Identification de systèmes biologiques dynamiques à effets aléatoires

Levy BATISTA[1,3], Thierry BASTOGNE[1,2,3], El-Hadi DJERMOUNE[1]

[1]CRAN CNRS UMR 7039
BP 70239, F-54506 Vandoeuvre-lès-Nancy Cedex, France

[2]INRIA BIGS
BP 70239, F-54506 Vandoeuvre-lès-Nancy Cedex, France

[3]CYBERnano
Telecom Nancy, 193 av. Paul Muller, 54602 Villers-lès-Nancy

levy.batista@univ-lorraine.fr
thierry.bastogne@univ-lorraine.fr
el-hadi.djermoune@univ-lorraine.fr

Résumé – L'identification de systèmes apparaît de plus en plus dans la modélisation de systèmes biologiques. Dans ce contexte d'application, chaque essai est toujours répété pour estimer la variabilité des réponses. Inférer les résultats à la population nécessite de prendre en compte la variabilité inter-individu dans la procédure de modélisation. Une solution consiste à utiliser des effets aléatoires mais jusqu'à maintenant il n'y a pas d'approche similaire dans le monde de l'identification de systèmes. Dans cet article nous proposons une méthode basée sur une structure ARX (*Auto Regressive model with eXternal inputs*) en utilisant l'algorithme EM (Espérance-Maximisation) pour estimer les paramètres du modèle . Des simulations montrent l'intérêt de cette approche en comparaison à une approche plus classique consistant à identifier les paramètres de chacun des individus indépendamment.

Abstract – System identification is a data-driven modeling approach more and more used in biology and biomedicine. In this application context, each assay is always repeated to estimate the response variability. The inference of the modeling conclusions to the whole population requires to account for the inter-individual variability within the modeling procedure. One solution consists in using random effects models but up to now no similar approach exists in the field of dynamical system identification. In this article, we propose a new solution based on an ARX (Auto Regressive model with eXternal inputs) structure using the EM (Expectation-Maximisation) algorithm for the estimation of the model parameters. Simulations show the relevance of this solution compared with a classical procedure of identification repeated for each subject.

1 Introduction

En pharmacologie, les études de pharmacocinétique et pharmocodynamie (PK, PD) produisent des signaux de mesure que l'on cherche à représenter par des modèles mathématiques pour caractériser les propriétés des médicaments. Le plus souvent, les méthodes d'analyse sont des méthodes statistiques. Ces dernières modélisent les réponses, qui sont souvent non-linéaires en les paramètres même pour des modèles à compartiments simples, [1]. L'intérêt de prendre en compte des répétitions individuelles et inter-individuelles dans les modèles a conduit à une expansion des travaux sur les modèles à effets aléatoires, [2, 3, 4, 5]. Ces modèles décomposent la variabilité des réponses en deux parties : une partie déterministe décrite par des effets fixes et une partie aléatoire permettant d'expliquer la variabilité inter-individus.

Par ailleurs, plusieurs méthodes d'identification des systèmes dynamiques ont été appliquées ces dernières années à des processus biologiques, [6, 7, 8, 9]. Plutôt que de modéliser seulement la réponse, c'est le système que l'on cherche à représenter en prenant en compte des signaux d'entrée tels que l'ajout du médicament, la modification du milieu, etc. Dans le cadre des modèles à compartiments, la structure des modèles utilisés est souvent linéaire par rapport aux paramètres contrairement à leurs réponses, [10].

Toutefois, à notre connaissance, il n'existe pas de méthode d'identification de systèmes dynamiques se reposant sur des structure de modèle à effets aléatoires lorsque l'on fait intervenir des mesures répétées. Cet article propose donc une méthode d'identification des paramètres d'une structure ARX intégrant des effets aléatoires.

Le reste de ce papier est organisé comme suit. Dans un premier temps, on présente le modèle ARX en y incorporant des effets aléatoires. Ensuite, on donnera une méthode d'estimation des paramètres par maximisation de la fonction de vraisemblance en utilisant l'algorithme EM. L'estimation des paramètres sera évaluée *in silico* sur des données de simulations par Monte-Carlo en faisant varier le nombre d'individus ainsi que le rapport signal-à-bruit. Les résultats sont comparés à l'approche plus classique consistant à identifier indépendamment chacun des individus. Enfin, nous conclurons par une discussion sur les avantages et les inconvénients de la méthode.

2 Modèle ARX à effets aléatoires

Nous avons choisi la structure de modèle ARX comme point de départ de notre étude, le modèle ARX d'ordres n_a et n_b comme décrit dans [11] s'écrit pour l'individu $i \in [1, ..., I]$:

$$y_i(t_k) = -a_{i,1}y_i(t_{k-1}) - ... - a_{i,n_a}y_i(t_{k-n_a}) + b_{i,1}u_i(t_{k-n_d}) + ... + b_{i,n_b}u_i(t_{k-n_b-n_d}) + e_i(t_k), \quad (1)$$

avec $y_i(t_k)$ le signal de sortie à l'instant $t_k = kT_e$, $k \in [0, ..., K-1]$ et T_e est la période d'échantillonnage. $e_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_e^2)$ et n_d le nombre d'échantillons de retard avant que l'action de l'entrée ait un effet sur la sortie. Les coefficients $a_{i,j}$ et $b_{i,l}$ sont les paramètres inconnus de ce modèle. Chaque individu étant différent on peut le situer au sein de la population en décomposant chaque paramètre comme suit :

$$\begin{cases} a_{i,j} &= a_{pop,j} + \tilde{a}_{i,j} \\ b_{i,l} &= b_{pop,l} + \tilde{b}_{i,l} \end{cases} \quad (2)$$

a_{pop} et b_{pop} étant les paramètres moyens de la population. Représenté sous forme vectorielle, on a :

$$\overline{y}_i = \phi_i \left(\theta_{pop} + \tilde{\theta}_i\right) + \overline{e}_i, \quad (3)$$

avec $\overline{y}_i$ et $\overline{e}_i \in \mathbb{R}^{(K-n_a)}$ sont respectivement les vecteurs de sortie et d'erreur de mesure : $\overline{y}_i = [y_i(t_{n_a}), ..., y_i(t_{K-1})]^T$, de même pour $\overline{e}_i$. $\phi_i = [-\underline{y}_{i,1}, ..., -\underline{y}_{i,n_a}, u_{i,n_d}, ..., u_{i,n_b+n_d}] \in \mathbb{R}^{(K-n_a)\times(n_a+n_b)}$ est la matrice de régression formée par les mesures de sortie passées et des entrées passées et présentes sachant que : $\underline{y}_{i,m} = [y_i(t_{n_a-m}), ..., y_i(t_{K-1-m})]^T$ et $u_{i,m} = [u_i(t_{n_a-m}), ..., u_i(t_{K-1-m})]^T$, sous l'hypothèse que : $n_a + 1 > n_b + n_d$. Enfin, $\theta_{pop} = [a_{pop,1}, ..., a_{pop,n_a}, b_{pop,1}, ..., b_{pop,n_b}]^T \in \mathbb{R}^{(n_a+n_b)}$ est le vecteur de paramètres des effets fixes, et $\tilde{\theta}_i = [\tilde{a}_{i,1}, ..., \tilde{a}_{i,n_a}, \tilde{b}_{i,1}, ..., \tilde{b}_{i,n_b}]^T \in \mathbb{R}^{(n_a+n_b)}$ le vecteur de paramètres des effets aléatoires.

L'équation (3) souligne un problème d'identifiabilité des paramètres θ_{pop} et $\{\tilde{\theta}_i\}_{i=1}^I$. Ce problème est évité en imposant une distribution de probabilité aux $\tilde{\theta}_i$:

$$\tilde{\theta}_i \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma) \quad (4)$$

où Σ est une matrice non nécessairement diagonale contenant les variances de chaque paramètre composant $\tilde{\theta}_i$ sur sa diagonale : $\text{diag}(\Sigma) = [\sigma_{a,1}^2, ..., \sigma_{a,n_a}^2, \sigma_{b,1}^2, ..., \sigma_{b,n_b}^2]^T$. Les paramètres à identifier sont alors :

$$\Theta = \left(\theta_{pop}, \Sigma, \sigma_e^2\right) \quad (5)$$

3 Estimation des paramètres par maximum de vraisemblance

3.1 Fonction de vraisemblance

Le modèle (3) peut être considéré comme un problème à données non-observées, puisque $\tilde{\theta}_i$ n'est pas mesuré ; on l'appelle aussi le « vecteur de données cachées ». On peut définir le vecteur des données complètes d'un individu par $y_i^{(c)} = \left[\overline{y}_i^T, \tilde{\theta}_i^T\right]^T$ dont la distribution est donnée par :

$$p(y_i^{(c)}) \triangleq p(\overline{y}_i, \tilde{\theta}_i; \Theta) = p(\overline{y}_i|\tilde{\theta}_i; \Theta)p(\tilde{\theta}_i; \Theta). \quad (6)$$

De plus, tous les individus étant indépendants les uns des autres on peut écrire :

$$p(\overline{y}, \tilde{\theta}; \Theta) = \prod_i p(\overline{y}_i|\tilde{\theta}_i; \Theta)p(\tilde{\theta}_i; \Theta) \quad (7)$$

$\overline{y}$ étant l'ensemble des mesures de sorties effectuées sur les I individus. De l'équation (7), on tire l'expression de la log-vraisemblance :

$$l(\Theta|\overline{y}, \tilde{\theta}) = -c - \frac{I(K-n_a)}{2}\ln(\sigma_e^2) - \frac{I}{2}\ln|\Sigma| - \sum_i \frac{1}{2\sigma_e^2}\overline{e}_i^T\overline{e}_i + \frac{1}{2}\tilde{\theta}_i^T\Sigma^{-1}\tilde{\theta}_i \quad (8)$$

où c est une constante indépendante de Θ. On peut montrer que la maximisation de l'équation (8) conduit aux estimées suivantes :

$$\hat{\theta}_{pop} = \left(\sum_i \phi_i^T\phi_i\right)^{-1} \sum_i \phi_i^T(\overline{y}_i - \phi_i\tilde{\theta}_i) \quad (9)$$

$$\hat{\Sigma} = \frac{1}{I}\sum_i \tilde{\theta}_i\tilde{\theta}_i^T \quad (10)$$

$$\hat{\sigma}_e^2 = \frac{1}{(K-n_a)I}\sum_i \overline{e}_i^T\overline{e}_i. \quad (11)$$

Comme $\tilde{\theta}_i$ n'est pas connu nous allons utiliser dans la section suivante l'algorithme EM pour l'estimer.

3.2 Algorithme EM

L'algorithme itératif Espérance-Maximisation (*Expectation-Maximisation*) est connu pour sa capacité à rendre possible l'estimation de paramètres dans le cadre de données incomplètes, [12]. En effet, l'algorithme EM estime la variable cachée, dans notre cas $\tilde{\theta}_i$, par son espérance conditionnellement à $\overline{y}_i$ et $\Theta^{(s-1)}$, s étant le numéro de l'itération.

3.2.1 Étape d'espérance

Le vecteur de données complètes étant distribué normalement on a :

$$y_i^{(c)} = \begin{pmatrix}\overline{y}_i \\ \tilde{\theta}_i\end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\phi_i\theta_{pop}^{(s)} \\ 0\end{bmatrix}, \begin{bmatrix}V^{(s)} & \phi_i\Sigma^{(s)} \\ \Sigma^{(s)}\phi_i^T & \Sigma^{(s)}\end{bmatrix}\right) \quad (12)$$

avec $V^{(s)} = \phi_i\Sigma^{(s)}\phi_i^T + \mathrm{I}_{K-n_a}\sigma_e^{2(s)}$, où I_N est la matrice identité $N \times N$. On peut montrer que l'espérance conditionnelle de $\tilde{\theta}_i$ est donnée par :

$$\hat{t}_{1,i}^{(s)} \triangleq \mathbb{E}(\tilde{\theta}_i|\overline{y}_i, \Theta^{(s)}) = \Sigma^{(s)}\phi_i^T V^{(s)-1}(\overline{y}_i - \phi_i\theta_{pop}^{(s)}), \quad (13)$$

de même pour la somme des carrés $\tilde{\theta}_i^T\tilde{\theta}_i$:

$$\hat{t}_{2,i}^{(s)} \triangleq \mathbb{E}(\tilde{\theta}_i\tilde{\theta}_i^T|\overline{y}_i, \Theta^{(s)}) = \hat{t}_{1,i}^{(s)}\hat{t}_{1,i}^{(s)T} + \Sigma^{(s)} - \Sigma^{(s)}\phi_i^T V^{(s)-1}\phi_i\Sigma^{(s)}. \quad (14)$$

En suivant ce raisonnement on a aussi :

$$\hat{t}_{3,i}^{(s)} \triangleq \mathbb{E}(\overline{e}_i^T \overline{e}_i | \overline{y}_i, \Theta^{(s)}) = \sigma_e^{4(s)} \left(V^{(s)-1}(\overline{y}_i - \phi_i \theta_{pop}^{(s)}) \right)^T \left(V_{11}^{-1}(\overline{y}_i - \phi_i \theta_{pop}^{(s)}) \right) + \mathrm{Tr}(\sigma_e^{2(s)} I_{K-n_a} - \sigma_e^{4(s)} V^{(s)-1}), \quad (15)$$

où $\mathrm{Tr}(.)$ est la trace d'une matrice.

3.2.2 Étape de maximisation

Dans l'étape de maximisation on utilise l'équation (11) dans laquelle on remplace les variables cachées par leurs statistiques $\hat{t}_{1,i}^{(s)}, \hat{t}_{2,i}^{(s)}$ et $\hat{t}_{3,i}^{(s)}$ calculées dans l'étape d'espérance :

$$\hat{\theta}_{pop}^{(s+1)} = \left(\sum_i \phi_i^T \phi_i \right)^{-1} \sum_i \phi_i^T (\overline{y}_i - \phi_i \hat{t}_{1,i}^{(s)}) \quad (16)$$

$$\hat{\Sigma}^{(s+1)} = \frac{1}{I} \sum_i \hat{t}_{2,i}^{(s)} \quad (17)$$

$$\hat{\sigma}_e^{2(s+1)} = \frac{1}{(K-n_a)I} \sum_i \hat{t}_{3,i}^{(s)} \quad (18)$$

On peut initialiser $\theta_{pop}^{(0)}$ et $\sigma_e^{2(0)}$ par une identification de type ARX sur un individu, et $\Sigma^{(0)}$ en donnant des valeurs raisonnables à ses coefficients.

4 Évaluation en simulation

4.1 Modèle à compartiment simple

La Figure 1 représente un modèle à un seul compartiment. Il décrit l'ajout d'un médicament dans un organe par l'intermédiaire de l'entrée $u(t_k)$ représentant le flux de médicament administré pondéré par b, le coefficient d'absorption ; $x(t_k)$ est la concentration de médicament absorbé par l'organe à l'instant t_k et, enfin, a est la constante d'élimination de l'organe. Ainsi, en considérant ces paramètres propres à chaque individu, on obtient :

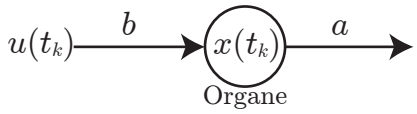

FIGURE 1 – Modèle à un compartiment

$$\begin{cases} x_i(t_{k+1}) = -a_i x(t_k) + b_i u_i(t_k) + e_i(t_{k+1}) \\ y_i(t_k) = x_i(t_k) \end{cases} \quad (19)$$

avec $e_i(t_k) \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_e^2)$ correspondant à l'erreur de mesure.

Chaque individu étant différent, on peut le situer par rapport à l'individu moyen :

$$y_i(t_k) = -(a_{pop} + \tilde{a}_i) y_i(t_{k-1}) + (b_{pop} + \tilde{b}_i) u_i(t_{k-1}) + e_i(t_k). \quad (20)$$

avec $\tilde{\theta}_i = [\tilde{a}_i, \tilde{b}_i]^T$ suivant une loi normale centrée et de matrice de variance-covariance Σ.

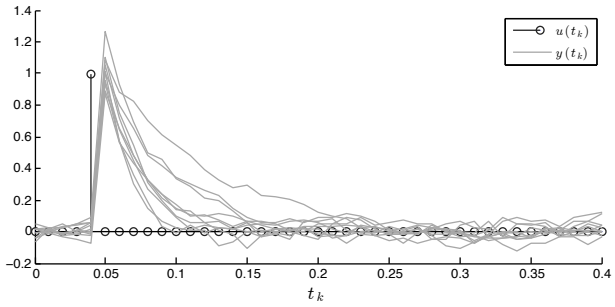

FIGURE 2 – Une simulation avec $I = 10$ et $\mathrm{RSB} = 50$.

4.2 Protocole expérimental

Dans le but de montrer l'apport de l'approche de population, nous allons la comparer à une méthode plus classique qui consisterait à calculer la moyenne des paramètres individuels et leur variance : $\hat{\theta}_{pop} = \frac{1}{I} \sum_{i=1}^{I} \theta_i$, $\hat{\Sigma}_{i,j} = \mathrm{cov}(\theta_i, \theta_j)$ et $\hat{\sigma}_e^2 = \frac{1}{I} \sum_{i=1}^{I} \sigma_{e,i}^2$.

Nous avons réalisé des simulations de Monte-Carlo en faisant varier deux paramètres qui sont le nombre d'individus et le rapport signal-à-bruit (RSB) indépendamment l'un de l'autre.

— Pour étudier l'impact du nombre d'individus on fait varier $I \in \{3, 5, 10, 25, 100\}$ en fixant le RSB à 50.
— Pour étudier l'impact du RSB on le fait varier $\mathrm{RSB} \in \{1, 10, 50, 100\}$ en fixant le nombre d'individus à 10.

Les autres paramètres ont été fixés : $a_{pop} = 0.7, b_{pop} = 1, \sigma_a^2 = 0.01, \sigma_b^2 = 0.01$. Chaque situation a été simulée 100 fois avec différentes réalisations de $e_i(t_k)$ et $\tilde{\theta}_i$. L'entrée $u(t_k)$ utilisée est un signal de Kronecker : $u(t_k) = \delta(t_k - \tau)$. Une simulation est représentée en Figure 2 avec $I = 10$ et $\mathrm{RSB} = 50$.

4.3 Résultats et discussion

Les résultats sont exposés en Figures 3 et 4. Lorsque l'on fait varier le nombre d'individus pour un RSB grand ($\mathrm{RSB} = 50$), on remarque que les effets fixes et la variance du bruit sont bien estimés, les deux méthodes sont sans biais, et avec des variances similaires. Le nombre d'individus est corrélé négativement à la variance des estimées. Pour les effets aléatoires, σ_a^2 et σ_b^2, on remarque que la méthode de moyenne *a posteriori* des estimations ARX individuelles est légèrement biaisée et que ce biais est inversement proportionnel au RSB. La méthode proposée *Random ARX* (RARX) est plus robuste au bruit, notamment elle donne de bonnes estimées même avec un rapport signal-à-bruit égal à 1. Toutefois, elle donne des estimées biaisées vers le bas pour un nombre d'individus faible, ceci est un résultat connu en statistique qui est dû à la non prise en compte de la perte de degré de liberté après estimation des effets fixes, [2]. En conclusion, la méthode RARX permet de mieux évaluer la variabilité au sein d'une population sous la contrainte d'avoir un nombre d'individus suffisamment grand. Dans le cas contraire il faudrait utiliser des méthodes dites de maximums de vraisemblance restreinte afin de prendre en compte la perte de degré de liberté. Ceci fera l'objet des développements à venir.

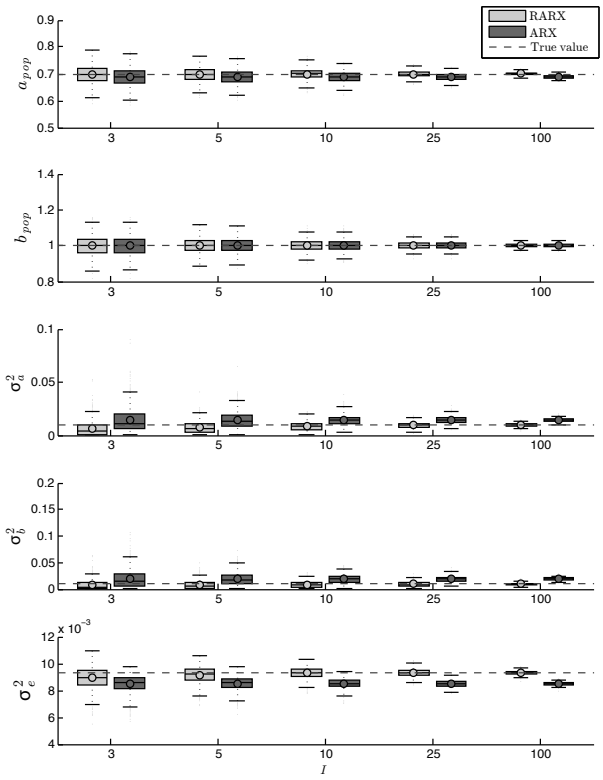

FIGURE 3 – Résultats des simulations de Monte-Carlo pour un nombre d'individus variant et un rapport signal à bruit fixé à 50. Avec notre méthode RARX (*Random effects ARX*) comparée à un ARX classique

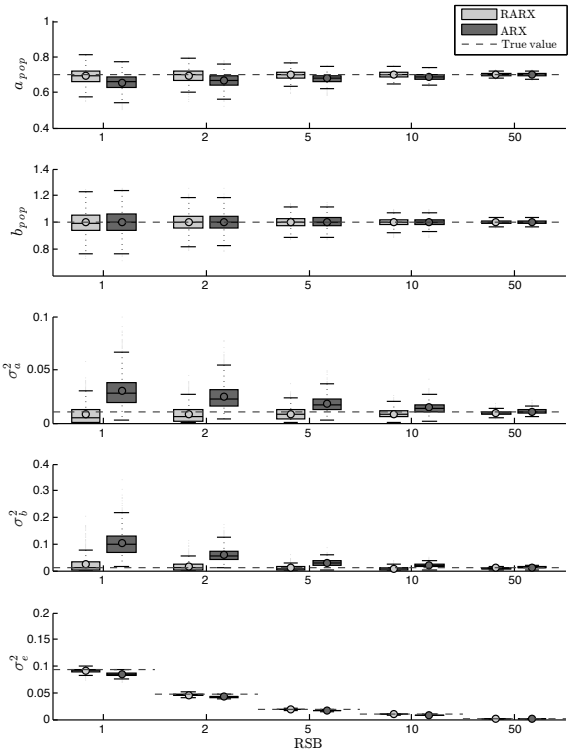

FIGURE 4 – Résultats des simulations de Monte-Carlo pour un rapport signal à bruit variant et un nombre d'individus fixé à 10. Avec notre méthode RARX (*Random effects ARX*) comparée à un ARX classique

Références

[1] M. Lavielle, *Mixed Effects Models for the Population Approach. Models, Tasks, Methods & Tools.* Chapman & Hall/CRC Biostatistics Series, 2014.

[2] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data.," *Biometrics*, vol. Biometrics, pp. pp. 963–974, Dec. 1982.

[3] T. Bastogne, A. Samson, P. Vallois, S. Wantz-Mézières, S. Pinel, D. Bechet, and M. Barberi-Heyob, "Phenomenological modeling of tumor diameter growth based on a mixed effects model," *Journal of Theoretical Biology*, vol. 262, no. 3, pp. 544 – 552, 2010.

[4] G. Verbeke and G. Molenberghs, *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics, Springer, 2009.

[5] S. Searle, G. Casella, and C. McCulloch, *Variance Components*. Wiley Series in Probability and Statistics, Wiley, 2009.

[6] D. Feng, S.-C. Huang, Z. Wang, and D. Ho, "An unbiased parametric imaging algorithm for nonuniformly sampled biomedical system parameter estimation," *Medical Imaging, IEEE Transactions on*, vol. 15, pp. 512–518, Aug 1996.

[7] N. D. Evans, R. J. Errington, M. Shelley, G. P. Feeney, M. J. Chapman, K. R. Godfrey, P. J. Smith, and M. J. Chappell, "A mathematical model for the *in vitro* kinetics of the anti-cancer agent topotecan," *Mathematical Biosciences*, vol. 189, no. 2, pp. 185 – 217, 2004.

[8] T. Bastogne, L. Tirand, D. Bechet, M. Barberi-Heyob, and A. Richard, "System identification of photosensitiser uptake kinetics in photodynamic therapy," *Biomedical Signal Processing and Control*, vol. Elsevier, pp. pp.217–225, 2007.

[9] J.-B. Tylcz, D. Bechet, T. Bastogne, H. Garnier, and M. Barberi-Heyob, "System identification of the intra-brain tumoral uptake of multifunctional nanoparticles," in *8th IFAC Symposium on Biological and Medical Systems, IFAC BMS 2012*, (Budapest, Hungary), p. CDROM, Aug. 2012.

[10] E. Walter and L. Pronzato, *Identification of Parametric Models from Experimental Data*. Springer, 1997.

[11] L. Ljung, *System Identification, Theory for the User*. PreWiley-Hall, 1987.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data *via* the EM algorithm," *Journal of the royal statistical society, Serie B*, vol. 39, no. 1, pp. 1–38, 1977.