

Détection d'anomalies non-supervisée sur des chocs de marché par piliers

Introduction

Dans le contexte de la finance de marché, il est crucial d'identifier des **valeurs aberrantes** (anomalies) dans les données de chocs de marché. Le dataset considéré comporte plus de 4 millions de lignes de chocs (variations) par *piliers* de maturité (ex : 6M, 1Y, 10Y, etc.), avec pour chaque point un identifiant de structure de taux (*rf_struct_id*), un identifiant de scénario (*scenario_id*) et une date de valorisation (*as_of_date*). L'objectif est de détecter sans *labels* (i.e. en **non-supervisé**) les chocs anormaux qui pourraient résulter d'erreurs de données ou d'événements inhabituels.

Détecter ces anomalies améliore la qualité des analyses de risque et de stress tests en évitant qu'un modèle soit faussé par des entrées aberrantes mdpi.com. Une *anomalie* est généralement définie comme « *une observation dont le comportement diffère significativement du reste des données* » mdpi.com. Dans un tel dataset, il peut s'agir par exemple d'un choc de courbe de taux extrêmement élevé sur un pilier donné, ou d'une structure de chocs incohérente par rapport aux tendances habituelles.

Nous passons en revue les principales familles de méthodes non-supervisées de **détection d'anomalies** applicables dans ce contexte : (1) approches *univariées classiques* sur chaque série individuelle, (2) méthodes de *séries temporelles* exploitant la dimension temporelle, (3) techniques *multivariées* considérant l'ensemble des piliers simultanément, (4) modèles *avancés* récents (deep learning, graphes, théorie des extrêmes, GANs, etc.), et enfin (5) méthodes *hybrides* ou en *ensemble*. Pour chaque famille, nous comparons les forces, limites et contextes d'application, avec des exemples tirés de la finance de marché (risques de marché, calibration de modèles, stress testing...). Des **tableaux comparatifs** synthétisent les caractéristiques des méthodes, et nous proposons en conclusion des recommandations pratiques pour les appliquer efficacement sur un grand volume de données (4M+ lignes).

Méthodes univariées classiques

Les approches univariées traitent chaque choc individuellement ou chaque série de chocs (par exemple la suite temporelle des chocs 10Y) séparément. Elles reposent souvent sur des statistiques simples pour repérer des valeurs exceptionnellement élevées ou faibles par rapport à la distribution habituelle. Ces techniques sont faciles à mettre en œuvre et peu coûteuses en calcul, ce qui est un atout pour un premier filtrage sur des millions de points. En revanche, elles ne capturent pas les relations entre piliers ni les comportements multidimensionnels. Voici les principales méthodes univariées :

- **Score Z (règle des 3 sigma)** – On calcule le z-score de chaque observation par rapport à la moyenne et l'écart-type de la série : $z = (x - \mu) / \sigma$. Sous l'hypothèse d'une distribution normale, ~99,7% des valeurs devraient tomber dans $\pm 3\sigma$ (d'où

la *règle des 3 sigma*). Un choc dont le $|z| > 3$ est alors considéré comme **outlier** potentiel. Cette méthode est simple mais suppose une distribution gaussienne; en présence de données asymétriques ou à queue épaisse, elle peut être inadéquate numberanalytics.com. En effet, un choc extrême peut **biais** la moyenne et σ et ainsi échapper à la détection ou au contraire provoquer de fausses alertes. Son avantage est la facilité d'interprétation (combien d'écarts-types du centre), mais sa **sensibilité aux outliers** mêmes en fait une approche à utiliser surtout sur des données proches de la normale.

- **Écart interquartile (IQR) et boîte à moustaches** – C'est une autre méthode classique : on calcule les 1er et 3e quartiles (Q1 et Q3) de la distribution des chocs, puis on définit les bornes à $Q1 - 1,5 \times IQR$ et $Q3 + 1,5 \times IQR$ (avec $IQR = Q3 - Q1$). Toute valeur en dehors de ces bornes est considérée comme aberrante. Cette règle de Tukey ne présuppose pas de forme de distribution et est plus **robuste** que le z-score aux distributions asymétriques. Cependant, si la série présente de la *saisonnalité* ou des changements de tendance, l'IQR global peut être trop large et masquer certaines anomalies locales. C'est un outil simple souvent visualisé via les *boîtes à moustaches* en statistique descriptive.
- **MAD (Median Absolute Deviation)** – Le MAD est un estimateur robuste de la dispersion : il s'agit de la médiane des écarts absolus à la médiane de la série. On peut définir un score robuste $z_{MAD} = 0.6745 \times \frac{|x - \text{médiane}|}{MAD}$ (0.6745 étant un facteur pour rendre ce score comparable à un z-score normal). Des seuils (typiquement ± 3.5) sur ce score permettent de détecter des outliers de manière **beaucoup moins influencée par les valeurs extrêmes** numberanalytics.com. En effet, la médiane et le MAD ne bougent pas significativement en présence d'anomalies isolées, contrairement à la moyenne et σ classiques. Ainsi, le MAD est recommandé dans les jeux de données financiers où quelques chocs extrêmes peuvent exister aux côtés d'une majorité de variations modérées. Il est efficace même si la distribution n'est pas gaussienne, et résiste bien aux **données à fortes queues** (courantes en finance) numberanalytics.com. Son inconvénient est qu'il faut fixer un seuil (comme 3.5) de manière un peu empirique, mais celui-ci est en général plus fiable que le 3σ lorsque des anomalies sont attendues numberanalytics.com.
- **Désaisonnalisation et tendances (STL)** – Lorsqu'une série univariée présente des composantes saisonnières ou de tendance, celles-ci peuvent masquer ou faire ressembler à des anomalies des variations normales. La méthode **STL** (*Seasonal and Trend decomposition using Loess*) permet de décomposer une série temporelle en trois composantes : tendance, saisonnalité et résidu neptune.ai. En soustrayant la saisonnalité et la tendance estimées, on obtient un

résidu stationnaire sur lequel on peut appliquer une détection d'outliers (par exemple via un seuil sur l'écart-type du résidu). Cette approche est utilisée par exemple par Twitter pour détecter des anomalies dans les métriques web. Elle est **robuste** et fournit une interprétation claire (les anomalies se voient dans la composante résiduelle)neptune.ai. STL est particulièrement utile si les chocs de marché ont des patterns réguliers (ex : effets de fin de mois, cyclicité économique) qu'on souhaite neutraliser avant de chercher des valeurs aberrantes. En revanche, son paramétrage est rigide : il faut choisir une période de saisonnalité, et à part le choix du seuil sur le résidu, il y a peu d'hyperparamètres pour ajuster la sensibiliténeptune.ai.

En résumé, les méthodes univariées offrent un **premier niveau de filtrage** simple : elles signalent les chocs isolés anormalement grands sur un pilier donné. Elles sont rapides sur un gros volume (on peut les calculer en une passe sur les données). Cependant, elles ne repèrent pas des anomalies plus subtiles comme une combinaison inhabituelle de mouvements modérés sur plusieurs piliers. De plus, en pratique, il faut parfois combiner plusieurs critères (par ex., un choc pourrait ne pas dépasser 3σ mais être le seul à avoir un signe opposé aux autres piliers du scénario).

Tableau 1 : Comparatif des méthodes univariées classiques

Méthode	Principe	Forces	Limites
Z-score (3σ)	Écart en σ par rapport à la moyenne	Simple, intuitif (suppose normalité)	Sensible aux outliers, suppose distribution gaussienne
IQR ($1.5 \times \text{IQR}$)	Intervalle interquartile (Tukey)	Aucune hypothèse de loi, facile à visualiser	Manque de sensibilité si distribution très étalée ou multimodale
MAD (écart médian absolu)	Dispersion via la médiane	Robuste aux valeurs extrêmes, adapté aux queues épaisses numberanalytics.com	Besoin d'un seuil calibré ($k \times \text{MAD}$), moins intuitif que σ
STL + seuil sur résidu	Décomposition saison/trend + outliers sur résidus	Gère saisonnalité, anomalies plus interprétables neptune.ai	Paramétrage rigide, suppose série suffisamment longue et périodique

Méthodes de séries temporelles

Au-delà des statistiques univariées simples, on peut exploiter la dimension temporelle des données de chocs. Ici, on modélise l'évolution d'une série dans le temps à l'aide

d'un modèle, puis on identifie comme anomalies les points qui s'écartent significativement de ce que prédit le modèle. L'idée est que les séries financières suivent en général certains patterns (retour à la moyenne, volatilité conditionnelle, etc.), et qu'un **écart de modèle** important peut signaler un outlier. Ces approches sont utiles pour détecter des *anomalies contextuelles* (par exemple un choc qui n'est pas aberrant en valeur absolue mais qui est inhabituel compte tenu de la tendance récente). Toutefois, elles nécessitent de bien spécifier le modèle et peuvent être coûteuses si on doit ajuster un modèle par série de pilier et par structure.

- **Modèles ARIMA (Auto-Regressive Integrated Moving Average)** – Les modèles ARIMA et variantes (ARMA, AR, MA) sont des modèles de prédiction linéaire largement utilisés en séries financières. Une approche classique d'anomalie consiste à estimer un ARIMA sur la série temporelle d'un pilier (éventuellement après différenciation pour stationnariser), puis à examiner les *résidus* du modèle : toute observation dont le résidu (erreur de prédiction) est statistiquement trop grand indique une anomalie. Des travaux comme Chen & Liu (1993) ont formalisé des tests pour détecter différents types d'**outliers temporels** dans un ARIMA (outlier additif vs innovateur, changement de niveau, etc.)efmaefm.org. En pratique, des outils comme la procédure *TSOUTLIERS* dans R utilisent ces tests pour identifier et corriger des points aberrants lors de l'ajustement d'un ARIMA. L'avantage d'ARIMA est de **profiler la "norme"** de la série en tenant compte de sa dynamique propre (autocorrélations, tendance). Il peut distinguer, par exemple, un choc isolé atypique d'une simple prolongation d'une tendance haussière normale. Cependant, cette approche souffre de **plusieurs limites** : (a) nécessité de choisir l'ordre du modèle (p, d, q) adéquat, ce qui peut être ardu sans intervention manuellemdpi.com ; (b) suppositions de stationnarité et de distribution souvent trop restrictives pour des séries financières hétérogènesmdpi.com ; (c) performance computationnelle – ajuster un ARIMA par pilier et par scénario peut être lourd sur 4M de lignes, à moins d'automatiser et de restreindre l'analyse à des sous-ensembles. En résumé, ARIMA est puissant pour détecter des *anomalies point-in-time* dans une série donnée, mais son déploiement sur un grand univers de séries doit être soigneusement calibré.
- **Modèles GARCH (Generalized Autoregressive Conditional Heteroskedasticity)** – Les modèles GARCH sont conçus pour modéliser la **volatilité variable dans le temps** des séries financières (volatility clustering). Ils pourraient ne pas sembler de prime abord des outils de détection d'anomalies, mais ils sont utiles pour **normaliser** les variations. En effet, un choc de marché peut paraître énorme en période calme mais normal en période de forte volatilité. Un GARCH ajusté sur la série de chocs d'un pilier fournit à chaque date une estimation de la variance conditionnelle σ^2_t ; on peut alors regarder les *résidus*

standardisés (rapport du choc à σ_t). Si ces résidus dépassent par exemple 3 ou 4 en valeur absolue, on a un outlier compte tenu du niveau de volatilité courant. Cette approche capte des anomalies **en tenant compte du régime de marché** – e.g. un mouvement de 50bps peut être banal lors d’une crise mais extraordinaire en temps normal. L’utilisation de lois à queue épaisse (Student-t) dans GARCH permet en outre de réduire la sensibilité aux outliers, car l’ajustement n’est pas trop influencé par quelques chocs extrêmes. Malgré cela, la présence de données aberrantes peut biaiser l’estimation des paramètres GARCH (d’où des variantes robustes dans la littérature). En pratique, on pourra ajuster un GARCH(1,1) sur chaque série de taux d’intérêt ou d’écart de crédit, puis **inspecter les pics de résidus**. Les travaux montrent par exemple que des procédures spécifiques peuvent identifier et traiter les outliers dans les GARCH econstor.eu/papers/ssrn.com. L’avantage est de détecter les anomalies *de retour* (un saut bien au-delà de la volatilité attendue) et de distinguer “volatilité accrue” vs “vraie anomalie”. La limite est que cela reste un modèle paramétrique avec des hypothèses (stationnarité hors anomalies, etc.) et qu’il faut avoir suffisamment de données historiques pour bien calibrer le GARCH avant de faire confiance aux signaux d’outliers.

- **Méthodes de prévision modernes (Prophet)** – *Prophet* est un outil de prévision développé à l’origine par Facebook, qui combine des termes de tendance, de saisonnalité et des holidays dans un modèle additif ajusté de manière robuste (via une vraisemblance Student). Prophet peut servir à la détection d’anomalies en tant que **filtre robuste** : on ajuste le modèle sur l’historique d’une série et on regarde où les prédictions s’écartent des observations. Prophet a l’avantage d’être **peu sensible aux outliers isolés** lors de l’ajustement, car il utilise par défaut une distribution d’erreur à queues épaisses (Student) qui leur assigne moins de poids facebook.github.io. Ainsi, un choc aberrant ne va pas fausser drastiquement la tendance estimée. Cependant, il ne “signale” pas explicitement l’anomalie : souvent, Prophet va absorber un gros choc en l’interprétant comme un changement de tendance (break) pour mieux coller aux données facebook.github.io. Cela peut conduire à des intervalles de prévision très larges ou à anticiper à tort d’autres changements de tendance similaires. La documentation conseille d’ailleurs de **retirer les outliers** identifiés pour améliorer la qualité des prévisions facebook.github.io. Dans notre optique, on peut donc utiliser Prophet en deux passes : ajuster le modèle, repérer les points ayant de grands résidus ou provoquant des ruptures de tendance artificielles, et les traiter en outliers (les marquer NA puis refitter éventuellement). Prophet est pratique pour des séries avec fortes composantes calendaires (saisonnalité annuelle, effets de fin de trimestre...) qu’on souhaite éliminer. Il est moins adapté à une détection fine des anomalies locales (il donnera plutôt une vision lissée).

En somme, c'est un **outil de modélisation robuste** qui peut faciliter la détection d'anomalies en fournissant un benchmark de "comportement normal" de la série (avec tendance/saison) puis en pointant les écarts inhabituels.

Tableau 2 : Méthodes basées sur des modèles temporels

Modèle	Idée	Points forts	Points faibles
ARIMA	Prévoir la série et détecter résidus	Tient compte de l'auto-corrélation, outliers classés par type (AO/IO...)	Paramétrage lourd (ordre du modèle), suppositions (stationnarité) souvent violées mdpi.com
GARCH	Estimer volatilité conditionnelle	Distingue choc extrême vs volatilité élevée attendue, robuste avec loi épaisse	Complexe à estimer, sensible aux erreurs de spécification (mauvaise distribution, etc.)
Prophet	Modèle additif robuste (tendance+saison)	Robustesse aux outliers isolés, capte saisons/trend, facile à utiliser	Ne signale pas explicitement l'anomalie (elle peut être absorbée) facebook.github.io , suppose patterns répétitifs stables

Remarque : D'autres approches de séries temporelles pourraient être mentionnées, comme les modèles à changement de régimes (Markov switching) qui détectent des régimes aberrants, ou les méthodes de filtrage de Kalman robustes. Cependant, ARIMA/GARCH/Prophet couvrent les principaux paradigmes (prévision linéaire, volatilité, modèle additif) utilisés couramment en pratique pour identifier des anomalies *a posteriori*.

Approches multivariées sur les vecteurs de tenors

Les méthodes multivariées considèrent **simultanément l'ensemble des piliers** d'un scénario (par exemple le vecteur de chocs [6M, 1Y, 10Y,...] à une date donnée). L'objectif est de détecter des scénarios dont la **forme de courbe de choc** est anormale, ou des combinaisons de mouvements entre tenors qui ne se produisent pas d'ordinaire. Contrairement aux approches précédentes, on ne traite plus chaque série isolément mais on regarde le vecteur d -dimensionnel (d = nombre de piliers) dans son ensemble. Ceci est crucial pour repérer des anomalies **structurelles** : par exemple, si tous les piliers $\leq 1Y$ montent fortement alors que ceux $\geq 5Y$ baissent fortement – ce pattern inhabituel pourrait être un outlier multivarié même si chaque composante prise individuellement n'est pas extrême.

Les données de marché ont souvent une **structure corrélée** (ex : les chocs sur la courbe des taux suivent des déformations en pente, parallèles, etc.). Les méthodes multivariées exploitent cette structure pour évaluer la plausibilité d'un vecteur entier. Elles incluent des techniques de réduction de dimension, des méthodes géométriques (distances) ou des algorithmes d'apprentissage non supervisé dédiés aux outliers.

- **Analyse en composantes principales (PCA)** – La PCA est souvent utilisée en finance pour représenter les variations d'une courbe de taux via quelques facteurs (ex: niveau, pente, courbure). Dans le contexte d'anomalie, on peut projeter chaque vecteur de chocs sur les premiers axes principaux et reconstruire une **valeur approchée** du vecteur. Si l'erreur de reconstruction est grande pour un scénario, c'est qu'il comporte une composante inhabituelle non expliquée par les facteurs principaux habituels – potentiellement une anomalie. En pratique, on pourrait calibrer une PCA sur l'historique des chocs de marché "normaux" et calculer la distance de chaque vecteur à l'espace factoriel. Cela revient à utiliser la **distance de Mahalanobis** : la PCA fournit la matrice de covariance des données, et la distance de Mahalanobis mesure à combien d'écarts-types multivariés se situe un point de cette distribution. Un scénario dont la distance de Mahalanobis est au-delà d'un certain quantile (par ex 99.97ème percentile correspondant à $\sim 3\sigma$ en multidimension) peut être flaggé. L'avantage de PCA est sa **simplicité et interprétabilité** : on identifie quelles composantes sont "anormales". Par exemple, on peut constater qu'un vecteur outlier a une contribution beaucoup trop forte sur un 4e composant habituellement négligeable. De plus, c'est peu coûteux si d (nb de piliers) est modeste. Dans notre dataset, d pourrait être $\sim 10-30$ (selon le nombre de maturités), ce qui est tout à fait gérable. Une limite est que la PCA est une méthode linéaire qui va mal capturer des relations non-linéaires entre tenors. De plus, elle requiert que les données d'apprentissage de la PCA soient représentatives : s'il y a déjà des anomalies dans l'échantillon, elles peuvent influencer les axes. On peut contourner cela en utilisant une PCA **robuste** (insensible aux outliers) ou en nettoyant itérativement. Malgré ces limites, PCA a été employée avec succès pour détecter des anomalies sur des données de marché complexes [mdpi.com](https://www.mdpi.com), notamment couplée avec des techniques d'imputation pour corriger les valeurs aberrantes et améliorer ensuite les mesures de risque.
- **Isolation Forest (IF)** – L'Isolation Forest est un algorithme d'ensemble (forêt aléatoire) conçu spécifiquement pour l'anomalie. Il fonctionne en isolant récursivement les observations à l'aide de partitions aléatoires : les anomalies, étant rares et différentes, tendent à être isolées en moins de coupures que les points normaux neptune.ai. En pratique, on construit de nombreux arbres de décision aléatoires; la profondeur moyenne à laquelle un point est isolé donne

un score d'anomalie. IF est **non paramétrique** (pas d'hypothèse distributionnelle) et gère bien les hautes dimensions en se basant sur des subdivisions aléatoires plutôt que sur le calcul de distances complexe. Il a montré son efficacité dans de nombreux domaines, y compris la finance pour détecter des transactions frauduleuses ou des comportements de marché inhabituels. Un atout clé est qu'il *n'efforce pas de modéliser les données normales*, il se concentre juste à isoler les points rares neptune.ai. Cela évite d'avoir à estimer une densité multi-dimensionnelle difficile. De plus, il peut naturellement gérer des jeux massifs en sous-échantillonnant les données pour chaque arbre. Pour 4M lignes, on peut par exemple construire chaque arbre sur un échantillon de 50k points, ce qui rend le calcul scalable. Parmi les limites : l'Isolation Forest nécessite de **fixer le taux d'anomalie** attendu (contamination) pour définir le seuil de score neptune.ai. Si ce paramètre est mal estimé, on peut avoir trop ou trop peu d'outliers signalés. Il faut souvent le régler par essais ou validation mdpi.com. Par ailleurs, IF ne donne pas directement une explication de pourquoi un point est isolé – c'est un modèle boîte noire comparé à PCA par exemple. Malgré cela, son **efficacité empirique** en fait un des algorithmes privilégiés pour la détection d'anomalies multivariées dans les jeux de données larges et bruités.

- **One-Class SVM (OC-SVM)** – Le SVM à une classe est une adaptation des SVM qui cherche à englober la majorité des données (classe “normale”) dans une frontière en espace des features, en excluant les outliers. Techniquement, il cherche un hyperplan dans un espace de haute dimension (défini par un kernel) qui sépare les données du **centre** (l'origine) avec une marge maximale, tolérant une certaine fraction d'erreurs (ν). Tout point situé *en dehors* de la frontière obtenue est considéré comme anomalie. L'OC-SVM peut être vu comme une estimation du **support** de la distribution des données. Ses points forts : il peut modéliser des frontières très complexes via le kernel (donc capturer des formes non linéaires de “normalité”) et il est relativement **robuste au bruit** dans les données d'entraînement via le paramètre ν (fraction d'outliers tolérée). Il a été utilisé en finance, par ex. pour détecter des configurations de marché inédites en apprentissage sur des périodes calmes. Cependant, l'OC-SVM souffre de la **malédiction de la dimension** si d est grand : il dépend de calculs sur des noyaux entre points, ce qui devient coûteux pour des millions de lignes (complexité souvent $> O(n \cdot d)$). Sur 4M points, un OC-SVM classique est pratiquement inapplicable sans réduction de dimension préalable ou sous-échantillonnage massif. De plus, ses résultats peuvent être sensibles au choix du kernel (gaussien RBF est le plus courant) et des hyperparamètres (ν , γ). Il faut souvent effectuer une validation croisée soignée. L'OC-SVM n'offre pas d'explication directe non plus, même s'il est possible de regarder quel vecteur de support

“définit” la frontière. En résumé, c’est un outil théoriquement élégant pour la *détection de nouveauté*, mais dont l’usage sur un très grand dataset multi-dimensionnel doit être limité soit à un sous-ensemble, soit à la sortie d’une réduction de dimension comme PCA pour en réduire le fardeau de calcul.

- **Autoencodeurs (Autoencoders)** – Les *autoencodeurs* sont des réseaux de neurones (généralement profonds) entraînés en non-supervisé à reconstruire leurs entrées en sortie. En les entraînant sur des données normales, l’autoencodeur apprend une représentation compressée (codage) des vecteurs de chocs et arrive à les reconstruire avec une faible erreur pour les configurations courantes. En revanche, un vecteur de chocs atypique ne pourra pas être correctement reconstruit – l’**erreur de reconstruction** servira alors de score d’anomalie medium.com. Cette approche non-linéaire est une généralisation de la PCA : là où la PCA ne peut capturer qu’une structure linéaire, un autoencodeur avec plusieurs couches et activations non-linéaires peut modéliser des relations complexes entre piliers analyticsvidhya.com. Par exemple, si l’effet d’un choc 6M sur le 10Y est non linéaire, un AE pourrait théoriquement l’apprendre. Ses avantages : il peut être très performant pour détecter des anomalies subtiles car il utilise la **puissance d’apprentissage du deep learning** ; il peut aussi intégrer davantage de features en entrée (par ex. inclure `rf_struct_id` encodé, ou d’autres variables) pour contextualiser la reconstruction. De plus, une fois entraîné, il est rapide à appliquer (une simple passe avant/arrière dans le réseau) même sur des millions de points. Néanmoins, ses limites sont à considérer : (a) il exige un effort d’entraînement non négligeable (il faut suffisamment de données “normales”, ce qui semble le cas ici, mais aussi calibrer l’architecture, le taux d’apprentissage, etc.) ; (b) comme tout modèle de réseau de neurones, il peut sur-apprendre et manquer de généralisation, surtout si certaines anomalies sont présentes dans les données d’entraînement – on peut introduire involontairement un biais ; (c) le résultat (erreurs de reconstruction) donne un score mais pas directement une cause, bien qu’en examinant les composantes qui contribuent le plus à l’erreur on puisse souvent cibler le pilier problématique. En pratique, les autoencodeurs ont été appliqués dans la *détection de fraudes* ou de défaillances industrielles avec succès, et en finance ils commencent à être utilisés pour surveiller des portefeuilles ou des séries économiques. Ils excellent quand la relation entre variables est trop complexe pour des méthodes linéaires. Enfin, notons qu’il existe des variantes avancées comme les *Variational Autoencoders (VAE)* ou *Sparse Autoencoders*, mais le principe reste le même pour la détection : utiliser l’erreur de reconstruction comme indicateur d’anomalie medium.com.

Tableau 3 : Méthodes d’anomalie multivariée (vecteurs de tenors)

Méthode	Atout principal	Limites	Contextes d'usage
PCA (Mahalanobis)	Détecte les formes de courbe inhabituelles via facteurs principaux, explicable	Linéaire seulement, sensible si données d'entraînement contiennent déjà anomalies	Courbes de taux, données corrélées où 2-3 facteurs expliquent l'essentiel
Isolation Forest	Non paramétrique, gère haute dimension, isole vite les outliers rares neptune.ai	Hyperparamètre contamination à estimer neptune.ai , "boîte noire" (peu explicatif)	Grands jeux de données bruités, anomalies globales ou locales, finance de marché (comportements atypiques)
One-Class SVM	Frontière flexible (kernel) englobant les données normales, approche statistique solide	Échelle mal au-delà de quelques centaines de milliers de points, tuning délicat (kernel, ν)	Détection de nouveautés sur variables réduites, validation d'anomalies sur un sous-échantillon représentatif
Autoencodeur (AE)	Capture des patterns non-linéaires complexes, améliore PCA, s'adapte en deep learning analyticsvidhya.com	Entraînement coûteux, possible sur-apprentissage, interprétation moins directe	Structures de taux complexes ou données multi-factorielles non linéaires, grands volumes où un modèle neuronal peut être formé

Modèles avancés (deep learning, graphes, EVT, GAN, etc.)

Les méthodes avancées englobent des approches plus récentes ou sophistiquées, qui vont au-delà des techniques "standards" listées précédemment. Elles peuvent combiner plusieurs dimensions (spatiales, temporelles), utiliser des réseaux de neurones profonds, ou reposer sur des fondements théoriques spécifiques (graphes, théorie des extrêmes, modèles génératifs). Dans le cadre de la détection d'anomalies sur des données financières, ces modèles avancés sont souvent utilisés pour *repérer des signaux faibles* ou des anomalies d'un type particulier, ou pour améliorer la **robustesse** de la détection. Voici quelques-unes des approches avancées pertinentes :

- **Réseaux de neurones profonds (LSTM, Transformers)** – Outre les autoencodeurs déjà évoqués, d'autres architectures de deep learning sont

utilisées pour les anomalies sur séries financières. Par exemple, les **LSTM** (Long Short-Term Memory networks) et autres RNN sont capables de modéliser des séquences temporelles en capturant les dépendances à long terme. Un LSTM entraîné à prédire la valeur suivante d'une série de chocs (ou à reconstruire la séquence glissante) peut signaler une anomalie lorsque l'erreur de prédiction est élevée sur un point. Ce type d'approche a été utilisé pour la détection d'anomalies dans des séries multivariées complexes, souvent en combinaison avec d'autres techniques (ex : un LSTM-VAE medium.com qui combine LSTM et autoencodeur variationnel). De même, les **Transformers** avec mécanisme d'attention commencent à s'imposer en série temporelle : des modèles comme **Anomaly Transformer** ou **TranAD** intègrent l'attention pour repérer des motifs inhabituels en considérant l'ensemble de la séquence et ses corrélations internes medium.com medium.com. L'attention peut par exemple révéler qu'à un instant donné, la relation entre les tenors est très différente de tout ce qu'elle a été historiquement, signalant une anomalie. Ces modèles profonds offrent souvent les meilleures performances en détection d'anomalies *complexes* d'après des benchmarks récents mdpi.com mdpi.com, surtout sur des données riches et de grande dimension. En revanche, ils présentent des inconvénients pratiques : (a) ils nécessitent énormément de données pour être entraînés correctement (cependant 4M de lignes peut suffire, surtout si on augmente en données synthétiques ou augmentation de données medium.com) ; (b) ils sont difficiles à interpréter (bien qu'il existe des techniques d'explicabilité, cela reste du deep learning) ; (c) ils demandent une infrastructure de calcul (GPU) et du temps pour l'entraînement et le tuning. Pour notre cas, on pourrait imaginer un LSTM multivarié qui ingère simultanément tous les piliers d'une structure à la fois, ou un Transformer qui modélise la *série temporelle de vecteurs de chocs*. Ces modèles pourraient détecter des anomalies du type *"ce scénario de stress n'était jamais arrivé auparavant sous cette forme"*. Un écueil : si les anomalies sont très rares, un réseau profond pourrait avoir du mal à les apprendre (on entraîne sur essentiellement du normal). Parfois, on entraîne sur du normal uniquement (comme autoencoder) et on détecte par écart, ou bien on utilise des techniques de **data augmentation** pour que le réseau voie des quasi-anomalies et apprenne à les reconnaître. Globalement, les deep nets sont **prometteurs** mais plus complexes à mettre en œuvre, et souvent justifiés pour des données à très haute dimension ou comportant de multiples modalités (ce qui peut être le cas ici si l'on intègre d'autres facteurs, ex. données de marché liées).

- **Méthodes basées sur les graphes** – Les données financières peuvent être représentées sous forme de **graphes** pour capter des relations structurelles. Par exemple, on peut représenter chaque maturité comme un nœud relié à ses voisins (graph linéaire de la courbe) ou relier différents piliers selon la corrélation

empirique. Les **Graph Neural Networks (GNN)** peuvent alors être appliqués pour détecter des anomalies prenant en compte la topologie du graphe. Une approche appelée **GDN (Graph Deviation Network)** propose de modéliser un ensemble de séries multivariées en construisant un graphe de dépendances puis en détectant des écarts via un réseau de neurones graphes [mdpi.com](https://www.mdpi.com). Ce modèle (Deng & Hooi 2021) a montré de très bons résultats sur des anomalies multivariées complexes en combinant l'apprentissage des relations entre variables avec la détection d'outliers sur les *signaux propagés dans le graphe*. Dans notre cas, une idée serait de construire un graphe où chaque *rf_struct_id* (structure de taux, peut-être par devise) est un nœud, relié aux autres structures similaires, et où les chocs extrêmes se propagent. Ou bien un graphe où les nœuds sont les piliers (6M, 1Y, etc.) avec des arêtes pondérées par la corrélation de leurs chocs – ainsi un choc qui rompe le schéma de corrélation habituel (par ex. un 30Y qui bouge sans aucun mouvement des autres tenors) pourrait être détecté par une GNN. Par ailleurs, *l'analyse de graphes* non-apprentissage peut aussi servir : par ex., calculer la *MST (arbre couvrant minimal)* des données et chercher les points reliés par des arêtes anormalement longues, ce qui révèle des outliers isolés. Il existe aussi des techniques d'anomalie pour graphes temporels (si on considère l'évolution d'un réseau). Toutefois, dans le cadre strict de notre dataset (courbes de taux/scénarios indépendants), l'approche GNN est encore expérimentale. Elle peut améliorer la détection en capturant explicitement des relations entre variables que d'autres méthodes ignorent. Son inconvénient principal est la **complexité** : il faut définir le graphe, entraîner un modèle spécifique (les GNN sont encore plus complexes à calibrer que les réseaux standard), et disposer d'assez de données pour apprendre les paramètres du graphe. C'est justifié surtout si l'on a de **nombreuses variables interdépendantes** (par ex. 50+ variables où la structure même d'interaction est informative). Ici, avec ~10–20 piliers, c'est envisageable mais peut-être pas nécessaire si PCA ou AE capturent déjà bien les corrélations.

- **Théorie des valeurs extrêmes (EVT)** – La théorie des valeurs extrêmes fournit un cadre statistique pour modéliser la distribution de la **queue** d'une distribution, au-delà d'un seuil élevé. Plutôt que de fixer un seuil de façon heuristique (ex: 3σ), on utilise l'EVT pour estimer à partir des données elles-mêmes quel seuil correspond à un certain niveau de rareté. L'approche courante est le **Peaks-Over-Threshold (POT)** : on choisit un quantile élevé (par ex 95e ou 99e), on prend toutes les valeurs au-dessus de ce seuil initial, puis on ajuste une loi de Pareto généralisée (GPD) sur ces dépassements [medium.com](https://www.medium.com). À partir de là, on peut extrapoler la probabilité (p-value) d'observer un point encore plus extrême. On étiquette comme anomalies ceux dont la p-value estimée est en dessous d'un petit seuil (par ex $< 0,001$). L'énorme avantage est que le seuil final n'est **pas**

arbitraire mais fondé sur une théorie asymptotique solide : on contrôle le taux de faux positifs selon un niveau de confiance choisi. De plus, l'EVT est **particulièrement adaptée aux données financières**, connues pour leurs lois à queues lourdes. Plutôt que de supposer une normalité et crier à l'outlier à 3σ , on reconnaît que les extrêmes suivent une loi différente et on les modélise. Par exemple, on peut trouver qu'au-delà de 4%, les chocs journaliers suivent une Pareto avec tel paramètre ξ ; on peut alors calculer que tel choc de +8% a une probabilité théorique de 0,0005 – assez pour le considérer comme *extrême anormal*. L'EVT peut être appliquée univariée (par pilier) ou même multivariée (il existe des EVT multivariées, plus complexes, ou on peut appliquer POT sur le score de Mahalanobis). Son inconvénient est la nécessité d'un échantillon suffisamment grand d'extrêmes pour bien ajuster la loi. Sur 4M de lignes, c'est probable qu'on en ait assez pour certains piliers. Un autre défi est de bien choisir le seuil initial : trop haut, on manque de données pour ajuster; trop bas, on viole l'hypothèse EVT. Des méthodes comme **SPOT (Streaming POT)** ajustent dynamiquement le seuil et peuvent être utiles en pratique pour la détection en fluxmedium.com. En résumé, l'EVT offre une approche **statistique rigoureuse** pour fixer des seuils d'anomalie extrême, et complète bien d'autres méthodes : par exemple, on peut coupler EVT avec Isolation Forest en appliquant POT sur les scores d'anomalie produits par la forêt afin de déterminer un cutoff optimisé medium.com. Dans un contexte de stress tests, EVT peut aider à identifier quelles *scénarios de chocs* sont "trop extrêmes" par rapport à l'historique des marchés (par ex., un choc de -40% sur un indice boursier pourrait être flaggé si historiquement même en 2008 on n'a pas eu pire que -30%).

- **Réseaux antagoniques génératifs (GANs)** – Les GANs, célèbres pour générer de fausses images, sont également utilisés pour la détection d'anomalies. L'idée est d'entraîner un **réseau générateur** $G(z)$ à produire des données semblables aux données normales, face à un **discriminateur** $D(x)$ qui apprend à distinguer vraies données de celles générées. Une fois entraîné sur des données "normales", on peut évaluer un nouvel exemple de deux manières : soit voir si le discriminateur le classifie comme *faux* (ce qui signifierait qu'il ne ressemble pas aux données d'entraînement), soit essayer de *reconstruire* cet exemple via le générateur en trouvant le z optimal et mesurer l'erreur (approche dite **AnoGAN**). Plusieurs variantes existent, notamment pour les séries temporelles : **MAD-GAN** (Multivariate Anomaly Detection GAN) applique un GAN sur des fenêtres de séries multivariées mdpi.com, **TAnoGAN** utilise l'attention avec GAN, **TadGAN** (Time series anomaly detection GAN) intègre à la fois un autoencodeur et un critic façon GAN. Ces méthodes combinent souvent l'idée de reconstruction et d'estimation de densité implicite. Le **point fort** des GANs est de pouvoir modéliser des distributions très complexes sans les supposer

explicitement, et d'en capturer les dépendances non linéaires. Par exemple, un GAN entraîné sur des vecteurs de chocs de taux pourrait apprendre le *manifold* des chocs réalistes; un vecteur en dehors de ce manifold sera repéré soit par un mauvais score du discriminateur, soit par une mauvaise reconstruction par le générateur [medium.com](https://medium.com/medium.com). De plus, les GANs peuvent générer de nouvelles données ressemblantes – ce qui peut servir à enrichir l'échantillon d'apprentissage ou à simuler des scénarios alternatifs. Les limites des GANs : (a) entraînement notoirement difficile (équilibre générateur/discriminateur à trouver, risque de mode collapse, etc.) ; (b) nécessité d'un volume de données important et d'un bon réglage – dans notre cas 4M lignes ça peut convenir, mais selon la dimension du vecteur; (c) moins d'interprétabilité encore que les autoencodeurs, on obtient un score ou une évaluation du discriminateur, qu'il faut traduire en décision d'outlier. Malgré cela, les études récentes montrent des **résultats prometteurs** : par exemple, un GAN entraîné sur des séries de capteurs a surpassé des méthodes classiques pour identifier des séquences aberrantes mdpi.com. Dans le domaine financier, on commence à voir l'application de GANs pour générer des scénarios extrêmes plausibles, et inversement pour détecter que tel scénario réel est improbable sous la distribution apprise (d'où alerte). TadGAN, développé par des chercheurs de NVIDIA et MIT, a été testé sur des données de séries économiques et a bien détecté des chutes anormales en combinant l'erreur de reconstruction et l'erreur adversariale. En pratique, l'usage des GANs pour l'anomalie reste complexe et requiert de l'expertise en deep learning, mais c'est une voie *avancée* pour maximiser l'utilisation de l'information contenue dans de larges datasets et appréhender des relations non linéaires subtiles.

Tableau 4 : Méthodes avancées et hybrides

Approche avancée	Description / Rôle	Intérêts	Précautions
Deep Learning séquentiel (LSTM, Transformer)	Réseaux apprenant la dynamique temporelle ou les dépendances complexes (attention)	Capture anomalies contextuelles complexes, état de l'art en performances mdpi.com	Exige gros volumes, tuning long, interprétabilité faible
Modèles à graphes (GNN, structures de corrélation)	Exploitent les relations structurelles entre variables via un graphe (ex: GDN)	Intègre explicitement les dépendances entre piliers, repère ruptures de corrélation	Complexité de modélisation, bénéfice réel si réseau de variables large mdpi.com

Approche avancée	Description / Rôle	Intérêts	Précautions
Théorie des extrêmes (EVT, POT)	Modélisation statistique des queues de distribution au-delà d'un seuil haut	Seuils d'anomalie fondés sur p-value cible, adapté aux queues épaisses (finance) medium.com	Suppose assez de données extrêmes, choix du seuil initial délicat
GAN génératifs	Apprentissage d'une distribution générative des données normales, évalue la probabilité qu'un point en provienne	Modélise des distributions complexes sans supposition, combine détection & génération (scénarios)	Entraînement instable, besoin de beaucoup de données et d'expertise deep
Hybride / Ensemble	Combinaison de plusieurs méthodes ou étapes	Cumul des forces des techniques, plus de robustesse	Complexité accrue, assemblage à valider (dépendances)

- Méthodes hybrides et ensemblistes** – En plus des familles isolées, on obtient souvent de meilleurs résultats en combinant les approches. Un exemple de méthode hybride issu de la recherche est celui de *Crépey et al. (2022)* qui combinent PCA et réseaux de neurones : ils extraient d'abord des caractéristiques via une PCA puis alimentent un perceptron pour estimer un score d'anomalie, optimisant simultanément un seuil de décision au sein de la fonction lossmdpi.commdpi.com. Cette approche PCA-NN a surpassé plusieurs algorithmes classiques (Isolation Forest, LOF, SVM, etc.) sur des données de marché, car elle allie la **réduction de dimension adaptée** aux données financières et la **flexibilité d'un réseau** pour calibrer le scoremdpi.commdpi.com. D'autres hybrides existent : par exemple, l'article de *Akyildirim et al.* propose *sig-IF*, qui combine l'extraction de features de trajectoires (signatures) avec une Isolation Forest pour détecter des anomalies de cheminmdpi.com. On peut également citer une approche *ANN-GARCH* (réseau de neurones + GARCH) pour détecter et corriger les outliers dans des séries financières volatiles papers.ssrn.com.

Les **ensembles de modèles** constituent une autre catégorie : au lieu de se fier à un seul algorithme, on agrège les détections de plusieurs. Par exemple, exécuter *Isolation Forest*, *One-Class SVM* et *Local Outlier Factor (LOF)* en parallèle sur le dataset, puis considérer comme anomalie tout point signalé par au moins deux de ces méthodes. L'idée est que les algorithmes ayant des biais différents, un point repéré par plusieurs a

plus de chances d'être un vrai outlier. Des techniques plus élaborées pondèrent les scores de différents détecteurs ou entraînent un méta-modèle sur ces scores. Cela améliore souvent la **fiabilité** et réduit les faux positifs dans les cas complexes milvus.io. Par exemple, un Isolation Forest peut rater un type d'anomalie très local que LOF (méthode par densité locale) attrapera, et vice versa. L'ensemble combinerait les deux. L'inconvénient est bien sûr le coût accru (il faut faire tourner N modèles au lieu d'un) et la difficulté à gérer des résultats divergents. On peut utiliser des approches de vote, de moyenne de scores, ou encore entraîner un modèle supervisé sur des données étiquetées artificiellement (en injectant de fausses anomalies connues) pour obtenir la meilleure combinaison.

En pratique, pour notre jeu de données volumineux, une démarche hybride raisonnable serait par exemple : **(i)** appliquer des filtres univariés robustes (MAD/EVT) pour éliminer les outliers flagrants par pilier (**nettoyage initial**), **(ii)** sur les données restantes, utiliser une méthode multivariée (PCA ou autoencodeur) pour repérer des scénarios suspects globalement, **(iii)** affiner localement avec une Isolation Forest ou un LOF sur les données standardisées afin d'attraper des cas particuliers, **(iv)** combiner les résultats : par exemple ne flagger anomalie que si au moins deux des trois méthodes (univarié, PCA, IF) sont d'accord. Une telle approche multi-échelle permet de tirer parti des atouts de chaque méthode tout en compensant leurs faiblesses respectives. Naturellement, la calibration doit être faite avec soin (par ex. éviter d'éliminer trop de points à l'étape univariée pour encore pouvoir calibrer correctement la PCA sur l'historique "nettoyé").

Comparaison générale et recommandations d'application

Chaque famille de méthodes présente donc des **avantages distincts** et des limites qu'il convient de connaître pour une utilisation judicieuse sur un dataset de chocs de marché :

- *Méthodes univariées* : Simples à implémenter et très rapides, elles conviennent en **premier filtre** pour attraper les erreurs grossières (par ex. une valeur aberrante due à un problème de saisie). Le **z-score** est facile à comprendre mais doit être utilisé sur des données déjà assez nettoyées ou transformées (log-return, etc.) pour ne pas être trompé par la non-normalité. Les méthodes **robustes** comme MAD ou IQR sont préférables dans un contexte financier où les distributions de chocs sont souvent asymétriques et leptokurtiques numberanalytics.com. Elles peuvent être appliquées par pilier indépendamment. Toutefois, elles ignorent toute relation inter-piliers : ainsi, elles manqueront des anomalies structurelles (courbe en forme étrange mais sans point individuellement extrême). On recommande de *toujours démarrer* par un profilage univarié (par ex. écarter tout point au-delà de $5 \times \text{MAD}$) afin d'assainir les données, ce qui facilitera ensuite l'analyse multivariée.

- Méthodes temporelles* : Elles exploitent la dimension historique. Si le dataset s'étend sur de nombreuses dates, il est pertinent d'analyser chaque série de chocs dans le temps. Un modèle **ARIMA** peut identifier un **saut additive** ou une **rupture de tendance** (outlier innovateur) qu'une simple méthode statique ne verrait pas. De même, un modèle **GARCH** aide à ne pas sur-signaliser des mouvements durant des marchés déjà volatils. Cependant, ces méthodes nécessitent un certain savoir-faire (choix des ordres, tests de stationnarité, etc.) et ne sont pas triviales à passer à l'échelle de dizaines de milliers de séries. Il peut être préférable de les réserver à une **analyse ciblée** : par exemple, après un filtre initial, appliquer ARIMA/GARCH sur les séries des piliers les plus sensibles (ex : taux courts) pour valider les outliers détectés et en détecter d'autres plus contextuels. **Prophet** ou des modèles de décomposition peuvent être utiles si des patterns saisonniers existent (ex : effet tournant de trimestre sur les spreads de crédit). Globalement, les approches temporelles sont surtout utiles si l'on a une **composante temporelle forte** dans les données de chocs (ce qui dépend du dataset – s'il s'agit de scénarios indépendants de stress, la notion de temps peut être moins pertinente que si l'on a des chocs journaliers). Pour un dataset de stress tests (scénarios calibrés à différentes dates), un ARIMA peut repérer qu'à telle date de calibration un outlier s'est glissé. En résumé, utiliser ces méthodes en complément pour contextualiser les anomalies dans le temps, et non comme seul outil sur 4M de lignes (sauf automatisation poussée).
- Méthodes multivariées* : Indispensables pour notre problème, car elles adressent la question « *Ce vecteur de chocs est-il cohérent dans son ensemble ?* ». Parmi celles-ci, **PCA/Mahalanobis** est un bon point de départ car elle fournit une mesure unique de l'écart d'un scénario par rapport aux précédents (distance dans l'espace des chocs) et indique quelles combinaisons de piliers posent problème. On la recommande pour identifier des anomalies de forme de courbe de taux ou de structure de chocs inter-marchés. De plus, la PCA est rapide même sur de gros volumes (possible de la mettre à jour incrémentalement sur 4M lignes). **Isolation Forest** est également très adapté aux grands datasets et souvent performant sans trop de réglages ; on le conseille pour balayer l'ensemble du jeu de données afin de détecter des points singuliers qu'aucun modèle paramétrique n'aurait anticipés. Il faudra toutefois bien calibrer le pourcentage d'anomalies attendu (on peut commencer par 0.5% par exemple et ajuster). **One-Class SVM** est potentiellement moins scalable – à utiliser si l'échantillon est réduit ou en second niveau (par ex. entraîner un OC-SVM sur les points marqués suspects par d'autres méthodes, afin de peaufiner la frontière entre normal/anormal). **Autoencodeurs** et réseaux neuronaux peuvent être envisagés si les relations entre piliers sont complexes ou si l'on souhaite intégrer d'autres features (par ex. chocs d'autres variables macro, etc.). Un autoencodeur

simple (quelques couches fully-connected) pourrait être entraîné sur une large portion des données (en excluant les extrêmes au préalable) et servir à scorer les scénarios par erreur de reconstruction. Cela peut compléter utilement PCA en captant d'éventuelles non-linéarités. En pratique, on pourrait combiner **PCA et autoencodeur** : utiliser PCA pour une première détection rapide et explicable, et l'autoencodeur pour valider ou trouver des anomalies supplémentaires plus subtiles.

- *Modèles avancés* : Les techniques de **deep learning** (LSTM, Transformer) apportent surtout de la valeur si l'on a de la donnée temporelle dense ou de très haute dimension. Par exemple, si notre dataset couvrirait des milliers de variables financières, un modèle attentionné pourrait repérer des liens anormaux. Dans le cadre de chocs de courbe, leur bénéfice marginal par rapport à des techniques plus simples peut être faible, à moins de chercher des *anomalies séquentielles* (comme une séquence inhabituelle de chocs successifs). On peut donc réserver ces modèles aux cas où les méthodes plus simples échouent à expliquer certaines anomalies observées, ou dans une démarche exploratoire (recherche de patterns complexes). Les méthodes à **graphes** pourraient être mises en œuvre si on suspecte qu'une représentation réseau (par ex. graphe de similarité entre scénarios, ou graphe de dépendance entre piliers) révèle mieux les outliers. Cela reste un domaine pointu ; une recommandation pragmatique serait d'utiliser un algorithme comme **LOF (Local Outlier Factor)**, qui est conceptuellement lié aux graphes de voisinage, pour détecter des anomalies locales en complément d'Isolation Forest. **EVT** est fortement recommandé pour *calibrer des seuils* : par exemple, utiliser POT sur la distribution des distances de Mahalanobis ou des scores d'autoencodeur afin de choisir un cutoff correspondant à une probabilité d'occurrence cible (par ex 0,1% de faux positifs). Cela donne un vernis statistique rassurant pour fixer une barre au-delà de laquelle on alerte sur un choc. Enfin, les **GANs** et modèles génératifs sont encore émergents dans ce contexte ; ils pourraient être explorés si l'on dispose des moyens et données suffisants, notamment pour générer artificiellement des scénarios extrêmes et voir si un scénario réel s'éloigne de ce que le GAN peut produire. Toutefois, compte tenu de la complexité, il peut être plus efficace dans un premier temps d'appliquer les autres techniques, et d'utiliser un GAN en *recherche et développement* pour valider la robustesse des conclusions.

Recommandations pratiques : Pour appliquer ces méthodes sur un dataset de >4 millions de lignes, il faut garder à l'esprit la **scalabilité** et la **complémentarité** des approches. Une stratégie efficace serait :

1. **Nettoyage univarié initial** – Filtrer les valeurs aberrantes univariées évidentes par pilier à l'aide de règles robustes (par ex : éliminer ou marquer tout choc au-

delà de $4 \times \text{MAD}$ sur chaque pilier). Ce pré-nettoyage améliore la qualité des étapes suivantes et peut éliminer quelques centaines/ milliers de points problématiques dès le départ à coût quasi nul.

2. **Détection multivariée globale** – Appliquer une méthode couvrant l'ensemble du vecteur de chocs pour chaque scénario. Idéalement, combiner une méthode linéaire explicable (PCA avec seuil sur score ou Mahalanobis) **et** une méthode non-linéaire (Isolation Forest ou autoencodeur). Comparer les résultats : si les deux méthodes concordent pour signaler certains points, ceux-ci sont de très bons candidats anomalies. S'ils divergent, investiguer pourquoi (par ex. IF détecte un point, mais PCA non – ce point était peut-être extrême dans une direction non couverte par les 2-3 premiers composants, d'où l'intérêt de IF).
3. **Affinage local et temporel** – Pour les points détectés, examiner les séries temporelles correspondantes : par exemple, si un scénario du 12/03/2020 est marqué anormal, regarder la série de chocs du pilier 10Y autour de cette date. On peut utiliser un petit ARIMA/GARCH pour confirmer que ce jour-là le choc était en effet en rupture avec la dynamique récente. Ou utiliser un LOF centré sur les voisins temporels pour voir si c'est un *point isolé* dans le temps. Ce genre d'analyse sert de validation croisée et aide aussi à **documenter l'anomalie** (ex : “ce jour-là, le modèle AR(1) des chocs 10Y avait un résidu de 5σ , confirmant l'aspect aberrant”).
4. **Ensemble et décision** – Combiner les évidences : un point repéré par plusieurs méthodes différentes et validé comme aberrant contextuellement sera finalement labellisé comme anomalie à traiter (exclusion du dataset ou marquage pour traitement spécial). A contrario, un point détecté par une seule méthode avec un score limite pourrait être jugé *tolérable* (peut-être s'agit-il d'un événement de marché rare mais légitime, comme un krach véridique). L'approche par ensemble peut être formalisée en attribuant un score global (moyenne pondérée des rangs sur chaque méthode, par exemple) ou en définissant des règles du type “doit être détecté par au moins 2 méthodes sur 3”. Cette redondance apporte de la **robustesse** et réduit le risque de faux positifs liés à un biais d'une méthode.
5. **Mise à l'échelle** – Pour gérer 4M+ lignes, penser à utiliser des techniques d'**échantillonnage** ou d'**incrémentation** : par exemple, Isolation Forest permet de traiter par batches (on peut construire plusieurs forêts sur des sous-échantillons de 500k lignes chacune, et agréger les résultats). La PCA peut être mise à jour en ligne (algorithme d'incrémentation) pour éviter de tout charger en mémoire. Les autoencodeurs peuvent être entraînés par mini-batch sur GPU. Il peut aussi être utile de **segmenter le dataset** par catégories homogènes (par ex. analyser séparément les chocs de courbe de taux en EUR et ceux en USD si ce

sont des distributions très différentes). Cela améliore la précision des modèles (une PCA par devise aura moins de variance non expliquée qu'une PCA mixant toutes devises).

En conclusion, il n'y a pas de méthode unique « meilleure » en toutes circonstances : la détection d'anomalies gagne à combiner plusieurs approches. Pour un dataset financier de chocs de marché, nous recommandons une approche **hybride** articulée autour d'une détection multivariée principale (PCA, Isolation Forest) renforcée par des analyses univariées robustes et des validations temporelles. Les méthodes avancées (deep learning, graphes, GANs) peuvent apporter un surplus de détection pour des anomalies particulièrement complexes ou rares, mais au prix d'une complexité de mise en œuvre – elles sont donc à réserver si les méthodes plus simples laissent des zones d'ombre. En pratique, commencer par les techniques **simples et explicables** permet déjà de capturer l'essentiel des anomalies (surtout les erreurs de données ou scénarios incohérents flagrants), puis affiner progressivement avec des méthodes plus sophistiquées pour traiter les cas limites. Cette démarche graduelle assurera une détection efficace et confiance dans les résultats, tout en maîtrisant le temps de calcul sur un grand volume de données. Ainsi outillés, les analystes risquent pourront identifier et traiter les chocs de marché aberrants, garantissant des calibrations de modèles et des analyses de stress plus fiables et pertinentes.