

Diffusion pour des trajectoires

Amina MANSEUR
ENSAE 2A

Année scolaire 2023/2024

Maîtres de stage : Badih GHATTAS, Georges OPPENHEIM
Aix Marseille School of Economics, Université d'Aix Marseille



Plan de la présentation

- 1 Introduction et problématique
- 2 Partie théorique : les modèles de diffusion
 - Idée générale
 - Un modèle adapté aux séries temporelles
- 3 Partie application : Données et génération
 - Données d'entraînement et faits stylisés
 - Simulations réalisées
 - Évaluation de la qualité de la génération
- 4 Conclusion
- 5 Annexe

Introduction et problématique

Introduction générale

Contexte : Données financières **rare**s et **coûteuses**.

Méthode : Modèles génératifs de diffusion, une alternative aux approches traditionnelles (GANs, VAEs).

Objectif : Produire des signaux unidimensionnels réalistes et diversifiés
=> **augmentation des données**.

Enjeux du projet

Explorer les modèles de diffusion pour séries temporelles.

Enjeux du projet

Explorer les modèles de diffusion pour séries temporelles.

Évaluer la qualité des signaux générés selon des critères usuels :

Enjeux du projet

Explorer les modèles de diffusion pour séries temporelles.

Évaluer la qualité des signaux générés selon des critères usuels :

- Similarité des distributions.

Enjeux du projet

Explorer les modèles de diffusion pour séries temporelles.

Évaluer la qualité des signaux générés selon des critères usuels :

- Similarité des distributions.
- Diversité des données produites.

Enjeux du projet

Explorer les modèles de diffusion pour séries temporelles.

Évaluer la qualité des signaux générés selon des critères usuels :

- Similarité des distributions.
- Diversité des données produites.
- Simplicité et interprétabilité du modèle.

Enjeux du projet

Explorer les modèles de diffusion pour séries temporelles.

Évaluer la qualité des signaux générés selon des critères usuels :

- Similarité des distributions.
- Diversité des données produites.
- Simplicité et interprétabilité du modèle.

Proposer des métriques adaptées aux caractéristiques statistiques particulières des signaux financiers.

Partie théorique : les modèles de diffusion

Modèle de diffusion : définition et objectifs

Problème : Générer de nouvelles données à partir d'une distribution p_0 inconnue.

Modèle de diffusion : définition et objectifs

Problème : Générer de nouvelles données à partir d'une distribution p_0 inconnue.

Objectif : Apprendre la distribution de données réelles p_0 à partir d'un échantillon $\{\mathbf{x}_0^i\}_{i=1}^N \subset \mathbb{R}^d$.

Modèle de diffusion : définition et objectifs

Problème : Générer de nouvelles données à partir d'une distribution p_0 inconnue.

Objectif : Apprendre la distribution de données réelles p_0 à partir d'un échantillon $\{\mathbf{x}_0^i\}_{i=1}^N \subset \mathbb{R}^d$.

Approche des modèles de diffusion en deux étapes :

Modèle de diffusion : définition et objectifs

Problème : Générer de nouvelles données à partir d'une distribution p_0 inconnue.

Objectif : Apprendre la distribution de données réelles p_0 à partir d'un échantillon $\{\mathbf{x}_0^i\}_{i=1}^N \subset \mathbb{R}^d$.

Approche des modèles de diffusion en deux étapes :

Bruitage : Ajout itératif d'un bruit à un signal initial \mathbf{x}_0 en T étapes.

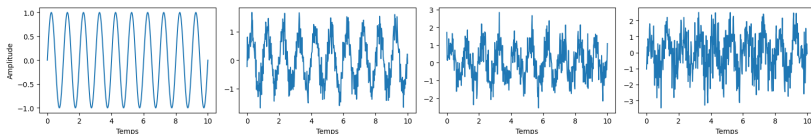


Figure – Illustration des processus de bruitage et de débruitage

Modèle de diffusion : définition et objectifs

Problème : Générer de nouvelles données à partir d'une distribution p_0 inconnue.

Objectif : Apprendre la distribution de données réelles p_0 à partir d'un échantillon $\{\mathbf{x}_0^i\}_{i=1}^N \subset \mathbb{R}^d$.

Approche des modèles de diffusion en deux étapes :

Bruitage : Ajout itératif d'un bruit à un signal initial \mathbf{x}_0 en T étapes.

Débruitage : Apprendre la transition entre un signal bruité à l'étape t et un signal à l'étape $t - 1$.

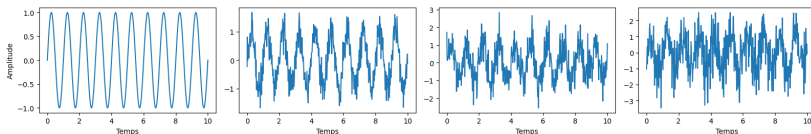


Figure – Illustration des processus de bruitage et de débruitage

Le modèle de Diffusion-TS

Un modèle adapté aux séries temporelles : *Diffusion-TS*, (YUAN et QIAO 2024).

Le processus de diffusion directe est :

- Distribution des échantillons inconnue : $p_0(\mathbf{x}_0)$

Le modèle de Diffusion-TS

Un modèle adapté aux séries temporelles : *Diffusion-TS*, (YUAN et QIAO 2024).

Le processus de diffusion directe est :

- Distribution des échantillons inconnue : $p_0(\mathbf{x}_0)$
- Ajout d'un bruit Gaussien entre chaque étape selon le noyau de transition :
 $p_{t|t-1}(\mathbf{x}_t|\mathbf{x}_{t-1}) = N(\sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}_d)$

$$\forall t \in \{1, \dots, T\} : \quad \mathbf{x}_t = \sqrt{1 - \beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\mathbf{z}_t \quad \text{où} \quad \mathbf{z}_t \stackrel{\text{i.i.d.}}{\sim} N(0, \mathbf{I}_d)$$

et (β_t) un ensemble fixé de paramètres contrôlant le bruit ajouté.

Le modèle de Diffusion-TS

Un modèle adapté aux séries temporelles : *Diffusion-TS*, (YUAN et QIAO 2024).

Le processus de diffusion directe est :

- Distribution des échantillons inconnue : $p_0(\mathbf{x}_0)$
- Ajout d'un bruit Gaussien entre chaque étape selon le noyau de transition :
 $p_{t|t-1}(\mathbf{x}_t|\mathbf{x}_{t-1}) = N(\sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}_d)$

$$\forall t \in \{1, \dots, T\} : \quad \mathbf{x}_t = \sqrt{1 - \beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\mathbf{z}_t \quad \text{où} \quad \mathbf{z}_t \stackrel{\text{i.i.d.}}{\sim} N(0, \mathbf{I}_d)$$

et (β_t) un ensemble fixé de paramètres contrôlant le bruit ajouté.

- Ajout de bruit en une étape, selon le noyau de transition :
 $p_{t|0}(\mathbf{x}_t|\mathbf{x}_0) = N(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I}_d)$, avec $\alpha_t = 1 - \beta_t$ et $\bar{\alpha}_t = \prod_{k=1}^t \alpha_k$

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\varepsilon_t \quad \text{où} \quad \varepsilon_t \stackrel{\text{i.i.d.}}{\sim} N(0, \mathbf{I}_d). \quad (1)$$

Le modèle de Diffusion-TS

Le processus de diffusion indirecte, en partant d'un échantillon \mathbf{x}_T de $p_T \approx N(0, \mathbf{I}_d)$, est :

- Chaîne de Markov : $p_{t-1|t}(\mathbf{x}_{t-1}|\mathbf{x}_t) = N(\mu_\theta(\mathbf{x}_t, t), \frac{1-\tilde{\alpha}_{t-1}}{1-\tilde{\alpha}_t}\beta_t\mathbf{I}_d)$.

Le modèle de Diffusion-TS

Le processus de diffusion indirecte, en partant d'un échantillon \mathbf{x}_T de $p_T \approx N(0, \mathbf{I}_d)$, est :

- Chaîne de Markov : $p_{t-1|t}(\mathbf{x}_{t-1}|\mathbf{x}_t) = N(\mu_\theta(\mathbf{x}_t, t), \frac{1-\tilde{\alpha}_{t-1}}{1-\tilde{\alpha}_t}\beta_t\mathbf{I}_d)$.
- Estimation de θ par maximum de vraisemblance.

Le modèle de Diffusion-TS

Le processus de diffusion indirecte, en partant d'un échantillon \mathbf{x}_T de $p_T \approx N(0, \mathbf{I}_d)$, est :

- Chaîne de Markov : $p_{t-1|t}(\mathbf{x}_{t-1}|\mathbf{x}_t) = N(\mu_\theta(\mathbf{x}_t, t), \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t\mathbf{I}_d)$.
- Estimation de θ par maximum de vraisemblance.
- Revient à minimiser l'écart entre : $\tilde{\mu}(\mathbf{x}_t, \mathbf{x}_0) = \sqrt{\alpha_t} \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \mathbf{x}_t + \sqrt{\frac{\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}} \mathbf{x}_0$
et $\mu_\theta(\mathbf{x}_t, t) = \sqrt{\alpha_t} \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \mathbf{x}_t + \sqrt{\frac{\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}} \hat{\mathbf{x}}_0(\mathbf{x}_t, t, \theta)$ pour tout t .

Le modèle de Diffusion-TS

Estimation de \mathbf{x}_0 par $\hat{\mathbf{x}}_0(\mathbf{x}_t, t, \theta)$, pour chaque étape t , obtenu par le modèle de paramètres θ , entraîné en minimisant la perte :

$$\mathcal{L}_\theta = \mathbb{E}_{t, \mathbf{x}_0} \left[w_t \left[\lambda_1 \|\mathbf{x}_0 - \hat{\mathbf{x}}_0(\mathbf{x}_t, t, \theta)\|^2 + \lambda_2 \|\mathcal{F}\mathcal{F}\mathcal{T}(\mathbf{x}_0) - \mathcal{F}\mathcal{F}\mathcal{T}(\hat{\mathbf{x}}_0(\mathbf{x}_t, t, \theta))\|^2 \right] \right] \quad (2)$$

avec $w_t = f(\beta_t, t)$ et λ_1, λ_2 des réels.

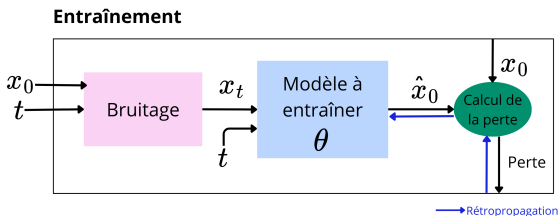


Figure – Principe de l'entraînement de Diffusion-TS

Le modèle de Diffusion-TS

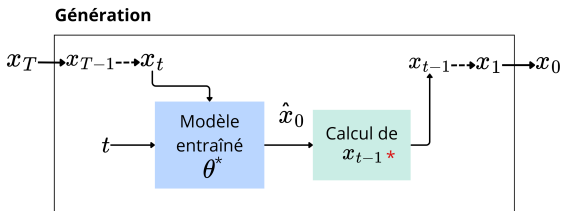


Figure – Principe de la génération de Diffusion-TS

Génération : à partir d'un bruit blanc gaussien \mathbf{x}_T , on débruite le signal de manière itérative, pour $t \in \{1, \dots, T\}$:

$$\mathbf{x}_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t} \hat{\mathbf{x}}_0(\mathbf{x}_t, t, \theta^*) + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_t + \sqrt{\frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}} \beta_t \mathbf{z}_t, \quad \mathbf{z}_t \stackrel{\text{i.i.d.}}{\sim} N(0, I_d)$$

Partie application : Données et génération

Données d'entraînement et faits stylisés

Les données :

- Évolution temporelle du prix des produits financiers les plus capitalisés du S&P500 : 2400 jours du 2 janvier 2015 au 17 juillet 2024.
- Évolution temporelle du prix du Bitcoin : 2 416 722 minutes du 1er janvier 2017 à 00h00 au 6 août 2021 à 06h42.

Les faits stylisés considérés :

- Absence d'autocorrélations des rendements.
- Existence de clusters de volatilité des rendements.
- Présence de queues de distributions des rendements lourdes.

Definition

Le rendements logarithmique à l'instant t est défini comme le logarithme du ratio des prix successifs :

$$r_t = \ln \left(\frac{P_t}{P_{t-1}} \right)$$

où P_t est le prix de l'actif à l'instant t .

Données d'entraînement et faits stylisés

Les données du Bitcoin :

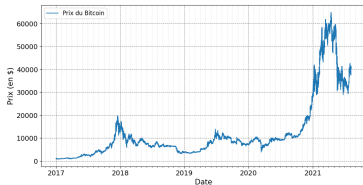


Figure – Evolution temporelle du prix

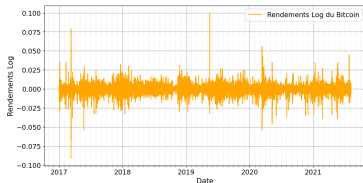


Figure – Evolution temporelle des rendements

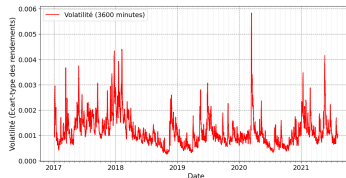


Figure – Volatilité des rendements (fenêtre glissante de 60h)

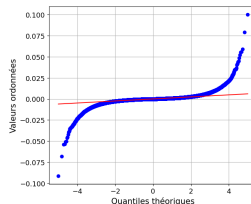


Figure – QQ plot des rendements

Simulations réalisées

Données d'entraînement : séries bidimensionnelles de longueur 120 (Bitcoin en minute et S&P500 en jours)

Les critères de qualité considérés

- **Critères graphiques** (évolutions temporelles des prix, rendements et volatilité, QQ plot, ACF, réduction de la dimensionnalité).
- **Tests statistiques** (ADF et Ljung-Box).
- **Caractéristiques statistiques** (Coefficient de Hurst et Kurtosis).
- **Distances** (L2, Divergence KL-Fourier, Wasserstein-Fourier (CAZELLES, ROBERT et TOBAR 2020))

Évaluation de la qualité de la génération

Méthode 1 : Comparaison simple entre signal initial et signal généré

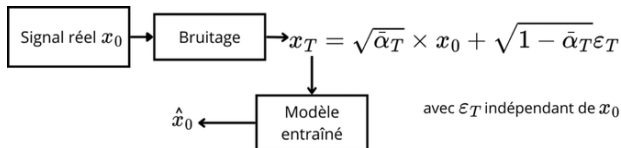


Figure – Schéma explicatif de la méthode 1

Limite de la méthode :

Le signal bruité x_T , est distribué selon une densité de probabilité gaussienne dont le bruit ajouté masque en grande partie le signal initial x_0 , x_T n'est donc pas associé de façon unique au signal x_0 , ce qui peut altérer la qualité de la comparaison.

Méthode 1 : Résultats

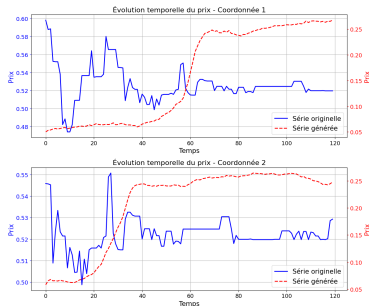


Figure – Évolution temporelle du prix du Bitcoin : en haut, la première coordonnée ; en bas, la deuxième coordonnée

Méthode 1 : Résultats

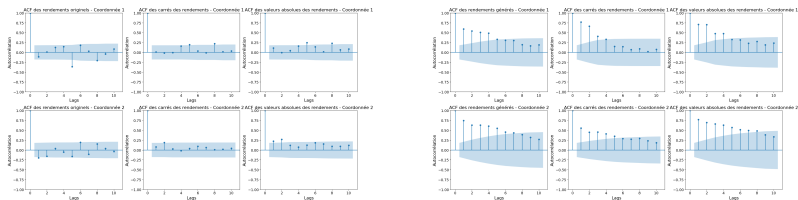


Figure – ACF des rendements : à gauche, série originelle ; à droite, série générée

Méthode 1 : Résultats

- **Points positifs :**
 - Relations temporelles entre les coordonnées de la série **bien capturées**.
 - Reproduction des queues de distribution des rendements lourdes (QQ plot et valeurs positives des Kurtosis des rendements).
 - Densités spectrales de puissance proches (Wasserstein-Fourier $\sim 10^{-3}$)
- **Points négatifs :** Autocorrélations des rendements plus importantes pour les données générées (ACF et test de Ljung-Box).

Méthode 2 : Génération multiple et moyenne des signaux

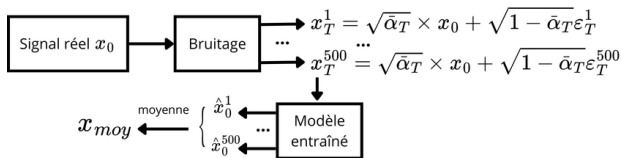


Figure – Schéma explicatif de la méthode 2

Limite de la méthode : :

La courbe moyenne obtenue est lissée et les caractéristiques importantes des séries comme les tendances, les saisonnalités et le bruit sont atténuées.

Méthode 2 : Résultats

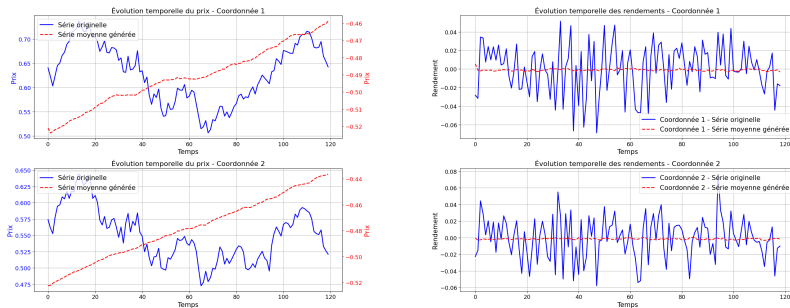


Figure – À gauche : Évolution temporelle des prix du S&P 500 ; à droite : Évolution temporelle des rendements du S&P 500

- Point négatif : Perte d'information concernant les fluctuations des prix et rendements dû au calcul de la moyenne des séries générées.

Méthode 3 : Génération multiple et comparaison individuelle des signaux générés

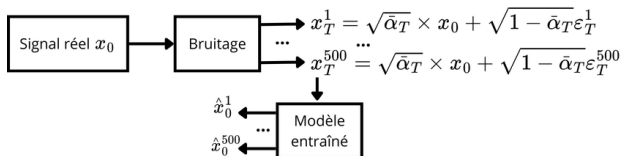


Figure – Schéma explicatif de la méthode 3

Limite de la méthode : :

- Grande diversité des signaux réels existants.

Méthode 3 : Résultats (Prix du Bitcoin)

L'idée :

- Observation des distributions des distances et des coefficients calculés.

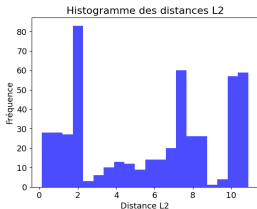


Figure – L2

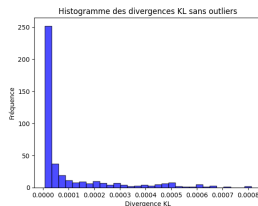


Figure – KL-Fourier

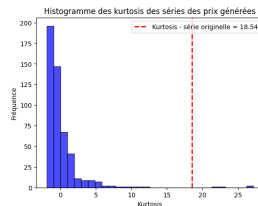


Figure – Kurtosis

Méthode 3 : Résultats

Points positifs :

- **Faibles distances** KL-Fourier et Wasserstein-Fourier entre les signaux.
- Capture la mémoire longue des séries et leur structure fréquentielle (avec coefficient de Hurst réel appartenant à la distribution des coefficients générés).

Points négatifs :

- **Écarts considérables** entre les kurtosis des rendements réels et issus de la génération (comportement des queues de distribution différent).
- **Variabilité importante** (avec écart-type de l'ordre de la moyenne) => la convergence vers le signal réel n'est pas toujours significative.

Méthode 4 : Comparaison agrégée entre données réelles et simulées

Le **principe de la méthode** est de :

- Comparer la distribution des données d'entraînement $\{\mathbf{x}_0^i\}_{i=1}^R$ et celle des données générées $\{\hat{\mathbf{x}}_0^i\}_{i=1}^S$ de manière agrégée.

Limite de la méthode :

- L'analyse agrégée peut masquer des variations spécifiques et des comportements particuliers des séries individuelles.

Méthode 4 : Résultats

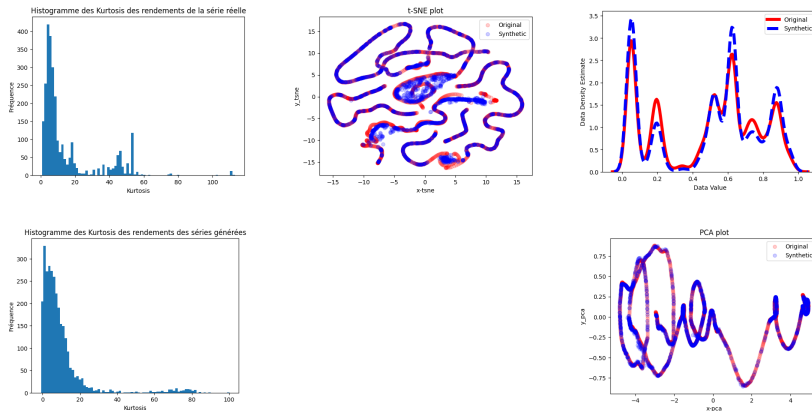


Figure – Distribution des Kurtosis (réelle et issue de la génération) ; Visualisations après réduction de dimension

Méthode 4 : Résultats

Points positifs :

- Diversité et structure des données d'entraînement **bien capturées**.
- Distribution des coefficients de Hurst (prix) et Kurtosis (prix et rendements) **semblables**.
- Faibles distances KL-Fourier et Wasserstein-Fourier.

Points négatifs :




- Distribution des coefficients de Hurst des rendements présentant des **différences notables**.
- Comportements individuels peuvent différer (avec certaines autocorrélations des rendements trop marquées).

Conclusion

Conclusion et discussion

- Exploration des modèles de diffusion pour générer des séries temporelles financières.
- Sélection d'approches et de métriques spécifiques pour mesurer la qualité de la génération.
- Étude des caractéristiques apprises malgré leur non présence dans le critère.
- Idées d'amélioration :
 - Affiner le critère d'apprentissage pour améliorer la qualité des données notamment par rapport aux autocorrélations des rendements.
 - Conditionnement.
 - Modèle de diffusion continu.
 - Utiliser d'autres modèles (Transfusion (SIKDER et AL. 2024))

Références I

-  CAZELLES, Elsa, Arnaud ROBERT et Felipe TOBAR (2020). "The Wasserstein-Fourier Distance for Stationary Time Series". In : *arXiv preprint arXiv :1912.05509*. URL : <https://arxiv.org/abs/1912.05509>.
-  SIKDER, Md Fahim et AL. (avr. 2024). "TransFusion : Generating Long, High Fidelity Time Series using Diffusion Models with Transformers". In : *arXiv. eprint : 2307.12667*. URL : <https://doi.org/10.48550/arXiv.2307.12667>.
-  YUAN, Xinyu et Yan QIAO (mars 2024). "Diffusion-TS : Interpretable Diffusion for General Time Series Generation". In : *arXiv. arXiv :2403.01742*. arXiv : 2403.01742. URL : <https://doi.org/10.48550/arXiv.2403.01742>.

Annexe

Le modèle de Diffusion-TS

Caractéristiques du modèle :

- Estimation du signal réel.
- Techniques de décomposition de la série selon tendance et saisonnalité.
- Fonction de perte basée sur les transformées de Fourier.
- Architecture de type transformer.

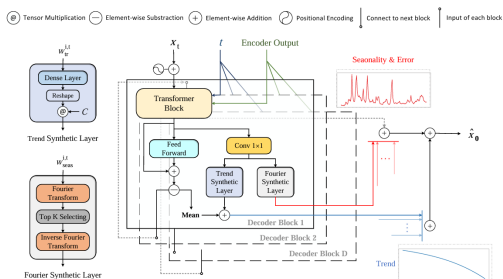


Figure – Architecture du décodeur de Diffusion-TS

Premières expériences

Paramètre	Valeur
Timestep T	500
Seq_length	120, 128 ou 250
Feature_size	2, 3 ou 16
Batch_size	128
n_layer_enc	3
n_layer_dec	2 ou 3
max_epochs	10000, 30000 ou 50000
save_cycle	max_epochs / 10
patience	2000, 6000 ou 10000

Figure – Valeurs des paramètres utilisées dans les expériences

Premières expériences

Paramètre	Définition
seq_length	Longueur de la série temporelle
feature_size	Nombre de coordonnées de la série
batch_size	Taille d'un lot de données dans le processus d'entraînement (nombre de séries temporelles dans le lot)
max_epochs	Nombre maximal d'epochs pendant lesquelles le modèle sera entraîné, c'est-à-dire le nombre maximum de passages complets à travers l'ensemble des données d'entraînement (S dans l'algorithme 1)
save_cycle	Détermine la fréquence à laquelle le modèle est sauvegardé pendant l'entraînement
timesteps	Nombre total d'étapes de bruitage dans le processus de diffusion
n_heads	Nombre de têtes d'attention dans le mécanisme d'attention multi-têtes des Transformers
n_layer_encod	Nombre de couches dans le bloc encodeur du modèle
n_layer_dec	Nombre de couches dans le bloc décodeur du modèle
patience	Nombre d'epochs à attendre sans amélioration avant d'arrêter l'entraînement (permet d'éviter l'overfitting)

Algorithme d'entraînement

Algorithm 1: Boucle d'entraînement

Input: S : Nombre de pas d'entraînement, N : Nombre de batchs avant mise à jour des paramètres, b : Taille d'un batch de données

$s \leftarrow 0$ **while** $s < S$ **do**

$total_loss \leftarrow 0$;

for $n = 1$ **to** N **do**

 Charger le batch de données $S_n : \{\mathbf{x}_0^i\}_{i=1}^b$;

 Calculer la perte sur le batch des données $S_n : loss/N$;

 Calculer et accumuler les gradients;

$total_loss \leftarrow total_loss + loss$;

 Mise à jour des paramètres θ_s avec l'optimiseur Adam;

 Ajuster le taux d'apprentissage η_s ;

$s \leftarrow s + 1$;

 Mettre à jour la moyenne mobile exponentielle : $\bar{\theta}_{s+1} \leftarrow \beta \bar{\theta}_s + (1 - \beta) \theta_s$;

Algorithme d'entraînement

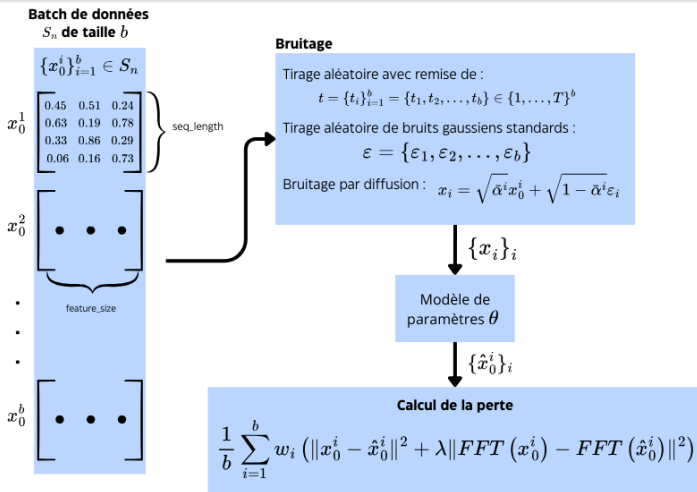


Figure – Calcul de la perte dans un batch de données

Les métriques utilisées

Fonction d'autocorrélation (ACF) : L'ACF d'une série temporelle $\{X_t\}_{t=1}^N$ à un lag k est définie par :

$$\text{ACF}(k) = \frac{\sum_{t=1}^{N-k} (X_t - \bar{X})(X_{t+k} - \bar{X})}{\sum_{t=1}^N (X_t - \bar{X})^2}$$

où $\bar{X} = \frac{1}{N} \sum_{t=1}^N X_t$.

L'ACF prend des valeurs entre -1 et +1. Elle est positive lorsqu'il existe une corrélation positive entre les valeurs à des moments espacés par k , négative lorsqu'il y a une corrélation inverse et nulle lorsqu'il n'y a pas de corrélation à ce lag.

Les métriques utilisées

Réduction de la dimensionnalité

- **ACP** : projette les données dans un espace de plus faible dimension tout en maximisant la variance des données projetées (projection sur le sous-espace des k vecteurs propres associés k plus grandes valeurs propres de la matrice de covariance).
- **t-SNE** : réduit les dimensions tout en préservant la structure locale en minimisant la divergence KL entre les distributions des distances dans les espaces original et réduit. Il minimise $KL(P\|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$ où p_{ij} est une probabilité basée sur la distance gaussienne et q_{ij} sur une distribution t de Student.
- **KDE** : méthode non paramétrique pour estimer la densité de probabilité d'une variable aléatoire en sommant des noyaux autour des données (utilisée après l'ACP)

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

où K est une fonction noyau (par exemple, gaussienne) et h est un paramètre de lissage.

Les métriques utilisées

Divergence de Kullback-Leibler (KL) : La divergence KL est une mesure qui quantifie la différence entre deux distributions de probabilité P (distribution réelle) et Q (distribution approximée). Elle indique la perte d'information lorsqu'on utilise Q pour représenter P . *Pour des distributions continues :*

$$D_{\text{KL}}(P\|Q) = \int_{-\infty}^{\infty} P(x) \log \left(\frac{P(x)}{Q(x)} \right) dx$$

Distance de Wasserstein : mesure le coût minimal pour transformer une distribution en une autre. *Cas unidimensionnel :* Pour deux distributions de probabilité unidimensionnelles P et Q , définies par leurs fonctions de répartition cumulées respectives $F_P(x)$ et $F_Q(x)$, la distance de Wasserstein d'ordre 1 est donnée par :

$$W_1(P, Q) = \int_{-\infty}^{\infty} |F_P(x) - F_Q(x)| dx$$

où :

- $F_P(x)$ et $F_Q(x)$ représentent les probabilités cumulées des distributions P et Q à chaque point x .

Les métriques utilisées

Coefficient de Hurst : noté H , c'est une mesure utilisée pour quantifier la persistance ou l'antipersistente d'une série temporelle. Il est défini mathématiquement à partir de l'analyse du *rapport range étendu sur l'écart type*, souvent noté R/S , et se base sur une loi de puissance.

$$\frac{R(n)}{S(n)} \propto n^H$$

où :

- $R(n)$: la portée (*range*) de la série sur une fenêtre de taille n , définie comme $\max(X_1, \dots, X_n) - \min(X_1, \dots, X_n)$, où X_1, \dots, X_n sont les valeurs de la série temporelle.
- $S(n)$: l'écart-type des données sur la même fenêtre.

Interprétation du coefficient H :

- $H = 0.5$: une série **aléatoire pure** (processus de type bruit blanc).
- $0.5 < H < 1$: série **persistante**, c'est-à-dire que des valeurs élevées sont suivies par des valeurs élevées, et des valeurs basses par des valeurs basses.
- $0 < H < 0.5$: série **antipersistante**, où des valeurs élevées sont souvent suivies par des valeurs basses, et vice versa.

Les métriques utilisées

Kurtosis : Ce coefficient mesure l'épaisseur des queues de distribution des rendements en comparant le coefficient calculé à celui de la distribution normale, qui vaut 3. La formule de la Kurtosis pour une série $X = \{X_t\}_{t=1}^N$ est

$$K = \frac{\frac{1}{N} \sum_{t=1}^N (X_t - \bar{X})^4}{\left(\frac{1}{N} \sum_{t=1}^N (X_t - \bar{X})^2 \right)^2}, \text{ avec } \bar{X} = \frac{1}{N} \sum_{t=1}^N X_t \text{ la moyenne empirique de la série.}$$

On lui soustrait habituellement la kurtosis de la distribution normale pour obtenir la kurtosis excédentaire

Interprétation de la Kurtosis excédentaire :

- Kurtosis excédentaire > 0 : queues plus épaisses
- Kurtosis excédentaire < 0 : queues plus légères
- Kurtosis excédentaire $= 0$: similaire à la normale.

Les métriques utilisées

Test de Dickey-Fuller augmenté (ADF) : utilisé pour vérifier si une série temporelle est stationnaire.

Hypothèses :

H_0 (hypothèse nulle) : La série a une racine unitaire (non stationnaire).

H_1 (hypothèse alternative) : La série est stationnaire.

Le test repose sur l'estimation du modèle suivant pour une série temporelle y_t :

$$\Delta y_t = \phi y_{t-1} + \sum_{i=1}^p \gamma_i \Delta y_{t-i} + \epsilon_t$$

où :

$\Delta y_t = y_t - y_{t-1}$ est la différence première, ϕ mesure la présence d'une racine unitaire.

$\sum_{i=1}^p \gamma_i \Delta y_{t-i}$ sont des termes de retard pour gérer la corrélation entre observations successives, et ϵ_t un terme d'erreur blanc.

Statistique de test : Le test ADF vérifie la nullité du coefficient ϕ :

$$t_{\text{stat}} = \frac{\hat{\phi}}{\text{SE}(\hat{\phi})}$$

où $\hat{\phi}$ est l'estimation de ϕ et $\text{SE}(\hat{\phi})$ est l'erreur standard associée.

Les métriques utilisées

Test de Ljung-Box : utilisé pour détecter la présence d'autocorrélations significatives dans une série temporelle, jusqu'à un certain décalage m .

Hypothèses

H_0 (hypothèse nulle) : La série est aléatoire (aucune autocorrélation significative).

H_1 (hypothèse alternative) : La série présente une autocorrélation significative.

Statistique de test Le test calcule la statistique Q basée sur les autocorrélations de la série :

$$Q = n(n+2) \sum_{k=1}^m \frac{\hat{\rho}_k^2}{n-k}$$

où n est le nombre d'observations, m le nombre de décalages considérés et $\hat{\rho}_k$ l'autocorrélation estimée au décalage k .

Sous H_0 , Q suit approximativement une distribution χ^2 avec m degrés de liberté.

Si Q dépasse une valeur critique de la distribution χ^2 , on rejette H_0 , indiquant la présence d'autocorrélations significatives. Sinon, on accepte H_0 , suggérant que la série est aléatoire.

L'architecture du code

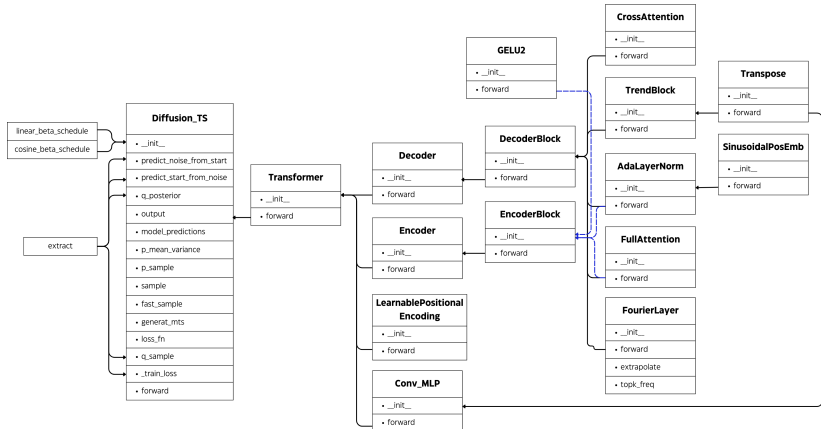


Figure – Architecture du code