal Massart	(Directeur de thèse)
M.	Thomas Duquesne	(Directeur de thèse)
M.	Gilles Celeux	(Président du jury)
M.	Denis Babusiaux	(Invité du jury)
Mme	Nathalie Alazard-Toux	(Invité du jury)

Remerciements

Après avoir soutenu sa thèse de doctorat sur l'estimation des réserves pétrolières, Vincent Lepez dessine les contours d'un sujet de thèse qui prolongerait le sien, sujet qui deviendra mon pain quotidien. Je tiens donc à saluer Vincent, c'est bien grâce à lui que je me suis lancé dans cette aventure pétrolière. Je le remercie aussi pour le soutien qu'il m'a apporté au début de ma thèse.

Je remercie très chaleureusement mes deux directeurs de thèse Pascal Massart et Thomas Duquesne. Toutes les discussions et les conseils dont j'ai pu profiter pendant ces quelques années m'ont permis de progressivement m'initier à la recherche. La bonne ambiance de ces réunions ne fut pas négligeable et m'a incité à persévérer dans les moments de doute, certains matins pluvieux d'hiver où l'on préférerait être dans le Risoux à approfondir le pas de un.

Il est parfois difficile dans le déroulement d'une thèse de naviguer entre le milieu universitaire et un milieu plus industriel tel que celui de l'Institut Français de Pétrole. Pour ma part, je tiens à saluer les membres de la Direction des Études Économiques de l'IFP, ils m'ont offert les meilleures conditions de travail possibles. Je remercie en particulier Denis Babusiaux, Nathalie Alazard, Armelle Sanière et Yves Mathieu pour leur disponibilité et leur confiance.

Je remercie Gérard Biau et Christophe Biernacki qui m'ont fait l'honneur de bien vouloir rapporter cette thèse.

Je tiens aussi à remercier tout particulièrement Cathy avec qui j'ai beaucoup travaillé ces deux dernières années. Il me semble que nous nous sommes bien épaulés et que nous sommes parvenus à progresser ensemble pendant cette collaboration. J'en suis très heureux, d'autant plus qu'il nous reste encore beaucoup de travail !

Ce manuscrit a été grandement amélioré par les remarques de relecture de Cathy, Nicolas et Sylvain. Vos conseils et nos discussions propos de ce qui suit, mais aussi sur nombreux autres sujets, m'auront été précieux. Un grand merci à toi, Amélie, pour avoir accepté de passer au détecteur de fautes de français ce pavé rempli de signes obscurs...

Mes remerciements vont aussi à Gilles Celeux, Jean-Michel Marin, Patrice Assouad, Patricia Raynaud qui ont accepté de prendre de leur temps pour répondre à mes questions et discuter avec moi.

Merci à tous les doctorants de Rueil-Malmaison et d'Orsay, je ne ferai pas de liste : un grand merci à tous pour votre bonne humeur et votre soutien !

J'ai eu la chance de conserver de très bons amis de mes années d'études à Orsay. Vous m'avez soutenu pendant ce long périple, un grand merci pour votre présence !

À mes parents pour leur attention discrète mais bien réelle, à mes frères Olivier et Renaud embarqués dans la même aventure, courage, le col n'est plus très loin !

À Amélie, pour tout ce qu'il reste devant nous.

Table des matières

Introduction **3**

1 Problématiques de la modélisation de l'activité pétrolière 3
2 Modélisation probabiliste de la formation des réserves et de l'exploration pétrolière 4
3 Modélisation de la production pétrolière et position du problème statistique . 7
4 Sélection de modèles pour l'estimation de densité 10
5 Heuristique de pente 15
6 Application à la production pétrolière 16
7 Éléments de conclusion et perspectives 17

1 Contexte Pétrolier **19**

1.1 Les réserves pétrolières 19
1.2 Exploration et production d'un bassin pétrolier 23
1.3 La courbe de Hubbert : une controverse 25
1.4 Le pic de production mondial 31
1.5 Objectifs et hypothèses générales de cette thèse 34

2 Modèle probabiliste pour la formation des réserves **37**

2.1 Modélisation de la formation des réserves pétrolières 37
2.2 Rappels mathématiques sur les partitions stables 41
2.3 Modélisation de la formation des réserves par le modèle de Bolthausen-Sznitman 48
2.4 Conclusions du chapitre 57

3 Modèle probabiliste pour l'exploration pétrolière **59**

3.1 Modélisation du processus de forage 59
3.2 Tirage proportionnel à la taille 62
3.3 Dynamique asymptotique de l'exploration 65
3.4 Conclusions du chapitre 67
3.5 Preuves 67

4 Modélisation de la production pétrolière et position du problème statistique **71**

4.1 Introduction 71
4.2 Réserves pétrolières 72

4.3 Exploration pétrolière 73
4.4 Production individuelle des gisements pétroliers 80
4.5 Politique de mise en production des champs 85
4.6 Synthèse des problématiques statistiques relevées 86

5 Sélection de modèle pour l'estimation de densité 87
5.1 Rappels sur la sélection de modèles pour l'estimation de densité 87
5.2 Sélection de modèles de mélanges gaussiens 92
5.3 Sélection d'un modèle d'exploration pétrolière 116

6 Heuristique de pente et applications 125
6.1 La méthode de la pente 125
6.2 Classification non supervisée et sélection de variables simultanés 127
6.3 Sélection d'un modèle d'exploration pétrolière 137
6.4 Conclusion du chapitre 144

7 Profils de production d'un bassin pétrolier 145
7.1 Modèle de production individuelle des gisements pétroliers 145
7.2 Profils de production de bassin pétrolier 152
7.3 Prolongement de l'exploration de la production d'un bassin 160

Appendices 175

A Bassins étudiés 175

B Rappels sur les processus de Poisson 179
B.1 Mesures de Poisson et processus ponctuels de Poisson 179
B.2 Processus de Lévy et subordinateurs stables 180

C Étude par simulations du critère pénalisé pour la sélection de modèles de mélanges gaussiens 183
C.1 Notations 183
C.2 Pénalités minimales 184
C.3 Comparaison à l'oracle et à d'autres critères 185
C.4 Base de données waveforms 187

D Annexe pour l'étude du processus d'exploration 193
D.1 Simulations 193
D.2 Validation 199

Bibliographie 202

Introduction

Cette thèse a pour objet la modélisation de la production pétrolière dans un bassin d'hydrocarbures. Nous proposons un modèle complet de la production dans un bassin qui s'appuie sur une description probabiliste des réserves, sur l'exploration des hydrocarbures et sur la mise en production des gisements découverts. Pour chacun des ces éléments, une modélisation et des procédures statistiques associées sont élaborées de façon à disposer d'un modèle utilisable en pratique pour étudier la forme des profils de production agrégés à l'échelle d'un bassin d'hydrocarbures.

1 Problématiques de la modélisation de l'activité pétrolière

Les réserves d'hydrocarbures se sont formées il y a plusieurs millions d'années dans des bassins sédimentaires à la suite d'un long processus de maturation, elles constituent aujourd'hui la première source d'énergie utilisée par l'humanité. Le caractère épuisable des sources d'énergie que sont le pétrole et le gaz fait de l'estimation des réserves d'hydrocarbures et de l'analyse de leur production deux questions fondamentales pour les politiques énergétiques nationales et mondiale. Le problème de l'estimation des ressources d'hydrocarbures intéressait déjà l'industrie pétrolière et les économistes au début du siècle. Dans une thèse soutenue récemment par Lepez (2002), une méthode originale est proposée pour l'estimation des réserves dans un bassin d'hydrocarbures en cours d'exploitation. Les résultats de Lepez constituent le point de départ de nos travaux personnels, ce qui a de plus permis de prolonger cette collaboration entre l'Institut Français du Pétrole et l'Université Paris Sud 11.

La modélisation de la production d'hydrocarbures est un problème qui se situe en aval de la problématique de l'estimation des réserves. Aujourd'hui, certains bassins pétroliers matures[1] atteignent, ou sont sur le point d'atteindre leur maximum de production. De nombreux économistes de l'énergie et industriels du secteur pétrolier s'interrogent sur la forme des courbes de production et en particulier les questions suivantes ne font pas l'objet d'un consensus :

- La courbe de production est-elle symétrique ?
- Quels sont les facteurs qui déterminent le plus la forme de la courbe de production du bassin ?
- Est-il possible de limiter le déclin de la production une fois franchi le maximum de la production ?

Il est possible de distinguer deux catégories de modèles pour la production d'hydrocarbures.

[1] C'est-à-dire produisant depuis plusieurs décennies.

La première "école", dans la continuité des travaux de Hubbert (1956), repose sur des ajustements graphiques effectués sur les courbes de production pétrolière. Pour les adeptes de Hubbert, la courbe de production est symétrique, et il existerait une forme générale pour les courbes de production de bassin. Une deuxième catégorie de travaux propose des modélisations avec un point de vue économétrique en expliquant les niveaux de production par de nombreuses variables explicatives telles que le prix du baril de brut ou les investissements effectués. Cette fois, plus aucune hypothèse n'est supposée sur la forme des courbes de production. Dans tous les cas, les travaux de modélisation de ces deux écoles ne s'intéressent jamais à la répartition du pétrole à l'intérieur de la population des gisements du bassin. Pourtant, la variable "taille" des gisements est de première importance pour comprendre comment sont découverts puis produits les gisements du bassin. La production pétrolière, c'est-à-dire l'extraction hors des gisements des ressources d'hydrocarbures qu'ils contiennent ne peut être comparée avec d'autres activités minières telles que la production du charbon ou de l'uranium. Les accumulations d'hydrocarbures ne sont pas réparties de façon diffuse dans le sous-sol du bassin, elles sont concentrées dans des structures géologiques appelées "pièges" qui ne forment pas une population homogène en taille. Dans ce travail de thèse, nous proposons de modéliser la production de pétrole d'un bassin en considérant avant toute chose la distribution des tailles de ses gisements et en respectant le principe naturel selon lequel seuls les gisements découverts peuvent être mis en production. Selon notre point de vue, la production pétrolière est le produit de la répartition des réserves dans le sous-sol, d'une campagne d'exploration dans le bassin, et d'une politique de mise en production des gisements découverts. Dans les chapitres qui suivent, nous proposons des modélisations et des procédures statistiques pour chacun de ces éléments. Notons que ces contributions ont un intérêt propre, au-delà de leur implication dans notre modèle de production pétrolière. Le modèle de production obtenu permet d'apporter des éléments de réponse aux trois questions exposées plus haut, il permet aussi de proposer des prolongements pour des courbes de production de bassins suffisamment matures. De nombreux perfectionnements restent cependant possibles, nous avançons à la fin de ce chapitre introductif quelques pistes de recherche pour améliorer les résultats exposés dans cette thèse.

2 Modélisation probabiliste de la formation des réserves et de l'exploration pétrolière

Les chapitres 2 et 3 proposent des modélisations probabilistes pour la formation des réserves et l'exploration pétrolière qui s'appuient sur la distribution de Poisson Dirichlet.

Formation des réserves

Avant de construire un modèle complet de la production d'un bassin pétrolier, il est essentiel de bien choisir une distribution de probabilité pour modéliser la taille des gisements qu'il contient. Nous étudions cette question dans le chapitre 2 en proposant un modèle probabiliste pour la formation des réserves pétrolières dans un bassin pétrolier.

Pour décrire la répartition spatiale des gisements pétroliers dans un bassin, on évoque souvent une organisation "en satellites" : à côté d'un très gros gisement se trouvent généralement quelques champs de tailles plus modestes, autour desquels on découvre de nombreux gisements plus petits. Ce type de structure spatiale observée dans les bassins pétroliers est compatible avec la propriété d'invariance stochastique qui caractérise la distribution de Lévy-Paréto. Cette modélisation a notamment été utilisée par Houghton (1988) et plus récemment par Lepez (2002). Les travaux de Lepez montrent que la distribution de Lévy-Pareto modélise convenablement les tailles des gisements d'un bassin pétrolier. La faiblesse principale de ce point de vue est que les distributions de Lévy-Pareto ne décrivent que les tailles de gisements en-deçà d'un certain seuil ε qui doit lui aussi être estimé. Ce seuil est généralement interprété comme un seuil de visibilité ou de rentabilité des gisements et celui-ci est donc susceptible d'évoluer au cours du temps. Nous proposons une nouvelle modélisation de la taille des gisements qui s'affranchit de cette difficulté. Les partitions aléatoires échangeables de $\mathbb{N}$ introduites par Kingman (1975) (voir aussi les notes de Saint-Flour de Pitman, 2006) peuvent être aussi utilisées pour modéliser la répartition des réserves à l'intérieur du bassin. Plus précisément, les tailles des gisements sont représentées par la suite (infinie) des tailles relatives des blocs aléatoires d'une partition dite "stable", et non plus par un échantillon (fini) de variables aléatoires d'une distribution fixée. Les tailles relatives de ces blocs peuvent être modélisés par les sauts normalisés d'un subordinateur stable, et les sauts du subordinateur représentent alors la quantité de pétrole contenue dans les gisements du bassin. La loi des sauts d'un subordinateur stable ordonnés de façon décroissante est appelée distribution de Poisson Dirichlet $PD(\alpha, 0)$. Les deux points de vue échantillon et subordinateur stable sont cependant cohérents car pour une partition stable d'indice α, la loi des tailles de bloc supérieures à un seuil ε est une loi de type Lévy-Paréto d'indice α.

Il est possible de justifier l'apparition de la distribution $PD(\alpha, 0)$ dans ce contexte pétrolier en proposant un modèle naïf pour la formation des réserves, qui est adapté du Random Energy Model (REM) introduit par Derrida (2000) pour étudier un modèle simplifié de verres de spin. Pour cela, on suppose que les gisements observés réellement au sein d'un même bassin peuvent se subdiviser en un grand nombre de petits gisements élémentaires $g_1, \ldots, g_M$ sur lesquels l'action géologique a agi de façon équivalente d'un point de vue statistique. L'action d'un certain facteur géologique sur la taille du gisement g_p est de multiplier la taille de g_p par le coefficient $\exp \epsilon_i^p$ où l'on suppose que les variables $(\epsilon_i^p; 1 \leq i \leq N, 1 \leq p \leq M)$ sont indépendantes et de même loi

$$\mathbb{P}(\epsilon_i^p = -v) = \mathbb{P}(\epsilon_i^p = v) = 1/2,$$

où v correspond à l'amplitude de l'événement géologique. L'action cumulée des N facteurs géologiques aboutit à multiplier la taille de g_p par le coefficient

$$\exp\left(\sum_{i=1}^{N} \epsilon_i^p\right).$$

Ne considérer que les facteurs géologiques qui ont permis de distinguer et de déterminer la

subdivision en M gisements élémentaires revient à supposer que N est de l'ordre de $\log M$. On pose $M = 2^N$ pour fixer les idées. La taille du gisement g_p admet alors pour expression $\langle t_{v,N} \rangle \exp\left(v\sqrt{N}G(g_p)\right)$, où $G(g_p)$ est une variable aléatoire de loi gaussienne centrée standard, et $\langle t_{v,N} \rangle$ est une taille typique commune à tous les gisements élémentaires, qui fixe l'échelle de grandeur dans laquelle les gisements élémentaires se situent. On peut alors montrer qu'il existe une amplitude critique v_c telle que pour $v > v_c = \sqrt{2\log(2)}$, et pour une bon choix de $\langle t_{v,N} \rangle$, les tailles de gisements ordonnés par ordre décroissant converge en loi vers la distribution $PD(\alpha, 0)$. Il est donc naturel de faire intervenir les distributions $PD(\alpha, 0)$ pour modéliser les tailles de gisements pétrolier.

Pour tenir compte du fait que v est susceptible d'évoluer au cours de la formation géologique du bassin (avec $v > v_c$), nous nous appuyons sur le modèle de Bolthausen et Sznitman qui permet de décrire des opérations de coalescence sur des partitions stables. Le processus "dual" que l'on obtient en renversant le temps est appelé fragmentation de Bolthausen-Sznitman. Ces deux processus ont la propriété remarquable d'être markoviens. L'évolution des réserves d'un bassin soumis à des opérations successives de fragmentation et de coalescence peut être décrite qualitativement par ce modèle : ces processus sont tels qu'à tout instant, la distribution des réserves est $PD(\alpha, 0)$ où seul α peut évoluer au cours du temps. Cette propriété remarquable nous conforte aussi dans le choix d'utiliser les lois de Lévy-Pareto pour modéliser les tailles de champs pétrolier.

Exploration pétrolière.

Le chapitre 3 est consacré à un premier modèle de l'exploration pétrolière, avec un point de vue volontairement qualitatif. Dans ce contexte, le temps désigne donc l'activité pétrolière et ne correspond plus comme précédemment au "temps géologique" de la formation du bassin. La modélisation des réserves d'un bassin à l'aide de partitions stables permet de décrire les découvertes successives dans le bassin. En supposant que la visibilité d'un gisement dépend de sa taille, nous modélisons le temps de découverte d'un gisement de taille x par une distribution conditionnelle exponentielle de paramètre $h(x)$ où h est appelée fonction de visibilité. Si la fonction de visibilité est choisie proportionnelle à la taille du gisement, les résultats de Pitman montrent d'une part que la suite des réserves restantes dans le bassin peut être décrite par une chaîne de Markov, et d'autre part que la géologie du sous-sol, à travers le coefficient α de la partition stable modélisant les réserves, détermine à quelle vitesse le sous-sol est progressivement épuisé. Ce modèle idéalisé de l'exploration pétrolière se révèle malheureusement difficile à utiliser en pratique. En particulier, le choix d'une fonction de visibilité proportionnelle à la taille des gisements ne semble pas suffisamment réaliste. Pour la suite de l'exposé, nous retenons de cette modélisation de l'exploration pétrolière que le temps de découverte d'un gisement pétrolier de taille x peut être modélisé par une loi exponentielle de paramètre $h(x)$. La modélisation générale de la production pétrolière que nous développons dans les chapitres suivants reposera, entre autres, sur cette hypothèse fondamentale.

3 Modélisation de la production pétrolière et position du problème statistique

Afin de construire un modèle pour la production d'un bassin qui soit le plus fidèle possible à la réalité, nous décrivons chacune des étapes intervenant dans la production des hydrocarbures. La production d'un bassin est définie comme l'agrégation des productions des gisements exploités à un instant donné. Dans un bassin contenant n gisements, en notant Y_u et L_u respectivement les réserves et la date de mise en production du gisement u, la production du bassin à l'instant t a pour expression

$$\mathrm{Prod}(t) = \sum_{u=1}^{n} \mathrm{prod}_u(t - L_u)$$

où $\mathrm{prod}_u(\tau)$ où désigne la production du champ u après τ années de production.

Pour décrire comment les réserves d'hydrocarbures sont réparties et découvertes dans le bassin, nous nous appuyons sur les conclusions des chapitres 2 et 4, avec les modifications nécessaires pour disposer de procédures d'estimations et rendre le modèle de production opérationnel. Il s'agit aussi de déterminer comment sont produits les hydrocarbures d'un gisement en exploitation. Le chapitre 4 est consacré à la description des modélisations retenues pour chacun des éléments rentrant dans le modèle complet de la production du bassin, et au relevé des principaux problèmes statistiques rencontrés.

Réserves et exploration pétrolières

D'après le chapitre 2, les tailles des gisements d'un bassin pétrolier de taille supérieures à un seuil ε peuvent être modélisées par un échantillon de variables aléatoires de distribution de Lévy-Paréto. Il serait plus délicat d'élaborer des procédures d'estimation en utilisant les subordinateurs stables et il est donc préférable à ce stade de revenir au point de vue "échantillon". Concernant l'estimation du paramètre α de la distribution de Lévy-Paréto, nous utilisons les travaux de Lepez (2002) qui traitent complètement de cette question et aucun travail statistique supplémentaire ne sera donc nécessaire.

Pour modéliser les dates de découverte des gisements du bassin, nous opérons une distinction entre les gisements les plus petits et le reste des gisements. Les gisements de taille supérieure à un seuil x_0 sont représentés par une distribution conditionnelle de loi exponentielle de paramètre $h(x)$, où h est une fonction appelée *fonction de visibilité des champs* qui est continue, croissante et affine par morceaux. Le fait que les gisements de petite taille (de taille inférieure au seuil x_0) soient en très grand nombre et possèdent une faible visibilité, permet de supposer que ceux-ci sont découverts selon un processus de Poisson homogène dans le temps.

La principale difficulté statistique associée à la modélisation présentée ci-dessus est l'estimation de la fonction de visibilité h. Nous précisons maintenant la nature de ce problème statistique. Lorsqu'un bassin est suffisamment mature, la taille maximale $x_{\max}$ de ses gisements est connue. Soit X la taille d'un gisement du bassin et D sa date de découverte, la

densité de la loi conditionnelle de (X, D) sachant $X \in [x_0, x_{\max}]$ est donnée par

$$g : (x,t) \longmapsto \alpha h(x) \exp\{-h(x)t\} \frac{x^{-\alpha-1}}{x_0^{-\alpha} - x_{\max}^{-\alpha}} \mathbb{1}_{t \geq 0,\, x_0 \leq x \leq x_{\max}}.$$

À la date $t^\star$ du présent, un gisement n'est observé que si $D \leq t^\star$. Soit un couple $(X^\star, D^\star)$ associé à un gisement découvert avant la date $t^\star$; celui-ci a pour distribution

$$(X^\star, D^\star) \stackrel{\mathcal{L}}{=} \left((X, D) \mid D \leq t^\star\right) .$$

La densité $g^\star$ de $(X^\star, D^\star)$ est définie pour tout $t \geq 0$ et tout $x \in \mathbb{R}^+$ par

$$g^\star(x,t) = \frac{g(x,t)}{P_{\text{dec}}(\alpha, t^\star, h)} \mathbb{1}_{t \leq t^\star},$$

où P_{dec} est la probabilité qu'un champ de taille dans $[x_0, x_{\max}]$ soit découvert avant $t^\star$. Pour une partition m en intervalles de $[x_0, x_{\max}]$ nous considérons le modèle S_m composé de ces densités $g^\star$ pour lesquelles les fonctions de visibilité h sont croissantes, continues et affines par morceaux sur la partition m, ce qui signifie que les ruptures de pente de h ont lieu au niveau des extrémités des intervalles qui composent m. Le problème statistique à résoudre peut donc être énoncé comme suit. Nous disposons d'un échantillon $((X_1^\star, D_1^\star), \ldots, (X_n^\star, D_n^\star))$ de densité inconnue s et correspondant à l'observation des gisements de taille supérieure à x_0 et découverts avant $t^\star$. Dans chaque modèle S_m, nous disposons d'un estimateur du maximum de vraisemblance pour s. Il s'agit donc de choisir un modèle S_m dans lequel l'estimation de s soit la meilleure possible. Cette question est traitée dans le chapitre 5.

Classification de profils de production pétroliers

La construction d'un modèle de production de bassin pétrolier nécessite aussi de modéliser comment les gisements exploités produisent leur pétrole au cours du temps. On appelle *profil de production* la courbe de la production d'un champ en fonction du temps, et *profil de production normalisé* la courbe de la production divisée par les réserves totales du champ. Dans l'industrie pétrolière, il existe un principe généralement admis selon lequel la forme d'un profil de production normalisé dépend essentiellement de la quantité de pétrole qu'il contient. Nous souhaitons valider ce principe à l'aide d'une procédure de classification non supervisée de courbes, avant de proposer (au chapitre 7) un modèle simple pour la production individuelle, qui est défini en fonction des réserves du gisement.

De nombreux auteurs ramènent le problème de la classification non supervisée[2] de courbes à la classification non supervisée "classique" en projetant les courbes sur une base fonctionnelle de fonctions splines ou de fonctions d'ondelettes. C'est par exemple le cas dans les travaux de Abraham *et al.* (2003), García-Escudero et Gordaliza (2005), Ma *et al.* (2006) et James et Sugar (2003). Dans tous ces articles, une transformation B-splines des courbes est effectuée et différentes méthodes de classification sont ensuite utilisées sur ces données transformées. En

[2]L'objectif de la classification dite "non supervisée" est de regrouper un ensemble de données en différents paquets homogènes, en ce sens que les données de chaque sous-ensemble partagent des caractéristiques communes. Celle-ci se distingue de la classification supervisée où les données d'apprentissage sont déjà étiquetées.

classification non supervisée, les modèles de mélanges gaussiens offrent un cadre statistique adéquat pour choisir le nombre de composantes selon un critère statistique. La densité d'un mélange gaussien à K composantes s'écrit sous la forme

$$t = \sum_{k=1}^{K} p_k \Phi(\cdot|\eta_k, \Sigma_k)$$

où les p_k sont les proportions du mélange et $\Phi(\cdot|\eta_k, \Lambda_k)$ désigne une densité gaussienne v-dimensionnelle de moyenne η_k et de matrice de covariance Σ_k. Sur la base d'une estimation de la densité du mélange, nous disposons ensuite d'une classification en utilisant la règle du maximum *a posteriori*. Ainsi le problème de classification est donc reposé comme un problème d'estimation de densité dans un modèle de mélange gaussien.

Dans de nombreuses situations en classification de courbes, le nombre de points de chaque courbe peut être de l'ordre de n, voire même beaucoup plus grand. Pour ce problème que l'on peut qualifier de *grande dimension*, la qualité de l'estimation peut donc être altérée si celle-ci est effectuée dans un modèle comportant trop de paramètres. Nous considérons alors des collections de modèles de mélanges gaussiens définis de la façon suivante : pour $\mathbf{v}$ un sous-ensemble de $\{1, \ldots, Q\}$ soit le modèle

$$S_{(K,\mathbf{v})} = \left\{ x \in \mathbb{R}^Q \mapsto f(x_{[\mathbf{v}]})\, \Phi(x_{[\mathbf{v}^c]} \,|\, 0, I_{Q-v}) \;;\;\; f \in \mathcal{L}_{(K,v)} \right\}$$

où $v = |\mathbf{v}|$, et $\mathcal{L}_{(K,v)}$ est une famille de mélanges gaussiens sur $\mathbb{R}^v$ à K composantes. La loi jointe des variables de classification est modélisée par une distribution de mélange gaussien alors que les variables restantes forment un vecteur de dimension $Q-v$, de loi normale centrée réduite. La sélection d'un modèle $S_{(K,v)}$ parmi une collection disponible conduit donc à une classification des données, mais aussi à une sélection d'un bloc de variables de classification. Notons que cette procédure de sélection de variables est aussi bénéfique pour la classification. En effet, certaines variables peuvent être inutiles, voire même jouer un rôle néfaste vis-à-vis de cet objectif. Cet argument est avancé par de nombreux auteurs pour développer des méthodes intégrant classification et sélection de variables. Citons par exemple les travaux de Law *et al.* (2004) où le concept de *feature saliency* est défini pour déterminer un ensemble de variables pertinentes pour la classification. C'est le cas aussi dans Raftery et Dean (2006) et Maugis *et al.* (2007) où les problèmes de sélection de variables et de classification non supervisé sont reposés là aussi comme un unique problème de sélection de modèles.

Concernant le problème de classification de profils pétroliers qui nous intéresse plus particulièrement ici, en considérant des observations rassemblant des informations sur les courbes (coefficients d'ondelettes par exemple) mais aussi des variables explicatives supplémentaires, la sélection d'un modèle $S_{(K,\mathbf{v})}$ permet aussi de déterminer quelles variables explicatives sont cohérentes avec la classification obtenue.

Comme dans le cadre des modèles d'exploration pétrolière, la minimisation de l'erreur d'estimation sur l'ensemble des estimateurs du maximum de vraisemblance associés à une collection de modèles $S_{(K,\mathbf{v})}$ apparaît comme un critère naturel pour choisir un modèle dans une collection fixée. Une résolution de ce problème statistique est proposée au chapitre 5.

Politique de mise en production des gisements

La politique de mise en production des gisements constitue le dernier élément qu'il nous faut étudier pour disposer d'un modèle complet de la production d'un bassin pétrolier. Soit $(L_u)_{u\geq 1}$ la suite des dates de mise en production de gisements, ces dates sont modélisées par un processus de Poisson non nécessairement homogène. À chaque instant L_u, il existe un stock de gisements découverts non encore exploités qui peuvent donc être potentiellement mis en production. Dans la réalité, il est difficile de reproduire fidèlement la façon avec laquelle un gisement est choisi dans le stock, mais pour des raisons économiques évidentes les gisements découverts les plus gros sont produits en priorité. Il est donc raisonnable de modéliser la politique de sélection d'un gisement dans le stock de champs disponibles par un tirage aléatoire biaisé par une fonction de la taille dans cette population.

Nous ne développons pas de procédure élaborée pour estimer ce modèle de mise en production. Notons que la question de l'estimation du biais de tirage est difficile car la population des gisements du stock évolue constamment. De plus, l'intensité de mise en production relève aussi de choix économiques et politiques que nous ne prétendons pas modéliser. À ce niveau nous touchons aux limites de notre modélisation et nous préférons adopter un point de vue plus prudent en considérant différents scénarios de politique de mise en production du bassin traduisant une volonté de développement du bassin plus ou moins marquée.

4 Sélection de modèles pour l'estimation de densité

Les deux principaux problèmes statistiques identifiés au chapitre 4 sont la sélection d'un modèle d'exploration d'une part, et la sélection d'un modèle de mélange gaussien pour la classification non supervisée de courbes d'autre part. Le chapitre 5 est consacré à la résolution de ces deux problèmes qui relèvent tous deux de la sélection de modèle en estimation de densité.

Soit $X_1, \ldots, X_n$ un échantillon i.i.d., avec $X_i \in \mathbb{R}^d$ de densité de probabilité s inconnue pour la mesure de Lebesgue sur $\mathbb{R}^d$. Soit $\mathcal{S}$ l'ensemble de toutes les densités pour la mesure de Lebesgue sur $\mathbb{R}^d$. La méthode du maximum de vraisemblance, qui consiste à trouver les paramètres d'un modèle qui maximisent la vraisemblance des observations, peut être réinterprétée comme une méthode de minimisation de contraste. Pour cela, nous considérons le contraste $\gamma(t,.) = -\ln\{t(\cdot)\}$. Soit le contraste empirique

$$\gamma_n(t) = -\frac{1}{n}\sum_{i=1}^{n} \ln\{t(X_i)\}$$

associé à γ. Dans ce contexte, la fonction de perte définie par $l(s,t) = \mathbb{E}[\gamma_n(t)] - \mathbb{E}[\gamma_n(s)]$ est exactement l'information de Kullback-Leibler. Celle-ci est définie pour deux densités f et g de $\mathcal{S}$ par

$$\mathrm{KL}(f,g) = \int \ln\left\{\frac{f(x)}{g(x)}\right\} f(x)\,dx$$

si $f\,dx$ est absolument continue par rapport à $g\,dx$, et $+\infty$ sinon. La densité s est l'unique

fonction de $\mathcal{S}$ telle que

$$s = \underset{t \in \mathcal{S}}{\operatorname{argmin}} \int \gamma(t,x)s(x)\,dx.$$

Soit S un sous-ensemble de $\mathcal{S}$, l'estimateur du maximum de vraisemblance (EMV) de s sur S est défini par

$$\hat{s} := \underset{t \in S}{\operatorname{argmin}}\, \gamma_n(t).$$

En remplaçant ainsi γ par γ_n et $\mathcal{S}$ par S, on s'attend à ce que l'estimateur obtenu soit proche de la véritable densité s, au moins dans le cas où s n'est pas "trop loin" du modèle S et pour n suffisamment grand.

Dans le contexte de la sélection de modèles qui est le nôtre, nous disposons d'une collection de modèles $(S_m)_{m \in \mathcal{M}_n}$ et d'un estimateur du maximum de vraisemblance $\hat{s}_m$ pour chacun d'entre eux. Comme les notations le suggèrent, la collection de modèles est autorisée à dépendre de la taille n de l'échantillon observé. Notons de plus qu'il n'est pas nécessaire de supposer que la densité s appartienne à l'un des modèles de la collection. Nous souhaitons utiliser l'EMV associé au "meilleur modèle", au sens d'un certain critère statistique. Dans le cadre de l'estimation de densité, un critère naturel est la minimisation de l'*erreur moyenne d'estimation* (ou *risque d'estimation*), que l'on définit pour un estimateur $\hat{s}_m$ par

$$\mathcal{R}(\hat{s}_m) = \mathbb{E}[\mathrm{KL}(s, \hat{s}_m)].$$

Idéalement, nous souhaiterions sélectionner le modèle minimisant cette quantité. Cependant, ceci est impossible en pratique car le risque dépend de la densité s qui est inconnue ; la densité $\tilde{m}$ qui minimise le risque d'estimation pour la collection $(S_m)_{m \in \mathcal{M}_n}$ est appelée *oracle*. Une procédure de sélection de modèle est considérée de bonne qualité si celle-ci permet de sélectionner un modèle dont l'EMV a les mêmes performances que celle de l'oracle. Une *inégalité oracle* permet de mettre en évidence de telles propriétés de façon non asymptotique :

$$\mathbb{E}[\mathrm{KL}(s, \hat{s}_m)] \leq C\mathbb{E}\left[\inf_{m\mathcal{M}_n} \{\mathrm{KL}(s, \hat{s}_m) + R(m,n)\}\right]$$

où C est une constante et $R(m,n)$ est un terme de reste qui ne doit pas être trop grand devant le terme de risque.

Puisque le risque de $\hat{s}_m$ est égal à $\mathbb{E}\left[\int \{\gamma(\hat{s}_m) - \gamma(s)\} s\,dx\right]$, l'objectif est donc de minimiser la quantité $\mathbb{E}[\int \gamma(\hat{s}_m) s\,dx]$. Une proposition naturelle pour sélectionner un modèle dans la collection serait de choisir celui pour lequel l'EMV minimise le critère $\gamma_n(\hat{s}_m)$. Cependant, cette méthode conduit à sous-estimer le risque $\mathcal{R}(m)$ et le critère obtenu sélectionnerait systématiquement les grands modèles. Cette sous-estimation (que l'on qualifie d'*erreur de substitution*) dépend en réalité de la complexité des modèles, les procédures de *pénalisation* consistent alors à considérer des critères de la forme

$$\mathrm{crit}(m) = \gamma_n(\hat{s}_m) + \mathrm{pen}(m)$$

où pen : $\mathcal{M}_n \mapsto \mathbb{R}$ est une fonction qui mesure la complexité des modèles et permet de pé-

naliser les modèles de trop grande complexité. Les premières procédures de pénalisation pour l'estimation de densité ont été proposées dans les années 70 par Akaike (1973). L'heuristique proposée par Akaike conduit à considérer une pénalité de la forme $\frac{D_m}{n}$ où D_m est la dimension du modèle S_m. Le point de vue d'Akaike peut être qualifié d'*asymptotique* puisque dans l'heuristique qu'il propose, la taille de l'échantillon est censée tendre vers l'infini indépendamment de la collection de modèles. Notons que les critères pénalisés ne sont pas propres à l'estimation de densité, ils sont aussi utilisés pour la classification et la régression. Pour ce dernier contexte, les premiers résultats sont dus à Mallows (1973) qui propose des pénalités de la forme $2\frac{D_m\sigma^2}{n}$ si la variance est connue.

Avec les travaux de Ledoux et Talagrand (voir Ledoux et Talagrand, 1991; Talagrand, 1995) sur le phénomène de concentration de la mesure, Birgé et Massart ont pu développer une approche non asymptotique de la pénalisation dont une présentation générale est disponible dans les notes de Saint-Flour de Massart (2007). En utilisant cette approche pour l'estimation de densité par des histogrammes, et pour des collections qui ne sont pas trop riches, Castellan (2003) montre qu'une pénalité de la forme $c_1\frac{D_m}{n}$ avec $c_1 > \frac{1}{2}$, permet d'obtenir une inégalité oracle pour l'estimateur pénalisé. Ce dernier possède de plus des propriétés d'adaptativité sur des classes de Hölder ; l'estimateur pénalisé réalise le risque minimax sur une famille de classes de Hölder sans utiliser pour autant la connaissance de cette classe.

De façon générale dans le contexte de l'estimation de densité, il est rarement possible de reproduire la démarche utilisée par Castellan qui exploite de façon fine le bon comportement des modèles d'histogrammes ou exponentiels vis-à-vis des inégalités de Talagrand. Pour obtenir des résultats de sélection de modèles hors du contexte étudié par Castellan, Massart (2007, section 7.4) propose un théorème général garantissant une inégalité oracle pour l'estimateur pénalisé. Plutôt que d'utiliser la dimension des modèles pour définir une pénalité convenable, ce résultat s'appuie sur la notion d'entropie à crochets, qui permet elle aussi de donner une mesure de la complexité d'un modèle. Les méthodes utilisées pour démontrer ce résultat ne permettent par d'évaluer avec précision les constantes en jeu dans la pénalité et l'inégalité oracle. La forme de la pénalité ainsi que la borne de risque non asymptotique obtenues dans ce théorème doivent être considérées d'un point de vue qualitatif. Essentiellement, ces résultats nous donnent la forme générale de la pénalité à utiliser pour la *méthode de la pente* (Birgé et Massart, 2006) qui permet dans un second temps de calibrer la pénalité en fonction des données. Pour pouvoir utiliser ce résultat, il nous faut effectuer des calculs techniques d'entropie à crochets pour les modèles associés aux deux problématiques qui nous intéressent dans ce travail de thèse, à savoir les collections de modèles $\mathcal{S}_{(K,v)}$ de mélanges gaussiens pour la classification non supervisée, et les collections de modèles d'exploration pétrolière.

Sélection de modèles de mélanges gaussiens

Rappelons que nous étudions des collections de modèles de la forme

$$S_{(K,\mathbf{v})} = \left\{x \in \mathbb{R}^Q \mapsto f(x_{[\mathbf{v}]})\,\Phi(x_{[\mathbf{v}^c]}\,|\,0, I_{Q-v})\ ;\ \ f \in \mathcal{L}_{(K,v)}\right\}$$

où l'ensemble $\mathcal{L}_{(K,v)}$ est composé d'une famille de densité de mélanges gaussiens de dimension $|v|$ à K composantes et dont la *forme*, c'est-à-dire le type de matrice variance-covariance autorisé dans les composantes du mélange, dépend de la collection considérée. Le vecteur des variables qui ne sont pas utilisées dans la structure de mélange est de distribution gaussienne multidimensionnelle centrée réduite. En nous appuyant sur la décomposition des matrices de variance-covariance des mélanges gaussiens proposée par Celeux et Govaert (1995), nous définissons les modèles $\mathcal{L}_{(K,v)}$ associés à trois collections de modèles :

– Collection $\mathcal{M}[LB_k]$:

$$\mathcal{L}_{(K,v)} = \left\{ \sum_{k=1}^{K} p_k \Phi(\cdot|\mu_k, \lambda\Sigma_k); \begin{array}{l} \lambda \in [{\sigma_{\mathrm{m}}}^2, {\sigma_{\mathrm{M}}}^2],\ \Sigma_k \in \Delta^1_{(v)}({\sigma_{\mathrm{m}}}^2, {\sigma_{\mathrm{M}}}^2) \\ \mu_k \in [-a,a]^v,\ 0 < p_k < 1,\ \sum\limits_{k=1}^{K} p_k = 1 \end{array} \right\}$$

où $\Delta^1_{(v)}({\sigma_{\mathrm{m}}}^2, {\sigma_{\mathrm{M}}}^2)$ désigne l'ensemble des matrices diagonales définies positives de déterminant 1 dont les valeurs propres appartiennent à l'intervalle $[{\sigma_{\mathrm{m}}}^2, {\sigma_{\mathrm{M}}}^2]$.

– Collection $\mathcal{M}[L_kB_k]$:

$$\mathcal{L}_{(K,v)} = \left\{ \sum_{k=1}^{K} p_k \Phi(\cdot|\mu_k, \Sigma_k); \begin{array}{l} \Sigma_k = \mathrm{diag}(\sigma^2_{k1}, \ldots, \sigma^2_{kv}),\ \sigma^2_{k1}, \ldots, \sigma^2_{kv} \in [{\sigma_{\mathrm{m}}}^2, {\sigma_{\mathrm{M}}}^2] \\ \mu_k \in [-a,a]^v,\ 0 < p_k < 1,\ \sum\limits_{k=1}^{K} p_k = 1 \end{array} \right\}.$$

– Collection $\mathcal{M}[L_kC_k]$:

$$\mathcal{L}_{(K,v)} = \left\{ \sum_{k=1}^{K} p_k \Phi(\cdot|\mu_k, \Sigma_k); \begin{array}{l} \Sigma_k \in \mathcal{D}^+_{(v)}({\sigma_{\mathrm{m}}}^2, {\sigma_{\mathrm{M}}}^2),\ \mu_k \in [-a,a]^v \\ 0 < p_k < 1,\ \sum\limits_{k=1}^{K} p_k = 1 \end{array} \right\}$$

où $\mathcal{D}^+_{(v)}({\sigma_{\mathrm{m}}}^2, {\sigma_{\mathrm{M}}}^2)$ désigne l'ensemble des matrices symétriques définies positives dont les valeurs propres appartiennent à l'intervalle $[{\sigma_{\mathrm{m}}}^2, {\sigma_{\mathrm{M}}}^2]$.

Ces trois collections de modèles permettent de traiter des situations variées dans la pratique. Les collections sont dites ordonnées si dans les modèles de la collection les blocs de variables de classification sont toujours de la forme $\mathbf{v} = \{1, \ldots, v\}$. Dans le cas contraire la collection est dite non ordonnée.

Pour chacune de ces trois collections, les calculs entropiques sur les ensembles $\mathcal{L}_{(K,v)}$ permettent de montrer que pour une pénalité choisie telle que

$$\mathrm{pen}(K,v) \geq \kappa \frac{D(K,v)}{n} \left\{ 2A \ln v + 1 - \ln\left(1 \wedge \left[\frac{D(K,v)}{n} A \ln v\right]\right)\right\} \tag{1}$$

dans le cas d'une collection ordonnée, et telle que

$$\mathrm{pen}(K,\mathbf{v}) \geq \kappa \frac{D(K,\mathbf{v})}{n} \left(2A \ln v - \ln\left(1 \wedge \left[\frac{D(K,v)}{n} A \ln v\right]\right) + \frac{1}{2} \ln\left[\frac{8eQ}{\{D(K,\mathbf{v})-1\} \wedge (2Q-1)}\right]\right)$$

dans le cas d'une collection non ordonnée, où A et κ sont des constantes, une inégalité oracle est alors obtenue pour l'estimateur pénalisé.

Contrairement aux situations classiques en statistique pour lesquelles le nombre de variables Q est fixé alors que n tend vers l'infini, les deux résultats précédents permettent de considérer les situations où Q augmente avec n. Pour les problèmes spécifiques où le nombre de variables est de l'ordre de n, voir même plus grand que n, les inégalités oracles obtenues montrent que le critère pénalisé est encore pertinent.

Bien les modèles LB_k soient des sous-modèles de la famille L_kB_k, qui sont eux mêmes des sous-modèles de la famille L_kC_k, ces résultats nécessitent une démonstration spécifique pour chacune des trois collections. En effet pour que le terme $D(K, v)$ dans la pénalité corresponde réellement au nombre de paramètres libres dans chacune des trois situations, il est nécessaire de mener les calculs d'entropie métrique pour chacun des trois types de modèles.

Si l'on ne retient que le terme prépondérant dans les termes de minoration, les pénalités à utiliser en pratique sont donc proportionnelles à la dimension. Les constantes dans ces résultats ne sont pas explicites et ce résultat permet surtout de justifier la forme de la pénalité à utiliser en pratique pour la méthode de la pente.

Sélection d'un modèle d'exploration pétrolière

Nous considérons une collection $(S_m^b)_{m\in\mathcal{M}_n}$ de modèles d'exploration pétrolière indexés par des partitions m de $[x_0, x_{\max}]$, telle que la collection de partitions $\mathcal{M}_n$ est autorisée à dépendre de la taille n de l'échantillon observé. Plus précisément, les modèles S_m^b sont des ensembles de densités tels que

$$S_m^b := \left\{ g^\star : (x,t) \mapsto \frac{h(x)\exp\{-h(x)t\}}{\mathrm{P}_{\mathrm{dec}}(\alpha, t^\star, h)} \frac{\alpha x^{-\alpha-1}}{x_0^{-\alpha} - x_{\max}^{-\alpha}} \mathbb{1}_{0\le t\le t^\star} \mathbb{1}_{x_1\le x\le x_{\max}} \,,\ h \in \mathcal{H}_m^b \right\}.$$

L'indice de Lévy-Paréto α est supposé connu et fixé. L'ensemble$\mathcal{H}_m^b$ désigne la famille des fonctions (de visibilité) croissantes, continues, affines par morceaux pour la partition m et bornées par $h_{\min}$ et $h_{\max}$ avec $0 < h_{\min} < h_{\max}$. Le nombre de paramètres libres de S_m^b est noté D_m, avec $D_m = k+1$ où k désigne le nombre d'intervalles de la partition m. Pour une partition $m \in \mathcal{M}_n$, on note l_m la longueur minimale des intervalles qui la composent. Il nous faut supposer l'hypothèse suivante sur les partitions m de $\mathcal{M}_n$:

$$(\mathbf{P_2}) : \quad \text{Il existe deux constantes } \gamma > 0 \text{ et } c > 0 \text{ telles que } \inf_{m\in\mathcal{M}_n} l_m \ge c\, n^{-\gamma}.$$

Nous reprenons les deux types de collections de partitions proposées par Castellan (1999) en considérant les deux situations suivantes :

- $(\mathbf{H_1})$ **Complexité polynomiale.** La collection $\mathcal{M}_n$ vérifie (P_2) et il existe deux constantes B_1 et β telles que le nombre de partitions de $\mathcal{M}_n$ composées de k intervalles est borné par $B_1 k^\beta$. Cette situation est par exemple celle d'une collection de partitions régulières, c'est-à-dire telle que la collection ne comporte que les partitions composées d'intervalles de même longueur $(x_{\max} - x_0)/k$, si la partition est de taille k.
- $(\mathbf{H_2})$: **Complexité exponentielle.** Soit une partition m_n de $[x_0, x_{\max}]$ composée de $N_n + 1$ intervalles avec $N_n + 1 \le B_2\, n/\ln^2 n$ et telle que m_n vérifie la condition (P_2), c'est-à-dire $l_{m_n} \ge c\, n^{-\gamma}$. On considère la grille formée des extrémités des intervalles de m_n et soit $\mathcal{M}_n$ un ensemble de partitions de $[x_0, x_{\max}]$ s'appuyant sur des points de cette

grille. Pour un nombre d'intervalles $k+1 \leq N_n$ donné, il existe au plus $C_{N_n}^k$ partitions possibles dans la collection $\mathcal{M}_n$.

Les calculs d'entropie à crochets sur les modèles S_m^b permettent de déterminer la forme des pénalités à utiliser en pratique. Sous les hypothèses (H_1) et (H_2), une condition de la forme

$$\mathrm{pen}(m) \geq \eta \frac{D_m}{n} \ln n$$

permet de garantir une inégalité oracle. Là encore, ces résultats sont à considérer d'un point de vue qualitatif car les constantes en jeu ne sont pas explicites.

5 Heuristique de pente

Pour rendre opérationnels les résultats de sélection de modèle obtenus pour les collections de modèles de mélanges gaussiens et les collections de modèles d'exploration pétrolière, nous utilisons au chapitre 6 la méthode de *l'heuristique de pente*. Cette méthode a été proposée par Birgé et Massart (2001; 2006) pour calibrer à partir des données des pénalités de formes connues. Dans les deux cas étudiés ici, les pénalités à utiliser sont proportionnelles à la dimension des modèles (à taille d'échantillon fixé), et l'heuristique de pente permet de calibrer la constante de proportionnalité.

Pour une pénalité proportionnelle à la dimension, la méthode de la pente dans sa version élémentaire consiste à ajuster une droite sur la courbe $D \mapsto -\gamma_n(\hat{s}_D)$ dans les grandes dimensions. La pente $\hat{\eta}$ de la droite ajustée permet de définir la pénalité par $\mathrm{pen}(D) = 2\hat{\eta}D$.

Une version plus élaborée de cette méthode repose sur la détection du "saut de dimension". Supposons que l'on s'intéresse à une pénalité de la forme $\eta\,\mathrm{pen}$ où pen est une fonction de D et n et où η est un paramètre à régler. Cette deuxième version de l'heuristique de pente repose sur les deux assertions suivantes :

- Il existe une pénalité minimale $\mathrm{pen}_{\min} = \eta_{\min}\,\mathrm{pen}$ dans la famille de fonctions retenues telle que toute pénalité inférieure sélectionne les modèles de grandes dimensions, et toute pénalité supérieure sélectionne des modèles de dimensions "raisonnables".
- Une pénalité choisie de l'ordre de $2\,\mathrm{pen}_{\min}$ permet de sélectionner un estimateur de risque comparable à celui de l'oracle.

Pour le moment, les deux assertions précédentes ont été vérifiés par Birgé et Massart (2006) dans le cadre de la régression sur design fixe avec bruit blanc gaussien homoscédastique, et par Arlot (2007) dans le contexte d'un bruit blanc hétéroscédastique sur design aléatoire, pour le cas des histogrammes. Des progrès importants ont été réalisés dans ce domaine depuis quelques années, et il est probable que cette heuristique soit encore valable dans de nombreuses autres situations. Pour ce qui nous concerne, les résultats de sélection de modèles obtenus au chapitre 5 ne permettent pas de mettre en évidence ce phénomène de saut de dimension, mais ils nous indiquent quelle forme de pénalité utiliser avec la méthode de la pente. Notons que celle-ci a déjà été utilisée dans des contextes où l'heuristique n'a pas été entièrement validée, tout en donnant en pratique de bons résultats. On peut citer par exemple des applications dans le domaine de la détection de ruptures par Lebarbier (2005), en génomique par Villers (2007), pour des modèles graphiques par Verzelen (2007), en classification non supervisée par

Baudry (2007) et aussi par Lepez (2002) pour l'estimation des réserves pétrolières.

La méthode de la pente est appliquée à un échantillon de courbes de production en mer du Nord. Nous effectuons une transformée en ondelettes discrète de chacun des profils de production normalisés. La méthode de la pente utilisée pour une collection de modèles de mélanges $\mathcal{M}[LB_k]$, une première fois sur les données de coefficients d'ondelettes uniquement, et une deuxième fois en ajoutant aussi des variables techniques décrivant les propriétés du gisement. Les résultats obtenus permettent de valider le principe selon lequel la forme d'un profil de production normalisé dépend essentiellement de la quantité de pétrole qu'il contient.

A la fin du chapitre 6 nous utilisons la méthode de la pente dans sa version "sauts de dimensions" pour estimer les fonctions de visibilité de trois bassins pétroliers en cours d'exploitation.

6 Application à la production pétrolière

Le chapitre 7 s'appuie sur les conclusions des chapitres précédents pour définir un modèle pour la production pétrolière dans un bassin exploité, qui soit utilisable en pratique. Au début du chapitre, nous proposons un modèle simple pour la production individuelle des gisements, comme fonction du temps et de la variable Réserves. Selon la modélisation retenue, la variable aléatoire de la production $\mathrm{prod}(x, \cdot)$ d'un gisement contenant une quantité x (déterministe ici) d'hydrocarbures vérifie

$$\mathbb{E}\left[\mathrm{prod}(x,t)\right] = x^{1-\beta}\,\mathrm{K}\left(\frac{t}{x^{\beta}}\right)$$

où la fonction K est le *profil élémentaire* et β, appelé *coefficient d'inertie*, contrôle la vitesse de production des champs en fonction de leur taille. La fonction K ainsi que le coefficient β sont estimés à partir d'un échantillon de courbes de production de gisements situés en mer du Nord. En particulier, la fonction K peut être estimée par des fonctions spline cubiques naturelles. Le couple (K, β) est finalement estimé par minimisation d'un critère de moindre carré pénalisant les fonctions splines présentant de trop fortes oscillations.

La seconde section du chapitre est consacrée à la description du modèle de production de bassin et à l'étude des courbes de production de bassin obtenues par simulations. Ce modèle de production de bassin repose sur l'ensemble des hypothèses suivantes :

1. La quantité de réserves X d'un gisement du bassin suit une loi de Lévy-Paréto.
2. Les champs sont découverts selon le modèle d'exploration suivant :
   - les dates de découverte des champs de taille supérieure au seuil x_0 sont de loi conditionnelle $(D \mid X) \sim \mathcal{E}\left(h(X)\right)$ où h est la fonction de visibilité des champs,
   - les gisements de taille inférieure à x_0 sont découverts selon un processus de Poisson homogène dans le temps.
3. Les champs en stock (disponible à la production) sont développés selon une "politique de mise en production" choisie par l'utilisateur.
4. Un gisement exploité produit ses réserves selon le modèle

$$\mathrm{prod}(x,t) = x^{1-\beta}\,\mathrm{K}\left(\frac{t}{x^{\beta}}\right)$$

où le profil élémentaire K et le coefficient d'inertie β ont tous deux été estimés dans la première section du chapitre.

Grâce aux simulations effectuées, des éléments de réponses peuvent être apportés aux questions générales formulées au début de cette introduction. Tout d'abord la courbe de production ne présente pas de symétrie systématique. Ensuite, pour une dynamique d'exploration fixée dans le bassin, l'intensité de mise en production des gisements a un effet limité sur la courbe de production du bassin. La distribution des réserves étant de type Lévy-Pareto, l'essentiel de la production du bassin est en réalité soutenue par quelques dizaines de gisements qui sont de plus découverts très rapidement. L'intensité de mise en production des gisements ne perturbe donc que peu la mise en production de ces gisements et au finale la courbe de production est peu sensible à cette intensité. De plus, une augmentation jointe de l'effort d'exploration et de mise en production du bassin a surtout pour effet d'avancer la date du pic et d'amplifier le déclin de la production une fois que ce dernier est franchi. Enfin, la mise en production massive des petits gisements ne permet pas de compenser le déclin au-delà du pic ; une telle politique de développement étant de plus extrêmement coûteuse, celle-ci a peu de chance d'être suivie dans le futur.

Dans la dernière section, une méthode de prolongement de l'exploration d'un bassin est détaillée et appliquée aux trois zones pétrolifères étudiées précédemment. Celle-ci nous permet finalement de proposer des scénarios de prolongement de la production de ces trois mêmes zones de production.

7 Éléments de conclusion et perspectives

À l'issu de ce travail de thèse, il nous est possible d'affirmer que les profils de production de bassin ne présentent pas de symétrie systématique. Nous avons aussi mis en évidence que le développement d'une grande quantité de gisements modestes ne suffirait pas à compenser le déclin de la production des gisements les plus importants. Enfin, ce n'est donc pas par la mobilisation d'un nombre important de petits gisements que la production pétrolière pourra être maintenue à des niveaux élevés. Dans les bassins matures, les pétroliers espèrent que le niveau moyen de récupération des gisements pourra être amélioré, permettant ainsi de prolonger la production des gisements les plus importants. Il est regrettable que notre modélisation ne puisse tenir compte de l'évolution du taux de récupération[3], il faudrait pour cela être capable de quantifier le biais systématique dû à la sous-estimation des ressources ultimes. Sur cette question, aucun travail de modélisation réaliste ne nous semble envisageable. Notons que le secteur pétrolier ne compte pas uniquement sur l'amélioration des taux de récupération pour compenser le déclin des plus gros gisements. Aujourd'hui la tendance observée est à la diversification des modes de production d'hydrocarbures avec le développement de projets pour des hydrocarbures comme les pétroles lourds et extra-lourds, les sables bitumeux ainsi que les schistes bitumineux ou encore l'off-shore ultra profond. Dans tous les cas, il paraît clair que l'ère du "pétrole facile à produire" touche à sa fin.

Cette thèse définit un modèle complet pour la production du pétrole dans un bassin

[3]Proportion des hydrocarbures extraits sur l'ensemble des hydrocarbures en place dans le gisement.

d'hydrocarbures, mais celui-ci n'est pas parfait et de nombreux perfectionnements pourront prolonger ce travail. Concernant la modélisation de l'exploration tout d'abord, la modélisation ne permet pas de faire varier l'effort global d'exploration pétrolière au cours du temps pour les gisements de taille supérieure à x_0. Plutôt que d'utiliser un échantillon de couples (X, D) (taille et date de découverte), nous pourrions modéliser l'exploration pétrolière par un unique processus de Poisson bidimensionnel non homogène $(D_i, X_i)_{i\geq 1}$ d'intensité $(t, x) \mapsto \mathrm{Int}(x, t)$. D'un point de vue statistique, cette intensité serait plus difficile à estimer que la fonction de visibilité h dans notre modélisation, mais ce point de vue plus élaboré faciliterait la création de scénarios de prolongement de l'exploration. Concernant la politique de mise en production des gisements, il serait plus satisfaisant de faire intervenir dans la modélisation des paramètres économiques pour affiner ces scénarios de développement du bassin. En effet, le développement des gisements dépend par exemple des investissements dont disposent les compagnies pétrolières[4]. Pour ce travail de thèse, nous n'avons pas souhaité développer cette question qui n'entre pas dans notre domaine de compétence. Nous espérons cependant pouvoir examiner ce problème grâce à une collaboration éventuelle avec un chercheur économiste.

Concernant les problèmes mathématiques étudiés, de nombreux prolongements à cette thèse peuvent être envisagés. À propos des résultats de sélection de modèles obtenus pour les collections de modèles de mélanges gaussiens, il serait plus satisfaisant de montrer que l'estimateur pénalisé obtenu possède des propriétés d'adaptativité en montrant que celui-ci atteint le risque minimax sur une large famille de classes fonctionnelles (classes de Hölder, de Sobolev, de Besov, ou sur des ellipsoïdes de $\mathbb{L}^2$). Pour cela, il nous faudrait contrôler, pour une densité s appartenant à l'un des espaces cités plus haut, le biais $\mathrm{KL}(s, S_{(K,v)})$ (ou pour n'importe quelle autre métrique) entre la densité s et l'espace fonctionnel composé des densités de mélanges gaussiens à K composantes. Nous espérons démontrer ultérieurement un tel résultat qui présenterait aussi un intérêt pour le domaine de l'approximation fonctionnelle.

Dans cette thèse, nous nous sommes intéressés à trois types de formes de mélanges gaussiens. En réalité, il serait possible d'étendre les résultats obtenus pour d'autres formes de mélanges. Quitte à rassembler ensuite plusieurs formes de mélanges dans une même collection, la méthode de la pente nous permettrait ainsi de sélectionner non seulement le nombre de variables de classification, le nombre de composantes mais aussi la forme du mélange. Toujours au sujet des collections de modèles de mélanges gaussiens, nous obtenons au chapitre 5 des résultats pour des collections non ordonnées de modèles de mélanges. Cependant en pratique, il est impossible d'estimer dans un temps raisonnable tous les modèles de telles collections dès que la dimension Q des vecteurs observés dépasse 10. Il serait intéressant de développer des stratégies de pré-sélection de modèles permettant de ne pas estimer tous les modèles, ce qui rendrait ainsi possible l'utilisation de ces collections de modèles. Notons enfin qu'il serait profitable d'adapter les procédures proposées au contexte des données manquantes.

[4]Notons que l'effort d'exploration dépend lui aussi des investissements.

Chapitre 1

Contexte Pétrolier

L'objet de ce premier chapitre est de préciser dans quel contexte nous étudions la production pétrolière. Nous commençons par préciser la notion de réserves pétrolières, puis nous discutons la "courbe de Hubbert" qui est la méthode la plus populaire pour modéliser et prolonger des profils de production pétroliers. Á la fin du chapitre, cet exposé préliminaire nous permet d'expliciter les objectifs et les hypothèses de notre travail de thèse.

1.1 Les réserves pétrolières

Cette partie s'attache à décrire la notion de réserves d'hydrocarbures. Nous renvoyons à Babusiaux et al. (2002) pour plus de détails à ce sujet.

1.1.1 Différents types d'hydrocarbures

Il existe des centaines de bruts de par le monde. Certains servent d'étalon pour établir le prix moyen du pétrole en provenance d'une région donnée. Les bruts les plus connus sont l'Arabian Light [1], le Brent [2] et le WTI[3].

Tous les pétroles ne possèdent pas les même propriétés chimiques. Il est possible de distinguer les différents types de pétrole selon leur densité (mesurée en degrés API), leur viscosité, leur teneur en soufre et autres impuretés (vanadium, mercure et sels). Ces caractéristiques permettent de préciser la *qualité* d'un pétrole. Il est aussi possible de classer les hydrocarbures en fonction de leur provenance : Golfe Persique, mer du nord, Venezuela, Nigeria, etc ... En effet le pétrole issu de gisements voisins a souvent des propriétés proches. Concernant le gaz, les classifications s'appuient non seulement sur la provenance, mais aussi sur la teneur en différentes classes d'hydrocarbures aussi présents dans les gisements concernés. Par exemple on distingue un gaz associé à des hydrocarbures liquides (huile, condensas) d'un gaz qui est le seul type d'hydrocarbure présent dans le gisement : on parle alors de *gaz sec*.

Les différents types d'hydrocarbures sont plus ou moins faciles à produire. Campbell *et al.* (1998) considèrent comme conventionnels les hydrocarbures qui peuvent être produits dans des conditions techniques et économiques actuelles et prévisibles dans le futur. Par opposition,

[1] brut de référence du Moyen-Orient.
[2] brut de référence européen.
[3] West Texas Intermediate, brut de référence américain.

on définit alors les hydrocarbures non conventionnels comme l'ensemble des hydrocarbures qui ne sont pas, aujourd'hui et dans le futur, techniquement exploitables à un coût raisonnable. Notons que tous les experts ne répartissent pas de la même façon les différents bruts entre ces deux catégories, ce qui explique parfois des différences importantes dans les estimations de réserves[4].

Cette distinction comporte donc une grande part de subjectivité puisqu'elle fait référence aux capacités technologiques et aux moyens financiers dont disposerons à l'avenir les compagnies pétrolières. Il est de toutes façons difficile de présumer des technologies d'exploration et de production pétrolière pour un horizon de plusieurs dizaines d'années. La frontière entre hydrocarbures conventionnels et non conventionnels évolue donc avec les progrès technologiques d'une part, et l'augmentation du prix du baril d'autre part. L'exemple des pétroles lourds et extra lourds du bassin de l'Orenoque au Vénézuela illustre bien ce phénomène. De 1967 à 1983, plusieurs centaines de Gbep sont passés du domaine non conventionnel au conventionnel. Nous reviendrons dans la partie suivante sur les propriétés évolutives de la notion de réserves. Dans tous les cas, les pétroles lourds et extra lourds, les sables bitumeux ainsi que les schiste bitumineux se rangent dans la catégorie des hydrocarbures non conventionnels.

Les hydrocarbures non conventionnels se caractérisent donc par des techniques de production différentes de celles utilisées pour l'exploitation plus classique du pétrole conventionnel. Le cadre de notre travail se limite à la production des hydrocarbures de types conventionnels et ne pourra être utilisé en dehors de ce cadre. En effet, notre modélisation s'appuie sur l'analyse des productions de bassins matures, c'est-à-dire dont les réservoirs sont bien connus et dont l'exploitation a débuté il y a plusieurs décennies. Celle-ci ne concerne donc que les seuls hydrocarbures considérés comme conventionnels aujourd'hui.

1.1.2 Gisements et réserves d'hydrocarbures

Dans un bassin pétrolier, les quantités d'hydrocarbures ne sont pas dispersées dans le sous-sol de façon homogène. Celles-ci sont concentrées dans des structures géologiques particulières appelées "pièges". Le champ (ou gisement) pétrolier peut être considéré comme une entité géologique élémentaire, supposée indivisible[5]. Nous verrons plus loin que cette organisation particulière distingue la production pétrolière d'autres productions de ressources primaires telles que le charbon ou l'uranium.

Les réserves d'un gisement pétrolier sont définies comme l'ensemble des ressources exploitables pour les besoins futurs et présents. Les réserves pétrolières sont donc une notion "dynamique". Celles-ci diminuent lorsque le champ est exploité, et augmentent si les techniques de production permettent d'augmenter le *taux de récupération* du gisement. Les quantités d'hydrocarbures contenues dans un gisement ne peuvent jamais être intégralement extraites du sous-sol. C'est pourquoi pour un même gisement, on parle

- de *réserves* pour les quantités d'hydrocarbures qui sont ou seront récupérables ;

[4]Une autre définition, assez répandue, considère le pétrole non conventionnel comme un pétrole produit ou extrait utilisant des techniques autres que la traditionnelle méthode de puits pétroliers.

[5]La réalité est souvent plus complexe. Ainsi, un même champ peut comporter plusieurs réservoirs, ou correspondre au regroupement de plusieurs petits champs.

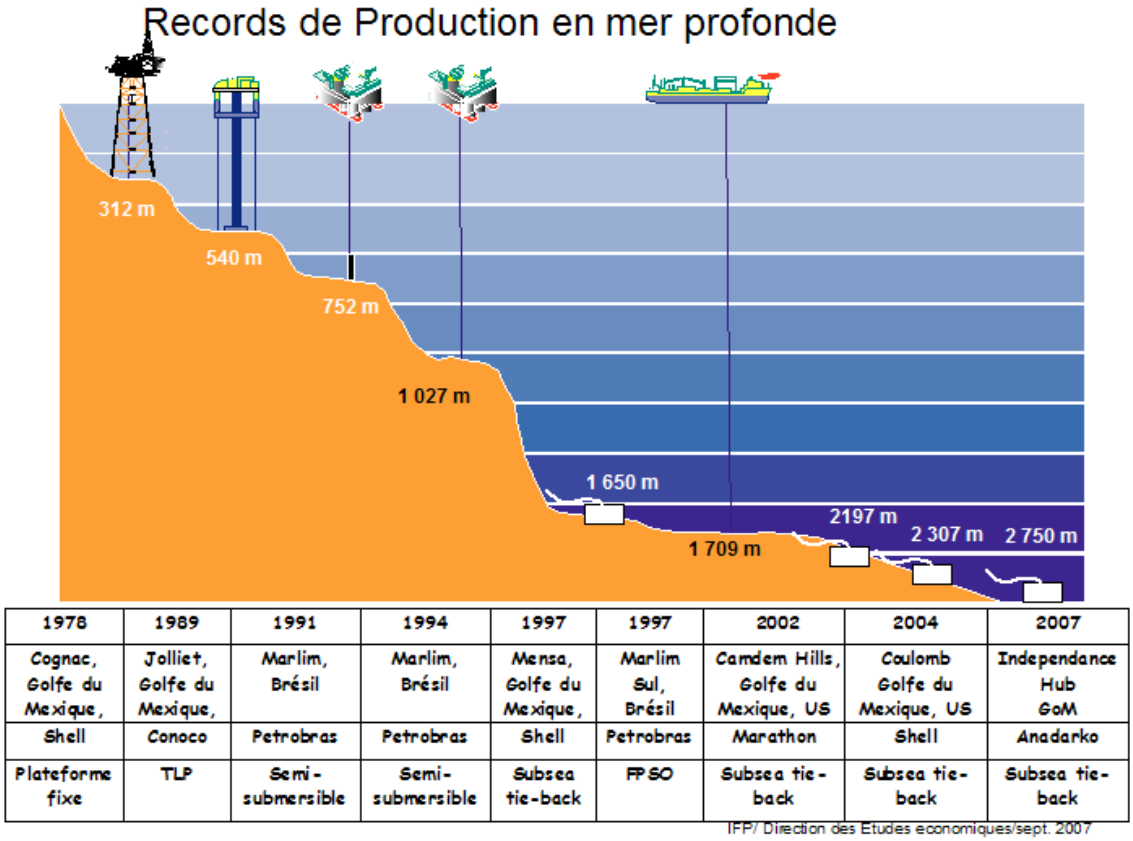


FIG. 1.1: Évolution des records de forages.

– de *ressources en place* pour les quantités d'hydrocarbures totales en place dans le gisement.

Le taux de récupération d'un gisement correspond ainsi au rapport des réserves par les ressources.

Après qu'un gisement a été découvert, les ressources de ce dernier sont d'abord évaluées. Les géologues et les économistes en déduisent alors les réserves en fonction des contraintes de production. Celles-ci peuvent être de natures différentes. Tout d'abord les limitations peuvent être imposées par la technologie. Il existe ainsi de nombreux gisements que l'on ne sait aujourd'hui pas mettre au production. C'est le cas notamment en exploitation "off-shore"[6] lorsque les profondeurs d'eau deviennent trop importantes. Comme le montre la figure 1.1, l'industrie pétrolière a progressivement réussi à exploiter des gisements situés à des profondeurs toujours plus grandes. Ceci illustre parfaitement le caractère évolutif de la notion de réserves pétrolières. La deuxième contrainte est d'ordre économique. En effet, si les coûts de production dépassent le total des recettes présumées des quantités d'hydrocarbures techniquement possibles à extraire, le champ ne sera évidemment pas produit.

Les géosciences permettent d'évaluer les réserves potentielles d'un gisement découvert. Du fait de l'incertitude importante que comportent ces estimations, celles-ci sont habituellement décrites en terme de distributions de probabilité. Ainsi, les réserves $2P$, $1P$ et $3P$ d'un gisement

[6]Exploitation en mer.

correspondent respectivement aux quantités de réserves que le champ a 50%, 90% et 10% de chances de dépasser. Un point de vue équivalent conçiste à distinguer les réserves prouvées des réserves possibles et probables avec la correspondance suivante :

- $1P$ = prouvées ;
- $2P$ = prouvées + probables ;
- $3P$ = prouvées + probables + possibles.

Ces définitions ont été fixées en 1997 par la SPE[7]. Mais en pratique, les organismes énergétiques et les compagnies pétrolières ne font pas toujours preuve d'une rigueur exemplaire. En particulier, il n'est pas toujours précisé à quel type de réserves ($1P$, $2P$ ou $3P$) correspondent leurs estimations publiées.

Cette définition de réserves est encore valable à l'échelle d'un bassin pétrolier, d'un pays ou de tout autre groupement de gisements d'hydrocarbures. Néanmoins, il est important de noter que les réserves d'un certain type ne peuvent être sommées de façon automatique. Par exemple, comme le souligne Lepez (2002, p.39), agréger des réserves $1P$ tend à sous-estimer les réserves du groupe alors que sommer les réserves $3P$ tend à les surestimer. Concernant les sommes de réserves $2P$, les deux biais sont possibles.

1.1.3 Réserves ultimes

Les réserves ultimes d'un gisement représentent l'ensemble des hydrocarbures qui en auront été extraits au terme de son exploitation. Dans la suite, nous emploierons aussi par abus le terme "taille" pour désigner les ressources ultimes d'un gisement. Avec le temps, la description géologique du gisement devient plus précise et les ressources ultimes sont alors mieux estimées. Les réserves prouvées correspondent donc à une estimation de la part non encore produite des ressources ultimes.

La définition des ressources ultimes est la même pour un bassin ou un pays. Á l'échelle d'un bassin pétrolier, les facteurs suivants sont responsables de réévaluations des ressources ultimes au cours du temps (Lepez, 2002, p.42) :

- les réévaluations des ressources en place pour chaque champ de façon individuelle,
- les nouvelles découvertes (nouveaux gisements),
- les progrès technologiques qui conduisent à améliorer les taux de récupération,
- l'impact de l'économie (un prix élevé du baril permet d'exploiter des champs plus coûteux à produire).

Proposer une modélisation permettant d'intégrer ces quatre facteurs pour estimer les réserves d'un bassin dans le futur est tout simplement trop ambitieux. Les deux derniers points, qui sont de nature technologique et économique, ne peuvent être pris en compte qu'à travers des scénarios de développement du bassin. Au-delà de quelques années, l'évolution des technologies n'est pas quantifiable. De même, il serait peu sérieux d'avancer des niveaux de prix du baril, ne serait-ce que pour les prochains mois. Les deux premiers points renvoient à la nature géologique du bassin. Dans un bassin suffisamment mature, les réservoirs sont généralement bien connus ; les réévaluations de ressources individuelles sont en réalité assez limitées et les réserves prouvées des champs déjà découverts peuvent être utilisées pour estimer les ressources

[7]Society of Petroleum Engineers.

ultimes des gisements connus. Concernant les nouvelles découvertes, les travaux de Lepez (2002) permettent d'estimer le nombre de champs non encore découverts dans un bassin.

Pessimistes et optimistes

Dans le débat sur l'estimation des réserves pétrolières, deux écoles de pensée radicalement opposées s'affrontent. L'école des pessimistes, tout d'abord, rappelle que les ressources d'hydrocarbures ne sont pas disponibles en quantité infinie dans le sous-sol. Et puisque notre consommation augmente avec le temps, notamment avec la forte croissance économique de certains pays comme la Chine et l'Inde, cette situation conduit naturellement à la raréfaction des réserves et à une augmentation du prix des hydrocarbures. Certes, le baril a atteint ces cinq dernières années des niveaux de prix records. Mais il est difficile d'évaluer dans quelle proportion la peur de la raréfaction des réserves a contribué à cette augmentation. En effet, de nombreux autres facteurs influencent le marché des hydrocarbures. De trop faibles capacités de production et de raffinage, la crainte de tensions politiques au Moyen-Orient, et d'importants phénomènes de spéculation sont principalement responsables de ces prix élevés. D'autre part, les pessimistes soulignent que les progrès technologiques ne signifient pas nécessairement une augmentation des réserves. Selon eux, les nouveaux procédés d'extraction conduisent aussi à produire les mêmes quantités de réserves en une durée plus courte. Dans ce cas, les progrès technologiques accélèrent encore plus la raréfaction des ressources !

L'école des optimistes est principalement composée d'économistes et d'industriels. Elle s'appuie sur le principe de la création de nouvelles réserves grâce aux progrès technologiques. Les périodes de prix élevés des hydrocarbures permettent, grâce aux fonds dégagés, le développement de nouvelles procédures de productions. Certains hydrocarbures passent ainsi du conventionnel au non conventionnel, et les prix ne peuvent atteindre ainsi des niveaux trop élevés. Selon ce point de vue, le marché des conventionnels n'est pas fermé puisque d'autres ressources deviennent disponibles au cours du temps, de façon automatique, avec la raréfaction des ressources conventionnelles. Ce mécanisme s'appelle le continuum du carbone fossile (Babusiaux, 2005).

Ces deux points de vue s'accordent au moins sur un point : les hydrocarbures produits à l'avenir seront de plus en plus difficiles à extraire. Des investissements importants sont aujourd'hui nécessaires, que ce soit pour produire des hydrocarbures non conventionnels pour les uns , ou pour développer de façon conséquente d'autres types d'énergie pour les autres.

1.2 Exploration et production d'un bassin pétrolier

Une dizaine d'années de travaux préparatoires est nécessaire avant que la production d'une zone vierge puisse réellement commencer, et tout débute par une campagne d'exploration. Cette nécessaire phase de prospection s'est progressivement accompagnée de moyens technologiques de plus en plus sophistiqués, mais la démarche reste la même : à partir des observations faites en surface, les géologues doivent extrapoler l'organisation de la roche et localiser ainsi les pièges où les hydrocarbures sont potentiellement retenus. Après 1945, la photographie aérienne est utilisée pour repérer les anticlinaux aux États-Unis. Rapidement

les méthodes de la géophysique ont permis d'améliorer encore l'étude du sous-sol. La sismique de réflexion est aujourd'hui la technique la plus couramment utilisée, elle consiste à envoyer des ondes élastiques qui se propagent dans le sous-sol et dont l'étude des échos permet de repérer les discontinuités de la roche. Elle peut être à 2, 3 (et même 4) dimensions, la sismique 3D étant surtout utilisée pour les campagnes de prospection en mer. Les géologues effectuent aussi des forages d'exploration[8] pour sonder le sous-sol. Aujourd'hui tous les pièges facilement visibles ont été localisés, et ces technologies plus perfectionnées permettent d'atteindre des gisements plus difficilement accessibles.

Les différents bassins pétroliers de la planète ont été mis en exploitation de façon progressive, en fonction de la demande mondiale et de la difficulté à extraire le pétrole dans la zone considérée. Dans certaines régions, les anticlinaux où le pétrole était piégé étaient faciles à repérer, c'est par exemple le cas aux États-Unis et au Moyen Orient où la production a pu débuter il y a déjà longtemps. Le Mexique et la Roumanie sont eux aussi de "vieux" pays producteurs. En revanche l'exploitation off-shore en Mer du Nord ne date que du début des années 1970.

L'accessibilité des hydrocarbures est donc un facteur essentiel de leur exploitation. Les politiques fiscales menées par les États influencent aussi de façon significative l'exploration et la production pétrolières. En effet, lorsqu'une compagnie pétrolière souhaite rechercher des hydrocarbures dans une zone vierge, elle doit tout d'abord obtenir l'autorisation d'effectuer une campagne d'exploration car excepté aux États-Unis, le sous-sol est la propriété des États. Dans le cadre du régime de *concession*, l'État accorde des permis d'exploration, et s'il y a découverte, des permis d'exploitation pour une zone bien définie. Les États peuvent eux aussi lancer des campagnes d'exploration supportées par leurs compagnies nationales.

Il existe un principe fondamental en exploration pétrolière : "Big stuff gets found first" : les plus gros gisements, qui sont aussi les plus visibles, sont généralement découverts en premier. Il est en effet naturel de trouver d'abord les structures les plus étendues. De plus, les géologues doivent justement découvrir des gisements de taille suffisante pour que la zone puisse être déclarée rentable. D'autre part, le développement d'une zone pétrolière est une opération financière très ambitieuse, et seules les grandes compagnies sont capables de les assumer. Si la phase d'exploration n'est pas la plus coûteuse, elle pèse en revanche de façon conséquente sur les comptes des entreprises ; une campagne d'exploration est un investissement très lourd. Par conséquent, pour revenir rapidement sur leur investissement, les compagnies ont intérêt à développer en premier lieu les plus gros gisements pour éviter une catastrophe financière. Pour ces deux raisons, la production d'un bassin débute généralement par le développement des champs de tailles les plus importantes relativement aux réserves disponibles dans la zone. Ainsi en mer du nord, les gisements de Brent, Ninian, Forties, Oseberg contiennent une grande part des réserves du bassin, et leur exploitation a débuté dès les années 1970.

Une fois que des champs de réserves rentables ont été identifiés, la production du bassin peut donc commencer, supportée par les productions croissantes des champs les plus volumineux. Après quelques années, celles-ci atteignent leur maximum, avant de décliner bientôt. Les compagnies commencent alors à exploiter des gisements plus modestes qui se situent autour

[8] Ces puits sont appelés "wild cat" dans le jargon des pétroliers.

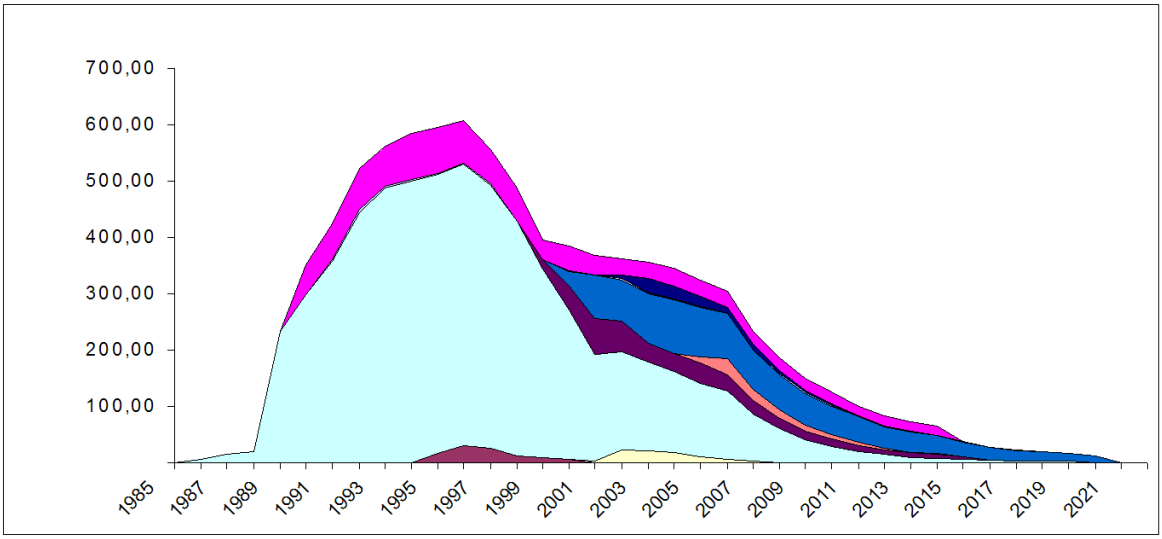


FIG. 1.2: Agrégation des productions (en Mbpd) des champs de la zone Oseberg en mer du nord.

des zones déjà en production (voir figure 1.2). En production off-shore, le réseau d'oléoducs s'organise donc en grappe autour des champs les plus importants, et la proximité d'une infrastructure pétrolière existante est un facteur essentiel dans le choix du développement d'un champ.

La mise en production de nombreux petits champs ne suffit pas forcément à compenser le déclin de la production des plus gros. Cette observation peut se justifier de façon plus rigoureuse en s'appuyant sur la distribution de la taille des champs. Le niveau de production des plus grands gisements constitue donc un indicateur pertinent de l'état de "maturité" d'un bassin. Un exemple récent illustre ce processus : en 2005 le champ Burgan au Koweit, qui contribue à la majeure partie de la production du pays, a entamé son déclin. Ce champ est considéré comme le second en taille sur l'ensemble de la planète derrière Gnawar, le célèbre champ saoudien. La Kuwait Oil Company a aussitôt déclaré vouloir augmenter la production d'autres champs pour compenser le déclin. Mais de nombreux experts doutent aujourd'hui de la capacité du Koweit à compenser la baisse de production de Burgan[9].

1.3 La courbe de Hubbert : une controverse

Bien que la production d'hydrocarbures constitue le pilier fondamental du développement industriel et économique mondial, il existe finalement assez peu de modèles de production pétrolier pour décrire le développement d'un bassin exploité. Le modèle le plus populaire mais aussi le plus contesté est celui de la *courbe de Hubbert*. Nous rappelons dans cette partie les principaux éléments historiques, économiques et géopolitiques qui animent depuis plusieurs décennies défenseurs et détracteurs de la courbe de Hubbert. Notons qu'il existe d'autres modèles pour la production pétrolière qui sont de nature économétriques. Moroney

[9]Cette nouvelle inattendue inquiète certains observateurs pour qui la même situation pourrait se produire bientôt en Arabie Saoudite, où l'essentiel de la production provient d'une poignée de *super géants* parmi lesquels Ghawar, Abqaiq, Safaniya et Berri, voir par exemple Simmons (2005).

et Berg (1999), et Kemp et Kasim (2003) étudient ainsi la production à l'aide de séries temporelles. Citons aussi les travaux de Cleveland et Kaufmann (1991) et de Kaufmann (1991) qui réconcilient le point de vue économétrique et les ajustements proposés par Hubbert.

1.3.1 King Hubbert et la théorie du déclin

King Hubbert est né le 5 octobre 1903 dans le Texas, où il passera toute son enfance. Il étudie à l'Université de Chicago la géologie mais aussi les mathématiques et la physique. Tout en préparant sa thèse de doctorat qu'il obtient en 1937, il travaille comme géologue notamment pour l'USGS [10] et l'Amerada Petroleum Corporation. En 1943 il est engagé par la compagnie Shell en tant que géophysicien, et quelques années plus tard il y devient directeur de la recherche en exploration-production. Il quitte Shell en 1963 pour se consacrer à la recherche au sein de l'USGS et d'universités telles que Berkeley et Stanford .

Lorsque Hubbert publie le texte *Nuclear energy and the fossil fuels* (Hubbert, 1956), il est déjà un géophysicien reconnu, spécialiste de sujets tels que l'écoulement des fluides, ou la résistance des roches sous pression. Dans cet article, il extrapole les productions nord-américaine et mondiale d'hydrocarbures à l'aide d'une courbe en cloche, la fameuse *courbe de Hubbert*. Par des méthodes graphiques, et en s'appuyant sur différentes estimations des réserves ultimes de pétrole et de gaz, il peut ainsi prévoir que le maximum de la production nord-américaine sera bientôt atteint. Selon l'un de ses scénarios, le pic de pétrole (*peak oil* en anglais), se produirait autour de 1970. Or, le secteur de l'industrie pétrolière crée déjà des profits considérables et annoncer le déclin ne peut être qu'impopulaire : Hubbert rencontre alors beaucoup d'hostilité de la part du milieu des affaires et du pétrole. Mais il est vrai aussi qu'avant lui, de nombreux experts s'étaient aventurés à annoncer à tort l'épuisement des réserves. En dépit de la meilleure qualité de ses estimations sur les réserves, il passe d'abord pour un énième prophète de l'apocalypse...

Au début des années 60, Hubbert rédige aussi un rapport pour le président Kennedy où il alerte les autorités sur le caractère épuisable des réserves. En effet, Hubbert est préoccupé par l'explosion de la consommation énergétique mondiale. Dans ce document, il s'attache à évaluer les quantités d'énergie encore disponibles à l'humanité avant que les différentes sources ne soient toutes épuisées ; une question aujourd'hui toujours aussi polémique. Selon lui, la croissance de la population et la consommation énergétique mondiale seront insoutenables à un tel rythme :

> "Human population growth is like nothing that has happened in all of geological history."

Il serait donc injuste de limiter la contribution de King Hubbert à la seule courbe qui porte désormais son nom, non seulement du fait des travaux scientifiques évoqués plus haut, mais aussi parce qu'il s'efforça de montrer, à raison, que l'approvisionnement énergétique deviendrait une problématique fondamentale des sociétés industrialisées.

Après 1971, la production des États-Unis décline effectivement ; heureusement l'Arabie Saoudite parvient très rapidement à développer de gigantesques capacités de production. En

[10] United States Geological Survey.

février 1975 un rapport de la National Academy of Sciences confirme les calculs de Hubbert sur la production et l'estimation des réserves nationales.

Depuis les années soixante-dix, la courbe de Hubbert est devenu populaire au delà des spécialistes de l'économie et de la géologie pétrolière. Plusieurs raisons peuvent être avancées pour expliquer la popularité de cette méthode. D'abord, le fait que la prédiction d'Hubbert pour la production des États-Unis se soit avérée exacte à un an près en est sans doute en partie responsable. De plus, la simplicité de la technique utilisée, un banal ajustement, a probablement aussi séduit de nombreux spécialistes à la recherche de calculs de prévisions faciles à mettre en oeuvre. Enfin, celle-ci dépend à la fois fortement et uniquement de l'estimation des ressources ultimes. Cette technique s'est donc automatiquement retrouvée en centre du débat entre pessimistes et optimistes au sujet des ressources ultimes. Le fait que de nombreux pessimistes soient des adeptes de l'utilisation de la courbe de Hubbert a probablement contribué à assimiler la courbe de Hubbert à cette communauté. Il est pourtant essentiel de souligner que ces deux questions sont de natures différentes : le débat sur les réserves ne doit être confondu avec celui de la modélisation de la production.

Après Hubbert

Récemment, Deffeyes (2001) a repris les méthodes de Hubbert pour estimer la date du pic mondial de production. Deffeyes préfère ajuster la production sur une gaussienne, car selon lui l'ajustement est meilleur[11]. Nous reviendrons sur la critique de ces différentes hypothèses dans la section suivante. Il obtient ainsi numériquement le meilleur ajustement possible pour des scénarios de 1800 et 2100 milliards de barils de réserves ultimes d'huile. Dans le premier cas, le maximum de la production se produit en 2003, et 2009 pour le second, mais notons que tout ceci suppose que la courbe de production a une forme gaussienne.

Laherrère (2003) et Campbell (2002) proposent eux aussi des scénarios de production qui s'inspirent des idées de Hubbert. En s'appuyant sur des propriétés graphiques de la fonction logistique, ils estiment les réserves et la date du pic de production par des ajustements linéaires. Les profils de production de certains pays, par exemple ceux de la France, présentent plusieurs pics de productions, et ceci ne rentre pas dans le cadre de la modélisation de Hubbert. Pour tenir compte des différentes phases de production d'un même pays, Laherrère (1997) propose de modéliser les scénarios de ce type par la superposition de plusieurs courbes de Hubbert correspondant à des cycles de production différents.

Dans un article de 2002, Bentley (2002) s'appuie sur les travaux de Laherrère et propose quelques idées pour justifier l'ajustement de "courbes à la Hubbert" sur les productions d'hydrocarbures. Il cherche notamment à mettre en évidence une courbe caractéristique pour l'agrégation de profils de production de champs de tailles décroissantes et lancés à intervalles réguliers.

[11] L'auteur ne le précise pas, mais il s'agit probablement de meilleur pour le critère des moindres carrés.

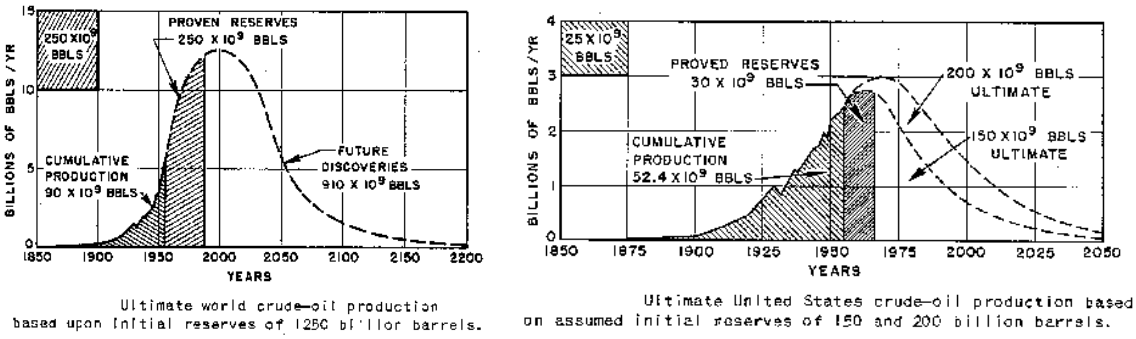


FIG. 1.3: Courbes de production de Hubbert mondiale (à gauche) et pour les États-Unis (à droite), tracées à la main par l'auteur dans Hubbert (1956).

1.3.2 Aspects mathématiques de la méthode de Hubbert

Dans son article de 1956, Hubbert propose un ajustement graphique pour prolonger le profil de production des États-Unis, ainsi que celui pour la production mondiale. Mais peu d'éléments sont donnés sur les justifications et les détails de la méthode. Comme le remarque Deffeyes (2001), Hubbert suppose uniquement les deux hypothèses suivantes :

- la courbe de production a la forme d'une cloche,
- la courbe est symétrique par rapport au pic.

En effet, Hubbert n'en dit pas plus dans ce premier article, l'aire située en dessous de la courbe devant uniquement traduire la taille des réserves ultimes, c'est-à-dire les réserves déjà consommées, présentes et restant à découvrir. En revanche, dans le rapport de 1962 (Hubbert, 1962), il détaille et justifie l'utilisation d'une courbe logistique en s'appuyant sur un modèle d'évolution de population. Ce dernier texte nous paraît beaucoup plus fondamental que celui de 1956 car Hubbert y expose plus que de simples calculs sur les réserves et les rythmes de production. Il s'agit d'une réflexion sur la consommation énergétique et les limites imposées par le sous-sol. Il y passe en revue les différents types d'énergie qui pourraient à l'avenir se substituer aux hydrocarbures, une problématique qui nous est aujourd'hui très familière...

Les scénarios de Hubbert reposent sur une estimation des réserves ultimes. Concernant le cas des États-Unis, il se réfère au rapport de deux géologues de l'époque : Weeks et Pratt. Hubbert remarque tout d'abord que les courbes de découvertes et de productions cumulées sont comparables, et qu'elles sont séparées d'un décalage Δ, comme l'illustrent les deux graphiques de la figure 1.3.2. Et puisque l'historique des découvertes est plus avancé que celui de la production, l'ajustement portera donc sur le premier.

Pour dresser ces courbes théoriques, Hubbert fait le choix de modélisation suivant :

> "Growth phenomena such as those(...), which start slowly, gradually accelerate, and finally level off to a maximum, are said to follow a logistic growth curve and described by an empirical equation of this form
>
> $$y(x) = \frac{h}{1 + a * e^{-bx}}\ ''.$$

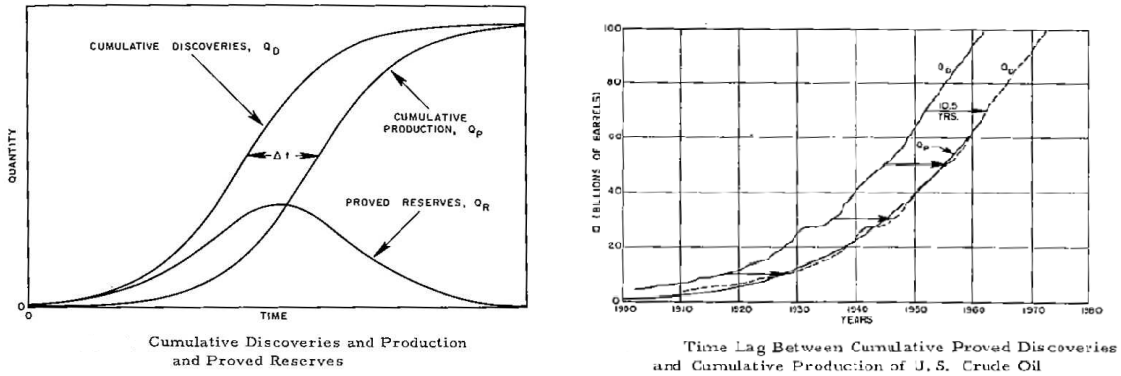


FIG. 1.4: Décalages entre les découvertes cumulées et les productions cumulées. La figure de gauche est une représentation théorique du décalage et la figure de droite correspond aux données des États-Unis.

L'équation logistique

$$\dot{y} = r * y(1 - y) \tag{1.1}$$

définit effectivement un modèle de population élémentaire. Selon celui-ci, la croissance de la population est proportionnelle au produit de la taille de la population par les ressources disponibles. Pour les découvertes cumulées, et en passant au logarithme :

$$\ln\left(\frac{Q_\infty}{Q_D(t)} - 1\right) = ln(a) - bt$$

où Q_∞ et $Q_D(t)$ représentent respectivement les découvertes cumulées jusqu'à l'instant t, et les réserves ultimes du bassin. D'où l'importance d'estimer correctement Q_∞ pour ajuster une droite sur la série temporelle $l = \left(\ln(\frac{Q_\infty}{Q_D} - 1)\right)$ et estimer ainsi les paramètres a et b. Par cette méthode, Hubbert trouve :

$$Q_D(t) = \frac{170 * 10^9}{1 + 46.8 * e^{-0.687*(t-1900)}}.$$

Et la production cumulée s'obtient par un décalage dans le temps de la fonction précédente :

$$Q_P(t) = \frac{170 * 10^9}{1 + 46.8 * e^{-0.687*(t-1910.5)}}.$$

La production est ensuite une simple dérivée de Q_P dont le maximum se calcule aisément. Tout dépend donc de la qualité de l'estimation des réserves ultimes. Hubbert utilise finalement un intervalle de confiance de 150Gb à 200Gb pour l'ensemble des réserves des États-Unis[12]. Aujourd'hui, on pense que la borne haute est proche de la réalité, et c'est justement ce scénario qui donne un pic pour 1971.

Hubbert applique aussi sa méthode à l'échelle mondiale : il trouve un pic pour l'année 2000 (voir figure 1.3.2). Pour cela, il estime les réserves ultimes de la planète à environ 1250 milliards de barils, or aujourd'hui, les géologues pensent qu'elles en atteindraient le double.

[12]hors Alaska.

1.3.3 Critique de la méthodologie de Hubbert

Il existe une littérature abondante et passionnée sur le pic de pétrole. L'ajustement graphique de Hubbert nécessitant une bonne connaissance des réserves ultimes, il en résulte une certaine confusion entre la méthode de Hubbert et le problème de l'état des réserves qui préoccupe les écologistes mais aussi de nombreux experts du monde pétrolier. Comme nous l'avons déjà remarqué, ces deux questions sont souvent mélangées à tort, il nous importe ici d'analyser la technique de Hubbert, sans pour autant rentrer dans le débat classique qui oppose les deux points de vue antagonistes des optimistes et pessimistes sur l'état des réserves mondiales.

La méthode de King Hubbert a le mérite d'être facile à mettre en oeuvre ; il suffit d'ajuster une gaussienne ou une courbe logistique sur les données disponibles. Cependant, aucun élément propre à la dynamique de la production pétrolière ne justifie une telle modélisation. D'ailleurs, Hubbert le reconnaît lui-même : son équation est empirique ; l'article de 1956 ne la mentionne même pas... Concernant le modèle logistique, l'analogie avec les théories de croissance de population est assez peu crédible ; quelle interprétation donner à l'équation logistique (1.1) ?

Pour expliquer la forme gaussienne de certains profils de bassin, le théorème de la Limite Centrale est parfois évoqué. La fonction de distribution représenterait ici le profil de production d'un champ, et la fonction de distribution de la somme des phénomènes serait proche d'une fonction gaussienne. Comme le souligne Lepez (Babusiaux et al., 2002), l'argument ne tient pas :

> "La fonction de distribution de la somme des phénomènes n'est pas la somme des fonctions de distributions[13]. Par ailleurs, le phénomène Hubbert ne s'inscrit pas dans le cadre d'application de ce théorème. En effet, d'une part, les profils que l'on somme ne sont évidemment pas indépendants, surtout dans une même zone géographique, et d'autre part, ce théorème concerne des distributions numériques et non temporelles, comme le sont celles du modèle de Hubbert. Il n'est donc pas plus justifié d'ajuster des courbes gaussiennes sur la production du bassin."

La deuxième hypothèse de Hubbert nous semble encore plus contestable. La courbe de production n'a aucune raison d'être symétrique. Tout d'abord, les profils de production individuels[14] sont fortement dissymétriques. Mais surtout, le processus de mise en production des champs ne présente aucune symétrie car un champ sera développé d'autant plus vite que ses réserves sont importantes. Il n'est pas exclu qu'au finale, les contributions de ces différents phénomènes aboutissent à une courbe de production symétrique, mais aucune justification évidente ne transparaît de la dynamique de mise en production du bassin. Au moment où la production des plus vieux et plus gros champs commence à décliner, il faudrait que les champs restants soient mobilisés exactement de façon de à ce que la courbe cumulée respecte la symétrie. Une telle situation n'a aucune raison de se réaliser dans la pratique ; les compagnies pétrolières cherchent avant tout à repousser le pic de production et à limiter le déclin.

On peut toutefois se demander dans quel cadre il est le plus approprié - ou le moins déraisonnable - d'utiliser des courbes de Hubbert. Les adeptes de cette méthode ont l'habitude

[13] La fonction de distribution de la somme est en fait la convolution des fonctions de distribution individuelles.

[14] Nous entendons par profil de production individuel la courbe de production d'un seul gisement, par opposition au profil de production d'une zone géographique, qui est une agrégation de productions de gisements.

de présenter des profils de production par pays. Or, la production d'un seul pays peut mélanger les contributions de zones qui n'ont pas été développées à la même époque, et qui n'ont pas été nécessairement exploitées avec la même intensité. Par exemple dans le Golfe du Mexique, le développement off-shore très profond est très récente et ne peut être comparé avec la production du début du siècle dans ce même pays. Il nous paraît plus justifié de cumuler les profils de production à l'échelle de bassins géologiques, en distinguant de plus les différents types d'hydrocarbures, surtout si leurs productions ont débuté à des époques différentes.

Enfin, comme le soulève Ryan (2003), l'estimation des réserves ultimes Q_∞ est très fluctuante au cours du temps. En effet, nous avons vu plus haut que les réserves pétrolières sont une notion dynamique évoluant avec la connaissance du réservoir et la capacité à extraire l'huile qui y est piégée. L'évaluation des réserves ultimes, que l'on ne connaîtra qu'une fois la zone totalement exploitée, est un travail très délicat. Et du fait des progrès technologiques et des efforts de l'exploration, les estimations ont tendance à croître avec le temps. Ce biais systématique est impossible à corriger. Cette remarque ne concerne pas uniquement la méthode de Hubbert. Toute modélisation, et c'est le cas de notre travail, s'appuyant sur une étude probabiliste de la distribution de la taille des champs rencontre nécessairement cette difficulté.

1.4 Le pic de production mondial

L'une des questions les plus polémiques de l'économie pétrolière est l'estimation du maximum de production mondial. Nous rappelons ici les enjeux principaux de ce débat.

1.4.1 Définitions et impact du pic de production

La majorité des experts pétroliers s'accorde sur le scénario général suivant. Dans un futur plus ou moins éloigné selon les avis, la production mondiale devrait atteindre son maximum, avant de décliner progressivement. Ce maximum de la production est couramment appelé *peak oil* (PO)[15], et son estimation est devenu un sujet de polémique bien au-delà du cercle des économistes et des géologues. Le débat sur l'état des réserves et l'imminence du *peak oil* mondial n'est en réalité pas récent, puisqu'Hubbert l'avait en quelque sorte amorcé avec les questions soulevées par ses travaux. Mais depuis 2002 et la flambée des cours du pétrole, la polémique a repris de plus belle.

Il est important de noter que le passage du pic de production ne signifie pas que les ressources sont épuisées. Cependant, dans le contexte mondial actuel où la demande en hydrocarbures est de plus en plus difficile à satisfaire, une chute de la production aurait des effets dévastateurs sur l'économie mondiale. En effet, aucune autre forme d'énergie n'est en mesure aujourd'hui de se substituer facilement aux hydrocarbures. On évoque souvent l'impact

[15]On trouve parfois dans des documents de vulgarisation sur la production pétrolière que le PO est aussi le moment où la moitié des ressources ont été consommés, comme c'est le cas pour la courbe de Hubbert, bien que rien ne justifie une telle propriété. Citons à titre de mauvais exemples la page "wikipedia" sur le *peak oil*, ou encore l'article de presse publié dans journal *Le Monde* du 26 octobre 2007 et intitulé "La moitié des réserves pétrolières seraient épuisées"

sur le domaine des transports, mais en réalité, beaucoup d'autres secteurs seraient directement affectés. Par exemple, l'industrialisation de l'agriculture, *la révolution verte*, qui a permis d'améliorer considérablement les rendements de production et de nourrir une population mondiale en forte expansion, repose sur une grosse consommation en hydrocarbures (fertilisation des sols, machines agricoles, transport...). Si la production mondiale venait à décliner significativement, la population mondiale devrait faire face à une crise alimentaire majeure. On comprend mieux dès lors que ce sujet soit aussi polémique.

Au début de l'année 2005, une étude complète (Hirsh *et al.*, 2005) sur la capacité de la technologie à atténuer l'impact du pic de production, a été remise au Département de l'Énergie des États-Unis. Le rapport souligne que la date du pic est difficile à évaluer et analyse donc plusieurs scénarios vraisemblables. Ses conclusions sont les suivantes : les politiques de développement de nouvelles technologies doivent être amorcées longtemps avant le passage du pic. Deux décennies seraient nécessaires pour réussir la transition énergétique qui fait suite au pic de production. Et selon cette étude, attendre le passage du pic pour amorcer une transition énergétique aurait des effets dramatiques sur l'économie mondiale.

1.4.2 Prévisions du pic mondial

On l'a déjà dit, peu d'outils théoriques sont disponibles pour modéliser la production des hydrocarbures. Il existe pourtant une profusion de scénarios sur la date du *peak oil* ; le tableau 1.1 présente une synthèse des dates avancées par les personnes et les organismes les plus actifs sur cette question. Le tableau montre à quel point les avis sont partagés sur l'imminence du *peak oil*. Prévoir le pic mondial nécessite une bonne connaissance de la géologie des bassins, des techniques de production, de l'économie pétrolière et de la géopolitique mondiale ; un calcul bien délicat en vérité. La majorité des estimations avancées ne reposent pas sur un travail de modélisation. Pour la plupart, elles sont le résultat d'une difficile collecte d'informations au niveau mondial sur les réserves, les productions, et les projets de développement. En effet, il paraît impossible de proposer une modélisation de la production au niveau mondial tant celle-ci est dépendante de la demande dans les pays émergents, des événements géopolitiques et des décisions politiques...

Fondée par Campbell, l'*Association for the Study of the Peak Oil* (ASPO) est très active dans le débat sur le *peak oil*. Parmi ses membres, on trouve notamment le français J. Laherrere. Les prévisions des membres de l'ASPO annoncent le passage du pic avant douze ans.

Simmons est un banquier et spécialiste de l'industrie pétrolière. Il est lui aussi persuadé que la production mondiale est maintenant proche du pic de production (Simmons, 2005). On considère que les deux tiers des réserves mondiales en huile conventionnelle seraient localisées dans le sous-sol du Moyen-Orient, dont la majeure partie en Arabie Saoudite. Simmons rappelle que le régime saoudien entretient l'opacité sur l'état de ses réserves et ses niveaux de production. Depuis des années, l'Arabie Saoudite revendique le rôle de pays ajusteur de la production mondiale (*world's oil swing producer*). Or les puits de forage ne sont pas de simples robinets que l'on peut ouvrir ou fermer sans conséquences, et les saoudiens auraient excessivement malmené leurs gigantesques réservoirs. Les taux de récupération seraient désormais plus faibles que prévu et l'Arabie Saoudite serait incapable de produire les quantités

Personne ou organisme	Estimations du PO
Deffeyes, K.[16]	2005
Bakhtiari, S.[17]	2007
Simmons, M.R. [18]	2007-2009
Bentley, R.[19]	$\approx$ 2010
Campbell, C. [20]	2010
Skrebowski, C.[21]	2010 +/- 1 an
Pang, X. and al.[22]	$\approx$ 2012
Lahererre, J.[23]	2010-2020
Wood Mackenzie [24]	$\approx$ 2020
Total	$\approx$ 2020
CERA[25]	bien après 2030
Brown, J. [26]	Impossible à prédire
ExxonMobil	Pas de pic attendu
Lynch, M.[27]	Pas de pic attendu
OPEP [28]	Récuse la théorie du *peak oil*

TAB. 1.1: Prédictions du *peak oil* mondial (Sources : rapport Hirsch (2007) pour le *U.S. Department of Energy.*)

[16]Professeur de Princeton à la retraite, ancien géologue de la compagnie Shell.
[17]Expert de l *Iranian National Oil Corporation.*
[18]Investisseur.
[19]Universitaire et analyste énergétique.
[20]Géologue à la retraite des compagnies Texaco et Amoco.
[21]Éditorialiste de la revue *Petroleum Review.*
[22]*China University of Petroleum.*
[23]Géologue à la retraite de la compagnie Total.
[24]Entreprise de conseil énergétique.
[25]*Cambridge Energy Research Associates* : entreprise de conseil énergétique.
[26]Expert de la compagnie PB CEO.
[27]Économiste de l'énergie.
[28]Organisation des pays exportateurs de pétrole.

nécessaires pour répondre à l'augmentation de la consommation mondiale.

D'autres experts ne sont pas de cet avis : l'école des optimistes, essentiellement formée d'économistes, affirme que le pic ne se produira pas avant au moins vingt ans. Selon Adelman et Lynch (2003) du M.I.T., les capacités d'innovation de l'industrie continueront de permettre le développement des nouvelles réserves. Ainsi, depuis les débuts de l'exploitation pétrolière, celles-ci ne cessent de croître grâce aux progrès des techniques qui permettent d'accéder à de nouvelles réserves et d'améliorer les taux de récupération. Lynch remarque de plus que la date du pic annoncée par les optimistes recule sans cesse à mesure que les réserves croissent.

Le CERA (*Cambridge Energy Research Associates*), une entreprise de conseil du groupe IHS, a, quant à elle, publié un rapport (Cera, 2005) selon lequel le maximum de la production mondial ne surviendrait pas avant 2030, et il précéderait un plateau plutôt qu'une phase de décroissance prononcée. On mesure l'écart entre les différents point de vue.

Depuis quelques années, on assiste à un resserrement des estimations du *peak oil* autour des années 2020. Les pessimistes repoussent leurs prévisions, et de leur côté, les optimistes considèrent désormais plus sérieusement l'estimation du pic. Pessimistes et optimistes s'accordent au moins sur une chose : l'ère du pétrole prendra fin pendant le XXIème siècle. Le débat porte plus sur la crédibilité de certaines données de réserves d'une part (voir Bentley, 2006), et d'autre part sur la capacité de l'industrie pétrolière à développer à temps les technologies permettant de basculer les hydrocarbures non conventionnels dans le domaine du conventionnel. Si les réserves de pétrole conventionnel sont assez importantes pour que leur exploitation dure suffisamment longtemps, permettant ainsi la mise au point des technologies indispensables à l'exploitation du non conventionnel, une crise majeure sera évitée. Au contraire, si les réserves disponibles s'épuisaient trop rapidement, il n'y aura pas de relais possible par un nouveau type d'énergie, un scénario catastrophe défendu par une partie des pessimistes.

1.5 Objectifs et hypothèses générales de cette thèse

Les remarques précédentes montrent qu'il n'est pas raisonnable de construire une modélisation de la production mondiale. En particulier, nous n'avons pas pour intention de proposer encore une nouvelle prévision de la date du *peak oil*. En revanche, à l'échelle du bassin pétrolier, nous verrons qu'une modélisation de la production est cette fois possible. Nous souhaitons développer une modélisation générale pour la production de pétrole dans bassin d'hydrocarbures de façon à apporter des éléments de réponses aux questions suivantes, qui ne font pas l'objet d'un consensus dans le secteur pétrolier :

- La courbe de production est elle symétrique ?
- Quels sont les facteurs qui déterminent le plus la forme de la courbe de production du bassin ?
- Est-il possible de limiter le déclin de la production une fois franchi le maximum de la production ?

Ce modèle à définir doit aussi nous permettre de proposer des prolongements de la production dans un bassin en cours d'exploitation.

Un modèle de production pétrolier doit être capable de décrire l'historique complet de l'exploitation d'un bassin d'hydrocarbures et les quantités produites au final sont donc exactement les ressources ultimes tels que nous les avons définies plus haut. Plus précisément, nous considérons les ressources ultimes des hydrocarbures aujourd'hui qualifiés de "conventionnels", en nous restreignant au cas de l'huile[29].

Afin de construire un modèle complet de la production d'un bassin, il nous faut comprendre comment les champs sont découverts et exploités à l'intérieur de celui-ci. Mais surtout, il est essentiel de choisir une distribution de probabilité correcte pour modéliser la taille des gisements pétroliers non seulement pour l'estimation des réserves, mais aussi pour décrire fidèlement l'exploration et la production des gisements car ces phénomènes sont fortement conditionnés par la taille des gisements. Le chapitre suivant est consacré au choix d'une loi de probabilité pour les réserves d'un bassin.

[29]Nous n'avons pas étudié spécifiquement la production de gaz, mais il est vraisemblable que cette production puisse être traitée selon une modélisation similaire à celle qui est présentée dans les chapitres qui suivent.

Chapitre 2

Modèle probabiliste pour la formation des réserves

Dans ce chapitre, nous proposons un modèle probabiliste permettant de décrire la formation des réserves pétrolières. La construction de ce modèle vise à justifier l'utilisation des distributions de Lévy-Pareto pour modéliser les tailles des gisements d'un bassin pétrolier. Le point de vue de ce chapitre est avant tout qualitatif : il s'agit de trouver un modèle probabiliste simple, c'est-à-dire comportant peu de paramètres, rendant compte de la structure générale de la taille des gisements au sein d'un même bassin d'hydrocarbures.

La première section rappelle d'abord les mécanismes géologiques qui aboutissent à la création de réserves d'hydrocarbures dans un bassin. Elle propose ensuite une description de la famille des tailles des gisements pétroliers du bassin en terme de distributions de probabilité. Nous parvenons ainsi à une modélisation probabiliste de la formation des réserves, dont les outils mathématiques nécessaires sont détaillés dans la deuxième section. Le modèle de fragmentation de Bolthausen-Sznitman est introduit dans la section 3 pour décrire la fragmentation progressive des réserves au cours de la formation du bassin.

2.1 Modélisation de la formation des réserves pétrolières

2.1.1 Formation des réserves pétrolières

Dans le cycle du carbone organique, une faible partie des déchets organiques issus de la décomposition des êtres vivants, essentiellement d'origine végétale, n'est pas détruite par les bactéries ; il s'agit de la matière organique sédimentaire. Les bassins sédimentaires sont des dépressions ou cuvettes partiellement recouvertes d'eau, où s'accumulent ces sédiments qui tendent ainsi à combler les creux. Une première décomposition de ces déchets organiques donne naissance au kérogène, prisonnier de la roche argileuse (roche mère). Du fait du remplissage progressif du bassin, on observe une superposition des couches géologiques : les dépôts de la base d'un bassin sont déposés en premier ; ceux qui les surmontent sont de plus en plus jeunes. La géodynamique interne de l'écorce terrestre conduit à un affaissement progressif du bassin (phénomène de subsidence). Pendant que les couches sédimentaires s'enfoncent (migration primaire), le kérogène se transforme en hydrocarbure par craquage thermique. Si

la température augmente rapidement pendant la descente, les hydrocarbures formés seront de nature gazeuse. Sous l'effet de la pression, les hydrocarbures sont alors expulsés de la roche mère. Plus légers que l'eau, ils ont tendance à monter vers la surface le long des fractures et des drains perméables. Au cours de cette migration dite secondaire, ils peuvent être bloqués par une couche supérieure imperméable. Pour que se constitue alors un gisement, il faut que les hydrocarbures s'accumulent dans une poche ou dans les fissures d'une roche réservoir. On distingue différents pièges : les pièges structuraux et les pièges stratigraphiques. Les pièges anticlinaux sont les plus nombreux, et les géologues ont appris à les repérer dès la fin du XIX-ème siècle sur le territoire des États-Unis. Dans le cas contraire, si le pétrole n'est pas piégé, il continue sa course jusqu'en surface, on parle alors de dysmigration. On appelle *système pétrolier* l'ensemble des facteurs géologiques aboutissant à l'accumulation d'hydrocarbures.

Le détail de la formation des gisements dans un bassin donné est sans doute très complexe : il est probable que de nombreux petits gisements coalescent puis se fragmentent puis coalescent de nouveau ...etc, pour aboutir à ce que l'on observe aujourd'hui. Pour des raisons expliquées plus loin, le modèle que l'on propose permet de s'affranchir d'une connaissance détaillée de ces successions de coalescences et de fragmentations. En effet, on explique dans la section consacrée au coalescent de Bolthausen-Sznitman qu'une fragmentation suivie d'une coalescence équivalente ne donne lieu à aucun changement de la statistique de la taille relative des gisements. Aussi dans un premier temps on adopte un point de vue qui centre la description sur le phénomène de fragmentation (entre autres choses parce qu'elle est plus facile à modéliser mathématiquement) et donc on considère (temporairement) que l'on part d'un gisement primitif qui a été uniquement fragmenté : même si cela ne reflète pas l'histoire réelle du gisement, cela permet d'expliquer l'apparition de certaines lois sur la répartition de la taille des gisements au sein d'un bassin. Une discussion plus détaillée est fournie dans la section introduisant le coalescent de Bolthausen-Sznitman. Concernant la fragmentation des gisements, nous supposons que celle-ci opère suivant les deux hypothèses suivantes :

- $\mathbf{H_a}$: Au cours du temps géologique, les phénomènes géologiques causant les fragmentations sont indépendants des gisements, et la fragmentation d'un gisement à un certain instant ne dépend pas des événements géologiques précédents.
- $\mathbf{H_b}$: Chaque gisement est fragmenté de la même façon à changement d'échelle près.

2.1.2 Distribution probabiliste des réserves

A l'intérieur d'un bassin pétrolifère, les hydrocarbures ne sont pas répartis de façon homogène dans le sous-sol ; ils sont piégés à l'intérieur de gisements. Considérer les réserves pétrolières à l'échelle d'un bassin permet de supposer que les caractéristiques géologiques du bassins sont identiques sur toute la zone étudiée. De plus, la population des champs d'un bassin pétrolier est suffisamment grande pour que des procédures d'estimation puissent être proposées. Il est donc naturel de définir une distribution de probabilité pour décrire les tailles des gisements à l'échelle d'un bassin, qui traduira ainsi les propriétés géologiques de la région.

La distribution des tailles de gisements a été l'objet de nombreuses études ces dernières décennies. Il semble que l'un des premiers articles sur ce sujet soit celui de Kaufman (1963), dans lequel l'auteur propose d'utiliser des distributions de type lognormal. Plus tard, Hough-

ton (1988) élabore un modèle probabiliste des découvertes d'hydrocarbures, construit à partir de tirages biaisés par la taille dans une population de gisements. La taille de ces gisements y est modélisée par des distributions de Lévy-Pareto, translatées et tronquées. Récemment et dans le même esprit, Lepez (2002) bâtit une procédure d'estimation du potentiel des réserves d'un bassin en supposant que la distribution des réserves suit une loi de Lévy-Pareto. Ces références ne constituent que quelques exemples parmi une littérature assez vaste privilégiant les distributions de type Lévy-Pareto ou lognormal pour modéliser la taille des gisements d'un bassin.

Comme l'expliquent Kaufman *et al.* (1975), il peut être difficile de distinguer une distribution lognormal d'une distribution de Lévy-Pareto. De plus, Attanasi et Charpentier (2002) ont montré que le choix entre l'une ou l'autre des deux distributions n'est pas sans conséquence sur les estimations de réserves du bassin. Les décisions économiques qui en découlent sont ainsi très sensibles à ce choix de modélisation.

Selon Lepez (2002), la distribution lognormal fournit une bonne description de la distribution des champs découverts à une date donnée, alors que la distribution de Lévy-Pareto modélise quant à elle la distribution des réserves de tous les champs du bassin, découverts ou encore inconnus. Tous les champs ne présentent pas la même probabilité d'être mis à jour, et la distribution de réserves des champs découverts n'est donc pas la même que celle des réserves pour la population complète. En effet, les champs les plus importants sont aussi les plus faciles à détecter, et ils sont de plus recherchés en priorité par les compagnies pétrolières. Les champs les plus modestes sont donc sous-représentés à l'intérieur de la famille des gisements découverts, ce qui explique pourquoi une distribution de type lognormal s'ajuste mieux que la distribution Lévy-Pareto sur les découvertes. Dans la suite, nous adopterons la distribution de Lévy-Pareto pour modéliser la taille des champs d'un bassin. Cependant, nous verrons que notre modèle pour la formation des réserves exposé plus loin permet de réconcilier les points de vue Lévy-Pareto et lognormal, en expliquant dans quelles situations l'une ou l'autre des deux distributions est observée.

Nous rappelons maintenant quelques définitions sur la loi de Lévy-Pareto, et nous renvoyons au premier chapitre de Lepez (2002) pour une description plus complète de l'utilisation de ces lois dans le contexte pétrolier.

Distribution de Lévy-Pareto

Il est possible de définir la distribution de Lévy-Paréto par la propriété suivante d'*invariance par changement d'échelle* :

Définition 2.1.1. Une variable X suit une loi de Lévy-Pareto μ ssi :
- il existe $\varepsilon > 0$ tel que le support de μ est $[\varepsilon, \infty)$;
- pour tous $\eta_1, \eta_2 \in [\varepsilon, \infty)$, la loi de $\eta_1^{-1}X$ sous $\mathbb{P}(\,\cdot\,|X > \eta_1)$ est la même celle de $\eta_2^{-1}X$ sous $\mathbb{P}(\,\cdot\,|X > \eta_2)$.

La définition entraîne immédiatement que $\log(\varepsilon^{-1}X)$ satisfait la propriété d'absence de mémoire qui caractérise les exponentielles et donc une variable X suit une loi de Lévy-Pareto ssi $X = \varepsilon \exp(Y)$ où Y est une variable aléatoire de loi exponentielle de paramètre α. La

densité d'une distribution de Lévy-Paréto a donc pour expression

$$f(x) = \alpha \frac{\varepsilon^\alpha}{x^{\alpha+1}} \mathbb{1}_{x \geq \varepsilon}$$

où α et ε sont respectivement l'indice (ou exposant) et le seuil de la distribution de Lévy-Pareto. Notons que cette distribution n'admet pas de moment d'ordre 1 lorsque $\alpha \leq 1$ (distribution à queue lourde).

Pour décrire la répartition spatiale des gisements pétroliers dans un bassin, on évoque souvent une organisation "en satellites" : à côté d'un très gros gisement, se trouvent généralement quelques champs de taille plus modeste, autour desquels on découvre de nombreux gisements plus petits. Ce type de structure spatiale observée dans les bassins pétroliers est compatible avec la propriété d'invariance stochastique rappelée ci-dessus. Cette remarque constitue l'un des arguments majeurs des géologues dans l'utilisation des lois de Lévy-Pareto pour modéliser la taille des réserves des gisements.

La définition 2.1.1 d'une distribution de Lévy-Pareto suppose qu'il existe un seuil ε en deçà duquel il n'existerait plus de gisements. En réalité, ce seuil correspond plus à une réalité technologique et économique qu'à une réalité géologique. En effet, les techniques de détection utilisées en prospection pétrolière ne permettent pas de localiser efficacement les gisements en-dessous d'une certaine taille. De plus, ces petits gisements ne sont pas souvent pas exploitables pour des raisons économiques. De ce fait, en-dessous d'un certain seuil, les champs d'hydrocarbures ne sont plus répertoriés. Cependant, ces explications montrent aussi que le seuil est susceptible d'évoluer au cours du temps, notamment avec les progrès technologiques et l'évolution du contexte économique. Par exemple, si le prix du brut augmente, certains petits gisements deviendront rentables et seront recherchés avec plus d'intérêt. L'un des objectifs de ce chapitre est de définir un modèle probabiliste qui s'affranchisse de n'importe quel type de seuil (économique ou technologique). Concernant la distribution des réserves, nous souhaitons valider l'hypothèse de distribution suivante :

$\mathbf{H_c}$: un gisement de taille supérieure à ε suit une loi de Lévy-Pareto sur l'intervalle $[\varepsilon, \infty[$.

Diagramme LogLog

Il est aussi possible de vérifier graphiquement que ce choix de modélisation est valide à l'aide de ce que l'on appelle un diagramme LogLog.

Définition 2.1.2. (Lepez, 2002, p. 54) Soit $\{z_1, \ldots, z_p\}$ une série de données réelles strictement positives. Soit σ la permutation de l'ensemble $\{1, \ldots, p\}$ qui à $\{z_1, \ldots, z_p\}$ associe la série $\{z_{\sigma(1)}, \ldots, z_{\sigma(p)}\}$ où $z_{\sigma(1)} \geq \ldots \geq z_{\sigma(p)}$. On appelle diagramme LogLog de la série $\{z_1, \ldots, z_p\}$ le graphique à double échelle logarithmique où sont portés les points $(i, z_{\sigma(i)})_{1 \leq i \leq p}$.

On peut montrer que le diagramme LogLog d'un échantillon qui suit une loi de Lévy-Pareto d'exposant α présente une tendance linéaire de pente $-\frac{1}{\alpha}$ (voir Lepez, 2002, p.75) ; la figure 2.1 illustre ce phénomène. La figure 2.2 est le diagramme LogLog des champs de la mer du Nord. La tendance linéaire est assez bien vérifiée pour les champs d'au moins 100 Mb, mais pour les plus petits gisements ce n'est pas le cas. La sous-représentativité de ces derniers, que

nous avons évoquée plus haut, est responsable de ce fléchissement que l'on n'observe pas sur les données simulées de la figure 2.1.

2.1.3 Modèle de fragmentation aléatoire

Nous proposons de modéliser les réserves pétrolières des gisements d'un bassin par les sauts d'un subordinateur stable d'indice α. Les subordinateurs sont des processus de Lévy à valeurs dans $[0, +\infty[$, un rappel mathématique sur ces objets est exposé dans la section suivante. Cette nouvelle modélisation présente plusieurs avantages. Tout d'abord, notons qu'elle est cohérente avec le point de vue "échantillon de Lévy-Pareto" puisque que les sauts d'un subordinateur stable de tailles supérieures à un seuil ε sont des variable aléatoires de distribution de Lévy-Pareto ; l'hypothèse $\mathbf{H_c}$ est donc vérifiée. De plus, nous verrons que cette modélisation est compatible avec l'idée selon laquelle la répartition actuelle du pétrole dans l'ensemble des gisements du bassin serait la réalisation d'une fragmentation aléatoire de la quantité initiale d'hydrocarbures. Les partitions aléatoires échangeables introduites par Kingman et Pitman offrent un cadre mathématique rigoureux pour modéliser de tels phénomènes. L'hypothèse $\mathbf{H_a}$ suggère que le processus de fragmentation a un comportement markovien, et l'hypothèse $\mathbf{H_b}$ précise le type de mécanisme de fragmentation des réserves. Le modèle de fragmentation de Bolthausen-Sznitman, c'est à dire le coalescent de Bolthausen-Sznitman retourné dans le temps, qui décrit la fragmentation progressive de partitions stables vérifie aussi ces deux autres conditions. Dans la section suivante, nous rappelons les notions mathématiques nécessaires pour définir rigoureusement les partitions aléatoires, tout en précisant au fur et à mesure comment ceux-ci sont employés pour la modélisation des réserves.

2.2 Rappels mathématiques sur les partitions stables

Les notions qui sont rappelées dans cette section relèvent du domaine des modèles combinatoires aléatoires ; le lecteur pourra consulter les notes de Saint-Flour de Pitman (2006) pour plus de détails à ce sujet.

2.2.1 Partitions aléatoires

On se place désormais sur un même espace de probabilité $(\Omega, \mathcal{F}, P)$. Soit $[n]$ l'ensemble des entiers $\{1, 2, \ldots, n\}$.

Définition 2.2.1.
- Une partition de $[n]$ est une collection non ordonnée $\{A_1, \ldots, A_k\}$ de sous-ensembles de $[n]$ disjoints, non vides et de réunion $[n]$. On note $\mathcal{P}_{[n]}$ l'ensemble de toutes les partitions de $[n]$.
- Une partition de $\mathbb{N}$ est une collection non ordonnée $\{A_1, \ldots, A_k \ldots\}$ de sous-ensembles de $\mathbb{N}$ disjoints, non vides et de réunion $\mathbb{N}$. On note $\mathcal{P}_\infty$ l'ensemble de toutes les partitions de $\mathbb{N}$.
- Une composition de n est une suite d'entiers $(n_1, \ldots, n_k)$ dont la somme vaut n.

Une partition de $[n]$ n'étant par définition pas ordonnée, il se pose le problème de l'énumération des blocs. Une première solution consiste à les indicer par ordre d'apparition, c'est-à-dire

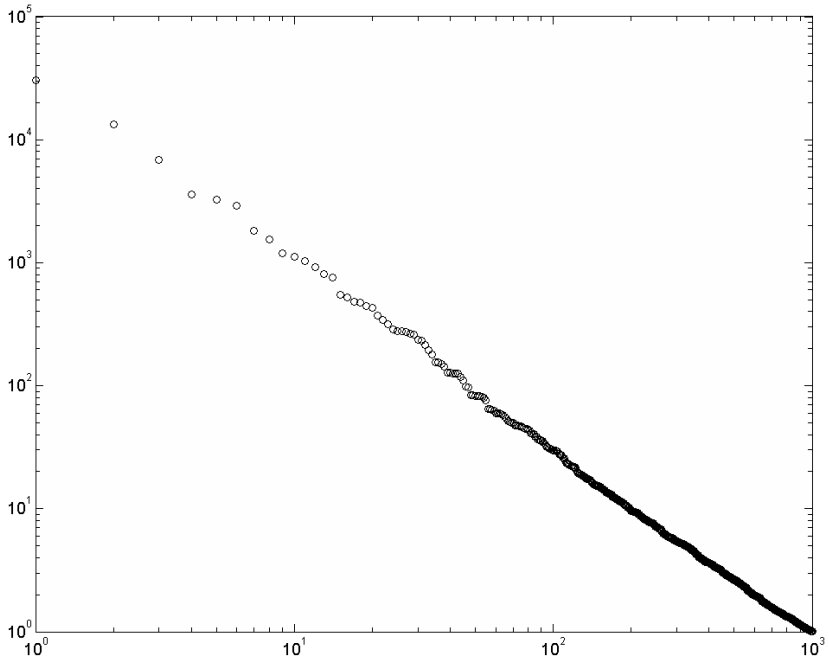


FIG. 2.1: Diagramme LogLog d'un échantillon i.i.d. de 1000 variables aléatoires de loi de Lévy-Pareto d'exposant $\alpha = 0.75$.

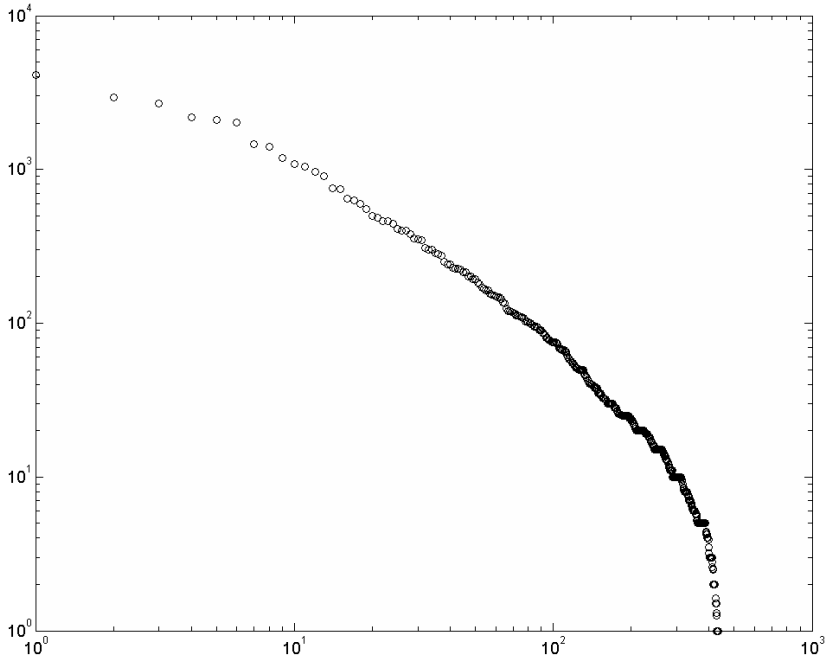


FIG. 2.2: Diagramme LogLog des tailles des champs du Viking Graben de mer du Nord (données Wood Mackenzie 2002 en Mb pour les champs de taille supérieure à 1 Mb).

par ordre croissant de leur plus petit élément. Le bloc contenant l'élément 1 est le premier de la liste, le suivant est le bloc qui contient 2 (sauf si 2 était déjà dans le premier bloc), etc... On note $(\tilde{N}_{n,1}, \ldots, \tilde{N}_{n,k_n})$ la suite des tailles des blocs rangés de cette façon. Une seconde façon d'énumérer les blocs consiste à les ranger par ordre décroissant de la taille des blocs, et $(N^{\downarrow}_{n,1}, \ldots, N^{\downarrow}_{n,k_n})$ désignera la composition de n déduite de la partition en ordonnant les blocs par ordre décroissant de leur taille. De façon générale, pour une partition Π, on notera $i \sim_{\Pi} j$ lorsque i et j sont dans un même bloc de Π.

Une partition aléatoire Π_n de l'ensemble $[n]$ est une variable aléatoire à valeur dans l'ensemble $\mathcal{P}_{[n]}$. On définit sur $\mathcal{P}_\infty$ la tribu

$$\mathcal{G} = \sigma\left(\{\pi_\infty \in \mathcal{P}_\infty \ : \ i \sim_{\pi_\infty} j\}, i, j \in \mathbb{N}\right) .$$

Une partition aléatoire de $\mathbb{N}$ est une variable aléatoire $\mathcal{G}$-mesurable. La numérotation devant rester arbitraire, on impose aux partitions d'être *échangeables* :

Définition 2.2.2.
1. Une partition aléatoire Π_n de $[n]$ est dite échangeable si sa distribution est invariante sous l'action du groupe des permutations de $[n]$.
2. Une partition aléatoire Π_∞ de $\mathbb{N}$ est dite échangeable si sa distribution est invariante sous l'action du groupe des permutations à support fini.

Pour une partition aléatoire de $[n]$, cette définition signifie que la distribution ne dépend pas des entiers présents dans chaque bloc, mais uniquement de la composition aléatoire de n déduite de la partition aléatoire :

$$\mathbb{P}(\Pi_n = \{A_1, \ldots, A_k\}) = p_n(|A_1|, \ldots, |A_k|) \tag{2.1}$$

où p_n est une fonction symétrique des compositions de n appelée *fonction de probabilité de la partition échangeable* (fppe).

Concernant les réserves pétrolières, ces partitions aléatoires représentent (d'une manière précisée plus loin) la famille des gisements du bassin. D'après l'analyse précédente des phénomènes géologiques en jeu, ces partitions doivent agir "de la même façon", quelle que soit la taille de l'ensemble à partitionner. Cette propriété s'appelle la *consistance* :

Définition 2.2.3. On dit qu'une suite (Π_n) de partitions aléatoires échangeables est consistante en distribution si

$$\forall m < n, \ \Pi_{m,n} \stackrel{d}{=} \Pi_m$$

où $\Pi_{m,n}$ est la restriction à $[m]$ de Π_n : si $\Pi_n := \{A_1, \ldots, A_K\}$ alors

$$\Pi_{m,n} = \{A_1 \cap [m], \ldots, A_K \cap [m]\}.$$

Si une suite (Π_n) est consistante, on a alors :

$$p_n(n_1, \ldots, n_k) = \sum_{j=1}^{k} p_{n+1}(\ldots, n_j + 1, \ldots) + p_{n+1}(n_1, \ldots, n_k, 1)$$

où $(n_1, \ldots, n_k)$ est une composition de n et p est défini par (2.1). En effet, si la partition aléatoire restreinte à $[n]$ a des blocs de tailles $(n_1, \ldots, n_k)$, alors la restriction à $[n+1]$ a des blocs de tailles à choisir parmi les compositions $(n_1, \ldots, n_k, 1), \ldots, (n_1, \ldots, n_k + 1), (n_1, \ldots, n_k, 1)$.

Notons que si Π_n est la restriction d'une partition Π_∞ de $\mathbb{N}$ aux n premiers entiers, alors Π_n est échangeable pour tout n, et la suite (Π_n) est consistante. Inversement, par le théorème de Kolmogorov, une suite consistante de partitions échangeables définit de manière unique en loi une partition Π_∞ sur $\mathbb{N}$.

Il existe plusieurs façons de construire des partitions aléatoires échangeables de $\mathbb{N}$, la méthode la plus naturelle consiste à utiliser une mesure de probabilité aléatoire pour définir la partition aléatoire, ce qui est expliqué dans la section suivante.

2.2.2 Partitions de Poisson-Kingman

Nous allons donner un procédé permettant de générer toutes le partions échangeables possibles. Pour cela quelques définitions sont nécessaires. Nous allons tout d'abord considérer des suites de variables aléatoires $X = (X_n, n \geq 0)$ définies sur $(\Omega, \mathcal{F})$, à valeurs dans $[0, 1]$ qui sont échangeables, c'est-à-dire qu'elles satisfont la condition suivante.

Définition 2.2.4. Une suite infinie $(X_0, X_1, \ldots)$ de variables aléatoires est dite *échangeable* si

$$(X_0, X_1, \ldots) \stackrel{d}{=} (X_{\sigma(0)}, X_{\sigma(1)}, \ldots)$$

pour toute permutation σ de $\mathbb{N}$ ne déplaçant qu'un nombre fini d'entiers.

Soit $\mathcal{M}_1([0, 1])$ l'ensemble des probabilités définies sur les boréliens de $[0, 1]$, équipé de la topologie de la convergence faible (qui est métrisable) et des boréliens associé à cette topologie.

Définition 2.2.5. Une mesure de probabilité aléatoire M est une fonction de Ω dans $\mathcal{M}_1([0, 1])$ qui est mesurable.

Soit M une mesure de probabilité aléatoire. Soit $(X_1, \ldots, X_n, \ldots)$ des variables aléatoires i.i.d. conditionellement à M et de loi conditionnelle M ; c'est-à-dire telles que pour toute suite de fonctions $F_n : [0, 1] \to \mathbb{R}$ mesurables bornées,

$$\forall n, \quad \mathbb{E}[F_1(X_1) \ldots, F_n(X_n)] = \int_\Omega \mathbb{P}(d\omega) \int_{[0,1]} M(\omega, dx_1) F_1(dx_1) \times \ldots \times \int_{[0,1]} M(\omega, dx_n) F_n(dx_n) \, .$$

La suite $(X_i)_{i \geq 0}$ est alors échangeable et on dit que $(X_1, \ldots, X_n, \ldots)$ est un M-mélange, et le théorème de De Finetti (voir par exemple Aldous, 1985) affirme que toute suite échangeable est un mélange.

Soit M une mesure de probabilité aléatoire et un M-mélange $X = (X_n, n \geq 0)$. On définit ensuite une partition aléatoire $\Pi_\infty(X)$ associée à la suite échangeable $(X_1, X_2, \ldots)$ par :

$$i \sim_{\Pi_\infty(X)} j \Longleftrightarrow X_i = X_j \, .$$

Il est clair que $\Pi_\infty(X)$ est donc échangeable, et la distribution de Π_∞ est déterminée par les atomes (aléatoires) de la mesure M. Notons enfin que si la probabilité aléatoire n'a p.s. pas d'atomes, alors la partition sera p.s. constituée de singletons.

Nous venons de voir que l'on peut générer une partition aléatoire échangeable infinie sur $\mathbb{N}$ à l'aide d'une mesure de probabilité aléatoire. La correspondance de Kingman établit que toutes les partitions aléatoires échangeables de $\mathbb{N}$ s'obtiennent de cette façon. Introduisons l'ensemble suivant :

$$\mathcal{S}_1 := \{\mathbf{p}^{\downarrow} = (p_n^{\downarrow}, n \geq 0)\,;\, \sum_{n=1}^{\infty} p_n^{\downarrow} \leq 1\},$$

muni de la topologie produit. Pour chaque suite d'atomes $\mathbf{p}^{\downarrow}$ de $\mathcal{S}_1$, on peut définir une mesure de probabilité aléatoire

$$\mu_{\mathbf{p}^{\downarrow}}(dx) = \sum_{i\geq 1} p_i^{\downarrow}\delta_{\frac{1}{i}}(dx) + (1 - \sum_{i\geq 1} p_i^{\downarrow})1_{[0,1]}dx,$$

et la fonction $\mathbf{p}^{\downarrow} \mapsto \mu_{\mathbf{p}^{\downarrow}}$ est mesurable.

Théorème 2.2.1. *Soit $\Pi_\infty := (\Pi_n)$ une partition aléatoire échangeable sur $\mathbb{N}$. Soit $N_{n,i}^{\downarrow}$ la suite des tailles des blocs de Π_n classés par ordre décroissant. Alors, il existe $\mathbf{P} = (P_i^{\downarrow})$ une variable aléatoire dans $\mathcal{S}_1$ telle que*

$$p.s.\ \forall i, \quad \frac{N_{n,i}^{\downarrow}}{n} \to_{n\to\infty} P_i^{\downarrow}\ .$$

De plus, si X est un $\mu_{\mathbf{P}^{\downarrow}}$-mélange, alors Π_∞ a même loi que $\Pi_\infty(X)$.

Dans notre cadre pétrolier, les blocs de la partition aléatoire représentent les gisements du bassin, et les fréquences des blocs correspondent à leurs tailles normalisées par la quantité totale d'hydrocarbures enfouie dans le sous-sol. Pour étudier la distribution des tailles relatives des champs, il est donc suffisant d'étudier les tailles relatives des blocs de $\Pi_\infty(X)$.

Le théorème de Kingman est l'analogue pour les partitions échangeables du théorème de De Finetti pour les variables échangeables, et sa démonstration s'appuie essentiellement sur ce résultat. Le théorème 2.2.1 établit la correspondance entre la partition aléatoire et la fréquence de ses blocs à l'aide des atomes d'une mesure aléatoire sur $\mathcal{S}_1$. Les mesures de Poisson fournissent un cadre agréable pour manipuler ces mesures aléatoires.

Constructions poissoniennes

Commençons par rappeler quelques définitions sur les mesures de Poisson, les propriétés élémentaires des mesures de Poisson et processus de Lévy sont rappelées dans l'annexe B. Soit E un espace Polonais et ν une mesure σ-finie sur E.

Définition 2.2.6. Soit Φ une mesure aléatoire sur E. On dit que Φ est une mesure de Poisson d'intensité ν si :

- Pour tout borélien A de E tel que $\nu(A) < \infty$, $\Phi(A)$ suit une loi de Poisson de paramètre $\nu(A)$.
- Pour tous boréliens deux à deux disjoints $B_1, ..., B_n$, alors $\Phi(B), ..., \Phi(B)$ sont des variables aléatoires indépendantes.

Les mesures de Poisson sur $(0,\infty)$ permettent de définir facilement des mesures de probabilité aléatoires de la façon suivante. Soit Λ une mesure sur $]0,+\infty[$ sans atome et telle que

$$\int_{]0,\infty[} (1 \wedge x)\, \Lambda(dx) < \infty \tag{2.2}$$

et

$$\Lambda(]0,\infty[) = \infty, \tag{2.3}$$

et soit M une mesure de Poisson de mesure aléatoire d'intensité Λ. L'hypothèse (2.2) permet de supposer qu'il y a un nombre fini de points dans les intervalles $[\varepsilon,\infty[$ pour $\varepsilon > 0$, et l'hypothèse (2.3) garantit qu'il y a une quantité infinie de points dans $]0,\infty[$, ce qui justifie l'écriture $M := \sum_{i=1}^{\infty} \delta_{a_i}$, avec $a_1 \geq \ldots a_i \geq \ldots > 0$. De plus, en utilisant la formule exponentielle (voir la proposition B.1.2 dans le chapitre suivant), on peut montrer que la condition (2.2) implique que $S_M := \sum_{i=1}^{\infty} a_i < \infty$ presque sûrement.

Sous les hypothèses précédentes, considérons la suite des fréquences aléatoires $(P_i)_{i\geq 1}$ obtenue par normalisation des atomes a_i de M de la façon suivante,

$$P_i := \frac{a_i}{S_M}.$$

La suite $(P_i)_{i\geq 1}$ définit de façon naturelle une mesure de probabilité aléatoire M. Une partition aléatoire de $\mathbb{N}$ induite par une mesure de probabilité M construite de cette façon est appelée *partition de Poisson-Kingman*. Pour la modélisation pétrolière, nous nous intéressons plus spécifiquement aux partitions stables, qui peuvent être obtenues de cette façon à partir des sauts d'un subordinateur stable.

2.2.3 Partitions stables

Définition 2.2.7. Un subordinateur (fini) T est un processus positif à accroissements indépendants homogènes.

Un subordinateur $(T_s)_{s\geq 0}$ est croissant, et dans la suite on le supposera toujours cadlag et tel que $T_0 = 0$. Pour tout $t > 0$, la loi de T_t est infiniment divisible et caractérisée par sa transformée de Laplace (voir la section B.2 en appendice pour plus de détails) ; son exposant de Laplace Φ vérifie la formule de Lévy-Khintchine suivante

$$\Phi(\lambda) = d\lambda + \int_{(0,\infty)} (1 - e^{-\lambda x})\Lambda(dx)$$

où la mesure de Levy Λ satisfait (2.2). Soit $(\Delta_s)_{t\geq 0}$ le processus des sauts associé à T, la mesure

$$M = \sum_{t:\Delta_t > 0} \delta_{(t,\Delta_t)}$$

est une mesure de Poisson sur $(0,\infty)^2$ d'intensité $dt \otimes \Lambda(dx)$. On a donc la représentation de Lévy-Ito

$$T_t = dt + \sum_{t\geq 0} \Delta_t\ .$$

Dans la suite, on suppose que le drift d est nul et que la mesure de Lévy est diffuse et satisfait (2.3). Pour un réel positif s_0, la mesure aléatoire

$$M_{s_0} = \sum_{\substack{t:\Delta_t>0 \\ t\in[0,s_0]}} \delta_{\Delta_t}$$

est une mesure de Poisson sur $(0,\infty)$ d'intensité $s_0\Lambda(dx)$. Notons que d'un point de vue théorique, il est toujours possible de se ramener au cas $s_0 = 1$ en changeant la mesure de Lévy par $s_0\Lambda$.
Nous aurons besoin dans la suite de la propriété suivante vérifiée par les subordinateurs.

Proposition 2.2.1. *Soient deux subordinateurs indépendants $S = (S_t, t \geq 0)$ et $T = (T_t, t \geq 0)$ d'exposants de Laplace respectifs Ψ et Φ. Alors $S \circ T = (S_{T_t}, t \geq 0)$ est un subordinateur d'exposant de Laplace $\Phi \circ \Psi$.*

Un subordinateur T dont la mesure de Lévy vérifie les propriétés précédentes permet alors de définir une partition aléatoire échangeable en considérant pour la suite des fréquences des blocs de la partition la suite des sauts normalisés du subordinateur :

$$P_i^{\downarrow} = \frac{\Delta_i^{\downarrow}}{T_{s_0}}$$

où $\Delta_i^{\downarrow}$ est la suite des sauts Δ_s non nuls pour $s \leq s_0$. Notons que l'ensemble des $s \geq 0$ tels que $\Delta T_s > \varepsilon$ est presque sûrement discret car $\Lambda([\varepsilon,\infty[) < \infty$ par hypothèse, et il est donc possible d'ordonner les sauts de façon décroissante.

Pour la modélisation pétrolière, T_{s_0} représente la totalité des quantités d'hydrocarbures contenues dans le sous-sol du bassin, et le saut $\Delta_i^{\downarrow}$ correspond à la taille du i-ème plus grand gisement du bassin. Il s'agit maintenant de choisir un subordinateur de façon à ce que la loi des sauts, c'est-à-dire la mesure aléatoire Λ corresponde à la distribution des tailles des gisements observés dans la réalité.

Subordinateur stable

On appelle subordinateur stable un subordinateur T de mesure de Lévy $\Lambda(x) = s\,\rho(x)dx$ où $s > 0$ et

$$\rho(x) = \frac{\alpha}{\Gamma(1-\alpha)}\frac{1}{x^{1+\alpha}} \ . \tag{2.4}$$

Pour ce subordinateur, l'exposant de Laplace a pour expression $\Phi(\lambda) = \lambda^{\alpha}$, et T vérifie la propriété de scaling

$$(T_s)_{s\geq 0} \stackrel{d}{=} (\lambda^{-1/\alpha}T_{\lambda s})_{s\geq 0}.$$

Soit $\varepsilon > 0$ et $(T_s)_{s\geq 0}$ un subordinateur α-stable. On note que

$$c_{\varepsilon,\alpha} := \Lambda((\varepsilon,\infty)) = \frac{\varepsilon^{-\alpha}}{\Gamma(1-\alpha)} < \infty,$$

et le subordinateur T vérifie donc les conditions suffisantes pour définir une partition aléatoire échangeable à partir des sauts normalisés de T. On appelle *partition stable* la partition aléatoire associée à ce subordinateur.

Puisque pour tout $\varepsilon > 0$ l'ensemble des $s \geq 0$ tels que $\Delta_s > \varepsilon$ est presque sûrement discret, on peut donc ordonner cet ensemble dans une suite croissante $(S_n)_{n\geq 1}$. D'après la proposition B.2.1 donnée en annexe, les variables aléatoires $S_{n+1} - S_n$ sont de loi exponentielle de paramètre $c_{\varepsilon,\alpha}$, et les variables aléatoires $X_n := \Delta_{S_n}$ sont i.i.d. de loi de densité $f(x) = \frac{s\,\rho(x)}{\Lambda([\varepsilon,\infty[)}\mathbb{1}_{x\geq\varepsilon}$ qui est la densité d'une loi de Lévy-Pareto sur (ε, ∞) de paramètre α.

Si l'on tronque les sauts du subordinateur stable, il ne reste plus qu'un nombre fini de sauts au-delà du seuil, et ceux-ci suivent des lois de Lévy-Pareto. Cette modélisation nous laisse donc la possibilité d'ajuster le niveau pour reproduire au mieux les données. Pour ces raisons, le subordinateur stable semble être le bon outil pour décrire les tailles des champs d'un bassin pétrolier. Le modèle de fragmentation de Bolthausen-Sznitman, que nous présentons dans la section suivante, fournit une modèle pour décrire la formation des réserves pétrolières au cours du temps dans un bassin, qui est de plus compatible avec la représentation des tailles des gisements par les sauts d'un subordinateur stable.

2.3 Modélisation de la formation des réserves par le modèle de Bolthausen-Sznitman

Dans cette section, nous adaptons le modèle de Bolthausen-Sznitman pour la formation des réserves pétrolières. Ce modèle a été introduit par Bolthausen et Sznitman (1998) pour étudier un modèle simplifié de verres de spin appelé Random Energy Model (REM), introduit par Derrida (2000).

2.3.1 REM et formation des réserves pétrolières

Pour la formation des gisements dans un bassin pétrolier, on soutient un modèle-jouet dont le but est d'expliquer qualitativement la forme des lois des tailles des gisements. Ce modèle se définit autour des trois hypothèses suivantes.

- (i) *Hypothèse de subdivision en gisements élémentaires équivalents.* On suppose que les gisements observés réellement au sein d'un même bassin peuvent se subdiviser en un grand nombre de petits gisements élémentaires $g_1, \ldots, g_M$ sur lesquels l'action géologique a agi de façon équivalente d'un point de vue statistique. Ces petits gisements élémentaires ont été susceptibles de se déplacer (par migration) et leur la taille a pu varier au cours du temps : augmenter si d'autres petites poches sont venues s'agréger, et diminuer si par exemple des failles ont permis à une partie de ces gisements élémentaires de migrer. On remarque qu'à chaque événement géologique la variation de la taille d'un gisement élémentaire va plutôt affecter la taille en proportion plutôt qu'en valeur absolue (par exemple, pour simplifier, une faille coupe le bassin en deux, la fusion avec un bassin de taille similaire multiplie la taille par deux).

– (ii) *Hypothèse de "champ moyen".* On suppose qu'un nombre important N de facteurs géologiques ont contribué à façonner ces gisements élémentaires et à déterminer leur taille et leur place finale. Nous n'avons aucune connaissance détaillée de ces facteurs géologiques à l'échelle des gisements élémentaires, aussi nous supposons que ces N facteurs géologiques jouent le même rôle, sont indépendants et agissent aléatoirement de la même façon et indépendamment sur chaque gisement élémentaire. Tout ceci constitue l'hypothèse de "champ[1] moyen".
On s'intéresse aux fluctuations créées par ces N facteurs sur la taille des gisements, par rapport à leur action moyenne. Plus précisément, nous choisissons le type d'action simplifiée suivant : on considère que le facteur géologique i, $1 \leq i \leq N$, a fait fluctuer la taille du gisement g_p d'un facteur $\exp(\epsilon_i^p)$, où ϵ_i^p peut prendre deux valeurs possibles $-v$ et v. L'action cumulée des N facteurs géologiques sur la taille de g_p est donc de multiplier la taille de g_p par le coefficient

$$\exp\left(\sum_{i=1}^{N} \epsilon_i^p\right).$$

On traduit les hypothèses de champ moyen ci-dessus en supposant donc que les variables $(\epsilon_i^p; 1 \leq i \leq N, 1 \leq p \leq M)$ sont indépendantes et de même loi

$$\mathbb{P}(\epsilon_i^p = -v) = \mathbb{P}(\epsilon_i^p = v) = 1/2\ .$$

– (iii) *Hypothèse d'économie de la subdivision.* Dans ce système, les N facteurs contribuant à façonner un bassin peuvent se voir comme les N facteurs qui ont permis de distinguer et de déterminer la subdivision en M gisements élémentaires. Cela revient à supposer que N est proche du nombre minimal de facteurs nécessaires pour discriminer les M gisements. Comme $\{-v, v\}^N$ représente les 2^N valeurs possibles des ϵ_i^p, il est cohérent de choisir N de l'ordre de $\log M$. Pour fixer les idées, on choisit N grand et

$$M = 2^N.$$

Pour tout bassin g_p on note

$$H_N^*(g_p) = \sum_{i=1}^{N} \epsilon_i^p.$$

Comme N est supposé grand, la loi de $(Nv^2)^{-1/2} H_N^*(g_p)$ est proche d'une Gaussienne centrée standard que l'on note $G(g_p)$. Dans la suite on prendra

$$H_N^*(g_p) \approx v\sqrt{N} G(g_p).$$

La taille du gisement g_p est donc $\langle t_{v,N} \rangle \exp\left(v\sqrt{N} G(g_p)\right)$, où $\langle t_{v,N} \rangle$ est une taille typique[2] commune à tous les gisements élémentaires, qui fixe l'échelle de grandeur dans

[1] Le terme "champ" correspond ici au vocabulaire des physiciens, et ne désigne pas les gisements pétroliers.
[2] Il serait difficile de trouver une signification physique à cette taille typique.

laquelle les gisements élémentaires se situent. La taille relative du gisement élémentaire g_p est donc

$$\mu_{N,v}(g_p) = \frac{\exp\left(v\sqrt{N}G(g_p)\right)}{\sum_{q=1}^{2^N}\exp\left(v\sqrt{N}G(b_q)\right)} . \tag{2.5}$$

La question est de savoir comment se comportent ces tailles relatives lorsque N est grand. Si v est grand c'est-à-dire si l'amplitude des phénomènes de fluctuation géologique est grande, les gisements élémentaires réellement observés sont ceux pour lesquels la variable $v\sqrt{N}G(g_p)$ prend des valeurs importantes. En revanche, si v est faible, cela ne va pas se produire : les gisements élémentaires ne s'organisent pas : le bassin est émietté si N est très grand. Il y a une valeur critique v_c au dessus de laquelle le premier comportement asymptotique prévaut et en dessous de la quelle le second prévaut.

Donnons une rapide heuristique expliquant ce qui se passe lorsque v est petit. La taille totale du bassin est alors donnée par

$$Z_{v,N} = \langle t_{v,N}\rangle \sum_{q=1}^{2^N} \exp\left(v\sqrt{N}G(g_p)\right) .$$

Un calcul simple montre que

$$\mathbb{E}[Z_{v,N}] = \langle t_{v,N}\rangle 2^N \exp(Nv^2/2) = \langle t_{v,N}\rangle \exp\left(N(\frac{v^2}{2} + \log(2))\right)$$

et

$$\mathrm{var}(Z_{v,N}) = \langle t_{v,N}\rangle \left(\exp(N(2v^2 + \log(2))) - \exp(N(v^2 + \log(2)))\right) .$$

Par conséquent

$$\frac{\mathrm{var}(Z_{v,N})}{(\mathbb{E}[Z_{v,N}])^2} \sim_{N\to\infty} \exp(N(v^2 - \log(2))) .$$

Par conséquent si $v < \sqrt{\log(2)}$, $Z_{v,N}/\mathbb{E}[Z_{v,N}]$ tend vers 1 en probabilité ; un argument plus précis utilisant le critère de Lindeberg, montre qu'un théorème central limite a lieu. Dans ce cas aucun gisement élémentaire n'est très gros. Cela provient du fait que les queues de distributions des variables log-normales $\exp(v\sqrt{N}G(g_p))$ ne sont pas trop lourdes et peuvent satisfaire un théorème central-limite. En revanche lorsque v est très grand, ces queues de distribution deviennent grandes et certains bassins vont être anormalement grands et beaucoup influer sur la distribution $\mu_{v,N}$.

Il existe des résultats mathématiques précis sur le comportement limite de $\mu_{v,N}$ qui a la même loi que la mesure de Gibbs issue du REM de Derrida. Dans les travaux sur le REM, v joue le rôle de l'inverse de la température β et dans la normalisation traditionnelle, il convient de prendre $\sqrt{2}\cdot v = \beta$. Les résultats que nous citons sont dus, sous des formes diverses, à Derrida (2000), Guerra (1995), Ruelle (1987), Bovier (2002) (voir aussi Bovier *et al.*, 2002; Bovier et Kurkova, 2003, pour des résultats plus fins), on pourra consulter le livre de Talagrand (2003) pour des preuves précises.

La valeur critique v_c n'est pas $\sqrt{\log(2)}$, comme semble le suggérer l'analyse succincte par la variance, mais elle en est très proche : on a $v_c = \sqrt{2\log(2)}$. Cette valeur numérique n'a pas de signification physique réelle pour notre modèle en raisons des nombreux choix faits pour gérer le moins de paramètres possible dans notre description qualitative. Mais le mérite de v_c est d'exister.

• Cas où $v < v_c$ *(l'amplitude des fluctuations géologiques est faible).* On considère la mesure

$$m_{v,N} = \sum_{p=1}^{2^n} \mu_{N,v}(g_p)\delta_{p2^{-N}},$$

qui est une mesure sur $[0,1]$. Alors presque surement, lorsque N tend vers l'infini,

$$m_{v,N} \longrightarrow \lambda$$

pour la convergence faible des mesures sur $[0,1]$, où λ désigne la mesure de Lebesgue (voir Bovier et Kurkova, 2003, Theorem 2.4). On remarque que la limite ne dépend pas de v et une façon d'interpréter ce résultat est de dire qu'aucun gisement élémentaire ne prévaut et que ces gisements ne s'organisent pas en clusters plus gros. Ainsi lorsque $v < v_c$, le bassin présente de nombreux gisements régulièrement dispersés dont la taille suit une loi log-normale. Plus N est grand, plus il y a de gisements.

• Cas où $v > v_c$ *(l'amplitude des fluctuations géologiques est forte).* On ordonne les tailles relatives $(\mu_{v,N}(g_p), 1 \leq p \leq 2^N)$ par ordre décroissant :

$$\mu_{v,N}^{\downarrow} = \left(P_1^{\downarrow,N,v} > P_2^{\downarrow,N,v} > \ldots > P_{2^N}^{\downarrow,N,v}\right)$$

où $\{P_n^{\downarrow,N,v}; 1 \leq n \leq 2^N\} = \{\mu_{v,N}(g_p); 1 \leq p \leq 2^N\}$. Lorsque N tend vers l'infini, on a la convergence en loi suivante :

$$\mu_{v,N}^{\downarrow} \longrightarrow P^{\downarrow,\alpha} = \left(\frac{\Delta_n^{\alpha\downarrow}}{T_1}\ ;\ n \geq 0\right) \tag{2.6}$$

où $(\Delta_n^{\alpha\downarrow}\ ;\ n \geq 0)$ représente la suite réordonnée des sauts d'un subordinateur stable $(T_t\ ;\ t \in [0,1])$ d'indice α donné par

$$\alpha = \sqrt{2\log(2)}/v\ . \tag{2.7}$$

Autrement dit les gisements élémentaires vont se regrouper pour former des clusters qui vont être les gisements réellement observés dans le bassin et les tailles relatives sont distribuées comme une loi $PD(\alpha, 0)$ (nous revenons dans la section suivante sur la description de cette distribution).

La convergence (2.6) donne la forme de la statistique des tailles relatives mais un examen de sa preuve (voir Talagrand (2003) ou Bovier (2002)) montre que les tailles des bassins élémentaires ordonnées dans l'ordre décroissant convergent vers les sauts réordonnés d'un subordinateur stable d'indice $\alpha = v_c/v$. Plus précisément, on fixe $v > v_c$ et α donné par (2.7).

On note $T^{v,N\downarrow}$ les tailles de gisements ordonnées dans l'ordre décroissant, c'est-à-dire :

$$T^{v,N\downarrow} = (T_1^{v,N\downarrow} > T_2^{v,N\downarrow} > \ldots > T_{2^N}^{v,N\downarrow})$$

et

$$\{T_p^{v,N\downarrow}\,;\, 1 \leq p \leq 2^N\} = \{\langle t_{v,N}\rangle \exp(v\sqrt{N}G(g_p))\,,\ 1 \leq p \leq 2^N\}.$$

Le principe d'économie de la subdivision fixe l'échelle de grandeur où se situent les gisements afin d'observer quelque chose macroscopiquement : on peut montrer qu'il est nécessaire de choisir :

$$-\alpha \log(\langle t_{v,N}\rangle) = -\frac{\sqrt{2\log 2}}{v}\log(\langle t_{v,N}\rangle) = 2\log 2^N - \frac{1}{2}\log\left(4\pi\log(2^N)\right) \tag{2.8}$$

Dans ce cas, on a (lorsque N tend vers ∞) la convergence en loi

$$T^{v,N\downarrow} \longrightarrow \left(\Delta_n^{\alpha\downarrow}\,;\, n \geq 0\right) \tag{2.9}$$

La preuve utilise un résultat sur les valeurs extrêmes des Gaussiennes. Pour en donner une idée nous démontrons la convergence pour le plus gros bassin, c'est-à-dire :

$$T_1^{v,N\downarrow} \longrightarrow \Delta_1^{\alpha\downarrow} \tag{2.10}$$

Preuve de (2.10). Il est établi dans la suite de l'exposé que $(\Delta_1^{\alpha\downarrow})^{-\alpha}$ suit une loi exponentielle (voir proposition 3.3.1) ; par conséquent, prouver (2.10) revient à prouver que $(T_1^{v,N\downarrow})^{-\alpha}$ converge en loi vers une exponentielle. On pose

$$M_N = \max_{1\leq p\leq 2^N} G(g_p)$$

et on remarque que

$$T_1^{v,N\downarrow} = \langle t_{v,N}\rangle \exp\left(v\sqrt{N}M_N\right).$$

Puisque α et v sont liés par (2.7), on a

$$\left(T_1^{v,N\downarrow}\right)^{-\alpha} = (\langle t_{v,N}\rangle)^{-\alpha}\exp\left(-\sqrt{2\log(2^N)}M_N\right).$$

On fixe $x > 0$. On a alors les égalités suivantes

$$\begin{aligned}\mathbb{P}\left(\left(T_1^{v,N\downarrow}\right)^{-\alpha} \geq x\right) &= \mathbb{P}\left(M_N \leq -\frac{\alpha\log\langle t_{v,N}\rangle}{\sqrt{2\log 2^N}} - \frac{\log x}{\sqrt{2\log 2^N}}\right)\\ &= \mathbb{P}\left(M_N \leq a(N,x)\right),\end{aligned}$$

où on a posé

$$a(N,x) = \sqrt{2\log 2^N} - \frac{\log x}{\sqrt{2\log 2^N}} - \frac{1}{2}\cdot\frac{\log\left(4\pi\log(2^N)\right)}{\sqrt{2\log 2^N}}.$$

Par conséquent on a

$$\mathbb{P}\left(\left(T_1^{v,N\downarrow}\right)^{-\alpha} \geq x\right) = \left[1 - u\left(a(N,x)\right)\right]^{2^N},$$

où on rappelle que

$$u(y) := \frac{1}{\sqrt{2\pi}} \int_y^\infty e^{-z^2/2} dz \sim_{y\to\infty} \frac{1}{y\sqrt{2\pi}} e^{-y^2/2}.$$

Un calcul élémentaire montre ensuite que pour x fixé et N tendant vers l'infini, on a

$$u\left(a(N,x)\right) = 2^{-N} x(1 + o(1)),$$

ce qui entraîne que

$$\left[1 - u\left(a(N,x)\right)\right]^{2^N} \sim_{N\to\infty} e^{-x},$$

et ce qui achève la preuve de (2.10). □

Ce modèle qualitatif permet de concilier les deux types de lois observées pour la taille des bassins dans différents gisements pétroliers. En effet, on considère la la queue de distribution de la taille des gisements dans un bassin donné et on observe par par un diagramme log-log que cette distribution décroît en général approximativement en $x^{-\alpha}$.

- Si $\alpha < 1$ (queue lourde) alors (2.7) permet de penser que l'on est dans le régime $v > v_c$ et que la distribution de Poisson Dirichlet prévaut. Les tailles des champs vont donc être les sauts d'un subordinateur stable d'indice α qui sont tronqués à un certain seuil (de détection ou de rentabilité) et vont donc être distribuées comme des loi de Lévy-Paréto.
- Si $\alpha > 1$ (queue légère), alors on est plutôt dans le régime $v < v_c$ et les tailles des gisements sont distribuées selon des lois log-normales.

Si ce modèle prétend expliquer qualitativement le phénomène, il est alors nécessaire de le considérer de façon dynamique, c'est-à-dire de considérer qu'il est susceptible de varier dans des échelles de temps géologiques. Autrement dit, l'amplitude v devient une fonction du temps $v(t)$. On distingue plusieurs cas de figures.

- *La fonction $t \mapsto v(t)$ reste en dessous du seuil v_c.* Dans ce cas, comme on l'a vu plus haut, l'amplitude $v(t)$ n'a qu'un faible impact sur la forme des lois de la taille des bassins qui suivent une loi log-normale. L'augmentation ou la diminution de $v(t)$ n'entraîne rien d'irréversible sur la statistique des bassins.
- *La fonction $t \mapsto v(t)$ franchit le seuil v_c en montant.* Dans ce cas les bassins sont alors brutalement organisés en clusters et cela entraîne un changement irréversible sur la statistique de la tailles des bassins. En effet si après avoir franchit v_c en montant, la fonction $t \mapsto v(t)$ franchit ultérieurement v_c en descendant, il y a une fragmentation brutale et il est raisonnable d'interpréter cela, si N est très grand, comme une dysmigration massive : après ce second franchissement vers le bas du seuil v_c, il ne reste vraisemblablement que de très petit gisements résiduels.

- $t \mapsto v(t)$ *évolue en restant au dessus de* v_c. Dans ce cas, on peut dire que lorsque $v(t)$ augmente les gisements élémentaires ont tendance à s'agréger plus souvent et lorsque $v(t)$ diminue à se fragmenter plus souvent, ceci sans entraîner une disparition totale du pétrole. Cette évolution en terme de coalescence-fragmentation de la statistique des tailles relatives des gisements est justifiée par une étude précise des distributions Poisson-Dirichlet $PD(\alpha, 0)$, qui d'après les analyses exposées ci-dessus, constituent les modèles limites. L'étude des variations de $v \mapsto \mu_{v,N}$ a motivé l'introduction du coalescent/fragmentation de Bolthausen-Sznitman, qui définit ce mécanisme d'évolution en terme de coalescence/fragmentation réversible sur les lois limites. C'est donc un processus $\alpha \mapsto P^{\downarrow,\alpha}$ qui est réversible. La définition de ce processus est l'objet de la section 2.3.3.

Dans la suite de cette section, nous allons décrire une dynamique sur les partitions stables qui rende compte des fluctuations de $v(t)$. Cette dynamique introduite par Bolthausen et Sznitman pour étudier le REM, est un processus de coalescence que l'on appelle coalescent de Bolthausen-Sznitman (BS). Le processus "dual" que l'on obtient en renversant le temps est appelé fragmentation de Bolthausen-Sznitman. Ces deux processus ont la propriété remarquable d'être markoviens. Pour les définir nous avons besoin d'introduire tout d'abord une classe de partitions aléatoires plus large que les partitions stables : le modèle de Poisson-Dirichlet à deux paramètres, introduits par Kingman (1975).

2.3.2 Le modèle à deux paramètres de Poisson-Dirichlet

Nous adoptons la notation suivante,

$$(x)_{n\uparrow\alpha} := x(x+\alpha)\dots(x+(n-1)\alpha).$$

Le résultat suivant est dû à Kingman, il peut être trouvé dans Pitman (2006).

Théorème 2.3.1. *Pour tout* $\alpha \in]0,1[$ *et tout* $\theta > -\alpha$, *il existe une partition aléatoire* Π_∞ *de* $\mathbb{N}$ *dont la fppe vérifie*

$$p_{\alpha,\theta}(n_1,\dots,n_k) = \frac{(\theta+\alpha)_{k-1\uparrow\alpha}\prod_{i=1}^{k}(1-\alpha)_{n_i-1\uparrow 1}}{(\theta+1)_{n-1\uparrow 1}}.$$

Dans ce cas, la suite des fréquences asymptotiques des blocs, rangées par ordre d'apparition de leur plus petit élément, admet la représentation suivante v

$$(\tilde{P}_1, \tilde{P}_2 \dots) = (W_1, \bar{W}_1 W_2, \bar{W}_1\bar{W}_2 W_3, \dots)$$

où $\bar{W}_i = 1 - W_i$, *et les* W_i *sont des variables aléatoires indépendantes de loi beta*$(1-\alpha, \theta+i\alpha)$.

En ordonnant par ordre décroissant la suite des $(\tilde{P}_i)_{i\geq 1}$ des fréquences aléatoires de la partition de $\mathbb{N}$, on obtient ainsi une suite $(P_i^{\downarrow})_{i\geq 1}$ dont la distribution appelée distribution de Poisson-Dirichlet, et notée PD(α,θ). Il est aussi possible de définir le modèle à deux paramètres à partir des subordinateurs stables, cette construction est due à Perman, Pitman et Yor, ce qui permet de vérifier que la distribution PD$(\alpha, 0)$ correspond effectivement à la

distribution des sauts du subordinateur stable. La correspondance entre les deux points de vue est détaillée dans Pitman (2006).

Nous sommes maintenant en position de définir le modèle de fragmentation de Bolthausen-Sznitman afin de proposer une modélisation de la formation des réserves.

2.3.3 Modèle de fragmentation de Bolthausen-Sznitman

Pour décrire la fragmentation progressive des réserves initiales au cours du temps, nous commençons par rappeler la définition des notions de fragmentation et de coagulation pour des partitions de $\mathbb{N}$. Sur ce sujet, nous renvoyons le lecteur à Pitman (2006) et Bertoin (2006) qui traitent largement de ces questions.

Définition 2.3.1. 1. Soit $\Pi = (A_1, A_2, \ldots)$ une partition de $\mathbb{N}$, et soit $\Pi^{(\cdot)} = (\Pi^{(i)}, i = 1, \ldots, n)$ une suite de partitions de $\mathbb{N}$. On note $A_j^{(i)}$ le j-ème bloc de la partition $\Pi^{(i)}$. Pour tout entier i, on considère la partition $\Pi^{(i)}_{|A_i}$ de A_i induite par la i-ème partition $\Pi^{(i)}$ de la suite $\Pi^{(\cdot)}$, c'est-à-dire

$$\Pi^{(i)}_{|A_i} = \left(A_j^{(i)} \cap A_i,\, j \in \mathbb{N}\right).$$

La partition obtenue en rassemblant l'ensemble de tous les blocs de toutes les partitions $\Pi^{(i)}_{|A_i}$ forme une partition de $\mathbb{N}$ que l'on note $\mathrm{Frag}(\Pi, \Pi^{(\cdot)})$.

2. Soient $\Pi = (A_1, A_2, \ldots)$ et $\Pi' = (A'_1, A'_2, \ldots)$ deux partitions de $\mathbb{N}$. On appelle coagulation de Π par Π' la partition notée $\mathrm{Coag}(\Pi, \Pi')$ de blocs B_j définis par

$$B_j := \bigcup_{i \in A'_j} A_i.$$

Les processus de coalescence et de fragmentation sont des processus de Markov à valeur dans l'espace $\mathcal{P}_\infty$.

Définition 2.3.2. Soit $\mathbf{\Pi} = (\Pi(t), t \geq 0)$ un processus de Markov à valeur dans $\mathcal{P}_\infty$ qui est continue en probabilité. On dit que $\mathbf{\Pi}$ est un processus de coalescence échangeable si son semi-groupe peut être décrit de la façon suivante. Pour tout t et t' positifs, la distribution conditionnelle de $\Pi(t + t')$ sachant $\Pi(t) = \pi$ est la même que celle de $\mathrm{Coag}(\pi, \Pi')$ où Π' est une partition aléatoire échangeable dont la loi ne dépend que de t'.

L'un des exemples de processus de coalescence échangeable les mieux connus est le coalescent de Bolthausen-Sznitman. Commençons par rappeler le résultat suivant (voir Bertoin, 2006, p.207).

Lemme 2.3.1. *On fixe α et β dans $]0, 1[$. Soit Π une partition aléatoire de distribution PD$(\alpha, 0)$, et Π' une partition aléatoire de distribution PD$(\beta, 0)$. Alors, la partition* $\mathrm{Coag}(\Pi, \Pi')$ *est une partition aléatoire (échangeable) de distribution PD$(\alpha\beta, 0)$.*

On peut définir le coalescent de Bolthausen-Sznitman $\mathbf{\Pi}^{\mathrm{BS}}$ par ses probabilités de transition P_t^{BS} en considérant l'opérateur suivant sur l'espace des fonctions continues $\Phi : \mathcal{P}_\infty \to \mathbb{R}$,

$$P_t^{\mathrm{BS}}(\Phi)(\Pi) = \mathbb{E}\left[\Phi\left(\mathrm{Coag}(\Pi, \Pi^{(e^{-t})})\right)\right], \Pi \in \Pi_{\mathbb{N}}$$

où $\Pi^{(e^{-t})}$ est une partition aléatoire de distribution $\mathrm{PD}(e^{-t}, 0)$.

La représentation abstraite d'une partition dans $\mathcal{P}_\infty$ est moins significative que les fréquences qui lui sont associées par le théorème de Kingman car ces fréquences représentent les tailles relatives des gisements. Il convient donc ici de reformuler dans ce contexte le mécanisme de coalescence de Bolthausen-Sznitman : soit $(T^1_s, \geq 0)$ et $(T^2_s, s \geq 0)$ deux subordinateurs indépendants d'exposants respectifs $\alpha_1 = e^{-t_1}$ et $\alpha_2 = e^{-t_2}$. D'après la proposition 2.2.1, $T^1 \circ T^2$ est un subordinateur stable d'indice $\alpha_1\alpha_2 = e^{-(t_1+t_2)}$. On pose les notations suivantes.

- On note $(\Delta_n^{\alpha_1,\downarrow}\,;\, n \geq 0)$ les sauts $\{T^1_s - T^1_{s-}\,;\, s \in [0, T^2_1]\}$ indexés dans l'ordre décroissant.
- On note $(\Delta_n^{\alpha_1\alpha_2,\downarrow}\,;\, n \geq 0)$ les sauts $\{(T^1 \circ T^2)_s - (T^1 \circ T^2)_{s-}\,;\, s \in [0,1]\}$ indéxés dans l'ordre décroissant.

On considère ensuite un coalescent de Bolthausen-Sznitman $(\Pi^{BS}_t, t \geq 0)$ dont la valeur initiale est $\Pi^{BS}(0) = \mathbf{e}$. On note

$$P^{t_1\downarrow} = \left(P_n^{t_1\downarrow}\,;\, n \geq 0\right) \quad \text{et} \quad P^{t_1+t_2\downarrow} = \left(P_n^{t_1+t_2\downarrow}\,;\, n \geq 0\right)$$

les fréquences associées à respectivement $\Pi^{BS}_{t_1}$ et $\Pi^{BS}_{t_1+t_2}$ par le théorème de Kingman. Alors on peut montrer que

$$\left[\left(\frac{\Delta_n^{\alpha_1,\downarrow}}{T^1_{T^2_1}}\,;\, n \geq 0\right)\,;\,\left(\frac{\Delta_n^{\alpha_1\alpha_2,\downarrow}}{T^1_{T^2_1}}\,;\, n \geq 0\right)\right] \overset{\text{(loi)}}{=} \left[P^{t_1\downarrow}\,;\, P^{t_1+t_2\downarrow}\right].$$

(voir Bertoin et Le Gall, 2000, pour plus de détails sur cette construction).

Nous souhaitons aussi considérer ce processus en inversant le temps, de façon à observer une fragmentation progressive des réserves. De façon générale, il n'existe pas de dualité parfaite entre les processus de fragmentation et les processus de coalescence. Dans le cas du coalescent de Bolthausen-Sznitman, on peut cependant énoncer le résultat suivant (voir Bertoin, 2006, p.210).

Proposition 2.3.2. *Le processus $\mathbf{\Pi}^{BS}_f$ défini par $\Pi^{BS}_f(u) := \mathbf{\Pi}^{BS}(-\ln u)$ obtenu en reversant le temps dans le coalescent de Bolthausen-Sznitman est un processus de Markov sur $\mathcal{P}_\infty$ inhomogène dans le temps. Ses probabilités de transition sont telles que pour $0 < u \leq u' \leq 1$, conditionnellement à $\Pi^{BS}_f(u) = \pi$, la partition $\Pi^{BS}_f(u')$ est distribuée comme $\mathrm{Frag}(\pi, \Pi^{(.)})$, où $\Pi^{(.)} = (\Pi^{(1)}, \Pi^{(2)}, \ldots)$ est une suite i.i.d de partitions aléatoire de distribution $PD(u', -u)$.*

Commentaires

Le modèle de fragmentation de Bolthausen-Sznitman fournit ainsi un processus tel qu'à tout instant $u \in]0,1]$, $\Pi^{\mathrm{BS}}_f(u)$ est de distribution $\mathrm{PD}(-\ln u, 0)$. Le processus n'est pas défini en 0, et de façon à compléter la modélisation, on peut supposer que $\Pi^{\mathrm{BS}}_f(0)$ est la partition de $\mathbb{N}$ égal à l'ensemble $\mathbb{N}$ lui-même. En effet au temps 0, toutes les réserves du bassin sont concentrées dans un seul gisement initial. Le temps u qui indice le processus de fragmentation de Bolthausen-Sznitman doit être considéré comme un temps géologique, celui-ci ne donne pas l'âge du bassin, pour un temps qui s'écoulerait de façon linéaire. Ce temps géologique permet de structurer la suite des fragmentations successives des ressources initiales.

Nous vérifions que le modèle de fragmentation de Bolthausen-Sznitman vérifie les deux hypothèses $\mathbf{H_a}$ et $\mathbf{H_b}$ que nous avons imposées dans la section 2.1.1. En termes mathématiques, l'hypothèse $\mathbf{H_a}$ signifie que la fragmentation des réserves a un comportement markovien dans le temps, ce qui est bien le cas du processus $\mathbf{\Pi}_f^{\mathrm{BS}}$. L'hypothèse $\mathbf{H_b}$ signifie que toutes les partitions sont fragmentées selon des partitions aléatoires $(\Pi^{(1)}, \Pi^{(2)}, \ldots)$ qui sont de même distribution. D'après la proposition 2.3.2, cette contrainte est bien respectée par le processus $\mathbf{\Pi}_f^{\mathrm{BS}}$; les partitions qui dirigent la fragmentation entre les temps u et u' sont toutes de distribution $\mathrm{PD}(u', -u)$. Idéalement, nous souhaiterions que le seul processus de fragmentation vérifiant $\mathbf{H_a}$, $\mathbf{H_b}$ et $\mathbf{H_c}$ soit le processus markovien décrit dans la proposition 2.3.2, mais à notre connaissance un tel résultat n'a jamais été démontré.

Il nous a paru plus simple pour l'exposé de commencer par décrire les choses uniquement du point de vue de la fragmentation qui est plus facile à étudier mathématiquement. Cependant, puisque nous supposons que $v > v_c$, cette modélisation autorise les deux mécanismes duaux de coalescence et de fragmentation à se produire successivement. Ceci est en effet plus réaliste que de supposer que qu'un grand gisement unique ait été progressivement fragmenté au cours du temps. Des processus de fragmentation et de coalescence successifs peuvent donc avoir lieu et ceci sans que ne change le type de loi décrivant la taille des gisements du bassin : $PD(\alpha, 0)$, où seul α varie au cours du temps. Au final, les bassins restent distribués comme des sauts de subordinateurs stables, et si on les tronque comme des variables aléatoires de Lévy-Paréto.

2.4 Conclusions du chapitre

Dans ce chapitre, nous avons abordé la problème de la formation des réserves pétrolières avec un point de vue qualitatif et justifié ainsi l'utilisation de la distribution de Lévy-Paréto pour modéliser les tailles de gisements d'un bassin pétrolier. Le modèle pour la formation des réserves adapté du REM montre qu'il est naturel de représenter ces tailles par les sauts d'un subordinateur stable. De plus, dans le cadre du modèle de Bolthausen-Sznitman, ce type de distribution n'est pas changé par des opérations de coalescence ou de fragmentation des réserves (pour $v > v_c$). En ne considérant que les gisements de tailles supérieurs à un seuil ε, il est donc naturel d'observer des distributions de Lévy-Paréto. Notons de plus que ce modèle s'affranchit de la difficulté du choix du seuil minimal des distributions de Lévy-Pareto.

Nous allons montrer dans le chapitre suivant comment les propriétés du subordinateur stable peuvent être utilisées pour décrire l'exploration pétrolière dans un bassin en exploitation.

Chapitre 3

Modèle probabiliste pour l'exploration pétrolière

Nous proposons dans ce chapitre un premier modèle pour l'exploration pétrolière dans un bassin. Ce modèle est défini de façon volontairement simpliste, afin de disposer de résultats théoriques permettant de décrire la dynamique des découvertes successives au cours du temps. Les résultats présentés ont donc avant tout un intérêt qualitatif, et ne feront pas l'objet de procédures d'estimation. À la différence du chapitre précédant, le temps ne désigne plus ici le temps géologique de la formation du bassin, mais correspond au "temps humain" de l'exploitation de la zone.

3.1 Modélisation du processus de forage

Pendant toute la durée de l'exploitation d'un bassin, les compagnies pétrolières recherchent constamment de nouveaux gisements. Les champs pétroliers ont une durée de vie de quelques années à quelques dizaines d'années, et pour conserver des niveaux de production suffisants, de nouveaux gisements doivent être régulièrement découverts. La sismique de réflexion est aujourd'hui la technique la plus couramment utilisée, elle consiste à envoyer des ondes élastiques qui se propagent dans le sous-sol et dont les échos permettent de repérer les discontinuités de la roche. Des forages d'exploration[1] sont ensuite effectués pour confirmer la présence de pétrole ou de gaz. La date de découverte d'un gisement correspond à la date de forage de ce premier puis, et les découvertes dans un bassin dépendent de façon logique de la succession des forages effectués dans la zone. Dans cette section, nous décrivons "l'empreinte" du processus des forages d'exploration sur la population des gisements.

La figure 3.1 présente les dates de découverte en fonction des quantités de pétrole mis à jour pour l'ensemble des champs connus de la mer du Nord. La découverte d'un champ pétrolier est conditionnée par sa visibilité. Le principe selon lequel un champ de grande taille a plus de chance d'être rapidement découvert semble tout à fait naturel. Cette idée a été retenue et exploitée par de nombreux auteurs (voir Kaufman *et al.*, 1975; Bickel *et al.*, 1992; Campbell et Laherrère, 1998; Kontorovich *et al.*, 2001; Lepez, 2002). En réalité, d'autres paramètres

[1] Appelés "*wild cats*" dans le jargon pétrolier

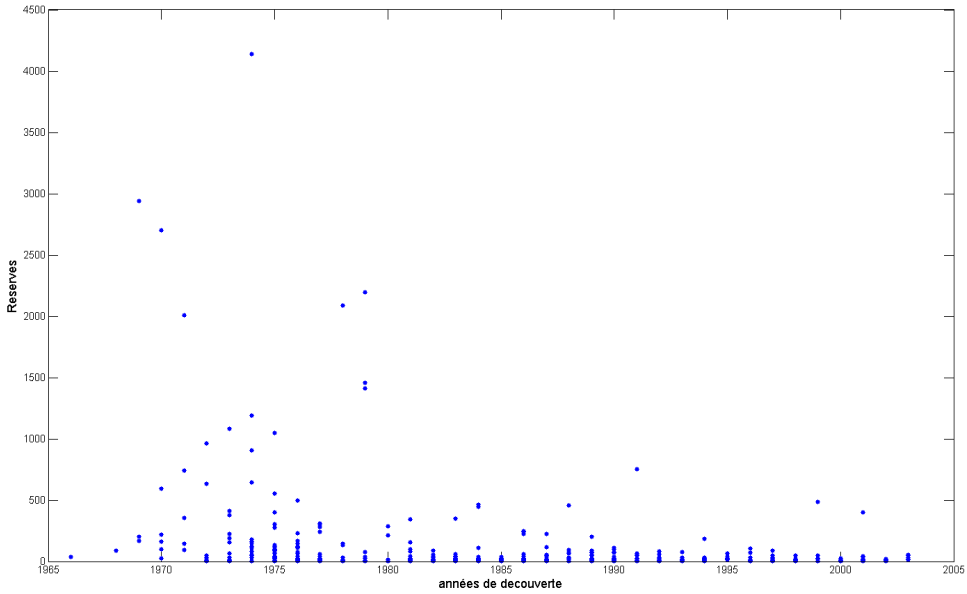


FIG. 3.1: Années de découverte et réserves de l'ensemble des champs en mer du Nord.

influencent la découverte d'un gisement, comme par exemple l'accessibilité géographique, la législation de la zone, la proximité d'autres gisements déjà découverts, etc... Il n'est évidemment pas possible d'intégrer de façon exhaustive tous ces éléments dans un unique travail de modélisation, et la taille du gisement reste le facteur le plus important de ce mécanisme. En conséquence, nous adoptons nous aussi le principe exposé plus haut ; nous supposons désormais que la fonction de visibilité, notée vis, est une fonction de la taille du gisement.

Nous avons vu au chapitre 2 que les tailles des champs d'un bassin pétrolier peuvent être représentées par les sauts d'un subordinateur stable T d'indice α. Soit $(\Delta_i^{\downarrow})_{i\geq 1}$ la suite des sauts de T, ordonnés par ordre décroissant. Nous considérons une fonction de visibilité vis définie de $\mathbb{R}^{+*}$ (l'espace des tailles de gisements) dans $\mathbb{R}^+$. La fonction vis est de plus croissante, de façon à nous accorder avec le principe rappelé plus haut. Soit Φ une mesure aléatoire définie sur $\mathbb{R}^+ \times \mathbb{R}^+$ telle que, conditionnellement au subordinateur T, Φ est une mesure de Poisson uniforme sur $[0, T(vis)] \times \mathbb{R}^+$ de mesure de Lévy $c\,dxdt$, où $T(v) = \sum_{i\geq 1} v(\Delta_i^{\downarrow})$. La mesure Φ représente l'empreinte du processus de forage sur la population des champs, et le coefficient c traduit l'intensité de l'effort d'exploration dans le bassin. L'intervalle $J_i := \left[\sum_{u=0}^{i-1} vis(\Delta_u^{\downarrow}), \sum_{u=0}^{i} vis(\Delta_u^{\downarrow})\right]$ est associé au champ d'indice i, et sa largeur est égale à la visibilité $vis(\Delta_i^{\downarrow})$ du champ, avec la convention $vis(\Delta_0^{\downarrow}) = 0$. Puisque la mesure de Poisson Φ est homogène, le nombre d'atomes dans l'intervalle $J_i \times [0, t]$ est en moyenne proportionnelle à la visibilité du champ i. La figure 3.2 illustre cette modélisation.

Pour tout $i > 0$, soit Φ_i la restriction de Φ à la tranche $J_i \times \mathbb{R}^+$ correspondant à l'exploration du champ i. Nous pouvons écrire $\Phi_i = \sum_{u\in J_i} \delta_{(x_u, s_{ui})}$ où les s_{ui} sont supposés ordonnés de façon croissante. Conditionnellement au subordinateur T, la mesure aléatoire Φ_i est une mesure de Poisson uniforme sur la tranche $J_i \times \mathbb{R}^+$, de mesure de Lévy $vis(\Delta_i^{\downarrow})c\,dxdt$. La date de découverte du champ est alors naturellement définie par $D_i := s_{1i}$. Par construction,

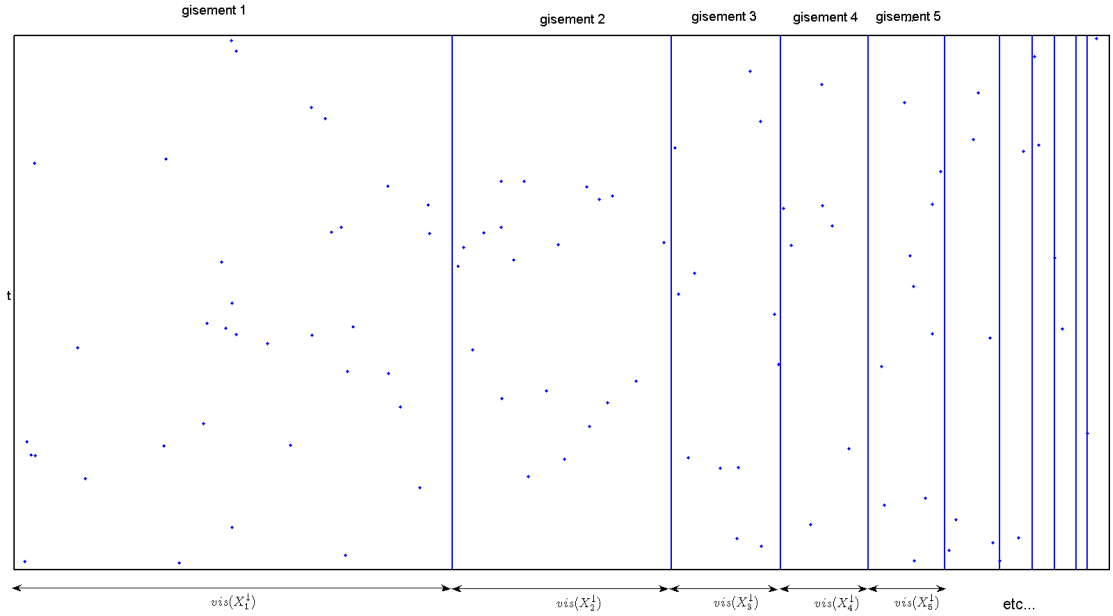


FIG. 3.2: Modélisation de l'empreinte du processus de forage sur la population des champs d'un bassin.

conditionnellement à T, les temps de découverte $(D_i)_{i\geq 1}$ sont indépendants et de distributions exponentielles de paramètres $c\,vis(\Delta_i^{\downarrow})$:

$$(D_i \,|\, T) \sim \mathcal{E}\left(c\,vis(\Delta_i^{\downarrow})\right), \tag{3.1}$$

Dans la suite, nous ne souhaitons pas distinguer c et vis, et nous posons $h := c\,vis$. Puisque h et vis sont identiques à une constante multiplicative près, nous qualifions encore h de fonction de visibilité.

Selon la distribution donnée en (3.1), le temps d'attente moyen, avant de découvrir un gisement de taille x, est donc égal à l'inverse de sa visibilité $h(x)$. Il est maintenant possible d'ordonner la famille des tailles des champs du bassin selon leur date de découverte : soit $(\tilde{\Delta}_j)_{j\geq 1}$ la suite des réserves $(\Delta_i^{\downarrow})_{i\geq 1}$ ordonnée dans l'ordre croissant des dates de découverte, et soit $(\tilde{D}_j)_{j\geq 1}$ la suite croissante des dates de découverte correspondantes. Nous formalisons la notion de permutation aléatoire associée à la fonction de visibilité de la façon suivante :

Définition 3.1.1. Soit A un ensemble quelconque et $(x_i)_{i\geq 1}$ une suite d'éléments de A. Soit une fonction H définie de A vers R^+ telle que $\sum_{i\geq 1} H(x_i) < \infty$. Soit $(I_j)_{j\geq 1}$ la suite des indices (aléatoires) correspondant à une permutation aléatoire biaisée par la taille de la suite $(H(x_i))_{i\geq 1}$. C'est-à-dire que pour le premier tirage, $\mathbb{P}(I_1 = i) = H(x_i)\left(\sum_{l\geq 1} H(x_l)\right)^{-1}$ puis pour $i \neq I_1$, $\mathbb{P}(I_2 = i|I_1) = H(x_i)\left(\sum_{l\neq I_1} H(x_l)\right)^{-1}$ pour le deuxième tirage, etc... On dit alors que la suite $\left(x_{I_j}\right)_{j\geq 1}$ est une permutation aléatoire de $(x_i)_{i\geq 1}$ biaisé selon la fonction H.

La proposition suivante montre que la modélisation précédente, basée sur le processus des forages, conduit à définir la suite des découvertes de gisements comme une permutation aléatoire de $(\Delta_i^{\downarrow})_{\geq 1}$, biaisée selon la fonction de visibilité.

Proposition 3.1.1. *Sous les hypothèses précédentes et conditionnellement au subordinateur T, la suite des réserves découvertes $(\tilde{\Delta}_i)_{i\geq 1}$ est de même loi que celle d'une permutation aléatoire de $(\Delta_i^{\downarrow})_{\geq 1}$ biaisée selon la fonction de visibilité h.*

Démonstration. Pour tout $i \geq 1$,

$$\begin{aligned}
\mathbb{P}(\tilde{\Delta}_1 = \Delta_i^{\downarrow} \,|\, T) &= \mathbb{E}\Big[\mathbb{P}(\forall u \neq i, D_u > D_i \,|\, D_i)|T\Big] \\
&= \mathbb{E}\left[\prod_{u\neq i} \exp\Big\{-h(\Delta_i^{\downarrow})D_u\Big\}\,|T\right] \\
&= \mathbb{E}\Big[\exp\Big\{-(T(h) - h(\Delta_i^{\downarrow}))D_i\Big\}\,|T\Big] \\
&= \frac{h(\Delta_i^{\downarrow})}{T(h)}\,.
\end{aligned}$$

On montre ensuite de la même façon que pour tout $j > 1$

$$\mathbb{P}(\tilde{\Delta}_j = \Delta_i^{\downarrow} \,|\, T, \tilde{\Delta}_1 \ldots, \tilde{\Delta}_{j-1}) = \frac{h(\Delta_i^{\downarrow})}{T(h) - \sum_{u=1}^{j} h(\tilde{\Delta}_u)}\,.$$

La suite $(\tilde{\Delta}_j)_{j\geq 1}$ est donc une permutation aléatoire de $(\tilde{\Delta}_i)$ biaisée selon la fonction de visibilité h. □

3.2 Tirage proportionnel à la taille

Pour de nombreux auteurs, comme par exemple Kaufman *et al.* (1975), Bickel *et al.* (1992), ou encore Kontorovich *et al.* (2001), il est naturel de supposer que le biais du tirage dans la population des champs du sous-sol est directement proportionnel à la taille des gisements, ce qui revient à choisir une fonction de visibilité proportionnelle à la taille. En effet, d'un point de vue géométrique, un champ pétrolier a une visibilité proportionnelle à l'aire de sa projection verticale à la surface du bassin. Si la roche réservoir est d'épaisseur constante sur l'ensemble de la zone, cette visibilité est alors directement proportionnelle à son volume, que l'on assimile à la quantité d'hydrocarbures qu'il contient. En réalité, les couches géologiques sont souvent inclinées, et les hydrocarbures sont confinés dans des pièges qui ne sont pas des pavés parfaits. Il convient donc de supposer plutôt que la visibilité est une fonction puissance de la taille des gisements,

$$h(x) = \gamma\, x^{\beta}, \tag{3.2}$$

où γ et β sont constantes positives inconnues. Dans la suite de cette section, nous n'étudions cependant que le cas $\beta = 1$. En effet, le lemme suivant permet de ramener une partie des situations à celle où les champs sont découverts selon un tirage biaisé par la taille.

Lemme 3.2.1. *Soit $T_t := \sum_{i\geq 1} \Delta_i^{\downarrow} \mathbb{1}_{U_i \geq t}$, $t \in [0,1]$ un subordinateur stable d'indice $\alpha \in]0,1[$, où les U_i sont des variables i.i.d. de loi uniforme sur $[0,1]$. Soit $\beta > \alpha$. Alors, le processus défini par*

$$T_t^{(\beta)} := \sum_{i\geq 1} \left(\Delta_i^{\downarrow}\right)^{\beta} \mathbb{1}_{U_i \geq t},\ t \in [0,1]$$

est un subordinateur stable d'indice $\alpha\beta^{-1}$.

Démonstration. Le processus T est un subordinateur stable, donc $\{U_i, \Delta_i\}$ est l'ensemble des points d'un processus ponctuel de Poisson d'intensité $ds\Lambda(dx)$ sur $[0,\infty[^2$, avec Λ de densité la fonction $s_0\rho$, où $\rho(x) = \frac{\alpha}{\Gamma(1-\alpha)}x^{-\alpha-1}\mathbb{1}_{x>0}$. Le nuage $\{(U_i, \Delta_i^\beta), i \leq 1\}$ est l'image de ces points par l'application mesurable $\Psi : (u,x) \mapsto (u, x^\beta)$. L'ensemble de ces points est donc distribué comme les points d'un processus de Poisson ponctuel d'intensité la mesure image de $ds\Lambda(dx)$ par Ψ, c'est-à-dire $ds\tilde{\Lambda}(dx)$ avec $\tilde{\Lambda}(dx)([c,d]) = \frac{s_0}{\Gamma(1-\alpha)}(c^{-\alpha/\beta} - d^{-\alpha/\beta})$. Le processus $T^{(\beta)}$ est bien un subordinateur stable d'indice $\alpha\beta^{-1}$. □

Quitte à considérer $T^{(\beta)}$ à la place de T, nous supposons donc que $\beta = 1$. Soit $(\tilde{\Delta}_j)_{j\geq 1}$ la permutation aléatoire biaisée par la taille de la suite des sauts $\Delta_i^\downarrow$ d'un subordinateur T d'indice α. Les $\tilde{\Delta}_j$ représentent donc la taille des champs pétroliers ordonnés selon leur date de découverte. Pour tout $k > 0$, la variable aléatoire $\tilde{T}_k := \sum_{j\geq k+1}\tilde{\Delta}_j$ désigne la quantité d'hydrocarbures encore à découvrir dans le bassin après la k-ème découverte. Le résultat suivant, dû à Perman *et al.* (1992), montre que la suite des découvertes restantes peut être décrite par une chaîne de Markov.

Théorème 3.2.1. *La suite* $(\tilde{T}_j)_{j\geq 1}$ *forme une chaîne de Markov de probabilités de transition stationnaires*

$$\mathbb{P}(\tilde{T}_{j+1} \in dx_1 | \tilde{T}_j = x) = \frac{\rho_*(x-x_1)}{x}\frac{\tau_\alpha(x_1)}{\tau_\alpha(x)}dx_1, \quad 0 \leq x_1 \leq x, \tag{3.3}$$

où τ_α *est la densité de la v.a.* T_1 *pour un subordinateur stable* T *et où* $\rho_*(x) = x\rho(x)$.

Ce résultat repose sur une propriété fondamentale des mesures de Poisson : la formule de Palm (Bertoin, 2006, p.79). Notons que le comportement markovien de la suite des $\tilde{T}_j$ peut être vérifié pour d'autres subordinateurs que le subordinateur stable, sous des conditions de régularité suffisantes (voir Perman *et al.*, 1992). En revanche, pour des fonctions de visibilité différentes de $h(x) = \gamma\, x$, nous ne disposons plus de ce comportement markovien.

3.2.1 Un théorème difficile à appliquer

À première vue, ce résultat semble l'outil idéal pour étudier le processus des découvertes puisqu'il en décrit complètement la loi. En réalité, le théorème se révèle difficile à exploiter pour plusieurs raisons. Tout d'abord, la chaîne de Markov concerne la suite des réserves restantes, et non la suite des découvertes successives. Celle-ci ne peut donc être utilisée en pratique que si le potentiel total du bassin T_1 est connu avec une bonne précision, ce qui est en fait bien délicat. La méthode proposée par Lepez (2002) peut en fournir une approximation pour des bassins suffisamment matures, mais il faut bien noter que la quantité ainsi obtenue correspond à une estimation de la somme des champs de taille supérieure à un certain seuil, et non la somme des réserves de tous les champs.

La mise en oeuvre de simulations basées sur ce résultat se heurte aussi à des difficultés sérieuses. Pour comprendre pourquoi la manipulation de ces probabilités de transition est délicate, nous rappelons ci-dessous l'expression de la densité de la v.a. T_1 pour un subordinateur

stable de mesure de Lévy de densité $\rho(x)$ (Pollard, 1946) :

$$\tau_\alpha(x) = \frac{1}{\pi}\sum_{k=0}^{\infty}\frac{(-1)^{k+1}}{k!}\sin(\pi\alpha k)\frac{\Gamma(\alpha k+1)}{x^{\alpha k+1}},\ x \geq 0. \tag{3.4}$$

Notons de plus que pour le subordinateur de mesure de Lévy $s_0\rho$, la v.a. T_1 admet pour densité

$$\tau_{\alpha,s_0}(x) = s_0^{-1/\alpha}\tau_\alpha(s_0^{-1/\alpha}x). \tag{3.5}$$

Pour être capable de simuler les lois de transition en utilisant par exemple une méthode d'acceptation-rejet, il faut obligatoirement pouvoir évaluer numériquement le rapport $\frac{\tau_\alpha(x_1)}{\tau_\alpha(x)}$. Dans l'expression de la densité donnée en (3.4), nous pouvons vérifier que pour x proche de 0, ce sont les termes pour des k très grands qui sont prépondérants, et ces termes oscillent entre des valeurs positives et négatives avec une très grande amplitude. Pour des valeurs de x trop faibles, il est alors impossible d'implémenter cette expression dans un logiciel de calcul. Or, une fois les champs les plus grands rapidement découverts, l'expression des probabilités de transition dans (3.3) nécessite ensuite une bonne précision de cette densité en 0. De plus, ce phénomène est encore accentué pour la densité (3.5) correspondant à la modélisation pétrolière qui nous intéresse ici. Une estimation grossière indique en effet que s_0 est de l'ordre du millier pour le cas de la mer du Nord.

Toutes ces remarques soulignent la difficile application du théorème 3.2.1. Nous achevons cette discussion avec la proposition suivante, qui montre de façon plus rigoureuse que la densité de T_1 tend vers 0 en 0 à grande vitesse.

Proposition 3.2.2. *Soit V une variable aléatoire de distribution uniforme sur $[0,\pi]$. Soit la fonction m définie sur $\mathbb{R}$ par*

$$m(v) = \left[\frac{\sin(\alpha v)}{\sin v}\right]^{\frac{\alpha}{1-\alpha}}\frac{\sin((1-\alpha)v)}{\sin v}.$$

Alors,

$$\tau_\alpha(x) = \frac{\alpha}{1-\alpha}\,x^{-\frac{1}{1-\alpha}}\,\mathbb{E}\left[m(V)\exp\left(-x^{-\frac{\alpha}{1-\alpha}}\,m(V)\right)\right] \tag{3.6}$$

et pour x proche de 0,

$$\tau_\alpha(x) \sim \sqrt{\alpha(1-\alpha)}\alpha^{\frac{\alpha}{2-2\alpha}}\sqrt{\frac{2}{\pi}}x^{-\frac{2-\alpha}{2-2\alpha}}\exp\left(-(1-\alpha)\left(\frac{\alpha}{x}\right)^{\frac{\alpha}{1-\alpha}}\right).$$

Démonstration. Il existe une représentation plus pratique de la variable aléatoire T_1 permettant de simuler des lois stables. Dans Chambers *et al.* (1976), les auteurs utilisent l'égalité en loi suivante

$$T_1 = \frac{\sin(\alpha(U+\pi/2))}{(\cos U)^{1/\alpha}}\left(\frac{\cos(U-\alpha(U+\pi/2))}{W}\right)^{\frac{1-\alpha}{\alpha}},$$

où U est une variable aléatoire de loi uniforme sur $[-\pi/2,\pi/2]$, et W est une variable aléatoire de loi exponentielle de paramètre 1. En posant $V = U+\pi/2$, ceci peut aussi s'écrire sous la

forme

$$T_1 = \left(\frac{m(V)}{W}\right)^{\frac{1-\alpha}{\alpha}},$$

où V est de loi uniforme sur $[0, \pi]$. La fonction de répartition de T_1 en x vaut donc

$$\begin{aligned}\mathbb{P}(T_1 \leq x) &= \mathbb{P}\left(x^{-\frac{\alpha}{1-\alpha}} m(V) \leq W\right) \\ &= \mathbb{E}\left[\exp\left\{-x^{-\frac{\alpha}{1-\alpha}} m(V)\right\}\right].\end{aligned}$$

On en déduit la densité de T_1 en $x > 0$,

$$\tau_\alpha(x) = \frac{\alpha}{1-\alpha} x^{-\frac{1}{1-\alpha}} \mathbb{E}\left[m(V) \exp\left(-x^{-\frac{\alpha}{1-\alpha}} m(V)\right)\right].$$

La preuve du deuxième point s'appuie sur la méthode de Laplace ; le calcul est détaillé dans la section 3.5.1 à la fin du chapitre. □

3.2.2 Loi des proportions découvertes successives

Comme nous l'avons remarqué plus haut, le comportement markovien décrit dans le théorème 3.2.1 concerne la suite des réserves restantes $\tilde{T}_j$, et non les découvertes successives. La suite des découvertes successives peut être décrire via les sauts normalisés

$$\tilde{P}_j = \frac{\tilde{\Delta}_j}{T_1},$$

où $\tilde{P}_j$ représente ici la proportion d'hydrocarbures du bassin contenue dans le j-ème gisement découvert. D'après la proposition 2.3.1 du chapitre précédent, la suite des $(\tilde{P}_j)$ admet la représentation suivante

$$\tilde{P}_j = W_j \prod_{i=1}^{j-1} (1 - W_i), \tag{3.7}$$

où les W_i sont des v.a. indépendantes de loi $\text{Beta}(1-\alpha, i\alpha)$. Cependant, comme précédemment, l'application de ce résultat implique une estimation de T_1. Plutôt que de chercher à obtenir des intervalles de confiance à partir de (3.7) qui ne seraient de toute façon pas exploitables en pratique, nous préférons adopter un point de vue plus qualitatif en étudiant maintenant le comportement asymptotique des découvertes successives.

3.3 Dynamique asymptotique de l'exploration

Nous rappelons dans cette partie quelques résultats sur le comportement asymptotique des proportions aléatoires P_j d'une distribution $\text{PD}(\alpha, 0)$. Ces résultats peuvent être partiellement retrouvés dans Pitman (2006). Par souci de clarté, nous avons complété en annexe les preuves correspondantes qui ne sont que partiellement exposées dans cet ouvrage.

Comme précédemment, les tailles de gisements pétroliers sont modélisées par les sauts $\Delta_j^{\downarrow}$ d'un subordinateur stable T de mesure de Lévy d'intensité $s_0\, \rho$.

Proposition 3.3.1. *1. Soit $(e_1, \ldots, e_k)$ un échantillon i.i.d. de variables aléatoires de loi exponentielle de paramètre $\frac{s_0}{\Gamma(1-\alpha)}$. Alors, pour tout $k \geq 1$, $\left(\Delta_1^{\downarrow -\alpha}, \Delta_2^{\downarrow -\alpha} \ldots, \Delta_k^{\downarrow -\alpha}\right)$ a même distribution que $(e_1, e_1 + e_2, \ldots, e_1 + \ldots + e_k)$.*

2. Avec probabilité 1,

$$\Delta_i^{\downarrow} i^{1/\alpha} \longrightarrow \left(\frac{s_0}{\Gamma(1-\alpha}\right)^{1/\alpha} \quad \textit{lorsque } i \longrightarrow +\infty. \tag{3.8}$$

Démonstration. Soit la fonction $L(t) = \int_t^\infty s_0 \rho(x)\, dx$. Le processus $\left(L(\Delta_i^{\downarrow})\right)_{i \leq 1}$ est alors un processus de Poisson homogène d'intensité 1 (voir par exemple Kingman, 1993, p.52). Les deux points de propositions se déduisent directement de la distribution de $\left(L(\Delta_i^{\downarrow})\right)_{i \geq 1}$. En particulier, (3.8) est une application directe de la loi des grands nombres. □

En passant au logarithme dans (3.8), nous voyons que la suite des sauts rangés par ordre décroissant a un comportement linéaire de pente $-\frac{1}{\alpha}$ dans un diagramme loglog. Nous retrouvons ici un phénomène déjà mentionné au début du chapitre 2 au sujet des échantillons de Lévy-Pareto (voir figure 2.1). Ceci est naturel puisque, comme nous l'avons remarqué auparavant, l'échantillon des sauts d'un subordinateur stable supérieurs à un seuil fixé, a la même distribution qu'un échantillon i.i.d. de variables aléatoires de loi de Lévy Pareto de paramètre l'indice du subordinateur, conditionnellement à la longueur de l'échantillon.

La proposition suivante détermine le comportement asymptotique des sauts normalisés, ordonnés de façon décroissante d'une part, et par ordre de découverte d'autre part.

Proposition 3.3.2. *Sous les mêmes notations que précédemment, et avec probabilité* 1,

$$P_i^{\downarrow} \sim \left(\frac{s_0}{\Gamma(1-\alpha)}\right)^{1/\alpha} \frac{1}{T_1}\, i^{-1/\alpha} \quad \textit{lorsque } i \to +\infty, \tag{3.9}$$

et

$$\left(1 - \sum_{i=1}^{k} \tilde{P}_i\right) \sim \frac{\alpha\, s_0^{1/\alpha}}{T_1}\, k^{1-1/\alpha} \quad \textit{lorsque } k \to +\infty. \tag{3.10}$$

La preuve de cette proposition dont les idées principales sont données par Pitman (2006) est exposée dans la section 3.5.2 à la fin du chapitre.

Commentaires

Le deuxième point de la proposition 3.3.2 établit un équivalent de la proportion restante dans le bassin au moment de la k-ième découverte. Nous considérons ce résultat avec le point de vue qualitatif suivant : la géologie du sous-sol, à travers le coefficient α, détermine à quelle vitesse le sous-sol est progressivement épuisé. Le fait que la géologie soit déterminante dans la dynamique des découvertes successives est le principe essentiel que nous retenons de ce modèle "idéalisé" de l'exploration pétrolière. Ainsi, pour un habitat dispersé (α proche de 1), l'épuisement est plus lent que pour un habitat concentré (α proche de 0). Ceci est naturel puisque dans ce dernier cas, quelques gisements contiennent à eux seuls l'essentiel des réserves, et ceux-ci sont trouvés rapidement.

Notons pour finir que les résultats de la proposition 3.3.2 sont ici encore difficiles à exploiter de façon quantitative. A la différence de l'équivalence (3.9) que nous avons pu vérifier dans des diagrammes loglog, l'équivalence (3.10) n'est validée que très grossièrement sur les données de bassins exploités. Différents éléments peuvent expliquer cet écart entre la théorie et les observations. Tout d'abord, les proportions sont définies en fonction de T_1, et celles-ci peuvent être en pratique mal estimées. Ensuite, il est possible que quelques centaines de champs découverts ne suffisent pas pour atteindre le "régime" de l'équivalence donnée en (3.10). Il est surtout probable que le choix d'une fonction de visibilité de type "proportionnelle" ou "puissance" ne soit pas pertinent.

3.4 Conclusions du chapitre

Le modèle probabiliste complet proposé dans ce chapitre et le précédent constitue, à notre connaissance, la première tentative de modélisation capable de décrire dans un unique cadre mathématique à la fois la formation des réserves, la distribution des tailles de gisements d'un bassin, et la dynamique des découvertes successives dans ce dernier. De ce fait, la cohérence du modèle est réellement satisfaisante.

Sur la base de résultats théoriques avérés, le modèle élémentaire pour l'exploration pétrolière met en évidence l'impact de la géologie sur la dynamique du processus des découvertes. Cependant, nous avons vu que ce modèle idéalisé se révèle difficile à utiliser en pratique. En particulier, le choix d'une fonction de visibilité proportionnelle à la taille des gisements ne semble pas suffisamment réaliste. Nous verrons dans le chapitre suivant qu'il est préférable de considérer le problème de l'estimation de la fonction de visibilité selon un point de vue non paramétrique. De l'étude de ce modèle probabiliste, nous retenons que le temps de découverte d'un gisement pétrolier de taille x peut être modélisé par une loi exponentielle de paramètre $h(x)$ où h est appelée fonction de visibilité. La modélisation générale de la production pétrolière présentée dans le chapitre suivant, plus opérationnelle que celle développée dans les chapitres 2 et 3, reposera sur cette hypothèse fondamentale.

3.5 Preuves

3.5.1 Preuve de la proposition 3.2.2

Le premier point de la proposition 3.2.2 peut s'écrire sous la forme suivante

$$\tau_\alpha(x) = \frac{\alpha}{1-\alpha} x^{-\frac{1}{1-\alpha}} \frac{1}{\pi} J_\alpha\left(x^{-\frac{\alpha}{1-\alpha}}\right) \tag{3.11}$$

avec

$$J_\alpha(\lambda) = \int_0^\pi m(v) \exp\left(-\lambda m(v)\right) dv.$$

Nous cherchons donc un équivalent de J_α en 0. Pour cela, nous commençons par vérifier le lemme suivant.

Lemme 3.5.1. *Sous les notations de la proposition 3.2.2,*

1. *La fonction m est strictement croissante sur $[0,\pi]$.*
2. *La fonction m admet le développement limité suivant en 0 :*

$$m(v) = \alpha^{\frac{1}{1-\alpha}}\left(1+\frac{v^2\alpha}{2}+O(v^4)\right).$$

Démonstration. Pour $\beta \in]0,1[$, on considère la fonction h définie sur $]0,\pi[$ par $h(v) = \frac{\sin(\beta v)}{\sin v}$. On a alors $h'(v) = \frac{k(v)}{\sin^2 v}$ avec $k(v) = \beta\cos(\beta v)\sin v - \cos v\sin(\beta v)$. Or, $k'(v) = ((1-\beta^2)\sin(\beta v)\sin v > 0$ sur $]0,\pi[$, et puisque $k(0)=0$, la fonction k est donc strictement positive sur $]0,\pi[$, de même que la fonction h'. Puisque $h'(0)=0$, on en déduit que la fonction h est donc strictement croissante sur $]0,\pi[$.

Nous commençons par calculer le développement limité de la fonction h en 0,

$$\begin{aligned} h(v) &= \frac{\beta v - \frac{\beta}{6}v^3 + O(v^4)}{v - \frac{v^3}{6} + O(v^5)} \\ &= \left(\beta - \frac{\beta}{6}v^2 + O(v^4)\right)\left(1+\frac{v^2}{6}+O(v^4)\right) \\ &= \beta + \beta\frac{1-\beta^2}{6}v^2 + O(v^4). \end{aligned}$$

On en déduit le développement limité de la fonction m en 0,

$$\begin{aligned} m(v) &= \left[\alpha + \frac{\alpha(1-\alpha^2)}{6}v^2 + O(v^4)\right]^{\frac{\alpha}{1-\alpha}}\left[1-\alpha+\frac{(1-\alpha)(1-(1-\alpha)^2)}{6}v^2+O(v^4)\right] \\ &= (1-\alpha)\alpha^{\frac{\alpha}{1-\alpha}}\left[1+\frac{(1-\alpha^2)}{6}v^2+O(v^4)\right]^{\frac{\alpha}{1-\alpha}}\left[1+\frac{1-(1-\alpha)^2}{6}v^2+O(v^4)\right] \\ &= (1-\alpha)\alpha^{\frac{\alpha}{1-\alpha}}\left[1+\frac{\alpha(1+\alpha)}{6}v^2+O(v^4)\right]\left[1+\frac{2\alpha-\alpha^2}{6}v^2+O(v^4)\right] \\ &= (1-\alpha)\alpha^{\frac{\alpha}{1-\alpha}}\left[1+\frac{\alpha}{2}v^2+O(v^4)\right]. \end{aligned}$$

□

Le lemme précédent nous permet donc d'appliquer la méthode de Laplace à la fonction (m)(voir par exemple Gourdon, 1994, p.161). Pour appliquer le résultat de cette référence, on peut considérer la fonction m sur $[-\pi,\pi]$, qui est paire sur cet intervalle. Ceci donne que pour λ tendant vers l'infini,

$$\begin{aligned} J_\alpha(\lambda) &\sim \sqrt{\frac{2\pi}{\lambda}}\exp\left(-\lambda m(0)\right)\frac{m(0)}{\sqrt{m''(0)}} \\ &\sim \sqrt{\frac{2\pi(1-\alpha)}{\lambda}}\alpha^{-\frac{2\alpha-1}{2-2\alpha}}\exp\left(-\lambda(1-\alpha)\alpha^{\frac{\alpha}{1-\alpha}}\right). \end{aligned}$$

En posant $\lambda = x^{-\frac{\alpha}{1-\alpha}}$ comme en (3.11), on obtient l'équivalent suivant de τ_α pour x proche de 0,

$$\tau_\alpha(x) \sim \sqrt{\alpha(1-\alpha)}\alpha^{\frac{\alpha}{2-2\alpha}}\sqrt{\frac{2}{\pi}}x^{-\frac{2-\alpha}{2-2\alpha}}\exp\left(-(1-\alpha)\left(\frac{\alpha}{x}\right)^{\frac{\alpha}{1-\alpha}}\right).$$

3.5.2 Preuve de la proposition 3.3.2

Le premier point de la proposition 3.3.2 est un corollaire immédiat de (3.8), et le second correspond à une version aléatoire du résultat obtenu par Karlin (1967) pour des fréquences déterministes.

Soit Π_∞ la distribution de masse aléatoire induite par la suite des sauts normalisées du subordinateur T. Par la représentation de Kingman, la distribution de masse Π_∞ peut être vue comme une partition aléatoire de $\mathbb{N}$, et la distribution conditionnelle de Π_∞ sachant $(P_i^\downarrow, i \geq 1)$ est la même que si les classes de Π_∞ étaient obtenues par tirage aléatoire à partir d'une distribution définie par les atomes $(P_i^\downarrow, i \geq 1)$. Nous établissons d'abord (3.10) sous la loi conditionnelle de Π_∞, sachant les fréquences $(P_i^\downarrow)$.

Considérons le schéma aléatoire suivant. Nous disposons d'une suite infinie de boîtes numérotées. Des boules sont placées aléatoirement dans les boîtes de façon indépendante et selon la même loi de probabilité. Plus précisément une boule est placée dans la boîte i avec la probabilité (déterministe) $p_i := P_i^\downarrow$. Les éléments de la suite $(\tilde{P}_j)_{j\geq 1}$ correspondent aux probabilités des boîtes ordonnées aléatoirement, selon l'ordre de remplissage progressif des boîtes. Soit $Z_N^\star$ le nombre de boîtes non vides après que N boules aient été placées. On considère aussi la quantité

$$D_N^\star = 1 - \sum_{j=1}^{Z_N^\star} \tilde{P}_j,$$

qui représente la somme des probabilités des boîtes encore vides à l'étape N.

On dit qu'une fonction r sur $\mathbb{R}$ est une fonction à variation lente si et seulement si $r(cx)/r(x) \to 1$ lorsque $x \to +\infty$. On pose aussi

$$\alpha(x) := \max\{i \,|\, p_i > 1/x\} \ .$$

Cette fonction joue un rôle important dans l'étude des moments de $Z_N^\star$ et $D_N^\star$. Nous utilisons maintenant les deux lemmes suivants démontrés par Karlin (1967). Le premier décrit le comportement asymptotique des espérances de $\mathbb{E}(Z_N^\star)$ et $\mathbb{E}(D_N^\star)$, et le deuxième nous donne une loi des grands nombres pour ces deux quantités.

Lemme 3.5.2. *Supposons que la suite des p_i soit telle que $\alpha(x) = x^\eta r(x)$, $0 < \eta < 1$ où $r(x)$ est une fonction à variation lente. Alors, avec probabilité* 1,

$$\mathbb{E}(Z_N^\star) \sim \Gamma(1-\eta) N^\eta \, r(N) \quad \textit{lorsque } N \to +\infty$$

, et

$$\mathbb{E}(D_N^\star) \sim \eta\, \Gamma(1-\eta) N^{\eta-1} \, r(N) \quad \textit{lorsque } N \to +\infty.$$

Lemme 3.5.3. *1. Avec probabilité 1, $Z_N^\star/\mathbb{E}(Z_N^\star) \to 1$ lorsque $N \to +\infty$.*

2. Supposons que $\alpha(x) = x^\eta r(x)$, $0 < \eta < 1$ où $r(x)$ est une fonction à variation lente. Alors, avec probabilité 1, *$D_N^\star/\mathbb{E}(D_N^\star) \to 1$ lorsque $N \to +\infty$.*

D'après (3.9), et sachant $(P_i^\downarrow)$, nous avons que $p_i \sim C i^{-1/\alpha}$ lorsque $i \to +\infty$, avec $C = \left(\frac{s_0}{\Gamma(1-\alpha)}\right)^{1/\alpha} \frac{1}{T_1}$. Notons que (3.9) nous dit aussi que T est déterministe sous la distribution

conditionnelle. Alors, $\alpha(x) \sim C^\alpha x^\alpha$ lorsque $x \to +\infty$ et $r = C^\alpha$ est une fonction constante et par conséquent à variation lente, ce qui nous permet d'appliquer les deux lemmes précédents. Sachant $(P_i^\downarrow)$, nous avons avec probabilité 1,

$$\begin{aligned} D_N^\star &= 1 - \sum_{i=1}^{Z_N^\star} \tilde{P}_i \\ &\sim \alpha\, C^\alpha\, \Gamma(1-\alpha)\, N^{\alpha-1} \quad \text{lorsque } x \to +\infty \\ &\sim \alpha\, C^\alpha\, \Gamma(1-\alpha) \left(\frac{Z_N^\star}{\Gamma(1-\alpha)} \right)^{1-1/\alpha} \quad \text{lorsque } x \to +\infty \\ &\sim \alpha\, \Gamma(1-\alpha)^{1/\alpha}\, C\, {Z_N^\star}^{1-1/\alpha} \quad \text{lorsque } x \to +\infty. \end{aligned}$$

Par définition, la suite $Z_N^\star$ décrit tous les entiers positifs. Sachant $(P_i^\downarrow)$, il vient qu'avec probabilité 1,

$$1 - \sum_{i=1}^{n} \tilde{P}_i \sim \alpha\Gamma(1-\alpha)^{1/\alpha}\, C\, n^{1-1/\alpha} \quad \text{lorsque } x \to +\infty. \tag{3.12}$$

Une intégration de (3.12) selon la distribution $\mathrm{PD}(\alpha, 0)$ de (Π_∞) aboutit finalement à 3.10.

Chapitre 4

Modélisation de la production pétrolière et position du problème statistique

Dans ce chapitre, nous présentons l'ensemble des éléments nécessaires pour modéliser de façon complète la production de pétrole dans un bassin exploité. Chacun de ces éléments est l'objet d'une modélisation propre, qui s'appuie notamment sur les conclusions des chapitres 2 et 3, avec les modifications qui s'imposent pour disposer de procédures d'estimation effectives pour rendre le modèle de production utilisable en pratique. Nous mettons ainsi en évidence les principaux problèmes statistiques associés aux modélisations retenues.

4.1 Introduction

En dehors du modèle de la courbe de Hubbert que nous avons discuté dans le premier chapitre, les modèles existants pour la production pétrolière sont exclusivement de nature économétriques. Moroney et Berg (1999), et Kemp et Kasim (2003) étudient ainsi la production à l'aide de séries temporelles. Citons aussi les travaux de Cleveland et Kaufmann (1991) et de Kaufmann (1991) qui réconcilient le point de vue économétrique et les ajustements proposés par Hubbert. Toutes ces méthodes ont pour objectif de mettre en évidence des relations entre la production et des variables explicatives. Cependant, aucun de ces modèles ne tient compte de la distribution probabiliste de la taille des gisements du bassin. Pourtant, de l'exploration à la production des gisements individuels, tous les phénomènes intervenant dans la production du pétrole sont fortement influencés par les tailles des gisements considérés. Nous proposons dans ce chapitre une modélisation complète de la production pétrolière d'un bassin qui prend en compte, à la différence des travaux précédents, la distribution probabiliste des tailles des gisements du bassin.

Il nous parait préférable de ne pas intégrer directement dans le modèle des facteurs de nature économique, tout en laissant la possibilité de les incorporer à l'arrivée, en considérant de multiples scénarios faisant varier la visibilité des champs, l'intensité de l'exploration pétrolière, ou encore l'intensité du processus de mise en production des gisements disponibles. Ce point

de vue nous parait plus raisonnable car même à l'échelle de quelques mois, il est impossible d'affirmer quel sera l'état de l'économie, alors que les propriétés géologiques du sous-sol sont au contraires immuables pendant la durée de l'exploitation.

La production d'un bassin est l'agrégation des productions de gisements individuels. Cette simple remarque soulève naturellement les questions suivantes,

- $\mathbf{Q_a}$: Comment sont réparties les réserves à l'intérieur du bassin ?
- $\mathbf{Q_b}$: Comment sont découverts les gisements du bassin au cours du temps ?
- $\mathbf{Q_c}$: Comment sont mis en production les gisements découverts au cours du temps ?
- $\mathbf{Q_d}$: Comment produit un champ les hydrocarbures qu'il contient au cours du temps ?

Dans les sections qui suivent, nous reconsidérons successivement chacune de ces questions en proposant à chaque fois une modélisation des phénomènes étudiés, et nous mettons en évidence les principaux problèmes statistiques associés aux modélisations retenues.

4.2 Réserves pétrolières

L'ensemble des réserves pétrolières produites par un gisement est appelé *ressources ultimes* du gisement. Dans la suite, nous emploierons aussi par abus le terme "taille" pour désigner les ressources ultimes d'un champ. La modélisation proposée au chapitre 2 justifie l'utilisation de la loi de Lévy-Paréto pour représenter les tailles des gisements du bassin. Il serait plus délicat d'élaborer des procédures d'estimation en utilisant les subordinateurs stables et il est donc préférable à ce stade de revenir au point de vue "échantillon". En mer du Nord, le seuil inférieur de la distribution de Lévy-Paréto des réserves peut être fixé à 1Mb. Dans la suite, nous adoptons ce seuil, et pour la variable aléatoire X de la taille d'un gisement du bassin,

$$X \sim \mathcal{P}\mathrm{ar}(\alpha),$$

ce qui revient à négliger les gisements de tailles inférieurs à 1Mb. Si le seuil est en réalité fixé à ε, il est possible de se ramener à la distribution standard $\mathcal{P}\mathrm{ar}(\alpha)$ quitte à diviser les tailles des gisements par le seuil minimal ε car les distributions de Lévy-Paréto sont invariantes par changement d'échelle (voir la section 2.1.2).

Concernant l'estimation du paramètre α de la distribution de Lévy-Paréto, nous utilisons les travaux de Lepez (2002) qui traitent complètement de cette question et aucun travail statistique supplémentaire ne sera donc nécessaire. Puisque nos résultats s'appuient sur cette méthode, nous présentons maintenant une courte synthèse des principes fondamentaux de ces travaux. Selon Lepez, la population des champs découverts à la date $t^\star$ du présent est le résultat d'un tirage sans remise et biaisé par la taille des gisements dans la population complète de tous les champs connus ou non dans le bassin. Plus précisément, un champ de réserves X est découvert ou non en $t^\star$ suivant la valeur d'une variable de censure ε de loi de Bernoulli de paramètre $\omega(X)$, conditionnellement à X. Le gisement i est découvert à la date

Zone	$\hat{\alpha}$
Graben Mer du Nord	0.75
Bassin de Sirte	0.92
Delta du Niger	0.90

TAB. 4.1: Paramètres de lois de Lévy-Paréto des réserves pétrolières de trois régions pétrolifères.

$t^\star$ si et seulement si $\varepsilon_i = 1$. De plus,

$$\begin{cases} \mathbb{P}(\varepsilon = 1|X) = \omega(X) \\ \mathbb{P}(\varepsilon = 0|X) = 1 - \omega(X) \end{cases} . \tag{4.1}$$

où ω est la probabilité d'inclusion du gisement dans la population des champs découverts à la date $t^\star$. Cette modélisation peut être qualifiée de "statique" puisqu'elle n'apporte d'information sur les découvertes dans le bassin que vis à vis du temps présent. Si la loi des tailles des gisements enfouis dans le bassin est une distribution de Lévy-Pareto, alors l'échantillon des log-tailles est une distribution exponentielle de paramètre α et l'échantillon $(Y_1, \ldots, Y_n)$ des log-tailles des gisements découverts à la date $t^\star$ est i.i.d. de densité

$$q_m(y) = \exp\left(\sum_{I \in m} (\theta_I + \log\alpha - \alpha y)\mathbb{1}_{I(y)}\right) . \tag{4.2}$$

où m est une partition qui définit des classes de tailles de gisements, et θ_I s'exprime en fonction du biais de tirage sur la classe I. Les équations de vraisemblances qui découlent de (4.2) peuvent être résolues de façon explicite en $(\theta_I)_{I \in m}$ et en α.

Pour choisir une fonction ω convenable, l'auteur considère une collection de modèles correspondant à une collection $\mathcal{M}$ de partitions. Pour sélectionner un modèle, c'est à dire une partition m, Lepez s'inspire des travaux de Castellan (2003, 1999) sur la sélection de modèles exponentiels. Nous disposons ainsi d'une procédure pour estimer au mieux[1] le paramètre α et la fonction de biais ω. Dans la suite, il sera donc possible de supposer que l'indice α de la distribution de Lévy-Pareto est connu.

Le tableau 4.1 donne les estimations fournies par le logiciel `Select` qui a été développé par Lepez, pour trois zones pétrolifères présentées dans l'annexe A. Les données de réserves correspondants à ces trois bassins proviennent de la base IHS 2002. Les résultats obtenus montrent que l'habitat en Mer du Nord est plus concentré que ceux des deux autres régions étudiées.

4.3 Exploration pétrolière

Dans cette section, nous proposons une nouvelle modélisation décrivant l'historique complet de l'exploration d'un bassin pétrolier et nous en déduisons un modèle statistique pour les

[1] C'est à dire en minimisant le risque d'estimation.

dates de découvertes ayant eu lieu avant aujourd'hui. À la différence du modèle d'exploration déjà présenté dans le chapitre précédent, ce qui suit a vocation à être appliqué à des bassins pétroliers réels, dans le but de proposer des prolongements des processus d'exploration[2] observés.

4.3.1 Présentation

Du modèle d'exploration pétrolière que nous avons présenté dans le chapitre précédant, nous retenons que la distribution du temps de découverte D d'un gisement peut être définie conditionnellement à la taille X du gisement concerné. Plus précisément, nous supposons que les dates de découvertes des champs sont indépendantes entre elles et de même loi conditionnelle

$$(D \,|\, X) \sim \mathcal{E}\left(h(X)\right), \tag{4.3}$$

où h est une fonction croissante de $\mathbb{R}^+$ dans $\mathbb{R}^+$ appelée fonction de visibilité et qu'il nous faut déterminer.

Il est possible de vérifier que cette modélisation est cohérente avec celle proposée par Lepez. En reprenant les notations de la section précédente, un champ est observé si sa variable de censure ε vérifie $\varepsilon = 1$, ce qui correspond à $D \leq t^\star$. Nous avons alors $\varepsilon = \mathbb{1}_{D \leq t^\star}$ et $\omega(x) = 1 - \exp\left(-h(x)t^\star\right)$.

En réalité, les petits gisements sont à la fois peu visibles et très nombreux, ce qui permet de proposer une modélisation simplifiée de la dynamique des découvertes successives pour cette catégorie particulière de champs. Cette simplification est fondée sur une approximation Poissonienne de la modélisation (4.3). Nous aboutissons ainsi à la modélisation *stratifiée* suivante :

- Les gisements les plus petits sont découverts selon un processus de Poisson homogène d'intensité μ_0. Soit $I_0 = [1, x_0]$ la classe de taille correspondante.
- Pour un gisement de taille $X \geq x_0$, le temps de découverte D de celui-ci a pour distribution conditionnelle

  $$(D \,|\, X) \sim \mathcal{E}\left(h(X)\right),$$

  où h est une fonction affine par morceaux et croissante.

Nous détaillons chacune de ces deux situations dans les sections suivantes. Des résultats de validation de cette modélisation sont disponibles dans l'annexe D.2 pour plusieurs bassins pétroliers réels.

4.3.2 Caractère Poissonien des petites découvertes

La classe I_0 correspond à l'ensemble des petits champs de tailles dans l'intervalle $[1, x_0]$. Même si ces gisements contiennent effectivement des hydrocarbures, pour l'exploration pétrolière leur découverte est plutôt considérée comme un échec. Ces "mauvaises" découvertes surviennent avec un taux r quasiment constant au cours du temps. Supposons que la suite des forages d'exploration effectués dans le bassin puisse être modélisée par un processus de

[2] Le terme "processus" est à prendre ici dans son sens le plus courant.

Poisson homogène d'intensité λ. Alors, les dates de découvertes de gisements dans la catégorie I_0 forment aussi un processus de Poisson homogène d'intensité $\mu_0 = r\lambda$. Selon cette représentation simple, il est naturel d'utiliser un processus de Poisson homogène pour modéliser les découvertes de cette catégorie de gisements. Notons que ce raisonnement serait mis en défaut s'il y avait un risque d'épuisement des réserves de cette classe de champs. En mer du Nord, nous savons qu'il y a de l'ordre d'un millier de gisements de tailles inférieures à 20Mb. Puisque l'on en découvre environ 6 par an, il n'y a donc aucun risque d'épuisement à l'échelle de quelques dizaines d'année.

Afin de justifier cette modélisation, nous allons montrer que le processus de Poisson homogène apparaît comme une situation "limite" de la dynamique définie en (4.3). Tout d'abord, puisque les tailles des gisements de cette catégorie sont très proches du seuil minimal de 1Mb, la visibilité de ceux-ci est très faible. À l'intérieur de cette classe, il est raisonnable de supposer que tous les champs possèdent quasiment la même visibilité h_0. Ensuite, les champs de cette catégorie sont en très grand nombre dans le bassin, soit N_0 le nombre total de champs de cette catégorie dans le bassin. D'après la modélisation (4.3), le nombre $n_0(t)$ des champs de cette classe découverts avant une date t, suit une loi binomiale de paramètres n_0 et $\omega(t) = 1 - \exp(-h_0 t) \approx h_0 t$. Si $h_0 t$ est "petit", alors $\omega(t)$ est petit lui aussi, et par une approximation classique[3], il est alors possible de supposer que $n_0(t)$ suit une loi de Poisson de paramètre $N_0 h_0 t$. La proposition suivante permet de donner une approximation de la loi des temps de découvertes des petits champs.

Proposition 4.3.1. *Soit h_l une suite de réels positifs qui tend vers 0 lorsque l tend vers l'infini. Pour tout l, soit une variable aléatoire Z_l telle que $Z_l \sim \mathcal{E}(h_l)$. Alors, pour tout $t > 0$,*

$$(Z_l | Z_l \leq t) \xrightarrow[l \to +\infty]{\mathcal{L}} U$$

où $U \sim \mathcal{U}([0\, t])$.

Démonstration. Soit $F_{t,l}$ la fonction de répartition de $(Z_l | Z_l \leq t)$. On a alors $F_{t,l}(u) = \frac{1-\exp(-h_l u)}{1-\exp(-h_l t)} \mathbb{1}_{0 \leq u \leq t}$. D'où, lorsque $l \to +\infty$,

$$F_{t,l} \to F_U(u) := \frac{u}{t} \mathbb{1}_{0 \leq u \leq t},$$

où F_U est la fonction de répartition de la variable aléatoire U. □

Sachant $n_0(t)$, les dates de découvertes des champs de la classe I_0 mis à jours avant la date t peuvent ainsi être modélisées par un échantillon i.i.d. de loi uniforme sur $[0, t]$. Tant que la probabilité d'inclusion reste faible, $n_0(t)$ suit une loi de Poisson de paramètre $N_0 h_0 t$. Il est alors justifié de considérer la suite des temps de découvertes des petits champs comme la réalisation d'un processus de Poisson homogène d'intensité μ_0.

Soit $(D_i^{(p)})_{i \geq 1}$ la suite des découvertes, à chacune de ces dates nous associons la taille du gisement découvert, que l'on note $X_i^{(p)}$. Les variables aléatoires $X_i^{(p)}$ sont i.i.d. de loi $\mathcal{P}\mathrm{ar}(\alpha, 1, x_0)$ la distribution de Lévy Pareto restreinte à $[1, x_0]$. Finalement, nous modélisons

[3] Il faut pour cela que $N_0 \omega(t) \approx 1$, on vérifie par exemple en mer du Nord que $N_0 \approx 10^3$ et $h_0 \approx 10^{-3}$

le processus des découvertes de petits champs par un processus de Poisson marqué sur $\mathbb{R}^+$, dont la loi des marques est la distribution $\mathcal{P}\mathrm{ar}(\alpha, 1, x_0)$. La suite des découvertes ayant eu lieu avant la date du présent $t^\star$ correspond donc à la restriction de ce processus sur $[0, t^\star]$.

4.3.3 Dynamique de découverte des gisements de tailles supérieures

Nous nous intéressons maintenant aux découvertes de champs de tailles supérieures à x_0. Soit N le nombre total de gisements du bassin de tailles supérieurs à x_0. La modélisation repose sur les deux hypothèses suivantes.

- Les tailles des gisements de cette catégorie sont représentées par un échantillon i.i.d. $(X_1, \ldots, X_N)$ de distribution commune $\mathcal{P}\mathrm{ar}(\alpha, x_0, +\infty)$.
- Les dates de découvertes D_i sont indépendantes, et pour tout $i = 1, \ldots, N$,

$$(D_i \mid X_i) \sim \mathcal{E}\left(h(X)\right). \tag{4.4}$$

Cependant, cette distribution ne correspond pas à l'échantillon des champs observés, c'est-à-dire découverts avant aujourd'hui.

Échantillon des découvertes observées avant $t^\star$

Les bassins dont nous souhaitons prolonger le processus d'exploration sont des bassins matures dont les gisements les plus importants ont été trouvés depuis longtemps. Dans ce cas, les gisements de cette catégorie sont tous de tailles dans l'intervalle $[x_0, x_{\max}]$, où $x_{\max}$ est la taille du plus grand gisement du bassin. Par conséquent, pour un champ de taille X de cette catégorie, nous avons

$$X \sim \mathcal{P}\mathrm{ar}(\alpha, x_0, x_{\max}). \tag{4.5}$$

D'après ce qui précède, la densité conditionnelle d'un couple (X, D) sachant $X \in [x_0, x_{\max}]$ est donnée par

$$g : (x, t) \longmapsto \alpha h(x) \exp\{-h(x)t\} \frac{x^{-\alpha-1}}{x_0^{-\alpha} - x_k^{-\alpha}} \mathbb{1}_{t \geq 0,\, x_0 \leq x \leq x_k} \tag{4.6}$$

Rappelons que $t^\star$ désigne la date du présent. À la date $t^\star$, un observateur ne voit en réalité que les champs tels que $D \leq t^\star$. La distribution d'un champ de taille dans $[x_0, x_{\max}]$ et découvert avant $t^\star$ correspond à la distribution de densité (4.6), tronquée au delà de $t^\star$. Soit le couple $(X^\star, D^\star)$ associé à l'un de ces gisements observés ; celui-ci a pour distribution

$$(X^\star, D^\star) \stackrel{\mathcal{L}}{=} \left((X, D) \mid D \leq t^\star\right).$$

La densité $g^\star$ de $(X^\star, D^\star)$ est définie pour tout $t \geq 0$ et tout $x \in \mathbb{R}^+$ par

$$g^\star(x, t) = \frac{g(x, t)}{P_{\mathrm{dec}}(\alpha, t^\star, h)} \mathbb{1}_{t \leq t^\star}, \tag{4.7}$$

où P_{dec} représente la probabilité qu'un champ de taille dans $[x_0, x_{\max}]$ soit découvert avant $t^\star$. Cette quantité vaut

$$\begin{aligned}
\mathrm{P}_{\text{dec}}(\alpha, t^\star, h) &= P\left(D \leq t^\star | X \in [x_0, x_{\max}]\right) \\
&= \frac{\alpha}{x_0^{-\alpha} - x_{\max}^{-\alpha}} \int_{x_0}^{x_{\max}} \int_0^{t^\star} h(x) \exp\left\{-h(x)t\right\} x^{-\alpha-1} dx dt \\
&= 1 - \frac{\alpha}{x_0^{-\alpha} - x_{\max}^{-\alpha}} \int_{x_0}^{x_{\max}} \exp\left\{-h(x)t^\star\right\} x^{-\alpha-1} dx \ .
\end{aligned}$$

Soit n le nombre (aléatoire) de gisements découverts à la date $t^\star$. Conditionnellement à n, l'échantillon $((X_1^\star, D_1^\star), \ldots, (X_n^\star, D_n^\star))_{i=1,\ldots,n}$ est i.i.d et chaque couple $(X_n^\star, D_n^\star)$ a pour loi jointe la densité donnée en (4.7). Dans toute la suite, afin de pouvoir construire des estimateurs de h, toutes les probabilités seront supposées conditionnelles à n, qui sera donc considéré comme une quantité déterministe. Il nous faut maintenant préciser la forme retenue pour la fonction de visibilité h.

Choix d'une forme de fonction de visibilité

Nous avons noté au chapitre précédent que le modèle idéalisé qui s'appuie sur une fonction de visibilité proportionnelle à la taille (ou même puissance de la taille) n'est pas suffisamment réaliste. Dans le modèle de Lepez, la fonction ω est une fonction constante par morceaux, la fonction de visibilité associée à ω est donc cette fois elle aussi constante par morceaux. Pour le sujet qui nous préoccupe ici, à savoir la modélisation de l'exploration, l'utilisation de fonctions de visibilité constantes par morceaux se confronte au problème suivant. D'une part, nous savons que les réserves seuillées des gisements d'un bassin suivent la loi de Lévy-Pareto. D'autre part, chacune des classes de la partition utilisée doit contenir un nombre suffisant de champs pour des raisons évidentes de qualité d'estimation. La distribution de Lévy-Pareto conduit donc à définir pour les plus gros champs une classe qui est de grande amplitude. À l'intérieur de celle-ci, on observe encore un effet de la taille sur les temps de découverte, les cinq plus gros (pour fixer les idées) sont en réalité découverts plus rapidement que les autres champs de cette classe. Choisir une fonction de visibilité constante par morceaux entraînerait alors une sous-estimation systématique des dates de découvertes de ces quelques champs. Or, au début de l'exploitation du bassin, la production totale est pour l'essentiel dirigée par cette petite population de champs *géants*. Ainsi modélisée, le début de la production serait alors lui aussi sous-estimé du fait du retard systématique des dates de découvertes des plus gros champs. Pour cette raison, dans la suite, nous ne pouvons utiliser de fonctions de visibilités correspondants rigoureusement avec la modélisation de Lepez[4].

Pour tenir compte des remarques précédentes, nous proposons d'utiliser une fonction de

[4]Notons que les critiques précédentes sur l'utilisation de fonctions constantes par morceaux ne sont pas pertinentes dans le contexte de Lepez. En effet, la procédure proposée par l'auteur vise avant tout à estimer le potentiel de réserves du bassin, et ne prétend pas décrire la dynamique du processus d'exploration. Le faible nombre de classes sélectionnées par cette méthode dans le cas de la mer du Nord (voir Lepez, 2002, p 191) montre bien que la modélisation avec une fonction ω constante par morceaux est suffisante pour estimer correctement les réserves et le paramètre α de la loi de Lévy-Pareto. Nous verrons qu'en pratique il est même préférable d'estimer préalablement le paramètre α par ce modèle pour simplifier les procédures d'estimation de la fonction de visibilité.

visibilité h affine par morceaux et croissante. Cette modélisation permet de conserver le point de vue non paramétrique de Lepez, tout en autorisant la visibilité à croître à l'intérieur des classes de la partition sur laquelle h est définie.

Soit $m = (I_1, \ldots, I_k)$ une partition de taille k de $[x_0, x_{\max}]$. La partition est composée des intervalles suivants,

$$I_1 =]x_0, x_1], \ldots, I_j =]x_{j-1}, x_j], \ldots, I_k =]x_{k-1}, x_{\max}],$$

avec $1 < x_0 < x_1 < \ldots x_j < \ldots < x_k = x_{\max}$. Associé à la partition m, soit l'ensemble $\mathcal{H}_m$ composé de fonctions continues et affines par morceaux,

$$\mathcal{H}_m = \left\{ h : x \in [x_0, x_{\max}] \longmapsto \sum_{j=1}^{k} \{a_j(x - x_{j-1}) + b_j\} \mathbb{1}_{x \in I_j},\ (A, B) \in (\mathbb{R}^+)^{2k} \right\} \cap \mathcal{C}^+([x_0, x_{\max}])$$

où $\mathcal{C}^+([x_0, x_{\max}])$ désigne l'ensemble des fonctions continues et strictement positives sur $[x_0, x_{\max}]$. Les vecteurs A et B désignent respectivement $(a_1, \ldots, a_k)$ et $(b_1, \ldots, b_k)$. Pour toute fonction $h \in \mathcal{H}_m$, la condition de continuité impose que,

$$\forall j \geq 2,\ b_j = \sum_{u=1}^{j-1} a_u(x_u - x_{u-1}) + b_1. \tag{4.8}$$

De plus, la condition de positivité équivaut à $b_1 > 0$. L'ensemble $\mathcal{H}_m$ est donc entièrement paramétré par le vecteur des pentes $A \in (\mathbb{R}^+)^k$ et le seul coefficient $b := b_1 > 0$. Notons qu'une fonction $h \in \mathcal{H}_m$ n'admet clairement qu'une seule écriture de la forme $\sum_{j=1}^{k} \{a_j(x - x_{j-1}) + b_j\} \mathbb{1}_{x \in I_j}$ avec les b_j vérifiant (4.8), ce qui permet d'assimiler la fonction h aux coefficients A et b. Dans la suite, il nous sera parfois utile de noter $h(A, b)$ ou encore $h(A, B)$ la fonction de visibilité h définie par (A, b) ou (A, B).

Écriture des modèles

Soit S_m l'ensemble des densités $g^\star$ définies comme en (4.7) à partir d'une partition m de $[x_0, x_{\max}]$,

$$S_m = \left\{ g^\star : (x, t) \mapsto \alpha \frac{h(x) \exp\{-h(x)t\}}{\mathrm{P}_{\mathrm{dec}}(\alpha, t^\star, h)} \frac{x^{-\alpha-1}}{x_0^{-\alpha} - x_k^{-\alpha}} \mathbb{1}_{0 \leq t \leq t^\star,\, x_1 \leq x \leq x_k},\ \alpha > 0,\ h \in \mathcal{H}_m \right\}.$$

Nous reviendrons au chapitre 6 sur la procédure d'estimation de $g^\star$ dans le modèle S_m par maximum de vraisemblance, mais notons dès maintenant que le paramètre α de la loi de Lévy-Paréto peut être supposé connu dans la définition des modèles ci-dessus. Nous pourrions construire un estimateur du maximum de vraisemblance qui s'appuie à la fois sur les $X_i^{(p)}$ et sur les $X_i^\star$. Une alternative plus simple consiste à estimer le paramètre α en utilisant la méthode de Lepez, que nous avons exposée dans la section précédente. En effet, nous avons rappelé que cette méthode s'appuie sur une résolution explicite d'un système d'équation de vraisemblance ce qui rend cette estimation plus fiable. Nous préférons donc estimer α préliminairement de cette façon ; ce paramètre est donc supposé connu à ce stade de l'étude. Nous considérons

donc les ensembles

$$S_m(\alpha) = \left\{ g^\star : (x,t) \mapsto \alpha \frac{h(x)\exp\{-h(x)t\}}{\mathrm{P}_{\mathrm{dec}}(\alpha, t^\star, h)} \frac{x^{-\alpha-1}}{x_0^{-\alpha} - x_k^{-\alpha}} \mathbb{1}_{0\le t\le t^\star,\, x_1\le x\le x_k},\ h \in \mathcal{H}_m \right\},$$

que nous noterons encore S_m dans la suite par abus de notation.

Proposition 4.3.2. *Le modèle S_m est identifiable vis à vis de la paramétrisation en h et α.*

Démonstration. Montrons que la paramétrisation du modèle est injective. Supposons fixés $t^\star$ ainsi qu'une partition m de $[x_0, x_{\max}]$ de taille k . Soient α_1 et α_2 deux réels positifs, et soient h_1 et h_2 deux fonctions de visibilité de l'ensemble $\mathcal{H}_m$ telles que pour tout $x \in [x_0, x_{\max}]$ et tout $t \in [0, t^\star]$,

$$\alpha_1 \frac{h_1(x)\exp\{-h_1(x)t\}}{\mathrm{P}_{\mathrm{dec}}(\alpha_1, t^\star, h_1)} \frac{x^{-\alpha_1-1}}{x_0^{-\alpha_1} - x_{\max}^{-\alpha_1}} = \alpha_2 \frac{h_2(x)\exp\{-h_2(x)t\}}{\mathrm{P}_{\mathrm{dec}}(\alpha_2, t^\star, h_2)} \frac{x^{-\alpha_2-1}}{x_0^{-\alpha_2} - x_{\max}^{-\alpha_2}}. \tag{4.9}$$

Pour $t = 0$, et pour $x \in [x_0, x_1]$ ceci nous donne

$$\alpha_1 \frac{a_j^{(1)} x + b_j^{(1)}}{\mathrm{P}_{\mathrm{dec}}(\alpha_1, t^\star, h_1)} \frac{x^{-\alpha_1-1}}{x_0^{-\alpha_1} - x_{\max}^{-\alpha_1}} = \alpha_2 \frac{a_j^{(2)} x + b_j^{(2)}}{\mathrm{P}_{\mathrm{dec}}(\alpha_2, t^\star, h_2)} \frac{x^{-\alpha_2-1}}{x_0^{-\alpha_2} - x_{\max}^{-\alpha_2}}.$$

L'identification des fonctions puissances implique que $\alpha_1 = \alpha_2 = \alpha$. Pour $x \in [x_0, x_{\max}]$ et tout $t \in [0, t^\star]$, l'identité (4.9) peut alors se simplifier sous forme suivante

$$\frac{h_1(x)\exp\{-h_1(x)t\}}{\mathrm{P}_{\mathrm{dec}}(\alpha, t^\star, h_1)} = \alpha \frac{h_2(x)\exp\{-h_2(x)t\}}{\mathrm{P}_{\mathrm{dec}}(\alpha_2, t^\star, h_2)}.$$

Fixons maintenant cette identité en un point x arbitraire. Les deux termes exponentiels, comme fonctions de t ont nécessairement le même exposant, ce qui donne $h_1(x) = h_2(x)$. Ceci est vrai pour tout $x \in [x_0, x_{\max}]$, nous avons donc $h_1 = h_2$. Nous avons vu plus haut que $h \in \mathcal{H}_m$ n'admet qu'une seule écriture, ce qui permet de conclure. □

Forme des partitions considérées

Nous expliquons maintenant pourquoi le seuil x_0 ne sera pas intégré dans la procédure de sélection d'une partition. Quitte à considérer la loi jointe de l'échantillon complet

$$\left(\left(X_1^{(p)}, D_1^{(p)} \right), \ldots, \left(X_{n_0}^{(p)}, D_{n_0}^{(p)} \right), (X_1^\star, D_1^\star), \ldots, (X_n^\star, D_n^\star) \right),$$

il serait pourtant tout à fait possible de proposer une famille de partitions $\tilde{m}$ de $[1, x_{\max}]$ permettant de faire varier x_0. Nous pourrions ainsi sélectionner le seuil optimal au sens d'un critère statistique. Cependant, il ne faut pas perdre de vue que l'un de nos objectifs principaux est de proposer des prolongements du processus d'exploration. Or, nous avons vu que l'approximation poissonnienne, présentée dans le section précédente, n'est valide que dans la mesure où la classe des petits champs ne risque pas d'être épuisée. Il y a donc un danger que la procédure de sélection de modèle choisisse un seuil x_0 tel que l'approximation poissonnienne ne soit plus valide dans le futur. C'est pourquoi il est préférable que le seuil x_0 soit choisi une

fois pour toutes par l'utilisateur[5].

Notre problème est donc de choisir la meilleure partition m possible de l'ensemble $[1, x_{\max}]$ au sens d'un critère statistique, afin d'estimer "au mieux" la dynamique de l'exploration pétrolière sur celle-ci. Cet objectif relève d'une problématique de sélection de modèles en estimation de densité. Le traitement statistique de ce problème est réalisé dans le chapitre suivant.

4.4 Production individuelle des gisements pétroliers

Dans cette section, nous nous intéressons aux production individuelle des gisements pétroliers, et plus particulièrement à la forme des courbes de production pétrolière.

4.4.1 Présentation du problème

On appelle *profil de production* la courbe de la production d'un champ en fonction du temps, et on appelle *profil de production normalisé* la courbe de la production divisée par les réserves totales du champ. Il est bien connu de l'industrie pétrolière que les profils de production normalisés des gisements importants n'ont pas la même allure que ceux des petits champs. Les champs de petites réserves ont tendance à produire leurs réserves en peu de temps, leur pic de production est atteint très tôt et la production décline ensuite rapidement. En revanche, les gros champs produisent plus lentement leurs réserves, un plateau peut même être observé au maximum de la production. La figure 4.1 confirme cette description pour une comparaison des productions de trois champs en mer du Nord.

Ce principe est tout à fait naturel pour plusieurs raisons. D'abord, l'extraction est effectuée par des puits de forage dont le diamètre et le débit sont nécessairement limités. Il est par conséquent difficile d'extraire rapidement une importante proportion des réserves d'un grand champ du fait de ces limitations techniques. Ensuite, l'écoulement des hydrocarbures à l'intérieur du gisement peut être plus au moins difficile, selon la viscosité des hydrocarbures et la porosité de la roche. Il faut alors forer à différents endroits du gisement, et ceci ne peut être fait instantanément. Enfin, on peut noter que seuls les champs suffisamment gros font l'objet de procédures de récupérations secondaires ou tertiaires permettant d'augmenter le taux de récupération, ce qui a pour effet de prolonger dans le temps la production du champ.

Nous proposons de valider ce principe en basant notre analyse sur une classification (non supervisée) d'un échantillon représentatif de courbes de production.

4.4.2 Classification de courbes non supervisée par modèle de mélange gaussien

Le cadre de la classification de courbes est généralement le suivant. Nous disposons d'un échantillon de n courbes dont nous souhaitons obtenir une classification, et chaque courbe est représentée par une suite de valeurs ordonnées dans le temps. Pour simplifier supposons que toutes les courbes sont composées d'un même nombre de points $\widetilde{Q}$. Pour résoudre le problème de classification, de nombreux auteurs proposent de se ramener à la classification

[5]Typiquement, x_0 est de l'ordre de 10 à 20 Mb pour les trois bassins étudiés.

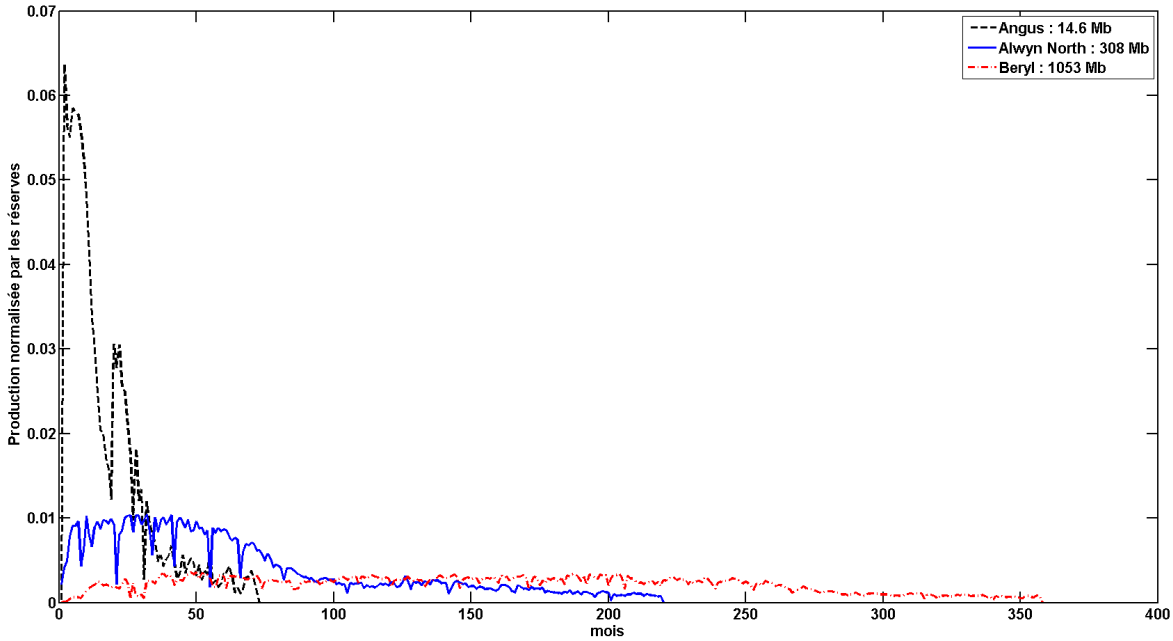


FIG. 4.1: Productions normalisées par les réserves des champs Angus (14.6 Mb), Alwyn North (308 Mb), et Beryl (1053 Mb) en mer du Nord.

non supervisée "classique" en projetant les courbes sur une base fonctionnelles de splines ou d'ondelettes. C'est par exemple le cas dans les travaux de Abraham *et al.* (2003), García-Escudero et Gordaliza (2005), Ma *et al.* (2006) et James et Sugar (2003). Dans tous ces articles, une transformation B-splines des courbes est effectuée et différentes méthodes de classification sont ensuite utilisées sur ces données transformées.

Dans l'ensemble des travaux que nous venons de citer, le problème du choix du nombre de composantes est soit omis, soit traité en utilisant des critères de choix qui ne s'appuient pas sur des résultats théoriques (sauf dans le cas de James et Sugar (2003) où les auteurs s'appuient sur une "fonction de distortion" pour choisir k). Les modèles de mélanges gaussiens offrent un cadre statistique adéquat pour choisir le nombre de composantes selon un critère statistique imposé. Nous rappelons maintenant les principes fondamentaux de cette méthode couramment utilisée en classification non supervisée.

Classification non supervisée par modèles de mélange gaussiens

Soit $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, un échantillon i.i.d de même distribution de probabilité inconnue f sur $\mathbb{R}^v$, et nous proposons d'estimer cette quantité par un modèle de mélanges gaussien sur $\mathbb{R}^v$. Chaque composante du mélange est modélisée par une loi normale, et représente une sous population qui est ainsi identifiée. La densité d'un mélange Gaussian à K composantes s'écrit donc sous la forme

$$t = \sum_{k=1}^{K} p_k \Phi(\cdot|\eta_k, \Sigma_k)$$

où les p_k sont les proportions du mélanges : pour tout k, $0 < p_k < 1$ avec $\sum_{k=1}^{K} p_k = 1$, et $\Phi(\cdot|\eta_k, \Lambda_k)$ désigne une densité gaussienne v-dimensionnelle de moyenne η_k et de matrice de

covariance Σ_k. La densité t est donc entièrement paramétrée par le vecteur des paramètres $(p_1, \ldots, p_K, \eta_1, \ldots, \eta_K, \Lambda_1, \ldots, \Sigma_K)$.

Ce modèle de mélange peut être vu comme une structure à données manquantes dont les données complètes seraient $((\mathbf{y}_1, \mathbf{z}_1), \ldots, (\mathbf{y}_n, \mathbf{z}_n))$ où $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ avec $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$ et $z_{ik} = 1$ si $\mathbf{y}_i$ appartient au k-ième groupe. Le vecteur des labels $\mathbf{z}$ définit la classification dont on souhaiterait idéalement disposer. Une estimation du vecteur des paramètres peut être obtenue grâce à un algorithme EM (Dempster *et al.*, 1977), ce qui fournit automatiquement une classification des données par la règle du maximum *a posteriori* (MAP) :

$$\hat{z}_{ik} = \begin{cases} 1 & \text{if } \hat{p}_k \Phi(\mathbf{y}_i | \hat{\eta}_k, \hat{\Lambda}_k) > \hat{p}_l \Phi(\mathbf{y}_i | \hat{\eta}_l, \hat{\Lambda}_l), \ \forall l \neq k \\ 0 & \text{sinon.} \end{cases}$$

Les modèles de mélanges gaussiens se différencient entre eux par la forme des matrices de covariance Σ_k intervenant dans chacune des composantes. Considérons les décomposition sur les sous-espaces propres de chacune des matrices Σ_k :

$$\Sigma_k = \lambda_k D_k \Delta_k {}^t D_k,$$

où Δ_k est une matrice de diagonale de déterminant 1, D_k est une matrice orthogonale et λ_k est un réel positif. D'après Banfield et Raftery (1993) et Celeux et Govaert (1995), il est possible de définir 23 formes particulières de modèles de mélange en imposant des contraintes particulières sur les matrices D_k, Δ_k et les coefficients λ_k. En adoptant les même notations que dans Biernacki *et al.* (2006), voici quelques exemples de mélanges gaussiens qui seront utilisés par la suite.

- Mélange $[LB_k]$: les matrices Σ_k sont diagonales et de même déterminant.
- Mélange $[L_k B_k]$: les matrices Σ_k sont diagonales et peuvent dépendre de k.
- Mélange $[L_k C_k]$: aucune contrainte sur les matrices Σ_k qui peuvent de plus être différentes.

4.4.3 Double intérêt de la sélection de variable

Les modèles de mélanges gaussiens permettent de ramener le problème de classification à un problème d'estimation de densité. Or, dans la plupart des situations concernées par la classification de courbes, le nombre de points $\widetilde{Q}$ de chaque courbe est de l'ordre de n, voir même beaucoup plus grand. Pour ce problème que l'on peut qualifier de "grande dimension", la qualité de l'estimation peut donc être altérée si celle-ci est effectuée dans un modèle comportant trop de paramètres. Le problème de la dimension du modèle sur lequel les courbes sont projetées est donc crucial pour aborder le problème de classification. Pour répondre à cette question, nous allons définir une collection de modèles de dimensions variées et un critère statistique pour choisir un modèle permettant de minimiser l'erreur d'estimation.

Soient $x_1, \ldots, x_n$ les données disponibles après projection des courbes dans un espace E de dimension Q. La projection peut correspondre à une transformation de Fourier, une transformée en ondelettes ou encore par B-splines des courbes initiales. Nous restons volon-

tairement abstraits sur cette étape préliminaire de transformation des courbes, car le choix de cette transformation des données dépend du contexte étudié. Nous imposons simplement à Q d'être suffisamment grand pour que la procédure de sélection de variables que nous allons détailler maintenant garde tout son sens. Les vecteurs $x_1, \ldots, x_n$ de $\mathbb{R}^Q$ sont considérés comme des réalisations indépendantes d'une même distribution de probabilité s. Nous supposons de plus que les données $x_1, \ldots, x_n$ sont centrées et réduites.

Soit $\mathbf{v}$ un sous-ensemble de $\{1, \ldots, Q\}$. Pour le vecteur x de $\mathbb{R}^Q$ correspondant à une courbe de l'échantillon, on note $x_{[\mathbf{v}]}$ le sous-vecteur de x composé des variables d'indices dans $\mathbf{v}$, et $x_{[\mathbf{v}^c]}$ le sous-vecteur de x composé des variables restantes. Les variables sont naturellement disposées dans les vecteurs $x_{[\mathbf{v}]}$ et $x_{[\mathbf{v}^c]}$ par ordre croissant de leur indice dans x. Pour une classification reposant sur les variables du bloc $\mathbf{v}$, nous définissons un modèle $S_{(K,\mathbf{v})}$ de la façon suivante,

$$S_{(K,\mathbf{v})} = \left\{ x \in \mathbb{R}^Q \mapsto f(x_{[\mathbf{v}]})\, \Phi(x_{[\mathbf{v}^c]} \,|\, 0, I_{Q-v}) \;;\;\; f \in \mathcal{L}_{(K,v)} \right\}.$$

où $v = |\mathbf{v}|$, et $\mathcal{L}_{(K,v)}$ est une famille de densités de mélange gaussien sur $\mathbb{R}^v$ à K composantes. La loi jointe des variables de classification est modélisée par une distribution de modèle de mélange gaussien alors que les variables restantes forment un vecteur de dimension $Q-v$ et de loi normale centrée réduite. La sélection d'un modèle $S_{(K,v)}$ parmi une collection disponible conduit donc à une classification des données, mais aussi à une sélection d'un bloc de variables classification.

Le recours à la sélection d'un bloc de variables de classification a été motivé plus haut par un argument de minimisation de l'erreur d'estimation de la densité s. Il est important de souligner que la sélection de variables est aussi bénéfique pour la classification. En effet, certaines variables peuvent être inutiles pour effectuer la classification, voir même jouer un rôle néfaste vis à vis de cet objectif. Cet argument est avancé par de nombreux auteurs pour développer des méthodes intégrant classification et sélection de variables. Citons par exemple les travaux de Law *et al.* (2004) où le concept de "feature saliency" est défini pour déterminer un ensemble de variables pertinentes pour la classification. C'est le cas aussi dans Raftery et Dean (2006) où le problème de la sélection de variables et de la classification sont reconsidérés comme un problème de sélection de modèles. Dans cet article, les variables dites "non pertinentes" sont expliquées par les variables pertinentes pour la classification par une régression linéaire. Une version améliorée de cette méthode est aussi proposée dans Maugis *et al.* (2007). Toujours dans le contexte des modèles de mélange gaussien, le problème de la classification et de sélection de variables simultanés est aussi étudié dans Bouveyron *et al.* (2007) qui propose une méthode de réduction de dimension sur une collection de modèles de mélanges. La méthode que nous proposons dans le chapitre suivant pour sélectionner un modèle $S_{(K,\mathbf{v})}$ se distingue des travaux précédents par la définition d'un nouveau critère pénalisé qui repose sur des résultats théoriques démontrés plus loin.

Rappelons que notre objectif est aussi de déterminer quelles variables explicatives sont cohérentes avec la classification de courbes obtenues. Quitte à rajouter à chaque individu son bloc de variables explicatives centrées réduites, une sélection de variables opérée sur l'ensemble

de toutes les variables permet de répondre aussi à ce deuxième objectif. Nous allons maintenant définir de façon plus précise des collections de modèles adaptées à notre problème.

4.4.4 Définition de collections de modèles adéquates

Nous considérons des collections de modèles pour les formes de mélanges gaussiens qui ont été définies à la section 4.4.2. Une collection est dite *ordonnée* si les blocs $\mathbf{v}$ de variables de classification sont des ensembles de la forme $\mathbf{v} = \{1, \ldots, v\}$, où $v = |\mathbf{v}|$. On note $\mathcal{M}$ une telle collection et dans le cas contraire la collection est dite *non ordonnée* et elle est notée $\mathcal{M}'$.

Collections ordonnées

Dans le cas d'une collection ordonnée, on note $S_{(K,v)}$ le modèle de mélange à K composantes et de bloc de variables de classification $\mathbf{v} = \{1, \ldots, v\}$ avec $|\mathbf{v}| = v$,

$$S_{(K,v)} = \left\{x \in \mathbb{R}^Q \mapsto f(x_1, \ldots, x_v)\, \Phi(x_{v+1}, \ldots, x_Q \,|\, 0, I_{Q-v}) \;;\;\; f \in \mathcal{L}_{(K,v)}\right\}.$$

Les ensembles $\mathcal{L}_{(K,v)}$ sont des familles de densité de mélanges gaussiens dont la *forme*, c'est-à-dire le type de matrice variance-covariance autorisé dans les composantes du mélange, dépend de la collection considérée. Nous définissons les ensembles $\mathcal{L}_{(K,v)}$ associés à trois collection de modèles ordonnés.

- Collection $\mathcal{M}[LB_k]$:

$$\mathcal{L}_{(K,v)} = \left\{\sum_{k=1}^{K} p_k \Phi(\cdot|\mu_k, \lambda\Sigma_k) \;;\; \lambda > 0,\, \Sigma_k \in \Delta^1_{(v)},\, 0 < p_k < 1,\, \sum_{k=1}^{K} p_k = 1\right\}$$

  où $\Delta^1_{(v)}$ désigne l'ensemble des matrices diagonales définies positives et de déterminant 1.
- Collection $\mathcal{M}[L_kB_k]$:

$$\mathcal{L}_{(K,v)} = \left\{\sum_{k=1}^{K} p_k \Phi(\cdot|\mu_k, \Sigma_k) \;;\; \Sigma_k = \mathrm{diag}(\sigma^2_{k1}, \ldots, \sigma^2_{kv}),\, 0 < p_k < 1,\, \sum_{k=1}^{K} p_k = 1\right\}.$$

- Collection $\mathcal{M}[L_kC_k]$:

$$\mathcal{L}_{(K,v)} = \left\{\sum_{k=1}^{K} p_k \Phi(\cdot|\mu_k, \Sigma_k) \;;\; \Sigma_k \in \mathcal{D}^+_{(v)},\, 0 < p_k < 1,\, \sum_{k=1}^{K} p_k = 1\right\}$$

  où $\mathcal{D}^+_{(v)}$ désigne l'ensemble des matrices symétriques définies positives de tailles $v \times v$.

Collections non ordonnés

Les modèles des collections non ordonnées sont notés $S_{(K,\mathbf{v})}$ où $\mathbf{v}$ désigne l'ensemble des variables de classification. Pour $v = |\mathbf{v}|$, un modèle $S_{(K,\mathbf{v})}$ s'obtient à partir du modèle $S_{(K,v)}$ de la collection ordonnée qui lui est associée en permutant l'ordre des variables. Il suffit pour

cela de noter que

$$S_{(K,\mathbf{v})} = \{x \in \mathbb{R}^Q \mapsto f \circ \tau(x),\ f \in S_{(K,v)}\},$$

où τ est une permutation telle que $(\tau(x)_1, \ldots, \tau(x)_v)' = x_{[\mathbf{v}]}$. Les collections de modèles non ordonnées sont notées $\mathcal{M}'[L_k B_k]$, $\mathcal{M}'[L_k C_k]$ et $\mathcal{M}'[LB_k]$.

4.5 Politique de mise en production des champs

Les nombreux éléments qui influencent la mise en production d'un champ ont été détaillées dans le chapitre 1. Citons par exemple la quantité de réserves contenues dans le gisement, la fiscalité en vigueur, le prix du brut, l'accessibilité des hydrocarbures, la proximité d'autres installations pétrolières (gisements, pipe-lines). Cependant, la première condition pour qu'un gisement puisse être produit est évidemment que celui-ci ait été déjà découvert. À chaque instant, les pétroliers disposent d'un stock de champs découverts et non encore produits, et nous appelons *politique de mise en production* la façon et le rythme avec lesquels les champs sont choisis dans le stock pour être mis en production.

Afin de définir plus rigoureusement le stock de champs disponibles à une date t, nous adoptons les notations suivantes. Soit $N_{\mathcal{B}}$ le nombre total de gisements contenus dans le bassin, le gisement i contient une quantités X_i de réserves en hydrocarbures, et soit D_i sa date de découverte. Pour $u \geq 1$, soient L_u la suite des dates de mise en production, et Y_u les réserves du gisement lancés à la date L_u. On note enfin $u(i)$ le rang de lancement du gisement i, et si celui-ci n'est jamais mis en production, alors par convention $u(i) = \infty$. À chaque instant t le stock de gisements, noté Stock_t, est défini par

$$\mathrm{Stock}_t = \left\{ i \in \{1, \ldots, N_{\mathcal{B}}\} \,|\, D_i \leq t < L_{u(i)} \right\}.$$

Il est très difficile de reproduire exactement par des simulations la politique de mise en production des compagnies pétrolières, il faudrait pour cela modéliser un grand nombre de facteurs économiques. Cependant, on peut vérifier sur des bassins connus que les dates L_u sont réparties dans le temps comme la réalisation d'un processus de Poisson dont l'intensité varie lentement. De plus, pour rentabiliser rapidement leurs investissements très importants, les compagnies pétrolières ont tout intérêt à produire en priorité les gisements les plus gros. Il parait donc raisonnable de conditionner le choix dans le stock d'un gisement à mettre en production, aux tailles de l'ensemble des champs composant le stock. Nous modélisons donc le choix du gisement à l'instant L_u par un tirage biaisé par la taille dans l'ensemble Stock_{L_u}. Soit ω_{St} la fonction de biais associée à ce tirage.

En pratique, il nous paraît très difficile de parvenir à estimer la fonction ω_{St} avec une bonne précision. En effet, le stock évolue au cours du temps, et il existe de plus des délais de mise en production incompressible qui viennent perturber la modélisation simplifiée que nous venons de présenter. Plutôt que de tenter d'estimer ω_{St}, nous préférons adopter un point de vue "exploratoire" en proposant des scénarios reposant sur des choix vraisemblables d'intensité du processus L, et de fonction de biais ω_{St}. Le chapitre 7 est consacré à cette étude.

4.6 Synthèse des problématiques statistiques relevées

De l'analyse des différents éléments nécessaires pour construire un modèle complet de la production pétrolière à l'échelle d'un bassin, il apparaît que deux problèmes statistiques principaux se posent pour la mise en application des modélisations proposées. Ces deux problèmes sont d'une part le choix d'une partition m pour estimer la distribution des dates de découvertes des gisements du bassin, et d'autre part du choix d'un modèle de mélange gaussiens $S_{(K,\mathbf{v})}$ pour obtenir une classification des profils de production normalisés. Dans les deux cas, il s'agit d'estimer une densité s inconnue à partir d'un échantillon composé de quelques centaines d'individus. Dans ce cadre, la minimisation de l'erreur d'estimation sur l'ensemble des estimateurs du maximum de vraisemblance associés à une collection de modèles apparaît comme un critère naturel pour choisir un modèle dans la collection. Dans le chapitre suivant, nous traitons ces deux problèmes statistiques qui relèvent tous deux de la sélection de modèles pour l'estimation de densité.

Chapitre 5

Sélection de modèle pour l'estimation de densité

Dans la première section de ce chapitre, nous rappelons le cadre statistique de la sélection de modèles pour l'estimation de densité et nous donnons un théorème général établi par Massart qui nous permet de déterminer des critères pénalisés d'abord pour des collections de modèles de mélanges gaussiens, et ensuite pour la collection de modèles définie dans le contexte de l'exploration pétrolière.

5.1 Rappels sur la sélection de modèles pour l'estimation de densité

5.1.1 Estimation de densité par maximum de vraisemblance et sélection de modèles

Soit $X_1, \ldots, X_n$ un échantillon i.i.d, avec $X_i \in \mathbb{R}^d$ de densité de probabilité s inconnue pour la mesure de Lebesgue sur $\mathbb{R}^d$. Soit $\mathcal{S}$ l'ensemble de toutes les densités pour la mesure de Lebesgue sur $\mathbb{R}^d$. La méthode du maximum de vraisemblance, qui consiste à trouver les paramètres d'un modèle qui maximisent la vraisemblance des observations, peut être réinterprétée comme une méthode de minimisation de contraste. Pour cela, nous considérons le contraste $\gamma(t, .) = -\ln\{t(\cdot)\}$. Soit le contraste empirique

$$\gamma_n(t) = -\frac{1}{n}\sum_{i=1}^{n} \ln\{t(X_i)\}$$

associé à γ. Dans ce contexte, la fonction de perte définie par $l(s,t) = \mathbb{E}[\gamma_n(t)] - \mathbb{E}[\gamma_n(s)]$ est exactement l'information de Kullback-Leibler. Celle-ci est définie pour deux densités f et g de $\mathcal{S}$ par

$$\mathrm{KL}(f,g) = \int \ln\left\{\frac{f(x)}{g(x)}\right\} f(x)\,dx$$

si $f\,dx$ est absolument continue par rapport à $g\,dx$, et $+\infty$ sinon. La densité s est l'unique fonction de $\mathcal{S}$ telle que

$$s = \operatorname*{argmin}_{t\in\mathcal{S}} \int \gamma(t,x)s(x)\,dx.$$

Soit S un sous-ensemble de $\mathcal{S}$, l'estimateur du maximum de vraisemblance (EMV) de s sur S est défini par

$$\hat{s} := \operatorname*{argmin}_{t\in S} \gamma_n(t).$$

En remplaçant ainsi γ par γ_n et $\mathcal{S}$ par S, on s'attend à ce que l'estimateur obtenu soit proche de la véritable densité s, au moins dans le cas où s n'est pas "trop loin" du modèle S et pour n suffisamment grand.

Dans le contexte de la sélection de modèles qui est le notre, nous disposons d'une collection de modèles $(S_m)_{m\in\mathcal{M}_n}$ et d'un estimateur du maximum de vraisemblance $\hat{s}_m$ pour chacun d'entre eux. Comme les notations le suggèrent, la collection de modèle est autorisée à dépendre de la taille n de l'échantillon observé. Notons de plus qu'il n'est pas nécessaire de supposer que la densité s appartiennent à l'un des modèles de la collection. Nous souhaitons utiliser l'EMV associé au "meilleur modèle", au sens d'un certain critère statistique. Dans le cadre de l'estimation de densité, un critère naturel est la minimisation de l'*erreur moyenne d'estimation* (ou *risque d'estimation*), que l'on définit pour un estimateur $\hat{s}_m$ par

$$\mathcal{R}(\hat{s}_m) = \mathbb{E}[\mathrm{KL}(s,\hat{s}_m)].$$

Idéalement, nous souhaiterions sélectionner le modèle minimisant cette quantité. Cependant, ceci est impossible en pratique car le risque dépend de la densité s qui est inconnu ; la densité $\tilde{m}$ qui minimise le risque d'estimation pour la collection $(S_m)_{m\in\mathcal{M}_n}$ est appelée *oracle*. Une procédure de sélection de modèle est considérée de bonne qualité si celle-ci permet de sélectionner un modèle dont l'EMV a les mêmes performances que celle de l'oracle. Une *inégalité oracle* permet de mettre en évidence de telles propriétés de façon non asymptotique :

$$\mathbb{E}[\mathrm{KL}(s,\hat{s}_m)] \leq C\mathbb{E}\left[\inf_{m\mathcal{M}_n}\{\mathrm{KL}(s,\hat{s}_m) + R(m,n)\}\right]$$

où C est une constante et $R(m,n)$ est un terme de reste qui ne doit pas être trop grand devant le terme de risque.

5.1.2 Critère pénalisé

Puisque le risque de $\hat{s}_m$ est égal à $\mathbb{E}\left[\int\{\gamma(\hat{s}_m)-\gamma(s)\}s\,dx\right]$, la minimisation de ce risque est équivalente à la minimisation la quantité $\mathbb{E}[\int\gamma(\hat{s}_m)s\,dx]$. Une proposition naturelle pour sélectionner un modèle dans la collection serait de choisir celui pour lequel l'EMV minimise le critère $\gamma_n(\hat{s}_m)$. Cependant, cette méthode conduit à sous-estimer le risque $\mathcal{R}(m)$ et le critère obtenu sélectionnerait systématiquement les grands modèles. Cette sous-estimation (que l'on qualifie d'*erreur de substitution*) dépend en réalité de la complexité des modèles, les procédures

de *pénalisation* consistent alors à considérer des critères de la forme

$$\mathrm{crit}(m) = \gamma_n(\hat{s}_m) + \mathrm{pen}(m)$$

où pen : $\mathcal{M}_n \mapsto \mathbb{R}$ est une fonction qui mesure la complexité des modèles et permet de pénaliser les modèles de trop grande complexité. Les premières procédures de pénalisation pour l'estimation de densité ont été proposés dans les années 70 par Akaike (1973). L'heuristique proposée par Akaike conduit à considérer une pénalité de la forme $\frac{D_m}{n}$ où D_m est la dimension du modèle S_m. Le point de vue d'Akaike peut être qualifié d'*asymptotique* puisque dans l'heuristique qu'il propose, la taille de l'échantillon est sensée tendre vers l'infini indépendamment de la collection de modèles.

Avec les travaux de Ledoux et Talagrand (voir Ledoux et Talagrand, 1991; Talagrand, 1995) sur le phénomène de concentration de la mesure, Birgé et Massart ont pu développer une approche non asymptotique de la pénalisation dont une présentation générale est disponible dans les notes de Saint-Flour de Massart (2007). Dans ce cadre, la taille n de l'échantillon est fixé et la collection de modèles peut être définie en fonction de n. L'objectif de l'approche non asymptotique est de définir des pénalités conduisant à des inégalités de type oracle.

Pour exploiter les résultats de concentration de la mesure dans le contexte de la sélection de modèles, le point de départ est le suivant (voir par exemple Massart, 2007, p.9) : pour tout $m \in \mathcal{M}$ et tout $s_m \in S_m$, nous pouvons écrire d'après les définitions précédentes que

$$\mathrm{KL}(s, \hat{s}_{\hat{m}}) \leq \mathrm{KL}(s, s_m) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + \bar{\gamma}_n(s_m) - \bar{\gamma}_n(\hat{s}_{\hat{m}}),$$

où $\bar{\gamma}_n$ est le processus empirique centré défini par $\bar{\gamma}_n(t) = \gamma_n(t) - \mathbb{E}[\gamma_n(t)]$. Pour obtenir une inégalité oracle, la fonction de pénalité doit être choisie de façon à annihiler les fluctuations de $\bar{\gamma}_n(s_m) - \bar{\gamma}_n(\hat{s}_{\hat{m}})$. Soit $Z = \zeta(X_1, \ldots, X_n)$ avec ζ une fonction mesurable de $\mathbb{R}^n$ dans $\mathbb{R}$. De façon générale une inégalité de concentration est une inégalité de la forme

$$\mathbb{P}[Z - \mathbb{E}(Z) \geq x] \leq \exp\left(-\frac{x^2}{2v}\right), \qquad \text{pour tout } x \geq x_0,$$

où v est un majorant de la variance de Z et x_0 est une fonction explicite de n et v. De même, une majoration de la même forme pour les déviations de Z à gauche de son espérance est encore appelée inégalité de concentration. Les inégalités de Talagrand permettent de contrôler par une inégalité exponentielle les déviations de $\bar{\gamma}_n(s_m) - \bar{\gamma}_n(\hat{s}_{\hat{m}})$, uniformément dans un modèle de la collection. Á ce stade deux situations sont possibles :

- Dans certains cas favorables, il est possible de contrôler finement l'espérance de $\bar{\gamma}_n(s_m)$ en fonction de la dimension des modèles. Les inégalités de Talagrand permettent alors d'aboutir à des inégalités oracle avec constantes explicites. C'est par exemple le cas pour les résultats obtenus par Castellan dans le cadre de l'estimation de densité par des histogrammes (Castellan, 1999), et pour l'estimation de densité dans des modèles exponentiels (Castellan, 2003). Notons que la méthode d'estimation développée par Lepez (2002) dans le contexte des réserves pétrolières s'appuie sur les travaux de Castellan.
- Dans les cas plus complexes, on ne peut exprimer facilement l'espérance de $\bar{\gamma}_n(s_m)$ en

fonction de la dimension des modèles. Les inégalités oracle ne permettent plus alors d'obtenir des bornes de risque raisonnables. Pour ces situations plus délicates, Massart (2007, section 7.4) propose une méthodologie s'appuyant sur un théorème général qui donne une condition suffisante pour garantir une inégalité oracle.

Le reste de cette section est consacrée à la description de la deuxième des deux alternatives présentées ci-dessus.

5.1.3 Sélection de modèles et entropie

Plutôt que de se baser directement sur la dimension des modèles pour définir une pénalité convenable, les résultats obtenus par Massart dans ce contexte s'appuient sur la notion l'entropie à crochets, qui permet elle aussi de donner une mesure de la taille d'un ensemble. Nous rappelons maintenant quelques définitions nécessaires avant d'énoncer le théorème général que nous utiliserons.

La norme $\|\sqrt{f}-\sqrt{g}\|_2$ entre deux fonctions positives f et g de $\mathbb{L}_1$ est notée $d_H(f,g)$. Lorsque f et g sont des densités par rapport à la mesure de Lebesgue sur $\mathbb{R}^Q$, $d_H^2(f,g)$ correspond au double de la distance de Hellinger au carré entre f et g. Dans la suite, $d_H(f,g)$ sera appelé par abus distance de Hellinger, même lorsque f et g ne sont pas des densités. Soit S un sous-ensemble de $\mathcal{S}$. Un ε-recouvrement de crochets de S pour d_H est un ensemble de paires de fonctions intégrables $(l_1,u_1),\ldots,(l_q,u_q)$ telles que

- pour tout $f\in S$, il existe $j\in\{1,\ldots,q\}$ tel que $l_j\le f\le u_j$,
- pour tout $j\in\{1,\ldots,q\}$, $d_H(l_j,u_j)\le\varepsilon$.

On note $\mathrm{N}_{[.]}(\varepsilon,S,d_H)$ le nombre minimal de ε-crochets nécessaire pour recouvrir S et l'entropie à crochets est définie par $\mathrm{H}_{[.]}(\varepsilon,S,d_H)=\ln\left\{\mathrm{N}_{[.]}(\varepsilon,S,d_H)\right\}$.

Soit $(S_m)_{m\in\mathcal{M}}$ une famille au plus dénombrable de modèles composés de densités de probabilité pour la mesure de Lebesgue sur $\mathbb{R}^d$. Afin d'éviter les problèmes de mesurabilité, nous supposons satisfaite l'hypothèse de séparabilité suivante. Pour tout modèle S_m, il existe un sous-ensemble dénombrable S'_m de S_m tel que pour tout $t\in S_m$, il existe une suite $(t_k)_{k\ge1}$ d'éléments de S'_m telle que, pour tout $x\in\mathbb{R}^d$, $\ln\{t_k(x)\}\to\ln\{t(x)\}$ lorsque $k\to+\infty$.

Ensuite, nous supposons aussi que pour tout m, la fonction $\varepsilon\mapsto\sqrt{\mathrm{H}_{[.]}(\varepsilon,S_m,d_H)}$ est intégrable en 0, et qu'il existe une fonction Ψ_m définie sur $\mathbb{R}_+$ vérifiant la propriété suivante,

> $(\mathbf{P_1})$: Ψ_m est croissante, $x\mapsto\Psi_m(x)/x$ est décroissante sur $]0,+\infty[$, et pour tout $\xi\in\mathbb{R}_+$ et tout $u\in S_m$,
>
> $$\int_0^\xi\sqrt{\mathrm{H}_{[.]}(x,S_m(u,\xi),d_H)}\,dx\le\Psi_m(\xi),$$
>
> où $S_m(u,\xi):=\{t\in S_m;d_H(t,u)\le\xi\}$.

Nous notons $\mathrm{KL}(s,S_m):=\inf_{t\in S_m}\mathrm{KL}(s,t)$ pour tout $m\in\mathcal{M}$. Sous les hypothèses précédentes, Massart (2007, Théorème 7.11) énonce le résultat suivant.

Théorème 5.1.1. *Soit $X_1,\ldots,X_n$ un échantillon i.i.d, avec $X_i\in\mathbb{R}^d$ de densité de probabilité s inconnue pour la mesure de Lebesgue sur $\mathbb{R}^d$. Pour tout m, $\hat{s}_m$ désigne l'estimateur du*

maximum de vraisemblance de s dans le modèle S_m. Soit $(\rho_m)_{m\in\mathcal{M}}$ une famille de poids positifs tels que

$$\sum_{m\in\mathcal{M}} e^{-\rho_m} = \Upsilon < \infty.$$

Pour tout $m \in \mathcal{M}$ soit une fonction Ψ_m satisfaisant la propriété $(\mathbf{P_1})$. On considère alors l'unique solution ξ_m de l'équation

$$\Psi_m(\xi) = \sqrt{n}\,\xi^2.$$

Pour une pénalité $\mathrm{pen} : \mathcal{M} \to \mathbb{R}_+$*, soit le critère de log-vraisemblance pénalisé*

$$\mathrm{crit}(m) = \gamma_n(\hat{s}_m) + \mathrm{pen}(m).$$

Sous les hypothèses précédentes, il existe des constantes absolues κ et C telles que, pour tout $m \in \mathcal{M}$, si

$$\mathrm{pen}(m) \geq \kappa \left(\xi_m^2 + \frac{\rho_m}{n}\right) \tag{5.1}$$

alors la variable aléatoire $\hat{m}$ qui minimise le critère crit *sur $\mathcal{M}$ existe, et de plus,*

$$\mathbb{E}\left[d_H^2(s, \hat{s}_{\hat{m}})\right] \leq C \left[\inf_{m\in\mathcal{M}} \{\mathrm{KL}(s, S_m) + \mathrm{pen}(m)\} + \frac{\Upsilon}{n}\right]. \tag{5.2}$$

5.1.4 Discussion

Les méthodes utilisées pour démontrer ce résultat ne permettent par d'évaluer avec précision les constantes en jeu dans la pénalité et l'inégalité oracle. C'est pourquoi, dans les applications de ce théorème à des situations particulières, nous ne chercherons pas non plus à donner de valeurs aux constantes rencontrées. La forme de la pénalité ainsi que la borne de risque non asymptotique dans le théorème 5.1.1 doivent être considérés d'un point de vue qualitatif. Essentiellement, ces résultats nous donnent la forme générale de la pénalité à utiliser pour la *méthode de la pente* (Birgé et Massart, 2006) qui permet dans un second temps de calibrer la pénalité en fonction des données. Cette méthode est exposée en détail dans la section 6.1.

L'inégalité (5.2) fait intervenir la distance de Hellinger dans le terme de droite, et l'information de Kullback-Leibler dans le terme de gauche ce qui peut d'abord paraître gênant car ces deux métriques ne sont pas rigoureusement équivalentes. Il est cependant possible de se ramener à une borne de risque entièrement exprimée à l'aide de la distance de Hellinger en utilisant le résultat suivant, donné dans ce but par Massart (2007, lemme 7.23). Pour P et Q deux mesures de probabilités telles que P est absolument continue par rapport à Q, alors

$$d_H^2(P, Q) \leq 2\,\mathrm{KL}(P, Q) \leq \left(2 + \left\|\frac{dP}{dQ}\right\|_\infty\right) d_H^2(P, Q).$$

Quitte à supposer qu'il existe une constante M telle que

$$\sup_{t \in \bigcup_{m\in\mathcal{M}} S_m} \left\|\frac{s}{t}\right\|_\infty \leq M,$$

nous obtenons alors une inégalité de la forme

$$\mathbb{E}\left[d_H^2(s,\hat{s}_{\hat{m}})\right] \leq C'\left[\inf_{m\in\mathcal{M}}\{d_H^2(s,S_m)+\mathrm{pen}(m)\}+\frac{\Upsilon}{n}\right], \tag{5.3}$$

où $d_H^2(s,S_m) := \inf_{t\in S_m} d_H^2(s,t)$.

L'un des intérêts essentiels du théorème 5.1.1 est que celui-ci ne nécessite pas que la vraie densité s soit dans la collection de modèles considérée. Cette propriété est très satisfaisante car en pratique, une densité n'est jamais "réellement" dans la collection de modèles. Notons que cette propriété dispense aussi de tout travail de validation de modèles. Évidemment, la borne de risque sera cependant d'autant plus fine que la densité s est bien approchée par la collection de modèles proposée pour l'information de Kulback-Leibler.

Les hypothèses du théorème 5.1.1 portent essentiellement sur le contrôle de l'entropie métrique à crochet des modèles considérés. Sans rentrer dans une description précise des outils théoriques qui permettent de démontrer le théorème, signalons que l'un des ingrédients principaux de la preuve est une inégalité maximale pour un processus empirique de la forme $Z = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(X_i)$ (voir Massart, 2007, théorème 6.3 p.193) où $\mathcal{F}$ est une classe de fonctions associée à l'un des ensembles de densités $S_m(u,\xi)$. Cette inégalité exponentielle fait directement intervenir l'entropie métrique à crochet de la classe $\mathcal{F}$, ce qui explique pourquoi le théorème 5.1.1 requiert aussi des hypothèses de cette nature. Notons de plus que l'hypothèse (P_1) touche à l'entropie métrique de façon locale puisqu'elle fait intervenir les ensembles $S_m(u,\xi)$. Nous verrons qu'il est souvent difficile de contrôler précisément ces entropies locales. Dans la plupart des cas, nous ne pouvons que majorer les entropies locales par l'entropie du modèle S_m tout entier. Nous discutons les conséquences de cet inconvénient sur la forme de la pénalité obtenue dans chacune des deux situations étudiées dans les sections 5.2 et 5.3.

Enfin, il est important de souligner que puisque les constantes κ et C ne dépendent pas de la collection de modèles $\mathcal{M}$, il est possible de choisir la collection de modèles en fonction de la taille de l'échantillon. Ce principe a notamment été utilisée auparavant par Castellan (1999) pour des résultats concernant l'estimation de densités par histogrammes, et nous suivons la même démarche pour définir des poids convenables dans le cas de la collection de modèles consacrée à l'étude de l'exploration pétrolière.

5.2 Sélection de modèles de mélanges gaussiens

En collaboration avec Cathy Maugis.

Les modèles de mélanges gaussiens $S_{(K,\mathbf{v})}$ présentés au chapitre précédent nous placent dans un contexte où l'application du théorème 5.1.1 est légitime. En effet, pour une densité de mélange, la log-vraisemblance fait apparaître le logarithme d'une somme, ce qui rend délicat une utilisation fine des inégalités de Talagrand. De plus, l'utilisation des inégalités de concentration (voir par exemple (5.50) p.170 dans Massart, 2007) nécessiterait de borner uniformément sur les modèles la quantité $\|\bar{\gamma}_n(s_m)-\bar{\gamma}_n(t)\|_\infty$, ce qui revient dans notre contexte à contrôler des rapports de densités de mélanges gaussiens uniformément sur R^Q. Or, la plus

plupart du temps, de tels rapports ne sont pas bornés si les matrices de variance-covariance sont distinctes. Ces différents obstacles nous empêchent d'utiliser directement les inégalités de concentrations ; nous allons donc nous appuyer sur le théorème 5.1.1 pour établir des résultats de sélection de modèle pour les familles de modèles de mélanges considérés.

5.2.1 Résultats principaux

Les deux théorèmes qui suivent exposent les résultats obtenus d'abord dans le cas ordonné, puis dans le cas non ordonné.

Cas ordonné

Dans le cas ordonné, le bloc des variables de classification est de la forme $\{1, \ldots, v\}$. Pour les trois types de collections de mélanges gaussiens considérés, les modèles $S_{(K,v)}$ sont tels que

$$S_{(K,v)} = \left\{ x \in \mathbb{R}^Q \mapsto f(x_1, \ldots, x_v)\, \Phi(x_{v+1}, \ldots, x_Q \,|\, 0, I_{Q-v}) \;;\;\; f \in \mathcal{L}_{(K,v)} \right\}.$$

Nous obtenons des résultats pour les collections de mélanges qui sont définies par les familles de densités $\mathcal{L}_{(K,v)}$ précisées ci-dessous. Chaque collection correspond à une certaine forme de matrice variance-covariance pour les composantes des mélanges choisis dans $\mathcal{L}_{(K,v)}$. Notons que par rapport aux définitions proposées au chapitre précédent, des hypothèses supplémentaires ont été rajoutées de façon à borner les paramètres des mélanges.

– Collection $\mathcal{M}[LB_k]$:

$$\mathcal{L}_{(K,v)} = \left\{ \sum_{k=1}^{K} p_k \Phi(\cdot|\mu_k, \lambda\Sigma_k); \begin{array}{l} \lambda \in [\sigma_{\mathrm{m}}^2, \sigma_{\mathrm{M}}^2],\ \Sigma_k \in \Delta^1_{(v)}(\sigma_{\mathrm{m}}^2, \sigma_{\mathrm{M}}^2) \\ \mu_k \in [-a, a]^v,\ 0 < p_k < 1,\ \sum\limits_{k=1}^{K} p_k = 1 \end{array} \right\}$$

où $\Delta^1_{(v)}(\sigma_{\mathrm{m}}^2, \sigma_{\mathrm{M}}^2)$ désigne l'ensemble des matrices diagonales définies positives de déterminant 1 dont les valeurs propres appartiennent à l'intervalle $[\sigma_{\mathrm{m}}^2, \sigma_{\mathrm{M}}^2]$.

– Collection $\mathcal{M}[L_kB_k]$:

$$\mathcal{L}_{(K,v)} = \left\{ \sum_{k=1}^{K} p_k \Phi(\cdot|\mu_k, \Sigma_k); \begin{array}{l} \Sigma_k = \mathrm{diag}(\sigma_{k1}^2, \ldots, \sigma_{kv}^2),\ \sigma_{k1}^2, \ldots, \sigma_{kv}^2 \in [\sigma_{\mathrm{m}}^2, \sigma_{\mathrm{M}}^2] \\ \mu_k \in [-a, a]^v,\ 0 < p_k < 1,\ \sum\limits_{k=1}^{K} p_k = 1 \end{array} \right\}.$$

– Collection $\mathcal{M}[L_kC_k]$:

$$\mathcal{L}_{(K,v)} = \left\{ \sum_{k=1}^{K} p_k \Phi(\cdot|\mu_k, \Sigma_k); \begin{array}{l} \Sigma_k \in \mathcal{D}^+_{(v)}(\sigma_{\mathrm{m}}^2, \sigma_{\mathrm{M}}^2),\ \mu_k \in [-a, a]^v \\ 0 < p_k < 1,\ \sum\limits_{k=1}^{K} p_k = 1 \end{array} \right\}$$

où $\mathcal{D}^+_{(v)}(\sigma_{\mathrm{m}}^2, \sigma_{\mathrm{M}}^2)$ désigne l'ensemble des matrices symétriques définies positives dont les valeurs propres appartiennent à l'intervalle $[\sigma_{\mathrm{m}}^2, \sigma_{\mathrm{M}}^2]$.

Dans le théorème qui suit, $\mathcal{M}$ désigne l'une des collections $\mathcal{M}[L_kB_k]$, $\mathcal{M}[L_kC_k]$ ou $\mathcal{M}[LB_k]$.

Théorème 5.2.1. *Il existe des constantes κ, A et C telles que, si*

$$\mathrm{pen}(K,v) \geq \kappa \frac{D(K,v)}{n} \left\{ 2A \ln v + 1 - \ln \left(1 \wedge \left[\frac{D(K,v)}{n} A \ln v \right] \right) \right\}$$

alors le modèle d'indice $(\hat{K}, \hat{v})$ qui minimise

$$\mathrm{crit}(K,v) = \gamma_n(\hat{s}_{(K,v)}) + \mathrm{pen}(K,v)$$

sur $\mathcal{M}$ existe, et

$$\mathbb{E}\left[d_H^2(s, \hat{s}_{(\hat{K},\hat{v})}) \right] \leq C \left[\inf_{(K,v)\in\mathcal{M}} \{ \mathrm{KL}(s, S_{(K,v)}) + \mathrm{pen}(K,v) \} + \frac{1}{n} \right]. \tag{5.4}$$

De plus, les constantes κ et C sont absolues, et la constante A ne dépend que de σ_m, σ_M et a.

La preuve de ce théorème est donnée dans la section 5.2.3, celle-ci s'appuie sur des calculs techniques d'entropie à crochets. Les bornes sur les entropies obtenues pour les modèles de chacune des collections $\mathcal{M}[L_kB_k]$, $\mathcal{M}[L_kC_k]$ et $\mathcal{M}[LB_k]$ sont respectivement démontrées dans les sections 5.2.5, 5.2.6 et 5.2.7.

Cas non ordonné

Rappelons que les modèles des collections non ordonnées sont notées $S_{(K,\mathbf{v})}$ où $\mathbf{v}$ désigne l'ensemble des variables de classification. Pour $v = |\mathbf{v}|$, un modèle $S_{(K,\mathbf{v})}$ s'obtient à partir du modèle $S_{(K,v)}$ de la collection ordonnée qui lui est associée en posant

$$S_{(K,\mathbf{v})} = \{ x \in \mathbb{R}^Q \mapsto f \circ \tau(x),\ f \in S_{(K,v)} \}, \tag{5.5}$$

où τ est une permutation telle que $(\tau(x)_1, \ldots, \tau(x)_v)' = x_{[\mathbf{v}]}$. Les collections de modèles non ordonnées sont notées $\mathcal{M}'[L_kB_k]$, $\mathcal{M}'[L_kC_k]$ et $\mathcal{M}'[LB_k]$. Dans le théorème suivant, l'ensemble $\mathcal{M}'$ correspond à l'une de ces trois collections.

Théorème 5.2.2. *Il existe des constantes κ, A et C telles que, si*

$$\mathrm{pen}(K,\mathbf{v}) \geq \kappa \frac{D(K,\mathbf{v})}{n} \left(2A \ln v - \ln \left(1 \wedge \left[\frac{D(K,v)}{n} A \ln v \right] \right) + \frac{1}{2} \ln \left[\frac{8eQ}{\{D(K,\mathbf{v}) - 1\} \wedge (2Q-1)} \right] \right) \tag{5.6}$$

alors le modèle $(\hat{K}, \hat{\mathbf{v}})$ qui minimise le critère

$$\mathrm{crit}(K,\mathbf{v}) = \gamma_n(\hat{s}_{(K,\mathbf{v})}) + \mathrm{pen}(K,\mathbf{v}) \tag{5.7}$$

sur la collection de modèles $\mathcal{M}'$ existe, et

$$\mathbb{E}\left[d_H^2(s, \hat{s}_{(\hat{K},\hat{\mathbf{v}})}) \right] \leq C \left[\inf_{(K,\mathbf{v})\in\mathcal{M}'} \{ \mathrm{KL}(s, S_{(K,\mathbf{v})}) + \mathrm{pen}(K,\mathbf{v}) \} + \frac{2}{n} \right]. \tag{5.8}$$

De plus, les constantes κ et C sont absolues et la constante A ne dépend que de σ_m, σ_M et a.

La preuve de ce résultat est donnée dans la section 5.2.3. D'après la relation (5.5), les modèles $S_{(K,\mathbf{v})}$ et $S_{(K,v)}$ ont même entropie métrique. La preuve du théorème 5.2.2 ne se distingue donc de celle du théorème 5.2.1 que par la définition de poids différents. Le terme logarithmique supplémentaire dans la pénalité (5.6) traduit la plus grande complexité de la collection $\mathcal{M}'$ dans le cas non ordonné.

Dans les deux théorèmes précédents, la constante A est identique pour les cas ordonné et non ordonné pour un type de mélange fixé. Puisque la constante κ qui provient du théorème 5.1.1 est de tout façon inconnue, il serait inutile de chercher à évaluer A précisément.

5.2.2 Discussion

Contrairement aux situations classiques en statistique pour lesquelles le nombre de variables Q est fixé alors que n tend vers l'infini, les deux résultats précédents permettent de considérer les situations où Q augmente avec n. Pour les problèmes spécifiques où le nombre de variables est de l'ordre de n, voir même plus grand que n, les deux inégalités oracles (5.8) et (5.4) montrent que le critère pénalisé est encore pertinent.

Bien les modèles LB_k soient des sous-modèles de la famille L_kB_k, qui sont eux mêmes des sous-modèles de la famille L_kC_k, les résultats des deux théorèmes précédents nécessitent une démonstration spécifique pour chacune des trois collections. En effet pour que le terme $D(K,v)$ dans la pénalité corresponde réellement au nombre de paramètres libres dans chacune des trois situations, il est nécessaire de mener les calculs d'entropie métrique pour chacun des trois types de modèles.

Nous avons vu dans la discussion à la suite du théorème 5.1.1 qu'il est possible de se ramener à une borne de risque qui ne fasse intervenir que la distance de Hellinger. Pour $\|x\| \to +\infty$, les densités de mélanges gaussiens des modèles ont tous le même comportement qu'une densité gaussienne. Pour garantir, que

$$\sup_{t\in S_{(K,v)}} \left\|\frac{s}{t}\right\|_\infty \leq M,$$

il est donc naturel de supposer que s est à support compact, dans ce cas la borne M dépend de a, σ_{m}, σ_{M} et $\|s\|_\infty$, mais il n'est pas nécessaire de la connaître pour utiliser le critère pénalisé en pratique.

Comme pour le théorème 5.1.1, les résultats obtenus doivent être interprétés de façon qualitative puisque leur principal intérêt est de fournir la forme de la pénalité à minimiser dans le critère. La méthode de la pente, une fois adaptée à ce contexte, permet dans un second temps de calibrer la pénalité. Nous donnons une description de la procédure correspondante dans la section 6.2.2.

Le terme $\frac{D(K,v)}{n} \ln\left\{1/(1 \wedge A\frac{D(K,v)}{n} \ln \mathbf{v})\right\}$ que nous retrouvons dans les collections ordonnées et non ordonnées n'est probablement pas nécessaire pour définir des pénalités optimales, c'est à dire conduisant à des estimateurs minimisant le risque d'estimation. La présence de ce terme est directement liée au fait que nous ne parvenons à contrôler l'entropie des modèles que de façon globale et non de façon locale comme il serait idéalement nécessaire pour appliquer le théorème 5.1.1. Il est en effet difficile d'effectuer des recouvrements fins des ensembles

$S_{(K,v)}(u,\xi) := \{t \in S_{(K,v)}; d_H(t,u) \leq \xi\}$; nous majorons l'entropie de ces ensembles par celle de $S_{(K,v)}$ tout entier. Nous verrons qu'en pratique, le facteur $\ln v$ dans le premier terme n'est pas non plus nécessaire pour définir des pénalités optimales. Une discussion est entièrement consacrée à la mise en pratique de nos résultats dans la section 6.2.5 du chapitre suivant, nous revenons à cette occasion sur la forme de la pénalité à utiliser en pratique.

Une façon de mettre en évidence qu'un estimateur est de "bonne qualité" est de considérer son risque maximal sur une classe $\mathcal{F}$ de fonctions de densités. Les performances de l'estimateur sont alors évaluées en comparant ce risque maximal au risque minimax sur la classe $\mathcal{F}$,

$$R_{\text{minimax}} := \inf_{\hat{s}} \sup_{s \in \mathcal{F}} \mathbb{E}\left(\|s - \hat{s}\|^2\right)$$

où la borne inférieure porte sur tous les estimateurs possibles $\hat{s}$. Généralement, on prend pour la classe $\mathcal{F}$ des classes fonctionnelles de Hölder, de Sobolev, de Besov, ou encore des ellipsoïdes de $\mathbb{L}^2$. Si l'estimateur pénalisé atteint la borne de risque minimax sur une large famille de classes fonctionnelles, sans utiliser la connaissance de la classe particulière à laquelle appartient ss, on dit que celui est adaptatif. Le fait que les inégalités oracle obtenues dans les théorèmes 5.2.2 et 5.2.1 soient valables pour toute densité s permet en particulier de majorer le risque maximal de $\hat{s}_{(\hat{K},Q)})$ (sans sélection de variables) pour la distance de Hellinger, sur une classe de fonction $\mathcal{F}$, par la quantité

$$C \sup_{s \in \mathcal{F}} \left[\inf_{(K,Q) \in \mathcal{M}} \{\mathrm{KL}(s, S_{(K,Q)}) + \mathrm{pen}(K)\} + \frac{1}{n} \right],$$

pour une collection de modèles ordonnés. Pour évaluer les performances de notre estimateur pénalisé en terme de risque minimax, la difficulté principale que nous rencontrons alors n'est pas d'ordre statistique, mais relève de l'approximation fonctionnelle. Pour une fonction s choisie dans l'une des classes citées plus haut, il n'existe pas à notre connaissance de résultat permettant de quantifier de façon précise le biais $\mathrm{KL}(s, S_{(K,Q)})$ (ou pour n'importe quelle autre métrique) entre la densité s et l'espace fonctionnel composé des densités de mélanges gaussiens à K composantes. Même en dimension 1, nous n'avons pu trouver un tel résultat. Á l'issue de ce travail de thèse, nous souhaitons réfléchir sur cette question afin d'améliorer les résultats obtenus ici. Notons qu'un résultat présenterait aussi un intérêt pour le domaine de l'approximation fonctionnelle.

5.2.3 Preuve des résultats principaux

Cette section est consacrée à la démonstration des résultats principaux. Nous montrons ici comment les bornes d'entropie à crochets calculées sur les modèles des différentes collections permettent de définir les pénalités. Pour cela, nous commençons par énoncer la proposition suivante.

Proposition 5.2.1. *Pour chacune des collections $\mathcal{M}[L_kB_k]$, $\mathcal{M}[L_kC_k]$ et $\mathcal{M}[LB_k]$, il existe*

une constante A_1 telle que pour tout $\varepsilon \in]0,1]$

$$H_{[.]}(\varepsilon, S_{(K,v)}, d_H) \leq D(K,v)\left[A_1 \ln v + \ln\left(\frac{1}{\varepsilon}\right)\right].$$

où la constante A_1 ne dépend que de σ_m, σ_M et a.

Ce résultat est démontré dans les sections suivantes pour chacune des trois collections. Puisque $\mathrm{H}_{[.]}(\varepsilon, S_{(K,\mathbf{v})}, d_H) = \mathrm{H}_{[.]}(\varepsilon, S_{(K,v)}, d_H)$ pour $|\mathbf{v}| = v$, la proposition est donc encore valable pour les collections de modèles $\mathcal{M}'[L_k B_k]$, $\mathcal{M}'[L_k B_k]$ et $\mathcal{M}'[L_k C_k]$. Nous montrons maintenant comment cette proposition permet de vérifier que les modèles $S_{(K,v)}$ satisfont la condition (P_1) qui est nécessaire pour utiliser le théorème 5.1.1 ; la démarche est exactement identique pour les modèles non ordonnés. Soit $\xi > 0$, en utilisant le lemme technique 5.2.18 donné en annexe dans la section 5.2.8, nous obtenons

$$\begin{aligned}
\int_0^{\xi} \sqrt{\mathrm{H}_{[.]}(x, S_{(K,v)}, d_H)}\, dx &\leq \sqrt{D(K,v)}\left[\xi\sqrt{A_1 \ln v} + \int_0^{\xi\wedge 1} \sqrt{\ln\left(\frac{1}{x}\right)}\, dx\right] \\
&\leq \xi\sqrt{D(K,v)}\left[\sqrt{A_1 \ln v} + \sqrt{\ln\left(\frac{1}{1\wedge\xi}\right)} + \sqrt{\pi}\right].
\end{aligned}$$

On pose $A_2 = \sqrt{A_1} + \sqrt{\pi}$ et la fonction

$$\Psi_{(K,v)} : \xi \in \mathbb{R}_+^\star \mapsto \xi\sqrt{D(K,v)}\left\{A_2\sqrt{\ln v} + \sqrt{\ln\left(\frac{1}{1\wedge\xi}\right)}\right\}$$

vérifie clairement la condition (P_1). Il nous faut ensuite trouver $\xi_\star$ tel que $\Psi_{(K,v)}(\xi_\star) = \sqrt{n}\,\xi_\star^2$ pour en déduire l'expression de la pénalité minimale. Ceci revient à résoudre

$$\sqrt{\frac{D(K,v)}{n}}\left\{A_2\sqrt{\ln v} + \sqrt{\ln\left(\frac{1}{1\wedge\xi_\star}\right)}\right\} = \xi_\star.$$

On note que la quantité $\tilde{\xi} = \sqrt{\frac{D(K,v)}{n}}\, A_2\sqrt{\ln v}$ vérifie $\tilde{\xi} \leq \xi_\star$, ce qui donne

$$\xi_\star \leq \sqrt{\frac{D(K,v)}{n}}\left\{A_2\sqrt{\ln v} + \sqrt{\ln\left(\frac{1}{1\wedge\tilde{\xi}}\right)}\right\},$$

et donc

$$\xi_\star^2 \leq \frac{D(K,v)}{n}\left\{2\,A_2^2 \ln v + \ln\left(\frac{1}{1\wedge\frac{D(K,v)}{n}\,A_2^2 \ln v}\right)\right\}. \tag{5.9}$$

Á ce stade, il ne reste qu'à définir des poids ρ_m convenables pour terminer la preuve des théorèmes 5.2.1 et 5.2.2.

Cas d'une collection $\mathcal{M}$ de modèles ordonnés

La famille de poids $\rho_{(K,v)} = D(K,v)$ vérifie $\sum_{(D,v)\in\mathcal{M}} e^{-\rho_{(K,v)}} = 1$. En effet,

$$\operatorname{card}\{(K,v) \in \mathbb{N}^\star \times \{1,\dots,Q\};\ D(K,v) = D\} \leq D$$

et $\sum\limits_{(K,v)} e^{-\rho_{(K,v)}} \leq \sum\limits_{D\geq 1} D\, e^{-D} \leq 1$. Il suffit alors d'appliquer le théorème 5.1.1 à chacune des collections $\mathcal{M}$ avec les poids $\rho_{(K,v)}$ et la majoration (5.9) pour achever la preuve du théorème 5.2.1.

Cas d'une collection $\mathcal{M}'$ de modèles non ordonnés

Nous définissons maintenant des poids pour les trois collections $\mathcal{M}'[L_kB_k]$, $\mathcal{M}'[L_kC_k]$ et $\mathcal{M}'[LB_k]$. Comme précédemment v désigne le cardinal de $\mathbf{v}$.

Lemme 5.2.2. *Pour chacune des trois collections, soit* $\mathcal{A}_D = \{(K,\mathbf{v}) \in \mathbb{N}^\star \times \mathcal{V};\ D(K,\mathbf{v}) = D\}$. *Alors,*

$$\operatorname{card}\mathcal{A}_D \leq \begin{cases} 2^Q & \textit{si } Q \leq \frac{D-1}{2} \\ \left(\frac{2eQ}{D-1}\right)^{\frac{D-1}{2}} & \textit{sinon} \end{cases}.$$

Démonstration. 1. Pour la collection $\mathcal{M}[L_kB_k]$, nous avons

$$\begin{aligned} \operatorname{card}\{(K,\mathbf{v}) \in \mathbb{N}^\star \times \mathcal{V};\ D(K,\mathbf{v}) = D\} &= \operatorname{card}[(K,\mathbf{v}) \in \mathbb{N}^\star \times \mathcal{V};\ K\{2\operatorname{card}(\mathbf{v})+1\} = D] \\ &= \sum_{K=1}^{\infty}\sum_{v=1}^{Q}\binom{Q}{v}\mathbb{1}_{K(2v+1)=D} \\ &\leq \sum_{v=1}^{\infty}\binom{Q}{v}\mathbb{1}_{v\leq Q\wedge\lfloor\frac{D-1}{2}\rfloor}. \end{aligned}$$

Si $Q \leq \lfloor\frac{D-1}{2}\rfloor$, alors $\sum\limits_{v=1}^{\infty}\binom{Q}{v}\mathbb{1}_{v\leq Q\wedge\lfloor\frac{D-1}{2}\rfloor} = 2^Q$. D'autre part, d'après la proposition 2.5 de Massart (2007), on a

$$\sum_{v=1}^{\infty}\binom{Q}{v}\mathbb{1}_{v\leq Q\wedge\lfloor\frac{D-1}{2}\rfloor} \leq f\left(\left\lfloor\frac{D-1}{2}\right\rfloor\right)$$

avec $f(x) = \left(\frac{eQ}{x}\right)^x$. Ceci nous donne que

$$\sum_{v=1}^{Q\wedge\lfloor\frac{D-1}{2}\rfloor}\binom{Q}{v} \leq \begin{cases} 2^Q & \text{si } Q \leq \frac{D-1}{2} \\ \left(\frac{2eQ}{D-1}\right)^{\frac{D-1}{2}} & \text{sinon} \end{cases}.$$

2. La collection de modèles $\mathcal{M}[LB_k]$ se traite de la même façon que la collection $\mathcal{M}[L_kB_k]$.

3. Pour la collection de modèles $\mathcal{M}[L_kC_k]$, $D(K,\mathbf{v}) = K\left[1+v+\frac{v\{v+1\}}{2}\right]$ et

$$\operatorname{card}\mathcal{A}_D \leq \sum_{v=1}^{Q}\binom{Q}{v}\mathbb{1}_{1+\frac{3}{2}v+\frac{v^2}{2}\leq D} \leq \sum_{v=1}^{Q}\binom{Q}{v}\mathbb{1}_{v\leq\frac{D-1}{2}}.$$

Nous retrouvons alors la même majoration que pour la collection $\mathcal{M}[L_kB_k]$.

□

Proposition 5.2.3. *Pour chacune des trois collections $\mathcal{M}'$ de modèles non ordonnés, on considère la famille de poids $\{\rho_{(K,\mathbf{v})}\}_{(K,\mathbf{v})\in\mathbb{N}^\star\times\mathcal{V}}$ définie par*

$$\rho_{(K,\mathbf{v})} = \frac{D(K,\mathbf{v})}{2}\ln\left[\frac{8eQ}{\{D(K,\mathbf{v})-1\}\wedge(2Q-1)}\right].$$

Alors, $\sum_{(K,\mathbf{v})\in\mathbb{N}^\star\times\mathcal{V}} e^{-\rho_{(K,\mathbf{v})}} \leq 2.$

Démonstration. D'après le lemme 5.2.2,

$$\begin{aligned}
\sum_{(K,\mathbf{v})\in\mathbb{N}^\star\times\mathcal{V}} e^{-\rho_{(K,\mathbf{v})}} &= \sum_{D=3}^{\infty}\exp\left[-\frac{D}{2}\ln\left\{\frac{8eQ}{(D-1)\wedge(2Q-1)}\right\}\right]\operatorname{card}\{(K,\mathbf{v});\ D(K,\mathbf{v})=D\}\\
&\leq \sum_{D=3}^{\infty}\exp\left[-\frac{D}{2}\ln\left\{\frac{8eQ}{(D-1)\wedge(2Q-1)}\right\}\right]\left\{2^Q\mathbb{1}_{Q\leq\frac{D-1}{2}}+\left(\frac{2eQ}{D-1}\right)^{\frac{D-1}{2}}\mathbb{1}_{\frac{D-1}{2}<Q}\right\}\\
&\leq \sum_{D=3}^{2Q}\exp\left\{-\frac{D}{2}\ln\left(\frac{8eQ}{D-1}\right)+\frac{D-1}{2}\ln\left(\frac{2eQ}{D-1}\right)\right\}\\
&+ \sum_{D=2Q+1}^{\infty}\exp\left\{-\frac{D}{2}\ln\left(\frac{8eQ}{2Q-1}\right)+Q\ln(2)\right\}.
\end{aligned}$$

La première somme vérifie,

$$\begin{aligned}
\exp\left\{-\frac{D}{2}\ln\left(\frac{8eQ}{D-1}\right)+\frac{D-1}{2}\ln\left(\frac{2eQ}{D-1}\right)\right\} &= \exp\left\{-\frac{D}{2}\ln(4)-\frac{1}{2}\ln\left(\frac{2eQ}{D-1}\right)\right\}\\
&\leq \exp\{-(D-1)\ln(2)\}
\end{aligned}$$

car $D\leq 2Q$. Pour la seconde somme, puisque $D\geq 2Q+1$, nous avons

$$\begin{aligned}
\exp\left\{-\frac{D}{2}\ln\left(\frac{8eQ}{2Q-1}\right)+Q\ln(2)\right\} &= \exp\left\{-\frac{3D}{2}\ln(2)+Q\ln(2)-\frac{D}{2}\ln\left(\frac{eQ}{2Q-1}\right)\right\}\\
&\leq \exp\left\{\left(Q-\frac{D-1}{2}\right)\ln(2)-(D-1)\ln(2)\right\}\\
&\leq \exp\{-(D-1)\ln(2)\}.
\end{aligned}$$

Et donc,

$$\begin{aligned}\sum_{(K,\mathbf{v})\in\mathbb{N}^\star\times\mathcal{V}} e^{-\rho(K,\mathbf{v})} &\leq \sum_{D=3}^{\infty}\left(\frac{1}{2}\right)^{D-1}\\ &\leq 2.\end{aligned}$$

□

On achève la preuve du théorème 5.2.1 en appliquant le théorème 5.1.1 à chacune des collections $\mathcal{M}'$ avec les poids $\rho_{(K,v)}$ et la majoration (5.9).

5.2.4 Résultats préliminaires pour le calcul des entropies

Avant de calculer les entropies des modèles des trois collections, nous commençons par donner ici quelques résultats que nous utiliserons à plusieurs reprises dans les sections suivantes. Soit $\mathcal{T}_{K-1}$ le simplexe de dimension $K-1$,

$$\mathcal{T}_{K-1} := \left\{\mathbf{p} = (p_1,\ldots,p_K),\ \forall k = 1,\ldots,K,\ p_k \geq 0,\ \sum_{k=1}^{K} p_k = 1\right\}.$$

Le lemme suivant est dû à Genovese et Wasserman (2000), il nous permet de majorer l'entropie métrique de $\mathcal{T}_{K-1}$.

Lemme 5.2.4. *Si $\varepsilon \leq 1$, alors*

$$N_{[.]}\left(\varepsilon, \mathcal{T}_{K-1}, d_H\right) \leq K(2\pi\, e)^{\frac{K}{2}}\left(\frac{1}{\varepsilon}\right)^{K-1}.$$

La proposition suivante permet de ramener le calcul de l'entropie d'une famille de mélanges au calcul de l'entropie pour les composantes du mélange. La démonstration de ce résultat peut également être trouvée dans Genovese et Wasserman (2000).

Proposition 5.2.5. *Soit $\mathcal{W}_K$ une famille de distributions de mélanges sur $\mathbb{R}^d$,*

$$\mathcal{W}_K := \left\{\sum_{k=1}^{K} p_k f_k\ ;\ \forall k = 1,\ldots,K,\ f_k \in \mathcal{F}_k\,,\ \boldsymbol{p} = (p_1,\ldots,p_K) \in \mathcal{T}_{K-1}\right\}$$

où les ensembles $\mathcal{F}_k$ sont des ensembles de densités pour la mesure de Lebesgue sur $\mathbb{R}^d$. Alors,

$$N_{[.]}(\varepsilon, \mathcal{W}_K, d_H) \leq N_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{T}_{K-1}, d_H\right)\prod_{k=1}^{K} N_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{F}_k, d_H\right),$$

et donc

$$H_{[.]}(\varepsilon, \mathcal{W}_K, d_H) \leq H_{[.]}\left(, \frac{\varepsilon}{3}\mathcal{T}_{K-1}, d_H\right) + \sum_{k=1}^{K} H_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{F}_k, d_H\right).$$

Nous ne pouvons utiliser une unique méthode pour calculer les entropies pour chacun des trois type de modèles de mélange, et il nous faut exposer des preuves pour chacune des trois situations dans les sections qui suivent.

5.2.5 Entropies pour les modèles de la collection $\mathcal{M}[L_k B_k]$

Cette section est consacrée à la preuve de la proposition 5.2.1 pour la collection de modèles $\mathcal{M}[L_k B_k]$. Dans ce but, nous commençons par majorer l'entropie métrique des ensembles $\mathcal{F}_{(v)}$ suivants,

$$\mathcal{F}_{(v)} = \left\{\Phi(\cdot|\mu,\Sigma);\ \mu \in [-a,a]^v,\ \Sigma = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_v^2),\ \sigma_1^2,\ldots,\sigma_v^2 \in [\sigma_{\mathrm{m}}^2,\sigma_{\mathrm{M}}^2]\right\}.$$

Proposition 5.2.6. *On pose* $c_1 = \frac{5}{8}\left(1-2^{-\frac{1}{4}}\right)$. *Pour tout* $\varepsilon \in]0,1]$,

$$H_{[.]}(\varepsilon,\mathcal{F}_{(v)},d_H) \leq v\ln\left(2\,a\sqrt{\frac{2}{c_1\,\sigma_m^2}}\right) + 2v\ln\left(8\frac{\sigma_M}{\sigma_m}\right) + 2v\ln(\sqrt{2}\,v) + 2v\ln\left(\frac{1}{\varepsilon}\right) \tag{5.10}$$

Démonstration. La preuve de l'inégalité (5.10) est une adaptation d'une preuve proposée dans Genovese et Wasserman (2000) qui montre des résultats similaires pour des familles de densités gaussiennes unidimensionnelles. L'idée principale consiste à définir un réseau sur l'espace des paramètres $\mathcal{B} = \left\{(\mu,\sigma_1^2,\ldots,\sigma_v^2) \in [-a,a]^v \times [\sigma_{\mathrm{m}}^2,\sigma_{\mathrm{M}}^2]^v\right\}$ pour en déduire un recouvrement de $\mathcal{F}_{(v)}$ par des crochets pour la distance de Hellinger.

Soit $\varepsilon \in]0,1]$ et $\delta = \frac{\varepsilon}{\sqrt{2}v}$. Pour tout $j \in \{2,\ldots,r\}$, on pose

$$b_j^2 = (1+\delta)^{1-\frac{j}{2}}\,\sigma_{\mathrm{M}}^2$$

avec $r = \left\lceil 2\,\frac{\ln\left\{\frac{\sigma_{\mathrm{M}}^2(1+\delta)}{\sigma_{\mathrm{m}}}\right\}}{\ln(1+\delta)}\right\rceil$ de façon à ce que $b_r^2 \leq \sigma_{\mathrm{m}}^2 < \sigma_{\mathrm{M}}^2 = b_2^2$. La notation $\lceil h \rceil$ désigne le plus petit entier supérieur ou égal à h. Ensuite, pour tout $J = (j(1),\ldots,j(v)) \in \{2,\ldots,r\}^v$, définissons la matrice B_J par

$$B_J = \mathrm{diag}(b_{j(1)}^2,\ldots,b_{j(v)}^2)\ .$$

Nous considérons aussi les vecteurs

$$\nu_J = (\nu_1^{(J)},\ldots,\nu_v^{(J)}) \in [-a,a]^v$$

tels que

$$\forall q \in \{1,\ldots,v\},\ \nu_q^{(J)} = \sigma_{\mathrm{M}}\sqrt{c_1}\,\delta\,(1+\delta)^{\frac{1-j(q)}{4}}\,s_q,$$

où $s_q \in \mathbb{Z} \cap [-A,A]$ avec $A = \left\lfloor \frac{a\,\delta^{-1}\,(1+\delta)^{-\frac{1-j(q)}{4}}}{\sigma_{\mathrm{M}}\sqrt{c_1}} \right\rfloor$. L'ensemble de tous les couples (ν_J, B_J) forme un réseau sur $\mathcal{B}$ que l'on note $\mathcal{R}(\varepsilon,v)$.

Cet ensemble $\mathcal{R}(\varepsilon,v)$ permet de construire des crochets qui définissent un ε-recouvrement de $\mathcal{F}_{(v)}$. Soit une densité $f(\cdot) = \Phi(\cdot|\mu,\Sigma)$ de $\mathcal{F}_{(v)}$, nous considérons les deux fonctions

$$\begin{cases} l(x) = (1+\delta)^{-v}\,\Phi(x|\nu_J,(1+\delta)^{-\frac{1}{4}}\,B_{J+1}) \\ u(x) = (1+\delta)^{v}\,\Phi(x|\nu_J,(1+\delta)\,B_J) \end{cases}.$$

Dans ces fonctions, le vecteur d'indices $J = (j(1), \ldots, j(v))$ est choisi pour que $b^2_{j(q)+1} \leq \sigma^2_q \leq b^2_{j(q)}$ pour tout q dans $\{1, \ldots, v\}$, et ν_J vérifie de plus

$$(\mu - \nu_J)' B^{-1}_{J+1} (\mu - \nu_J) \leq c_1 \, v \, \delta^2 \tag{5.11}$$

avec la notation $J+1 := (j(1)+1, \ldots, j(v)+1)$. Nous vérifions maintenant que $[l, u]$ contient f. L'inégalité (5.11) implique que

$$(\mu - \nu_J)' B^{-1}_J (\mu - \nu_J) \leq \frac{v}{4} \delta^2. \tag{5.12}$$

La proposition 5.2.19 permet de majorer le rapport de deux densités gaussiennes multidimensionnelles. Ce résultat, avec l'inégalité (5.12) nous donne que

$$\begin{aligned}
\frac{f(x)}{u(x)} &= \frac{\Phi(x|\mu, B)}{(1+\delta)^v \, \Phi(x|\nu_J, (1+\delta) \, B_J)} \\
&\leq (1+\delta)^{-\frac{v}{4}} \exp\left[\frac{1}{2\,\delta} (\mu - \nu_J)' \, B_J^{-1} \, (\mu - \nu_J)\right] \\
&\leq 1.
\end{aligned}$$

La fonction $h : \delta \mapsto 1 - (1+\delta)^{-\frac{1}{4}}$ étant concave, nous avons donc que $1 - (1+\delta)^{-\frac{1}{4}} \geq \delta(1 - 2^{-\frac{1}{4}})$. Ceci, avec la proposition 5.2.19 et l'inégalité (5.11) permet de montrer que

$$\begin{aligned}
\frac{l(x)}{f(x)} &= \frac{(1+\delta)^{-v} \Phi(x|\nu_J, (1+\delta)^{-\frac{1}{4}} \, B_{J+1})}{\Phi(x|\mu, B)} \\
&\leq (1+\delta)^{-\frac{5v}{8}} \exp\left[\frac{(\mu - \nu_J)' \, B^{-1}_{J+1} (\mu - \nu_J)}{2[1 - (1+\delta)^{-\frac{1}{4}}]}\right] \\
&\leq 1.
\end{aligned}$$

Par conséquent, le crochet $[l, u]$ contient la fonction f.

Il nous faut maintenant calculer $d_H(l, u)$ et montrer que $[l, u]$ est un ε-crochet.D'après le corollaire 5.2.20,

$$\begin{aligned}
d^2_H(l, u) &= d^2_H\left((1+\delta)^{-v} \Phi(\cdot|\nu_J, (1+\delta)^{-\frac{1}{4}} \, B_{J+1}), (1+\delta)^v \Phi(x|\nu_J, (1+\delta) \, B_J)\right) \\
&= (1+\delta)^{-v} + (1+\delta)^v - 2\left\{\frac{2}{(1+\delta)^{-\frac{7}{8}} + (1+\delta)^{\frac{7}{8}}}\right\}^{\frac{v}{2}} \\
&= \underbrace{2\,\mathrm{ch}(v \ln[1+\delta]) - 2}_{(i)} + \underbrace{2 - 2\left[\mathrm{ch}\left\{\frac{7}{8} \ln(1+\delta)\right\}\right]^{-\frac{v}{2}}}_{(ii)}.
\end{aligned}$$

En majorant séparément les termes (i) et (ii), nous obtenons que

$$\begin{aligned}
d^2_H(l, u) &\leq \left\{\mathrm{sh}(1) + \frac{49}{128}\right\} v^2 \delta^2 \\
&\leq 2\, v^2 \delta^2 \\
&\leq \varepsilon^2.
\end{aligned}$$

En conséquence, la famille de paramètres $\mathcal{R}(\varepsilon, v)$ induit un recouvrement de $\mathcal{F}_{(v)}$ par une famille de ε-crochets.

Le cardinal de $\mathcal{R}(\varepsilon, v)$ fournit une majoration de $\mathrm{N}_{[.]}(\varepsilon, \mathcal{F}_{(v)}, d_H)$.

$$\begin{aligned}
\mathrm{N}_{[.]}(\varepsilon, \mathcal{F}_{(v)}, d_H) &\leq \mathrm{Card}(\mathcal{R}(\varepsilon, v)) \\
&\leq \sum_{J \in \{2,\ldots,r\}^v} \prod_{q=1}^{v} \left\{ \frac{2\,a}{\sigma_{\mathrm{M}}\,\sqrt{c_1}\,\delta\,(1+\delta)^{\frac{1-j(q)}{4}}} \right\} \\
&\leq \left\{ \frac{2\,a(1+\delta)^{\frac{r-1}{4}}}{\sigma_{\mathrm{M}}\,\sqrt{c_1}\delta} \right\}^v (r-1)^v.
\end{aligned}$$

D'après la définition de r, $(1+\delta)^{\frac{r-1}{4}} \leq \frac{\sigma_{\mathrm{M}}}{\sigma_{\mathrm{m}}} \sqrt{(1+\delta)}$. Donc,

$$\begin{aligned}
\mathrm{N}_{[.]}(\varepsilon, \mathcal{F}_{(v)}, d_H) &\leq \left(\frac{2\,a}{\delta\,\sigma_{\mathrm{m}}} \sqrt{\frac{1+\delta}{c_1}} \right)^v \left[2\,\frac{\ln\left\{ \frac{\sigma_{\mathrm{M}}^2}{\sigma_{\mathrm{m}}^2}(1+\delta) \right\}}{\ln(1+\delta)} \right]^v \\
&\leq \left(\frac{2\sqrt{2}\,a}{\sigma_{\mathrm{m}}\,\sqrt{c_1}} \right)^v \left(\frac{8\,\sigma_{\mathrm{M}}^2}{\sigma_{\mathrm{m}}^2} \right)^v \delta^{-2v} \\
&\leq \left(\frac{2\sqrt{2}\,a}{\sigma_{\mathrm{m}}\,\sqrt{c_1}} \right)^v \left(\frac{8\,\sigma_{\mathrm{M}}^2}{\sigma_{\mathrm{m}}^2} \right)^v \left(\frac{\sqrt{2}v}{\varepsilon} \right)^{2v},
\end{aligned}$$

ce qui donne l'inégalité annoncée dans la proposition 5.2.6. □

Fin de la preuve de la proposition 5.2.1 pour la collection $\mathcal{M}[L_k B_k]$
Rappelons que

$$S_{(K,v)} = \left\{ x \in \mathbb{R}^Q \mapsto f(x_1, \ldots, x_v)\,\Phi(x_{v+1}, \ldots, x_Q \,|\, 0, I_{Q-v})\ ;\ \ f \in \mathcal{L}_{(K,v)} \right\},$$

ce qui implique que

$$\mathrm{H}_{[.]}(\varepsilon, \mathcal{S}_{(K,v)}, d_H) = \mathrm{H}_{[.]}(\varepsilon, \mathcal{L}_{(K,v)}, d_H).$$

D'après la proposition 5.2.5, l'entropie de $\mathcal{L}_{(K,v)}$ peut être majorée de la façon suivante,

$$\mathrm{H}_{[.]}(\varepsilon, \mathcal{L}_{(K,v)}, d_H) \leq \mathrm{H}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{T}_{K-1}, d_H\right) + K\,\mathrm{H}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{F}_v, d_H\right).$$

Nous déduisons alors de la proposition 5.2.6 et du lemme 5.2.4 que

$$\begin{aligned}
\mathrm{H}_{[.]}(\varepsilon, \mathcal{L}_{(K,v)}, d_H) \leq \quad & Kv\left[\ln\left(2\,a\sqrt{\frac{2}{c_1\,\sigma_{\mathrm{m}}^2}}\right) + 2\ln\left(8\frac{\sigma_{\mathrm{M}}}{\sigma_{\mathrm{m}}}\right) + 2\ln(\sqrt{2}\,v)\right] + \ln K + \frac{K}{2}\ln(2\pi\,e) \\
& + [K(2v+1)-1]\ln\left(\frac{3}{\varepsilon}\right).
\end{aligned}$$

La dimension du modèle $S_{(K,\mathbf{v})}$ a pour valeur $D(K,v) = K(2v+1)-1$, et il est alors clair que

$$\mathrm{H}_{[.]}(\varepsilon, \mathcal{L}_{(K,v)}, d_H) \leq \left(A_1 \ln v + \ln\frac{1}{\varepsilon}\right) D(K,v)$$

où la constante A_1 ne dépend que de σ_{m}^2, σ_{M}^2 et a. Ceci achève la preuve de la proposition 5.2.1 pour la collection de modèles $\mathcal{M}[L_kB_k]$.

5.2.6 Entropies pour les modèles de la collection $\mathcal{M}[L_kC_k]$

Cette section est consacrée à la preuve de la proposition 5.2.1 pour la collection de modèles $\mathcal{M}[L_kC_k]$. Nous adoptons la même démarche que celle utilisée pour la collection $\mathcal{M}[L_kB_k]$ en majorant l'entropie métrique des ensembles $\mathcal{F}_{(v)}$ suivants,

$$\mathcal{F}_{(v)} = \left\{\Phi(\cdot|\mu,\Sigma)\ ;\ \mu\in[-a,a]^v,\ \Sigma\in\mathcal{D}^+_{(v)}(\sigma_{\mathrm{m}}^2,\sigma_{\mathrm{M}}^2)\right\}$$

Proposition 5.2.7. *Pour tout* $\varepsilon\in]0,1]$,

$$\begin{aligned} H_{[.]}\left(\varepsilon,\mathcal{F}_{(v)},d_H\right) &\leq \frac{v(v+1)}{2}\ln\left(\frac{6\sqrt{3}\,\sigma_M}{\sigma_m}\right)+v\ln\left(\frac{6\,a}{\sigma_m}\right)\\ &+\ 2\left\{\frac{v(v+1)}{2}+v\right\}\ln(v)+\left\{\frac{v(v+1)}{2}+v\right\}\ln\left(\frac{1}{\varepsilon}\right). \end{aligned}$$

Une idée naturelle pour démontrer ce résultat serait d'utiliser la décomposition en valeurs propres de la matrice Σ pour se ramener à la situation des matrices diagonales que nous avons traitée dans le cadre de la collection $\mathcal{M}[L_kB_k]$. Mais pour cela, il faudrait aussi être capable de construire un réseau sur les matrices orthogonales qui interviennent dans l'opération de changement de bases. Or, pour $v>3$, il n'existe pas de paramétrage évident des matrices orthogonales sur lequel s'appuyer pour construire un tel réseau. Une autre démarche doit donc être proposée pour démontrer la proposition 5.2.7.

L'idée principale de la démonstration consiste à définir un recouvrement adéquat de $\mathcal{D}^+_{(v)}(\sigma_{\mathrm{m}}^2,\sigma_{\mathrm{M}}^2)$ pour la norme uniforme, et d'utiliser ensuite celui-ci pour construire un recouvrement par des crochets de $\mathcal{F}_{(v)}$ pour la distance de Hellinger. Dans ce qui suit, plusieurs résultats techniques sont d'abord démontrés avant de définir la famille de crochets permettant de recouvrir $\mathcal{F}_{(v)}$. La preuve de la proposition 5.2.7 est ensuite achevée par des arguments de dénombrement. La fin de cette section est consacrée à la preuve de la proposition 5.2.1 pour les modèles de la collection $\mathcal{M}[L_kC_k]$.

Dans la suite, nous adoptons les notations suivantes : $\|B\|_\infty=\max\limits_{1\leq i,j\leq v}|B_{ij}|$ et $|||B|||=\sup\limits_{\|x\|_2=1}|x'Bx|=\sup\limits_{\lambda\in\mathrm{vp}(B)}|\lambda|$, où $\mathrm{vp}(B)$ désigne l'ensemble des valeurs propres de B.

Définition du réseau de matrices de covariance

Soit $\beta>0$, nous définissons le β-réseau $\mathcal{R}(\beta)$ sur $\mathcal{D}^+_{(v)}(\sigma_{\mathrm{m}}^2,\sigma_{\mathrm{M}}^2)$ pour la norme uniforme de la façon suivante,

$$\mathcal{R}(\beta)=\left\{A=(A_{ij})_{1\leq i,j\leq v};\ A_{ij}=a_{ij}\beta;\ a_{ij}=a_{ji}\in\mathbb{Z}\cap\left[-\left\lfloor\frac{\sigma_{\mathrm{M}}^2}{\beta}\right\rfloor,\left\lfloor\frac{\sigma_{\mathrm{M}}^2}{\beta}\right\rfloor\right]\right\}.$$

Notons d'une part que $\mathcal{R}(\beta)$ est composé exclusivement de matrices symétriques, et d'autre part que pour tout Σ dans $\mathcal{D}^+_{(v)}(\sigma_{\mathrm{m}}^2, \sigma_{\mathrm{M}}^2)$, il existe A dans $\mathcal{R}(\beta)$ tel que

$$\|A - \Sigma\|_\infty \leq \beta. \tag{5.13}$$

Pour ces deux matrices, le lemme suivant permet de comparer les valeurs propres de A et de Σ.

Lemme 5.2.8. *Soit $\Sigma \in \mathcal{D}^+_{(v)}(\sigma_m^2, \sigma_M^2)$ et $A \in \mathcal{R}(\beta)$ tels que $\|\Sigma - A\|_\infty \leq \beta$. Soient $\sigma_1^2, \ldots, \sigma_v^2$ et $\tau_1^2, \ldots, \tau_v^2$ les valeurs propres respectives de Σ et A, rangées par ordre croissant et comptées avec leur multiplicité. Alors, pour tout $q \in \{1, \ldots, v\}$, nous avons*

$$\tau_q^2 - \beta v \leq \sigma_q^2 \leq \tau_q^2 + \beta v.$$

Démonstration. Puisque $\|\Sigma - A\|_\infty \leq \beta$, nous avons $|||\Sigma - A||| \leq \beta v$. De plus, d'après le théorème de Rayleigh, (voir Serre, 2002, théorème 3.3.2 p.49),

$$\sigma_q^2 = \min_{\dim(F)=q} \max_{x \in F \setminus \{0\}} \frac{x'\Sigma x}{\|x\|_2^2} \quad \text{et} \quad \tau_q^2 = \min_{\dim(F)=q} \max_{x \in F \setminus \{0\}} \frac{x'Ax}{\|x\|_2^2},$$

où F est un sous-espace linéaire de $\mathbb{R}^v$. On en déduit que pour tout $q \in \{1, \ldots, v\}$, $\tau_q^2 - \beta v \leq \sigma_q^2 \leq \tau_q^2 + \beta v$. □

Définition de la famille de ε-crochets sur $\mathcal{F}_{(v)}$

Nous construisons maintenant des ε-crochets à partir de l'ensemble $\mathcal{R}(\beta)$. Soit $f = \Phi(\cdot|\, \mu, \Sigma)$ une fonction de $\mathcal{F}_{(v)}$ avec $\mu \in [-a, a]^v$ et $\Sigma \in \mathcal{D}^+_{(v)}(\sigma_{\mathrm{m}}^2, \sigma_{\mathrm{M}}^2)$. D'après (5.13), pour tout $\beta > 0$ il existe une matrice $A \in \mathcal{R}(\beta)$ telle que $\|A - \Sigma\|_\infty \leq \beta$. Considérons les deux fonctions suivantes,

$$u(x) = (1 + 2\delta)^v \, \Phi\left(x|\, \nu, (1 + \delta)A\right) \tag{5.14}$$

et

$$l(x) = (1 + 2\delta)^{-v} \, \Phi\left(x|\, \nu, (1 + \delta)^{-1}A\right) \tag{5.15}$$

où les vecteurs ν et la quantité δ seront précisés plus loin de façon à ce que $[l, u]$ soit un ε-crochet de $\mathcal{F}_{(v)}$ qui contienne f.

La proposition 5.2.19 donnée plus loin dans la section 5.2.8 permet de majorer le quotient de deux densités gaussiennes multidimensionnelles. Le lemme suivant garantit que les hypothèses de la proposition 5.2.19 sont bien vérifiées.

Lemme 5.2.9. *On suppose que $0 < \beta < \frac{\sigma_m^2}{3v}$ et on pose $\delta = \frac{3\beta v}{\sigma_m^2}$. Alors, $(1 + \delta)A - \Sigma$ et $\Sigma - (1 + \delta)^{-1}A$ sont deux matrices symétriques définies positives. De plus, pour tout $x \in \mathbb{R}^v$,*

$$x'\{(1 + \delta)A - \Sigma\}x \geq \beta v \|x\|_2^2 \tag{5.16}$$

et

$$x'\{\Sigma - (1 + \delta)^{-1}A\}x \geq \beta v \|x\|_2^2. \tag{5.17}$$

Démonstration. Puisque $|||A-\Sigma||| \leq v\beta$, pour tout $x \neq 0$,

$$\begin{aligned} x'\{(1+\delta)A-\Sigma\}x &= (1+\delta)x'(A-\Sigma)x + \delta x'\Sigma x \\ &\geq -(1+\delta)\,|||A-\Sigma|||\,\|x\|_2^2 + \delta\,\overset{2}{\sigma}_{\mathrm{m}}\,\|x\|_2^2 \\ &\geq \{\delta\,\overset{2}{\sigma}_{\mathrm{m}} - (1+\delta)v\beta\}\|x\|_2^2 \\ &\geq \left(\frac{2}{3}\delta\,\overset{2}{\sigma}_{\mathrm{m}} - v\beta\right)\|x\|_2^2 \end{aligned}$$

car $v\beta \leq \frac{\sigma_{\mathrm{m}}^2}{3}$. Alors, $x'\{(1+\delta)A-\Sigma\}x \geq v\beta\|x\|_2^2 > 0$ d'après la définition de δ. De façon similaire, nous avons

$$\begin{aligned} x'\{\Sigma-(1+\delta)^{-1}A\}x &= (1+\delta)^{-1}x'(\Sigma-A)x + \{1-(1+\delta)^{-1}\}x'\Sigma x \\ &\geq \left(\frac{\delta\,\sigma_{\mathrm{m}}^2 - v\beta}{1+\delta}\right)\|x\|_2^2 \\ &\geq \frac{2v\beta}{1+\delta}\|x\|_2^2 \\ &\geq v\beta\|x\|_2^2 > 0. \end{aligned}$$

□

Lemme 5.2.10. *On suppose que* $\beta < \frac{\sigma_m^2}{3v}$ *et on pose* $\delta = \frac{3\beta v}{\sigma_m^2}$. *Alors,*

$$\frac{f(x)}{u(x)} \leq (1+2\delta)^{-\frac{v}{2}} \exp\left(\frac{\|\mu-\nu\|_2^2}{2\beta v}\right)$$

et

$$\frac{l(x)}{f(x)} \leq (1+2\delta)^{-\frac{v}{2}} \exp\left(\frac{\|\mu-\nu\|_2^2}{2\beta v}\right).$$

Démonstration. Le lemme 5.2.9 garantie que sous ces hypothèses, $(1+\delta)A-\Sigma$ est une matrice symétrique définie positive. Donc, d'après la proposition 5.2.19, nous avons

$$\frac{f(x)}{u(x)} \leq (1+2\delta)^{-1}\sqrt{\frac{|(1+\delta)A|}{|\Sigma|}}\ \exp\left[\frac{1}{2}(\mu-\nu)'\{(1+\delta)A-\Sigma\}^{-1}(\mu-\nu)\right].$$

De plus, l'inégalité (5.16) implique que $|||\{(1+\delta)A-\Sigma\}^{-1}||| = \{\inf\lambda\}^{-1} \leq (\beta v)^{-1}$ où la borne inférieure porte sur toutes les valeurs propres de $(1+\delta)A-\Sigma$. Puisque

$$(\mu-\nu)'\{(1+\delta)A-\Sigma\}^{-1}(\mu-\nu) \leq |||\{(1+\delta)A-\Sigma\}^{-1}|||\,\|\mu-\nu\|_2^2,$$

on en déduit que

$$(\mu-\nu)'\{(1+\delta)A-\Sigma\}^{-1}(\mu-\nu) \leq \frac{\|\mu-\nu\|_2^2}{v\beta}.$$

Ensuite, par le lemme 5.2.8,

$$\begin{aligned}
\frac{|(1+\delta)A|}{|\Sigma|} &= (1+\delta)^v \prod_{q=1}^{v} \frac{\tau_q^2}{\sigma_q^2} \\
&\leq (1+\delta)^v \prod_{q=1}^{v} \left(1+\frac{\beta v}{\sigma_q^2}\right) \\
&\leq (1+\delta)^v \left(1+\frac{\beta v}{\sigma_{\mathrm{m}}^2}\right)^v \\
&\leq (1+2\delta)^v,
\end{aligned}$$

d'où

$$\frac{f(x)}{u(x)} \leq (1+2\delta)^{-\frac{v}{2}} \exp\left(\frac{\|\mu-\nu\|_2^2}{2\beta v}\right).$$

On montre de façon similaire en utilisant la proposition 5.2.19, l'inégalité (5.17) et le lemme 5.2.8 que

$$\frac{l(x)}{f(x)} \leq (1+2\delta)^{-\frac{v}{2}} \ \exp\left(\frac{\|\mu-\nu\|_2^2}{2\beta v}\right).$$

□

La proposition suivante achève la construction de la famille des ε-crochets permettant de recouvrir $\mathcal{F}_{(v)}$.

Proposition 5.2.11. *Pour tout $\varepsilon \in]0,1]$, on pose $\delta = \frac{\varepsilon}{\sqrt{3}\,v}$ et $\beta = \frac{\sigma_m^2\,\varepsilon}{3\sqrt{3}v^2}$. L'ensemble de crochets suivant*

$$\left\{ [l,u]\,; \begin{array}{l} u(x) = (1+2\delta)^v\,\Phi\left(x|\,\nu,(1+\delta)A\right) \\ l(x) = (1+2\delta)^{-v}\,\Phi\left(x|\,\nu,(1+\delta)^{-1}A\right) \end{array} \ ;\ A \in \mathcal{R}(\beta),\ \nu \in \mathcal{X}(\varepsilon,a,\sigma_m^2,v) \right\}$$

où

$$\mathcal{X}(\varepsilon,a,\sigma_m^2,v) = \left\{ \nu = (\nu_1,\ldots,\nu_v);\ \nu_q = \frac{\sigma_m\,\varepsilon}{3v}\,s_q;\ s_q \in \mathbb{Z} \cap \left[-\left\lfloor \frac{3\,a\,v}{\sigma_m\,\varepsilon} \right\rfloor, \left\lfloor \frac{3\,a\,v}{\sigma_m\,\varepsilon} \right\rfloor\right] \right\},$$

forme un recouvrement par des ε-crochets de l'ensemble $\mathcal{F}_{(v)}$.

Démonstration. Soit $f(x) = \Phi(x|\mu,\Sigma)$ une fonction de $\mathcal{F}_{(v)}$ avec $\mu \in [-a,a]^v$ et $\Sigma \in \mathcal{D}_{(v)}^{+}(\sigma_{\mathrm{m}}^2,\sigma_{\mathrm{M}}^2)$. Il existe une matrice A dans $\mathcal{R}(\beta)$ telle que $\|\Sigma - A\|_\infty \leq \beta$ et ν dans $\mathcal{X}(\varepsilon,a,\sigma_{\mathrm{m}}^2,v)$ qui satisfait pour tout q de $\{1,\ldots,v\}$ l'inégalité $|\mu_q - \nu_q| \leq \frac{\sigma_{\mathrm{m}}\varepsilon}{3v}$. Considérons les deux fonctions l et u définies respectivement comme en (5.14) et (5.15). Puisque $\|\mu-\nu\|_2^2 \leq \frac{\sigma_{\mathrm{m}}^2\,\varepsilon^2}{9v}$, en utilisant le lemme 5.2.10, nous obtenons

$$\frac{f(x)}{u(x)} \leq (1+2\delta)^{-\frac{v}{2}} \exp\left(\frac{\sqrt{3}\,\varepsilon}{6}\right).$$

On note ensuite que pour x dans $[0,2]$, $\ln(1+x) \geq \frac{x}{2}$, ce qui donne

$$\begin{aligned}
\ln\left\{\frac{f(x)}{u(x)}\right\} &\leq -\frac{v}{2}\ln\left(1+\frac{2\,\varepsilon}{\sqrt{3}v}\right)+\frac{\sqrt{3}\,\varepsilon}{6} \\
&\leq -\frac{v}{2}\frac{\varepsilon}{\sqrt{3}v}+\frac{\varepsilon}{2\sqrt{3}} \\
&\leq 0.
\end{aligned}$$

On montre de la même façon que $\ln\left\{\frac{l(x)}{f(x)}\right\} \leq 0$, et donc $l(x) \leq f(x) \leq u(x)$ pour $x \in \mathbb{R}^v$. Il reste à évaluer la longueur du crochet $[l,u]$ pour la distance de Hellinger. En utilisant la proposition 5.2.20, nous obtenons que

$$\begin{aligned}
d_H^2(l,u) &= (1+2\delta)^v+(1+2\delta)^{-v}-\left\{2-d_H^2(\Phi(\cdot|\nu,(1+\delta)A),\Phi(\cdot|\nu,(1+\delta)^{-1}A))\right\} \\
&= 2\left(\mathrm{ch}\{v\ln(1+2\delta)\}-1+1-[\mathrm{ch}\{\ln(1+\delta)\}]^{-\frac{v}{2}}\right) \\
&\leq 2\left(\mathrm{sh}(1)v^2\delta^2+\frac{1}{4}v^2\delta^2\right) \\
&\leq 3v^2\delta^2=\varepsilon^2.
\end{aligned}$$

□

Fin de la preuve de la proposition 5.2.7

La famille des ε-crochets qui recouvre $\mathcal{F}_{(v)}$ est définie à partir des ensembles $\mathcal{R}(\beta)$ et $\mathcal{X}(\varepsilon,a,\sigma_{\mathrm{m}}^2,v)$. Nous pouvons donc majorer le cardinal du nombre d'entropie de $\mathcal{F}_{(v)}$ de la façon suivante,

$$\begin{aligned}
\mathcal{N}_{[.]}\left(\varepsilon,\mathcal{F}_{(v)},d_H\right) &\leq \mathrm{card}\{\mathcal{R}(\beta)\}\times\mathrm{card}\{\mathcal{X}(\varepsilon,a,\sigma_{\mathrm{m}}^2,v)\} \\
&\leq \left(\frac{2\sigma_{\mathrm{M}}^2}{\beta}\right)^{\frac{v(v+1)}{2}}\left(\frac{6\,av}{\sigma_{\mathrm{m}}\,\varepsilon}\right)^v \\
&\leq \left(\frac{6\sqrt{3}\,\sigma_{\mathrm{M}}^2v^2}{\sigma_{\mathrm{m}}^2\varepsilon}\right)^{\frac{v(v+1)}{2}}\left(\frac{6\,av}{\sigma_{\mathrm{m}}\,\varepsilon}\right)^v.
\end{aligned}$$

Nous en déduisons que

$$\begin{aligned}
\mathrm{H}_{[.]}\left(\varepsilon,\mathcal{F}_{(v)},d_H\right) &\leq \frac{v(v+1)}{2}\ln\left(\frac{6\sqrt{3}\,\sigma_{\mathrm{M}}^2}{\sigma_{\mathrm{m}}^2}\right)+v\ln\left(\frac{6\,a}{\sigma_{\mathrm{m}}}\right) \\
&+ 2\left\{\frac{v(v+1)}{2}+v\right\}\ln(v)+\left\{\frac{v(v+1)}{2}+v\right\}\ln\left(\frac{1}{\varepsilon}\right).
\end{aligned}$$

Fin de la preuve de la proposition 5.2.1 pour la collection $\mathcal{M}[L_kC_k]$

Pour compléter la preuve de la proposition 5.2.1, nous procédons comme pour la collection $\mathcal{M}[L_kB_k]$ en notant tout d'abord que

$$\mathrm{H}_{[.]}(\varepsilon,\mathcal{F}_{(v)},d_H)=\mathrm{H}_{[.]}(\varepsilon,\mathcal{L}_{(K,v)},d_H).$$

D'après la proposition 5.2.5,

$$\mathrm{H}_{[.]}(\varepsilon, \mathcal{L}_{(K,v)}, d_H) \leq \mathrm{H}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{T}_{K-1}, d_H\right) + K\,\mathrm{H}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{F}_{(v)}, d_H\right).$$

La proposition 5.2.5 et le lemme 5.2.4 nous donnent que

$$\begin{aligned}\mathrm{H}_{[.]}(\varepsilon, \mathcal{L}_{(K,v)}, d_H) \leq \quad & \ln K + \frac{K}{2}\ln(2\pi e) + (K-1)\ln\frac{3}{\varepsilon} \\ + \quad & \frac{Kv(v+1)}{2}\ln\left(\frac{6\sqrt{3}\sigma_{\mathrm{M}}^2}{\sigma_{\mathrm{m}}^2}\right) + Kv\ln\left(\frac{6\,a}{\sigma_{\mathrm{m}}}\right) \\ + \quad & 2K\left\{\frac{v(v+1)}{2} + v\right\}\ln(Q) + K\left\{\frac{v(v+1)}{2} + v\right\}\ln\left(\frac{3}{\varepsilon}\right).\end{aligned}$$

La dimension du modèle $S_{(K,\mathbf{v})}$ a pour valeur $D(K,v) = Kv(1+\frac{v+1}{2}) + K - 1$, il est alors clair que

$$\mathrm{H}_{[.]}(\varepsilon, \mathcal{L}_{(K,v)}, d_H) \leq (A_1 \ln v + \ln\frac{1}{\varepsilon})D(K,v)$$

où la constante A_1 ne dépend que de σ_{m}, σ_{M} et a, ce qui achève la preuve de la proposition 5.2.1 pour la collection $\mathcal{M}[L_kC_k]$.

5.2.7 Entropies pour les modèles de la collection $\mathcal{M}[LB_k]$

Pour cette dernière collection de modèles, il n'est pas possible de se ramener à des sous familles $\mathcal{F}_{(v)}$ comme pour les collections de modèles $\mathcal{M}[L_kB_k]$ et $\mathcal{M}[L_kC_k]$. En effet, pour une densité de mélange t d'un modèle $\mathcal{L}_{(K,v)}$, les matrices de covariance intervenant dans le mélange t sont cette fois toutes de même volume, et il n'est plus possible de décomposer l'ensemble $\mathcal{L}_{(K,v)}$ comme dans la proposition 5.2.5.

Nous considérons cette fois les ensembles $\mathcal{F}_{(K,v)}$ suivants,

$$\mathcal{F}_{(K,v)} = \left\{ (\Phi(\cdot|\mu_1, \lambda\Sigma_1), \ldots, \Phi(\cdot|\mu_K, \lambda\Sigma_K)) \ ; \ \begin{array}{l} \sigma_{\mathrm{M}}^{-2} \leq \lambda \leq \sigma_{\mathrm{M}}^2,\ k = 1\ldots K \\ \mu_k \in [-a,a]^v,\ \Sigma_k \in \Delta^1_{(v)}(\sigma_{\mathrm{m}}^2, \sigma_{\mathrm{M}}^2) \end{array} \right\}.$$

Soient deux fonctions l et u de $\mathbb{R}^v$ dans $\mathbb{R}^K$ telles que $l_k \leq u_k$ pour $k = 1\ldots K$, avec les notations $l(x) = (l_1(x), \ldots, l_K(x))$ et $u(x) = (u_1(x), \ldots, u_K(x))$. L'ensemble $[l,u]$ est composé de toutes les fonctions $f = (f_1, \ldots, f_K)$ de $\mathbb{R}^v$ dans $\mathbb{R}^K$ telles que $l_k \leq f_k \leq u_k$ pour $k = 1\ldots K$. Dans ce contexte, on dit que l'ensemble $[l,u]$ est un ε-crochet si $d_H(l_k, u_k) \leq \varepsilon$ pour tout $k \in \{1, \ldots, K\}$. Nous définissons alors de façon naturel le nombre d'entropie de $\mathcal{F}_{(K,v)}$ comme le cardinal minimal des familles de ε-crochets de la forme $[l,u]$ recouvrant l'espace $\mathcal{F}_{(K,v)}$.

Pour démontrer la proposition 5.2.1 dans le cas de la collection $\mathcal{M}[LB_k]$, nous utilisons la propriété suivante qui est une version modifiée de la proposition 5.2.5.

Proposition 5.2.12. *Soient $K \geq 2$ et $v \geq 1$ des entiers. On suppose que $\{[a_i, b_i]\ ,\ i \in I\}$ est une famille de $\frac{\varepsilon}{3}$-crochets qui forme un recouvrement du simplexe τ_{K-1}, où $a_i := (a_{1i}, \ldots, a_{Ki})$ et $b_i := (b_{1i}, \ldots, b_{Ki})$. On suppose de plus qu'il existe une famille de $\frac{\varepsilon}{3}$-crochets $\{[l_j, u_j]\ ,\ j \in J\}$*

formant un recouvrement de $\mathcal{F}_{(K,v)}$, avec $l_j := (l_{1j}, \ldots, l_{Kj})$ et $u_j := (u_{1j}, \ldots, u_{Kj})$. Pour tout $(i,j) \in I \times J$, soient $L_{ij} = \sum_{k=1}^{K} a_{ki} l_{kj}$ et $U_{ij} = \sum_{k=1}^{K} b_{ki} u_{kj}$. Alors, la famille $\{[L_{ij}, U_{ij}]\,,\ (i,j) \in I \times J\}$ forme un recouvrement de l'ensemble $\mathcal{L}_{(K,v)}$ par des ε-crochets.

La démonstration de ce résultat est exactement identique à celle de la proposition 5.2.5 qui est démontrée dans Genovese et Wasserman (2000). La majoration de l'entropie métrique de $\mathcal{L}_{(K,v)}$ (ou de $S_{(K,v)}$) se ramène donc à une majoration de l'entropie métrique de $\mathcal{F}_{(K,v)}$.

Proposition 5.2.13. *Il existe une constante A_2 telle que pour tout $\varepsilon \in]0,1]$,*

$$N_{[.]}\left(\mathcal{F}_{(K,v)}\right) \leq \left(A_2 \frac{v}{\varepsilon}\right)^{K(2v-1)+1},$$

et la constante A_2 ne dépend que de σ_m, σ_M et a.

Nous allons démontrer ce résultat à l'aide de plusieurs lemmes techniques successifs. La preuve de la proposition 5.2.1 est donnée à la suite, à la fin de cette section. Notons toute d'abord que si $v = 1$, les entropies des modèles $S_{(K,1)}$ peuvent se calculer facilement en imitant la méthode utilisée pour la collection $\mathcal{M}[L_k B_k]$. Nous pouvons donc supposer dans la suite que $v \geq 2$.

Soit $\varepsilon \in]0,1]$, et supposons fixés $K \geq 2$ et $v \geq 2$. On pose $\delta = \frac{\varepsilon}{3v}$. Pour $j = 1 \ldots r$, soit

$$b_j^2 = (1+\delta)^{1-\frac{j}{2}}\, {\sigma_{\mathrm{M}}}^2$$

où $r = \left\lceil 2\, \frac{\ln\left\{\frac{{\sigma_{\mathrm{M}}}^2}{{\sigma_{\mathrm{m}}}^2}(1+\delta)\right\}}{\ln(1+\delta)} \right\rceil$, de façon à ce que $b_r^2 \leq {\sigma_{\mathrm{m}}}^2$ et $b_2^2 = {\sigma_{\mathrm{M}}}^2$. De même, pour $z = 0 \ldots r'$, soit

$$\lambda_z = (1+\delta)^{-z} {\sigma_{\mathrm{M}}}^2,$$

où $r' = \left\lceil \frac{\ln\left\{\frac{{\sigma_{\mathrm{M}}}^2}{{\sigma_{\mathrm{m}}}^2}\right\}}{\ln(1+\delta)} \right\rceil$, de façon à ce que $\lambda_{r'} \leq {\sigma_{\mathrm{m}}}^2$ et $\lambda_0 = {\sigma_{\mathrm{M}}}^2$. Pour tout vecteur $J = (j(1), \ldots, j(v-1)) \in \{2, \ldots, r\}^{v-1}$, les matrices B_J^l et B_J^u sont définies par

$$B_J^l = \mathrm{diag}\left(b_{j(1)+1}^2, \ldots, b_{j(v-1)+1}^2, {\sigma_{\mathrm{M}}}^{-2v+2}(1+\delta)^{\frac{S_J}{2}-(v-1)}\right),$$

et

$$B_J^u = \mathrm{diag}\left(b_{j(1)}^2, \ldots, b_{j(v-1)}^2, {\sigma_{\mathrm{M}}}^{-2v+2}(1+\delta)^{\frac{S_J}{2}-\frac{v-1}{2}}\right),$$

avec $S_J = \sum_{q=1}^{v-1} j(q)$. Dans la suite, pour $q \in \{1, \ldots, v\}$, les notations $B_{J,q}^l$ et $B_{J,q}^u$ désignent les coefficients (q,q) des matrices B_J^l et B_J^u.

Lemme 5.2.14. *Soit $\Sigma \in \Delta_{(v)}^1({\sigma_M}^{-2}, {\sigma_M}^2)$, et $\lambda \in [{\sigma_m}^2, {\sigma_M}^2]$. Alors, il existe $J \in \{2, \ldots, r\}^{v-1}$ et $z \in \{0, \ldots, r'\}$ tels que pour tout $q \in \{1, \ldots, v\}$,*

$$\lambda_{z+1}\, B_{J,q}^l \leq \lambda\, \Sigma_{qq} \leq \lambda_z\, B_{J,q}^u.$$

Démonstration. Soit z l'unique entier dans $\{0, \ldots, r'\}$ tel que $\lambda_{z+1} < \lambda \leq \lambda_z$. Par définition des b_j, on peut trouver un vecteur J tel que $B_{J,q}^l \leq \Sigma_{qq} \leq B_{J,q}^u$ pour $q = 1, \ldots, v-1$; ce qui

donne l'encadrement voulu pour les $v-1$ premières coordonnées. Puisque $\prod_{q=1}^{v} \Sigma_{qq} = 1$, on a donc $\Sigma_{vv} = \prod_{q=1}^{v-1} \Sigma_{qq}^{-1}$. On en déduit que

$$\prod_{q=1}^{v-1} \left\{B_{J,q}^{u}\right\}^{-1} \leq \Sigma_{vv} \leq \prod_{q=1}^{v-1} \left\{B_{J,q}^{l}\right\}^{-1},$$

ce qui donne l'encadrement annoncé pour la dernière coordonnée. □

Les matrices de covariance ainsi associées au couple (z, J) nous sont utiles pour définir les fonctions extrémités des crochets. Il nous faut aussi définir des paramètres de centrage. Pour un couple (z, J) fixé, on considère le maillage régulier de paramètres de centrage $\nu^{(zJ)} = \left(\nu_1^{(zJ)}, \ldots, \nu_v^{(zJ)}\right) \in [-a, a]^v$ tels que pour $q \in \{1, \ldots, v-1\}$,

$$\nu_q^{(zJ)} = (1+\delta)^{-\frac{j(q)+1}{4} - \frac{z}{2}} \sigma_{\mathrm{M}}^{2} \sqrt{c_1}\, \delta s_q,$$

avec $s_q \in \{-N_q, \ldots, N_q\}$ où $N_q = \left\lfloor \frac{a(1+\delta)^{\frac{j(q)+1}{4} + \frac{z}{2}}}{\sqrt{c_1}\, \sigma_{\mathrm{M}}{}^{2} \delta} \right\rfloor$, et

$$\nu_v^{(zJ)} = (1+\delta)^{\frac{S_J}{4} - \frac{v+z}{2}} \sigma_{\mathrm{M}}^{2-v} \sqrt{c_1}\, \delta s_v,$$

avec $s_v \in \{-N_v, \ldots, N_v\}$ où $N_v = \left\lfloor \frac{a(1+\delta)^{\frac{v+z}{2} - \frac{S_J}{4}}}{\sqrt{c_1}\, \sigma_{\mathrm{M}}{}^{2-v} \delta} \right\rfloor$, et $c_1 := \frac{1-2^{-\frac{1}{4}}}{2}$. Pour un couple (z, J) donné, ceci garantie que pour tout $\mu \in [-a, a]$, il existe un vecteur $\nu^{(zJ)}$ du maillage tel que

$$\left\{\sum_{q=1}^{v-1} \left(\nu_q^{(zJ)} - \mu_q\right)^2 (1+\delta)^{\frac{j(q)+1}{2} + z} \sigma_{\mathrm{M}}^{-4}\right\} + \left(\nu_v^{(zJ)} - \mu_v\right)^2 (1+\delta)^{-\frac{S_J}{2} + v + z} \sigma_{\mathrm{M}}^{2v-4} \leq c_1\, v\, \delta^2. \tag{5.18}$$

Pour un couple (z, J), et un paramètre $\nu^{(zJ)}$ de la famille ci-dessus, on considère finalement les deux fonctions suivantes,

$$\begin{cases} l(x) = (1+\delta)^{-2v}\, \Phi\left(x \mid \nu^{(zJ)}, (1+\delta)^{-\frac{1}{4}}\, \lambda_{z+1}\, B_J^{l}\right) \\ u(x) = (1+\delta)^{2v}\, \Phi\left(x \mid \nu^{(zJ)}, (1+\delta)\, \lambda_z\, B_J^{u}\right). \end{cases} \tag{5.19}$$

Soit $\lambda \in [\sigma_{\mathrm{m}}^{2}, \sigma_{\mathrm{M}}^{2}]$, $\Sigma \in \Delta_{(v)}^{1}(\sigma_{\mathrm{m}}^{2}, \sigma_{\mathrm{M}}^{2})$, $\mu_k \in [-a, a]^v$, et on pose $\Phi := \Phi(\cdot \mid \mu, \lambda\Sigma)$. Il existe λ_z et J tels que pour tout $q \in \{1, \ldots, v\}$,

$$\lambda_{z+1}\, B_{J,q}^{l} \leq \lambda\, \Sigma_{qq} \leq \lambda_z\, B_{J,q}^{u}.$$

Il existe aussi un vecteur $\nu^{(zJ)}$ du maillage, associé au couple (z, J), tel que la contrainte (5.18) soit vérifiée pour le vecteur μ. Pour z, J et ce vecteur $\nu^{(zJ)}$, on considère u et l les deux fonctions définies comme en (5.19).

Lemme 5.2.15. *Sous les hypothèses précédentes, pour tout $x \in \mathbb{R}^d$, $l(x) \leq \Phi(x) \leq u(x)$.*

Démonstration. En appliquant le lemme 5.2.19 qui permet de majorer des quotients de den-

sités gaussiennes, nous obtenons

$$\begin{aligned}\frac{\Phi(x)}{u(x)} &\leq (1+\delta)^{-2v}\sqrt{\frac{|(1+\delta)\lambda_z B_J^u|}{|\lambda\Sigma|}}\exp\left[\frac{1}{2}(\nu^{(zJ)}-\mu)'\left\{(1+\delta)\lambda_z B_J^u-\lambda\Sigma\right\}^{-1}(\nu^{(zJ)}-\mu)\right]\\ &\leq (1+\delta)^{-\frac{v+1}{2}}\exp\left\{\frac{1}{2\delta}(\nu^{(zJ)}-\mu)'(\lambda_z B_J^u)^{-1}(\nu^{(zJ)}-\mu)\right\}.\end{aligned}$$

Pour que $\Phi \leq u$, il suffit donc que $(\nu^{(zJ)}-\mu)'(\lambda_z B_J^u)^{-1}(\nu^{(zJ)}-\mu) \leq 2\delta\frac{v+1}{2}\ln(1+\delta)$, ou encore que

$$(\nu^{(zJ)}-\mu)'(\lambda_z B_J^u)^{-1}(\nu^{(zJ)}-\mu) \leq \frac{\delta^2}{2}(v+1). \tag{5.20}$$

En appliquant de nouveau le lemme 5.2.19 à Φ et l, nous obtenons cette fois

$$\begin{aligned}\frac{l(x)}{\Phi(x)} &\leq (1+\delta)^{-2v}\sqrt{\frac{|\lambda\Sigma|}{|(1+\delta)^{-\frac{1}{4}}\lambda_{z+1}B_J^l|}}\exp\left[\frac{1}{2}(\nu^{(zJ)}-\mu)'\left\{\lambda\Sigma-(1+\delta)^{-\frac{1}{4}}\lambda_{z+1}B_J^l\right\}^{-1}(\nu^{(zJ)}-\mu)\right]\\ &\leq (1+\delta)^{-\frac{7}{8}v-\frac{1}{2}}\exp\left\{\frac{1}{2\delta(1-2^{-\frac{1}{4}})}(\nu^{(zJ)}-\mu)'(\lambda_{z+1}B_J^l)^{-1}(\nu^{(zJ)}-\mu)\right\},\end{aligned}$$

en utilisant la concavité de la fonction $\delta \mapsto 1-(1+\delta)^{-\frac{1}{4}}$. Pour que $l \leq \Phi$, il suffit donc que $(\nu^{(zJ)}-\mu)'(\lambda_{z+1}B_J^l)^{-1}(\nu^{(zJ)}-\mu) \leq (\frac{7}{8}v+\frac{1}{2})\ln(1+\delta)(2\delta)(1-2^{-\frac{1}{4}})$, ou encore que

$$(\nu^{(zJ)}-\mu)'(\lambda_{z+1}B_J^l)^{-1}(\nu^{(zJ)}-\mu) \leq \delta^2(1-2^{-\frac{1}{4}})\frac{v}{2}. \tag{5.21}$$

On vérifie finalement que la condition (5.18) implique les inégalités (5.20) et (5.21). □

Lemme 5.2.16. *Sous les hypothèses précédentes, $d_H(u,l) \leq \varepsilon$.*

Démonstration. D'après le lemme 5.2.20 donné dans la section 5.2.8, nous avons

$$\begin{aligned}d_H^2(l,u) &= (1+\delta)^{-2v}+(1+\delta)^{2v}-\left\{\prod_{q=1}^{v-1}2\frac{(1+\delta)^{-\frac{1}{8}}\sqrt{\lambda_{z+1}}b_{j(q)+1}(1+\delta)^{\frac{1}{2}}\sqrt{\lambda_z}b_{j(q)}}{(1+\delta)^{-\frac{1}{4}}\lambda_{z+1}{b_{j(q)+1}}^2+(1+\delta)\lambda_z{b_{j(q)}}^2}\right\}^{\frac{1}{2}}\\ &\quad\times\left\{2\frac{(1+\delta)^{-\frac{1}{8}}\sqrt{\lambda_{z+1}}B_{J,v}^l(1+\delta)^{\frac{1}{2}}\sqrt{\lambda_z}B_{J,v}^u}{(1+\delta)^{-\frac{1}{4}}\lambda_{z+1}\left[B_{J,v}^l\right]^2+(1+\delta)\lambda_z\left[B_{J,v}^u\right]^2}\right\}^{\frac{1}{2}}\\ &= (1+\delta)^{-2v}+(1+\delta)^{2v}-2\left\{\frac{2}{(1+\delta)^{-\frac{11}{8}}+(1+\delta)^{\frac{11}{8}}}\right\}^{\frac{v-1}{2}}\left\{\frac{2}{(1+\delta)^{-\frac{5+v}{4}}+(1+\delta)^{\frac{5+v}{4}}}\right\}^{\frac{1}{2}}\\ &= 2\cosh(2v\ln(1+\delta))-2\left\{\cosh\left(\frac{11}{8}\ln(1+\delta)\right)\right\}^{-\frac{v-1}{2}}\left\{\cosh\left(\frac{5+v}{4}\ln(1+\delta)\right)\right\}^{-\frac{1}{2}}\\ &= [2\cosh(2v\ln(1+\delta))-2]+\left[2-2\left\{\cosh\left(\frac{11}{8}\ln(1+\delta)\right)\right\}^{-\frac{v-1}{2}}\right]\\ &\quad+2\left\{\cosh\left(\frac{11}{8}\ln(1+\delta)\right)\right\}^{-\frac{v-1}{2}}\left[1-\left\{\cosh\left(\frac{5+v}{4}\ln(1+\delta)\right)\right\}^{-\frac{1}{2}}\right]\end{aligned}$$

Et donc,

$$\begin{aligned} d_H^2(l,u) &\leq 4\sinh(1)v^2\delta^2 + 2\frac{v-1}{2}\frac{11}{8}\delta^2 + 2\frac{5+v}{4}\frac{1}{2}\delta^2 \\ &\leq 9v^2\delta^2 \\ &\leq \varepsilon^2. \end{aligned}$$

□

Nous sommes maintenant en position de construire des crochets sur l'ensemble $\mathcal{F}_{(K,v)}$. Soit $(\Phi_1,\ldots,\Phi_K) \in \mathcal{F}_{(K,v)}$, avec $\Phi_k = \Phi(\cdot\,|\,\mu_k, \lambda\Sigma_k)$ pour $k = 1,\ldots,K$. Soient λ_z, $J_1,\ldots,J_k$ tels que pour tout $k = 1 \in \{1,\ldots,K\}$ et tout $q \in \{1,\ldots,v\}$,

$$\lambda_{z+1}\, B^l_{J_k,q} \leq \lambda\, \Sigma_{k,qq} \leq \lambda_z\, B^u_{J_k,q},$$

où $\Sigma_{k,qq}$ désigne le coefficient (q,q) de la matrice Σ_k. Pour tout k, il existe un vecteur $\nu^{(zJ_k)}$ du maillage, associé au couple (z, J_k), tel que la contrainte (5.18) soit vérifiée pour le vecteur μ_k. Pour z, J_k et ce vecteur $\nu^{(zJ_k)}$, on nomme u_k et l_k les deux fonctions définies comme en (5.19). Soient les fonctions $l := (l_1,\ldots,l_K)$ et $u := (u_1,\ldots,u_K)$ de $\mathbb{R}^v$ dans $\mathbb{R}^K$, l'ensemble de tous les crochets $[l,u]$ que l'on peut définir de cette façon est noté $\mathcal{R}(\varepsilon,K,v)$.

Lemme 5.2.17. *Il existe une constante A_2 qui ne dépend que de σ_m, σ_M et a telle que*

$$\text{card } \mathcal{R}(K,\varepsilon,v) \leq (A_2\frac{v}{\varepsilon})^{K(2v-1)+1}. \tag{5.22}$$

Démonstration. La définition de l'ensemble $\mathcal{R}(K,\varepsilon,v)$ permet de majorer son cardinal de la façon suivante.

$$\text{card } \mathcal{R}(K,\varepsilon,v) \leq \sum_{z=0}^{r'}\left[\sum_J \left\{\prod_{q=1}^{v-1} 1 \vee \frac{2\,a\,(1+\delta)^{\frac{j(q)+1}{4}+\frac{z}{2}}}{\sqrt{c_1}\sigma_{\mathrm{M}}^2\,\delta}\right\}\left\{1 \vee \frac{2\,a\,(1+\delta)^{\frac{v+z}{2}}}{\sqrt{c_1}\sigma_{\mathrm{M}}^{2-v}(1+\delta)^{\frac{S_J}{4}}\,\delta}\right\}\right]^K$$

Si $2a \geq \sqrt{c_1}(\sigma_{\mathrm{M}}^2 \vee \sigma_{\mathrm{M}}^{2-v})$ alors,

$$\begin{aligned} \text{card } \mathcal{R}(K,\varepsilon,v) &\leq r'\left[(r-1)^{v-1}\left\{\frac{2\,a}{\sqrt{c_1}\sigma_{\mathrm{M}}^2}\frac{(1+\delta)^{\frac{r+1}{4}+\frac{r'}{2}}}{\delta}\right\}^{v-1}\left\{\frac{2\,a}{\sqrt{c_1}\sigma_{\mathrm{M}}^{2-v}}\frac{(1+\delta)^{\frac{v+r'}{2}}}{\delta}\right\}\right]^K \\ &\leq \frac{r'(r-1)^{K(v-1)}}{\delta^{vK}}(1+\delta)^{\frac{Kv(r'-1)}{2}+\frac{K(v-1)(r-1)}{4}+Kv+\frac{K(v-1)}{2}}\left(\frac{2\,a}{\sqrt{c_1}\,\sigma_{\mathrm{M}}}\right)^{Kv} \end{aligned}$$

On note que $\delta \leq 1$. Ensuite, les définitions de r et r' montrent d'une part que $r-1 \leq \frac{4(1+\delta)\,\sigma_{\mathrm{M}}^2}{\delta\sigma_{\mathrm{m}}^2}$ et $r' \leq \frac{2\,\sigma_{\mathrm{M}}^2}{\delta\sigma_{\mathrm{m}}^2}$, et d'autre part que $(1+\delta)^{r'} \leq 4\left(\frac{\sigma_{\mathrm{m}}}{\sigma_{\mathrm{M}}}\right)^4$ et $(1+\delta)^r \leq 8\left(\frac{\sigma_{\mathrm{m}}}{\sigma_{\mathrm{M}}}\right)^4$. On peut donc trouver une constante c_2 qui ne dépend que de σ_{M}, σ_{m} et a telle que

$$\begin{aligned} \text{card } \mathcal{R}(K,\varepsilon,v) &\leq \frac{{c_2}^{Kv}}{\delta^{K(2v-1)+1}} \\ &\leq (A_2\frac{v}{\varepsilon})^{K(2v-1)+1} \end{aligned}$$

avec A_2 qui ne dépend que de σ_{M}, σ_{m} et a. On note enfin que si $2a \leq \sqrt{c_1}(\sigma_{\mathrm{M}}^2 \vee \sigma_{\mathrm{M}}^{2-v})$, la majoration (5.22) est encore vérifiée quitte à changer de constante c_2. □

Les lemmes 5.2.15 et 5.2.16 montrent que la famille de crochets $\mathcal{R}(K, \varepsilon, v)$ forme un ε-recouvrement de $\mathcal{F}_{(K,v)}$. Le lemme 5.2.17 nous donne finalement la majoration du nombre d'entropie annoncée dans la proposition 5.2.13 en utilisant que $v \leq Q$.

Fin de la preuve de la proposition 5.2.1 pour la collection $\mathcal{M}[LB_k]$
D'après la proposition 5.2.12,

$$\begin{aligned} \mathrm{H}_{[.]}\left(\varepsilon, \mathcal{S}_{(K,v)}, d_H\right) &= \mathrm{H}_{[.]}\left(\varepsilon, \mathcal{L}_{(K,v)}, d_H\right) \\ &\leq \mathrm{H}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{T}_{K-1}, d_H\right) + \mathrm{H}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{F}_{(K,v)}, d_H\right). \end{aligned}$$

On utilise ensuite le lemme 5.2.4 et la proposition 5.2.13, il vient

$$\begin{aligned} \mathrm{H}_{[.]}\left(\varepsilon, \mathcal{L}_{(K,v)}, d_H\right) &\leq \ln K + \frac{K}{2}\ln(2\pi e) + (K-1)\ln\frac{3}{\varepsilon} + [K(2v-1)+1]\left(\ln A_2 + \ln v + \ln\frac{3}{\varepsilon}\right) \\ &\leq D(K,v)\left[\ln\frac{1}{\varepsilon} + A_1 \ln v\right]. \end{aligned}$$

où la constante A_1 ne dépend que de σ_{m}, σ_{M} et a, ce qui achève la preuve de la proposition 5.2.1 pour la collection $\mathcal{M}[LB_k]$.

5.2.8 Résultats annexes

Un lemme utile

Lemme 5.2.18. *Pour tout* $\varepsilon \in]0,1]$, $\int_0^\varepsilon \sqrt{\ln\left(\frac{1}{x}\right)}\,dx \leq \varepsilon\left\{\sqrt{\ln\left(\frac{1}{\varepsilon}\right)} + \sqrt{\pi}\right\}$.

Démonstration. Cette inégalité se déduit d'un intégration par partie et une l'inégalité de concentration suivante pour une variable gaussienne centrée réduite (Massart, 2007, p.19), $P(Z \geq c) \leq e^{-\frac{c^2}{2}}$ pour tout $c > 0$. □

Rapport de deux densités gaussiennes

Proposition 5.2.19. *Soient* $\Phi(\cdot|\mu_1, \Sigma_1)$ *et* $\Phi(\cdot|\mu_2, \Sigma_2)$ *deux densités gaussiennes. Si* $\Sigma_2 - \Sigma_1$ *est une matrice définie positive, alors pour* $x \in \mathbb{R}^Q$,

$$\frac{\Phi(x|\mu_1, \Sigma_1)}{\Phi(x|\mu_2, \Sigma_2)} \leq \sqrt{\frac{|\Sigma_2|}{|\Sigma_1|}}\exp\left\{\frac{1}{2}(\mu_1 - \mu_2)'(\Sigma_2 - \Sigma_1)^{-1}(\mu_1 - \mu_2)\right\}.$$

Démonstration. Le rapport de deux densités gaussiennes est égal à

$$\frac{\Phi(x|\mu_1, \Sigma_1)}{\Phi(x|\mu_2, \Sigma_2)} = \frac{|2\pi\Sigma_1|^{-\frac{1}{2}}}{|2\pi\Sigma_2|^{-\frac{1}{2}}}\exp\left[-\frac{1}{2}\left\{(x-\mu_1)'\Sigma_1^{-1}(x-\mu_1) - (x-\mu_2)'\Sigma_2^{-1}(x-\mu_2)\right\}\right].$$

La matrice $\Sigma_1^{-1} - \Sigma_2^{-1} = \Sigma_1^{-1}(\Sigma_2 - \Sigma_1)\Sigma_2^{-1}$ est inversible car Σ_1, Σ_2 et $\Sigma_2 - \Sigma_1$ sont des matrices définies positives. On pose $\mu^\star = (\Sigma_1^{-1} - \Sigma_2^{-1})^{-1}(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2)$, d'où

$$(x-\mu_1)'\Sigma_1^{-1}(x-\mu_1)-(x-\mu_2)'\Sigma_2^{-1}(x-\mu_2) = (x-\mu^\star)'(\Sigma_1^{-1}-\Sigma_2^{-1})(x-\mu^\star)+(\mu_1-\mu_2)'(\Sigma_2-\Sigma_1)^{-1}(\mu_1-\mu_2)\,.$$

Finalement,

$$\begin{aligned}\frac{\Phi(x|\mu_1,\Sigma_1)}{\Phi(x|\mu_2,\Sigma_2)} &= \sqrt{\frac{|\Sigma_2|}{|\Sigma_1|}}\exp\left[-\frac{1}{2}\left\{(x-\mu^\star)'(\Sigma_1^{-1}-\Sigma_2^{-1})(x-\mu^\star)+(\mu_1-\mu_2)'(\Sigma_2-\Sigma_1)^{-1}(\mu_1-\mu_2)\right\}\right]\\ &\leq \sqrt{\frac{|\Sigma_2|}{|\Sigma_1|}}\exp\left\{\frac{1}{2}(\mu_1-\mu_2)'(\Sigma_2-\Sigma_1)^{-1}(\mu_1-\mu_2)\right\}.\end{aligned}$$

□

Distance de Hellinger entre deux densités gaussiennes

Proposition 5.2.20. *Soient* $\Phi_1 = \Phi(\cdot|\mu_1,\Sigma_1)$ *et* $\Phi_2 = \Phi(\cdot|\mu_2,\Sigma_2)$ *deux densités gaussiennes sur* $\mathbb{R}^Q$*, alors*

$$d_H^2(\Phi_1,\Phi_2) = 2\left[1 - 2^{\frac{Q}{2}}|\Sigma_1\Sigma_2|^{-\frac{1}{4}}|\Sigma_1^{-1}+\Sigma_2^{-1}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{4}(\mu_1-\mu_2)'(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2)\right\}\right].$$

Démonstration. D'après la définition de la distance de Hellinger,

$$d_H^2(\Phi(\cdot|\mu_1,\Sigma_1),\Phi(\cdot|\mu_2,\Sigma_2)) = 2 - 2\int\sqrt{\Phi(x|\mu_1,\Sigma_1)\,\Phi(x|\mu_2,\Sigma_2)}\,dx\,.$$

De plus,

$$\Phi(x|\mu_1,\Sigma_1)\,\Phi(x|\mu_2,\Sigma_2) = (2\pi)^{-Q}\,|\Sigma_1\Sigma_2|^{-\frac{1}{2}}\,\exp\left[-\frac{1}{2}\left\{(x-\mu_1)'\Sigma_1^{-1}(x-\mu_1)+(x-\mu_2)'\Sigma_2^{-1}(x-\mu_2)\right\}\right].$$

On pose $\mu^\star = (\Sigma_1^{-1}+\Sigma_2^{-1})^{-1}(\Sigma_1^{-1}\mu_1+\Sigma_2^{-1}\mu_2)$ et on en déduit que

$$(x-\mu_1)'\Sigma_1^{-1}(x-\mu_1)+(x-\mu_2)'\Sigma_2^{-1}(x-\mu_2) = (x-\mu^\star)'(\Sigma_1^{-1}+\Sigma_2^{-1})(x-\mu^\star)+(\mu_1-\mu_2)'(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2).$$

Finalement,

$$\begin{aligned}d_H^2(\Phi(\cdot|\mu_1,\Sigma_1),\Phi(\cdot|\mu_2,\Sigma_2)) &= 2 - 2(2\pi)^{-\frac{Q}{2}}|\Sigma_1\Sigma_2|^{-\frac{1}{4}}\exp\left\{-\frac{1}{4}(\mu_1-\mu_2)'(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2)\right\}\\ &\quad\times\int\exp\left\{-\frac{1}{4}(x-\mu^\star)'(\Sigma_1^{-1}+\Sigma_2^{-1})(x-\mu^\star)\right\}dx\\ &= 2 - 2(2\pi)^{-\frac{Q}{2}}|\Sigma_1\Sigma_2|^{-\frac{1}{4}}\exp\left\{-\frac{1}{4}(\mu_1-\mu_2)'(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2)\right\}\\ &\quad\times(4\pi)^{\frac{Q}{2}}|\Sigma_1^{-1}+\Sigma_2^{-1}|^{-\frac{1}{2}}\end{aligned}$$

ce qui conclut la preuve. □

5.3 Sélection d'un modèle d'exploration pétrolière

Les modèles présentés au chapitre précédant pour la modélisation du processus d'exploration pétrolier ne sont pas des modèles de type exponentiel comme dans le cas de Castellan. Comme pour les collections de modèles de mélanges gaussiens étudiés auparavant, nous utilisons le théorème 5.1.1 pour obtenir des résultats de sélection de modèles en majorant l'entropie métrique à crochet des modèles considérés.

5.3.1 Notations et hypothèses

Soient $\alpha > 0$ et $t^\star > 0$ fixés, ainsi que x_0 et $x_{\max}$ tels que $1 < x_0 < x_{\max}$. Soit m une partition d'intervalles de $[x_0, x_{\max}]$, l'ensemble $\mathcal{H}_m$ a été défini au chapitre précédent comme l'ensemble des fonctions continues et affines par morceaux sur la partition m. Une fonction de $\mathcal{H}_m$ est entièrement caractérisée par le vecteur A des pentes sur les intervalles de m, et le réel positif $b = h(x_0)$; la notation $h(A, b)$ désigne la fonction de $\mathcal{H}_m$ associée à ces paramètres. Nous supposons de plus que les fonctions h sont bornées par deux niveaux limite $h_{\min}$ et $h_{\max}$, avec $0 < h_{\min} < h_{\max}$ et on note

$$\mathcal{H}_m^b := \{h \in \mathcal{H}_m \,|\, h_{\min} \leq h \leq h_{\max}\}\,.$$

L'ensemble des densités associé à $\mathcal{H}_m^b$ est défini par

$$S_m^b := \left\{ g^\star : (x,t) \mapsto \frac{h(x)\exp\{-h(x)t\}}{\mathrm{P}_{\mathrm{dec}}(\alpha, t^\star, h)} \frac{\alpha x^{-\alpha-1}}{x_0^{-\alpha} - x_{\max}^{-\alpha}} \mathbb{1}_{0 \leq t \leq t^\star} \mathbb{1}_{x_1 \leq x \leq x_{\max}}\,,\ h \in \mathcal{H}_m^b \right\}.$$

Le nombre de paramètres libres du modèle S_m^b est noté D_m, avec $D_m = k + 1$ où k désigne le nombre d'intervalles de la partition m.

Soit $\mathcal{M}_n$ une collection de partitions de $[x_0, x_{\max}]$, où n est la taille de l'échantillon observé. Pour une partition $m \in \mathcal{M}_n$, on note l_m la longueur minimale des intervalles qui la composent. Dans la suite, nous supposerons vérifiée la propriété suivante :

$$(\mathbf{P_2}) : \quad \text{Il existe deux constantes } \gamma > 0 \text{ et } c > 0 \text{ telles que } \inf_{m \in \mathcal{M}_n} l_m \geq c\, n^{-\gamma}.$$

Nous reprenons les deux types de collections de partitions proposées par Castellan (1999) en considérant les deux situations suivantes.

- $(\mathbf{H_1})$ **Complexité polynomiale.** Soit une collection $\mathcal{M}_n$ de partitions de $[x_0, x_{\max}]$ vérifiant (P_2). Nous supposons qu'il existe deux constantes B_1 et β telles que le nombre de partitions de $\mathcal{M}_n$ composées de k intervalles est borné par $B_1 k^\beta$. Cette situation est par exemple celle d'une collection de partitions régulières, c'est à dire telle que la collection ne comporte que les partitions composées d'intervalles de même longueur $(x_{\max} - x_0)/k$, si la partition est de taille k.
- $(\mathbf{H_2})$: **Complexité exponentielle.** Nous considérons une partition m_n de $[x_0, x_{\max}]$ composée de $N_n + 1$ intervalles, avec $N_n + 1 \leq B_2\, n / \ln^2 n$, et telle que m_n vérifie la condition (P_2), c'est à dire $l_{m_n} \geq c\, n^{-\gamma}$. On considère la grille formée des extrémités des intervalles de m_n et soit $\mathcal{M}_n$ un ensemble de partitions de $[x_0, x_{\max}]$ s'appuyant sur

des points de cette grille. Pour un nombre d'intervalles $k+1 \leq N_n$ donné, il existe au plus $C_{N_n}^k$ partitions possibles dans la collection $\mathcal{M}_n$.

5.3.2 Énoncé du théorème

Le résultat suivant correspond à l'application du théorème 5.1.1 pour le cas particulier des modèles S_m^b.

Théorème 5.3.1. *Soit un échantillon $((X_1^\star, D_1^\star), \ldots, (X_n^\star, D_n^\star))$ de variables aléatoires indépendantes et de loi commune une densité s pour la mesure de Lebesgue sur $\mathbb{R}^2$. Soit $\mathcal{M}_n$ une collection dénombrable de partitions de $[x_0, x_{max}]$ pour l'une des deux situations (H_1) ou (H_2), et une famille de poids positifs $(\rho_m)_{m\in\mathcal{M}_n}$ telle que pour tout n*

$$\sum_{m\in\mathcal{M}_n} e^{-\rho_m} = \Upsilon < \infty. \tag{5.23}$$

Dans chacun des modèles S_m^b, soit $\hat{g}_m^\star$ l'estimateur du maximum de vraisemblance pour l'échantillon observé. Alors il existe des constantes positives η et C telles que pour des fonctions de pénalités vérifiant

$$\mathrm{pen}(m) \geq \eta \frac{D_m}{n}\left(\ln n + \frac{\rho_m}{D_m}\right), \tag{5.24}$$

le modèle $\hat{m}$ minimisant

$$\mathrm{crit}(m) = \gamma_n(\hat{g}_m^\star) + \mathrm{pen}(m)$$

sur la famille $\mathcal{M}_n$ existe. Pour une telle pénalité, nous avons de plus

$$\mathbb{E}\left[d_H^2\left(s, \hat{g}_{\hat{m}}^\star\right)\right] \leq C\left[\inf_{m\in\mathcal{M}_n}\left\{\mathrm{KL}\left(s, S_m^b\right) + \mathrm{pen}(m)\right\} + \frac{\Upsilon}{n}\right].$$

De plus, η ne dépend que de h_{max}, x_{max}, $t^\star$, c et γ et C est une constante absolue.

La preuve de ce théorème ainsi que les bornes sur les entropies à crochets nécessaires pour obtenir ces résultats sont données dans la section 5.3.4. L'essentiel de la preuve repose sur des calculs techniques d'entropie à crochets. En particulier, il s'agit d'exprimer ces entropies en fonction du nombre de paramètres D_m des modèles.

Choix des poids

La forme de la pénalité minimale (5.24) tient compte d'une part de la taille des modèles via leur dimension, et d'autre part de la complexité de la collection $\mathcal{M}_n$ via les poids ρ_m. Nous reprenons les arguments de Castellan (1999) pour choisir des poids pour les deux collections de partitions proposées ci-dessus.

- (H_1) : Dans ce cas, choisir des poids de la forme $\rho_m = LD_m$, avec $L > 0$ permet de satisfaire la condition de sommabilité (5.23).
- (H_2) : Castellan montre que dans cette situation, les poids $\rho_m = LD_m \ln n$, avec $L > 0$, permettent à $\mathcal{M}_n$ de satisfaire (5.23).

5.3.3 Discussion

Comme il a déjà été noté à la suite du théorème 5.1.1, quitte à supposer que

$$\sup_{g^\star \in \bigcup_{m\in\mathcal{M}_n} S_m^b} \left\| \frac{s}{g^\star} \right\|_\infty \leq M \tag{5.25}$$

nous déduisons de (5.3) la borne de risque qui suit,

$$\mathbb{E}\left[d_H^2\left(s, \hat{g}_{\hat{m}}^\star\right)\right] \leq C' \left[\inf_{m\in\mathcal{M}_n} \left\{ d_H^2\left(s, S_m^b\right) + \mathrm{pen}(m) \right\} + \frac{\Upsilon}{n} \right],$$

où C' dépend de C et de M. Notons que dans notre contexte, la condition (5.25) est réalisée dès que s est bornée sur $\mathbb{R}^2$, et dans ce cas M dépend de $h_{\min}$, $h_{\max}$ et $\|s\|_\infty$.

Sous les deux hypothèses (H_1) et (H_2), les pénalités minimales déduites de (5.24) sont de la forme

$$\mathrm{pen}(m) = \eta \frac{D_m}{n} \ln n, \tag{5.26}$$

et nous parvenons à la borne de risque suivante, valable dans les deux cas,

$$\mathbb{E}\left[d_H^2(s, \hat{g}_{\hat{m}}^\star)\right] \leq C' \left[\inf_{m\in\mathcal{M}_n} \left\{ \mathrm{KL}(s, S_m^b) + \frac{D_m}{n} \ln n \right\} \right]. \tag{5.27}$$

Dans les deux situations (H_1) et (H_2), nous obtenons dans la pénalité et donc dans la borne de risque (5.27) un terme supplémentaire en $\ln n$ par rapport aux résultats obtenus par Castellan. La présence de ce terme logarithmique est directement liée à notre incapacité à contrôler l'entropie à crochets localement, comme le nécessiterait idéalement le théorème 5.1.1. Des poids plus complexes sont aussi proposés par Castellan pour obtenir des estimateurs optimaux[1] dans le cadre de l'estimation de densité par des histogrammes. Dans notre cas, ces poids ne permettent pas de se débarrasser du terme en $\ln n$ additionnel. Pour pouvoir discuter de l'optimalité de l'estimateur pénalisé $\hat{g}_{\hat{m}}^\star$, il nous faudrait comparer les vitesses minimax sur certaines classes fonctionnelles aux vitesses atteintes par l'estimateur pénalisé. Ceci pourra faire l'objet d'une recherche ultérieure à ce travail de thèse.

Comme il a déjà été dit au sujet des résultats du théorème 5.3.1, la borne de risque (5.27) ainsi que la pénalité minimale (5.26) doivent être considérés d'un point de vue qualitatif. Essentiellement, ces résultats nous donnent la forme générale de la pénalité à utiliser. La méthode de la pente, mise en pratique pour ces collections de modèles à la section 6.3, nous permet dans un second temps de calibrer la pénalité en fonction des données.

5.3.4 Preuve du théorème 5.3.1

La preuve du théorème 5.3.1 repose sur les majorations d'entropie à crochets des modèles S_m^b que l'on donne dans la proposition suivante.

Proposition 5.3.1. *Soit m une partition de $[x_0, x_{max}]$ composée de k intervalles, on note l_m la longueur minimale de ces derniers. Soient $\alpha > 0$ et $\varepsilon \in]0,1[$. Il existe une constante $\tilde{C}_1$ ne*

[1] au sens où ceux-ci atteignent la vitesse minimax sur certaines classes de densités.

dépendant que de h_{min}, h_{max}, x_{max}, et $t^\star$ telle que

$$N_{[.]}\left(\varepsilon, S_m^b, d_H\right) \leq \left(\frac{\tilde{C}_1}{\varepsilon}\right)^{k+1} \left(\frac{1}{l_m}\right)^k .$$

Nous en déduisons que,

$$H_{[.]}\left(\varepsilon, S_m^b, d_H\right) \leq (k+1)\left(C_1 + \ln^+ \frac{1}{l_m} + \ln \frac{1}{\varepsilon}\right),$$

avec $C_1 = \ln \tilde{C}_1$ et où $\ln^+ x$ désigne la partie positive du logarithme de $\ln x$.

La preuve de cette proposition est détaillée dans la section 5.3.5. Nous montrons maintenant comment le théorème 5.3.1 se démontre à partir de la proposition 5.3.1.

Soit $\mathcal{M}_n$ une famille de partitions de $[x_0, x_{\max}]$ pour l'une des situations (H_1) ou (H_2). Soit $m \in \mathcal{M}_n$, il s'agit de proposer une fonction Ψ_m vérifiant les hypothèses du théorème 5.1.1. Pour cela, il nous faut majorer une intégrale de la racine de l'entropie de l'ensemble $S_m^b(u,\xi) := \{t \in S_m^b; d_H(t,u) \leq \xi\}$. Il est techniquement difficile d'effectuer ces recouvrements locaux. Nous n'avons alors pas d'autres solutions que d'effectuer des recouvrements de l'espace S_m^b tout entier, et d'utiliser la majoration naïve

$$\mathrm{H}_{[.]}\left(\varepsilon, S_m^b(u,\xi), d_H\right) \leq \mathrm{H}_{[.]}\left(\varepsilon, S_m^b, d_H\right),$$

ainsi que la borne donnée par la proposition 5.3.1. On note D_m le nombre de paramètres libres de S_m^b, et donc $D_m = k+1$ où k désigne le nombre d'intervalles composant la partition m. Pour tout $\xi > 0$, en utilisant le lemme technique 5.2.18 donné à la section 5.2.8, nous obtenons

$$\begin{aligned}
\int_0^\xi \sqrt{\mathrm{H}_{[.]}(x, S_m^b, d_H)}\, dx &\leq \sqrt{D_m}\left\{\xi\left(\sqrt{C_1} + \sqrt{\ln^+ \frac{1}{l_m}}\right) + \int_0^{\xi\wedge 1} \sqrt{\ln \frac{1}{x}}\, dx\right\} \\
&\leq \sqrt{D_m}\,\xi\left\{\sqrt{C_1} + \sqrt{\ln^+ \frac{1}{l_m}} + \sqrt{\ln^+ \frac{1}{\xi}} + \sqrt{\pi}\right\} \\
&\leq \sqrt{D_m}\,\xi\left\{C_1' + \sqrt{\ln^+ \frac{1}{l_m}} + \sqrt{\ln^+ \frac{1}{\xi}}\right\}.
\end{aligned}$$

La fonction

$$\Psi_m : \xi \in \mathbb{R}_+^\star \mapsto \sqrt{D_m}\,\xi\left\{C_1' + \sqrt{\ln^+ \frac{1}{l_m}} + \sqrt{\ln^+ \frac{1}{\xi}}\right\}$$

satisfait la condition (P_1) du théorème 5.1.1. Il nous faut trouver $\xi_\star$ tel que $\Psi_m(\xi_\star) = \sqrt{n}\,\xi_\star^2$ pour déterminer l'expression de la pénalité minimale. Il s'agit de résoudre l'équation suivante

$$\sqrt{\frac{D_m}{n}}\left\{C_1' + \sqrt{\ln^+ \frac{1}{l_m}} + \sqrt{\ln^+ \frac{1}{\xi_\star}}\right\} = \xi_\star . \tag{5.28}$$

En remarquant que la quantité $\tilde{\xi} = \sqrt{\frac{D_m}{n}}\left(C_1' + \sqrt{\ln^+ \frac{1}{l_m}}\right)$ vérifie $\tilde{\xi} \leq \xi_\star$, ceci nous donne

$$\xi_\star \leq \sqrt{\frac{D_m}{n}}\left\{C_1' + \sqrt{\ln^+ \frac{1}{l_m}} + \sqrt{\ln^+ \frac{1}{\tilde{\xi}}}\right\}.$$

En utilisant le fait que la partition m satisfait $l_m \geq c\, n^{-\gamma}$ par la propriété (P_2), il vient

$$\begin{aligned}
\xi_\star^2 &\leq \frac{D_m}{n}\left\{4\, C_1'^2 + 4 \ln^+ \frac{1}{l_m} + \ln^+ \frac{n}{D_m(C_1'^2 + \ln^+ \frac{1}{l_m})}\right\} \\
&\leq \frac{D_m}{n}\left\{C_1'' + \gamma \ln n + \ln^+ \frac{n}{D_m}\right\} \\
&\leq C_1''' \frac{D_m}{n} \ln n.
\end{aligned}$$

Ceci achève la preuve du théorème 5.3.1.

5.3.5 Preuve de la proposition 5.3.1

Cette section est dédiée à la preuve de la proposition 5.3.1. On considère l'ensemble $\mathcal{A} = [0, \frac{h_{\max}}{l_m}]^k \times [h_{\min}, h_{\max}]$. On vérifie facilement que l'ensemble des fonctions $h(A,b)$ avec $(A,b) \in \mathcal{A}$ contient l'ensemble $\mathcal{H}_m^b$. Pour tout $\beta \in]0,1]$, on considère le maillage régulier $\tilde{R}(\beta)$ de pas β sur l'ensemble $\mathcal{A}$, défini à partir de l'élément $(0,\ldots,0,h_{\min})$. Plus précisément,

$$\tilde{R}(\beta) := \left\{(Q,q)\beta\ ;\ Q \in [\![0, \lfloor h_{\max}/(\beta l_m)\rfloor]\!]^k,\ (q - h_{\min}) \in [\![0, \lfloor (h_{\max} - h_{\min})/\beta \rfloor]\!]\right\},$$

où la notation $\lfloor x \rfloor$ désigne la partie entière de x. Puisque les fonctions de l'ensemble $\mathcal{H}_m^b$ sont bornées par $h_{\max}$, nous pouvons nous restreindre au réseau suivant,

$$R(\beta) = \left\{(A,b) \in \tilde{R}(\beta)\ ;\ h(A,b) \leq h_{\max}\right\}.$$

Toutes les fonctions $h(A,b)$ construites sur le maillage $R(\beta)$ vérifient donc $h_{\min} \leq h(a,b) \leq h_{\max}$, et de plus,

$$\operatorname{card}\{R(\beta)\} \leq \left(1 + \frac{h_{\max}}{l_m \beta}\right)^k \left(1 + \frac{h_{\max}}{\beta}\right). \tag{5.29}$$

Dans la suite, il nous sera pratique de considérer aussi des fonctions $h(A,b)$ étendues à $\mathbb{R}^+$ de la façon suivante : pour $h \in \mathcal{H}_m^b$, on définit la fonction $\tilde{h}$ par

$$\tilde{h}(x) = \begin{cases} \frac{x}{x_0} h(x_0) & \text{pour } x \in [0, x_0] \\ h(x) & \text{pour } x \in [x_0, x_{\max}] \\ h(x_{\max}) + (x - x_{\max}) & \text{pour } x \geq x_{\max} \end{cases}.$$

Prolonger ainsi les fonctions h permet de disposer de fonctions $\tilde{h}$ croissantes dont l'image recouvre $\mathbb{R}^+$ tout entier. Si $(h_1, h_2) \in \left(\mathcal{H}_m^b\right)^2$ alors $\left\|\tilde{h}_1 - \tilde{h}_2\right\|_\infty = \|h_1 - h_2\|_\infty$. Si de plus, $h_1 \leq h_2$, on a encore $\tilde{h}_1 \leq \tilde{h}_2$.

Soit $(A_1, b_1) \in R(\beta)$, on considère l'élément $(A_2, b_2) := (A_1, b_1) + (\beta, \ldots, \beta)$. Á l'élement (A_i, b_i), nous associons la fonction $h_i := h(a_i, b_i)$ et son prolongement $\tilde{h}_i$, pour $i = 1, 2$. La fonction ϕ est définie sur $(\mathbb{R}^+)^2$ par $\phi(u, t) = u \exp(-ut)$. Dans les définitions de fonctions qui suivent, on adopte la convention $[x, x] = \emptyset$. Soient les fonctions

$$\begin{aligned}
\tilde{\Phi}_d(x,t) &:= \min\left\{\phi\left(\tilde{h}_1(x), t\right), \phi\left(\tilde{h}_2(x), t\right)\right\}, \\
\tilde{\Phi}_u(x,t) &:= \phi\left(\tilde{h}_2(x), t\right) \mathbb{1}_{[0, x_2(t)]}(x,t) + \tfrac{e^{-1}}{t} \mathbb{1}_{[x_2(t), x_1(t)]}(x,t) + \phi\left(\tilde{h}_1(x), t\right) \mathbb{1}_{[x_1(t), +\infty[}(x,t)
\end{aligned}$$

où $x_i(t) := \inf\{x\,;\, \tilde{h}_i(x) = \frac{1}{t}\}$, pour $i = 1, 2$. Le fait de prolonger h_i en $\tilde{h}_i$ permet de disposer d'un inverse généralisé en $1/t$ pour toutes les fonctions de $h \in \mathcal{H}_m^b$. On note encore

$$\begin{aligned}
\Phi_d &:= \tilde{\Phi}_d \mathbb{1}_{x \in [x_0, x_{\max}]}, \\
\Phi_u &:= \tilde{\Phi}_u \mathbb{1}_{x \in [x_0, x_{\max}]}.
\end{aligned}$$

Nous pouvons finalement définir le crochet d'extrémités

$$\begin{aligned}
G_d(x,t) &:= \tfrac{\alpha x^{-\alpha-1}}{x_0^{-\alpha} - x_{\max}^{-\alpha}} \tfrac{1}{\mathrm{P}_{\mathrm{dec}}(\alpha, t^\star, h_2)} \Phi_d(x,t) \mathbb{1}_{[0,t^\star]}(t), \\
G_u(x,t) &:= \tfrac{\alpha x^{-\alpha-1}}{x_0^{-\alpha} - x_{\max}^{-\alpha}} \tfrac{1}{\mathrm{P}_{\mathrm{dec}}(\alpha, t^\star, h_1)} \Phi_u(x,t) \mathbb{1}_{[0,t^\star]}(t).
\end{aligned}$$

Lemme 5.3.2. *Avec les notations précédentes,*

1. *l'ensemble des crochets $[\Phi_d, \Phi_u]$ recouvre l'ensemble $\left\{(x,t) \mapsto \phi(h(x), t),\ h \in \mathcal{H}_m^b\right\}$,*
2. *l'ensemble des crochets $[G_d, G_u]$ recouvre S_m^b.*

Démonstration. 1. Soit $h \in \mathcal{H}_m^b$, et $(A, b) \in \mathcal{A}$ tel que $h = h(A, b)$. Il existe $(A_1, b_1) \in R(\beta)$ tel que $(A_1, b_1) \leq (A, b) \leq (A_2, b_2)$, avec $(A_2, b_2) := (A_1, b_1) + (\beta, \ldots, \beta)$. Pour $i = 1, 2$, on considère ensuite $h_i := h(A_i, b_i)$ et son prolongement $\tilde{h}_i$. On a alors $\tilde{h}_1 \leq \tilde{h} \leq \tilde{h}_2$.

On fixe $t \in [0, t^\star]$, la fonction $\phi(\cdot, t)$ est croissante sur $[0, 1/t]$ et décroissante sur $[1/t, +\infty[$. Les éléments $x_1(t)$ et $x_2(t)$ définis plus haut permettent de décrire les variations des fonctions $\phi(h(\cdot), t)$, $\phi(h_1(\cdot), t)$ et $\phi(h_2(\cdot), t)$ et aussi de les comparer entre elles. On note que $x_2(t) \geq x(t) \geq x_1(t)$ où $x(t) := \inf\{x\,;\, \tilde{h}(x) = \frac{1}{t}\}$.

- Sur $[0, x_2(t)]$, nous avons $\tilde{h}_1(x) \leq \tilde{h}(x) \leq \tilde{h}_2(x) \leq \frac{1}{t}$. Ceci implique que $\phi(h_1(x), t) \leq \phi(h(x), t) \leq \phi(h_2(x), t)$.
- Sur $[x_1(t), +\infty[$, nous avons cette fois $\frac{1}{t} \leq \tilde{h}_1(x) \leq \tilde{h}(x) \leq \tilde{h}_2(x)$ et donc $\phi(h_2(x), t) \leq \phi(h(x), t) \leq \phi(h_1(x), t)$.
- Sur $[x_2(t), x(t)]$, nous avons $\tilde{h}_1(x) \leq \tilde{h}(x) \leq \frac{1}{t}$ et donc $\phi(h_1(x), t) \leq \phi(h(x), t)$.
- De même, $\phi(h(\cdot), t)$ est minorée par $\phi(h_2(\cdot), t)$ sur $[x(t), x_1(t)]$.
- On note enfin que la fonction $\phi(\cdot, t)$ est majorée par $\frac{1}{t} e^{-1}$.

Tous ces inégalités montrent que le crochet $[\tilde{\Phi}_d(\cdot, t), \tilde{\Phi}_u(\cdot, t)]$ contient la fonction $\phi(h(\cdot), t)$. Ceci est vérifié pour tout $t \in [0, t^\star]$ et le premier point du lemme est démontré en restreignant la variable x au domaine $[x_0, x_{\max}]$.

2. Soit $g^\star \in S_m^b$ et la fonction $h \in \mathcal{H}_m^b$ qui lui est associée. Le crochet $[\Phi_d, \Phi_u]$ défini comme ci-dessus encadre la fonction $(x, t) \mapsto \phi(h(x), t)$. La croissance de la probabilité $\mathrm{P}_{\mathrm{dec}}$ en

la fonction h implique que $\mathrm{P}_{\mathrm{dec}}(\alpha, t^\star, h_1) \leq \mathrm{P}_{\mathrm{dec}}(\alpha, t^\star, h) \leq \mathrm{P}_{\mathrm{dec}}(\alpha, t^\star, h_2)$. Ceci, avec le premier point, montre que $[G_d, G_u]$ contient la fonction $g^\star$.

□

Lemme 5.3.3. *Sous les hypothèses précédentes,*

$$\forall (x,t) \in [x_0, x_{max}] \times [0, t^\star], \quad \left|\sqrt{\Phi_u}(x,t) - \sqrt{\Phi_d}(x,t)\right| \leq c_1 \left\|h_2 - h_1\right\|_\infty^2 \tag{5.30}$$

et

$$d_H^2(G_d\,, G_u) \leq c_2 \left\|h_2 - h_1\right\|_\infty . \tag{5.31}$$

où c_1 et c_2 sont des constantes qui ne dépendent que de h_{min}, h_{max}, x_{max} et $t^\star$.

Démonstration. Pour tout $(u,t) \in (R^+)^2$, nous avons $\left|\frac{\delta}{\delta u}\phi(u,t)\right| \leq 1$. Ceci nous donne

$$\left|\phi\left(\tilde{h}_2(x), t\right) - \phi\left(\tilde{h}_1(x), t\right)\right| \leq \left\|h_2 - h_1\right\|_\infty .$$

Nous avons vu dans la démonstration du lemme précédent que

$$\begin{aligned}\tilde{\Phi}_d(x,t) := \quad & \phi\left(\tilde{h}_1(x), t\right) \mathbb{1}_{[0, x_2(t)]}(x,t) + \min\left\{\phi\left(\tilde{h}_1(x), t\right), \phi\left(\tilde{h}_2(x), t\right)\right\} \mathbb{1}_{[x_2(t), x_1(t)]}(x,t) \\ & + \phi\left(\tilde{h}_2(x), t\right) \mathbb{1}_{[x_1(t), +\infty[}(x,t).\end{aligned}$$

Sur les deux segments $[0, x_2(t)]$ et $[x_1(t), +\infty[$, nous avons donc $\left|\tilde{\Phi}_u(x,t) - \tilde{\Phi}_d(x,t)\right| \leq \left\|h_2 - h_1\right\|_\infty$. Nous vérifions maintenant cette inégalité sur le segment central $[x_2(t), x_1(t)]$. Fixons $t \in [0, t^\star]$. Sur $[x_2(t), x_1(t)]$, la fonction $\phi\left(\tilde{h}_1(\cdot), t\right)$ croît jusqu'en son maximum $\phi\left(\tilde{h}_1(x_1(t)), t\right) = \frac{e^{-1}}{t}$, alors que la fonction $\phi\left(\tilde{h}_2(\cdot), t\right)$ décroît depuis le même maximum $\phi\left(\tilde{h}_2(x_2(t)), t\right) = \frac{e^{-1}}{t}$. La quantité $\min\left\{\phi\left(\tilde{h}_1(x), t\right), \phi\left(\tilde{h}_2(x), t\right)\right\}$ est donc minimale en l'une des deux extrémités $x_2(t)$ ou $x_1(t)$. Nous avons alors pour tout $x \in [x_2(t), x_1(t)]$,

$$\begin{aligned} 0 \quad &\leq \quad \frac{e^{-1}}{t} - \min\left\{\phi\left(\tilde{h}_1(x), t\right), \phi\left(\tilde{h}_2(x), t\right)\right\} \\ &\leq \quad \max\left\{\left|\phi\left(\tilde{h}_2(x_2(t)), t\right) - \phi\left(\tilde{h}_1(x_2(t)), t\right)\right|, \left|\phi\left(\tilde{h}_1(x_1(t)), t\right) - \phi\left(\tilde{h}_2(x_1(t)), t\right)\right|\right\} \\ &\leq \quad \left\|h_2 - h_1\right\|_\infty , \end{aligned}$$

d'après ce que nous avons montré ci dessus. En restreignant la variable x au segment $[x_0, x_{\max}]$, nous obtenons $|\Phi_u - \Phi_d| \leq \left\|h_2 - h_1\right\|_\infty$. Notons que pour $i = 1$ et 2,

$$h_{\min} \leq h_i \leq h_{\max} + \beta \leq h'_{\max} := h_{\max} + 1 \tag{5.32}$$

et donc $\Phi_d \geq \Phi_{\min} := h_{\min} \exp\{-h'_{\max} t^\star\}$. Nous en déduisons que

$$\begin{aligned} \left|\sqrt{\Phi_u} - \sqrt{\Phi_d}\right| \quad &= \quad \frac{|\Phi_u - \Phi_d|}{\sqrt{\Phi_u} + \sqrt{\Phi_d}} \\ &\leq \quad \frac{\left\|h_2 - h_1\right\|_\infty}{2\sqrt{\Phi_{\min}}} \end{aligned}$$

ce qui nous donne (5.30).

Il nous reste à évaluer la longueur du crochet $[G_d\,, G_u]$,

$$\begin{aligned} d_H^2(G_d\,, G_u) &= \int_0^{t^\star}\int_{x_0}^{x_{\max}} \frac{\alpha x^{-\alpha-1}}{x_0^{-\alpha}-x_{\max}^{-\alpha}} \left\{\sqrt{\frac{\Phi_d(x,t)}{\mathrm{P}_{\mathrm{dec}}(h_2)}} - \sqrt{\frac{\Phi_u(x,t)}{\mathrm{P}_{\mathrm{dec}}(h_1)}}\right\}^2 dx\,dt \\ &\leq 2\int_0^{t^\star}\int_{x_0}^{x_{\max}} \frac{\alpha x^{-\alpha-1}}{x_0^{-\alpha}-x_{\max}^{-\alpha}}\,\Phi_d(x,t) \left\{\frac{1}{\sqrt{\mathrm{P}_{\mathrm{dec}}(h_2)}} - \frac{1}{\sqrt{\mathrm{P}_{\mathrm{dec}}(h_1)}}\right\}^2 dx\,dt \\ &\quad + 2\int_0^{t^\star}\int_{x_0}^{x_{\max}} \frac{\alpha x^{-\alpha-1}}{x_0^{-\alpha}-x_{\max}^{-\alpha}}\,\frac{1}{\mathrm{P}_{\mathrm{dec}}(h_1)} \left\{\sqrt{\Phi_d(x,t)} - \sqrt{\Phi_u(x,t)}\right\}^2 dx\,dt. \end{aligned}$$

Notons d'une part que $|\mathrm{P}_{\mathrm{dec}}(h_2) - \mathrm{P}_{\mathrm{dec}}(h_1)| \leq t^\star \|h_2 - h_1\|_\infty$ et d'autre part que $\mathrm{P}_{\mathrm{dec}}(h_1) \geq \mathrm{p}_{\mathrm{dec}} := 1 - \exp\{-h_{\min}t^\star\} > 0$. De plus, (5.32) nous donne que $\Phi_d \leq h'_{\max}$. Ces trois résultats, avec (5.30), conduisent à la majoration suivante,

$$d_H^2(G_d\,, G_u) \leq \left(\frac{h'_{\max}t^{\star 3}}{\mathrm{p}_{\mathrm{dec}}{}^3} + 2\frac{t^\star c_1}{\mathrm{p}_{\mathrm{dec}}}\right) \|h_2 - h_1\|_\infty^2\,.$$

□

Pour deux vecteurs (A_1, b_1) et (A_2, b_2) de l'ensemble $\mathcal{A}$, nous avons

$$\begin{aligned} \|h(A_1, b1) - h(A_2, b_2)\|_\infty &\leq |b_2 - b_1| + \sum_{j=1}^{k}(x_j - x_{j-1})|A_{2j} - A_{1j}| \\ &\leq x_{\max}\,\|(A_1, b_1) - (A_2, b_2)\|_\infty \\ &\leq x_{\max}\beta \end{aligned}$$

car $x_0 \geq 1$. D'après (5.31), les crochets $[G_d, G_u]$ définis à partir du réseau $R(\beta)$ vérifient donc

$$d_H^2(G_d\,, G_u) \leq c_2\, x_{\max}\beta^2.$$

Soit $\varepsilon > 0$; posons $\beta = \frac{\varepsilon}{\sqrt{c_2\, x_{\max}}}$. D'après les deux lemmes précédents, l'ensemble des crochets $[G_d, G_u]$ définis à partir du maillage $R(\beta)$ forme un ε-recouvrement de S_m^b. Il suffit pour finir de noter que le nombre de ces crochets est égal au cardinal de $R(\beta)$. On conclut finalement la preuve de la proposition 5.3.1 en utilisant (5.29).

Chapitre 6

Heuristique de pente et applications

Les résultats théoriques obtenus au chapitre précédent pour la sélection de modèles de mélanges gaussiens et de modèles d'exploration pétrolière ne peuvent être appliqués directement. En effet, la forme des pénalités a été déterminée mais il reste à calibrer les constantes pour pouvoir utiliser en pratique les critères pénalisés associés. Dans ce chapitre, nous nous appuyons sur une méthode de calibration de pénalité appelée "méthode de la pente" pour déterminer des pénalités optimales à partir des données. La première section est un rappel sur la méthode de la pente. Les sections qui suivent présentent les applications aux collections de modèles de mélanges gaussiens et aux collections de modèles d'exploration pétrolière.

6.1 La méthode de la pente

L'objet de cette première section de cette section est de présenter les aspects théoriques et pratiques de la méthode de la pente. Cette méthode a été proposée par Birgé et Massart (2001; 2006) pour calibrer à partir des données des pénalités de formes connues. De nombreuses applications de cette méthode ont été développées, notamment dans le domaine de la détection de ruptures par Lebarbier (2005), en génomique par Villers (2007), pour des modèles graphiques par Verzelen (2007) et aussi en classification non supervisée par Baudry (2007).

6.1.1 Heuristique de pente

Dans de nombreuses situations en sélection de modèles, la collection de modèles considérée peut contenir plusieurs modèles de même dimension. Dans ce cas, nous considérons la collection $(S_D)_{D\in\mathcal{D}}$ obtenue en rassemblant dans un seul modèle les modèles de même dimension. Dans la suite, la collection de modèles peut donc être supposée indexée par la dimension des modèles.

Soit $X_1, \ldots, X_n$ un échantillon d'une même distribution inconnue s. Comme au chapitre 5, γ and γ_n désignent respectivement le contraste associé à l'information de Kullback-Leibler et le contraste empirique qui lui est associé. On suppose que pour tout D dans $\mathcal{D}$, il existe deux densités s_D et $\hat{s}_D$ qui minimisent respectivement $\mathrm{KL}(s,.)$ et $\gamma_n(\cdot)$ sur S_D ; la densité $\hat{s}_D$ est donc l'estimateur du maximum de vraisemblance de s sur S_D. Le modèle que l'on souhaiterait idéalement sélectionner dans la collection est celui pour lequel le risque $\mathbb{E}\left[\mathrm{KL}(s,\hat{s}_D)\right]$ est

minimum, ce qui est impossible à réaliser directement car cette quantité dépend de la densité s inconnue. Pour chaque estimateur $\hat{s}_D$, nous considérons la décomposition suivante du risque

$$\mathbb{E}\left[\mathrm{KL}(s, \hat{s}_D)\right] = b_D + \mathbb{E}(V_D)$$

où $V_D := \int \ln(s_D/\hat{s}_D)\, s\, dx$ est un terme de variance et $b_D := \mathrm{KL}(s, s_D)$ est un terme de biais. Notons que la quantité V_D augmente et que le biais b_D diminue lorsque la dimension D augmente.

La sélection de modèles par critère pénalisé consiste à sélectionner le modèle de la collection $(S_D)_{D\in\mathcal{D}}$ qui minimise un critère de la forme

$$\mathrm{crit}(D) = \gamma_n(\hat{s}_D) + \mathrm{pen}(D). \tag{6.1}$$

En posant $\hat{b}_D := \gamma_n(s_D) - \gamma_n(s)$ et $\widehat{V}_D := \gamma_n(s_D) - \gamma_n(\hat{s}_D)$, il est alors clair que le modèle sélectionné selon le critère (6.1) est aussi un minimiseur de

$$\begin{aligned} \gamma_n(\hat{s}_D) - \gamma_n(s) + \mathrm{pen}(D) &= \hat{b}_D - \widehat{V}_D + \mathrm{pen}(D) \\ &= \mathrm{KL}(s, \hat{s}_D) + (\hat{b}_D - b_D) - (V_D + \widehat{V}_D) + \mathrm{pen}(D)\ . \end{aligned} \tag{6.2}$$

D'après la loi des grands nombres, $\hat{b}_D - b_D \approx 0$ presque sûrement. De plus, des arguments de concentration permettent de montrer que la quantité $\mathrm{KL}(s, \hat{s}_D)$ est proche de son espérance, qui est justement le risque de $\hat{s}_D$. Pour que le critère (6.2) soit proche du risque $\mathbb{E}\left[\mathrm{KL}(s, \hat{s}_D)\right]$, la pénalité *optimale* doit donc être choisie telle que

$$\mathrm{pen}_{\mathrm{opt}}(D) = V_D + \widehat{V}_D\ .$$

Ensuite, l'hypothèse principale de l'heuristique de pente est de supposer que $\widehat{V}_D \approx V_D$. Pour justifier ceci, on peut remarquer que dans les expressions de V_D et $\widehat{V}_D$, la mesure de probabilité et la mesure empirique jouent un rôle "symétrique". Si l'on échange ces deux mesures dans les expressions de V_D et $\widehat{V}_D$, ainsi que dans les expressions de s_D et $\hat{s}_D$ alors V_D devient $\widehat{V}_D$ et réciproquement. Cette hypothèse conduit à considérer une pénalité de la forme $\mathrm{pen}(D) = 2\widehat{V}_D$ et il reste à déterminer $\widehat{V}_D$ à partir des données. Par définition de $\widehat{V}_D$, nous avons

$$\widehat{V}_D = \hat{b}_D + \gamma_n(s) - \gamma_n(\hat{s}_D).$$

Or, pour les modèles de grande dimension, le terme de biais finit par se stabiliser et la fonction de $D \mapsto \widehat{V}_D$ peut donc être connue par l'intermédiaire de la quantité $-\gamma_n(\hat{s}_D)$.

Dans de nombreuses situations en statistiques, les théorèmes de sélection de modèles non asymptotiques conduisent à considérer des pénalités proportionnelles à $\frac{D}{n}$. Dans ce cas, il est alors facile de vérifier que la fonction $-\gamma_n(\hat{s}_D)$ est bien une fonction linéaire de $\frac{D}{n}$ pour D assez grand. Pour n fixé, l'estimation $\hat{\eta}$ de la pente de cette fonction conduit finalement à choisir la pénalité suivante

$$\mathrm{pen}(D) = 2\hat{\eta}D.$$

6.1.2 Méthode de la pente et sauts de dimension

Pour compléter cette description de la méthode de la pente, nous en rappelons maintenant la présentation plus classique proposée dans Birgé et Massart (2006) et Arlot (2007). Supposons que l'on s'intéresse à une pénalité de la forme $\eta\,\mathrm{pen}$ où pen est une fonction de D et n et où η est un paramètre à régler. L'heuristique de pente repose sur les deux assertions suivantes.

- Il existe une pénalité minimale $\mathrm{pen}_{\min} = \eta_{\min}\,\mathrm{pen}$ dans la famille de fonctions retenues telle que toute pénalité inférieure sélectionne les modèles de grandes dimensions, et toute pénalité supérieure sélectionne des modèles de dimensions "raisonnables".
- Une pénalité choisie de l'ordre de $2\,\mathrm{pen}_{\min}$ permet de sélectionner un estimateur de risque comparable à celui de l'oracle.

Pour des pénalités proportionnelles à $\frac{D}{n}$, nous vérifions que cette présentation de la méthode de la pente et la description précédente sont cohérentes en considérant (à n fixé) pour pénalité minimale $\mathrm{pen}_{\min} = \hat{\eta}D$ où $\hat{\eta}$ est l'estimation de la pente définie plus haut.

Les deux assertions précédentes ont été vérifiés par Birgé et Massart (2006) dans le cadre de la régression sur design fixe avec bruit blanc gaussien homoscédastique, et par Arlot (2007) dans le contexte d'un bruit blanc hétéroscédastique sur design aléatoire, pour le cas des histogrammes. Des progrès importants ont été réalisés dans ce domaine depuis quelques années, et il est probable que cette heuristique soit encore valable dans de nombreuses autres situations. Nous utilisons donc cette méthode dans les sections suivantes pour calibrer les pénalités dont les formes générales ont été déterminées au chapitre précédent.

6.2 Classification non supervisée et sélection de variables simultanés

Dans cette section, la méthode de la pente est appliquée aux collections modèles de mélanges gaussiens $(S_{(K,\mathbf{v})})$.

6.2.1 Quelle forme de pénalité utiliser en pratique ?

Au chapitre 5, nous avons démontré que pour des collections de modèles ordonnées et des pénalités de la forme

$$\mathrm{pen}_{\min}(K,v) \geq \kappa \frac{D(K,v)}{n} \left\{ 2A \ln v + 1 + \ln \left(\frac{1}{1 \wedge \frac{D(K,v)}{n} A \ln v} \right) \right\}, \tag{6.3}$$

une inégalité oracle est réalisée par l'estimateur pénalisé $\hat{s}_{(\hat{K},\hat{v})}$. La borne inférieure donnée dans (6.3) ne constitue pas en toute rigueur ce que l'on appelle une pénalité minimale. Il faudrait pour cela qu'il y ait un comportement explosif de la dimension du modèle sélectionné pour des pénalité inférieur à la borne proposée, ce qui est impossible à démontrer car les constantes en jeu ne sont de toutes façons pas raisonnables. Cependant, l'inégalité (6.3) suggère d'utiliser des pénalités (qu'elle soit minimale ou non) de la forme du terme de droite. Ensuite, omme nous l'avons déjà mentionné dans le chapitre précédent, il est probable que

l'on puisse se passer du terme $\kappa\frac{D(K,v)}{n}\ln\left(\frac{1}{1\wedge\frac{D(K,v)}{n}A\ln v}\right)$ pour définir des pénalités efficaces. Ceci nous conduit donc à considérer des pénalités de la forme $\kappa\frac{D(K,v)}{n}(1+A\ln v)$. Les simulations présentées en annexe indiquent de plus que l'on peut remplacer le terme $\ln v$ par une constante (voir section C.2). De plus, dans les simulations de la section C.2, pour une pénalité choisie proportionnelle à $\frac{D(K,v)}{n}$, les estimateurs de pénalité calibrée par la méthode de la pente ont des performances proches de celle de l'oracle. Nous considérons donc des pénalités proportionnelles à $\frac{D(K,\mathbf{v})}{n}$ pour les applications présentées plus loin.

6.2.2 Utilisation de la méthode de la pente

Nous adoptons ici la première version de la méthode la pente telle qu'est elle est présentée dans la section 6.1.1. En effet, la dimension des modèles dépend ici de K et v, et ceci rend parfois plus difficile l'observation du saut de dimension sur lequel repose la deuxième version de la méthode de la pente (section 6.1.2). Nous décrivons maintenant en détail comment la méthode de la pente est utilisée pour définir des pénalités efficaces et sélectionner un modèle dans une collection $\left(S_{(K,\mathbf{v})}\right)_{(K,\mathbf{v})\in\mathcal{M}}$ avec $\mathcal{M}:=\{(K,\mathbf{v});\ 2\leq K\leq K_{\max},\mathbf{v}\in\mathcal{V}\}$ où $\mathcal{V}$ est un ensemble de parties de $\{1,\ldots,Q\}$.

La méthode de la pente se déroule en trois étapes :

1. *Étape d'estimation* : L'estimateur du maximum de vraisemblance est calculé pour chaque modèle $S_{(K,\mathbf{v})}$. Une densité dans un modèle $S_{(K,\mathbf{v})}$ est de la forme

$$t(x)=f(x_{[\mathbf{v}]})\,\Phi\left(x_{[\mathbf{v}^c]}\,|\,0_{Q-v},I_{Q-v}\right)$$

   où f est un mélange gaussien à K composantes et de dimension v dont les matrices de covariance sont de forme imposée par la collection. Les paramètres $(\hat{p}_1,\ldots,\hat{p}_K,\hat{\mu}_1,\ldots,\hat{\mu}_K,\widehat{\Sigma}_1,\ldots,\widehat{\Sigma}_K)$ du mélange sont estimés par un algorithme EM (Estimation Maximization) (Dempster *et al.*, 1977) en utilisant le logiciel MIXMOD (Biernacki *et al.*, 2006).

2. *Étape de détermination de la pénalité* : Une nouvelle collection de modèles $(S_D)_{D\in\mathcal{D}}$ est formée en réunissant tous les modèles $S_{(K,\mathbf{v})}$ de même dimension D dans un même modèle S_D. Un seuil D_0 est choisi tel qu'au delà de celui-ci la fonction $D\mapsto-\gamma_n(\hat{s}_D)$ ait un comportement linéaire en D. Soit $\hat{\eta}$ une estimation de la pente correspondante en utilisant une régression robuste. Nous obtenons finalement la pénalité

$$\mathrm{pen}_{\mathrm{opt}}(D)=2\hat{\eta}D\ .$$

3. *Étape de sélection de modèle* : Soit $\hat{D}$ dans $\mathcal{D}$ qui minimise le critère $D\mapsto\gamma_n(\hat{s}_D)+2\hat{\eta}D$, et soit $(\hat{K},\hat{\mathbf{v}})$ le modèle de la collection initiale qui vérifie

$$D(\hat{K},\hat{\mathbf{v}})=\hat{D}\ \text{ et }\ \hat{s}_{(\hat{K},\hat{\mathbf{v}})}=\hat{s}_{\hat{D}}.$$

   Au final, la partie mélange gaussien de la densité $\hat{s}_{(\hat{K},\hat{\mathbf{v}})}$ fournit automatiquement une classification des données par la règle du MAP (voir section 4.4.2).

De façon à valider la procédure, il est aussi recommandé de vérifier que la nappe du nombre de paramètres $(K, v) \mapsto D(K, v)$ s'ajuste bien sur la nappe des log-vraisemblances $(K, v) \mapsto \gamma_n(\hat{s}_{(K,v)})$.

6.2.3 Classification de courbes par modèles de mélanges gaussiens

La méthode de sélection de modèles de mélanges gaussiens qui est maintenant totalement décrite permet en particulier de faire de la classification non supervisée de courbes. Pour cela, les données de courbes initiales sont tout d'abord projetées dans l'espace correspondant aux coefficients d'une décomposition de Fourier, une décomposition en ondelettes, ou encore une décomposition spline. Cet espace est choisi de dimension suffisante Q pour que la procédure de sélection de variables soit réellement utile.

Dans le cadre de la classification de courbes, les collections non ordonnées ne peuvent généralement pas être utilisées, sauf à réduire à une dizaine le nombre de variables considérées. Les décompositions de Fourier et en ondelettes présentent l'avantage de fournir un ordonnancement (même partiel) des variables et pour comparer l'allure générale des courbes, il est naturel de privilégier les basses fréquences ou les informations de faible résolution. Ainsi, en pratique nous préférons utiliser ces deux types de décomposition. Notons que les décompositions en ondelettes ont déjà été utilisées en classification de courbes, citons par exemple les travaux de Berlinet *et al.* (2008) dans un contexte de classification supervisée.

Pour nous placer dans le cadre des collections de modèles $(S_{(K,\mathbf{v})})$ définies à la section 4.4.4, les coefficients obtenus sont ensuite centrés et réduits dans chaque variable. Nous appliquons la méthode de la pente à une collection de modèles de mélanges afin de déterminer finalement une classification des courbes initiales.

6.2.4 Application à la classification de profils de production pétroliers

Dans cette section, la procédure de classification de courbes est utilisée pour étudier un échantillon de profils de production pétroliers. La première section est consacrée au détail de la construction de la base de donnée.

6.2.4.1 Construction de la base de données

Les états anglais et norvégiens publient les productions mensuelles des plusieurs centaines de champs qu'ils possèdent[1]. Cette situation est assez exceptionnelle car pour de nombreuses zones exploitées il est nécessaire d'acheter ces informations à des sociétés qui les commercialisent. Pour comparer les formes des profils sans tenir compte de leur amplitude, toutes les courbes sont initialement normalisées par une estimation des réserves de chacun des champs[2].

[1] Site internet norvégien du Norwegian Petroleum Directorate : `www.npd.no/engelsk/cwi/pbl/en/index.htm`, et site internet anglais du Department of Trade and Industry : `www.og.dti.gov.uk/fields/fields_index.htm`.

[2] En pratique, cette opération peut s'avérer délicate ! Par exemple, la base de données des productions anglaises ne fournit pas d'estimations de réserves des gisements, il faut alors utiliser des bases telles que IHS et Wood Mackenzie. Or, les champs pétroliers des différentes bases ne correspondent pas toujours, certains champs sont par exemple regroupés en une même entité.

Nous obtenons ainsi une base initiale de courbes homogènes représentant les taux de production de chacun des champs.

Idéalement, la procédure de classification devrait être effectuée sur des profils de production de gisements dont l'exploitation est achevée, de façon à étudier des courbes de production complètes. En réalité, ceci est impossible à réaliser car la grande majorité des champs de la mer du Nord sont toujours en production aujourd'hui. Beaucoup de plus petits gisements ne produisent que depuis peu d'années. Quand aux champs les plus gros, même s'ils sont exploités depuis déjà plusieurs décennies, leur production n'est pas pour autant terminée. Cependant, les débuts de profil sont connus, et comme on peut le vérifier sur la figure 4.1, les premiers mois de la production révèlent déjà des différences de forme importantes entre les courbes. Dans la suite, nous ne considérons que les 64 premiers mois de l'exploitation, et tous les champs n'ayant pas produits sur une période assez longue sont retirés de l'échantillon initial. Une dizaine de gisements dont la production présente des anomalies (arrêts de production dûs à des accidents) sont aussi éliminés de la base. Nous aboutissons ainsi à une famille de 180 courbes homogènes entre elles et toutes définies sur une période de 64 mois.

Transformation discrète d'ondelettes des profils de production normalisés. De façon à obtenir des informations en temps et en fréquence, nous utilisons une transformée en ondelettes discrète des profils de production normalisés. Nous rappelons maintenant les définitions et les résultats nécessaires pour présenter rigoureusement les décompositions en ondelettes effectuées sur chacune des 180 courbes. Pour plus d'informations sur ce sujet, le lecteur pourra par exemple consulter l'ouvrage de Percival et Walden (2000). Soit $\mathbf{Y}$ une série temporelle de longueur $N = 2^J$. On note $\mathcal{W}$ la matrice $N \times N$ associée à la transformée en ondelettes et $\mathbf{W}$ le vecteur de $\mathbb{R}^n$ des coefficients d'ondelettes de $\mathbf{Y}$,

$$\mathbf{W} = \mathcal{W}\,\mathbf{Y}.$$

La matrice $\mathcal{W}$ peut s'écrire sous la forme suivante

$$\mathcal{W} = \begin{bmatrix} \mathcal{V}_J \\ \mathcal{W}_J \\ \mathcal{W}_{J-1} \vdots \\ \mathcal{W}_1 \end{bmatrix}.$$

Pour tout $j \in \{1, \ldots, J\}$, la sous-matrice $\mathcal{W}_j$ est de taille $N/2^j \times N$. Il est possible de décrire les lignes de $\mathcal{W}_j$ à partir des fonctions d'ondelettes

$$\Psi_{j,k} = 2^{-j/2}\mathbf{\Psi}\left(2^{-j}x - k2^j\right), \quad j = 1, \ldots, J,\ k = 0, \ldots, 2^{J-j} - 1$$

où $\mathbf{\Psi}$ est appelée ondelette mère de la famille. Pour ce qui nous concerne, nous utilisons l'ondelette de Haar $\mathbf{\Psi} = \mathbb{1}_{[0,1/2]} - \mathbb{1}_{[1/2,1]}$. La k-ème ligne de $\mathcal{W}_j$ est définie par le vecteur des valeurs prises par $\Psi_{j,k}$ aux points entiers $\{0, \ldots, N-1\}$. On peut alors montrer que la

transformée en ondelettes est une transformation orthogonale et donc

$$\mathbf{Y} = \mathcal{W}'\mathbf{W}.$$

Le vecteur $\mathbf{W}$ se décompose de même que $\mathcal{W}$, en posant

$$\mathbf{W} = \left(\mathbf{V}'_J, \mathbf{W}'_J, \ldots, \mathbf{W}'_1\right)'$$

où $\mathbf{W_j} = \mathcal{W}_j\mathbf{Y}$ contient l'ensemble des coefficients d'ondelettes associés au niveau j. Les coefficients $\mathbf{W_j}$ correspondent aux variations de $\mathbf{Y}$, à l'échelle 2^{j-1}, alors que le coefficient $\mathbf{V}_J$ est égal à $\bar{\mathbf{Y}}/\sqrt{N}$.

Dans notre cas, les profils de production sont de taille $N = 64$. Chaque courbe est donc décrite par un vecteur $\mathbf{W}_{\cdot i}$ de longueur 64, formé des coefficients d'ondelettes. Ce vecteur se décompose en 6 blocs correspondant à des niveaux de résolution différents, pour la courbe i :

$$\mathbf{W}_{\cdot i} = \left(\mathbf{V}'_{6i}, \mathbf{W}'_{6i}, \ldots, \mathbf{W}'_{1i}\right)'.$$

Les familles de modèles de mélanges gaussiens utilisées par notre méthode reposent sur la distinction entre le vecteur des variables de classification qui a la distribution d'un mélange gaussien multivarié et le vecteur du reste des variables qui a la distribution d'un vecteur gaussien centré réduit. De façon à faciliter l'ajustement d'une distribution de vecteur gaussien sur le second bloc, toutes les variables de coefficients d'ondelettes sont préalablement centrées et réduites. Soit $\widetilde{\mathbf{W}}_{\cdot i}$ le vecteur des coefficients d'ondelettes centrés et réduits, celui-ci se décompose encore ainsi

$$\widetilde{\mathbf{W}}_{\cdot i} = \left(\widetilde{\mathbf{V}}'_{6i}, \widetilde{\mathbf{W}}'_{6i}, \ldots, \widetilde{\mathbf{W}}'_{1i}\right)'.$$

Variables techniques. De façon à déterminer quelles variables "techniques" sont cohérentes avec la classification obtenue, nous considérons les trois variables descriptives suivantes.

- **Réserves** du gisement. Pour les champs norvégiens, des estimations sont disponibles dans la même base que celle fournissant les courbes de production. Concernant le côté anglais, nous utilisons les bases IHS et Wood Mackenzie.
- **Profondeur** du haut du réservoir. Certains champs sont composés de plusieurs réservoirs, dans ce cas la variable correspond au réservoir enfoui le moins profondément (Source utilisée : base IHS).
- **Densité** des hydrocarbures du gisement. Celle-ci se mesure en degrés API ; plus un brut est léger (plus sa densité est faible), et plus son indice API est élevé. Le degré API peut varier à l'intérieur d'un même gisement, en particulier si celui-ci est composé de plusieurs réservoirs. L'indice API utilisé ici provient de la base IHS et correspond à un indice moyen sur l'ensemble du gisement.

Notons $\mathbf{y}_{\mathrm{res}\,i}$, $\mathbf{y}_{\mathrm{pro}\,i}$ et $\mathbf{y}_{\mathrm{den}\,i}$ les trois variables techniques décrites ci-dessus, centrées et réduites.

6.2.4.2 Classification de courbes de production

Afin d'obtenir une classification de courbes qui ne soit pas influencée par les variables techniques, nous ne considérons pour commencer qu'une base de données constituée des coefficients d'ondelettes. Nous n'allons cependant pas utiliser les vecteurs d'ondelettes $\widetilde{\mathbf{W}}_{\cdot i}$ complets pour appliquer la méthode de la pente. En effet, utiliser tous les coefficients d'ondelettes conduirait à considérer des modèles de très grandes dimensions vis à vis de la taille de l'échantillon, et dans l'étape d'estimation de nombreux modèles ne parviennent pas à être estimés. Nous retirons les variables d'ondelettes correspondant aux deux résolutions les plus fines ($j = 1$ et $j = 2$). Puisque nous souhaitons classer les courbes en fonction de leur forme générale, il est naturel de retenir pour l'analyse les coefficients d'ondelettes associés aux échelles les plus larges. De plus, nous allons voir que pour construire la classification, 6 variables parmi les 16 retenues pour l'analyse sont suffisantes, ce qui nous conforte dans notre choix d'oublier les autres coefficients de la décomposition. Enfin, notons qu'en se ramenant à des observations de dimension $Q = 16$, nous diminuons significativement les temps de calculs nécessaires pour effectuer la procédure complète de la méthode de la pente. Nous utilisons dans la procédures les vecteurs observés

$$\tilde{\mathbf{y}}_{\cdot i} = \left(\widetilde{\mathbf{V}}'_{6i}, \widetilde{\mathbf{W}}'_{6i}, \ldots, \widetilde{\mathbf{W}}'_{3i}\right)'.$$

Dans la suite la notation $\mathbf{y}_q$ désigne la q-ème variable de coefficient d'ondelette centrée et réduite.

Idéalement, nous souhaiterions considérer une collection de modèles non ordonnée pour que tous les choix de blocs de variables de classification soient possibles. Malheureusement la collection serait alors trop riche pour que la procédure de sélection de modèle soit effectuée dans un temps de calcul réaliste ; il nous faut donc trouver une façon d'ordonner les variables pour nous ramener à une collection de modèles ordonnée. Une idée naturelle est d'ordonner les variables selon leur résolution d'abord, puis selon le temps à résolution égale, c'est à dire en conservant l'ordre des variables à l'intérieur des vecteurs $\tilde{\mathbf{y}}_{\cdot i}$. Nous allons utiliser une seconde solution, qui consiste à ranger les variables par ordre décroissant de leur adéquation à une variable gaussienne centrée réduite, et permet ainsi de sélectionner des modèles de dimension inférieure. Pour cela nous effectuons pour chaque variable d'ondelettes un test d'adéquation à une variable gaussienne centrée réduite, et nous ordonnons finalement les variables par ordre décroissant des statistiques de test. On note $\mathbf{y}_{\cdot i}$ les vecteurs obtenues en permutant les variables de cette façon dans chacun des vecteurs $\tilde{\mathbf{y}}_{\cdot i}$.

Nous utilisons la collection de modèles $\mathcal{M}[LB_k]$, ce choix permet de disposer d'une famille de distributions suffisamment riche tout en limitant les risques de dégénérescence lors de la procédure d'estimation. Notons que l'hypothèse d'indépendance portant sur les coefficients d'ondelettes est en partie justifiée car les coefficients d'ondelettes d'un même niveau de résolution sont peu corrélés. Rappelons que les résultats démontrés au chapitre 5 ne nécessitent de toutes façons pas que la "vraie densité" soit dans la collection de modèles utilisée.

La figure 6.1 illustre l'application de la méthode de la pente sur l'échantillon de vecteurs $(\mathbf{y}_{\cdot 1}, \ldots, \mathbf{y}_{\cdot 180})$. Notons que le nuage de points de la figure 6.1 présente dans les grandes dimensions le comportement linéaire attendu. La régression robuste permet de calibrer la

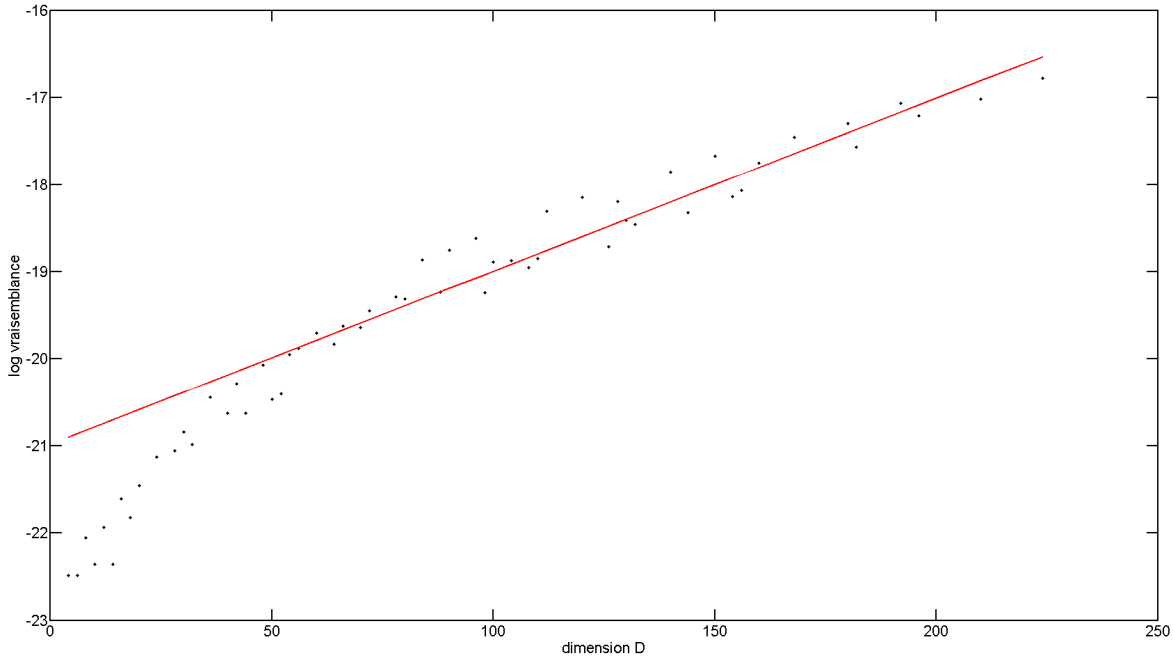


FIG. 6.1: Méthode de la pente appliquée à la base de données composée des variables d'ondelettes (sans les variables explicatives).

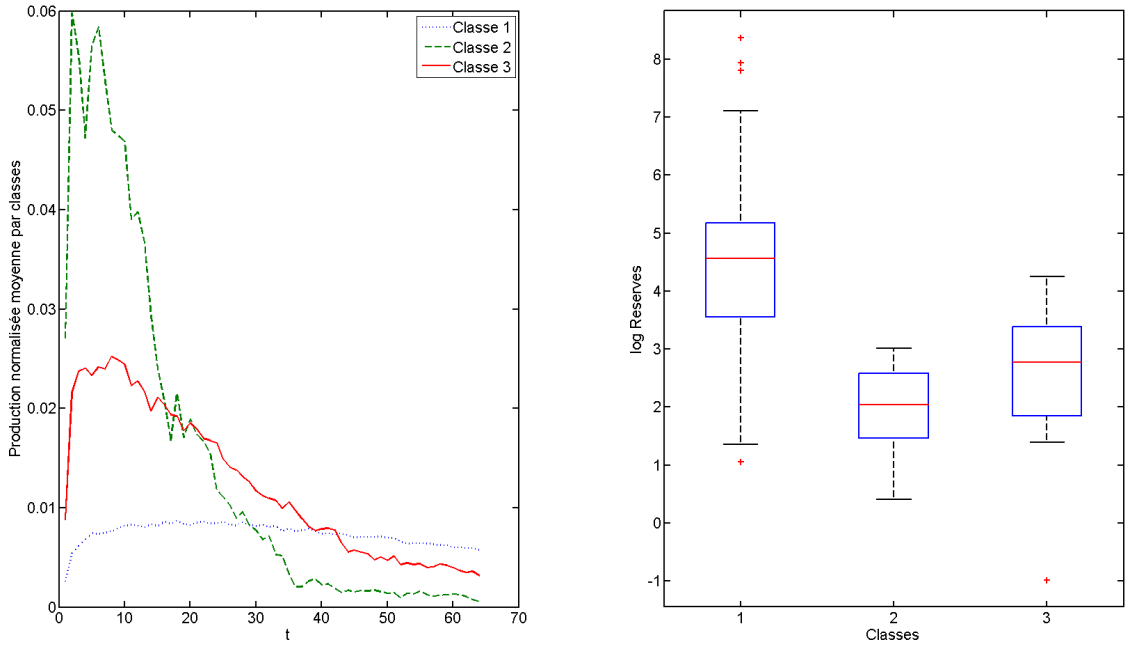


FIG. 6.2: Classification obtenue (sans variables techniques) : profils normalisés moyens de chacune des 3 classes (à gauche) et boxplots du logarithme des tailles de gisements pour chacune des trois classes (à droite).

pénalité, et le modèle finalement sélectionné est celui comportant 6 variables informatives et 3 composantes dans le mélange. Le tableau suivant donne les effectifs de la classification déduite de la règle du maximum a posteriori appliquée à la partie mélange de la densité $\hat{s}_{(\hat{K},\hat{v})}$.

Groupes	1	2	3
Effectifs	147	8	25

Nous pouvons ensuite retrouver quelle forme de profil moyen correspond à chacune des classes obtenues. La figure 6.2 (à gauche) permet de comparer les profils normalisés moyens de chacune des 3 classes. Ces groupes correspondent de façon naturelle à des taux de productions différents. Sur le graphique de droite, les boîtes à moustaches du logarithme de la variable "Réserves" montrent que la première classe correspond surtout aux gisements les plus gros alors que la deuxième contient plutôt des champs de petites tailles et la troisième des gisements de tailles intermédiaires. Nous retrouvons ainsi le principe évoqué dans la présentation du problème à la section 4.4.1 du chapitre 4.

6.2.4.3 Classification de courbes et sélection de variables techniques

Nous souhaitons maintenant déterminer quelles variables techniques expliquent la classification de courbes. Pour cela, nous utilisons notre procédure pour effectuer simultanément une classification de courbes, une sélection de variables d'ondelettes et une sélection de variables techniques. Nous ajoutons aux 16 variables d'ondelettes les 3 variables techniques elles aussi centrées réduites, nous disposons maintenant de 180 individus et de 19 variables. Le vecteur correspondant à l'observation du gisement i est de la forme $(\mathbf{y}'_{\cdot i}, \mathbf{y}_{\mathrm{res}\, i}, \mathbf{y}_{\mathrm{pro}\, i}, \mathbf{y}_{\mathrm{den}\, i})'$.

Une nouvelle collection de modèles est définie de la façon suivante. Pour un modèle, nous notons $\mathbf{v}$ l'ensemble des indices des variables de classification. La collection de modèles que nous utilisons est composée des modèles pour lesquels les variables de classification sont précisées par des ensembles d'indices $\mathbf{v}$ de la forme

$$\mathbf{v} = \mathbf{v}_{\mathrm{tec}} \cup \{1, \ldots, v'\},$$

où $\mathbf{v}_{\mathrm{tec}}$ est choisi dans l'ensemble des parties de $\{\mathrm{res}, \mathrm{pro}, \mathrm{den}\}$. Ainsi, vis à vis des variables techniques, tous les modèles possibles sont considérés. On note encore v le cardinal de $\mathbf{v}$. Pour un couple (K, v) donné, il existe plusieurs modèles à K composantes et v variables de classification dans la famille.

L'application de la méthode de la pente est illustrée par la figure 6.3, et la figure 6.4 permet de vérifier que la nappe des paramètres $(K, v) \mapsto D(K, v)$ s'ajuste bien sur la nappe des log-vraisemblances $(K, v) \mapsto \gamma_n(\hat{s}_{(K,v)})$. La procédure aboutit à la sélection du modèle à 4 composantes et 7 variables informatives, dont la variables "Réserves" et 6 variables d'ondelettes. Ceci confirme que la variable "Réserves" est cohérente avec la classification globale, alors que les variables Profondeur et Gravité ne sont pas sélectionnées. Nous avons ainsi validé le principe selon lequel la forme d'un profil normalisé peut s'expliquer grâce à la quantité de d'hydrocarbures qu'il contient. Nous obtenons cette fois les effectifs suivants pour la classification.

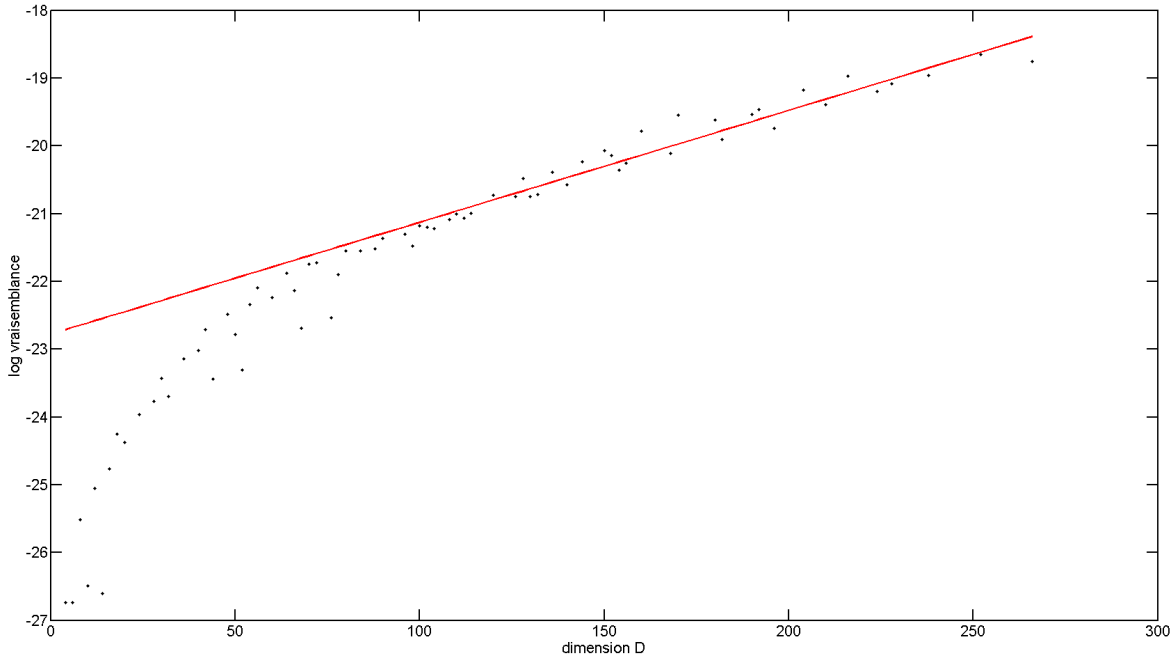


FIG. 6.3: Méthode de la pente appliquée à la base de données composée des variables d'ondelettes et des variables explicatives.

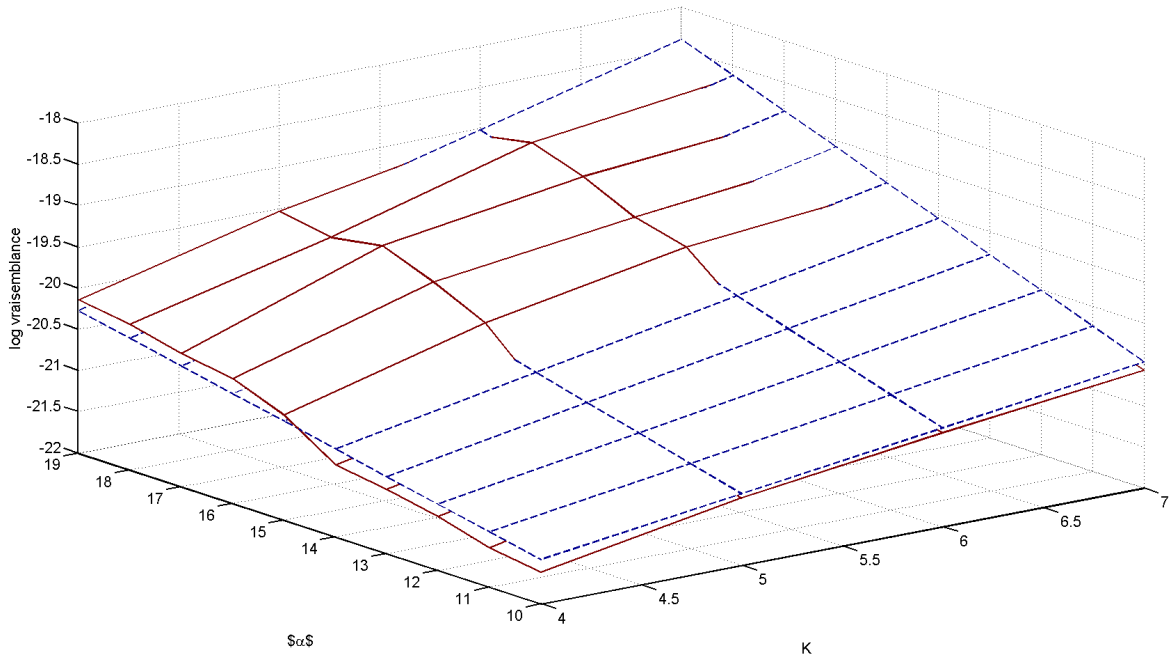


FIG. 6.4: Ajustement de la nappe des paramètres sur la nappe de la log-vraisemblance pour les modèles de grandes dimensions.

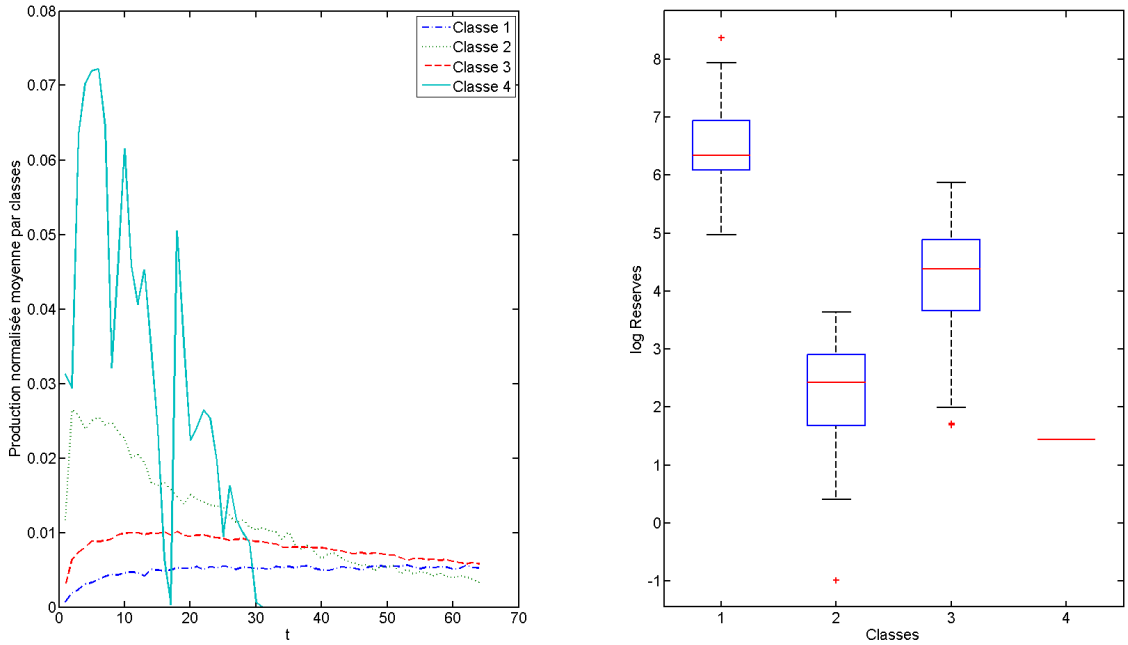


FIG. 6.5: Classification obtenue (avec variables techniques) : profils normalisés moyens de chacune des 4 classes (à gauche) et boxplots du logarithme des tailles de gisements pour chacune des quatre classes (à droite)

Groupes	1	2	3	4
Effectifs	23	37	119	1

La classe 4 ne contient qu'un seul profil, dont la production est très rapide. Notons que le fait de retirer l'unique courbe de la classe 4 de l'échantillon étudié ne permet pas de diminuer le nombre de classes sélectionnées. L'incorporation de la variable "Réserves" dans la procédure a pour effet de renforcer le regroupement par classes de tailles, comme le confirme la figure 6.5.

6.2.5 Discussion

L'étude des profils pétroliers ainsi que les simulations détaillées en annexe C illustrent les bonnes propriétés de cette nouvelle méthode de classification et de sélection de variables simultanés, ce qui vient renforcer les résultats théoriques obtenus au chapitre précédent. Notons qu'il serait possible de démontrer des résultats similaires pour d'autres types de mélanges gaussiens que les trois formes utilisées ici. Nous disposons ainsi d'un large choix de collections de modèles et il est facile de vérifier en pratique si une collection de mélanges est adaptée au problème considéré en vérifiant que la nappe des paramètres s'ajuste bien sur la nappe de la log-vraisemblance dans les grandes dimensions. L'exemple des waveforms présenté dans l'annexe C.4 illustre ce propos. L'application de la méthode de la pente nécessite d'estimer suffisamment de modèles pour lesquels le "régime" linéaire de $\gamma_n(\hat{s}_D)$ est atteint. Lorsqu'un grand nombre de composantes ou de variables de classification est nécessaire pour observer le comportement linéaire $\gamma_n(\hat{s}_D)$, il est alors conseillé d'utiliser une collection de modèles plus riche afin d'appliquer la méthode de la pente plus facilement.

Notons que notre méthode peut aussi être appliquée à la classification non supervisée sans sélection de variables. Les résultats théoriques obtenus au chapitre précédent suggèrent alors de considérer des pénalités proportionnelles à la dimension, qui est dans ce cas une fonction de K uniquement. En considérant une collection de modèles $\left(S_{(K,Q)}\right)_{K\geq 2}$, le critère proposé permet de sélectionner le nombre de classes, et il est alors un concurrent possible des critères BIC ou ICL dans un cadre plus classique d'utilisation de ces critères.

Concernant l'interprétation du modèle sélectionné par notre méthode, il est important de rappeler que l'optimisation du critère pénalisé ne correspond pas exactement avec la recherche simultanée d'une classification des données et du bloc de toutes les variables cohérentes avec celles-ci. En effet, nous avons décidé de privilégier comme objectif la minimisation du risque d'estimation. En conséquence, l'optimisation du critère pénalisé n'aboutit pas nécessairement à la sélection de toutes les variables cohérentes avec la classification retenue. Par exemple, dans une situation où Q est de l'ordre de n, certaines variables, qui sont pourtant cohérentes avec la classification ne seront pas retenues de façon à limiter l'erreur d'estimation dans le modèle sélectionné. Dans le cas d'une collection non ordonnée de modèles, nous pouvons cependant affirmer qu'il n'existe pas de bloc de variables $\mathbf{v}$ de même taille que $\hat{\mathbf{v}}$ telle que la vraisemblance de $s_{(\hat{K},\mathbf{v})}$ soit supérieure à celle de $\hat{s}_{(\hat{K},\hat{\mathbf{v}})}$. Dans ce sens, le bloc de variables $\hat{\mathbf{v}}$ est optimale. Dans le cas ordonné, ce n'est pas nécessairement le cas et les variables ne devront pas être rangées dans un ordre arbitraire de façon à faciliter l'interprétation de la sélection de variables .

La classification de courbes constitue un domaine d'application privilégié de notre méthode. En effet, le nombre des coefficients d'ondelettes (ou pour toute autre décomposition) peut être beaucoup plus grand que n, et dans ce cadre la sélection de variables prend tout son sens. De plus, le fait que les variables d'ondelettes n'aient pas vocation à expliquer la classification obtenue modère le problème de l'interprétation de la sélection de variables que nous avons signalé au paragraphe précédent. Le problème principal rencontré dans l'application de cette nouvelle méthode de classification de courbes est l'ordonnancement des variables, qui est obligatoire du fait des temps de calculs irréalistes dans le cas des collections non ordonnées. L'adéquation à une distribution normale centrée réduite est un critère possible pour ordonner les variables : les premières variables sont alors celles dont les distributions empiriques s'éloignent le plus de la distribution normale centrée réduite.

En ce qui concerne plus particulièrement l'étude des profils pétroliers, nous avons validé le principe selon lequel les profils normalisés ont une forme qui dépend des quantités d'hydrocarbures contenues dans le gisement. Seule la variable "Réserves" a été sélectionnée parmi les variables techniques, ce qui signifie donc que la variable "Réserves" explique mieux que les autres variables techniques la classification obtenue.

6.3 Sélection d'un modèle d'exploration pétrolière

Dans cette section, nous appliquons la méthode de la pente pour sélectionner une partition sur laquelle est estimée la fonction de visibilité des gisements. La première section détaille la procédure d'estimation sur une partition fixée et la seconde est dédiée à l'application de la

méthode de la pente sur des données de bassins réels.

6.3.1 Estimation de la fonction de visibilité sur une partition fixée

Nous reprenons les notations déjà utilisées au chapitre 4. Soit $(X_1^\star, \ldots, X_n^\star)$ l'échantillon des tailles de gisements dans l'intervalle $[x_0, x_{\max}]$ et découverts avant la date $t^\star$, et soit $(D_1^\star, \ldots, D_n^\star)$ l'échantillon des dates de découvertes correspondantes. Dans le modèle S_m associé à une partition m de $[x_0, x_{\max}]$, la log-vraisemblance a pour expression

$$\begin{aligned}
\gamma_n(g^\star) &= \gamma_n(h) \\
&= n \ln \alpha - n \ln(x_0^{-\alpha} - x_{\max}^{-\alpha}) + \sum_{i=1}^n \ln h(X_i^\star) - \sum_{i=1}^n h(X_i^\star) D_i^\star \\
&\quad -(\alpha+1) \sum_{i=1}^n \ln X_i^\star - n \ln \mathrm{P}_{\mathrm{dec}}(h).
\end{aligned}$$

Il s'agit donc de maximiser la log-vraisemblance en les paramètres $A = (a_1, \ldots, a_k)$ et b, ou par abus, en la fonction h. Le problème se ramène donc à maximiser en (A, b) la fonction q définie sur $(\mathbb{R}_+)^k \times \mathbb{R}_+^*$ par

$$q(A,b) := \sum_{i=1}^n \ln \{h(A,b)(X_i^\star)\} - \sum_{i=1}^n D_i^\star \, h(A,b)(X_i^\star) - n \ln \{\mathrm{P}_{\mathrm{dec}}(h(A,b))\}.$$

Proposition 6.3.1. *La fonction q est strictement concave et admet un unique maximum sur son domaine de définition.*

La preuve de la proposition est donnée dans la section 6.3.3. Le calcul des dérivés partielles de q montre qu'il n'existe pas de solution explicite au problème de la maximisation de cette fonction. Bien que le domaine $[0, +\infty[^k \times]0, +\infty[$ ne soit pas compact, la proposition précédente permet de définir de façon rigoureuse l'estimateur du maximum de vraisemblance $\hat{g}_m^\star$ dans le modèle S_m. La concavité de la fonction q garantie de plus que celle-ci n'admet pas de maximum local.

La fonction h est régulière en (A, b), ce qui nous place dans une situation favorable pour évaluer $\hat{g}_m^\star$ en utilisant une méthode de descente de gradient. Nous utilisons la fonction `fmincon` du logiciel `MATLAB` (Coleman et Li, 1996) qui permet de plus d'imposer les contraintes de positivité nécessaires sur les coefficients A et b. La procédure nécessite de nombreuses évaluations de la quantité $\mathrm{P}_{\mathrm{dec}}$. Cette probabilité peut être estimée en utilisant une estimation par la méthode de Monte Carlo. Les performances de cette procédure d'estimation pour une partition fixée sont étudiées dans l'annexe D.1.1 sur des échantillons simulés.

6.3.2 Méthode de la pente appliquée à l'exploration pétrolière

Nous utilisons ici la méthode de la pente en détectant le saut de dimension correspondant à la pénalité minimale (voir section 6.1.2). Dans les situations $\mathbf{H_1}$ et $\mathbf{H_2}$, le théorème 5.3.1

suggère d'utiliser une pénalité minimale de la forme

$$\mathrm{pen}_{\min}(m) = \eta \frac{D_m}{n} \ln n. \tag{6.4}$$

Adapté à notre contexte, l'algorithme proposé par Arlot (2007, p.35) se déroule de la façon suivante,

1. Déterminer pour tout η le modèle $\hat{m}(\eta)$ minimisant le critère

$$\mathrm{crit}(m) = \gamma_n(\hat{g}_m^\star) + \eta \frac{D_m}{n} \ln n. \tag{6.5}$$

2. Déterminer la constante $\eta_{\min}$ telle que les modèles sélectionnés sont de "grandes" dimensions lorsque $\eta < \eta_{\min}$ et de dimensions "raisonnables" lorsque $\eta > \eta_{\min}$.
3. Choisir le modèle $\hat{m}$ minimisant le critère (6.5) pour la pénalité $2\,\eta_{\min}$.

Pour que cette méthode puisse repérer le saut de dimension correspondant à la pénalité minimale, il est donc préférable que la famille contienne quelques modèles de grande dimension. En pratique, cette procédure est illustrée par un graphique sur lequel on porte la constante de pénalité η en abcisses et la dimension du modèle sélectionné $D(\eta)$ en ordonnées. Les simulations exposées dans l'annexe D.1.2 montrent que la procédure permet effectivement de sélectionner un modèle proche du modèle oracle.

Application à des bassins réels

Les zones pétrolières que nous étudions maintenant sont présentées dans l'annexe A. Les données de la mer du Nord et du bassin de Sirte correspondent exactement aux données utilisées pour estimer l'indice α des distributions de Lévy-Pareto (voir tableau 4.1). En revanche, les données du bassin du Delta du Niger sont limitées ici aux champs en off-shore, sans tenir compte des champs exploités en "off-shore" ultra-profond. Ce choix se justifie par le fait que les champs en "on-shore", en "off-shore" et en "off-shore" ultra profond n'ont pas été développés simultanément. Il n'est donc pas possible d'intégrer ces trois composantes dans un unique échantillon dans lequel se superposerait trois dynamiques d'exploration distinctes.

Pour chacun des trois bassins, nous avons obtenu des estimations de l'indice de Pareto α par la méthode d'estimation de Lepez au chapitre 4 (voir tableau 4.1). Nous estimons la fonction de visibilité pour les champs de tailles supérieures à x_0 en utilisant la méthode de la pente pour sélectionner une partition convenable sur laquelle h est estimée. Les figures 6.6, 6.7 et 6.8 présentent les résultats de la procédure pour chacune des zones. Le tableau 6.1 donne les paramètres des fonctions de visibilité estimées.

6.3.3 Preuve de la proposition 6.3.1

1. Commençons par montrer que la fonction q est strictement concave. Puisqu'une somme de fonction strictement concave est elle même une fonction strictement concave, il suffit de montrer que pour tout $x \in [x_0, x_{\max}]$ et tout $\tau \in [0, t^\star]$,

$$(A, b) \mapsto \ln\{h(A, b)(x)\} - \tau h(A, b)(x) - \ln \mathrm{P}_{\mathrm{dec}(A,b)}$$

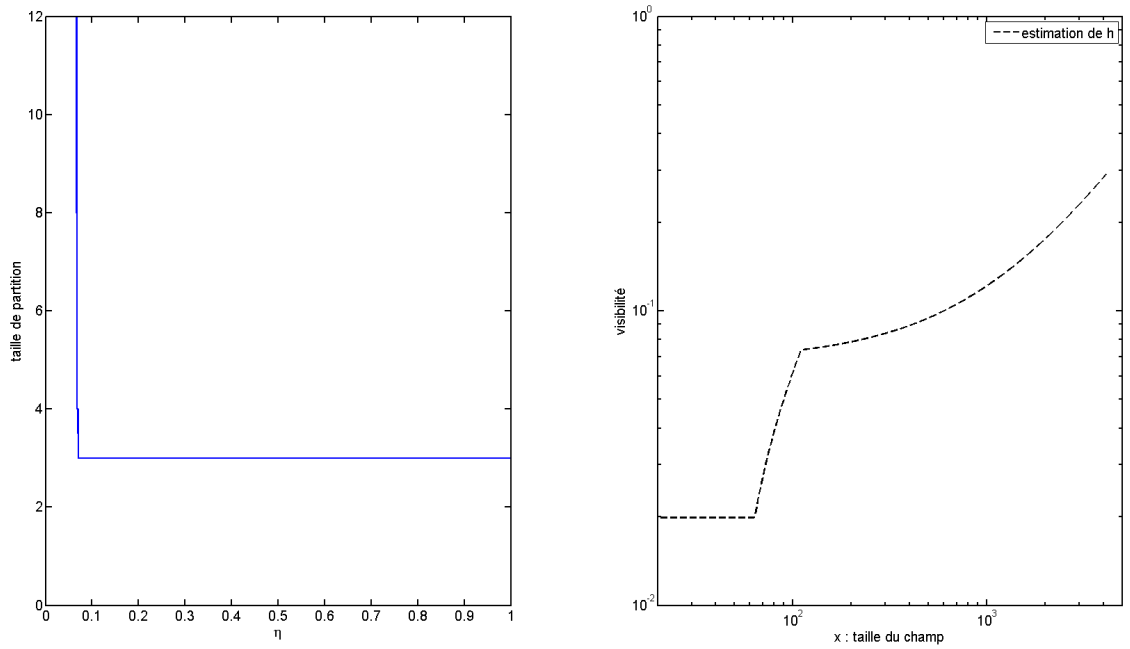


FIG. 6.6: Méthode de pente pour la province géologique du Graben de la mer du Nord. Le graphique de gauche indique la dimension du modèle sélectionné en fonction de la constante multiplicative η dans la fonction de pénalité. Le graphique de droite présente la fonction de visibilité des champs dans un diagramme à double échelle logarithmique.

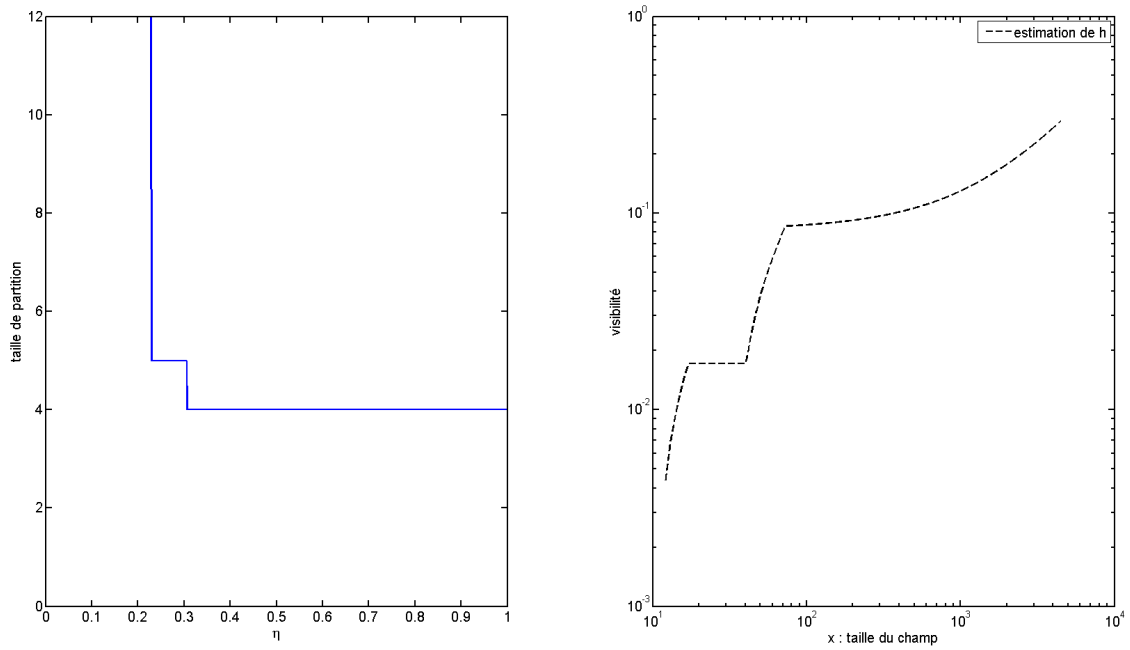


FIG. 6.7: Méthode de pente pour le bassin de Sirte. Le graphique de gauche indique la dimension du modèle sélectionné en fonction de la constante multiplicative η dans la fonction de pénalité. Le graphique de droite présente la fonction de visibilité des champs dans un diagramme à double échelle logarithmique.

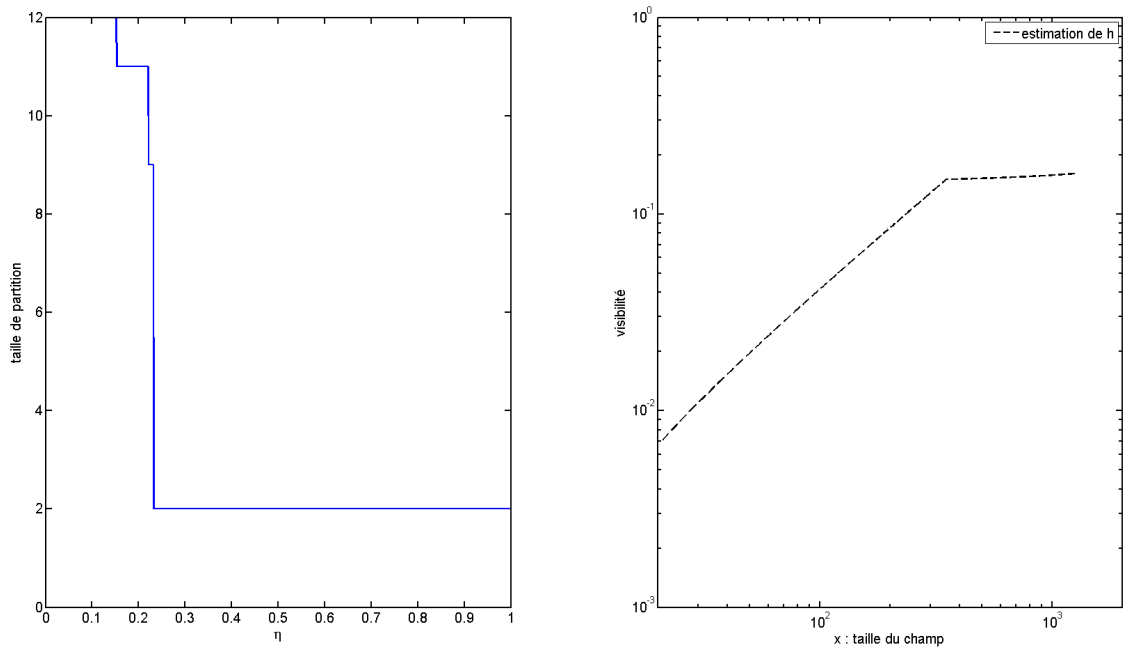


FIG. 6.8: Méthode de pente pour le bassin du Delta du Niger (Nigeria off-shore). Le graphique de gauche indique la dimension du modèle sélectionné en fonction de la constante multiplicative η dans la fonction de pénalité. Le graphique de droite présente la fonction de visibilité des champs dans un diagramme à double échelle logarithmique.

Zone	Partition Sélectionnée	Estimations de h	
		A	b
Sirte	[12 17 40.5 72.5 4482]	$2.63\,10^{-3}$ $2.69\,10^{-10}$ $2.14\,10^{-3}$ $4.73\,10^{-2}$	$4.26\,10^{-3}$
mer du N.	[20 63.5 110 4139]	$1.15\,10^{-11}$ 0.0012 $5.39\,10^{-5}$	$1.99\,10^{-2}$
Nigéria	[20 350 1250]	$4.35\,10^{4}$ $1.16\,10^{-5}$	$6.89\,10^{-3}$

TAB. 6.1: Estimations de la fonction de visibilité des champs pour trois zones de production pétrolières.

est strictement concave. Soient $x \in [x_0, x_{\max}]$ et $\tau \in [0, t^\star]$. La fonction $h(A,b)(x)$ est linéaire en (A,b), on en déduit alors que $(A,b) \mapsto \ln\{h(A,b)(x)\} - \tau h(A,b)(x)$ est strictement concave. Il suffit donc de vérifier que $(A,b) \mapsto -\ln \mathrm{P}_{\mathrm{dec}}(A,b)$ est concave. Soient (A,b) et (A',b') deux couples dans $[0,+\infty[^k \times]0,+\infty[$ et soit $\lambda \in [0,1]$. D'après la convexité de la fonction exponentielle, nous avons

$$\begin{aligned}\exp\left\{-h\left(\lambda A + (1-\lambda)A', \lambda b + (1-\lambda)b'\right)(x)t^\star\right\} &\leq \lambda \exp\{-h(A,b)(x)t^\star\} \\ &+ (1-\lambda)\exp\left\{-h(A',b')(x)t^\star\right\}.\end{aligned}$$

Nous en déduisons que

$$\begin{aligned}&\frac{\alpha x^{-\alpha-1}}{x_0^{-\alpha} - x_{\max}^{-\alpha}}\left[1 - \exp\left\{-h\left(\lambda A + (1-\lambda)A', \lambda b + (1-\lambda)b'\right)(x)t^\star\right\}\right] \\ \geq\ & \lambda \frac{\alpha x^{-\alpha-1}}{x_0^{-\alpha} - x_{\max}^{-\alpha}}\left[1 - \exp\{-h(A,b)(x)\}\right] + (1-\lambda)\frac{\alpha x^{-\alpha-1}}{x_0^{-\alpha} - x_{\max}^{-\alpha}}\left[1 - \exp\left\{-h(A',b')(x)\right\}\right].\end{aligned}$$

En intégrant cette inégalité en x sur $[x_0, x_{\max}]$, nous obtenons ainsi que

$$\mathrm{P}_{\mathrm{dec}}\left(\lambda A + (1-\lambda)A', \lambda b + (1-\lambda)b'\right) \geq \lambda \mathrm{P}_{\mathrm{dec}}(A,b) + (1-\lambda)v\mathrm{P}_{\mathrm{dec}}(A',b').$$

Finalement, nous utilisons la convexité de la fonction logarithme pour montrer que $-\ln \mathrm{P}_{\mathrm{dec}}$ est une fonction concave de (A,b), ce qui termine la preuve de la stricte concavité de q.

2. Pour montrer que q admet un maximum, commençons par montrer que q est majorée en 0. La fonction q peut se décomposer de la façon suivante,

$$q(A,b) = \sum_{j=1}^{k} q_j(a_j, b_j),$$

où b_j est défini par $b_j = \sum_{u=1}^{j-1} a_u(x_u - x_{u-1}) + b_1$ et

$$q_j(a_j, b_j) := \sum_{i|X_i^\star \in I_j} \left\{\ln h(X_i^\star) - h(X_i^\star)D_i^\star - \ln\left(\mathrm{P}_{\mathrm{dec}}\right)\right\}.$$

Pour tout $h \in S_m$, nous avons

$$\begin{aligned}\mathrm{P}_{\mathrm{dec}}(h) &= \frac{\alpha}{x_0^{-\alpha} - x_{\max}^{-\alpha}} \int_{x_0}^{x_{\max}} \left[1 - \exp\{-h(x)t^\star\}\right] x^{-\alpha-1} dx \\ &= \frac{\alpha}{x_0^{-\alpha} - x_{\max}^{-\alpha}} \sum_{j=1}^{k} \int_{x_{j-1}}^{x_j} \left[1 - \exp\{-h(x)t^\star\}\right] x^{-\alpha-1} dx\end{aligned}$$

et par croissance de la fonction h, il vient

$$\begin{aligned}
\mathrm{P}_{\mathrm{dec}}(h) &\geq \frac{\alpha}{x_0^{-\alpha} - x_{\max}^{-\alpha}} \sum_{j=1}^{k} \left[1 - \exp\left\{-h(x_{j-1})t^\star\right\}\right] \int_{x_{j-1}}^{x_j} x^{-\alpha-1} dx \\
&\geq \sum_{j=1}^{k} \frac{x_{j-1}^{-\alpha} - x_j^{-\alpha}}{x_0^{-\alpha} - x_{\max}^{-\alpha}} \left[1 - \exp\left\{-h(x_{j-1})t^\star\right\}\right].
\end{aligned}$$

De plus, en utilisant l'inégalité, $1 - \exp u \geq u \exp(-u)$ pour $u \geq 0$, il vient

$$\mathrm{P}_{\mathrm{dec}} \geq \sum_{j=1}^{k} c_j \, h(x_{j-1}) \exp\left\{-h(x_{j-1})t^\star\right\},$$

où les c_j sont des constantes strictement positives qui ne dépendent pas des a_j et des b_j. Ceci permet de majorer q_j pour tout j de la façon suivante,

$$\begin{aligned}
q_j(a_j, b_j) &\leq \sum_{i|X_i^\star \in I_j} \left\{ \ln \frac{h(X_i^\star)}{\sum_{r=1}^{k} c_r h(x_{r-1}) \exp\left\{-h(x_{r-1})t^\star\right\}} - h(X_i^\star) D_i^\star \right\} \\
&\leq \sum_{i|X_i^\star \in I_j} \left\{ \ln \frac{h(X_i^\star)}{c_j h(x_{j-1}) \exp\left\{-h(x_{j-1})t^\star\right\}} - h(X_i^\star) D_i^\star \right\} \\
&\leq \sum_{i|X_i^\star \in I_j} \left\{ \ln \frac{h(X_i^\star)}{h(x_{j-1})} + h(x_{j-1})t^\star - \ln c_j - h(X_i^\star) D_i^\star \right\} \\
&\leq \sum_{i|X_i^\star \in I_j} \left\{ \ln \frac{h(X_i^\star)}{h(x_{j-1})} + h(x_{j-1})t^\star - \ln c_j - h(X_i^\star) D_i^\star \right\} \\
&\leq \sum_{i|X_i^\star \in I_j} \left\{ \ln \left(1 + \frac{X_i^\star - x_{j-1}}{x_{j-1}}\right) + h(x_{j-1})t^\star - \ln c_j \right\}.
\end{aligned}$$

La dernière inégalité est obtenue en utilisant le fait que $h(x) = a_j(x - x_{j-1}) + b_j$ sur l'intervalle I_j. On note de plus que $h(x)$ tend vers 0 lorsque (A, b) tend vers 0, ce qui permet de montrer que q_j est majorée en 0.

Ensuite, on vérifie facilement pour pour tout j, la fonction q_j tend vers $-\infty$ lorsque $\|(A, b)\| \to +\infty$. En effet, pour tout i tel que $X_i^\star \in I_j$, on a

$$\ln\left[a_j(X_i^\star - x_{j-1}) + b_j\right] - \left[a_j(X_i^\star - x_{j-1}) + b_j\right] \|(A, b)\| \overrightarrow{\to} +\infty \; -\infty.$$

On a de plus $h \to +\infty$, uniformément sur $[x_0, x_{\max}]$, lorsque $\|(A, b)\| \to +\infty$ ce qui implique que $\lim\limits_{\|(A,b)\| \to +\infty} \mathrm{P}_{\mathrm{dec}} = 1$, d'où la limite q_j annoncée ci-dessus.

La fonction q est continue, majorée en 0 et tend vers $-\infty$ lorsque $\|(A, b)\|$ tend vers $+\infty$. On en déduit donc que celle-ci admet un maximum sur $[0, +\infty[^k \times]0, +\infty[$. L'unicité du maximum provient enfin de la stricte concavité de la fonction q.

6.4 Conclusion du chapitre

Dans ce chapitre, nous avons développé les procédures permettant d'exploiter en pratique les résultats théoriques obtenus dans le chapitre précédent. Concernant l'étude des profils pétroliers, nous avons montré que les profils normalisés se déforment en fonction des réserves ultimes du gisement, en utilisant une méthode originale de classification de courbes. Ceci justifie l'introduction au chapitre suivant d'un modèle pour la production individuelle des gisements dans lequel la production dépend directement de la quantité de pétrole contenue dans le gisement considéré. Quant au problème de l'exploration pétrolière, l'estimation de la visibilité, qui d'un point de vue statistique constituait la difficulté principale, va nous permettre dans la suite de proposer des prolongements du processus d'exploration.

Chapitre 7

Profils de production d'un bassin pétrolier

Ce chapitre est consacré à l'étude des courbes de production de bassins pétroliers. Afin de pouvoir modéliser complètement la production d'un bassin, nous proposons tout d'abord un modèle élémentaire de la production individuelle des gisements qui s'appuie sur les conclusions du chapitre précédent. La section 2 est consacrée à la définition du modèle de production de bassin et à l'étude de l'impact de différents facteurs sur la forme de la courbe de production. La section finale propose des prolongements de processus d'exploration et de courbes de production pour des bassins connus.

7.1 Modèle de production individuelle des gisements pétroliers

La classification des profils de production exposée dans le chapitre précédent nous a révélé que la forme d'un profil de production normalisé s'expliquait principalement par la taille du gisement correspondant. Nous proposons maintenant un modèle paramétrique simple de la production individuelle, comme fonction du temps et de la variable Réserves.

7.1.1 Présentation du modèle

Les classes de profils obtenues au chapitre précédent (voir figure 6.5) suggèrent que les gisements les plus riches en réserves sont produits de façon beaucoup plus lente que les champs plus petits. La figure 7.1 illustre encore ce phénomène. Pour un échantillon de courbes de production en mer du Nord, les durées de production des champs, et les niveaux des pics de production sont portés dans ces deux graphiques en fonction des réserves contenues dans chacun des gisements, selon des échelles logarithmiques. Notons que 90% des gisements correspondants sont toujours en production, et les durées de production considérées ne sont majoritairement que des estimations fournies par la base de données. En revanche, le pic de production d'un gisement est atteint rapidement, et très peu de pics de production nécessitent d'être estimés dans cet échantillon. Il est facile d'ajuster une droite sur le nuage de points du graphique inférieur, ce qui suggère que le niveau du pic est une puissance de la variable Réserves. L'ajustement est de moins bonne qualité sur le graphique supérieur, mais comme

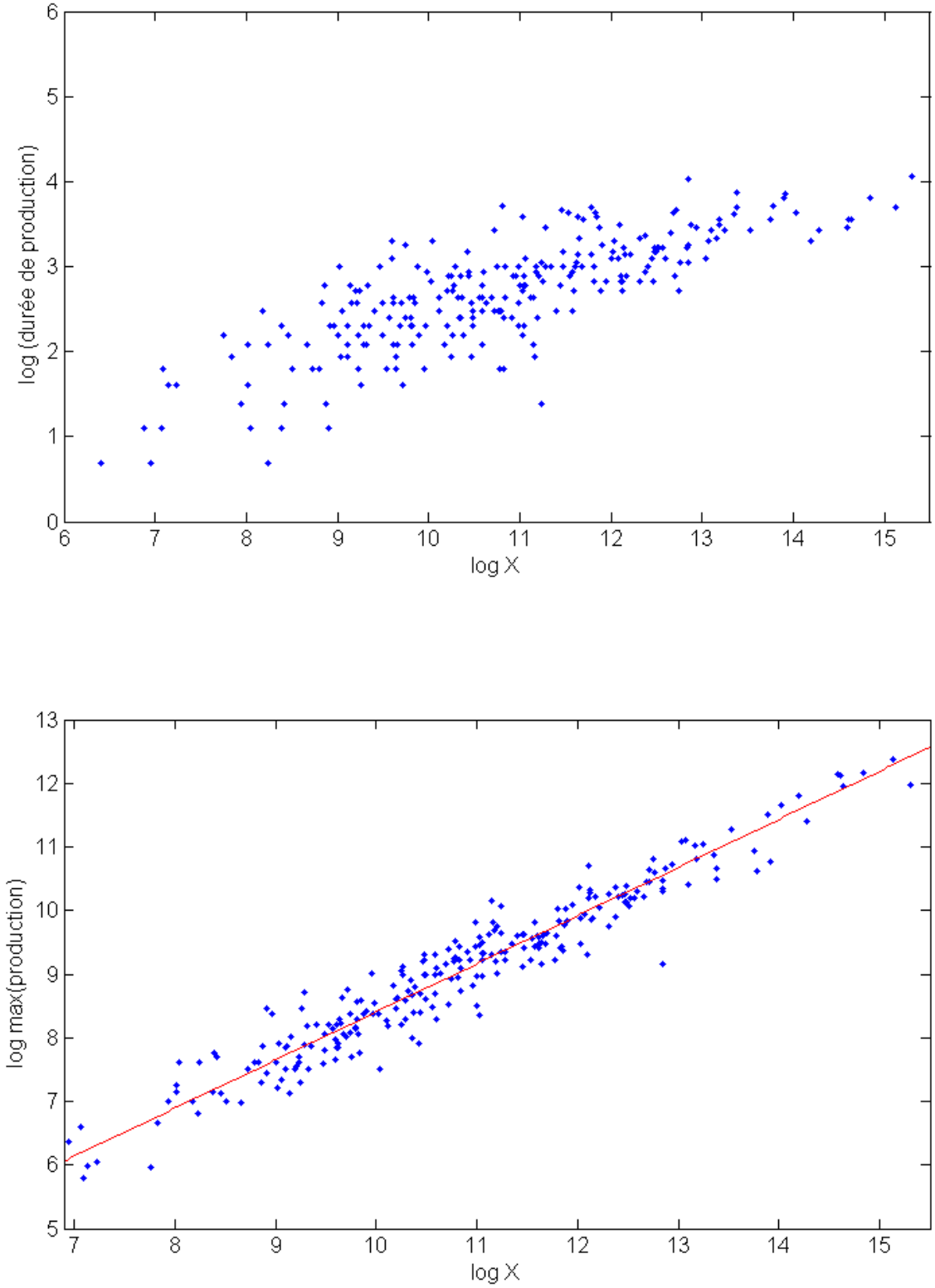


FIG. 7.1: Logarithmes des durées de production (graphique supérieur) et logarithmes des productions maximales (graphique inférieur) en fonction des réserves des gisements, pour un échantillon de 250 champs productifs en mer du Nord. Données issues de la base Wood MacHenzie 2004, les longueurs des profils sont des estimations fournies par la base.

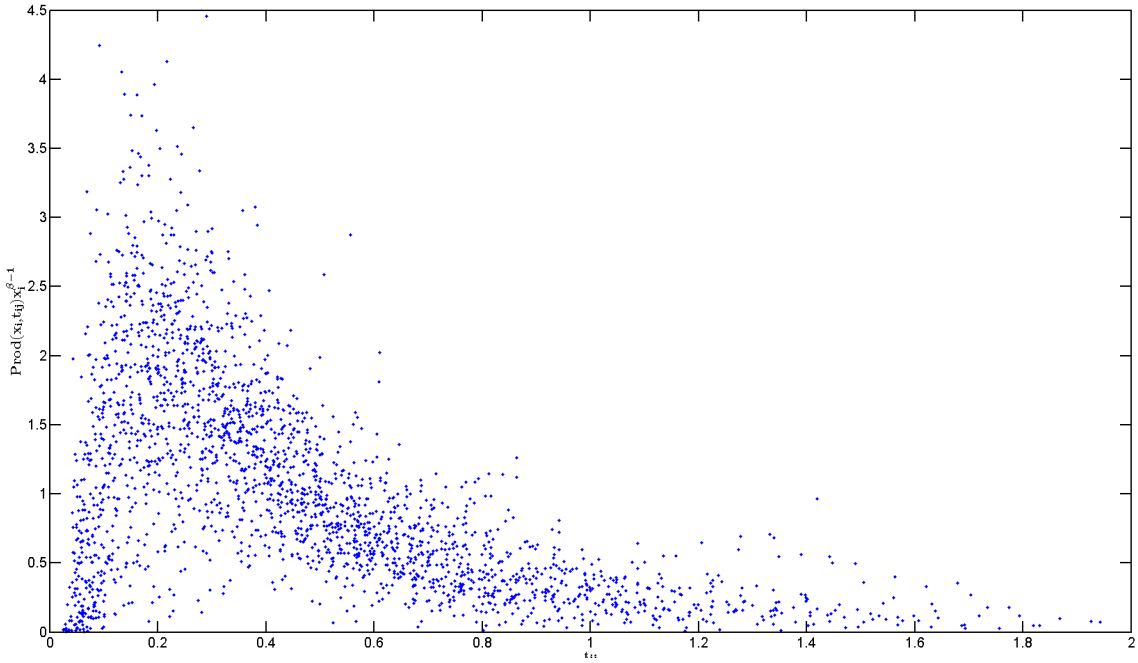


FIG. 7.2: Quantités $x_i^{\beta-1}\text{prod}(x_i, t_{ij})$ en fonction des points $\frac{t_{ij}}{x_i}$ pour l'ensemble des gisements en mer du Nord. La quantité $\text{prod}(x_i, t_{ij})$ correspond à la production d'un champ de taille x_i au temps t_{ij}.

nous venons de l'expliquer, les données utilisées dans ce cas sont surtout des estimations. On vérifie tout de même qu'un champ produit d'autant plus longtemps que celui-ci renferme des réserves importantes.

La taille du gisement, notée x, ainsi que le temps t, sont ici des variables déterministes. Soit $\text{prod}(x, t)$ la variable aléatoire de la production au temps t d'un gisement de taille x. Les remarques précédentes nous conduisent à supposer que

$$\mathbb{E}\left(\frac{1}{x}\text{prod}(x,t)\right) = \frac{1}{x^\beta}\,\text{K}\left(\frac{t}{x^\beta}\right), \tag{7.1}$$

où la fonction K est appelée profil élémentaire. Le coefficient β, que nous appelons coefficient d'inertie, contrôle la vitesse de production des champs en fonction de leur taille. La figure 7.2 permet de valider graphiquement cette modélisation. Sur celle-ci, nous avons porté les quantités $x_i^{\beta-1}\text{prod}(x_i, t_{ij})$ en fonction des points $\frac{t_{ij}}{x_i}$. La quantité $\text{prod}(x_i, t_{ij})$ correspond à la production du i-ème champ au temps t_{ij}. Le coefficient d'inertie β a été estimé de façon grossière en utilisant la régression linéaire effectuée sur le nuage de points du graphique inférieur de la figure 7.1, ce qui donne $\beta = 0.25$. La forme du profil élémentaire K se dessine clairement à travers le nuage de points $\left(x_i^{\beta-1}\text{prod}(x_i, t_{ij}), \frac{t_{ij}}{x_i}\right)$.

La figure 7.2 suggère que le bruit agit de façon multiplicative sur la variable prod. Nous aboutissons ainsi au modèle de production suivant :

$$\frac{1}{x}\text{prod}(x,t) = \frac{1}{x^\beta}\,\text{K}\left(\frac{t}{x^\beta}\right)\tilde{\epsilon}_{x,t} \tag{7.2}$$

où $\epsilon_{x,t} := \log(\tilde{\epsilon}_{x,t})$ une variable gaussienne centrée de variance σ^2.

Supposer que le bruit agit additivement et non pas multiplicativement ne modifie que peu l'estimation de β. En revanche, les résidus que l'on obtiendrait pour le modèle additif présentent une plus forte hétéroscédasticité que pour le cas multiplicatif.

7.1.2 Principe d'estimation

En passant au logarithme[1]dans (7.2), le modèle s'écrit sous la forme suivante

$$\ln \frac{\mathrm{prod}(x,t)}{x} = -\beta \ln x + \mathrm{H}\left(\frac{t}{x^\beta}\right) + \epsilon_{x,t}, \tag{7.3}$$

avec $\mathrm{H} := \ln \mathrm{K}$. Les productions du gisement i (de taille x_i) sont connues aux temps $t_{i1}, \ldots, t_{ij}$, $\ldots, t_{iq(j)}$, où $q(j)$ correspond au nombre d'années pendant lesquelles le gisement a produit. A chaque couple possible (i,j), nous associons un nouvel indice $u = u(i,j)$. Nous pouvons alors écrire (7.3) pour les observations disponibles sous la forme suivante

$$Z_u = -\beta \ln x_u + \mathrm{H}\left(\tau_u^{(\beta)}\right) + \epsilon_u, \tag{7.4}$$

où $Z_u := \ln \frac{\mathrm{prod}(x_i,t_{ij})}{x_i}$, $x_u := x_i$, et $\tau_u^{(\beta)} := \frac{t_{ij}}{x_i^\beta}$. Les variables aléatoires ϵ_u sont i.i.d. de loi gaussienne centrée de variance σ^2. Cette indexation ne fait plus de distinction entre les différents gisements et nous permet de nous replacer dans un contexte d'estimation statistique plus classique.

Pour estimer H dans le modèle de régression (7.3), nous utilisons une méthode d'interpolation spline. Une fonction *spline cubique* est une fonction continue polynomiale par morceaux de degré au plus 3, définie sur une partition d'intervalles délimités par des noeuds, et telle que celle-ci ait ses dérivés première et seconde continues. Un problème courant dans l'utilisation des splines cubiques est leur tendance à "exploser" aux deux extrémités de leur intervalle de définition. Les fonctions *splines cubiques naturelles* permettent de résoudre ce problème en imposant à la fonction d'avoir de plus un comportement linéaire au-delà des deux extrémités. Pour un ensemble de k noeuds fixés, l'ensemble des fonctions splines cubiques naturelles définies sur ces noeuds forme un espace vectoriel de dimension k. Le lecteur pourra se référer à Hastie *et al.* (2001) pour une description détaillée de l'interpolation spline.

Dans la suite, nous supposons que le nombre k de noeuds est fixé, et que les noeuds ξ sont régulièrement espacés sur $[0, \tau_{\max}]$, où $\tau_{\max}$ est une valeur maximale pour les τ_u observés. Soit $m = (m_r)_{r=1,\ldots,k}$ une base de l'espace $\mathcal{S}_{cn}(\xi)$ des fonctions splines naturelles construites sur les noeuds ξ. Une fonction $\mathrm{H} \in \mathcal{S}_{cn}(\xi)$ s'écrit donc sous la forme $\mathrm{H} = \sum_{r=1}^k \theta_r\, m_r$, avec $\theta \in \mathbb{R}^k$.

De façon à éviter que la fonction estimée sur-ajuste les observations, le critère pénalisé des

[1]Cette transformation exclut donc les production nulles que l'on observe lorsque l'exploitation est terminée. Pour l'échantillon de mer du Nord utilisé ici, seules les fins de production d'une vingtaine de gisements sont concernées, et l'impact est donc très faible sur l'estimation de K. Pour éviter d'exclure ces données, il serait toujours possible d'utiliser le modèle avec bruit additif tout en ayant à l'esprit que cette modélisation est moins réaliste.

moindres-carrés ci-dessous est minimisé en $\mathrm{H} \in \mathcal{S}_{cn}(\xi)$ et en β,

$$\mathrm{crit}(\mathrm{H}, \beta, \lambda) = \sum_{u=1}^{N} \left\{ Z_u + \beta \ln x_u - \mathrm{H}\left(\tau_u^{(\beta)}\right) \right\}^2 + \lambda \int_0^{\tau_{\max}} \{\mathrm{H}(\tau)\}^2 \, d\tau. \tag{7.5}$$

Les fonctions H qui présentent des oscillations trop prononcées sont pénalisées par le second terme ; de façon générale la fonction choisie par ce critère est appelée *spline de lissage*. Dans ce qui suit, nous adaptons la procédure détaillée dans Hastie *et al.* (2001, p.127) à notre cas particulier mêlant estimations de H et de β. Définissons la matrice $\mathbf{M}^{(\beta)}$ par $\mathbf{M}_{ur}^{(\beta)} = m_r\left(\tau_u^{(\beta)}\right)$, posons $W_u^{(\beta)} := Z_u + \beta \ln x_u$. Le critère (7.5) peut aussi s'exprimer sous la forme matricielle suivante

$$\mathrm{crit}(\theta, \beta, \lambda) = {}^t\left(W^{(\beta)} - \mathbf{M}^{(\beta)}\theta\right)\left(W^{(\beta)} - \mathbf{M}^{(\beta)}\theta\right) + \lambda {}^t\theta \Omega^{(m)} \theta, \tag{7.6}$$

où la matrice $\Omega^{(m)}$ est définie selon la base m par

$$\Omega_{rl}^{(m)} = \int m_r''(t)\, m_l''(t)\, dt.$$

Pour des paramètres λ et β fixés, le critère (7.6) est minimum en

$$\hat{\theta}(\beta, \lambda) = \left({}^t\mathbf{M}^{(\beta)}\mathbf{M}^{(\beta)} + \lambda\Omega^{(m)}\right)^{-1} {}^t\mathbf{M}^{(\beta)} W^{(\beta)}.$$

La minimisation de (7.6) en β est effectuée de façon numérique en calculant les valeurs $\mathrm{crit}\left(\hat{\theta}(\beta), \beta, \lambda\right)$ sur une grille fine de valeurs de β ; nous notons $\hat{\beta}(\lambda)$ le β optimal correspondant. Pour ajuster le paramètre λ, il est souvent préconisé d'effectuer une validation croisée de façon à minimiser le risque de prédiction. L'échantillon utilisé ici est de longueur environ 2000. Il est suffisant et plus simple dans notre cas de séparer aléatoirement l'échantillon en deux blocs de données distincts. Le premier bloc B_1 est utilisé pour calculer les estimateurs $\hat{\theta}$ et $\hat{\beta}$ pour différentes valeurs de λ. Le risque correspondant à chaque choix de λ est ensuite estimé en utilisant le deuxième bloc de données B_2. Nous choisissons le paramètre de lissage optimal qui minimise le risque estimé

$$\hat{R}(\lambda) = \sum_{u \in B_2} \left[Z_u + \hat{\beta} x_u - \hat{\mathrm{H}}\left(\tau_u^{(\hat{\beta})}\right) \right]^2$$

où $\hat{\mathrm{H}} = \sum_{r=1}^{k} \hat{\theta}(\hat{\beta}, \lambda) m_r$.

7.1.3 Résultats

La base de donnée utilisée ici est la base Wood Mackenzie 2004. Le nombre de noeuds est fixé à $k = 50$. Il serait inutile et plus coûteux en temps de calcul de s'appuyer sur un maillage plus fin. Nous utilisons la fonction `smooth.spline` du logiciel `R` pour estimer H à β fixé. Cette procédure `R` présente l'avantage de laisser à l'utilisateur la possibilité de régler le nombre de noeuds sur lesquels s'appuie la base de splines cubiques naturelles. La régression linéaire proposée sur la figure 7.1 permet de préciser un intervalle raisonnable sur lequel rechercher

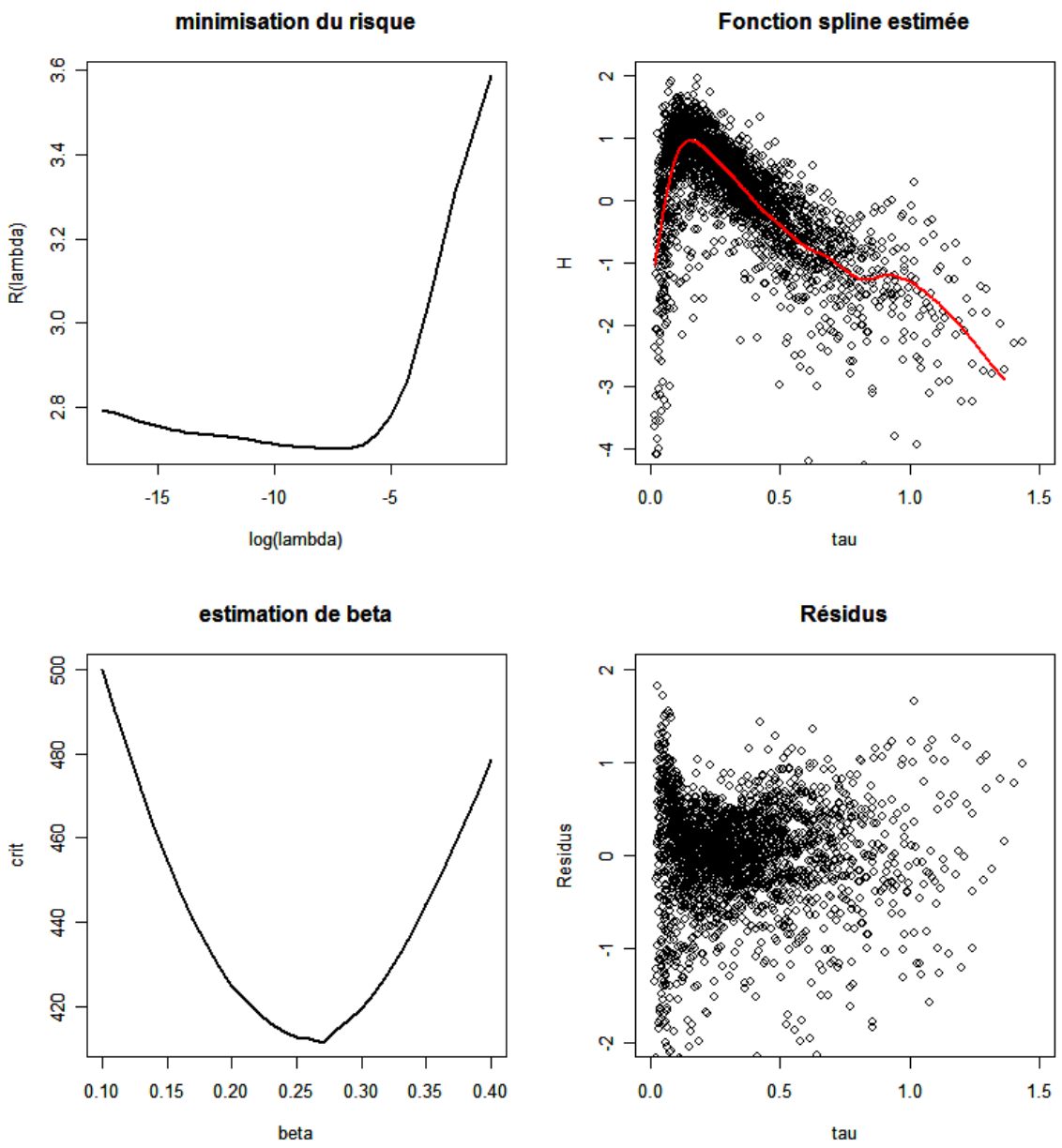


FIG. 7.3: Procédure complète d'estimation du paramètre β et de la fonction spline cubique naturelle.

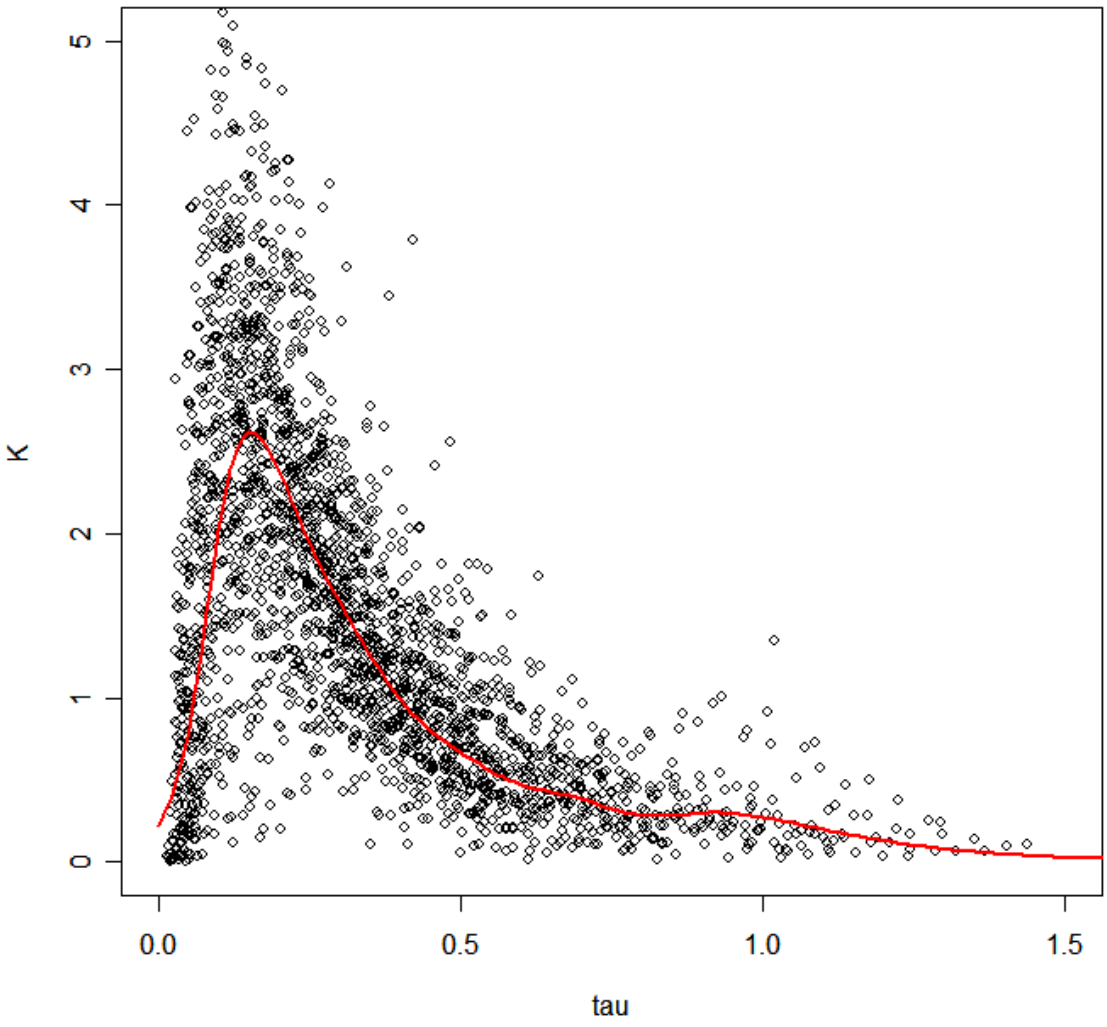


FIG. 7.4: Estimation du profil élémentaire $\hat{K}$.

$\hat{\beta}$; la minimisation du critère est effectuée pour β entre 0.1 et 0.4.

La figure 7.3 est composée de quatre graphiques qui résument les résultats obtenus. La figure en haut à gauche représente l'estimation du risque en fonction du paramètre de lissage. Juste en dessous, le critère pénalisé, pour le paramètre de lissage optimal, est calculé en fonction du paramètre β. Nous trouvons ainsi $\hat{\beta} = 0.27$, et la fonction spline $\hat{\mathrm{H}}$ correspondant à $\hat{\beta}$ est tracée sur le graphique en haut à droite. Nous avons aussi affiché les résidus correspondant au couple $(\hat{\beta}, \hat{\theta})$ retenu. Si l'on exclut la zone temporelle correspondant au début de la production, l'observation des résidus montre que ce modèle simpliste explique cependant convenablement les données observées. Selon notre modèle, le bruit est ici multiplicatif et lorsque la production est faible, la variance est théoriquement faible elle aussi. Or, les quantités observées ici sont cumulées par année de production, et le début de l'exploitation d'un gisement peut avoir lieu au début ou à la fin de la première années de production. Ceci explique en partie la variance importante des résidus en 0.

Pour finir, le profil élémentaire est estimé naturellement par $\hat{\mathrm{K}} = \exp\left(\hat{\mathrm{H}}\right)$. Cette fonction est représentée sur la figure 7.4. Il est intéressant de noter que la décroissance de cette fonction, au-delà du pic, est atténuée par un léger rebond en toute fin de production. Ceci s'explique par les techniques de récupérations secondaires et tertiaires mises en place sur une partie des gisements de l'échantillon, et qui permettent d'accéder à de nouvelles réserves dans le gisement. Il est satisfaisant que l'estimation du profil élémentaire K intègre ce phénomène par ce rebond de production.

D'après le modèle défini en (7.1), la fonction élémentaire est une fonction d'intégrale 1. Or, la procédure que nous avons détaillée ci-dessus n'impose pas cette contrainte, qui aurait nettement compliqué l'estimation de H. Pour combler cette lacune, il suffit de restreindre le support de la fonction $\hat{\mathrm{K}}$ de façon que celle-ci vérifie la contrainte. Nous obtenons finalement un couple $(\hat{\mathrm{K}}, \hat{\beta})$ que nous pourrons utiliser au chapitre suivant pour la modélisation de la production du bassin.

7.2 Profils de production de bassin pétrolier

Dans cette section, le modèle de production de bassin que nous sommes maintenant en capacité d'utiliser est complètement défini. Soit $(X_1, \ldots, X_{N_\mathcal{B}})$ l'échantillon des tailles des gisements et soit $(D_1, \ldots, D_{N_\mathcal{B}})$ l'échantillon des dates de découverte correspondantes, où $N_\mathcal{B}$ désigne le nombre total de gisements contenus dans le bassin. Pour $u \geq 1$, nous notons $(L_u)_{u\geq 1}$ la suite des dates de mise en production, et soit Y_u les réserves du gisement lancé à la date L_u. L'entier $u(i)$ désigne le rang de lancement du gisement i, et si celui-ci n'est jamais mis en production, alors par convention $u(i) = +\infty$ avec $L_\infty = +\infty$. Si $u(i) < \infty$, nous avons alors $Y_{u(i)} = X_{i(u)}$. Le stock de gisements disponibles à la date t a pour expression

$$\mathrm{Stock}_t = \left\{ i \in \{1, \ldots, N_\mathcal{B}\} \,|\, D_i \leq t < L_{u(i)} \right\}.$$

De façon général, nous appelons "politique de mise en production" une fonction de L_u et de Stock_t (et éventuellement d'autres paramètres externes) qui détermine le gisement à sélectionner dans Stock_{L_u} à la date L_u.

La synthèse suivante énumère l'ensemble des hypothèses sur lesquelles repose le modèle de la production d'un bassin .

1. Les réserves X d'un gisement du bassin suit une loi de Lévy-Paréto $\mathcal{P}\mathrm{ar}(\alpha, 1, x_{\max})$.
2. Les champs sont découverts selon le modèle d'exploration suivant :
   - les dates de découvertes des champs de taille supérieure à un seuil x_0 sont de loi conditionnelle $(D \mid X) \sim \mathcal{E}\left(h(X)\right)$ où h est la fonction de visibilité des champs,
   - les gisements de taille inférieure à x_0 sont découverts selon un processus de Poisson homogène dans le temps.
3. Les champs du stock découverts sont développés selon une "politique de mise en production".
4. Un gisement exploité produit ses réserves selon le modèle

$$\mathrm{prod}(x,t) = x^{1-\beta}\,\mathrm{K}\left(\frac{t}{x^{\beta}}\right)$$

   où le profil élémentaire K et le coefficient d'inertie β ont tous deux été estimés dans la section précédente.

Sous cet ensemble d'hypothèses, la production du bassin a pour expression

$$\mathrm{Prod}(t) = \sum_{u=1}^{\infty} Y_u^{1-\beta}\,\mathrm{K}\left(\frac{t-L_u}{Y_u^{\beta}}\right).$$

Il serait également possible de supposer que X suit une loi de Lévy-Pareto $\mathcal{P}\mathrm{ar}(\alpha)$ non restreinte. Pour que la production $\mathrm{Prod}(t)$ soit intégrable en tout t, il faudrait alors des hypothèses supplémentaires sur la loi de (X, D) et sur β. Les simulations qui suivent visent à déterminer des courbes moyennes de production, pour éviter toute discussion sur l'intégrabilité de $\mathrm{Prod}(t)$, nous supposerons donc que X est majoré. Pour les prolongement de courbes de production proposés dans la section suivante, cette hypothèse est naturelle car pour des bassins matures la taille du plus gros gisement est connue. Concernant le dernier point, dans l'objectif de déterminer des courbes de production moyennes, il n'est pas nécessaire d'incorporer de bruit multiplicatif dans le modèle production individuel des gisements que nous utilisons ici.

Nous étudions maintenant l'impact de la politique de mise en production des champs et de la géologie sur la courbe de production du bassin.

7.2.1 Impact de la politique de mise en production

Pour étudier l'influence de la gestion du stock sur la forme de la courbe de production du bassin, plusieurs scénarios de gestion sont considérés dans les protocoles de simulation qui suivent. Pour chacun des protocoles, 100 simulations sont effectuées pour évaluer un profil de production moyen du bassin.

Étude de l'impact de la politique de sélection

A chaque date L_u, un champ est choisi dans la population de Stock_{L_u} selon un certain critère. Il est difficile de modéliser fidèlement la politique de sélection d'un gisement dans

le stock. En effet les décisions de mise en production sont influencées par de nombreux paramètres économiques et géographiques alors que dans le cadre de notre modélisation, nous ne connaissons pour chaque gisement que la quantité de réserves qu'il contient et la date de sa découverte. Cependant, pour des raisons économiques évidentes, il est clair qu'il y a une grande préférence à mettre en production les gisements du stock qui possèdent les plus grandes réserves.

Protocole 1 : Nous supposons ici que la suite des dates L_u est modélisée par un processus de Poisson d'intensité λ constante dans le temps. Le tableau 7.5a détaille les caractéristiques du processus d'exploration simulé qui sont communes pour toutes les simulations effectuées. A chaque date L_u, un gisement est sélectionné dans le stock selon l'un des modes de tirage suivants,

- tirage non biaisé,
- tirage biaisé selon une puissance γ de la taille (ici $\gamma = 1$ et $\gamma = 3$),
- sélection du gisement de plus grande taille.

Les courbes de production moyennes correspondant à chacune de ces situations sont représentées sur la figure 7.5c. Notons tout d'abord que plus le tirage dans le stock avantage les grands champs, plus le pic survient tôt et avec une grande amplitude. En revanche, si le biais est faible, la production est globalement mieux répartie sur les premières décennies. De plus, nous notons que pour des puissance $\gamma \geq 1$, les profils de bassin sont quasiment identiques.

Impact de l'intensité de mise en production

Les 3 protocoles suivants permettent d'évaluer l'impact de l'intensité de mise en production des champs sur la forme des courbes de production du bassin.

Protocole 2 : Nous commencer par considérer plusieurs scénarios d'intensité constante dans le temps, pour les mêmes distributions de réserves et d'exploration (voir tableau 7.6a). Le tableau 7.1 et la figure 7.6c présentent les résultats obtenus dans chacun des scénarios. Notons que pour ces simulations, la visibilité et l'intensité μ_0 sont suffisamment élevées pour que le stock ne soit jamais épuisé au cours de l'exploitation du bassin. L'augmentation de l'intensité a un effet important sur la courbe tant que les niveaux d'intensité restent faibles. Pour des niveaux d'intensité plus grands, l'impact de l'intensité sur la courbe de production est beaucoup moins important et les productions cumulées du tableau 7.1 confirment cette observation. La distribution des réserves est à l'origine de ce phénomène. En effet, pour $\alpha = 0.75$, environ 70% des réserves sont en moyenne contenues dans l'ensemble des gisements de taille supérieure à 100Mb. Ces gisements ne sont que quelques dizaines et l'essentiel d'entre eux est découvert sur les 30 premières années. Le fait d'augmenter l'intensité de la mise en production des gisements ne peut donc que faiblement augmenter les réserves disponibles à la production. De plus, la figure 7.6c montre que l'augmentation de λ a pour effet de produire l'essentiel des réserves plus tôt, ce qui accentue naturellement le déclin de la production au-delà du pic.

$N_{\mathcal{B}} = 2500$

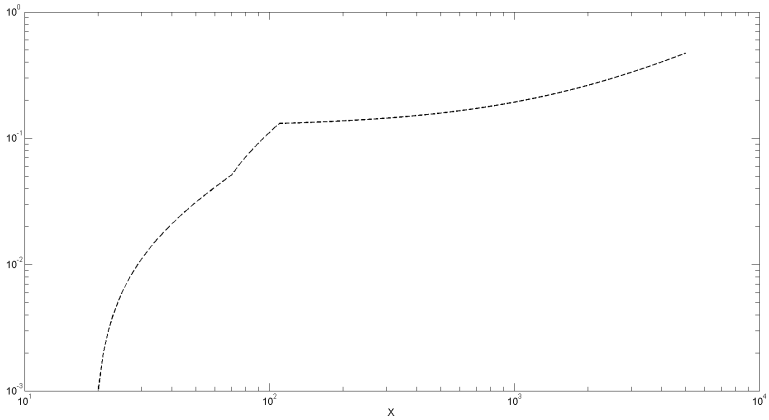
$\alpha = 0.75$

$$h : \begin{cases} \text{Partition :} & [20\,,\,70\,,\,110\,,\,5000] \\ A = & [10^{-3}\,,\,2\,10^{-3}\,,\,7\,10^{-5}] \\ b = & 10^{-3} \end{cases}$$

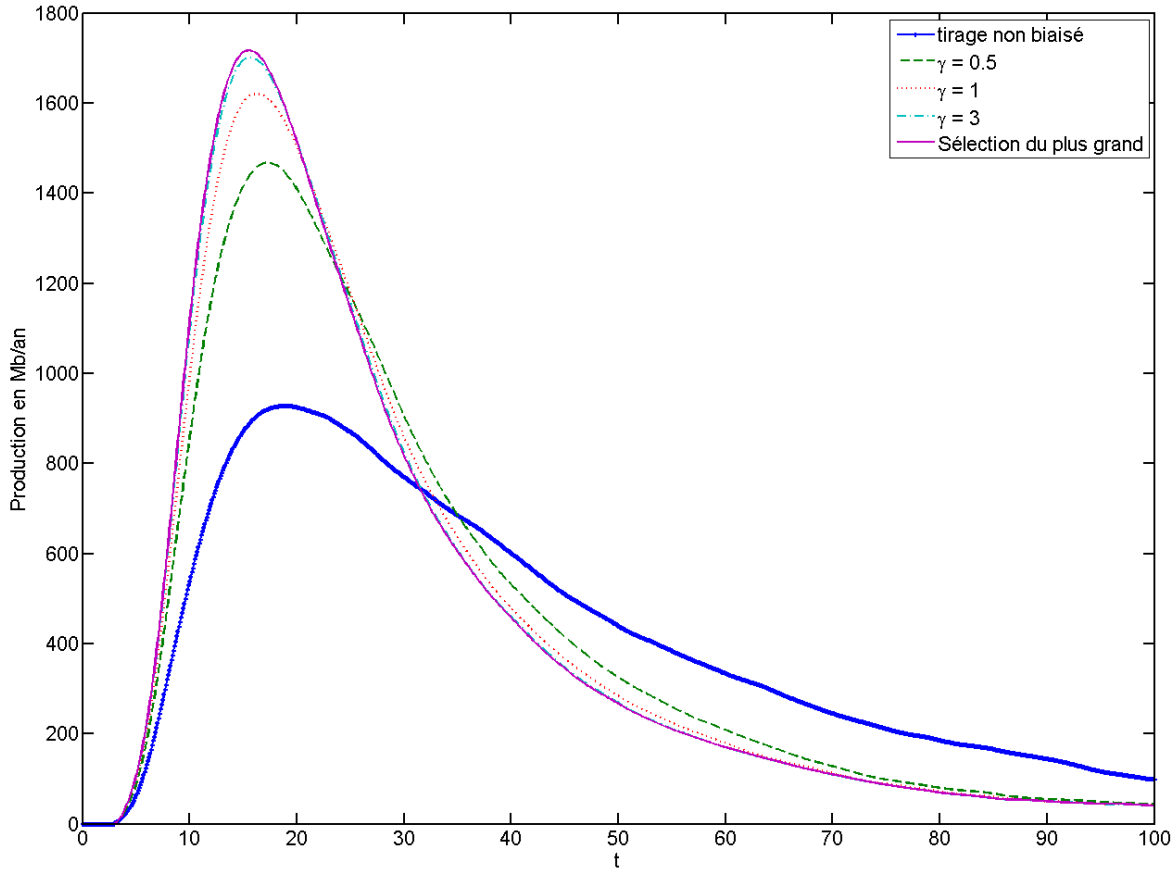
$\mu_0 = 6$

Intensité de mise en production : $\lambda = 6$

(a) Paramètres du protocole.



(b) Fonction de visibilité associée.



(c) Courbes de production moyennes pour différents types de tirage dans le stock.

FIG. 7.5: Impact de la politique de sélection (protocole 1).

$N_{\mathcal{B}} = 2500$

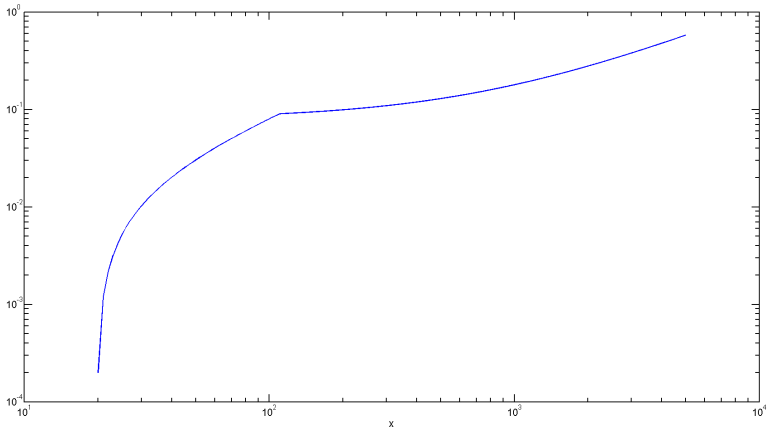
$\alpha = 0.75$

$$h : \begin{cases} \text{Partition :} & [20\,,\,70\,,\,110\,,\,5000] \\ A = & [10^{-3}\,,\,10^{-3}\,,\,10^{-4}] \\ b = & 2\,10^{-4} \end{cases}$$

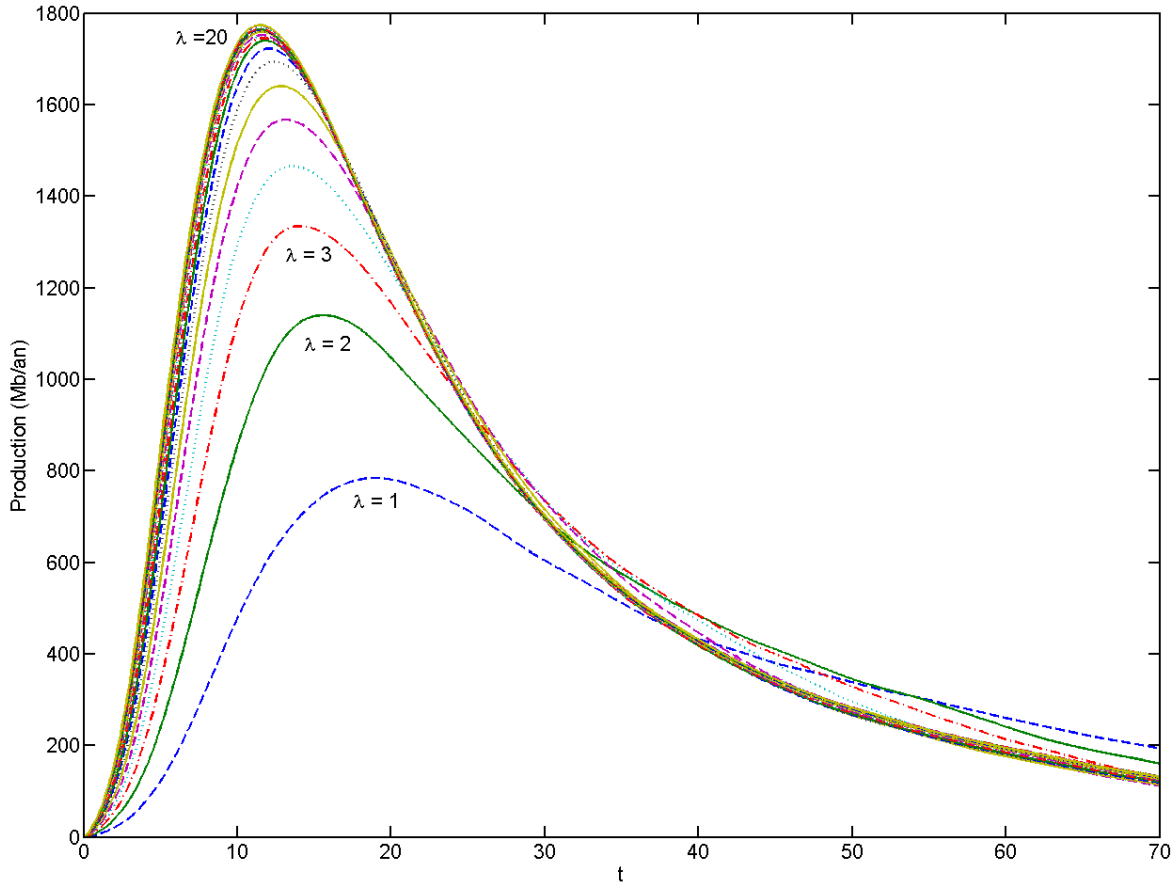
$\mu_0 = 6$

Tirage dans le stock : $\gamma = 2$

(a) Paramètres du protocole.



(b) Fonction de visibilité associée.



(c) Courbes de production moyennes pour différentes hypothèses d'intensité de mise en production.

FIG. 7.6: Impact de l'intensité de mise en production (protocole 2).

λ	Quantité moyenne de champs produits	Production cumulée moyenne (10^9 b/an)
2	140	373
6	420	440
10	700	455
15	1050	465
20	1400	469

TAB. 7.1: Description des productions de bassin pour cinq scénarios d'intensité différentes. La période de production considérée est $[0, 70]$.

Protocole 3 : Nous considérons maintenant un scénario dans lequel le nombre de gisements augmente avec le temps, ceci afin d'étudier dans quelle mesure il est possible d'atténuer le déclin au-delà du pic de production. Cette situation peut être modélisée par un processus de Poisson inhomogène d'intensité $\lambda(t) = \tilde{\lambda} + \frac{t}{2}$. Ici $\tilde{\lambda} = 6$ et le tableau 7.7a donne les valeurs des autres paramètres fixés pour cet exemple. Sous ces hypothèses, le nombre de gisements lancés chaque année augmente jusqu'à atteindre environ 40 gisements par an au bout de 70 ans. Au total, plus d'un millier de gisements supplémentaires sont mis en production par rapport la situation où l'intensité reste à $\tilde{\lambda} = 6$. Le graphique 7.7c permet de comparer les deux courbes de production moyennes correspondant aux deux situations λ et $\tilde{\lambda}$. En réalité, cet effort important de mise en production ne parvient à augmenter la production cumulée que d'environ 7%, et les vitesses de déclin sont comparables, avec des niveaux de production légèrement supérieurs dans le cas de l'intensité λ. Il ne suffit donc pas d'augmenter l'effort de mise en production des gisements pour compenser le déclin de la production au-delà du pic.

Protocole 4 : Dans cette expérience, nous effectuons quatre types de simulations pour mesurer l'effet de l'augmentation de la visibilité des champs combinée avec une augmentation de l'effort de mise en production. Le détail du protocole de simulation est présenté dans la figure 7.8. L'effet sur la courbe de production est plus important que dans la situation du protocole 2 où l'effort d'exploration était identique pour toutes les intensités de mise en production considérées. Mais l'augmentation de l'activité pétrolière (exploration et production) a surtout pour effet d'avancer la date du pic et d'accentuer la vitesse de déclin au-delà de celui-ci, sans découvrir beaucoup de nouveaux gisements : le scénario 4 (voir la légende de la figure 7.8c) mobilise 4 fois plus de gisements que le scénario 1 mais produit à peine 20% de plus que le premier scénario.

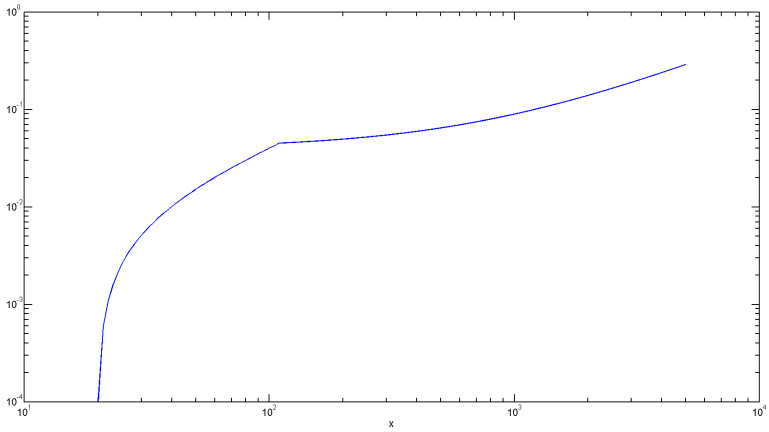
7.2.2 Impact de la loi des réserves

Afin d'illustrer l'influence de la distribution des réserves sur la forme du profil du bassin, nous proposons le schéma de simulation suivant.

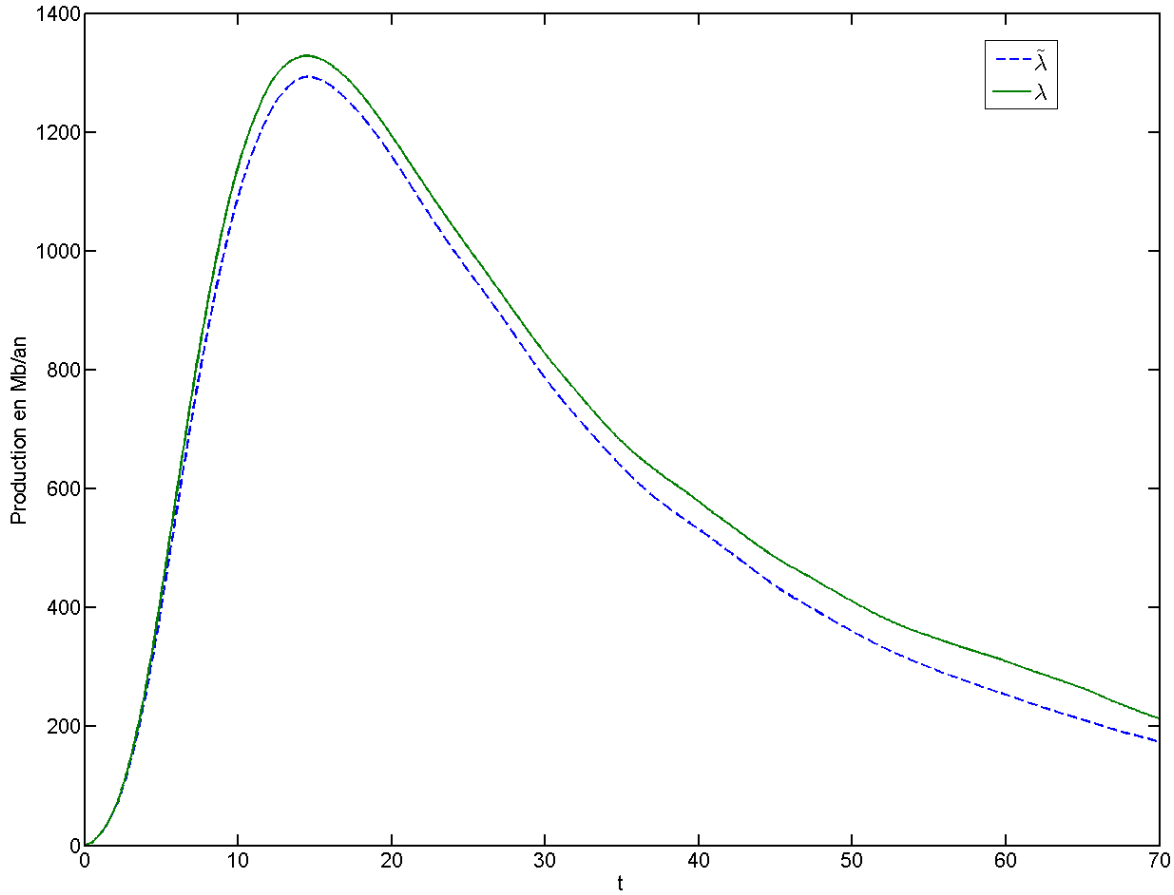
Protocole 5 : Les paramètres du protocole sont donnés dans le tableau 7.9a, le paramètre α de la loi de Lévy-Pareto varie entre 0.3 et 0.9 afin des décrire des situations de concentrations d'habitat très différentes. La figure 7.9c présente les courbes de production moyennes obtenues.

$N_{\mathcal{B}} =$	2500	
$\alpha =$	0.75	
h :	Partition :	$[20\,,\,70\,,\,110\,,\,5000]$
	$A =$	$[5\,10^{-4}\,,\,5\,10^{-4}\,,\,5\,10^{-5}]$
	$b =$	10^{-4}
$\mu_0 =$	20	
Tirage dans le stock :	$\gamma = 2$	

(a) Paramètres du protocole.



(b) Fonction de visibilité associée.



(c) Courbe de production moyenne du bassin pour les deux hypothèses d'intensité λ et $\tilde{\lambda}$.

FIG. 7.7: Impact de l'intensité de mise en production (protocole 3).

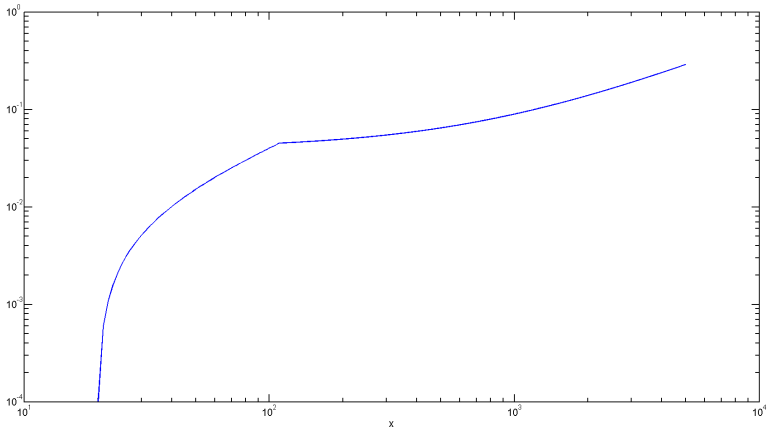
$N_{\mathcal{B}} = 2500$

$\alpha = 0.75$

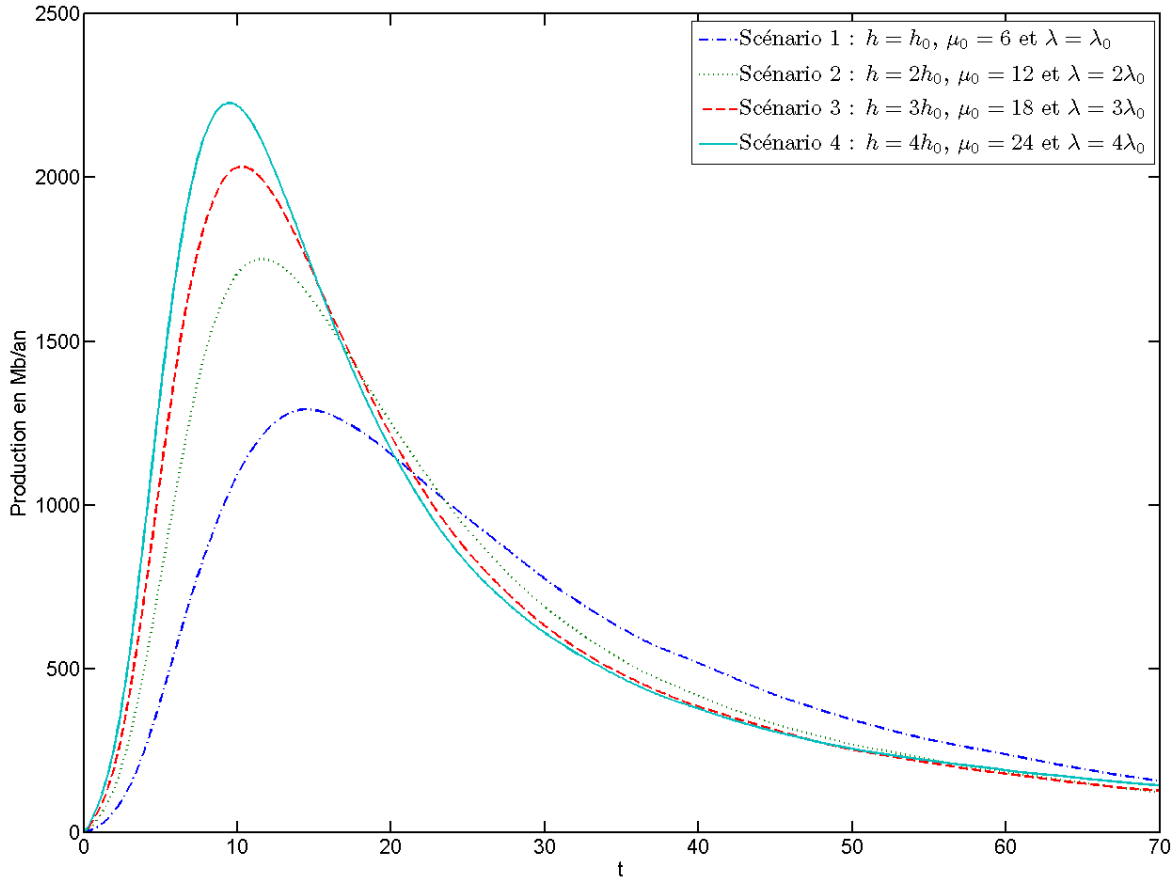
$$h_0 : \begin{cases} \text{Partition :} & [20\,,\,70\,,\,110\,,\,5000] \\ A = & [5\,10^{-4}\,,\,5\,10^{-4}\,,\,5\,10^{-5}] \\ b = & 10^{-4} \end{cases}$$

Tirage dans le stock : $\gamma = 2$

(a) Paramètres du protocole.



(b) Fonction de visibilité associée.



(c) Courbe de production moyenne du bassin pour des scénarios d'exploration et de mise en production de différentes intensités.

FIG. 7.8: Impact de l'intensité de mise en production (protocole 4).

Nous observons tout d'abord que l'amplitude de la courbe de production augmente avec la concentration de l'habitat, c'est-à-dire lorsque α diminue. Ceci est naturel puisque l'espérance d'une variable aléatoire de distribution de Lévy-Pareto de paramètre α restreinte à $[1, x_{\max}]$ est une fonction décroissante de α. En revanche, une fois normalisées par la valeur de leur maximum, ces courbes ont globalement la même allure, comme le montre la figure 7.10. On note enfin que le pic de production survient plus tard pour un habitat concentré. Ce phénomène est naturel car il y a beaucoup de gisements de grande taille lorsque α est faible, et la production de ces grands gisements s'étale alors plus dans le temps que lorsqu'il n'y en a que quelques-uns, et ceci a pour effet de retarder le passage du pic.

7.3 Prolongement de l'exploration de la production d'un bassin

Dans cette section, nous proposons des scénarios pour l'exploration et la production des hydrocarbures dans des bassins en cours d'exploitation. Pour être capable de proposer ces prolongements, nous devons tout d'abord estimer l'intensité du processus de découverte des petits champs, ainsi que le nombre de gisements de tailles supérieurs à x_0 restants dans le bassin.

7.3.1 Estimation de l'intensité du processus de découvertes des petits champs

Soit Q_p le nombre des champs de taille dans $[1, x_0]$ et découverts pendant l'année p. On note $t^\star$ l'année actuelle, alors $(Q_1, \ldots, Q_{t^\star})$ forme un échantillon variables aléatoires de distribution de Poisson de paramètre l'intensité μ_0 du processus de découverte. L'estimation du maximum de vraisemblance de μ_0 a pour expression

$$\hat{\mu}_0 = \frac{1}{t^\star} \sum_{p=1}^{t^\star} Q_p,$$

le tableau 7.2 présente les estimations obtenues pour zones de production. Rappelons qu'une description de ces bassins est disponible dans l'annexe A.

Zone pétrolifère	I_0	$t^\star$	$n_0(t^\star)$	$\hat{\mu}_0$
Sirte	[1 , 12]	42	104	2.36
mer du Nord	[1 , 20]	31	226	6.85
Nigeria off-shore	[1 , 20]	39	103	2.5

TAB. 7.2: Estimations des intensités des processus de Poisson modélisant les temps de découverte dans la classe des petits champs pour les trois bassins étudiés.

7.3.2 Effectifs des gisements de taille supérieure à x_0

Soit $(X_1, \ldots, X_N)$ un échantillon de variables aléatoires de distribution de Lévy-Pareto restreinte à $[x_0, x_{\max}]$. Pour tout $i \in \{1, \ldots, N\}$ soit D_i la variable aléatoire de loi condition-

$N_{\mathcal{B}} = 2500$

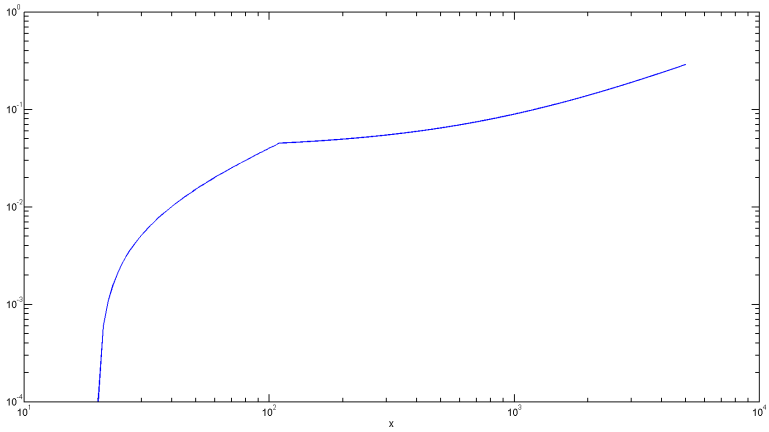
$$h : \begin{cases} \text{Partition :} & [20\,,\,70\,,\,110\,,\,5000] \\ A = & [5\,10^{-4}\,,\,5\,10^{-4}\,,\,5\,10^{-5}] \\ b = & 10^{-4} \end{cases}$$

$\mu_0 = 8$

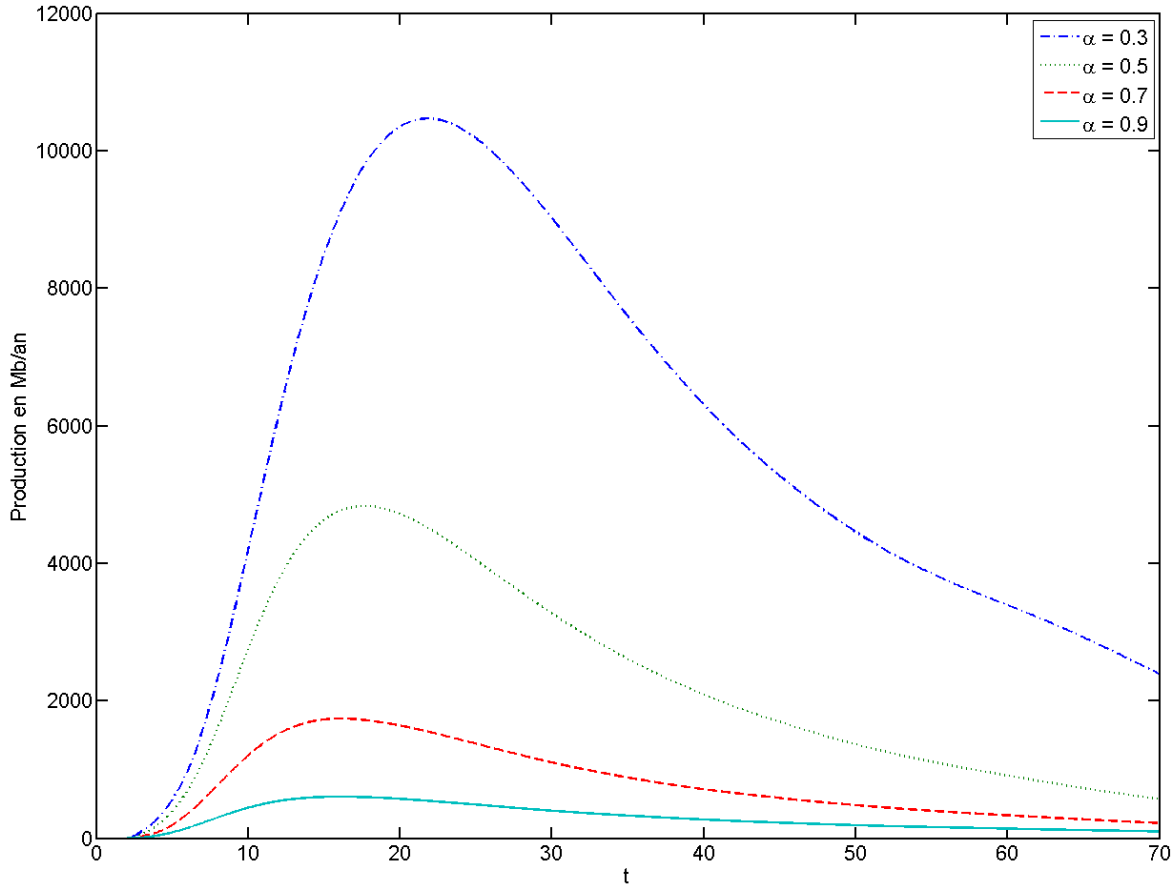
Tirage dans le stock : $\gamma = 2$

Intensité de mise en production : $\lambda = 8$.

(a) Paramètres du protocole.



(b) Fonction de visibilité associée.



(c) Courbes de production moyennes pour différents différents coéfficient α de la loi de Lévy-Pareto.

FIG. 7.9: Impact de la politique de la loi des réserves (protocole 5).

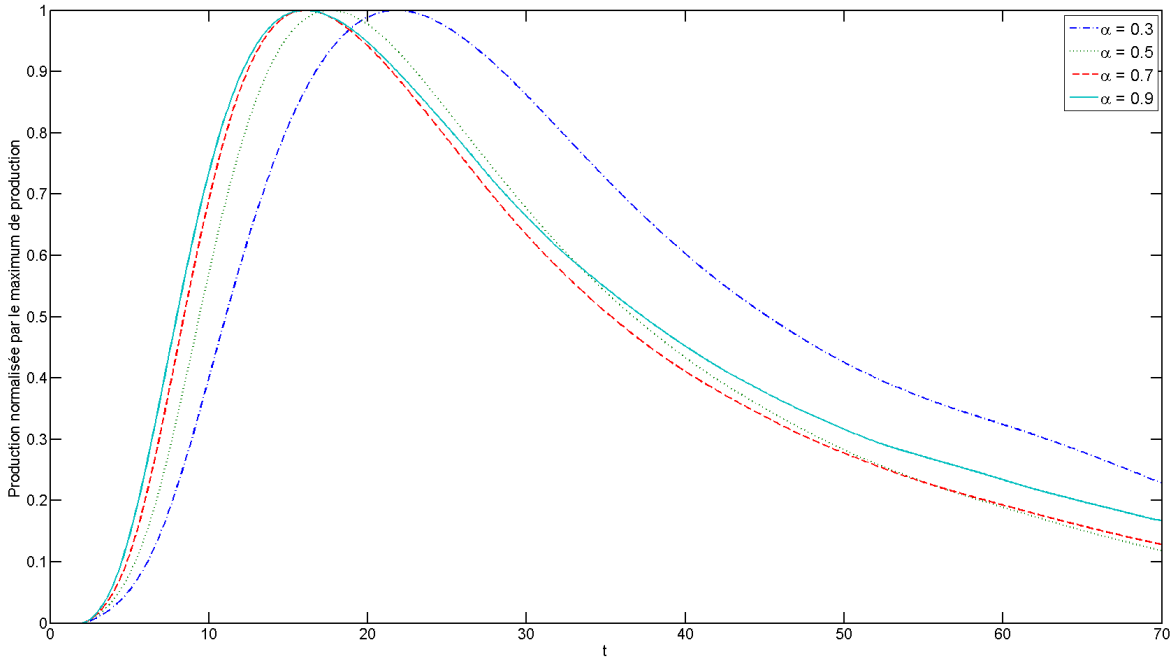


FIG. 7.10: Courbes de production moyennes de la figure 7.9c normalisées par la valeur du maximum de production.

nelle $(D_i \,|\, X_i) \sim \mathcal{E}(h(X_i))$ avec $h \in \mathcal{H}_m$ où $m = (I_1, \ldots, I_k)$ est une partition d'intervalles de $[x_0, x_{\max}]$ (voir section 4.3.3). Comme précédemment, X_i et D_i représentent respectivement la taille et la date de découverte du gisement i. Soit n le nombre des champs de l'échantillon découverts avant la date $t^\star$, nous avons

$$n = \sum_{i=1}^{N} \mathbb{1}_{D_i < t^\star}.$$

Soit N_j le nombre total de gisements de la classe I_j enfouis dans le bassin, et n_j désigne le nombre de champs de cette catégorie découverts avant $t^\star$. Nous pouvons écrire de même que

$$n_j = \sum_{i \,|\, X_i \in I_j} \mathbb{1}_{D_i < t^\star}. \tag{7.7}$$

Nous déduisons de (7.7) la relation suivante

$$\mathbb{E}(n_j) = N_j \, \mathbb{P}(D < t^\star | X \in I_j).$$

La probabilité dans le terme de droite correspond à la probabilité d'inclusion ω_j d'un champ de la catégorie I_j. Pour $I = [x_{j-1}\,,\, x_j]$, celle-ci a pour expression

$$\omega_j = 1 - \int_{x_{j-1}}^{x_j} \exp\{-h(x)t^\star\} \, \frac{\alpha x^{-\alpha-1}}{x_j^{-\alpha} - x_{j+1}^{-\alpha-1}} \, dx. \tag{7.8}$$

Cette probabilité d'inclusion peut être estimée en remplaçant dans (7.8) la fonction de visibilité h par l'estimateur $\hat{h}$ qui est associé à l'estimateur pénalisé $\hat{g}_{\hat{m}}^\star$. Notons $\hat{\omega}_j$ l'estimateur de ω_j

ainsi obtenu. Soit I_j un intervalle de la partition $\hat{m}$, le nombre N_j de champs de cette catégorie dans le sous-sol peut alors être estimé par la quantité

$$\hat{N}_j = \frac{n_I}{\hat{\omega}_I}.$$

Le nombre total des champs de taille supérieure à x_0 est finalement estimé par $\hat{N} = \sum_{j=1}^{k} \hat{N}_j$.

Il est possible de proposer un second estimateur des effectifs. Dans le contexte des sondages sous probabilités d'inclusion inégales, l'estimateur de Horvitz-Thompson (Horvitz et J., 1952) est couramment utilisé. Cet estimateur a notamment été employé par Lepez (2002) pour l'estimation des réserves pétrolières. Adapté à notre modélisation, son expression est la suivante :

$$\hat{N}_j^{\text{HS}} = \sum_{i \,|\, X_i^\star \in I_j} \frac{1}{\hat{\omega}(X_i)}$$

où $\hat{\omega}(x) = 1 - \exp\{-\hat{h}\,(x)t^\star\}$ est la probabilité d'inclusion d'un champ de taille x. Comme précédemment, nous définissons l'estimateur du nombre total de champs N par $\hat{N}^{\text{HS}} = \sum_{j=1}^{k} \hat{N}_j^{\text{HS}}$.

Des simulations sont proposées dans l'annexe D.1.3 pour étudier les performances de $\hat{N}$ et de $\hat{N}^{\text{HS}}$ à l'aide de jeux de données simulés. Pour ces simulations, les deux estimateurs ont des performances comparables. En revanche, celles-ci montrent aussi que l'estimation des effectifs est très sensible à la qualité de l'estimation de α. Ceci nous conforte dans le choix d'utiliser le modèle de Lepez pour estimer le paramètre α, plutôt que de l'incorporer dans les équations de vraisemblance de notre modèle.

Application

Les tableaux 7.3, 7.4 et 7.5 présentent les estimations obtenues pour les effectifs des trois bassins étudiés.

j	I_j	$\hat{\omega}_j$	n_j	$\hat{N}_j - n_j$
1	[20 , 63.5]	0.49	111	115
2	[63.5 , 110]	0.73	37	13
3	[110 , 4139]	0.95	74	4

TAB. 7.3: Estimations du nombre de champs restant à découvrir par classes de taille pour le bassin de la mer du Nord.

j	I_j	$\hat{\omega}_j$	n_j	$\hat{N}_j - n_j$
1	[12.1 , 17]	0.35	23	43
2	[17 , 40.5]	0.53	43	38
3	[40.5 , 72.5]	0.81	22	5
4	[72.5 , 4482]	0.98	44	1

TAB. 7.4: Estimations du nombre de champs restant à découvrir par classes de taille pour le bassin de Sirte.

j	I_j	$\hat{\omega}_j$	n_j	$\hat{N}_j - n_j$
1	[20.5 , 350]	0.52	89	81
2	[350 , 1250]	0.99	43	0

TAB. 7.5: Estimations du nombre de champs restant à découvrir par classes de taille pour le Nigeria off-shore (Delta du Niger).

7.3.3 Prolongements de l'exploration

Dans la classe des petits champs, les découvertes futurs sont modélisées en prolongeant naturellement le processus de Poisson marqué dont l'intensité à été estimée précédemment. La proposition suivante permet de modéliser la suite des découvertes pour la classe des champs de taille supérieure à x_0.

Proposition 7.3.1. *Soit X une variable aléatoire de distribution de Lévy-Pareto restreinte à $[x_0, x_{max}]$ et soit D une variable aléatoire de loi conditionnelle $(D \mid X) \sim \mathcal{E}(h(X))$ avec $h \in H_m$ où m est une partition de $[x_0, x_{max}]$. Pour $I = [x_{j-1}\,,\, x_j]$ un intervalle de la partition m, soit $(\tilde{X}, \tilde{D})$ un vecteur aléatoire de même loi que $(X, D \mid D > t^\star, X \in I)$. On note $P_{dec}(I) := \mathbb{P}(D \leq t^\star | X \in I)$. La loi jointe de $(\tilde{X}, \tilde{D})$ a pour densité la fonction définie sur $I \times \mathbb{R}_+$ par*

$$(x,t) \longmapsto \frac{1}{1 - P_{dec}(I)} \frac{\alpha x^{-\alpha-1}}{x_{j-1}^{-\alpha} - x_j^{-\alpha}} h(x) \exp\{-h(x)u\} \mathbb{1}_{x \in I, t > t^\star}.$$

En particulier, la densité marginale de $\tilde{X}$ vaut

$$x \mapsto \frac{\exp\{-h(x)\, t^\star\}}{1 - P_{dec}(I)} \frac{\alpha x^{-\alpha-1}}{x_{j-1}^{-\alpha} - x_j^{-\alpha}} \mathbb{1}_{x \in I}\ , \tag{7.9}$$

et la loi conditionnelle de $\tilde{D}$ sachant $\tilde{X}$ est de distribution exponentielle de paramètre $h(\tilde{X})$ tronquée à gauche en $t^\star$.

Démonstration. Soit z une fonction mesurable et bornée définie sur $I \times \mathbb{R}_+$ et à valeurs dans $\mathbb{R}$. On a alors

$$\begin{aligned}
\mathbb{E}\left(z(\tilde{X}, \tilde{D})\right) &= \frac{1}{\mathbb{P}(D > t^\star | X \in I)} \int_{t \geq t^\star} \int_{x \in I} \frac{\alpha x^{-\alpha+1}}{x_{j-1}^{-\alpha} - x_j^{-\alpha}} h(x) \exp\left\{-h(x)t\right\}\, z(x,t)\, dxdt \\
&= \frac{1}{1 - \mathrm{P_{dec}}(I)} \int_{x \in I} \int_{u \geq 0} \frac{\alpha x^{-\alpha+1}}{x_{j-1}^{-\alpha} - x_j^{-\alpha}} h(x) \exp\{-h(x)(u + t^\star)\}\, z(x,t)\, dxdt.
\end{aligned}$$

La loi marginale de $\tilde{X}$ s'obtient ensuite directement en intégrant la densité jointe en u, et la loi conditionnelle de $\tilde{D}$ se lit sur l'expression de la loi jointe. □

La loi des réserves restant à découvrir n'est donc pas de type Lévy-Pareto. Pour une classe fixée I, le terme exponentiel dans l'expression (7.9) montre que la densité de la loi de $\tilde{X}$ prend des valeurs plus importantes sur les plus petits champs de I que ne prend la densité d'une loi de Pareto restreinte à I. Ceci est naturel puisqu'à l'intérieur de la classe I, la visibilité

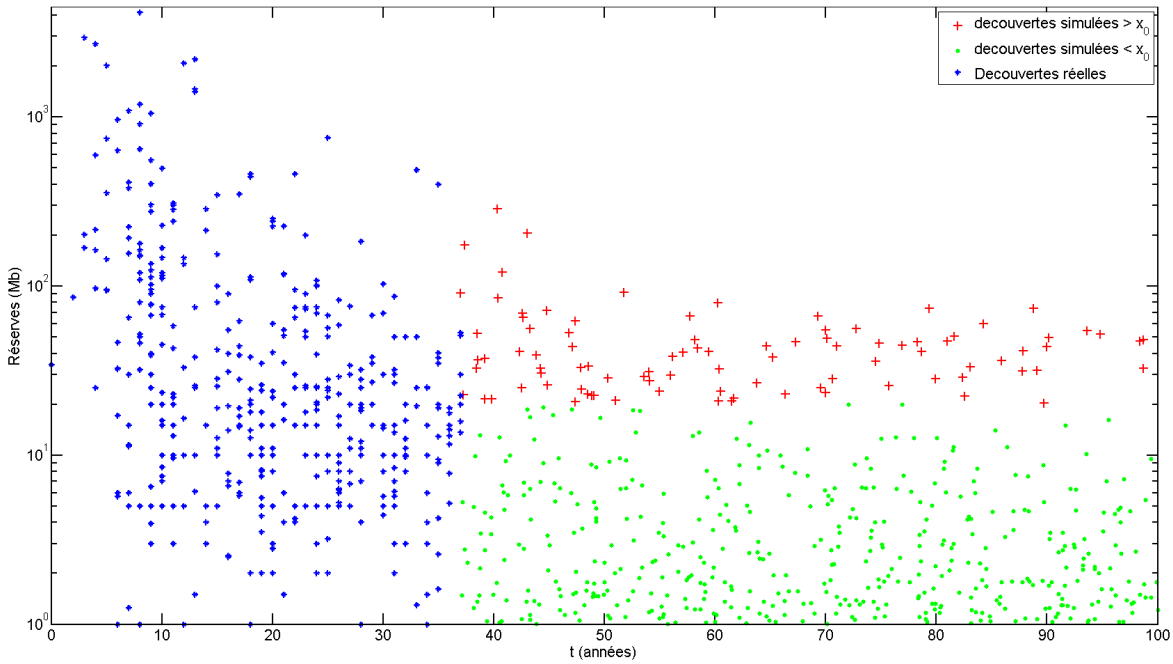


FIG. 7.11: Prolongement du processus des découvertes en mer du Nord. Les réserves sont représentées en échelle logarithmique pour mieux visualiser les petites découvertes.

d'un champ augmente avec sa taille, et les plus grands gisements de cette classe ont été plus facilement découverts que les autres.

Soit $S_j(t) = \sum_{i \,|\, X_i \in I_j} X_i \mathbb{1}_{t^\star < D_i \leq t}$ la variable aléatoire des réserves totales découvertes cumulées sur la période $[t^\star, t]$, pour les champs de la catégorie I_j. D'après la proposition précédente,

$$\mathbb{E}\left(S_j(t) \,|\, N_j - n_j\right) = (n_j - N_j)\,\frac{1}{1 - \mathrm{P}_{\mathrm{dec}}(I_j)} \int_{x \in I_j} \int_{t \geq t^\star} \frac{\alpha x^{-\alpha}}{x_{j-1}^{-\alpha} - x_j^{-\alpha}} h(x) \exp\{-h(x)t\}\, dx dt$$

Pour étudier l'évolution des quantités découvertes au cours du temps, nous définissons le taux de découverte moyen (en Mb par an) par la fonction τ suivante, pour $t > t^\star$,

$$\begin{aligned} \tau(t) &:= \frac{1 - x_0^{-\alpha+1}}{1 - x_0^{-\alpha}} \frac{\alpha}{1-\alpha} \mu_0 + \sum_{j=1}^{k} \frac{d}{dt} \mathbb{E}\left[S_j(t)\right] \\ &= \mu_0 + \sum_{j=1}^{k} (n_j - N_j)\, \frac{1}{1 - \mathrm{P}_{\mathrm{dec}}(I_j)} \int_{x \in I_j} \frac{\alpha x^{-\alpha}}{x_{j-1}^{-\alpha} - x_j^{-\alpha}}\, h(x) \exp\{-h(x)t\}\, dx \end{aligned}$$

Cette quantité permet de mesurer le déclin des quantités de réserves découvertes au cours du temps.

Applications

La figure 7.11 présente une simulation de prolongement pour la mer du Nord. Les gisements découverts dans la classe $[x_0, x_{\max}]$ ont été simulés en utilisant l'estimation de la fonction de visibilité $\hat{h}$ sur la partition $\hat{m}$ sélectionnée par la méthode de la pente et la proposition 7.3.1.

Les gisements découverts dans la catégorie $[1, x_0]$ ont été obtenus en prolongeant le processus de Poisson marqué des petites découvertes. Notons que sur la période prolongée, les très petits champs (de 1Mb à 3Mb) sont découverts en plus grand nombre que sur la période connue de l'exploration du bassin. Ce défaut n'est en réalité pas gênant car les quantités de réserves en jeu sont négligeables. De plus, il est probable qu'à l'avenir ces très petits gisements soient recherchés plus activement par les pétroliers. Dans section suivante, nous utilisons de tels prolongements de processus d'exploration pour proposer des prolongements de la production d'un bassin.

Sur la figure 7.12, le taux de découverte moyen τ représentée correspond à la zone de la mer du Nord. Sur le figure 7.13, des bornes de confiance pour la variable aléatoire $S_x(t)$ sont représentés à partir d'un grand nombre de simulations de prolongements Au-delà de la ligne verticale correspondant au présent, la fonction $t \mapsto \mathbb{E}(S_x(t))$ est encadrée par les quantiles 5% et 95% de $S_x(t)$ obtenus pour 1000 simulations. Les figures 7.14 et 7.15 représentent respectivement les taux moyens de découvertes et les découvertes cumulées pour le bassin de Sirte et pour le Nigeria off-shore (Delta du Niger). Sur ces figures, il est clair que l'exploration dans le Bassin de Sirte apportera significativement moins de découvertes qu'en mer du Nord ou dans le Nigeria off-shore.

Nous tenons à souligner de nouveau que notre modélisation est conçue pour décrire le cycle d'exploration correspondant à un certain type d'hydrocarbures. Le modèle ne prétend donc pas en compte les découvertes importantes correspondant aux hydrocarbures qualifiés de "deep oil", notamment pour la zone du Nigeria off-shore.

Les prolongements obtenus semblent respecter correctement la dynamique des processus d'exploration des trois bassins. En particulier, les scénarios de taux moyens de découverte prolongent de façon satisfaisante l'historique des quantités de pétrole découvertes dans le passé.

7.3.4 Prolongements de profils de production

Pour proposer un scénario de prolongement de la production d'un bassin pétrolier en cours d'exploitation, il suffit d'utiliser un prolongement de l'exploration du bassin et de définir une politique de gestion du stock. Chaque nouveau champ découvert est lancé selon la politique de gestion du stock et produit selon le modèle présenté à la section 7.1. Il est difficile d'inférer quelle politique de mise en production régit l'exploitation d'un bassin en cours de production, car le tirage effectué dans la population des gisements du stock évolue au cours du temps. Cependant, nous avons vu dans la section précédente que pour des tirages dans le stock biaisé par une puissance de la taille suffisamment importante, les profils de production du bassin ont globalement la même forme de courbe de production. Nous utilisons donc un tirage biaisé par la taille ($\gamma = 1$) comme politique de gestion du stock.

Les données de la base Wood Mackenzie 2004 proposent des prolongements des courbes de production de champs déjà exploités aujourd'hui. Il nous serait possible d'utiliser le modèle de production individuelle proposé dans la section 7.3.3 pour effectuer des prolongements des production individuelles de gisements, mais il est préférable d'utiliser la base Wood Mackenzie 2004 dont les scénarios de prolongement individuel sont basés sur les prévisions des compagnies

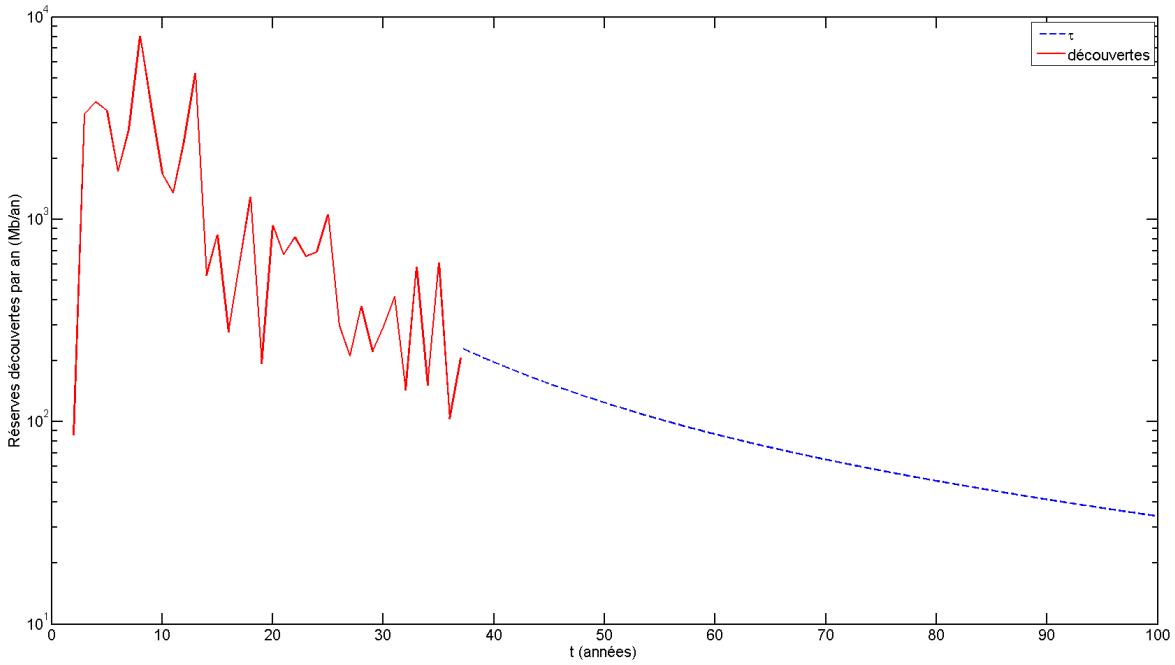


FIG. 7.12: Quantités annuelles découvertes en mer du Nord. La courbe en trait plein correspondant aux découvertes dans le passé est prolongée par la fonction τ du taux de découvertes moyen estimé dans le futur.

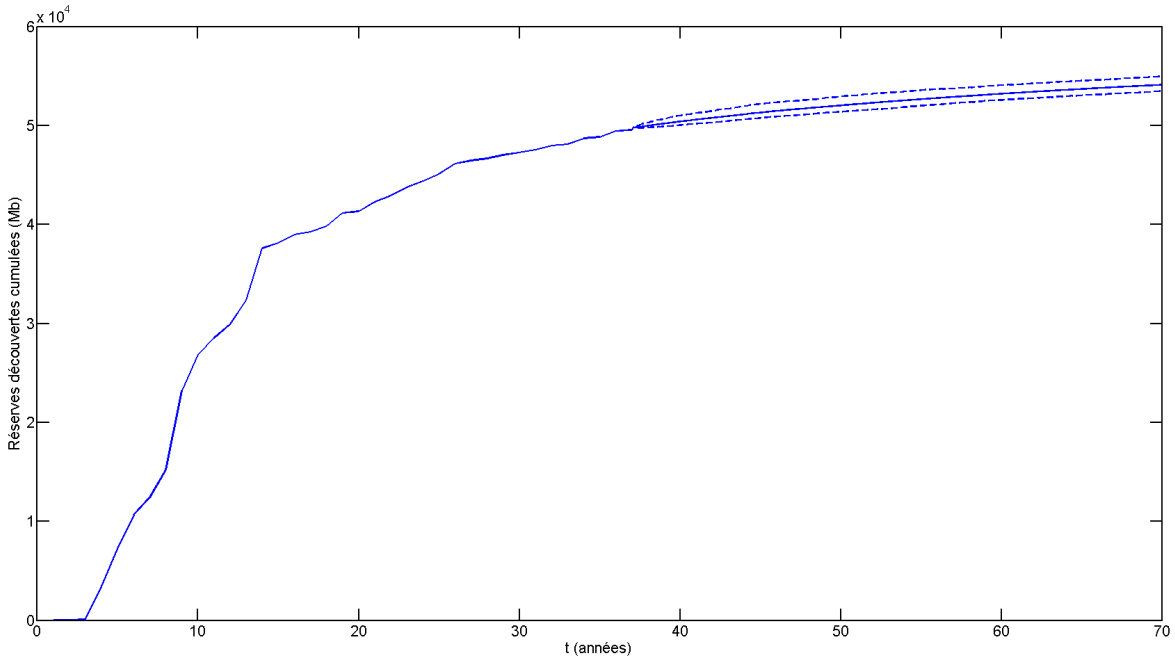


FIG. 7.13: Réserves découvertes cumulées en mer du Nord. Au delà du temps présent, la courbe prolongée correspond à la moyenne de $S_x(t)$, encadrée par les quantiles 5% et 95% pour 1000 simulations de $S_x(t)$.

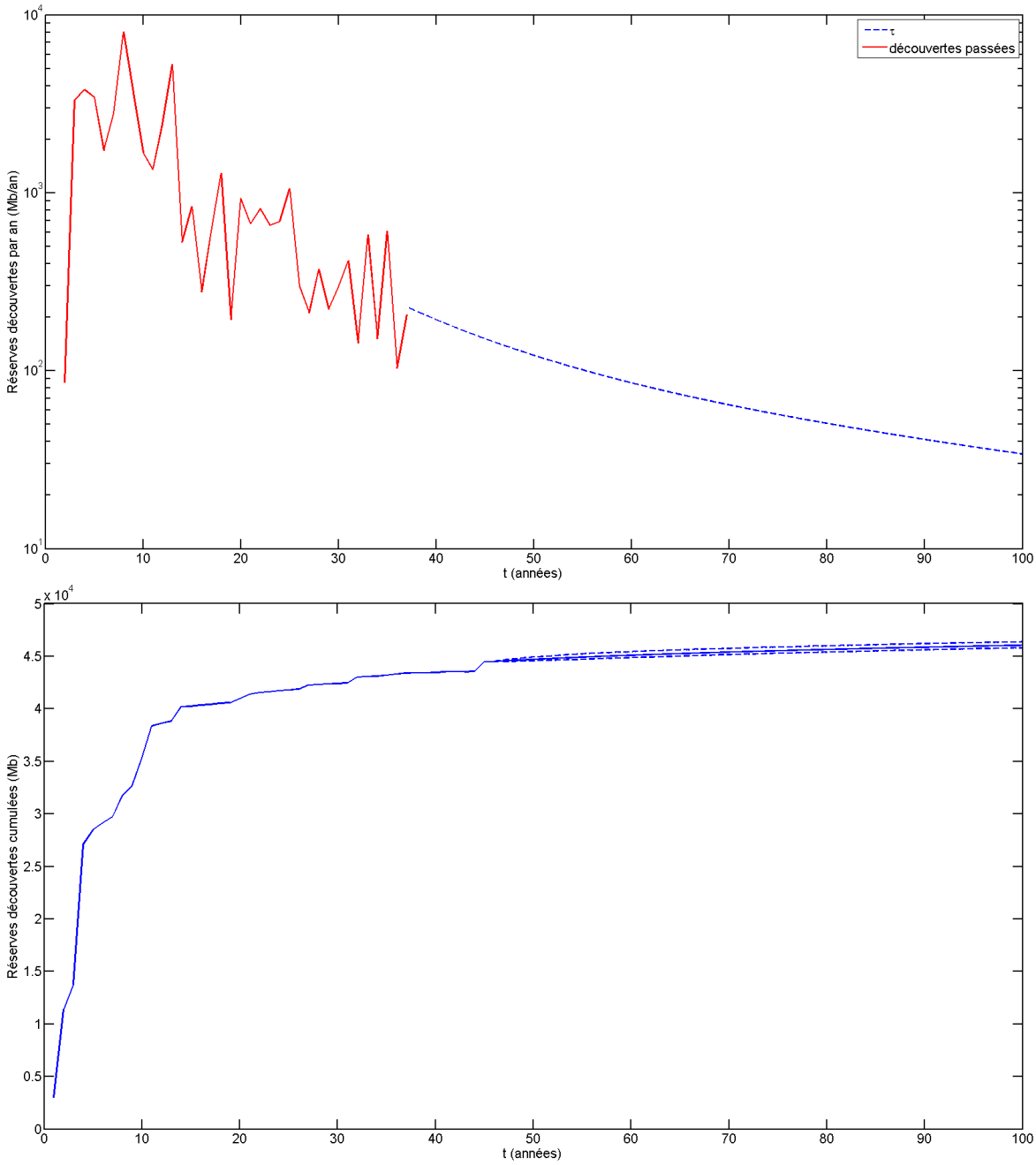


FIG. 7.14: Le graphique supérieur représente les quantités annuelles découvertes dans le bassin de Sirte. Le graphique inférieur donne les réserves découvertes cumulées.

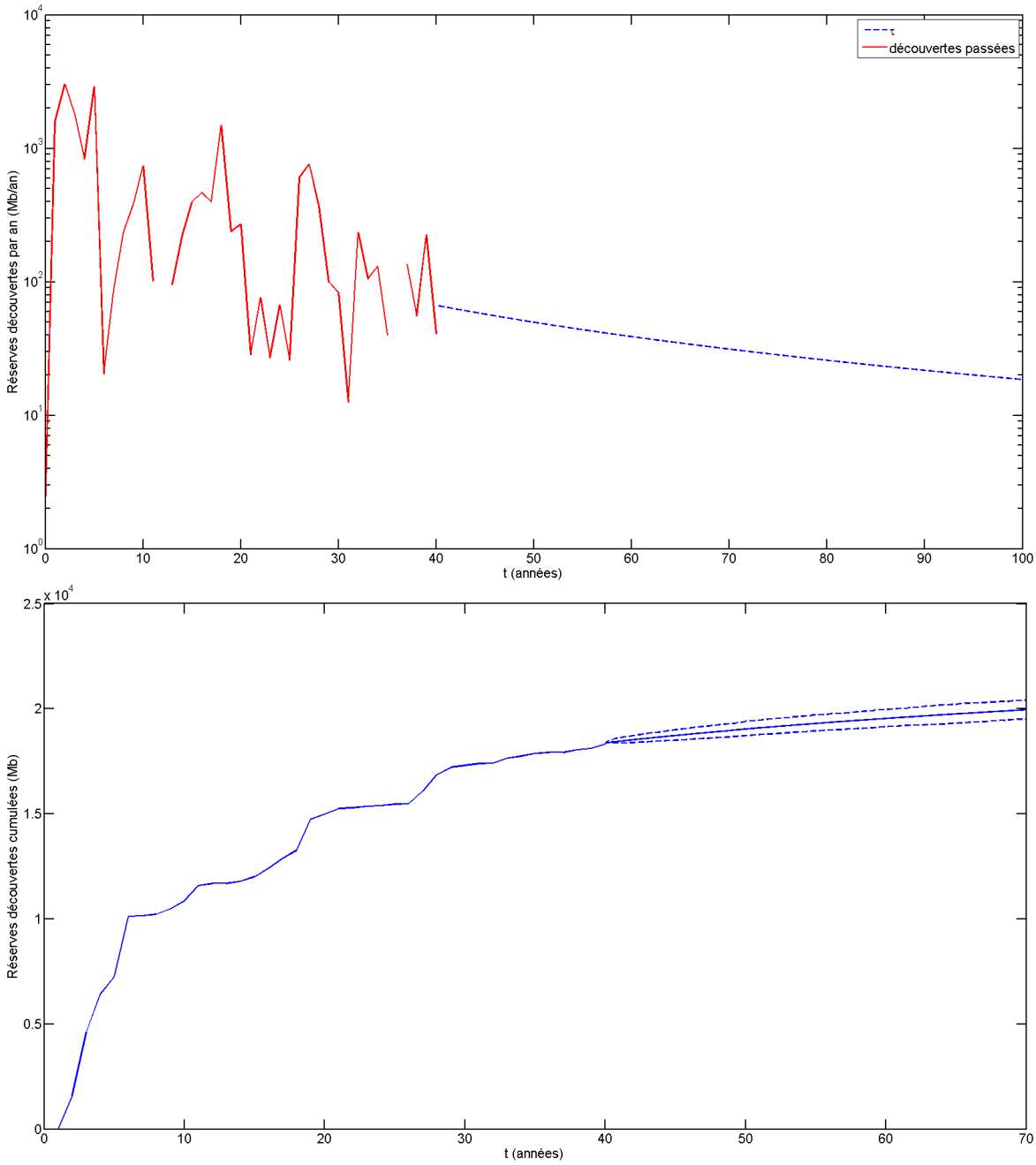


FIG. 7.15: Le graphique supérieur représente les quantités annuelles découvertes dans le Nigeria off-shore(sans l'off-shore ultra-profond). Le graphique inférieur donne les réserves découvertes cumulées.

pétrolières.

En réalité, un champ ne peut pas naturellement pas commencer à produire aussitôt après sa découverte. Nous imposons donc pour les simulations un délai incompressible de trois années entre la date de découverte d'un champ et la date de mise en production de celui-ci. Nous n'avons pu isoler la production passée dans le Bassin de Sirte car pour la Lybie, la base de donnée Wood Mackenzie 2004 regroupe les productions de gisements par compagnies pétrolières, ce qui ne permet donc pas d'isoler la production restreinte au Bassin de Sirte. La figure 7.16 présente l'évolution du stock en mer du Nord jusqu'en 2004. Nous proposons un premier scénario qui correspond au rythme de mise en production de ces dernières années, c'est-à-dire en considérant une intensité $\lambda = 8$. Le second scénario est basé sur une intensité croissante $\tilde{\lambda}(t) = 0.7t + 8$. Les prolongements correspondants sont tracés sur la figure 7.17. L'évolution du stock de gisements pour l'off-shore Nigeria (sans deep-water) est présenté sur la figure 7.18. Les mêmes scénarios de mise en production qu'en mer du Nord sont considérés pour prolonger la production sur la figure 7.19.

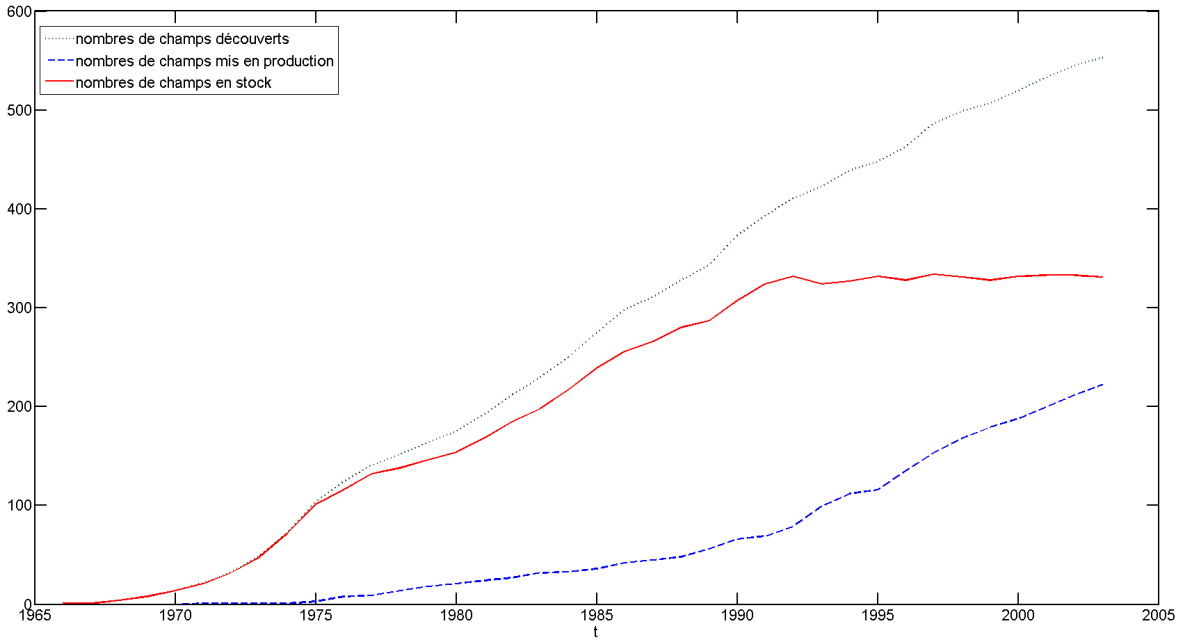


FIG. 7.16: Évolution du stock de gisements disponibles en mer du Nord.

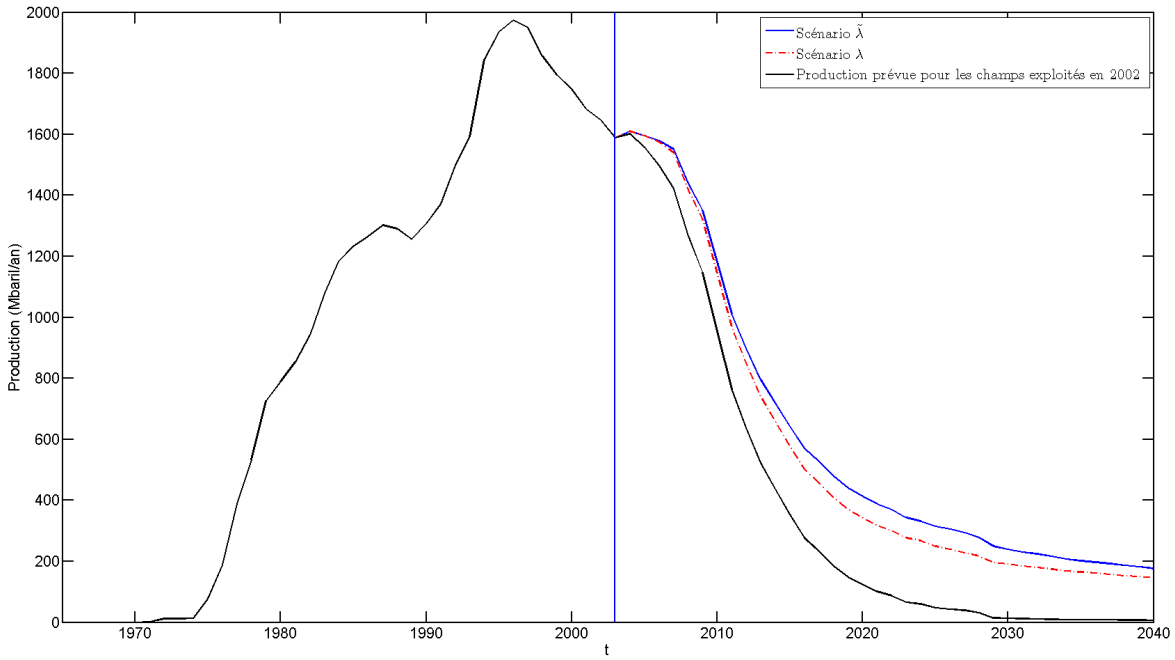


FIG. 7.17: Prolongement de la production en mer du Nord.

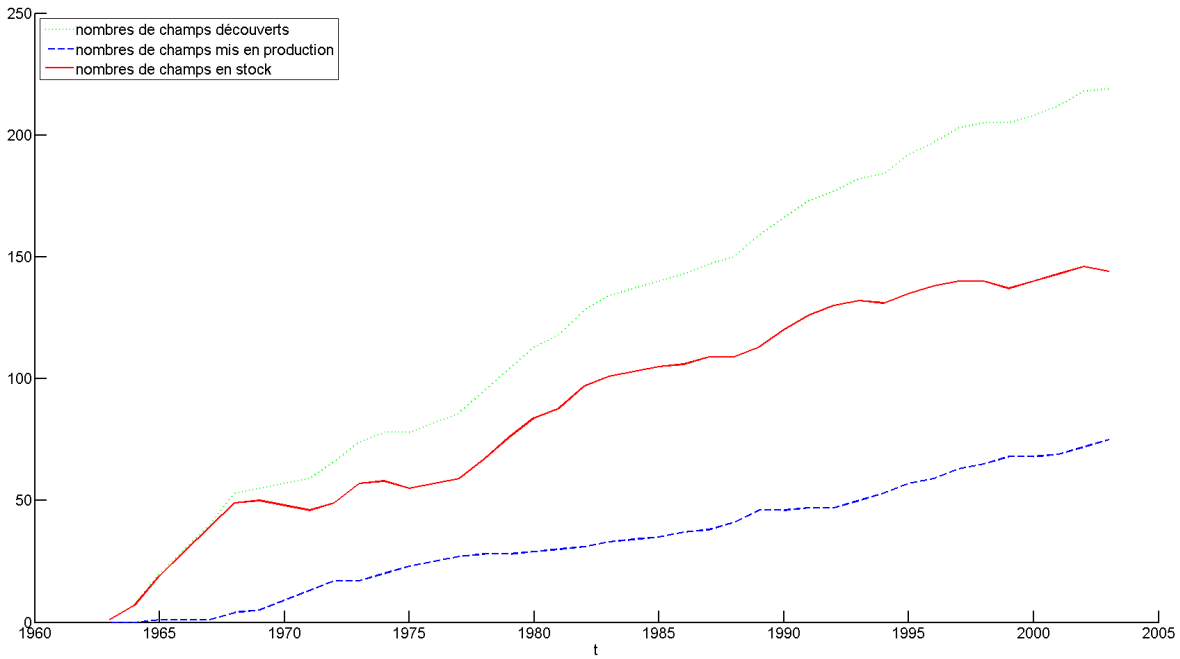


FIG. 7.18: Évolution du stock de gisements disponibles : gisements off-shore du Nigeria (sans deep-water).

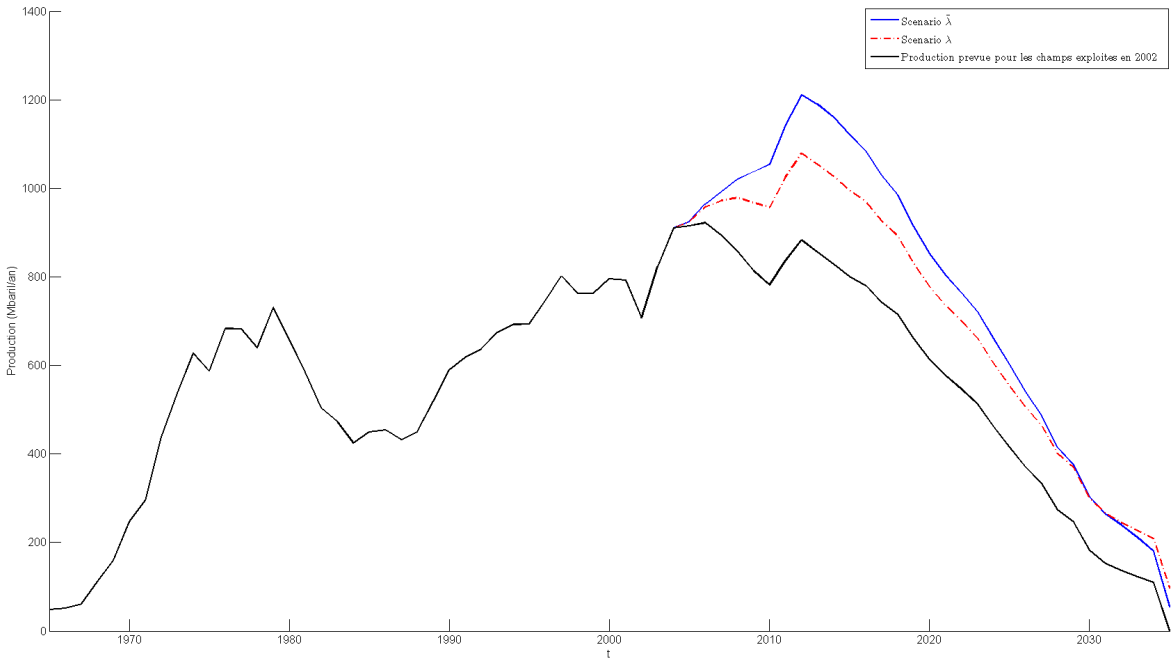


FIG. 7.19: Prolongement de la production off-shore du Nigeria (sans deep-water).

Appendices

Annexe A

Bassins étudiés

Pour illustrer les méthodes proposées dans cette thèse, nous étudions les trois bassins pétroliers suivants : la mer du Nord, le bassin de Sirte ainsi que le le Delta du Niger.

La province géologique du **Graben de la Mer du Nord** se situe entre les côtes anglaises et norvegiennes (voir figure A.1). L'ensemble des accumulations d'hydrocarbures de cette zone, riche en matières organiques d'origine marine, provient d'une même roche mère formée à la fin du Jurassique et au début du Crétacé. D'après la classification géologique proposée par l'USGS (2000), ce système pétrolier peut être divisé en trois sous-entités géologiques : les zones du Viking Graben, du Central Graben et du Moray Firth. Si ces zones proviennent de la fragmentation d'une même roche mère, celles-ci n'ont cependant pas eu rigoureusement la même histoire géologique. En particulier, les hydrocarbures de chacune de cette zones n'ont pas connu la même maturation. Les pièges ne sont pas non plus du même type sur l'ensemble de la province du Graben. Cependant, nous considérons la province dans son ensemble car c'est à cette échelle que la Mer du Nord a été exploitée. De plus, ce choix permet de disposer d'une quantité de champs en nombre suffisant pour les procédures d'estimation statistiques.

Le Bassin de Sirte (voir figure A.2) situé en Lybie contient plusieurs centaines de champs en production on-shore. Les accumulations d'hydrocarbures de cette zone se sont fragmentées à partir d'une roche mère située à l'origine dans les couches du haut Crétacé. Des champs ont été découverts et sont exploités dans seulement deux des quatre zones du découpage du bassin de Sirte proposé par l'USGS.

La dernière zone étudiée est le **bassin du Delta Niger** (voir figure A.2). D'après USGS (2000), les accumulations d'hydrocarbures de cette zone proviendraient d'une roche mère formée au début du Crétacé. Pour l'estimation de la distributions des tailles de gisements de cette région, nous utilisons les données des champs du Nigeria, dont une partie se situe en mer et une autre fraction à proximité des côtes.

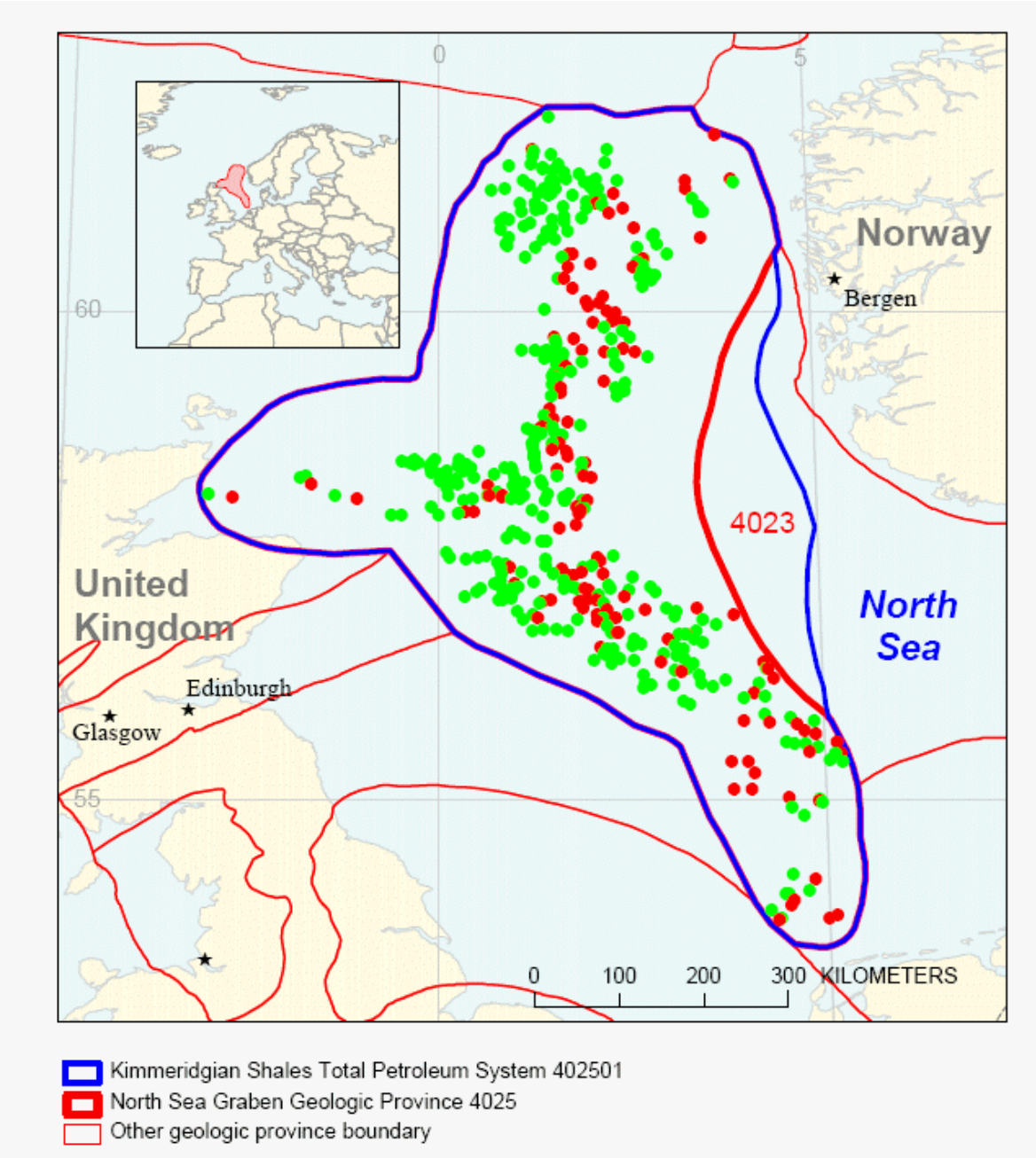


FIG. A.1: Province géologique du Viking Graben en Mer du Nord (sources : USGS (2000))

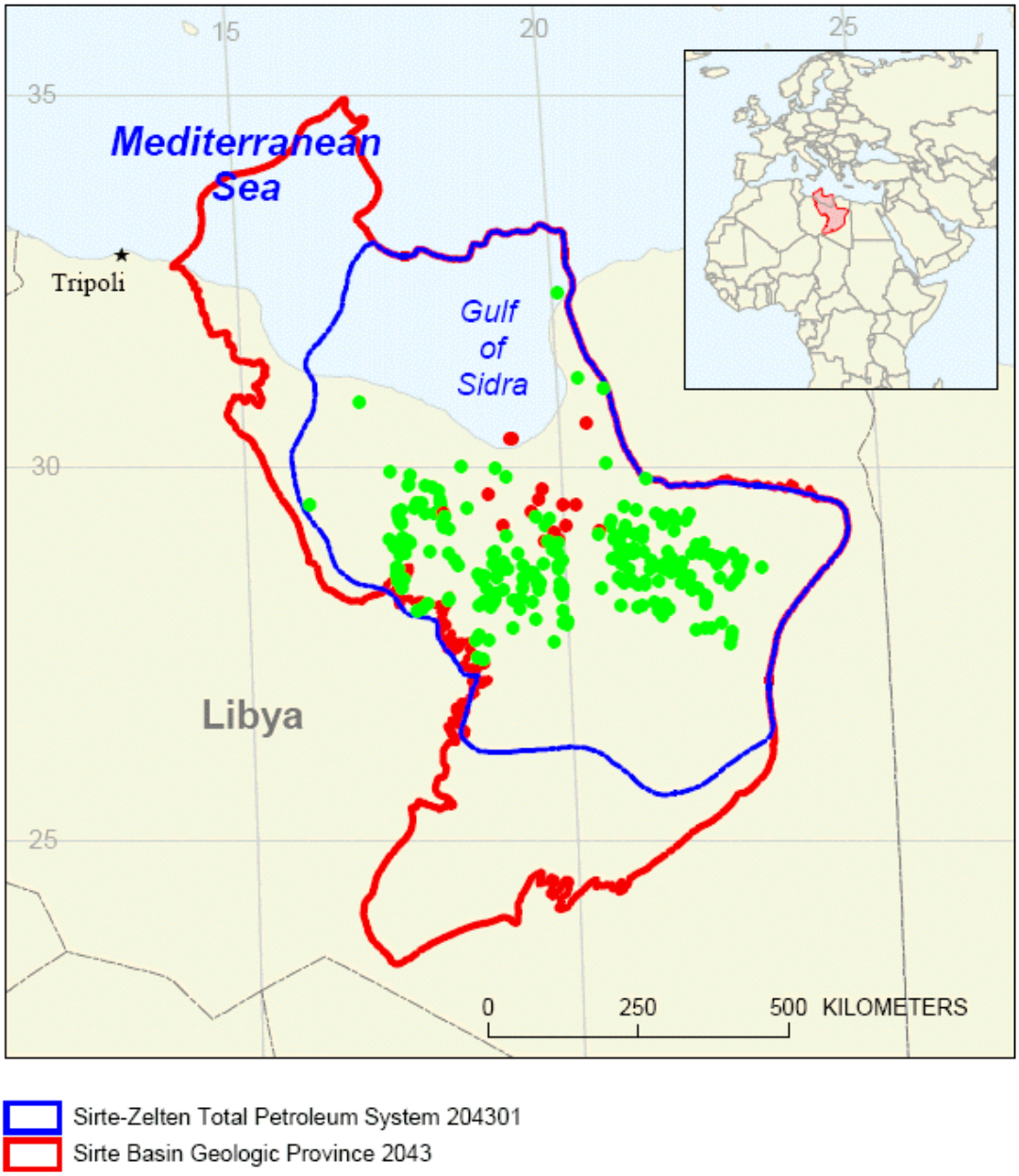


FIG. A.2: Province géologique du Bassin de Sirte (sources : USGS (2000))

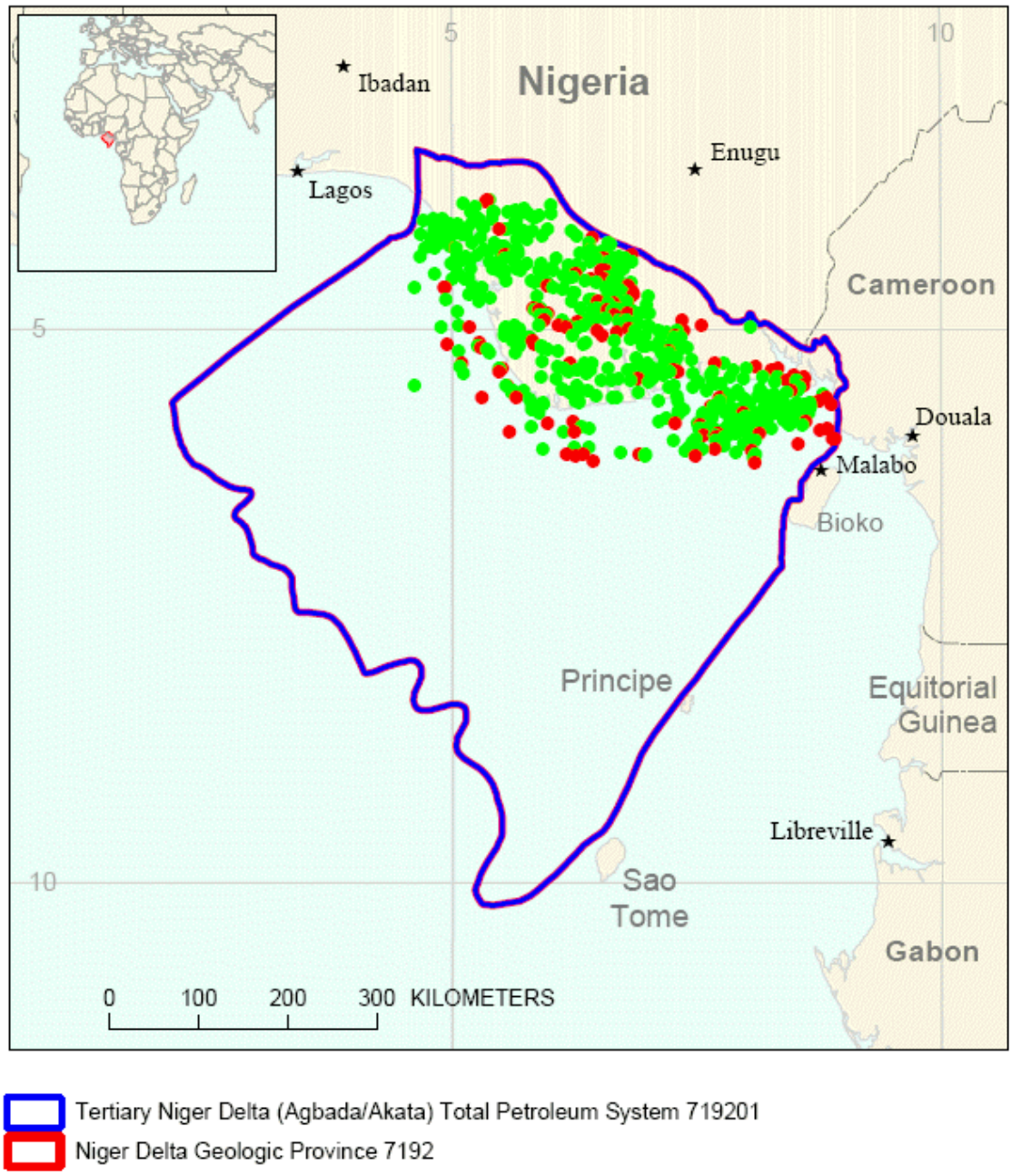


FIG. A.3: Province géologique du Delta du Niger (sources : USGS (2000))

Annexe B

Rappels sur les processus de Poisson

Nous rappelons ici les notions essentielles sur les mesures de Poisson en relation avec les processus de Lévy. Toutes les démonstrations des résultats énoncés ci dessous sont bien connus et peuvent être trouvées dans Kingman (1975) et Bertoin (1996).

B.1 Mesures de Poisson et processus ponctuels de Poisson

Soit E un espace Polonais et ν une mesure σ-finie sur E.

Définition B.1.1. Soit Φ une mesure aléatoire sur E. On dit que Φ est une mesure de Poisson d'intensité ν si :

- Pour tout borélien A de E tel que $\nu(A) < \infty$, $\Phi(A)$ suit une loi de Poisson de paramètre $\nu(A)$.
- Pour tous boréliens deux à deux disjoints $B_1, ..., B_n$, alors $\Phi(B), ..., \Phi(B)$ sont des variables aléatoires indépendantes.

Une mesure de Poisson Φ peut s'écrire comme une somme dénombrable de masses de Dirac :

$$\Phi = \sum_{i \in I} \delta_{e_i}(de).$$

En ajoutant une composante temporelle, c'est-à-dire en considérant la mesure de Poisson $\bar{\Phi}$ sur l'espace produit $[0, \infty[\times E$ d'intensité $\mu = dt \otimes \nu$, il est possible de définir un processus $(e(t))_{t \geq 0}$ ordonnant les atomes de Φ. On note

$$\bar{\Phi} := \sum_{i \in I} \delta_{(t_i, e_i)}(dt, de) \ ,$$

et on peut de plus montrer que tous les t_i sont p.s. distincts.

Soit γ un point cimetière, le *le processus ponctuel de Poisson* d'intensité ν est défini par :

$$e(t) = \begin{cases} e_i & \text{si } t = t_i, i \in I \\ \gamma & \text{si} t \notin \{t_i, i \in I\}. \end{cases}$$

On a donc

$$\{t_i, i \in I\} = \{t \geq 0, e(t) \neq \gamma\}.$$

On utilisera aussi la notation abusive $\sum_{t\geq 0} \delta_{e(t)}(de)$ pour $\sum_{i\in I} \delta_{e_i}(de)$.

Proposition B.1.1. *Soit $\sum_{t\geq 0} \delta_{(t,e(t))}$ un nuage Poissonnien d'intensité $dt \otimes \nu(de)$. Alors*

- *si $\nu(E) = \infty$, $\{t_i, i \in I\}$ est dense dans $[0, +\infty)$,*
- *si $\nu(E) < \infty$, les sauts forment un processus ponctuel de Poisson d'intensité $\nu(E)$ et l'ensemble des temps de saut est discret. De plus, les temps d'attente entre les sauts T_n sont des variables indépendantes de loi exponentielle de paramètre $\nu(E)$. Enfin, les sauts $(e(T_n))_{n\geq 1}$ sont i.i.d. de loi $\frac{\nu(.)}{\nu(E)}$.*

Citons maintenant deux formules importantes, tout d'abord, la formule exponentielle est un résultat très utile pour manipuler les processus ponctuels de Poisson.

Proposition B.1.2. *Soit f une fonction borélienne sur $E \cup \{\gamma\}$, à valeurs complexes, telle que $f(\gamma) = 0$ et $\int_E \nu(d\epsilon)|1 - e^{f(\epsilon)}| < \infty$. Alors, $\forall t \geq 0$:*

$$\mathbb{E}\Big(\exp\{\sum_{0\leq s\leq t} f(e(s))\}\Big) = \exp\Big\{-t\int_E \nu(d\epsilon)(1 - e^{f(\epsilon)})\Big\}.$$

Nous aurons aussi besoin de la formule de Palm (voir Bertoin, 2006, p.79).

Proposition B.1.3. *Soit f une fonction borélienne sur $E \cup \{\gamma\}$ à valeurs positives, et $\Phi = \sum_{i\in I} \delta_{e_i}(de)$ une mesure de Poisson d'intensité ν. On note $\mathcal{M}_E$ l'ensemble des mesures σ-finies sur E. Soit $F : \mathcal{M}_E \to \mathbb{R}_+$ une fonction mesurable. Alors :*

$$\int \Phi(de)f(e)F(\Phi - \delta_e) = \mathbb{E}\big[F(\Phi)\big] \times \int \nu(de)f(e).$$

Ou encore

$$\mathbb{E}\big[\sum_{i\in I} f(e_i)F(\sum_{j\in I-\{i\}} \delta_{e_j})\big] = \mathbb{E}\big[F(\Phi)\big] \times \int \nu(de)f(e).$$

Cette formule nous apprend qu'un point retiré "au hasard" du nuage Poissonnien selon $\Phi(de)$ est indépendant du nuage restant, et ce dernier a même loi que Φ.

B.2 Processus de Lévy et subordinateurs stables

Définition B.2.1. La loi d'une variable aléatoire Y est dite infiniment divisible si, pour tout n, il existe des variables aléatoires $Y_{n,1}, ..., Y_{n,n}$ i.i.d. telles que

$$Y = Y_{n,1} + ... + Y_{n,n}$$

On rappelle que l'exposant caractéristique Ψ de la loi ν d'une variable aléatoire Y est défini par

$$\mathbb{E}(e^{i\langle Y,u\rangle}) = e^{-\Psi(u)}.$$

Le point de départ de l'étude des lois infiniment divisibles est la formule de Lévy-Khintchine qui caractérise les exposants caractéristiques des lois infiniment divisibles.

Théorème B.2.1. *Une fonction Ψ est l'exposant caractéristique d'une loi infiniment divisible sur $\mathbb{R}^d$ si et seulement s'il existe $a \in \mathbb{R}^d$, Q une forme quadratique semi définie positive sur $\mathbb{R}^d$ et Λ une mesure sur $\mathbb{R}^d - \{0\}$ vérifiant $\int (1 \wedge |x|^2)\Lambda(dx) < \infty$ tels que pour tout $u \in \mathbb{R}^d$,*

$$\Psi(u) = i\langle a, u\rangle + \frac{1}{2}Q(u) + \int_{\mathbb{R}^d} \left(1 - e^{i\langle u,x\rangle} + i\langle u, x\rangle 1_{|x|<1}\right)\Lambda(dx) \tag{B.1}$$

Définition B.2.2. Un processus de Lévy $(Z_t, t \geq 0)$ est un processus à accroissements indépendants stationnaires.

On supposera de plus ici que $Z_0 = 0$, et que Z est cadlag (continue à droite et possédant des limite à gauche en tout point), ce qui est toujours possible. Alors, pour tout $t \geq 0$, la loi de Z_t est infiniment divisible car :

$$\forall n \geq 0, \quad Z_t = Z_{\frac{t}{n}} + (Z_{\frac{2t}{n}} - Z_{\frac{t}{n}}) + \ldots + (Z_t - Z_{\frac{(n-1)t}{n}}).$$

Notons Ψ l'exposant caractéristique de Z_1. Par le même argument, on montre que

$$\mathbb{E}[e^{iuZ_t}] = e^{-t\Psi(u)}.$$

La loi de Z est donc entièrement caractérisée par Ψ, que l'on appellera aussi exposant caractéristique du processus de Lévy. Réciproquement, à toute loi infiniment divisible ν on peut associer un processus de Lévy issu de 0 tel que Z_1 ait pour loi ν. L'exposant caractéristique d'un processus de Lévy Z peut donc s'écrire sous la forme (B.1), la mesure Λ et la forme quadratique Q sont appelés respectivement *mesure de Levy* et *coefficient Gaussien* du processus de Lévy.

La propriété suivante montre comment les mesures de Poisson interviennent dans la théorie des processus de Lévy.

Proposition B.2.1. *Soit $(Z_t, t \geq 0)$ un processus de Lévy. On pose $X_t := Z_t - Z_{t^-}$. Alors, le processus des sauts $(X_t)_{t\geq 0}$ est un processus ponctuel de Poisson d'intensité Λ. Autrement dit, $\sum_{t\geq 0} \delta_{(t,X_t)}(dt.dx)$ est un nuage Poissonien d'intensité $dt \otimes \Lambda$.*

Annexe C

Étude par simulations du critère pénalisé pour la sélection de modèles de mélanges gaussiens

C.1 Notations

Cette section est consacrée à l'étude par simulations des performances du critère pénalisé proposé à la section 5.2 du chapitre 5 pour la sélection de modèles de mélanges gaussiens. Nous montrons que le critère, combiné avec la méthode de la pente, permet de définir un estimateur pénalisé dont l'erreur de prédiction est comparable à celle de l'oracle. Nous comparons aussi l'estimateur pénalisé aux estimateurs déduits pour les critères AIC, BIC and ICL. Avec les notations du chapitre 5.2, rappelons que ces critères sont définis par

$$\mathrm{crit}_{\mathrm{AIC}}(D) = \gamma_n(\hat{s}_D) + \frac{D}{n}$$

$$\mathrm{crit}_{\mathrm{BIC}}(D) = \gamma_n(\hat{s}_D) + \frac{D\ln(n)}{2n}$$

$$\mathrm{crit}_{\mathrm{ICL}}(D) = \mathrm{crit}_{\mathrm{BIC}} - \frac{\mathrm{ENT}}{n}$$

avec

$$\mathrm{ENT} = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\ln(t_{ik})$$

où z est donné par la règle du MAP (voir section 4.4.2) et

$$t_{ik} = \frac{\hat{p}_k\Phi(y_i|\hat{\mu}_k,\hat{\Sigma}_k)}{\sum_{l=1}^{K}\hat{p}_l\Phi(y_i|\hat{\mu}_l,\hat{\Sigma}_l)}.$$

Le lecteur pourra consulter Akaike (1973, 1974), Schwarz (1978) et Biernacki *et al.* (2000) pour plus de détails sur ces critères.

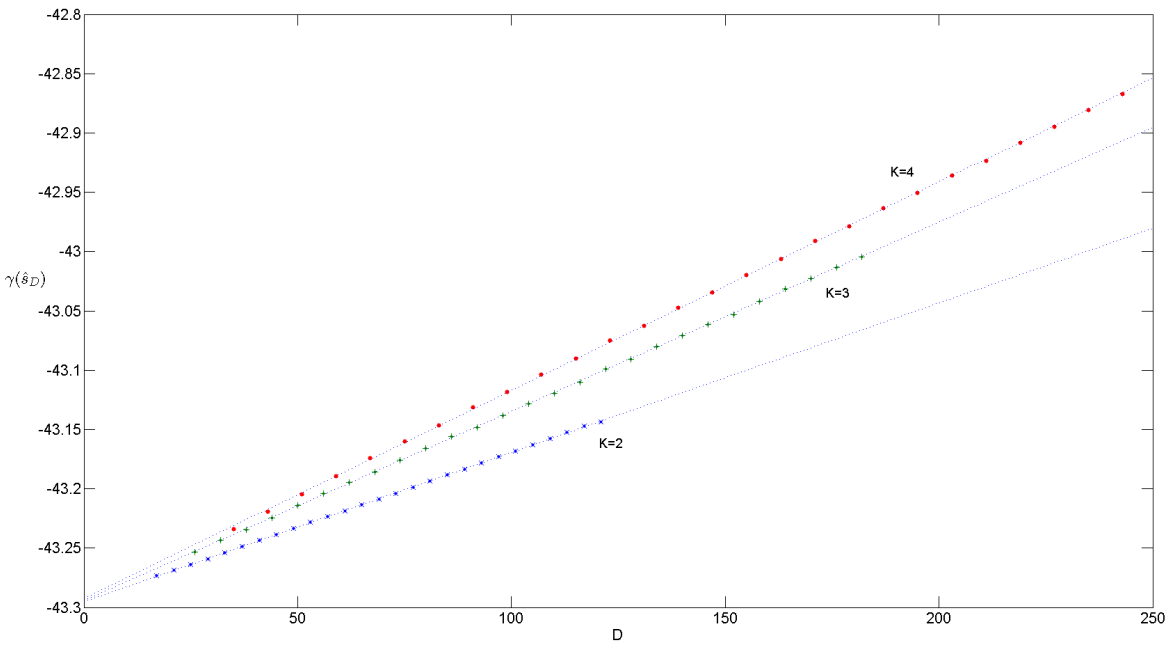


FIG. C.1: Estimation de la fonction $v \mapsto -\mathbb{E}\left[\gamma_n(\hat{s}_{(K,v)})\right]$ obtenue pour $K = 2, 3$ et 4.

C.2 Pénalités minimales

Dans le chapitre 6, et dans les simulations qui suivent, nous utilisons des pénalités de la forme $\eta\frac{D(K,v)}{n}$. L'objectif de cette première simulation est de s'assurer sur un exemple simple que la pénalité minimale ne comporte pas de terme multiplicatif supplémentaire en "$\ln v$, c'est-à-dire que la pénalité minimale n'est pas de la forme $\eta\frac{D(K,v)}{n}\ln v$ comme pourraient le faire croire les résultats théoriques obtenus au chapitre 5. De façon rigoureuse, la pénalité minimale est définie dans l'heuristique de pente par (voir par exemple Arlot, 2007, p.38)

$$\text{pen}_{\min} = \mathbb{E}\left[\gamma_n(s_{(K,v)}) - \gamma_n(\hat{s}_{(K,v)})\right] \tag{C.1}$$

où $s_{(K,v)} = \underset{t\in S_{(K,v)}}{\operatorname{argmin}}\,\text{KL}(s,t)$.

Considérons un échantillon $(Y_1, \ldots, Y_{200})$ où Y_i est un vecteur de dimension $Q = 30$ de distribution de densité s définie comme suit. Pour tout i, le vecteur (Y_{i1}, Y_{i2}, Y_{i3}) a la distribution d'un mélange gaussien sur $\mathbb{R}^3$ à deux composantes. Plus précisément, les variables de classification sont simulées à partir de deux distributions gaussiennes équiprobables $\mathcal{N}(\mu_k, \Sigma_k)$ avec

$$\mu_1 = -\mu_2 = (1,1,1) \text{ et } \Sigma_1 = \Sigma_2 = I_3.$$

Un bloc de 27 variables indépendantes de loi $\mathcal{N}(0,1)$ est ajouté à la suite pour chaque individu.

Nous utilisons la collection de modèles ordonnés $\mathcal{M}[L_kB_k]$. Notons que pour $K \geq 2$ et $v \geq 3$, $s \in S_{(K,v)}$ et donc $s_{(K,v)} = s$. Pour de tels couples (K,v), la pénalité minimale définie en (C.1) a donc pour expression $\text{pen}_{\min} = \gamma(s) - \mathbb{E}\left[\gamma_n(\hat{s}_{(K,v)})\right]$. Pour estimer cette espérance, nous simulons 300 fois l'échantillon $(Y_1, \ldots, Y_200)$ et par une procédure de Monte-Carlo nous obtenons ainsi facilement une estimation de la pénalité minimale pour $K \geq 2$

et $v \geq 3$. La figure C.1 représente pour $K = 2, 3$ et 4 l'estimation obtenue pour la fonction $v \mapsto -\mathbb{E}\left[\gamma_n(\hat{s}_{(K,v)})\right]$. Si la pénalité minimale était effectivement en $\frac{D(K,v)}{n} \ln v$, nous devrions observer une stricte concavité pour chacune des courbes de la figure C.1, ce qui n'est pas le cas. Cette première simulation nous conforte donc dans l'utilisation de pénalités proportionnelles à la dimension des modèles.

C.3 Comparaison à l'oracle et à d'autres critères

Pour ce deuxième exemple, les données sont composées d'un échantillon de $n = 2000$ individus décrits par par $Q = 22$ variables dont les dix premières sont les variables de classification qui sont de plus indépendantes entre elles. Plus précisément, les variables de classification sont simulées à partir de quatre distributions gaussiennes équiprobables $\mathcal{N}(\mu_k, \Sigma_k)$ avec

$$\begin{aligned} \mu_1 &= (3, 2, 1, 0.7, 0.3, 0.2, 0.1, 0.07, 0.05, 0.025)\ , \mu_2 = 0_{10}\ , \mu_3 = -\mu_1\ , \\ \mu_4 &= (3, -2, 1, -0.7, 0.3, -0.2, 0.1, -0.07, -0.05, -0.025)\ , \end{aligned}$$

et

$$\Sigma_1 = \Sigma_3 = \Sigma_4 = I_{12} \text{ et } \Sigma_2 = \mathrm{Diag}(2, 1.9, 1.8, \ldots, 1.1).$$

Le vecteur 0_{10} désigne le vecteur nul de longueur 10. Un bloc de 12 variables indépendantes de loi $\mathcal{N}(0, 1)$ est ajouté à la suite pour chaque individu. En conséquence, le "vrai modèle" correspond au couple $(K_0, v_0) = (4, 10)$. Notons que le niveau de discrimination décroît le long des 22 variables. En d'autres termes, les quatre sous-populations du mélange sont progressivement réunies en une seule distribution gaussienne, comme le montre la figure C.2.

Nous utilisons la collection ordonnée de modèles $\mathcal{M}[L_k B_k]$. Après l'étape d'estimation, la fonction $D \mapsto -\gamma_n(\hat{s}_D)$ est tracée (voir figure C.3). Pour $D \geq D_0 = 140$, on observe que la fonction $-\gamma_n(\hat{s}_D)$ a un comportement linéaire, comme le prévoit l'heuristique de pente qui a été exposée à la section 6.1. Une estimation de $\hat{C}$ est ainsi obtenue, ce qui permet de calibrer la pénalité. Le modèle sélectionné pour ce premier jeu simulé est le modèle pour lequel $\hat{K}_{\text{pente}} = 4$ et $v_{\text{pente}} = 8$.

Afin de comparer le comportement de l'estimateur de la pente avec les estimateurs correspondant aux critères AIC, BIC and ICL, cette procédure est répétée 1000 fois avec un nouvel échantillon de longueur 2000 simulé à chaque fois. Puisque la véritable densité est connue, il est possible ici d'évaluer par une procédure de Monte-Carlo la valeur de l'oracle : $K_{\text{oracle}} = 4$ et $v_{\text{oracle}} = 9$. Les résultats obtenus sont portés dans le tableau C.1. Les deux critères "asymptotiques" BIC et ICL sélectionnent la plupart du temps un modèle de mélange à 4 composantes et 6 variables de classification. Concernant le critère BIC, on considère généralement que celui-ci vise à trouver le "vrai" modèle. Ainsi, les travaux de Keribin (2000) montrent que la procédure de sélection par BIC est consistante vis-à-vis du nombre de composantes d'un mélange gaussien. Cependant, ce résultat de consistance n'a pas été démontré pour le problème de sélection de variable et de classification qui nous intéresse ici. D'après le tableau C.1, le modèle sélectionné par BIC est loin du vrai modèle. S'il y a réellement consistance du critère

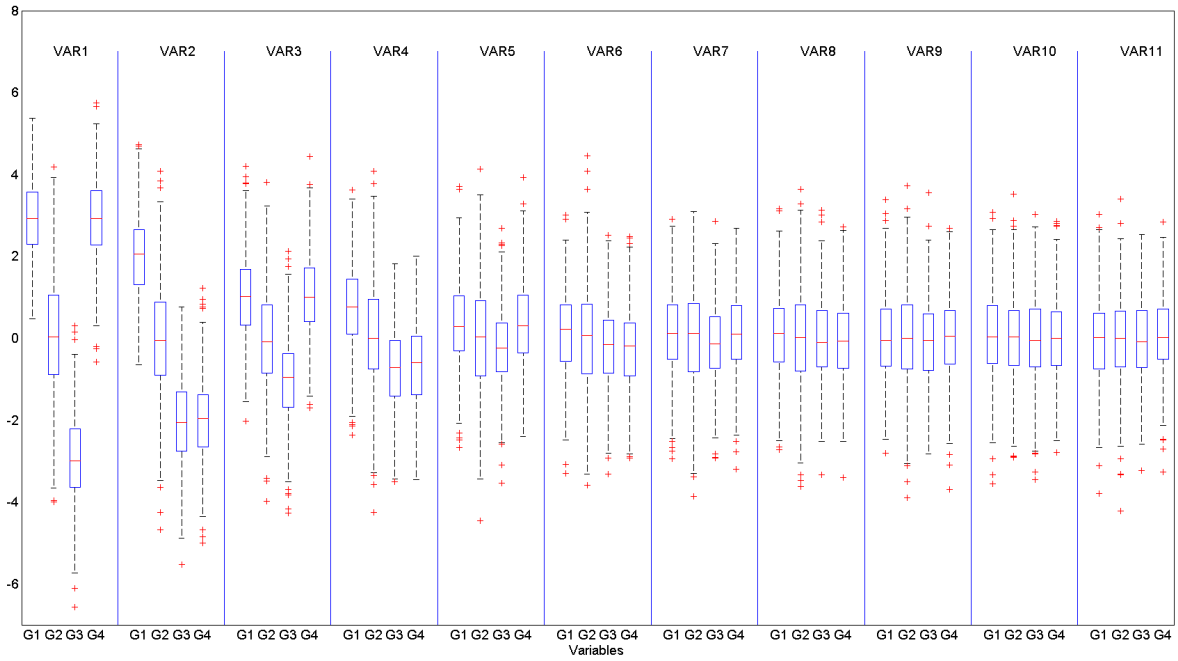


FIG. C.2: Boxplots des 11 premières variables (VAR1,...,VAR11) des quatre composantes (G1,G2,G3,G4) du mélange.

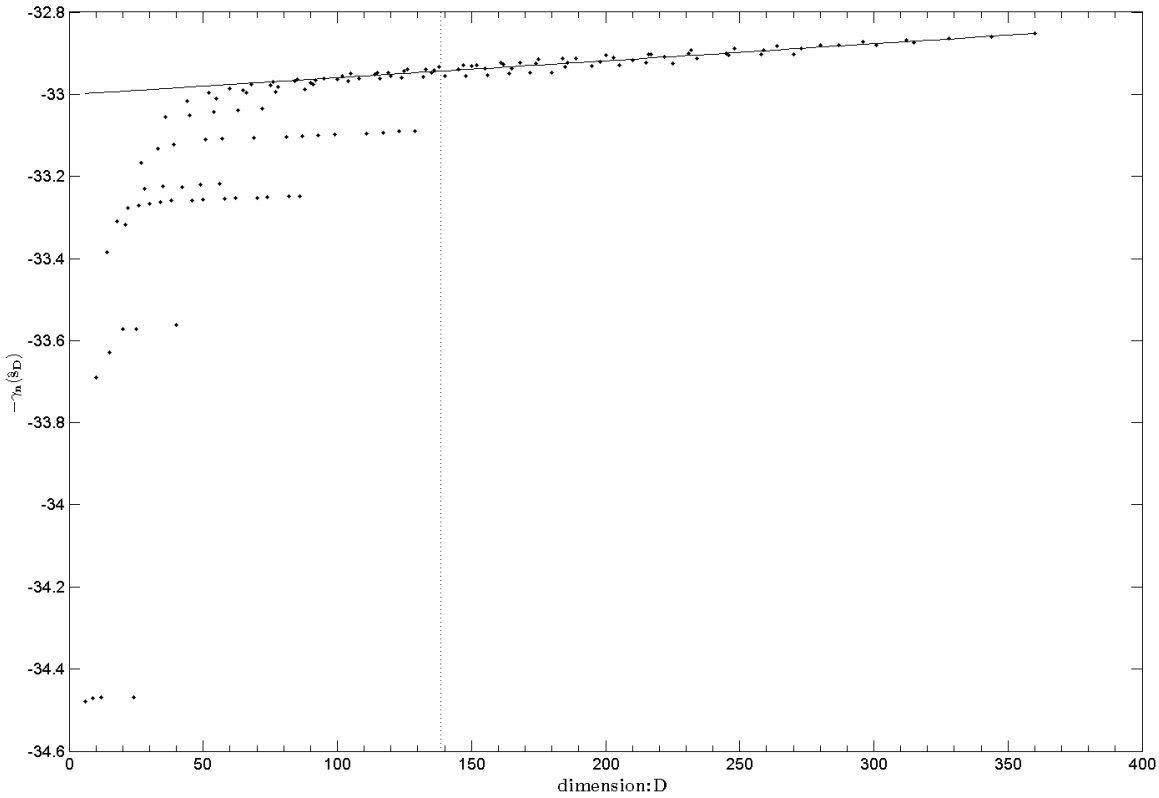


FIG. C.3: Graphique de la fonction$(K, v) \rightarrow \gamma_n(\hat{s}_{(K,v)})$.

criterion	K	v								
		5	6	7	8	9	10	11	12	≥ 13
ICL	4	22	792	184	2					
BIC	4	29	859	111	1					
AIC	≤ 5			2	21	26	11	3	2	
	6			7	42	62	26	9	2	3
	≥ 7			57	155	237	170	93	41	41
Méth. pente	4		43	417	456	58	1			
	5		8	13	13	4				

TAB. C.1: Synthèse des modèles sélectionnés par chaque méthode. Le tableau donne le nombre de simulations parmi les 1000 qui conduisent à la sélection d'un certain modèle (K, v).

BIC dans ce contexte, il semble que l'échantillon considéré ne soit pas de taille suffisante. Le comportement du critère ICL n'est pas surprenant car l'objectif de ce critère est avant tout de déterminer un modèle de mélange qui fournisse une partition des données convenable vis-à-vis de l'entropie. Le critère AIC conduit à sélectionner des modèles de trop grandes dimensions. Dans une telle situation ce dernier critère n'est pas du tout performant.

Finalement, du point de vue de la classification, les méthode BIC, ICL et la pente ont des performances comparables ici. L'intérêt principal de cette première étude par simulation est de permettre de comparer les comportements des différents critères entre eux. Comme attendu, nous vérifions que la méthode de la pente sélectionne un modèle qui est proche de celui de l'oracle.

C.4 Base de données waveforms

La base de données "waveforms" est composée d'un échantillon de 5000 observations réparties en trois groupes. Une description complète de la construction de la base est disponible dans Breiman *et al.* (1984) et les données sont disponibles sur le site de l'UCI[1]. Les données sont créées à partir de combinaisons convexes des trois fonctions triangles h_1, h_1 et h_3 représentées sur la figure C.4. Chaque observation est décrite par 40 variables, qui sont définies comme suit. Soient la variable aléatoire $U \sim \mathcal{U}([0,1])$ et $\varepsilon_1, \ldots, \varepsilon_{21}$ un échantillon de même distribution $\mathcal{N}(0,1)$. Pour un individu i de la classe 1, et pour tout $j \in \{1 \ldots 21\}$ on pose

$$Y_{ij} = U h_2(j) + (1 - U)\ h_3(j) + \varepsilon_j$$

et pour les deux autres classes les variables Y_{ij}, pour $j \in \{1 \ldots 21\}$, sont définies de la même façon à partir des fonctions h_1, h_1 et h_3. Les variables Y_{ij} pour $j \in \{22, \ldots, 40\}$ sont indépendantes de même distribution $\mathcal{N}(0,1)$ quelle que soit la classe de l'individu i. De plus, pour tout i le vecteur $(Y_{i1}, \ldots, Y_{i21})$ est indépendant du vecteur $(Y_{i22}, \ldots, Y_{i40})$. Il y a donc $v_0 = 19$ "véritables" variables de classification. Toutes les variables sont ensuite centrées, mais pour

[1]Voir `http://archive.ics.uci.edu/ml/datasets/Waveform+Database+Generator+`

cet exemple particulier elles ne sont pas réduites. En effet, les 40 variables ne sont pas naturellement ordonnées car il n'est pas évident de savoir lesquels des 19 premières variables sont les plus pertinentes pour la classification et nous proposons de réordonner les 19 premières variables par ordre décroissant de variance.

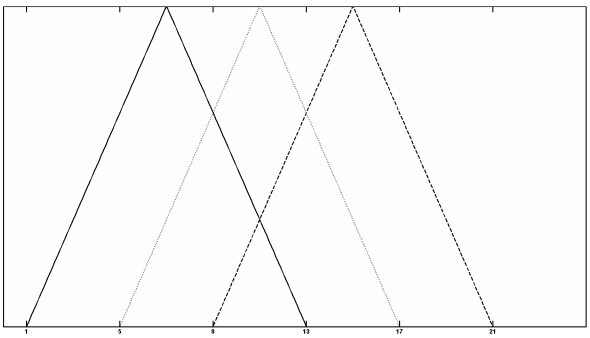


FIG. C.4: Les trois fonctions "triangle" utilisées pour définir les données.

Puisque les variables ne sont pas réduites, nous utilisons ici des collections de modèles de la forme

$$S_{(K,v)} = \left\{x \in \mathbb{R}^Q \mapsto f(x_1, \ldots, x_v)\, \Phi(x_{v+1}, \ldots, x_Q \,|\, 0, \omega^2 I_{Q-v})\ ;\ \ f \in \mathcal{L}_{(K,v)}\right\}$$

où ω^2 est la variance commune des variables non utilisées pour la classification. Il est possible de démontrer pour ce type de collection de modèles des résultats similaires à ceux qui ont été obtenus au chapitre 5 ; ces collections sont spécifiquement étudiées dans Maugis et Michel (2008a) et Maugis et Michel (2008b). Les densités de mélange de l'ensemble $\mathcal{L}_{(K,v)}$ sont choisies de la forme $[p_k L_k B_k]$ (matrices diagonales définies positives) et $[p_k L_k C_k]$ (matrices définies positives), en reprenant les notations utilisées dans Biernacki *et al.* (2006). Les graphiques des fonctions $D \mapsto -\gamma_n(\hat{s}_D)$ pour l'estimation de la pente $\hat{\eta}$ sont représentés dans les figures C.6 et C.7 pour les deux collections. Nous vérifions sur les figures C.8 et C.9 pour chacune des collections $[p_k L_k B_k]$ et $[p_k L_k C_k]$ que la nappe du nombre de paramètres $(K, v) \mapsto D(K, v)$ s'ajuste correctement sur la nappe des log-vraisemblances $(K, v) \mapsto \gamma_n(\hat{s}_{(K,v)})$. L'ajustement est bon pour la collection des mélanges "diagonaux" alors qu'il est satisfaisant pour la collection des mélanges "généraux". La méthode de la pente appliquée à la collection des mélanges "généraux" conduit à sélectionner un modèle à $\hat{K} = 3$ composantes et les 19 "véritables" variables de classification. Pour la collection de modèles de mélanges diagonaux $[p_k L_k B_k]$, la procédure sélectionne encore 19 variables mais avec cette fois 10 composantes.

Cette simulation permet de souligner l'importance du choix d'une collection de mélanges convenable pour appliquer la méthode de la pente. Les variables de classification de la vraie densité n'ont pas pour distribution un mélange gaussien et par construction, celles-ci sont dépendantes entre elles. Il est donc naturel que la collection de modèles de mélanges diagonaux $[p_k L_k B_k]$ ne soit pas suffisamment riche pour traiter le problème. De façon générale, le tracé de l' ajustement de la nappe du nombre de paramètres sur la nappe des log-vraisemblances $(K, v) \mapsto \gamma_n(\hat{s}_{(K,v)})$ permet de s'assurer que la collection de modèles de mélanges a été bien choisie.

	cl1	cl2	cl3	total
gp1	1331	185	176	1692
gp2	95	99	1459	1653
gp3	65	1494	96	1655
total	1491	1778	1731	5000

TAB. C.2: Tableau de contingence de la classification obtenue avec la collection de modèles de mélanges $[p_k L_k C_k]$.

Cette simulation permet aussi d'illustrer l'intérêt de la sélection de variables en ce qui concerne la classification. Le tableau C.2 montre qu'avec la sélection de variables, les trois sous-populations sont identifiées avec un taux d'erreur de classement de 14.3%. La figue C.5 représente le taux d'erreur en fonction du nombre v de variables de classification utilisées et chaque courbe correspond à un nombre de composantes K fixé. Il est clair sur le graphique que sélectionner plus de 19 variables a un effet néfaste sur la qualité de la classification.

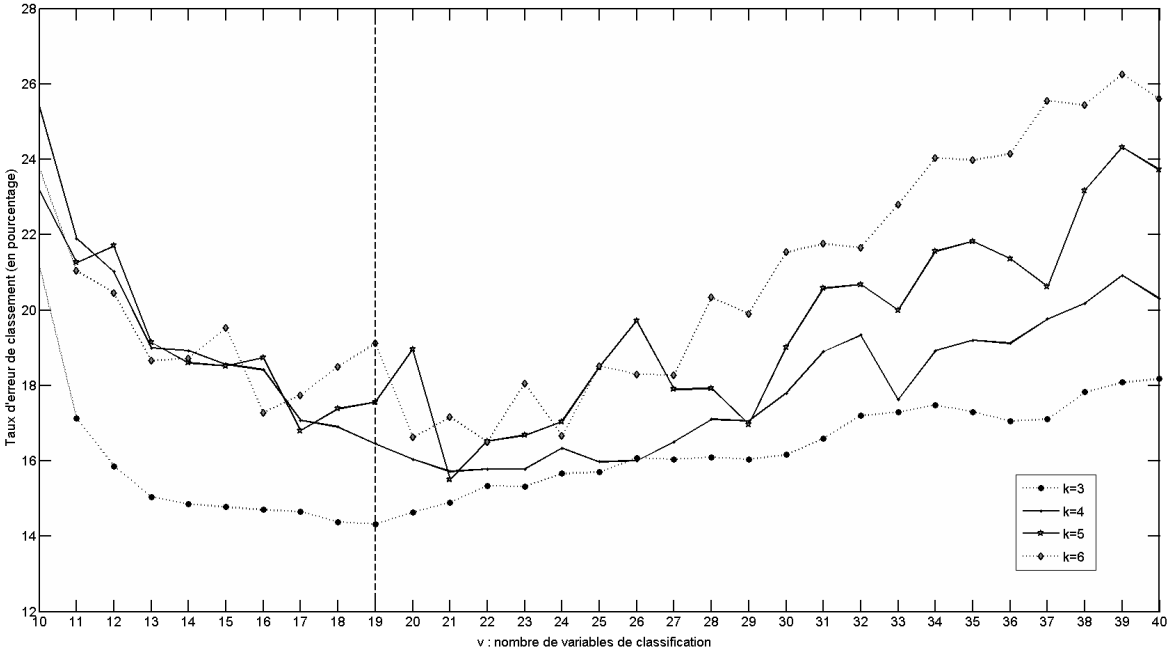


FIG. C.5: Taux d'erreur de classement en fonction du nombre v de variables de classification utilisées. Chaque courbe correspond à un nombre de composantes K fixé.

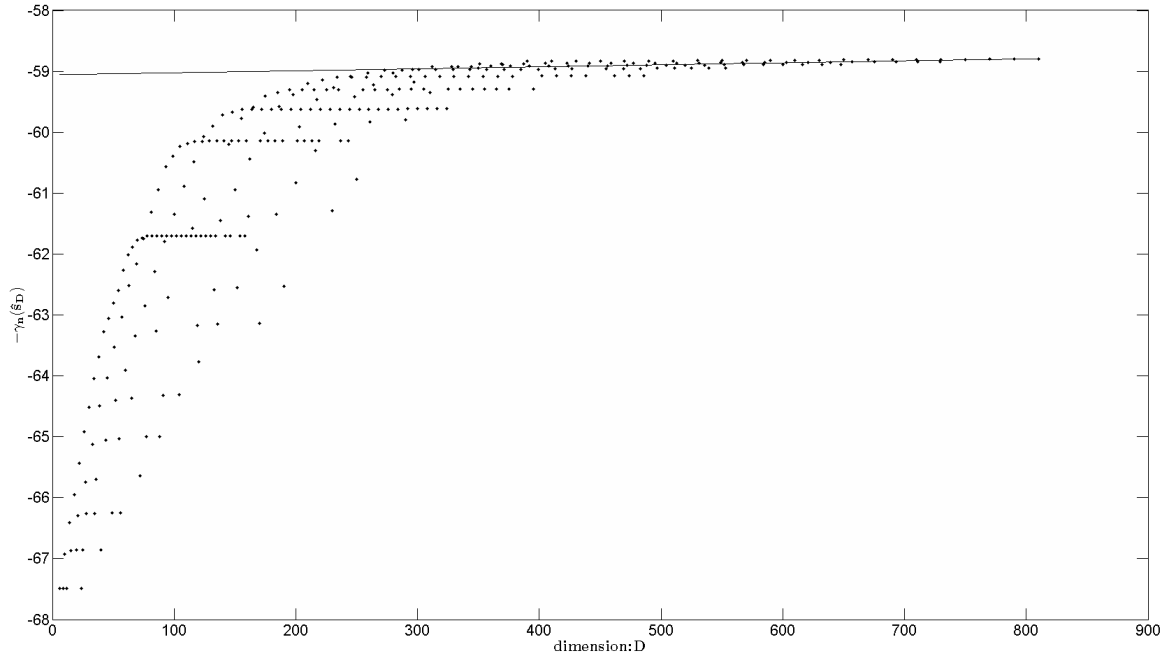


FIG. C.6: Fonction $D \mapsto -\gamma_n(\hat{s}_D)$ pour l'estimation de $\hat{\eta}$ dans le cas de la collection de modèles de mélange diagonaux.

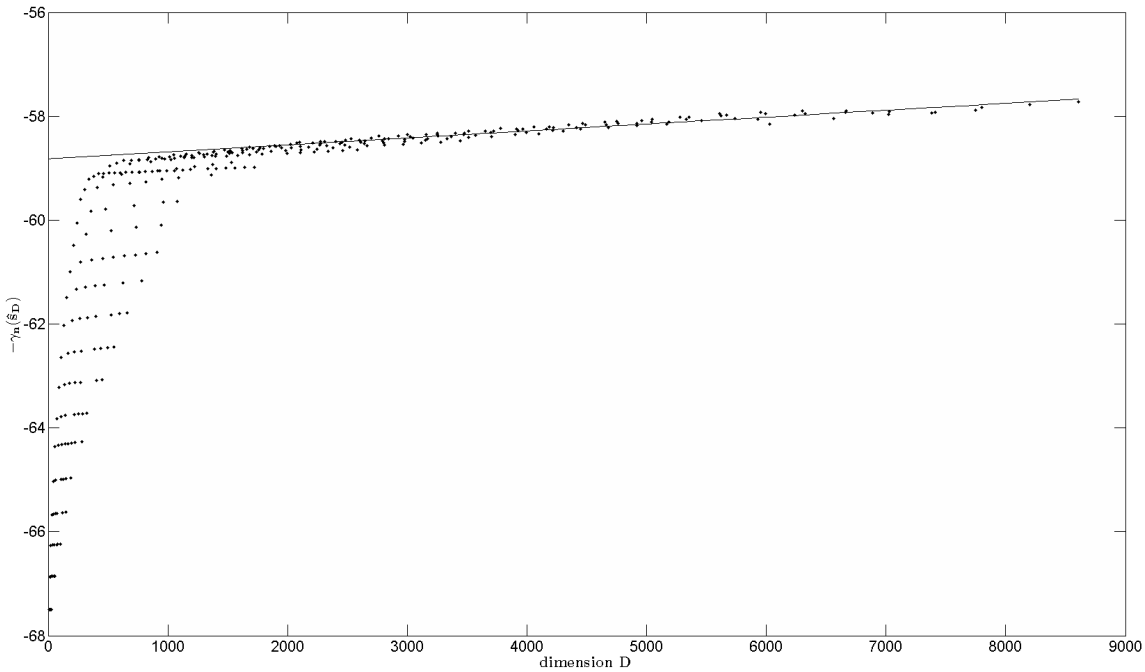


FIG. C.7: Fonction $D \mapsto -\gamma_n(\hat{s}_D)$ pour l'estimation de $\hat{\eta}$ dans le cas de la collection de modèles de mélange généraux.

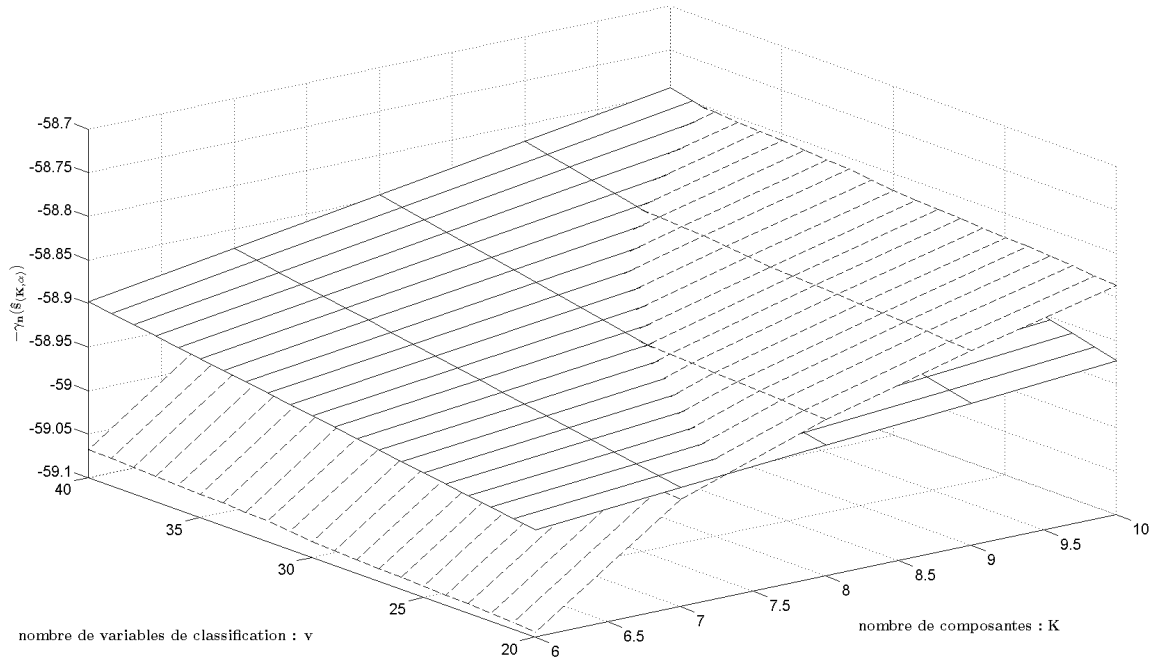


FIG. C.8: Ajustement de la nappe du nombre de paramètres sur la nappe $(K, v) \in [|6, 10|] \times [|20, 40|] \mapsto -\gamma_n(\hat{s}_{(K,v)})$ pour le cas de la collection de modèles de mélanges diagonaux $[p_k L_k B_k]$.

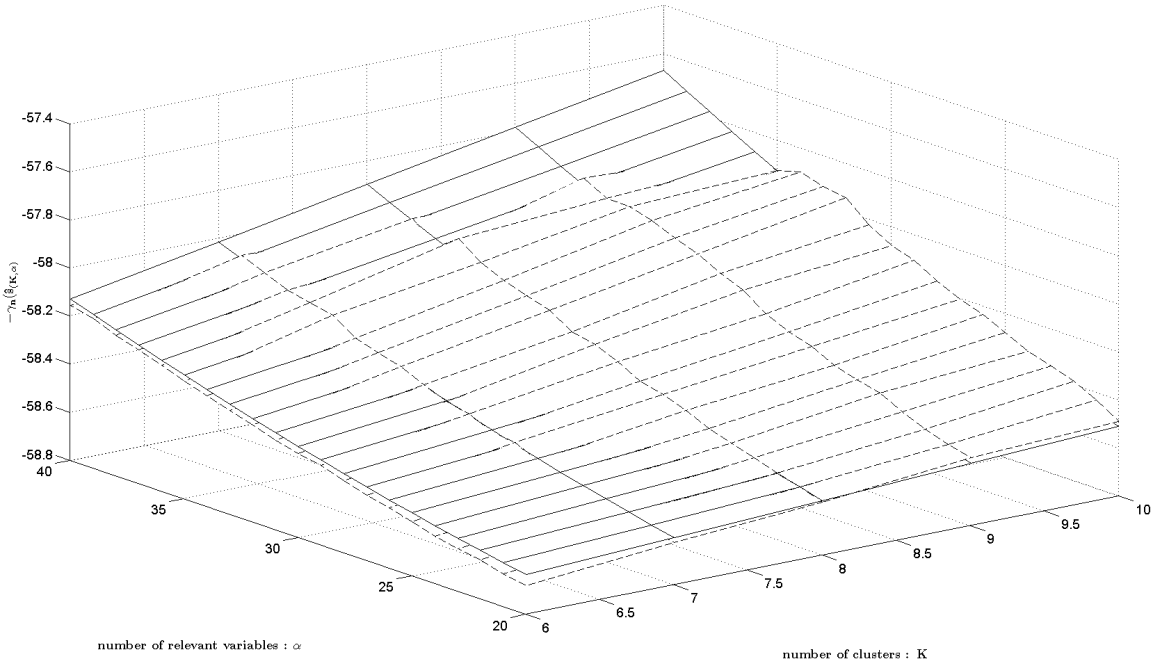


FIG. C.9: Ajustement de la nappe du nombre de paramètres sur la nappe $(K, v) \in [|6, 10|] \times [|20, 40|] \mapsto -\gamma_n(\hat{s}_{(K,v)})$ pour le cas de la collection de modèles de mélanges généraux $[p_k L_k C_k]$.

Annexe D

Annexe pour l'étude du processus d'exploration

D.1 Simulations

Cette section est consacrée à des études par simulations des procédures d'estimation associées aux modèles d'exploration pétrolière.

D.1.1 Estimation pour une partition fixée

Nous étudions tout d'abord les performances de la procédure d'estimation de la fonction de visibilité h sur une partition m fixée. Dans toutes les situations considérées ci-dessous, nous simulons 200 fois un échantillon de taille n_{s} de couples $(X^\star, D^\star)$ de loi de densité $g^\star$ comme en (4.7), avec

- un indice de Pareto $\alpha_{\mathrm{s}} = 0.75$,
- une durée d'exploration de 40 ans : $t^\star = 40$,
- une fonction h définie sur la partition [20 50 140 5000] par les paramètres

$$b = 10^{-4} \quad , a_1 = 3\,10^{-4} \quad , a_2 = 2\,10^{-4}, \quad a_3 = 10^{-4} \ .$$

La loi des dates de découverte correspondante est réaliste puisque la fonction de visibilité est comparable à celle qui est observée pour les découvertes en mer du Nord. Pour évaluer la robustesse de la méthode, nous estimons aussi la fonction h pour des coefficients α et des partitions différentes des valeurs utilisées pour la simulation. Les tableaux D.1 et D.2 synthétisent les résultats obtenus.

Les estimations de h obtenues pour des paramètres α choisis au voisinage de la vraie valeur sont acceptables, même pour des partitions elles aussi différentes de la partition utilisée pour simuler les échantillons. Les figures D.1 et D.2 représentent les fonctions de visibilité obtenues pour des simulations d'échantillon de type 2 (voir tableau D.1). La figure D.3 donne les représentations graphiques des fonctions de visibilité associées aux échantillons 6, 7 et 8 du tableau D.2.

	n_s	α	Partition	Estimations			
				$\mathbb{E}\left\lvert\frac{a_1-\hat{a}_1}{a_1}\right\rvert$	$\mathbb{E}\left\lvert\frac{a_2-\hat{a}_2}{a_2}\right\rvert$	$\mathbb{E}\left\lvert\frac{a_3-\hat{a}_3}{a_3}\right\rvert$	$\mathbb{E}\left\lvert\frac{b-\hat{b}}{b}\right\rvert$
1	200	0.75	$[0, 50, 140, 5000]$	0.19	0.27	0.20	2.13
2	400	0.75	$[0, 50, 140, 5000]$	0.13	0.24	0.12	1.66
3	1000	0.75	$[0, 50, 140, 5000]$	0.09	0.12	0.09	1.30
4	400	0.70	$[0, 50, 140, 5000]$	0.14	0.22	0.11	1.46
5	400	0.85	$[0, 50, 140, 5000]$	0.13	0.25	0.15	1.54

TAB. D.1: Écarts relatifs moyens entre les coefficients simulés et estimés. La partition utilisée pour l'estimation de h correspond à la partition réelle sur laquelle est définie h.

	n_s	α	Partition	Estimations			
				$\mathbb{E}(\hat{a}_1)$	$\mathbb{E}(\hat{a}_2)$	$\mathbb{E}(\hat{a}_3)$	$\mathbb{E}(\hat{b})$
6	400	0.75	$[20, 70, 110, 5000]$	$2.86\,10^4$	$1.95\,10^4$	$1.03\,10^4$	$2.46\,10^4$
7	400	0.75	$[20, 40, 300, 5000]$	$3.15\,10^4$	$1.75\,10^4$	$0.93\,10^4$	$1.17\,10^4$
8	400	0.85	$[20, 40, 300, 5000]$	$2.26\,10^4$	$1.56\,10^4$	$1.02\,10^4$	$0.84\,10^4$

TAB. D.2: Estimations moyennes de h pour des partitions différentes de la partition utilisée pour la simulation. La partition utilisée pour l'estimation de h n'est pas la partition réelle sur laquelle est définie h.

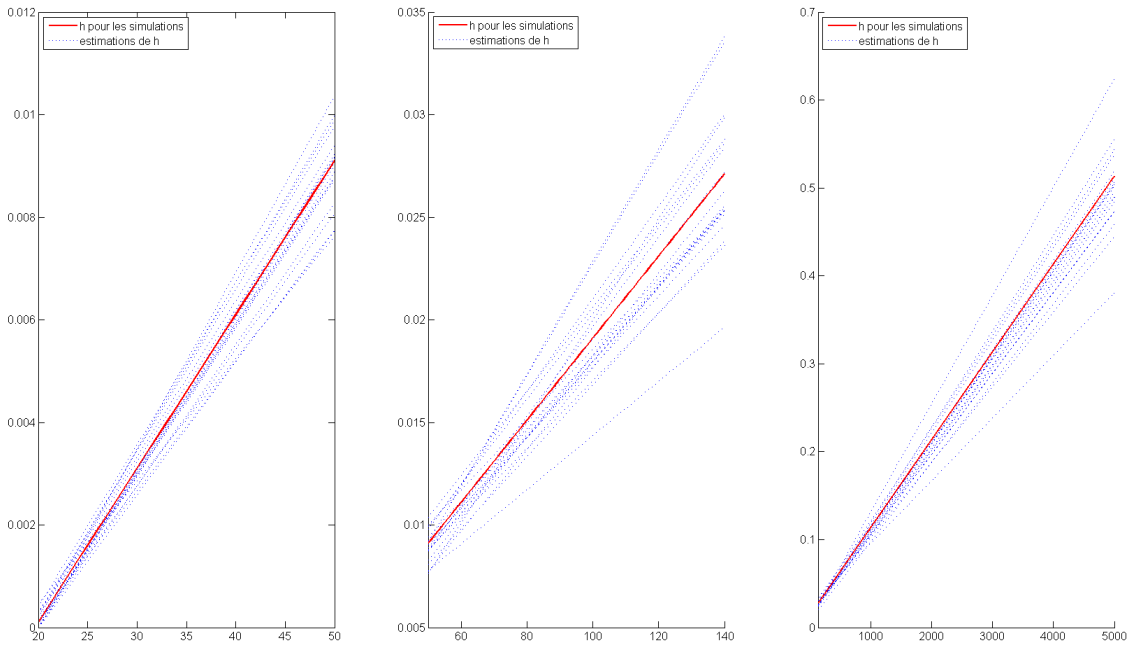


FIG. D.1: Fonction de visibilité utilisée et ses estimations pour 20 simulations d'échantillons de type 2 (voir tableau D.1). Chaque graphique correspond à un intervalle de la partition de $[x_0, x_{\max}]$ utilisée.

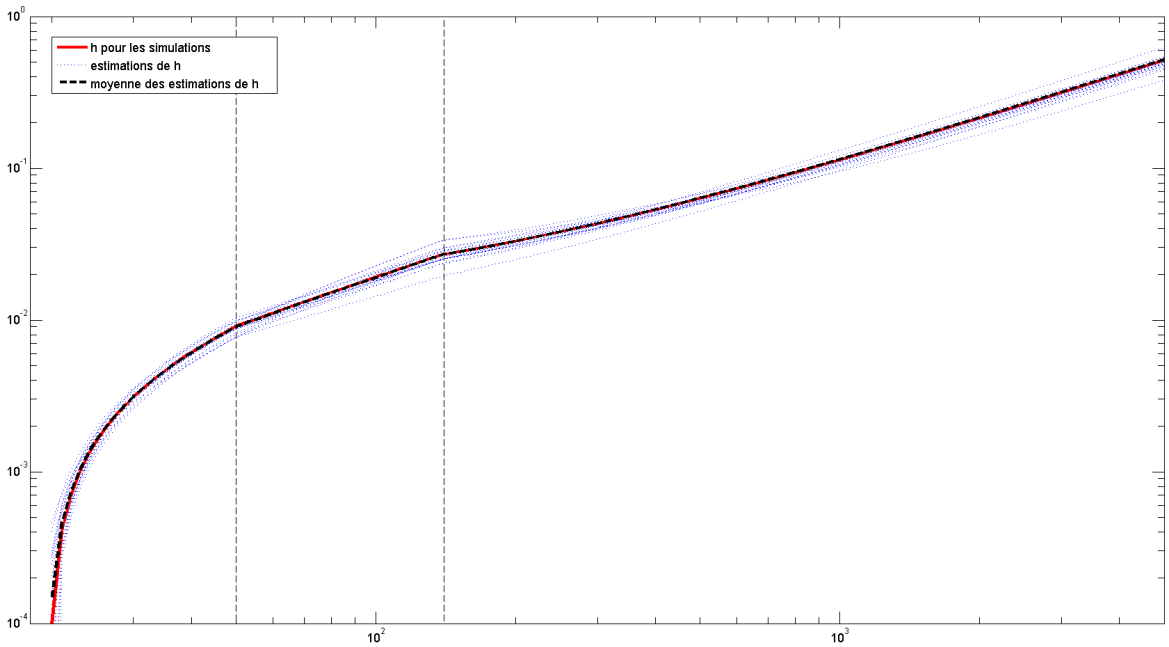


FIG. D.2: Représentation de la fonction de visibilité et de ses estimations pour 20 simulations d'échantillons de type 2 (voir tableau D.1) dans un graphique à double échelle logarithmique.

Notons que le coefficient b est plus difficile à estimer que les autres coefficients ; un échantillon de grande taille est nécessaire pour parvenir à l'estimer correctement. Ceci s'explique par le fait que peu de variables $X^\star$ simulées sont proches du seuil x_0. Ainsi, sur notre exemple, moins de 5% des variables $X^\star$ sont dans l'intervalle [20 30]. Or, ce sont ces observations qui permettent d'améliorer le plus l'estimation de b. Il est donc naturel que ce coefficient soit mal estimé pour des échantillons de l'ordre de quelques centaines d'observations. Heureusement ce phénomène a peu de conséquences sur l'estimation globale de la fonction h comme le montrent les figures D.1 et D.2.

D.1.2 Sélection d'une partition par la méthode de la pente : un exemple simulé

Cette section étudie les performances de la méthode de la pente appliquée à une collection de modèles S_m fixée. Pour définir la collection de modèles, considérons la grille suivante sur $[10, 5000]$

$$\mathcal{G} = [10, 23, 35, 50, 90, 200, 600, 5000].$$

La collection contient tous les modèles S_m associés à une partition m de $[10, 5000]$ construite sur cette grille. Nous obtenons ainsi les modèles de la collection de dimensions entre $D = 2$ (pour $k = 1$) et $D = 8$ (pour $k = 7$). En revanche, un seul modèle est considéré pour les dimensions allant de 9 (pour $k = 8$) à 15 (pour $k = 14$). L'ensemble de toutes ces partitions fournit une collection $\mathcal{M}$ de modèles permettant d'appliquer efficacement la méthode de la pente. Celle-ci comporte à la fois une famille assez riche de modèles de dimensions raisonnables, mais aussi quelques modèles de plus grandes dimensions qui permettent de détecter plus facilement les sauts de dimension. La collection s'inscrit dans le cadre de l'hypothèse (H_2)

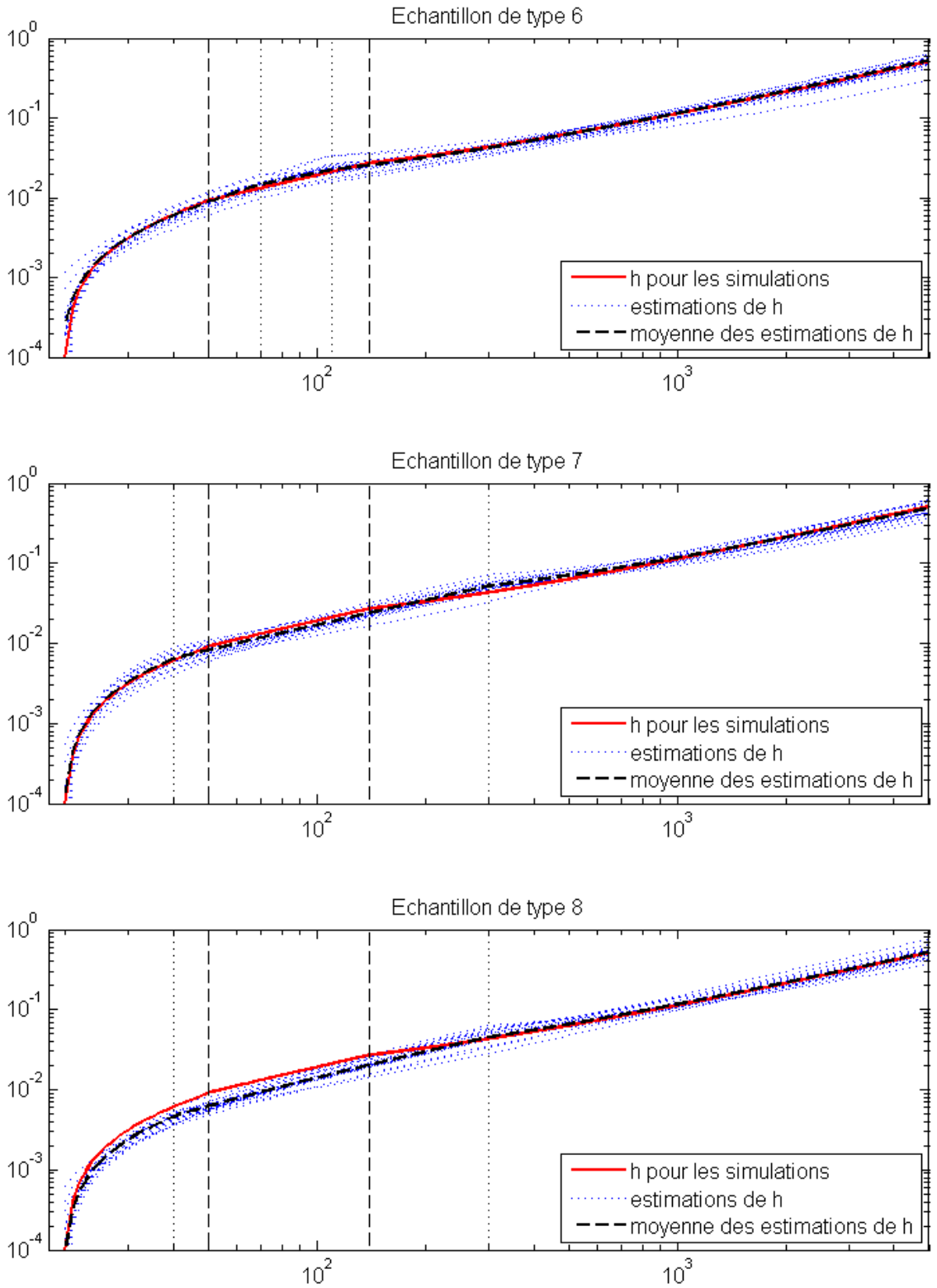


FIG. D.3: Représentations des échantillons 6, 7 et 8 (voir tableau D.2) dans un graphique à double échelle logarithmique.

définie à la section 5.3.1 ce qui nous suggère de rechercher des pénalités minimales de la forme $\text{pen}_{\min}(m) = \eta \frac{D_m}{n} \ln n$.

La procédure suivante est répétée 200 fois.

1. Un échantillon $(X_i^\star, D_i^\star)_{i=1\ldots500}$ est simulé sous la distribution de densité $g^\star$ comme en (4.7), avec les paramètres suivants : l'indice de Pareto des réserves est fixé à 0.70 et la fonction de visibilité est définie sur l'intervalle $[10, 40, 100, 5000]$ avec les paramètres

$$a_1 = 4\ \ 10^{-4}, \quad a_2 = 5\ \ 10^{-6}, \quad a_3 = 2\ \ 10^{-4} \quad \text{et} \quad b = 10^{-4}.$$

2. Dans chacun des modèles de la famille, la fonction de visibilité h est estimée sur la partition correspondante, avec un indice de Pareto fixé cette fois à 0.75. Le paramètre α que nous imposons pour les estimations est volontairement différent de celui que nous utilisons pour simuler les échantillons($\alpha = 0.70$). Ainsi, nous prenons en compte le fait qu'en pratique, le paramètre α imposé dans les procédures d'estimation de h n'est qu'une estimation du paramètre "réel" α de la loi des réserves.
3. Pour chacun des échantillons simulés, la méthode de la pente est utilisée pour sélectionner un modèle de la collection.

La figure D.4 illustre la mise en pratique de la méthode de la pente pour un échantillon simulé particulier (graphique de droite). Les deux fonctions de visibilité réelles et estimées peuvent être comparées sur le graphique de droite. Sur cet exemple, le modèle sélectionné correspond à une fonction de visibilité définie sur une partition à 4 composantes. Notons qu'il n'est pas toujours aussi facile de localiser le saut de dimension et de déterminer ainsi la pénalité minimale ; dans certains cas heureusement relativement rares, plusieurs sauts successifs sont observés. Pour décider ce qu'est une "dimension raisonnable" dans une telle situation, Arlot (2007, p.94) conseille de choisir un saut correspondant à une dimension $D \ll \frac{\ln n}{n}$.

La procédure de sélection de modèles vise à approcher, en risque de Kullback-Leibler, les performances de l'oracle. Pour cet exemple entièrement simulé, toutes les distributions sont connues et il est donc possible d'évaluer le risque de chacun des estimateurs $\hat{g}_m^\star$ par des procédures de Monte-Carlo. La figure D.5 représente le risque minimal par dimension et le tableau D.3 synthétise les résultats pour l'ensemble des 200 simulations. Nous voyons que dans plus de 90% des cas, des partitions de tailles 3, 4 ou 5 sont sélectionnées, ce qui est satisfaisant d'après l'estimation du risque représentée sur la figure D.5.

D.1.3 Étude des performances de $\hat{N}$ et de $\hat{N}^{\text{HS}}$ par simulation

Nous proposons d'étudier les estimateurs $\hat{N}$ et de $\hat{N}^{\text{HS}}$ à l'aide des jeux de données simulés suivants. Nous simulons 200 fois un échantillon de longueur $N = 1000$ de couples (X_i, D_i) avec

$$X \sim \mathcal{P}\text{ar}(\alpha, x_0, x_{\max}) \text{ et } (D \,|\, X) \sim \mathcal{E}\left(h(X)\right).$$

Les paramètres de la distribution du couple sont les suivants :

$$\alpha_s = 0.75, \quad m_s = [x_0 = 10, 40, 300, x_{\max} = 5000], \quad a_s = 10^{-4} \times [4\,,\, 0.05\,,\, 2], \quad b_s = 10^{-4}.$$

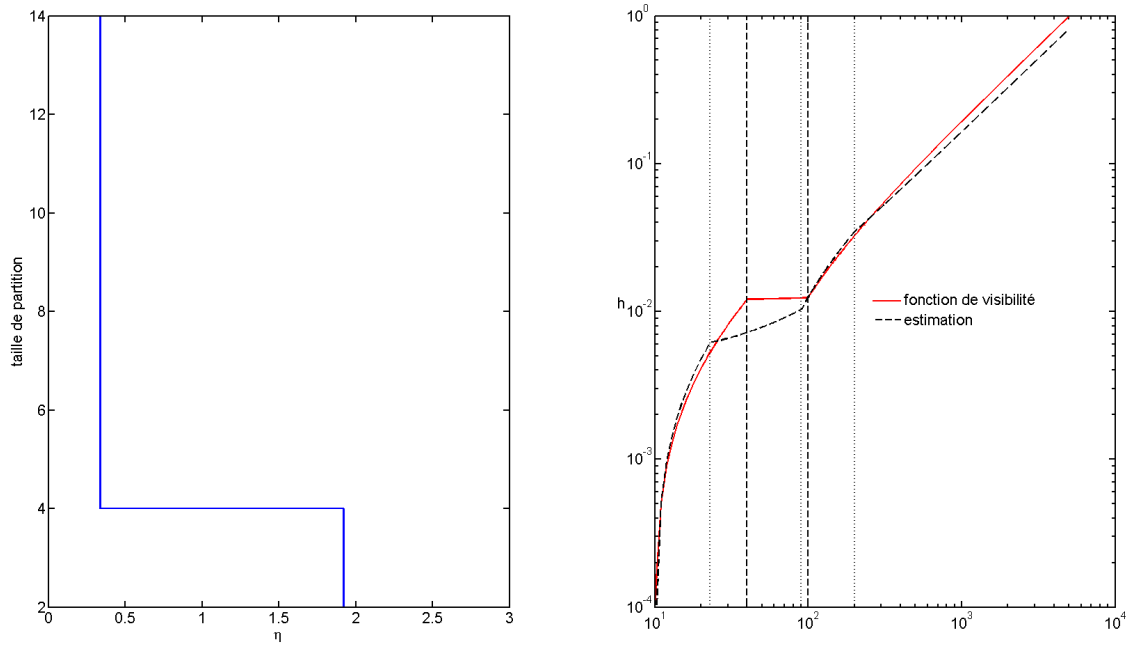


FIG. D.4: Illustration de la méthode de la pente pour l'un des échantillons simulés (gauche). La figure de droite permet de comparer la fonction de visibilité réelle avec celle qui est estimée dans le modèle sélectionné.

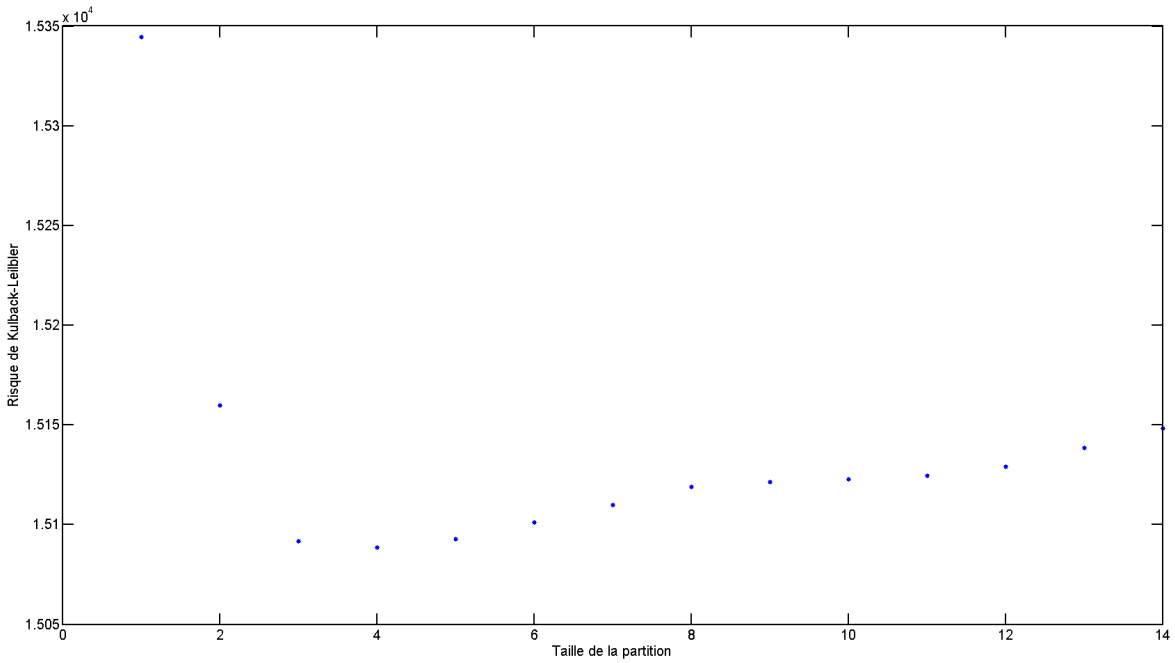


FIG. D.5: Risque minimal (sur l'ensemble des modèles de même dimension) en fonction de la dimension des modèles.

taille de la partition	Nombre de sélections	pourcentage de sélection
1	0	0 %
2	6	3 %
3	88	44 %
4	68	34 %
5	29	14.5 %
6	8	4 %
7	0	0 %
8	1	0.5 %

TAB. D.3: Modèles sélectionnés pour les 200 simulations.

α	$\mathbb{E}\left(\frac{\lvert\hat{N}-N\rvert}{N}\right)$	$\mathbb{E}\left(\frac{\lvert\hat{N}^{\mathrm{HS}}-N\rvert}{N}\right)$
0.75	0.11	0.09
0.65	0.22	0.22
0.85	0.34	0.35

TAB. D.4: Erreurs relatives moyennes du nombre estimé de champs de tailles supérieures à x_0.

Nous fixons la durée de l'exploitation à $t^\star = 50$ ans, et nous formons ensuite, pour chaque échantillon simulé, l'échantillon des couples $(X_i^\star, D_i^\star)$ des champs découverts avant $t^\star$. L'espérance de la longueur n de l'échantillon des gisements découverts est inférieure à 300 ; moins d'un tiers des gisements sont donc découverts pour la plupart des simulations. Ensuite, nous utilisons la méthode de la pente et nous obtenons ainsi l'estimateur pénalisé $\hat{g}_{\hat{m}}^\star$ et la fonction de visibilité $\hat{h}$ qui lui est associée. Ceci nous permet pour chaque simulation de calculer $\hat{N}$ et $\hat{N}^{\mathrm{HS}}$. Les résultats obtenus pour ces deux estimateurs sont donnés dans le tableau D.4. Pour les simulations de X, nous avons utilisé différentes valeurs pour α. Dans tous les cas, les deux estimateurs ont des performances comparables. Pour $\alpha = \alpha_s$, l'estimation de n donne de bons résultats. En revanche, si α est sensiblement différent de α_s, la qualité des estimations de N est dégradée. Ceci nous conforte dans le choix d'utiliser le méthode de Lepez pour estimer le paramètre α, plutôt que de l'incorporer dans les équations de vraisemblance de notre modèle d'exploration pétrolière.

D.2 Validation

Cette section s'attache à montrer que le modèle stratifié présenté au chapitre 4 correspond bien à la réalité du processus des découvertes tel qu'il peut être observé dans des bassins connus.

Zone pétrolifère	I_0	$t^\star$	$n_0(t^\star)$	$\hat{\mu}_0$	p-value
Sirte	[1 , 12]	42	104	2.36	0.90
mer du Nord	[1 , 20]	31	226	6.85	0.67
Nigeria off-shore	[1 , 20]	39	103	2.51	0.79

TAB. D.5: Estimations des intensités des processus de Poisson modélisant les temps de découvertes dans la classe des petits champs pour les trois bassins étudiés. La dernière colonne donne le résultat du test du χ^2 pour l'adéquation de la loi du nombre de découvertes annuelles à une loi de Poisson.

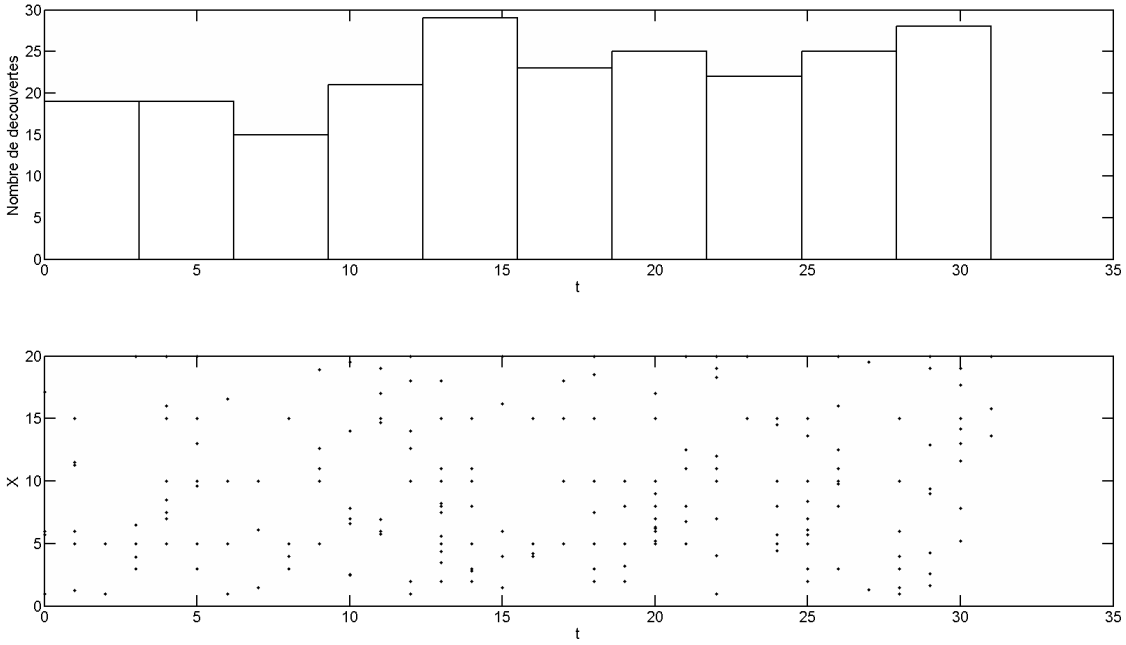


FIG. D.6: Le graphique supérieur donne le nombre de champs de taille inférieure à 20Mb découverts en mer du Nord au cours du temps. Le graphique inférieur représente le processus des découvertes dans la classe I_0.

D.2.1 Classe I_0 des petits champs

Nous vérifions que les champs de petites tailles sont découverts à taux constant au cours de l'exploration pétrolière, comme l'illustre le graphique supérieur de la figure D.6 pour la mer du Nord. Pour cela, nous comparons la distribution empirique du nombre annuel de champs découverts dans cette catégorie à celle d'une loi de Poisson de paramètre $\hat{\mu}_0$. Pour valider la modélisation, un test d'adéquation du χ^2 est effectué et les p-values ainsi que les intensités estimées $\hat{\mu}_0$ sont données dans le tableau D.5. Le seuil supérieur x_0 a été choisi de façon à ce que l'hypothèse testée soit largement acceptée tout en conservant une large proportion des champs du bassin dans I_0. Les résultats portés dans le tableau montrent que l'hypothèse de modélisation sur les petits champs est valide pour les trois bassins étudiés.

Bassin	I	$a/10^{-4}$	$b/10^{-4}$	P-value du K.S. test
mer du Nord	$[300, 4139]$	1.95	2.15	0.2
	$[100, 300]$	18.4	17.4	0.5
	$[20, 50]$	1.24	0.81	0.6
Sirte	$[300, 4482]$	1.42	0.88	0.10
	$[40, 200]$	18.4	17.4	0.40
Nigeria off-shore	$[400, 1250]$	3.32	15.8	0.10
	$[100, 400]$	10.1	9.70	0.15
	$[20, 100]$	1.46	1.00	0.20

TAB. D.6: Test de Kolmogorov-Smirnov pour la loi conditionnelle de $(D^\star \mid X)$.

D.2.2 Gisements de taille supérieure à x_0

La méthode de sélection de modèles que nous utilisons ne nécessite pas que la "véritable" densité de $(X^\star, D^\star)$ soit dans l'un des modèles S_m. Nous montrons tout de même que dans les bassins étudiés, pour une unique classe de taille I donnée, l'hypothèse

$$(D^\star \mid X^\star = x \in I) \sim \mathcal{E}(ax + b) \tag{D.1}$$

peut être validée. Les paramètres a et b sont estimés par maximum de vraisemblance en utilisant une méthode de descente de gradient. Le tableau D.6 fournit les résultats des tests de Kolmogorov-Smirnov (K.S.) et les estimations des fonctions h pour différentes classes de tailles dans chacun des bassins. Soit n_I est la longueur de l'échantillon de la classe I, et soit n_s la longueur de l'échantillon simulé pour effectuer le test. Dans tous les cas, nous avons $n_I\, n_s/(n_I + n_s) \gg 4$, ce qui est généralement préconisé pour que le test soit performant. Le tableau exhibe des intervalles pour lesquels il est possible d'accepter (D.1) avec un niveau de test toujours supérieur à 10%.

Néanmoins, plusieurs écarts entre les distributions théoriques et les observations sont relevés. Ces écarts permettent de mettre à jour les limitations suivantes de la modélisation que nous avons proposée pour la loi conditionnelle de $D^\star$.

- Toutes les classes de tailles ne peuvent pas toujours être validées.
- Sur l'ensemble de la classe $[x_0, x_{\max}]$, il est parfois nécessaire de ne conserver que les découvertes apparues après un certain laps de temps. Ceci est naturel puisqu'au début de l'exploration, l'activité de forage n'atteint pas immédiatement son "régime courant".
- Dans certains bassins, les champs de tailles modestes sont parfois découverts plus tard que ne le prévoit la distribution exponentielle proposée. La figure D.7 illustre ce phénomène pour la classe des champs de mer du Nord de taille dans l'intervalle $[20\,,\,50]$. Le graphique de gauche correspond aux observations complètes et le graphique de droite présente les données translatées de 5 ans, et pour celles-ci les quelques champs découverts dans les cinq premières années de l'exploration ont été retirés de l'échantillon. La loi estimée s'ajuste beaucoup mieux sur les observations dans le second cas. Heureusement, cette remarque ne concerne pas la population des plus gros champs dont la distribution

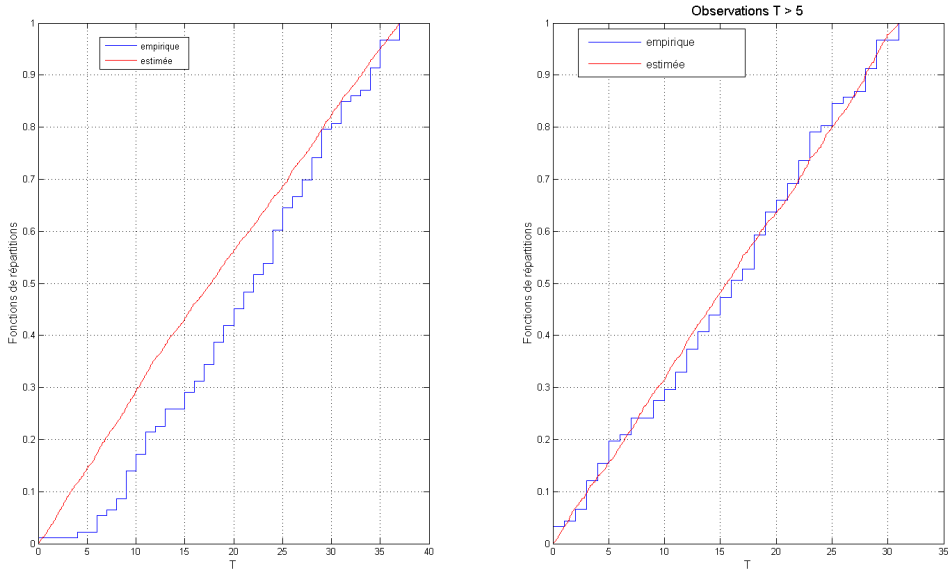


FIG. D.7: Fonctions de répartitions empiriques et estimées de la loi de $(D^\star \,|\, X^\star \in [20\,,\,50])$ sur les données de la mer du Nord.

conditionnelle des temps de découverte peut être plus facilement validée. Notre modélisation a donc tendance à découvrir des champs modestes un peu plus rapidement que dans la réalité. Pour deux raisons distinctes, cette observation est heureusement sans conséquences sur la modélisation de la production. Tout d'abord, seuls les champs de taille modeste sont concernés, et les réserves en jeu le sont donc elles aussi. En second lieu, ces champs de petite taille ne sont pas mis en production au début de l'exploration, et leur découverte n'aura donc que peu d'influence sur les niveaux de production modélisés.

Bibliographie

Abraham, C., Cornillon, P. A., Matzner-Løber, E. et Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scand. J. Statist.*, 30:581–595.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *In Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control*, AC-19:716–723. System identification and time-series analysis.

Aldous, D. J. (1985). Exchangeability and related topics. *In École d'été de probabilités de Saint-Flour, XIII—1983*, volume 1117 de *Lecture Notes in Math.*, pages 1–198. Springer, Berlin.

Arlot, S. (2007). *Rééchantillonnage et sélection de modèles.* Thèse de doctorat, Université Paris-Sud XI.

Attanasi, E. et Charpentier, R. (2002). Comparison of two probability distributions used to model sizes of undiscovered oil and gas accumulations : does the tail wag the assessment ? *Mathematical Geology*, vol34:767–777.

Babusiaux, D. (2005). Quelles productions et quels prix à l'avenir ? *Petroles & Techniques*, 456.

Babusiaux, D. et al. (2002). *Recherche et production du pétrole et du gaz : réserves, coûts, contrats.* Technip. Centre économie et gestion de l'Ecole du Pétrole et des Moteurs.

Banfield, J. D. et Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821.

Baudry, J.-P. (2007). Clustering through model selection criteria. Poster session at One Day Statistical Workshop in Lisieux. http ://www.math.u-psud.fr/ baudry.

Bentley, R. (2002). Global oil and gas depletion : an overview. *Energy Policy*, 30:189–205.

Bentley, R. (2006). Global oil and gas depletion - a letter to the energy modelling community. *IAEE Newsletter*, Second Quarter:6–14.

Berlinet, A., Biau, G. et Rouvière, L. (2008). Functional classification with wavelets. Rapport technique, Annales de l'ISUP.

BERTOIN, J. (1996). *Lévy processes.* Cambridge University Press.

BERTOIN, J. (2006). *Random fragmentation and coagulation processes.* Cambridge University Press.

BERTOIN, J. et LE GALL, J.-F. (2000). The Bolthausen-Sznitman coalescent and the genealogy of continuous-state branching processes. *Probab. Theory Related Fields*, 117:249–266.

BICKEL, P. J., NAIR, V. N. et WANG, P. C. C. (1992). Nonparametric inference under biased sampling from a finite population. *Ann. Statist.*, 20:853–878.

BIERNACKI, C., CELEUX, G. et GOVAERT, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:719–725.

BIERNACKI, C., CELEUX, G., GOVAERT, G. et LANGROGNET, F. (2006). Model-based cluster and discriminant analysis with the MIXMOD software. *Comput. Statist. Data Anal.*, 51: 587–600.

BIRGÉ, L. et MASSART, P. (2006). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138:33–73.

BIRGÉ, L. et MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3:203–268.

BOLTHAUSEN, E. et SZNITMAN, A. (1998). On Ruelle's probability cascades and abstract cavity method. *Comm. Math. Phys.*, 197:247–276.

BOUVEYRON, C., GIRARD, S. et SCHMID, C. (2007). High-dimensional data clustering. *Computational Statistics and Data Analysis*, 52:502–519.

BOVIER (2002). *Statistical mechanics of disordered spin.* Aarhus University.

BOVIER, A. et KURKOVA, I. (2003). Rigorous results on some simple spin glass models. *Markov Process. Relat. Fields*, 9:209–242.

BOVIER, A., KURKOVA, I. et LÖWE, M. (2002). Fluctuations of the free energy in the rem and the p-spin sk models. *Ann. Probab.*, 30:605–651.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. et STONE, C. J. (1984). *Classification and regression trees.* Wadsworth Advanced Books and Software.

CAMPBELL, C. (2002). Forcasting global oil supply. Rapport technique, Hubbert Center.

CAMPBELL, C. et LAHERRÈRE, J. (1998). The end of cheap oil. *Scientific American*, March: 80–85.

CAMPBELL, C., LAHERRÈRE, J. et A., P. (1998). The world's on conventional oil and gas. *Petroleum economist*, March.

CASTELLAN, G. (1999). Modified akaike's criterion for histogram density estimation. Rapport technique, Université de Paris Sud.

CASTELLAN, G. (2003). Density estimation via exponential model selection. *IEEE Trans. Inform. Theory*, 49:2052–2060.

CELEUX, G. et GOVAERT, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28:781–793.

CERA (2005). Technology and higter prices drive a supply buildup. Rapport technique, Cambridge Energy Research Associates.

CHAMBERS, J. M., MALLOWS, C. L. et STUCK, B. W. (1976). A method for simulating stable random variables. *J. Amer. Statist. Assoc.*, 71:340–344.

CLEVELAND, C. et KAUFMANN, R. (1991). Forecasting ultimate oil recovery and its rate of production : Incorporating economic forces into the model of m. king hubbert. *The Energy Journal*, 12(2):12–46.

COLEMAN, T. F. et LI, Y. (1996). An interior trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optim.*, 6:418–445.

DEFFEYES, K. (2001). *Hubbert's peak : the impending world oil shortage.* Princenton University Press.

DEMPSTER, A. P., LAIRD, N. M. et RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39:1–38.

DERRIDA, B. (2000). Random energy model : an exactly solvable model of disordered system. *Phys. Rev B. 24*, 24:2613–2626.

GARCÍA-ESCUDERO, L. A. et GORDALIZA, A. (2005). A proposal for robust curve clustering. *J. Classification*, 22:185–201.

GENOVESE, C. R. et WASSERMAN, L. (2000). Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.*, 28:1105–1127.

GOURDON, G. (1994). *Analyse.* elipses.

GUERRA, F. (1995). Fluctuations and thermodynamic variables in mean field spin glass models. Albeverio, S. (ed.) et al., Stochastic processes, physics and geometry II. Proceedings of the 3rd international conference held in Locarno, Switzerland, 24-29 June 1991. Singapore : World Scientific. 333-352 (1995).

HASTIE, T., TIBSHIRANI, R. et FRIEDMAN, J. (2001). *The elements of statistical learning.* Springer-Verlag.

HIRSCH (2007). Peaking of world oil production : recent forcasts. Rapport technique, National Energy Technology Laboratory - U.S. Department of Energy.

HIRSH, L., BESDEK, R. et WENDING, R. (2005). Peaking of world oil production : impacts, mitigation, & risk management.

HORVITZ, D. G. et J., T. D. (1952). A generalization of sampling without remplacement from finite universe. *Journal of the american statistical association*, 47:663–685.

HOUGHTON, J. (1988). Use of the truncated shifted pareto distribution in assessing size distribution of oil and gas fields. *Mathematical Geology*, vol20:907–937.

HUBBERT, M. (1956). Nuclear energy and fossil fuels. *Am. Petrol. Inst. Drilling and Production Practice*, pages 7–25.

HUBBERT, M. (1962). Energy resources. *National Research Sciences.*

JAMES, G. M. et SUGAR, C. A. (2003). Clustering for sparsely sampled functional data. *J. Amer. Statist. Assoc.*, 98:397–408.

KARLIN, S. (1967). Central limit theorems for certain infinite urn schemes. *J. Math. Mech.*, 17:373–401.

KAUFMAN, G. (1963). *Statistical decision and related techniques in oil and gas exploration.* N.J. :Prentice Hall, Englewood Cliffs.

KAUFMAN, G., BALCER, Y. et KRUYT, D. (1975). A probabilistic model of oil and gas discovery.

KAUFMANN, R. (1991). Oil production in the lower 48 states : Reconciling curve fitting and econometric models. *Resources and Energy*, 13:111–127.

KEMP, A. et KASIM, S. (2003). An econometric model of oil and gas exploration development and production in the uk contiental shelf : A systems approach. *The Energy Journal*, 24:113–136.

KERIBIN, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā Ser. A*, 62:49–66.

KINGMAN, J. F. C. (1975). Random discrete distribution. *J. Roy. Statist. Soc. Ser. B*, 37:1–22. With a discussion by S. J. Taylor, A. G. Hawkes, A. M. Walker, D. R. Cox, A. F. M. Smith, B. M. Hill, P. J. Burville, T. Leonard and a reply by the author.

KINGMAN, J. F. C. (1993). *Poisson processes.* The Clarendon Press Oxford University Press.

KONTOROVICH, A. E., DYOMIN, V. I. et R., L. V. (2001). Size distribution and dynamics of oil and gas discovories in petroleum basins. *The american association of petroleum geologists*, 85:1609–1622.

LAHERRÈRE, J. (1997). Multi-hubbert modeling.

LAHERRÈRE, J. (2003). Oil and natural gas resource assessment : Production growth cycle models. *Encyclopedia of Energy.*

LAW, M. H., JAIN, A. K. et FIGUEIREDO, M. A. T. (2004). Simultaneous Feature Selection and Clustering Using Mixture Models. *IEEE*, 26:1154–1166.

LEBARBIER, E. (2005). Detecting multiple change-points in the mean of gaussian process by model selection. *Signal Processing*, 85:717–736.

LEDOUX, M. et TALAGRAND, M. (1991). *Probability in Banach spaces*. Springer-Verlag.

LEPEZ, V. (2002). *Potentiel de réserves d'un bassin pétrolier : modélisation et estimation*. Thèse de doctorat, Université Paris Sud.

LYNCH, M. (2003). The new pessimism about petroleum resources : Debunking the hubbert model (and hubbert modelers). *Mineral and Energy*, 18.

MA, P., CASTILLO-DAVIS, C., ZHONG, W. et LIU, J. S. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Research*, 34:1261–1269.

MALLOWS, C. (1973). Some comments on c_p. *Technometrics*, 37:362–372.

MASSART, P. (2007). *Concentration inequalities and model selection*. Springer.

MAUGIS, C., CELEUX, G. et MARTIN-MAGNIETTE, M. (2007). Variable selection for clustering with Gaussian mixture models. Rapport technique RR6211, INRIA.

MAUGIS, C. et MICHEL, B. (2008a). A penalized criterion for Gaussian mixture model selection. Rapport technique 6549, INRIA.

MAUGIS, C. et MICHEL, B. (2008b). Slope heuristics for variable selection and clustering via Gaussian mixtures. Rapport technique 6550, INRIA.

MORONEY, J. et BERG, M. (1999). An integred model of oil production. *The Energy Journal*, 20:105–124.

PERCIVAL, D. B. et WALDEN, A. T. (2000). *Wavelet methods for time series analysis*. Cambridge University Press.

PERMAN, M., PITMAN, J. et YOR, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Related Fields*, 92:21–39.

PITMAN, J. (2006). *Combinatorial stochastic processes*. Springer-Verlag.

POLLARD, H. (1946). The representation of e^{-x^λ} as a Laplace integral. *Bull. Amer. Math. Soc.*, 52:908–910.

RAFTERY, A. E. et DEAN, N. (2006). Variable selection for model-based clustering. *J. Amer. Statist. Assoc.*, 101:168–178.

RUELLE, D. (1987). A mathematical reformulation of Derrida's REM and GREM. *Comm. Math. Phys.*, 108:225–239.

RYAN, M. (2003). Hubbert's peak : deja vu all over again. *IAEE Newsletter*, Second Quarter:9–12.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6:461–464.

SERRE, D. (2002). *Matrices*. Springer-Verlag.

SIMMONS, M. R. (2005). *Twiligt in the Desert : the coming Saudi oil shock and the world economy*. John Wiley and Sons.

TALAGRAND, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Inst. Hautes Études Sci. Publ. Math.*, 81:73–205.

TALAGRAND, M. (2003). *Spin glasses : a challenge for mathematicians*. Springer-Verlag.

USGS (2000). World petroleum assement 2000.

VERZELEN, N. (2007). Model selection for graphical models. In preparation.

VILLERS, F. (2007). *Tests et sélection de modèles pour l'analyse de données protéomiques et transcriptomiques*. Thèse de doctorat, Université Paris-Sud XI.

Modélisation de la production d'hydrocarbures dans un bassin pétrolier

Résumé : Cette thèse a pour objet la modélisation de la production pétrolière dans un bassin d'hydrocarbures. Le modèle proposé s'appuie sur une description probabiliste des réserves, de l'exploration des hydrocarbures et de la mise en production des gisements découverts. L'utilisation de la loi de Levy-Paréto pour décrire les tailles des gisements s'appuie d'une part sur une description probabiliste de la formation des réserves au cours de l'évolution du temps géologique, et d'autre part sur les propriétés d'invariance de la distribution de Poisson-Dirichlet pour des processus de coalescence et de fragmentation, dans le cadre du modèle de Bolthausen-Sznitman. Deux principaux problèmes statistiques, relevant tous les deux d'une problématique de choix de modèle en estimation de densité, sont identifiés. Le premier concerne l'estimation d'un modèle d'exploration pétrolière et le second est une étude de courbes de production qui repose sur une classification non supervisée et une sélection de variables pertinentes effectuées via la sélection d'un modèle de mélange Gaussien. Dans les deux cas, un critère de maximum de vraisemblance pénalisé est défini pour obtenir une inégalité de type oracle. Le modèle global de production pétrolière d'un bassin ainsi obtenu permet d'une part de préciser la forme des profils de production de bassin, et d'autre part de proposer des scénarios de prolongement de la production de bassins en cours d'exploitation.

OIL PRODUCTION MODELLING IN AN HYDROCARBON BASIN

Abstract : This thesis proposes a modelling of the oil production in a hydrocarbon basin. The model is built on a probabilistic description of reserves, of the exploration process and of the launching process of the discovered fields. The use of the Levy-Paréto distribution to model field sizes is justified first by a probabilistic modelling of the reserves creation during the evolution of the geologic time, and second by the invariance properties of the Poisson-Dirichlet distribution under coalescence and fragmentation operations, within the Bolthausen-Sznitman model framework. Two main statistical problems of model selection in the density estimation framework are identified. The first topic is about the estimation of the oil exploration model and the second is a production curve study which is carried out with a clustering and a variable selection obtained by the selection of a Gaussian mixture model. In both cases, a penalized maximum likelihood criterion is defined in order to make the selected estimator achieve an oracle inequality. The complete model for oil production in a basin allows to specify the shape of basin production profiles. It also allows to propose production scenarios in the future for producing basins.

AMS Classification : 62H30,62G07,60C05,60G09,82C31