# Étude longitudinale de la Qualité de Vie en cancérologie par mélanges de modèles mixtes

Antoine Barbieri[1,3,4] & Florence Cousson-Gélie[2,3,5] & Sophie Gourgou[3] & Caroline Mollevi[3] & Christian Lavergne[2,4]

[1] *Université de Montpellier, Place Eugène Bataillon - 34095 Montpellier Cedex 5 (Antoine.Barbieri@umontpellier.fr)*

[2] *Université Paul-Valéry Montpellier 3 (Christian.Lavergne@univ-montp3.fr)*

[3] *Institut régional du Cancer Montpellier / Val d'Aurelle - 208 Rue des Apothicaires, 34298 Montpellier (Sophie.Gourgou@icm.unicancer.fr, Caroline.Mollevi@icm.unicancer.fr)*

[4] *Institut Montpelliérain Alexander Grothendieck*

[5] *Laboratoire Epsylon EA 4556 (Florence.Cousson-Gelie@icm.unicancer.fr)*

**Résumé.** En cancérologie, la qualité de vie relative à la santé (QdV) n'est pas seulement un critère de jugement secondaire dans les essais cliniques, mais fait l'objet de nombreuses études afin d'améliorer la prise en charge des patients atteints d'un cancer. En effet, les caractéristiques des individus mesurées lors d'études psycho-sociales peuvent expliquer des différentes évolutions de QdV. Dans ce contexte, nous nous intéressons à un mélange de modèles mixtes pour l'étude longitudinale de la QdV. Les composantes du mélange traduisent les différentes trajectoires de QdV des patients des sous-populations homogènes de la population d'intérêt, et les modèles mixtes utilisés permettent la prise en compte de la variabilité induite par les mesures répétées au cours du temps grâce aux effets aléatoires. Une fois que la classification de la population d'intérêt est effectuée suivant les trajectoires de QdV des patients, nous cherchons dans un second temps à expliquer l'appartenance aux classes par l'intermédiaire d'un modèle de régression logistique (multinomiale si $K > 2$) suivant des variables explicatives permettant de caractériser un trait commun de la sous-population homogène.

**Mots-clés.** Trajectoires, Qualité de vie, Classes latentes, Modèles mixtes.

**Abstract.** In oncology, health-related quality of life (HRQoL) is not only a secondary endpoint in clinical trials, but is also the object of numerous studies aiming at improving patient management. Indeed, the subjects' characteristics measured in psycho-social studies may explain the different HRQoL evolutions. In this context, the analysis of finite mixture mixed models for the longitudinal study of HRQoL was considered. The mixture components correspond to the HRQoL trajectories associated with each homogeneous subpopulation. The dependence among repeated measurements over time from the same subject is taken into account through the random effect models defined for each component. Once the classification of the population of interest was performed through the patients' trajectories, a regression logistic model was used to explain the class-membership through explanatory variables which characterize a common trait of the subpopulations.

**Keywords.** Trajectories, Quality of life, Latent classes, Mixed models.

# 1 Introduction

En cancérologie, la qualité de vie relative à la santé (QdV) est un critère de jugement secondaire dans les essais cliniques. Hors de ce contexte, la QdV fait également l'objet de nombreuses études, par exemples psycho-sociales, afin d'améliorer la prise en charge des patients atteints de cancer. En effet, ce type d'études est réalisé afin de déterminer l'impact de certaines caractéristiques cliniques, socio-démographiques ou personnelles des patients sur l'évolution de la QdV. La connaissance de l'influence de ces covariables sur les trajectoires de QdV peut ainsi améliorer la prise en charge initiale du patient. Afin de distinguer des profils particuliers d'individus, nous faisons l'hypothèse que la population d'intérêt n'est pas homogène, et qu'il existe $K$ sous-populations homogènes non contrôlées qui se distinguent par différentes trajectoires moyennes de QdV. Dans ce travail, l'objectif est double :

- Déterminer les sous-populations homogènes ainsi que les trajectoires latentes de QdV qui les caractérisent (classification) ;
- Expliquer l'appartenance aux classes par l'intermédiaire de variables explicatives.

Lors du recrutement d'un essai clinique ou d'une étude de cohorte, une sélection précise de patients est effectuée suivant des critères d'inclusion et d'exclusion prédéfinis. Dans ce contexte, l'analyse de la QdV est réalisée sur une population homogène particulière répondant à des critères objectifs. Une étude sur ce type de données induit d'ores et déjà un trait commun de la population d'intérêt. L'objectif de la classification revient alors à déterminer si la population est homogène, et si elle ne l'est pas, établir une partition de sous-populations homogènes. Un mélange de modèles mixtes sera alors utilisé dans ce but. La vocation d'un mélange est de modéliser des données hétérogènes, ou supposées comme telles. Pour cela, une nouvelle source de variabilité est considérée : celle du modèle (Celeux et al., 2005). De nouveaux paramètres sont alors introduits pour modéliser cette hétérogénéité au sein de la population, provenant de l'existence de différents sous-groupes non contrôlés. Cette approche consiste en la modélisation de la variable aléatoire réponse et en la classification en sous-échantillons de la population d'intérêt.

Dans un deuxième temps, une régression logistique (multinomiale si $K > 2$) modélisera l'appartenance à une classe latente, qui est caractérisée par une trajectoire de QdV particulière, en fonction des variables explicatives telles que des observations caractérisant les traits de personnalité des individus.

# 2 Modélisation

En cancérologie, la QdV des patients $i (i = 1, \ldots, I)$, est évaluée par des questionnaires à différentes visites $v (v = 1, \ldots, n_i)$ tout au long de leur prise en charge. La référence en

Europe est le questionnaire EORTC [1] QLQ-C30. Il est composé de 30 items à plusieurs catégories de réponse $c(c = 1, \ldots, C_j)_{j=1,\ldots,30}$, regroupés en 15 dimensions uni ou multi-items (Fayer et al. 2001). Les réponses aux items (qualitatives ordinales) nous permettent ainsi d'évaluer indirectement la QdV des patients. En pratique, un score est calculé par dimension, chaque dimension étant ensuite analysée indépendamment les unes des autres. Dans ce travail, nous considérerons deux scores : $S^{(s)}$ le "score-somme" qui correspond à la somme des réponses aux items de la dimension considérées (Gorter et al. 2015) ; et $S^{(m)}$, recommandé par l'EORTC, qui correspond à la moyenne des réponses aux items standardisée sur une échelle de 0 à 100. En pratique, la modélisation des scores $S^{(m)}$ et $S^{(s)}$ est réalisée par l'intermédiaire de deux modèles mixtes différents. Pour $S^{(m)}$, nous utilisons un modèle linéaire mixte (LMM) tel que :

$$S_{iv}^{(m)} = \theta_i\left(t_{iv}\right) + \varepsilon_{iv}^{(m)}, \tag{1}$$

où $\varepsilon_{iv}^{(m)} \sim \mathcal{N}\left(0, \sigma_{\varepsilon_m}^2\right)$ et $\theta_i\left(t_{iv}\right)$ est le niveau de QdV du sujet $i$ au temps $t_{iv}$.

En revanche, le score $S^{(s)}$ est discret ordinal, et même qualitatif ordinal si la dimension ne comprend qu'un seul item. Dans ce cas, un modèle linéaire généralisé mixte (GLMM) pour données ordinales de la famille des modèles cumulatifs est utilisé (Proust-Lima et al., 2015) :

$$\Pr\left(S_{iv}^{(s)} \leq c | \theta_i\left(t_{iv}\right)\right) = F\left(\delta_c - \theta_i\left(t_{iv}\right)\right), \tag{2}$$

où $c = 0, \ldots, C$ et $F$ est la fonction de répartition de la loi normale centrée réduite. Un processus latent sous-jacent au modèle est caractérisé par :

$$\tilde{\theta}_i\left(t_{iv}\right) = \theta_i\left(t_{iv}\right) + \varepsilon_{iv}^{(s)},$$

avec $\varepsilon_{iv}^{(s)} \sim \mathcal{N}\left(0, \sigma_{\varepsilon_s}^2\right)$ et de telle sorte que,

$$\left\{S_{iv}^{(s)} = c\right\} \Leftrightarrow \tilde{\theta}_i\left(t_{iv}\right) \in [\delta_c, \delta_{c+1}[\ ,$$

où $-\infty = \delta_0 < \delta_1 < \ldots < \delta_C = +\infty$.

Pour partitionner la population d'intérêt suivant la trajectoire de QdV des patients, nous utilisons un mélange de modèles mixtes. Cette approche est basée sur l'hypothèse d'un mélange de $K$ distributions représentant chacune une classe $(\mathcal{C}_k)_{k=1,\ldots,K}$. Nous supposons que toutes les observations $S_i = (S_{i1}, \ldots, S_{in_i})$ d'un même individu $i$ sont issues d'une même composante du mélange, et la densité du mélange pour cet individu est :

$$f\left(S_i | \lambda, p\right) = \sum_{k=1}^{K} p_k f_k\left(S_i | \lambda_k\right),$$

1. European Organisation for Research and Treatment of Cancer

où $p = (p_1, \ldots, p_K)$ sont les proportions du mélange tel que $p_k > 0$ et $\sum_k p_k = 1$, et $f_k(S_i|\lambda_k)$ la densité marginale associée à la loi de probabilité considérée dans le mélange de paramètre $\lambda_k$. Les différentes classes correspondent aux différents profils de patient qui émergent suivant les différentes évolutions de QdV. Les composantes se différentient de part les paramètres qui caractérisent la trajectoire de QdV. Pour cela, nous faisons l'hypothèse que les trajectoires latentes ont une structure de modèles linéaires mixtes et que les sous-populations se distinguent via leur trajectoire moyenne de QdV. La trajectoire d'un individu $i (i = 1 \ldots, I)$ appartenant à $\mathcal{C}_k$ est caractérisée le processus latent $\theta_{ik}(t)$ défini tel que :

$$\theta_{ik}(t_{iv}) = x_{1i}(t_{iv})' \beta_k + x_{2i}(t_{iv})' \nu + u_i(t_{iv})' \xi_i, \quad (3)$$

où pour l'équation (3) :

- $\beta_k$ est le vecteur des $q_1$ effets fixes associé à $\mathcal{C}_k$, $x_{1i}(t_{iv})$ son vecteur design au temps $t_{iv}$ ;
- $\nu$ est le vecteur des $q_2$ effets fixes commun à toutes les classes latentes, $x_{2i}(t_{iv})$ son vecteur design au temps $t_{iv}$ ;
- $\xi_i$ est le vecteur des $q_\xi$ effets aléatoires individuels normalement distribués associé à $\mathcal{C}_k$, de moyenne nulle et de matrice de variance covariance $\Sigma_k$, et $u_i(t_{iv})$ son vecteur design associé au temps $t_{iv}$.

Les modèles de mélange sont donc de bons outils pour classifier les données sur la base des trajectoires de QdV. L'évolution linéaire au cours du temps ne paraissant pas adéquate pour modéliser ces trajectoires, nous prenons en considération une relation polynomiale ($\theta_{ik}(t_{iv}) = \sum_{q=0}^{3} t_{iv}^q \beta_{kq} + \sum_{q=0}^{3} t_{iv}^q \xi_{iq}$) permettant une évolution plus flexible au cours du temps de la QdV et une distinction plus efficace des classes par leur trajectoire.

Enfin, le choix du nombre de classes latentes $K$ est déterminé par l'intermédiaire du critère $ICL$ (Biernacki et al., 2000) qui correspond au $BIC$ auquel est ajouté une pénalisation relative à la séparabilité des classes :

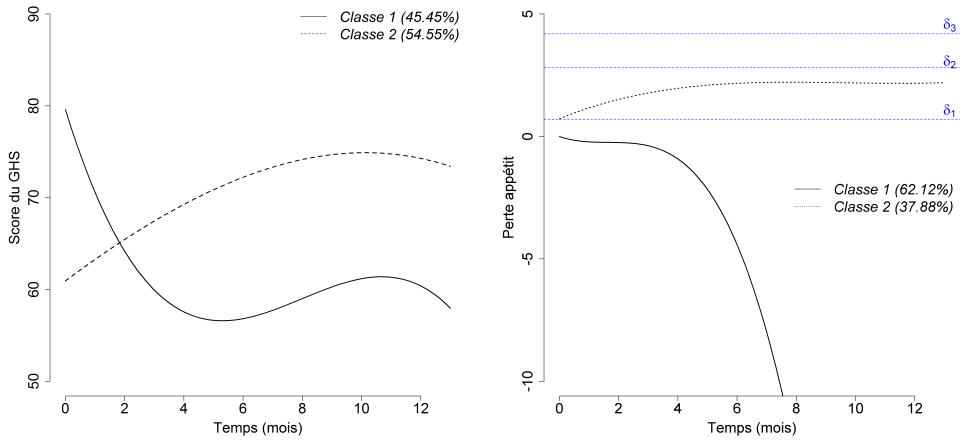
$$ICL = BIC - \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\pi_{ik}),$$

où $z_{ik}$ est la variable indicatrice d'appartenance à la classe $k$ égale à 1 si $\pi_{ik}$ est le maximum a posteriori, et 0 sinon. Soit, $\forall k \in \{1, \ldots, K\}$,

$$\pi_{ik} = \frac{p_k f_k(S_i|\lambda_k)}{\sum_{l=1}^{K} p_l f_l(S_i|\lambda_l)}.$$

# 3 Résultats

Cette approche est illustrée sur les données de l'étude Moral concernant 132 patientes atteintes d'un cancer du sein. Les mesures de QdV sont issues de la collecte de l'auto-




(a) GHS : Trajectoires latentes (b) AP : Trajectoires latentes

FIGURE 1 – Trajectoires associées aux différentes classes latentes pour la classification de la population suivant les scores de QdV des dimensions GHS et AP.

questionnaire EORTC QLQ-C30 à différentes visites au cours du traitement et du suivi. Ce questionnaire décompose la QdV en dimensions fonctionnelles et symptomatiques et en un statut global de santé. Le recueil des données incluant des variables socio-démographiques (âge, niveau d'études, situation familiale,...), cliniques (taille de la tumeur, grade,...), ainsi que des données psycho-sociales (anxiété, dépression, soutien social, optimisme,...) est réalisé à 6 visites (à la chirurgie puis à 1, 4, 7, 10 et 13 mois). Les deux types de mélanges présentés dans la partie précédente peuvent engendrer des différences sur le choix de $K$, sur la partition de la population, ou encore sur la modélisation des trajectoires latentes caractérisant les différentes sous-populations obtenues. Avec l'appui d'une étude de simulation, nous remarquons que le choix du nombre de classes $K$ diffère suivant le mélange considéré, et cela particulièrement pour les dimensions ne comprenant qu'un seul item. En effet, l'approche par mélange de LMM a tendance à surestimer $K$. Cependant, pour un $K$ fixé, la classification des individus ainsi que l'information portée par les trajectoires de QdV de chaque classes latentes sont très similaires.

La Figure 1 montre les différentes trajectoires associées aux deux classes latentes qui ressortent pour les deux dimensions de QdV suivantes : statut global de santé (GHS) et la perte d'appétit (AP). Concernant la dimension GHS qui comprend deux items à sept modalités de réponses, un mélange de LMM pour modéliser $S^{(m)}$ a été réalisé. En revanche, la dimension AP regroupe un seul item à quatre modalités de réponse (1 "Rien", 2 "Un peu", 3 "Assez", 4 "Beaucoup"). Le mélange de GLMM a été ici utilisé pour modéliser la variable $S^{(s)}$. Les seuils associés aux différentes catégories de réponse sont visibles sur la Figure 1b.

Enfin, un modèle de régression logistique multinomiale a été utilisé pour expliquer

l'appartenance aux différentes classes latentes et ainsi proposer un profil type pouvant expliquer ou prédire l'évolution de la QdV. Concernant les dimensions de QdV « insomnie » et « perte d'appétit », l'état anxieux de la patiente à la chirurgie (baseline) est prédictif de l'appartenance à la classe latente ayant la trajectoire de QdV la moins favorable (la classe 2 pour la Figure 1b). Les patientes plus jeunes ont une fonction émotionnelle plus faible à la baseline (p=.001) que les patientes plus âgées. Les patientes avec un niveau d'étude supérieur ou égal au Baccalauréat sont plus exposées à une dégradation de leur capacité fonctionnelle (p=.003) et de leur niveau de QdV global (p=<.001).

# 4 Conclusion

Le choix du modèle mixte utilisé dans le mélange est important et dépend également du score considéré. Un mélange de GLMM pour données ordinales est le plus adapté pour la classification des données longitudinale de QdV, en particulier pour déterminer le nombre de sous-populations homogènes. Nous discuterons également de la pertinence du critère ICL versus le critère BIC dans notre cas d'étude.

# Bibliographie

[1] Fayers, P.M., Aaronson, N.K., Bjordal, K., Groenvold, M., Curran, D. et Bottomley, A. (2001), EORTC QLQ-C30 Scoring Manual (3rd edition).
[2] Gorter, R., Fox, J.P. et Twisk, JW. Why item response theory should be used for longitudinal questionnaire data analysis in medical research, BMC Medical Research Methodology, 2015 ; 15(1) : 55.
[3] Proust-Lima, C., Philipps, V., et Liquet, B. (2015). Estimation of extended mixed models using latent classes and latent processes : the R package lcmm, arXiv :1503.00890.
[4] Celeux, G., Martin, O., et Lavergne, C. (2005), Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments, Statistical Modelling, 5(3) :243–267.
[5] Biernacki, C., Celeux, G., et Govaert, G. (2000). Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood, IEEE Trans, Pattern Anal. Mach. Intell., 22(7) :719–725.