Stage d’été – Semaine 3

QGNN-TimeCausality
Avancement du projet

Université de Sherbrooke

UNIVERSITÉ TÉLUQ

01

Avancement dans l’implémentation

Jeu de données partagés par M. Miloud

- Référence financière basée sur les rendements boursiers, extraite de S. Kleinberg (Finance CPT) et prétraitée.

Simulated financial time series

From Causality, Probability, and Time

Basic features 25 portfolios, 10 causal structures, two 4,000 day time periods. There are 20 separate data files (10 structures x 2 time periods). The two time periods allow one to test the robustness of results as the same relationships should be found in both.

Structures simulated 1) no dependency between portfolios, 2-6) 20 random relationships with a lag of 1 time unit, 7) 40 random relationships with a lag of 1 time unit, 8) 20 random relationships with random lags of 1-3 time units, 9) 40 random relationships with random lags of 1-3 time units, and 10) many-to-one relationships at a lag of one time unit

Getting the data The data along with a more detailed README and files with the true embedded structure for each dataset can be downloaded here: [FinanceCPT.tar.gz]

Citation Kleinberg, S. (2012). *Causality, Probability, and Time*. Cambridge University Press.
The data are discussed in depth and analyzed in chapter 7.

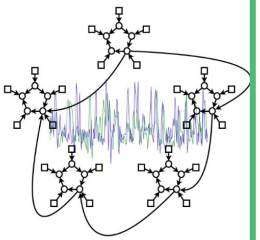
License The data are available under a Creative Commons Attribution-NonCommercial 3.0 Unported License. Basically, you're free to share and adapt it with proper attribution, but the data cannot be used for commercial purposes.

- Référentiel neuroscientifique de données FMRI issues de réseaux cérébraux, extraites de Smith et al. prétraitées.

NetSim - Evaluation of Network Modelling Methods for FMRI

FMRIB Analysis Group, Oxford

Please send feedback to steve@fmrib.ox.ac.uk



There is great interest in estimating brain "networks" from FMRI data. This is often attempted by identifying a set of functional "nodes" (e.g., spatial ROIs or ICA maps) and then conducting a connectivity analysis between the nodes, based on the FMRI timeseries associated with the nodes. Analysis methods range from very simple measures that consider just two nodes at a time (e.g., correlation between two nodes' timeseries) to sophisticated approaches that consider all nodes simultaneously and estimate one global network model (e.g., Bayes net models). Many different methods are being used in the literature, but almost none has been carefully validated or compared *for use on FMRI timeseries data*. In this work we generate rich, realistic simulated FMRI data for a wide range of underlying networks, experimental protocols and problematic confounds in the data, in order to compare different connectivity estimation approaches. Our results show that in general correlation-based approaches can be quite successful, methods based on higher-order statistics are less sensitive, and lag-based approaches perform very poorly. More specifically: there are several methods that can give high *sensitivity to network connection detection* on good quality FMRI data, in particular, partial correlation, regularised inverse covariance estimation and several Bayes net methods; however, accurate estimation of *connection directionality* is more difficult to achieve, though Patel's τ can be reasonably successful. With respect to the various confounds added to the data, the most striking result was that the use of functionally inaccurate ROIs (when defining the network nodes and extracting their associated timeseries) is extremely damaging to network estimation; hence, results derived from inappropriate ROI definition (such as via structural atlases) should be regarded with great caution.

Dataset simulé de séries temporelles financières (tiré de Kleinberg, 2013)

- Ce dataset contient les retours journaliers de **25** portefeuilles sur **4000** jours.
- Il est fourni en deux périodes temporelles : j3000–j7000 et j8000–j12000
- Les relations entre les portefeuilles sont définies dans un fichier séparé .csv selon **différents scénarios** :
  - Pas de causalité (nocause)
  - Causalité aléatoire avec lags fixes (ex: 20 relations avec 1 jour de décalage)
  - Causalité avec lags variables (1 à 3 jours)
  - Plusieurs entrées causales vers une seule sortie (many-to-one)
- Ces relations servent de **ground truth** pour tester les performances d'algorithmes de découverte causale.

📁 Exemple utilisé :

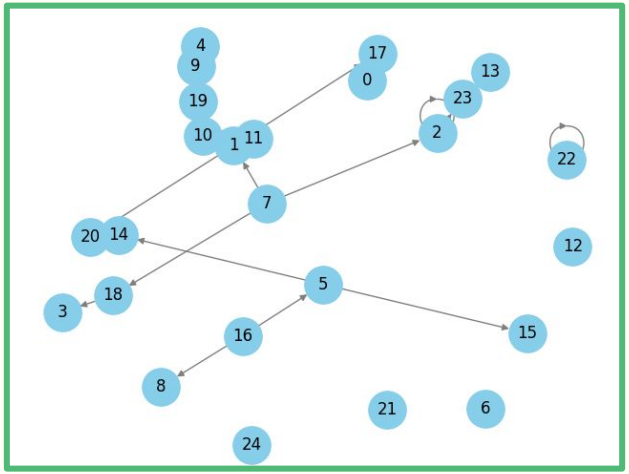
- random-rels_20_1A_returns30007000.csv → données
- random-rels_20_1A.csv → structure causale correspondante

Raisonnement & Objectif

- Chaque **noeud** représente **un portefeuille**.
  - On a 25 Portefeuilles ((P0 à P24)
- Chaque **arête dirigée** encode une **relation causale entre deux portefeuilles**, avec un **décalage temporel (lag)**
  - Les arêtes prennent chaque ligne du fichier ‘rels’
  - une arête va du nœud cause à effect, avec un décalage temporel.

📁 Exemple : cause=5, effect=14, lag=2 → ajoute une arête **de P5(t-2) vers P14(t)**

- Ces graphes sont construits **à chaque instant t (l’indice de jour)**



Démarche & Code

- Étapes :
  - Chargement des séries temporelles dans un dataframe df (dimensions 4000 x 25).
  - Chargement des relations causales sous forme de triplets (cause, effect, lag).
- Fonction principale : **build_graph_at_time_t(t)**
- Pour chaque instant t :
  - Création d'un graphe dirigé G.
  - Ajout des 25 noeuds avec leur valeur de retour à l'instant t comme attribut "feature".
  - Ajout des arêtes selon les relations causales, si le temps t – lag existe.
  - L'arête contient l'information du décalage (lag) et la valeur passée du portefeuille causal.
- Boucle :
  - On parcourt tous les instants à partir de t = max(lag) jusqu'à la fin du dataset.
  - À chaque t, on construit un graphe et on l'ajoute à une liste graphs.
- Visualisation :
  - Le graphe est affiché avec NetworkX, chaque nœud représentant un portefeuille.
  - Cela permet de vérifier visuellement que les connexions sont bien créées.

02

Questions & Prochaines étapes

Questions

- Est-ce que je continue me concentrer sur ce dataset ou je switch vers sur l'autre ?
- Est-ce que je continue avec ce raisonnement ?
- Est-ce que j'étudie les 2 datasets simultanément ou je continue avec seulement 1 pour le moment et j'utilise l'autre pour l'entrainement ?

Prochaines etapes

- Vérifier un peu plus mon raisonnement et mon code
- Rendre l'encodage plus prêt pour les QGNNs (avec PyTorch ou d'autres outils)

Merci de votre attention!

Université de Sherbrooke

UNIVERSITÉ TÉLUQ