

Hedge Funds Replication and Factor Models

Serge Darolles¹, Gulden Mero²

Preliminary Version

This Draft: May 2007

Abstract

In this paper, we present a four-step dynamic approach allowing replication of Hedge Fund returns. Instead of fixing the number of factors and their definitions in the replicating portfolio, our approach permits to choose at each point in time the best factors to use. Hence we get a deep comprehension of the factor structure underlying Hedge Funds returns. In particular, we first obtain the risk dimension, i.e. the optimal number of latent risk factors. Second, we assess the economic interpretation of these factors and finally, we build an optimal replicating portfolio. This approach is used to first analyse individual Hedge Funds belonging to the Equity Hedge strategy, and then build a clone of the aggregated index.

¹ SGAM Alternative Investments and CREST-INSEE, Paris.

² Allocataire de Recherche en Finance: Institut de Gestion de Rennes (IGR-IAE) & CREM UMR CNRS n° 6211 : 2, Rue Richard Lenoir, 35000 Rennes ; Tel : 06 30 01 33 70 ; Email : guldenmero@hotmail.com

1. Introduction

In the last decade, interest in Hedge Funds from both academics and investors has grown dramatically. Because hedge funds are typically organized as private investment vehicles for wealthy individuals and institutional investors, they do not have to disclose their activities publicly. Hence, little is known about the risk in hedge fund strategies. Academics are intrigued by the unconventional performance characteristics in hedge funds, while investors are attracted by the option-like returns and the low correlation with different asset classes. However, high performance fees charged by the managers, lack of liquidity, lack of transparency and fear for style risk have raised the question whether it is possible to generate hedge-fund-like returns with less effort using more liquid assets such as stock indexes, bonds, currency, commodity and interest rate futures.

The idea is to employ statistical methods and factor models to replicate the hedge fund mean returns. Thus, clones will capture on average, the main performance trends of the hedge funds and provide to investors similar return characteristics with a much lower cost. But in which way can hedge funds be replicated? There exist two main approaches. One possible approach aims to replicate hedge fund returns using several observed factors in order to capture the main risk exposures of the hedge fund returns. For example Lo and Hasanhobvic (2006) adopt a dynamic linear approach of return replication using stock indexes, bonds, and commodity indexes. They used a 24 months rolling window which counts for the variability of hedge fund risk exposures. However, the factor selection is *ad hoc* and is not based on any theoretical framework.

Kat and Palaro (2005) criticize factor-based approaches explaining that since we generally use portfolio of traditional assets indices, factor models for hedge funds explain only a small proportion of their returns. These authors propose another approach which relies on statistical tools in order to simply replicate the statistical properties of Hedge Fund returns without aiming to replicate month-to-month returns. For a further discussion on that point refer to Kat and Palaro (2005).

In this paper we are interested in the factor-based replicating approach. Our aim is to deal with the lack of actual replicating methods concerning the factor selection mechanism. We adopt a four-step dynamic approach which relies on the multifactor model framework. This paper proposes three main contributions. First, we use the recent econometric models of Bai and Ng (2002, 2003, 2006) to determine the number of latent factors that drive the co-

movements of hedge fund returns. We focus on the individual Equity Hedge Fund returns of the HFR Database. This step allows assessing the risk dimension of this strategy. The question we try to answer here is whether the equity-oriented risk exposure is multidimensional. We find that equity-oriented fund returns are driven by at least two latent factors. If the first one behave in a similar way as the stock indices, it is more difficult to understand which economic factors drive the behaviour of the second latent factor.

The second contribution concerns the factor selecting mechanism in order to replicate the fund's returns. Once the latent factors are estimated from the data, we use Bai and Ng (2006) to asses the adequacy between the latent and the observed factors. The question we try to answer here is which economic forces drive hedge fund returns. The two last steps use the observed factors selected in the second step to replicate hedge fund returns. The particularity of this approach is that it allows including in the replication process only the "useful" factors for a given period. We find that the hedge fund clone index constructed by our methodology behaves better than the "naïve" clone index constructed by a methodology consisting of an *ad hoc* factor selection, as in Lo and Hasanhobvic (2006).

The last contribution of this paper concerns the percentage of hedge fund returns we can really explain using factor-based replication. In general, the replicated portfolio is calibrated directly from the targeted hedge fund index returns. We prefer to use the set of individual hedge fund returns to infer the latent factors structure driving the index returns. As pointed out by Chan, Getmansky and Lo (2005), a disaggregated approach may yield additional insights not apparent from index-based risk models. Moreover, this approach allows us to get more information concerning the latent structure thanks to an important cross-sectional dimension while the time series dimension remains quite moderate. It is obvious that this type of asymptotic analysis is more in line with the dynamic properties we want to give to our approach.

The paper is organized as follows. Section 2 provides a literature review on multifactor asset pricing models and their wide application in the evaluation of stock returns. We discuss the possibilities to transpose some existing econometric models on the hedge fund framework. The models used in this paper and Monte Carlo simulations are described in Section 3. Section 4 discusses our methodology and our first empirical results. Section 5 concludes.

2. Literature review

A substantial part of the research effort in finance is directed toward improving our understanding of how investors value risky cash flows. Several capital asset pricing models have been suggested in the literature that attempt to assess the risk exposures that drive the covariation of asset returns. We can distinguish two main approaches. The first one is based on the arbitrage pricing theory framework developed by Ross (1976). Statistical methods, such as the factor analysis or the principal components analysis are used to estimate latent factors from the covariance structure of the data. The theory does not determine the number, or the nature of the latent factors. The most important drawback of this approach is the lack of economical interpretation of asset's risk exposures.

A more popular approach is to rely on intuition and theory as guides to select a set of observed variables as proxies of the unobserved theoretical factors. For example in the CAPM (Sharpe (1964), Linter (1965) and Black (1972)) analysis the equal-weighted and value-weighted market returns are used as proxies of the unobserved theoretical market factor. But if Fama and Macbeth (1973) validate the model empirically using the US stock market data, more recent studies highlight the lack of empirical support of the CAPM. For example, in their widely cited study, Fama and French (1992) present evidence of the inability of the CAPM to explain the stock returns. They pointed out the robustness of the size and the book-to-market effect. The empirical lack of one-factor approach motivates the development of multifactor asset pricing models.

Chen et al. (1986) found that the factors in APT are related to macroeconomic variables while Jagannathan and Wang (1996, 1998) use the return on labor income to improve the proxy of market portfolio and the credit spread as a proxy for the systematic risk instability. Perhaps the most well-known of observable risk factors are the three discussed in Fama and French (1993): the market excess return (MKT), small minus big factor (SMB), and high minus low factor (HML).

Both, statistical and observed factors are widely used to assess the risk exposures of equity returns. In fact, several academic studies³ bring empirical evidence on the ability of multifactor models to explain the cross-section of stock returns. Moreover the need to use

³ Fama and French (1993, 1995, 1996, 1998) for size and value factors, Jagannathan and Wang (1996, 1998) for economic factors such as the credit spread and the return on human capital, Chen et al. (1986) for macroeconomic variables, ...

multifactor models in assessing the risk of stock returns is also demonstrated by the commercial success of firms like BARRA, which provide beta estimates for risk management and valuation purposes, using elaborate time series models.

The growth of hedge fund industry this last two decades has reoriented the asset pricing efforts toward alternative returns offered by hedge funds. In fact, investors looking for alternative returns must be concerned by how do hedge funds managers deliver return characteristics that are different from the returns of the very asset classes they are trading. A significant gap has emerged between the expectations of institutional investors and the hedge-fund managers. Hedge fund managers rarely provide position-level transparency. Furthermore, they routinely impose lock-ups of one to three years and charge very important performance fees. These points raise the natural question of whether it is possible to obtain hedge-fund-like returns using liquid asset classes without investing in hedge funds. Another issue of hedge fund return replication is to construct a plausible benchmark providing a better evaluation of the alpha creation of a particular hedge fund.

An extensive literature has documented that hedge fund returns differ from the returns of the traditional asset classes. Mutual fund returns have high and positive correlation with asset class returns which suggest that they behave as a “buy and hold” strategy. Hedge fund returns seem to have low and sometimes negative correlation⁴ with asset class returns, which suggest that they behave as if deploying a dynamic strategy including short sells and leverage. For example, Fung and Hsieh (1997a, 1997b) pointed out the option-like characteristics for the returns of several hedge funds strategies, such as the trend followers and the merger arbitrage. Note that option-like feature is a possible explanation of the observed fat-tailed hedge fund return distribution.

However another, less documented, plausible cause of the fat-tailed return distribution is that the moments of the returns distribution are time-varying. Agarwal and Naik (2000c) examine the persistence of the hedge fund returns which can be directly linked to serial correlation. The serial correlation of hedge funds has also been studied by Lo and Marakov (2001) and Chan, Lo and Getamansky (2005). Thus, the replication of the hedge fund returns needs to adopt dynamic approaches that count for the dynamic risk exposures of the alternative returns. Different approaches have been developed.

⁴ Note that several recent studies have challenged the uncorrelatedness of hedge fund returns with market indexes, arguing that the standard methods of assessing their risks may be misleading. For example, Asness, Kraill and Liew (2001) show that in several cases where hedge funds purport to be market neutral, including both contemporaneous and lagged market returns as regressors and summing the coefficients yields significantly higher market exposure.

The approach proposed by Kat and Palaro (2005), uses option replication theory to dynamically allocate between a risky asset and cash. When a bank sell an option, with a fixed payoff function, she executes in the same time the dynamic trading strategy allowing it to hedge the risk related to this short option position. Kat uses the same type of reasoning in characterising the target fund return distribution by a payoff function of a single risky asset. The investment horizon is taken equal to the synthetic option maturity. Once this payoff function and the maturity are fixed, we only have to use the dynamic replicating strategy corresponding to this payoff to obtain at the maturity the fund return distribution. Finally, Kat argues that the sequence of returns is of no importance for investors as long as replication can give us a return distribution with the desired statistical properties, i.e. mean, variance, correlation with market indexes ...

Different approaches consisting in using options to replicate the hedge fund returns exist in the literature. Fung and Hsieh (2001) showed that the returns from trend following strategies can be replicated by a dynamically managed option-based strategy known as a “lookback option”. Mitchell and Pulvino (2001) modelled the return to merger arbitrage funds by using announced transactions from 1963 to 1998 to construct the return of a specific merger arbitrage strategy. Agarwal and Naik (2001) combine passive buy-and-hold strategies and option-based strategies to characterize the risks of different hedge fund strategy indexes. These replicating strategies count for nonlinearities in hedge fund returns using linear replicating methods. However, Amin and Kat (2003) point out that the replicating strategies involving derivatives seem difficult to be implemented in practice for two reasons: *“First, it is not clear how many options and which strike prices should be included... Second, since only a small number of ordinary puts and calls can be included, there is a definite limit to the range and type of non-linearities that can be captured”*.

For this reason, another issue of interest is to use linear portfolio of liquid assets – such as several traded indexes on stocks, bonds, commodities and interest rates. Fung and Hsieh (2004) showed that equity-oriented hedge fund indexes have two major exposures: the equity market as a whole⁵, and the spread between small cap and large cap stocks⁶. Thus, the multifactor approaches used to assess the risk exposures of equity returns are being transposed in a dynamic risk exposure framework. But to analyse the hedge fund returns, one has to count for the fact that their risk exposures are likely to change very frequently. For example, as pointed out by Fung and Hsieh (1997a), the trend follower returns are positively correlated

⁵ As proxied by S&P 500 index.

⁶ As proxied by the difference between the Wilshire Small Cap 1750 and the Wilshire Large Cap 750 index.

with the stock market in situations of bullish market and negatively correlated in bear markets.

Thus, to replicate hedge fund returns using liquid asset classes, it is important to employ dynamic replicating approaches in order to count for dynamic risk exposures of the alternative returns. Lo and Hasanhobvic (2006) use several observed factors – such as the S&P500 index, the USD return index, the Bond Index, ..., - in order to replicate the returns of more than 1610 individual hedge funds extracted from the TASS Database. They employ a dynamic approach which consists in estimating the risk exposures using a 24 months rolling window. This dynamic approach takes into account the variability of hedge fund risk exposures. Another important issue of their technique is the estimation of beta coefficients using individual hedge funds rather than hedge fund indexes. As pointed out by Chan, Getmansky and Lo (2005), a disaggregated approach may yield additional insights not apparent from index-based risk models. However, as discussed in the next section, the methodology employed by Lo and Hasanhobvic (2006) has drawbacks as well. The factor selection is rather *ad hoc* and does not count for the covariance structure of the fund returns.

3. Determining the number of factors : the hedge fund case

To answer the question whether or not a hedge fund strategy can be replicated, Lo & Hasanhodzic (2006) adopted an *ad hoc* methodology. They perform a time-series regression for each individual hedge fund of a given strategy, regressing the hedge fund's monthly returns on the whole set of the selected risk factors. Using the parameter estimates from these regressions, they construct – for each strategy of hedge funds – two types of clones: the fixed weighted clones (in-sample) and the rolling-window clones (out-of-sample). This methodology has two major drawbacks: (i) The authors use the same factors for each rolling window: this technique does not take into account the dynamic of risk exposures for different hedge fund strategies; (ii) The factor selection is *ad hoc* and does not directly count for the covariance structure of the individual hedge funds. Replicating hedge fund returns using factors with weak explanatory power may generate greater estimation error.

These critics are not new. For example Kat and Palaro (2005), supports their option-like replication approach using this kind of arguments. They argue that, even if we are using rolling windows to estimate the replication portfolio components, we still remain in a space spanned by traditional assets. The consequence is a high correlation between replicating

portfolios and traditional asset classes on the long run, and then no diversification benefits. Hence, only the replication of well diversified hedge fund indexes works and this approach fails to replicate more specialized hedge fund indices. More fundamentally, replication techniques can be seen as the “implementation of yesterday” as the underlying used in the replicating portfolio comes from historical estimation.

To deal with these remarks, we adopt a four-step dynamic approach to replicate the returns of the Equity Hedge Fund strategy. We use a T -month rolling window to construct linear clones for the N individual hedge funds of the given strategy. In the two first steps, we use recent asymptotic theories developed by Bai & Ng (2002, 2003, and 2006) for factor selection. These asymptotic models allow us to estimate the factor structure from the data and, then, to select the observed variables that match the best the latent factors for a given period⁷. Thus, the variables with weak explanatory power, for a given period, are excluded from the analysis. The third step consists in estimating the regression coefficients, for each individual hedge fund, using the $(T-1)$ first months of each rolling window. We regress the hedge fund returns on the set of the risk factors selected previously. The last step consists in replicating the individual hedge fund returns out-of-sample, using the regression coefficients estimated in the third step and the factor observations of month T of the rolling window.

3.1. Determining the number of latent factors from the covariance structure of the data

Let X_{it} be the observed return of the data for the i th cross-section unit at time t , for $i = 1, \dots, N$, and $t = 1, \dots, T$. Consider the following model which is generated by r common factors:

$$X_{it} = \lambda_i' F_t + e_{it} \quad (1)$$

In this equation, F_t is a $(r \times 1)$ vector of common factors, λ_i is a $(r \times 1)$ vector of factor loadings for the fund i , and e_{it} is the i th element of the t th column of the idiosyncratic component matrix (e). The idiosyncratic components are supposed to have a mean zero and covariance matrix Σ . Equation (1) is then the factor representation of the data. Note that the factors, their loadings, as well as the idiosyncratic errors are not observable.

⁷ The $(T-1)$ first months of each T -month rolling window.

Determining the number of factors in approximate factor⁸ models is an important issue when dealing with large panel data in both cross-sectional (N) and time-series (T) dimensions. The correct specification of the number of factors is crucial to both the theoretical and empirical validity of factor models. This parameter is often assumed rather than determined by the data. As Bai and Ng (2002) pointed out, even when trying to estimate r ⁹, the essential of classical factor analysis carries over to the case when one of two dimensions (N or T) goes to infinity while the other one is fixed. For example, Connor & Korajczyk (1993) developed a test under sequential limit assumptions, i.e., N converges to infinity with a fixed T and then T converges to infinity. Furthermore covariance stationary and homoscedasticity are crucial for the validity of their test. Stock and Watson (1998) showed that a modification of the BIC can be used to select the number of factors optimal for forecasting a single series¹⁰. Their criterion is restrictive not only because it requires $N \gg T$, but also because there can be factors that are pervasive for a set of data and yet have not predictive ability for an individual data series. Thus, their rule may not be appropriate outside of a forecasting framework. Cragg and Donald (1997) showed that these methods tend to perform poorly for moderately large N and T .

To deal with these problems, Bai & Ng (2002) develop an econometric theory for factor models of large dimensions. The focus is the determination of the number of latent factors. The authors first establish the convergence rate for the factors estimates that will allow for consistent estimation of r . They then propose some panel criteria and show that the number of factors can be consistently estimated using the criteria. The determination of factors is set up as a model selection problem. In consequence, the proposed criteria depend on the usual trade-off between good fit and parsimony. However, the problem considered by Bai and Ng (2002) is not standard not only because account needs to be taken of the sample size in both the cross-sectional and time-series dimensions, but also because the factors are not observed. They demonstrate that the penalty for overfitting must be a function of both N and T in order to consistently estimate the number of factors. Consequently the usual AIC and BIC criteria, which are functions of N or T alone, do not work when both dimensions of the panel are large. Their theory does not rely on sequential limits and the results hold under

⁸ Approximate factor models relax the hypothesis of iid normally distributed idiosyncratic errors to allow for serial and cross-section correlations: for example, an approximate factor model in the sense of Chamberlain and Rothschild (1983) allows for the presence of cross-section correlation on the idiosyncratic errors.

⁹ Recall that r represents the number of the latent factors that drive the covariation structure of the data.

¹⁰ Assuming that $N, T, \sqrt{N}/T \rightarrow \infty$.

heteroscedasticity and weak cross-section and serial correlations¹¹ in the idiosyncratic components. The latent factors are estimated by the method of asymptotic principal components (PCA)¹² on the $T \times T$ covariance matrix since the $N \times N$ problem can be turned into a $T \times T$ problem, as noted by Connor and Korajczyk (1993) and others¹³.

We focus on the criteria proposed by Bai & Ng (2002) to determine the number of latent factors that better describe the covariance structure of our data. As we will see later, the simulation results show that these criteria yield more efficient results than AIC and BIC. Recall that Bai and Ng (2002) focus on large N and T panels. Their simulations yield good results for large panel data in both dimensions. However, our study focuses on a dynamic replicating approach which consists of using a T -month rolling window. We have to deal with the problem of choosing the minimal length of the rolling window (T) ensuring the convergence of the estimated parameters. Thus, T depends on the trade-off between a high dynamics of our approach and good finite sample properties of the criteria. If T is too small the convergence is not achieved and the selected criteria will not yield good estimates of the number of latent factors. If T is too large, our approach will lose its dynamic character. Furthermore, Bai and Ng (2002) show that the criteria proposed in their paper do not behave in the same way as N and T vary simultaneously.

In order to observe the behaviour of different criteria as N and T vary (especially in cases of large N but moderately large T), allowing for the presence of both, cross-section and serial correlations in the idiosyncratic errors, we performed Monte Carlo simulations. Following Bai and Ng (2002), we simulate data from the following model¹⁴:

$$X_{it} = \sum_{j=1}^r \lambda_{ij} F_{jt} + \sqrt{\theta} e_{it} = c_{it} + \sqrt{\theta} e_{it} \quad (2)$$

where the idiosyncratic errors are generated by the equation (3) to allow for serial and cross-section correlation.

¹¹ The allowance for weak cross-section correlation in the idiosyncratic components leads to the approximate factor structure of Chamberlain and Rothschild (1983). It is more general than a strict factor model which assumes the idiosyncratic errors are uncorrelated across different funds.

¹² When N is small the parameters are estimated by maximum likelihood (under normality assumption): see, for example, Stock and Watson (1989). For large N , the drawback of maximum likelihood estimation is that because the numbers of parameters to be estimated increases with N , computational difficulties make it necessary to abandon information on many series even though they are available (Gregory, Head and Raynauld, 1997). Principal component analysis (PCA) allows the estimation of a number of factors [$\min(T, N)$] that is much larger than that permitted by ML method: Connor and Korajczyk (1986, 1988) considered the method for fixed T and Stock and Watson (1998) for large T .

¹³ Cf. Bai & Ng (2002, 2003, 2004, and 2006).

¹⁴ All computations are performed using Matlab Version 7.0. The programs used for Monte Carlo simulations and test statistic computations are available upon request.

$$e_{it} = \rho e_{it-1} + v_{it} + \sum_{j \neq 0, j=-J}^J \beta v_{i-jt} \quad (3)$$

The factors are $T \times r$ matrices of $N(0, 1)$ variables, and the factor loadings are $N(0, 1)$ variables. Hence, the common component of X_{it} , denoted by c_{it} , has variance r . Our model assumes that the idiosyncratic component has the same variance¹⁵ as the common component (i.e. $\theta = r$). The parameters ρ and β represent respectively the serial and the cross-section correlation parameters. Following Bai and Ng (2002), we set $\rho = 0.50$, $\beta = 0.20$ and J ¹⁶ to $\max\{N/20, 10\}$.

We consider fifteen configurations of the data. The first five simulate plausible asset pricing applications with two years of monthly data ($T = 24$) on 100 to 300 asset returns. We then increase T to 36 months. The last five configurations are more general and are used as a reminder of the results obtained by Bai and Ng (2002). We estimate the common factors by the PCA method and then use different criteria to estimate the number of common factors (noted by \hat{r}). We considered the criteria proposed by Bai and Ng (2002) given as follows:

$$PC1(k) = V(k, \tilde{F}^k) + k \hat{\sigma}^2 \left(\frac{N+T}{NT} \right) \ln \left(\frac{NT}{N+T} \right);$$

$$PC2(k) = V(k, \tilde{F}^k) + k \hat{\sigma}^2 \left(\frac{N+T}{NT} \right) \ln C_{NT}^2;$$

$$PC3(k) = V(k, \tilde{F}^k) + k \hat{\sigma}^2 \left(\frac{\ln C_{NT}^2}{C_{NT}^2} \right);$$

$$IC1(k) = \ln(V(k, \tilde{F}^k)) + k \left(\frac{N+T}{NT} \right) \ln \left(\frac{NT}{N+T} \right);$$

$$IC2(k) = \ln(V(k, \tilde{F}^k)) + k \left(\frac{N+T}{NT} \right) \ln C_{NT}^2;$$

$$IC3(k) = \ln(V(k, \tilde{F}^k)) + k \left(\frac{\ln C_{NT}^2}{C_{NT}^2} \right);$$

¹⁵ Bai and Ng (2002) also performed simulations allowing for the variance of the idiosyncratic component to be larger than that of the common component and yield similar results for the finite sample properties of their criteria.

¹⁶ J is the number of the cross-correlated idiosyncratic errors.

where $C_{NT} = \min(\sqrt{N}, \sqrt{T})$, \tilde{F} is used to denote the estimated common factors by PCA method, $V(k, \tilde{F}^k) = N^{-1} \sum_{i=1}^N \hat{\sigma}_i^2$, $\hat{\sigma}_i^2 = \tilde{e}_i \tilde{e}_i' / T$ and $\hat{\sigma}^2 = V(k \max, \tilde{F}^{k \max})$. As in Bai and Ng (2002), we set¹⁷ $kmax = 8$.

Reported in Table I and II are the averages of the estimated number of factors over 1000 replications, for $r = 2$ and 3 respectively. While the presence of cross-section and serial correlation reduces the precision of the estimates somewhat, the results generally confirm that a small degree of correlation in the idiosyncratic errors will not affect the finite sample properties of the estimates. The IC criteria are more robust than PC criteria for different model configurations. For example, for $r = 2$, the PC criteria overestimate – on average – the number of factors up to 5 for $T = 24$ and up to 4 for $T = 36$. On the other hand, the IC criteria (especially the IC1 and the IC2) infer – on average – 2 factors when T is at least 36.

For small T ($T = 24$) the six criteria lose their precision, even for large N . For $T = 36$, IC criteria allows, on average, to infer the number of common factors used to generate the data. When T and N are both small, the criteria are no more efficient so that they can not be efficiently used when dealing with small sample data. For example, for $N = 30$ and $T = 40$, both sets of criteria overestimate – on average – the number of latent factors. Finally, in all cases considered here, the PC and IC criteria are more robust than the classical AIC and BIC criteria¹⁸.

3.2. Matching the macroeconomic variables with latent factors

The factors are the common shocks that underlie the covariation of asset returns. Several authors have replaced the unobserved factors with statistically estimated ones. For example, Lehman and Modest (1988) used factor analysis, while Connor and Korajczyk (1998) adopted the method of principal components. The drawback is that the statistical factors do not have immediate economic interpretation.

The question of interest is whether some observable economic variables are in fact the underlying observed factors. A more popular approach is to select a set of observed variables

¹⁷ The first three criteria are called Panel Criteria because they are similar to those developed by Mallows (1973) in a strict time-series or cross-section model framework. The last three criteria are called Information Panel Criteria. The main advantage of these three information criteria is that they do not depend on the choice of $kmax$ through $\hat{\sigma}^2$, which could be desirable in practice.

¹⁸ In presence of cross and serial correlation, AIC and BIC criteria tend to overestimate the number of the latent factors to 8. Our results concerning the AIC and BIC criteria are similar to those of Bai and Ng (2002) and not reported here.

as proxies of the unobserved theoretical factors. For example in the CAPM analysis the equal-weighted and value-weighted market returns are used as proxies of the unobserved theoretical market factor. Chen et al. (1986) found that the factors in APT are related to macroeconomic variables. Perhaps the most well-known of observable risk factors are the three discussed in Fama and French (1993): the market excess return (MKT), small minus big factor (SMB), and high minus low factor (HML).

There is a certain appeal in associating the latent factors with the observed variables as this facilitates the economic interpretation. But as pointed out by Shanken (1992), estimation of betas using proxy factors is meaningful only if the fundamental factors are spanned by the observed factors. Yet such a condition will be violated even if a pure measurement error is added to a perfect proxy. The problem is not so much that the common factors are unobserved because in principle, if we observe indicators of the factors, we can estimate the factors from the data. As discussed in Bai and Ng (2002, 2003, 2004 and 2006), the problem is that latent factors estimated from a small number of indicators are imprecise, and in theory, consistent estimation of the latent factors cannot be achieved under the traditional assumption that T is large and N fixed, or vice versa.

Bai & Ng (2006) consider statistics to compare the observed variables with estimates of the unobserved factors. The key to their analysis is that the space spanned by the latent factors can be consistently estimated when the sample size is large in both the cross-section and the time series dimensions¹⁹. Their analysis combines the statistical approach of Lehmann and Modest (1988) and Connor and Korajczyk (1998), with the economic approach of using observed variables as proxies.

In this paper, we implement the tests developed in Bai and Ng (2006) to assess the adequacy between the latent factors and the observed variables. Only the variables that seem to span the same space as the common factors estimated from the data are taken into account in the replication process.

Let consider the model representation for a panel of data X_{it} ($i = 1, \dots, N, t = 1, \dots, T$), given by equation (1):

$$X_{it} = \lambda_i' F_t + e_{it}$$

¹⁹The rate of consistency of the estimated latent factors was studied by Bai & Ng (2002). The rate of convergence and the limiting distributions for the estimated factors, factor loadings and common components, estimated by the principal component method (PCA) was developed by Bai & Ng (2003).

where F_t is a $(r \times 1)$ vector of common factors, λ_i is a $(r \times 1)$ vector of factor loadings for the unit i , and e_{it} is the idiosyncratic component of X_{it} . As opposed to classical factor analysis, Bai and Ng (2006) consider that both N and T go to infinity, and that the idiosyncratic errors are serially and cross-sectional correlated rather than i.i.d. over t and independent²⁰ over i . These assumptions match much better the properties of hedge fund returns, i.e., we dispose a large number of cross-section units with moderately large history data and allowing for two-dimensional residual correlations.

Suppose we observe G_t , an $(m \times 1)$ vector of economic variables. We want to know if its m elements are generated by (or are linear combinations of) the r^{21} latent factors, F_t . It is of particular interest to know if a given G_{it} ($i = 1, \dots, m$), is in fact a common factor. In this paper, we consider three particular statistics proposed by Bai and Ng (2006) to assess the adequacy of the estimated latent factors with each observed factor at one time²². These statistics are given respectively by equations (4), (6) and (7).

$$A(j) = \frac{1}{T} \sum 1^* (|\hat{\tau}_t(j)| > \phi_\alpha), \quad (4)$$

In this equation, $\tau_t(j) = \frac{(\hat{G}_{jt} - G_{jt})}{(\text{var}(\hat{G}_{jt}))^{1/2}}$ where a consistent estimate of $\text{var}(\hat{G}_{jt})$ is given by $\frac{1}{N} \hat{\gamma}_j' \tilde{V}^{-1} \tilde{\Gamma}_t \tilde{V}^{-1} \hat{\gamma}_j$. To allow for the serial correlated errors, $\tilde{\Gamma}_t$ is given²³ as follows:

$$\tilde{\Gamma}_t = \frac{1}{N} \sum_{i=1}^N \tilde{e}_{it}^2 \tilde{\lambda}_i \tilde{\lambda}_i' \quad (5)$$

Proposition²⁴ 1 in Bai and Ng (2006) says that, under the null hypothesis that G_{jt} is an exact factor ($G_{jt} = \delta' F_t$) and as N, T go to infinity, $A(j) \xrightarrow{p} 2\alpha$. Thus²⁵, if $\alpha = 2,5\%$, then $A(j)$ is the frequency that $|\hat{\tau}_t(j)|$ exceeds the 5% asymptotic critical value of a standard normally distributed variable ($\phi_{2,5\%} = 1,96$).

²⁰ Following the classical factor analysis the covariance matrix of the idiosyncratic errors must be diagonal.

²¹ In general r is an unknown parameter which can be estimated by the IC or Pc criteria proposed by Bai and Ng (2002).

²² Bai and Ng (2006) also developed some econometric statistics testing the adequacy of observed factors as a set. The results of these statistics will be provided in a future version of this working paper.

²³ This equation counts for the serial correlation of the idiosyncratic components (Bai and Ng, 2006).

²⁴ For further discussion of the proposition 1 and its proof refer to Bai and Ng (2006).

²⁵ Remember that tildes are used to denote the estimated parameters from the data.

But, as discussed in Bai and Ng (2006), requiring that G_{jt} be an exact linear combination of the latent factors is rather strong. An observed series might match the variations of the latent factors very closely, and yet is not an exact factor in a mathematical sense. Measurement error and time aggregation, for example, could be responsible for deviations between the observed variables and the latent factors, as discussed in Breeden et al. (1989). For that reason, we consider two more tests (named approximate tests) developed by Bai and Ng (2006). These tests allow for an approximate relation between the observed and the latent factors given by: $G_{jt} = \delta' F_t + \varepsilon_{jt}$.

$$NS(j) = \frac{\hat{\text{var}}(\hat{\varepsilon}(j))}{\hat{\text{var}}(\hat{G}(j))} \quad (6)$$

$$R^2(j) = \frac{\hat{\text{var}}(\hat{G}(j))}{\hat{\text{var}}(G(j))} \quad (7)$$

The $NS(j)$ statistic is simply the noise-to-signal ratio. If G_{jt} is an exact factor, the population value of $NS(j)$ is zero. A large $NS(j)$ thus indicates important departures of G_{jt} of the latent factors. But, as pointed out by Bai and Ng (2006), this statistic leaves open the question of what is small and what is large. For this reason, we also consider $R^2(j)$. The higher is this statistic, the higher is the adequacy between G_{jt} and the latent factors. Remember that tildes are used to denote the estimated parameters.

We performed Monte Carlo simulations to assess the finite sample properties of the tests for the same data configurations as previously ($T = 24, 36$ and $N = 100, 150, 200, 260, 300$). The statistics considered here are $A(j)$ (exact test), $NS(j)$ and $R^2(j)$ (approximate tests). We assume $F_{kt} \sim N(0, 1)$, $k = 1, \dots, r$, and $e_{it} \sim N(0, \sigma_e^2(i))$, where e_{it} is uncorrelated²⁶ with e_{jt} for $i \neq j, i, j = 1, \dots, N$. The factor loadings are standard normal, i.e. $\lambda_{ij} \sim N(0, 1)$, $j = 1, \dots, r, i = 1, \dots, N$. The data are generated as $X_{it} = \lambda_i' F_t + e_{it}$. As in Bai and Ng (2006), we assume that there are $r = 2$ factors and that this is known²⁷. The data are standardized to have mean zero and unit variance prior to estimation of the factors by the method of principal components.

²⁶ Monte Carlo simulations counting for cross section and serial correlation in the idiosyncratic component will be available in a future version of this paper.

²⁷ In finite samples, the criteria developed in Bai and Ng (2002) for selecting the number of factors are excellent even under heteroscedasticity, mild, weak and cross-correlation.

The observed factors are generated as $G_{it} = \delta_j' F_t + \varepsilon_{jt}$, where δ_j' is a $r \times 1$ vector of weights, and $\varepsilon_{jt} \sim \sigma_\varepsilon(j)N(0, \text{var}(\delta_j' F_t))$. As in Bai and Ng (2006), we test $m = 7$ observed variables parameterized as follows:

Parameters for G_{jt}

J	1	2	3	4	5	6	7
δ_{j1}	1	1	1	1	1	1	0
δ_{j2}	1	0	0	1	0	1	0
σ_ε	0	0	0.2	0.2	2	2	1

The first two factors, G_{1t} and G_{2t} , are exact factors since $\sigma_\varepsilon = 0$. Factors three to six are linear combinations of the two latent factors but are contaminated by errors. The variance of this error is small relative to the variations of the latent factors for G_{3t} and G_{4t} , but is large for G_{5t} and G_{6t} . Finally, G_{7t} is an irrelevant factor as it is simply a random variable $N(0, 1)$.

The Monte Carlo simulation results are reported in Table III. The test statistics are averaged over 1000 simulations. Concerning the $A(j)$ statistic, we set $\alpha = 0.025$. According to the theory, $A(j)$ should be 2α if G_{jt} is a true factor and unity if the factor is irrelevant. Indeed, for G_{1t} and G_{2t} , the $A(j)$ statistic is close to 0.05, even when T is set to 24. For the irrelevant factor G_{7t} , the test rejects the null hypothesis with high probabilities. The $NS(j)$ and $R^2(j)$ statistics reinforce the previous result. When the observed factors are contaminated by errors, Table III shows that the higher is the variance of this error, the worst is the efficiency of the tests considered here. Finally, the test precision is higher for $T = 36$ than for $T = 24$. This result confirms the asymptotic properties of the tests discussed in Bai and Ng (2006).

3.3. Estimation of risk exposures and replication of hedge fund returns

The two previous steps help us identify and select the observed factors that explain the best the common variations of the data. Recall that this approach allows us to take into account only the most significant risk exposures, excluding from the analysis, for each rolling

window, the useless factors²⁸. Let denote m and g respectively, the total number of the observed factors and the number of the factors selected for each rolling period. We use the selection matrix – denoted S – to select only g out of the m factors. These factors will then be used to replicate the individual Equity-Hedge fund returns for a given rolling period.

. Thus, to estimate the regression coefficients, we perform the least-squares regressions as follows:

$$R_{it} = \beta_{i1}G_{1t} + \dots + \beta_{ig}G_{gt} + \varepsilon_{it}, \quad t = 1, \dots, T$$

We use R_{it} to denote the return of fund i in t , G for the observed factors and g the number of the selected observed factors for a given T -month rolling window.

The estimated regression coefficients $\{\beta_{ig}^*\}$ are then used as portfolio weights for the g factors. Hence, the replicated returns for the fund i are the equivalent to the fitted values $\{R_{it}^*\}$ of the regression equation.

$$R_{it}^* = \beta_{i1}^*G_{1t} + \dots + \beta_{ig}^*G_{gt}$$

We can now reformulate the question of whether or not the Equity Hedge Fund strategy can be cloned as a question about how much of a hedge fund's expected return is due to risk premia from identifiable factors. Another important question is whether or not our replication method using the factor selecting matrix S , outperforms the “naïve” replication strategy which consists to including in the replication process all the observed factors (even the useless ones). The behaviour of the time series of different clones constructed by both methods is discussed in the next section.

4. Empirical applications

In this paper we investigate the characteristics of a sample of individual hedge funds drawn from the HFR database. The database contains only the funds that are still “alive”. These funds are considered to be active as of the end of our sample period, December 2005. We acknowledge that the database suffers from the survivorship bias. However, the importance of such a bias for our application is tempered by the fact that many successful funds leave the sample as well as the poor performers, reducing the upward bias in expected returns. In particular, Fung and Hsieh (2000) estimate the magnitude of survivorship bias to be 3.00% per year, and Liang's (2000) estimate is 2.24% per year. Furthermore, the focus of

²⁸ Factors that do not satisfy the selecting criteria of step 2 are called here, “useless” in the sense of Bai and Ng, 2006.

our study is on the relative performance of hedge funds versus relatively passive portfolios of liquid securities, and as long as our cloning process is not selectively applied to a peculiar subset of funds in the HFR database, any survivorship bias should impact both funds and clones identically, leaving their relative performances unaffected²⁹. HFR database classifies funds into one of 17 different investment styles, listed below.

Fund repartition by strategy for the HFR Database at December 2005.

	<i>Strategy</i>	<i>Fund Number</i>
1	Convertible Arbitrage	109
2	Distressed Securities	127
3	Emerging Markets	269
4	Equity Hedge	1232
5	Equity Market Neutral	282
6	Equity Non-Hedge	146
7	Event-Driven	225
8	Fixed Income	310
9	Foreign Exchange	68
10	Fund of Funds	2011
11	Macro	277
12	Managed Futures	337
13	Market Timing	25
14	Merger Arbitrage	46
15	Relative Value Arbitrage	268
16	Sector	279
17	Short Selling	23
	<i>Total</i>	<i>6034</i>

We limit our analysis on the individual funds of a single strategy: the Equity Hedge. We focus on this strategy for two reasons. First, because this strategy involves quite homogenous equity-oriented funds investing on both the long and the short sides of the market. Thus, we will expect the Equity Hedge funds to be more sensible to equity-based risk factors included in our analysis. Second, the number of funds (N) disposing full data for our sample period is large, which will improve the finite sample properties of our tests.

We limit our analysis to the sample period from January 1997 to December 2005 because this is the timespan for which we have full data for a large number of Equity Hedge Funds. Recall that large N is needed to improve finite sample properties of the tests for factor selection, as described in the previous section. Of these funds, we drop those that: (i) do not

²⁹ For a further discussion on how to deal with the survivorship bias refer to Lo and Hasanhodzic (2006).

report net-of-fee returns; (ii) report returns in currencies other than the U.S. dollar; (iii) report returns less frequently than monthly; (iv) have at least \$10 Millions of assets under management . These filters yield a final sample of 680 Equity Hedge Funds. The summary statistics for our sample data are reported below.

Summary statistics for our sample data (680 Equity Hedge Funds)

	Annualized Mean	Annualized SD	First-order Autocorelation	Ljung-Box p-Value
Mean	15,7%	12,2%	9,2%	32%
SD	14,5%	7,9%	16,5%	29%

An important feature of the data is the positive average return-autocorrelation, which remains significant even for an equity-oriented hedge fund strategy. Lo and Getamansky (2001), Lo and Marakov (2004) have shown that such high serial correlation in hedge fund returns is likely to be an indication of illiquidity exposure³⁰.

4.1. Estimating the number of latent factors that drive the Equity Hedge Fund returns

Using our sample data, we perform the asymptotic tests developed in Bai and Ng (2002, 2006) which are exposed in the previous section. The first step of our methodology consists of using the Bai and Ng (2002) criteria to select the number of latent factors estimated from the data.

To asses the finite sample properties of these criteria, we performed Monte Carlo simulations as described in the previous section. To choose the most efficient criterion given our data configuration, we consider the simulation results shown in Table I and II for two reasons. First, in presence of cross-sectional and serial correlation in the idiosyncratic errors, the IC criteria conserve their good finite sample properties while the PC criteria appear to be less efficient. These results confirm those reported by Bai and Ng (2002). From the three IC criteria, the IC2 seems to behave better than the two others. Table I and II show that IC2 yields, in general³¹, 2 factors for $r = 2$, and 3 factors when $r = 3$. That is why we use this criterion in our analysis to estimate the number of latent factors that describe the best the

³⁰ As Lo and Hasanhodzic (2006) pointed out, this is a legitimate and often lucrative source of expected return, but illiquidity exposure is typically accompanied by capacity limits.

³¹ Exception is made for the case of small T (T=24).

covariance structure of our data. Second, Table I and Table II show that for $T = 24$, the six criteria lose their finite sample properties and the estimation of the number of latent factors became less efficient, even in presence of large cross-section dimensions. For example, when the data is generated by three factors ($r = 3$) all the criteria considered here (even the IC criteria) tend to overestimate the number of factors when $T = 24$ months. For $T = 36$ months, the IC criteria behave better and infer quite efficiently the number of the factor structure. These results motivate us to use, at least, a 36-month rather than a 24-month rolling window.

The length of the entire sample allows us to form 72 rolling windows of length 37 months for each one. The first rolling window goes from January 1997 to January 2000, the second from February 1997 to February 2000, ..., the last one extends from December 2002 to December 2005. The first 36 months³² of each rolling window are used to perform the factor selecting tests and to estimate the beta coefficients, while the last (the 37th) observation is used to replicate the fund's returns.

For each rolling window, we drop the funds that do not have full data for the given period. This allows us to count progressively for the new funds that enter in the database. The size of our sample³³ varies from 97 for the first rolling window to 388 for the last one. Large N , ensures having good finite sample properties for the asymptotic tests used in this paper.

The first 36 months of each rolling window are used to estimate the factor structure by the principal component method. The fund's returns are standardized previously within the 36 first months of each rolling period. Throughout, we use "tilde" to denote the principal component estimates. Let X be the T by N matrix of the Equity Hedge returns of our sample data such that the i th column is the time series of the i th cross section. Let \tilde{V} be a $r \times r$ diagonal matrix consisting of the r largest eigenvalues of XX'/NT . Let $\tilde{F} = (\tilde{F}_1, \dots, \tilde{F}_T)'$ be the principal component estimates of F under the normalization that $\frac{F'F}{T} = I_r$. Then \tilde{F} is comprised of the r eigenvectors (multiplied by \sqrt{T}) associated with the r largest eigenvalues of the matrix XX'/NT in the decreasing order. Let $\Lambda = (\lambda_1, \dots, \lambda_N)'$ be the matrix of factor loadings. The principal component estimator of Λ is $\tilde{\Lambda} = X'\tilde{F}/T$. By definition, $\tilde{e}_{it} = X_{it} - \tilde{\lambda}_i' \tilde{F}_t$.

³² The length of our rolling window seems to be a good compromise between the dynamic feature of our replication approach and good finite sample properties of the asymptotic tests considered here.

³³ More detailed information concerning the evolution of the number of funds during the entire sample period is available upon request.

Using the principal component estimates, we calculate the six criteria proposed by Bai and Ng (2002) to estimate the number of latent factors³⁴. Three important remarks can be done. First, the IC criteria seem to yield more stable estimations than PC criteria. The IC2 criterion is the one we use to estimate the number of latent factors while performing the tests of the next step. Second, the number of the estimated factors provided by IC2 criterion (denoted \hat{r}) varies between 2 and 4, which seems to be quite realistic for this kind of strategy. Finally, if we refer to IC2 criterion, the number of latent factors is always more than one, which highlights the lack of the CAPM in explaining the hedge fund returns even for an equity-oriented strategy.

Thus, the Equity-Hedge Fund returns are exposed to a multidimensional risk. Figure I and Figure II match the two first estimated latent factors with the two equity indexes used in our analysis: the S&P500 and the Russell 2000. It is obvious that the first latent factor returns behave closely with each of the two equity indexes, while the second seems to be uncorrelated with the equity market. Thus, even if the equity market factor seems to play an important role in explaining the cross-section of Equity Hedge Fund returns, we still are missing an important part of risk exposure represented by the second estimated latent factor. This raises the question of how to explain the risk dimension left unexplained by the market factor. This work is left to the next step of our methodology which consists in assessing the adequacy between the observed variables and the estimated latent factors.

4.2. Matching the observed risk factors with the estimated latent variables

Once the latent factors \tilde{F}_t and their number \hat{r} estimated, we have implemented tests proposed by Bai and Ng (2006) in order to match the m observed risk factors with the \hat{k} estimated latent variables. We considered the following ten factors ($m = 10$): (1) USD: the U.S. Dollar Index return; (2) BOND: the return on the Moody's Bond Index Corporate AA; (3) CREDIT: the spread between the Moody's BAA Corporate Bond Index return and the US Government 10-year yield; (4) S&P500: the S&P 500 total return; (5) Russell 2000: the Russell 2000 total return; (6) CMDTY: the Goldman Sachs Commodity Index (GSCI) total return; the three factors of Fama et French provided by the web site of Kenneth French: (7) the MKT : the return of the portfolio containing all the stocks of CRSP³⁵; (8) SMB (small

³⁴ The results for each of 72 rolling windows are available upon request.

³⁵ Centre for Research in Security Prices of the University of Chicago.

minus big): the spread between small and big capitalizations; (9) HML (high minus low) : the spread between high and low Book-to-Market stocks; and (10) MOM: the short-term reversal factor (momentum) provided by the web site of Kenneth French. The last four factors are not traded liquid benchmarks. We include them in our analysis only for academic purposes. For all the factors we dispose full data for the sample period.

After having determined the number of latent factors (\hat{r}), we can construct the tests proposed by Bai and Ng (2006). We consider here only the statistics that allow testing each observed factor (denoted G_i) at one time: the $A(j)$ for the exact tests and $NS(j)$ and $R^2(j)$ for the approximate tests. These three statistics are computed³⁶ using respectively equations (4), (6) and (7). Recall that, for each rolling period, the factors, as well as the fund's returns are standardized using the first 36 months.

Concerning the $A(j)$ statistic, we set $\alpha = 0.025$. Recall that according to the theory, $A(j)$ should be 2α if G_{jt} is a true factor and unity if the factor is irrelevant. The results we obtained show that the probability to reject the null hypothesis of exact factors, exceeds 0.05 in most cases³⁷. Thus, even if the market indexes seem to match quite efficiently the latent factors, they fail to be exact factors³⁸. However, $A(j)$ test seems to be too strict. What interests to us is to choose among the ten candidates, the factors that match the better the covariance structure of our data.

For this reason, we turn our attention toward the NS ³⁹ criterion to select the factors to include in the replication process. Given the results of Table III, we set the critical value of $NS(j)$ to 10^{40} . Thus the observed factors having a $NS(j)$ value less than 10 for a given rolling period are included in the analysis. Concerning the three equity-oriented factors (i.e. S&P 500, Russell 2000 and MKT) the $NS(j) < 10$. From these three factors, only the S&P 500 Index is included in the replication process. At the end, this procedure yields the selecting matrix, S reported in Table IV. Each of the 72 rows of this matrix corresponds to a particular rolling window. The factors to be included in the analysis for a given period are those having $NS(j) < 10$. Each element S_{jt} ($j = 1, \dots, m$, and $t = 1, \dots, 72$) equals to one if the j^{th} factor

³⁶ The test results for each of the 72 rolling windows can be available upon request.

³⁷ For example, for non-equity market factors, like USD, GSCI, CREDIT, BOND $A(j)$ statistic is close to unity showing that these factors fail to be exact factors. Concerning SMB, HML and MOM, $A(j)$ even if less important, still exceeds the 0.05 critical value. Even for the equity indexes (Russell 2000, S&P 500 and MKT) the $A(j)$ value rarely respect the upper bounder of 0.05.

³⁸ Even for the equity indexes (Russell 2000, S&P 500 and MKT) the $A(j)$ value rarely respect the upper bounder of 0.05.

³⁹ We also considered the R^2 statistics and obtained similar results.

⁴⁰ We also used other critical values but they yield similar results.

satisfies the restriction imposed on the $NS(j)$ criterion for the t^{th} rolling period, and zero otherwise.

The selecting matrix showed in Table IV highlights three important things concerning the risk exposures of the Equity Hedge Fund returns: (i) the equity index that seems to match the latent factor the most efficiently is the MKT factor of Fama and French and the Russell 2000. These two indexes are larger than S&P 500 and seem to be better proxies of the market factor; (ii) SMB, HML and MOM factors are rarely excluded from the analysis. HML factor have the lowest $NS(j)$ values after the equity indexes; (iii) factors like the USD or GSCI seems to be irrelevant for the most cases while the CREDIT factor becomes relevant later, at the beginning of 2000.

The aim of the next step is to replicate the Equity Hedge Fund returns using the observed factors selected by S , for each rolling period.

4.3. “Linear clone” construction and replication results

Our analysis uses rolling-periods to construct clones. This method seems more practically relevant because it avoids the most obvious forms of look-ahead-bias. To construct a rolling-window linear clone for fund i , for each month t ($t = 1, \dots, 72$), we use a 37-month rolling window. The first 36 months are used to construct different tests for factor selecting and to estimate the regression coefficients for the selected factors. Thus, we regress the fund’s returns on g of the m factors selected by S for each rolling period.

$$R_{i,t-h} = \beta_{i1} G_{1t-h} + \dots + \beta_{ig} G_{gt-h} + \varepsilon_{i,t-h}, \quad h = 1, \dots, 36$$

Note that the coefficients are indexed by both i and t since we repeat this process each month for every fund i . We use R_{it} to denote the return of fund i in t , G for the observed factors and g the number of the selected observed factors for a given rolling window.

Following Lo and Hasanhodzic (2006), we omit the intercept because our objective is to estimate a weighted average of the factors that best replicates the fund’s returns. Dropping the constant term forces the least-squares algorithm to use the factor means to fit the mean of the fund, an important feature of replicating hedge-fund expected returns with factor risk premia. We do not constrain the beta coefficients to sum to one, as Lo and Hasanhodzic (2006) do for a better interpretation of the weights.

The estimated regression coefficients $\{\beta_{ig}^*\}$ are then used as portfolio weights for the g factors, hence the replicated returns for the fund i are the equivalent to the fitted values $\{R_{it}^*\}$ of the regression equation. However, as in Lo and Hasanhodzic (2006), we implement an additional renormalization so that the resulting replicated return $\{\hat{R}_{it}\}$ has the same volatility as the original fund's return series:

$$R_{it}^* = \beta_{it1}^* G_{1t} + \dots + \beta_{itg}^* G_{gt}$$

$$\hat{R}_{it} = \gamma_{it} R_{it}^* \quad \gamma_{it} \equiv \frac{\sqrt{\sum_{h=1}^{36} (R_{it-h} - \bar{R}_{it})^2 / 35}}{\sqrt{\sum_{h=1}^{36} (R_{it-h}^* - \bar{R}_{it}^*)^2 / 35}}$$

$$\bar{R}_{it} = \frac{1}{36} \sum_{h=1}^{36} R_{it-h}, \quad \bar{R}_{it}^* = \frac{1}{36} \sum_{h=1}^{36} R_{it-h}^*$$

Following the authors, the motivation for this renormalization is to create a fair comparison between the clone and the fund by equalizing volatilities. Renormalizing is equivalent to changing the leverage of the clone portfolio, since the sum of the renormalized betas $\gamma_i \sum_g \beta_{ig}^*$ will equal the renormalization factor, not one. If the renormalization factor exceeds one, then positive leverage is required, and if less than one, the portfolio is not fully invested in the g factors. Note that the renormalization factors γ_{it} are indexed by time t to reflect the fact that they are also computed within the rolling window. This implies that for a given clone i , the volatility of its returns over the entire history will not be identical to the volatility of its matching fund because the renormalization process is applied only to rolling windows, not to the entire history of returns. However, as pointed out by Lo and Hasanhodzic (2006), as volatilities do not shift dramatically over time, the rolling-window renormalization process should yield clones with similar volatilities.

The results yielded from our four-step dynamic approach are compared to those of a “naïve” replicating strategy which consists of including in the regression analysis the whole set of the observed factors. To facilitate the comparison, we construct the track of returns yielded by an equally-weighted portfolio of Equity Hedge Fund dynamic clones constructed by both methodologies considered here. We denote IDS and INS the monthly returns of a “dynamic factor selecting clone index” and a “naïve factor selecting clone index” respectively. The former is the clone constructed by our methodology and the latter is the one

constructed in a similar way as Lo and Hasanhobvic (2006). Both clone indexes are compared to the Equity Hedge equally weighted index containing all the funds of our sample. Figure III shows the cumulated returns for the three indexes concerning the replicating⁴¹ period which extends from January 2000 to December 2005 (72 months). The main summary statistics for different indexes are reported below. The results are compared with S&P500 Index returns.

Summary statistics for replication results

	Equity Hedge Index	Dynamic Selecting Replica Index (IDS)	Naive Selecting Replica Index (INS)	S&P500
Annualized Return	9,50%	6,70%	5,14%	0,02%
Annualized SD	7,69%	12,36%	13,08%	15,23%

The equally-weighted Equity Hedge Index outperforms both the clone indexes and the S&P500. The “Dynamic Factor Selecting Clone Index” outperforms the “Naïve Factor Selecting Index”. The annualized average returns for our clone is 6.70% (with an annualized standard deviation of 12.14%) against 5.14% for the “naïve” clone (with an annualized standard deviation of 13.70%).

The results reported in this paper show that the funds of the selected strategy are sensible to 2-3 latent factors (rather than one) that capture the covariation of the fund returns. If the first latent factor seems to match with the S&P500 index returns, our analysis shows that it is not sufficient to explain all the common variation of the hedge fund returns even for the equity-oriented hedge fund strategy considered here.

The four-step dynamic approach developed in this paper has permitted to implement recent asymptotic theory developments (Bai and Ng, 2002, 2003, 2006) in a multifactor modelling framework for equity returns. The results seem to be encouraging. A better adaptation of the asymptotic tests implemented here is left to a future version of this work.

⁴¹ Not to be confused with the full history beginning in January 1997. The 36 first months of our history data are used to perform estimations for the first rolling period.

5. Concluding Remarks

We have shown in this article that it is possible to link a new market practice – hedge fund replication, to some useful and well known financial theory – factor modelling of equity returns. Doing so, we get a more precise comprehension on the underlying factor structure that drives the covariation of Equity Hedge Fund returns. Our approach is based on individual hedge fund returns, instead of just index returns. This choice allows us to go further in the comprehension of the latent factor structure we use in the replication process. In particular, the information on the common behaviour of fund returns depends not only on the past historical data we use (time series dimension), but also on the number of funds we observe (cross-sectional dimension). Then, we can use recent asymptotic theories⁴² for factor selection. The asymptotic tests behave well for large N and moderately large T . This approach is clearly more in line with the dynamic factor selection objective we fix.

In practice, for the Equity Hedge Funds belonging to the HFR Database, a simple 2 factors structure is sufficient to capture the common behaviour in hedge fund returns. If the first factor is closed to the equity market index, the second factor is often more difficult to understand and illustrate the style rotation employed by hedge fund managers. Finally, the economic interpretation of the risk factors allows building replicating portfolio as they are proposed by practitioners.

⁴² Bai and Ng (2002, 2003, 2006).

Bibliography

AGARWAL Vikas, Narayan Y. NAIK, 2001, “Characterizing Hedge Fund Risks with Buy-and-Hold and Option-Based Strategies”, Working Paper, June 6, 2001.

AMIN Gaurav S., KAT Harry M., 2003, “Hedge Fund Performance 1990-2000: Do the Money Machines Really Add Value?”, *Journal of Financial and Quantitative Analysis*, 38, No.2, June 2003.

BAI Jushan, Serena NG, 2006, “Evaluating Latent and Observed Factors in Macroeconomics and Finance”, *Journal of Econometrics*, 131, 507-537.

BAI Jushan, Serena NG, 2006, “Confidence Intervals for Diffusion Index Forecasts with a Large Number of Predictors”, Working Paper, August 2004.

BAI Jushan, Serena NG, 2003, “Inferential Theory for Factor Models of Large Dimensions”, *Econometrica*, 71, No.1 (January, 2003), 135-171.

BAI Jushan, Serena NG, 2002, “Determining the Number of Factors in Approximate Factor Models”, *Econometrica*, 70, No.1 (January, 2002), 191-221.

BLACK F., M. C. JENSEN and M. SCHOLES (1972), “The Capital Asset Pricing Model: some new Empirical Tests”, *Studies in the Theory of Capital Markets*, Michael C. Jensen (Preager, New York).

CARHART M. (1997), “On Persistence in Mutual Fund Performance”, *Journal of Finance*, pp. 57-92.

CHAMBERLAIN, G., ROTHSCILD, M., 1983, “Arbitrage, Factor Structure and Mean-Variance Analysis in Large Asset Markets”, *Econometrica* 51, 1305-1324.

CHAN Nicholas, GETMANSKY Mila, HAAS Shane M. and LO Andrew W., 2005, “Systemic Risk and hedge Funds”, Working Paper, August 1 – 2005.

CHEN, N., ROLL, R., ROSS, S., 1986, “Economic Forces and the Stock Market”, *Journal of Business* 59, 383-403.

CONNOR Gregory, Robert A KORAJCZYK, 1986, “Performance Measurement with the APT: A New Framework for Analysis”, *Journal of Financial Economics*, 15, 373-394.

CONNOR Gregory, Robert A KORAJCZYK, 1988, “Risk and Return in an Equilibrium APT: Application to a New Test Methodology”, *Journal of Financial Economics* 21, 255-289.

CONNOR Gregory, Robert A KORAJCZYK, 1993, “A Test for the Number of Factors in an Approximate Factor Model”, *Journal of Finance*, 48, 1263-1291.

CORIELLI Francesco, Massimiliano MARCELLINO, 2006, “Factor Based Index Tracking”, *Journal of Banking and Finance*, 30, 2215-2233.

DONALDS, S., 1997, “Inference Concerning the Number of Factors in a Multivariate Nonparametric Relationship”, *Econometrica*, 65, 103-132.

FAMA, Eugene F. and James D. MACBETH, 1973. « Risk, Return, and Equilibrium : Empirical Tests.” *Journal of Political Economy*. 81:3, pp. 607-636.

- FAMA, Eugene F. and Kenneth R. FRENCH. 1992. "The Cross-Section of Expected Stock Returns." *Journal of Finance*. 47:2, pp. 427-465.
- FAMA, Eugene F. and Kenneth R. FRENCH. 1993. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics*. 33:1, pp. 3-56.
- FUNG, William, David A. HSIEH, 2004, "Extracting Portable Alphas from Equity Long-Short Hedge Funds", *Journal of Investment Management*, Vol.2, N°4, pp. 1-19.
- FUNG, William, David A. HSIEH, 2002a, "Asset-Based Style Factors for Hedge Funds", *Financial Analysis Journal* 58, pp. 16-27.
- FUNG, William, David A. HSIEH, 2001, "The Risk in Hedge Fund Strategies: Theory and Evidence from Trend Followers", *The Review of Financial Studies*, Vol. 14, No 2, pp. 313 – 341.
- FUNG, William, David A. HSIEH, 1997a, "Empirical Characteristics of Dynamic Trading Strategies: The Case of Hedge Funds", *The Review of Financial Studies*, Vol.10, N°2, pp. 275-302 .
- GETAMANSKY, M., LO, A. and I. MARAKOV, 2004, "An Econometric Analysis of Serial Correlation and Illiquidity in Hedge-Fund Returns", *Journal of Financial Economics* 74, 529-609.
- HASANHODZIC Jasmina, Andrew W. LO, 2006, "Can Hedge-Funds Be Replicated?: The Linear Case", Working Paper, August, 2006.
- JAGANNATHAN R. and Z. WANG. 1996. "The Conditional CAPM and the Cross-Section of Expected Returns." *Journal of Finance*. 51, pp. 3-54.
- KAT Harry M., Helder P. PALARO, 2005, "Who needs Hedge Funds? A Copula-Based Approach to Hedge Fund Return Replication", Working Paper, Alternative Investment Research Centre, Cass Business School, City University London, (downloadable from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=855424).
- LEHMANN, B., MODEST, D., 1988, "The Empirical Foundations of the Arbitrage Pricing Theory", *Journal of Financial Economics* 21, 213-254.
- Liang, B., 2000, "Hedge Funds: The Living and the Dead", *Journal of Financial and Quantitative Analysis* 35, 309-326.
- MITCHELL Mark, Todd PULVINO, 2001, "Characteristics of Risk and Return in Risk Arbitrage", *The Journal of Finance*, 56, N° 6, December 2001, 2135-275.
- ROLL R., S. A. ROSS, 1980, "An Empirical Investigation of the Arbitrage Pricing Theory", *Journal of Finance*, vol. 35, pp. 1073-1103.
- RONCALLI Thierry and TEILETCHE Jérôme, 2007, "An Alternative Approach to Alternative Beta", Working Paper, SGAM Alternative Investments, April 2007.
- ROSS, Stephen A. 1976. "The Arbitrage Theory of Capital Asset Pricing", *Journal of Economic Theory*, 13:3, 341-360.
- SHANKEN J., 1992, "On the Estimation of Beta Pricing Models", *Review of Financial Studies*, 5, 1-33.

Appendix

TABLE I

$$\text{DGP: } X_{it} = \sum_{j=1}^r \lambda_{ij} F_{ij} + \sqrt{\theta} e_{it} = c_{it} + \sqrt{\theta} e_{it}; e_{it} = \rho e_{it-1} + v_{it} + \sum_{j \neq 0, j=-J}^J \beta v_{i-jt}; r = 2; \theta = 2;$$

$$\rho = 0.50; \beta = 0.20; J = \max\{N/20, 10\}.$$

N	T		PC1	PC2	PC3		IC1	IC2	IC3
300	24		5,80	5,61	6,25		2,08	2,02	2,30
260	24		6,67	6,46	7,10		2,59	2,31	3,56
200	24		6,30	6,01	6,89		2,24	2,05	2,92
150	24		6,54	6,23	7,25		3,00	2,48	4,68
100	24		6,66	6,24	7,60		4,26	3,42	6,20
300	36		3,32	3,08	3,93		2,00	2,00	2,02
260	36		3,04	2,82	3,74		2,00	2,00	2,01
200	36		5,78	5,45	6,70		2,60	2,29	4,19
150	36		5,14	4,72	6,22		2,35	2,15	3,76
100	36		5,46	5,02	6,68		3,33	2,66	5,31
100	60		5,57	5,03	7,12		4,35	3,29	6,63
200	60		4,67	4,16	6,31		2,27	2,07	4,52
100	100		5,88	5,27	7,97		5,25	4,12	7,92
200	100		5,55	4,77	7,93		3,44	2,39	7,84
30	40		1,00	1,00	1,00		1,00	1,00	1,00

The results are averaged over 1000 replications.

TABLE II

$$\text{DGP: } X_{it} = \sum_{j=1}^r \lambda_{ij} F_{ij} + \sqrt{\theta} e_{it} = c_{it} + \sqrt{\theta} e_{it}; e_{it} = \rho e_{it-1} + v_{it} + \sum_{j \neq 0, j=-J}^J \beta v_{i-jt}; r = 3; \theta = 3;$$

$$\rho = 0.50; \beta = 0.20; J = \max\{N/20, 10\}.$$

N	T		PC1	PC2	PC3		IC1	IC2	IC3
300	24		4,73	4,54	5,16		2,79	2,75	2,89
260	24		4,62	4,41	5,13		2,82	2,75	2,88
200	24		5,71	5,45	6,34		2,63	2,48	2,94
150	24		6,19	5,84	6,94		2,86	2,52	3,81
100	24		6,48	6,06	7,53		3,56	2,80	5,55
300	36		5,80	5,56	6,43		3,26	3,10	3,97
260	36		5,47	5,22	6,19		3,13	3,00	3,74
200	36		5,00	4,67	5,92		2,98	2,89	3,45
150	36		4,16	3,78	5,22		2,99	2,96	3,16
100	36		6,06	5,63	7,19		4,10	3,32	6,34
100	60		6,26	5,74	7,76		5,26	4,12	7,59
200	60		5,18	4,68	6,86		3,26	3,06	5,48
100	100		6,66	6,01	7,99		6,17	4,97	7,99
200	100		6,13	5,32	7,99		4,40	3,39	7,98
30	40		1,00	1,00	1,00		1,00	1,00	1,00

The results are averaged over 1000 replications.

TABLE III

Finite sample properties of test statistics: Monte Carlo simulations

<i>N</i>	<i>T</i>		<i>A(j)</i>	<i>NS</i>	<i>R</i> ²	<i>N</i>	<i>T</i>		<i>A(j)</i>	<i>NS</i>	<i>R</i> ²
300	36	G1	0,003	0,009	0,991	300	24	G1	0,005	0,009	0,991
300	36	G2	0,003	0,009	0,991	300	24	G2	0,006	0,009	0,991
300	36	G3	0,078	0,085	0,923	300	24	G3	0,128	0,085	0,923
300	36	G4	0,040	0,046	0,956	300	24	G4	0,066	0,046	0,956
300	36	G5	0,781	20,338	0,163	300	24	G5	0,807	33,421	0,188
300	36	G6	0,709	5,697	0,246	300	24	G6	0,756	8,640	0,262
300	36	G7	0,885	567,107	0,057	300	24	G7	0,885	46,731	0,087
250	36	G1	0,003	0,011	0,990	250	24	G1	0,006	0,012	0,989
250	36	G2	0,003	0,011	0,989	250	24	G2	0,006	0,011	0,989
250	36	G3	0,068	0,087	0,920	250	24	G3	0,107	0,085	0,922
250	36	G4	0,033	0,049	0,954	250	24	G4	0,057	0,049	0,954
250	36	G5	0,763	25,327	0,159	250	24	G5	0,785	18,144	0,187
250	36	G6	0,685	6,616	0,243	250	24	G6	0,732	29,687	0,267
250	36	G7	0,871	128,404	0,059	250	24	G7	0,874	60,498	0,086
200	36	G1	0,003	0,013	0,987	200	24	G1	0,005	0,013	0,987
200	36	G2	0,003	0,013	0,987	200	24	G2	0,004	0,014	0,986
200	36	G3	0,054	0,089	0,919	200	24	G3	0,090	0,089	0,919
200	36	G4	0,028	0,052	0,951	200	24	G4	0,045	0,052	0,951
200	36	G5	0,731	18,900	0,163	200	24	G5	0,764	28,367	0,191
200	36	G6	0,655	5,554	0,240	200	24	G6	0,700	6,084	0,269
200	36	G7	0,859	127,262	0,055	200	24	G7	0,855	80,598	0,086
150	36	G1	0,003	0,018	0,982	150	24	G1	0,005	0,019	0,981
150	36	G2	0,003	0,017	0,983	150	24	G2	0,005	0,020	0,981
150	36	G3	0,041	0,097	0,913	150	24	G3	0,066	0,093	0,916
150	36	G4	0,022	0,057	0,946	150	24	G4	0,035	0,057	0,947
150	36	G5	0,699	29,742	0,158	150	24	G5	0,730	23,360	0,192
150	36	G6	0,612	5,856	0,245	150	24	G6	0,659	5,701	0,266
150	36	G7	0,831	97,851	0,059	150	24	G7	0,836	75,996	0,088
100	36	G1	0,003	0,027	0,974	100	24	G1	0,005	0,028	0,973
100	36	G2	0,003	0,027	0,974	100	24	G2	0,005	0,029	0,972
100	36	G3	0,028	0,105	0,905	100	24	G3	0,047	0,106	0,905
100	36	G4	0,015	0,066	0,938	100	24	G4	0,025	0,067	0,938
100	36	G5	0,646	18,603	0,154	100	24	G5	0,686	19,214	0,184
100	36	G6	0,552	21,352	0,234	100	24	G6	0,607	7,830	0,260
100	36	G7	0,803	102,409	0,057	100	24	G7	0,802	83,840	0,085

$A(j)$ is the frequency that $|\hat{\tau}_t(j)|$ exceeds the critical value of 1.96 in the sample of size T . $NS(j)$ is the noise-to-signal ratio, see (6). R^2 is defined in (7). All the values reported here are averaged over 1000 replications.

TABLE IV

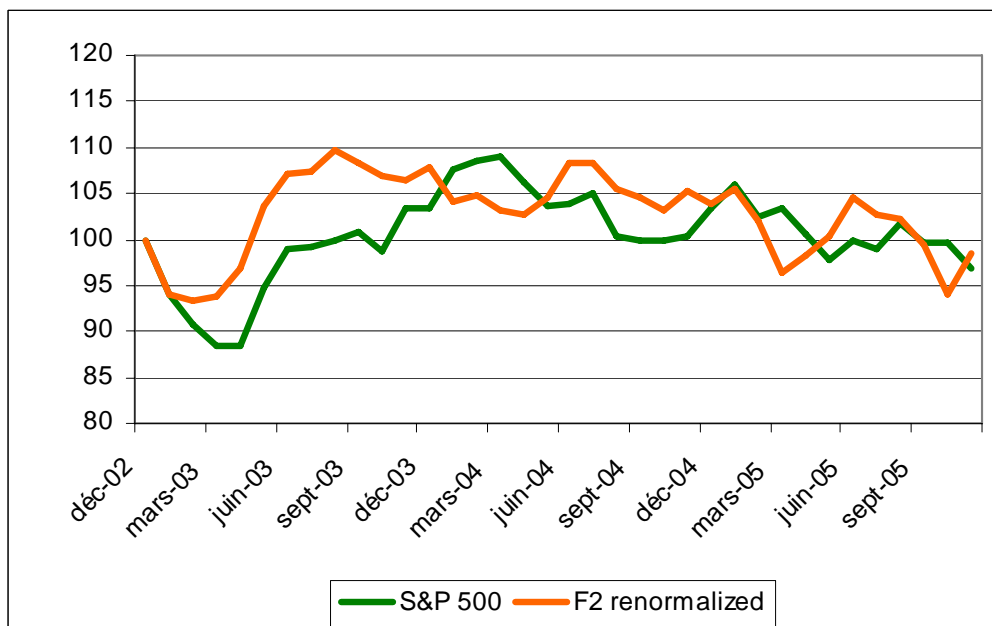
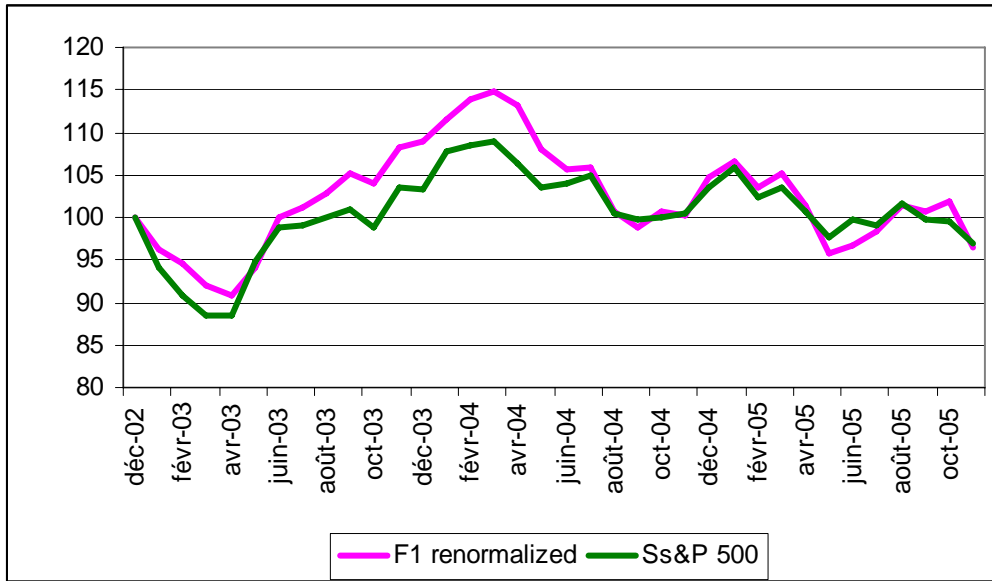
The selecting matrix, S

	MKT	SMB	HML	MOM	CREDIT	BOND	Russell 2000	USD	S&P500	GSCI
1	0	1	1	1	0	0	0	0	1	1
2	0	1	1	1	0	0	0	0	1	1
3	0	1	1	1	0	0	0	0	1	1
4	0	1	1	1	0	0	0	0	1	1
5	0	1	1	1	0	0	1	0	1	1
6	0	1	1	1	0	0	1	0	1	0
7	0	1	1	1	0	0	1	0	1	0
8	0	1	1	1	0	0	1	0	1	0
9	0	1	1	1	0	0	1	0	1	0
10	0	1	1	1	0	0	1	0	1	0
11	0	1	1	1	0	0	1	0	1	0
12	0	1	1	1	0	0	1	0	1	0
13	0	1	1	1	0	0	1	0	1	0
14	0	1	1	1	0	0	1	0	1	0
15	0	1	1	1	0	0	0	0	1	0
16	0	1	1	1	0	0	0	0	1	0
17	0	1	1	1	0	0	0	0	1	0
18	0	1	1	1	0	0	0	0	1	0
19	0	1	1	1	0	0	0	0	1	0
20	0	1	1	1	0	0	0	0	1	0
21	0	1	1	1	0	0	0	0	1	0
22	0	1	1	1	1	0	0	0	1	0
23	0	1	1	1	1	0	0	0	1	0
24	0	1	1	1	1	0	0	0	1	0
25	0	1	1	1	1	0	0	0	1	1
26	0	1	1	1	1	0	0	0	1	1
27	0	1	1	1	1	0	0	0	1	0
28	0	1	1	1	1	0	0	0	1	0
29	0	1	1	1	1	0	0	0	1	0
30	0	1	1	1	1	0	0	0	1	0
31	0	1	1	1	1	0	0	0	1	0
32	0	1	1	1	1	0	0	0	1	0
33	0	1	1	1	1	0	0	0	1	0
34	0	1	1	1	1	0	1	0	1	1
35	0	1	1	1	1	0	1	0	1	1
36	0	1	1	1	1	0	1	0	1	0
37	0	1	1	1	1	0	0	0	1	0
38	0	1	1	1	1	0	0	0	1	0
39	0	1	1	1	1	0	0	0	1	0
40	0	1	1	1	1	1	1	1	1	0
41	0	1	1	1	1	0	1	0	1	1
42	0	1	1	1	1	0	1	0	1	1
43	0	1	1	1	1	0	1	0	1	1
44	0	1	1	1	1	0	1	0	1	1
45	0	1	1	1	1	0	1	0	1	0
46	0	1	1	1	1	0	1	0	1	0
47	0	1	1	1	1	1	1	0	1	0
48	0	1	1	1	1	0	1	0	1	0
49	0	1	1	1	1	0	1	0	1	0

50	0	1	1	1	1	0	1	0	1	0
51	0	1	1	1	1	0	1	0	1	0
52	0	1	1	1	1	0	1	0	1	0
53	0	1	1	1	1	0	1	1	1	0
54	0	1	1	1	1	0	1	0	1	0
55	0	1	1	1	1	0	1	1	1	0
56	0	1	1	1	1	0	1	1	1	0
57	0	1	1	1	1	0	1	0	1	0
58	0	1	1	1	1	0	1	1	1	1
59	0	1	1	1	1	1	1	0	1	1
60	0	1	1	1	1	1	1	0	1	0
61	0	1	1	1	1	0	1	0	1	0
62	0	1	1	1	1	1	1	0	1	0
63	0	1	1	1	1	0	1	0	1	0
64	0	1	1	1	1	1	1	1	1	1
65	0	1	1	1	1	1	1	1	1	1
66	0	1	1	1	1	1	1	1	1	1
67	0	1	1	1	1	0	1	1	1	1
68	0	1	1	1	1	0	1	0	1	1
69	0	1	1	1	1	0	1	0	1	1
70	0	1	1	1	1	0	1	0	1	1
71	0	1	1	1	1	0	1	1	1	1
72	0	1	1	1	0	1	1	1	1	1

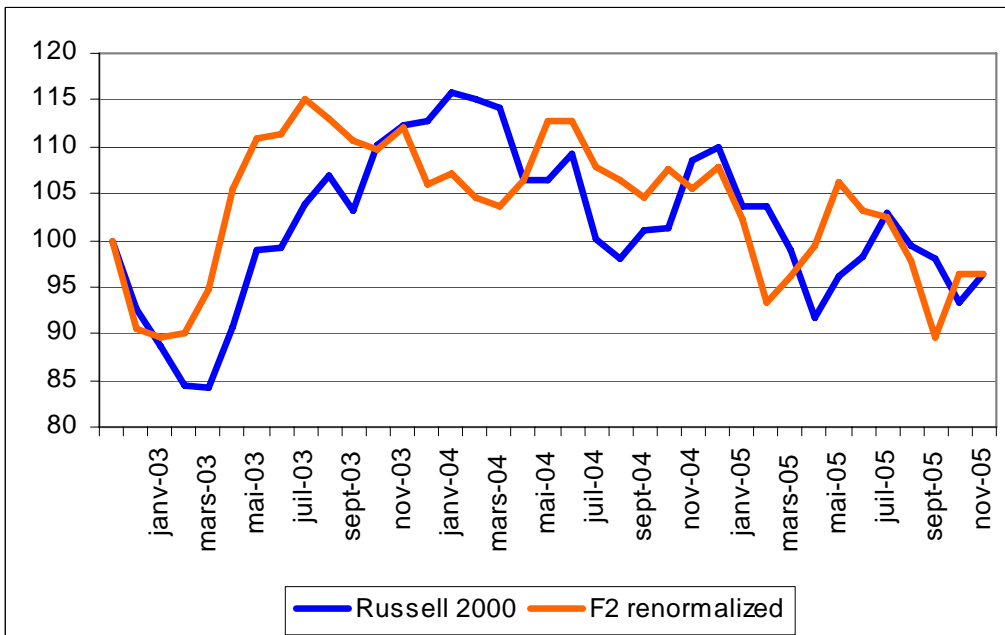
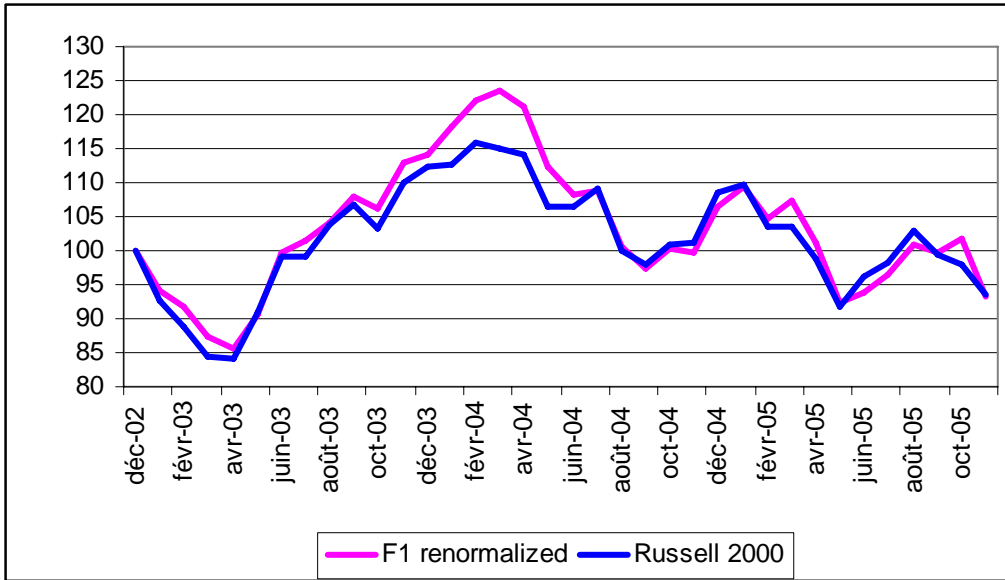
Each row corresponds to one of the 72 rolling windows of our sample data. For a given rolling window, are included in the analysis, only the factors having $S_{jt} = 1$ ($j = 1, \dots, m$ and $t = 1, \dots, 72$).

FIGURE I



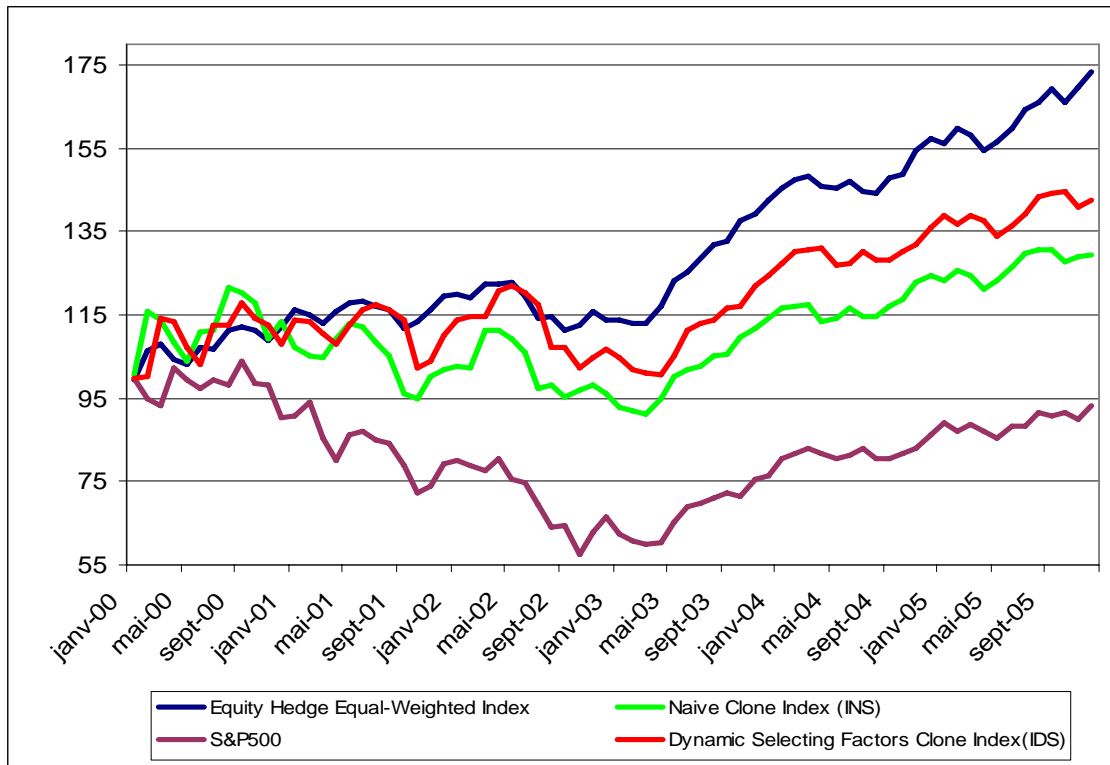
F1 and F2 are the estimated latent factors corresponding to the two largest eigenvalues of the covariance matrix of the fund's returns of our sample having full data for the last rolling period which extends from December 2002 to December 2005. Recall that by definition, F1 and F2 are standardized variables having mean zero and unit variance. F1 and F2 have been renormalized in order to obtain the same standard deviation as for S&P 500 Index. S&P500 Index returns have been centred by their mean in order to facilitate the comparison with F1 and F2.

FIGURE II



F1 and F2 are the estimated latent factors corresponding to the two largest eigenvalues of the covariance matrix of the fund's returns of our sample having full data for the last rolling period which extends from December 2002 to December 2005. Recall that by definition, F1 and F2 are standardized variables having mean zero and unit variance. F1 and F2 have been renormalized in order to obtain the same standard deviation as for Russell 2000 Index. Russell 2000 Index returns have been centred by their mean in order to facilitate the comparison with F1 and F2.

FIGURE III



We denote IDS and INS respectively the monthly returns of a “dynamic factor selecting clone index” and a “naïve factor selecting clone index”. The former is the clone constructed by our methodology and the latter is the one constructed in a similar way as Lo and Hasanbobic (2006).