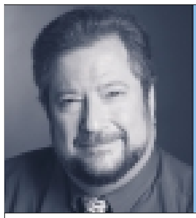

Aaron Brown

Hard-Target Search

Searching for risk sure is a lot more fun when you know what you're talking about ...

The most memorable scene in the 1993 movie *The Fugitive* is Tommy Lee Jones' speech to his searchers: "Alright, listen up, people. Our fugitive has been on the run for ninety minutes. Average foot speed over uneven ground barring injuries is four miles per hour. That gives us a radius of six miles. What I want from each and every one of you is a hard-target search of every gas station, residence, warehouse, farmhouse, henhouse, outhouse, and doghouse in that area. Checkpoints go up at fifteen miles. Your fugitive's name is Dr. Richard Kimble. Go get him."

If you want to be a successful risk manager, study this speech. It's a masterpiece of tough-sounding, take-charge direction. It uses simple quantitative data and computation to arrive at precise instructions. There's no doubt that the guy knows what he's doing and should be promoted, whether Kimble escapes or not.

If you want to be a useful risk manager, study this speech for traps to avoid. Tommy has two pieces of quantitative information. One is the average foot speed over uneven ground barring injuries. We don't know where this comes from, and we don't know that it's relevant. Is our fugitive average? Is he on foot and traveling over broken ground? Is he moving in a straight line? The other piece of information is that Kimble has had a 90-minute headstart.

Mr Jones proved an instant hit with the risk management team

Tommy multiplies these two numbers correctly, deftly adjusting for the difference in units. This is an impressive achievement in a big-budget movie. But he misuses the conclusion. For one thing, he sets the boundaries of investigation at the average level. For another, he neglects dynamics. Kimble can keep moving during the search, either to get outside the six-mile limit or to move to a place that was previously searched.

More important is the quant's fallacy, to focus on what you can compute instead of what you want to know. It is far more efficient to fan out from Kimble's last known location than to search intensively but at random anywhere he might be. Every time you find evidence that Kimble passed through an area, you can recenter your search. Just because you can compute a boundary for his likely location doesn't mean that's how to organize the search.

The computation of potential area is useful, however, for a different reason. The fictional location of the search is a semi-rural area in Illinois. Such an area is likely to have a density of 1,000 residences per square mile (the scene was filmed in near Dillsboro, North Carolina, 317.2 residences per square mile). πr2 tells us that's 36,000 houses, even using the lower figure from the rural area of filming, and 100,000 buildings of all types. Tommy is talking to about 25 people, so he's just assigned about 4,000 searches to each one.

As far as I can tell, there is no such thing as a "hard-target search." "Hard target" is a military term with several related definitions. The most relevant to this context is a target that requires specific attention to destroy, due to fortification, camouflage or armament. In contrast, a soft target is something likely to be destroyed by any mass attack, such as a bomb in its general vicinity, and that is not likely to present a d anger. By analogy and context, a "hard-target search" would mean looking for someone who might be hiding, and who might be dangerous, as opposed to walking around yelling the name of a lost hiker.

My guess is that you would want at least two armed and trained searchers to take at least half an hour to do a hard target search of a house, not counting getting the search warrant. If the house is unoccupied, they have to break in. If it's occupied, they have to negotiate with residents, and worry about hostages. They have to look in closets and attics and under beds, and be alert for ambush at all times. 100,000 of these searches are going to take quite a bit of time.

Tommy's problem is that it would take too long even to perform a soft-target search. If his people split up and walked at four miles per hour on optimal paths, and automatically found Kimble if they came within 100 yards of him, it would take about ten hours to cover the area (during which Kimble could move 40 miles, which would take three weeks to search, during which he could get anywhere in the United States or Canada).

Tommy gives no quantitative reasoning for his other instruction, to set up road checkpoints at 15 miles. By analogy with his search radius, it suggests that he thinks that ten miles per hour is the average speed over roads. That's a high value on foot (so is four miles per hour on broken ground) but not impossible for a serious runner. It's obviously much too low a value if Kimble gets a ride, or steals a car or even a bicycle. Tommy also has neglected the possibility that Kimble would start out on a road to go, say, ten miles, then go off-road to find a hiding place.

Dimensionality is the friend of the fugitive. If Kimble is restricted to a linear path, say a road, the searchers' task is easy. Two people can search the road as rapidly as Kimble can run, or they can pick two boundaries and wait. In two dimensions, it takes 25 people ten hours to cover the area that Kimble can get to in 90 minutes. Add a third dimension and it would take 250 people 70 hours, moving at four miles per hour, to get within 100 yards of every point in a six-mile sphere. By the time we get up to seven dimensions, it will take almost 12 million searcher-years to cover Kimble's possible range.

This assumes that all the dimensions are "flat." Houses and other hiding places can be thought of as extra "curled-up" dimensions within a two-dimensional search. The reason that the slow hard-target search is necessary is that it's not enough to get within 100 yards of Kimble; you must know additional things about his location. These additional things can be thought of as dimensions, but ones that vary from place to place within the two flat dimensions. At one place is a house, with extra dimensions like what room he is in and his position within the room. Another place might be in the woods, with extra dimensions like whether he is hiding in a hole, or up a tree.

Tommy gives no quantitative reasoning for his other instruction, to set up road checkpoints at 15 miles

In risk management, we are searching for plausible future events that can hurt us. In the easiest case, we can do a soft-target search of a low-dimensional space. That's appropriate for a portfolio that depends smoothly on a small number of market factors. A single-currency, unlevered, investment-grade bond portfolio is an example. In other cases, we may have complex, levered instruments which require extensive calculation at of possible future scenarios. That's like a hard-target search. In still other cases, our portfolio depends on a lot of market factors, so we have to search a high-dimensional space. The most difficult cases combine both issues.

High-dimensional searches are possible only due to a wonderful idea, invented by Stanislaw Ulam, Nicholas Metropolis and John von Neumann: Monte Carlo. Suppose that instead of hunting fugitives, Tommy was counting trees (a duller movie plot). He could divide his six-mile circle into 3.5 million 10-yard squares, and pick 100 of them at random. His searchers could go out and count the number of trees in the selected squares, four squares per searcher – much easier than 4,000 hard-target searches for a dangerous and clever fugitive. The total multiplied by 35,000 should be a good estimate of the number of trees in the circle. The great thing about Monte Carlo, the reason it is said to lift the curse of dimensionality, is that the accuracy of this method is independent of the number of dimensions, whether flat dimensions or curled-up ones. As long as you select your locations at random, having 100 locations is equally good in high- and low-dimensional space.

Unfortunately, selecting possible future values of market factors at random is tricky in the multivariate case. Let's start with one variable: how much the one-year USD treasury rate might move tomorrow. From 1962 to 2008, the standard deviation of the move has been nine basis points. If it had a normal distribution, the probability is 0.000063 of a move more than four standard deviations on any day, so we would expect 0.36 examples over the 26 years. Actually, there were 124 such days. More extreme values than predicted by a normal distribution is typical of financial data – it is the "fat tail" problem I discussed in the last issue of *Wilmott*.

The ten-year USD treasury rate had 95 days with moves beyond four standard deviations over the same period. If I assume that the one-year and ten-year treasury rate changes have a bivariate normal distribution, then I can draw an ellipse, using the standard deviations of the two variables plus their correlation, such that only 0.000063 fraction of the observations should be outside the ellipse. In fact, 229 days were outside the ellipse.

At first glance, you might add the 124 days that the one-year rate moved more than four standard deviations to the 95 days that the ten-year rate did the same, to get 219 and say that there were only ten additional outliers when we look at both variables at once. But on 56 days, both the one-year and ten-year rates moved more than four standard deviations, so there were only 163 days on which either the one-year or ten-year rate, or both, moved more than four standard deviations. That leaves 66 days on which both the one- and ten-year rates moved less than four standard deviations, but their combined move was outside the bivariate normal ellipse. December 7, 2001 was one such day. The one-year rate fell nine basis points, while the ten-year rate went up 16. Both moves were less than two standard deviations, but the combination was nearly impossible under a bivariate normal distribution, given the 0.74 correlation between the two. The move was not large, but it was in a direction in which the ellipse is narrow.

In principle, one could fit a multidimensional distribution that had non-elliptical curves. There are two problems with that

If I add in an overnight rate as well, the number of outliers jumps to 328. That includes 52 days in which none of the three rates individually moved more than four standard deviations. When I include a five-year rate, there are 366 outliers (108 "new" ones in which none of the four individual rates moved more than four standard deviations, but fewer than 70 days on which individual or joint moves were outside the lower-dimensional ellipse but inside the higher dimensional).

This process of additional outliers continues as you add more market factors. Whenever I see a covariance matrix with hundreds of rows and columns, I assume that that every single simulated point is closer to the mean in cumulative probability than any actual observation, even if there are thousands of simulated and actual points. If you don't believe me, ask yourself why no one ever tells you this obvious and easy-to-compute measure of simulation quality.

This is different from the problem of fat tails. Fat tails mean that if you simulate assuming a normal distribution, you will miss a few extreme movements seen in actual data. The multivariate problem is that you will miss the entire distribution. Your simulated points will have no overlap with the actual data used to fit it (and, of course, no chance of overlap with what happens tomorrow).

People sometimes use copulas to attack this problem. Copulas can address the problem of fat tails for each univariate distribution, but the multivariate problem arises from the correlation structure. A Gaussian copula has exactly the same issues as the multivariate normal distribution, regardless of what marginal distributions you use. In most alternatives to Gaussian, people preserve the elliptical shape of the iso-probability curves, and therefore cannot solve the problem.

If you graph multivariate financial data you will virtually always see a structure too complex to draw an ellipse whose boundary points have the same cumulative probability distance from the mean. Unless you draw a more complex shape, you'll have some regions with lots of points far outside the boundary and other regions with all the points much closer to the mean than the shape boundary. All you have to do is look at any set of financial data to make this point obvious. Many quants spend too much time running code on data and not enough time looking at it.

In principle, one could fit a multidimensional distribution that had non-elliptical curves. There are two problems with that. The first is mathematical tractability. The second is parameters: a covariance matrix, with $N(N+1)/2$ parameters for N variables. Shapes that imply higher-order dependence than correlation require more parameters by a factor of N for each increase in order. In most applications, we don't have enough data for reliable covariance matrices. Without some sound theory about what shapes should be, and no one has a clue about that, this approach is hopeless.

Another solution is to restrict analysis to low-dimensional models. This is sufficient for a lot of instrument pricing, and even for portfolio and risk management within a single security type. But some instruments require many parameters, some portfolios attempt simultaneously to optimize within and across asset types, and some risk management is done at a cross-market level.

I have one simple approach that works well in many financial applications. The multivariate normal distribution specifies that the conditional distribution of any variable always has the same standard deviation. Using the data above, for example, the expected move in the ten-year treasury rate is 0.55 times the move in the one-year treasury rate. The standard deviation of the residual, ten-year move minus 0.55 times one-year move, should be 4.5 basis points, independent of how much the one-year treasury moves.

One-year treasury rate move	Standard deviation of residual (ten-year move minus 0.55 times one-year move)
Less than 1 standard deviation	3.7 basis points
1 to 2 standard deviations	6.2 basis points
2 to 3 standard deviations	8.4 basis points
3 to 4 standard deviations	9.3 basis points
More than 4 standard deviations	12.5 basis points

The actual data look quite different:

Rather than the conditional standard deviation being constant, it seems more like it is proportional to the one-year treasury rate move. To put it geometrically, the direction of a move is independent of its direction. In multivariate normal distributions, large moves have to be close to the expected direction, because a small error in the angle makes a large total residual. The biggest difference between real financial data and multivariate normal data is that the market makes large moves far from the expected direction. This is true even if you correct the univariate distributions for fat tails, and is a fatal error of assuming normality in most financial applications.

It's quite easy to correct this defect. You can simulate a future market move by picking two historical days at random. Take the direction from one day and the size of the move from the other. Repeat (with replacement) until you get as many samples as you need.

You will get some implausible simulated values. Some very small moves will be in unlikely directions. It's easily possible for the one-year treasury rate to move up one basis point while the ten-year treasury rate moves down two basis points; but it's highly unlikely for the one-year treasury rate to move up 20 basis points while the ten-year treasury moves down 40 basis points. I like to leave these implausible simulated points in. There aren't very many of them, and it rarely hurts to explore a bit beyond what you expect. But if you prefer, you can put a lower limit on the move size for the day from which you select the direction. That will reduce, but not eliminate, the problem. In most risk management applications, having a few implausibly extreme points doesn't hurt. The multivariate normal distribution, which in high-dimensional applications produces mainly, or only, points less extreme than any actual observation, hurts a lot.

The only tricky part of the algorithm is defining move size. One way is to form the covariance matrix of all the variables, and invert it. If you pre- and post-multiply the inverse covariance matrix by the vector of returns for a day, then take the square root of the result, you will get a scalar corresponding to the move size. I don't like this, because I don't trust covariance matrices or matrix inversion, but it gives reasonable results in many cases. A simpler, and more robust, way is to divide each univariate move by its standard deviation, square the results, add them up across all variables, and then take the square root of the result. Ideally, the definition of size would correspond to dimensions of specific interest to your application, concentrating on the market factors and spreads of most importance to your P&L. But any reasonable definition of size should work pretty well, since only the univariate distribution of relative size matters.

Once you have a move size for each day, you select two days at random. Form the ratio of the move size of day 1 divided by the move size of day 2, and call it R. Multiply all the returns from day 2 by R, and you have your simulated move.

There is even some economic intuition for this approach. It was Benoît Mandelbrot who first noticed that financial time series look the same at a wide range of time scales. If someone shows you a financial time series with the scales removed and the tick size obscured, it's hard to tell if you're looking at one minute or 50 years of data.

It's quite easy to correct this defect. You can simulate a future market move by picking two historical days at random

There are traders watching each trade trying to make a tick, long-term fundamental investors looking at ten-year Sharpe ratios, and everyone in between – all trying to buy low and sell high at different scales. It's plausible that the interaction of market forces will produce similar results (i.e., a similar distribution of vector directions of market moves) and all scales (i.e., at all move sizes). This is called "self-similarity" or "scale invariance." While it is not exactly true in financial data, it's a better simple approximation than multivariate normal.

In *The Fugitive*, Kimble gets away. In financial modeling, my approach only corrects for a gross defect of most cross-market financial modeling; it doesn't give a clear prediction of the future. It forces consideration of large moves in directions previously only seen on small move days. That's better than not considering such moves, but it won't protect you from all surprises.

W