Statistical Methods & Applications (2004) 13: 103–127
DOI: 10.1007/s10260-003-0081-z

Pitfalls in linear models for style analysis*

Francesco Corielli[1] **and Attilio Meucci**[2]

[1] Istituto di Metodi Quantitativi, Università Bocconi, Milano, Italy
(e-mail: francesco.corielli@uni-bocconi.it)
[2] Relative Value International, Greenwich, CT, USA (e-mail: ameucci@rvifunds.com)

Abstract. We discuss the statistical properties of return-based OLS style analysis introduced by Sharpe (1992). The aim of style analysis is to infer a fund manager's investment decisions using only publicly available data on the fund performance and on the time evolution of market indexes. We show that the model proposed by Sharpe suffers of relevant drawbacks, most notably that it fails to yield correct results even in the simple case of a buy-and-hold strategy that only invests in the market indexes. Under this hypothesis we show that a model linear in index levels, as opposed to index returns, estimated via a Kalman filter avoids Sharpe's model drawbacks. We further extend our analysis to strategies where the fund manager policy changes with time and the asset classes in which the fund manager invests are not known exactly. In this last case we show that a style analysis is possible only conditional to either an orthogonality hypothesis on the "active" investment strategy, or by the introduction of suitable instrumental variables.

Key words: Linear models, financial statistics, instrumental variables

1. Sharpe model for style analysis

The analysis of the "style" of a fund manager is one of the most interesting field of application for statistical methods in finance. The basic problem of style analysis is quite simple. In Europe and the US thousands of investment funds are available for subscription either by private or institutional investors. It is of the utmost interest, from the point of view of the investor but also from the point of view of state sponsored regulatory agencies (e.g. SEC in the US and Consob in Italy) and of privately owned rating companies (e.g. Morningstar, Moody's, S&P etc.), to be

* The authors are grateful to the editor and an anonymous referee for many comments which greatly helped in improving the paper. The authors are, obviously, fully responsible for any remaining error.

able to assess the historical evolution of the fund investment policy and to track its day-by-day changes.

For any subscriber in the fund and for control authorities, there is a relevant legal interest in knowing whether the fund actually follows the investment policy he purports to hold. For rating agencies and for institutional fund of funds managers, the main interest is that of evaluating the quality of the fund manager investment decisions: the ability of the fund manager to invest in the "right" asset classes. This is the basis for the evaluation of a fund performance and, consequently, for the decision of including it in a fund of funds product, or for awarding the fund a low versus high rating.

Availability of effective asset exposures for investment funds would solve the problem of style analysis. Asset exposures, however, are known with sufficient precision and frequency only for some classes of funds in regulated markets (e.g., Italian investment funds available for public subscription) and often even this data is available only with significant delay. On the contrary funds quotas, i.e., the value of a fund unit of investment, are usually known with detail and negligible delay.

The above described investment requirements and lack of detailed information opens the way for statistical based style analysis.

Style analysis comes in two flavors: "strong" and "weak". The kind of problems described in this introduction require strong style analysis: the need is to make inference on the actual strategy of the fund manager. In this paper we concentrate on this kind of style analysis. On the other hand, weak style analysis only requires the estimation of a trading strategy, maybe based on different securities, which can mimic, at least for some time, the behaviour of the fund value. This problem is strongly connected with the problem of tracking market indexes and is discussed by a related strain of the literature (see e.g. Beasley and others (2001) and Corielli and Marcellino (2002)).

The idea of inferring a fund composition regressing its market returns on the returns of suitably chosen asset classes is a "folk" idea in quantitative finance (see, e.g. the detailed discussion in Brown and Goetzmann, 1997). However the merit for a simple and direct first description of return based style analysis can be ascribed to the work of Sharpe (1992). In his paper, Sharpe tries to infer a mutual fund composition by observing only the level of the fund performance and the performance of some market indexes. Sharpe's model generated a huge mass of empirical applications. It is reasonable to argue that return based style analysis is now one of the most frequently used, if not the most frequently used, statistical tool in the field of finance.

In recent years style analysis is enjoying a renewed academic interest jointly with the study of the hedge fund industry (see e.g. Fung and Hsieh, 1997, 1998; Brown and Goetzmann, 2001).

In this paper we introduce Sharpe's assumptions and discuss the merits and statistical coherence of Sharpe's model.

Any statistical style analysis begins with the definition of a set of style indexes that is, with the definition of a set of market indexes to which the style analyzer tries to reduce the investment style of the fund manager. Obviously the fund manager cannot be required, a priori, to invest only in the indexes chosen by the style analyzer.

It is then necessary to give an accounting definition of the fund manager strategy in order to pinpoint the strategy elements that the style analyzer wants to estimate.

For this reason the paper begins introducing an accounting definition of a fund manager strategy which separates an “active” from a “passive” investment strategy. In our context a “passive” strategy is simply the investment in the set of indexes used for style analysis, the “active” strategy is the complement to the passive investment. To make possible the identification of the two strategies we suppose that each strategy is self financing. We choose this definition because it closely mimics, and formalizes, commonly shared market intuition. Other accounting definitions are possible and, on the basis of these definitions, other methods for style analysis can be suggested. We stress the fact that, in any case, the choice of some accounting definition of the term “style” is necessary a-priori in any style analysis.

Given the definition of active and passive strategy, we begin our analysis with the very simple setting where the fund manager’s investments are actually concentrated in the set of indexes used for style analysis. Moreover the fund manager’s strategy is completely static, that is, the fund manager buys some amount of each index at fund inception and never acts again on the fund up to fund liquidation. Even in this very simple setting, which idealizes the tracking of a set of indexes (a benchmark strategy), we show that Sharpe’s model results in biased estimates of the style parameters. The basic reason for this is the specification of the model in terms of returns. We introduce a level based model which gives exact estimates.

We consider then the general setting, where the strategy of the fund manager goes beyond the set of indexes used for style analysis and is a mix of passive and active strategy. Moreover we allow for investments to be time varying. In order to tackle the problem of style analysis in this case, we introduce a level based model in statespace form and we define a set of hypotheses under which it can be estimated using the standard Kalman filter.

Among these hypotheses we pinpoint an assumption which is unlikely to hold in general: that is the assumption of non correlation between the values of the passive and the active portions of the fund. We argue that this hypothesis fairly well characterizes active strategies in the class of “market neutral” strategies but that we cannot suppose all active strategies to be in this class. From this observation we show that, implicit in any style analysis, is the need for some characterization of the active strategy of the fund manager in terms of a set of observable variables hortogonal to it. These variables, which can be naturally chosen for many common active strategies, take the role of instruments in the estimation of the style model and allow for unbiased estimates of the style parameters. We stress again the fact that the choice of the set of instruments is an essential part of style analysis, it cannot be avoided. In fact, the use of no instruments is equivalent to the implicit assumption of a market neutral active strategy.

The rest of the paper is organized as follows: in Sect. 2 we define the assumptions underlying our analysis. Section 3 discusses the simple buy and hold case.

In Sects. 4 and 5 we consider the more complex general case of a time varying investment in both the passive and active portfolio.

In Sect. 6 we discuss some possible ways of choosing instrumental variables.

In Sect. 7 we run some empirical analyses based on Sharpe's original data which pinpoint the properties of the models discussed in the paper.

In Sect. 8 we conclude.

2. Basic modelling assumptions

We start by defining the basic quantities in our analysis and making some assumptions which shall be shared by all models we shall consider in this paper.

1. There exists a set of m securities whose prices at time t we indicate with $S_j(t)$. There exists a set of l exhaustive, mutually exclusive easy to track indexes. Indexes are simply portfolios of securities, that is the value of index i at time t is given by:

$$I_i(t) = \sum_{j=1}^{m} n_{ij}(t) S_j(t)$$

   Exhaustive means that each security available for investment is included in at least one index. Exclusive means that no two indexes contain the same security. These two conditions imply that $n_{ij}(t) \neq 0$ only and exactly once for each pair ij. For the sake of simplicity we suppose $n_{ij}(t)$ to be constant in time: $n_{ij}(t) = n_{ij}$.
2. The prices $I_1(t), \ldots, I_l(t)$ at time t of the indexes are publicly available. We suppose that the data series for the indexes $I_1(t), \ldots, I_l(t)$ do not suffer of collinearity problems. This does not seem a very strong hypothesis and embodies Sharpe's idea of "exclusiveness" of each index. From the point of view of the style analyzer it can always be satisfied by an accurate selection of the indexes. On the other hand, if the fund manager chooses to invest a part of the fund in collinear indexes, any attribution of the style to someone of the collinear indexes or to a linear combination of the same could be easily considered as equivalent.
3. Each fund portfolio $\Pi(t)$ can be seen as the sum of a passive and an active portfolio:

$$\Pi(t) = P(t) + A(t) \tag{1}$$

   The value of the portfolio is observable:
4. The passive portfolio is a combination of the indexes, so its value is:

$$P(t) = \sum_{i=1}^{l} n_i(t) I_i(t) \tag{2}$$

   where $n_i(t)$ is the unknown number of shares of the i-th index at time t. .
5. The active portfolio summarizes any difference between the actual fund manager position, and a pure exposition to the indexes I_i. So the "active" position is simply the remainder of the fund position after concentrating out the index exposition. This definition is consistent with common risk management practice where the fund performance is split between the "benchmark" (in our case the indexes)

and the "active investment" (i.e., non benchmark). For normalization purposes usually the active investment is initialized to 0 at some time $t_0 = 0$ but this is not relevant in our context. Net value at time t of the active part of the portfolio is:

$$A(t) = \sum_{j=1}^{m} m_j(t) S_j(t) \tag{3}$$

where m_j is the (possibly negative) unknown number of shares of the j-th security, whose price we denote by S_j.

6. We assume that no cashflow occurs between the passive and active portion of the portfolio and in or out of the portfolio. In other words we assume that both the passive and the active portfolio are self-financing. The fund manager's strategy simply consists in asset reallocations inside both portfolios. If t_k are the event times when price changes (we use t_k for specific dates while t denotes a generic date) this is equivalent to imposing that:

$$P(t_k) = \sum_{i=1}^{l} n_i(t_{k-1}) I_i(t_k) = \sum_{i=1}^{l} n_i(t_k) I_i(t_k) \tag{4}$$
$$A(t_k) = \sum_{j=1}^{m} m_j(t_{k-1}) S_j(t_k) = \sum_{j=1}^{m} m_j(t_k) S_j(t_k)$$

We stress again that (1) is a stylized representation of a fund manager's approach to investing: it is unlikely that a manager split this process into investment in indexes and investment in single securities, instead of simply deciding which securities to hold in his portfolio. Therefore (1), as well as Assumptions 1–6, should be seen as an arbitrary accounting reinterpretation of the portfolio manager's holdings.

Sharpe's purpose is to fairly evaluate the performance of the manager. Since only the value of the fund and of the indexes I_i are observed, this evaluation can only be based on an estimate of the composition of the passive portfolio at each time t, i.e. of the unknowns $n_i(t)$. Equivalently, defining the relative weights as follows:

$$\omega_i(t) = \frac{n_i(t) I_i(t)}{P(t)} \tag{5}$$

we have to determine the unknowns $\omega_i(t)$. Notice that Sharpe assumes the $\omega_i - s$ to satisfy the constraints:

$$0 \leq \omega_i(t) \leq 1, \quad \sum_{i=1}^{l} \omega_i(t) = 1 \tag{6}$$

This is meaningful only if the value of the active part of the strategy is identically 0. Some consequences of this assumption shall be discussed in what follows.

Let us now consider the time evolution of returns. Denoting the return (+1) of a security (or a portfolio) with price p between time t_{k-1} and time t_k as: $R_p(t_k) = \frac{p(t_k)}{p(t_{k-1})}$, for a passive portfolio (that is a portfolio with $A(t) = 0$) it follows that

$$R_{\Pi}(t_k) = R_P(t_k) = \sum_{i=1}^{l} \omega_i(t_{k-1}) R_{I_i}(t_k).$$

In the general case, when $A(t) \neq 0$, from the above assumptions we obtain:

$$R_{\Pi}(t_k) = u(t_k) \sum_{i=1}^{l} \omega_i(t_{k-1}) R_{I_i}(t_k) + \epsilon(t_k), \tag{7}$$

where

$$u(t_k) = \frac{\sum_{i=1}^{l} n_i(t_{k-1}) I_i(t_{k-1})}{\sum_{i=1}^{l} n_i(t_{k-1}) I_i(t_{k-1}) + \sum_{j=1}^{m} m_j(t_{k-1}) S_j(t_{k-1})}$$

$$\epsilon(t_k) = \frac{\sum_{j=1}^{m} m_j(t_k) S_j(t_k)}{\sum_{i=1}^{l} n_i(t_{k-1}) I_i(t_{k-1}) + \sum_{j=1}^{m} m_j(t_{k-1}) S_j(t_{k-1})}$$

The Eq. (7) is an identity derived from certain accounting definitions, in order to apply to this identity any statistical procedure we must make some hypothesis on the probabilistic nature of its terms. To fix ideas we start by recalling Sharpe hypotheses.

In order to infer the value of ω, Sharpe (1992) collects equally spaced (monthly, in the example described in his paper) data for $I_1, \ldots, I_l$ and Π, runs the regression

$$R_{\Pi}(t_k) = \sum_{i=1}^{l} \omega_i R_{I_i}(t_k) + \epsilon(t_k), \tag{8}$$

and estimates $\hat{\omega}$ using constrained OLS to satisfy (6).

It is clear that, for the estimates to have good statistical properties, at least three hypotheses must hold and are implicit in Sharpe's approach:

7. The vector ω is constant in time.
8. The errors $\epsilon(t_k)$ are not correlated with $R_{I_i}(t_k)$, their expected value is 0 or, at least, is constant and their second moments exist.
9. The multiplicative term $u(t)$ in (7) is approximately equal to 1.

In the next section we shall see how these assumptions are unlikely to hold exactly even under the most simple of the trading strategies: the buy and hold strategy, where the asset allocation is decided at fund inception and no further trades are operated during the fund life.

Indeed, Assumptions 8 and 9 basically require the portfolio to be invested only in the indexes; Moreover Assumption 7 can hold only under the very unlikely hypothesis that the fund manager changes the asset allocation continuously and in a "contrarian" way, selling the securities with better performances and buying the securities with worse performances.

This notwithstanding, it is interesting to assess whether Sharpe's model can be considered as a good approximation to reality and, if this is not the case, find alternative models. Since most of the problems with Sharpe's model come from the choice of a parametrization based on returns, we shall suggest to reparametrize the model in levels and show how this simple alternative can overcome a large parte of Sharpe's style regression problems.

3. Static investment in indexes

In analyzing Sharpe's assumptions we start from the hypothesis that the portfolio manager only invests in the observed indexes, that is $A(t) = 0$ and that the investment strategy is completely static, that is $n_i(t) = n_i$. This simple strategy, a buy and hold strategy, is the most common strategy for index tracking funds as, for instance, the ETF funds recently introduced in Italy. We can consider this as a benchmark case, meaning that a good style model should work properly at least in this case.

Under the hypothesis of a buy and hold strategy with investments only in the indexes I_i we have that $u(t) = 1$ and $\epsilon(t) = 0$ so that assumptions 8 and 9 are automatically satisfied.

Moreover, an exact relation holds for prices and returns:

$$\Pi(t_k) = \sum_{i=1}^{l} n_i I_i(t_k) \tag{9}$$

$$R_\Pi(t_k) = \sum_{i=1}^{l} \omega_i(t_{k-1}) R_{I_i}(t_k)$$

However the weights

$$\omega_i(t_{k-1}) = \frac{n_i I_i(t_{k-1})}{\sum_{i=1}^{l} n_i I_i(t_{k-1})} \tag{10}$$

change with time so that Assumption 7 is not valid.

We can still interpret Sharpe's regression as an attempt to estimate the following model:

$$R_\Pi(t_k) = \sum_{i=1}^{l} \widetilde{\omega}_i R_{I_i}(t_k) + \epsilon(t_k) \tag{11}$$

$$\epsilon(t_k) = \sum_{i=1}^{l} (\omega_i(t_{k-1}) - \widetilde{\omega}_i) R_{I_i}(t_k)$$

where $\widetilde{\omega}_i$ can be interpreted as an "average" value of $\omega_i(t_{k-1})$.

To apply OLS methods to these equations we need non correlation between $\epsilon(t_k)$ and $R_{I_i}(t_k)$ and constant expectation for $\epsilon(t_k)$.

If we make the very strong, but acceptable, hypothesis of independence between $R_{I_i}(t_k)$ and the $I_i(t_{k-1})$, we have that $R_{I_i}(t_k)$ shall be independent of $\omega_i(t_{k-1})$ hence

$$\begin{aligned} E\left[\epsilon(t_k) R_{I_i}(t_k)\right] &= E\left[\sum_{i=1}^{l}(\omega_i(t_{k-1}) - \widetilde{\omega_i})R^2_{I_i}(t_k)\right] \\ &= E\left[\sum_{i=1}^{l}(\omega_i(t_{k-1}) - \widetilde{\omega_i})\right] E\left[R^2_{I_i}(t_k)\right] \end{aligned}$$

But, from the same hypothesis, we have

$$E\left[\epsilon(t_k)\right] = E\left[\sum_{i=1}^{l}(\omega_i(t_{k-1}) - \widetilde{\omega_i})\right] E\left[R_{I_i}(t_k)\right]$$

So that, even under the strong hypothesis of independence between $R_{I_i}(t_k)$ and the $I_i(t_{k-1})$, no correlation between $\epsilon(t_k)$ and $R_{I_i}(t_k)$ would require the variance of $R_{I_i}(t_k)$ to be equal to 0.

As far as the constant expectation hypothesis is concerned, if $R_{I_i}(t_k)$ is independent of all the $R_{I_i}(t_k)$ then (provided the expected values exist)

$$E(\epsilon(t_k)) = \sum_{i=1}^{l} E(\omega_i(t_{k-1}) - \widetilde{\omega_i})E(R_{I_i}(t_k))$$

and the constant expected value hypothesis for $\epsilon(t_k)$ becomes equivalent to the hypothesis of constant expectation for $R_{I_i}(t_k)$ and for $\omega_i(t_{k-1})$ separately.

A constant expected value for $R_{I_i}(t_k)$ is compatible with the simplest assumptions on asset prices returns made in empirical literature. However this is not true for the hypothesis of constant expected value for $\omega_i(t_{k-1})$. In fact, if $n_i(t) = n_i$ this implies a strong mean reversion property for the price processes and this is not a commonly accepted hypothesis.

In short: both hypotheses necessary for a reasonable use of OLS do not seem to hold in this context, hence interpretation of OLS estimates as "average weights" does not hold.

Another possible way to attack the problem is to explicitly consider the time varying nature of the $\omega_i(t_{k-1})$.

In his paper, Sharpe recognizes that the weights $\omega_i(t)$ can vary over time. Surprisingly, the only reason he mentions this fact is an active intervention on the part of the manager. No consideration is given to the fact that $\omega_i(t)$ may change with constant n_i simply because of non proportional movements of the different indexes. Sharpe's suggestion for dealing with this problem is to run a "rolling-window" regression, i.e., a regression on a set of data that moves along with the observed sample. The idea is that, for short time periods, the weights $\omega_i(t_{k-1})$ should not vary too much and the error term

$$\epsilon(t_k) = \sum_{i=1}^{l}(\omega_i(t_{k-1}) - \widetilde{\omega_i})R_{I_i}(t_k)$$

should be negligible. However if $\omega_i(t_{k-1})$ is approximately constant over a given time span and, since $n_i(t) = n_i$, this means that the return series for all indexes are almost perfectly correlated. In this case, OLS estimation based on return series becomes impossible due to singularity in the regressors matrix. So, a window based regression should work if the corresponding $\epsilon(t_k)$ are small, but the smaller are the $\epsilon(t_k)$ the greater is the correlation across the $R_{I_i}(t_k)$.

To summarize, even in the simplified case of a purely passive buy-and-hold strategy and under the strong hypothesis of independency between $R_{I_i}(t_k)$ and all of the $I_i(t_{k-1})$, Sharpe's constant-coefficient linear model does not seem to yield an OLS estimate. If we interpret it as an approximate model with "average", the invalidity of standard OLS hypotheses implies biased and inconsistent estimates. If, instead, we concentrate on short time window, an approximately constant value of $\omega_i(t_{k-1})$ implies an approximately collinear regressor matrix.

Nonetheless, these problems can be easily overcome by changing the parametrization of the model. Two different but very simple alternative solutions exist and both give exact estimates of the exposures to the indexes under the hypotheses of this section.

The first of these solutions is to specify the model in levels:

$$\Pi(t_k) = \sum_{i=1}^{l} n_i I_i(t_k)$$

Provided the $I_i(t_k)$ are not collinear, the simple use of OLS in this model shall give us exact estimates (up to measurement error of the $I_i(t_k)$ values) of the n_i and the estimation of the relative weights follows as

$$\hat{\omega}_i(t) = \frac{\hat{n}_i I_i(t)}{I(t)}.$$

Alternatively we can preserve the return formulation of the model as:

$$R_\Pi(t_k) = \sum_{i=1}^{l} \omega_i(t_{k-1}) R_{I_i}(t_k)$$

$$\omega_i(t_{k-1}) = \frac{n_i I_i(t_{k-1})}{\sum\limits_{i=1}^{l} n_i I_i(t_{k-1})}$$

but, in this case, we should directly estimate n_i with a nonlinear least squares technique by fitting observed $R_\Pi(t_k)$ to $\sum_{i=1}^{l} \omega_i(t_{k-1}) R_{I_i}(t_k)$ and compute the $\omega_i(t_{k-1})$ from

$$\omega_i(t_{k-1}) = \frac{n_i I_i(t_{k-1})}{\sum\limits_{i=1}^{l} n_i I_i(t_{k-1})}$$

Again, except for measurement errors, this simple technique shall give exact estimates of n_i.

Under the hypotheses of this section, i.e. pure buy and hold investment in the indices, the two suggested parametrizations shall completely solve the problem The purpose of the following sections is to find conditions under which analogous solutions exist under more general hypotheses.

4. The general case: time varying investment in both active and passive portfolios

In this section we assume that the investment $n_i(t)$ in each index may change in time and we allow for the presence of an "active" strategy. On the other hand we suppose that the only information available to the researcher is given by the value of the fund. $\Pi\,(t_k)$ and the values of the indexes $I_i\,(t_k)$. In this case the model in both level and return terms becomes:

$$\Pi\,(t_k) = \sum_{i=1}^{l} n_i(t_k) I_i\,(t_k) + \sum_{j=1}^{m} m_j(t_k) S_j\,(t_k) \tag{12}$$

$$R_{\Pi}\,(t_k) = \sum_{i=1}^{l} \omega_i(t_{k-1}) R_{I_i}(t_k) + \sum_{j=1}^{m} \phi_j(t_{k-1}) R_{Sj}\,(t_k)$$

where the weights now read:

$$\omega_i(t_{k-1}) = \frac{n_i(t_{k-1}) I_i\,(t_{k-1})}{\sum\limits_{i=1}^{l} n_i(t_{k-1}) I_i\,(t_{k-1}) + \sum_{j=1}^{m} m_j(t_{k-1}) S_j\,(t_{k-1})}. \tag{13}$$

$$\phi_j(t_{k-1}) = \frac{m_j(t_{k-1}) S_j\,(t_{k-1})}{\sum\limits_{i=1}^{l} n_i(t_{k-1}) I_i\,(t_{k-1}) + \sum_{j=1}^{m} m_j(t_{k-1}) S_j\,(t_{k-1})} \tag{14}$$

The purpose of the researcher is to estimate the proportion $\omega_i(t_{k-1})$ of the fund value invested in each index. Given the results of the previous subsection we abandon Sharpe suggestion of simply running a linear regression on returns and look for a more proper model.

However before analyzing the new setting, it is useful to briefly discuss a relevant suggestion made in Sharpe (1992). The strategy suggested by Sharpe is to use a constrained version of the OLS estimate where the "weights" ω_i are supposed to be positive and sum to 1. This suggestion, which is consistent with the setting discussed in the previous section, has been unconditionally applied by practitioners to cases where the very strict assumption of an investment strategy fully concentrated on indexes does not hold.

In fact the hypothesis of "weights" ω_i summing to 1 is clearly incompatible with (12) as it would imply $\sum_{j=1}^{m} m_j(t_k) S_j\,(t_k)$ equal to 0 for all t_k. (For a detailed discussion of the problems arising from the constraint on the total values of the weights see DeRoon, Nijman and TerHost, 2000)

Under (12) it is still perfectly sensible to speak of ω_i as weights but, in order to compute these weights, it is first necessary to evaluate the amount of the fund invested in indexes. This simple observation stresses the relevance of expressing the style model in level and not in return terms. While it is still possible to conceive a highly nonlinear model expressed in return terms capable to jointly estimate both the value of the part of the fund invested in indexes and the $\omega_i - s$ this model would be very difficult to estimate and its statistical properties would be very difficult to evaluate. On the other hand for a level specification it is very easy to suggest estimation procedures and to study statistical properties of the results.

The equation

$$\Pi\left(t_k\right) = \sum_{i=1}^{l} n_i(t_k) I_i\left(t_k\right) + \sum_{j=1}^{m} m_j(t_k) S_j\left(t_k\right)$$

easily suggests a state space interpretation of (12) as

$$\begin{aligned} \Pi\left(t_k\right) &= \sum_{i=1}^{l} n_i(t_k) I_i\left(t_k\right) + \varepsilon(t_k) \\ n_i(t_k) &= g(\mathcal{I}_{t_k}) \end{aligned} \qquad (15)$$

where $\mathcal{I}_{t_k}$ represents some set of information available at t_k. The first line of (15) is usually termed "observation equation" and the second "state equation". The first hypothesis we make is on the structure of the state equation.

State evolution hypothesis. the state equation is given by:

$$n_i(t_k) = n_i(t_{k-1}) + u(t_k) \qquad (16)$$

where the terms $u(t_k)$ are uncorrelated zero mean random variables with constant variance.

This is a very strong hypothesis and for many reasons cannot be accepted as a reasonable description of an investing behaviour. Other assumptions involving only n_i could be considered, such as an i.i.d. hypothesis:

$$n_i(t_k) = n_i + \varepsilon_i(t_k) \quad i = 1, ..., l-1, \qquad (17)$$

or a stable AR(1) hypothesis:

$$n_i(t_k) = \rho_i n_i(t_{k-1}) + \varepsilon_i(t_k) \quad |\rho| < 1, \;\; i = 1, ..., l-1, \qquad (18)$$

Moreover it is indeed reasonable to assume that investment decisions at time t_k depend on a set of available economic indicators expressing investment opportunities. In this case a specification of the kind

$$n_i(t_k) = g(X(t_k)) + u(t_k)$$

where $X(t_k)$ represents a set of economic observables, would be much more sensible and, indeed, the study of g could be much more interesting then the simple evaluation of the n_i required by style analysis.

However since the primary aim of style analysis is to evaluate n_i and a detailed study of proper expressions for the function g would require a totally different approach to the problem, we suppose that (16) is at least a sufficiently good approximation of the investment trajectory and concentrate on the observation equation.

The basic problem in estimating the equation

$$\Pi\left(t_k\right) = \sum_{i=1}^{l} n_i(t_k) I_i\left(t_k\right) + \varepsilon(t_k)$$

is that, contrary to standard specifications of state space models, the index processes $I_i\left(t_k\right)$ are stochastic.

Observation equation hypotheses. the standard set of hypotheses for the observation equation are (see e.g. Hamilton (1994) ch. 13.8):

O1-If $I(t_k)$ is the matrix containing all the observations on the indexes I_i up to time t_k and Π_{t_k} is the vector containing all the observations on the fund value up to time t_k then $E(\Pi\left(t_k\right)|I(t_k), \Pi_{t_{k-1}}) = \sum_{i=1}^{l} n_i(t_k) I_i\left(t_k\right)$ or, equivalently $E(\varepsilon(t_k)|I(t_k), \Pi_{t_{k-1}}) = 0$

O2- conditional on $I(t_k), \Pi_{t_{k-1}}$, $\varepsilon(t_k)$ and $u(t_k)$ are independent Gaussian random variables with constant variance.

Under the state and evolution equation hypotheses we can apply the standard result:

Proposition. *If the state and observation equation hypotheses are true, and the variances of u and ε are known then the best linear unbiased estimate of $n_i(t_k)$ is given by the Kalman filter (see, e.g., Hamilton, 1994, ch. 13).*

When the variances of ε and u are unknown it is possible, under some additional hypotheses we do not discuss here (see e.g. Hamilton (1994)), to derive a consistent estimate of the best linear unbiased estimate of $n_i(t_k)$ given by the Kalman filter.

In the end, under some approximate hypothesis for the time evolution of n_i, O1 and O2 and other technical conditions, there exist a viable solution to the basic problem of style analysis, that is the estimation of n_i. The stance of this solution depends on how much hypotheses O1 and O2 can be considered a realistic picture of the problem and this is the point we discuss in the next section.[1]

[1] Since the investment strategy must be self financing we should add the constraint:

$$n_l(t_k) = n_l(t_{k-1}) - \sum_{i=1}^{l-1} \varepsilon_i(t_k) I_i\left(t_k\right)/I_l(t_k).$$

However, for clarity, while we use this constraint in the empirical part of the paper we prefer to omit it from the general discussion.

5. On the observation equation hypotheses

We start with hypothesis O2 as this is the less problematic to deal with.

Since

$$\varepsilon(t_k) = \sum_{j=1}^{m} m_j(t_k) S_j(t_k)$$

the hypothesis of independence between $\varepsilon(t_k)$ and $u(t_k)$ (conditional on $I(t_k)$, $\Pi_{t_{k-1}}$) is equivalent to the assumption of independence between the value of the actively managed part of the portfolio and the innovations of the n_i for the passive part of the portfolio. While it is possible to consider investment strategies where active and passive strategies are correlated, these strategies do not seem to be frequent in the investment community. Alternatively we could choose this hypothesis as a complement to the accounting definition of the difference between active and passive strategies we gave in (4).

The second and last requirement of O2 is that of a (conditionally) constant variance. From the financial point of view this is not an unreasonable assumption. We could derive this from the hypothesis that the fund manager's must follow risk management rules that imply limits to the risk of active investing. Even if this hypothesis should be considered too strong, the main consequence of its assumption would be on the efficiency of estimates and, at least at the level of an exploratory analysis, its effect could be considered as negligible, as the effect of possible autocorrelation of the $\varepsilon(t_k)$.

We do not comment on the gaussian assumption, which is required both in order to derive the optimality of estimators for stochastic I_i and to estimate the variances of ε and u when these are unknown. The role and the limit of this hypothesis in the context of finance has been dealt with in depth in many papers. Our position is to accept this as a working hypothesis, while perfectly conscious of its weakness.

From our point of view the real problems are connected with Hypothesis O1.

In order to see why, suppose for the moment that both $n_i(t_k)$ and $m_j(t_k)$ are time invariant.

If we let: W be the (lxm) matrix with (constant) elements n_{ij}, N be the $(1xl)$ vector with elements n_i, M the $(1xm)$ vector with elements the m_j, $S(t_k)$ the $(mx1)$ of the $S_j(t_k)$ and $I(t_k)$ the $(lx1)$ vector with elements the $I_i(t_k)$. From:

$$Cov(NI(t_k), MS(t_k)) = 0$$

we derive:

$$NWCov(S(t_k))M' = 0 \tag{19}$$

We see that (19) can hold, for a given set of index weights W, only if the covariance matrix of security prices $S(t_k)$, the passive allocations N and the active allocations M, satisfy a strict relation which is totally unrealistic to suppose valid for all funds under analysis. The most simple case requires that $Cov(S(t_k)) = f(t_k)C$, where C is any positive definite (mxm) matrix such that:

$$NWCM' = 0.$$

and $f(.)$ is a function of time. Since there is no reason for supposing that the manager of the fund under scrutiny follows such a specific investment policy, or that security prices adapt to the manager's strategy, the Kalman filter, which in this case reduces to a simple OLS regression of $\Pi\,(t_k)$ on the $I_i\,(t_k)$ shall, in general, yield biased and inconsistent estimates of the n_i.

We can expect this problem to be present also in the general case when $n_i(t_k)$ and $m_j(t_k)$ evolve with time.

Under hypothesis O1 we can write

$$E(\varepsilon_i(t_k)|I(t_k), \Pi_{t_{k-1}}) = 0 \tag{20}$$

That is: we assume a (conditional) constant (0) expectation for the value of the active part of the portfolio.

Moreover:

$$\varepsilon_i(t_k) = \sum_{j=1}^{m} m_j(t_k) S_j\,(t_k)$$

and the indexes I_i are linear combinations of the S_j :

$$I_i(t_k) = \sum_{j=1}^{m} n_{ij}(t_k) S_j(t_k)$$

where $n_{ij}(t_k)$ are non stochastic and known. In other words, in order for hypothesis O1 to hold, we should assume that a (random) linear combination of the $S_j\,(t_k)$, that is the active strategy, and the values of other linear combinations of the same prices, that is the indexes and the past value of the portfolio, are regressively independent.

This, in general, is not an admissible hypothesis as it implies a constraint on the fund manager active strategy which shall hold only in very specific cases. The crucial point is not the independence of the active strategy from past values of the portfolio, which could be accepted as a sufficiently general hypothesis but the independence of the active strategy from the set of indexes.

It is interesting to notice that there actually exist an important class of active strategies whose behaviour can be formalized by hypothesis O1.

"Market neutral" strategies are loosely defined as active management strategies where the fund manager position "does not depend on the market". Assumption O1 is a way to formalize this: if "the market" is represented by the indexes I_i then (20) requires regressive independence of the active strategy from the market.

While it is by itself an interesting fact the correspondence of a statistical hypothesis, derived from the requirement of finding reasonably good estimates for a set of parameters, with an actual investment strategy, the market neutral strategy, frequently adopted by hedge fund managers, is by no means the strategy of choice of all fund managers who mix passive and active strategies. If the active strategy of a fund manager is not market neutral (for example if the choice is for a trend following or a contrarian strategy) O1 shall not hold and the estimates of $n_i(t_k)$ shall be biased.

Notice that the delicate point of O1 is the requirement of constant (in fact equal to 0) *conditional* expected value for $\varepsilon(t_k)$. The simple requirement of constant unconditional expected value could easily be accounted for, at least approximately, by adding to the states $n_i(t_k)$ a new state $n_0(t_k)$ representing the time varying expectation of $\varepsilon(t_k)$.

In the end we can summarise the results of this section with the following proposition.

Proposition. *Hypotheses O1 and O2 imply that a Kalman filtering procedure shall yield optimal estimates of $n_i(t_k)$ only if the active strategy of the fund manager is market neutral.*

As hinted at before, when the variances of ε and u are unknown we can derive a consistent estimate of the optimal estimate of $n_i(t_k)$ given by the Kalman filter. For the meaning of "optimal" in this context we refer again to Hamilton (1994) ch. 13.8.

This result, while quite straightforward, is of central importance for any style estimation effort. Its meaning is that no style estimation is possible without an hypotesis on the active behaviour of the fund manager. A model specified in levels and estimated with the Kalman filter can give good estimates of the passive strategy n_i only when the active strategy is of the market neutral kind (a very special case of this is when there is no active strategy). Obviously the use of a different model, such as a return based model, cannot solve the problem. In fact it shall make it worse by adding to it the nonlinearities discussed in the preceding section.

However O1 is not the only hypothesis under which style analysis is possible. An equivalent analysis can be based on

O1'- Let $I(t_k)$ is the matrix containing all the observations on the indexes I_i up to time t_k, Π_{t_k} the vector containing all the observations on the fund value up to time t_k and Z_{t_k} a matrix containing at least l processes observed up to t_k such that $E(\Pi\,(t_k)\,|I(t_k), \Pi_{t_{k-1}}, Z_{t_k}) = \sum_{i=1}^{l} n_i(t_k)\widehat{I_i}\,(t_k)$ where $\widehat{I_i}\,(t_k)$ is the expected value of $I(t_k)$ given Z_{t_k} and is not constant.

Hypothesis O1' requires for the existence of an (observable) set of instruments for the indexes I_i, that is a set of observable processes correlated with the indexes but conditional to which the errors ε have constant (zero) expectation.

The main result of this section is this: since we assume the active part of the fund manager strategy not to be observable, style analysis is only possible jointly with the choice of a set of observable variables which are correlated with the indexes I_i but uncorrelated with the active strategy. If this set is empty, we are assuming either that no active investment exists or that it is uncorrelated with the indexes that is: the active strategy is market neutral. Any other choice of "instruments" allows for style analysis of the passive part of the investment by constraining the active strategy. This choice can be seen as a set of identifying restriction necessary in order to obtain an unique solution out of the style model and cannot be avoided.

In the following section we shall summarize our results and comment about possible choices for the set of instruments.

6. A summary of the results and a comment on instruments

All style analysis implementations we are aware of are based on return models where the "weights" $\omega_i(t_{k-1})$ are estimated directly and not through a previous estimate of the $n_i(t_k)$. Often the $\omega_i(t_{k-1})$ are constrained to sum to 1 and to be positive. Moreover any active strategy is supposed to be either totally absent or to be of the market neutral kind (actually, this problem does not seem to have been considered before).

The results in the preceding sections show that these are three critical points which could heavily bias the result of a style model. How big this bias is mostly depends on unobservables like the actual portfolio composition. However the bias shall tend to be greater when the indexes are not very much correlated (as it should be if they are "exclusive" as required by Sharpe) and when they vary considerably in value during the sample period and these are observable properties. Moreover while a good fit of the style model to the data is not a guarantee of good style parameter estimates, a considerable lack of fit is a clue for the presence of a strong active component in the investment strategy which, if not orthogonal to the index set, shall imply biased estimates for the style parameters. Notice that these are qualitative results motivated by extensive empirical experiments of which we shall give some example in the next section. Nonlinearities implicit in the traditional return based model actually make any general theoretical result very difficult to establish.

The use of a state space, level based, model allows us, in the preceding section, to find reasonable conditions under which style analysis is possible. The most relevant of these conditions, and also the most interesting from both the statistical and the financial point of view, is the requirement for the choice of a set of instrumental variables correlated with the index set but orthogonal to ε, the active strategy value.

This choice is equivalent to characterizing the active strategy (if any) followed by the fund manager. We stress the fact that our definition of active strategy is just an accounting convention (the most intuitively clear in our opinion) needed for defining the meaning of the word "style" when the fund manager does not invest only in indexes. Other accounting conventions could be conceived and to each of these conventions shall correspond a different definition of style. However, the intuition behind our choice of an accounting definition allows us to describe several possible choices of instruments which correspond to a formal definition of a set of possible investing strategies.

There does not exist a proper catalogue of active investment strategies. In fact, widely different strategies are often tagged with the same name. Here we consider the three most frequently discussed active strategies and suggest ways for deriving from their description a suitable set of instruments.

Market neutral strategies. This strategy was described in the previous section. The basic property of this strategy is that of being orthogonal to a given set of market indexes. If these indexes are the same I_i which define the fund manager style the assumption that the active strategy of the fund manager is market neutral is equivalent to the assumption of regressive independence of the active strategy result (ε) from the indexes I_i. This is not the only possible market neutral strategy.

The active strategy could be orthogonal to a set of market indexes different than the set defining the possible styles (maybe with non empty intersection). If this is the case, instrumentation of the I_i would be necessary in order to derive reasonably good estimates of the n_i.

Relative value strategies. The simplest example of a strategy of this kind is the partitioning of the S_j contained in an index in two subsets: the subset of overvalued securities and the subset of the undervalued securities. The method used in order to partition the set of securities is not relevant for our analysis and, in fact, a whole set of suggestions, ranging from the use of econometric models to the use of black sorcery has been put forward. The relevant point is that relative value strategies usually require for the value of the investment to be independent on the value of the partitioned index so that this set of strategies usually falls in the set of market neutral strategies.

Macro strategies. In this case the investment rule selects securities depending on the value of a set of economic variables. For instance, signals of a positive cycle in the economy shall induce the fund manager to increase the investment in capital goods producing firms and correspondingly to close positions in fixed income securities. If these investment choices are performed through a modification of the n_i that is: through a change of style, these strategies shall not impact the active part of the portfolio. The problem arises when these strategies imply the buying or selling of other securities than the style indexes. A possible solution is to characterize the macro strategy using a set of economic variables from which the value of the strategy linearly depends. If this is possible, then any other set of economic variables orthogonal to this characterizing set and correlated with the indexes I_i shall be a viable set of instruments.

While it is possible to suggest a reasonable choice of instruments describing the most frequently implemented active strategies, we do not want to suggest that this possibility solves the problems of style analysis. Any fund manager could use any one of the three described strategies, mix them or follow other investing criteria and the consequence of the choice of the wrong set of instruments shall be an error in style estimation. The lack of consideration of this problem can result in an overvaluation of the power of style analysis, especially when this technique is applied to investment funds which are very likely to follow strategies not restricted to a simply defined set of indexes. As an instance of this we cite the paper of Fung and Hsie (1997) where style analysis is performed on a set of hedge funds, which are likely to be characterized by a strong component of active strategy, while no care is given to possible correlation of the passive and the active part of the strategy.

There are cases when the correlation problem still allows for a useful application of style analysis. A major instance of this is the use of style analysis as a tool for verifying the correspondence between the investment strategy of the fund manager and the style agreed with the client or described in the fund prospectus. If the official fund strategy is described in terms of reference indexes (benchmarks) and if the question posed to the style analyzer is simply "is the fund manager faithful to the official fund investment policy" the answer can simply be based on a style model.

However, in the case of a negative answer, the style model, by itself, shall not be sufficient to characterize the actual strategy followed by the fund manager.

An alternative way for finding good instruments can be based on tests in the Hausman (1978) class (see also Davidson and Mc Kinnon, 1989). We do not follow this suggestion in this paper for two reasons. First, the need for identifying a considerable number of instruments (at least as many as the indexes I_i) creates problems with the use of Hausman test which is more indicated for testing the "exogeneity" of single series, this requires detailed study by itself. Second, since the choice of a set of instruments is equivalent to the specification of an active strategy style, we believe that this choice should be based on a direct study of the available information concerning the fund. We intend to pursue this line of research in a future paper.

In summary: style analysis is possible, and statistical models for performing it exist. Linear models in returns can be considered, at most, as rough approximations in a very particular case. However even the hypotheses underlying more reasonably level based statespace models cannot, in general, be accepted uncritically as they imply a specific characterization of the active investment strategy. If this assumption does not correspond to the actual behaviour of the fund manager, the estimation of style parameters shall be biased. By choosing sets of instrumental variables many different active investment strategies can be taken into account. This choice, as well as the choice of the set of indexes I_i, constitute the main contribution and responsibility of the analyzer.

7. Empirical test

This section does not present a thorough investigation of the style behaviour of a set of funds. Its only aim is to offer some examples of the empirical relevance of the techniques we suggested in this paper as compared to standard return based style analysis. In order to do so we base our analysis on a dataset which is substantially identical to the one used in Sharpe (1992). The dataset contains 11 different indexes with monthly return data from January 1985 to December 1989. The indexes are: short-, medium- and long-term U.S. Government bond (Tbills, Intbds, Lngbds), corporate bonds (Crpbds), foreign bonds (Forbds), large capitalization stocks, mid-capitalization stocks, growth stocks and small stocks (Valstx, Medval, Medgth, Smlstx), euro stocks and Japanese stocks (Eurstx, Jpnstx). The only differences with Sharpe's choice are that we drop the mortgage index and we split the mid-size stock into value and growth while Sharpe does this for the large capitalization stocks. The first choice is made in order to avoid collinearity and the second is more in line with the current use in style analysis. Since our purpose is that of comparing different statistical procedures for style analysis we do not discuss further the definitions of the indexes used in this section and refer for this to Sharpe (1992).

We notice, however, that the analysis of this dataset is quite tricky for at least two reasons: the number of observations is quite small compared to the number of indexes (only 60 observations) and, while not being collinear, the different series show relevant correlations (with the exception of small stocks, correlations range between 0.8 and 0.95). In this case the indexes do not seem to satisfy the "mutually exclusive" requirement of Sect. 2. Indexes do not overlap, that is: no two of them contain the same securities, however indexes show a lot of comovement, they seem driven by similar economic factors. We shall see that this feature of the data does not create problems to well specified models. While we stick to this set of indicators in order to compare our results to the results of Sharpe, our suggestion is for a more restricted set of indexes.

Based on this data we perform two kinds of analysis, the first is based on fictitious funds of which we do know the composition, the second is based on a real world fund of which we do not know the composition.

Example 1. We derive index levels from December 1984 to December 1989 and with these we build two fictitious mutual funds with no active investment as defined in (3). The allocation in the two funds is, in a sense, extreme. The first fund follows a simple buy-and-hold strategy where the fund manager invests an equal amount of funds in each index at fund inception. The second fund follows the same strategy up to mid-sample (obs. 30), but at mid-sample 50% of the investment in the last 5 indexes is transferred to an investment in the first 6 indexes: the allocation is then unchanged until the end of the sample.

We analyze the style of these two funds using five different models. The first model is the standard return-based (constrained) OLS regression: the returns of the fund are regressed on the returns of the indexes. The second model is a return-based rolling-window (constrained) OLS regression: the returns of the fund are regressed on the returns of the indexes but only a moving window of two years of observations is used to take into account possible parameter evolution. The third model is a simple level-based OLS: the value of the fund is regressed on the values of the indexes. The fourth model is the state space model (15), which is estimated using an extended Kalman filter: we assume that the errors u and ε are uncorrelated with constant variance (for u) and covariance (for ε) and the unknown parameters in the model are estimated using iteratively the Kalman smoother and maximum likelihood as described in Koopman et alii (1999). For comparison with the Kalman smoother we produce a fifth estimate of level-based model where we estimate the parameters with a kernel regression. The kernel weight for time t_i is $exp(-((t_i - t_k)^2)/(2h))$, where the choice for the bandwidth $h = 10$ implies that 99% of data weight is concentrated in the ± 10 months interval around t_k. This method is an easy-to-implement approximation of the Kalman filter algorithm: indeed, in our applications the results of the two methods are very similar. All algorithms are implemented using Eviews 4.0.

We would expect the following results. Both return-based OLS and rolling-window return-based OLS should give poor estimates of investment weights for both funds. Level-based OLS should give almost exact estimates for the first (constant allocation) fund but should not perform too well for the second (jump allocation) fund. The state-space model and the kernel regression on levels should give almost exact estimates for the constant-allocation fund whereas their usefulness is debatable for the jump-allocation fund.

The results can be summarized in many ways. We choose to present two tables: one for each strategy. In both tables, each row corresponds to a different index,

Table 1. No changes in portfolio composition*

	OLS rets	WOLS rets	OLS price	K.Filt. price	Kernel price
TBills	0.0561(0)	0.0394 (0)	0.2388 (−13)	0.1623 (−12)	0.1197 (−10)
IntBds	0.0694	0.0512	0.1602	0.1685	0.1700
LngBds	0.0390	0.0324	0.2357	0.1015	0.0852
CrpBds	0.0621	0.0515	0.1746	0.2316	0.1113
ValStx	0.0343	0.0250	0.4447	0.1391	0.0512
MedVal	0.0293	0.0316	0.3408	0.1719	0.0646
MedGth	0.0468	0.0268	0.1103	0.1232	0.0462
SmlStx	0.0491	0.0288	0.2382	0.1430	0.0543
ForBds	0.0347	0.0236	0.1147	0.1096	0.0689
EurStx	0.0184	0.0153	0.3279	0.0876	0.0632
JpnStx	0.0434	0.0241	0.3049	0.0420	0.0368

* The numbers in brackets are power of 10 to be multiplied by the entries in each column

each column corresponds to a different estimation method. The numbers in the cells correspond to the mean absolute deviation between the series of the true weights (ω_j) and the series of the estimated weights of each index in the portfolio. A negative number in parentheses at the top of the column means that the values in that column values must be multiplied by 10 raised to that number.

We begin by examining the results in Table 1. In this case the n_j are constant but the ω_j change in time as in (9). The initial condition is $\omega_j(0) = 1/11$. The first and second column report the mean absolute errors as computed by standard style analysis using a simple return based OLS regression and rolling OLS regression. The errors for OLS are quite sizable as they are between 20% and 70% of the value of the (initial) weights. A rolling regression gives slightly better results: the errors are now between 10% and 60% of the initial weights.

On the contrary, and as expected, OLS on levels, the state-space model and the kernel regression give an almost exact result with only minimal numerical errors. In summary, even in an ideal setting with a buy and hold strategy, the OLS models on returns and its rolling counterpart give bad estimates of the investment weights. On the other hand all thee level model result in an exact evaluations of the investment weights.

Table 2 displays the results for the case of abrupt change in investment policy. In this case we do not expect to be able to correctly track the evolution of investment weights even with the Kalman filter and the kernel model. The standard return-based OLS and rolling-window return-based OLS give unsatisfactory results, with errors comparable to those in Table 1. On the other hand, the report for the level-based OLS, Kalman filter and kernel regression are only marginally better. In fact, the summary indicator is misleading: a detailed analysis of errors over time shows that, while the return models display a biased valuation of the investment weights for the whole sample length, the Kalman filter and the kernel model are quite good in tracking the weights in the first and last part of the sample but undergo a huge degradation of results at mid sample. The change in policy is too abrupt and it can only be partially tracked by models which allow only for a slow evolution of the weights.

Table 2. Sudden jump: sell 50% of each of (MedGth, SmlStx, ForBds, EurStx, JpnStx) to buy 50% more of each of (IntBds, LngBds, CrpBds, ValStx, MedVal). The change in TBills makes the trade self-financing

	OLS rets	WOLS rets	OLS price	K.Filt. price	Kernel price
TBills	0.0452	0.0578	0.0547	0.0493	0.0311
IntBds	0.0909	0.0596	0.0265	0.0349	0.0268
LngBds	0.0441	0.0428	0.0113	0.0328	0.0297
CrpBds	0.0602	0.0542	0.0381	0.0482	0.0380
ValStx	0.0482	0.0465	0.0511	0.0573	0.0156
MedVal	0.0760	0.0550	0.0509	0.0590	0.0456
MedGth	0.0565	0.0374	0.0121	0.0210	0.0117
SmlStx	0.0792	0.0457	0.0134	0.0322	0.0202
ForBds	0.0470	0.0186	0.0542	0.0422	0.0260
EurStx	0.0239	0.0146	0.0139	0.0167	0.0151
JpnStx	0.0250	0.0256	0.0295	0.0182	0.0207

Table 3. Style analysis of the Vanguard index

	S.S. Av.	Sharpe	Av. abs. diff.
TBills	0.0149	−0.0723	0.0873
IntBds	0.1688	0.4638	0.2943
LngBds	0.0930	−0.0156	0.1087
CrpBds	−0.2607	−0.1655	0.0959
ValStx	0.9963	0.8718	0.1242
MedVal	−0.4637	−0.2374	0.2265
MedGth	0.4359	0.3847	0.0515
SmlStx	−0.0254	−0.0546	0.0291
ForBds	0.0212	−0.0063	0.0275
EurStx	0.0589	0.0186	0.0401
JpnStx	−0.0393	0.0036	0.0426

Example 2. We now compare style estimates derived by Sharpe technique and by our statespace level model for a real world fund: the Vanguard 500. This fund is particularly interesting because its investment strategy is known: it tracks the Standard and Poors 500 (S&P500) stock index. While not an exact replica of the S&P500 index the Vanguard 500 fund is, in practice, an index fund. We analyze the style of this fund using both Sharpe's model and our model and Sharpe's dataset, which does not contain the S&P500 index, but different stock indexes. Our main interest here is to compare the results of the two models in the case of a slight mispecification. For the sake of brevity in this case we only compare Sharpe's return based model with a statespace specification whose parameters are estimated using the Kalman smoother and maximum likelihood as in Example 1.

We report the results in Table 3.

Since our aim is that of comparing parameter estimates obtained using the two models, in Table 3 we report the mean weights for the state space model (S.S. Av.) the values for Sharpe model (Sharpe) and the average absolute differences for the series of investment weights derived using the two models. Notice that, in this case, we do not constrain any estimate to be positive. This is due to the fact that the Vanguard fund officially tracks the S&P 500 index which is not an index considered in the style set. On the other hand, Sharpe's index set contains several style indexes for stocks (large cap, growth etc.) and some bond indexes a linear combination of which can effectively track the S&P500.

We expect the fund to be concentrated on the value stock index (large stocks) with some investment in the other three stock indexes (maybe a long/short spread) and possibly some position in corporate bonds (which are usually highly correlated with the corresponding stocks) compensated by an opposite sign position in long bonds (since the Vanguard fund should not be exposed to changes in interest rates).

In fact the results in the first and second columns of Table 3 roughly correspond to expectations. The main position is in the value stocks index, to this position it must be added a spread position long in medium growth and short in medium value stocks, negligible positions in the small stocks and foreign stocks indexes and a spread position short in corporate bonds and long in medium/long maturity government bonds.

The stock positions correspond to expectations: the S&P500 is an index containing high capitalization stocks, so the main weight is on the value index. On the other hand, over time, the index shall change composition due to the entrance of fast growing stocks and the exit of companies which lag behind the market. This explains the long position in medium growth and the short position in medium value stocks. A similar explanation can be given to the medium/long term government bonds against corporate bonds. Such a position is not dependent on the interest rate level while it depends on the credit of the firms issuing bonds. In practice such spread is equivalent to selling stocks of the firms with worsening credit. Since stocks of firms with credit problems tend to lag behind stocks of more reliable firms, this spread position could be considered as a second correction to the value index which takes into account the possible exit from the S&P 500 index of large firms with credit problems.

From the results in columns 1 and 2 of Table 3 we see that, qualitatively, the average styles as estimated by the Sharpe model and the state space model are quite similar, however column 3 shows that the time evolution of portfolio weights estimates are considerably different. The most striking differences can be found in the value stock, and in the government bonds weights.

On average the absolute difference between Sharpe and state space estimates of the value stock weights is of 0.1242 and, by a comparison of columns 1 and 2 we see that this is due to the fact that the state space model estimates an higher concentration of the Vanguard fund investment in this class of stocks.

For what concerns the government bond weights Sharpe model concentrates all the weight on intermediate bonds while the state space model divides the weight between medium and long term bonds.

A more thorough study of the S&P500 index is required in order to take a stance about the appropriateness of Sharpe of the state space model. A clear feature of the results, however, is that, while qualitatively similar, the results from the two model show relevant specific differences. Since the linear in returns model of Sharpe can

Table 4. Statespace level based models estimates over n =1000 Monte Carlo runs (true $n_i = 0.25$)

		n_1	n_2	n_3
Statespace model	average	0.316	0.316	0.308
no instruments	std.dev	0.0056	0.0048	0.0050
Statespace models	average	0.243	0.240	0.241
with instruments	std.dev	0.0040	0.0043	0.0039

be considered as a particular case of the linear in levels state space model, the fact that a direct estimation of this model leads to estimates which differ from those derived from Sharpe model seems to imply that the correctness of a return based model is, at least in this case, under questioning.

Example 3. In both Example 1 and 2 we relied, either by construction or by reasonable hypothesis, on the fact that the strategy of the fund is actually concentrated in the set of indexes used for the analysis. In this case no provision has to be made regarding the choice of instrumental variables. In the last example we build an artificial dataset in order to show how, in the general case, the choice of proper instruments is relevant in order to derive good style estimates.

We build an artificial dataset containing 70 simulated observations on four "security prices" defined in this way:

- Three factor series $F_i(t)$ are generated, the returns in each series are i.i.d. standard gaussian and the series are non cross correlated.
- An error series $E(t)$ is generated, the observations are i.i.d. and the distribution is standard gaussian. This series is not correlated with the factor series.
- Three index series $I_i(t)$ are generated by summing each $F_i(t)$ to the $E(t)$.
- The portfolio value series is generated by multiplying each index series and the error series times .25 and summing the four resulting series.

Suppose now that only data on $I_i(t)$ $i = 1, ..., 3$ and on the portfolio values are available, the error series can be considered as the result of an active strategy. Using this dataset we want to estimate the values of $n_i(t)$ $i = 1, ..., 3$. (0.25). By construction the omitted term is correlated with the three included indexes. However, if the $F_i(t)i = 1, ..., 3$ are observable, we can apply model (15) using $F_i(t)i = 1, ..., 3$ as instruments.

In order to evaluate the performance of the various models described in this paper we perform a Monte Carlo experiment. The above described data are simulated in 1000 runs. For each run we estimate the $n_i(t)$ with the Kalman filter both with and without instrumental variables. In Table 4 we display the sample average (both over time and over the different Monte Carlo runs) of the estimate of each $n_i(t)$.

Not unsurprisingly, the estimates not based on instrumental variables show an upward bias of about 0.06 which is more than ten times bigger than the Monte Carlo standard deviation, on the other hand the instrumental variables version of the Kalman filter show values within 0.01 from the theoretical value.

It is probably not fair to compare Sharpe's estimates to those obtained by using level models as in this example both return weights are not constant and the error is correlated with the index variables. However it is interesting to notice that, when converted into $n_i(t)$ terms, the Monte Carlo standard deviation of Sharpe's estimates is bigger then the corresponding standard deviation for both level models by a factor greater than 6.

While very simple and schematic this example shows how a possible correlation between indexes used for style estimation and omitted components of the portfolio value can degrade the estimation result.

8. Conclusions and further developments

Style analysis is one of the most common applications of statistics in both theoretical and applied finance.

The most widespread approach to style analysis was defined by Sharpe (1992) and is based on a standard or a rolling-window OLS regression of fund returns on index returns. In this paper we show that, under different assumptions, both lax and restrictive, the usefulness of models linear in returns to estimate the composition of a fund is questionable. A simple variant of Sharpe model, i.e. a model linear in levels, solves the problem when the investor actually invests only in the index set. In the general context, where a general active strategy is allowed, we describe a general statespace level based model for style analysis. However we show that the problem cannot be solved using only a set of indexes, unless the active strategy is orthogonal to the indexes.

In the, more than likely, case of correlation between strategies, an appropriate set of instrumental variables needs to be added to the regressors. The instrumental variables choice is not arbitrary but implies an assumption on the fund manager active behaviour. We show how three widely applied active strategies can be converted in the choice of instruments.

We conclude the paper with a short empirical section describing the performance of the different models introduced in the paper.

In this paper we concentrated on two main objectives: to show that a return based model cannot work in a reasonably general setting and to show that any style analysis requires (or implies) the choice of a set of instrumental variables. In doing so we left out many interesting developments which deserve further study.

We would like to just hint to two of these topics. The first, already mentioned in the paper, is the study of a set of statistical tools, based on the Hausman test procedure, for choosing appropriate instruments. The second is the specification of the model in difference terms. While the use of returns creates difficult estimation problems, the use of level differences does not introduce non linearities in the estimation procedure and could be the appropriate choice in the case of non stationary (integrated) active strategies.

References

Beasley JE, Meade N, Chang TJ (2001) An evolutionary euristic for the index tracking problem. Mimeo, Imperial College, London

Brown SJ, Goetzmann WN (1997) Mutual funds styles, Journal of Financial economics 43: 373–399

Brown SJ, Goetzmann WN (2001) Hedge funds with style. W.P. N.B.E.R n. 8173

Corielli F, Marcellino M (2002) Factor based index tracking. D.P. C.E.P.R. n. 3265

DeRoon FA, Nijman TE, TerHost JR (2000) Evaluating style analysis. W.P., Tilburg University

Davidson R, Mc Kinnon JG (1989) Testing for consistency using artificial regressions. Econometric theory 5: 363–384

Fung W, Hsieh DA (1997) Empirical characterizations of dynamic trading strategies: the case of hedge funds. Review of Financial Studies 10: 275–302

Fung W, Hsieh DA (1998) Performance attribution and style analysis: from mutual funds to hedge funds. W.P., Fuqua school of business, Duke University

Hamilton JD (1994) Time series analysis. Princeton University Press, Princeton

Hausman JA (1978) Specification tests in econometrics. Econometrica 46: 1251–1272

Koopman SJ, Shephard N, Doornik JA (1999) Statistical algorithm for models in state space using Sskpacl 2.2. Econometrics Journal 2(1): 107–160

Sharpe WF (1992) Asset allocation: management style and performance measurement. J. of Portfolio Management 18: 7–19

Swinkels L, Van Der Sluis PJ (2001) Return based style analysis with time varying exposures. W.P., Tilburg University