Modélisation dynamique de la causalité entre processus latents : application aux sphères anatomique, cognitive et fonctionnelle dans la maladie d'Alzheimer.

Bachirou Taddé [*,1] & Cécile Proust-Lima [*,2]

[*] *INSERM U1219, 146 rue Léo Saignat, 33076 Bordeaux Cedex, France*
[1] *Bachirou.Tadde@isped.u-bordeaux2.fr* [2] *Cecile.Proust@isped.u-bordeaux2.fr*

Résumé. La maladie d'Alzheimer est une maladie multidimensionnelle qui atteint progressivement plusieurs composantes : la sphère anatomique avec des atrophies cérébrales, la sphère cognitive avec un déclin des différentes fonctions cognitives (mémoire, langage, mémoire visuo-spatiale, etc.) et la sphère fonctionnelle avec des atteintes progressives des activités de la vie quotidienne. Récemment, des schémas hypothétiques ont été proposés qui appréhendent la maladie dans son ensemble en mettant en exergue les aspects dynamique et multidimensionnel de la maladie. Mais à cause de leur complexité, ces schémas n'ont pas encore été traduits sous forme de modèles statistiques considérant simultanément les aspects dynamique et multidimensionnel des atteintes. Nous proposons donc une nouvelle approche statistique pour décrire de façon dynamique les différents processus impliqués dans la maladie et explorer leurs relations causales. Ce modèle est inspiré à la fois des réseaux bayésiens dynamiques dépendant du temps et des modèles de régression pour données longitudinales. Il consiste à décrire le réseau de composantes au temps t en fonction du réseau au temps $t-1$ par le biais d'une matrice de transition. La matrice de transition est modélisée de façon souple en fonction du temps par une base de B-splines. Ce modèle est estimé par maximum de vraisemblance. Appliqué aux données de la cohorte populationnelle sur le vieillissement "3Cités", il permet de décrire l'évolution des processus au cours du temps et d'explorer leurs relations de causalité.

Mots-clés. modèle causal, processus latents, Alzheimer.

Abstract. Alzheimer's disease is a multidimensional disease that gradually affects several components : the anatomical sphere with brain atrophies, the cognitive sphere with a decline in various cognitive functions (memory, language, visuo-spatial memory, etc.) and the functional sphere with progressive damages to the activities of daily living. Recently hypothetical schemes have been proposed to understand the disease taken as a whole, highlighting the dynamic and multidimensional aspects of the disease. But because of their complexity, these schemes have not yet been translated into statistical models that simultaneously combine the dynamic and multidimensional aspects. We thus propose a new statistical approach to dynamically describe the various processes involved in the disease and explore their causal relationships. This model is inspired both by time-varying

dynamic bayesian networks and by regression models for longitudinal data. It allows to describe the network components at time t based on the network at time $t-1$ through a transition matrix. The transition matrix is modeled in a flexible way as a function of time using a basis of B-splines. This model is estimated by maximum likelihood. Applied to data from the population-based aging cohort "3Cités" it allows to describe the evolution of the processes over time and explore their causal relationships.

Keywords. causal model, latent processes, Alzheimer's disease.

1 Contexte et Objectif

La maladie d'Alzheimer est une maladie neurodégénérative qui touche aujourd'hui plus de 800.000 personnes en France et dont l'incidence a été estimée en 2006 à 225.000 nouveaux cas par an (Plan «Alzheimer et maladies apparentées» 2008-2012). C'est une maladie multidimensionnelle qui atteint plusieurs composantes : la sphère anatomique avec des atrophies cérébrales, la sphère cognitive avec un déclin des différentes fonctions cognitives (mémoire, langage, mémoire visuo-spatiale, etc.) et la sphère fonctionnelle avec des atteintes progressives des activités de la vie quotidienne. Récemment, des schémas hypothétiques ont été proposés qui appréhendent la maladie dans son ensemble (Jack et al., 2013 et McLaughlin, et al., 2010). Ces schémas posent des hypothèses sur les aspects dynamique, multidimensionnel et les relations causales entre les différentes dimensions de la maladie. Mais à cause de leur complexité, ces schémas n'ont pas encore été traduits en modèles statistiques considérant simultanément les aspects dynamique et multidimensionnel des atteintes.

L'objectif de notre étude est de proposer un nouveau modèle statistique permettant de comprendre l'évolution des liens de causalité entre les sphères anatomique, cognitive et fonctionnelle dans le vieillissement en population âgée. Jusqu'à présent soit l'aspect dynamique (Proust-Lima, et al., 2013) soit la relation causale entre dimensions (Mungas, et al., 2005) avaient été explorés par des modèles statistiques. Mais aucun modèle statistique n'avait vraiment allié les deux aspects.

D'un point de vue statistique, Dagum et al. (1992) ont proposé une approche par réseaux bayésiens dynamiques qui permet de modéliser une structure de causalité temporelle des processus. Mais, la structure de causalité a l'inconvénient de ne pas dépendre du temps. Song et al. (2009) ont étendu cette approche à la prise en compte d'une structure de causalité qui peut varier au cours du temps au prix d'une grande complexité de calcul.

Inspirés par ces travaux, notre approche consiste à définir un réseau bayésien dynamique dépendant du temps au niveau des processus sous-jacents de chaque dimension dans lequel la matrice de transition est modélisée de façon souple en fonction du temps par une base de splines, et à allier un modèle de régression pour décrire la trajectoire moyenne de chaque processus. Ce modèle peut être estimé par maximum de vraisemblance. Appliqué aux données de la cohorte populationnelle sur le vieillissement"3Cités", il permettra de

décrire l'évolution des processus anatomique, cognitif et fonctionnel au cours du temps et d'explorer leurs relations de causalité.

2 Méthodologie

2.1 Modèle statistique en présence de données complètes

Nous nous intéressons à un réseau de K processus latents $Y = (Y_1, Y_2, ..., Y_K)$ définis en temps discret. Notons Y(t) le vecteur associé au temps t .

Supposons un échantillon de n individus i $(i = 1, 2, \ldots, n)$ pour lesquels des observations de K marqueurs, chacun associé à un des K processus, sont collectés aux temps t, $t \in \tau$ où, τ représente l'ensemble de tous les temps de mesures : $\tau = \{0, 1, \ldots, T\}$. Le vecteur $X_i(t) = [X_{i1}(t), X_{i2}(t), \ldots, X_{iK}(t)]^{'}$ représente les observations des K marqueurs à l'instant t chez l'individu i. $X_i = [X_i(0), X_i(1), X_i(t), \ldots, X_i(T)]^{'}$ représente l'ensemble des observations successivement aux temps $0, 1, \ldots, T$ chez l'individu i. Enfin, $X = [X_1, X_2, \ldots, X_n]^{'}$ formalise l'échantillon des n individus supposés indépendants.
La relation temporelle entre les processus et le modèle d'observation peuvent être définis par :

$$\begin{cases} Y_i(t) = & \mu_i(t) + A(t) \times (Y_i(t-1) - \mu_i(t-1)) + \delta_i(t), \quad \forall t \in \tau \\ X_i(t) = & Y_i(t) + \epsilon_i(t) \end{cases}$$

où $\mu_i(t)$ est le vecteur d'espérances des processus latents au temps t. Le processus $(\delta_i(t))$ est un processus Gaussien d'espérance nulle. Nous considérons dans la suite des erreurs stochastiques multivariées indépendantes : $\delta_i(t) \sim \mathcal{N}(0, B)$, avec B non structurée ou des erreurs stochastiques de variance dépendant du temps :$\delta_i(t) \sim \mathcal{N}(0, B_t)$ avec par exemple $B_t = B \times \log(2+t)$. D'autres structures de corrélation pourraient cependant être considérées dans d'autres applications. $\epsilon_i(t)$ est le vecteur d'erreurs de mesure : $\epsilon_i(t) \sim \mathcal{N}(0, \Sigma)$ avec $\Sigma = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2)$. Notons $\sigma = [\sigma_1, \sigma_2, \ldots, \sigma_K]^{'}$.

L'état initial $(Y_i(0), X_i(0))$ est donné par : $\begin{cases} Y_i(0) = & \mu_i(0) + \delta_i(0) \\ X_i(0) = & Y_i(0) + \epsilon_i(0) \end{cases}$

L'espérance $\mu_i(t)$ peut être modélisée à l'aide d'une régression linéaire : $\mu_i(t) = x_i^{'}(t)\beta$ où $x_i(t)$ est un vecteur de variables explicatives et fonctions du temps.

Une des complexités de notre approche est la modélisation des matrices de transition $A(t)$ définissant le lien entre chaque composante $k(k = 1, \ldots, K)$ au temps $t-1$ et chaque composante $j(j = 1, \ldots, K)$ au temps t :

$$A(t) = \begin{pmatrix} a_{11}(t) & \dots & a_{1k}(t) & \dots & a_{1K}(t) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{j1}(t) & \dots & a_{jk}(t) & \dots & a_{jK}(t) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{K1}(t) & \dots & a_{Kk}(t) & \dots & a_{KK}(t) \end{pmatrix}$$

Nous proposons de modéliser les éléments $a_{jk}(t)$ de la matrice $A(t)$ en fonction du temps par une base de B-splines cubiques $(S_l)_{l=0,L}$: $a_{jk}(t) = \sum_{l=0}^{L} \alpha_{jkl} S_l(t), \quad \forall t \geq 1$, où α_{jkl} sont des coefficients de régression. Notons $\alpha = [\alpha_{jk0}, \alpha_{jk1}, \dots, \alpha_{jkL}]'$ pour $1 \leq j, k \leq K$.

Par récurrence, le modèle peut se réécrire sous la forme :

$$\begin{cases} Y_i(t) = & \mu_i(t) + \delta_i(t) + \left(\sum_{l=0}^{t-1} \{ \prod_{j=l+1}^{t} A(j) \} \delta_i(l) \right) \mathbb{1}_{t \geq 1}, \quad \forall t \in \tau \\ X_i(t) = & Y_i(t) + \epsilon_i(t) \end{cases}$$

Le vecteur $X_i(t)$ est donc multivarié Gaussien de moyenne $\mu_i(t)$ et de variance $V_i(t)$ définies comme suit pour $t \in \tau$:

$$\mu_i(t) = x_i'(t)\beta \quad ; \quad V_i(t) = B_t + \left(\sum_{l=0}^{t-1} \{ \prod_{j=l+1}^{t} A(j) \} B_l \{ \prod_{m=l+1}^{t} A(m) \}' \right) \mathbb{1}_{t \geq 1} + \Sigma \qquad (2)$$

Le vecteur $X_i = [X_i(0), X_i(1), \dots, X_i(T)]'$ est aussi un vecteur Gaussien $X_i \sim \mathcal{N}(\mu_i \ ; V_i)$ tel que : $\mu_i = [\mu_i(0), \mu_i(1), \dots, \mu_i(T)]'$ et

$$\begin{cases} V_i(t,t) = & \left(B_t + \left(\sum_{l=0}^{t-1} \{ \prod_{j=l+1}^{t} A(j) \} B_l \{ \prod_{m=l+1}^{t} A(m) \}' \right) \mathbb{1}_{t \geq 1} + \Sigma \right)_{t \in \tau} \\ V_i(t,u) = & \left(B_t \{ \prod_{m=t+1}^{u} A(m) \}' + \left(\sum_{l=0}^{t-1} \{ \prod_{j=l+1}^{t} A(j) \} B_l \{ \prod_{m=l+1}^{u} A(m) \}' \right) \mathbb{1}_{t \geq 1} \right)_{t<u;(t,u) \in \tau^2} \\ V_i(u,t) = & \left(V_i(t,u)' \right)_{t<u;(t,u) \in \tau^2} \end{cases} \qquad (3)$$

2.2 Prise en compte de données manquantes aléatoires

Dans la section 2.1, nous avons considéré que les $X_i(t)$ étaient observés à tous les temps $t \in \tau$. En pratique, chaque sujet est observé dans un intervalle de temps $\tau_i \subset \tau$ si bien que les vecteurs $X_i(t) \sim \mathcal{N}(\mu_i(t) \ ; V_i(t))$ sont définis pour $t \in \tau_i$ avec $\mu_i(t)$ et $V_i(t)$ donnés en équations (2).

De plus, toutes les composantes constituant le vecteur $X_i(t)$ ne sont pas systématiquement observées. On suppose que seules $k_i(t)$ composantes sont observées formant le

vecteur d'observation ${X_i}^{obs}(t)$. Ces observations partielles peuvent être prises en compte par une matrice d'observation $H_i(t)$ de dimension $k_i(t) \times K$ telle que $X_i^{obs}(t) = H_i(t)X_i(t)$. Le vecteur ${X_i}^{obs}(t)$ est donc multivarié Gaussien de moyenne $\mu_i^{obs}(t)$ et de variance $V_i^{obs}(t)$ définies comme suit pour $t \in \tau_i$: $\mu_i^{obs}(t) = H_i(t)\mu_i(t)$ et $V_i^{obs}(t) = H_i(t)V_i(t)H_i(t)^{'}$.

Le vecteur X_i^{obs} des vecteurs $X_i^{obs}(t)$, pour $t \in \tau_i$ est aussi un vecteur Gaussien de moyenne μ_i^{obs} et de variance V_i^{obs} telles que : ${\mu_i}^{obs} = ({\mu_i}^{obs}(t))_{t\in\tau_i}$ et $V_i(t,u)^{obs} = \left(H_i(t)V_i(t,u)H_i(u)^{'}\right)_{t,u\in\tau_i}$ où $V_i(t,u)$ est définie en équations (3).

3 Estimation par maximum de vraisemblance

Le vecteur de paramètres $\theta = (\beta; vec(B); \sigma; \alpha)$ peut être estimé par maximum de vraisemblance. La vraisemblance $L(X^{obs}, \theta) = \prod_{i=1}^{n} L_i(X_i^{obs}, \theta)$ où la contribution de l'individu i $L_i(X_i^{obs}, \theta)$ est la densité multivariée normale de moyenne μ_i^{obs} et de variance V_i^{obs}.

Dans la pratique, c'est la log-vraisemblance $\mathcal{L}(X^{obs}, \theta) = \log L(X^{obs}, \theta) = \sum_{1}^{n} \log L_i(X_i, \theta)$ qui est maximisée numériquement par un algorithme itératif de type quasi-newton. L'estimation des paramètres par maximum de vraisemblance est implémentée sous R.

4 Données et Résultats attendus

4.1 Données

Cette approche est appliquée aux données de la cohorte épidémiologique en population âgée "3C"(Etude "3Cités",2003). L'objectif de cette cohorte, initiée en 1999, est de comprendre les liens entre les facteurs de risque vasculaires et neurodégératifs et la survenue de plusieurs événements morbides : maladie coronaire, accidents vasculaires cérébraux et démences dont la maladie d'Alzheimer. Les sujets inclus dans la cohorte 3C ont été tirés au sort dans la population générale de 3 villes de France (Dijon, Bordeaux, Montpellier). Ils devaient avoir 65 ans ou plus et vivre à leur domicile à l'inclusion. Des données très complètes sur leur santé, leur comportement, leurs expositions ainsi que des évaluations cognitives et fonctionnelles ont été collectées à l'inclusion puis à 2 ans, à 4 ans, à 7 ans, à 10 ans et à 12 ans. A 1 an, 4 ans et 10 ans, un sous-échantillon de sujets de la cohorte ont aussi passé un examen d'IRM (Imagerie par Résonance Magnétique)pour évaluer la structure de leur cerveau.Dans cette application, nous nous concentrons sur les trois dimensions :

1. dimension anatomique avec un score résumé des volumes relatifs des hippocampes gauche et droite ;
2. dimension cognitive avec un score résumé de la batterie de tests psychométriques ;

3. dimension fonctionnelle avec un score résumé des items d'activités de la vie quotidienne (ADL) et d'activités instrumentales de la vie quotidienne (IADL).

L'analyse est réalisée en fonction de l'âge avec une discrétisation tous les un ou deux ans.

4.2 Résultats attendus

Cette application permettra de mieux comprendre le lien entre les trois dimensions principales impliquées dans la maladie d'Alzheimer. Notamment les intensités de transition apporteront un nouvel éclairage sur les hypothèses selon lesquelles les sphères anatomique, cognitive et fonctionnelle sont atteintes de façon progressive ou si des relations temporelles plus complexes interviennent.

5 Conclusion et perspectives

Notre approche apporte un nouveau regard sur la maladie d'Alzheimer, en mettant en exergue à la fois l'aspect dynamique et les relations causales entre les processus impliqués. Deux extensions consisteront à inclure des variables explicatives dans la matrice de transition et à modéliser simultanément les risques de décès et de démence.

Bibliographie

[1] Dagum, P., Galper, A., Horvitz, E. (1992). Dynamic net work mo dels for forecasting. In Proceedings of the Eighth A nnual Conference on Uncertainty in Artificial Intel ligence.
[2] Jack, C. R., Knopman, D. S., Jagust, W. J., Petersen, R. C., Weiner, M. W.,... Trojanowski, J. Q. (2013).Tracking pathophysiological processes in Alzheimer's disease : an updated hypothetical model of dynamic biomarkers. Lancet Neurology, 12(2), 207-216.
[3] McLaughlin, T., Buxton, M., Mittendorf, T., Redekop, W., Mucha, L., Darba, J., ... Leibman, C. (2010). Assessment of potential measures in models of progression in Alzheimer disease. Neurology, 75(14), 1256- 1262.
[4] Mungas, D. M., Harvey, D., Reed, B. R., Jagust, W. J., DeCardi, C., Beckett, L., Mack, W. J., Kramer, J. H., Weiner, M. W., Schuff, N. and Chui, H. C. (2005). Longitudinal volumetric MRI change and rate of cognitive decline. Neurology, 65(4) : 565–571.
[5] Proust-Lima, C., Amieva, H. and Jacqmin-Gadda, H. (2013). Analysis of multivariate mixed longitudinal data : A flexible latent process approch. British Journal of Mathematical and Statistical Psychology, 66, 470-487.
[6] Song, L., Kolar, M., Xing, E. P. (2009). Time-Varying Dynamic Bayesian Networks. http ://papers.nips.cc/paper/3716-time-varying-dynamic-bayesian-networks.pdf