Université de Montréal

APPROCHE BAYÉSIENNE À LA CLASSIFICATION DE PATIENTS SÉROPOSITIFS

par

Anne-Marie Robert

Département de mathématiques et de statistique

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures

En vue de l'obtention du grade de

Maître ès sciences (M. Sc.)
en mathématiques

Août 1999

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

APPROCHE BAYÉSIENNE À LA CLASSIFICATION DE PATIENTS SÉROPOSITIFS

présenté par

Anne-Marie Robert

a été évalué par un jury composé des personnes suivantes :

Yves Lepage
(président-rapporteur)

Jean-François Angers
(directeur de recherche)

Martin Bilodeau
(membre du jury)

Mémoire accepté le :
15 *Novembre 1999*

SOMMAIRE

Depuis la découverte du virus de l'immunodéficience humaine (VIH), de nombreux chercheurs ont élaboré des traitements pour contrer les effets du sida. À chaque année, plusieurs personnes atteintes du VIH sont enrôlées dans des études visant à comprendre le comportement du VIH vis-à-vis divers médicaments.

Dans ce mémoire, nous tentons de modéliser le comportement de patients traités pour le VIH afin de pouvoir classer ceux-ci en deux groupes distincts. Le premier groupe est composé des patients dont le traitment est efficace et l'autre est composé de ceux qui ne répondent pas aussi bien au traitement. Pour ce faire, nous essayons tout d'abord de modéliser la charge virale (nombre de copies du virus dans le sang) qui explique en grande partie l'évolution du patient. Nous utilisons des splines pour estimer une fonction permettant d'exprimer la charge virale en fonction du temps. Par la suite, nous vérifions si ces splines permettent de regrouper correctement les patients en utilisant l'analyse discriminante.

Puis, nous utilisons des modèles contaminés qui sont formés d'un mélange de deux distributions, une pour chaque groupe, pour classifier les patients. Afin d'estimer les paramètres de ces modèles contaminés, l'algorithme EM est utilisé. Nous regardons ensuite si ces modèles réussissent à bien classer les patients.

REMERCIEMENTS

Je remercie d'abord Monsieur Serge Tardif qui, à l'origine, a accepté de travailler sur ce projet de maîtrise. Il a toujours été d'une grande gentillesse et son support m'a été d'un grand recours.

J'aimerais plus particulièrement remercier mon directeur de recherche, Monsieur Jean-François Angers, pour sa disponibilité et sa grande patience. Il a toujours été très accueillant lors de nos nombreuses rencontres dans son bureau. Je le remercie aussi pour les conseils qu'il m'a donnés.

Je suis reconnaissante au Fonds du Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) pour la bourse qu'il m'a octroyée.

Je voudrais aussi remercier Monsieur Chris Tsoukas, médecin au Centre de traitement de l'immunodéficience à l'Hôpital général de Montréal, pour m'avoir permis d'utiliser les données provenant d'une étude clinique et pour son support financier.

Enfin, je tiens également à remercier mes compagnons de travail qui m'ont souvent aidée et conseillée tout au long de ces dernières années d'étude.

Table des matières

Table des figures

Liste des tableaux

INTRODUCTION

Depuis plusieurs années, les compagnies pharmaceutiques mettent au point des traitements qui détruisent de façon significative le virus de l'immunodéficience humaine (VIH) permettant ainsi aux patients atteints du VIH de ralentir leur progression vers le sida. Plusieurs études tentent d'établir des modèles permettant de pouvoir prédire le comportement des patients traités avec une nouvelle combinaison de médicaments. Dans ce mémoire, nous analysons les données d'une étude clinique effectuée auprès de patients infectés par le VIH. Nous cherchons à savoir si les patients répondant bien au traitement ont le même comportement que ceux dont le traitement n'a pas été aussi efficace. Cette étude nous servira d'application dans les chapitres qui suivent.

Pour débuter, au premier chapitre, nous présentons d'abord l'étude clinique puis nous introduisons quelques techniques qui servent dans les méthodes d'analyses que nous développons aux deux chapitres suivants. Nous voyons dans l'ordre suivant : l'analyse discriminante, l'algorithme EM et le modèle contaminé. Afin de faciliter la compréhension de ces méthodes, un exemple simple est présenté et illustre les trois techniques expliquées. Cet exemple est donné après chaque section.

Ensuite, dans le deuxième chapitre, nous tentons de trouver une fonction modélisant la charge virale (copies du virus contenues dans le sang) en fonction de la variable temps dans le but de prédire le comportement des patients. Nous utilisons une spline pour estimer cette fonction. Tout d'abord avec la méthode

classique, nous utilisons une spline de lissage basé sur une série de Taylor tronquée. Nous nous servons surtout des travaux de Wahba (1990) et de Eubank (1988) pour le développement de la spline de lissage. Par la suite, nous utilisons l'approche bayésienne, qui tient compte de l'information *a priori* sur la fonction à estimer et celle fournie par les observations de notre étude, pour calculer une spline estimant la fonction que nous cherchons. Cette fois-ci, nous faisons appel à la méthode proposée par O'Hagan (1978). Pour chacune de ces approches, nous voyons s'il est possible de classifier les patients en deux groupes, selon l'évolution de leur charge virale, à l'aide des splines en utilisant l'analyse discriminante. Cette analyse construit une fonction qui classifie les patients et calcule le taux d'erreur de classification. Enfin, à la dernière section de ce chapitre, nous comparons les résultats obtenus par l'approche classique avec ceux obtenus dans un contexte bayésien.

Enfin, le troisième chapitre est consacré au développement d'un modèle servant à classifier les patients selon l'impact du traitement sur la charge virale en utilisant un modèle contaminé. Ce modèle est constitué d'un mélange de deux distributions de la variable charge virale, une pour chacun des deux groupes de patients. Le premier groupe représente les patients qui répondent bien au traitement et le deuxième groupe représente les autres patients. Dans un premier temps, nous développons un modèle contaminé de façon classique en utilisant l'algorithme EM pour estimer les paramètres de ce modèle. Nous nous basons principalement sur les travaux de Titterington, Smith et Makov (1985) pour estimer ces paramètres. Puis, nous trouvons un modèle contaminé bayésien en ajoutant l'information *a priori* en nous basant sur un modèle proposé par Zellner (1971). L'algorithme EM est encore utilisé pour estimer les paramètres. Par la suite, une fois les paramètres estimés pour les deux modèles contaminés, nous vérifions si les modèles

permettent une bonne classification des patients. Nous calculons le pourcentage de patients mal classés. Encore une fois, une comparaison des résultats des deux modèles contaminés est présentée à la dernière section du chapitre.

Chapitre 1

TECHNIQUES UTILISÉES

Dans ce premier chapitre, nous débutons par une description du jeu de données qui nous servira d'application pour les méthodes d'analyse que nous utilisons dans les chapitres 2 et 3. Les données proviennent d'une étude clinique effectuée sur des patients séropositifs et traités avec un médicament qui détruit le VIH. Par la suite, nous expliquons brièvement quelques techniques qui sont utilisées dans le développement des méthodes d'analyse. Ces techniques ne sont pas décrites dans leur complète généralité mais plutôt dans le contexte de notre jeu de données. Nous décrivons d'abord l'analyse discriminante puis l'algorithme EM. Afin de mieux expliquer l'algorithme, nous introduisons aussi le modèle contaminé. Ce dernier est discuté plus en détail au chapitre 3 où nous nous servons des modèles contaminés pour mieux représenter nos observations.

1.1. DESCRIPTION DES OBSERVATIONS

Une infection au VIH cause une détérioration progressive du système immunitaire. Il existe des traitements médicaux qui peuvent détruire largement le VIH dans l'organisme.

Un échantillon de 58 patients, composé de 6 femmes et 52 hommes âgés en moyenne de 42,8 ans (avec un écart type de 9,3), ayant le VIH et étant asymptomatiques, a été suivi pendant une période de 52 semaines. Ces patients étaient

traités continuellement avec du sulfate d'indinavir. Ils ont été recrutés à l'aide d'un protocole effectué au Centre de traitement de l'immunodéficience à l'Hôpital général de Montréal par le médecin Chris Tsoukas.

Le protocole de cette étude est une évaluation prospective du traitement à l'indinavir chez des patients ayant le VIH. L'hypothèse de cette étude est que le bénéfice thérapeutique de l'indinavir chez les patients affectés par le VIH peut être affecté par un traitement antérieur avec l'inhibiteur de protéase saquinavir. Plus particulièrement, nous pouvons observer chez les patients une diminution de la charge virale (nombre de copies du virus par millilitre de sang) et une augmentation des cellules CD4 (cellules cibles du VIH). La majorité des patients (43 sur 58) ont été préalablement traités avec du saquinavir pour une période d'au moins 6 mois et les autres (15) n'avaient jamais été traités au saquinavir. Aucun des patients enrôlés dans l'étude n'était traité avec un inhibiteur de protéase au moment de son entrée dans l'étude. La plupart des patients ont pris la dose recommandée d'indinavir, c'est-à-dire 800 mg à toutes les 8 heures, et quelques patients ont pris une dose inférieure, c'est-à-dire 400 mg à toutes les 8 heures, sur recommandation de leur médecin. Une fois les valeurs de départ observées (au jour de recrutement et au jour 0 du traitement), des mesures ont été prises au jour 3, à la première semaine, à la deuxième semaine, à la quatrième semaine et par la suite à toutes les quatre semaines jusqu'à la semaine 52. Chaque patient a donc un vecteur d'observations pour chacun des temps suivants: jour 0, jour 3, semaine 1, semaine 2, semaine 4, semaine 8, semaine 12, semaine 16, semaine 20, semaine 24, semaine 28, semaine 32, semaine 36, semaine 40, semaine 44, semaine 48 et semaine 52. Étant donné que certains patients n'ont pas atteint la fin de l'étude au moment de l'analyse, nous ne considérerons dans ce mémoire que les mesures

récoltées jusqu'en août 1997. Ainsi certains patients ont des valeurs manquantes pour certains temps.

Après avoir observé globalement le jeu de données, nous remarquons que la variable charge virale fluctue beaucoup au début de l'étude, c'est-à-dire pour les jours 0 et 3 et pour les semaines 1 et 2. Cette fluctuation peut être expliquée par le changement de médicaments. Afin d'améliorer les résultats des méthodes que nous utilisons pour classer les patients, nous avons volontairement enlevé les observations de tous les patients pour ces quatre temps. De plus, étant donné que la charge virale a une très grande variation (voir plus loin pour plus de détails), nous avons également standardisé cette variable en la divisant par 400. Ce nombre représente la valeur minimale détectable du nombre de copies du virus.

L'objectif principal de cette étude était de vérifier s'il y a une différence entre les deux groupes de patients dans la diminution de la charge virale. Nous avons regroupé les patients d'une nouvelle façon, tel que recommandé par Dr Chris Tsoukas. Dans un groupe, nous retrouvons les patients qui ont atteint le niveau minimal détectable de charge virale, c'est-à-dire 400 copies/ml, et dans l'autre, les patients qui n'ont pas atteint ce niveau, c'est-à-dire pour qui l'inhibiteur de protéase indinavir n'a pas été aussi efficace. Ainsi, dans le premier groupe, nous retrouvons 26 patients ayant atteint une charge virale de 400 copies/ml et le deuxième groupe comprend 32 patients. Pour ce mémoire, nous nous attarderons plutôt à vérifier, dans un premier temps, s'il est possible de modéliser la charge virale en fonction du temps, afin de pouvoir prédire le comportement du virus pour un patient et donc par le fait même, de pouvoir prédire à quel groupe appartient ce patient. Nous avons effectué un bref survol dans la littérature médicale afin de voir ce que les auteurs utilisent comme méthode pour modéliser la charge virale.

Plusieurs auteurs calculent la relation entre la charge virale et la progression vers la maladie en utilisant l'estimateur de Kaplan-Meier et des modèles de régression proportionnels au taux de panne. En autre, Sabin *et al.* (1998) ont trouvé qu'un haut niveau de charge virale est associé avec une progression plus rapide vers le sida et un temps de survie plus court en utilisant un modèle de Cox univarié. Des modèles non linéaires avec liens ont également été utilisés par Drusano et Stein (1998) pour modéliser l'influence de la charge virale sur la progression vers la maladie. D'autres auteurs optent plutôt pour la comparaison des mesures prises à certains points fixes dans le temps pour voir si le traitement a un effet sur la charge virale. Par exemple, Tamalet *et al.* (1997) compare le niveau de la charge virale des patients au début d'un traitement avec quatre antiviraux, avec celui obtenu après huit semaines de traitement. Dans ce mémoire, nous proposons une nouvelle approche pour modéliser la charge virale afin de déterminer la progression d'un patient vers la maladie. Nous utilisons une spline pour exprimer la charge virale en fonction du temps.

Dans un deuxième temps, nous tenterons, toujours à l'aide de la charge virale, de voir s'il est possible de bien classifier un patient dans son groupe respectif. Notre survol de la littérature médicale ne nous a pas permis de trouver des auteurs traitant des données dans le but de classifier des patients. Nous ne tenons donc compte que du fait qu'il est préférable d'utiliser la charge virale pour classifier les patients étant donné que plusieurs auteurs dont Smol'skaia *et al.* (1999) prétendent que c'est la variable qui représente le plus la progression vers la maladie du sida pour des patients ayant le VIH. Dans cet article, les chercheurs expliquent que la détermination du niveau de la charge virale est un critère fiable indiquant la progression d'un patient vers la maladie. Ainsi, un patient ayant une faible charge virale progresse moins vite vers la maladie qu'un patient ayant une

charge virale élevée. L'analyse des données nous permettra de prédire l'évolution de nouveaux patients qui seront enrôlés dans une étude similaire. En effet, une fois le modèle construit, celui-ci pourra être appliqué à d'autres patients. Ainsi, à l'aide de quelques mesures prises sur des nouveaux sujets, nous serons en mesure de décider si ces patients devraient être traités avec les mêmes médicaments ou non. Par conséquent, plusieurs patients pourraient éviter d'être traités à l'aide de médicaments pour lequels aucun effet bénéfique n'est observé. Ceci pourrait grandement ralentir la détérioration de l'état de santé chez certains patients.

Pour parvenir à ces deux objectifs, nous tentons donc de modéliser la charge virale en fonction du temps. Nous utilisons ces deux variables parce que, tout d'abord, la charge virale est le meilleur outil qui existe jusqu'à présent pour représenter l'évolution d'un patient (voir Smol'skaia *et al.* (1999)) et puis parce que nous voulons regarder le comportement de cette variable dans le temps pour pouvoir prédire le comportement des patients.

Le jeu de données comporte 13 variables. La première est le numéro d'identification du patient, la deuxième est le groupe auquel appartient le patient (0 s'il a atteint une charge virale de 400 et 1, sinon) et la troisième est la variable temps, c'est-à-dire le numéro de la semaine où l'observation a été prise. Les six variables suivantes sont des variables qui indiquent des quantités de cellules qui sont des cellules cibles du VIH. Il y en a trois : les cellules CD4, les cellules CD8 et les CD3. Il y a d'abord la quantité d'un type de cellules par ml de sang et la colonne qui suit représente le pourcentage dans le sang pour cette même sorte de cellules. Puis, les trois variables suivantes sont aussi des cellules cibles du VIH mais les chercheurs ne semblent pas en être certains. Ce sont les cellules CD38, DR et CD28. Pour celles-ci, nous n'avons que le pourcentage de cellules présentes dans le sang. Enfin, la dernière variable, la plus importante, est la charge virale.

Nous avons également besoin de vérifier si la charge virale suit une loi normale afin de savoir si nous pouvons appliquer des méthodes d'analyse basées sur la normalité des données. Nous avons donc effectué des tests de normalité en utilisant la statistique de Kolmogorov-Smirnov pour chacun des temps. Le tableau 1.1.1 indique la valeur de la statistique et la valeur p correspondante. Après un bref coup d'oeil aux résultats, nous concluons que la charge virale ne peut être issue d'une loi normale. Nous devons donc utiliser des méthodes non paramétriques pour analyser nos observations. Par contre, nous voyons à la section 1.3 et au chapitre 3 que la loi de la charge virale peut être représentée par un mélange de deux lois normales. En effet, pour simplifier le développement des estimateurs au chapitre 3, nous supposons que la charge virale provient d'un modèle contaminé de lois normales. Après plusieurs essais avec d'autres lois qui n'amélioraient pas les résultats, nous avons décidé d'opter pour la loi normale. Ainsi, la charge virale dans chacun des deux groupes est supposée normale. De plus, sans tenir compte du temps, nous avons calculé les principales statistiques descriptives pour chacun des groupes. Les tableaux 1.1.2 et 1.1.3 présentent les résultats obtenus. En regardant ces tableaux, nous remarquons que le groupe 1 est plus dispersé que le groupe 0. Nous tenons compte de cette observation dans les sections et les chapitres qui suivent où nous tentons de modéliser la charge virale.

1.2. Analyse discriminante

Après avoir modélisé la charge virale comme une fonction du temps, nous tentons de classifier les patients dans leur groupe respectif en utilisant le modèle estimé. Pour ce faire, nous utiliserons l'analyse discriminante. Cette sorte

Tableau 1.1.1: *Tableau des tests de normalité de Kolmogorov-Smirnov pour la charge virale*

	Valeur de la statistique	Valeur p
semaine 4	0,3900	0
semaine 8	0,3416	0
semaine 12	0,3231	0
semaine 16	0,3690	0
semaine 20	0,3673	0
semaine 24	0,3601	0
semaine 28	0,3367	0
semaine 32	0,3593	0
semaine 36	0,3555	0
semaine 40	0,3745	0
semaine 44	0,3032	0
semaine 48	0,3177	0
semaine 52	0,3685	0

Tableau 1.1.2: *Tableau des statistiques descriptives de la charge virale du groupe 0*

Minimum	1er quartile	Médiane	Moyenne	3e quartile	Maximum	Écart type
400	400	400	4308	568	168800	15950

Tableau 1.1.3: *Tableau des statistiques descriptives de la charge virale du groupe 1*

Minimum	1er quartile	Médiane	Moyenne	3e quartile	Maximum	Écart type
400	2087	10730	42110	34480	617900	86438

d'analyse calcule une fonction discriminante qui permet de classer des sujets. Définissons la notion de fonction discriminante.

Définition 1.2.1

Une fonction discriminante est une fonction qui, à l'aide d'observations dont le groupe d'appartenance est connu, construit un critère de classification qui permet de classer de nouvelles observations dont le groupe d'appartenance est inconnu.

En effet, à l'aide d'une variable identifiant le groupe d'appartenance (dans notre cas, c'est la variable groupe qui prend la valeur 0 ou 1) et d'une ou de plusieurs variables quantitatives, cette analyse construit un critère de classification qui classe chaque observation dans un des groupes.

Afin de construire ce critère de classification, nous devons avoir un échantillon dont nous connaissons l'appartenance aux groupes. Cet échantillon, appelé échantillon d'entraînement, permet de construire le critère de classification. Une fois le critère établi, celui-ci est appliqué à toutes les observations de l'échantillon d'entraînement et nous calculons ainsi le taux d'erreur de classification.

Étant donné que l'hypothèse de la normalité des données n'est pas vérifiée (voir à la section 1.1), nous utilisons une méthode non paramétrique appelée ν *plus proches voisins* introduite par Parzen (1962) qui estime la densité des observations pour chaque groupe. Cette méthode fixe à ν le nombre d'observations de l'échantillon d'entraînement utilisées pour classer chaque observation x de dimension p de l'échantillon d'entraînement. Pour chaque observation x, la méthode calcule la distance de Mahalanobis entre l'observation x et les autres observations y de l'échantillon d'entraînement comme suit :

$$d(x, y) = \sqrt{(x - y)^t V^{-1} (x - y)}, \tag{1.2.1}$$

où $(x - y)^t$ est la transposée de $(x - y)$ et où V est la matrice de covariance *combinée* pour l'ensemble des observations et est définie comme suit :

$$V = \frac{(n_0 - 1)S_0 + (n_1 - 1)S_1}{n_0 + n_1 - 2},$$

où S_0 est la matrice de covariance du groupe 0 et S_1 la matrice de covariance du groupe 1 et n_t $(t = 0, 1)$ est le nombre d'observations x de l'échantillon d'entraînement appartenant au groupe t.

La méthode des ν *plus proches voisins* conserve seulement les ν plus petites distances pour estimer la densité au point x. La densité au point x pour le groupe t ($t = 0, 1$ dans notre cas) est alors estimée par :

$$f_t(x) \propto \frac{\nu_t}{n_t \, v_\nu(x)}, \tag{1.2.2}$$

où ν_t ($0 \leq \nu_t \leq \nu$) est le nombre de distances qui sont associées au groupe t parmi les ν plus petites distances conservées et $v_\nu(x) = r_\nu^p(x)|V|^{\frac{1}{2}} v_0$ est le volume de l'ellipsoïde borné par $\{x | x'V^{-1}x = r_\nu^2(x)\}$ où $v_0 = \pi^{\frac{p}{2}}/\Gamma(\frac{p}{2}+1)$ et $r_\nu(x)$ est la distance entre x et son ν^e plus proche voisin. Étant donné que la matrice de covariance *combinée* est utilisée dans le calcul des distances, le volume $v_\nu(x)$ ne dépend pas du groupe d'appartenance.

Supposons que l'observation z est une nouvelle observation dont nous ne connaissons pas le groupe auquel elle appartient. La probabilité *a posteriori* de l'appartenance de z au groupe t selon le théorème de Bayes est obtenue par :

$$p(t|z) = \frac{q_t f_t(z)}{f(z)}, \tag{1.2.3}$$

où $f(z) = \sum_{i=0}^{1} q_i f_i(z)$ est la densité marginale estimée de la nouvelle observation et où q_t est la probabilité *a priori* que nous avons fixée pour chacun des groupes. La densité $f_i(z)$ est estimée par l'équation (1.2.2). Deux probabilités sont ainsi calculées pour chaque nouvelle observation z et nous classons celle-ci dans le groupe t correspondant à la plus grande probabilité obtenue. Lorsque les deux probabilités d'appartenance sont égales, l'observation est classée dans un groupe artificiel appelé groupe autre. Lorsque z appartient à l'échantillon d'entraînement, nous pouvons ainsi vérifier si la classification de l'observation z est bonne.

Afin de mieux comprendre l'analyse discriminante, appliquons ce que nous venons de voir à l'aide d'un exemple. Nous avons généré un échantillon de sept

observations à partir d'une population de loi normale avec moyenne 0 et variance 1, notée $N(0,1)$ qui forment le groupe 0. Le groupe 1 est composé de trois observations issues d'une loi normale avec moyenne 0 et variance 5. Les observations de ce groupe sont visuellement plus distants de 0 que les observations du groupe 0 afin de rendre cet exemple plus intéressant. Nous avons obtenu l'échantillon suivant :

$$\begin{aligned} x &= (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}) \\ &= (-0{,}539; -0{,}178; -0{,}529; -0{,}513; 0{,}289; 0{,}572; 0{,}052; -3{,}814; 3{,}576; 4{,}003) \end{aligned}$$

où x_1,x_2,x_3,x_4,x_5,x_6 et $x_7 \sim N(0,1)$ et x_8,x_9 et x_{10} sont les trois observations avec une plus grande variance. Les dix premières lignes et colonnes du tableau 1.2.1 indiquent les distances de Mahalanobis, telles que définies par l'équation (1.2.1), en prenant le carré des observations. Nous utilisons le carré des observations car pour une loi normale dont la moyenne est égale à 0, nous pourrions estimer la variance en prenant le carré des observations. Ainsi, dans notre exemple, les observations -3,814 et 3,576 ont une distance de Mahalanobis petite et par conséquent, elles sont classées dans le même groupe.

Étant donné que la taille de notre échantillon est très petite, nous choisissons $\nu = 3$ pour estimer la densité au point x_i, $i = 1, \ldots, 10$. Le tableau 1.2.2 indique les probabilités *a posteriori* d'appartenance pour le groupe 0 et pour le groupe 1 calculées à l'aide de l'équation (1.2.3). Le fait d'avoir choisi les trois observations du groupe 1 de manière à ce qu'elles soient éloignées des observations du groupe 0, nous permet d'avoir un bon taux de classification.

Une fois les probabilités *a posteriori* calculées, nous pouvons classer une nouvelle observation en calculant de nouveau les probabilités d'appartenance pour chacun des deux groupes et ainsi classer cette observation dans le groupe où la

Tableau 1.2.1: *Distances de Mahalanobis*

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	y
x_1	0,000	0,106	0,004	0,011	1,084	0,015	0,117	5,817	5,100	6,420	0,103
x_2		0,000	0,101	0,094	0,021	0,121	0,012	5,923	5,205	6,526	0,003
x_3			0,000	0,007	0,080	0,019	0,113	5,822	5,104	6,425	0,099
x_4				0,000	0,073	0,026	0,106	5,828	5,111	6,431	0,092
x_5					0,000	0,099	0,033	5,902	5,184	6,505	0,018
x_6						0,000	0,132	5,802	5,085	6,405	0,118
x_7							0,000	5,935	5,217	6,538	0,015
x_8								0,000	0,718	0,603	5,920
x_9									0,000	1,321	5,203
x_{10}										0,000	6,523
y											0.000

Tableau 1.2.2: *Probabilités a posteriori*

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	y
$p(0\|x_i)$	1	1	1	1	1	1	1	0	0	0	1
$p(1\|x_i)$	0	0	0	0	0	0	0	1	1	1	0

probabilité est la plus grande. Nous avons donc généré une nouvelle observation $y = 0{,}196$ à partir d'une loi normale $N(0,1)$ et nous avons calculé les distances de Mahalanobis entre cette observation et chacune des dix observations x_i en prenant le carré des observations, afin de déterminer les 3 plus proches voisins de l'observation y. La dernière colonne du tableau 1.2.1 indique ces distances. Les probabilités *a posteriori* d'appartenance pour l'observation y sont inscrites dans la dernière colonne du tableau 1.2.2. Ainsi, l'observation y est classée dans le groupe 0.

1.3. Modèle contaminé et classification

Étant donné que nous sommes en présence de deux groupes dans notre jeu de données, nous utilisons un modèle de mélange de deux distributions tel que présenté dans Titterington, Smith et Makov (1985). Posons y, la variable aléatoire représentant la charge virale. Alors la fonction de densité de cette variable est :

$$f(y) = \varepsilon f_0(y) + (1-\varepsilon) f_1(y), 0 \leq \varepsilon \leq 1,$$

où $f_0(y)$ est la fonction de densité pour les patients dans le groupe 0 et $f_1(y)$ pour les patients dans le groupe 1. La fonction de densité $f(y)$ est alors un mélange de ces deux densités. Si ε est proche de 1, $f_0(y)$ est souvent appelée la fonction de densité d'intérêt et $f_1(y)$, la fonction de densité contaminante. Les paramètres de la densité $f_1(y)$ sont ajustés de telle sorte que les observations appartenant au groupe 1 soient des valeurs extrêmes ou inhabituelles par rapport à la densité $f_0(y)$. Dans notre cas, nous avons vu à la section 1.1 que les deux densités ont la même moyenne mais des variances différentes telles que $\sigma_1^2 > \sigma_0^2$. Nous ajustons donc notre modèle pour tenir compte de la dispersion de la variance du groupe 1 par rapport à celle du groupe 0. Afin de simplifier la présentation, nous supposons ici que les deux densités sont normales.

En définissant les paramètres pour chacune des densités, nous obtenons :

$$f(y|\underset{\sim}{\eta}) = \varepsilon f_0(y|\underset{\sim}{\eta_0}) + (1-\varepsilon) f_1(y|\underset{\sim}{\eta_1}),$$

où $\underset{\sim}{\eta} = (\varepsilon, \underset{\sim}{\eta_0}, \underset{\sim}{\eta_1})$ représente le vecteur de paramètres à estimer.

Une fois les paramètres estimés (voir à la section suivante pour la discussion sur la méthode d'estimation proposée), nous serons en mesure de classifier les patients dans leur groupe respectif. En effet, pour chacun des sujets, nous évaluons

les densités $f_j(\cdot|\underset{\sim}{\eta_j})$ pour $j = 0, 1$. La classification se fait à l'aide de l'algorithme suivant :

- si $f_0(y_i|\underset{\sim}{\eta_0}) > f_1(y_i|\underset{\sim}{\eta_1})$, alors le patient i est classé dans le groupe 0;
- si $f_0(y_i|\underset{\sim}{\eta_0}) \leq f_1(y_i|\underset{\sim}{\eta_1})$, alors le patient i est classé dans le groupe 1.

Étant donné que l'appartenance aux groupes est connue pour tous les patients de notre échantillon, nous pourrons aussi calculer le taux de mauvaise classification ainsi obtenu.

Nous avons supposé que les deux densités f_0 et f_1 sont deux lois normales univariées avec la même moyenne μ mais avec des variances différentes telles que $\sigma_1^2 > \sigma_0^2$ où $\sigma_1^2 = k\sigma_0^2$, $k > 1$. La densité de la charge virale y devient alors :

$$f(y|\mu, \sigma_0^2, k) = \varepsilon f_0(y|\mu, \sigma_0^2) + (1 - \varepsilon) f_1(y|\mu, k\sigma_0^2)$$

ou bien

$$\begin{aligned} f(y|\mu, \sigma_0^2, k) &= \varepsilon(2\pi\sigma_0^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_0^2}(y-\mu)^2\right] \\ &\quad +(1-\varepsilon)(2\pi k\sigma_0^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2k\sigma_0^2}(y-\mu)^2\right]. \end{aligned}$$

Les paramètres $\varepsilon, \mu, \sigma_0^2$ et k sont estimés de manière à maximiser la fonction de vraisemblance. La fonction de vraisemblance du modèle contaminé pour la charge virale a la forme suivante :

$$L(y|\underset{\sim}{\eta}) = \prod_{i=1}^{N} [\varepsilon f_0(y_i|\mu, \sigma_0^2) + (1 - \varepsilon) f_1(y_i|\mu, k\sigma_0^2)],$$

où f_0 est la densité pour le groupe 0, f_1 est la densité pour le groupe 1 et $\underset{\sim}{\eta}$ est le vecteur de paramètres $(\varepsilon, \mu, \sigma_0^2, k)$. Nous maximiserons plutôt le logarithme de la fonction de vraisemblance $l(y|\underset{\sim}{\eta})$; mais, étant donné qu'il est difficile de maximiser cette fonction par rapport au vecteur $\underset{\sim}{\eta}$ directement, nous utilisons un algorithme

introduit par Dempster, Laird et Rubin (1977) pour calculer les estimateurs des paramètres $\varepsilon, \mu, \sigma_0^2$ et k.

1.4. ALGORITHME EM

Nous supposons maintenant que l'appartenance des patients aux groupes n'est pas connue. Introduisons une variable latente Z qui suit une distribution de Bernoulli de paramètre $1-\varepsilon$, notée $Z \sim Bin(1, 1-\varepsilon)$ qui indique de quel groupe proviennent les y_i. La variable Z prendra la valeur 0 si le i^e patient provient du groupe 0 avec probabilité ε et la valeur 1 s'il provient du groupe 1 avec probabilité 1-ε.

La densité conditionnelle de Y_i étant donné $\underset{\sim}{\eta}$ et Z_i, $i = 1, \ldots, N$, notée $Y_i|\underset{\sim}{\eta}, Z_i$ s'écrit alors comme :

$$f(y_i|\underset{\sim}{\eta}, z_i) = f_0(y_i|\mu, \sigma_0^2)^{1-z_i} f_1(y_i|\mu, k\sigma_0^2)^{z_i}.$$

Trouvons maintenant la distribution de Z_i étant donné Y_i et $\underset{\sim}{\eta}$.

Théorème 1.4.1

La distribution de $Z_i|Y_i, \underset{\sim}{\eta}$ est une densité de Bernoulli de paramètre

$$\varepsilon_i = \frac{(1-\varepsilon) f_1(y_i|\mu, k\sigma_0^2)}{\varepsilon f_0(y_i|\mu, \sigma_0^2) + (1-\varepsilon) f_1(y_i|\mu, k\sigma_0^2)}, \tag{1.4.1}$$

notée $Z_i|Y_i, \underset{\sim}{\eta} \sim Bin\left(1, \frac{(1-\varepsilon) f_1(y_i|\mu, k\sigma_0^2)}{\varepsilon f_0(y_i|\mu, \sigma_0^2) + (1-\varepsilon) f_1(y_i|\mu, k\sigma_0^2)}\right)$.

Démonstration

La distribution de Z_i étant donné Y_i et $\underset{\sim}{\eta}$ se trouve comme suit :

$$\begin{aligned} f(z_i|y_i,\underset{\sim}{\eta}) &= \frac{f(y_i|z_i,\underset{\sim}{\eta})f(z_i|\underset{\sim}{\eta})}{f(y_i|\underset{\sim}{\eta})} \\ &= \frac{\varepsilon^{1-z_i}(1-\varepsilon)^{z_i} f_0(y_i|\mu,\sigma_0^2)^{1-z_i} f_1(y_i|\mu,k\sigma_0^2)^{z_i}}{\varepsilon f_0(y_i|\mu,\sigma_0^2)+(1-\varepsilon)f_1(y_i|\mu,k\sigma_0^2)}. \end{aligned}$$

Voir Desgagné (1998) section 2.2.2 pour plus de détails. □

L'algorithme EM tel qu'expliqué par Titterington, Smith et Makov (1985) et par Tanner (1993) génère à partir d'une valeur initiale $\underset{\sim}{\eta}^{(0)}$ une série d'estimateurs $\{\underset{\sim}{\eta}^{(m)}\}_{m\geq 1}$. Cet algorithme est une méthode itérative comportant deux étapes à chaque itération :

i. Étape E (étape de l'espérance) :
Nous évaluons $Q(\underset{\sim}{\eta},\underset{\sim}{\eta}^{(m)}) = \mathbb{E}^{Z|Y,\underset{\sim}{\eta}^{(m)}}[\log[L(x|\underset{\sim}{\eta})]]$ où $\mathbb{E}^{Z|Y,\underset{\sim}{\eta}^{(m)}}$ est l'espérance conditionnelle de Z étant donné Y et $\underset{\sim}{\eta}^{(m)}$ et x est le vecteur des observations complètes c'est-à-dire $x=(y,z)$ et $\log[L(x|\underset{\sim}{\eta})]$ a la forme suivante :

$$\begin{aligned} \log[L(x|\underset{\sim}{\eta})] &= \log\left[\prod_{i=1}^{N} f_0(y_i|\mu,\sigma_0^2)^{1-z_i} f_1(y_i|\mu,k\sigma_0^2)^{z_i}(1-\varepsilon)^{z_i}\varepsilon^{1-z_i}\right] \\ &= \sum_{i=1}^{N}\big[(1-z_i)\log f_0(y_i|\mu,\sigma_0^2)+z_i\log f_1(y_i|\mu,k\sigma_0^2) \\ &\quad +z_i\log(1-\varepsilon)+(1-z_i)\log\varepsilon\big] \end{aligned}$$

ii. Étape M (étape de maximisation) :
Nous trouvons $\underset{\sim}{\eta}=\underset{\sim}{\eta}^{(m+1)}$ qui maximise $Q(\underset{\sim}{\eta},\underset{\sim}{\eta}^{(m)})$.

L'algorithme est répété jusqu'à ce que $\|\underset{\sim}{\eta}^{(m+1)} - \underset{\sim}{\eta}^{(m)}\|$ soit suffisamment petit (dans notre cas, nous arrêtons lorsque $\|\underset{\sim}{\eta}^{(m+1)} - \underset{\sim}{\eta}^{(m)}\| < 0{,}01$) où $\|\cdot\|$ représente la norme euclidienne donnée par la définition suivante.

Définition 1.4.1

Soit $x = (x_1, \ldots, x_n)$ un vecteur dans $\mathbb{R}^n$. Alors $\|x\| = \sqrt{x^t x} = \left(\sum_{i=1}^{N} x_i^2\right)^{\frac{1}{2}}$ est appelé la norme de x.

Appliquons cet algorithme à la fonction de vraisemblance $l(x|\underset{\sim}{\eta})$. Tout d'abord, calculons $Q(\underset{\sim}{\eta}, \underset{\sim}{\eta}^{(m)})$ pour l'étape E.

Théorème 1.4.2

À l'étape E, $Q(\underset{\sim}{\eta}, \underset{\sim}{\eta}^{(m)})$ se calcule de la façon suivante :

$$\begin{aligned} Q(\underset{\sim}{\eta}, \underset{\sim}{\eta}^{(m)}) &= \sum_{i=1}^{N} \Big[(1-\varepsilon_i^{(m)})[\log f_0(y_i|\mu, \sigma_0^2) + \log\varepsilon] \\ &\quad + \varepsilon_i^{(m)}[\log f_1(y_i|\mu, k\sigma_0^2) + \log(1-\varepsilon)]\Big] \end{aligned}$$

où

$$\varepsilon_i^{(m)} = \frac{(1-\varepsilon^{(m)}) f_1(y_i|\mu^{(m)}, k^{(m)}\sigma_0^{2(m)})}{\varepsilon^{(m)} f_0(y_i|\mu^{(m)}, \sigma_0^{2(m)}) + (1-\varepsilon^{(m)}) f_1(y_i|\mu^{(m)}, k^{(m)}\sigma_0^{2(m)})}. \tag{1.4.2}$$

Démonstration

En utilisant le théorème 1.4.1, nous trouvons

$$\begin{aligned} Q(\underset{\sim}{\eta}, \underset{\sim}{\eta}^{(m)}) &= \int_Z \log[L(x|\underset{\sim}{\eta})] p(Z|\underset{\sim}{\eta}^{(m)}, Y) dZ \\ &= \mathbb{E}^{Z|Y,\underset{\sim}{\eta}^{(m)}} \left[\sum_{i=1}^{N} [(1-z_i)\log f_0(y_i|\mu, \sigma_0^2) + z_i \log f_1(y_i|\mu, k\sigma_0^2) \right. \\ &\quad \left. + z_i \log(1-\varepsilon) + (1-z_i)\log\varepsilon]\right] \\ &= \sum_{i=1}^{N} \Big[(1-\varepsilon_i^{(m)})\log f_0(y_i|\mu, \sigma_0^2) + \varepsilon_i^{(m)} \log f_1(y_i|\mu, k\sigma_0^2) \end{aligned}$$

$$+ \varepsilon_i^{(m)} \log(1-\varepsilon) + (1-\varepsilon_i^{(m)}) \log \varepsilon \Big]$$

$$= \sum_{i=1}^{N} \Big[(1-\varepsilon_i^{(m)})[\log f_0(y_i|\mu, \sigma_0^2) + \log \varepsilon]$$

$$+ \varepsilon_i^{(m)} [\log f_1(y_i|\mu, k\sigma_0^2) + \log(1-\varepsilon)] \Big]$$

où $\varepsilon_i^{(m)}$ est définie par l'équation (1.4.2). □

Puis, pour l'étape M, trouvons $\underset{\sim}{\eta}$ qui maximise $Q(\underset{\sim}{\eta}, \underset{\sim}{\eta}^{(m)})$.

Théorème 1.4.3

Les estimateurs qui maximisent $Q(\underset{\sim}{\eta}, \underset{\sim}{\eta}^{(m)})$ sont les suivants :

$$\varepsilon^{(m+1)} = \frac{\sum_{i=1}^{N}(1-\varepsilon_i^{(m)})}{N}, \tag{1.4.3}$$

$$\mu^{(m+1)} = \frac{\sum_{i=1}^{N}(1-\frac{(k^{(m)}-1)}{k^{(m)}}\varepsilon_i^{(m)})y_i}{\sum_{i=1}^{N}(1-\frac{(k^{(m)}-1)}{k^{(m)}}\varepsilon_i^{(m)})}, \tag{1.4.4}$$

$$\sigma_0^{2(m+1)} = \frac{\sum_{i=1}^{N}(1-\frac{(k^{(m)}-1)}{k^{(m)}}\varepsilon_i^{(m)})(y_i-\mu^{(m+1)})^2}{N}, \tag{1.4.5}$$

$$k^{(m+1)} = \frac{\sum_{i=1}^{N}\varepsilon_i^{(m)}(y_i-\mu^{(m+1)})^2}{\sigma_0^{2(m+1)}\sum_{i=1}^{N}\varepsilon_i^{(m)}}, \tag{1.4.6}$$

où $\varepsilon_i^{(m)}$ est définie par l'équation (1.4.2).

Voir Desgagné (1998) pour une démonstration de ce théorème.

Afin d'illustrer cet algorithme, revenons à notre exemple avec les sept observations provenant d'une population de loi $N(0,1)$ et les trois observations d'une

population de densité normale avec une variance plus grande que 1. Pour commencer, nous devons spécifier les valeurs de départ du vecteur $\underset{\sim}{\eta} = (\varepsilon, \mu, \sigma_0^2, k\sigma_0^2)$. Ainsi, nous avons choisi $\varepsilon^{(0)} = 0{,}7$ (7 observations sur 10 font partie du groupe 0), $\mu^{(0)} = 0{,}292$ qui est la moyenne des 10 observations prises ensemble, $\sigma_0^{2(0)} = 0{,}196$ qui est la variance du groupe 0 et $k^{(0)} = 98{,}523$ qui est le rapport de la variance du groupe 1 sur la variance du groupe 0. À l'aide des équations (1.4.3) à (1.4.6) et après cinq itérations, nous trouvons les estimateurs suivants :

$$\hat{\varepsilon} = 0{,}618;$$

$$\hat{\mu} = -0{,}127;$$

$$\hat{\sigma}_0^2 = 0{,}159;$$

$$\hat{k} = 73{,}540.$$

Ces estimateurs ont été calculés à l'aide du logiciel Splus en utilisant les programmes en annexe.

Par conséquent, la fonction de densité peut alors s'écrire comme :

$$f(x) = 0{,}618 f_0(x) + 0{,}382 f_1(x),$$

où $f_0(x)$ est une loi normale de paramètres $-0{,}127$ et $0{,}159$ tandis que $f_1(x)$ est une loi normale de paramètres $-0{,}127$ et $11{,}693$. Ces chiffres ont été calculés avec l'algorithme EM.

De plus, les ε_i, $i = 1, \ldots, 10$, tels que définis par l'équation (1.4.1) sont estimés afin de classifier chacune des observations. Chaque ε_i représente la probabilité d'appartenance au groupe 1 étant donné Y_i et $\underset{\sim}{\eta}$. Ainsi, si $\varepsilon_i < 0{,}5$, alors l'observation est classée dans le groupe 0 et si $\varepsilon_i \geq 0{,}5$, alors elle est classée dans le groupe 1. Nous obtenons le vecteur suivant :

$$(0{,}109; 0{,}068; 0{,}106; 0{,}102; 0{,}110; 0{,}247; 0{,}074; 1{,}000; 1{,}000; 1{,}000)$$

Ainsi, les sept premières observations sont classées dans le groupe 0 et les trois dernières, dans le groupe 1.

Dans ce premier chapitre, nous avons présenté le jeu de données que nous appliquerons aux méthodes d'analyse des chapitres suivants. Puis, à l'aide d'un exemple, nous avons vu comment classifier des observations en deux groupes à l'aide de l'analyse discriminante. De plus, nous avons introduit le modèle de mélange de deux lois et nous avons expliqué comment estimer les paramètres d'un tel modèle à l'aide de l'algorithme EM.

Chapitre 2

SPLINES

Dans ce second chapitre, nous tentons de trouver une fonction qui modélise la charge virale en fonction de la variable temps. Soient Y la variable charge virale et T la variable temps. Nous voulons trouver une fonction f telle que $Y = f(T) + \varepsilon$. Afin d'avoir un aperçu de la forme de cette fonction, les figures 2.0.1 et 2.0.2 illustrent la charge virale en fonction du temps pour deux patients représentatifs pour chacun des groupes (les patients numéro 11 et 14 pour le groupe 1 et les patients numéro 56 et 58 pour le groupe 0). Étant donné l'allure de ces graphiques, nous estimons la fonction à l'aide d'une spline. Dans les pages qui suivent, nous décrivons la spline de lissage qui permet d'estimer la fonction f et nous calculons les coefficients de cette spline pour chacun des patients. Puis, nous effectuons une analyse discriminante sur ces coefficients afin de déterminer si ces coefficients permettent de classer les patients dans leur groupe respectif. Nous débutons d'abord avec un modèle classique et par la suite, nous trouvons la spline de façon bayésienne. Enfin, nous comparons les taux d'erreur obtenus avec l'analyse discriminante pour les deux modèles.

Figure 2.0.1: *Graphique de la charge virale en fonction du temps pour 2 patients du groupe 0*

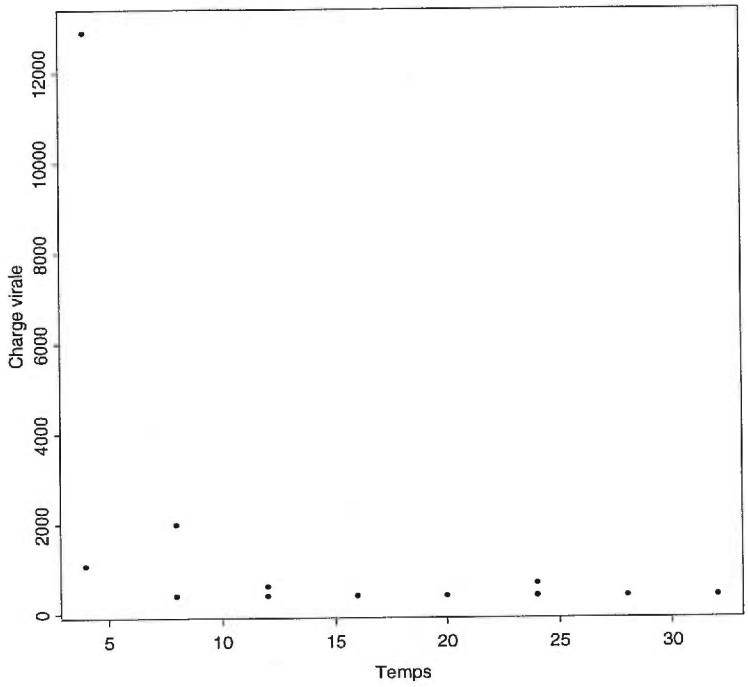
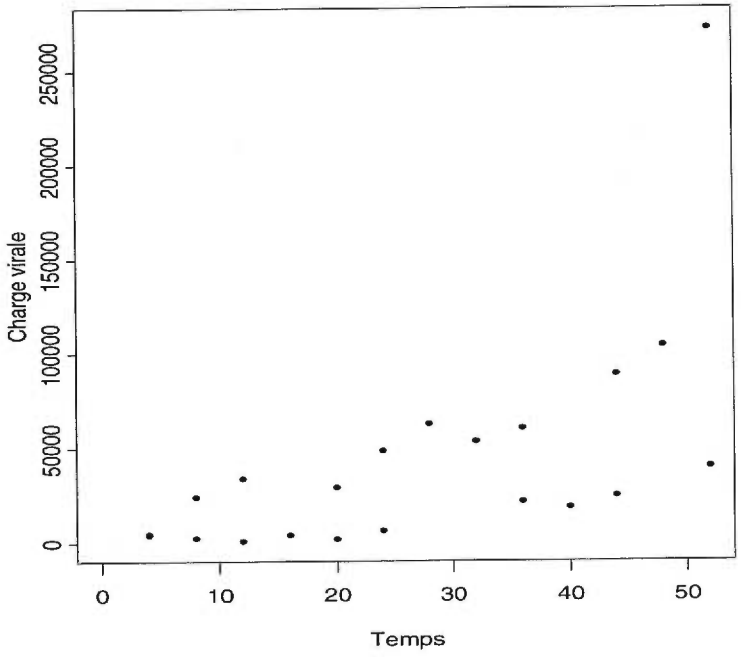


Figure 2.0.2: *Graphique de la charge virale en fonction du temps pour 2 patients du groupe 1*



2.1. Splines classiques

2.1.1. Splines de lissage

Soient (t_i, y_i), $i = 1, \ldots, n$, un ensemble fini de points générés par le modèle suivant :

$$y_i = f(t_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1.1}$$

où les ε_i, $i = 1, \ldots, n$ sont indépendantes et identiquement distribuées (i.i.d.) de moyenne 0 et de variance commune σ^2 inconnue et où les t_i, $i = 1, \ldots, n$ et ε_i, $i = 1, \ldots, n$ sont indépendantes. Supposons que la fonction f de l'équation (2.1.1) appartient à $W_2^m[a, b]$, où

$$W_2^m[a, b] = \{f : f^{(k)} \text{ absolument continue sur } [a, b],\ k = 0, \ldots, m-1 \quad \text{et} f^{(m)} \in L_2[a, b]\} \tag{2.1.2}$$

et

$$L_2[a, b] = \{f : \int_a^b [f(x)]^2 dx < \infty\}. \tag{2.1.3}$$

où $f^{(k)}$ est la k^e dérivée de f avec $f^{(0)}$ égale à la fonction elle-même.

La fonction f est estimée par $f_{n,\lambda}$ qui est la solution minimisant la fonction de perte quadratique pénalisée suivante :

$$\min_{f \in W_2^m[a,b]} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(t_i))^2 + \lambda \int_a^b [f^{(m)}(t)]^2 dt. \tag{2.1.4}$$

La fonction $f_{n,\lambda}$ est une spline de lissage définie par la définition 2.1.1 d'ordre $2m$. Le paramètre λ de l'équation (2.1.4) contrôle le degré de lissage de la fonction $f_{n,\lambda}$. La mesure standard de la qualité de l'ajustement est calculée par $n^{-1} \sum_{i=1}^{n} (y_i - f(t_i))^2$ et le degré de lissage est donné par $\int_a^b [f^{(m)}(t)]^2 dt$. La proposition 2.1.1

présente le résultat obtenu par Eubank (1988) qui minimise l'équation (2.1.4) pour $f \in W_2^m[a,b]$.

Définition 2.1.1

Une spline de lissage S d'ordre $k-1$ est une fonction appartenant à $\Psi_{k,t}$ qui peut s'écrire sous la forme

$$S(t) = \sum_{i=0}^{k-1} \alpha_i t^i + \sum_{i=1}^{n} \beta_i (t - t_i)_+^{k-1},$$

où $u_+ = u$ si $u > 0$ et 0 sinon. L'ensemble $\Psi_{k,t}$ représente toutes les fonctions polynomiales par morceaux dont les $k-2$ premières dérivées sont continues. Cet ensemble peut s'écrire sous la forme $\Psi_{k,t} = P_{k,t} \cap C^{k-2}[a,b]$ où

$$P_{k,t} = \{f : f(x) = p_i(x)\chi_{I_i}(x) \text{ où } p_i \in P_k \text{ pour } i = 1, \ldots, n\}$$

est l'ensemble des fonctions polynomiales par morceaux P_k de degré $k-1$ au point de cassure t et où χ_{I_i} est la fonction indicatrice pour l'intervalle $I_i = [t_i, t_{i+1})$, $i = 1, \ldots, n-1$ et

$$C^{k-2}[a,b] = \{f : \text{la } j^e \text{ dérivée de } f, \text{ pour } j = 1, \ldots, k-2, \text{ existe et est continue en chaque } t \in [a,b]\}.$$

2.1.2. Développement du lisseur

Revenons au modèle donné par l'équation (2.1.1). Supposons que la fonction f appartient à $W_2^m[a,b]$ tel que défini par les équations (2.1.2) et (2.1.3). Eubank (1988) montre que l'équation (2.1.4) peut être résolue en solutionnant $\min n^{-1} \sum_{i=1}^{n} (y_i - f(t_i))^2$, sous la contrainte $\int_a^b [f^{(m)}(t)]^2 dt \leq c$, où c est une constante dépendante de m.

La démonstration de la proposition suivante se trouve dans Eubank(1988).

Proposition 2.1.1

Si $n \geq m$, alors la solution de l'équation (2.1.4) est de la forme

$$f_{n,\lambda}(t_i) = \sum_{j=1}^{m} \theta_{\lambda j} x_j,$$

où $\theta_\lambda = (\theta_{\lambda 1}, \ldots, \theta_{\lambda m})^t$ est la solution de l'équation

$$(X^t X + n\lambda\Omega)\theta_\lambda = X^t Y,$$

avec $X = \{X_j(t_i)\}_{\substack{i=1,\ldots,n \\ j=1,\ldots,m}}$ et $\Omega = \{\int_a^b X_i^{(m)}(t) X_j^{(m)}(t)dt\}_{i,j=1,\ldots,m}$ où $X_j(t_i)$ peut prendre plusieurs formes.

2.1.3. Approximation par une série de Taylor tronquée

En choisissant $X_j(t_i) = t_i^{j-1}$ tel que proposé par Wahba et Kimeldorf (1971), nous pouvons écrire la solution de l'équation (2.1.4) comme une série de Taylor tronquée

$$f(t) = \sum_{i=0}^{k-1} \frac{t^i}{i!}\theta_i, \tag{2.1.5}$$

où $\theta_j = f^{(j)}(0)$ est la j^e dérivée de f au point 0. Ainsi, en dérivant m fois la fonction f, nous obtenons

$$f^{(m)}(t) = \sum_{i=m}^{k-1} \frac{t^{i-m}}{(i-m)!}\theta_i, \quad m \leq k-1,$$

et donc

$$\int_0^1 [f^{(m)}(t)]^2 dt = \sum_{i=m}^{k-1}\sum_{j=m}^{k-1} \frac{\theta_i\theta_j}{(i+j-2m+1)(i-m)!(j-m)!}.$$

En utilisant l'équation (2.1.5), l'équation (2.1.4) peut être approximée par

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{i=0}^{k-1}\frac{t^i}{i!}\theta_i\right)^2 + \lambda\sum_{i=m}^{k-1}\sum_{j=m}^{k-1}\frac{\theta_i\theta_j}{(i+j-2m+1)(i-m)!(j-m)!} \qquad (2.1.6)$$

ou bien sous forme matricielle

$$\frac{1}{n}(Y - X\theta)^t(Y - X\theta) + \lambda\theta^t\Omega\theta,$$

où

$$\begin{aligned} X &= (X_1, X_2, \ldots, X_n)^t, \\ X_i &= \left(1, t_i, \frac{t_i^2}{2!}, \ldots, \frac{t_i^{k-1}}{(k-1)!}\right)^t, \\ \theta &= (\theta_0, \theta_1, \ldots, \theta_{k-1})^t \end{aligned}$$

pour $i = 1, 2, \ldots, n$ et Ω est une matrice $k \times k$ telle que $\Omega = \{\omega_{ij}\}_{i,j=1,2,\ldots,k}$

où

$$\omega_{ii} = \begin{cases} \frac{1}{[2(i-1)-2m+1][((i-1)-m)!]^2} & \text{si } i = m+1, \ldots, k, \\ 0 & \text{sinon,} \end{cases}$$

$$\omega_{ij} = \begin{cases} \frac{1}{2}\frac{1}{[(i-1)+(j-1)-2m+1][(i-1)-m)]![(j-1)-m]!} & \text{si } i = m+1, \ldots, k, \\ & \text{et } j = i+1, \ldots, k, \\ 0 & \text{sinon.} \end{cases}$$

La proposition suivante, dont la démonstration se trouve dans Bennaghmouch (1992), nous donne la forme du lisseur $f_{n,\lambda}$.

Proposition 2.1.2

Pour $m < k-1$, le minimum de l'équation (2.1.6) est de la forme

$$f_{n,\lambda} = X\hat{\theta_\lambda},$$

où $\hat{\theta_\lambda} = (X^tX + n\lambda\Omega)^{-1}X^tY$.

2.1.4. Paramètre de lissage

Afin de choisir le paramètre λ, nous appliquons l'algorithme proposé par Wahba et Wold (1975). Cet algorithme trouve la valeur de λ qui minimise la fonction suivante :

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_\lambda^{[i]}(t_i))^2, \tag{2.1.7}$$

où $f_\lambda^{[i]}$ est la spline ajustée pour l'ensemble des observations en enlevant la i^e observation. Cette méthode est nommée validation croisée ordinaire (CV). Étant donné que notre fonction $f_{n,\lambda}$ est linéaire en chaque observation, le théorème 4.2.1 de Wahba (1990) nous permet d'écrire l'équation (2.1.7) comme :

$$\begin{aligned} CV(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - f_\lambda^{[i]}(t_i))^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - f_\lambda(t_i)}{1 - A_\lambda(i,i)}\right]^2 \end{aligned} \tag{2.1.8}$$

où $A_\lambda(i,i)$ est l'élément (i,i) de la matrice A_λ qui est l'unique matrice satisfaisant :

$$\begin{pmatrix} f_\lambda(t_1) \\ \vdots \\ f_\lambda(t_n) \end{pmatrix} = A(\lambda) \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

c'est-à-dire $A(\lambda) = X(X^tX + n\lambda\Omega)^{-1}X^t$.

Nous choisissons la valeur de λ de façon à minimiser l'équation (2.1.8). Pour ce faire, nous utilisons la méthode de Newton-Raphson telle que programmée dans le logiciel Splus.

Appliquons ce que nous venons de voir à notre problème de charge virale.

Afin de déterminer le degré $k-1$ de notre spline, nous avons fait varier la valeur k de 2 à 10 et nous avons calculé la fonction de validation croisée (CV)

Tableau 2.1.1: *Tableau pour le choix de k*

k	2	3	4	5	6	7	8	9	10
groupe 0	15	4	2	0	2	1	1	0	1
groupe 1	16	6	1	2	3	2	2	0	0
total	31	10	3	2	5	3	3	0	1

Tableau 2.1.2: *Tableau pour un deuxième choix de k*

k	2	3	4	5	6	7	8	9	10
groupe 0	0	11	3	5	1	3	3	0	0
groupe 1	0	15	6	1	3	1	3	3	0
total	0	26	9	6	4	4	6	3	0

minimale par rapport à λ pour chacun des patients. Le tableau 2.1.1 présente le nombre de patients pour chacune des valeurs de k pour lesquelles cette valeur donne une CV minimale. Nous avons aussi trouvé une deuxième valeur du k qui donne la valeur la plus proche du minimum de CV. Les résultats sont dans le tableau 2.1.2. En regardant les deux tableaux, nous remarquons que les valeurs pour k qui donnent le plus souvent une CV minimale sont 2 et 3. Nous avons donc calculé les coefficients $\theta = (\theta_0, \theta_1)$ pour $k = 2$ et $\theta = (\theta_0, \theta_1, \theta_2)$ pour $k = 3$ pour chacun des patients.

2.1.5. Analyse discriminante des coefficients pour la charge virale

Une fois les coefficients $\theta = (\theta_0, \theta_1, \theta_2)$ calculés pour chacun des 58 patients, nous effectuons une analyse discriminante (voir à la section 1.2 pour une brève explication de cette procédure) afin de déterminer le taux de mauvaise classification obtenu à l'aide des coefficients. Nous avons fixé les probabilités *a priori* à $\frac{26}{58}$ pour le groupe 0, étant donné qu'il y a 26 patients dans le groupe 0 et à $\frac{32}{58}$ pour

Tableau 2.1.3: *Tableau de classification de l'analyse discriminante pour la spline classique avec $k = 2$*

Réel / Classé	0	1	Total
0	24	2	26
1	3	29	32
Total	30	28	58

Tableau 2.1.4: *Tableau de classification de l'analyse discriminante pour la spline classique avec $k = 3$*

Réel / Classé	0	1	Total
0	25	1	26
1	3	29	32
Total	30	28	58

le groupe 1, puisqu'il y a 32 patients dans ce dernier groupe. Après avoir essayé plusieurs valeurs pour le nombre de plus proches voisins ν, les plus petits taux d'erreur de classification que nous avons obtenus sont de 0,069 pour une valeur de $k = 3$ et de 0,086 pour $k = 2$. Notons que ces taux ont été obtenus en utilsant les 3 plus proches voisins. Cette valeur est conservée pour les splines bayésiennes afin de pouvoir comparer les deux sortes de spline. De plus, les tableaux de classification 2.1.3 et 2.1.4 présentent la façon dont les patients ont été classés pour $k = 2$ et $k = 3$ respectivement. Dans notre cas, les deux types d'erreur de classification peuvent être définies de la façon suivante: classer un patient malade (du groupe 1) dans le même groupe que les patients qui vont bien (groupe 0) et classer un patient qui va bien (groupe 0) dans le même groupe que ceux qui sont malades (groupe 1). D'un point de vue médical, il est plus grave de classer un patient du groupe 0 dans le groupe 1. Ceci est dû au fait que, dans le cas de patients

séropositifs, les médecins cherchent à trouver la meilleure combinaison de médicaments pour un patient donné permettant d'enrayer au maximum la charge virale. Lorsque cette combinaison est trouvée, le patient conserve celle-ci aussi longtemps que l'effet persiste. Par conséquent, si un patient a une petite charge virale mais qu'il est jugé malade, à la fin de l'étude le traitement sera interrompu malgré le fait que la combinaison de médicaments était la bonne. En regardant les tableaux 2.1.3 et 2.1.4, nous remarquons que pour une valeur de $k = 2$, un patient de plus est mal classé qu'en utilisant $k = 3$. Pour cette raison, nous conservons donc 3 pour la valeur du k dans le cas classique comme dans le cas bayésien. Les valeurs des coefficients $\theta = (\theta_0, \theta_1, \theta_2)$ pour $k = 3$ sont inscrites dans le tableau 2.1.5. Les figures 2.1.1 et 2.1.2 montrent les splines classiques et bayésiennes pour les mêmes deux patients des figures 2.0.1 et 2.0.2. Les splines classiques sont illustrées par des lignes continues et les splines bayésiennes (voir à la section suivante pour le développement de ce type de spline) par des lignes pointillées. Nous observons que pour un des patients du groupe 0, la spline classique ne réussit pas très bien à approximer la fonction. Par contre, la spline bayésienne approxime bien la fonction pour ce patient. À la section 2.3, nous discutons un peu plus de la classification lorsque nous comparons le modèle classique au modèle bayésien.

Tableau 2.1.5: Tableau des coefficients pour la spline classique

Coefficients de la spline classique				
Patient	Groupe	$\hat{\theta}_{i0}$	$\hat{\theta}_{i1}$	$\hat{\theta}_{i2}$
1	1	0,257	-1,557	9,202
2	1	265,624	255,017	-222,266
3	1	80,938	0,611	-48,826
4	1	2,305	-5,014	7,268
5	1	49,528	180,150	114,174

Coefficients de la spline classique				
Patient	Groupe	$\hat{\theta}_{i0}$	$\hat{\theta}_{i1}$	$\hat{\theta}_{i2}$
6	1	1,259	3,521	2,869
7	1	10,001	6,409	-6,499
8	1	28,332	30,762	-44,949
9	1	8,520	2,459	-9,767
10	1	13,371	31,261	34,694
11	1	32,425	-86,005	433,499
12	1	104,710	-750,356	1330,322
13	1	-6,581	111,813	-127,501
14	1	12,934	-38,747	112,494
15	1	-1,894	0,874	183,869
16	1	16,263	-97,926	297,497
17	1	-27,279	127,838	-121,725
18	1	131,413	308,903	-70,145
19	1	2,874	-3,742	3,355
20	1	3,864	5,606	12,728
21	1	2,292	-7,159	14,869
22	1	40,155	19,035	44,719
23	1	-0,047	3,941	34,279
24	1	27,057	141,559	500,940
25	1	39,490	43,117	-72,823
26	1	15,724	232,446	604,213
27	1	193,023	-138,070	6,136
28	1	3,037	11,133	-11,255
29	1	4,831	26,778	9,103
30	1	39,192	74,923	29,369
31	1	-148,483	809,816	-797,036
32	1	24,509	-18,421	24,648
33	0	7,345e-2	2,309	9,976
34	0	1,011	1,100e-2	-8,000e-2
35	0	1,841	-0,876	0,203

Coefficients de la spline classique				
Patient	Groupe	$\hat{\theta}_{i0}$	$\hat{\theta}_{i1}$	$\hat{\theta}_{i2}$
36	0	1,000	2,442e-14	-3,286e-14
37	0	1,011	-3,794e-2	5,111e-2
38	0	1,205	-0,206	7,268e-2
39	0	1,466	-1,226	1,471
40	0	1,014	-3,205e-2	3,043e-2
41	0	-0,328	0,727	8,203
42	0	1,960	-1,567	1,193
43	0	3,058	-4,215	3,949
44	0	39,354	123,873	-199,691
45	0	0,497	1,029	-0,516
46	0	-0,763	2,721	20,178
47	0	0,145	1,241	4,273
48	0	-0,386	-4,503e-2	7,298
49	0	0,588	1,750	-1,392
50	0	2,488	-2,677	2,241
51	0	1,655	-0,744	0,149
52	0	1,351	1,344	-2,327
53	0	1,000	1,976e-14	-1,421e-14
54	0	0,832	0,634	-0,634
55	0	1,000	-1,332e-15	1,332e-15
56	0	2,310	-3,680	4,727
57	0	1,175	-0,169	-6,030e-2
58	0	5,684	-2,323	-7,101

Il est à noter qu'une alternative à l'approximation par une série de Taylor pour la spline classique aurait été d'utiliser une spline de type B telle que définie par De Boor (1978) au chapitre 9.

Figure 2.1.1: *Graphique de la spline classique et de la spline bayesienne de la charge virale pour 2 patients du groupe 0*

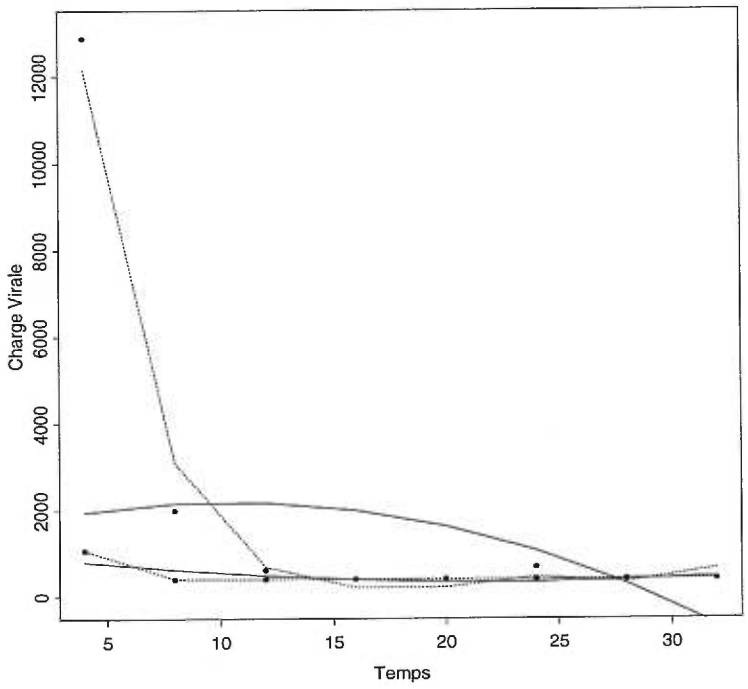
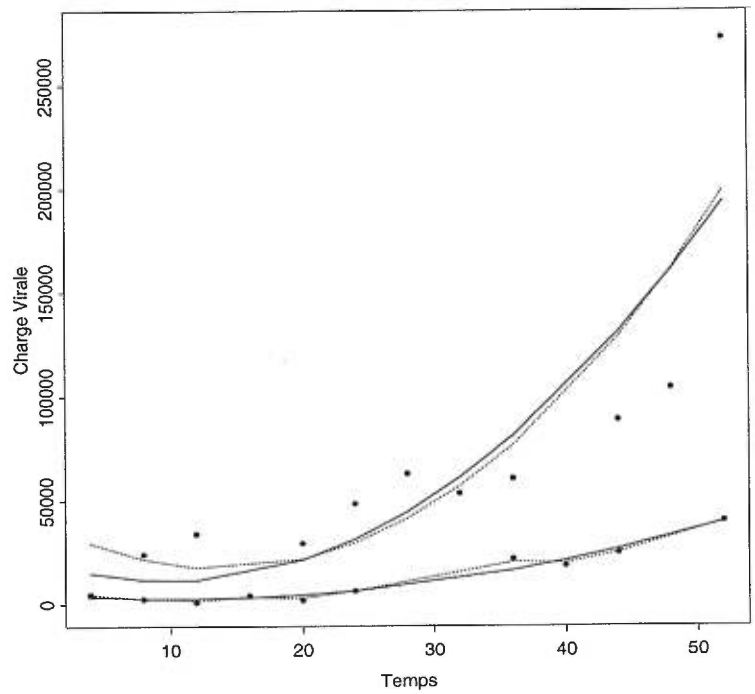


Figure 2.1.2: *Graphique de la spline classique et de la spline bayesienne de la charge virale pour 2 patients du groupe 1*



2.2. Splines bayésiennes

2.2.1. Description du modèle bayésien

Reprenons le modèle défini par l'équation (2.1.1) pour les splines classiques. Soient (t_i, y_i), $i = 1, \ldots, n$, un ensemble fini de points générés par le modèle suivant :

$$y_i = f(t_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

où cette fois-ci $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ est issu d'une population multinormale de dimension n, avec moyenne $\underset{\sim}{0}$ et matrice de variance-covariance $\sigma^2 I$ notée $\varepsilon \sim N_n(\underset{\sim}{0}, \sigma^2 I)$ et où $t = (t_1, \ldots, t_n)$ et ε sont indépendants. Dans cette section, la fonction f est estimée en utilisant une approche bayésienne. En effet, nous utilisons l'information *a priori* sur f et les données observées pour trouver une approximation de la fonction f.

2.2.2. Splines ajustées de O'Hagan

La méthode discutée dans cette section pour estimer la fonction f a été développée par O'Hagan (1978). Supposons que f peut s'écrire comme :

$$f(t) = g(t)^t \beta(t), \tag{2.2.1}$$

où $g(t)$ est un vecteur d'ordre q avec g qui est une fonction pouvant prendre différentes formes et $\beta(t)$ est un vecteur d'ordre q. Dans cette approche, les coefficients β dépendent du temps, contrairement au cas précédent. Afin de pouvoir comparer les résultats avec ceux obtenus à l'aide de la spline classique, nous choisissons $g(t) = \left(1, t_i, \frac{t_i^2}{2!}, \ldots, \frac{t_i^{k-1}}{(k-1)!}\right)^t$ comme dans le modèle classique.

Étant donné que $\varepsilon \sim N_n(\underset{\sim}{0}, \sigma^2 I)$, la distribution de Y, étant donné t et $\beta(t)$, est multinormale de dimension n, de moyenne $g(t)^t\beta(t)$ et de matrice de variance-covariance $\sigma^2 I$ notée

$$Y|t, \beta(t) \sim N_n(g(t)^t\beta(t), \sigma^2 I).$$

Nous supposons que $\beta(t)$ et $\beta(t^*)$ sont fortement corrélés lorsque $|t - t^*|$ est petit car nous croyons que la courbe de régression est localement stable. Afin de simplifier le problème, nous supposons que l'information *a priori* sur $\beta(t)$ est la même pour toutes les valeurs de t. Plus particulièrement, la moyenne *a priori*

$$\mathbb{E}[\beta(t)|b_0] = b_0 \tag{2.2.2}$$

est indépendante de t et la matrice de variance-covariance de $\beta(t)$,

$$\mathbb{E}[(\beta(t) - b_0)(\beta(t^*) - b_0)^t] = \rho(|t - t^*|)B_0, \tag{2.2.3}$$

où $\rho(d)$ est une fonction monotone décroissante avec $0 \leq d < \infty$ et $\rho(0) = 1$, ne dépendant que de $|t - t^*|$ et $B_0 = \tau^2 I_n$ où n est le nombre d'observations pour un patient. (Nous discutons du choix de τ^2 à la section 2.2.3.) La distribution de $\beta(t)$ est donc un processus gaussien stationnaire de deuxième ordre dont les moments sont donnés par les équations (2.2.2) et (2.2.3). Rappelons qu'un processus stationnaire est défini de la façon suivante.

Définition 2.2.1

Un processus stationnaire de deuxième ordre $\beta(t)$ est un processus tel que la moyenne est indépendante de la variable t et tel que la variance ne dépend que du vecteur différence $(t - t^*)$ c'est-à-dire

$$\mathbb{E}[\beta(t)] = b_0 \text{ pour tout } t$$

et

$$Cov[\beta(t), \beta(t^*)] = h(t - t^*)$$

où h est une fonction monotone quelconque.

O'Hagan (1978) suggère la forme normale $\rho(d) = \exp\left(-d^2/2\sigma_\rho^2\right)$ si f est une fonction polynomiale. Nous choisissons plutôt $\rho_c(|t-t^*|) = \exp\left(-c|t-t^*|\right)$ pour simplifier les expressions (voir Angers et Delampady (1992)). Le choix du c est discuté un peu plus loin (voir section 2.2.3). Ainsi, plus la fonction $\rho(\cdot)$ décroît lentement, plus $\beta(t)$ devient stable. De plus, le modèle de régression défini par l'équation (2.2.1), avec une distribution *a priori* définie par les équations (2.2.2) et (2.2.3), est appelé modèle de régression localisée par O'Hagan (1978).

Trouvons maintenant la distribution *a posteriori* des $\beta(t)$.

Théorème 2.2.1

Les $\beta(t)$ ont une distribution a posteriori étant donné Y conjointe multinormale de moyenne

$$b_1(t) = \mathbb{E}[\beta(t)|y_1, \ldots, y_n, b_0] = S(t)^t A^{-1} y + Q(t)^t b_0, \tag{2.2.4}$$

et de matrice de variance-covariance

$$\begin{aligned} B_1(t,t^*) &= \mathbb{E}[(\beta(t) - b_1(t))(\beta(t^*) - b_1(t^*))^t | y_1, \ldots, y_n, b_0] \\ &= \rho(|t-t^*|)B_0 - S(t)^t A^{-1} S(t^*), \end{aligned}$$

où

$Q(t) = I_q - GA^{-1}S(t)$ *est une matrice* $q \times q$,

$y = (y_1, y_2, \ldots, y_n)^t$ *est un vecteur de longueur* n,

$G = (g(t_1), g(t_2), \ldots, g(t_n))$ *est une matrice* $n \times q$,

$A = \sigma^2 I_n + C(t)$ *est une matrice* $n \times n$,

$S(t) = (\rho(|t-t_1|)g(t_1)B_0, \ldots, \rho(|t-t_1|)g(t_n)B_0)^t$ *est une matrice* $n \times q$

et $C(t)$ *est une matrice* $n \times n$ *dont l'élément* (i,j) *est donnée par*

$$c_{i,j}(t) = \rho(|t_i - t_j|)g(t_i)^t B_0 g(t_j).$$

La démonstration du théorème précédent se trouve dans O'Hagan (1978).

Les valeurs futures de la variable Y peuvent être prédites en utilisant la distribution *a posteriori* de f. Il peut être montré facilement que la distribution de f étant donné Y est multinormale de dimension n,

$$f(t)|Y \sim N_n(g(t)^t b_1(t), \sigma^2 + g(t)^t B_1(t,t)g(t)).$$

où σ^2 est la variance de la variable y. La fonction f peut donc être estimée par la moyenne $\hat{f}(t) = g(t)^t b_1(t)$.

En analysant les données dans un contexte bayésien, l'inférence *a posteriori* est un compromis entre l'inférence suggérée par l'information *a priori* et celle suggérée par les données. Par conséquent, pour estimer $f(t_i)$ par $\hat{f}(t_i) = g(t_i)^t b_1(t_i)$ pour un certain i, nous avons un compromis entre la moyenne *a priori* $g(t_i)^t b_0$ et la valeur de y_i suggérée par les données. Mais les autres y_j $(j \neq i)$ procurent aussi de l'information sur $\beta(t_i)$, étant donné les corrélations données par $S(t_i)$. C'est la structure de ces corrélations qui détermine la part que les autres observations jouent dans la moyenne $\hat{f}(t) = g(t)^t b_1(t)$. Le modèle de régression localisée crée une structure particulière pour représenter notre croyance *a priori* que $\beta(t)$ est approximativement localement constant.

La moyenne $\hat{f}(t) = g(t)^t b_1(t)$, qui estime la fonction f, dépend du paramètre b_1, qui lui même dépend du paramètre *a priori* b_0 (voir l'équation (2.2.4)). Nous aimerions que la spline $\hat{f}(t)$ ne soit déterminée que par les observations y_i. Afin de remédier à ce problème, nous supposons que le paramètre b_0 est sujet à une distribution *a priori*. La distribution *a priori* de b_0 est choisie comme étant une loi multinormale de dimension q avec moyenne b^* et matrice de variance-covariance

kB^*, $b_0 \sim N_q(b^*, kB^*)$. Laissons maintenant tendre k vers l'infini. La distribution *a posteriori* de $\beta(t)$ étant donné Y reste multinormale mais avec une nouvelle moyenne et une nouvelle matrice de variance-covariance qui, lorsque $k \to \infty$, deviennent

$$b_2(t) = \mathbb{E}[\beta(t)|y_1, \ldots, y_n] = S(t)^t A^{-1} y + Q(t)\hat{b}_0, \tag{2.2.5}$$

où $\hat{b}_0 = (G^t A^{-1} G)^{-1} G^t A^{-1} Y$ et

$$\begin{aligned} \Xi(t, t^*) &= \mathbb{E}[(\beta(t) - b_2(t))(\beta(t^*) - b_2(t^*))|y_1, \ldots, y_n] \\ &= B_1(t, t^*) + Q(t)^t (G^t A^{-1} G)^{-1} Q(t^*). \end{aligned} \tag{2.2.6}$$

Il est à noter que cette distribution limite est indépendante du choix de b^* et de kB^*.

Ainsi, la spline estimant la fonction f devient $\hat{f}(t) = g(t)^t b_2(t)$. La preuve de ce résultat se trouve dans O'Hagan (1978).

2.2.3. Analyse discriminante des coefficients pour la charge virale

Étant donné que, dans le modèle classique, nous avons utilisé 2 comme valeur pour le degré de la spline $(k-1)$ pour calculer les coefficients, nous conservons la même valeur pour le k. Ainsi, nous pouvons comparer les résultats obtenus à partir du modèle classique et ceux donnés par le modèle bayésien.

Les probabilités *a priori* utilisées dans l'analyse discriminante ont été fixées aux mêmes valeurs que pour la spline classique c'est-à-dire à $\frac{26}{58}$ pour le groupe 0 et à $\frac{32}{58}$ pour le groupe 1. Une première discrimination, avec $c = 1$, nous a donné un taux de mauvaise classification de 0,138 en utilisant les 3 plus proches voisins. Afin d'améliorer ce taux et donc d'avoir une meilleure précision de notre

estimateur $\hat{f}$, nous avons ajouté la somme des résidus au carré, c'est-à-dire

$$\sum_{i=1}^{n}(y_i - g(t_i)^t b_2(t_i))^2$$

dans l'analyse discriminante. Ceci nous a permis d'obtenir un taux de 0,086. Toujours dans le but d'avoir un meilleur taux de classification, nous avons calculé la variance minimale de $\hat{f}(t)$ en minimisant la trace de cette variance par rapport à la constante c. La distribution de $\hat{f}$, étant donné Y, est multinormale de dimension n,

$$\hat{f}|Y \sim N_n(g(t)^t b_2(t), \sigma^2 + g(t)^t \Xi(t,t) g(t))$$

où $b_2(t)$ et $\Xi(t,t)$ sont définies par les équations (2.2.5) et (2.2.6). Nous avons donc minimisé la trace de $g(t)^t \Xi(t) g(t)$. Une analyse discriminante des coefficents de la spline avec le c minimisant la trace de la variance n'a pas amélioré le taux d'erreur de classification, même en utilisant la somme des résidus au carré. Pour cette raison, nous conservons 1 comme valeur de la constante c.

Par la suite, nous avons fait varier τ^2 afin de trouver une valeur qui minimise le taux de mauvaise classification. Après plusieurs tentatives d'essais et erreurs, nous avons observé que la valeur du τ^2 n'avait pas beaucoup d'influence sur le taux de mauvaise classification. Toutefois, une valeur de 730 pour τ^2 nous a donné un minimum local de 0,069, toujours en utilisant la somme des résidus au carré dans l'analyse discriminante. C'est donc ce taux d'erreur que nous conservons. Le tableau 2.2.2 présente les coefficients $\beta(t)$ moyens obtenus en utilisant $\tau^2 = 730$. En effet, étant donné que nous possédons une série d'observations pour chaque patient, nous utilisons plutôt la moyenne des coefficients $b_2(t)$ pour effectuer l'analyse discriminante des coefficients de la fonction. Le tableau 2.2.1 indique les résultats de la classification de l'analyse discriminante.

Tableau 2.2.1: *Tableau de classification de l'analyse discriminante pour la spline bayésienne*

Réel / Classé	0	1	Total
0	24	2	26
1	2	30	32
Total	30	28	58

Tableau 2.2.2: Tableau des coefficients moyens pour la spline bayésienne

Coefficients moyens de la spline bayésienne avec $\tau^2 = 730$				
Patient	Groupe	$\hat{\beta}_{i0}$	$\hat{\beta}_{i1}$	$\hat{\beta}_{i2}$
1	1	3,645	-23,734	68,700
2	1	373,767	478,364	-1016,813
3	1	97,095	-53,644	-45,745
4	1	2,011	-8,039	20,706
5	1	-154,393	1746,191	-2467,795
6	1	1,165	8,984	-0,218
7	1	73,051	-63,228	-15,129
8	1	36,185	64,638	-175,786
9	1	11,894	15,479	-55,824
10	1	7,011	138,987	-95,526
11	1	104,674	-467,795	1720,099
12	1	123,559	-1306,265	3662,611
13	1	-4,015	172,651	-324,735
14	1	15,407	-73,741	314,496
15	1	48,211	-249,412	841,437
16	1	20,180	-177,779	818,411
17	1	-19,649	163,994	-235,485
18	1	2,599	1665,486	-2523,013
19	1	5,351	-16,108	24,610
20	1	12,957	36,520	-50,114
21	1	2,018	-13,725	44,376

Coefficients moyens de la spline bayésienne avec $\tau^2 = 730$				
Patient	Groupe	$\hat{\beta}_{i0}$	$\hat{\beta}_{i1}$	$\hat{\beta}_{i2}$
22	1	42,143	21,641	108,665
23	1	-0,513	23,064	46,804
24	1	-234,098	1984,110	-2103,124
25	1	34,424	167,265	-385,316
26	1	-195,295	1445,741	-773,805
27	1	384,036	-971,149	1288,904
28	1	3,150	19,339	-30,431
29	1	-27,112	227,133	-166,394
30	1	15,733	338,210	-377,928
31	1	-170,806	1322,021	-1984,318
32	1	33,180	-56,811	90,933
33	0	-125,915	904,596	-1582,148
34	0	0,985	0,103	-0,187
35	0	8,159	-24,660	36,450
36	0	1,000	-1,363e-15	4,510e-16
37	0	1,006	-5,067e-2	0,130
38	0	0,965	0,920	-1,917
39	0	1,996	-3,572	5,318
40	0	1,027	-0,104	0,160
41	0	24,533	-176,874	429,109
42	0	5,145	-15,280	23,618
43	0	5,721	-17,100	25,857
44	0	-18,010	552,308	-1185,687
45	0	0,754	1,842	-2.885
46	0	-45,067	198,586	-253,914
47	0	-1,501	20,382	-31,447
48	0	6,558	-60,848	163,137
49	0	0,737	4,147	-7,165
50	0	3,052	-6,783	9,891
51	0	6,734	-21,336	32,403

Coefficients moyens de la spline bayésienne avec $\tau^2 = 730$				
Patient	Groupe	$\hat{\beta}_{i0}$	$\hat{\beta}_{i1}$	$\hat{\beta}_{i2}$
52	0	-1,332	22,420	-42,969
53	0	1,000	7,531e-15	-2,061e-14
54	0	0,973	0,467	-0,777
55	0	1,000	9,062e-15	-1,759e-14
56	0	3,723	-9,421	13,789
57	0	0,936	0,819	-1,557
58	0	47,807	-153,207	219,672

2.3. Comparaison des résultats de l'analyse discriminante à l'aide des méthodes classique et bayésienne

En comparant les deux taux de mauvaise classification, nous remarquons que les deux sortes de spline sont équivalents. Afin de pouvoir voir en détail les différences entre les deux splines, nous regardons les observations qui ont été mal classifiées. Le tableau 2.3.1 nous indique respectivement le numéro du patient (de 1 à 58), le groupe auquel il appartient, le groupe dans lequel il a été classifié. Nous avons aussi reporté la probabilité *a posteriori* de l'appartenance au vrai groupe en faisant l'analyse discriminante basée sur les splines de lissage et les splines ajustées de O'Hagan. Ces deux dernières informations sont présentes dans les quatre dernières colonnes du tableau, les deux premières pour la spline classique et les deux dernières pour la spline bayésienne. Rappelons que ces informations pour les splines bayésiennes ont été obtenues en utilisant les coefficients moyens dans l'analyse discriminante.

En regardant les patients dans le groupe 1, nous remarquons que les patients 1, 4 et 19 ont été mal classés par la spline classique. Par contre, le spline bayésienne classe bien le patient 1, mais donne la même probabilité *a posteriori* que la spline

classique pour les patients 4 et 19 et ceux-ci sont donc aussi classés dans le groupe 0.

Dans le groupe 0, la spline classique et la spline bayésienne classent le patient 44 dans le groupe 1. De plus, le patient 48 est bien classé par la spline classique mais pas par la spline bayésienne.

Dans ce deuxième chapitre, nous avons tenté de modéliser la charge virale en fonction du temps à l'aide de splines. Nous avons estimé deux sortes de splines. Pour estimer ces splines, nous avons tout d'abord utilisé un modèle classique puis un modèle bayésien. Après avoir effectué une analyse discriminante des coefficients, nous avons conclu que les deux splines permettaient de bien classer les patients dans leur groupe respectif. En effet, les deux méthodes ont obtenu un taux de mauvaise classification de 0,069. Les deux méthodes sont donc équivalentes malgré le fait que la classification des patients n'est pas tout à fait pareille pour ces deux méthodes.

Tableau 2.3.1: Tableau de classification des patients à l'aide de splines classiques et bayésiennes

Patient	Vrai groupe	Spline classique		Spline bayésienne	
		Groupe classé	Probabilité *a posteriori*	Groupe classé	Probabilité *a posteriori*
1	1	0	0,333	1	1,000
2	1	1	1,000	1	1,000
3	1	1	0,667	1	1,000
4	1	0	0,333	0	0,333
5	1	1	1,000	1	1,000
6	1	1	0,667	1	0,667
7	1	1	0,667	1	0,667
8	1	1	1,000	1	1,000

Patient	Vrai groupe	Spline classique		Spline bayésienne	
		Groupe classé	Probabilité *a posteriori*	Groupe classé	Probabilité *a posteriori*
9	1	1	0,667	1	0,667
10	1	1	1,000	1	1,000
11	1	1	1,000	1	1,000
12	1	1	1,000	1	1,000
13	1	1	1,000	1	1,000
14	1	1	1,000	1	0,667
15	1	1	1,000	1	1,000
16	1	1	1,000	1	1,000
17	1	1	0,667	1	0,667
18	1	1	1,000	1	1,000
19	1	0	0,333	0	0,333
20	1	1	0,667	1	1,000
21	1	1	0,667	1	1,000
22	1	1	1,000	1	1,000
23	1	1	0,667	1	0,667
24	1	1	1,000	1	1,000
25	1	1	0,667	1	1,000
26	1	1	1,000	1	1,000
27	1	1	1,000	1	1,000
28	1	1	0,667	1	0,667
29	1	1	1,000	1	1,000
30	1	1	1,000	1	1,000
31	1	1	1,000	1	0,667
32	1	1	1,000	1	0,667
33	0	0	1,000	0	1,000
34	0	0	1,000	0	1,000
35	0	0	1,000	0	1,000
36	0	0	1,000	0	1,000
37	0	0	1,000	0	1,000
38	0	0	1,000	0	1,000

Patient	Vrai groupe	Spline classique		Spline bayésienne	
		Groupe classé	Probabilité *a posteriori*	Groupe classé	Probabilité *a posteriori*
39	0	0	1,000	0	1,000
40	0	0	1,000	0	1,000
41	0	0	0,667	0	0,667
42	0	0	1,000	0	0,667
43	0	0	0,667	0	0,667
44	0	1	0,333	1	0,333
45	0	0	1,000	0	1,000
46	0	0	1,000	0	0,667
47	0	0	0,667	0	1,000
48	0	0	0,667	1	0,333
49	0	0	1,000	0	1,000
50	0	0	0,667	0	1,000
51	0	0	1,000	0	0,667
52	0	0	1,000	0	1,000
53	0	0	1,000	0	1,000
54	0	0	1,000	0	1,000
55	0	0	1,000	0	1,000
56	0	0	0,667	0	1,000
57	0	0	1,000	0	1,000
58	0	0	0,667	0	0,667

Étant donné que chaque observation a été utilisée deux fois, c'est-à-dire une fois pour construire la fonction discriminante et une autre fois pour vérifier la classification, l'erreur de classification s'en trouve sous-estimée. Pour remédier à ce problème, nous aurions pu utiliser la validation croisée pour effectuer la classification. Cette méthode se fait comme suit. Tout d'abord, nous calculons les probabilités d'appartenance au groupe 0 et au groupe 1 d'un patient donné en utilisant tous les patients sauf celui-ci pour établir la fonction discriminante. Puis, nous classons le patient dans le groupe où la probabilité est la plus grande.

Nous effectuons ceci pour tous les patients et nous calculons le taux d'erreur de classification. Cette méthode aurait cependant nécessité des calculs complexes.

Chapitre 3

MODÈLES CONTAMINÉS

Dans ce dernier chapitre, nous essayons de modéliser l'échantillon contenant les deux groupes de patients en utilisant un modèle contaminé introduit à la section 1.3. Afin d'estimer les paramètres de ce modèle, nous utilisons l'algorithme EM présenté à la section 1.4. Nous estimons ces paramètres tout d'abord à l'aide d'un modèle classique puis de manière bayésienne. Une fois les paramètres estimés, nous utilisons l'algorithme décrit à la section 1.4 afin de pouvoir classifier les patients en deux groupes. Encore une fois, nous comparons les résultats obtenus par la classification classique et ceux obtenus par la classification bayésienne.

3.1. MÉTHODE CLASSIQUE

3.1.1. Modèle contaminé

À la section 1.3, nous supposons que les deux groupes proviennent d'une distribution normale univariée pour introduire le modèle contaminé. Or, chaque patient i possède un vecteur d'observations $\underset{\sim}{y_i}$ pour la charge virale. Nous verrons à la section 3.1.2 que si nous tenons compte du groupe les résidus de la régression des $\underset{\sim}{y_i}$ sur les X_i, que nous définissons un peu plus loin, sont issus d'une population de densité relativement symétrique. Pour simplifier les calculs, nous supposerons

donc que la densité des erreurs est normale. Par conséquent, nous pouvons supposer que le vecteur $\underset{\sim}{Y}_i$ provient d'une population normale multivariée de dimension n_i c'est-à-dire

$$\underset{\sim}{Y}_i \sim N_{n_i}(X_i\underset{\sim}{\beta}_i, k\sigma_0^2 I),$$

où n_i est le nombre d'observations pour le patient i, $X_i = (X_{i1}, X_{i2}, \ldots, X_{in_i})^t$ est une matrice de dimension $n_i \times 3$ avec $X_{ij} = \left(1, t_{i1}, \frac{t_{i2}^2}{2!}\right)^t$ et où k vaut 1 si le patient i est classé dans le groupe 0 et $k > 1$ sinon.

La densité de la charge virale $\underset{\sim}{y}_i$ est alors :

$$f(\underset{\sim}{y}_i|X_i\underset{\sim}{\beta}_i, \sigma_0^2, k) = \varepsilon f_0(\underset{\sim}{y}_i|X_i\underset{\sim}{\beta}_i, \sigma_0^2 I) + (1-\varepsilon) f_1(\underset{\sim}{y}_i|X_i\underset{\sim}{\beta}_i, k\sigma_0^2 I),$$

où $f_l(\underset{\sim}{y}_i|\underset{\sim}{a}, A)$, $l = 0, 1$, est la densité d'une normale de moyenne $\underset{\sim}{a}$ et de matrice de covariance A. Ceci peut aussi être écrit sous la forme :

$$\begin{aligned} f(\underset{\sim}{y}_i|X_i\underset{\sim}{\beta}_i, \sigma_0^2, k) &= \varepsilon \frac{1}{(2\pi\sigma_0^2)^{n_i/2}} \exp\left(-\frac{1}{2\sigma_0^2}(\underset{\sim}{y}_i - X_i\underset{\sim}{\beta}_i)^t(\underset{\sim}{y}_i - X_i\underset{\sim}{\beta}_i)\right) \\ &\quad + (1-\varepsilon)\frac{1}{(2\pi k\sigma_0^2)^{n_i/2}} \exp\left(-\frac{1}{2k\sigma_0^2}(\underset{\sim}{y}_i - X_i\underset{\sim}{\beta}_i)^t(\underset{\sim}{y}_i - X_i\underset{\sim}{\beta}_i)\right). \end{aligned}$$

Nous estimons les paramètres $\varepsilon, k, \sigma_0^2$ et $\underset{\sim}{\beta}_i$, $i = 1, \ldots, N$ en utilisant la méthode du maximum de vraisemblance. Nous devons donc maximiser la fonction suivante :

$$L(\underset{\sim}{y}|\underset{\sim}{\eta}) = \prod_{i=1}^{N} \left[\varepsilon f_0(\underset{\sim}{y}_i|X_i\underset{\sim}{\beta}_i, \sigma_0^2 I) + (1-\varepsilon) f_1(\underset{\sim}{y}_i|X_i\underset{\sim}{\beta}_i, k\sigma_0^2 I)\right].$$

Afin de simplifier les calculs, nous maximisons le logarithme de la fonction de vraisemblance. Mais, étant donné qu'il est difficile de maximiser directement cette fonction par rapport aux paramètres, nous utilisons l'algorithme EM introduit à la section 1.4.

3.1.2. Algorithme EM appliqué au modèle classique

Lorsque nous introduisons la variable latente Z dans notre modèle, la fonction de vraisemblance du modèle contaminé devient :

$$\begin{aligned} L(\underset{\sim}{y}|\underset{\sim}{\eta},\underset{\sim}{z}) &= \prod_{i=1}^{N} f_0(\underset{\sim}{y_i}|X_i\underset{\sim}{\beta_i},\sigma_0^2 I)^{1-z_i} f_1(\underset{\sim}{y_i}|X_i\underset{\sim}{\beta_i},k\sigma_0^2 I)^{z_i} \\ &= \prod_{i=1}^{N} \frac{1}{(2\pi\sigma_0^2)^{(1-z_i)n_i/2}} \left[\exp\left(-\frac{1}{2\sigma_0^2}(\underset{\sim}{y_i}-X_i\underset{\sim}{\beta_i})^t(\underset{\sim}{y_i}-X_i\underset{\sim}{\beta_i})\right)\right]^{(1-z_i)} \\ &\quad \times \frac{1}{(2\pi k\sigma_0^2)^{z_i n_i/2}} \left[\exp\left(-\frac{1}{2k\sigma_0^2}(\underset{\sim}{y_i}-X_i\underset{\sim}{\beta_i})^t(\underset{\sim}{y_i}-X_i\underset{\sim}{\beta_i})\right)\right]^{z_i} \\ &= \prod_{i=1}^{N} \frac{1}{(2\pi\sigma_0^2)^{n_i/2}k^{z_i n_i/2}} \\ &\quad \times \exp\left(-\frac{1}{2\sigma_0^2}\left[(1-z_i)+\frac{z_i}{k}\right](\underset{\sim}{y_i}-X_i\underset{\sim}{\beta_i})^t(\underset{\sim}{y_i}-X_i\underset{\sim}{\beta_i})\right), \end{aligned}$$

où $Z_i|\varepsilon \sim Bin(1,1-\varepsilon)$. Le logarithme de la fonction de vraisemblance $l(\underset{\sim}{y}|\underset{\sim}{\eta},\underset{\sim}{z})$ devient alors :

$$\begin{aligned} l(\underset{\sim}{y}|\underset{\sim}{\eta},\underset{\sim}{z}) &= -\frac{1}{2}\sum_{i=1}^{N} n_i \log(2\pi\sigma_0^2) - \frac{\log k}{2}\sum_{i=1}^{N} z_i n_i \\ &\quad -\frac{1}{2\sigma_0^2}\sum_{i=1}^{N}\left[(1-z_i)+\frac{z_i}{k}\right]\|\underset{\sim}{y_i}-X_i\underset{\sim}{\beta_i}\|^2. \end{aligned}$$

où $\|\cdot\|$ est la norme telle qu'énoncée par la définition 1.4.1.

Pour calculer l'étape E de l'algorithme EM, nous avons besoin de la loi de Z_i étant donné $\underset{\sim}{Y_i}$ et $\underset{\sim}{\eta_i}$. Nous avons vu à la section 1.4 que

$$Z_i|\underset{\sim}{Y_i},\underset{\sim}{\eta_i} \sim Bin\left(1, \frac{(1-\varepsilon)f_1(\underset{\sim}{y_i}|X_i\underset{\sim}{\beta_i},k\sigma_0^2 I)}{\varepsilon f_0(\underset{\sim}{y_i}|X_i\underset{\sim}{\beta_i},\sigma_0^2 I)+(1-\varepsilon)f_1(\underset{\sim}{y_i}|X_i\underset{\sim}{\beta_i},k\sigma_0^2 I)}\right).$$

Rappelons les deux étapes de l'algorithme EM.

i. Étape E (étape de l'espérance) :

Évaluer $Q(\underset{\sim}{\eta}, \underset{\sim}{\eta}^{(m)}) = \mathbb{E}^{Z|\underset{\sim}{Y},\underset{\sim}{\eta}^{(m)}}[\log[L(\underset{\sim}{x}|\underset{\sim}{\eta})]]$, où $\underset{\sim}{x} = (\underset{\sim}{y}, \underset{\sim}{z})$.

ii. Étape M (étape de maximisation) :

Trouver $\underset{\sim}{\eta} = \underset{\sim}{\eta}^{(m+1)}$ qui maximise $Q(\underset{\sim}{\eta}, \underset{\sim}{\eta}^{(m)})$.

À l'étape de l'espérance, nous calculons $\mathbb{E}^{Z|\underset{\sim}{Y},\underset{\sim}{\eta}^{(m)}}[\log[L(\underset{\sim}{x}|\underset{\sim}{\eta})]]$.

Théorème 3.1.1

À l'étape E, nous calculons $Q(\underset{\sim}{\eta}, \underset{\sim}{\eta}^{(m)})$ comme suit :

$$\begin{aligned} Q(\underset{\sim}{\eta}, \underset{\sim}{\eta}^{(m)}) &= \log\varepsilon \sum_{i=1}^{N}(1-\varepsilon_i^{(m)}) + \log(1-\varepsilon)\sum_{i=1}^{N}\varepsilon_i^{(m)} - \frac{1}{2}\sum_{i=1}^{N} n_i \log(2\pi\sigma_0^2) \\ &\quad -\frac{\log k}{2}\sum_{i=1}^{N}\varepsilon_i^{(m)} n_i - \frac{1}{2\sigma_0^2}\sum_{i=1}^{N}\left[(1-\varepsilon_i^{(m)}) + \frac{\varepsilon_i^{(m)}}{k}\right]\|\underset{\sim}{y}_i - X_i\underset{\sim}{\beta}_i\|^2, \end{aligned}$$

où

$$\varepsilon_i^{(m)} = \frac{(1-\varepsilon^{(m)}) f_1(\underset{\sim}{y}_i | X_i\underset{\sim}{\beta}_i^{(m)}, k^{(m)}\sigma_0^{2(m)} I)}{\varepsilon^{(m)} f_0(\underset{\sim}{y}_i | X_i\underset{\sim}{\beta}_i^{(m)}, \sigma_0^{2(m)} I) + (1-\varepsilon^{(m)}) f_1(\underset{\sim}{y}_i | X_i\underset{\sim}{\beta}_i^{(m)}, k^{(m)}\sigma_0^{2(m)} I)}. \quad (3.1.1)$$

Pour arriver à ce résultat, nous avons utilisé le théorème 1.4.2 en remplacant μ par $X_i\underset{\sim}{\beta}_i$, σ_0^2 par $\sigma_0^2 I$ et $k\sigma_0^2$ par $k\sigma_0^2 I$ étant donné que le problème est multivarié.

Enfin, à l'étape de maximisation, nous trouvons $\underset{\sim}{\eta}$ qui maximise $Q(\underset{\sim}{\eta}, \underset{\sim}{\eta}^{(m)})$. À la section 1.4, les paramètres du modèle contaminé étaient des scalaires. Dans ce

présent chapitre, pour trouver l'estimateur du maximum de vraisemblance $\underset{\sim}{\beta_i}$ qui maximise $Q(\underset{\sim}{\eta}, \underset{\sim}{\eta}^{(m)})$, il faut utiliser le lemme suivant pour passer au cas vectoriel.

Lemme 3.1.1

La dérivée de la norme du vecteur $\underset{\sim}{y_i} - X_i\underset{\sim}{\beta_i}$ *est:*

$$\frac{\partial}{\partial \underset{\sim}{\beta_i}} \|\underset{\sim}{y_i} - X_i\underset{\sim}{\beta_i}\|^2 = -2X_i^t\underset{\sim}{y_i} + 2X_i^tX_i\underset{\sim}{\beta_i}.$$

Démonstration

Nous calculons la dérivée de la manière suivante:

$$\begin{aligned}
\frac{\partial}{\partial \underset{\sim}{\beta_i}} \|\underset{\sim}{y_i} - X_i\underset{\sim}{\beta_i}\|^2 &= \frac{\partial}{\partial \underset{\sim}{\beta_i}} (\underset{\sim}{y_i} - X_i\underset{\sim}{\beta_i})^t(\underset{\sim}{y_i} - X_i\underset{\sim}{\beta_i}) \\
&= \frac{\partial}{\partial \underset{\sim}{\beta_i}} (\underset{\sim}{y_i}^t\underset{\sim}{y_i} - \underset{\sim}{y_i}^tX_i\underset{\sim}{\beta_i} - \underset{\sim}{\beta_i}^tX_i^t\underset{\sim}{y_i} + \underset{\sim}{\beta_i}^tX_i^tX_i\underset{\sim}{\beta_i}) \\
&= \frac{\partial}{\partial \underset{\sim}{\beta_i}} (\underset{\sim}{y_i}^t\underset{\sim}{y_i} - 2\underset{\sim}{\beta_i}^tX_i^t\underset{\sim}{y_i} + \underset{\sim}{\beta_i}^tX_i^tX_i\underset{\sim}{\beta_i}) \\
&= -2X_i^t\underset{\sim}{y_i} + 2X_i^tX_i\underset{\sim}{\beta_i}.
\end{aligned}$$

□

Théorème 3.1.2

Les estimateurs maximisant $Q(\underset{\sim}{\eta}, \underset{\sim}{\eta}^{(m)})$ *sont les suivants:*

$$\varepsilon^{(m+1)} = \frac{\sum_{i=1}^{N}(1 - \varepsilon_i^{(m)})}{N}, \tag{3.1.2}$$

$$\underset{\sim}{\beta_i}^{(m+1)} = (X_i^tX_i)^{-1}X_i^t\underset{\sim}{y_i}, \tag{3.1.3}$$

$$\sigma_0^{2(m+1)} = \sum_{i=1}^{N}\left[(1 - \varepsilon_i^{(m)}) + \frac{\varepsilon_i^{(m)}}{k^{(m)}}\right] \|\underset{\sim}{y_i} - X_i\underset{\sim}{\beta_i}^{(m)}\|^2 \Big/ \sum_{i=1}^{N} n_i\,, \tag{3.1.4}$$

$$k^{(m+1)} = \frac{\sum_{i=1}^{N} \varepsilon_i^{(m)} \|\underset{\sim}{y}_i - X_i \underset{\sim}{\beta}_i^{(m)}\|^2}{\sigma_0^{2(m)} \sum_{i=1}^{N} \varepsilon_i^{(m)} n_i}, \tag{3.1.5}$$

où $\varepsilon_i^{(m)}$ *est telle que définie par l'équation (3.1.1).*

Notons ici que les $\underset{\sim}{\beta}_i^{(m)}$ ne dépendent pas de m et ils ne varient donc pas durant les itérations. Pour arriver à ces résultats, nous avons dérivé $Q(\underset{\sim}{\eta}, \underset{\sim}{\eta}^{(m)})$ par rapport à chacun des paramètres et nous avons solutionné le système d'équations tel que décrit à la section 1.4 en utilisant le lemme 3.1.1 pour le paramètre $\underset{\sim}{\beta}_i$.

Nous appliquons cet algorithme à notre jeu de données jusqu'à convergence, c'est-à-dire jusqu'à ce que $\|\underset{\sim}{\eta}^{(m+1)} - \underset{\sim}{\eta}^{(m)}\|$ soit suffisamment petit, afin d'estimer les paramètres $\varepsilon, k, \sigma_0^2$ et $\underset{\sim}{\beta}_i$, $i = 1, \ldots, N$. Les valeurs de départ sont : $\varepsilon^{(0)} =$ 0,448 (26 patients sur 58 sont dans le groupe 0), les $\underset{\sim}{\beta}_i = (\beta_{i0}, \beta_{i1}, \beta_{i2})$ sont les coefficients de régression qui se trouvent dans le tableau 3.1.1 et ne varient pas, $\sigma_0^{2(0)} = 198{,}862$ est la variance du groupe 0 et $k^{(0)} = 21{,}514$ est le rapport de la variance du groupe 1 sur la variance du groupe 0. À l'aide des équations 3.1.2 à 3.1.5, nous trouvons les estimateurs pour chacun des paramètres. Le tableau 3.1.1 présente les résultats. Ainsi, la densité de la charge virale s'écrit comme :

$$f(\underset{\sim}{y}_i | X_i \underset{\sim}{\beta}_i, \sigma_0^2, k) = 0{,}500 f_0(\underset{\sim}{y}_i | X_i \underset{\sim}{\beta}_i; 7{,}740I) + 0{,}500 f_1(\underset{\sim}{y}_i | X_i \underset{\sim}{\beta}_i; 24976{,}895I),$$

où les $\underset{\sim}{\beta}_i$ sont les coefficients de régression de la charge virale sur la variable temps.

En regardant le tableau 3.1.1, nous remarquons que certains patients ont des valeurs extrêmes pour les coefficients $\underset{\sim}{\beta}_i$. Ces patients sont indiqués avec un + en exposant dans la première colonne du tableau. Pour expliquer ces valeurs, nous avons fait des nuages de points à l'aide de toutes les variables (voir la section 1.1 pour une description des variables) et nous avons trouvé que, pour certaines variables, les patients qui ont de grandes valeurs de $\underset{\sim}{\beta}_i$ se retrouvent un peu à l'écart des autres patients. En effet, pour le groupe 0, ce sont les variables CD4,

CD3 et CD28 qui permettent d'isoler les patients ayant des valeurs à l'écart pour $\underset{\sim}{\beta_i}$ tandis que pour le groupe 1, ce sont les variables CD4, CD8 et CD38. Les figures 3.1.1 et 3.1.2 indiquent les patients ayant des valeurs extrêmes avec des "+" et les autres patients avec des "·". Ainsi, nous observons en général que des variables affectées par la charge virale des patients ayant des valeurs à l'écart pour les $\underset{\sim}{\beta_i}$ ne semblent pas avoir un comportement semblable aux autres patients. Ceci expliquerait pourquoi ces patients ont des coefficients de régression de la charge virale différents de la majorité des patients, étant donné que leur évolution vers la maladie est également différente.

Tableau 3.1.1: Tableau des estimateurs du modèle classique

Estimateurs du modèle contaminé classique				
$\hat{\varepsilon} = 0{,}500$		$\hat{\sigma}_0^2 = 7{,}740$		$\hat{k} = 3226{,}989$
Patient	Groupe	$\hat{\beta}_{i0}$	$\hat{\beta}_{i1}$	$\hat{\beta}_{i2}$
1	1	3,866	-18,405	51,607
2^+	1	374,257	475,634	-1011,531
3	1	98,887	-60,405	-36,489
4	1	2,305	-8,147	19,193
5^+	1	-154,592	1746,010	-2466,213
6	1	0,606	11,268	-3,028
7	1	74,970	-70,818	-3,844
8	1	41,122	47,107	-151,301
9	1	18,579	-1,491	-6,605
10	1	8,776	128,271	-73,590
11^+	1	105,772	-467,963	1713,849
12^+	1	124,363	-1310,608	3670,621
13	1	-8,317	190,532	-352,851
14	1	12,934	-62,964	297,053
15	1	51,177	-256,241	842,838

Estimateurs du modèle contaminé classique				
Patient	Groupe	$\hat{\beta}_{i0}$	$\hat{\beta}_{i1}$	$\hat{\beta}_{i2}$
16	1	23,482	-193,834	846,690
17	1	-27,279	191,757	-273,882
18^{+}	1	2,938	1664,781	-2522,874
19	1	3,421	-8,381	12,726
20	1	16,905	17,883	-17,170
21	1	2,769	-12,901	37,208
22	1	40,155	28,553	100,619
23	1	0,446	6,494	90,058
24^{+}	1	-234,022	1983,489	-2101,660
25	1	37,271	155,260	-366,713
26^{+}	1	-195,285	1445,461	-772,993
27^{+}	1	380,934	-959,159	1271,474
28	1	-0,660	40,367	-71,752
29	1	-27,641	226,759	-161,387
30	1	15,385	339,880	-380,681
31^{+}	1	-173,303	1331,003	-1996,508
32	1	31,621	-51,826	85,680
33^{+}	0	-126,832	908,885	-1589,113
34	0	1,011	8,250e-03	-0,045
35	0	6,198	-17,028	24,865
36	0	1,000	1,865e-14	-1,243e-14
37	0	1,011	-0,062	0,135
38	0	1,467	-1,293	1,698
39	0	1,642	-2,127	3,120
40	0	1,014	-0,052	0,080
41^{+}	0	29,325	-184,478	422,854
42	0	3,277	-7,604	11,319
43	0	3,749	-9,249	13,812
44^{+}	0	-15,755	545,779	-1179,891
45	0	0,281	3,216	-4,373

Estimateurs du modèle contaminé classique				
Patient	Groupe	$\hat{\beta}_{i0}$	$\hat{\beta}_{i1}$	$\hat{\beta}_{i2}$
46^+	0	-47,885	205,091	-258,592
47	0	-5,906	32,237	-43,115
48^+	0	9,665	-58,949	137,422
49	0	-0,302	8,862	-15,198
50	0	2,674	-5,094	7,136
51	0	4,319	-11,507	17,474
52	0	0,872	14,210	-31,834
53	0	1,000	-1,377e-14	3,286e-14
54	0	0,832	1,030	-1,675
55	0	1,000	-1,643e-14	2,665e-15
56	0	2,892	-6,155	8,873
57	0	1,175	-0,169	-0,060
58^+	0	37,830	-116,418	166,688

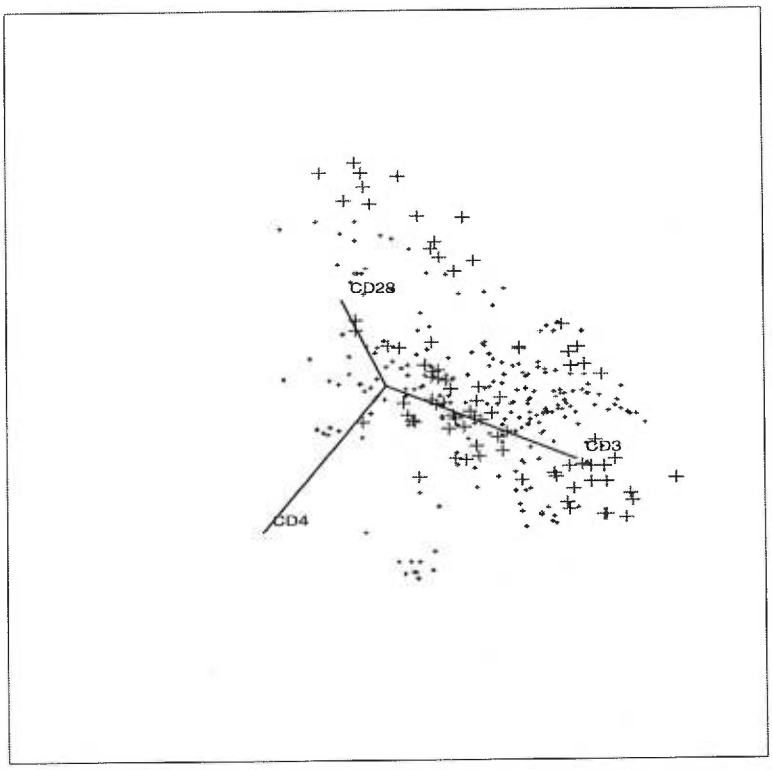
Nous avons également calculé les résidus de la régression des $\underset{\sim}{y}_i$ sur les X_i afin de vérifier s'ils proviennent d'une distribution normale. Les histogrammes des résidus (voir les figures 3.1.3 et 3.1.4) pour le groupe 0 et le groupe 1 montrent que la distribution des résidus pour chacun des deux groupes est relativement symétrique. Par conséquent, le choix de la densité normale semble donc raisonnable.

3.1.3. Classification des patients

En calculant les valeurs des ε_i, nous pouvons déterminer dans quel groupe chaque patient est classé après chaque itération. Par conséquent, le ε_i obtenu à la dernière itération nous indique dans quel groupe est classé chacun des patients. La classification se fait comme suit :

- si $\varepsilon_i < 0{,}5$, alors le patient est classé dans le groupe 0;
- si $\varepsilon_i \geq 0{,}5$, alors le patient est classé dans le groupe 1.

Figure 3.1.1: *Graphique pour les variables CD4, CD3 et CD28 du groupe 0 (+ patients avec grandes valeurs pour $\underset{\sim}{\beta}_i$, · autres patients)*



Cet algorithme de classification vient du fait que le ε_i est défini comme suit :

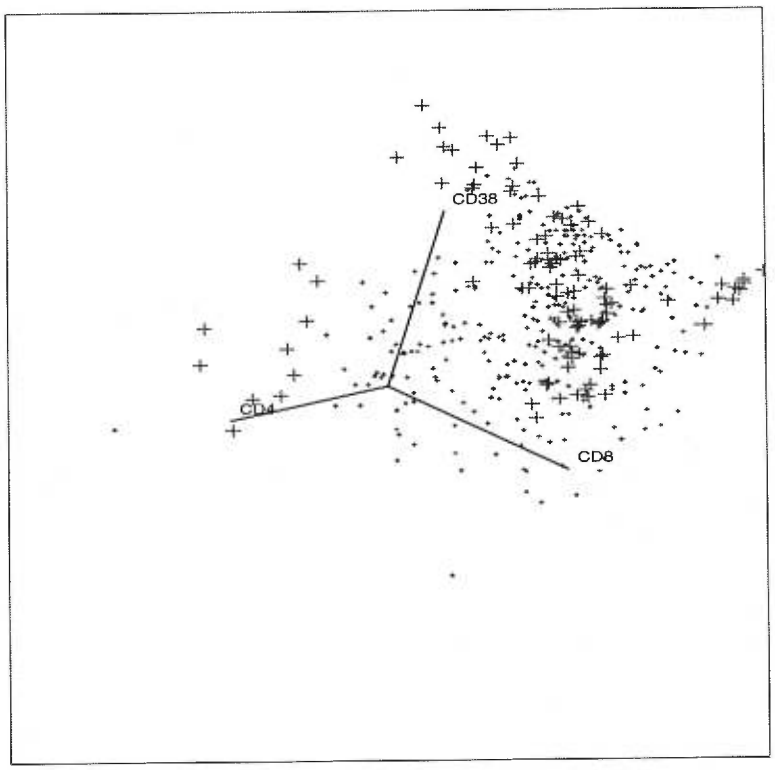
$$\varepsilon_i = \frac{(1-\varepsilon) f_1(\underset{\sim}{y}_i | X_i \underset{\sim}{\beta}_i, k\sigma_0^2 I)}{\varepsilon f_0(\underset{\sim}{y}_i | X_i \underset{\sim}{\beta}_i, \sigma_0^2 I) + (1-\varepsilon) f_1(\underset{\sim}{y}_i | X_i \underset{\sim}{\beta}_i, k\sigma_0^2 I)}$$

et de l'algorithme de la section 1.3 qui est tel que

- si $f_0(y_i|\underset{\sim}{\eta}_0) > f_1(y_i|\underset{\sim}{\eta}_1)$, alors le patient est classé dans le groupe 0;
- si $f_0(y_i|\underset{\sim}{\eta}_0) \leq f_1(y_i|\underset{\sim}{\eta}_1)$, alors le patient est classé dans le groupe 1.

Ainsi, l'algorithme réussit à classer 47 patients dans leur groupe respectif et le taux de mauvaise classification est donc de 0,190. À la section 3.3, nous voyons plus en détail la classification des patients. En effet, le tableau 3.3.1 permettant

Figure 3.1.2: *Graphique pour les variables CD4, CD8 et CD38 du groupe 1 (+ patients avec grandes valeurs pour $\underset{\sim}{\beta}_i$, · autres patients)*



de comparer les méthodes classique et bayésienne indique pour chaque patient la valeur du ε_i à la dernière itération et le groupe dans lequel il est classé.

Nous remarquons que le taux de bonne classification est inférieur à celui obtenu par les splines au chapitre 2. Ceci peut s'expliquer par le fait que nous sommes en présence d'observations provenant de deux lois normales avec la même moyenne mais avec des variances différentes. En effet, il est difficile dans ce contexte d'obtenir un bon taux de classification car les deux lois normales ont un grand recouvrement, dû au fait qu'elles ont la même moyenne et il est donc difficile de bien classer une observation. En calculant un intervalle de confiance pour

Figure 3.1.3: *Histogramme des résidus pour le groupe 0*

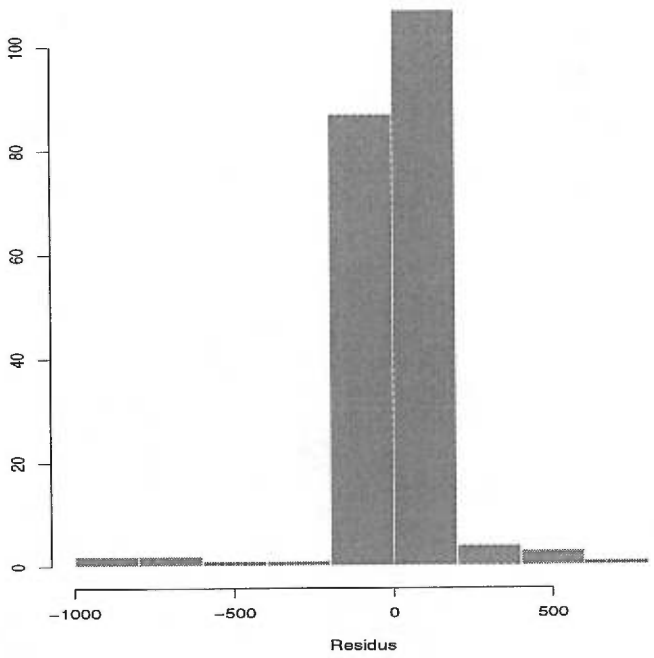
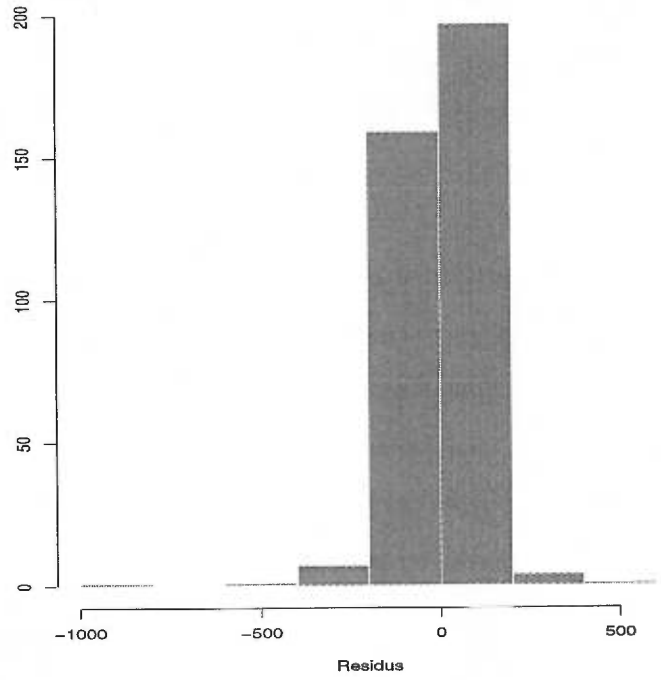


Figure 3.1.4: *Histogramme des résidus pour le groupe 1*



chacun des deux groupes, nous obtenons le résultat suivant: 95% des valeurs de la charge virale du groupe 0 se situent dans l'intervalle [-5370,987;57154,267] et ce même intervalle contient 28% des valeurs du groupe 1. Par conséquent, pour des observations entre [-5370,987;57154,267], il devient difficile de bien déterminer le groupe d'appartenance.

3.2. MÉTHODE BAYÉSIENNE

3.2.1. Modèle contaminé

Rappelons la distribution du vecteur $\underset{\sim}{Y}_i$. La distribution de $\underset{\sim}{Y}_i$ est multinormale de dimension n_i,

$$\underset{\sim}{Y}_i \sim N_{n_i}(X_i \underset{\sim}{\beta}_i, k\sigma_0^2 I),$$

où n_i est le nombre d'observations pour le patient i, $X_i = (X_{i1}, X_{i2}, \ldots, X_{in_i})^t$ est une matrice de dimension $n_i \times 3$ avec $X_{ij} = \left(1, t_{i1}, \frac{t_{i2}^2}{2!}\right)^t$ et où k vaut 1 si le patient i est dans le groupe 0 et $k > 1$ sinon.

Dans un contexte bayésien, nous ne considérons pas les paramètres $\underset{\sim}{\beta}_i$ comme fixes, mais comme étant des variables aléatoires dont la densité est spécifiée plus loin. En effet, nous trouvons la distribution *a posteriori* de ces paramètres qui est un compromis entre l'information *a priori* modélisée par la densité *a priori* et l'information suggérée par les observations contenue dans la fonction de vraisemblance. Pour simplifier, étant donné que la distribution de la variable $\underset{\sim}{y}_i$ est supposée multinormale, nous supposerons également que les coefficents $\underset{\sim}{\beta}_i$ sont issus d'une distribution normale. Pour la densité *a priori* des $\underset{\sim}{\beta}_i$, nous utilisons un modèle proposé par Zellner (1971). Ainsi, la distribution *a priori* de $\underset{\sim}{\beta}_i$, $i = 1, \ldots, n$, est multinormale de dimension 3,

$$\underset{\sim}{\beta}_i | Z_i = 0 \sim N_3(\underset{\sim}{\beta}_0, \frac{\sigma_0^2}{n_0}(X_i^t X_i)^{-1}) \text{ si le patient } i \text{ est dans le groupe 0,}$$

$\underset{\sim}{\beta}_i | Z_i = 1 \sim N_3(\underset{\sim}{\beta}_1, \frac{k\sigma_0^2}{n_0}(X_i^t X_i)^{-1})$ si le patient i est dans le groupe 1,

où n_0 représente notre confiance dans l'information *a priori* des paramètres $\underset{\sim}{\beta}_i$, c'est-à-dire que, plus nous sommes certains de notre information *a priori*, plus n_0 sera choisi grand. Notre choix pour n_0 est discuté à la section 3.2.2.

La densité de la charge virale $\underset{\sim}{y}_i$ devient :

$$\begin{aligned} f(\underset{\sim}{y}_i, \underset{\sim}{\beta}_i | \underset{\sim}{\beta}_0, \underset{\sim}{\beta}_1, \sigma_0^2, k) &= \left[\varepsilon f_0(\underset{\sim}{y}_i | X_i \underset{\sim}{\beta}_i, \sigma_0^2 I) + (1-\varepsilon) f_1(\underset{\sim}{y}_i | X_i \underset{\sim}{\beta}_i, k\sigma_0^2 I)\right] \\ &\quad \times \left[\varepsilon g_0(\underset{\sim}{\beta}_i | \underset{\sim}{\beta}_0, \sigma_0^2 I) + (1-\varepsilon) g_1(\underset{\sim}{\beta}_i | \underset{\sim}{\beta}_1, k\sigma_0^2 I)\right], \end{aligned}$$

où $g_l(\underset{\sim}{y}_i | \underset{\sim}{a}, A)$, $l = 0, 1$, est la densité d'une normale de moyenne $\underset{\sim}{a}$ et de matrice de covariance A. Cette dernière équation peut également être écrite sous la forme :

$$\begin{aligned} & f(\underset{\sim}{y}_i, \underset{\sim}{\beta}_i | \underset{\sim}{\beta}_0, \underset{\sim}{\beta}_1, \sigma_0^2, k) \\ &= \left[\varepsilon \frac{1}{(2\pi\sigma_0^2)^{n_i/2}} \exp\left(-\frac{1}{2\sigma_0^2}\|\underset{\sim}{y}_i - X_i \underset{\sim}{\beta}_i\|^2\right)\right. \\ &\quad \left. + (1-\varepsilon)\frac{1}{(2\pi k\sigma_0^2)^{n_i/2}} \exp\left(-\frac{1}{2k\sigma_0^2}\|\underset{\sim}{y}_i - X_i \underset{\sim}{\beta}_i\|^2\right)\right] \\ &\quad \times \left[\varepsilon \frac{n_0^{3/2}}{(2\pi\sigma_0^2 |(X_i^t X_i)^{-1}|)^{3/2}} \exp\left(-\frac{n_0}{2\sigma_0^2}\|\underset{\sim}{\beta}_i - \underset{\sim}{\beta}_0\|^2_{X_i^t X_i}\right)\right. \\ &\quad \left. + (1-\varepsilon)\frac{n_0^{3/2}}{(2\pi k\sigma_0^2 |(X_i^t X_i)^{-1}|)^{3/2}} \exp\left(-\frac{n_0}{2k\sigma_0^2}\|\underset{\sim}{\beta}_i - \underset{\sim}{\beta}_1\|^2_{X_i^t X_i}\right)\right], \end{aligned}$$

où $\|a\|_A^2 = a^t A a$. Les paramètres $\varepsilon, k, \sigma_0^2, \underset{\sim}{\beta}_0, \underset{\sim}{\beta}_1$ et $\underset{\sim}{\beta}_i$, $i = 1, \ldots, N$ sont estimés de manière à maximiser la fonction de vraisemblance. Nous maximisons donc la fonction suivante :

$$\begin{aligned} L(\underset{\sim}{y} | \underset{\sim}{\eta}) &= \prod_{i=1}^{N} \left[\left(\varepsilon f_0(\underset{\sim}{y}_i | X_i \underset{\sim}{\beta}_i, \sigma_0^2 I) + (1-\varepsilon) f_1(\underset{\sim}{y}_i | X_i \underset{\sim}{\beta}_i, k\sigma_0^2 I)\right)\right. \\ &\quad \left. \times \left(\varepsilon g_0(\underset{\sim}{\beta}_i | \underset{\sim}{\beta}_0, \sigma_0^2) + (1-\varepsilon) g_1(\underset{\sim}{\beta}_i | \underset{\sim}{\beta}_1, k\sigma_0^2)\right)\right]. \end{aligned}$$

Pour simplifier les calculs, le logarithme de la fonction de vraisemblance est maximisé. Tout comme dans le contexte classique, l'algorithme EM, introduit à la section 1.4, est utilisé puisqu'il est difficile de maximiser directement cette fonction par rapport aux paramètres.

3.2.2. Algorithme EM appliqué au modèle bayésien

Après avoir introduit la variable latente Z dans le modèle, la fonction de vraisemblance du modèle contaminé devient :

$$
\begin{aligned}
& L(\underset{\sim}{y}, \underset{\sim}{z})|\underset{\sim}{\eta} \\
&= \prod_{i=1}^{N} \Bigg\{ (1-\varepsilon)^{z_i} \varepsilon^{1-z_i} f_0(\underset{\sim}{y_i}|X_i\underset{\sim}{\beta_i}, \sigma_0^2 I)^{1-z_i} f_1(\underset{\sim}{y_i}|X_i\underset{\sim}{\beta_i}, k\sigma_0^2 I)^{z_i} g_0(\underset{\sim}{\beta_i}|\underset{\sim}{\beta_0}, \sigma_0^2)^{1-z_i} \\
&\quad \times g_1(\underset{\sim}{\beta_i}|\underset{\sim}{\beta_1}, k\sigma_0^2)^{z_i} \Bigg\} \\
&= \prod_{i=1}^{N} \Bigg\{ (1-\varepsilon)^{z_i} \varepsilon^{1-z_i} \frac{1}{(2\pi\sigma_0^2)^{(1-z_i)n_i/2}} \left[\exp\left(-\frac{1}{2\sigma_0^2} \|\underset{\sim}{y_i} - X_i\underset{\sim}{\beta_i}\|^2 \right) \right]^{(1-z_i)} \\
&\quad \times \frac{1}{(2\pi k\sigma_0^2)^{z_i n_i/2}} \left[\exp\left(-\frac{1}{2k\sigma_0^2} \|\underset{\sim}{y_i} - X_i\underset{\sim}{\beta_i})\|^2 \right) \right]^{z_i} \\
&\quad \times \frac{n_0^{3(1-z_i)/2}}{(2\pi\sigma_0^2 |(X_i^t X_i)^{-1}|)^{3(1-z_i)/2}} \left[\exp\left(-\frac{n_0}{2\sigma_0^2} \|\underset{\sim}{\beta_i} - \underset{\sim}{\beta_0}\|^2_{X_i^t X_i} \right) \right]^{(1-z_i)} \\
&\quad \times \frac{n_0^{3z_i/2}}{(2\pi k\sigma_0^2 |(X_i^t X_i)^{-1}|)^{3z_i/2}} \left[\exp\left(-\frac{n_0}{2k\sigma_0^2} \|\underset{\sim}{\beta_i} - \underset{\sim}{\beta_1}\|^2_{X_i^t X_i} \right) \right]^{z_i} \Bigg\} \\
&= \prod_{i=1}^{N} \Bigg\{ \frac{n_0^{3/2} (1-\varepsilon)^{z_i} \varepsilon^{1-z_i}}{(2\pi\sigma_0^2)^{(n_i+3)/2} k^{(z_i n_i + 3z_i)/2} |(X_i^t X_i)^{-1}|^{3/2}} \\
&\quad \times \exp\left(-\frac{(1-z_i)}{2\sigma_0^2} \left[\|\underset{\sim}{y_i} - X_i\underset{\sim}{\beta_i}\|^2 + n_0 \|\underset{\sim}{\beta_i} - \underset{\sim}{\beta_0}\|^2_{X_i^t X_i} \right] \right) \\
&\quad \times \exp\left(-\frac{z_i}{2k\sigma_0^2} \left[\|\underset{\sim}{y_i} - X_i\underset{\sim}{\beta_i}\|^2 + n_0 \|\underset{\sim}{\beta_i} - \underset{\sim}{\beta_1}\|^2_{X_i^t X_i} \right] \right) \Bigg\}
\end{aligned}
$$

où $Z_i|\varepsilon \sim Bin(1, 1-\varepsilon)$.

Le logarithme de la fonction de vraisemblance $l(\underset{\sim}{y}, \underset{\sim}{z}|\underset{\sim}{\eta})$ devient

$$\begin{aligned} l(\underset{\sim}{y}, \underset{\sim}{z}|\underset{\sim}{\eta}) &= \log\varepsilon \sum_{i=1}^{N}(1-z_i) + \log(1-\varepsilon)\sum_{i=1}^{N} z_i - \frac{3}{2}\sum_{i=1}^{N}\log|(X_i^t X_i)^{-1}| \\ &\quad + \frac{3}{2}\log n_0 - \frac{1}{2}\sum_{i=1}^{N}(n_i+3)\log(2\pi\sigma_0^2) - \frac{1}{2}\sum_{i=1}^{N}(n_i+3) z_i \log k \\ &\quad - \frac{1}{2\sigma_0^2}\sum_{i=1}^{N}(1-z_i)\left[\|\underset{\sim}{y}_i - X_i\underset{\sim}{\beta}_i\|^2 + n_0\|\underset{\sim}{\beta}_i - \underset{\sim}{\beta}_0\|^2_{X_i^t X_i}\right] \\ &\quad - \frac{1}{2k\sigma_0^2}\sum_{i=1}^{N} z_i\left[\|\underset{\sim}{y}_i - X_i\underset{\sim}{\beta}_i\|^2 + n_0\|\underset{\sim}{\beta}_i - \underset{\sim}{\beta}_1\|^2_{X_i^t X_i}\right]. \end{aligned}$$

Nous appliquons l'algorithme EM et à l'étape de l'espérance, nous obtenons le résultat suivant.

Théorème 3.2.1

À l'étape E, nous trouvons $Q(\underset{\sim}{\eta}, \underset{\sim}{\eta}^{(m)})$ de la manière suivante :

$$\begin{aligned} Q(\underset{\sim}{\eta}, \underset{\sim}{\eta}^{(m)}) &= \log\varepsilon \sum_{i=1}^{N}(1-\varepsilon_i^{(m)}) + \log(1-\varepsilon)\sum_{i=1}^{N}\varepsilon_i^{(m)} - \frac{3}{2}\sum_{i=1}^{N}\log|(X_i^t X_i)^{-1}| \\ &\quad - \frac{1}{2}\sum_{i=1}^{N}(n_i+3)\log(2\pi\sigma_0^2) - \frac{1}{2}\sum_{i=1}^{N}(n_i+3)\varepsilon_i^{(m)}\log k + \frac{3}{2}\log n_0 \\ &\quad - \frac{1}{2\sigma_0^2}\sum_{i=1}^{N}(1-\varepsilon_i^{(m)})\left[\|\underset{\sim}{y}_i - X_i\underset{\sim}{\beta}_i\|_I^2 + n_0\|\underset{\sim}{\beta}_i - \underset{\sim}{\beta}_0\|^2_{X_i^t X_i}\right] \\ &\quad - \frac{1}{2k\sigma_0^2}\sum_{i=1}^{N}\varepsilon_i^{(m)}\left[\|\underset{\sim}{y}_i - X_i\underset{\sim}{\beta}_i\|_I^2 + n_0\|\underset{\sim}{\beta}_i - \underset{\sim}{\beta}_1\|^2_{X_i^t X_i}\right]. \end{aligned}$$

où ε_i est donné par l'équation (3.1.1).

Pour parvenir à ce résultat, nous avons utilisé le théorème 1.4.2, comme à la section précédente en remplaçant μ par $X_i\underset{\sim}{\beta}_i$, σ_0^2 par $\sigma_0^2 I$ et $k\sigma_0^2$ par $k\sigma_0^2 I$.

Puis, nous trouvons les estimateurs qui maximisent $Q(\underset{\sim}{\eta}, \underset{\sim}{\eta}^{(m)})$ à l'étape de maximisation.

Théorème 3.2.2

Les estimateurs maximisant $Q(\underset{\sim}{\eta}, \underset{\sim}{\eta}^{(m)})$ sont les suivants :

$$\varepsilon^{(m+1)} = \frac{\sum_{i=1}^{N}(1-\varepsilon_i^{(m)})}{N}, \tag{3.2.1}$$

$$\underset{\sim}{\beta}_i^{(m+1)} = \frac{1}{1+n_o}(X_i^t X_i)^{-1} X_i^t \underset{\sim}{y}_i + \frac{n_0}{1+n_0}\frac{(1-\varepsilon_i^{(m)})\underset{\sim}{\beta}_0^{(m)} + \frac{\varepsilon_i^{(m)}}{k^{(m)}}\underset{\sim}{\beta}_1^{(m)}}{1-\varepsilon_i^{(m)} + \frac{\varepsilon_i^{(m)}}{k^{(m)}}}, \tag{3.2.2}$$

$$\underset{\sim}{\beta}_0^{(m+1)} = \left(\sum_{i=1}^{N}(1-\varepsilon_i^{(m)})X_i^t X_i\right)^{-1} \sum_{i=1}^{N}(1-\varepsilon_i^{(m)})X_i^t X_i \underset{\sim}{\beta}_i^{(m)}, \tag{3.2.3}$$

$$\underset{\sim}{\beta}_1^{(m+1)} = \left(\sum_{i=1}^{N}\varepsilon_i^{(m)} X_i^t X_i\right)^{-1} \sum_{i=1}^{N}\varepsilon_i^{(m)} X_i^t X_i \underset{\sim}{\beta}_i^{(m)}, \tag{3.2.4}$$

$$\begin{aligned}\sigma_0^{2(m+1)} = &\left[\sum_{i=1}^{N}(1-\varepsilon_i^{(m)})\left[\|\underset{\sim}{y}_i - X_i\underset{\sim}{\beta}_i^{(m)}\|^2 + n_0\|\underset{\sim}{\beta}_i^{(m)} - \underset{\sim}{\beta}_0^{(m)}\|^2_{X_i^t X_i}\right]\right.\\ &\left.+\frac{1}{k}\sum_{i=1}^{N}\varepsilon_i^{(m)}\left[\|\underset{\sim}{y}_i - X_i\underset{\sim}{\beta}_i^{(m)}\|^2 + n_0\|\underset{\sim}{\beta}_i^{(m)} - \underset{\sim}{\beta}_1^{(m)}\|^2_{X_i^t X_i}\right]\right]\\ &\Bigg/\left[\sum_{i=1}^{N}(n_i+3)\right],\end{aligned} \tag{3.2.5}$$

$$k^{(m+1)} = \frac{\sum_{i=1}^{N}\varepsilon_i^{(m)}\left[\|\underset{\sim}{y}_i - X_i\underset{\sim}{\beta}_i^{(m)}\|^2 + n_0\|\underset{\sim}{\beta}_i^{(m)} - \underset{\sim}{\beta}_1^{(m)}\|^2_{X_i^t X_i}\right]}{\sigma_0^{2(m)}\sum_{i=1}^{N}\varepsilon_i^{(m)}(n_i+3)}, \tag{3.2.6}$$

où $\varepsilon_i^{(m)}$ est telle que définie par l'équation (3.1.1).

L'algorithme est appliqué au jeu de données jusqu'à convergence afin d'estimer les paramètres $\varepsilon, k, \sigma_0^2, \underset{\sim}{\beta}_0, \underset{\sim}{\beta}_1$, et $\underset{\sim}{\beta}_i$, $i = 1, \ldots, N$. Les valeurs de départ sont

les mêmes que pour le modèle contaminé classique c'est-à-dire : $\varepsilon^{(0)} = 0{,}448$ (26 patients sur 58 sont dans le groupe 0), $\underset{\sim}{\beta}_i^{(0)} = (\beta_{i0}^{(0)}, \beta_{i1}^{(0)}, \beta_{i2}^{(0)})$ sont les coefficients de régression linéaire, $\underset{\sim}{\beta}_0$ est la moyenne des moindres carrés du groupe 0, $\underset{\sim}{\beta}_1$ est la moyenne des moindres carrés du groupe 1, $\sigma_0^{2(0)} = 198{,}862$ et $k^{(0)} = 21{,}514$ (voir la section 3.1.2 pour une justification de ces valeurs). À l'aide des équations (3.2.1) à (3.2.6), nous calculons les estimateurs pour chacun des paramètres. Nous avons aussi fait varier la valeur du n_0 afin de maximiser le taux de bonne classification. Nous avons obtenu une classification optimale avec $n_0 = 0,025$. Les résultats sont donnés dans le tableau 3.2.1. Par conséquent, nous pouvons écrire la densité de la charge virale comme :

$$f(\underset{\sim}{y}_i) = 0{,}500 f_0(\underset{\sim}{y}_i | X_i \underset{\sim}{\beta}_i; 8{,}464 I) + 0{,}500 f_1(\underset{\sim}{y}_i | X_i \underset{\sim}{\beta}_i; 19808{,}875 I),$$

où les $\underset{\sim}{\beta}_i$ sont les coefficients obtenus par l'algorithme EM (voir le tableau 3.2.1). Nous remarquons que les valeurs de $\hat{\sigma}_0^2$ et de $\hat{k}$ sont légèrement supérieures à celles obtenues avec le modèle classique.

Tableau 3.2.1: Tableau des estimateurs du modèle bayésien

Estimateurs du modèle contaminé bayésien				
$\hat{\varepsilon} = 0{,}500 \quad \hat{\sigma}_0^2 = 8{,}464 \quad \underset{\sim}{\hat{\beta}}_0 = (3{,}302; -7{,}475; 20{,}473)$				
$\hat{k} = 2340{,}368 \qquad \underset{\sim}{\hat{\beta}}_1 = (25{,}248; 239{,}914; -182{,}752)$				
Patient	Groupe	$\hat{\beta}_{i0}$	$\hat{\beta}_{i1}$	$\hat{\beta}_{i2}$
1	1	3,771	-17,956	50,349
2^+	1	365,128	464,033	-986,860
3	1	96,475	-58,932	-35,599
4	1	2,249	-7,948	18,725
5^+	1	-150,821	1703,425	-2406,061
6	1	0,592	10,993	-2,954

Estimateurs du modèle contaminé bayésien				
Patient	Groupe	$\hat{\beta}_{i0}$	$\hat{\beta}_{i1}$	$\hat{\beta}_{i2}$
7	1	73,141	-69,091	-3,751
8	1	40,119	45,958	-147,611
9	1	18,126	-14,548	-6,444
10	1	8,562	125,142	-71,796
11^+	1	103,192	-456,549	1672,048
12^+	1	121,330	-1278,642	3581,093
13	1	-8,114	185,885	-344,245
14	1	12,619	-61,428	289,808
15	1	49,929	-249,991	822,281
16	1	22,909	-189,107	826,039
17	1	-26,613	187,080	-267,202
18^+	1	2,866	1624,176	-2461,341
19	1	3,337	-8,177	12,416
20	1	16,493	17,447	-16,751
21	1	2,702	-12,586	36,301
22	1	39,176	27,856	98,165
23	1	0,435	6,336	87,862
24^+	1	-228,314	1935,111	-2050,400
25	1	36,362	151,473	-357,768
26^+	1	-190,522	1410,206	-754,140
27^+	1	371,643	-935,764	1240,462
28	1	-0,644	39,382	-70,002
29	1	-26,967	221,228	-157,451
30	1	15,010	331,590	-371,397
31^+	1	-169,076	1298,539	-1947,813
32	1	30,849	-50,562	83,590
33^+	0	-123,738	886,717	-1550,354
34	0	0,986	8,049e-3	-4,390e-2
35	0	6,047	-16,613	24,259
36	0	0,976	6,932e-15	0,000

Estimateurs du modèle contaminé bayésien				
Patient	Groupe	$\hat{\beta}_{i0}$	$\hat{\beta}_{i1}$	$\hat{\beta}_{i2}$
37	0	0,986	-6,014e-2	0,132
38	0	1,431	-1,261	1,656
39	0	1,602	-2,075	3,044
40	0	0,990	-5,081e-2	7,841e-2
41^{+}	0	28,610	-179,979	412,541
42	0	3,197	-7,418	11,043
43	0	3,658	-9,024	13,475
44^{+}	0	-15,371	532,467	-1151,113
45	0	0,274	3,137	-4,267
46^{+}	0	-46,717	200,089	-252,285
47	0	-5,762	31,451	-42,064
48^{+}	0	9,429	-57,512	134,070
49	0	-0,295	8,646	-14,828
50	0	2,609	-4,970	6,962
51	0	4,214	-11,226	17,048
52	0	0,851	13,864	-31,058
53	0	0,976	-2,080e-14	0,000
54	0	0,812	1,005	-1,634
55	0	0,976	-2,773e-14	0,000
56	0	2,828	-6,005	8,656
57	0	1,146	-0,165	-5,883e-2
58^{+}	0	36,907	-113,578	162,622

De plus, la même explication pour les valeurs extrêmes des $\underset{\sim}{\beta}_i$ classiques est valable ici dans le cas bayésien (voir la sous-section 3.1.2) étant donné qu'à la section suivante, nous affirmons que les coefficients $\underset{\sim}{\beta}_i$ du modèle classique et ceux du modèle bayésien sont semblables.

3.2.3. Classification des patients

Encore une fois, nous pouvons déterminer dans quel groupe chaque patient est classé en calculant les ε_i. Ainsi, la valeur du ε_i obtenue à la dernière itération nous indique dans quel groupe est classé chacun des patients. Dans un contexte bayésien, l'algorithme réussit également à classer 47 patients dans leur groupe respectif comme le modèle classique. Nous obtenons donc encore un taux de mauvaise classification de 0,190. La valeur du ε_i à la dernière itération est présente dans le tableau 3.3.1 à la section 3.3 qui permet de comparer les deux méthodes.

3.3. Comparaison des résulats de la classification des patients à l'aide des méthodes classique et bayésienne

Lorsque nous comparons le taux de bonne classification, nous voyons que les deux modèles contaminés classique et bayésien sont équivalents. En effet, les deux méthodes nous permettent de classifier correctement 47 patients dans leur groupe. Le tableau 3.3.1 nous présente respectivement le numéro de patient (de 1 à 58), le groupe auquel il appartient, le groupe dans lequel il a été classé par les méthodes classique et bayésienne et la valeur du ε_i à la dernière itération de l'algorithme EM. Encore une fois, ces deux dernières informations sont écrites dans les quatre dernières colonnes du tableau, les deux premières pour le modèle contaminé classique et les deux dernières pour le modèle contaminé bayésien. En regardant ce tableau, nous remarquons que les deux modèles classent exactement tous les patients de la même façon. En effet, les derniers ϵ_i donnés par les deux modèles sont de même ordre de grandeur pour tous les patients.

De plus, afin de comparer les coefficients $\underset{\sim}{\beta_i}$ obtenus par la méthode classique et ceux obtenus par l'approche bayésienne, nous avons représenté graphiquement

les résultats. Les graphiques 3.3.1, 3.3.2 et 3.3.3 indiquent en abscisse les coefficients $\underset{\sim}{\beta}_i$ classiques et en ordonnée les coefficients $\underset{\sim}{\beta}_i$ bayésiens et chaque graphique présente une composante β_{ij}, $j = 0, 1, 2$, des vecteurs $\underset{\sim}{\beta}_i$. Les patients classés dans le groupe 0 sont représentés par des $\cdot$ et les patients classés dans le groupe 1 par des $^+$. L'allure des trois graphiques nous permet de conclure que les coefficients des deux modèles sont semblables, quoique l'étendue des estimateurs bayésiens est moindre que celle des estimateurs classiques.

Enfin, pour terminer cette comparaison entre les deux modèles contaminés, nous comparons les estimateurs des paramètres ε, σ_0^2 et k. La valeur du ε est de 0,500 pour le modèle classique et aussi de 0,500 pour le modèle bayésien. Nous concluons donc que ces valeurs sont pareilles pour les deux modèles. Le modèle classique et le modèle bayésien ont comme valeur de σ_0^2, 7,740 et 8,464 respectivement. Nous pouvons donc conclure que les deux estimateurs sont aussi similaires. Enfin, la valeur du k obtenue du modèle classique est de 3226,989 tandis que le modèle bayésien a donné une valeur de 2340,368. Notons ici que la valeur du k obtenue à l'aide du modèle contaminé bayésien est plus petite que celle obtenue avec l'approche classique. La similitude des estimateurs nous permet de conclure que les deux modèles sont équivalents, quoique le modèle bayésien donne une valeur de k inférieure.

Dans ce dernier chapitre, nous avons présenté deux approches afin de classifier les patients dans leur groupe respectif en utilisant un modèle contaminé. Nous avons tout d'abord estimé les paramètres du modèle contaminé à l'aide de l'approche classique. Puis, nous avons utilisé l'approche bayésienne qui introduit de l'information *a priori* au modèle classique, pour trouver les estimateurs des paramètres du modèle contaminé. Nous avons ensuite comparé le taux de bonne classification obtenu par les deux méthodes et nous avons conclu que ces deux

méthodes sont équivalentes, étant donné que chacune d'elles a obtenu un taux de mauvaise classification de 0,190. Finalement, nous avons comparé les coefficients des paramètres $\underset{\sim}{\beta}_i$ classiques et bayésiens ainsi que les estimateurs des paramètres ε, σ_0^2 et k et nous avons conclu que tous ces paramètres sont semblables sauf le paramètre k.

Tableau 3.3.1: Tableau de classification des patients à l'aide de modèles contaminés classique et bayésien

		Classique		Bayésien	
Patient	Vrai groupe	Groupe classé	Dernier ε_i	Groupe classé	Dernier ε_i
1	1	0	1,457e-17	0	5,330e-17
2	1	1	1,000	1	1,000
3	1	1	1,000	1	1,000
4	1	0	2,042e-23	0	1,628e-22
5	1	1	1,000	1	1,000
6	1	0	6,467e-13	0	1,983e-12
7	1	1	1,000	1	1,000
8	1	1	1,000	1	1,000
9	1	1	1,000	1	0,999
10	1	1	1,000	1	1,000
11	1	1	1,000	1	1,000
12	1	1	1,000	1	1,000
13	1	1	1,000	1	1,000
14	1	0	1,951e-9	0	3,305e-9
15	1	1	1,000	1	1,000
16	1	1	1,000	1	1,000
17	1	1	1,000	1	1,000
18	1	1	1,000	1	1,000
19	1	0	2,117e-23	0	1,682e-22
20	1	1	1,000	1	1,000

Patient	Vrai groupe	Classique		Bayésien	
		Groupe classé	Dernier ε_i	Groupe classé	Dernier ε_i
21	1	0	8,612e-21	0	4,938e-20
22	1	1	1,000	1	1,000
23	1	1	1,000	1	1,000
24	1	1	1,000	1	1,000
25	1	1	1,000	1	1,000
26	1	1	1,000	1	1,000
27	1	1	1,000	1	1,000
28	1	0	5,242e-9	0	2,468e-9
29	1	1	1,000	1	1,000
30	1	1	1,000	1	1,000
31	1	1	1,000	1	1,000
32	1	1	1,000	1	1,000
33	0	1	1,000	1	1,000
34	0	0	3,031e-11	0	7,969e-11
35	0	0	2,608e-16	0	1,071e-15
36	0	0	2,945e-18	0	1,476e-17
37	0	0	1,668e-16	0	7,114e-16
38	0	0	5,458e-20	0	3,200e-19
39	0	0	5,399e-13	0	1,667e-12
40	0	0	2,945e-18	0	1,476e-17
41	0	1	1,000	1	1,000
42	0	0	1,166e-21	0	7,901e-21
43	0	0	1,213e-21	0	8,196e-21
44	0	1	1,000	1	1,000
45	0	0	1,526e-21	0	1,010e-20
46	0	1	1,000	1	1,000
47	0	0	3,693e-9	0	4,447e-9
48	0	0	1,112e-3	0	2,632e-4
49	0	0	5,073e-19	0	2,456e-18
50	0	0	3,092e-11	0	8,118e-11

Patient	Vrai groupe	Classique		Bayésien	
		Groupe classé	Dernier ε_i	Groupe classé	Dernier ε_i
51	0	0	2,587e-16	0	1,063e-15
52	0	0	9,460e-14	0	1,943e-13
53	0	0	1,668e-16	0	7,114e-16
54	0	0	5,449e-20	0	3,194e-19
55	0	0	2,945e-18	0	1,476e-17
56	0	0	5,639e-13	0	1,734e-12
57	0	0	3,055e-11	0	8,029e-11
58	0	0	4,526e-9	0	5,487e-9

Figure 3.3.1: *Graphique des coefficients β_{i0} (· groupe 0, + groupe 1)*

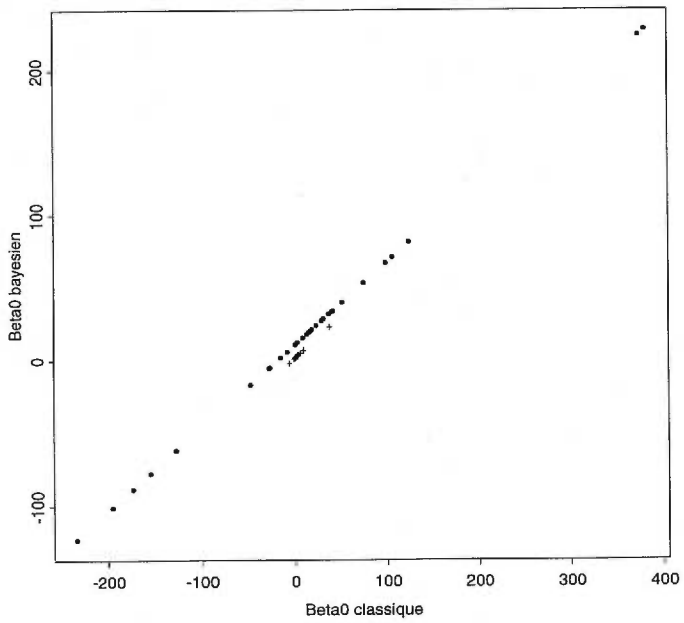


Figure 3.3.2: *Graphique des coefficients β_{i1} (· groupe 0, + groupe 1)*

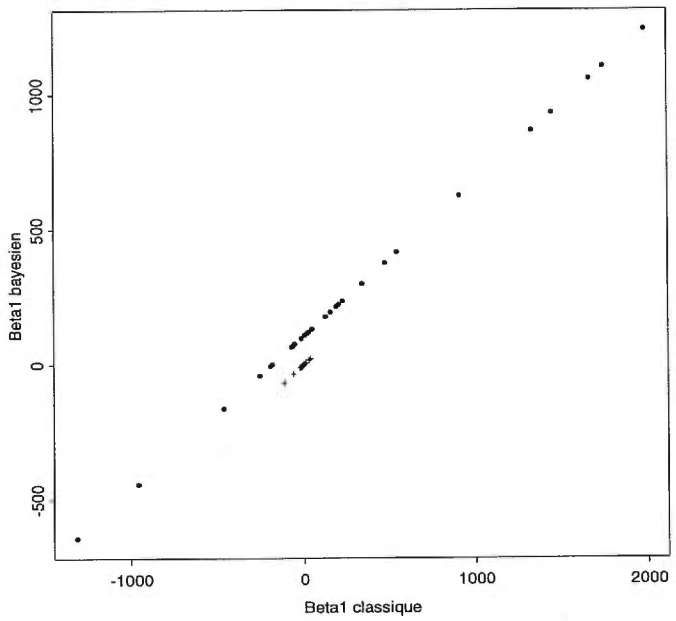
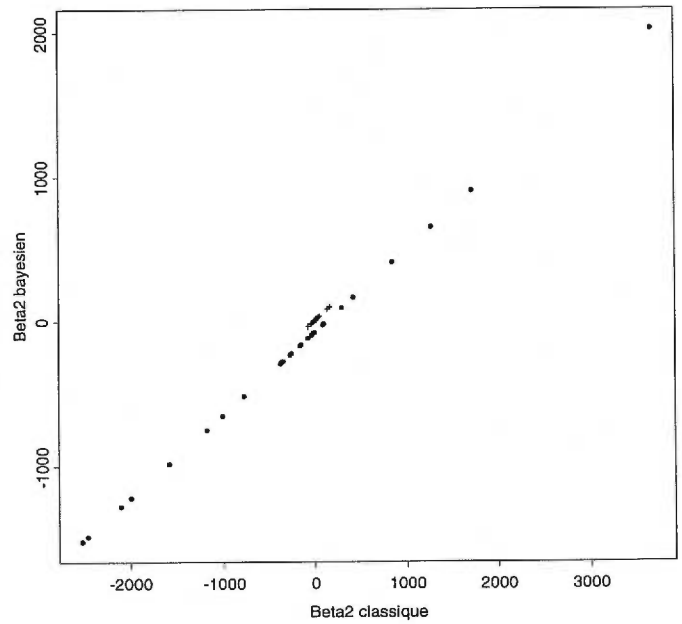


Figure 3.3.3: *Graphique des coefficients β_{i2} (· groupe 0, + groupe 1)*



CONCLUSION

Dans ce mémoire, nous avons présenté quelques méthodes permettant de classifier des patients séropositifs dans leur groupe respectif, selon leur évolution vers la maladie du sida. Les deux principales méthodes que nous avons expliquées ont été développées dans des contextes classique et bayésien.

Tout d'abord, nous avons modélisé la charge virale en fonction du temps à l'aide d'une spline estimant cette fonction. Dans le contexte classique, l'estimation de cette fonction s'est faite en utilisant une spline de lissage. Cette spline a été développée grâce à une approximation de la fonction par une série de Taylor tronquée. Nous avons trouvé la spline de lissage minimisant la fonction de perte quadratique pénalisée. Après avoir calculé les coefficients de cette spline pour chacun des patients, nous avons effectué une analyse discriminante afin de savoir si les coefficients permettent une bonne classification des patients. Cette analyse discriminante construit une fonction qui permet de classifier un patient selon la valeur de ses coefficients. La classification se fait comme suit : la fonction classe un patient dans le même groupe que les patients qui ont des coefficients les plus proches des coefficients de ce patient. Le taux de mauvaise classification est ensuite calculé car l'appartenance au groupe est connue pour tous les patients. La spline de lissage développée a permis un taux d'erreur de classification de 0,069. Avec l'approche bayésienne, nous avons construit une spline tenant compte de l'information *a priori* de la fonction et celle donnée par les observations. Nous avons utilisé la méthode proposée par O'Hagan qui base le développement de la

spline sur un modèle de régression. Une fois les coefficients de cette spline calculés, nous avons obtenu un taux de mauvaise classification de 0,069 tout comme dans le cas classique. Par conséquent, nous avons conclu que les deux approches sont équivalentes pour l'estimation de la fonction à l'aide d'une spline.

Puis, toujours dans le but de classifier les patients selon la charge virale, nous avons développé un modèle qui tient compte de l'évolution différente des deux groupes de patients, en utilisant un mélange de deux lois, appelé modèle contaminé. Ce modèle a également été construit en utilisant des approches classique et bayésienne. Pour estimer les paramètres de ce modèle, nous avons utilisé l'algorithme EM. Une fois les paramètres estimés, nous avons effectué une classification des patients. Pour classifier les patients, nous avons évalué la densité au point donné par la valeur de la charge virale pour chacune des deux densités, puis nous avons classé chaque patient dans le groupe pour lequel la densité était plus grande. Pour l'approche classique, nous avons réussi à classer 47 patients dans leur groupe. Avec l'approche bayésienne, nous avons aussi classé correctement les mêmes 47 patients. En comparant les résultats des deux approches, nous avons vu qu'en général, les paramètres estimés à l'aide des deux approches sont semblables et par conséquent, nous avons conclu que les approches classique et bayésienne sont équivalentes pour cette méthode.

Pour terminer, à la lumière des résultats, nous pouvons conclure dans notre cas que la spline et l'analyse discriminante permettent de mieux classifier les patients dans leur groupe respectif que le modèle contaminé et la classification. Cette conclusion demeure vraie dans un contexte classique comme dans un contexte bayésien. En effet, comme nous l'avons vu, les taux de mauvaise classification obtenus à l'aide de splines sont inférieurs à ceux obtenus à partir de modèles contaminés.

Annexe A

PROGRAMMES INFORMATIQUES

```
#Tous les programmes qui suivent sont definis pour le logiciel
#Splus.


#1.PROGRAMME POUR LA SPLINE CLASSIQUE DE LA CHARGE VIRALE

#a)FONCTION pour la matrice de covariance de X appelee omega (voir
#section 2.1.3) :

omega_function(n){
w_matrix(1:n*n,ncol=n,nrow=n)
 for (i in 1:n){
  for (j in 1:n){
   if (i!=j)
    {w[i,j]_1/(2*((i-1)+(j-1)+1)*fact(i-1)*fact(j-1))}
   else
    {w[i,i]_1/((2*(i-1)+1)*(fact(i-1)^2))}
  w[j,i]_w[i,j]
  }
 }
return(w)
}

#b)FONCTION pour la matrice X (voir section 2.1.3) :

#pt: matrice contenant les variables temps et charge virale
#k: degre de la spline
```

```
lesx_function(pt,k){
XX_matrix(1:length(pt[,1])*k,ncol=k,nrow=length(pt[,1]))
XX[,1]_1
 for (i in 1:length(pt[,1])){
  for (j in 2:k){
   XX[i,j]_pt[i,1]^(j-1)/gamma(j)
  }
 }
return(XX)
}

#c)FONCTION pour la matrice H, f=X*teta=HY ou
#H=X*inverse(X'X+n*lambda*omega)*X' :

#lambda: parametre de lissage

math_function(pt,k,lambda){
#trouve les valeurs et les vecteurs propres pour l'inverse
hv_eigen(t(lesx(pt,k))%*%lesx(pt,k)+length(pt[,1])*lambda
*omega(k))
hval_hv$values
hvect_hv$vectors
h_lesx(pt,k)%*%hvect%*%diag(1/hval)%*%t(hvect)%*%t(lesx(pt,k))
return(h)
}

#d)FONCTION pour faire la validation croisee (voir equation
#2.1.6) :

GCV_function(pt,k,lambda){
h_math(pt,k,lambda)
#numerateur
terme1_(diag(length(pt[,1]))-h)%*%pt[,2]
#denominateur
terme2_diag(diag(length(pt[,1]))-h)
GCV_mean((terme1/terme2)^2)
return(GCV)
}
```

```
#e)COMMANDES pour trouver le k donnant une CV minimum et le lambda
#correspondant :

#group1 et group0: vecteurs contenant le numero des patients
#ap: jeu de donnees

simul1_for (i in c(group1,group0)){
#construit une matrice contenant les variable temps et charge
#virale standardisees
pt_cbind(ap[ap[,1]==i,3]/max(ap[ap[,1]==i,3]),
(ap[ap[,1]==i,13])/400)
 for (k in 2:10){
  CV_function(i){GCV(pt,k,i)}
  #trouve le minimum en partant a lambda=0.5
  mini1_nlmin(CV,0.5)
  print(c(i,k,mini1$x,CV(mini1$x)))
 }
}

#f)COMMANDES pour calculer les thetas pour k=3 (voir proposition
#2.1.2) :

#lambda3: vecteur contenant les valeurs de lambda donnant une CV
#minimum

mateta3_matrix(0,nrow=3,ncol=length(lambda3))
tetas3_for (j in c(group1,group0)){
pt_cbind(ap[ap[,1]==j,3]/max(ap[ap[,1]==j,3]),
(ap[ap[,1]==j,13])/400)
#trouve les valeurs et les vecteurs propres pour l'inverse
h_eigen(t(lesx(pt,3))%*%lesx(pt,3)+length(pt[,1])*lambda3[j]
*omega(3))
hval_h$values
hvect_h$vectors
mateta3[1:3,j]_hvect%*%diag(1/hval)%*%t(hvect)%*%t(lesx(pt,3))
%*%pt[,2]
}
```

```
#2.PROGRAMME POUR LA SPLINE BAYESIENNE DE LA CHARGE VIRALE

#a)FONCTION pour les coefficents de regression lineaire, la somme
#des residus de la regression et le nombre d'observations pour
#chaque patient :

patient_function(i){
#attribue a j le numero du patient
j_group[i]
pt_cbind(ap[ap[,1]==j,3]/max(ap[ap[,1]==j,3]),
(ap[ap[,1]==j,13])/400)
#construit la matrice X
un_rep(1,length(pt[,1]))
matX_cbind(un,pt[,1])
matXX_cbind(matX,(pt[,1])^2/2)
matXX_matrix(matXX,ncol=3,byrow=F)
#calcule le nombre d'observations
lengtmp_length(pt[,1])
#calcule les coefficients de regression
beta_solve(t(matXX)%*%matXX)%*%t(matXX)%*%pt[,2]
#calcule la somme des residus
sommekn_sum((pt[,2]-matXX%*%beta)^2)
return(c(beta,sommekn,lengtmp))
}

#b)COMMANDES pour storer les coefficents de regression, la somme
#des residus et le nombre d'observations :

ebr_matrix(0,58,4)
leng_c(1:58)

for(i in 1:58){
 creg_patient(i)
 #store les coefficients et la somme des residus
 ebr[i,]_creg[1:4]
 #store le nombre d'observations
 leng[i]_creg[5]
}
```

```
#c)COMMANDES pour les coefficients b2 avec c=1 (voir equation
#2.2.6) :

splineoh_matrix(0,ncol=9,nrow=58)
for(l in 1:58){
 i_group[l]
 #standardise les variables temps et charge virale
 temps_ap[ap[,1]==i,3]/max(ap[ap[,1]==i,3])
 y_ap[ap[,1]==i,13]/400
 cc_1
 res_rep(0,length(temps))
 b2_matrix(0,ncol=3,nrow=length(temps))
 #construit la matrice C
 mat1_outer(temps,temps,FUN="*")
 mat2_exp(-cc*abs(outer(temps,temps,FUN="-")))
 CC_(1+mat1+0.25*(mat1^2))*(730*mat2)
 #construit la matrice G
 gmat_cbind(rep(1,length(temps)),temps,(temps^2)/2)
 #calcule une estimation de sigma carre
 sig2_ebr[l,4]/(leng[l]-3)
 #calcule la matrice A et son inverse
 A_sig2*diag(length(temps))+CC
 sA_svd(A)
 ssA_sA$v%*%diag(1/sA$d)%*%(t(sA$u))
 #calcule le vecteur bo
 bo_solve(t(gmat)%*%ssA%*%gmat)%*%t(gmat)%*%ssA%*%y
  for(j in 1:length(temps)){
   #calcule la matrice S
   St_mat2[j,]*(730*gmat)
   #calcule la matrice Q
   Qt_diag(3)-t(gmat)%*%ssA%*%St
   #calcule les coefficients b2
   b2[j,]_t(St)%*%ssA%*%y+t(Qt)%*%bo
   res[j]_y[j]-gmat[j,]%*%b2[j,]
  }
 #store la moyenne des coefficients et la somme des residus
 splineoh[l,]_cbind(l,gr[l],mean(b2[,1]),mean(b2[,2]),
 mean(b2[,3]),sum(res^2))
}
```

```
#3.PROGRAMME POUR LE MODELE CONTAMINE CLASSIQUE

#a)COMMANDES pour les valeurs de depart pour k et sigma carre
#(voir section 3.1.2) :

sse0_rep(0,length(group0))
estsig0_rep(0,length(group0))
leng0_rep(0,length(group0))
for(j in 33:58){
 i_group[j]
 pt_cbind(ap[ap[,1]==i,3]/max(ap[ap[,1]==i,3]),
 (ap[ap[,1]==i,13])/400)
 un_rep(1,length(pt[,1]))
 matX_cbind(un,pt[,1])
 matXX_cbind(matX,(pt[,1])^2/2)
 matXX_matrix(matXX,ncol=3,byrow=F)
 #calcule les valeurs predites pour y
 ych_matXX%*%solve(t(matXX)%*%matXX)%*%t(matXX)%*%pt[,2]
 #calcule les residus
 res_ych-pt[,2]
 #calcule la variance estimee des patients du groupe 0
 sse0[j-32]_sum(res^2)
 estsig0[j-32]_(1/(sum(length(pt[,1])-3)))*sse0[j-32]
 leng0[j-32]_length(pt[,1])
}

sse1_rep(0,length(group1))
estsig1_rep(0,length(group1))
leng1_rep(0,length(group1))
for(j in 1:32){
 i_group[j]
 pt_cbind(ap[ap[,1]==i,3]/max(ap[ap[,1]==i,3]),
 (ap[ap[,1]==i,13])/400)
 un_rep(1,length(pt[,1]))
 matX_cbind(un,pt[,1])
 matXX_cbind(matX,(pt[,1])^2/2)
 matXX_matrix(matXX,ncol=3,byrow=F)
 ych_matXX%*%solve(t(matXX)%*%matXX)%*%t(matXX)%*%pt[,2]
 res_ych-pt[,2]
```

```
 #calcule la variance estimee des patients du groupe 1
 sse1[j]_sum(res^2)
 estsig1[j]_(1/(sum(length(pt[,1])-3)))*sse1[j]
 leng1[j]_length(pt[,1])
}

#calcule la variance estimee du groupe 0 et du groupe 1
sum0_(1/(sum(leng0)-26*3))*sum(estsig0)
sum1_(1/(sum(leng1)-32*3))*sum(estsig1)

#Valeurs de depart :
sigma_sum0
k_sum1/sum0

#b)COMMANDES pour pour les valeurs des parametres :

kk_0
ssigma_0
eeps_0
esper_c(1:58)
eesper_c(1:58)
#initialise les parametres
k_sum1/sum0
sigma_sum0
#ici, les coefficients beta ne varient pas
beta_ebr[,1:3]
eps_0.448
#ici, la somme des residus ne varie pas non plus
sommekn_ebr[,4]
num_c(1:58)
den_c(1:58)
for(h in 1:500){
 #conserve la derniere valeur afin de verifier s'il y a
 #convergence
 kk_k
 ssigma_sigma
 eeps_eps
 eesper_esper
  for(i in 1:58){
```

```
   j_group[i]
   pt_cbind(ap[ap[,1]==j,3]/max(ap[ap[,1]==j,3]),
   (ap[ap[,1]==j,13])/400)
   #calcule le epsilon i (voir equation 3.1.1)
   num[i]_(1-eps)*exp(-1*sommekn[i]/2/k/sigma)/k^(leng[i]/2)
   den[i]_num[i]+eps*exp(-1*sommekn[i]/2/sigma)
   esper[i]_num[i]/den[i]
  }
 #calcule le epsilon (voir equation 3.1.2)
 eps_mean(1-esper)
 #calcule le k (voir equation 3.1.5)
 k_max(1,sum(esper*sommekn)/sigma/sum(esper*leng))
 #calcule le sigma carre (voir equation 3.1.4)
 sigma_sum(sommekn*(1-esper+esper/k))/sum(leng)
 #verifie s'il y a convergence des parametres
 erreurk_abs(kk-k)/(1+kk)
 erreurs_abs(ssigma-sigma)/(1+ssigma)
 erreure_abs(eeps-eps)/(1+eeps)
 errmax_max(erreurk,erreurs,erreure)
 print(c(h,k,sigma,eps))
 #verifie la condition de convergence
  if(errmax<0.01)
   {break}
}

#4.PROGRAMME POUR LE MODELE CONTAMINE BAYESIEN

#a)COMMANDES pour les valeurs des parametres :

#ici, on a besoin de matrices de 3 dimensions
new_array(0,dim=c(3,3,58))
new1_array(0,dim=c(3,1,58))
new2_array(0,dim=c(3,1,58))
new3_array(0,dim=c(3,3,58))
new4_array(0,dim=c(3,3,58))
for(i in 1:58){
 j_group[i]
 pt_cbind(ap[ap[,1]==j,3]/max(ap[ap[,1]==j,3]),
 (ap[ap[,1]==j,13])/400)
```

```
 un_rep(1,length(pt[,1]))
 matX_cbind(un,pt[,1])
 matXX_cbind(matX,((pt[,1])^2)/2)
 matXX_matrix(matXX,ncol=3,byrow=F)
 #construit les matrices X'X
 new[,,i]_t(matXX)%*%matXX
 #initialise les matrices et les vecteurs pour le calcule de beta0
 #et de beta1
 new1[,,i]_t(matXX)%*%matXX%*%beta[i,]
 new2[,,i]_t(matXX)%*%matXX%*%beta[i,]
 new3[,,i]_t(matXX)%*%matXX
 new4[,,i]_t(matXX)%*%matXX
}

kk_0
ssigma_0
eeps_0
bbetta0_c(1:3)
bbetta1_c(1:3)
eesper_c(1:58)
esper_c(1:58)
betaj_ebr[,1:3]
bbetaj_matrix(0,ncol=3,nrow=58)
k_sum1/sum0
sigma_sum0
#initialise les valeurs de beta0 et de beta1
betta0_c(mean(ebr[1:32,1]),mean(ebr[1:32,2]),mean(ebr[1:32,3]))
betta1_c(mean(ebr[33:58,1]),mean(ebr[33:58,2]),mean(ebr[33:58,3]))
eps_0.448
num_c(1:58)
den_c(1:58)
#valeur de la constante n0
n0_0.025
numbeta_array(0,dim=c(3,3,58))
denbeta_array(0,dim=c(3,1,58))
for(h in 1:500){
 z_rep(0,length(esper))
 sommekn_rep(0,58)
 #conserve la derniere valeur afin de verifier s'il y a
```

```
#convergence
kk_k
ssigma_sigma
eeps_eps
bbetta0_betta0
bbetta1_betta1
bbetaj_betaj
sommebeta0_0
sommebeta1_0
sumesp0_0
sumesp1_0
 for(i in 1:58){
  l_group[i]
  pt_cbind(ap[ap[,1]==1,3]/max(ap[ap[,1]==1,3]),
  (ap[ap[,1]==1,13])/400)
  un_rep(1,length(pt[,1]))
  matX_cbind(un,pt[,1])
  matXX_cbind(matX,((pt[,1])^2)/2)
  matXX_matrix(matXX,ncol=3,byrow=F)
  y_pt[,2]
  #calcule la somme des residus
  matrice_sum((y-matXX%*%betaj[i,])^2)
  num[i]_(1-eps)*exp(-1*matrice/2/k/sigma)/k^(leng[i]/2)
  den[i]_num[i]+eps*exp(-1*matrice/2/sigma)
  esper[i]_num[i]/den[i]
  sommekn[i]_matrice
   if(esper[i]>=.5)
    #le patient est classe dans le groupe 1
    {z[i]_1}
   else
    #le patient est classe dans le groupe 0
    {z[i]_0}
  #calcule les beta i (voir equation 3.2.2)
  betaj[i,]_solve(new[,,i])%*%(t(matXX)%*%y)/(1+n0)
  +(n0*(1-esper[i])*betta0+esper[i]*betta1/k)/
  ((1+n0)*(1-esper[i]+esper[i]/k))
  #calcule les (1-epsilon i)X'X pour beta0 (voir equation 3.2.3)
  new1[,,i]_(1-esper[i])*new[,,i]%*%betaj[i,]
  #calcule les (epsilon i)X'X pour beta1 (voir equation 3.2.4)
```

```
  new2[,,i]_esper[i]*new[,,i]%*%betaj[i,]
  #calcule les (1-epsilon i)X'Xbeta i pour beta0 (voir equation
  #3.2.3)
  new3[,,i]_(1-esper[i])*new[,,i]
  #calcule les (epsilon i)X'Xbeta i pour beta1 (voir equation
  #3.2.4)
  new4[,,i]_esper[i]*new[,,i]
  #calcule les (epsilon i)(beta i-beta1) pour k (voir equation
  #3.2.6)
  sommebeta1_sommebeta1+esper[i]*(t(betaj[i,]-betta1)%*%new[,,i]
  %*%(betaj[i,]-betta1))
  #calcule les (1-epsilon i)(beta i-beta0) pour sigma carre (voir
  #equation 3.2.5)
  sommebeta0_sommebeta0+(1-esper[i])*(t(betaj[i,]-betta0)
  %*%new[,,i]%*%(betaj[i,]-betta0))
 }
#calcule le epsilon (voir equation 3.2.1)
eps_mean(1-esper)
#calcule le k (voir equation 3.2.6)
k_max(1,(sum(esper*sommekn)+n0*sommebeta1)/sigma
/(sum(esper*(leng+3))))
#calcule le sigma carre (voir equation 3.2.5)
sigma_(sum(sommekn*(1-esper+esper/k))+n0*sommebeta0
+n0*sommebeta1/k)/(sum(leng+3))
#calcule le beta0 (voir equation 3.2.3)
betta0_solve(apply(new3,c(1,2),sum))%*%apply(new1,c(1,2),sum)
#calcule le beta1 (voir equation 3.2.4)
betta1_solve(apply(new4,c(1,2),sum))%*%apply(new2,c(1,2),sum)
#verifie s'il y a convergence des parametres
erreurk_abs(kk-k)/(1+kk)
erreurs_abs(ssigma-sigma)/(1+ssigma)
erreure_abs(eeps-eps)/(1+eeps)
erreurb0_abs(bbetta0-betta0)/(1+bbetta0)
erreurb1_abs(bbetta1-betta1)/(1+bbetta1)
errmax_max(erreurk,erreurs,erreure,erreurb0,erreurb1)
print(c(h,k,sigma,betta0,betta1,eps,errmax))
 if(errmax<0.01)
  {break}
}
```

BIBLIOGRAPHIE

Angers, J.-F. et Delampady, M. (1992). Hierarchical bayesian curve fitting and smoothing. *The Canadian Journal of Statistics, 20:35–49.*

Bennaghmouch, Z. (1992). *Estimation bayésienne d'une fonction avec contraintes.* Mémoire de maîtrise, Département de mathématiques et de statistique, Université de Montréal.

De Boor, C. (1978). *A practical guide to splines.* Springer-Verlag, Berlin.

Dempster, A. P., Laird, N. M. et Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society (série B), 39:1–38.*

Desgagné, A. (1998). *Comparaison bayésienne de coûts entre deux traitements pour des mélanges de lois.* Mémoire de maîtrise, Département de mathématiques et de statistique, Université de Montréal.

Drusano, G. L. et Stein D. S. (1998). Mathematical modeling of the interrelationship of CD4 lymphocyte count and viral load changes induced by the protease inhibitor indinavir. *Antimicrobial Agents and Chemotherapy, 42(2):358-361.*

Eubank, R. L. (1988). *Smoothing splines and non parametric regression.* Marcel Decker, New York.

O'Hagan, A. (1978). Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society (série B), 40:1–42.*

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics, 33:1065–1076.*

Sabin, C. A., Devereux, H., Phillips, A. N., Janossy, G., Loveday, C. et Lee, C. A. (1998). Immune markers and viral load after HIV-1 seroconversion as predictors of disease progression in a cohort of haemophilic men. *AIDS, 12(11):1347–1352.*

Smol'skaia, T. T., Sizova, N. V., Korovina, G. I., Maslov, V. P., Kevlova, N. A., Novikova, V. A. et Bogoiavlenskii, G. V. (1999). The use of a method for the quantitative determination of

HIV-1 RNA for assessing the severity and prognosis of the development of the disease. *Zhurnal Mikrobiologii, Epidemiologii i Immunobiologii, 1:57–59.*

Tamalet, C., Lafeuillade, A., Fantini, J., Poggi, C. et Yahi, N. (1997). Quantification of HIV-1 viral load in lymphoid and blood cells: assessment during four-drug combination therapy. *AIDS, 11(7):895–901.*

Tanner, M. A. (1993). *Tools for statistical inference.* Springer-Verlag, New York, second edition.

Titterington, D., Smith, A. et Makov, U. (1985). *Statistical analysis of finite mixture distributions.* Wiley, New York.

Wahba, G. (1990). *Spline models for observational data.* SIAM, Philadelphia.

Wahba, G. et Wold, S. (1975). Fitting splines functions by cross validation. *Communications in Statistics, 4 (1):1–17.*

Wahba, G. et Kimeldorf, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications, 33:82–95.*

Zellner, A. (1971). *An introduction to bayesian inference in econometrics.* Wiley, New York.