IMPROVED PORTFOLIO DIVERSIFICATION THROUGH UNSUPERVISED LEARNING

Michael Lewis

Original version: August 6, 2019

This version: August 29, 2019

IMPROVED PORTFOLIO DIVERSIFICATION THROUGH UNSUPERVISED LEARNING

ABSTRACT

This paper introduces the Recursive Clustering Risk Parity (RCRP) approach. The RCRP method builds on the philosophy first introduced in Hierarchical Risk Parity (HRP) in López de Prado [2016], leveraging unsupervised learning to recursively build an optimal portfolio.

HRP introduced the concept of building a diversified portfolio using the inherent structure of the covariance matrix, utilizing graph theory and hierarchical clustering. In doing so, HRP avoided inverting the covariance matrix, a numerically unstable procedure when performed on the notoriously ill-conditioned, if not singular, covariance matrix. However, the procedure relies on sorting the underlying instruments and, in doing so, creating a quasi-diagonalization of the matrix, thereby foregoing the underlying tree structure. RCRP builds on this premise by utilizing this underlying tree structure, leveraging improved clustering techniques and inversion approximations to enhance overall performance.

Keywords: Risk parity, tree graph, cluster, recursive, inverse rank-one update, metric space.

JEL Classification: G0, G1, G2.

AMS Classification: 91G10, 91G60, 91G70, 60E.

PORTFOLIO OPTIMIZATION: A BACKGROUND

This paper proposes a method to obtain two desirable portfolios: the minimum variance portfolio and the maximum sharpe ratio portfolio. We consider both the case where there are limits on shorts, and when there aren't; initially, we will focus on the case without limits.

For completeness, a portfolio is a vector of weights w across N market instruments. When considering the case of no shorting limits, the weights sum to unity, or $1^T w = 1$, while in the case with limits the absolute weights sum to unity, or $\sum_i |w_i| = 1$. Given the covariance matrix Σ, expected excess returns vector μ, and portfolio weights w, the variance of the portfolio is $w^T \Sigma w$ and the expected excess return of the portfolio is $w^T \mu$.

With that terminology in place, we will be discussing the minimum variance portfolio without shorting limits, which satisfies equation (1),

$$\mathit{argmin}_w w^T \Sigma w$$

$$\text{such that } 1^T w = 1 \quad (1)$$

and the maximum sharpe ratio portfolio without shorting limits, which satisfies equation (2).

$$\mathit{argmax}_w \frac{w^T \mu}{\sqrt{w^T \Sigma w}}$$

$$\text{such that } 1^T w = 1 \quad (2)$$

The known optimal solution to equation (1) is setting $w = \frac{\Sigma^{-1}1}{1^T\Sigma^{-1}1}$, or equivalently $w \propto \Sigma^{-1}1$, whereas the known optimal solution to equation (2) is setting $w \propto \Sigma^{-1}\mu$. Put more concisely, the solutions set $w \propto \Sigma^{-1}a$, where $a = 1$ for the minimum variance portfolio, and $a = \mu$ for the max sharpe ratio. Unfortunately, both solutions require inverting the covariance matrix, a numerically unstable task when it's possible, and impossible when the covariance matrix is singular, as is often the case in practice when involving thousands of stocks. Aside from this inherent numerical instability, empirical covariance matrices can themselves be nosy when evaluated on a finite number of financial returns, further complicating matters. A method that is robust to noise and does not require ill-conditioned matrix inversion is desirable.

One notable option for the minimum variance portfolio is the inverse variance portfolio. In this case, the weight w_i given to instrument i is set to be inversely proportional to its variance $\sigma_i^2 = \Sigma_{ii}$, or $w_i \propto \frac{1}{\sigma_i^2}$. Observe that this is the optimal solution in the special case where the instruments are independent, or equivalently $\Sigma = \Sigma_{diag}$ is a diagonal matrix with $\sigma_{ij} = \Sigma_{ij} = 0$ for $i \neq j$. HRP notably uses this result when recursively evaluating the portfolio weights subsequent to quasi-diagonalizing the covariance matrix. However, in finance it is often the case that instruments are not independent, and for this reason the inverse variance portfolio is suboptimal.

IMPROVED INVERSE APPROXIMATION

A notable improvement to the inverse variance portfolio can be found in López de Prado and Lewis [2018]. We assume a constant correlation ρ between each instrument i, j with $i \neq j$, then $\sigma_{ij} = \rho \sigma_i \sigma_j$, and thus

$$\Sigma = (1 - \rho)\Sigma_{diag} + \rho \sigma \sigma^T,$$

where I is the identity matrix, and σ is the column vector of standard deviations. Given this is a rank-one update to the matrix $(1 - \rho)\Sigma_{diag}$, we can use the Sherman-Morrison Identity to evaluate its analytic inverse.

$$\Sigma^{-1} = \frac{1}{1-\rho}\Sigma_{diag}{}^{-1} - \frac{\rho}{(1-\rho)(1+\rho(N-1))}\left(\frac{1}{\sigma}\right)\left(\frac{1}{\sigma}\right)^T$$

In this equation $\left(\frac{1}{\sigma}\right)$ is, with slight abuse of notation, the vector with component i set to $\frac{1}{\sigma_i}$. For a covariance matrix with this form, the optimal minimum variance portfolio takes the form

$$w_i \propto \frac{1}{\sigma_i{}^2} - \frac{\rho \sum_j \frac{1}{\sigma_j}}{(1+\rho(N-1))\sigma_i},$$

while the optimal maximum sharpe ratio portfolio has

$$w_i \propto \frac{\mu_i}{\sigma_i{}^2} - \frac{\rho \sum_j \frac{\mu_j}{\sigma_j}}{(1+\rho(N-1))\sigma_i}.$$

This predictably reduces to the inverse variance portfolio in the case where $\rho = 0$. In practice, we can set ρ to equal the off-diagonal average correlation if we wish to approximate the covariance matrix this way.

PORTFOLIO ENHANCEMENT VIA RECURSIVE CLUSTERING

While the above enhancement is a useful approximation in cases where the off-diagonal correlations are all similar, this isn't the typical case. We therefore reconsider use of this method to better leverage this result.

To facilitate discussion, let us first consider a portfolio of stocks that are exclusively in the Industrial or the Technology sector. In this scenario, we expect Industrial stocks to have returns similar to other Industrial stocks, and likewise Technology stocks returns similar to other Technology stocks. However, Industrial stock performance will be less similar to Technology stocks. If we sort the indices to place the Industrials first followed by the Technology stocks, we should expect the correlation matrix matrix to have a block-like formation. See Exhibit 1 for such an example of a hypothetical correlation matrix. Consider first the portfolio of just Industrials as portfolio 1, and the portfolio of just Technology stocks as portfolio 2. We could optimize portfolio 1, then do similarly for portfolio 2, and finally optimize this portfolio of optimized portfolios. While technically suboptimal, its performance would likely be near optimal.

This thought experiment brings up a few discussion questions:

1. Some Industrial stocks performs more similarly to Technology stocks than other Industrials, so why delineate by sector?
2. Within the Technology sector, stocks A and B may perform more similarly to one another and less like stock C, so why not further dissect the stocks within each sector?

3. Why limit the analysis to a portfolio of just Industrials and Technology stocks?

Thus, for a more general portfolio, we will generalize this optimization procedure

We refer to this generalized optimal portfolio procedure as Recursive Clustering Risk Parity (RCRP), which takes as inputs a covariance matrix of returns Σ and vector a, with $a = 1$ in the case of the minimum variance portfolio and $a = \mu$ in the max sharpe ratio portfolio. The procedure can be described as follows:

1. Form the correlation matrix ρ from the covariance matrix Σ for a certain set of instruments S.
2. Use advanced unsupervised learning techniques to cluster ρ and partition S into some optimal k disjoint sets of instruments S_i.
3. For each set S_i, extract the covariance matrix Σ_i and vector a_i associated with those instruments
4. If Σ_i is too small for clustering, find the optimal portfolio vector w_i via the improve inverse portfolio optimization above using Σ_i and a_i. Otherwise, set $w_i = RCRP(\Sigma_i,\ a_i)$. Note that w_i has dimension $N_i = |S_i|$ and $1^T w_i = 1$.
5. Using w_i to define the change in basis elements, create the compressed covariance matrix Σ^{comp} and compressed vector a^{comp} for our portfolio of portfolios. This is equivalent to treating the performance of the optimized portfolios as individual instruments. Observe that $a_i^{comp} = a_i^T w_i$ and $\Sigma_{ij}^{comp} = w_i^T \Sigma_{ij} w_j$, where Σ_{ij} is the submatrix of Σ with rows related to instruments in S_i and columns related to

instruments in S_j. Find the cluster weights vector v via the improved inverse portfolio optimization using Σ^{comp} and a^{comp}. Note that v has dimension k, and $1^T v = 1$.

6. For $i = 1, \ldots, k$, set $w_i \rightarrow v_i * w_i$, then concatenate the weight vectors w_i into a single vector w, and return w. Note that $1^T w = 1$ as desired.

For step (2), we consider the clustering method formulated in López de Prado and Lewis [2018], though this can be enhanced should improved clustering techniques arise.

INCORPORATING LIMITS ON SHORT POSITIONS

The above analysis and procedure focused on the scenario where there are no shorting limits. We now discuss extending both to the case with limits.

NUMERICAL EXAMPLES:

The code for this analysis, written in Python, can be found at: https://github.com/mlewis1729

A key takeaway is that there is significant improved performance to industry alternatives.

CONCLUSIONS

We have discussed a new procedure.

REFERENCES

López de Prado, Marcos, Building Diversified Portfolios that Outperform Out-of-Sample (May 23, 2016). Journal of Portfolio Management, 2016; https://doi.org/10.3905/jpm.2016.42.4.059. . Available at SSRN: https://ssrn.com/abstract=2708678 or http://dx.doi.org/10.2139/ssrn.2708678

López de Prado, Marcos and Lewis, Michael J., Detection of False Investment Strategies Using Unsupervised Learning Methods (August 18, 2018). Available at SSRN: https://ssrn.com/abstract=3167017 or http://dx.doi.org/10.2139/ssrn.3167017

Vim (https://quant.stackexchange.com/users/19004/vim), Maximum Sharpe portfolio (no short selling restrictions) Quantitative Finance Stack Exchange, URL: https://quant.stackexchange.com/questions/43999/maximum-sharpe-portfolio-no-short-selling-restrictions (version: 8/26/2019)

Garey, M. R. and Johnson, D. S., Computers and Intractability. A Guide to the Theory of NP-Completeness (1979).

APPENDIX

A.1. OPTIMAL PORTFOLIOS WITHOUT SHORTING LIMITS

In López de Prado [2016], a proof is given for the optimal minimum variance portfolio construction. The following proof for the optimal max sharpe portfolio construction is given for completeness. A similar proof is given by Vim [2019] on the Quantitative Finance Stack Exchange.

As stated previously, our goal is to solve

$$argmax_{\ w} \ \frac{w^T \mu}{\sqrt{w^T \Sigma w}}$$

$$such\ that\ 1^T w = 1$$

Let

$$g_1(w) = w^T \mu$$

$$g_2(w) = w^T \Sigma w$$

$$h(x_1, x_2) = \frac{x_1}{\sqrt{x_2}}$$

$$\mathrm{f(w)}= \frac{w^T \mu}{\sqrt{w^T \Sigma w}} = h(g_1(w), g_2(w))$$

We set the derivative of $f(w)$ to 0, and find

$$\frac{1}{\sqrt{w^T \Sigma w}} \mu_i - \frac{w^T \mu}{2(w^T \Sigma w)^{3/2}} (2\Sigma w)_i = 0 \ \forall i \ \Rightarrow \Sigma w = \frac{w^T \Sigma w}{w^T \mu} \mu = C\mu \,,$$

implying the optimal portfolio $w = C\Sigma^{-1}\mu$. Observe that $f(w) = f(Cw)$ for $C > 0$; thus, the optimal max sharpe portfolio has $w \propto \Sigma^{-1}\mu$.

Note that this same argument extends to the case where short positions are limited, in which case the condition $\Sigma_i |w_i| = 1$ holds. Thus, the max sharpe portfolio with no shorting limits differs from the case with limits only by the normalization.

A2. OPTIMAL PORTFOLIOS WITH SHORTING LIMITS

We now discuss the scenario where short positions on the portfolio are limited. To do so, we exchange the condition $\Sigma_i w_i = 1$ with $\Sigma_i |w_i| = 1$. Section A.1 makes clear the max sharpe ratio portfolio with shorting limits differs only in normalization from the case without limits. In this section, we focus our attention on the minimum variance portfolio.

We consider the problem

$$argmax_{w}\ w^T \Sigma w$$

$$such\ that\ \sum_i |w_i| = 1\,.$$

Using Lagrange multipliers, we set the derivative of $w^T\Sigma w + \lambda(\sum_i |w_i| - 1)$ to 0, and find

$$\Sigma w = Cs \Rightarrow w = C\Sigma^{-1}s\,,$$

where $s_i = \sigma(w_i)$, the sign of w_i, and C is a normalizing constant. Observe that

$s^T w = \sum_i |w_i| = 1 = Cs^T\Sigma^{-1}s$, and thus $C = \frac{1}{s^T\Sigma^{-1}s}$. Therefore, our optimal minimum variance portfolio is $w = \frac{\Sigma^{-1}s}{s^T\Sigma^{-1}s}$, and the minimum variance is $w^T\Sigma w = C^2 s^T\Sigma^{-1}\Sigma\Sigma^{-1}s = C$.

The difficulty of this problem depends on discovering the appropriate s. However, interpreting C is far simpler, as it is the smallest variance it can be, so we focus our attention on that. Let $S = \{s | s_i \in \{-1, 1\} \forall i\}$ be the set of all possible sign vectors, and thus $s \in S$. Thus, our problem reduces to finding

$$argmax_{s \in S}\, s^T \Sigma^{-1} s .$$

This is challenging to analyze for general covariance matrix Σ, so we reduce our consideration to the approximate matrix listed earlier, namely

$$\Sigma = (1 - \rho)\Sigma_{diag} + \rho\sigma\sigma^T .$$

Note that Σ **is positive semi-definite. By the Matrix determinant lemma,**

$$det(\Sigma) = det((1-\rho)\Sigma_{diag})(1 + \frac{\rho\sigma^T\Sigma_{diag}^{-1}\sigma}{(1-\rho)}) \geq 0 \Rightarrow \frac{\rho\sigma^T\Sigma_{diag}^{-1}\sigma}{(1-\rho)} = \frac{N\rho}{(1-\rho)} \geq -1 \text{,}$$

implying that $\rho \geq \frac{-1}{N-1}$ is a necessary condition. Next, observe that $s^T \Sigma_{diag}{}^{-1} s = \sum_i \frac{1}{\sigma_i{}^2} \forall s \in S$. Third, let $K = \frac{\rho}{(1+\rho(N-1))}$, and thus $\sigma(K) = \sigma(\rho)$. Given that

$$\Sigma^{-1} = \frac{1}{1-\rho}\Sigma_{diag}{}^{-1} - \frac{\rho}{(1-\rho)(1+\rho(N-1))}(\frac{1}{\sigma})(\frac{1}{\sigma})^T ,$$

we find

$$s^T\Sigma^{-1}s = \frac{1}{1-\rho}(\sum_i \frac{1}{\sigma_i^2}) - \frac{K}{(1-\rho)}((\frac{1}{\sigma})^T s)^2 .$$

If $\rho < 0$, then $K < 0$, and thus maximizing $s^T\Sigma^{-1}s$ requires maximizing $|(\frac{1}{\sigma})^T s|$. Given that $\sigma_i > 0 \ \forall i$, the optimal $s = \pm 1$. If $\rho > 0$, then $K > 0$, and thus maximizing $s^T\Sigma^{-1}s$ requires minimizing $|(\frac{1}{\sigma})^T s|$.

Since s is a vector of ± 1s, we see that minimizing $|(\frac{1}{\sigma})^T s|$ can be observed as the optimization version of the partition problem, though with real numbers rather than integers, which is known to be NP-hard (Garey and Johnson [1979]). For small N, we can check all $s \in S$, noting that $|S| = 2^N$. For improved computation, note that the variance is the same for $\pm s$; therefore, consider only s such that $s_1 = 1$, thereby reducing computation by a factor of 2. For larger N, an approximate solution must be found.

Consider the following greedy algorithm: let $s_1 = 1$, and let $y = \frac{1}{\sigma_i}$ hold the running sum. For $n > 1$, if $y > 0$, then set $s_n =- 1$, and $y \leftarrow y - \frac{1}{\sigma_n}$; otherwise, set $s_n = 1$, and $y \leftarrow y + \frac{1}{\sigma_n}$. When $n = N$, observe that $y = \sum_j \frac{s_j}{\sigma_j}$. We can then repeat this greedy procedure multiple times by permuting the σ_i, running this procedure, performing the inverse permutation on the resulting s, and ultimately choose the s that obtains the smallest $|y|$.

Finally, there is the matter of confirming that, given the optimal s, the condition $s_i = \sigma(w_i)$ is satisfied. Given that $w = C\Sigma^{-1}s$, and $C > 0$, we observe

$$w_i \propto \frac{s_i}{\sigma_i^2} - \frac{K}{\sigma_i} \sum_j \frac{s_j}{\sigma_j} \Rightarrow |w_i| = s_i w_i \propto \frac{1}{\sigma_i} - K s_i \sum_j \frac{s_j}{\sigma_j} \geq 0$$

is a necessary and sufficient condition. We see this condition is trivially satisfied when $\rho < 0 \Rightarrow K < 0$, with optimal $s = \pm 1$. Observe that $\rho \in [0, 1] \Rightarrow K \in [0, \frac{1}{N}]$. If $s_i \sum_j \frac{s_j}{\sigma_j} \leq 0$, then this condition is satisfied trivially; therefore, consider the case $s_i \sum_j \frac{s_j}{\sigma_j} \geq 0$. Observe that, if $s_i \geq 0$, then $\sum_j \frac{s_j}{\sigma_j} \geq 0$. In this scenario, if $\sum_j \frac{s_j}{\sigma_j} \geq \frac{1}{\sigma_i}$, the sum could be decreased by setting $s_i =- 1$, and thus is not the optimal s. Consequently, in this scenario, $\sum_j \frac{s_j}{\sigma_j} < \frac{1}{\sigma_i}$, and thus the condition $Ks_i \sum_j \frac{s_j}{\sigma_j} \le \frac{1}{\sigma_i}$ is satisfied. A similarly argument confirms the case for $s_i < 0$, and thus we confirm the condition $s_i = \sigma(w_i)$ is satisfied for optimal s.