diffusion-réaction, fractales et dynamique du CAC 40

Table des matières

Résumé 4

Systèmes diffusion-réaction, fractales et dynamique du CAC 40 6

1. Le point de départ : les systèmes diffusion-réaction et l'auto-similarité 6

2. Solutions auto-similaires 7

3. Lien avec les processus de branchement 7

4. Fractales naturelles et auto-organisation 7

5. Passage aux systèmes économiques 8

6. Analyse empirique du CAC 40 8

7. Dimension effective et lois d'échelle 9

8. Reconstruction de l'espace des phases 9

Résultat principal 11

Résumé quantitatif 11

9. Analyse en trois dimensions 13

10. Réduction de dimension par PCA 13

11. Interprétation des deux variables principales 15

12. Vers un modèle dynamique 15

13. Équation pour la densité 16

14. Extension diffusion-réaction 16

15. Vision d'ensemble 16

16. Conclusion 17

Modélisation stochastique du signal du CAC 40 projeté sur la première composante principale 17

1. Point de départ : un signal complexe, mais essentiellement porté par une variable effective 17

2. L'hypothèse centrale : une diffusion effective dans un paysage de potentiel 18

3. Pourquoi un simple bruit brownien ne suffit pas 20

4. La mémoire de volatilité et la construction du noyau 20

5. Le forçage coloré : pourquoi il faut une seconde mémoire 21

6. Le forçage hybride : addition de la composante colorée et de la composante blanche22

7. L'équation complète du simulateur 23

8. Comment lire les différentes composantes du modèle 24

9. Ce que les simulations montrent 24

Lecture physique 25

10. Interprétation conceptuelle de l'ensemble 26

11. Équation de Fokker-Planck associée et portée de la modélisation 26

12. Conclusion générale 27

INTÉRÊT DU MODÈLE EN DATA SCIENCE 27

Résumé

Ce chapitre propose de relier la dynamique du CAC 40 aux cadres théoriques des systèmes diffusion-réaction, des processus de branchement et des structures fractales. Le point de départ est l'idée que de nombreux phénomènes naturels auto-organisés, comme les réseaux de rivières, les dendrites ou certains phénomènes turbulents, obéissent à des équations de diffusion-réaction capables de produire des formes auto-similaires. Le texte suggère qu'une partie des propriétés observées sur les marchés financiers, comme les queues lourdes, la volatilité en paquets, l'intermittence et les structures multi-échelles, peut être comprise dans un cadre conceptuel analogue.

L'analyse empirique du CAC 40 confirme d'abord deux faits majeurs : les rendements se comportent approximativement comme un bruit sans mémoire, tandis que la volatilité présente une mémoire longue marquée. L'étude spectrale et les lois d'échelle conduisent à estimer une dimension effective et un ordre de réaction proches de valeurs quadratiques, ce qui suggère que la dynamique collective du marché pourrait être dominée par des interactions effectives de type binaire, analogues à des mécanismes d'imitation ou de contagion entre agents.

Le chapitre reconstruit ensuite l'espace des phases du CAC 40 à l'aide de coordonnées retardées. En deux dimensions, on distingue déjà deux grands bassins d'attraction et un corridor de transition. En trois dimensions, la structure devient plus riche et révèle environ cinq bassins dynamiques correspondant à différents régimes du marché : régime bas stable, transition haussière, régime haut stable, sommet ou retournement, et régime de crise. Cette géométrie montre que le marché ne se déplace pas aléatoirement dans l'espace des états, mais visite préférentiellement certaines régions, comme le ferait un système dynamique à régimes multiples.

L'analyse en composantes principales montre ensuite que cette dynamique tridimensionnelle est presque entièrement contenue dans un plan de faible dimension. Deux variables principales dominent l'essentiel de la variance : l'une liée au niveau général du marché, l'autre aux accélérations et retournements de régime. À partir de là, le texte propose une modélisation stochastique de type Langevin, dans laquelle la dynamique du marché est décrite comme celle d'une variable évoluant dans un paysage de potentiel à plusieurs cuvettes, sous l'effet d'un bruit aléatoire. Ce potentiel est reconstruit directement à partir de la densité empirique, selon l'idée classique d'un potentiel effectif dérivé de $-\log p(x)$-\log p(x)$-\log p(x)$.

Mais un bruit brownien simple ne suffit pas à reproduire les données. Le modèle est donc enrichi par deux mécanismes supplémentaires. Le premier est une volatilité latente à mémoire longue, construite comme un champ corrélé, qui reproduit l'intermittence et le

volatility clustering. Le second est un forçage hybride mêlant une composante colorée lente et une composante blanche rapide, afin de restituer à la fois les basses fréquences du signal et sa microstructure locale. Le simulateur final combine ainsi trois briques : la topologie du potentiel, la mémoire de volatilité et la couleur fréquentielle du forçage.

Le résultat conceptuel est fort : malgré la complexité apparente du CAC 40, une grande partie de sa dynamique peut être représentée par un modèle effectif de faible dimension, lisible physiquement et statistiquement cohérent. Le texte souligne enfin l'intérêt de ce cadre en data science. Il permet de passer d'une logique purement prédictive à une logique explicative et générative, en reliant directement la densité aux puits de potentiel, l'autocorrélation de la volatilité au noyau mémoire et le spectre du signal à la structure du forçage. Le modèle devient ainsi un pont entre réduction de dimension, inférence stochastique et simulation générative interprétable.

Systèmes diffusion-réaction, fractales et dynamique du CAC 40

1. Le point de départ : les systèmes diffusion-réaction et l'auto-similarité

De nombreux phénomènes naturels présentent des structures auto-similaires, c'est-à-dire des structures qui se ressemblent à différentes échelles. On retrouve ce comportement dans :

- réseaux de rivières
- les branches d'arbres
- les structures de cristallisation
- les dendrites de solidification
- les colonies bactériennes
- certains phénomènes turbulents.

Un cadre mathématique simple permettant de décrire ces phénomènes est celui des équations de diffusion-réaction :

$$\frac{\partial s}{\partial t} = D\Delta s + R(s)$$

où

- s(x,t) est une grandeur (densité, concentration, activité),
- D est un coefficient de diffusion,
- R(s) est un terme de réaction.

Dans beaucoup de systèmes naturels, la réaction peut être approximée par une loi de puissance

$$R(s) = s^p$$

ce qui donne

$$\frac{\partial s}{\partial t} = D\Delta s + s^p$$

2. Solutions auto-similaires

Ce type d'équation admet souvent des solutions auto-similaires de la forme

$$s(x,t) = t^{-\alpha} F\left(\frac{x}{t^{\beta}}\right)$$

Cette structure signifie que l'évolution du système peut être décrite par une fonction de forme universelle F dont l'échelle spatiale et l'amplitude évoluent avec le temps.

En substituant cette forme dans l'équation, on obtient une équation ordinaire pour F. L'existence de telles solutions montre que l'équation possède une invariance d'échelle.

C'est précisément cette propriété qui conduit souvent à des structures fractales.

3. Lien avec les processus de branchement

Une autre manière d'interpréter ces équations est de partir d'un processus de branchement.

Dans un processus de branchement :

- un élément peut se multiplier
- ses descendants se diffusent
- les interactions locales peuvent produire des cascades.

Lorsqu'on passe à la limite continue, la densité moyenne des branches obéit souvent à une équation diffusion-réaction.

Ainsi :

processus de branchement ⇒ équation diffusion-réaction.

Lorsque le processus est proche d'un régime critique, la structure résultante devient auto-similaire.

C'est ce qui explique pourquoi de nombreuses formes naturelles fractales sont liées à des processus de croissance ou de branchement.

4. Fractales naturelles et auto-organisation

Dans la nature, on observe souvent :

- diffusion d'une ressource
- amplification locale
- croissance ou reproduction.

Cette combinaison produit des structures ramifiées ou fractales.

Le cadre diffusion-réaction constitue donc un modèle minimal permettant de décrire des systèmes auto-organisés.

5. Passage aux systèmes économiques

L'idée centrale de l'éconophysique est que les marchés financiers sont aussi des systèmes complexes adaptatifs.

Ils présentent en effet plusieurs propriétés remarquables :

- distributions à queue lourde
- clustering de volatilité
- structures multi-échelles
- intermittence.

Ces propriétés ressemblent à celles observées dans de nombreux systèmes physiques.

On peut donc se demander si un cadre conceptuel proche de celui des systèmes diffusion-réaction ou des processus de branchement pourrait également décrire certaines propriétés des marchés.

6. Analyse empirique du CAC 40

Pour explorer cette idée, nous avons analysé une série temporelle longue du CAC 40.

Les premières étapes ont consisté à étudier les propriétés d'échelle du signal.

Deux objets ont été examinés :

- les rendements

- la volatilité.

Les résultats sont classiques mais fondamentaux.

Rendements: Le coefficient de Hurst est proche de $H \approx 0.5$, ce qui correspond à un comportement proche d'un bruit blanc.

Volatilité: La volatilité présente en revanche une mémoire longue : $H \approx 0.7 – 0.8$ Cela signifie que les périodes de forte volatilité tendent à se regrouper.

Cette propriété est connue sous le nom de volatility clustering.

7. Dimension effective et lois d'échelle

En étudiant les propriétés spectrales et les corrélations, nous avons estimé un exposant spectral β. Dans un cadre inspiré des équations diffusion-réaction, il est possible de relier cet exposant à deux paramètres :

- une dimension effective d_{eff}
- un ordre de réaction p.

Les estimations obtenues pour le CAC sont approximativement : $p \approx 2.1 – 2.3$ et $d_{eff} \approx 2.2 – 2.5$.

Ces valeurs sont remarquables. Elles suggèrent que la dynamique collective du marché pourrait être dominée par des interactions effectives proches du quadratique.

8. Reconstruction de l'espace des phases

Pour comprendre la structure dynamique du système, nous avons reconstruit un espace des phases à l'aide de coordonnées retardées : $(x(t), x(t+\tau))$ puis $(x(t), x(t+\tau), x(t+2\tau))$.

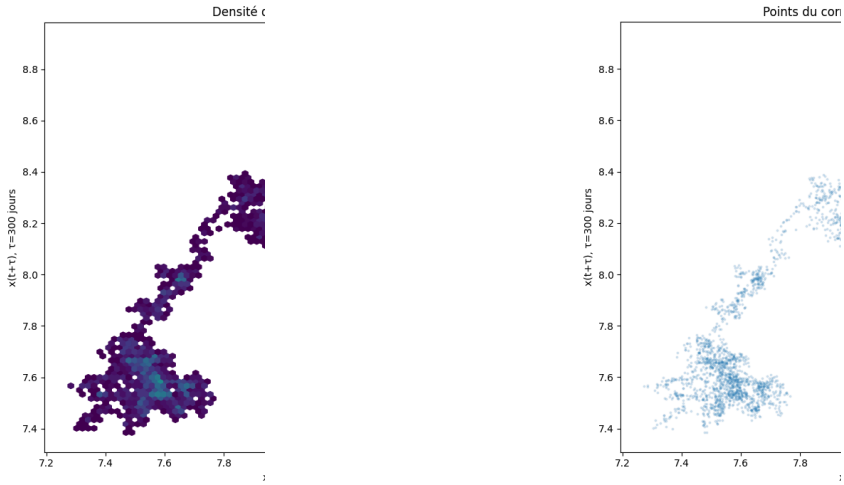
Cette technique provient de la théorie des systèmes dynamiques. Elle permet de reconstruire la géométrie de l'attracteur sous-jacent.

Dans le cas du CAC, cette reconstruction révèle plusieurs caractéristiques :

- des zones de forte densité
- des corridors de transition

- une structure multi-régime.

Autrement dit, le système ne remplit pas l'espace uniformément. Il visite préférentiellement certaines régions.



Ce que montrent les graphes

- Dans l'espace retardé

$$(x(t), x(t + \tau)), \tau = 300 jours$$

la densité n'est pas uniforme.

- On distingue bien **deux zones de forte densité** :
  - **Bassin 1** autour de (8.49, 8.54)
  - **Bassin 2** autour de (8.08, 8.22)
- Entre les deux, on voit une **bande de transition oblique**, que j'ai extraite comme **corridor de transition**.

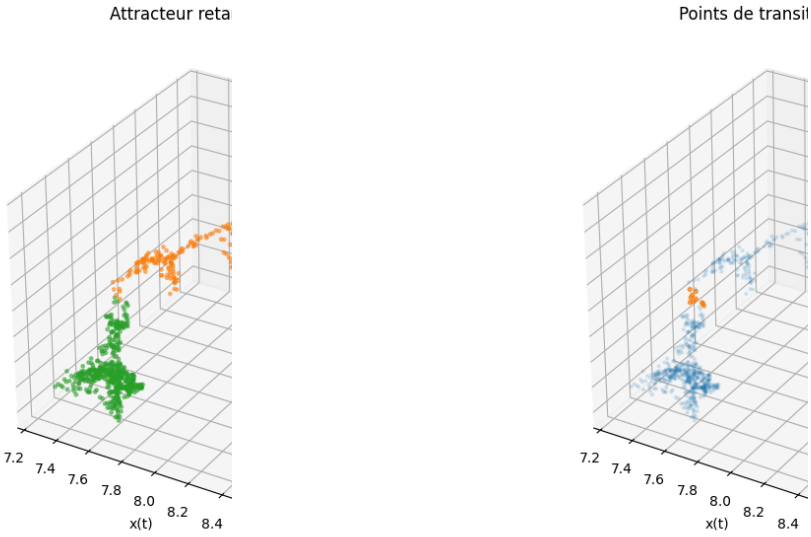
Résumé quantitatif

- **Bassin 1**
  - 3684 points
  - volatilité locale moyenne ≈ **0,0112**
- **Bassin 2**

- 4232 points
- volatilité locale moyenne ≈ **0,0139**

- **Corridor**
  - 498 points
  - volatilité locale moyenne ≈ **0,0110**

Interprétation pédagogique
Ce que cela veut dire, c'est que le CAC 40 ne remplit pas le plan (x(t),x(t+τ))(x(t),x(t+\tau))(x(t),x(t+τ)) comme le ferait un bruit purement aléatoire.
Au contraire, il visite préférentiellement **certaines régions** de cet espace.



Nous avons étudié l'espace retardé :

$$(x(t), x(t + \tau), x(t + 2\tau)), \tau = 300 jours$$

$$x(t) = log(CAC(t))$$

Résultat principal

Le critère BIC sélectionne **5 bassins** en 3D, pas seulement 2.

Autrement dit :

- en **2D**, on voyait surtout **deux grands bassins**
- en **3D**, on découvre que ces bassins se **décomposent en plusieurs sous-régimes**

C'est très important, car cela signifie que le marché n'oscille pas seulement entre "deux états", mais entre **plusieurs configurations dynamiques**.

Résumé quantitatif

Les 5 bassins détectés sont :

- **Bassin 1**
  - 612 points
  - volatilité locale moyenne ≈ **0,0116**
- **Bassin 2**
  - 2330 points
  - volatilité locale moyenne ≈ **0,0137**
- **Bassin 3**
  - 1392 points
  - volatilité locale moyenne ≈ **0,0115**
- **Bassin 4**
  - 2023 points
  - volatilité locale moyenne ≈ **0,0105**
- **Bassin 5**
  - 1259 points
  - volatilité locale moyenne ≈ **0,0148**
- **Corridor / faible confiance**
  - 1523 points
  - volatilité locale moyenne ≈ **0,0106**

Voici un **tableau synthétique** qui résume les 5 bassins identifiés dans l'espace retardé 3D, avec leur **signification économétrique** et les valeurs estimées dans notre analyse.

Bassin	**Signification économétrique**	**Volatilité moyenne (60 j)**	**HvolH_{vol}Hvol**	**βvol\beta_{vol}βvol**	**ppp (sid=2d=2d=2)**	**deffd_{\mathrm{eff}}deff (sip=2p=2p=2)**	**Lecture économique**
Bassin 1	Régime **diffus / instable** proche des retournements	0.0117	0.61	0.37	2.10	2.19	Zone de sommet ou transition vers un nouveau régime
Bassin 2	Régime **de transition haussière**	0.0132	0.71	0.50	2.15	2.25	Phase de reconfiguration ou reprise du marché
Bassin 3	Régime **bas stable**	0.0115	0.72	0.51	2.15	2.25	Marché relativement calme à niveau bas
Bassin 4	Régime **haut stable**	0.0103	0.72	0.75	2.24	2.38	Marché installé dans une tendance haussière durable
Bassin 5	Régime **de stress / crise**	0.0146	0.73	0.98	2.34	2.49	Phase de turbulence ou de crise

Bassin	Signification économétrique	Volatilité moyenne (60 j)	H_{vol}	β_{vol}	p_{pp} (sid=2d=2d=2)	d_{eff} (sip=2p=2p=2)	Lecture économique financière

Cela suggère que la dynamique fondamentale du marché est proche d'une **interaction quadratique**, ce qui correspond à :

- interactions binaires entre agents
- mécanisme d'imitation / contagion.

C'est exactement ce que prédisent plusieurs modèles d'économophysique.

9. Analyse en trois dimensions

En ajoutant une troisième coordonnée retardée, nous avons obtenu un attracteur tridimensionnel. L'analyse statistique montre l'existence d'environ cinq bassins principaux.

Ces bassins correspondent à différents régimes dynamiques du marché :

- régime bas stable
- transition haussière
- régime haut stable
- sommet ou retournement
- régime de crise.

Ces états ne sont pas fixes : le système passe de l'un à l'autre par des zones de transition.

10. Réduction de dimension par PCA

Pour comprendre la structure géométrique de cet attracteur, nous avons effectué une analyse en composantes principales (PCA).

Cette méthode consiste à trouver les directions dans lesquelles la variance des données est maximale.

Dans notre cas, les résultats sont frappants :

PC1 : 83 %

PC2 : 12 %

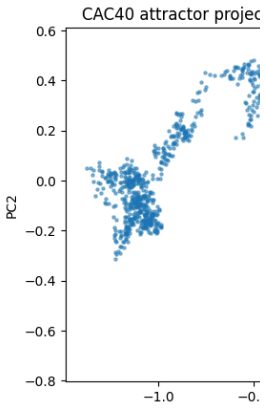
PC3 : 4 %

Ainsi :

$$PC_1 + PC_2 \approx 96\%$$

Cela signifie que la dynamique tridimensionnelle du système est presque entièrement contenue dans un plan bidimensionnel.

En pratique, deux variables dominent la dynamique collective du marché.



Quand on projette l'attracteur sur ces deux

---

axes :

on retrouve :

- plusieurs **amas de points**
- des **corridors de transition**

Donc la PCA confirme que :

- l'attracteur 3D peut être vu comme une **surface légèrement courbée**
- dans un espace **effectivement 2D**.

La dynamique collective du marché pourrait être dominée par **deux variables lentes** :

1️⃣**niveau moyen du marché** (tendance macro-financière)

2️⃣**accélération / retournement** (dynamique de régime)

Les autres fluctuations seraient essentiellement du **bruit ou des perturbations rapides**.

11. Interprétation des deux variables principales

Les deux directions principales peuvent être interprétées comme :

- une variable lente liée au niveau global du marché
- une variable liée à l'accélération ou aux retournements de tendance.

Les autres fluctuations apparaissent essentiellement comme des perturbations rapides ou du bruit.

12. Vers un modèle dynamique

Les résultats précédents suggèrent que le marché peut être décrit comme un système dynamique stochastique de faible dimension.

Une première formulation naturelle est un modèle de type Langevin :

$$dz = b(z)dt + \Sigma(z)dW_t$$

où

$$z = \left(PC_1, PC_2\right)$$

- b(z) est le drift
- Σ(z) la diffusion
- dW_t un bruit.

Dans certains cas, le drift peut être approximé par le gradient d'un potentiel :

$$b(z) = -\nabla U(z)$$

Le paysage de potentiel U(z) possède alors plusieurs minima correspondant aux bassins observés.

13. Équation pour la densité

L'équation de Langevin correspond à une équation pour la densité de probabilité des états :

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot (b\rho) + \nabla \cdot (D\nabla\rho)$$

C'est l'équation de Fokker-Planck.

Elle décrit l'évolution de la densité d'occupation dans l'espace des phases.

14. Extension diffusion-réaction

Si l'on introduit des effets collectifs ou d'imitation, on peut enrichir ce modèle en ajoutant un terme non linéaire :

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot (b\rho) + \nabla \cdot (D\nabla\rho) + R(\rho)$$

Un terme de réaction proche de

$$R(\rho) \sim \rho^{p}, p \approx 2$$

C' est cohérent avec les estimations obtenues à partir des lois d'échelle.

Dans ce cadre, les bassins correspondent à des cuvettes de potentiel renforcées par la dynamique collective.

15. Vision d'ensemble

L'ensemble du travail conduit à une image cohérente :

- les marchés financiers sont des systèmes complexes adaptatifs
- leur dynamique peut être décrite par quelques variables collectives
- ces variables évoluent dans un paysage dynamique possédant plusieurs régimes.

Le modèle le plus naturel combine :

- une dynamique de type Langevin
- une diffusion liée à la volatilité
- éventuellement une non-linéarité collective proche du quadratique.

16. Conclusion

La démarche suivie relie plusieurs niveaux d'analyse :

- théorie des systèmes diffusion-réaction et auto-similarité
- processus de branchement et fractales naturelles
- analyse empirique des lois d'échelle du CAC 40
- reconstruction de l'espace des phases
- réduction de dimension par PCA
- construction d'un modèle dynamique stochastique

Elle suggère que la dynamique collective des marchés pourrait être décrite par un système de faible dimension gouverné par des interactions effectives proches du quadratique.

Autrement dit, malgré la complexité apparente des marchés financiers, leur comportement pourrait émerger d'un petit nombre de mécanismes fondamentaux similaires à ceux observés dans de nombreux systèmes naturels auto-organisés.

Modélisation stochastique du signal du CAC 40 projeté sur la première composante principale

1. Point de départ : un signal complexe, mais essentiellement porté par une variable effective

L'étude commence par une réduction de dimension. Le CAC 40, pris comme système fluctuant, est un objet de grande complexité. Pourtant, lorsqu'on projette les données sur leur première composante principale, on constate que cette direction capte l'essentiel de la variance, de l'ordre de 80 %. Cela signifie qu'une grande partie de la dynamique utile peut être résumée par une seule variable effective, notée $x(t)$, qui représente le signal du marché projeté sur la direction principale de la PCA.

Cette observation est fondamentale. Elle ne signifie pas que le marché est véritablement unidimensionnel, mais qu'à l'échelle statistique qui nous intéresse ici, il existe une variable d'ordre dominante. Autrement dit, la dynamique collective du système peut être décrite, au premier ordre, comme le mouvement d'un point dans un espace d'état à une dimension. C'est exactement la situation dans laquelle la physique statistique devient pertinente : un grand nombre de degrés de liberté microscopiques peuvent être remplacés par une dynamique effective mésoscopique.

À partir de cette variable $x(t)$, on introduit sa dérivée empirique, notée $y(t)$, ainsi que son carré $y^2(t)$. En pratique, comme les données sont discrètes, on travaille avec la version journalière

$$y_n = x_{n+1} - x_n$$

ou, si l'on souhaite expliciter le pas de temps Δt,

$$y_n = \frac{x_{n+1} - x_n}{\Delta t}.$$

Dans nos simulations, le pas de temps vaut un jour de bourse, de sorte que l'on peut souvent prendre $\Delta t = 1$ sans perte de généralité. La variable y_n mesure les fluctuations rapides du signal, tandis que y_n^2 mesure l'énergie locale de ces fluctuations. C'est précisément y_n^2 qui porte la mémoire du système.

L'une des premières constatations empiriques a été que y se comporte approximativement comme un bruit centré, ou en tout cas comme une variable peu autocorrélée, alors que y^2 possède une autocorrélation lente et robuste. Cette dissymétrie entre le premier ordre et le second ordre est typique des séries financières. Les incréments sont peu prévisibles, mais leur intensité, elle, est fortement structurée dans le temps.

2. L'hypothèse centrale : une diffusion effective dans un paysage de potentiel

Puisque x résume l'essentiel de la dynamique, on fait l'hypothèse qu'il peut être modélisé par une équation différentielle stochastique de type diffusion. La forme générale de départ est

$$dX_t = F(X_t)dt + B_t dW_t,$$

où $F(X_t)$ est le drift, c'est-à-dire la composante déterministe de la dynamique, B_t est l'amplitude instantanée du bruit, et W_t est un mouvement brownien standard.

Cette équation est la forme continue la plus simple d'un système soumis à une force de rappel et à une agitation aléatoire. Si le coefficient B_t était constant, on serait dans le cadre classique d'une diffusion d'Itô. Mais les données montrent très vite que l'amplitude du bruit ne peut pas être constante, car la volatilité du système est intermittente et mémorielle.

Avant même de traiter la mémoire, il faut comprendre le drift $F(x)$. Pour cela, on exploite la densité empirique stationnaire de x, notée $p(x)$. Dans la théorie des diffusions, la densité stationnaire et le drift sont liés par l'équation de Fokker-Planck stationnaire. Dans le cas d'un bruit additif de variance constante, l'équation de Fokker-Planck s'écrit

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial}{\partial x}(F(x)p(x,t)) + \frac{D}{2}\frac{\partial^2 p(x,t)}{\partial x^2},$$

où D est le coefficient de diffusion, relié à l'amplitude du bruit par $D = B^2$ si B est constant. À l'état stationnaire, on a $\partial_t p = 0$. Si l'on suppose en plus un flux nul, ce qui est l'analogue d'un équilibre statistique effectif, alors on obtient

$$F(x)p(x) - \frac{D}{2}\frac{\partial p(x)}{\partial x} = 0.$$

En divisant par $p(x)$, on trouve

$$F(x) = \frac{D}{2}\frac{\partial}{\partial x} log p(x).$$

Cette formule donne immédiatement le drift à partir de la densité empirique. Si l'on introduit alors un potentiel effectif

$$U(x) = - \, log p(x),$$

on obtient

$$F(x) = - \frac{D}{2} U'(x).$$

Autrement dit, la dynamique peut être vue comme celle d'une particule qui évolue dans un paysage de potentiel $U(x)$, avec une force de rappel

$$- \frac{dU}{dx}.$$

Dans la pratique, nous avons utilisé une version réduite de cette relation, écrite sous la forme

$$F(x) = \gamma \frac{d}{dx} log p(x),$$

où γ est un coefficient d'échelle ajustable. Ce paramètre absorbe à la fois le niveau effectif de diffusion, les approximations faites dans la reconstruction de $p(x)$ par noyau, et le fait que l'on travaille sur un système empirique qui n'est pas un équilibre thermodynamique strict. La forme fonctionnelle du drift est donc entièrement dictée par la densité de probabilité de x, et non imposée a priori.

La reconstruction de $U(x) = - \, log p(x)$ à partir des données montre un paysage multi-cuvettes. Ces cuvettes sont des minima locaux de potentiel. Elles correspondent à des zones où le signal tend à séjourner plus longtemps. Ce sont des régimes préférentiels, ou, si l'on veut parler comme en physique, des puits de potentiel.

3. Pourquoi un simple bruit brownien ne suffit pas

Si l'on s'arrêtait à l'équation

$$dX_t = F\left(X_t\right)dt + BdW_t$$

avec B constant, on obtiendrait une diffusion classique dans un potentiel. Mais les observations montrent immédiatement que ce modèle est insuffisant.

La première raison est que l'intensité des fluctuations n'est pas constante. Les données de y^2 présentent des paquets de volatilité, des épisodes de calme alternant avec des épisodes d'agitation. Cela signifie que le bruit qui pousse x ne peut pas avoir une variance stationnaire constante. Il faut donc remplacer B par une amplitude fluctuante.

La seconde raison est spectrale. Le signal x réel possède un spectre de puissance très proche de $\frac{1}{f^2}$ sur une large bande de fréquences. Cela signifie que le forçage ne peut pas être purement blanc. Un bruit blanc injecté directement dans x produirait un signal trop rugueux ou, selon les réglages, mal réparti fréquentiellement. Il faut donc distinguer deux choses : l'intensité locale du bruit, qui est liée à la volatilité, et la couleur fréquentielle du forçage, qui détermine le contenu en basses et hautes fréquences.

C'est cette double nécessité qui conduit à décomposer le terme aléatoire en deux briques distinctes : une amplitude aléatoire S_t, ou σ_t si l'on préfère cette notation, et un forçage normalisé u_t. Le modèle devient alors

$$dX_t = F(X_t)dt + \sigma_t u_t dt$$

dans une écriture continue simplifiée, ou en temps discret

$$x_{n+1} = x_n + F(x_n)\Delta t + \sigma_n u_n.$$

Le terme σ_n mesure l'intensité locale des fluctuations. Le terme u_n porte la texture fréquentielle.

4. La mémoire de volatilité et la construction du noyau

L'analyse empirique de y^2 montre que son autocorrélation décroît selon une loi de puissance

$$C_{y^2}(\tau) \sim A\tau^{-m},$$

avec un exposant m mesuré numériquement. Cette décroissance lente signifie que la volatilité n'a pas une mémoire exponentielle simple. Il ne s'agit pas d'une relaxation mono-échelle. Au contraire, on est en présence d'une mémoire distribuée sur de nombreuses échelles de temps.

Pour traduire cette propriété, on introduit un champ latent gaussien corrélé, noté ici z_n, dont la fonction d'autocorrélation suit elle-même une loi de puissance

$$Corr(z_n, z_{n+\tau}) \approx A\tau^{-m}.$$

En pratique, ce champ est construit numériquement par synthèse spectrale, en partant d'une séquence de corrélation cible $\rho(\tau)$, puis en utilisant un plongement circulant dans l'espace de Fourier. Cela permet de fabriquer un processus gaussien corrélé ayant la mémoire voulue.

Une fois ce champ z_n obtenu, on construit la volatilité latente sous la forme lognormale

$$\sigma_n^2 = \sigma'^2 exp\left(A_\sigma z_n - \frac{1}{2}A_\sigma^2\right).$$

Ici, σ'^2 est le niveau moyen de variance, estimé à partir des données, et A_σ est un paramètre d'amplitude. Dans plusieurs essais, on a fixé $A_\sigma = 1$, de manière à garder une construction minimale et contrôlable.

Le terme $-\frac{1}{2}A_\sigma^2$ est introduit pour recentrer l'exponentielle et préserver la moyenne. En effet, si z_n est centré et de variance unitaire, alors

$$E[e^{A_\sigma z_n - \frac{1}{2}A_\sigma^2}) = 1.$$

On a donc

$$E[\sigma_n^2) = \sigma'^2.$$

Cette construction donne naturellement une volatilité intermittente, avec des épisodes de forte agitation et des épisodes de calme, tout en incorporant une mémoire longue. C'est précisément ce que l'on cherche à reproduire dans y^2.

L'idée de noyau mémoire apparaît ici de manière naturelle. Si l'on écrivait le modèle de façon formelle, on pourrait dire que z_n est le résultat de la convolution d'un bruit blanc avec un noyau mémoire K,

$$z_n = \sum_{k\geq 0} K_k \varepsilon_{n-k},$$

où K_k décroît lentement, en loi de puissance. La synthèse spectrale que nous avons utilisée est une manière numérique efficace de produire un tel objet sans avoir à tronquer explicitement la somme ni à gérer directement toutes les convolutions.

5. Le forçage coloré : pourquoi il faut une seconde mémoire

Le seul fait de donner une mémoire à σ_n ne suffit pas à reproduire la structure fréquentielle de x. Le spectre du signal réel projeté sur PCA1 montre une pente proche de -2,

$$S_x(f) \sim \frac{1}{f^2},$$

sur une large bande de fréquence. Cette propriété suggère que la composante aléatoire qui pousse x possède elle-même une structure lente, rouge ou colorée, et non pas purement blanche.

Pour reproduire cela, on introduit un second processus, noté v_n, qui représente la composante colorée du forçage. On le construit comme un processus de type Ornstein-Uhlenbeck discrétisé :

$$v_{n+1} = \rho v_n + \sqrt{1 - \rho^2}\eta_n,$$

où η_n est un bruit blanc gaussien centré réduit, et où

$$\rho = e^{-\frac{1}{\tau}}.$$

Le paramètre τ est la constante de temps de corrélation du bruit coloré. Dans les différents essais, nous avons testé des valeurs de l'ordre de 100, 200 ou 300 jours. Une valeur de τ grande donne un forçage plus lent, plus riche en basses fréquences.

Ce processus v_n a une autocorrélation exponentielle

$$Corr\left(v_n, v_{n+\tau}\right) \sim e^{-\frac{\tau}{\tau_c}},$$

et un spectre de puissance en forme de passe-bas, qui décroît comme $\frac{1}{f^2}$ au-delà de la fréquence de coupure. C'est précisément la propriété recherchée pour reproduire le caractère très rouge du signal x.

6. Le forçage hybride : addition de la composante colorée et de la composante blanche

Les essais numériques ont montré qu'un bruit trop coloré rendait x réaliste sur le plan des basses fréquences, mais faisait perdre à y une partie de sa texture rapide. Inversement, un bruit trop blanc rendait y plus vivant, mais dégradait la structure fréquentielle de x. La solution naturelle a donc été de construire un forçage hybride, somme d'une composante colorée et d'une composante blanche.

On écrit ainsi

$$u_n^{raw} = av_n + b\xi_n,$$

où v_n est la composante colorée, ξ_n la composante blanche, et a et b sont deux coefficients qui règlent leur importance relative.

Dans la simulation retenue ici, nous avons pris

$$a = 1, b = 0.2,$$

c'est-à-dire

$$u_n^{raw} = v_n + 0.2\xi_n.$$

Comme nous souhaitons que le forçage conserve une variance contrôlée, nous normalisons ensuite cette somme :

$$u_n = \frac{v_n^{raw} - \langle v^{raw} \rangle}{std\left(v^{raw}\right)}.$$

Cette normalisation permet de comparer les simulations sur une base stable, sans que les coefficients a et b modifient automatiquement la variance totale du forçage. Leur rôle devient alors essentiellement fréquentiel et morphologique.

Le spectre de $av + b\xi$ montre très bien la logique du modèle. À basse fréquence, la composante v domine. À haute fréquence, la composante blanche ξ empêche le spectre de s'effondrer trop vite. Le forçage hybride contient donc à la fois la lenteur nécessaire pour reproduire la structure de x et la rugosité locale nécessaire pour approcher la texture de y.

7. L'équation complète du simulateur

À ce stade, le modèle peut être écrit de manière complète. En temps discret, l'équation d'évolution du signal simulé s'écrit

$$x_{n+1} = x_n + \gamma \frac{d}{dx} logp\left(x_n\right) + \sigma_n u_n,$$

avec

$$\sigma_n^2 = \sigma'^2 exp\left(A_\sigma z_n - \frac{1}{2}A_\sigma^2\right),$$

où z_n est un processus gaussien corrélé ayant une loi de corrélation en puissance,

$$Corr\left(z_n, z_{n+\tau}\right) \sim A\tau^{-m},$$

et

$$u_n = \frac{av_n + b\xi_n - \langle av + b\xi \rangle}{std(av + b\xi)},$$

avec

$$v_{n+1} = \rho v_n + \sqrt{1-\rho^2}\eta_n, \rho = e^{-\frac{1}{\tau_c}}.$$

La dynamique de x est donc déterminée par trois briques : la géométrie du potentiel via $\gamma \frac{d}{dx} logp(x)$, l'intermittence de volatilité via σ_n, et la structure fréquentielle du forçage via u_n.

En version continue formelle, le même modèle peut être pensé comme

$$dX_t = F(X_t)dt + \sigma_t u_t dt,$$

avec

$$F(x) = \gamma \frac{d}{dx} logp(x),$$

et où u_t n'est plus un bruit blanc idéal mais un processus coloré-blanc hybride.

8. Comment lire les différentes composantes du modèle

Le drift dérivé de $p(x)$ fixe la topologie du paysage. Il détermine où sont les puits de potentiel et les barrières entre eux. La dynamique y est contrainte mais pas figée.

La volatilité latente σ_n agit comme une température interne du système. Lorsqu'elle est faible, la trajectoire reste proche du drift. Lorsqu'elle est forte, les fluctuations s'amplifient et les barrières deviennent plus facilement franchissables.

Le forçage hybride u_n est le moteur fréquentiel du système. Sa composante colorée donne des structures lentes, des déplacements larges, des variations de régime. Sa composante blanche redonne au signal sa granularité fine, sa rugosité locale et une partie de son imprédictibilité instantanée.

La combinaison des trois permet d'obtenir un simulateur complet, à la fois topologique, intermittent et fréquentiellement structuré.

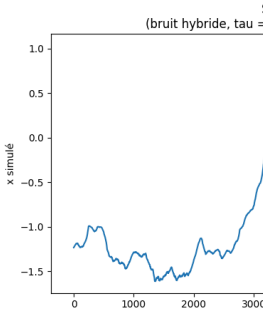
9. Ce que les simulations montrent

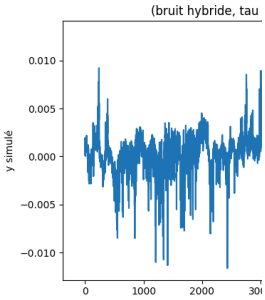
L'enchaînement des simulations a mis en évidence plusieurs comportements importants. Lorsque le bruit est purement gaussien blanc avec une amplitude insuffisante, le système reste piégé dans un puits. Lorsque l'on augmente trop brutalement l'amplitude, il explore mieux les puits, mais au prix d'un signal x trop rugueux et d'un contenu fréquentiel insuffisamment rouge.

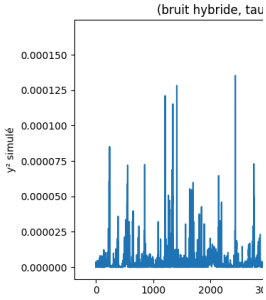
Lorsque l'on introduit un bruit coloré pur à longue constante de temps, x devient très réaliste sur le plan basse fréquence, mais y perd une partie de sa texture fine, et la trajectoire peut devenir trop polarisée entre deux grands puits.

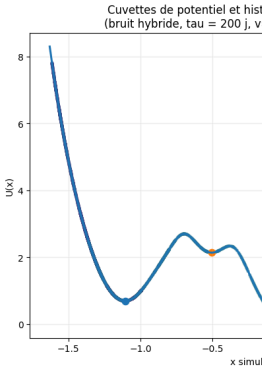
Le modèle hybride permet de résoudre une partie de cette tension. Avec une composante colorée dominante et une composante blanche modeste, le signal simulé devient capable de circuler entre les principales cuvettes du potentiel, tout en conservant une microstructure plus réaliste dans y. En parallèle, la queue de distribution de y^2 reste proche des exposants observés empiriquement, ce qui montre que l'intermittence est correctement reproduite.

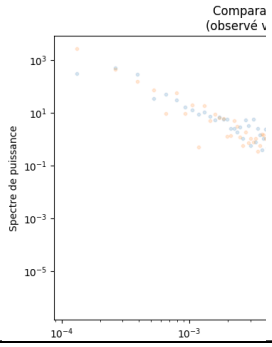
La simulation finale avec v comme composante colorée de coefficient 1, ξ comme composante blanche de coefficient 0,2, une constante de temps de 200 jours pour v, un coefficient de sigma de 0,2, et une mémoire de volatilité construite à partir de l'autocorrélation de y^2, donne une image topologiquement convaincante du signal. La trajectoire visite les grands puits du potentiel, la texture de y reste plausible, la structure de y^2 est intermittente, et le spectre du forçage $av + b\xi$ rend visible le compromis entre basses et hautes fréquences.











J'ai refait la simulation avec votre correction :

- **composante colorée** v(n)v(n)v(n) avec coefficient **1**
- **composante blanche** ξ(n)\xi(n)ξ(n) avec coefficient **0,2**

- **coefficient de sigma = 0,2**

$$\tau = 200 jours$$

Plage de xxx simulé :

$$x_{min} \approx -1.61, x_{max} \approx 1.03$$

Donc, cette fois, la trajectoire explore bien une grande partie du paysage.

Répartition dans les puits :

- puits vers −1.106- : **39.0 %**
- puits vers −0.507 : **2.7 %**
- puits vers 0.000. : **5.4 %**
- puits vers 0.4490.9 : **52.9 %**

Donc le système circule surtout entre les deux puits principaux, avec peu d'occupation des puits intermédiaires.

Queue de y2 simulé :

$$CCDF \approx 1.72$$

avec

$$R^2 \approx 0.985$$

C'est un résultat assez proche de votre cible, surtout du côté Hill.

Lecture physique

Ce réglage est intéressant parce qu'il donne un compromis crédible :

- x explore bien les deux grands puits ;
- y garde une texture rapide grâce à

la composante blanche ;

- le spectre du forçage av+bξa montre bien le mélange entre :
    - basses fréquences dominées par v,
    - hautes fréquences enrichies par ξ.

10. Interprétation conceptuelle de l'ensemble

Ce modèle montre que les propriétés principales du signal du CAC 40 projeté sur PCA1 sont, à un niveau effectif, quasi unidimensionnelles. La complexité du système ne disparaît pas, mais elle se concentre dans trois mécanismes clairement séparés.

Le premier mécanisme est géométrique. Il est porté par la densité empirique de x, donc par le potentiel $U(x)$. Il fixe les régimes préférentiels et les barrières de transition.

Le deuxième mécanisme est énergétique. Il est porté par σ_n, donc par la mémoire longue de la volatilité. Il détermine quand le système est calme et quand il devient agité.

Le troisième mécanisme est fréquentiel. Il est porté par le forçage hybride u_n. Il règle le partage entre les fluctuations rapides et les structures lentes.

Ce découplage est très précieux, car il permet d'interpréter séparément les effets. Si le signal manque de transitions, ce n'est pas nécessairement le potentiel qui est en cause. Si y manque de microstructure, ce n'est pas nécessairement σ_n qui doit être modifié. Si le spectre de x est trop plat, c'est peut-être le poids de la composante colorée qui est insuffisant.

Le simulateur n'est donc pas seulement un modèle numérique. C'est aussi une grille de lecture structurée du signal.

11. Équation de Fokker-Planck associée et portée de la modélisation

Si l'on fige temporairement σ_t et si l'on idéalise u_t comme un bruit brownien effectif, alors la densité de probabilité $p(x, t)$ du processus vérifie une équation de Fokker-Planck

$$\frac{\partial p}{\partial t} = -\frac{\partial}{\partial x}(F(x)p) + \frac{1}{2}\frac{\partial^2}{\partial x^2}(D(x,t)p),$$

où $D(x, t)$ est la diffusion effective, reliée à σ_t^2.

Dans notre cadre complet, le système n'est pas markovien en x seul, puisque σ_t porte une mémoire longue et que le forçage u_t a lui-même une corrélation temporelle. La description exacte devrait donc être formulée dans un espace d'état élargi, comprenant au moins x, la variable de volatilité latente, et l'état du bruit coloré. Mais la logique de Fokker-Planck reste essentielle pour la reconstruction du drift : c'est elle qui justifie le passage de la densité empirique de x au potentiel effectif.

Il faut donc comprendre le modèle comme une généralisation empirique d'une diffusion de Fokker-Planck, enrichie par une volatilité à mémoire et un forçage hybride.

12. Conclusion générale

Le travail conduit à un résultat important. Le signal du CAC 40 projeté sur la première composante principale peut être représenté, de manière acceptable au plan topologique et statistique, par un simulateur unidimensionnel enrichi. Ce simulateur associe un drift dérivé d'un potentiel empirique, une volatilité intermittente à mémoire longue, et un forçage hybride mélangeant une composante colorée et une composante blanche.

Les équations du modèle sont suffisamment simples pour être interprétables, mais assez riches pour reproduire les propriétés essentielles du signal. Le système contient des cuvettes de potentiel, il présente une intermittence robuste sur y^2, et la variable y peut, avec un réglage adéquat, tendre vers la texture observée.

Ce point est très fort conceptuellement. Il signifie que, derrière la complexité apparente du marché, il existe une dynamique effective relativement compacte, dont les propriétés principales sont portées par un petit nombre de mécanismes identifiables. La topologie vient de la densité de x. La mémoire vient de y^2. La couleur fréquentielle vient du forçage hybride. L'ensemble forme un simulateur complet, cohérent et physiquement lisible.

INTÉRÊT DU MODÈLE EN DATA SCIENCE

Ce modèle a un intérêt réel en data science, justement parce qu'il se situe à l'interface entre trois manières de travailler qui sont souvent séparées : la réduction de dimension, l'inférence statistique non paramétrique et la modélisation générative.

Le premier intérêt est qu'il transforme un problème de marché très complexe en une chaîne de travail lisible. On part d'un objet empirique très riche, le CAC 40, puis on le projette sur une variable dominante via l'analyse en composantes principales (PCA). On reconstruit ensuite un drift à partir de la densité de probabilité p(x), puis on identifie la mémoire de second ordre à partir de la série y^2, et enfin on construit un simulateur cohérent avec ces contraintes.

Pour un data scientist, c'est précieux, parce que cela force à passer d'une logique de simple prédiction à une logique de modèle explicatif et génératif. Autrement dit, on ne cherche pas seulement à ajuster une courbe, on cherche à reconstruire un mécanisme latent qui explique pourquoi le signal a telle topologie, telle intermittence et telle structure fréquentielle.

Cette manière de faire est très proche de l'esprit des modèles de diffusion et de l'inférence mécaniste en sciences des données. Les travaux de non-paramétrie sur les diffusions montrent précisément comment remonter des observations discrètes vers les coefficients d'un processus stochastique.

Le deuxième intérêt est qu'il donne une méthode de travail très structurée pour un data scientist. D'abord, il oblige à séparer les niveaux d'analyse. La densité de x renseigne sur la géométrie des régimes. L'autocorrélation de y^2 renseigne sur la mémoire de volatilité. Le spectre de x renseigne sur la couleur du forçage.

Cette décomposition en couches est extrêmement utile, car elle évite de tout mélanger dans une seule boîte noire. Concrètement, cela apprend au data scientist à travailler par diagnostics successifs : distribution, corrélation, spectre, puis seulement simulation.

Troisième intérêt : ce modèle s'inscrit dans une littérature scientifique abondante.

Le potentiel est reconstruit à partir de la densité empirique selon l'équation :

$$U(x) = -\ logp(x)$$

La dynamique stochastique peut alors être écrite sous la forme d'une équation différentielle stochastique :

$$dX_t = F\left(X_t\right)dt + B\left(X_t\right)dW_t$$

Les modèles de volatilité stochastique introduisent une dynamique supplémentaire, par exemple :

$$dv_t = -\frac{1}{\tau}v_t dt + \sigma_v dW_t$$

Dans de nombreux modèles multifractals, la volatilité suit une structure log-normale :

$$\sigma_n^2 = \sigma'^2 exp\left(Az_n - \frac{1}{2}A^2\right)$$

où z_n est un processus gaussien corrélé.

Ces équations montrent comment la dynamique peut être séparée entre drift, volatilité et structure du bruit.

Ce que ce modèle apporte à un data scientist est aussi méthodologique. Il apprend à relier directement les objets statistiques aux objets dynamiques :

densité → potentiel
autocorrélation de y^2 → noyau mémoire
spectre de x → structure du forçage

Un bon modèle doit reproduire simultanément :

- la topologie des puits de potentiel
- la forme de la densité
- la mémoire de la volatilité
- le spectre de puissance
- la texture locale du signal

Ce type de modèle forme ainsi un pont entre la data science et la physique de la complexité.

Il permet de construire un jumeau statistique génératif et interprétable qui peut servir à comprendre la structure profonde du signal, générer des scénarios plausibles, tester des hypothèses ou réaliser des analyses de stress.

Au fond, ce modèle propose une manière de faire de la data science qui articule exploration, théorie et simulation dans un même cadre scientifique.