

REVUE DE STATISTIQUE APPLIQUÉE

J. Y. MARY

A. J. VALLERON

Méthode de Monte-Carlo pour estimer la précision des estimateurs des paramètres de modèles de systèmes biologiques. Applications en cinétique cellulaire

Revue de statistique appliquée, tome 23, n° 4 (1975), p. 77-84

http://www.numdam.org/item?id=RSA_1975__23_4_77_0

© Société française de statistique, 1975, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

MÉTHODE DE MONTE CARLO POUR ESTIMER LA PRÉCISION DES ESTIMATEURS DES PARAMÈTRES DE MODÈLES DE SYSTÈMES BIOLOGIQUES

APPLICATIONS EN CINÉTIQUE CELLULAIRE (1)

J. Y. MARY (2) – A. J. VALLERON (3)

INTRODUCTION

Quand on dispose pour un phénomène biologique d'un modèle mathématique en accord avec les données expérimentales, on estime à partir de celles-ci les valeurs inconnues des paramètres du modèle. La connaissance des variances des estimateurs est nécessaire pour comparer statistiquement les valeurs trouvées dans deux situations expérimentales différentes ou pour tester si un des paramètres diffère d'une valeur théorique connue par ailleurs.

Si l'on connaît la formulation mathématique du modèle, la méthode classique du maximum de vraisemblance fournit moyennant des hypothèses peu restrictives (4) les estimations des paramètres et leur intervalle de confiance.

Cependant certaines techniques de modélisation ne cherchent pas à connaître la formulation mathématique du modèle, alors même que les hypothèses sont toutes précisées. Donnons un exemple simple : soit une population cellulaire dans laquelle chaque cellule a une durée de vie T_c de densité de probabilité $f(T_c)$. A la fin de sa vie elle se divise par mitose en deux cellules fille dont la durée de vie suit la même loi ; on suppose que la cellule ancêtre est née au temps 0. La théorie des processus de renouvellement (6) nous indique que le nombre de cellules au temps t vérifie l'équation intégrale suivante

$$N(t) = 1 - F(t) + 2 \int_0^t N'(t-x) (1 - F(x)) dx$$

où N' est la dérivée du nombre de cellules N , F la fonction de répartition de T_c .

La résolution de cette équation permet d'obtenir la fonction mathématique $N(t)$. A partir de valeurs expérimentales on estime la moyenne et la variance de la distribution de T_c ainsi que leur intervalle de confiance par la méthode du maximum de vraisemblance.

On peut également obtenir le nombre de cellules vivant à la date t correspondant à une certaine distribution de la durée de la vie T_c par un procédé de

(1) Article remis en Novembre 1974, révisé en mars 1975.

(2) Chargé de Recherches U.54 INSERM

(3) Chargé de Recherches U.21 INSERM

type Monte-Carlo : on dresse un tableau représentant les dates de naissance de chaque cellule appartenant à une lignée issue d'une cellule ancêtre vivante au temps zéro : les durées de vie de chaque cellule sont obtenues en prenant des nombres au hasard dans la distribution de probabilité $f(T_c)$, sa date de naissance est obtenue en ajoutant à celle de sa mère la durée de vie de celle-ci. Moyennant une bonne organisation des données en mémoire (9), on conçoit qu'on pourra par un simple procédé de comptage retrouver combien de cellules de chaque lignée ainsi simulée sont présentes au temps t . On peut ainsi connaître $N(t)$ avec la précision désirée, en ayant pris en compte toutes les hypothèses du modèle mais sans avoir sa formulation mathématique puisqu'alors on n'a pas eu besoin de poser ni résoudre l'équation intégrale ci-dessus. Connaissant les données expérimentales on pourra estimer les paramètres du modèle en retenant ceux qui donnent les résultats simulés les plus proches des résultats expérimentaux au sens d'un certain critère de ressemblance par une méthode d'optimisation (5).

La méthode que nous présentons ici permet l'estimation des variances des estimations des paramètres d'un tel modèle où la réponse est connue par simulation alors que la formule explicite n'est pas utilisée. La méthode sera présentée dans le cas d'un problème de cinétique cellulaire pour lequel elle a été développée. Elle sera ensuite comparée à la méthode classique du maximum de vraisemblance dans un cas théorique simple.

I.1. Notions biologiques nécessaires.

Le cycle cellulaire est l'intervalle de temps qui sépare la naissance d'une cellule de sa mitose. On distingue dans le cycle cellulaire quatre phases distinctes : la présynthèse (notée G_1), la synthèse (notée S) pendant laquelle la cellule double sa quantité d'ADN, la postsynthèse (notée G_2) et la mitose (notée M) ; les quatre phases ont des durées variables d'une cellule à l'autre dans une population donnée.

Soit $\mu G_1, \mu S, \mu G_2, \mu M$ les durées moyennes, $\sigma G_1, \sigma S, \sigma G_2, \sigma M$ les écarts-types correspondants. Pour mesurer ces paramètres on utilise de façon générale l'expérience du relevé au cours du temps du pourcentage de mitoses marquées (PMM) après contact de la population cellulaire avec un précurseur spécifique de l'ADN, usuellement la thymidine tritiée : en effet, seules les cellules en synthèse au moment de ce contact sont marquées, le restent et transmettent ce marquage à leur descendantes. On observe ensuite au cours du temps le pourcentage de cellules marquées parmi celles qui sont en mitose $PMM(t)$. Si on se restreint à l'étude de populations cellulaires en croissance stationnaire, sans mort ni migration, où toutes les cellules participent à la croissance de la population (pas de cellules quiescentes) la fonction $PMM(t)$ ne dépend que des huit paramètres $\mu G_1, \mu S, \mu G_2, \mu M, \sigma G_1, \sigma S, \sigma G_2$ et σM . En l'absence de variabilité ($\sigma G_1 = \sigma S = \sigma G_2 = \sigma M = 0$) et si la mitose dure un temps très bref la courbe $PMM(t)$ à l'allure en créneaux décrite sur la figure 1. La variabilité de la durée des phases modifie la courbe comme le montre la figure 2. Les expérimentateurs estiment souvent manuellement la durée moyenne des phases $\mu G_1, \mu S$, et μG_2 en considérant les intersections de la courbe expérimentale avec l'horizontale 50 % (cf. fig. 2).

Cette méthode ne donne que des valeurs approximatives pour les durées moyennes des phases et rien sur leur variabilité.

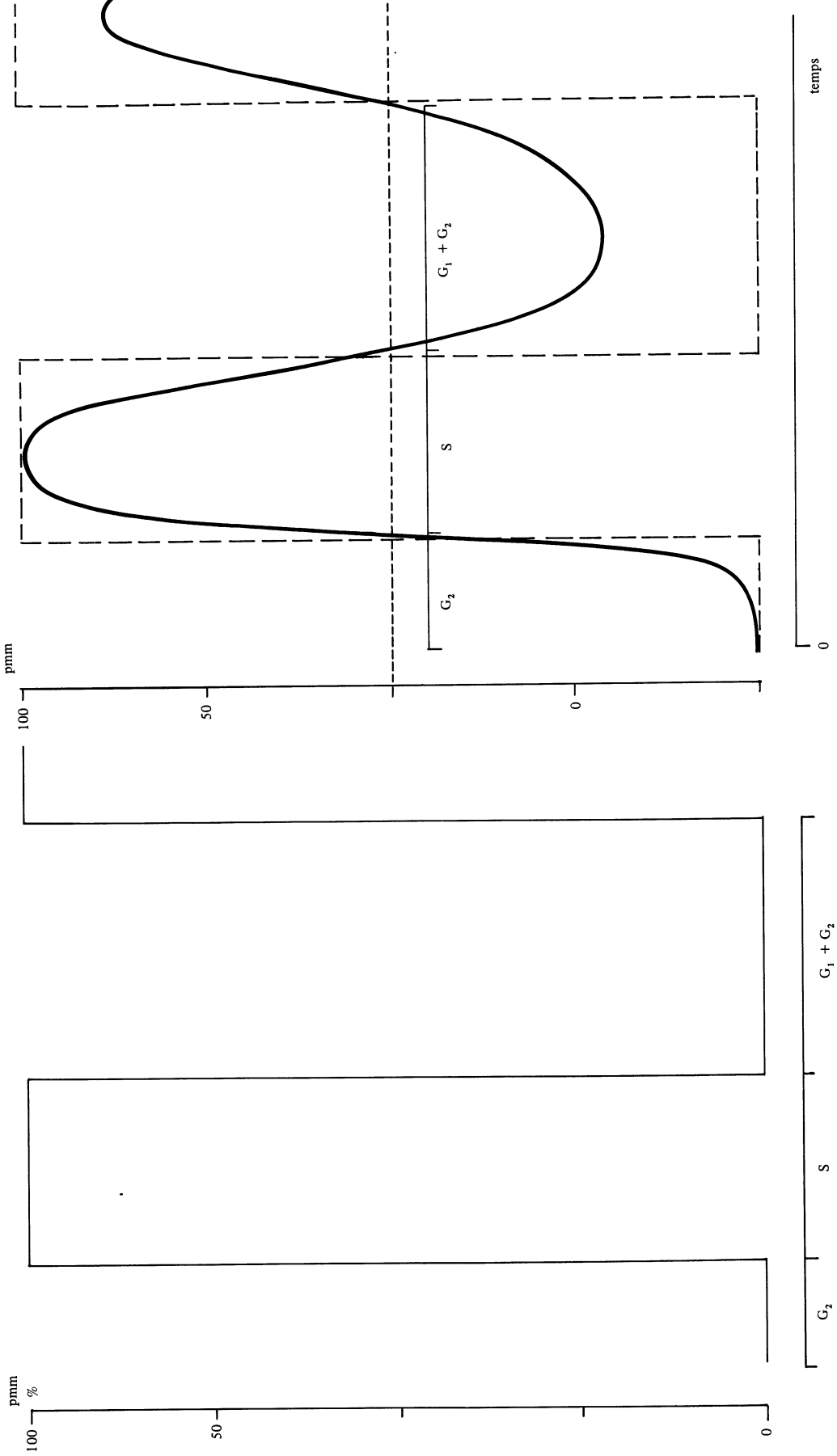


Figure 1 - Pourcentage de mitoses marquées en fonction du temps en absence de variabilité de la durée des phases.

Figure 2 - Courbe théorique de mitoses marquées en présence de variabilités des durées des phases du cycle dans la population. L'estimation graphique des durées moyennes est indiquée à l'ordonnée 50%.

I.2. Modèle utilisé

Un modèle de simulation sur ordinateur numérique a été développé (9) pour connaître point par point la courbe théorique PMM(t) correspondant à un ensemble de six paramètres μ_{G_1} , μ_S , μ_{G_2} , σ_{G_1} , σ_S , σ_{G_2} (μ_M et σ_M sont supposés connus dans les applications).

Des suites de nombres pseudoaléatoires choisis dans les distributions des durées des phases (supposées lognormales) permettent d'obtenir par simulation la fonction PMM(t) point par point sans connaître sa formulation mathématique. Soit une courbe expérimentale PMM constituée de K points. Le i ème pourcentage de mitoses marquées mesuré au temps t_i vaut Y_i , valeur moyenne de n_i observations de variance s_i^2 ; soit y_i le pourcentage théorique au temps t_i obtenu par simulation pour un ensemble donné de paramètres (μ_{G_1} , μ_S , μ_{G_2} , σ_{G_1} , σ_S , σ_{G_2}).

Les valeurs des paramètres retenus sont ceux qui correspondent à un ensemble de valeurs $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_k$ minimisant la quantité

$$D = \sum_{i=1}^K \frac{n_i}{s_i^2} (Y_i - y_i)^2$$

choisie comme critère d'écart entre les observations expérimentales (Y_i) et les réponses du modèle (y_i). Le procédé numérique de minimisation que nous avons choisi est décrit ailleurs (10).

La figure 3 donne le principe de la méthode : pour chaque temps t_i est figuré l'intervalle de confiance des points moyens expérimentaux Y_i ; dans la suite, nous supposons que les Y_i sont distribués normalement de moyenne \hat{y}_i et de variance s_i^2/n_i .

Dans chacune de ces K distributions liées, on tire au sort une première série de K points "pseudo expérimentaux" Y_{is} (carrés blancs sur la figure 3). Ceux-ci sont obtenus par la formule

$$Y_{is} = \hat{y}_i + t \frac{s_i}{\sqrt{n_i}}$$

où t est une valeur tirée au sort dans une loi de Student (3) à $n_i - 1$ degrés de liberté. L'optimisation pour ces K premiers points "pseudo expérimentaux" fournit les estimations $\mu_{G_1}^1, \mu_S^1, \mu_{G_2}^1, \sigma_{G_1}^1, \sigma_S^1, \sigma_{G_2}^1$

On tire au sort une deuxième série de points "pseudo-expérimentaux" (carrés noirs sur la figure 3); on obtient six nouvelles estimations des moyennes et écart-types des durées des phases du cycle cellulaire.

En itérant le procédé avec n séries de points "pseudo expérimentaux" on obtient n sextuplets d'estimations $\mu_{G_1}^j, \mu_S^j, \mu_{G_2}^j, \sigma_{G_1}^j, \sigma_S^j, \sigma_{G_2}^j$ des paramètres du cycle. Leur fluctuation est l'image de la fluctuation des points "pseudo expérimentaux". C'est la matrice de variance covariance calculée à partir des n sextuplets obtenus que nous utiliserons comme estimation de la matrice de variance covariance des estimateurs.

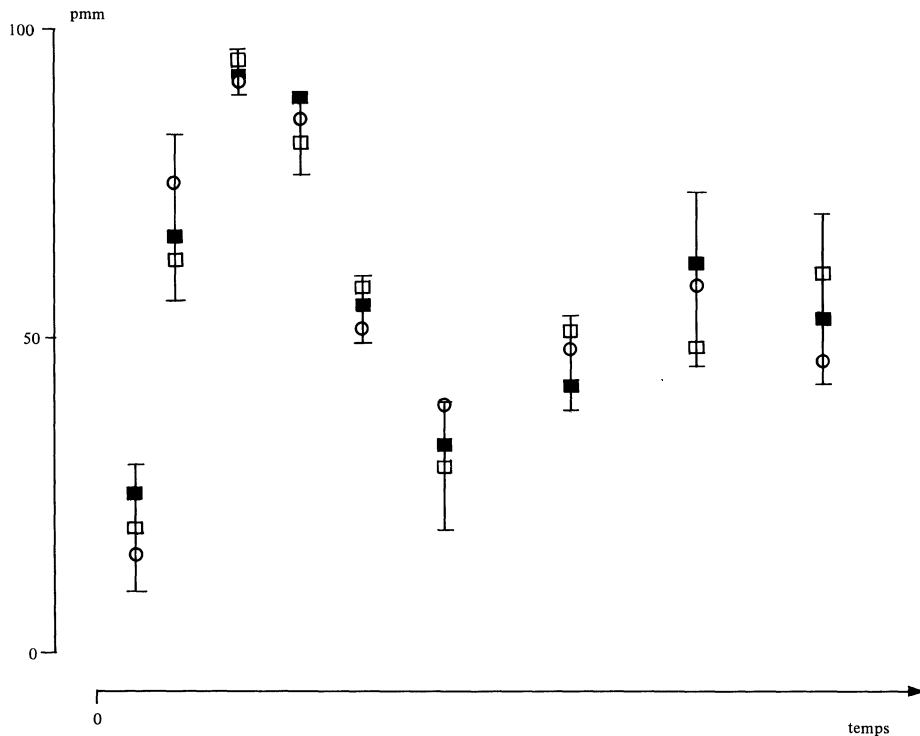


Figure 3 — ■ □ ○ indiquent trois séries de points pseudoexpérimentaux tirés au sort dans les intervalles de confiance des données expérimentales d'une courbe de mitoses marquées.

1.3. Résultats

20 courbes "pseudoexpérimentales" ont été utilisées pour analyser les données de la figure 4 suivant la méthode décrite ci-dessus. Les résultats concernant la précision des estimations sont fournis sur le tableau I. L'observation des coefficients de variabilité permet ainsi de voir quels paramètres sont connus les plus précisément. La dispersion du réseau des 20 "meilleures" réponses du modèle représentées sur la figure 4 donne une image de l'imprécision des estimations des paramètres sur la courbe théorique des mitoses marquées.

Tableau I

Précision en heures des estimations des paramètres
du cycle cellulaire dans l'exemple de la figure 4

	G1		S		G2	
Estimations	$\mu = 22.50$	$\sigma = 25.12$	$\mu = 11.03$	$\sigma = 4.83$	$\mu = 2.63$	$\sigma = 1.79$
Ecart type de ces esti- mations	2.43	5.06	0.66	1.15	0.25	0.60

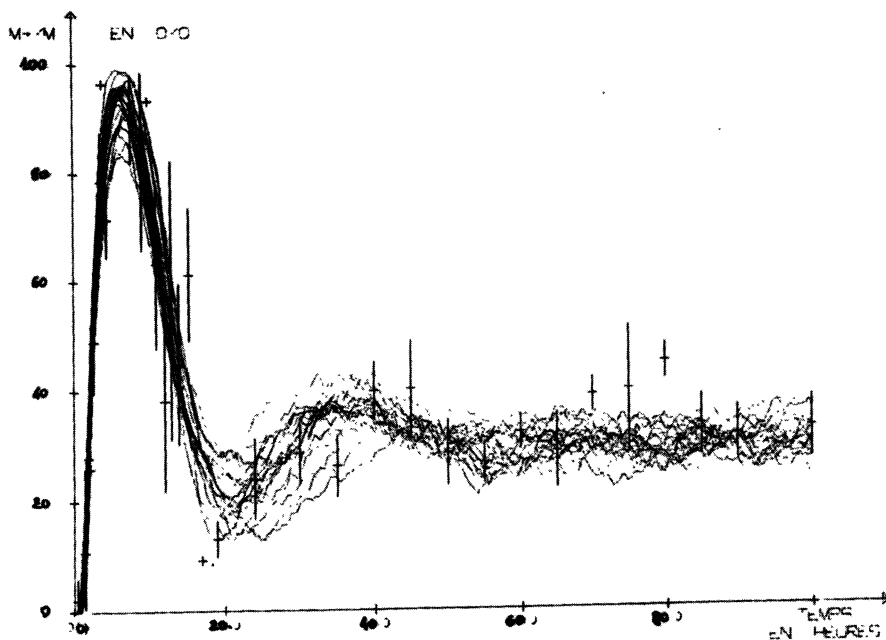


Figure 4 – Points expérimentaux munis de leur intervalle de confiance et “meilleures” réponses du modèle pour 20 séries de points pseudoexpérimentaux.

II – EVALUATION DE LA METHODE PAR RAPPORT A CELLE DU MAXIMUM DE VRAISEMBLANCE

II.1. Nous avons vérifié que lorsque la méthode classique était applicable les résultats que nous obtenions lui étaient comparables. Pour cela nous avons choisi arbitrairement le modèle de régression linéaire suivant :

$$y_{ij} = x_i + 2 + e_{ij}$$

Les x_i sont les variables contrôlées fixées arbitrairement à 1, 2, 3, 10. Les e_{ij} sont supposés indépendants et distribués suivant la loi normale centrée réduite. y_{ij} est la jème observation de la variable y pour $x = x_i$. Dans un premier temps, nous avons “fabriqué” des données “expérimentales” correspondant à ce modèle puis nous les avons analysées d’une part en utilisant la méthode du maximum de vraisemblance, d’autre part en utilisant notre méthode.

Les données expérimentales ont été obtenues en effectuant sur ordinateur pour chaque valeur de x_i 10 tirages au sort indépendants dans une loi normale de moyenne $x_i + 2$ et de variance 1. Le tableau II fournit pour chaque valeur de x_i les 10 valeurs y_{ij} .

L’analyse statistique classique de ces données fournit les estimations \hat{A} de la pente 1.043 et \hat{B} de l’ordonnée à l’origine 1.998. La variance résiduelle vaut 0.857 avec 90 degrés de liberté et le test de déviation à la linéarité (0.962 à comparer à F_{90}^8) est non significatif. La variance de l’estimation de la pente

Tableau II

Données expérimentales obtenues par tirage au sort pour y dans une loi normale de moyenne $x_i + 2$ et de variance 1.

i	x_i	valeurs de y_{ij}										moyenne des y_{ij}	variance des y_{ij}
1	1,0	1.975	2.487	1.811	4.948	2.897	3.659	3.233	3.620	2.819	2.831	3.029	0.828
2	2.0	3.927	5.462	2.435	3.845	2.693	4.978	4.978	3.861	4.468	3.494	3.986	0.917
3	3,0	4.879	3.437	6.059	4.742	4.488	4.297	6.168	6,102	4.098	5.156	4.943	0.864
4	4.0	5.830	6.413	6.682	5.639	6.616	6.616	4.635	8.342	6.737	8.819	6.300	0.920
5	5.0	7.782	6.388	5.306	8.536	8.479	8.835	8.102	7.583	7.376	7.481	7.557	1.100
6	6.0	6.734	8.363	7.929	8.433	7.874	9.253	8.570	6.824	8.015	8.144	8.014	0.582
7	7.0	8.685	9.338	10.553	9.330	8.670	9.572	9.037	11.064	10.654	8.807	9.572	0,771
8	8.0	9.637	11.313	10.176	10.227	8.465	10.891	9.504	10.305	10.293	11.469	10.229	0.794
9	9.0	10.589	9.288	11.800	12.124	11.261	13.210	10.972	10.546	11.933	13.132	11.486	1.460
10	10.0	11.540	12.263	12.423	12.021	12.056	12.529	12.439	11.787	13.872	11.794	12.243	0.322

vaut 0.103×10^{-2} , celle de l'ordonnée à l'origine 0.400×10^{-1} et la corrélation entre ces deux estimations – 0.886 .

La méthode du maximum de vraisemblance nous a donné les estimations \hat{A} et \hat{B} des paramètres du modèle ce qui permet de calculer les valeurs $\hat{y}_i = \hat{A}x_i + \hat{B}$. Soit Y_i la i ème moyenne observée. On sait que Y_i doit fluctuer dans une loi normale de moyenne \hat{y}_i et de variance $0.857/10 = 0.0857$.

100 séries de 10 points pseudoexpérimentaux ont été tirés au sort successivement. Les résultats sont résumés sur le tableau III. On constate qu'ils sont très proches de ceux obtenus par le maximum de vraisemblance et que la convergence de la méthode est rapide. Cet exemple simple nous a permis de vérifier que la méthode proposée fournissait dans ce cas des résultats en accord avec la méthode classique. Il est évident qu'en pratique il est hors de question d'utiliser cette méthode pour analyser un modèle de régression linéaire. La justification de la méthode est fournie dans l'application précédente, alors que le modèle n'est connu que par simulation sur calculateur numérique.

Tableau III

Résultats comparatifs de l'analyse des données du Tableau II par la méthode proposée et la méthode classique

nombre d'itérations	Méthode Proposée					Maximum de vraisemblance
	20	40	60	80	100	
écart-type de la pente	0.031	0.029	0.032	0.030	0.031	0.032
écart-type de l'ordonnée	0.258	0.225	0.250	0.232	0.225	0.200
corrélation pente ordonnée	- 0.840	- 0.836	- 0.874	- 0.861	- 0.886	- 0.886

III – DISCUSSION CONCLUSION

Pour interpréter une courbe expérimentale de mitoses marquées PMM, il a été proposé de nombreuses méthodes biomathématiques (1), (2), (8). Seul Macdonald (7) fournit grâce au maximum de vraisemblance la précision des estimations des paramètres du cycle cellulaire. Cependant pour mener à bien ses calculs l'auteur doit faire des hypothèses sur la nature des distributions des phases du cycle : loi gamma ou normale. La méthode que nous proposons calcule également ces précisions mais sa mise en œuvre peut se faire quelle que soit l'hypothèse faite sur les distributions des durées des phases.

Ce cas est très fréquent en biologie. La complexité mathématique des modèles nécessite souvent des hypothèses restrictives sans lien avec le problème biologique pour rendre les calculs possibles. La méthode que nous proposons adaptée à chaque problème particulier peut être utilisée dans de tels cas où le maximum de vraisemblance n'est pas applicable.

BIBLIOGRAPHIE

- (1) BARRETT J.C. – Optimized parameters for the mitotic cycle. *Cell Tissue Kinet.*, 1970, 3, 349.
- (2) GILBERT C.W. – The labelled mitoses curve and the estimation of the parameters of the cell cycle. *Cell Tissue Kinet.*, 1970, 3, 1.
- (3) HILL C.W. – Algorithm 395 Student's distribution. *Communications of the A.C.M.*, 1970, 10, 617.
- (4) KENDALL M.G., STUART A. – The advanced theory of statistics 1964, Vol. II, Griffin Ed. London.
- (5) KOWALIK J. and OSBORNE M.R. – Modern analytic and computational methods. *Science and Mathematics* Bellman Edit. 1968.
- (6) LOTKA A.J. – Elements of physical biology. Dover publications, New York, 1926.
- (7) P.D.M. MACDONALD – Statistical Inference from the fraction labelled mitosis curve. *Biometrika*, 1970, 57, 489.
- (8) TAKAHASHI M., HOGG J.D. and MENDELSON M.L. – The automatic analysis of PLM curves. *Cell Tissue Kinetics*, 1971, 4, 505.
- (9) VALLERON A.J., FRINDEL E. – Computer simulation of growing cell populations, *Cell Tissue Kinetics*, 1973, 1, 69.
- (10) VALLERON A.J., MARY J.Y., FRINDEL E. – Méthode d'analyse sur ordinateur des courbes de mitoses marquées. *Biomedicine*, 1973, 18, 118.