

Mining of Switching Sparse Networks for Missing Value Imputation in Multivariate Time Series

Kohei Obata

SANKEN, Osaka University
Osaka, Japan
obata88@sanken.osaka-u.ac.jp

Yasuko Matsubara

SANKEN, Osaka University
Osaka, Japan
yasuko@sanken.osaka-u.ac.jp

Koki Kawabata

SANKEN, Osaka University
Osaka, Japan
koki@sanken.osaka-u.ac.jp

Yasushi Sakurai

SANKEN, Osaka University
Osaka, Japan
yasushi@sanken.osaka-u.ac.jp

Abstract

Multivariate time series data suffer from the problem of missing values, which hinders the application of many analytical methods. To achieve the accurate imputation of these missing values, exploiting inter-correlation by employing the relationships between sequences (i.e., a network) is as important as the use of temporal dependency, since a sequence normally correlates with other sequences. Moreover, exploiting an adequate network depending on time is also necessary since the network varies over time. However, in real-world scenarios, we normally know neither the network structure nor when the network changes beforehand. Here, we propose a missing value imputation method for multivariate time series, namely MissNET, that is designed to exploit temporal dependency with a state-space model and inter-correlation by switching sparse networks. The network encodes conditional independence between features, which helps us understand the important relationships for imputation visually. Our algorithm, which scales linearly with reference to the length of the data, alternatively infers networks and fills in missing values using the networks while discovering the switching of the networks. Extensive experiments demonstrate that MissNET outperforms the state-of-the-art algorithms for multivariate time series imputation and provides interpretable results.

CCS Concepts

- Information systems → Data mining; • Mathematics of computing → Time series analysis.

Keywords

Multivariate time series, Missing value imputation, Network inference, State-space model, Graphical lasso

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '24, August 25–29, 2024, Barcelona, Spain

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 979-8-4007-0490-1/24/08
<https://doi.org/10.1145/3637528.3671760>

ACM Reference Format:

Kohei Obata, Koki Kawabata, Yasuko Matsubara, and Yasushi Sakurai. 2024. Mining of Switching Sparse Networks for Missing Value Imputation in Multivariate Time Series . In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3637528.3671760>

1 Introduction

With the development of the Internet of Things (IoT), multivariate time series are generated in many real-world applications, such as motion capture [37], and health monitoring [8]. However, there are inevitably many values missing from these data, and this has many possible causes (e.g., sensor malfunction). As most algorithms assume an intact input when building models, missing value imputation is indispensable for real-world applications [3, 40].

In time series data, missing values often occur consecutively, leading to a missing block in a sequence, and can happen simultaneously to multiple sequences [27]. To effectively reconstruct missing values from such partially observed data, we must exploit both temporal dependency, by taking account of past and future values in the sequence, and inter-correlation, by using the relationship between different sequences (i.e., a network) [10, 52]. Here, a network does not necessarily mean the spatial proximity of sensors but rather underlying connectivity (e.g., Pearson correlation or partial correlation). Moreover, as time series data are normally non-stationary, so is the network; an adequate network must be exploited depending on time [45, 46]. We collectively refer to a group of time points with the same network as a “regime”. Fig. 1 shows an illustrative example where missing blocks randomly exist in a multivariate time series consisting of three features (i.e., A, B, and C). Each time point belongs to either of two regimes with different networks (i.e., #1 and #2), where the thickness of the edge indicates the strength of the interplay between features. It is appropriate to use the values of feature C to impute the block missing from feature B in regime #1 since the network has an edge between B and C. On the other hand, in regime #2, it is preferable to use feature A, as the network suggests. Nevertheless, in real-world scenarios, we often have no information about the data; that is, we do not know the structure of the network, let alone when the network changes. Thus, given a partially observed multivariate time series, how can we infer networks and allocate each time point to the correct regime? How

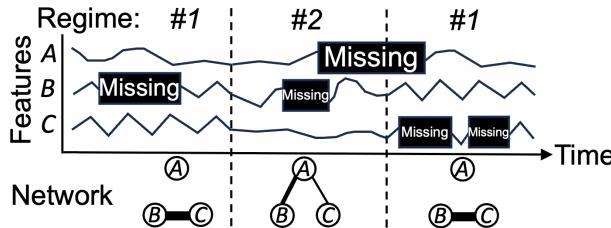


Figure 1: An illustrative example of a multivariate time series including missing blocks, where each time point of the data is allocated to two regimes, each with a distinct network, where the edges show relationships between features.

can we achieve an accurate imputation exploiting both temporal dependency and inter-correlation?

There have been many studies on time series missing value imputation with machine learning and deep learning models [23]. Many of them employ time-variant latent/hidden states, as used in a State-Space Model (SSM) or a Recurrent Neural Network (RNN), to learn the dynamic patterns of time series [7, 9, 25, 54]. However, they do not make full use of inter-correlation, while the sequences of a multivariate time series usually interact. Although they implicitly capture inter-correlation in latent space, they potentially capture spurious correlations under the presence of missing values, leading to inaccurate imputation. To tackle this problem, several studies explicitly utilize the network [5, 6, 10, 17, 21, 50]. However, they require a predefined network, even though we rarely know it in advance.

In this work, we propose MissNET¹, *Mining of switching sparse Networks* for multivariate time series imputation, which repeatedly infers sparse networks from imputed data and fills in missing values using the networks while allocating each time point to regimes. Specifically, our model has three components: (1) a regime-switching model based on a discrete Markov process for detecting the change point of the network. (2) An imputation model based on an SSM for exploiting temporal dependency and inter-correlation by latent space and the networks. (3) A network inference model for inferring sparse networks via graphical lasso, where each network encodes pairwise conditional independencies among features, and the lasso penalty helps avoid capturing spurious correlations. Our proposed algorithm maximizes the joint probability distribution over the above components.

Our method has the following desired properties:

- *Effective*: MissNET, which exploits both temporal dependency and inter-correlation by inferring switching sparse networks, outperforms the state-of-the-art algorithms for missing value imputation for multivariate time series.
- *Scalable*: Our proposed algorithm scales linear with regard to the length of the time series and is thus applicable to a long-range time series.
- *Interpretable*: MissNET discovers regimes with distinct sparse networks, which help us interpret data, e.g., what is the key relationship in the data, and why does a particular regime distinguish itself from others?

¹Our source code and datasets are publicly available: <https://github.com/KoheiObata/MissNet>.

Table 1: Capabilities of MissNET, Matrix Factorization (MF), State-Space Model (SSM), Deep Learning (DL), and Graphical Lasso (GL)-based approaches.

	MF	SSM	DL	GL					
	SoftImpute [30]	SLDS [36]	DynaMMo [25]	DCMF [5]	BRITS [7]	POGEVON [50]	TICC [19]	MNCL [48]	MissNET
Temporal dependency	-	-	✓	✓	✓	✓	-	-	✓
Inter-correlation	-	✓	-	-	✓	-	✓	-	✓
Time-varying inter-correlation	-	-	-	-	-	✓	-	✓	✓
Missing value imputation	✓	✓	✓	-	✓	✓	✓	-	✓
Segmentation	-	-	-	✓	-	-	✓	-	✓
Sparse network inference	-	-	-	-	-	-	✓	✓	✓

2 Related work

We review previous studies closely related to our work. Table 1 shows the relative advantages of MissNET. Current approaches fall short with respect to at least one of these desired characteristics.

Time series missing value imputation. Missing value imputation for time series is a very rich topic [23]. We roughly classify missing value imputation methods as Matrix Factorization (MF)-based, SSM-based, and Deep Learning (DL)-based approaches.

MF-based methods, such as SoftImpute [30] based on Singular Value Decomposition (SVD), recover missing values from low-dimensional embedding matrices of partially observed data [22, 35, 52, 56]. For example, SoRec [28], proposed as a recommendation system, constrains MF with a predefined network to exploit inter-correlation. Since MF is limited in capturing temporal dependency, TRMF [55] uses an Auto-Regressive (AR) model and imposes temporal smoothness on MF.

SSMs, such as Linear Dynamical Systems (LDS) [16], use latent space to capture temporal dependency, where the data point depends on all past data points [9, 11, 17, 39]. To fit more complex time series, Switching LDS (SLDS) [14, 36] switches multiple LDS models. SSM-based methods, such as DynaMMo [25], focus on capturing the dynamic patterns in time series rather than inter-correlation implicitly captured through the latent space. To use the underlying connectivity in multivariate time series, DCMF [5], and its tensor extension Facets [6] use SSM constrained with a predefined network, which is effective, especially when the missing rate is high. However, they assume that the network is accurately known and fixed, while it is usually unknown and may change over time in real-world scenarios.

Recently, extensive research has focused on DL-based methods, employing techniques including graph neural networks [10, 13, 21, 38], self-attention [12, 41], and, most recently, diffusion models [1, 44, 51], to harness their high model capacity [2, 53]. For example, BRITS [7] and M-RNN [54] impute missing values according to hidden states from bidirectional RNN. To utilize dynamic inter-correlation in time series, POGEVON [50] requires a sequence of networks and imputes missing values in time series and missing edges in the networks, assuming that the network varies over time. Although DL-based methods can handle complex data, the imputation quality depends heavily on the size and the selection of the training dataset.

Sparse network inference. From another perspective, our method infers sparse networks from time series containing missing values and discovers regimes (i.e., clusters) based on networks. Inferring a sparse inverse covariance matrix (i.e., network) from data helps us to understand feature dependency in a statistical way. Graphical lasso [15], which maximizes the Gaussian log-likelihood imposing an ℓ_1 -norm penalty, is one of the most commonly used techniques for estimating a sparse network from static data. Städler and Bühlmann [42] have tackled inferring a sparse network from partially observed data according to conditional probability. However, the network varies over time [29, 32, 33]; thus, TVGL [18] infers time-varying networks by considering the time similarity with a network belonging to neighboring segments. To infer time-varying networks in the presence of missing values, MMGL [48], which employs TVGL, uses the expectation-maximization (EM) algorithm to repeat the inference of time-varying networks and missing value imputation based on conditional probability under the condition that each segment has the same observed features. Discovering clusters based on networks [34], such as TICC [19] and TAGM [47], provides interpretable results that other traditional clustering methods cannot find. However, they cannot handle missing values.

As a consequence, none of the previous studies have addressed missing value imputation for multivariate time series by employing sparse network inference and segmentation based on the network.

3 Preliminaries

3.1 Problem definition

In this paper, we focus on the task of multivariate time series missing value imputation. We use a multivariate time series with N features and T timesteps $X = \{x_1, x_2, \dots, x_T\} \in \mathbb{R}^{N \times T}$. To represent the missing values in X , we introduce the indicator matrix $W \in \mathbb{R}^{N \times T}$, where $W_{i,t}$ indicates the availability of feature i at timestep t : $W_{i,t}$ being 0 or 1 indicates whether $X_{i,t}$ is missing or observed. Thus, the observed entries can be described as $\tilde{X} = W \circ X$, where \circ is a Hadamard product. Our problem is formally written as follows:

PROBLEM 1 (MULTIVARIATE TIME SERIES MISSING VALUE IMPUTATION). *Given:* a partially observed multivariate time series \tilde{X} ; *Recover:* its missing parts indicated by W .

3.2 Graphical lasso

We use graphical lasso [15] to infer the network for each regime. Given X , graphical lasso estimates the sparse Gaussian inverse covariance matrix (i.e., network) $\Theta \in \mathbb{R}^{N \times N}$, also known as the precision matrix. The network encodes pairwise conditional independencies among N features, e.g., if $\Theta_{i,j} = 0$, then features i and j are conditionally independent given the values of all the other features. The optimization problem is expressed as follows:

$$\text{minimize}_{\Theta \in S_{++}^p} \lambda \|\Theta\|_{od,1} - \sum_{t=1}^T ll(x_t, \Theta), \quad (1)$$

$$\begin{aligned} ll(x_t, \Theta) = & -\frac{1}{2}(x_t - \rho)' \Theta (x_t - \rho) \\ & + \frac{1}{2} \log \det(\Theta) - \frac{N}{2} \log(2\pi), \end{aligned} \quad (2)$$

where Θ must be symmetric positive definite (S_{++}^p). $ll(x_t, \Theta)$ is the log-likelihood and $\rho \in \mathbb{R}^N$ is the empirical mean of X . $\lambda \geq 0$ is a hyperparameter for determining the sparsity level of the network, and $\|\cdot\|_{od,1}$ indicates the off-diagonal ℓ_1 -norm. This is a convex optimization problem that can be solved via the alternating direction method of multipliers (ADMM) [4].

4 Proposed MissNet

In this section, we present our proposed MissNet for missing value imputation. We give the formal formulation of the model and then provide the detailed algorithm to learn the model.

4.1 Optimization model

MissNet infers sparse networks and fills in missing values using the networks while discovering regimes. We first introduce three interacting components of our model: a regime-switching model, an imputation model, and a network inference model. Then, we define the optimization formulation.

Regime-switching model. MissNet describes the change of networks by regime-switching. Let K be the number of regimes, $F = \{f_1, \dots, f_T\} \in \mathbb{R}^{K \times T}$ be regime assignments, and f_t be a one-hot vector² that indicates $\tilde{x}_t \in \mathbb{R}^N$ belongs to f_t^{th} -regime. We assume regime-switching to be a discrete first-order Markov process:

$$p(f_1) = \pi_0, \quad p(f_{t+1}|f_t) = \Pi_{f_{t+1}f_t}, \quad (3)$$

where $\Pi \in \mathbb{R}^{K \times K}$ is the Markov transition matrix and $\pi_0 \in \mathbb{R}^K$ is the initial state distribution.

Imputation model. MissNet imputes missing values exploiting temporal dependency and inter-correlation indicated by the networks. We assume that the temporal dependency is consistent throughout all regimes and captured in the latent space of an SSM, which allows us to consider long-term dependency. We define the latent states $Z = \{z_1, z_2, \dots, z_T\} \in \mathbb{R}^{L \times T}$ corresponding to \tilde{X} , where L is the number of latent dimensions, and $z_{t+1} \in \mathbb{R}^L$ is linear to z_t through the transition matrix $B \in \mathbb{R}^{L \times L}$ with the covariance Σ_Z , shown in Eq. (5). As defined in Eq. (4), the first timestep of latent state z_1 is defined by the initial state z_0 and the covariance Ψ_0 .

$$z_1 \sim \mathcal{N}(z_0, \Psi_0), \quad (4)$$

$$z_{t+1}|z_t \sim \mathcal{N}(Bz_t, \Sigma_Z). \quad (5)$$

We define the observation \tilde{x}_t assigned to f_t^{th} -regime as being linear to the latent state z_t through the observation matrix of f_t^{th} -regime $U^{(f_t)} \in \mathbb{R}^{N \times L}$ with the covariance $\Sigma_{X^{(f_t)}}$:

$$\tilde{x}_t|z_t, U^{(f_t)}, f_t \sim \mathcal{N}(U^{(f_t)} z_t, \Sigma_{X^{(f_t)}}). \quad (6)$$

MissNet captures inter-correlation by adding a constraint on $U^{(k)}$. Let it be assumed that the contextual matrix of the k^{th} -regime $S^{(k)} \in \mathbb{R}^{N \times N}$ encodes the inter-correlation of X belonging to the k^{th} -regime ($1 \leq k \leq K$). We define the contextual latent factor of the k^{th} -regime $V^{(k)} \in \mathbb{R}^{L \times N}$, and the j -th column ($1 \leq j \leq N$) of the contextual matrix $s_j^{(k)}$ as linear to $v_j^{(k)}$ through the observation matrix $U^{(k)}$ with the covariance $\Sigma_{S^{(k)}}$, formulated in Eq. (7). In

²We use it interchangeably with the index of the regime (i.e., f_t^{th} -regime).

Eq. (8), we place zero-mean spherical Gaussian priors on $\mathbf{v}_j^{(k)}$ under the assumption that $-1 \leq S_{i,j}^{(k)} \leq 1$.

$$\mathbf{s}_j^{(k)} | \mathbf{v}_j^{(k)}, \mathbf{U}^{(k)} \sim \mathcal{N}(\mathbf{U}^{(k)} \mathbf{v}_j^{(k)}, \Sigma_{S^{(k)}}), \quad (7)$$

$$\mathbf{v}_j^{(k)} \sim \mathcal{N}(0, \Sigma_{V^{(k)}}). \quad (8)$$

To avoid our model overfitting the observed entries since many entries are missing, we simplify the covariances by assuming that all noises are independent and identically distributed (i.i.d.). Thus, the parameters $\Sigma_Z, \Sigma_{X^{(k)}}, \Sigma_{S^{(k)}}, \Sigma_{V^{(k)}}$ are reduced to $\sigma_Z^2, \sigma_{X^{(k)}}^2, \sigma_{S^{(k)}}^2, \sigma_{V^{(k)}}^2$, respectively (i.e., $\Sigma_Z = \sigma_Z^2 \mathbf{I}$, $\Sigma_{X^{(k)}} = \sigma_{X^{(k)}}^2 \mathbf{I}$, $\Sigma_{S^{(k)}} = \sigma_{S^{(k)}}^2 \mathbf{I}$, and $\Sigma_{V^{(k)}} = \sigma_{V^{(k)}}^2 \mathbf{I}$).

With the above imputation model, the imputed time series $\tilde{\mathbf{x}}_t$ at timestep t is recovered as follows:

$$\tilde{\mathbf{x}}_t = \mathbf{W}_{:,t} \circ \tilde{\mathbf{x}}_t + (1 - \mathbf{W}_{:,t}) \circ \mathbf{U}^{(f_t)} \mathbf{z}_t. \quad (9)$$

Network inference model. MissNET infers the network for each regime to exploit inter-correlation. We define a Gaussian distribution and ℓ_1 -norm for the imputed data belonging to each regime to allow us to estimate networks accurately:

$$\hat{\mathbf{x}}_t | f_t \sim \mathcal{N}(\boldsymbol{\rho}^{(f_t)}, \Theta^{(f_t)})^{-1}, \quad s.t., \|\Theta^{(f_t)}\|_{od,1} \leq \frac{e}{\lambda}, \quad (10)$$

where e is any constant value for convenience, and λ is a hyper-parameter that controls the sparsity of the network (i.e., inverse covariance matrix) $\Theta^{(f_t)}$ with ℓ_1 -norm, which helps avoid capturing spurious correlations.

Optimization formulation. Our goal is to estimate the model parameters $\theta = \{B, z_0, \Psi_0, \sigma_Z, \sigma_X, \sigma_S, \sigma_V, \dot{\rho}, \dot{\Theta}, \pi_0, \Pi\}$ and find the latent factors $Z, \dot{V}, \dot{U}, \dot{S}, F$, where the letters with a dot indicate a set of K vectors/matrices/scalars (e.g., $\dot{S} = \{S^{(k)}\}_{k=1}^K$), that maximizes the following joint probability distribution:

$$\begin{aligned} \arg \max p(\tilde{X}, Z, \dot{V}, \dot{U}, \dot{S}, F) &= p(z_1) \underbrace{\prod_{t=2}^T p(z_t | z_{t-1})}_{\text{temporal dependency}} \\ &\quad \underbrace{\prod_{t=1}^T p(\tilde{x}_t | z_{t-1}, \mathbf{U}^{(f_t)}, f_t) \prod_{k=1}^K \left\{ \underbrace{\prod_{j=1}^N p(v_j^{(k)})}_{\text{time series}} \underbrace{\prod_{j=1}^N p(s_j^{(k)} | \mathbf{U}^{(k)}, v_j^{(k)})}_{\text{network constraint}} \right\}}_{\text{network constraint}} \\ &\quad p(f_1) \underbrace{\prod_{t=2}^T p(f_t | f_{t-1})}_{\text{regime-switching}} \underbrace{\prod_{t=1}^T p(\tilde{x}_t | \boldsymbol{\rho}^{(f_t)}, \Theta^{(f_t)}, f_t),}_{\text{network inference}} \quad (s.t. \|\Theta^{(k)}\|_{od,1} \leq \frac{e}{\lambda}) \underbrace{\prod_{j=1}^N p(s_j^{(k)} | \mathbf{U}^{(k)}, v_j^{(k)})}_{\text{network sparsity}}. \end{aligned} \quad (11)$$

4.2 Algorithm

It is difficult to find the global optimal solution of Eq. (11), for the following reasons: (i) As a constraint for \dot{U}, \dot{S} has to be fixed and encode inter-correlation; (ii) \dot{U}, Z and F jointly determine \tilde{X} , and \dot{V} and \dot{U} jointly determine \dot{S} ; (iii) Calculating the correct F is NP-hard. Hence, we aim to find its local optimum instead, following the EM algorithm, where the graphical model for each iteration is shown in Fig. 2. Specifically, to address the aforementioned difficulties, we

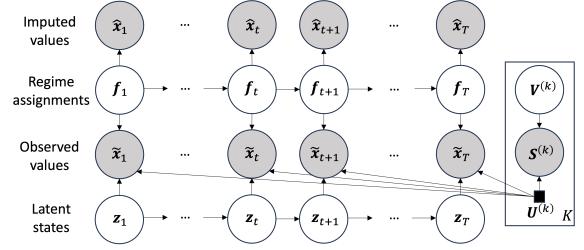


Figure 2: Graphical model of MissNET at each iteration.

employ the following steps: (i) we consider \dot{S} is observed in each iteration, and we update it at the end of the iteration using $\dot{\Theta}$; (ii) we regard \dot{U} as a model parameter, thus $\{\tilde{X}, Z, F\}$ are independent with $\{\dot{S}, \dot{V}\}$. We alternate the inference of $\{Z, F\}$ and \dot{V} and the update of the model parameters; (iii) we employ a Viterbi approximation and infer the most likely F .

4.2.1 E-step. Given \dot{U} , we can infer $\{Z, F\}$ and \dot{V} independently.

Let \mathbf{o}_t denote the indices of the observed entries of $\tilde{\mathbf{x}}_t$. The observed-only data $\tilde{\mathbf{x}}_t$ and the corresponding observed-only observation matrix $\bar{\mathbf{U}}_t^{(k)}$ are defined as follows:

$$\begin{aligned} \mathbf{o}_t &= \{i | W_{i,t} > 0, i = 1, \dots, N\}, \\ \tilde{\mathbf{x}}_t &= \tilde{\mathbf{x}}_t(\mathbf{o}_t), \quad \bar{\mathbf{U}}_t^{(k)} = \mathbf{U}^{(k)}(\mathbf{o}_t, :). \end{aligned} \quad (12)$$

Inferred F and Z . F and Z are coupled, and so must be jointly determined. We first use a Viterbi approximation to find the most likely regime assignments F that maximize the log-likelihood Eq. (11). The likelihood term obtained during the calculation of F also acts as a Kalman Filter (forward algorithm). Then, we infer Z with a Rauch-Tung-Streibel (RTS) smoother (backward algorithm).

In a Viterbi approximation, finding F requires the partial cost $J_{t,t-1,k,l}$ when the switch is to regime k at time t from regime l at time $t-1$. To calculate the partial cost, we define the following LDS state and variance terms:

$$\begin{aligned} \boldsymbol{\mu}_{t|t-1,k,l} &= B\boldsymbol{\mu}_{t-1|t-1,k,l}, \\ \boldsymbol{\mu}_{t|t,k,l} &= \boldsymbol{\mu}_{t|t-1,k,l} + D_{t,k,l}(\tilde{\mathbf{x}}_t - \bar{\mathbf{U}}_t^{(k)} B\boldsymbol{\mu}_{t|t-1,k,l}), \\ \boldsymbol{\Psi}_{t|t-1,k,l} &= B\Psi_{t-1|t-1,k,l} B' + \sigma_Z^2 \mathbf{I}, \\ \boldsymbol{\Psi}_{t|t,k,l} &= (\mathbf{I} - D_{t,k,l} \bar{\mathbf{U}}_t^{(k)}) \boldsymbol{\Psi}_{t-1|t-1,k,l}, \\ D_{t,k,l} &= \boldsymbol{\Psi}_{t|t-1,k,l} \bar{\mathbf{U}}_t^{(k)'} (\bar{\mathbf{U}}_t^{(k)} \boldsymbol{\Psi}_{t|t-1,k,l} \bar{\mathbf{U}}_t^{(k)'} + \sigma_{X^{(k)}}^2 \mathbf{I})^{-1}, \end{aligned} \quad (13)$$

with the initial state:

$$\begin{aligned} \boldsymbol{\mu}_{1|1,k} &= z_0 + D_{1,k}(\tilde{\mathbf{x}}_1 - \bar{\mathbf{U}}_1^{(k)} z_0), \\ \boldsymbol{\Psi}_{1|1,k} &= (\mathbf{I} - D_{1,k} \bar{\mathbf{U}}_1^{(k)}) \Psi_0, \\ D_{1,k} &= \Psi_0 \bar{\mathbf{U}}_1^{(k)'} (\bar{\mathbf{U}}_1^{(k)} \Psi_0 \bar{\mathbf{U}}_1^{(k)'} + \sigma_{X^{(k)}}^2 \mathbf{I})^{-1}, \end{aligned} \quad (14)$$

where $\boldsymbol{\mu}_{t|t-1,k,l}$ and $\boldsymbol{\mu}_{t|t,k,l}$ are the one-step predicted LDS state and the best-filtered state estimates at t , respectively, given the switch is in regime k at time t and in regime l at time $t-1$ and only the $t-1$ measurements are known. Similar definitions are used for $\boldsymbol{\Psi}_{t|t-1,k,l}$ and $\boldsymbol{\Psi}_{t|t,k,l}$.

The partial cost is obtained by calculating the logarithm of Eq. (11) related to f_t :

$$\begin{aligned} J_{t,t-1,k,l} &= \frac{1}{2} (\tilde{\mathbf{x}}_t - \bar{\mathbf{U}}_t^{(k)} \boldsymbol{\mu}_{t|t-1,k,l})' \mathbf{R} (\tilde{\mathbf{x}}_t - \bar{\mathbf{U}}_t^{(k)} \boldsymbol{\mu}_{t|t-1,k,l}) \\ &\quad - \frac{1}{2} \log \det(\mathbf{R}) + \frac{L}{2} \log(2\pi) \\ &\quad + \frac{1}{2} (\hat{\mathbf{x}}_t - \boldsymbol{\rho}^{(k)})' \boldsymbol{\Theta}^{(k)} (\hat{\mathbf{x}}_t - \boldsymbol{\rho}^{(k)}) \\ &\quad - \frac{1}{2} \log \det(\boldsymbol{\Theta}^{(k)}) + \frac{N}{2} \log(2\pi) - \log(\Pi_{k,l}), \quad (15) \\ \mathbf{R} &= (\bar{\mathbf{U}}_t^{(k)} \Psi_{t|t-1,k,l} \bar{\mathbf{U}}_t^{(k)'} + \sigma_{X^{(k)}}^2 \mathbf{I})^{-1}. \end{aligned}$$

Once all partial costs are obtained, it is well-known how to apply a Viterbi inference to a discrete Markov process to obtain the most likely regime assignments F [49].

Then, we infer Z . Let the posteriors of \mathbf{z}_t be as follows:

$$\begin{aligned} p(\mathbf{z}_t | \tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_t) &= \mathcal{N}(\mathbf{z}_t | \boldsymbol{\mu}_t, \Psi_t), \\ p(\mathbf{z}_t | \tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_T) &= \mathcal{N}(\mathbf{z}_t | \hat{\boldsymbol{\mu}}_t, \hat{\Psi}_t). \end{aligned} \quad (16)$$

We now obtain $\boldsymbol{\mu}_t, \Psi_t, \Psi_{t|t-1}$ using F ; thus, in practice, we have conducted a Kalman Filter. We apply the RTS smoother to infer Z .

$$\begin{aligned} \hat{\boldsymbol{\mu}}_t &= \boldsymbol{\mu}_t + C_t (\hat{\boldsymbol{\mu}}_{t+1} - B \boldsymbol{\mu}_t), \\ \hat{\Psi}_t &= \Psi_t + C_t (\hat{\Psi}_{t+1} - \Psi_{t|t-1}) C_t', \\ C_t &= \Psi_t B' (\Psi_{t|t-1})^{-1}, \quad (17) \\ \mathbb{E}[\mathbf{z}_t] &= \hat{\boldsymbol{\mu}}_t, \\ \mathbb{E}[\mathbf{z}_t \mathbf{z}'_{t-1}] &= \hat{\Psi}_t C'_t + \hat{\boldsymbol{\mu}}_t \hat{\boldsymbol{\mu}}'_{t-1}, \\ \mathbb{E}[\mathbf{z}_t \mathbf{z}'_t] &= \hat{\Psi}_t + \hat{\boldsymbol{\mu}}_t \hat{\boldsymbol{\mu}}'_{t-1}. \end{aligned} \quad (18)$$

Inferring \dot{V} . We apply Bayes' theorem to Eq. (7) and (8) to obtain the posteriors $p(\mathbf{v}_j^{(k)} | s_j^{(k)}) = \mathcal{N}(\mathbf{v}_j^{(k)} | \mathbf{v}_j^{(k)}, \Upsilon^{(k)})$:

$$\begin{aligned} \mathbf{v}_j^{(k)} &= \mathbf{M}^{(k)-1} \mathbf{U}^{(k)'} \mathbf{s}_j^{(k)}, \\ \Upsilon^{(k)} &= \sigma_{S^{(k)}}^2 \mathbf{M}^{(k)-1}, \\ \mathbf{M}^{(k)} &= \mathbf{U}^{(k)'} \mathbf{U}^{(k)} + \sigma_{V^{(k)}}^{-2} \sigma_{S^{(k)}}^2 \mathbf{I}, \\ \mathbb{E}[\mathbf{v}_j^{(k)}] &= \mathbf{v}_j^{(k)}, \\ \mathbb{E}[\mathbf{v}_j^{(k)} \mathbf{v}_j^{(k)'}] &= \Upsilon^{(k)} + \mathbf{v}_j^{(k)} \mathbf{v}_j^{(k)'} \quad (19) \end{aligned}$$

4.2.2 M -step. After obtaining $\{Z, F\}$ and \dot{V} , we update the model parameters to maximize the expectation of the log-likelihood:

$$\theta^{new} = \arg \max_{\theta} Q(\theta, \theta^{old}), \quad (20)$$

$$Q(\theta, \theta^{old}) = \mathbb{E}_{Z, F, \dot{V} | \theta^{old}} [\log p(\tilde{\mathbf{x}}, Z, \dot{V}, \dot{U}, \dot{S}, F | \theta)] - \sum_{k=1}^K \lambda \|\boldsymbol{\Theta}^{(k)}\|_{od,1}$$

where we incorporate the ℓ_1 -norm constraint.

Parameters for the imputation model, \dot{U} and $\{\mathbf{B}, \mathbf{z}_0, \Psi_0, \sigma_Z, \sigma_X, \sigma_S, \sigma_V\}$, can be obtained by taking the derivative of $Q(\theta, \theta^{old})$. For $\mathbf{U}^{(k)}$,

we update each row $\mathbf{U}_{i,:}^{(k)}$ individually:

$$(\mathbf{U}_{i,:}^{(k)})^{new} = \mathbf{A}_1^{(k)} \mathbf{A}_2^{(k)-1}, \quad (21)$$

$$\begin{aligned} \mathbf{A}_1^{(k)} &= \alpha \sigma_{S^{(k)}}^{-2} \sum_{j=1}^N S_{i,j}^{(k)} \mathbb{E}[\mathbf{v}_j^{(k)'}] + (1-\alpha) \sigma_{X^{(k)}}^{-2} \sum_{t=1, f_t \in k}^T \mathbf{W}_{i,t} \tilde{\mathbf{x}}_{i,t} \mathbb{E}[\mathbf{z}_t'], \\ \mathbf{A}_2^{(k)} &= \alpha \sigma_{S^{(k)}}^{-2} \sum_{j=1}^N \mathbb{E}[\mathbf{v}_j^{(k)} \mathbf{v}_j^{(k)'}] + (1-\alpha) \sigma_{X^{(k)}}^{-2} \sum_{t=1, f_t \in k}^T \mathbf{W}_{i,t} \mathbb{E}[\mathbf{z}_t \mathbf{z}_t'], \end{aligned}$$

where $0 \leq \alpha \leq 1$ is a hyperparameter employed as a trade-off for the contributions of inter-correlation and temporal dependency. The details of updating $\{\mathbf{B}, \mathbf{z}_0, \Psi_0, \sigma_Z, \sigma_X, \sigma_S, \sigma_V\}$ are presented in Appendix A.3.

For the network inference model, we calculate \dot{X} with Eq. (22) and then update $\boldsymbol{\rho}^{(k)}$ by calculating the empirical mean of \dot{X} belonging to the k^{th} -regime and $\boldsymbol{\Theta}^{(k)}$ by solving the graphical lasso problem shown in Eq. (23) via ADMM.

$$\dot{\mathbf{x}}_t = \mathbf{W}_{:,t} \circ \tilde{\mathbf{x}}_t + (1 - \mathbf{W}_{:,t}) \circ (\mathbf{U}^{(f_t)})^{new} \hat{\boldsymbol{\mu}}_t, \quad (22)$$

$$\text{minimize}_{\boldsymbol{\Theta}^{(k)} \in S_{++}^P} \lambda \|\boldsymbol{\Theta}^{(k)}\|_{od,1} - \sum_{t=1, f_t \in k}^T ll(\dot{\mathbf{x}}_t, \boldsymbol{\Theta}^{(k)}). \quad (23)$$

For the regime-switching model, the initial state distribution and the Markov transition matrix are updated as follows:

$$\boldsymbol{\pi}_0 = f_1, \quad \Pi = \left(\sum_{t=2}^T f_t f_{t-1}' \right) \text{diag} \left(\sum_{t=2}^T f_t \right)^{-1}. \quad (24)$$

4.2.3 Update \dot{S} . We update $S^{(k)}$ at the end of each iteration. As shown in Eq. (25), we define the off-diagonal elements of $S^{(k)}$ as partial correlations calculated from the network $\boldsymbol{\Theta}^{(k)}$ to encode the inter-correlation.

$$S_{i,j}^{(k)} = \begin{cases} 1 & (i=j) \\ -\left(\frac{\Theta_{i,j}^{(k)}}{\sqrt{\Theta_{i,i}^{(k)}} \sqrt{\Theta_{j,j}^{(k)}}} \right) & (i \neq j) \end{cases} \quad (25)$$

4.2.4 Overall algorithm. We have the overall algorithm shown as Alg. 1 to obtain a local optimal solution of Eq. (11). Given a partially observed multivariate time series $\tilde{\mathbf{x}}$, an indicator matrix \mathbf{W} , the dimension of latent state L , the number of regimes K , the network parameter α , and the sparse parameter λ , our algorithm aims to find the latent factors $Z, \dot{V}, \dot{U}, \dot{S}, F$, other model parameters in θ , and imputed time series $\dot{\mathbf{x}}$.

The MissNET algorithm starts by initializing $\dot{\mathbf{x}}$ with a linear interpolation, and by randomly initializing $Z, \dot{V}, \dot{U}, \dot{S}, F$, and θ (Line 3). Then, it alternately updates the latent factors and parameters until they converge. In each iteration, we consider \dot{S} to be given and \dot{U} to be a model parameter. In an iteration, we first conduct a Viterbi approximation to calculate the most likely regime assignments F (Line 5-12). Then, we infer the expectations of Z and \dot{V} (Line 13-15 and Line 16-18), and we update \dot{U} and model parameters θ (Line 20), and at the end of the iteration, we update \dot{S} (Line 21).

4.3 Complexity analysis

LEMMA 1. *The time complexity of MissNET is $O(\#iter \cdot (K^2 \sum_{t=1}^T (L^3 + L^2 N_t + LN_t^2 + N_t^3) + KL^2 N^2 + KTL^2 N + KN^3))$.*

Algorithm 1 MissNET ($\tilde{X}, W, L, K, \alpha, \lambda$)

```

1: Input: (a) partially observed multivariate time series  $\tilde{X}$ ,
   (b) indicator matrix  $W$ ,
   (c) and hyperparameters  $L, K, \alpha, \lambda$ 
2: Output: latent factors  $Z, \hat{V}, \hat{U}, \hat{S}, F$  model parameters  $\theta$ , and  $\hat{X}$ 
3: Initialize  $\hat{X}$  with linear interpolation,  $Z, \hat{V}, \hat{U}, \hat{S}, F$ , and  $\theta$ ;
4: repeat
5:   for  $t = 1 : T$  do
6:     for  $k = 1 : K$  do
7:       for  $l = 1 : K$  do
8:         Calculate partial cost  $J_{t,t-1,k,l}$  using Eq. (15)
9:       end for
10:      end for
11:    end for
12:    Infer  $F$  and obtain  $\mu_t, \Psi_t$  based on Eq. (13), (14)
13:    for  $t = T : 1$  do
14:      Infer  $\hat{\mu}_t, \hat{\Psi}_t, \mathbb{E}[z_t z'_{t-1}], \mathbb{E}[z_t z'_t]$  by Eq. (17), (18)
15:    end for
16:    for  $j = 1 : N$  do
17:      Infer  $\mathbb{E}[v_j^{(k)}], \mathbb{E}[v_j^{(k)} v_j^{(k)\prime}]$  using Eq. (19)
18:    end for
19:    Set  $Z = \{\hat{\mu}_1, \dots, \hat{\mu}_T\}, \{V^{(k)} = \{v_1^{(k)}, \dots, v_N^{(k)}\}\}_{k=1}^K$ 
20:    Update  $\hat{U}, \theta$  and  $\hat{X}$  by Eq. (21) - (24)
21:    Update  $\hat{S}$  based on Eq. (25)
22:  until convergence;
23: return  $\{Z, \hat{V}, \hat{U}, \hat{S}, F, \theta, \hat{X}\}$ ;

```

PROOF. See Appendix A.2. \square

N_t represents the number of observed features of \tilde{x}_t . Note that $K^2 \sum_{t=1}^T (L^3 + L^2 N_t + LN_t^2 + N_t^3)$ is upper bounded by $K^2 T N^3$. In practice, the length of the time series (T) is often orders of magnitude greater than the number of features (N). Hence, the actual running time of MissNET is dominated by the term related to T , which is linear in T .

LEMMA 2. *The space complexity of MissNET is $O(TN + K^2 TL^2 + KL^2 N + KN^2)$.*

PROOF. See Appendix A.2. \square

5 Experiments

In this section, we empirically evaluate our approach against state-of-the-art baselines on 12 datasets. We present experimental results for the following questions:

Q1. Effectiveness: How accurate is the proposed MissNET for recovering missing values?

Q2. Scalability: How does the proposed algorithm scale?

Q3. Interpretability: How can MissNET help us understand the data?

5.1 Experimental setup

5.1.1 *Datasets.* We use the following datasets.

Synthetic. We generate two types of synthetic data, PatternA and PatternB, five times each, by defining Z and \hat{U} . We set $T = 1000, N =$

50 and $L = 10$ (Appendix B.1). PatternA has one regime ($K = 1$), and in PatternB ($K = 2$), two regimes switch at every 200 timesteps.

MotionCapture. This dataset contains nine types of full body motions of MotionCapture database³. Each motion measures the positions of 41 bones in the human body, resulting in a total of 123 features (X, Y, and Z coordinates).

Motes. This dataset consists of temperature measurements from the 54 sensors deployed in the Intel Berkeley Research Lab⁴. We use hourly data for the first two weeks (03-01 ~ 03-14). Originally, 9.6% of the data is missing, including a blackout from 03-10 to 03-11 where all the values are missing.

5.1.2 *Data preprocessing.* We generate a synthetic missing block at a length of $0 \sim 5\%$ of the data length and place it randomly until the total missing rate reaches $\{10, 20, \dots, 80\}$. Thus, a missing block can be longer than $0.05T$ when it overlaps. An additional 10% of missing values are added for hyperparameter tuning. Each dataset feature is normalized independently using a z-score so that each dataset has a zero mean and a unit variance.

5.1.3 *Comparison methods.* We compare our method with state-of-the-art imputation methods ranging from classical baselines (Linear and Quadratic), MF-based methods (SoftImpute, CDRec and TRMF), SSM-based methods (DynaMMo and DCMF), to DL-based methods (BRITS, SAITS and TIDER).

- Linear/Quadratic⁵ use linear/quadratic equations to interpolate missing values.
- SoftImpute [30] first fills in missing values with zero, then alternates between recovering the missing values and updating the SVD using the recovered matrix.
- CDRec [22] is a memory-efficient algorithm that iterates centroid decomposition (CD) and missing value imputation until they converge.
- TRMF [55] is based on MF that imposes temporal dependency among the data points with the AR model.
- DynaMMo [25] first fills in missing values using linear interpolation and then uses the EM algorithm to iteratively recover missing values and update the LDS model.
- DCMF [5] adds a contextual constraint to SSM and captures inter-correlation by a predefined network. As suggested in the original paper, we give the cosine similarity between each pair of time series calculated after linear interpolation as a predefined network. This method is similar to MissNET if we set $K = 1$, employ a predefined network that is fixed throughout the algorithm, and eliminate the effect of regime-switching and network inference models from MissNET.
- BRITS [7] imputes missing values according to hidden states from bidirectional RNN.
- SAITS [12] is a self-attention-based model that jointly optimizes imputation and reconstruction to perform the missing value imputation of multivariate time series.
- TIDER [26] learns disentangled seasonal and trend representations by employing a neural network combined with MF.

³<http://mocap.cs.cmu.edu>

⁴<https://db.csail.mit.edu/labdata/labdata.html>

⁵<https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.interpolate.html>

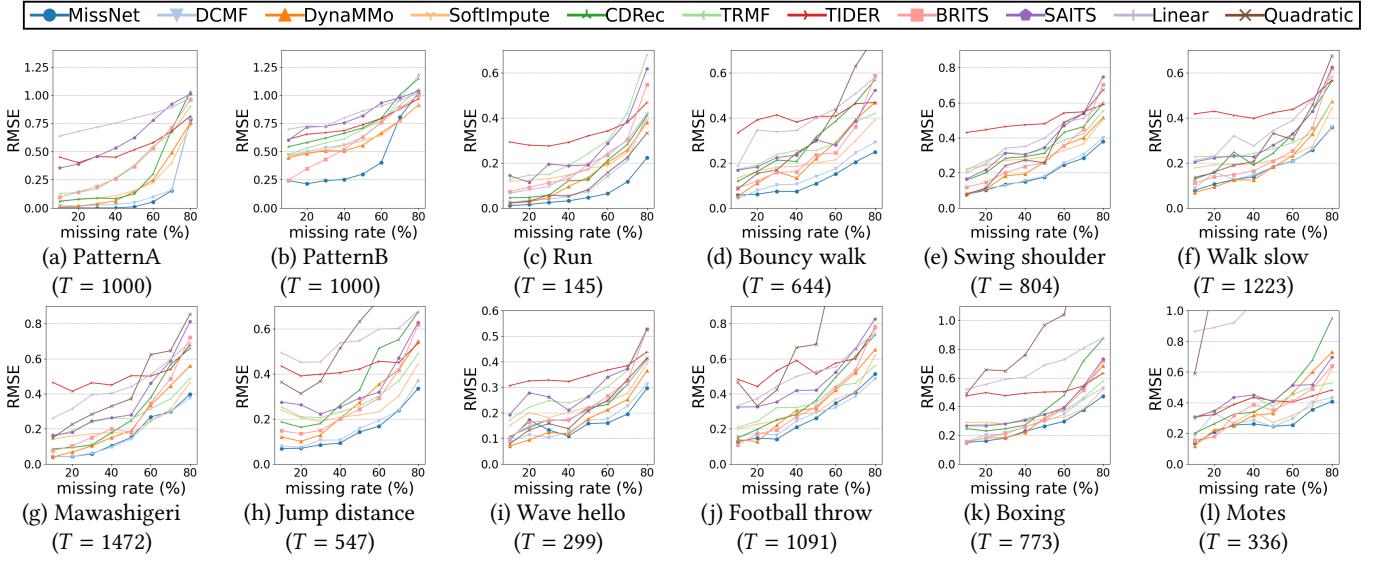


Figure 3: RMSE of (a), (b) Synthetic ($N = 50$), (c) ~ (k) MotionCapture ($N = 123$) and (l) Motes ($N = 54$) datasets.

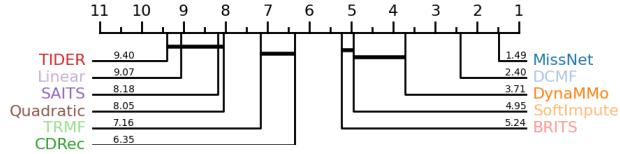


Figure 4: Critical difference diagram of real-world datasets.

5.1.4 Hyperparameter setting. For MissNET, we use the latent dimensions of 10, 30 and 15 for Synthetic, MotionCapture and Motes, respectively, and we set $\lambda = 1.0, \alpha = 0.5$ for all datasets. We set the correct number of regimes on Synthetic datasets; we vary $K = \{1, 2, 3\}$ for other datasets. We list the detailed hyperparameter settings for the baselines in Appendix B.2.

5.1.5 Evaluation metric. To evaluate the effectiveness, we use Root Mean Square Error (RMSE) of the observed time series [5].

5.2 Results

5.2.1 Q1. Effectiveness. We show the effectiveness of MissNET over baselines in missing value imputation.

Synthetic. Fig. 3 (a) and (b) show the results obtained with Synthetic datasets. SSM and MF-based methods perform worse with PatternB than with PatternA due to the increased complexity of data. DL-based methods, especially BRITS, are less affected thanks to their high modeling power. MissNET significantly outperforms DCMF for PatternB although it produces similar results for PatternA. This is because DCMF fails to capture inter-correlation with PatternB since it can only use one predefined network and cannot afford a change of network. Meanwhile, MissNET can capture the inter-correlation for two different regimes thanks to our regime-switching model. However, MissNET fails to discover the correct transition when the missing rate exceeds 70%, and RMSE becomes similar to DCMF.

MotionCapture and Motes. The results for MotionCapture and Motes datasets are shown in Fig. 3 (c) ~ (l). We can see that MissNET

and DCMF constantly outperform other baselines thanks to their ability to exploit inter-correlation explicitly.

Fig. 4 shows the corresponding critical difference diagram for all missing rates based on the Wilcoxon-Holm method [20], where methods that are not connected by a bold line are significantly different in average rank. This confirms that MissNET significantly outperforms other methods, including DCMF, in average rank. Our algorithm for repeatedly inferring networks and the use of ℓ_1 -norm enables the inference of adequate networks for imputation, contributing to better results than DCMF, which uses cosine similarity as a predefined network that may contain spurious correlations in the presence of missing values. Note that MissNET and DCMF exhibit only minor differences when the missing rate is low (10%) because a plausible predefined network can be calculated from observed data.

Classical Linear and Quadratic baselines are unsuitable for imputing missing blocks since they impute missing values mostly from neighboring observed points and cannot capture temporal patterns when there are large gaps. DL-based methods lack sufficient training data and are not suitable for the data we use here, making them perform particularly poorly at a high missing rate, as also noted in [23]. MF-based methods, SoftImpute and CDRec, have a higher RMSE than SSM-based methods since they do not model the temporal dynamics of the data. TRMF utilizes temporal dependency with the AR model, however, it can only capture certain lags specified on the hyperparameter of the AR model. SSM-based methods are superior in imputation to other groups owing to their ability to capture temporal dependency in latent space. DynaMMo implicitly captures inter-correlation in latent space. Therefore, it is no match for MissNET or DCMF.

Fig. 5 demonstrates the results for the MotionCapture Run dataset (missing rate = 60%). We compare the imputation result for the sensor at RKNE provided by the top five methods in terms of average rank, including MissNET, in Fig. 5 (a). BRITS and SoftImpute fail to capture the dynamics of time series while providing a good fit to observed values. The imputation of DynaMMo is smooth, but

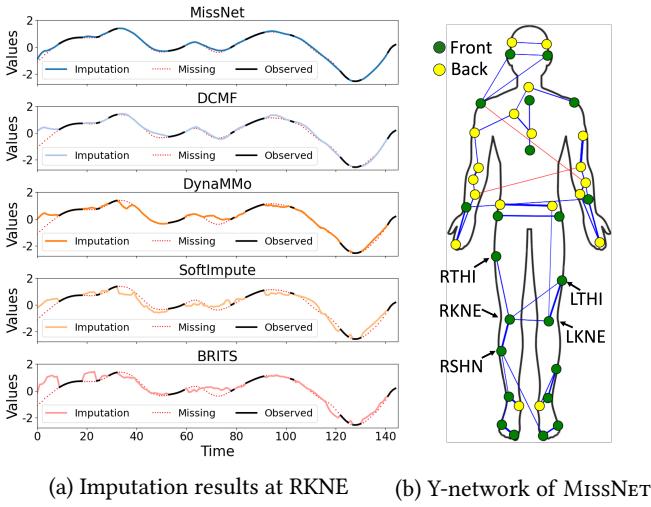


Figure 5: Case study on MotionCapture Run dataset.

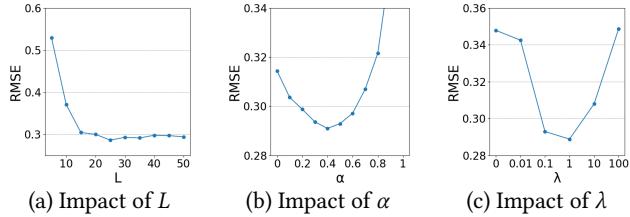


Figure 6: Hyperparameter sensitivity results.

some parts are imprecise since it cannot explicitly exploit inter-correlation. MissNet and DCMF can effectively exploit other observed features associated with RKNE, thereby accurately imputing missing values where other methods fail (e.g., $T = 20 \sim 40, 60 \sim 80$). Fig. 5 (b) shows the sensor network of Y-coordinate values obtained by MissNet plotted on the human body, where a green/yellow dot (node) indicates a sensor placed on the front/back of the body and the thickness and color (blue/red) of the edges are the value and sign (positive/negative) of partial correlations, respectively. We can see that the sensors located close together have edges, meaning they are conditionally dependent given all other sensor values. For example, the sensor at RKNE has edges between RTHI, RSHN, LKNE, and LTHI. They are located to RKNE nearby and show similar dynamics, thus it is reasonable to consider that they are connected. Since MissNet can infer such a meaningful network from partially observed data, the imputation of MissNet is more accurate than that of DCMF.

5.2.2 Hyperparameter sensitivity. We take the Motes dataset and show the impact of hyperparameters: the latent dimension L , the network parameter α , and the sparse parameter λ . We show the mean RMSE of all missing rates.

Latent dimension. Fig. 6 (a) shows the impact of L . As L becomes larger, the model's fitting against the observed data increases. As we can see, the RMSE is constantly decreasing as L increases and stabilizes after 15. This shows that MissNet does not overfit the observed data even for a large L .

Network parameter. α determines the contributions of inter-correlation and temporal dependency to learning \hat{U} . If $\alpha = 0$, the

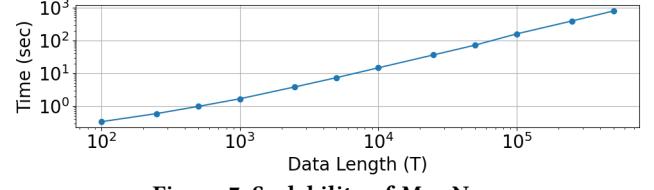


Figure 7: Scalability of MISSNET.

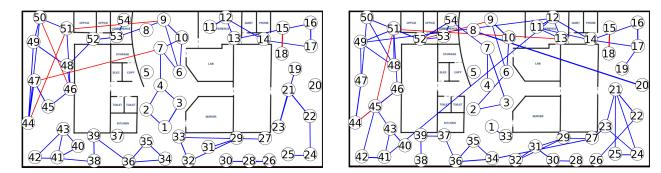
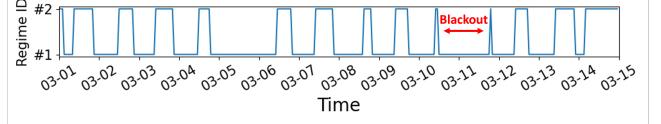


Figure 8: Case study on Motes dataset.

contextual matrix \hat{S} is ignored. If $\alpha = 1$, only \hat{S} is considered for learning \hat{U} . Fig. 6 (b) shows the results of varying α and they are robust except when $\alpha = 1$ ($RMSE = 0.76$). We can see that $\alpha = 0.4$ shows the best result, indicating that both temporal dependency and inter-correlation are important for precise imputation.

Sparse parameter. λ controls the sparsity of the networks $\hat{\Theta}$ through ℓ_1 -norm. The bigger λ becomes, the more sparse the networks become, resulting in MissNet considering only strong interplay. By contrast, when λ is small, MissNet considers insignificant interplays. Fig. 6 (c) shows the impact of λ . We can see that the sparsity of the networks affects the accuracy, and the best λ exists between 0.1 and 10. Thus, the ℓ_1 -norm constraint helps MissNet to exploit important relationships.

5.2.3 Q2. Scalability. We test the scalability of the MissNet algorithm by changing the number of the data length (T) in PatternA. Fig. 7 shows the computation time for one iteration plotted with the data length. As it shows, our proposed MissNet algorithm scales linearly with regard to the data length T .

5.2.4 Q3. Interpretability. We demonstrate how MissNet helps us understand data. We have shown an example with the MotionCapture Run dataset in Fig. 5 (b) where MissNet provides an interpretable network. Here, we demonstrate the results on the Motes dataset (missing rate = 30%) of MissNet ($K = 2$). Fig. 8 (a) shows the regime assignments F , and MissNet mostly assigns night hours to regime #1 and working hours (about 9 am. to 10 pm.) to regime #2, suggesting that they have different networks. Fig. 8 (b) and (c) show the networks for regimes #1 and #2 obtained by MissNet plotted on the building layout. The sensor numbers in the figure are plotted on the actual sensor deployments. As we can see, the two regimes have different networks, and a common feature is that the neighboring sensors tend to form edges, which aligns with our expectations, considering that the sensors measure temperature and, thus, neighboring sensors correlate. The network of regime #2 has more edges than that of #1, and the edges are 1.2 times longer

on average, which might be caused by air convection due to the movement of people during working hours. Consequently, MissNET provides regime assignments and sparse networks, which help us understand how data can be separated and important relationships for imputation.

6 Conclusion

In this paper, we proposed an effective missing value imputation method for multivariate time series, namely MissNET, which captures temporal dependency based on latent space and inter-correlation by the inferred networks while discovering regimes. Our proposed method has the following properties: (a) *Effective*: it outperforms the state-of-the-art algorithms for multivariate time series imputation. (b) *Scalable*: the computation time of MissNET scales linearly with regard to the length of the data. (c) *Interpretable*: it provides sparse networks and regime assignments, which help us understand the important relationships for imputation visually. Our extensive experiments demonstrated the above properties of MissNET.

Acknowledgments

The authors would like to thank the anonymous referees for their valuable comments and helpful suggestions. This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research Number JP21H03446, JP22K17896, NICT JPJ012368C03501, JST-AIP JP-MJCR21U4, JST CREST JPMJCR23M3.

References

- [1] Juan Miguel Lopez Alcaraz and Nils Strodthoff. 2022. Diffusion-based time series imputation and forecasting with structured state space models. *arXiv preprint arXiv:2208.09399* (2022).
- [2] Parikshit Bansal, Prathamesh Deshpande, and Sunita Sarawagi. 2021. Missing Value Imputation on Multidimensional Time Series. *Proc. VLDB Endow.* 14, 11 (jul 2021), 2533–2545.
- [3] Jeroen Berrevoets, Fergus Imrie, Trent Kyono, James Jordon, and Mihaela van der Schaa. 2023. To impute or not to impute? missing data in treatment effect estimation. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 3568–3590.
- [4] Stephen P. Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. 2011. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Found. Trends Mach. Learn.* 3, 1 (2011), 1–122.
- [5] Yongjie Cai, Hanghang Tong, Wei Fan, and Ping Ji. 2015. Fast mining of a network of coevolving time series. In *Proceedings of the 2015 SIAM International Conference on Data Mining*. SIAM, 298–306.
- [6] Yongjie Cai, Hanghang Tong, Wei Fan, Ping Ji, and Qing He. 2015. Facets: Fast comprehensive mining of coevolving high-order time series. In *KDD*. 79–88.
- [7] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. 2018. Brits: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems* 31 (2018).
- [8] Stanislas Chambon, Mathieu N Galtier, Pierrick J Arnal, Gilles Wainrib, and Alexandre Gramfort. 2018. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26, 4 (2018), 758–769.
- [9] Xinyu Chen and Lijun Sun. 2021. Bayesian temporal factorization for multidimensional time series prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 9 (2021), 4659–4673.
- [10] Andrea Cini, Ivan Marisca, and Cesare Alippi. 2021. Filling the g_ap_s: Multivariate time series imputation by graph neural networks. *arXiv preprint arXiv:2108.00298* (2021).
- [11] Joel Janeil Dabrowski, Ashfaqur Rahman, Andrew George, Stuart Arnold, and John McCulloch. 2018. State Space Models for Forecasting Water Quality Variables: An Application in Aquaculture Prawn Farming (*KDD '18*). Association for Computing Machinery, New York, NY, USA, 177–185.
- [12] Wenjie Du, David Côté, and Yan Liu. 2023. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications* 219 (2023), 119619.
- [13] Yangxin Fan, Xuanji Yu, Raymond Wieser, David Meakin, Avishai Shaton, Jean-Nicolas Jaubert, Robert Flottemesch, Michael Howell, Jennifer Braid, Laura Bruckman, Roger French, and Yinghui Wu. 2023. Spatio-Temporal Denoising Graph Autoencoders with Data Augmentation for Photovoltaic Data Imputation. *Proc. ACM Manag. Data* 1, 1, Article 50 (may 2023), 19 pages. <https://doi.org/10.1145/3588730>
- [14] Emily Fox, Erik Sudderth, Michael Jordan, and Alan Willsky. 2008. Nonparametric Bayesian learning of switching linear dynamical systems. *Advances in neural information processing systems* 21 (2008).
- [15] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 3 (2008), 432–441.
- [16] Zoubin Ghahramani. 1998. *Learning dynamic Bayesian networks*. Springer Berlin Heidelberg, Berlin, Heidelberg, 168–197.
- [17] N Hairi, Hanghang Tong, and Lei Ying. 2019. NetDyna: mining networked coevolving time series with missing values. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 503–512.
- [18] David Hallac, Youngsuk Park, Stephen P. Boyd, and Jure Leskovec. 2017. Network Inference via the Time-Varying Graphical Lasso. In *KDD*. 205–213.
- [19] David Hallac, Sagar Vare, Stephen P. Boyd, and Jure Leskovec. 2017. Toeplitz Inverse Covariance-Based Clustering of Multivariate Time Series Data. In *KDD*. 215–223.
- [20] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2019. Deep learning for time series classification: a review. *Data mining and knowledge discovery* 33, 4 (2019), 917–963.
- [21] Baoyu Jing, Hanghang Tong, and Yada Zhu. 2021. Network of tensor time series. In *Proceedings of the Web Conference 2021*, 2425–2437.
- [22] Mourad Khayati, Michael Böhlen, and Johann Gamper. 2014. Memory-efficient centroid decomposition for long time series. In *2014 IEEE 30th International Conference on Data Engineering*. IEEE, 100–111.
- [23] Mourad Khayati, Alberto Lerner, Zakhari Tymchenko, and Philippe Cudré-Mauroux. 2020. Mind the gap: an experimental evaluation of imputation of missing values techniques in time series. In *Proceedings of the VLDB Endowment*, Vol. 13. 768–782.
- [24] Mladen Kolar and Eric P Xing. 2012. Estimating sparse precision matrices from data with missing values. In *International Conference on Machine Learning*. 635–642.
- [25] Lei Li, James McCann, Nancy S Pollard, and Christos Faloutsos. 2009. Dynammo: Mining and summarization of coevolving sequences with missing values. In *KDD*. 507–516.
- [26] SHUAI LIU, Xiucheng Li, Gao Cong, Yile Chen, and YUE JIANG. 2022. Multivariate Time-Series Imputation with Disentangled Temporal Representations. In *The Eleventh International Conference on Learning Representations*.
- [27] Yuehua Liu, Tharam Dillon, Wenjin Yu, Wenny Rahayu, and Fahed Mostafa. 2020. Missing value imputation for industrial IoT sensor data with large gaps. *IEEE Internet of Things Journal* 7, 8 (2020), 6855–6867.
- [28] Hao Ma, Haixuan Yang, Michael R Lyu, and Irwin King. 2008. Sorec: social recommendation using probabilistic matrix factorization. In *Proceedings of the 17th ACM conference on Information and knowledge management*. 931–940.
- [29] Solt Kovács Malte Londschen and Peter Bühlmann. 2021. Change-Point Detection for Graphical Models in the Presence of Missing Values. *Journal of Computational and Graphical Statistics* 30, 3 (2021), 768–779.
- [30] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. 2010. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research* 11 (2010), 2287–2322.
- [31] Karthik Mohan, Palma London, Maryam Fazel, Daniela Witten, and Su-In Lee. 2014. Node-Based Learning of Multiple Gaussian Graphical Models. *J. Mach. Learn. Res.* 15, 1 (jan 2014), 445–488.
- [32] Ricardo Pio Monti, Peter Hellyer, David Sharp, Robert Leech, Christoforos Anagnostopoulos, and Giovanni Montana. 2014. Estimating time-varying brain connectivity networks from functional MRI time series. *NeuroImage* 103 (2014), 427–443.
- [33] A. Namaki, A.H. Shirazi, R. Raei, and G.R. Jafari. 2011. Network analysis of a financial market based on genuine correlation and threshold method. *Physica A: Statistical Mechanics and its Applications* 390, 21 (2011), 3835–3841.
- [34] Kohei Obata, Koki Kawabata, Yasuko Matsubara, and Yasushi Sakurai. 2024. Dynamic Multi-Network Mining of Tensor Time Series. In *Proceedings of the ACM on Web Conference 2024 (WWW '24)*. 4117–4127.
- [35] Spiros Papadimitriou, Jimeng Sun, and Christos Faloutsos. 2005. Streaming pattern discovery in multiple time-series. (2005).
- [36] Vladimir Pavlovic, James M Rehg, Tat-Jen Cham, and Kevin P Murphy. 1999. A dynamic Bayesian network approach to figure tracking using learned dynamic models. In *Proceedings of the seventh IEEE international conference on computer vision*. Vol. 1. IEEE, 94–101.
- [37] Zhen Qin, Yibo Zhang, Shuyu Meng, Zhiguang Qin, and Kim-Kwang Raymond Choo. 2020. Imaging and fusing time series for wearable sensor-based human activity recognition. *Information Fusion* 53 (2020), 80–87.
- [38] Xiaobin Ren, Kaiqi Zhao, Patricia Riddle, Katerina Taškova, Liyan Li, and Qingyi Pan. 2023. DAMR: Dynamic Adjacency Matrix Representation Learning for Multivariate Time Series Imputation. *SIGMOD* (2023). <https://doi.org/10.1145/3589333>
- [39] Mark Rogers, Lei Li, and Stuart J Russell. 2013. Multilinear dynamical systems for tensor time series. *Advances in Neural Information Processing Systems* 26 (2013).

- [40] Tolou Shadbahr, Michael Roberts, Jan Stanczuk, Julian Gilbey, Philip Teare, Sören Dittmer, Matthew Thorpe, Ramon Viñas Torné, Evis Sala, Pietro Lió, Mishal Patel, Jacobus Preller, Ian Selby, Anna Breger, Jonathan R. Weir-McCall, Effrossyni Gkrania-Klotsas, Anna Korhonen, Emily Jefferson, Georg Langs, Guang Yang, Helmut Prosch, Judith Babar, Lorena Escudero Sánchez, Marcel Wassin, Markus Holzer, Nicholas Walton, Pietro Lió, James H. F. Rudd, Tuomas Mirtti, Antti Sakari Rannikko, John A. D. Aston, Jing Tang, and Carola-Bibiane Schönlieb. 2023. The impact of imputation quality on machine learning classifiers for datasets with missing values. *Communications Medicine* 3, 1 (2023).
- [41] Satya Narayan Shukla and Benjamin M Marlin. 2021. Multi-time attention networks for irregularly sampled time series. *arXiv preprint arXiv:2101.10318* (2021).
- [42] Nicolas Städler and Peter Bühlmann. 2012. Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing* 22 (2012), 219–235.
- [43] Nicolas Städler and Peter Bühlmann. 2012. Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing* 22 (2012), 219–235.
- [44] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. 2021. Csd: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems* 34 (2021), 24804–24816.
- [45] Federico Tomasi, Veronica Tozzo, and Annalisa Barla. 2021. Temporal Pattern Detection in Time-Varying Graphical Models. In *ICPR*. 4481–4488.
- [46] Federico Tomasi, Veronica Tozzo, Saverio Salzo, and Alessandro Verri. 2018. Latent Variable Time-varying Network Inference. In *KDD*. 2338–2346. <https://doi.org/10.1145/3219819.3220121>
- [47] Veronica Tozzo, Federico Ciech, Davide Garbarino, and Alessandro Verri. 2021. Statistical Models Coupling Allows for Complex Local Multivariate Time Series Analysis. In *KDD*. 1593–1603.
- [48] Veronica Tozzo, Davide Garbarino, and Annalisa Barla. 2020. Missing Values in Multiple Joint Inference of Gaussian Graphical Models. In *Proceedings of the 10th International Conference on Probabilistic Graphical Models (Proceedings of Machine Learning Research, Vol. 138)*, Manfred Jaeger and Thomas Dyrhe Nielsen (Eds.). PMLR, 497–508.
- [49] Andrew Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory* 13, 2 (1967), 260–269.
- [50] Dingyu Wang, Yuchen Yan, Ruizhong Qiu, Yada Zhu, Kaiyu Guan, Andrew Margenot, and Hanghang Tong. 2023. Networked time series imputation via position-aware graph enhanced variational autoencoders. In *KDD*. 2256–2268.
- [51] Xu Wang, Hongbo Zhang, Pengkun Wang, Yudong Zhang, Binwu Wang, Zhengyang Zhou, and Yang Wang. 2023. An Observed Value Consistent Diffusion Model for Imputing Missing Values in Multivariate Time Series. In *KDD*. 2409–2418.
- [52] Xiwen Yi, Yu Zheng, Junbo Zhang, and Tianrui Li. 2016. ST-MVL: filling missing values in geo-sensory time series data. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*.
- [53] Jinsung Yoon, James Jordan, and Mihaela Schaar. 2018. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*. PMLR, 5689–5698.
- [54] Jinsung Yoon, William R Zame, and Mihaela van der Schaar. 2018. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Transactions on Biomedical Engineering* 66, 5 (2018), 1477–1490.
- [55] Hsiang-Fu Yu, Nikhil Rao, and Inderjit S. Dhillon. 2016. Temporal Regularized Matrix Factorization for High-dimensional Time Series Prediction. In *Advances in Neural Information Processing Systems* 28.
- [56] Dejiao Zhang and Laura Balzano. 2016. Global convergence of a grassmannian gradient descent algorithm for subspace estimation. In *Artificial Intelligence and Statistics*. PMLR, 1460–1468.

A Proposed model

A.1 Symbols

Table 2 lists the main symbols we use throughout this paper.

A.2 Complexity analysis

Proof of Lemma 1.

PROOF. The overall time complexity is composed of four parts by taking the most time-consuming part of equations for each iteration considering $N > N_t, L$: the complexity for the inference of Z and F is $O(K^2 \sum_{t=1}^T (L^3 + L^2 N_t + LN_t^2 + N_t^3))$ related to Eq. (13) and Eq. (15); the inference of \dot{V} is $O(KL^2 N^2)$ (Eq. (19)); M step is $O(KTL^2 N)$ related to the calculation of \dot{U} (Eq. (21)); and the update of $\dot{\Theta}$ is $O(KN^3)$ (Eq. (23)). Thus, the overall time complexity is $O(\#iter \cdot (K^2 \sum_{t=1}^T (L^3 + L^2 N_t + LN_t^2 + N_t^3) + KL^2 N^2 + KTL^2 N + KN^3))$. \square

Proof of Lemma 2.

PROOF. The space complexity is composed of three parts: storing input dataset X is $O(TN)$; intermediate values in E step are $O(K^2 TL^2 + KL^2 N)$; and storing parameter set is $O(KN^2)$. Thus, the overall space complexity is $O(TN + K^2 TL^2 + KL^2 N + KN^2)$. \square

Table 2: Symbols and definitions.

Symbol	Definition
X	Multivariate time series $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\} \in \mathbb{R}^{N \times T}$
W	Indicator matrix $W \in \mathbb{R}^{N \times T}$
\tilde{X}	Partially observed multivariate time series $\tilde{X} = W \circ X$
\hat{X}	Inputted multivariate time series
Z	Time series latent states $Z \in \mathbb{R}^{L \times T}$
$V^{(k)}$	Contextual latent factor of k^{th} -regime $V^{(k)} \in \mathbb{R}^{L \times N}$
$S^{(k)}$	Contextual matrix of k^{th} -regime $S^{(k)} \in \mathbb{R}^{N \times N}$
B	Transition matrix $B \in \mathbb{R}^{L \times L}$
$U^{(k)}$	Observation matrix of k^{th} -regime $U^{(k)} \in \mathbb{R}^{N \times L}$
F	Regime assignments $F \in \mathbb{R}^{K \times T}$
$\rho^{(k)}$	Mean vector of k^{th} -regime $\rho^{(k)} \in \mathbb{R}^N$
$\Theta^{(k)}$	Inverse covariance matrix (i.e., network) of k^{th} -regime $\Theta^{(k)} \in \mathbb{R}^{N \times N}$
N	Number of features
T	Number of timesteps
L	Number of latent dimensions
K	Number of regimes
α	Trade-off between temporal dependency and inter-correlation
λ	Parameter for ℓ_1 -norm that regulates network sparsity

A.3 Updating parameters

The parameters are updated as follows:

$$\begin{aligned} B^{new} &= \left(\sum_{t=2}^T \mathbb{E}[z_t z'_{t-1}] \right) \left(\sum_{t=2}^T \mathbb{E}[z_{t-1} z'_{t-1}] \right)^{-1}, \\ z_0^{new} &= \mathbb{E}[z_1], \\ \Psi_0^{new} &= \mathbb{E}[z_1 z'_1] - \mathbb{E}[z_1] \mathbb{E}[z'_1], \end{aligned}$$

$$\begin{aligned}
(\sigma_Z^2)^{\text{new}} &= \frac{1}{(T-1)L} \text{tr} \left(\sum_{t=2}^T \mathbb{E}[z_t z'_t] - B \sum_{t=2}^T \mathbb{E}[z_t z'_t] \right), \\
(\sigma_{X^{(k)}}^2)^{\text{new}} &= \frac{1}{\sum_{t=1}^T (f_t=k) \sum_{i=1}^N W_{i,t}} \\
&\quad \left[\sum_{t=1}^T (f_t=k) \text{tr} \left((\bar{U}_t^{(k)})^{\text{new}} \mathbb{E}[z_t z'_t] (\bar{U}_t^{(k)})^{\text{new}} \right) \right. \\
&\quad \left. + \sum_{t=1}^T (f_t=k) \left((\bar{x}_t)' \bar{x}_t - 2(\bar{x}_t)' ((\bar{U}_t^{(k)})^{\text{new}} \mathbb{E}[z_t]) \right) \right], \\
(\sigma_{S^{(k)}}^2)^{\text{new}} &= \frac{1}{N^2} \left[\sum_{j=1}^N \left(s_j^{(k)'} s_j^{(k)} - 2s_j^{(k)'} ((U^{(k)})^{\text{new}} \mathbb{E}[v_j^{(k)}]) \right) \right. \\
&\quad \left. + \text{tr} \left((U^{(k)})^{\text{new}} \left(\sum_{j=1}^N \mathbb{E}[v_j^{(k)} v_j^{(k)'}] \right) (U^{(k)})^{\text{new}} \right) \right], \\
(\sigma_{V^{(k)}}^2)^{\text{new}} &= \frac{1}{NL} \sum_{j=1}^N \text{tr}(\mathbb{E}[v_j^{(k)} v_j^{(k)'}]). \tag{26}
\end{aligned}$$

B Experiments

B.1 Synthetic data generation

We first generate a latent factor containing a linear trend, a sinusoidal seasonal pattern, and a noise, $Z_{i,t} = \sin(2\pi \frac{\beta}{T} t) + \gamma t + \eta$, s.t. $1 < \beta < 20, 0.3 < |\gamma| < 1, \eta \sim \mathcal{N}(0, 0.3)$, where $Z \in \mathbb{R}^{L \times T}$. We then project the latent factor with object matrix $X = U^{(k)}Z$, where $X \in \mathbb{R}^{N \times T}$, and $U^{(k)} \in \mathbb{R}^{N \times L}$ is a random graph created as follows [31]:

- (1) Set $U^{(k)}$ equal to the adjacency matrix of an Erdős-Rényi directed random graph, where every edge has a 20% chance of being selected.
- (2) Set selected edge $U_{i,j}^{(k)} \sim \text{Uniform}([-0.6, -0.3] \cup [0.3, 0.6])$, where $U_{i,j}^{(k)}$ denotes the weight between variables i and j .

We set $T = 1000, N = 50, L = 10$. We generate two types of synthetic data, PatternA and PatternB, five times each, where $K = 1, 2$, respectively. In PatternB, the regime switches every 200 timesteps.

B.2 Hyperparameters

We describe the hyperparameters of the baselines. For Synthetic datasets, we give a latent dimension of 10 for all baselines. For a fair comparison, we set the latent dimension of the SSM-based methods at the same value as MissNet. For the MF-based methods, we vary the latent dimension $\{3, 5, 10, 15, 20, 30, 40\}$. We vary the AR parameter for TRMF $\{[1, 2, 3, 4, 5], [1, 24]\}$. To learn the DL-based methods, we add 10% of the data as missing values for training the model. We vary the window size $\{16, 32, T\}$. Other hyperparameters are the same as the original codes.

C Discussion

While MissNet achieved superior performance against state-of-the-art baselines in missing value imputation, here, we mention two limitations of MissNet in terms of sparse network inference and data size.

As mentioned in Section 5.2.1, MissNet fails to discover the correct transition when the missing rate exceeds 70%. However, we claim that MissNet failing to discover the correct transition when the missing rate exceeds 70% is reasonable; rather, correctly discovering transition up to 60% is valuable. Several studies [24, 43, 48] tackled the sparse network inference under the existence of missing values. They aim to infer the correct network and, thus, only utilize the observed value for the network inference. Since observing a complete pair at a high missing rate is rare, it is difficult to infer the correct network. Therefore, the maximum missing rate in their experiments is 30%. Although the experimental settings are different from ours, we can say that the task of sparse network inference in the presence of missing values itself is challenging.

As shown in the experiments, MissNet performs well even when a relatively small number of samples (T) and a large number of features (N) since MissNet is a parametric model and we assume the sparse networks to capture inter-correlation. This cannot be achieved by DL models, which contain a massive number of parameters that require a large amount of T , especially when N is large since all the relationships between features need to be learned. However, unlike DL models, the increased number of samples may not greatly improve MissNet's performance as it has a much smaller number of parameters than DL models, even though switching sparse networks increases the model's flexibility.