Matching Portfolios

Kevin Bartz
Harvard University
Statistics Department
bartz@fas.harvard.edu

Dave Kane
Harvard University
IQSS Fellow
dkane@iq.harvard.edu

August 7, 2013

Abstract

We propose a portfolio performance measure based on comparison with a *matched portfolio* sharing the same exposures but holding different stocks. Under the framework of the Rubin Causal Model, the matched portfolio provides a view of the *counterfactual*, an alternative portfolio that the manager could have chosen but did not. We provide a methodology for constructing a matched portfolio using the holdings of the original portfolio and a set of stock characteristics. Treated as a benchmark, a matched portfolio provides a more precise estimate of alpha, or stock-picking ability, because it shares the same characteristics as the original portfolio. It is also more flexible than other benchmarks because it can be matched on any number of characteristics, and can mimic the weighting structure (e.g., long-short, 130/30) of the original portfolio.[1]

1 Introduction

All performance measurement techniques set up a counterfactual world in which the portfolio invested in similar — but not identical — securities. The counterfactual represents something an investor could have done, usually simply, without the need for an active manager. The portfolio is said to be successful if it outperforms the counterfactual portfolio, commonly called a benchmark.

In practice, the most common benchmark is a simple, well-known passive portfolio like the S&P 500. Here the counterfactual is straightforward: the money could have been invested in the S&P 500 instead of the portfolio, and the benchmark shows how that alternative strategy would have performed.

The problem with a simple benchmark like the S&P 500 is that it can easily give an inaccurate view of performance. First, the manager can easily lie about his benchmark, picking one after the fact to make his performance look as good as possible. Second, the manager may never set a benchmark in the first place.

[1] Comments welcome. An R package, which reproduces the results of this paper, is available at `www.kevinbartz.com/matching`.

Third, he may specify a poor benchmark, one that does not reflect his strategy. Fourth, there may be no clear benchmark to specify, as with a long-short equity portfolio.

One approach is to focus on stock characteristics, as in Daniel et al. (1997) and Daniel and Titman (1997). Characteristics like size, country and sector can drive a portfolio's return. Only by accounting for these factors can a benchmark measure the manager's stock-picking ability. In a characteristic-based approach, the benchmark mimics the original portfolio's exposures to a set of characteristics. This corresponds to a counterfactual in which the manager chooses a different set of stocks with the same characteristics. Comparing this benchmark's performance to the original portfolio's yields an estimate of stock-picking ability that is free from bias due to characteristics.

Recent comparisons suggest that the characteristic-based benchmarks perform well. Kothari and Warner (2001) compare several performance measurements and find that a characteristic-based measure performs the best. Their measure is based on classifying the portfolio's components into a 5 x 5 x 5 matrix, with a benchmark for each cell. "alpha" is the weighted average difference of the portfolio returns and the per-cell benchmarks.

There are also problems with Daniel et al. (1997)'s characteristic-based benchmarks. The method requires that there be stocks in every cell of the characteristic matrix. This may be impossible in situations with more characteristics than three, or to situations with more levels per characteristic than five. The method does not scale well because of the inevitably limited number of stocks in the universe. Additionally, it is unclear how the approach applies to portfolios with short positions.

Our approach applies the statistical matching literature (Wilks (1932), Cochran and Rubin (1973), Rubin (1973a), Rubin (1973b)), which examines techniques for matching one subject to another based on a vector of covariates. This is exactly the problem we face: for each security in the portfolio, we seek a match, a security not in the portfolio with a similar vector of stock-specific characteristics. Using the causal model of Rubin (1974), we reframe the problem of identifying a manager's alpha to that of identifying causal effects. Our estimate of alpha is the difference between the portfolio's return and the return of a characteristic-matched benchmark.

Our paper introduces a more general framework for constructing characteristic-matched portfolios, providing more precision and flexibility. It is more precise because We allow for an arbitrary number of characteristics and levels per characteristic. It is more flexible because we allow for alternative portfolio weightings, such as long-short, 130/30 and so on.

2 Background

Consider an investor with $100 million to invest. She selects a manager. That manager could invest in stocks, bonds, real estate, trading cards, Beanie Babies, etc. Indeed, he can invest the money that the investor has given him in an almost

an infinite number of items. Do we need data on all of these items and their returns in order to evaluate the manager's performance? The answer lies in the restrictions the investor places on the manager. In the extreme, she could allow the manager to invest in anything, but no institutional investor behaves that way. Instead, in consultation with the manager, she creates a *universe* of permitted investments.

A typical first requirement is that money be placed in exchange-traded assets. But even within this, there is a diverse set of possibilities. Few portfolio managers have the freedom to invest in both complicated derivatives and thinly traded emerging market equities. Additionally, the manager may not be aware of all exchange-traded assets, or he may not have data on all of them. Assets without data are not possibilities.

Finally, the manager may specialize in only a certain segment of what remains. He may be comfortable investing in only American large-cap stocks. The universe is defined before any investments are made. Bodie et al. (2001) define a *comparison universe* as a group of securities with similar risk characteristics. This forms the set of securities from which we will construct counterfactual portfolios for performance evaluation. As such, it is important that the universe be neither too broad nor too narrow. If it is too broad, the counterfactual portfolios will contain securities never under active consideration by the manager. If it is too narrow, then some actively considered alternative investments will never be included in the portfolio.

2.1 Benchmarks

A benchmark can be viewed as a counterfactual portfolio: a set of securities that the manager *could have* invested in, but chose not to. These securities must come from the universe defined above.

A common benchmark is a passive portfolio like the S&P 500. The first problem with this is the makeup of the S&P 500, which is often different from the universe of investments the manager actually considered. Even if the S&P 500 is contained in the universe, the stocks in the S&P may be very different from the portfolio: They may have larger capitalization or come from different countries and sectors. Relative to the S&P 500, outperformance or underperformance could then be explained in terms of sector and cap assignment, as opposed to stock-picking ability. A benchmark that is matched by sector, cap and country would provide a better indicator of the manager's stock-picking ability.

The appropriate benchmark depends on the source of the manager's claimed advantage. The benchmark should be similar to the portfolio in every characteristic *except* those the manager claims as his advantage. It can be possibly dissimilar in characteristics the manager does not claim as his advantage. In this way, the benchmark should have the same constraints the manager had, but without the advantage of the manager's ability. This framework isolates and tests the manager's claim.

In the most common case in which stock-picking ability is the presumed advantage, the benchmark should be as closely matched to the original portfolio

as possible on any observable characteristics. Here, the manager's claim is that stocks just like his will do worse than his stocks. Comparison with an identically characterized benchmark isolates the manager's stock-picking ability. It removes competing explanations for the manager's excess return. Constructing a matched portfolio for use as a performance benchmark is the subject of this paper.

Some managers make conscious bets on sector or some other characteristic. Here, a benchmark matched on the same characteristic will not reveal their puported timing ability. The benchmark must *not* be matched on the characteristic of presumed advantage. However, it is still desirable that the benchmark be matched on all other characteristics. For instance, if a claimed sector-timer bets on large-cap energy stocks, a benchmark which matches on capitalization would help isolate his ability to time sectors.

2.2 Rubin Causal Model

We can view the manager's claimed ability as a causal treatment effect; the manager believes his decisions will cause excess returns. Viewed like this, performance evaluation falls neatly into the Rubin Causal Model, a statistical framework for causal inference. Rubin (1973a) develops matching techniques in which treated units are compared with near-identical untreated units to measure the causal effect of a treatment.

To understand the Rubin Causal Model, consider an experimental setting in which a researcher desires to estimate a drug's ability to cure headaches. Ideally, he would run a randomized experiment, randomly assigning subjects with headaches to drug and placebo (or control) treatments. This makes the treated and control populations identical in every other way, so that any difference in headache incidence after treatment is a result of the drug.

But often, a randomized experiment is prohibitively expensive or unethical. Then the researcher must resort to an observational study, in which treated and control units are found wherever possible (e.g., from patient beds in a hospital). In this setting, treatment assignment is non-random. Comparing the headache rate in the treated and untreated population is no longer proof of causality because confounding factors could be at work. For instance, perhaps males are more likely both to take the drug and to have headaches go away naturally. This gender effect could be mistaken as the drug's treatment effect.

Rubin (1973a), Rubin (1973b), Rubin (1974) and Rosenbaum and Rubin (1983) develop a series of matching techniques to counteract bias in an observational study. For every treated unit, matching identifies a near-identical control unit, based on every observable covariate (e.g., race, gender, income, weight, etc.). The comparison in headache rates is made across matched units, where the treatment effect is measured as the difference in response between the treated and matched control units. Intuitively, the goal is to make the observational study look as much like a randomized experiment as possible.

Venture capital provides a useful financial analogy. The venture capitalist spreads his money around a variety of small companies. The infusion of capital

is a treatment, and the resulting growth (or decline) in the companies is a treatment effect. The venture capitalist cannot observe the *potential outcomes*: how unchosen companies would have performed had they received capital, or how chosen companies would have performed had he not invested in them. He can instead use a matching method to determine the causal effect of his capital. He can compare two similar companies, one that received venture capital and one that did not. Their difference in outcomes constitutes an estimate of the treatment effect.

Identifying stock-picking skill presents a similar challenge. The inclusion of a stock in a portfolio can be considered a treatment. The treatment effect is the excess return experienced by a stock in the portfolio, all characteristics being equal. Matching the portfolio holdings to similar non-holdings provides a precise estimate of the treatment effect.

There is also an important difference between this problem and the Rubin Causal Model framework. In an observational study, potential outcomes are not observed. In the portfolio problem, potential outcomes are observed because returns are available for all stocks, even those outside the portfolio. We can easily determine the return a manager would have enjoyed on a matched portfolio (ignoring the price impact of his own trading), even though the investor did not actually commit her money to it. We use this difference to our advantage in our performance evaluation, computing the potential returns on a collection of matched portfolios.

One assumption of the Rubin Causal Model is the *stable unit treatment value assumption*, which holds that treatment assignment has no effect on potential outcome. For our evaluation this means we must assume that investment choices have no direct effect on stock returns. For most smaller managers, this is a reasonable assumption, but the trading of large institutional managers often has nontrivial effects on stock prices.

2.3 Characteristics

The authors whose work most closely resembles ours are Daniel et al. (1997). Their measure is based on constructing a characteristic-based benchmark with the same exposures to three characteristics: size, book-to-market ratio and prior-year return. Dividing these into a 5 x 5 x 5 matrix, they compute the portfolio's total weight in each of the 125 cells. They then construct a style-matched portfolio as an equivalently weighted average of 125 benchmarks correspondings to the cells. "alpha" is defined as the difference in returns between this and the original portfolio.

A key disadvantage is the method's coarseness. For example, if we had five characteristics, each with ten levels, we would have 100,000 individual cells in our matrix. No universe of stocks is large enough to provide even a single match for each cell in such a large matrix. Dealing with this curse of dimensionality is one of the primary reasons that our method is more precise than approaches like Daniel et al. (1997) and Ferson and Qian (2004).

2.3.1 Holdings

Throughout, we assume the existence of holdings information for the portfolio at the time under consideration. Our methodology is not applicable if only returns are available, as assumed by the models of Fama and French (1992) and Jensen (1968).

Using holdings information has several advantages. First, regression-based models typically require a long period of time to accumulate enough returns that the benchmark has predictive value. Oftentimes a divergence between the regressors occurs only in rare market environments. By using holdings directly, a meaningful benchmark can be found with only a single period of returns. This is important in practice when analyzing the value of signals without enough history to provide a well-fitting multi-factor regression. Other authors who have used holdings include Grinblatt and Titman (1989), Grinblatt and Titman (1993) and Kosowski et al. (2006).

Additionally, Daniel et al. (1997) points out that time-variable fees, trading costs and turnovers can contaminate a model that uses only the portfolio-level returns. Evaluating holdings insolates performance measurement from fees, which can be considered separately.

2.4 Random Portfolios

A simple approach to creating a counterfactual portfolio is to "throw darts" at the universe, picking random stocks for a portfolio. This was suggested as long ago Lorie (1965). Since then, several variants have been proposed, such as the "Average Investment Performance Index" of Cohen and Fitch (1966), which simply assigns random positive weights to every stock in the universe, rescales them to add to 1, and treats the result as a single draw of a random portfolio. Repeating this many times, they measure the performance of mutual funds against the average of the "random investments." Burns (2004) and Dawson and Young (2003) further refine the idea by using the same universe of stocks that the manager likely used.

Matching portfolios represent a concrete counterfactual, but do not share the characteristics of the original portfolio. Instead, the portfolios have the same exposures on average as the universe. Their average return is the same as that of an equal-weighted benchmark spanning the universe.

3 StarMine

As a concrete example, we will focus on a portfolio for a single month using the ratings provided by StarMine, a San Fransisco-based research company that creates quantitative equity models for stock selection.

The ratings come from January 31, 1995, and were made available before the market opened on February 1. The company had data on 4,476 stocks from a global universe. Of these, they produced stock ratings for 2,791 for which they had the greatest expertise. They did not rate the other 1,685.

StarMine's indicator, "SMI," is a rating from 1 to 100, with 100 their predicted best performers. The ratings are intended to be percentiles and are roughly uniformly distributed. Table 1 shows a few sample records. The data include the **country** in which the company is headquaratered, the **sector** in which a majority of the company's business takes place, the **market capitalization** and the **one-month forward return**. A summary of the distribution of these attributes in the data appears in Figure 1.

	Coun.	Sector	Cap. (M)	SMI	1-mo. Ret.
Pacific Metals Co Ltd	JPN	Manuf	868.0	NA	-0.108
Jcr Pharmaceuticals Co Ltd	JPN	Hlth	558.0	NA	-0.155
Titan International Inc	USA	Manuf	195.0	NA	0.014
Brite Voice Systems Inc	USA	HiTec	123.0	95	0.234
Pharmaceutical Marketing Svc	USA	HiTec	122.0	23	-0.092
Sainsbury (j) Plc	GBR	Shops	12171.0	NA	-0.035

Table 1: Sample records from the `starmine` data.

The object of our analysis is a simple long-only portfolio using the SMI score, splitting our position equally among the 100 top-rated stocks in the universe, representing roughly the top decile of StarMine's ratings. Figure 2 shows this portfolio's primary exposures. The portfolio returns 2.5% over the following month.

Figure 2 compares this portfolio's exposures to those of the universe. This portfolio is relatively heavily weighted in high-tech., shopping and durable goods stocks, as well as mid-caps and Australian stocks.

Figure 3 shows the performance of the portfolio next to that of various benchmarks. Compared to the S&P, the portfolio looks bad; compared to the equal-weighted complete universe, it looks great. The equal-weighted return on the rated universe also underperforms the actual portfolio.

4 Matching Portfolio

In this section we give steps to build a "matching portfolio" that matches the exposures of the StarMine portfolio, but holds different stocks. We start by examining the portfolio's holdings at the time under consideration. Associated with each stock is the set of characteristics listed above: country, sector and market cap. Note that our method can incorporate any number of characteristics.

We take an approach similar to that of Daniel et al. (1997) in that it measures performance against a benchmark with the same characteristics. Ours is more finely matched, however, pairing each holding with a non-holding with the same characteristics. Our framework is also more flexible, allowing for any number of characteristics.

In the sections below, we first illustrate the goal of matching and then provide computational details. To make matches we utilize the propensity score of

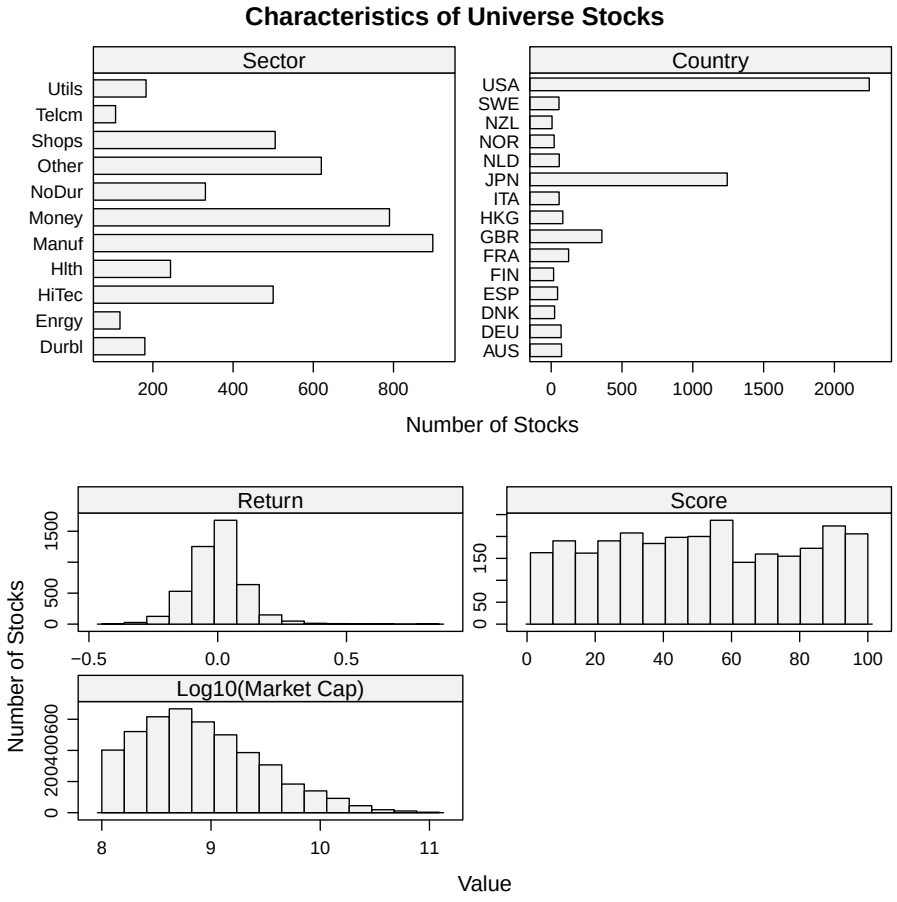


Figure 1: Histograms representing the makeup of the universe of stocks by country, sector, return, StarMine score and market capitalization in U.S. dollars. The complete universe of stocks is represented here, including both StarMine-rated stocks.

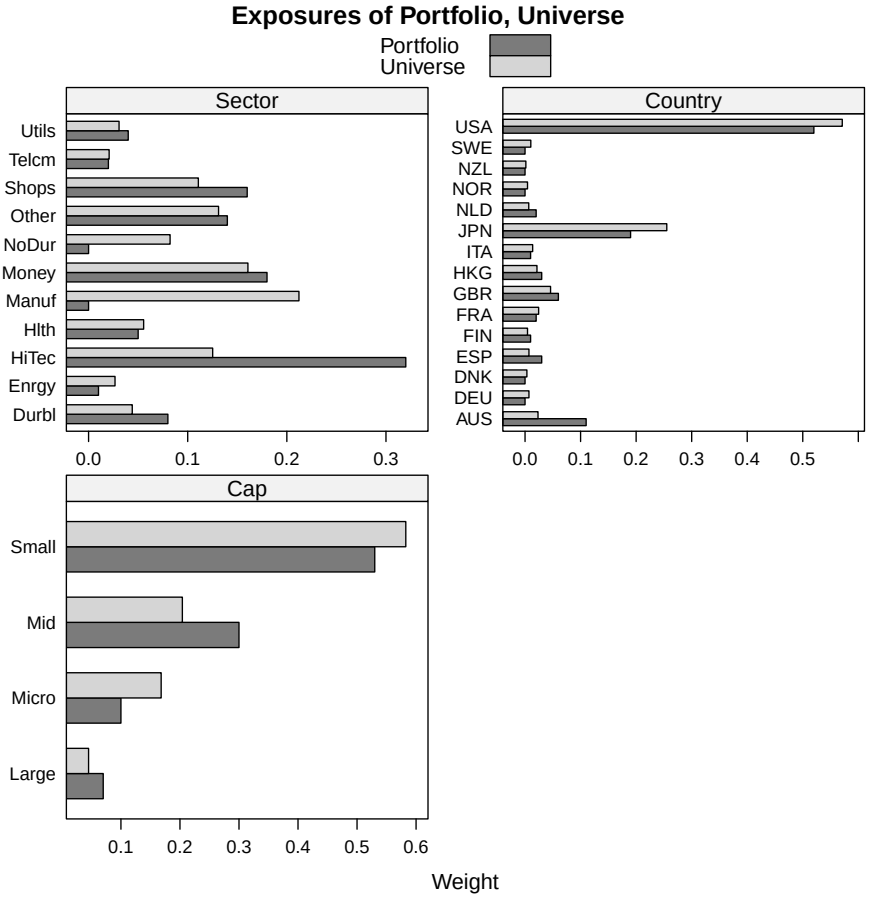


Figure 2: Exposures of the StarMine portfolio to sector, country and market cap. Also shown are exposures of an equal-weighted index of stocks in the rated universe.

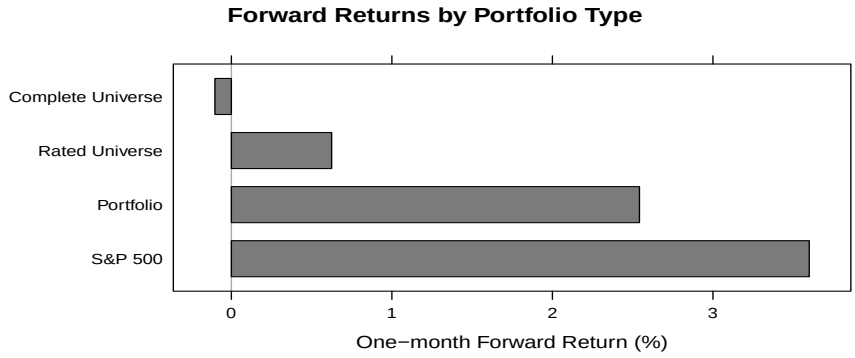


Figure 3: Performance of the StarMine portfolio compared to performance of various benchmarks: the S&P 500, the equal-weighted complete universe, and the equal-weighted rated universe.

Rosenbaum and Rubin (1983), who used it to match treated subjects to near-equivalent control subjects in an observational study. Finally we demonstrate how matched portfolios can be used as benchmarks, using the StarMine portfolio as an example.

4.1 Motivation

The simplest way to create matching portfolios is to examine the universe of securities in the multi-dimensional characteristic space. Figure 4 shows a set of two-dimensional cross-sections by country, each showing securities by sector and market cap. The black dots represent holdings and the gray dots non-holdings.

For most portfolio positions, there's an abundance of very close possible matches. A few holdings might have to reach a bit farther for a match, such as the $70 billion durable goods holding (Toyota Motor) on the bottom row of the Japan panel. The nearest match with the same sector and country is Nissan Motor ($18 billion). Even if we allow non-Japanese matches, the closest we can get is General Motors ($28 billion). Thankfully, either of these is a plausible match, and most holdings have many more possible matches.

4.2 Propensity Score

It is clear from Figure 4 how one would execute matching as a visual exercise. We next discuss how to approach matching computationally.

Matching in an arbitrarily dimensional stock characteristic space is made challenging by the curse of dimensionality. Looking at Figure 4, one's first intuition is to find a "nearby" stock for each holding, under some notion of distance. Rubin (1973a) and Rubin (1973b) investigate a variety of matching

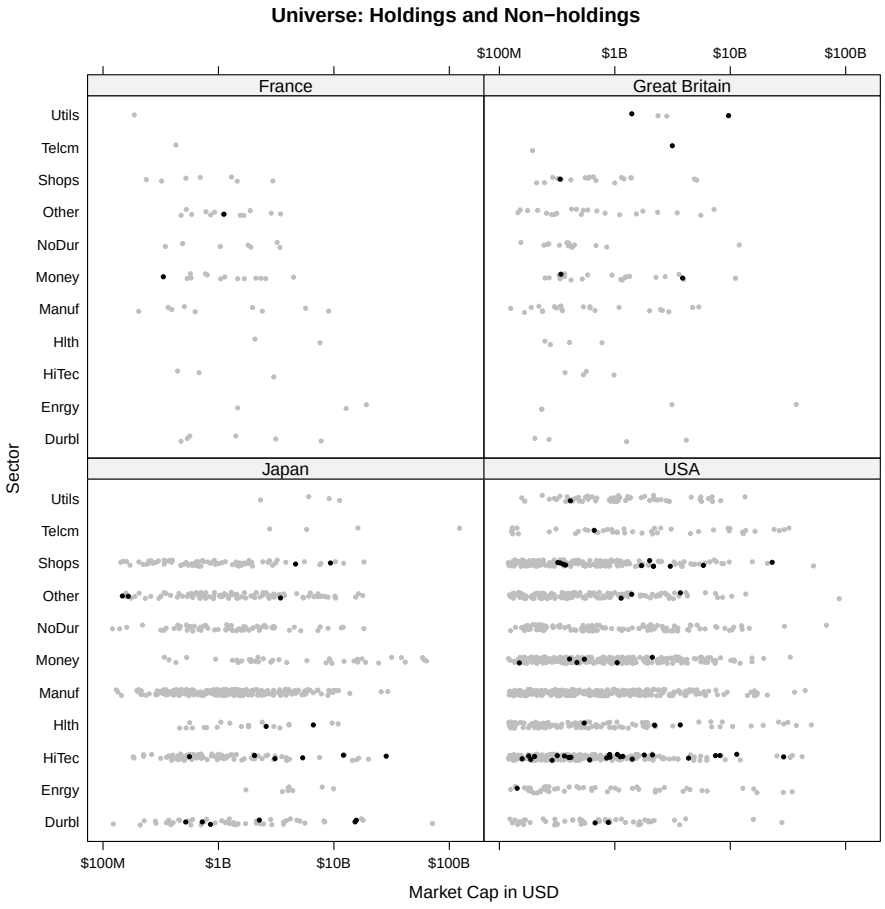


Figure 4: Cross-sections of universe stocks by market cap and sector. There is one panel for each of the four countries, the U.S., Japan, Great Britain and Australia, with the most universe securities. Black dots are portfolio holdings and grey dots non-holdings. The vertical positions of the dots are jittered slightly so as to appear separate. Most securities have a plethora of possible matches, but some do not, like Toyota Motor Corp. represented by the dot in the lower right corner in the Japan panel. It would have to stretch all the way to Nissan Motor, the nearest gray dot, for a suitable match.

approaches based on distance metrics, but find that these make increasingly poor matches as the number of dimensions grows.

Rosenbaum and Rubin (1983) solve the problem with the *propensity score*. The propensity score is the estimated probability that a given security appears in the portfolio, given only its characteristics. The *fundamental balancing property of the propensity score* shows that matching two units based on the propensity score is sufficient, on average, for matching them on all their characteristics.

Mathematically, define $I_i = 1$ when stock i is in the original portfolio and $I_i = 0$ otherwise, with $i \in \{1, ..., n\}$ for all n stocks in the universe. Let $X_i \in \mathbb{R}^p$ represent the p-dimensional vector of stock characteristics associated with stock i. The propensity score is defined as $P(I_i = 1|X_i)$. Typically this probability is estimated using a logistic regression of the binary outcome vector I on the characteristics X as the covariate matrix, as described in Rubin and Thomas (1996).

Rosenbaum and Rubin (1983) show that the propensity score is a "balancing score" by the definition of Dawid (1979): a single-dimensional value that acts to "balance" the values of the entire vector X_i. Statistically, this means $I_i \perp X_i|b(X_i)$; i.e., there is independence between I_i and X_i given the balancing score $b(X_i)$. In words, this means that if we match stocks to each other using only the estimated probabilities $b(X_i) = \hat{P}(I_i = 1|X_i)$, then we have effectively matched them on the basis of all the characteristics. A multidimensional problem is thus reduced to one dimension.

Estimating a propensity score model for the StarMine portfolio, we obtain the fitted coefficients in Table 2. The sector and country variables are swept out as a set of 1/0 indicators with an excluded level for each: "AUS" for the country indicators and "Durbl" for the sectors. Indicators corresponding to countries unrepresented in the portfolio are mathematically $-\inf$ and are excluded from the table.

The model's positive coefficients correspond to characteristics to which the portfolio is over-exposed relative to the universe. All of the country indicators are negative, showing exposure to the excluded level, Australia. Likewise, the positive high-tech coefficient indicates exposure to that sector.

4.3 Computation

Next we use the propensity score to match holdings to nearby non-holdings with similar propensity scores. There are several ways of finding matches, and Abadie and Imbens (2006) provides a nice review of the possibilities. The simplest, proposed by Cochran and Rubin (1973), is *nearest-available matching*, sometimes called "greedy matching."

For each portfolio holding $i \in \{1, ..., n\}$,

1. Find the security j that minimizes the absolute difference in propensity scores $|b(X_i) - b(X_j)|$. Add it to the matched portfolio with the same weight as holding i.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.6447	0.4969	-7.33	0.0000
1(country = USA)	0.3221	0.2611	1.23	0.2173
1(country = France)	0.2401	0.6905	0.35	0.7281
1(country = Great Britain)	1.0172	0.4513	2.25	0.0243
1(sector = Enrgy)	-1.8188	0.9656	-1.88	0.0597
1(sector = HiTec)	0.3614	0.3782	0.96	0.3393
1(sector = Hlth)	-0.7223	0.5312	-1.36	0.1740
1(sector = Money)	-1.2159	0.4559	-2.67	0.0077
1(sector = Other)	-1.1677	0.4779	-2.44	0.0146
1(sector = Shops)	-0.3180	0.4169	-0.76	0.4457
1(sector = Telcm)	-1.0039	0.7379	-1.36	0.1738
1(sector = Utils)	-0.8080	0.6358	-1.27	0.2039
1(cap = Small)	0.6089	0.3576	1.70	0.0887
1(cap = Mid)	1.4591	0.3834	3.81	0.0001
1(cap = Large)	1.5832	0.4951	3.20	0.0014

Table 2: Coefficients from the fitted logistic regression for propensity score using indicator variables from the StarMine portfolio.

2. Remove security j from the universe of possible matches. This is to ensure that matched holdings are selected *without replacement*. This ensures that no match is given double weight in the matched portfolio.

Greedy matching is imperfect, and one could benefit from rearranging matches optimally. The gains from this would be small in our case because the size of our universe relative to the portfolio ensures a large reservoir of possible matches.

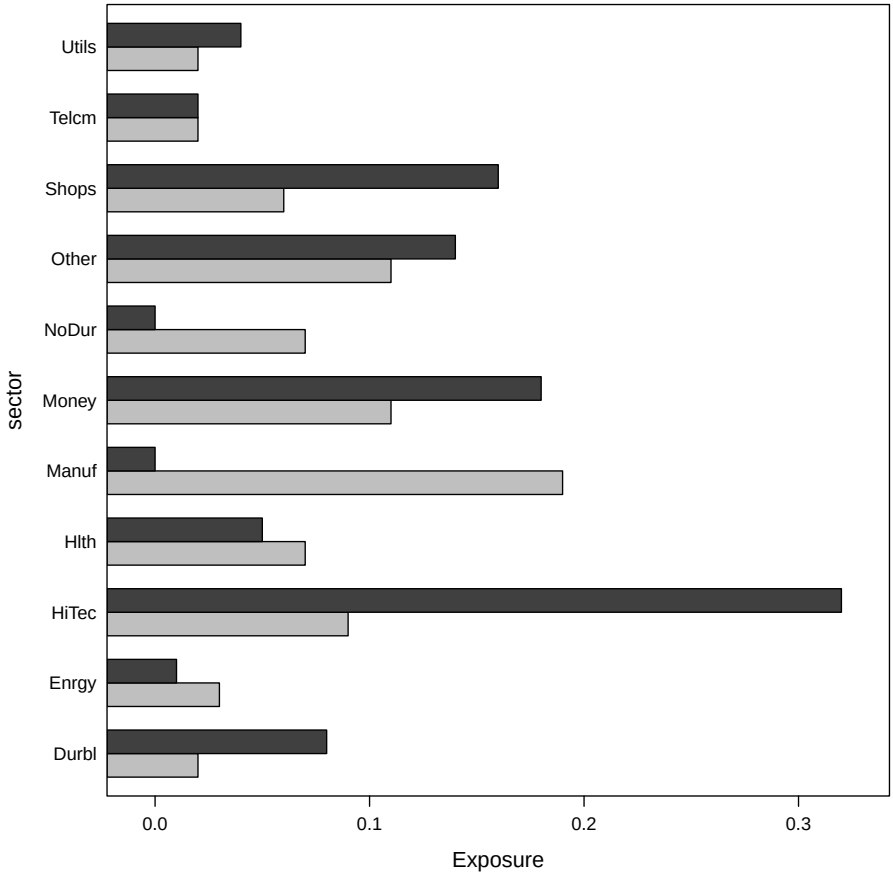
The propensity score's balancing behavior is a distributional property. Balance is realized asymptotically, not at the level of the individual match. While individual matches may not appear matched, the entire matched portfolio's exposures usually come close to matching the original's. For this reason we do not report any specific stocks' matches, but rather focus on the exposures of the entire matched portfolio.

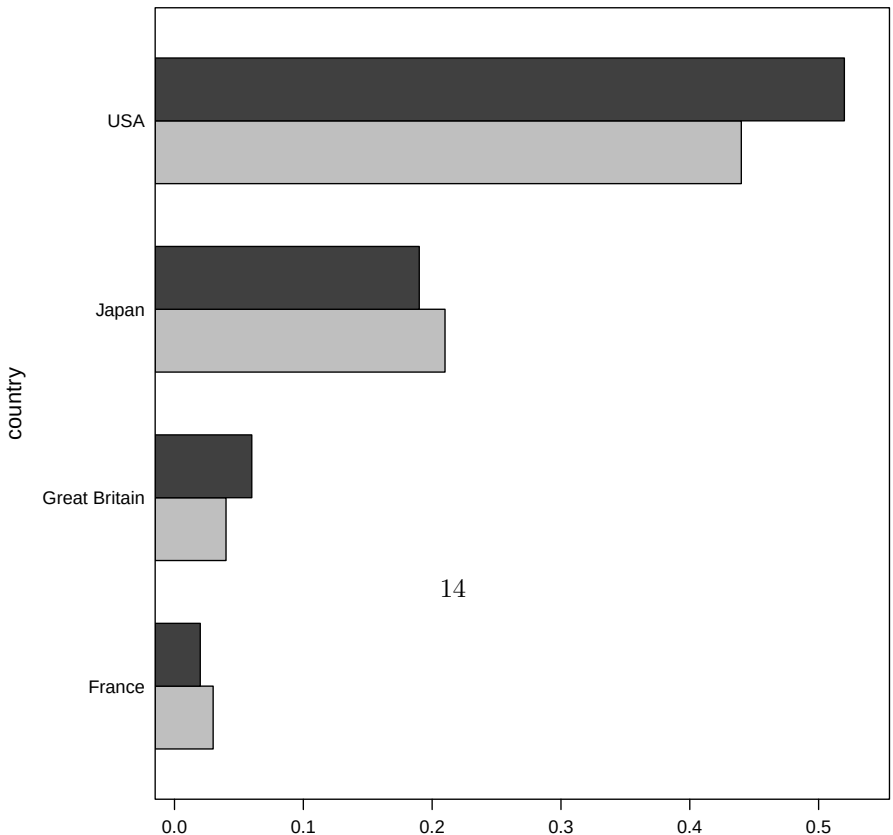
To execute the matching procedure, we use the "MatchIt" package of Ho et al. (2005b) and explained in detail in Ho et al. (2005a). Figure 5 compares the exposures of the original portfolio with those of the propensity score-matched portfolio. There are no points of significant disagreement. The matched portfolio has nearly exactly the same characteristics as the original.

4.4 Performance

After we have found a matching portfolio, outperformance can be expressed as a simple difference of returns from the original to the matched portfolio. Mathematically, this can be expressed as

$$\sum_i w_i r_i - \sum_i \tilde{w}_i r_i,$$


Utils
Telcm
Shops
Other
NoDur
Money
Manuf
Hlth
HiTec
Enrgy
Durbl
sector
0.0
0.1
0.2
0.3
Exposure


USA
Japan
Great Britain
France
country
0.0
0.1
0.2
0.3
0.4
0.5

where r_i represents the forward return on security i, $\{w_i\}$ the portfolio weights of securities in the original portfolio and $\{\tilde{w}_i\}$ the portfolio weights of securities in the matched portfolio. Here i indexes all stocks in the universe, $i \in \{1, ..., n\}$, so that most $w_i = 0$.

For the StarMine portfolio, the matching portfolio attains a return of 1.00% versus a return of 2.5% for the original portfolio, so the portfolio slightly outperforms. The simple weighted average of the universe yields a somewhat higher outperformance, but this is narrowed when matching portfolios are used as benchmarks. This implies that some of the StarMine portfolio's total performance comes from favorable sector, country and market cap bets.

5 Random Portfolios

Instead of creating a single matched portfolio, consider creating 1,000. Consider Figure 4. Note how each stock in the target portfolio has many possible matches. Yet these matches may not all tell the same story. Inevitably, some will have higher return than others. The set of returns on these portfolios, each containing a different set of matched securities, provides some measure of uncertainty in our performance measurement. Each matched portfolio provides an estimate of "alpha" as the difference between its return and the original portfolio's return. This yields a distribution for "alpha."

The relative position of the original portfolio's return among the matched portfolios' returns is the primary quantity of interest. The relative position allows us to conduct a nonparametric hypothesis test of the claim that $alpha = 0$ based on the empirical distribution of matching portfolio returns.

Our approach to random portfolios is distinguished by its emphasis on matching portfolios. Burns (2004) and Dawson and Young (2003) use a similar procedure, but do not require the random portfolios to be matched to the original. As such, their random portfolios reflect the makeup of the universe more than the makeup of the target portfolio, and could have widely varying cap, sector and country exposures. We propose a stricter framework: limiting portfolios to those within a certain distance from the original.

5.1 Procedure

Before we can sample portfolios, we need some guidelines on how far a matching portfolio is allowed to be from the original. The tradeoff is between quality and quantity of the matched portfolios. The closer we require portfolios to be matched, the fewer of them available. The simplest approach is to set a treshold in the propensity score difference allowed for an individual stock to be considered a match, as done by Rosenbaum and Rubin (1985). We consider two such thresholds, 0.005 and 0.05 (recall that this is relative to the space of propensity scores, on $(0, 1)$) in the two figures below.

Next we must decide how we will sample a non-holding from among each holding's matches. To do this, we must induce a distribution among the matches.

The simplest possibility is to choose a matching non-holding uniformly at random, and that is the practice we adopt below. Also note that not all holdings have matches within a specified threshold. For these, the nearest available is selected, non-randomly. Hence the procedure collapses to simple matching as the threshold nears zero and collapses to Burns (2004)-style random portfolios as the threshold grows to 1. Finally, the matches are sampled sequentially, one holding at a time.

To recap, for our experiment we build $K = 100$ random portfolios as follows:

1. For each portfolio holding, define a "match" as any non-holding within a threshold distance of the holding.
2. If the holding has at least one match, choose a non-holding randomly from the set of matches. If the holding has zero matches, simply pick the nearest available non-holding.
3. Repeat until all portfolio holdings have matches.

Figure 6 shows the sector exposures of 100 random portfolios under two different propensity score difference thresholds, 0.005 and 0.05. For each portfolio we also computed "total absolute bias," defined as the sum of the absolute differences in sector exposures between the original and matched portfolio. Quintiles of total absolute bias appear on the X axis, demonstrating that the 0.005 threshold leads to somewhat less bias than the 0.05 threshold. The tick marks on the Y axis correspond to the sector exposures of the original portfolio.

Since most of the random portfolios' sector exposures line up closely with the original's, it is clear that the random portfolio procedure is indeed producing matched portfolios. There is visibly less bias with the tighter threshold, though the difference is moderate. For the returns results related in the following sections we use the 0.005-threshold random portfolio collection.

5.1.1 Results

We next compare the returns enjoyed by the random, matching portfolios to those of the original. Figure 7 shows the distribution of matched portfolio returns along with a vertical line representing the original portfolio's return. The StarMine portfolio outperforms 100.0% of the random, matching portfolios. This provides some evidence for outperformance of the StarMine portfolio, but not enough to achieve statistical significance.

6 Long-Short Portfolios

One great advantage of the matching portfolio framework is that it effortlessly allows for different weighting strategies, such as non-equal-weighted long portfolios and long-short portfolios. Managers have struggled to find decent benchmarks for long-short portfolios. Jacobs et al. (1999) write that there are no inherent benchmarks suitable for a long-short portfolio, besides the return on

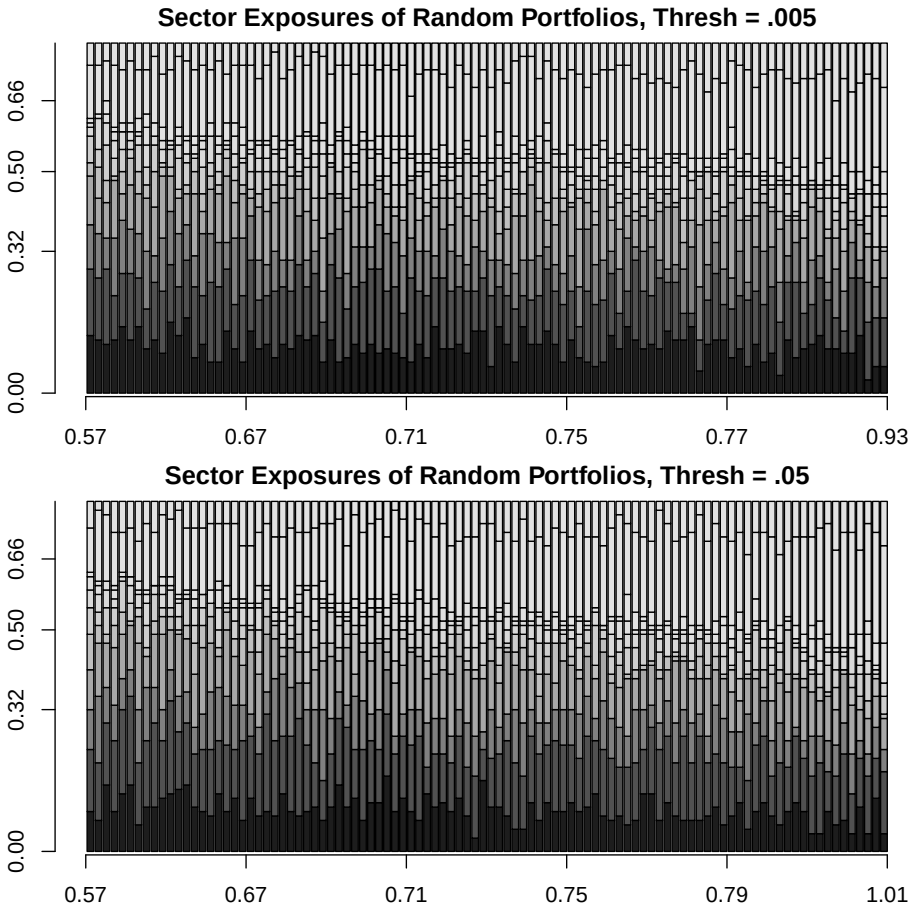


Figure 6: This figure shows the distribution of sector exposures of 100 random portfolios. The Y axis has tick marks at the cumulative exposures of the original portfolio. The X axis reflects quintiles in the distribution of absolute bias across the random portfolios. The portfolios are sorted from left to right in ascending order on total absolute bias, so the leftmost portfolios are closest to the original and the rightmost are farthest. There is a slight gain in closeness evident in the set of matching portfolios based on a tighter propensity score bound.

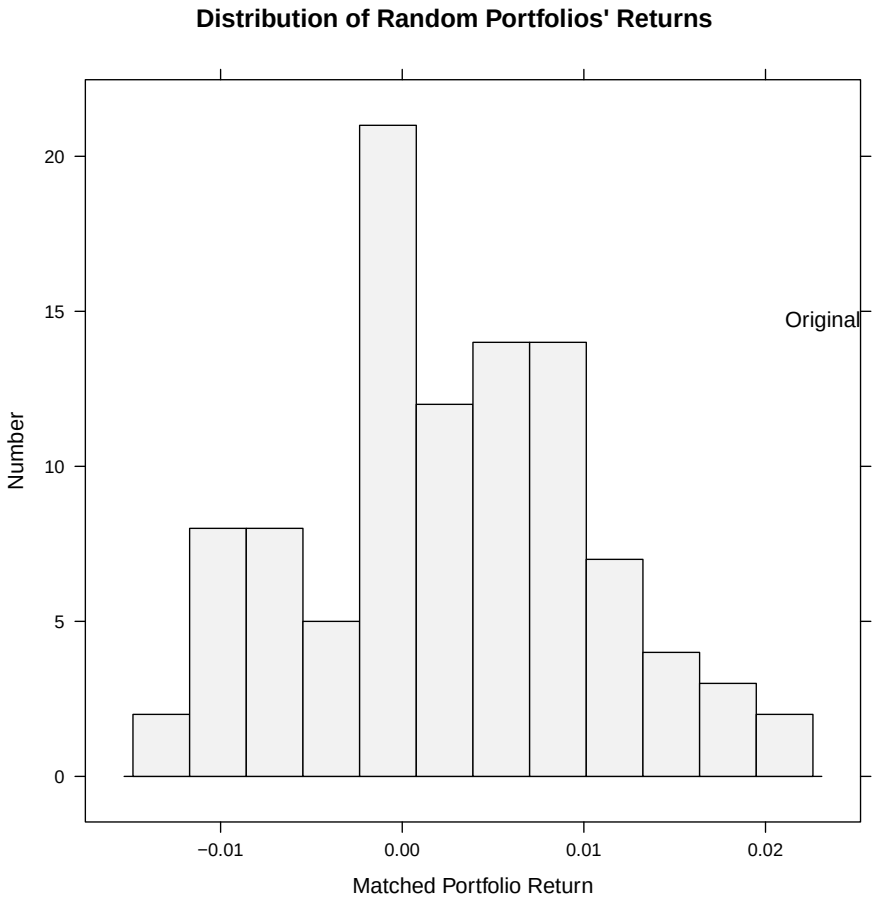


Figure 7: Histogram showing the returns of random, matched portfolios for the StarMine data. The thick vertical line represents the return realized by the StarMine decile portfolio. It stands at the 100.0% percentile of the random returns.

cash. In our framework, we simply create a matching portfolio, where each stock gets the same as its matching stock holds in the original portfolio.

In this section we build a long-short version of the StarMine portfolio based on the StarMine score. We then describe how to construct matching portfolios when the portfolio is not equally-weighted long using the "generalized propensity" score due to Imai and van Dyk (2004), Hirano and Imbens (2004) and Lu et al. (2001). Finally, we sample random portfolios and compare the performance results to other measures.

For the tests in this section we create a long-short portfolio based on the StarMine score. We divide the long position equally between stocks among the top 10% of StarMine scores and the short position between stocks in the bottom 10%. The total short position is 75% of the long position, creating a 100/75 long-short portfolio.

The long and short exposures of this portfolio are given in Table 8. Overall, the portfolio returns 5.0%.

Some of the mechanics of matching must change. Previously we followed the Rubin Causal Model by analogizing a stock's inclusion in the portfolio to what Rosenbaum and Rubin (1983) call a "treatment effect" (see Appendix). The variable I_i associated with this treatment is defined to be one when stock i is in the portfolio and zero otherwise. Matching on the propensity score $P(I_i = 1|X_i)$ yields matched stocks with the same probability of inclusion in the portfolio, which guarantees exposure balance between the original and matched portfolio. The definition of this probability depends on the treatment indicator I_i being a binary random variable.

But in a long-short portfolio, inclusion takes two forms: long and short. For a portfolio with unequal weights, inclusion can refer to any nonzero weighting. Generally, the treatment is best described as a continuous random variable, where $I_i = w_i$, stock i's weight in the original portfolio. Of course, since I_i is no longer binary, the propensity score method cannot be applied directly.

To match with continuous weights, we employ the "generalized propensity score" of Imai and van Dyk (2004). In this approach, we fit a model for $b(X_i) = E[I_i|X_i]$ as a function of X_i. The most popular choice is a linear regression, which writes $\hat{b}(X_i) = \hat{\beta}'X_i$ for a fitted coefficient p-vector $\hat{\beta}$. Both Imai and van Dyk (2004) and Hirano and Imbens (2004) show that the expected value acts as a balancing score, ensuring $I_i \perp X_i|b(X_i)$. This means that we can create a characteristic-matched portfolio by matching on the univariate score $b(X)$, just as we can for the propensity score in the binary treatment case.

Next we fit $\hat{b}(X_i)$ using a linear regression of the weights $I_i = w_i$ on the covariates X_i representing the same set of indicator variables as before. The R^2 from the fit is poor, only 0.06. Yet Lu et al. (2001) has shown that even a poor fit for the propensity score can lead to well-balanced matches.

To find matches, we adopt much the same procedure as we did in the simple propensity score case, with the added tweak that matched non-holdings receive the weight given their original holding. The propensity score procedure for the equal-weighted case can be considered a sub-case of the current methodology.

1. Match each holding i to a non-holding $\mathcal{P}(i)$ using the estimated generalized propensity score $\hat{\beta}$.

2. Assign non-holding $\mathcal{P}(i)$ a weight in the matched portfolio equal to i's weight in the original portfolio; i.e., $\tilde{w}_{\mathcal{P}(i)} = w_i$.

3. Continue matching until every holding has a matching non-holding.

In our case, Figure 8 shows the long and short exposures of the original and matched portfolios. We can see that both the long and short bets in the original portfolio line up very nicely with the matched portfolio's.

The matched portfolio returns -2.3% compared with 5.0% for the original portfolio. This suggests that the StarMine long-short portfolio exhibits some stock-picking ability; its returns are more than can be explained by its sector, country and market cap bets.

We can form random matching portfolios much the same way we did using the simple propensity score. We simply define a threshold distance for the generalized propensity score, within which every non-holding is taken to be a match. The scale of the propensity score is the same as the scale of the weights it models. In our case, we find a threshold of 0.0001, a single basis point, to lead to high-quality matches without sacrificing diversity.

The returns of the matching portfolios are shown in the histogram in Figure 9. Since the original portfolio's return falls at the 100.0% percentile of the random portfolios' returns, it turns out to be a little worse than we thought. We cannot reject the hypothesis that the stock-picking ability is zero.

7 Conclusion

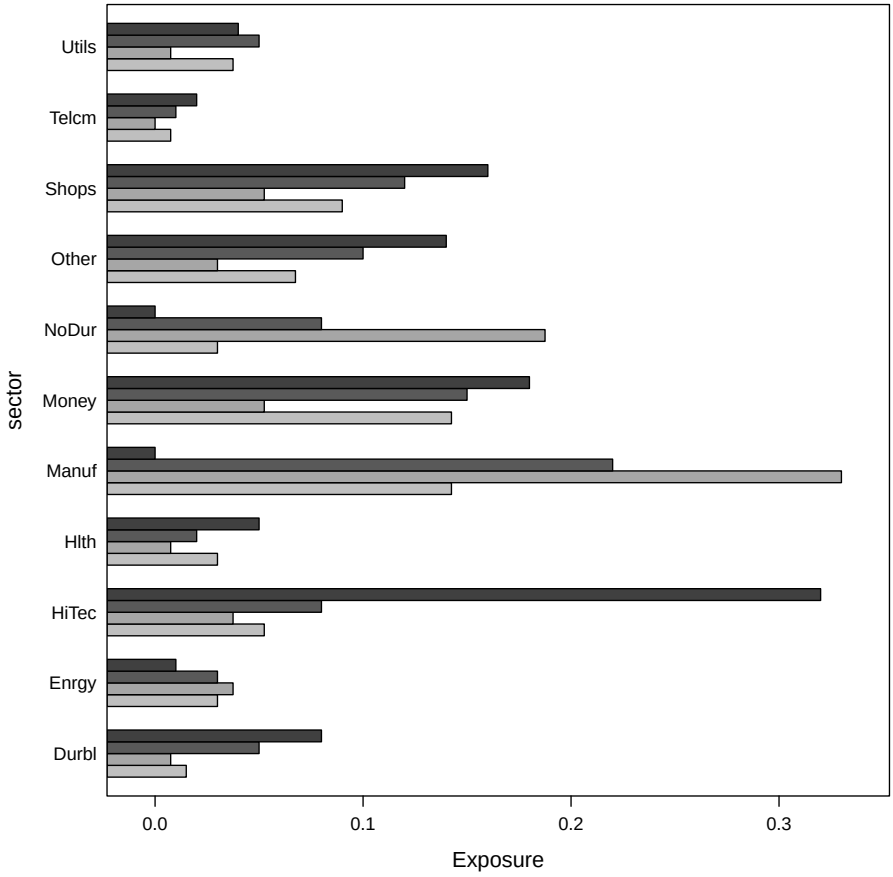
Our method evaluates performance by comparing the returns of the target portfolio to a counterfactual portfolio with the same characteristics but different holdings. This comparison isolates a manager's stock-picking ability from the effect of characteristic exposures. We have shown that the propensity score provides a flexible and precise means of constructing matching portfolios. Our methodology extends naturally to creating matching portfolios under any weighting scheme.

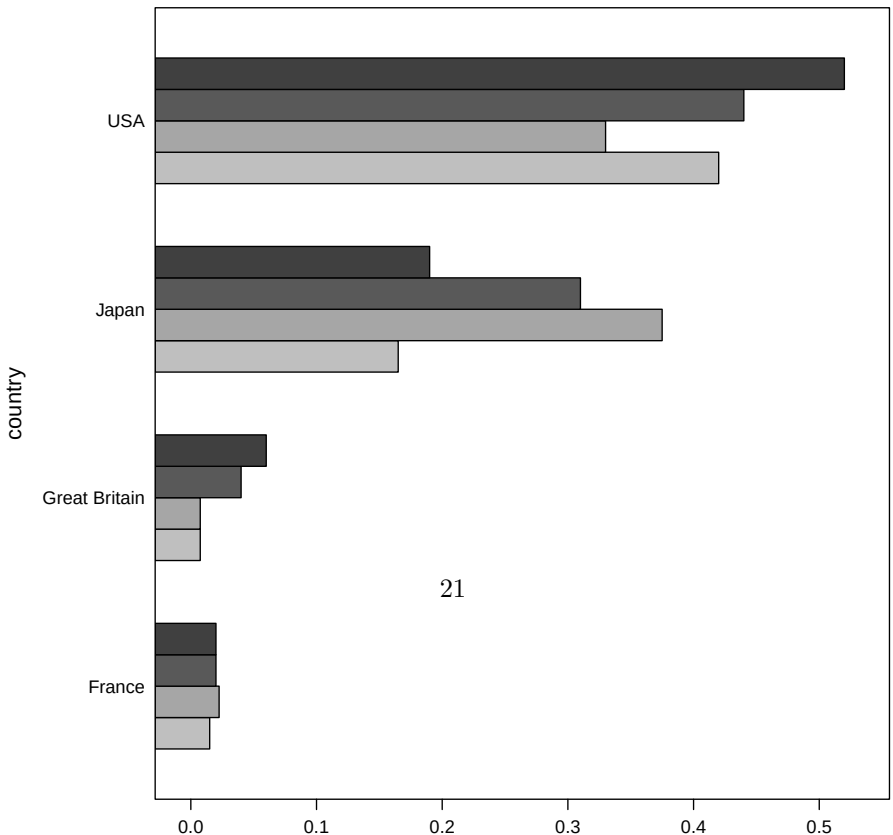
References

Alberto Abadie and Guido Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

Zvi Bodie, Alex Kane, and Alan J. Marcus. *Investments*, pages 811–812. McGraw-Hill-Irwin, 6 edition, 2001.

Patrick Burns. Performance measurement via random portfolios, 2004. URL `http://www.burns-stat.com/pages/Working/perfmeasrandport.pdf`.


Utils
Telcm
Shops
Other
NoDur
Money
Manuf
Hlth
HiTec
Enrgy
Durbl
sector
0.0
0.1
0.2
0.3
Exposure


USA
Japan
Great Britain
France
country
0.0
0.1
0.2
0.3
0.4
0.5

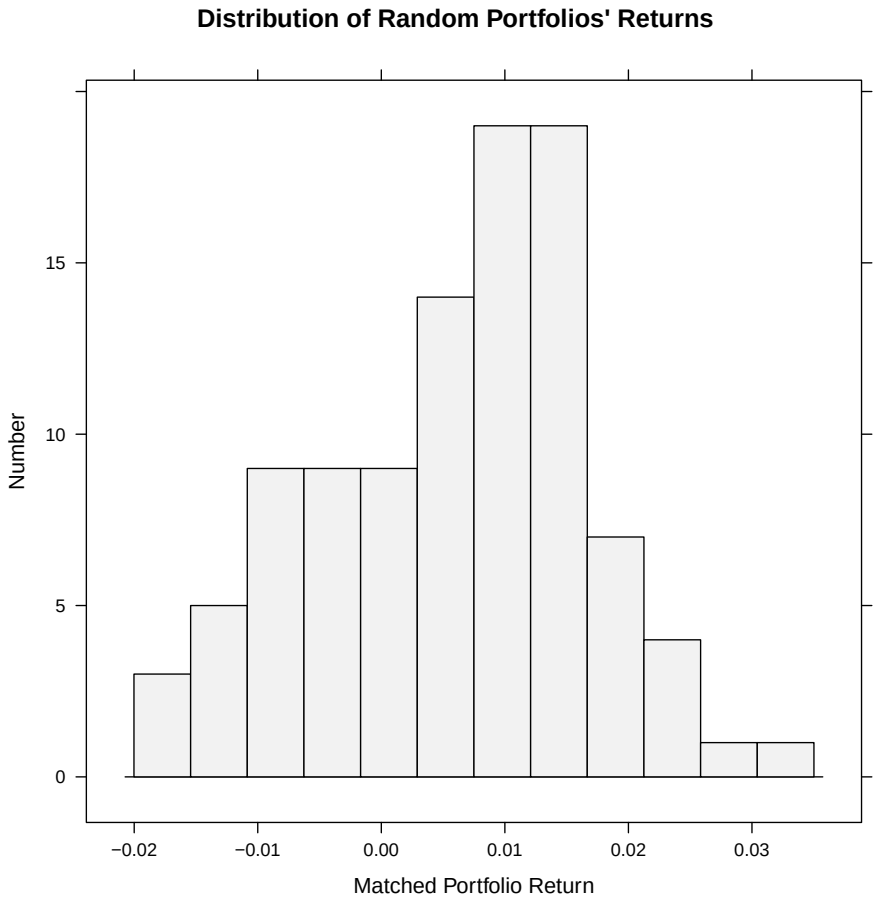


Figure 9: Histogram showing the returns of random portfolios matched to the long-short StarMine portfolio. The thick vertical line represents the return realized by the StarMine decile portfolio. It stands at the 100.0% percentile of the random returns, suggesting that the StarMine portfolio's excess return is indeed due to stock-picking ability.

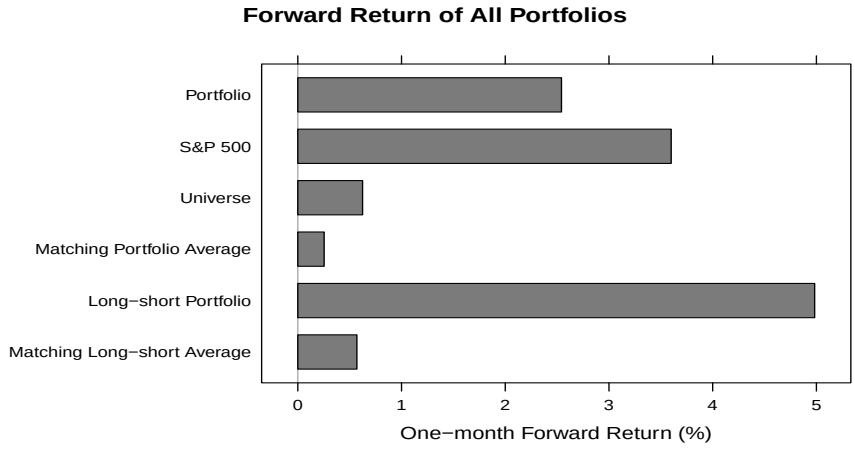


Figure 10: Comparisons of returns from the actual StarMine portfolio with returns from benchmark portfolios: the S&P 500; the simple average return of rated universe stocks; the average return of propensity score-based matching long-only portfolios; the long-short portfolio constructed based on the SMI score; and the average return of matching long-short portfolios.

W.G. Cochran and D.B. Rubin. Controlling bias in observational studies. *THE INDIAN JOURNAL OF STATISTICS SERIES A*, 35:417–446, Dec 1973.

Kalman J. Cohen and Bruce P. Fitch. The average investment performance index. *Management Science*, 12(6):B195–B215, feb 1966. ISSN 0025-1909.

Kent Daniel and Sheridan Titman. Evidence on the characteristics of cross sectional variation in stock returns. *The Journal of Finance*, 52(1):1–33, mar 1997. ISSN 0022-1082.

Kent Daniel, Mark Grinblatt, Sheridan Titman, and Russ Wermers. Measuring mutual fund performance with characteristic-based benchmarks. *The Journal of Finance*, 52(3):1035–1058, jul 1997. ISSN 0022-1082.

A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31, 1979. ISSN 0035-9246.

R. Dawson and R. Young. *Advances in Portfolio Construction and Implementation*, chapter Near-uniformly distributed, stochastically generated portfolios. Butterworth–Heinemann, 2003.

Eugene F. Fama and Kenneth R. French. The cross-section of expected stock returns. *The Journal of Finance*, 47(2), 1992.

Wayne E. Ferson and Meijun Qian. *Conditional Performance Evaluation, Revisited*. The Research Foundation of CFA Institute, 2004.

Mark Grinblatt and Sheridan Titman. Portfolio performance evaluation: Old issues and new insights. *The Review of Financial Studies*, 2(3):393–421, 1989. ISSN 0893-9454.

Mark Grinblatt and Sheridan Titman. Performance measurement without benchmarks: An examination of mutual fund returns. *The Journal of Business*, 66(1):47–68, jan 1993. ISSN 0021-9398.

Keisuke Hirano and Guido W. Imbens. The propensity score with continuous treatments, February 2004. A draft of a chapter for an upcoming book.

Daniel Ho, Kosuke Imai, Gary King, and Elizabeth Stuart. Matching as nonparametric preprocessing for parametric causal inference, 2005a. URL `http://gking.harvard.edu/files/matchp.pdf`.

Daniel Ho, Kosuke Imai, Gary King, and Elizabeth Stuart. *MatchIt: Nonparametric Preprocessing for Parametric Casual Inference*, 2005b. URL `http://gking.harvard.edu/matchit`. R package version 2.2-5.

Kosuke Imai and David A. van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467), September 2004.

Bruce I. Jacobs, Kenneth N. Levy, and David Starer. Long-short portfolio management: An integrated approach. *The Journal of Portfolio Management*, Winter:21–32, 1999.

Michael C. Jensen. The performance of mutual funds in the period 1945-1964. *The Journal of Finance*, 23(2), 1968.

Robert Kosowski, Allan Timbermann, Russ Wemers, and Hal White. Can mutual fund "stars" really pick stocks? new evidence from a bootstrap analysis. *The Journal of Finance*, 61(6):2551–2595, 2006.

S. P. Kothari and Jerold B. Warner. Evaluating mutual fund performance. *The Journal of Finance*, 56(5):1985–2010, 2001.

James H. Lorie. Current controversies on the stock market, 1965.

Bo Lu, Elaine Zanutto, Robert Hornik, and Paul R. Rosenbaum. Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association*, 96(456):1245–1253, December 2001.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, apr 1983. ISSN 0006-3444.

Paul R. Rosenbaum and Donald B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, February 1985.

Donald B. Rubin. Matching to remove bias in observational studies. *Biometrics*, 29(1):159–183, mar 1973a. ISSN 0006-341X.

Donald B. Rubin. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29(1):185–203, mar 1973b. ISSN 0006-341X.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

Donald B. Rubin and Neal Thomas. Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, 52:249–264, 1996.

S. S. Wilks. On the distribution of statistics in samples from a normal population of two variables with matched sampling of one variable. *Metron*, 9:87–126, 1932.