# COMPOSITION , FACTORISATION ET CONDITIONS D'OPTIMALITE (FAIBLE, FORTE) DANS LA METHODE DE PARTAGE DES POIDS ; APPLICATION A L'ENQUETE SUR LE TOURISME EN BRETAGNE.

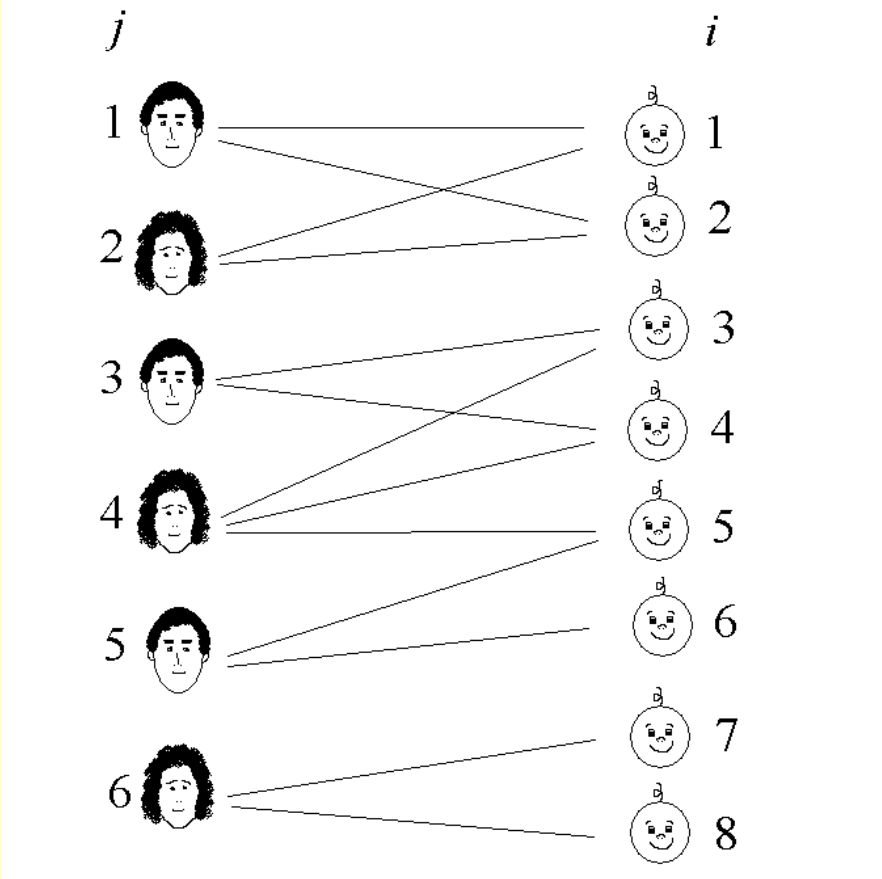
Jean-Claude DEVILLE (CREST/ENSAI) , Pierre LAVALLEE (Statistique Canada) et Myriam MAUMY (Université de Strasbourg).

# La méthode de partage des poids (ou échantillonnage indirect)

Deux populations *A* et *B* sont liées par un graphe dont les flèches relient des éléments de *A* à des éléments de *B*. On possède , pour *A* ,une ‘bonne’ base de sondage permettant de faire un ‘bon’ sondage probabiliste conduisant à un système de poids $w^i$ .

En revanche, on ne dispose pas de base pour *B,* et, seule, les flèches du graphe nous livrent un moyen d’attraper des unités *k* de *B*.

**Figure 1: Population *A* de parents et population *B* d'enfants et liens entre elles.**




Grappes particulières: en 'Un pour tous' ou en 'Tous pour un'

Les flèches du graphe sont chargées par des nombres positifs $\theta_{ik}$
Un système de poids sans biais $w_i$ sur A se transforme en un système de poids $w_k$ sur $B$. Matriciellement, si on représente les poids par des vecteurs-lignes $w^A$ et $w^B$ et que la matrice $T$ est celle des $\theta_{ik}$ on pose $w^B = w^A T$. Les poids sur B sont sans biais si et seulement si $\mathbf{1}^A T = \mathbf{1}_B$ où $\mathbf{1}^{\mathbf{A\ resp\ B}}$ est le vecteur ligne de 1 construit sur *A resp B*. Autrement dit la somme en $i$ des à $k$ fixé est égale à 1 pour tout $k$ .

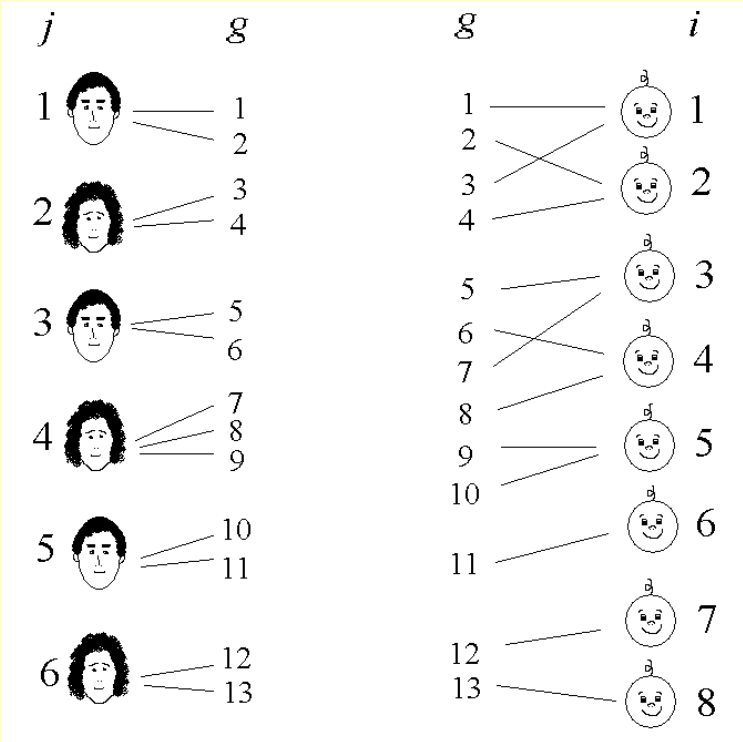
Si on connaît la matrice $D^A$ des variances/covariances sur $A$ ( c'est à dire $Var(w^A)$ ) alors la variance du système de poids sur $B$ vaut $D^B = T' D^A T$ .

# Composition

Si, de plus, $B$ est reliée à une troisième population $C$ par un second système de flèches , la méthode peut être itérée : on définit des charges $\sigma_{km}$ et la matrice $S$ correspondante ayant les mêmes propriétés (en particulier de donner naissance à des poids sans biais, soit $\mathbf{1}^C = \mathbf{1}^B \mathbf{S}$ ).

L'estimateur définit par les poids $w^C = w^B S = w^A TS$ est sans biais et sa variance vaut $D^C = S'T'D^A TS$ .

# Factorisation canonique




Un pour tous  Tous pour un

Dans un graphe ‘Un pour tous’ les charges valent 1:y’a pas le choix.
Toute optimisation des charges ne concerne que la partie ‘Tous pour un’.

# Optimalité faible

On a, si les poids sont sans biais: $E(w_k \,|\, k \in s_B) = 1/\pi_k$

L'optimalité faible consiste à minimiser la variance conditionnelle, Soit:

$$\theta_k^{'} \; D_{kk}^{G} \theta_k$$

sous la contrainte $\boldsymbol{1_k \theta_k} = 1$.

où $\theta_k$ est le vecteur des $\theta_{gk}$ cherché

$D_{kk}^{G}$ la sous-matrice de variance de $G$ correspondant aux lignes et colonnes qui ont des indices $g$ pointant sur $k$

et $\boldsymbol{1_k}$ le vecteur ligne de 1 ayant la dimension du nombre de flèches pointant sur $k$

Solution: $\theta_k = \lambda_{kk} D_{kk}^{G\,-1} 1_k$ avec

$$\lambda_{kk} = (1_k' D_{kk}^{G\,-1} 1_k)^{-1} = \theta_k' D_{kk}^{G} \theta_k = Var(w_k).$$

Remarque 1:
C'est la minimisation de la variance d'estimation d'une variable 'bizarre' qui vaut 1 chez monsieur $k$ et zéro ailleurs.

Remarque 2:
C'est assez facile à appliquer, même si on va voir qu'on peut souvent appliquer une optimisation approchée TRES SIMPLE.

# Optimalité forte

Pour une variable d'intérêt quelconque, $y$ , on cherche à minimiser la variance :

$$\sum_k \sum_l y_k y_l \theta_k^{'} D_{kl}^G \theta_l$$

sous les contraintes $\mathbf{1}_k \boldsymbol{\theta}_k = 1$ pour tout $k$.

où $D_{kl}^G$ est la matrice extraite de $D^G$ dont les lignes correspondent aux flèches qui pointent vers $k$, les colonnes à celles qui pointent vers $l$.

D'où les relations (encore Lagrange!)

$$y_k \sum_l D_{kl}^G \theta_l y_l = \lambda_k^y \mathbf{1}_l{}'$$

La solution dépend de $y$.
Il y a optimalité forte quand ce n'est plus le cas.

On obtient les conditions nécessaires et suffisantes suivantes pour l'optimalité forte:

$$D^G_{kl}\theta_l = \phi^{kl}\mathbf{1}_k$$

(valides aussi si $k=l$ ) =optimalité faible

Soit $\Phi$ la matrice d'éléments $\Phi_{ii'}$ .
La variance optimale est donnée par $\mathbf{Y}'\Phi\mathbf{Y}$ , où $\mathbf{Y}$ est le vecteur des $y_k$ .

En pratique, donc , on cherchera l'optimalité faible qui est facile à vérifier et nécessaire à la forte.

# APPLICATION A L'ENQUETE SUR LE TOURISME EN BRETAGNE

Pour attraper des touristes en Bretagne on se sert de plusieurs sondages relatifs à leurs activités qui se ramène au schéma suivant :

-Un sondage sur les visites de certains sites (dont le péage de La Gravelle !).On réalise un sondage à deux degrés par site : premier degré le jour d'enquête, deuxième degré :le ménage touristique.

-Un sondage à trois degrés sur les achats en boulangerie ( premier degré :la boulangerie, puis le jour, puis le ménage).

Chacun de nos touristes $k$ est attrapé par un seul de ces dispositifs $d$, ce qui lui confère un 'poids de base' $w_d$ . Il a effectué $n_b$ achats en boulangerie, et visité certains des sites sélectionnés. Soit $L_k$ la liste des $n_k$ occurrences où notre ménage touristique aurait pu être attrapé. Nous devons construire la matrice $n_k x\ n_k$ des covariances entre ces occurrences . Or les dispositifs sont indépendants, et donc les éléments de matrices hors diagonale correspondant sont nuls. Sur la diagonale, nous avons une quantité du type $\frac{\pi_i - \pi_i^2}{\pi_i^2}$ connue par le plan dans le dispositif correspondant .

Pour les $nb\ x\ nb$ entrées venant des boulangeries, la partie diagonale s'obtient de la même façon. Comme $nb$ est très petit devant le nombre total d'achats de la 'base de sondage', les éléments hors diagonale sont négligeables devant les éléments diagonaux. Notre matrice est donc très voisine d'une matrice diagonale d'éléments :

$$\frac{\pi_i - \pi_i^2}{\pi_i^2} \cong \frac{1}{\pi_i}$$

Et donc, approximativement:

$$\theta_{ik} = \frac{\pi_i}{\sum_{l \in L_k} \pi_l}$$

Le poids du ménage touristique sera donc égal à

$$w_d \ \theta_{dk} = \frac{1}{\sum_{l \in L_k} \pi_l} \quad .$$

Ce résultat est,somme toute, assez naturel. En particulier , si un touriste ne déclare qu'une seule occurrence ayant permis de l'attraper, son poids est celui d'Horvitz-Thompson .

Exemple: 4 achats en boulange et deux sites:

$$D_{kk} = \begin{pmatrix} \frac{N}{n}(1-\frac{n}{N}) & \frac{(n-1)N}{n(N-1)} & \frac{(n-1)N}{n(N-1)} & \frac{(n-1)N}{n(N-1)} & 0 & 0 \\ \frac{(n-1)N}{n(N-1)} & \frac{N}{n}(1-\frac{n}{N}) & \frac{(n-1)N}{n(n-1)} & \frac{(n-1)N}{n(N-1)} & 0 & 0 \\ \frac{(n-1)N}{n(N-1)} & \frac{(n-1)N}{n(N-1)} & \frac{N}{n}(1-\frac{n}{N}) & \frac{(n-1)N}{n(N-1)} & 0 & 0 \\ \frac{(n-1)N}{n(N-1)} & \frac{(n-1)N}{n(N-1)} & \frac{(n-1)N}{n(n-1)} & \frac{N}{n}(1-\frac{n}{N}) & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{N_1}{n_1}(1-\frac{n_1}{N_1}) & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{N_2}{n_2}(1-\frac{n_2}{N_2}) \end{pmatrix}$$

$$\cong \begin{pmatrix} N/n & 1 & 1 & 1 & 0 & 0 \\ 1 & N/n & 1 & 1 & 0 & 0 \\ 1 & 1 & N/n & 1 & 0 & 0 \\ 1 & 1 & 1 & N/n & 0 & 0 \\ 0 & 0 & 0 & 0 & N_1/n_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & N_2/n_2 \end{pmatrix}$$

On doit résoudre $D_{kk}\theta_k=\mathbf{1_k}$ à un facteur près pour que la somme des éléments du vecteur $\theta_k$ fasse 1, ce qui nous donne approximativement $(n/N)/S$ pour les 4 premières coordonnées, puis $(n_1/N_1)/S$ et $(n_2/N_2)/S$ pour les deux dernières, avec $S=4*n/N+n_1/N_1+n_2/N_2$.

Où que notre client ait été attrapé, il recevra donc le poids $1/S$.

Et voilà le travail…..

Merci de votre patience!