Rapport de stage : Bio-informaticien à l'Institut National de la Recherche Agronomique

Tuteur dans l'entreprise : Romain Yvinec
Tuteur enseignant : Hermine Biermé

Quilly Erwan 2015/2016

Résumé :
Abstarct :

Sommaire

Introduction :

Les principaux sujets de ma formation sont les statistiques et leurs applications dans divers domaines de la biologie. J'ai donc du trouver un stage dans une entreprise ayant un rapport avec ces aspects. J'ai envoyé une demande de stage au centre INRA de Tours, ce centre travaillant principalement dans la recherche agronomique, ce qui est d'un grand intérêt pour moi.

J'ai travaillé pendant 2 mois à l'INRA de Tours en tant que biostatisticien dans une équipe de bio-mathématiques. J'ai principalement fait de la programmation à l'aide du logiciel R dont le but était de calculer le biais entre 2 jeux de données et de représenter ces données grâce à des méthodes d'optimisation.

Dans ce rapport, je vais d'abord présenter l'INRA, ses missions et l'équipe où j'ai travaillé. Ensuite, j'expliquerai comment fonctionne les programmes que j'ai réalisés.

1. Présentation de l'entreprise

1.1 Les missions de l'INRA

L'Institut National de la Recherche Agronomique (INRA) est la première entreprise de recherche agronomique d'Europe. Elle mène de nombreuses recherches dans différents domaines. On peut résumer le travail de l'INRA en 4 mots : Explorer, Comprendre, Expérimenter et Anticiper.

L'une des actions de l'INRA est de produire et diffuser des connaissances afin de contribuer à la compréhension du monde et de ses évolutions, au développement et à la gestion des biens publics, au bien-être des citoyens ainsi qu'à l'innovation socio-économique. L'INRA utilise aussi ces connaissances pour déboucher sur des innovations qui bénéficieront aux autres entreprises (agricoles, industrielles, services).

Les actvités de l'INRA ont un impact dans de nombreux domaines. Les chercheurs et techniciens de l'institut cherchent à nourir la France et le monde, travaillent sur le réchauffement climatique, à réduire la dépendance aux pesticides et aux engrais, à la sauvegarde des espèces et également à conserver la biodiversité génétique des plantes et des animaux. Ils coopèrent aussi avec l'enseignement supérieur afin d'acceuillir et former les futurs chercheurs.

Les différents impacts peuvent être classés en cinq catégories : économique, environnemental, politique, sanitaire, et territorial/social. Ils sont mesurés à long terme car il faut souvent plusieurs années (temps moyen : 20 ans) pour qu'une innovation soit réalisée, mise en application et ait des conséquences sérieuses sur la société.

L'impact économique consiste évidemment à mesurer les contributions des différents acteurs d'une innovation et de les comparer aux retombés économiques pour les bénéficiaires. L'impact environnemental consiste à mesurer les effets, en bien ou en mal, sur l'environnement des innovations (énergies, déchets, ...).

1.2 Les pôles de recherche

Au centre INRA du Val de Loire où j'ai fait mon stage, il y a plusieurs pôles de recherche qui regroupe les activités du centre :

- Biologie animale intégrative et gestion durable des productions animales
- Biologie intégrative des arbres et des organismes associées
- Dynamique des sols et gestion de l'environnement

- Santé animale et santé publique

L'objectif du pôle Biologie animale intégrative et gestion durable des productions animales, en plus de la production de connaissances, est d'améliorer la durabilité des systèmes d'élevage dans les domaines économique, social et environnemental. Dans le domaine de l'économie, le but est d'assurer la compétitivité des filières, dans le domaine social, le but est de permettre la coexistence des petites et grandes exploitations et, dans le domaine environnemental, le but est de limiter l'usage des intrants (engrais, produits chimiques, hormones de synthèse, …) et gérer les paysages.

C'est pour garantir un avenir pour l'importante ressource naturelle que sont les forêts que le pôle Biologie intégrative des arbres et des organismes associées existe. On cherche à y approfondir les connaissances sur la diversité génétique des espèces forestières majeures afin d'optimiser la conservation de ces ressources et d'identifier de nouvelles espèces. On étudie aussi les insectes vivant dans ces forêts.

Le pôle Dynamique des sols et gestion de l'environnement est particulièrement important pour le développement durable. Les sols intervenant à la fois dans la production agricole et dans la protection de l'environnement. Deux approches complémentaires sont appliquées. D'une part, on inventorie et surveille la qualité des sols et, d'autre part, on étudie les déterministes du fonctionnement des sols. Tout cela ayant pour but de protéger les sols contre le tassement et l'érosion, d'améliorer l'alimentation en eau des cultures et de réduire les émissions de gaz à effet de serre par les sols.

Le pôle Santé animale et santé publique est dédié à l'étude et à la recherche sur les agents pathogènes animaux, plus spécifiquement de ceux qui sont communs à l'homme et aux animaux. Les recherches menées doivent contribuer à la maîtrise des infections des animaux d'élevage (maladies infectieuses et parasitaires) constituant un risque pour la santé publique (consommation d'aliments d'origine animale).

Pour finir, l'unité dont fait partie l'équipe au sein de laquelle j'ai travaillé (équipe BIOS) est l'unité de Physiologie de la Reproduction et des Comportements (PRC). C'est une unité de recherche qui mène des recherches fondamentales et appliquées sur la fonction de reproduction, sur les comportements émotionnels, sociaux et sexuels et sur les mécanismes d'adaptation des individus et des populations à des environnements changeants.

1.3 L'équipe BIOS

L'équipe BIOS (Biologie et Bio-informatique des Systèmes de Signalisation) regroupe des chercheurs ayant des compétences en mathématiques et en biologie pour étudier des données fournies par les expérimentateurs. On y développe et utilise des méthodes statistiques et des logiciels appropriés à cela.

Les objectifs de l'équipe sont de comprendre et modéliser les réseaux de signalisation induits par des hormones, comprendre l'action de ces hormones et tirer profit de ces connaissances pour développer des substituts à l'utilisation de ces hormones.

2. Mon stage

2.1 Motivation biologique

Les programmes réalisés pendant mon stage s'appliquent aux réseaux de signalisation intracellulaire. Ce sont des systèmes complexes de communication qui gèrent l'ensemble des signaux entre les cellules et coordonnent les actions entre ces cellules. Ces signaux existent principalement entre les les cellules d'un même organisme mais il peut toutefois y en avoir entre les cellules de 2 organismes distincts. Par exemple, un embryon communique avec les cellules de l'utérus de sa mère.

Les signaux peuvent se faire par contact direct entre les cellules, sur de courtes distances ou sur de longues distances. Les cellules reçoivent ces signaux grâce à des protéines appelées récepteurs. Ces récepteurs sont sont situés sur la surface de la cellule et chacun d'entre eux est activé par un type de signal particulier (hormones, neurotransmetteurs, …).

2.2 Présentation du modèle

La réaction biochimique Ligand-Récepteur :

Le type de réaction que j'ai étudié est une réaction impliquant un récepteur Rc situé sur la surface d'une cellule et un ligand A situé à l'extérieur de la cellule et qui va se lier au récepteur Rc. Ils forment ainsi un complexe Ligand-Récepteur ARc qui va provoquer une cascade de phénomènes à l'intérieur de la cellule dont le but est de transformer un substrat S en un produit, via une ou plusieurs réactions enzymatique. E est la molécule effectrice de cette réaction. C'est donc la quantité d'élément E que l'on veut mesurer. On

notera que la formation du complexe ARc et du substrat S en produit E sont réversible. On a les réactions chimiques suivantes :

$A + Rc \xrightarrow{k1} ARc$

k1 étant le coefficient de réaction de la formule et k2 celui de la formule inverse

$ARc \xrightarrow{k2} A + Rc$

$n \cdot ARc + S \xrightarrow{k3} n \cdot ARc + E$

k3 est le coefficient de réaction et n est un entier ≥ 1

$E \xrightarrow{k4} S$

k4 est le coefficient de réaction

Les coefficients de réaction permettront de calculer la vitesse de réaction des produits et réactifs, c'est à dire leur vitesse d'apparition/disparition.

Le modèle opérationnel :

Dans le cas où la quantité de molécules E présente quand [A]=0 est nulle, le modèle le plus couramment utilisé dans le domaine de la modélisation biochimique est : $E = \frac{Em \cdot [A]^n \cdot \tau^n}{[A]^n \cdot \tau^n + ([A] + Ka)^n}$

La réponse du modèle est la concentration [E] de la molécule E. [A] est la concentration du Ligand L et correspond à la variable explicative du modèle. Em, τ et Ka sont des paramètres que l'on doit estimer. On peut les exprimer en fonction de n, des conditions initiales de la réaction et des coefficients définis précédemment qui sont connus. On peut prouver l'existence de ce modèle grâce aux calculs de vitesse de réaction suivants :

$$\frac{d[A]}{dt} = -k1 \cdot [A] \cdot [Rc] + k2 \cdot [ARc]$$
$$\frac{d[Rc]}{dt} = -k1 \cdot [A] \cdot [Rc] + k2 \cdot [ARc]$$
$$\frac{d[ARc]}{dt} = k1 \cdot [A] \cdot [Rc] - k2 \cdot [ARc]$$
$$\frac{d[E]}{dt} = k3 \cdot [ARc]^n \cdot [S] - k4 \cdot [E]$$
$$\frac{d[S]}{dt} = -k3 \cdot [ARc]^n \cdot [S] + k4 \cdot [E]$$

On a $\frac{d[Rc]}{dt}+\frac{d[ARc]}{dt}=0$ donc, d'après la loi de conservation, on peut en déduire que $[Rc](t)+[ARc](t)=Rtot$ pour tout temps t. De même, on a $[E](t)+[S](t)=Etot$ pour tout temps t. La quantité totale de récepteurs Rc et ARc et de substrat-produit est constante. A l'équilibre, on a $\frac{d[A]}{dt}=0$, on obtient donc les équations suivantes :

$k2\cdot[ARc]=k1\cdot[A]\cdot[Rc]$
$[Rc]=Rtot-[ARc]$

On en déduit donc que : $k2\cdot[ARc]=k1\cdot[A]\cdot(Rtot-[ARc])$

=> $[ARc]=\frac{Rtot\cdot k1\cdot[A]}{k2+k1\cdot[A]}$

On cherche maintenant à déterminer [E] en fonction de [ARc], on a $\frac{d[E]}{dt}=0$ à l'équilibre donc :

$k4\cdot[E]=k3\cdot[ARc]^n\cdot[S]$
$[S]=Etot-[E]$

=> $k4\cdot[E]=k3\cdot[ARc]^n\cdot(Etot-[E])$ => $[E]=\frac{Etot\cdot k3\cdot[ARc]^n}{k4+k3\cdot[ARc]^n}$

On remplace [ARc] par l'expression calculée précédemment et on obtient ainsi :

$$[E]=\frac{Etot\cdot(\frac{k3\cdot Rtot^n}{k4})\cdot[A]^n}{[A]^n\cdot(\frac{k3\cdot Rtot^n}{k4})+([A]+\frac{k2}{k1})^n}$$

On a alors une expression du même type que celle du modèle opérationnel :

$$E=\frac{Em\cdot[A]^n\cdot\tau^n}{[A]^n\cdot\tau^n+([A]+Ka)^n}$$

On peut identifier dans le cas présent :

$Em=Etot$ $\tau=\sqrt[n]{\frac{k3}{k4}}\cdot Rtot$ $Ka=\frac{k2}{k1}$

Em est la quantité maximale de substrat-produit. τ est l'efficacité de transduction du signal, elle ne dépend que des paramètres en rapport avec la

consommation/formation de substrat et de produit. Ka est la constante de dissociation d'équilibre entre le Ligand et le Récepteur, elle ne dépend que des paramètres concernant la formation du complexe Ligand-Récepteur.

Autres schémas :

Il existe d'autres réactions Ligand-Récepteur. Par exemple, on a :

(1)

$$n \cdot A + Rc \rightarrow ARc$$
$$ARc \rightarrow n \cdot A + Rc$$
$$ARc + S \rightarrow ARc + E$$
$$E \rightarrow S$$

(2)

$$A + n \cdot Rc \rightarrow ARc$$
$$ARc \rightarrow A + n \cdot Rc$$
$$ARc + S \rightarrow ARc + E$$
$$E \rightarrow S$$

Ces réactions étant différentes de la première, le modèle opérationnel est également différent. En gardant les mêmes coefficients de réaction et en appliquant la même méthode de calcul, on obtient pour la réaction (1) :

$$[E] = \frac{Em \cdot \tau^n \cdot [A]^n}{\tau^n \cdot [A]^n + ([A] + Ka)^n}$$

avec $Em = Etot$ $\quad Ka = \sqrt[n]{\frac{k2}{k1}}$ $\quad \tau = \sqrt[n]{\frac{k3 \cdot Rtot}{k4}}$

On peut aussi considérer que la quantité d'enzyme E présent quand [A]=0 n'est pas nulle. On rajoute alors le paramètre Basal qui mesure cette quantité et on obtient l'équation suivante :

(3) $$E = Basal + \frac{(Em - Basal) \cdot \tau^n \cdot [A]^n}{\tau^n \cdot [A]^n + ([A] + Ka)^n}$$

2.3 Biais de signalisation

Les ligands produisent des signaux favorisant ou limitant certains phénomènes chimiques de la réaction. Cela peut modifier la quantité de produit

E générée par la réaction. Une cause possible de ces changements est la concentration du ligand ou des protéines. En fonction de cela, le nombre de gènes activé est différent. Parmi les réponses E calculées, la plupart sont donc biaisées.

On mesure un biais à partir de 2 réponses en comparant leurs paramètres τ et Ka respectifs, ou plus précisément le rapport log(τ/Ka) des 2 réponses car individuellement ces 2 paramètres sont difficilement identifiable. On appelle log(τ/Ka) le ratio de transduction, il représente une estimation de la puissance et de l'efficacité de la réaction. Parmi les 2 réponses, l'une sert de référence.

2.4 Utilisation du modèle et quantification du biais :

2.4.1 Optimisation du modèle :

La fonction de vraisemblance est une description des données d'une loi statistique en fonction de ses paramètres. La vraisemblance du modèle est la probabilité que les données expérimentales correspondent au modèle. On veut donc la maximiser.

Calcul de la vraisemblance :

On doit d'abord estimer les paramètres du modèle (3) (Em, Basal, n, τ et Ka) à partir des données de dose-réponse pour que la vraisemblance soit maximale. On utilise pour cela la méthode du maximum de vraisemblance. On fait varier la concentration Ai du ligand et on mesure la réponse Yi correspondante où i=1 … m. On suppose que Yi = E(Ai) + ε où ε~N(0,σ²) avec σ connue. Avec ces données, la vraisemblance L du modèle est :

$$L = \prod \left(\frac{1}{\sigma \cdot \sqrt{2 \cdot \Pi}} \cdot \exp\left(\frac{-\left(Yi - E\left(Ai \right) \right)^2}{\sigma^2} \right) \right)$$

On veut donc trouver les paramètres qui maximisent cette fonction mais, pour simplifier la méthode, on cherchera plutôt à minimiser la fonction l = -logL :

$$l = \frac{m \cdot \log\left(\sigma^2 \cdot 2 \cdot \Pi \right)}{2} + \frac{1}{2 \cdot \sigma^2} \cdot \sum \left(Yi - E\left(Ai \right) \right)^2$$

algorithme de calcul des paramètres :

Puisque l'on travaille sur R, on utilise une boucle While pour programmer un algorithme qui, à chaque itération, calcule le gradient G de l et le pas tk puis $\theta(k+1) = \theta(k) + tk*G$ où $\theta(k) = (Em, Basal, n, \tau, Ka)$

Le gradient G est calculé par la formule $G=J^t*f$ où f est un vecteur de taille m tel que f(i) = Yi-E(Ai) et J est la jacobienne de l, c'est à dire une matrice de taille m x 3 tel que J(i,j) est la dérivée de f(i) par rapport à la jième variable. La boucle s'arrête après un certain nombre d'itération ou si l'écart entre $\theta(k)$ et $\theta(k+1)$ ne dépasse plus un certain seuil.
Une fois que l'on a calculé l'estimateur θ, on calcule la valeur du modèle E pour chaque concentration Ai et en remplaçant les paramètres inconnus par les valeur de l'estimateur correspondante. On compare ainsi ces résultats avec les valeurs Yi dans un même graphe :

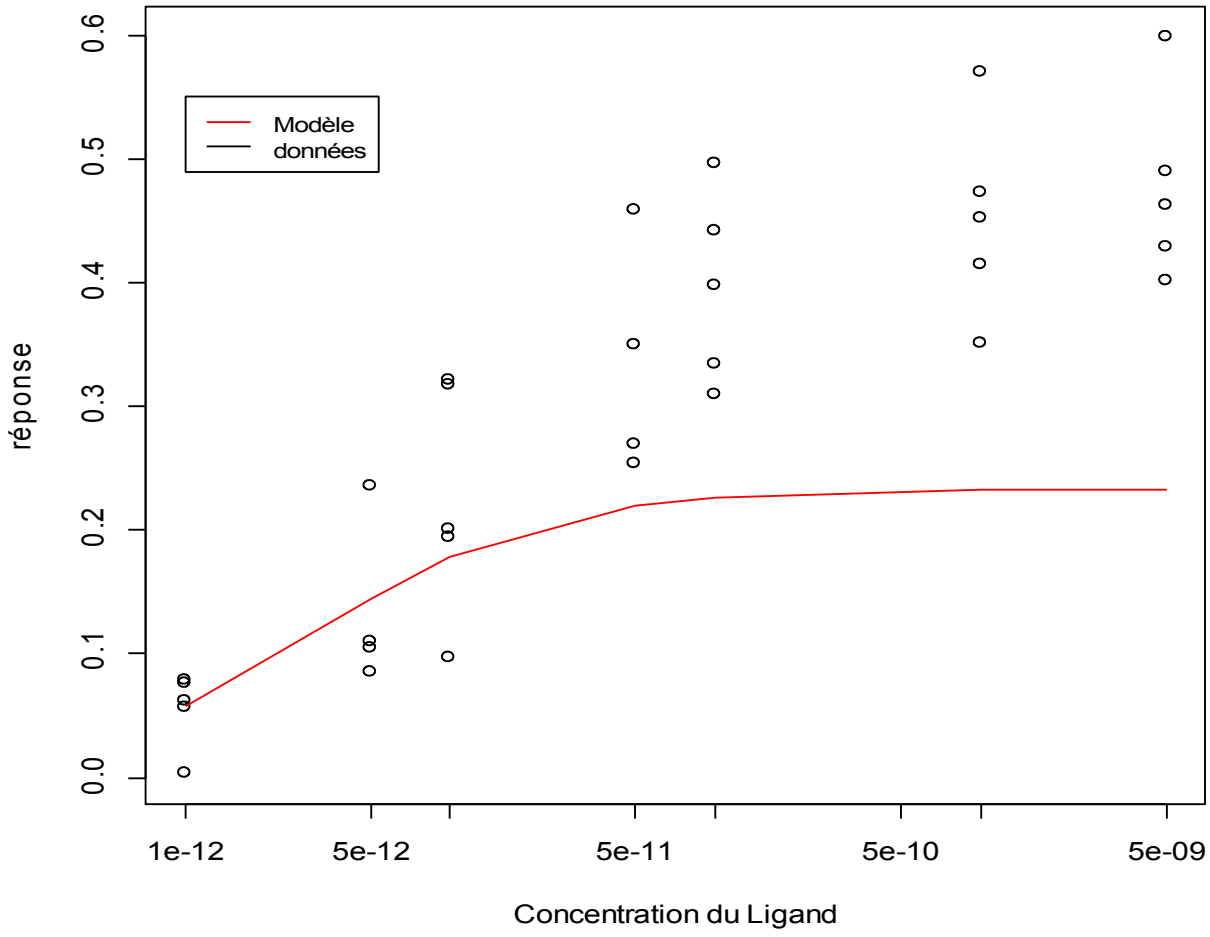


Illustration 1: Algorithme + Gradient

Reparamétrage et méthode optimx :

On remarque que, bien que la courbe du modèle et les points Yi aient des variations semblables, il y a un net écart entre eux. On va donc changer de méthode d'optimisation et reparamétrer le modèle E. Au lieu d'estimer les paramètres τ et Ka, on prend comme paramètres log(Ka) et log(τ/Ka) que l'on simplifie en log(R). Ce dernier est le paramètre le plus important et est appelé ratio de transduction. On garde toujours Em, Basal et n comme paramètres. Le modèle E se présente alors sous le forme suivante :

$$E = Basal + \frac{Em - Basal}{1 + \frac{\left(\frac{[A]}{10^{\log(Ka)}} + 1\right)^{n}}{\left(10^{\log(R)} \cdot [A]\right)^{n}}}$$

On change aussi de méthode pour minimiser la fonction de log vraisemblance. On utilise désormais la fonction optimx utilisable dans R. Elle se présente sous la forme :

x=optimx(ohm, L, gr=NULL, hess=NULL, lower=c(0, 0, 0, 5, -3), upper=c(1, 0.1, 4, 12, 3), method=c("L-BFGS-B"), itnmax=100, hessian=TRUE)

ohm est le vecteur contenant les conditions initiales des paramètres à estimer, L est la fonction à minimiser, gr=NULL et hess=NULL peuvent être remplacé par des fonctions calculant le gradient et la hessienne de L si besoin selon la méthode de calcul indiquée par method, lower et upper indiquent les valeurs minimales et maximales des paramètres, itnmax indique le nombre maximal d'itération et hessian=TRUE permet de calculer la hessienne de L une fois le calcul des paramètres effectués. On choisit comme bornes à priori pour les paramètres :
Em Є [0,1] ; Basal Є [0,0.1] ; n Є [0,4] ; log(R) Є [5,12] ; log(Ka) Є [-3,3]

coef(x) nous renvoie ensuite les paramètres estimés par la fonction optimx. On calcule aussi la vraisemblance pour ces paramètres. Pour être sûr d'avoir une bonne estimation des paramètres, on lance cette optimisation plusieurs fois en changeant les valeurs initiales des paramètres et on garde les paramètres qui ont la vraisemblance la plus faible car sinon on risque de trouver seulement un minimum local au lieu d'un minimum global. On retrace alors la courbe du modèle E et elle correspond bien mieux aux données Yi :

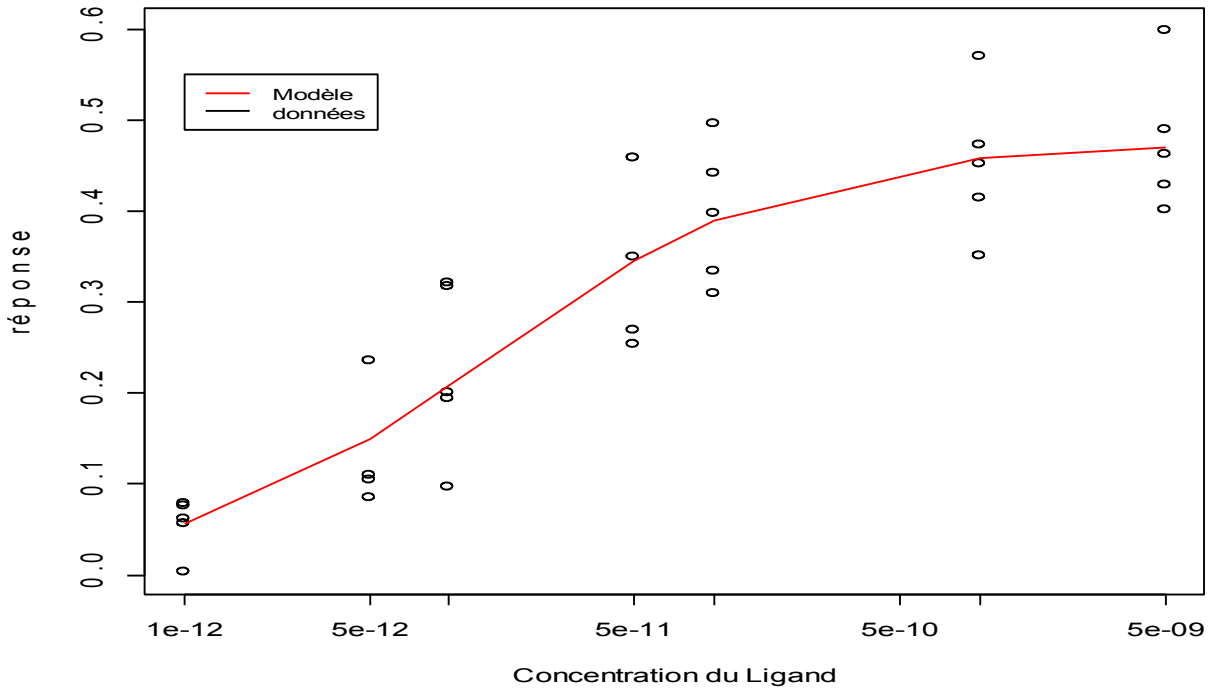
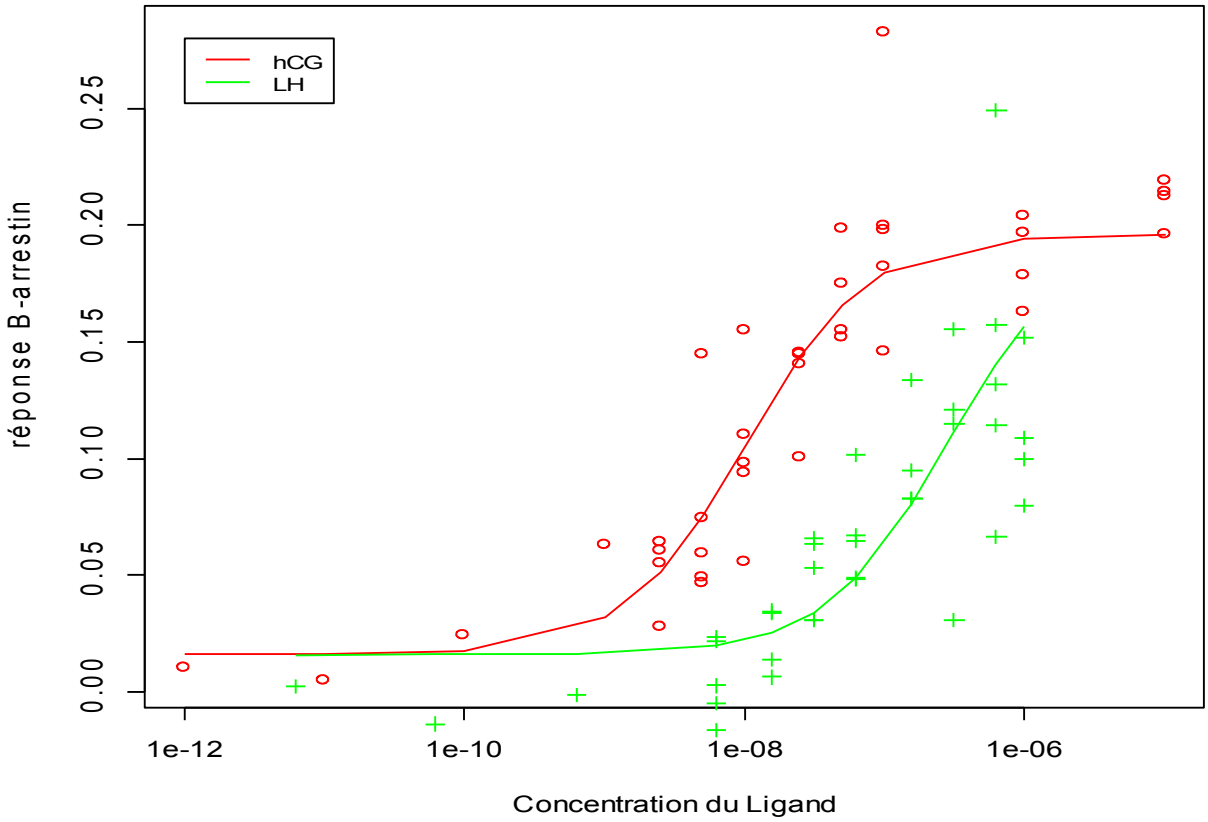


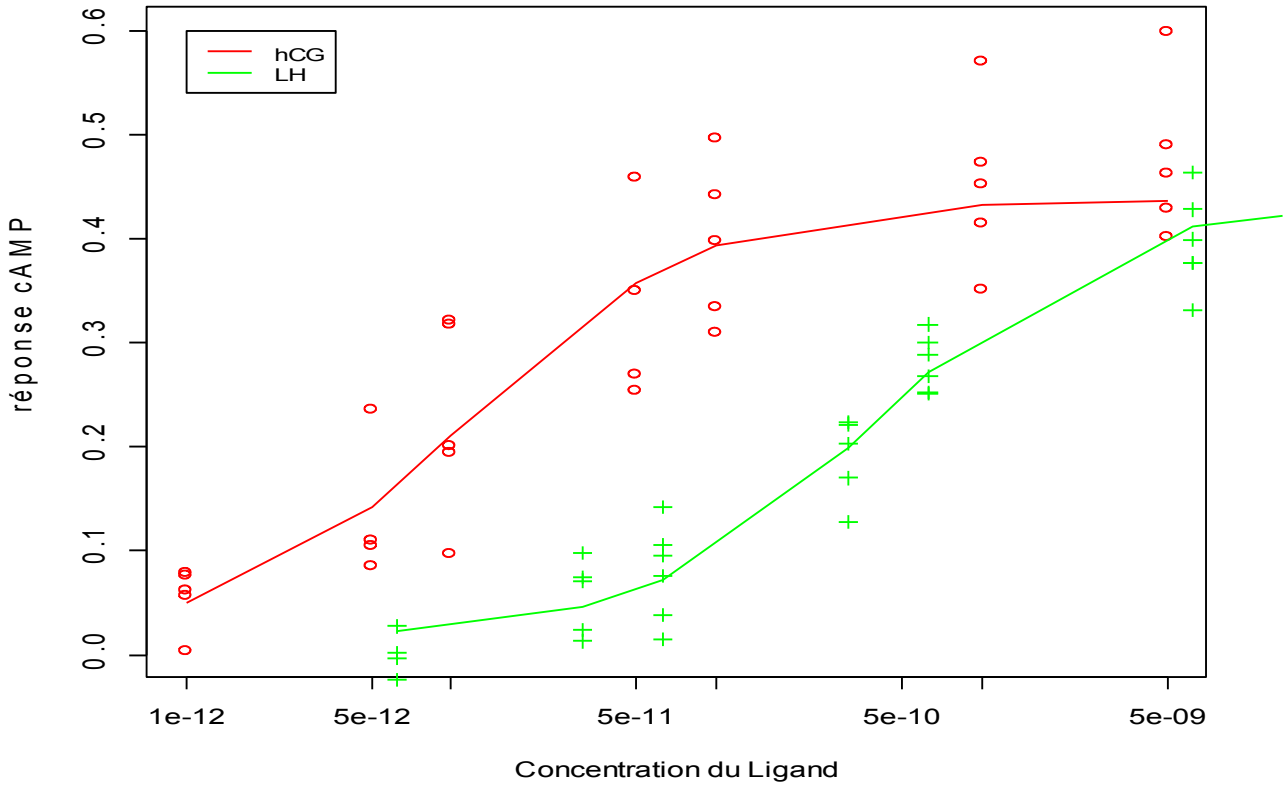
Illustration 2: Optimx + Reparamétrisation

2.4.2 Calcul du biais et des intervalles de confiance:

On travaille maintenant sur 4 jeux de données Yi différents. Les 2 premiers correspondent à une même réponse mesurée (cAMP) tout comme les 2 autres (β-arrestin) et on leur applique aussi une hormone qui joue le rôle de ligand (CG ou LH). On veut calculer le biais entre ces 2 réponses qui s'exprime par le paramètre log(R). On considère que pour une même réponse les paramètres Em, Basal et n ne changent pas selon l'hormone appliquée contrairement à log(R) et log(Ka). On a donc 14 paramètres à estimer :
Em1, Basal1, n1, log(R1), log(Ka1), log(R2), log(Ka2) pour cAMP
Em3, Basal3, n3, log(R3), log(Ka3), log(R4), log(Ka4) pour β-arrestin

Pour estimer ces paramètres, on applique la même méthode que précédemment sauf que la fonction l est maintenant la somme des vraisemblances des 4 jeux de données. De plus, θ contient alors 14 paramètres au lieu de 5. Une fois ces paramètres estimés, on peut calculer et tracer un modèle par jeu de données :


hCG
LH
réponse B-arrestin
0.00
0.05
0.10
0.15
0.20
0.25
1e-12
1e-10
1e-08
1e-06
Concentration du Ligand


hCG
LH
réponse cAMP
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1e-12
5e-12
5e-11
5e-10
5e-09
Concentration du Ligand

On voit très nettement qu'il y a un écart entre les différentes courbes, cela est dû au biais que l'on va calculer à partir de nos 4 valeurs de log(R). On commence par calculer l'écart entre log(R1) et log(R2) et l'écart entre log(R3) et log(R4). Le biais est la différence entre ces 2 écarts :
OR1=log(R1)-log(R2)
OR2=log(R3)-log(R4)
biais=OR1-OR2

On veut ensuite déterminer les intervalles de confiance des 14 paramètres. On a 2 méthodes pour cela.

Intervalles de confiance Méthode 1 :

Pour la première méthode, on suppose que les données suivent une loi normale et que l'effectif est infini. Soit θestim le vecteur contenant les paramètres estimés et H la hessienne de l. On cherche :
$\{\theta ; l(\theta)-l(\theta estim)<seuil\}$
ce qui est équivalent à $\{\theta ; 1/2*(\theta-\theta estim)^t *H(\theta estim)*(\theta-\theta estim)<seuil\}$
d'après l'approximation de l(θ) par la formule de Taylor. On obtient H(θestim) grâce à l'option hessian=TRUE de optimx et l'intervalle de confiance [θ- ;θ+] correspondant à cet ensemble se calcule de la façon suivante :

θ-(i) = θestim(i) – (seuil*C(i,i))^1/2
θ+(i) = θestim(i) + (seuil*C(i,i))^1/2
où C=2*H(θestim)^(-1) et le seuil est le quantile de la loi χ^2 de niveau α et à 1 degré de liberté. On choisit ici α=95%.

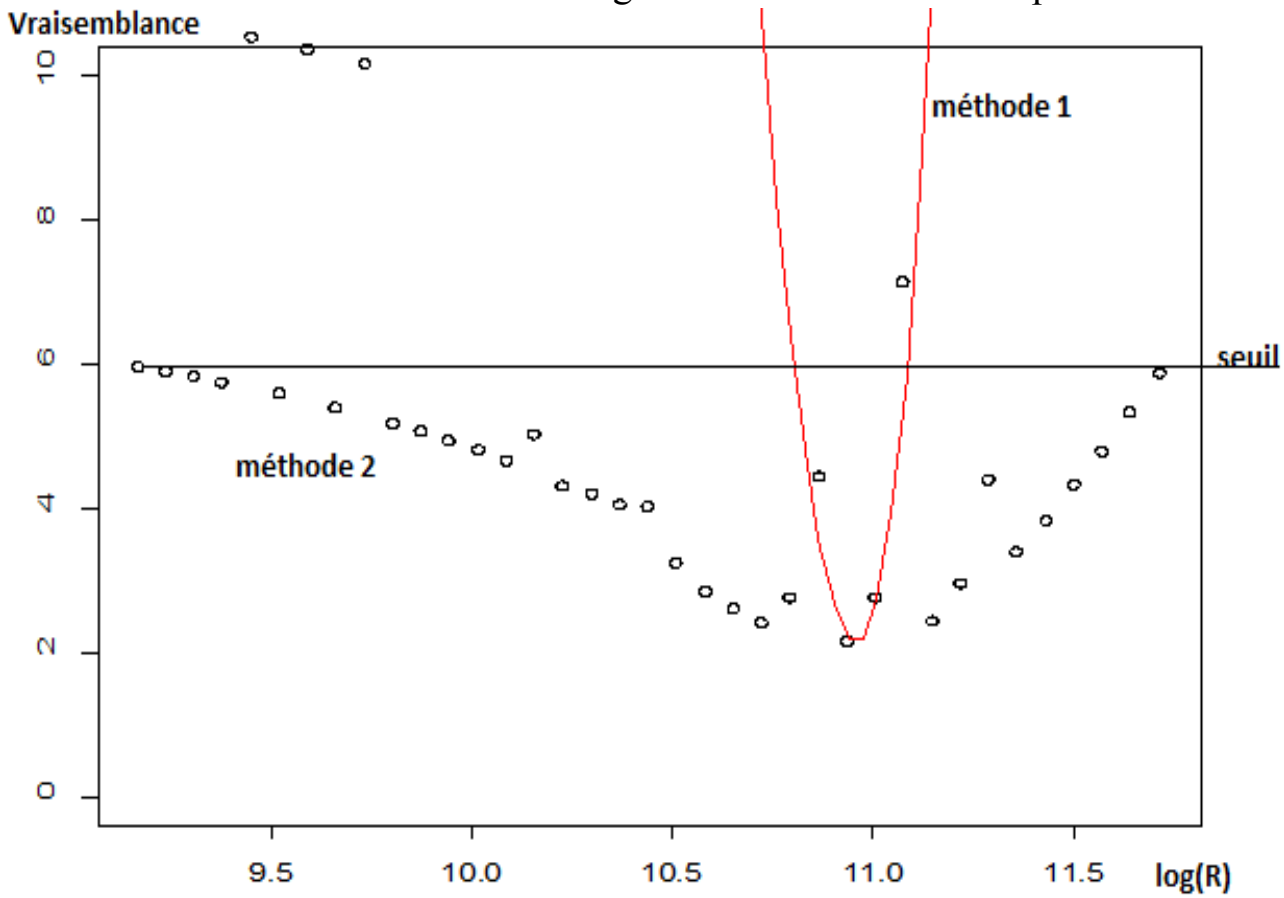
Intervalles de confiance Méthode 2 :

Pour la deuxième méthode, on n'a pas à faire l'hypothèse que l'effectif est infini, ce qui la rend préférable à la première car elle est plus juste. Pour chaque paramètre θ(i), on cherche $\{\theta(i) ; l(\theta, \theta(i) \text{ fixé})-l(\theta estim)<seuil\}$

On utilise le même programme de minimisation de l que précédemment mais en ayant pris une valeur fixe pour θ(i). Une fois les 13 autres paramètres estimés par la méthode optimx, on calcule la vraisemblance associée. On relance ce même procédé pour un grand nombre de valeurs possibles de θ(i) et celles dont la vraisemblance est inférieure à l(θestim)+seuil forment l'intervalle de confiance.

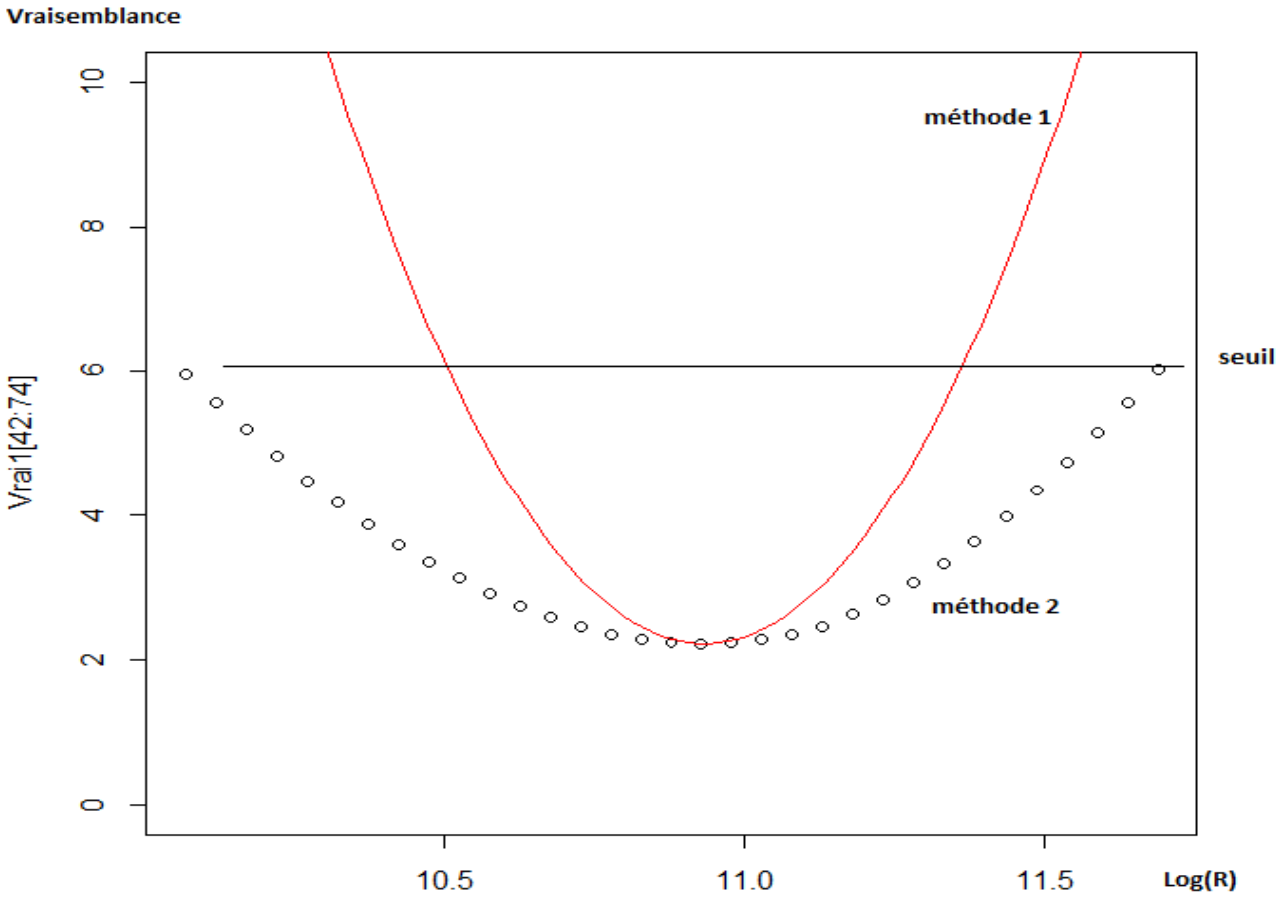
Comparaison des 2 méthodes :

On peut comparer les 2 méthodes en traçant les courbes représentant la vraisemblance en fonction de θ(i) selon chaque méthode. L'intervalle de confiance de la 2e méthode est, en général, plus large que celui de la 1ere méthode. Ils sont représentés par l'ensemble des points situés en dessous de la courbe du seuil. On note aussi des irrégularités dans la courbe représentant ce



second intervalle :

On peut améliorer cet algorithme en fixant log(Ka1)=log(Ka3) =0 et n1=n3=1 au début de l'algorithme. Cela rend les données plus facilement identifiable même si cela n'est pas totalement juste :



2.5 Tests statistiques de significativité

On cherchera surtout à déterminer les intervalles de confiance des paramètres log(R). On va donc maintenant faire un test statistique pour évaluer la significativité des paramètres log(R) estimés, des intervalles de confiance et du biais. On considère que chaque log(R) suit une loi normale. On fait un test sur la moyenne des log(R) avec une variance inconnue.

On doit donc estimer individuellement la variance σi de chaque log(Ri) pour ensuite calculer la variance totale σ des log(R).
$\sigma^2 = \sum \sigma i^2$

On peut estimer σi de 2 façons différentes grâce aux 2 intervalles de confiance calculés précédemment, la méthode 2 restant préférable. Selon la méthode 1, on a :
IC1 = [θestim(i) – (seuil*C(i,i))^1/2 ; θestim(i) + (seuil*C(i,i))^1/2]
et on considère donc que $\sigma i^2 = C(i,i)$ d'après la normalité de θ(i).

Pour la méthode 2, on utilise le 2e intervalle de confiance calculé. Toujours d'après la normalité de θ(i), on a :

IC2 = [θestim(i) – (seuil*σi²)^1/2 ; θestim(i) + (seuil*σi²)^1/2]
IC2 = [θestim(i) – Δi ; θestim(i) + Δi]

On peut calculer Δi facilement car on connaît déjà les valeurs de θestim(i) et de IC2 et on a donc :
σi² = Δi²/seuil

Maintenant que l'on a une valeur estimée pour la variance globale σ des log(R). On peut passer au test sur la moyenne μ des log(R).

H0 : μ=0 H1 :μ ≠ 0
Si l'hypothèse nulle H0 est vraie alors on pourra considérer les données et le biais comme non significative. La statistique de test est :
T= ∑Yi/(σ*m^1/2) T~St(m-1) où m est le nombre de replica, c'est à dire le nombre de fois où l'on la même concentration Ai de Ligand. Dans ce cas, on considère que m=4.

Si |T|>t(α,m-1) alors on rejette H0
pour α=5% et m=4 t(α,m-1)=2,35
On trouve T= 6,98 donc on rejette H0
On calcule aussi la p-valeur :
p-valeur=1 – P(|Z|<T)=1 – P(Z<T) + P(Z<-T) = 0,006 Z~St(n-1)

La p-valeur est très proche de 0 donc on peut en déduire H0 est presque toujours fausse quelque soit α. On a donc bien un biais et des intervalles de confiance significatifs.

Le second test nous permet de comparer si le biais est nul ou non. On prend donc comme hypothèses :

H0 : biais = 0 H1 : biais ≠ 0

On utilise le test du rapport de vraisemblance, on estime la vraisemblance L0 sous H0 et la vraisemblance L1 sous H1 et on compare les 2 valeurs -2*log(L0)+2*k0 et -2*log(L1)+2*k1 où k0 et k1 sont respectivement le nombre de paramètres estimés sous H0 et H1. On rajoute cela au calcul de la vraisemblance afin de pénaliser une hypothèse si elle a plus de paramètre que l'autre.

Sous H1, on ne pose pas de contrainte pour l'optimisation de θ. On utilise

la méthode optimx et on obtient une valeur de θ et de L1 semblable à celles de l'hypothèse H0 du précédent test. Sous H0, pour que l'on ait un biais nul, on pose une contrainte pour l'un des paramètres log(R). On rappelle que
biais = (log(R1) − log(R2)) − (log(R3) − log(R4))
On pose donc comme contrainte log(R4) = log(R2) − log(R1) + log(R3)
On a donc un paramètre en moins sous H0 et on estime les autres paramètres avec la même méthode. La statistique de test est la différence entre les 2 valeurs -2*log(L0)+2*k0 et -2*log(L1)+2*k1. On rejette H0 si la statistique de test est supérieure au quantile de χ^2 à 1 degré de liberté de niveau 1−α. On calcule aussi la p-valeur comme précédemment :
p-valeur=1 – P(Z<T) $Z \sim \chi^2(1)$
On trouve une p-valeur de 0,17. Plus précisément, on rejettera H0 si le seuil α est supérieur à 0,17 et sinon, on rejettera H1. On peut donc supposer que H1 est fausse car on prend rarement α > 0,1.

On réalise aussi un dernier test statistique pour assurer que les paramètres log(R) estimés grâce à la fonction optimx sont bien les meilleurs optimisations. On utilise pour cela le test du rapport de vraisemblance pour comparer ces paramètres avec d'autres valeurs possibles. On a donc comme hypothèses :

H0 : θ = θestim H1 : θ = θ1

La valeur des paramètres log(R) de θ1 est choisie de manière aléatoire dans le même intervalle que précédemment (log(R) Є [5,12]). Les autres paramètres de θ1 sont ré-optimisés en fonction de ces log(R) toujours par la méthode optimx et en étant toujours contenus dans les mêmes intervalles (Em Є [0,1] ; Basal Є [0,1] ; n Є [0,4] ; log(Ka) Є [-3,3]).

Une fois la valeur de θ1 obtenue, on calcule le rapport entre la vraisemblance L1 des données si θ=θ1 et la vraisemblance L0 des données si θ=θestim :

T=L1(Y1, … , Ym)/L0(Y1, … , Ym)

La statistique de test T est équivalente à une loi χ^2 à p degrés de liberté où p est le nombre de contraintes (ici p=4). On rejette H0 si la statistique de test est supérieure au quantile de χ^2 au niveau 1-α et à p degrés de liberté. On calcule aussi la p-valeur :
p-valeur=1 – P(Z<T) $Z \sim \chi^2(p)$

On refait ce test une centaine de fois avec différentes valeurs de log(R)

pour θ1. On obtient à chaque fois une statistique de test très proche de 0 et une p-valeur supérieure à 0,99. On en déduit donc que l'hypothèse H0 ne peut être rejetée, ce qui confirme la première estimation de θ. Ce test est une vérification de la méthode optimx.