JBF 2236
22 November 2005 Disk Used
ARTICLE IN PRESS
No. of Pages 19, Model 1+

Available online at www.sciencedirect.com
SCIENCE DIRECT®

Journal of Banking & Finance xxx (2005) xxx–xxx

Journal of BANKING & FINANCE
www.elsevier.com/locate/jbf

Factor based index tracking

Francesco Corielli [a], Massimiliano Marcellino [b,*]

[a] IMQ – Bocconi University, Italy
[b] IEP – Bocconi University, IGIER and CEPR, Via Salasco 5, 20136 Milano, Italy

Received 8 June 2005; accepted 4 July 2005

Abstract

Stock index tracking requires to build a portfolio of stocks (a *replica*) whose behavior 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 or integrated (persistent) components in the tracking error. 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. Moreover, most existing methods do not take into account the known structure of the index weight system. In this paper we represent the index components with a dynamic factor model. In this model the price of each stock in the index is driven by a set of common and idiosyncratic factors. Factors can be either integrated or stationary. We develop a procedure that, in a first step, builds a replica that is driven by the same persistent factors as the index. This procedure is grounded in recent results which suggest the application of principal component analysis for factor estimation even for integrated processes. In a second step, it is also possible to refine the replica so that it minimizes a specific loss function, as in the traditional approach. In both steps the replica weights depend on the existing information on the index weights system. An extended set of Monte Carlo simulations and an application to the most widely used index in the European stock market, the EuroStoxx50 index, provide substantial support for our approach.
© 2005 Elsevier B.V. All rights reserved.

JEL classification: C30; C53; G10

Keywords: Index tracking; Cointegration; Unit roots; Factor models

* Corresponding author. Tel.: +39 02 5836 3327; fax: +39 02 5836 3302.
E-mail address: massimiliano.marcellino@uni-bocconi.it (M. Marcellino).

0378-4266/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.jbankfin.2005.07.012

UNCORRECTED PROOF

1. Introduction

Stock index tracking underlies a big slice of the fund management industry world wide. Index tracking can be the explicit strategy of a fund, e.g., the Vanguard 500 index with a total asset value of dozens of 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 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.

The purpose of this paper is to clarify some problems related to index replication, to state conditions under which a buy and hold replication strategy may succeed, 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 substantial, 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. (2003) 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

UNCORRECTED PROOF

suggest diagnostics based on the study of the autocorrelation 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 Rotemberg (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, existing methods allow to test for cointegration among a limited set of series while an index can include a very large number of components. Third, 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. Finally, 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. Moreover, in order to avoid relevant transaction costs, the coefficients should be constant for sufficiently long time periods and this implies the use of buy and hold tracking portfolios. Hence, it is by no means obvious that an acceptable solution, i.e., a solution whose error is free of low frequency/integrated components, 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 (2002), Forni and Reichlin (1996, 1998), Forni et al. (2000), Bai and Ng (2004), Bai (2003). 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 or at least minimizes 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 show that a low frequency free 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. However, this condition can hold only when a linear factor model in price levels holds, and this is not the most widely accepted model for price dynamics. Therefore, index replication seems at odds with commonly accepted price models.

UNCORRECTED PROOF

It is possible to consider several other 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 concentrate on the linear model in price levels. In a related paper, Corielli and Marcellino (2005), we consider the linear in returns model, which, however, requires a dynamic trading strategy.

The structure of the paper is as follows: In Section 2 we shortly describe the definitions and characteristics of the most commonly studied stock market indexes. In Section 3 we discuss the properties of the most widely used tracking error measures and we show how these measures, without additional requirements, cannot distinguish between good and bad index replications. In Section 4 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 5 we propose a practical sequential procedure for building a replicating portfolio. In Section 6 we report the results of a set of Monte Carlo experiments to evaluate the performance of our approach. In Section 7 the method is applied to construct a replica of the EuroStoxx50 index. Section 8 summarizes the main results and contains suggestions for further research.[1]

2. Remarks on stock indexes

In this paper we consider index replication using buy and hold strategies. It is then relevant for us to analyze the definitions of the most widely used classes of stock indexes, in order to see whether these can actually be considered as buy and hold portfolios with possible periodic (predetermined) revisions. When this is the case, the problem of portfolio replication, discussed in this paper, and that of portfolio updating, not discussed in this paper, can be discussed independently. Moreover, when this is true, widely used replication methods based on regression on returns shall give biased answers as the return of a buy and hold index cannot be expressed as a constant weights linear combination of the component returns with constant weights.

In this section, and in this paper, we only consider price indexes. For each price index there usually exists a total return index, that is, an index which takes into account cash

[1] An extended version of this paper, with added considerations about a non-linear model based on returns is available from the authors.

UNCORRECTED PROOF

outflows, as dividends, and inflows. Price indexes come basically in three kinds: price weighted, cap weighted and constant weighted (most often equally weighted). The value of a unit of a price weighted index is proportional to the value of a portfolio containing one share of each stock considered by the index. The value of a unit of a cap weighted index is proportional to the value of a portfolio containing all the existing shares of each stock considered by the index. Therefore, up to index revision determined both by firm actions (e.g., stock splits) or by the choice of a new set of stocks, both price and cap weighted indexes correspond to buy and hold strategies where a fixed number of shares for each stock are bought and hold.

The most widely known example of a price weighted index is the Dow Jones, where the prices of 30 stocks are simply summed. The most widely known example of cap weighted index is the S&P500, where the total capitalization of 500 companies is summed. The EuroStoxx50 index we consider in this paper is a cap weighted index with a constraint on the maximum weight of each stock.

Constant weighted indexes, in principle, do not correspond to buy and hold portfolios, as the composition of the portfolio must change each day in order for the relative amounts of money invested in each stock to be kept constant. Obviously the continuous rebalancing of the index portfolio would make such an index useless as a comparison benchmark, due to the (unaccounted for) transaction costs. In practice, constant weighted indexes resort to some compromise. The most popular example is the S&P500 equally weighted index. At the beginning of each quarter the index portfolio is built anew in such a way that the amount of money invested in each stock is the same. The portfolio then evolves as a buy and hold strategy for a quarter and it is updated at the beginning of the following quarter. This means that during the quarter the proportion of money invested in each stock is NOT kept constant.

3. Measures of tracking error

The most widely used measures of tracking error are based on some summaries of returns differences between the tracking portfolio and the index, the same measures being used when returns are defined as log ratios or as percentage variations. The, by far, most widely used tracking error measure in theoretical and practical work are the tracking error variance and the tracking mean square error defined as

$$S = \frac{1}{n}\sum_{i=1}^{n}(r_i - R_i - (\bar{r}_i - \bar{R}_i))^2, \quad M = \frac{1}{n}\sum_{i=1}^{n}(r_i - R_i)^2,$$

where r_i and R_i are, respectively, the return from the tracking portfolio and from the index and overbars are used to indicate means.

Beasley et al. (2003) argue against the use of S as a measure of tracking error because S can be zero while the returns of the two portfolios differ by a constant so that portfolios values drift apart. We agree with this criticism, however the simple choice of M as a tracking error measure is clearly not enough in order to discriminate between "good" and "bad" tracking. In fact, for a given constant value of M, the values of a tracking portfolio and of an index can behave in very different ways, while the subscriber of a tracking fund should be able of exchanging, in any moment, an investment in the fund with a (nearly) equal investment in the index. Consequently, a successful tracking strategy requires the values of the tracking portfolio and of the index to be close. Deviations between the values of the fund and of the index are admissible, but the two processes should cross frequently.

Concentrating on a day by day analysis of squared return differences is perfectly legitimate but it is not enough to achieve good tracking behavior. The temporal ordering of the return differences and their persistence are relevant for assessing the success of a tracking strategy, but both S and M do not take into account such information.

4. Necessary conditions for index tracking

In this section we discuss necessary conditions for the construction of a parsimonious index, $\tilde{I}_t$, whose implied tracking error shows limited low frequency components. We discuss first the definition of the index and the model for the index components, next define the index replica, and finally give conditions for the construction of a reliable index replica. The focus is not on the minimization of a particular tracking error summary measure (e.g., the S or M loss functions defined in the previous section), but on the procedure to follow in order to build a replica whose low frequency tracking error behavior is largely independent of the chosen summary measure. However, since in general the necessary conditions only identify a set of possible replicas and not a specific replica, the minimization of a particular tracking error measure can be suggested as a tool to choose among the admissible replicas.

4.1. The index and the model for prices

Let us write the generic index I_t in matrix terms as

$$I_t = wp_t, \tag{1}$$

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 $t = 1,\ldots,T$. The weights can themselves be time varying provided they are known for each time t but, for simplicity, we assume that they are constant. I_t can be a price or a cap weighted index in the terminology of Section 2, depending on the choice of w.

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 basic set of hypotheses we make on (2) are

(1) the factors f_t are uncorrelated among themselves and with the e_t,
(2) the factors f_t are integrated or stationary processes,
(3) the errors e_t are stationary and may be correlated in time and across prices.

Under these hypotheses we have an integrated dynamic approximate factor model of the kind studied by Bai (forthcoming) and Bai and Ng (2004), to whom we refer for more details and precise conditions on the correlation structure of f_t and e_t. The model with stationary f_t was studied by Stock and Watson (2002) and Forni et al. (2000), but this hypothesis is too restrictive when modelling stock prices.

UNCORRECTED PROOF

Usually, in finance the distinction between f_t and e_t is the distinction between "common factors" and "idiosyncratic factors", and 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. If 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, the common factors will be substantially more persistent than the idiosyncratic errors. Yet, in the event of persistent idiosyncratic errors, they should also be taken into consideration to build working tracking portfolios.

Notice also that the model in (2) is a linear factor model with constant weights matrix Λ and additive "error" structure for the price processes. In finance models of this kind are most frequently specified for log prices or for returns. We do not suggest here that a linear factor model in prices is more correct as a description of price behavior then a linear model in returns or in price logs. The price level factor model is useful because it is the only specification which allows for exact index tracking with a buy and hold strategy and with tracking errors independent of the f_t processes. In Corielli and Marcellino (2005) we examine a return based factor model and show that a solution of the tracking problem can be achieved only for dynamic (not buy and hold) strategies, and a similar analysis can be performed for log price models. Therefore, if a linear factor model for price levels does not hold, buy and hold strategies are not a right choice for index replication.

4.2. The index replica and the tracking error

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

$$\tilde{I}_t = \beta S p_t, \tag{3}$$

where β is a $1 \times q$ vector of weights, $q \geqslant r$, 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, $\tilde{I}_t$ 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. As stressed above, 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.

It is also worth stressing that the index replica must be a portfolio of securities, i.e., it cannot be a function of leads, lags or differences of prices (as instead, e.g., when estimating a conditional expectation given a set of prices). This provides a strong reason for specifying the factor model for the price levels, as we did in the previous section, see Bai (forthcoming) for a similar observation in a related context. An *admissible* replica is a buy and

UNCORRECTED PROOF

hold portfolio $\tilde{I}_t$ such that $I_t - \tilde{I}_t$ is independent of f_t. In other words, both the index and the replica must share the same f_t structure.

4.3. Index tracking

Let us assume for the moment that the choice of assets represented in $\tilde{I}_t$ 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 $\tilde{I}_0 = I_0$. Under these hypotheses we can deduce a simple but strong result which is independent of the choice of the tracking error measure.

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

$$\beta^* = w(\Lambda \vdots p_0)(S\Lambda \vdots Sp_0)^+ + h[I_q - (S\Lambda \vdots Sp_0)(S\Lambda \vdots Sp_0)^+], \tag{4}$$

where $A \vdots B$ *indicates the matrix* $(A\ B)$, h *is any* $1 \times q$ *vector*, I_q *is the* $q \times q$ *identity matrix and* $(\cdot)^+$ *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 $\tilde{I}_t$ are, respectively, $w\Lambda f_t$ and $\beta S\Lambda f_t$, while $I_0 = wp_0$ and $\tilde{I}_0 = \beta Sp_0$. Hence, we can form the system $\beta(S\Lambda \vdots Sp_0) = w(\Lambda \vdots p_0)$. The general representation of the solution (if it exists) is β^* in (4) (see, e.g., Graybill, 1983, chapter 7.3). □

Thus, a set of linear constraints on β^* are required to avoid the leakage of f_t in $I_t - \tilde{I}_t$. 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).

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 β^* 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 β^* coincides with the cointegration vector for (I_t, Sp_t), when there exists cointegration.

The constraint in (4) does not, in general, identify one set of admissible weights. From the point of view of (4) any choice of h which satisfies the constraint is equivalent. Hence, the choice of h gives the necessary degrees of freedom to optimize the replica with respect to a specific choice of the loss function. Possible choices are the S and M functions considered in Section 3. Yet, minimization of M and S would require the use of non-linear

UNCORRECTED PROOF

methods and this would hamper an extensive numerical evaluation based on Monte Carlo simulation. For the sake of simplicity, we adopt a quadratic loss function in price levels, $L = (I - \tilde{I})(I - \tilde{I})'$, where $I = (I_1, \ldots, I_T)$ and $\tilde{I}$ is similarly defined. In this case we can find an explicit formula for the optimal value of h which, when $q = r + 1$ and $(BSPP'S'B')^{-1}$ exists, is given by

$$h^* = -(ASPP'S'B' - IP'S'B')(BSPP'S'B')^{-1}, \tag{5}$$

where $A = w(\Lambda \vdots p_0)(S\Lambda \vdots Sp_0)^+$, $B = [1 - (S\Lambda \vdots Sp_0)(S\Lambda \vdots Sp_0)^+]$, and P is the $N \times T$ matrix of prices. This choice of h is adopted in the simulation experiments and in the empirical application below.

An explicit formula for h permits to evaluate more directly the effects of the restrictions imposed by (4) on the replica weights. To do this we derive the OLS solution to the tracking problem when the sum of squared errors are defined on index levels and the constraints given in (4) are not considered. The tracking error can be written as $\varepsilon_t = I_t - \tilde{I}_t = (w - \beta S)p_t$, and the loss function to be minimized with respect to the $1 \times q$ vector β (assuming S and q are known) is

$$(I - \tilde{I})(I - \tilde{I})' = (w - \beta S)PP'(w - \beta S)'. \tag{6}$$

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

$$\beta_{\text{RLS}} = w(PP'S' \vdots p_0)C^+ + h^*[I - CC^+],$$

where $C = (SPP'S' \vdots Sp_0)$. The main differences between β^* in (4) and β_{RLS} can be best appreciated if we set $q = r$ and disregard the condition $\tilde{I}_0 = I_0$. In this case the (supposed unique) vector satisfying (4) is given by

$$\beta^* = w\Lambda(S\Lambda)^{-1},$$

while the OLS solution simplifies to

$$\beta_{\text{OLS}} = w(\Lambda FF'\Lambda'S' + ee'S')(S\Lambda FF'\Lambda'S' + See'S')^{-1}. \tag{7}$$

The tracking error when using β^* is given by $\varepsilon^* = (w - w\Lambda(S\Lambda)^{-1}S)e$. Since β^* 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 $\beta_{\text{OLS}}S\Lambda F$. 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 β^*. This happens in the factor model in Stock and Watson (1988) where the factors are pure random walks. Actually, in this case β_{OLS} behaves asymptotically as

$$\beta_{\text{OLS}} \approx w(\Lambda FF'\Lambda'S')(S\Lambda FF'\Lambda'S')^{-1} = w\Lambda(S\Lambda)^{-1} = \beta^*. \tag{8}$$

5. 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, Eq. (2). The estimated loading matrix, $\widehat{\Lambda}$, can then be used to construct β^* in (4). Estimation is discussed in the first subsection. Second, we have to choose how many and which stocks are to be used in the replication. We suggest a procedure to select the

UNCORRECTED PROOF

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 stocks, 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 (4). In particular, since in general it is $q > r + 1$, we can use the degrees of freedom in h to minimize a given tracking error measure.

5.1. Estimation of the factor structure

Following Stock and Watson (2002) and Bai (forthcoming), we estimate the factors F and their loadings Λ relying on a principal component analysis of the matrix of historical prices P. More precisely, our aim is to decompose the stock prices movements into orthogonal components ordered by their contribution to the *low* frequency variance of the prices. 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.[2]

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 $\Lambda KK^{-1}f_t$, where K is a generic $r \times r$ matrix with full rank, the expression for β^* in (4) does not change.

For the choice of the number of factors, r, Stock and Watson (2002) 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 (2002) 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

UNCORRECTED PROOF

---

[2] Principal component based estimators are consistent for (the space spanned by) the factors, when the latter are either stationary or integrated but with stationary idiosyncratic errors, see Bai (forthcoming). Bai and Ng (2004) develop a procedure for the case of integrated e_t, based on a factor model for the differenced series. This is an unlikely event in our context, since series within the same sector tend to be driven by the same non-stationary forces. Therefore, we focus on the model in the price levels.

hand, suggested 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 6, 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.

5.2. 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 a loss function, subject to the constraint that the solution satisfies Proposition 1. However, this approach becomes quickly computationally 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. This is its main advantage, even though more sophisticated approaches could be developed in the future.

As mentioned above, for the construction of the optimal weights β^* 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 that avoids 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.

UNCORRECTED PROOF

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 best suited for our purposes.

6. 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), OLS on levels, and factor based tracking.

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:

$$p_t = \Lambda_1 f_{1t} + \Lambda_2 f_{2t} + e_t, \tag{9}$$

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.

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.

The number of factors is determined so that at least 99.9% of the variability of the series is explained. We also experimented with the Bai and Ng (2002) 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.[3]

The variables to be included in the index replica are selected according to the procedure described in Section 5.2, with the value for the minimum R^2 in step 2 set at 0.80. The factor based index replication is then built, using the formulae in Section 4. The same variables are used for the OLS on returns and levels replications. The model is estimated over the

[3] Similarly, the fraction of explained variance is decreased to 90% when $r_1 = 2$ and $r_2 = 8$ otherwise a (too) large number of factors is selected in practice.

UNCORRECTED PROOF

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.

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). Notice that standard procedures to evaluate the statistical significance of the difference in loss functions from different tracking methods cannot be applied in our context. In fact, these procedures, e.g., Diebold and Mariano (1995), are suited to compare a set of h-step ahead forecasts, with the forecast horizon h fixed, while we have one forecast for each h, with $h = 1, 2, \ldots, 500$.

The first vertical panel of Table 1 reports the average value of the loss functions over 5000 replications for each tracking 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 on returns both in-sample and out of sample, according to any criterion, and the gains are substantial. Second, 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. Third, OLS on levels performs much better than OLS on returns, but it is still beaten, in almost all the out-of-sample tests, by the factor based replica. Fourth, the tracking performance deteriorates with the forecast horizon, less so for the factor based replications. Fourth, the selected number of factors is on average about 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 7.

To evaluate the relative role of stationary and integrated factors, the second and third vertical panels of Table 1 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 on returns methods improves with a lower number of integrated factors. This happens 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, the number of integrated factors seems to have a marginal effect on the relative results for the OLS on levels and factor based replications. Third, on average, the selected number of factors is about 3 when $r_1 = 2$ and to 5 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 9 for both $r_1 = 2$ and $r_1 = 8$, which suggests that the total number of factors matters more than that of integrated or stationary factors.

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 fix the number of factors to the true value of 10. Third, we compare the performance of the methods using the median rather than the mean over the simulations. The results are reported in the three vertical panels of Table 2.

A first finding is 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 versus 7 when $N = 50$.

UNCORRECTED PROOF

Table 1
Monte Carlo comparison of factor and OLS based index replication ($N = 50$)

	$r_1 = 5, r_2 = 5$			$r_1 = 2, r_2 = 8$			$r_1 = 8, r_2 = 2$		
	Returns	Levels	Factors	Returns	Levels	Factors	Returns	Levels	Factors
In sample									
MEAN	25.26	0.57	0.47	10.87	0.94	0.88	29.90	0.24	0.25
	6.06	*0.03*	*0.02*	*1.19*	*0.08*	*0.04*	*7.33*	*0.01*	*0.01*
STD.DEV.	42.63	4.15	3.87	18.78	5.01	4.31	50.59	3.31	3.48
	4.04	*0.04*	*0.03*	*1.30*	*0.09*	*0.05*	*5.96*	*0.03*	*0.03*
MAD	35.04	3.33	3.10	15.66	4.05	3.46	44.24	2.67	2.83
	3.29	*0.03*	*0.03*	*1.09*	*0.08*	*0.04*	*5.11*	*0.02*	*0.02*
SUPMOD	142.19	12.88	11.93	62.62	15.79	13.29	139.89	10.23	10.93
	14.95	*0.11*	*0.10*	*3.61*	*0.28*	*0.16*	*23.84*	*0.08*	*0.08*
NCOMP			3.63			2.67			4.92
			0.01			*0.01*			*−2.41*
NVAR			7.29			8.84			9.00
			0.07			*0.14*			*0.06*
Out of sample 1									
MEAN	−19.38	−0.20	−0.13	3.44	−0.24	−0.09	−19.52	0.13	0.21
	12.66	*0.17*	*0.17*	*1.17*	*0.22*	*0.19*	*7.31*	*0.17*	*0.17*
STD.DEV.	59.57	6.98	6.63	19.25	6.52	5.93	97.83	7.01	7.31
	7.13	*0.08*	*0.08*	*0.79*	*0.13*	*0.09*	*15.13*	*0.08*	*0.08*
MAD	49.12	5.76	5.47	15.90	5.35	4.87	95.13	5.95	6.00
	5.79	*0.07*	*0.07*	*0.63*	*0.11*	*0.08*	*11.52*	*0.07*	*0.06*
SUPMOD	199.34	25.23	23.99	68.80	23.48	21.53	265.49	25.41	26.01
	24.28	*0.28*	*0.28*	*3.30*	*0.46*	*0.34*	*39.78*	*0.30*	*0.30*
Out of sample 2									
MEAN	−15.07	−0.18	−0.22	2.76	−0.15	0.09	9.85	−0.02	0.19
	9.62	*0.12*	*0.13*	*0.92*	*0.16*	*0.14*	*8.04*	*0.12*	*0.14*
STD.DEV.	46.54	5.26	5.01	14.42	5.22	4.80	68.80	5.34	5.47
	7.20	*0.05*	*0.05*	*0.69*	*0.08*	*0.06*	*8.88*	*0.04*	*0.05*
MAD	38.81	4.31	4.10	11.93	4.24	3.90	41.58	4.32	4.44
	6.16	*0.05*	*0.04*	*0.57*	*0.07*	*0.05*	*6.43*	*0.05*	*0.05*
SUPMOD	152.42	18.68	17.97	49.57	18.05	16.89	159.45	18.73	18.92
	22.23	*0.20*	*0.20*	*2.44*	*0.30*	*0.25*	*21.35*	*0.17*	*0.15*
Out of sample 3									
MEAN	−23.09	−0.22	−0.05	4.03	−0.32	−0.25	−9.02	0.36	0.42
	15.42	*0.24*	*0.23*	*1.44*	*0.30*	*0.25*	*16.28*	*0.27*	*0.26*
STD.DEV.	45.22	5.65	5.36	15.06	5.37	4.90	75.18	5.48	5.73
	5.59	*0.06*	*0.06*	*0.73*	*0.09*	*0.07*	*11.98*	*0.05*	*0.06*
MAD	38.15	4.64	4.39	12.40	4.38	3.98	60.30	4.53	4.70
	4.81	*0.05*	*0.05*	*0.58*	*0.08*	*0.06*	*14.05*	*0.05*	*0.05*
SUPMOD	195.57	24.56	23.33	66.75	22.70	20.70	360.64	24.04	25.30
	24.18	*0.28*	*0.27*	*3.19*	*0.45*	*0.33*	*40.90*	*0.25*	*0.29*

Notes: Factor indicates the factor based replica. Levels and returns indicate the OLS based replica, using the same variables as in factor. 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. Out of sample 1, 2, 3 refer, respectively, to observations 501–1000, 501–750, and 751–1000.
The figures are averages over 5000 replications. Monte Carlo standard errors are reported in italicised fonts.

UNCORRECTED PROOF

Table 2
Monte Carlo comparison, sensitivity analysis

	$n = 100$, $r_1 = 5$, $r_2 = 5$			$n = 50$, 10 factors, $r_1 = 5$, $r_2 = 5$			$n = 50$, median, $r_1 = 5$, $r_2 = 5$		
	Returns	Levels	Factors	Returns	Levels	Factors	Returns	Levels	Factors
In sample									
MEAN	11.97	0.58	0.62	13.18	0.00	0.00	5.93	0.17	0.16
	3.42	*0.02*	*0.03*	*2.88*	*0.00*	*0.00*	*32.08*	*0.04*	*0.02*
STD.DEV.	37.51	3.97	4.25	34.88	0.47	0.46	17.34	3.60	3.43
	2.49	*0.03*	*0.05*	*2.79*	*0.00*	*0.00*	*13.77*	*0.05*	*0.04*
MAD	27.43	3.22	3.40	29.01	0.37	0.37	14.18	2.88	2.72
	1.61	*0.03*	*0.03*	*2.29*	*0.00*	*0.00*	*11.25*	*0.04*	*0.03*
SUPMOD	92.06	11.98	13.11	113.39	1.47	1.45	57.15	11.69	10.97
	7.69	*0.10*	*0.13*	*8.08*	*0.01*	*0.01*	*67.84*	*0.16*	*0.13*
NCOMP			3.68			10.00			3.81
			−11.17			*0.00*			*0.01*
NVAR			9.63			36.97			7.39
			0.16			*0.04*			*0.06*
Out of sample 1									
MEAN	15.57	−0.04	−0.15	−8.74	−0.01	−0.01	1.31	0.12	−0.07
	5.07	*0.16*	*0.18*	*5.38*	*0.01*	*0.00*	*46.15*	*0.20*	*0.20*
STD.DEV.	50.23	6.70	7.17	46.70	0.55	0.54	19.59	4.69	4.46
	5.38	*0.07*	*0.08*	*4.52*	*0.00*	*0.00*	*30.52*	*0.09*	*0.08*
MAD	39.70	5.73	5.99	39.09	0.44	0.43	16.47	3.78	3.59
	3.69	*0.06*	*0.08*	*3.87*	*0.00*	*0.00*	*25.23*	*0.07*	*0.07*
SUPMOD	144.37	23.79	24.97	156.78	1.95	1.89	65.57	17.44	16.50
	16.15	*0.27*	*0.29*	*14.29*	*0.01*	*0.01*	*114.57*	*0.31*	*0.30*
Out of sample 2									
MEAN	8.27	0.01	0.02	−0.75	−0.01	−0.01	0.10	0.00	−0.11
	2.67	*0.14*	*0.14*	*2.99*	*0.00*	*0.00*	*26.26*	*0.15*	*0.15*
STD.DEV.	31.86	5.36	5.03	31.55	0.53	0.53	14.12	3.95	3.80
	3.20	*0.06*	*0.07*	*2.47*	*0.00*	*0.00*	*16.18*	*0.07*	*0.06*
MAD	21.80	4.34	4.18	26.52	0.42	0.42	11.66	3.15	3.05
	2.88	*0.05*	*0.04*	*2.10*	*0.00*	*0.00*	*13.82*	*0.05*	*0.05*
SUPMOD	111.00	18.84	18.45	102.83	1.72	1.67	45.92	14.12	13.40
	8.60	*0.20*	*0.21*	*7.77*	*0.01*	*0.01*	*62.20*	*0.24*	*0.23*
Out of sample 3									
MEAN	9.77	−0.08	−0.05	−15.64	−0.01	−0.02	2.25	0.14	−0.10
	8.14	*0.21*	*0.25*	*8.20*	*0.01*	*0.01*	*67.06*	*0.27*	*0.27*
STD.DEV.	31.19	5.64	5.37	35.23	0.53	0.53	13.78	4.01	3.79
	3.85	*0.07*	*0.05*	*3.45*	*0.00*	*0.00*	*26.02*	*0.06*	*0.06*
MAD	27.24	4.63	4.37	29.58	0.43	0.42	11.26	3.21	3.04
	3.14	*0.05*	*0.05*	*2.95*	*0.00*	*0.00*	*22.34*	*0.05*	*0.05*
SUPMOD	171.08	25.07	23.00	153.20	1.87	1.82	62.31	16.47	15.71
	16.09	*0.33*	*0.31*	*14.23*	*0.01*	*0.01*	*114.11*	*0.31*	*0.30*

Notes: Factor indicates the factor based replica. Levels and returns indicate the OLS based replica, using the same variables as in factor. 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. Out of sample 1, 2, 3 refer, respectively, to observations 501–1000, 501–750, and 751–1000. The figures are averages over 5000 replications in the first two vertical panels, medians in the third one.
Monte Carlo standard errors are reported in italicised fonts.

UNCORRECTED PROOF

When the true number of factors is imposed, 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 versus 9 before. The higher number of variables likely underlies the improved performance of OLS on levels, which becomes even closer to factor based replication, and the worse performance of the OLS on returns method, due to collinearity problems because of the large number of regressors. Yet, 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.

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.

In summary, factor based index tracking performs much better than conventional OLS on returns based on a variety of loss functions. The gains are smaller with respect to OLS on levels, but still present and systematic. Moreover, the factor based approach provides an easy procedure for the selection of the variables in the replica. Their number is substantially smaller than that in the index, usually only slightly larger than twice the number of selected factors. Finally, it is worth recalling that the factor based approach provides more flexibility than OLS on levels or returns when the composition or the weights of the index change. In fact given the new set of weights or titles the factor based replica can be instantaneously modified without the need for a new estimation process while any OLS based method requires the computation of new estimates.

7. Tracking the EuroStoxx50

In this section we evaluate the performance of our factor based methods for tracking the EuroStoxx50 index, relative to the common OLS on returns and to OLS on prices. We use a daily dataset from January 1997 to June 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.[4] We split the sample so that the replica portfolio weights are estimated using the first set of 350 observations, while index tracking is carried out on the sample 351–890.

UNCORRECTED PROOF

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

Table 3
EuroStoxx50, comparison of factor and OLS based index replications

	OLS returns	OLS level	Factor
In sample (1–350)			
MEAN	−6.687	−0.020	−0.0025
STD.DEV.	3.871	0.654	0.6387
MAD	3.006	0.524	0.5145
SUPMOD	14.782	2.031	1.7651
NCOMP		10	
NVAR	24	24	24
Out of sample (351–890)			
MEAN	2.039	−5.365	−2.8373
STD.DEV.	12.729	7.103	6.009
MAD	10.694	6.285	5.2213
SUPMOD	29.326	21.157	18.3402
Out of sample (351–650)			
MEAN	−6.562	1.333	2.57
STD.DEV.	5.294	2.591	2.6472
MAD	4.611	2.241	2.2735
SUPMOD	18.616	7.715	9.4822
Out of sample (651–890)			
MEAN	9.454	−11.138	−7.4987
STD.DEV.	12.605	4.008	3.7693
MAD	11.655	3.315	2.857
SUPMOD	29.326	21.157	18.3402

Notes: 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 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. The comparison is made with the OLS on returns and on price levels replications using the same variables as in the factor based approach.

The percentages of variance explained by the eigenvectors associated with the largest 10 eigenvalues are, respectively, 90.39, 5.27, 2.78, 0.94, 0.28, 0.17, 0.08, 0.04, 0.03, 0.01. These values suggest that the variance of the first three factors is relatively more concentrated on the low frequencies than the variance of the other factors. 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 estimated factors.[5]

In accordance with the simulation design, we select 10 factors in order to explain about 99.9% of the variance of all prices. The number of variables required to mimic the factor structure of the index is to 21. The performance of the resulting index replications is summarized in Table 3. From Table 3, the factor method is the best according to all criteria in

[5] Bai and Ng (2004) 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.

UNCORRECTED PROOF

sample, over the whole forecast samples, and in the second forecast subsamples. The OLS in levels is better in the first forecast sample, but the differences with the factor replication are minor. In all cases, the OLS on returns, the most common method in practice, is by far the worst.

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, or its persistent deviations from the mean.

In order to give an idea of the absolute average quality of the replicating portfolio, since the mad in the replica is 5.221 for the factor method, 6.285 for OLS on levels and 10.694 for OLS on returns, while the average of the index is about 234 out of sample, the factor replica shows an error of about 2.23% of the average value of the index, while this goes up to about 2.69% for OLS on levels and 4.47% for the commonly used OLS on returns.

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

8. 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 EuroStoxx50 index. The results are quite encouraging, and emphasize the importance of a statistical approach to index tracking.

9. Uncited references

Chamberlain and Rothschild (1983), Clements and Hendry (1999), Rudolf et al. (1999).

Acknowledgements

We are grateful to two anonymous referees and seminar participants at Bocconi University for useful comments on a previous draft. Marcellino thanks the Italian MIUR for partial financial support.

UNCORRECTED PROOF

References

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.

Anderson, T.W., 1963. The asymptotic theory for principal component analysis. Annals of Mathematical Statistics 343, 122–148.

Bai, J., 2003. Inferential theory for factor models of large dimensions. Econometrica 71, 135–171.

Bai, J., forthcoming. Estimating cross-section common stochastic trends in non-stationary panel data. Journal of Econometrics.

Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70, 191–223.

Bai, J., Ng, S., 2004. A panic attack on unit roots and cointegration. Econometrica 72, 1127–1177.

Beasley, J.E., Meade, N., Chang, T.J., 2003. An evolutionary euristic for the index tracking problem. European Journal of Operational Research 148, 621–643.

Brillinger, D.R., 1981. Time Series Analysis: Data Analysis and Theory. Holt, Rinehart and Winston, New York.

Chamberlain, G., Rothschild, M., 1983. Arbitrage factor structure, and mean variance analysis of large asset markets. Econometrica 51, 1281–1304.

Clements, M., Hendry, D.F., 1999. Forecasting Non-Stationary Economic Time Series. MIT Press, Cambridge, MA.

Corielli, F., Marcellino, M., 2005. Index tracking: The linear in returns case, Mimeo.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. Journal of Business and Economic Statistics 13, 253–263.

Forni, M., Reichlin, L., 1996. Dynamic common factors in large cross-sections. Empirical Economics 21, 27–42.

Forni, M., Reichlin, L., 1998. Let's get real: A dynamic factor analytical approach to disaggregated business cycle. Review of Economic Studies 65, 453–474.

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2000. The generalised factor model: Identification and estimation. The Review of Economic and Statistics 82, 540–554.

Graybill, F.A., 1983. Matrices with Applications in Statistics. Wadsworth, Belmont, CA.

Kaiser, H.F., 1960. The application of electronic computers to factor analysis. Educational and Psychological Measurement 2, 141–151.

Pindyck, R.S., Rotemberg, J.J., 1992. The comovement of stock prices. Quarterly Journal of Economics 108, 1073–1103.

Pope, P.F., Yadav, P.K., 1994. Discovering errors in tracking error. The Journal of Portfolio Management 20, 27–32.

Rohweder, H.C., 1998. Implementing stock selection ideas: Does tracking error optimization do any good? The Journal of Portfolio Management 24, 49–59.

Rudolf, M., Wolter, H.-J., Zimmermann, H., 1999. A linear model for tracking error minimization. Journal of Banking and Finance 23, 85–103.

Stock, J.H., Watson, M.W., 1988. Testing for common trends. Journal of the American Statistical Association 83, 1097–1107.

Stock, J.H., Watson, M.W., 2002. Macroeconomic forecasting using diffusion indexes. Journal of Business and Economic Statistics 20, 147–162.

Wang, M.Y., 1999. Multiple benchmark and multiple portfolio optimization. Financial Analyst Journal 55, 63–72.

UNCORRECTED PROOF