Prediction of missing Bid-Ask spread values

Challenge ENS-CFM 2022

Yoach Lacombe
MVA 21/22
ENS Paris-Saclay

March 22, 2022

Sommaire

Rappel du problème

Prédiction de valeurs manquantes

Nous sommes intéressés par la prédiction de spreads journaliers manquants sur un ensemble d'instruments financiers.

- Le bid-ask spread est la différence entre le prix d'offre et de demande.
- Il est un indicateur de la liquidité du marché.

On s'intéresse ici au bid-ask spread journalier de futures, c'est-à-dire de contrats garantissant la vente à un horizon de temps T à un prix donné.

Les données

On dispose de 895 915 entrées, réparties en 365 futures, avec environ 27% de spreads manquants par instruments et entre 226 et 2538 observations par contrats

Les features

- La date actuelle
- La date d'expiration du future
- Les prix d'ouverture et de fermeture du marché
- Les prix les plus élevés et bas de la journée
- Le tick size - la plus petite unité d'échange du future.
- Le volume échangé durant la journée
- L'Open Interest (OI) i.e le nombre de contrats actifs à la fin de la journée
- TTM (time-to-maturity) i.e le nombre de jours avant la prochaine expiration
- High-Low spread, Close-Open spread

Note

Chaque future est défini par son **underlying** (l'asset sur lequel porte le contrat) et son **niveau de maturité** (liquidity rank) qui va de celui qui va expirer le plus rapidement à celui qui a été émis le plus tard.

Observations initiales - 1

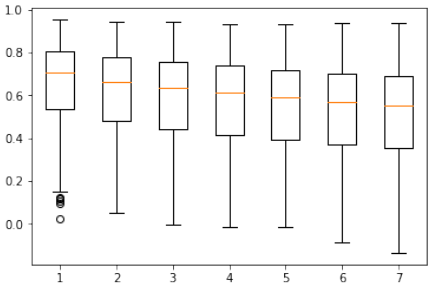

Figure: Boxplot de l'autocorrélation des spreads.

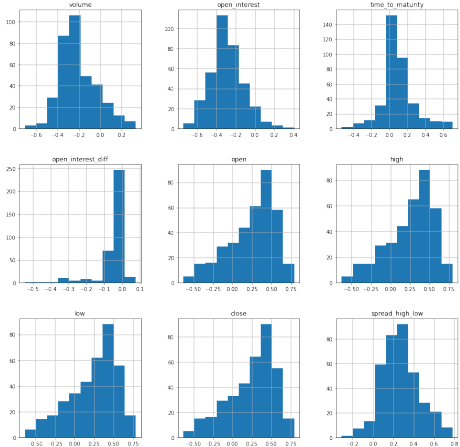

Figure: Histogramme des corrélations entre le spread et les autres features.

Observations initiales - 2

La principale difficulté sera donc de trouver un algorithme qui s'appuie à la fois sur la dépendance **spatiale** et la dépendance **temporelle**.
La dépendance spatiale est d'autant plus **importante** qu'on observe des jours à activité fortement **inhabituelle**.

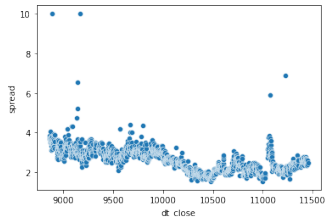

Figure: Spread journalier du future numéro 150659072, de rang de liquidité 6.

Les principales approches dans la littérature

- Méthodes basées sur les voisins (**Neighbor-based**) - KNN-imputer
- Modélisation avec contraintes (**Constraint-based**)
- **Interpolation** - Linear interpolation
- Méthodes de **régression** - ARIMA
- Méthodes statistiques (**Statistical based**) - Moyenne ou médiane
- Factorization de matrices (**Matrix Factorization - MF**) - SVD/PCA/PPCA
- Méthodes d'Expectation-Maximization (**EM**) - TRMF/PPCA
- Apprentissage profond (**Deep-Learning**) - basées sur des RNN, Transformers, GNN etc.

Résultat intéressant

Le papier "Do we really need deep learning models for time series forecasting?" soutient qu'un algorithme de gradient boosting couplé à un feature engineering bien pensé surpasse la plupart des méthodes de DL.

Du dataset initial aux séries temporelles multivariées

On passe dynamiquement du dataset original representé par un ensemble de tuples $\{(d_i, future_j, feature_k)\}$ à une matrice $\mathbf{X} \in \mathbb{R}^{T \times F}$ où la ligne correspond à un jour donnée et la colonne à un tuple future/feature donné.

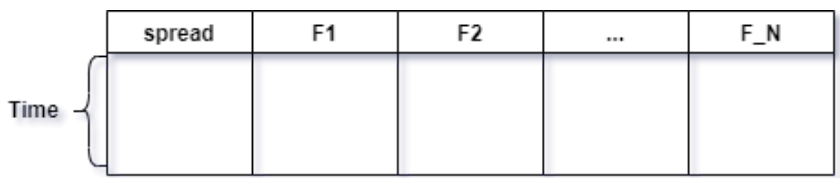

Figure: Diagramme de la répresentation en matrice. Chaque bloc est de taille Tx365 (une colonne par future).

Des séries temporelles multivariées aux séquences de graphes

On peut aussi vouloir représenter les interactions entre futures par un graphe.
Pour cela, on peut créer une matrice de similarité à partir de la corrélation du spread entre les contrats.
Pour ce faire, j'utilise le k-nearest neighbors (KNN) sur la valeur absolue des corrélations.

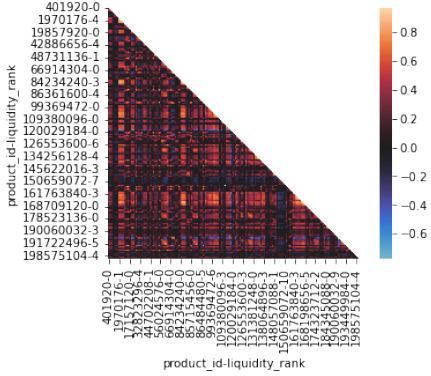

Figure: Heatmap de la matrice de corrélation des spreads.

Probabilistic PCA - PPCA

PPCA

- La PCA est une méthode de réduction de dimension qui projette linéairement les données sur de nouveaux axes qui réduisent la variance dans l'espace latent.
- La PPCA suppose que les variables latentes et que les données originales conditionnellements aux variables latentes sont normalement distribuées et applique un EM.
- Les données manquantes sont prédites en maximisant leur vraisemblance par au modèle.

Inconvénient

Les entrées sont supposées indépendantes. Autrement dit, il n'y a pas de dépendances temporelles prise en compte.

Multiple Imputation by Chained Equations (MICE)

MICE est un algorithme simple et efficace qui **fit itérativement** un modèle de régression sur chaque spread à partir des autres features dont les valeurs manquantes ont été imputé à partir d'une procédure simple (j'ai choisi l'interpolation linéaire) ou d'une version antérieure de l'algorithme. J'ai utilisé un algorithme de Gradient Boosting (XGBRegressor), très efficace et rapide sur des données tabulaires.

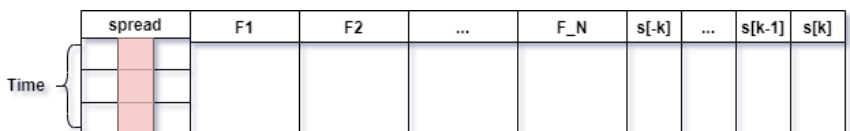

Figure: Diagramme de la répresentation utilisée par MICE. La colonne rouge est le spread cible qu'on va prédire à partir des autres données. On rajoute également les spreads précédents et suivants.

GRIN

- La plupart des approches DL sont faites soit pour des séries temporelles univariées ou pour des datasets sans dépendances temporelles.
- GRIN dépasse cette limite en incorporant une série temporelle de graphes.
- Testé seulement avec les spreads.

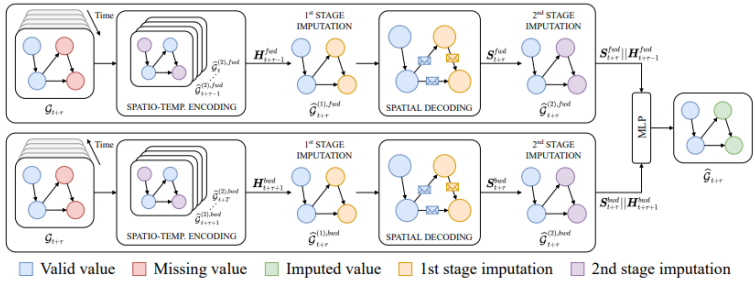

Figure: Aperçu de l'architecture de GRIN.

Résultats

Method	Parameters	Public score
MICE	2 iter,200 trees,depth 3	0.651
KNN-Imputer	5 neighbors	0.689
PPCA	K=100, tol=1e-4	0.720
GRIN	with no global attention	0.723
Linear Interpolation	**BENCHMARK**	**0.891**
XGBRegressor	Naive approach	0.953
TRMF	K=30, 500 iter	1.12

Table: Mes meilleurs résultats sur le leaderboard public. Résultat sur le leaderboard privé - **0.6546**

Commentaires - sur la dépendance spatiale

Meilleurs modèles

- Mes meilleurs résultats sont MICE, KNN-Imputer et PPCA.
- Les 3 approches sont principalement basés sur la **dépendance spatiale** plutôt que temporelle.
- Comme mentionné au début, c'est probablement parce que les spreads les plus difficiles à prédire sont ceux correspondant à des journées exceptionnelles.

Pour aller plus loin

- **Etude d'ablation** - MICE seulement avec les spreads performe seulement à 0.731.
- GRIN devrait beaucoup mieux marcher en intégrant les autres features.
- Une étude rapide de l'importance des features de XGBRegressor sur MICE confirme la diversité des spreads.

Conclusion

MICE

Malgré la simplicité de son approche, MICE est un algorithme très puissant lorsqu'il est couplé à un régresseur de qualité et à des features assez informatives.

Adaptation d'algorithmes de prédiction

Certaines approches SOTA de prédiction (ex: Temporal Fusion Transformers) peuvent être adapté avec un peu d'effort en algorithme d'imputation.

Modélisation

Une étude plus poussée de la structure des spreads pourrait produire des résultats intéressants. Il semble y avoir une composante **périodique** (=modélisable avec les séries temporelles) couplé à des **excitations/inhibitions ponctuelles** (=modélisable avec les dépendances spatiales).

The End