A Multi-Layer Perceptron for imputing missing values in financial panel/time series data[1]

Philip Kokic[2]

Working Paper Series
No. 5, April 2002

QANTARIS GmbH

Hostatostraße 25
D-65929 Höchst
Frankfurt, Germany
Tel.: ++49-69/3140 2311
Fax: ++49-69/3140 2323
Email: qantaris@freenet.de

[1] EUREDIT WP 5.7: Part C of deliverables D 5.7.1 and D 5.7.2.

[2] The author wishes to thank Dennis Hauser for his help in preparing the computer programs for this part of the EUREDIT project.

1 Introduction

The purpose of this note is to describe how a simple Multi-Layer Perceptron (MLP) can be used to impute missing values in financial time series.

The data examined in the Eur€dit project is confined to shares, European style call and put options on certain of these shares, and bonds (non-callable), as well as several indexes. Given that the best results out of the more sophisticated models tested in Kokic (2002) was for the R.1 cross-sectional regression model, we will use the same set of covariates as inputs for the MLP, that is the index variables. This considerably simplifies the process of fitting the MLP, because the index data is complete. Also, as for the R.1 model, all data will be pre-transformed by taking log-returns. Since MLPs are highly flexible models, we would hope that they also work quite well for predicting the prices of options (recall that the EM algorithm for the R.1 model did not converge in this case).

In the following section we present the MLP model and then describe the methodology that was used to estimate the parameters in the model. Unlike the traditional approach used in most Neural Network software and texts, see for example Schere (1997) or Rehkugler and Zimmermann (1994), we present a very statistical approach which has the advantage of clearly highlighting the modelling assumptions being made, as well as leading to some worthwhile improvements in implementation and notation. In subsequent sections we present a heuristic algorithm for the estimation of the parameters, and finally, a similar analysis of the imputed data to those performed in Kokic (2001) and Kokic (2002) will be undertaken.

2 The MLP Model

What is distinct about the approach in this paper is that the MLP is presented as a statistical model, and the estimation of the model parameters is done by minimising a corresponding least-squares criterion. Note that the standard back-fitting algorithm often applied to estimate parameters in MLPs need not necessarily minimise this objective function. Indeed it may only find a local minimum or it may implicity use some other objective function. What is important to note here is that for the standard MLP approach it is often unclear exactly what modelling assumptions are being made, in which case one can not properly assess whether the model is valid in a statistical sense, which means that its predictive ability cannot be relied upon.

We begin with the statistical specification of the MLP model. Suppose that there are n observations in the dataset, and for the i^{th} observation, $i = 1, \ldots, n$, y_i is a univariate response (output) variable, and $x_i = (x_{i1}, \ldots, x_{ip})'$ is a $p \times 1$ vector ($p < n$) of explanatory (input) variables. We assume that y_i is related to x_i according to the nonlinear model:

$$y_i = \mathrm{f}(x_i|\theta) + \varepsilon_i, \tag{2.1}$$

where $\varepsilon_i \sim \mathrm{NID}(0, \nu^2)$ (i.e. independent normal random variables), and ν is a unknown

constant. For a single layer MLP f is defined as:

$$\mathrm{f}(x_i|\theta) = \sum_{k=1}^{K} \beta_k \tanh\left(\sum_{j=1}^{p} \gamma_{kj} \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}\right), \tag{2.2}$$

where β_k, and γ_{kj}, $k = 1, \ldots, K$, $j = 1, \ldots, p$, are the unknown parameters making up θ, and $\hat{\mu}_j$ and $\hat{\sigma}_j^2$ are the mean and variance over the j^{th} column of the $n \times p$ input data matrix $X = (x_1, \ldots, x_n)'$, respectively. Note that K is the number of nodes in the single intermediate layer of this MLP. In vector notation,

$$\mathrm{f}(x_i|\theta) = \beta' \tanh\left(\Gamma' \hat{\Sigma}_0^{-1/2}(x_i - \hat{\mu})\right), \tag{2.3}$$

where $\beta = (\beta_1, \ldots, \beta_K)'$ is a $K \times 1$ vector and $\Gamma = (\gamma_{kj})'$ is a $p \times K$ matrix of unknown constants, and $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_p)'$ is a $p \times 1$ vector and $\hat{\Sigma}_0 = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ is a $p \times p$ matrix of known constants.

To significantly reduce the problem of collinearity during estimation and consequently improve stability, in place of $\hat{\Sigma}_0$ we will use the full covariance matrix of X:

$$\hat{\Sigma} = (n-1)^{-1}(X - \hat{\mu})'(X - \hat{\mu}), \text{ where } \hat{\mu} = n^{-1}1'X. \tag{2.4}$$

That is, we replace (2.3) by

$$\mathrm{f}(z_i|\theta) = \beta' \tanh\left(\Gamma' z_i\right), \tag{2.5}$$

where $z_i = \hat{\Sigma}_i^{-1/2}(X - \hat{\mu})$ and $\theta = \{\Gamma, \beta\}$. Defining $Z = (z_1, \ldots, z_n)' = (X - \hat{\mu})\hat{\Sigma}^{-1/2}$, we may rewrite (2.5) as

$$f(Z|\theta) \equiv (f(x_1|\theta), \ldots, f(x_n|\theta))' = \tanh(Z\Gamma)\beta.$$

Hence the model (2.1) may be specified in the more convenient vectorised form:

$$y = f(Z|\theta) + \varepsilon = \tanh(Z\Gamma)\beta + \varepsilon, \tag{2.6}$$

where $y = (y_1, \ldots, y_n)'$, the n-vector $\varepsilon \sim N(0, \nu^2 I_n)$ and I_n is a $n \times n$ identity matrix.

It is straightforward to see how (2.6) can be generalised to a multi-layer perceptron. For example, for a two-layer MLP $f(Z|\theta) = \tanh(\tanh(Z\Gamma_1)\Gamma_2)\beta$, where in this case Γ_1 is $p \times K_1$, Γ_2 is $K_1 \times K_2$, β is $K_2 \times 1$, and $\theta = \{\Gamma_1, \Gamma_2, \beta\}$. However, for reasons of simplicity, we confine attention in this paper to the single-layer model (2.6).

3 Estimation of the Model Parameters and Imputation

3.1 Estimation of the parameters

We will assume throughout that there are no missing values in the input X data matrix. However, there may be missing values in the output data vector y. The intention, of course, is to use the MLP (2.6) to impute these values. Although it may be possible to use the EM-algorithm to estimate the parameters in (2.6), we adopt the simpler, although less efficient solution, of removing these observations entirely from the estimation process.

Since at (2.6) a normal assumption with constant error variance is made, it is most appropriate to use the least-squares criterion:

$$\hat{\theta} \equiv \{\hat{\Gamma}, \hat{\beta}\} = \text{argmin}_{\{\Gamma,\beta\}} (y - \tanh(Z\Gamma)\beta)' (y - \tanh(Z\Gamma)\beta) . \tag{3.1}$$

The problem of estimating the these parameters is then reduced to a minimisation problem, where a variety of standard approaches can be used. We shall not go into details in this paper, but rather assume that an appropriate procedure can be applied to the current situation. Note, however, that minimising (3.1) may not be entirely straightforward because the objective function is potentially quite complex and could have many local minima.

A short algorithm for estimation of the MLP parameters is as follows:

1. Let $\hat{\Gamma}_0$ be some starting estimate for Γ. A sensible choice is to set up a matrix of polynomial values, G, of dimensions $p \times K$, with k^{th} column equal to $[1:p]^k$. Since many of the values in this matrix are likely to be too large, we should rescale it by a suitable constant. We propose using the maximum eigenvalue of $G'G$, $\lambda_{\max}$ say. That is, set $\hat{\Gamma}_0 = \lambda_{\max}^{-1} G$.
2. Let $i = 1$
3. Let $\hat{X}_{i-1} = \tanh(Z\hat{\Gamma}_{i-1})$
4. Use the least squares estimate for the current value of β: $\hat{\beta}_i = (\hat{X}'_{i-1}\hat{X}_{i-1})^{-1}\hat{X}'_{i-1}y$.
5. Find the current estimate of Γ, $\hat{\Gamma}_i$ say, by minimising (3.1) while keeping $\beta = \hat{\beta}_i$ fixed.
6. Set $i = i + 1$ and repeat from step 3 until convergence.

3.2 Cross validation and imputation

Perhaps a more difficult problem than estimation itself is the optimal choice for the number of nodes K on the intermediate layer of the MLP. The total number of parameters in the model is $(K + 1)p$ (plus 1 for the dispersion parameter ν), which increases quickly with K. Hence one quickly moves into a situation where over-fitting can become a serious problem.

The approach we propose using here is to minimise a least-squares criteria for a pre-selected subset of y-values, denoted by s, which are not used to estimate the parameters themselves. In effect, imputation is being used in the cross-validation process. We simply predict a particular y-value using:

$$\tilde{y}_i = \text{f}(z_i|\tilde{\theta}) = \tanh(z'_i\tilde{\Gamma})\tilde{\beta}, \tag{3.2}$$

where $\tilde{\Gamma}$ and $\tilde{\beta}$ have been estimated from the observations in $\{1, \ldots, n\} \setminus s$. The cross-validation criterion is then,

$$\sum_{i \in s} (y_i - \tilde{y}_i)^2 = (y^{(s)} - \tanh(Z^{(s)}\tilde{\Gamma})\tilde{\beta})'(y^{(s)} - \tanh(Z^{(s)}\tilde{\Gamma})\tilde{\beta}), \tag{3.3}$$

where $y^{(s)} = (y_i; i \in s)'$ is the vector of y-values in s, and $Z^{(s)} = (z'_i; i \in s)'$ is the input data sub-matrix obtained by selecting the rows $i \in s$ from Z.

3.3 Algorithm for cross-validation and estimation

Given the above discussion, it is quite simple to define an algorithm for estimating the parameters in the MLP:

1. Using (2.4), pre-transform the input data X according to $Z = \hat{\Sigma}^{-1/2}(X - \hat{\mu})$.
2. Specify the cross-validation subset $s \subset U = \{1, \ldots, n\}$.
3. Set $K = 1$.
4. Using the algorithm in subsection 3.1, estimate $\{\Gamma, \beta\}$ by minimising the objective function:
$$(y^{(U\setminus s)} - \tanh(Z^{(U\setminus s)}\Gamma)\beta)'(y^{(U\setminus s)} - \tanh(Z^{(U\setminus s)}\Gamma)\beta).$$
Denote the estimated values by $\{\tilde{\Gamma}, \tilde{\beta}\}$.
5. Compute CV_K using equation (3.3).
6. Repeat from step 3 for $K = 2, 3, 4, 5, 6, 7, 8, 9, 10$[3].
7. Set $K \in \{1, \ldots, 10\}$ to the value which minimises CV_K.
8. Estimate $\{\Gamma, \beta\}$ using all observations in U via the objective function (3.1).

3.4 Application of the MLP to the time series data

Let us denote the price or index time series (with missing data) by $\{P_{ti}, t = 1, \ldots, T\}$ where t is time (days) and $i = 1, \ldots, I$ is an instrument or index label, and let $P = (P_{ti})$ be the matrix of all these values. In all cases $P_{ti} \in \mathbb{R}^+ \cup \{\cdot\}$, i.e. the values are either positive real numbers, or missing, denoted by "$\cdot$". In the EurEdit project the dimension of P is 1304×99, that is there are 1304 daily values for 99 time series.

Let W be the log-transformed data and $W = W^{(1)}, \ldots, W^{(5)}$, where $W^{(1)}$ are US shares, $W^{(2)}$ are UK shares, $W^{(3)}$ are UK bonds (Gilts), $W^{(4)}$ are UK derivatives and $W^{(5)}$ are stock indexes and exchange rates. The following algorithm specifies how we can apply the MLP from the previous section in the current situation.

Specification MLP:

1. Perform the log-return pre-transformation as described in section 5.1 of Kokic (2002) on P to obtain $W = [W^{(1)}, \ldots, W^{(4)}]$. Let C be the number of columns in W, and $c = 1$.
2. Let $y = W(:, c)$ and $X = W^{(5)}$. Note that y may contain missing values because it is constructed from the missing value dataset P.
3. Let s be the subset of y-indices where missing values occur in the first 3 years of the y-time series. For $i \in s$, replace y_i by its true value from the true-value dataset. Remove the remaining missing values from y and the corresponding rows of X. Let n be the number of observations in y, and $U = \{1, \ldots, n\}$.

[3] A different set of K-values could be used if the one proposed does not prove appropriate

4. Use the MLP algorithm described in subsection 3.3 to form estimates $\{K, \hat{\Gamma}, \hat{\beta}\}$.
5. Using these estimates, $X = W^{(5)}$ and equation (3.2), impute the missing values in $W(:, c)$.
6. Repeat from step 2 for $c = 2, \ldots, C$.
7. Perform the post-transformation described in section 5.1 of Kokic (2002) on W to obtain the imputed price time series.

4 Assessment Results

4.1 Shares and Bonds

Analysis of the shares and bonds data was performed using the financial panel/time series data from the Eur$^{\text{E}}$dit project. For a full description of this data and how missing observations were generated see the associated documentation with these data. A total of 51 daily time series covering the time period from the beginning of 1995 to the end of 1999 were used in the analysis. The MLP method was applied to the data as well as the simple last-value carried forward (LVCF), see Kokic (2001), and the non-parametric (NP100) methods, see Kokic (2002), purely for purposes of comparison.

According to the statistical definition, MLPs are univariate because they model a single response variable at a time, although one could easily derive a multivariate generalisation from (2.6), this is not the purpose of the current paper. Consequently, 51 univariate MLPs were independently fitted to the shares and bonds time series data. During each fitting process cross validation was performed, and in nearly all 51 cases it produced an optimal K value of either 2 or 3. When $K = 2$ there are a total of 26 parameters in the MLP, and in the case $K = 3$ there are 39 parameters. Such a result is interesting, and probably as one would expect, as it indicates that a more parsimonious model is completely adequate for predicting the shares and bonds prices.

Assessment was performed on the basis of two criteria, distributional accuracy and predictive accuracy as defined in Chambers (2000). Note that a fuller set of assessments will be performed in a later stage of the Eur$^{\text{E}}$dit project. *In all cases assessment was performed on the pretransformed log-return data because, on practical grounds, this is the most sensible to use.* In addition, observations where the log return of the non-missing data equals zero were excluded from the analysis, because they have already been imputed at their original source and it would bias the results in favour of the LVCF technique if they were included in the assessment. In fact, excluding these observations only had an impact on some of the distribution assessment results.

For the first assessment criterion the Wald statistic was used, see expression (14) of Chambers (2000). Specifically, this statistic and the corresponding p-value, computed on the basis of a χ^2 approximation, was determined over all imputed observations separately for each time series. The resulting set of p-values were then summarised using box plots as shown in figures 1 – 2 in the appendix. Note that in these figures small values of p close to zero indicate a significant departure from preservation of distribution. For predictive accuracy expression (19) in Chambers (2000) with $w_i = 1$ was used. This statistic can be interpreted as the average error of imputation. In effective it is a relative measure because the log-return data is a rate of change variable. Again the statistic was

computed separately for each time series and then the set of results were summarised using box plots (see figures 3 – 4).

The results indicate that the MLP and the NP100 approaches are both considerably better than the LVCF method in terms of distributional accuracy, with the MLP approach marginally better than the NP100 method. In particular, the slightly better performance of the MLP approach is maintained consistently across degree of missingness categories (figure 2). In terms of predictive accuracy there is very little to distinguish between the methods. If anything, the MLP method is slightly worse than the other approaches, while the NP100 method is slightly better.

4.2 Options

In section 7 of Kokic (2002), for imputing missing option prices, it was found that the sophisticated and complex EM algorithm (the so-called BSEM approach), produced little if any improvement over the relatively simple BSLVCF approach, where the missing implied volatilities were imputed by carrying the last observable implied volatility forward in time. It was noted therein that this could be due to the modelling assumptions implicit in the EM algorithm, i.e. either the assumption of multivariate normality or of linearity between the log-return values of the implied volatilities.

The second of these two assumptions are considerably relaxed by the MLP model (as it would indeed be by any other non-linear statistical model). Because of multivariate normality, the EM-algorithm makes a strong assumption about linearity between the dependent variables, whereas MLPs are reputably a very flexible class of non-linear models[4]. Thus if the linearity assumptions is the reason for the poorer-than-expected performance of the BSEM approach, then one would expect some improvement using an MLP.

The way the MLP was applied in this situation is as follows. As for the other BS methods, implied volatilities were computed where an option price was available, which resulted in missing implied volatilities only where there are missing option prices. The log returns of the implied volatilities were computed, and a single log-return implied volatility time series was imputed at a time using the remaining log-return implied volatility time series as explanatory (input) variables[5]. All values were transformed back to the original scale using the approach described in section 5.1 of Kokic (2002). Finally, the Black-Scholes pricing formula was used to impute the missing option prices from the imputed implied volatilities. In the remaining part of this paper we refer to this method of imputation as the BSMLP approach.

Results for the BSMLP approach, and for comparative purposes the BSEBASE, BSLVCF and BSEM approaches, are presented in figures 5 - 6 in the appendix. These results clearly show that, as predicted above, in terms of distributional accuracy the BSMLP approach is superior to the BSEM method. Out of the four methods examined here, for low and moderate degrees of missingness, distributional properties are best preserved by the BSMLP approach, but the BSLVCF is slightly better for high degrees

[4]Note that by modifying the loss function used in 3.1, the normality assumption could also be removed from the MLP model, although we do not pursue this idea here.

[5]Where a missing value occurred in one of the explanatory variables, the LVCF approach was used to input its value (strictly speaking a zero was imputed because the log-returns of the volatilities have already been taken).

of missingness. In terms of predictive accuracy, it is hard to tell any difference between the BSLVCF and BSMLP approaches. However, for high degrees of missingness the BSEM approach is superior to the other three approaches.

References

Chambers, R. (2000). Evaluation Criteria for Statistical Editing and Imputation. EUREDIT working paper, University of Southampton, Southampton, UK.

Kokic, P. (2001). Standard methods for imputing missing values in financial panel/time series data. Working paper 2, QANTARIS GmbH, Frankfurt am Main.

Kokic, P. (2002). The EM Algorithm for a Multivariate Regression Model: including its application to a non-parametric regression model and a multivariate time series model. Working paper 4, QANTARIS GmbH, Frankfurt am Main.

Rehkugler, H. and H.-G. Zimmermann (Eds.) (1994). *Neuronale Netze in der Ökonomie.* München: Verlag Franz Vahlen.

Schere, A. (1997). *Neuronale Netze.* Wiesbaden: Vieweg & Sohn.

A Figures

A.1 Wald statistics

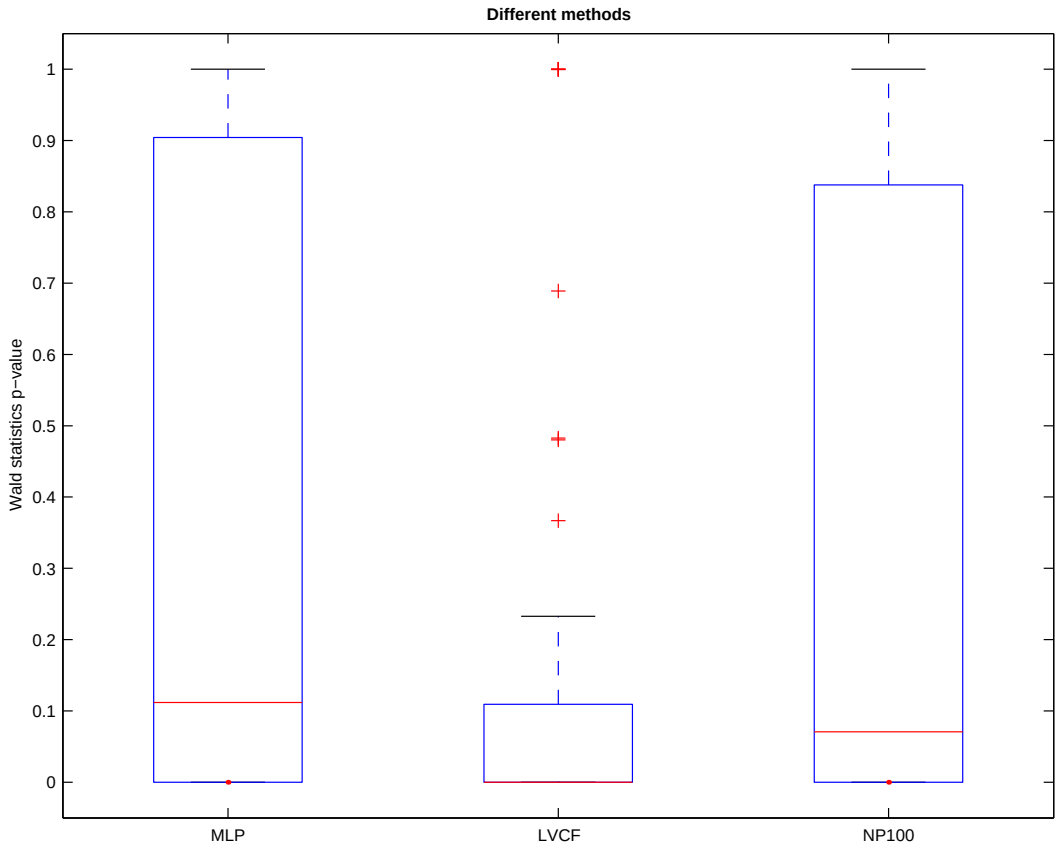

Figure 1: Distributional accuracy of the log-return imputed values

A.2 Distance statistics

A.3 Results for options

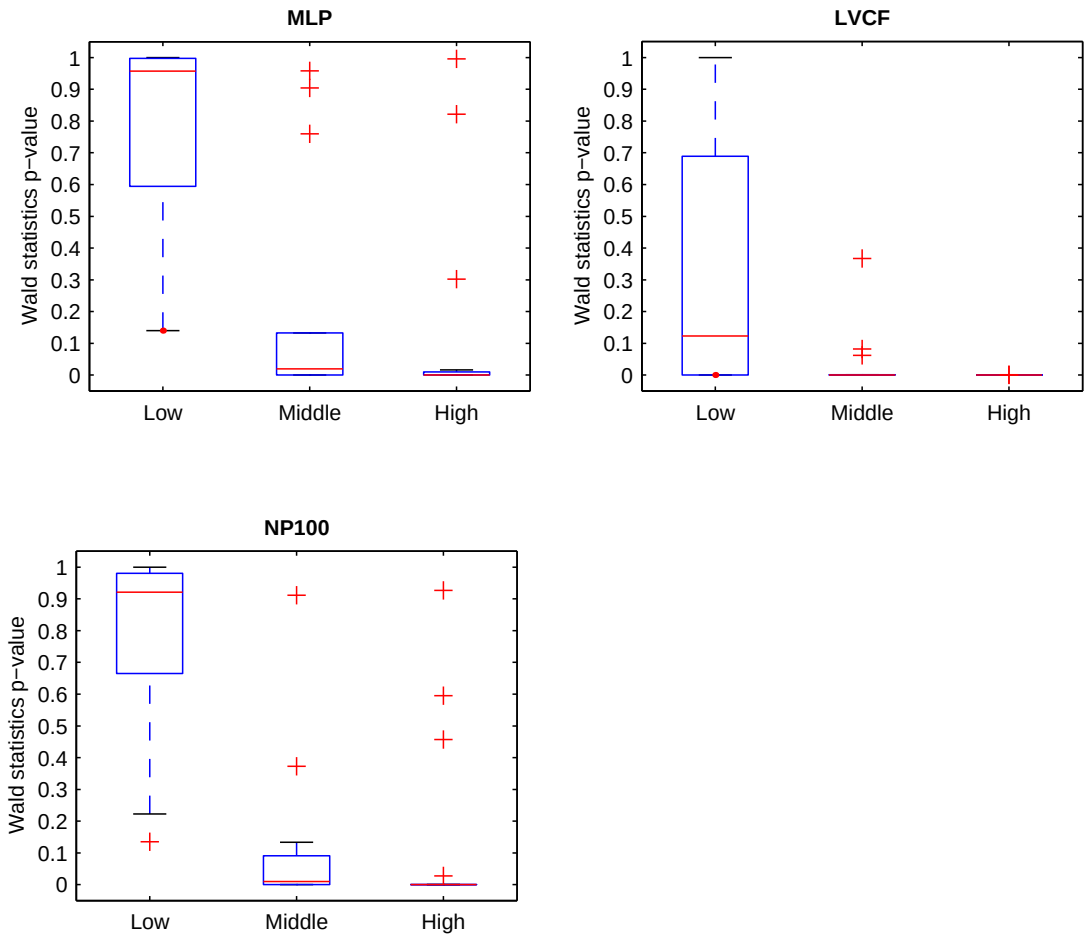

Figure 2: Distributional accuracy of the log-return imputed values by degree of missingness

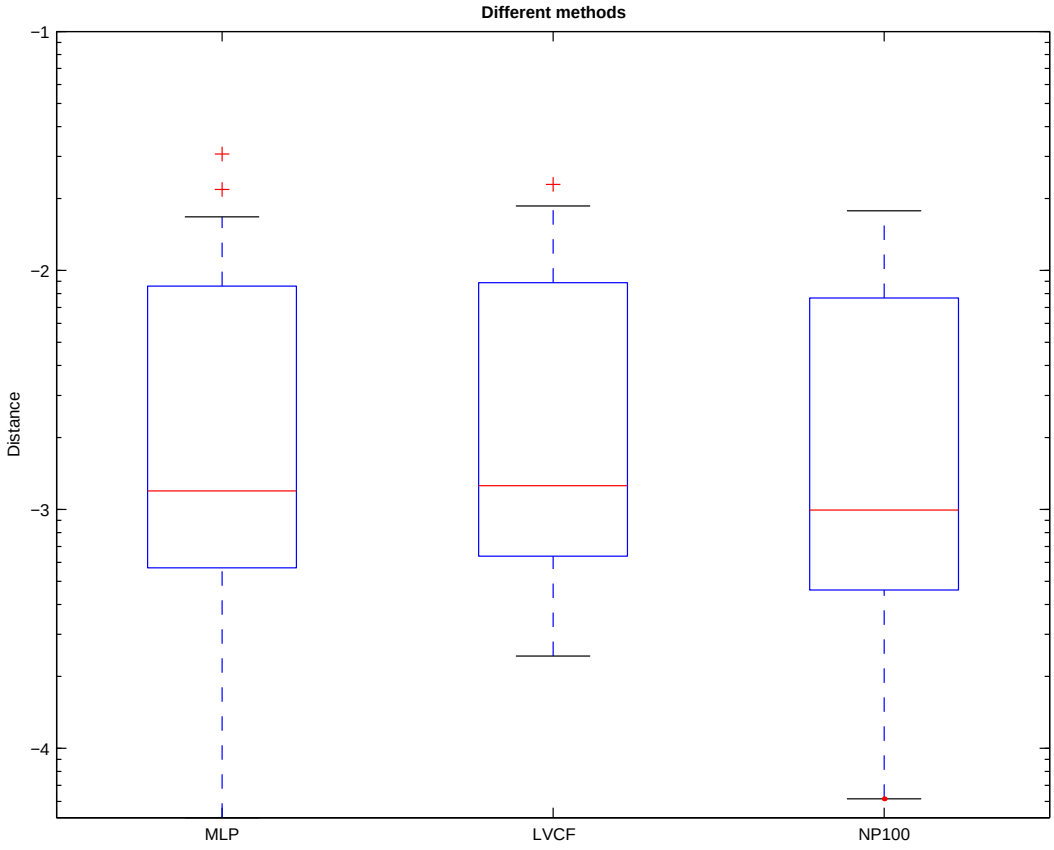

Figure 3: Relative accuracy of the log-return imputed values by method of imputation (log scale)

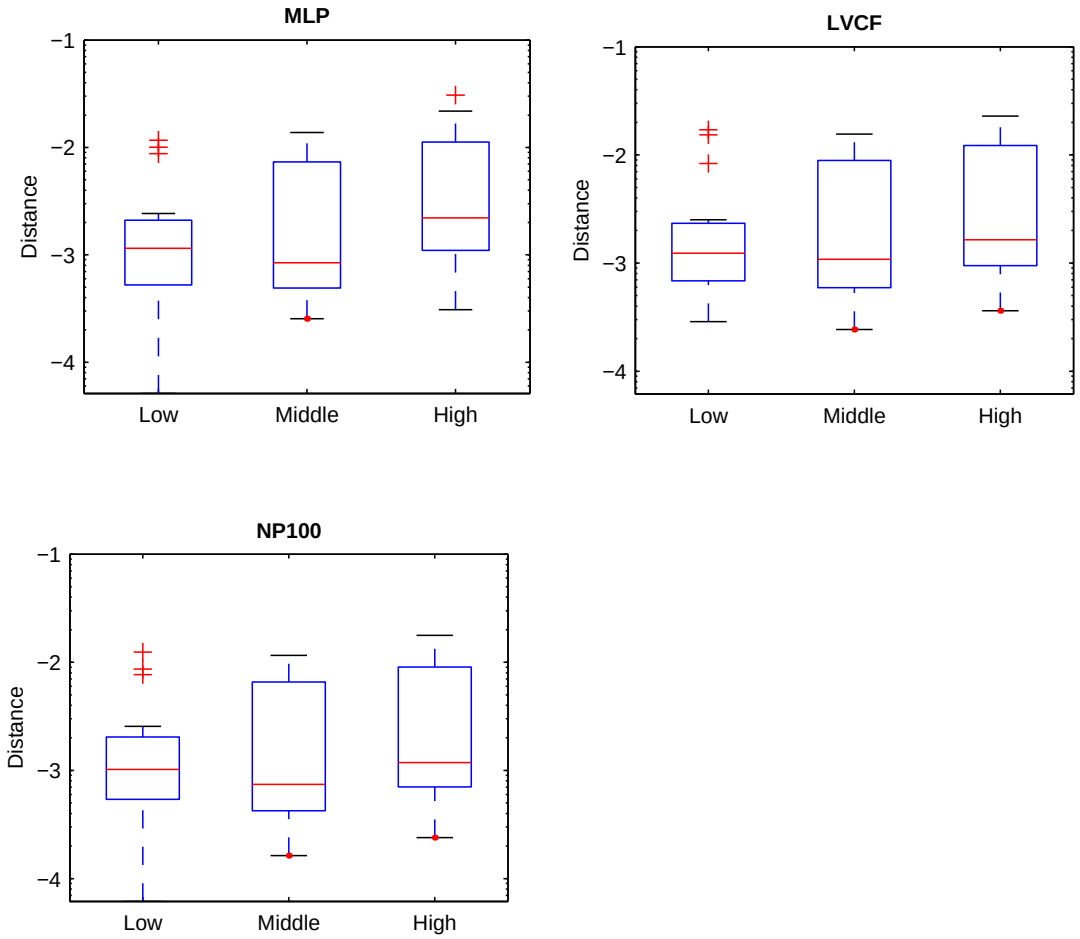

Figure 4: Relative accuracy of the log-return imputed values by degree of missingness (log scale)

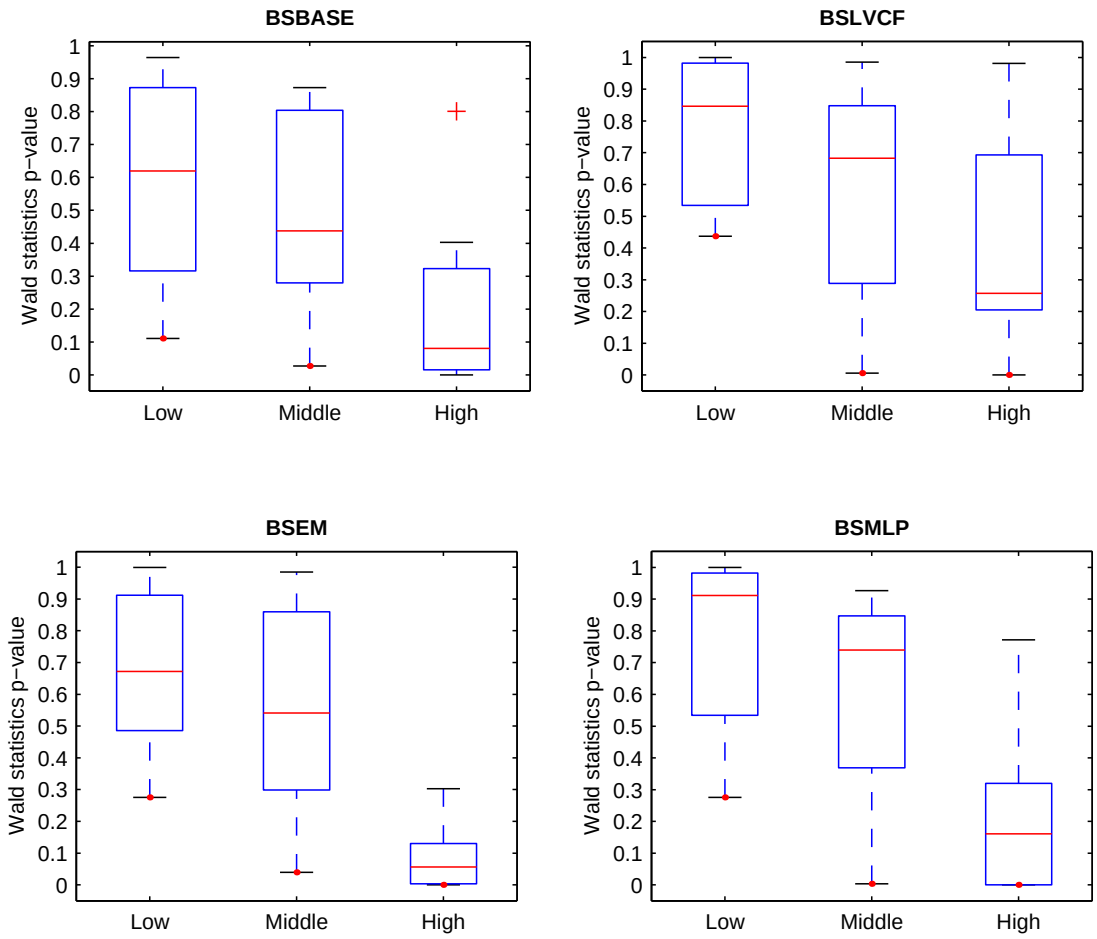

Figure 5: Distributional accuracy by degree of missingness and by method of imputation

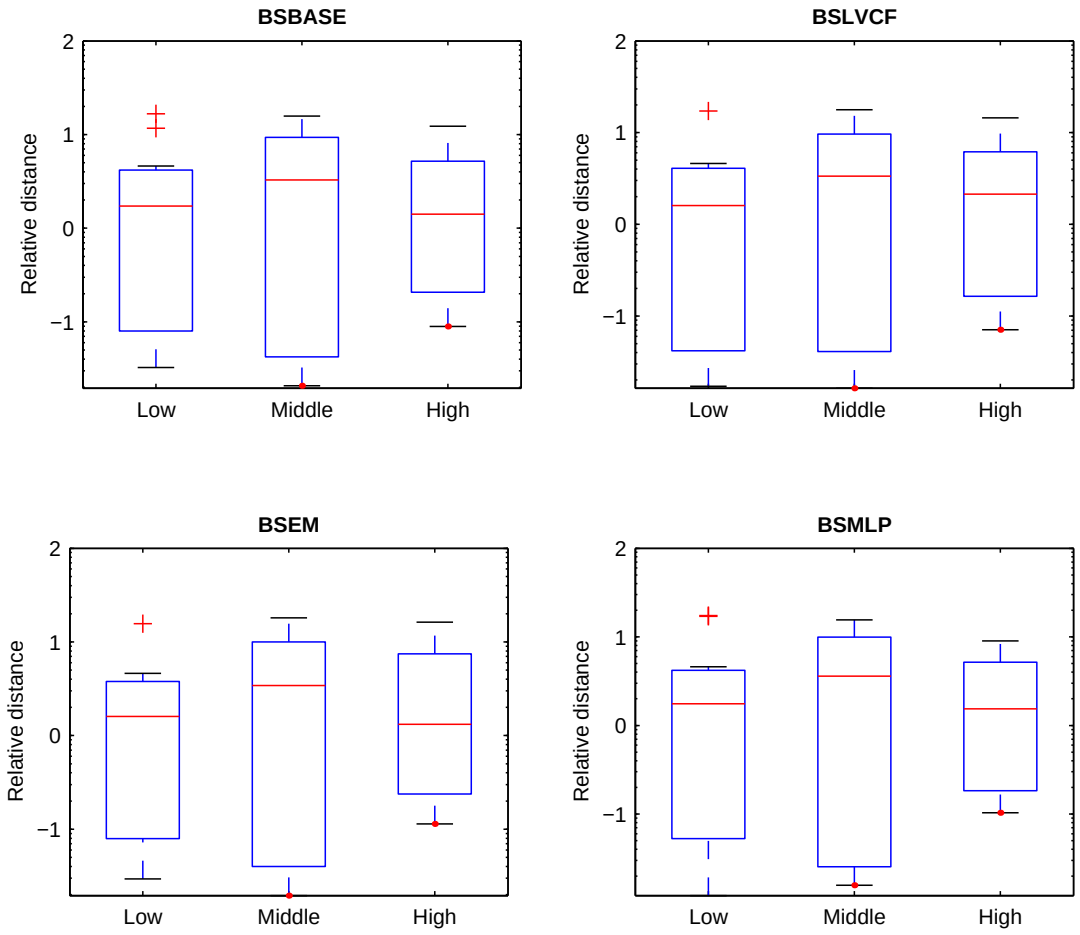

Figure 6: Relative accuracy by degree of missingness and by method of imputation (log scale)