Portfolio Performance Evaluation via Characteristic-Matched Random Portfolios

Jack Luby

11/20/2019

Abstract

The portfoliowalkr package develops and implements a new benchmark construction methodology which allows for the comparison of long-only, single currency equity portfolios to randomly generated characteristic-matching benchmarks. The set of portfolios which match the characteristics of a target is represented by the complete solution space to Ax = b and the N-simplex. In representing an exhaustive set of alternative portfolios which have no allocation bias, randomly sampling from this space provides a fair counterfactual set of investments. We propose that comparison against a large set of randomly generated, characteristic-matched counterfactual benchmarks provides an accurate metric for a manager's selection ability.

1 Introduction

The assessment of a portfolio's performance principally requires the establishment of a benchmark against which the portfolio can be compared. Typically, a portfolio is said to be successful if its return is greater than that of the benchmark. As such, the establishment of a fair and reasonable benchmark is important.

A benchmark serves to control for market movement beyond the manager's control. Consider a portfolio which has returned 5% in the past year. Were this portfolio to exist in a universe where the average asset returned 25%, this portfolio's performance would be considered abysmal. Conversely, were it to exist in a universe where the average asset lost 15%, its performance would be excellent. The orientation of a portfolio's performance within its own universe is intuitively necessary. A manager should be judged on stock selection ability, not general market fluctuation.

This intuition can be extended to benchmarks of greater nuance. Consider a mutual fund which, by mandate of its patriotic primary financier, invests only in American assets. Over the past year, this fund returned 10%. In that same period of time, American assets vastly underperformed booming international market growth of 25%, actually slipping by 5%. Should this fund's manager be criticized for failing to match the international market's growth, or lauded for managing 15% overperformance of the declining American market? Certainly the latter. In this case, the fund's benchmark should not be the whole of the global asset universe. It should be limited to match the characteristics of their portfolio.

This phenomenon applies yet more generally. A "stock-picking" manager only claims to be advantaged in their ability to pick superior stocks from a set of apparently similar ones. These managers should not be praised for, say, happening to be heavily exposed to successful industrial assets. Nor should they be criticized for being underexposed to them. Unless an advantage beyond stock-picking has been claimed and pre-declared, a portfolio's characteristic exposures should not advtantage or disadvatage the portfolio relative to a benchmark. Portfolio performance should be considered on a basis relative to benchmarks which control for passively allocated characteristics. Several approaches to performance assessment have sought to address this consideration, namely matching portfolio (MP) benchmark construction (Bartz and Kane (2010)) and regression-based performance attribution (Lu and Kane (2013)).

The **portfoliowalkr** package develops and implements a new benchmark construction methodology which allows for the comparison of long-only, single currency equity portfolios to randomly generated characteristic-matching benchmarks. The set of portfolios which match the characteristics of a target is represented by the complete solution space to Ax = b and the N-simplex. In representing an exhaustive set of alternative portfolios which have no allocation bias, randomly sampling from this space provides a fair counterfactual set of investments. We propose that comparison against a large set of randomly generated, characteristic-matched counterfactual benchmarks provides an accurate metric for a manager's selection ability. We will look to compare this metric to other characteristic-matching performance measurements and demonstrate the package's use.

2 Matching Portfolios

Bartz and Kane (2010) develop a benchmark construction methodology which matches assets to near relatives on selected covariates. A nearest match, as defined by Rosenbaum and Rubin (1983) propensity score, is found for each asset in the target portfolio. Following from the Rubin (1973) causal model, the authors propose that these matched assets serve as asset-level counterfactuals. Because the matching portfolio (MP) benchmark contains a near-identical non-holding for each holding, the target portfolio's excess return is attributable to the selection 'treatment.' Therefore, any difference in return between the portfolio and matched benchmark is attributable to the manager's stock-picking ability.

The Rubin model makes use of propensity score matched units to directly control for covariates and their interactions in observational studies. Because each untreated observational unit has a highly similar treated match, any difference between the pair is attributable to treatment. Most commonly, propensity scoring is applied to characteristics such as age, race, gender, health status, and other simple attributes among a large pool of observational units. In these cases, a close match is readily available. Each unit either remains unweighted or is weighted relative to its

demographic proportion of the population.

Despite the robust implications of the Rubin causal model, it is not a perfect analog for the stock market. In benchmark creation applications, a close propensity score match may not be available due to the complexity of characteristics and limited size of the pool of assets. In these cases, attempting to match at the unit-level comes at the expense of bias. Placing a close but not-so-similar match into the benchmark portfolio creates significant susceptibility to bias, especially when a given unit weight is large (Austin, Jembere, and Chiu (2018)). Though the authors attempt to address the small sample / large weight concern with a bootstrapped sampling approach, iteratively creating new matched portfolios with next closest matches, this approach magnifies bias. As the matched units become less and less similar to the target, their applicability as comparison units becomes dubious. Bootstrapped matched sampling in this context is likely to sacrifice much of the interaction effect capturing power of propensity scoring and at the limit simply become a metric of performance relative to the universe.

A final theoretical consideration lies in the weighting of MP units. Propensity scoring traditionally applies to a context in which a unit's weight is non-variable. In building a portfolio, the weight applied to a given security can range from 0%-100%. Imagine a portfolio which happens to have .1% of its weight allocated to each of the top 100 securities by return and 1% allocated to the next 90 top returners. With matched portfolio benchmarking, this portfolio would be considered to have done as well as physically possible. However, there is still a large sum of 'lost' returns which could have been claimed had the weights been better allocated. Being invested in the best possible assets does not mean one has created the best possible portfolio. Matching portfolio benchmark creations seem to fail to represent the whole of return opportunity.

3 Regression-Based Performance Attribution

Lu and Kane (2013) make use of a linear regression approach to disentangle the extent to which a portfolio's excess return is attributable to allocation effects. This approach conceptualizes return as a linear function of various factors and their interactions. The extent to which a portfolio's assets' returns differ from their linear expectation is considered to be its excess return.

Though this approach sensibly looks to control for universe-wide return covariates, it is likely to put too much confidence in linear regression. A standard asset universe fails many of the assumptions of linear regression. First, a linear regression assumes a linear relationship. There is no guarantee of this relationship between a characteristic and returns (mid-sized company assets falter relative to large and small). Second, all observations are assumed to be independent of each other. Unless this has been prevented, many asset universes are likely to contain several securities from the same company, obscuring true effects. Finally, a linear regression ought to

avoid extreme values which could suggest a relationship where there is none. In an asset universe, the boom or bust of a security on the edge of a characteristic distribution may harmfully obscure true relationships.

Given that this approach conceptualizes excess return as residual from the linear fit, obscuring factors such as those described above are likely to significantly effect return estimates. Were a portfolio's assets evenly distributed throughout all characteristics, this approach might serve as an accurate metric for performance. However, this assumption is unlikely to be true for most portfolios, and its necessity would likely render the method's usefulness null.

4 Characteristic-Matched Random Portfolios

Burns (2004) discusses vital issues in the widespread and widely accepted method of manager performance attribution using traditional benchmarks. His argument centers upon the volatility of individual benchmarks which are susceptible, even when well diversified, to the random swings of its most heavily weighted assets. Instead, he argues for the use of constrained random portfolios in measuring manger performance. A collection of random portfolios, when held under the same constraints as the portfolio manager, allows for an unbiased representation of a manager's true stock-picking ability. These portfolios are selected with a neutral perspective to any measure of utility and are therefore serve as a much more meaningful representation of expected performance. Further, constrained random portfolios allow greater ability (a) for investor and fund manager alike to set flexible and individually preferred investment mandates and (b) to establish robust measures of alpha.

We suggest a new method of benchmark creation, characteristic-matched random portfolios, which lends heavily from Burns's work and integrates constraints for passively allocated characteristics. Our method samples the space of all portfolios which exactly match the target portfolio's characteristics. As a result, this approach is able to directly explore a complete counterfactual set of investments, controlling for the effects of matched characteristics.

This method's primary advantage, somewhat ironically given its computational demands, lies in its simplicity and interpretability. This method is not reliant on any of the assumptions of linear regression. Nor is it prone to the biases incurred by the unit-level matching of MPs. Instead, it is centered around the minimization of allocation biases. In essence, the establishment of the characteristic-matching space answers the question of "what returns could this manager have achieved given their passive exposures?"

The space of characteristic matches is defined by the intersection of $Ax = b$ and the N-simplex, where A represents the universe of assets, x represents a vector of weights, b represents a vector of target portfolio characteristics, and the N-simplex

bounds the weights to a sum of 1. In order to sample from this space, we have created the **portfoliowalkr** package, which serves as an equity portfolio-centric wrapper for the **walkr** function. The **walkr** function institutes two Monte-Carlo Markov Chain (MCMC) algorithms, Dikin walk and hit-and-run. For additional information on these algorithms and their limitations, see Yao and Kane (2017).

4 Data

We demonstrate the use of the **portfoliowalkr** package using data from the **pa** package (Lu and Kane (2013)). This data comes from MSCI Barra's Global Equity Model II (GEM2). The modified data pulled from **pa** contains three modified datasets (year, quarter, jan), containing 3,000 securities each. Because **portfoliowalkr** is currently limited to single-period analyses, the single-period jan dataset is of most use.

```
data(jan)
names(jan)
```

```
##  [1] "barrid"    "name"      "return"    "date"      "sector"
##  [6] "momentum"  "value"     "size"      "growth"    "cap.usd"
## [11] "yield"     "country"   "currency"  "portfolio" "benchmark"
```

Per **pa**, the characteristics contained in the dataset are as follows:

- barrid: security identifier by Barra.
- name: name of a security.
- return: monthly total return in trading currency.
- date: the starting date of the month to which the data belong.
- sector: consolidated sector categories based on the GICS.[1]
- momentum: capture sustained relative performance.
- value: capture the extent to which a stock is priced inexpensively in the market.
- size: differentiate between large and small cap companies.
- growth: capture stock's growth prospects.
- cap.usd: capitalization in model base currency USD.

[1] Global Industry Classification Standard

- yield: dividend of a security.
- country: the country in which the company is traded.
- currency: currency of exposure.
- portfolio: top 200 securities based on `value` scores in January are selected as portfolio holdings and are held through December 2010. This is to avoid the complexity of trading in the analyses.
- benchmark: top 1000 securities based on size each month. The benchmark is cap-weighted.

A sample of the dataset is presented below with modified weights and selected columns. We will use this sampled dataset for a micro demonstration of **portfoliowalkr** functionality.

```
sample
```

```
##                                    name   return       date      sector
## 1  DAIKIN INDUSTRIES                     -0.08333 2010-01-01 Industrials
## 2  MARK HOLDINGS INC                      0.00000 2010-01-01   Materials
## 3  PGNIG                                  0.00528 2010-01-01      Energy
## 4  CHECK POINT SOFTWARE                   0.16739 2010-01-01    InfoTech
## 5  PHIL.LONG DIST.TEL.-PR                 0.00000 2010-01-01    TeleSvcs
## 6  DELL INC                               0.00000 2010-01-01    InfoTech
## 7  ACTIVISION BLIZZARD INC               -0.08551 2010-01-01    InfoTech
## 8  AL JADEED TXTL (PKR10)                -0.12970 2010-01-01   ConDiscre
## 9  QUALCOMM INC                          -0.15280 2010-01-01    InfoTech
## 10 ROYAL DUTCH SHELL PLC [A] - ADR       -0.07852 2010-01-01      Energy
##    momentum   size growth country portfolio
## 1    -0.008  0.205  1.847     JPN      0.25
## 2    -0.830  0.000 -0.124     USA      0.00
## 3    -0.930  0.326 -0.228     POL      0.25
## 4     0.706  0.225 -0.353     ISR      0.00
## 5    -0.547  1.335 -0.337     PHL      0.25
## 6    -0.626  0.196 -0.056     USA      0.00
## 7    -0.866 -0.270  0.850     USA      0.00
## 8    -2.968  0.000 -0.109     PAK      0.00
## 9    -0.352  0.860  0.788     USA      0.25
## 10   -0.008  0.931 -0.483     GBR      0.00
```

To limit complexity, we will use this sample frame in our analysis. In this case, we have assigned 25% weighting to four securities, `DAIKIN INDUSTRIES`, `PGNIG`, `PHIL.LONG DIST.TEL.-PR`, and `QUALCOMM INC`. We would like to see how well this

portfolio performs within the space of all possible portfolios which share its exposures. We load the **portfoliowalkr** library and sample 10,000 benchmarks which match our target portfolio's growth, size, and momentum. We will use the Dikin walk method (described in Yao and Kane (2017)) with 10 chains to analyze convergence.

```
library(portfoliowalkr)

walked <- portfoliowalkr(sample,
                         match = c('growth', 'size', 'momentum'),
                         portfolio.weight = "portfolio",
                         ret.var = "return",
                         points = 10000,
                         chains = 10,
                         method = 'dikin')
```

This function returns 5 outputs in a list. They are:

- $plot: A histogram showing the distribution of returns of characteristic-matched random portfolios. Visualizes the relative performance of the target portfolio and an evenly weighted baseline.
- $summary: A list of summary statistics describing returns of the target and an evenly weighted baseline relative to the space of characteristic-matches.
- $frame: The initial 'universe' frame input to the function, with characteristic-matching random portfolio weights attached. length(universe) + points columns in width.
- $returns: A dataframe containing the returns of each portfolio, including an evenly weighted baseline and the target.
- $explore: The list of chains output from the walkr function. Primarily useful for exploring MCMC random walk convergence

First, we can examine our base portfolio's exposures and randomly sample from our benchmarks to ensure that the characteristic matching has worked.

```
check_exposures(walked$frame,
                match = c('growth', 'size', 'momentum'),
                portfolio.weight = "portfolio",
                n = 4)
```

```
##          portfolio weight2257 weight5606 weight5915 weight8315
## growth     0.51750    0.51750    0.51750    0.51750    0.51750
## size       0.68150    0.68150    0.68150    0.68150    0.68150
```

```
## momentum  -0.45925   -0.45925   -0.45925   -0.45925   -0.45925
```

As desired, all portfolios match the characteristic exposures of the target portfolio.

We make use of the `explore_walkr` function from **walkr** to analyze the convergence of the MCMC random walk. This function initializes a `shiny` interface from **shinystan** which contains metrics and visualizations to describe the walk's chain convergence. We run the function on the $explore output, which contains a list of the sampled chains.

```
explore_walkr(walked$explore)
```

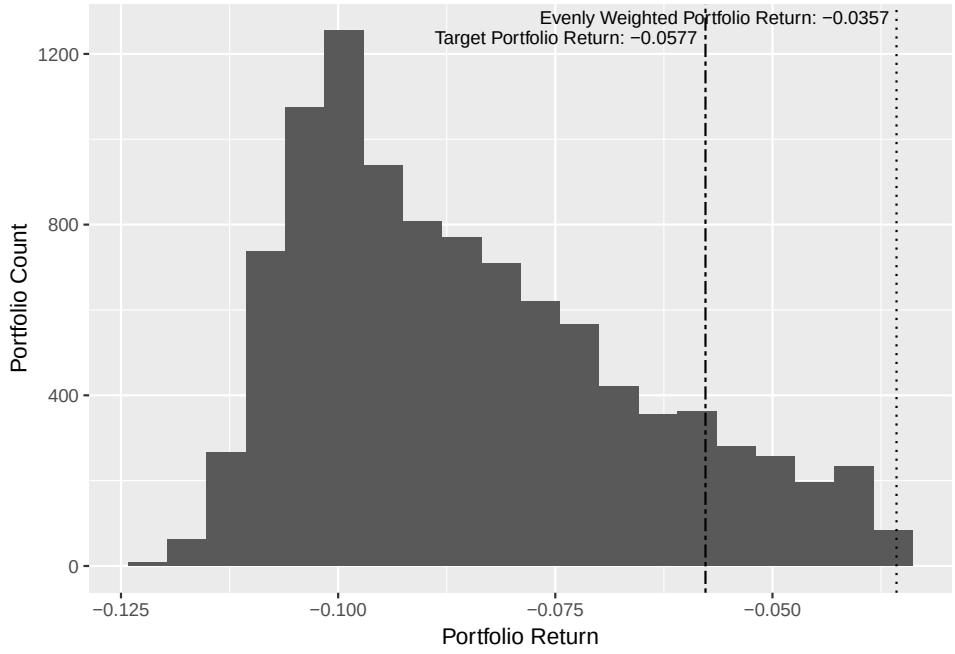
Because all ^R parameters are less than 1.1 (walkr would otherwise throw a warning), we can be confident that we have representatively sampled the space.

We can now assess the performance of our portfolio relative to the space of characteristic matches.

```
walked$summary
```

```
## $Target
##          Return Percentile Rank
##      -0.0577125       0.8865000
##
## $Universe
##          Return Percentile Rank
##       -0.035719        0.999100
##
## $Quantiles
##          0%         25%         50%         75%         100%
## -0.12076757 -0.10036988 -0.08889656 -0.07253724 -0.03498837
```

```
walked$plot
```



The target portfolio has underperformed an evenly weighted portfolio by -2.20%. Given this sample baseline, one might conclude that the manager has done poorly. But the portfolio has apparently been disadvantaged by its characteristics. 99.9% of portfolios with these exposures underperform the evenly weighted benchmark.

In fact, controlling for the characteristics of the portfolio, our manager has performed their job as "stock-picker" well. Among the set of portfolios with these characteristics, the target portfolio performed better than 88.6%. The target portfolio outperformed the median random portfolio by 3.12%.

The portfolio analyzed herein was inherently disadvantaged relative to the universe by its passive characteristics. The manager does not claim to be advantaged in their ability to predict the performance of securities based on size, growth, or momentum. The manager claims to be talented in achieving excess returns beyond the macro movements of the market. Therefore, they should be assessed against the counterfactual space of benchmarks with their portfolio's exact exposures. The figures above demonstrate the necessity for dynamic benchmarking of manager performance. We believe that characteristic-matched random portfolios provide the most robust benchmark available.

Conclusion

The **portfoliowalkr** package makes use of two Monte-Carlo Markov Chain (MCMC) algorithms, Dikin walk and hit-and-run, to sample from the space of portfolios which bear the same exposures as a target. In doing so, this package provides the tools to control for allocation biases and assess manager skill.

The package is currently limited to single-period equity portfolio sampling across all available assets. Future improvements could be made by creating tools for multiple time periods or allowing users to specify a maximum number of assets in the benchmark. These improvements are likely to improve the applicability of the package to real-world performance assessment.

Bibliography

Austin, Peter C, Nathaniel Jembere, and Maria Chiu. 2018. "Propensity Score Matching and Complex Surveys." *Statistical Methods in Medical Research* 27 (4): 1240–57. doi:10.1177/0962280216658920.

Bartz, Kevin, and Dave Kane. 2010. "Matching Portfolios." http://dx.doi.org/10.2139/ssrn.1135361.

Burns, Patrick. 2004. "Performance Measurement via Random Portfolios." http://www.burns-stat.com/pages/Working/perfmeasrandport.pdf.

Lu, Yang, and David Kane. 2013. "Performance Attribution for Equity Portfolios." In.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70 (1): 41–55. doi:10.1093/biomet/70.1.41.

Rubin, Donald B. 1973. "Matching to Remove Bias in Observational Studies." *Biometrics* 29 (1). [Wiley, International Biometric Society]: 159–83. http://www.jstor.org/stable/2529684.

Yao, Andy, and David Kane. 2017. "Walkr: MCMC Sampling from Non-Negative Convex Polytopes." *The Journal of Open Source Software* 2 (March). doi:10.21105/joss.00061.