

Costruzione di portafogli diversificati mediante algoritmi di *clustering*

INDICE

- Premessa e obiettivi
- Sviluppo del lavoro
- Allocazione di portafoglio
- Portafoglio di tangenza
- *K-Means*
- *Hierarchical Risk Parity*
- *Bounded K-Means*
- *Backtest* dei portafogli
- Valutazione degli algoritmi di *clustering*
- Conclusioni e possibili sviluppi

Premessa e obiettivi

- Scopo del progetto è quello di valutare la *performance* di tre portafogli diversificati costruiti mediante algoritmi di *clustering* partizionale e gerarchico.
- In un'ottica di allocazione di portafoglio, la *clustering analysis* può essere utilizzata per selezionare, attraverso un processo di apprendimento non-supervisionato, gruppi di *asset* omogenei sulla base della correlazione fra rendimenti.
- Il capitale viene allocato fra i *cluster* di *asset* prodotti utilizzando una strategia di ottimizzazione in media-varianza e la *performance* del portafoglio così ottenuto viene valutata rispetto ad un portafoglio *benchmark*.

Sviluppo del lavoro

- Si selezionano casualmente circa 200 titoli azionari provenienti dall'indice Nasdaq.
- Si utilizzano diversi algoritmi di *clustering* partizionale e gerarchico per suddividere i rendimenti in *cluster* omogenei sulla base della correlazione.
- La bontà di ogni algoritmo di *clustering* viene valutata in base alla correlazione infra- e intra-*cluster* e alla numerosità dei *cluster* prodotti.
- I *cluster* ottenuti vengono aggregati in portafogli equiponderati, su cui si applica una strategia di ottimizzazione in media-varianza al fine di ottenere i pesi ottimali dei singoli *asset*.
- Nell'allocazione statica, i pesi ottimali di portafoglio vengono mantenuti costanti durante tutto il *test set*, pari all'ultimo anno.
- Nell'allocazione *rolling*, i pesi di portafoglio sono calcolati su finestre *rolling* giornaliere di ampiezza annuale, che simulano un ribilanciamento giornaliero del portafoglio.
- Per ogni algoritmo di *clustering*, i risultati dell'allocazione statica e *rolling* sono confrontati al fine di calcolare l'extrarendimento dovuto al ribilanciamento.
- La *performance* dei portafogli di *clustering* viene infine valutata rispetto a due portafogli *benchmark*: il portafoglio di tangenza e un portafoglio *equally weighted*.

Allocazione di portafoglio

- I *cluster* prodotti da ciascun algoritmo vengono utilizzati per formare dei portafogli *equally weighted* contenenti gli *asset* appartenenti ad ogni *cluster*.
- A questi portafogli si applica una strategia di allocazione in media-varianza al fine di ottenere un singolo portafoglio che contiene tutti gli *asset* disponibili.
- I pesi ottimali sono calcolati massimizzando la seguente funzione:

$$\max_{\omega} \frac{\omega' \mu - rf}{(\omega' \Sigma \omega)}, \text{ s.t. } \omega' 1 = 1$$

dove ω è il vettore di pesi ottimali di portafoglio, μ è il rendimento atteso del portafoglio, Σ è la matrice di covarianza dei rendimenti e rf è il tasso privo di rischio.

- Nell'allocazione statica, i pesi di portafoglio ottimali vengono mantenuti costanti durante tutto il *test set*, pari all'ultimo anno.
- Nell'allocazione *rolling*, i pesi di portafoglio sono calcolati su finestre *rolling* giornaliere di ampiezza annuale.

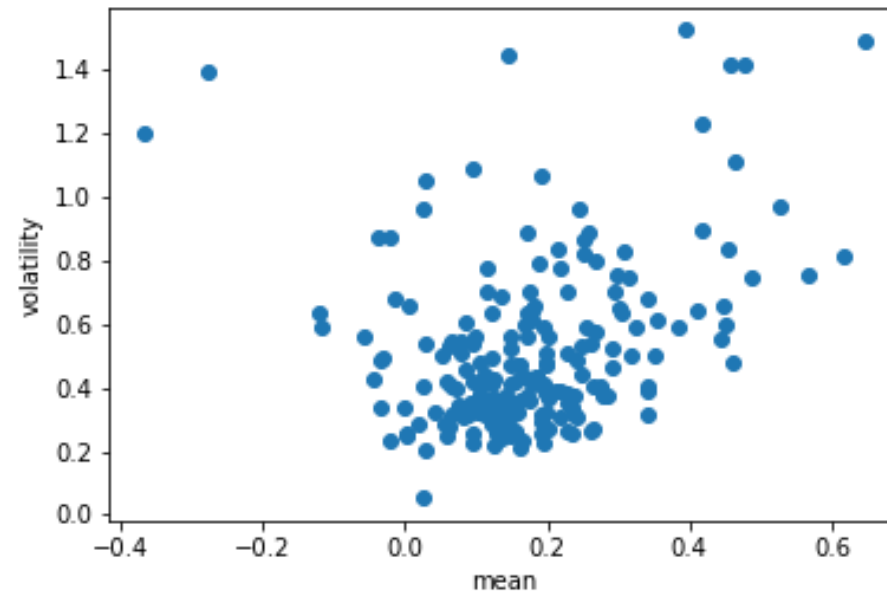
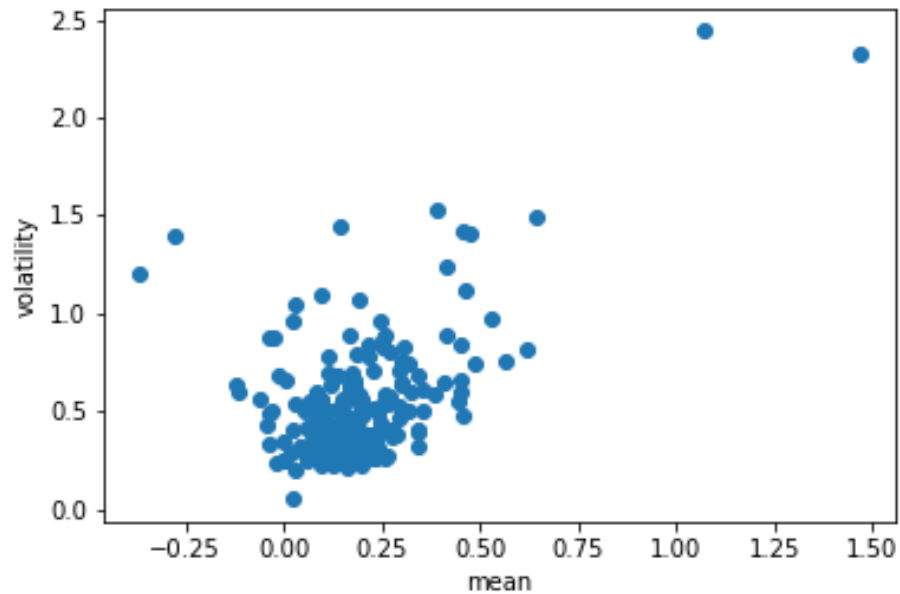
Portafoglio di tangenza

Pesi del portafoglio *rolling*

date	AAL	ACAD	ACGL	ACTG	ACUR	ADMP	AEGN	AGNC	AINV
2019-12-31	0.0	0.1	6.24	0.0	0.27	0.0	0.0	0.0	0.0
2020-01-02	0.0	0.0	6.76	0.0	0.32	0.0	0.0	0.0	0.0
2020-01-03	0.0	0.0	7.29	0.0	0.3	0.0	0.0	0.0	0.0
2020-01-06	0.0	0.0	7.86	0.0	0.29	0.0	0.0	0.0	0.0
2020-01-07	0.0	0.0	7.09	0.0	0.24	0.0	0.0	0.0	0.0
2020-01-08	0.0	0.0	5.98	0.0	0.22	0.0	0.0	0.0	0.0
2020-01-09	0.0	0.0	5.46	0.0	0.29	0.0	0.0	0.0	0.0
2020-01-10	0.0	0.0	6.36	0.0	0.18	0.0	0.0	0.0	0.0
2020-01-13	0.0	0.0	5.28	0.0	0.15	0.0	0.0	0.0	0.0
2020-01-14	0.0	0.0	4.05	0.0	0.19	0.0	0.0	0.0	0.0

K-Means

- Gli *outlier* presenti nei rendimenti azionari sono preventivamente eliminati al fine di evitare che *K-Means* produca *cluster* contenenti *asset* singoli.
- Gli *outlier* sono identificati mediante uno *scatterplot* della media e della volatilità dei rendimenti annualizzati.

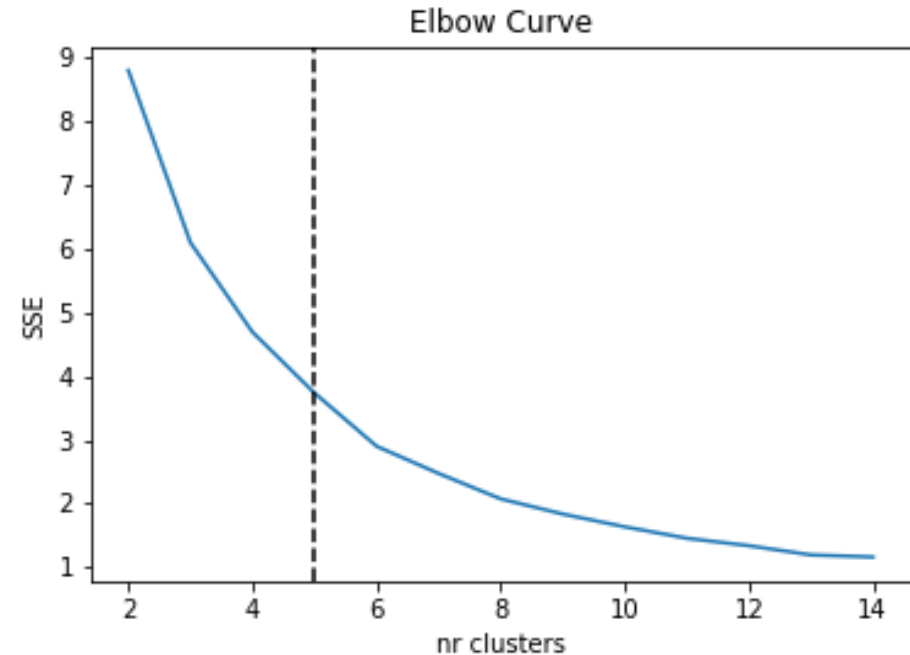


K-Means

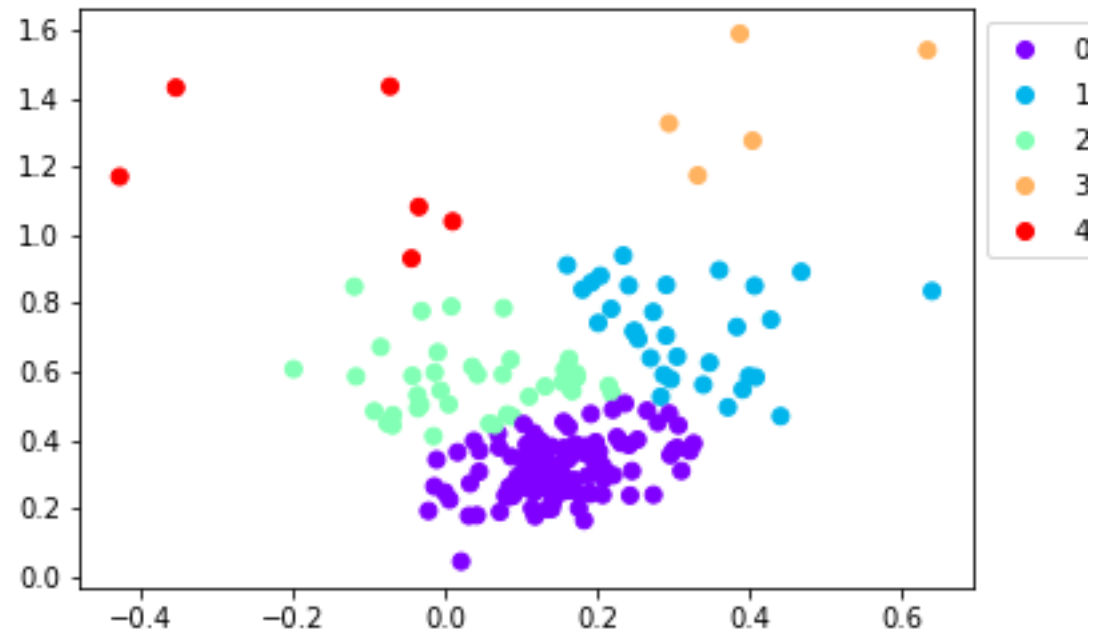
- L'algoritmo *K-Means* divide un insieme di N campioni X in C cluster disgiunti, ciascuno descritto dalla media μ_i dei campioni nel *cluster* (il centroide).
- L'algoritmo mira a scegliere i centroidi che minimizzano la somma dei quadrati all'interno del cluster (SSE):

$$\min_{\mu_i \in C} \sum_{i=0}^n (|x_i - \mu_i|)^2$$

- Il numero ottimale di *cluster* è selezionato come il punto in cui il tasso di decrescita del SSE rispetto al numero di *cluster* subisce un rallentamento significativo (*elbow rule*).



K-Means



Pesi del portafoglio statico

cluster	nr elements	static weights
0.0	114.0	0.0
1.0	32.0	60.0
2.0	39.0	26.0
3.0	5.0	14.0
4.0	6.0	0.0

K-Means

Pesi del portafoglio *rolling*

date	0	1	2	3	4
2019-12-31	0.0	88.0	0.0	12.0	0.0
2020-01-02	0.0	88.0	0.0	12.0	0.0
2020-01-03	0.0	88.0	0.0	12.0	0.0
2020-01-06	0.0	87.0	0.0	13.0	0.0
2020-01-07	0.0	88.0	0.0	12.0	0.0
2020-01-08	0.0	88.0	0.0	12.0	0.0
2020-01-09	0.0	87.0	0.0	13.0	0.0
2020-01-10	0.0	88.0	0.0	12.0	0.0
2020-01-13	0.0	88.0	0.0	12.0	0.0
2020-01-14	0.0	86.0	0.0	14.0	0.0

Hierarchical Risk Parity

- Si applica un algoritmo di *clustering* gerarchico alla matrice di covarianza dei rendimenti azionari, utilizzando una matrice di distanza \tilde{d} basata sull'indice di correlazione dei rendimenti degli *asset*:

$$\tilde{d}_{ij} = \sqrt{\sum_{k=1}^N (d_{ki} - d_{kj})^2}$$

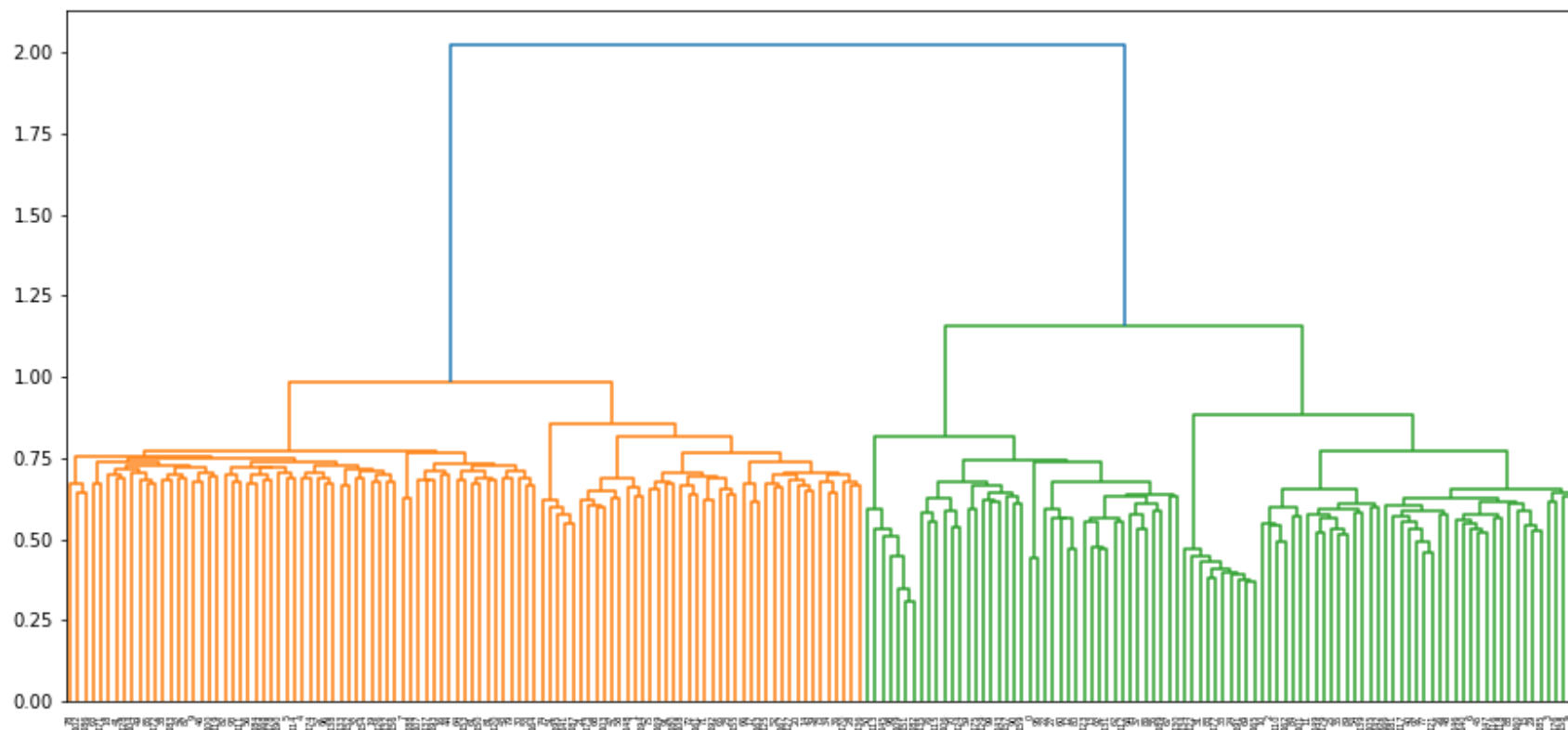
dove $d_{ij} = \sqrt{0.5(1 - \rho_{ij})}$.

- Si forma poi il primo *cluster* $U[1] = (i^*, j^*) = \operatorname{argmin}_{(i,j)} \tilde{d}_{ij}$. La matrice \tilde{d} viene quindi aggiornata rimuovendo le righe e le colonne corrispondenti al primo *cluster* e calcolando la distanza fra il *cluster* e l'*i*-esimo *asset*:

$$\tilde{d}(i, U[1]) = \min(\tilde{d}(i, i^*), \tilde{d}(i, j^*))$$

- La procedura viene ripetuta ricorsivamente fino ad ottenere il *cluster* finale contenente tutti gli *asset* disponibili.
- Utilizzando l'ordine dei *cluster*, si riorganizza la matrice di covarianza in modo da ottenere una matrice quasi-diagonale in cui gli *asset* simili sono posti nei pressi della diagonale.
- Si ottengono i pesi ottimali utilizzando l'approccio *inverse variance*, in cui il peso è inversamente proporzionale al rischio degli *asset*, e bisezionando ricorsivamente la matrice di covarianza seguendo l'ordinamento dei *cluster*.

Hierarchical Risk Parity



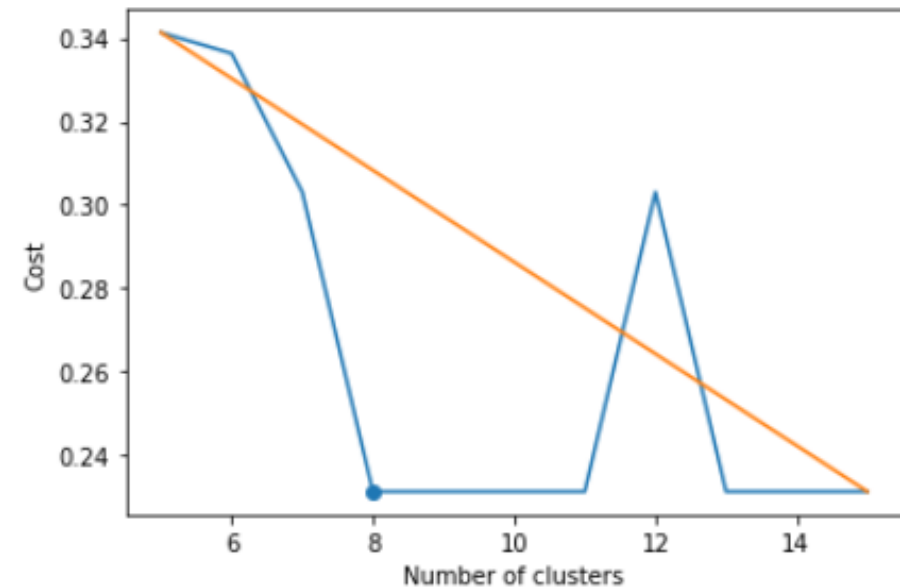
Hierarchical Risk Parity

Pesi del portafoglio *rolling*

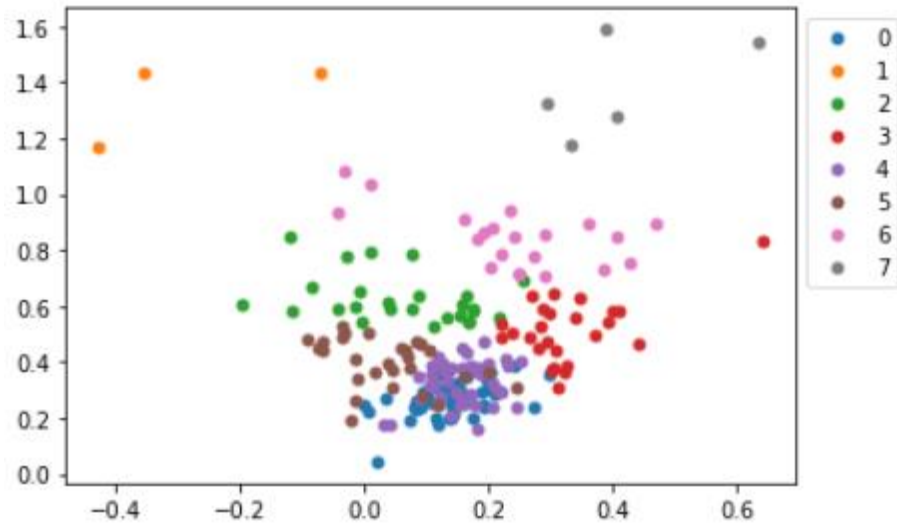
date	AAL	ACAD	ACGL	ACTG	ACUR	ADMP	AEGN	AGNC	AINV
2019-12-31	0.03	0.02	0.19	0.19	0.02	0.07	0.07	0.63	0.33
2020-01-02	0.04	0.03	0.18	0.22	0.02	0.08	0.07	0.71	0.2
2020-01-03	0.03	0.04	0.19	0.22	0.02	0.08	0.08	0.6	0.28
2020-01-06	0.05	0.03	0.19	0.17	0.02	0.07	0.06	0.6	0.33
2020-01-07	0.04	0.03	0.18	0.16	0.01	0.07	0.07	0.52	0.42
2020-01-08	0.06	0.03	0.2	0.19	0.02	0.06	0.08	0.5	0.44
2020-01-09	0.05	0.03	0.19	0.16	0.02	0.07	0.06	0.52	0.43
2020-01-10	0.06	0.03	0.18	0.17	0.02	0.07	0.06	0.52	0.4
2020-01-13	0.04	0.03	0.17	0.16	0.02	0.06	0.07	0.55	0.34
2020-01-14	0.04	0.03	0.17	0.16	0.02	0.06	0.07	0.54	0.34

Bounded K-Means

- *K-Means* non permette di imporre vincoli alla numerosità dei *cluster* \rightarrow *Bounded K-Means*
- $$c_j^{(t)} = \{x_p : \|x_p - \mu_j^{(t)}\|^2 \leq \|x_p - \mu_i^{(t)}\|^2, |c_j^{(t)}| < \zeta_j, 1 \leq i \leq k\}.$$
- Dove x_p indica un generico punto, la prima disequazione paragona la SSE del *cluster* j -esimo con quella del *cluster* i -esimo, e la seconda disequazione permette il controllo sulla numerosità massima del *cluster* ζ_j .
- Il numero ottimale di *cluster* è determinato come per *K-Means* tramite la cosiddetta «*elbow rule*»



Bounded K-Means



Pesi del portafoglio statico

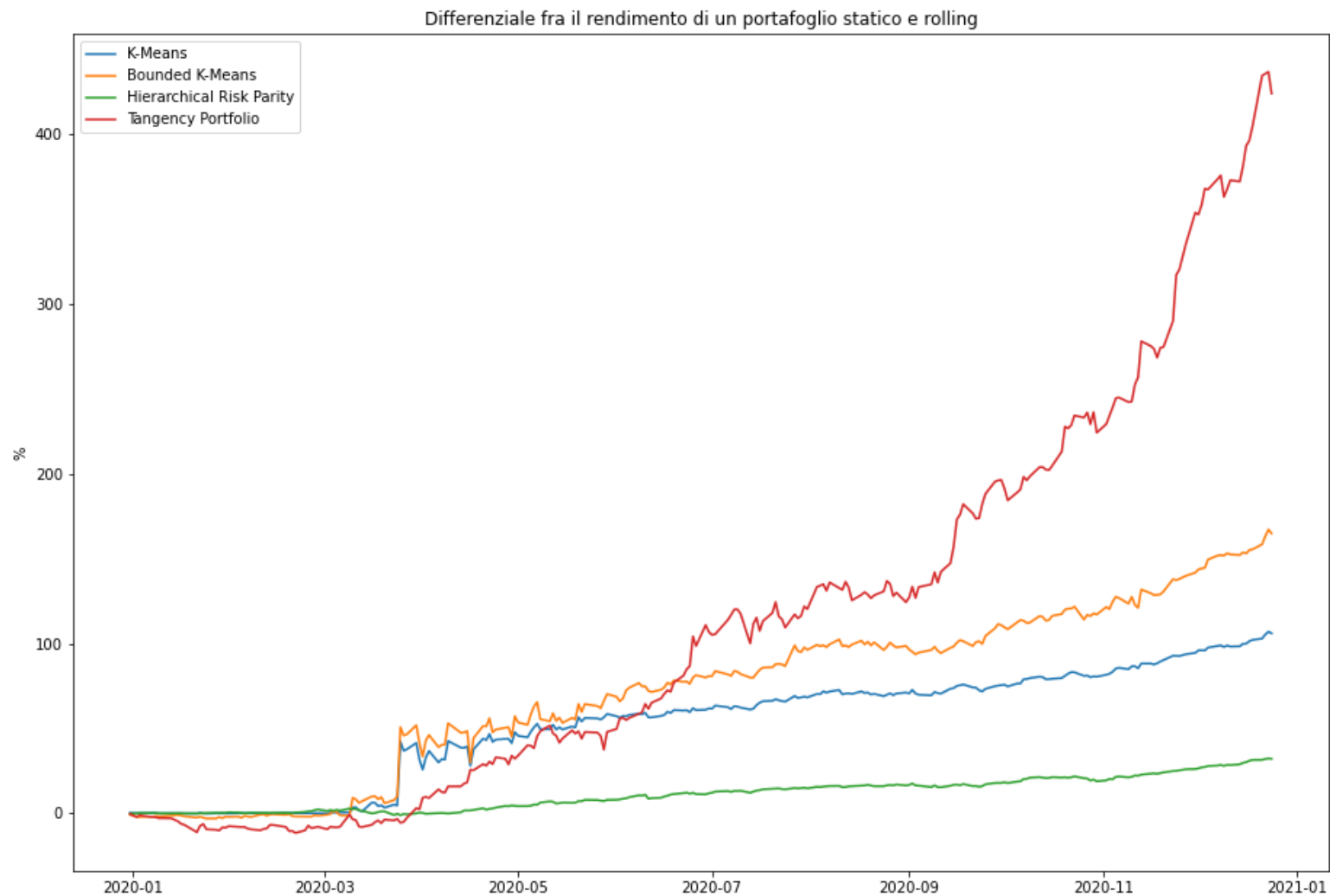
cluster	nr elements	static weights
0.0	44.0	0.0
1.0	3.0	0.0
2.0	26.0	52.0
3.0	25.0	0.0
4.0	44.0	0.0
5.0	28.0	0.0
6.0	21.0	35.0
7.0	5.0	14.0

Bounded K-Means

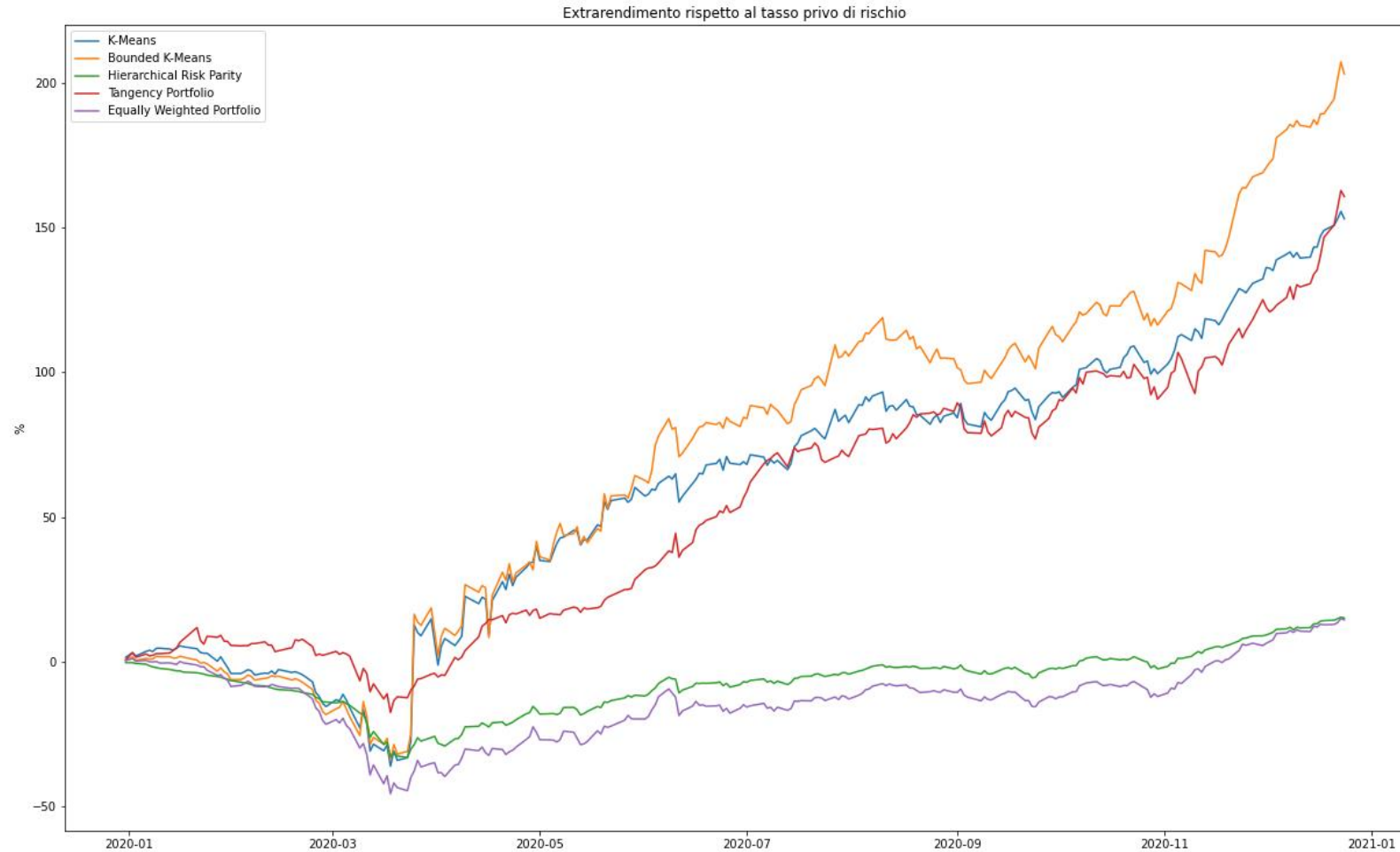
Pesi del portafoglio *rolling*

date	0	1	2	3	4	5	6	7
2019-12-31	18.0	0.0	20.0	35.0	0.0	0.0	17.0	10.0
2020-01-02	17.0	0.0	13.0	42.0	0.0	0.0	18.0	10.0
2020-01-03	28.0	0.0	10.0	31.0	0.0	0.0	20.0	10.0
2020-01-06	29.0	0.0	5.0	33.0	0.0	0.0	22.0	11.0
2020-01-07	26.0	0.0	4.0	31.0	0.0	0.0	27.0	11.0
2020-01-08	31.0	0.0	1.0	32.0	0.0	0.0	25.0	11.0
2020-01-09	28.0	0.0	0.0	32.0	0.0	0.0	29.0	12.0
2020-01-10	29.0	0.0	0.0	29.0	0.0	0.0	31.0	11.0
2020-01-13	42.0	0.0	0.0	25.0	0.0	0.0	23.0	10.0
2020-01-14	45.0	0.0	0.0	26.0	0.0	0.0	18.0	11.0

Backtest dei portafogli



Backtest dei portafogli



Backtest dei portafogli

	K-Means	Bounded K-Means	HRP	Tangency	Equal Weights
Cumulative Returns	153.0	203.0	15.0	161.0	14.0
Ann. Sharpe Ratios	1.63	1.83	0.79	4.2	0.56
Ann. Expected Returns	1.4	1.61	0.42	0.9	0.46
Ann. Volatility	0.7	0.74	0.21	0.24	0.37

Valutazione del clustering

- Il coefficiente *Silhouette* misura la distanza media fra un'osservazione e gli altri elementi dello stesso *cluster* (a) rispetto alla distanza media fra l'osservazione e gli elementi degli altri *cluster* (b). Il suo valore è compreso fra -1 e +1.
- L'indice di *Calinski-Harabasz* è dato dal rapporto fra la dispersione infra-*cluster* B_k e la dispersione intra-*cluster* W_k . Più il suo valore è elevato, migliore è la *performance* dell'algoritmo.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$ch(k) = \frac{trB_k/(k-1)}{trW_k/(n-k)}$$

	K-Means	Bounded K-Means
Silhouette coefficient	0.45	0.16
Calinski-Harabasz Index	186.63	135.6

Conclusioni e possibili sviluppi

- È stato dimostrato che gli algoritmi di *clustering* possono essere utilizzati per costruire portafogli diversificati che evidenziano una buona *performance out of sample*.
- Nell'ultimo anno, il portafoglio *Bounded K-Means* avrebbe garantito l'extrarendimento maggiore, mentre il portafoglio di tangenza avrebbe dato luogo allo *Sharpe Ratio* più elevato.

Possibili estensioni del progetto:

- Inclusione di titoli di Stato, obbligazioni *corporate* o indici di mercato;
- Utilizzo di una strategia di ottimizzazione che consenta di aprire posizioni *short*;
- Utilizzo di una strategia di ottimizzazione in media-varianza o di allocazione del rischio (*risk parity*, *inverse variance*) all'interno dei *cluster*;
- Utilizzo di altri algoritmi di *clustering*;
- *Backtest* dei portafogli su fasi rialziste del mercato e/o su fasi ribassiste precedenti (crisi finanziaria, crisi del debito sovrano);
- Verifica di come la scelta del numero ottimale di *cluster* influenzi la *performance* dei portafogli.

Riferimenti bibliografici

- De Prado, M. L. (2016). *Building diversified portfolios that outperform out of sample*. The Journal of Portfolio Management, 42(4), 59-69.
- Ganganath, N., Cheng, C. T., & Chi, K. T. (2014). *Data clustering with cluster size constraints using a modified k-means algorithm*. In *2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery* (pp. 158-161). IEEE.
- Markowitz H. (1959). *Portfolio Selection: Efficient Diversification of Investment*. (J. Wiley, New York).
- Tola, V., Lillo, F., Gallegati, M., & Mantegna, R. N. (2008). *Cluster analysis for portfolio optimization*. Journal of Economic Dynamics and Control, 32(1), 235-258.