PORTFOLIO THEORY – LECTURE NOTES

Dr. Andrea Rigamonti

MEAN-VARIANCE OPTIMIZATION

From an economist's point of view, an investor that optimizes a portfolio is trying to maximize a utility function. An extremely simple utility function is the **linear utility function**:

$$U(V) = a + bV, \quad b > 0$$

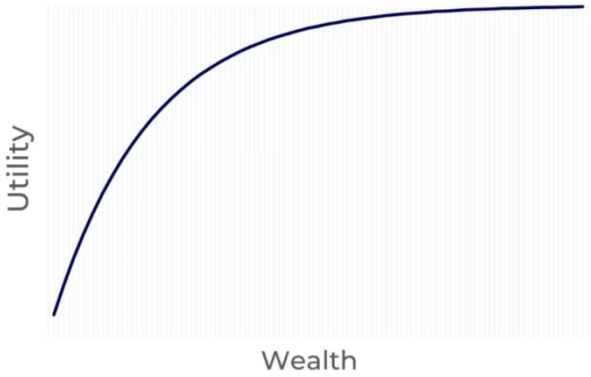
where $U(V)$ is the utility that the investor gets depending on the value V of the portfolio. This function simply says that the higher the wealth, the higher the utility of the investor. Its shape is a line, and the parameter b determines how much the utility increases following a wealth increase. b is assumed be positive, otherwise the investor would be indifferent ($b = 0$) or less satisfied ($b < 0$) when wealth increases. In other words, only the mean return of the portfolio matters.

Markowitz (1952) revolutionized the field by adding risk to the equation. His standard approach assumes that an investor cares not only about the mean but also the variance of the portfolio returns, i.e. the investor has a **mean-variance utility**. Given a certain mean return, the utility increases as the variance (which quantifies risk) gets lower. Equivalently, given a certain level of variance, the utility increases with a higher mean return. The decision in the trade-off between return and risk is quantified by a **risk aversion parameter** γ. A higher γ means that the investor is more risk averse and will therefore require a higher compensation for an increased risk. A lower γ means the investor has a lower risk-aversion and will be willing to take more risks. In other words, given a set of assets with a certain mean and variance, the lower the γ of the investor, the more he will create an optimal portfolio with a higher mean return but also a higher variance.

To model such preferences, a **quadratic utility function** is used:

$$U(V) = V - \frac{\gamma}{2}V^2, \quad \gamma > 0$$

γ is assumed to be positive so that the utility function is concave:



Source: https://financestu.com

This implies that the investor is risk-averse. With $\gamma = 0$ the investor would be indifferent to risk, while $\gamma < 0$ would mean that the investor is risk taker (i.e., prefers more risk for the same amount of wealth, which is obviously not realistic).

Remember that the expected return of a portfolio is given by $\mu_P = \boldsymbol{w}'\boldsymbol{\mu} = V$, where $\boldsymbol{w}$ is the vector of portfolio weights and $\boldsymbol{\mu}$ is the vector of mean return of the single assets. Moreover, recall that the variance is the expected value of the squared deviation from the mean.

So, the utility function becomes:

$$U(\boldsymbol{w}) = \boldsymbol{w}'\boldsymbol{\mu} - \frac{\gamma}{2}\boldsymbol{w}'\Sigma\boldsymbol{w}$$

Therefore, given a risk-free asset and a set of N risky assets with mean returns $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and a certain risk aversion coefficient γ, the investor we are considering wants to select the weights $\boldsymbol{w}$ in a way that maximizes the following utility function:

$$\max_{\boldsymbol{w}} \boldsymbol{w}'\boldsymbol{\mu} - \frac{\gamma}{2}\boldsymbol{w}'\Sigma\boldsymbol{w}$$

This is an unconstrained optimization problem easy to solve. We just need to set the first-order condition, i.e., take the partial derivative with respect to $\boldsymbol{w}$ and set it equal to zero:

$$\frac{\partial U(\boldsymbol{w})}{\partial \boldsymbol{w}} = \boldsymbol{\mu} - \frac{2\gamma}{2}\Sigma\boldsymbol{w} = \boldsymbol{\mu} - \gamma\Sigma\boldsymbol{w} = \mathbf{0}$$

We then solve for $\boldsymbol{w}$:

$$\Sigma\boldsymbol{w} = \frac{1}{\gamma}\boldsymbol{\mu}$$

$$\boldsymbol{w} = \frac{1}{\gamma}\Sigma^{-\mathbf{1}}\boldsymbol{\mu}$$

So, the closed-form solution that gives the optimal weights for the risky assets is:

$$\boldsymbol{w_U} = \frac{1}{\gamma}\Sigma^{-\mathbf{1}}\boldsymbol{\mu}$$

where "U" stands for "Utility", while the weight for the risk-free asset is equal to $1 - \boldsymbol{w_U}'\mathbf{1}$, where $\mathbf{1}$ is a vector of ones with length equal to the number of risky assets.

The resulting optimal expected utility is:

$$U(\boldsymbol{w_U}) = \frac{1}{2\gamma}\boldsymbol{\mu}'\Sigma^{-\mathbf{1}}\boldsymbol{\mu}$$

Notice that we do not need to explicitly include the risk-free asset in the asset menu, as it is equivalent and simpler to work with excess returns, i.e. with the returns of the risky assets from which we subtracted the risk-free rate.

A more difficult version is the one where we impose a **full investment in the risky assets**. In other words, the sum of the weights of the risky assets must be equal to 1, and nothing is invested in the risk-free asset. Hence, we have to solve the following constrained optimization problem:

$$\max_{\boldsymbol{w}} \boldsymbol{w}'\boldsymbol{\mu} - \frac{\gamma}{2}\boldsymbol{w}'\Sigma\boldsymbol{w}$$

subject to:

$$\boldsymbol{w}'\mathbf{1} = 1$$

To solve this problem, we use the method of Lagrange multipliers.

First, we need to define the Lagrangian function, i.e., a modified version of the objective function that incorporates the constraint in this way:

$$L(\boldsymbol{w}, \lambda) = \boldsymbol{w}'\boldsymbol{\mu} - \frac{\gamma}{2}\boldsymbol{w}'\Sigma\boldsymbol{w} + \lambda[1 - \boldsymbol{w}'\mathbf{1}]$$

where λ is the Lagrange multiplier.

By including this additional term we can now solve an unconstrained problem instead of a constrained one. Therefore, we set the first order conditions for the Lagrangian function. The conditions involve two simultaneous equations, as we have to compute the partial derivative both with respect to $\boldsymbol{w}$ and to λ.

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{\mu} - \frac{2\gamma}{2}\Sigma\boldsymbol{w} - \lambda\mathbf{1} = \boldsymbol{\mu} - \gamma\Sigma\boldsymbol{w} - \lambda\mathbf{1} = \mathbf{0}$$

$$\frac{\partial L}{\partial \lambda} = 1 - \boldsymbol{w}'\mathbf{1} = 0$$

We start by solving the first equation for $\boldsymbol{w}$:

$$\gamma\Sigma\boldsymbol{w} = \boldsymbol{\mu} - \lambda\mathbf{1}$$

$$\boldsymbol{w} = (\gamma\Sigma)^{-1}(\boldsymbol{\mu} - \lambda\mathbf{1})$$

Now we can plug this into the second equation:

$$1 - [(\gamma\Sigma)^{-1}(\boldsymbol{\mu} - \lambda\mathbf{1})]'\mathbf{1} = 0$$

Remember that $(\boldsymbol{AB})' = \boldsymbol{B}'\boldsymbol{A}'$, and therefore we have:

$$(\boldsymbol{\mu} - \lambda\mathbf{1})'((\gamma\Sigma)^{-1})'\mathbf{1} = 1$$

$$(\boldsymbol{\mu} - \lambda\mathbf{1})'\left(\frac{\Sigma^{-1}}{\gamma}\right)'\mathbf{1} = 1$$

The transpose of the sum of matrices (or vectors) is the sum of the transpose of those matrices: $(\boldsymbol{A} + \boldsymbol{B})' = \boldsymbol{A}' + \boldsymbol{B}'$. Moreover, a scalar is unaffected by the transpose: $(c\boldsymbol{A})' = c\boldsymbol{A}'$. Also notice that Σ^{-1} is a symmetric matrix, and therefore its transpose is still Σ^{-1}. Therefore:

$$(\boldsymbol{\mu}' - \lambda\mathbf{1}')\frac{\Sigma^{-1}}{\gamma}\mathbf{1} = 1$$

$$(\boldsymbol{\mu}' - \lambda\mathbf{1}')\Sigma^{-1}\mathbf{1} = \gamma$$

$$\boldsymbol{\mu}'\Sigma^{-1}\mathbf{1} - \lambda\mathbf{1}'\Sigma^{-1}\mathbf{1} = \gamma$$

$$\lambda = \frac{\boldsymbol{\mu}'\Sigma^{-1}\mathbf{1} - \gamma}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}$$

We can now plug this into the first equation, finally obtaining the solution we were looking for:

$$\boldsymbol{w} = (\gamma\Sigma)^{-1}(\boldsymbol{\mu} - \lambda\mathbf{1})$$

$$\boldsymbol{w} = (\gamma\Sigma)^{-1}\left(\boldsymbol{\mu} - \frac{\boldsymbol{\mu}'\Sigma^{-1}\mathbf{1} - \gamma}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}\mathbf{1}\right)$$

With some minimal rearrangement, we have therefore the following set of optimal weights that maximize the mean-variance utility given the constraint of full investment in the risky assets:

$$\boldsymbol{w}_{U^*} = \frac{\Sigma^{-1}}{\gamma}\left(\boldsymbol{\mu} + \frac{\gamma - \boldsymbol{\mu}'\Sigma^{-1}\mathbf{1}}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}\mathbf{1}\right)$$

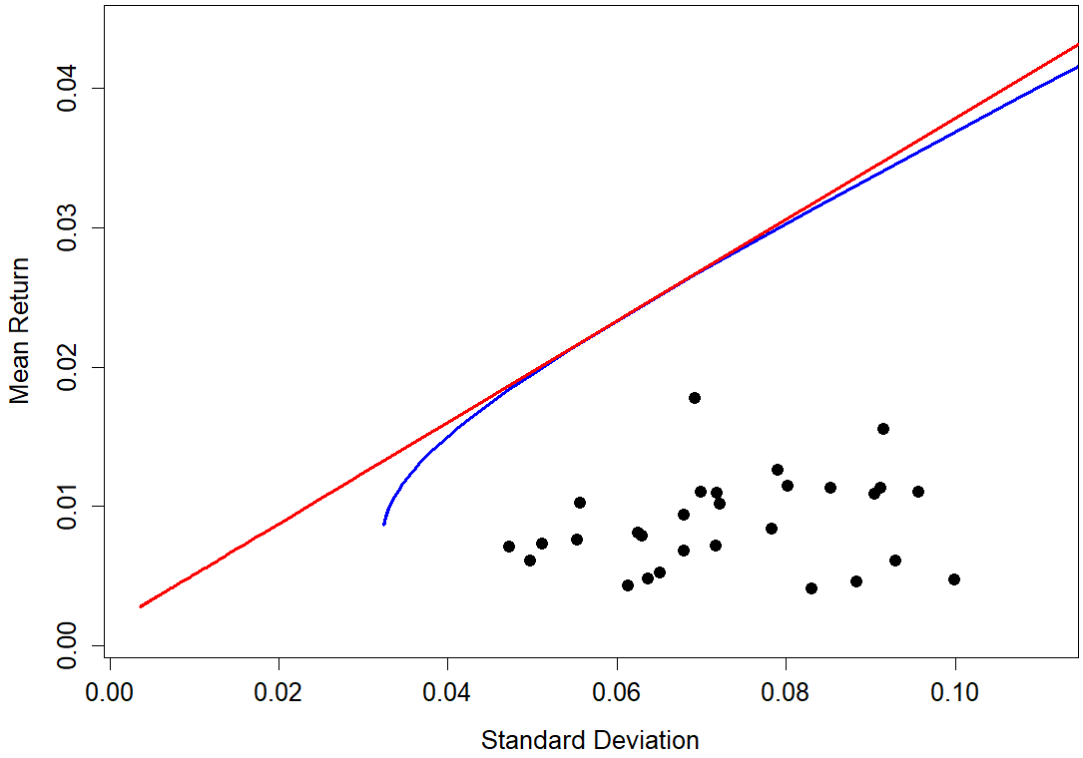
Given a vector of mean $\boldsymbol{\mu}$ and a covariance matrix Σ, there will be different sets of optimal weights $\boldsymbol{w}_U$ and $\boldsymbol{w}_{U^*}$ for each different value of the risk-aversion parameter γ. However, while in $\boldsymbol{w}_{U^*}$ the relative wealth allocated to each risky asset changes with each different value of γ, in $\boldsymbol{w}_U$ all the weights scale up or down by the same proportion. In other words, when there is a risk-free asset,

the only thing changing with different values of γ is the amount of wealth allocated to the risky assets, while the proportions within $\boldsymbol{w_U}$ do not change.

For example, imagine we have three risky assets and a risk-free asset, and the optimal weights with $\gamma = 3$ are $\boldsymbol{w_U'} = [0.2 \quad 0.4 \quad 0.3]$. This means that 90% of the wealth is invested in the risky assets, and 10% in the risk-free asset. If the risk aversion parameter increases to $\gamma = 6$, the new risky weights will be $\boldsymbol{w_U'} = [0.1 \quad 0.2 \quad 0.15]$. Now 45% of the wealth is invested in the risky assets and 55% in the risk-free. However, the weights for the risky assets maintained the same proportions: they were all cut in half.

With the formulas that we computed, we can obtain the **efficient frontier**, i.e., the set of portfolios that have the most efficient mean-variance combination (in other words, the highest possible utility) for each level of γ. However, the efficient frontier that we can draw in this way is incomplete, as the minimum variance portfolio is only reached with an infinite value for the risk aversion parameter.

The following picture provides an illustration of the efficient frontier obtained from 29 asset with (red line) and without (blue curve) the possibility to also invest in a risk-free asset. The black dots are the individual stocks that have been combined to obtain the portfolio allocations that lie on the efficient frontier(s).



The higher γ, the more the investor chooses an investor toward the left side of the plot, i.e., with lower mean and lower standard deviation.

We could also extend this plot by plotting the frontier allocations obtained with a negative γ. In this case the red line and the blue curve would be horizontally mirrored, and have therefore a negative inclination. Such allocation would be seek by an investor who is risk-seeking, i.e., who prefers more risk for a given mean return. As this is unreasonable and not "efficient" in any sensible way, those allocations are generally not considered and are not part of the efficient frontier.

Providing a specific value for γ might be difficult in practice. Moreover, it can be difficult to interpret the meaning of the specific utility value associated with a certain portfolio. While we can say that a portfolio with a certain utility is preferable to another with a lower utility, it is not obvious how good it is in a more general sense. In short, the value itself, in the case of utility, is not very informative.

A more intuitive approach is to specify the preferences through a desired mean portfolio return Re instead of a level of risk aversion. In this case, the goal becomes to **minimize the variance given the desired mean return**. This is a constrained optimization problem:

$$\min_{\boldsymbol{w}} \boldsymbol{w}'\Sigma\boldsymbol{w}$$

subject to:

$$\boldsymbol{w}'\boldsymbol{\mu} + (1 - \boldsymbol{w}'\mathbf{1})R_f = Re$$

In the constraint, $\boldsymbol{w}'\boldsymbol{\mu}$ is the return of the risky assets, and $(1 - \boldsymbol{w}'\mathbf{1})R_f$ is the return of the risk-free asset. Together they give the return of the portfolio, which, as stated, must be equal to Re.

As with mean-variance utility maximization, it is possible (and preferable) to work with excess returns instead of explicitly including the risk-free asset in the asset menu. In this case Re is the desired excess return, and the problem simplifies to:

$$\min_{\boldsymbol{w}} \boldsymbol{w}'\Sigma\boldsymbol{w}$$

subject to:

$$\boldsymbol{w}'\boldsymbol{\mu} = Re$$

The Lagrangian function is:

$$L(\boldsymbol{w}, \lambda) = \boldsymbol{w}'\Sigma\boldsymbol{w} + \lambda[Re - \boldsymbol{w}'\boldsymbol{\mu}]$$

In order to get a more convenient first order condition, it is common to multiply the first term by 0.5, which does not alter the result (because minimizing half the variance is equivalent to minimizing the variance). So the Lagrangian becomes:

$$L(\boldsymbol{w}, \lambda) = \frac{1}{2}\boldsymbol{w}'\Sigma\boldsymbol{w} + \lambda[Re - \boldsymbol{w}'\boldsymbol{\mu}]$$

The first order conditions are:

$$\frac{\partial L}{\partial \boldsymbol{w}} = \Sigma\boldsymbol{w} - \lambda\boldsymbol{\mu} = \mathbf{0}$$
$$\frac{\partial L}{\partial \lambda} = Re - \boldsymbol{w}'\boldsymbol{\mu} = 0$$

We start by solving for $\boldsymbol{w}$ in the first equation:

$$\boldsymbol{w} = \lambda\Sigma^{-1}\boldsymbol{\mu}$$

Then we plug it into the second equation and we solve for λ:

$$Re = \boldsymbol{w}'\boldsymbol{\mu}$$
$$Re = (\lambda\Sigma^{-1}\boldsymbol{\mu})'\boldsymbol{\mu}$$
$$Re = \lambda\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}$$
$$\lambda = \frac{Re}{\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}}$$

We can now substitute this back in the first equation and we get the solution we wanted:

$$\boldsymbol{w} = \frac{Re}{\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}}\Sigma^{-1}\boldsymbol{\mu}$$

Using the subscript “mv” (for “mean-variance”) to univocally identify the formula, we have:

$$\boldsymbol{w_{mv}} = \frac{Re}{\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}}\Sigma^{-1}\boldsymbol{\mu}$$

Obviously, it is also possible to specify a given level of variance and maximize the mean return:

$$\max_{\boldsymbol{w}} \boldsymbol{w}'\boldsymbol{\mu}$$

subject to:

$$\boldsymbol{w}'\Sigma\boldsymbol{w} = \sigma^2$$

We write the Lagrangian and the first order conditions:

$$L(\boldsymbol{w}, \lambda) = \boldsymbol{w}'\boldsymbol{\mu} + \lambda[\sigma^2 - \boldsymbol{w}'\Sigma\boldsymbol{w}]$$

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{\mu} - \boldsymbol{2}\lambda\Sigma\boldsymbol{w} = \boldsymbol{0}$$

$$\frac{\partial L}{\partial \lambda} = \sigma^2 - \boldsymbol{w}'\Sigma\boldsymbol{w} = 0$$

We solve the first equation for $\boldsymbol{w}$:

$$\boldsymbol{2}\lambda\Sigma\boldsymbol{w} = \boldsymbol{\mu}$$

$$\boldsymbol{w} = \frac{\Sigma^{-\boldsymbol{1}}\boldsymbol{\mu}}{2\lambda}$$

Now we plug into the second equation:

$$\sigma^2 = \boldsymbol{w}'\Sigma\boldsymbol{w}$$

$$\sigma^2 = \left(\frac{\Sigma^{-\boldsymbol{1}}\boldsymbol{\mu}}{2\lambda}\right)' \Sigma \frac{\Sigma^{-\boldsymbol{1}}\boldsymbol{\mu}}{2\lambda} = \frac{\boldsymbol{\mu}'\Sigma^{-\boldsymbol{1}}}{2\lambda}\Sigma\frac{\Sigma^{-\boldsymbol{1}}\boldsymbol{\mu}}{2\lambda}$$

$$\sigma^2 = \frac{\boldsymbol{\mu}'\boldsymbol{I}}{2\lambda}\frac{\Sigma^{-\boldsymbol{1}}\boldsymbol{\mu}}{2\lambda} = \frac{\boldsymbol{\mu}'\Sigma^{-\boldsymbol{1}}\boldsymbol{\mu}}{4\lambda^2}$$

$$\sigma = \frac{\sqrt{\boldsymbol{\mu}'\Sigma^{-\boldsymbol{1}}\boldsymbol{\mu}}}{2\lambda}$$

$$\lambda = \frac{\sqrt{\boldsymbol{\mu}'\Sigma^{-\boldsymbol{1}}\boldsymbol{\mu}}}{2\sigma}$$

And finally we replace this in the first equation to get the solution:

$$\boldsymbol{w} = \frac{\Sigma^{-\boldsymbol{1}}\boldsymbol{\mu}}{2\lambda} = \frac{\Sigma^{-\boldsymbol{1}}\boldsymbol{\mu}}{2\left(\frac{\sqrt{\boldsymbol{\mu}'\Sigma^{-\boldsymbol{1}}\boldsymbol{\mu}}}{2\sigma}\right)} = \frac{\Sigma^{-\boldsymbol{1}}\boldsymbol{\mu}}{\sqrt{\boldsymbol{\mu}'\Sigma^{-\boldsymbol{1}}\boldsymbol{\mu}}}\sigma = \frac{\sigma}{\sqrt{\boldsymbol{\mu}'\Sigma^{-\boldsymbol{1}}\boldsymbol{\mu}}}\Sigma^{-\boldsymbol{1}}\boldsymbol{\mu}$$

Mathematically, these are two equivalent optimization problems (notice the obvious similarities between the two solutions). However, it is more intuitive and more common to specify the desired mean and minimize the variance.

Also for this problem we can have a more complicated version with the additional **constraint of weights for the risky assets summing up to 1**:

$$\min_{\boldsymbol{w}} \boldsymbol{w}'\Sigma\boldsymbol{w}$$

subject to:

$$\boldsymbol{w}'\boldsymbol{\mu} = Re$$

$$\boldsymbol{w}'\mathbf{1} = 1$$

As usual, we have to write the Lagrangian. As there are two constraints, this time there are two Lagrange multipliers:

$$L(\boldsymbol{w}, \lambda_1, \lambda_2) = \frac{1}{2}\boldsymbol{w}'\Sigma\boldsymbol{w} + \lambda_1[Re - \boldsymbol{w}'\boldsymbol{\mu}] + \lambda_2[1 - \boldsymbol{w}'\mathbf{1}]$$

The first order conditions are:

$$\frac{\partial L}{\partial \boldsymbol{w}} = \Sigma\boldsymbol{w} - \lambda_1\boldsymbol{\mu} - \lambda_2\mathbf{1} = \mathbf{0}$$

$$\frac{\partial L}{\partial \lambda_1} = Re - \boldsymbol{w}'\boldsymbol{\mu} = 0$$

$$\frac{\partial L}{\partial \lambda_2} = 1 - \boldsymbol{w}'\mathbf{1} = 0$$

We solve the first equation for $\boldsymbol{w}$:

$$\Sigma\boldsymbol{w} - \lambda_1\boldsymbol{\mu} - \lambda_2\mathbf{1} = \mathbf{0}$$

$$\Sigma\boldsymbol{w} = \lambda_1\boldsymbol{\mu} + \lambda_2\mathbf{1}$$

$$\boldsymbol{w} = \Sigma^{-1}(\lambda_1\boldsymbol{\mu} + \lambda_2\mathbf{1})$$

$$\boldsymbol{w} = \lambda_1\Sigma^{-1}\boldsymbol{\mu} + \lambda_2\Sigma^{-1}\mathbf{1}$$

We need to get a formula for λ_1 and one for λ_2 that do not contain each other among their terms.

To this end, notice that if we pre-multiply each side of the equation by $\boldsymbol{\mu}'$ we get

$$\boldsymbol{\mu}'\boldsymbol{w} = \lambda_1\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu} + \lambda_2\boldsymbol{\mu}'\Sigma^{-1}\mathbf{1}$$

$$Re = \lambda_1\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu} + \lambda_2\boldsymbol{\mu}'\Sigma^{-1}\mathbf{1}$$

Likewise, if we pre-multiply each side by $\mathbf{1}'$ we get

$$\mathbf{1}'\boldsymbol{w} = \lambda_1\mathbf{1}'\Sigma^{-1}\boldsymbol{\mu} + \lambda_2\mathbf{1}'\Sigma^{-1}\mathbf{1}$$

$$1 = \lambda_1\mathbf{1}'\Sigma^{-1}\boldsymbol{\mu} + \lambda_2\mathbf{1}'\Sigma^{-1}\mathbf{1}$$

Therefore, we get a system of two equations:

$$Re = \lambda_1\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu} + \lambda_2\boldsymbol{\mu}'\Sigma^{-1}\mathbf{1}$$

$$1 = \lambda_1\mathbf{1}'\Sigma^{-1}\boldsymbol{\mu} + \lambda_2\mathbf{1}'\Sigma^{-1}\mathbf{1}$$

Notice that $\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}$, $\boldsymbol{\mu}'\Sigma^{-1}\mathbf{1}$, $\mathbf{1}'\Sigma^{-1}\boldsymbol{\mu}$ and $\mathbf{1}'\Sigma^{-1}\mathbf{1}$ are scalars. For conveniency, we name them as

$$A = \boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu} \qquad B = \boldsymbol{\mu}'\Sigma^{-1}\mathbf{1} = \mathbf{1}'\Sigma^{-1}\boldsymbol{\mu} \qquad C = \mathbf{1}'\Sigma^{-1}\mathbf{1}$$

So the system of two equations is

$$\lambda_1 A + \lambda_2 B = Re$$

$$\lambda_1 B + \lambda_2 C = 1$$

which in matrix form is

$$\begin{bmatrix} A & B \\ B & C \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} Re \\ 1 \end{bmatrix}$$

We solve the system for λ_1 and λ_2:

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} A & B \\ B & C \end{bmatrix}^{-1} \begin{bmatrix} Re \\ 1 \end{bmatrix}$$

The inverse of 2×2 matrix $\boldsymbol{M} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is given by a simple formula:

$$\boldsymbol{M^{-1}} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Hence, our system becomes

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \frac{1}{AC - B^2} \begin{bmatrix} C & -B \\ -B & A \end{bmatrix} \begin{bmatrix} Re \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \frac{1}{AC - B^2} \begin{bmatrix} CRe - B \\ -BRe + A \end{bmatrix}$$

So we have obtained the formulas we were looking for, which written in plain form are

$$\lambda_1 = \frac{CRe - B}{AC - B^2}$$

$$\lambda_2 = \frac{A - BRe}{AC - B^2}$$

We can finally plug these terms back in the first equation we obtained for $\boldsymbol{w}$:

$$\boldsymbol{w} = \lambda_1 \Sigma^{-1} \boldsymbol{\mu} + \lambda_2 \Sigma^{-1} \mathbf{1}$$

$$\boldsymbol{w} = \frac{CRe - B}{AC - B^2} \Sigma^{-1} \boldsymbol{\mu} + \frac{A - BRe}{AC - B^2} \Sigma^{-1} \mathbf{1}$$

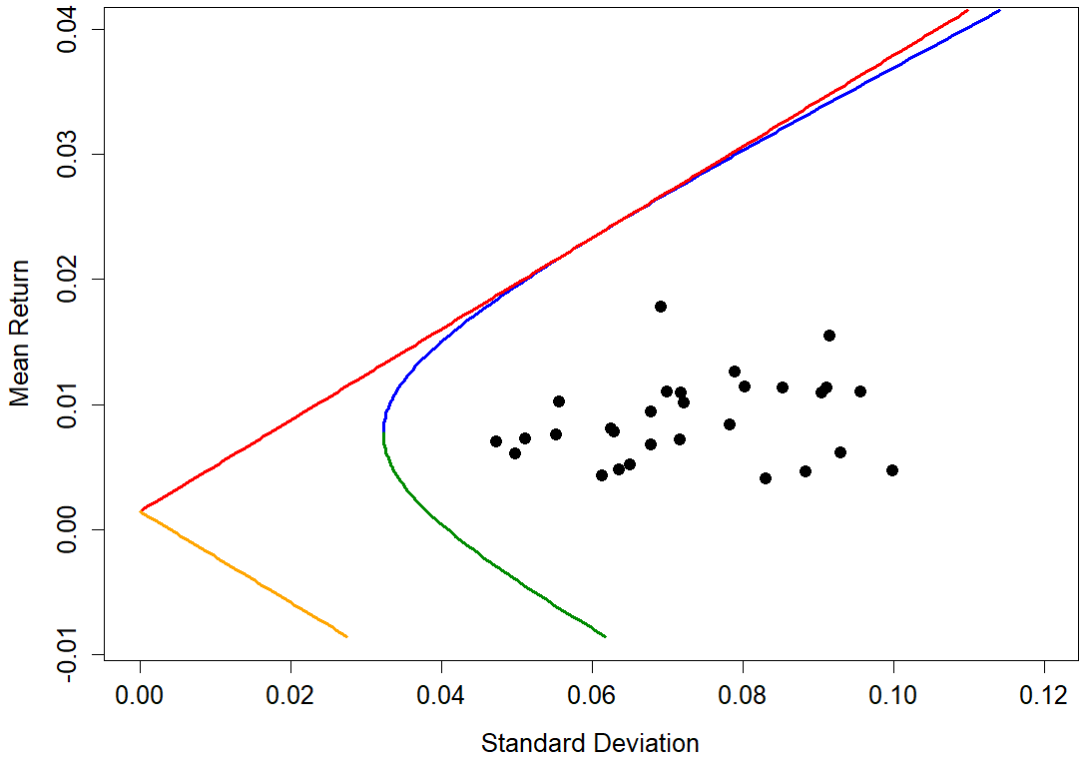
Hence, with some minimal rearrangement, the set of optimal weights that minimize the variance given a target return and the constraint of full investment in the risky assets is:

$$\boldsymbol{w_{mv^*}} = \Sigma^{-1} \left[\frac{CRe - B}{AC - B^2} \boldsymbol{\mu} + \frac{A - BRe}{AC - B^2} \mathbf{1} \right]$$

where $A = \boldsymbol{\mu}' \Sigma^{-1} \boldsymbol{\mu}$, $B = \mathbf{1}' \Sigma^{-1} \boldsymbol{\mu}$ and $C = \mathbf{1}' \Sigma^{-1} \mathbf{1}$.

Analogously to what happens with the weights that maximize the utility, given a certain $\boldsymbol{\mu}$ and Σ, changing the target return alters the relative wealth allocation between the risky assets in $\boldsymbol{w_{mv^*}}$, but in $\boldsymbol{w_{mv}}$ only the value of the sum of the weights of the risky assets changes, while the proportions stay the same.

By applying with different target returns the formulas that we computed, we can obtain the **efficient frontier**, which again will be a line when it is possible to invest also in a risk-free asset, or a curve if all the wealth must be placed in the risky assets. As with the one that can be obtained by maximizing utility, the full frontier also includes inefficient allocations. In the picture below, obtained from 29 assets, we plot in red and blue the efficient frontier with and without a risk-free asset respectively, and in orange and green the inefficient frontier allocations. The black dots are individual stocks.



The frontier obtainable by minimizing the variance given a target return (or by maximizing the mean return given a target variance) is, of course, identical to the one obtainable by maximizing the mean-variance utility. In both cases we are facing the same trade-off between mean and variance and, given a certain $\boldsymbol{\mu}$ and Σ, the set of achievable efficient portfolios is the same. The problem of maximizing utility is a bit easier to solve because the risk-aversion parameter γ directly tells us how to weigh mean and variance in this trade-off. The value γ is however not very meaningful in practice, and so in a practical setting we need to solve the slightly more complex (but based on the same premises) problem of minimizing the variance given a target mean or vice versa.

As you can see, there is a point at which the efficient frontiers with and without a risk-free asset touch each other. That corresponds to the mean and standard deviation of the **tangency portfolio**. This portfolio is the one portfolio of risky assets which has the highest possible Sharpe ratio. The **Sharpe ratio** of an investment can be geometrically interpreted as the slope of the line that connects the risk-free rate with an investment in the plot above. Therefore, the highest Sharpe ratio can always be achieved with the optimal weights given by the optimization procedures with a risk-free asset that we derived. The weights $\boldsymbol{w_U}$ or $\boldsymbol{w_{mv}}$ are used for the risky assets, and $1 - \boldsymbol{w_U}'\mathbf{1}$ or $1 - \boldsymbol{w_{mw}}'\mathbf{1}$ is the weigh for the risk-free asset (which can also be negative). This result (i.e., that the combination of the tangency portfolio and a risk-free asset gives the highest utility for a given risk-aversion level) is known as **two-fund separation theorem**, and dates back to Tobin (1958). It played a large role in the development of CAPM, where the tangency portfolio is identified (under a series or rather stringent assumptions) as the market portfolio.

But what if we want or can invest only in the risky assets, and we still want to achieve the highest possible Sharpe ratio? In other words, how can we compute the weights for the tangency portfolio?

To this end, working with excess returns, we need to solve the following optimization problem:

$$\max_{\boldsymbol{w}} \frac{\boldsymbol{w}'\boldsymbol{\mu}}{\sqrt{\boldsymbol{w}'\Sigma\boldsymbol{w}}}$$

subject to:

$$\boldsymbol{w}'\mathbf{1} = 1$$

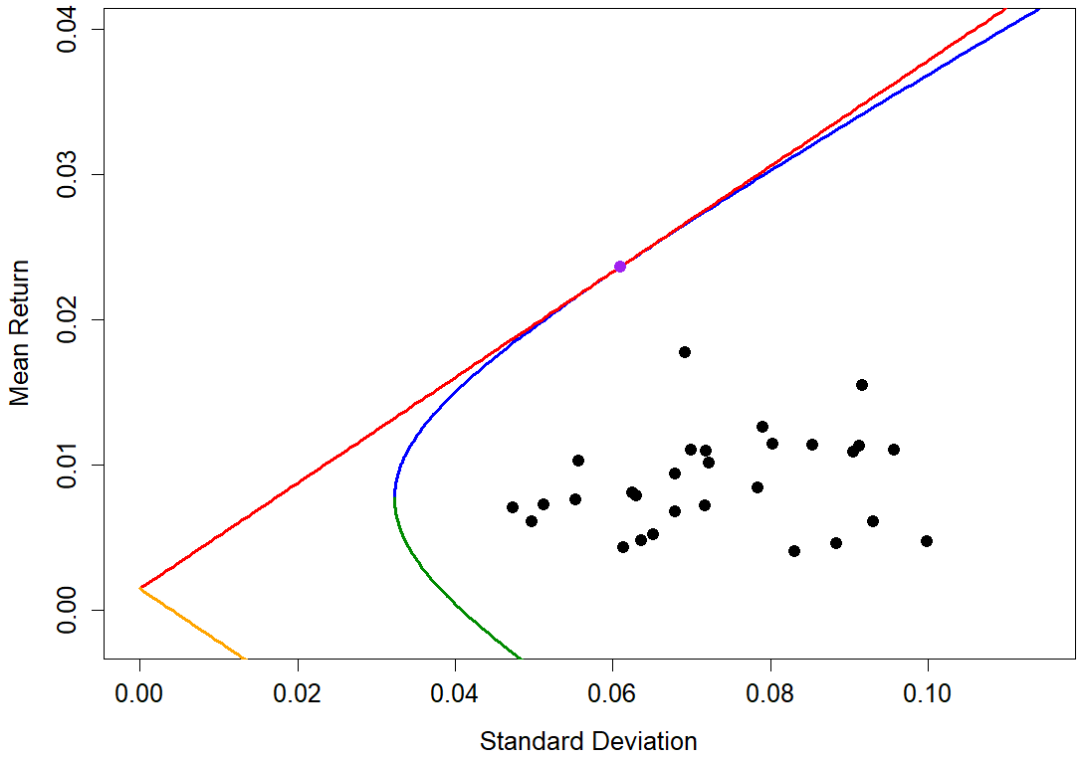
This can be solved with the usual Lagrange multiplier method, but the computations are rather long and intricate, so we do not show them. The resulting closed form solution is:

$$\boldsymbol{w}_{tan} = \frac{\Sigma^{-1}\boldsymbol{\mu}}{\mathbf{1}'\Sigma^{-1}\boldsymbol{\mu}}$$

However, there is another very easy way to arrive at this solution. Notice that the weights of the tangency portfolio are simply the weights of the portfolio that maximizes the mean-variance utility with any given value of γ in the presence of the risk-free asset, normalized so that they sum to 1. This is because $\boldsymbol{w}_U$ (and $\boldsymbol{w}_{mv}$) are in fact just the weights of the tangency portfolio scaled according to the risk preferences of the investor: they sum to less than 1 or more than 1 depending on how risk-averse is the investor, but their proportions do not change. Therefore, we can simply take the formula for $\boldsymbol{w}_U$ and divide it by its own sum:

$$\boldsymbol{w}_{tan} = \frac{\frac{1}{\gamma}\Sigma^{-1}\boldsymbol{\mu}}{\mathbf{1}'\left(\frac{1}{\gamma}\Sigma^{-1}\boldsymbol{\mu}\right)} = \frac{\Sigma^{-1}\boldsymbol{\mu}}{\mathbf{1}'\Sigma^{-1}\boldsymbol{\mu}}$$

We plot in purple the tangency portfolio in the graph with the frontier:



GLOBAL MINIMUM VARIANCE AND EQUALLY WEIGHTED PORTFOLIO

So far we treated the inputs $\boldsymbol{\mu}$ and Σ as if they are given, but in practice they need to be estimated. The simplest approach is called **plug-in approach**: the sample estimates of the inputs are computed from past data, and are then plugged into the optimization problem as if they were the true values. Obviously, this is not really the case, as sample estimates can be poor estimates of the true parameter values. This typically causes theoretically optimal mean-variance optimized portfolios to perform poorly out-of-sample.

In particular, the estimation error in the sample mean is typically so big that minimizing the variance while ignoring $\boldsymbol{\mu}$ usually leads to portfolios with a higher Sharpe ratio than those computed via mean-variance optimization. Therefore, the investor might want to compute the **global minimum variance portfolio (GMV)**, also simply called **minimum variance portfolio**. Obviously, we need to impose the constraint that the sum of the weights of the risky assets is equal to 1, which means that nothing is invested in the risk-free asset. Otherwise, everything would be invested in the risk-free asset. Hence, we have to solve the following constrained optimization problem:

$$\min_{\boldsymbol{w}} \boldsymbol{w}'\Sigma\boldsymbol{w}$$

subject to:

$$\boldsymbol{w}'\mathbf{1} = 1$$

We write the Lagrangian function:

$$L(\boldsymbol{w}, \lambda) = \boldsymbol{w}'\Sigma\boldsymbol{w} + \lambda[1 - \boldsymbol{w}'\mathbf{1}]$$

As done before, we multiply the first term by 0.5, so the Lagrangian becomes:

$$L(\boldsymbol{w}, \lambda) = \frac{1}{2}\boldsymbol{w}'\Sigma\boldsymbol{w} + \lambda[1 - \boldsymbol{w}'\mathbf{1}]$$

The first order conditions are:

$$\frac{\partial L}{\partial \boldsymbol{w}} = \Sigma\boldsymbol{w} - \lambda\mathbf{1} = \mathbf{0}$$

$$\frac{\partial L}{\partial \lambda} = 1 - \boldsymbol{w}'\mathbf{1} = 0$$

Through some simple rearrangement we get:

$$\boldsymbol{w} = \lambda\Sigma^{-1}\mathbf{1}$$

$$\boldsymbol{w}'\mathbf{1} = 1$$

In the first equation, we can multiply both sides by $\mathbf{1}'$**,** obtaining:

$$\mathbf{1}'\boldsymbol{w} = \lambda\mathbf{1}'\Sigma^{-1}\mathbf{1}$$

From the second equation we know that:

$$\boldsymbol{w}'\mathbf{1} = \mathbf{1}'\boldsymbol{w} = 1$$

Hence, the first equation becomes:

$$1 = \lambda\mathbf{1}'\Sigma^{-1}\mathbf{1}$$

$$\lambda = \frac{1}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}$$

So, finally, we can take this last result and replace λ in $\boldsymbol{w} = \lambda\Sigma^{-1}\mathbf{1}$, obtaining:

$$\boldsymbol{w} = \frac{1}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}\Sigma^{-1}\mathbf{1}$$

Therefore, the closed form-solution that gives the minimum variance weights is:

$$\boldsymbol{w}_v = \frac{1}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}\Sigma^{-1}\mathbf{1}$$

As nothing is invested in the risk-free rate (since we require that the weights for the risky assets must sum up to 1), it is equivalent to work with returns or excess returns. However, it might be convenient to still work with excess returns, so that the results will be easily comparable with those obtained by the mean-variance portfolio.

A further improvement can come from restricting the minimum variance portfolio to only have long positions. In other words, we add another constraint that prohibits short selling positions, to get a **long-only minimum variance portfolio**:

$$\min_{\boldsymbol{w}} \boldsymbol{w}'\Sigma\boldsymbol{w}$$

subject to:

$$\boldsymbol{w}'\mathbf{1} = 1$$

$$\boldsymbol{w} \geq \mathbf{0}$$

This problem does not have a closed form solution, but it is a quadratic programming problem (i.e., a problem with a quadratic objective function subject to linear constraints) that can easily be solved with computer programs using various algorithms.

Notice that this solution is theoretically sub-optimal: if we actually knew the true parameters, it would lead to a loss compared to the unconstrained minimum variance portfolio (which is itself already theoretically sub-optimal compared to mean-variance portfolio). However, because it limits the impact of parameter uncertainty, disallowing short selling generally leads to a performance increase out-of-sample.

When a value for γ is specified, another strategy that mitigates the impact of the estimation error is the **1/N rule**.[1] In this case we do not ignore the mean, as we want to somehow take into account the mean-variance preferences given by the value of γ. What we do instead is using the (sample) estimates of $\boldsymbol{\mu}$ and Σ to optimally allocate the wealth between the risk-free asset and the equally weighted risky assets. Remember that the mean-variance utility optimization problem is

$$\max_{\boldsymbol{w}} \boldsymbol{w}'\boldsymbol{\mu} - \frac{\gamma}{2}\boldsymbol{w}'\Sigma\boldsymbol{w}$$

If all the risky assets must have the same weight, it means we are imposing that $\boldsymbol{w} = c\mathbf{1}$, where c is a scalar that determines the weight of the assets. Therefore, what we need to find is the value of c. Hence, we substitute $\boldsymbol{w} = c\mathbf{1}$ in the original problem:

$$\max_{c} \ (c\mathbf{1})'\boldsymbol{\mu} - \frac{\gamma}{2}(c\mathbf{1})'\Sigma(c\mathbf{1})$$

$$\max_{c} c\mathbf{1}'\boldsymbol{\mu} - \frac{\gamma}{2}c^2\mathbf{1}'\Sigma\mathbf{1}$$

[1] The name "1/N rule" is often used to refer to a naive rule where one simply places all the wealth on risky assets with all weights equal to 1/N, without estimating any input. In these notes we refer instead to the more elaborate rule described in the text.

So we are back to an unconstrained problem. The first order condition is:

$$\frac{\partial U(c)}{\partial c} = \mathbf{1}'\boldsymbol{\mu} - \gamma c \mathbf{1}'\Sigma\mathbf{1} = \mathbf{0}$$

from which we easily get

$$c = \frac{1}{\gamma}\frac{\mathbf{1}'\boldsymbol{\mu}}{\mathbf{1}'\Sigma\mathbf{1}}$$

We then get the optimal weights for the risky assets by simply plugging this expression into $\boldsymbol{w} = c\mathbf{1}$

$$\boldsymbol{w}_{\mathbf{1}/N} = \frac{1}{\gamma}\frac{\mathbf{1}'\boldsymbol{\mu}}{\mathbf{1}'\Sigma\mathbf{1}}\mathbf{1}$$

while the weight for the riskless asset is given by $1 - \boldsymbol{w}'_{\mathbf{1}/N}\mathbf{1}$.

For example, if $N = 5$ and this rule returns a weight of 0.15 for each risky asset, we equally divide 75% of our wealth among the risky assets (i.e., 15% on each risky asset), and then place the remaining 25% on the risk-free asset.

FACTOR INVESTING WITH LONG-SHORT PORTFOLIOS

Computing optimal weights is not the only possibility. An alternative approach involves creating **long-short portfolios**. Suppose we want to invest into N assets. First, assets are ranked according to their predicted return. We then assemble a portfolio with two legs: a long leg which contains a given number of assets with the highest predicted returns, and a short leg with a given number of assets predicted to have the lowest returns. Within a certain leg the assets are often equally weighted, although other weighting systems that assign different weights based on the predicted return are of course possible.

One of the advantages of this approach is that it allows the investor to consider predicted returns without amplifying the effects of the estimation error in the mean. Optimization procedures like the mean-variance one are in fact error-maximizing: errors in the inputs lead to extreme weights, which can lead to abysmal performance. This is why standard mean-variance optimization rarely works well and is generally replaced either with more advanced techniques that limit extreme allocations, or with a minimum variance portfolio. A long-short portfolio does not have theoretically optimal weights, but it can still work better by avoiding this error-maximization trap.

The other advantage is that a long-short portfolio can be self-financing: the money obtained from shorting the assets predicted to perform poorly is used to go long on the assets with predicted high return. As short positions tend to be more risky than long positions, using a partially self-financing portfolio is also common. In this case, less than half (e.g., 30%) of the wealth is placed on the short leg, and the long leg is financed partly from the shorting and partly from the investor's initial wealth (in our example where 70% of the invested sum goes to the long leg, 30% of the money comes for the shorting and the other 40% from the investor's funds).

A practical disadvantage of such a portfolio is that in the real world it is generally difficult to short a large number of stocks. So in practice N has to be relatively small, and therefore it can be more risky, as it is not very well diversified. A long-short portfolio can also be particularly vulnerable in turbulent market conditions, when the price of virtually all stocks are either increasing or decreasing at the same time. The latter problem can be eased by making the portfolio construction more flexible (e.g., by varying the number of stocks and/or the amount of wealth in the long and short leg depending on the market conditions).

Obviously, a pre-condition for creating a long-short portfolio is having a ranking based on how we expect the assets to perform. How can we obtain it? Using the sample means is not appropriate, as we pointed out that such estimates are too unreliable. A much better alternative is to use a factor-based approach. In fact, long-short portfolios are a typical way **factor investing** is performed. This generally involves computing expected returns using multifactor models.

Remember that the formula of a multifactor model with k factors is:

$$R_i = \alpha_i + b_{i1}f_1 + b_{i2}f_2 + \cdots + b_{ik}f_k + \varepsilon_i$$

In practice, the expected return of a stock given a certain multifactor model is computed as:

$$E[R_i] = \alpha_i + b_{i1}\gamma_1 + b_{i2}\gamma_2 + \cdots + b_{ik}f\gamma_k$$

where γ is the factor risk premium.[2]

Therefore, we need to estimate the loadings b_{ik} and the risk premia. This is typically done using the **Fama-MacBeth regression**. It is a two-stage linear regression. Consider an estimation sample with N assets and T periods.

In the first stage, the loadings are estimated by regressing the returns of each asset i on the k factors, using the entire set of T periods:

$$\begin{gathered}
R_{1t} = \alpha_1 + b_{11}f_{1t} + b_{12}f_{2t} + \cdots + b_{1k}f_k \\
R_{2t} = \alpha_2 + b_{21}f_{1t} + b_{22}f_{2t} + \cdots + b_{2k}f_k \\
\vdots \\
R_{it} = \alpha_i + b_{i1}f_{1t} + b_{i2}f_{2t} + \cdots + b_{ik}f_k \\
\vdots \\
R_{Nt} = \alpha_N + b_{N1}f_{1t} + b_{N2}f_{2t} + \cdots + b_{Nk}f_{kt}
\end{gathered}$$

The estimated loadings are then used as explanatory variables in a second regression that, for each period t, regresses the asset returns of the entire set of N assets:

$$\begin{gathered}
R_{i1} = \gamma_{10} + \gamma_{11}\widehat{b_{\imath 1}} + \gamma_{12}\widehat{b_{\imath 2}} + \cdots + \gamma_{1k}\widehat{b_{\imath k}} \\
R_{i2} = \gamma_{20} + \gamma_{21}\widehat{b_{\imath 1}} + \gamma_{22}\widehat{b_{\imath 2}} + \cdots + \gamma_{2k}\widehat{b_{\imath k}} \\
\vdots \\
R_{it} = \gamma_{t0} + \gamma_{t1}\widehat{b_{\imath 1}} + \gamma_{t2}\widehat{b_{\imath 2}} + \cdots + \gamma_{tk}\widehat{b_{\imath k}} \\
\vdots \\
R_{iT} = \gamma_{T0} + \gamma_{T1}\widehat{b_{\imath 1}} + \gamma_{T2}\widehat{b_{\imath 2}} + \cdots + \gamma_{Tk}\widehat{b_{\imath k}}
\end{gathered}$$

Ideally we should use the true loadings, but their value is of course unknown in practice.

To compute the expected returns of each asset i we need the loadings, estimated in the first regression, and the risk premia, estimated in the second regression. Notice however that the risk premia are time-varying. A common approach is to compute their average value over the T periods (just like it is common to compute the average market excess return when using the CAPM). The expected return of asset i according to the chosen multifactor model is given by (we omit the ^ to keep the notation light):

$$E[R_i] = b_{i1}\gamma_1 + b_{i2}\gamma_2 + \cdots + b_{ik}\gamma_k$$

For greater clarity, let us consider how this works with the Fama-French three-factor model, which is probably the most important factor model.

[2] In the CAPM, and in single factor models in general, we can directly use the factor value (the excess market return in the case of CAPM). In multifactor models we cannot do this, and we need to use the risk premia of the factors instead.

Recall that the model is:

$$R_i = R_f + b_{i1}\left(R_m - R_f\right) + b_{i2}SMB + b_{i3}HML$$

In practice the expected return of asset i will be computed as:

$$E[R_i] = R_f + b_{i1}\gamma_{(R_m - R_f)} + b_{i2}\gamma_{SMB} + b_{i3}\gamma_{HML}$$

We use the Fama-MacBeth regression to estimate the loadings and the risk premia. Usually, the excess return is used as dependent variable, to focus on the component of the return that is dependent on factor exposure. Therefore, the first stage regression for each asset i is:

$$R_{it} - R_{ft} = \alpha_i + b_{i1}\left(R_{mt} - R_{ft}\right) + b_{i2}SMB_t + b_{i3}HML_t$$

To simplify the notation, we indicate the first factor as MKT:

$$R_{it} - R_{ft} = \alpha_i + b_{i1}MKT_t + b_{i2}SMB_t + b_{i3}HML_t$$

As explained before, this regression needs to be carried out separately for each of the N assets.

Now that we have the estimates for the loadings, we can set up the second stage regression:

$$R_i - R_f = \gamma_{t0} + \gamma_{t1}\widehat{b_{i1}} + \gamma_{t2}\widehat{b_{i2}} + \gamma_{t3}\widehat{b_{i3}}$$

This regression needs to be carried out separately for each of the T periods in the estimation window, obtaining T values for γ_{t1}, γ_{t2} and γ_{t3} . We then compute their average in order to have a single value. We rename the average of γ_{t1}, γ_{t2} and γ_{t3} as γ_{MKT}, γ_{SMB} and γ_{HML} respectively, for better clarity. We also compute the average risk-free rate in order to have a single value for R_f.[3]

We can now compute the expected return of each asset i as:

$$E[R_i] = R_f + b_{i1}\gamma_{MKT} + b_{i2}\gamma_{SMB} + b_{i3}\gamma_{HML}$$

It is now straightforward to create the long-short portfolio. We simply rank the assets according to their expected return, and take a long position on those positioned in the upper part of the ranking, and a short position on those in lower part of the ranking.

Each time the portfolio has to be updated, we need to compute new estimates of the expected returns. So, for example, if we want to update the portfolio monthly, we ne need to repeat the procedure every month, using the up-to-date data.

IMPROVING PORTFOLIO OPTIMIZATION

Let us consider again the portfolio optimization problem. Whatever metrics we will use to evaluate the results, we need to compare them with appropriate **benchmarks** in order to know if the results are satisfying. A benchmark portfolio surprisingly difficult to beat is the one created with a **naïve 1/N rule**.[4] This portfolio simply assigns equal weights to all the assets in all periods, and therefore does not require to estimate any input and to perform any optimization procedure. This is indeed one of its main strengths: it is completely immune to estimation errors. Another advantage is that it has a very low turnover, which translates into very low transaction costs.

[3] Computing the average value of the risk premia and of the risk-free rate is a reasonable approach, and the one commonly used. However, it is not "the" right approach. Other approaches might also be appropriate depending on the specific situation.

[4] We will refer to this rule/portfolio as "naive" or "naïve 1/N", to not confuse it with the 1/N rule we described before. However, it is also commonly called simply "1/N" rule/portfolio in the literature.

Another possible benchmark is the **market**, or more precisely the returns of a large stock market index, like the S&P 500. While it is not possible to buy an index, it is possible to buy ETFs. An **ETF (Exchange-Traded Fund)** is a fund that is traded on the financial markets and which tries to replicate a certain index. By investing in such fund, the investor can invest in a certain index without having to trade al the stocks contained in such index. Stock markets sometimes go through periods of poor performance that can last years, but over the long run they provide good returns (in countries with a solid economy). Therefore, such passive investing solution, which also does not require to perform estimation and optimization procedures, is another reasonable benchmark (over long enough periods of time).

Basic mean-variance optimization with sample estimates often struggles to beat common benchmarks, especially the naïve rule mentioned above. To improve optimization performance we need to reduce the severity of the estimation errors in the inputs, and/or reduce the impact of such errors on portfolio formation. Plenty of solutions have been proposed. We focus on some of them that proved to be effective and not too difficult to apply.

On the side of parameter estimation, we consider techniques developed to improve the estimation of the covariance matrix. While in general the error in the vector of sample means in more severe than the error in the sample covariance matrix, solutions proposed to improve over the latter are simpler and more effective. Moreover, while the number of parameters in $\boldsymbol{\mu}$ is equal to the number of assets N, the number of parameters in Σ is equal to N^2. Therefore, as the portfolio gets larger, estimation error in Σ becomes worse much faster than in does in $\boldsymbol{\mu}$, which might jeopardize the performance benefit that we theoretically get from a larger N thanks to the greater diversification potential.

The approach arguably more successful in dealing with the estimation of the covariance matrix is given by **shrinkage estimators**. Ledoit & Wolf (2004) define the shrinkage estimator

$$\Sigma_{\boldsymbol{LW}} = \delta \boldsymbol{I} \mu_{\hat{s}} + (1 - \delta)\Sigma$$

where Σ is the usual sample covariance matrix, $\boldsymbol{I}$ is the identity matrix, $\mu_{\hat{s}}$ is the average sample variance of all the variables (so the product $\boldsymbol{I}\mu_{\hat{s}}$ gives a diagonal matrix whose elements on the diagonal are equal to the average sample variance and all the other elements are equal to zero), and δ is the shrinkage parameter whose value is between 0 and 1.

It is called "shrinkage" estimator because the sample estimate is shrunk toward a target matrix. $\Sigma_{\boldsymbol{LW}}$ is basically the result of a weighted average between the sample covariance matrix and the target matrix. The intensity of the shrinkage is controlled by the shrinkage parameter δ. Ledoit and Wolf (2004) select an optimal value for δ with a procedure whose details are beyond our scope here. This shrunk estimator improves over the sample estimator, which translate in better performance of the mean-variance and minimum variance portfolios (and portfolios computed using the covariance matrix as input in general). Moreover, it gives a nonsingular covariance matrix even when the number of periods T used for estimation is smaller than the number of assets N. The sample covariance matrix, on the contrary, in such case is singular and therefore not invertible, which makes it impossible to perform the optimization procedures. It is also very easy to use this estimator in common statistical environments like R. Hence, it is now standard to use this estimator instead of the sample covariance in many applications, including portfolio optimization.

We now turn to solutions that mitigate the impact of estimation errors on portfolio formation. One of such solutions relies on a logic similar to the one we just described, and consists in computing **shrinkage portfolios**. The idea, proposed by Tu and Zhou (2011), is that we can improve over the mean-variance portfolio by shrinking it toward the naïve 1/N portfolio. In this way we can combine

a portfolio that optimizes the weights but suffers from estimation errors with one that is not optimized but is also immune from estimation errors, obtaining a portfolio that improves over both.

The weights of such portfolio are given by

$$\boldsymbol{w}^* = \delta \boldsymbol{w}_{\boldsymbol{NAIVE}} + (1 - \delta)\boldsymbol{w}$$

where $\boldsymbol{w}_{\boldsymbol{NAIVE}}$ is the vector of weights of the naïve 1/N portfolio, $\boldsymbol{w}$ is the vector of weights of the optimized portfolio (usually the mean-variance portfolio, but can be also other portfolios), and δ is again the parameter that controls the shrinkage intensity. The value of δ can be chosen using optimization rules, heuristics, or cross-validation.

Another solution to improve over the standard mean-variance portfolio is the **grouping strategy** proposed by Branger et al. (2019). The idea is that since the performance of mean-variance (or minimum variance) optimization suffers more and more as N increases, due to the growing number of parameters to estimate, one could achieve better performance by grouping together the assets in a certain number of groups. The optimization procedure is than performed between the groups, while within a group the assets are equally weighted. This reduces the dimension of the problem, and therefore the number of parameters to estimate. In other words, it is another strategy that combines the benefits of the naïve 1/N rule (which is applied within groups) and of optimization (which is applied between groups). The stocks can be grouped according to how similar they are in terms of estimated mean, variance or beta, and the number of groups can be chosen using optimization rules, heuristics, or cross-validation. The higher the number of groups, the closer we get to the usual optimization; the smaller the number of groups, the closer we get to the naïve 1/N rule. In the extreme case where we only have one group we obtain the naïve 1/N portfolio; in the extreme case where the number of groups is equal to N, we get the usual optimized portfolio.

DOWNSIDE RISK MEASURES

So far we measured risk using the variance (or the standard deviation, which is simply its square root). This is based on the assumption that returns are symmetrically distributed, or at least that the investor only cares about volatility as a whole, without distinguishing between upside and downside movements. While this is not realistic (because investors want to minimize the losses but not the gains), it greatly simplifies optimization procedures. Moreover, **downside risk measures** tend to be more difficult to estimate as inputs for optimization procedures, which may lead to worse performance despite targeting a more appropriate measure of risk. For these reasons, here we do not consider portfolio optimization techniques targeting downside risk. However, it is still important to be familiar with the most popular downside risk measures, as they are useful for performance evaluation, and some of them are also employed by regulatory authorities supervising the banking sector.

The downside risk measure most closely related to the variance is the **semivariance**, which is defined as:[5]

$$\sigma_B^2 = \frac{1}{T}\sum_{t=1}^{T}[\mathrm{Min}(R_t - B, 0)]^2$$

where T is the number of periods in the estimation window, and B is the benchmark below which the investor considers volatility to account as risk. To apply this formula, one has to replace all the

[5] Technically, this is the downside semivariance, as it is also possible to compute an upside semivariance by replacing Min with Max in the formula. However, we are generally interested in the downside semivariance, which we therefore simply call "semivariance".

portfolio returns above the benchmark with 0, and then the computations are exactly the same done to compute the variance.

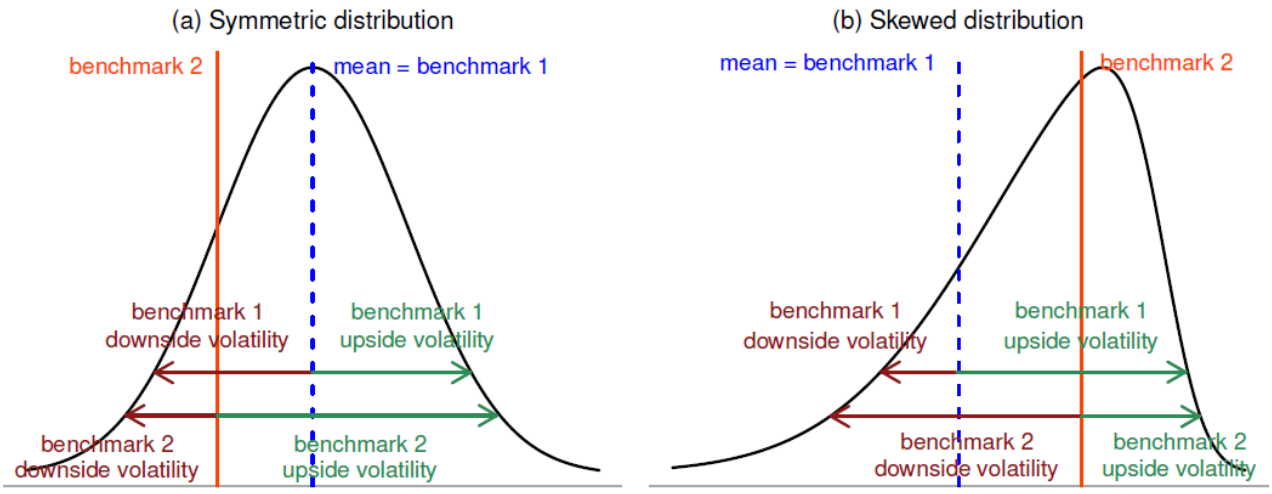
B depends on the preferences of the investor. It is convenient (and also has some nice theoretical properties) to set B equal to the risk-free rate. In this way if one works with excess returns, B can be treated as equal to zero. However, in principle, it can be set to any value.

The square root of σ_B^2 is called **downside deviation**, which we indicate with σ_B. The downside deviation is to the semivariance, what the standard deviation is to the variance.

Theoretically, one can compute an optimal mean-semivariance or minimum semivariance portfolio by simply replacing the covariance matrix with the semicovariance matrix (the analogous to the covariance matrix in a downside risk setting) in the optimization procedures. However, estimating this matrix presents several challenges, and therefore we do not address this topic, but we still provide some intuition regarding what it means to target the semivariance. We distinguish between different scenarios:

- If the distribution is symmetric and the benchmark is equal to the sample mean, targeting the variance or the semivariance is always equivalent. One should therefore target the former, as sample estimates for it are more accurate than the sample estimates for the latter.
- If the distribution is symmetric but the benchmark is not equal to the sample mean, targeting the variance or the semivariance is only equivalent if we set a target return (i.e., mean-semivariance optimization). Minimize the variance or the semivariance without a target return is not equivalent in this setting.
- If the distribution is not symmetric, targeting the variance is never equivalent to targeting the semivariance.

The following figure provides a graphical illustration.



To compute the risk-adjusted return in this context, the Sharpe ratio should be replaced by the **Sortino ratio**, which is similar to the Sharpe ratio but replaces the risk-free rate with the benchmark B, and the standard deviation with the downside deviation σ_B:

$$\text{Sortino} = \frac{\overline{R} - B}{\sigma_B}$$

Another popular downside risk measure is the **Value at Risk (VaR)**. VaR measures the maximum potential loss that an investor can suffer over a certain period, with a $1-\alpha$ confidence level. α is set by the investor; for example, an $\alpha = 0.05$ corresponds to a 95% confidence level.

More formally, given a profit and loss distribution Y we can define VaR as:

$$VaR_{\alpha}(Y) = -\inf\{y \in \mathrm{R}: (Y \leq y) > \alpha\}$$

For example, if we set $\alpha = 0.05$ and when evaluating a set of returns we get a $VaR = 0.04$, it means that we have a 5% chance of losing 4% or more in one period over the time horizon considered.

VaR can be computed in different ways. The most commonly used is the historical method: we simply rank the historical returns in increasing order and then check the (typically negative) return that we have at the α percentile. Another possibility is the parametric method: we assume that returns follow a certain distribution and we compute the loss at the chosen percentile. Simulation ("Monte Carlo") approaches are also possible.

The main problem with VaR is that it is not a **coherent risk measure**. Consider the outcomes V_1 and V_2 of two investments. A risk measure is said to be coherent if it possesses the following desirable properties:

- Monotonicity: if V_1 is larger or equal to V_2 in every possible scenario, then the risk of V_1 must be lower than V_2. Formally: if $V_1 \geq V_2$, then $Risk(V_1) < Risk(V_2)$.
- Translation invariance: for any outcome V, adding an additional outcome C with a certain return reduces the risk by that amount. Formally: $Risk(V + C) = Risk(V) - C$.
- Positive homogeneity: multiplying all outcomes by a constant should result in a scaling of the risk measure by the same constant. In other words, if we invest, say, twice the original amount, the risk measure should also double. Formally: $Risk(\lambda V) = \lambda Risk(V)$.
- Subadditivity: the risk of a combination of two risky positions should be lower or equal to the risk of the individual positions. In other words, diversifying by combining different assets should reduce risk, or at worst leave it unaffected, but it cannot increase it. Formally: $Risk(V_1 + V_2) \leq Risk(V_1) + Risk(V_2)$.

The Value at Risk satisfies the first three conditions, but not the last one. As it violates subadditivity, risk quantified using VaR can sometimes increase with greater diversification, which is not very meaningful.

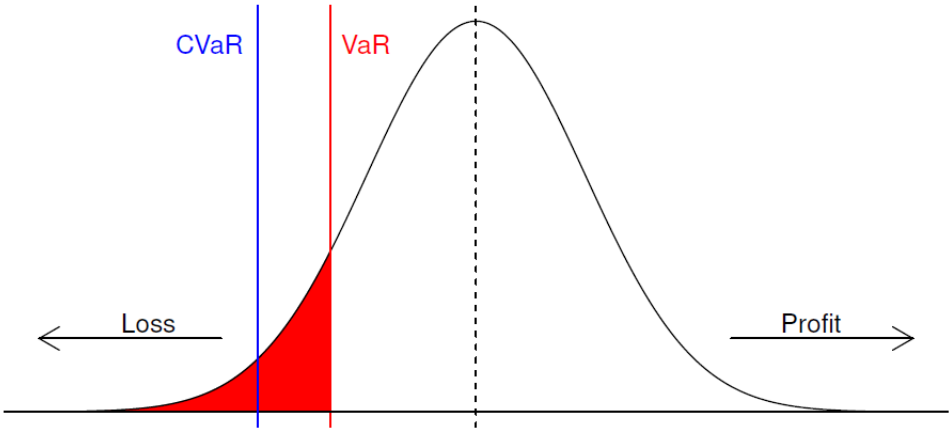
To overcome this problem, the **Conditional Value at Risk (CVaR)**, also known as **Expected Shortfall (ES)**, has been proposed:

$$CVaR_{\alpha}(Y) = -\frac{1}{\alpha}\int_{0}^{\alpha} VaR_{u}du$$

where u is just the variable of integration and du is the differential of this variable (i.e., we are integrating from 0 to α using infinitesimal increments in u from 0 until we reach α).

In more intuitive terms, the CVaR measures the average (the "expected") loss that we get, given that the loss exceeds the VaR. As it is a coherent measure of risk, it is preferred and more commonly used than the VaR. Of course, in order to compute the CVaR, you first need to compute the VaR.

The following figure provides a graphical intuition of VaR and CVaR:



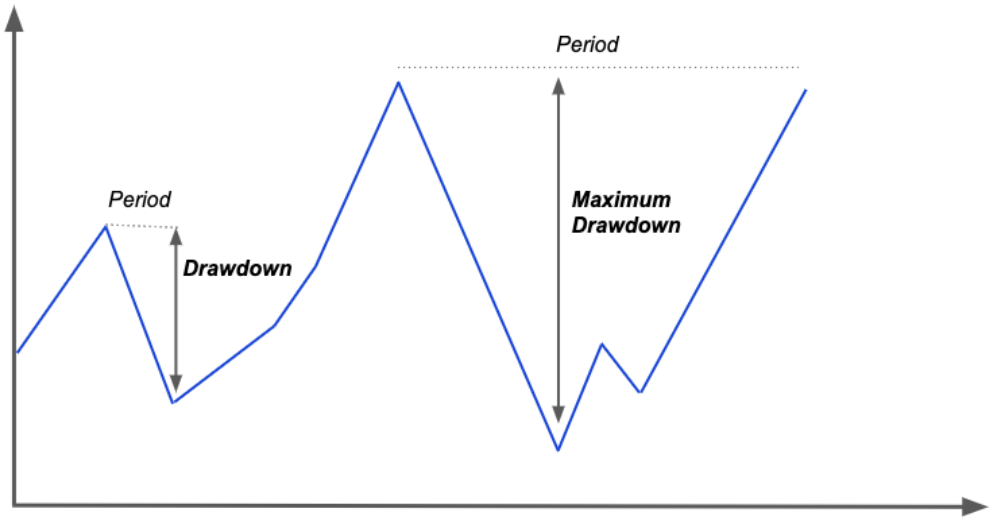
CVaR is always lower than the VaR, because it is the value that we get by computing the average loss that we have when we find ourselves in the red area left of the VaR.

Finally, another popular measure of downside risk is the **drawdown (DD)**. The drawdown is the decline in the value of an investment from a peak to a low point. Different drawdown measures can be computed. A popular and easy to compute one is the **maximum drawdown (MDD)**:

$$MDD = \frac{Trough\ Value - Peak\ Value}{Peak\ Vaue}$$

where the "Trough Value" is the lowest point in the series that is reached after the highest peak.

Obviously, a lower MDD is preferable to a higher MDD. In the worst possible case, MDD is equal to 100%, i.e., the value of the investment drops to zero.



Source: https://financetrain.com

MDD fails to consider the frequency and duration of losses, and does not account for the size of any gains. To account for the gains, we can use a more informative measure called **Calmar Ratio**:

$$Calmar = \frac{\overline{R} - r_f}{MDD}$$

This is similar to the Sharpe ratio, but the MDD is used instead of the standard deviation.

PERFORMANCE EVALUATION

After we obtain the series of portfolio returns, we need to appropriately evaluate results in order to gauge how well our investment strategy performed. We may start with some basic statistics about the distribution of the returns:

- Mean: the higher the better
- Standard deviation: the lower the better
- Skewness: a positive value is preferable
- (Excess) Kurtosis: a lower value is preferable unless the skewness is significantly positive

We then compute the Sharpe ratio to quantify the risk-adjusted return.

If we are interested in evaluating downside risk we may compute, instead or in addition to the standard deviation and the Sharpe ratio, some or all of these measures:

- Downside deviation: the lower the better
- CVaR: the lower (in absolute value) the better
- (Maximum) drawdown: the lower the better

We then quantify the risk-adjusted return with an appropriate measure, like the Sortino ratio.

Another important financial indicator is the **alpha**, used to check if the returns of the investments are explained by a given asset pricing model. The alpha is obtained as the intercept in a regression of the portfolio returns over the returns of the factors of the model considered. Usually, the CAPM (in which case the alpha is called "Jensen's alpha") or the Fama-French three-factor model are used. In the first case the regression takes the form:

$$R = \alpha + \beta\left(R_{Mkt} - R_f\right)$$

while in the second case it is:

$$R = \alpha + b_1\left(R_{Mkt} - R_f\right) + b_2 SMB + b_3 HML$$

If α is significantly greater than zero, it means that our strategy achieves returns higher than those predicted by the model based on the portfolio exposure to the factors. In order to check if we have a significant positive α, we compute its standard errors and then perform a test of hypothesis. The usual standard errors for the linear regression are generally not appropriate, as they assume homoskedasticity (i.e., constant variance) and no autocorrelation (i.e., no temporal dependency in the standard errors). These conditions are usually not met in financial time series, where heteroskedasticity and/or autocorrelation are often observed. To account for this, we may use instead the **Newey-West standard errors**. The technical details of such estimator are somewhat complicated, but not relevant here, and Newey-West standard errors can be easily computed in R. Using them, we can test whether the α of our investment is significantly greater than zero.

A proper evaluation, however, should also account for the **turnover**, i.e., how much trading the strategy requires. The higher the turnover, the higher the transaction costs, which of course translates into lower net returns. To get an idea of the amount of trading required we can compute the average turnover. The turnover at a certain period t is given by:

$$TO_t = \sum_{i=1}^{N} \left| w_{i,t} - w_{i,t-1} \right|$$

Basically, for each stock we compute the absolute value of the change in the corresponding weight compared to the previous period, and we sum all these N values.

We do this for each of the T periods in which we applied our strategy, and then we compute the mean. This gives us the average turnover.

However, applying this formula using the set of weights we selected for the previous period as the value for $w_{i,t-1}$ is not entirely correct. This is because when we update the weights at the beginning of each new period we need to account for the fact that, due to the realized returns during the period that just ended, the allocation of wealth changed compared to what was at the beginning of the previous period. We clarify this with an example.

Suppose we have a portfolio with two assets updated monthly, and the weights at time $t-1$ were 0.5 and 0.5, while now at time t we want to change them to 0.4 and 0.6. We might think that the turnover is $|0.4-0.5|+|0.6-0.5|=0.2$, which means we have to trade 20% of our wealth to update the portfolio. If the price of the two assets did not change over the month that just ended, this would indeed be correct. Suppose however that during that month the first asset experienced a +10% return, and the second one a -20% return. When we update the portfolio at time t, we no longer have the two original weights, but $0.5+0.5\times 0.1=0.55$ for the first asset and $0.5-0.5\times 0.2=0.4$ for the second. The weights computed like this do not sum up to 1 because the total value of the portfolio changed compared to period $t-1$. We need to account for this by dividing both weights by their sum. So in this example where they sum to $0.55+0.4=0.95$ we have $0.55/0.95\approx 0.58$ for the first weight and $0.4/0.95\approx 0.42$ for the second. Therefore, the actual turnover is $|0.4-0.58|+|0.6-0.42|=0.36$.

We might express this concept by rewriting the formula for the turnover as

$$TO_t=\sum_{i=1}^{N}\left|w_{i,t}-w_{i,t-1}^{+}\right|$$

where $w_{i,t-1}^{+}$ indicates that we are considering the returns from the previous period after accounting for the redistributing effect of the realized returns. We can then compute the average turnover as before.

While the turnover certainly provides some useful information, we can get an even better figure by considering the portfolio **returns net of transaction costs**. Transaction costs can be fixed or proportional to the amount of trading. It is generally considered more appropriate to use proportional transaction costs. These can be accounted for by multiplying the turnover of each asset for the proportional cost. Frazzini (2012) suggests using transaction costs equal to 10 basis points (bp). A basis point is equal to 0.01%. Therefore, for example, if we need to buy or sell 5% of the positions we have in a certain asset, we face transaction costs equal to $0.05\times 0.001=0.00005$, which means that 0.005% of the money invested in that position is lost in transaction costs.

Once we have the portfolio returns net of transaction costs, we can use them to compute all the other statistics we listed before.

Finally, it is useful to visualize the value V of the portfolio over time, which can be computed as

$$V_T=V_0+\sum_{t=1}^{T}(V_{t-1}R_t)$$

It is appropriate to compute the value both ignoring and net of transaction costs. The result is then plotted in a graph, which provides a visual representation of the effectiveness of our strategy.

We might want to also compute the evolution of real wealth in addition to the nominal wealth. In other words, we might want to account for the inflation. We can do this by dividing the value of the portfolio over time by the deflator.

We can compute the deflator D using a formula analogous to the one used to compute the value of the portfolio, simply replacing the return with the inflation rate I:

$$D_T = D_0 + \sum_{t=1}^{T} (D_{t-1} I_t)$$

Of course, the two series need to have the same starting value (e.g., 1 unit of wealth), and the same frequency (e.g., monthly).