Table des matières

.

Chapitre 1

Introduction

Au cours de ce projet nous allons étudier les triangles de développement de sinistres en assurance (run-off triangles) et tenter de définir un modèle de prédiction de ces triangles.

En assurance les triangles de développement peuvent être vus comme un récapitulatif des sinistres survenus au cours de l dernières années. Ceux-ci sont classés dans un tableau en fonction de l'année du sinistre et de l'année de déclaration du sinistre. Les sinistres n'étant, souvent, pas déclarés l'année oú ils surviennent, l'assureur se doit de créer des provisions.

Voici comment se présente un triangle de développement:

	Années d'occurence des sinistres						
Année de développement	$N_{1,1}$	$N_{1,2}$	$N_{1,3}$		$\cdots$		$N_{1,l}$
	$N_{2,1}$	$N_{2,2}$	$N_{2,3}$		$\cdots$	$N_{2,l-1}$	
	$N_{3,1}$	$N_{3,2}$	$N_{3,3}$	$\cdots$	$N_{3,l-2}$		
	$\vdots$	$\vdots$					
	$N_{l-1,1}$	$N_{l-1,2}$					
	$N_{l,1}$						

Si l'on se réfère au cas de l'assurance invalidité, par exemple, voici l'interpretation que l'on peut tirer d'un triangle de développement. $N_{i,j}$ représente le nombre de sinistres survenu durant l'année j et déclarés dans le courant de l'année i. Ainsi, j indique l'année de l'accident tandis que i est l'année oú une rente d'invalidité est déclarée.

Pour effectuer des provisions un assureur doit pouvoir anticiper le nombre de sinsitres en suspens qui vont être déclaré durant les années à venir. Cela revient à estimer la partie inférieure du triangle de provision.

Pour ce faire, un modèle particultièrement simple consiste à supposer que l'estimation du nombre de sinistres annoncés pendant une année est le pro-

duit de deux facteurs, le premier dépendant de l'année d'occurence du sinistre et l'autre de l'année de développement. Ce modèle part de l'hypothèse que la rapidité avec laquelle les sinistres sont annoncés est strictement la meme pour chacunes des années d'occurence. Nous allons contruire un modèle dans lequel on suppose que le développement a une composante aléatoire spécifique à chaque année d'occurence. Plus précisemment, nous allons tenter de modéliser le nombre total de sinistres N_j ($N_j = \sum_{i=1}^{l} N_{ij}$)par année d'occurence j, en supposant que le développement de chaque année de sinistres a un niveau de risque $\vec{\Theta}$ qu'il faut estimer. Bien évidemment, les sinistres déjà déclarés pour une années de sinistres seront révélateurs du risque $\vec{\Theta}$ encouru.

Dans une première partie, nous allons énoncer et détailler les propriétés des distributions multinomiales et Dirichlet dont nous aurons besoin pour les chapitres suivants.

Le troisième chapitre propose un modèle de prévision du niveau de risque $\vec{\Theta}$ en supposant N_j connu. Pour ce modèle nous allons supposer que les sinistres suivent une distribution multinomial sachant le risque $\vec{\Theta}$. D'autre part nous tenterons de résoudre le problème avec et sans un l'hypothèse que $\vec{\Theta}$ suit une distribution de Dirichlet

Finalement, le dernier chapitre propose la mise en place d'un modèle pour prédire N_j. Ce chapitre décrit le raisonnement qui nous a mené à la solution, tout en démontrant sa validité.

Chapitre 2

Les distributions

Dans ce chapitre, nous allons présenter la distribution multinomiale et la distribution de Dirichlet ainsi que quelques propriétes importantes les concernant.

2.1 La distribution multinomiale

La distribution multinomiale n'est autre que la généralisation multidimentionnelle de la distribution binomiale. Elle est définie de la manière suivante:

Considérons une série de n tirages indépendants tels que chacun de ces tirages puissent se réaliser dans un état $A_i \in \{A_1, \ldots, A_l\}$ avec une probabilité p_i, avec $0 < p_i < 1$, $i = 1, \ldots, l$, et $\sum_{i=1}^{l} p_i = 1$. Si N_i dénote le nombre d'occurences de l'état A_i lors des n tirages, alors on dit que le vecteur aléatoire $\overrightarrow{N} = (N_1, \ldots, N_l)$ possède une distribution multinomiale de paramètres $(n, p_1, \ldots, p_l)$.

La fonction de probabilité de ce vecteur aléatoire s'écrit:

$$P(N_1 = n_1, \ldots, N_l = n_l) = \frac{n!}{n_1! n_2! \cdots n_l!} \prod_{i=1}^{l} {p_i}^{n_i}$$

Voici alors différents moments de cette distribution:

$$E(N_i) = np_i$$

$$Var(N_i) = np_i(1 - p_i)$$

$$Cov(N_i, N_j) = n(-p_i p_j), \text{ pour } i \neq j$$

On peut aussi calculer la fonction génératrice des moments:

$$
\begin{aligned}
M_{\overrightarrow{N}}(\overrightarrow{t}) &= E(e^{\overrightarrow{t}.\overrightarrow{N}}) = E(e^{t_1N_1+\cdots+t_lN_l}) = E(\prod_{i=1}^{l} e^{t_iN_i}) \\
&= \sum_{n_1,\ldots,n_l t.q.\ \sum_i n_i=n} \frac{n!}{n_1!\cdots n_l!} {p_1}^{n_1}\cdots {p_l}^{n_l} \prod_{i=1}^{l} e^{t_in_i} \\
&= \sum_{n_1,\ldots,n_l t.q.\ \sum_i n_i=n} \frac{n!}{n_1!\cdots n_l!} \prod_{i=1}^{l} (p_ie^{t_i})^{n_i} \\
&= (\sum_i p_ie^{t_i})^n \sum_{n_1,\ldots,n_l t.q.\ \sum_i n_i=n} \frac{n!}{n_1!\cdots n_l!} \prod_{i=1}^{l} (\frac{p_ie^{t_i}}{\sum_i p_ie^{t_i}})^{n_i} \\
&= (\sum_i p_ie^{t_i})^n \\
&= (p_1e^{t_1}+\cdots+p_le^{t_l})^n
\end{aligned}
$$

Il nous est utile de mentionner les propriétés suivantes:

1. Les distributions marginales des variables aléatoires X_i, $i = 1,\ldots,l$ sont binomiales (n,p_i).
2. On mentionne cette propriété sous forme de proposition:

   Proposition 1 *Soit* $\overrightarrow{N} = (N_1,\ldots,N_l) \sim Multinomiale(n,p_1,\ldots,p_l)$ *et soit* $M_k = N_1+\cdots+N_k$. *Alors, on a que* $M_k \sim Binomiale(n,p_1+\cdots+p_k)$.

   Proof
   On utilise la fonction génératrice des moments calculée précédemment.

$$
\begin{aligned}
M_{M_k}(t) = E(e^{tM_k}) &= E(e^{t(N_1+\cdots+N_k)}) \\
&= E(e^{t(N_1+\cdots+N_k)+o(N_{k+1}+\cdots+N_l)}) \\
&= (p_1e^t+\cdots+p_ke^t+p_{k+1}e^0+\cdots+p_le^0)^n \\
&= ((p_1+\cdots+p_k)e^t+(p_{k+1}+\cdots+p_l)e^0)^n \\
&= (pe^t+(1-p))^n \text{ , où } p = p_1+\cdots+p_k
\end{aligned}
$$

   Ce qui est justement la fonction génératrice des moments d'une loi binomiale de paramètres (n,p).

   □

2.2 La distribution de Dirichlet

La distribution de Dirichlet est en fait elle aussi la généralisation multidimentionnelle de la distribution Bêta. Elle est définie de la manière suivante:

Un vecteur aléatoire $\overrightarrow{\Theta} = (\Theta_1,\ldots,\Theta_l)$ possède une distribution de Dirichlet de paramètres $(\alpha_1,\ldots,\alpha_l)$ lorsque chaque variable θ_i est définie par $\theta_i = Z_i/\sum_{j=1}^{l} Z_j$ avec les v.a. Z_j telles que $Z_j \sim Gamma(\alpha_j,1)$, pour j=1,...,l. On a alors que $\sum_{j=1}^{l} \Theta_j = 1$.

La fonction de densité de ce vecteur aléatoire s'écrit ainsi:

$$f(\theta_1,\ldots,\theta_l) = \frac{\Gamma(\alpha_1+\cdots+\alpha_l)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_l)} \prod_{j=1}^{l} \theta_j^{\alpha_i-1} \qquad \theta_j \geq 0$$

On peut aisément calculer différents moments de cette distribution (où $\alpha = \sum_{j=1}^{n} \alpha_j$, et $i \neq k$):

$$E(\Theta_k) = \frac{\alpha_k}{\alpha}$$

$$E(\Theta_k^2) = \frac{\alpha_k(\alpha_k+1)}{\alpha(\alpha+1)}$$

$$E(\Theta_i\Theta_k) = \frac{\alpha_k\alpha_i}{\alpha(\alpha+1)}$$

$$E(\Theta_k(1-\Theta_k)) = \frac{\alpha_k(\alpha-\alpha_k)}{\alpha(\alpha+1)}$$

$$Var(\Theta_k) = \frac{\alpha_k(\alpha-\alpha_k)}{\alpha^2(\alpha+1)}$$

$$Cov(\Theta_i,\Theta_k) = -\frac{\alpha_k\alpha_i}{\alpha^2(\alpha+1)}$$

Pour la suite, il est interessant de mentionner aussi les deux rapports suivants qui valent α:

$$\alpha = \frac{E(\Theta_k(1-\Theta_k))}{Var(\Theta_k)}$$

$$\alpha = -\frac{E(\Theta_i\Theta_k)}{Cov(\Theta_i,\Theta_k)}$$

On a alors les propriétés suivantes:

1. Lorsque l=2, on a que $(\Theta_1,\Theta_2 = 1-\Theta_1) \sim Beta(\alpha_1,\alpha_2)$

2. On mentionne cette propriété sous forme de proposition:

Proposition 2 *Si* $(\Theta_1,\ldots,\Theta_l) \sim Dirichlet(\alpha_1,\ldots,\alpha_l)$*, et si* $\gamma_1,\ldots,\gamma_l$ *sont des entiers tels que* $0 < \gamma_1 < \gamma_2 < \ldots < \gamma_n = l$*, alors on a que*

$$(\sum_{j=1}^{\gamma_1}\Theta_j, \sum_{j=\gamma_1+1}^{\gamma_2}\Theta_j, \ldots, \sum_{j=\gamma_{n-1}+1}^{\gamma_n}\Theta_j) \sim Dirichlet(\sum_{j=1}^{\gamma_1}\alpha_j, \sum_{j=\gamma_1+1}^{\gamma_2}\alpha_j, \ldots, \sum_{j=\gamma_{n-1}+1}^{\gamma_n}\alpha_j)$$

.

Proof
La démonstration découle directement des propriétés de l'additivité de la loi Gamma. cf: Johnson and Kotz (1972) Distribution in statistics. Continuous Multivariate Distributions. Wiley, New York.

□

Chapitre 3

Modèle de prédiction de $\vec{\Theta}$

3.1 Introduction

Au cours de ce chapitre nous allons proposer un modèle de prédiction des sinistres pour les triangles de développement (Run-off triangles). Ces triangles sont de la forme:

$$\begin{array}{ccccccc} N_{1,1} & N_{1,2} & N_{1,3} & & \cdots & & N_{1,l} \\ N_{2,1} & N_{2,2} & N_{2,3} & & \cdots & N_{2,l-1} & \\ N_{3,1} & N_{3,2} & N_{3,3} & \cdots & N_{3,l-2} & & \\ \vdots & \vdots & & & & & \\ N_{l-1,1} & N_{l-1,2} & & & & & \\ N_{l,1} & & & & & & \end{array}$$

On suppose que pour chaque année de sinistre le montant total des sinistres est connus (que l'on noteras N_i pour l'année de sinistre i). Cette hypothèse est peu réaliste, mais elle permet de simplifier beaucoup le modèle, dans le but de mieux cerner les difficultés de la prédiction du nombres de sinistres. L'étude d'un modèle sans supposer le nombres total des sinistres connus est développé au chapitre suivant.

$$\text{Notons que} \quad N_j = \sum_{i=1}^{l} N_{ij}, \quad \forall j$$

Notre but est de prédire le niveau de risque de l'année de développement pour une année d'occurence données (ie: on cherche le niveau de risque Θ_{kj} avec $k = 1, \ldots, l$ pour un j fixé). Les prédictions doivent se faire en fonction du nombres de sinistres déjà déclarés lors des années précédentes.

On cherche donc à calculer:

$$E(\Theta_{kj}|N_{1j}=n_{1j},N_{2j}=n_{2j},\cdots,N_{(k-1)j}=n_{(k-1)j})$$

Notre modèle ne tenant compte que d'une année de sinistre, nous allons donc, par soucis de clarté, omettre de mentionner l'année de sinistre dans les indices (ie : $N_j \equiv$ et $N_{ij} \equiv N_i$).

3.2 Le modèle multinomial - Dirichlet

Dans cette partie nous supposons que la distribution du nombre de sinistres survenues par années de développement $\vec{N}$ est connue si l'on connait le niveau de risque $\vec{\Theta}$ de l'année de sinistre. Cette distribution est supposée multinomiale de paramètres:

$$\vec{N}|\vec{\Theta} \sim Multinomial(N,\Theta_1,\Theta_2,\ldots,\Theta_k)$$

D'autre part, nous allons supposer que le niveau de risque $\vec{\Theta}$ suit une distribution de Dirichlet:

$$\vec{\Theta} \sim Dirichlet(\alpha_1,\alpha_2,\ldots,\alpha_k)$$

Les fonctions de répartition de $\vec{N}|\vec{\Theta}$ et $\vec{\Theta}$ sont nommée respectivement, $f(\vec{n};\vec{\theta})$ et $u(\vec{\theta})$.

Fort de ces hypothèses, nous allons déterminer la distribution à postériori, c'est à dire la distribution de $\vec{\Theta}|\vec{N}$.

Pour cela, on utilise la formule de Bayes qui nous donne:

$$\begin{aligned} u(\vec{\Theta}|\vec{N}) &= \frac{u(\vec{\theta})f(\vec{n};\vec{\theta})}{\int u(\vec{\theta})f(\vec{n};\vec{\theta})d\theta} \\ &\propto u(\vec{\theta})f(\vec{n};\vec{\theta}) \\ &\propto \prod_{i=1}^{l}\theta_i^{\alpha_i-1}\prod_{i=1}^{l}\theta_i^{n_1} \\ &\propto \prod_{i=1}^{l}\theta_i^{\alpha_i+n_i-1} \end{aligned}$$

On remarque que $u(\vec{\Theta}|\vec{N})$ est une fonction de répartition d'une distribution de Dirichlet de paramètres $\tilde{\alpha}_1,\tilde{\alpha}_2,\ldots,\tilde{\alpha}_l$ tels que:

$$\begin{aligned}\tilde{\alpha}_1 &= \alpha_1 + N_1,\\ \tilde{\alpha}_2 &= \alpha_2 + N_2,\\ &\dots,\\ \tilde{\alpha}_l &= \alpha_l + N_l\end{aligned}$$

On note aussi:

$$\tilde{\alpha} = \sum_{i=1}^{l} \tilde{\alpha}_i$$

On obtient donc:

$$u(\vec{\Theta}|\vec{N}) = \frac{\Gamma(\tilde{\alpha})}{\prod_{i=1}^{l}\Gamma(\tilde{\alpha}_i)}\theta_1^{\tilde{\alpha}_1-1}\dots\theta_l^{\tilde{\alpha}_l-1}$$

Cette propriété propre à la conjonction des distributions multinomiale et Dirichlet nous permet de retrouver une distributiuon à postériori suivant la loi de Dirichlet. Remarquons que nous avons choisi ces deux lois dans le but de bénéficier de cette propriété.

On peut ainsi aisément déterminer $E(\vec{\Theta}|\vec{N})$:

$$\begin{aligned}E(\vec{\theta_k}|\vec{N}) &= \frac{\tilde{\alpha}_k}{\tilde{\alpha}},\ \ \forall k = 1,2,\dots,l\\ &= \frac{\alpha_k + N_k}{\alpha + N}\end{aligned}$$

3.3 Le modèle multinomial

Maintenant, nous nous intéressons au modèle plus général, où N est toujours donné, mais pas la distribution de $\vec{\Theta}$. N'ayant pas d'hypothèse sur la distribution de $\vec{\Theta}$ on tente de linéariser chaque Θ_k en posant:

$$\hat{\Theta}_k = a_0 + a_1N_1 + a_2N_2 + \dots + a_kN_k$$

Le but est alors d'évaluer les a_i de façon à trouver un $\hat{\Theta}_k$ le plus réaliste possible. Pour cela, on utilise le critère qui consiste à minimiser $Q = E[(\hat{\Theta}_k - \Theta_k)^2]$:

$$\begin{aligned}Q &= E[(\hat{\Theta}_k - \Theta_k)^2]\\ &= E[(a_0 + a_1N_1 + a_2N_2 + \dots + a_kN_k - \Theta_k)^2]\\ &= E[\sum_{j=0}^{k} a_jN_j - \Theta_k)^2]\end{aligned}$$

Tout d'abord nous déterminons a_0:

$$
\begin{aligned}
& \frac{\delta E}{\delta a_0} = 0 \\
\Leftrightarrow \quad & E[2(a_0 + a_1 N_1 + a_2 N_2 + \cdots + a_k N_k - \Theta_k)] = 0 \\
\Leftrightarrow \quad & a_0 + E[\textstyle\sum_{j=1}^{k} a_j N_j] - E[\Theta_k] = 0 \\
\Leftrightarrow \quad & a_0 + \textstyle\sum_{j=1}^{k} a_j E[N_j] - E[\Theta_k] = 0 \\
\Leftrightarrow \quad & a_0 = E[\Theta_k] - \textstyle\sum_{j=1}^{k} a_j E[N_j]
\end{aligned}
$$

On peut donc remplacer a_0 dans l'équation de départ par l'expression trouvée et on obtient:

$$
\begin{aligned}
Q &= E[(\hat{\Theta}_k - \Theta_k)^2] \\
&= E[(a_0 + E[\sum_{j=1}^{k} a_j N_j] - \Theta_k)^2] \\
&= E[(E[\Theta_k] - \sum_{j=1}^{k} a_j E[N_j] + E[\sum_{j=1}^{k} a_j N_j] - \Theta_k)^2] \\
&= E[(\sum_{j=1}^{k} a_j (N_j - E[N_j]) - (\Theta_k - E[\Theta_k]))^2]
\end{aligned}
$$

Maintenant nous dérivons par rapport à a_i:

$$
\begin{aligned}
& \frac{\delta E}{\delta a_i} = 0 \\
\Leftrightarrow \quad & E[2(N_i - E[N_i])(\textstyle\sum_{j=1}^{k} a_j (N_j - E[N_j]) - (\Theta_k - E[\Theta_k])] = 0 \\
\Leftrightarrow \quad & \textstyle\sum_{j=1}^{k} a_j cov(N_j, N_i) = cov(\Theta_k, N_i) \\
\Leftrightarrow \quad & cov(\hat{\Theta}_k, N_k) = cov(\Theta_k, N_i)
\end{aligned}
$$

On doit donc résoudre le système d'équation suivant pour déterminer les a_i propres à chaque Θ_k:

$$
(3.1) \begin{cases}
\sum_{j=1}^{k} a_j cov(N_j, N_1) = cov(\Theta_k, N_1) \\
\sum_{j=1}^{k} a_j cov(N_j, N_2) = cov(\Theta_k, N_2) \\
\vdots \qquad\qquad\qquad\quad \vdots \\
\sum_{j=1}^{k} a_j cov(N_j, N_k) = cov(\Theta_k, N_k)
\end{cases}
$$

Ceci revient donc à résoudre les équations normales:

$$\begin{aligned} E[\hat{\Theta}_k] &= E[\Theta_k] \\ cov(\hat{\Theta}_k,N_i) &= cov(\Theta_k,N_i) \quad \forall i = 1,\ldots,k \end{aligned}$$

Notons que l'on peut encore développé les termes $cov(N_j,N_i)$ et $cov(\Theta_k,N_i)$, cela pour tout i=1,...,k. Le fait que $N|\vec{\Theta}$ suive une loi multinomiale nous permet de dire que l'on a l'égalité:

$$\begin{aligned} cov(N_j,N_i) &= E[cov(N_j,N_i|\vec{\Theta})] + cov(E[N_j|\vec{\Theta}],E[N_i|\vec{\Theta}]) \\ &= E[N\Theta_i(\delta_{ij} - \Theta_j)] + cov(N\Theta_j,N\Theta_i) \\ &= NE[\Theta_i(\delta_{ij} - \Theta_j)] + N^2 cov(\Theta_j,\Theta_i) \end{aligned}$$

De meme que:

$$\begin{aligned} cov(\Theta_k,N_i) &= E[cov(\Theta_k,N_i|\vec{\Theta})] + cov(E[\Theta_k|\vec{\Theta}],E[N_i|\vec{\Theta}]) \\ &= 0 + cov(\Theta_k,N\Theta_i) \\ &= Ncov(\Theta_k,\Theta_i) \end{aligned}$$

En remplaçant les résultats obtenus dans les système (3.1), et en simplifiant par N, on obtient:

$$(3.2) \begin{cases} \sum_{j=1}^k a_j\{E[\Theta_i(\delta_{ij} - \Theta_j)] + Ncov(\Theta_j,\Theta_1)\} &= cov(\Theta_k,\Theta_1) \\ \sum_{j=1}^k a_j\{E[\Theta_i(\delta_{ij} - \Theta_j)] + Ncov(\Theta_j,\Theta_2)\} &= cov(\Theta_k,\Theta_2) \\ \qquad\qquad\qquad\qquad \vdots & \qquad\qquad \vdots \\ \sum_{j=1}^k a_j\{E[\Theta_i(\delta_{ij} - \Theta_j)] + Ncov(\Theta_j,\Theta_k)\} &= cov(\Theta_k,\Theta_k) \end{cases}$$

On obtient donc le système (3.2) de k équations à k inconnues. On sait qu'il existe une soultion unique si la matrice de covariance de $\vec{N}$ est régulière. Mais il reste à trouver la solution. Pour cela nous allons nous intéresser à un système plus simple. Dans le cas particulier oú $\vec{\Theta}$ a une distribution de Dirichlet, l'espérance de Θ_k donné $\vec{N}$ est linéaire, c'est donc également la regression de Θ_k sur $\vec{N}$.

3.3.1 Regression de Θ_k sur N_k

On pose le problème simplifié suivant, correspondant à la solution dans le cas Dirichlet:

$$\hat{\Theta}_k = b_k + c_k N_K$$

En procédant comme dans le cas général on obtient les deux valeurs suivantes pour b_k et c_k:

$$b_k = E[\Theta_k](1 - c_k N)$$

$$\begin{aligned} & c_k var(N_k) = cov(\Theta_k, N_k) \\ \Leftrightarrow \quad & c_k = \frac{cov(\Theta_k, N_k)}{var(N_k)} \\ \Leftrightarrow \quad & c_k = \frac{var(\Theta_k)}{N var(\Theta_k) + E[\Theta_k(1-\Theta_k)]} \end{aligned}$$

On pose:

$$e_k = \frac{E[\Theta_k(1 - \Theta_k)]}{var(\Theta_k)}$$

On a alors:

$$\begin{aligned} c_k &= \frac{1}{N + e_k} \\ b_k &= \frac{e_k}{N + e_k} E[\Theta_k] \\ \Rightarrow \hat{\Theta}_k &= \frac{N_k + e_k E[\Theta_k]}{N + e_k} \\ &= \frac{N}{N + e_k} \frac{N_K}{N} + \frac{e_k}{N + e_k} E[\Theta_k] \end{aligned}$$

On obtient donc une expression analytique de Θ_k, ce qui signifie que l'on a uniquement déterminé Θ_k dans le cas particulier oú il n'y a qu'un seul N_k connu (ie: $\hat{\Theta}_k = b_k + c_k N_k$).

Si l'on essaye de se ramener au cas précédent oú Θ_k suit une distribution de Dirichlet on obtient:

$$e_k = \alpha$$

par une des propriétés de la loi de Dirichlet que l'on a vu au cours du deuxième chapitre. Finalement on trouve:

$$\begin{aligned}\hat{\Theta}_k &= \frac{N_k + \alpha E[\Theta_k]}{N+\alpha} \\ &= \frac{N}{N+\alpha}\frac{N_k}{N} + \frac{\alpha}{N+\alpha}E[\Theta_k]\end{aligned}$$

Remarque: en général e_k dépend de k. C'est justement le fait que, le rapport exprimé par e_k soit constant, qui rend la distribution de Dirichlet si particulière. Par exemple si l'on cesse de supposer que Θ_k suit une distribution de Dirichlet on a :

$$\Theta_1 \equiv 1/2 \;\Rightarrow e_1 = +\infty \;\; et \;\; c_k = 0$$

3.3.2 Regression de Θ_k sur N_1,N_2,...,N_k

Revenons à nouveau au cas général. On a vu que pour une regression linéaire avec une seul N_k donné on arrive à obtenir une solution. Maintenant, faisons la regression linéaire sur N_1,N_2,...,N_k:

$$\hat{\Theta}_k = a_{k0} + \sum_{j=1}^{k} a_{kj} N_j$$

on a le système suivant à résoudre:

$$(3.3)\begin{cases} \sum_{j=1}^{k} a_{kj} cov(N_j,N_1) &= cov(\Theta_k,N_1) \\ \sum_{j=1}^{k} a_{kj} cov(N_j,N_2) &= cov(\Theta_k,N_2) \\ \vdots & \vdots \\ \sum_{j=1}^{k} a_{kj} cov(N_j,N_k) &= cov(\Theta_k,N_k) \end{cases}$$

La question que l'on se pose est de savoir si la solution de cette regression est identique à la précédente. Ce qui voudrait dire que:

$$\begin{cases} a_{kj} = 0 & \forall j = 1,\cdots,l \;\; j \neq k \\ a_{kk} = c_k & \\ a_{k0} = b_k & \end{cases}$$

Est-ce une solution du système (3.3)? C'est ce que l'on va tenter de prouver.

Pour la k^{eme} équation du système, pas de problème, l'égalité est vérifée. En effet:

$$\begin{aligned} & a_{kk}cov(N_k,N_k) = cov(\Theta_k,N_k) \\ \Leftrightarrow \quad & c_k var(N_k) = cov(\Theta_k,N_k) \\ \Leftrightarrow \quad & c_k = \frac{cov(\Theta_k,N_k)}{var(N_k)} \end{aligned}$$

Ce qui est la définition de c_k.
Voyons ce que l'on obtient pour $i \neq k$:

$$\begin{aligned} & a_{kk}cov(N_k,N_i) = cov(\Theta_k,N_i) \\ \Leftrightarrow \quad & a_{kk} = \frac{Ncov(\Theta_k,\Theta_i)}{N^2 cov(\Theta_k,\Theta_i) - NE[\Theta_k\Theta_i]} \\ \Leftrightarrow \quad & a_{kk} = \frac{1}{N+fik} \end{aligned}$$

Si les a_{kj} posés sont justes on doit avoir:

$$f_{ik} = \frac{-E[\Theta_k(1-\Theta_k)]}{cov(\Theta_k,\Theta_i)} = e_k$$

Malheureusement ce n'est pas le cas. En effet, supposons que pour un i Θ_i soit déterminé, on a alors un contre exemple:

$$\Theta_i \text{ déterminé } \Rightarrow f_{ik} = +\infty,\ a_{kk} = 0$$

Par contre, si l'on se penche à nouveau sur le cas Dirichlet on a par une des propriétés enoncée dans le deuxième chapitre:

$$\frac{-E[\Theta_k(1-\Theta_k)]}{cov(\Theta_k,\Theta_i)} = \alpha = \frac{E[\Theta_k(1-\Theta_k)]}{var(\Theta_k)}$$

Ce qui confirme le fait que cette propriété n'est pas vérifiée lorsque l'on ne suppose pas que $\vec{\Theta}$ suit une distribution de Dirichlet.

3.3.3 Remarque

Au cours de ce chapitre on a déterminé une solution de Θ_k pour le cas simple d'une régression sur N_k, puis on a supposé que la solution pour une regression de Θ_k sur $N_1,N_2,\ldots,N_k$ ressemble fortement à celle trouvée pour le cas simple. Mais ce n'est pas les cas, loin de là. Pourtant lorsque l'on suppose que $\vec{\Theta}$ suit une loi de Dirichlet, la solution du cas simple est vérifée pour une regression sur $N_1,N_2,\ldots,N_k$.

L'hypothèse $\vec{\Theta}$ suit une distribution de Dirichlet nous offre la possibilité de trouver une solution au problème de la regression de Θ_k sur N_1,N_2,...,N_k. Sans cette hypothèse le problème devient beaucoup plus difficile à résoudre. Pour la suite de notre projet nous allons supposer que $\vec{\Theta}$ suit une distribution de Dirichlet.

La recherche d'une solution sans l'hypothèse Dirichlet pourrait faire l'objet d'une étude différente que celle qui est faite pour le chapitre suivant.

Chapitre 4

Prédiction de N avec le modèle multinomial-Dirichlet

4.1 Énoncé du modèle

Notre but est de donner une estimation $\hat{N}_j$ de chaque $N_j := \sum_{i=1}^{l} N_{ij}$ en utilisant les données que l'on a à disposition. Le présent modèle vise donc à obtenir une telle estimation. Les hypothèses pour ce modèle sont les suivantes:

- Pour chaque année d'occurence j,
  1. Étant donnés N_j et $\vec{\Theta}_j = \Theta_{1j}, \dots \Theta_{lj}$, on a que $(N_{1j}, \dots, N_{lj}) \sim multinomiale(N_j, \vec{\Theta}_j)$.
  2. N_j et $\vec{\Theta}_j$ sont indépendants.
- Pour l'ensemble des années d'occurence,
  1. les vecteurs aléatoires $\vec{\Theta}_1, \dots, \vec{\Theta}_l$ sont i.i.d et suivent chacun une distribution de Dirichlet de paramètres $(\alpha_1, \dots, \alpha_l)$.
  2. Les variables aléatoires N_j sont indépendantes et possèdent une distribution de Poisson de paramètre $V_j\lambda$, où V_j est un volume connu. On a alors que $E(N_j) = var(N_j) = V_j\lambda$.

On travaille désormais pour une colonne j déterminée et on note par $k := l - j + 1$ le nombre d'éléments de cette colonne. Sachant ceci, on n'écrira plus l'indice j par la suite (N_j sera donc noté par N et les données N_{ij} seront notées par N_i).

Nous avons choisi de chercher un estimateur linéaire en les données que nous possédons, c'est à dire un estimateur de la forme suivante:

$$\hat{N} = a_0 + \sum_{i=1}^{k} a_i N_i$$

et on impose que cet estimateur optimise le critère suivant:

$$\text{minimiser } E[(\hat{N} - N)^2]$$

Le but est donc de trouver l'expression des a_i en fonction des moments de la distribution de N et en fonction des paramètres α_i (qui définissent la distribution du "vecteur de risque" $\vec{\Theta}$) pour que la condition d'optimalité imposée soit remplie. En d'autre termes, nous faisons une régression linéaire de N sur $N_1, \ldots, N_k$.

Le problème est donc le suivant:

$$\begin{gathered}\text{chercher } \hat{N} = a_0 + \sum_{i=1}^{k} a_i N_i \text{ t.q.} \\ E[(\hat{N} - N)^2] \text{ soit minimale}\end{gathered}$$

4.2 Établissement des équations

Nous avons vu dans le chapitre précédent que ceci revient à résoudre les équations normales suivantes:

$$E(\hat{N}) = E(N) \tag{4.1}$$

$$cov(\hat{N},N_i) = cov(N,N_i) \quad \text{pour } i = 1,\ldots,k \tag{4.2}$$

En développant l'équation (4.1) on obtient:

$$\begin{aligned}
& E(\hat{N}) = E(N) \\
\Leftrightarrow \quad & E(a_0 + \sum_{i=1}^{k} a_i N_i) = E(N) \\
\Leftrightarrow \quad & a_0 = E(N) - \sum_{i=1}^{k} a_i E(N_i) \\
\Leftrightarrow \quad & a_0 = E(N) - \sum_{i=1}^{k} a_i E(E(N_i|\vec{\Theta})) \\
\Leftrightarrow \quad & a_0 = E(N) - \sum_{i=1}^{k} a_i E(N\Theta_i)
\end{aligned}$$

$$
\begin{aligned}
&\Leftrightarrow \quad a_0 = E(N) - \sum_{i=1}^{k} a_i E(N) E(\Theta_i) \\
&\Leftrightarrow \quad a_0 = E(N)(1 - \sum_{i=1}^{k} a_i E(\Theta_i)) \\
&\Leftrightarrow \quad a_0 = E(N)(1 - \sum_{i=1}^{k} a_i \frac{\alpha_i}{\alpha})
\end{aligned}
$$

Développons maintenant l'équation (4.2):

$$
\begin{aligned}
& cov(\hat{N},N_i) = cov(N,N_i) \quad \text{pour } i = 1,\ldots,k \\
\Leftrightarrow \quad & cov(a_0 + \sum_{j=1}^{k} a_j N_j, N_i) = cov(N,N_i) \quad \text{pour } i = 1,\ldots,k \\
\Leftrightarrow \quad & \sum_{j=1}^{k} a_j cov(N_j,N_i) = cov(N,N_i) \quad \text{pour } i = 1,\ldots,k
\end{aligned}
$$

Pour le terme de gauche, on a:

$$
\begin{aligned}
cov(N_j,N_i) &= cov(E(N_j|\vec{\Theta}),E(N_i|\vec{\Theta})) + E[cov(N_j,N_i|\vec{\Theta})] \\
&= E(N^2\Theta_j\Theta_i) - E(N\Theta_j)E(N\Theta_i) + E[N(-\Theta_j\Theta_i + \delta_{ij}\Theta_i)] \\
&= E(N^2)E(\Theta_j\Theta_i) - E(N)^2E(\Theta_j)E(\Theta_i) - E(N)[E(\Theta_j\Theta_i) - \delta_{ij}E(\Theta_i)] \\
&= \begin{cases} E(N^2)\frac{\alpha_i\alpha_j}{\alpha(\alpha+1)} - E(N)^2\frac{\alpha_i\alpha_j}{\alpha^2} - E(N)\frac{\alpha_i\alpha_j}{\alpha(\alpha+1)} & \text{pour } i \neq j \\ E(N^2)\frac{\alpha_i(\alpha_i+1)}{\alpha(\alpha+1)} - E(N)^2\frac{\alpha_i^2}{\alpha^2} - E(N)(\frac{\alpha_i(\alpha_i+1)}{\alpha(\alpha+1)} - \frac{\alpha_i}{\alpha}) & \text{pour } i = j \end{cases}
\end{aligned}
$$

Pour le terme de droite, on a:

$$
\begin{aligned}
cov(N,N_i) &= cov(E(N|\vec{\Theta}),E(N_i|\vec{\Theta})) + E[cov(N,N_i|\vec{\Theta})] \\
&= cov(N,N\Theta_i) \\
&= E(N^2\Theta_i) - E(N)E(N\Theta_i) \\
&= E(N^2)E(\Theta_i) - E(N)^2E(\Theta_i) \\
&= E(\Theta_i)var(N) \\
&= \frac{\alpha_i}{\alpha}var(N)
\end{aligned}
$$

L'équation (4.2) se résume alors au grand système matriciel suivant:

$$\begin{pmatrix} \text{bloc 1} & & \vdots & \\ & \ddots & \text{bloc 2} & \dots \\ \dots & \text{bloc 2} & \ddots & \\ & \vdots & & \text{bloc 1} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix} = \begin{pmatrix} \frac{\alpha_1}{\alpha} var(N) \\ \frac{\alpha_2}{\alpha} var(N) \\ \vdots \\ \frac{\alpha_k}{\alpha} var(N) \end{pmatrix}$$

où les éléments du bloc 1 sont du type:

$$E(N^2)\frac{\alpha_i(\alpha_i+1)}{\alpha(\alpha+1)} - E(N)^2\frac{\alpha_i^2}{\alpha^2} - E(N)(\frac{\alpha_i(\alpha_i+1)}{\alpha(\alpha+1)} - \frac{\alpha_i}{\alpha})$$

et les éléments du bloc 2 sont du type:

$$E(N^2)\frac{\alpha_i\alpha_j}{\alpha(\alpha+1)} - E(N)^2\frac{\alpha_i\alpha_j}{\alpha^2} - E(N)\frac{\alpha_i\alpha_j}{\alpha(\alpha+1)}$$

avec les indices i et j respectivement adaptés aux indices de la ligne et de la colonne auxquelles se trouve le terme.

Et l'équation (4.1) nous donne la condition permettant de déduire a_0 à partir des a_i:

$$a_0 = E(N)(1 - \sum_{i=1}^{k} a_i\frac{\alpha_i}{\alpha})$$

4.3 Solution du problème

Il nous faut maintenant résoudre ce système. Malheureusement l'informatique ne nous donne pas de solution satisfaisante pour un tel problème. Nous allons donc traiter des cas "limites", afin d'essayer de deviner la solution. On verra que cette approche intuitive nous permettra effectivement de trouver la solution du problème puisque nous finirons par démontrer qu'on a bien obtenu la bonne solution.

1. Le premier cas considéré est le cas de la l-ième colonne. Cette colonne ne possède qu'une seule donnée N_1, c'est pourquoi le problème se résume à trouver un estmateur $\hat{N}$ de la forme $\hat{N} = a_0 + a_1N_1$. Dans ce cas, le système d'équations et la condition pour a_0 découlant des équation normales s'écrivent:

$$\begin{cases} a_1(E(N^2)\frac{\alpha_1(\alpha_1+1)}{\alpha(\alpha+1)} - E(N)^2\frac{\alpha_1^2}{\alpha^2} - E(N)(\frac{\alpha_1(\alpha_1+1)}{\alpha(\alpha+1)} - \frac{\alpha_1}{\alpha}) = \frac{\alpha_1}{\alpha}var(N) \\ \\ a_0 = E(N)(1 - a_1\frac{\alpha_1}{\alpha}) \end{cases}$$

c'est à dire:

$$\begin{aligned} a_1 &= \frac{var(N)}{E(N^2)\frac{\alpha_1+1}{\alpha+1} - E(N)^2\frac{\alpha_1}{\alpha} + E(N)\frac{\alpha-\alpha_1}{\alpha+1}} \\ a_0 &= E(N)(1 - a_1\frac{\alpha_1}{\alpha}) \end{aligned}$$

Dans ce cas, la solution est donc parfaitement déterminée.

2. Le deuxième cas considéré est celui de la première colonne, c'est à dire le cas où on possède toutes les observations necéssaires. On cherche donc maintenant un estimateur de la forme $\hat{N} = a_0 + \sum_{i=1}^{l} a_i N_i$. Mais dans ce cas, on sait que l'estimateur que nous devons obtenir est $\hat{N} = \sum_{i=1}^{l} N_i$, puisque l'on cherche à estimer la somme des données et que nous les possédons toutes.

En tenant compte de ces deux cas spéciaux, on peut poser (pour une colonne j quelconque possdant k éléments) une solution de la forme:

$$\begin{aligned} a_i &= \frac{var(N)}{E(N^2)\frac{\sum\alpha_m+1}{\alpha+1} - E(N)^2\frac{\sum\alpha_m}{\alpha} + E(N)\frac{\alpha-\sum\alpha_m}{\alpha+1}} \\ a_0 &= E(N)(1 - \sum_{i=1}^{k} a_i\frac{\alpha_i}{\alpha}) \quad \text{pour } m \text{ variant de 1 à k et pour } i = 1,\ldots,k \end{aligned}$$

On retrouve bien alors la solutions de chacun des deux cas spéciaux lorsqu'on remplace k par 1 et l respectivement. Ce qui est étonnant de remarquer est que l'on "obtient" (ou que l'on suppose simplement pour le moment) une solution avec les a_i tous égaux!

4.4 Démonstration de la validité de notre solution

Il nous faut bien entendu encore démontrer que la solution que nous avons énoncée est correcte. Pour cette démonstration, nous avons besoin de deux résultats que nous fournissent les propositions 1 et 2. voici ces deux rsultats:

1. Si $(N_1,\ldots,N_l|\vec{\Theta}) \sim multinomiale(\Theta_1,\ldots,\Theta_l,N)$, alors $(\sum_{i=1}^{k} N_i, \sum_{i=k+1}^{l} N_i) \sim multinomiale(\sum_{i=1}^{k}\Theta_i, \sum_{i=k+1}^{l}\Theta_i, N)$. Ce résultat est fourni par la proposition 1.

2. Si $(\Theta_1,\ldots,\Theta_l) \sim Dirichlet(\alpha_1,\ldots,\alpha_l)$, alors $(\sum_{i=1}^{k}\Theta_i, \sum_{i=k+1}^{l}\Theta_i) \sim Dirichlet(\sum_{i=1}^{k}\alpha_i, \sum_{i=k+1}^{l}\alpha_i)$. Ce résultat est fourni par la proposition 2.

Il nous faut d'abord démontrer que les a_i sont bien tous égaux, comme ce que l'on avait supposé. Pour cela, on a besoin du corollaire suivant:

Corollaire 1 *Si, étant donné $N = \sum_{i=1}^{l} N_i$, on a que $N_1,\ldots,N_l$ est mutinomial-Dirichlet de paramètres $(\alpha_1,\ldots,\alpha_l)$, alors, étant donné $M_k = \sum_{i=1}^{k} N_i$, on a que $N_1,\ldots,N_k$ est mutinomial-Dirichlet de paramètres $(\alpha_1,\ldots,\alpha_k)$.*

Proof

Donné $N = n$:

- Étant donné M_k et $\vec{\Theta}$, alors on a que $(N_1,\ldots,N_k)$ suit une distribution multinomiale de paramètres $(M_k, \frac{\Theta_1}{\tilde{\Theta}_k},\ldots,\frac{\Theta_k}{\tilde{\Theta}_k})$, où $\tilde{\Theta}_k = \sum_{j=1}^{k}\Theta_j$. En effet, on a:

$$\begin{aligned}
& Pr[N_1 = x_1,\ldots,N_k = x_k|\tilde{\Theta}_k = m,\vec{\Theta}] \\
= & \frac{Pr[N_1 = x_1,\ldots,N_k = x_k|\vec{\Theta}]}{Pr[\tilde{\Theta}_k = m|\vec{\Theta}]} \quad \text{avec } x_1 + \ldots + x_K = m \\
= & \frac{\frac{n!}{x_1!\cdots x_k!(n-m)!}\Theta_1^{x_1}\cdots\Theta_k^{x_k}(1-\tilde{\Theta}_k)^{n-m}}{\frac{n!}{m!(n-m)!}\tilde{\Theta}_k^{m}(1-\tilde{\Theta}_k)^{n-m}} \\
= & \frac{m!}{x_1!\cdots x_k!}(\frac{\Theta_1}{\tilde{\Theta}_k})^{x_1}\cdots(\frac{\Theta_k}{\tilde{\Theta}_k})^{x_k}
\end{aligned}$$

  C'est bien ce que l'on voulait montrer.

- Étant donné $\tilde{\Theta}_k = y$, on a que $(\frac{\Theta_1}{\tilde{\Theta}_k}\cdots\frac{\Theta_k}{\tilde{\Theta}_k})$ suit une loi de Dirichlet de paramètres $(\alpha_1,\ldots,\alpha_k)$. En effet, soient $x_1,\ldots,x_k$ tels que $x_1+\ldots+x_k = 1$, alors on a:

$$\begin{aligned}
& dP(\frac{\Theta_1}{\tilde{\Theta}_k} = x_1,\ldots,\frac{\Theta_k}{\tilde{\Theta}_k} = x_k|\tilde{\Theta}_k = y) \\
= & \frac{dP(\Theta_1 = x_1 y,\ldots,\Theta_k = x_k y)}{dP(\tilde{\Theta}_k = y)} \\
= & \frac{\frac{\Gamma(\alpha)y^{k-1}}{\prod_{j=1}^{k}\Gamma(\alpha_j)\Gamma(\alpha-\tilde{\alpha}_k)}(x_1 y)^{\alpha_1-1}\ldots(x_k y)^{\alpha_k-1}(1-y)^{\alpha-\tilde{\alpha}_k-1}}{\frac{\Gamma(\alpha)}{\Gamma(\tilde{\alpha}_k)\Gamma(\alpha-\tilde{\alpha}_k)}y^{\tilde{\alpha}_k-1}(1-y)^{\alpha-\tilde{\alpha}_k-1}} \\
= & \frac{\Gamma(\tilde{\alpha}_k)}{\prod_{j=1}^{k}\Gamma(\alpha_j)}x_1^{\alpha_1-1}\cdots x_k^{\alpha_k-1}
\end{aligned}$$

  Cette expression ne dépend plus de y, et c'est bien la fonction de densité d'une loi Dirichlet de paramètres $(\alpha_1\ldots\alpha_k)$.

□

On va maintenant supposer que les a_j sont tous égaux ($a_1 = \ldots = a_k = a$), et voir que nous obtenons dans ce cas une solution pour a qui ne dépend justement pas de j. On pourra donc conclure que cette solution satisfait bien notre problème, et comme notre problème possède une solution unique, on pourra alors conclure que notre solution est la bonne.

On part de l'équation (4.2):

$$\begin{aligned}
& \sum_{j=1}^{k} a_j cov(N_j,N_i) = cov(N,N_i) \quad \text{pour } i = 1,\ldots,k \\
\Leftrightarrow \quad & a \sum_{j=1}^{k} cov(N_j,N_i) = cov(N,N_i) \quad \text{pour } i = 1,\ldots,k \\
\Leftrightarrow \quad & a \cdot cov(\sum_{j=1}^{k} N_j,N_i) = cov(N,N_i) \quad \text{pour } i = 1,\ldots,k \\
\Leftrightarrow \quad & a \cdot cov(M_k,N_i) = cov(N,N_i) \quad \text{pour } i = 1,\ldots,k \\
\Leftrightarrow \quad & a = \frac{cov(N,N_i)}{cov(M_k,N_i)} \\
& \text{où } M_k = \sum_{j=1}^{k} N_j
\end{aligned}$$

On avait déjà calculé:

$$cov(N,N_i) = \frac{\alpha_i}{\alpha} var(N)$$

et donc, par le corollaire 1:

$$cov(M_k,N_i) = \frac{\alpha_i}{\alpha^*} var(M_k) \quad \text{où } \alpha^* = \sum_{j=1}^{k} \alpha_j$$

On a alors finalement:

$$a = \frac{cov(N,N_i)}{cov(M_k,N_i)} = \frac{\frac{\alpha_i}{\alpha} var(N)}{\frac{\alpha_i}{\alpha^*} var(M_k)} = \frac{\frac{\alpha^*}{\alpha} var(N)}{var(M_k)}$$

et (en utilisant les résultats du début de section):

$$var(M_k) \quad = \quad var(E(M_k|\vec{\Theta})) + E(var(M_k|\vec{\Theta}))$$

$$
\begin{aligned}
&= var(N\sum_{j=1}^{k}\Theta_j) + E[N\sum_{j=1}^{k}\Theta_j(1-\sum_{j=1}^{k}\Theta_j)] \\
&= E(N^2(\sum_{j=1}^{k}\Theta_j)^2) - [E(N\sum_{j=1}^{k}\Theta_j)]^2 + E(N)E[\sum_{j=1}^{k}\Theta_j(1-\sum_{j=1}^{k}\Theta_j)] \\
&= E(N^2)E((\sum_{j=1}^{k}\Theta_j)^2) - E(N)^2(E(\sum_{j=1}^{k}\Theta_j))^2 + E(N)E[\sum_{j=1}^{k}\Theta_j(1-\sum_{j=1}^{k}\Theta_j)] \\
&= E(N^2)\frac{\alpha^*(\alpha^*+1)}{\alpha} - E(N)^2(\frac{\alpha^*}{\alpha})^2 + E(N)\frac{\alpha^*(\alpha-\alpha^*)}{\alpha(\alpha+1)}
\end{aligned}
$$

donc

$$
\begin{aligned}
a &= \frac{\frac{\alpha^*}{\alpha}var(N)}{var(M_k)} \\
&= \frac{\frac{\alpha^*}{\alpha}var(N)}{E(N^2)\frac{\alpha^*(\alpha^*+1)}{\alpha} - E(N)^2(\frac{\alpha^*}{\alpha})^2 + E(N)\frac{\alpha^*(\alpha-\alpha^*)}{\alpha(\alpha+1)}}
\end{aligned}
$$

□

Pour résumer, dans le cas d'un triangle de développement, le résultat est donc le suivant:

$$
\begin{aligned}
\hat{N}_j &= a_{j0} + a_{j1}\tilde{N}_{l+1-j} \\
\text{avec} \quad & \tilde{N}_{l+1-j} = N_{1j} + \ldots + N_{l+1-j,j} \\
& a_{j0} = E(N)(1 - a_{j1}\sum_{i=1}^{l+1-j}\frac{\alpha_i}{\alpha}) \\
& a_{j1} = \frac{var(N)}{E(N^2)\frac{\sum\alpha_m+1}{\alpha+1} - E(N)^2\frac{\sum\alpha_m}{\alpha} + E(N)\frac{\alpha-\sum\alpha_m}{\alpha+1}} \\
& \text{pour } m \text{ variant de 1 à } l+1-j
\end{aligned}
$$

Chapitre 5

Conclusion

Le dernier chapitre nous démontre une formule de prédiction de N avec le modèle multinomial - Dirichlet. Nous avons donc réussi à aboutir au résultat recherché au départ de notre projet.

Pour cela nous avons commencé par décrire les deux principales distribution dont nous avons besoin, c'est à dire la distribution mulitnomial et la distribution de Dirichlet.

Puis, dans le troisième chapitre, nous avons mis en place un modèle de prédiction de $\vec{\Theta}$ avec N donné. Au cours de chapitre, nous nous sommes rendu compte que l'hypothèse que $\vec{\Theta}$ suit une distribution de Dirichlet nous procure des propriétés qui nous permettent d'expliciter un modèle de prévision de $\vec{\Theta}$. En effet, nous avons remarqué que sans cette hypothèse, le modèle est beaucoup plus difficle à calculer.

Dans le quatrième chapitre, nous avons développé et démontré un modèle de prédictione de N. Les conclusions du chapitre précédent sur l'hypothèse Dirichlet, nous ont conduit à nous intéresser au cas mutlinomial - Dirichlet exclusivement. Le modèle que l'on a trouvé possède donc des propriétés qui découlent directement du 'bon comportement' de la distribution de Dirichlet.

Dans le cadre d'une étude moins théorique du problème, il serait intéressant de disposer d'un triangle de développement composé de données observées par un assureur. On pourrait ainsi mettre en place un modèle statistique d'estimation des paramètres α_i et λ de notre modèle de prédiction de N. L'évaluation de ces paramètres nous permettrait de tester la validité de notre modèle de prédiction.