uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Amay Sok Muy Cheam
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.Sc. (mathématiques)
GRADE / DEGREE

Département de mathématiques et statistique
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Modélisation de la courbe ROC à partir des distributions de Pearson

TITRE DE LA THÈSE / TITLE OF THESIS

André Dabrowski
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

Pierre-Jérôme Bergeron

Nicholas Birkett

Patrick Ferrell

Gary W. Slater
Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

Modélisation de la courbe ROC à partir des distributions de Pearson

Amay S.M. Cheam

Thèse soumise à la Faculté des études supérieures et postdoctorales
en vue de l'obtention de la Maîtrise ès science en biostatistiques [1]

Département de mathématiques et de statistiques
Faculté des sciences
Université d'Ottawa

1. le programme de maîtrise est un programme conjoint avec l'Université Carleton, administré par l'Institut d'études supérieures et de recherche en mathématiques et en statistiques d'Ottawa-Carleton

Your file Votre référence
ISBN: 978-0-494-79664-1
Our file Notre référence
ISBN: 978-0-494-79664-1

Abstract

The Receiver Operating Characteristic curve (ROC) is frequently used in medical studies to assess the efficiency of a diagnostic test. There exists two different ways to model the curve theoretically: direct and indirect. The direct method consists of constructing the curve based on the observed data. This method is less appealing in that the curve obtained is not easily interpretable. In the indirect method, a distribution is assumed for both the diseased and non-diseased patients. The normal distribution has been used to model both types of patients owing to the ease of its manipulation. In this thesis, we propose the use of the Pearsonian system of distributions in order to select the distribution for the diseased and non-diseased patients. An approach using a Monte Carlo simulation provides a confidence band for the derived ROC. The approach is evaluated using normal, gamma and beta distributed data. It is also tested on a real data set. It is seen that the Pearson based estimation of the ROC is at least as accurate as the normal theory based approach and often superior.

Key words : ROC curve, Pearson distribution, AUC, pAUC, trapezoidal rule, Mann-Whitney U-Stat, Monte-Carlo simulation.

Résumé

La courbe ROC, soit *Receiver Operating Characteristic*, est souvent utilisée comme un outil pour évaluer l'efficacité d'un diagnostic médical. Il existe deux approches de modélisations: directe et indirecte. La première consiste à modéliser directement la courbe ROC par une forme fonctionnelle quelconque. Cette approche est moins favorisée due à sa faiblesse intuitive. La seconde repose sur une hypothèse de la distribution des populations malades et non-malades. Notre étude propose la famille de distribution de Pearson étant donné sa flexibilité à caractériser des données jusqu'au quatrième moment. L'absence d'une écriture analytique nous amène à employer une technique de simulation par Monte-Carlo permettant, en même temps, de construire un intervalle de confiance autour de la courbe ROC. Cette approche est évaluée sur les données simulées à partir des lois normale, gamma et bêta. Elle est aussi testée sur les données réelles. Les résultats affichent une performance égale et souvent supérieure à l'approche binormale.

Mots clés : Courbe ROC, distribution de Pearson, AUC, pAUC, méthodes de trapèze et de Mann-Whitney, simulation de Monte-Carlo.

Remerciements

Tout d'abord, je tiens à remercier mon défunt superviseur, André Dabrowski, qui croyait en moi, et qui s'est acharné et a persisté pour mon admission au programme de Biostatistique. De même, je remercie mon superviseur, Mayer Alvo, de m'avoir prise sous son aile après cette perte affligeante. Grâce à son expertise et son attention envers mes intérêts, il a réussi à cibler un sujet devenu de jour en jour intéressant et captivant. Je le remercie aussi pour sa patience, ses encouragements et ses conseils judicieux. Je remercie également le professeur David Sankoff qui est toujours prêt à m'aider et m'encourager dans toutes les situations.

Je remercie aussi *toute* ma famille pour leur support et leurs précieux conseils tout au long de ma vie. Je remercie spécialement mon frère, Tim. pour avoir toujours été à mes côtés. Il n'y a pas de mots pour exprimer ma gratitude autre que *merci d'être mon frère.* Sans oublier mes trois petites et rayonnantes nièces qui embellissent ma vie.

Je remercie ma chère amie, Noémi Couture Guindon, de m'avoir non seulement aidée dans la correction de cette thèse, mais aussi pour sa précieuse amitié. Finalement, une salutation à mon amie, Marie-Pascal Berthelot, avec qui j'ai partagé des rires et des pleurs, sans oublier le support mutuel que nous nous sommes apporté durant mes années universitaires.

Dédicace

À ma précieuse famille.

Table des matières

Liste des figures

Liste des tableaux

Chapitre 1

Introduction

La courbe ROC, plus précisément *Receiver Operating Characteristic*, a fait ses premières apparitions dans les années 1950 dans le domaine de la théorie de détection des signaux [14]. Pendant la deuxième Guerre Mondiale, des opérateurs de radar tentent de discerner si le point clignotant sur leur image de radar représente un missile envoyé par l'ennemi ou simplement un bruit. Ainsi le dilemme se résume à trouver le meilleur réglage de radars : si le réglage était trop sensible même un oiseau pouvait déclencher l'alarme, par contre, si le réglage était moins sensible, mais plus spécifique, le radar manquerait le passage du missile [22]. Par la suite, deux chercheurs en psychologie, David M. Green et John A. Swets, critiquent la méthode employée par leurs collègues pour distinguer les stimuli des perceptions subjectives des patients, en d'autres mots pour différencier le monde physique des sensations. En 1966, Green et Swets ont eu la brillante idée d'introduire la théorie de détection en psychophysique afin d'identifier un seuil de séparation. En 1971, Lee B. Lusted a écrit un article dans le journal *Science* qui influencera à tout jamais le domaine médical. Il a postulé qu'avant de mesurer la valeur diagnostique d'un test, il faut avant tout évaluer la performance du test diagnostique dans un modèle de classification binaire [28]. En basant ses recherches sur des tests radiographiques, Lusted argumenta que les courbes ROC sont

d'une grande utilité pour étudier la performance et la précision d'un test diagnostique. Dès cet instant, la courbe ROC gagne plus de popularité dans plusieurs disciplines médicales comme un outil d'évaluation diagnostique. Par exemple, en radiologie, la courbe ROC permet de cerner la méthode d'une imagerie diagnostique. En fait, la plupart des travaux méthodologiques sur l'analyse du ROC a été effectué dans le cadre de la radiologie, car le diagnostic d'une imagerie varie d'un radiologiste à un autre [18].

Récemment, il y a eu un essor spectaculaire de l'utilisation de la courbe ROC comme outil : d'évaluation de l'efficacité d'un test diagnostique traitant des variables continues basées sur des observations indépendantes, de sélection d'un seuil décisionnel optimal et de comparaison entre plusieurs tests. Malheureusement, la courbe ROC empirique obtenue directement des données n'est pas toujours désirable pour la simple raison qu'elle respecte rarement les propriétés théoriques d'une courbe ROC. Plusieurs travaux de modélisation de la courbe ROC furent entamer afin de résoudre ce problème. Dans la littérature, la modélisation de la courbe ROC peut se diviser en deux approches : directe et indirecte.

La première approche, l'approche *directe*, peu populaire, ne dépend aucunement des hypothèses de distribution. Le concept est de construire explicitement une courbe ROC à partir des scores de la population malade et non-malade [31, 46]. Cependant, l'apparence de la courbe ROC empirique est en forme d'escalier rocailleux, *i.e.* non lisse. À cause de ceci, certaines propriétés théoriques de la courbe ROC furent violer. Une solution possible, pour contourner cette infraction, est d'estimer non-paramétriquement les fonctions de densité de chaque population par l'entremise de la méthode d'estimation du noyau [17, 31, 32, 41, 50]. Ainsi le problème se résume à choisir une largeur de bande asymptotiquement optimale. Plusieurs études cherchent à développer une solution à ce dilemme [31, 38, 46]. Lloyd [31] suggèra d'utiliser le *bootstrap* afin de minimiser des distorsions dans le lissage de la courbe ROC.

La seconde approche, dite *indirecte*, modélise les scores des deux populations en

supposant une distribution quelconque comme gaussienne, gamma ou bêta. À partir de cette distribution théorique des scores, une forme fonctionnelle de la courbe ROC sera dérivée implicitement. Dans le but d'arriver à construire une courbe ROC, des méthodes paramétriques et semi-paramétriques furent suggérer.

Une des méthodes paramétriques est d'assumer que la population malade et non-malade suivent une certaine famille de distribution comme la normale qui est un choix évident et simple, la logistique [36], la Lomax [5], la gamma [8] ou d'autres [51]. Dans le cas de la normale, Goddard et Hinberg [13] pointèrent qu'elle n'est pas toujours un choix adéquat dans certaines situations telle que le cancer de la prostate. L'auteur souligna explicitement qu'une application inconsidérée et insouciante de la méthode est déconseillée, car elle dépend sérieusement des conjectures des distributions. De plus, Zhou *et al.* [48] cita la nécessité de vérifier consciencieusement la cohérence des données avec les hypothèses. L'alternative à la méthode précédente serait de spécifier une forme fonctionnelle de la courbe ROC au lieu d'assumer une distribution. Egan proposa implicitement une forme fonctionnelle hyperbolique de la courbe ROC [9] en assumant que les deux populations suivent une distribution logistique de même variance. Quant à England, il proposa un modèle exponentiel à deux paramètres [10]. Les deux techniques paramétriques sont très analogues, car la distribution des résultats des tests détermine entièrement la forme de la courbe ROC. L'avantage primordial d'un tel système paramétrique est incontestablement la simplicité, le lissage de la courbe ROC et la particularité de jongler avec un petit nombre de paramètres.

La méthode semi-paramétrique est particulièrement attrayante au niveau de sa flexibilité due à la présence non-paramétrique de certaines composantes en plus des caractéristiques paramétriques. Le modèle binormal [14] est sans doute un choix flagrant. La procédure s'effectue en assumant la binormalité des résultats d'un test diagnostique après une transformation monotone croissante quelconque [19]. Ainsi le problème se réduit à estimer les paramètres, *i.e.* la pente et l'intercepte. Un éventail de solutions propose différentes techniques telles que la méthode des moindres carrés

généralisés [21]. du maximum de vraisemblance ou pseudo-vraissemblance [4, 49, 47] et d'autres. Par exemple, afin d'obtenir une courbe ROC binormale lisse, Metz *et al.* [35] développa un algorithme, LABROC4. L'algorithme va grouper les données continues en nombre fini de catégories ordonnées et utilise après l'algorithme du maximum de vraisemblance de Dorfman-Alf [7] pour les données ordinales. Une variation à cette méthode a été suggérée par Li *et al.* [30] où ils modélisent les résultats d'un test diagnostique pour la population non-malade et malade non-paramétriquement et paramétriquement, respectivement. Par contre, aucune relation fonctionnelle est supposée entre ces deux distributions. Quant à Qin et Zhang [40], au lieu de modéliser directement les distributions des résultats diagnostiques des deux populations sachant le statut véritable de la maladie, ils modélisent la probabilité du statut de la maladie sachant les résultats diagnostiques à l'aide du modèle de régression logistique. Évidemment comme tout problème d'estimation un effet de manque d'ajustement (*lack-of-fit*) peut se produire avec la méthode semi-paramétrique.

Nous proposons une approche indirecte, plus précisément semi-paramétrique, basée sur la distribution de Pearson qui va combiner tous les avantages, mais va éviter certains désavantages des approches précédentes. L'idée centrale est de remplacer la distribution gaussienne par celle de Pearson pour la construction des courbes ROC puisque cette dernière offre théoriquement plusieurs avantages intéressants. Néanmoins, les deux critères de base pour juger un modèle de la courbe ROC sont la flexibilité et l'ajustement de la courbe ROC issue du modèle par rapport à celle empirique. L'objectif principal de notre étude est d'examiner la performance du modèle ROC basée sur la distribution de Pearson relative au choix classique du modèle ROC basée sur la distribution gaussienne. Pour ce faire, notre étude portera sur deux types de données ou populations : simulées et réelles. Les populations simulées sont générées aléatoirement par des lois normale, gamma et bêta. Les données réelles sont celles sur l'étude de l'infertilité masculine provenant d'Aziz *et al.* [1].

Cette thèse est organisée comme suit. Dans le chapitre 2, les préliminaires sur

les notions de classification sont introduits avec un exemple concret sur le cancer du sein. Par la suite, dans le chapitre 3, nous nous orientons vers la théorie derrière la courbe ROC tels ses définitions et ses propriétés mathématiques fondamentales. Au chapitre 4, nous nous concentrons davantage sur les approches de la modélisation de la courbe ROC : binormale (BIN) et *Monte-Carlo Pearson* (MCP). Nous exposons étape par étape la méthodologie de la courbe MCP tout en introduisant la méthode de simulation par Monte-Carlo. Dans le chapitre 5, nous discutons des mesures de performance afin de comparer la performance de notre modèle MCP *versus* la BIN. Finalement dans le chapitre 6, nous abordons l'analyse des résultats provenant des simulations et des données médicales sur l'infertilié masculine.

Chapitre 2

Préliminaires

Le principal atout de la courbe ROC est sans doute son pouvoir d'afficher et de résumer la performance des règles de classifications par l'entremise d'une courbe. Avant de se lancer profondément dans la théorie de la courbe ROC, il serait sage de connaître ses origines, *i.e.* la classification. Pour une meilleure illustration, nous allons considérer des tests diagnostiques binaires et un exemple d'application médicale en mammographie.

2.1 Classification pour les résultats diagnostiques binaires

Un problème de classification se résume à classer correctement un patient à son groupe ou population. Ainsi plusieurs recherches ont été effectuées dans le but de développer des règles de classification. Dans cette section, nous allons seulement considérer des résultats dichotomes.

Définition 2.1.1 *Soit D une variable indicatrice désignant le statut véritable de la*

maladie

$$D = \begin{cases} 1, & \textit{si la maladie est présente;} \\ 0, & \textit{sinon.} \end{cases}$$

Définition 2.1.2 *Soit S le résultat d'un test diagnostique. Par convention, plus la valeur de S est grande plus elle indique la présence de la maladie.*

$$S = \begin{cases} 1, & \textit{si positif;} \\ 0, & \textit{sinon.} \end{cases}$$

TABLE 2.1 La classification des résultats d'un test diagnostique par le statut de la maladie

	Statut véritable de la maladie	
Diagnostic d'un test	D=0	D=1
S=0	Vrai négatif	Faux négatif
S=1	Faux positif	Vrai positif

Les résultats d'un test peuvent être classifiés comme vrai positif, vrai négatif, faux positif et faux négatif comme illustré dans la table 2.1. On définit :

i. un vrai positif, *TPR*, la probabilité qu'un sujet malade obtienne un résultat d'un test positif : $P(S = 1 \mid D = 1)$,

ii. un faux positif, *FPR*, la probabilité qu'un sujet non-malade obtienne un résultat d'un test positif : $P(S = 1 \mid D = 0)$,

iii. un vrai négatif, *TNR*, la probabilité qu'un sujet non-malade obtienne un résultat d'un test négatif : $P(S = 0 \mid D = 0)$,

iv. un faux négatif, *FNR*, la probabilité qu'un sujet malade obtienne un résultat d'un test négatif : $P(S = 0 \mid D = 1)$.

Remarque 2.1.3 Ces termes peuvent varier d'un domaine à un autre. Dans un contexte d'une hypothèse statistique, supposons que

$$H_0 : D = 0 \quad \text{vs} \quad H_1 : D = 1$$

alors FPR est appelé niveau de significance, α, et TPR est la puissance (*power*), $1 - \beta$. La table 2.2 nous suggère quelques termes selon leur domaine.

TABLE 2.2 Les terminologies selon les domaines de recherches

Domaine	FPR	TPR
Biomédical	1-Spécificité	Sensibilité
Ingénerie	Taux de fausse alarme	Taux de réussite
Statistique	Niveau de signification (α)	Puissance ($1 - \beta$)

2.2 Exemple

Afin d'illustrer les calculs, considérons un exemple de données ordinaux d'un diagnostic de mammographie de 60 patients qui se présentent pour un dépistage du cancer du sein. Les données de la table 2.3 proviennent de sept études rétrospectives enquêtant sur la précision du dépistage de la mammographie [39, 48]. L'échantillon est composé de 30 patients avec une pathologie cancéreuse et 30 patients ayant une mammographie normale durant les deux dernières années. Un diagnostic d'une mammographie est considérée positive si le radiologiste requiert des examens diagnostiques supplémentaires.

Supposons que le radiologiste utilise une échelle catégorique afin de classer ses résultats diagnostiques comme le présente la table 2.3. Afin de calculer la sensibilité et la spécificité, il décide de prendre le seuil décisionnel à bénin. Ainsi tous les patients dans les catégories probablement bénin, suspect et malin sont considérés positifs comme le présente la table 2.6. Des 30 patients reconnus d'avoir le cancer du sein, 29 obtiennent un résultat positif et furent contactés pour d'autres examens additionnels. Ainsi il y a 29 vrais positifs et un faux négatif. La sensibilité est de $Se = \frac{29}{30} = 0.967$. Des 30 patients ne présentant aucun indice cancéreux du sein, 11

TABLE 2.3 Résultats mammographiques du cancer du sein avec une échelle catégorique

	Statut véritable de la maladie	
Diagnostic d'un test	D=0	D=1
Normal	9	1
Bénin	2	0
Probablement bénin	11	6
Suspect	8	11
Malin	0	12

obtiennent un résultat négatif. Ainsi la spécificité est de $Sp = \frac{11}{30} = 0.367$. En variant le seuil, on obtient différentes valeurs pour notre α et $1 - \beta$, voir le tableau 2.8. Avec ces couples, on construit une courbe ROC empirique, voir figure 2.1.

TABLE 2.4 Résultats mammographiques du cancer du sein dont le seuil est à suspect

	Statut véritable de la maladie	
Diagnostic d'un test	D=0	D=1
S=0	30	18
S=1	0	12
Total	30	30

TABLE 2.5 Résultats mammographiques du cancer du sein dont le seuil est à probablement bénin

	Statut véritable de la maladie	
Diagnostic d'un test	D=0	D=1
S=0	22	7
S=1	8	23
Total	30	30

TABLE 2.6 Résultats mammographiques du cancer du sein dont le seuil est à bénin

	Statut véritable de la maladie	
Diagnostic d'un test	D=0	D=1
S=0	11	1
S=1	19	29
Total	30	30

TABLE 2.7 Résultats mammographiques du cancer du sein dont le seuil est à normal

	Statut véritable de la maladie	
Diagnostic d'un test	D=0	D=1
S=0	9	1
S=1	21	29
Total	30	30

TABLE 2.8 Valeurs de spécificité$(1-\alpha)$ et sensibilité$(1-\beta)$ aux différents seuils pour les résultats mammographiques du cancer du sein

Seuil	**Sp** $(1-\alpha)$	**Se** $(1-\beta)$
Suspect	1	12/30
Probablement bénin	22/30	23/30
Bénin	11/30	29/30
Normal	9/30	29/30

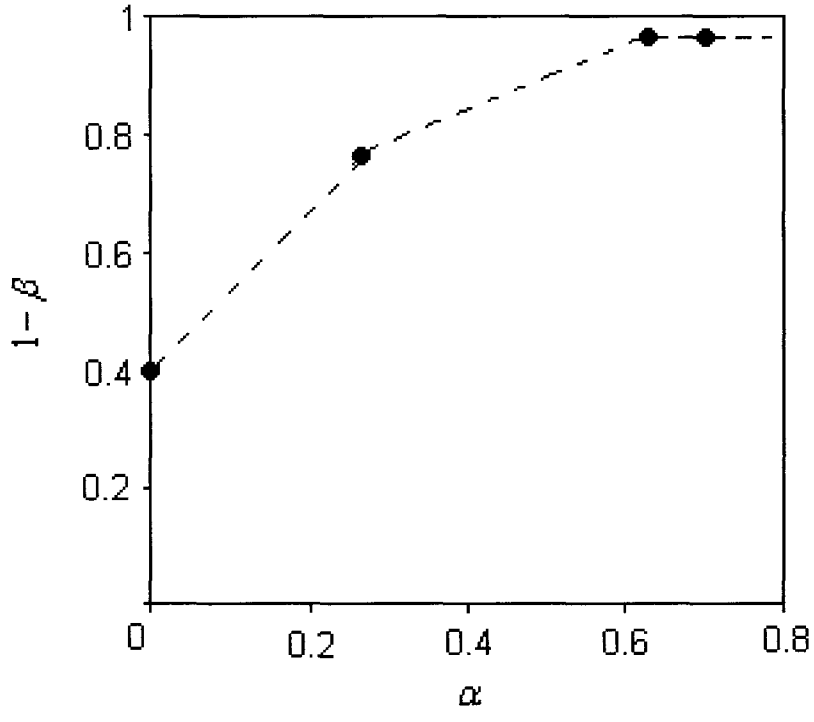

FIGURE 2.1 La courbe ROC empirique pour les résultats mammographiques du cancer du sein

Chapitre 3

Théorie de la courbe ROC

Dans le chapitre précédent, une introduction succincte du domaine de classification fut abordée. Brièvement, la courbe ROC est un sommaire sophistiqué pour le rassemblement d'informations sur la performance d'une classification autour d'un seuil décisionnel variant. Qu'en est-il du fondement derrière cet outil éminent ? Désormais, nous allons mettre l'accent sur les concepts théoriques qui surlignent la courbe ROC afin de mieux comprendre son fonctionnement et son application. Les points saillants sont : la formulation mathématique et les propriétés de la courbe ROC empirique.

3.1 Définition de la courbe ROC

L'origine de son nom, *Receiver Operating Characteristic*, dérive de la notion que connaissant la courbe, le receveur de l'information peut opérer à n'importe quel point sur la courbe en employant un seuil décisionnel approprié. Avant de définir la courbe ROC, reprenons quelques termes de la section 2.1, mais dans un cadre général.

3.1.1 Terminologie

Les résultats d'un test diagnostique ne sont pas toujours de nature dichotomes. Par exemple, les biomarqueurs pour le cancer comme la concentration sérique se présente sous forme continue.

Notation 3.1.1 Soient D_0 la population des sujets non-malades et D_1 des patients présentant une maladie.

Notation 3.1.2 Soient X un vecteur de variables descriptives continues et $S(X)$ une fonction de score représentant les résultats d'un test diagnostique de chaque patient.

Remarque 3.1.3 Dans un contexte médical, le vecteur X peut représenter un item d'un questionnaire portant sur le mode de vie, la nutrition ou encore l'historique médical du patient.

Définition 3.1.4 *Une classification est considérée idéale ou parfaite, si*

$$D_0 = \{S(X) \leq t : t \in \mathbb{R}\}$$

et

$$D_1 = \{S(X) > t : t \in \mathbb{R}\}.$$

Par convention, plus la maladie est présente, plus la valeur de $S(X)$ est élevée.

En relation aux quatre probabilités de la section 2.1, on peut les redéfinir de la manière suivante :

i. TPR ou $Se = P(S(X) > t \mid D_1)$,

ii. $FPR = P(S(X) > t \mid D_0)$,

iii. TNR ou $Sp = P(S(X) \leq t \mid D_0)$,

iv. $FNR = P(S(X) \leq t \mid D_1)$.

En variant t et en calculant les quatre probabilités ci-dessus, nous obtiendrons des informations pertinentes permettant d'évaluer la performance d'un classement. Puisque $TPR + FNR = 1$ et $FPR + TNR = 1$, nous n'avons pas besoin d'autant d'informations. Soulignons qu'une performance adéquate, dans la majorité des applications cliniques, requiert un taux élevé de *vrais* et un taux bas de *faux*.

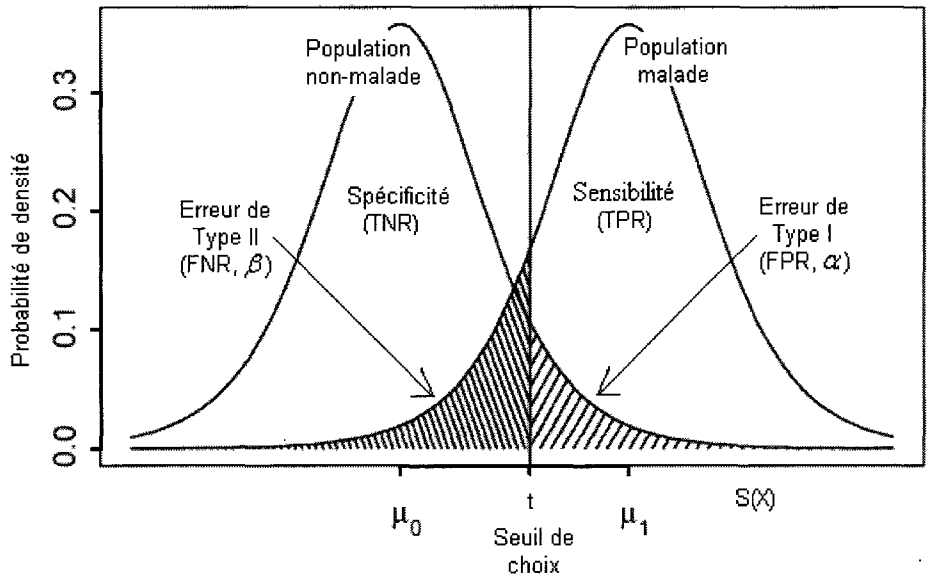

FIGURE 3.1 Distribution d'un test diagnostique $S(X)$ pour D_0 et D_1 avec un seuil décisionnel $t \in \mathbb{R}$

3.1.2 Définition générale

Le rôle de la courbe ROC est de procurer les mêmes informations que la table de classification, mais d'une façon plus élégante et élaborée. La courbe ROC est un graphe du couple (FPR, TPR). La courbe se situe dans un carré unitaire définit par ces points : $(0,0)$, $(0,1)$, $(0,1)$ et $(1,1)$. Chaque point du graphe est généré par la variation du seuil décisionnel. Ici, *1-spécificité* ou FPR est la valeur sur l'axe horizontal, l'abscisse, et la sensibilité ou TPR est sur l'axe vertical, l'ordonnée. La

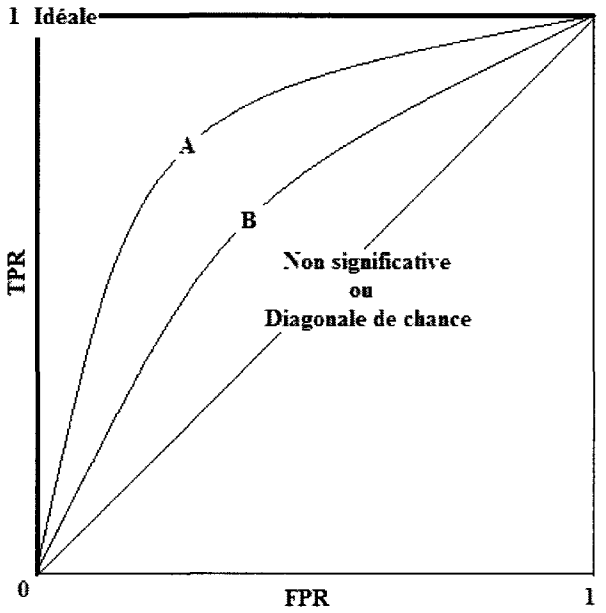

FIGURE 3.2 Les courbes ROC de deux tests A et B, avec la diagonale de chance, où le test A est plus performant que B

figure 3.2 illustre trois différentes courbes.

Définition 3.1.5 *Une courbe ROC est l'ensemble de toutes les pairs envisageables de* (FPR, TPR) *à différent seuil* t*, i.e.*

$$ROC = \{(P(S > t \mid D_0), P(S > t \mid D_1)), t \in \mathbb{R}\} \tag{3.1.1}$$

$$= \{(FPR(t), TPR(t)), t \in \mathbb{R}\} \tag{3.1.2}$$

Pour une meilleure idée visuelle de la construction de la courbe ROC, voir l'exemple 2.2 du chapitre 2 et la figure 3.3.

Définition 3.1.6 *Soient* S_{D_0} *et* S_{D_1} *les résultats diagnostiques ou les scores des patients dans la population non-malade et malade, respectivement. Supposons que les densités de probabilités de* S_{D_0}*, noté* $f_{S_0}(x)$*, et de* S_{D_1}*, noté* $f_{S_1}(x)$*, sont connus.*

Alors on les définit comme suit

$$S_{D_0} \sim f_{S_0}(x) \quad et \quad S_{D_0} \sim f_{S_1}(x).$$

Définition 3.1.7 *Soient* $S_{D_0} \sim f_{S_0}(x)$ *et* $S_{D_0} \sim f_{S_1}(x)$, *alors pour* $t \in \mathbb{R}$ *donné*

$$FPR(t) = \int f_{S_0}(x) I\left(s(x) - t\right) dx = P\left(S > t \mid D_0\right) \tag{3.1.3}$$

$$TPR(t) = \int f_{S_1}(x) I\left(s(x) - t\right) dx = P\left(S > t \mid D_1\right) \tag{3.1.4}$$

où

$$I(u) = \begin{cases} 1, & si\ u > 0, \\ 0, & si\ u \leq 0. \end{cases}$$

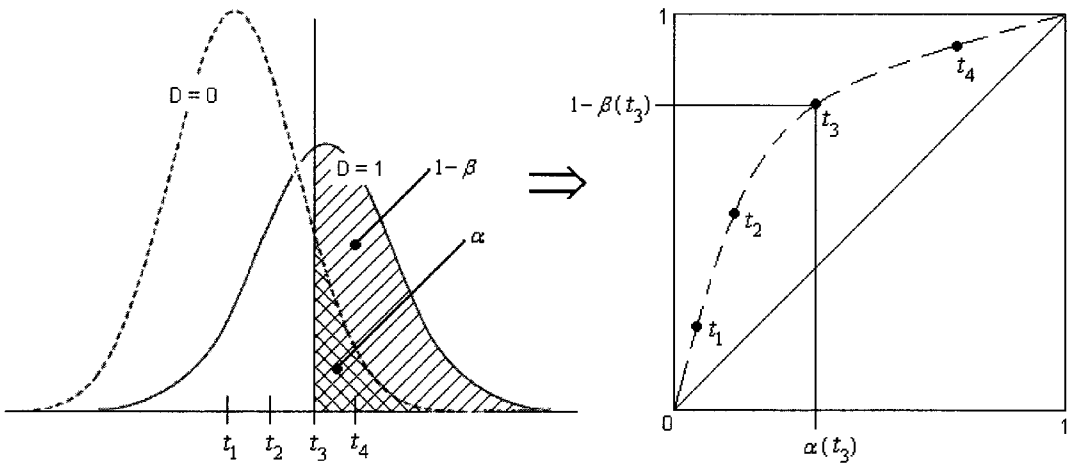

FIGURE 3.3 Construction de la courbe ROC à partir des population D_0 et D_1 à différent seuil, $t \in \mathbb{R}$

Clairement une classification est jugée selon l'écart entre les deux distributions, $f_{S_0}(x)$ et $f_{S_1}(x)$. Généralement, plus la dispersion est flagrante, moins les courbes se chevauchent entre elles. Alors la classification est moins susceptible de commettre des erreurs de classement. Ainsi le taux de réussite est plus élevé. À l'opposé, un

recouvrement de deux distributions indique une attribution faussement effectuée, voir la figure 3.4 provennant de l'article de Brumback *et al.* [3].

Considérons les cas extrêmes. Une classification est dite triomphante lorsque les deux densités de probabilités sont entièrement séparées, *i.e.* $f_{S_0}(x) \neq f_{S_1}(x)$. Le scénario plausible est qu'il y ait au moins une valeur t pour laquelle un sujet a été bien classé. Ainsi pour un tel t, on a $FPR = 0$ et $TPR = 1$. Puisque la courbe ROC accentue seulement les probabilités que $s > t$ des deux populations, alors pour toutes valeurs t

i. petites, nous obtiendrons $TPR = 1$ quand $0 \leq FPR \leq 1$

ii. larges, nous obtiendrons $FPR = 1$ quand $0 \leq TPR \leq 1$.

Inversement, une classification est dite *non informative* lorsque les deux densités de probabilités sont exactement réciproques, *i.e.* $f_{S_0}(x) = f_{S_1}(x) = f(x)$. Dans ce cas, la probabilité cumulative d'un individu de la population D_1 est la même qu'un sujet provenant de D_0 ou D_1. Ainsi la valeur de cette probabilité dépend uniquement du seuil t. Au fur et à mesure que t varie, TPR est toujours égale à FPR et la courbe ROC se résume par une ligne droite reliant les points $(0,0)$ et $(1,1)$, $ROC(t) = t$. Cette ligne est souvent appelée la *diagonale de chance* qui représente essentiellement une attribution aléatoire des sujets à une des deux populations.

En pratique, la courbe ROC est une courbe continue se situant entre les deux extrêmes, plus formellement dans le triangle supérieur du graphe, sans perte de généralité. Une performance d'une classification estimée satisfaisante possède une courbe se rapprochant le plus du coin supérieur gauche. Constatons que si une courbe ROC se localise dans le triangle inférieur, alors la fonction de score est faussement orientée et une inversion est nécessaire, *i.e.* un individu se classe dans D_1 si $s < t$ au lieu de $s > t$. La figure 3.2 affiche trois courbes incluant la ligne de chance. Le test A est sans équivoque plus performant, car pour un point FPR donné, son TPR est le plus élevé que le test B.

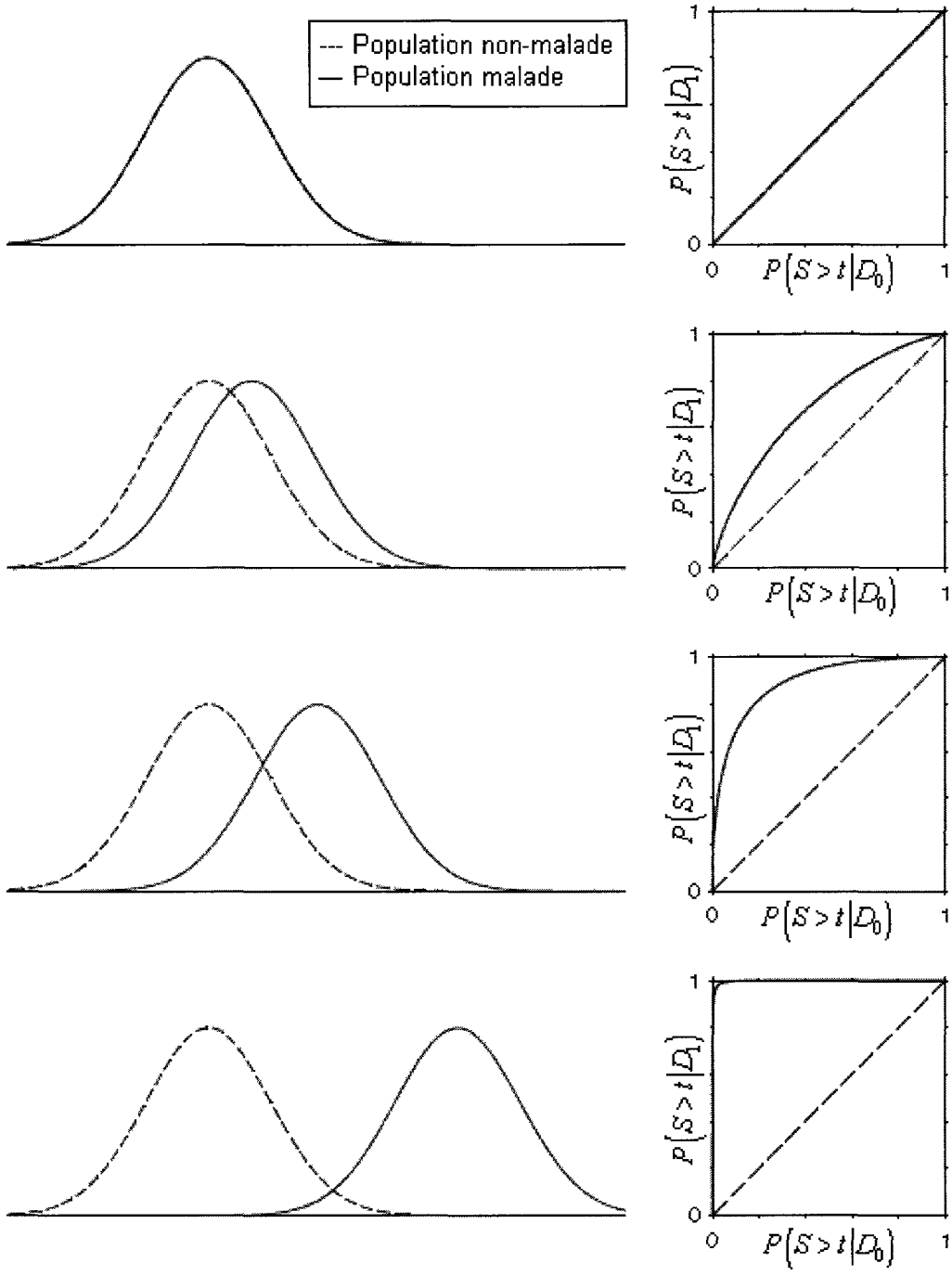

FIGURE 3.4 Relation entre différentes distributions des populations D_0 et D_1, et leurs courbes ROC correspondantes [3]

3.2 Propriétés de la courbe ROC

Maintenant que la définition de la courbe ROC fut énoncer, nous allons explorer les propriétés qui entourent la courbe ROC.

Propriété 3.2.1 La courbe ROC est une fonction croissante monotone dans le quadrant positif se situant entre $(0,0)$ et $(1,1)$.

Démonstration: Observons que lorsque le seuil t augmente, les valeurs de $FPR(t)$ et $TPR(t)$ diminuent. Supposons que $t = \infty$, on a

$$\lim_{t \to +\infty} FPR(t) = 0,$$

$$\lim_{t \to +\infty} TPR(t) = 0.$$

Inversement, posons $t = -\infty$, alors

$$\lim_{t \to -\infty} FPR(t) = 1,$$

$$\lim_{t \to -\infty} TPR(t) = 1.$$

∎

Propriété 3.2.2 La courbe ROC est invariante si les résultats diagnostiques subissent une transformation strictement croissante.

Démonstration: Supposons que $U = \varphi(S)$ est une transformation strictement croissante, *i.e.*

$$S_2 > S_1 \quad \Leftrightarrow \quad U_2 = \varphi(S_2) > U_1 = \varphi(S_1).$$

De plus, l'inverse de U existe, soit

$$S = \varphi^{-1}(U) = \varphi^{-1}(\varphi(S)).$$

Considérons un point de la courbe ROC pour S au seuil t, soit $v = \varphi(t)$. Alors, il succède que

$$P(U > v \mid D_1) = P(\varphi(S) > \varphi(t) \mid D_1) = P(S > t \mid D_1)$$

et

$$P(U > v \mid D_0) = P(\varphi(S) > \varphi(t) \mid D_0) = P(S > t \mid D_0)$$

afin qu'un même point existe sur la courbe ROC pour U. En appliquant l'argument inverse à chaque point de la courbe ROC pour U, ceci démontre que les deux courbes sont identiques. ∎

Propriété 3.2.3 La pente de la courbe ROC à un point au seuil $t \in \mathbb{R}$ est bien définie et donnée par

$$\frac{dTPR}{dFPR} = \frac{P(t \mid D_1)}{P(t \mid D_0)}. \tag{3.2.1}$$

Démonstration: Rappelons par la définition 3.1.5 que

$$\begin{aligned} TPR(t) &= P(S > t \mid D_1) \\ &= 1 - \int_{-\infty}^{t} P(s \mid D_1) ds \end{aligned}$$

alors,

$$\frac{dTPR}{dt} = -P(t \mid D_1).$$

Ainsi

$$\frac{dTPR}{dFPR} = \frac{dTPR}{dt}\frac{dt}{dFPR} = -P(t \mid D_1)\frac{dt}{dFPR}. \tag{3.2.2}$$

Encore par la définition 3.1.5, on a

$$\begin{aligned} FPR(t) &= P(S > t \mid D_1) \\ &= 1 - \int_{-\infty}^{t} P(s \mid D_0) ds \end{aligned}$$

et donc

$$\frac{dFPR}{dt} = -P(t \mid D_0). \tag{3.2.3}$$

En remplaçant ceci dans (3.2.2)

$$\frac{dt}{dFPR} = 1/\frac{dFPR}{dt}$$

l'énoncé est prouvé. ∎

Remarque 3.2.4 La propriété 3.2.3 montre que la pente de la courbe à un point au seuil $t \in \mathbb{R}$ est égale au rapport de vraisemblance

$$\lambda(t) = \frac{P(t \mid D_1)}{P(t \mid D_0)}.$$

Il y a une relation intéressante entre ceci et la théorie des tests d'hypothèse, voir annexe A.

3.3 Définition pour les résultats diagnostiques continus

Antérieurement, nous nous sommes concentrés uniquement sur le cas général d'une courbe ROC sans se préoccuper du type d'échelle de mesure des résultats diagnostiques. À présent, nous allons cibler uniquement les scores continus et développer une forme fonctionnelle de la courbe ROC.

Théorème 3.3.1 *Supposons que les densités de probabilités, f_{S_0} et f_{S_1}, sont continues et connues, alors la forme fonctionnelle de la courbe ROC peut se formuler comme suit*

$$TPR = 1 - F_{S_1}[F_{S_0}^{-1}(1 - FPR)] \quad 0 \leq FPR \leq 1. \tag{3.3.1}$$

où F_{S_0} et F_{S_1} sont les fonctions cumulatives des scores de la population D_0 et D_1, respectivement.

Démonstration: Par la définition 3.1.5, pour un $t \in \mathbb{R}$ donné, on a

$$FPR(t) = 1 - F_{S_0}(t)$$

$$\Leftrightarrow \quad F_{S_0}(t) = 1 - FPR(t)$$

$$\Leftrightarrow \quad t = F_{S_0}^{-1}(1 - FPR)$$

Donc,

$$\begin{aligned} TPR &= P(S \geq t \mid D_1) \\ &= 1 - P(S \leq t \mid D_1) \\ &= 1 - F_{S_1}(t) \qquad \text{(par la définition d'une fonction cumulative)} \\ &= 1 - F_{S_1}(F_{S_0}^{-1}(1 - FPR)). \end{aligned}$$

où $0 \leq FPR \leq 1$. ∎

Chapitre 4

Modélisation de la courbe ROC

Maintenant que les concepts théoriques furent abordés et appréhendés, nous allons s'attaquer au coeur du problème, *i.e.* la modélisation de la courbe ROC tout en respectant ces propriétés théoriques. Dans ce chapitre, nous présentons deux modèles de la courbe ROC en posant une hypothèse sur la nature distributionnelle des scores des deux populations, soit S_{D_0} et S_{D_1}. Nous commençons par présenter le modèle binormal. Par la suite, nous proposons une approche alternative, soit la courbe ROC par les distributions de Pearson.

4.1 Courbe ROC par le modèle binormal

Le modèle binormal est sans équivoque la méthode la plus explorée. La forme binormale procure une approximation satisfaisante de la courbe ROC dans diverses situations. Tout comme la distribution normale est un modèle classique pour les fonctions de répartition, la courbe binormale est un modèle standard pour les courbes ROC. Tel que mentionné dans l'introduction, il existe une vaste littérature autour de la binormale que ce soit paramétriquement ou semi-paramétriquement [4, 19, 21, 47]. Son atout est sans hésitation sa simplicité et son aisance de jongler avec un petit

nombre de paramètres.

Proposition 4.1.1 *Supposons que* $S_{D_0} \sim N(\mu_0, \sigma_0^2)$ *et* $S_{D_1} \sim N(\mu_1, \sigma_1^2)$. *La forme fonctionnelle d'une courbe ROC binormale pour les variables continues est définie comme*

$$TPR = \Phi(a + b\Phi^{-1}(FPR)) \tag{4.1.1}$$

où

$$a = \frac{\mu_1 - \mu_0}{\sigma_1}, \quad b = \frac{\sigma_0}{\sigma_1}. \tag{4.1.2}$$

et Φ *désigne la distribution cumulative normale standardisée.*

Démonstration: Selon la définition 3.1.4, on assume que $\mu_1 > \mu_0$. De plus,

$$\frac{S - \mu_0}{\sigma_0} \sim N(0,1) \quad \text{et} \quad \frac{S - \mu_1}{\sigma_1} \sim N(0,1).$$

Par la définition 3.1.5, on a

$$\begin{aligned} FPR(t) &= P(S > t \mid D_0) \\ &= P\left(Z > \frac{t - \mu_0}{\sigma_0}\right) \\ &= P\left(Z \leq \frac{\mu_0 - t}{\sigma_0}\right) \qquad \text{(par symétrie de la normale)} \\ &= \Phi\left(\frac{\mu_0 - t}{\sigma_0}\right) \end{aligned}$$

où Z et $\Phi(\cdot)$ représentent la variable normale standardisée et la distribution normale cumulative, respectivement. Donc, si z_{FPR} est la valeur de Z, alors

$$z_{FPR} = \Phi^{-1}[FPR(t)] = \frac{\mu_0 - t}{\sigma_0}$$

impliquant

$$t = \mu_0 - z_{FPR}\sigma_0. \tag{4.1.3}$$

Par conséquent, la courbe ROC à ce point FPR est

$$TPR = P(S > t \mid D_1)$$

$$= P\left(Z > \frac{t-\mu_1}{\sigma_1}\right)$$
$$= \Phi\left(\frac{\mu_1 - t}{\sigma_1}\right),$$

et en substituant la valeur t obtenue par (4.1.3), on a

$$TPR = \Phi\left(\frac{\mu_1 - \mu_0 + z_{FPR}\sigma_0}{\sigma_1}\right).$$

Donc, la courbe ROC est de la forme

$$TPR = \Phi(a + bz_{FPR}) \Longleftrightarrow \Phi^{-1}(TPR) = a + b\Phi^{-1}(FPR)$$

où

$$a = \frac{\mu_1 - \mu_0}{\sigma_1} \quad \text{et} \quad b = \frac{\sigma_0}{\sigma_1}.$$

Par l'hypothèse précédente, on a que $a > 0$ tandis que b est non-négative par définition. ∎

Remarque 4.1.2 L'intercepte et la pente de la courbe ROC binormale sont symbolisées par la valeur de a et b, respectivement.

L'avantage payant de ce modèle est son aisance et sa simplicité dans la dérivée et la forme analytique d'AUC, présentée prochainement dans la section 5.1.3.

Remarque 4.1.3 La propriété 3.2.2 cite que la courbe ROC est invariante à une transformation monotone croissante. Supposons que $S_{D_0} \sim N(\mu_0, \sigma_0^2)$ et $S_{D_1} \sim N(\mu_1, \sigma_1^2)$. Soient les transformations,

$$U_0 = \phi(S_{D_0}), \qquad U_1 = \phi(S_{D_1}),$$

où ϕ est une fonction monotone strictement croissante. Alors la courbe ROC de U_0 et U_1 est de la forme (4.1.1), *i.e.* une courbe ROC binormale. Inversement, supposons que

la courbe ROC est binormale, alors pour une transformation donnée, ϕ, strictement croissante, on a que $\phi(S_{D_0})$ et $\phi(S_{D_1})$ suivent une distribution gaussienne. Cette hypothèse est considérée légèrement faible, car ce modèle peut s'appliquer à certaines données non-gaussiannes [45, 19]. La courbe ROC binormal est par conséquent simple et très utilisée dans l'analyse ROC.

4.2 Courbe ROC par les distributions de Pearson

L'objectif ultime de la modélisation de la courbe ROC est de répliquer la courbe ROC empirique. Pourquoi cherchons-nous à reproduire cette courbe ROC empirique au lieu de travailler directement avec elle ? Pour la simple raison qu'elle ne respecte pas nécessairement les propriétés asymptotiques et la complexité des calculs.

4.2.1 Approche proposée

Dans l'introduction, nous avons introduit deux approches possibles pour modéliser la courbe ROC. Le schéma 4.1 synthétise ces deux approches. Due à la faiblesse intuitive et l'incertitude autour de la méthode de lissage utilisée, l'approche directe est moins attrayante. Par ce fait, notre approche proposée s'inscrit dans l'approche indirecte. Dans la section précédente, nous avons detaillé et exploré le modèle binormal. Un de ses désavantages est la supposition que les scores suivent une distribution gaussienne. Cependant, l'hypothèse de normalité des données est difficilement observée pour des données réelles.

Comme méthode alternative, nous proposons de construire la courbe ROC par une distribution Pearson, élaborée dans la sous-section 4.2.2. Contrairement à la gaussienne, la distribution de Pearson offre une plus grande flexibilité, car elle permet de capturer non seulement les déformations de la moyenne et de la variance, mais également celles d'asymétrie (*skewness*) et d'aplatissement (*kurtosis*) des données. Il

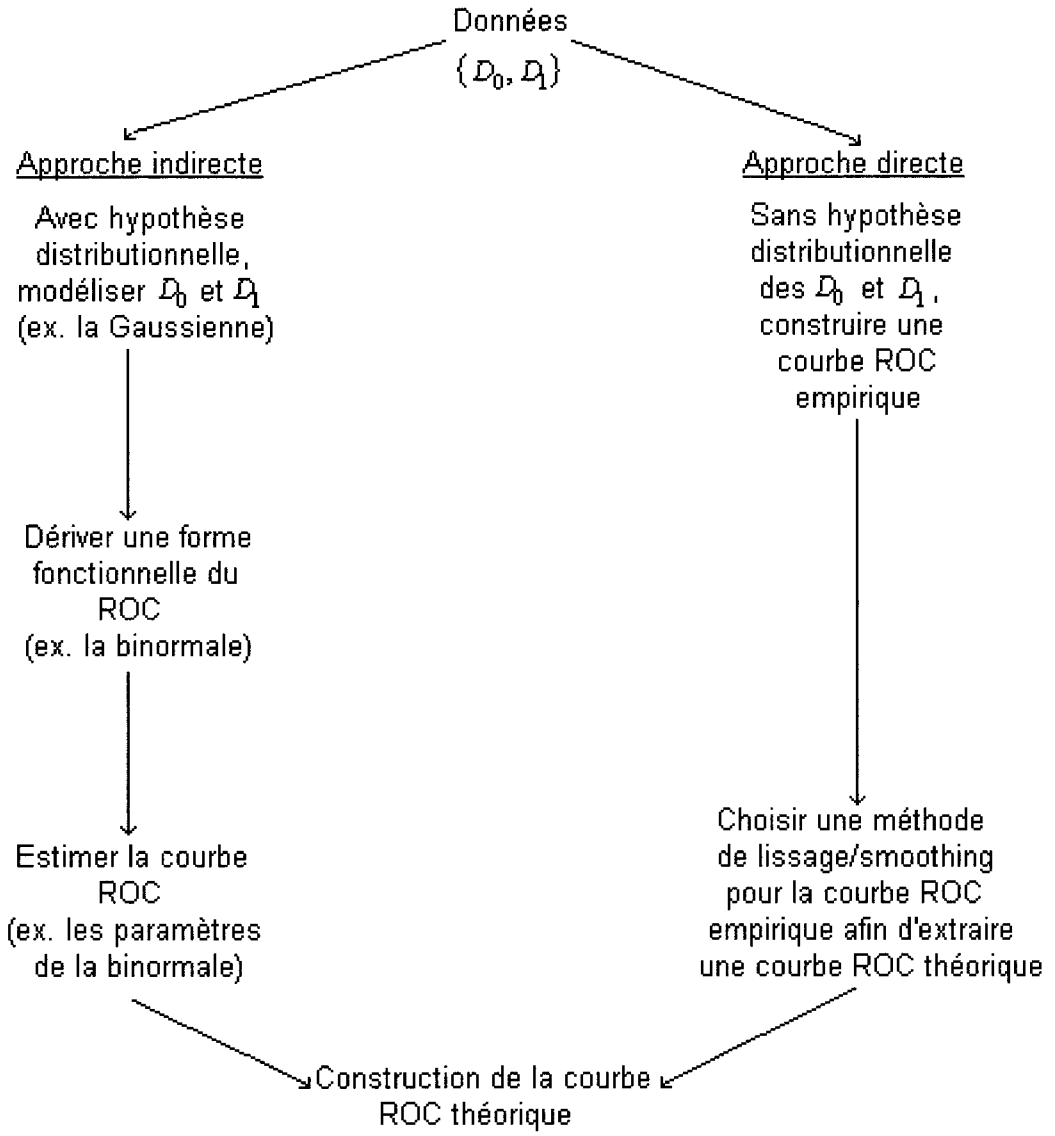

FIGURE 4.1 Approches de la modélisation de la courbe ROC

est à noter que la distribution de Pearson n'est pas une distribution, mais un système qui engendre un nombre considérable de distributions dont la gaussienne. Ainsi la distribution de Pearson peut non seulement capturer les données normales mais aussi d'autres. L'idée est au lieu de supposer que $S_{D_0} \sim N(\mu_0, \sigma_0^2)$ et $S_{D_1} \sim N(\mu_1, \sigma_1^2)$, on suppose que les scores suivent une distribution de Pearson comme paramètres les quatre premiers moments, soit

$$S_{D_0} \sim Pearson(\theta_0) \quad \text{et} \quad S_{D_1} \sim Pearson(\theta_1).$$

où $\theta_0 = \{m1_0, m2_0, m3_0, m4_0\}$ et $\theta_1 = \{m1_1, m2_1, m3_1, m4_1\}$.

Tel mentionné, la distribution de Pearson est un sytème de distributions. Ainsi il n'existe pas de formule fermée dans la modélisation en deux dimensions, *i.e.* une certaine *bi-Pearson* à la manière de la binormale. Par exemple, dans les circonstances où S_{D_0} et S_{D_1} suivent la même distribution, soit gaussienne, gamme ou bêta, il est possible de dériver une formule analytique de la courbe ROC. Mais, dans la situation où $f_{S_0} \neq f_{S_1}$, la solution n'est pas aussi évidente et accessible.

Afin de contourner ce problème, nous avons recours à un technique numérique dont la méthode de simulation par Monte-Carlo, discutée dans la sous-section 4.2.3. Par la simulation de Monte-Carlo, nous allons simuler un ensemble de distributions de Pearson des scores S_{D_0} et S_{D_1} pour ensuite construire un ensemble de courbes ROC. Puisque la forme fonctionnelle de chaque courbe ROC simulée est inconnue, nous utiliserons les propriétés de Monte-Carlo pour extraire différentes statistiques et intervalles de confiance.

Méthodologie

Supposons qu'on dispose entre les mains les scores de chaque population D_0 et D_1, soient S_{D_0} et S_{D_1}, respectivement. Ces scores peuvent provenir des données réelles ou générées aléatoirement sans hypothèse distributionnelle. Ci-dessous, nous décrivons étape par étape notre méthodologie.

1. Directement des données, S_{D_0} et S_{D_1}, on construit la courbe ROC empirique à partir de la défintion 3.1.5 (voir l'annexe B pour la construction d'une courbe ROC) ;
2. On calcule les quatres premiers moments θ_0 et θ_1 à partir de S_{D_0} et S_{D_1}, respectivement ;
3. Avec les deux premiers moments, *i.e.* la moyenne et la variance, on construit la courbe ROC binormale (BIN) à partir de l'équation (4.1.1) en supposant

$$S_{D_0} \sim N(\mu_0, \sigma_0^2) \quad \text{et} \quad S_{D_1} \sim N(\mu_1, \sigma_1^2)$$

4. Avec θ_0 et θ_1, on génère deux ensembles de données à partir de la distribution de Pearson, notée S'_{D_0} et S'_{D_1}, respectivement, en supposant

$$S'_{D_0} \sim Pearson(\theta_0) \quad \text{et} \quad S'_{D_1} \sim Pearson(\theta_1)$$

5. À partir de S'_{D_0} et S'_{D_1}, on construit une première courbe *Monte-Carlo Pearson* (MCP) de la définition 3.1.5, *i.e.*

$$ROC = \{(FPR(t), TPR(t)), t \in \mathbb{R}\}$$

6. En utilisant la technique de Monte-Carlo (voir l'annexe C), on répète les étapes 4 et 5, M fois, donc on obtient M courbes MCP ;
7. Par corollaire de la technique de simulation par Monte-Carlo, on construit un intervalle de confiance sur les valeurs TPR pour chaque point FPR.

Remarque 4.2.1 La construction de l'intervalle de confiance pour la courbe MCP est réalisable grâce à l'interpolation par spline cubique, présentée dans la sous-section 5.3.2. L'interpolation par spline cubique nous permet l'obtention d'une courbe lisse et continue. Donc, on obtient un ensemble de points (FPR, TPR) à intervalle constant sur l'axe des FPR. Cela nous permet de comparer les courbes ROC sur un ensemble de points FPR fixés.

4.2.2 Distributions de Pearson

La famille de Pearson est un rival menaçant à la distribution gaussienne puisqu'elle procure une flexibilité d'adjustement des données intéressante et elle engendre un nombre considérable de distributions. À présent, explorons en détail le fonctionnemet de la distribution de Pearson.

Dans les années 1894-1895, Karl Pearson développa un système de distribution de probabilité, communément appelé les courbes de Pearson, dans le but de décrire les distributions non normales rencontrées dans ses recherches en biométrie. Autrement dit, Pearson chercha à identifier une famille de distributions qui conviendra aux données observées. En plus d'ajuster des modèles pour des distributions de fréquences observées, les distributions de Pearson sont utilisées notamment dans l'approximation de d'autres distributions théoriques et dans l'étude des effets de la non normalité sur les distributions d'échantillonnage.

Pour éviter toutes confusions, les notations dans cette section ne correspondent aucunement aux autres chapitres ou sections.

Dans cette section, nous allons nous référer particulièrement au manuel de Stuart et Ord [44]. Les distributions de Pearson sont grandement basées sur les quatre premiers moments dont la *moyenne*, la *variance*, le *moment de troisième ordre* et le *moment de quatrième ordre*. Ces quatre moments permettront de trouver les quatre paramètres de l'équation suivante

$$\frac{df}{dx} = \frac{(x-a)f}{b_0 + b_1 x + b_2 x^2}. \tag{4.2.1}$$

Cette équation différentielle de premier ordre est appelée les distributions de Pearson. La moindre variation dans les quatre paramètres des distributions de Pearson peut produire une gamme de distributions unimodales. Avant de s'aventurer davantage dans les différents types de distributions de Pearson, nous allons détailler étape par étape l'obtention des solutions explicites possibles de cette équation.

Remarque 4.2.2 Le mode de l'équation (4.2.1) se trouve au point $x = a$.

En réarrangeant l'équation (4.2.1), on obtient

$$(b_0 + b_1 x + b_2 x^2)df = (x - a)f dx$$
$$x^n(b_0 + b_1 x + b_2 x^2)\frac{df}{dx}dx = x^n(x - a)f dx$$

En supposant l'existence des intégrales, on intègre par parties le terme à gauche. Donc, on obtient

$$\left[x^n(b_0 + b_1 x + b_2 x^2)f\right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} nb_0 x^{n-1} + (n-1)b_1 x^n + (n+2)b_2 x^{n+1} f dx$$
$$= \int_{-\infty}^{\infty} x^{n+1} f dx - a\int_{-\infty}^{\infty} x^n f dx. \tag{4.2.2}$$

Maintenant supposons que

$$\lim_{n\to\pm\infty} x^{n+2} f \to 0.$$

Notons que $x^{n+2}f$ vient de l'expression à l'extrémité de la distribution entre crochets. Par la suite, en substituant les moments dans (4.2.2), on obtient une équation récursive sur a

$$0 - nb_0\mu'_{n-1} - (n+1)b_1\mu'_n - (n+2)b_2\mu'_{n+1} = \mu'_{n+1} + a\mu'_n$$

$$nb_0\mu'_{n-1} + [(n+1)b_1 - a]\mu'_n + [(n+2)b_2 + 1]\mu'_{n+1} = 0$$

où μ'_n représente le moment de $n^{ième}$ ordre,

$$\mu'_n = \int (x - a)^n f dx.$$

Remarque 4.2.3 Le coefficient d'asymétrie (*skewness*) et le coefficient d'aplatissement (*kurtosis*) sont définis respectivement comme suit :

$$\gamma_1 = \frac{\mu'_3}{(\mu'_2)^{3/2}}, \quad \text{et} \quad \gamma_2 = \frac{\mu'_4}{(\mu'_2)^2}.$$

Dans notre cas, on définit $n = 1, 2, 3, 4$. De ce fait, les termes $a, b_0, b_1, b_2, \mu_0(\equiv 1)$ et μ_1' ne sont que le fruit des combinaisons des moments. Inversement, les constantes a, b_0, b_1 et b_2 peuvent être réécrites, à son tour, en termes des moments μ_1' à μ_4'. Par contre, si on définit μ_1' comme l'origine, *i.e.* $\mu_1' = 0$, alors ces quatre constantes sont réécrites en termes des moments μ_2' à μ_4'. Avec ces hypothèses, on peut trouver les équations de ces quatre constantes définies comme suit

$$b_1 = a = -\frac{\mu_3(\mu_4 + 3\mu_2^2)}{A} = -\frac{\sqrt{\mu_2}\sqrt{\beta_1}(\beta_2 + 3)}{A'} \tag{4.2.3}$$

$$b_0 = -\frac{\mu_2(4\mu_2\mu_4 - 3\mu_3^2)}{A} = -\frac{\mu_2(4\beta_2 - 3\beta_1)}{A'} \tag{4.2.4}$$

$$b_2 = -\frac{(2\mu_2\mu_4 - 3\mu_3^2 - 6\mu_2^3)}{A} = -\frac{(2\beta_2 - 3\beta_1 - 6)}{A'} \tag{4.2.5}$$

où

$$A = 10\mu_4\mu_2 - 12\mu_3^2 - 18\mu_2^3,$$

$$A' = 10\beta_2 - 12\beta_1 - 18,$$

$$\beta_1 = \mu_3^2,$$

$$\beta_2 = \mu_2\mu_4.$$

Notons que si $\beta_1 = 0$, alors la distribution est symétrique.

À présent, supposons que le mode a est l'origine au lieu de μ_1' et posons $X = x - a$, où $a = 0$, alors le système de Pearson devient

$$\frac{\partial \log f}{\partial X} = \frac{X}{B_0 + B_1 X + B_2 X^2}. \tag{4.2.6}$$

En résolvant $b_0 + b_1 x + b_2 x^2 \equiv B_0 + B_1(x - a) + B_2(x - a)^2$ et en posant $b_1 = a$ de (4.2.3), on a

$$B_0 = b_0 + a^2(1 + b_2),$$

$$B_1 = a(1 + 2b_2),$$

$$B_2 = b_2.$$

L'expression explicite de la fonction de densité f requiert l'intégration du membre droit de (4.2.6). Voici quelques critères afin de déterminer le type de Pearson

i. si $B_2 > 0$, posons $K = \frac{B_1^2}{4B_0B_2}$. Si les racines quadratiques du dénominateur de (4.2.6) sont réelles et de signe opposé, alors $K \leq 0$ (avec $K = 0 \Leftrightarrow B_1 = 0$) et on obtient la famille de distribution bêta, classée Type I.

ii. si $B_2 > 0$, $B_1 = 0$ et $B_0 < 0$, alors on obtient la famille de distribution Type II où la fonction de densité est donnée par

$$f = c\left(1 - \frac{x^2}{a^2}\right)^m, -a < x < a.$$

Notons que dans ce cas-ci, $\beta_1 = 0$ et $\beta_2 < 3$.

iii. si $B_1 = B_2 = 0$ et $B_0 < 0$, alors on obtient la distribution gaussienne. Notons ici que $\beta_1 = 0$ et $\beta_2 = 3$.

iv. si $B_2 = 0$, on obtient la distribution gamma.

En application, le système de Pearson permet d'identifier la meilleure famille de distribution afin d'ajuster ou de calibrer les données d'un échantillon. Ainsi les étapes de sélection et d'ajustement de la distribution de Pearson sont :

i. calculer les valeurs des quatre premiers moments ainsi que β_1 et β_2 des données,

ii. déterminer le type de Pearson selon les critères,

iii. égaliser les moments observés aux moments du type de distribution de Pearson,

iv. résoudre les équations résultantes pour ces paramètres.

4.2.3 Technique de simulation par Monte-Carlo

Bien que la distribution de Pearson soit un choix judicieux, elle demeure complexe au niveau de la dérivée d'une formule analytique en deux dimensions. Par conséquent, la construction de la courbe ROC est faite de manière numérique, soit la technique de simulation par Monte-Carlo.

La méthode de Monte-Carlo est un outil essentiel dans plusieurs domaines comme en mathématique, statistique, physique et ingénierie. Cette méthode encadre toute technique numérique de résolution de problème utilisant des procédés aléatoires. Les pionniers furent John Von Neumann et Stanislaw Ulam qui développèrent cette méthode dans le cadre de la fabrication de la bombe atomique durant la Seconde Guerre mondiale. C'est Nicholas Metropolis qui donna le nom de cette méthode en relation avec les jeux de hasard pratiqués au casino de Monaco [6, 26]. C'est simplement à l'apparition des ordinateurs que la méthode commença à faire fureur.

Remarque 4.2.4 Dans ce qui suit, nous introduisons succinctement les théories autour de la technique de simulation par Monte-Carlo dans le cadre d'une variable aléatoire univariée. Ces théories sont aussi valables pour le cadre multivariées. Le lecteur pourra se référer [27] pour plus de détails.

Pour éviter toutes confusions, les notations dans cette section sont indépendentes des autres chapitres ou sections.

La méthode de Monte-Carlo cherche à résoudre un problème stochastique comme le calcul d'espérance d'une variable aléatoire X. Afin d'y parvenir, une simulation est effectuée sur la variable aléatoire suivant toute la loi de X. Ceci ramène à un problème de calcul d'intégral de la forme

$$I = \int_{[0,1]^d} f(u_1, \ldots, u_d) du_1 \ldots du_d.$$

À présent, posons $X = (f(U_1, \ldots, U_d))$ où les $U_1, \ldots, U_d$ sont des variables aléatoires indépendantes suivant la loi uniforme sur l'intervalle $[0, 1]$. Alors on obtient

$$E(X) = E(f(U_1, \ldots, U_d)) = \int_{[0,1]^d} f(u_1, \ldots, u_d) du_1 \ldots du_d$$

par définition de la loi du n-uplet, $(U_1, \ldots, U_d)$.

Pour la partie de simulation, soit U_i où $i \geq 1$ une suite de variables aléatoires indépendantes suivant une loi uniforme sur $[0, 1]$. Posons $X_1 = f(U_1, \ldots, U_d), X_2 =$

$f(U_{d+1}, \ldots, U_{2d})$, etc. Or, la suite X_i où $i \geq 1$ est aussi une suite de variables aléatoires indépendantes suivant toutes la loi de X. On peut donc approximer I par

$$I \approx \frac{1}{N} (X_1 + \ldots + X_N) .$$

Le fondement théorique de la méthode de Monte-Carlo est basée sur la loi forte des grands nombres et le théorème central limite.

Théorème 4.2.5 *(La loi forte des grands nombres) Soit X_i où $i \geq 1$ une suite de variables aléatoires indépendantes suivant toutes la même loi que la variable aléatoire X. On suppose que $E(|X|) < +\infty$. Alors, pour presque tout ω, i.e. il existe $N \subset \Omega$, avec $P(N) = 0$ et que si $\omega \notin N$,*

$$E(X) = \lim_{n \to +\infty} \frac{1}{n} (X_1(\omega) + \ldots + X_n(\omega)) .$$

Ce théorème inflige l'hypothèse d'intégrabilité des variables aléatoires et justifie la convergence de la méthode. L'autre théorème essentiel est celui de la limite centrale qui affirme que les propriétés statistiques de l'échantillon se rapprochent de plus en plus de celles d'une population gaussienne lorsque la taille de l'échantillon est grande. De plus, elle permet l'obtention des intervalles de confiance.

Théorème 4.2.6 *(Théorème central limite) Soit X_i où $i \geq 1$ est une suite de variables aléatoires indépendantes suivant toutes la même loi que X. On suppose que $E(X^2) < +\infty$. On note σ^2 la variance de X et μ, l'espérance de X. Alors la suite*

$$\frac{\sqrt{n}}{\sigma} \left(\bar{X}_n - \mu\right) \text{ converge en loi vers } G$$

où G est une variable de la loi gaussienne centrée réduite $\mathcal{N}(0, 1)$.

De ce théorème, un corollaire peut être déduit.

Corollaire 4.2.7 *Soit X_i où $i \geq 1$ est une suite de variables aléatoires indépendantes suivant toutes la même loi que X d'espérance μ et de variance σ^2. Alors pour toute*

fonction $h : \mathbb{R} \to \mathbb{R}$ continue bornée, si N représente une variable aléatoire gaussienne $\mathcal{N}(0,1)$

$$\lim_{n\to+\infty} E\left[h\left(\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)\right)\right] = E(h(N)) = \int_{-\infty}^{+\infty} h(x)\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx.$$

De plus pour tout couple de nombre réel $a < b$, on a

$$\lim_{n\to+\infty} P\left(\frac{\sigma}{\sqrt{n}}a \leq \bar{X}_n - \mu \leq \frac{\sigma}{\sqrt{n}}b\right) = \int_a^b \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx.$$

Corollaire 4.2.8 *Une table de fonction de répartition d'une loi gaussienne centrée réduite montre que si N est $\mathcal{N}(0,1)$, $P(|N| \leq 1.96) = 0.95$. Pour un n assez grand, on déduit*

$$P\left(\left|\bar{X}_n - E(X)\right| \leq 1.96\frac{\sigma}{\sqrt{n}}\right) \approx 0.95.$$

Ainsi l'intervalle de confiance de $E(X)$ à 95% est défini comme suit

$$\bar{X}_n \pm 1.96\frac{\sigma}{\sqrt{n}}.$$

Remarque 4.2.9 Le dernier corollaire est important pour notre étude. Il nous permet de construire des intervalles de confiance sur les courbes ROC simulées à partir de la distribution de Pearson. Ainsi la distribution asymptotique des points TPR de ces courbes ROC simulées converge vers une loi normale. En d'autres mots, pour chaque point sur l'axe des FPR, on a M points TPR des M courbes ROC simulées. La distribution de ces points TPR converge asymptotiquement vers une loi normale par le corollaire ci-dessus.

Supposons que nous avons M courbes ROC similuées par la méthode MCP telle que une courbe ROC de la m-ième simulation est donnée par :

$$ROC(m) = \{(FPR_k(m), TPR_k(m)); k = 0, 1, ..., K-1\},$$

$$FPR_k = \frac{k}{K-1}, K > 0.$$

L'algorithme permettant de calculer les intervalles de confiance des courbes simulées par la méthode MCP est comme suit :

. Pour $k = 0, 1, 2, ..., k-1$,

i. $\mu(TPR)(1, k+1) = \mu(TPR)_k = \text{moyenne}(TPR_k(1), TPR_k(2), ..., TPR_k(M))$.

ii. $\sigma(TPR)(1, k+1) = \sigma(TPR)_k = \text{écart-type}(TPR_k(1), TPR_k(2), ..., TPR_k(M))$.

La courbe ROC-MCP moyenne est donnée par :

$$\bar{ROC} = \{(FPR_k, \mu(TPR)_k); k = 0, 1, ..., K-1\}.$$

Les deux intervallecourbes ROC-MCP à un niveau de confiance à 95% est donnée par :

$$ROC_{95\%} = \left\{ (FPR_k, \mu(TPR)_k) \pm 1.96 \frac{\sigma(TPR)_k}{\sqrt{M}}; k = 0, 1, ..., K-1 \right\}.$$

Chapitre 5

Les mesures de performance

Dans le chapitre précédent, nous nous sommes concentré uniquement dans la modélisation de la courbe ROC. À présent que la courbe est construite, la prochaine étape serait la comparaison entre les deux modèles de la courbe ROC soit la binormale (BIN) et la *Monte-Carlo Pearson* (MCP). À savoir, laquelle des courbes reproduit mieux l'empirique. Mais comment les comparer ? En utilisant des mesures de performance comme l'aire sous la courbe ROC, l'aire sous la courbe partielle ROC et l'erreur quadratique moyenne.

5.1 Aire sous la courbe ROC

Il est souvent pratique de résumer la performance d'un test par un simple chiffre dans certaines circonstances, par exemple, lorsqu'il est impraticable de tracer la courbe ROC ou de comparer plusieurs classifications. C'est ainsi que plusieurs attentions se sont détournées vers les indices quantitatifs. L'utilité d'un tel indice est similaire à d'autres mesures telles que la moyenne et la variance qui décrivent la distribution d'une probabilité.

5.1.1 Définition d'AUC

Jusqu'à présent, nous avons simplement prôné les mérites de la courbe ROC en omettant son indice sommaire, l'aire sous la courbe ROC, couramment abrégée *AUC*. La courbe ROC elle seule est bénéfique. mais avec l'ajout d'AUC elle devient substantielle dans l'analyse d'un test diagnostique.

Afin d'alléger l'écriture, nous définissons la courbe ROC sous une autre notation tel que $y = h(x)$ où x et y désignent FPR et TPR, respectivement.

Définition 5.1.1 *L'AUC est définie comme*

$$AUC = \int_0^1 y(x)dx \tag{5.1.1}$$

où $0 \leq AUC \leq 1$.

Un test diagnostique accompli, celui où la courbe ROC est dite idéale, possède une valeur $AUC = 1$. Inversement, un test non informatif, où $y(x) = x$, détient un $AUC = 0.5$. La majorité des tests diagnostiques ont une valeur qui tombe dans cet intervalle.

Remarque 5.1.2 Si deux tests sont ordonnés de façon que le test A soit uniformément meilleur que le test B, voir la figure 3.2, i.e.

$$y_A(x) \geq y_B(x) \quad \Rightarrow \quad AUC_A \geq AUC_B \qquad \forall x \in (0,1).$$

Cependant, l'inverse n'est pas nécessairement tangible, par exemple, dans le cas où les deux courbes se croisent, voir la figure 5.1.

5.1.2 Interprétations

Plusieurs auteurs se sont intéressés à l'indice de la courbe ROC, AUC. Certains parmi eux ont commencé à proposer diverses interprétations d'AUC. La première interprétation est basée sur les propriétés du calcul et de la théorie des probabilités,

la définition 5.1.1, et par le fait que l'aire totale du domaine du ROC est 1. Ceci dit, AUC est une moyenne du TPR pour toutes les valeurs du FPR possibles entre $(0, 1)$.

Une autre interprétation souvent utilisée est qu'AUC égale à la probabilité que le score d'un individu sélectionné aléatoirement et indépendamment d'une population D_1 soit plus élevé qu'un sujet désigné aléatoirement d'une population D_0 [2, 20].

Corollaire 5.1.3 *Soient S_{D_0} et S_{D_1} des scores attribués à un individu sélectionné aléatoirement et indépendamment de D_1 et D_0, respectivement, alors*

$$AUC = P(S_{D_1} \geq S_{D_0}). \tag{5.1.2}$$

Démonstration: Donc, par la définition 5.1.1, on a

$$\begin{aligned}
AUC &= \int_0^1 y(x)dx \\
&= \int_{\infty}^{-\infty} y(t)\frac{dx}{dt}dt && \text{(par changement de var. et prop. 3.2.1)} \\
&= -\int_{\infty}^{-\infty} P(S > t \mid D_1)P(t \mid D_0)dt && \text{(par déf. 3.1.5 et éq. (3.2.3))} \\
&= \int_{-\infty}^{\infty} P(S > t \mid D_1)P(t \mid D_0)dt \\
&= \int_{-\infty}^{\infty} P(S_{D_1} > t \cap S_{D_0} = t)dt && \text{(par hypothèse d'indépendance)} \\
&= \int_{-\infty}^{\infty} P(S_{D_1} > S_{D_0} \mid t)dt \\
&= P(S_{D_1} > S_{D_0}) && \text{(par théorème de la probabilité totale)}
\end{aligned}$$

∎

Selon Hanley et McNeil, la caractéristique importante de la preuve est qu'il n'y a pas de supposition sur la forme des distributions de S_{D_0} et S_{D_1} [20]. Bamber nota que l'aire sous la courbe ROC empirique est équivalente au test de Wilcoxon-Mann-Whitney qui compte le nombre total de paires (X, Y) dans lesquelles $Y > X$ [2]. Cette relation permit à Hanley et McNeil d'utiliser les propriétés de la statistique de Wilcoxon pour prédire les propriétés statistiques d'AUC [20].

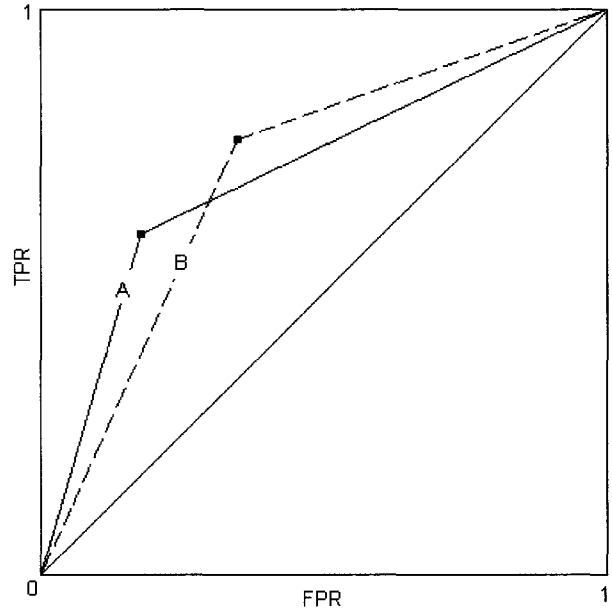

FIGURE 5.1 Deux courbes distinctes ayant la même valeur d'AUC

Remarque 5.1.4 Bien que la mesure AUC soit simple à comprendre, elle n'est pas toujours robuste. En effet, il est possible mathématiquement de construire deux courbes ROC distinctes ayant exactement la même valeur AUC, voir la figure 5.1. Par conséquent, il est préférable d'avoir recours à d'autres mesures dont la mesure de l'aire sous la courbe ROC partielle ou *pAUC*.

5.1.3 AUC pour le modèle binormal

Dans la section 4.1, nous avons défini le modèle binormal de la courbe ROC. Le bénéfice d'utiliser le modèle binormal est dû à sa forme fermée, voir proposition 4.1.1. Grâce à ceci, une formule analytique de l'AUC est définie comme ci-dessous.

Corollaire 5.1.5 *L'aire sous la courbe ROC binormale est définie par*

$$AUC = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) \tag{5.1.3}$$

où les valeurs a et b sont déterminées par l'équation (4.1.2).

Démonstration: Par le corollaire 5.1.3, on a que

$$\begin{aligned} AUC &= P(S_{D_1} > S_{D_0}) \\ &= P(S_{D_1} - S_{D_0} > 0) \qquad \text{(par indépendance de } S_{D_1} \text{ et } S_{D_0}) \end{aligned}$$

Par la théorie de probabilité, si

$$S_{D_0} \sim N(\mu_0, \sigma_0^2) \quad \perp \quad S_{D_1} \sim N(\mu_1, \sigma_1^2)$$

alors,

$$S_{D_1} - S_{D_1} \sim N(\mu_1 - \mu_0, \sigma_1^2 + \sigma_0^2).$$

Si Z est une variable aléatoire normale standardisée, donc

$$\begin{aligned} AUC &= P\left(Z > \frac{0 - (\mu_1 - \mu_0)}{\sqrt{\sigma_1^2 + \sigma_0^2}}\right) \\ &= 1 - \Phi\left(\frac{-\mu_1 + \mu_0}{\sqrt{\sigma_1^2 + \sigma_0^2}}\right) \\ &= \Phi\left(\frac{\mu_1 - \mu_0}{\sqrt{\sigma_1^2 + \sigma_0^2}}\right) \\ &= \Phi\left(\frac{a}{\sqrt{1 + b^2}}\right) \qquad \text{(en divisant par } \sigma_1) \end{aligned}$$

∎

5.1.4 Méthodes d'estimation d'AUC

Contrairement au modèle de la binormale, l'approche proposée pour la construction de la courbe ROC par les distributions de Pearson ne possède pas de forme fermée. Puisque nous n'avons pas de formule analytique pour le calcul de l'AUC pour

la courbe MCP, nous utiliserons deux méthodes d'estimation soit numériquement par la méthode du trapèze et statistiquement par la statistique U de Mann-Whitney.

Rappelons que l'objectif ultime est de reproduire la courbe ROC empirique à partir des deux modèles présentés. Ainsi, nous ne cherchons pas à comparer quelle mesure est la plus performante, mais plutôt laquelle des valeurs AUC se rapproche le plus à celle de l'empirique.

Méthode du trapèze

La méthode du trapèze est une technique numérique simple qui approxime la valeur d'une intégrale définie. Soit l'intégrale définie

$$\int_a^b h(x)dx. \tag{5.1.4}$$

On assume que $h(x)$ est continue sur $[a, b]$. Par la suite, on partitionne l'intervalle $[a, b]$ en n sous-intervalles de même longueur où

$$\Delta x = \frac{b-a}{n}. \tag{5.1.5}$$

En utilisant $n+1$ points, on a

$$\begin{aligned} x_0 &= a \\ x_1 &= a + \Delta x \\ x_2 &= a + 2\Delta x \\ &\vdots \\ x_n &= a + n\Delta x = b. \end{aligned}$$

Ainsi on peut calculer la valeur de $h(x)$, soit

$$\begin{aligned} y_0 &= h(x_0) \\ &\vdots \\ y_n &= h(x_n). \end{aligned}$$

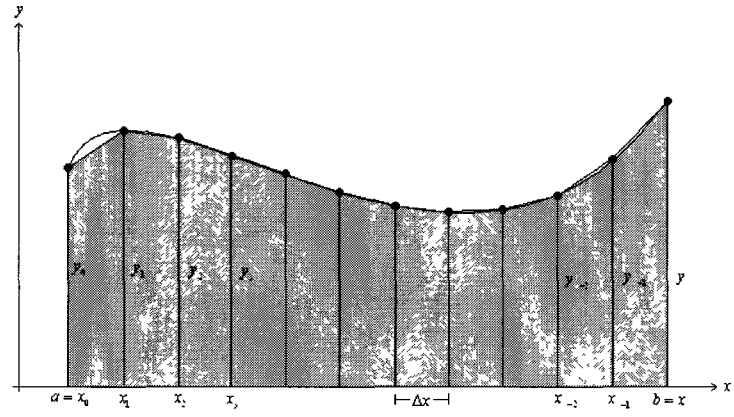

FIGURE 5.2 Division par segment d'une courbe sur l'intervalle $[a, b]$

Par la suite, on approxime l'intégrale en utilisant les n trapèzes formés par les segments du point (x_{i-1}, y_{i-1}) jusqu'à (x_i, y_i) pour $1 \leq i \leq n$ tel qu'illustré par la figure 5.2.

En additionnant l'aire des rectangles et des triangles, voir la figure 5.3, on obtient l'aire d'un trapèze, soit

$$\begin{aligned} A_t &= y_0 \Delta x + \frac{1}{2}(y_1 - y_0)\Delta x \\ &= \frac{(y_0 + y_1)\Delta x}{2}. \end{aligned}$$

Donc, l'approximation est obtenue en additionnant l'aire des n trapèzes, *i.e.*

$$\begin{aligned} \int_a^b h(x)dx &\cong \frac{(y_0 + y_1)\Delta x}{2} + \cdots + \frac{(y_{n-1} + y_n)\Delta x}{2} \\ &\approx \frac{\Delta x}{2}\left(y_0 + 2y_1 + \cdots + 2y_{n-1} + y_n\right). \end{aligned}$$

Remarque 5.1.6 La définition ci-haut décrit un cas général. Pour le calcul de l'AUC, l'intervalle serait $[0, 1]$ puisque la courbe ROC est bornée sur $[0, 1]$.

Statistique U de Mann-Whitney

La statistique U de Mann-Whitney, aussi connue sous le nom de Wilcoxon-Mann-Whitney, est une méthode d'estimation d'AUC non-paramétrique. Cette statistique

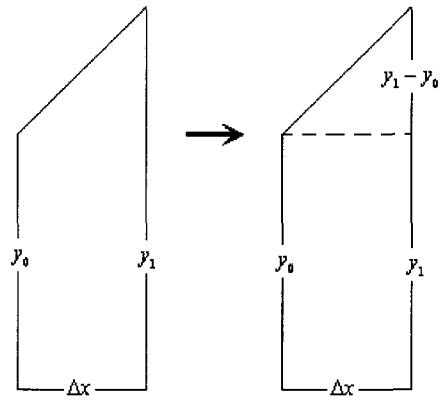

FIGURE 5.3 Illustration du calcul de l'aire d'un trapèze

cherche à vérifier si les éléments de D_0 précèdent D_1 dans une classification par ordre croissant sur une même échelle ordinale. Le choix d'estimation de l'AUC par la statistique U de Mann-Whitney est due aux conclusions apportés par Bamber [2]. L'auteur nota que l'aire sous la courbe ROC est étroitement liée à la statistique U de Mann-Whitney.

Définition 5.1.7 *La statistique U de Mann-Whitney est définie comme suit*

$$U = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \left[I\left(s_{1_j} > s_{0_i}\right) + \frac{1}{2} I\left(s_{1_j} = s_{0_i}\right) \right] \tag{5.1.6}$$

où n_0, n_1 et s_0, s_1 sont la taille de l'échantillon et les scores de D_0 et D_1, respectivement.

Puisque nous nous concentrons uniquement sur les variables aléatoires continues, alors la probabilité d'obtenir des noeuds est théoriquement nulle. Ainsi comme Bamber [2] le pointa

$$\begin{aligned} AUC &= P(S_1 > S_0) && \text{(par 5.1.3)} \\ &= E(U). \end{aligned}$$

Donc U est un estimateur non-biaisé d'AUC.

5.2 Aire sous la courbe partielle ROC

L'élégance de l'AUC est due à sa simplicité et sa commodité. McClish pointa que l'AUC attribue le même poids pour toutes les valeurs possibles du seuil décisionnel [34]. Ainsi, l'AUC n'est pas toujours un choix judicieux tel indiqué dans la remarque 5.1.4. Dans le cas où les courbes ROC se croisent, l'AUC ignore la possibilité qu'un test diagnostique performe mieux à un certain intervalle donné tandis que l'autre test est supérieur à l'intervalle restant. Dans un essai clinique, ceci peut s'interpréter par le fait qu'un diagnostic fonctionne uniquement pour certains types de patients. Un autre scénario plausible est lorsque le spécialiste s'intéresse particulièrement à une portion de FPR. Pour ces deux situations, l'aire sous la courbe partielle, pAUC, serait une mesure désirable puisqu'elle se concentre sur un intervalle de FPR restreint.

Pour alléger l'écriture, revenons à cette notation de la courbe ROC, *i.e.* $y = h(x)$ où x et y désignent FPR et TPR, respectivement.

Définition 5.2.1 *De manière analogue au calcul AUC, on a pour* $b > a$ *et* $a, b \in [0, 1]$,

$$pAUC = \int_a^b y(x)dx.$$

La méthode du trapèze s'avère un candidat pratique pour l'estimation du pAUC. Il est à souligner que l'intervalle est entre $0 \leq a < b \leq 1$. Par contre, le calcul du pAUC ne peut s'effectuer par la statistique U de Mann-Whitney. Le pAUC mesure l'aire en sous de la courbe par segment, alors que Mann-Whitney est une statistique de S_{D_0} et S_{D_1}. En autres mots, pAUC travaille numériquement avec la courbe, alors que Mann-Whitney travaille directement avec les données.

5.3 Erreur quadratique moyenne

Rappelons que l'objectif n'est pas simplement d'obtenir une valeur quantitative qui se rapproche à celle de l'empirique, mais d'obtenir aussi une courbe avoisinante à celle-ci. Puisque la comparaison quantitative peut s'effectuer à l'aide d'AUC et du pAUC, mais qu'en est-il du côté graphique ? Une solution possible serait l'erreur quadratique moyenne (MSE).

5.3.1 Définition du MSE

L'idée est de comparer point par point la distance entre les deux courbes théoriques. soit la courbe binormale (BIN) et la courbe MCP, *versus* la courbe empirique.

Définition 5.3.1 *Soient la courbe ROC empirique,*

$$ROC_E = \{(FPR, TPR); FPR, TPR \in [0,1]\},$$

et la courbe ROC théorique (BIN ou MCP),

$$ROC_T = \{(\widetilde{FPR}, \widetilde{TPR}); \widetilde{FPR}, \widetilde{TPR} \in [0,1]\}.$$

On définit MSE des points sur l'axe des TPR de la courbe ROC pour un ensemble de points de l'axe des FPR fixés comme suit

$$MSE = \frac{1}{K}\sum_{k=1}^{K}(TPR_k - \widetilde{TPR}_k)^2 \tag{5.3.1}$$

où k représente le nombre de découpes sur l'axe des FPR.

Remarque 5.3.2 L'idée est de calculer une distance euclidienne entre deux courbes sur un ensemble fini K points à intervalle régulier sur l'axe des x (ou FPR). Par exemple, nous voulons comparer la similitude ou l'ajustement de la courbe binormale par rapport à la courbe empirique. Les étapes du calcul sont ci-dessous. Supposons qu'on a un ensemble K points sur l'axe des FPR tels que $FPR_k = \frac{k}{K-1}$ pour $k = 0, 1, ..., K-1$.

$ERR = 0$

Pour $k = 0 : K - 1$

i. Calcule le couple (FPR_k, TPR_k) de la courbe empirique.

ii. Calcule le couple $(\widetilde{FPR_k}, \widetilde{TPR_k})$ de la courbe binormale.

iii. $ERR = ERR + (TPR_k - \widetilde{TPR_k})^2$

Fin

$MSE = ERR/K$

En ce qui a trait du calcul de MSE pour la méthode MCP, on calcule un MSE pour chaque courbe ROC simulée. Supposons qu'on a M courbes simulées, on aura M calculs de MSE. Ainsi on peut calculer une moyenne et une variance comme suit :

$$\mu(MSE) = \frac{1}{M}\sum_{m=1}^{M} MSE_m,$$

$$\sigma^2(MSE) = \frac{1}{M}\sum_{m=1}^{M}(MSE_m - \mu(MSE))^2.$$

Toujours dans le même concept, nous allons utiliser MSE pour comparer chaque aire partielle de la courbe ROC théorique *versus* l'empirique.

Définition 5.3.3 *Soit un ensemble* $\{a_1 < a_2 < \ldots < a_l < \ldots < a_L\}$ *où* $a_l \in [0,1] \forall l$ *où* $l = 1, \ldots, L$ *représente l'indice des seaux ou sous-intervalles sur l'axe des FPR de la courbe ROC tel que* $FPR_{l+1} - FPR_l = 1/L$. *On définit MSE de l'aire sous la courbe ROC partielle pour des intervalles sur l'axe des FPR prédéfinies comme suit*

$$MSEpAUC = \frac{1}{L}\sum_{l=1}^{L}(pAUC_{E_l} - pAUC_{T_l})^2 \tag{5.3.2}$$

où $pAUC_E$ *est le pAUC de la courbe empirique et* $pAUC_T$ *est celle théorique.*

Remarque 5.3.4 Pour le calculer du pAUC, nous allons diviser l'intervalle $[0,1]$ (l'axe des FPR) en L sous-intervalles. Ainsi, on obtient L valeurs pAUC. Encore une

fois, on cherche à comparer le rapprochement de la valeur du pAUC de la binormale à celle de l'empirique. Les étapes du calcul de MSEpAUC sont décrites ci-dessous.

$ERR = 0$

Pour $l = 0 : L - 1$

i. Calcule le pAUC sur $]\frac{l}{L}, \frac{l+1}{L}]$ l'axe des FPR de la courbe empirique.

ii. Calcule le pAUC sur $]\frac{l}{L}, \frac{l+1}{L}]$ l'axe des FPR de la courbe binormale.

iii. $ERR = ERR + (pAUC_{E_l} - pAUC_{T_l})^2$

Fin

$MSE = ERR/L$

Quant à la méthode MCP, on calcule le MSEpAUC pour chaque courbe ROC simulée. Avec M simulations, on obtient M MSEpAUC. De ceci, on peut calculer une moyenne et une variance comme suit :

$$\mu(MSEpAUC) = \frac{1}{M}\sum_{m=1}^{M} MSEpAUC_m,$$

$$\sigma^2(MSEpAUC) = \frac{1}{M}\sum_{m=1}^{M}(MSEpAUC_m - \mu(MSEpAUC))^2.$$

5.3.2 Interpolation par spline cubique

En application, il est impératif d'obtenir des courbes régulières passant par un grand nombre de points. Malheureusement, cet énoncé ne s'applique pas pour notre approche de la courbe MCP. Au contraire, pour chaque courbe ROC simulée, un ensemble de points discontinus à intervalle irrégulier est obtenu. Afin d'obtenir un point précis, non défini par la simulation, nous avons recours à l'interpolation par spline cubique. En ayant les coordonnées (FPR, TPR) relative aux points simulés, le calcul du pAUC et du MSE deviennent moins lourds.

La régularité d'une fonction peut se mesurer par ses dérivées. Assurément, plus une fonction est différentiable, plus sa courbe est lisse et donc la fonction est régulière. L'objectif du spline cubique est d'obtenir une fonction d'interpolation qui est lisse au dérivée première et continue au dérivée seconde. La théorie de cette section provient majoritairement du manuel de Shikin et Plis [43].

Définition

Pour éviter toutes confusions, les notations dans cette section ne correspondent aucunement aux autres chapitres ou sections.

Soit (x_i, y_i) des points d'interpolation où $i = 0, \ldots, m$. Posons, $a = x_0$ et $b = x_m$. Une fonction, $S(x)$, définie sur l'intervalle $[a, b]$ est appelée une fonction d'interpolation du spline cubique si la fonction

i. est un polynôme cubique

$$S(x) = S_i(x) = a_0^{(i)} + a_1^{(i)}(x - x_i) + a_2^{(i)}(x - x_i)^2 + a_3^{(i)}(x - x_i)^3$$

ii. est deux fois continûment dérivable, *i.e.* $S(x) \in C^2[a, b]$

iii. satisfait aux conditions

$$S(x_i) = y_i, \quad \text{où} \quad i = 0, 1, \ldots, m.$$

La fonction de spline, $S(x)$, est un polynôme cubique sur chaque intervalle partiel $[x_i, x_{i+1}]$ où $i = 0, 1, \ldots, m - 1$ et donc, elle est définie sur l'intervalle par quatre coefficients. Le nombre total d'intervalle partiel est égal à m. Ainsi lorsque les $4m$ valeurs des coefficients seront trouvées

$$a_0^{(i)}, a_1^{(i)}, a_2^{(i)}, a_3^{(i)} \quad \text{où} \quad i = 0, 1, \ldots, m - 1$$

la fonction de spline désirée sera définie.

La condition où $S(x) \in C^2[a,b]$ suggère la continuité de la fonction $S(x)$ et ses dérivées, $S'(x)$ et $S''(x)$ à tous les points intérieurs $x_1, \ldots, x_{m-1}$. Ainsi, il existe $3(m-1)$ conditions ou équations pour les coefficients désirés. À partir du troisième critère ci-haut, le nombre total de conditions ou équations équivaut à $3(m-1)+(m+1) = 4m-2$.

Les conditions sur les bornes

Le problème dans une fonction d'interpolation est le choix des conditions sur les bornes. Ci-dessous sont décrits deux autres conditions qui sont données sous forme de contrainte sur les valeurs du spline et/ou les valeurs de leurs dérivées sur les bornes a et b.

i. Les valeurs de la dérivée première sont données aux bornes de l'intervalle $[a,b]$

$$S'(a) = f'(a), \quad S'(b) = f'(b).$$

ii. Les valeurs de la dérivée seconde sont données aux bornes de l'intervalle $[a,b]$

$$S''(a) = f''(a), \quad S''(b) = f''(b).$$

iii. Les valeurs respectives de la dérivée première et seconde sont égaux aux bornes l'intervalle $[a,b]$

$$S'(a) = S'(b), \quad S''(a) = S''(b).$$

Remarque 5.3.5 La troisième condition est appelée périodique. On choisit cette condition lorsque $f(x)$ est une fonction interpolante périodique avec une période égale à $T = b - a$.

Le choix des conditions des bornes est souvent déterminé par la disponibilité d'information additionnelle sur le comportement de la fonction $f(x)$ estimée. La première condition est applicable si les valeurs de la dérivée première $f'(x)$ au borne de l'intervalle $[a,b]$ sont connues. Similairement, la deuxième condition est convenable si les

valeurs de la dérivée seconde $f''(x)$ au borne de l'intervalle $[a, b]$ sont connues. Si $f(x)$ est une fonction périodique le choix revient à la troisième condition.

Remarque 5.3.6 Dans une situation où l'information additionnelle est manquante, les conditions *naturelles* des bornes, $S''(a) = 0$ et $S''(b) = 0$, sont fréquemment utilisées.

La construction de la fonction d'interpolation du spline cubique

Une méthode de calculs est décrite ci-dessous avec $m + 1$ valeurs de coefficients au lieu de $4m$. La fonction d'interpolation du spline dans chaque intervalle $[x_i, x_{i+1}]$ où $i = 0, 1, \ldots, m-1$ est calculé comme suit

$$S(x) = S(x_i) = y_i(1-t)^2(1+2t) + y_{i+1}t^2(3-2t) + n_i h_i t(1-t)^2 - n_{i+1}t^2(1-t), \quad (5.3.3)$$

où

$$h_i = x_{i+1} - x_i, \quad t = \frac{x - x_i}{h_i}$$

et les nombres n_i où $i = 0, 1, \ldots, m$ sont une solution du système linéaire algébrique. La forme du sytème dépend grandement du choix des conditions sur les bornes.

Pour la première et la deuxième condition, le système prend la forme comme suit

$$\begin{cases} 2n_0 + \mu_0^* n_1 = c_0^*, \\ \lambda_i n_{i-1} + 2m_i + \mu_i n_{i+1} = c_i, \quad i = 1, 2, \ldots, m-1, \\ \lambda_m^* n_{m-1} + 2n_m = c_m^*, \end{cases}$$

où

$$c_i = 3\left(\mu_i \frac{y_{i+1} - y_i}{h_i} + \lambda_i \frac{y_i - y_{i-1}}{h_{i-1}}\right).$$

Les coefficients $\mu_0^*, c_0^*, \lambda_m^*$ et c_m^* dépendent du choix des conditions sur les bornes. Ainsi avec la première condition, on a

$$\mu_0^* = 0, \quad c_0^* = 2y_0',$$

$$\lambda_m^* = 0, \quad c_m^* = 2y'_m.$$

et avec la deuxième condition, on a

$$\mu_0^* = 1, \quad c_0^* = 3\frac{y_1 - y_0}{h_0} - \frac{h_0}{2}y''_0,$$

$$\lambda_m^* = 1, \quad c_m^* = 3\frac{y_m - y_{m-1}}{h_{m-1}} - \frac{h_{m-1}}{2}y''_m.$$

Pour la troisième condition, le système linéaire pour les nombres n_i où $i = 1, 2, \ldots, m$ peut s'écrire comme suit

$$\begin{cases} 2n_1 + \mu_1 n_2 + \lambda_1 n_m = c_1, \\ \lambda_i n_{i-1} + 2n_i + \mu_i n_{i+1} = c_i, \quad i = 1, 2, \ldots, m-1, \\ \mu_m n_1 + \lambda_m n_{m-1} + 2n_m = c_m, \end{cases}$$

Puisque la fonction est périodique, alors $n_0 = n_m$. Ainsi le système se retrouve avec m inconnus.

Chapitre 6

Résultats et discussions

Dans ce chapitre, nous procédons à l'analyse graphique et quantitative des résultats de notre étude. Rappelons que notre objectif est de comparer la flexibilité et l'ajustement de la courbe ROC des modèles binormale (BIN) et *Monte-Carlo Pearson* (MCP) sur les données réelles et simulées. En autres mots, nous cherchons à vérifier laquelle des courbes théoriques reproduit mieux la courbe empirique. L'analyse se fera en deux parties : la première sur l'étude de simulation et la seconde sur l'étude sur les données réelles. Pour une meilleure idée globale sur l'analyse des résultats, voir la figure 6.1.

6.1 Étude sur les données simulées

Les données simulées sont générées aléatoirement en supposant que

$$S_{D_0} \sim f_{S_0}(x) \quad \text{et} \quad S_{D_1} \sim f_{S_1}(x)$$

où $f_{S_0}(x)$ et $f_{S_1}(x)$ sont des densités de probabilités des distributions gaussienne, gamma ou bêta. Une simulation de 2500 valeurs pour les populations non-malades et malades fut exécutée. Le choix d'une taille d'échantillon large nous permet de capturer les propriétés asymptotiques des distributions simulées. Afin de comparer les deux

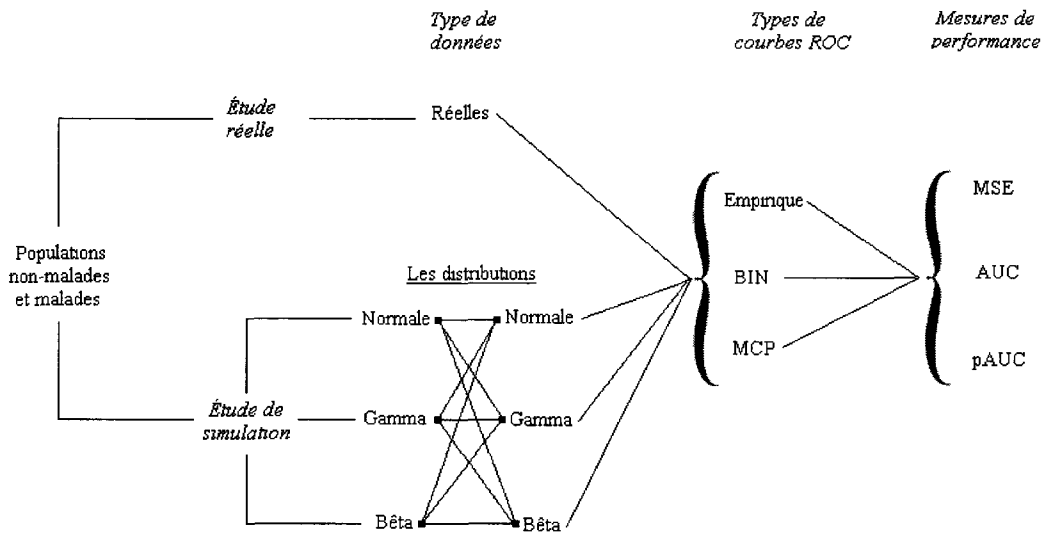

FIGURE 6.1 Schéma de l'analyse des résultats

méthodes, soient BIN et MCP, nous avons choisi trois ensembles de paramètres où la discrimination entre les deux populations est forte, modérée et faible.

6.1.1 Analyse graphique

Le but de notre étude est d'identifier quel modèle, entre BIN et MCP, réplique mieux la courbe ROC empirique. Une analyse graphique nous est utile afin d'avoir une idée générale de la forme des courbes théoriques par rapport à celle empirique. Notons que dans cette section, la courbe MCP se réfère à la courbe MCP moyenne des $M = 1000$ courbes MCP simulées par la technique de Monte-Carlo avec un intervalle de confiance à 95%, voir la remarque 4.2.9.

Simulation des distributions gaussiennes

Supposons que $S_{D_0} \sim N(0,1)$. Dans le but d'étudier le comportement des deux méthodes selon le degré de discrimination, nous avons varié les paramètres de $f_{S_1}(x)$. Une forte dispersion est observée dans l'histogramme 6.2 lorsque $S_{D_1} \sim N(5,1)$. Dans

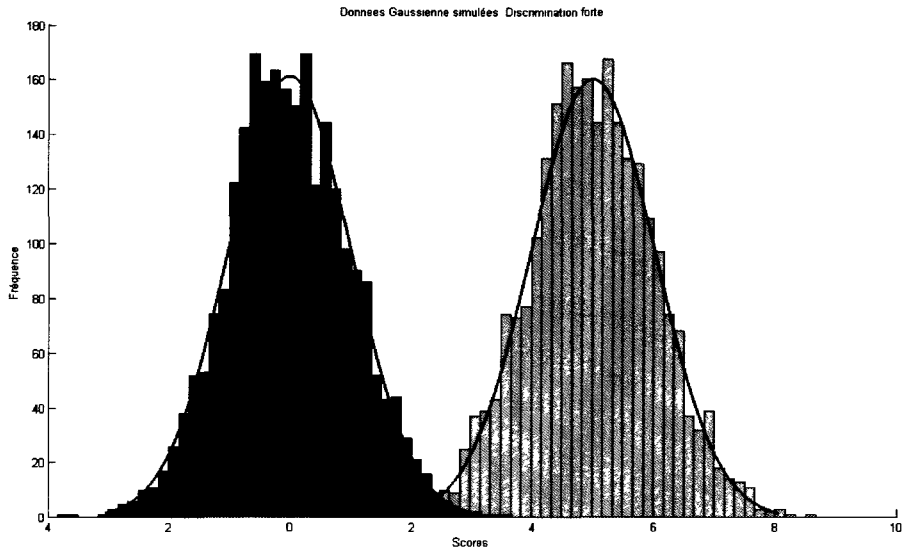

FIGURE 6.2 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim N(0,1)$ et $S_{D_1} \sim N(5,1)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

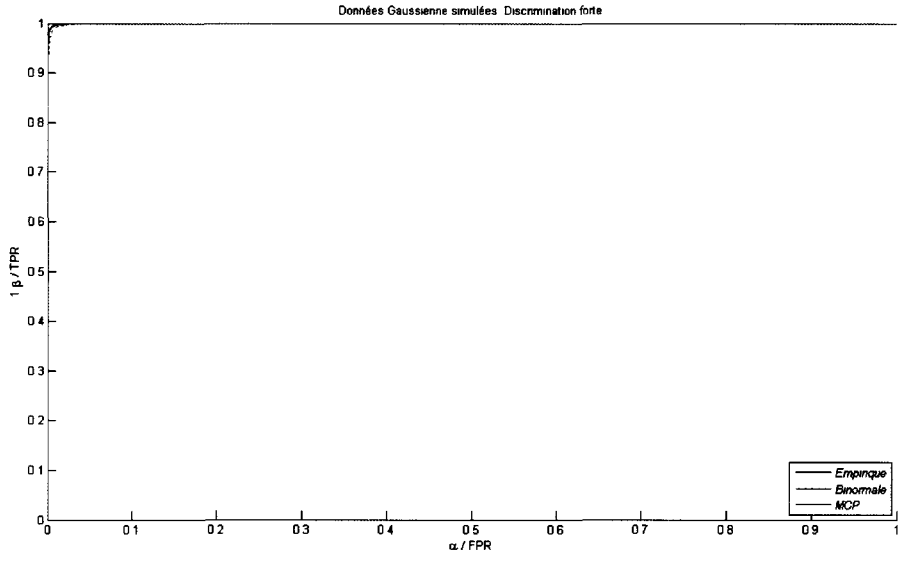

FIGURE 6.3 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim N(0,1)$ et $S_{D_1} \sim N(5,1)$

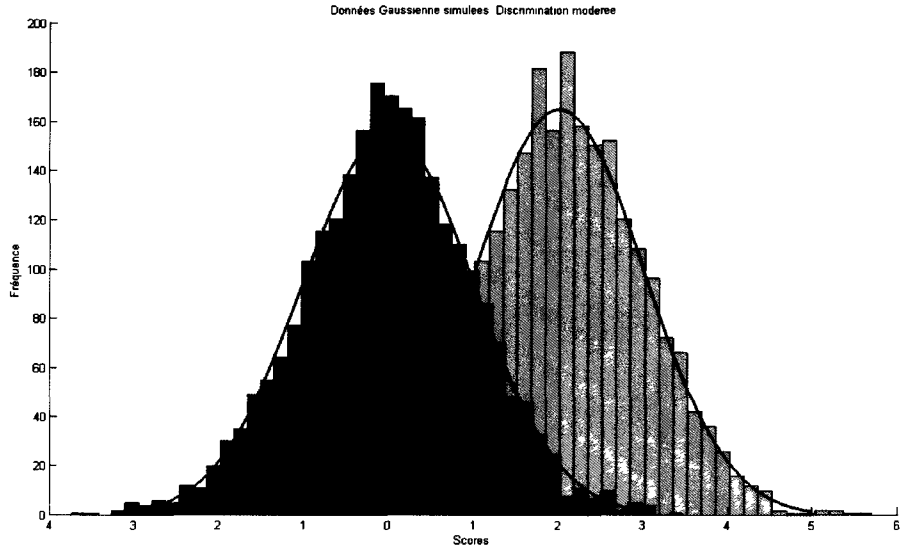

FIGURE 6.4 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim N(0,1)$ et $S_{D_1} \sim N(2,1)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

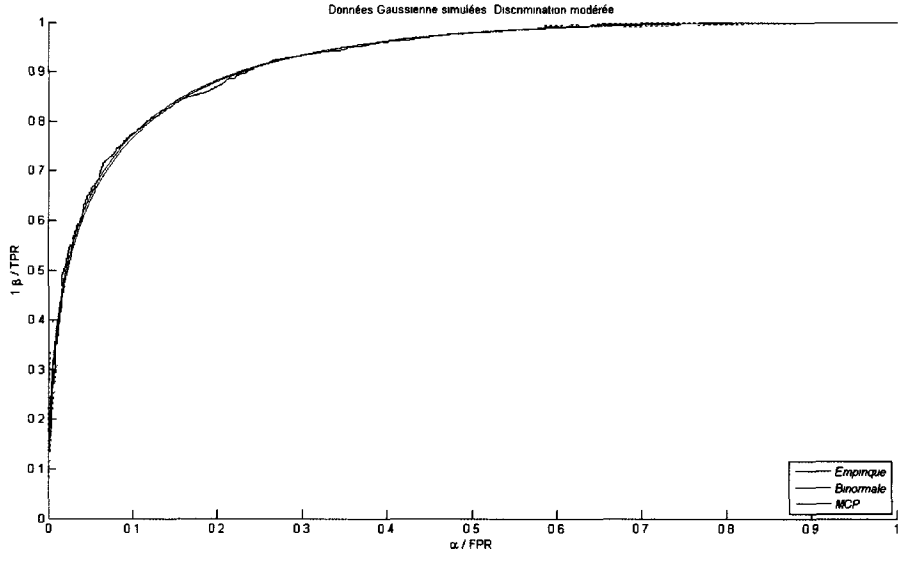

FIGURE 6.5 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim N(0,1)$ et $S_{D_1} \sim N(2,1)$

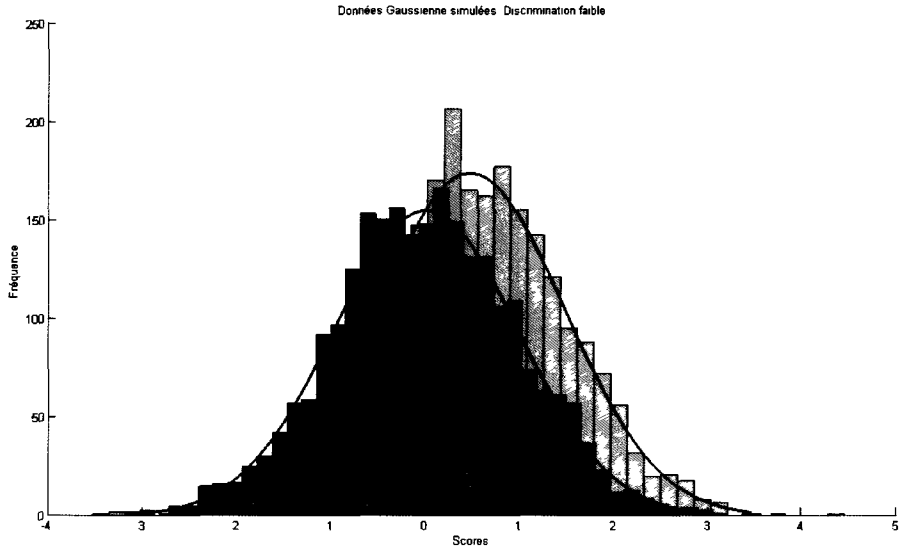

FIGURE 6.6 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim N(0,1)$ et $S_{D_1} \sim N(0.5,1)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

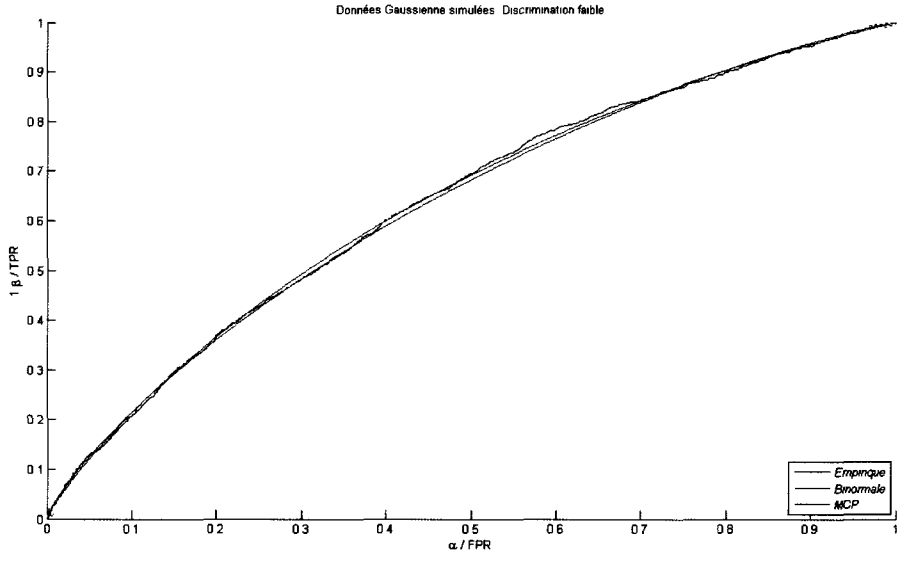

FIGURE 6.7 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim N(0,1)$ et $S_{D_1} \sim N(0.5,1)$

les histogrammes 6.4 et 6.6, une discrimination modérée et faible est obtenue en supposant $S_{D_1} \sim N(2,1)$ et $S_{D_1} \sim N(0.5,1)$, respectivement. Ces histogrammes nous montre que la distribution des scores de D_0 et D_1 s'apparente bien à une distribution gaussienne comme espérée. Les figures 6.3, 6.5 et 6.7 illustrent un exemple d'un intervalle de confiance à 95% de la courbe MCP dont la courbe empirique se trouve entre les bornes supérieure et inférieure. À première vue, les deux courbes théoriques respectent bien la propriété 3.2.1, *i.e.* la montonie d'une fonction croissante, à l'opposé, de la courbe empirique. Tel soupçonné, la courbe BIN réplique agréablement la courbe empirique puisque les scores des deux populations proviennent d'une distribution gaussienne. Cependant, la courbe MCP est un adversaire redoutable, puisqu'elle la suit de très près. Dans le cas d'une forte discrimination, on peut même devancer que les trois courbes ROC sont fidèlement superposées. Une légère variation entre les courbes BIN et MCP est observée lorsque la discrimination est modérée et faible. Puisque cette différence est négligeable, on peut affirmer que les deux modèles perfoment aussi bien l'un que l'autre. Afin de trancher une conclusion, une étude des mesures de performance serait d'une grande utilité. Mais avant de se lancer dans l'analyse quantitative, il est intéressant de voir le comportement du modèle BIN pour des données autre que la gaussienne. Demeure-t-il aussi performant que le modèle MCP ?

Simulation des distributions gamma

Supposons $S_{D_0} \sim Gamma(1.25,2)$. Dans l'histogramme 6.8, une grande dispersion entre les deux populations est obtenue lorsque $S_{D_1} \sim Gamma(10,4)$. Selon les histogrammes 6.10 et 6.12, une discrimination modérée et faible est obtenue quand $S_{D_1} \sim Gamma(5,3)$ et $S_{D_1} \sim Gamma(2.75,2)$, respectivement. La distribution des scores de D_0 et D_1 s'apparente bien à une distribution gamma comme souhaitée. D'après les figures 6.9, 6.11 et 6.13, les courbes ROC théoriques respectent la pro-

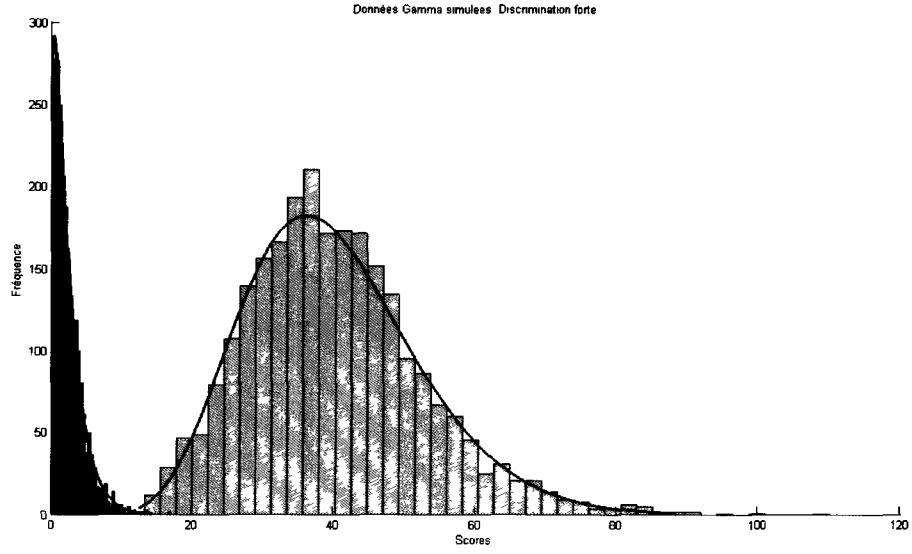

FIGURE 6.8 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim Gamma(1.25, 2)$ et $S_{D_1} \sim Gamma(10, 4)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

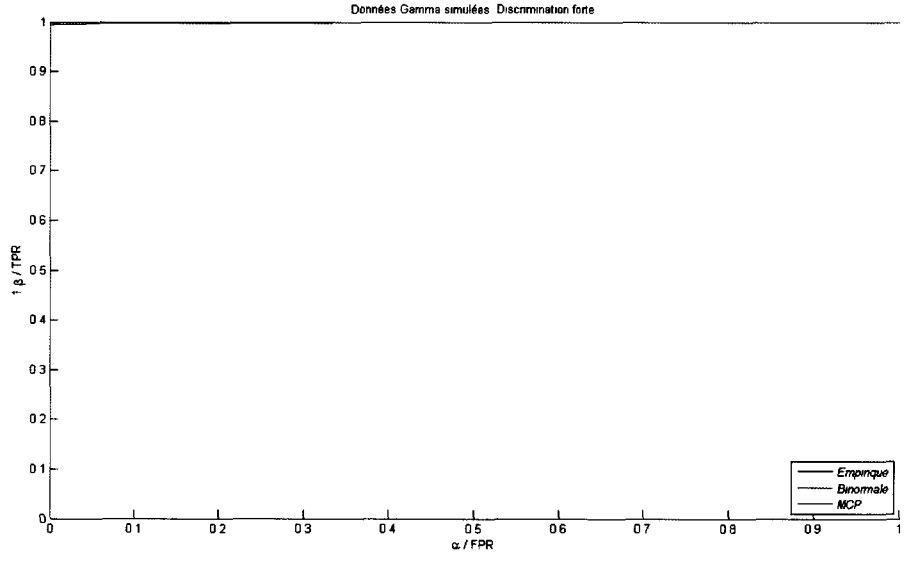

FIGURE 6.9 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} Gamma(1.25, 2)$ et $S_{D_1} \sim Gamma(10, 4)$

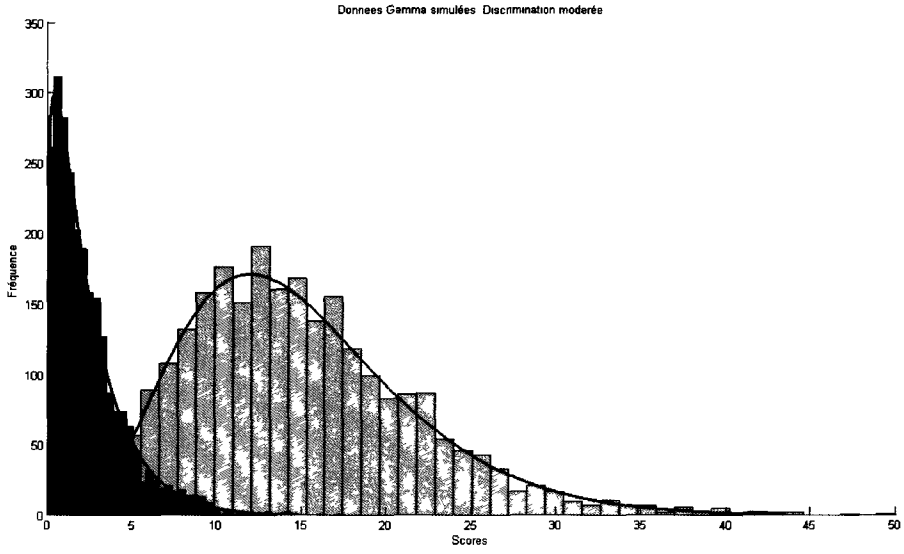

FIGURE 6.10 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim Gamma(1.25, 2)$ et $S_{D_1} \sim Gamma(5, 3)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

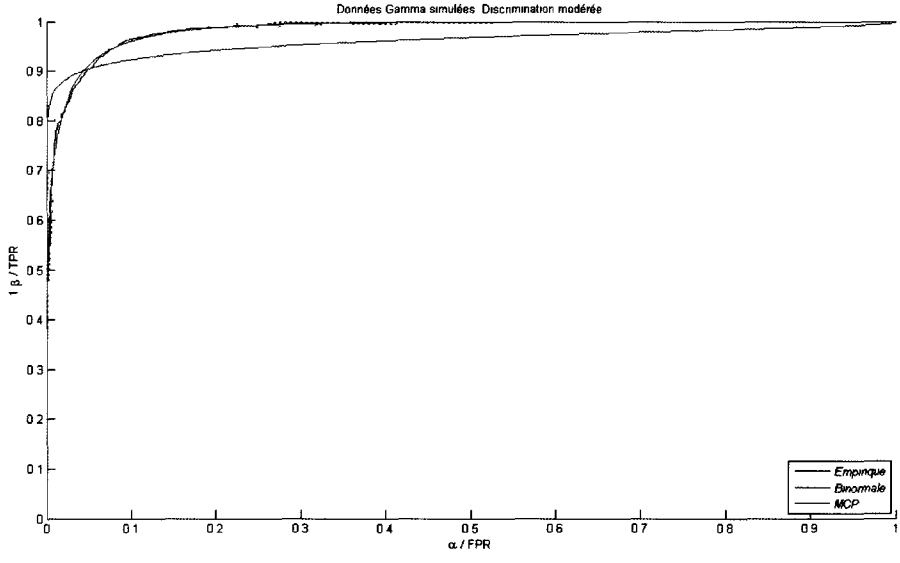

FIGURE 6.11 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim Gamma(1.25, 2)$ et $S_{D_1} \sim Gamma(5, 3)$

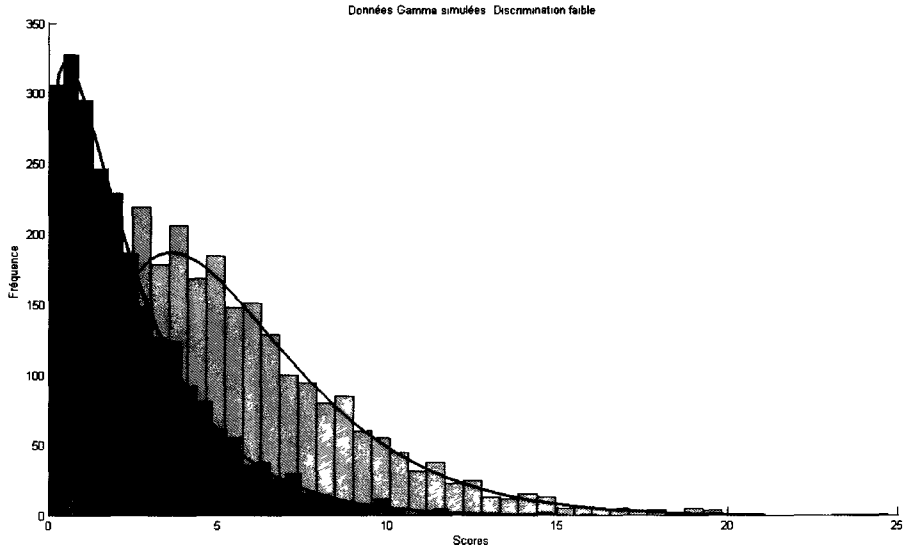

FIGURE 6.12 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim Gamma(1.25, 2)$ et $S_{D_1} \sim Gamma(2.75, 2)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

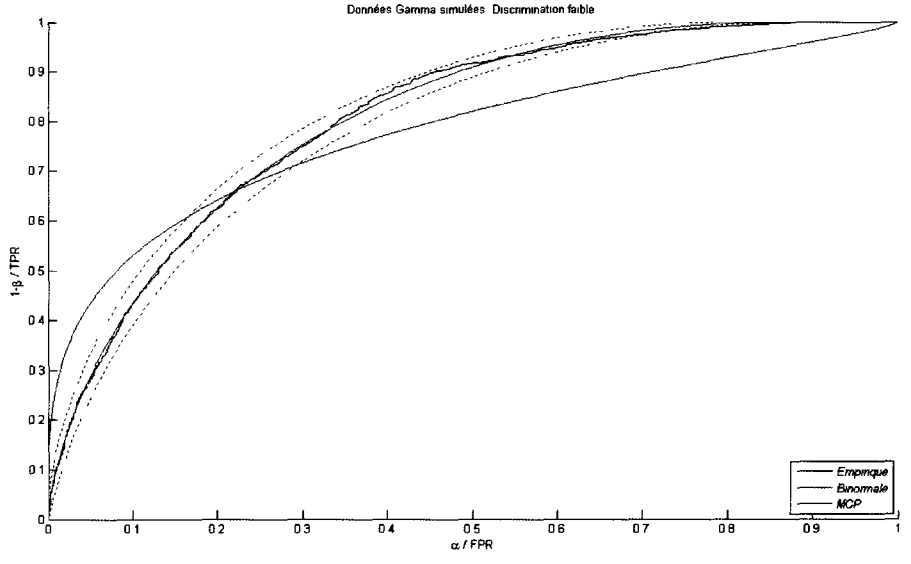

FIGURE 6.13 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim Gamma(1.25, 2)$ et $S_{D_1} \sim Gamma(2.75, 2)$

priété 3.2.1 contrairement à la courbe empirique. Encore une fois, pour le scénario où les deux populations sont complètement distinctes, voir la figure 6.9, les courbes BIN et MCP sont superposées. Selon les figures 6.11 et 6.13, il est intéressant de souligner qu'au fur à mesure que la discrimination diminue, la courbe BIN reproduit vulnérablement l'empirique. Par contre, à notre grand étonnement, la courbe MCP est *quasi* superposée par dessus la courbe empirique. En terme d'ajustement de la courbe empirique, la courbe MCP est sans équivoque un clone indéniable.

Simulation des distributions bêta

Une séparation complète des populations non-malades et malades est présente quand $S_{D_0} \sim Beta(1.25, 14)$ et $S_{D_1} \sim Beta(12, 2)$, voir l'histogramme 6.14. Selon les histogrammes 6.16 et 6.18, une discrimination modérée et faible est obtenue lorsque $S_{D_0} \sim Beta(2, 10)$ et $S_{D_1} \sim Beta(4, 2)$, et $S_{D_0} \sim Beta(5, 8)$ et $S_{D_1} \sim Beta(3.5, 1.25)$, respectivement. Selon la figure 6.15, les courbes BIN et MCP sont incontestablement superposées. Par contre, tel deviné, à la présence d'interaction entre les deux populations, la courbe BIN calque maladroitement l'empirique à l'inverse de la courbe MCP, voir les figures 6.17 et 6.19. De plus, la courbe empirique tombe entre l'intervalle de confiance à 95% de la courbe MCP. On constate que la méthode MCP ajuste mieux la courbe empirique que la BIN lorsque la distribution des scores est non-gaussienne dans un environnement d'interférence.

Simulation des distributions gaussienne-gamma

Antérieurement nous avons étudié le cas où S_{D_0} et S_{D_1} suivent la même famille de distribution. À présent, examinons le comportement des deux modèles, BIN et MCP, dans ces cas extrêmes, *i.e.* lorsque la distribution des scores ne découle pas de la même famille. Supposons que $S_{D_0} \sim N(2, 6)$. Selon l'histogramme 6.20, les deux populations sont complètement séparées quand $S_{D_1} \sim Gamma(7, 10)$. Une diminution de la dis-

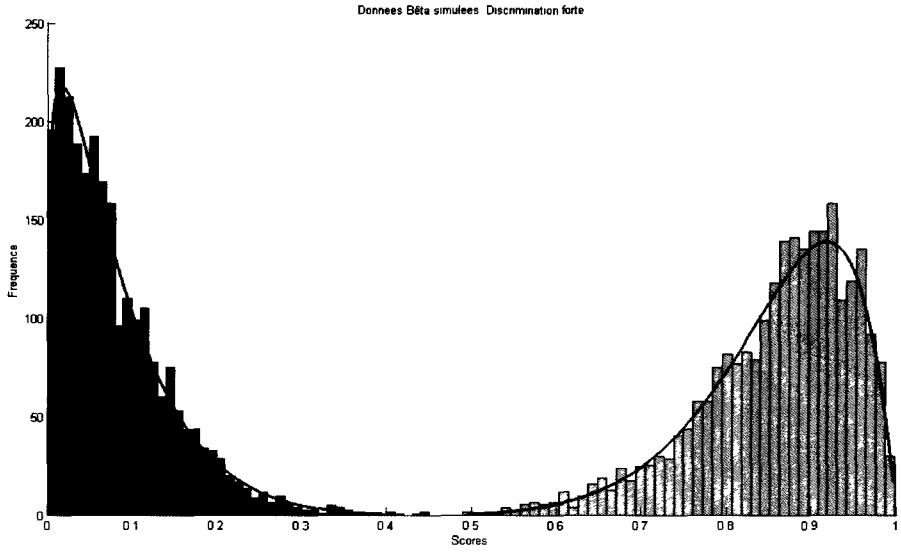

FIGURE 6.14 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim Beta(1.25, 14)$ et $S_{D_1} \sim Beta(12, 2)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

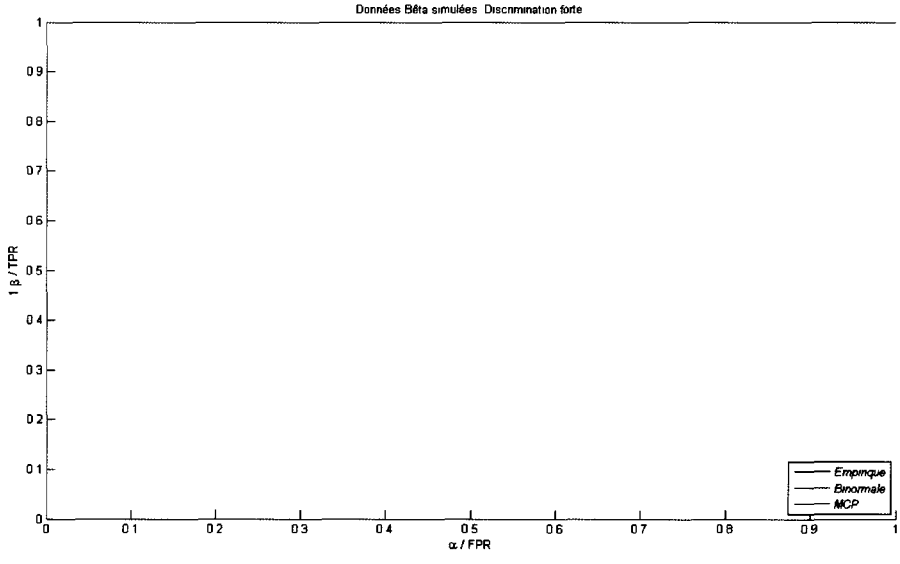

FIGURE 6.15 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim Beta(1.25, 14)$ et $S_{D_1} \sim Beta(12, 2)$

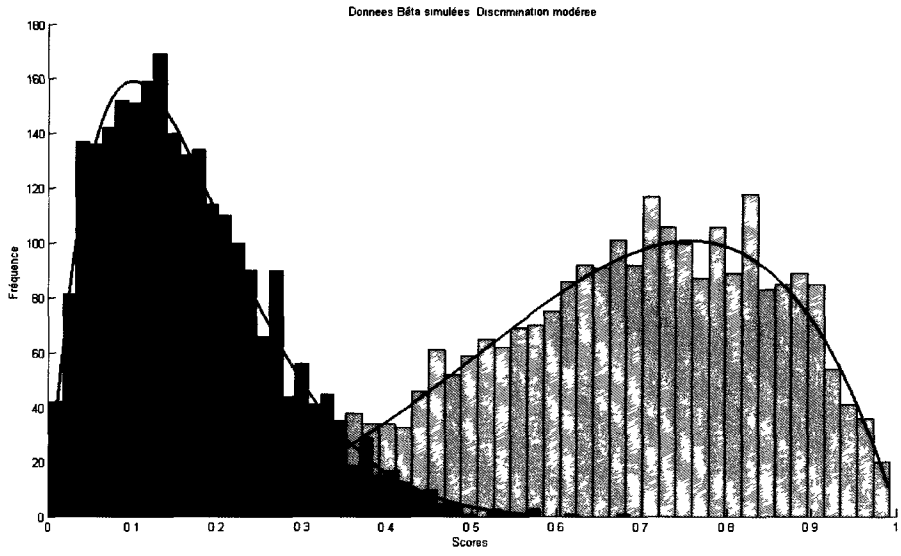

FIGURE 6.16 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim Beta(2, 10)$ et $S_{D_1} \sim Beta(4, 2)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

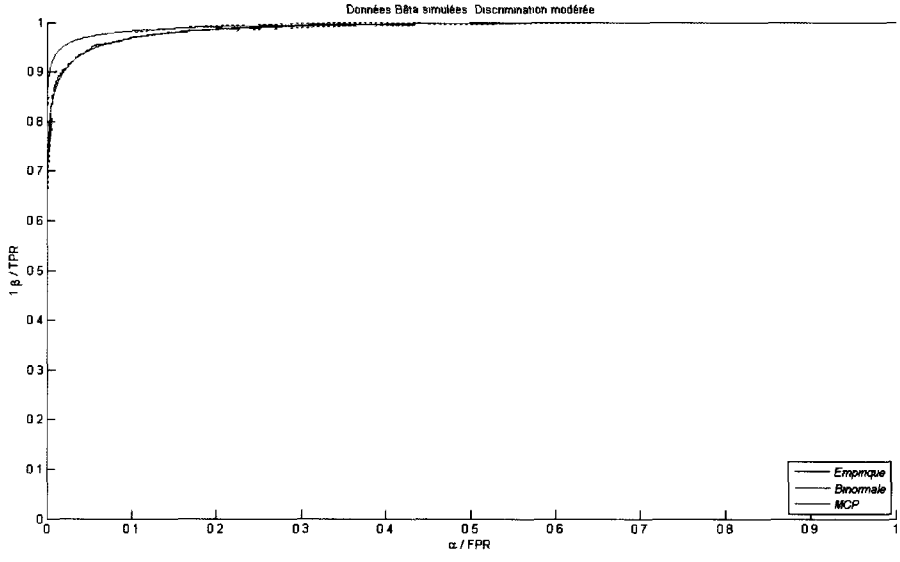

FIGURE 6.17 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim Beta(2, 10)$ et $S_{D_1} \sim Beta(4, 2)$

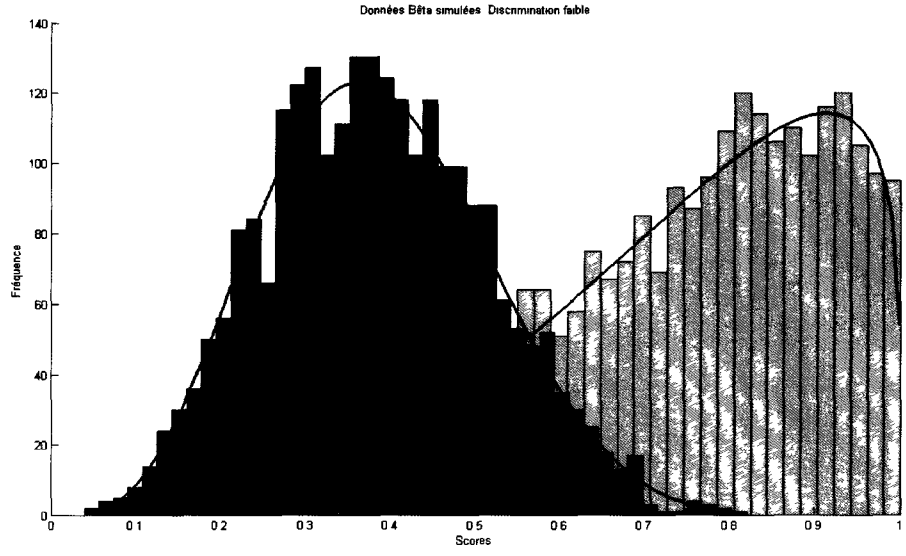

FIGURE 6 18 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim Beta(5, 8)$ et $S_{D_1} \sim Beta(3\,5, 1.25)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

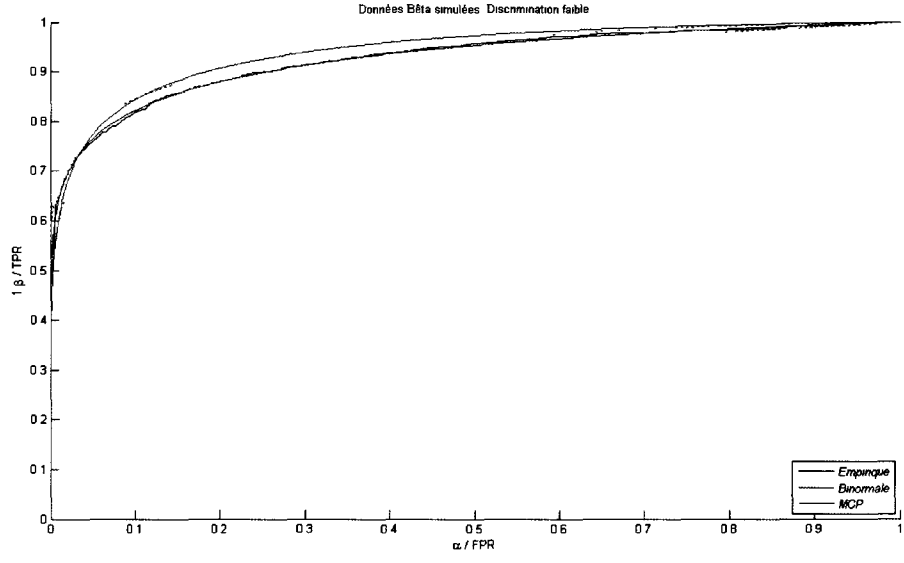

FIGURE 6.19 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim Beta(5, 8)$ et $S_{D_1} \sim Beta(3\,5, 1.25)$

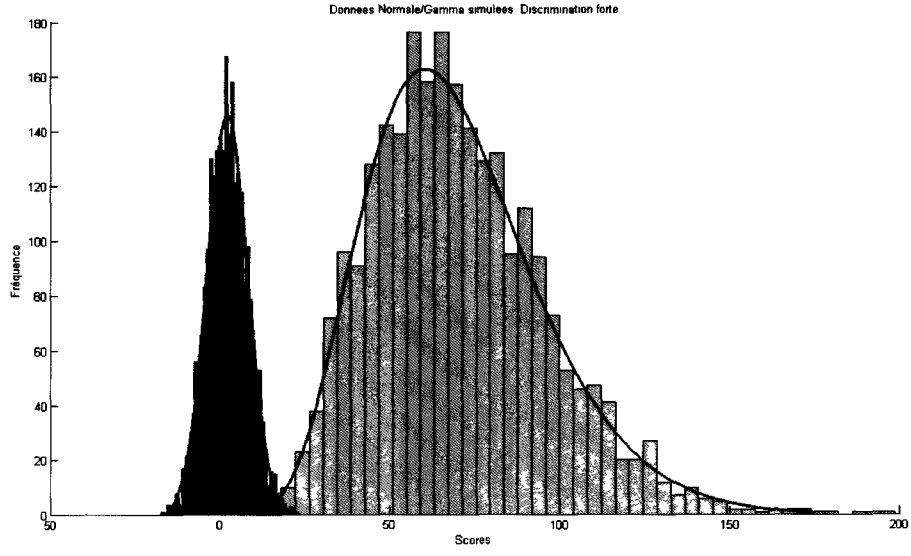

FIGURE 6.20 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim N(2,6)$ et $S_{D_1} \sim Gamma(7,10)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

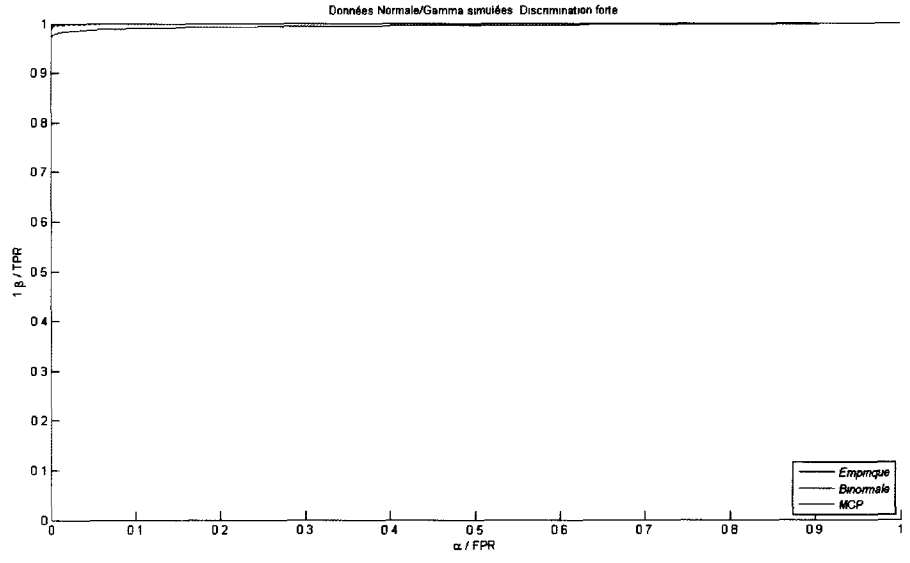

FIGURE 6.21 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim N(2,6)$ et $S_{D_1} \sim Gamma(7,10)$

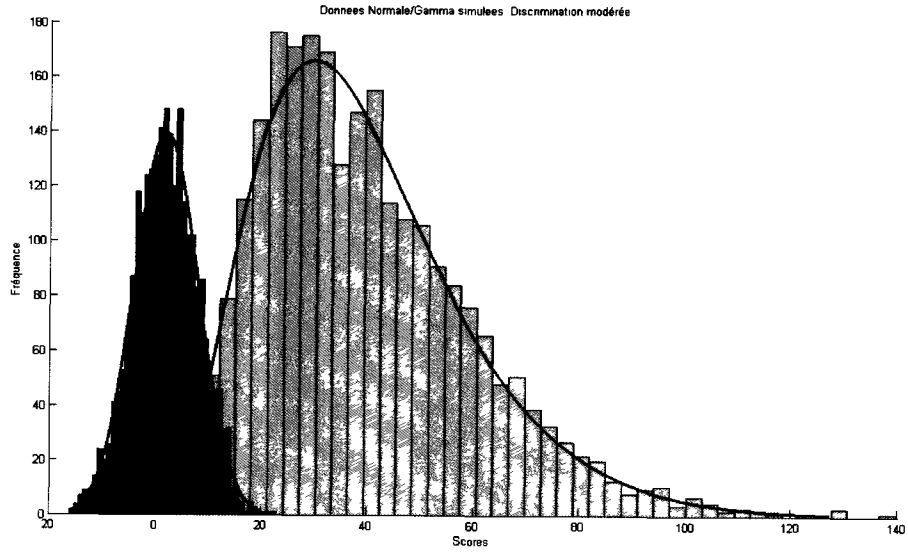

FIGURE 6.22 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim N(2,6)$ et $S_{D_1} \sim Gamma(4,10)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

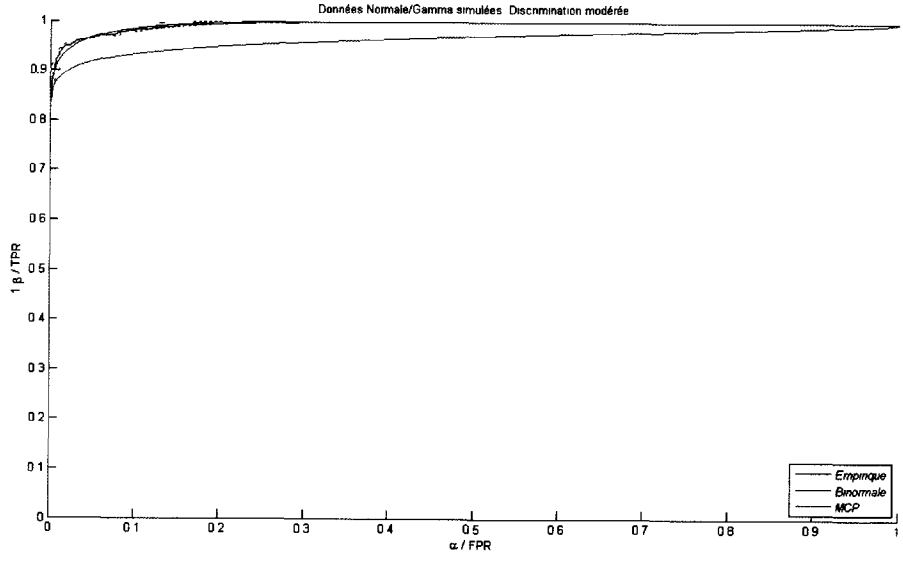

FIGURE 6.23 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim N(2,6)$ et $S_{D_1} \sim Gamma(4,10)$

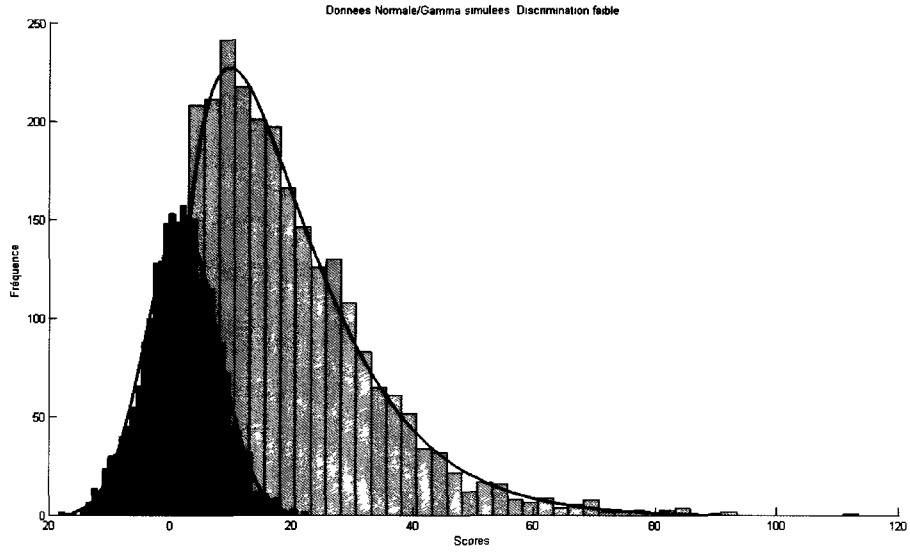

FIGURE 6.24 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim N(2,6)$ et $S_{D_1} \sim Gamma(2,10)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

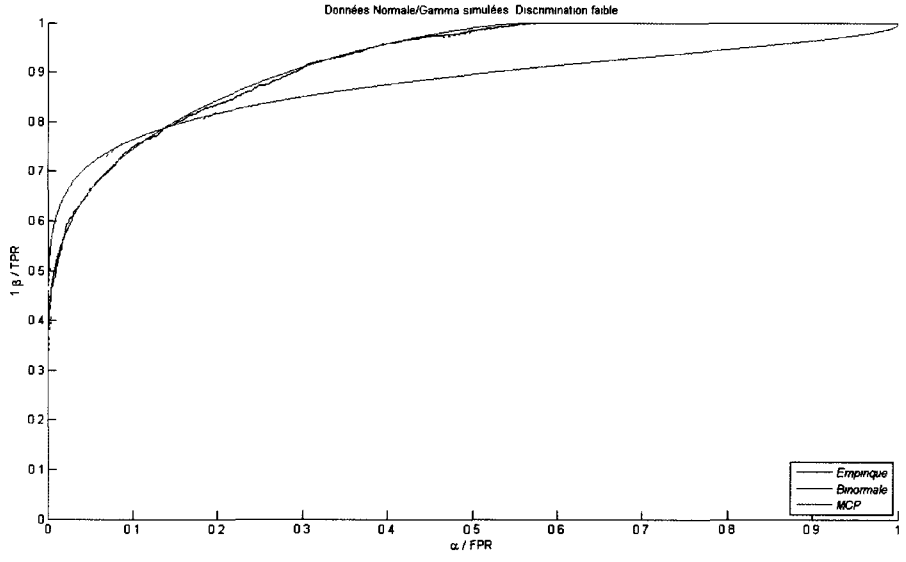

FIGURE 6.25 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim N(2,6)$ et $S_{D_1} \sim Gamma(2,10)$

crimination est observée lorsque $S_{D_1} \sim Gamma(4, 10)$ et $S_{D_1} \sim Gamma(2, 10)$, voir les histogrammes 6.22 et 6.24. Ces histogrammes nous montre que la distribution des scores de D_0 et D_1 s'apparente bien à une distribution gaussienne et gamma, respectivement. Les propriétés énoncées à la section 3.2 d'une courbe ROC sont respectées par les deux courbes théoriques. Contrairement à la méthode MCP, la BIN semble éprouver de la difficulté à répliquer l'empirique même lorsque les populations sont dispersées. Cette remarque s'applique aussi lorsque les deux distributions se chevauchent entre elles, voir les figures 6.23 et 6.25. Ainsi l'hypothèse que la méthode MCP est plus flexible que la BIN est toujours valide.

Simulation des distributions gaussienne-bêta

D'après l'histogramme 6.26, les populations D_0 et D_1 sont entièrement distinctes lorsque $S_{D_0} \sim N(-0.7, 0.25)$ et $S_{D_1} \sim Beta(5, 1)$. Une interaction modérée est observée quand $S_{D_0} \sim N(-0.3, 0.35)$ et $S_{D_1} \sim Beta(5, 1.25)$, voir l'histogramme 6.28. Lorsque $S_{D_0} \sim N(0.1, 0.25)$ et $S_{D_1} \sim Beta(3, 1.5)$, on obtient une faible discrimination comme illustrée par l'histogramme 6.30. La figure 6.27 montre que les deux courbes sont superposées. Lorsque la dispersion est flagrante, aucune conclusion ne peut être soutirer. car les deux méthodes performent aussi bien l'une que l'autre. Par contre, dans un environnement où la discrimination décroît, la méthode MCP réplique scrupuleusement l'empirique contrairement au modèle binormal, voir les figures 6.29 et 6.29. Ainsi notre hypothèse demeure vérédique.

Simulation des distributions gamma-gaussienne

Supposons que $S_{D_0} \sim Gamma(1.25, 3)$. Les histogrammes 6.32 et 6.34 représentent une discrimination forte et modérée lorsque $S_{D_1} \sim N(30, 3)$ et $S_{D_1} \sim N(25, 6)$, respectivement. Par contre, en supposant que $S_{D_0} \sim Gamma(1.5, 1)$ et $S_{D_1} \sim N(3.5, 1)$, on obtient une faible discrimination. Par la figure 6.33, on constate que les méthodes

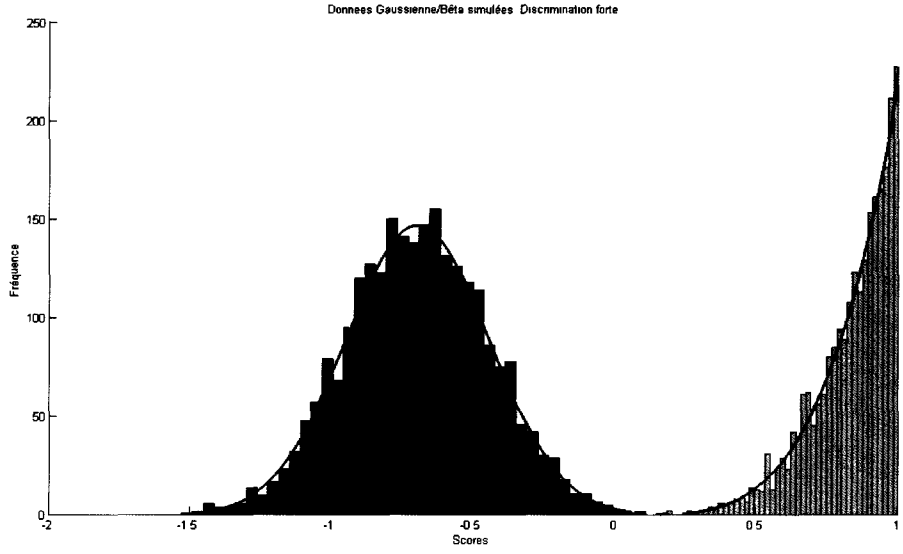

FIGURE 6.26 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim N(-0.7, 0.25)$ et $S_{D_1} \sim Beta(5, 1)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

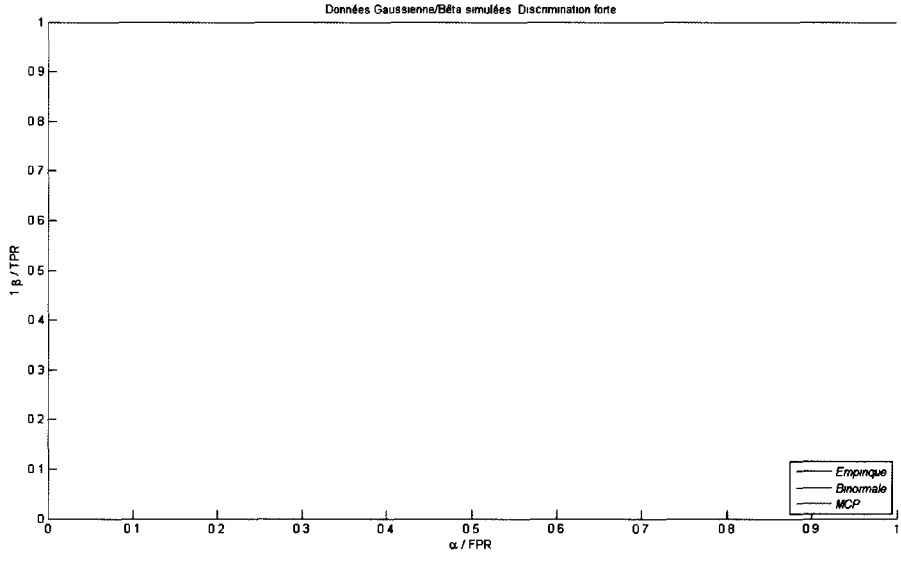

FIGURE 6.27 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim N(-0.7, 0.25)$ et $S_{D_1} \sim Beta(5, 1)$

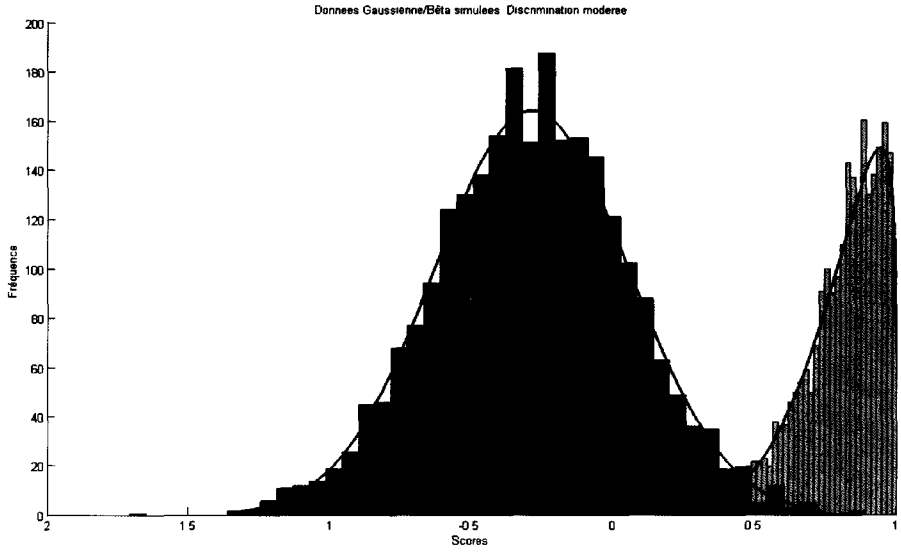

FIGURE 6.28 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim N(-0.3, 0.35)$ et $S_{D_1} \sim Beta(5, 1.25)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

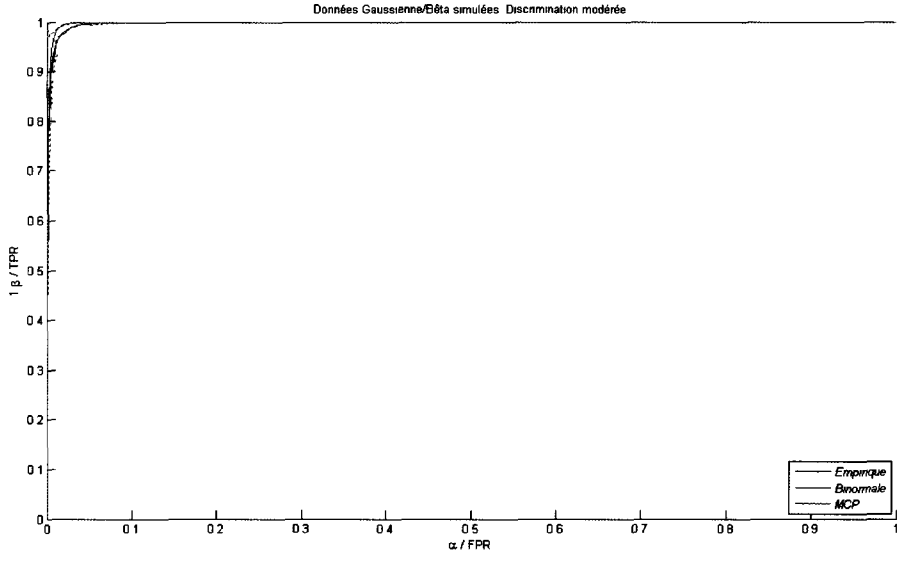

FIGURE 6.29 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim N(-0.3, 0.35)$ et $S_{D_1} \sim Beta(5, 1\ 25)$

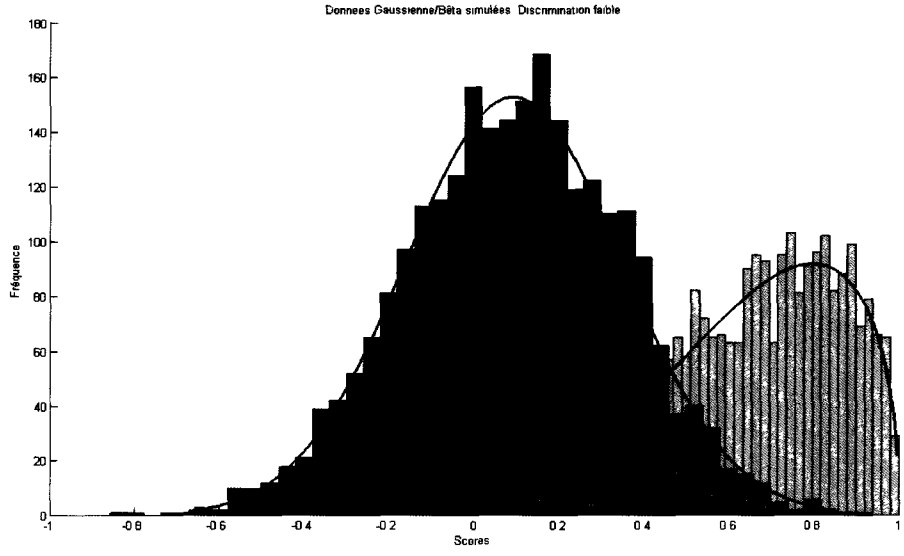

FIGURE 6.30 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim N(0.1, 0.25)$ et $S_{D_1} \sim Beta(3, 1.5)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

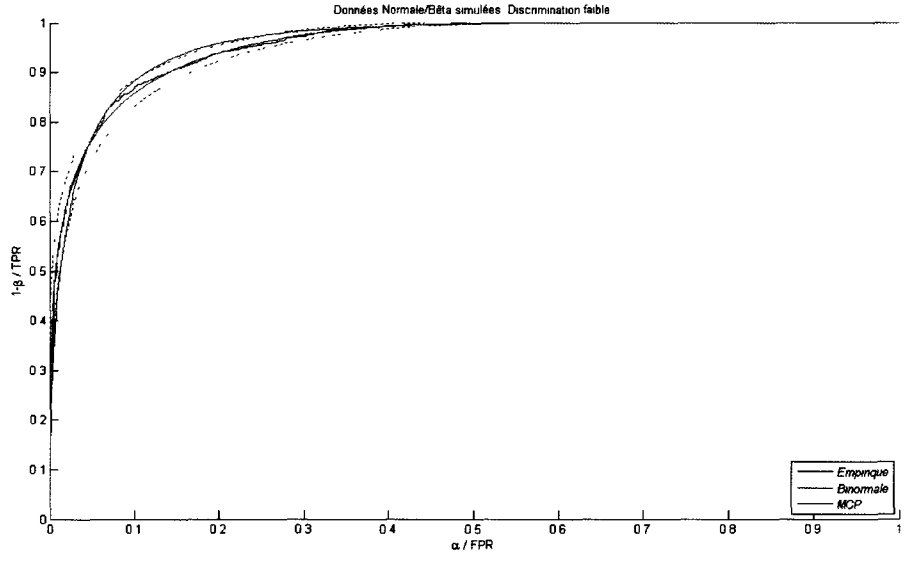

FIGURE 6.31 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim N(0.1, 0.25)$ et $S_{D_1} \sim Beta(3, 1.5)$

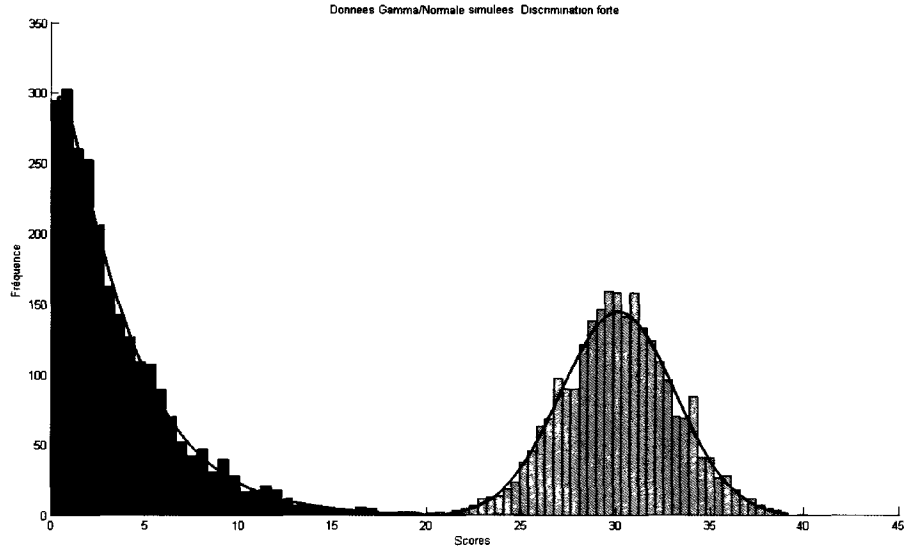

FIGURE 6.32 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim Gamma(1.25, 3)$ et $S_{D_1} \sim N(30, 3)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

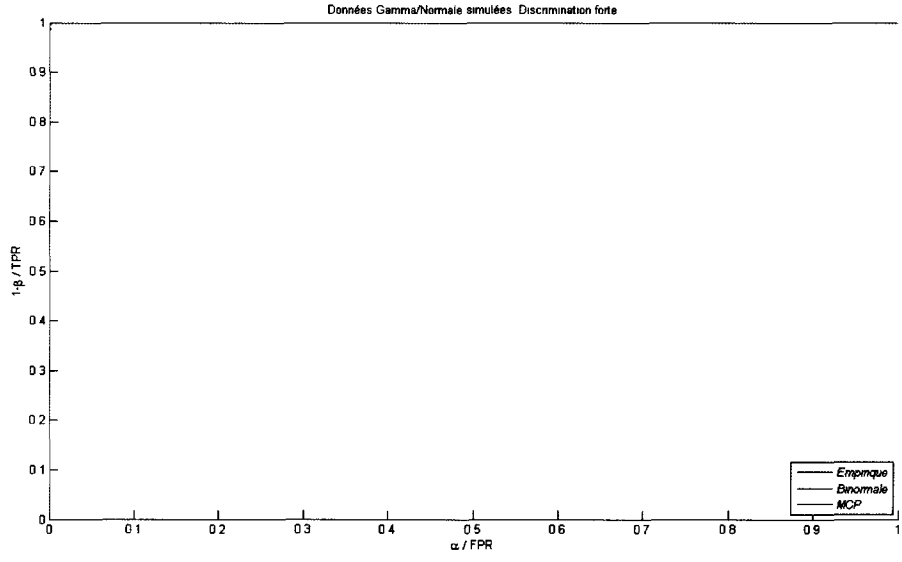

FIGURE 6.33 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim Gamma(1.25, 3)$ et $S_{D_1} \sim N(30, 3)$

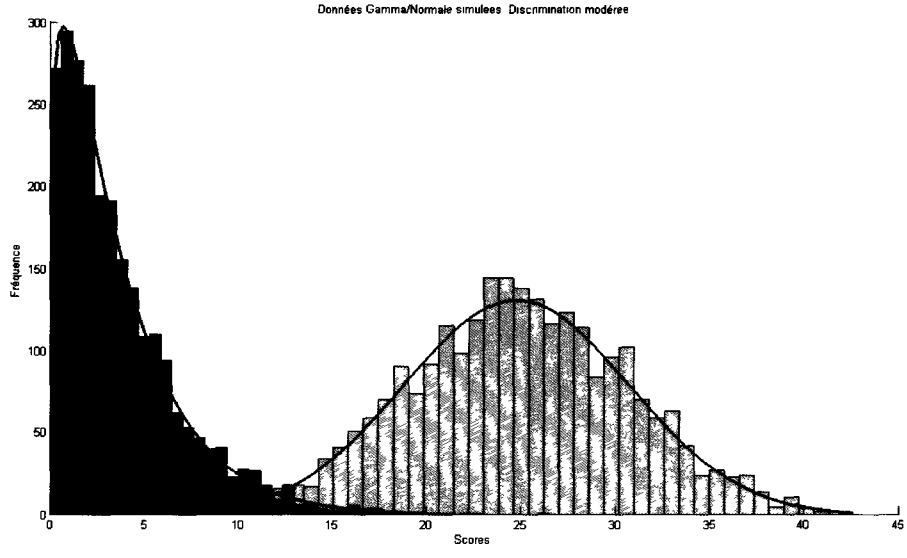

FIGURE 6.34 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim Gamma(1.25, 3)$ et $S_{D_1} \sim N(25, 6)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

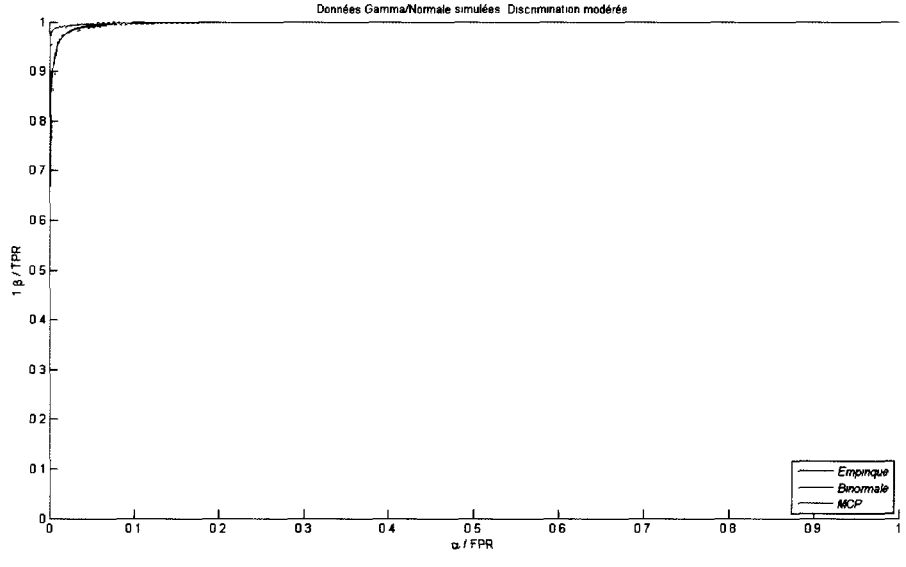

FIGURE 6.35 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim Gamma(1.25, 3)$ et $S_{D_1} \sim N(25, 6)$

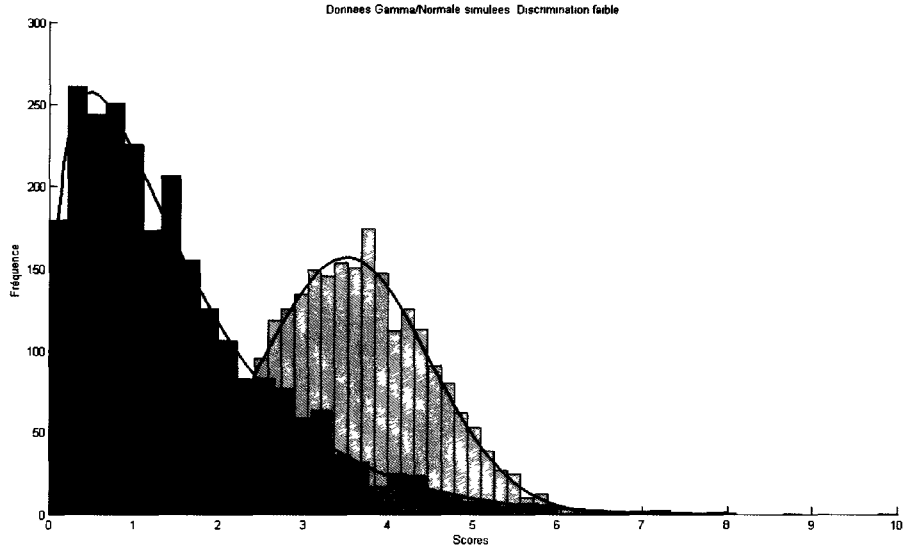

FIGURE 6.36 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim Gamma(1.5, 1)$ et $S_{D_1} \sim N(3.5, 1)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

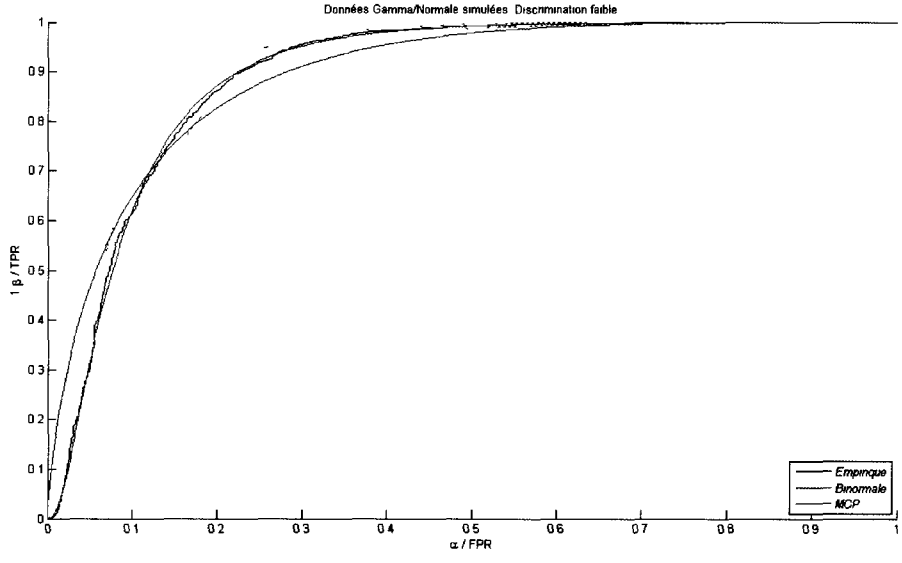

FIGURE 6.37 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim Gamma(1.5, 1)$ et $S_{D_1} \sim N(3.5, 1)$

BIN et MCP sont équivalentes, car les deux courbes sont superposées. Puisque ce cas nous est peu informatif, regardons le scénario où les populations se chevauchent entre elles. La figure 6.35 montre la courbe de la binormale se rapproche plus du coin supérieur gauche, soit le point $(0, 1)$, donc elle serait plus performante que la courbe de la MCP en terme d'un test de classification, voir la sous-section 3.1.2. Cependant, rappelons que notre objectif est d'ajuster les courbes théoriques à celles de l'empirique. Alors, on se soucie moins de la performance, mais plutôt de la réplication. Ainsi notre intuition demeure vérifiée puisque la courbe MCP excelle mieux que la courbe BIN en terme d'ajustement à l'empirique, voir la figure 6.37.

Simulation des distributions gamma-bêta

L'histogramme 6.38 montre une dispersion flagrante entre les distributions quand $S_{D_0} \sim Gamma(1, 0.1)$ et $S_{D_1} \sim Beta(12, 1)$. Par contre, les histogrammes 6.40 et 6.42 illustrent une discrimination modérée et faible quand $S_{D_0} \sim Gamma(1.1, 0.125)$ et $S_{D_1} \sim Beta(4, 1.05)$, et $S_{D_0} \sim Gamma(2, 0.2)$ et $S_{D_1} \sim Beta(7, 5)$, respectivement. Encore une fois, les courbes de la figure 6.39 nous sont peu informatives. Cependant, les figures 6.41 et 6.43 affichent clairement la supériorité de la méthode MCP à celle de la BIN. La courbe BIN bifurque significativement de l'empirique surtout dans un milieu où la discrimination est défaillante. D'une vue globale, on voit que la courbe MCP dépasse grandement la courbe BIN en terme d'ajustement à la courbe empirique.

Simulation des distributions bêta-gaussienne

Supposons que $S_{D_0} \sim Beta(3, 1)$. En variant les paramètres de la distribution de S_{D_1}, nous avons étudié le comportement des courbes sous différents niveaux de dispersion. Les histogrammes 6.44, 6.46 et 6.48 illustrent le degré de discrimination entre les populations soit : forte quand $S_{D_1} \sim N(1.75, 0.25)$, modérée quand $S_{D_1} \sim N(1.35, 0.25)$ et faible quand $S_{D_1} \sim N(1.075, 0.25)$, respectivement. Selon les figures

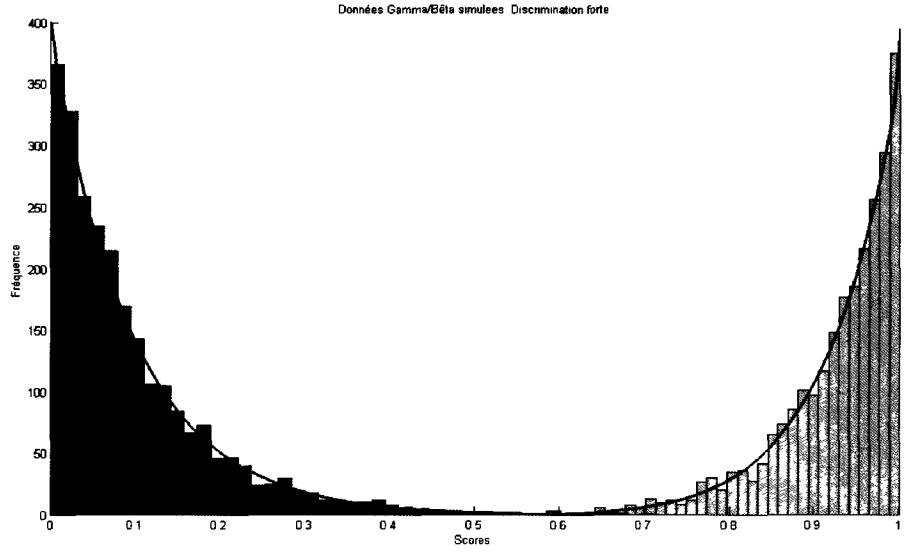

FIGURE 6.38 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim Gamma(1, 0.1)$ et $S_{D_1} \sim Beta(12, 1)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

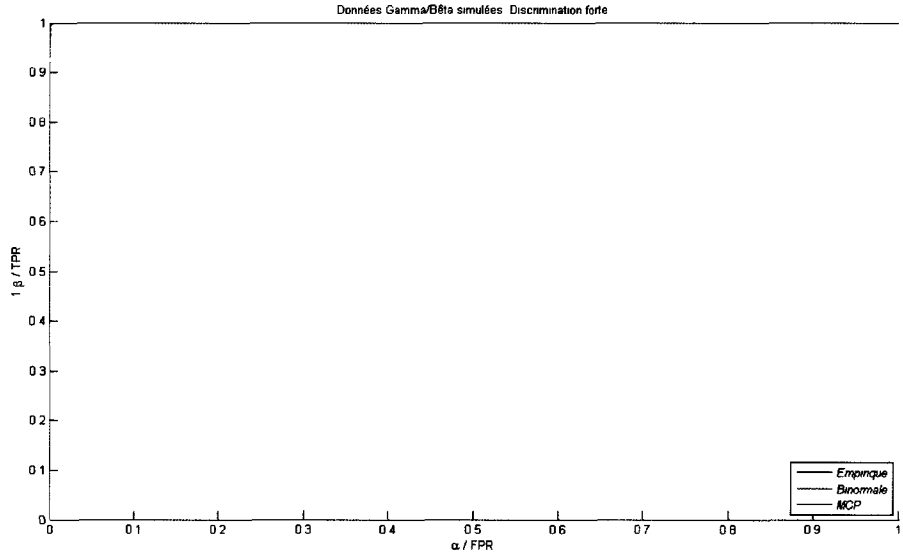

FIGURE 6.39 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim Gamma(1, 0.1)$ et $S_{D_1} \sim Beta(12, 1)$

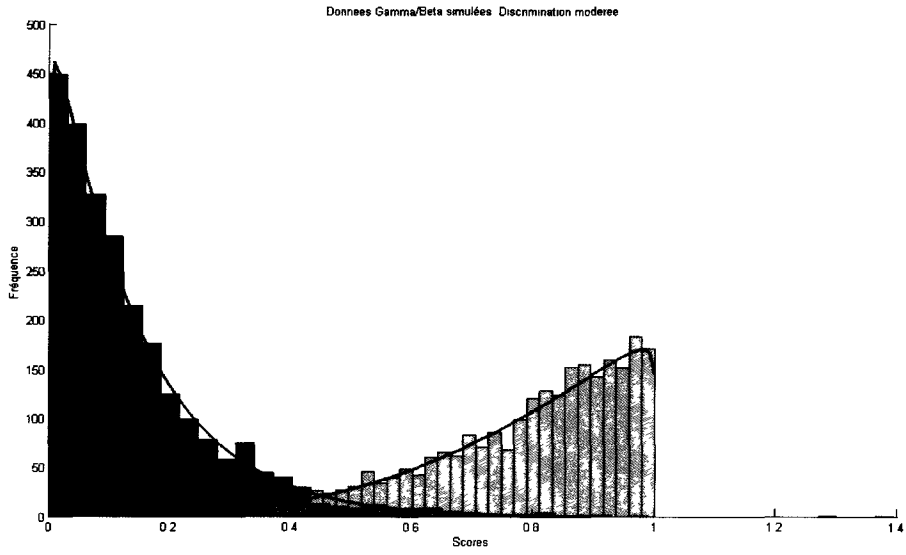

FIGURE 6.40 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim Gamma(1.1, 0.125)$ et $S_{D_1} \sim Beta(4, 1.05)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

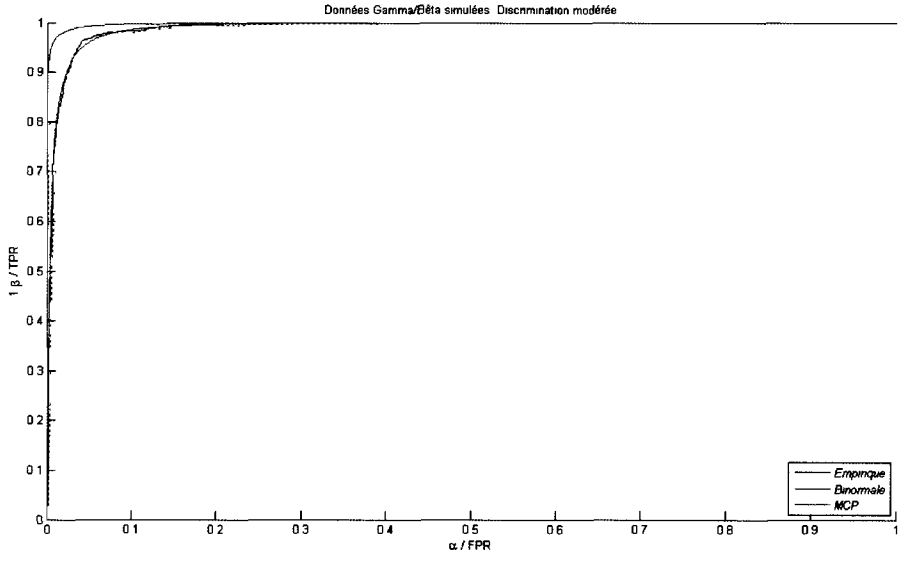

FIGURE 6.41 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim Gamma(1.1, 0.125)$ et $S_{D_1} \sim Beta(4, 1.05)$

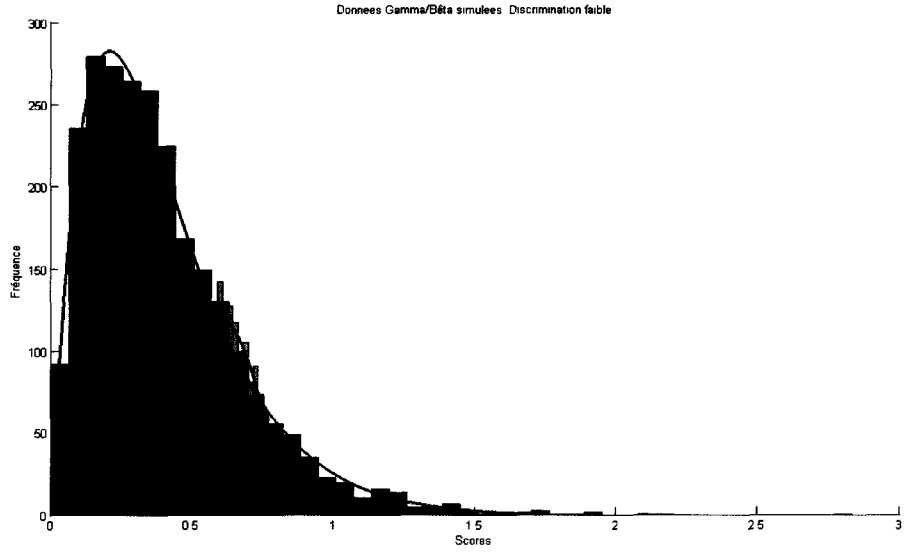

FIGURE 6.42 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim Gamma(2, 0.2)$ et $S_{D_1} \sim Beta(7, 5)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

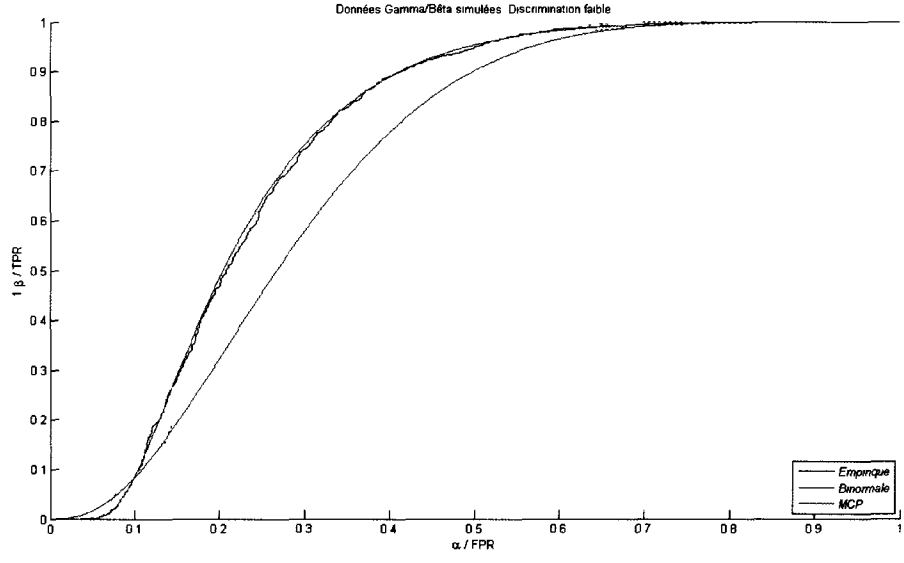

FIGURE 6.43 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim Gamma(2, 0.2)$ et $S_{D_1} \sim Beta(7, 5)$

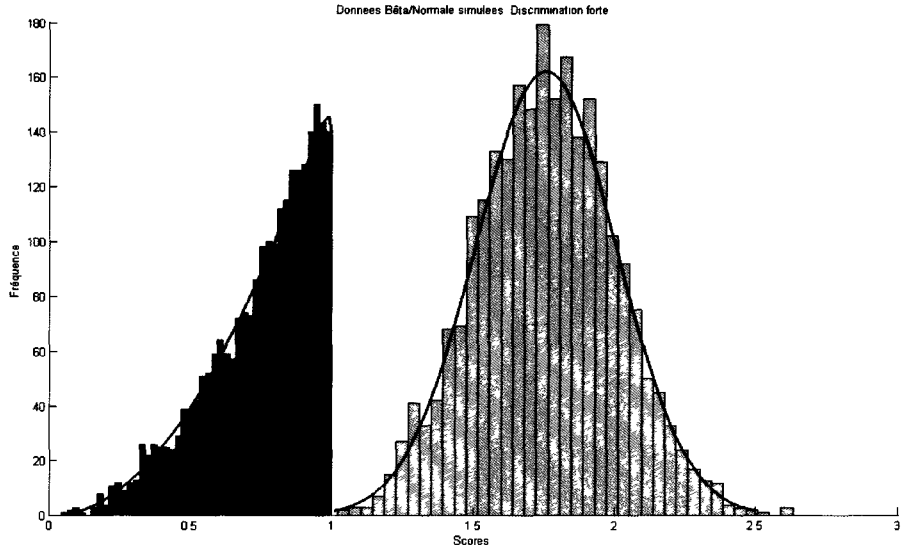

FIGURE 6.44 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim Beta(3,1)$ et $S_{D_1} \sim N(1.75, 0.25)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

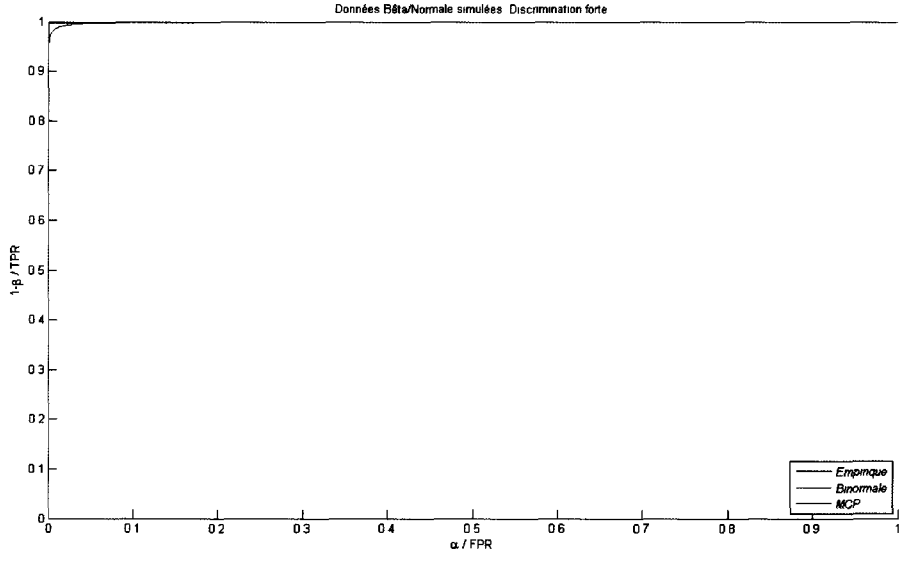

FIGURE 6.45 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim Beta(3,1)$ et $S_{D_1} \sim N(1.75, 0.25)$

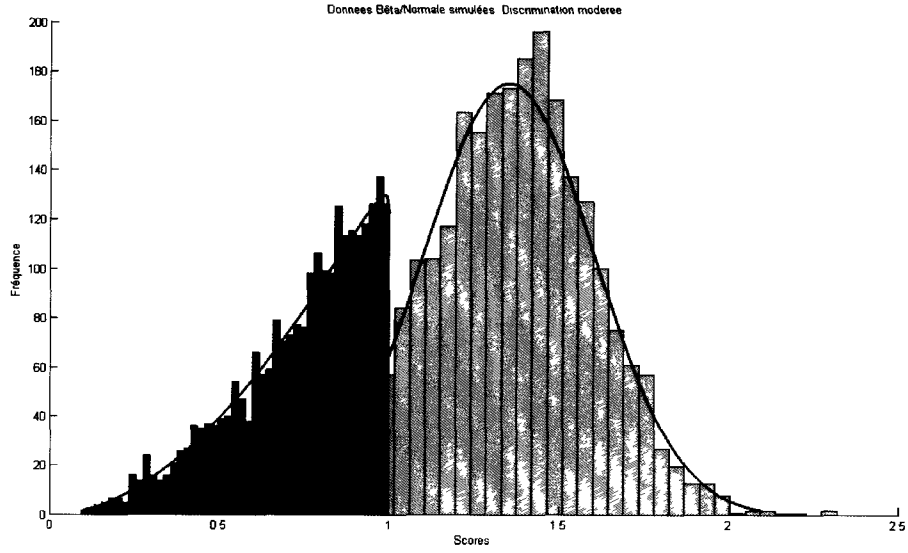

FIGURE 6.46 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim Beta(3,1)$ et $S_{D_1} \sim N(1.35, 0.25)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

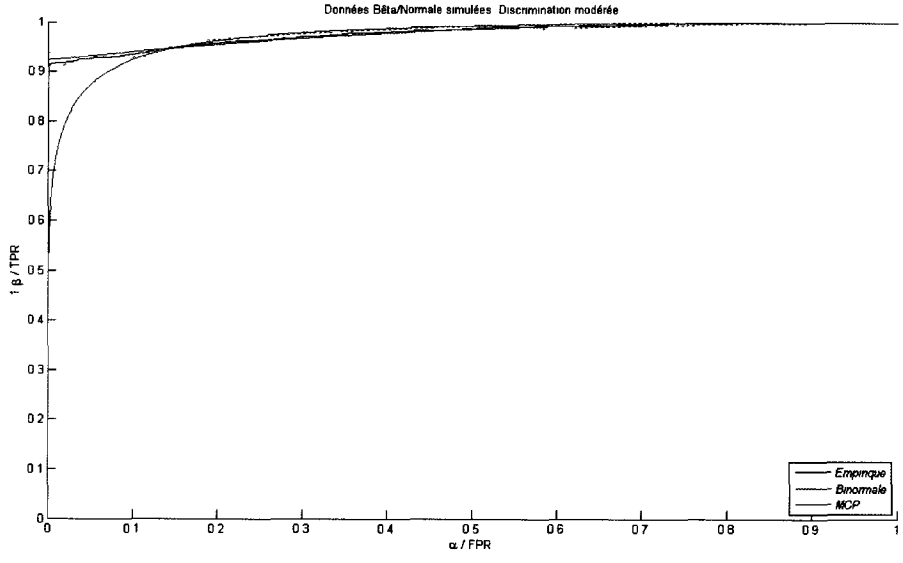

FIGURE 6.47 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim Beta(3,1)$ et $S_{D_1} \sim N(1.35, 0.25)$

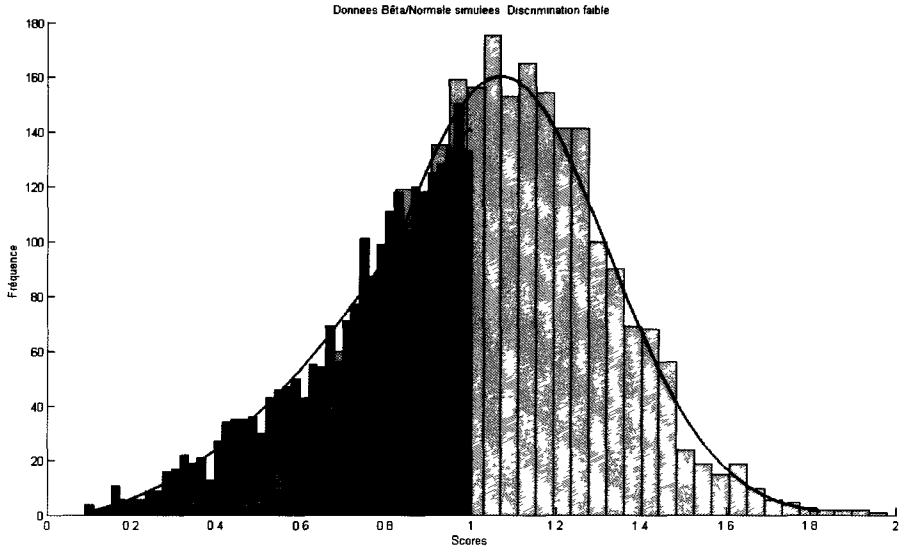

FIGURE 6.48 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim Beta(3,1)$ et $S_{D_1} \sim N(1.075, 0.25)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

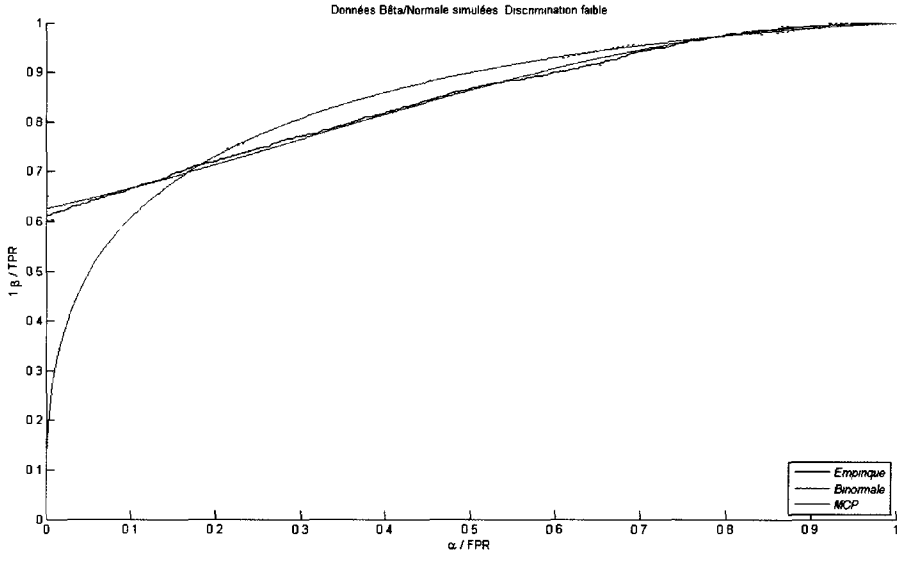

FIGURE 6.49 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim Beta(3,1)$ et $S_{D_1} \sim N(1.075, 0.25)$

6.45, 6.47 et 6.49, on remarque que plus les distributions se chevauchent entre elles, plus la courbe BIN arrive difficilement à capturer l'empirique. Contrairement à la BIN, la méthode MCP est indéniablement une copie transcendante de l'empirique.

Simulation des distributions bêta-gamma

Supposons que $S_{D_0} \sim Beta(3.5, 1)$. Les populations D_0 et D_1 sont complètement séparées quand $S_{D_1} \sim Gamma(10, 0.35)$, voir l'histogramme 6.50. Quand $S_{D_1} \sim Gamma(7, 0.35)$ et $S_{D_1} \sim Gamma(5.75, 0.35)$, on obtient une discrimination modérée et faible tel illustré par les histogrammes 6.52 et 6.54, respectivement. Selon les figures 6.51, 6.53 et 6.55, pour les trois catégories de discrimination, la courbe MCP ajuste mieux l'empirique que la courbe BIN. Sans oublier que la courbe empirique tombe dans l'intervalle de la courbe MCP. Globalement, notre hypothèse est valide.

6.1.2 Analyse quantitative

Selon l'analyse visuelle, dans un environnement où la dispersion est flagrante, six cas sur neuf sont non conclusifs, car les courbes BIN et MCP sont superposées. Par contre, les trois cas restants (gaussienne-gamma, bêta-gaussienne, bêta-gamma) favorisent la méthode MCP à celle de la BIN. Quant à la discrimination modérée et faible, huit cas sur neuf démontrent la supériorité de la méthode MCP à celle de la BIN, à l'exception des données gaussiennes où les deux méthodes sont *quasi* équivalentes. Afin de justifier davantage notre hypothèse, une analyse numérique reste tout de même nécessaire. Pour ce faire, les six tables à venir présentent les résultats des trois mesures de performance décrites au chapitre 5 selon le degré de discrimination.

Avant de débuter l'analyse, revisons quelques notions des mesures de performance. Pour la mesure MSE, selon la définition 5.3.1, nous avons pris $K = 1000$ découpes sur l'axe des FPR. Notons qu'avec la méthode MCP, nous aurons $M = 1000$ courbes MCP. Ainsi la mesure MSE représente une moyenne de MSE des courbes

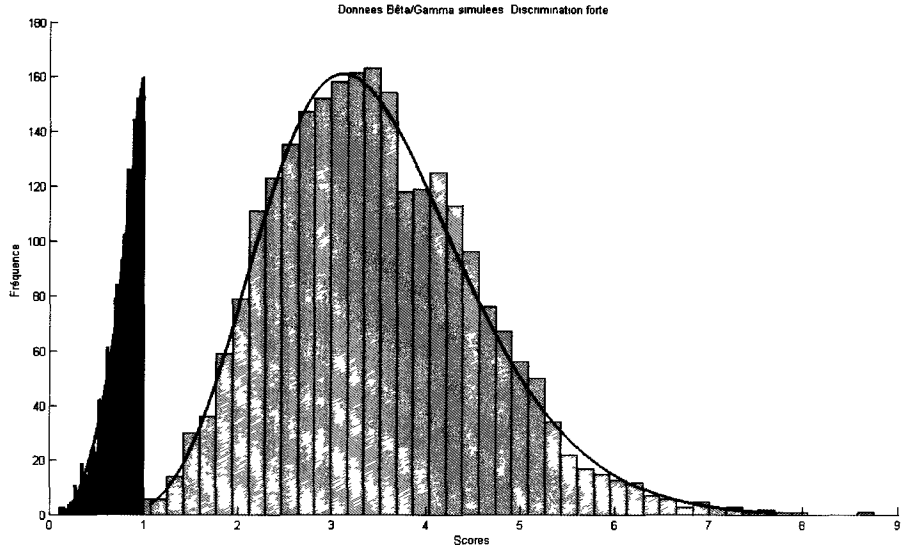

FIGURE 6.50 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim Beta(3.5, 1)$ et $S_{D_1} \sim Gamma(10, 0.35)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

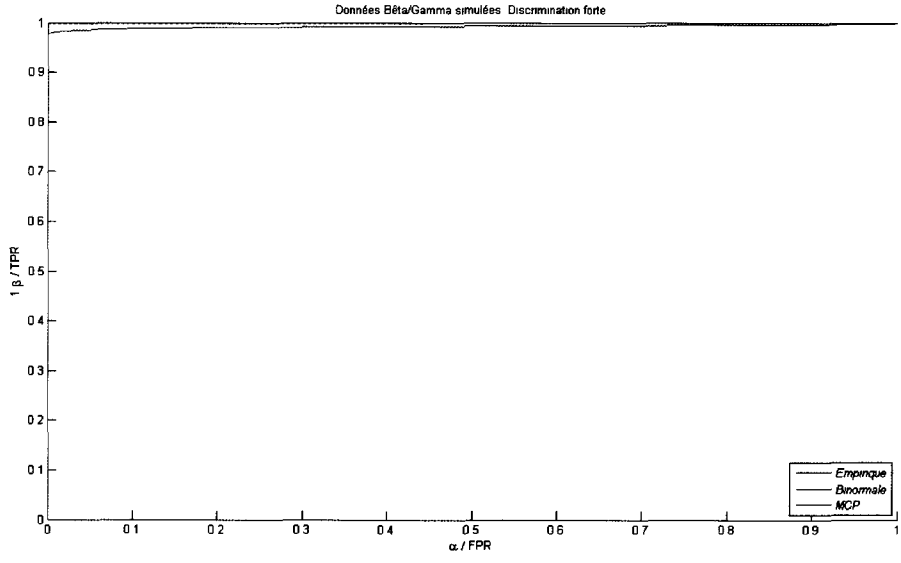

FIGURE 6.51 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim Beta(3.5, 1)$ et $S_{D_1} \sim Gamma(10, 0.35)$

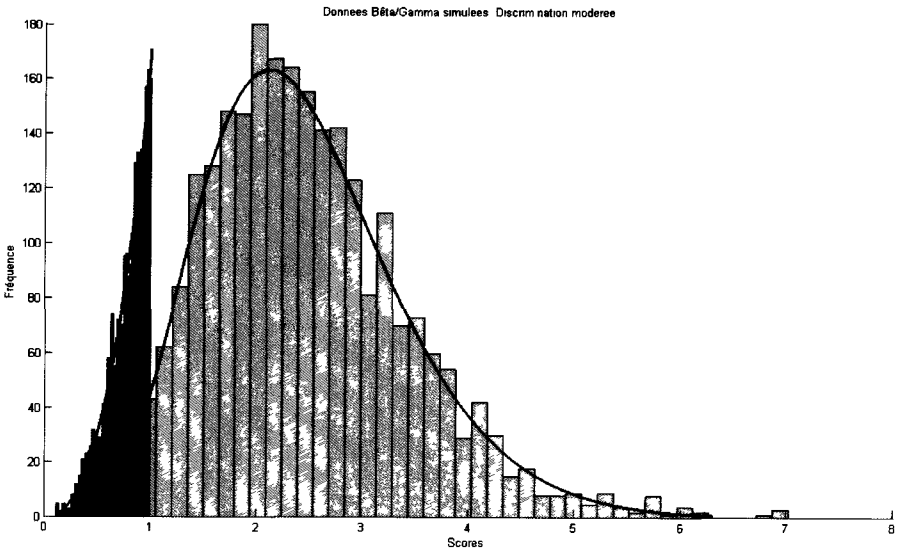

FIGURE 6 52 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim Beta(3\,5, 1)$ et $S_{D_1} \sim Gamma(7, 0\,35)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

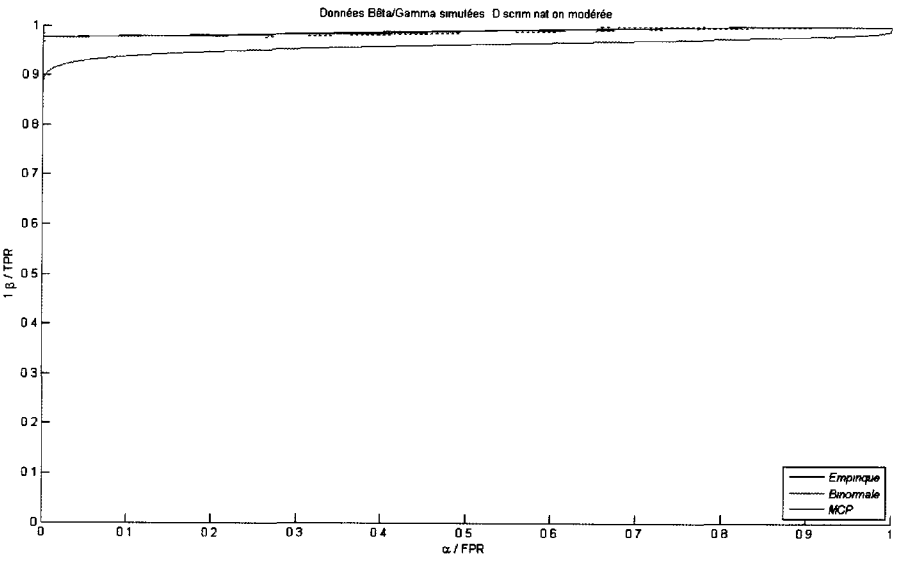

FIGURE 6 53 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim Beta(3\,5, 1)$ et $S_{D_1} \sim Gamma(7, 0\,35)$

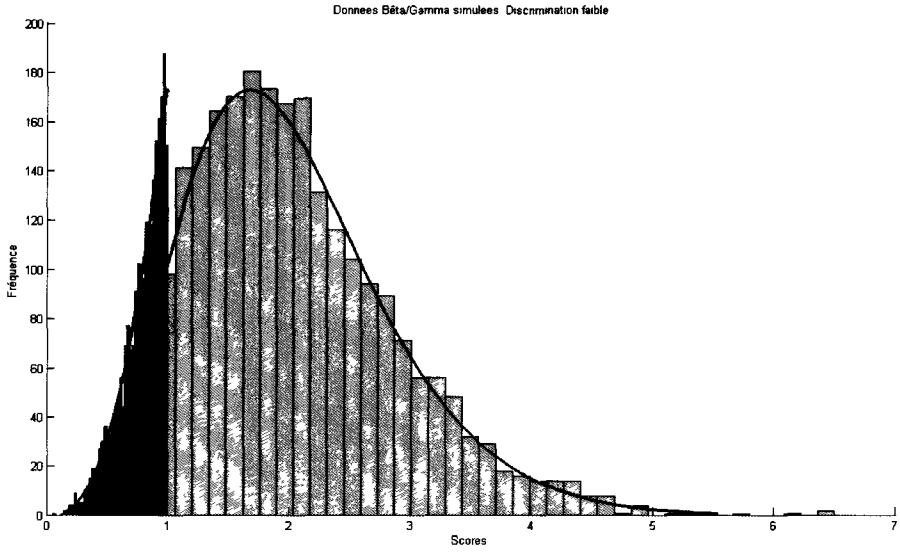

FIGURE 6.54 Histogramme des populations non-malades (en bleu) et malades (en rose) où $S_{D_0} \sim Beta(3.5, 1)$ et $S_{D_1} \sim Gamma(5.75, 0.35)$, respectivement avec $n_{D_0} = n_{D_1} = 2500$

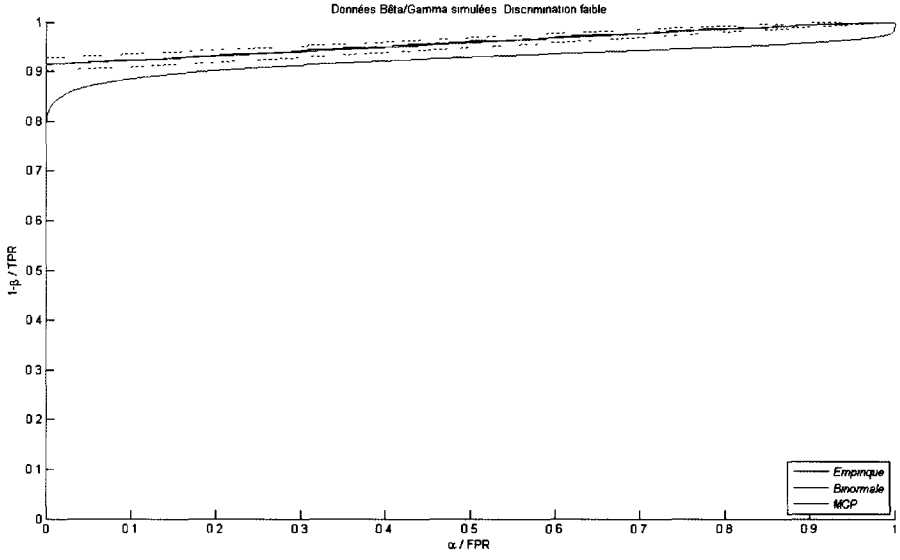

FIGURE 6.55 Courbes ROC empiriques, binormale et MCP avec un intervalle de confiance à 95% lorsque $S_{D_0} \sim Beta(3.5, 1)$ et $S_{D_1} \sim Gamma(5.75, 0.35)$

MCP comme expliqué dans la remarque 5.3.2. Pour le modèle binormal, l'AUC se calcule directement avec l'équation (5.1.3). Par contre, le modèle MCP utilise la méthode du trapèze et la statistique U de Mann-Whitney, voir la sous-section 5.1.4, puisque la distribution de Pearson n'a pas de forme fermée. Les méthodes d'estimation d'AUC est aussi utilisées pour l'empirique, car on ne connaît pas la nature distributionnelle des données. Pour le calcul du pAUC, nous avons divisé l'axe des FPR en $L = 10$ sous-intervalles ou *buckets*. Par la suite, nous avons calculé le MSEpAUC, voir l'équation (5.3.2). Il est à noter que le pAUC se calcule avec la méthode du trapèze, voir la section 5.2. De plus, soulignons que l'écart-type est obtenue grâce aux résultats de simulation de Monte-Carlo, voir remarque 4.2.9.

À présent, commençons notre analyse pour le cas où $S_{D_0} \sim N(\mu_0, \sigma_0^2)$ et $S_{D_1} \sim N(\mu_1, \sigma_1^2)$. Rappelons que graphiquement, les deux courbes sont réciproques, voir les figures 6.3, 6.5 et 6.7. Lorsque le degré de discrimination est élevé, la valeur de MSE pour la méthode BIN est approximativement six fois inférieure ($1.02 \cdot 10^{-6}/1.02 \cdot 10^{-7}$) à celle de la MCP, voir la table 6.1. Selon les tables 6.3 et 6.5, lorsqu'il y a une présence de chevauchement cet écart diminue, soit trois fois plus petit à celle de la MCP. Pour une discrimination faible, on remarque que l'AUC de la MCP est équivalente ou plus proche de l'empirique que la BIN. Quant à la mesure du MSEpAUC, la BIN est entre une à cinq fois plus petite que la MCP. Tel constaté, cette différence des mesures MSE et MSEpAUC, entre les deux modèles, peut être qualifiée négligeable. Ces résultats ne sont guère étonnants puisqu'on estime des données gaussiennes avec une distribution binormale. Malgé cette victoire mineure de la méthode BIN, la MCP demeure un candidat exceptionnel.

Dans une situation de dispersion complète entre $S_{D_0} \sim Beta(\alpha_0, \beta_0)$ et $S_{D_1} \sim Beta(\alpha_1, \beta_1)$, le modèle BIN est approximativement 1.5 fois inférieur (1.86/1.12) que celui du MCP, au niveau du MSE. Cependant, on obtient une égalité pour la valeur d'AUC, d'après la table 6.1. Notons qu'ici l'$AUC = 1$, ce qui signifie que la courbe ROC est idéale, donc on a une séparation totale des deux populations, voir l'histo-

gramme 6.14. Il est flagrant que le MSEpAUC favorise la méthode BIN à celle de la MCP, soit 90 fois inférieur ($1.24 \cdot 10^{-16}/1.12 \cdot 10^{-14}$). Lorsque la méthode BIN est exposée à un recouvrement entre les deux populations, elle succombe à la MCP dans les trois mesures de performance. Contrairement à la BIN, la méthode MCP arrive à mieux capturer l'empirique grâce à la flexibilité du système de distributions de Pearson.

Pour une séparation entière entre $S_{D_0} \sim N(\mu_0, \sigma_0^2)$ et $S_{D_1} \sim Beta(\alpha_1, \beta_1)$, la mesure MSE avantage la méthode MCP à celle de la BIN. Une égalité de l'AUC est observée entre les deux parties. Par contre, le modèle BIN est privilégié par le MSEpAUC. Inversement, dans un milieu de discrimination modérée, le MSE défavorise la méthode MCP, mais les mesures AUC et MSEpAUC sont en sa faveur. Rappelons que notre objectif principal est de répliquer minutieusement la courbe empirique à partir des méthodes BIN et MCP. Non seulement les résultats numériques nous soient substantiels, mais aussi le comportement des courbes. Même si la méthode BIN devance négligemment la MCP en terme numérique, mais graphiquement, selon le figure 6.29, elle la triomphe. De plus, lorsque la discrimination est faible, les trois mesures de performance favorisent la MCP que la BIN.

Supposons $S_{D_0} \sim Gamma(k_0, \theta_0)$. Pour le cas où $S_{D_1} \sim N(\mu_1, \sigma_1^2)$ et $S_{D_1} \sim Beta(\alpha_1, \beta_1)$, dans un environnement où la discrimination est observée, les mesures MSE et MSEpAUC favorisent la méthode BIN à celle de la MCP, d'après la table 6.2. Quant à l'AUC, la différence entre les deux méthodes est insignifiante. Par contre, lorsque les populations commencent à se chevaucher, la MCP excelle grandement la binormale, selon les tables 6.4 et 6.6. En pratique, il est peu fréquent d'observer une dispersion complète entre les groupes de patients. Le but d'une courbe ROC est de vérifier l'exactitude et la précision d'un test à classer les sujets non-malades des malades. Alors on s'intéresse davantage au comportement des courbes BIN et MCP dans un milieu où la discrimination est restreinte. Donc, encore une fois, la méthode MCP est certes supérieure à celle de la BIN.

Pour les autres combinaisons, la méthode MCP dépasse inconstestablement la binormale pour les trois mesures de performance dans tous les types de discrimination : forte, modérée et faible. On remarque que la binormale devance légèrement la MCP seulement dans un milieu où la discrimination est forte. Tel mentionné précédemment, cet environnement nous charme peu. Graphiquement, la méthode MCP domine clairement la BIN dans toutes les circonstances. Numériquement, plus le degré de discrimination diminue, plus la MCP gagne de l'avancement par rapport à la BIN. En somme, la méthode MCP se révèle plus flexible et ajuste mieux la courbe empirique que la BIN.

TABLE 6.1 Discrimination forte : mesures de performances des données simulées (l'écart-type)

S_{D_0}/S_{D_1}	*Modèle du courbe ROC*	**MSE**	**AUC** Trapèze	**AUC** MW	**MSEpAUC**
Normale	Empirique	-	0.9998	0.9998	-
	BIN	**1.67E-07**	0.9998		**3.91E-10**
	MCP	1.02E-06	0.9998	0.9998	5.67E-10
		(2.34E-06)	(9.53E-05)	(1.12E-04)	(1.01E-09)
Gamma	Empirique	-	1	1	-
	BIN	3.32E-06	0.9984		3.25E-08
	MCP	**5.41E-08**	**1**	**0.9999**	**1.76E-10**
		(9.71E-08)	(7.01E-05)	(7.09E-05)	(5.04E-10)
Bêta	Empirique	-	1	1	-
	BIN	**1.12E-10**	1		**1.24E-16**
	MCP	1.86E-10	1	1	1.12E-14
		(2.05E-09)	(2.24E-06)	(3.25E-06)	(1E-13)
Normale / Gamma	Empirique	-	1	1	-
	BIN	4.17E-05	0.9946		4.05E-07
	MCP	**1.57E-07**	**0.9999**	**0.9999**	**9.96E-10**
		(1.88E-07)	(1.03E-04)	(1.03E-04)	(1.6E-09)
Normale / Bêta	Empirique	-	1	1	-
	BIN	1.09E-10	1		**1.07E-16**
	MCP	**2.83E-11**	1	1	1.37E-15
		(1.81E-10)	(9E-07)	(1.3E-06)	(1.28E-14)

TABLE 6.2 Discrimination forte : mesures de performances des données simulées (l'écart-type) (cont.)

S_{D_0}/S_{D_1}	*Modèle du courbe ROC*	**MSE**	**AUC** Trapèze	**AUC** MW	**MSEpAUC**
	Empirique	-	1	1	-
Gamma /	BIN	**3.04E-09**	1		**9.53E-14**
Normale	MCP	1.63E-07	1	0.9999	7.86E-12
		(1.3E-06)	(6.94E-05)	(9.91E-05)	(7.09E-11)
	Empirique	-	1	1	-
Gamma /	BIN	**6.29E-08**	**1**		**1.69E-11**
Bêta	MCP	3.86E-07	0.9999	0.9999	2.32E-11
		(3.01E-06)	(9.3E-05)	(1.32E-04)	(1.32E-10)
	Empirique	-	0.9996	0.9996	-
Bêta /	BIN	6.59E-06	0.9993		1.06E-08
Normale	MCP	**2.85E-07**	**0.9997**	**0.9997**	**2.61E-09**
		(1.79E-07)	(2.42E-04)	(2.42E-04)	(1.79E-09)
	Empirique	-	0.9997	0.9997	-
Bêta /	BIN	6.29E-05	0.9924		6.17E-07
Gamma	MCP	**4.68E-07**	**0.9994**	**0.9994**	**4.46E-09**
		(6.34E-07)	(4.32E-04)	(4.33E-04)	(6.24E-09)

TABLE 6.3 Discrimination modérée : mesures de performances des données simulées (l'écart-type)

S_{D_0}/S_{D_1}	*Modèle du courbe ROC*	**MSE**	**AUC** Trapèze	MW	**MSEpAUC**
Normale	Empirique	-	0.9238	0.9236	-
	BIN	**4.27-05**	**0.0.9231**		**1.47E-07**
	MCP	1.48E-04	**0.9244**	0.9241	8.17E-07
		(9.14E-05)	(0.0047)	(0.0047)	(7.88E-07)
Gamma	Empirique	-	0.9842	0.9840	-
	BIN	0.0015	0.9610		9.25E-06
	MCP	**6.01E-05**	**0.9840**	**0.9838**	**1.93E-07**
		(3.92E-05)	(0.0017)	(0.0017)	(2.32E-07)
Bêta	Empirique	-	0.9884	0.9883	-
	BIN	1.95E-04	0.9931		1.29E-06
	MCP	**2.11E-05**	**0.9885**	**0.9883**	**8.51E-08**
		(1.48E-05)	(0.0014)	(0.0014)	(1.05E-07)
Normale / Gamma	Empirique	-	0.9944	0.9943	-
	BIN	0.0011	0.9648		1.05E-05
	MCP	**1.10E-05**	**0.9949**	**0.9948**	**4.77E-08**
		(6.5E-06)	(7.32E-04)	(7.35E-04)	(4.39E-08)
Normale / Bêta	Empirique	-	0.9977	0.9975	-
	BIN	**2.02E-05**	0.9983		4.67E-08
	MCP	5.35E-05	**0.9978**	**0.9976**	**2.65E-08**
		(8.60E-05)	(6.43E-04)	(7.01E-04)	(4.34E-08)

TABLE 6.4 Discrimination modérée : mesures de performances des données simulées (l'écart-type) (cont.)

S_{D_0}/S_{D_1}	*Modèle du courbe ROC*	**MSE**	**AUC** Trapèze	**AUC** MW	**MSEpAUC**
	Empirique	-	0.9978	0.9977	-
Gamma /	BIN	8.81E-05	0.9991		1.42E-07
Normale	MCP	**2.04E-05**	**0.9978**	**0.9976**	**1.69E-08**
		(2.36E-05)	(5.17E-04)	(5.46E-04)	(2.28E-08)
	Empirique	-	0.9893	0.9891	-
Gamma /	BIN	0.002	0.9982		5.25E-06
Bêta	MCP	**1.91E-04**	**0.9893**	**0.989**	**2.25E-07**
		(1.70E-04)	(0.0017)	(0.0017)	(3.31E-07)
	Empirique	-	0.9774	0.9774	-
Bêta /	BIN	0.0013	0.9725		6.23E-06
Normale	MCP	**2.20E-05**	**0.9773**	**0.9773**	**1.98E-07**
		(1.63E-05)	(0.0026)	(0.0026)	(1.60E-07)
	Empirique	-	0.9891	0.9891	-
Bêta /	BIN	9.25E-04	0.9601		9.05E-06
Gamma	MCP	**7.92E-06**	**0.9897**	**0.9897**	**7.51E-08**
		(7.48E-06)	(0.002)	(0.002)	(7.44E-08)

TABLE 6.5 Discrimination faible : mesures de performances des données simulées (l'écart-type)

S_{D_0}/S_{D_1}	*Modèle du courbe ROC*	**MSE**	**AUC** Trapèze	**AUC** MW	**MSEpAUC**
Normale	Empirique	-	0.6357	0.6355	-
	BIN	**7.05E-05**	0.6325		**5.96E-07**
	MCP	2.63E-04	**0.6361**	**0.6358**	2.29E-06
		(2.14E-04)	(0.0105)	(0.0105)	(2.11E-06)
Gamma	Empirique	-	0.8074	0.8073	-
	BIN	0.0062	0.7781		6.03E-05
	MCP	**2.19E-04**	**0.8077**	**0.8073**	**1.78E-06**
		(1.59E-04)	(0.0076)	(0.0076)	(1.53E-06)
Bêta	Empirique	-	0.9267	0.9266	-
	BIN	3.59E-04	0.9412		2.84E-06
	MCP	**7.44E-05**	**0.9263**	**0.9261**	**5.44E-07**
		(5.05E-05)	(0.0049)	(0.0049)	(4.86E-07)
Normale / Gamma	Empirique	-	0.9207	0.9206	-
	BIN	0.0038	0.877		3.70E-05
	MCP	**9.39E-05**	**0.9226**	**0.9224**	**7.13E-07**
		(7.06E-05)	(0.0044)	(0.0044)	(6.95E-07)
Normale / Bêta	Empirique	-	0.9567	0.9565	-
	BIN	2.19E-04	0.9591		1.07E-06
	MCP	**1.08E-04**	**0.9566**	**0.9563**	**4.95E-07**
		(6.58E-05)	(0.0033)	(0.0031)	(5.43E-07)

TABLE 6.6 Discrimination faible : mesures de performances des données simulées (l'écart-type) (cont.)

S_{D_0}/S_{D_1}	*Modèle du courbe ROC*	**MSE**	**AUC** Trapèze	**AUC** MW	**MSEpAUC**
	Empirique	-	0.8940	0.8938	-
Gamma /	BIN	0.0021	0.8931		1.66E-05
Normale	MCP	**2.76E-04**	**0.8938**	**0.8935**	**1.94E-06**
		(2.36E-04)	(0.0062)	(0.0063)	(1.95E-06)
	Empirique	-	0.7619	0.7617	-
Gamma /	BIN	0.0065	0.7115		6.21E-05
Bêta	MCP	**3.39E-04**	**0.7641**	**0.7638**	**2.82E-06**
		(3.38E-04)	(0.009)	(0.0092)	(3.21E-06)
	Empirique	-	0.8459	0.8458	-
Bêta /	BIN	0.0045	0.8425		3.43E-05
Normale	MCP	**1.27E-04**	**0.8458**	**0.8457**	**1.16E-06**
		(1.01E-04)	(0.0071)	(0.007)	(9.99E-07)
	Empirique	-	0.9591	0.9591	-
Bêta /	BIN	0.0012	0.9249		1.20E-05
Gamma	MCP	**2.65E-05**	**0.9582**	**0.9582**	**2.51E-07**
		(2.79E-05)	(0.0039)	(0.0039)	(2.77E-07)

6.2 Étude sur les données réelles

Les données proviennent de l'étude d'Aziz *et al.* [1] sur l'indice de la déformation des spermes, SDI. Cet indice est défini comme le nombre moyen des déformations par sperme. On compte 158 patients qui subissent un traitement de fécondation in vitro où 73% des sujets sont reconnus fertiles, *i.e.* non-malades. On a $n_{D_0} = 116$

et $n_{D_0} = 42$. La fertilité masculine est mesurée par la réussite de grossesse de leur partenaire.

6.2.1 Analyse graphique

Contrairement aux données simulées, dans une étude clinique, la nature distributionnelle des scores nous est inconnue. L'histogramme 6.56 montre que les scores sont légèrement dispersés. À première vue, selon la figure 6.57, on constate que les courbes théoriques respectent la propriété de la monotonie d'une fonction croissante contrairement à la courbe empirique. De plus, la courbe empirique se situe majoritairement dans l'intervalle de confiance à 95% de la courbe MCP, voir la remarque 4.2.9. à l'exception de certain endroit. Visuellement parlant, il est difficile de soutirer une conclusion. On constate qu'au départ la courbe MCP est adjacent à l'empirique, mais qu'après la BIN devance et ainsi de suite. Dans cette circonstance, une analyse quantitative serait cruciale.

6.2.2 Analyse quantitative

Soulignons que pour la mesure de MSE, nous avons aussi pris 1000 découpes sur l'axe des FPR comme pour les données simulées. Quant au calcul du pAUC, nous avons coupé en 10 sous-intervalles.

Visuellement, il est infaisable de déterminer laquelle réplique mieux l'empirique. Espérons que les résultats quantitatifs nous soient plus informatifs. Les résultats démontrent que la valeur de MSE pour la méthode BIN est approximativement une fois plus petite (0.0021/0.002) à celle de la MCP. Par contre, cette différence est tellement petite qu'elle est négligeable. Ainsi en moyenne les points de la courbe MCP est aussi proche de l'empirique que ceux de la BIN. Les chiffres de l'AUC pour la MCP, par la méthode du trapèze ou MW, sont nettement proches de l'empirique que la BIN. La valeur du MSEpAUC de la BIN est 1.5 fois inférieure ($1.21 \cdot 10^{-5}/7.50 \cdot 10^{-6}$)

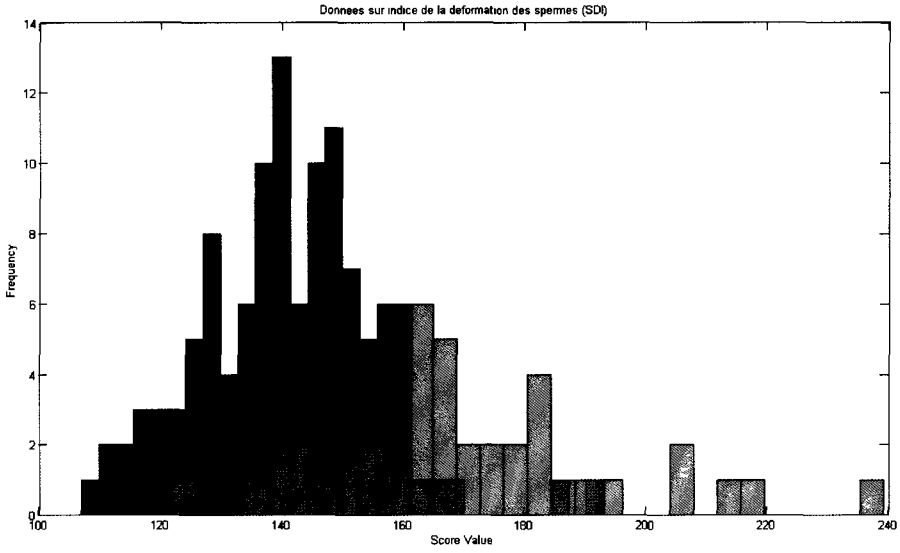

FIGURE 6.56 Histogramme des populations non-malades (en bleu) et malades (en rose) où la distribution des scores est inconnue et $n_{D_0} = 116$ et $n_{D_1} = 42$

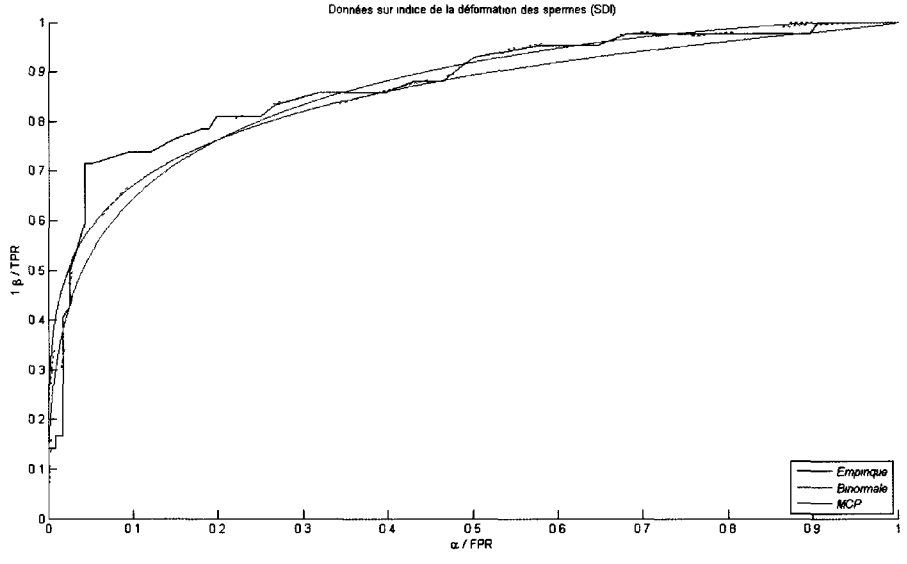

FIGURE 6.57 Courbes ROC empirique, binormale et MCP avec un intervalle de confiance à 95% sur l'indice de la déformation (SDI)

que la MCP.

Une des explications possible est que les données peuvent être de nature gaussienne. Afin de soutenir notre hypothèse, nous avons effectuer un test de normalité. D'après le test de normalité de Lilliefors sur D_0 ($p - value = 0.4128$) et D_1 ($p - value = 0.1418$), les données sont effectivement gaussiennes. Il est impératif qu'on obtient ces résultats puisqu'on estime des données gaussiennes avec une distribution binormale. Par contre, la différence du MSE et du MSEpAUC peut être qualifiée négligeable. Ainsi les deux méthodes performent aussi bien l'une que l'autre.

TABLE 6.7 Mesures de performances des données simulées où l'écart-type est donné entre parenthèses

	Modèle du		**AUC**		
S_{D_0}/S_{D_1}	*courbe ROC*	**MSE**	Trapèze	MW	**MSEpAUC**
Indice de	Empirique	-	0.878	0.8754	-
déformations	BIN	**0.002**	0.8539		**7.50E-06**
des spermes	MCP	0.0021	**0.8632**	**0.8629**	1.21E-05
(SDI)		(5.14E-04)	(0.0063)	(0.0063)	(4.75E-06)

Chapitre 7

Conclusion

La courbe ROC ou *Receiver Operating Characteristic*, a été développée dans les années 1950 dans le domaine de la théorie de détection des signaux à des fins militaires [14]. De nos jours, la courbe ROC s'est répandue dans plusieures domaines comme médical, psychologie, finance et même *machine learning*. La courbe ROC est un outil transcendant pour : évaluer l'efficacité et la précision d'un test diagnostique, sélectionner un seuil décisionnel optimal et comparer deux ou plusieurs courbes, par exemple, comparer un nouveau test par rapport au test existant.

La courbe ROC se construit à partir des scores de deux types de populations, soit non-malades (D_0) et malades (D_1). Généralement, la courbe ROC est décrite comme suit

$$\begin{aligned} ROC &= \{(P(S > t \mid D_0), P(S > t \mid D_1)), t \in \mathbb{R}\} \\ &= \{(FPR(t), TPR(t)), t \in \mathbb{R}\} \end{aligned}$$

où t désigne le seuil décisionnel. Une des propriétés de la courbe ROC est la monotonie d'une fonction croissante. Grâce à cette propriété, la sélection d'un seuil décisionnel optimal est accessible. Malheureusement, la courbe ROC empirique respecte rarement cette propriété. Ainsi nous cherchons à déterminer une courbe ROC théorique respectant les propriétés théoriques d'une courbe ROC.

Il existe essentiellement deux stratégies pour la modélisation de la courbe ROC. La première, dite *directe*, consiste à construire explicitement une courbe ROC empirique directement à partir des scores de la population non-malade et malade. soit S_{D_0} et S_{D_1}, respectivement, et ce, sans hypothèse distributionnelle. Toutefois, la courbe empirique obtenue est indésirable, car elle ne respecte pas certaines propriétés théoriques d'une courbe ROC. Afin de contourner ce problème, plusieurs auteurs proposent d'utiliser diverses techniques de lissage de la courbe empirique ou tout simplement en supposant une forme fonctionnelle quelconque [17, 31, 32, 41, 50]. En pratique, cette approche est peu attrayante due à sa faiblesse intuitive.

La seconde approche, dite *indirecte*, consiste à modéliser les scores des deux populations en supposant une distribution quelconque, par exemple, gaussienne, gamma ou bêta. Ensuite, il s'agit de dériver implicitement une forme fonctionnelle de la courbe ROC à partir des distributions théoriques des scores. Pour certaines distributions, il existe une formule analytique directe pour la loi jointe comme la binormale, bi-gamma ou bi-bêta. Dans la littérature, la majorité des recherches se concentrent sur l'estimation des paramètres du modèle binormal [4, 14, 19, 21, 49, 47].

La faiblesse du modèle binormal est causée par la non-gaussienne des donnéees expérimentales. Dans notre étude, nous proposons de construire la courbe ROC en supposant que la distribution des scores appartienne à la famille de Pearson. Par conséquent, notre approche s'inscrit dans la deuxième catégorie, *i.e.* indirecte. Le choix de la distribution de Pearson est théoriquement plus attirante, puisqu'elle permet de capturer les quatre premiers moments des données. Par ailleurs, les lois normale, gamma et bêta sont des cas particuliers de la loi de Pearson. Toutefois, il n'existe pas de forme fermée à la manière de la binormale, par exemple. Néanmoins la puissance computationnelle des ordinateurs d'aujourd'hui permet aisément d'utiliser la technique de simulation par Monte-Carlo pour construire numériquement la courbe ROC théorique. De plus, cette approche, dite MCP ou *Monte-Carlo Pearson*, nous permet de calculer diverses statistiques autour de la courbe ROC telle que les

intervalles de confiance. Intuitivement, nous espérons que notre stratégie offre relativement de meilleur résultat que la binormale en terme de réplication de la courbe empirique.

Afin de valider notre hypothèse, nous avons comparé la performance de la méthode MCP avec celle de la binormale (BIN), souvent utilisée comme point de repère ou *benchmark* dans la littérature. Plus précisement, l'objectif central est d'étudier la flexibilité et la performance d'ajustement des deux méthodes à capturer une courbe ROC empirique quelconque. Sur ce, notre étude se fait sur deux types de données : simulées et réelles, avec une taille échantillonnage $n_{D_0} = 1000$ et $n_{D_1} = 500$ pour la population non-malade et malade, respectivement.

Dans l'étude de simulation, on a appliqué les deux méthodes sur des compositions de trois types de distributions soit gaussienne, gamma et bêta. Par la suite, en variant les paramètres des distributions, on obtient différent degré de discrimination soit forte, modérée et faible. En utilisant des mesures de performance, voir chapitre 5, comme MSE, AUC et MSEpAUC, nous avons comparé la performance des méthodes MCP et BIN sur les neuf compositions possibles des données et chacun sous trois types de discrimination, donc au total 27 cas. Graphiquement, lorsque la dispersion est flagrante, la méthode MCP performe aussi bien que la BIN. Tel soupçonné, la méthode MCP capture mieux des cas extrêmes comme : $S_{D_0} \sim N(2,6)$ et $S_{D_1} \sim Gamma(7,10)$, $S_{D_0} \sim Beta(3,1)$ et $S_{D_1} \sim N(1.75, 0.25)$, et $S_{D_0} \sim Beta(3.5,1)$ et $S_{D_1} \sim Gamma(10, 0.35)$. Au fur et à mesure que le degré de discrimination diminue, la méthode MCP affiche une supériorité en terme de réplication de la courbe empirique, à l'exception du cas où les deux distributions de scores suivent une gaussienne. Par contre, on observe que la courbe MCP est *quasi* superposée à la courbe BIN. Visuellement, il est très difficile de distinguer laquelle des deux courbes reproduisent mieux l'empirique. Par contre, numériquement, la méthode BIN devance légèrement la MCP. Cette conclusion n'est certes surprenante puisqu'on estime des données gaussiennes avec une distribution binormale. La méthode MCP demeure un

adversaire redoutable puisque ses résultats performent mieux que la BIN. La grande question, maintenant, est de savoir si la méthode BIN performe aussi bien que la MCP pour les autres cas ? Numériquement, la BIN semble gagner du terrain dans un milieu où la discrimination est forte. Cinq cas sur neuf favorisent la BIN pour les mesures MSE et MSEpAUC. Quant à l'AUC, on observe soit une égalité soit supériorité de la MCP à la BIN. Au fur et à mesure que le degré de discrimination diminue, la méthode MCP conquiert la BIN. En pratique, il est peu fréquent d'observer une séparation complète entre les groupes de patients. Ainsi on s'intéresse davantage au comportement des courbes BIN et MCP dans un milieu où la discrimination est restreinte. En somme, pour les données simulées, la méthode MCP surpasse significativement la méthode BIN dans la flexibilité et l'adjustement de la courbe empirique.

Dans l'étude réelle, nous avons pris les données d'Aziz *et al.* [1] sur l'indice de la déformation des spermes, SDI. Contrairement aux données simulées, la nature distributionnelle de ces données nous est inconnue et la taille d'échantillon est petite, soit $n_{D_0} = 116$ et $n_{D_1} = 42$. Selon l'aspect visuelle, voir la figure 6.57, il semble que la courbe MCP reproduit mieux l'empirique que la courbe BIN. Numériquement, la méthode BIN performe légèrement mieux que la méthode MCP. Cette performance de la BIN peut s'expliquer par la nature gaussienne, à priori des données selon le test de normalité de Lilliefors. Ainsi nous pouvons conclure que les deux méthodes performent aussi bien l'une que l'autre.

En somme, la performance de la modélisation de la courbe ROC par simulation de Monte-Carlo de la distribution de Pearson offre une alternative attrayante à la méthode classique, la binormale. En se basant sur les résultats graphiques et quantitatifs, des données simulées et réelles, on remarque la flexibilité que procure la méthode MCP à capturer les déformations des données. À l'opposé, la méthode BIN arrive difficilement à répliquer la courbe empirique lorsque les données ne sont pas gaussiennes. En terme général, la méthode MCP démontre une supériorité dans la flexibilité et l'adjustement de la courbe empirique que la BIN.

Une des faiblesses à notre approches est sans équivoque le recours à la méthode de simulation par Monte-Carlo puisqu'il n'existe pas de forme fermée en deux dimensions pour la distribution de Pearson. Contrairement à la méthode binormale qui possède une formule analytique de la courbe ROC, la méthode MCP construit la courbe ROC numériquement, voir section 4.2. Mais grâce à la puissance computationnelle de nos jours, la technique de Monte-Carlo se fait promptement.

Notre étude, bien qu'elle présente des résultats intéressants, peut être améliorée de plusieurs façons. Comme dans le cas de la binormale où plusieurs auteurs se sont attardés à développer de meilleures méthodes pour estimer les paramètres de manière plus robustes, il est possible d'en faire autant pour les quatre paramètres de la distribution de Pearson. Par exemple, une approche baysienne peut être effectuée. Pour augmenter l'efficacité de la simulation d'un système de Pearson, la technique de Monte-Carlo peut être améliorée en introduisant, dans l'algorithme de simulation, une technique de réduction de variance quelconque. Un autre aspect que notre étude ne s'est pas attardée est la construction d'un intervalle de confiance bi-dimensionnel autour de la courbe ROC comme certains auteurs ont fait pour la méthode binormale. Cet intervalle de confiance bi-dimensionnelle permet de caractériser les deux axes FPR et TPR de la courbe ROC. Enfin, il serait aussi intéressant de comparer la méthode MCP à d'autres méthodes sur plusieurs ensembles de données réelles.

Annexe A

Relation entre la courbe ROC et le lemme de Neyman-Pearson

Le rapport de vraisemblance nous donne la probabilité qu'une valeur t survienne dans la population D_1 est plus probable que dans D_0. Par contre, la corrélation entre la pente de la courbe ROC et la théorie des tests d'hypothèse est encore plus attrayante que le ratio de vraisemblance. Considérons le test d'hypothèse suivant :

$$H_0 : \text{patient appartient à } D_0 \qquad vs \qquad H_1 : \text{patient appartient à } D_1$$

La fonction de score, S, permet de classer un sujet dans une population quelconque. Ainsi, la construction des données se fait grâce à cette fonction de score. Soit Ω, l'ensemble de toutes les valeurs possibles de S pour lesquelles nous rejetons l'hypothèse nulle, autrement dit nous attribuons un individu à la population D_1. Le lemme de Neyman-Pearson énonce que le test le plus puissant de taille α possède une région Ω comprennant toutes les valeurs s de S tel que

$$\lambda(s) = \frac{P(s \mid D_1)}{P(s \mid D_0)} \geq k,$$

où $k \in [0, \infty)$. La valeur de k peut être déterminée par la condition $\alpha = P(S \in \Omega \mid D_0)$. Cependant pour l'intérêt du sujet, α se définit comme suit :

$$\begin{aligned}
\alpha &= P(\text{Erreur de type I}) \\
&= P(\text{Rejeter } H_0 \mid H_0 \text{ vrai}) \\
&= P(S > t \mid D_0) \\
&= FPR
\end{aligned}$$

et la puissance d'un test,

$$\begin{aligned}
1 - \beta &= 1 - P(\text{Erreur de type II}) \\
&= 1 - P(\text{Accepter } H_0 \mid H_1 \text{ vrai}) \\
&= P(S > t \mid H_1) \\
&= TPR
\end{aligned}$$

Donc, le lemme de Neyman-Pearson montre que pour un FPR fixe, TPR peut être maximisé par un classement dont l'ensemble des scores S est donné par $\lambda(S) \geq k$.

Annexe B

Algorithme de construction de la courbe ROC

Algorithm Method for calculating ROC points

Inputs : $S_{D_0}, S_{D_1}, mean_0, mean_1, n_0, n_1$ the scores, mean and size of healthy and diseased patients, respectively,

Outputs : A vectors of FPR and TPR

Require : $n_0 > 0$ and $n_1 > 0$

$X \leftarrow$ combine S_{D_0} and S_{D_1} in one vector

$X_{sorted} \leftarrow$ sort X in ascending order

$t \leftarrow$ find unique value of X_{sorted}

$n_t \leftarrow$ size of t

create an empty array A

for $i = 1$ to n_t

  if $mean_0 < mean_1$ **then**

    $FPR(i) = \frac{\text{count}(S_{D_0} > t(i))}{n_0}$

    $TPR(i) = \frac{\text{count}(S_{D_1} > t(i))}{n_1}$

  else

    $FPR(i) = \frac{\text{count}(S_{D_0} < t(i))}{n_0}$

    $TPR(i) = \frac{\text{count}(S_{D_1} < t(i))}{n_1}$

  endif

  add (FPR, TPR) onto A

endfor

Annexe C

Algorithme de la technique de simulation par Monte-Carlo

Algorithm Monte-Carlo simulation

Inputs : $S_{D_0}, S_{D_1}, m1_0, m2_0, m3_0, m4_0, m1_1, m2_1, m3_1, m4_1, n_0, n_1, M$: the scores, first four moments, size and nb of simulation ;

Outputs : A vectors of FPR, TPR and SIM.

1 : create an empty array SIM

2 : create an empty array A

3 : **for** $i = 1$ to M

$S'_{D_0} \leftarrow$ generate a vector of $S_{D_0} \sim Pearson(\theta_0)$ with size n_0

$S'_{D_1} \leftarrow$ generate a vector of $S_{D_1} \sim Pearson(\theta_1)$ with size n_1

add (S'_{D_0}, S'_{D_1}) onto SIM

call Algorithm Method for calculating ROC points (FPR, TPR) from S'_{D_0} and S'_{D_1}

add (FPR, TPR) onto A

endfor

Bibliographie

[1] N. Aziz, I. Buchan, C. Taylor, C.R. Kingsland and I. Lewis-Jones (1996). Sperm deformity index : A reliable predictor of the outcome of fertilization in vitro, *Fertility and Sterility*, 66 :1000-1008.

[2] Donald Bamber (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph, *Journal of Mathematical psychology*, 12 :387-415.

[3] L.C. Brumback, M.S. Pepe and T.A. Alonzo (2006). Using the ROC curve for gauging treatment effect in clinical trials, *Statistics in Medicine*, 25 :575-590.

[4] T. Cai and C.S. Moskowitz (2004). Semi-parametric estimation of the binormal ROC curve for a continuous diagnostic test, *Biostatistics*, 5 :573-586.

[5] G. Campbell and M.V. Ratnaparkhi (1993). An application of Lomax distributions in receiver operating characteristic (ROC) curve analysis, *Communications in Statistics*, 22 :1681-1697.

[6] Y. Dodge and G. Melfi (2008). *Premiers pas en simulation*, Springer, France.

[7] D.D. Dorfman and E. Jr. Alf (1968). Maximum likelihood estimation of parameters of signal detection theory - a direct solution, *Psychometrika*, 33 :117-124.

[8] D.D. Dorfman, K.S. Berbaum, C.E. Metz, R.V. Lenth, J.A. Hanley and H.A. Dogga (1997). Proper receiver operating characteristic analysis : the bigamma model, *Academic Radiology*, 4 :138-149.

[9] James P. Egan (1975). *Signal detection theory and ROC analysis*, Academic Press, New York.

[10] William L. England (1988). An exponential model used for optimal threshold selection on ROC curves, *Medical Decision Making*, 8 :120-131.

[11] Tom Fawcett (2006). An introduction to ROC analysis, *Pattern Recognition Letters*, 27 :861-874.

[12] D. Faraggi and B. Reiser (2002). Estimation of the area under the ROC curve, *Statistics in Medicine*, 21 :3093-3106.

[13] M.J. Goddard and I. Hinberg (1990). Receiver operating characteristic (ROC) curves and non-normal data : an empirical study, *Statistics in Medicine*, 9 :325-337.

[14] D.M. Green and J. Swets (1966). *Signal Detection Theory and Psychophysics*, John Wiley and Sons, New York.

[15] J. Gu, S. Ghosal and A. Roy (2008). Bayesian bootstrap estimation of ROC curve, *Statistics in Medicine*, 27 :5407-5420.

[16] K.O. Hajian-Tilaki, J.A. Hanley, L. Joseph and J-P. Collet (1997). A Comparaison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests, *Medical Decision Making*, 17 :94-102.

[17] P.G. Hall and R.J. Hyndman (2003). Improved methods for bandwidth selection when estimating ROC curves, *Statistics & Probability Letters*, 64 :181-189.

[18] James A. Hanley (1989). Receiver Operating Characteristic (ROC) Methodology : The State of the Art, *Clinical Reviews in Diagnostic Imaging*, 29 :307-335.

[19] James A. Hanley (1996). The use of the binormal model for parametric ROC analysis of quantitative diagnostic tests. *Statistics in Medicine*, 15 :1575-1585.

[20] J.A. Hanley and B.J. McNeil (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143 :29-36.

[21] F. Hsieh and B.W. Turnbull (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve, *The Annals of Statistics*, 24 :25-40.

[22] M.Huguier and A. Flahault (2003). *Biostatistiques au quotidien*, Éditions scientifiques et médicales Elsevier SAS, Paris.

[23] Adrien Jamain (2004). *Meta-Analysis of Classification Methods*, Unpublished PhD Thesis, Departement of Mathematics, Imperial College, London.

[24] W.J. Krzanowski and D.J. Hand (2009). *ROC Curves for Continuous Data*, Taylor and Francis Group, Florida.

[25] Kenneth Lange (1999). *Numerical Analysis for Statisticians*, Springer, New York.

[26] B. Lapeyre, E. Pardoux and R. Sentis (1998). *Méthodes de Monte-Carlo pour les équations de transport et de diffusion*, Springer, Berlin.

[27] Pierre L'Écuyer (2009). *cours IFT6561, Simulation : aspects stochastiques*, Université de Montréal, http .//www.iro.umontreal.ca/ lecuyer/ift6561.html.

[28] Lee B. Lusted (1971). Signal Detectability and Medical Decision-Making, *Science*, 171 :1217-1219.

[29] Erich L. Lehmann (2006). *Nonparametrics Statistical Methods Based on Ranks*, Springer, Berkeley.

[30] G. Li, R.C. Tiwari and M.T. Wells (1999). Semiparametric inference for a quantile comparison function with applications to receiver operating characteristic curves, *Biometrika*, 86 :487-502.

[31] Chris J. Lloyd (1998). Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems, *Journal of the American Statistical Association*, 93 :1356-1364.

[32] I. López-de-Ullibarri, R. Cao, C. Cadarso-Suárez and M.J. Lado (2008). Nonparametric estimation of conditional ROC curves : application to discrimination

tasks in computerized detection of early breast cancer, *Computational Statistics & Data Analysis*, 52 :2623-2631.

[33] J.H. Mathews and K.D. Fink (2004). *Numerical Methods Using MATLAB*, 4^{th} Edition, Prentice-Hall Inc, New Jersey.

[34] Donna Katzman McClish (1989). Analyzing a Portion of the ROC Curve, *Medical Decision Making*, 9 :190-195.

[35] C.E. Metz, B.A. Herman and J-H. Shen (1998). Maximum Likelihood Estimation of Receiver Operating Characteristic (ROC) Curves from Continuously-Distributed Data, *Statistics in Medicine*, 17 :1033-1053.

[36] J.C. Ogilvie and C.D. Creelman (1968). Maximum likelihood estimation of ROC curve parameters, *Journal of Mathematical Psychology*, 5 :377-391.

[37] Magaret Sullivan Pepe (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, New York.

[38] L. Peng and X.-H. Zhou (2004). Local linear smoothing of receiver operating characteristic (ROC) curves, *Journal of Statistical Planning and Inference*, 118 :129-143.

[39] K. Powell, N. Obuchowski, W.A. Chilcote, M.W. Barry, S.N. Ganobcik and G. Cardenosa (1999). Clinical evaluation of digital versus film-screen mammograms : Diagnostic accuracy and patient management, *Am.J. Roentgenol*, 173 :889-894.

[40] J. Qin and B. Zhang (2003). Using logistic regression procedures for estimating receiver operating characteristic curves, *Biometrika*, 90 :585-596.

[41] P. Qiu and C. Le (2001). ROC curve estimation based on local smoothing, *Journal of Statistical Computation and Simulation*, 70 :55-69.

[42] Bruno Scherrer (1984). *Biostatistique*, Gaëtan Morin Éditeur, Québec.

[43] E.V. Shikin and A.I. Plis (1995). *Handbook on splines for the user*, CRC Press Inc, Florida.

[44] Alan Stuart and J. Keith Ord (1987). *Kendall's Advanced Theory of Statistics : Volume 1*, Oxford University Press, New York.

[45] J.A. Swets and R.M. Pickett (1982). *Evaluation of diagnostic systems : Methods from signal detection theory*, Academic Press, New York.

[46] X.-H. Zhou and J. Harezlak (2002). Comparison of bandwidth selection methods for kernel smoothing of ROC curves, *Statistics in Medicine*, 21 :2045-2055.

[47] X.-H. Zhou and H. Lin (2008). Semi-parametric maximum likelihood estimates for ROC curves of continuous-scale tests, *Statistics in Medicine*, 27 :5271-5290.

[48] X.-H. Zhou, N.A. Obuchowski and D.K. McClish (2002). *Statistical Methods in Diagnostic Medicine*, Wiley, New York.

[49] K.H. Zou and W.J. Hall (2000). Two transformation models for estimating an ROC curve derived from continuous data, *Journal of Applied Statistics*, 27 :621-631.

[50] K.H. Zou, W.J. Hall and D.E. Shapiro (1997). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests, *Statistics in Medicine*, 16 :2143-2156.

[51] M.H. Zweig and G. Campbell (1993). Receiver Operating Characteristic (ROC) Plots : A Fundamental Evaluation Tool in Clinical Medicine, *Clinical Chemistry*, 39 :561-577.