Notes on Itô calculus & quantitative trading

Sal Elder
sal.elder@yale.edu

August 2017

1 Maximizing profit while minimizing risk

Let's say there are several stocks i, each of which has a random return r_i over some fixed time period. Also say you have a good idea of the expected returns, $\bar{r}_i$, and also the covariances, v_{ij}, on these returns. If I buy fractions (or weights) x_i of the respective stocks, then my profit, P, is itself a random variable. The *expected* return on my investment is

$$\mathbb{E}(P) = \sum_i x_i \bar{r}_i. \tag{1}$$

We can model risk as the standard deviation squared, or variance, of R.

$$\mathbb{V}(P) = \sum_i \sum_j x_i x_j v_{ij}, \tag{2}$$

which is the general form for the variance of a weighted sum of correlated variables.

The critical idea of modern portfolio theory (MPT) is that for a given level of expected return, ρ, there is some minimum, unavoidable risk. Portfolios which minimize risk for a particular level of profit are said to lie on the "efficient frontier."

Such portfolios solve the optimization problem below.

$$\text{minimize } f(x) = \sum_i \sum_j x_i x_j v_{ij} \tag{3}$$

$$\text{such that } g(x) = \sum_i x_i = 1 \tag{4}$$

$$\text{and } h(x) = \sum_i x_i \bar{r}_i = \rho. \tag{5}$$

We can transform this problem into a set of linear equations—then any matrix solver will do. To transform a constrained maximization into a set of equations, use the method of Lagrange multipliers. If x is optimal, then ∇f at x must be perpendicular to the constraint surface defined by $g(x) = 1, h(x) = \rho$. This is equivalent to requiring

$$\nabla f = \lambda \nabla g + \mu \nabla h \tag{6}$$

for some unknown values of λ and μ (the "Lagrange multipliers").

Getting the gradient of f is a little tricky because of the double sum, but the standard calculus rules still apply. The tricky part is that there are many terms that depend on x_k (where k is a particular index). Some are of the form $x_k x_i$, others are of the form $x_i x_k$, and one is of the form x_k^2. If you like, you can write out the double sum (with ...'s as appropriate) and think about the different derivatives. Instead, I'm going to use a convenient notation called the Kronecker delta. This is a symbol δ_{ij} defined by

$$\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j. \end{cases} \tag{7}$$

That is, we define δ_{ij} as 1 if i and j are the same, and 0 if they are different. This way we can write a single expression for

$$\frac{\partial}{\partial x_k} (x_i x_j) \tag{8}$$

that will hold regardless of whether i, j, both, or neither are equal to k. The idea is that

$$\frac{\partial x_i}{\partial x_k} = \delta_{ik}. \tag{9}$$

When you say it in words, this is a simple idea. It just says that the partial derivative of a variable is 1 if it's the same variable you're taking a derivative with respect to, and 0 if they're different variables. In fact, you can think of (9) as stating what a partial derivative means: "treat other variables as constants."

Then apply the product rule for each term:

$$\frac{\partial f}{\partial x_k} = \sum_i \sum_j \left(\delta_{ik} x_j v_{ij} + x_i \delta_{jk} v_{ij}\right). \tag{10}$$

Generally speaking, a Kronecker delta inside a (single) sum will pick out the one term of the sum for which the indices match. In our case, the first delta picks out $i = k$ in the sum over i, and the second delta picks out $j = k$ in the sum over j. Thus

$$\frac{\partial f}{\partial x_k} = \sum_j x_j v_{kj} + \sum_i x_i v_{ik} \tag{11}$$

$$= 2 \sum_i v_{ki} x_i. \tag{12}$$

To go to the second line, I used the fact that covariances are symmetric: $v_{ik} = v_{ki}$. Then the two sums are the same, just with different names on their indices. You can rename both to i, since each sum is only over one variable. That's where the factor of two comes from.

Rewriting (6) in terms of the gradients we calculated, I get

$$2 \sum_i v_{ki} x_i = \lambda + \mu \bar{r}_k, \tag{13}$$

which must hold for all k. Each k represents a row of the vector equation (6), or in other words, a component of the gradients involved.

We are now in a position to reformulate the optimization as a matrix equation! To do that, first recall the definition of a matrix–vector product. Let's say that some matrix A has coefficients A_{ki}, indexed by row k and column i. If a vector $\vec{x}$ is indexed by x_i, then the kth row of the vector $A\vec{x}$ is given by

$$\sum_i A_{ki} x_i. \tag{14}$$

With this in mind, we can write a large "composite" matrix equation that will satisfy Eqs. (4), (5), and (13). It will help to mentally move $\lambda + \mu\bar{r}_k$ to the other side of the equation first. Also bear in mind that x, λ, and μ are the unknown quantities to appear in the matrix equation.

$$\begin{bmatrix} 2v_{11} & \cdots & 2v_{1N} & -1 & -\bar{r}_1 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 2v_{N1} & \cdots & 2v_{NN} & -1 & -\bar{r}_N \\ 1 & \cdots & 1 & 0 & 0 \\ \bar{r}_1 & \cdots & \bar{r}_N & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_N \\ \lambda \\ \mu \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \rho \end{bmatrix} \tag{15}$$

This matrix equation can be solved by any of the various tools out there (MATLAB, Python, R, Excel, etc.). The solutions for various values of ρ give you efficient portfolios for different levels of expected return. In the above formulation, there is no requirement that $x_i \geq 0$. In practice, negative weights may indeed arise. They correspond to short-selling, which we may want to avoid for various reasons. In that case the optimization problem to solve is

$$\text{minimize } \sum_i \sum_j x_i x_j v_{ij} \tag{16}$$

$$\text{such that } \sum_i x_i = 1, \tag{17}$$

$$\sum_i x_i \bar{r}_i = \rho, \tag{18}$$

$$\text{and } x_i \geq 0 \text{ for all } i. \tag{19}$$

This system cannot be solved with Lagrange multipliers, due to the presence of inequality constraints. It can be solved by a "quadratic programming" (QP) package, such as the `qp` module of `cnxopt` for Python. Just be aware that the QP folks often write the sums above in terms of matrices and vectors. The representations are equivalent, of course. In particular,

$$\sum_i \sum_j x_i x_j v_{ij} = x^T V x, \tag{20}$$

where V and x are the matrix and column vector versions of σ_{ij} and x_i, respectively.

2 Itô calculus

The QP formalism is only useful if we can estimate $\bar{r}_i$ and v_{ij} from historical price data. In order to do that, we need a model of correlated, random stock prices. One method of dealing with functions of random variables is the Itô calculus, which I describe here.

The basic building block of stochastic processes is Brownian motion, also called the Wiener process. You can arrive at Brownian motion from a limiting process of discrete random walks. Suppose we take n steps, where each step is an increment $s_i = \pm 1$ with equal probability of each direction. Let's also say that each step is independent of all the previous ones. The expected value of each step is 0, and the variance is $0.5(+1-0)^2 + 0.5(-1-0)^2 = 1$.

By the linearity of expectations, the expected value of the walk's final destination is 0:

$$\mathbb{E}\left[\sum_{i=1}^{n} s_i\right] = \sum_{i=1}^{n} 0 = 0. \tag{21}$$

Similarly, the rule for addition of variances gives us

$$\mathbb{V}\left[\sum_{i} s_i\right] = \sum_{i} 1 = n. \tag{22}$$

In order to get at the continuous version, we instead specify a total time t, which is split into n steps. We then take $n \to \infty$. If we suppose the variance of each step is t/n, then the variance of the walk is

$$\sum_{i} t/n = t, \tag{23}$$

irrespective of n. This variance also holds in the limit as $n \to \infty$. Standard Brownian motion is commonly denoted by B_t, with $B_0 = 0$. [1] It satisfies a few important properties. If (t, t') and (u, u') are non-overlapping time intervals, then $B_{t'} - B_t$ and $B_{u'} - B_u$

- are independent
- are normally distributed

[1] Sometimes I'll write $B(t)$ or just B.

- have 0 mean
- have variance $t' - t$ (or $u' - u$, respectively)

Brownian motion has no knowledge of the future. It is therefore a good choice for modeling things such as future stock prices. Furthermore, only the current value matters in predicting the future. Given knowledge of the entire trajectory all the way up to some current value B_t, our expectation for any future value is just B_t. This is called the martingale property.

It's worth mentioning that while standard Brownian motion has variance t, you can easily scale this. For some constant σ,

$$\mathbb{V}[\sigma B_t] = \sigma^2 \mathbb{V}[B_t] = \sigma^2 t. \tag{24}$$

The first equality is a generic property of variances, and the second is one of the properties of standard Brownian motion. (The earlier time is 0, where we know $B_0 = 0$.)

Although B_t is continuous everywhere, it is differentiable *nowhere*. Therefore, we have to be particularly careful when trying to use calculus on functions of B_t.

As an example, suppose Y was an ordinary, continuous, differentiable function of t. Also suppose we have some continuous and differentiable function $F(t, Y)$. The multivariable chain rule says that

$$\frac{\mathrm{d}F}{\mathrm{d}t} = \frac{\partial F}{\partial t} + \frac{\partial F}{\partial Y}\frac{\mathrm{d}Y}{\mathrm{d}t}, \tag{25}$$

which can also be written as

$$\mathrm{d}F = \frac{\partial F}{\partial t}\,\mathrm{d}t + \frac{\partial F}{\partial Y}\,\mathrm{d}Y. \tag{26}$$

The problem is that for functions $F(t, B_t)$, we can't apply (25). That's because $\mathrm{d}B/\mathrm{d}t$ does not exist! In order to calculate $\mathrm{d}F$ for functions of Brownian motion, we must use Itô's lemma. Here it is: [2]

$$\boxed{\mathrm{d}F = \left(\frac{\partial F}{\partial t} + \frac{1}{2}\frac{\partial^2 F}{\partial B^2}\right)\mathrm{d}t + \frac{\partial F}{\partial B}\,\mathrm{d}B.} \tag{27}$$

[2] Some authors give a more general version, but we will only need the formula above.

The only new feature in this rule is the term involving a second derivative. One way to gain intuition for Itô's lemma is to think of it as a Taylor series, where we have replaced $\mathrm{d}B^2$ with $\mathrm{d}t$ and neglected higher powers of $\mathrm{d}t$. This is not a derivation, but it can help you remember the formula and get a better idea of what's going on.

To see why you might want to replace $\mathrm{d}B^2$ with $\mathrm{d}t$, first recall the following rule for variances.

$$\mathbb{V}[A] = \mathbb{E}[A^2] - \mathbb{E}[A]^2, \tag{28}$$

where A is "anything." Applying this rule to an interval Δt of Brownian motion, we have

$$\mathbb{V}[\Delta B] = \mathbb{E}[\Delta B^2] - \mathbb{E}[\Delta B]^2. \tag{29}$$

We already know that the variance is Δt, and that the expectation of ΔB is 0. Thus

$$\mathbb{E}[\Delta B^2] = \Delta t, \tag{30}$$

which is always true for finite intervals. So it isn't *too* much of a stretch to apply it to infinitesimals as well.

An equivalent way of writing Itô's lemma is the integral version,

$$F(t') - F(0) = \int_0^{t'} \left(\frac{\partial F}{\partial t} + \frac{1}{2}\frac{\partial^2 F}{\partial B^2} \right) \mathrm{d}t + \int_0^{t'} \frac{\partial F}{\partial B} \, \mathrm{d}B_t. \tag{31}$$

The first integral is over time. The meaning of the second integral is not at all obvious. It has a differential, $\mathrm{d}B_t$, which is itself a random variable. It can be defined like the ordinary Riemann integral, where we take a limit of increasingly narrow rectangles:

$$\int_0^{t'} g(t, B_t) \, \mathrm{d}B_t = \lim_{n\to\infty} \sum_i^n g(t_i, B_i) \, \Delta B_i. \tag{32}$$

Here, the width, ΔB_i, as well as the height, $g(t_i, B_i)$, of the rectangles are random. Itô's lemma tells us how to perform such integrals, by putting them in terms of things which we already know how to perform.

3 Geometric Brownian motion

You can also use Itô's lemma to solve stochastic differential equations, which are differential equations written in terms of stochastic variables. For example, stock prices are commonly modeled by

$$\mathrm{d}S = \mu S \,\mathrm{d}t + \sigma S \,\mathrm{d}B. \tag{33}$$

Here, changes in price are proportional to the current price, which makes sense: Larger companies should expect larger profits, losses, and random fluctuations. Itô's lemma gives us conditions on S. Generally, any $S(t, B_t)$ satisfies

$$\mathrm{d}S = \left(\frac{\partial S}{\partial t} + \frac{1}{2}\frac{\partial^2 S}{\partial B^2}\right)\mathrm{d}t + \frac{\partial S}{\partial B}\,\mathrm{d}B. \tag{34}$$

Equating the $\mathrm{d}t$ and $\mathrm{d}B$ terms of (33) and (34) gives us two conditions on S:

$$\frac{\partial S}{\partial B} = \sigma B, \tag{35}$$

and

$$\frac{\partial S}{\partial t} + \frac{1}{2}\frac{\partial^2 S}{\partial B^2} = \mu S. \tag{36}$$

Although $S(t, B_t)$ is a function of a random variable, at this point we are only concerned with the form of S in terms of its explicit B– and t–dependence. In other words, we have used Itô's lemma to remove all the stochastic-ness from the problem. Now it's just like any other coupled system; you don't need any stochastic calculus at all to solve (35) and (36).

Start with (35): its solution is

$$S = c(t)e^{\sigma B} \tag{37}$$

for some unknown function c of the other variable, t. (It's a "constant" with respect to B.) Plug this result into (36) to get

$$c'(t)e^{\sigma B} + \frac{1}{2}c(t)\sigma^2 e^{\sigma B} = \mu c(t)e^{\sigma B}. \tag{38}$$

A little bit of algebra goes a long way in making this more transparent. We have, equivalently,

$$c'(t) = \left(\mu - \frac{\sigma^2}{2}\right)c(t). \tag{39}$$

That's an *ordinary* (not partial) differential equation in t, and it has the solution

$$c(t) = k \exp\left[\left(\mu - \frac{\sigma^2}{2}\right) t\right] \tag{40}$$

for any constant k. Combining (37) and (40) and putting in the initial condition $S(0)$ gives you the solution to (33).

$$\boxed{S(t) = S(0) \exp\left[\left(\mu - \frac{\sigma^2}{2}\right) t + \sigma B_t\right].} \tag{41}$$

Evidently, the solution exhibits geometric Brownian motion: it's the exponential of Brownian motion with a drift term. It's also said to have a lognormal distribution; the *logarithm* of $S(t)$ is normally distributed with mean $\ln S(0)+(\mu-\sigma^2/2)t$ and variance $\sigma^2 t$.

4 Measuring model parameters

I still haven't addressed how we can calculate $\bar{r}_i$ and v_{ij}, or even measure μ_i and σ_{ij}. The whole reason we needed a concrete model was that the expected returns and covariances we want to use are *future* variables—we cannot sample them! Instead, we can observe the logarithms of historical price data, which should have the form

$$\log S_i(t) = \log S_i(0) + \left(\mu_i - \frac{\sigma_i^2}{2}\right) t + \sigma_i B_i(t) \tag{42}$$

for each stock i. In practice, the data are separated by some time-discretization, Δt. Then the increments are

$$\Delta \log S_i = \left(\mu_i - \frac{\sigma_i^2}{2}\right) \Delta t + \sigma_i \Delta B_i. \tag{43}$$

These increments form the set of observations we need! Now we have many numbers pulled from the same distribution, and you can use them to calculate the sample covariances as well.

Actually, it's not immediately clear what it means to have correlated Brownian walks. One way to *build* them is to start with a collection of independent, standard Brownain walks, then combine them in different proportions. Concretely, let's say we have several independent walks $\{X_i(t)\}$, from which we create two related processes $A(t)$ and $B(t)$.

$$A(t) = \sum_i a_i X_i(t) \tag{44}$$

$$B(t) = \sum_i b_i X_i(t) \tag{45}$$

Now we can calculate the covariance between ΔA and ΔB over the same interval Δt. It is

$$\mathbb{E}\left[\Delta A \Delta B\right] = \mathbb{E}\left[\sum_i \sum_j a_i b_j \Delta X_i \Delta X_j\right] \tag{46}$$

The first observation to make is that most of these terms have 0 expectation. That's because the ΔX_i's were chosen to be independent. In fact, the *definition* of independence says that

$$\mathbb{E}\left[\Delta X_i \Delta X_j\right] = 0 \tag{47}$$

if $i \neq j$. We know the expectation for terms where $i = j$, too. The expectation of ΔX_i^2 is the variance of ΔX_i, and that's Δt. Combining these facts with the linearity of $\mathbb{E}$, the covariance between ΔA and ΔB is

$$\Delta t \sum_i a_i b_i. \tag{48}$$

There are two important conclusions to take away here.

1. There is such a thing as correlated Brownian motion.
2. Its covariance is proportional to Δt.

That's nice, but you may wonder what the point is *procedurally*. Here's the recipe. First calculate the sample covariance matrix of $\Delta \log S_i$. Then when you're done, divide it by Δt! That will give you σ_{ij}. Now for any future length of time t for which we care to hold our stocks, the log–prices should have covariance $\sigma_{ij} t$.

To get the parameters μ_i, observe that the sample means have the form $(\mu_i - \sigma_i^2/2)\Delta t$. So once we have the σ_i's, it's easy to do a little algebra and get the μ_i's.

Up to this point we have measured μ_i and σ_{ij}, but not $\bar{r}_i$ and v_{ij}. The first two are GBM parameters characterizing—in some sense—the random motion of the stock prices. The other two tell us what we really want—the expected returns and covariances on the actual stock prices! The final ingredient, therefore, is to relate the model parameters to the price statistics.

First let's take care of the expected return, $\bar{S}(t)/S(0)$. The deterministic factor in (41) will merely scale it, so let's start with the expected value of $\exp\left[\sigma B_t\right]$, where B_t is normally distributed with mean 0 and variance t. Recalling the normal distribution function, we can write the integral explicitly:

$$\mathbb{E}\left[e^{\sigma B}\right] = \int_{-\infty}^{\infty} \mathrm{d}B\, p(B)\, e^{\sigma B} \tag{49}$$

$$= \frac{1}{\sqrt{2\pi t}} \int_{-\infty}^{\infty} \mathrm{d}B\, e^{-B^2/2t+\sigma B}. \tag{50}$$

There is a trick to solving this integral; it involves completing the square on the exponent. Observe that

$$-\frac{B^2}{2t} + \sigma B = -\frac{1}{2t}\Big(B - \sigma t\Big)^2 + \frac{\sigma^2 t}{2}. \tag{51}$$

Pulling the B–independent term out of the integral, we are left with

$$\mathbb{E}\left[e^{\sigma B}\right] = e^{\sigma^2 t/2}\left[\frac{1}{\sqrt{2\pi t}} \int_{-\infty}^{\infty} \mathrm{d}B\, e^{-(B-\sigma t)^2/2t}\right]. \tag{52}$$

There is a curious interpretation of the factor in large square brackets. It is the integral of some strange new normal distribution with mean σt and variance t. I don't see any particular reason why that distribution is meaningful, but like all probability distributions, it must integrate to 1! We are therefore left with

$$\mathbb{E}\left[e^{\sigma B}\right] = e^{+\sigma^2 t/2}. \tag{53}$$

To get the expected return of the stock, tack back on the deterministic factor from (41). We are left with a very nice result:

$$\mathbb{E}\Big[S_i(t)/S_i(0)\Big] = \boxed{\bar{r}_i = \exp(\mu t).} \tag{54}$$

Evidently the mean return is slightly greater than the median, or *most likely*, return. That would correspond to $B = 0$ and is given by $\exp[(\mu - \sigma^2/2)t]$.

Although the logarithm of S is normally distributed, the distribution of S itself is asymmetric; the mean occurs a bit downstream from the peak.

The covariances are more tedious to derive.[3] Anyway, the result we need is

$$\operatorname{Cov}\left[\frac{S_i(t)}{S_i(0)}, \frac{S_j(t)}{S_j(0)}\right] = \boxed{v_{ij} = e^{(\mu_i+\mu_j)t}\left(e^{\sigma_{ij}t} - 1\right).} \tag{55}$$

[3] You can look up the properties of lognormal distributions in general, then apply them to our case. The fact that our variables arise from GBM is not relevant once you know the distribution.