Imputation of multivariate time series data - performance benchmarks for multiple imputation and spectral techniques

Johannes Bauer[1]*, Orazio Angelini[1], Alexander Denev[1]

Abstract

Missing data is a problem appearing ubiquitously across many fields and needs to be dealt with systematically. For multivariate time series data imputation can be a challenging problem. We consider the particular case of credit default swap time series, where missing data can pose a considerable problem preventing important value at risk estimates. We present a new framework to structure and classify missingness patterns, and generate suitable realistic test sets. We then benchmark the performance of a number of state-of-the-art imputation techniques, both stochastic multiple imputation (MI) approaches and deterministic spectral decomposition techniques. We demonstrate that for the missingness patterns under consideration, the MI package Amelia based on the expectation maximisation algorithm performs most robustly and reliably, however, other techniques like multiple singular spectral analysis can also perform well. Our results can serve as a valuable guideline for researchers and practicioners working with incomplete multivariate time series.

Keywords

Multivariate times series — Multiple Imputation — Empirical Orthogonal Functions – Matrix interpolation – Multiple Singular Spectral Analysis

[1] *IHS Markit, London, United Kingdom*

***Corresponding author**: johannes.bauer@ihsmarkit.com

Contents

Introduction

Across many different fields, ranging from finance and economics over energy and transportation, to geophysical, meteorological and sensor data, one of the challenges when working with data is that it is rarely complete. For instance, about 28% of publications in finance between 1995 and 1999 are reported to contain on average about 20% missing values [1]. As analyzed in [2], a sample of more than 100 papers in medical research between 2008 and 2013, typically contain missingness fractions exceeding 20%. The reasons for data to be incomplete are manifold and usually domain specific. Possibilities include faulty sensors, incomplete records, mistakes in data collection, unavailability to report certain information, or other very specific reasons. Often it is also not known exactly why data is missing. In most cases it is not possible to recover missing values through additional data collection

Electronic copy available at: https://ssrn.com/abstract=2996611

or measurements. Therefore when building data applications one has to accept incomplete data as the norm, and devise appropriate strategies of dealing with it.

The purpose of this publication is twofold: (i) we introduce a systematic approach to deal with missing data for multivariate time series and (ii) we benchmark a number of advanced techniques for imputation. The approach is relatively general and with minor modifications can also be applied in other domains.

As a first step in our procedure we analyse missingness patterns in the data. In principle, there can be systematic reasons that particular data points are not reported, or data can be missing without any pattern, i.e., essentially at random. Thus, as a first step we test for the missingness mechanism. Then we extract features of the missingness patterns, and perform a cluster analysis of the missingness patterns. This is very important to provide an overview over the missingness space, and it also feeds into the generation of a realistic train/validation set. This is done by superimposing the different classes of observed missingness patterns on completed data.

Once the test data has been generated, we benchmark the performance of different imputation techniques. We use state-of-the-art MI techniques based on IP [impute-posterior, like multiple imputation with chained equations (mice)] and EM (expectation maximisation, like the R package Amelia) vs state-of-the-art multivariate time series techniques [data interpolation with empirical orthogonal functions (DINEOF), multiple singular spectral analysis (MSSA)] on multivariate time series data. We will discuss advantages and disadvantages of the different methods. Depending on the application and the underlying data, one might prefer one over the other. For instance, deterministic techniques may be able to more accurately reconstruct a certain pattern in the data and hence fill values with higher accuracy. However, MI approaches preserve the statistical properties more accurately. As an example we use incomplete credit default swap (CDS) time series data with daily quotes, which show interesting missingness patterns and correlations. Our approach and benchmark results will provide valuable guidance for researchers and practitioners in the field.

Due to the ubiquitous nature of the missing data problem, one can find an extensive literature on the subject. One simple minded approach - referred to as list-wise deletion - is to omit all observations which are partially incomplete. In certain cases this can be a viable option, but more often this constitutes a very costly procedure as a lot of data is discarded. In addition this can introduce a particular bias to the remaining data [1]. In some cases it is entirely impractical, for instance, for the CDS data discussed in this article, we would loose a lot of valuable data. Therefore, list-wise deletion is nowadays usually dismissed in favour of more sophisticated techniques. More flexible and powerful strategies are ones where we predict missing data from the observed one. Generally, one can distinguish **deterministic** from **stochastic** approaches for data imputation.

A basic deterministic approach is to impute missing values for a particular feature by a simple guess, such as the mean of the observed values of this feature or the majority value (mode). This can be a successful strategy, in particular, if the missing fraction is very small. There are, however, two problems with this approach: (i) mean or mode imputation can be inaccurate, (ii) as discussed extensively in the literature [3, 4], this simple imputation technique alters the statistical properties of the data. For instance, the variance of a variable is decreased through mean imputation. A more sophisticated approach to impute missing data is to use a common machine learning technique to impute missing values from the observed ones. In this paper we also test one of these techniques called Random Forest (RF) [5] for the imputation. The approach is described in more detail in Sec. B.3.

Over the years a statistical framework has emerged, which is termed **multiple imputation (MI)**. The general idea of this framework is to deduce distribution functions from which the imputed data can be sampled. The data imputation is then non-deterministic, and multiple imputation sets can be generated. For predictive analytics on top of the completed data set, statistics for the predicted quantities can be computed, and the uncertainty about the imputation can be properly accounted for. Moreover, these imputation techniques ensure that statistical properties of the data, such as the mean and variance are not altered by the imputation.

Often, time series are dealt with by deterministic techniques, which for instance extract trend and seasonal behavior. We can split time series data into univariate and multivariate cases. Typical imputation techniques for univariate time series include linear interpolation, moving average smoothing and imputation, low pass filters, ARIMA decomposition, splines, wavelet expansion, or Kalman filters, or singular spectrum analysis (SSA). These techniques are particularly successful when the stretches of missing data are short and if the time series contains a good signal to noise ratio. Imputation for multivariate time series can in principle also be performed by these techniques, however, when correlations exists it can be beneficial to use those cross-correlations for imputations. These can be taken into account by matrix decomposition techniques such as DINEOF or its extension, MSSA. Importantly, also MI imputation techniques provide multivariate time series imputation support using lags, leads and explicit time covariates. We point out that imputations in this paper focus on working directly with the levels (values) and not on returns (first differences). Working with returns constitutes a different analysis and reconstruction of the levels may require stitching the integrated series or approaches like Brownian bridge. Preliminary analysis of this alternative approach did not suggest a strong performance, but it can be revisited in future work.

The paper is structured as follows: we first introduce our notation in Sec. 1 and define some imputation performance measures. Then in Sec. 2 we describe our approach to miss-

Electronic copy available at: https://ssrn.com/abstract=2996611

ingness characterisation and classification. This is followed by Sec. 3 with a description of MI techniques and Sec. 4 with the deterministic spectral decomposition techniques. In Sec. 5, we present a detailed discussion of the results of the performance of these techniques, followed by conclusions.

1. Notation and definitions

We use a description in terms of a standard data matrix $X_{N\times P}$ with N observations and P features. This means that x_{np} has observations along the first index (rows) and different features along the second index (columns). Since we are dealing with multivariate time series P corresponds to the number of time series components and time stamps increase along the columns. It is noteworthy that a lot of what we discuss also applies to data in different format such as heterogeneous data with P different features or image data with P pixels. All observations for a particular time series component p can be written as a column vector $\boldsymbol{x}_p$. The row vector $(x_{n1},\dots,x_{nP})$ collects all values of the components for a particular observation and we define an observation vector by $\boldsymbol{x}^{(n)} = (x_{n1},\dots,x_{nP})^T$. Explicitly, the matrix X has the following form,

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1P} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2P} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & x_{N3} & \dots & x_{NP} \end{bmatrix}.$$

A typical matrix with missing data (na) then looks like this,

$$X_o = \begin{bmatrix} x_{11} & \text{na} & x_{13} & \dots & x_{1P} \\ x_{21} & x_{22} & \text{na} & \dots & \text{na} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{na} & x_{N2} & x_{N3} & \dots & x_{NP} \end{bmatrix}.$$

It is useful to define a missing matrix M to describe the position of missing data points, for the example above it is of the following form,

$$M = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 0 \end{bmatrix}.$$

This matrix helps to keep track of the position of the missing data and can also be used to analyse all the missingness patterns appearing. For a large number of features P, we can see that one of the challenges of filling missing data is that a large number of missing data patterns can appear, and it is not clear a priori which variables to use to predict the missing ones, since the predictors might as well contain missing values. As we will discuss below for the CDS data discussed in this article we have $P = 11$.

In order to quantify the quality of data imputations, we define the following metrics:

1. **Root mean square error (RMSE):** This is an *absolute* measure, which is frequently used in the literature. Denote by S_p the set of missing observations for component p, $N_m = \sum_p |S_p|$,[1] and x_{np} the true and $\hat{x}_{np}$ the imputed value, then RMSE is

$$d_{\text{RMSE}} = \sqrt{\frac{1}{N_m}\sum_{p=1}^{P}\sum_{n\in S_p}(x_{np}-\hat{x}_{np})^2}. \tag{1}$$

2. **Mean relative deviation (MRD):** This is a *relative* measure, which can be more suitable when the values under consideration vary over different magnitudes:

$$d_{\text{MRD}} = \frac{1}{N_m}\sum_{p=1}^{P}\sum_{n\in S_p}\left|\frac{x_{np}-\hat{x}_{np}}{x_{np}}\right|. \tag{2}$$

In situations where x_{np} may take zero as a value or values close to zero, one needs to be very careful when using this metric. In the literature this quantity is sometimes referred to as *mean absolute percentage error* (MAPE). In the performance analysis in Sec. 5 we will focus on MRD.

3. **True vs predicted R squared coefficient:** R squared is a measure which appears frequently in linear regression analysis and is also often used to gauge the accuracy of data imputations. It is best to split it into separate quantities for each component p, so that values of different magnitudes do not get mixed,

$$d_{R^2,p} = 1 - \frac{\sum_{n\in S_p}(x_{np}-\hat{x}_{np})^2}{\sum_{n\in S_p}(x_{np}-\mu_p)^2}, \tag{3}$$

where μ_p is the mean for x_{np}. Then we define the weights $w_p = |S_p|/N_m$ and compute the micro-averaged R squared,

$$d_{R^2} = \sum_p w_p d_{R^2,p}. \tag{4}$$

Note that for the multiple imputation techniques all these metrics have multiple values, one for each realisation. One can analyse the mean, standard deviation, best, or worst result for each of them.

2. Missing data classification

Since missingness patterns can appear in very different forms, which can impact the imputation strategy, it is useful to initially analyse possible missing mechanisms as well as common patterns. In this section we describe a framework for how to do that.

[1] We denote by $|S|$ the number of elements of a set.

Electronic copy available at: https://ssrn.com/abstract=2996611

2.1 Definition of missingness mechanisms

In the statistical literature one usually considers the data being generated by a distribution function, $g(X|\theta)$, with unknown parameters θ. The functional form of g may or may not be known. It is then of interest to clarify how the missingness pattern M is generated and how it is related to the observed data, i.e., what general form the conditional distribution function $f(M|X,\phi)$ has; ϕ is a collection of unknown parameters. Formally, we can separate the data into observed and missing part, $X = (X^{\text{obs}}, X^{\text{miss}})$. This is meant to be understood as follows: there exists a complete data set X, but we only observe values X^{obs}. The values X^{miss} are not observed, so usually we would not know them. However, for the following reasoning it is very useful to consider their values and their relation to the missingness patterns as well. In the literature typically the following distinction is made [3]:

1. **Missing Completely at Random (MCAR):** Missingness patterns do not depend on any observed or non-observed data values,

$$f(M|X,\phi) = f(M|\phi). \tag{5}$$

2. **Missing at Random (MAR):** Missingness patterns depend on observed, but not on non-observed data values,

$$f(M|X,\phi) = f(M|X^{\text{obs}},\phi). \tag{6}$$

   One may find the term MAR confusing, since the missingness pattern M is **not** random, but rather depends on the observed values. It is, however, commonly used in the literature.

3. **Missing not at Random (MNAR):** Missingness patterns depend on both observed and non-observed data values,

$$f(M|X,\phi) = f(M|X^{\text{obs}}, X^{\text{miss}},\phi). \tag{7}$$

An example for MAR is a survey where income quotes are missing for respondents above a certain age. An example for MNAR would be that in a survey income values are more likely to be missing if these values are below a certain threshold, and age (observed) is above a certain value. In other words, respondents leave out income if they are old and earn little.

The distinction has the following consequences: MCAR and MAR belong to a class of missingness which is called **ignorable** and which makes it applicable for MI approaches. Roughly speaking the non-observed values can be integrated out in these cases [3]. In contrast, treating MNAR carefully is more difficult, since in principle we cannot predict the missing values only from the observed ones. In these situations extra data collection or additional insights from domain experts can be useful. Formally, one can then introduce suitable priors to deal with the imputation. Some of the MI packages allow for that. We will not deal with MNAR situations in this paper.

2.2 MCAR test, Little's test

Unfortunately, it is generally not possible without additional insights to determine unambiguously the missingness class for a given data set. It is, however, possible to test whether the missingness patterns of the data are consistent with the MCAR hypothesis. One way to test this explicitly is to compare the basic statistics, for instance, the mean for observed and missing data. For instance, consider the variable v_1 and compute the mean vector $\boldsymbol{\mu}_m$ of the other variables v_i, $i \neq 1$ for the cases where v_1 is missing and $\boldsymbol{\mu}_o$, i.e., the mean vector when v_1 is observed separately. If the means in these two cases differ substantially, $|\boldsymbol{\mu}_m - \boldsymbol{\mu}_o| > \varepsilon$, then we should reject the MCAR hypothesis, since the missingness pattern for v_1 correlates with the values of observed data in contrast to the MCAR assumption. This can be formulated as a statistical t-test [6]. The disadvantage with this procedure is that we have to run these t-tests separately for all appearing missingness patterns and results may become difficult to interpret.

Little [6] found a way of combining these tests into a single framework. This is now referred to as Little's test and it essentially performs such tests in a combined version. The assumption is that the data is distributed according to a multivariate normal (MVN) with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.[2] Then all the missingness patterns are identified and labelled by $m = 1, \ldots N_m$. For each pattern one can define the subset of observed values $\boldsymbol{x}_{\text{obs,m}}$ and calculate their mean vector $\bar{\boldsymbol{x}}_{\text{obs,m}}$. This can be compared with the global mean vector for these variables, i.e., the projection for $\boldsymbol{\mu}$ on these variables which is denoted by $\boldsymbol{\mu}^*_{\text{obs,m}}$. Σ_m is the corresponding covariance matrix. This is combined and analysed in the following log-likelihood form,

$$d^2 = \sum_m w_m (\bar{\boldsymbol{x}}_{\text{obs,m}} - \boldsymbol{\mu}^*_{\text{obs,m}}) \Sigma_m^{-1} (\bar{\boldsymbol{x}}_{\text{obs,m}} - \boldsymbol{\mu}^*_{\text{obs,m}})^T, \tag{8}$$

where w_m is the weight proportional to the number of cases pattern m occurs.

Little showed the following: (i) d^2 is the log-likelihood ratio for the data to be MVN distributed with global mean $\boldsymbol{\mu}$ and covariance matrix Σ and (ii) d^2 follows a χ^2-distribution with f degrees of freedom, $f = \sum_{m=1}^{N_m} p_m - P$, where p_m is the number of observed variables for pattern m. The MCAR hypothesis is equivalent to the data following an MVN distribution with global $\boldsymbol{\mu}$ and Σ and hence requires that d^2 is relatively small. MCAR is then tested as a χ^2-test. We reject the MCAR hypothesis if the computed value for d^2 is very large, since it is very unlikely that MCAR data lies deep in the tail of the χ^2-distribution (small p-value). Note that the global covariance matrix of the data, Σ, is usually not known initially, so it has to be estimated. This can be achieved with the help of the expectation-maximisation (EM) algorithm.

[2] As discussed in the paper for large samples the assumption of normality can be relaxed by application of the central limit theorem.

2.3 Missing data patterns classification and clustering

When working with missing data it is very useful to identify the most prevalent patterns and to understand whether they can be collected into similar groups (clusters).

The following procedure can be applied to find and characterise missingness patterns: As a first step we extract features of the missing data. The following numerical quantities can be useful:

1. Total fraction of missing values
2. Fraction of missing data in particular features, for instance, for the CDS data separately in short medium or long maturities
3. Statistics about the length of runs of consecutive missing values for the different features (min, max, mean, standard deviation)
4. Other data specific measures

Once the feature space is constructed, dimensionality reduction [e.g., principal component analysis (PCA)] can be performed, followed by clustering (e.g., K-means). We will present the results for the clustering for the CDS data in the next section.

2.4 CDS data and test data generation

To test the performance of imputation techniques we use Credit default swap (CDS) time series data.[3] We started with a collection of over 4000 CDS entities for different maturities and doc clauses over a period of nearly two years. In order to produce a comparable sample we narrowed the data down by focusing on tickers based in the US, which are traded in USD and possess a tier of higher seniority. This resulted in a sample of 741 tickers with 11 maturities, 6 months to 30 years (6M- 30Y). Data samples and typical missingness patterns will be shown in Sec. 5. Missing values appear quite frequently for longer maturities (15Y, 20Y, 30Y), and occasionally also for short maturities (6M, 1Y, 2Y), whereas the central maturities (5Y, 7Y) are usually observed. As discussed below, this comes from the fact that missingness is related to the liquidity, and the central maturities are the ones most commonly traded. For data characterisation, we performed standard MVN tests (Henze-Zirkler, Royston, Mardia) for a subsample for 200 tickers with very small missingness fraction. We found that the data is not consistent with the MVN hypothesis, and instead shows considerable deviations from being MVN distributed.

We ran the Little's test on the actual missingness patterns and the MCAR hypothesis was rejected in the majority of the cases with a very low p-value. This can be partly attributed to deviations from the MVN distribution initially assumed in Little's test. However, as discussed before for large enough sample size these deviations from MVN are tolerable and the test result is still meaningful. The cases were the missingness pattern is found consistent with the MCAR hypothesis usually correspond to a low missingness fraction ($< 1\%$). In such cases it can be difficult to distinguish MCAR from not MCAR. We further consulted with domain experts about the reasons for missing data. The main underlying cause quoted was liquidity, i.e., insufficient trading data to produce reliable price quotes. There was no particular evidence that the missing mechanism is MNAR and we concluded that MAR, and in some cases MCAR, is a suitable assumption for this data set.

Then we performed the feature extraction and clustering analysis as introduced above. After some exploratory work we focused on the following four features: percentage of missing data in the four longest maturities, percentage of missing data in the four shortest maturities, a standard measure for the length consecutive missing streaks for the four longest maturities and the variance of this quantity. We then used Gaussian mixture models [7] for clustering in this 4-dimensional space. The results are summarised in Fig. 1.

We show different tickers separated by dashed lines and grouped by increasing maturity (6M-30Y). The black regions indicate missing values. We identified five different clusters: (1) relatively small fraction of missing values, (2) missing values mainly for long maturities and with relatively short and alternating stretches of consecutive missing values, (3) missing values in long streaks for longer maturities, (4) patterns with considerable amount of missing data and substantial variation, (5) patterns with large amount of missing data with uniform long stretches, often covering all maturities. As we can see in the histogram the majority of the patterns (around 70%) lie within cluster 1 and 2. For the first three missing patterns we will show the imputation results in Sec. 5.

For cluster 1 about 15% of the samples were found to be consistent with the MCAR hypothesis. These usually possess a very low missingness fraction. For clusters 2 and 3 none of the samples were found to be consistent with the MCAR hypothesis.

An important question for the evaluation of the performance of data imputation techniques is how to produce a suitable training and validation set. Using data with actually missing values would be favourable in the sense that it contains realistic missingness patterns, however, it is problematic, since it does not allow us to estimate how close the imputed values are to the true values, since they are not known. For MAR ideally one would start with a complete data set and use a missingness generator of the form $f(M|X_{\mathrm{obs}}, \phi)$ to create realistic missingness patterns M. Then imputation routines can be applied to the training data and imputed and true data can be compared. However, it is generally not easy to build such a generator and moreover we usually are not in the possession of a complete data set.

We approached the problem as follows: We extracted a subset of the tickers, which have very few missing values (all 11 maturities with less then 1% missing fraction). The values which were missing were imputed by linear interpolation in order not to introduce a particular bias. This data serves as

[3] More information about the data can be found at http://www.markit.com/Product/Pricing-Data-CDS

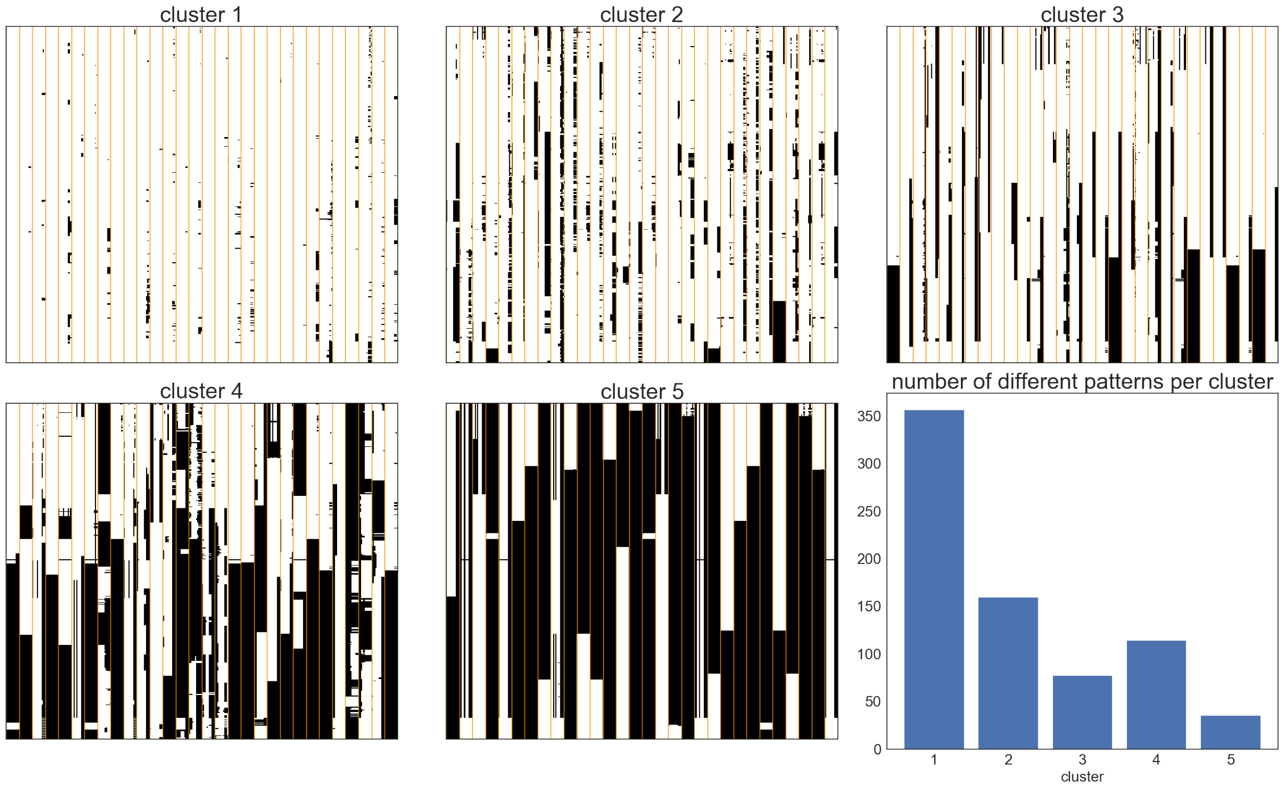

Figure 1. Clustering for CDS time series data: (1) relatively small fraction of missing values, (2) missing values mainly for long maturities and with relatively short stretches of consecutive missing values, (3) missing values in long streaks for longer maturities, (4) patterns with considerable amount of missing data and substantial variation, (5) patterns with large amount of missing data with uniform long stretches, often covering all maturities. Histogram of number of occurrences of missingness patterns for the different clusters.

ground truth for testing. With this procedure we generated 200 samples of ground truth CDS data each with 11 maturities.

As a next step we had to impose missingness pattern. A simple procedure, often found in the literature, is to randomly remove data points, however, that is problematic, since the Little test discussed above showed that the data is not consistent with MCAR. Therefore our procedure was to impose realistic missingness patterns M on this data, i.e., we remove values according to those predefined patterns. As discussed, we found five most prevalent patterns. Here we focus on the imputation of the clusters 1, 2, and 3, as clusters 4,5 contain longer stretches of completely absent data which would be better filled by a proxy. The test sets for each cluster are generated by applying the patterns on the 200 ground truth examples. In the case where the cluster does not contain enough different patterns we draw from the available patterns with repetition. A typical block for cluster 2 (ticker number 1) is shown in the bottom of Fig. 3. In this case the long maturities (10Y, 20Y, 30Y) are missing for a considerable amount of the time steps. The described combination of complete underlying data set and imposed missingness pattern leads to semi-synthetic data sets on which we can run the imputation routines, and since we have the ground truth all the performance metrics can be computed. This procedure leads to a relatively realistic missingness representation for test purposes, which can be generated with little effort.

3. Multiple imputation methods

Multiple imputation (MI) is a statistical framework for data imputation. The objective is to determine a good approximation for the distribution functions for the data $f(X)$, both observed and unobserved. This is usually achieved by an iterative mechanism. Once $f(X)$ is found, imputations can be generated by sampling from the conditional distribution functions for the various missingness patterns which occur. The conditional distributions can be derived from the general joined distributions either explicitly or they are made accessible implicitly by a Monte Carlo sampling procedure.

A particular framework is termed *multiple imputations by chained equations* (mice). Chained equations refer to an iterative procedure, by which data values and parameter values are generated in series of steps. The general assumption is that the (complete) data is generated from a multivariate distribution function, $p(X|\theta)$, where θ is a collection of parameters, which is not known. In certain cases, the distribution function p can be assumed to have a particular form. For instance, a common assumption is that the complete data is generated by a MVN, i.e., $X \sim \mathcal{N}(\boldsymbol{\mu},\Sigma)$ and $\theta = (\boldsymbol{\mu},\Sigma)$. Then all the distribution functions can be given explicitly and the procedure becomes a bit more transparent [8]. We focus on the description of this case in this section. The general description based on a Markov Chain Monte Carlo sampling approach is described in [9].

3.1 MVN case

The basic assumption of this section is that the data (both observed and missing) is described by a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, formally,

$$X \sim \mathcal{N}(\boldsymbol{\mu},\Sigma). \tag{9}$$

Then the conditional distribution functions used to impute data are also MVNs [7]. The distribution functions from which the parameters can be sampled are given as follows: As described [8] the distribution function for the MVN case for the covariance matrices has the following form,

$$p(\Sigma|\nu,\Lambda) = \mathscr{W}^{-1}(\Sigma, \nu = N-1, \Psi = \Lambda). \tag{10}$$

$\mathscr{W}^{-1}$ is the inverse Wishart function, ν is the number of degrees of freedom, Ψ is a positive definite scale matrix. The matrix Λ is specified below. The mean vector is sampled from,

$$p(\boldsymbol{\mu}|\boldsymbol{\mu}^*,\Sigma^*) = \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\mu}^*,\Sigma_0) \tag{11}$$

The matrix Σ_0 is specified below. The parameters for these distribution functions are estimated iteratively from previous imputation results.

With these specifications we can describe the algorithm in more detail. It has an explicit form and does not rely on Gibbs sampling. It has the **imputation (I)** - **posterior (P)** form.

First, based on a some initial estimate the generating distributions $p(\Sigma|\nu,\Lambda)$ and $p(\boldsymbol{\mu}|\boldsymbol{\mu}^*,\Sigma^*)$ are specified. We can draw parameters $\theta^{(1)} = (\boldsymbol{\mu}^{(1)},\Sigma^{(1)})$ from these distributions.

I step: We can impute data based on these parameters (so called **I-step**). We have to do this for all missingness patterns separately. To predict missing values for variable X_k, we have to determine the conditional distribution

$$p(X_k|X_{-k},\boldsymbol{\mu}^{(1)},\Sigma^{(1)}), \tag{12}$$

where X_{-k} denotes the collection of observed variables excluding X_k. This can be achieved in two equivalent ways: (i) We can sample from (12) to impute values for X_k, and so forth for the other variables. (ii) Instead of sampling from the conditional MVN in equation (12), we can also derive linear regression equations directly from $\boldsymbol{\mu}^{(1)}$ and $\Sigma^{(1)}$ and add a stochastic variance term. There are different versions of doing this regression. The ones which are most commonly used are (a) Bayesian linear regression (called norm in the **mice** package), and(b) **predictive mean matching** (pmm). Once all values are imputed the I-step has finished and the so-called **P-step** (for posterior in the Bayesian framework) follows.

P step: In this step new distribution functions for the parameters θ are estimated. This is usually done entirely in a Bayesian framework. Certain assumptions are made for the priors, and the likelihood and posterior functions are computed from the observed and previously imputed data. The posterior distribution for Σ is found to be the one in Eq. (10), where Λ is the sample covariance matrix of the completed data set. If we denote the drawn matrix Σ^* then the new distribution function for μ is the one in Eq. (11) with μ^* the vector of sample means using the completed data and $\Sigma_0 = \Sigma^*/N$ [8]. Once the distributions are specified new parameters θ can be obtained by sampling.

3.2 Expectation maximisation (EM) procedure

If the data is MVN distributed, instead of the I-P procedure discussed in the last section, the parameters $\theta = (\boldsymbol{\mu},\Sigma)$ can also be estimated by maximum likelihood estimation (MLE) using the expectation maximisation (EM) algorithm.

The procedure can be described as follows: Consider we have the data collected into a matrix X which can be split into observed and missing, $X = (X^{\text{miss}}, X^{\text{obs}})$. The log-likelihood can be written as,

$$l = \log p(X^{\text{obs}}|\theta) = \sum_n \log \sum_{\boldsymbol{x}^{\text{miss},(n)}} \left[p(\boldsymbol{x}^{\text{obs},(\text{n})}, \boldsymbol{x}^{\text{miss},(n)}|\theta) \right] \tag{13}$$

This is difficult to maximise directly, but constitutes a situation which can be treated with EM. The idea is to compute the parameters $\theta = \theta^{(t)}$ iteratively. We first need an initial estimate to compute θ, either by just using complete data rows, or by using a simple imputation scheme, e.g. mean imputation. Then we can compute $\theta^{(0)}$ from the MLE.

E step: Once we have some estimate for $\theta^{(t-1)}$ we can compute the expectation value

$$Q(\theta^{(t)},\theta^{(t-1)}) = E\Big[\sum_n \log \mathcal{N}(\boldsymbol{x}^{(n)}|\boldsymbol{\mu},\Sigma)\Big]\Big|(X,\theta^{(t-1)}), \tag{14}$$

where the expectation value is conditioned on $(X,\theta^{(t-1)})$. This can be simplified and reduced to compute expectation of the form $\sum_n E[\boldsymbol{x}^{(n)}]$ and $\sum_n E[\boldsymbol{x}^{(n)}[\boldsymbol{x}^{(n)}]^T]$, where we omitted the conditioning for notational simplicity. These are called **expected sufficient statistics**. In order to calculate those we need to use relations of multivariate normal conditional probability densities (see Ref. [7] p.374).

M step: In the maximisation step we compute new parameters $\theta^{(t)}$. This is done be computing appropriate derivatives of the function Q and solving for $\boldsymbol{\mu}$ and Σ, $\nabla Q = 0$. One finds,

$$\boldsymbol{\mu}^{(t)} = \frac{1}{N}\sum_n E[\boldsymbol{x}^{(n)}] \tag{15}$$

and

$$\Sigma^{(t)} = \frac{1}{N}\sum_n E[\boldsymbol{x}^{(n)}[\boldsymbol{x}^{(n)}]^T] - \boldsymbol{\mu}^{(t)}[\boldsymbol{\mu}^{(t)}]^T. \tag{16}$$

Note that this approach is quite careful to take into account the variance of the data. Once this is computed we can return to the E step and iterate.

Once $\theta = (\boldsymbol{\mu}, \Sigma)$ are estimated, missing values can be imputed by sampling from the appropriate conditional distribution. A data vector can usually be split up into missing and observed part $\boldsymbol{x} = (\boldsymbol{x}^{\text{miss}}, \boldsymbol{x}^{\text{obs}})$. Missing values $\boldsymbol{x}^{\text{miss}}$ can be predicted by sampling from the conditional distribution function,

$$p(\boldsymbol{x}^{\text{miss}} | \boldsymbol{x}^{\text{obs}}, \theta), \tag{17}$$

as in Eq. (12). To account for the uncertainty in the parameters θ a bootstrap approach can be used, which is done in the implementation of the R package Amelia [10].

4. Deterministic and EOF based techniques

As discussed in the introduction, rather than using the MI frameworks data imputation can also be achieved by deterministic techniques. One approach is to use machine learning techniques to predict missing data from the observed one. We used one popular approach based on random forest. Some details about the algorithm and the software library which was used can be found in the appendix. Other deterministic approaches are ones based on spectral decompositions and empirical orthogonal functions (EOFs). We will give a brief introduction to such techniques.

4.1 Brief recap of singular value decomposition

Consider a matrix $X_{N\times P}$. Then there exist orthonormal matrices $U_{N\times N}$, $V_{P\times P}$, such that

$$X = USV^T, \tag{18}$$

where S ($N \times P$) is a matrix with singular values $\sqrt{\lambda_i}$ on the diagonal and the rest is filled with zeros. We always consider situations where the singular values are ordered in terms of their magnitude. The matrix U can be written as a collection of column vectors,

$$U = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N], \tag{19}$$

and similar for V. They satisfy,

$$XX^T \boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i, \tag{20}$$

with $i = 1, \ldots, P$, and

$$X^T X \boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i. \tag{21}$$

We call $\boldsymbol{v}_i$ the right eigenvectors of X and $\boldsymbol{u}_i$ the left eigenvectors. Since these vectors span a suitable space related to the empirical data, they are referred to as empirical orthogonal functions (EOF). We can write the SVD decomposition explicitly as,

$$X = \sum_{k=1}^{q} \sqrt{\lambda_k} \boldsymbol{u}_k \boldsymbol{v}_k^T \tag{22}$$

where q is the number of non-zero singular values. This expression is a sum of rank 1 matrices.

4.2 Data interpolation with empirical orthogonal functions (DINEOF)

The DINEOF approach was introduced in the context of time resolved geological data [11], for instance, consider spatio-temporal field $f(t_i, \boldsymbol{r_j})$ and relate it to the data matrix X by

$$X_{ij} = f(t_i, \boldsymbol{r}_j). \tag{23}$$

The strategy to fill missing data is in the spirit of matrix completion via a decomposition of the form,

$$X = AB, \tag{24}$$

where A is $N \times K$ matrix and B a $K \times P$ matrix. K corresponds to a latent dimension carrying the essential information about the data. In the DINEOF approach this **matrix factorisation** is constructed iteratively using the EOF basis obtained via SVD. We start by imputing a first guess for the missing values, e.g., mean values, and then compute the EOFs for the completed data matrix. The reconstruction in the DINEOF is based on a subset of the EOFs $n_{\text{EOF}} < q$,

$$X^{(n_{\text{EOF}})} = \sum_{k=1}^{n_{\text{EOF}}} \sqrt{\lambda_k} \boldsymbol{u}_k \boldsymbol{v}_k^T, \tag{25}$$

where we iterate $n_{\text{EOF}} = 1, \ldots, N_{\max}$ and $N_{\max}$ is an upper boundary for the number of EOFs to be used. For a given number of EOFs, n_{EOF}, one has an inner loop to iterate for convergence for the imputed values. One typically measures the accuracy and convergence of the imputations by initially removing a small random subset of otherwise known data points and computes RMSE (true vs predicted), see Eq. (1). Convergence is assumed when RMSE does not decrease any further. A problem with this convergence assessment is that the randomly removed points might follow a quite different pattern than the data which actually needs to be imputed, and the imputation may not therefore be optimal.

This approach works quite well when there is enough structure in the data, and P must not be too small. For illustration, we give an example of DINEOF imputation for synthetic two-dimensional data field[4] in Fig. 2.

Two comparisons are shown: The first one has true data from which randomly 50% of the pixels are removed. With $n_{\text{EOF}} = 8$ EOF basis functions we can get an accurate reconstruction. In the second case additional noise is added to the data, and again a good reconstruction can be achieved. Note that the DINEOF approach aims to only include EOFs as long as they add signal. However, to achieve a clear signal/noise separation is numerically not easy.

[4] This example is reproduced from `http://menugget.blogspot.co.uk/2012/10/dineof-data-interpolating-empirical.html` with the synthetic data introduced in [11].

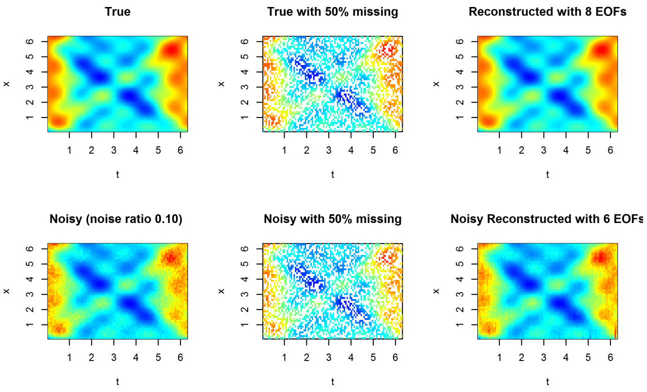

Figure 2. Example of DINEOF imputation for synthetic 2d data.

4.3 Multiple singular spectral analysis (MSSA)

Singular spectral analysis (SSA) is a more advanced decomposition technique, which is very successful for time series analysis as well as images (see Ref. [12] and references therein). We describe the technique for time series, but the extension to images is formally relatively straight forward. The basic idea is to construct objects which contain time-lagged covariances up to a certain window length L. For these objects we perform SVD and then the time series can be decomposed and reconstructed with the dominant modes and EOFs.

First consider the case of a univariate time series $X = X_t$, $P = 1$. We first describe how to do the time series decomposition for a complete data set formally: for a given window size $L \leq N$ and $K = N - L + 1$, construct the trajectory matrix T_X,

$$T_X = \begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_K \\ x_2 & x_3 & x_4 & \dots & x_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & x_{L+2} & \dots & x_N \end{bmatrix}.$$

T_X is a $L \times K$ matrix with the same time series points on the anti-diagonals. Note that for the transformation $L \to K$, $K \to L$, the corresponding trajectory matricies satisfy $T_X \to T_X^T$. The trajectory matrix can be used to compute the time lagged covariance matrix,

$$C = T_X T_X^T\,. \tag{26}$$

This is a symmetric matrix and has the explicit form,

$$C = \begin{bmatrix} \sum_{i=1}^K x_i^2 & \sum_{i=1}^K x_i x_{i+1} & \dots & \sum_{i=1}^K x_i x_{i+L-1} \\ \sum_{i=1}^K x_i x_{i+1} & \sum_{i=2}^{K+1} x_i^2 & \dots & \sum_{i=2}^{K+1} x_i x_{i+L-1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^K x_i x_{i+L-1} & \sum_{i=2}^{K+1} x_i x_{i+L-1} & \dots & \sum_{i=K}^N x_i^2 \end{bmatrix}.$$

We can see from this how time lags up to length L are considered. In other words modes with maximal period L can be identified. The time lagged trajectory matrix is just used for illustration which time correlation are picked up by SSA, the approach works directly with the trajectory matrix T_X [12].

The next step is to perform SVD on T_X, such that the following reconstruction can be given,

$$T_X = \sum_{k=1}^{q} \sqrt{\lambda_k} \boldsymbol{u}_k \boldsymbol{v}_k^T\,. \tag{27}$$

Typically, one groups eigenvalues into certain subsets $I_1, \dots, I_m$, for instance, oscillatory modes appear as paired eigenvectors with very similar singular values. The partial reconstruction is then written as

$$T_X^{\text{rec}} = \sum_{h=1}^{R} \sum_{k \in I_h} \sqrt{\lambda_k} \boldsymbol{u}_k \boldsymbol{v}_k^T\,, \tag{28}$$

for a particular choice of $\{I_h\}$. There is some subjectivity involved in this step. For instance, in time series analysis, one might want to focus on a trend and only two oscillatory modes.

The final step is to map the reconstructed trajectory matrix, back to the time series. We do this by averaging over the antidiagonals. Denote by Δ_k the set of anti-diagonal index pairs (i, j), e.g., $\Delta_1 = \{(1,1)\}$, $\Delta_2 = \{(2,1),(1,2)\}$, etc., and $|\Delta_k|$ the number of elements. Then the reconstructed time series reads, $k = 1, \dots, N$,

$$x_k^{\text{rec}} = \frac{1}{|\Delta_k|} \sum_{(i,j) \in \Delta_k} [T_X^{\text{rec}}]_{ij}\,. \tag{29}$$

The data imputation based on SSA follows the same logic as for time series decomposition, except that the EOF basis for the reconstruction and the imputed values are determined iteratively. One starts by filling the missing values by an initial guess. Then T_X is constructed and the SVD computed. A partial reconstruction with n_{EOF} EOFs,

$$T_X = \sum_{k=1}^{n_{\text{EOF}}} \sqrt{\lambda_k} \boldsymbol{u}_k \boldsymbol{v}_k^T\,, \tag{30}$$

is used to fill the missing values for the reconstructed time series. This is iterated to convergence for a fixed n_{EOF} as above in the DINEOF approach. The algorithm successively adds more EOFs until no further improvement for the imputations can be achieved or a maximal number of EOFs is reached. The improvement is typically measured by randomly removing a small set of otherwise known data points and computing RMSE (true vs predicted). The same possible problems due to different patterns missingness patterns as in the DINEOF approach apply here.

The multivariate case (MSSA) is formally very similar to the univariate case, but numerically more involved. For each time series $\{X_{n,p}\}$ a trajectory matrix T_{X_p} can be computed,

$$T_{X_p} = \begin{bmatrix} x_{1p} & x_{2p} & x_{3p} & \dots & x_{Kp} \\ x_{2p} & x_{3p} & x_{4p} & \dots & x_{K+1p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{Lp} & x_{L+1p} & x_{L+2p} & \dots & x_{Np} \end{bmatrix}.$$

These trajectory matrices are stacked together into a combined trajectory matrix,

$$T_X = [T_{X_1}, \ldots, T_{X_P}]. \tag{31}$$

This is a $L \times PK$ block-Hankel matrix. Note that the corresponding lagged covariance matrix includes cross correlations terms for the different time series. Once the trajectory matrix is defined the formalism proceeds essentially as described above in the univariate case. Data imputation based on MSSA was proposed and tested in Ref. [13].

5. Results for the CDS data

We will now discuss the performance of the different imputation techniques on the CDS time series data introduced in Sec. 2. An example of a completed ground truth series is displayed in Fig. 3. It belongs to an issuer in the consumer goods sector and possesses a “modified restructuring” doc clause.

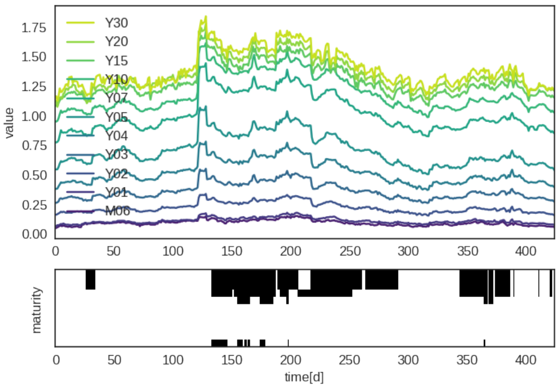

Figure 3. Top; Example of complete time series data (ticker 1, cluster 2). The lower part shows the missingness pattern which is imposed on the complete data.

We can see the daily quotes for a period of nearly two years. We observe a hierarchy of values, which are ordered by the maturity. Since the CDS price is a measure for the market view on the probability of default of a certain underlying asset within a defined time period (maturity), this is expected to be the case. There are no strong trend or seasonality patterns, but the times series are also not strictly stationary. We can see that the values for the different maturities are relatively well correlated. It is therefore intuitive that if values for some of the maturities are missing, they can be inferred from the others. A typical missingness pattern in cluster 2 class is shown in the lower part of Fig. 3. Stretches of values for longer maturities are missing in certain intervals, and also some values of shorter maturities are missing, whereas the central maturities are complete. For the test data generation introduced in Sec. 2, we can think of this as a mask, which blocks out the respective values.

Before we provide a detailed comparison of imputed values with the the true values for the different imputation techniques, we first give a full overview of the results for the imputation performance in the different clusters. For each method we did extensive initial testing to identify hyper-parameters and input data adjustments for favorable performance. We have used the following techniques (see also appendix): The multiple imputation techniques **mice**, based on chained equations and conditional sampling, and **Amelia**, which uses the EM algorithm to determine the joined data distribution function. As discussed in Sec. 3.2 Amelia uses the assumption of the data being MVN distributed, however we have seen in Sec. 2 that this is not the case. As pointed in Refs. [10, 4] and illustrated by the following results the MVN violation does not hamper imputations with good performance. For **mice** we manually included one step time leads and lags in the imputation procedure. We checked both the Bayesian linear regression option (norm) as well as predictive mean matching (pmm), but report results only for the former here, which showed better performance. In **Amelia** we used the options to include time lagged and lead data and and explicit time covariate to second order. For both MI we computed five imputations and take the average values as prediction to evaluate the performance metrics.

As deterministic routines we tested Random Forest (MissForest), DINEOF, and MSSA. We added an explicit time variable in the RF imputation, but no leads and lags. When all values were missing for a particular time step we first interpolated the four central maturities linearly. This was also done for the DINEOF approach. In DINEOF we first subtracted the mean for each time series and then added it again after the imputation. For MSSA we did not perform any prior linear interpolation. We chose the window length for the patterns in cluster 1 as 10 time steps and for ones in cluster 2,3 as 40 time steps. It is noteworthy that both EOF based techniques depend on the starting values, sometimes quite sensitively. We also tested an approach, where MSSA was initialised with Amelia results, which avoids very inaccurate starting values (Amelia+MSSA).

With these choices we computed the imputations for 200 cases of missingness patterns overlayed on the ground true values for each cluster 1, 2 and 3, enough to get reliable statistics. In this section we focus on the MRD performance measure defined in Eq. (2), a relative measure suitable for the comparison of values with different magnitude. To get a global comparative view of the performance we computed the summary statistics (mean, standard deviation, minimum, maximum) of the 200 MRD values for each pattern, imputed vs ground truth. The summary statistics for cluster 1 can be found in table 1, for cluster 2 in table 2, and for cluster 3 in table 4.

For cluster 1 the imputations are quite accurate and the MRD is usually between 1-3%, with a few exceptions. The best performance is obtained with MSSA and Amelia, but the other techniques produce comparable results. The patterns in

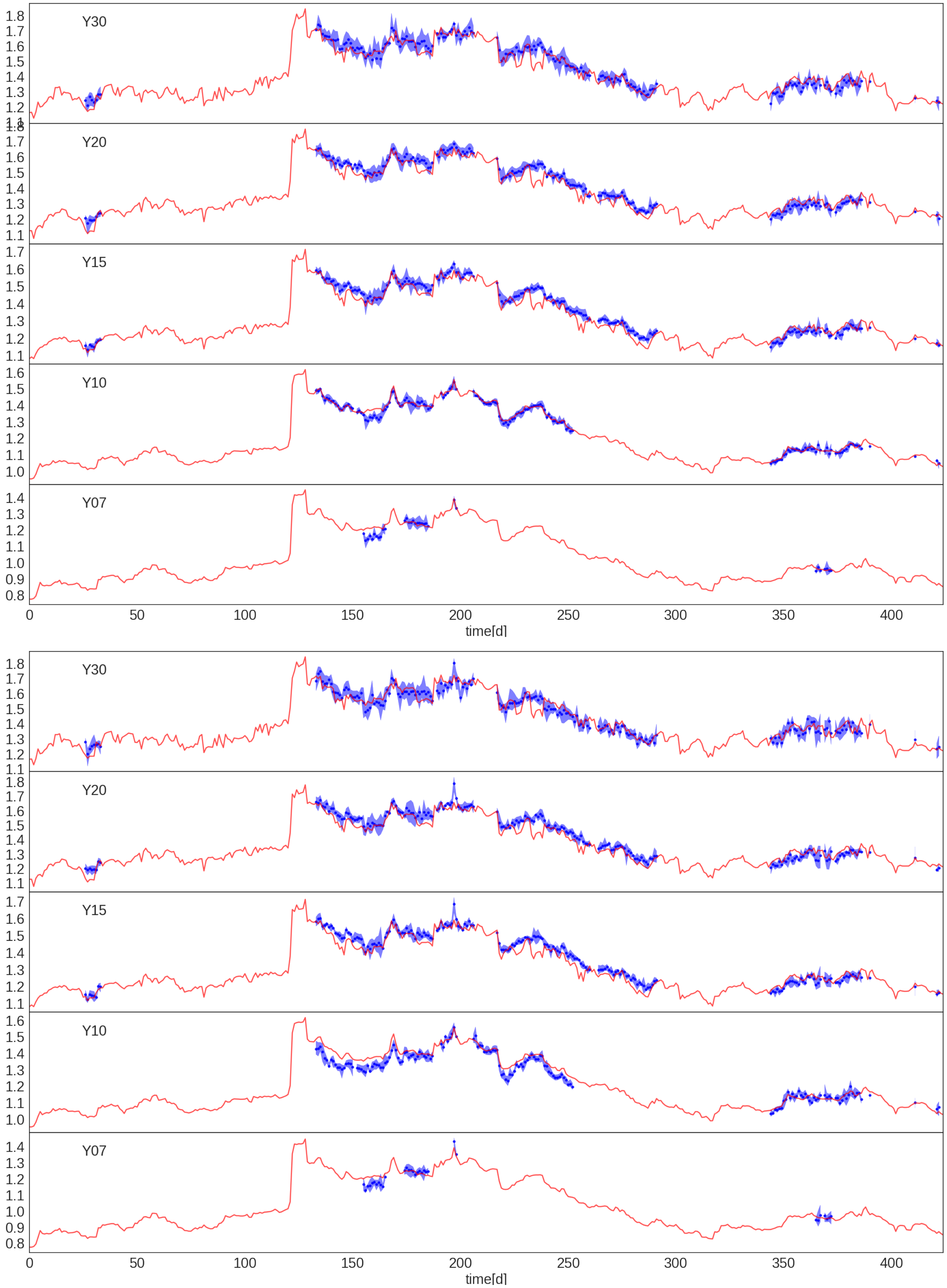

Figure 4. Amelia (top) and mice (bottom) imputed time series for data in Fig. 3 (dots), compared with the ground truth (lines) for the longer maturities. The shaded region indicates minimum and maximum for 5 imputations

	Amelia	DINEOF	mice	RF	MSSA
mean	0.017	0.024	0.031	0.019	0.016
std	0.010	0.019	0.032	0.014	0.011
min	0.002	0.001	0.002	0.000	0.001
max	0.057	0.141	0.374	0.077	0.102

Table 1. Summary statistics for MRD metrics for cluster 1 in comparison: Random Forest (RF), DINEOF, MSSA, and average result out of 5 imputation for Amelia, mice

	Amelia	DINEOF	mice	RF	MSSA
mean	0.035	0.064	0.052	0.046	0.048
std	0.035	0.053	0.056	0.057	0.056
min	0.005	0.011	0.009	0.002	0.005
max	0.328	0.384	0.497	0.483	0.492

Table 2. Summary statistics for MRD metrics for cluster 2.

cluster 1 have relatively few missing values (1.5% on average) and they come in short stretches such that the imputation is fairly straightforward.

A more challenging situation occurs for the patterns in cluster 2, which have a larger missingness fraction of 13% on average. The MRD results in table 2 are also still quite accurate with typical values 2-7%. Amelia shows the strongest performance followed by RF and MSSA. The matrix factorizaton approach (DINEOF) is less successful imputing the patterns here. We will see this in more detail later, when we directly compare the imputations with the ground truth values.

	Amelia	DINEOF	mice	RF	MSSA
mean	0.028	0.064	0.046	0.037	0.041
std	0.015	0.054	0.052	0.032	0.041
min	0.005	0.011	0.009	0.002	0.005
max	0.104	0.384	0.497	0.256	0.342

Table 3. Summary statistics for MRD metrics for cluster 2 where patterns were filtered out if they have rows entirely missing.

Cluster 2 contains 20 patterns which have stretches where observations are missing for all maturities for a number of consecutive time steps. These cases are particularly difficult to impute with the methods discussed here. An imputation based on a proxy, i.e. external data which is directly related, is likely to be more successful. This is beyond the scope of this paper. It is of interest to look at the metrics when these patterns are filtered out. The summary statistics for this are shown in table 3. We can see that the performance improves a little bit for all of the techniques except for DINEOF.

The patterns in cluster 3 typically have more missing values (about 19% on average) and long stretches of missing values for the long maturities. The summary statistics for MRD can be found in table 4. The values for MRD are spread typically between 3-20% with means of the order of 10%. As in cluster 2 Amelia shows the strongest performance with an average of about 9%, followed by RF, mice and MSSA and DINEOF.

	Amelia	DINEOF	mice	RF	MSSA
mean	0.093	0.141	0.111	0.098	0.128
std	0.135	0.121	0.158	0.103	0.125
min	0.009	0.012	0.010	0.014	0.008
max	0.980	0.728	1.522	0.650	0.739

Table 4. Summary statistics for MRD metrics for cluster 3 in comparison.

	Amelia	DINEOF	mice	RF	MSSA
mean	0.061	0.135	0.095	0.092	0.126
std	0.084	0.124	0.155	0.104	0.129
min	0.009	0.012	0.010	0.014	0.008
max	0.705	0.728	1.522	0.650	0.739

Table 5. Summary statistics for MRD metrics for cluster 3 where patterns were filtered out if they have rows entirely missing.

Cluster 3 contains 23 patterns which have stretches where observations are missing for all maturities for a number of consecutive time steps. The results for when we remove those are shown in table 5. The performance improves considerably for Amelia and moderately for the others.

Both EOF based approaches depend on the starting values. We have also tested a combined Amelia+MSSA approach, where Amelia predictions are used as starting values for the MSSA algorithm. For cluster 3 we found mean MRD of 0.099, which is a considerable improvement over the pure MSSA approach (0.128) with naive mean starting values.

We now compare in more detail the predictions of the different techniques. As an example we choose ticker 1 (cluster 2) for which the complete data and the missingness mask were shown in Fig. 3. It misses a lot of consecutive values for the longer maturities and the total missingness fraction is around 17%. The Amelia imputations are shown in Fig. 4 (top). From the 5 imputations we computed the average values shown as dots, and the shaded region indicates the spread between maximum and minimum for the imputation range. The full line shows the ground truth. We can see that the imputated data follows relatively well the general structure of the data, which is inferred well from the correlation with the other series. The spread for the higher maturities is a bit larger. Hence, Amelia learns the the correlations with the other (more complete) time series well and imputes both the temporal structure as well as the magniture of the values accurately. The value for MRD is only 0.02.

The mice imputations for the same time series are shown in Fig. 4 (bottom) and are quite similar to the Amelia results in this case. Again we can see that the imputated data follows relatively well the general structure of the data. The spread appears to be a bit larger than in the case of Amelia. The value

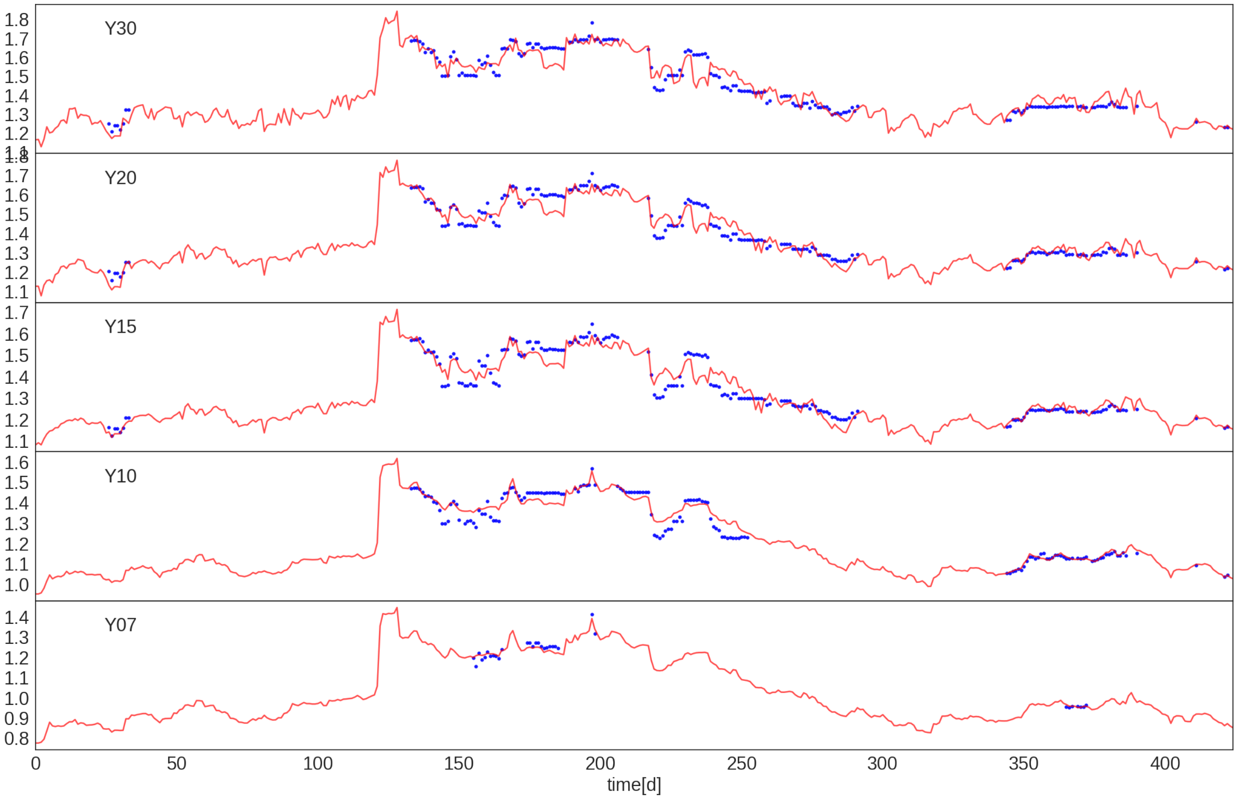

Figure 5. RF imputation (dots) for data in Fig. 3, compared with the ground truth (lines) for the longer maturities.

for MRD is with 0.024 slightly worse than for Amelia.

The imputations for RF, DINEOF, and MSSA are collected in Figs. 5 and 6. The values for MRD are 0.025, 0.044, 0.019, respectively. A number of observations can be made: RF imputes the magnitudes of the values relatively well, however, it does not follow the temporal structure very faithfully, and produces somewhat artificial results. These can be nearly constant over certain periods or possess unexpected discontinuities. In contrast, the DINEOF approach reproduces the overall temporal structures quite well, but does not predict the magnitude of the values as accurately as the other techniques. DINEOF systematically underestimates the magnitude of the values in many cases. EOF based matrix interpolation techniques work best when there are enough data points in the vicinity of the missing values and the data has clear enough structure with variations on not too short scales. As shown in Sec. 4 this can work very well for images, however, the CDS data and missingness patterns here possess different features, for which DINEOF does not perform so well. Finally, MSSA reproduces the structures and values most accurately and performs competitively with the MI technique Amelia. Since it is based on an EOF basis expansion, it tends to smooth the curves somewhat.

For this example in cluster 2 the overall performance of the imputations is quite strong. Since we have patterns which are alternating between missing and observed data the algorithms can learn the structure well. The situation is different in cluster 3, where we have much longer streaks of missing values. We show an example in Fig. 7. The data for the long maturities is missing over almost the entire time period and partially for shorter maturities. The total missing fraction is 45%, which is rather high.

The Amelia imputation is shown in Fig. 8 for the longer maturities[5] MRD is with 0.166 comparatively high. We can see that in contrast to the previous case the short periods towards the end where all time series are observed are not enough to learn the correlation pattern well enough to impute the long stretches of missing data in the past. For the maturities (Y07, Y10) which are closer to the observed data of shorter maturities the time period between 280-400 is imputed quite well, whereas for the longer maturities (Y15, Y20, Y30) the imputations systematically underestimate the true result. The imputations for the time period 0-180 are not satisfactory in all cases, and do not describe the trend correctly. This demonstrates limitations for cases where correlations can not be learned well enough from the observed data. It is worth noting that some of the temporal structure is described fairly well, whilst the overall values are estimated inaccurately in many cases.

For the same data we also show the result for the MSSA imputation in Fig. 9. MRD is 0.22. We find that the interme-

[5] The shorter maturities are imputed accurately and are not shown.

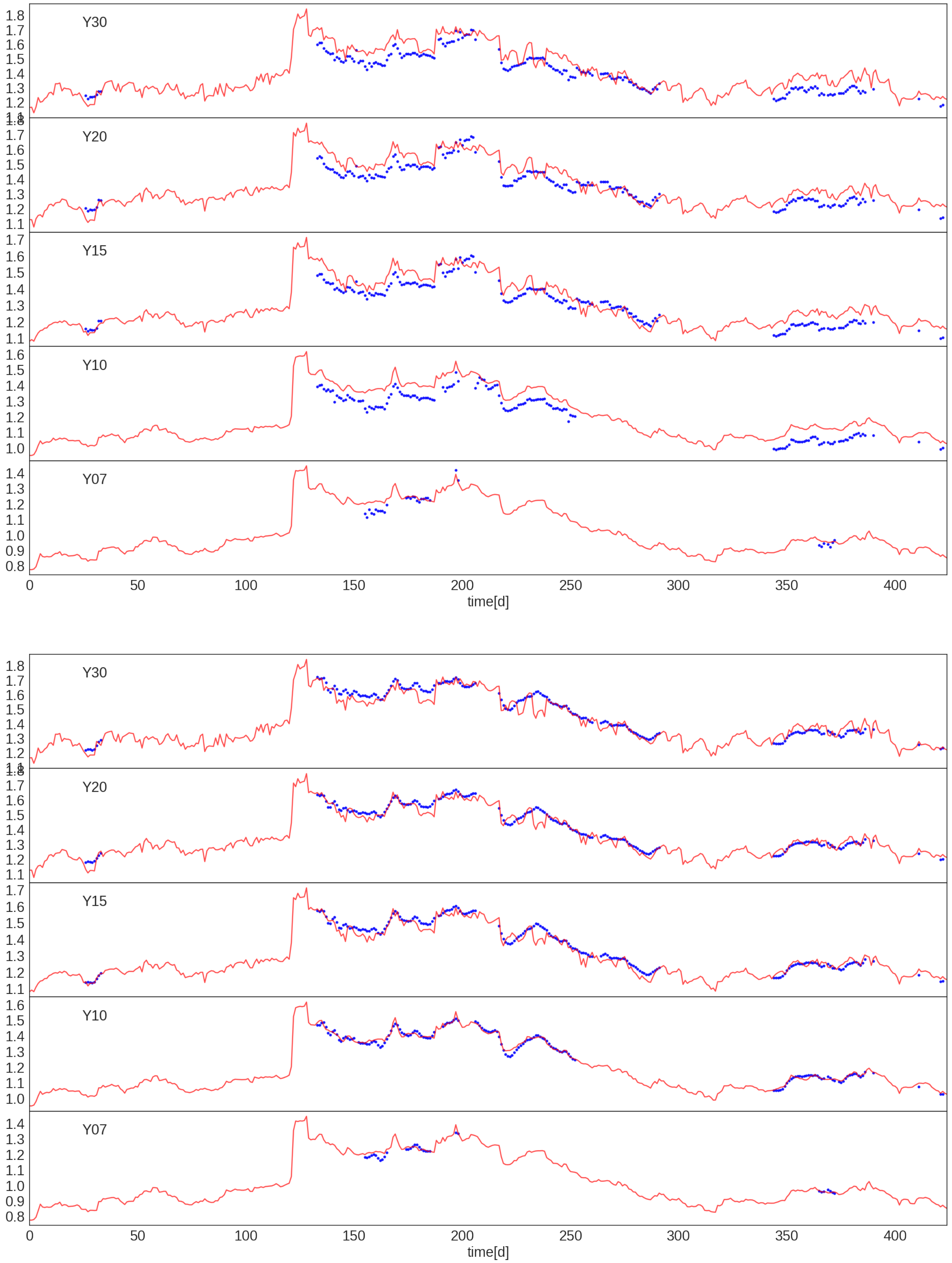

Figure 6. DINEOF (top) and MSSA (bottom) imputation (dots) for data in Fig. 3, compared with the ground truth (lines) for the longer maturities.

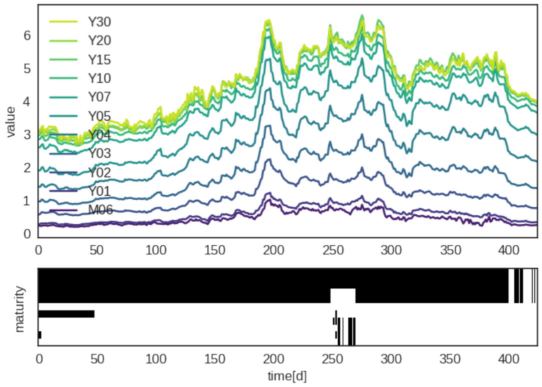

Figure 7. Example of complete time series data (ticker 40, cluster 3). Lower part shows the missingness pattern which is imposed on the complete data.

diate maturites where some observations are present in the middle of the time period give satisfactory results for the time period between 280-400, whereas for the longest maturities the correlations have not been learned well enough from the few observed values and thus the imputations yield too little variation. The other techniques (mice, RF, DINEOF) have similar difficulties to impute this data set and are not shown.

In summary we therefore conclude that when there are few common observations and long stretches with missing data the generic methods used here do not perform very favorably. One therefore has to accept inaccuracies of about 20%. Domain specific techniques including prior knowledge about the data can be more performant then.

6. Conclusions

We have introduced a structured framework for approaching and benchmarking the problem of filling missing data for multivariate time series. As an example we used a large sample of CDS daily quotes over a period of nearly two years. We demonstrated how missing data patterns can be classified and similar patterns can be collected into clusters, an approach which is not found in the literature. We also devised a strategy to generate test sets with realistic missingness patterns, which can be used for performance tests.

We introduced and described a variety of state-of-the-art stochastic MI techniques and deterministic, mainly EOF based techniques. We then compared the imputation performance of these techniques. As a performance measure we mainly focused on the mean relative deviation (MRD), which calculates by what percentage the imputations differ from the true values. We ran imputations for three different clusters with different missingness characteristics for 200 samples each. For the patterns in the first cluster with small overall missingness fraction (1.5%) the performance of all methods was comparable and values for MRD of about 0.02 could be achieved.

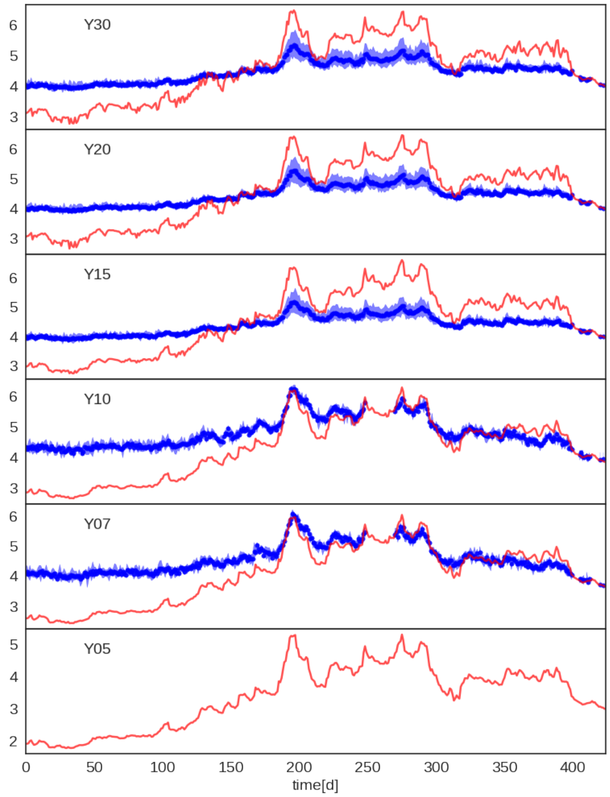

Figure 8. Amelia imputed time series for data in Fig. 7 (dots), compared with the ground truth (lines) for the longer maturities.

For missingness patterns with higher missingness fraction the performance of the techniques varied considerably. We found strong and robust performance for the EM based algorithm, Amelia. It has good time series support natively including leads and lags, and as such is easier to apply for multivariate time series than mice, the other MI package tested thoroughly. We attribute the unfavorable performance of mice to less well-developed time series support, however, further developments might overcome this problem. It is noteworthy that Amelia works very well, in spite of the fact that the data is not normally distributed in contrast to the basic assumption in this framework. This corroborates the findings in the literature [10] that deviations from MVN distribution are tolerable when using Amelia.

For the EOF based techniques, we typically find that MSSA outperforms the DINEOF approach, as it can resolve the intra and intertimes series patterns more precisely. Both of these techniques tend to smoothen the curves, which might be undesirable for specific applications. They also depend on the starting values used in the iterative framework as well as the held out validation set, which is randomly chosen in the beginning of the algorithm. We found that a combined approach, Amelia+MSSA, where Amelia's imputation are used as starting values for MSSA can be a strong performer.

As a well-known machine learning technique for imputa-

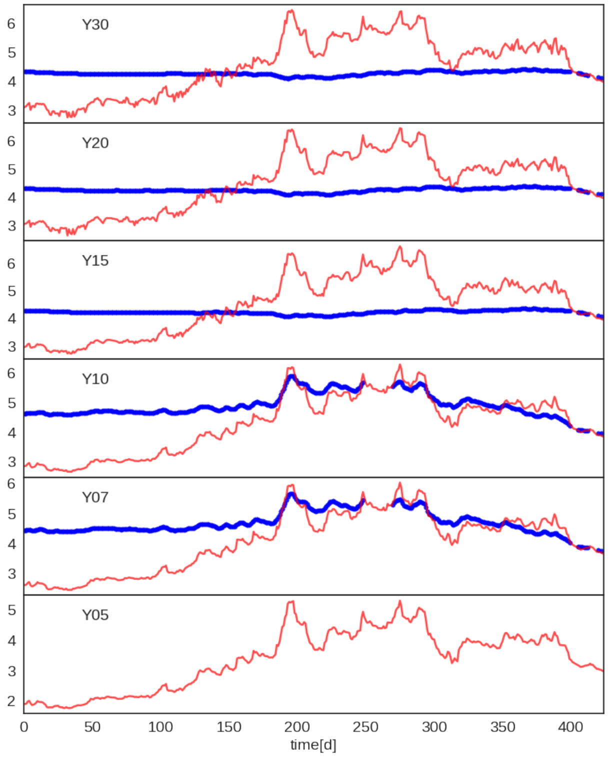

Figure 9. MSSA imputed time series for data in Fig. 7 (dots), compared with the ground truth (lines) for the longer maturities.

tion we tested Random Forest. It showed good performance in terms of the MRD metric, but it can supply somewhat artificial imputations, i.e., curve shapes which are too flat or discontinuous.

In conclusion, Amelia usually outperforms the other techniques and can be seen as the method of choice given that it is also quite fast. Only in certain situation MSSA can capture the temporal structure and correlation properties more accurately and yield lower MRD values. Our findings can serve as a guideline and should be useful in many other comparable settings for multivariate time series data, where missing values need to be imputed.

Acknowledgments

We thank Daniel Jones for helpful discussion.

Appendix

1. General description of MI procedure

As introduced in Sec. 3 a particular MI framework is termed *multiple imputations by chained equations*. Chained equations refers to a an iterative procedure, by which features and parameter values are generated in series of steps.

The general assumption is that the (complete) data is generated from a multivariate distribution function, $p(X|\theta)$, where θ is a collection of parameters, which is not known. In certain cases, the distribution function p can be assumed to have a particular form, e.g., MVN, which we discussed in in more detail in Sec. 3. If not specified explicitly, it must at least implicitly be accessible for sampling. In the following we describe a generic procedure by which parameters are estimated from data, and in turn new data estimates are generated from the corresponding distributions [9]. To keep the description very generic in this section it is described as a pure Monte Carlo sampling approach. In particular, the description is a type of sampling, where we step-by-step draw values from a multivariate conditional distribution function $p(X_1,\ldots,X_p|\theta)$, which is called a Gibbs sampler. It is a member of the Markov Chain Monte Carlo (MCMC) family.

This chained equation procedure can be described as follows:

At iteration t we determine $\theta_1^{(t)}$ by sampling from a distribution,

$$p_\theta(\theta_1|X_1^{\text{obs}},X_2^{(t-1)},\ldots,X_P^{(t-1)}), \tag{32}$$

where $X_i^{(t)} = (X_i^{\text{obs}}, X_i^{\text{imp},(t)})$ includes both observed and imputed data, and $X_2^{(t-1)},\ldots,X_P^{(t-1)}$ were determined at step $t-1$. For the first step some initial guess must be used. We can imagine the distribution in (32) as a being derived from a prior and a likelihood function in a Bayesian framework.

We call the value sampled from Eq. (32) $\theta_1^{(t)}$. New imputation values for the first feature X_1 are then obtained by sampling from the distribution,

$$p_x(X_1|X_1^{\text{obs}},X_2^{(t-1)},\ldots,X_P^{(t-1)},\theta_1^{(t)}). \tag{33}$$

So we take into account the previously sampled parameter vector $\theta_1^{(t)}$. The next step is to sample for θ_2 and X_2 in similar fashion as above. The only difference is that we take the imputed values for X_1, i.e., $X_i^{(t)}$ into account. Hence, the order in which imputations are made matters. This continues for all P features, i.e., we sample θ_P from

$$p_\theta(\theta_P|X_P^{\text{obs}},X_1^{(t)},X_2^{(t)},\ldots,X_{P-1}^{(t)}), \tag{34}$$

and the new values for X_P from,

$$p_x(X_P|X_P^{\text{obs}},X_1^{(t)},X_2^{(t)},\ldots,X_{P-1}^{(t)},\theta_P^{(t)}). \tag{35}$$

Once this has finished we can start with iteration $t+1$.

A particular example for this procedure occurs, when we can assume that the complete data is generated by a MVN, i.e., $X \sim \mathcal{N}(\boldsymbol{\mu},\Sigma)$ and $\theta = (\boldsymbol{\mu},\Sigma)$. Then all the distribution functions can be given explicitly and the procedure becomes a bit more transparent as shown in Sec. 3.

2. Software libraries used in this publication

We used a number of software package for the imputations performed in this paper. Most of them are freely available package, which we detail and reference below. For multiple imputation there are two packages based on chained equations **MI** [14] and **MICE** [9]. We initially tested both of them, but for the more involved performance studies focussed on **MICE**, due to the slightly simpler API. The standard R package for MI based on EM is Amelia II [10].

B.1 Mice

Mice is an R package which is available from comprehensive R repository CRAN. Its functionality is documented in Ref. [9]. We used it for the multivariate time series data as specified in Sec. 5 explicitly including leads and lags. We also used an option to choose predictor variables for the imputation based on availability and aimed for using only variables which are present in more than 50% of the cases. Predictor models, such as Bayesian linear regression and predictive mean matching have been discussed in the main text.

B.2 Amelia II

The R package Amelia II is available from CRAN and follows the algorithm described in Sec. 3.2. It has direct time series support with a number of different options. One of them is to include time polynomials up to third order as additional variables in the covariance matrix. Another option is using time lagged variables (lead, lagged). The way this is works is that we not only use variables but also variables shifted by one unit of time into account and thus enlarge the covariance matrix accordingly. Amelia II usse a bootstrap approach to account for the variance of the paramters $\theta = (\boldsymbol{\mu}, \Sigma)$.

B.3 MissForest: Random Forest imputation

Random Forest (RF) is a very successful technique for regression and classification, which learns feature interactions well and naturally handles different data types [5]. It has been proposed as a suitable tool for data imputation as well [15]. The algorithm proceeds as follows: We start with an initial guess for the missing values. Then for each feature (or time series component) p which contains missing values we train a RF prediction model from the available data. This can be used to generate improved imputations. We iterate through all the p features, which contain missing values. In the next iteration we use the imputed values of the last iteration. The iterations stop once the imputed values do not change much anymore from iteration to iteration. These converged values are used as imputations for the missing data.

Random Forest data imputation is available as the R library **MissForest** [15], which is available from CRAN. It is well documented. It is noteworthy that the default setting for the subset of features used for the imputation is $\sqrt{P}$. For the imputations in this paper we increased this to $\sim P/2$ to improve the accuracy.

B.4 DINEOF

The matrix interpolation approach DINEOF is based on SVD of the data matrix and a suitable reconstruction. We used the R package **sinkr**, which is available from the repository `https://github.com/menugget/sinkr`. It is worth pointing out that this package assumes that the data has zero mean. Therefore, we subtracted the mean from the input data and added it again after the imputation.

B.5 MSSA

As discussed in the main text the MSSA approach is based on SVD of the trajectory matrices and reconstruction employing the EOFs. This can be technically be achieved by the R package **RSSA** [12]. To treat the multivariate time series we found it favourable to use the ”2dSSA” option rather than the ”MSSA” option, even though the latter also works. We are not aware of a full implementation to treat the general missing data problem. Therefore, we wrote our own routine performing the iterations as described in Sec. 4 and in Ref. [13].

References

[1] P. Kofman and Ian G. Sharpe. Using Multiple Imputation in the Analysis of Incomplete Observations in Finance. *Journal of Financial Econometrics*, 1(2):216–249, jun 2003.

[2] Panteha Hayati Rezvan, Katherine J Lee, and Julie A Simpson. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Medical Research Methodology*, 15(30):1–14, 2015.

[3] Roderick J a Little and Donald B Rubin. *Statistical Analysis with Missing Data*. Wiley, 2nd ed. edition, 2002.

[4] J.L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman&Hall, first edition, 1997.

[5] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[6] Roderick J.A. A. Little. A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*, 83(404):1198, dec 1988.

[7] Kevin P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, 2012.

[8] Craig Enders. *Applied missing Data Analysis*. The Guidford Press, 2010.

[9] Stef van Buuren and Karin Groothuis-Oudshoorn. mice : Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3):1–35, 2011.

[10] James Honaker, Gary King, and Matthew Blackwell. AMELIA II : A Program for Missing Data. *Journal Of Statistical Software*, 45(7):1–54, 2011.

[11] J. M. Beckers and M. Rixen. EOF calculations and data filling from incomplete oceanographic datasets. *Journal of Atmospheric and Oceanic Technology*, 20(12):1839–1856, 2003.

[12] Nina Golyandina, Anton Korobeynikov, Alex Shlemov, and Konstantin Usevich. Multivariate and 2D Extensions of Singular Spectrum Analysis with the <b>Rssa</b> Package. *Journal of Statistical Software*, 67(2):1–78, oct 2015.

[13] D. Kondrashov and M Ghil. Spatio-temporal filling of missing points in geophysical data sets. *Nonlinear Processes in Geophysics*, 13(1):151–159, 2006.

[14] Yu-Sung Su, Jennifer Hill, Andrew Gelman, and Masanao Yajim. Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box. *Journal Of Statistical Software*, 45(2):1–31, 2011.

[15] Daniel J Stekhoven and Peter Bühlmann. MissForest—non-parametric missing value imputation for mixed-type data. *BIOINFORMATICS*, 28(1):112–118, 2012.