Approches graphiques pour la modélisation statistique de la dépendance entre activités journalières

J.-B. Durand[†] P. Fernique[⋆]

Inria ([†]Mistis, Grenoble; [⋆]Virtual Plants, Montpellier);
[†]Grenoble Inp; [⋆]CIRAD, Université Montpellier 2

5 novembre 2013

Enjeux

Analyse de données : activités et transports.

- Analyser et modéliser la dépendance entre les motifs de trajet d'une même personne au cours d'une journée. *Est-ce que faire 2 trajets pour des courses exclut de faire plus de 2 trajets pour des loisirs ?*
- Dépendance entre activités et modes de transport. *Est-ce que la voiture est privilégiée si on doit aller au travail et faire les courses ?*
- Dépendance entre activités et mode de transport pour des personnes d'un même foyer. *La voiture est-elle plus utilisée par l'un des membres d'un couple si les deux vont travailler ?*
- Effet de l'âge, du sexe, ... sur le nombre d'activités de chaque type réalisées et les modes de transport utilisés.
- Effet de l'âge, du sexe, ... sur la dépendance entre le nombre d'activités de chaque type réalisées et les modes de transport utilisés.

Motifs possibles :
1 (domicile), 2 (travail), ..., 9 (autres), 10 (tournée professionnelle).

- $N_i^{(p)}$: nombre de trajets effectués en une journée par la personne p au motif i.
- Modélisation du vecteur $(N_1^{(p)}, \ldots, N_{10}^{(p)})$: **loi, dépendances**, ...
- Données de comptage multivariées (variables discrètes).
- Si on s'intéresse au temps $T_i^{(p)}$ associé à un motif de trajet plutôt qu'au nombre d'occurrences, on a des données multivariées continues $(T_1^{(p)}, \ldots, T_{10}^{(p)})$.
- Exemple : *Est-ce que faire 2 trajets pour des courses exclut de faire plus de 2 trajets pour des loisirs ?*

Motifs possibles :
1 (domicile), 2 (travail), ..., 10 (tournée professionnelle).
Modes possibles :
A (marche), B (vélo), ..., E (voiture).

- Modélisation du vecteur de comptages $(N_1^{(p)}, \ldots, N_{10}^{(p)}, N_A^{(p)}, \ldots, N_E^{(p)})$.
- Variante : vecteur des temps $(T_1^{(p)}, \ldots, T_{10}^{(p)}, T_A^{(p)}, \ldots, T_E^{(p)})$.
- Exemple : *Est-ce que la voiture est privilégiée si on doit aller au travail et faire les courses ?*

Remarque : en fait dans les données, on sait associer un mode de transport et un motif précis à **chaque** déplacement. On pourrait modéliser le nombre de fois où une personne a pris sa voiture pour aller au travail, le vélo pour aller au travail, le bus pour aller faire les courses, etc. Mais on passerait de $10 + 5 = 15$ à $10 \times 5 = 50$ variables (par personne...).

Motifs possibles : 1 (domicile), 2 (travail), ..., 10 (tournée professionnelle)
Interaction entre les motifs de trajet de deux personnes d'un foyer f.

- Vecteur de comptages $(H_1^{(f)}, \ldots, H_{10}^{(f)})$ pour la personne H.
- Vecteur de comptages $(F_1^{(f)}, \ldots, F_{10}^{(f)})$ pour la personne F.
- Modélisation du vecteur $(H_1^{(f)}, \ldots, H_{10}^{(f)}, F_1^{(f)}, \ldots, F_{10}^{(f)})$ pour le foyer f.
- Variante : vecteur des temps au lieu des nombres d'occurrences.
- Exemple : *Est-ce que si F prend la voiture pour aller au travail, alors H fait les courses avec une probabilité moindre ?*

Approche classique

Par la suite : hypothèse de $(N_1^{(p)}, \ldots, N_{10}^{(p)})$ dépendants dans ce vecteur mais vecteurs indépendants en p (sauf p de même foyer ?), de même loi.
Modèle classique pour $(N_1, \ldots, N_{10})$:

- Régression séparée des N_i.
- Inclusion possible d'autres covariables (âge A, sexe S, ...).

Exemple :

$$P(N_i = n_i | A = a, S = s) = \exp(-\lambda_i(a,s)) \frac{\lambda_i(a,s)^{n_i}}{n_i!}$$

$$\ln \lambda_i(a,s) = \beta_0 a + \beta_s + \varepsilon_i$$

où ε_i est de loi Gamma (indépendant des autres variables).
Pour prendre en compte les interactions entre $(N_1, \ldots, N_{10})$, on peut inclure les $(N_j)_{j=1,\ldots,10; j\neq i}$ dans le modèle :

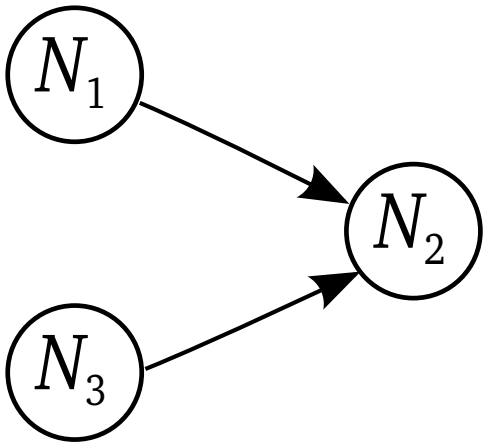
$$\ln \lambda_i(a,s,n_1,\ldots,n_{10}) = \beta_0 a + \beta_s + \sum_{j\neq i} \gamma_j n_j + \varepsilon_i$$

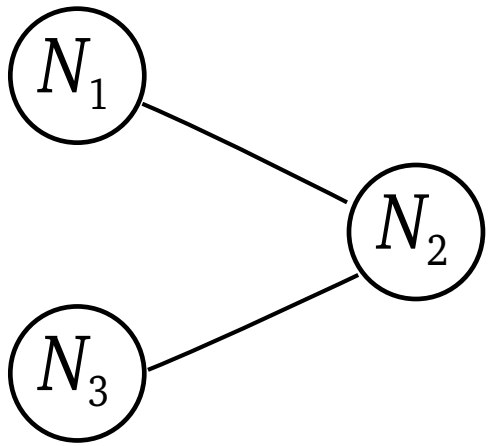
associé à $P\left(N_i = n_i | A = a, S = s, (N_j = n_j)_{j=1,\ldots,10; j\neq i}\right)$.

Limites de l'approche classique

Disposer de trois modèles $P(N_1|A = a, S = s, N_2 = n_2, N_3 = n_3)$, $P(N_2|A = a, S = s, N_1 = n_1, N_3 = n_3)$ et $P(N_3|A = a, S = s, N_1 = n_1, N_2 = n_2)$:

- ne définit pas forcément une loi unique $P(N_1, N_2, N_3|A = a, S = s)$;
- ne permet pas forcément de se rendre compte que N_1 et N_3 sont indépendantes, ou ne sont dépendantes qu'à travers leur dépendance à N_2, etc. ;
- ne permet pas de percevoir globalement les interactions entre ces variables (difficulté d'interprétation).





Modèle probabiliste pour comptages multivariés

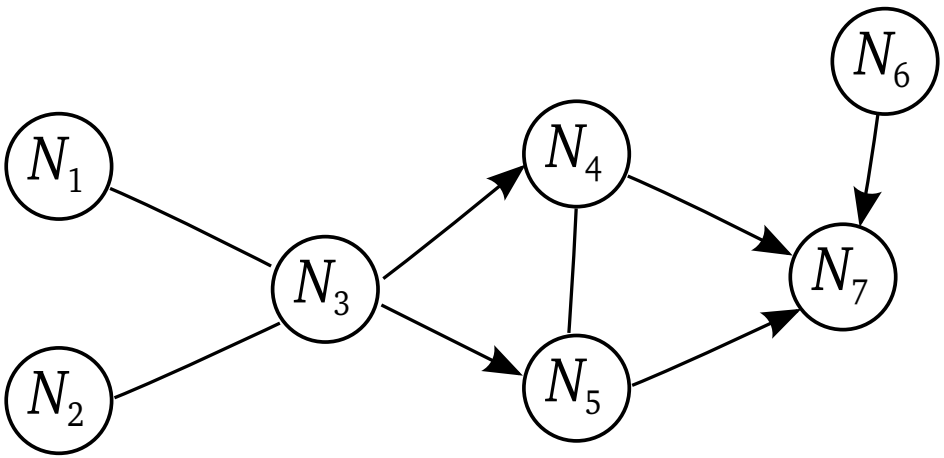
Objectifs :

- définir un modèle probabiliste pour des comptages multivariés permettant de prédire $(N_1, \ldots, N_K)$ (loi jointe) ;
- modèle avec peu de paramètres (si les N_k peuvent valoir $1, 2, \ldots, 15 = M$, éviter d'estimer les probabilités par $M^K = 15^K$ fréquences ;
- modèle permettant une identification simple des indépendances (conditionnelles ou non).

Modèles probabilistes pour comptages multivariés

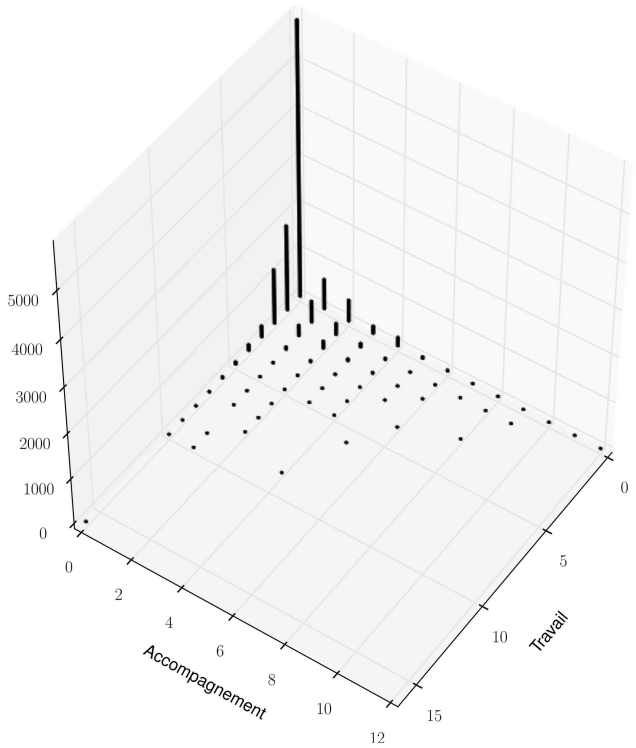
Principe :

- Représenter des relations d'indépendance conditionnelle entre les variables par un graphe partiellement orienté, dont les sommets sont les variables.
- Indépendance conditionnelle : représentée par la séparation par des (groupes de) sommets.
- Indépendance : représentée par l'absence de chemin orienté (+ subtilités...)
- Notations : indépendance $N_1 \perp\!\!\!\perp N_2$, dépendance $N_1 \not\perp\!\!\!\perp N_2$, indépendance conditionnelle $N_1 \perp\!\!\!\perp N_2 | N_3$, ...



$N_6 \perp\!\!\!\perp N_4$
$N_6 \not\perp\!\!\!\perp N_4 | N_7$
$N_1 \not\perp\!\!\!\perp N_2$
$N_1 \perp\!\!\!\perp N_2 | N_3$

Problèmes typiques avec les comptages multivariés



Données de comptage multivariées

1. Les histogrammes sont souvent creux (la plupart des fréquences de la grille sont nulles ou presque).
2. $(0, \ldots, 0)$ (resp. zéro) est une valeur fréquente (resp. fréquente marginalement).
3. Les vecteurs fréquents ont une majorité de composantes nulles (par exemple $(3, 0, 0, 1); (1, 2, 0, 0), \ldots$)

Requiert un modèle paramétrique discret (non-gaussien).

Factorisation induite par le graphe

Pour qu'une loi $P(N_1, \ldots, N_K)$ respecte les relations d'indépendance associée à un graphe donné, il suffit qu'elle se factorise en accord avec le graphe :

$$P\left[N_1 = n_1, ..., N_K = n_K\right] = P[\mathbf{N} = \mathbf{n}]$$
$$= \prod_{\mathcal{C}} P[\mathbf{N}_{\mathcal{C}} = \mathbf{n}_{\mathcal{C}} | \mathbf{N}_{pa(\mathcal{C})} = \mathbf{N}_{pa(\mathcal{C})}]$$

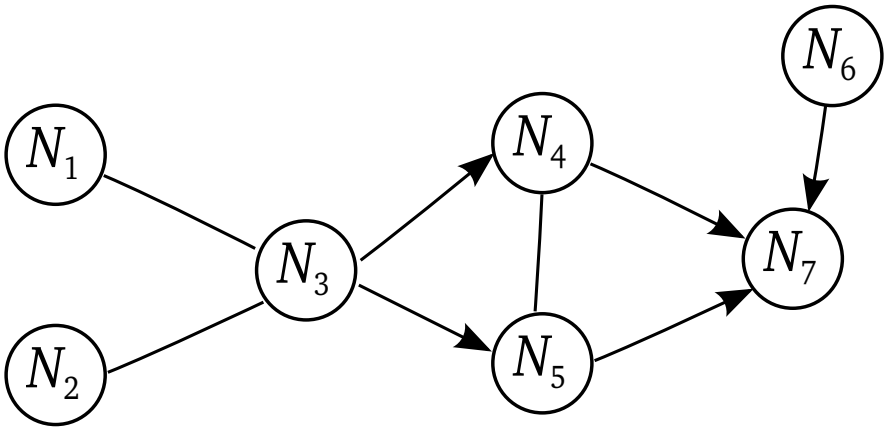
avec :

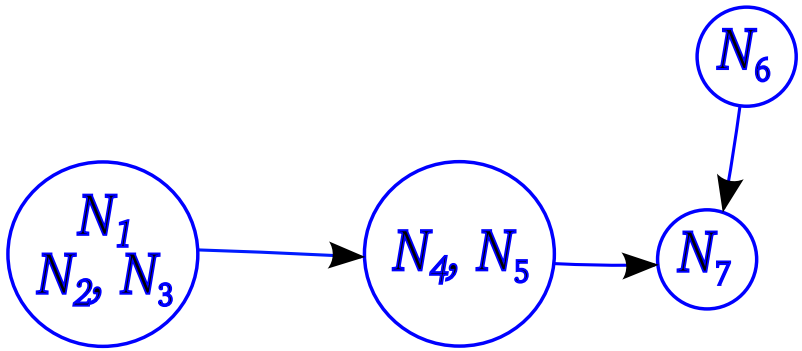
- $\mathcal{C}$ sous-ensemble maximal de sommets du graphe connecté par des arêtes (non-orientées) ;
- $pa(\mathcal{C})$ ensemble des parents de $\mathcal{C}$;
- la loi des variables $\mathbf{N}_{\mathcal{C}}$ de $\mathcal{C}$ se factorise aussi en accord avec le graphe non-orienté $\mathcal{C}$.

Permet de passer d'un nombre de probabilités à estimer de M^K à $CM^{\max(K_1+K_2)}$, où $K_1 + K_2$ représente le nombre maximal de variables dans un ensemble $\mathcal{C}$ avec ses parents $pa(\mathcal{C})$, et C le nombre d'ensembles $\mathcal{C}$.

Exemple de factorisation

$$
\begin{aligned}
&P(N_1, ..., N_7) \\
&= P(N_7|N_4, N_5, N_6)P(N_6)P(N_4, N_5|N_1, N_2, N_3)P(N_1, N_2, N_3) \\
&= P(N_7|N_4, N_5, N_6)P(N_6)P(N_4, N_5|N_1, N_2, N_3)P(N_1, N_3)P(N_2, N_3)/P(N_3)
\end{aligned}
$$





Paramétrisation du modèle

$$P(N_1, ..., N_7)$$
$$= P(N_7|N_4, N_5, N_6)P(N_6)P(N_4, N_5|N_1, N_2, N_3)P(N_1, N_3)P(N_2, N_3)/P(N_3)$$

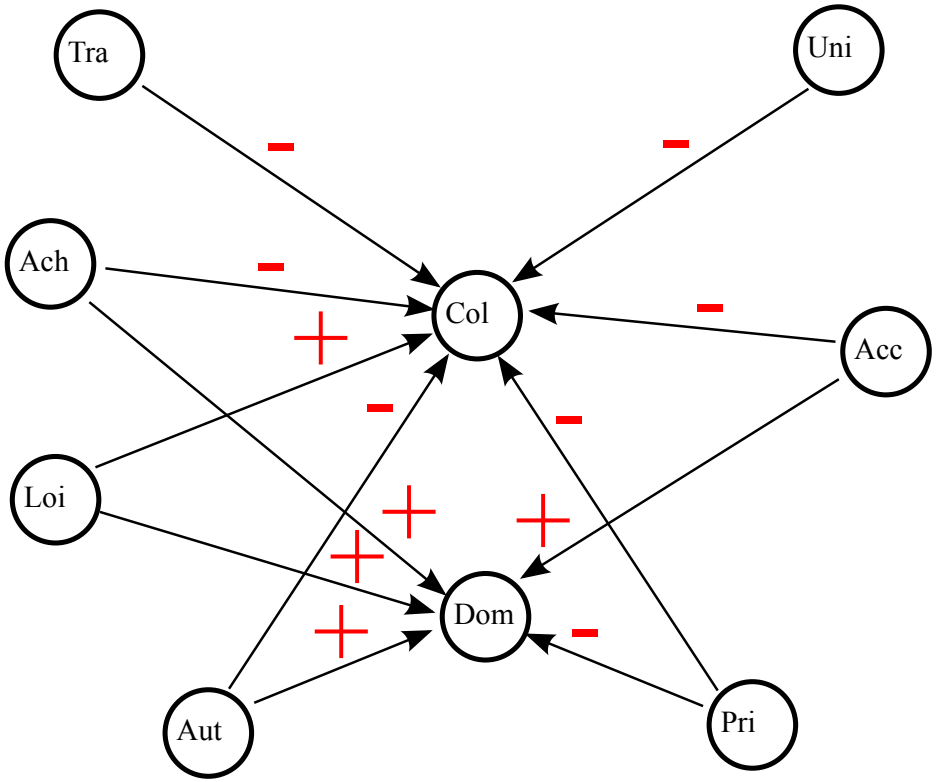
- Nombre de paramètres encore trop important malgré la simplification opérée.
- Exemple précédent : $800\,000$ au lieu de $800\,000 \times 212 = 170\,000\,000$.
- Mettre des lois paramétriques sur les différents facteurs : loi univariée pour $P(N_6)$, loi multivariée pour $P(N_1, N_3)$, régression univariée pour $P(N_7|N_4, N_5, N_6)$, régression multivariée pour $P(N_1, N_3)$.
- On se donne un catalogue de lois paramétriques : Poisson, binomiale, binomiale négative, multinomiale (tronquée, composée, négative), mélanges, régressions analogues...
- Ces lois doivent être compatibles avec le graphe considéré (pas d'hypothèse d'indépendance en trop !).

Démarche complète pour identifier le modèle

- À graphe connu : on teste toutes les familles paramétriques (Poisson, binomiale, ...) et on garde le maximum de vraisemblance.
- À graphe connu : en pratique, on sait trouver un modèle paramétrique et calculer sa vraisemblance (max.) dans la famille de lois choisie.
- Identification du graphe : on part d'un graphe initial, on trouve le meilleur modèle paramétrique associé, on évalue l'ajustement aux données, puis on modifie le graphe itérativement pour en trouver un meilleur (opérations sur les arcs, arêtes, sommets).
- Ajustement aux données mesuré par un critère statistique de sélection de modèle (compromis vraisemblance / nombre de paramètres du modèle) – par ex. BIC.

Programme d'activité individuel

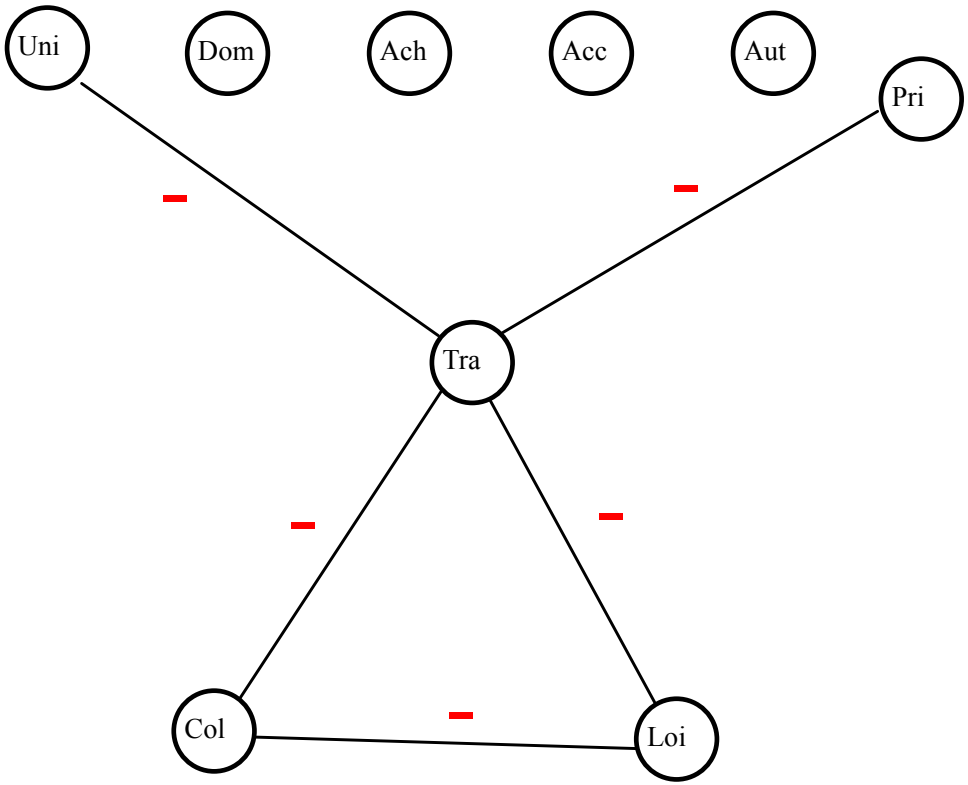
Modèle graphique orienté pour les nombres de déplacements $(N_1, \ldots, N_9)$ (N_{10} quasiment toujours à 0) :

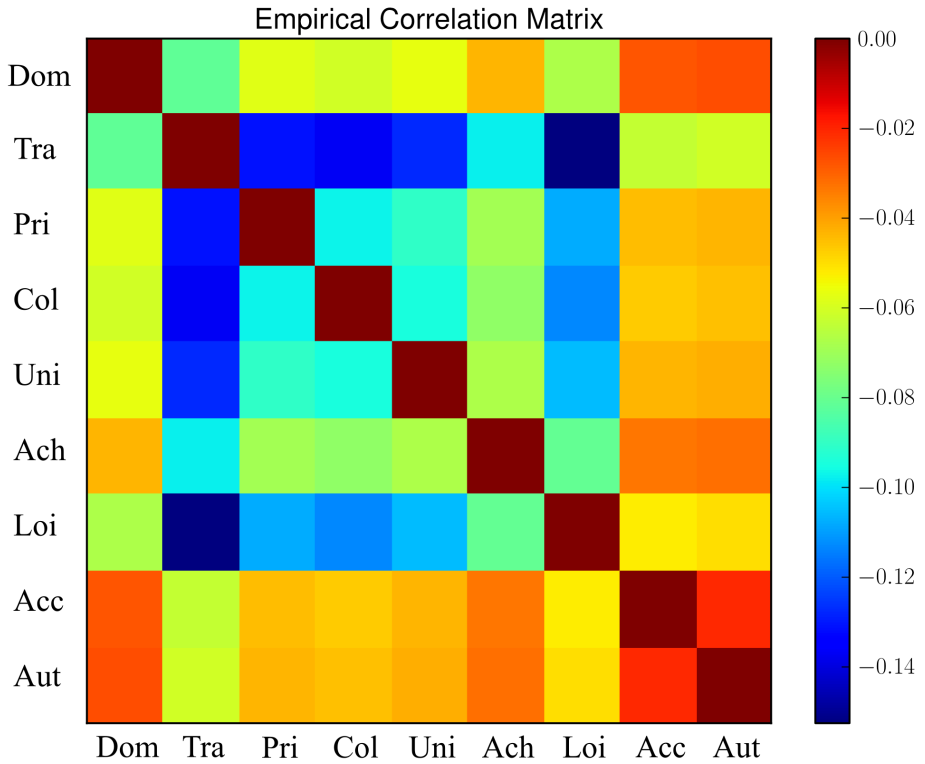


- Indépendance nombre de déplacements pour Tra, Pri, Uni, Ach, Loi, Acc, Aut.
- Dépendance entre ces variables sachant Col.
- Dépendance entre Ach, Loi, Acc, Pri, Aut. via Dom.
- Dépendance entre Col et Dom via Ach, Loi, Acc, Pri, Aut.
- ...

Durées de déplacement – échelle individuelle

Modèle graphique non-orienté pour $(T_1, \ldots, T_9)$:

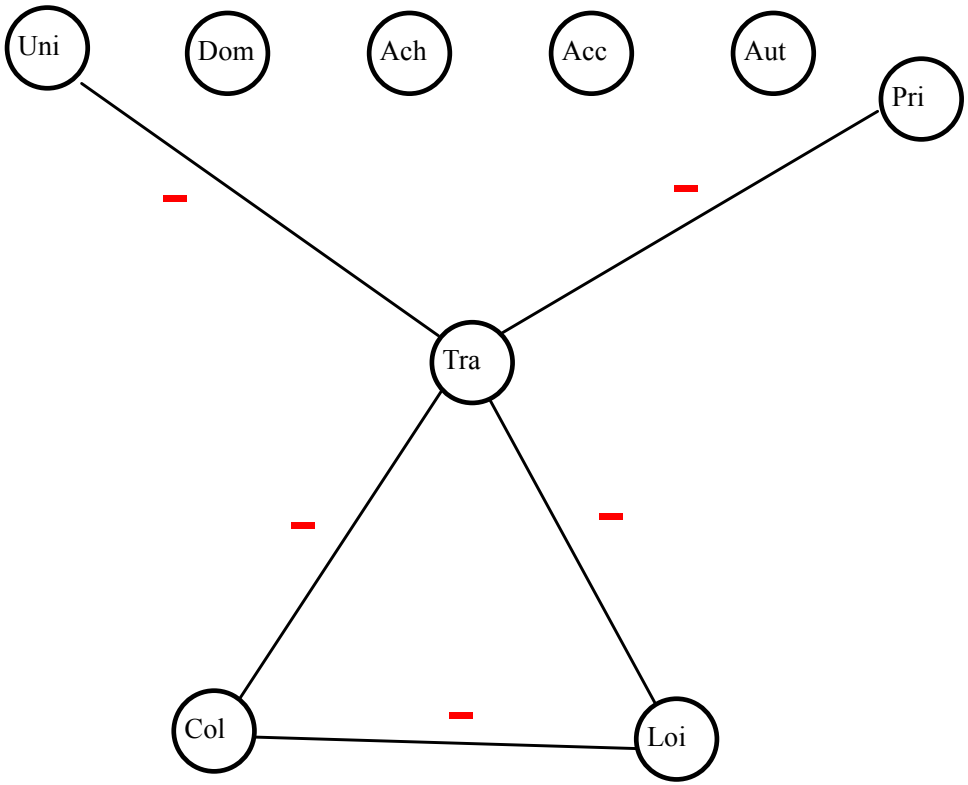


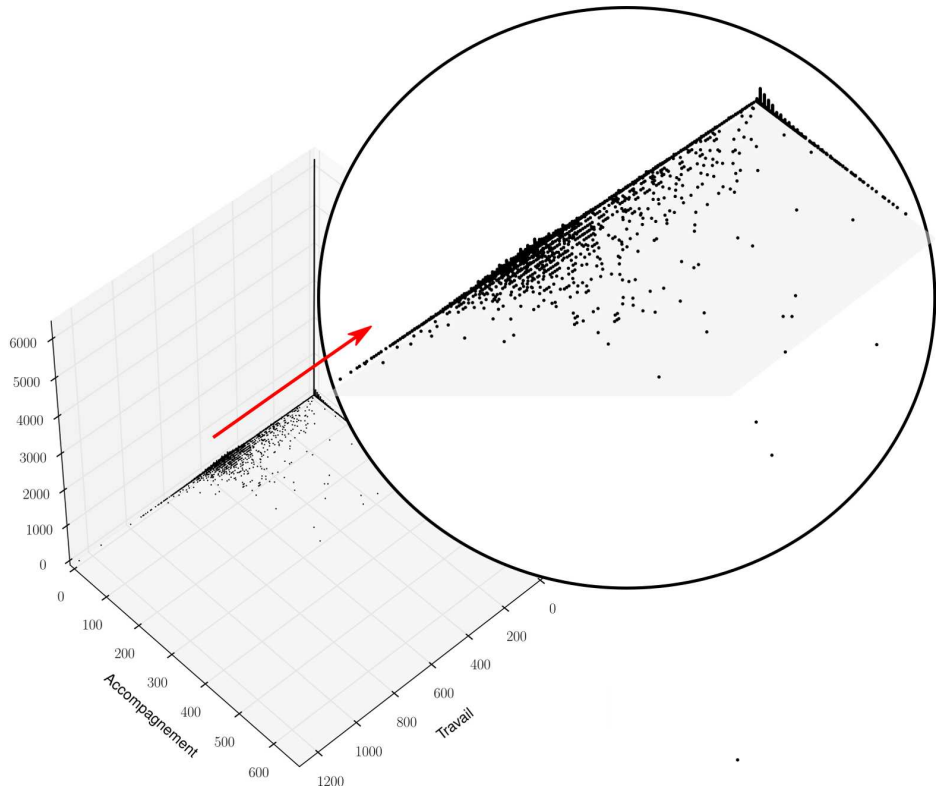


Modèle obtenu sous une hypothèse gaussienne (modélisation de la matrice de covariance inverse).

Durées de déplacement – échelle individuelle

Modèle graphique non-orienté pour $(T_1, \ldots, T_9)$:





Modèle obtenu sous une hypothèse gaussienne (modélisation de la matrice de covariance inverse).

Travaux en cours

- Variables continues T_k : lois non-gaussiennes (ex. gamma).
- Variables discrètes : familles paramétriques plus complexes en termes d'indépendances conditionnelles.
- Prises en compte de covariables.
- Analyse des données à l'échelle du foyer (vs. individu).