Derivatives Packaging

Sterling Xie

February 2025

1 Overview

This project analyzes the capacity to group various derivatives into "bins", "groups", or "clusters" considering transaction costs, execution on such groups, and strike price movement. To complete this, we work across three grouping methodologies. The first is hierarchical clustering which prioritizes strike price distance. Hierarchical implementations of agglomerative clustering are inherently distance-based and are corrective across neighboring clusters. Strikes are assigned to the closest representative strike in the group. The second is K-Means clustering which partitions strikes into a predefined number of clusters by minimizing variance within each group. K-Means ensures minimum variance by iteratively adjusting the centroid of each bin. The third is strict or fixed binning which assumes the distribution of strikes is uniform and sequentially assigns ordered strikes into bins of the size of a set maximum distance. This approach does not consider the variance of the set. Because it lacks predictive power, it is used as an error benchmark on the first two clustering methods. It is also caveated that both K-Means clustering and binning will default to generating a bin for each individual option when the maximum distance is set at None.

2 Hierarchical Clustering

2.1 Advantages

Ward's method of agglomerative clustering dynamically adjusts the number of clusters based on the dataset which allows for minimization of clusters starting from a pre-set n clusters where n is the number of items in the set. Because this maximizes the notional within each cluster, it both maximizes execution performance and minimizes transaction costs. This method also maintains the relationships between close strike prices (in case there are similar factors impacting pricing movements between various options) because it clusters in an iterative tree model. Finally, agglomerative clustering iteratively corrects groups by merging the closest clusters at each level of the tree, ensuring that strike prices are grouped based on their natural proximity. This means if initial groupings

result in an uneven density of strikes across clusters, merges will adjust these clusters to gradually correct such imbalances. Such correction provides well-rounded clusters for more even option transaction schemes.

2.2 Disadvantages

The algorithm has a time complexity of $O(n^2)$ which makes it inefficient for scaling to larger datasets. Unlike K-Means, cluster assignments in agglomerative clustering are greedy so if outliers are grouped, it cannot be undone.

3 K-Means Clustering

3.1 Advantages

K-Means clustering prioritizes variance minimization via centroid adjustment that ensures each strike is assigned to the closest representative strike that minimizes variance. K-Means also runs in $O(nk)$ time, which makes it more computationally scalable and converges faster than hierarchical methods.

3.2 Disadvantages

The algorithm requires a pre-specified number of clusters. In this project's implementation, this is defined by $\frac{max\_strike - min\_strike}{max\_distance}$. K-Means also assumes a Gaussian distribution of strikes which is not universally true.

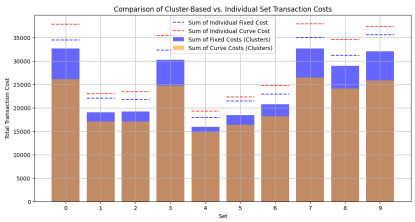
4 Performance Comparison

4.1 Approach

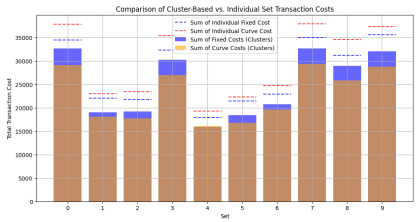
The order of evaluation is cluster tightness, then transaction costs relative to the cost of purchasing each option individually, and finally notional value per group. Cluster tightness is defined by inertia and cluster similarity is defined by its silhouette score. This order of evaluation is chosen because transaction cost acts as a precursor to profitability over notional value. Even if increasing notional value increases the probability the trade executes, it is likely worth holding if the transaction cost per unit is high. Also, cluster tightness is important because highly variant clusters will lead to more orders getting filled with worse prices whereas similar strikes would decrease risk of slippage. Thus, we evaluate on cluster tightness (which enhances hedge exposure), then evaluate first on transaction cost minimization then on notional maximization. Transaction cost minimization and notional maximization are inherently correlated because lower relative transaction cost under both the fixed and curve models occurs on sets with fewer groups and therefore higher notional value per group.

4.2 Method Selection

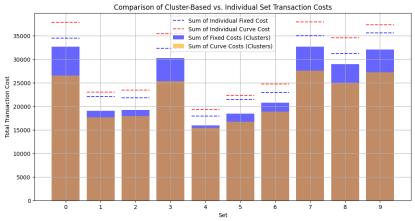
We take a case study of the clustering at a maximum distance setting of 0.02.



(a) K-Means Clustering



(b) Hierarchical Clustering



(c) Fixed Bin Clustering

Figure 1: Cost Models

Most differences between the clustering methods are relatively insignificant, implying the data is well-ordered. However, for sets 3, 7, 8 and 9, K-Means clustering performs significantly better than both alternatives. Being able to customize the number of clusters probably attributes to this: variance across strike prices is low regardless of the cluster size so creating larger clusters is more effective. Thus, while the agglomerative method created 7 clusters in set 9, for example, the K-Means method only created 3. It is true that inertia was sacrificed as the K-Means method had a score of 0.01646454363698803 compared to the agglomerative method's 0.005593591940476201. However, this increase in variance is balanced by the reduced cluster count and transaction cost.

The gaps in transaction costs become more evident across all sets and pronounced in magnitude as the maximum distance increases, likely because increasing maximum distance decreases the number of sets our implementation of K-Means creates. This would increase the notional and generally decrease net transaction cost.

Note: Additional materials and graphics for all other maximum distance settings can be accessed in the Jupyter notebooks in this repository.

4.3 Algorithm Recommendation

Because "execution" is very difficult to quantify, it is difficult to provide an algorithm recommendation without knowledge of market conditions (i.e. the order book depth on vanilla options at any specific time) which would then be used to weight the metrics described in the approach. Modeling of vanilla option pricing trends and risk thresholds would also be used to determine what variance of strikes would still meet our criteria.

Tentatively, the recommendation would be to use some hybridization of K-Means and Agglomerative clustering. K-Means would be used for lower maximum distance restrictions while agglomerative clustering would be used if there are no restrictions and for higher restrictions.