FOSA: Full Information Maximum Likelihood (FIML) Optimized Self-Attention Imputation for Missing Data

Ou Deng
Human Sciences
Waseda University
dengou@toki.waseda.jp

Qun Jin
Human Sciences
Waseda University
jin@waseda.jp

ABSTRACT

In data imputation, effectively addressing missing values is pivotal, especially in intricate datasets. This paper delves into the FIML Optimized Self-attention (FOSA) framework, an innovative approach that amalgamates the strengths of Full Information Maximum Likelihood (FIML) estimation with the capabilities of self-attention neural networks. Our methodology commences with an initial estimation of missing values via FIML, subsequently refining these estimates leveraging the self-attention mechanism. Our comprehensive experiments on both simulated and real-world datasets underscore FOSA's pronounced advantages over traditional FIML techniques, encapsulating facets of accuracy, computational efficiency, and adaptability to diverse data structures. Intriguingly, even in scenarios where the Structural Equation Model (SEM) might be mis-specified, leading to suboptimal FIML estimates, the robust architecture of FOSA's self-attention component adeptly rectifies and optimizes the imputation outcomes. Our empirical tests reveal that FOSA consistently delivers commendable predictions, even in the face of up to 40% random missingness, highlighting its robustness and potential for wide-scale applications in data imputation.

***K*eywords** Full information maximum likelihood · Self-attention · Deep learning · Missing data

1 Introduction

In data analysis and modeling, missing data consistently presents a formidable challenge. Basic techniques often resort to simple imputations like mean substitution or directly omit incomplete data points. Sadly, these strategies can introduce bias, reduce precision, and sometimes lead to incorrect conclusions [1, 2, 3]. Against this backdrop, various sophisticated methods have been developed. Notably, the Maximum Likelihood Estimation (MI) lineage stands out, with Full Information Maximum Likelihood (FIML) as a prime example.

FIML estimation is a principled approach that utilizes all available information in a dataset for parameter estimation, even with missing values. Unlike simpler imputation methods, FIML tackles missingness head-on during parameter estimation, ensuring that data complexities are retained. Experimental evidence suggests that FIML sets a robust foundation for further modeling.

Conversely, the self-attention mechanism, pivotal in modern neural network designs, excels in capturing long-term data dependencies. By weighting data points differently, it highlights the most relevant relationships, enabling a detailed understanding of complex datasets.

In this paper, we present FOSA, or FIML Optimized Self-Attention. This innovative method merges FIML estimation with self-attention. After initializing missing data with column means and refining these with Structural Equation Modeling (SEM) and FIML, we obtain refined input for the neural model. The self-attention mechanism then processes this input, detecting subtle data dependencies for improved predictions.

Combining FIML and self-attention offers dual benefits. While FIML ensures the retention of essential information despite missingness, the self-attention mechanism delves into the data, identifying and exploiting patterns, particularly

in complex multivariate time-series datasets. Collectively, they forge a powerful framework for managing datasets with missing values and predicting outcomes.

In the following sections, we explore FOSA in detail, shedding light on its theoretical foundations, practical application, and advantages over traditional methods. Through this, we emphasize the potential of merging classical statistical techniques with advanced deep learning to tackle persistent data analysis challenges.

The source code for the experiments is publicly available at: `https://github.com/oudeng/FOSA/`.

2 Related Work

Missing data imputation has been an enduring challenge, driving comprehensive academic investigations. Rubin's seminal work in 1976 provided a cornerstone for missing data classification, distinguishing between missing completely at random (MCAR) and missing at random (MAR) [4]. Since most real-world data aligns with the MAR category, it has naturally drawn extensive research focus.

FIML estimation, emerging as a front-runner in this domain, has been lauded for its robustness and precision. Key works by Enders and Bandalos explored FIML's prowess within structural equation modeling [1]. Their empirical findings vouched for FIML's unbiased nature, especially under ignorable missing data conditions. Comprehensive reviews and insights from Schafer and Graham further reinforced FIML's superiority over traditional methods [2]. Lazar's, Fielding et al.'s, and Graham's subsequent investigations continued to underscore FIML's robust adaptability and practical implementations [5, 6, 7].

Recent literature, steered by researchers like Baraldi and Enders, White et al., Dong and Peng, Mazza, and Enders and Nelson, has widened the discourse on missing data analyses, encompassing both FIML and multiple imputation techniques [8, 9, 10, 11, 12, 13]. These works have invariably echoed the theme of FIML's advantages over conventional imputation techniques.

With the advent of deep learning, there has been a shift towards harnessing these methodologies for missing data. Early in this paradigm shift, Stekhoven et al. introduced MissForest, a non-parametric approach rooted in random forests, demonstrating promising results on mixed-type data [14].

Deep learning methodologies for missing data imputation have primarily bifurcated into two trajectories. The first trajectory revolves around recurrent networks. This domain has seen significant contributions, including Che et al.'s approach for multivariate time series [15], Cao et al.'s bidirectional recurrent neural network-based methodology [16], and Chai et al.'s U-Net convolutional neural network-based approach [17].

The second trajectory is driven by the Generative Adversarial Networks (GAN) framework. GANs, inspired by their success in image generation, have been tailored to discern and impute missing values in multivariate time series datasets. Yoon et al.'s GAIN methodology and Luo et al.'s integration of Gate Recurrent Unit (GRU) within the GAN framework stand out as quintessential examples [18, 19].

Despite the promise of these advanced techniques, challenges persist. Deep learning methodologies, especially those rooted in recurrent networks, occasionally falter in capturing intricate multivariate relationships. GAN-based methods, demanding abundant high-quality data, often find their hands tied in specific scenarios. Recognizing this gap, we introduce the FOSA framework. FOSA, an innovative amalgamation of traditional FIML, represents a harmonious blend of FIML's robust foundational principles with the advanced capabilities of self-attention mechanisms, offering a comprehensive solution for the challenges of missing data imputation.

3 Methodology

We introduce FOSA as a two-step strategy that synergistically integrates the robustness of FIML estimation with the capabilities of the self-attention mechanism. This section elucidates the key components and principles underpinning FOSA.

3.1 Step-I: FIML Estimation

FIML operates by conducting a maximum likelihood estimation for nonlinear simultaneous equations. It comprises N equations, accounting for N endogenous variables. These equations can be represented implicitly. FIML shines in scenarios where error terms follow a normal distribution, making it one of the most effective techniques for models with non-linear coefficients.

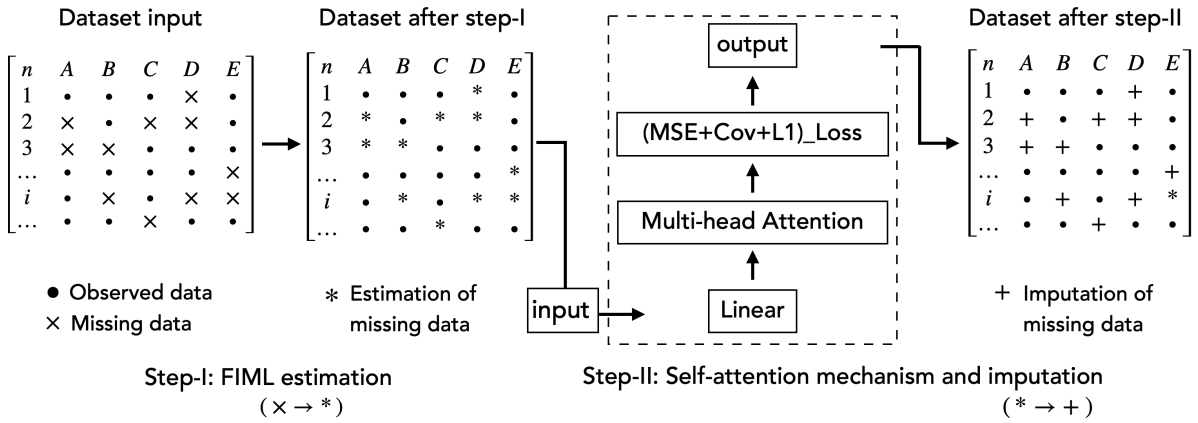

Figure 1: Overview of the FIML Optimized Self-attention (FOSA) framework. The process starts with an initial estimation of missing values using FIML. This estimation is then refined using a self-attention mechanism.

In our FOSA framework (Figure 1), FIML lays the groundwork, providing robust parameter estimates for the dataset Y. The likelihood function $L(\theta; Y)$ for a parameter set θ is:

$$L(\theta; Y) = \prod_{i=1}^{n} f(y_i; \theta) \tag{1}$$

Where, n presents the number of observations and $f(y_i; \theta)$ denotes the probability density function of the observed data y_i for the given parameters θ. The goal of FIML is to determine the parameter values θ^* that maximize this function:

$$\theta^* = \arg\max_{\theta} L(\theta; Y) \tag{2}$$

3.2 Multihead Self-attention Structure

The self-attention model within FOSA gracefully amalgamates model complexity and computational efficiency. Inputs are first transformed into tensors suitable for training and testing. For each attention head, the mechanism can be summarized by:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{3}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors. In the context of self-attention, $Q = K = V = X$. Within this module, each feature is compared with every other feature to discern its interrelations. This is achieved by computing the similarity between each feature (termed as "query") and all other features (referred to as "keys"). Subsequently, these similarity scores are employed to weigh the input features (designated as "values"), resulting in a new representation that considers the interrelationships among all features.

The outputs from all heads are concatenated and linearly transformed to produce the attention output. Thus the entire process can be summarized as:

$$Z = \text{Linear}(\text{Dropout}(\text{ReLU}(\text{MultiheadAttention}(X)))) \tag{4}$$

3.3 Step-II: Self-attention Mechanism

Our tailored loss function combines Mean Squared Error (MSE) with the covariance matrix, guiding the model toward accurate predictions that respect the dataset's covariance structure. The MSE, $\text{MSE}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, serves as the primary loss function. The covariance loss ensures alignment between the predicted and target values' covariance matrices:

$$\begin{aligned}\text{cov\_loss} &= \|\text{cov}(\hat{y}) - \text{cov}(y)\|_F \\ &= \left\| \frac{\text{predicted}^\top \cdot \text{predicted}}{n} - \frac{\text{target}^\top \cdot \text{target}}{n} \right\|_F \end{aligned} \tag{5}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of the difference between the covariance matrices of predictions and targets[1].

Incorporating L1 regularization prevents overfitting, represented as $\text{L1\_loss} = \lambda \sum_{i=1}^{n} |w_i|$, where λ is the regularization coefficient and w_i represents the model weights. The composite loss ensures predictions are accurate while maintaining the original data's covariance structure:

$$\text{Total\_loss} = \text{MSE\_loss} + \text{cov\_loss} + \text{L1\_loss} \tag{6}$$

3.4 Summary of FOSA Methodology

The self-attention model's output, depicted in Figure 1, is a refined representation considering the interrelationships among input features. This output can serve various purposes, including predicting missing values. FOSA seamlessly bridges FIML with advanced neural components. With FIML's robustness, parameter estimates remain unbiased even in the presence of missing data. Coupled with self-attention mechanisms, FOSA captures intricate data relationships, providing a holistic approach to missing data.

4 Experiment

We conducted experiments on two datasets. The first one is simulated and designed under the premise of established causal relationships among variables. The second one is derived from real-world observations where the causal relationships among variables are unknown or not clear enough.

For the first dataset of Experiment-I, our primary objective was to evaluate the theoretical imputation performance of FOSA. We crafted a dataset encompassing five variables, with their inter-variable causal relationships depicted in Figure 2. This design allows us to assess the improvements in the prediction accuracy of FOSA compared to the baseline FIML method under identical conditions. Furthermore, we can conveniently alter the causal relationships among variables and the data generation process. This flexibility enables us to scrutinize the performance of FOSA under a myriad of potential causal relationships and various primary relationships between variables, be they linear or nonlinear.

The second dataset of Experiment-II is sourced from the publicly available Personal Key Indicators of Heart Disease dataset on the Kaggle site [2], referred to as the CDC dataset. The primary aim of employing this real-world dataset is to gauge the imputation performance of FOSA in the context of health-related data. After data preprocessing, this dataset is cleaned, without missing values. To evaluate the capabilities of FOSA, we artificially introduced missing values in various patterns. This approach allows us to compare the performance of FOSA against the standalone FIML method in terms of imputing missing values.

In practical data processing, we place greater emphasis on the model's performance on the test set, as it more accurately reflects the model's generalization capability on unseen data. The experimental results presented in this paper, specifically the final predictions for missing values, are all based on the mentioned test datasets. Furthermore, the self-attention model used in the experiment was configured with the following primary hyper-parameters: Multi-head(4), Hidden layer dimension(64), Number of epochs(200), Dropout rate(0.5), and Learning rate(0.001).

4.1 Experiment-I on Simulated Dataset

As depicted in Figure 2 and Table 1, we assess the performance of both FIML and FOSA models across a spectrum of data characteristics. This is achieved by modulating the functions f_C, f_D, and f_E in patterns (a) and (b). Specifically,

[1]The Frobenius norm is a variant of matrix norms, employed to measure the magnitude of a matrix. For any given matrix, its Frobenius norm is defined as the square root of the sum of squares of all its entries. When referring to the Frobenius norm of the difference between the covariance matrices of predictions and targets, we are essentially quantifying the degree of disparity between the two covariance matrices.

[2]The CDC refers to the Centers for Disease Control and Prevention of the United States. The dataset in question is titled "2020 annual CDC survey data of 400k adults related to their health status" and can be accessed at `https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease`.

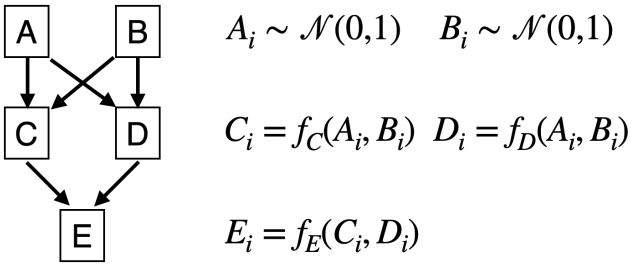

Figure 2: Experiment-I: design of simulated dataset. Based on the causality defined in the experimental design, a simulated dataset is generated with five variables, each containing a mount of (e.g. 500) data points. Variables A and B are independent random variables. Variables C and D act as intermediary variables influenced by A and B, with their causal relationships being represented by functions f_C and f_D respectively. The variable E serves as the target variable, determined by variables C and D.

Table 1: Adjusting the data characteristics by f_C, f_D and f_E

Variables	Pattern (a): strong linear	Pattern (b): strong non-linear
C_i	$A_i + B_i$	$A_i + 2B_i^2$
D_i	$A_i - B_i$	$A_i \times B_i$
E_i	$2C_i + 3D_i + \epsilon_i$	
Note:	The noise $\epsilon_i \sim \mathcal{N}(0, 0.5)$	
	Initialize the SEM of FIML as $E \sim A + B + C + D$	

pattern (a) embodies pronounced linear attributes, while pattern (b) accentuates robust non-linear traits. Concurrently, we maintain a consistent f_E to ascertain the differential outcomes on the target variable E_i.

In our experimental evaluation, we employed the following metrics to rigorously compare the imputation capabilities of FIML and FOSA on an identical dataset. The Kullback-Leibler (KL) Divergence serves as a measure to quantify the divergence between two probability distributions. A smaller KL Divergence indicates that the predicted distribution closely aligns with the true distribution, with its optimal value being zero. The Mean Squared Error (MSE) gauges the discrepancy between predicted and actual values, where a lower MSE signifies a reduced difference between these values. The R^2 metric, ranging between 0 and 1, assesses the predictive power of a model, with higher values denoting superior predictive performance. The Mean Absolute Percentage Error (MAPE) quantifies the relative prediction error as a percentage of the actual value, and a diminished MAPE suggests a smaller prediction error[3]. The adoption of a diverse set of evaluation metrics ensures a more objective and comprehensive assessment of model performance. Furthermore, these metrics were consistently applied in our evaluations of Experiment-II CDC datasets.

In evaluating the FIML and FOSA imputation techniques, Table 2 offers a thorough comparison across different variables within two distinct data patterns: the linear Pattern (a) and the non-linear Pattern (b).

KL Divergence: Serving as a metric to gauge the divergence between true and predicted probability distributions, FOSA frequently matches or surpasses FIML, hinting at FOSA's proficiency in rendering an accurate probabilistic portrayal of missing data.

MSE: FOSA consistently excels in terms of Mean Squared Error across both patterns, particularly for variables C, D, and E.

R^2: FOSA typically achieves higher R^2 values, highlighting its prowess in capturing data variance. Notably, in the context of variable E, while FIML struggles, FOSA's performance is nearly flawless for linear data and remains commendable for non-linear datasets.

MAPE: FOSA frequently outperforms FIML, especially for non-linear Pattern (b) variables A and D.

[3]The evaluation metrics selected are as follows:
Kullback-Leibler divergence (KL Divergence) $= D_{KL}(P \,||\, Q) = \sum_i P(i) \log(\frac{P(i)}{Q(i)})$, where $P(i)$ is the probability of the i-th event in distribution P, and $Q(i)$ is its probability in distribution Q.
Mean Squared Error (MSE) $= \frac{1}{n}\sum(y_i - \hat{y}_i)^2$, Mean Absolute Percentage Error (MAPE) $= \frac{100\%}{n}\sum\left|\frac{y_i - \hat{y}_i}{y_i}\right|$,
Coefficient of Determination$(R^2) = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$, where y_i represents the observed data, $\bar{y}_i$ denotes the mean of the observed data, $\hat{y}_i$ signifies the predicted data, and n is the total number of data points.

Table 2: Experiment-I, FIML and FOSA Imputation Evaluation on Missing Data

Pattern (a)	KL Divergences		MSE		R^2		MAPE(%)	
Variables *	FIML	FOSA	FIML	FOSA	FIML	FOSA	FIML	FOSA
A	178.87	198.373	0.088	0.142	0.896	0.831	84.326	138.772
B	153.057	127.454	1.433	0.397	-0.213	0.664	174.755	71.068
C	165.099	164.813	1.600	0.410	-0.315	0.663	247.087	73.174
D	125.775	106.01	3.093	0.739	-0.457	0.652	337.446	168.151
E	9.646	23.575	22.754	1.325	0.245	0.956	84.817	64.227
Pattern (b)	KL Divergences		MSE		R^2		MAPE(%)	
Variables *	FIML	FOSA	FIML	FOSA	FIML	FOSA	FIML	FOSA
A	246.98	124.479	1.166	0.578	-0.386	0.313	267.299	190.947
B	115.187	161.53	1.538	1.008	-0.303	0.146	174.490	115.923
C	61.573	53.135	12.035	2.432	-0.309	0.735	142.867	106.062
D	297.216	276.978	1.395	0.731	-2.762	-0.971	829.412	243.778
E	18.484	20.827	23.337	13.031	-0.208	0.325	146.585	223.259

Note: (*) Pattern (a) with the strong linear characteristics, and Pattern (b), non-linear characteristics, which is described in Table 1.

Table 3: FIML and FOSA imputation evaluation on missing data of Experiment-II.

CDC dataset	KL Divergences		MSE		R^2		MAPE(%)	
Variables *	FIML	FOSA	FIML	FOSA	FIML	FOSA	FIML	FOSA
HeartDisease	471.026	492.835	0.703	0.252	-7.808	-2.151	inf**	inf
BMI	12.297	1.029	811.821	196.25	-20.732	-4.254	99.929	39.373
PhysicalHealth	-2.521	8.223	73.123	72.179	-0.154	-0.139	inf	inf
MentalHealth	-1.579	1.861	79.678	74.616	-0.214	-0.137	inf	inf
SleepTime	0.362	9.007	53.144	11.050	-22.369	-3.859	100.000	36.629

Note: (*) The first 50,000 entries of the total 320,000 in the CDC dataset. The missing rate takes 40%.
(**) The 'inf' values in the MAPE column arise due to the division by zero error when the true value is zero. This is the one inherent limitation of MAPE.

Pattern (a) - Linear Data: Variables like B and E particularly benefit from FOSA, outstripping FIML in terms of MSE and KL Divergence.

Pattern (b) - Non-linear Data: FOSA's edge is accentuated, especially for variables A and C, where it significantly eclipses FIML's MSE. This indicates FOSA's robustness and adaptability in capturing complex, non-linear relationships in the data.

Variable E, as depicted in Figure 2, stands out due to its complex causal relationships. Despite initializing the FIML-estimated model with an $E \sim A + B + C + D$ SEM structure, FIML's imputation fell short. However, the self-attention mechanism in FOSA, adept at deciphering intricate relationships, considerably improved the imputation. This enhancement is lucidly portrayed in Figure 3.

In summary, while both FIML and FOSA showcase their merits, FOSA consistently stands out across various metrics and data patterns, emphasizing its versatility in addressing both linear and intricate non-linear data patterns.

4.2 Experiment-II on Actual Dataset

In Experiment-II, we employed the identical experimental code, encompassing both the model architecture and its parameters, with the sole modification being the substitution of the simulated dataset with the CDC dataset. Contrary to the former, the causal relationships between variables in the latter remain elusive. Endeavoring under the premise of solely observed data, we leveraged FOSA to discern inter-variable relationships and impute missing values. These missing entries were pre-emptively excised from the comprehensive dataset, enabling us to juxtapose the imputed results against the actual values. To facilitate a comparison with Experiment-I, we meticulously selected numerical attributes from the CDC dataset, namely: 'HeartDisease', 'BMI', 'PhysicalHealth', 'MentalHealth', and 'SleepTime'. For non-numerical attributes, we employed a One-Hot encoding approach.

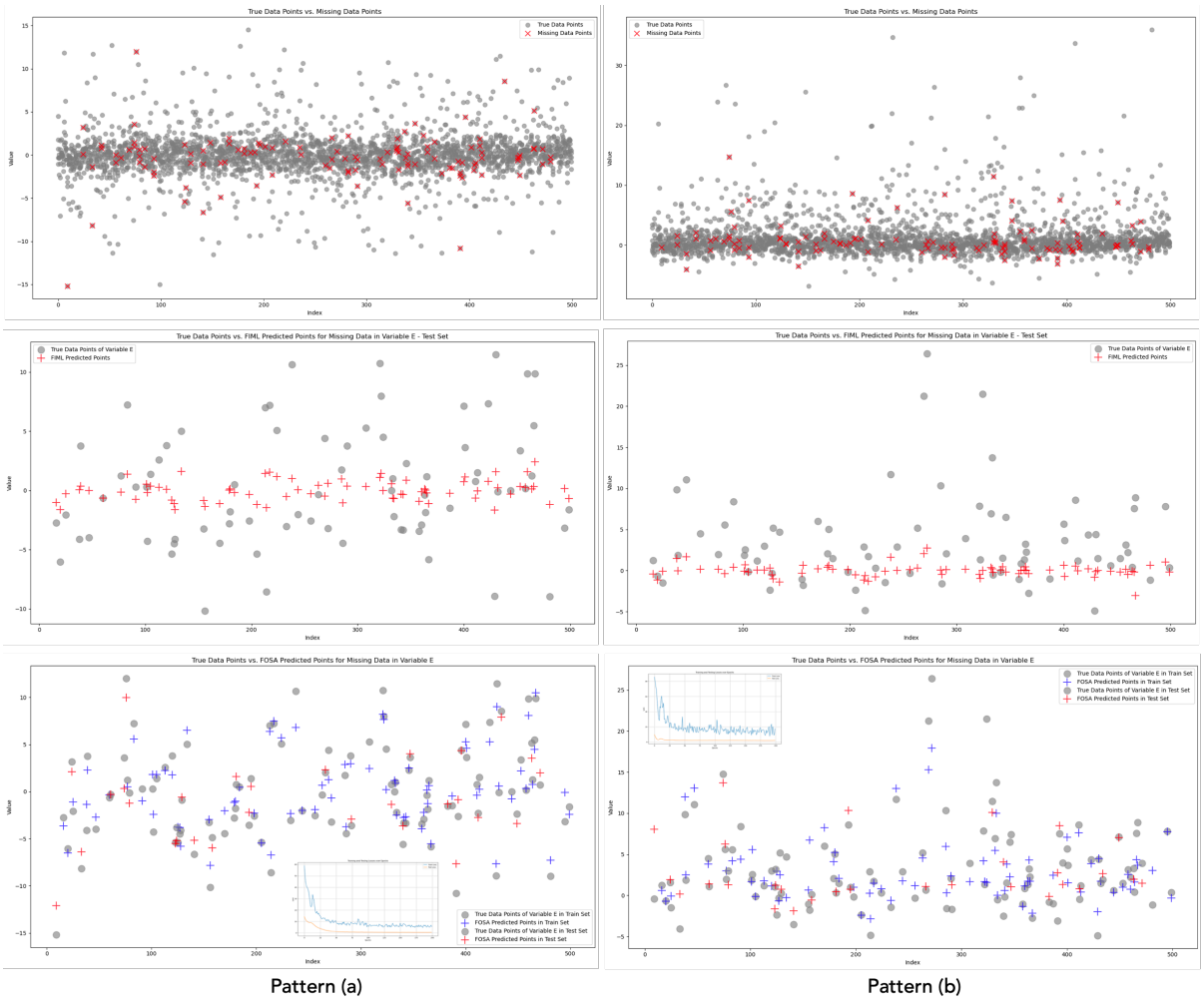

Figure 3: Imputation performance comparison of FIML and FOSA on both Pattern (a) and (b) for variable E missing data. In the top row, we present the genuine data from the simulated dataset alongside the intentionally designed missing values, which account for approximately 20%. The middle row showcases the results of the FIML estimation for the missing values of variable E in Step-I. It is evident that the FIML estimations are suboptimal for both Pattern (a) and Pattern (b). The bottom row delineates the predictions for the missing values of variable E post-training with the Self-attention mechanism, juxtaposed with the actual values. In comparison to the FIML estimations, there is a marked enhancement in the outcomes.

We discern notable disparities in the imputation outcomes between FIML and FOSA, shown in Table 3. Across all variables, FOSA consistently outperforms FIML in terms of the metrics provided. This is evident from the lower KL Divergences, reduced MSE, improved R^2 values, and in most cases, lower MAPE percentages when using FOSA compared to FIML.

KL Divergence: For the variables HeartDisease, BMI, and SleepTime, FOSA has a higher KL Divergence than FIML, indicating a greater divergence from the true distribution. However, for PhysicalHealth and MentalHealth, FOSA exhibits a more favorable KL Divergence, suggesting a closer alignment with the true distribution.

MSE: The MSE for BMI under FIML is notably high at 811.821. This suggests that the FIML method may not be suitable for predicting missing values for BMI. In contrast, FOSA significantly reduces this error to 196.25, showcasing its superior performance.

$\mathbf{R^2}$: Negative R^2 values are unusual and typically indicate that the model is a poor fit for the data. The negative R^2 values across both methods, especially for HeartDisease and SleepTime, suggest that the models might not be capturing the underlying data patterns effectively. However, it's worth noting that FOSA consistently improves the R^2 values relative to FIML, indicating a better fit to the data.

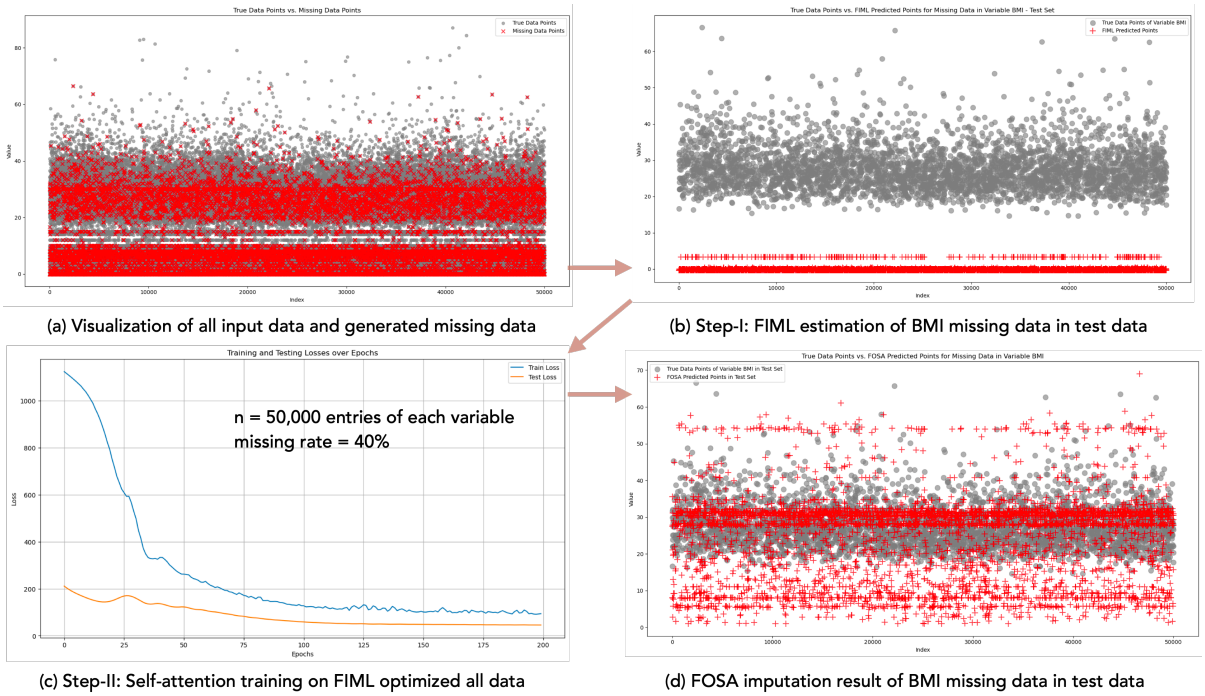

Figure 4: Experiment-II processes and BMI missing data imputation comparison (FIML vs FOSA). Subfigures (a) through (d) provide a detailed depiction of the FOSA imputation process for missing BMI values. Figure (a) displays all variables, including BMI, and their respective missing value statuses. The missing values, represented by red 'X' markers, account for 40% of the actual values, depicted as gray dots. The visualization reveals intricate non-linear interrelationships (causal relationships) among the variables. Figure (b) presents the preliminary imputation of BMI's missing values using FIML. As can be discerned from the metric evaluations in Table 3, this initial estimation is not at an optimal level. Figure (c) showcases the training status of the self-attention mechanism across the entire input dataset, encompassing both observed values and FIML-imputed values. Over 200 epochs, both training and testing losses exhibit a consistent decline. Figure (d) represents the final output for missing BMI values from the FOSA framework. Despite the suboptimal preliminary imputation results from Step-I using FIML, FOSA, leveraging the self-attention mechanism's capability to learn complex data relationships, produces results within our acceptable threshold.

MAPE: The 'inf' values in the MAPE column arise due to the division by zero error when the true value is zero. This limitation of MAPE is highlighted in the table's footnote. Excluding these 'inf' values, FOSA generally exhibits lower MAPE percentages than FIML, indicating more accurate predictions.

Variables' Characteristics: (HeartDisease) Given its binary nature (presence or absence), predicting its missing values can be challenging. The negative R^2 values indicate potential model misfit, but FOSA offers a relative improvement. (BMI) It's a continuous variable, and the substantial reduction in MSE by FOSA suggests its efficacy in handling such data. (PhysicalHealth and MentalHealth) These might be ordinal or categorical, and while both methods struggle, FOSA shows a relative advantage. (SleepTime) As a continuous variable representing hours, the drastic reduction in MSE by FOSA is noteworthy.

For datasets like the CDC, which consists of non-temporal single-entry collections, there is no causal relationship in terms of sequence between each entry. Consequently, rather than focusing on the absolute prediction error, we place greater emphasis on the relative prediction error. Specifically, we are interested in whether FOSA demonstrates an improvement in prediction metrics compared to the baseline method, FIML, and whether the KL divergence remains within a reasonable range when compared to the true values.

Upon visual inspection of the results, we observe that if the predictions of missing values by FOSA maintain a KL divergence below 1, an MSE below 200, an R^2 value not less than -5, and MAPE below 40% then the prediction accuracy meets the requirements for our subsequent specific research endeavors. From Table 3, examining the FOSA imputation results for missing BMI data, within this acceptable threshold, the visualization in Figure 4 (d) demonstrates that the imputed predictions and the actual values highly coincide in the regions of densest distribution. This essentially aligns with our probabilistic accuracy requirements for the specific research topic at hand. Additionally, we observe that due to the inherent bias in FIML's estimation for BMI data, FOSA also manifests a deviation towards lower values

compared to the true values. This serves as a cautionary note, emphasizing the need for heightened vigilance when processing subsequent data and drawing data inferences.

In conclusion, while both FIML and FOSA exhibit certain limitations in predicting missing values on the CDC dataset, FOSA consistently demonstrates a relative improvement across all metrics. This underscores its potential as a robust method for imputation in real-world datasets. In the subsequent ablation experiments, we will further elucidate this perspective. Such acceptability thresholds for prediction values often correspond to a data missingness level of approximately 40% at random.

5 Discussion

5.1 On the SEM Configuration for FIML

When leveraging FIML for estimation, a predefined structural equation model (SEM) for the variables is essential. In our Experiment-I, which used a simulated dataset, the SEM was deliberately designed as a core part of the experimental setup. This design allowed FIML to naturally exhibit its strengths, resulting in impressive performance. However, during Experiment-II, due to the undefined causal relationships between the variables, the effectiveness of FIML estimation might be compromised.

Interestingly, our observations suggest that even if FIML's estimations are not pinpoint accurate, the self-attention mechanism's capacity to understand complex relationships can counterbalance the adverse effects of an inaccurate SEM, enhancing the final predictions. We also investigated the consequences of intentionally altering the SEM input for FIML (e.g., replacing $E \sim A + B + C + D$ with $A \sim B + C + D + E$. This deliberate error mainly impacted the preliminary estimates in FIML's Step-I. However, during Step-II, the self-attention mechanism effectively corrected these erroneous inputs. The resulting predictions were strikingly similar to those obtained using the correct SEM inputs. This data-driven proficiency was especially evident in predictions for the missing values of the 'E' variable in Experiment-I and the 'BMI' and 'SleepTime' variables in Experiment-II.

These outcomes not only highlight the robust learning capabilities of self-attention neural networks but also hint at the potential for new methodologies in data causality discovery by examining the differences between FIML and FOSA outputs. Identifying discrepancies or inaccuracies from these output variations might provide insights into areas of prior knowledge that are lacking or incorrect. We aim to delve deeper into this promising area in our future research.

5.2 Recognition of Complex Data Patterns via the Self-attention Mechanism

The SelfAttentionModel class encapsulates the self-attention mechanism, primarily through its multihead-attention module, which discerns and captures interrelations among input features. The specific flow of this implementation is detailed below:

Linear Embedding: The journey commences with the input data passing through a shared linear layer. This layer transforms each feature from its original dimensionality to an embedding dimension designated as $hidden\_dim$. Such transformation facilitates the transition of raw data into a more conducive embedding or feature space, setting the stage for the subsequent self-attention mechanism (elaborated in Section 3.2). While this step doesn't directly unravel the relationships between features, it equips the self-attention mechanism with a richer representation of the data. To illustrate, consider our Experiment-I: each variable (e.g., A, B, C, D, E) constitutes a distinct feature of the input. Upon feeding into the model, these features collectively traverse the linear layer, ensuring a unified initial transformation across all variables.

Multihead-attention Processing: After undergoing linear embedding, the data is channeled into the multihead-attention module, marking its initiation into the core self-attention process.

Subsequent Processing: Once processed by the self-attention mechanism, the data navigates through a sequence of linear layers interspersed with nonlinear activation functions, culminating in the generation of the final output.

In essence, the multihead-attention module within the SelfAttentionModel class in the FOSA experimental framework is entrusted with the vital role of identifying and internalizing the interdependencies among input features. By virtue of the self-attention mechanism, the model acquires the capability to fathom and depict the nuanced patterns inherent in the input data.

Table 4: Ablation studies of FOSA imputation for missing BMI data across diverse data scales and missing rate patterns.

Round	n	Missing rate	KL Divergences	MSE	R^2	MAPE(%)
#1	1,000	20%	12.378	149.817	-3.067	30.699
	10,000		1.111	169.953	-3.220	33.864
	100,000		0.557	144.305	-2.553	30.760
	200,000		0.632	152.172	-2.982	32.146
	300,000		0.470	146.191	-2.597	30.342
#2	1,000	40%	9.537	182.366	-3.647	36.287
	10,000		1.003	206.117	-4.417	40.243
	100,000		0.629	195.875	-3.827	37.263
	200,000		0.455	195.896	-3.986	36.783
	300,000		0.930	190.974	-3.753	36.458
#3	1,000	60%	18.542	229.869	-5.565	42.542
	10,000		1.568	287.238	-6.671	50.139
	100,000		0.667	223.725	-4.597	45.388
	200,000		0.669	219.165	-4.400	43.458
	300,000		0.907	253.689	-5.268	47.545

Note: The parameter n denotes the number of input entries corresponding to each variable, namely HeartDisease, BMI, PhysicalHealth, MentalHealth, and SleepTime, as elaborated in Section 4.2.

5.3 Enhancing Interpretability of Evaluation Metrics and Results

In our research, we have adopted a holistic approach to evaluation metrics, striving for a broader perspective than many preceding studies. Our selection of metrics aims to navigate and counteract potential "cognitive pitfalls" linked to intricate causal relationships amongst variables. This approach sets the stage for in-depth causal investigations. We amalgamated traditional measures of predictive error, such as MSE, R-squared, and MAPE, with metrics like KL divergence that probe into differences between probability distributions.

It's crucial to understand that each metric addresses unique facets of the data, leading to the possibility of observing contrasting trends. For instance, there were scenarios where predictive error metrics showcased excellent results, while KL divergence indicated considerable differences. Let's delve into potential explanations:

Divergence between Predictive Error and Distributional Differences: Metrics like MSE, R-squared, and MAPE quantify the difference between predicted and actual values. For instance, if predictions for variable B are consistently closer to the actual values compared to those for variable A, these metrics would naturally indicate superior performance for B. In juxtaposition, KL divergence delves deeper, assessing disparities between the probability distributions of predictions and actual values. Therefore, even if on average predictions for B closely match the actual values, their respective probability distributions might differ significantly.

Interplay of Non-linearity and Outliers: Data characteristics such as non-linear relationships or the presence of outliers can have varied impacts on the metrics. For example, a model might exhibit impeccable predictions for the bulk of the dataset but struggle with outliers or specific regions, resulting in a satisfactory MSE but a discordant KL divergence.

Inherent Distributional Traits of Data: When the genuine data distribution is multimodal, even slight predictive errors can lead to a scenario where the predicted distribution diverges from the actual distribution, leading to increased KL divergence. This discrepancy is especially noticeable in data with multiple prominent patterns or peaks.

Model Assumptions: Underlying assumptions integrated within the model can also play a pivotal role in metric disparities.

To encapsulate, a multifaceted metric evaluation is indispensable for a comprehensive understanding of model performance. Our research accentuates not just metric-based assessments but also the power of visualization to render intricate relationships and findings with enhanced clarity.

5.4 Ablation Studies on Data Scale and Missing Rate

The performance of imputation methods can be influenced by factors such as the scale of data and the rate of missing values. These factors can subsequently affect prediction accuracy, computation time, and overall robustness of the

model. To delve deeper into these aspects, we conducted ablation experiments, the results of which are presented in Table 4.

Our first investigation focused on the scale of data. Using the CDC Dataset, which boasts nearly 320,000 entries, we established a suitable testing ground for scale assessments. We defined six scale tiers, ranging from 1,000 to 300,000 entries, by extracting subsets from the CDC dataset. While keeping the model parameters consistent, we observed the effects of the data scale on the evaluation metrics. Interestingly, the variations in scale had a minimal bearing on the metrics for missing value prediction. This phenomenon can be attributed to the nature of the CDC Dataset, which contains health indicators for individuals without temporal aspects, and only displays causal relationships between variables. Additionally, FOSA's proficiency in capturing inter-variable relationships likely contributes to the consistent performance across different scales.

Our next exploration centered on the rate of missing data. We designed an experiment using a subset of n=100,000 entries, keeping the model parameters constant, while progressively increasing the missing rate from 10% to 90% in increments of 10%. Our findings revealed that FIML displayed stable metrics for missing rates up to 60%. Beyond this threshold, its performance experienced a sharp decline. Throughout these tests, FOSA consistently outshined FIML. Notably, even with a missing rate of 40%, FOSA's results remained commendable. At a missing rate of 60%, despite FIML's initial predictions being subpar in Step-I, FOSA in Step-II managed to produce satisfactory outcomes for certain variables. Given the dataset's intrinsic properties, we surmise that FOSA can sustain acceptable performance for scenarios with missing rates up to 60%—although a threshold of 40% is preferable. For missing rates exceeding 80%, FOSA's performance begins to wane, likely marking its operational limit. It's worth noting that datasets with missing rates above 60% are rarely leveraged directly in research endeavors.

6 Conclusion

In our investigation, we delved deep into the realm of data imputation, unveiling the novel FOSA Framework. This method synergistically marries the robustness of FIML estimation with the advanced capabilities of the self-attention mechanism. Unlike conventional methods that resort to discarding missing data or employing rudimentary imputation techniques like mean replacement, FIML leverages all available information to derive estimates for missing values, laying a solid foundation for neural network models. Concurrently, the self-attention mechanism excels in capturing long-term dependencies in data—an indispensable trait, especially when dealing with intricate multivariate time series data. The fusion of these two methodologies equips us with a potent tool for proficiently predicting and handling datasets riddled with missing values. Our exhaustive experiments, spanning both simulated and real-world datasets, unequivocally highlight the superior performance of FOSA over traditional FIML methods.

One of the standout attributes of FOSA is its versatility. Whether faced with linear or nonlinear simulated datasets, FOSA consistently outperforms in the imputation realm. With real-world datasets, FOSA doesn't merely meet the performance benchmarks set by FIML—it often eclipses them. Our results suggest that even when FIML is challenged by nonlinear relationships or intricate data patterns, FOSA's resilient architecture effectively discerns underlying patterns, delivering precise and accurate imputations. This enhanced accuracy ensures that the resulting data becomes a reliable bedrock for subsequent analytical pursuits.

In addition to its accuracy, our study also emphasizes FOSA's efficiency. It not only excels in imputation accuracy but also demonstrates a distinct edge in computational speed, especially with large-scale datasets. To illustrate, for the CDC dataset—with its roughly 300,000 entries totaling 1.5 million data points—even a modest computing setup like a Dual-core Intel i7 (3.5GHz) can complete the FOSA imputation procedure (refer to Figure 1) in just about 280 seconds. With more powerful GPUs, such as the NVIDIA RTX A6000, this processing time shrinks to about 10 seconds. Given that this dataset's scale is comparable to the yearly basic health data of residents in a regional Japanese city, FOSA's computational prowess makes it an ideal candidate for extensive combination analyses and data mining tasks on similar large-scale datasets.

7 Limitation and Future work

While FOSA has demonstrated robust performance in imputation experiments, certain limitations inherent to its components warrant acknowledgment. As an integrative model, FOSA assimilates both FIML's and the self-attention mechanism's constraints. FIML, for instance, operates under the presumption of multivariate normal data distribution. Should the data deviate from this assumption, FIML's estimates could be suboptimal. Moreover, FIML is especially proficient when data is either missing completely at random (MCAR) or missing at random (MAR) concerning observed data. Any deviation from these random missing patterns might introduce biases into FIML's estimations. Additionally, FIML can occasionally grapple with convergence challenges, especially when faced with datasets exhibiting model

misspecifications or pronounced multicollinearity. Such challenges could potentially affect the quality of input destined for the self-attention phase.

On the other hand, the self-attention mechanism, while powerful, introduces its own set of challenges, notably in the domain of hyper-parameter selection and optimization. Fine-tuning these parameters is crucial, and any missteps could complicate the evaluation and interpretation processes.

Recognizing these challenges, we have designed the FOSA architecture with a keen sense of adaptability. The two primary components, FIML and self-attention, have been conceived as distinct modules. This modular approach offers flexibility, enabling the replacement of FIML with alternative imputation techniques, such as multiple imputations or random forests if deemed necessary. Subsequently, the self-attention mechanism can refine these preliminary imputations. We're eager to explore and expound upon this adaptive strategy in our forthcoming publications. As we chart our future course, we are resolute in our commitment to testing and enhancing FOSA across a broader array of real-world datasets.

References

[1] Craig K Enders and Deborah L. Bandalos. The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 8:430 – 457, 2001.

[2] Joseph L. Schafer and John W. Graham. Missing data: our view of the state of the art. *Psychological methods*, 7 2:147–77, 2002.

[3] Kristel J. M. Janssen, A. Rogier T. Donders, Frank E. Harrell, Yvonne Vergouwe, Qingxia Chen, Diederick E. Grobbee, and Karel G. M. Moons. Missing covariate data in medical research: to impute is better than to ignore. *Journal of clinical epidemiology*, 63 7:721–7, 2010.

[4] Donald B. Rubin. Inference and missing data. *Psychometrika*, 1975:19, 1975.

[5] Nicole A. Lazar. Statistical analysis with missing data. *Technometrics*, 45:364 – 365, 2003.

[6] S Fielding, P M Fayers, J H Loge, M S Jordhøy, and S Kaasa. Methods for handling missing data in palliative care research. *Palliat Med*, 20(8):791–798, Dec 2006.

[7] John W. Graham. Missing data analysis: making it work in the real world. *Annual review of psychology*, 60:549–76, 2009.

[8] Amanda N Baraldi and Craig K Enders. An introduction to modern missing data analyses. *J Sch Psychol*, 48(1):5–37, Feb 2010.

[9] Ian R. White, Patrick Royston, and Angela M. Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30, 2011.

[10] Yiran Dong and Chao-Ying Joanne Peng. Principled missing data methods for researchers. *SpringerPlus*, 2, 2013.

[11] Gina L. Mazza, Craig K. Enders, and Linda S. Ruehlman. Addressing item-level missing data: A comparison of proration and full information maximum likelihood estimation. *Multivariate Behavioral Research*, 50(5):504–519, 2015. PMID: 26610249.

[12] Craig K Enders. Multiple imputation as a flexible tool for missing data handling in clinical research. *Behav Res Ther*, 98:4–18, Nov 2017.

[13] Timothy D. Nelson, Rebecca L. Brock, Sonja Yokum, Cara C. Tomaso, Cary R. Savage, and Eric Stice. Much ado about missingness: A demonstration of full information maximum likelihood estimation to address missingness in functional magnetic resonance imaging data. *Frontiers in Neuroscience*, 15, 2021.

[14] Daniel J. Stekhoven and Peter Bühlmann. Missforest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28 1:112–8, 2011.

[15] Zhengping Che, S. Purushotham, Kyunghyun Cho, David A. Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8, 2016.

[16] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[17] Xintao Chai, Hanming Gu, Feng Li, Hongyou Duan, Xiaobo Sharon Hu, and Kai Lin. Deep learning for irregularly and regularly missing data reconstruction. *Scientific Reports*, 10, 2020.

[18] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Gain: Missing data imputation using generative adversarial nets. *International conference on machine learning. PMLR*, pages page 5689–5698, 2018.

[19] Yonghong Luo, Xiangrui Cai, Y. Zhang, Jun Xu, and Xiaojie Yuan. Multivariate time series imputation with generative adversarial networks. In *Neural Information Processing Systems*, 2018.