Sujet : Arbitrage Statistique Dynamique sur le Crack Spread

Equipe BlaCS-Scholes

Avertissement

Ce document n'est pas un énoncé au sens classique. L'intérêt de ce projet est d'acquérir de l'expérience, par la lecture de l'état de l'art et par votre autoformation aux méthodes présentées. Au regard de votre expérience et de vos résultats intermédiaires, vous pouvez et êtes encouragés à modifier ou adapter la stratégie, et à redéfinir les objectifs.

1 Introduction et Objectifs du projet

Type de projet : Quant Research & Quant Trading.
Ce projet de recherche s'accompagne d'une application concrète : l'objectif est de mettre le modèle en production et de simuler un P&L en bout de chaîne.

Ce projet définit une stratégie d'arbitrage statistique sur la marge de raffinage (*crack spread*). L'approche fondamentale classique consiste à trader un spread fixe 3 CL − 2 RB − 1 HO. Ce ratio 3:2:1 repose sur le rendement physique théorique d'une raffinerie standard : le craquage de 3 barils de brut (WTI) produit historiquement environ 2 barils d'essence (RBOB) et 1 baril de distillat (Heating Oil).
Cependant, ce ratio fixe n'est pas nécessairement cointégré. Les rendements réels fluctuent en raison des modernisations technologiques des infrastructures de raffinage, des ajustements saisonniers de production (mélanges été/hiver) et de la dynamique d'offre et de demande propre à chaque produit. Par conséquent, l'utilisation de poids statiques impose une relation rigide qui s'éloigne de la réalité du marché, exposant le spread à des dérives structurelles et produisant une série temporelle pas nécessairement stationnaire.

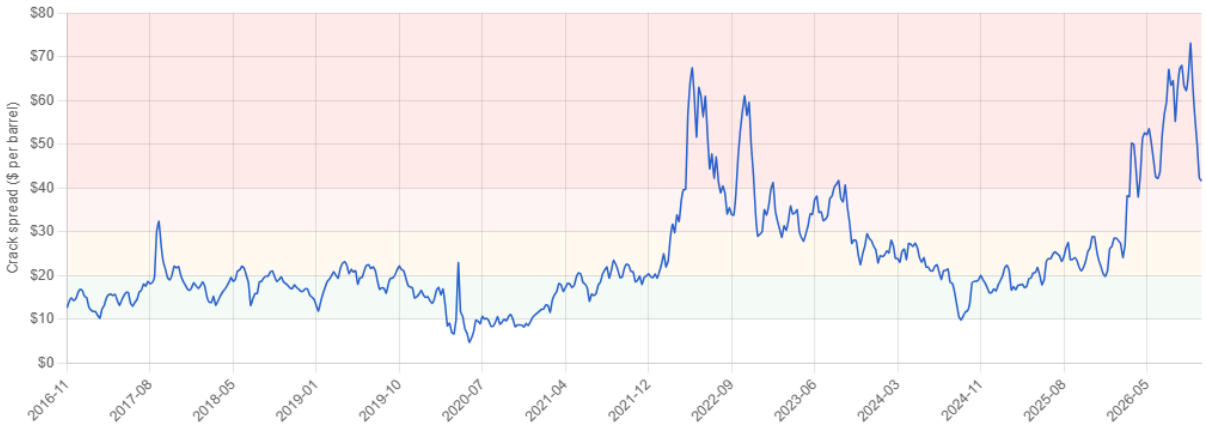


Figure 1: Crack spread 3:2:1 sur 10 ans

Au lieu de présupposer un modèle de production théorique (le spread 3:2:1), l'approche consiste à estimer dynamiquement les pondérations pricées par le marché. L'estimation du vecteur de cointégration vise à extraire la combinaison linéaire de ces trois actifs qui neutralise leurs tendances stochastiques marginales. Cette étape permet alors d'obtenir un portefeuille synthétique stationnaire, sur lequel on peut modéliser les résidus avec des modèles connus, afin de mesurer et quantifier les écarts, pour pouvoir prendre des positions en conséquence sur les marchés.

2 Démarche proposée

Étape 1 : Préparation des données et alignement
Les séries de prix du brut (CL), de l'essence (RB) et du fioul (HO) doivent être normalisées dans la même unité (dollars par baril). Les séries historiques doivent être construites sous forme de contrats continus pour traiter l'impact des sauts de prix lors des roulements des contrats futures. Les données peuvent être obtenues via yfinance (gratuit), london strategic edge, ou encore sur le site du NYMEX. Les éventuels coûts d'accès à des bases de données premium doivent être justifiés au préalable.

Étape 2 : Test de Johansen et Vecteur de Cointégration
La recherche de cointégration s'effectue via un modèle vectoriel à correction d'erreur (VECM) sur le vecteur des prix $Y_t = (P_t^{\text{CL}}, P_t^{\text{RB}}, P_t^{\text{HO}})^\top$:

$$\Delta Y_t = \Pi Y_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta Y_{t-j} + \varepsilon_t$$

Étapes de validation :

- Sélection du nombre de retards optimal p via les critères d'information (AIC ou BIC).
- Détermination du rang de la matrice $\Pi = \alpha\beta^\top$ par les tests statistiques de la Trace et de la Valeur Propre Maximale.

Si la cointégration est validée, le vecteur β fournit les poids de l'allocation optimale. Le portefeuille $Z_t = \beta^\top Y_t$ est la combinaison linéaire stationnaire des trois actifs.

Étape 3 : Modélisation du Résidu (Retour à la moyenne)
Une fois le spread Z_t identifié, sa dynamique est modélisée par un processus d'Ornstein-Uhlenbeck, caractérisant la force de rappel des raffineries :

$$dZ_t = \theta(\mu - Z_t)dt + \sigma dW_t$$

Où on a :

- θ : Vitesse de retour à la moyenne.
- μ : Moyenne de long terme du spread.
- σ : Volatilité du processus.

Points techniques : Calcul de la demi-vie (*half-life*) du processus pour calibrer la durée de détention anticipée des positions :

$$t_{1/2} = \frac{\ln(2)}{\theta}$$

Validation de la stationnarité de Z_t via un test de Dickey-Fuller Augmenté in-sample et out-of-sample.

Étape 4 : Signaux de Trading et Z-score
Le résidu est transformé en *z-score* pour normer les signaux d'entrée et de sortie :

$$z_t = \frac{Z_t - \mu}{\sigma}$$

La mécanique de trading s'articule autour de seuils (Δ) :

- **Vente (Short)** : $z_t > \Delta_{sup}$ (Spread surévalué).
- **Achat (Long)** : $z_t < -\Delta_{inf}$ (Spread sous-évalué).
- **Clôture** : Retour de z_t à 0 (ou proche de 0).

Étape 5 : Gestion du Risque (Filtre de Régimes Markoviens)
Des chocs physiques peuvent briser temporairement la force de rappel de la cointégration. Un filtre de régimes markoviens (HMM - Hidden Markov Model) à deux états cachés, $S_t \in \{1, 2\}$, est calibré spécifiquement sur la volatilité du résidu. Les transitions d'un état à l'autre sont régies par une matrice de probabilités $P = (p_{ij})$ où $p_{ij} = \mathbb{P}(S_t = j \mid S_{t-1} = i)$.
L'innovation du processus, définie par $\varepsilon_t = Z_t - \mathbb{E}[Z_t \mid \mathcal{F}_{t-1}]$, a une espérance nulle. Conditionnellement à l'état latent $S_t = k$, seule la variance change, et l'innovation suit la distribution :

$$\varepsilon_t \mid (S_t = k) \sim \mathcal{N}(0, \sigma_k^2)$$

Le modèle identifie deux régimes distincts par leur variance ($\sigma_1^2 \ll \sigma_2^2$) :

- **État 1 ($S_t = 1$, Variance faible)** : Régime de convergence classique. La probabilité filtrée d'être dans ce régime $\mathbb{P}(S_t = 1 \mid \mathcal{F}_t)$ est dominante. On peut faire tourner le modèle car il y a peu de perturbation dans la relation de cointégration.
- **État 2 ($S_t = 2$, Variance forte)** : Rupture de la relation de cointégration. L'augmentation brutale de la variance σ_2^2 active la fonction de coupe-circuit et suspend les signaux lorsque $\mathbb{P}(S_t = 2 \mid \mathcal{F}_t)$ franchit un seuil critique.

3 Jalons et Résultats Intermédiaires

Lors de la mise en œuvre du projet, la production de résultats intermédiaires est obligatoire afin de valider chaque maillon de la chaîne.

Jalon	Livrable intermédiaire	Échéance indicative
J1	Revue de littérature (synthèse d'articles) et plan de travail	Semaines 1-2
J2	Données collectées/nettoyées et statistiques descriptives	Semaines 3-4
J3	Baseline bout en bout (version simpliste) et premiers résultats	Semaines 5-6
J4	Version améliorée, analyse de robustesse et étude des échecs	Semaines 7-9
J5	Livrable final : code reproductible, rapport, présentation	Semaine 10

Table 1: Calendrier de suivi du projet

4 Pièges à Éviter

Afin de garantir la validité de la stratégie jusqu'à la simulation du P&L, portez une attention particulière aux biais suivants lors de vos backtests :

- **Look-ahead bias :** Veillez à n'utiliser aucune information non disponible à la date de décision (attention à l'alignement temporel des données et à l'estimation sur fenêtre glissante).

- **Survivorship bias et biais de sélection :** Prêtez attention aux biais induits par le choix de la période ou des contrats spécifiques.
- **Surapprentissage (p-hacking) :** Évitez la multiplication des essais non corrigés. Privilégiez une validation *walk-forward* et conservez un jeu de test indépendant.
- **Fuite de données (data leakage) :** Empêchez toute interférence entre les périodes d'apprentissage et de test, un risque majeur avec les séries temporelles autocorrélées.
- **Coûts et réalisme :** Intégrez de manière réaliste les coûts de transaction, la liquidité, le roulement des contrats (roll) et le glissement (slippage).
- **Non-stationnarité :** Gardez à l'esprit qu'un signal performant dans le passé peut disparaître; la stabilité temporelle doit être rigoureusement testée.