Rapport de stage

Mesure d'audience croisée multi-media : Mesure du taux de visionnage d'une publicité étendue sur plusieurs appareils

Auteur : Ghiles Kemiche

Nom de l'organisme d'accueil : Fondation du risque
Maître De Stage : Pr. Guillaume Lecué

Licence double diplôme Mathématiques - Informatique

2021 - 2022

Remerciements

Je tiens tout d'abord à adresser mes sincères remerciements à mon maitre de stage, Guillaume Lecué, professeur à l'ENSAE, pour avoir accepté de m'accorder sa confiance et de me compter parmi ses stagiaires. Tout au long du stage, il a toujours fait preuve d'un professionalisme irréprochable, accompagné d'une expertise dont découlait de précieux conseils, le tout dans un environnement saint et chaleureux .

Je tiens aussi à remercier le Professeur Nicolas Chopin pour ses disponibilités et le partage de son savoir faire durant ce stage.

(G)

Table des matières

Table des figures

1

Introduction

En 2013, la WFA (Fédération mondiale des annonceurs) a exprimé le besoin des grands annonceurs d'une mesure d'efficacité des campagnes de publicités multi-médiaet multisupport, uniformisée au niveau mondial. En effet, le problème posé est le suivant : un utilisateur peut visionner cette publicité sur plusieurs devices, ce qui fait que la mesure d'audience est souvent surestimée .

Google a immédiatement proposé une approche sur 3 papiers (KSML16) (KSV13) (SK19). : Cependant, l'approche de google pose quelques problèmes :
Tout d'abord, Google est à la fois sur le marché de mesure et sur le marché des annonceurs, il deviendrait alors juge et partie de ses propres compagnes de publicités, ce qui pose un problème déontologique. Ensuite, ces trois papiers ne sont pas faciles à lire en raison de beaucoup d'erreurs de notations et de typos, de plus, aucun de ces papiers n'a été publié. Enfin, Google a supposé le fait qu'un utilisateur soit atteint sur un device est indépendant du fait qu'il soit atteint sur un autre device durant la même compagne publicitaire, ce qui est peu realiste et non justifié.

1.1 Notations et définitions

- **Reach :** Nombre de foyers ou personnes d'une strate exposés au moins une fois à une publicité lors d'une campagne publicitaire.
- **Frequency :** Nombre moyen d'expositions à une publicité reçues par les individus d'une strate de population ayant été atteint au moins une fois lors d'une campagne.
- **Impression :** On parle d'une impression quand une publicité est apparue dans un encart publicitaire.

- **Editeur :** Un endroit où on peut mettre des publicités et potentiellement connaître des informations sur l'utilisateur qui les a vues (sur internet ça sera un site).
- **Annonceur :** Une marque qui achète les espaces publicitaires.
- **PPD (Publisher Provided data) :** Ce sont les données sur les utilisateurs recueillies par les éditeurs.
- **WFA :** World Federation of Advertisers
- **CESP (Centre d'étude des Supports de Publicité) :** Le CESP est l'organisme interprofessionnel des acteurs de la communication concernés par l'étude de l'audience des médias et la mesure de leur efficacité : annonceurs, agences et médias. Le CESP audite toutes les mesures d'audience de référence pour le compte de ses adhérents, quel que soit le média : Internet, TV, presse, communication extérieure, radio et cinéma.

1.2 Jeu de données

Pour commencer nous allons d'abord effectuer une **stratification** des bases de données en fonction des cibles d'audience comme représenté dans la Figure 1.1.

strate	Sexe	Age_5	CSP_3
1	Homme	15-24 ans	CSP+
2	Homme	15-24 ans	CSP-
3	Homme	15-24 ans	Inactifs
4	Homme	25-34 ans	CSP+
5	Homme	25-34 ans	CSP-
6	Homme	25-34 ans	Inactifs
7	Homme	35-49 ans	CSP+
8	Homme	35-49 ans	CSP-Inactifs
9	Homme	50-64 ans	CSP+
10	Homme	50-64 ans	CSP-
11	Homme	50-64 ans	Inactifs
12	Homme	65 ans et +	CSP+
13	Homme	65 ans et +	CSP-Inactifs
14	Femme	15-24 ans	CSP+
15	Femme	15-24 ans	CSP-
16	Femme	15-24 ans	Inactifs
17	Femme	25-34 ans	CSP+
18	Femme	25-34 ans	CSP-
19	Femme	25-34 ans	Inactifs
20	Femme	35-49 ans	CSP+
21	Femme	35-49 ans	CSP-Inactifs
22	Femme	50-64 ans	CSP+
23	Femme	50-64 ans	CSP-
24	Femme	50-64 ans	Inactifs
25	Femme	65 ans et +	CSP+
26	Femme	65 ans et +	CSP-Inactifs

Figure 1.1 : Stratification des bases de données

On considère $U = U_1 \sqcup \cdots \sqcup U_D$ la population totale stratifiée où D est le nombre de strates et U_d l'ensemble de tous les individus de la strate d et $Imp \subset U$ est l'ensemble des individus de la population totale U ayant été impressionés lors de cette campagne. Pour tout individu $k \in U$ on note par $n_k \in \mathbb{N}$ le nombre de fois où l'individu k a été impressionné. En particulier, on a $Imp = \{k \in U : n_k > 0\}$. On peut aussi donner de manière aggrégée le Reach : comme le nombre d'impressions uniques (= le reach) des hommes. Il peut aussi être rendu une répartition des $(n_k)_{k \in U_1}$.

L'estimation de l'audience d'une campagne de publicités digitale consiste à renvoyer un tableau par strate comme dans la Figure 1.2 :

strate	Reach (couverture)	Frequency (répétition)	GRP
(d=1) H 15-24 CSP+	$R_1 = \sum_{k \in U_1} I(k \in Imp)$	$f_1 = \frac{1}{R_1} \sum_{k \in U_1} n_k$	$R_1 \times f_1$
...	...	...	...
(d=D) F $\geq$ 65 CSP—-inactifs	R_D	f_D	$R_D \times f_D$

Figure 1.2 : Rendu de la mesure d'audience d'une campagne de publicité.

Etant donnée une campagne digitale, on cherche à estimer les D couples $(R_1, f_1), \cdots, (R_D, f_D)$, pour cela nous avons besoin de 4 bases de données :

1) le nombre total d'impressions de la campagne

2) la totalité des cookies des annonceurs participant à cette campagne exposés ou non à cette campagne et les données (en majorié déclaratives) des cookies de ces annonceurs recueillies par les annonceurs eux-mêmes (les **PPD**)

3) les données du dernier recensement

4) les données d'un panel et en particulier les panélistes ayant un cooky de ses annonceurs qui ont ou pas été impréssionnés par cette campagne.

1.3 Modélisation du problème

Dans notre problème on considère que P est la taille de notre population et pour chaque utilisateur u dans notre population on note z_u le type de ce dernier sachant qu'on a en tout K types d'utilisateurs, on peut représenter le type d'un uilisateur par son groupe socio-démographique par exemple. Et soit J le nombre d'appareils (devices) sur lesquels la compagne s'est tenue.

Pour tout $u \in [P]$ et pour tout $j \in [J]$ on notera N_{uj} le nombre de cookies atteints issus par le user u en regardant la publicité sur le device j. On pourra ainsi construire pour chaque compagne publicitaire la matrice cookies-to-user comme suit :

User Id	device 1	...	device j	...	device J
1	$N_{1,1}$	...	$N_{1,j}$	...	$N_{1,J}$
$\vdots$	$\vdots$	...	$\vdots$	$\vdots$	$\vdots$
u	$N_{u,1}$	...	N_{uj}	...	N_{uJ}
$\vdots$	$\vdots$	...	$\vdots$	$\vdots$	$\vdots$
P	$N_{P,1}$	...	N_{Pj}	...	N_{PJ}

Table 1.1 : Matrice cookies-to-user

Ce qui nous importe dans notre problème est la quantité d'utilisateurs atteints ; un utilisateur est atteint lorsque un de ses cookies a été atteint. Nous aboutissons donc à la définition suivante du Reach :

$$\text{Reach } := \sum_{u\in[P]} I\left(\sum_{j\in[J]} N_{uj} > 0\right) \tag{1.1}$$

où $I(A)$ est la fonction indicatrice qui prend la valeur 1 si A est satisfait sinon 0, ce qui fait que pour chaque utilisateur, si ce dernier a été atteint au moins une fois sur l'un des appareils alors il sera compté comme atteint.

Ainsi, nous disposons de N_{sample} campagnes antérieures (qui devraient être proches/équivalentes à la campagne donnée). Les données disponibles pour chaque campagne passée peuvent être résumées dans un tableau représenté dans la Table 1.1. Dans cette configuration, nous considérons ces données comme un modèle de régression dont la sortie est le reach estimé $\hat{R}$ et l'entrée est le vecteur du nombre total de cookies atteints par chaque appareil, on le notera $c = (c_j)_{j=1}^{J}$ avec $c_j = \sum_{u=1}^{P} N_{uj}$ .C'est dans cette configuration que nous devrions essayer d'optimiser la reach-curve qui est une fonction definie comme suit :

$$\begin{cases} (\mathbb{R}^+)^J & \to & [0,1] \\ \mathrm{c} = (\mathrm{c}_j)_{j=1}^{J} & \to & Reach \end{cases}$$

Ici le reach est donné sous forme de proportion de P ie $Reach := \frac{N_U}{P}$ où N_U est le nombre moyen d'utilisateurs uniques atteints.

2

Modèle de Google

De la même façon, dans le modèle de Google, les auteurs ont supposé qu'ils connaissent N_{sample} compagnes sur une population de taille P distribuées de la même façon que la compagne publicitaire et que pour chaque compagne ils obsevent la proportion d'utilisateurs atteints (ie le reach).

L'approche naturelle pour construire un modèle statistique pour un vecteur de comptages $(N_{uj})_{j\in[J]}$ est de considérer des variables de Poisson. Mais comme nous pouvons anticiper que la valeur 0 peut être plus souvent observée, nous pouvons lui attribuer un poids supplémentaire. Nous considérons donc un mélange entre un Dirac en 0 et une variable de Poisson :

$$N_{uj} \sim \eta_{kj}\delta_0 + (1-\eta_{kj})\ \text{Poisson}\,(\theta_{kj})$$

Mais comme le reach ne prend en compte que les deux événements complémentaires $\{N_{uj}=0\}$ et $\{N_{uj}>0\}$, on pourrait donc ajuster le modèle en posant $\eta_{kj}=0$ on aura alors :

$$N_{uj} \sim\ \text{Poisson}\,(\theta_{kj})$$

2.1 Acitvity Distribution Function Model

Pour obtenir le modèle ADF (Acitvity Distribution Function) de Google, nous aurons besoin de calculer l'espérance du Reach qu'on a défini dans (1.1) :

$$\mathbb{E}[\text{ Reach }] = \mathbb{E}\left(\sum_{k\in[K]} \sum_{u\in[P] \text{ de type } k} I\left(\sum_{j\in[J]} N_{uj} > 0\right)\right)$$

$$= \sum_{k\in[K]} \sum_{u\in[P] \text{ de type } k} \mathbb{P}\left(\sum_{j\in[J]} N_{uj} > 0\right)$$

$$= \sum_{k\in[K]} \sum_{u\in[P] \text{ de type } k} \mathbb{P}\left(\exists j \in [J] : N_{uj} > 0\right)$$

$$= \sum_{k\in[K]} \sum_{u\in[P] \text{ de type } k} \left(1 - \mathbb{P}\left(\forall j \in [J] : N_{uj} = 0\right)\right)$$

$$= P - \sum_{k\in[K]} \sum_{u\in[P] \text{ de type } k} \mathbb{P}\left(\forall j \in [J] : N_{uj} = 0\right)$$

$$= P - \sum_{k\in[K]} \sum_{u\in[P] \text{ de type } k} \prod_{j\in[J]} \mathbb{P}\left(N_{uj} = 0\right)$$

Ici les auteurs ont supposé que les N_{uj} sont indépendants sur les j (nous allons rejeter ette hypothèse dans le modèle que nous proposerons car celle-ci n'est pas réaliste)

$$\begin{aligned}
\mathbb{E}[\text{ Reach }] &= P - \sum_{k\in[K]} \sum_{u\in[P] \text{ de type } k} \prod_{j\in[J]} \exp\left(-\theta_{kj}\right) \\
&= P\left(1 - \sum_{k\in[K]} \frac{|\{ \text{ users de type } k\}|}{P} \exp\left(-\sum_{j\in[J]} \theta_{kj}\right)\right) \\
&= P\left(1 - \sum_{k\in[K]} \alpha_k \exp\left(- < x_k, t >\right)\right) = P(1 - \mathbb{E}[\exp(- < X, t >)]) \\
&= PR(t)
\end{aligned}$$

où $\alpha_k = \frac{|\{ \text{ users de type } k\}|}{P}$, $\theta_{kj} = x_{kj}t_j$ et $R(t) = 1 - \mathbb{E}[\exp(- < X, t >)]$
Pour trouver la reach-curve sous le modèle DMM pour cette définition de x_{kj} et t_j on doit prendre $t = (t_j)_{j\in[J]} = (c_j/P)_{j\in[J]}$.

2.2 Dirac Mixture Model

Dans ce premier calcul de l'espérance $\mathbb{E}[\text{Reach}]$ sous le modèle ADF, seuls les N_{uj} sont des variables aléatoires. Par conséquent, cet espérance est prise par rapport à ces variables aléatoires et donc ne peut pas être une fonction des c_j . Pour retrouver exactement le modèle de la Reach curve sous l'hypothèse DMM en fonction des cj , nous devons calculer l'espérance du Reach sachant ce qu'on a observé sur les c_j , ce qui nous amène à calculer la fonction de l'espérance conditionnelle suivante :

$$(c_j)_{j\in[J]} \in \mathbb{N}^J \to \frac{1}{P}\mathbb{E}\left[\text{Reach} \mid \sum_{u\in[P]} N_{uj} = c_j, j \in [J]\right] \tag{2.1}$$

Cette fonction est la Reach curve dans le modèle DMM.

$$\begin{aligned}
&\mathbb{E}\left[\text{Reach} \mid \textstyle\sum_{u\in[P]} N_{uj} = c_j, j \in [J]\right] \\
&= P - \sum_{k\in[K]} \sum_{u\in[P]\text{de type } k} \mathbb{P}\left(\sum_{j\in[J]} N_{uj} = 0 \mid \sum_{u\in[P]} N_{uj} = c_j, j \in [J]\right) \\
&= P - \sum_{k\in[K]} \sum_{u\in[P]\text{de type } k} \mathbb{P}\left(\forall j \in [J], N_{uj} = 0 \mid \sum_{u\in[P]} N_{uj} = c_j, j \in [J]\right)
\end{aligned}$$

Comme tout à l'heure on suppose que les N_{uj} sont indépendants sur les j, on a alors :

$$\mathbb{P}\left(\forall j \in [J], N_{u_0 j} = 0 \mid \sum_{u\in[P]} N_{uj} = c_j, j \in [J]\right) = \frac{\mathbb{P}\left(\forall j \in [J], N_{u_0 j} = 0 \text{ and } \sum_{u\in[P]} N_{uj} = c_j\right)}{\mathbb{P}\left(\sum_{u\in[P]} N_{uj} = c_j, j \in [J]\right)}$$

$$= \prod_{j\in[J]} \frac{\mathbb{P}\left(N_{u_0 j} = 0 \text{ and } \sum_{u\in[P]} N_{uj} = c_j\right)}{\mathbb{P}\left(\sum_{u\in[P]} N_{uj} = c_j\right)} = \prod_{j\in[J]} \frac{\mathbb{P}\left(N_{u_0 j} = 0\right)\mathbb{P}\left(\sum_{u\neq u_0} N_{uj} = c_j\right)}{\mathbb{P}\left(\sum_{u\in[P]} N_{uj} = c_j\right)}$$

$$= \prod_{j\in[J]} \left(1 - \frac{\theta_{u_0 j}}{\sum_u \theta_{uj}}\right)^{c_j}$$

Où on pose $\theta_{uj} = \theta_{kj}$ quand u est de type k.

Puisqu'on a que les N_{uj} sont des variables de Poisson indépendantes , alors leur somme est une variable de Poisson dont son paramètre la somme des paramètre de ces variables, ce qui nous donne :

$$N_{u_0 j} \sim \text{Poisson}\left(\theta_{u_0 j}\right), \sum_{u\neq u_0} N_{uj} \sim \text{ Poisson } \left(\sum_{u\neq u_0} \theta_{uj}\right) \text{ et } \sum_{u\in[P]} N_{uj} \sim \text{Poisson}\left(\sum_u \theta_{uj}\right)$$

On a donc :

$$\mathbb{E}\left[\text{Reach} \mid \sum_{u\in[P]} N_{uj} = c_j, j \in [J]\right] = P\left(1 - \sum_{k\in[K]} \frac{\mid \text{ users de type } k\} \mid}{P} \prod_{j\in[J]} \left(1 - \frac{\theta_{kj}}{\sum_u \theta_{uj}}\right)^{c_j}\right)$$

$$= P(1 - \mathbb{E}\exp(-\langle X, t\rangle)) = PR(t)$$

où $t = (c_j/P)_{j\in[J]}$ et X est une variable aléatoire à valeurs dans $\mathbb{R}^J$ tel que pour tout $k \in [K]$ et $j \in [J]$ on a :

$$\mathbb{P}[X = x_k] = \alpha_k = \frac{\mid \{ \text{ users de type } k\} \mid}{P} \quad \text{et} \quad x_{kj} = -P\log\left(1 - \frac{\theta_{kj}}{\sum_u \theta_{uj}}\right) \geq 0$$

On prend $X \sim \sum_{k=1}^K \alpha_k \delta_{x_k^0}$, dans ce cas, la Reach curve s'écrit :

$$R : t \in (\mathbb{R}_+)^J \to 1 - \sum_{k=1}^K \alpha_k \exp\left(- < x_k^0, t >\right)$$

où :

- $\alpha_k > 0$ et $\sum_k \alpha_k = 1$
- $x_1^0, \ldots, x_K^0$ sont des vecteurs dans $(\mathbb{R}_+)^J$

Ainsi les paramètres à estimer de cette fonction sont : $K, (\alpha_k)_{k=1}^K$ et $\left(x_k^0\right)_{k=1}^K$

2.3 Gaussian Mixture Model

L'algorithme consiste à ajuster un mélange gaussien (GMM) sur $(\mathbb{R}_+)^J$ qui essaie d'ajuster au mieux les données $\left(\hat{R}_t, c^{(t)}\right)_{t=1}^{N_{sample}}$, pour cela on suit les étapes suivantes :

- Les $x_k^0, \forall k \in [K]$ sont pris aléatoirement dans $(\mathbb{R})^J$.
- Les paramètres $(\alpha_k)_{k=1}^K$ sont estimés par les moindres carrés sous contrainte.
- Le nombre K est incrémenté d'au plus l à chaque étape.
- Seuls les centres x_k^0 associés à des poids (α_k) seront conservés.

2.4 Critiques du modèle

- Le choix de de la Reach curve $R(t) = 1-\mathbb{E}[\exp(- < X, t >)]$ n'est pas justifié et la signification du vecteur centre X n'est pas expliquée.
- La supposition de l'indépendance des $N_{uj},\ u \in [P],\ j \in [J]$ sur j n'est pas fondée (on verra plus tard qu'on pourra supposer que les $N_{uj} \mid z_u = k$ sont indépendants).
- Prendre $X \sim \sum_{k=1}^{K} \alpha_k \delta_{x_k^0}$ signifie que $\mathbb{P}\left[X = x_k^0\right] = \alpha_k$ ce qui fait qu'on regarde seulement un seul type de cookies.

2.5 Solutions proposées

La première solution consiste comme dans le modèle de google à avoir les données $\left(\hat{R}_t, c^{(t)}\right)_{t=1}^{N_{sample}}$ où pour tout $t \in [N_{sample}],\ \hat{R}_t$ est le each estimé sur la Population $[P]$ sur la compagne t et où $c^{(t)} = \left(c_j^{(t)}\right)_{j=1}^{J}$ est le vecteur des cookies atteints sur l'ensemble de la population pour chaque type de cookie sur la compagne t . Dans cette configuration, nous considérons ces données comme un modèle de régression dont la sortie est le reach estimée $\hat{R}_t$ et l'entrée est le vecteur du nombre total de cookies atteints sur un appareil $c^{(t)}$. Nous devrons alors apprendre une certaine Reach-curve, soit par les moindres carrés, soit par une approche de maximum de vraisemblance.

La deuxième solution consiste à utiliser une nouvelle approche bayésienne où nous allons utiliser un échantillonneur de Gibbs.

Dans ce stage nous allons seulement nous concentrer sur la première solution.

3

Level Correlated Models

Dans les modèles LCM (Level Correlated Model), on designe par $\theta = \left(q_k, \Sigma^{(k)}\right)_{k\in[K]}$ l'ensemble desparamètres du modèle et par $\mathbb{P}_\theta$ (resp. $\mathbb{E}_\theta$) la probabilité conditionnelle (resp. l'ésperance conditionnelle) données dans un des modèles LCM.
Notre premier objectif est de déterminer la reach curve en fonction de notre paramètre θ :

$$R_\theta : (c_j)_{j\in[J]} \in \mathbb{N}^J \to \mathbb{E}_\theta \left[\text{Reach} \mid \sum_{u\in[P]} N_{uj} = c_j, j \in [J] \right] \tag{3.1}$$

Ensuite on estimera notre paramètre en partant des données $\left(\hat{R}_t, \left(c_j^{(t)}\right)_{j\in[J]}\right)_{t=1}^{N_{sample}}$ où pour tout $t \in [N_{sample}]$, $\hat{R}_t$ est le each estimé sur la population $[P]$ sur la compagne t et où $c^{(t)} = \left(c_j^{(t)}\right)_{j=1}^{J}$ est le vecteur des cookies atteints sur l'ensemble de la population pour chaque type de cookie sur la compagne t . Pour ce faire, on doit minimiser la fonction :

$$loss : \begin{cases} \left(\mathbb{R} \times S^{J\times J}\right)^K & \longrightarrow \\ \theta & \longrightarrow \sum_{i=1}^{N_{sample}} \left(\hat{R}_i - R_\theta\left(\left(c_j^{(t)}\right)_{j\in[J]}\right)\right)^2 \end{cases} \tag{3.2}$$

où $S^{d\times d}$ est la cône des matrices symétriques semi-definies positives. Pour y arriver,on doit utiliser des méthodes du style descente de gradient, Nelder-Mead ou une optimisation bayésienne .

3.1 Reach Curve

Le but étant d'expliciter R_θ definie comme suit :

$$\begin{aligned} R_\theta\left((cj)_{j=1}^J\right) &= \frac{1}{P}\mathbb{E}_\theta\left[\text{Reach} \mid \sum_{u\in[P]} N_{uj} = c_j,\ \forall j \in [J]\right] \\ &= \frac{1}{P}\sum_{u_0\in[P]} \mathbb{P}_\theta\left(\sum_{j\in[J]} N_{u_0 j} > 0 \mid \sum_{u\in[P]} N_{uj} = c_j,\ \forall j \in [J]\right) \\ &= \frac{1}{P}\sum_{u_0\in[P]} 1 - \mathbb{P}_\theta\left(N_{u_0 j} = 0 \quad \forall j \in [J] \mid \sum_{u\in[P]} N_{uj} = c_j,\ \forall j \in [J]\right) \end{aligned}$$

Remarquons qu'on a :

$$\mathbb{P}_\theta\left(N_{u_0 j} = 0, \forall j \in [J] \mid \sum_{u\in[P]} N_{uj} = c_j, \forall j \in [J]\right) = \frac{\mathbb{P}_\theta\left(N_{u_0 j} = 0, \forall j \in [J], \sum_{u\in[P]} N_{uj} = c_j, \forall j \in [J]\right)}{\mathbb{P}_\theta\left(\sum_{u\in[P]} N_{uj} = c_j, \forall j \in [J]\right)}$$

car on sait que $\mathbb{P}_\theta\left(\sum_{u\in[P]} N_{uj} = c_j,\ \forall j \in [J]\right) > 0$

3.1.1 LCM 1

On prendera dans ce modèle le vecteur aléatoire $(\alpha_{kj})_{kj}$ dans $\mathbb{R}^{K\times J}$ et la variable aléatoire z_u à valeurs dans $[K]$ qui désigne le type de chaque utilisateur $u \in [P]$

3.1.1.1 Hypothèses

On posera les hypothèses suivantes :

1) $\left(N_{uj} \mid \left(z_u = k, (\alpha_{kj})_{k,j}\right) \sim \text{Poisson}\ (\exp(\alpha_{kj}))\right.$

2) $((\alpha_{kj} : j \in [J]))_{k\in[K]}$ sont indépendants et $(\alpha_{kj})_{j\in[J]} \sim \mathcal{N}\left(0, \Sigma^{(k)}\right)$, $\Sigma^{(k)}$ étant semi-définie positive

3) $\mathbb{P}[z_u = k] = q_k\ \forall k \in [K]$

On pourra également observer que $\mathbb{E}[N_{uj} \mid z_u = k] = \exp\left(\frac{\Sigma_{jj}^{(k)}}{2}\right)$ et
$\text{cov}(N_{up}, N_{ul} \mid z_u = k) = \exp\left(\frac{\Sigma_{pp}^{(k)} + \Sigma_{ll}^{(k)}}{2}\right)\left(\exp\left(\Sigma_{pl}^{(k)}\right) - 1\right)$

Ce qui nous permettera d'avoir des corrélations positives et/ou négatives dans le comptage des cookies

3.1.1.2 Calcul de la Reach Curve

1) Calculons $\mathbb{P}_\theta\Big(N_{u_0 j} = 0 \ \ \forall j \in [J] \ \ , \ \ \sum_{u\in[P]} N_{uj} = c_j \ \ \forall j \in [J]\Big)$:

A l'aide de la formule de probabilités totales

$$f_X(x) = \int f_{X,Y}(x,y)dy = \int f_{X|Y}(x \mid y) f_Y(y) dy \tag{3.3}$$

On obtient :

$$\mathbb{P}_\theta\Big(N_{u_0 j} = 0 \ \ \forall j \in [J] \ \ , \ \ \sum_{u\in[P]} N_{uj} = c_j \ \ \forall j \in [J]\Big)$$
$$= \sum_{(k_u)u} \int_{(\alpha_{kj})_{kj}} \mathbb{P}_\theta\Big(N_{u_0 j} = 0 \ \ \forall j \in [J] \ \ , \ \ \sum_{u\in[P]} N_{uj} = c_j \ \ \forall j \in [J] \ \Big| \ z_u = k_u \ \ , \ \ (\alpha_{kj})_{kj}\Big)$$
$$\mathbb{P}_\theta(z_u = k_u \ , \ (\alpha_{kj})_{kj}) \ d_{\alpha_{11}} ... d_{\alpha_{KJ}}$$

Comme z_u et $(\alpha_{kj})_{kj}$ sont indépendants alors :
$$= \sum_{(k_u)u} \int_{\alpha_{kj}} \mathbb{P}_\theta\Big(N_{u_0 j} = 0 \ \ \forall j \in [J] \ \ , \ \ \sum_{u\in[P]} N_{uj} = c_j \ \ \forall j \in [J] \ \Big| \ z_u = k_u \ \ , \ \ (\alpha_{kj})_{kj}\Big)$$
$$\mathbb{P}_\theta(z_u = k_u) \ f_\alpha((\alpha_{kj})_{kj}) \ d_{\alpha_{11}} ... d_{\alpha_{KJ}}$$

D'après l'hypothèse 3 :
$$= \sum_{(k_u)u} \int_{\alpha_{kj}} \mathbb{P}_\theta\Big(N_{u_0 j} = 0 \ \ \forall j \in [J] \ \ , \ \ \sum_{u\in[P]} N_{uj} = c_j \ \ \forall j \in [J] \ \Big| \ z_u = k_u \ \ , \ \ (\alpha_{kj})_{kj}\Big)$$
$$\prod_{u\in[P]} q_{k_u} \ f_\alpha((\alpha_{kj})_{kj}) \ d_{\alpha_{11}} ... d_{\alpha_{KJ}}$$

Sous l'hypothèse d'indépendance des $\Big(N_{uj} \mid \Big(z_u = k, (\alpha_{kj})_{kj}\Big)\Big)_{j\in[J]}$:
$$= \sum_{(k_u)u} \int_{\alpha_{kj}} \prod_{j=1}^{J} \mathbb{P}_\theta\Big(N_{u_0 j} = 0 \ \ , \ \ \sum_{u\in[P]} N_{uj} = c_j \ \Big| \ z_u = k_u \ \ , \ \ (\alpha_{kj})_{kj}\Big)$$
$$\prod_{u\in[P]} q_{k_u} \ f_\alpha((\alpha_{kj})_{kj}) \ d_{\alpha_{11}} ... d_{\alpha_{KJ}}$$

$$= \sum_{(k_u)u} \int_{\alpha_{kj}} \prod_{j=1}^{J} \mathbb{P}_\theta\Big(N_{u_0 j} = 0 \ \ , \ \ \sum_{u\neq u_0} N_{uj} = c_j \ \Big| \ z_u = k_u \ \ , \ \ (\alpha_{kj})_{kj}\Big)$$
$$\prod_{u\in[P]} q_{k_u} \ f_\alpha((\alpha_{kj})_{kj}) \ d_{\alpha_{11}} ... d_{\alpha_{KJ}}$$

$$= \sum_{(k_u)u} \int_{\alpha_{kj}} \prod_{j=1}^{J} \Bigg(\mathbb{P}_\theta \Big(N_{u_0 j} = 0 \;\Big|\; z_u = k_u \;\;,\;\; (\alpha_{kj})_{kj} \Big)$$

$$\mathbb{P}_\theta \Big(\sum_{u \neq u_0} N_{uj} = c_j \;\Big|\; z_u = k_u \;\;,\;\; (\alpha_{kj})_{kj} \Big) \Bigg) \prod_{u \in [P]} q_{k_u} \; f_\alpha((\alpha_{kj})_{kj}) \; d_{\alpha_{11}} ... d_{\alpha_{KJ}}$$

Comme on a $(N_{uj} | (z_u = k, (\alpha_{kj})_{kj}) \sim Poisson(exp(\alpha_{k_u j}))$ alors
$\mathbb{P}_\theta \Big(N_{u_0 j} = 0 \;\Big|\; z_u = k_u \;\;,\;\; (\alpha_{kj})_{kj} \Big) = exp(-exp(\alpha_{k_u j}))$ et
$\Big(\sum_{u \neq u_0} N_{uj} \;\Big|\; z_u = k_u \;\;,\;\; (\alpha_{kj})_{kj} \Big) \sim Poisson(\sum_{k=1}^{K} n_{k_0} exp(\alpha_{k_u j}))$
avec $n_{k_0} = |\{u \in [P] : z_u = k \;\; et \;\; u \neq u_0\}|$ alors :

$$\mathbb{P}_\theta \Big(\sum_{u \neq u_0} N_{uj} = c_j \;\Big|\; z_u = k_u \;\;,\;\; (\alpha_{kj})_{kj} \Big) = exp(\, -(\sum_{k=1}^{K} n_{k_0} exp(\alpha_{k_u j})) \,) \; \frac{(\sum_{k=1}^{K} n_{k_0} exp(\alpha_{k_u j}))^{c_j}}{c_j!}$$

D'après l'hypothèse 2 $(\alpha_{kj})_{j \in [J]} \sim \mathcal{N}(0, \Sigma^{(k)})$ alors :

$$f_\alpha((\alpha_{kj})_{kj}) = \prod_{i=1}^{K} \frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(i)})}} \; exp(-\frac{1}{2} \; \alpha_i^T (\Sigma^{(i)})^{-1} \alpha_i)$$

où $\alpha_i^T = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{iJ})$

En remplaçant on obtient :

$$\mathbb{P}_\theta \Big(N_{u_0 j} = 0 \;\; \forall j \in [J] \;\;,\;\; \sum_{u \in [P]} N_{uj} = c_j \;\; \forall j \in [J] \Big)$$

$$= \sum_{(k_u)u} \int_{\alpha_{kj}} \prod_{j=1}^{J} \Bigg(exp(-exp(\alpha_{k_u j})) \;\; exp(\, -(\sum_{k=1}^{K} n_{k_0} exp(\alpha_{k_u j})) \,) \; \frac{(\sum_{k=1}^{K} n_{k_0} exp(\alpha_{k_u j}))^{c_j}}{c_j!} \Bigg)$$

$$\prod_{u \in [P]} q_{k_u} \; \frac{1}{\sqrt{(2\pi)^{KJ} det(\prod_{i=1}^{K} \Sigma^{(i)})}} \; exp(-\frac{1}{2} \; \sum_{i=1}^{K} \alpha_i^T (\Sigma^{(i)})^{-1} \alpha_i) \quad d_{\alpha_{11}} ... d_{\alpha_{KJ}}$$

$$= \sum_{(k_u)u} \int_{\alpha_{kj}} exp(-\sum_{j=1}^{J} exp(\alpha_{k_u j})) \;\; exp(\, -\sum_{j=1}^{J} (\sum_{k=1}^{K} n_{k_0} exp(\alpha_{k_u j})) \,) \; \frac{\prod_{j \in [J]} (\sum_{k=1}^{K} n_{k_0} exp(\alpha_{k_u j}))^{c_j}}{\prod_{j=1}^{J} c_j!}$$

$$\prod_{u \in [P]} q_{k_u} \; \frac{1}{\sqrt{(2\pi)^{KJ} det(\prod_{i=1}^{K} \Sigma^{(i)})}} \; exp(-\frac{1}{2} \; \sum_{i=1}^{K} \alpha_i^T (\Sigma^{(i)})^{-1} \alpha_i) \quad d_{\alpha_{11}} ... d_{\alpha_{KJ}}$$

$$= \frac{1}{\prod_{j \in [J]} c_j! \;\; \sqrt{(2\pi)^{KJ} det(\prod_{i=1}^{K} \Sigma^{(i)})}} \int_{\alpha_{kj}} \sum_{(k_u)u} exp(-\sum_{j=1}^{J} exp(\alpha_{k_u j})) \;\; exp(\, -\sum_{j=1}^{J} (\sum_{k=1}^{K} n_{k_0} exp(\alpha_{k_u j})) \,)$$

$$\prod_{j \in [J]} (\sum_{k=1}^{K} n_{k_0} exp(\alpha_{k_u j}))^{c_j} \prod_{u \in [P]} q_{k_u} \;\; exp(-\frac{1}{2} \; \sum_{i=1}^{K} \alpha_i^T (\Sigma^{(i)})^{-1} \alpha_i) \quad d_{\alpha_{11}} ... d_{\alpha_{KJ}}$$

$$= \frac{1}{\prod_{j\in[J]} c_j! \ \sqrt{(2\pi)^{KJ} det(\prod_{i=1}^{K} \Sigma^{(i)})}} \int_{\alpha_{kj}} \prod_{u\in[P]} q_{k_u} \ \ exp(-\frac{1}{2} \ \sum_{i=1}^{K} \alpha_i^T (\Sigma^{(i)})^{-1} \alpha_i)$$
$$\sum_{(k_u)u} exp(-\sum_{j=1}^{J} exp(\alpha_{k_u j})) \ \ exp(-\sum_{j=1}^{J} (\sum_{k=1}^{K} n_{k_0} exp(\alpha_{k_u j}))) \ \prod_{j\in[J]} (\sum_{k=1}^{K} n_{k_0} exp(\alpha_{k_u j}))^{c_j} \quad d_{\alpha_{11}} ... d_{\alpha_{KJ}}$$

On notera

$$I_1 = \int_{\alpha_{kj}} \prod_{u\in[P]} q_{k_u} \ \ exp(-\frac{1}{2} \ \sum_{i=1}^{K} \alpha_i^T (\Sigma^{(i)})^{-1} \alpha_i)$$
$$\sum_{(k_u)u} exp(-\sum_{j=1}^{J} exp(\alpha_{k_u j})) \ \ exp(-\sum_{j=1}^{J} (\sum_{k=1}^{K} n_{k_0} exp(\alpha_{k_u j}))) \ \prod_{j\in[J]} (\sum_{k=1}^{K} n_{k_0} exp(\alpha_{k_u j}))^{c_j} \quad d_{\alpha_{11}} ... d_{\alpha_{KJ}}$$

2) Calculons $\mathbb{P}_\theta\Big(\ \sum_{u\in[P]} N_{uj} = c_j \ \ \forall j \in [J]\Big)$ **:**

De la même façon en utilisant la formule de probabilités totales introduite dans (3.3)

$$\mathbb{P}_\theta\Big(\sum_{u\in[P]} N_{uj} = c_j \ \ \forall j \in [J]\Big)$$
$$= \sum_{(k_u)u} \int_{\alpha_{kj}} \mathbb{P}_\theta\Big(\sum_{u\in[P]} N_{uj} = c_j \ \ \forall j \in [J] \ \Big| \ z_u = k_u \ \ , \ \ (\alpha_{kj})_{kj}\Big) \ \mathbb{P}_\theta(z_u = k_u \ , \ (\alpha_{kj})_{kj}) \ d_{\alpha_{11}} ... d_{\alpha_{KJ}}$$

Comme z_u et $(\alpha_{kj})_{kj}$ sont indépendants alors :

$$= \sum_{(k_u)u} \int_{\alpha_{kj}} \mathbb{P}_\theta\Big(\sum_{u\in[P]} N_{uj} = c_j \ \ \forall j \in [J] \ \Big| \ z_u = k_u \ \ , \ \ (\alpha_{kj})_{kj}\Big) \ \mathbb{P}_\theta(z_u = k_u) \ f_\alpha((\alpha_{kj})_{kj}) \ d_{\alpha_{11}} ... d_{\alpha_{KJ}}$$

Daprès l'hypothèse 2

$$= \sum_{(k_u)u} \int_{\alpha_{kj}} \mathbb{P}_\theta\Big(\sum_{u\in[P]} N_{uj} = c_j \ \ \forall j \in [J] \ \Big| \ z_u = k_u \ \ , \ \ (\alpha_{kj})_{kj}\Big)$$
$$\prod_{u\in[P]} q_{k_u} \ \frac{1}{\sqrt{(2\pi)^{KJ} det(\prod_{i=1}^{K} \Sigma^{(i)})}} \ exp(-\frac{1}{2} \ \sum_{i=1}^{K} \alpha_i^T (\Sigma^{(i)})^{-1} \alpha_i) \quad d_{\alpha_{11}} ... d_{\alpha_{KJ}}$$

Sous l'hypothèse d'indépendance des $\Big(N_{uj} \mid \Big(z_u = k, (\alpha_{kj})_{kj}\Big)\Big)_{j\in[J]}$ et l'hypothèse 1 :

$$= \sum_{(k_u)u} \int_{\alpha_{kj}} \prod_{j=1}^{J} \left(exp(\ -(\sum_{k=1}^{K} n_k exp(\alpha_{k_u j}))) \ \frac{(\sum_{k=1}^{K} n_k exp(\alpha_{k_u j}))^{c_j}}{c_j!} \right)$$
$$\prod_{u\in[P]} q_{k_u} \ \frac{1}{\sqrt{(2\pi)^{KJ} det(\prod_{i=1}^{K} \Sigma^{(i)})}} \ exp(-\frac{1}{2} \ \sum_{i=1}^{K} \alpha_i^T (\Sigma^{(i)})^{-1} \alpha_i) \quad d_{\alpha_{11}} ... d_{\alpha_{KJ}}$$

$$= \int_{\alpha_{kj}} \sum_{(k_u)u} exp(-\sum_{j=1}^{J}(\sum_{k=1}^{K} n_k exp(\alpha_{k_u j}))\,)\ \frac{\prod_{j\in[J]}(\sum_{k=1}^{K} n_k exp(\alpha_{k_u j}))^{c_j}}{\prod_{j=1}^{J} c_j!}$$

$$\prod_{u\in[P]} q_{k_u}\ \frac{1}{\sqrt{(2\pi)^{KJ} det(\prod_{i=1}^{K}\Sigma^{(i)})}}\ exp(-\frac{1}{2}\ \sum_{i=1}^{K}\alpha_i^T(\Sigma^{(i)})^{-1}\alpha_i) \quad d_{\alpha_{11}}...d_{\alpha_{KJ}}$$

$$= \frac{1}{\prod_{j=1}^{J} c_j!\ \sqrt{(2\pi)^{KJ} det(\prod_{i=1}^{K}\Sigma^{(i)})}} \int_{\alpha_{kj}} \sum_{(k_u)u} exp(-\sum_{j=1}^{J}(\sum_{k=1}^{K} n_k exp(\alpha_{k_u j}))\,)$$

$$\prod_{j\in[J]}(\sum_{k=1}^{K} n_k exp(\alpha_{k_u j}))^{c_j}\ \prod_{u\in[P]} q_{k_u}\ exp(-\frac{1}{2}\ \sum_{i=1}^{K}\alpha_i^T(\Sigma^{(i)})^{-1}\alpha_i) \quad d_{\alpha_{11}}...d_{\alpha_{KJ}}$$

$$= \frac{1}{\prod_{j=1}^{J} c_j!\ \sqrt{(2\pi)^{KJ} det(\prod_{i=1}^{K}\Sigma^{(i)})}} \int_{\alpha_{kj}} \prod_{u\in[P]} q_{k_u}\ exp(-\frac{1}{2}\ \sum_{i=1}^{K}\alpha_i^T(\Sigma^{(i)})^{-1}\alpha_i)$$

$$\sum_{(k_u)u} exp(-\sum_{j=1}^{J}(\sum_{k=1}^{K} n_k exp(\alpha_{k_u j}))\,)\ \prod_{j\in[J]}(\sum_{k=1}^{K} n_k exp(\alpha_{k_u j}))^{c_j} \quad d_{\alpha_{11}}...d_{\alpha_{KJ}}$$

Avec $n_k = |\{u \in [P] : z_u = k\}$ et on notera :

$$I_2 = \int_{\alpha_{kj}} \prod_{u\in[P]} q_{k_u}\ exp(-\frac{1}{2}\ \sum_{i=1}^{K}\alpha_i^T(\Sigma^{(i)})^{-1}\alpha_i)$$
$$\sum_{(k_u)u} exp(-\sum_{j=1}^{J}(\sum_{k=1}^{K} n_k exp(\alpha_{k_u j}))\,)\ \prod_{j\in[J]}(\sum_{k=1}^{K} n_k exp(\alpha_{k_u j}))^{c_j} \quad d_{\alpha_{11}}...d_{\alpha_{KJ}}$$

Notre fonction dans le modèle LCM1 devient :

$$\frac{1}{P}\sum_{u_0\in[P]} 1 - \frac{I_1}{I_2}$$

On remarquera que les deux intégrales I_1 et I_2 sont particulièrement compliqués à calculer et à manipuler avec une machine

3.1.2 LCM 2

Comme dans LCM1 on prendera la variable aléatoire z à valeurs dans $[K]$ qui désigne le type de chaque utilisateur $u \in [P]$ et le vecteur aléatoire $(\alpha_j)_j$ dans $\mathbb{R}^J$

3.1.2.1 Hypothèses

On posera les hypotheses suivantes :

1) $N_{uj} \mid (\alpha_j)_{j\in[J]} \sim \text{poisson}\left(\exp\left(\alpha_j\right)\right) \ \ \forall j \in [J]$

2) $(\alpha_j)_{j\in[J]} \mid z \sim \mathcal{N}\left(0, \Sigma^{(z)}\right)$

3) $\mathbb{P}\left[z = k\right] = q_k \ \ \forall k \in [K]$

4) Les $N_{uj} \mid (\alpha_j)_{j\in[J]}$ sont indépendants

5) Les vecteurs $N_u := (N_{uj})_{j\in[J]}, u \in [N]$ sont i.i.d

3.1.2.2 Calcul de la Reach Curve

Remarquons que sous l'hypothèse 5 on :

$$\mathbb{P}_\theta\left(\forall j \in [J], N_{u_0 j} = 0 \text{ and } \sum_{u\in[P]} N_{uj} = c_j\right) = \mathbb{P}_\theta\left(N_{u_0 j} = 0 \ \forall j \in [J]\right) \ \ \mathbb{P}_\theta\left(\sum_{u\neq u_0} N_{uj} = c_j \ \forall j \in [J]\right)$$

1) Calculons $\mathbb{P}_\theta\left(N_{u_0 j} = 0 \ \forall j \in [J]\right)$ **:**

En utilisant la formule (3.3) on a :

$$\begin{aligned}
&\mathbb{P}_\theta\left(N_{u_0 j} = 0 \ \ \forall j \in [J]\right) \\
&= \int_{\alpha_j} \mathbb{P}_\theta\left(N_{u_0 j} = 0 \ \forall j \in [J] \mid (\alpha_j)_j\right) f_\alpha((\alpha_j)_j) \ \ d_{\alpha_1}...d_{\alpha_J}
\end{aligned}$$

Sous les hypothèses 1 et 4 :

$$= \int_{\alpha_j} \prod_{j\in[J]} exp(-exp(\alpha_j)) f_\alpha((\alpha_j)_j) \ \ d_{\alpha_1}...d_{\alpha_J}$$

En appliquant une deuxième fois la formule (3.3) et sous l'hypothèse 3 :

$$= \int_{\alpha_j} \prod_{j\in[J]} exp(-exp(\alpha_j)) \sum_{z\in[K]} q_k f_{\alpha|z}((\alpha_j)_j \mid z) \ \ d_{\alpha_1}...d_{\alpha_J}$$

Sous l'hypothèse 2 :

$$= \int_{\alpha_j} \prod_{j\in[J]} exp(-exp(\alpha_j)) \sum_{z\in[K]} q_k \frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}} exp\left(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2}\right) \ \ d_{\alpha_1}...d_{\alpha_J}$$

$$= \int_{\alpha_j} exp\left(\sum_{j\in[J]} -exp(\alpha_j)\right) \sum_{z\in[K]} q_k \frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}} exp\left(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2}\right) \ \ d_{\alpha_1}...d_{\alpha_J}$$

$$= \int_{\alpha_j} \sum_{z\in[K]} exp\left(\sum_{j\in[J]} -exp(\alpha_j)\right) q_k \frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}} exp\left(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2}\right) \ \ d_{\alpha_1}...d_{\alpha_J}$$

$$= \sum_{z\in[K]} \int_{\alpha_j} exp\left(\sum_{j\in[J]} -exp(\alpha_j)\right) q_k \frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}} exp\left(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2}\right) \ d_{\alpha_1}...d_{\alpha_J}$$

$$= \sum_{z\in[K]} q_k \frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}} \int_{\alpha_j} exp\left(\sum_{j\in[J]} -exp(\alpha_j)\right) exp\left(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2}\right) \ d_{\alpha_1}...d_{\alpha_J}$$

$$= \sum_{z\in[K]} q_k \frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}} \int_{\alpha_j} exp\left(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2} - \sum_{j\in[J]} exp(\alpha_j)\right) \ d_{\alpha_1}...d_{\alpha_J}$$

On posera :

$$I_1^{(z)} = \frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}} \int_{\alpha_j} exp\left(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2} - \sum_{j\in[J]} exp(\alpha_j)\right) \ d_{\alpha_1}...d_{\alpha_J}$$

2) Calculons $\mathbb{P}_\theta\left(\sum_{u\neq u_0} N_{uj} = c_j \ \forall j \in [J]\right)$:

En procédant de la même façon on a :

$$\mathbb{P}_\theta\left(\sum_{u\neq u_0} N_{uj} = c_j \ \forall j \in [J]\right)$$

$$= \int_{\alpha_j} \mathbb{P}_\theta\left(\sum_{u\neq u_0} N_{uj} = c_j \ \forall j \in [J] \mid (\alpha_j)_j\right) f_\alpha((\alpha_j)_j) \ d_{\alpha_1}...d_{\alpha_J}$$

$$= \int_{\alpha_j} \prod_{j\in[J]} exp(\,-((P-1)exp(\alpha_j))\,) \ \frac{((P-1)exp(\alpha_j))^{c_j}}{c_j!} f_\alpha((\alpha_j)_j) \ d_{\alpha_1}...d_{\alpha_J}$$

$$= \int_{\alpha_j} \prod_{j\in[J]} exp(\,-((P-1)exp(\alpha_j))\,) \ \frac{((P-1)exp(\alpha_j))^{c_j}}{c_j!} \sum_{z\in[K]} q_k f_\alpha((\alpha_j)_j \mid z) \ d_{\alpha_1}...d_{\alpha_J}$$

$$= \int_{\alpha_j} \prod_{j\in[J]} exp(\,-((P-1)exp(\alpha_j))\,) \ \frac{((P-1)exp(\alpha_j))^{c_j}}{c_j!}$$

$$\sum_{z\in[K]} \left(q_k \frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}} exp(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2})\right) \ d_{\alpha_1}...d_{\alpha_J}$$

$$= \int_{\alpha_j} exp(\,-\sum_{j\in[J]} ((P-1)exp(\alpha_j))\,) \ \frac{exp\left(\sum_{j\in[J]}(ln(P-1)+\alpha_j)c_j\right)}{\prod_{j\in[J]} c_j!}$$

$$\sum_{z\in[K]} \left(q_k \frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}} exp(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2})\right) \ d_{\alpha_1}...d_{\alpha_J}$$

$$= \frac{(P-1)^{\sum_{j\in[J]} c_j}}{\prod_{j\in[J]} c_j!} \int_{\alpha_j} exp\left(-(P-1)\sum_{j\in[J]} exp(\alpha_j)\right) \ exp\left(\sum_{j\in[J]} \alpha_j c_j\right)$$

$$\sum_{z\in[K]}\left(q_k\frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}}exp(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2})\right)\ d_{\alpha_1}...d_{\alpha_J}$$
$$=\frac{(P-1)^{\sum_{j\in[J]}c_j}}{\prod_{j\in[J]}c_j!}\int_{\alpha_j}exp\left(\sum_{j\in[J]}-(P-1)exp(\alpha_j)\ +\alpha_j c_j\right)$$
$$\sum_{z\in[K]}\left(q_k\frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}}exp(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2})\right)\ d_{\alpha_1}...d_{\alpha_J}$$
$$=\frac{(P-1)^{\sum_{j\in[J]}c_j}}{\prod_{j\in[J]}c_j!}\int_{\alpha_j}\sum_{z\in[K]}\left(exp\left(\sum_{j\in[J]}-(P-1)exp(\alpha_j)\ +\alpha_j c_j\right)q_k\frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}}\right.$$
$$\left.exp(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2})\ d_{\alpha_1}...d_{\alpha_J}\right.$$
$$=\frac{(P-1)^{\sum_{j\in[J]}c_j}}{\prod_{j\in[J]}c_j!}\sum_{z\in[K]}\int_{\alpha_j}exp\left(\sum_{j\in[J]}-(P-1)exp(\alpha_j)\ +\alpha_j c_j\right)q_k\frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}}$$
$$exp(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2})\ d_{\alpha_1}...d_{\alpha_J}$$
$$=\frac{(P-1)^{\sum_{j\in[J]}c_j}}{\prod_{j\in[J]}c_j!}\sum_{z\in[K]}q_k\frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}}\int_{\alpha_j}exp\left(\sum_{j\in[J]}-(P-1)exp(\alpha_j)\ +\alpha_j c_j\right)$$
$$exp(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2})\ d_{\alpha_1}...d_{\alpha_J}$$
$$=\frac{(P-1)^{\sum_{j\in[J]}c_j}}{\prod_{j\in[J]}c_j!}\sum_{z\in[K]}q_k\frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}}$$
$$\int_{\alpha_j}exp\left(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2}+\sum_{j\in[J]}-(P-1)exp(\alpha_j)\ +\alpha_j c_j\right)\ d_{\alpha_1}...d_{\alpha_J}$$

On posera

$$I_2=\frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}}\int_{\alpha_j}exp\left(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2}+\sum_{j\in[J]}-(P-1)exp(\alpha_j)+\alpha_j c_j\right)d_{\alpha_1}...d_{\alpha_J}$$

3) Calculons $\mathbb{P}_\theta\left(\sum_{u\in[P]}N_{uj}=c_j\ \forall j\in[J]\right)$:

En procédant égalemment de la même manière on a :

$$\mathbb{P}_\theta\left(\sum_{u\in[P]}N_{uj}=c_j\ \forall j\in[J]\right)$$
$$=\int_{\alpha_j}\mathbb{P}_\theta\left(\sum_{u\in[P]}N_{uj}=c_j\ \forall j\in[J]\mid(\alpha_j)_j\right)f_\alpha((\alpha_j)_j)\ d_{\alpha_1}...d_{\alpha_J}$$

$$= \int_{\alpha_j} \prod_{j\in[J]} exp(\,-(Pexp(\alpha_j))\,)\ \frac{(Pexp(\alpha_j))^{c_j}}{c_j!} f_\alpha((\alpha_j)_j)\ \ d_{\alpha_1}...d_{\alpha_J}$$

$$= \int_{\alpha_j} \prod_{j\in[J]} exp(\,-(Pexp(\alpha_j))\,)\ \frac{(Pexp(\alpha_j))^{c_j}}{c_j!} \sum_{z\in[K]} q_k f_\alpha((\alpha_j)_j \mid z)\ \ d_{\alpha_1}...d_{\alpha_J}$$

$$= \int_{\alpha_j} \prod_{j\in[J]} exp(\,-(Pexp(\alpha_j))\,)\ \frac{(Pexp(\alpha_j))^{c_j}}{c_j!}$$

$$\sum_{z\in[K]} \left(q_k \frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}} exp(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2}) \right)\ \ d_{\alpha_1}...d_{\alpha_J}$$

$$= \int_{\alpha_j} exp(\,-\sum_{j\in[J]} (Pexp(\alpha_j))\,)\ \frac{exp\left(\sum_{j\in[J]}(ln(P)+\alpha_j)c_j\right)}{\prod_{j\in[J]} c_j!}$$

$$\sum_{z\in[K]} \left(q_k \frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}} exp(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2}) \right)\ \ d_{\alpha_1}...d_{\alpha_J}$$

$$= \frac{P^{\sum_{j\in[J]} c_j}}{\prod_{j\in[J]} c_j!} \int_{\alpha_j} exp\left(\,-P\sum_{j\in[J]} exp(\alpha_j)\right)\ exp\left(\sum_{j\in[J]} \alpha_j c_j\right)$$

$$\sum_{z\in[K]} \left(q_k \frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}} exp(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2}) \right)\ \ d_{\alpha_1}...d_{\alpha_J}$$

$$= \frac{P^{\sum_{j\in[J]} c_j}}{\prod_{j\in[J]} c_j!} \int_{\alpha_j} exp\left(\sum_{j\in[J]} -Pexp(\alpha_j)\ + \alpha_j c_j\right)$$

$$\sum_{z\in[K]} \left(q_k \frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}} exp(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2}) \right)\ \ d_{\alpha_1}...d_{\alpha_J}$$

$$= \frac{P^{\sum_{j\in[J]} c_j}}{\prod_{j\in[J]} c_j!} \int_{\alpha_j} \sum_{z\in[K]} \left(exp\left(\sum_{j\in[J]} -Pexp(\alpha_j)\ + \alpha_j c_j\right) q_k \frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}} \right.$$

$$\left. exp(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2}) \right)\ \ d_{\alpha_1}...d_{\alpha_J}$$

$$= \frac{P^{\sum_{j\in[J]} c_j}}{\prod_{j\in[J]} c_j!} \sum_{z\in[K]} \int_{\alpha_j} exp\left(\sum_{j\in[J]} -Pexp(\alpha_j)\ + \alpha_j c_j\right) q_k \frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}}$$

$$exp(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2})\ \ d_{\alpha_1}...d_{\alpha_J}$$

$$= \frac{P^{\sum_{j\in[J]} c_j}}{\prod_{j\in[J]} c_j!} \sum_{z\in[K]} q_k \frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}} \int_{\alpha_j} exp\left(\sum_{j\in[J]} -Pexp(\alpha_j) + \alpha_j c_j \right)$$
$$exp(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2}) \; d_{\alpha_1}...d_{\alpha_J}$$

$$= \frac{P^{\sum_{j\in[J]} c_j}}{\prod_{j\in[J]} c_j!} \sum_{z\in[K]} q_k \frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}}$$
$$\int_{\alpha_j} exp\left(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2} + \sum_{j\in[J]} -Pexp(\alpha_j) + \alpha_j c_j \right) \; d_{\alpha_1}...d_{\alpha_J}$$

On posera

$$I_3^{(z)} = \frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(z)})}} \int_{\alpha_j} exp\left(-\frac{(\alpha_1...\alpha_J)(\Sigma^{(z)})^{-1}(\alpha_1...\alpha_J)^T}{2} + \sum_{j\in[J]} -Pexp(\alpha_j) + \alpha_j c_j \right) \; d_{\alpha_1}...d_{\alpha_J}$$

Notre fonction dans le modèle LCM2 devient alors :

$$R_\theta\left((cj)_{j=1}^J\right) = \frac{1}{P} \sum_{u_0\in[P]} 1 - (\frac{P-1}{P})^{\sum_{j\in[J]} c_j} \quad \frac{\sum_{z\in[K]} q_k I_1^{(z)} \quad \sum_{z\in[K]} q_k I_2^{(z)}}{\sum_{z\in[K]} q_k I_3^{(z)}}$$

Puisque aucun terme ne dépend de u_0 alors :

$$R_\theta\left((cj)_{j=1}^J\right) = 1 - (\frac{P-1}{P})^{\sum_{j\in[J]} c_j} \quad \frac{\sum_{z\in[K]} q_k I_1^{(z)} \quad \sum_{z\in[K]} q_k I_2^{(z)}}{\sum_{z\in[K]} q_k I_3^{(z)}} \tag{3.4}$$

Dans ce modèle on remarque que les trois intégrales I_1, I_2 et I_3 sont plus manipulables mais reste toujours difficiles à calculer.

3.1.3 LCM 3

On prendera dans ce modèle le vecteur aléatoire $(\alpha_{uj})_{uj}$ dans $\mathbb{R}^{P\times J}$ et la variable aléatoire z_u à valeurs dans $[K]$ qui désigne le type de chaque utilisateur $u \in [P]$.

3.1.3.1 Hypothèses

On posera les hypothèses suivantes :

1) les $P \times J$ variables $\left(N_{uj} \mid (\alpha_{uj})_{u,j}\right)_{u,j}$ sont indépendantes

2) $N_{uj} \mid (\alpha_{uj})_{u,j} \sim \text{Poisson}\left(\exp\left(\alpha_{uj}\right)\right)$

3) Les $\left((\alpha_{uj})_{j=1}^{J} \mid z_u = k\right)_u$ sont P vecteurs indépendants dans $\mathbb{R}^J$

4) $(\alpha_{uj})_{j=1}^{J} \mid z_u = k \sim \mathcal{N}\left(0, \Sigma^{(k)}\right)$

3.1.3.2 Calcul de la Reach Curve

1) Calculons $\mathbb{P}_\theta\left(\forall j \in [J], N_{u_0 j} = 0 \text{ and } \sum_{u\in[P]} N_{uj} = c_j\right)$:

$$\mathbb{P}_\theta\left(\forall j \in [J], N_{u_0 j} = 0 \text{ and } \sum_{u\in[P]} N_{uj} = c_j\right)$$

$$= \int_{\alpha_{uj}} \mathbb{P}_\theta\left(\forall j \in [J], N_{u_0 j} = 0 \text{ and } \sum_{u\in[P]} N_{uj} = c_j \mid (\alpha_{uj})_{uj}\right) f_\alpha((\alpha_{uj})_{uj}) \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \int_{\alpha_{uj}} \mathbb{P}_\theta\left(N_{u_0 j} = 0 \mid (\alpha_{uj})_{uj} \;\; \forall j \in [J]\right) \quad \mathbb{P}_\theta\left(\sum_{u\in[P]} N_{uj} = c_j \mid (\alpha_{uj})_{uj} \;\; \forall j \in [J]\right) f_\alpha((\alpha_{uj})_{uj}) \;\; d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \int_{\alpha_{uj}} \prod_{j\in[J]} \mathbb{P}_\theta\left(N_{u_0 j} = 0 \mid (\alpha_{uj})_{uj}\right) \quad \mathbb{P}_\theta\left(\sum_{u\in[P]} N_{uj} = c_j \mid (\alpha_{uj})_{uj}\right) f_\alpha((\alpha_{uj})_{uj}) \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \int_{\alpha_{uj}} f_\alpha((\alpha_{uj})_{uj}) \prod_{j\in[J]} \mathbb{P}_\theta\left(N_{u_0 j} = 0 \mid (\alpha_{uj})_{uj}\right) \quad \mathbb{P}_\theta\left(\sum_{u\in[P]} N_{uj} = c_j \mid (\alpha_{uj})_{uj}\right) \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \int_{\alpha_{uj}} \sum_{k\in[K]} \left(f_\alpha((\alpha_{uj})_{uj} \mid z_u = k)\mathbb{P}(z_u = k)\right) \prod_{j\in[J]} \mathbb{P}_\theta\left(N_{u_0 j} = 0 \mid (\alpha_{uj})_{uj}\right)$$

$$\mathbb{P}_\theta\left(\sum_{u\in[P]} N_{uj} = c_j \mid (\alpha_{uj})_{uj}\right) \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \int_{\alpha_{uj}} \sum_{k\in[K]} \left(f_\alpha((\alpha_{uj})_{uj} \mid z_u = k) \;\; q_k\right) \prod_{j\in[J]} \mathbb{P}_\theta\left(N_{u_0 j} = 0 \mid (\alpha_{uj})_{uj}\right)$$

$$\mathbb{P}_\theta\left(\sum_{u\in[P]} N_{uj} = c_j \mid (\alpha_{uj})_{uj}\right) \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \int_{\alpha_{uj}} \sum_{k\in[K]} \left(\prod_{u\in[P]} f_\alpha((\alpha_{uj})_j \mid z_u = k) \;\; q_k\right) \prod_{j\in[J]} \mathbb{P}_\theta\left(N_{u_0 j} = 0 \mid (\alpha_{uj})_{uj}\right)$$

$$\mathbb{P}_\theta\left(\sum_{u\in[P]} N_{uj} = c_j \mid (\alpha_{uj})_{uj}\right) \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \int_{\alpha_{uj}} \sum_{k\in[K]} \left(\prod_{u\in[P]} f_\alpha((\alpha_{uj})_j \mid z_u = k) \;\; q_k\right) \prod_{j\in[J]} exp(-exp(\alpha_{u_0 j}))$$

$$exp(-\sum_{u\neq u_0} exp(\alpha_{uj})) \;\; \frac{(\sum_{u\neq u_0} exp(\alpha_{uj}))^{c_j}}{c_j!} \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \int_{\alpha_{uj}} \sum_{k\in[K]} \left(\prod_{u\in[P]} \frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(k)})}} \, exp(-\frac{1}{2}\, \alpha_u^T (\Sigma^{(k)})^{-1} \alpha_u) \;\; q_k \right)$$

$$\prod_{j\in[J]} exp(-exp(\alpha_{u_0 j})) \quad exp(-\sum_{u\neq u_0} exp(\alpha_{uj})) \;\; \frac{(\sum_{u\neq u_0} exp(\alpha_{uj}))^{c_j}}{c_j!} \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \int_{\alpha_{uj}} \sum_{k\in[K]} \left(\frac{1}{\sqrt{(2\pi)^{PJ} det(\Sigma^{(k)})^P}} \, exp(-\frac{1}{2} \sum_{u\in[P]} \alpha_u^T (\Sigma^{(k)})^{-1} \alpha_u) \;\; q_k \right)$$

$$exp(-\sum_{j\in[J]} exp(\alpha_{u_0 j})) \quad exp(-\sum_{j\in[J]} \sum_{u\neq u_0} exp(\alpha_{uj})) \;\; \prod_{j\in[J]} \frac{(\sum_{u\neq u_0} exp(\alpha_{uj}))^{c_j}}{c_j!} \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \sum_{k\in[K]} \int_{\alpha_{uj}} \frac{1}{\sqrt{(2\pi)^{PJ} det(\Sigma^{(k)})^P}} \, exp(-\frac{1}{2} \sum_{u\in[P]} \alpha_u^T (\Sigma^{(k)})^{-1} \alpha_u) \;\; q_k \;\; exp(-\sum_{j\in[J]} exp(\alpha_{u_0 j}))$$

$$exp(-\sum_{j\in[J]} \sum_{u\neq u_0} exp(\alpha_{uj})) \;\; \prod_{j\in[J]} \frac{(\sum_{u\neq u_0} exp(\alpha_{uj}))^{c_j}}{c_j!} \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \sum_{k\in[K]} \int_{\alpha_{uj}} \frac{1}{\sqrt{(2\pi)^{PJ} det(\Sigma^{(k)})^P}} \, exp(-\frac{1}{2} \sum_{u\in[P]} \alpha_u^T (\Sigma^{(k)})^{-1} \alpha_u) \;\; q_k$$

$$exp(-\sum_{j\in[J]} exp(\alpha_{u_0 j})) \quad exp(-\sum_{j\in[J]} \sum_{u\neq u_0} exp(\alpha_{uj})) \;\; \prod_{j\in[J]} \frac{exp(c_j \;\; ln(\sum_{u\neq u_0} exp(\alpha_{uj})))}{c_j!} \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \frac{1}{\prod_{j\in[J]} c_j!} \sum_{k\in[K]} \int_{\alpha_{uj}} \frac{1}{\sqrt{(2\pi)^{PJ} det(\Sigma^{(k)})^P}} \, exp(-\frac{1}{2} \sum_{u\in[P]} \alpha_u^T (\Sigma^{(k)})^{-1} \alpha_u) \;\; q_k$$

$$exp(-\sum_{j\in[J]} exp(\alpha_{u_0 j})) \quad exp(-\sum_{j\in[J]} \sum_{u\neq u_0} exp(\alpha_{uj})) \;\; exp(\sum_{j\in[J]} c_j \; ln(\sum_{u\neq u_0} exp(\alpha_{uj}))) \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \frac{1}{\sqrt{(2\pi)^{PJ}} \prod_{j\in[J]} c_j!} \sum_{k\in[K]} \frac{q_k}{\sqrt{det(\Sigma^{(k)})^P}}$$

$$\int_{\alpha_{uj}} exp\left(-\frac{1}{2} \sum_{u\in[P]} \alpha_u^T (\Sigma^{(k)})^{-1} \alpha_u - \sum_{j\in[J]} \left(exp(\alpha_{u_0 j}) + \sum_{u\neq u_0} exp(\alpha_{uj}) - c_j ln(\sum_{u\neq u_0} exp(\alpha_{uj})) \right) \right) d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

On posera

$$I_1 = \int exp\left(-\frac{1}{2} \sum_{u\in[P]} \alpha_u^T (\Sigma^{(k)})^{-1} \alpha_u - \sum_{j\in[J]} \left(exp(\alpha_{u_0 j}) + \sum_{u\neq u_0} exp(\alpha_{uj}) - c_j ln(\sum_{u\neq u_0} exp(\alpha_{uj})) \right) \right) d\alpha$$

1) Calculons $\mathbb{P}_\theta \left(\sum_{u\in[P]} N_{uj} = c_j, j \in [J] \right)$:

$$\mathbb{P}_\theta \left(\sum_{u\in[P]} N_{uj} = c_j, j \in [J] \right)$$

$$= \int_{\alpha_{uj}} \mathbb{P}_\theta \left(\sum_{u\in[P]} N_{uj} = c_j \;\; \forall j \in [J] \mid (\alpha_{uj})_{uj} \right) f_\alpha((\alpha_{uj})_{uj}) \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \int_{\alpha_{uj}} \prod_{j\in[J]} \mathbb{P}_\theta \left(\sum_{u\in[P]} N_{uj} = c_j \mid (\alpha_{uj})_{uj} \right) f_\alpha((\alpha_{uj})_{uj}) \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \int_{\alpha_{uj}} \prod_{j\in[J]} exp(-\sum_{u\in[P]} exp(\alpha_{uj})) \ \frac{(\sum_{u\in[P]} exp(\alpha_{uj}))^{c_j}}{c_j!} \quad f_\alpha((\alpha_{uj})_{uj}) \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \int_{\alpha_{uj}} exp(-\sum_{j\in[J]} \sum_{u\in[P]} exp(\alpha_{uj})) \ \prod_{j\in[J]} \frac{(\sum_{u\in[P]} exp(\alpha_{uj}))^{c_j}}{c_j!} \quad f_\alpha((\alpha_{uj})_{uj}) \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \int_{\alpha_{uj}} exp(-\sum_{j\in[J]} \sum_{u\in[P]} exp(\alpha_{uj})) \ \prod_{j\in[J]} \frac{(\sum_{u\in[P]} exp(\alpha_{uj}))^{c_j}}{c_j!}$$

$$\sum_{k\in[K]} q_k \ \prod_{u\in[P]} f_{\alpha|z_u}((\alpha_{uj})_j \mid z_u = k) \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \int_{\alpha_{uj}} exp(-\sum_{j\in[J]} \sum_{u\in[P]} exp(\alpha_{uj})) \ \prod_{j\in[J]} \frac{(\sum_{u\in[P]} exp(\alpha_{uj}))^{c_j}}{c_j!}$$

$$\sum_{k\in[K]} q_k \ \prod_{u\in[P]} \frac{1}{\sqrt{(2\pi)^J det(\Sigma^{(k)})}} \ exp(-\frac{1}{2} \ \alpha_u^T (\Sigma^{(k)})^{-1} \alpha_u) \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \int_{\alpha_{uj}} exp(-\sum_{j\in[J]} \sum_{u\in[P]} exp(\alpha_{uj})) \ \prod_{j\in[J]} \frac{(\sum_{u\in[P]} exp(\alpha_{uj}))^{c_j}}{c_j!}$$

$$\sum_{k\in[K]} q_k \ \frac{1}{\sqrt{(2\pi)^{PJ} det(\Sigma^{(k)})^P}} \ exp(-\sum_{u\in[P]} \frac{1}{2} \ \alpha_u^T (\Sigma^{(k)})^{-1} \alpha_u) \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \frac{1}{\prod_{j\in[J]} c_j! \ \sqrt{(2\pi)^{PJ}}} \int_{\alpha_{uj}} exp(-\sum_{j\in[J]} \sum_{u\in[P]} exp(\alpha_{uj}))$$

$$\prod_{j\in[J]} (\sum_{u\in[P]} exp(\alpha_{uj}))^{c_j} \quad \sum_{k\in[K]} q_k \ \frac{1}{\sqrt{det(\Sigma^{(k)})^P}} \ exp(-\sum_{u\in[P]} \frac{1}{2} \ \alpha_u^T (\Sigma^{(k)})^{-1} \alpha_u) \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \frac{1}{\prod_{j\in[J]} c_j! \ \sqrt{(2\pi)^{PJ}}} \int_{\alpha_{uj}} exp(-\sum_{j\in[J]} \sum_{u\in[P]} exp(\alpha_{uj})) \ exp\left(\sum_{j\in[J]} c_j \ ln(\sum_{u\in[P]} exp(\alpha_{uj})) \right)$$

$$\sum_{k\in[K]} q_k \ \frac{1}{\sqrt{det(\Sigma^{(k)})^P}} \ exp(-\sum_{u\in[P]} \frac{1}{2} \ \alpha_u^T (\Sigma^{(k)})^{-1} \alpha_u) \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \frac{1}{\prod_{j\in[J]} c_j! \ \sqrt{(2\pi)^{PJ}}} \int_{\alpha_{uj}} \sum_{k\in[K]} q_k \ exp(-\sum_{j\in[J]} \sum_{u\in[P]} exp(\alpha_{uj})) \ exp\left(\sum_{j\in[J]} c_j \ ln(\sum_{u\in[P]} exp(\alpha_{uj})) \right)$$

$$\frac{1}{\sqrt{det(\Sigma^{(k)})^P}} \ exp(-\sum_{u\in[P]} \frac{1}{2} \ \alpha_u^T (\Sigma^{(k)})^{-1} \alpha_u) \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \frac{1}{\prod_{j\in[J]} c_j! \ \sqrt{(2\pi)^{PJ}}} \sum_{k\in[K]} \int_{\alpha_{uj}} q_k \ exp(-\sum_{j\in[J]} \sum_{u\in[P]} exp(\alpha_{uj})) \ exp\left(\sum_{j\in[J]} c_j \ ln(\sum_{u\in[P]} exp(\alpha_{uj})) \right)$$

$$\frac{1}{\sqrt{det(\Sigma^{(k)})^P}} \ exp(-\sum_{u\in[P]} \frac{1}{2} \ \alpha_u^T (\Sigma^{(k)})^{-1} \alpha_u) \quad d_{\alpha_{11}} ... d_{\alpha_{PJ}}$$

$$= \frac{1}{\prod_{j\in[J]} c_j! \ \sqrt{(2\pi)^{PJ}}} \sum_{k\in[K]} \frac{q_k}{\sqrt{det(\Sigma^{(k)})^P}} \int_{\alpha_{uj}} exp(-\sum_{j\in[J]} \sum_{u\in[P]} exp(\alpha_{uj}))$$

$$exp\left(\sum_{j\in[J]} c_j \ ln(\sum_{u\in[P]} exp(\alpha_{uj}))\right) \quad exp(-\sum_{u\in[P]} \frac{1}{2}\ \alpha_u^T(\Sigma^{(k)})^{-1}\alpha_u) \quad d_{\alpha_{11}}...d_{\alpha_{PJ}}$$
$$= \frac{1}{\prod_{j\in[J]} c_j!\ \sqrt{(2\pi)^{PJ}}} \sum_{k\in[K]} \frac{q_k}{\sqrt{det(\Sigma^{(k)})^P}}$$
$$\int_{\alpha_{uj}} exp\left(-\sum_{u\in[P]} \frac{1}{2}\ \alpha_u^T(\Sigma^{(k)})^{-1}\alpha_u \ + \sum_{j\in[J]}\left(c_j \ ln(\sum_{u\in[P]} exp(\alpha_{uj})) - \sum_{u\in[P]} exp(\alpha_{uj})\right)\right) \quad d_{\alpha_{11}}...d_{\alpha_{PJ}}$$

On posera

$$I_2 = \int_{\alpha_{uj}} exp\left(-\sum_{u\in[P]} \frac{1}{2}\ \alpha_u^T(\Sigma^{(k)})^{-1}\alpha_u \ + \sum_{j\in[J]}\left(c_j\ ln(\sum_{u\in[P]} exp(\alpha_{uj})) - \sum_{u\in[P]} exp(\alpha_{uj})\right)\right) \quad d_{\alpha_{11}}...d_{\alpha_{PJ}}$$

Notre fonction dans le modèle LCM3 devient alors :

$$R_\theta\left((cj)_{j=1}^J\right) = \frac{1}{P}\sum_{u_0\in[P]} 1 - \frac{I_1}{I_2}$$

Comme dans le modèle LCM1 les deux intégrales I_1 et I_2 sont particulièrement compliquées à calculer et à manipuler avec une machine.

3.2 Choix du modèle

En ayant remarqué que les modèles LCM1 et LCM3 ont des termes diffciles à manipuler numériquement on choisira le modèle LCM2, néanmoins les trois intégrales I_1, I_2 et I_3 sont diffciles à calculer, nous allons donc essayer d'exprimer ces trois intégrales sous forme d'espérances qu'on essaiera d'approximer avec des méthodes numériques.

3.3 Approximation et Quasi-Monte-Carlo

On essaiera d'exprimer les trois intégrales I_1, I_2 et I_3 du LCM2 sous forme d'espérances, vu qu'on a $\alpha|z = k \sim \mathcal{N}(0, \Sigma^{(k)})$ on remarquera donc :

- $I_1 = \int f_{\alpha|z}(\alpha)\ exp(-\sum_{j\in[J]} exp(\alpha_j))\ d\alpha = \mathbb{E}\left[exp(-\sum_{j\in[J]} exp(\alpha_j))\right]$
- $I_2 = \int f_{\alpha|z}(\alpha)\ exp\left(-\sum_{j\in[J]}(P-1)exp(\alpha_j)\ -\alpha_j c_j\right)\ d\alpha$
  $= \mathbb{E}\left[exp\left(-\sum_{j\in[J]}(P-1)exp(\alpha_j)\ -\alpha_j c_j\right)\right]$
- $I_3 = \int f_{\alpha|z}(\alpha)\ exp\left(-\sum_{j\in[J]} Pexp(\alpha_j)\ -\alpha_j c_j\right)\ d\alpha = \mathbb{E}\left[exp\left(-\sum_{j\in[J]} Pexp(\alpha_j)\ -\alpha_j c_j\right)\right]$

Pour approcher ces espérances nous allons utiliser la méthode de Quasi-Monte-Carlo qui est très efficace pour les petits nombres, cette méthode consiste à générer N points d'une suite de sobol (on essaiera de choisir N une puissance de 2 pour garder les propriétés de la suite de sobol, ex : $N = 1024 = 2^{10}$), on essaiera ensuite de générer N vecteurs gaussiens avec les points de la suite de sobol et de calculer leur moyenne. Contrairement à la méthode Monte-Carlo classique, les vecteurs gaussiens générés ne sont pas pseudo-aléatoires mais sont déterministes et forment une suite à faible discrepance, ayant une meilleure distribution uniforme. On procède comme suit :

Soit U une matrice de taille $N \times J$ tel que $\forall$ i $\in [N]$ U_i est un vecteur généré par suite de sobol à valeurs dans $[0,1]^J$, on pose $Z_i = \phi^{-1}(U_i)$ pour tout $i \in [N]$ avec ϕ^{-1} la fonction reciproque de la desité gaussienne multivariée.
Pour avoir nos N vecteurs gaussiens $(\beta_1 ... \beta_i ... \beta_N)$ il suffirait de prendre $\beta_i = Z_i \ L^T$ pour tout $i \in [N]$ et tel que $\Sigma = LL^T$, ainsi on aura donc pour tout $k \in [K]$
avec $\alpha | z = k \sim \mathcal{N}(0, \Sigma^{(k)})$:

- $I_1^{(k)} = \mathbb{E}\left[exp(-\sum_{j\in[J]} exp(\alpha_j^{(k)}))\right] \approx \frac{1}{N}\sum_{i=1}^{N} exp(-\sum_{j\in[J]} exp(\beta_{ij}^{(k)}))$
- $I_2^{(k)} = \mathbb{E}\left[exp\left(-\sum_{j\in[J]}(P-1)exp(\alpha_j^{(k)}) - \alpha_j^{(k)}c_j\right)\right]$
  $\approx \frac{1}{N}\sum_{i=1}^{N} exp\left(-\sum_{j\in[J]}(P-1)exp(\beta_{ij}^{(k)}) - \beta_{ij}^{(k)}c_j\right)$
- $I_3^{(k)} = \mathbb{E}\left[exp\left(-\sum_{j\in[J]} Pexp(\alpha_j^{(k)}) - \alpha_j^{(k)}c_j\right)\right] \approx \frac{1}{N}\sum_{i=1}^{N} exp\left(-\sum_{j\in[J]} Pexp(\beta_{ij}^{(k)}) - \beta_{ij}^{(k)}c_j\right)$

Nous avons réussi à approximer ces trois intégrales et nous suffirait donc de remplacer dans la fonction (3.4) qui pourra ainsi être evaluée.

4

Simulations numériques et apprentissage du modèle

Dans le modèle LCM2 on estimera notre paramètre $\theta = \left(q_k, \Sigma^{(k)}\right)_{k\in[K]}$ en partant de N_{sample} compagnes publicitaires antérieures données sous forme $\left(\left(c_j^{(t)}\right)_{j\in[J]}, \hat{R}_t\right)_{t=1}^{N_{sample}}$, Pour cela nous allons minimiser la fonction introduite dans (3.2) avec des méthodes numériques où R_θ est notre reach curve dans le modèle LCM2.

4.1 Reach Curve et stabilité numérique :

Remarquons que dans l'expression de la reach curve (3.4) les termes I_1, I_2 et I_3 sont très petits et sont souvent approchés par 0 par la machine, ce qui cause beaucoup de problèmes dans la fraction car une fois les termes approchés par 0, non seulement on aura des problèmes sur la vraie évaluation mais la machine essaierait aussi d'évaluer $\frac{0}{0}$ ce qui affichera pleins d'erreurs.

Ce problème est du à la composition de deux exponentielles, ce qui devient très grand et qui produit un nombre très petit en inverse, pour y remédier nous allons utiliser la logsumexp_trick définie comme suit :

$$log \sum_i exp(\gamma_i) = \gamma_* + log \sum_i exp(\gamma_i - \gamma_*) \tag{4.1}$$

Où $\gamma_* = \max_i \gamma_i$, cette formule est particulièrement intéressante, en effet, pour des γ_i très grand $exp(\gamma_i)$ explose et la machine ne serait pas le calculer, sauf qu'avec cette formule il serait facile pour une machine de calculer $exp(\gamma_i - \gamma_*)$ car la quantité $(\gamma_i - \gamma_*)$ n'est pas grande, ce qui est donc stable numériquement. Nous allons donc exprimer $\log(N \sum_{k=1}^K I_1^{(k)})$, $\log(N \sum_{k=1}^K I_2^{(k)})$ et $\log(N \sum_{k=1}^K I_3^{(k)})$ sous cette forme.

$$\begin{aligned}
\log(N\sum_{k=1}^{K} I_1^{(k)}) &= \log\sum_{k=1}^{K} q_k \sum_{i=1}^{N} exp\left(-\sum_{j\in[J]} exp(\beta_{ij}^{(k)})\right) \\
&= \log\sum_{k=1}^{K}\sum_{i=1}^{N} \; q_k \; exp\left(-\sum_{j\in[J]} exp(\beta_{ij}^{(k)})\right) \\
&= \log\sum_{k=1}^{K}\sum_{i=1}^{N} exp\left(\log(q_k) - \sum_{j\in[J]} exp(\beta_{ij}^{(k)})\right) \\
&= \log\sum_{k=1}^{K} exp\left(\log\sum_{i=1}^{N} exp\left(\log(q_k) - \sum_{j\in[J]} exp(\beta_{ij}^{(k)})\right)\right)
\end{aligned}$$

On remarque que la forme appraît deux fois, il nous suffira donc d'appliquer deux fois la formule (4.1), on obtiendra alors :

$$\log(N\sum_{k=1}^{K} I_1^{(k)}) = \gamma_* + log\sum_{k=1}^{K} exp(\gamma_k - \gamma_*)$$

où $\gamma_k = \lambda_* + log\sum_{i=1}^{N} exp(\lambda_i - \lambda_*)$ et $\gamma_* = \max_k \gamma_k$
avec $\lambda_i = \log(q_k) - \sum_{j\in[J]} exp(\beta_{ij}^{(k)})$ et $\lambda_* = \max_i \lambda_i$

De la même façon :

$$\begin{aligned}
\log(N\sum_{k=1}^{K} I_2^{(k)}) &= \log\sum_{k=1}^{K} q_k \sum_{i=1}^{N} exp\left(-\sum_{j\in[J]} (P-1)exp(\beta_{ij}^{(k)}) \; - \beta_{ij}^{(k)} c_j\right) \\
&= \log\sum_{k=1}^{K}\sum_{i=1}^{N} \; q_k \; exp\left(-\sum_{j\in[J]} (P-1)exp(\beta_{ij}^{(k)}) \; - \beta_{ij}^{(k)} c_j\right) \\
&= \log\sum_{k=1}^{K}\sum_{i=1}^{N} exp\left(\log(q_k) - \sum_{j\in[J]} (P-1)exp(\beta_{ij}^{(k)}) \; - \beta_{ij}^{(k)} c_j\right) \\
&= \log\sum_{k=1}^{K} exp\left(\log\sum_{i=1}^{N} exp\left(\log(q_k) - \sum_{j\in[J]} (P-1)exp(\beta_{ij}^{(k)}) \; - \beta_{ij}^{(k)} c_j\right)\right)
\end{aligned}$$

En appliquant deux fois la formule (4.1), on obtient :

$$\log(N\sum_{k=1}^{K} I_2^{(k)}) = \gamma_* + log\sum_{k=1}^{K} exp(\gamma_k - \gamma_*)$$

où $\gamma_k = \lambda_* + log\sum_{i=1}^{N} exp(\lambda_i - \lambda_*)$ et $\gamma_* = \max_k \gamma_k$
avec $\lambda_i = \log(q_k) - \sum_{j\in[J]} (P-1)exp(\beta_{ij}^{(k)}) \; - \beta_{ij}^{(k)} c_j$ et $\lambda_* = \max_i \lambda_i$

De la même sorte égalemment :

$$\begin{aligned}
\log(N\sum_{k=1}^{K} I_3^{(k)}) &= \log\sum_{k=1}^{K} q_k \sum_{i=1}^{N} exp\left(-\sum_{j\in[J]} Pexp(\beta_{ij}^{(k)}) \ -\beta_{ij}^{(k)} c_j\right) \\
&= \log\sum_{k=1}^{K}\sum_{i=1}^{N} \ q_k \ exp\left(-\sum_{j\in[J]} Pexp(\beta_{ij}^{(k)}) \ -\beta_{ij}^{(k)} c_j\right) \\
&= \log\sum_{k=1}^{K}\sum_{i=1}^{N} exp\left(\log(q_k) - \sum_{j\in[J]} Pexp(\beta_{ij}^{(k)}) \ -\beta_{ij}^{(k)} c_j\right) \\
&= \log\sum_{k=1}^{K} exp\left(\log\sum_{i=1}^{N} exp\left(\log(q_k) - \sum_{j\in[J]} Pexp(\beta_{ij}^{(k)}) \ -\beta_{ij}^{(k)} c_j\right)\right)
\end{aligned}$$

En appliquant deux fois la formule (4.1), on obtient :

$$\log(N\sum_{k=1}^{K} I_3^{(k)}) = \gamma_* + log\sum_{k=1}^{K} exp(\gamma_k - \gamma_*)$$

où $\gamma_k = \lambda_* + log\sum_{i=1}^{N} exp(\lambda_i - \lambda_*)$ et $\gamma_* = \max_k \gamma_k$
avec $\lambda_i = \log(q_k) - \sum_{j\in[J]} Pexp(\beta_{ij}^{(k)}) \ -\beta_{ij}^{(k)} c_j$ et $\lambda_* = \max_i \lambda_i$

Notre reach curve dans le modèle LCM2 devient alors :

$$\begin{aligned}
R_\theta\left((cj)_{j=1}^{J}\right) &= 1 - (\frac{P-1}{P})^{\sum_{j\in[J]} c_j} \quad \frac{\sum_{k\in[K]} q_k I_1^{(k)} \quad \sum_{k\in[K]} q_k I_2^{(k)}}{\sum_{k\in[K]} q_k I_3^{(k)}} \\
&= 1 - (\frac{P-1}{P})^{\sum_{j\in[J]} c_j} \quad \frac{1}{N}\frac{N\sum_{k\in[K]} q_k I_1^{(k)} \quad N\sum_{k\in[K]} q_k I_2^{(k)}}{N\sum_{k\in[K]} q_k I_3^{(k)}} \\
&= 1 - (\frac{P-1}{P})^{\sum_{j\in[J]} c_j} \quad \frac{1}{N}\exp\left(\log\frac{N\sum_{k\in[K]} q_k I_1^{(k)} \quad N\sum_{k\in[K]} q_k I_2^{(k)}}{N\sum_{k\in[K]} q_k I_3^{(k)}}\right) \\
&= 1 - (\frac{P-1}{P})^{\sum_{j\in[J]} c_j} \quad \frac{1}{N}\exp\left(\log(N\sum_{k=1}^{K} I_1^{(k)}) + \log(N\sum_{k=1}^{K} I_2^{(k)}) - \log(N\sum_{k=1}^{K} I_3^{(k)})\right)
\end{aligned}$$

Il nous suffirait donc de remplacer $\log(N\sum_{k=1}^{K} I_1^{(k)})$, $\log(N\sum_{k=1}^{K} I_2^{(k)})$ et $\log(N\sum_{k=1}^{K} I_3^{(k)})$ par les quantités calculées en haut, ainsi notre fonction serait stable numériquement.

4.2 Apprentissage du modèle

Pour la minimisation de la fonction (3.2) on utilisera l'algorithme de la descente de gradient, pour cela on doit calculer le gradient de la fonction $\nabla_\theta loss$, seulement le calcul de ce gradient s'avère

long (une petite erreur de clacul conduirait à un résultat complètement faux) et coûteux pour la machine, nous choisiront alors d'utiliser le framework Pytorch qui est un outil puissant et optimal pour l'auto-differentiation et nous pouvant aussi par la même occasion utiliser un GPU avec ce framework pour augmenter la puissance de calcul et ainsi réduire le temps de l'apprentissage.

4.3 Algorithmes

4.3.1 Génération du jeux de données

Nous allons utiliser le modèle LCM2 pour la génération de N_{sample} matrices N_{uj}

Algorithm 1 data_LCM2

Input : K, J, P, N_{sample}, q, Σ
for u = 1 to P **do**
  $Z[u] \leftarrow k \in [K]$ tel que pour tout $u \in [P]$ $\mathbb{P}(Z[u] = k) = q_k$
for k = 1 to K **do**
  $\alpha^{(k)} \sim \mathcal{N}(0, \Sigma^{(k)})$
for i = 1 to N_{sample} **do**
  $N \leftarrow 0_{P \times J}$
  for u = 1 to P **do**
    $k \leftarrow Z[u]$
    for j = 1 to J **do**
      $Nuj \sim Poisson(exp(\alpha_j^{(k)}))$
  $Tableaux[i] \leftarrow N$
Return : Tableaux

4.3.2 Calcul du reach

Calculer le reach à partire d'une matrice N_{uj} donnée

Algorithm 2 calcul reach

Input : N_{uj}
$Reach \leftarrow 0$
for u = 1 to P **do**
  $sum \leftarrow 0$
  for j = 1 to J **do**
    $sum \leftarrow sum + N_{uj}$
  if $sum > 0$ **then**
    $Reach \leftarrow Reach + 1$
Return : Reach

4.3.3 Calcul des c_j

Comme avant nous allons calculer $(c_j)_{j=1}^{J}$ à partire d'une matrice N_{uj} donnée

Algorithm 3 calcul cj

Input : N_{uj}
$C \leftarrow 0_J$
for j = 1 to J **do**
  $sum \leftarrow 0$
  for u = 1 to P **do**
    $sum \leftarrow sum + N_{uj}$
  $C[j] \leftarrow sum$
Return : C

4.3.4 Vecteurs gaussiens

A partir d'une suite de sobol U donnée de taille $N \times J$ nous allons générer N vecteurs gaussien de moyenne m et de covariance Σ à discrépance faible.

Algorithm 4 Vecteurs_gaussiens

Input : U, m, Σ
$Z \leftarrow 0_{N \times J}$
for u = 1 to P **do**
  $Z[i] \leftarrow \phi^{-1}(U[i])$
$L \leftarrow$ Cholesky(Σ)
Return : m + $ZL^{\intercal}$

4.3.5 Logsumexp_trick

Pour un vecteur V donné, nous allons calculer $\log \sum_i exp(V_i)$ d'une façon à garder la stabilité numérique.

Algorithm 5 Logsumexp_trick

Input : V
$V_* \leftarrow \max_i V_i$
$sumexp \leftarrow 0$
for i = 1 to Size(V) **do**
  $sumexp \leftarrow sumexp + \exp(V_i - V_*)$
Return : $V_* + \log(sumexp)$

4.3.6 Reach Curve

Nous allons écrire une fonction qui évaluera la reach curve définie par :

$$R_\theta\left((cj)_{j=1}^J\right) = 1 - (\frac{P-1}{P})^{\sum_{j\in[J]} c_j} \quad \frac{1}{N}\exp\left(\log(N\sum_{k=1}^K I_1^{(k)}) + \log(N\sum_{k=1}^K I_2^{(k)}) - \log(N\sum_{k=1}^K I_3^{(k)})\right)$$

tout en gardant la stabilité numérique et en vectorisant le code.

Algorithm 6 reach_curve

Input : P, J, K, C, q, Σ
$N \leftarrow 2^{10} = 1024$
$U \leftarrow Sobol(N * J)$
$sum\_k1 \leftarrow 0_K$
$sum\_k2 \leftarrow 0_K$
$sum\_k3 \leftarrow 0_K$
$\leftarrow$
for k = 1 to K **do**
  $\alpha \leftarrow$ Vecteurs_gaussiens$(U, 0, \Sigma^{(k)})$
  $sum\_j1 \leftarrow 0_J$
  $exp\_a\_sum \leftarrow 0_J$
  for 1 = 1 to N **do**
    for j = 1 to J **do**
      $sum\_j1[j] \leftarrow sum\_j1[j] + \alpha[i][j] * C[j]$
      $exp\_a\_sum[j] \leftarrow exp\_a\_sum[j] + exp(\alpha[i][j])$
  $sum\_j2 \leftarrow P * exp\_a\_sum$
  $sum\_j3 \leftarrow (P-1) * exp\_a\_sum$
  $sum\_j4 \leftarrow exp\_a\_sum$
  $sum\_i3 \leftarrow \log(q[k]) + sum\_j1 - sum\_j2$
  $sum\_i2 \leftarrow \log(q[k]) + sum\_j1 - sum\_j3$
  $sum\_i1 \leftarrow \log(q[k]) - sum\_j4$
  $sum\_k3[k] \leftarrow$ Logsumexp_trick(sum_i3)
  $sum\_k2[k] \leftarrow$ Logsumexp_trick(sum_i2)
  $sum\_k1[k] \leftarrow$ Logsumexp_trick(sum_i1)
$log1 \leftarrow$ Logsumexp_trick(sum_k1)
$log2 \leftarrow$ Logsumexp_trick(sum_k2)
$log3 \leftarrow$ Logsumexp_trick(sum_k3)

Return : $1 - (\frac{P-1}{P})^{\sum_{j\in[J]} c_j} \quad \frac{1}{N}\exp\left(log1 + log2 - log3\right)$

4.3.7 Descente de gradient

Nous allons minimiser la fonction de coût définie dans (3.2) en utilisant l'algorithme de la descente du gradient et trouver :

$$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N_{sample}} \left(\hat{R}_i - R_\theta \left(\left(c_j^{(t)} \right)_{j \in [J]} \right) \right)^2$$

Algorithm 7 Gradient descent

Input : K, J, P, N_{sample}, $\theta = \left((q_k)_{init}, (\Sigma^{(k)})_{init}\right)_{k \in [K]}$, $(R^{(t)}, C^{(t)})_{t=1}^{N_{sample}}$, epochs, lr

for epoch = 1 to epochs **do**

$Reachs\_preds \leftarrow (\text{reach\_curve}(P, J, K, C^{(t)}, \theta))_{t=1}^{N_{sample}}$

$loss \leftarrow \sum_{t=1}^{N_{sample}} (Reachs\_preds^{(t)} - R^{(t)})^2$

$\theta \leftarrow \theta - lr * \nabla_\theta loss$

Return : θ

5

Conclusion et travail restant à faire

Au cours de ce stage, j'ai pu appliquer l'algorithme de descente de gradient et ainsi trouver le bon paramètre θ de notre modèle, cependant, je n'ai pu effectuer des tâches importantes qui sont :

- Appliquer l'algorithme de Nelder-Mead et celui de l'optimisation bayésienne et comparer l'éfficacité des trois méthodes.
- Tester l'algorithme sur des données réelles.

Ce stage fut une expérience très enrichissante tant sur le plan professionnel que sur le plan personnel. J'ai pu apprendre aux côtés du Professeur Guillaume Lecué de nombreux concepts de modélisation, calcul probabiliste, approximation et stabilité numérique.
J'ai particulièrement apprécié la démarche rigoureuse menée durant ce stage dans lequel j'ai développé grâce aux précieux conseils du Professeur Guillaume lecue, un esprit critique ainsi qu'une grande persévérance dans la résolution des problèmes rencontrés. Cette expérience m'a permis d'appliquer mes connaissances théoriques acquises pendant mon parcours scolaire.

Pour conclure, ces 3 mois de stage m'ont permis de mieux définir le poste vers lequel je souhaiterais me tourner lors de mon parcours professionnel.

Références

[KSML16] Jim Koehler, Evgeny Skvortsov, Sheng Ma, and Song Liu. Measuring cross-device online audiences. Technical report, Google, Inc., 2016. 1

[KSV13] Jim Koehler, Evgeny Skvortsov, and Wiesner Vos. A method for measuring online audiences. Technical report, Google Inc, 2013. 1

[SK19] Evgeny Skvortsov and Jim Koehler. Virtual people : Actionable reach modeling. Technical report, 2019. 1