randPort vignette

Mike Flynn, Angel Zhou, Dave Kane

July 11, 2013

A simple case to consider is one with only 3 stocks in the universe, say GE, IBM and Coca-Cola (KO). Assume that there is some data available about these stocks:

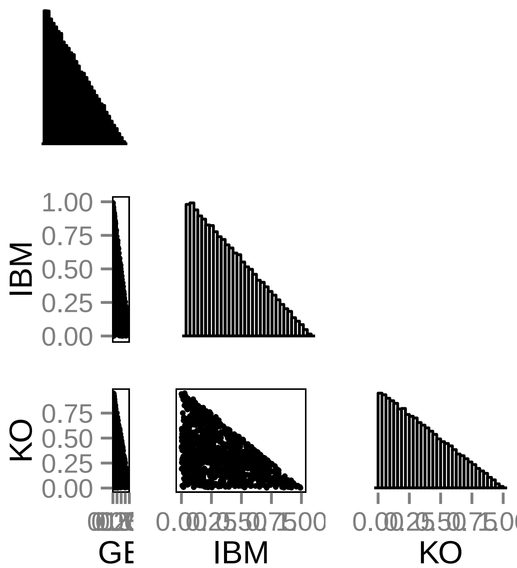
```
set.seed(40)
stocks <- data.frame(ticker = c("GE", "IBM", "KO"), sizerank = c(1, 0, -1))
stocks

##   ticker sizerank
## 1     GE        1
## 2    IBM        0
## 3     KO       -1
```

The most basic constraint is the long-only constraint i.e. the weights of the stock must add up to one and be positive. This can be represented by a linear equation: let the weights of "GE","IBM", and "KO", be x, y, and z, respectively. Then $x + y + z = 1$, $x, y, z \geq 0$ are the only constraints. The `randPort` package is designed to sample underdetermined equations of the form $Ax = b$, with $x \geq 0$ and so it is ideal for this set of constraints. In this case $A = [1, 1, 1]$ and $b = 1$. With `hitandrun`, we can easily obtain a set of weights that are randomly distributed and fulfill the constraints.

```
A <- matrix(c(1, 1, 1), nrow = 1, ncol = 3)
A

##      [,1] [,2] [,3]
## [1,]    1    1    1

b <- 1

w <- hitandrun(A, b, n = 1000)
points <- as.data.frame(t(w))
colnames(points) <- stocks$ticker
head(points)
```

```
##         GE     IBM     KO
## 1 0.10506 0.03877 0.8562
## 2 0.07990 0.83320 0.0869
## 3 0.09234 0.80191 0.1057
## 4 0.14397 0.44635 0.4097
## 5 0.43460 0.26308 0.3023
## 6 0.13072 0.19263 0.6766
```

The pairwise scatters display the fact that any pair of variables has a sum bounded by 1, but can be anywhere between that line and the axes.



Perhaps now we want to match a portfolio that has an exposure to size rank of 0.5. This is the same thing as adding a new row to A and b, corresponding to the column "sizerank" of our stock data and 0.5 respectively.

```
A <- rbind(A, stocks$sizerank)
b <- c(b, 0.5)

w2 <- hitandrun(A, b, n = 1000)
points2 <- as.data.frame(t(w))
colnames(points2) = colnames(points)
head(points2)

##         GE     IBM     KO
## 1 0.10506 0.03877 0.8562
## 2 0.07990 0.83320 0.0869
## 3 0.09234 0.80191 0.1057
## 4 0.14397 0.44635 0.4097
```

```
## 5 0.43460 0.26308 0.3023
## 6 0.13072 0.19263 0.6766
```

Running this code is equivalent to sampling from the positive solution space of $x+y+z=1$ and $x-z=.5$, which corresponds to the line segment parameterized by $r(t)=[t,1.5-2t,t-.5]$ with $.5\le t\le 0.75$. This is very clear when looking at pairwise plots of the variables:

Geometric Nature of the Problem

It is easier to picture the space of each problem when looking at things geometrically. Each row of Ax=b corresponds to a linear equation $c_1x_1+c_2x_2+\cdots+c_nx_n=b_m$. Geometrically, this is an (n-1)-plane in $\mathbb{R}^n$ which our solution must lie in. When there are m rows, our solution must lie in the intersection of those m planes which is itself an $(n-m)$ plane. A concrete example is the 3 stock case with only the long-only constraint. This is the peice of the $(3-1=2)$-plane $x+y+z=1$ that is in the positive quadrant. It forms a triangle. If another constraint is added, we can see this as the intersection of two 2-planes, also known as a line, as you can see via these graphs. When we sample with our function, the random points are distributed uniformly along these shapes.

Caption1: Consider the case with only 3 stocks and the only constraint is that we are fully invested, long-only. Here are pairwise scatterplots of the variables, as well as histograms for each variable on the diagonals. For the scatterplots, 1000 samples were taken. For the histograms, 100,000 samples were taken to increase resolution. If all the portfolios are distributed uniformly, we should have that all the distributions are identical. Notice that it seems like the probability distribution seems to decrease linearly as each variable increases. The reason can be figured out by looking at the pairwise plots. Each variable is a function of the other 2. For example: $z=1-x+y$. The most probable value of $x+y$ corresponds to the line $y=c-x$ with the most points on it, or the longest line. As can be seen in the distributions, the longest such line is $y=1-x$, therefore the most probable value for $x+y$ is 1 and the most probable value for z is 0.

Caption2: Perhaps we add another constraint that we must be "sizerank" neutral. Here are scatterplots and histograms again with the sample sample sizes for each. This effectively slices our current plane by the plane $x-z=0$. As derived in the text, this curve can be parameterized by $r(t)=[t,1-2t,t]$ with $0\le t\le 0.5$. This curve can be checked by looking at the pairwise scatter plots (check the end points). Since there is no difference between the number of points at one point of the line vs. another, all the variables are uniformly distributed as can be seen by the histograms.

Caption3: Here it is easy to visualize the simplex, mapped to 2-d coordinates. Every point on the simplex is fully invested, long only. Each vertex is the single position portfolio for GE, IBM, or KO, respectively. The points generated are uniformly distributed on this triangle.

Caption4: This is the corresponding slice of the simplex. As you can see, only certain points, on a line, are allowed. Full investment in IBM is allowed because it has a natural exposure to "sizerank" of 0, and the half and half portfolio with GE and KO is allowed because their average "sizerank" is 0. The line connecting them is the set of valid portfolios.