Machine Learning Asset Allocation

Prof. Marcos López de Prado

Advances in Financial Machine Learning

ORIE 5256

Key Points

- Convex optimization solutions tend to be unstable, to the point of entirely offsetting the benefits of optimization.
  - For example, in the context of financial applications, it is known that portfolios optimized in-sample often underperform the naïve (equal weights) allocation out-of-sample.
- This instability can be traced back to two sources:
  - noise in the input variables
  - signal structure that magnifies the estimation errors in the input variables.
- There is abundant literature discussing noise-induced instability.
- In contrast, signal-induced instability is often ignored or misunderstood.
- **We introduce a new optimization method that is robust to signal-induced instability.**
- For additional details, see the full paper at: https://ssrn.com/abstract=3469961

SECTION I
Problem Statement

The Problem

- Consider a system with N random variables, where the expected value of draws from these variables is represented by an array μ, and the variance of these draws is represented by the covariance matrix V.
- We would like to minimize the variance of the system, measured as $\omega' V \omega$, subject to achieving a target $\omega' a$, where a characterizes the optimal solution.
- The problem can be stated as

$$\min_{\omega} \frac{1}{2} \omega' V \omega$$
$$\text{s.t.: } \omega' a = 1$$

The Solution (1/2)

- This problem can be expressed in lagrangian form as

$$L[\omega, \lambda] = \frac{1}{2}\omega' V \omega - \lambda(\omega' a - 1)$$

with first order conditions

$$\frac{\partial L[\omega, \lambda]}{\partial \omega} = V\omega - \lambda a$$

$$\frac{\partial L[\omega, \lambda]}{\partial \lambda} = \omega' a - 1$$

- Setting the first order (necessary) conditions to zero, we obtain that $V\omega - \lambda a = 0 \Longrightarrow \omega = \lambda V^{-1} a$, and $\omega' a = a' \omega = 1 \Longrightarrow \lambda a' V^{-1} a = 1 \Longrightarrow \lambda = \frac{1}{a' V^{-1} a}$, thus

$$\omega^* = \frac{V^{-1} a}{a' V^{-1} a}$$

The Solution (2/2)

- The second order (sufficient) condition confirms that this solution is the minimum of the lagrangian,

$$\begin{vmatrix} \frac{\partial L^2[\omega,\lambda]}{\partial\omega^2} & \frac{\partial L^2[\omega,\lambda]}{\partial\omega\partial\lambda} \\ \frac{\partial L^2[\omega,\lambda]}{\partial\lambda\partial\omega} & \frac{\partial L^2[\omega,\lambda]}{\partial\lambda^2} \end{vmatrix} = \begin{vmatrix} V' & -a' \\ a & 0 \end{vmatrix} = a'a \geq 0$$

- The issue is, this solution is mathematically correct, but impractical.

Numerical Instability

- The common approach to estimating ω^* is to compute

$$\hat{\omega}^* = \frac{\hat{V}^{-1}\hat{a}}{\hat{a}'\hat{V}^{-1}\hat{a}}$$

where $\hat{V}$ is the estimated V, and $\hat{a}$ is the estimated a.

- In general, replacing each variable with its estimate will lead to unstable solutions, that is, solutions where a small change in the inputs will cause extreme changes in $\hat{\omega}^*$.
- This is problematic, because in many practical applications there are material costs associated with the re-allocation from one solution to another.

SECTION II
Noise-induced Instability

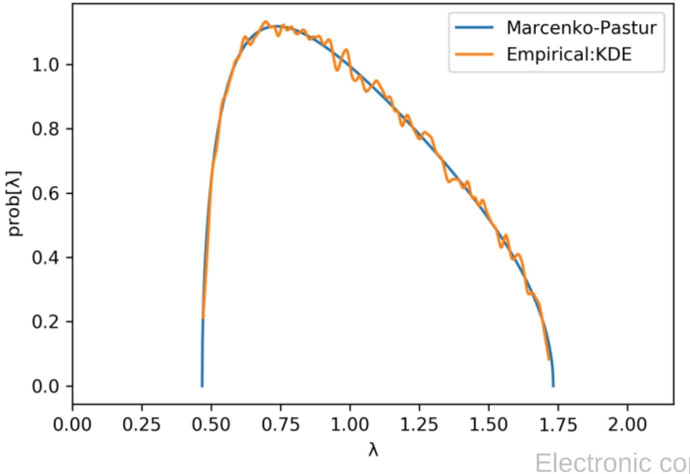
The Marcenko-Pastur Distribution (1/2)

- Consider a matrix of independent and identically distributed random observations X, of size TxN, where the underlying process generating the observations has zero mean and variance σ^2.
- The matrix $C = T^{-1}X'X$ has eigenvalues λ that asymptotically converge (as $N \rightarrow +\infty$ and $T \rightarrow +\infty$ with $1 < {}^{T}/_{N} < +\infty$) to the Marcenko-Pastur probability density function (PDF),

$$f[\lambda] = \begin{cases} \frac{T}{N}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\lambda\sigma^2} & \text{if} \quad \lambda \in [\lambda_-, \lambda_+] \\ 0 & \text{if} \quad \lambda \notin [\lambda_-, \lambda_+] \end{cases}$$

The Marcenko-Pastur Distribution (2/2)

… where the maximum expected eigenvalue is $\lambda_{+} = \sigma^2\left(1+\sqrt{\frac{N}{T}}\right)^2$, and the minimum expected eigenvalue is $\lambda_{-} = \sigma^2\left(1-\sqrt{\frac{N}{T}}\right)^2$. When $\sigma^2 = 1$, then C is the correlation matrix associated with X.



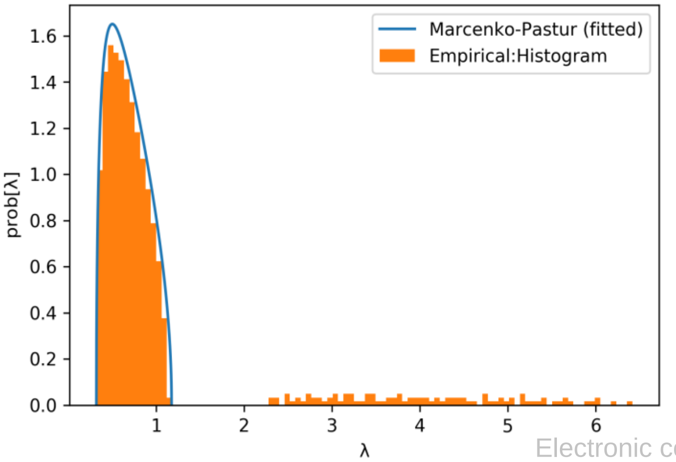
Eigenvalues $\lambda \in [\lambda_{-}, \lambda_{+}]$ are consistent with random behavior, and eigenvalues $\lambda \notin [\lambda_{-}, \lambda_{+}]$ are consistent with non-random behavior. Specifically, we associate eigenvalues $\lambda \in [0, \lambda_{+}]$ with noise.

Problem: In empirical covariance matrices, most of the eigenvalues fall under the Marcenko-Pastur distribution, and are insignificant.

The implication is that neither C^{-1} nor V^{-1} can be computed robustly. Solutions are only optimal in-sample, not out-of-sample.

Fitting the Marcenko-Pastur PDF

- Laloux et al. [2005] argue that, since only part of the variance is caused by random eigenvectors, we can adjust σ^2 accordingly in the above equations.
  - For instance, if we assume that the eigenvector associated with the highest eigenvalue is *not* random, then we should replace σ^2 with $\sigma^2\left(1-\frac{\lambda_+}{N}\right)$ in the above equations.
- In fact, we can fit the function $f[\lambda]$ to the empirical distribution of the eigenvalues to derive the implied σ^2.



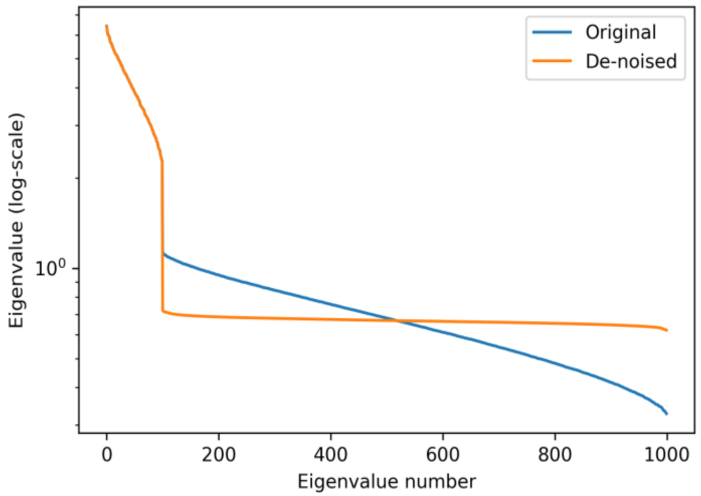
That will give us the variance that is explained by the random eigenvectors present in the correlation matrix, and it will determine the cut-off level λ_+, adjusted for the presence of non-random eigenvectors.

Key point: Because we know what eigenvalues are associated with noise, we can shrink only those, without diluting the signal!

SECTION III
De-Noising and De-Toning

The Constant Residual Eigenvalue Method

- Let $\{\lambda_n\}_{n=1,\dots,N}$ be the set of all eigenvalues, ordered descending, and i be the position of the eigenvalue such that $\lambda_i > \lambda_+$ and $\lambda_{i+1} \leq \lambda_+$.
- Then we set $\lambda_j = \frac{1}{N-i}\sum_{k=i+1}^{N} \lambda_k$, $j = i+1,\dots,N$, hence preserving the trace of the correlation matrix.



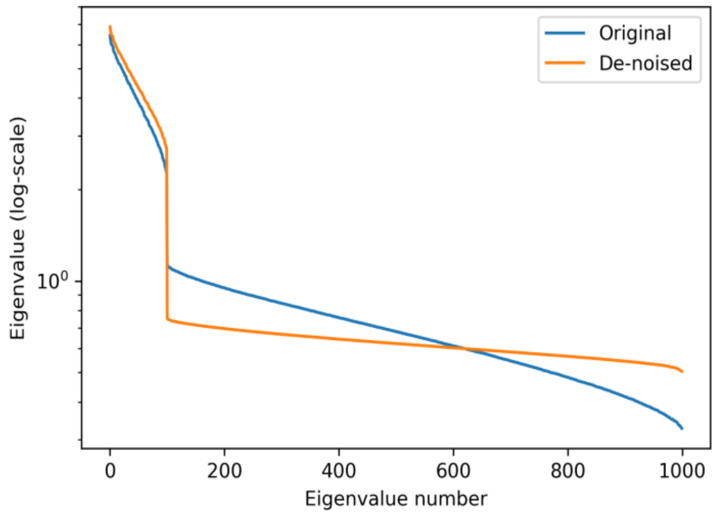
Given the eigenvector decomposition $VW = W\Lambda$, we form the de-noised correlation matrix C_1 as

$$\tilde{C}_1 = W\tilde{\Lambda}W'$$
$$C_1 = \left(\mathrm{diag}[\tilde{C}_1]\right)^{-1/2}\tilde{C}_1\left(\mathrm{diag}[\tilde{C}_1]\right)^{-1/2}$$

where $\tilde{\Lambda}$ is the diagonal matrix holding the corrected eigenvalues. The reason for the second transformation is to re-scale the matrix $\tilde{C}_1$, so that the main diagonal of C_1 is an array of 1s.

The Targeted Shrinkage Method

- The numerical method described earlier is preferable to shrinkage, because it removes the noise while preserving the signal.
- Alternatively, we could target the application of the shrinkage strictly to the random eigenvectors. Consider the correlation matrix C_1



$$C_1 = W_L \Lambda_L W_L' + \alpha W_R \Lambda_R W_R' + (1-\alpha)\mathrm{diag}[W_R \Lambda_R W_R']$$

where W_R and Λ_R are the eigenvectors and eigenvalues associated with $\{n \mid \lambda_n \leq \lambda_+\}$, W_L and Λ_L are the eigenvectors and eigenvalues associated with $\{n \mid \lambda_n > \lambda_+\}$, and α regulates the amount of shrinkage among the eigenvectors and eigenvalues associated with noise ($\alpha \to 0$ for total shrinkage).

De-Toning (1/3)

- Financial correlation matrices usually incorporate a market component.
- The market component is characterized by the first eigenvector, with loadings $W_{n,1} \approx N^{-\frac{1}{2}}, n = 1, \dots, N$.
- Accordingly, a market component affects every item of the covariance matrix.
- By removing the market component, we allow a greater portion of the correlation to be explained by components that affect specific subsets of the securities.
- <u>Intuition</u>: De-toning is similar to removing a loud tone that prevents us from hearing other sounds.

De-Toning (2/3)

- We can remove the market component from the de-noised correlation matrix, C_1, to form the de-toned correlation matrix,

$$\tilde{C}_2 = C_1 - W_M \Lambda_M W_M' = W_D \Lambda_D W_D'$$
$$C_2 = \left(\text{diag}\left[\tilde{C}_2\right]\right)^{-1/2} \tilde{C}_2 \left(\text{diag}\left[\tilde{C}_2\right]\right)^{-1/2}$$

where W_M and Λ_M are the eigenvectors and eigenvalues associated with market components (usually only one, but possibly more), and W_D and Λ_D are the eigenvectors and eigenvalues associated with non-market components.

De-Toning (3/3)

- The de-toned correlation matrix is singular, as a result of eliminating (at least) one eigenvector.
  - This is not a problem for clustering applications, as most approaches do not require the invertibility of the correlation matrix.
- Still, a de-toned correlation matrix C_2 cannot be used directly for mean-variance portfolio optimization.
- Instead, we can optimize a portfolio on the selected (non-zero) principal components, and map the optimal allocations f^* back to the original basis.
- The optimal allocations in the original basis are

$$\omega^* = W_+ f^*$$

where W_+ contains only the eigenvectors that survived the de-toning process (i.e., with a non-null eigenvalue), and f^* is the vector of optimal allocations to those same components.

Experimental Results (1/2)

- We generate a vector of means and a covariance matrix out of 10 blocks of size 50 each, where off-diagonal elements within each block have a correlation of 0.5.
  - This covariance matrix is a stylized representation of a “true” (non-empirical) de-toned correlation matrix of the S&P 500, where each block is associated with an economic sector.
  - Without loss of generality, the variances are drawn from a uniform distribution bounded between 5% and 20%, and the vector of means is drawn from a Normal distribution with mean and standard deviation equal to the standard deviation from the covariance matrix.
  - This is consistent with the notion that in an efficient market all securities have the same expected Sharpe ratio.
- We use this means vector and covariance matrix to draw 1,000 random matrices X of size $TxN = 1000x500$, compute the associated empirical covariance matrices and vectors of means, and evaluate the (empirical) optimal portfolios.
- We compute the root-mean-square error (RMSE) between the “empirical” and “true” optimal portfolios.

Experimental Results (2/2)

	Not De-Noised	De-Noised
Not Shrunk	4.95E-03	1.99E-03
Shrunk	3.45E-03	1.70E-03

Minimum Variance Portfolio

	Not De-Noised	De-Noised
Not Shrunk	9.48E-01	5.27E-02
Shrunk	2.77E-01	5.17E-02

Maximum Sharpe Ratio Portfolio

De-noising is much more effective than shrinkage: the de-noised minimum variance portfolio incurs only 40.15% of the RMSE incurred by the minimum variance portfolio without de-noising. That is a 59.85% reduction in RMSE from de-noising alone, compared to a 30.22% reduction using Ledoit-Wolf shrinkage. Shrinkage adds little benefit beyond what de-noising contributes. The reduction in RMSE from combining de-noising with shrinkage is 65.63%, which is not much better than the result from using de-noising only.

The de-noised maximum Sharpe ratio portfolio incurs only 0.04% of the RMSE incurred by the maximum Sharpe ratio portfolio without de-noising. That is a 94.44% reduction in RMSE from de-noising alone, compared to a 70.77% reduction using Ledoit-Wolf shrinkage. While shrinkage is somewhat helpful in absence of de-noising, it adds no benefit in combination with de-noising. This is because shrinkage dilutes the noise at the expense of diluting some of the signal as well.

SECTION IV
Signal-induced Instability

The Condition Number (1/2)

- Certain covariance structures can make the mean-variance optimization solution unstable.
- Consider a correlation matrix between two securities,

$$C = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

where ρ is the correlation between their returns.

- Matrix C can be diagonalized as $CW = W\Lambda$ as follows, where

$$W = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}, \Lambda = \begin{bmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{bmatrix}$$

The Condition Number (2/2)

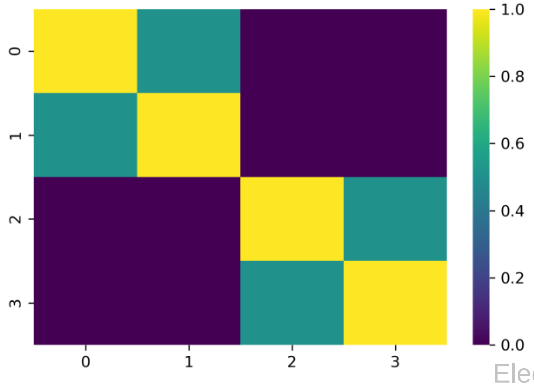
- The trace of C is $tr(C) = \Lambda_{1,1} + \Lambda_{2,2} = 2$, so ρ sets how big one eigenvalue gets at the expense of the other.
- The determinant of C is given by $|C| = \Lambda_{1,1}\Lambda_{2,2} = (1+\rho)(1-\rho) = 1-\rho^2$.
  - The determinant reaches its maximum at $\Lambda_{1,1} = \Lambda_{2,2} = 1$, which corresponds to the uncorrelated case, $\rho = 0$.
  - The determinant reaches its minimum at $\Lambda_{1,1} = 0$ or $\Lambda_{2,2} = 0$, which corresponds to the perfectly correlated case, $|\rho| = 1$.
- The inverse of C is $C^{-1} = W\Lambda^{-1}W' = \frac{1}{|C|}\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}$.
- The implication is that, the more ρ deviates from zero, the bigger one eigenvalue becomes relative to the other, causing $|C|$ to approach zero, which makes the values of C^{-1} explode. ***This happens regardless of the N/T ratio***.

Markowitz's Curse

- Matrix C is just a standardized version of V, and the conclusions we drew on C^{-1} apply to the V^{-1} used to compute ω^*.
- When securities within a portfolio are highly correlated ($-1 < \rho \ll 0$ or $0 \ll \rho < 1$), C has a high condition number, and the values of V^{-1} explode.
- This is problematic in the context of portfolio optimization, because ω^* depends on V^{-1}, and unless $\rho \approx 0$, we must expect an unstable solution to the convex optimization program.
- In other words, Markowitz's solution is guaranteed to be numerically stable only if $\rho \approx 0$, which is precisely the case when we don't need it!
- The reason we needed Markowitz was to handle the $\rho \not\approx 0$ case, but the more we need Markowitz, the more numerically unstable is its estimation of ω^*.

Signal-induced Instability in Finance

- When a subset of securities exhibits greater correlation among themselves than to the rest of the investment universe, that subset forms a cluster within the correlation matrix.
- Clusters appear naturally, as a consequence of hierarchical relationships.
- When K securities form a cluster, they are more heavily exposed to a common eigenvector, which implies that the associated eigenvalue explains a greater amount of variance.



But because the trace of the correlation matrix is exactly N, that means that an eigenvalue can only increase at the expense of the other $N - K$ eigenvalues, resulting in a condition number greater than 1.

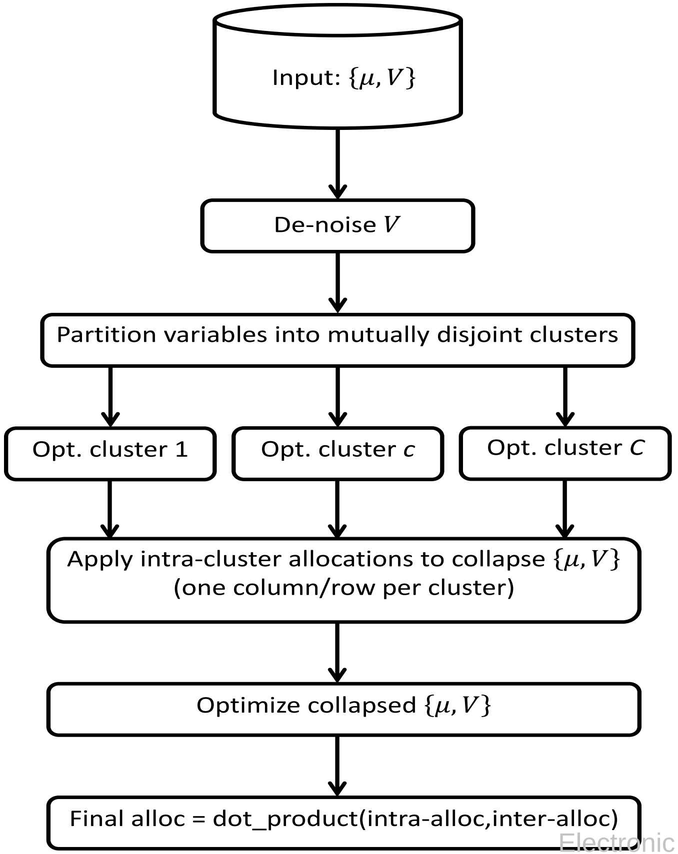
Accordingly, the greater the intra-cluster correlation is, the higher the condition number becomes.

SECTION V
The NCO Algorithm

NCO’s Structure (1/2)

- The Nested Clustered Optimization (NCO) algorithm is composed of five steps:
  1. **Correlation Clustering**: Find the optimal number of clusters.
     - One possibility is to apply the ONC algorithm, however NCO is agnostic as to what particular algorithm is used for determining the number of clusters.
     - For large matrices, where T/N is relatively low, it is advisable to de-noise the correlation matrix prior to clustering, following the method described in Section III.
  2. **Intra-Cluster Weights**: Compute optimal intra-cluster allocations, one optimal allocation per cluster, using the de-noised covariance matrix. This operation can be parallelized.
  3. **System Reduction**: Use the intra-cluster weights to reduce the system (one row/column per cluster).
  4. **Inter-Cluster Weights**: Compute optimal inter-cluster allocations, using the reduced covariance matrix.
  5. **Dot Product**: The final allocation per security results from multiplying intra-cluster weights with the inter-cluster weights.

NCO's Structure (2/2)



<u>Why does NCO beat Markowitz?</u>

By construction, the reduced covariance matrix is close to a diagonal matrix, and the optimization problem is close to the ideal Markowitz case.

In other words, the clustering and intra-cluster optimization steps have allowed us to transform a "Markowitz-cursed" problem ($|\rho| \gg 0$) into a well-behaved problem ($\rho \approx 0$).

Experimental Results (1/2)

- We generate a vector of means and a covariance matrix out of 10 blocks of size 5 each, where off-diagonal elements within each block have a correlation of 0.5.
  - This covariance matrix is a stylized representation of a “true” (non-empirical) de-toned correlation matrix of the S&P 500, where each block is associated with an economic sector.
  - Without loss of generality, the variances are drawn from a uniform distribution bounded between 5% and 20%, and the vector of means is drawn from a Normal distribution with mean and standard deviation equal to the standard deviation from the covariance matrix.
  - This is consistent with the notion that in an efficient market all securities have the same expected Sharpe ratio.
- We use this means vector and covariance matrix to draw 1,000 random matrices X of size $TxN = 1000x50$, compute the associated empirical covariance matrices and vectors of means, and evaluate the (empirical) optimal portfolios.
- We compute the root-mean-square error (RMSE) between the “empirical” and “true” optimal portfolios.

Experimental Results (2/2)

	Markowitz	NCO
Raw	7.95E-03	4.21E-03
Shrunk	8.89E-03	6.74E-03

Minimum Variance Portfolio

	Markowitz	NCO
Raw	7.02E-02	3.17E-02
Shrunk	6.54E-02	5.72E-02

Maximum Sharpe Ratio Portfolio

NCO computes the minimum variance portfolio with 52.98% of Markowitz's RMSE, i.e. a 47.02% reduction in RMSE. While Ledoit-Wolf shrinkage helps reduce the RMSE, that reduction is relatively small, around 11.81%. Combining shrinkage and NCO yields a 15.30% reduction in RMSE, which is better than shrinkage but worse than NCO alone. The implication is that NCO delivers substantially lower RMSE than Markowitz's solution, even for a small portfolio of only 50 securities, and that shrinkage adds no value.

NCO computes the maximum Sharpe ratio portfolio with 45.17% of Markowitz's RMSE, i.e. a 54.83% reduction in RMSE. The combination of shrinkage and NCO yields a 18.52% reduction in the RMSE of the maximum Sharpe ratio portfolio, which is better than shrinkage but worse than NCO. Once again, NCO delivers substantially lower RMSE than Markowitz's solution, and shrinkage adds no value. It is easy to test that NCO's advantage widens for larger portfolios.

SECTION VI
Robustness Analysis via Monte Carlo

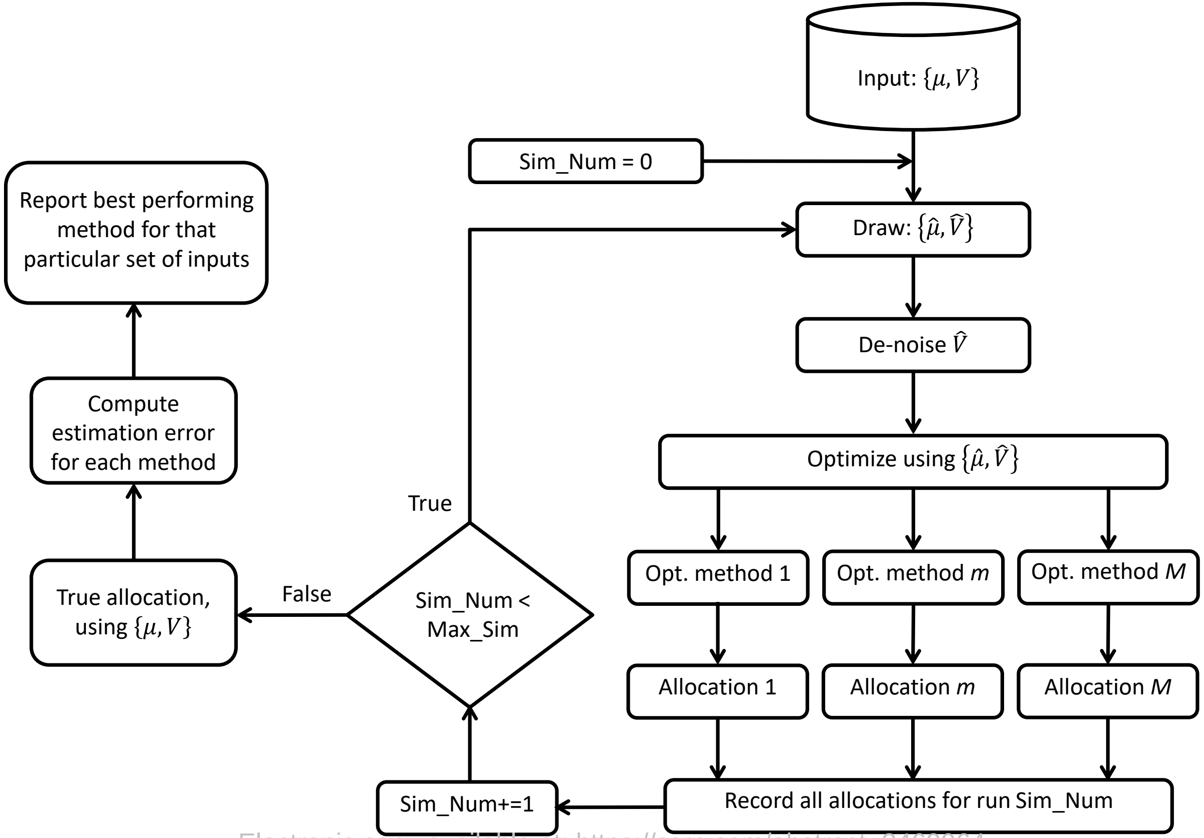
Monte Carlo Optimization Selection (MCOS)

- There is no reason to believe that one particular optimization method is the most robust under all conditions.
- The interactions between noise and signal-induced instabilities make it hard to determine *a priori* what is the most robust optimization approach for a particular problem.
- Thus, rather than relying always on one particular approach, researchers should apply opportunistically whatever optimization method is best suited to a particular setting.
- We introduce a Monte Carlo approach that derives the estimation error produced by various optimization methods on a particular set of input variables.
- The goal is to determine what method is most robust to a particular case.

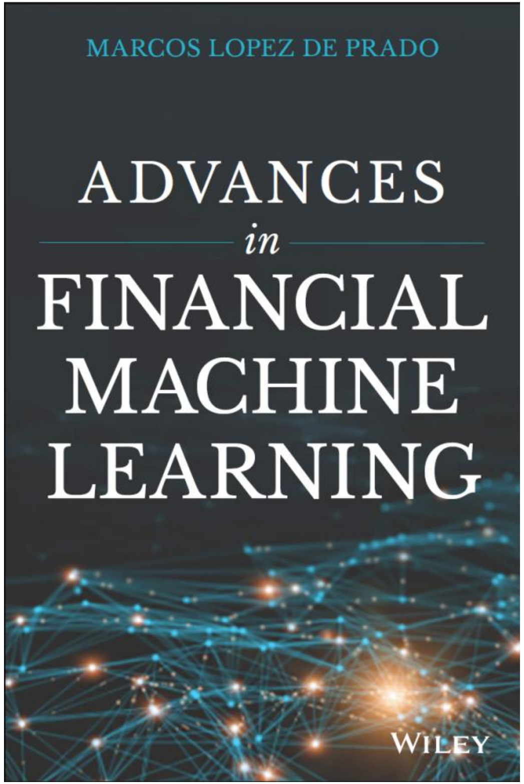
MCOS' Structure (1/2)

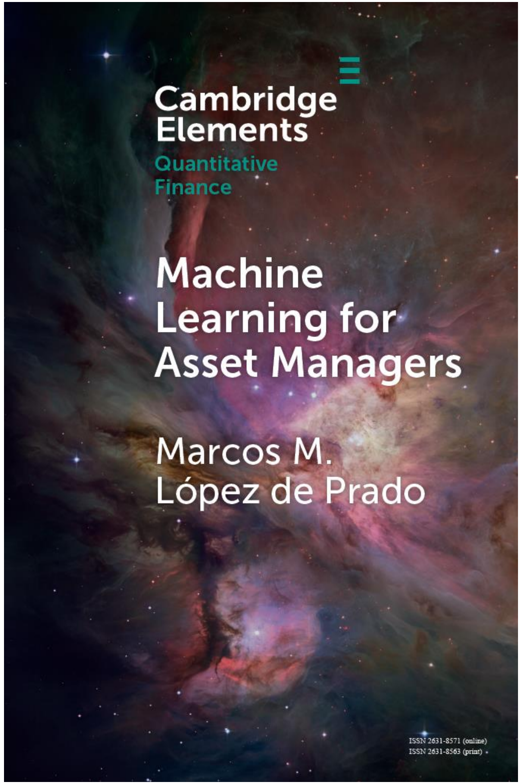
1. **Compute input set**: Get the input variables that characterize the problem, $\{\mu, V\}$.
2. **Random draw**: Draws T observations from the data-generating process characterized by $\{\mu, V\}$, and derive from those observations the pair $\{\hat{\mu}, \hat{V}\}$.
3. **Optimal allocations**: Compute optimal allocations for $\{\hat{\mu}, \hat{V}\}$, applying M alternative methods, resulting in M alternative allocations.
4. **Storage**: Record the M allocations associated with this run of the MCOS method.
5. **Loop**: Repeat steps 1-4 a user-defined number of times.
6. **Benchmark**: Compute the true allocation derived from $\{\mu, V\}$.
7. **Estimation error**: Compute the estimation error associated with each of the M alternative methods.
8. **Report**: Report the method that yields the most robust allocation for the particular set of inputs $\{\mu, V\}$.

MCOS’ Structure (2/2)



For Additional Details





The first wave of quantitative innovation in finance was led by Markowitz optimization. Machine Learning is the second wave and it will touch every aspect of finance. López de Prado's Advances in Financial Machine Learning is essential for readers who want to be ahead of the technology rather than being replaced by it.
— Prof. **Campbell Harvey**, Duke University. Former President of the American Finance Association.

Financial problems require very distinct machine learning solutions. Dr. López de Prado's book is the first one to characterize what makes standard machine learning tools fail when applied to the field of finance, and the first one to provide practical solutions to unique challenges faced by asset managers. Everyone who wants to understand the future of finance should read this book.
— Prof. **Frank Fabozzi**, EDHEC Business School. Editor of The Journal of Portfolio Management.

Disclaimer

- The views expressed in this document are the authors’ and do not necessarily reflect those of the organizations he is affiliated with.
- No investment decision or particular course of action is recommended by this presentation.
-

www.QuantResearch.org