Abstract — Multi-factor models constitute a useful tool to explain cross-sectional covariance in equities returns. We propose in this paper the use of irregularly spaced returns in the multi-factor model estimation and provide an empirical example with the 389 most liquid equities in the Brazilian Market. The market index shows itself significant to explain equity returns while the US$/Brazilian Real exchange rate and the Brazilian standard interest rate does not. This example shows the usefulness of the estimation method in further using the model to fill in missing values and to provide interval forecasts.

Keywords: Multi-Factor Model, Missing Data

JEL Classification: C32, C53 e G13

1 - Introduction

Emergent markets frequently suffer from low liquidity and tend to concentrate most of transactions on a few liquid assets. Subramanian (2001, p.77), for example, shares this view for bond markets. As to equity markets, the Brazilian one comprises about 1190 different stocks but almost 40% have not been negotiated in the last year and 32% of the remaining has been negotiated less than once a month. As a consequence, for many stocks there will be no price for a large proportion of days. There will be still more days for which one cannot compute daily returns, since they require the existence of transactions in two subsequent days.

Financial institutions, however, do need those prices and returns everyday in order to fulfil regulatory requirements and implement their methodologies of quantitative
analysis of risk and return. A market model can provide the expected values of missing prices and returns. Sharpe (1964) proposed the Capital Asset Pricing Model (CAPM) to explain asset returns. However, a number of papers provided empirical evidence against the CAPM. For example, Bhandari (1988) and Chan, Hamao and Lakonishok (1991). Furthermore Fama and French (1992) and Jegadeesh (1992) show that the market beta has little power in explaining cross-sectional asset returns, meaning that some common "factor" could further explain the returns. Ross (1976) proposed the Arbitrage Pricing Theory (APT), which allowed more than one factor to explain the assets returns and consequently diversify risk premia. However, these factors are non-observable. Reinganum (1981) and Mei (1993) use autoregressive approach to explain the hidden factors. Chen, Roll and Ross (1986) introduced macroeconomic variables to explain stock returns in a multi-factor linear regression. It is common practice in financial institutions to use this model to explain assets returns, where the factors are called risk factors.

Then a circular problem may arise: these market models must be estimated before computing the expected value of missing prices and returns, but if the proportion of missing data is high many numerical problems may surge and bias the estimation. A biased estimation will lead to poor price filling. Note that models are usually for returns since they are statistically more attractive than prices, so that it is the number of missing daily returns that matters. In this paper, we propose an estimation method to the multi-factor model which makes use of irregularly spaced returns, enabling the use of every historical price available and thus increasing efficiency. In other words, it permits calculations with percentage of existing daily (closing) prices as low as 5% in a year of data, whereas traditional estimation do not.

The multi-factor model can then be used to estimate missing prices and also to estimate the h-step ahead interval forecast for returns, \( h = 1, 2, 3 \). The motivation to use up to three steps ahead comes from the fact that the Brazilian Market liquidity system determines that parties involved in a transaction have three days to liquidate it.

To evaluate the estimation method, we ran a back test with 389 stock prices over the period ranging from August 25, 1998 to February 28, 2001. Two main important
aspects of the estimated model, with practical implications to the users of the data were examined. First, we examine its ability to produce good estimation of missing prices. This was done comparing each observed price to the estimated value that would be produced by the model if this particular price was missing. Second, the h steps ahead interval forecasts were evaluated by comparing the nominal and observed frequencies of values in the 5% tails, which is an important measure in risk analysis.

For a comparative evaluation, the same experiment was run to some other imputation methodologies frequently used by practitioners such as repeating the last price or mimicking the market index return, as well as the same multi-factor model estimated only with the available daily returns. All these are nested under the same multi-factor class of models. We conclude that the multi-factor model estimated using irregular returns globally outperforms all the others, including the traditional regular returns estimation. Furthermore, the Brazilian Equity Market Index show to significantly explain equity returns but the USS - Brazilian Real exchange rate and the Brazilian inter-bank overnight interest rate do not.

The next Section presents the notation used further in this paper and the methodology proposed here, while Section 3 shows an empirical example with Brazilian equity data. Section 4 offers some concluding remarks.

2 - Notation and Methodology

In this Section we describe the notation and the methodology used in this paper. The price of equity \(j\) at time \(t\) is denoted by \(P_{jt}\) and the (log-)return of the same equity at time \(t\) by:

\[
R_{jt} = \ln \left( \frac{P_{jt}}{P_{jt-1}} \right) = \ln(P_{jt}) - \ln(P_{jt-1}),
\]

where \(j = 1, ..., J\) \((J = 389\) in the experiment) and \(t = 1, ..., T\). The advantage of using log-returns instead of simple returns is that the log-return yielded in \(\tau\) days \((\tau\) integer and greater than one) is the sum of the daily log-returns over the \(\tau\)-days period.

2.1 - Multi-Factor Model
A Multi-Factor model describes a linear relation between the return of equity \( j \) and some market indicators, taking the form

\[
R_{t}^{j} = (\beta_{t}^{j})^{T} X_{t} + \epsilon_{t}^{j},
\]

where \( \beta_{t}^{j} \) is the vector of coefficients of equity \( j \) at time \( t \), \( X_{t} \) is the vector of market indicators (possibly log-returns) at time \( t \) and \( \epsilon_{t}^{j} \) errors supposedly iid, \( \epsilon_{t}^{j} \sim N(0, \sigma_{\epsilon}^{2}) \).

We allow the coefficients to vary slowly and smoothly in time to let the relationship between returns and indicators be dynamic. However, the study of this variation is not in the scope of the paper. To deal with this variation in a simple fashion, we estimate all models within the exponential smoothing (EWMA) framework (J.P. Morgan, 1995), which is briefly explained in Section 2.3. This procedure is common standard in practice within financial institutions.

For instance, if the return of a market index and the overnight inter-bank interest rate\(^2\) are used as risk factors (or simply factors), as well as a constant, the model becomes thus:

\[
R_{t}^{i} = \beta_{0,t}^{i} + \beta_{1,t}^{i} X_{1,t} + \beta_{2,t}^{i} X_{2,t} + \epsilon_{t}^{i},
\]

where \( X_{1,t} \) is the market index return and \( X_{2,t} \) the interest rate at time \( t \). In the notation of Equation (2), \( X_{t} = \begin{bmatrix} 1 \\ X_{1,t} \\ X_{2,t} \end{bmatrix} \) and \( \beta_{t}^{i} = \begin{bmatrix} \beta_{0,t}^{i} \\ \beta_{1,t}^{i} \\ \beta_{2,t}^{i} \end{bmatrix} \).

Given there is no missing value in the estimation window, the estimation is obtained via weighted least squares, with the weights exponentially decreasing with the age of the data. The missing value of equity \( j \) occurs when there is no trade of the equity \( j \) at time (day) \( t \). Its imputation at time \( t \) based on this model uses the returns expected values given the current (daily) market indicators at time \( t \). The VaR (Value at Risk, which "measures the worst expected loss over a given time interval under normal market conditions at a given confidence level", Jorion, 1997, p. xiii) estimation \( h \) days ahead, however, requires that the market indexes be forecast.

\(^2\)To have the interest rate measured in log-returns one must construct a curve from the original percentage interest rate and then compute the log-returns. However, for the overnight interest rate, the difference is negligible.
2.2 – Irregularly Spaced Returns

Emergent markets contain a number of illiquid equities which are not negotiated every day, unlike the market indicators. For these equities, the estimation window is scattered with missing values. Considering the existence of these windows, one must pay attention to returns which interval is not one day, but say \( \tau \) days. However, data are available every day. Figure 1 illustrates these non-full estimation windows and the irregular returns.

Figure 1: Estimation window with missing data and irregular returns

Thus, one needs to define the \( \tau \)-days return at time \( t \) for equity \( j \). Let it be:

\[
R_{t, \tau}^j = \ln P_t^j - \ln P_{t-\tau}^j .
\]

Moreover, equalling the right-hand sides of equations (1) and (2) and rearranging yields:

\[
\ln P_t^j = \ln P_{t-\tau}^j + (\beta_j^\tau)^\top X_t + \epsilon_t^j .
\]

Applying Equation (4) to \( t = t, t-1, t-2, \ldots, t-\tau+1 \) and summing up gives:

\[
\ln P_t^j = \ln P_{t-\tau}^j + \sum_{i=0}^{\tau-1} (\beta_{j,i}^\tau)^\top X_{t-i} + \sum_{i=0}^{\tau-1} \epsilon_{t-i}^j .
\]
If for practical purposes $\beta^i_j$ and $\sigma^2_j$, are considered (almost) constant on the interval $[t-k, t]$, and since $\varepsilon^i_j$ is iid, then, using (3), (5) rewrites as:

$$R^i_{t,k} = (\beta^i_j)^T \sum_{i=0}^{t-k} X_{i+j} + w^i_{t,k}, \quad (6)$$

where

$$w^i_{t,k} = \sum_{i=0}^{t-k} \varepsilon^i_j; \quad w^i_{t,k} \sim N(0, \tau \sigma^2_j).$$

### 2.3 - Exponentially Weighted Moving Average (EWMA)

The EWMA framework is also known as RiskMetrics (J.P. Morgan, 1995) and its motivation is that the dynamics of what is under study (e.g., equity data) changes with time but no assumption is to be made upon these changes besides the change is smooth. So, the older are the data the less weight it must have attached to. These weights decrease exponentially with time, driven by the smoothing parameter $\lambda$, $0 < \lambda < 1$. The day $t^*$ return is thus given $\lambda$ times the weight given to the return of $t^*+1$. The parameter $K$, in turn, determines the number of effective days to be used in the estimation window. The mean of equity $j$ returns at time $t$ is then estimated by:

$$\hat{\mu}^j_t = \sum_{i=0}^{K-1} \lambda^i (1-\lambda) R^j_{t-i} . \quad (7)$$

The covariances between returns of equities $i$ and $j$ at time $t$, as well as the variance of equity $i$ are estimated by:

$$\begin{align*}
\hat{\sigma}^2_{i,j} &= \sum_{m=0}^{K-1} \lambda^m (1 - \lambda) (R^i_{t-m} R^j_{t-m}) - \hat{\mu}^i_t \hat{\mu}^j_t \\
\hat{\sigma}^2_{i,i} &= \sum_{m=0}^{K-1} \lambda^m (1 - \lambda) (R^i_{t-m})^2 - (\hat{\mu}^i_t)^2
\end{align*} \quad (8)$$

Depending on $K$, the right-hand sides of Equations (7) and (8) must be multiplied by a constant $C$ so that the sum of weights $C\lambda^m(1-\lambda)$ is the unity ($C \to 1$ as $K \to \infty$). For the sake of simplicity, the constant $C$ was omitted from Equations (7) and (8). We do not take into account the loss of a degree of freedom in calculating the mean. Little effect though there will be in our results since we use $K = 252$ (corresponding to one year of

---

3 There is no data in between $t-k$ and $t$ to estimate how $\beta$ and $\sigma^2$ change. Furthermore, if $\tau$ is small the coefficients $\beta$ and the variance $\sigma^2$ are believed to change little.
data). It is usual to consider the mean of returns as being zero, simplifying more these equations. However, this assumption is not made here. An overview on EWMA is given in Alexander (1996).

2.4 - Methodology
2.4.1 - Estimation with Irregular Returns

Some equities have trading prices for all days in the estimation window. In these cases, Equation (2) is considered and a weighted least squares procedure, with EWMA weights, yields the estimates of $\beta_i$ for each equity $j$ at the current time $t$. This is done through the equation:

$$\hat{\beta}_i^j = (X^T WX)^{-1} X^T WR_i^j,$$

where

$$R_i^j = R_{i-K+1}^j R_{i-K+2}^j \ldots R_{i-1}^j R_i^j$$

is the vector of equity $j$ returns;

$$X = X_{K(t+1)} = \begin{bmatrix} (X_{i-K+1})^T \\ \vdots \\ (X_{i})^T \end{bmatrix} = \begin{bmatrix} 1 & X_{i-K+1} & \ldots & X_{i-K+1} \\ 1 & X_{i-K+2} & \ldots & X_{i-K+2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_i & \ldots & X_i \end{bmatrix}$$

is the design matrix (the same for all equities); and

$$W = W_{KxK} = \begin{bmatrix} \lambda^{K-1} & 0 & \ldots & 0 \\ 0 & \lambda^{K-2} & \vdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \ldots & 0 & 1 \end{bmatrix}$$

is the weights matrix. In this case, $p$ market indicators are considered.

The lines of $R_i^j$, $X$ and $W$ correspond to one day period (daily returns). However, whether there is a missing value in the estimation window, (6) instead of (2) must be considered, so that all lines corresponding to a same $\tau$-days return will collapse into one single line in the redefined matrices $R_i^j$, $X$ and $W$. The redefined matrices are as follows:

---

4 In fact, if all the weights in $W$ are multiplied by the same constant, the estimates do not change, as the change in the inversion of $X^T WX$ compensates the change in $X^T WR_i^j$. That is why the weights do not match with those in 2.3.
\[ R^j = R_{K_j+1}^j = \begin{bmatrix} R_{j,t_1}^{j}\, & R_{j,t_2}^{j}\, & \cdots\, & R_{j,t_{K_j}}^{j} \end{bmatrix}, \quad X = X_{K_j+t} \]

\[ W = W_{K_j} = \frac{\lambda_0^0}{\tau_1} \begin{bmatrix} 0 & \cdots & 0 \\ \frac{\lambda_0^1}{\tau_2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\lambda_0^{K_j}}{\tau_{K_j}} \end{bmatrix} \]

where there are \( K_j+1 \) prices (\( K_j \) possibly irregularly spaced returns) available in the estimation window for equity \( j \); \( t_k, k = 1, 2, \ldots, K_j \), is the interval of the \( k \)-th return in the estimation window; and \( i_k \) is the time past between the \( k \)-th return of equity \( j \) and the current time \( t \) (in other words, the age of the return, measured in days in the present implementation). So, in addition to account for exponentially decaying weights, \( W \) assigns weights inversely proportional to the variances of the errors \( w_{r,t}^j \).\(^5\)

2.4.2 - Missing Values Imputation

In the more general case, the imputation of missing values can occur at any point of the estimation window, although the results with real data shown here are only from imputing data at the end of it (current day). For the sake of simplicity, consider \( p_t = \ln p_t^t \). Following (5), the expected value of \( p_t \), conditional on \( \beta_t^t \), the vector of market indicators \( \bar{X} \) and the next existing price (forward or backward respectively for equations 11.A and 11.B) is given by:

\[ E(p_t \, | \, p_{r-t}, \beta_t^t, \bar{X}) = p_{r-t} + (\beta_t^t)^T \sum_{i=0}^{r-t-1} X_{r-i} \]  

(11.A)

\(^5\) In order to minimize the estimator variance (see, for example, Neter et al, 1996, p.400).
Using (11.A) and (11.B), replacing \( \beta' \) by its estimate \( \hat{\beta}' \), we reach the estimates of a missing price conditional on the next (forward or backward) existing price and the risk factors. The estimates are as follows:

\[
\hat{p}_{j_{i-i}} = p_{j_{i-i}} + (\hat{\beta}' j_{i-i})^T \sum_{i=0}^{t-i} X_{i_{j-i}} ,
\]

(12.A)

\[
\hat{p}_{j_{i+i}} = p_{j_{i+i}} - (\hat{\beta}' j_{i+i})^T \sum_{i=0}^{t-i} X_{i_{j+i}} .
\]

(12.B)

Thus, if the missing value to be imputed is at the end of the estimation window, (12.A) is used. If it is in the beginning, (12.B) is used. Otherwise, (12.A) and (12.B) are combined so as to yield minimum variance, using the fact that the errors with \( j \) fixed (same equity) are iid as follows:

\[
\hat{p}_j = \gamma \hat{p}_{j_{i-i}} + (1-\gamma) \hat{p}_{j_{i+i}} = \frac{\tau_1}{\tau_1 + \tau_2} \hat{p}_{j_{i-i}} + \frac{\tau_2}{\tau_1 + \tau_2} \hat{p}_{j_{i+i}} .
\]

(13)

For details on the minimum variance estimator see Neter et al (1996, p.400).

2.4.3 – Estimation of Interval Forecasts

We consider here the interval forecast (IF) estimation as an ad hoc procedure, ignoring uncertainties of many sorts and making some further simplifying assumptions. The exact parametric predictive density given the specified model would require the account of the variance of estimated coefficients, even if the assumptions hold. If one is to consider the problem through the view of the clearinghouse, which guarantees the transactions, two-sided IF’s must be supplied. This is because both parties may default, which means that the clearinghouse may have to buy or sell the equity in the market three days after the transaction is agreed. The IF is computed as the following. Let \( E(R_{t+\tau, t}) \) be the expected value of equity \( j \) \( \tau \)-days return at time \( t+\tau \) given \( \beta' \) and \( \hat{\beta}' \):

\[
E(R_{t+\tau, t} | \beta' , \hat{\beta}' j_{i}, \hat{X} j_{i}) = E \left( \sum_{i=0}^{t} (\beta' j_{i})^T X_{i_{j}, i} + w_{i, \tau} \right)
\]

(14)
The first simplifying assumption is that the coefficients $\beta$ are independent in relation to the factors $X_t$. Note that this assumption may not hold since the $\beta$'s may be influenced by the factors magnitude. But considering this assumption true, it is straightforward to estimate the expected value of the $\tau$-days ahead return given the information until time $t$ by:

$$
\hat{E}(R_{i,\tau,t} | t) = (\hat{\beta}_{i}')^T \sum_{i=1}^{L} \tilde{X}_{i,t}
$$  \hspace{1cm} (15)

To use (15), the risk factors must be forecast. However, the method used to forecast the risk factors is beyond the scope of the paper. Now we must estimate the variance $\tau$-days ahead, and to facilitate computations one more assumption is made. The second simplifying assumption is that the risk factors are homoskedastic, or at least that their variances vary slowly in the $\tau$-days ahead period. Alternatively to making this assumption, this variance may be viewed as conditional on the current circumstances concerning volatility and risk. As it was assumed before that the coefficients $\beta$ are (almost) constant and the errors are serially independent, the variance of equity $j$ $\tau$-days return at time $t+\tau$ may be approximated by:

$$
\sigma^{2}_{R_{j,\tau,t}} \equiv \tau\left[\sigma^{2}_{j,t} + \hat{\beta}_{i}' \text{COV}(X_j)\hat{\beta}_{i}'\right],
$$  \hspace{1cm} (16)

Ignoring the uncertainty caused by the estimation of the coefficients $\beta$, this variance can be estimated at time $t$ as:

$$
\hat{\sigma}^{2}_{R_{j,\tau,t}} = \tau\left[\hat{\sigma}_{j,t}^{2} + \hat{\beta}_{i}' \text{COV}(X_j)\hat{\beta}_{i}'\right].
$$  \hspace{1cm} (17)

The risk factors variance-covariance matrix (VCV) must be estimated and there are a number of methods to do so. Alexander and Leigh (1997) study the accuracy of some of these methods considering the proportion of returns that fall below the estimated VaR. The IF must then be based on a predictive distribution, which exact form is non-trivial to obtain. If the betas were known, as well as the future values of $X_t$, the distribution would be a Student's $t$ with $K_j - p - 1$ degrees of freedom, but the betas are estimated and the risk factors forecast. Nonetheless, this parametric form will be used in place of the exact one, just incorporating the risk factors uncertainty in the variance as in (17). The lower and upper bounds for a $(1-\alpha)100\%$ IF are given then by:
\[
L = \exp[p_1/(p_2 + E(R^2_{t+s,t} + 1/(\alpha/2,K_j - p - 1))\tilde{\sigma}_{R,t+s,t})] \quad \text{and}
\]
\[
U = \exp[p_1/(p_2 + E(R^2_{t+s,t} + 1/(1-\alpha/2,K_j - p - 1))\tilde{\sigma}_{R,t+s,t})],
\]
where \(t^1(.,v)\) is the inverse cumulative distribution function of a Student's \(t\) random variable with \(v\) degrees of freedom, and \(K_j\) is the number of observations available for equity \(j\) at the estimation window. The lower bound, if desired, can be calibrated using the semi-variance (see Gastineau e Kristzman, 1996, p. 250, for example) to estimate \(\text{VAR}(X_{i,j})\). The semi-variance is the mean of the squared negative deviations from the mean. The VaR of 1-\(\alpha/2\) confidence level is given by the lower bound of the 1-\(\alpha\) IF.

3 – Empirical Example with Brazilian Equity Data

The database consists of 389 stock prices (the most liquid in the Brazilian Market) over the period ranging from August 25, 1998 to February 28, 2001. The results are divided in two sections, the first dedicated to the imputation of missing values and the second to the IF estimation. To access the accuracy of the missing values imputation, each single observation in the database within specified patterns is deleted at a time and filled in as is schematised in Figure 2. This enables the computation of error statistics, where the error is the difference between the filled in and the observed data. In turn, the IF is estimated for every observation in the database within the same specified patterns, one two and three steps ahead, and its coverage verified. The specified patterns are: two subsequent prices available; a price available at \(t\) and \(t+2\); and a price available at \(t\) and \(t+3\). These patterns are motivated by the fact that the parties involved have three days to liquidate the transaction, and thus a clearinghouse must estimate a IF for the equity price up to three days ahead in order to access its risk in the case of default.

3.1 – Missing Values

The imputation of missing values was considered in the present, that is, the data posterior to the day to fill in missing values was considered unknown, but the current market index known. The multi-factor model estimated with irregular returns is compared with the following methods: the \textit{naif}, where the last price available is repeated to fill in a missing value (predicted returns are zero); and the market proxy (the São Paulo Stock
Exchange Market Index), where the equity is supposed to follow the market return in the absence of the real price. Note that both methods are nested in the multi-factor (naif: $\beta_{i,t} = 0, i = 0, 1, \ldots, p$; proxy: $\beta_{k,t} = 1$, where $\beta_{k,t}$ is the coefficient referring to the market index return, and $\beta_{i,t} = 0, i \neq k$). A further comparison includes the multi-factor estimated conventionally, that is, only with regular (daily) returns. In a companion paper, Souza and Veiga (2001) implemented an E-M algorithm (Dempster, Laird and Rubin, 1977) with principal component analysis to compete against the multi-factor with irregular returns on the same database. Their conclusion is that the multi-factor outperforms the E-M for the equities with less than 95% of data available.

Figure 2: Schematisation of the experiment.

To find the best multi-factor configuration, we test three values of the EWMA smoothing constant, $\lambda = 0.98, 0.99$ and 1. These high values of $\lambda$ are justified by the size of the estimation window (252 business days, approximately one year of data), which in turn is justified by the inclusion of many low liquidity equities in the comparison$^6$.

$^6$ Some require one year to have, say, 6 or 10 days in which they are negotiated.
The factors used to explain the equities returns were the Brazilian Market Index (IBV), based on the most liquid equities negotiated in the São Paulo Stock Market; the Brazilian standard overnight interest rate (CDI); and the US Dollar/Brazilian Real exchange rate (US$). Different combinations of these three factors were compared against each other. The comparison is done by means of the following statistics:

**RMSSE:** root mean square of standardized errors

\[ RMSSE(\tau) = \sqrt{\frac{1}{N} \sum_{t=1}^{N} \left( \frac{\hat{R}_{t,\tau} - R_{t,\tau}}{\hat{\sigma}_{R_{t,\tau}}} \right)^2} \]  

where \( N \) is the total number of cases of interest over time and equities and \( \tau \) is the time horizon, \( \tau = 1, 2, 3, \ldots \). In order to keep the comparison fair, \( \hat{\sigma}_{R_{t,\tau}} \) is estimated by EWMA and is the same for all methods. The standardization by a volatility estimate is because some equities are more volatile than others, and so their prediction errors can be compared without some equities dominating the statistic in spite of others.

**DC:** direction of change statistic

\[ DC(\tau) = \frac{1}{N} \sum_{t=1}^{N} \text{sgn}(\hat{R}_{t,\tau}) \times \text{sgn}(R_{t,\tau}) \times \frac{1}{1}, \quad \text{if } a > 0 \]
\[ \quad 0, \quad \text{if } a = 0 \]
\[ \quad -1, \quad \text{if } a < 0 \]  

where \( \text{sgn}(a) = \begin{cases} 
1, & \text{if } a > 0 \\
0, & \text{if } a = 0 \\
-1, & \text{if } a < 0 
\end{cases} \)

This statistic is related to the proportion of times the predicted return has the same sign than the actual return (predict the equity price will rise and it indeed rises or predict it will fall and it falls). It measures the difference between the number of times the method predicts the direction of change correctly and the number of times the direction of change is predicted incorrectly, relatively to the total number of cases.

**MAPE:** mean absolute percentage error

\[ MAPE(\tau) = \frac{1}{N} \sum_{t=1}^{N} \left| \frac{\hat{P}_{t,\tau} - P_{t,\tau}}{P_{t,\tau}} \right| \]  

While the RMSSE measures the error in the return prediction, the MAPE measures the percentage error in imputed prices. The former has a more statistical approach while the latter a more financial approach.

Tables 1 and 2 (one day ahead) and Figures 3-5 (three days ahead) show the results for the naive, the proxy and the multi-factor with irregular returns using the
following factors: no factor – only the constant (of), the market index return (IBV), IBV and the one day standard interest rate (IBV & CDI), and the IBV and the return of the US$/Brazilian Real exchange rate (IBV & US$). All the multi-factor results refer to $\lambda = 0.99$. The results referring to the remaining values of $\lambda$, 0.98 and 1, are not shown as they are in general worse than those with $\lambda = 0.99$. The results are grouped by the adjusted $R^2$ (of IBV) as it was the feature that most explained the difference between methods.

Table 1: RMSSE one day ahead by adjusted $R^2$.

<table>
<thead>
<tr>
<th>adj. R²</th>
<th>naïf</th>
<th>proxy</th>
<th>of</th>
<th>IBV</th>
<th>IBV &amp; CDI</th>
<th>IBV &amp; US$</th>
<th># cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>-20% - 0</td>
<td>1.03</td>
<td>1.12</td>
<td>1.04</td>
<td>1.03</td>
<td>1.05</td>
<td>1.05</td>
<td>6339</td>
</tr>
<tr>
<td>0 - 5%</td>
<td>0.96</td>
<td>1.01</td>
<td>0.97</td>
<td>0.95</td>
<td>0.96</td>
<td>0.96</td>
<td>23839</td>
</tr>
<tr>
<td>5 - 15%</td>
<td>0.94</td>
<td>0.96</td>
<td>0.94</td>
<td>0.90</td>
<td>0.91</td>
<td>0.91</td>
<td>31151</td>
</tr>
<tr>
<td>15 - 30%</td>
<td>0.93</td>
<td>0.88</td>
<td>0.93</td>
<td>0.84</td>
<td>0.84</td>
<td>0.85</td>
<td>24975</td>
</tr>
<tr>
<td>30 - 45%</td>
<td>0.87</td>
<td>0.77</td>
<td>0.88</td>
<td>0.73</td>
<td>0.74</td>
<td>0.74</td>
<td>13053</td>
</tr>
<tr>
<td>45 - 55%</td>
<td>0.89</td>
<td>0.71</td>
<td>0.90</td>
<td>0.67</td>
<td>0.68</td>
<td>0.71</td>
<td>5339</td>
</tr>
<tr>
<td>55 - 65%</td>
<td>0.81</td>
<td>0.63</td>
<td>0.82</td>
<td>0.61</td>
<td>0.62</td>
<td>0.63</td>
<td>3332</td>
</tr>
<tr>
<td>65 - 75%</td>
<td>0.79</td>
<td>0.55</td>
<td>0.80</td>
<td>0.54</td>
<td>0.55</td>
<td>0.55</td>
<td>1679</td>
</tr>
<tr>
<td>75 - 85%</td>
<td>0.77</td>
<td>0.47</td>
<td>0.79</td>
<td>0.46</td>
<td>0.47</td>
<td>0.48</td>
<td>1070</td>
</tr>
<tr>
<td>85 - 90%</td>
<td>0.83</td>
<td>0.42</td>
<td>0.86</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>440</td>
</tr>
<tr>
<td>90 - 95%</td>
<td>0.77</td>
<td>0.30</td>
<td>0.77</td>
<td>0.29</td>
<td>0.29</td>
<td>0.29</td>
<td>343</td>
</tr>
<tr>
<td>95 - 100%</td>
<td>0.73</td>
<td>0.19</td>
<td>0.77</td>
<td>0.20</td>
<td>0.21</td>
<td>0.20</td>
<td>141</td>
</tr>
<tr>
<td>all</td>
<td>0.88</td>
<td>0.85</td>
<td>0.90</td>
<td>0.81</td>
<td>0.83</td>
<td>0.84</td>
<td>111701</td>
</tr>
</tbody>
</table>

Table 2: MAPE one day ahead by adjusted $R^2$.

<table>
<thead>
<tr>
<th>adj. R²</th>
<th>naïf</th>
<th>proxy</th>
<th>of</th>
<th>IBV</th>
<th>IBV &amp; CDI</th>
<th>IBV &amp; US$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-20% - 0</td>
<td>0.048</td>
<td>0.054</td>
<td>0.050</td>
<td>0.051</td>
<td>0.054</td>
<td>0.052</td>
</tr>
<tr>
<td>0 - 5%</td>
<td>0.038</td>
<td>0.042</td>
<td>0.039</td>
<td>0.040</td>
<td>0.041</td>
<td>0.041</td>
</tr>
<tr>
<td>5 - 15%</td>
<td>0.030</td>
<td>0.032</td>
<td>0.031</td>
<td>0.030</td>
<td>0.031</td>
<td>0.031</td>
</tr>
<tr>
<td>15 - 30%</td>
<td>0.028</td>
<td>0.027</td>
<td>0.028</td>
<td>0.026</td>
<td>0.026</td>
<td>0.027</td>
</tr>
<tr>
<td>30 - 45%</td>
<td>0.028</td>
<td>0.025</td>
<td>0.028</td>
<td>0.024</td>
<td>0.024</td>
<td>0.025</td>
</tr>
<tr>
<td>45 - 55%</td>
<td>0.029</td>
<td>0.024</td>
<td>0.030</td>
<td>0.023</td>
<td>0.024</td>
<td>0.024</td>
</tr>
<tr>
<td>55 - 65%</td>
<td>0.029</td>
<td>0.023</td>
<td>0.030</td>
<td>0.023</td>
<td>0.023</td>
<td>0.024</td>
</tr>
<tr>
<td>65 - 75%</td>
<td>0.031</td>
<td>0.022</td>
<td>0.032</td>
<td>0.023</td>
<td>0.023</td>
<td>0.023</td>
</tr>
<tr>
<td>75 - 85%</td>
<td>0.037</td>
<td>0.021</td>
<td>0.038</td>
<td>0.022</td>
<td>0.022</td>
<td>0.023</td>
</tr>
<tr>
<td>85 - 90%</td>
<td>0.033</td>
<td>0.018</td>
<td>0.034</td>
<td>0.017</td>
<td>0.017</td>
<td>0.017</td>
</tr>
<tr>
<td>90 - 95%</td>
<td>0.028</td>
<td>0.011</td>
<td>0.029</td>
<td>0.011</td>
<td>0.012</td>
<td>0.012</td>
</tr>
<tr>
<td>95 - 100%</td>
<td>0.040</td>
<td>0.010</td>
<td>0.041</td>
<td>0.011</td>
<td>0.011</td>
<td>0.011</td>
</tr>
<tr>
<td>all</td>
<td>0.0320</td>
<td>0.0326</td>
<td>0.0328</td>
<td>0.0308</td>
<td>0.0315</td>
<td>0.0317</td>
</tr>
</tbody>
</table>
**Figure 3:** RMSSE three days ahead by adjusted $R^2$.

![Graph showing RMSSE three days ahead by adjusted $R^2$.](image)

**Figure 4:** MAPE three days ahead by adjusted $R^2$.

![Graph showing MAPE three days ahead by adjusted $R^2$.](image)

**Figure 5:** DC three days ahead by adjusted $R^2$.

![Graph showing DC three days ahead by adjusted $R^2$.](image)
The RMSSE and the MAPE show that overall the multi-factor with irregular returns using only the market index returns performs best. This means that in general the market index is significant to explain equity returns, whereas the US$/Brazilian Real exchange rate and the Brazilian inter-bank overnight interest rate are not so. However, the reader must keep in mind that 50 of these equities form the index, so that the experiment is biased in favour of this index (each of the two assets with most weight make up around 10% of the index, while the following five make up between 4% and 6% each\(^7\)). On the other hand, the majority of these equities have a negligible weight in the index.

The naïf outperforms the multi-factor by slight margin where the adjusted \(R^2\) is low, and so does the proxy where the \(R^2\) is high. However, as the \(R^2\) begins to rise the naïf tends to be outperformed by all others, and furthermore the proxy is the worst method for low \(R^2\) (which is within the expected since the \(R^2\) is related to the explication coefficient for the IBV only, and the proxy uses the IBV with \(\beta = 1\)). The Direction of Change statistic points to the proxy (the market index return) as the best indicator for the rise or fall of each equity price three days ahead. The results for one, two (not shown) and three days ahead are qualitatively similar.

### 3.1.1 - Irregular Versus Regular Returns

In this Section, the irregular returns multi-factor model is compared with the conventional multi-factor with daily returns. The simple fact that the conventional multi-factor needs two subsequent days of trade to provide a daily return compares favourably to the multi-factor with irregular returns. It is because the regular returns multi-factor discards the information of single prices (with no negotiation of the equity in the previous or in the next day) and part of the information brought by a price in the end of a block of prices, whilst the irregular returns version uses all prices. For this reason, there can be equities whose coefficients \(\beta\) can be estimated by the irregular returns version but not conventionally. There are cases where no daily return can be computed in one year, but a

\(^7\) Source: Algorithmics Brazil (http://www.algorithmics.com.br/servicos/risco/AtivosIbovespa.asp).
number of irregular ones are available\(^8\). As long as there are degrees of freedom enough to reasonably estimate the coefficients \(\beta\), the irregular returns multi-factor can be used, and in the present paper we considered 10 returns (less than 4% of data available, considering a 252 days estimation window) as a lower bound to estimate the regression. Table 3 compares the number of cases in the database for which there were enough data so that each version could estimate the coefficients \(\beta\). Note that the regular returns enable the estimation only one third of the times the irregular returns do when there are at most 5% of data (prices) available, and approximately half of the times when there are between 5% and 15% of data available. From 65% of data available on both enable the same number of estimations in the database. The comparison reported below takes into account only the cases where both versions were able to estimate the coefficients. Since we stipulated 10 returns as a minimum to estimate the regression, the case where between 0% and 5% of the prices are available is restricted to 11 or 12 prices existing in the 252 days estimation window. In view of this, 33 cases out of 100 where 11 or 12 prices were available and 10 daily returns could be computed seem too many, as one would expect existing prices scattered randomly over the estimation window. However, there are many cases in the database where a liquid equity started being negotiated during the period under study, having less than 13 prices in one year past because it was less than 13 days old then.

Table 3: frequency of estimation windows with more than 10 returns available (with an existing price in the end), for regular and irregular returns multi-factor. The cases are grouped by percentage of existing prices within the window.

<table>
<thead>
<tr>
<th></th>
<th>0 - 5</th>
<th>5 - 15</th>
<th>15 - 30</th>
<th>30 - 45</th>
<th>45 - 55</th>
<th>55 - 65</th>
<th>65 - 75</th>
<th>75 - 85</th>
<th>85 - 95</th>
<th>90 - 95</th>
<th>95 - 100</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>regular</td>
<td>33</td>
<td>824</td>
<td>3628</td>
<td>6073</td>
<td>5027</td>
<td>5976</td>
<td>7049</td>
<td>10260</td>
<td>7262</td>
<td>9131</td>
<td>24248</td>
<td>30822</td>
</tr>
<tr>
<td>irregular</td>
<td>100</td>
<td>1446</td>
<td>4006</td>
<td>6278</td>
<td>5976</td>
<td>6003</td>
<td>7049</td>
<td>10260</td>
<td>7262</td>
<td>9131</td>
<td>24248</td>
<td>30822</td>
</tr>
</tbody>
</table>

Figures 6 - 9 show the RMSSE and the MAPE for filled in data 1 and 3 days ahead. Note that the results are now grouped by percentage of existing prices in the estimation window, unlike the comparison with the naïf and the proxy. The irregular

\(^8\) In a degenerate case, an equity can be negotiated each second day, having thus availability of 50% of the daily prices but unavailability of daily returns. Even if one changes the frequency of observation, this would not be useful in the majority of cases, where the time between days in which the equity is traded is variable.
returns version yields almost always better results. The greatest difference appears when there are between 5% and 55% of prices available for imputation 1 day ahead and between 0% and 75% of prices available for imputation 3 days ahead. When the horizon increases to 3 days ahead, the advantage of the irregular returns version over the conventional one becomes more apparent. When all the prices are available, both versions are equal by definition and consequently their results are so. This also means that their results approximate more the more data exist in the estimation window.

Figure 6: RMSSE of filled in data from regular and irregular returns multi-factor, 1 day ahead, grouped by percentage of existing prices in the estimation window.

It is clear by these results that the multi-factor with irregular returns outperforms the conventional multi-factor, in addition to enable the estimation in cases where the conventional cannot be applied. As pointed out before, as the proportion of existing data approaches the unity, both estimation methods (with only regular and with irregular returns) are more similar and hence have more similar results.

---

9 Between 0% and 5% of prices available and 1 day ahead, the regular returns performed better, but note that this result is based on only 33 cases and that there were other 67 cases where only the irregular returns version could be used that were left out of the comparison.
Figure 7: MAPE of filled in data from regular and irregular returns multi-factor, 1 day ahead, grouped by percentage of existing prices in the estimation window.

Figure 8: RMSSE of filled in data from regular and irregular returns multi-factor, 3 days ahead, grouped by percentage of existing prices in the estimation window.

Figure 9: MAPE of filled in data from regular and irregular returns multi-factor, 3 days ahead, grouped by percentage of existing prices in the estimation window.
3.2 – IF estimation

Clearly the irregular multi-factor outperforms the naïf in filling in missing values, especially for higher values of the adjusted $R^2$. The proxy is slightly outperformed, especially for lower values of that statistic. From the previous results, we chose the best configuration to be the one that uses only the IBV as a factor. The 90% two-sided IF estimates (obtained via equation (18)) from this configuration are tested in this Section, together with a benchmark. The benchmark is the use of plain EWMA ($\lambda = 0.99$) to estimate the IF of the equity price, under the normal assumption. The IF estimates from equation (18), however, need estimates for the risk factors volatilities. A volatility proxy using the highest and the lowest prices (assuming the log-price is a Brownian Motion) is tried (denoted by “irreg MF u-d” in the Tables 4 and 5), as well as the squared returns (denoted by “irreg MF”). Both were computed using EWMA weights and $\lambda = 0.90$ (other values were tried and yielded no better results). The volatility proxy using the highest and lowest prices is taken from Parkinson (1976) and is given by:

$$\sigma^2 = \frac{(u - d)^2}{4 \ln 2},$$

where $u$ and $d$ are the high and low prices normalized by the closing price of the previous day. A number of volatility proxies using the highest, the lowest, the opening and the
closing prices are found in Garman and Klass (1980). All volatility proxies in the experiment use only data up to t-1 to estimate the volatility at time t. We also tried the semi-variance but the results were no better and are not shown.

The percentage of cases the price fell below (above) the 90% central IF is shown in Tables 4 and 5, grouped by adjusted $R^2$. The starred numbers are significantly ($\alpha = 0.05$) different from the nominal percentage of 5%. The significance was obtained using approximate 95% confidence intervals based on the DeMoivre-Laplace Central Limit Theorem. More powerful tools to check coverage of IF's are given in Christoffersen (1998), but we felt them unnecessary for the present analysis.

**Table 4:** Percentage of cases below or above the 90% one day ahead central IF grouped by adjusted $R^2$.

<table>
<thead>
<tr>
<th>Adj. R2</th>
<th>Below 90% IF</th>
<th>Above 90% IF</th>
<th># cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>irreg MF</td>
<td>irreg MF u-d</td>
<td>EWMA</td>
</tr>
<tr>
<td>-20% - 0%</td>
<td>0.042*</td>
<td>0.044*</td>
<td>0.065*</td>
</tr>
<tr>
<td>0 - 5%</td>
<td>0.037*</td>
<td>0.039*</td>
<td>0.058*</td>
</tr>
<tr>
<td>5 - 15%</td>
<td>0.033*</td>
<td>0.037*</td>
<td>0.056*</td>
</tr>
<tr>
<td>15 - 30%</td>
<td>0.034*</td>
<td>0.033*</td>
<td>0.055*</td>
</tr>
<tr>
<td>30 - 45%</td>
<td>0.033*</td>
<td>0.030*</td>
<td>0.047</td>
</tr>
<tr>
<td>45 - 55%</td>
<td>0.037*</td>
<td>0.034*</td>
<td>0.051</td>
</tr>
<tr>
<td>55 - 65%</td>
<td>0.030*</td>
<td>0.023*</td>
<td>0.044</td>
</tr>
<tr>
<td>65 - 75%</td>
<td>0.034*</td>
<td>0.026*</td>
<td>0.040</td>
</tr>
<tr>
<td>75 - 85%</td>
<td>0.038</td>
<td>0.025*</td>
<td>0.037</td>
</tr>
<tr>
<td>85 - 90%</td>
<td>0.051*</td>
<td>0.033*</td>
<td>0.054</td>
</tr>
<tr>
<td>90 - 95%</td>
<td>0.027</td>
<td>0.013*</td>
<td>0.040</td>
</tr>
<tr>
<td>95 - 100%</td>
<td>0.035</td>
<td>0.026</td>
<td>0.044</td>
</tr>
</tbody>
</table>

**Table 5:** Percentage of cases below or above the 90% three days ahead central IF grouped by adjusted $R^2$.

<table>
<thead>
<tr>
<th>Adj. R2</th>
<th>Below 90% IF</th>
<th>Above 90% IF</th>
<th># cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>irreg MF</td>
<td>irreg MF u-d</td>
<td>EWMA</td>
</tr>
<tr>
<td>-20% - 0%</td>
<td>0.016*</td>
<td>0.016*</td>
<td>0.017*</td>
</tr>
<tr>
<td>0 - 5%</td>
<td>0.017*</td>
<td>0.018*</td>
<td>0.046*</td>
</tr>
<tr>
<td>5 - 15%</td>
<td>0.021*</td>
<td>0.023*</td>
<td>0.053*</td>
</tr>
<tr>
<td>15 - 30%</td>
<td>0.034*</td>
<td>0.029*</td>
<td>0.059*</td>
</tr>
<tr>
<td>30 - 45%</td>
<td>0.035*</td>
<td>0.029*</td>
<td>0.057*</td>
</tr>
<tr>
<td>45 - 55%</td>
<td>0.049</td>
<td>0.036*</td>
<td>0.069*</td>
</tr>
<tr>
<td>55 - 65%</td>
<td>0.045</td>
<td>0.025*</td>
<td>0.052</td>
</tr>
<tr>
<td>65 - 75%</td>
<td>0.051</td>
<td>0.025*</td>
<td>0.062*</td>
</tr>
<tr>
<td>75 - 85%</td>
<td>0.066</td>
<td>0.030*</td>
<td>0.052</td>
</tr>
<tr>
<td>85 - 90%</td>
<td>0.087*</td>
<td>0.054</td>
<td>0.084*</td>
</tr>
<tr>
<td>90 - 95%</td>
<td>0.050</td>
<td>0.027</td>
<td>0.060</td>
</tr>
<tr>
<td>95 - 100%</td>
<td>0.044</td>
<td>0.009*</td>
<td>0.044</td>
</tr>
</tbody>
</table>
Tables 4 and 5 show that the multi-factor model estimated with irregular returns brings an improvement to the plain EWMA in the central IF estimation, the exception being the upper bound three days ahead. The EWMA is a benchmark used in many financial institutions and showed to be conservative on the present data. In general, the methods tend to be more conservative in the IF lower bound than in the upper bound, which means that an asymmetric predictive distribution could do better. However, using the semi-variance approach did not yield any better result.

As to the market index volatility proxy for the multi-factor, using the squared returns (irreg MF) performs better than using highs and lows (irreg MF u-d) for three days ahead, while the inverse occurs for one day ahead. As the irreg MF u-d seems to be too liberal for three days ahead, yielding an excessive number of cases above and below the IF when the adjusted $R^2$ is above 45%, we recommend using the squared returns (weighted by EWMA) to estimate the factor volatility in the irregular returns multi-factor.

4 – Conclusion

In this paper we proposed the use of returns which are not regularly spaced in time to estimate the multi-factor model for equity returns. The multi-factor model is a simple but efficient tool to explain cross-sectional covariance in equities returns. The model showed itself useful to estimate missing data as well as to provide interval forecasts for future returns. Furthermore, the use of irregular returns enables the estimation in cases where using only regular (daily) returns would not.

An empirical example with data from the 389 most liquid equities in the Brazilian Market confirmed the superiority of the multi-factor estimated with irregular returns over the traditional regular returns version, as well as two benchmark methods (the naif and mimicking the return of the market index). Moreover, the market index showed itself significant to explain equities returns whereas the US$/Brazilian Real exchange rate and the Brazilian inter-bank overnight interest rate did not.
Acknowledgements

This work was carried out in 2001 when both authors were working as consultants to Algorithmics Brazil and through Algorithmics Brazil to the Sao Paulo Exchange Market (BOVESPA) and its clearinghouse (CBLC). The second author (Souza) greatly acknowledges the financial support received afterwards by FAPERJ, which enabled to write the academic version of the paper.

References


FUNDACÃO GETULIO VARGAS
BIBLIOTECA MÁRIO HENRIQUE SIMONSEN
4º. 312023 0002 - 18/11/2002
famou p/ erante
N.Cham. P/EPGE SA S729u
Autor: Souza, Leonardo
Título: Using irregularly spaced returns to estimate multi-f

FGV - BMHS

Pat.: 312023/02