Resampling DEA estimates of investment fund performance

This item was submitted to Loughborough University’s Institutional Repository by the/an author.


Additional Information:

- This article was submitted for publication in the serial European Journal of Operational Research [© Elsevier]. The definitive version is available at: http://dx.doi.org/10.1016/j.ejor.2012.07.015

Metadata Record: https://dspace.lboro.ac.uk/2134/14796

Version: Submitted for publication

Publisher: © Elsevier

Please cite the published version.
This item was submitted to Loughborough’s Institutional Repository (https://dspace.lboro.ac.uk/) by the author and is made available under the following Creative Commons Licence conditions.

For the full text of this licence, please go to:
http://creativecommons.org/licenses/by-nc-nd/2.5/
Resampling DEA estimates of investment fund performance

John D. Lamb
University of Aberdeen Business School, Aberdeen AB24 3QY

Kai-Hong Tee
Loughborough University Business School, Loughborough LE11 3TU

Data envelopment analysis (DEA) is attractive for comparing investment funds because it handles different characteristics of fund distribution and gives a way to rank funds. There is substantial literature applying DEA to funds, based on the time series of funds’ returns. This article looks at the issue of uncertainty in the resulting DEA efficiency estimates, investigating consistency and bias. It uses the bootstrap to develop stochastic DEA models for funds, derive confidence intervals and develop techniques to compare and rank funds and represent the ranking. It investigates how to deal with autocorrelation in the time series and considers models that deal with correlation in the funds’ returns.

1 Introduction

Consider an investor or fund manager who has data on the returns of \( n \) investment funds over \( T \) recent time periods and wishes to know if fund \( o \) is likely to perform well in future. They have a performance measure \( \phi_o \). It is not known exactly and can only be estimated from the data as \( \phi_{o,T} \). We consider here the case where \( \phi_o \) is a (population) efficiency from a data envelopment analysis (DEA) model whose decision-making units are the \( n \) investment funds. But, in general, \( \phi_o \) can be another performance measure such as the Sharpe, Calmar or Sortino (Lhabitant, 2004) ratio. And the methods we describe apply whenever DEA efficiencies are estimated from a time series of observed data.

Our purpose is to investigate the following. First, is \( \phi_{o,T} \) a consistent estimator of \( \phi_o \)? If it is, is it biased? And if it is biased, how can we estimate the bias? Second, how can we obtain a confidence interval for each \( \phi_o \)? Third, how can we compare funds statistically and rank them according to their performance? We investigate these assuming initially no autocorrelation in the time series and later allowing for autocorrelation. We test the methods we develop on both standard and diversification-consistent (Lamb and Tee, 2011) DEA models, which are arguably more appropriate for investment funds.

Investment funds (especially hedge funds) are often characterised by strong skewness and kurtosis (Brooks and Kat, 2002) and not just mean and variance. These makes DEA attractive because it can handle multiple properties of the fund distributions and provide a means to rank them. Ranking is important because it helps identify funds that perform well and so, for example, helps form a fund of hedge funds portfolio efficiently. However, the accuracy of the ranking requires the investigation of the statistical properties of the DEA models, the main purpose of the paper.

Various studies (e.g. Murthi et al. (1997); Gregoriou and Zhu (2005); Chen and Lin (2006); Lozano and Gutiérrez (2008a); Kumar et al. (2010)) apply deterministic DEA to the problem of measuring the efficiency of investment funds. Studies have considered both choice of model (e.g. Gregoriou and Zhu (2005); Lozano and Gutiérrez (2008a,b); Briec and Kerstens (2010)) and choice of measure (e.g. Eling (2006)). Our findings are valid for a wide range of models and measures and so we do not discuss these in detail. Lamb and Tee (2011) discuss what models and risk (input) and return (output)
measures are appropriate for modelling funds.

As far as we know, stochastic DEA has not been applied before to modelling funds. However, several methods for stochastic DEA are developed in other contexts. In particular we note the methods developed by Simar and Wilson (1998); Löthgren and Tambour (1999a,b); Löthgren and Tambour (2000); Kuosmanen and Johnson (2010), which have been applied to modelling Italian banking data (Ferrier and Hirschberg, 1997) and Spanish public services (González and Miles, 2002). These methods use the bootstrap (Efron and Tibshirani, 1998) to deal with uncertainty, and Souza et al. (2011) use a similar method for confidence intervals. These methods assume only one observation of each input and output for each decision making unit (DMU), which means the bootstrap data generating processes cannot include every DMU in each replication. By contrast, investment fund data has many observations for each fund (or DMU) and so we can use a more conventional bootstrap method. We know of one example of stochastic DEA with multiple observations for each DMU: Kao and Liu (2009) model Taiwanese banks with five values for each input and output but do not use the bootstrap. Dyson and Shale (2010) discuss more generally the problems that arise in stochastic DEA and how to deal with them.

Investment funds are typically ranked (see, for example, Morningstar (2011)). Ranking is straightforward if we know the values of the performance measure exactly, as in deterministic DEA with precisely known inputs and outputs. In stochastic DEA ranking must be reassessed. The exact rank of a fund cannot be known for sure and may change from one resample to the next. In this case, ranking will be based on the average value of the efficiency scores out of the entire re-sampling. However, using the average value could bias the accuracy of the rank position, an issue which we will address in the paper. Still, the effect of biases could complicate the position ranking of two funds, especially when the average values of their scores are very similar. Therefore, an alternative approach we adopt in this paper is to use partial ranking to ascertain the relative positions of funds. Partial ranking ensures that one fund is ranked above another whenever its performance is statistically significantly better. Partial ranking is a partial ordering \( \preceq \) that we can best show in figures like Fig. 2 as a Hasse diagram (Rutherford, 1965). A Hasse diagram shows an arrow from \( i \) to \( j \) if \( a) \phi_i \preceq \phi_j \) and \( b) \) there is no \( k \) with \( \phi_i \preceq \phi_k \) and \( \phi_k \preceq \phi_j \).

Section 2 presents basic DEA models and shows that \( \phi_{o,T} \) is, in general, a consistent estimator of \( \phi_o \). Section 3 describes a bootstrap data generating process and how we can use it to produce an unbiased estimator of \( \phi_o \). It then shows how we can find bootstrap confidence intervals for DEA efficiency and how we can compare and rank funds statistically. Section 4 discusses how to deal with autocorrelation. Section 5 applies the methods to DEA models that allow the possibility of risk reduction through diversification. Furthermore, Sections 3, 4 and 5 discuss fund ranking and related issues in the resampling process. And Section 6 presents some discussion and conclusions.

2 Consistency of DEA efficiency estimates

A sequence of estimators \( \theta_T \) \( (T = 1, 2, \ldots) \) for a parameter \( \theta \) is consistent if \( \theta_T \) converges in probability to \( \theta \) as \( T \to \infty \). Clearly we want DEA estimates to be consistent. We show they are when \( T \) is the number of observations for each fund. Contrast this with Banker (1993), whose result is of the form \( \phi_{o,n} \to \phi_o \) in probability as \( n \to \infty \), where \( n \) is the number of DMUs.

We now describe the main DEA models we use. Section 3 discusses the specific measures we use. We assume we have investment funds \( f_1, \ldots, f_n \) whose returns for time periods \( 1, \ldots, T \) are given as \( r_{jt} \) \( (j = 1, \ldots, n, t = 1, \ldots, T) \). We assume we have nonnegative risk measures \( x_1, \ldots, x_m \) and nonnegative return measures \( y_1, \ldots, y_r \) that we can estimate for any fund. Let \( x_{ij} = x_i(f_j) \) \( (i = 1, \ldots, m, j = 1, \ldots, n) \) and let \( y_{rj} = y_r(f_j) \) \( (r = 1, \ldots, s, j = 1, \ldots, n) \). We use the risk and return measures to estimate an efficiency \( \phi_{o,T} \) for each fund \( f_o \) \( (o = 1, \ldots, n) \).

We use the following input-oriented model. Define \( \phi_{o,T} \) \( (o = 1, \ldots, n) \) as follows. If \( (x_{1o}, \ldots, x_{mo}) \leq 0 \) put \( \phi_{o,T} = 1 \). Otherwise, choose \( \phi_{o,T}, \lambda_1, \ldots, \lambda_n \) to minimise \( \phi_{o,T} \) subject to
We denote by \( \phi \) the population efficiencies in the range \([0,1]\). It is the model of A simplified by setting \( \hat{n} = 0 \).

We use an output-oriented \( \text{nirs} \) model that gives an efficiency \( \phi_{o,T} \) in the range \([0,1]\). It is the model of \( A \) simplified by setting \( \hat{n} = 0 \).

We denote by \( \phi \) and \( \phi' \) the population efficiencies we would get if we could replace the risk and return measures with the corresponding population statistics.

The following results help establish when \( \phi_{o,T} \) and \( \phi'_{o,T} \) are consistent. Let \( X \) be the \( m \times n \) matrix whose \((i,j)\)th entry is \( x_{ij} \). Let \( Y \) be the \( s \times n \) matrix whose \((r,j)\)th entry is \( y_{rj} \). Call a function \( f(X,Y) \) \( p\)-continuous if it is continuous with respect to perturbations in \( X, Y \) that leave elements that take the value zero unchanged.

**Proposition 1** The efficiencies \( \phi \) and \( \phi' \) are \( p\)-continuous for almost every value of \( X \) and \( Y \).

Proof. Note that if we add a fund all of whose risk and return measures are zero to either model, we get a variable returns to scale model. Then the result for \( \phi \) follows *a fortiori* from Scheel and Scholtes (2002, Proposition 3). The result of Scheel and Scholtes (2002) requires a linear programme with a bounded feasible set and so cannot be used directly for \( \phi' \). But if \((y_{1o},\ldots,y_{so}) = 0\) then \( \phi' \) is \( p\)-continuous with probability 1 by definition. Otherwise, define \( \psi_{ro} (r = 1,\ldots,s) \) by the linear programme: choose \( \lambda_1,\ldots,\lambda_n \) to

\[
\psi_{ro} = \sum_{j=1}^{n} \lambda_j y_{rj}
\]

subject to constraints (9) (with \( \hat{n} = 0 \)) and (4). (Banker (1993) uses this programme to estimate frontier points when \( s = 1 \).) Put \( \eta_{ro} = y_{ro}/\psi_{ro} \) if \( \psi_{ro} > 0 \) otherwise. Then \( \eta_{ro} = \max_{r=1,\ldots,s} \eta_{ro} \) is the solution to the output-oriented \( \text{nirs} \) model. The linear programme has a bounded feasible set and so (see Scheel and Scholtes (2003); Robinson (1977)) each \( \eta_{ro} \) is continuous at \( X \) and \( Y \). Since the functions \( \max \) and \( \eta_{ro} \) are \( p\)-continuous almost everywhere it follows that \( \phi'_{o} = 1/\eta_{ro} \) is also \( p\)-continuous almost everywhere.

The following well-known result of Mann and Wald (1943) is called the continuous mapping theorem.

**Proposition 2** Let \( g : S \to T \) be continuous almost everywhere and let \( w_n \) be a sequence of random vectors on \( S \) such that \( w_n \) converges in probability to \( w \). Then \( g(w_n) \) converges in probability to \( g(w) \).

**Proposition 3** Suppose \( x_1,\ldots,x_m, y_1,\ldots,y_s \) are consistent measures. Suppose that each estimate of \( x_i(f) \) \((i = 1,\ldots,m)\) is either always zero or zero with probability 0 and that each estimate of \( y_r(f) \) \((r = 1,\ldots,s)\) is either always zero or zero with probability 0. Then \( \phi_{o,T} \) and \( \phi'_{o,T} \) are consistent estimators of \( \phi \) and \( \phi' \) for \( o = 1,\ldots,n \).

Proof. We can consider \( \phi \) and \( \phi' \) as functions of the elements of \( X, Y \) that are not always zero. Then they are continuous almost everywhere by Proposition 1 and so...
\(\phi_{o,T}\) and \(\phi'_{o,T}\) are consistent estimators by Proposition 2.

Although Proposition 3 is proved for nirs models, it is straightforward to check that it holds also for vrs models.

3 Bootstrap, bias, confidence intervals and comparisons

We have shown \(\phi_{o,T}\) is consistent but wish to know more about its properties. The bootstrap is a method for investigating the properties of a sample statistic (for example, estimating bias and standard error) such as \(\phi_{o,T}\) when the distribution of the sample statistic is unknown. It is particularly useful when parametric methods are inappropriate. Clearly the distribution of \(\phi_{o,T}\) is bounded below by zero and above by one and plausibly may take either with nonzero probability. We know of no parametric method that can be used in such circumstances.

3.1 Bootstrap data generating process

The bootstrap needs a data generating process (Efron and Tibshirani, 1998). Our data generating process is as follows. We generate a number of replications. These are data sets of the form \(r_{it} = r_{ip}(t)\) (\(i = 1, \ldots, n, \ t = 1, \ldots, T\)) where \(\rho(1), \ldots, \rho(t)\) are generated uniformly at random with replacement from \(\{1, \ldots, T\}\). This is a standard approach for generating data sets of vectors and preserves correlations between returns of different funds. We use 2000 replications in each test we carry out because Efron and Tibshirani (1998) indicate this is a reasonable number for each bootstrap method we use. We estimate the DEA efficiencies for each replication and use them in various bootstrap procedures.

3.2 Specific data and measures

We use data comprising 60 monthly returns from 30 hedge funds for the period 2000–2004. The funds come from Center for International Securities and Derivatives Markets (2010) and are broadly classified as market neutral (MN), long/short strategy (LS) and global macro (GM).

Although we may use any risk or return measures, we illustrate our method with measures that satisfy a property called convexity consistency (Lamb and Tee, 2011) because we need such measures in Section 5. Convexity consistency is closely related to the well-known notion of a coherent measure of risk (Artzner et al., 1999; Rockafellar and Uryasev, 2002). We use measures based on two particular coherent measures of risk. The first is conditional value at risk: \(\text{cvvar}_{\alpha}\) is the expected loss on an investment conditional on that loss not exceeding \(\text{var}_{\alpha}\), where \(\text{var}_{\alpha}\), the value at risk at \(\alpha\), is the negative of the 100\(\alpha\)% percentile of the distribution of returns. Both \(\text{cvvar}\) and \(\text{var}\) are widely used in the finance literature (Jorion, 2007). The second measure is the lower semideviation minus the mean (Ogryczak and Ruszczyński, 1999). We use the same four specific risk and return measures for every test we carry out. The risk measures are \(\max(\text{cvvar}_{0.2}, 0)\) and \(\max(\text{sd}, 0)\), where \(\text{sd}\) is the lower semideviation minus the mean. The return measures are the mean return and \(\max(−\text{cvvar}_{0.1}, 0)\).

![Figure 1: Efficient frontiers of simplified nirs and diversification-consistent nirs models](image)

Fig. 1 plots mean against \(\max(\text{cvvar}_{0.2}, 0)\) for each fund and the solid line shows the efficient frontier from the simplified nirs DEA model that uses only these two measures. Note that the frontier does not depend on whether the model is input-oriented or output-oriented.

3.3 Bias

We first estimate the bias and derive bias-corrected estimates of the DEA efficiencies. We use the bootstrap estimate of bias of Efron and Tibshirani (1998, Sec-
Although there are better bootstrap methods to estimate bias (Efron, 1990), we cannot adapt them for DEA because the efficiencies are given by linear programmes rather than an explicit formula.

Columns 2–4 of Table 1 show, for each fund, the raw efficiency (labelled $\phi_T$, raw efficiency for short) bootstrap bias-corrected DEA efficiency (labelled $\phi_{bc,T}$) and bootstrap standard error (labelled se) for the input-oriented model. The bootstrap estimate of bias can be found as

$$\hat{\text{bias}} = \phi_T - \phi_{bc,T}.$$  

(6)

We estimate also the bootstrap standard error using Efron and Tibshirani (1998, Algorithm 6.1). The table is ordered by bias-corrected efficiency and we note that the ordering is noticeably different from that given by the raw efficiencies.

The bootstrap bias we observe is invariably less than the standard error, but is large enough to suggest it may be better to use bias-corrected efficiency estimates. The bias is negative for funds that have efficiency close to 1 though some funds have positive bias. This is expected. The basic DEA model must produce some funds with efficiency 1 but it unlikely that these funds will achieve efficiency 1 in every replication.

Section 5 discusses columns 5–7 of Table 1.

### 3.4 Confidence intervals

There are several bootstrap methods for confidence intervals (Efron and Tibshirani, 1998). We cannot use the parametric bootstrap because it assumes we know the form of the distribution of the DEA efficiencies. So we seek a nonparametric bootstrap method. Suppose a method gives a $1 - 2\alpha$ confidence interval $(\phi_\alpha, \phi_{1-\alpha})$ satisfying

$$P(\phi < \phi_\alpha) - \alpha = O(T^{-r})$$  

and

$$P(\phi > \phi_{1-\alpha}) - \alpha = O(T^{-r}),$$  

(7)

for some $r$, where $T$ is the sample size and $O$ is the Landau symbol. The method is first-order accurate if it satisfies Eq. (7) with $r = 1$ and second-order accurate if it satisfies it with $r = 2$. Ideally we want a second-order accurate method. The bias-corrected accelerated (BCa) method (Efron and Tibshirani, 1998) is second-order accurate depending on the assumption (Efron, 1987) that there is a strictly increasing function $g$ and constants $z_0$ and $a$ such that

$$g(\phi_T) = g(\phi) + (1 + ag(\phi))(Z - z_0), \quad Z \sim N(0, 1),$$  

(8)

where $\phi$ is the population efficiency and $\phi_T$ is sample efficiency. However, $\phi_{0,T}$ is bounded below by zero and above by one, and we find that in practice it may take either value with positive probability. So the assumption of a strictly-increasing $g$ satisfying Eq. (8) is not plausibly even approximately true for DEA efficiency estimates.

The bootstrap percentile interval (Efron and Tibshirani, 1998) is first order accurate and does not depend on assumptions about the distribution of efficiency estimates, and so we report percentile confidence intervals for the efficiencies. In practice, we find the percentile and BCa intervals are very similar.

Table 2 presents the funds in the same order as Table 1. Columns 2–3 show the 95% percentile bootstrap confidence intervals for each fund using the input-oriented model. Columns 2–3 of Table 3 show the 95% percentile bootstrap confidence intervals using the output-oriented model. Sections 4 and 5 discuss the remaining columns of both tables.

### 3.5 Comparisons

The confidence intervals of Table 2 are wide. Nonetheless, we may reasonably hope to observe significant differences between many pairs of funds if we can use a matched pairs test. Such a test is possible because we have matching observations for each fund in each time period. We can test individual differences in efficiency between funds using a one-sample bootstrap hypothesis test (Efron and Tibshirani, 1998, Chapter 16) based on the difference between efficiencies. As in Section 3.4, we use bootstrap percentiles for the differences in efficiency to construct individual hypothesis tests.

As usual, when we carry out multiple matched-pairs tests, some tests may appear significant by chance alone (our example has 435 tests). The usual adjustments are inappropriate because a transitive comparison rela-
We modify the procedure slightly by using the more accurate šidák (1967) equation rather than the Bonferroni approximation for the experimentwise significance level. We also consider in Section 5 a version of Shaf-fer’s procedure, which is called least significant differences (LSD). It uses a single significance level and identifies all individually significant differences that do not violate the transitivity assumption. In all the comparisons we use a familywise significance level of 5%.

Fig. 2 (first diagram) should be compared with Table 1. It shows as a Hasse diagram (see Section 1) the ranking of the funds in the input-oriented model using Shaffer’s procedure. The same comparison procedure with the output-oriented model shows far fewer significant differences.

Table 1: dea efficiencies and bootstrap bias-corrected values

<table>
<thead>
<tr>
<th>FUND</th>
<th>$\phi_T$</th>
<th>$\phi_{bc,T}$</th>
<th>SE</th>
<th>$\phi_T$</th>
<th>$\phi_{bc,T}$</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM45</td>
<td>1.000</td>
<td>0.931</td>
<td>0.195</td>
<td>0.391</td>
<td>0.196</td>
<td>0.334</td>
</tr>
<tr>
<td>GM2693</td>
<td>0.881</td>
<td>0.727</td>
<td>0.313</td>
<td>0.031</td>
<td>0.052</td>
<td>0.130</td>
</tr>
<tr>
<td>GM366</td>
<td>0.860</td>
<td>0.695</td>
<td>0.306</td>
<td>0.030</td>
<td>0.046</td>
<td>0.112</td>
</tr>
<tr>
<td>LS90</td>
<td>1.000</td>
<td>0.659</td>
<td>0.502</td>
<td>0.773</td>
<td>0.442</td>
<td>0.519</td>
</tr>
<tr>
<td>MN629</td>
<td>0.720</td>
<td>0.584</td>
<td>0.322</td>
<td>0.025</td>
<td>0.028</td>
<td>0.063</td>
</tr>
<tr>
<td>MN2681</td>
<td>0.535</td>
<td>0.564</td>
<td>0.325</td>
<td>0.173</td>
<td>0.106</td>
<td>0.200</td>
</tr>
<tr>
<td>GM2695</td>
<td>0.691</td>
<td>0.520</td>
<td>0.406</td>
<td>0.456</td>
<td>0.331</td>
<td>0.394</td>
</tr>
<tr>
<td>GM2640</td>
<td>1.000</td>
<td>0.492</td>
<td>0.662</td>
<td>1.000</td>
<td>0.440</td>
<td>0.712</td>
</tr>
<tr>
<td>MN2540</td>
<td>0.618</td>
<td>0.491</td>
<td>0.305</td>
<td>0.022</td>
<td>0.024</td>
<td>0.055</td>
</tr>
<tr>
<td>LS40</td>
<td>0.582</td>
<td>0.416</td>
<td>0.347</td>
<td>0.362</td>
<td>0.430</td>
<td>0.411</td>
</tr>
<tr>
<td>MN2547</td>
<td>0.470</td>
<td>0.365</td>
<td>0.234</td>
<td>0.017</td>
<td>0.020</td>
<td>0.044</td>
</tr>
<tr>
<td>GM2889</td>
<td>0.315</td>
<td>0.305</td>
<td>0.254</td>
<td>0.011</td>
<td>0.027</td>
<td>0.085</td>
</tr>
<tr>
<td>MN2704</td>
<td>0.245</td>
<td>0.180</td>
<td>0.133</td>
<td>0.009</td>
<td>0.010</td>
<td>0.032</td>
</tr>
<tr>
<td>LS6</td>
<td>0.092</td>
<td>0.131</td>
<td>0.194</td>
<td>0.003</td>
<td>0.049</td>
<td>0.156</td>
</tr>
<tr>
<td>LS78</td>
<td>0.160</td>
<td>0.125</td>
<td>0.112</td>
<td>0.006</td>
<td>0.008</td>
<td>0.021</td>
</tr>
<tr>
<td>MN2999</td>
<td>0.131</td>
<td>0.107</td>
<td>0.104</td>
<td>0.005</td>
<td>0.007</td>
<td>0.033</td>
</tr>
<tr>
<td>LS17</td>
<td>0.118</td>
<td>0.105</td>
<td>0.113</td>
<td>0.004</td>
<td>0.011</td>
<td>0.046</td>
</tr>
<tr>
<td>LS39</td>
<td>0.007</td>
<td>0.101</td>
<td>0.261</td>
<td>0.000</td>
<td>0.069</td>
<td>0.239</td>
</tr>
<tr>
<td>GM3434</td>
<td>0.069</td>
<td>0.098</td>
<td>0.174</td>
<td>0.002</td>
<td>0.029</td>
<td>0.123</td>
</tr>
<tr>
<td>LS79</td>
<td>0.037</td>
<td>0.098</td>
<td>0.216</td>
<td>0.001</td>
<td>0.050</td>
<td>0.185</td>
</tr>
<tr>
<td>MN301</td>
<td>0.117</td>
<td>0.097</td>
<td>0.099</td>
<td>0.004</td>
<td>0.007</td>
<td>0.040</td>
</tr>
<tr>
<td>LS64</td>
<td>0.102</td>
<td>0.097</td>
<td>0.133</td>
<td>0.004</td>
<td>0.014</td>
<td>0.071</td>
</tr>
<tr>
<td>GM6876</td>
<td>0.100</td>
<td>0.094</td>
<td>0.107</td>
<td>0.004</td>
<td>0.007</td>
<td>0.031</td>
</tr>
<tr>
<td>GM59</td>
<td>0.079</td>
<td>0.086</td>
<td>0.148</td>
<td>0.003</td>
<td>0.022</td>
<td>0.114</td>
</tr>
<tr>
<td>MN2777</td>
<td>0.000</td>
<td>0.083</td>
<td>0.253</td>
<td>0.000</td>
<td>0.062</td>
<td>0.237</td>
</tr>
<tr>
<td>LS20</td>
<td>0.040</td>
<td>0.073</td>
<td>0.169</td>
<td>0.001</td>
<td>0.029</td>
<td>0.149</td>
</tr>
<tr>
<td>LS12</td>
<td>0.042</td>
<td>0.070</td>
<td>0.150</td>
<td>0.001</td>
<td>0.025</td>
<td>0.129</td>
</tr>
<tr>
<td>MN2639</td>
<td>0.027</td>
<td>0.052</td>
<td>0.139</td>
<td>0.001</td>
<td>0.021</td>
<td>0.134</td>
</tr>
<tr>
<td>MN147</td>
<td>0.043</td>
<td>0.050</td>
<td>0.002</td>
<td>0.010</td>
<td>0.097</td>
<td>0.085</td>
</tr>
<tr>
<td>GM2068</td>
<td>0.023</td>
<td>0.041</td>
<td>0.092</td>
<td>0.001</td>
<td>0.007</td>
<td>0.068</td>
</tr>
</tbody>
</table>
We may assume that returns from investment funds are available as a time series of vectors \( r_t = (r_{1t}, \ldots, r_{nt}) \) \((t = 1, \ldots, t)\). In general we might reasonably expect this time series to be stationary but to show some autocorrelation. Exploratory data analysis in the form of sequence plots and Ljung–Box (Ljung and Box, 1978) tests suggests that our data have these properties. Several bootstrap methods are available (Bühlmann, 2002) for stationary vector time series. These deal with autocorrelation either by resampling blocks of data or by resampling the (assumed iid) residuals of a model of the data. We consider three: the block bootstrap (Kün-
We tried the method of Carlstein et al. (1998) with block sizes 2–6. Fig. 2 (third diagram) shows a Hasse diagram illustrating the significant differences using Shaffer’s procedure with the input-oriented model. The problem we note with this method is that it assumes that there is a hidden set of discrete performance states and individual funds move from time to time among them. But we have no reason to assume this.

The block bootstrap works by resampling blocks $r_i, \ldots, r_{i+l-1}$ of $l$ consecutive vectors, which reduces any autocorrelation effects. We resample with replacement as described in Künsch (1989). Hall et al. (1995) indicate a block size $l = 2–4$ should be reasonable and we have tried block sizes up to $l = 10$ with similar results, especially for smaller block sizes. Columns 6–7 of Tables 2 and 3 show 95% percentile bootstrap confidence intervals for the input-oriented and output-oriented models with block size $l = 3$. Fig. 2 (second diagram) shows a Hasse diagram illustrating the significant differences using Shaffer’s procedure with the input-oriented model.

The matched-block bootstrap uses blocks of fixed size but uses a transition matrix to choose consecutive blocks. We tried the method of Carlstein et al. (1998) with block sizes 2–6. Fig. 2 (third diagram) shows a Hasse diagram illustrating the significant differences using Shaffer’s procedure with the input-oriented model. The problem we note with this method is that it assumes that there is a hidden set of discrete performance states and individual funds move from time to time among them. But we have no reason to assume this.

We consider also a $\text{SVAR}$ bootstrap model, based closely on the vector autoregressive ($\text{VAR}$) bootstrap model of Section 3 of Inoue and Kilian (2002). A $\text{VAR}(k)$ model with vectors of length $n$ requires $n \times n$ matrices $A_1, \ldots, A_k$ of coefficients. We use an $\text{SVAR}$ model rather than a $\text{VAR}$ model because we have 60 observations for 30 funds, which only allows us to estimate the coefficients of a $\text{VAR}(k)$ model for $k = 1$. We can, however, estimate $\text{SVAR}(k)$ models. These have the same coefficient matrices as the $\text{VAR}(k)$ model but fewer coefficients because many of the entries of $A_1, \ldots, A_n$ are set to zero. We estimate $\text{SVAR}(k)$ models for $k = 2$ and 3 as described by Lütkepohl (2007) using the Akaike information criterion to decide which coefficients to set to zero. Fig. 2 (right diagram) shows the ranking of the funds using Shaffer’s procedure with the input-oriented model.
We note that the method is problematic for the data we have. The matrices $A_1, \ldots, A_k$ are sparse for both $k = 2$ and $k = 3$. And we have no theoretical grounds for supposing a particular structure for them.

We note that the ranking of funds is broadly similar to the original ranking for all three methods we consider and suggest the block bootstrap with small block size is appropriate when there is reason to doubt the assumptions needed for the other methods.

5 Models that deal with diversification

Consider Fig. 1. The solid line shows the standard DEA frontier. The grey dashed line shows another frontier. It shows the maximum return that an investor wealthy enough to hold a portfolio of hedge funds could have achieved at any given risk level. The reason these frontiers are different is that standard DEA does not allow for the very real possibility of reducing risk by diversification: that is, choosing a portfolio of imperfectly correlated investments. Several DEA models allow for diversification (Morey and Morey, 1999; Joro and Na, 2006; Lozano and Gutiérrez, 2008a, b; Briec and Kerstens, 2009, 2010). Most use nonlinear programmes. Lamb and Tee (2011) discuss them and introduce the diversification-consistent model, which we use here because it uses a linear programme and does not prescribe specific risk and return measures.

Lamb and Tee (2011) describe the input-oriented diversification-consistent model, and A describes the output-oriented diversification-consistent (nirs) model and the iterative procedure we use to compute efficiencies using the models.

Columns 5–7 of Table 1 show the raw and bootstrap bias-corrected efficiency estimates and bootstrap standard errors for our funds with the input-oriented diversification-consistent model. As before, we observe substantial bias in many raw efficiency estimates. Columns 4–5 of Table 2 show the 95% bootstrap percentile confidence intervals for our funds with the input-oriented diversification-consistent model. Columns 8–9 show the corresponding intervals with block (size 3) bootstrap. Again, the upper and lower bounds are invariably reduced relative to those from the corresponding standard model. But the reduction is much less and very few lower bounds are zero.
output-oriented model. And we noted in Section 3.5 that the input-oriented standard model produced more significant differences than the output-oriented model. Fig. 1 helps illustrates why we might observe these effects. Given a point on the standard frontier, we can always find a point on the diversification-consistent frontier with at least as much return on each return measure and at most as much risk on each risk measure. So, any solution of a standard model gives a feasible solution for a corresponding diversification-consistent model and so in each bootstrap resample the diversification-consistent model efficiency of a fund must be at most the standard model efficiency. Thus both the lower and upper bounds of a percentile bootstrap efficiency confidence interval from the diversification-consistent model must be at most those of the standard model.

Consider a fund $o$. Any solution $\lambda_1, \ldots, \lambda_n, \hat{\lambda}_1, \ldots, \hat{\lambda}_h$ (see constraint (10)) of the diversification-consistent output-oriented model must have at least as much re-

<table>
<thead>
<tr>
<th>FUND</th>
<th>STANDARD BOOTSTRAP</th>
<th>BLOCK (SIZE 3) BOOTSTRAP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>basic</td>
<td>2.5%</td>
</tr>
<tr>
<td>GM45</td>
<td>0.647</td>
<td>1.000</td>
</tr>
<tr>
<td>GM2693</td>
<td>0.460</td>
<td>1.000</td>
</tr>
<tr>
<td>GM366</td>
<td>0.454</td>
<td>1.000</td>
</tr>
<tr>
<td>LS90</td>
<td>0.439</td>
<td>1.000</td>
</tr>
<tr>
<td>MN229</td>
<td>0.340</td>
<td>1.000</td>
</tr>
<tr>
<td>MN2681</td>
<td>0.458</td>
<td>1.000</td>
</tr>
<tr>
<td>GM2695</td>
<td>0.379</td>
<td>1.000</td>
</tr>
<tr>
<td>GM2640</td>
<td>0.503</td>
<td>1.000</td>
</tr>
<tr>
<td>MN2540</td>
<td>0.344</td>
<td>1.000</td>
</tr>
<tr>
<td>LS40</td>
<td>0.489</td>
<td>1.000</td>
</tr>
<tr>
<td>MN2547</td>
<td>0.269</td>
<td>0.972</td>
</tr>
<tr>
<td>GM2889</td>
<td>0.282</td>
<td>1.000</td>
</tr>
<tr>
<td>MN2704</td>
<td>0.117</td>
<td>0.536</td>
</tr>
<tr>
<td>LS6</td>
<td>0.111</td>
<td>0.979</td>
</tr>
<tr>
<td>LS75</td>
<td>0.140</td>
<td>0.754</td>
</tr>
<tr>
<td>MN299</td>
<td>0.095</td>
<td>0.733</td>
</tr>
<tr>
<td>LS57</td>
<td>0.140</td>
<td>0.767</td>
</tr>
<tr>
<td>LS39</td>
<td>0.000</td>
<td>0.776</td>
</tr>
<tr>
<td>GM3434</td>
<td>0.056</td>
<td>0.897</td>
</tr>
<tr>
<td>LS97</td>
<td>0.091</td>
<td>1.000</td>
</tr>
<tr>
<td>MN301</td>
<td>0.081</td>
<td>0.710</td>
</tr>
<tr>
<td>LS64</td>
<td>0.075</td>
<td>0.833</td>
</tr>
<tr>
<td>GM6876</td>
<td>0.125</td>
<td>0.767</td>
</tr>
<tr>
<td>GM59</td>
<td>0.095</td>
<td>0.887</td>
</tr>
<tr>
<td>MN2777</td>
<td>0.000</td>
<td>0.472</td>
</tr>
<tr>
<td>LS20</td>
<td>0.000</td>
<td>0.598</td>
</tr>
<tr>
<td>LS32</td>
<td>0.000</td>
<td>0.603</td>
</tr>
<tr>
<td>MN2639</td>
<td>0.010</td>
<td>0.434</td>
</tr>
<tr>
<td>MN147</td>
<td>0.023</td>
<td>0.353</td>
</tr>
<tr>
<td>GM2068</td>
<td>0.058</td>
<td>0.597</td>
</tr>
</tbody>
</table>

Table 3: Bootstrap percentile confidence intervals for basic and diversification-consistent (dc) output-oriented models.
turn in each measure as the solution $\lambda_1, \ldots, \lambda_n$ of the standard model. But the return is not infinite and so, provided $\phi$ produces positive return in some measure, $\phi' o > 0$ in the diversification-consistent model. By contrast a solution of the diversification-consistent input-oriented model has at most as much risk as the corresponding solution in the standard model and may have zero risk and so zero efficiency provided $o$ has positive risk. In practice, we find this happens rather often in bootstrap replications for low-return funds, making it hard to distinguish statistically among them.

6 Discussion

The ranking of investment funds is important in supporting fund selection, especially for hedge funds. And DEA allows us to compare and rank funds based on the characteristics of funds’ return distributions. But, the accuracy of the ranking also depends on the statistical properties of DEA. We investigate these properties and conclude the following.

First, the DEA efficiency estimates of investment fund performance typically have substantial bias. Worse, this bias may be markedly different for different funds. In particular, high-risk high-return funds may show greater bias than low-risk low return funds. Bootstrap methods help us deal with this bias.

Second, the bootstrap allows us to estimate the accuracy of DEA efficiency estimates. In particular, it gives us confidence intervals for efficiency scores.

Third, resampling gives us a better way than using simple DEA scores to approximately rank investment funds. The ranking should nearly always be a partial ordering rather than a total ordering, and we find Hasse diagrams indicate the ranking parsimoniously.

Fourth, the substantial bias, wide confidence intervals and partial orderings whose precise structure depends on the bootstrap method used suggest we should be modest in how we interpret DEA efficiencies. This is especially true for the model suggested by Lamb and Tee (2011) that accounts for correlation between the performance of different funds. We can identify approximately how well we expect an individual fund to perform in future compared to its peers. But it is difficult to predict with confidence precisely what its future efficiency will be.

We note that the methods we use can be applied to many different DEA models or even to comparing funds using more conventional methods such as Sharpe ratios. We anticipate the same issues of bias, accuracy and ranking will arise.

A Output-oriented diversification-consistent DEA model and iterative procedure

Lamb and Tee (2011) describe an input-oriented diversification-consistent DEA model. The corresponding output-oriented model is as follows. A portfolio is...
a combination of funds $\sum_{j=1}^{n} \mu_j f_j$ with $\sum_{j=1}^{n} \mu_j \leq 1$ and $\mu_j \geq 0$ for $j = 1, \ldots, n$. Suppose we have portfolios $f_1, \ldots, f_n$, which we call notional units. Let $\hat{x}_{ij}$ be the risk measure estimate using measure $x_j$ for $f_j$ ($i = 1, \ldots, m$, $j = 1, \ldots, \hat{n}$). Let $\hat{y}_{rj}$ be the risk measure estimate using measure $y_r$ for $f_j$ ($r = 1, \ldots, s$, $j = 1, \ldots, \hat{n}$). Define $\phi_{o,T}^\prime$ ($o = 1, \ldots, n$) as follows. If $(y_{1o}, \ldots, y_{so}) = 0$ put $\phi_{o,T}^\prime = 1$ or 0 according as $(x_{1o}, \ldots, x_{mo}) = 0$ or $(x_{1o}, \ldots, x_{mo}) > 0$. Otherwise, choose $\eta_o, \lambda_1, \ldots, \lambda_n, \hat{\lambda}_1, \ldots, \hat{\lambda}_{\hat{n}}$ to maximise $\eta_o$ subject to

$$\sum_{j=1}^{n} x_{ij} \lambda_j + \sum_{j=1}^{\hat{n}} \hat{x}_{ij} \hat{\lambda}_j \leq x_{io}, \quad (i = 1, \ldots, m) \quad (9)$$

$$\sum_{j=1}^{n} y_{rj} \lambda_j + \sum_{j=1}^{\hat{n}} y_{rj} \hat{\lambda}_j \geq y_{ro} \eta_o, \quad (r = 1, \ldots, s) \quad (10)$$

$$\sum_{j=1}^{n} \lambda_j + \sum_{j=1}^{\hat{n}} \hat{\lambda}_j \leq 1, \quad (11)$$

and

$$\lambda_j \geq 0 \quad (j = 1, \ldots, n),$$

$$\hat{\lambda}_j \geq 0 \quad (j = 1, \ldots, \hat{n}), \quad (12)$$

and put $\phi_{o,T}^\prime = \eta_o$.

This model and the input-oriented diversification-consistent model are not generally diversification consistent if they include no notional units (Lamb and Tee, 2011): that is, its efficient frontier does not take into account the possibility of improving performance by diversification. So we use the iterative procedure of Fig. 1 to add notional units and approximate a diversification-consistent model arbitrarily accurately provided the risk and return measures are convexity consistent (see Section 3.2 and Lamb and Tee (2011)).

Lamb and Tee (2011) suggest choosing $\alpha = 0.5$. However, here we choose $\alpha$ using 10 iterations of golden section search (see Winston and Albright (2001)) because we find it gives more accurate approximations without slowing the iterative process too much. We need many computations and to solve large numbers of linear programmes during the iterative procedure. We coded C++ programmes and libraries for the computation and use cplex 12.1 (IBM, 2010) to solve the linear programmes. The code is available from http://www.abdn.ac.uk/~cms127/code.html together with documentation. We carried out all the computations on a gnu/Linux operating system with a 2.4GHz Intel quad core processor. We stop the iterative procedure when the average reduction in efficiency over ten improving iterations falls below $10^{-4}$. In practice, it takes about two hours to calculate efficiencies from 2000 bootstrap replications.

**References**


Banker, R. D., 1993. Maximum likelihood, consistency


