INTEGRATED ASSET SELECTION AND OPTIMIZATION USING SUPPORT VECTORS AND MEAN VARIANCE OPTIMIZATION

Current Status

Note: I will write this document in an informal style until the full development of the ideas.

September 24th, 2021

This work aims to develop a methodology for portfolio optimization that uses support vector machines within the context of mean-variance optimization. The mean-variance approach pioneered by Markowitz (1952) is a pillar in financial theory. Markowitz (1952) formulates the problem as a tradeoff between portfolio risk and return. However, the resulting portfolios do have their disadvantages. First of all, the most apparent drawback of the mean-variance approach is that the resulting portfolios are extremely sensitive to estimates of their input parameters. Secondly, a practical concern with portfolio management is transaction costs that erode profits over time due to rebalancing the portfolio. As a result, it is common for investors to specify a limit on the total number of securities in the portfolio to combat transaction costs.

Several existing methods combat estimation errors. Michaud and Michaud (2007) uses a data resampling and portfolio averaging approach to reduce the effects of estimation error. Anis and Kwon (2020) and Costa and Kwon (2019) advocate for using optimization models that are not dependent on the estimated mean return since the mean returns are subject to the most estimation error. Butler and Kwon (2021) reduces the impact of estimation error in the mean by learning a predictive model for the mean return such that the mean-variance objective is optimized. The rationale given by Butler and Kwon (2021) is as follows; since no prediction model exhibits perfect performance, is there a parameterization of the model that would lead to improved decision making. Under traditional paradigms, predictive models are estimated by maximum likelihood (or some other criterion differing from the task criterion), whereas in the integrated prediction and optimization paradigm the

Goldfarb and Iyengar (2003) devises several robust formulations of portfolio optimization models where the resulting portfolios are immunized against input parameters that are allowed to vary within uncertainty sets. Furthermore, Blanchet et al. (2018) shows that a distributionally robust mean-variance problem is equivalent to adding a regularization term to the objective function. Before the distributionally robust linkage was discovered, it was well known that adding a regularization term to the objective is another popular method that is shown to reduce the impacts of estimation error Carrasco and Noumon, 2011.

Tillmann et al. (2021) provides a comprehensive review of cardinality constrained optimization techniques. A popular heuristic to find cardinality-constrained cardinality solutions is to penalize the objective function by a norm of the decision variables, which is a particular case of the distributionally robust framework by Blanchet et al. (2018). For example, penalization with an l_1 norm is a commonly used heuristic to find portfolios of a specified cardinality since the l_1 norm promotes sparsity. Most works on cardinality constrained portfolio optimization have focused on heuristic solution methods due to the problem's Np-Hardness; see (Chang et al., 2000) for example. There have been fewer works devoted to solving the optimization problem exactly using a dual form of the problem combined with a branch and bound procedure; see (Shaw et al., 2008) for example.

Although portfolio optimization is popular, some other methodologies and frameworks assist in making portfolio decisions. A promising framework proposed by Fan and Palaniswami (2001) suggests the use of a support vector machine (SVM) to determine a decision rule that decides whether or not to allocate capital to a particular stock. Fan and Palaniswami (2001) proceeds by classifying stocks into two groups: an exceptional return group and a mediocre return group. The stocks exhibiting exceptional returns are defined to be those in the top 25 % of the one-year return distribution. Then, using financial information of each company as features for the stocks, Fan and Palaniswami (2001) trains an SVM on yearly data and evaluates the performance of the support vector out of sample. Finally, the capital is allocated equally among each stock predicted to out-perform, and the corresponding portfolio return is calculated. (Paiva et al., 2019) build on the work by Fan and Palaniswami, 2001 by introducing a portfolio optimization step in the mean-variance approach by training an SVM to predict whether or not a stock will meet the return threshold designated by the investor and then subsequently allocating capital amongst the stocks

predicted to meet the return threshold according to the SVM. The capital allocation step is done by optimizing the mean-variance objective function. (Paiva et al., 2019) suggests that adding an SVM decision step to the portfolio selection process may improve risk-adjusted returns.

However, the work by (Paiva et al., 2019) only learns the SVM based on the pre-specified return preference. There may be a support vector that would include assets in the portfolio, which would improve the out-of-sample mean-variance objective. This work proposes three approaches to arrive at a support vector decision rule that will optimize the mean-variance objective function. The first approach will leverage results from the literature on cardinality constrained portfolio optimization and will incorporate the asset constraints directly into the optimization problem. The second approach will treat the support vector machine as a predictive model that will parameterize the mean-variance objective function such that if a given asset is classified in the negative class, it will not be included in the optimal portfolio. Then similar to Butler and Kwon (2021) and Donti et al. (2017), the realized mean-variance objective function is optimized over the predictive model's parameters (support vector in this case). The third approach also uses task based learning, however, the parameterized (by the support vector) mean-variance problem contains binary 0-1 variables.

1 Background

Support Vector Machines

Given a dataset of n pairs $\{(y_1, u_1), (y_2, u_2), \ldots, (y_n, u_n)\}$, with $y_i \in \mathbb{R}^p$ and $u_i \in \{-1, 1\}$. One can define a classification rule by determining which side a given input lies on a hyperplane (taking the sign of a hyperplane):

$$G(x) = \text{sign}[y^\intercal w + b] \tag{1}$$

Friedman (2017) show that for the given classification rule and under the assumption of linear separability, the hyperplane that creates the most margin between the positive and negative classes

is given by the following optimization problem:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad & u_i(w^\intercal y_i + b)) \geq 1 \quad \forall i = 1, 2, \ldots, n \end{aligned} \tag{SVM-1}$$

Furthermore, in the non-separable case by introducing $\xi_i \geq 0$, and relaxing the constraint $u_i(w^\intercal y_i + b)) \geq 1$ to $u_i(w^\intercal y_i + b)) \geq 1 - \xi_i$. It follows that SVM-1 is equivalent to the following problem:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \\ \text{s.t.} \quad & u_i(w^\intercal y_i + b)) \geq 1 - \xi_i \quad \forall i = 1, 2, \ldots, n \\ & \xi_i \geq 0 \quad \forall i = 1, 2, \ldots, n \end{aligned} \tag{SVM-2}$$

Where C is a parameter that penalizes miss-classification. To make the problem unconstrained ξ_i can be written as the hinge loss $\max(0, 1 - u_i(w^\intercal y_i + b)))$ and SVM-2 is equivalent to

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2 + C\sum_i \max(0, 1 - u_i(w^\intercal y_i + b)))^\intercal x \tag{SVM-3}$$

Menon (2009), argues that the primal is of interest when tackling large-scale problems because its lack of constraints allows it to be tackled by gradient descent algorithms. The dual of SVM-2 has historically been of interest because it lends itself easier to the "kernel trick" since it explicitly makes use of dot products between transformed input vectors [1]. Furthermore, the dual formulation also has simple box constraints for the dual variables. In this work, the primal version of the SVM is considered because it lends itself to gradient descent methods.

Integrated Prediction and Optimization

First, suppose there are N assets available. A portfolio is represented by $x \in \mathbb{R}^N$ where its elements denote the proportions of the portfolio invested in a particular asset. Let $R\mathbb{N}$ denote the random

[1] The kernel trick is a technique used to project the input data space onto a higher dimension. It is more likely that the data will be linearly separable in a transformed higher dimensional space. See Friedman (2017) for more details. The dual formulation of SVM-2 does not required the specific transformation mapping but only requires an expression for the dot product of the transformed vectors.

single-period return of all the assets. Then, also let $\mu \in \mathbb{R}^N$ and $\Sigma \in \mathbb{R}^{N \times N}$ denote the mean and covariance of the random returns variable. The traditional mean-variance approach is to select x as follows:

$$\begin{aligned} \min_{x} \quad & \lambda x^\intercal \Sigma x - \mu^\intercal x \\ \text{s.t.} \quad & \mathbf{1}^\intercal x = 1 \\ & x \geq 0 \end{aligned} \tag{MVO}$$

where λ denotes the investor's aversion to risk. As discussed above, estimation error leads to sup-optimal decision making within the mean-variance framework and investor's can have a preference for lower cardinally portfolios making modifications and extensions of (MVO) necessary for practical purposes.

No prediction model exhibits perfect performance. Since all models are imperfect, it begs the question: does a parameterized model exist that leads to better decision making? Donti et al. (2017) show, there exist predictive models, that when used to guide stochastic programs, lead to a better decision. Furthermore, Butler and Kwon (2021) show that for the case of mean-variance optimization, it is possible to calibrate a model such that the realized mean-variance criterion exceeds that of the implied by a model estimated based on maximum likelihood. The standard practice is to model an uncertain output $u = f(y, w)$ that depends on input data y_i and is parameterized by $w \in \mathbb{R}^p$, and then, use the estimated model with the input y to make a portfolio decision to minimize a given cost function. Let the cost function be denoted by $c(x, u)$ where $x \in \mathcal{D}$ is the portfolio decision on a feasible set, and u is the uncertain output. Mathematically, this is stated as:

$$x^*(y, w) = \arg\min_{x \in \mathcal{D}} \quad \mathbb{E}_{u \sim f(y,w)}[c(x, u)] \tag{Nominal}$$

where it is clear that x^* is a function of the input data y and the model parameters w. It is often the case that samples of y and u are available. Let $\{y^{(k)}, u^{(k)}\}_{k=1}^M$ denote the samples. The integrated prediction and optimization approach (IPO) proceeds by setting w to minimize the realized costs

$c(x^*(y^{(k)}, w), u^{(k)})$ under the distribution defined by the samples:

$$\begin{aligned} \min_{w} \quad & \frac{1}{M} \sum_{k} c(x^*(y^{(k)}, w), u^{(k)}) \\ \text{s.t.} \quad & x^*(y^{(k)}, w) = \arg\min_{x \in \mathcal{D}} \quad \mathbb{E}_{u \sim f(y^{(k)}, w)}[c(x, u)] \quad \forall k = 1, 2, \ldots, M \end{aligned} \tag{IPO}$$

In general, (IPO) is a non-convex problem due to the dependency of x^* on the argmin operator. However, there has been recent success solving (IPO) with neural networks in the case that the minimization present in the constraints is a convex program (Agrawal et al., 2019).

2 Formulations

This section contains the exposition of three ideas which aim at integrating the selection of a separating hyper plane for assets with the capital allocation procedure according to a mean-variance objective. For conciseness let $\mathcal{C} = \{x \in \mathbb{R}^N \mid \mathbf{1}^\intercal x = 1 \text{ and } x \geq 0\}$

Model # 1: A Single Joint Program

Let $y_i \in \mathbb{R}^p$ denote a feature vector for the ith asset; this vector could contain information such as company earnings, debt to equity ratios, and other valuation metrics. It is assumed that there is a biasing feature included in y_i as well (a column of 1's). A stock i will be accepted into the mean-variance capital allocation phase if $w^\intercal y_i \geq 0$. One potential approach is to simultaneously solve for the separating hyperplane and the mean-variance portfolio by solving the following mixed-integer quadratic programming problem:

$$\begin{aligned} \min_{x,z} \quad & \lambda x^\intercal \Sigma x - \mu^\intercal x + \frac{1}{2} \|w\|^2 \\ \text{s.t.} \quad & x_i \leq z_i && \forall i = 1, 2 \ldots, N \\ & y_i^\intercal w \leq M z_i && \forall i = 1, 2 \ldots, N \\ & -M(1 - z_i) \leq y_i^\intercal w && \forall i = 1, 2 \ldots, N \\ & x \in \mathcal{C} \quad z \in \{0, 1\}^N \end{aligned} \tag{MVO-SVM1}$$

where z_i is 1 if the ith asset is accepted into the capital allocation stage, and M is a big M constant for modelling the disjunctive constraint that z_i is zero (one) if it falls on the negative (positive) side of the hyperplane. The norm penalty on w is meant to select a classifier w such that it is of the maximum margin.

There are a couple of key points about the above formulation. First off, assume that all combinations of eligible assets (positive and negative classes) are linearly separable. If that is the case, then the solution to (MVO-SVM1) is to solve the mean-variance optimization problem without consideration for the separating hyperplane and then solve train the SVM on the labels generated by the mean-variance solution. That is, first solve (MVO), and then using the solution x^* separate the assets into two classes: $u_i = +1$ if $x_i > 0$ and $u_i = -1$ otherwise. Using the given y_i and the labels u_i solve (SVM-2) to obtain w. If the assets are not linearly separable, (MVO-SVM1) restricts the portfolios to include assets that can be linearly separated. One question of interest is: is it the case that this type of regularization on the portfolio could improve out of sample performance? Also, what are the relations between (MVO-SVM1) and the dual of (SVM-1)? I think that if the kernel trick is used to make the problem separable, then the procedure to solve would again be reduced down to solving the portfolio optimization and the SVM separately.

Integrated Predict then Optimize - Mean Variance

Although constraining the set of feasible solutions to consist of assets that are linearly separable may be of value, it is not reflective of how an SVM would be used in practice for portfolio allocation. For example, in Paiva et al. (2019), the SVM is trained before the portfolio optimization occurs. It is instructive to think about how the SVM will be used when trying to optimize the investor's objective. First, the SVM decides based on given data y what assets to include in the portfolio optimization (u). Second, the optimization procedure decides the optimal mean-variance capital mix x. Third, capital is then allocated as prescribed, and lastly, a random return for the assets R is observed. The full allocation process as described lends itself well to the IPO framework described in Section 1. Let the input data y_i, stacked row-wise for each asset be denoted by $Y = [y_{ij}]\ \mathbb{R}^{N\times p}$. The input data Y, along with asset returns R can be observed on a consistent basis; let $\{(Y^k, R^k)\}_{k=1}^{M}$

denote M observations of the asset features and returns respectively. For a given k it follows that the binary vector of assets selected by w is given by $\mathbf{1}_{\{Yw\geq 0\}}$. It follows that within the an IPO framework for mean-variance optimization the goal would be to select w according to:

$$\begin{aligned}
\min_{w} \quad & \frac{1}{M}\sum_{k}\Big[-(R^{(k)})^{\intercal}\, x^*(Y^{(k)},w) + \lambda\, x^*(Y^{(k)},w)^{\intercal}(R^{(k)}-\hat{R})(R^{(k)}-\hat{R})^{\intercal}x^*(Y^{(k)},w)\Big] + \|w\|^2 \\
\text{s.t.} \quad & x^*(Y^{(k)},w) = \arg\min_{x\in\mathcal{D}} \quad \mathbb{E}_{u\sim f(y^{(k)},w)}[c(x,u)] \quad \forall k = 1,2,\ldots,M
\end{aligned} \tag{MVO-IPO}$$

where the norm is added as a regularization term to reflect the desire for maximum margin solutions. The following two sub-sections present two parameterizations of the nested minimization problem in the constraints. In any case, the nested problem must enforce the condition $x_i \approx 0$ if $w^{\intercal}y_i < 0$

Integrated Program # 1: Penalty approach

The most laid-back and intuitive approach would be to form the following optimization problem for a given $Y^{(k)}$:

$$x^*(Y^{(k)},w) = \arg\min_{x\in\mathcal{C}} \quad \lambda x^{\intercal}\Sigma^{(k)}x - (\mu^{(k)})^{\intercal}x + C\mathbf{1}_{\{Y^{(k)}w<0\}}x \tag{IPO-1}$$

where C is essentially a return penalty for the assets that are on the negative side of the hyper plane, and $\mu^{(k)}$ and $\Sigma^{(k)}$ are the estimates for the mean and covariance of the returns. A large enough value for C effectively forces the assets in the negative class to obtain zero portfolio weight in the allocation. Assuming that there is a separating hyperplane the particular scaling for the margin does not matter (Friedman, 2017). Therefore, the condition for separating the assets into the positive and negative class can also be written as $w^y \geq 1 \implies$ class $+1$ and $w^y \leq -1 \implies$ class -1. This scaling leads to the hinge loss definition in the objective of (SVM-3). Interestingly, it is also the case that the hinge loss $\max(1+v,0)$ is a convex surrogate for the indicator function $\mathbf{1}_{\{v\geq 0\}}$, meaning that the the hinge loss is a convex approximation that is always greater than or equal to the indicator function (Rigollet, 2015). This allows (IPO-1) to be written in the convenient form:

$$x^*(Y^{(k)}, w) = \underset{x \in \mathcal{C}}{\arg\min} \quad \lambda x^\intercal \Sigma^{(k)} x - (\mu^{(k)})^\intercal x + C \sum_{i=1}^{N} \max(0, 1 - w^\intercal y_i^{(k)} x_i) \tag{IPO-2}$$

This form above will allow for simpler computations of the gradient $\frac{\partial x^*}{\partial w}$ for the backward pass of training the network.

Integrated Program # 2: Constraint Approach

The last IPO program formulation is based on the joint optimization problem (MVO-SVM1); however, the hyperplane w is fixed and is used to constrain the binary asset inclusion decisions z_i

$$\begin{aligned} x^*(Y^{(k)}, w) = \underset{x,z,w}{\arg\min} \quad & \lambda x^\intercal \Sigma x - \mu^\intercal x \\ \text{s.t.} \quad & x_i \leq z_i && \forall i = 1, 2 \ldots, N \\ & w^\intercal y_i^{(k)} \leq M z_i && \forall i = 1, 2 \ldots, N \\ & -M(1 - z_i) \leq y_i^\intercal w && \forall i = 1, 2 \ldots, N \\ & x \in \mathcal{C} \quad z \in \{0, 1\}^N \end{aligned} \tag{IPO-3}$$

The challenge with (IPO-3) is that it is a binary program, and it is not clear how to differentiate through it to get the sensitivities with respect to w. To be continued...

2.1 Note on Data

To generate the y vectors, I think https://simfin.com/data/bulk should have the required fundamentals data. Otherwise, there are fundamentals data on Quandl for $24 dollars a month.

3 Next Steps

September 24th, 2021

There are several areas of further investigation/items to address. One item is to explore (MVO-SVM1) further. Does constraining the combinations of assets to be linearly separable provide a desirable regularization effect? Also, it may be worthwhile to explore how to include kernels

in (MVO-SVM1). Another action item is to read up on the literature of cardinality constrained portfolio optimization; for example, (Shaw et al., 2008). With regards to the IPO programs, one action item is to explore differentiating the mixed-integer quadratic programs to solve (IPO-3); for example Paulus et al. (2020), is one reference to explore. Another action item is to proceed ahead with learning the coding to solve IPO-2.

References

Agrawal, A., Amos, B., Barratt, S., Boyd, S., Diamond, S., & Kolter, Z. (2019). Differentiable convex optimization layers.

Anis, H., & Kwon, R. (2020). Cardinality constrained risk parity portfolios. *Available at SSRN 3805592*.

Blanchet, J., Chen, L., & Zhou, X. Y. (2018). Distributionally robust mean-variance portfolio selection with wasserstein distances. *arXiv preprint arXiv:1802.04885*.

Butler, A., & Kwon, R. (2021). Integrating prediction in mean-variance portfolio optimization. *Available at SSRN 3788875*.

Carrasco, M., & Noumon, N. (2011). Optimal portfolio selection using regularization. *Citeseer, Tech. Rep.*

Chang, T.-J., Meade, N., Beasley, J. E., & Sharaiha, Y. M. (2000). Heuristics for cardinality constrained portfolio optimisation. *Computers & Operations Research*, *27*(13), 1271–1302.

Costa, G., & Kwon, R. H. (2019). Risk parity portfolio optimization under a markov regime-switching framework. *Quantitative Finance*, *19*(3), 453–471.

Donti, P. L., Amos, B., & Kolter, J. Z. (2017). Task-based end-to-end model learning in stochastic optimization. *arXiv preprint arXiv:1703.04529*.

Fan, A., & Palaniswami, M. (2001). Stock selection using support vector machines. *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)*, *3*, 1793–1798.

Friedman, J. H. (2017). *The elements of statistical learning: Data mining, inference, and prediction*. springer open.

Goldfarb, D., & Iyengar, G. (2003). Robust portfolio selection problems. *Mathematics of operations research*, *28*(1), 1–38.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77–91.

Menon, A. K. (2009). Large-scale support vector machines: Algorithms and theory.

Michaud, R. O., & Michaud, R. (2007). Estimation error and portfolio optimization: A resampling solution. *Available at SSRN 2658657*.

Paiva, F. D., Cardoso, R. T. N., Hanaoka, G. P., & Duarte, W. M. (2019). Decision-making for financial trading: A fusion approach of machine learning and portfolio selection. *Expert Systems with Applications*, *115*, 635–655.

Paulus, A., Rolinek, M., Musil, V., Amos, B., & Martius, G. (2020). Fit the right {np}-hard problem: End-to-end learning of integer programming constraints. *Learning Meets Combinatorial Algorithms at NeurIPS2020*. https://openreview.net/forum?id=-3qCWheZhxU

Rigollet, P. (2015). Lecture 8: Convexification [MIT OpenCourseWare]. *18.657 mathematics of machine learning*. hhttps://ocw.mit.edu/courses/mathematics/18-657-mathematics-of-machine-learning-fall-2015/index.htm

Shaw, D. X., Liu, S., & Kopman, L. (2008). Lagrangian relaxation procedure for cardinality-constrained portfolio optimization. *Optimisation Methods & Software*, *23*(3), 411–420.

Tillmann, A. M., Bienstock, D., Lodi, A., & Schwartz, A. (2021). Cardinality minimization, constraints, and regularization: A survey.

Appendix