Exploring Machine Learning Optimizatios for Pair Selection in Pairs Trading

Ausarbeitung

Pairstrading ist eine nichtdirektionale Handelsstrategie, die unabhängig von der Marktrichtung funktioniert (Vidyamurthy 2004, p. 8). Sie basiert auf historischen Zusammenhängen zwischen zwei Wertpapieren (Zeitreihen) $X_i(t)$ und $X_j(t)$ und nutzt temporäre Abweichungen in deren Preisbeziehung zur Gewinnerzielung aus. Im verbreiteten, statistischen Pairstrading werden zwei Hauptansätze unterschieden: kointegrationsbasierte und korrelationsbasierte Verfahren. Engle et al. 1987 legten die theoretischen Grundlagen der Kointegration, Vidyamurthy et al. 2004 entwickelten deren Anwendung im Pairstrading, und Gatev et al. 2006 etablierten den korrelationsbasierten Ansatz. Pairstrading umfasst zwei fundamentale Schritte: das Identifizieren geeigneter Paare und deren Handel.

Der Kointegrationsansatz

Die Cointegration hat nichts mit der “Integration” zu tun wie der Name vermuten lässt. Nach Engle et al. 1987 werden Paare auf eine langfristige Gleichgewichtsbeziehung getestet. Die Zeitreihen können kurzfristig voneinander abweichen, werden jedoch in der thorie durch ökonomische Kräfte langfristig zusammengehalten.

Cointegrierte Paare finden: Um solche Beziehungen zu identifizieren, wird zunächst eine Regressionsgleichung geschätzt:

$$X_i(t) = \alpha_{i,j} + \beta_{i,j} X_j(t) + \varepsilon_{i,j}(t)$$

Die Residuen $\varepsilon_{i,j}(t)$ dieser Regression müssen stationär sein, auch wenn die ursprünglichen Zeitreihen $X_i(t)$ und $X_j(t)$ selbst nichtstationär sind. Ein stationärer Prozess kehrt zu seinem Mittelwert zurück und weist konstante Varianz auf. Mit dem Engle-Granger-Verfahren (cf. Engle Granger 1987) wird diese Stationarität der Residuen mittels Augmented Dickey-Fuller-Test (cf. Dickey 1979) geprüft. Die Nullhypothese lautet, dass die Residuen eine Einheitswurzel besitzen (nichtstationär sind). Wird diese Nullhypothese bei einem Signifikanzniveau α_{sig} verworfen (p-Wert $< \alpha_{\text{sig}}$), liegt Kointegration vor.

Cointegrierte Paare handeln: Aufgrund ihrer statistischen Eigenschaften (stabilisieren die Varianz) werden die Preise logarithmisch mit $\log X_i(t)$ und $\log X_j(t)$ angewandt. Um das optimale Verhältnis für einen Trade zu finden, wird die Hedge Ratio durch lineare Regression bestimmt:

$$\log X_i(t) = \alpha_{i,j} + \beta_{i,j} \log X_j(t) + \varepsilon_{i,j}(t)$$

Wobei $\alpha_{i,j}$ die Konstante (Intercept der Regression), $\beta_{i,j}$ das Hedge Ratio (gibt an, wie viele Einheiten von X_j pro Einheit X_i gehandelt werden) und $\varepsilon_{i,j}(t)$ der Fehlerterm (Residuen) ist. Das Hedge Ratio $\beta_{i,j}$ hat den Zweck, dass Long- und Short-Positionen marktneutral sind. Mit dem Hedge Ratio kann nun ein marktneutraler Spread konstruiert werden:

$$Spread_{i,j}(t) = \log X_i(t) - \beta_{i,j} \log X_j(t)$$

Dieser Spread eliminiert die gemeinsame Marktbewegung. Damit Trading damit aber profitabel ist, muss der Spread mean-reverting (mittelwertrückkehrend) sein:

$$Spread_{i,j}(t) = \mu_{i,j} + u_{i,j}(t)$$

wobei $u_{i,j}(t)$ ein stationärer Prozess mit $E[u_{i,j}(t)] = 0$ ist. Im Handel können profitable Abweichungen nur realisiert werden, wenn der Spread zum Mittelwert zurückkehrt. Um feststellen zu können, wann der Spread zu weit von seinem Mittelwert entfernt ist, wird er normalisiert:

$$Z_{i,j}(t) = \frac{Spread_{i,j}(t) - \mu_{i,j}}{\sigma_{i,j}}$$

Der Z-Score macht Abweichungen vergleichbar und definiert klare Ein- und Ausstiegssignale. Als Beispiel bei $\theta = 2$: Long bei $Z_{i,j}(t) < -\theta$ mit Long X_i, Short $\beta_{i,j} \cdot X_j$ oder Short bei $Z_{i,j}(t) > +\theta$ dann vice versa Short X_i, Long $\beta_{i,j} \cdot X_j$. Der Exit solcher Trades erfolgt bei $Z_{i,j}(t) \to 0$, da der Spread zur historischen Mitte zurückkehrt.

Der Korrelationsansatz

Nach Gatev et al. 2006 basiert der korreltaionsansatz auf der Annahme, dass $X_i(t)$ und $X_j(t)$ mit historisch hoher Korrelation auch in Zukunft eine ähnliche Preisbewegung aufweisen werden. Im Gegensatz zu kointegrationsbasierten Ansätzen wird hier keine langfristige Gleichgewichtsbeziehung vorausgesetzt sondern kurzfristige Korrelationsbezeihungen angenommen.

Korrealierende Paare finden: Bei korrelationsbasierten Verfahren erfolgt die Paarauswahl nach Sharpe 1964 mittels der Pearson-Korrelation in einer Formationsperiode der Länge t_f:

$$\rho_{i,j} = \frac{Cov(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}} = \frac{\sum_{t=1}^{t_f}(X_i(t) - \bar{X}_i)(X_j(t) - \bar{X}_j)}{\sqrt{\sum_{t=1}^{t_f}(X_i(t) - \bar{X}_i)^2}\sqrt{\sum_{t=1}^{t_f}(X_j(t) - \bar{X}_j)^2}}$$

wobei $\bar{X}_i$ und $\bar{X}_j$ die Mittelwerte der jeweiligen Zeitreihen in der Formationsperiode darstellen. Paare mit $\rho_{i,j} > \rho_{\min}$ oberhalb eines definierten Schwellenwerts werden für das Trading ausgewählt.

Korrelierte Paare handeln: Für das Handeln der Paare nach Gatev et al. 2006 werden die Zeitreihen zunächst normalisiert. Die kumulative Summe[1] der logarithmierten Renditen wird durch ihre historische Standardabweichung geteilt:

$$Z_k(t) = \frac{\sum_{s=1}^{t} \log \frac{X_k(s)}{X_k(s-1)}}{\sigma_k} \quad \text{für } k \in i, j$$

wobei σ_i und σ_j die Standardabweichungen der logarithmierten Renditen in der Formationsperiode sind. Der Spread zwischen den normalisierten Zeitreihen wird als Divergenz definiert: $D_{i,j}(t) = Z_i(t) - Z_j(t)$. Wenn die Divergenz einen Schwellenwert δ überschreitet werden entsprechende Handelssignale erzeugt. Long von X_i und Short X_j bei $D_{i,j}(t) < -\delta$ und Long von X_j und Short X_i bei $D_{i,j}(t) < +\delta$. Die Positionen werden geschlossen, wenn die Divergenz gegen null konvergiert ($D_{i,j}(t) \to 0$).

Ziel oder Motivation

Drei zentrale Faktoren reduzieren die Profitabilität statischer Pairstrading-Anwendungen im modernen Marktumfeld:

1. **Popularität:** Je mehr Marktteilnehmer dieselbe Strategie nutzen, desto schneller verschwinden die Arbitragemöglichkeiten (cf. Jacobs 2015). Die Anwendung von Pairs trading steigt seit Anfang der

2000er Jahre nachweisbar bei professional traders, institutional investors, and hedge fund managers (Miao 2014, p. 96).

2. **Eliminierung von Ineffizienzen:** Miao et al. 2014 verdeutlicht, wie High Frequency Handel (*HFT*) durch technologischen Fortschritt Markteffizienzen eliminiert. Während bei der Entwicklung der kointegrationsbasierten und korrelationsbasierten Ansätze Positionen über Wochen oder Tage gehalten wurden, werden heute Arbitragemöglichkeiten binnen Stunden, Minuten oder sogar Sekunden ausgenutzt. Die Ineffizienzen, auf denen klassisches Pairs Trading basiert, werden schneller erkannt und eliminiert, bevor traditionelle Ansätze reagieren können.

3. **Instabilität:** Chen et al. 2019 zeigen, dass Kointegrationsbeziehungen zwischen Aktienpaaren nicht stabil sind, sondern sich durch Regime-Wechsel grundlegend verändern oder vollständig verschwinden können. Marktbedingungen wechseln zwischen verschiedenen Zuständen - von kointegrierenden Beziehungen hin zu Black-Scholes-ähnlichen Märkten ohne statistische Arbitragemöglichkeiten.

Diese Herausforderungen eröffnen spezifische Ansatzpunkte zur Weiterentwicklung. Aufgrund der Ressourcen institutioneller High Frequency Handelsysteme macht es im Kontext der Markteffizienzen keinen Sinn, an den Handelsstrategien für vorhandene Paarfindungsstrategien direkt anzusetzen. Die Annahme ist, dass sobald eine Ineffizienz besteht, verschiedene Handelsstrategien zu ihr führen können, diese aber immer von Ressourcen (Rechen- und Übertragungsgeschwindigkeit) geschlagen werden können. Kommulieren eine großer Anteil der Marktteilnehmer durch die verwendung Dokumentierter und in der Vergangenheit erfolgreicher Verfahren wie der Cointegration, Korrelationen oder ihren Weiterentwicklungen, entstehen “crowded trades”, die bei ersten Verlusten zu gleichzeitigen Liquidationen und Verlustververstärkungen führen. Daher lautet die Hypthese dieser Arbeit die Profitabilitat und Optimizierung des Pairs Trading darin liegen kann, Paare finden zu können die durch diese traditionell statistischen Ansätze sonst verborgen bleiben. Zur Bewältigung der Instabilitätsproblematik werden adaptive Ansätze mit rollenden Fenstern und dynamischen Parametern eingesetzt, die sich kontinuierlich an veränderte Marktregime anpassen können.

Das führt zur folgenden Forschungsfrage: Inwieweit verbessert ein maschineller Lernansatz zur Paaridentifikation die Performance von Pairs Trading Strategien im Vergleich zu traditionellen kointegrationsbasierten Verfahren? Als explorative Arbeit ist die Vorgabe der maschinellen Lernansätze bewusst offen gehalten. Am Ende konnten zwei Ansätze erarbeitet werden: ein Affinity propagation Clustering und Gradient Boost Regressor Verfahren. Diese werden mit einem kointegrationsbasierten, nach Engle et al. 1987 implementierten Ansatz verglichen. Dieser erhält jedoch zur Vergleichbarkeit ebenfalls die Anwendung des Sliding Window Ansatzes. Vergleichsstudien begründen die Auswahl durch validierte Überlegenheit der Kointegrationsverfahren gegenüber der Korrelation (Rad 2016); (Carrasco Blázquez 2018); (Ma 2022).

Data

Um eine Teilmenge an Zeitreihen (Wertpapieren, Aktien oder sonstigem) zu erhalten auf die dann Verfahren zur Paarbildung angewendet werden können verwenden vergleichbare Arbeiten Vorauswahlen wie den Global Industry Classification Standard (GICS) (Do 2010, p. 8).

Abbildung 1: Systematische Darstellung der Wirtschaftsstruktur mit den Klassifikationsebenen: 1. Global Industry Classification Standard (*GICS*); 2. Nationale Sektoren (Primär-, Sekundär, Tertiärer-sektor); 3. Hierarchische Industriestruktur (Industriezweig; Industrie, Subindustrie); 4. Inländische Aktienindizes im Kontext der globalen und nationalen Wirtschaft.

GICS bildet die Basis für S&P und MSCI Finanzmarkt-Indizes, in denen jede Firma gemäß ihrer Hauptgeschäftstätigkeit genau einer Subindustrie und damit einer Industrie, einem Industrie-Zweig und einem Sektor zugewiesen ist. Die Eingrenzung an potenziellen Paaren bildet sich dann in den jeweils durch die Klassifizierung festgelegten Sektoren bzw. Subsektoren (siehe Abbildung 1, siehe 3). So können zwar wirtschaftliche Synergien in den Paaren vorliegen, jedoch können potenzielle Verbindungen von vornherein fälschlicherweise ausgeschlossen werden. Um diese Ausgrenzung größtmöglich zu abstrahieren, geschieht die Vorauswahl auf nationaler volkswirtschaftlicher Ebene (siehe Abbildung 1, siehe 2). Diese bewusste Entscheidung gegen sektorspezifische Vorfilterung folgt der Hypothese, dass maschinelle Lernverfahren auch branchenübergreifende Zusammenhänge identifizieren können, die durch traditionelle GICS-basierte Einschränkungen ausgeschlossen würden. Im Rahmen dieser Arbeit geschieht dies im Rahmen des

amerikanischen und britischen Marktes in Form des *NASDAQ-100* und *FTSE-100* (im Nachfolgenden abgekürzt als N_{100} und F_{100}) (siehe Abbildung 1, siehe 4).

Indizes: Die Aktiendaten beider Indizes erstrecken sich über den Zeitraum $T = [t_{\text{start}}, t_{\text{end}}]$, wobei t_{start} dem 01.01.2020 und t_{end} dem 01.01.2025 entspricht. Für jede Aktie wird an jedem Handelstag ein vollständiger Datensatz erfasst, der das Handelssymbol, das Datum sowie die wesentlichen Kursinformationen umfasst: Eröffnungskurs, Tageshöchstkurs, Tagestiefstkurs, Schlusskurs und Handelsvolumen. Da keine weiteren Einschränkungen bezüglich Marktkapitalisierung, Liquidität oder Branchenzugehörigkeit vorliegen, ergibt sich aus dem Pool aller 100 Aktien eines Index $A = a_1, a_2, \ldots, a_{100}$ die Menge aller zulässigen Paare durch $P = (a_i, a_j) \mid a_i, a_j \in A \land i < j$. Dies entspricht $\binom{100}{2} = 4950$ möglichen Paarkandidaten pro Index.

Abbildung 2: Die Abbildung zeigt die normalisierten Marktindizes F_{100} und N_{100} (2020-2025, Basis=100) mit Gesamtwachstum (F_{100}: +30,1%, N_{100}: +107,1%) und hervorgehobenem Wachstum im Jahr 2024 von F_{100}: +7,9%, N_{100}: +15,8%.

Datenbereinigung: Zeitreihen, die nicht die vollständige Anzahl an Datenpunkten für den Zeitraum $T = [t_{\text{start}}, t_{\text{end}}]$ vorweisen können, wurden ausgelassen. Das resultierte für N_{100} in 94 und F_{100} in 98 Aktien pro Index. Der Umgang mit vereinzelt fehlenden Werten erfolgt anhand der Kurszeitreihen $X_i(t)$ für jede Aktie i. Eine Vorwärtsinterpolation ersetzt fehlende Werte durch den letzten verfügbaren Wert derselben Zeitreihe: $X_i(t) = X_i(t - k)$, wobei k die kleinste positive Zahl ist, für die $X_i(t - k)$ beobachtet wurde. Anschließend werden bestehende Lücken durch nachfolgende Werte aufgefüllt: $X_i(t) = X_i(t + m)$, wobei m die kleinste positive Zahl ist, für die $X_i(t + m)$ beobachtet wurde. Im Vergleich zu anderen Methoden wie der linearen oder polynomialen Interpolation bietet die Vorwärts- und Rückwärtsinterpolation eine robuste Lösung für Finanzzeitreihen, da sie ausschließlich beobachtete Werte verwendet und so die Kontinuität der Daten bewahrt (Moritz 2017). Der bereinigte, normalisierter Kursverlauf aller vorhanden Werte pro Indices kann Abbildung 2 entnommen werden.

ML1

ML2

Window

Data

Cite Example

bla Zur Kostenfunktion Machine Learning () ist wichtig.

(cf. Engle Granger 1987) (Engle Granger et al. 1987, S. 255)

Laut cf. Jacobs et al. 2015, S. 255 ist…

Literatur

[Sharpe 1964] Sharpe, W. F. (1964). Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk. The Journal of Finance, 19(3), 425-442. https://doi.org/10.1111/j.1540-6261.1964.tb02865.x
[Dickey 1979] Dickey, D. A., & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. Journal of the American Statistical Association, 74, 427-431. https://api.semanticscholar.org/CorpusID:56458593

[Engle 1987] Engle, R. F., & Granger, C. W. J. (1987). Co-integration and Error Correction: Representation, Estimation, and Testing. Econometrica, 55(2), 251-276. https://ideas.repec.org/a/ecm/emetrp/v55y1987i2p251-76.html
[Vidyamurthy 2004] Vidyamurthy, G. (2004). Pairs Trading: Quantitative Methods and Analysis. John Wiley & Sons, Hoboken, New Jersey.
[Gatev 2006] Gatev, E., Goetzmann, W. N., & Rouwenhorst, K. G. (2006). Pairs Trading: Performance of a Relative Value Arbitrage Rule. Yale ICF Working Paper No. 08-03. https://doi.org/10.2139/ssrn.141615
[Do 2010] Do, B., & Faff, R. (2010). Does simple pairs trading still work? Financial Analysts Journal, 66(4), 83-95. https://doi.org/10.2469/faj.v66.n4.1
[Miao 2014] Miao, G. J. (2014). High frequency and dynamic pairs trading based on statistical arbitrage using a two-stage correlation and cointegration approach. International Journal of Economics and Finance, 6, 96. https://api.semanticscholar.org/CorpusID:54175142
[Jacobs 2015] Jacobs, H., & Weber, M. (2015). On the determinants of pairs trading profitability. Journal of Financial Markets, 23, 75-97. https://doi.org/10.1016/j.finmar.2014.12.001
[Rad 2016] Rad, H., Low, R. K. Y., & Faff, R. (2016). The profitability of pairs trading strategies: distance, cointegration and copula methods. Quantitative Finance, 16(10), 1541-1558. https://doi.org/10.1080/14697688.2016.1164337
[Moritz 2017] Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: Time series missing value imputation in R. The R Journal, 9(1), 207-218. https://doi.org/10.32614/RJ-2017-009
[Carrasco Blázquez 2018] Carrasco Blázquez, M., De la Orden De la Cruz, C., & Prado Román, C. (2018). Pairs trading techniques: An empirical contrast. European Research on Management and Business Economics, 24(3), 160-167. https://doi.org/10.1016/j.iedeen.2018.05.002
[Chen 2019] Chen, K., Chiu, M. C., & Wong, H. Y. (2019). Time-consistent mean-variance pairs-trading under regime-switching cointegration. SIAM Journal on Financial Mathematics, 10(2), 632-665. https://doi.org/10.1137/18M1209611
[Ma 2022] Ma, B., & Ślepaczuk, R. (2022). The profitability of pairs trading strategies on Hong-Kong stock market: distance, cointegration, and correlation methods. Working Papers 2022-02, Faculty of Economic Sciences, University of Warsaw. https://ideas.repec.org/p/war/wpaper/2022-02.html

---

1. Die kumulative Summe der Log-Renditen ist die laufende Aufsummierung aller bisherigen Renditen und zeigt, wie sich der Preis insgesamt vom Startpunkt verändert hat. ↩

back to top