Model Agnostic Time Series Analysis via Matrix Estimation

ANISH AGARWAL, Massachusetts Institute of Technology, USA
MUHAMMAD JEHANGIR AMJAD, Massachusetts Institute of Technology, USA
DEVAVRAT SHAH, Massachusetts Institute of Technology, USA
DENNIS SHEN, Massachusetts Institute of Technology, USA

We propose an algorithm to impute and forecast a time series by transforming the observed time series into a matrix, utilizing matrix estimation to recover missing values and de-noise observed entries, and performing linear regression to make predictions. At the core of our analysis is a representation result, which states that for a large class of models, the transformed time series matrix is (approximately) low-rank. In effect, this generalizes the widely used Singular Spectrum Analysis (SSA) in the time series literature, and allows us to establish a rigorous link between time series analysis and matrix estimation. The key to establishing this link is constructing a Page matrix with non-overlapping entries rather than a Hankel matrix as is commonly done in the literature (e.g., SSA). This particular matrix structure allows us to provide finite sample analysis for imputation and prediction, and prove the asymptotic consistency of our method. Another salient feature of our algorithm is that it is model agnostic with respect to both the underlying time dynamics and the noise distribution in the observations. The noise agnostic property of our approach allows us to recover the latent states when only given access to noisy and partial observations a la a Hidden Markov Model; e.g., recovering the time-varying parameter of a Poisson process *without knowing* that the underlying process is Poisson. Furthermore, since our forecasting algorithm requires regression with noisy features, our approach suggests a matrix estimation based method—coupled with a novel, non-standard matrix estimation error metric—to solve the error-in-variable regression problem, which could be of interest in its own right. Through synthetic and real-world datasets, we demonstrate that our algorithm outperforms standard software packages (including R libraries) in the presence of missing data as well as high levels of noise.

ACM Reference Format:
Anish Agarwal, Muhammad Jehangir Amjad, Devavrat Shah, and Dennis Shen. 2018. Model Agnostic Time Series Analysis via Matrix Estimation. *Proc. ACM Meas. Anal. Comput. Syst.* 2, 3, Article 40 (December 2018), 39 pages. https://doi.org/10.1145/3287319

1 INTRODUCTION

Time series data is of enormous interest across all domains of life: from health sciences and weather forecasts to retail and finance, time dependent data is ubiquitous. Despite the diversity of applications, time series problems are commonly confronted by the same two pervasive obstacles: interpolation and extrapolation in the presence of noisy and/or missing data. Specifically, we

Authors' addresses: Anish Agarwal, Massachusetts Institute of Technology, 32-D666 Vassar St. Cambridge, MA, 02139, USA, anish90@mit.edu; Muhammad Jehangir Amjad, Massachusetts Institute of Technology, 32-D560 Vassar St. Cambridge, MA, 02139, USA, mamjad@mit.edu; Devavrat Shah, Massachusetts Institute of Technology, 32-D670 Vassar St. Cambridge, MA, 02139, USA, devavrat@mit.edu; Dennis Shen, Massachusetts Institute of Technology, 32-D560 Vassar St. Cambridge, MA, 02139, USA, deshen@mit.edu.

2476-1249/2018/12-ART40 $15.00
https://doi.org/10.1145/3287319

consider a discrete-time setting with $t \in \mathbb{Z}$ representing the time index and $f : \mathbb{Z} \to \mathbb{R}$[1] representing the latent discrete-time time series of interest. For each $t \in [T] := \{1, \dots, T\}$ and with probability $p \in (0, 1]$, we observe the random variable $X(t)$ such that $\mathbb{E}[X(t)] = f(t)$. While the underlying mean signal f is of course strongly correlated, we assume the per-step noise is independent across t and has uniformly bounded variance. Under this setting, we have two objectives: (1) interpolation, i.e., estimate $f(t)$ for all $t \in [T]$; (2) extrapolation, i.e., forecast $f(t)$ for $t > T$. Our interest is in designing a generic method for interpolation and extrapolation that is applicable to a large model class while being agnostic to the time dynamics and noise distribution.

We develop an algorithm based on matrix estimation, a topic which has received widespread attention, especially with the advent of large datasets. In the matrix estimation setting, there is a "parameter" matrix $\boldsymbol{M}$ of interest, and we observe a sparse, corrupted signal matrix $\boldsymbol{X}$ where $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{M}$. The aim then is to recover the entries of $\boldsymbol{M}$ from noisy and partial observations given in $\boldsymbol{X}$. For our purposes, the attractiveness of matrix estimation derives from the property that these methods are fairly model agnostic in terms of the structure of $\boldsymbol{M}$ and distribution of $\boldsymbol{X}$ given $\boldsymbol{M}$. We utilize this key property to develop a model and noise agnostic time series imputation and prediction algorithm.

1.1 Overview of contributions

Time series as a matrix. We transform the time series of observations $X(t)$ for $t \in [T]$ into what is known as the Page matrix (cf. [23]) by placing contiguous segments of size $L > 1$ (an algorithmic hyper-parameter) of the time series into non-overlapping columns; see Figure 1 for a caricature of this transformation.

As the key contribution, we establish that—in expectation—this generated matrix is either exactly or *approximately* low-rank for a large class of models f. Specifically, f can be from the following families:

Linear Recurrent Formulae (LRF): $f(t) = \sum_{g=1}^{G} \alpha_g f(t-g)$.

Compact Support: $f(t) = g(\varphi(t))$ where $\varphi : \mathbb{Z} \to [-C_1, C_1]$ has the form $\varphi(t+s) = \sum_{l=1}^{G} \alpha_l a_l(t) b_l(s)$ with $\alpha_l \in [-C_2, C_2], a_l : \mathbb{Z} \to [0,1], b_l : \mathbb{Z} \to [0,1]$ for some $C_1, C_2 > 0$; and $g : [-C_1, C_1] \to \mathbb{R}$ is $\mathcal{L}$-Lipschitz [2] [3].

Sublinear: $f(t) = g(t)$ where $g : \mathbb{R} \to \mathbb{R}$ and $\left|\frac{dg(s)}{ds}\right| \leq Cs^{-\alpha}$ for some $\alpha, C > 0$, and $\forall s \in \mathbb{R}$.

Over the past decade, the matrix estimation community has developed a plethora of methods to recover an exact or approximately low-rank matrix from its noisy, partial observations in a noise and model agnostic manner. Therefore, by applying such a matrix estimation method to this transformed matrix, we can recover the underlying mean matrix (and thus $f(t)$ for $t \in [T]$) accurately. In other words, we can interpolate and de-noise the original corrupted and incomplete time series without any knowledge of its time dynamics or noise distribution. Theorem 4.1 and Corollary 4.1 provide finite-sample analyses for this method and establish the consistency property of our algorithm, as long as the underlying f satisfies Property 4.1 and the matrix estimation method satisfies Property 2.1. In Section 5, we show that any additive mixture of the three function classes listed above satisfies Property 4.1. Effectively, Theorem 4.1 establishes a *statistical reduction* between

[1] We denote $\mathbb{R}$ as the field of real numbers and $\mathbb{Z}$ as the integers.

[2] We say $g : \mathbb{R} \to \mathbb{R}$ is $\mathcal{L}$-Lipschitz if there exists a $\mathcal{L} \geq 0$ such that $\|g(x) - g(y)\| \leq \mathcal{L}\|x - y\|$ for all $x, y \in \mathbb{R}$ and $\|\cdot\|$ denotes the standard Euclidean norm on $\mathbb{R}$.

[3] It can be verified that if φ is an LRF satisfying $\varphi(t) = \sum_{h=1}^{H} \gamma_h \varphi(t-h)$, then it satisfies the form $\varphi(t+s) = \sum_{g=1}^{G} \alpha_g a_g(t) b_g(s)$ for $G = H$ with appropriately defined constants α_g, functions a_g, b_g; see Proposition D.2 of Appendix D for details.

time series imputation and matrix estimation. Our key contribution with regards to imputation lies in establishing that a large class of time series models (see Section 5) satisfies Property 4.1.

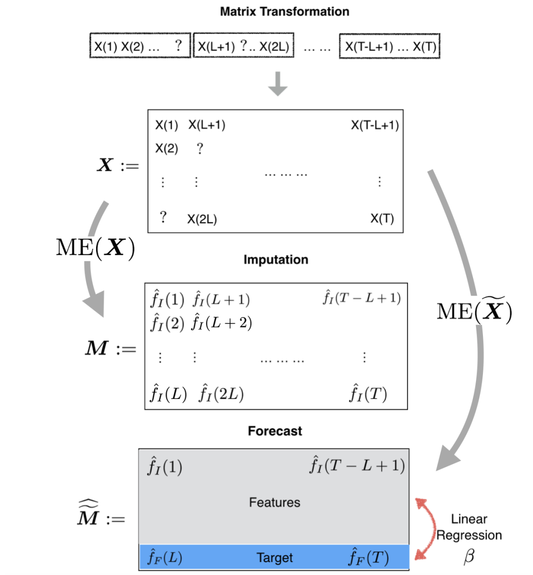

Matrix Transformation
X(1) X(2) ... ?
X(L+1) ?.. X(2L)
... ...
X(T-L+1) ... X(T)
$\boldsymbol{X} :=$
X(1) X(L+1) X(T-L+1)
X(2) ?
... ...
? X(2L) X(T)
$\text{ME}(\boldsymbol{X})$
Imputation
$\boldsymbol{M} :=$
$\hat{f}_I(1)$ $\hat{f}_I(L+1)$ $\hat{f}_I(T-L+1)$
$\hat{f}_I(2)$ $\hat{f}_I(L+2)$
$\hat{f}_I(L)$ $\hat{f}_I(2L)$ $\hat{f}_I(T)$
$\text{ME}(\widetilde{\boldsymbol{X}})$
Forecast
$\widetilde{\widetilde{\boldsymbol{M}}} :=$
$\hat{f}_I(1)$ $\hat{f}_I(T-L+1)$
Features
$\hat{f}_F(L)$ Target $\hat{f}_F(T)$
Linear Regression
β

Fig. 1. Caricature of imputation and forecast algorithms. We first transform the noisy time series $X(t)$ (with "?" indicating missing data) into a Page matrix $\boldsymbol{X}$ with non-overlapping entries. For imputation, we apply a matrix estimation (ME) algorithm with input $\boldsymbol{X}$ to obtain the estimates $\hat{f}_I(t)$ for the de-noised and filled-in entries. For forecasting, we first apply ME to $\widetilde{\boldsymbol{X}}$ (i.e., $\boldsymbol{X}$ excluding the last row), and then fit a linear model β between the last row and all other rows to obtain the forecast estimates $\hat{f}_F(t)$.

It is clear that for LRF, the last row of the mean transformed matrix can be expressed as a linear combination of the other rows. An important representation result of the present paper, which generalizes this notion, is that an *approximate* LRF relationship holds for the other two model classes. Therefore, we can forecast $f(t)$, say for $t = T + 1$, as follows: apply matrix estimation to the transformed data matrix as done in imputation; then, linearly regress the last row with respect to the other rows in the matrix; finally, compute the inner product of the learnt regression vector with the vector containing the previous $L - 1$ values that were estimated via the matrix estimation method. Theorem 4.2 and Corollary 4.2 imply that the mean-squared error of our predictions decays to zero provided the matrix estimation method satisfies Property 2.2 and the underlying model f satisfies Property 4.2. Similar to the case of imputation, establishing that Property 4.2 holds for the three function classes is novel (see Section 5).

Noisy regression. Our proposed forecasting algorithm performs regression with noisy and incomplete features. In the literature, this is known as error-in-variable regression. Recently, there has been exciting progress to understand this problem especially in the *high-dimensional* setting [11, 24, 39]. Our algorithm offers an alternate solution for the high-dimensional setting through the lens of matrix estimation: first, utilize matrix estimation to de-noise and impute the feature observations, and then perform least squares with the pre-processed feature matrix. We demonstrate that if the true, underlying feature matrix is (approximately) low-rank, then our algorithm provides a consistent estimator to the true signal (with finite sample guarantees). Our analysis further suggests the usage of a non-standard error metric, the max row sum error (MRSE) (see Property 2.2 for details).

Class of applicable models. As aforementioned, our algorithm enjoys strong performance guarantees provided the underlying mean matrix induced by the time series f satisfies certain structural properties, i.e., Properties 4.1 and 4.2. We argue that a broad class of commonly used time series models meets the requirements of the three function classes listed above.

LRFs include the following important family of time series: a finite sum of products of exponentials ($\exp\{\alpha t\}$), harmonics ($\cos(2\pi\omega t + \phi)$), and finite degree polynomials ($P_m(t)$) [29], i.e., $f(t) = \sum_{g=1}^{G} \exp\{\alpha_g t\} \cos(2\pi\omega_g t + \phi_g) P_{m_g}(t)$. Further, since stationary processes and L_2 integrable functions are well approximated by a finite summation of harmonics (i.e., sin and cos), LRFs encompass a vitally important family of models. For this model, we show that indeed the structural properties required from the time series matrix for both imputation and prediction are satisfied.

However, there are many important time series models that do not admit a finite order LRF representation. A few toy examples include $\cos(\sin(t))$, $\exp\{\sin^2(t)\}$, $\log t$, $\sqrt{t}$. Time series models with compact support, on the other hand, include models composed of a finite summation of periodic functions (e.g., $\cos(\sin(t))$, $\exp\{\sin^2(t)\}$). Utilizing our low-rank representation result, we establish that models with compact support possess the desired structural properties. We further demonstrate that sublinear functions, which include models that are composed of a finite summation of non (super-)linear functions (e.g., $\log t$, $\sqrt{t}$), also possess the necessary structural properties. Importantly, we argue that the finite mixture of the above processes satisfy the necessary structural properties.

Recovering the hidden state. Our algorithm, being noise and time-dynamics agnostic, makes it relevant to recover the hidden state from its noisy, partial observations as in a Hidden Markov-like Model. For example, imagine having access to partial observations of a time-varying truncated Poisson process[4] *without* knowledge that the process is Poisson. By applying our imputation algorithm, we can recover time-varying parameters of this process accurately and, thus, the hidden states. If we were to apply an Expectation-Maximization (EM) like algorithm, it would require knowledge of the underlying model being Poisson; moreover, theoretical guarantees are not clear for such an approach.

Sample complexity. Given the generality and model agnostic nature of our algorithm, it is expected that its sample complexity for a specific model class will be worse than model aware optimal algorithms. Interestingly, our finite sample analysis suggests that for the model classes stated above, the performance loss incurred due to this generality is minor. See Section 5.6 for a detailed analysis.

Experiments. Using synthetic and real-world datasets, our experiments establish that our method *outperforms existing standard software packages* (including R) for the tasks of interpolation and extrapolation in the presence of noisy and missing observations. When the data is generated synthetically, we "help" the existing software package by choosing the correct parametric model and algorithm while our algorithm remains oblivious to the underlying model; despite this disadvantage, our algorithm continues to *outperform* the standard packages with missing data.

Further, our empirical studies demonstrate that our imputation algorithm accurately recovers the hidden state for Hidden Markov-like Models, verifying our theoretical imputation guarantees (see Theorem 4.1). All experimental findings can be found in Section 6.

[4] Let C denote a positive, bounded constant, and X a Poisson random variable. We define the *truncated* Poisson random variable Y as $Y = \min\{X, C\}$.

1.2 Related works

There are two related topics: matrix estimation and time series analysis. Given the richness of both fields, we cannot do justice in providing a full overview. Instead, we provide a high-level summary of known results with references that provide details.

Matrix estimation. Matrix estimation is the problem of recovering a data matrix from an incomplete and noisy sampling of its entries. This has become of great interest due to its connection to recommendation systems (cf. [18–20, 25, 34–36, 38, 41]), social network analysis (cf. [1–3, 8, 32]), and graph learning (graphon estimation) (cf. [5, 14, 15, 54]). The key realization of this rich literature is that one can estimate the true underlying matrix from noisy, partial observations by simply taking a low-rank approximation of the observed data. We refer an interested reader to recent works such as [14, 19] and references there in.

Time series analysis. The question of time series analysis is potentially as old as civilization in some form. Few textbook style references include [16, 17, 30, 43]. At the highest level, time series modeling primarily involves viewing a given time series as a function indexed by time (integer or real values) and the goal of model learning is to identify this function from observations (over finite intervals). Given that the space of such functions is complex, the task is to utilize function form (i.e., "basis functions") so that for the given setting, the time series observation can fit a sparse representation. For example, in communication and signal processing, the harmonic or Fourier representation of a time series has been widely utilized, due to the fact that signals communicated are periodic in nature. The approximation of stationary processes via harmonics or ARIMA has made them a popular model class to learn stationary-like time series, with domain specific popular variations, such as 'Autoregressive Conditional Heteroskedasticity' (ARCH) in finance. To capture non-stationary or "trend-like" behavior, polynomial bases have been considered. There are rich connections to the theory of stochastic processes and information theory (cf. [22, 28, 42, 47]). Popular time series models with latent structure are Hidden Markov Models (HMM) in probabilistic form (cf. [10, 33] and Recurrent Neural Networks (RNN) in deterministic form (cf. [44]).

The question of learning time series models with missing data has received comparatively less attention. A common approach is to utilize HMMs or general State-Space-Models to learn with missing data (cf. [26, 48]). To the best of the authors' knowledge, most work within this literature is restricted to such class of models (cf. [27]). Recently, building on the literature in online learning, sequential approaches have been proposed to address prediction with missing data (cf. [9]).

Time series and matrix estimation. The use of a matrix structure for time series analysis has roughly two streams of related work: SSA for a single time series (as in our setting), and the use of multiple time series. We discuss relevant results for both of these topics.

Singular Spectrum Analysis (SSA) of time series has been around for some time. Generally, it assumes access to time series data that is not noisy and fully observed. The core steps of SSA for a given time series are as follows: (1) create a Hankel matrix from the time series data; (2) perform a Singular Value Decomposition (SVD) of it; (3) group the singular values based on user belief of the model that generated the process; (4) perform diagonal averaging for the "Hankelization" of the grouped rank-1 matrices outputted from the SVD to create a set of time series; (5) learn a linear model for each "Hankelized" time series for the purpose of forecasting.

At the highest level, SSA and our algorithm are cosmetically similar to one another. There are, however, several key differences: (i) *matrix transformation*—while SSA uses a Hankel matrix (with repeated entries), we transform the time series into a Page matrix (with non-overlapping structure); (ii) *matrix estimation*—SSA heavily relies on the SVD while we utilize general matrix estimation procedures (with SVD methods representing one specific procedural choice); (iii) *linear* *regression*—SSA assumes access to fully observed and noiseless data while we allow for corrupted and missing entries.

These differences are key in being able to derive theoretical results. For example, there have been numerous recent works that have attempted to apply matrix estimation methods to the Hankel matrix inspired by SSA for imputation, but these works do not provide any theoretical guarantees [45, 46, 49]. In effect, the Hankel structure creates strong correlation of noise in the matrix, which is an impediment for proving theoretical results. Our use of the Page matrix overcomes this challenge and we argue that in doing so, we still retain the underlying structure in the matrix. With regards to forecasting, the use of matrix estimation methods that provide guarantees with respect to MRSE rather than standard MSE is needed (which SSA provides no theoretical analysis for). While we do not explicitly discuss such methods in this work, such methods are explored in detail in [4]. With regards to imputation, SSA does not provide direction on how to group the singular values, which is instead done based on user belief of the generating process. However, due to recent advances in matrix estimation literature, there exist algorithms that provide data-driven methods to perform spectral thresholding (cf. [19]). Finally, it is worth nothing that to the best of the authors' knowledge, the classical literature on SSA seem to be lacking finite sample analysis in the presence of noisy observations, which we do provide for our algorithm.

Multiple time series viewed as matrix. In a recent line of work [6, 7, 21, 40, 51, 53], multiple time series have been viewed as a matrix with the primary goal of imputing missing values or de-noising them. Some of these works also require prior model assumptions on the underlying time series. For example in [53], as stated in Section 1, the second step of their algorithm changes based on the user's belief in the model that generated the data along with the multiple time series requirement.

In summary, to the best of our knowledge, ours is the first work to give rigorous theoretical guarantees for a matrix estimation inspired algorithm for a *single, univariate* time series.

Recovering the hidden state. The question of recovering the hidden state from noisy observations is quite prevalent and a workhorse of classical systems theory. For example, most of the system identification literature focuses on recovering model parameters of a Hidden Markov Model. While Expectation-Maximization or Baum-Welch are the go-to approaches, there is limited theoretical understanding of it in generality (for example, see a recent work [52] for an overview) and knowledge of the underlying model is required. For instance, [13] proposed an optimization based, statistically consistent estimation method. However, the optimization "objective" encoded knowledge of the precise underlying model.

It is worth comparing our method with a recent work [6] where the authors attempt to recover the hidden time-varying parameter of a Poisson process via matrix estimation. Unlike our work, they require access to multiple time series. In essence, our algorithm provides the solution to the same question *without* requiring access to any other time series!

1.3 Notation

For any positive integer N, let $[N] = \{1, \dots, N\}$. For any vector $v \in \mathbb{R}^n$, we denote its Euclidean (ℓ_2) norm by $\|v\|_2$, and define $\|v\|_2^2 = \sum_{i=1}^n v_i^2$. In general, the ℓ_p norm for a vector v is defined as $\|v\|_p = \left(\sum_{i=1}^n |v_i|^p \right)^{1/p}$.

For a $m \times n$ real-valued matrix $\boldsymbol{A} = [A_{ij}]$, its spectral/operator norm, denoted by $\|\boldsymbol{A}\|$, is defined as $\|\boldsymbol{A}\|_2 = \max_{1 \le i \le k} |\sigma_i|$, where $k = \min\{m, n\}$ and σ_i are the singular values of $\boldsymbol{A}$ (assumed to be in decreasing order and repeated by multiplicities). The Frobenius norm, also known as the Hilbert-Schmidt norm, is defined as $\|\boldsymbol{A}\|_F^2 = \sum_{i=1}^m \sum_{j=1}^n A_{ij}^2 = \sum_{i=1}^k \sigma_i^2$. The max-norm, or sup-norm, is defined as $\|\boldsymbol{A}\|_{\max} = \max_{i,j} \left|A_{ij}\right|$. The Moore-Penrose pseudoinverse $\boldsymbol{A}^{\dagger}$ of $\boldsymbol{A}$ is defined as

$$\boldsymbol{A}^{\dagger} = \sum_{i=1}^{k}(1/\sigma_i)y_i x_i^T, \quad \text{where} \quad \boldsymbol{A} = \sum_{i=1}^{k} \sigma_i x_i y_i^T,$$

with x_i and y_i being the left and right singular vectors of $\boldsymbol{A}$, respectively.

For a random variable X we define its sub-gaussian norm as

$$\|X\|_{\psi_2} = \inf\Big\{t > 0 : \mathbb{E}\exp\big(X^2/t^2\big) \leq 2\Big\}.$$

If $\|X\|_{\psi_2}$ is bounded by a constant, we call X a sub-gaussian random variable.

Let f and g be two functions defined on the same space. We say that $f(x) = O(g(x))$ if and only if there exists a positive real number M and a real number x_0 such that for all $x \geq x_0$, $|f(x)| \leq M|g(x)|$. Similarly, we say $f(x) = \Omega(g(x))$ if and only if for all $x \geq x_0$, $|f(x)| \geq M|g(x)|$.

1.4 Organization

In Section 2, we list the desired properties needed from a matrix estimation estimation method in order to achieve our theoretical guarantees for imputation and prediction. In Section 3, we formally describe the matrix estimation based algorithms we utilize for time series analysis. In Section 4, we identify the required properties of time series models f under which we can provide finite sample analysis for imputation and prediction performance. In Section 5, we list a broad set of time series models that satisfy the properties in Section 4, and we analyze the sample complexity of our algorithm for each of these models. Lastly, in Section 6, we corroborate our theoretical findings with detailed experiments.

2 MATRIX ESTIMATION

2.1 Problem setup

Consider an $m \times n$ matrix $\boldsymbol{M}$ of interest. Suppose we observe a random subset of the entries of a noisy signal matrix $\boldsymbol{X}$, such that $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{M}$. For each $i \in [m]$ and $j \in [n]$, the (i, j)-th entry X_{ij} is a random variable that is observed with probability $p \in (0, 1]$ and is missing with probability $1 - p$, independently of all other entries. Given $\boldsymbol{X}$, the goal is to produce an estimator $\widehat{\boldsymbol{M}}$ that is "close" to $\boldsymbol{M}$. We use two metrics to quantify the estimation error:

(1) mean-squared error,

$$\text{MSE}(\widehat{\boldsymbol{M}}, \boldsymbol{M}) := \mathbb{E}\Big[\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(\hat{M}_{ij} - M_{ij})^2\Big]; \tag{1}$$

(2) max row sum error,

$$\text{MRSE}(\widehat{\boldsymbol{M}}, \boldsymbol{M}) := \mathbb{E}\Big[\frac{1}{\sqrt{n}}\max_{i\in[m]}\Big(\sum_{j=1}^{n}(\hat{M}_{ij} - M_{ij})^2\Big)^{1/2}\Big]. \tag{2}$$

Here, $\hat{M}_{ij}$ and M_{ij} denote the (i, j)-th elements of $\widehat{\boldsymbol{M}}$ and $\boldsymbol{M}$, respectively. We highlight that the MRSE is a non-standard matrix estimation error metric, but we note that it is a stronger notion than the RMSE$(\widehat{\boldsymbol{M}}, \boldsymbol{M})$[5]; in particular, it is easily seen that MRSE$(\widehat{\boldsymbol{M}}, \boldsymbol{M}) \geq$ RMSE$(\widehat{\boldsymbol{M}}, \boldsymbol{M})$. Hence, for any results we prove in Section 4 regarding the MRSE, any known lower bounds for RMSE of matrix estimation algorithms immediately hold for our results. We now give a definition of a matrix estimation algorithm, which will be used in the following sections.

[5] $\text{RMSE}(\widehat{\boldsymbol{M}}, \boldsymbol{M}) := \mathbb{E}\Big[\frac{1}{\sqrt{mn}}\Big(\sum_{i=1}^{m}\sum_{j=1}^{n}(\hat{M}_{ij} - M_{ij})^2\Big)^{1/2}\Big]$.

Definition 2.1. *A matrix estimation algorithm, denoted as ME* : $\mathbb{R}^{m\times n} \longrightarrow \mathbb{R}^{m\times n}$, *takes as input a noisy matrix* $\boldsymbol{X}$ *and outputs an estimator* $\widehat{\boldsymbol{M}}$.

2.2 Required properties of matrix estimation algorithms

As aforementioned, our algorithm (Section 3.3) utilizes matrix estimation as a pivotal "blackbox" subroutine, which enables accurate imputation and prediction in a model and noise agnostic setting. Over the past decade, the field of matrix estimation has spurred tremendous theoretical and empirical research interest, leading to the emergence of a myriad of algorithms including spectral, convex optimization, and nearest neighbor based approaches. Consequently, as the field continues to advance, our algorithm will continue to improve in parallel. We now state the properties needed of a matrix estimation algorithm ME$(\cdot)$ to achieve our theoretical guarantees (formalized through Theorems 4.1 and 4.2); refer to Section 1.3 for matrix norm definitions.

Property 2.1. *Let ME satisfy the following: Define* $\boldsymbol{Y} = [Y_{ij}]$ *where* $Y_{ij} = X_{ij}$ *if* X_{ij} *is observed, and* $Y_{ij} = 0$ *otherwise. Then, for all* $p \geq \max(m,n)^{-1+\zeta}$ *and some* $\zeta \in (0,1)$, *the produced estimator* $\widehat{\boldsymbol{M}} = ME(\boldsymbol{X})$ *satisfies*

$$\left\| \hat{p}\widehat{\boldsymbol{M}} - p\boldsymbol{M} \right\|_F^2 \leq \frac{1}{mn} C_1 \left\| \boldsymbol{Y} - p\boldsymbol{M} \right\| \left\| p\boldsymbol{M} \right\|_*. \tag{3}$$

Here, $\hat{p}$ [6] *denotes the proportion of observed entries in* $\boldsymbol{X}$ *and* C_1 *is a universal constant.*

We argue the two quantities in Property 2.1, $\|\boldsymbol{Y} - p\boldsymbol{M}\|$ and $\|\boldsymbol{M}\|_*$, are natural. $\|\boldsymbol{Y} - p\boldsymbol{M}\|$ quantifies the amount of noise corruption on the underlying signal matrix $\boldsymbol{M}$; for many settings, this norm concentrates well (e.g., a matrix with independent zero-mean sub-gaussian entries scales as $\sqrt{m}+\sqrt{n}$ with high probability [50]). $\|\boldsymbol{M}\|_*$ quantifies the inherent model complexity of the latent signal matrix; this norm is well behaved for an array of situations, including low-rank and Lipschitz matrices (e.g., for low-rank matrices, $\|\boldsymbol{M}\|_*$ scales as $\sqrt{rmn}$ where r is the rank of the matrix, see [19] for bounds on $\|\boldsymbol{M}\|_*$ under various settings). We note the universal singular value thresholding algorithm proposed in [19] is one such algorithm that satisfies Property 2.1. We provide more intuition for why we choose Property 2.1 for our matrix estimation methods in Section 4.2, where we bound the imputation error.

Property 2.2. *Let ME satisfy the following: For all* $p \geq p^*(m,n)$, *the produced estimator* $\widehat{\boldsymbol{M}} = ME(\boldsymbol{X})$ *satisfies*

$$\mathrm{MRSE}(\widehat{\boldsymbol{M}}, \boldsymbol{M}) \leq \delta_3(m,n) \tag{4}$$

where $\lim_{m,n\to\infty} \delta_3(m,n) = 0$.

Property 2.2 requires the normalized max row sum error to decay to zero as we collect more data. While spectral thresholding and convex optimization methods accurately bound the average mean-squared error, minimizing norms akin to the normalized max row sum error require matrix estimation methods to utilize "local" information, e.g., nearest neighbor type methods. For instance, [54] satisfies Property 2.2 for generic latent variable models (which include low-rank models) with $p^*(m,n) = 1$; [36] also satisfies Property 2.2 for $p^*(m,n) \gg \min(m,n)^{-1/2}$; [14] establishes this for low-rank models as long as $p^*(m,n) \gg \min(m,n)^{-1}$.

[6] Precisely, we define $\hat{p} = \max\{\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \mathbb{1}_{X_{ij} \text{ observed}}, \frac{1}{mn}\}$.

3 ALGORITHM

3.1 Notations and definitions

Recall that $X(t)$ denotes the observation at time $t \in [T]$ where $\mathbb{E}[X(t)] = f(t)$. We shall use the notation $X[s : t] = [X(s), \dots, X(t)]$ for any $s \leq t$. Furthermore, we define $L > 1$ to be an algorithmic hyperparameter and $N = \lfloor T/L \rfloor - 1$. For any $L \times N$ matrix $\boldsymbol{A}$, let $A_L = [A_{Lj}]_{j \leq N}$ represent the the last row of $\boldsymbol{A}$. Moreover, let $\widetilde{\boldsymbol{A}} = [A_{ij}]_{i<L, j \leq N}$ denote the $(L-1) \times N$ submatrix obtained by removing the last row of $\boldsymbol{A}$.

3.2 Viewing a univariate time series as a matrix.

We begin by introducing the crucial step of transforming a single, univariate time series into the corresponding Page matrix. Given time series data $X[1 : T]$, we construct L different $L \times N$ matrices $\boldsymbol{X}^{(k)}$ defined as

$$\boldsymbol{X}^{(k)} = [X_{ij}^{(k)}] = [X(i + (j-1)L + (k-1))]_{i \leq L, j \leq N}, \tag{5}$$

where $k \in [L]$[7]. In words, $\boldsymbol{X}^{(k)}$ is obtained by dividing the time series into N non-overlapping contiguous intervals each of length L, thus constructing N columns; for each $k \in [L]$, $\boldsymbol{X}^{(k)}$ is the k-th shifted version with starting value $X(k)$. For the purpose of imputation, we shall only utilize $\boldsymbol{X}^{(1)}$. In the case of forecasting, however, we shall utilize $\boldsymbol{X}^{(k)}$ for all $k \in [L]$. We define $\boldsymbol{M}^{(k)}$ analogously to $\boldsymbol{X}^{(k)}$ using $f(t)$ instead of $X(t)$.

3.3 Algorithm description

We will now describe the imputation and forecast algorithms separately (see Figure 1).
Imputation. Due to the matrix representation $\boldsymbol{X}^{(1)}$ of the time series, the task of imputing missing values and de-noising observed values translates to that of matrix estimation.

(1) Transform the data $X[1 : T]$ into the matrix $\boldsymbol{X}^{(1)}$ via the method outlined in Subsection 3.2.
(2) Apply a matrix estimation method (as in Definition 2.1) to produce $\widehat{\boldsymbol{M}}^{(1)} = \text{ME}(\boldsymbol{X}^{(1)})$.
(3) Produce estimate: $\widehat{f}_I(i + (j-1)L) := \widehat{M}_{ij}^{(1)}$ for $i \in [L]$ and $j \in [N]$.

Forecast. In order to forecast future values, we first de-noise and impute via the procedure outlined above, and then learn a linear relationship between the the last row and the remaining rows through linear regression.

(1) For each $k \in [L]$, apply the imputation algorithm to produce $\widehat{\widetilde{\boldsymbol{M}}}^{(k)}$ from $\widetilde{\boldsymbol{X}}^{(k)}$.
(2) For each $k \in [L]$, define $\hat{\beta}^{(k)} = \arg\min_{v \in \mathbb{R}^{L-1}} \left\| X_L^{(k)} - (\widehat{\widetilde{\boldsymbol{M}}}^{(k)})^T v \right\|_2^2$.
(3) Produce the estimate at time $t > T$ as follows:
  i) Let $v_t = [X(t - L + 1) : X(t-1)]$ and $k = (t \mod L) + 1$.
  ii) Define $\alpha_t = \arg\min_{\alpha \in \mathbb{R}^N} \left\| v_t - \widehat{\widetilde{\boldsymbol{M}}}^{(k)} \alpha \right\|_2^2$.
  iii) Let $v_t^{\text{proj}} = \widehat{\widetilde{\boldsymbol{M}}}^{(k)} \alpha_t$.
  iv) Produce the estimate: $\hat{f}_F(t) = (v_t^{\text{proj}})^T \cdot \hat{\beta}^{(k)}$.

Why $\boldsymbol{X}^{(k)}$ is necessary for forecasting: For imputation, we are attempting to de-noise all observations made up to time T; hence, it suffices to only use $\boldsymbol{X}^{(1)}$ since it contains all of the relevant information. However, in the case of making predictions, we are only creating an estimator for the

[7] Technically, to define each $\boldsymbol{X}^{(k)}$, we need access to $T' = T + L$ time steps of data. To reduce notational overload and since it has no bearing on our theoretical analysis, we let $T' = T$.

last row. Thus, if we take $X^{(1)}$ for instance, then it is not hard to see that our prediction algorithm only produces estimates for $X(L), X(2L), X(3L), \ldots$, and so on. Therefore, we must repeat this procedure L times in order to produce an estimate for each entry.

Choosing the number of rows L: Theorems 4.1 and 4.2 (and the associated corollaries) suggest L should be as large as possible with the requirement $L = o(N)$. Thus, it suffices to let $N = L^{1+\delta}$ for any $\delta > 0$, e.g., $N = L^2 = T^{2/3}$.

4 MAIN RESULTS

4.1 Properties

We now introduce the required properties for the matrices $\boldsymbol{X}^{(k)}$ and $\boldsymbol{M}^{(k)}$ to identify the time series models f for which our algorithm provides an effective method for imputation and prediction. Under these properties, we state Theorems 4.1 and 4.2, which establish the efficacy of our algorithm. The proofs of these theorems can be found in Appendices B and C, respectively. In Section 5, we argue these properties are satisfied for a large class of time series models.

Property 4.1. (r, δ_1)-imputable
Let matrices $\mathbf{X}^{(1)}$ and $\mathbf{M}^{(1)}$ satisfy the following:

A. *For each $i \in [L]$ and $j \in [N]$:*

1. *$X_{ij}^{(1)}$ are independent sub-gaussian random variables[8] satisfying $\mathbb{E}[X_{ij}^{(1)}] = M_{ij}^{(1)}$ and $\left\| X_{ij}^{(1)} \right\|_{\psi_2} \leq \sigma$.*
2. *$X_{ij}^{(1)}$ is observed with probability $p \in (0, 1]$, independent of other entries.*

B. *There exists a matrix $\mathbf{M}_{(r)}$ of rank r such that for $\delta_1 \geq 0$,*

$$\left\| \boldsymbol{M}^{(1)} - \boldsymbol{M}_{(r)} \right\|_{\max} \leq \delta_1.$$

Property 4.2. (C_β, δ_2)-forecastable
For all $k \in [L]$, let matrices $\boldsymbol{X}^{(k)}$ and $\boldsymbol{M}^{(k)}$ satisfy the following:

A. *For each $i \in [L]$ and $j \in [N]$:*

1. *$X_{ij}^{(k)} = M_{ij}^{(k)} + \epsilon_{ij}$, where ϵ_{ij} are independent sub-Gaussian random variables satisfying $\mathbb{E}[\epsilon_{ij}] = 0$ and $Var(\epsilon_{ij}) \leq \sigma^2$.*
2. *$X_{ij}^{(k)}$ is observed with probability $p \in (0, 1]$, independent of other entries.*

B. *There exists a $\beta^{*(k)} \in \mathbb{R}^{L-1}$ with $\left\|\beta^{*(k)}\right\|_1 \leq C_\beta$ for some constant $C_\beta > 0$ and $\delta_2 \geq 0$ such that*

$$\left\| M_L^{(k)} - (\widetilde{\boldsymbol{M}}^{(k)})^T \beta^{*(k)} \right\|_2 \leq \delta_2.$$

For forecasting, we make the more restrictive additive noise assumption since we focus on linear forecasting methods. Such methods generally require additive noise models. If one can construct linear forecasters under less restrictive assumptions, then we should be able to lift the analysis of such a forecaster to our setting in a straightforward way.

[8] Recall that this condition only requires the per-step noise to be independent; the underlying mean time series f remains highly correlated.

4.2 Imputation

The imputation algorithm produces $\hat{f}_I = [\hat{f}_I(t)]_{t=1:T}$ as the estimate for the underlying time series $f = [f(t)]_{t=1:T}$. We measure the imputation error through the relative mean-squared error:

$$\text{MSE}(\hat{f}_I, f) := \frac{\mathbb{E}\left\|\hat{f}_I - f\right\|_2^2}{\|f\|_2^2}. \tag{6}$$

Recall from the imputation algorithm in Section 3.3 that $\boldsymbol{M}^{(1)}$ is the Page matrix corresponding to f and $\widehat{\boldsymbol{M}}^{(1)}$ is the estimate ME produces; i.e. $\widehat{\boldsymbol{M}}^{(1)} = \text{ME}(\boldsymbol{X}^{(1)})$. It is then easy to see that for any matrix estimation method we have

$$\text{MSE}(\hat{f}_I, f) = \frac{\mathbb{E}\left\|\widehat{\boldsymbol{M}}^{(1)} - \boldsymbol{M}^{(1)}\right\|_F^2}{\left\|\boldsymbol{M}^{(1)}\right\|_F^2}. \tag{7}$$

Thus, we can immediately translate the (un-normalized) MSE of *any* matrix estimation method to the imputation error $\text{MSE}(\hat{f}_I, f)$ of the corresponding time series.

However, to highlight how the rank and the low-rank approximation error δ_1 of the underlying mean matrix $\boldsymbol{M}^{(1)}$ (induced by f) affect the error bound, we rely on Property 2.1, which elucidates these dependencies through the quantity $\|\boldsymbol{M}\|_*$. Thus, we have the following theorem that establishes a precise link between time series imputation and matrix estimation methods.

Theorem 4.1. *Assume Property 4.1 holds and ME satisfies Property 2.1. Then for some $C_1, C_2, C_3, c_4 > 0$,*

$$\text{MSE}(\hat{f}_I, f) \leq \frac{C_1\sigma}{p}\left(\frac{LN\delta_1}{\|f\|_2^2} + \frac{\sqrt{rLN\delta_1}}{\|f\|_2^2} + \frac{\sqrt{rN}}{\|f\|_2}\right) + \frac{C_2(1-p)}{pLN} + C_3 e^{-c_4 N}. \tag{8}$$

Theorem 4.1 states that any matrix estimation subroutine ME that satisfies Property 2.1 will accurately filter noisy observations and recover missing values. This is achieved provided that the rank of $\boldsymbol{M}_{(r)}$ and our low-rank approximation error δ_1 are not too large. Note that knowledge of r is not required apriori for many standard matrix estimation algorithms. For instance, [19] does not utilize the rank of $\boldsymbol{M}$ in its estimation procedure; instead, it performs spectral thresholding of the observed data matrix in an adaptive, data-driven manner. Theorem 4.1 implies the following consistency property of $\hat{f}_I$.

Corollary 4.1. *Let the conditions for Theorem 4.1 hold. Let $\|f\|_2^2 = \Omega(T)$ [9]. Further, suppose f is $(C_5 L^{1-\epsilon_2}, C_6 L^{-\epsilon_1})$-imputable for some $\epsilon_1, \epsilon_2 \in (0, 1)$ and $C_5, C_6 > 0$. Then for $p \gg L^{-\min\left(2\epsilon_1, \epsilon_2\right)}$*

$$\lim_{T\to\infty} \text{MSE}(\hat{f}_I, f) = 0.$$

We note that Theorem 4.1 follows in a straightforward manner from Property 2.1 and standard results from random matrix theory [50]. However, we again highlight that our key contribution lies in establishing that the conditions of Corollary 4.1 hold for a large class of time series models (Section 5).

[9] Note the condition $\|f\|_2^2 = \Omega(T)$ is easily satisfied for any time series f by adding a constant shift to every observation $f(t)$.

4.3 Forecast

Recall $\hat{f}_F(t)$ can only utilize information until time $t-1$. For all $k \in [L]$, our forecasting algorithm learns $\hat{\beta}^{(k)}$ with the previous $L-1$ time steps. We measure the forecasting error through:

$$\text{MSE}(\hat{f}_F, f) := \frac{1}{T-L+1}\mathbb{E}\left\|\hat{f}_F - f\right\|_2^2. \tag{9}$$

Here, $\hat{f}_F = [\hat{f}_F(t)]_{t=L:T}$ denotes the vector of forecasted values. The following result relies on a novel analysis of how applying a matrix estimation pre-processing step affects the prediction error of error-in-variable regression problems (in particular, it requires analyzing a non-standard error metric, the MRSE).

Theorem 4.2. *Assume Property 4.2 holds and ME satisfies Property 2.2, with $p \geq p^*(L,N)$*[10]*. Let $\hat{r} := \max_{k\in[L]} \text{rank}(\widehat{\widehat{\boldsymbol{M}}}^{(k)})$. Then,*

$$\text{MSE}(\hat{f}_F, f) \leq \frac{1}{N-1}\left((\delta_2 + \sqrt{C_\beta N}\delta_3)^2 + 2\sigma^2\hat{r}\right).$$

Note that $\hat{r}$ is trivially bounded by $L = o(N)$ by assumption (see Section 3). If the underlying matrix $\boldsymbol{M}$ is low-rank, then ME algorithms such as the USVT algorithm (cf. [19]) will output an estimator with a small $\hat{r}$. However, since our bound holds for general ME methods, we explicitly state the dependence on $\hat{r}$.

In essence, Theorem 4.2 states that any matrix estimation subroutine ME that satisfies Property 2.2 will produce accurate forecasts from noisy, missing data. This is achieved provided the linear model approximation error δ_2 is not too large (recall $\delta_3 = o(1)$ by Property 2.2). Additionally, Theorem 4.2 implies the following consistency property of $\hat{f}_F$.

Corollary 4.2. *Let the conditions for Theorem 4.2 hold. Suppose f is $(C_1, C_2\sqrt{N}L^{-\epsilon_1})$-forecastable for any $\epsilon_1, C_1, C_2 > 0$ and $N = L^{1+\delta}$ for any $\delta > 0$. Then for $p \geq p^*(L,N)$, such that $\lim_{L,N\to\infty}\delta_3(L,N) = 0$ for $p^*(L,N)$,*

$$\lim_{T\to\infty} \text{MSE}(\hat{f}_F, f) = 0.$$

Similar to the case of imputation, a large contribution of this work is in establishing that the conditions of Corollary 4.2 hold for a large class of time series models (Section 5). Effectively, Corollary 4.2 demonstrates that learning a simple linear relationship among the singular vectors of the de-noised matrix is sufficient to drive the empirical error to zero for a broad class of time series models. The simplicity of this linear method suggests that our estimator will have low generalization error, but we leave that as future work.

We should also note that for auto-regressive processes (i.e., $f(t) = \sum_{g=1}^{G}\alpha_g f(t-1) + \epsilon(t)$ where $\epsilon(t)$ is mean zero noise), previous works (e.g., [37]) have already shown that simple linear forecasters are consistent estimators. For such models, it is easy to see that the underling mean matrix $\boldsymbol{M}^{(k)}$ is not (approximately) low-rank, and so it is not necessary to pre-process the data matrix via a matrix estimation subroutine as we propose in Section 3.3.

5 FAMILY OF TIME SERIES THAT FIT OUR FRAMEWORK

In this section, we list out a broad set of time series models that satisfy Properties 4.1 and 4.2, which are required for the results stated in Section 4. The proofs of these results can be found in Appendix D. To that end, we shall repeatedly use the following model types for our observations.

[10] Refer to Section 2.2 for lower bounds on $p^*(L, N)$ for various ME algorithms. The dependence of the bound on p is implicitly captured in δ_3.

Model Type 1. For any $t \in \mathbb{Z}$, let $X(t)$ be a sequence of independent sub-gaussian random variables with $\mathbb{E}[X(t)] = f(t)$ and $\|X(t)\|_{\psi_2} \leq \sigma$. Note the noise on $f(t)$ is generic (e.g., non-additive).

Model Type 2. For $t \in \mathbb{Z}$, let $X(t) = f(t) + \epsilon(t)$ where $\epsilon(t)$ are independent sub-gaussian random variables with $\mathbb{E}[\epsilon(t)] = 0$ and $\mathrm{Var}(\epsilon(t)) \leq \sigma^2$.

5.1 Linear recurrent functions (LRFs)

For $t \in \mathbb{Z}$, let

$$f^{\mathrm{LRF}}(t) = \sum_{g=1}^{G} \alpha_g f(t-g). \tag{10}$$

Proposition 5.1.

(i) Under Model Type 1, *f^{LRF} satisfies Property 4.1 with $\delta_1 = 0$ and $r = G$*[11].

(ii) Under Model Type 2, *f^{LRF} satisfies Property 4.2 with $\delta_2 = 0$ and $C_\beta = CG$ for all $k \in [L]$ where $C > 0$ is an absolute constant.*

By Proposition 5.1, Theorems 4.1 and 4.2 give the following corollaries:

Corollary 5.1. *Under* Model Type 1, *let the conditions of Theorem 4.1 hold. Let $N = L^{1+\delta}$ for any $\delta > 0$. Then for some $C > 0$, if*

$$T \geq C \cdot \left(\frac{G}{\delta_{error}^2} \right)^{2+\delta},$$

we have $\mathrm{MSE}(\hat{f}_I, f^{\mathrm{LRF}}) \leq \delta_{error}$.

Corollary 5.2. *Under* Model Type 2, *let the conditions of Theorem 4.2 hold. Let $N = L^{1+\delta}$ for any $\delta > 0$. Then for some $C > 0$, if*

$$T \geq C \cdot \left(\frac{\sigma^2}{\delta_{error} - G\delta_3^2} \right)^{\frac{2+\delta}{\delta}},$$

we have $\mathrm{MSE}(\hat{f}_F, f^{\mathrm{LRF}}) \leq \delta_{error}$.

We now provide the rank G of an important class of time series methods—a finite sum of the product of polynomials, harmonics, and exponential time series functions.

Proposition 5.2. *Let P_{m_a} be a polynomial of degree m_a. Then,*

$$f(t) = \sum_{a=1}^{A} \exp\{\alpha_a t\} \cos(2\pi\omega_a t + \phi_a) P_{m_a}(t)$$

admits a representation as in (10). Further the order G of $f(t)$ is independent of T, the number of observations, and is bounded by

$$G \leq A(m_{\max} + 1)(m_{\max} + 2)$$

where $m_{\max} = \max_{a \in A} m_a$.

[11] To see this, take $G = 2$ for example. WLOG, let us consider the first column. Then $f(3) = f(2) + f(1)$, which in turn gives $f(4) = f(3) + f(2) = 2f(2) + f(1)$ and $f(5) = f(4) + f(3) = 3f(2) + 2f(1)$. By induction, it is not hard to see that this holds more generally for any finite G.

5.2 Functions with compact support

For $t \in \mathbb{Z}$, let

$$f^{\text{Compact}}(t) = g(\varphi(t)) \tag{11}$$

where $\varphi : \mathbb{Z} \to [-C_1, C_1]$ takes the form $\varphi(t+s) = \sum_{l=1}^{G} \alpha_l a_l(t) b_l(s)$ with $\alpha_l \in [-C_2, C_2]$, $a_l : \mathbb{Z} \to [0,1]$, $b_l : \mathbb{Z} \to [0,1]$; and $g : [-C_1, C_1] \to \mathbb{R}$ is $\mathcal{L}$-Lipschitz for some $C_1, C_2 > 0$.

Proposition 5.3. *For any $\epsilon \in (0,1)$,*

(i) Under Model Type 1, *f^{Compact} satisfies Property 4.1 with $\delta_1 = \frac{C\mathcal{L}}{L^{\epsilon}}$ and $r = L^{G\epsilon}$ for some $C > 0$.*
(ii) Under Model Type 2, *f^{Compact} satisfies Property 4.2 with $\delta_2 = 2\delta_1\sqrt{N}$ and $C_\beta = 1$ for all $k \in [L]$.*

Using Proposition 5.3, Theorems 4.1 and 4.2 immediately lead to the following corollaries.

Corollary 5.3. *Under* Model Type 1, *let the conditions of Theorem 4.1 hold. Let $N = L^{1+\delta}$ for any $\delta > 0$. Then for some $C > 0$ and any $\epsilon \in (0,1)$, if*

$$T \geq C\left(\left(\frac{1}{\delta_{error}}\right)^{\frac{2}{1-G\epsilon}} + \left(\frac{\mathcal{L}}{\delta_{error}}\right)^{\frac{1}{\epsilon}}\right)^{2+\delta},$$

we have $\text{MSE}(\hat{f}_I, f^{\text{LRF}}) \leq \delta_{error}$.

Corollary 5.4. *Under* Model Type 2, *let the conditions of Theorem 4.2 hold. Let $N = L^{1+\delta}$ for any $\delta > 0$. Then for some $C > 0$ and any $\epsilon \in (0,1)$, if*

$$T \geq C\left(\frac{\sigma^2}{\delta_{error} - \left(\frac{\mathcal{L}}{L^{\epsilon}} + \delta_3\right)^2}\right)^{\frac{2+\delta}{\delta}},$$

we have $\text{MSE}(\hat{f}_F, f^{\text{LRF}}) \leq \delta_{error}$.

As the following proposition will make precise, any Lipschitz function of a periodic time series falls into this family.

Proposition 5.4. *Let*

$$f^{\text{Harmonic}}(t) = \sum_{r=1}^{R} \varphi_r\Big(\sin(2\pi\omega_r t + \phi)\Big), \tag{12}$$

where φ_r is $\mathcal{L}_r$-Lipschitz and ω_r is rational, admits a representation as in (11). Let x_{lcm} denote the fundamental period.[12] *Then the Lipschitz constant $\mathcal{L}$ of $f^{\text{Harmonic}}(t)$ is bounded by*

$$\mathcal{L} \leq 2\pi \cdot \max_{r \in R}(\mathcal{L}_r) \cdot \max_{r \in R}(\omega_r) \cdot x_{lcm}.$$

[12] The "fundamental period", x_{lcm}, of $\{\omega_1, \ldots, \omega_G\}$ is the smallest value such that $x_{\text{lcm}}/(q_a/p_a)$ is an integer for all $a \in A$. Let $S \equiv \{q_a/p_a : g \in G\}$ and let p_{lcm} be the least common multiple (LCM) of $\{p_1, \ldots, p_G\}$. Rewriting S as $\left\{\frac{q_1 * p_{\text{lcm}}/p_1}{p_{\text{lcm}}}, \ldots, \frac{q_G * p_{\text{lcm}}/p_G}{p_{\text{lcm}}}\right\}$, we have the set of numerators, $\{q_1 * p_{\text{lcm}}/p_1, \ldots, q_G * p_{\text{lcm}}/p_A\}$ are all integers and we define their LCM as d_{lcm}. It is easy to verify that $x_{\text{lcm}} = d_{\text{lcm}}/p_{\text{lcm}}$ is indeed a fundamental period. As an example, consider $x = \{n, n/2, n/3, \ldots, n/n-1\}$, in which case the above computation results in $x_{\text{lcm}} = n$.

5.3 Finite sum of sublinear trends

Consider $f^{\text{Trend}}(t)$ such that

$$\left|\frac{df^{Trend}(t)}{dt}\right| \leq C_* t^{-\alpha} \tag{13}$$

for some $\alpha, C_* > 0$.

Proposition 5.5. *Let* $\left|\frac{df^{\text{Trend}}(t)}{dt}\right| \leq C_* t^{-\alpha}$ *for some* $\alpha, C_* > 0$. *Then for any* $\epsilon \in (0, \alpha)$,

(i) Under Model Type 1, f^{Trend} *satisfies Property 4.1 with* $\delta_1 = \frac{C_*}{L^{\epsilon/2}}$ *and* $r = L^{\epsilon/\alpha} + \frac{L - L^{\epsilon/\alpha}}{L^{\epsilon/2}}$.

(ii) Under Model Type 2, f^{Trend} *satisfies Property 4.2 with* $\delta_2 = 2\delta_1\sqrt{N}$ *and* $C_\beta = 1$ *for all* $k \in [L]$.

By Proposition 5.5 and Theorems 4.1 and 4.2, we immediately have the following corollaries on the finite sample performance guarantees of our estimators.

Corollary 5.5. *Under* Model Type 1, *let the conditions of Theorem 4.1 hold. Let* $N = L^{1+\delta}$ *for any* $\delta > 0$. *Then for some* $C > 0$, *if*

$$T \geq C \cdot \left(\frac{1}{\delta_{error}^{2(\alpha+1)/\alpha}}\right)^{2+\delta},$$

we have $\text{MSE}(\hat{f}_I, f^{\text{LRF}}) \leq \delta_{error}$.

Corollary 5.6. *Under* Model Type 2, *let the conditions of Theorem 4.2 hold. Let* $N = L^{1+\delta}$ *for any* $\delta > 0$. *Then for some* $C > 0$ *and for any* $\epsilon \in (0, \alpha)$, *if*

$$T \geq C \cdot \left(\frac{\sigma^2}{\delta_{error} - (L^{-\epsilon/2} + \delta_3)^2}\right)^{\frac{2+\delta}{\delta}},$$

we have $\text{MSE}(\hat{f}_F, f^{\text{LRF}}) \leq \delta_{error}$.

Proposition 5.6. *For* $t \in \mathbb{Z}$ *with* $\alpha_b < 1$ *for* $b \in [B]$,

$$f^{Trend}(t) = \sum_{b=1}^{B} \gamma_b t^{\alpha_b} + \sum_{q=1}^{Q} \log(\gamma_q t) \tag{14}$$

admits a representation as in (13).

5.4 Additive mixture of dynamics

We now show that the imputation results hold even when we consider an additive mixture of any of the models described above. For $t \in \mathbb{Z}$, let

$$f^{\text{Mixture}}(t) = \sum_{q=1}^{Q} \rho_q f_q(t). \tag{15}$$

Here, each f_q is such that under *Model Type 1* with $\mathbb{E}[X(t)] = f_q(t)$, Property 4.1 is satisfied with $\delta_1 = \delta_q$ and $r = r_q$ for $q \in [Q]$.

Proposition 5.7. *Under* Model Type 1, f^{Mixture} *satisfies Property 4.1 with* $\delta_1 = \sum_{q=1}^{Q} \rho_q \delta_q$ *and* $r = \sum_{q=1}^{Q} r_q$.

Proposition 5.7 and Corollary 4.1 imply the following.

Corollary 5.7. *Under* Model Type 1, *let the conditions of Theorem 4.1 hold. For each* $q \in [Q]$, *let* $\delta_q \leq C'_q L^{-\epsilon_q}$ *and* $r_q = o(L)$ *for some* $\epsilon_q, C'_q > 0$. *Then,* $\lim_{T\to\infty} \text{MSE}(\hat{f}_I, f^{\text{Mixture}}) = 0$.

In summary, Corollaries 5.1, 5.3, 5.5 and 5.7 imply that for any additive mixture of time series dynamics coming from f^{LRF}, f^{Compact}, and f^{Trend}, the algorithm in Section 3.2 produces a consistent estimator for an appropriate choice of L.

5.5 Hidden State

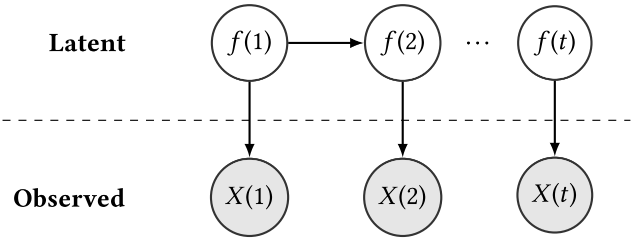


Fig. 2. Hidden State Model with $\mathbb{E}[X(t)] = f(t)$ and $\|X(t)\|_{\psi_2} \leq \sigma$.

A common problem of interest is to uncover the hidden dynamics of latent variables given noisy observations. For example, consider the problem of estimating the true weekly demand rate of umbrellas at a retail store given its weekly sales of umbrellas. This can be mathematically described as uncovering the underlying parameters of a time varying truncated Poisson process [13] whose samples are the weekly sales reports, (cf. [6]). In general, previous methods to learn the hidden states either require multiple time series as inputs or require that the underlying noise model is known (refer to Section 1.2 for a detailed overview).

In contrast, by viewing $f(t)$ as the time-varying latent variables (see Figure 2), we are well equipped to handle more generic noise distributions and complicated hidden dynamics. Specifically, our imputation and forecast algorithms can uncover the latent dynamics if: (i) per-step noise is sub-gaussian (additive noise is needed for forecasting); (ii) $\mathbb{E}[X(t)] = f(t)$. Moreover, our algorithm is model and noise agnostic, robust to missing entries, and comes with strong theoretical consistency guarantees (Theorems 4.1 and 4.2). Given these findings, our approach is likely to become a useful gadget in the toolkit for dealing with scenarios pertinent to uncovering latent states a la Hidden Markov-like models. We corroborate our findings through experiments in Section 6.

5.6 Sample complexity

As discussed, our algorithm operates for a large class of models—it is not tailored for a specific model class (e.g., sum of harmonics). In particular, for a variety of model classes, our algorithm provides consistent estimation for imputation while the forecasting MSE scales with the quality of the matrix estimation algorithm δ_3. Naturally, it is expected that to achieve accurate performance, the number of samples T required will scale relatively poorly compared to model specific optimal algorithms. Corollaries 5.1 - 5.6 provide finite sample analysis that quantifies this "performance loss" and indicates that this loss is minor. As an example, consider imputation for any periodic time series with periods between $[n]$. By proposition 5.2, it is easy to see that the order G of such a time series is $2n$. Thus, corollary 5.1 indicates that the MSE decays to 0 with $T \sim n^{2+\delta}$ for any $\delta > 0$ as $n \longrightarrow \infty$. For such a time series, one expects such a result to require $T \sim n \log n$ even for a model aware optimal algorithm.

[13] Recall that a *truncated* Poisson random variable $Y(t)$ is defined as $Y(t) = \min\{X(t), C\}$, where C denotes a positive, bounded constant and $X(t) = \text{Poisson}(f(t))$.

6 EXPERIMENTS

We conduct experiments on real-world and synthetic datasets to study the imputation and prediction performance of our algorithm for mixtures of time series processes under varying levels of missing data. Additionally, we present the applicability of our algorithm to the hidden state setting (see Section 5.5).

Mixtures of time series processes. For the synthetically generated datasets, we utilize mixtures of harmonics, trend, and auto-regressive (AR) processes with Gaussian additive noise (since AR is effectively a *noisy* version of LRF). When using real-world datasets, we are unaware of the underlying time series processes; nevertheless, these processes appear to display periodicity, trend, and auto-regression.

Comparisons. For forecasting, we compare our algorithm to the state-of-the-art time series forecasting library of R, which decomposes a time series into stationary auto-regressive, seasonal, and trend components. The library learns each component separately and combines them to produce forecasts. Given that our synthetic and real-world datasets involve additive mixtures of these processes, this serves as a strong baseline to compare against our algorithm. We note that we do not outperform optimal model-aware methods for single model classes with all of the data present, at least as implemented in the R-package. However, these methods are not necessarily optimal with missing data and/or when the data is generated by a mixture of multiple model types, which is the setting in which we see our model agnostic method outperform the R-package. For our imputation experiments, we compare our algorithm against AMELIA II ([31]), which is another R-based package that is widely believed to exhibit excellent imputation performance.

Metric of evaluation. Our metric of comparison is the root mean-squared error (RMSE).

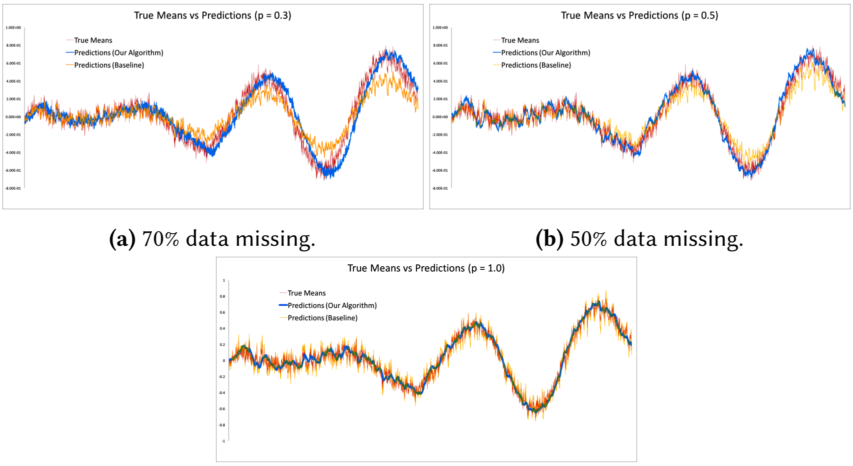
Algorithmic hyper-parameters. For both imputation and forecasting, we apply the Universal Singular Value Thresholding (USVT) algorithm ([19]) as our matrix estimation subroutine. We use a data-driven approach to choose the singular value threshold μ and the number of rows in the time series matrix L in our algorithm. Specifically, we reserve 30% of our training data for cross-validation to pick μ and L.

Summary of results. Details of all experiments are provided below. Recall that p is the probability of observation of each datapoint.

Synthetic data: For forecasting, we determine the forecast RMSE of our algorithm and R's forecast library (see below for how the synthetic data was generated). Our experimental results demonstrate that we outperform R's forecast library, especially under high levels of missing data and noise. For imputation, we outperform the imputation library AMELIA under all levels of missing data.

Real-world data: We test against two real world datasets: (i) Bitcoin price dataset from March 2016 at 30s intervals; (ii) Google flu trends data for Peru from 2003-2012. In both cases, we introduce randomly missing data and then use our algorithm and R's forecast library to forecast into the future. Corroborating the results from the synthetic data experiments, our algorithm's forecast RMSE continues to be lower than that of the R library.

Hidden State Model: We generate a time series according to a Poisson process with latent time-varying parameters. These parameters evolve according to a mixture of time series processes, i.e., sum of harmonics and trends. Our interest is in estimating these time-varying hidden parameters



(a) 70% data missing.

(b) 50% data missing.

(c) No data missing.

Fig. 3. Plots for three levels of missing data ($p \in \{0.3, 0.5, 1\}$) showing the original time series (means) and forecasts produced by the R-library (baseline) and our algorithm.

using one realization of integer observations, of which several are randomly missing. For p ranging from 0.3 to 1.0, the imputation RMSE is always < 0.2 while the R^2 is always > 0.8, which should be considered excellent. This illustrates the versatility of our algorithm in solving a diverse set of problems.

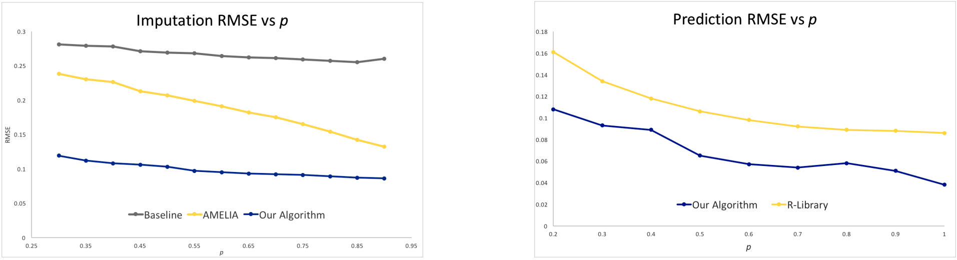
6.1 Synthetically generated data

We generate a mixture process of harmonics, trend, and auto-regressive components. The first 70% of the data points are used to learn a model (training) and point-predictions, i.e., forecasts are performed on the remaining 30% of the data. In order to study the impact of missing data, each entry in the training set is observed independently with probability $p \in (0, 1]$.

Forecasts. Figures 3a-3c visually depict the predictions from our algorithm when compared to the state-of-the-art time series forecasting library in R. We provide the R library the number of lags of the AR component to search over, in effect making its job easier. It is noticeable that the forecasts from the R library always experience higher variance. As p becomes smaller, the R library's forecasts also contain an apparent bias. These visual findings are confirmed in Figure 4b, which shows that our algorithm produces a lower RMSE than that of the R forecasting library when working with mixtures of AR, harmonic, and trend processes; in particular, our algorithm's RMSE ranges from $[0.03, 0.11]$ vs. $[0.09, 0.16]$ for R's forecasting library.

Imputation. Figure 4a shows that our algorithm outperforms the state-of-the-art AMELIA library for multiple time series imputation under all levels of missing data. The RMSE of our algorithm ranged between $[0.09, 0.13]$ vs. $[0.14, 0.24]$ for AMELIA. Note that AMELIA is much better than the baseline, i.e., imputing all missing entries with the mean.

Note that this experiment involved multiple time series where the outcome variable of interest and the log of its squared power were also included. The additional time series components were included to help AMELIA impute missing values because it is unable to impute missing entries in a



(a) Imputation RMSE (mixture AR, harmonic, trend).

(b) Prediction RMSE (mixture AR, harmonic, trend).

Fig. 4. Plots showing the Imputation and Prediction RMSE as a function of p.

single time series. However, our algorithm *did not* use these additional time series; instead, our algorithm was only given access to the original time series with missing, noisy observations.

6.2 Real-world data

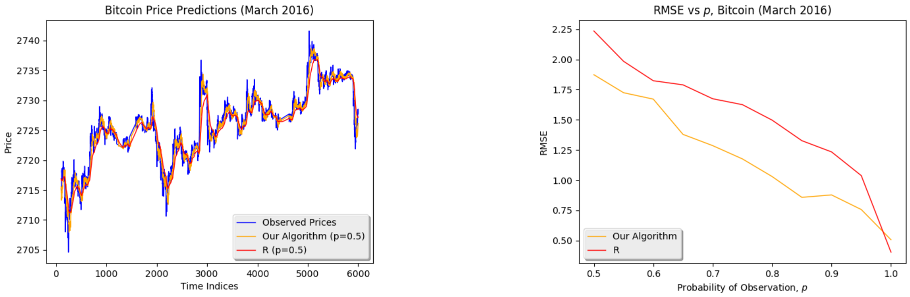
We use two real-world datasets to evaluate the performance of our algorithm in situations where the identities of the time series processes are unknown. This set of experiments is intended to highlight the versatility of our algorithm and applicability to practical scenarios involving time series forecasting. We again highlight that for the following datasets, we do not know the true mean processes. Therefore, it is not possible to generate the metric of interest (RMSE) using the means. Instead, we use the observations themselves as the reference to compute the metric.

Bitcoin. Figures 5a and 5b show the forecasts for Bitcoin prices (in Yuans) in March 2016 at regular 30s time intervals, which demonstrates classical auto-regressive properties. We provide a week's data to learn and forecast over the next two days. Figure 5a shows that our algorithm and the R library appear to do an excellent job of predicting the future even with 50% data missing. Figure 5b shows the RMSE of the predictions for our algorithm and the R library as a function of p; our algorithm had RMSE's in the range $[0.55, 1.85]$ vs $[0.48, 2.25]$ for the R library, for p ranging from 1.0 to 0.5 (note that prices are not normalized). This highlights our algorithm's strength in the presence of missing data.

Google flu trends (Peru). Figures 6a and 6b show the forecasts for Google flu search-trends in Peru which shows significant seasonality. We provide weekly data from 2003-2012 to learn and then forecast for each week in the next three years. Figure 6a shows that our algorithm outperforms R when predicting the future with 30% data missing. Figure 6b shows the RMSE of the predictions as a function of p indicating outperformance of our algorithm under all levels of missing data; our algorithm had RMSE's in range $[8.0, 17.5]$ vs. $[9.0, 26.0]$ for the R library, with p ranging from 1.0 to 0.5 (note that prices are not normalized).

6.3 Hidden state

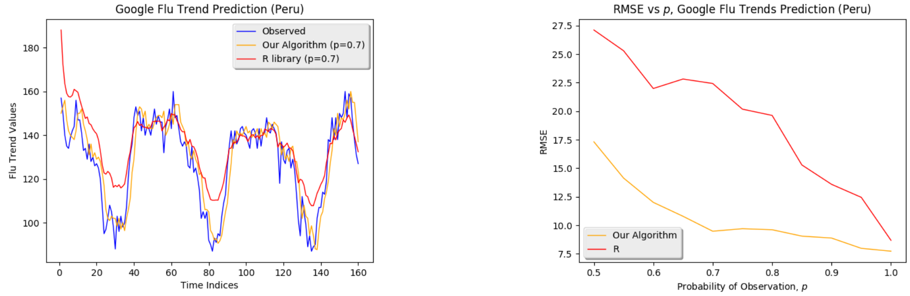
We generate a time series from a Poisson process with time-varying parameters, which are hidden. These parameters evolve according to a mixture of sums of harmonics and trends. Our interest is in estimating these time-varying hidden parameters using one realization of integer observations, of which several are randomly missing. Specifically, each point in the original time series is a Poisson



(a) Price predictions for Bitcoin.

(b) RMSE for Bitcoin predictions.

Fig. 5. Bitcoin price forecasts and RMSE as a function of p.



(a) Flu trends predictions (Peru).

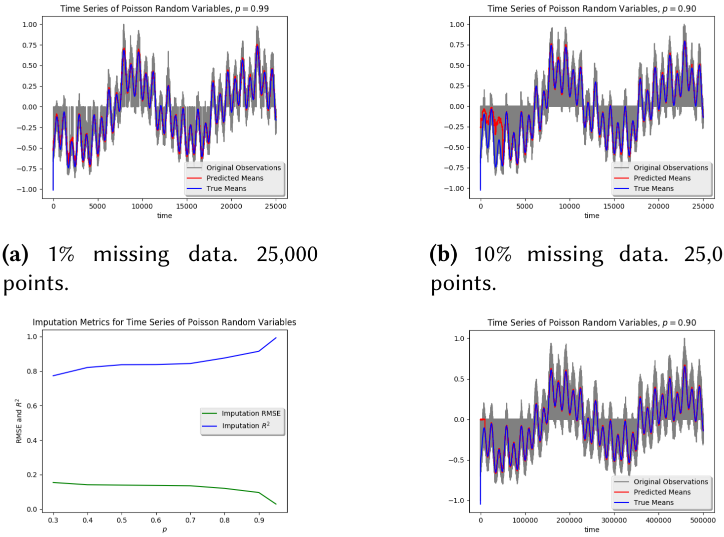
(b) RMSE flu trend predictions.

Fig. 6. Peru's Google flu trends forecasts and RMSE as a function of p.

random variable with parameter $\lambda(t)$, i.e., $X(t) \sim \text{Poisson}(\lambda(t))$. Further, we let $\lambda(t) = f(t)$, where $f(t)$ is a time-dependent sum of harmonics and logarithmic trend components. Each $X(t)$ is then observed independently with probability p to produce a random variable $Y(t)$. We normalize all parameters and observations to lie between $[-1, 1]$. Observe that $\mathbb{E}[Y(t)] = p\lambda(t)$. Note that this is similar to the settings described earlier in this work. It is important to highlight that we have imposed a generic noise model as opposed to an additive noise model. Our goal is to estimate the mean time series process under randomly missing data profiles.

Figures 7a-7b show the mean time series process can be estimated via imputation using the algorithm proposed in our work. These two plots show the original time series (with randomly missing data points set to 0), the true means and our estimation. With only 1% missing data, our algorithm is able to impute the means accurately with the performance degrading slightly with 10% missing data. We note that these are relatively small datasets with only 25,000 points. Figure 7d shows the same process under 10% missing data but for 50,000 data points. As expected, our algorithm performs better when given access to a greater number of data points.

Figure 7c shows plots of RMSE and R^2 for the imputed means of the process. Note these apply to the smaller time series of 25,000 data points. The metrics are computed only on the data points that were missing. Observe that the R^2 value rises while the RMSE falls as p increases. Both of these profiles confirm our intuition that the imputation improves as a function of p. Overall, our performance is fairly robust (RMSE < 0.2 and $R^2 > 0.8$) under all levels of missing data.



(a) 1% missing data. 25,000 points.

(b) 10% missing data. 25,000 points.

(c) RMSE and R^2 vs p. 25,000 points.

(d) 10% missing data. 50,000 points.

Fig. 7. Imputation of the means of a Poisson time series. The first three plots correspond to the time series with 25,000 data points and a resulting matrix of dimension 50×500. The last figure is for the same process, but with twice as much data and matrix dimensions of 100×500. Note that the randomly missing observations are set to 0 and the entire process is normalized to lie between $[-1, 1]$.

7 CONCLUSION

In this paper, we introduce a novel algorithm for time series imputation and prediction using matrix estimation methods, which allows us to operate in a model and noise agnostic setting. At the same time, we offer an alternate solution to the error-in-variables regression problem through the lens of matrix estimation. We provide finite sample analysis for our algorithm, and identify generic conditions on the time series model class under which our algorithm provides a consistent estimator. As a key contribution, we establish that many popular model classes and their mixtures satisfy these generic conditions. Using synthetic and real-world data, we exhibit the efficacy of our algorithm with respect to a state-of-the-art software implementation available through R. Our experimental results agree with our finite sample analysis. Lastly, we demonstrate that our method can provably recover the hidden state of dynamics, which could be of interest in its own right.

ACKNOWLEDGEMENTS

During this work, the authors were supported in part by a Draper, IDSS WorldQuant, and Thompson Reuters Fellowship, NSF CMMI-1462158, NSF CMMI-1634259 and collaboration with KAIST through MIT project 6937985.

REFERENCES

[1] Emmanuel Abbe and Colin Sandon. 2015. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*. IEEE, 670–688.

[2] Emmanuel Abbe and Colin Sandon. 2015. Recovering communities in the general stochastic block model without knowing the parameters. In *Advances in neural information processing systems*.

[3] Emmanuel Abbe and Colin Sandon. 2016. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic BP, and the information-computation gap. *Advances in neural information processing systems* (2016).

[4] Anish Agarwal, Devavrat Shah, Dennis Shen, and Dogyoon Song. 2018. Supervised Learning in High Dimensions via Matrix Estimation. *Working Paper* (2018).

[5] Edo M Airoldi, Thiago B Costa, and Stanley H Chan. 2013. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*. 692–700.

[6] Muhammad J Amjad and Devavrat Shah. 2017. Censored Demand Estimation in Retail. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 1, 2 (2017), 31.

[7] Muhammad Jehangir Amjad, Devavrat Shah, and Dennis Shen. 2017. Robust synthetic control. *arXiv preprint arXiv:1711.06940* (2017).

[8] Animashree Anandkumar, Rong Ge, Daniel Hsu, and Sham Kakade. 2013. A tensor spectral approach to learning mixed membership community models. In *Conference on Learning Theory*. 867–881.

[9] Oren Anava, Elad Hazan, and Assaf Zeevi. 2015. Online Time Series Prediction with Missing Data. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, David Blei and Francis Bach (Eds.). JMLR Workshop and Conference Proceedings, 2191–2199. http://jmlr.org/proceedings/papers/v37/anava15.pdf

[10] Leonard E Baum and Ted Petrie. 1966. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics* 37, 6 (1966), 1554–1563.

[11] Alexandre Belloni, Mathieu Rosenbaum, and Alexandre B Tsybakov. 2017. Linear and conic programming estimators in high dimensional errors-in-variables models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79, 3 (2017), 939–956.

[12] Sergei Bernstein. 1946. *The Theory of Probabilities*. Gastehizdat Publishing House.

[13] Dimitris Bertsimas, David Gamarnik, and John N Tsitsiklis. 1999. Estimation of time-varying parameters in statistical models: an optimization approach. *Machine Learning* 35, 3 (1999), 225–245.

[14] Christian Borgs, Jennifer Chayes, Christina E Lee, and Devavrat Shah. 2017. Thy Friend is My Friend: Iterative Collaborative Filtering for Sparse Matrix Estimation. In *Advances in Neural Information Processing Systems*. 4718–4729.

[15] Christian Borgs, Jennifer T Chayes, Henry Cohn, and Shirshendu Ganguly. 2015. Consistent nonparametric estimation for heavy-tailed sparse graphs. *arXiv preprint arXiv:1508.06675* (2015).

[16] Jenkins Box and Reinsel. 1994. *Time Series Analysis, Forecasting and Control* (3rd ed.). Prentice Hall, Englewood Clifs, NJ.

[17] Peter J Brockwell and Richard A Davis. 2013. *Time series: theory and methods*. Springer Science & Business Media.

[18] Emmanuel J Candès and Terence Tao. 2010. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory* 56, 5 (2010), 2053–2080.

[19] Sourav Chatterjee. 2015. Matrix estimation by universal singular value thresholding. *The Annals of Statistics* 43, 1 (2015), 177–214.

[20] Yudong Chen and Martin J Wainwright. 2015. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025* (2015).

[21] Zhe Chen and Andrzej Cichocki. 2005. Nonnegative matrix factorization with temporal smoothness and/or spatial decorrelation constraints. In *Laboratory for Advanced Brain Signal Processing, RIKEN, Tech. Rep*.

[22] Thomas M Cover. 1966. *BEHAVIOR OF SEQUENTIAL PREDICTORS OF BINARY SEQUENCES*. Technical Report. DTIC Document.

[23] A.A.H Damen, P.M.J Van den Hof, and A.K Hajdasinskit. 1982. Approximate realization based upon an alternative to the Hankel matrix: the Page matrix. *Systems and Control Letters* 2, 4 (1982), 202.

[24] Abhirup Datta and Hui Zou. 2017. Cocolasso for high-dimensional error-in-variables regression. *The Annals of Statistics* 45, 6 (2017), 2400–2426.

[25] Mark A Davenport, Yaniv Plan, Ewout van den Berg, and Mary Wootters. 2014. 1-bit matrix completion. *Information and Inference* 3, 3 (2014), 189–223.

[26] William Dunsmuir and PM Robinson. 1981. Estimation of time series models in the presence of missing data. *J. Amer. Statist. Assoc.* 76, 375 (1981), 560–568.

[27] James Durbin and Siem Jan Koopman. 2012. *Time series analysis by state space methods*. Vol. 38. OUP Oxford.

[28] Meir Feder, Neri Merhav, and Michael Gutman. 1992. Universal prediction of individual sequences. *Information Theory, IEEE Transactions on* 38, 4 (1992), 1258–1270.

[29] Nina Golyandina, Vladimir Nekrutkin, and Anatoly A Zhigljavsky. 2001. *Analysis of time series structure: SSA and related techniques*. Chapman and Hall/CRC.

[30] James Douglas Hamilton. 1994. *Time series analysis*. Vol. 2. Princeton university press Princeton.

[31] James Honaker, Gary King, and Matthew Blackwell. 2015. AMELIA II: A Program for Missing Data. https://cran.r-project.org/web/packages/Amelia/vignettes/amelia.pdf

[32] Samuel B Hopkins and David Steurer. 2017. Efficient Bayesian estimation from few samples: community detection and related problems. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*. IEEE, 379–390.

[33] Rudolph Emil Kalman et al. 1960. A new approach to linear filtering and prediction problems. *Journal of basic Engineering* 82, 1 (1960), 35–45.

[34] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. 2010. Matrix completion from a few entries. *IEEE Transactions on Information Theory* 56, 6 (2010), 2980–2998.

[35] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. 2010. Matrix completion from noisy entries. *Journal of Machine Learning Research* 11, Jul (2010), 2057–2078.

[36] Christina E. Lee, Yihua Li, Devavrat Shah, and Dogyoon Song. 2016. Blind Regression: Nonparametric Regression for Latent Variable Models via Collaborative Filtering. In *Advances in Neural Information Processing Systems 29*. 2155–2163.

[37] Yuval Nardi and Alessandro Rinaldo. 2011. Autoregressive process modeling via the lasso procedure. *Journal of Multivariate Analysis* 102, 3 (2011), 528–549.

[38] Sahand Negahban and Martin J Wainwright. 2011. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics* (2011), 1069–1097.

[39] Loh Po-ling and Martin J Wainwright. 2012. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *The Annals of Statistics* 40 (2012), 1637–1664.

[40] Swati Rallapalli, Lili Qiu, Yin Zhang, and Yi-Chao Chen. 2010. Exploiting temporal stability and low-rank structure for localization in mobile networks. In *Proceedings of the sixteenth annual international conference on Mobile computing and networking*. ACM, 161–172.

[41] Benjamin Recht. 2011. A simpler approach to matrix completion. *Journal of Machine Learning Research* 12, Dec (2011), 3413–3430.

[42] Jorma Rissanen. 1984. Universal coding, information, prediction, and estimation. *Information Theory, IEEE Transactions on* 30, 4 (1984), 629–636.

[43] David S. Stoffer Robert H. Shumway. 2015. *Time Series Analysis and It's Applications* (3rd ed.). Blue Printing.

[44] Jürgen Schmidhuber. 1992. Learning complex, extended sequences using the principle of history compression. *Neural Computation* 4, 2 (1992), 234–242.

[45] David H Schoellhamer. 2001. Singular spectrum analysis for time series with missing data. *Geophysical Research Letters* 28, 16 (2001), 3187–3190.

[46] Y Shen, F Peng, and B Li. 2015. Improved singular spectrum analysis for time series with missing data. *Nonlinear Processes in Geophysics* 22, 4 (2015), 371–376.

[47] Paul C Shields. 1998. The interactions between ergodic theory and information theory. In *IEEE Transactions on Information Theory*. Citeseer.

[48] Robert H Shumway and David S Stoffer. 1982. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of time series analysis* 3, 4 (1982), 253–264.

[49] Grigorios Tsagkatakis, Baltasar Beferull-Lozano, and Panagiotis Tsakalides. 2016. Singular spectrum-based matrix completion for time series recovery and prediction. *EURASIP Journal on Advances in Signal Processing* 2016, 1 (2016), 66.

[50] Roman Vershynin. 2010. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* (2010).

[51] Christopher Xie, Alex Talk, and Emily Fox. 2016. A Unified Framework for Missing Data and Cold Start Prediction for Time Series Data. In *Advances in neural information processing systems Time Series Workshop*.

[52] Fanny Yang, Sivaraman Balakrishnan, and Martin J Wainwright. 2017. Statistical and computational guarantees for the Baum-Welch algorithm. *The Journal of Machine Learning Research* 18, 1 (2017), 4528–4580.

[53] 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*. 847–855.

[54] Yuan Zhang, Elizaveta Levina, and Ji Zhu. 2015. Estimating network edge probabilities by neighborhood smoothing. *arXiv preprint arXiv:1509.08588* (2015).

A USEFUL THEOREMS

Theorem A.1. Bernstein's Inequality. [12]
Suppose that $X_1, \ldots, X_n$ are independent random variables with zero mean, and M is a constant such that $|X_i| \leq M$ with probability one for each i. Let $S := \sum_{i=1}^n X_i$ and $v := Var(S)$. Then for any $t \geq 0$,

$$\mathbb{P}(|S| \geq t) \leq 2\exp\left(-\frac{3t^2}{6v + 2Mt}\right).$$

Theorem A.2. Norm of matrices with sub-gaussian entries. [50]
Let $\boldsymbol{A}$ be an $m \times n$ random matrix whose entries A_{ij} are independent, mean zero, sub-gaussian random variables. Then, for any $t > 0$, we have

$$\|\boldsymbol{A}\| \leq CK(\sqrt{m} + \sqrt{n} + t)$$

with probability at least $1 - 2\exp(-t^2)$. Here, $K = \max_{i,j} \|A_{ij}\|_{\psi_2}$.

B IMPUTATION ANALYSIS

Lemma B.1. *Let $\boldsymbol{X}$ be an $L \times N$ random matrix (with $L \leq N$) whose entries X_{ij} are independent sub-gaussian entries where $\mathbb{E}[X_{ij}] = M_{ij}$ and $\|X_{ij}\|_{\psi_2} \leq \sigma$. Let $\boldsymbol{Y}$ denote the $L \times N$ matrix whose entries Y_{ij} are defined as*

$$Y_{ij} = \begin{cases} X_{ij} & w.p.\ p, \\ 0 & w.p.\ 1-p, \end{cases}$$

for some $p \in (0, 1]$. Let $\hat{p} = \max\left\{\frac{1}{LN}\sum_{i=1}^L \sum_{j=1}^N \mathbb{1}_{X_{ij}\ observed}, \frac{1}{LN}\right\}$. Define events E_1 and E_2 as

$$E_1 := \left\{|\hat{p} - p| \leq p/20\right\}, \tag{16}$$

$$E_2 := \left\{\|\boldsymbol{Y} - p\boldsymbol{M}\| \leq C_1\sigma\sqrt{N}\right\}. \tag{17}$$

Then, for some positive constant c_1

$$\mathbb{P}(E_1) \geq 1 - 2e^{-c_1 LNp} - (1-p)^{LN}, \tag{18}$$

$$\mathbb{P}(E_2) \geq 1 - 2e^{-N}. \tag{19}$$

Proof. Let $\hat{p}_0 = \frac{1}{LN}\sum_{i=1}^L \sum_{j=1}^N \mathbb{1}_{X_{ij}\text{ observed}}$, which implies $\mathbb{E}[\hat{p}_0] = p$. We define the event $E_3 := \{\hat{p}_0 = \hat{p}\}$. Thus, we have that

$$\begin{aligned} \mathbb{P}(E_1^c) &= \mathbb{P}(E_1^c \cap E_3) + \mathbb{P}(E_1^c \cap E_3^c) \\ &= \mathbb{P}(|\hat{p}_0 - p| \geq p/20) + \mathbb{P}(E_1^c \cap E_3^c) \\ &\leq \mathbb{P}(|\hat{p}_0 - p| \geq p/20) + \mathbb{P}(E_3^c) \\ &= \mathbb{P}(|\hat{p}_0 - p| \geq p/20) + (1-p)^{LN}, \end{aligned}$$

where the final equality follows by the independence of observations assumption and the fact that $\hat{p}_0 \neq \hat{p}$ only if we do not have any observations. By Bernstein's Inequality, we have that

$$\mathbb{P}(|\hat{p}_0 - p| \leq p/20) \geq 1 - 2e^{-c_1 LNp}.$$

Furthermore, since $\mathbb{E}[Y_{ij}] = pM_{ij}$, Theorem A.2 yields

$$\mathbb{P}(E_2) \geq 1 - 2e^{-N}.$$

□

Corollary B.1. *Let $E := E_1 \cap E_2$. Then,*

$$\mathbb{P}(E^c) \le C_1 e^{-c_2 N}, \tag{20}$$

where C_1 and c_2 are positive constants independent of L and N.

Proof. By DeMorgan's Law and the Union Bound, we have that

$$\begin{aligned}\mathbb{P}(E^c) &= \mathbb{P}(E_1^c \cup E_2^c)\\ &\le \mathbb{P}(E_1^c) + \mathbb{P}(E_2^c)\\ &\le C_1 e^{-c_2 N},\end{aligned} \tag{21}$$

where $C_1, c_2 > 0$ are appropriately defined, but are independent of L and N. □

Lemma B.2. *Let $\boldsymbol{M}^{(1)}$ be defined as in Section 4.1 and satisfy Property 4.1. Then,*

$$\left\|\boldsymbol{M}^{(1)}\right\|_* \le L\sqrt{N}\delta_1 + \sqrt{rLN}\delta_1 + \sqrt{r}\|\boldsymbol{M}\|_F.$$

Proof. By the definition of $\boldsymbol{M}^{(1)}$ and the triangle inequality property of nuclear norms,

$$\begin{aligned}\left\|\boldsymbol{M}^{(1)}\right\|_* &\le \left\|\boldsymbol{M}^{(1)} - \boldsymbol{M}_{(r)}\right\|_* + \left\|\boldsymbol{M}_{(r)}\right\|_*\\ &\overset{(a)}{\le} \sqrt{L}\left\|\boldsymbol{M}^{(1)} - \boldsymbol{M}_{(r)}\right\|_F + \left\|\boldsymbol{M}_{(r)}\right\|_*\\ &\overset{(b)}{\le} L\sqrt{N}\delta_1 + \left\|\boldsymbol{M}_{(r)}\right\|_*.\end{aligned}$$

Note that (a) makes use of the fact that $\|\boldsymbol{Q}\|_* \le \sqrt{\operatorname{rank}(\boldsymbol{Q})}\|\boldsymbol{Q}\|_F$ for any real-valued matrix $\boldsymbol{Q}$ and (b) utilizes Property 4.1. Since $\operatorname{rank}(\boldsymbol{M}_{(r)}) = r$, we have $\left\|\boldsymbol{M}_{(r)}\right\|_* \le \sqrt{r}\left\|\boldsymbol{M}_{(r)}\right\|_F$. Applying triangle inequality and Property 4.1 again further yields

$$\left\|\boldsymbol{M}_{(r)}\right\|_F \le \left\|\boldsymbol{M}_{(r)} - \boldsymbol{M}\right\|_F + \|\boldsymbol{M}\|_F \le \sqrt{LN}\delta_1 + \|\boldsymbol{M}\|_F.$$

This completes the proof. □

Theorem (4.1). *Assume Property 4.1 holds and ME satisfies Property 2.1. Then for some $C_1, C_2, C_3, c_4 > 0$,*

$$\text{MSE}(\hat{f}_I, f) \le \frac{C_1\sigma}{p}\left(\frac{LN\delta_1}{\|f\|_2^2} + \frac{\sqrt{rLN}\delta_1}{\|f\|_2^2} + \frac{\sqrt{rN}}{\|f\|_2}\right) + \frac{C_2(1-p)}{pLN} + C_3 e^{-c_4 N}.$$

Proof. By (7), it suffices to analyze the time series imputation error by measuring the relative mean-squared error of $\widehat{\boldsymbol{M}}^{(1)}$. For notational simplicity, let us drop the superscripts on $\widehat{\boldsymbol{M}}^{(1)}$ and $\boldsymbol{M}^{(1)}$. Let $E := E_1 \cap E_2$, where E_1 and E_2 are defined as in Lemma B.1. By the law of total probability, we have that

$$\mathbb{E}\left\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\right\|_F^2 \le \mathbb{E}\left[\left\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\right\|_F^2 \mid E\right] + \mathbb{E}\left[\left\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\right\|_F^2 \mid E^c\right]\mathbb{P}(E^c). \tag{22}$$

We begin by bounding the first term on the right-hand side of (22). By Property 2.1 and assuming E occurs, we have that

$$\left\|\hat{p}\widehat{\boldsymbol{M}} - p\boldsymbol{M}\right\|_F^2 \le C_1\|Y - p\boldsymbol{M}\|\,\|p\boldsymbol{M}\|_* \le C_2\sigma\sqrt{N}\,\|\boldsymbol{M}\|_*.$$

Therefore,

$$p^2\left\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\right\|_F^2 \le C_3\hat{p}^2\left\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\right\|_F^2$$

$$\begin{aligned}&\leq C_3\left\|\hat{p}\widehat{\boldsymbol{M}} - p\boldsymbol{M}\right\|_F^2 + C_3(\hat{p}-p)^2\|\boldsymbol{M}\|_F^2\\ &\leq C_4 p\sigma\sqrt{N}\,\|\boldsymbol{M}\|_* + C_3(\hat{p}-p)^2\|f\|_2^2\end{aligned}$$

for an appropriately defined C_4. Observe that $\mathbb{E}(\hat{p}-p)^2 = p(1-p)/LN$. Thus using Corollary B.1 and taking expectations, we obtain

$$\mathbb{E}\left\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\right\|_F^2 \leq C_4 p^{-1}\sigma\sqrt{N}\,\|\boldsymbol{M}\|_* + \frac{C_3(1-p)\|f\|_2^2}{pLN} + C_5\|f\|_2^2 e^{-c_6 N}.$$

Normalizing by $\|f\|_2^2$ gives

$$\mathrm{MSE}(\hat{f}_I, f) \leq \frac{C_4\sigma\sqrt{N}\,\|\boldsymbol{M}\|_*}{p\,\|f\|_2^2} + \frac{C_3(1-p)}{pLN} + C_5 e^{-c_6 N}.$$

Invoking Lemma B.2, we obtain

$$\mathrm{MSE}(\hat{f}_I, f) \leq \frac{C_4\sigma}{p}\left(\frac{LN\delta_1}{\|f\|_2^2} + \frac{\sqrt{r}LN\delta_1}{\|f\|_2^2} + \frac{\sqrt{r}N}{\|f\|_2}\right) + \frac{C_3(1-p)}{pLN} + C_5 e^{-c_6 N}.$$

The proof is complete after relabeling constants.

□

C FORECAST ANALYSIS

Let us begin by analyzing the forecasting error for any $k \in [L]$.

Lemma C.1. *For each $k \in [L]$, assume Property 4.2 holds and $ME(\cdot)$ satisfies Property 2.2. Then,*

$$\mathbb{E}\left[\sum_{t\in S_k}\left(\hat{f}_F(t) - f(t)\right)^2\right] \leq \left(\delta_2 + \sqrt{C_\beta N}\delta_3\right)^2 + 2\sigma^2\hat{r}_k. \tag{23}$$

Here, $S_k := \{t \in [T] : (t \mod L) + 1 = k\}$ and $\hat{r}_k := rank(\widehat{\widetilde{\boldsymbol{M}}}^{(k)})$.

Proof. Observe that we can write

$$\mathbb{E}\left\|M_L^{(k)} - (\widehat{\widetilde{\boldsymbol{M}}}^{(k)})^T\hat{\beta}^{(k)}\right\|_2^2 \equiv \mathbb{E}\left[\sum_{t\in S_k}\left(\hat{f}_F(t) - f(t)\right)^2\right]. \tag{24}$$

For notational simplicity, let $\boldsymbol{Q} := (\widetilde{\boldsymbol{M}}^{(k)})^T$ and $\widehat{\boldsymbol{Q}} := (\widehat{\widetilde{\boldsymbol{M}}}^{(k)})^T$. Similarly, we will drop all superscripts (k) throughout this analysis for notational ease. Recall $X_L = M_L + \epsilon_L$. Then note that by the definition of the optimization in step 2 of the forecast algorithm,

$$\begin{aligned}\left\|X_L - \widehat{\boldsymbol{Q}}\hat{\beta}\right\|_2^2 &\leq \left\|X_L - \widehat{\boldsymbol{Q}}\beta^*\right\|_2^2\\ &= \left\|M_L - \widehat{\boldsymbol{Q}}\beta^*\right\|_2^2 + \|\epsilon_L\|_2^2 + 2\epsilon_L^T(M_L - \widehat{\boldsymbol{Q}}\beta^*).\end{aligned} \tag{25}$$

Moreover,

$$\left\|X_L - \widehat{\boldsymbol{Q}}\hat{\beta}\right\|_2^2 = \left\|M_L - \widehat{\boldsymbol{Q}}\hat{\beta}\right\|_2^2 + \|\epsilon_L\|_2^2 - 2\epsilon_L^T(\widehat{\boldsymbol{Q}}\hat{\beta} - M_L). \tag{26}$$

Combining (25) and (26) and taking expectations, we have

$$\mathbb{E}\left\|M_L - \widehat{\boldsymbol{Q}}\hat{\beta}\right\|_2^2 \leq \mathbb{E}\left\|M_L - \widehat{\boldsymbol{Q}}\beta^*\right\|_2^2 + 2\mathbb{E}[\epsilon_L^T\widehat{\boldsymbol{Q}}(\hat{\beta} - \beta^*)]. \tag{27}$$

Let us bound the final term on the right hand side of (27). Under our independence assumptions, observe that

$$\mathbb{E}[\epsilon_L^T \widehat{\boldsymbol{Q}}]\beta^* = \mathbb{E}[\epsilon_L^T]\mathbb{E}[\widehat{\boldsymbol{Q}}]\beta^* = 0. \tag{28}$$

Recall $\hat{\beta} = \widehat{\boldsymbol{Q}}^\dagger X_L = \widehat{\boldsymbol{Q}}^\dagger M_L + \widehat{\boldsymbol{Q}}^\dagger \epsilon_L$. Using the cyclic and linearity properties of the trace operator (coupled with similar independence arguments), we further have

$$\begin{aligned}
\mathbb{E}[\epsilon_L^T \widehat{\boldsymbol{Q}}\hat{\beta}] &= \mathbb{E}[\epsilon_L^T \widehat{\boldsymbol{Q}}\widehat{\boldsymbol{Q}}^\dagger]M_L + \mathbb{E}[\epsilon_L^T \widehat{\boldsymbol{Q}}\widehat{\boldsymbol{Q}}^\dagger \epsilon_L] \\
&= \mathbb{E}\Big[\mathrm{Tr}\Big(\epsilon_L^T \widehat{\boldsymbol{Q}}\widehat{\boldsymbol{Q}}^\dagger \epsilon_L\Big)\Big] \\
&= \mathbb{E}\Big[\mathrm{Tr}\Big(\widehat{\boldsymbol{Q}}\widehat{\boldsymbol{Q}}^\dagger \epsilon_L \epsilon_L^T\Big)\Big] \\
&= \mathrm{Tr}\Big(\mathbb{E}[\widehat{\boldsymbol{Q}}\widehat{\boldsymbol{Q}}^\dagger] \cdot \mathbb{E}[\epsilon_L \epsilon_L^T]\Big) \\
&\leq \sigma^2 \mathbb{E}\Big[\mathrm{Tr}\Big(\widehat{\boldsymbol{Q}}\widehat{\boldsymbol{Q}}^\dagger\Big)\Big].
\end{aligned} \tag{29}$$

Let $\widehat{\boldsymbol{Q}} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$ be the singular value decomposition of $\widehat{\boldsymbol{Q}}$. Then

$$\begin{aligned}
\widehat{\boldsymbol{Q}}\widehat{\boldsymbol{Q}}^\dagger &= \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{S}^\dagger\boldsymbol{U}^T \\
&= \boldsymbol{U}\tilde{\boldsymbol{I}}\boldsymbol{U}^T.
\end{aligned} \tag{30}$$

Here, $\tilde{\boldsymbol{I}}$ is a block diagonal matrix where its nonzero entries on the diagonal take the value 1. Plugging in (30) into (29), and using the fact that the trace of a square matrix is equal to the sum of its eigenvalues,

$$\sigma^2 \mathbb{E}\Big[\mathrm{Tr}\Big(\widehat{\boldsymbol{Q}}\widehat{\boldsymbol{Q}}^\dagger\Big)\Big] = \sigma^2 \mathbb{E}[\mathrm{rank}(\widehat{\boldsymbol{Q}})]. \tag{31}$$

We now turn our attention to the first term on the right hand side of (27). By Property 4.2, we obtain

$$\begin{aligned}
\Big\| M_L - \widehat{\boldsymbol{Q}}\beta^* \Big\|_2 &= \Big\| M_L - (\boldsymbol{Q} - \boldsymbol{Q} + \widehat{\boldsymbol{Q}})\beta^* \Big\|_2 \\
&\leq \| M_L - \boldsymbol{Q}\beta^* \|_2 + \Big\| (\boldsymbol{Q} - \widehat{\boldsymbol{Q}})\beta^* \Big\|_2 \\
&\leq \delta_2 + \Big\| (\boldsymbol{Q} - \widehat{\boldsymbol{Q}})\beta^* \Big\|_2.
\end{aligned}$$

Thus we have that

$$\mathbb{E}\Big\| (\boldsymbol{Q} - \widehat{\boldsymbol{Q}})\beta^* \Big\|_2 = \mathbb{E}\Big\| (\widetilde{\boldsymbol{M}} - \widehat{\widetilde{\boldsymbol{M}}})^T \beta^* \Big\|_2 \tag{32}$$

$$\leq \sum_{i=1}^{L-1} \left|\beta_i^*\right| \cdot \mathbb{E}\left[\left(\sum_{j=1}^{N} (\hat{M}_{ij} - M_{ij})^2\right)^{1/2}\right] \tag{33}$$

$$\leq \|\beta^*\|_1 \cdot \mathbb{E}\left[\left(\max_{1 \leq i < L} \sum_{j=1}^{N} (\hat{M}_{ij} - M_{ij})^2\right)^{1/2}\right] \tag{34}$$

$$=: C_\beta \sqrt{N} \cdot \mathrm{MRSE}(\widehat{\widetilde{\boldsymbol{M}}}, \widetilde{\boldsymbol{M}}). \tag{35}$$

Putting everything together, we obtain our desired result. □

Theorem (4.2). *Assume Property 4.2 holds and ME satisfies Property 2.2, with $p \geq p^*(L, N)$. Let $\hat{r} := \max_{k \in [L]} \operatorname{rank}(\widehat{\widetilde{\boldsymbol{M}}}^{(k)})$. Then,*

$$\text{MSE}(\hat{f}_F, f) \leq \frac{1}{N-1}\Big((\delta_2 + \sqrt{C_\beta N}\delta_3)^2 + 2\sigma^2 \hat{r}\Big).$$

Proof. For simplicity, define $\delta(k) := (\delta_2 + \sqrt{N}\delta_3)^2 + 2\sigma^2 \hat{r}_k$. By Lemma C.1, for all $k \in [L]$ we have

$$\mathbb{E}\left[\sum_{t \in S_k} \left(\hat{f}_F(t) - f(t)\right)^2\right] \leq \delta(k). \tag{36}$$

Let $\delta_{\max} := (\delta_2 + \sqrt{C_\beta N}\delta_3)^2 + 2\sigma^2 \hat{r}$. Recall $S_k := \{t \in [T] : (t \mod L) + 1 = k\}$. Then, it follows that

$$\text{MSE}(\hat{f}_F, f) \leq \frac{\delta_{\max}}{N-1}.$$

□

D MODEL ANALYSIS

We first define a somewhat technical Property D.1, that will aid us in proving that the various models in Section 5 satisfy Property 4.1 and 4.2. Recall f is the underlying time series we would like to estimate. Define $\eta_k : \mathbb{Z} \times \mathbb{Z} \to \mathbb{R}$ such that

$$\eta_k(\theta_i, \rho_j) := f(i + (j-1)L + (k-1)), \tag{37}$$

where $\theta_i = i$ and $\rho_j = (j-1)L + (k-1)$.

Intuitively, (37) is representing $f(t)$ as a function of two parameters: $\theta_i = i$ and $\rho_j = (j-1)L+(k-1)$. As a result, we can express f as a latent variable model, a representation which is very amenable to theoretical analysis in the matrix estimation literature. Specifically, $[M_{ij}^{(k)}] = [\eta_k(\theta_i, \rho_j)]$ by the construction of $\boldsymbol{M}^{(k)}$. Effectively, the latent parameters (θ_i, ρ_j) encode the amount of shift in the argument to $f(t)$ so as to obtain the appropriate entry in the matrix $\boldsymbol{M}^{(k)}$.

Property D.1. *For all $k \in [L]$, let matrices $\boldsymbol{X}^{(k)}$ and $\boldsymbol{M}^{(k)}$ satisfy the following:*

A. *For each $i \in [L]$ and $j \in [N]$:*

1. $X_{ij}^{(k)}$ are independent sub-gaussian random variables with $\mathbb{E}[X_{ij}^{(k)}] = M_{ij}^{(k)}$ and $\left\|X_{ij}^{(k)}\right\|_{\psi_2} \leq \sigma$.

2. $X_{ij}^{(k)}$ is observed with probability $p \in (0, 1]$, independently.

B. *There exists $\boldsymbol{M}_{(r)} \in \mathbb{R}^{L \times N}$ such that:*

1. $\boldsymbol{M}_{(r)}$ has r_4 distinct rows where $r_4 < L$.

2. $\left\|\boldsymbol{M}^{(k)} - \boldsymbol{M}_{(r)}\right\|_{\max} \leq \delta_4$.

We begin with Proposition D.1, which motivates the use of linear methods in forecasting.

Proposition D.1. *For all $k \in [L]$, let $\boldsymbol{M}^{(k)}$, defined as in Section 4.1, satisfy Property D.1. Then, there exists a β^* such that*

$$\left\|M_L^{(k)} - (\widetilde{\boldsymbol{M}}^{(k)})^T \beta^*\right\|_2 \leq 2\delta_4\sqrt{N},$$

where $\|\beta^*\|_0 = 1$.

Proof. We drop the dependence on k from $\boldsymbol{M}^{(k)}$ and η_k for notational convenience. Furthermore, we prove it for the case of $k = 1$ since the proofs for a general k follow from identical arguments after first making an appropriate shift in the entries of the matrix of interest. Assume we have access to data from $X[1\colon T + r_4 - 1]$. Let us first construct a matrix with overlapping entries,

$\overline{\boldsymbol{M}} = [\overline{M}_{ij}] = [f(i+j-1)]$, of dimension $L \times (T + r_4 - 1)$. We have $\overline{M}_{ij} = \eta(\bar{\theta}_i, \bar{\rho}_j)$ with $\bar{\theta}_i = i$ and $\bar{\rho}_j = (j-1)$, where η is as defined in (37). By construction, the skew-diagonal entries from left to right of $\overline{\boldsymbol{M}}$ are constant, i.e.,

$$\overline{M}_{ki} := \{\overline{M}_{k-j,i+j} : 1 \leq k - j \leq L, 1 \leq i + j \leq T + r_4 - 1\}. \tag{38}$$

Under this setting, we note that the columns of $\boldsymbol{M}$ are subsets of the columns of $\overline{\boldsymbol{M}}$. Specifically, for all $0 \leq j < N$ and $k \leq L$,

$$\overline{M}_{k,jL+1} = M_{k,j+1}. \tag{39}$$

Analogously to how $\overline{\boldsymbol{M}}$ was constructed with respect to $\boldsymbol{M}$, we define $\overline{\boldsymbol{M}}_{(r)}$ with respect to $\boldsymbol{M}_{(r)}$.

Observe that by construction, every entry within $\overline{\boldsymbol{M}}$ exists within $\boldsymbol{M}$. Hence, $\overline{M}_{i,j} = M_{i',j'}$, $\overline{M}^{(r)}_{i,j} = M^{(r)}_{i',j'}$ for some (i', j'), and

$$\begin{aligned} \left|\overline{M}_{i,j} - \overline{M}^{(r)}_{i,j}\right| &= \left|M_{i,j} - M^{(r)}_{i,j}\right| \\ &\leq \left\|\boldsymbol{M} - \boldsymbol{M}_{(r)}\right\|_{\max} \\ &\leq \delta_4, \end{aligned}$$

where the inequality follows from Condition B.2 of Property D.1.

By Condition B.1 of Property D.1 and applying the Pigeonhole Principle, we observe that within the last $r_4 + 1$ rows of $\boldsymbol{M}_{(r)}$, at least two rows are identical. Without loss of generality, let these two rows be denoted as $M^{(r)}_{L-r_1} = [M^{(r)}_{L-r_1,i}]_{i \leq N}$ and $M^{(r)}_{L-r_2} = [M^{(r)}_{L-r_2,i}]_{i \leq N}$, respectively, where $r_1 \in \{1, \ldots, r_4 - 1\}$, $r_2 \in \{2, \ldots, r_4\}$, and $r_1 < r_2$. Consequently, it must be the case that the same two rows in $\overline{\boldsymbol{M}}_{(r)}$ are also identical; i.e., for all $i \leq T + r_4 - 1$,

$$\overline{M}^{(r)}_{L-r_1,i} = \overline{M}^{(r)}_{L-r_2,i}. \tag{40}$$

Using this fact, we have that for all $i \leq T + r_4 - 1$,

$$\left|\overline{M}_{L-r_1,i} - \overline{M}_{L-r_2,i}\right| \leq \left|\overline{M}_{L-r_1,i} - \overline{M}^{(r)}_{L-r_1,i}\right| + \left|\overline{M}_{L-r_2,i} - \overline{M}^{(r)}_{L-r_2,i}\right| + \left|\overline{M}^{(r)}_{L-r_1,i} - \overline{M}^{(r)}_{L-r_1,i}\right| \leq 2\delta_4, \tag{41}$$

where the last inequality follows from (40) and the construction of $\overline{\boldsymbol{M}}_{(r)}$. Additionally, by the skew-diagonal property of $\overline{\boldsymbol{M}}$ as described above by (38), we necessarily have the following two equalities:

$$\overline{M}_{Li} = \overline{M}_{L-r_1,r_1+i} \tag{42}$$

$$\overline{M}_{L-\Delta_r,i} = \overline{M}_{L-r_2,r_1+i}, \tag{43}$$

where $\Delta_r = r_2 - r_1$. Thus, by (41), (42), and (43), we obtain for all $i \leq T$,

$$\begin{aligned} \left|\overline{M}_{Li} - \overline{M}_{L-\Delta_r,i}\right| &= \left|\overline{M}_{L-r_1,r_1+i} - \overline{M}_{L-r_2,r_1+i}\right| \\ &\leq 2\delta_4. \end{aligned} \tag{44}$$

Thus, applying (39) and (44), we reach our desired result, i.e., for all $i \leq N$,

$$\left|M_{Li} - M_{L-\Delta_r,i}\right| \leq 2\delta_4. \tag{45}$$

Recall $\widetilde{\boldsymbol{M}} = [M_{ij}]_{i<L,j\leq N}$ excludes the last row of $\boldsymbol{M}$. From above, we know that there exists some row $\ell := L - \Delta_r < L$ such that $\|M_L - M_\ell\|_2 \leq 2\delta_4\sqrt{N}$. Clearly, we can express

$$M_\ell = \widetilde{\boldsymbol{M}}^T \beta^*, \tag{46}$$

where $\beta^* \in \mathbb{R}^{L-1}$ is a 1-sparse vector with a single nonzero component of value 1 in the ℓth index. This completes the proof. □

Corollary D.1. *For all $k \in [L]$, let $\boldsymbol{M}^{(k)}$, defined as in Section 4.1, satisfy Property D.1 with δ_4, r_4. Then $\boldsymbol{M}^{(k)}$ obeys,*

(i) Under Model Type 1, *Property 4.1 is satisfied with $\delta_1 = \delta_4$ and $r = r_4$.*
(ii) Under Model Type 2, *Property 4.2 is satisfied with $\delta_2 = 2\delta_4\sqrt{N}$.*

Proof. Condition A of both Property 4.1 and 4.2 is satisfied by definition. (i) Condition B.1, B.2 of Property D.1 together imply Condition B of Property 4.1 for the same δ_1, r_4. (ii) Proposition D.1 implies Condition B of Property 4.2 by scaling δ_4 with $2\sqrt{N}$. □

D.1 Proof of Proposition 5.1

Proposition (5.1).

(i) Under Model Type 1, *f^{LRF} satisfies Property 4.1 with $\delta_1 = 0$ and $r = G$;*
(ii) Under Model Type 2, *f^{LRF} satisfies Property 4.2 with $\delta_2 = 0$ and $C_\beta = C \cdot G$ where $C > 0$ is an absolute constant.*

Proof. Let $f(t) = f^{\mathrm{LRF}}$. By definition of $f(t)$, we have that for all $i \in \{G+1, \ldots, L\}$ and $j \in \{1, \ldots N\}$,

$$\begin{aligned} M_{ij}^{(k)} &= f(i + (j-1)L + (k-1)) \\ &= \sum_{g=1}^{G} \alpha_g f((i-g) + (j-1)L + (k-1)) \\ &= \sum_{g=1}^{G} \alpha_g M_{(i-g)j}^{(k)}. \end{aligned}$$

In particular, $M_{Lj}^{(k)} = \sum_{g=1}^{G} \alpha_g M_{(L-g)j}^{(k)}$ for all $j \in \{1, \ldots N\}$, and so we immediately have condition (ii) of the Proposition with $C = \max_{g \in G} \alpha_g$. Since every row from $G+1, \ldots, L$ is a linear combination of the rows above, the rank of $\boldsymbol{M}^{(k)}$ is at most G. Ergo, we have condition (i) of the Proposition. □

Proposition D.2. *Let $f(t) = f^{LRF}$ be defined as in* (5.1)*. Then, for any given $L \geq 1$ and $N \geq 1$, for all $1 \leq s \leq L$, $1 \leq t \leq N$, f admits decomposition*

$$f(t+s) = \sum_{g=1}^{G} \alpha_g a_g(t) b_g(s) \tag{47}$$

for some scalars α_g and functions $a_g : [L] \to \mathbb{R}$, $b_g : [N] \to \mathbb{R}$.

Proof. Let $T = LN$, consider f restricted to $\{1, \ldots, T = LN\}$. Now, by Proposition 5.1, we have that the rank of $\boldsymbol{M}^{(k)}$ is at most G. Thus, the singular value decomposition of $\boldsymbol{M}^{(k)}$ has the form

$$\boldsymbol{M}^{(k)} = \sum_{g=1}^{G} \alpha_g a_g b_g^T,$$

where α_g are the singular values, and a_g, b_g are the corresponding left and right singular vectors of $\boldsymbol{M}^{(k)}$, respectively. Therefore, the (i,j)-th entry of $\boldsymbol{M}^{(k)}$ has the form

$$M_{ij}^{(k)} = f(i + (j-1)L + (k-1)) = \sum_{g=1}^{G} \alpha_g a_g(i) b_g(j), \tag{48}$$

where $a_g(i)$ corresponds to the i-th entry of the g-th left singular vector, and $b_g(j)$ corresponds to the j-th entry of the g-th right singular vector. Thus, $a_g : [L] \to \mathbb{R}$ and $b_g : [N] \to \mathbb{R}$. □

Corollary (5.1). *Under* Model Type 1*, let the conditions of Theorem 4.1 hold. Let $N = L^{1+\delta}$ for any $\delta > 0$. Then for some $C > 0$, if*

$$T \geq C\left(\frac{G}{\delta_{error}^2}\right)^{2+\delta},$$

we have $\text{MSE}(\hat{f}_I, f^{\text{LRF}}) \leq \delta_{error}$.

Proof. By Proposition 5.1, we have for some $C_1, C_2, C_3, c_4 > 0$

$$\text{MSE}(\hat{f}_I, f^{LRF}) \leq \frac{C_1\sigma}{p}\sqrt{\frac{G}{L}} + C_2\frac{(1-p)}{LNp} + C_3e^{-c_4N}.$$

We require the r.h.s of the term above to be less than δ_{error}. Thus, we have that

$$\begin{aligned}\frac{C_1\sigma}{p}\sqrt{\frac{G}{L}} + C_2\frac{(1-p)}{LNp} + C_3e^{-c_4N} &\overset{(a)}{\leq} C\left(\sqrt{\frac{G}{L}} + \frac{1}{LN} + e^{-c_4N}\right)\\ &\overset{(b)}{\leq} C\left(\sqrt{\frac{G}{L}}\right)\end{aligned}$$

where (a) follows for appropriately defined $C > 0$ and by absorbing p, σ into the constant; (b) follows since $\frac{1}{LN} \leq \frac{G}{L}$ and $e^{-c_4N} \leq \sqrt{\frac{G}{L}}$ for sufficiently large L, N and by redefining C. Hence, it suffices that $\delta_{\text{error}} \geq C\left(\sqrt{\frac{G}{L}}\right) \implies T \geq C\left(\frac{G}{\delta_{\text{error}}^2}\right)^{2+\delta}$. □

Corollary (5.2). *Under* Model Type 2*, let the conditions of Theorem 4.2 hold. Let $N = L^{1+\delta}$ for any $\delta > 0$. Then for some $C > 0$, if*

$$T \geq C\left(\frac{\sigma^2}{\delta_{error} - G\delta_3^2}\right)^{\frac{2+\delta}{\delta}}$$

we have $\text{MSE}(\hat{f}_F, f^{\text{LRF}}) \leq \delta_{error}$.

Proof. By Proposition 5.1, we have

$$MSE(\hat{f}_F, f^{LRF}) \leq \frac{1}{N-1}(G\delta_3^2\,N + 2\sigma^2\hat{r}).$$

We require the r.h.s of the term above to be less than δ_{error}. Since $\frac{1}{N}\sigma^2\hat{r} \leq \frac{1}{L^\delta}\sigma^2$, it suffices that

$$\begin{aligned}\delta_{\text{error}} &\overset{(a)}{\geq} C\left(G\delta_3^2 + \frac{1}{L^\delta}\sigma^2\right)\\ \implies L^\delta &\overset{(b)}{\geq} C\left(\frac{\sigma^2}{\delta_{\text{error}} - G\delta_3^2}\right)\\ \implies T &\geq C\left(\frac{\sigma^2}{\delta_{\text{error}} - G\delta_3^2}\right)^{\frac{2+\delta}{\delta}}\end{aligned}$$

where (a) and (b) follow for an appropriately defined $C > 0$. □

D.2 Proof of Proposition 5.2

Proposition (5.2). *Let P_{m_a} be a polynomial of degree m_a. Then,*

$$f(t) = \sum_{a=1}^{A} \exp\{\alpha_a t\} \cos(2\pi\omega_a t + \phi_a) P_{m_a}(t)$$

admits a representation as in (10). Further the order G of $f(t)$ is independent of T, the number of observations, and is bounded by

$$G \leq A(m_{\max} + 1)(m_{\max} + 2)$$

where $m_{\max} = \max_{a \in A} m_a$.

Proof. This proof is adapted from [29]; we state it here for completeness. First, observe that if there exists latent functions $\psi_l : \{1, \ldots, L\} \to \mathbb{R}$ and $\rho_l : \{1, \ldots, N\} \to \mathbb{R}$ for $l \in [G]$ such that for all $(i, j) \in [L] \times [N]$

$$f(i + j) = \sum_{l=1}^{G} \psi_l(i) \rho_l(j), \tag{49}$$

then each $\boldsymbol{M}^{(k)}$ (induced by f for $k \in [L]$) has rank at most G.

Second, observe that time series that admit a representation of the form in (49) form a linear space, which is closed with respect to term-by-term multiplication, i.e.,

$$f(i + j) = f^{(1)} \circ f^{(2)} = \Big(\sum_{l=1}^{G_1} \psi_l^{(1)}(i)\, \rho_l^{(1)}(j) \Big) \Big(\sum_{l=1}^{G_2} \psi_l^{(2)}(i)\, \rho_l^{(2)}(j) \Big), \tag{50}$$

where G_1 and G_2 are the orders of the $f^{(1)}$ and $f^{(2)}$ respectively.

Given the two observations above, it suffices to show separately that $f^{(1)}(t) = \exp\{\alpha t\} \cos(2\pi\omega t + \phi)$ and $f^{(2)}(t) = P_m(t)$ have a representation of the form in (49).

We begin with $f^{(1)}(t) = \exp\{\alpha t\} \cos(2\pi\omega t + \phi)$. For $(i, j) \in [L] \times [N]$,

$$\begin{aligned} f^{(1)}(i + j) &= \exp\{\alpha(i + j)\} \cos(2\pi\omega(i + j) + \phi) \\ &\overset{(a)}{=} \exp\{\alpha i\} \cos(2\pi\omega i) \cdot \exp\{\alpha j\} \cos(2\pi\omega j + \phi) \\ &\quad - \exp\{\alpha i\} \sin(2\pi\omega i) \cdot \exp\{\alpha j\} \sin(2\pi\omega j + \phi) \\ &:= \psi_1(i)\rho_1(j) + \psi_2(i)\rho_2(j), \end{aligned}$$

where in (a) we have used the trigonometric identity $\cos(a + b) = \cos(a)\cos(b) - \sin(a)\sin(b)$. Thus, for $f^{(1)}(t)$, we have $G = 2$.

For $f^{(2)}(t) = P_m(t)$, with $(i, j) \in [L] \times [N]$, we have $P_m(i + j) = \sum_{l=0}^{m} c_l (i + j)^l$. By expanding $(i + j)^l$, it is easily seen (using the Binomial theorem) that there are $l + 1$ unique terms involving powers of i and j. Hence, for $f^{(2)}(t)$, $G \leq \sum_{l=1}^{m+1} l = \frac{(m+1)(m+2)}{2}$ [14].

Now we bound G for $f(t) = \sum_{a=1}^{A} \exp\{\alpha_a t\} \cos(2\pi\omega_a t + \phi_a) P_{m_a}(t)$. For $f^{(1)}(t) = \exp\{\alpha t\} \cos(2\pi\omega t + \phi)$, we have $G^{(1)} = 2$. For $f^{(2)}(t) = P_{m_a}(t)$, we have $G^{(2)} \leq \frac{(m_a+1)(m_a+2)}{2} \leq \frac{(m_{\max}+1)(m_{\max}+2)}{2}$. By (50), it is clear that the order, $G^{(1,2)}$, for $f^{(1)} \circ f^{(2)}$ is bounded by $G^{(1)} \cdot G^{(2)} \leq (m_{\max} + 1)(m_{\max} + 2)$. Since there are A such terms, it follows immediately that for $f(t)$, we have $G \leq A(m_{\max} + 1)(m_{\max} + 2)$, which completes the proof. □

[14] To build intuition, consider $f(t) = t^2$, in which case $f(i + j) = i^2 + j^2 + (2i)(j) := \psi_1(i)\rho_1(j) + \psi_2(i)\rho_2(j) + \psi_3(i)\rho_3(j)$. Here, $G = 3$.

D.3 Proof of Proposition 5.3

Proposition (5.3). *For any $\epsilon \in (0,1)$,*

(i) Under Model Type 1, f^{Compact} *satisfies Property 4.1 with $\delta_1 = \frac{C\mathcal{L}}{L^{\epsilon}}$ and $r = L^{G\epsilon}$ for some $C > 0$.*
(ii) Under Model Type 2, f^{Compact} *satisfies Property 4.2 with $\delta_2 = 2\delta_1\sqrt{N}$ and $C_\beta = 1$.*

Proof. Recall $f^{Compact} = g(\varphi(t))$ where $\varphi : \mathbb{Z} \to [-C_1, C_1]$ takes the form $\varphi(t+s) = \sum_{l=1}^{G} \alpha_l a_l(t) b_l(s)$ with $\alpha_l \in [-C_2, C_2]$, $a_l : \mathbb{Z} \to [0,1]$, $b_l : \mathbb{Z} \to [0,1]$ for some $C_1, C_2 > 0$; and $g : [-C_1, C_1] \to \mathbb{R}$ is $\mathcal{L}$-Lipschitz. Without loss of generality, we drop the dependence of k on η_k to decrease notational overload. Recall that η (as defined in (37)) has row and column parameters $\{\theta_1 \cdots \theta_L\}$ and $\{\rho_1 \cdots \rho_N\}$, which denote shifts in an integer time index.

For some $\delta > 0$, we define the set $P(\frac{\delta}{C_2\mathcal{L}}) \subset [0,1]^G$ such that for all $i \in [0,1]^G$, there exists an $i' \in P(\frac{\delta}{C_2\mathcal{L}})$ where $\|i - i'\|_1 \le \frac{\delta}{C_2\mathcal{L}}$. It is easily shown that we can construct this set such that $\left|P(\frac{\delta}{C_2\mathcal{L}})\right| \le (\frac{3C_2\mathcal{L}}{\delta})^G$.

For any $i \in [L]$, let $\bar{a}(i) = [a_1(i), \dots, a_G(i)]$. Thus, from the construction of $P(\frac{\delta}{C_2\mathcal{L}})$, there must exist an $\bar{a}^*(i) = [a_1^*(i), \dots, a_G^*(i)] \in P(\frac{\delta}{C_2\mathcal{L}})$ such that $\|\bar{a} - \bar{a}^*\|_1 \le \frac{\delta}{C_2\mathcal{L}}$. Therefore, for any $(i,j) \in [L]\times[N]$, we have

$$\begin{aligned}
\left|\eta(i,(j-1)L) - g\Big(\sum_{l=1}^{G} \alpha_l a_l^*(i) b_l((j-1)L)\Big)\right| &= \left|f(i+(j-1)L) - g\Big(\sum_{l=1}^{G} \alpha_l a_l^*(i) b_l((j-1)L)\Big)\right| \\
&= \left|g\Big(\sum_{l=1}^{G} \alpha_l a_l(i) b_l((j-1)L)\Big) - g\Big(\sum_{l=1}^{G} \alpha_l a_l^*(i) b_l((j-1)L)\Big)\right| \\
&\le \mathcal{L}\left|\sum_{l=1}^{G} \alpha_l a_l(i) b_l((j-1)L) - \sum_{l=1}^{G} \alpha_l a_l^*(i) b_l((j-1)L)\right| \\
&= \mathcal{L}\left|\sum_{l=1}^{G} \alpha_l\,(a_l(i) - a_l^*(i)) \cdot b_l((j-1)L)\right| \\
&\le \mathcal{L}\sum_{l=1}^{G} \left|\alpha_l\,(a_l(i) - a_l^*(i)) \cdot b_l((j-1)L)\right| \\
&\le C_2\mathcal{L}\sum_{l=1}^{G} \left|a_l(i) - a_l^*(i)\right| \\
&= C_2\mathcal{L}\,\|\bar{a}(i) - \bar{a}^*(i)\|_1 \\
&\le \delta.
\end{aligned}$$

For each $(i,j) \in [L] \times [N]$, we define $\eta^*(i,(j-1)L) = g\Big(\sum_{l=1}^{G} \alpha_l a_l^*(i) b_l((j-1)L)\Big)$. Let $\boldsymbol{M}_{(r)}$ be the matrix whose (i,j)-th element is $\eta^*(i,(j-1)L)$. Consequently, we have for all k

$$\left\|\boldsymbol{M}^{(k)} - \boldsymbol{M}_{(r)}\right\|_{\max} \le \delta.$$

Observe that for $i_1, i_2 \in [L]$, if $\bar{a}(i_1)$ and $\bar{a}(i_2)$ map to the same element $\bar{a}^*(i) \in P(\frac{\delta}{C_2\mathcal{L}})$, then rows i_1, i_2 in $\boldsymbol{M}_{(r)}$ will be identical. Therefore, there are at most $\left|P(\frac{\delta}{C_2\mathcal{L}})\right|$ distinct rows in $\boldsymbol{M}_{(r)}$. For an appropriately defined $C > 0$, choosing $\delta = C\mathcal{L}L^{-\epsilon}$ gives $\left|P(\frac{\delta}{C_2\mathcal{L}})\right| \le L^{G\epsilon}$.

Hence, Property D.1 is satisfied with $\delta_4 = C\mathcal{L}L^{-\epsilon}$ and $r_4 = L^{G\epsilon}$. By Corollary D.1, we have: under Model Type 1, Property 4.1 is satisfied with $\delta_1 = \delta_4$ and $r = r_4$; under Model Type 2, Property 4.2 is satisfied with $\delta_2 = 2\delta_1\sqrt{N}$. This completes the proof. □

Corollary (5.3). *Under* Model Type 1, *let the conditions of Theorem 4.1 hold. Let $N = L^{1+\delta}$ for any $\delta > 0$. Then for some $C > 0$ and any $\epsilon \in (0,1)$ if*

$$T \geq C\left(\left(\frac{1}{\delta_{error}}\right)^{\frac{2}{1-G\epsilon}} + \left(\frac{\mathcal{L}}{\delta_{error}}\right)^{\frac{1}{\epsilon}}\right)^{2+\delta}$$

we have $\text{MSE}(\hat{f}_I, f^{\text{LRF}}) \leq \delta_{error}$.

Proof. By Proposition 5.3, for any $\epsilon \in (0,1)$ and some $C_1, C_2, C_3, c_4 > 0$,

$$MSE(\hat{f}_I, f^{Compact}) \leq \frac{C_1\sigma}{p}\left(\frac{\mathcal{L}}{L^{\epsilon}} + \frac{1}{L^{(1-G\epsilon)/2}}\right) + C_2\frac{(1-p)}{LNp} + C_3 e^{-c_4 N}.$$

We require the r.h.s of the term above to be less than δ_{error}. Thus, we have

$$\begin{aligned}
&\frac{C_1\sigma}{p}\left(\frac{\mathcal{L}}{L^{\epsilon}} + \frac{1}{L^{(1-G\epsilon)/2}}\right) + C_2\frac{(1-p)}{LNp} + C_3 e^{-c_4 N} \\
&\overset{(a)}{\leq} C\left(\frac{\mathcal{L}}{L^{\epsilon}} + \frac{1}{L^{(1-G\epsilon)/2}} + \frac{1}{LNp} + e^{-c_4 N}\right) \\
&\overset{(b)}{\leq} C\left(\frac{\mathcal{L}}{L^{\epsilon}} + \frac{1}{L^{(1-G\epsilon)/2}}\right)
\end{aligned}$$

where (a) follows for an appropriately defined $C > 0$ and by absorbing p, σ into the constant; (b) follows since $\frac{1}{LN} \leq \frac{\mathcal{L}}{L^{\epsilon}}, e^{-c_4 N} \leq \frac{\mathcal{L}}{L^{\epsilon}}$ for sufficiently large L, N and by redefining C.

To have $\frac{C}{L^{(1-G\epsilon)/2}} \leq \delta_{\text{error}}/2$, it suffices that $L \geq \left(\frac{2C}{\delta_{\text{error}}}\right)^{2/(1-G\epsilon)}$. Similarly, we solve $\frac{C\mathcal{L}}{L^{\epsilon}} \leq \delta_{\text{error}}/2$ to get $L \geq \left(\frac{2C\mathcal{L}}{\delta_{\text{error}}}\right)^{\frac{1}{\epsilon}}$. Thus for appropriately defined C, we require L to be

$$L \geq C\left(\left(\frac{1}{\delta_{\text{error}}}\right)^{\frac{2}{1-G\epsilon}} + \left(\frac{\mathcal{L}}{\delta_{\text{error}}}\right)^{\frac{1}{\epsilon}}\right) \tag{51}$$

$$\implies T \geq C\left(\left(\frac{1}{\delta_{\text{error}}}\right)^{\frac{2}{1-G\epsilon}} + \left(\frac{\mathcal{L}}{\delta_{\text{error}}}\right)^{\frac{1}{\epsilon}}\right)^{2+\delta}. \tag{52}$$

□

Corollary (5.4). *Under* Model Type 2, *let the conditions of Theorem 4.2 hold. Let $N = L^{1+\delta}$ for any $\delta > 0$. Then for some $C > 0$ and any $\epsilon \in (0,1)$ if*

$$T \geq C\left(\frac{\sigma^2}{\delta_{error} - \left(\frac{\mathcal{L}}{L^{\epsilon}} + \delta_3\right)^2}\right)^{\frac{2+\delta}{\delta}}$$

we have $\text{MSE}(\hat{f}_F, f^{\text{LRF}}) \leq \delta_{error}$.

Proof. By Proposition 5.3, for any $\epsilon \in (0,1)$ and some $C > 0$,

$$\text{MSE}(\hat{f}_F, f^{Compact}) \le \frac{1}{N-1}\left(\left(\frac{C\mathcal{L}}{L^\epsilon} + \delta_3\right)^2 N + 2\sigma^2 \hat{r}\right).$$

We require the r.h.s of the term above to be less than δ_{error}. Since $\frac{1}{N}\sigma^2\hat{r} \le \frac{1}{L^\delta}\sigma^2$, it suffices that

$$\delta_{\text{error}} \overset{(a)}{\ge} C\left(\left(\frac{\mathcal{L}}{L^\epsilon} + \delta_3\right)^2 + \frac{1}{L^\delta}\sigma^2\right)$$
$$\implies L^\delta \overset{(b)}{\ge} C\frac{\sigma^2}{\delta_{\text{error}} - \left(\frac{\mathcal{L}}{L^\epsilon} + \delta_3\right)^2}$$
$$\implies T \ge C\left(\frac{\sigma^2}{\delta_{\text{error}} - \left(\frac{\mathcal{L}}{L^\epsilon} + \delta_3\right)^2}\right)^{\frac{2+\delta}{\delta}}$$

where (a) and (b) follow for an appropriately defined $C > 0$. □

Proposition (5.4).

$$f^{\text{Harmonic}}(t) = \sum_{r=1}^{R} \varphi_r(\sin(2\pi\omega_r t + \phi))$$

where φ_r is $\mathcal{L}_r$-Lipschitz and ω_r is rational, admits a representation as in (11). Let x_{lcm} denote the fundamental period. Then the Lipschitz constant $\mathcal{L}$ of $f^{\text{Harmonic}}(t)$ is bounded by

$$\mathcal{L} \le 2\pi \cdot \max_{r\in R}(\mathcal{L}_r) \cdot \max_{r\in R}(\omega_r) \cdot x_{lcm}.$$

Proof. The fact that $f^{Harmonic}$ has a representation as in (11) follows immediately. It remains to show the explicit dependence of $\mathcal{L}$ on the parameters of $f^{Harmonic}$. Observe that

$$f^{Harmonic}(t) = f^{Harmonic}(\psi(t)),$$

where $\psi(t) = t \mod x_{\text{lcm}}$. By bounding the derivative of $f^{Harmonic}(t)$, it is easy to see that

$$\mathcal{L} \le 2\pi \cdot \max_{r\in R}(\mathcal{L}_r) \cdot \max_{r\in R}(\omega_r) \cdot x_{\text{lcm}}.$$

This completes the proof. □

D.4 Proof of Proposition 5.5

Proposition (5.5). *Let $\left|\frac{df^{\text{Trend}}(t)}{dt}\right| \le C_* t^{-\alpha}$ for some $\alpha, C_* > 0$. Then for any $\epsilon \in (0,\alpha)$,*

(i) Under Model Type 1, f^{Trend} *satisfies Property 4.1 with* $\delta_1 = \frac{C_*}{L^{\epsilon/2}}$ *and* $r = L^{\epsilon/\alpha} + \frac{L - L^{\epsilon/\alpha}}{L^{\epsilon/2}}$
(ii) Under Model Type 2, f^{Trend} *satisfies Property 4.2 with* $\delta_2 = 2\delta_1\sqrt{N}$ *and* $C_\beta = 1$.

Proof. Without loss of generality, we drop the dependence of k on η_k to decrease notational overload. Let $f(t) = f^{\text{Trend}}$. We construct our mapping $p : [L] \longrightarrow [L]$ in two steps:
Step 1: For $i < L^{\epsilon/\alpha}$, with $\epsilon \in (0,\alpha)$, let $p(i) = i$ (i.e., the i-th row of $\boldsymbol{M}_{(r)}$ is equal to the i-th row of $\boldsymbol{M}^{(k)}$).

Step 2: For rows $i \geq L^{\epsilon/\alpha}$, we construct the following mapping (similar to [19]). Let R and D refer to the set of row and column parameters of the sub-matrix of $\boldsymbol{M}^{(k)}$ corresponding to its last $L - i + 1$ rows, $\{\theta_{L^{\epsilon/\alpha}}, \cdots, \theta_L\}$ and $\{\rho_{1,} \cdots, \rho_N\}$, respectively.

Let f' denote the derivative of f, and $\theta \in (\min(i, i') + (j-1)L, \max(i, i') + (j-1)L)$. Then, we have that for all $i, i' \in R$

$$\begin{aligned} |\eta(i, (j-1)L) - \eta(i', (j-1)L)| &= |f(i + (j-1)L) - f(i' + (j-1)L)| \\ &\overset{(a)}{\leq} |f'(\theta)| \cdot |i + (j-1)L - (i' + (j-1)L)| \\ &\overset{(b)}{\leq} C_*(L^{\epsilon/\alpha})^{-\alpha} \cdot |i - i'| \\ &= C_* L^{-\epsilon} \cdot |i - i'|, \end{aligned}$$

where (a) follows from the Mean Value Theorem, and (b) uses the fact that $|f'(\theta)| \leq C_* \min(i, i')^{-\alpha} \leq C_*(L^{\epsilon/\alpha})^{-\alpha}$.

We define a partition $P(\epsilon)$ of R into continuous intervals of length $L^{\epsilon/2}$. Then, for any $A \in P(\epsilon)$, we have $|\theta - \theta^{'}| \leq L^{\epsilon/2}$ (recall that $\theta_i = i$) whenever $\theta, \theta^{'} \in A$. It follows that $|P(\epsilon)| = (L - L^{\epsilon/\alpha})/L^{\epsilon/2} = L^{1-\epsilon/2} - L^{\epsilon(\frac{1}{\alpha} - \frac{1}{2})}$.

Let T be a subset of R that is constructed by selecting exactly one element from each partition in $P(\epsilon)$, i.e., $|T| = |P(\epsilon)|$. For each $\theta \in R$, let $p(\theta)$ be the corresponding element from the same partition in T. Therefore, it follows that for each $\theta \in R$, we can find $p(\theta) \in T$ so that θ and $p(\theta)$ belong to the same partition of $P(\epsilon)$.

Hence, we can define the (i, j)-th element of $\boldsymbol{M}_{(r)}$ in the following way: (1) for all $i < L^{\epsilon/\alpha}$, let $p(\theta_i) = \theta_i$ such that $M^{(r)}_{ij} = \eta(\theta_i, \rho_j)$; (2) for $i \geq L^{\epsilon/\alpha}$, let $\boldsymbol{M}^{(r)}_{ij} = \eta(p(\theta_i), \rho_j)$. Consequently for all k,

$$\begin{aligned} \left\| \boldsymbol{M}^{(k)} - \boldsymbol{M}_{(r)} \right\|_{\max} &\leq \max_{i \in [L], j \in [N]} |\eta(\theta_i, \rho_j) - \eta(p(\theta_i), \rho_j)| \\ &= \max_{i \in [j \geq L^{\epsilon/\alpha}], j \in [N]} |\eta(\theta_i, \rho_j) - \eta(p(\theta_i), \rho_j)| \\ &\leq \max_{i \in [j \geq L^{\epsilon/\alpha}]} |\theta_i - p(\theta_i)| \, L^{-\epsilon} C_* \\ &\leq C_* L^{-\epsilon/2}. \end{aligned}$$

Now, if θ_i and θ_j belong to the same element of $P(\epsilon)$, then $p(\theta_i)$ and $p(\theta_j)$ are identical. Therefore, there are at most $|P(\epsilon)|$ distinct rows in the last $L - L^{\epsilon/\alpha}$ rows of $\boldsymbol{M}_{(r)}$ where $|P(\epsilon)| = L^{1-\epsilon/2} - L^{\epsilon(\frac{1}{\alpha} - \frac{1}{2})}$. Let $\mathcal{P}(\theta) := \{p(\theta_i) : i \in [L]\} \subset \{\theta_1, \ldots, \theta_L\}$. By construction, since $\epsilon \in (0, \alpha)$, we have that $|\mathcal{P}(\theta)| = L^{\epsilon/\alpha} + |P(\epsilon)| = o(L)$.

Hence, Property D.1 is satisfied with $\delta_1 = \frac{C_*}{L^{\epsilon/2}}$ and $r = L^{\epsilon/\alpha} + \frac{L - L^{\epsilon/\alpha}}{L^{\epsilon/2}}$. By Corollary D.1, we have: under Model Type 1, Property 4.1 is satisfied with $\delta_1 = \delta_4$ and $r = r_4$; under Model Type 2, Property 4.2 is satisfied with $\delta_2 = 2\delta_1\sqrt{N}$. This completes the proof. □

Corollary (5.5). *Under* Model Type 1, *let the conditions of Theorem 4.1 hold. Let $N = L^{1+\delta}$ for any $\delta > 0$. Then for some $C > 0$, if*

$$T \geq C\left(\frac{1}{\delta_{error}^{(2(\alpha+1)/\alpha)}}\right)^{2+\delta}$$

we have $\text{MSE}(\hat{f}_I, f^{\text{LRF}}) \leq \delta_{error}$.

Proof. By Proposition 5.5, for any $\epsilon \in (0, \alpha)$ and some $C_1, C_2, C_3, c_4 > 0$,

$$MSE(\hat{f}_I, f^{Trend}) \leq \frac{C_1 \sigma}{p} \left(\frac{C_*}{L^{\epsilon/2}} + \frac{1}{(L^{1-\epsilon/\alpha} + L^{\epsilon/2})^{1/2}} \right)$$
$$+ C_2 \frac{(1-p)}{LNp} + C_3 e^{-c_4 N}.$$

We require the r.h.s of the term above to be less than δ_{error}. We have,

$$\begin{aligned} &\frac{C_1 \sigma}{p} \left(\frac{C_*}{\sqrt{p} L^{\epsilon/2}} + \frac{1}{\sqrt{p}(L^{1-\epsilon/\alpha} + L^{\epsilon/2})^{1/2}} \right) + C_2 \frac{(1-p)}{LNp} + C_3 e^{-c_4 N} \\ &\overset{(a)}{\leq} C \left(\frac{1}{L^{\epsilon/2}} + \frac{1}{(L^{1-\epsilon/\alpha} + L^{\epsilon/2})^{1/2}} + \frac{1}{LN} + e^{-c_4 N} \right) \\ &\overset{(b)}{\leq} C \left(\frac{1}{L^{\epsilon/2}} + \frac{1}{(L^{1-\epsilon/\alpha} + L^{\epsilon/2})^{1/2}} \right) \\ &\leq C \left(\frac{1}{L^{\epsilon/2}} + \frac{1}{(L^{1-\epsilon/\alpha})^{1/2}} \right) \end{aligned}$$

where (a) follows for an appropriately defined $C > 0$ and by absorbing p, σ into the constant; (b) follows since $\frac{1}{LN} \leq \frac{1}{L^{\epsilon/2}}, e^{-c_4 N} \leq \frac{1}{L^{\epsilon/2}}$ for sufficiently large L, N and by redefining C.

Setting $\frac{\epsilon}{2} = \frac{1-\epsilon/\alpha}{2}$, we get $\epsilon = \frac{\alpha}{\alpha+1} < \alpha$, which satisfies the condition that $\epsilon \in (0, \alpha)$ in Proposition 5.5. Therefore, it suffices that $\delta_{\text{error}} \geq CL^{\frac{\alpha}{2(\alpha+1)}} \implies T \geq C \Big(\frac{1}{\delta_{\text{error}}^{\frac{2(\alpha+1)}{\alpha}}} \Big)^{2+\delta}$.

□

Corollary (5.6). *Under* Model Type 2, *let the conditions of Theorem 4.2 hold.. Let $N = L^{1+\delta}$ for any $\delta > 0$. Then for some $C > 0$ and for any $\epsilon \in (0, \alpha)$ if*

$$T \geq C \left(\frac{\sigma^2}{\delta_{error} - \left(\frac{1}{L^{\epsilon/2}} + \delta_3 \right)^2} \right)^{\frac{2+\delta}{\delta}}$$

we have $\text{MSE}(\hat{f}_F, f^{\text{LRF}}) \leq \delta_{error}$.

Proof. By Proposition 5.5, for any $\epsilon \in (0, \alpha)$,

$$MSE(\hat{f}_F, f^{Trend}) \leq \frac{1}{N-1} \Big((\frac{C_*}{L^{\epsilon/2}} + \delta_3)^2 N + 2\sigma^2 \hat{r} \Big).$$

We require the r.h.s of the term above to be less than δ_{error}. Since $\frac{1}{N}\sigma^2 \hat{r} \leq \frac{1}{L^\delta}\sigma^2$, it suffices that

$$\begin{aligned} \delta_{\text{error}} &\overset{(a)}{\geq} C \left(\left(\frac{1}{L^{\epsilon/2}} + \delta_3 \right)^2 + \frac{1}{L^\delta} \sigma^2 \right) \\ \implies L^\delta &\overset{(b)}{\geq} C \frac{\sigma^2}{\delta_{\text{error}} - \left(\frac{1}{L^{\epsilon/2}} + \delta_3 \right)^2} \end{aligned}$$

$$\implies T \geq C\left(\frac{\sigma^2}{\delta_{\text{error}} - \left(\frac{1}{L^{\epsilon/2}} + \delta_3\right)^2}\right)^{\frac{2+\delta}{\delta}}$$

where (a) and (b) follow for an appropriately defined $C > 0$. □

Proposition (5.6). *For $t \in \mathbb{Z}$ with $\alpha_b < 1$ for $b \in [B]$,*

$$f^{Trend}(t) = \sum_{b=1}^{B} \gamma_b t^{\alpha_b} + \sum_{q=1}^{Q} \log(\gamma_q t).$$

admits a representation as in (13).

Proof. The proof follows immediately from the definition of f^{Trend}. □

D.5 Proof of Proposition 5.7

Proposition (5.7). *Under* Model Type 1, *f^{Mixture} satisfies Property 4.1 with $\delta_1 = \sum_{q=1}^{Q} \rho_q \delta_q$ and $r = \sum_{q=1}^{Q} r_q$.*

Proof. Let $\boldsymbol{M}_g^{(1)}$ refer to the underlying mean matrix induced by each $X_g(t)$. Similarly, as defined in Property 4.1, let $\boldsymbol{M}_{g,(r)}$ be the low rank matrix associated with $\boldsymbol{M}_g^{(1)}$. We have

$$\boldsymbol{M}^{(1)} = \sum_{g}^{G} \alpha_g \boldsymbol{M}_g^{(1)}.$$

We define $\boldsymbol{M}_{(r)}$ as

$$\boldsymbol{M}_{(r)} = \sum_{g}^{G} \alpha_g \boldsymbol{M}_{g,(r)}.$$

As a result, we have that $\text{rank}(\boldsymbol{M}_{(r)}) \leq \sum_{g}^{G} r_g$, and

$$\begin{aligned}\left\|\boldsymbol{M}^{(1)} - \boldsymbol{M}_{(r)}\right\|_{\max} &= \left\|\sum_{g}^{G} \alpha_g \boldsymbol{M}_g^{(1)} - \sum_{g}^{G} \alpha_g \boldsymbol{M}_{g,(r)}\right\|_{\max} \\ &\leq \sum_{g}^{G} \alpha_g \left\|\boldsymbol{M}_g^{(1)} - \boldsymbol{M}_{g,(r)}\right\|_{\max} \\ &= \sum_{g}^{G} \alpha_g \delta_g.\end{aligned}$$

This completes the proof. □

Received August 2018; revised October 2018; accepted December 2018