Imputation of Missing Values in Time Series with Lagged Correlations

Shah Atiqur Rahman, Yuxiao Huang
Stevens Institute of Technology, NJ, USA
{*srahman1, yhuang23*}*@stevens.edu*

Jan Claassen
Columbia University, NY, USA
jc1439@cumc.columbia.edu

Samantha Kleinberg
Stevens Institute of Technology, NJ, USA
skleinbe@stevens.edu

***Abstract*—Missing values are a common problem in real world data and are particularly prevalent in biomedical time series, where a patient's medical record may be split across multiple institutions or a device may briefly fail. These data are not missing completely at random, so ignoring the missing values can lead to bias and error during data mining. However, current methods for imputing missing values have yet to account for the fact that variables are correlated and that those relationships exist across time. To address this, we propose an imputation method (FLk-NN) that incorporates time lagged correlations both within and across variables by combining two imputation methods, based on an extension to k-NN and the Fourier transform. This enables imputation of missing values even when all data at a time point is missing and when there are different types of missingness both within and across variables. In comparison to other approaches on two biological datasets (simulated glucose in Type 1 diabetes and multi-modality neurological ICU monitoring) the proposed method has the highest imputation accuracy. This was true for up to half the data being missing and when consecutive missing values are a significant fraction of the overall time series length.**

***Keywords*-missing data; correlated data with time-lag; extended k-NN imputation; Fourier imputation;**

I. Introduction

Missing values occur in almost all real world data, and this problem is particularly prevalent in clinical data [1] due to equipment errors, varied sampling granularity or fragmented data. It may be that a sensor became disconnected, different sensors record values at different intervals, a patient changed hospitals during their medical care, or a study subject refused to answer questions.

However, simply ignoring these missing values can lead to computational problems such as bias (if an expensive lab test is only ordered when a doctor suspects it will be positive), difficulties in model learning (when different subsets of variables are present for different patients), and reduced power (if many cases with missing values are not used). Hence, methods for handling missing data are increasingly important even as the amount of available data grows.

One challenge is that there are multiple types of missingness, that each require different imputation strategies:
Missing completely at random (MCAR) means the probability of a variable's data being missing, $P(V)$, is independent of both the other observed variables, O, and V. That is:

$$P(V|V,O) = P(V). \quad (1)$$

For example, a person wearing a sensor that communicates through wireless signals may be in and out of wireless coverage for a period of time.
Missing at random (MAR) is when the probability of data for V being missing is dependent on variables other than V. Mathematically,

$$P(V|V,O) = P(V|O). \quad (2)$$

For instance, the likelihood of a particular test being done (and its value being recorded) may depend in part on a patient's health insurance.
Not missing at random (NMAR) data are those that are neither MCAR nor MAR. In this case, the probability of a variable being missing may depend on the missing variable itself. For instance, people with normal blood pressure may measure their blood pressure less frequently than people with high blood pressure. Thus, blood pressure would be NMAR as its presence depends on itself rather than on other measured variables. With MAR and MCAR, one can focus on correlations between missing and observed data, while NMAR requires specification of the missingness model.

A number of approaches have been developed for estimating missing values, but the existing methods have failed to address a few key issues: correlations between variables across time, multiple types of missingness within a variable, and timepoints where all data are missing.

Take figure 1, which shows three variables over time. Say variable x is correlated with both y and z at two different lags (l_{xy} and l_{xz} respectively). If x is missing at time t, then existing methods impute this value using the values

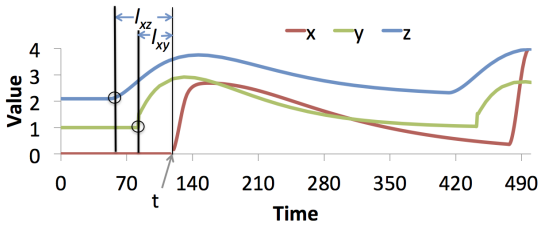


Figure 1: Variables with lagged correlation. x is missing at t, l_{xy} and l_{xz} are time lag of x with y and z respectively

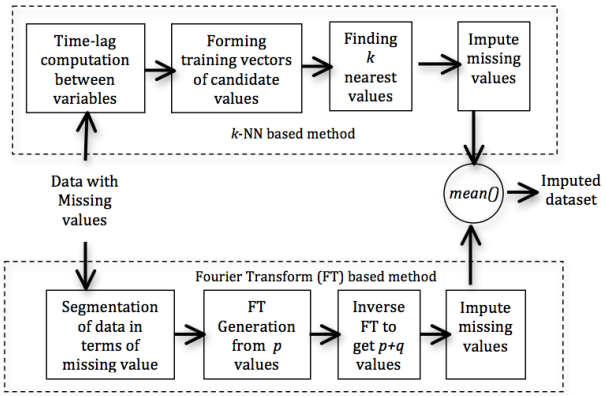


Figure 2: System block diagram. Here, k is the number of nearest neighbors, p is the number of observed values from beginning to prior data point of a missing value, q is the number of missing values after those observed p values.

of y and z at time t. However this is incorrect due to the lagged correlation. Instead, imputation should be based on the values of y and z at $t - l_{xy}$ and $t - l_{xz}$ respectively.

Second, there may be multiple types of missingness in a variable, yet most current methods (e.g. [2], [3]) assume that each variable can only have one type of missingness. In biomedical data, the presence of a variable is often dependent on both the variable itself and other observed variables (making a single variable both MAR and NMAR). For instance, the amount of glucose in a person's stomach at a particular time is divided into two parts, liquid and solid glucose. The amount of liquid glucose depends on both, so missing values can be both MAR and NMAR.

Finally, single devices are often used to measure multiple signals (e.g. cellphone accelerometer and GPS, laboratory panel), making it likely that multiple values will be missing at a single instance. This poses challenges for many methods, which require some non-missing data to impute values for a particular instance and thus may fail to impute a value for some instances.

In this paper we propose a method, FLk-NN, to impute missing data from continuous-valued time series, where there may be lagged correlations between variables, data may be both MAR and NMAR, and entire time points may be missing. We compare the approach to others on multiple datasets from the biological domain (one simulated and one real-world dataset), demonstrating that FLk-NN has the highest imputation accuracy for all ratios of missing data on both datasets, even with up to 50% of the dataset missing and while being able to impute values for all missing points.

II. Related work

We briefly describe existing methods for handling missing data and refer the reader to [4], [5] for a full review.

Table I: Comparison of methods, with ours being FLk-NN.

Name	Impute empty instances	Lagged relationships	Multi missingness in a variable
MEI	Yes	No	No
k-NN	No	No	No
Model-based	Yes	No	No
EM	Yes	No	No
Probabilistic EM	No	No	No
MICE	Yes	No	No
FLk-NN	Yes	Yes	Yes

A. Ignoring Missing Values

The simplest way of handling missing data is to remove the data instances that contain missing elements. Common examples are list-wise deletion, which removes all instances having at least one missing value, and pairwise deletion, which removes an instance if the variables currently being used contain missing values [2]. In this method, results may be biased when data are non-MCAR and the statistical power is reduced due to the deletion of information.

B. Single Imputation (SI)

These methods replace each missing value with a single imputed value. The simplest and most efficient SI method is mean or mode imputation (MEI), which fills the missing values by either attribute mean (for continuous values) or mode (for nominal values) [6], [7]. However, this method assumes that data are MCAR, which is not true in most real world biologic or other data, and it overestimates the precision of measurements [4].

k-Nearest Neighbor (k-NN) based methods identify the k most similar instances to the one with a missing value based on the observed values of the other variables at that instance. They then apply a predefined rule (e.g. weighted average [8]) or kernel function (e.g. exponential kernel [9]) to impute a value based on these instances. The case where $k = 1$ is called hot deck. Usually, k-NN based methods are more accurate than MEI but require enough complete instances to identify the neighbors [10], and one must determine the appropriate number of neighbors to use. Most critically, these methods cannot impute if all variables at an instance are missing. This is a major limitation when measurements come from one device or when they are always either all present or absent.

Model based methods [11] learn a model from non-missing attributes. The learning task is then classification for nominal attributes and regression for continuous attributes, and the model is used to impute values. These methods are time intensive for the learning process and the model structure is problem specific.

Statistical methods exist based on maximum likelihood, and expectation maximization [12]. Expectation Maximiza-

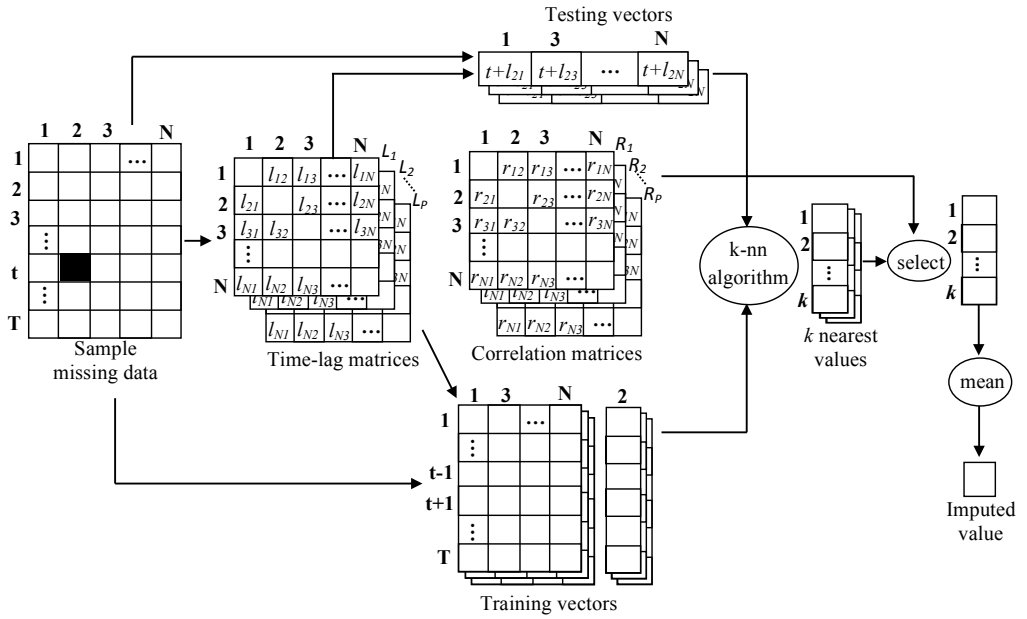


Figure 3: An example of Lk-NN for a single missing value (indicated by the black cell), where N is the number of variables, T is the number of time-instances, L_i is the i^{th} time lag matrix, l_{xy} is the time lag from x to y variable, p is the number of lag and correlation matrices, and k is the number of nearest neighbors.

tion (EM) methods iteratively impute missing values and update distribution parameters. These methods give higher classification accuracy compared with model based imputations [13]. However, EM algorithms are computationally expensive and problem specific for the iterative E-M steps.

In clustering based SI methods [14], [15], data are first clustered using the non-missing values and then missing values are imputed using the instances of the cluster that contain the missing value instance. A hybrid clustering and model based method was proposed by Nishanth et al. [16] where they combine k-means with artificial neural network (ANN) and found that the method is more accurate than individual model based techniques (e.g. ANN) on financial data. However, the performance decreases when there are fewer complete instances and a higher missing rate.

C. Multiple Imputation (MI)

Multiple imputation contains two phases, i) imputation, where multiple estimates of a missing value are generated, and ii) combination, where inferences from each of the imputed datasets are combined [17].

Methods used in the imputation phase can be divided into two categories: i) multi-variate normal (MVN) model that assumes that the variables are continuous and normally distributed and ii) ICE or MICE (Multivariate Imputation by Chained Equations) which uses a chained equation to fill the missing values [18], [19]. MICE has several advantages over MVN such as mixed variable type (e.g. continuous, categorical), and skewed continuous variables, shown experimentally by Bouhlila et al. [20]. MICE can impute when variables have different types of missingness, but not when multiple types of missingness occur within a single variable.

The combination of results can be done by averaging [21], [22], Bagging [22], and boosting [23]. Schomaker et al. [22] experimented on simulated data and showed that model averaging can give stable estimation of different parameters like standard errors, and confidence interval.

The existing methods make two primary assumptions that may not hold, and are particularly problematic with biomedical data. First, in MAR, variables are assumed to be correlated with no time lag. Second, each variable is assumed to have only one type of missingness, whereas we often need to impute missing data whose value depends on both the missing variable and other variables (i.e. missingness is both MAR and NMAR). Moreover, likelihood based methods (e.g. k-NN) are not able to impute at all if all the values along an instance are missing. A brief comparison of our approach and others is shown in Table I, where methods are compared in terms of: ability to impute completely missing time instances, inclusion of time lags for correlations, and ability to handle variables that are both MAR and NMAR.

III. Method

We now introduce a new method for imputing missing values in time series data with lagged correlations and multiple types of missingness within a variable. Our proposed method is a combination of two imputation methods: i) an extension of k-NN imputation with lagged correlations and

ii) the Fourier transform. The system block diagram is shown in figure 2.

First, we develop an extension to k-NN with time lagged correlations using cross-correlation. Since correlations may persist for a period of time and time measurements may be uncertain, we introduce lagged k-NN (Lk-NN), which has two parameters: k, the number of nearest neighbors, and p the number of time lags. Thus we take the p lags with the strongest correlation for each pair of variables and then later the k nearest neighbors across all lags (weighted by the strength of the correlation), averaging the results. This enables better handling of MAR data by improving the handling of time-dependent correlations. However, this does not take into account correlations within a variable and cannot be used when all data at the lagged timepoints are missing, so we also develop an imputation approach based on the Fourier transform, which uses only the data for each variable to impute its values.

Results from the two methods are then averaged for each value. Combining Lk-NN with the Fourier-based method overcomes the limitation of nearest neighbors methods requiring some data present at each instance and improves accuracy by handling both MAR and NMAR missing data.

A. Lk-NN Method

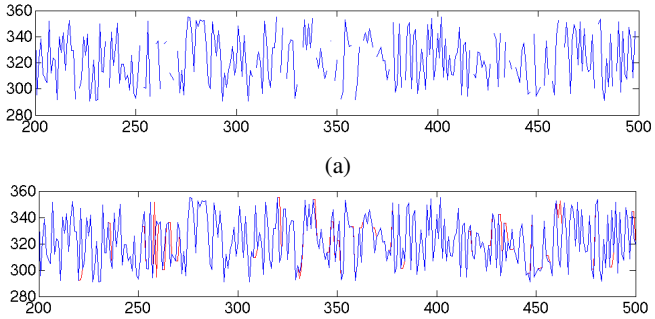
Normally, k-NN finds similar instances by, say comparing the values of variables at time 1 to those at time 10. However, correlations may occur across time. For example, insulin does not affect blood glucose immediately and weight and exercise are correlated at multiple timescales. This was shown in figure 1, where there is a lag between a change in one variable's value and the response in another. To handle this, we develop a new approach for constructing the test and training vectors using lagged correlations, where the time lags can differ between pairs of variables. This is illustrated in figure 3.

1) Calculating Time Lags: To form the test and training vectors, we first identify the correlation between the variables and their timing using cross-correlation. Cross-correlation is a similarity measure of two time-series as a function of a time delay applied to one of them [24]. The cross-correlation, r_{xy}, between two variables, x and y, for time delay d is:

$$r_{xy}(d) = \frac{c_{xy}(d)}{\sqrt{c_{xx}(0)c_{yy}(0)}}, \tag{3}$$

$$c_{xy}(d) = \begin{cases} \frac{1}{T-d}\sum_{t=1}^{T-d}(x_t-\bar{x})(y_{t+d}-\bar{y}), & \text{if } d \geq 0 \\ \frac{1}{T+d}\sum_{t=1-d}^{T}(x_t-\bar{x})(y_{t+d}-\bar{y}), & \text{otherwise} \end{cases} \tag{4}$$

where T is the length of the series, $\bar{x}$ and $\bar{y}$ are the mean of x and y respectively, d varies from $-(D-1)$ to $(D-1)$, and D is the maximum time delay. Since values for x and y may be missing at different timepoints, we use only the instances where both are present in this calculation.



(a)

(b)

Figure 4: An example of Fourier based imputation for one variable, (a) with simulated missing data points, (b) the actual data (in blue) with the imputed data (in red).

Matrices are constructed for each of the p lags, with the correlations ordered from $1 \ldots p$ by decreasing strength. Thus, for each pair of variables L_1 contains the lag, d, with the strongest correlation (max $|r_{xy}|$) and L_p the lag with the weakest. Each L is an $N \times N$ matrix, where elements represent the time lags for each correlation between the N variables. An element l_{xy} can be positive (values of variable y have a delayed response in time unit l_{xy} to values of x) or negative (values of variable x have a delayed response of time unit l_{xy} for values of y) and $l_{xy} = -l_{yx}$. The diagonal elements of the matrix are not computed since those elements give the auto-correlation of the signal and are not used in this algorithm. For all l_{xy}, the corresponding correlation values, $|r_{xy}|$, are stored in the matrices $R_1 \ldots R_p$, which are used in the neighbor selection step.

2) Forming Vectors: Formation of vectors with Lk-NN is more complex than for k-NN since we must account for multiple lags that differ across variable pairs. Instead we create a set of test and training vectors for each of the p lags. Below we describe how to create the vectors for a single lag.

Say a variable, x, is missing at time t and x has a time lagged relationship with variables y and z, with lags l_{xy} and l_{xz} respectively. The test vector is then formed using the values of y and z at $t+l_{xy}$ and $t+l_{xz}$. Training vectors are formed in similar way and the values of x, which are the candidate values for imputation, are stored separately. Training vectors are generated from the existing values of x and the time instances resulting after adding the lags must be within 1 to T (length of data). This makes the boundary of time instances of training vectors for a missing value:

$$[\max(1, 1-\min(l_{x1} \ldots l_{xN})), \min(T, T-\max(l_{x1} \ldots l_{xN}))] \tag{5}$$

where $l_{x1}, \ldots, l_{xN}$ are the time lags of correlations between x and all N variables for the current lag matrix.

Algorithm 1 Fourier transform based imputation

Input:
Data matrix, $Y = \{V_1, V_2, ..., V_N\}$, is a set of variables, where each $V_i = \{v_1, v_2, ..., v_T\}$, and v_j is the j^{th} data point;
Output:
Data matrix, Y with imputed values
for each V in Y **do**
  t_s = $min(j)$, where v_j is missing, $1 \leq j \leq T$;
  while $t_s \neq \emptyset$ **do**
    t_e = $min(j)$, where v_j is non-missing, $t_s \leq j \leq T$;
    F = DFT$(v_1, v_2, \ldots, v_{(t_s-1)})$;
    u = IDFT(F, t_e);
    $v_j = u_j$, where $t_s \leq j \leq t_e$;
    t_s = $min(j)$, where v_j is missing and $1 \leq j \leq T$;
  end while
end for
return Y

3) Finding Neighbors and Imputing Missing Values: Once the lags are found and vectors formed, the next step is finding the nearest neighbors for each missing instance. Since the strength of the correlation between variables and across the p lags may differ substantially, we incorporate a weight into our distance measure. Note that each neighbor may be based on different variables (if some are missing), so this accounts for the correlation of the variables actually present. This ensures that neighbors based on highly correlated variables are given more weight than those with weakly correlated ones.

Most current methods use the Euclidean distance as a proximity measure, but this does not incorporate the differing correlations. Instead we propose a weighted modification of the Euclidean distance that is similar to the Mahalanobis distance but can handle missing values in both test and training vectors. First,

$$d(x,y) = \frac{\sqrt{\sum_{i=1}^{N}(x_i \wedge y_i) \times (x_i - y_i)^2}}{\sum_{i=1}^{N}(x_i \wedge y_i)} \tag{6}$$

where N is the number of variables and we are calculating the distance between x and y. This is the average Euclidean distance between two vectors computed for non-missing pairs of values, where we also keep track of non-missing pairs of variables. The result is p sets of k nearest neighbors (one set of neighbors for each L matrix).

Next, the distances are weighted by the average correlations of the non-missing variables. With the distance between an instance with missing values and one of its neighbors being d, the weighted distance, d_w, is:

$$d_w = d \times (2 - mean_{u \in U}(r_{vu})) \tag{7}$$

where v is the variable whose missing value is being imputed and U is the set of variables that were present during computation of d. Suppose for a missing value of a variable, v, there is a neighbor where the distance between vectors are computed by non-missing pairs of variables, x, y and z. Then the distance is multiplied by the weight $(2 - \text{mean}(r_{vx}, r_{vy}, r_{vz}))$. We then average the values for the k neighbors with the lowest weighted distance (out of the set of $p \times k$ neighbors).

Table II: Variables in DSIM dataset

	Name
G	Glucose concentration
G_p	Glucose mass in plasma
G_t	Glucose mass in tissue
I	Insulin concentration
I_p	Insulin mass in plasma
I_t	Insulin mass in tissue
U_t	Glucose Utilization
X_t	Insulin in the interstitial fluid
EGP	Endogenous glucose production
R_a	Glucose rate of appearence
Q_{sto1}	Solid glucose in stomach
Q_{sto2}	Liquid glucose in stomach
Q_{gut}	glucose mass in the intestine
R_i	Rate of appearence of insulin in plasma
I_{sc1}	Nonmonomeric insulin in subcutaneous space
I_{sc2}	Monomeric insulin in subcutaneous space

B. Fourier Method

While Lk-NN takes into account correlations between variables, we also need a way of accounting for patterns within a variable, in order to handle data that are NMAR. To do this, we develop an imputation method based on the Fourier transform that uses past values of each variable to impute each missing value.

First, a data segment is formed with the data points from the beginning of the signal up to the last non-missing data point. Where values v_1 through v_{p-1} are present (or imputed), and $v_p \ldots v_q$ are missing, the Fourier descriptors are obtained with:

$$F_k = \sum_{j=1}^{p-1} v_j \times \mathrm{e}^{-\frac{2i\pi}{p-1}(j-1)(k-1)} \tag{8}$$

where, F_k is the k^{th} Fourier descriptor with $1 \leq k \leq (p-1)$, and $i = \sqrt{-1}$.

Then, the imputed value for time m, where $p \leq m \leq q$, can be calculated from the Fourier descriptors with:

$$v_m = \frac{1}{p-1} \sum_{k=1}^{p-1} F_k \times \mathrm{e}^{\frac{2i\pi}{p-1}(m-1)(k-1)} \tag{9}$$

where, the notation is same as equation (8). Algorithm 1 shows the process, where DFT(v) generates Fourier descriptors for a variable, v, and IDFT(F, t) regenerates a signal of length t from the Fourier descriptors, F. An example of the result on a set of simulated data is shown in figure 4 where most of the imputed data points are near the actual value.

C. Combining the methods for FLk-NN

For each missing data point, we impute two values using the described methods and then combine these. Since model

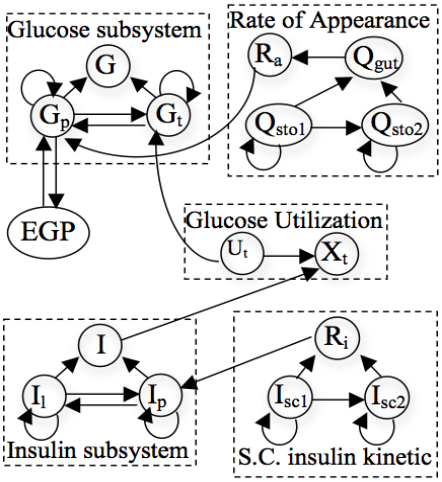


Figure 5: Simulated glucose data variables and relationships.

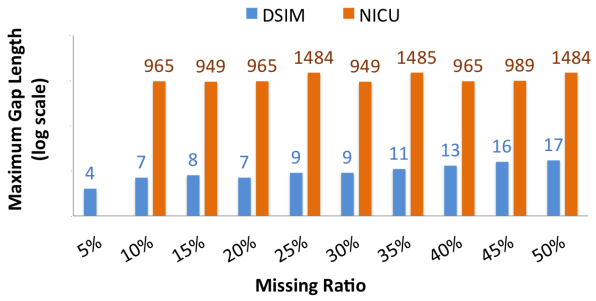


Figure 6: Maximum gap length of DSIM and NICU datasets. Note that the NICU data begins at 10% due to the existing missing values.

averaging gives a more stable and unbiased result compared with other approaches such as bagging and weighted mean [22], we average the value estimated by the two methods, and call the resulting combined approach FLk-NN.

D. Time complexity

The computational complexity of Lk-NN is a combination of two processes: cross-correlation and k-NN. For two time series of the same length, T, and maximum delay, D, the complexity is $O(DT)$ for cross-correlation, which results in $O(\binom{N}{2}DT)$ for N variables. The complexity of k-NN for x missing values is $O(xTN)$. Therefore, the total time complexity of Lk-NN is $O(\binom{N}{2}DT + xTN)$. Note that the efficiency of this method can be improved by a look-up table of distance between instances. In our Fourier method, we used the fast Fourier transform (FFT) algorithm, which has the complexity $O(T \log T)$. Thus the complexity of imputing x missing values with the Fourier method is $O(xT \log T)$. Hence, the complexity of FLk-NN is $O(\binom{N}{2}DT + xTN + xT \log T)$.

Table III: Baseline level of missing data in NICU dataset

Patient	# of variables	original missing
P1	11	0.1%
P2	14	9.37%
P3	16	3.28%
P4	14	8.16%
P5	16	4.62%
P6	18	8.68%
P7	13	9.96%
P8	16	6.57%
P9	18	4.54%

IV. Experimental Results

A. Data

We compared the proposed approach to others on two biomedical datasets that have time lagged correlations between the variables.[1]

Simulated diabetes (DSIM) dataset: We used the glucose-insulin simulation model developed by Dalla-Man et al. [25] to construct a simulated dataset, DSIM. The model describes the physiological events occurring after a meal and was created by fitting the major metabolic fluxes estimated (endogenous glucose production, meal rate of appearance, glucose utilization, and insulin secretion) in a model-independent way on a wide population [25]. This model has been validated with human subjects [25] and approved by the FDA for use in pre-clinical trials [26], and is thus more realistic than examples such as random networks. The model contains a set of submodules that affect one another with varying delays. We generated one day of data for each of 10 patients by randomly selecting patient parameters (e.g. body weight, meal amount and timing, and insulin dose) within realistic ranges (e.g. body weight within 50kg-120kg). Data was recorded at every minute, yielding 1440 time points for the 16 variables listed in Table II. We added Gaussian noise to make the data more similar to real-world cases. The relationships embedded in the model are shown in figure 5.

Real-world NICU dataset: In the second experiment we used physiologic data collected from a set of subarachnoid hemorrhage (SAH) patients admitted to the Neurological intensive care unit (NICU) at Columbia University [27]. Data on cardiac and respiratory variables, and brain perfusion, oxygenation, and metabolism were continuously collected from 48 patients. However, the set of variables collected (a max of 22) differed for each patient as did the number of timepoints, as it covered the duration of ICU stay. Data duration ranged from 2.5 to 24.7 days, with a mean of 12.33 days. The majority of data were recorded at 5 second intervals, which were then minute-averaged so that all recordings

[1]The DSIM data, code, and instructions for replicating results are available at https://github.com/kleinberg-lab/FLK-NN. The NICU data cannot be shared due to HIPAA privacy regulations.

Table IV: Mean (m) and standard deviation (s) of MAE for DSIM dataset.

(a) Missing ratios 5% - 25%.

	5%		10%		15%		20%		25%	
Method	m	s	m	s	m	s	m	s	m	s
BPCA	0.046	0.059	0.047	0.060	0.049	0.062	0.051	0.064	0.052	0.066
EM	0.057	0.064	0.054	0.063	0.053	0.064	0.053	0.065	0.053	0.067
Hot deck	0.053	0.076	0.055	0.080	0.057	0.086	0.059	0.090	0.063	0.096
Inpaint	73.4	342.3	79.8	372.5	82.0	382.1	81.7	389.1	88.4	417.7
k-NN	0.044	0.061	0.046	0.065	0.047	0.069	0.049	0.071	0.051	0.076
MEI	0.179	0.143	0.178	0.142	0.178	0.143	0.178	0.143	0.178	0.142
MICE	0.063	0.091	0.065	0.092	0.065	0.092	0.065	0.091	0.068	0.095
Fourier	0.048	0.073	0.049	0.080	0.050	0.083	0.049	0.079	0.051	0.082
Lk-NN	**0.041**	0.057	0.043	0.060	0.044	0.062	0.045	0.063	0.047	0.066
FLk-NN	**0.041**	0.060	**0.042**	0.066	**0.043**	0.070	**0.043**	0.066	**0.044**	0.068

(b) Missing ratios 30% - 50%.

	30%		35%		40%		45%		50%	
Method	m	s	m	s	m	s	m	s	m	s
BPCA	0.053	0.066	0.055	0.070	0.057	0.070	0.059	0.073	0.061	0.075
EM	0.053	0.067	0.055	0.070	0.056	0.071	0.058	0.073	0.060	0.075
Hot deck	0.069	0.106	0.081	0.122	0.095	0.138	0.108	0.151	0.116	0.15
Inpaint	89.9	423.8	98.8	474.4	100.9	476.7	105.3	499.9	109.1	523.4
k-NN	0.056	0.082	0.064	0.092	0.076	0.104	0.088	0.116	0.098	0.125
MEI	0.178	0.143	0.178	0.141	0.178	0.141	0.178	0.142	0.178	0.142
MICE	0.069	0.097	0.072	0.101	0.074	0.103	0.076	0.106	0.081	0.111
Fourier	0.051	0.083	0.053	0.087	0.053	0.089	0.056	0.095	0.058	0.099
Lk-NN	0.048	0.066	0.048	0.067	0.049	0.069	0.052	0.071	0.057	0.083
FLk-NN	**0.045**	0.069	**0.046**	0.071	**0.047**	0.076	**0.049**	0.083	**0.051**	0.083

were synchronized to the same time points. This resulted in an average of 17,771 time points for each patient, with a standard deviation of 10,216. As the amount of missing data differed widely due to factors such as interventions, device malfunctions and loss of connectivity between the device and network, we selected a subset of 9 patients with fewer missing values and used 3 days of data. It was necessary to ensure a sufficient amount of data present at the start, as we later removed varying amounts of data to test the methods and compare imputed to actual values. Table III gives the baseline amount of missing data for each subject. For the simulated missing data, the missing ratios indicate the total fraction of missing values (original + simulated).

B. Procedure

We created synthetic missing data by deleting randomly selected values. If the selected data point was already missing (which can occur in the NICU dataset), we select another and repeat this until the target missing ratio is reached. The ratios are 5% to 50% and 10% to 50% in increments of 5% for DSIM and NICU respectively. The maximum length of consecutively missing values (gaps) for both the datasets are shown in figure 6. The maximum gap length is 17 for DSIM and 1,485 for NICU.

We compared our system with several commonly used state-of-the-art methods from different categories.

MEI [7]: Missing values are imputed by computing the mean of non-missing values of a variable.

Hot deck and k-NN [8]: Euclidean distance is used to find the k neighbors and the weighted average of these is used to impute. For k-NN, we used k = 5, which gives the best result for this algorithm and for Hot Deck k is always 1.

BPCA [12]. This probabilistic method applies Bayesian principle component analysis prior to the conventional E-M process. We used the authors' BPCAfill.m code[2] with two parameters set to their default values, k = number of variable -1 and $maxepoch = 200$.

EM [28]: This iterated linear regression analysis replaces the conditional maximum likelihood estimation of regression parameters in the traditional E-M algorithm with a regularized estimation method. We used the RegEM package[3] with the default values for the parameters (e.g. maximum number of iteration: 30, regression method used: multiple ridge regression).

Inpaint [29]: This statistical model based approach extrapolates non-missing elements using an iterative process. We used the authors' code[4] with the default value for number of iterations, which is 100.

[2]http://ishiilab.jp/member/oba/tools/BPCAFill.html

[3]http://www.clidyn.ethz.ch/imputation/

[4]http://www.mathworks.com/matlabcentral/fileexchange/27994-inpaint-over-missing-data-in-n-d-arrays

Table V: Mean and standard deviation of MAE for NICU dataset.

(a) Missing ratios 10% - 30%

	10%		15%		20%		25%		30%	
Method	m	s	m	s	m	s	m	s	m	s
BPCA	0.082	0.214	0.124	1.087	0.146	0.541	0.177	0.675	0.197	0.986
EM	0.046	0.064	0.049	0.068	0.049	0.069	0.051	0.070	0.053	0.073
Hot-deck	0.026	0.056	0.031	0.066	0.039	0.078	0.049	0.090	0.064	0.102
Inpaint	1.410	2.328	1.491	2.604	1.569	2.880	1.642	3.075	1.731	3.352
k-NN	0.024	0.042	0.027	0.049	0.031	0.054	0.036	0.060	0.045	0.067
MEI	0.089	0.094	0.092	0.095	0.091	0.095	0.091	0.095	0.091	0.095
MICE	0.057	0.089	0.062	0.095	0.064	0.097	0.066	0.098	0.069	0.101
Fourier	0.025	0.102	0.025	0.072	0.027	0.134	0.028	0.124	0.030	0.134
Lk-NN	**0.019**	0.035	0.022	0.041	0.024	0.044	0.027	0.046	0.029	0.050
FLk-NN	0.020	0.080	**0.021**	0.050	**0.022**	0.084	**0.024**	0.086	**0.026**	0.101

(b) Missing ratios 35% - 50%

	35%		40%		45%		50%	
Method	m	s	m	s	m	s	m	s
BPCA	0.203	0.903	0.190	0.863	0.190	1.197	0.199	1.467
EM	0.055	0.074	0.057	0.075	0.060	0.076	0.062	0.078
Hot-Deck	0.078	0.115	0.091	0.122	0.101	0.127	0.110	0.132
Inpaint	1.798	3.535	1.852	3.671	1.950	3.968	2.017	4.120
k-NN	0.055	0.075	0.066	0.084	0.076	0.091	0.084	0.098
MEI	0.091	0.096	0.091	0.095	0.091	0.095	0.091	0.095
MICE	0.073	0.103	0.076	0.105	0.080	0.107	0.084	0.109
Fourier	0.030	0.119	0.032	0.133	0.035	0.161	0.040	0.196
Lk-NN	0.032	0.052	0.035	0.055	0.039	0.059	0.044	0.065
FLk-NN	**0.027**	0.077	**0.030**	0.089	**0.033**	0.110	**0.038**	0.146

MICE [18]: As a multiple imputation method we used MICE, which employs chained equation to impute. We used the mice R package[5] with all parameters set to their defaults.

FLk-NN: We used $k = 5$ since it gives the highest accuracy, $D = 60$ (i.e. 1 hour) as most of the biological effects will occur within one hour, and $p = 3$ to enable multiple lags without drastically increasing computational complexity.

We used the authors' code for each algorithm when available and implemented MEI, hot deck, and k-NN ourselves.

We evaluate the performance of each approach based on how close the imputed values are to the actual values, using the mean absolute error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |d_i^{ac} - d_i^{imp}| \tag{10}$$

where n is the number of missing data points, and d_i^{ac} and d_i^{imp} are the i^{th} normalized actual and normalized imputed values respectively. For normalization, we used min-max normalization for each variable, ignoring the missing values from the actual data. MAE is computed for each subject individually (10 for DSIM, 9 for NICU), and then the average and standard deviation are calculated.

[5]http://cran.r-project.org/web/packages/mice/index.html

C. Results

DSIM: Table IV shows the mean and standard deviation of the MAE for each method, with the lowest error rates highlighted. For all missing ratios our combined method, FLk-NN, gives the lowest average MAE with a small standard deviation. Further, Lk-NN has lowest MAE for the 5% missing ratio and is ranked second for all other ratios. Figure 7 shows the number of times each method gives the highest imputation accuracy out of the 100 total datasets, with FLk-NN yielding the highest accuracy in 89 cases and Lk-NN the highest in the other 11 cases. Thus, including lagged correlations in k-NN allows more accurate imputation of missing values when data have temporal correlations and there are potentially significant amounts of missing data. Other methods did not consider time lagged correlation at all and thus their imputation accuracy is lower than the proposed method.

Among the existing methods, k-NN and BPCA had better results for lower missing ratios but their accuracy decreases significantly as the missing ratio increases. On the other hand, EM was less accurate for lower missing ratios but the accuracy did not decrease as significantly as the missing ratio increased and in fact it gave better accuracy than k-NN and BPCA for higher missing ratios.

Note that the accuracy of the combined approach, FLk-NN, is higher than the individual approaches, Fourier and

Lk-NN, for every missing ratio since the combined approach includes relationships within and across variables, and the DSIM data has auto-correlations with lagged correlations, as shown in figure 5. For example, in figure 5, liquid glucose in the stomach (Q_{sto2}) depends on Q_{sto1} and itself.

Figure 6 shows the maximum number of consecutive occurrences of missing values (i.e. gap of values within observed values) where DSIM has a maximum gap length of 17. Large gaps have an impact on Fourier but less influence on Lk-NN, which uses lagged correlations with other variables and leads to better results when the methods are combined.

Our Lk-NN can impute if some of the variables are missing in test vector cannot if all the lagged values are missing (e.g. a subject wearing sensors went out of network coverage for a longer period of time) whereas the Fourier method can impute in this situation. On the other hand, Fourier cannot impute missing values that occur before the first observed value (e.g. due to starting delay of a device) while Lk-NN can handle this. Across the DSIM datasets an average of 1.27% of missing values could not be imputed by Lk-NN, while FLk-NN imputed all missing values.

A two tailed un-paired t-test (for unequal variance) found that for all missing ratios, the MAE of FLk-NN is significantly different from that of other methods ($p < 0.0009$) besides Lk-NN. FLk-NN and Lk-NN are significantly different for 20% to 50% ($p < 0.000004$) and 15% ($p < 0.041$) missing ratios, but not for 5% and 10% using the threshold $p < 0.05$.

NICU: For this dataset, we compute MAE for the simulated missing data points only. Table V shows the mean and standard deviation of MAEs of NICU. The best mean values for each missing ratio are highlighted in bold. Our proposed methods out-performed all other methods, where Lk-NN has lowest mean MAE for the 10% missing ratio and FLk-NN was best for all other missing ratios. Figure 7 shows the number of times each method gives the highest imputation accuracy for this dataset. FLk-NN has highest proportion (39 out of 81), with Lk-NN being second (21 of 81), and Fourier third (11 of 81).

Compared with the DSIM dataset, the accuracy of many other methods such as BPCA deteriorated significantly due to the increased amount of non randomly-generated missing values whereas our method's accuracy improved. k-NN and EM had the best accuracy of the existing methods but their accuracy drops significantly as the amount of missing data increases, while FLk-NN showed a more gradual decrease in accuracy as the ratio increased.

For the NICU dataset, Lk-NN could not impute an average of 1.71% of missing values, and for k-NN the amount is 1.99%, while FLk-NN imputed all missing values. The p-value of the difference between our approach and the others using an unpaired t-test was significant for all methods from 15% to 50% missing ratios ($p < 0.0005$). For the 10% missing ratio, all other methods are significantly different ($p < 0.0007$) except Lk-NN.

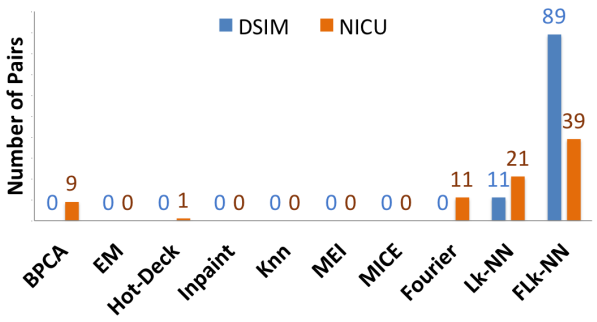


Figure 7: Dataset missing-ratio pair comparison.

D. Imputation with missing rows

One of the key benefits of our proposed approach is that the combined method enables imputation when an entire row is missing, meaning that all variables at a particular time are missing. This is a realistic challenge with biomedical data where measurements may come from a single device or there's a loss in connectivity preventing recording.

To evaluate this, we created another simulated missing dataset using the DSIM data. Here for each subject, 10% of rows were deleted. All imputation methods were applied and evaluated using the same approach as described earlier. Note that BPCA, hot deck, and k-NN cannot impute at all in this case. For Lk-NN, though the time instances are fully missing for a missing value, the test vector may not be empty because of the use of time lags, where the lagged values may be present. However, this did not occur and Lk-NN was able to impute all missing values.

The MAE for the remaining methods across the 10 datasets is shown in Table VI, which shows that our proposed method, FLk-NN, has the highest accuracy and lowest standard deviation. Lk-NN and Fourier were second and third respectively. A t-test shows that the MAE of FLk-NN is significantly different from that of other methods ($p < 0.0051$) other than Lk-NN. Note that the accuracy of EM and MEI is the same here since EM first initializes missing values using MEI and then optimizes those values, but in this situation it did not optimize.

V. Conclusion

Missing values are common in big data, where often many variables have correlations across time. Further, these data are rarely missing completely at random, especially when multiple signals are collected from a single device that may face errors or malfunction. At the same time, current guidelines for biomedical research recommend using complete datasets and imputing missing values, making

Table VI: Mean and Standard deviation of MAE for simulated data (DSIM) where 10% of rows are missing, meaning all variables are absent for the missing instances. Methods that cannot handle such cases are indicated with a dash.

Method	Mean	Standard Deviation
BPCA	-	-
EM	0.182	0.146
Hot deck	-	-
Inpaint	28.39	136.52
k-NN	-	-
MEI	0.182	0.146
MICE	0.206	0.190
Fourier	0.049	0.072
Lk-NN	0.045	0.061
FLk-NN	**0.043**	0.060

accurate imputation in these data a priority [30]. Here we propose a novel imputation method that incorporates varying time lags between correlated variables and auto-correlations within the variables. The main contributions of this paper are two-fold: i) it incorporates time lagged correlations between the variables during imputation and ii) it can handle multiple types of missingness occurring in a single variable, whereas existing methods cannot handle these cases. Moreover, the proposed system is able to impute with high accuracy in the case of empty instances while some of the state-of-the-art methods cannot impute values at all. The system obtained the best accuracy in terms of MAE for both simulated and real world biological datasets and outperformed other benchmark methods. Experimental results show that the system can impute plausible data even if 50% of a dataset is missing with many consecutively missing values and in the presence of fully empty instances in the data.

ACKNOWLEDGMENT

This work was supported in part by the NLM of the NIH under Award Number R01LM011826.

REFERENCES

[1] G. Molenberghs and M. G. Kenward, *Missing Data in Clinical Studies*. Wiley, 2007.

[2] M. Hua and J. Pei, "Cleaning disguised missing data: A heuristic approach," in *ACM SIGKDD*, 2007, pp. 950–958.

[3] B. Marlin, R. Zemel, S. Roweis, and M. Slaney, "Recommender systems: Missing data and statistical model estimation," in *IJCAI*, 2011, pp. 2686–2691.

[4] Y. He, "Missing data analysis using multiple imputation: Getting to the heart of the matter," *Circulation: Cardiovascular Quality and Outcomes*, vol. 3, no. 1, pp. 98–105, 2010.

[5] A. Johansson and M. Karlsson, "Comparison of methods for handling missing covariate data," *AAPS Journal*, vol. 15, no. 4, pp. 1232–1241, 2013.

[6] C. Bishop, *Pattern Recognition and Machine Learning*. Wiley, 2002.

[7] P. D. Allison, *Missing Data*. Sage Publication, 2001.

[8] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[9] T. Yu, H. Peng, and W. Sun, "Incorporating nonlinear relationships in microarray missing value imputation," *IEEE/ACM Trans Comput Biol Bioinfo*, vol. 8, no. 3, pp. 723–731, 2011.

[10] S. Zhang, Z. Jin, and X. Zhu, "Missing data imputation by utilizing information within incomplete instances," *J Sys Soft*, vol. 84, no. 3, pp. 452 – 459, 2011.

[11] F. V. Nelwamondo, D. Golding, and T. Marwala, "A dynamic programming approach to missing data estimation using neural networks," *Info Sci*, vol. 237, pp. 49 – 58, 2013.

[12] S. Oba, M. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii, "A bayesian missing value estimation method for gene expression profile data," *Bioinformatics*, vol. 19, no. 16, pp. 2088–2096, 2003.

[13] X. Su, K. T.M., and G. R., "Using imputation techniques to help learn accurate classifiers," in *IEEE ICTAI*, Nov 2008, pp. 437–444.

[14] S. Gunnemann, E. Muller, S. Raubach, and T. Seidl, "Flexible fault tolerant subspace clustering for data with missing values," in *ICDM*, Dec 2011, pp. 231–240.

[15] M. Ouyang, W. Welsh, and P. Georgopoulos, "Gaussian mixture clustering and imputation of microarray data," *Bioinformatics*, vol. 20, no. 6, pp. 917–923, 2004.

[16] K. J. Nishanth, V. Ravi, N. Ankaiah, and I. Bose, "Soft computing based imputation and hybrid data and text mining: The case of predicting the severity of phishing alerts," *Expert Sys. Appl*, vol. 39, no. 12, pp. 10 583 – 10 589, 2012.

[17] D. Rubin, *Multiple imputation for nonresponse in surveys*. Wiley, 1987.

[18] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *J. Stat Soft*, vol. 45, no. 3, pp. 1–67, 2011.

[19] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: Issues and guidance for practice," *Stat Med*, vol. 30, no. 4, pp. 377–399, 2011.

[20] D. Bouhlila and F. Sellaouti, "Multiple imputation using chained equations for missing data in timss: a case study," *Large-scale Assess. Edu.*, vol. 1, no. 1, pp. 1–4, 2013.

[21] L. Nanni, A. Lumini, and S. Brahnam, "A classifier ensemble approach for the missing feature problem," *Artificial Intell Med*, vol. 55, no. 1, pp. 37 – 50, 2012.

[22] M. Schomaker and C. Heumann, "Model selection and model averaging after multiple imputation," *Comput StatData Anal*, vol. 71, pp. 758 – 770, 2014.

[23] A. Farhangfar, L. Kurgan, and W. Pedrycz, "A novel framework for imputation of missing values in databases," *IEEE Trans. Syst. Man Cybern. A., Syst. Humans*, vol. 37, no. 5, pp. 692–709, 2007.

[24] C. Chatfield, *The Analysis of Time Series, An Introduction*. New York: Chapman & Hall, 2004.

[25] C. Dalla Man, M. D. Breton, and C. Cobelli, "Physical activity into the meal glucose-insulin model of type 1 diabetes: in silico studies." *J Diabetes Sci Technol*, vol. 3, no. 1, pp. 56–67, Jan 2009.

[26] B. Kovatchev, M. Breton, C. Dalla Man, and C. Cobelli, "In silico preclinical trials: A proof of concept in closed-loop control of type 1 diabetes," *J Diabetes Sci. Tech.*, vol. 3, no. 1, pp. 44–55, 2009.

[27] J. Claassen, A. Perotte, D. Albers, S. Kleinberg, J. M. Schmidt, B. Tu, N. Badjatia, H. Lantigua, L. J. Hirsch, S. A. Mayer, E. S. Connolly, and G. Hripcsak, "Nonconvulsive seizures after subarachnoid hemorrhage: Multimodal detection and outcomes," *Annals of Neurology*, vol. 74, no. 1, pp. 53–64, 2013.

[28] T. Schneider, "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values," *Journal of Climate*, vol. 14, no. 5, pp. 853–871, 2001.

[29] D. Garcia, "Robust smoothing of gridded data in one and higher dimensions with missing values," *Comput. Stat Data Anal*, vol. 54, no. 4, pp. 1167 – 1178, 2010.

[30] T. Li, S. Hutfless, D. O. Scharfstein, M. J. Daniels, J. W. Hogan, R. J. Little, J. A. Roy, A. H. Law, and K. Dickersin, "Standards should be applied in the prevention and handling of missing data for patient-centered outcomes research: a systematic review and expert consensus," *Journal of clinical epidemiology*, vol. 67, no. 1, pp. 15–32, 2014.