Portfolio Optimization with Market Regime Classification using Gaussian Mixtures

Pengbo Hao
Cornell University
ph396@cornell.edu

Jingyu Shen
Cornell University
js2836@cornell.edu

Kornelia Wang
Cornell University
xw488@cornell.edu

May 17, 2022

Abstract

Portfolio optimizations always involve the tasks of parameter estimations and formulations of the optimization problem. With the abundance of highly optimized off-the-shelf optimization softwares and packages, accurate estimations of volatile market parameters (e.g. return and variance in Markowitz) are becoming the bottleneck of portfolio optimizations nowadays. Classifying market regimes is one of the most popular ways that ease parameter estimation. Market regimes are periods of persistent market conditions, where investment decisions can be made with similar logic and reasoning's because of similar market behaviors within such periods. We therefore propose the use of Gaussian mixtures with Expectation-Maximization (EM) algorithm for market regime classification to assist in estimating the parameters of returns and covariance matrices used in Markowitz portfolio optimization. Experiments are conducted on the U.S. and European stock markets, using S&P 500 and FTSE 350 constituent stocks, and we later interpret the backtesting performances of our proposed algorithm for the overall investment horizon and during times of market crises.

I. Introduction

Investment decisions are often made with a forecast on how markets will evolve in the near future, and this forecasting process often involves estimating future returns and volatilities. Parameter estimation in the domain of quantitative finance is generally difficult due to the unpredictable nature of stock markets and irrational behavior of investors. Therefore, one of the biggest motivations for classifying market regimes comes to its ability to simplify the parameter estimation process by limiting the analysis to only the most relevant market regime that is mostly likely to be dominant.

In this project, we propose the use of Gaussian mixtures clustering algorithm along with Expectation Maximization to cluster market regimes in the U.S. and European stock markets in the time spans from January 1995 to March 2022 and from May 2006 to March 2022 respectively. A trading strategy is developed using ARIMA to predict the one step ahead market regime, and subsequently using that to estimate the parameters of mean returns and covariance matrices used in Markowitz portfolio optimizations. A variety of configurations (i.e. numbers of clusters, lengths of the look-back period, features used for clustering, and lengths of the holding period) are experimented. We analyze how much risk adjusted returns each trading strategy generates overall and how they are capable of control drawdowns in times of market crises (i.e. Dot-com Bubble, Subprime Mortgage Crisis, and 2020 Stock Market Crash).

II. Literature Review

Market regime clustering has always been a popular area of quantitative finance, and a wide range of literature have already paved ways for algorithms to detect market regimes. Haase & Neuenkirch, have used the traditional principal component analysis and shrinkage methods to aggregate high-dimensional macroeconomic and financial variables, which are then used to build one-step Markov-switching models with time-varying transition probabilities to make predictions on the stock market. It is concluded by Haase & Neuenkirch, that such methods are able to classify bull and bear markets more accurately than simply relying on common benchmarks. Munnix et al., has adopted the method of K-means and used the sum of absolute values of stocks' correlation coefficients as the similarity measure, and has pointed out that this method can generate regimes that can only persist for very short periods of time and are sparsely embedded in time. In order to find clusters that are representative of the overall market, Tang, Xu, & Zhou, has developed the algorithm of correlation block-model that clusters assets into groups with the maximum joint correlation with all other assets. It is concluded by Tang et al., that picking a certain number of stocks from each cluster instead of invest in all of them can still ensure portfolio diversification, and such strategies have outperformed benchmark returns under various scenarios.

Clustering market regimes is gradually gaining its popularity in the field of portfolio optimization, where parameter estimation plays a big role. Some literature are going one step further to explore ways of predicting future market regime and thus parameters used in optimization models using clustering algorithms. Procacci & Aste, and their paper's corresponding extension Wang & Aste, , have developed more novel approaches of identifying market regimes by a reference sparse precision matrix or inverse covariance matrix along with a vector of expectation values, after which clusters are generated by minimizing a penalized Mahalanobis distance. They have concluded that using Mahalanobis distance not only makes the algorithm computationally efficient but also helps produce distinguishable pre- and post-crisis clusters. The clusters can then be used to estimate parameters (i.e. return vector and covariance matrices) used in standard Markowitz.

Despite all the previous novel approaches to cluster and predict market regimes, our project aims to fill the gap for the lack of extensive research on Gaussian mixtures and its efficacy (i.e. generating soft assignments and more complex decision boundaries as will be later discussed in Section IV) in the field of quantitative finance.

III. Data

In this section we will describe the raw data that are obtained from Yahoo Finance in Section i, and will describe in detail the input features used in our Gaussian mixture model in Section ii.

i. Raw Data Description

The datasets that are used in our project are the historical prices of the constituent stocks of two market indices - S&P 500 and FTSE 350. In regards to some constituent stocks being relatively new and some becoming out of fashion and therefore dropped out of the indices lately, the asset pool is filtered in the following way in order to make our empirical analysis more relevant: for both indices, the most liquid 300 stocks as of March 17, 2022 are selected out of each market index, after which 200 stocks that have survived the longest are selected to our asset pool used for empirical analysis. Note here that with this way of constructing asset pool, evaluating investment strategies using market indices as benchmarks is inherently biased. To deal with this issue, equally weighted portfolios are used as benchmarks in Section V. After the filtering procedure described above, we have left with 200 stocks from the S&P 500 index spanning from November 1999 to March 2022 and 200 stocks from the FTSE 350 index spanning from April 2006 to March 2022.

ii. Feature Engineering

Using Gaussian mixtures to perform the task of clustering requires input features. Here we will conduct experiments on a set of features that can potentially characterize the market - return (R), volume-weighted exponential moving average of returns (EMA), moving average convergence divergence (MACD), and relative strength index (RSI). All the features are in a daily frequency and are weighted by their trading volumes. The calculations of these features are shown in Table 1 below.

Feature	Formula
$Return(t)$	$R(t)$
$EMA(t)$	$EMA(t-1) + K(R(t) - EMA(t-1))$
	$K = \frac{2}{30+1}$
$MACD(t)$	$shortema(t) - longema(t)$
	$shortema(t) = 0.15 \cdot p(t) + 0.85 \cdot shortema(t-1)$
	$longema(t) = 0.075 \cdot p(t) + 0.925 \cdot longema(t-1)$
$RSI(t)$	$100 \cdot \frac{upavg(t)}{upavg(t)+dnavg(t)}$
	$upavg(t) = \frac{upavg(t-1)\cdot(14-1)+up(t)}{14}$
	$dnavg(t) = \frac{dnavg(t-1)\cdot(14-1)+dn(t)}{14}$
	$up(t) = max\{close(t) - close(t-1), 0\}$
	$dn(t) = max\{close(t-1) - close(t), 0\}$

$Return(t), EMA(t), MACD(t), RSI(t) \in \mathbb{R}^N$, where $N = 200$ representing the 200 stocks in each market; All the features are element-wise multiplied by a vector $\bar{V}(t) \in \mathbb{R}^N$, where the i^{th} element is $\frac{V(i,t)}{\sum_{j=1}^{N} V(j,t)}$

Table 1: *Input Features*

IV. Methods

i. Gaussian Mixture

Gaussian mixture is the model that we will use extensively throughout this project to cluster market regimes, and here we will roughly go through the underlying mathematics, under the context of this project, behind Gaussian mixture as well as the EM algorithm that is used to perform the unsupervised learning process.

Suppose there is an observed random variable X, the feature of the market regime (1 out of the 4 features as presented in Table 1), and a latent discrete random variable Z, the numbering of the random market regime in our case, which is in the form of integer labels such as 0, 1, 2 etc. Then the data for day t can be written as $(x(t), z(t))$, and the probability for any given day t can be formulated as:

$$P(x(t), z(t)) = P(x(t)|z(t))P(z(t)) \tag{1}$$

$$where \quad z(t) \sim Multinomial(\Phi) \tag{2}$$

$$x(t)|z(t) = i \sim N(\mu_i, \Sigma_i) \tag{3}$$

We assume the market regime follows a multinomial distribution with a probability vector $\Phi \in \mathbb{R}^n$, where n is the number of possible market regimes. We also assume that for any given market regime, the feature vector $x(t)$ is normally distributed with a mean vector μ_i and a covariance matrix Σ_i. These assumptions arise naturally when the stock returns follow a normal distribution, because all the proposed features are more or less a linear combinations of stock returns.

Now if we do know the market regime $z(t)$ that each daily feature $x(t)$ is assigned to, we can then estimate the parameters of μ_i and Σ_i using maximum likelihood estimation, where the loss function can be written as:

$$l(\Phi, \mu, \Sigma) = \sum_{t=1}^{T} log\ p(x(t), z(t); \Phi, \mu, \Sigma) \tag{4}$$

and the maximum likelihood estimators are then as follows:

$$\Phi_i = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{z(t) = i\} \tag{5}$$

$$\mu_i = \frac{\sum_{t=1}^{T} \mathbb{1}\{z(t) = i\} x(t)}{\sum_{t=1}^{T} \mathbb{1}\{z(t) = i\}} \tag{6}$$

$$\Sigma_i = \frac{\sum_{t=1}^{T} \mathbb{1}\{z(t) = i\}(x(t) - \mu_i)(x(t) - \mu_i)^T}{\sum_{t=1}^{T} \mathbb{1}\{z(t) = i\}} \tag{7}$$

However, now we do not know the exact assignment of clusters for each daily feature, and so Expectation-Maximization algorithm is then used iteratively to find the optimal cluster assignment as well as the location and distribution of the clusters. The Expectation step (E-step) of the EM algorithm is as follows:

$$\text{E-step:} \quad w_i(t) = P(z(t) = i | x(t); \Phi, \mu, \Sigma) \tag{8}$$

$$= \frac{P(x(t)|z(t) = i) P(z(t) = i)}{\sum_{j=1}^{k} P(x(t)|z(t) = j) P(z(t) = j)} \tag{9}$$

where $w_i(t)$ is the probability for the feature data point $x(t)$ to be in the market regime i, and this probability is re-written using the Bayes rule. $P(x(t)|z(t) = i)$ is the probability density function for a normal random variable that has a mean of μ_i and a covariance matrix of Σ_i. $P(z(t) = i)$ is the probability density of a multinomial random variable with value Φ_i.

The Maximization step (M-step) then follows the E-step for each iteration to update the estimated parameters:

$$\text{M-step:} \quad \Phi_i = \frac{1}{T}\sum_{t=1}^{T} w_i(t) \tag{10}$$

$$\mu_i = \frac{\sum_{t=1}^{T} w_i(t) x(t)}{\sum_{t=1}^{T} w_i(t)} \tag{11}$$

$$\Sigma_i = \frac{\sum_{t=1}^{T} w_i(t)(x(t) - \mu_i)(x(t) - \mu_i)^T}{\sum_{t=1}^{T} w_i(t)} \tag{12}$$

Note that before we start the EM algorithm, the parameters of Φ_i, μ_i, and Σ_i are randomly initialized. This EM algorithm is guaranteed to converge but the detailed proof is not shown here for simplicity. Also, the time horizon T that has been frequently used in the above formulations still remains to be decided. From Equations 10 to 12, it is clear that the distribution for each cluster is normal, which in turn provides more complex nonlinear decision boundaries, compared to that of K-means' linear decision boundaries.

ii. Parameter Estimation

By using the Gaussian mixture with EM algorithm as described in Section IV i, the assignments of market regimes to each day in the look-back window can be found. In order to use the market regimes to predict the future returns and covariance matrices of the stocks, we will first need to predict the one step ahead market regimes. Here we propose the use of autoregressive integrated moving average (ARIMA) model to model the historical market regimes as a time series data. In an $ARIMA(p, d, q)$ model, for a time series of $\{z(t)|t = 1, ..., T\}$, $\Delta^d z(t)$ (expressed as $y(t)$ below), the time series of $z(t)$ after performing a total of d differencing's, can be expressed as an $ARMA(p, q)$ process:

$$y(t) - \mu = \sum_{i=1}^{p} \phi_i (y(t-i) - \mu) + \sum_{j=0}^{q} \theta_i \epsilon(t-i) \tag{13}$$

where μ is the long term mean of $y(t)$, $\epsilon(t)$ is the white noise at time t, and $\phi(t)'s$ and $\theta(t)'s$ are the coefficients to be calculated using maximum likelihood estimation. Note here that the forecasts made by the ARIMA model are floats instead of integer values for the market regime, and so we will round these floats to the nearest integers so they can represent valid market regimes. In some extremely rare cases, the ARIMA model will predict a market regime that has never appeared or does not exist (e.g. predicting a market regime of 4 when there are only 3 available regimes). This is caused by the market regimes being too random, which is rarely the case and most of the predictions made by ARIMA are the same as the current regime. For the sake of completeness, if such scenario occurs, we will use the most frequent market regime in the look-back window as the predicted regime.

The Akaike Information Criterion (AIC) is used here to measure the quality of the ARIMA model in order to find the optimal set of $\{p, d, q\}$. AIC is calculated by:

$$AIC = -2log\{L(\hat{\theta})\} + 2dim(\theta) \tag{14}$$

where $L(\theta)$ is the likelihood for a model with the set of parameters θ (not to be confused with the θ parameters in the ARIMA model as in Equation 13), and $\hat{\theta}$ is the maximum likelihood estimator of the parameters. AIC balances the likelihood as well as the complexity of the model.

Now after the one step ahead market regime is predicted, the predicted market regime is used to estimate the future returns of covariances. Here we will assume that the returns of covariance matrices of the stocks will persist and remain the same for any particular market regime, and these parameters will be fed into the subsequent optimization models to generate the optimal set of portfolio weights.

iii. Portfolio Optimization

As the main focus of this project is to study the dynamics of market regimes and the corresponding return and covariance profile of the stocks, the portfolio optimization model is made to be simple and standard. Here we will use the classical Markowitz portfolio optimization to find the optimal weights. Classical Markowitz optimization solves the optimization problem of:

$$\begin{aligned} &\textit{maximize } \mu^T w - \gamma w^T \Sigma w \\ &\textit{subject to } \mathbf{1}^T w = 1, w \in W \end{aligned}$$

where $w \in \mathbb{R}^N$ is the 200-dimensional optimization variable. W is the set of allowed portfolios, and we will use $W = \mathbb{R}_+^N$, which serves as a non-negativity constraint for our portfolio. γ is a user-defined risk aversion factor, and can be varied for different risk-adjusted returns. During our implementation, we will enumerate over the γ values of 0.01, 0.04, 0.13, 0.46, 1.67, 5.99, 21.5, 77.4, 278.3, and 1000 to get the portfolio that has the highest Sharpe ratio. These values are distributed evenly on a log scale. This step is necessary because maximizing Sharpe ratio can sometimes be non-convex and the algorithm can fail to converge.

V. Experiments

i. Evaluation Metrics

The experiment that is conducted in this project is an empirical analysis, or a backtest, of the combined performance of Gaussian mixture clustering algorithm and the trading/rebalancing strategy that follows. The analyses are conducted separately on the historical price data on the 200 constituent stocks of the S&P500 and FTSE350 stock indices. The evaluation metrics that we use are net asset value, yearly Sharpe ratio, and yearly Calmar ratio. The formulae for the yearly Sharpe and Calmar ratios are:

$$SR(t) = \frac{R_p(t)}{\sigma_p(t)} \tag{15}$$

$$CR(t) = \frac{R_p(t)}{MDD_p(t)} \tag{16}$$

$$\textit{where } MDD_p(t) = \max_{i,j \in [t,t+1],\ j>i} \left\{ \frac{NAV(i) - NAV(j)}{NAV(i)} \right\} \tag{17}$$

Note here that yearly Sharpe and Calmar ratios are treated as time series instead of a single number for the entire backtesting period. The performances of the trading strategies are evaluated both for the entire period as a whole and at major stock market crises to draw insights on the efficacy of clustering market regimes. The time spans for these crises are listed in Table 2 below.

Crisis	Start Date	End Date
Dotcom Bubble	Mar. 10, 2000	Oct. 4, 2002
Subprime Mortgage Crisis	Aug. 1, 2007	Jul. 1, 2009
2020 Stock Market Crash	Feb. 12, 2020	Aug. 18, 2020

Table 2: *Crises Dates*

ii. General Backtesting Framework

This project is framed similar to a machine learning project where we have a set of potential hyperparameters, and the data are broken into training set and test set. The hyperparameters, or the configurations, of the backtesting framework are the number of market regimes, the feature that characterizes the regime, the length of the look-back window for Gaussian mixture clustering (or the length of the training set), and the lengths of the holding period. The words hyperparameter and configuration are used interchangeably throughout the rest of the report. All the possible values to be tested for each of the aforementioned hyperparameters are listed in Table 3 below:

Hyperparameter	Possible Values
Number of Regimes	2, 3, 4, 5
Feature	VW Returns, VW EMA Returns, VW MACD, VW RSI
Training Set Length (years)	0.5, 1, 2, 3, 4
Holding Period (days)	1, 5, 21, 63, 252

Table 3: *Possible Configurations*

All these hyperparameters will form a total of 400 possibilities, but some of them are filtered out by evaluating the persistence of generated market regimes as will be later discussed in Section V iii, and a more refined set of hyperparamters are listed in a later section.

Suppose we choose the number of regimes to be n, the feature to be X, the training set length to be T, and the holding period to be t. And now the backtesting procedure is as follows:

1. Perform the EM algorithm on Gaussian mixtures with n clusters for the last T years to get the assignments of each trading day to its corresponding market regime.

2. Run these assignments with an ARIMA model that maximizes the AIC, and obtain the ARIMA's one step ahead forecast.

3. With the forecasted market regime, say i, look back at the training set and calculate the return (μ_i) and covariance matrix (Σ_i) of the 200 stocks on the days where the market is assigned to regime i. (The notations of μ_i and Σ_i are not to be confused with the parameters in the EM algorithm described in Equation 10 to 12.)

4. Using the calculated μ_i and Σ_i, perform a Markowitz portfolio optimization and obtain the optimal portfolio weights.

5. Rebalance the current portfolio with the generated new weights, and hold them for the next t days of the holding period.

6. Once the holding period ends, the above steps are repeated.

The following Figure 1 gives an illustration of our backtesting framework. The training set shown in white is used to find clusters, predict the future market regime, and estimate parameters, and the testing set shown in red is used to realize the profit/loss of the portfolio weights generated by the optimization model. And this procedure is repeated as a sliding window.

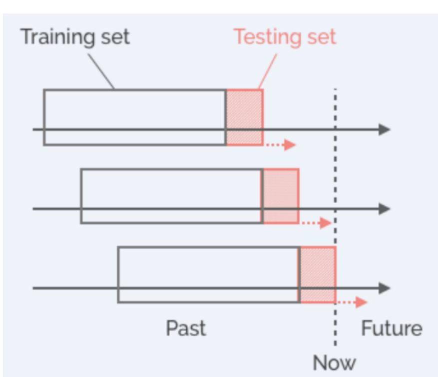

Figure 1: *Backtesting with Sliding Window*

With the described rebalancing trading strategy above, a net asset value curve can be obtained, and performance indicators like Sharpe ratio and Calmar ratio can be calculated using the NAV curve.

iii. Market Regime Persistence Evaluation

Before the actual backtesting described in the above section begins, the persistence of market regimes are first evaluated in order to get a general sense of the behavior of the clustering algorithm with different configurations as in Table 3. This evaluation also serves as a preliminary filtering for the configurations. Here we define the measure of switching score to measure the frequency of regime switching. The switching score is calculated as:

$$P = \frac{1}{T} \sum_{t=2}^{T} \mathbb{1}\{z(t) \neq z(t-1)\} \tag{18}$$

where z(t) is the market regime at time t, and T is the length of the look-back window. The switching score roughly measures the percentage of time during the look-back window where the market is switching, and the inverse of this number can be interpreted as the average length of any particular market regime. Naturally, neither very low switching scores (the market regimes barely switch and provide little useful information) nor very high switching scores (the market regimes switch too often and the switching behavior can mostly be caused by random noise) are desirable.

The formulation presented in Equation 18 only measures the switching score for a particular look-back window (or equivalently a particular training set as in Figure 1). For comparison purposes, we will compute the average of switching scores for all these look-back windows in a backtesting time horizon.

iv. Implementation Details

All the experiments are conducted in the programming language Python. The Gaussian mixture model is fitted and cluster assignments are made using the `GaussianMixture` class in the `Scikit-Learn` package. The $ARIMA(p,d,q)$ model is fitted by the `pmdarima` package, where there is a feature of using `auto_arima` to enumerate some possible values of p, d, and q to find the model that maximizes AIC. The classical Markowitz problem is solved by using the `cvxpy` package. However, while using `cvxpy`, certain solvers can encounter errors when they fail to converge in some occasions, and to solve this issue during the backtesting process, we have tried the solvers of `ECOS`, `ECOS_BB`, `OSQP`, and `SCIPY`. If all these solvers still fail to generate a converged solution, an equally weighted portfolio is used as a placeholder for the backtesting. By randomly choosing certain scenarios, all these aforementioned solvers give almost identical solutions to any optimization problems, and it is also extremely rare that all these solvers all together fail to give a converging solution to a particular problem.

VI. Results & Discussion

i. Market Regime Persistence

The first set of experiments that are conducted are the evaluation of market regime persistence because this evaluation serves as a pilot run and helps reduce the number of all possible strategies that we will use for backtesting. By doing so the required total computation time for the project is greatly reduced. All the combinations of the hyperparameters listed in Table 3 are tested by evaluating their corresponding average market state persistences. Because there are 400 total possibilities, we will only focus on the average values of the switching scores for each value of a hyperparameter. For example, in order to evaluate whether 2 market regimes is appropriate, we will fix the number of regimes to be 2 and average over all the possibilities where the number of regimes is 2.

# of Regimes	S&P 500	FTSE 350
2	0.069	0.057
3	0.121	0.098
4	0.158	0.128
5	0.188	0.152

Table 4: *Average Switching Score for # of Regimes*

From Table 4 we can see that by using only 2 regimes, the average persistence is the highest and the persistence gradually weakens when the number of regimes increases. This can be a result of data points being assigned to an increased number of clusters and the markets tend to switch more often when there are more regimes to switch to. However, this observation is not anomalous and all the possible numbers of regimes are kept.

Feature	S&P 500	FTSE 350
VW Returns	0.173	0.115
VW EMA Returns	0.035	0.008
VW MACD	0.042	0.063
VW RSI	0.286	0.249

Table 5: *Average Switching Score for Features*

From Table 5, we can see that the features of volume-weighted exponential moving average returns and volume-weighted MACD result in very low switching scores, with VW EMA returns having a switching score of as low as 0.008. These low scores can be concerning because they indicate the phenomenon that the market regimes are barely switching, and the features do not capture enough information about the variations in the market. Such phenomenon is also consistent with the nature of exponential moving average and MACD because they are greatly smoothed out average of times series, and do not react to market shifts in a timely manner. Therefore, we will exclude the hyperparamters of VW EMA Returns and VW MACD for experiments in backtesting.

Holding Period (days)	S&P 500	FTSE 350
1	0.133	0.110
5	0.133	0.110
21	0.134	0.109
63	0.134	0.109
252	0.135	0.106

Table 6: *Average Switching Score for Holding Periods*

Table 6 shows that all the possible holding periods give similar and moderate switching scores - around 0.13 for stocks in S&P 500 and around 0.11 for stocks in FTSE 350, and none of them are showing anomalous behaviors, so all these frequencies are kept for backtesting.

Training Length (years)	S&P 500	FTSE 350
0.5	0.156	0.116
1	0.142	0.108
2	0.132	0.110
3	0.124	0.106
4	0.117	0.104

Table 7: *Average Switching Score for Training Lengths*

Similarly, Table 7 shows the choice of training set length does not impact the average switching scores very much, where the scores are all within the range of 0.10 and 0.16. Therefore all these training lengths are kept.

ii. Final Modification to Hyperparameter Sets

Furthermore, in the process of calculating the switching scores for each possible set of hyperparameters, we have found that the configurations with rebalancing frequencies of 1 or 5 days and those with 5 market regimes incurs high computational costs and are overly time-consuming their the backtest. Also, these rebalancing frequencies are too often too high for long-term investments, and also 5 market regimes can be difficult to interpret. Therefore, these configurations are also excluded going forward, and the remaining configurations are shown in Table 8 below:

Hyperparameter	Possible Values
Number of Regimes	2, 3, 4
Features	VW Returns, VW RSI
Training Set Length (years)	0.5, 1, 2, 3, 4
Holding Period (days)	21, 63, 252

Table 8: *Remaining Configurations*

Additionally, it is naturally preferable to have a test set shorter than the training set because intuitively, the insights generated from a short period is not likely to hold true for a longer future investment period. And so we also exclude the possibilities where the training set length is shorter than the holding period. Now, the number of possible hyperparameter sets comes down to a total of 84.

iii. Overall Backtesting Results

Because our project is experimenting on all 84 possibilities for each market, we have generated a massive amount of results where it is unwieldy to show all the individual results. Hence, we will only highlight the most interesting findings from the experimental results. In this report, we will only study the strategies whose net asset values throughout the entire backtesting period have outperformed the equally weighted portfolio. By using this metric, a total of 29 out of the 84 configurations have outperformed the equally weighted portfolio for stocks in the S&P 500 index. However, in the case of stocks in the FTSE 350 index, a total of 64 out of the 84 configurations have outperformed overall.

The 10 top performing configurations for S&P 500 along with their return outperformance, Sharpe outperformance ratio, and Calmar outperformance ratio are shown in Table 9 and 10.

Rank	# of MRs	Feature	TSL	HP
1	3	VW Return	0.5	63
2	3	VW RSI	0.5	63
3	3	VW Returns	0.5	21
4	2	VW RSI	0.5	63
5	3	VW RSI	2	252
6	2	VW Returns	0.5	63
7	2	VW Returns	0.5	21
8	4	VW RSI	0.5	63
9	3	VW RSI	2	63
10	2	VW RSI	2	252

MR: Market Regime; TSL: Training Set Length; HP: Holding Period

Table 9: *Outperforming Configurations for S&P 500*

Rank	R OP	Sharpe OPR	Calmar OPR
1	18.57%	53.57%	46.43%
2	15.97%	64.29%	57.14%
3	15.74%	50.00%	42.86%
4	15.11%	53.57%	53.57%
5	19.00%	57.69%	61.54%
6	11.50%	67.86%	39.29%
7	11.00%	53.57%	57.14%
8	10.38%	46.43%	39.29%
9	8.06%	46.15%	50.00%
10	8.05%	57.69%	53.85%

OP: Outperformance; OPR: Out-Performance Ratio

Table 10: *Outperformances of S&P 500 Configurations*

The return outperformances are percentages describing how much the strategy has outperformed the equally weighted portfolio on average during the backtesting period. The Sharpe outperformance ratio is the percentage of years when the strategy's Sharpe ratio has exceeded that of equally weighted portfolio's Sharpe ratio, and similarly for the Calmar outperformance ratio. Note here that sometimes the return outperformances (R OP) do not always align with the rank (same for the results in Table 12) because due to the differences in the training set length, meaning the backtesting horizon is not the same for all the possible strategies. Some strategies have higher net asset values simply because their investment horizons are longer. However, such occurrences do not impact our analysis because the top performing 10 portfolios are the same regardless.

From Table 9, we can see that the 4 top performing strategies all use 3 market regimes to cluster the data, and only one strategy out of the 10 top performing ones has used 4 as the number of market regimes. Also, although some top performing strategies use 2 years as the training set length, 7 out of the 10 have used half a year as the training set length, strongly indicating that fitting to the most recent data, in our case half a year instead of longer periods, can provide more useful information about the assignments of market regimes. This can also indirectly support our choice using the historical return and covariance matrices for stocks as the predicted parameters because more accurate market regime assignments have led to more profitable portfolios, meaning the Markovitz optimization solutions indeed provide solutions that maximize the risk-adjusted returns. Moreover, by looking at the Sharpe and Calmar outperformance ratios, it is evident that during about 50% of the time, the top 10 performing strategies have better risk-adjusted returns than the equally weighted portfolio. Although this ratio merely breaks even, the overall net asset value can greatly outperform the benchmark.

The following Figure 3 shows the net asset values for the

10 top performing portfolios, with the equally weighted portfolio as the benchmark plotted as the dashed line. Note here that although some strategies have reached net asset values that well exceed 150 in the 26-year time span, the way we pick our stocks are inherent biased as mentioned earlier in Section III i. One should not expect to replicate these profits going forward into the future using our methods. We only use these net asset values as relative comparisons to the equally weighted portfolios to comment on the efficacy of the hyperparameters used for our market regime clustering algorithm.

Rank	# of MRs	Feature	TSL	HP
1	4	VW RSI	1	21
2	4	VW RSI	1	252
3	2	VW RSI	0.5	21
4	4	VW RSI	0.5	63
5	4	VW RSI	3	21
6	2	VW Returns	0.5	21
7	3	VW RSI	1	252
8	4	VW RSI	0.5	21
9	3	VW RSI	0.5	63
10	2	VW Returns	4	63

MR: Market Regime; TSL: Training Set Length; HP: Holding Period

Table 11: *Outperforming Configurations for FTSE 350*

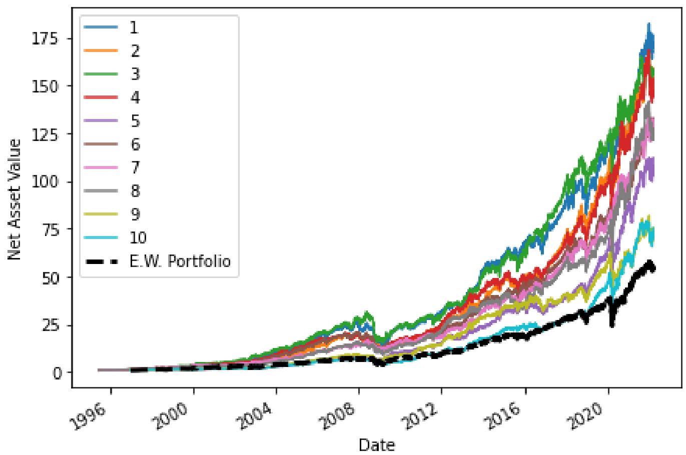

Figure 2: *Top Performing Portfolios for S&P 500*

Rank	R OP	Sharpe OPR	Calmar OPR
1	154.55%	62.50%	62.50%
2	94.04%	58.82%	64.71%
3	86.78%	70.59%	64.71%
4	48.48%	50.00%	50.00%
5	84.40%	70.59%	70.59%
6	123.14%	81.25%	87.50%
7	77.20%	64.71%	64.71%
8	70.02%	64.71%	58.82%
9	64.85%	69.23%	69.23%
10	32.44%	71.43%	71.43%

OP: Out-Performance; OPR: Out-Performance Ratio

Table 12: *Outperformances of FTSE 350 Configurations*

According to this line chart of net asset values, the strategies performances are quite correlated with the benchmark, especially during times of the Subprime Mortgage Crisis around 2008 and the 2020 Stock Market Crash. At a first glance, we might make the observation that the performances of our developed strategies are magnified versions of the equally weighted benchmark, where our strategies lose higher values during the times of crises and grow much faster during stock market recoveries and booms. A much detailed analysis on the crises will be made later in the next subsection.

The 10 top performing configurations for stocks in the FTSE 350 index along with their evaluation metrics are shown in Table 11 and 12.

Here, there is no clear pattern on the more preferable number of market regimes to use for clustering because the number of 2, 3, and 4 have all appeared in the top 10 performing portfolio configurations equally frequent. However, we do see that the volume-weighted relative strength index is a more desirable feature to use for clustering for the FTSE 350 stocks as 8 out of the 10 top performing portfolio use VW RSI as the input feature. On the other hand, similar to the S&P 500 stocks, half a year is still the most preferable training set length because half of the 10 top performing portfolios use half a year to train the clusters. This finding also agrees with the claim made earlier that fitting to the most recent data can be more informative about the assignments of market regimes. The results are showing indifference about the lengths of the holding period.

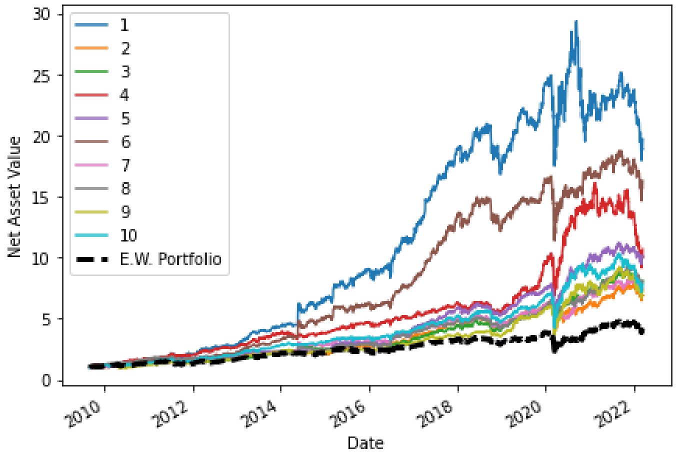

Figure 3: *Top Performing Portfolios for S&P 500*

The line chart for the strategies on FTSE 350 stocks shows similar behaviors where the net asset values of the developed strategies are seemingly magnified versions of the equally weighted portfolio.

iv. Backtesting Results during Crises

As market regimes are often used to identify the phases of stock markets to avoid catastrophic loss during crises, much effort of this project will focus on the behavior of our strategies during the three stock market crises of the Dotcom Bubble, the Subprime Mortgage Crisis, and the 2020 Stock Market Crash. Because of the limited access to the FTSE data, only the most recent crisis of the 2020 Stock Market Crash is studied on the European stock market. This subsection is broken down into three parts, where every part focuses on one market crisis. More emphasis is given to the evaluation metrics of Sharpe ratio and Calmar ratio when we study crises.

iv.1 Dotcom Bubble

The Dotcom Bubble was created by a rapid rise in U.S. technology stock equity valuations fueled by investments in Internet-based companies in the late 1990s, which later burst and dragged the stock market into a bear market when the speculative investing and venture capital fundings failed to make a profit. The time span for the Dotcom Bubble used in this project is from Mar. 10, 2000 to Oct. 4, 2002.

The 10 top performing strategy configurations for S&P 500 in terms of Sharpe ratio are presented in Table 13 below:

Rank	# of MRs	Feature	TSL	HP	SR
1	3	VW Returns	0.5	63	4.41
2	2	VW Returns	1	63	3.92
3	3	VW Returns	1	63	3.55
4	4	VW RSI	4	63	3.46
5	3	VW RSI	1	63	3.37
6	3	VW Returns	1	21	3.31
7	3	VW RSI	4	252	3.31
8	2	VW RSI	0.5	63	3.21
9	3	VW RSI	0.5	21	3.19
10	2	VW RSI	1	63	3.16
Equally Weighted Portfolio					1.34

MR: Market Regime; TSL: Training Set Length; HP: Holding Period; SR: Sharpe Ratio

Table 13: *Outperforming Configurations at Dotcom Bubble (SR)*

Out of the top 10 performing configurations, the training set length of 1 year and a holding period of 63 days (quarterly rebalancing) appear dominant, whereas the input features have shown no clear pattern. Additionally, it is interesting to note that only 1 out of the 10 configurations uses 4 as the number market regimes, where the rest are 2 and 3 with equal occurrences. The net asset value curves for these strategy configurations are plotted in Figure 4 where the equally weighted portfolio is drawn with a bold dashed line.

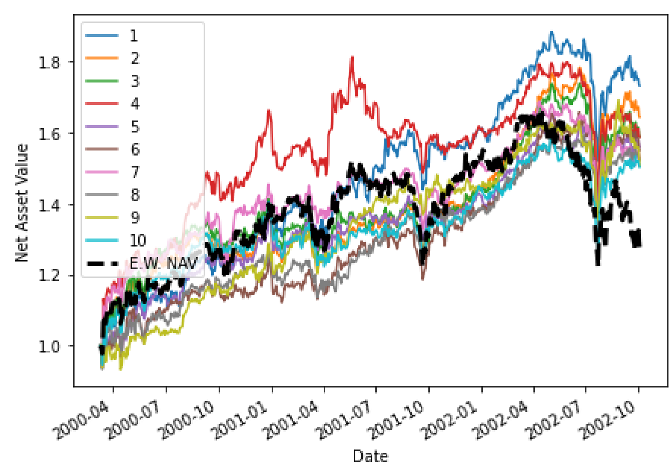

Figure 4: *NAVs for Highest-Sharpe Configurations*

For the portfolios that have higher Sharpe ratio than the equally weighted portfolio, they either have much higher returns at some time, or much lower volatilities overall. In general, these 10 portfolios, while compared to the equally weighted portfolio, have much smoother net asset value curves, especially during the dips. For the two dips that happened during October 2001 and mid 2002, our strategies have significantly prevent these major dips compared to the equally weighted portfolio.

The 10 top performing strategy configurations for S&P 500 in terms of Calmar ratio are presented in Table 14 below:

Rank	# of MRs	Feature	TSL	HP	CR
1	3	VW Return	0.5	63	3.79
2	2	VW RSI	0.5	63	3.37
3	2	VW Return	1	63	3.22
4	3	VW RSI	1	63	3.13
5	2	VW Return	1	21	2.97
6	3	VW RSI	0.5	21	2.96
7	3	VW RSI	1	252	2.90
8	3	VW Return	1	21	2.89
9	3	VW Return	0.5	21	2.85
10	3	VW Return	1	63	2.81
Equally Weighted Portfolio					1.05

MR: Market Regime; TSL: Training Set Length; HP: Holding Period; CR: Calmar Ratio

Table 14: *Outperforming Configurations at Dotcom Bubble (CR)*

Here we observe that the 10 top performing portfolios in terms of Calmar ratio only use 2 or 3 as the number of market regimes, and it is evident that for both Sharpe and Calmar ratio during the Dotcom crisis, 2 and 3 market regimes are more desirable. The training set lengths of 0.5 and 1 year appear more desirable and the holding periods of 63 days (quarterly rebalancing) and 21 days (monthly rebalancing) are more dominant. Note here that the holding period of 252 (yearly rebalancing) only appeared once out of the 10, and it is intuitive that more frequent rebalancings are needed when the market enters into a crisis. And here again we do not see clear pattern on the more desirable features to choose during the Dotcom Bubble. The net asset value curves for these strategy configurations are plotted in Figure 5 with the equally weighted portfolio as a benchmark.

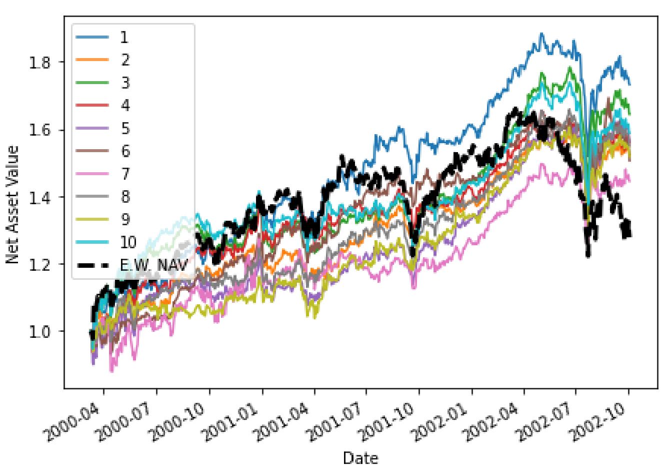

Figure 5: *NAVs for Highest-Calmar Configurations*

Here we see that all these curves are more smoothed out because the calmar ratio puts more emphasis on minimizing the maximum drawdown of net asset values. We see that all the net asset value curves have smaller drawdowns during the 2 major dips and a minor one by the end of October 2002.

To sum up the analysis for the Dotcom Bubble, in order to obtain the highest risk-adjusted returns and control drawdowns, it is more preferable to use 2 or 3 market regimes for clustering, to use one or half a year as the look-back window for training clusters, and to have a monthly or quarterly rebalancing frequency. The choice of feature for clustering market regimes does not make much of a difference in terms of correctly assigning data points to regimes.

iv.2 Subprime Mortgage Crisis

The Subprime Mortgage Crisis was a result of housing bubble burst that led to mortgage delinquencies and major devaluation of housing-related securities. In some sense this crisis is similar to the Dotcom Bubble where the shattering of economic illusions caused panics and negative sentiments in stock markets. The time span for the Subprime Mortgage Crisis used in this project are from Aug. 1, 2007 to Jul. 1, 2009.

Here only 8 strategy configurations have outperformed the equally weighted portfolio for S&P 500 in terms of Sharpe ratio during the Subprime Mortgage Crisis, and they are all presented in Table 15 below:

Rank	# of MRs	Feature	TSL	HP	SR
1	3	VW RSI	2	252	0.53
2	2	VW Returns	0.5	21	0.11
3	4	VW RSI	4	252	-0.15
4	3	VW RSI	2	63	0.24
5	4	VW RSI	4	63	-0.29
6	4	VW RSI	3	63	-0.39
7	4	VW Returns	4	63	-0.41
8	3	VW Returns	0.5	63	-0.41
Equally Weighted Portfolio					-0.47

MR: Market Regime; TSL: Training Set Length; HP: Holding Period; SR: Sharpe Ratio

Table 15: *Outperforming Configurations at Subprime Mortgage Crisis (SR)*

Out of the 8 outperforming configurations, we see that smaller number of market regimes tend to generate better clustering assignments and thus higher Sharpe ratios. There are no clear evidence on the optimal training set length and holding period to use. Similar to the Dotcom Bubble, the choices of features for clustering do not make much of a difference here as well. However, it is noteworthy that 2 of the 3 top performing strategy configurations use a holding period of 252 days (yearly rebalancing), indicating that keeping the portfolio turnover relatively low can sometimes help prevent great losses during crisis because the insights learned from the price data during the crisis can be misleading. The net asset value curves are plotted in Figure 6 below:

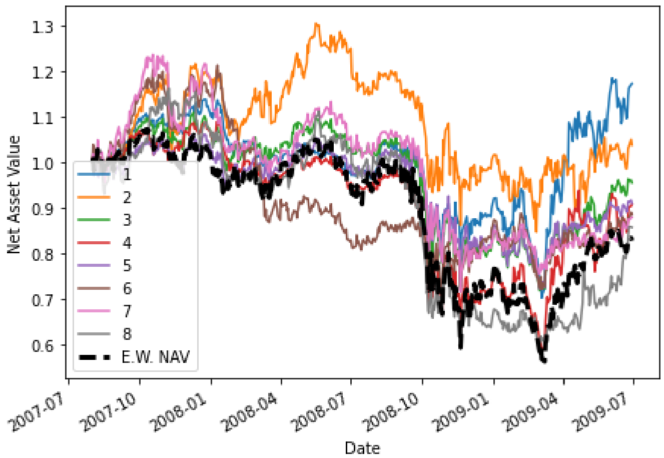

Figure 6: *NAVs for Highest-Sharpe Configurations*

The top performing portfolio, plotted in blue, has stayed relatively smooth during the entire period of the crisis, although it also experienced the dip, but to a much lesser extent.

Similarly, only 8 strategy configurations have outperformed the benchmark for S&P 500 in terms of Calmar ratio, and are listed in Table 16 below:

Rank	# of MRs	Feature	TSL	HP	CR
1	3	VW RSI	2	252	0.50
2	2	VW Returns	0.5	21	0.11
3	4	VW RSI	4	252	-0.13
4	3	VW RSI	2	63	-0.20
5	4	VW RSI	4	63	-0.28
6	4	VW RSI	3	63	-0.28
7	4	VW Returns	4	63	-0.29
8	3	VW Returns	0.5	63	-0.29

MR: Market Regime; TSL: Training Set Length; HP: Holding Period; CR: Calmar Ratio

Table 16: *Outperforming Configurations at Subprime Mortgage Crisis (CR)*

Here we note that these are the exact same strategies presented in Table 15, and so the net asset values are not plotted because they are same as the ones in Figure 6, even for the ranks.

To sum up the analysis on Subprime Mortgage Crisis, it is generally difficult to outperform the benchmark for this particular crisis, although such difficulty can be partly caused by the non-negativity constraint of our optimization problems that forces our portfolio to be in a 100% long position of stocks. Also, not many insights on how the strategy configurations should be chosen are observed. The only useful observation we have drawn from the results is that frequent rebalancings might not generate better portfolio, especially when the most recent volatile historical price data are results of irrational behaviors of market players.

iv.3 2020 Stock Market Crash

The 2020 Stock Market Crash was a short-lived crisis caused by global fears of the spread of coronavirus and drops in oil prices, and it has caused the most significant point plunge for the Dow Jones Industrial Average historically. Here we will study the performance of our strategies for both the S&P 500 and FTSE 350 indices.

For the S&P 500 markets, the 10 top performing strategy configurations in terms of Sharpe ratio are shown in Table 17:

Rank	# of MRs	Feature	TSL	HP	SR
1	4	VW RSI	0.5	21	0.70
2	4	VW RSI	0.5	63	0.62
3	2	VW Returns	0.5	21	0.55
4	2	VW RSI	0.5	21	0.51
5	2	VW RSI	0.5	63	0.51
6	2	VW Returns	1	21	0.43
7	4	VW RSI	2	21	0.42
8	3	VW RSI	0.5	21	0.40
9	4	VW RSI	4	21	0.39
10	3	VW Returns	3	63	0.36
Equally Weighted Portfolio					-0.026

MR: Market Regime; TSL: Training Set Length; HP: Holding Period; SR: Sharpe Ratio

Table 17: *Outperforming Configurations at 2020 Stock Market Crash for S&P 500 (SR)*

Here 0.5 year is the most preferable training set length because the 5 best performing portfolios all have such the training length. It also strongly shows that monthly and quarterly are the more preferable rebalancing frequencies. Here no clear pattern on which choice of number of market regimes or feature is obverved. The net asset value curves are plotted in Figure 7 below:

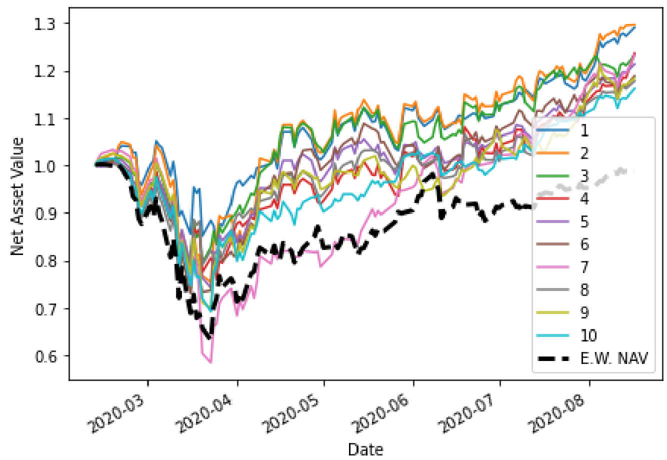

Figure 7: *NAVs for Highest-Sharpe Configurations for S&P 500*

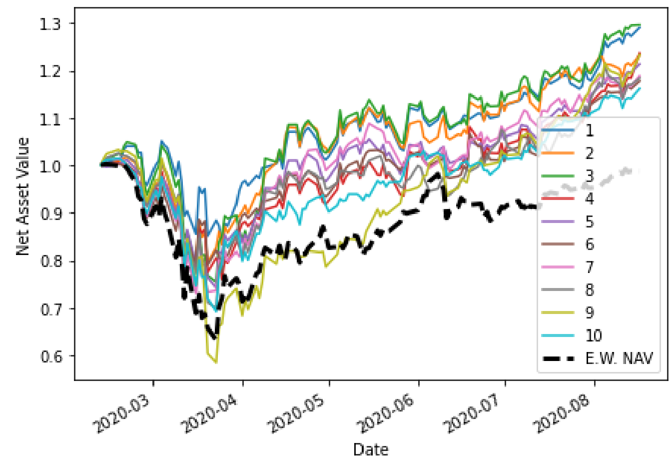

Figure 8: *NAVs for Highest-Calmar Configurations for S&P 500*

The 10 top performing strategy configurations for U.S. stock market in terms of Calmar ratio are shown in Table 18 and their corresponding net asset value curves are plotted in Figure 8:

Rank	# of MRs	Feature	TSL	HP	CR
1	4	VW RSI	0.5	21	1.40
2	2	VW Returns	0.5	21	1.06
3	4	VW RSI	0.5	63	1.05
4	2	VW RSI	0.5	21	0.96
5	2	VW RSI	0.5	63	0.80
6	3	VW RSI	0.5	21	0.77
7	2	VW Returns	1	21	0.68
8	4	VW RSI	4	21	0.58
9	4	VW RSI	2	21	0.53
10	3	VW Returns	3	63	0.51
Equally Weighted Portfolio					-0.034

MR: Market Regime; TSL: Training Set Length; HP: Holding Period; CR: Calmar Ratio

Table 18: *Outperforming Configurations at 2020 Stock Market Crash for S&P 500 (CR)*

From both Figure 7 and 8, we can see that all the portfolios have suffered a huge drawdown during late March, and none of these portfolios are better off than the others. However, when the market starts to bounce back right after the sharp decline, these better performing portfolios have all had some rapid yet steady growth. The configurations have again suggested the appropriateness of using half a year as the look-back window for clustering and using monthly or quarterly rebalancing frequencies.

Now we turn our attention to the European market, which is also affected by the threat posed by the coronavirus pandemic. For stocks in FTSE 350, the 10 top performing strategy configurations in terms of Sharpe ratio along with their net asset value curves are shown in Table 19 and Figure 9 below:

Rank	# of MRs	Feature	TSL	HP	SR
1	4	VW RSI	3	63	0.92
2	3	VW RSI	3	63	0.85
3	4	VW RSI	3	21	0.70
4	3	VW RSI	4	63	0.66
5	4	VW RSI	3	252	0.62
6	2	VW RSI	3	63	0.56
7	4	VW Returns	3	21	0.49
8	4	VW RSI	1	21	0.49
9	3	VW RSI	3	21	0.46
10	4	VW RSI	4	63	0.46
Equally Weighted Portfolio					-0.34

MR: Market Regime; TSL: Training Set Length; HP: Holding Period; SR: Sharpe Ratio

Table 19: *Outperforming Configurations at 2020 Stock Market Crash for FTSE 350 (SR)*

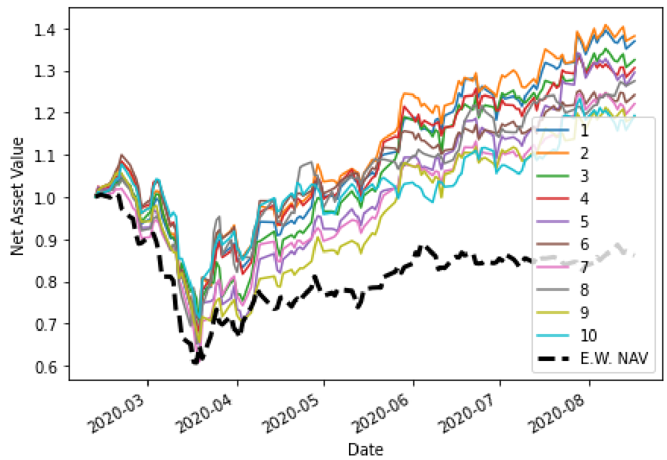

Figure 9: *NAVs for Highest-Sharpe Configurations for FTSE 350*

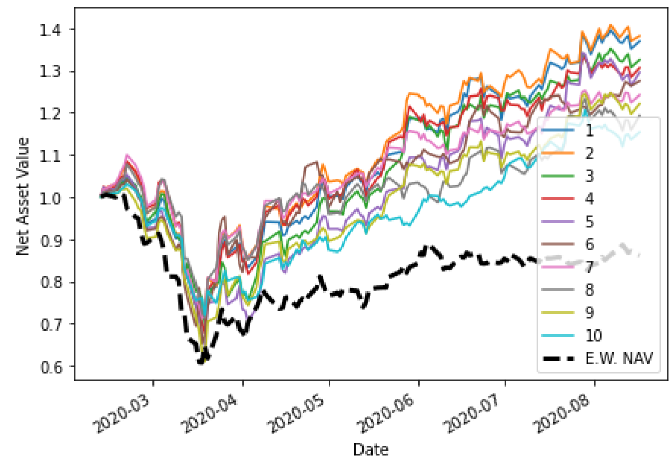

Figure 10: *NAVs for Highest-Calmar Configurations for FTSE 350*

From Table 19, we have observed major differences between the U.S. and European stock market for the performance of our strategies. Here for the first time we have observed that the preferred feature for clustering is the volume-weighted relative strength index, whereas no feature is preferred over the other for the U.S. stock markets. As the relative strength index serves as a momentum indicator of the oversold/overbought levels, this preference for using VW RSI can imply the overvalued/undervalued levels of stocks during crises give better representations of the European stock market compared to returns. Furthermore, 9 out of the 10 top performing configurations used 3 and 4 years as the training set length, and this can mean that the market landscape for the European market does not shift as often as that of the U.S. stock market and longer analyses on the historical data can still be relevant and more informative.

The table and plots for using Calmar ratio as the evaluation metric are presented below:

Rank	# of MRs	Feature	TSL	HP	CR
1	4	VW RSI	3	63	1.14
2	3	VW RSI	3	63	1.12
3	4	VW RSI	3	21	0.98
4	3	VW RSI	4	63	0.82
5	4	VW RSI	3	252	0.77
6	4	VW RSI	1	21	0.74
7	2	VW RSI	3	63	0.69
8	4	VW RSI	4	63	0.57
9	4	VW Returns	3	21	0.55
10	2	VW Returns	0.5	21	0.51
Equally Weighted Portfolio					-0.35

MR: Market Regime; TSL: Training Set Length; HP: Holding Period; CR: Calmar Ratio

Table 20: *Outperforming Configurations at 2020 Stock Market Crash for FTSE 350 (CR)*

Similarly here, most of the top performing portfolios use the volume-weighted relative strength index as the feature and use longer periods of 3 or 4 years as the training set length.

In summary, although in both the U.S. and European stock markets, none of our strategies have prevented the dramatic drawdown happening in late February 2020, especially for the European stock market, some of them have bounced back to a much greater extent compared to the equally weighted portfolio when the market recovers afterwards. Similar to our previous analyses on the U.S. stock market during crises, the best performing strategies on this market also use relatively short look-back window (0.5 years) and the relatively frequent rebalancing (monthly and quarterly). Moreover, the best performing strategies for the European stock market also use monthly or quarterly rebalancings. On the other hand, to perform well in the European stock market, longer (around 3 to 4 years) look-back windows need to be used when training the Gaussian mixture models, and the relatively strength index, compared to returns, is shown to be more explanatory for market regime clustering.

VII. Conclusion

In this paper, we have experimented a method of using Gaussian mixture with Expectation-Maximization algorithm to cluster market regimes, the assignments of which are subsequently used to estimate future return and variances used in Markowitz portfolio optimizations. A wide range of combination of strategy configurations (i.e. the number of market regimes, clustering features, training set lengths, and holding period lengths) are tested in terms of their performances on both the U.S. and European stock markets. The performance metrics used are net asset value, Sharpe ratio, and Calmar ratio, and different strategies are evaluated not only on the entire investment horizon overall (from Nov. 1999 for stocks in S&P 500 and from April 2006 for stocks in FTSE 350), but also during market crises (i.e. Dotcom Bubble, Subprime Mortgage Crisis, and 2020 Stock Market Crash). Our strategies on European stocks are only evaluated during the 2020 Stock Market Crash due to the limited access to European stock data.

To perform well overall (i.e. for the entire investment horizon), we have concluded that the strategies using 3 market regimes along with half a year of training set length tend to perform well for both markets. Besides, the strategy performances are in general indifferent of the choice of clustering features or rebalancing frequencies. With regards to performances during crises, it is preferable to choose relatively short training set lengths (one or half a year), and frequent rebalancings (monthly or quarterly) during all three stock market crises for stocks in S&P 500. It is also worthwhile to note that during the Subprime Mortgage Crisis, strategies with yearly rebalancing have performed unexpectedly well possible due to market information being too random. On the other hand, in order for the strategies to perform well during the 2020 Stock Market Crisis on European stocks, relative strength index is the more preferable clustering feature and longer training set lengths (3 and 4 years) are preferred for training the clusters.

VIII. Appendix

All the data and scripts used for this project are uploaded on Github along with instructions on how to replicate the results, which can be accessed here.

References

Haase, F., Neuenkirch, M. (n.d.). Predictability of bull and bear markets: A new look at forecasting stock market regimes (and returns) in the us. *Research Papers in Economics*.

Munnix, M., Shimada, T., Schafer, R., Leyvraz, F., Seligman, T. H., Guhr, T., Stanley, H. E. (2012). Identifying states of a financial market. *Scientific Reports*, *2*(644). doi: https://doi.org/10.1038/srep00644

Procacci, P. F., Aste, T. (2019, jul). Forecasting market states. *Quantitative Finance*, *19*(9), 1491–1498. Retrieved from `https://doi.org/10.1080%2F14697688.2019.1622313` doi: 10.1080/14697688.2019.1622313

Tang, W., Xu, X., Zhou, X. Y. (2021). *Asset selection via correlation blockmodel clustering.* arXiv. Retrieved from `https://arxiv.org/abs/2103.14506` doi: 10.48550/ARXIV.2103.14506

Wang, Y., Aste, T. (2021). *Dynamic portfolio optimization with inverse covariance clustering.* arXiv. Retrieved from `https://arxiv.org/abs/2112.15499` doi: 10.48550/ARXIV.2112.15499