Mathematics of Finance, Courant Institute, Spring 2019
https://www.math.nyu.edu/faculty/goodman/teaching/MathFin2019/MathFinance.html

Mean variance analysis
(Incomplete version posted April 5, 2019)

Asset allocation in finance is the problem of deciding how to allocate your assets among a menu of risky investments. *Mean variance* analysis is a framework this. The *return* on an investment is the profit, expressed as a percentage (return = profit/investment). You have to do the allocation (make your investments) before the return is known. It is treated as a random variable with a mean and variance. The mean, which is the *expected return*, is something the investor wants to be large. The variance of the return is a measure of risk, which the investor wants to be small. Asset allocation (investing) is seen as a tradeoff between risk (variance) and return (expected return).

You can think of asset allocation theory as a systematic approach to *diversification*. Suppose there are two stocks S_1 and S_2 that each have $\mu = 10\%$ expected return and variance σ^2. Suppose S_1 and S_2 are independent random variables (an extreme case, real stocks aren't independent). If you invest \$100 on S_1 or S_2, then you have mean 10% and variance σ^2. If you invest \$50 each on S_1 and S_2, then your mean is still 10%, but your variance is $\frac{1}{2}\sigma^2$ (calculations below). Diversification, spreading your money around, has given you the same expected return with less risk. Mean/variance analysis is for (somewhat) more realistic situations. Suppose that $\mu_2 < \mu_1$, then there is a cost (lost expected return) to investing in S_2. There is a tradeoff. Reduced risk comes at the cost of reduced expected return.

The goal of mean variance analysis is to choose the investment strategy that maximizes expected return for a given risk. Equivalently (we will see this) you can minimize risk (variance) for a given expected return. A *portfolio* (asset allocation) is called *efficient* if it satisfies these criteria. There is more than one efficient portfolio. Some have high return and high risk. Others have less return in exchange for less risk. Different investors will choose different points on the risk/return curve. But every investor should have an efficient portfolio. No investor wants to have less expected return than they could have for a specified risk. The *efficient frontier* is the set of all efficient portfolios.

The simple mean variance analysis covered here is just the beginning of asset allocation and investment. The ideas explained here, along with others, are used in most fancier theories of asset allocation or investment and trading strategies.

Review of variance and covariance

If X is a random variable, the expected value is $\mathrm{E}[X]$. It is common (but not universal) to use capital letters for random variables and lower case letters for values they might take. If X has a probability density $p(x)$, then

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x p(x)\, dx \ .$$

A random variable with a probability density is called *continuous*. The probability density of a continuous random variable doesn't have to be a continuous function of x. The expectation may be written μ or μ_X (to emphasize the random variable it is the expectation of) or $\overline{X}$.

Some random variables take values only in a certain list of values. The possible stock prices in the binary or binomial tree model are examples. A random variable like this is called *discrete*. Suppose the possible values are called x_j and $\Pr(X = x_j) = p_j$. Then

$$\mathrm{E}[X] = \sum_k x_j p_j \ .$$

The expectation has some mathematical properties that don't depend on how it is defined. The expectation is *linear*. If X and Y are any two random variables, then

$$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y] \ .$$

If c is a constant (not random), then

$$\mathrm{E}[cX] = c\mathrm{E}[X] \ .$$

The expected value of a constant is that constant.

The *variance* of X is

$$\mathrm{var}(X) = \mathrm{E}\Big[\, (X - \mu_X)^2 \Big] \ .$$

The variance is often called σ^2, or σ_X^2. The *standard deviation* is

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{\mathrm{var}(X)} \ .$$

The variance is easier to calculate (because it doesn't have a square root), but it may be less meaningful. The standard deviation measures how far X is likely to be from its mean.

Suppose X and Y are two continuous random variables. We write $p_X(x)$ for the probability density of X, and $p_Y(y)$ for the probability density of Y. The joint probability density is $p_{XY}(x, y)$. If X and Y are independent, then $p_{XY}(x, y) = p_X(x)p_Y(y)$. It is uncommon that two random variables in finance are independent.

The *covariance* of X and Y is

$$\mathrm{cov}(X, Y) = \mathrm{E}[\, (X - \mu_X)\, (Y - \mu_Y)] \ .$$

If X and Y are independent, then (check this!) $\mathrm{cov}(X, Y) = 0$. If $\mathrm{cov}(X, Y) \neq 0$, then X and Y are not independent. We say they are *correlated*. If $\mathrm{cov}(X, Y) >$

0, we say that X and Y are *positively correlated*. Almost every pair of stock prices is positively correlated. If $\text{cov}(X, Y) < 0$, we say X and Y are *negatively* correlated, or *anti-correlated*. It is often said that stock and bond prices are anti-correlated. If bond prices go up, then people sell their stocks to buy bonds (the story goes).

The variance is a *special case* of covariance. Look at the formulas for variance and covariance. You will see that

$$\text{cov}(X, X) = \text{var}(X) .$$

The variance is the expected value of a positive quantity. Therefore $\text{var}(X) > 0$ unless X is not random. If X is constant, then $X = \mu_X$ and $(X - \mu_X) = 0$ always. This makes the expected value equal to zero. Otherwise, the expected value is positive. Any truly random variable is positively correlated with itself.

An asset allocation is a sum of several investments. The total return is the sum of the returns on the individual investments. These individual returns are correlated random variables. The variance of the total return, which represents its risk, is the variance of a sum of correlated random variables. We need a formula for this.

Suppose X and Y are correlated random variables. The mean of $Z = X + Y$ is

$$\mu_{X+Y} = \text{E}[\,X + Y] = \mu_X + \mu_Y .$$

The variance of $Z = X + Y$ is

$$\begin{aligned}
\text{var}(X + Y) &= \text{E}\Big[\,(X + Y - \mu_{X+Y})^2\Big] \\
&= \text{E}\Big[\,([X - \mu_X] + [Y - \mu_Y])^2\Big] \\
&= \text{E}\big[\,(X - \mu_X)^2 + 2(X - \mu_x)(Y - \mu_Y) + (Y - \mu_Y)^2\big] \\
&= \text{E}\big[\,(X - \mu_X)^2\big] + 2\text{E}[\,(X - \mu_X)(Y - \mu_Y)] + \text{E}\big[\,(Y - \mu_Y)^2\big] \\
\text{var}(X + Y) &= \text{var}(X) + 2\text{cov}(X, Y) + \text{var}(Y) . \qquad (1)
\end{aligned}$$

A variance is like a square. The variance sum formula is like the square sum formula (the binomial theorem) except that X^2 becomes the variance and XY becomes the covariance.

We need the formula for the variance of a sum of n random variables. Suppose X_j are random variables (all correlated) and w_i are *weights* (numbers that are not random). The *weighted sum* is

$$Z = \sum_{j=1}^{n} w_j X_j . \qquad (2)$$

In the application, X_j represents the value of one "share" (economists call it one *unit*) of asset j. The weight w_j is the number of shares of asset j in the portfolio. The variance of Z represents the risk of this portfolio. The formula that corresponds to (1) involves the variances and covariances. We use a slightly different notation:

$$\begin{aligned} \sigma_{jj} &= \sigma^2_{X_j} = \text{var}(X_j) \\ \sigma_{jk} &= \text{cov}(X_j, X_k) \ . \end{aligned}$$

Here is a trick with the indices that leads to a simple formula for var(Z). Suppose a_j are numbers and

$$s = \left(\sum_{j=1}^{n} a_j \right)^2 \ .$$

This can be written as

$$s = \left(\sum_{j=1}^{n} a_j \right) \left(\sum_{k=1}^{n} a_k \right) \ .$$

The two sums on the right are equal. The only difference is the letter we use to represent the summation index. But the second sum can be written as

$$s = \sum_{j=1}^{n} \sum_{k=1}^{n} a_j a_k \ . \tag{3}$$

The square of a single sum has been written as a double sum of all products $a_j a_k$.

Suppose $n = 2$, so $s = (a_1 + a_2)^2$. The double sum formula is

$$\begin{aligned} s &= a_1 a_1 + a_1 a_2 + a_2 a_1 + a_2 a_2 \\ &= a_1^2 + 2a_1 a_2 + a_2^2 \ . \end{aligned}$$

The coefficient 2 in the variance sum formula (1) arises from the fact that $a_1 a_2 = a_2 a_1$. In the full sum (3), the *diagonal* terms are the ones with $j = k$. These have value a_j^2. The *off diagonal* terms are the ones with $j \neq k$. These come in pairs, since $a_j a_k = a_k a_j$. We could combine these by taking only the term with $k > j$ (or $j > k$, but not both). The result would be a sum of diagonal and off diagonal terms, with the factor of 2:

$$s = \sum_{j=1}^{n} a_j^2 + 2 \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} a_j a_k \ .$$

This formula requires some attention to detail in the off diagonal sum. The j variable goes from $j = 1$ to $j = n - 1$ because there are no off diagonal terms with $j = n$ and $k > j$. The k variable starts at $j + 1$ because that is the first off diagonal term. We often prefer the formula (1) because it is simpler. But whenever you use it, keep in mind that each off diagonal term appears twice.

Back to portfolios. A general portfolio is a weighted sum of assets (2). The expected value of the portfolio is

$$\mu_Z = \sum_{j=1}^n w_j \mu_j \; .$$

To see this,

$$\begin{aligned}
\mu_Z &= \mathrm{E}[Z] \\
&= \mathrm{E}\left[\sum_{j=1}^n w_j X_j \right] \\
&= \sum_{j=1}^n w_j \mathrm{E}[X_j] \qquad \text{(Expectation is linear)} \\
\mu_Z &= \sum_{j=1}^n w_j \mu_j \; .
\end{aligned}$$

The variance of the portfolio is

$$\mathrm{var}(Z) = \sum_{j=1}^n \sum_{k=1}^n w_j w_k \mathrm{cov}(X_j, X_k) = \sum_{j=1}^n \sum_{k=1}^n w_j w_k \sigma_{jk} \; . \tag{4}$$

The algebra behind this uses the square of the sum trick (3):

$$\begin{aligned}
\mathrm{var}(Z) &= \mathrm{E}\Big[\left(Z - \mu_z\right)^2 \Big] \\
&= \mathrm{E}\left[\left(\sum_{j=1}^n w_j X_j - \sum_{j=1}^n w_j \mu_j \right)^2 \right] \\
&= \mathrm{E}\left[\left(\sum_{j=1}^n w_j \left(X_j - w_j \mu_j\right) \right)^2 \right] \\
&= \mathrm{E}\left[\left(\sum_{j=1}^n w_j X_j - w_j \mu_j \right) \left(\sum_{j=k}^n w_k X_k - w_k \mu_k \right) \right] \\
&= \mathrm{E}\left[\sum_{j=1}^n \sum_{k=1}^n w_j w_k \left(X_j - \mu_j\right)\left(X_k - \mu_k\right) \right] \qquad \text{(double sum trick of (3))} \\
&= \sum_{j=1}^n \sum_{k=1}^n w_j w_k \mathrm{E}[\left(X_j - \mu_j\right)\left(X_k - \mu_k\right)] \qquad \text{(expectation is linear)} \\
\mathrm{var}(Z) &= \sum_{j=1}^n \sum_{k=1}^n w_j w_k \sigma_{jk} \; . \qquad \text{(the formula (4))}
\end{aligned}$$

The covariances σ_{jk} form the elements of an $n \times n$ symmetric matrix called the *covariance* matrix (or the *variance covariance* matrix, because the diagonal elements σ_{jj} are variances). We will call this matrix C. The (j,k) entry of C is σ_{jk}. It will be helpful to write the variance formula (4) in the notation of linear algebra. Let $w \in \mathbb{R}^n$ be the $n-$component column vector whose components are the weights w_j:

$$w = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} .$$

The transpose of w is the row vector with the same components

$$w^t = (w_1, w_2, \cdots, w_n) .$$

The sum in (4) may be written

$$\text{var}(Z) = w^t C w . \tag{5}$$

This abstract version of the variance formula (4) simplifies the analysis and the programming.

We verify the matrix/vector formula (5) first for $n = 2$ and then in general. For $n = 2$, the calculation is (note $\sigma_{12} = \sigma_{21} = \text{cov}(X_1, X_2)$):

$$\begin{aligned} w^t C w &= \begin{pmatrix} w_1 & w_2 \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \\ &= \begin{pmatrix} w_1 & w_2 \end{pmatrix} \begin{pmatrix} \sigma_{11} w_1 + \sigma_{12} w_2 \\ \sigma_{12} w_1 + \sigma_{22} w_2 \end{pmatrix} \\ &= w_1 \left(\sigma_{11} w_1 + \sigma_{12} w_2 \right) + w_2 \left(\sigma_{12} w_1 + \sigma_{22} w_2 \right) \\ &= w_1^2 \sigma_{11} + 2 w_1 w_2 \sigma_{12} + w_2^2 \sigma_{22} \ . \end{aligned}$$

This is the same as the formula (4).

For general n, define the vector $v = Cw$. The components of v are

$$v_j = \sum_{k=1}^{n} \sigma_{jk} w_k .$$

We also have (because matrix/vector multiplication is associative)

$$\begin{aligned} w^t C w &= w^t v \\ &= \sum_{j=1}^{n} w_j v_j \\ &= \sum_{j=1}^{n} w_j \left(\sum_{k=1}^{n} \sigma_{jk} w_k \right) \\ &= \sum_{j=1}^{n} \sum_{k=1}^{n} w_j w_k \sigma_{jk} . \end{aligned}$$

This shows that the "scalar sum" form (4) is equivalent to the matrix/vector form (5).

Basic one period model

In the simplest one period model, a total wealth M is to be allocated among n *risky assets*. The price of asset j is 1 today and X_j "tomorrow". We may invest w_j on asset j, which costs w_j today and yields $w_j X_j$ tomorrow. The numbers X_j are random and not known at the time the asset allocation is made. The only information we have is the expectations (means), variances, and covariances. The means are

$$\mathrm{E}[\,X_j] = \mu_j \; . \tag{6}$$

The variances are

$$\mathrm{var}(X_j) = \sigma_j^2 = \sigma_{jj} \; .$$

The covariances are

$$\mathrm{cov}(X_j, X_k) = \sigma_{jk} \; .$$

The wealth "tomorrow" is

$$Z = \sum_{j=1}^{n} w_j X_j \; .$$

The expected wealth tomorrow is

$$\mu_Z = \mathrm{E}[Z] = \sum_{j=1}^{n} \mu_j w_j = \mu^t w \; .$$

Here, $\mu \in \mathbb{R}^n$ is a column vector with components μ_j:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} \; .$$

The variance of the wealth tomorrow is (see the previous section)

$$\sigma_Z^2 = \mathrm{var}(Z) = w^t C w \; .$$

The entries of C are the covariances σ_{jk}.

The goal of mean variance analysis is to maximize the expected return μ_Z with a constraint on the variance σ_Z^2 and the total investment

$$M = \sum_{j=1}^{n} w_j \; .$$

This is equivalent (we will see) to minimizing the variance with a constraint on the expected return and the total investment. We want to write everything in the language of linear algebra, vectors, matrices and such. The total investment constraint is just a sum, but it can be put in linear algebra form using the vector of all ones:

$$\mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} .$$

The total investment constraint (also called the *budget constraint* may be written

$$M = \mathbf{1^t w} .$$

Gradients and Lagrange multipliers

Cauchy Schwarz inequality

The *Cauchy Schwarz inequality* is a simple theorem about vectors and inner products in n dimensional space. Suppose u and v are n component column vectors, the inequality is

$$(u^t v)^2 \leq (u^t u)\,(v^t v) . \tag{7}$$

Moreover, the inequality is *strict* (meaning $(u^t v)^2 < (u^t u)\,(v^t v)$) unless u and v are "in the same direction". If $u \neq 0$ and $v \neq 0$, being in the same direction means that there is a scaling s so that $u = sv$. In components, this means that $u_i = sv_i$ for $i = 1, \ldots, n$.

The proof of the Cauchy Schwarz inequality is a clever trick. Look at

$$m(s) = (u - sv)^t (u - sv) = u^t u - 2su^t v + s^2 v^t v .$$

"Clearly" $m(s) \geq 0$ for all s, because if x is any vector, then $x^t x = \sum x_i^2 \geq 0$. Maybe there is an s_* so that $m(s_*) = 0$. In that case $u - s_* v = 0$, which means u and v point in the same direction. Otherwise $m(s) > 0$ for all s.

Choose s_* to minimize m. Take the derivative with respect to s and set it to zero. The result is

$$-2u^t v + 2s_* v^t v = 0 \implies s_* = \frac{u^t v}{v^t v} .$$

We calculate

$$m(s_*) = u^t u - \frac{\left(u^t v\right)^2}{v^t v} .$$

This is positive (because $m(s)$ is always positive), so

$$u_t u > \frac{\left(u^t v\right)^2}{v^t v} \implies (u^t u)(v^t v) > (u^t v)^2 .$$

This is the Cauchy Schwarz inequality. It is strict – we have $>$, not $\geq$. It is strict because $m(s_*) > 0$, which is because u and v are not in the same direction.

The gradient vector and descent direction

Suppose $f(x) = f(x_1, \ldots, x_n)$ is a function of n variables. You can think of the variables x_j as the components of an n component column vector x. The *gradient* of f is the n component column vector made of first partial derivatives:

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_j} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

Recall from multivariate calculus that the gradient leads to a first derivative approximation to f. Consider two nearby "points" x and $x + \Delta x$. The column vector Δx has components Δx_j. Then

$$f(x + \Delta x) - f(x) \approx \sum_{j=1}^{n} \frac{\partial f(x)}{\partial x_j} \Delta x_j \ .$$

In vector notation, this may be written

$$f(x + \Delta x) - f(x) \approx (\Delta x)^t \, \nabla f(x) \ . \tag{8}$$

A *constrained optimization* problem is to find the maximum or minimum of $f(x)$, but with *constraints* that may be written

$$g_i(x) = a_i \ , \quad \text{for } i = 1, \ldots k \ .$$

Constraints like this are *equality constraints*. There are also *inequality constraints*, which take the form $g_i(x) \geq a_i$. Basic mean/variance analysis involves only equality constraints.

An *optimality condition* is an equation that we can solve to help find the optimal x. We write x_* for "the" optimal x (there may be more than one). For *unconstrained* optimization (no constraints, $k = 0$), the optimality condition is that $\nabla f(x_*) = 0$. To see this, suppose $\nabla f(x_*) \neq 0$. Choose a small *step size*, s, and take $\Delta x = s \nabla f(x_*)$. The first derivative approximation formula gives

$$f(x_* + \Delta x) - f(x) \approx s \, (\nabla f(x_*))^t \, \nabla f(x_*) \ .$$

The inner product on the right is

$$(\nabla f(x_*))^t \, \nabla f(x_*) = \sum_{j=1}^{n} \left(\frac{\partial f(x_*)}{\partial x_j} \right)^2 \ .$$

This is positive unless all the partial derivatives are zero. This means we can make f a little bigger by taking $s > 0$ and f a little smaller by taking $s < 0$. Either way, x_* is not the optimal x.

You can imagine that $f(x)$ is the "height" of a surface over the "x plane" (though x may have more than two components). Then $\nabla f(x)$ is a vector that points in the steepest uphill direction. If $\nabla f(x) \neq 0$, then you are on the side of a hill. The function gets bigger (higher) in one direction and lower in the other direction.

For minimization, the negative gradient $-\nabla f$ is a *descent direction*, f decreases if you go in that direction, at least if you don't go too far.

One equality constraint

The situation is more complicated for equality constrained optimization. Suppose that there is only one constraint $g(x) = a$ and that x_* satisfies it. We want to see whether there are nearby x values that satisfy the constraint and have better (larger or smaller) f. For this, we need to choose Δx that stays on the *constraint surface* (the set of x values that satisfy the constraint). If we go from x_* to $x_* + \Delta x$, the constraint changes according to the first derivative approximation

$$g(x_* + \Delta x) - g(x_*) \approx (\Delta x^t) \nabla g(x_*) \ .$$

We want to see whether $\Delta f = f(x+\Delta x) - f(x)$ can be made positive or negative with perturbations Δx that stay on the constraint surface:

$$(\Delta x^t) \nabla g(x_*) = 0 \ . \tag{9}$$

In two dimensions, the constraint set $g(x) = a$ is a curve, $\nabla g(x_*)$ is normal to this curve, the condition (9) says that Δx is perpendicular to $\nabla(x_*)$. This means Δx is tangent to the constraint curve. In more than two dimensions, there is a constraint surface and (9) says that Δx is tangent to this surface.

Now suppose $\nabla f(x_*) \neq 0$ and $\nabla g * (x_*) \neq 0$ and try to find a direction tangent to the constraint surface, condition (9), that improves f. One way to seek such a Δx is to modify $\nabla f(x_*)$ to get something that satisfies the constraint. We can subtract the "component" of ∇f in the ∇g direction. That is, try

$$\Delta x = s(\nabla f(x_*) - a \nabla g(x_*)) \ .$$

Substituting this into the constraint condition (9) gives

$$\left(\nabla f(x_*) - a \nabla g(x_*)\right)^t \nabla g(x_*) = 0 \ .$$

This leads to

$$a = \frac{\nabla f^t \nabla g}{\nabla g^t \nabla g} \ .$$

And from there

$$\Delta x = s \left(\nabla f - \frac{\nabla f^t \nabla g}{\nabla g^t \nabla g} \nabla g \right) \ .$$

and

$$\Delta f \approx s \left[(\nabla f)^t (\nabla f) - \left(\frac{\nabla f^t \nabla g}{\nabla g^t \nabla g} \right) (\nabla g)^t \nabla f \right]$$
$$= s \left[\nabla f^t \nabla f - \left(\frac{(\nabla f^t \nabla g)^2}{\nabla g^t \nabla g} \right) \right] .$$

The Cauchy Schwarz inequality above says that the quantity in square braces $[\cdots]$ is positive unless ∇f is in the same direction as ∇g.

Here's the conclusion: If there is a λ with $\nabla f(x_*) = \lambda \nabla g(x_*)$, fine. Otherwise, it is possible to move x along the constraint surface $g(x) = a$ to either increase or decrease f. If x_* maximizes or minimizes (optimizes) f on the constraint surface, then

$$\nabla f(x_*) = \lambda \nabla g(x_*) .$$

This λ is the *Lagrange multiplier*.

Lagrange multipliers for mean variance allocation

Suppose $x \in \mathbb{R}^n$ is an n component vector. You can think of x as a column vector if you're going to do linear algebra on it (multiply by a matrix). Otherwise, just know that x consists of the numbers x_1, ..., x_n. Mathematicians use the term *scalar* for numbers, or one component vectors. (The term comes from physics, where it means something more specific.) Suppose $f(x)$ is a *scalar valued* function of the vector x. That means that the numbers x_1, ..., x_n are combined in some way to get a single number $f(x)$. For example, if a_1, ..., a_n are other numbers, then there is the *linear* function

$$f(x) = a_1 x_1 + \cdots + a_n x_n = a^t x .$$

The expected return function is a linear function of the portfolio weights. The budget constraint also involves a linear function of the weights, with the specific *coefficient vector* $a = \mathbf{1}$.

A *quadratic* function is defined by an $n \times n$ symmetric matrix A as

$$f(x) = \sum_{j=1}^{n} \sum_{k=1}^{n} a_{jk} x_j x_k = x^t A x .$$

This formula doesn't require A to be symmetric. It makes sense even if $a_{jk} \neq a_{kj}$ (which is the same as $A \neq A^t$).