DISCUSSION PAPER SERIES

No. 3265

FACTOR BASED INDEX TRACKING

Francesco Corielli and
Massimiliano Marcellino

FINANCIAL ECONOMICS

Centre for Economic Policy Research

www.cepr.org

Available online at: **www.cepr.org/pubs/dps/DP3265.asp**

ISSN 0265-8003

FACTOR BASED INDEX TRACKING

Francesco Corielli, Università Bocconi
Massimiliano Marcellino, Università Bocconi, IGIER and CEPR

Discussion Paper No. 3265
March 2002

Centre for Economic Policy Research
90–98 Goswell Rd, London EC1V 7RR, UK
Tel: (44 20) 7878 2900, Fax: (44 20) 7878 2999
Email: cepr@cepr.org, Website: www.cepr.org

CEPR Discussion Paper No. 3265

March 2002

ABSTRACT

Factor Based Index Tracking*

Index tracking requires building a portfolio of stocks (a replica) whose behaviour is as close as possible to that of a given stock index. Typically, much fewer stocks should appear in the replica than in the index, and there should be no low-frequency (persistent) components in the tracking error. Unfortunately, the latter property is not satisfied by many commonly used methods for index tracking. These are based on the in-sample minimization of a loss function, but do not take into account the dynamic properties of the index components. Instead, we represent the index components with a dynamic factor model, and develop a procedure that, in a first step, builds a replica that is driven by the same persistent factors as the index. In a second step, it is also possible to refine the replica so that it minimizes a loss function, as in the traditional approach. Both Monte Carlo simulations and an application to the EuroStoxx50 index provide substantial support for our approach.

JEL Classification: C43, C53 and G10
Keywords: factor models, index tracing, replica and stock index

Francesco Corielli
IMQ Universitá Bocconi
Viale Isonzo
20100 Milano
ITALY
Email: francesco.corielli@uni-bocconi.it

Massimiliano Marcellino
IGIER
Universitá Bocconi
Via Salasco, 5
20136 Milano
ITALY
Tel: (39 02) 5836 3327
Fax: (39 02) 5836 3302
Email: massimiliano.marcellino@uni-bocconi.it

* We are grateful to seminar participants at Bocconi University for useful comments. Marcellino thanks MURST for financial support.

Submitted 22 February 2002

1 Introduction

Stock index tracking underlies a big slice of the fund management industry world wide. There are funds whose explicit strategy is to replicate an index or a pool of indexes, as for example the Vanguard 500 index fund which, on May 2001, showed a total asset value of about 93 billions of dollars. More generally, the existence and use of benchmarks for performance valuation compels the manager, from time to time, to "index" her strategy. Moreover, convex strategies as, for instance, portfolio insurance strategies, are based on the dynamic replication of an option whose underlying is usually a market index. When a futures on the index does not exist, or its use is forbidden by the rules of the fund, the manager will implement the synthetic hedging policy using an index replication. Long/short (also called market neutral) strategies, where a trader takes a long position in an index and a short position in some subset of the same, also require the accurate replication of the underlying index. Basis trading between a stock index futures and the underlying is another example of the need for good index replication.

Index tracking is also relevant for the development and enforcement of regulation policies. For example, if a benchmarking policy is imposed to fund managers by law (e.g., in Italy) or by use, the degree of effective replicability of the chosen benchmarks has important influences on the fund manager choices and, by consequence, on the investors ability to evaluate the fund strategy and performance. A purely passive tracking policy, for instance, could imply relevant deviations from the benchmark when replication of the latter is difficult. On the other hand, small differences from the benchmark could imply, in case of easy replication, a relevant discretional behavior of the fund manager.

The purpose of this paper is to clarify some problems related to index replication, and to propose and test a statistical method for building a portfolio made of a limited number of stocks which, with good probability, will

track a given index.

We are interested in cases where a full replication of the index is not a feasible strategy for the fund manager. A full replication is feasible when the net asset value of the fund is relevant, the structure of the index is kept constant for long stretches of time, and no big inflows or outflows of money affect the fund. When these three conditions are not met, the manager must consider a partial replication strategy and address the problem of tracking error. This is the case, for example, when considering convex strategy funds, strategies based on variable weight allocation across a set of index funds, index funds replicating indexes based on partially illiquid markets, etc.

While the construction of a replicating strategy is a practically important problem, the literature about this topic is still quite undeveloped. A recent survey in Beasley et al. (2000) reports less than 15 papers specifically dedicated to the topic. Most of this literature assumes that the solution of the tracking problem lies in some simple tracking error measure minimization, and concentrates on efficient ways for implementing this minimization, under investment constraints and transaction costs. While these are important problems indeed, we believe other relevant issues should be considered *before* tracking error minimization.

In this paper we concentrate on the study of the relationships among the time dynamics of the price series for the securities contained in the original index, the index itself, and the replica. While the dynamics of most securities and securities indexes are characterized by trend-like low frequency behavior, the dynamics of an acceptable tracking error, i.e., of the difference between the index and the tracking portfolio, should be characterized by a trendless high frequency behavior.

The dynamic properties of the stocks composing the index are taken into consideration by Pope and Yadav (1994). They go beyond simple tracking error minimization, and suggest diagnostics based on the study of the au-

tocorrelation function of the tracking error. Alexander (1999) takes a step further and suggests a tracking procedure based on the study of cointegration between the index and the component series. This method could be extended by borrowing from the literature on cointegration of stock prices, see e.g. Pindyck and Rothemberg (1992). In fact, the main procedures suggested in the present paper are directly connected to those considered in the literature on common trends and cointegration, a particularly relevant reference for our work being Stock and Watson (1988).

On the other hand, we cannot simply apply multiple cointegration theory to our problem for several reasons. First, while in cointegration analysis no explicit relations are known to hold across the series, in the tracking case we *know* the structure of the index, and this is important information. Second, the cointegration literature focuses on an extreme form of non-stationarity in the series: integration. Other non stationary behavior, as for instance low frequency cycles, are also relevant for us. For this reason, we prefer to speak in general of "low frequency components", as characterized by a low frequency peaked periodogram, and do not make any distinction between the particular origin of this behavior. Third, the inclusion of some specific assets in the replica and the minimization of a given loss function can be also required.

An additional complication for index tracking is that the fund manager is not free to choose among functional forms depending, possibly in a non-linear way, on past and present values of the component stocks. The index replication must be a *tradable* asset and the only shape it can take is that of a portfolio, i.e., a linear combination of contemporaneous values for a limited number of securities. Hence, it is by no means obvious that a feasible solution exists for the replication problem under this constraint, no matter which tracking error measure is chosen. The existence and the properties of a solution depend on the assumptions on the dynamics of the securities

involved in the index, and on the weight structure of the index itself.

In this paper we consider the case where the prices evolve according to a linear dynamic factor model, see e.g. Stock and Watson (1998), Forni and Reichlin (1996, 1998), Forni et al. (2000). We distinguish between long term and short term factors. Long term factors are dynamic components of the prices which exhibit a low frequency dominated behavior. As a limiting case, these factors could exhibit integrated behavior. Short term factors exhibit high frequency behavior.

Our main objective is to build a tracking portfolio which excludes the long term/low frequency factors from the tracking error, using a small number of securities. Traditional procedures based on the minimization of some tracking error measure do not necessarily achieve this aim, even when it is possible.

We argue that a useful replication is possible only when few long term factors drive the evolution of all the prices of the components of the index, and the weights of these components in the full index and in the replicating portfolio are the same. This is a necessary condition in order to avoid the undesirable presence of low frequency or even integrated components in the tracking error.

It is possible to consider many examples where these conditions are not satisfied. The low frequency factors could be a complicated combination of leads and lags of the prices in the index, while we can only use contemporaneous values for the replication. Or the total number of low frequency factors could be larger than the number of stocks that we are willing to include in the replication. Hence, we also suggest methods to clarify the source of a poor tracking performance and develop possible remedies. Moreover, since a replication free of low frequency components in the error could be impossible to build with a small number of securities, our approach to index tracking endogenously warns the user when this problem comes up.

Index tracking is, by its nature, a multi-period dynamic problem. We must then distinguish between linear factor models based on returns and on price levels. Most factor models for security prices are linear in returns, see for example the APT class of models. Yet, static replicas of indexes, the most common method in which the fund manager buys and holds without rebalancing a portfolio of stocks, implicitly hypothesize linear factor models in price levels. We discuss the implications of these alternative assumptions, and suggest solutions in both cases. Also, though we only require the tracking error to have no low frequency components, additional constraints, e.g. minimization of a particular loss function, can be added.

The structure of the paper is as follows. In section 2 we formulate a linear factor model for prices, and discuss a necessary condition for the tracking error not to have low frequency components. In section 3 we suggest a similar model for returns. In section 4 we discuss in details a sequential procedure for building a replicating portfolio. In section 5 we report the results of a set of Monte Carlo experiments to evaluate the performance of our approach. In section 6 the method is applied to construct a replica of the EURO STOXX50 index. Finally, section 7 summarizes the main results and contains suggestions for further research.

2 Necessary conditions for index tracking

In this section we discuss necessary conditions for the construction of a parsimonious index, $\widetilde{I}_t$, whose implied tracking error shows limited low frequency components. The focus is not on the minimization of a particular tracking error summary measure (e.g., the sum of squares of return differences or of index levels), but on the procedure to follow in order to build a replica whose low frequency tracking error behavior is largely independent on the chosen summary measure. Here we work with index levels, the following section considers the case of returns.

Let us write the generic index I_t in matrix terms as:

$$I_t = w\, p_t, \tag{1}$$

where p_t is an $N \times 1$ vector of prices, w is a $1 \times N$ vector of known weights (numbers of shares), and $t = 1, ..., T$. The weights can themselves be time varying provided they are known for each time t but, for simplicity, we assume that they are constant.

The prices are assumed to evolve according to the factor model:

$$p_t = \Lambda f_t + e_t, \tag{2}$$

where f_t is an $r \times 1$ vector of factors, whose loadings are grouped in the $N \times r$ matrix Λ, and e_t is an $N \times 1$ vector of disturbances. The factors f and the disturbances e are allowed to be correlated in time, and e can also be correlated across prices. Hence, (2) is a dynamic approximate factor model, see Stock and Watson (1998) and Forni et al. (2000) for more details and precise conditions on the correlation structure of f and e.

The above model is common to many descriptions of asset price behavior and there are in principle an infinite number of ways to decompose p_t into factors. Usually, in finance the distinction between f_t and e_t is the distinction between "common factors" and "idiosyncratic factors". In this case the factors are constructed to yield residuals maximally uncorrelated across different stocks. These factors can be either observable economic variables (as in the APT model) or "synthetic" variables expressed as portfolios of the stocks themselves. There are, however, many other useful ways of interpreting and then identifying the decomposition implicit in the model (2).

For our purposes, the interesting distinction is that between "long term factors" and "short term factors". An approximate but empirically useful definition of "long term factor" is that of a series whose periodogram is

concentrated on low frequencies. By contrast a "short term factor" indicates a series whose periodogram peaks on the high frequencies.

It may well be possible, and indeed it seems to be the case in empirical studies, that the long term factors coincide with the common factors of the financial literature. This will be the case, for instance, if we believe that the common behavior across assets is induced by the influence of fundamentals and other slow varying economic variables, while idiosyncratic factors represent short term deviations from the common path of evolution of the market. It is quite reasonable, however, that low frequency idiosyncratic components exist, and it is also possible for common factors to show high frequency behavior.

Our interest is not in the economic interpretation of the factors, but in the assessment of the possibility to impute the long run dynamics of prices to a small number of driving series. Only when this is possible, a reliable replica of the index can be built. A detailed analysis of the economic interpretation of the factor representation should take into account its non-uniqueness. For our purposes, however, any decomposition which distinguishes between high and low frequency components is equivalent.

A similar interpretation of model (2) is provided by Stock and Watson (1988): p_t is a vector of N processes satisfying $N - r$ cointegration relations, and driven by r independent integrated stochastic processes f_t. We also include in f_t non integrated processes with a dominating low frequency component.

We must now define the replica of the index I_t. It is another index, $\widetilde{I}_t$, based on a choice of q out of the N stocks in the original index:

$$\widetilde{I}_t = \omega \, S \, p_t, \tag{3}$$

where ω is a $1 \times q$ vector of weights, $q \geq r$, and S is a $q \times N$ selection matrix, namely, a matrix that selects only q out of the N prices in p_t. Hence, $\widetilde{I}_t$ is made up of a subset of the prices in I_t, with different weights.

In practice, it is possible that the replica also contains shares not included in the index. For instance, these shares could be already existent in the trader portfolio and she could be unwilling to sell them while, at the same time, she could require her portfolio to track a specified index. The suggestions of this paper can be easily extended to this case. Moreover, we concentrate on static replicas, i.e., portfolios where the number of shares for each stock is kept constant. We do this for two reasons. First, the index itself is usually a constant weights portfolio (or, at least, weights are changed only in a predetermined way and on predetermined dates). Second, a time varying structure of the replica requires explicit consideration of transaction costs, something we want to avoid in this paper. However, in Section 3 we examine a case where a dynamic replica is required by the assumptions made on stock dynamics, and we suggest a solution to this problem.

Any definition of loss function useful for evaluating the tracking error $I_t - \widetilde{I}_t$ should give negative weight to the fact that the error process contains a low frequency component. Otherwise, the tracking error can be persistent and even diverge, and this, when the drift direction cannot be forecasted, is not a good property for a tracking portfolio.

The most widely used loss functions are of the type

$$g(\varepsilon, q) = \sum_{t=0}^{T} d(\varepsilon_t) + c(q), \tag{4}$$

where ε_t is either $I_t - \widetilde{I}_t$ or the difference of the returns, and $d(.)$ is typically the square function, the absolute value, the positive part of the difference, etc. The function $d(\varepsilon_t)$ is in general non increasing in the number of assets, q, in the replica $\widetilde{I}_t$. Yet, increasing q can be expensive for the investor. Hence, a penalty $c(q)$ is sometimes added to the loss function, as it is common in the construction of information criteria in the time series literature to penalize overparametrized models. Other tracking error measures, which cannot be written as a sum but are of relevance for portfolio replication, are based on

the supremum over time of the difference of the values of the indexes or of returns.

In order to find the optimal replication index $\widetilde{I}_t$, we could minimize the objective function $g(\varepsilon, q)$ with respect to q, S, and ω over a stretch of past data. This agrees with common practice but, in general, it is not enough to assure that low frequency components (f) do not appear in the tracking error. We want to develop a tracking procedure that avoids this problem.

Let us assume for the moment that the choice of assets represented in $\widetilde{I}_t$ is already made, i.e., q and the selection matrix S are fixed. We further suppose that Λ is known. We add the obvious restriction that, at some date $t = 0$, the index and the replica have the same value $\widetilde{I}_0 = I_0$.

Under these hypotheses, a simple result follows:

Proposition 1 *For the index and the replica to share the same factor structure the optimal weights must satisfy:*

$$\omega^* = w(\Lambda \vdots p_0)(S\Lambda \vdots Sp_0)^+ + h[I_q - (S\Lambda \vdots Sp_0)(S\Lambda \vdots Sp_0)^+], \tag{5}$$

where h is any $1 \times q$ vector, I_q is the $q \times q$ identity matrix and $(.)^+$ indicates a generalized (Moore-Penrose) inverse. If $[(S\Lambda \vdots Sp_0)'(S\Lambda \vdots Sp_0)]^{-1}$ exists, we can write it as:

$$(S\Lambda \vdots Sp_0)^+ = [(S\Lambda \vdots Sp_0)'(S\Lambda \vdots Sp_0)]^{-1}(S\Lambda \vdots Sp_0)'.$$

Proof. The common components of I_t and $\widetilde{I}_t$ are, respectively, $w\Lambda F_t$ and $\omega\, S\Lambda F_t$, while $I_0 = wp_0$ and $\widetilde{I}_0 = \omega\, Sp_0$. Hence, we can form the system $\omega(\, S\Lambda \vdots Sp_0) = w(\Lambda \vdots p_0)$. The general representation of the solution (if it exists) is ω^* in (5) (see e.g. Graybill (1983) ch. 7.3). ■

Thus, a set of linear constraints on ω^* are required to avoid the leakage of low frequency factors in $I_t - \widetilde{I}_t$. These constraints depend on the choice of the components for the replica (S), the weights of each stock in the original

index (w), and the loadings of the long term factors (Λ). Moreover, this condition is independent from the choice of the tracking error measure.

If $q = r + 1$ and $|S\Lambda \vdots Sp_0| \neq 0$, the generalized inverse simplifies to $(S\Lambda \vdots Sp_0)^{-1}$, and the solution is unique. The intuition underlying such a result is that in this case the factors can be expressed as a linear combination of all the q selected stocks.

Otherwise, in general, given a factor structure Λ, for some choices of S there can be infinite solutions (the q elements of the vector of weights ω^* can be all expressed as functions of $q - r - 1$ free parameters), while for other choices of S there may be no solution. This happens, for example, when the S matrix selects stocks whose behavior is unaffected by the factors f_t (i.e., the corresponding rows of Λ are equal to zero).

Notice also that if the factor model for the prices in (2) coincides with the common trend model in Stock and Watson (1988), i.e., if the factors are pure orthogonal random walks, then ω^* coincides with the cointegration vector for (I_t, Sp_t), when there exists cointegration.

In general, the choice of h gives the degrees of freedom necessary to optimize the replica with respect to a specific choice of the loss function $g(.)$. As an example, for a quadratic loss function, i.e., $g = (I - \widetilde{I})(I - \widetilde{I})'$, the optimal value of h, when unique, is given by:

$$h^* = -(ASPP'S'B' - IP'S'B')(BSPP'S'B')^{-1}, \tag{6}$$

where

$$\begin{aligned} A &= w(\Lambda \vdots p_0)(S\Lambda \vdots Sp_0)^+, \\ B &= [1 - (S\Lambda \vdots Sp_0)(S\Lambda \vdots Sp_0)^+]. \end{aligned}$$

This choice of h is adopted in the simulation experiments and in the empirical application below.

For comparison, it is instructive to derive the OLS solution to the tracking problem when the sum of squared errors are defined on index levels

unconditional to the constraints given in (5). The tracking error can be written as $\varepsilon_t = I_t - \widetilde{I}_t = (w - \omega\, S\,)p_t$, and the loss function to be minimized with respect to the $1 \times q$ vector ω (assuming S and q are known) is

$$(I - \widetilde{I})(I - \widetilde{I})^{'} = (w - \omega S\,)PP^{'}(w - \omega\, S\,)^{'}. \tag{7}$$

Adding the constraint $I_0 = \widetilde{I}_0$, the solution is

$$\begin{aligned} \omega_{RLS} &= w(PP^{'}S^{'}\vdots p_0)C^{+} + h^{*}[I - CC^{+}], \\ C &= (SPP^{'}S^{'}\vdots Sp_0). \end{aligned} \tag{8}$$

The main differences between ω^{*} in (5) and ω_{RLS} in (8) can be best appreciated if we set $q = r$ and disregard the condition $\widetilde{I}_0 = I_0$. In this case the (supposed unique) vector satisfying (5) is given by

$$\omega^{*} = w\Lambda(S\Lambda)^{-1}, \tag{9}$$

while the OLS solution simplifies to

$$\omega_{OLS} = w(\Lambda FF^{'}\Lambda^{'}S^{'} + ee^{'}S^{'})(S\Lambda FF^{'}\Lambda^{'}S^{'} + See^{'}S^{'})^{-1}. \tag{10}$$

The tracking error when using ω^{*} is given by

$$\varepsilon^{*} = (w - w\Lambda(S\Lambda)^{-1}S)e. \tag{11}$$

Since ω^{*} satisfies Proposition 1, ε^{*} does not dependent on F. This in general is not the case for the OLS weights, since $w\Lambda F$ is different from $\omega_{OLS}S\Lambda F$. Hence, even if the OLS weights could provide a better in-sample fit, their performance can deteriorate substantially out-of-sample.

There is a case where the OLS weights coincide asymptotically with the optimal weights ω^{*}. This happens in the common factor model in Stock and Watson (1988) where, as mentioned before, the factors are pure random walks. Actually, in this case ω_{OLS} behaves asymptotically as

$$\omega_{OLS} \approx w(\Lambda FF^{'}\Lambda^{'}S^{'})(S\Lambda FF^{'}\Lambda^{'}S^{'})^{-1} = w\Lambda(S\Lambda)^{-1} = \omega^{*}. \tag{12}$$

Finally, it is worth noting that for the construction of the OLS weights, we have considered a regression of index levels on price levels. Yet, frequently in practice the OLS regression is run between the returns of I and those of Sp, see e.g. Pope and Yadav (1994), Rohweder (1998), Wang (1999). The resulting weights are used as amounts to be invested in each stock at time 0 for the replica of the index. This is inconsistent with a linear factor model in prices. If the loadings of the factors on the stocks and the weights of the stocks in the index are constant, a linear relation can exist between the returns of the index and the returns of the stocks. However, the coefficients of this relation are not constant, hence the OLS estimate is meaningless. Intuitively, the returns of those stocks whose prices grew more than other stocks, weight progressively more on the index returns. The weights remain constant only if the stocks undergo identical percentage variations. We consider in more details issues related to modelling the returns in the next Section.

3 Modelling returns

In the financial literature it is usually assumed that returns rather than prices follow a linear factor model, see e.g. Chamberlain and Rotschild (1983). Similarly, replicas are often built to track the returns of the index rather than the index itself, see e.g. Bealsley et al (2000). In this section we consider the required modifications to the previous analysis to deal with this case.

Besides the OLS method mentioned before, another common approach to tracking index returns, e.g. Rudolf et al. (1999), is to choose the subset of stocks in the replica, J, according to some criterion (e.g. size, sector,

etc.), and then find the weights ω_j as the minimizers of

$$\sum_{t=0}^{T} d(R_{I_t} - \sum_{j \in J} \omega_j R_{p_{jt}}), \tag{13}$$

where R indicates the return and d is usually given by the square function or the absolute value.

An additional problem in this case is that there is no replica portfolio, containing a constant number of shares for each stock, such that $\sum_{j \in J} \omega_j R_{p_{jt}}$ in (13) can be equal to the index replica returns, $R_{\widetilde{I}_t}$. In fact, we can write the replica returns as

$$R_{\widetilde{I}_t} = \sum_{j \in J} \frac{w_j p_{jt-1}}{\widetilde{I}_{t-1}} R_{p_{jt}}, \tag{14}$$

so that, for $\sum \omega_j R_{p_{jt}}$ to be equal to $R_{\widetilde{I}_t}$, ω_j should be

$$\omega_{jt} = w_j \frac{\widetilde{I}_{t-1}}{p_{jt-1}}.$$

In practice, to keep ω_j constant we need to sell some shares of a given stock if it outperforms the index I_t, and buy if it underperforms. As a consequence, it can happen that, even when the replica and the index contain the *same* stocks, it is $\widetilde{I}_t \neq I_t$, independently of the choice of d. Moreover, the variability of the weights and the connected transaction costs become non negligible when, as in the recent past, subsets of stocks are highly favored, or left behind, for rather long periods by the market.

This drawback can be overcome by a proper reformulation of the loss function as

$$\sum_{t=0}^{T} d(R_{I_t} - \sum_{j \in J} \frac{\omega_j p_{jt-1}}{\widetilde{I}_{t-1}} R_{p_{jt}}), \tag{15}$$

which leads to a nonlinear optimization problem. Here the loss function is defined in terms of the returns, but the ω_j are the number of shares for each stock in the constant weights replica portfolio.

This rather obvious problem extends to the case where we require a replica portfolio to share the factor structure of the index. Suppose that stock returns follow a linear factor model:

$$R_t = \Lambda f_t + e_t. \tag{16}$$

Proposition 2 *For the index and the replica to share the same factor structure the optimal weights must satisfy:*

$$\begin{aligned} \widetilde{\omega}_t &= \left[B_t \vdots 0\right] C_t^+ + h\left[I - C_t C_t^+\right], \\ C_t &= \left[SA_t \vdots p_t - p_{t-1}\right] \end{aligned} \tag{17}$$

where $^+$ *indicates a generalized inverse,* h *is an arbitrary* $1 \times q$ *vector (which may depend on* t*), and* A_t *and* B_t *are defined below.*

Proof. In this case prices follow a non linear factor model. Building a constant weights replica portfolio such that the returns of the replica and those of the index share the same factor structure is equivalent to solving the following problem:

$$\begin{aligned} I_{t-1} \sum_{j \in J} \omega_j p_{j,t-1} \lambda_{j,l} &= \widetilde{I}_{t-1} \sum_j w_j p_{j,t-1} \lambda_{j,l}, \qquad l = 1, ..., r, \\ \omega_j &= 0, \qquad \forall j \notin J, \end{aligned} \tag{18}$$

where the $\lambda_{j,l}$ are elements of Λ in (16). This system cannot be solved, in general, if the weights ω_j are not allowed to be time varying. If the weights are allowed to be time varying the problem can be written as:

$$\omega_t \left[SA_t \vdots p_t - p_{t-1}\right] = \left[B_t \vdots 0\right], \tag{19}$$

where

$$A_t = \left\{ \begin{array}{cccc} p_{1,t-1}\lambda_{1,1} & p_{1,t-1}\lambda_{1,2} & ... & p_{1,t-1}\lambda_{1,r} \\ p_{2,t-1}\lambda_{2,1} & p_{2,t-1}\lambda_{2,2} & ... & p_{2,t-1}\lambda_{2,r} \\ ... & ... & & ... \\ p_{N,t-1}\lambda_{N,1} & p_{N,t-1}\lambda_{N,2} & ... & p_{N,t-1}\lambda_{N,r} \end{array} \right\},$$

$$B_t = \frac{\widetilde{I}_{t-1}}{I_{t-1}} \left\{ \begin{array}{c} \sum_j w_j p_{j,t-1} \lambda_{j,1} \\ \sum_j w_j p_{j,t-1} \lambda_{j,2} \\ \ldots \\ \sum_j w_j p_{j,t-1} \lambda_{j,r} \end{array} \right\}, \quad \omega_t = \left\{ \begin{array}{c} \omega_{1t} \\ \ldots \\ \omega_{qt} \end{array} \right\}.$$

S is the usual selection matrix and the second part of the partitioned system imposes the self financing property for the tracking portfolio. Then, the solution of the system is $\widetilde{\omega}_t$ in (17). ■

The solution to the problem is not so satisfactory as in the previous section. In fact, the tracking error on returns is:

$$\widetilde{\varepsilon}_t = \frac{I_t}{I_{t-1}} - \frac{\widetilde{I}_t}{\widetilde{I}_{t-1}} = \frac{\sum_j w_j p_{j,t-1} e_{j,t}}{I_{t-1}} - \frac{\sum_{j \in J} \widetilde{\omega}_{jt} p_{j,t-1} e_{j,t}}{\widetilde{I}_{t-1}}. \qquad (20)$$

Thus, even when $\widetilde{\omega}_{jt}$ solves the linear problem, long term factors may affect, in a multiplicative way, the tracking error trough p_t and $\widetilde{\omega}_{jt}$. This implies a (known but difficult to compute) low frequency time evolution for the *variance* of the tracking error. Moreover, with a time varying portfolio, transaction costs can no longer be assumed negligible and their minimization should be considered among the asset manager objectives. We intend to pursue these topics in further research, while in the remaining part of the paper we focus on the construction of a buy and hold tracking portfolio, that is on the case of a linear factor model for prices.

4 Implementing index tracking

In order to implement the procedure described in the previous Sections, two main practical issues must be addressed.

First, it is necessary to estimate the factor model for the prices, equation (2). The estimated loading matrix, $\widehat{\Lambda}$, can then be used to construct ω^* in (5). In the first subsection, we discuss three versions of a method based on a slight modification of principal components analysis.

Second, we have to choose how many and which stocks to be used in the replication. In the second subsection, we suggest a procedure to select the component stocks on the basis of their ability in reconstructing the estimated factors, with a predetermined error margin. While we do not need to reconstruct the estimated factors in order to satisfy Proposition 1, a selection of the set of stocks to be included in the replica based on "factor tracking" seems quite sensible.

The problem of stock selection can be complicated by the need to include in, or exclude from, the replica portfolio a set of stocks. Often the maximum and minimum value of the investment in each single stock are also given and, for some stock, even the exact amount to be purchased is a constraint. All these requirements can be smoothly added to the procedure we suggest by choosing, if it exists, a proper h in (5). In particular, since in general it is $q > r + 1$, we will use the degrees of freedom in h to minimize a given tracking error measure, g.

Let us now consider in more details each step of the procedure for index tracking.

4.1 Estimation of the factor structure

We suggest three methods to estimate F and Λ based on a principal component analysis of the matrix of historical prices P. They differ in the transformation of P used for the analysis. These methods are based on those suggested by Stock and Watson (1988, 1998) (SW henceforth), see also Bai (2001a, 2001b). In particular, given the number of factors, r, and assuming stationarity of the modelled series, SW showed that the space spanned by the factors can be consistently estimated by the principal components of the original variables. Bai (2001b) proved that the result remains valid for integrated series.

Our aim is to decompose the stock prices movement into orthogonal com-

ponents ordered by their contribution to the *low* frequency variance of the prices (the factors F). Principal components rank orthogonal constrained linear combination of prices in terms of total variance, disregarding the distribution of this variance across frequencies. Yet, when studying stock prices it is reasonable to hypothesize that, at least for the first components, the bulk of the variance is concentrated on the low frequencies. This is consistent with the idea of "trends" in the market and is sometimes expressed by assessing that prices are integrated. In this case, the first principal components provide a good approximation to the low frequency dominating factors.

A greater care should be given to the estimation of lower order factors. Now it is no longer reasonable to assume that low frequency components dominate linear combinations variances. To address this issue, we suggest two modifications of the principal components analysis outlined above.

1) If we want to identify as factors the linear combinations of p_t that contribute most to the low frequency component of the series variance, and if we roughly measure this using the covariance between p_t and p_{t-1}, a proper procedure is to use as estimators the eigenvectors associated with the largest eigenvalues of $T^{1/2}PP'_{-1}$ instead of $T^{1/2}PP'$, where P_{-1} is the $T \times N$ matrix of lagged prices. While traditional principal components maximize the variance of each orthogonal component, given that the sum of squared coefficient is 1, in this case we maximize the first order autocovariance of the component series.

2) A second method to enhance the low frequency content of principal components is to filter the prices P with a low band pass filter in order to remove high frequency components before extracting classical principal components.

Having estimated a proper factor space, the second step is the estimation of the factor loadings on the original variables. This can be achieved by an OLS regression of the variables on the estimated factors. Note that Λ and

f in (2) are not identified unless additional restrictions are imposed, e.g., $f'f/T = I$, which explains why the principal components are only consistent for the space spanned by f. While the lack of identification can be an important problem for structural analysis, it is not particularly relevant in our context. Actually, even if the common component of the factor model (2) is rewritten as $\Lambda K K^{-1} f_t$, where K is a generic $r \times r$ matrix with full rank, the expression for ω^* in (5) does not change.

For the choice of the number of factors, r, SW suggested to start with a large enough value, and then use a particular information criterion to select the number of factors in an equation of interest. As long as the assumed number of factors is larger than the true one, consistency of the principal components is preserved. Bai and Ng (2000) proposed a multivariate information criterion to determine r, which seems to perform quite well when the sample size is long enough. Forni *et al.* (2000), on the other hand, suggest to include as many factors as necessary to explain a fixed percentage of the variability of p_t, say 90%. On the basis of the simulation experiments in Section 5, the latter method seems to work better in the context of index tracking, as we will discuss in more details below.

Finally, notice that the framework developed by Forni et al. (2000), where the factors are estimated using Brillinger's (1981) dynamic principal components, is not suited in our context, since the factors are combinations not only of the contemporaneous values of the series but also of their leads and lags.

4.2 Reconstruction of the factor structure

We want to replicate the index using a limited number of stocks and we must choose them. We could select the component stocks by minimizing the objective function $g(\varepsilon, q)$ in (4), subject to the constraint that the solution satisfies Proposition 1. However this approach becomes quickly computa-

tionally cumbersome, as the solution should be found in a large, partially discrete, space. Hence, we suggest a stepwise ad hoc procedure that gives good results both with real data and in simulations.

As mentioned above, for the construction of the optimal weights ω^* we do not need an exact reconstruction of low frequency factors based on a limited number of stocks. Yet, a choice of the stocks based on their factor replicating ability seems a sensible strategy. For example, when the total number of stocks to be included in the replica is fixed and small, it is important to exclude those stock which are not influenced by the factors. Otherwise, as we mentioned, there could be no solution to the index tracking problem such as to avoid the presence of persistent components in the tracking error.

To implement this idea we start by ranking the estimated factors according to their correlation with the index. We then proceed to reconstruct all the reordered factors using the stocks in the index, or in a restricted or expanded set of stocks available to the fund manager. Each factor is reconstructed up to a predefined residual variance.

The outcome of this procedure is a ranking of the available stocks based on their ability in replicating the estimated factors. The choice of the stocks in which to invest is then made according to how many factors are included in the replica. We assume their number to be equal to that selected for modelling all the stocks in the index.

The steps to be followed in this procedure are:

1. Order the factors according to their correlation with the index.
2. Choose a minimum R^2 for the replication of each factor.
3. Start from factor 1.
4. Rank the shares in correlation order with the factor.
5. Regress the factor on the first share.

6. If the R^2 of the regression is greater then the objective, skip next step.

7. Add to the regression the share with the highest correlation with the residual and go to step 6).

8. Regress the next factor on all the variables included in the analysis up to now then go to step 6).

9. If all the desired factors are replicated with the desired accuracy, stop

Finally, since any choice of h results in a portfolio satisfying the requirements of Proposition 1, we are free to minimize with respect to h the tracking error measure g best suited for our purposes.

5 Monte Carlo experiments

In this section we run a set of simulation experiments to compare the performance of OLS on returns (e.g. Pope and Yadav (1994), Rohweder (1998), Wang (1999)) and factor based tracking, and evaluate the relative merits of alternative methods for factor estimation.

OLS on returns is selected as a benchmark because stratified sampling by firm characteristics, the other popular method for the construction of tracking portfolios, is not suited for an "automatic" implementation. Moreover, in the OLS regressions we can use the same variables chosen for factor based tracking, so that the two approaches only differ for the choice of the weights of the stocks in the replica.

The N price series are generated according to the following factor model:

$$p_t = \Lambda_1 f_{1t} + \Lambda_2 f_{2t} + e_t, \tag{21}$$

where f_{1t} and f_{2t} are, respectively, r_1 and r_2 integrated and stationary factors, while e_t is an idiosyncratic i.i.d. standard normal error. More precisely, f_{2t} are i.i.d. standard normal, while f_{1t} are pure independent

random walks, driven by i.i.d. standard normal innovations. The elements of the loading matrices Λ_1 and Λ_2 are independent draws from a uniform distribution over the interval zero-one.

This data generating process is quite extreme in its behavior as it only considers integrated and i.i.d. components. However, such a specification already highlights all the relevant characteristics of the tracking procedure we suggest. Notice also that in this context OLS on levels would yield asymptotically the same results as factor based tracking, while this would not be the case if f_{2t} were persistent.

As detailed in Section 3, a linear factor model on levels is necessary for a constant coefficient replication to be possible. However, as mentioned in Section 2, this data generating model implies that a constant coefficient replication based on OLS on returns is a misspecified solution to the tracking problem. More precisely, the OLS on returns replication will work reasonably well when the simulated series, by chance, do not show any relevant trending behavior. This is more probable in simulations where the number of integrated factors is small.

To mimic the values in the empirical application in the next section, we set $N = 50$ and $T = 1000$, where T is the sample size. We consider three possible factor structures: $r_1 = 5$ and $r_2 = 5$, $r_1 = 2$ and $r_2 = 8$, $r_1 = 8$ and $r_2 = 2$. The index to be replicated is then constructed as an average of all the N prices with equal weights.

Following subsection 4.1, we consider three alternative estimators for the factors. In the base case, no transformation is applied to the price series and the factors are extracted from their variance covariance matrix (BASE). Then we experiment with extracting the factors using the first lag of the autocovariance matrix (AC), and with transforming the series with a low band pass filter (BPF), a one-sided moving average of length 50 with equal weights. Different lengths and decreasing weights do not alter sensibly

the results.

The number of factors is determined so that at least 99.9% of the variability of the series is explained when using the low band pass filtered data, in order to focus on the low frequency variability. We also experimented with the Bai and Ng (2000) selection criteria, and with the Bartlett (see Anderson (1963)) and Kaiser (1960) tests. These methods often selected a larger number of factors. This is because in a sample as long as ours, they can correctly identify as different from zero even very small eigenvalues. Yet, the contribution of the associated eigenvectors to explaining the variability of the series is so small that it is offset by the cost of having to include a larger number of variables in the replica in order to mimic the factor structure of the index.

The variables to be included in the index replica are selected according to the procedure described in subsection 4.2, with the value for the minimum R^2 in step 2 set at 0.80. Three different sets of variables are selected for each of the three sets of estimated factors. The factor based index replication is then built, using the formulae in section 2. The same variables are used for the OLS on returns replications.

The model is estimated over the period $1-500$, and the performance of the replica indices are evaluated in sample and out of sample, over the periods $501-1000$, $501-750$, $751-1000$. An evaluation over such a long out of sample period is required by the use of high frequency data in practice, combined with the need of the fund manager to keep the weights in the replica constant over long periods of time to reduce transaction costs.

Four loss functions are used to compare the performance of the alternative index replications: the mean tracking error (MEAN), the standard deviation (STD.DEV.), the mean absolute deviation (MAD), and the supremum of the absolute value of the errors (SUPMOD).

Table 1 reports the average value of the loss functions over 5000 replica-

tions for each method, with Monte Carlo standard errors, for the case $r_1 = 5$, $r_2 = 5$. Six comments are in order. First, the factor based replications work better than OLS both in-sample and out of sample, according to any criterion, and the gains are substantial. Second, among the factor methods, BASE works best, AC is a close second best, while BPF performs worst. Third, the standard error for the OLS on returns cases are much larger than those for factor based tracking. This is consistent with the fact that this model is misspecified for the data generating process we use. Fourth, the tracking performance deteriorates with the forecast horizon, less so for the factor based replications. Fifth, the selected number of factors is on average equal to 4, which is close to the number of non-stationary factors in the data generating process. Finally, the number of variables included in the replica, selected in order to match as closely as possible the factor structure of the index, is equal on average to 8 for BASE and AC, and to 4 for BPF (the same variables are used in the OLS on returns replications). The fact that fewer variables are included in the replica index when using BPF is the most plausible reason for the worse performance of this method.

To evaluate the relative role of stationary and integrated factors, Tables 2 and 3 report results for, respectively, the cases $r_1 = 2$, $r_2 = 8$ and $r_1 = 8$, $r_2 = 2$. It is worth making four comments on them. First, the performance of the OLS methods improves with a lower number of integrated factors, while that of the factor methods deteriorates substantially. Actually, when $r_1 = 2$, $r_2 = 8$ the average gains from the factor methods are only about 10%. Moreover, in general, the standard error of the results decreases for the OLS on returns methods but increases for the factors methods, with comparable values when $r_1 = 2$, $r_2 = 8$. These facts happen because a smaller number of integrated factors implies, in our data generating process, a smaller long term variance and a higher probability of trendless simulated series. Second, among the factor methods, the AC performs slightly better than BASE, but

the differences are very small. Third, on average, the selected number of factors is equal to 2 when $r_1 = 2$ and to 6 when $r_1 = 8$. This is in line with the more important role of the integrated factors in explaining the variability of the series. Finally, the number of variables in the replica index is about 6 for $r_1 = 2$ and 11 for $r_1 = 8$, which reflects the higher number of factors in the latter case.

As a check on the robustness of the results we got, using $r_1 = 5$ and $r_2 = 5$, we conduct three additional experiments. First, we increase the number of variables from 50 to 100. Second, we compare the performance of the methods using the median rather than the mean over the simulations. Third, we fix the number of factors to the true value of 10.

The figures in Table 4 indicate that a larger number of variables does not alter the results. The ranking of the methods and the size of the gains remain basically the same as for $N = 50$; the same number of factors is selected; the only minor difference is that now a slightly larger number of variables is included in the replica index, about 10 for BASE and 13 for AC, versus 8 when $N = 50$.

When the median rather than the average over the 5000 replications is used to compare the methods, the performance of the OLS on returns based replications improves, but it still remains substantially worse than that of the factor based replications, using any criterion and both in sample and out of sample, see Table 5.

When the true number of factors is imposed, Table 6 shows that the gap between the factor and the OLS on returns based replications widens substantially. The former provides a sensibly more accurate tracking while the performance of the latter deteriorates with respect to the case where the number of factors is determined with the percentage of explained variance criterion. The cost of the improved factor based tracking is the much higher number of variables to be included on average in the replication: 36 out of

50 for the best factor method (BASE) versus 8 before. Since the aim of the replication is to reduce substantially the number of stocks in the portfolio, this major increase in the number of selected variables more than offsets the benefits in terms of reduced loss. The worse performance of the OLS on returns method is likely due to collinearity problems because of the large number of regressors.

In summary, factor based index tracking performs better than conventional methods, with the magnitude of the gains increasing with the number of non-stationary factors driving the variables. Extracting the factors from the variance covariance matrix of the raw data is in general performing well, with only minor gains in a few cases from using the autocovariance matrix. The selected number of factors using our approach is smaller than the true value, but close to the number of integrated factors and capable of explaining more than 99.9% of the variability of all the series in the index. Finally, the number of variables in the replica is substantially smaller than that in the index, usually only slightly larger than twice the number of selected factors.

6 Tracking the EURO STOXX50

In this section we evaluate the performance of our factor based methods for tracking the EURO STOXX50 index, relative to the common OLS on returns and to OLS on prices. We use a daily dataset from Jan 1997 to Jun 2000, for a total of 890 observations. This period contains several problematic events, including the Asian crisis of summer 1998 and the rise and early fall of the tech market bubble, which makes the analysis particularly interesting.

The dataset contains all the component stocks which were in the index during the sample period. In the spirit of this paper, we recomputed the index so that the weights of each stock are constant over the whole period, and the component stocks are the same.[1] We split the sample so that the

[1] During the sample period stocks were added to and deleted from the index (see

replica portfolio weights are estimated using the first set of 350 observations, while index tracking is carried out on the sample 351-890.

The procedures applied for factor estimation, reconstruction using a subset of the variables, and computation of the weights in the replica are the same as those described in the previous Section. In particular, the number of factors is selected so that the total variance explained is at least 99.8%, and we consider the three cases BASE (raw data, factors extracted from variance matrix), AC (raw data, factors extracted from first lag of autocovariance matrix) and BPF (smoothed data, factors extracted from variance matrix). The comparison is made with the OLS on returns and on price levels replications.

In Table 7 we report the percentages of variance explained by the eigenvectors associated with the largest 10 eigenvalues. The use of the autocovariance matrix does not change significantly the relative weights of the factors, but the prefiltering of the series has an interesting impact. Actually, while the first two values are larger and the third one is of comparable magnitude, the other figures are substantially smaller. A reasonable interpretation of this result is that the variance of the first three factors is relatively more concentrated on the low frequencies than the variance of the other factors, hence it is less affected by the low band pass filters. This interpretation is supported by the outcome of standard Dickey-Fuller tests, the null hypothesis of a unit root is not rejected only for the first three factors.[2]

According to our selection criterion, 6 factors are required. The perfor-

http://www.stoxx.com/index.html for a detailed description of the index computing procedures). We decided to compute the new index using the full set of these stocks. For this reason our index contains 54 stocks instead of 50 as the actual one. As stated above, we could have used the actual weights, since weight modifications are announced in advance of their implementation, with additional computational costs.

[2]Bai and Ng (2001) prove that the limiting distribution of the unit root test is not affected by factor estimation. They estimate the integrated factors by cumulating their first differences while we directly estimate the levels, but this fact does not alter the result.

mance of the resulting index replications is summarized in Table 8, while in Figure 1 we plot the out of sample values of the index, the BASE factor based and the OLS replications, that are the best performing in their class.

From Table 8, depending on the factor extraction method, the number of variables chosen to mimic the factor structure of the index ranges from 6 (BPF) to 21 (AC). When the same variables as in the BPF case are used for the OLS replications, (Naive1 and Level1), the factor method yields the lowest std.dev. and mad both in sample and in the two forecast subsamples, while looking at the whole forecast sample Level1 is better. In all cases, the OLS on returns, the most common method in practice, is by far the worst.

Focusing instead on the overall best performing factor method, BASE, in-sample the ranking based on either the std.dev. or the mad is OLS on levels, factor, and OLS on returns. Yet, out of sample, the factor method becomes the best, in all subperiods and according to any criterion. The OLS on returns remains the worst tracking procedure.

The performance of all methods is in general substantially better over the first forecast subsample, 351-650, than over the second one, 651-890. Though a certain deterioration is expected, since we are forecasting further ahead in the future, another cause for this temporal pattern is the upward trend in the index over most of the second subsample, compare Figure 1. The Figure particularly highlights the bad tracking performance of OLS on returns. Actually, as we mentioned earlier, the problems of OLS on returns are exacerbated when the index presents a trending behavior, or persistent deviations from the mean (as it happens at the beginning of the first forecast subsample, in coincidence with the Asian crisis).

In order to give an idea of the absolute average quality of the replicating portfolio, since the mad using 6 series in the replica is 16.05 for the BPF factors and 23.68 for the corresponding OLS on returns, while the average of the index is about 234 out of sample, the factor replica shows an error

of about 6.86% of the average value of the index, about 10.12% for the commonly used OLS on returns. The corresponding values with 19 variables and BASE tracking are 1.63% and 5.84%.

In summary, the good performance of the factor based index tracking that emerged in simulation experiments is confirmed also with real data, in particular when the index shows a trending or highly persistent behavior.

7 Conclusions

In this paper we have proposed a statistical method for building tracking portfolios. An interesting property of our procedure is that it can take into account the constraints commonly imposed to fund managers. Moreover, it does not require the index weights to be constant and can be applied to a wider universe of securities than those included in the index.

The starting point is a detailed analysis of the component stocks which, combined with a proper choice of the replica weights, can avoid tracking errors contaminated by low frequency components. We have analyzed in details the cases of a linear factor model for prices and for returns. In the second setting, however, a tracking portfolio with constant weights cannot be built. This highlights the importance of dynamic hedging and of accounting for transaction costs, whose examination is left for further research.

Our procedure is tested against competing methods by means of simulation experiments and an application to the well known EURO STOXX50 index. The results are quite encouraging, and emphasize the importance of a statistical approach to index tracking.

References

[1] Alexander, C. (1999), "Optimal hedging using cointegration", *Philosophical Transactions of the Royal Society of London, Series A - Mathematical Physical and Engineering Sciences*, 357, 2039-2058.

[2] Anderson, T. W. (1963), "The asymptothic theory for principal component analysis", *Annals of Mathematical Statistics*, 343, 122-148.

[3] Bai, J. (2001a), "Inference on factor models of large dimension", *mimeo*, Boston College.

[4] Bai, J. (2001b), "Estimating cross-section common stochastic trends in non-stationary panel data", *mimeo*, Boston College.

[5] Bai, J. and S. Ng (2000), "Determining the Number of Factors in Approximate Factor Models", *Econometrica*, (forthcoming).

[6] Bai, J. and S. Ng (2001), "A panic attack on unit roots and cointegration", *mimeo*, Boston College

[7] Beasley, J.E., Meade, N., Chang, T. J.(2000),"Index Tracking", mimeo, Imperial College, London.

[8] Brillinger, D.R. (1981), *Time series analysis: Data analysis and theory*, New York: Holt, Rinehart and Winston.

[9] Chamberlain, G. and M. Rothschild (1983), "Arbitrage factor structure, and mean variance analysis of large asset markets", *Econometrica*, 51, 1281-1304.

[10] Clements, M. and D.F. Hendry (1999), *Forecasting Non-Stationary Economic Time Series*. Cambridge (MA):.MIT Press.

[11] Forni, M. and L. Reichlin (1996), "Dynamic common factors in large cross-sections", *Empirical Economics*, 21, 27-42.

[12] Forni, M. and L. Reichlin (1998), "Let's get real: A dynamic factor analytical approach to disaggregated business cycle", *Review of Economic Studies*, 65, 453-474.

[13] Forni, M., Hallin, M., Lippi, M. and L. Reichlin (2000), "The generalised factor model: identification and estimation", *The Review of Economic and Statistics*, 82, 540-554.

[14] Graybill, F.A. (1983), *Matrices with applications in statistics*, Wadsworth.

[15] Kaiser, H. F. (1960), "The application of electronic computers to factor analysis", *Educational and Psycological Measurement*, 2, 141-151.

[16] Pindyck, R.S. and J.J. Rotemberg (1992), "The comovement of stock prices", *Quarterly Journal of Economics*, 108, 1073-1103.

[17] Pope, P.F., and P.K. Yadav (1994), "Discovering errors in tracking error", *The Journal of Portfolio Management*, 20, 27-32.

[18] Rohweder, H.C. (1998), "Implementing stock selection ideas: does tracking error optimization do any good?", *The Journal of Portfolio Management*, 24, 49-59.

[19] Rudolf, M., Wolter, H.-J., and H. Zimmermann (1999), "A linear model for tracking error minimization", *Journal of Banking & Finance*, 23, 85-103.

[20] Stock, J. H., and M. W. Watson (1988), "Testing for common trends", *Journal of the American Statistical Association*, 83, 1097-1107.

[21] Stock, J.H. and M.W. Watson (1998), "Diffusion Indexes, NBER Working Paper #6702.

[22] Wang, M.Y. (1999), "Multiple benchmark and multiple portfolio optimization", *Financial Analyst Journal*, 55, 63-72.

Table 1 - Monte Carlo comparison of Factor and OLS on retuns based index replication (N=50, r1=5, r2=5)

	Naive 1	*Naive 2*	*Naive 3*	*Factor 1*	*Factor 2*	*Factor 3*
			In sample (1-500)			
mean	20.65	2.50	1.43	1.25	0.63	0.54
	4.53	15.90	16.60	0.05	0.03	0.03
std.dev.	49.91	46.29	49.14	7.05	4.28	4.19
	4.73	7.13	7.59	0.05	0.04	0.04
mad	41.45	38.47	40.69	5.68	3.44	3.36
	3.93	5.87	6.23	0.04	0.03	0.03
supmod	163.51	165.56	181.15	22.39	13.64	13.40
	14.35	32.84	35.83	0.15	0.13	0.12
ncomp				3.81		
				0.40		
nvar	3.89	8.38	7.54	3.89	8.38	7.54
	0.58	6.21	4.64	0.58	6.21	4.64
			Out of sample (501-1000)			
mean	-3.26	33.99	27.05	-0.14	0.11	-0.04
	8.27	24.45	23.47	0.26	0.16	0.15
std.dev.	54.28	61.16	65.51	10.89	6.26	6.09
	4.64	15.34	15.80	0.10	0.08	0.07
mad	45.13	50.95	54.39	9.02	5.17	5.01
	3.75	12.70	13.05	0.09	0.07	0.06
supmod	186.80	213.87	221.59	38.85	22.58	22.04
	17.43	58.71	58.73	0.36	0.27	0.25
			Out of sample (501-750)			
mean	-0.45	21.16	19.00	-0.21	0.15	0.01
	7.15	13.93	13.41	0.19	0.12	0.11
std.dev.	39.52	40.68	44.29	8.28	4.97	4.89
	3.51	7.91	8.50	0.07	0.05	0.05
mad	32.69	34.14	37.22	6.80	4.06	3.98
	2.84	6.73	7.26	0.06	0.05	0.04
supmod	136.45	143.71	152.05	29.26	17.63	17.35
	14.47	31.24	32.13	0.25	0.19	0.18
			Out of sample (751-1000)			
mean	-6.08	46.83	35.10	-0.07	0.07	-0.09
	10.00	35.16	34.14	0.36	0.22	0.21
std.dev.	40.07	47.04	51.75	8.33	4.96	4.86
	4.11	13.19	13.59	0.07	0.05	0.05
mad	33.13	39.64	43.53	6.84	4.04	3.96
	3.20	11.36	11.67	0.06	0.05	0.04
supmod	180.92	209.28	215.09	37.52	21.78	21.24
	17.13	58.69	58.48	0.36	0.27	0.25

Notes

FactorX indicates the factor based replica using factors extracted from the variance covariance matrix of raw data (Base, X=3), of the first autocovariance matrix (AC, X=2), or from the variance covariance matrix of a 50 period moving average of the data (BPF, NaïveX indicates the OLS on returns based replica, using the same variables as in FactorX, X=1,2,3.

Mean is the mean error of replications, std.dev. the standard deviation, mad the mean absolute deviation, supmod the sup of the mo

Ncomp is the number of factors included in the factor model, nvar the number of variables included in the replica

The figures are averages over 5000 replications. Monte Carlo standard errors are reported in smaller fonts.

Table 2 - Monte Carlo comparison of Factor and OLS on returns based index replication (N=50, r1=2, r2=8)

	Naive 1	*Naive 2*	*Naive 3*	*Factor 1*	*Factor 2*	*Factor 3*
			In sample (1-500)			
mean	-7.85	23.23	25.01	-10.55	22.07	23.26
	20.88	11.81	11.94	20.84	11.80	11.93
std.dev.	75.56	57.54	59.25	71.92	54.86	55.85
	18.78	8.98	9.05	18.78	8.98	9.03
mad	62.27	48.14	49.53	59.30	45.95	46.77
	15.06	7.59	7.64	15.05	7.59	7.62
supmod	237.64	183.52	190.21	222.74	172.98	177.04
	54.10	27.98	28.33	54.06	27.97	28.21
ncomp				1.93		
				1.51		
nvar	2.48	7.07	5.73	2.48	7.07	5.73
	2.79	10.36	9.40	2.79	10.36	9.40
			Out of sample (501-1000)			
mean	-29.39	-4.95	-6.33	-29.65	-4.81	-5.35
	20.96	5.16	5.31	20.95	5.15	5.16
std.dev.	60.71	48.09	49.93	55.97	44.87	45.37
	12.29	5.23	5.35	12.27	5.23	5.25
mad	50.29	40.08	41.55	46.27	37.38	37.78
	10.05	4.38	4.46	10.04	4.37	4.39
supmod	213.27	162.08	168.26	197.82	151.07	152.58
	46.06	16.86	17.31	46.02	16.84	16.92
			Out of sample (501-750)			
mean	-16.96	-1.97	-3.37	-17.26	-1.59	-1.87
	15.14	3.80	4.01	15.12	3.78	3.79
std.dev.	50.03	34.92	36.08	46.97	32.62	33.11
	12.31	3.81	3.86	12.30	3.80	3.82
mad	41.05	28.98	29.91	38.47	27.05	27.45
	9.92	3.16	3.20	9.91	3.15	3.17
supmod	171.90	118.29	123.63	161.34	110.41	112.07
	43.82	12.09	12.53	43.79	12.06	12.15
			Out of sample (751-1000)			
mean	-41.82	-7.92	-9.30	-42.05	-8.04	-8.82
	27.34	7.64	7.74	27.32	7.63	7.65
std.dev.	42.56	36.48	38.46	39.07	34.20	34.74
	7.32	4.61	4.81	7.29	4.60	4.64
mad	34.93	30.26	31.93	31.94	28.34	28.77
	5.89	3.83	4.01	5.87	3.83	3.85
supmod	205.22	156.70	162.62	190.45	146.29	147.66
	43.92	16.42	16.88	43.88	16.41	16.49

Notes

FactorX indicates the factor based replica using factors extracted from the variance covariance matrix of raw data (Base, X=3), of the first autocovariance matrix (AC, X=2), or from the variance covariance matrix of a 50 period moving average of the data (BPF, NaïveX indicates the OLS on returns based replica, using the same variables as in FactorX, X=1,2,3.

Mean is the mean error of replications, std.dev. the standard deviation, mad the mean absolute deviation, supmod the sup of the mo

Ncomp is the number of factors included in the factor model, nvar the number of variables included in the replica

The figures are averages over 5000 replications. Monte Carlo standard errors are reported in smaller fonts.

Table 3 - Monte Carlo comparison of Factor and OLS on returns based index replication (N=50, r1=8, r2=2)

	Naive 1	*Naive 2*	*Naive 3*	*Factor 1*	*Factor 2*	*Factor 3*
			In sample (1-500)			
mean	22.65	52.32	34.91	0.51	0.22	0.22
	28.85	26.96	20.41	0.02	0.01	0.01
std.dev.	112.85	91.91	57.20	5.82	3.19	3.23
	43.92	29.66	12.00	0.03	0.03	0.03
mad	96.39	75.57	47.34	4.66	2.55	2.58
	38.94	24.06	9.82	0.03	0.02	0.02
supmod	340.30	304.81	195.17	18.30	10.26	10.33
	106.30	95.56	49.68	0.11	0.08	0.08
ncomp				5.57		
				0.76		
nvar	5.88	11.00	10.70	5.88	11.00	10.70
	0.93	5.95	5.60	0.93	5.95	5.60
			Out of sample (501-1000)			
mean	-107.04	-65.78	-1.74	-0.22	0.10	0.07
	98.86	64.68	3.36	0.27	0.14	0.14
std.dev.	101.96	106.19	72.72	11.66	5.62	5.65
	28.96	29.93	17.77	0.10	0.06	0.06
mad	81.35	86.62	59.20	9.72	4.64	4.66
	21.01	23.54	13.20	0.09	0.06	0.05
supmod	396.11	391.88	239.70	40.48	20.14	20.21
	149.01	132.41	56.11	0.34	0.23	0.22
			Out of sample (501-750)			
mean	-90.73	-52.64	9.50	-0.09	0.10	0.10
	87.14	64.27	10.98	0.19	0.10	0.10
std.dev.	87.68	93.78	57.07	8.57	4.38	4.42
	32.39	34.82	17.89	0.07	0.05	0.05
mad	71.06	79.36	47.56	7.09	3.59	3.63
	25.02	29.88	14.93	0.06	0.04	0.04
supmod	315.50	328.94	188.13	29.72	15.32	15.50
	138.56	131.83	55.94	0.24	0.16	0.16
			Out of sample (751-1000)			
mean	-123.34	-78.92	-12.98	-0.36	0.10	0.03
	110.73	66.71	10.17	0.38	0.20	0.20
std.dev.	72.16	72.85	50.79	8.58	4.35	4.38
	17.94	19.21	10.56	0.07	0.05	0.04
mad	59.87	60.32	42.60	7.10	3.56	3.58
	14.44	15.63	8.81	0.06	0.04	0.04
supmod	382.87	362.56	212.06	39.07	19.47	19.50
	148.50	124.96	35.10	0.34	0.22	0.22

Notes

FactorX indicates the factor based replica using factors extracted from the variance covariance matrix of raw data (Base, X=3), of the first autocovariance matrix (AC, X=2), or from the variance covariance matrix of a 50 period moving average of the data (BPF, NaïveX indicates the OLS on returns based replica, using the same variables as in FactorX, X=1,2,3.

Mean is the mean error of replications, std.dev. the standard deviation, mad the mean absolute deviation, supmod the sup of the mo

Ncomp is the number of factors included in the factor model, nvar the number of variables included in the replica

The figures are averages over 5000 replications. Monte Carlo standard errors are reported in smaller fonts.

Table 4 - Monte Carlo comparison of Factor and OLS on returns based index replication (N=100, r1=5, r2=

	Naive 1	*Naive 2*	*Naive 3*	*Factor 1*	*Factor 2*	*Factor 3*
			In sample (1-500)			
mean	7.43	24.49	16.88	1.40	0.74	0.64
	7.74	7.24	4.51	0.04	0.03	0.03
std.dev.	45.76	41.48	41.41	7.21	4.38	4.35
	3.66	4.60	3.06	0.05	0.05	0.04
mad	37.63	34.17	34.14	5.81	3.52	3.49
	2.90	3.73	2.47	0.04	0.04	0.03
supmod	155.90	142.75	141.56	22.83	13.99	13.93
	14.24	18.16	11.75	0.15	0.16	0.13
ncomp				3.83		
				0.39		
nvar	3.87	12.04	9.18	3.87	12.04	9.18
	0.51	12.90	9.75	0.51	12.90	9.75
			Out of sample (501-1000)			
mean	0.06	10.56	-5.58	-0.18	-0.08	-0.02
	6.91	7.74	6.18	0.25	0.16	0.16
std.dev.	52.42	51.83	46.16	11.15	6.46	6.30
	4.18	7.75	3.14	0.10	0.09	0.08
mad	43.77	43.55	38.43	9.25	5.33	5.18
	3.45	6.64	2.58	0.09	0.08	0.07
supmod	174.87	171.72	158.78	39.73	23.01	22.77
	14.10	23.56	12.09	0.34	0.30	0.27
			Out of sample (501-750)			
mean	0.62	5.71	-5.44	-0.02	-0.07	0.05
	5.28	3.87	6.01	0.18	0.12	0.11
std.dev.	39.19	37.64	35.35	8.39	5.02	4.98
	3.77	5.22	3.13	0.07	0.06	0.05
mad	32.73	31.38	29.47	6.88	4.10	4.05
	3.19	4.26	2.62	0.06	0.05	0.05
supmod	130.09	122.58	120.38	29.64	17.61	17.62
	12.45	14.71	11.40	0.23	0.22	0.19
			Out of sample (751-1000)			
mean	-0.51	15.42	-5.71	-0.33	-0.10	-0.08
	9.06	13.49	6.67	0.36	0.23	0.22
std.dev.	36.32	34.50	34.20	8.53	5.10	5.05
	2.61	4.02	2.47	0.07	0.06	0.05
mad	30.20	28.49	28.60	7.01	4.17	4.12
	2.17	3.21	2.10	0.06	0.05	0.05
supmod	168.64	165.18	153.01	38.42	22.25	22.00
	13.57	23.26	11.71	0.34	0.30	0.27

Notes

FactorX indicates the factor based replica using factors extracted from the variance covariance matrix of raw data (Base, X=3), of the first autocovariance matrix (AC, X=2), or from the variance covariance matrix of a 50 period moving average of the data (BPF, NaïveX indicates the OLS on returns based replica, using the same variables as in FactorX, X=1,2,3.

Mean is the mean error of replications, std.dev. the standard deviation, mad the mean absolute deviation, supmod the sup of the mo

Ncomp is the number of factors included in the factor model, nvar the number of variables included in the replica

The figures are averages over 5000 replications. Monte Carlo standard errors are reported in smaller fonts.

Table 5 - Monte Carlo comparison of Factor and OLS on returns based index replication (N=50, r1=5, r2=5, median)

	Naive 1	*Naive 2*	*Naive 3*	*Factor 1*	*Factor 2*	*Factor 3*
			In sample (1-500)			
mean	8.16	5.74	5.70	0.67	0.16	0.16
std.dev.	18.33	16.58	16.37	6.25	3.78	3.55
mad	15.18	13.57	13.55	5.00	3.01	2.83
supmod	59.96	55.10	55.42	19.89	12.13	11.43
ncomp				3.81		
nvar	3.89	8.38	7.54	3.89	8.38	7.54
			Out of sample (501-1000)			
mean	0.14	0.77	1.41	0.17	0.08	0.10
std.dev.	20.25	18.88	19.09	9.10	4.87	4.53
mad	16.66	15.62	15.87	7.40	3.92	3.66
supmod	67.82	63.25	64.16	32.95	17.91	17.01
			Out of sample (501-750)			
mean	0.40	0.70	0.70	-0.09	0.06	0.05
std.dev.	14.50	13.71	13.90	7.01	4.13	3.87
mad	11.95	11.30	11.40	5.68	3.31	3.10
supmod	49.06	46.23	46.17	25.17	14.74	13.84
			Out of sample (751-1000)			
mean	0.57	1.07	1.90	-0.01	0.04	0.06
std.dev.	14.69	13.65	13.50	7.01	4.13	3.91
mad	12.06	11.22	11.14	5.68	3.31	3.14
supmod	65.21	60.17	60.41	31.68	17.01	16.15

Notes

FactorX indicates the factor based replica using factors extracted from the variance covariance matrix of raw data (Base, X=3),
of the first autocovariance matrix (AC, X=2), or from the variance covariance matrix of a 50 period moving average of the data (BPF, X=1).
NaïveX indicates the OLS on returns based replica, using the same variables as in FactorX, X=1,2,3.
Mean is the mean error of replications, std.dev. the standard deviation, mad the mean absolute deviation, supmod the sup of the modulus
Ncomp is the number of factors included in the factor model, nvar the number of variables included in the replica
The figures are medians over 5000 replications.

Table 6 - Monte Carlo comparison of Factor and OLS on returns based index replication (N=50, r1=5, r2=5, ncomp

	Naive 1	*Naive 2*	*Naive 3*	*Factor 1*	*Factor 2*	*Factor 3*
			In sample (1-500)			
mean	25.80	18.61	-162.34	0.021	0.003	0.003
	5.73	4.08	127.53	0.001	0.000	0.000
std.dev.	56.03	39.89	101.50	1.382	0.489	0.479
	5.34	2.57	39.27	0.005	0.001	0.001
mad	46.80	32.83	76.85	1.102	0.390	0.382
	4.54	2.07	27.54	0.004	0.001	0.001
supmod	179.83	136.47	437.15	4.491	1.585	1.551
	16.23	9.49	200.58	0.017	0.005	0.005
ncomp				10.00		
				0.00		
nvar	17.03	36.41	36.81	17.03	36.41	36.81
	2.59	2.30	2.97	2.59	2.30	2.97
			Out of sample (501-1000)			
mean	17.89	7.69	-96.27	-0.052	0.012	0.002
	11.49	8.44	61.83	0.020	0.004	0.004
std.dev.	58.47	51.37	124.12	1.585	0.545	0.531
	5.26	4.23	53.46	0.007	0.001	0.002
mad	49.12	42.93	101.45	1.271	0.435	0.424
	4.49	3.49	43.31	0.006	0.001	0.001
supmod	205.34	171.56	461.94	5.719	1.864	1.822
	21.06	15.62	210.73	0.030	0.006	0.007
			Out of sample (501-750)			
mean	17.54	7.63	-45.87	-0.032	0.014	0.002
	8.78	6.52	26.93	0.016	0.003	0.003
std.dev.	43.99	37.13	78.53	1.489	0.532	0.521
	4.49	3.74	29.65	0.006	0.001	0.002
mad	36.45	31.09	65.60	1.189	0.424	0.416
	3.62	3.21	24.93	0.005	0.001	0.001
supmod	159.77	127.19	263.25	5.032	1.695	1.671
	18.99	14.04	99.56	0.025	0.006	0.006
			Out of sample (751-1000)			
mean	-0.51	15.42	-5.71	-0.332	-0.096	-0.082
	14.59	10.78	96.78	0.027	0.005	0.005
std.dev.	41.16	35.56	107.73	1.485	0.533	0.519
	3.19	2.45	48.91	0.006	0.001	0.002
mad	34.09	29.53	90.25	1.185	0.425	0.414
	2.63	2.02	41.42	0.005	0.001	0.001
supmod	199.58	166.61	457.68	5.457	1.780	1.733
	20.99	15.42	210.73	0.031	0.006	0.007

Notes

FactorX indicates the factor based replica using factors extracted from the variance covariance matrix of raw data (Base, X=3), of the first autocovariance matrix (AC, X=2), or from the variance covariance matrix of a 50 period moving average of the data (BPF, X=1).
NaïveX indicates the OLS on returns based replica, using the same variables as in FactorX, X=1,2,3.
Mean is the mean error of replications, std.dev. the standard deviation, mad the mean absolute deviation, supmod the sup of the modulus
Ncomp is the number of factors included in the factor model, nvar the number of variables included in the replica
The figures are averages over 5000 replications. Monte Carlo standard errors are reported in smaller fonts.

Table 7 - Percentage of variance explained by each factor

Factor	*1*	*2*	*3*
1	90.39	88.23	87.96
2	5.27	4.21	4.22
3	2.78	3.59	3.63
4	0.94	1.66	1.63
5	0.28	0.61	0.62
6	0.17	0.28	0.31
7	0.08	0.23	0.24
8	0.04	0.19	0.22
9	0.03	0.19	0.19
10	0.01	0.16	0.19

FactorX indicates the factor based replica using factors extracted from the variance covariance matrix of raw data (Base, X=3),
of the first autocovariance matrix (AC, X=2), or from the variance covariance matrix of a 2 month moving average of the data (BPF, X=1).

Table 8 - Data on Eurostoxx50, comparison of Factor and OLS on returns based index replication

	Naive 1	*Naive 2*	*Naive 3*	*Factor 1*	*Factor 2*	*Factor 3*	*Level 1*	*Level 2*	*Level 3*
					In sample (1-350)				
mean	-19.62	-11.12	-5.04	0.10	-0.06	0.02	-1.18	0.02	0.03
std.dev.	6.21	5.37	2.83	3.57	1.04	0.88	5.20	0.96	0.86
mad	5.18	4.20	2.44	2.93	0.83	0.68	4.06	0.78	0.66
supmod	33.20	21.24	11.05	11.70	3.76	2.76	15.83	2.89	3.01
ncomp				6.00	6.00	6.00			
nvar	6.00	21.00	19.00	6.00	21.00	19.00	6.00	21.00	19.00
					Out of sample (351-890)				
mean	-13.46	-5.36	3.44	20.11	-26.84	1.08	14.74	-24.86	2.30
std.dev.	29.48	13.67	16.73	19.56	22.28	5.25	18.16	19.10	6.23
mad	23.68	11.44	13.67	16.05	18.98	3.81	14.16	16.41	4.78
supmod	72.96	28.13	53.48	79.40	82.80	18.12	75.18	68.95	23.73
					Out of sample (351-650)				
mean	-36.55	-13.68	-9.63	5.92	-7.56	-0.93	4.01	-8.00	-1.53
std.dev.	16.54	7.41	5.68	8.51	5.73	2.74	9.26	5.37	3.09
mad	13.86	6.35	4.93	6.86	4.74	2.14	7.31	4.56	2.43
supmod	72.96	28.13	21.16	26.12	19.22	7.82	21.84	18.35	9.30
					Out of sample (651-890)				
mean	6.51	1.84	14.75	32.38	-43.52	2.82	24.03	-39.45	5.62
std.dev.	22.90	13.76	14.78	18.07	17.23	6.20	18.85	14.02	6.35
mad	20.57	12.69	13.10	15.28	14.99	4.47	15.88	12.34	4.92
supmod	56.28	26.20	53.48	79.40	82.80	18.12	75.18	68.95	23.73

Notes

FactorX indicates the factor based replica using factors extracted from the variance covariance matrix of raw data (Base, X=3),
of the first autocovariance matrix (AC, X=2), or from the variance covariance matrix of a 50 period moving average of the data (BPF, X=1).
NaïveX indicates the OLS on returns based replica, using the same variables as in FactorX, X=1,2,3.
LevelX indicates the OLS on prices based replica, using the same variables as in FactorX, X=1,2,3.
Mean is the mean error of replications, std.dev. the standard deviation, mad the mean absolute deviation, supmod the sup of the modulus
Ncomp is the number of factors included in the factor model, nvar the number of variables included in the replica

Figure 1: Out of sample tracking of replication methods