Appendix A: Page vs. Hankel mSSA

This section discusses the benefits and drawbacks of using the Page matrix representation, as we propose in our variant, instead of the Hankel representation used in the original mSSA. Recall the key steps of the original SSA method in Section 2. The extension to mSSA is done by stacking the Hankel matrices induced by each of the N time series either column-wise (horizontal mSSA) or row-wise (vertical mSSA) Hassani and Mahmoudvand (2018). In this section, we will use mSSA to denote our mSSA variant, and hSSA/vSSA to denote the original horizontal/vertical mSSA. In what follows, we will compare our mSSA variant with hSSA/vSSA in terms of their: (i) theoretical analysis; (ii) computational complexity; and (iii) empirical performance.

Theoretical analysis. We re-emphasize that to the best of our knowledge, the theoretical analysis of the mSSA algorithm, both hSSA and vSSA, have been absent from the literature, despite their popularity. We do a comprehensive theoretical analysis of the variant of mSSA we propose. By utilizing the Page matrix, it allows us to invoke results from random matrix theory to prove our imputation and forecasting results. However, extending our analysis to the Hankel matrix representation is challenging as the Hankel matrix has repeated entries of the same time series observation. This leads to correlation in the noise in the observation of the entries of the Hankel matrix, which prevents us from invoking the results from random matrix theory in a straightforward way. The Page matrix representation does not have repeated entries of the same observation, and thus allows us to circumvent this issue in our theoretical analysis.

Computational complexity. Our mSSA variant is computationally far more efficient than both hSSA and vSSA. This is because the Page matrix representation of a multivariate time series with N time series and T time steps is a matrix of dimension $\sqrt{NT} \times \sqrt{NT}$ (with $L = \sqrt{NT}$)., i.e., it has a total of $O(NT)$ entries. In contrast, the Hankel matrix representation is of dimension $T/4 \times 3NT/4$ for hSSA and $NT/4 \times 3T/4$ for vSSA (we set the parameter L to $T/4$ as recommended in Hassani and Mahmoudvand (2018)), i.e., both variants of the Hankel matrix have $O(NT^2)$ entries. This makes computing the SVD (the most computationally intensive step of mSSA) prohibitive for hSSA and mSSA even for the standard time series benchmarks we consider in Section 6.

To empirically demonstrate the computational efficiency of our variant of mSSA, we compare its training time to that of hSSA and vSSA. Specifically, we measure the training time for mSSA, hSSA, and vSSA as we increase the number of time steps $T \in [400, 10000]$. We perform this experiment on two datasets: (i) the synthetic dataset; (ii) a subset of the electricity dataset, where we choose

only 50 of the available 370 time series. Both datasets are described in details in Appendix B. Figure 10 shows that in both datasets, the training time of both hSSA and vSSA can be as 600-1000x as high as the training time of our mSSA variant as we increase T.

Empirical performance. Here, we compare the forecasting performance of mSSA to that of hSSA and vSSA. We report performance in terms of the NRMSE of the three methods as we increase the number of time steps $T \in [400, 10000]$ in the aforementioned synthetic and electricity dataset. The goal in the synthetic dataset is to predict the next 50 time steps using one step ahead forecasts, while the goal in the electricity dataset is to predict the next three days using day-ahead forecasts. For hSSA and vSSA, we choose $L = T/4$ as recommended in Hassani and Mahmoudvand (2018); and for mSSA, we choose $L = \lfloor \sqrt{NT} \rfloor$. For all three methods, we choose the number of retained singular values based on the thresholding procedure outlined in Gavish and Donoho (2014).

Figures 11 shows the performance of the three methods in both datasets. We find that initially, with few data points ($T < 600$ in the synthetic data and $T < 4000$ in the electricity data), both hSSA and vSSA outperform mSSA. As we increase T, mSSA performance significantly improves and eventually outperforms vSSA. In the electricity dataset, mSSA performs similar to hSSA for $T = 10000$. These experiments suggest that if only a few observations were available, hSSA and vSSA might provide better performance. However, if the number of observations were relatively large, then the performance of mSSA is superior to vSSA and relatively similar to hSSA.

Importantly, the electricity dataset experiment illustrates a critical advantage of our mSSA variant. Specifically, when T is large such that running hSSA or vSSA is computationally infeasible, then one can achieve better accuracy using mSSA. For example, while we could not run the hSSA and vSSA on the electricity dataset with $T = 20000$ due to memory constraints, we were able to run mSSA and achieve a lower NRMSE. This suggests that our mSSA variant is the more practical mSSA algorithm when it comes to efficiently utilizing large multivariate time series.

Appendix B: Experiment Details

In Appendix B.1, we describe the datasets utilized. In Appendix B.2, we describe the various algorithms we compare with as well as the choice of hyper-parameters used for each of them.

B.1. Datasets

We use four real-world datasets and one synthetic dataset. The description and preprocessing we do for each of these datasets are as follows.

Electricity Dataset. This is a public dataset obtained from the UCI repository which shows the 15-minutes electricity load of 370 households Trindade (2014). As was done in Yu et al. (2016),Sen et al. (2019),Salinas et al. (2019), we aggregate the data into hourly intervals and use the first 25824 time-points for training, the next 288 points for validation, and the last 168 points for testing in the forecasting experiments. Specifically, in our testing period, we do 24-hour ahead forecasts for the next seven days (i.e. 24-step ahead forecast). See Table 3 for more details.

Traffic Dataset. This public dataset obtained from the UCI repository shows the occupancy rate of traffic lanes in San Francisco Trindade (2014). The data is sampled every 15 minutes but to be consistent with previous work in Yu et al. (2016), Sen et al. (2019), we aggregate the data into hourly data and use the first 10248 time-points for training, the next 288 points for validation, and the last 168 points for testing in the forecasting experiments. Specifically, in our testing period, we do 24-hour ahead forecasts for the next seven days (i.e. 24-step ahead forecast). See Table 3 for more details.

Financial Dataset. This dataset is obtained from the Wharton Research Data Services (WRDS) and contains the average daily stocks prices of 839 companies from October 2004 till November 2019 WRDS (2021). The dataset was preprocessed to remove stocks with any null values, or those with an average price below 30$ across the aforementioned period. This was simply done to constrain the number of time series for ease of experimentation and we end up with 839 time series (i.e. stock prices of listed companies) each with 3993 readings of daily stock prices. In our forecasting experiments, we train on the first 3693 time points, validate on the next 120 time points, while for testing we consider the task of predicting 180 time-points ahead one point at a time. That is, the goal here is to do one-day ahead forecasts for the next 180 days (i.e. 1-step ahead forecast). We choose to do so as this is a standard goal in finance. See Table 3 for more details.

M5 Dataset. This public dataset obtained from Kaggle's M5 Forecasting competition include daily sales data of 30490 items across different Walmart stores for 1941 days Makridakis et al. (2020). The dataset was preprocessed to only include items that has more than zero sales in at least 500 days. For forecasting, as is the goal in the Kaggle competition, we consider the task of predicting the sales for the next 28 days (i.e. 28-step ahead forecast). We use the first 1829 points for training, the next 84 points for cross validation, and the last 28 points for testing.

Synthetic Dataset. We generate the observation tensor $X \in \mathbb{R}^{n\times m\times T}$ by first randomly generating the two matrices $U \in \mathbb{R}^{r\times n} = [u_1, \ldots, u_n]$ and $V \in \mathbb{R}^{r\times m} = [v_1, \ldots, v_m]$; we do so by randomly

sampling each coordinate of U, V independently from a standard normal. Then, we generate r mixtures of harmonics where each mixture $g_k(t), k \in [r]$, is generated as: $g_k(t) = \sum_{h=1}^{4} \alpha_h \cos(\omega_h t/T)$ where the parameters α_h, ω_h are selected uniformly at randomly from the ranges $[-1, 10]$ and $[1, 1000]$, respectively. Then each value in the observation tensor is constructed as follows: $X_{i,j}(t) = \sum_{k=1}^{r} u_{ik} v_{jk} g_k(t)$, where r is the tensor rank, $i \in [n]$, $j \in [m]$. In our experiment, we select $n = 5$, $m = 10$, $T = 15000$, and $r = 4$. This gives us $N = n \times m = 50$ time series each with 15000 observations per time series. In the forecasting experiments, we use the first 13700 points for training, the next 300 points for validation, while for testing, we do 10-step ahead forecasts for the final 1000 points. See Table 3 for more details.

Table 3 Dataset and training/validation/test split details.

Dataset	No.time series	Observations per time series	Forecast horizon (h)	Training period	No. validation windows W_{val}	Validation period	No. test windows	Test period
Electricity	370	26136	24	1 to 25824	2	25825 to 25968	7	25969 to 26136
Traffic	963	10560	24	1 to 10248	2	10249 to 10392	7	10393 to 10560
Synthetic	50	15000	10	1 to 13700	10	13701 to 14000	100	14001 to 15000
Financial	839	3993	1	1 to 3693	40	3694 to 3813	180	3814 to 3993
M5	15678	1941	28	1 to 1829	1	1830 to 1913	1	1914 to 1941

B.2. Algorithms.

In this section, we describe the algorithms used throughout the experiments in more detail and the hyper-parameters/implementation used for each method.

mSSA & SSA. Note that since the SSA's variant described in Agarwal et al. (2018) is a special case of our proposed mSSA algorithm, we use our mSSA's implementation to perform the SSA experiments; key difference in SSA is that we do not ''stack'' the various Page matrices induced by each time series. For all experiments we choose the parameters through the cross validation process detailed in Appendix B.3, where we perform a grid search for the following parameters:

1. *The number of retained singular values, k.* This parameter is chosen using one of the following data-driven methods: (i) we choose k based on the thresholding procedure outlined in Gavish and Donoho (2014), where the threshold is determined by the median of the singular values and the shape of the matrix; (ii) we choose k as the minimum number of singular values that capture a fraction τ of the spectral energy, with $\tau \in \{0.7, 0.8, 0.9, 0.95\}$; (iii) we additionally try fixed ranks $k \in \{1, 3, 5, 10, 25\}$.
2. *The shape of the Page matrix.* The Page matrix shape is controlled either by setting L directly or through a column-to-row ratio $\rho = M/L$. For mSSA, $L \in \{10, 500, 700, 800, 1000, 1250, 2000\}$

(dataset-dependent) or $\rho \in \{2, 5, 10, 20, 500\}$. For SSA, $L \in \{10, 30, 40, 50, 80, 100, 150\}$ (dataset-dependent) or $\rho \in \{2, 5, 10\}$.

3. *Missing values initialization.* Initializing the missing values is done according to one of two methods: (i) set the missing values to zero; (ii) perform forward filling where each missing value is replaced by the nearest preceding observation, followed by backward filling to accommodate the situation when the first observation is missing.

DeepAR. We use the ''DeepAREstimator'' algorithm provided by the GluonTS package. We choose the parameters through a grid search for the following parameters:

1. *Context length*. This parameter determines the number of steps to unroll the RNN for before computing predictions. We choose this from the set $\{h \text{ (default)}, 2h, 3h\}$, where h is the prediction horizon.
2. *Number of Layers.* This parameter determines the number of RNN layers. We choose this from the set $\{2 \text{ (default)}, 3\}$.

TRMF. We use the implementation provided by the authors in the Github repository associated with the paper (Yu et al. (2016)). We choose the parameters through a grid search, as suggested by the authors in their codebase, for the following parameters:

1. *Matrix rank k*. This parameter represents the chosen rank for the $T \times N$ time series matrix, we choose k from the set $\{5, 10, 20, 40, 60\}$.
2. *Regularization parameters $\lambda_f, \lambda_x, \lambda_w$*. We choose these parameters from $\{0.05, 0.5, 5, 50\}$ as suggested in the authors repository.

For the lag indices , we include the last day and the same weekday in the last week for the traffic and electricity data, the last 30 points for the financial and synthetic dataset, and the last 10 points for the M5 dataset.

LSTM. Across all datasets, we use an LSTM network with $H \in \{2, 3, 4\}$ hidden layers each, with 45 neurons per layer, as is done in Sen et al. (2019). We use the Keras implementation of LSTM. As with other methods' parameters, H is chosen via cross validation.

Prophet. We used Prophet's Python library with the parameters selected using a grid search of the following parameters as suggested in Facebook (2020):

1. *Changepoint prior scale*. This parameter determines how much the trend changes at the detected trend changepoints. We choose this parameter from $\{0.001, 0.05, 0.2\}$.

2. *Seasonality prior scale.* This parameter controls the magnitude of the seasonality. We choose this parameter from $\{0.01, 10\}$.
3. *Seasonality Mode.* Which is chosen to be either 'additive' or 'multiplicative'.

VAR. We used the VAR estimator in the python package ''statsmodels'' (Seabold and Perktold (2010)). We apply the method on the first difference of the time series and verify that the series are not non-stationary using a unit root test (specifically, Augmented Dickey–Fuller test). For all datasets except M5, we choose the best value for the parameter `max_lag` $\in \{1, 2, 5, 10, 20, 50\}$. This parameter corresponds to the maximum number of lags used in fitting the VAR process. For M5, we choose `max_lag` $\in \{1, 2, 5\}$, as fitting the model for larger values is computationally infeasible.

tSSA, mSSA and ME in Section 6.3. In this experiment, we use HSVT as the the matrix estimation subroutine for both ME in mSSA. For both methods, we choose the number of singular components retained based on the the thresholding procedure outlined in Gavish and Donoho (2014). For tSSA, we use ALS as the tensor estimation subroutine. Therein, we choose the best performing rank among the follwoing options: (i) the rank suggested by the thresholding procedure outlined in Gavish and Donoho (2014) for the stacked Page matrix used in mSSA; (ii) the rank suggested by the same procedure for the Page matrix of one of the time series (specifically the first); (iii) the fixed values $\lfloor L/2 \rfloor$ and $\lfloor L/3 \rfloor$.

B.3. Parameters Selection

In all experiments, we choose the hyperparameters for out method and for the baselines by using cross-validation. Below, we detail the procedure for both imputation and forecasting experiments.

Imputation Experiments. To select the parameters in our imputation experiments, we additionally mask 10% of the observed data uniformly at random. Then, we evaluate the performance of each parameter choice in recovering these additionally masked observations. This process is repeated 3 times, and the choice of parameters that achieves the best performance (in NRMSE) across these runs is selected. In our results, we report the accuracy of the selected parameters in recovering the original missing values.

Forecasting Experiments. For parameters selection in the forecasting experiments, we use cross-validation on a rolling basis as typically used in time-series forecasting models Hyndman and Athanasopoulos (2018). In this procedure, there are multiple validation sets. For each validation set, we train the model only on previous observations. That is, no future observations can be used

in training the model, which will occur when a typical cross-validation procedure is followed for time series data. In our experiments, we start with a subset of the data used for training, then we forecast the first validation set using h-step ahead forecasts for W_{val} windows , where the horizon h and the number of validation windows W_{val} are detailed in Table 3. We do this for three validation sets, each of length $h \times W_{val}$, and select the choice of parameters that achieves the best performance (in NRMSE) for evaluation on the test set. When evaluating on the test set, both the training and validation periods are used for training.

Appendix C: Time-varying Recommendation Systems

In tSSA, we considered the setting where the $N \times T$ matrix $\boldsymbol{M}$ induced by the latent time series $f_1(\cdot), \ldots, f_N(\cdot)$ is low-rank; in particular, Property 1 captures this spatial structure across these N time series. However, in many settings there is *additional* spatial structure across the N time series.

Recommendation systems – time-varying matrices/tensors. For example, in recommendation systems, for each $t \in T$, there is a $N_1 \times N_2$ matrix, $\boldsymbol{M}^{(t)} \in \mathbb{R}^{N_1 \times N_2}$ of interest. The n_1-th row and n_2-th column of $\boldsymbol{M}^{(t)}$ denotes the latent rating user n_1 has for product n_2, i.e., $\boldsymbol{M}^{(t)}_{n_1,n_2}$ denotes the value of the latent time series $f_{n_1,n_2}(\cdot)$ at time step t. To capture the latent structure across users and products, one typically assumes that each $\boldsymbol{M}^{(t)}$ is low-rank. More generally, at each time step t, $\boldsymbol{M}^{(t)} \in \mathbb{R}^{N_1 \times N_2, \ldots, \times N_d}$ could be an order-d tensor. That is, $\boldsymbol{M}^{(t)}_{n_1,\ldots,n_d}$ denotes the value of the latent time series $f_{n_1,\ldots,n_d}(\cdot)$ at time step t for $n_1, \ldots, n_d \in [N_1] \times \cdots \times [N_d]$. For example, if $d = 3$, $\boldsymbol{M}^{(t)}$ might represent the t-th measurement for a collection of (x, y, z)-spatial coordinates. Let $\boldsymbol{N} \in \mathbb{R}^{N_1 \times N_2, \ldots, \times N_d \times T}$ denote the $d+1$ order tensor induced by viewing each order-d tensor $\boldsymbol{M}^{(t)}$ as the t-th 'slice' of $\boldsymbol{N}$, for $t \in [T]$. Again, to capture the spatial and temporal structure of these latent time series, we posit the following spatio-temporal model for $\boldsymbol{N}$, which is a higher-order analog of the model assumed in Property 1.

PROPERTY 13. Let $\boldsymbol{N}$ have CP-rank at most R. That is, for any $n_1, \ldots, n_d \in [N_1] \times \cdots \times [N_d]$

$$\boldsymbol{N}_{n_1,\ldots,n_d,t} = \sum_{r=1}^{R} U_{n_1,r} \ldots U_{n_d,r} \; W_{rt},$$

where the factorization is such that $|U_{n_1,r}|, \ldots |U_{n_d,r}| \leq \Gamma_1$, $|W_{rt}| \leq \Gamma_2$ for constants $\Gamma_1, \Gamma_2 > 0$.

As before, to explicitly model the temporal structure, we continue to assume Property 2 holds for the latent time factors $W_{r\cdot}$ for $r \in [R]$.

Order-$d+2$ Page tensor representation. We now consider the following order-$d+2$ Page tensor representation of $\boldsymbol{N}$. In particular, given the hyper-parameter $L \geq 1$, define $\mathbf{HT} \in \mathbb{R}^{N_1 \times \cdots \times N_d \times L \times T/L}$ such that for $n_1, \ldots, n_d \in [N_1] \times \cdots \times [N_d],\ \ell \in [L],\ s \in [T/L]$,

$$\mathbf{HT}_{n_1,\ldots,n_d,\ell,s} = f_{n_1,\ldots,n_d}((s-1) \times L + \ell).$$

The corresponding observation tensor, $\mathbb{HT} \in (\mathbb{R} \cup \{\star\})^{N_1 \times \cdots \times N_d \times L \times T/L}$, is

$$\mathbb{HT}_{n_1,\ldots,n_d,\ell,s} = X_{n_1,\ldots,n_d}((s-1) \times L + \ell). \tag{10}$$

Recall from (1) that $X_{n_1,\ldots,n_d}(t)$ is the noisy, missing observation we get of $f_{n_1,\ldots,n_d}(t)$. $\mathbf{HT}$ and $\mathbb{HT}$ then have the following property:

PROPOSITION 13. *Let Properties 13, 2, and 3 hold. Then, for any $1 \leq L \leq \sqrt{T}$, $\mathbf{HT}$ has CP-rank at most $R \times G$. Further, all entries of $\mathbb{HT}$ are independent random variables with each entry observed with probability $\rho \in (0, 1]$, and $\mathbb{E}[\mathbb{HT}] = \rho \mathbf{HT}$.*

Analogous to Proposition 3, Proposition 13 also establishes that order-$d+2$ Page tensor representation of the various latent time series $f_{n_1,\ldots,n_d}(\cdot)$ has CP-rank that continues to be bounded by $R \times G$. Proof of Proposition 13 can be found in Appendix L.

Higher-order tensor singular spectrum analysis (htSSA). Proposition 13 motivates the following algorithm, which exploits the further spatial structure amongst the N time series. We now define the ''meta'' htSSA algorithm. The two algorithmic hyper-parameters are $L \geq 1$ (defined in (5)) and TE_{d+2} (the order-$d+2$ tensor estimation algorithm one chooses). First, using the observations $X_{n_1,\ldots,n_d}(t)$ for $n_1, \ldots, n_d \in [N_1] \times \cdots \times [N_d], t \in [T]$ we construct the higher-order Page tensor $\mathbb{HT}$ as in (10). Second, we obtain $\widehat{\mathbf{HT}}$ as the output of $\mathsf{TE}_{d+2}(\mathbb{HT})$, and read off $\hat{f}_{n_1,\ldots,n_d}(t)$ by selecting the appropriate entry in $\widehat{\mathbf{HT}}$.

Relative effectiveness of mSSA, htSSA, and tensor estimation (TE). Again, for ease of exposition, we consider the case where $\rho = 1$. We now briefly discuss the relative effectiveness of htSSA, mSSA,and ''vanilla'' tensor estimation (TE) in imputing $X_{n_1,\ldots,n_d}(\cdot)$ to estimate $f_{n_1,\ldots,n_d}(\cdot)$. mSSA and htSSA have been previously described. In TE, one directly de-noises the original order-$d+1$ tensor induced by the noisy observations, which we denote $\boldsymbol{X} \in \mathbb{R}^{N_1 \times N_2, \ldots, \times N_d \times T}$, where $\boldsymbol{X}_{n_1,\ldots,n_d,t} = X_{n_1,\ldots,n_d}(t)$. In particular, one produces an estimate of $\widehat{\boldsymbol{N}} = \mathsf{TE}_{d+1}(\boldsymbol{X})$, and then produces the estimates $\hat{f}_{n_1,\ldots,n_d}(t)$ by reading off the appropriate entry of $\widehat{\boldsymbol{N}}$. Let $\mathsf{ImpErr}(N, T; \text{htSSA})$,

$\mathsf{ImpErr}(N,T;\text{mSSA})$, and $\mathsf{ImpErr}(N,T;\text{TE})$ denote the imputation error for htSSA, mSSA, and TE, respectively. Now if we assume Property 7 holds, we have

$$\mathsf{ImpErr}(N,T;\text{htSSA}) = \tilde{\Theta}\left(\frac{1}{\min\left(N_1,\ldots,N_d,\sqrt{T}\right)^{\lceil\frac{d+2}{2}\rceil}}\right),$$
$$\mathsf{ImpErr}(N,T;\text{mSSA}) = \tilde{\Theta}\left(\frac{1}{\sqrt{\min(N,T)T}}\right),$$
$$\mathsf{ImpErr}(N,T;\text{TE}) = \tilde{\Theta}\left(\frac{1}{\min\left(N_1,\ldots,N_d,T\right)^{\lceil\frac{d+1}{2}\rceil}}\right).$$

Then just as was done in the proof of Proposition 6, for any given d, one can reason about the relative effectiveness of htSSA, mSSA, and TE for different asymptotic regimes of the relative ratio of N and T.

Appendix D: Proof of Proposition 7

Below, we present the proof of Proposition 7. First we define the stacked Hankel matrix of N time series over T time steps. Precisely, given N latent time series $f_1,\ldots,f_N$, consider the stacked Hankel matrix induced by each of them over T time steps, $[T]$, defined as follows. It is $\mathsf{SH} \in \mathbb{R}^{\lfloor T/2\rfloor \times N\lfloor T/2\rfloor}$ where its entry in row $i \in [\lfloor T/2\rfloor]$ and column $j \in [N\lfloor T/2\rfloor]$, SH_{ij}, is given by

$$\mathsf{SH}_{ij} = f_{n(i,j)}(i + (j \mod \lfloor T/2\rfloor) - 1), \quad \text{where } n(i,j) = \left\lceil \frac{j}{\lfloor T/2\rfloor} \right\rceil.$$

We now establish Proposition 14, which immediately implies Proposition 7 – the stacked Page matrix can be viewed as a sub-matrix of SH, by selecting the appropriate columns.

PROPOSITION 14. *Let Properties 1 and 8 hold for N latent time series of interest, $f_1,\ldots,f_N$. Then for any $T \geq 1$, the stacked Hankel Matrix of these N time series has ϵ'-approximate rank $R \times G$ with $\epsilon' = R\Gamma_1\epsilon$.*

We have N latent time series $f_1,\ldots,f_n$ satisfying Properties 1 and 8. Consider their stacked Hankel matrix over $[T]$, $\mathsf{SH} \in \mathbb{R}^{\lfloor T/2\rfloor \times N\lfloor T/2\rfloor}$. By definition for $i \in [\lfloor T/2\rfloor]$ and $j = (n-1)\times\lfloor T/2\rfloor + j'$ for $j' \in [\lfloor T/2\rfloor]$, we have

$$\mathsf{SH}_{ij'} = f_n(i + j' - 1).$$

That is,

$$\begin{aligned}\mathsf{SH}_{ij} &= f_n(i+j'-1) \\ &= \sum_{r=1}^{R} U_{nr} W_{r(i+j'-1)}. \end{aligned} \tag{11}$$

Let $H(r) \in \mathbb{R}^{\lfloor T/2 \rfloor \times \lfloor T/2 \rfloor}$ be the Hankel matrix associated with $W_{r\cdot}$ over $[T]$. Due to Property 8, there exists a low-rank matrix $M(r) \in \mathbb{R}^{\lfloor T/2 \rfloor \times \lfloor T/2 \rfloor}$ such that (a) $\text{rank}(M(r)) \le G$, (b) $\|H(r) - M(r)\|_\infty \le \epsilon$. That is, for any $i, j' \in [\lfloor T/2 \rfloor]$, we have that $M(r)_{ij'} = \sum_{g=1}^{G} a^r_{ig} b^r_{j'g}$ for some $a^r_{i\cdot}, b^r_{j'\cdot} \in \mathbb{R}^G$. Therefore, for any $i, j' \in [\lfloor T/2 \rfloor]$, we have that

$$\begin{aligned} W_{r(i+j'-1)} = H(r)_{ij'} \; &= \; M(r)_{ij'} + (H(r)_{ij'} - M(r)_{ij'}) \\ &= \sum_{g=1}^{G} a^r_{ig} b^r_{j'g} + (H(r)_{ij'} - M(r)_{ij'}). \end{aligned} \tag{12}$$

From (11) and (12), we conclude that

$$\begin{aligned} \mathsf{SH}_{ij} &= \sum_{r=1}^{R} \sum_{g=1}^{G} U_{nr} a^r_{ig} b^r_{j'g} + \sum_{r=1}^{R} U_{nr} (H(r)_{ij'} - M(r)_{ij'}) \\ &= \sum_{(r,g) \in [R] \times [G]} a^r_{ig} \times (U_{nr} b^r_{j'g}) + \sum_{r=1}^{R} U_{nr} (H(r)_{ij'} - M(r)_{ij'}). \end{aligned}$$

Define matrix $\mathsf{M} \in \mathbb{R}^{\lfloor T/2 \rfloor \times N \lfloor T/2 \rfloor}$ with its entry for row $i \in [\lfloor T/2 \rfloor]$ and column $j = (n-1) \times \lfloor T/2 \rfloor + j'$ for $j' \in [\lfloor T/2 \rfloor]$ given by

$$\begin{aligned} \mathsf{M}_{ij} &= \sum_{(r,g) \in [R] \times [G]} a^r_{ig} \times (U_{nr} b^r_{j'g}) \\ &= \sum_{(r,g) \in [R] \times [G]} \alpha_{i(r,g)} \beta_{j(r,g)}, \end{aligned}$$

where $\alpha_{i(r,g)} = a^r_{ig}$ and $\beta_{j(r,g)} = U_{nr} b^r_{j'g}$. Further,

$$\begin{aligned} |\mathsf{SH}_{ij} - \mathsf{M}_{ij}| &\le \sum_{r=1}^{R} |U_{nr}| |(H(r)_{ij'} - M(r)_{ij'})| \\ &\le \sum_{r=1}^{R} \Gamma_1 \|H(r) - M(r)\|_\infty \; \le \; R\Gamma_1 \epsilon. \end{aligned}$$

That is, the stacked Hankel matrix SH of N time series of $[T]$ has ϵ'-approximate rank $G \times R$ with $\epsilon' = R\Gamma_1 \epsilon$. This completes the proof.

Appendix E: Proofs For Section 5

E.1. Proof of Proposition 8

Let f_1, f_2 have a (G_1, ϵ_1) and (G_2, ϵ_2)-Hankel representation, respectively. For any $T \geq 1$, let $\boldsymbol{H}_1, \boldsymbol{H}_2 \in \mathbb{R}^{\lfloor T/2 \rfloor \times \lfloor T/2 \rfloor}$ be the Hankel matrices of f_1, f_2, respectively, over the time interval $[T]$. By definition, there exists matrices $M_1, M_2 \in \mathbb{R}^{\lfloor T/2 \rfloor \times \lfloor T/2 \rfloor}$ such that $\text{rank}(\boldsymbol{M}_1) \leq G_1$, $\|\boldsymbol{M}_1 - \boldsymbol{H}_1\|_\infty \leq \epsilon_1$ and $\text{rank}(\boldsymbol{M}_2) \leq G_2$, $\|\boldsymbol{M}_2 - \boldsymbol{H}_2\|_\infty \leq \epsilon_2$.

Component-wise addition. Note the Hankel matrix of $f_1 + f_2$ over $[T]$ is $\boldsymbol{H}_1 + \boldsymbol{H}_2$. Then, matrix $\boldsymbol{M} = \boldsymbol{M}_1 + \boldsymbol{M}_2$ has rank at most $G_1 + G_2$ since for any two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, it is the case that $\text{rank}(\boldsymbol{A}+\boldsymbol{B}) \leq \text{rank}(\boldsymbol{A}) + \text{rank}(\boldsymbol{B})$. Further, $\|\boldsymbol{H}_1 + \boldsymbol{H}_2 - (\boldsymbol{M}_1 + \boldsymbol{M}_2)\|_\infty \leq \epsilon_1 + \epsilon_2$. Therefore it follows that $f_1 + f_2$ has $(G_1 + G_2, \epsilon_1 + \epsilon_2)$-Hankel representation.

Component-wise multiplication. For $f_1 \circ f_2$, its Hankel over $[T]$ is given by $\boldsymbol{H}_1 \circ \boldsymbol{H}_2$ where we abuse notation of $\circ$ in the context of matrices as the Hadamard product of matrices. Let $\boldsymbol{M} = \boldsymbol{M}_1 \circ \boldsymbol{M}_2$. Then $\text{rank}(M) \leq G_1 \times G_2$ since for any two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\text{rank}(\boldsymbol{A} \circ \boldsymbol{B}) \leq \text{rank}(\boldsymbol{A})\text{rank}(\boldsymbol{B})$. Now

$$
\begin{aligned}
\|\boldsymbol{H}_1 \circ \boldsymbol{H}_2 - \boldsymbol{M}_1 \circ \boldsymbol{M}_2\|_\infty &\leq \|\boldsymbol{H}_1 \circ \boldsymbol{H}_2 - \boldsymbol{H}_1 \circ \boldsymbol{M}_2\|_\infty + \|\boldsymbol{H}_1 \circ \boldsymbol{M}_2 - \boldsymbol{M}_1 \circ \boldsymbol{M}_2\|_\infty \\
&\leq \|\boldsymbol{H}_1\|_\infty \|\boldsymbol{H}_2 - \boldsymbol{M}_2\|_\infty + \|\boldsymbol{M}_2\|_\infty \|\boldsymbol{H}_1 - \boldsymbol{M}_1\|_\infty \\
&\leq \|f_1\|_\infty \epsilon_2 + (\|\boldsymbol{M}_2 - \boldsymbol{H}_2\|_\infty + \|\boldsymbol{H}_2\|_\infty)\epsilon_1 \\
&\leq \|f_1\|_\infty \epsilon_2 + (\|f_2\|_\infty + \epsilon_2)\epsilon_1 \\
&= \|f_1\|_\infty \epsilon_2 + \|f_2\|_\infty \epsilon_1 + \epsilon_1 \epsilon_2 \ \leq \ 3 \max(\epsilon_1, \epsilon_2) \max(\|f_1\|_\infty, \|f_2\|_\infty).
\end{aligned}
$$

This completes the proof of Proposition 8.

E.2. Proof of Proposition 9

Proof is immediate from Definitions 4 and 5.

E.3. Proof of Proposition 10

E.3.1. Helper Lemmas for Proposition 10 We begin by stating some classic results from Fourier Analysis. To do so, we introduce some notation. Throughout, we have $R > 0$.

$C[0, R]$ and $L^2[0, R]$ functions. $C[0, R]$ is the set of real-valued, continuous functions defined on $[0, R]$. $L^2[0, R]$ is the set of square integrable functions defined on $[0, R]$, i.e. $\int_0^R f^2(t)dt \leq \infty$

Inner Product of functions in $L^2[0, R]$. $L^2[0, R]$ is a space endowed with inner product defined as $\langle f, g \rangle := \frac{1}{R} \int_0^R f(t)g(t)dt$, and associated norm as $\|f\| := \sqrt{\frac{1}{R} \int_0^R f^2(t)dt}$.

Fourier Representation of functions in $L^2[0,R]$. For $f \in L^2[0,R]$, define its $G \geq 1$-order Fourier representation, $\mathcal{F}(f,G) \in L^2[0,R]$ as

$$\mathcal{F}(f,G)(t) = a_0 + \sum_{g=1}^{G} (a_g \cos(2\pi g t/R) + b_g \cos(2\pi g t/R)), \quad t \in [0,R], \tag{13}$$

where a_0, a_g, b_g with $g \in [G]$ are called the Fourier coefficients of f, defined as

$$\begin{aligned}
a_0 &:= \langle f, 1 \rangle = \frac{1}{R} \int_0^R f(t)dt, \\
a_g &:= \langle f, \cos(2\pi g t/R) \rangle = \frac{1}{R} \int_0^R f(t) \cos(2\pi g t/R)dt, \\
b_g &:= \langle f, \sin(2\pi g t/R) \rangle = \frac{1}{R} \int_0^R f(t) \sin(2\pi g t/R)dt.
\end{aligned}$$

We now state a classic result from Fourier analysis.

THEOREM 7 (**Grafakos (2008)**). *Given $k \geq 1, R > 0$, let $f \in C^k(R, PER)$. Then, for any $t \in [0,R]$ (or more generally $t \in \mathbb{R}$),*

$$\lim_{G \to \infty} \mathcal{F}(f,G)(t) \to f(t).$$

We next argue that if $f \in C^k(R, \mathrm{PER})$, then its Fourier coefficients decay rapidly.

LEMMA 1. *Given $k \geq 1, R > 0$, let $f \in C^k(R, PER)$. Then, for $j \in [k]$, the G-order Fourier coefficient of $f^{(j)}$, the j-th derivative of f, recursively satisfy the following relationship: for $g \in [G]$,*

$$a_g^{(j)} = -\left(\frac{2\pi g}{R}\right) b_g^{(j-1)}, \qquad b_g^{(j)} = \left(\frac{2\pi g}{R}\right) a_g^{(j-1)}. \tag{14}$$

We establish (14) for $a_g^{(1)}$, $g \in [G]$. Notice that an identical argument applies to establish (14) for any $a_g^{(j)}, b_g^{(j)}$ for $j \in [k]$ and $g \in [G]$.

$$\begin{aligned}
a_g^{(1)} = \langle f^{(1)}, \cos(2\pi g t/R) \rangle &= \frac{1}{R} \int_0^R f^{(1)}(t) \cos(2\pi g t/R)dt \\
&\overset{(a)}{=} \frac{1}{R} \left(\left[f(t) \cos(2\pi g t/R) \right]_0^R - \frac{2\pi g}{R} \left[\frac{1}{R} \int_0^R f(t) \sin(2\pi g t/R)dt \right] \right) \\
&= -\left(\frac{2\pi g}{R}\right) b_g^{(0)}.
\end{aligned}$$

(a) follows by integration by parts.

E.3.2. Completing Proof of Proposition 10

For $G \in \mathbb{N}$, let $\mathcal{F}(f,G)$ be defined as in (13). Then for $t \in \mathbb{R}$

$$\begin{aligned}
|f(t) - \mathcal{F}(f,G)(t)| &\overset{(a)}{=} \Big| \sum_{g=G+1}^{\infty} \left(a_g \cos(2\pi g t/R) + b_g \cos(2\pi g t/R)\right)\Big| \\
&\leq \sum_{g=G+1}^{\infty} |a_g| + |b_g| \\
&\overset{(b)}{\leq} \sum_{g=G+1}^{\infty} \left(\frac{R}{2\pi g}\right)^k \left(|a_g^{(k)}| + |b_g^{(k)}|\right) \\
&\overset{(c)}{\leq} \sqrt{2}\left(\frac{R}{2\pi}\right)^k \sqrt{\sum_{g=G+1}^{\infty} \left(\frac{1}{g}\right)^{2k}} \sqrt{\sum_{g=G+1}^{\infty} \left(|a_g^{(k)}|^2 + |b_g^{(k)}|^2\right)} \\
&\overset{(d)}{\leq} \sqrt{2}\left(\frac{R}{2\pi}\right)^k \frac{1}{G^{k-0.5}} \sqrt{\sum_{g=G+1}^{\infty} \left(|a_g^{(k)}|^2 + |b_g^{(k)}|^2\right)} \\
&\overset{(e)}{\leq} \sqrt{2}\left(\frac{R}{2\pi}\right)^k \frac{\|f^{(k)}\|}{G^{k-0.5}} \\
&= C(k,R)\frac{\|f^{(k)}\|}{G^{k-0.5}},
\end{aligned}$$

where $C(k,R)$ is a constant that depends only on k and R; (a) follows from Theorem 7; (b) follows from Lemma 1; (c) follows from Cauchy-Schwarz inequality and fact that $(\alpha+\beta)^2 \leq 2(\alpha^2+\beta^2)$ for any $\alpha, \beta \in \mathbb{R}$; (d) $\sum_{g=G+1}^{\infty} g^{-2k} \leq \int_G^{\infty} x^{-2k} dx$ which can be bounded as $G^{-2k+1}/(2k-1)$ which is at most G^{-2k+1} since $k \geq 1$; (e) follows from Bessel's inequality, i.e. $\|f^{(k)}\|^2 \geq \sum_{g=0}^{\infty}(|a_g^{(k)}|^2 + |b_g^{(k)}|^2)$.

Thus, for any $t \in \mathbb{R}$, we have a uniform error bound for f being approximated by $\mathcal{F}(f,G)$ which is a sum of $2G$ harmonics. Noting $2G$ harmonics can be represented by an order-$4G$ LRF (by Proposition 1),we complete the proof.

E.4. Proof of Proposition 11

This analysis is adapted from Xu (2017).

Step 1: Partitioning the space $[0,1)^K$. Consider an equal partition of $[0,1)^K$. Precisely, for any $k \in \mathbb{N}$, we partition the the set $[0,1)$ into $1/k$ half-open intervals of length $1/k$, i.e, $[0,1) = \cup_{i=1}^{k} [(i-1)/k, i/k)$. It follows that $[0,1)^K$ can be partitioned into k^K cubes of forms $\otimes_{j=1}^{K} \left[(i_j-1)/k, i_j/k\right)$ with $i_j \in [k]$. Let $\mathcal{E}_k$ be such a partition with $I_1, I_2, \ldots, I_{k^K}$ denoting all such cubes and $z_1, z_2, \ldots, z_{k^K} \in \mathbb{R}^K$ denoting the centers of those cubes.

Step 2: Taylor Expansion of $g(\cdot,\omega)$. Consider a fixed ω. To reduce notational overload, we suppress dependence of g on ω, and abuse notation by using $g(\cdot) = g(\cdot,\omega)$ in what follows.

For every I_i with $1 \le i \le k^K$, define $P_{I_i,\ell}(x)$ as the degree-ℓ Taylor's series expansion of $g(x)$ at point z_i:

$$P_{I_i,\ell}(x) = \sum_{\kappa:|\kappa|\le\ell} \frac{1}{\kappa!}(x-z_i)^\kappa \nabla_\kappa g(z_i), \tag{15}$$

where $\kappa = (\kappa_1,\ldots,\kappa_d)$ is a multi-index with $\kappa! = \prod_{i=1}^K \kappa_i!$, and $\nabla_k g(z_i)$ is the partial derivative defined in Section 5.2. Note similar to g, $P_{I_i,\ell}(x)$ really refers to $P_{I_i,\ell}(x,\omega)$.

Now we define a degree-ℓ piecewise polynomial

$$P_{\mathcal{E}_k,\ell}(x) = \sum_{i=1}^{k^K} P_{I_i,\ell}(x)\mathbb{1}(x \in I_i).$$

For the remainder of the proof, let $\ell = \lfloor \alpha \rfloor$ (recall $\lfloor \alpha \rfloor$ refers to the largest integer strictly smaller than α). Since $f \in \mathcal{H}(\alpha, L)$, it follows that

$$\begin{aligned}
&\sup_{x\in[0,1)^K} \left|g(x) - P_{\mathcal{E}_k,\ell}(x)\right| = \max_{1\le i\le k^K} \sup_{x\in I_i} \left|g(x) - P_{I_i,\ell}(x)\right| \\
&\overset{(a)}{=} \max_{1\le i\le k^K} \sup_{x\in I_i} \left| \sum_{\kappa:|\kappa|\le\ell-1} \frac{\nabla_\kappa g(z_i)}{\kappa!}(x-z_i)^\kappa + \sum_{\kappa:|\kappa|=\ell} \frac{\nabla_\kappa g(\tilde{z}_i)}{\kappa!}(x-z_i)^\ell - P_{I_i,\ell}(x) \right| \\
&\overset{(b)}{=} \max_{1\le i\le k^K} \sup_{x\in I_i} \left| \sum_{\kappa:|\kappa|=\ell} \frac{\nabla_\kappa g(\tilde{z}_i)}{\kappa!}(x-z_i)^\ell - \sum_{\kappa:|\kappa|=\ell} \frac{\nabla_\kappa g(z_i)}{\kappa!}(x-z_i)^\ell \right| \\
&= \max_{1\le i\le k^K} \sup_{x\in I_i} \left| \sum_{\kappa:|\kappa|=\ell} \frac{\nabla_\kappa g(\tilde{z}_i) - \nabla_\kappa g(z_i)}{\kappa!}(x-z_i)^\ell \right| \\
&\overset{(c)}{\le} \max_{1\le i\le k^K} \sup_{x\in I_i} \|x-z_i\|_\infty^\ell \sup_{x\in I_i} \sum_{\kappa:|\kappa|=\ell} \frac{1}{\kappa!} \left|\nabla_\kappa g(\tilde{z}_i) - \nabla_\kappa g(z_i)\right| \\
&\overset{(d)}{\le} \mathcal{L} k^{-\alpha}.
\end{aligned} \tag{16}$$

where (a) follows from multivariate version of Taylor's theorem (and using the Lagrange form for the remainder) and $\tilde{z}_i \in [0,1)^K$ is a vector that can be represented as $z_i + cx$ for $c \in (0,1)$; (b) follows from (15); (c) follows from Holder's inequality; (d) follows from Definition 7.

Step 3: Construct Low-Rank Approximation of Time Series Hankel Using $P_{\mathcal{E}_k,\ell}$. Recall the Hankel matrix, $\boldsymbol{H} \in \mathbb{R}^{\lfloor T/2\rfloor \times \lfloor T/2\rfloor}$ induced by the original time series over $[T]$, where $\boldsymbol{H}_{ts} =$

$g(\theta_t, \omega_s),\ t, s \in [\lfloor T/2 \rfloor]$ with $g(\cdot, \omega) \in \mathcal{H}(\alpha, \mathcal{L})$ for any ω. We now construct a low-rank approximation of it using $P_{\mathcal{E}_k,\ell} = P_{\mathcal{E}_k,\ell}(\cdot, \omega)$. Define $\widetilde{\boldsymbol{H}} \in \mathbb{R}^{\lfloor T/2 \rfloor \times \lfloor T/2 \rfloor}$, where $\widetilde{\boldsymbol{H}}_{ts} = P_{\mathcal{E}_k,\ell}(\theta_t, \omega_s),\ t, s \in [\lfloor T/2 \rfloor]$.

By (16), we have that for all $t, s \in [\lfloor T/2 \rfloor]$,

$$\left| \boldsymbol{H}_{ts} - \widetilde{\boldsymbol{H}}_{ts} \right| \leq \mathcal{L} k^{-\alpha}.$$

It remains to bound the rank of $\widetilde{\boldsymbol{H}}$. Note that since $P_{\mathcal{E}_k,\ell}(\cdot, \omega)$ is a piecewise polynomial of degree $\ell = \lfloor \alpha \rfloor$ for any given ω, it has the following decomposition: for $t, s \in [\lfloor T/2 \rfloor]$,

$$\widetilde{\boldsymbol{H}}_{ts} = P_{\mathcal{E}_k,\ell}(\theta_t, \omega_s) \ = \sum_{i=1}^{k^K} \langle \Phi(\theta_t), \beta_{I_i,s} \rangle \mathbb{1}(\theta_t \in I_i)$$

where for any $\theta \in \mathbb{R}^K$,

$$\Phi(\theta) = \left(1, \theta_1, \ldots, \theta_K, \ldots, \theta_1^\ell, \ldots, \theta_K^\ell\right)^T,$$

the vector of all monomials of degree less than or equal to ℓ, and $\beta_{I_i,s}$ is a vector collecting the corresponding coefficients. The number of such monomials is easily shown to be equal to $C(\alpha, K) := \sum_{i=1}^{\lfloor \alpha \rfloor} \binom{i+K-1}{i}$. That is, $\widetilde{\boldsymbol{H}}_{ts} = u_t^T v_s$ where u_t, v_s are of dimension at most $k^K C(\alpha, K)$ for each $t, s \in [\lfloor T/2 \rfloor]$. That is, $\widetilde{\boldsymbol{H}}$ has rank at most $k^K C(\alpha, K)$. Setting $k = \left\lceil \frac{1}{\epsilon} \right\rceil$ completes the proof.

Appendix F: Helper Lemmas

We recall known concentration and perturbation inequalities that will be useful throughout.

THEOREM 8 **(Bernstein's Inequality Bernstein (1946))**. *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 9 **(Norm of matrices with sub-gaussian entries Vershynin (2010))**. *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, $K = \max_{i,j} \|A_{ij}\|_{\psi_2}$.

LEMMA 2 **(Maximum of sequence of random variables Vershynin (2010))**. *Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables, which are not necessarily independent, and satisfy $\mathbb{E}[X_i^{2p}]^{\frac{1}{2p}} \leq Kp^{\frac{\beta}{2}}$ for some $K, \beta > 0$ and all i. Then, for every $n \geq 2$,*

$$\mathbb{E}\max_{i\leq n}|X_i| \leq CK\log^{\frac{\beta}{2}}(n).$$

We note that Lemma 2 implies that if $X_1, \ldots, X_n$ are ψ_α random variables with $\|X_i\|_{\psi_\alpha} \leq K_\alpha$ for all $i \in [n]$, then

$$\mathbb{E}\max_{i\leq n}|X_i| \leq CK_\alpha\log^{\frac{1}{\alpha}}(n).$$

LEMMA 3 **(Modified Hoeffding Inequality Agarwal et al. (2020))**. *Let $X \in \mathbb{R}^n$ be random vector with independent mean-zero sub-Gaussian random coordinates with $\|X_i\|_{\psi_2} \leq K$. Let $a \in \mathbb{R}^n$ be another random vector that satisfies $\|a\|_2 \leq b$ almost surely for some constant $b \geq 0$. Then for all $t \geq 0$,*

$$\mathbb{P}\Big(\Big|\sum_{i=1}^n a_iX_i\Big| \geq t\Big) \leq 2\exp\Big(-\frac{ct^2}{K^2b^2}\Big),$$

where $c > 0$ is a universal constant.

LEMMA 4 **(Modified Hanson-Wright Inequality Agarwal et al. (2020))**. *Let $X \in \mathbb{R}^n$ be a random vector with independent mean-zero sub-Gaussian coordinates with $\|X_i\|_{\psi_2} \leq K$. Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a random matrix satisfying $\|\boldsymbol{A}\|_2 \leq a$ and $\|\boldsymbol{A}\|_F^2 \leq b$ almost surely for some $a, b \geq 0$. Then for any $t \geq 0$,*

$$\mathbb{P}\left(|X^T\boldsymbol{A}X - \mathbb{E}[X^T\boldsymbol{A}X]| \geq t\right) \leq 2\cdot\exp\Big(-c\min\Big(\frac{t^2}{K^4b}, \frac{t}{K^2a}\Big)\Big).$$

LEMMA 5 **(Weyl's inequality)**. *Given $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$, let σ_i and $\widehat{\sigma}_i$ be the i-th singular values of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively, in decreasing order and repeated by multiplicities. Then for all $i \in [m \wedge n]$,*

$$|\sigma_i - \widehat{\sigma}_i| \leq \|\boldsymbol{A} - \boldsymbol{B}\|_2.$$

Appendix G: Matrix Estimation via HSVT

This section describes and analyzes a well-known matrix estimation method, Hard Singular Value Thresholding (HSVT). While the analysis utilizes known arguments from the literature, we need to adapt it for the setting where the underlying 'signal' is only approximately low-rank.

G.1. Setup, Notations

Setup. Given a deterministic matrix $\boldsymbol{M} \in \mathbb{R}^{q \times p}$ with $p, q \in \mathbb{N}$ and $q \leq p$, a random matrix $\boldsymbol{Y} \in \mathbb{R}^{q \times p}$ is such that all of its entries, $Y_{ij},\ i \in [q],\ j \in [p]$ are mutually independent and for any given $i \in [q],\ j \in [p]$,

$$Y_{ij} = \begin{cases} M_{ij} + \boldsymbol{\varepsilon}_{ij} & \text{w.p. } \rho, \text{ (i.e. observed)} \\ 0 & \text{w.p. } 1-\rho, \text{ (i.e. not observed)} \end{cases}$$

for some $\rho \in (0, 1]$ with $\boldsymbol{\varepsilon}_{ij}$ are independent random variables with $\mathbb{E}[\boldsymbol{\varepsilon}_{ij}] = 0$ and $\|\boldsymbol{\varepsilon}_{ij}\|_{\psi_2} \leq \sigma$. Given this, we have $\mathbb{E}[\boldsymbol{Y}] = \rho \boldsymbol{M}$. Defineff

$$\widehat{\rho} = \max\Big(1/(q\ p), \big(\sum_{i=1}^{q}\sum_{j=1}^{p} \mathbf{1}(Y_{ij} \text{ is obs.})\big)/(q\ p)\Big).$$

Goal of Matrix Estimation. The goal of matrix estimation is to produce an estimate $\widehat{\boldsymbol{M}}$ from observation $\boldsymbol{Y}$ so that $\widehat{\boldsymbol{M}}$ is close to $\boldsymbol{M}$. In particular, we will be interested in bounding the error between $\widehat{\boldsymbol{M}}$ and $\boldsymbol{M}$ using the following metric: $\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\|_{2,\infty}$.

G.2. Matrix Estimation using HSVT

Hard Singular Value Thresholding (HSVT) Map. We define the HSVT map. For any $q, p \in \mathbb{N}$, consider a matrix $\boldsymbol{B} \in \mathbb{R}^{q \times p}$ such that $\boldsymbol{B} = \sum_{i=1}^{q \wedge p} \sigma_i(\boldsymbol{B}) x_i y_i^T$. Here for $i \in [q \wedge p]$, $\sigma_i(\boldsymbol{B})$ is the ith largest singular value of $\boldsymbol{B}$ and x_i, y_i are the corresponding left and right singular vectors respectively. Then, for given any $\lambda > 0$, we define the map $\mathsf{HSVT}_\lambda : \mathbb{R}^{q \times p} \to \mathbb{R}^{q \times p}$, which simply shaves off the singular values of the input matrix that are below the threshold λ. Precisely,

$$\mathsf{HSVT}_\lambda(\boldsymbol{B}) = \sum_{i=1}^{q \wedge p} \sigma_i(\boldsymbol{B}) \mathbb{1}(\sigma_i(\boldsymbol{B}) \geq \lambda) x_i y_i^T.$$

Matrix Estimating using HSVT map. We define a matrix estimation method using the HSVT map that is utilized by mSSA for imputation. Precisely, we estimate $\boldsymbol{M}$ from $\boldsymbol{Y}$ as follows: given parameter $k \geq 1$,

$$\widehat{\boldsymbol{M}} = \frac{1}{\widehat{\rho}} \mathsf{HSVT}_{\lambda_k}(\boldsymbol{Y}). \tag{17}$$

where $\lambda_k = \sigma_k(\boldsymbol{Y})$, i,e. the kth largest singular value of $\boldsymbol{Y}$.

G.3. A Useful Linear Operator

We define a linear map associated to HSVT. For a specific choice of $\lambda \geq 0$, define $\varphi_\lambda^{\boldsymbol{B}} : \mathbb{R}^p \to \mathbb{R}^p$ as follows: for any vector $w \in \mathbb{R}^p$ (i.e. $w \in \mathbb{R}^{p\times 1}$),

$$\varphi_\lambda^{\boldsymbol{B}}(w) = \sum_{i=1}^{q\wedge p} \mathbb{1}(\sigma_i(\boldsymbol{B}) \geq \lambda) y_i y_i^T w. \tag{18}$$

Note that $\varphi_\lambda^{\boldsymbol{B}}$ is a linear operator and it depends on the tuple $(\boldsymbol{B}, \lambda)$; more precisely, the singular values and the right singular vectors of $\boldsymbol{B}$, as well as the threshold λ. If $\lambda = 0$, then we will adopt the shorthand notation: $\varphi^{\boldsymbol{B}} = \varphi_0^{\boldsymbol{B}}$. The following is a simple, but curious relationship between $\varphi_\lambda^{\boldsymbol{B}}$ and HSVT_λ that will be useful subsequently.

LEMMA 6 **(Lemma 35 of Agarwal et al. (2019, 2021))**. *Let $\boldsymbol{B} \in \mathbb{R}^{q\times p}$ and $\lambda \geq 0$ be given. Then for any $j \in [q]$,*

$$\varphi_\lambda^{\boldsymbol{B}}\left(\boldsymbol{B}_{j\cdot}^T\right) = \mathrm{HSVT}_\lambda(\boldsymbol{B})_{j\cdot}^T,$$

where $\boldsymbol{B}_{j\cdot} \in \mathbb{R}^{1\times p}$ represents the jth row of $\boldsymbol{B}$, and $\mathrm{HSVT}_\lambda(\boldsymbol{B})_{j\cdot} \in \mathbb{R}^{1\times p}$ represents the jth row of the matrix obtained after applying HSVT over $\boldsymbol{B}$ with threshold λ.

By (18), the orthonormality of the right singular vectors and noting $\boldsymbol{B}_{j\cdot}^T = \boldsymbol{B}^T e_j$ with $e_j \in \mathbb{R}^p$ with jth entry 1 and everything else 0, we have

$$\begin{aligned}
\varphi_\lambda^{\boldsymbol{B}}\left(\boldsymbol{B}_{j\cdot}^T\right) &= \sum_{i=1}^{q\wedge p} \mathbb{1}(\sigma_i(\boldsymbol{B}) \geq \lambda) y_i y_i^T \boldsymbol{B}_{j\cdot}^T = \sum_{i=1}^{q\wedge p} \mathbb{1}(\sigma_i(\boldsymbol{B}) \geq \lambda) y_i y_i^T \boldsymbol{B}^T e_j \\
&= \sum_{i=1}^{q\wedge p} \mathbb{1}(\sigma_i(\boldsymbol{B}) \geq \lambda) y_i y_i^T \Big(\sum_{i'=1}^{q\wedge p} \sigma_{i'}(\boldsymbol{B}) x_{i'} y_{i'}^T\Big)^T e_j = \sum_{i,i'=1}^{q\wedge p} \sigma_{i'}(\boldsymbol{B}) \mathbb{1}(\sigma_i(\boldsymbol{B}) \geq \lambda) y_i y_i^T y_{i'} x_{i'}^T e_j \\
&= \sum_{i,i'=1}^{q\wedge p} \sigma_{i'}(\boldsymbol{B}) \mathbb{1}(\sigma_i(\boldsymbol{B}) \geq \lambda) y_i \delta_{ii'} x_{i'}^T e_j = \sum_{i=1}^{q\wedge p} \sigma_i(\boldsymbol{B}) \mathbb{1}(\sigma_i(\boldsymbol{B}) \geq \lambda) y_i x_i^T e_j \\
&= HSVT_\lambda(\boldsymbol{B})^T e_j = HSVT_\lambda(\boldsymbol{B})_{j\cdot}^T.
\end{aligned}$$

G.4. HSVT based Matrix Estimation: A Deterministic Bound

We state the following result about property of the estimator.

LEMMA 7. *For $k \geq 1$, let $\boldsymbol{M} = \boldsymbol{M}_k + \boldsymbol{E}_k$ with $rank(\boldsymbol{M}_k) = k$. Let $\varepsilon = \max(\widehat{\rho}/\rho, \rho/\widehat{\rho}) \geq 1$. Then, the HSVT estimate $\widehat{\boldsymbol{M}}$ with parameter k is such that for all $j \in [q]$,*

$$\begin{aligned}
\|\widehat{\boldsymbol{M}}_{j\cdot}^T - \boldsymbol{M}_{j\cdot}^T\|_2^2 \leq{}& \frac{2\|\boldsymbol{Y} - \rho\boldsymbol{M}\|_2^2 + 2\rho^2\|\boldsymbol{E}_k\|_2^2}{\left(\sigma_k(\rho\boldsymbol{M}_k)\right)^2}\left(2\left\|[\boldsymbol{M}_k]_{j\cdot}^T\right\|_2^2 + \frac{4\varepsilon^2\left(\|\boldsymbol{Y}_{j\cdot}^T - \rho\boldsymbol{M}_{j\cdot}^T\|_2\right)^2}{\rho^2}\right) \\
&+ \frac{4\varepsilon^2}{\rho^2}\left\|\varphi^{\boldsymbol{M}_k}(\boldsymbol{Y}_{j\cdot}^T - \rho\boldsymbol{M}_{j\cdot}^T)\right\|_2^2 + 2(\varepsilon-1)^2\|\boldsymbol{M}_{j\cdot}^T\|_2^2 + 2\left\|[\boldsymbol{E}_k]_{j\cdot}^T\right\|_2^2.
\end{aligned}$$

We prove our lemma in four steps.

Step 1. Decomposing $\widehat{\boldsymbol{M}}_{j\cdot}^T - \boldsymbol{M}_{j\cdot}^T$ in two terms. Fix a row index $j \in [q]$. Let λ_k be the kth largest singular value of $\boldsymbol{Y}$, as used by HSVT algorithm with parameter $k \geq 1$.

$$\widehat{\boldsymbol{M}}_{j\cdot}^T - \boldsymbol{M}_{j\cdot}^T = \left(\widehat{\boldsymbol{M}}_{j\cdot}^T - \varphi_{\lambda_k}^{\boldsymbol{Y}}\left(\boldsymbol{M}_{j\cdot}^T\right)\right) + \left(\varphi_{\lambda_k}^{\boldsymbol{Y}}\left(\boldsymbol{M}_{j\cdot}^T\right) - \boldsymbol{M}_{j\cdot}^T\right).$$

By definition per (18), $\varphi_{\lambda_k}^{\boldsymbol{Y}} : \mathbb{R}^p \to \mathbb{R}^p$ is the projection operator onto $\text{span}\{u_1, \ldots, u_k\}$, the span of top k right singular vectors of $\boldsymbol{Y}$, denoted as $u_1, \ldots, u_k$. Therefore,

$$\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T) - \boldsymbol{M}_{j\cdot}^T \in \text{span}\{u_1, \ldots, u_k\}^{\perp}.$$

By design, $\text{rank}(\widehat{\boldsymbol{M}}) = k$. Therefore, by Lemma 6

$$\widehat{\boldsymbol{M}}_{j\cdot} - \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T) = \frac{1}{\widehat{\rho}}\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{Y}_{j\cdot}^T) - \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T) \in \text{span}\{u_1, \ldots, u_k\}.$$

Therefore, $\langle \widehat{\boldsymbol{M}}_{j\cdot}^T - \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T), \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T) - \boldsymbol{M}_{j\cdot}^T \rangle = 0$, and hence

$$\left\|\widehat{\boldsymbol{M}}_{j\cdot}^T - \boldsymbol{M}_{j\cdot}^T\right\|_2^2 = \left\|\widehat{\boldsymbol{M}}_{j\cdot}^T - \varphi_{\lambda_k}^{\boldsymbol{Y}}\left(\boldsymbol{M}_{j\cdot}^T\right)\right\|_2^2 + \left\|\varphi_{\lambda_k}^{\boldsymbol{Y}}\left(\boldsymbol{M}_{j\cdot}^T\right) - \boldsymbol{M}_{j\cdot}^T\right\|_2^2 \tag{19}$$

by the Pythagorean theorem.

Step 2. Bounding Term 1, $\left\|\widehat{\boldsymbol{M}}_{j\cdot}^T - \varphi_{\lambda_k}^{\boldsymbol{Y}}\left(\boldsymbol{M}_{j\cdot}^T\right)\right\|_2$. We begin by bounding the first term on the right hand side of (19). By Lemma 6,

$$\begin{aligned}\widehat{\boldsymbol{M}}_{j\cdot} - \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T) &= \frac{1}{\widehat{\rho}}\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{Y}_{j\cdot}^T) - \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T) = \varphi_{\lambda_k}^{\boldsymbol{Y}}\left(\frac{1}{\widehat{\rho}}\boldsymbol{Y}_{j\cdot}^T - \boldsymbol{M}_{j\cdot}^T\right) \\ &= \frac{1}{\widehat{\rho}}\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{Y}_{j\cdot}^T - \rho\boldsymbol{M}_{j\cdot}^T) + \frac{\rho - \widehat{\rho}}{\widehat{\rho}}\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T).\end{aligned}$$

Using the Parallelogram Law (or, equivalently, combining Cauchy-Schwartz and AM-GM inequalities), we obtain

$$\begin{aligned}\|\widehat{\boldsymbol{M}}_{j\cdot}^T - \varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot})^T\|_2^2 &= \|\frac{1}{\widehat{\rho}}\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T - \rho\boldsymbol{M}_{j\cdot}^T) + \frac{\rho - \widehat{\rho}}{\widehat{\rho}}\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T)\|_2^2 \\ &\leq 2\,\|\frac{1}{\widehat{\rho}}\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{Y}_{j\cdot}^T - \rho\boldsymbol{M}_{j\cdot}^T)\|_2^2 + 2\,\|\frac{\rho - \widehat{\rho}}{\widehat{\rho}}\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{M}_{j\cdot}^T)\|_2^2 \\ &\leq \frac{2}{\widehat{\rho}^2}\|\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{Y}_{j\cdot}^T - \rho\boldsymbol{M}_{j\cdot}^T)\|_2^2 + 2\left(\frac{\rho - \widehat{\rho}}{\widehat{\rho}}\right)^2\|\boldsymbol{M}_{j\cdot}^T\|_2^2 \\ &\leq \frac{2\varepsilon^2}{\rho^2}\|\varphi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{Y}_{j\cdot}^T - \rho\boldsymbol{M}_{j\cdot}^T)\|_2^2 + 2(\varepsilon - 1)^2\|\boldsymbol{M}_{j\cdot}^T\|_2^2. \end{aligned}\tag{20}$$

From definition of ε, $\frac{1}{\widehat{\rho}} \leq \frac{\varepsilon}{\rho}$ and $\left(\frac{\rho-\widehat{\rho}}{\widehat{\rho}}\right)^2 \leq (\varepsilon-1)^2$. The first term of (20) can be decomposed as,

$$\begin{aligned}
&\|\varphi^{\boldsymbol{Y}}_{\lambda_k}(\boldsymbol{Y}^T_{j\cdot} - \rho \boldsymbol{M}^T_{j\cdot})\|_2^2 \\
&\leq 2\left\|\varphi^{\boldsymbol{Y}}_{\lambda_k}(\boldsymbol{Y}^T_{j\cdot} - \rho \boldsymbol{M}^T_{j\cdot}) - \varphi^{\boldsymbol{M}_k}(\boldsymbol{Y}^T_{j\cdot} - \rho \boldsymbol{M}^T_{j\cdot})\right\|_2^2 + 2\left\|\varphi^{\boldsymbol{M}_k}(\boldsymbol{Y}^T_{j\cdot} - \rho \boldsymbol{M}^T_{j\cdot})\right\|_2^2. \qquad (21)
\end{aligned}$$

In above, we have used notation $\varphi^{\boldsymbol{M}_k} = \varphi_0^{\boldsymbol{M}_k}$. Given that $\boldsymbol{M}_k$ is rank k matrix, $\varphi^{\boldsymbol{M}_k} : \mathbb{R}^p \to \mathbb{R}^p$ is the projection operator mapping any element in $\mathbb{R}^p$ to the projection onto the subspace spanned by $\{\mu_1, \ldots, \mu_k\}$, where $\mu_1, \ldots, \mu_k \in \mathbb{R}^p$ are the k non-trivial right singular vectors of $\boldsymbol{M}_k$. Similarly, by definition $\varphi^{\boldsymbol{Y}}_{\lambda_k}$ is a map $\mathbb{R}^p \to \mathbb{R}^p$ mapping any element in $\mathbb{R}^p$ to its projection onto the subspace spanned by $\{u_1, \ldots, u_k\}$, the top k right singular vectors of $\boldsymbol{Y}$–this can be seen by noting $\lambda_k = \sigma_k(\boldsymbol{Y})$ is the k-th top singular value of $\boldsymbol{Y}$. Recall $\sigma_j(\boldsymbol{Y}),\ j \in [q \wedge p]$ is the jth largest singular value of $\boldsymbol{Y}$.

Next, we bound the first term on the right hand side of (21). To that end, by Wedin $\sin\Theta$ Theorem (see Davis and Kahan (1970), Wedin (1972)) and recalling $\text{rank}(\boldsymbol{M}_k) = k$,

$$\begin{aligned}
\left\|\varphi^{\boldsymbol{Y}}_{\lambda_k} - \varphi^{\boldsymbol{M}_k}\right\|_2 &\leq \frac{\|\boldsymbol{Y} - \rho \boldsymbol{M}_k\|_2}{\sigma_k(\rho \boldsymbol{M}_k)} \\
&\leq \frac{\|\boldsymbol{Y} - \rho \boldsymbol{M}\|_2}{\sigma_k(\rho \boldsymbol{M}_k)} + \frac{\rho\|\boldsymbol{M} - \boldsymbol{M}_k\|_2}{\sigma_k(\rho \boldsymbol{M}_k)} \\
&\leq \frac{\|\boldsymbol{Y} - \rho \boldsymbol{M}\|_2}{\sigma_k(\rho \boldsymbol{M}_k)} + \frac{\rho\|\boldsymbol{E}_k\|_2}{\sigma_k(\rho \boldsymbol{M}_k)}. \qquad (22)
\end{aligned}$$

Then it follows that

$$\begin{aligned}
\left\|\varphi^{\boldsymbol{Y}}_{\lambda_k}(\boldsymbol{Y}^T_{j\cdot} - \rho \boldsymbol{M}^T_{j\cdot}) - \varphi^{\boldsymbol{M}_k}(\boldsymbol{Y}^T_{j\cdot} - \rho \boldsymbol{M}^T_{j\cdot})\right\|_2 &\leq \|\varphi^{\boldsymbol{Y}}_{\lambda_k} - \varphi^{\boldsymbol{M}_k}\|_2 \|\boldsymbol{Y}^T_{j\cdot} - \rho \boldsymbol{M}^T_{j\cdot}\|_2 \\
&\leq \frac{\left(\|\boldsymbol{Y} - \rho \boldsymbol{M}\|_2 + \rho\|\boldsymbol{E}_k\|_2\right)\left(\|\boldsymbol{Y}^T_{j\cdot} - \rho \boldsymbol{M}^T_{j\cdot}\|_2\right)}{\sigma_k(\rho \boldsymbol{M}_k)}. \qquad (23)
\end{aligned}$$

Using (21) and (23) in (20),

$$\begin{aligned}
\|\widehat{\boldsymbol{M}}_{j\cdot} - \varphi^{\boldsymbol{Y}}_{\lambda_k}(\boldsymbol{M}^T_{j\cdot})\|_2^2 \leq \frac{4\varepsilon^2}{\rho^2} &\frac{\left(\|\boldsymbol{Y} - \rho \boldsymbol{M}\|_2 + \rho\|\boldsymbol{E}_k\|_2\right)^2\left(\|\boldsymbol{Y}^T_{j\cdot} - \rho \boldsymbol{M}^T_{j\cdot}\|_2\right)^2}{\left(\sigma_k(\rho \boldsymbol{M}_k)\right)^2} \\
&+ \frac{4\varepsilon^2}{\rho^2}\left\|\varphi^{\boldsymbol{M}_k}(\boldsymbol{Y}^T_{j\cdot} - \rho \boldsymbol{M}^T_{j\cdot})\right\|_2^2 + 2(\varepsilon-1)^2\|\boldsymbol{M}^T_{j\cdot}\|_2^2. \qquad (24)
\end{aligned}$$

Step 3. Bounding Term 2, $\left\|\varphi^{\boldsymbol{Y}}_{\lambda_k}(\boldsymbol{M}^T_{j\cdot}) - \boldsymbol{M}^T_{j\cdot}\right\|_2^2$. Recall $\boldsymbol{M} = \boldsymbol{M}_k + \boldsymbol{E}_k$ and using (22),

$$\left\|\varphi^{\boldsymbol{Y}}_{\lambda_k}(\boldsymbol{M}^T_{j\cdot}) - \boldsymbol{M}^T_{j\cdot}\right\|_2^2 = \left\|\varphi^{\boldsymbol{Y}}_{\lambda_k}([\boldsymbol{M}_k]^T_{j\cdot} + [\boldsymbol{E}_k]^T_{j\cdot}) - [\boldsymbol{M}_k]^T_{j\cdot} - [\boldsymbol{E}_k]^T_{j\cdot}\right\|_2^2$$

$$
\begin{aligned}
&\leq 2\left\|\varphi^{\boldsymbol{Y}}_{\lambda_k}\left([\boldsymbol{M}_k]^T_{j\cdot}\right) - [\boldsymbol{M}_k]^T_{j\cdot}\right\|_2^2 + 2\left\|\varphi^{\boldsymbol{Y}}_{\lambda_k}\left([\boldsymbol{E}_k]^T_{j\cdot}\right) - [\boldsymbol{E}_k]^T_{j\cdot}\right\|_2^2 \\
&= 2\left\|\varphi^{\boldsymbol{Y}}_{\lambda_k}\left([\boldsymbol{M}_k]^T_{j\cdot}\right) - \varphi^{\boldsymbol{M}_k}_{\lambda_k}\left([\boldsymbol{M}_k]^T_{j\cdot}\right)\right\|_2^2 + 2\left\|\varphi^{\boldsymbol{Y}}_{\lambda_k}\left([\boldsymbol{E}_k]^T_{j\cdot}\right) - [\boldsymbol{E}_k]^T_{j\cdot}\right\|_2^2 \\
&\leq 2\left\|\varphi^{\boldsymbol{Y}}_{\lambda_k} - \varphi^{\boldsymbol{M}_k}_{\lambda_k}\right\|_2^2 \left\|[\boldsymbol{M}_k]^T_{j\cdot}\right\|_2^2 + 2\left\|[\boldsymbol{E}_k]^T_{j\cdot}\right\|_2^2 \\
&\leq 2\frac{\left(\|\boldsymbol{Y}-\rho\boldsymbol{M}\|_2 + \rho\|\boldsymbol{E}_k\|\right)^2}{\left(\sigma_k(\rho\boldsymbol{M}_k)\right)^2}\left\|[\boldsymbol{M}_k]^T_{j\cdot}\right\|_2^2 + 2\left\|[\boldsymbol{E}_k]^T_{j\cdot}\right\|_2^2. \qquad (25)
\end{aligned}
$$

Step 4. Putting everything together. Inserting (24) and (25) back to (19), we have that for each $j \in [q]$,

$$
\begin{aligned}
\left\|\widehat{\boldsymbol{M}}^T_{j\cdot} - \boldsymbol{M}^T_{j\cdot}\right\|_2^2 &\leq 2\frac{\left(\|\boldsymbol{Y}-\rho\boldsymbol{M}\|_2 + \rho\|\boldsymbol{E}_k\|_2\right)^2}{\left(\sigma_k(\rho\boldsymbol{M}_k)\right)^2}\left\|[\boldsymbol{M}_k]^T_{j\cdot}\right\|_2^2 + 2\left\|[\boldsymbol{E}_k]^T_{j\cdot}\right\|_2^2 \\
&\quad + \frac{4\varepsilon^2}{\rho^2}\frac{\left(\|\boldsymbol{Y}-\rho\boldsymbol{M}\|_2 + \rho\|\boldsymbol{E}_k\|_2\right)^2\left(\|\boldsymbol{Y}^T_{j\cdot} - \rho\boldsymbol{M}^T_{j\cdot}\|_2\right)^2}{\left(\sigma_k(\rho\boldsymbol{M}_k)\right)^2} \\
&\quad + \frac{4\varepsilon^2}{\rho^2}\left\|\varphi^{\boldsymbol{M}_k}(\boldsymbol{Y}^T_{j\cdot} - \rho\boldsymbol{M}^T_{j\cdot})\right\|_2^2 + 2(\varepsilon-1)^2\|\boldsymbol{M}^T_{j\cdot}\|_2^2 \\
&\leq \frac{2\|\boldsymbol{Y}-\rho\boldsymbol{M}\|_2^2 + 2\rho^2\|\boldsymbol{E}_k\|_2^2}{\left(\sigma_k(\rho\boldsymbol{M}_k)\right)^2}\left(2\left\|[\boldsymbol{M}_k]^T_{j\cdot}\right\|_2^2 + \frac{4\varepsilon^2\left(\|\boldsymbol{Y}^T_{j\cdot} - \rho\boldsymbol{M}^T_{j\cdot}\|_2\right)^2}{\rho^2}\right) \\
&\quad + \frac{4\varepsilon^2}{\rho^2}\left\|\varphi^{\boldsymbol{M}_k}(\boldsymbol{Y}^T_{j\cdot} - \rho\boldsymbol{M}^T_{j\cdot})\right\|_2^2 + 2(\varepsilon-1)^2\|\boldsymbol{M}^T_{j\cdot}\|_2^2 + 2\left\|[\boldsymbol{E}_k]^T_{j\cdot}\right\|_2^2,
\end{aligned}
$$

where we used $(a+b)^2 \leq 2a^2 + 2b^2$. This completes the proof.

G.5. HSVT based Matrix Estimation: Deterministic To High-Probability

Next, we convert the bound obtained in Lemma 7 to a bound in expectation (as well as one in high-probability) for our metric of interest: $\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\|_{2,\infty}$. In particular, we establish

THEOREM 10. *For $k \geq 1$, let $\boldsymbol{M} = \boldsymbol{M}_k + \boldsymbol{E}_k$ with $rank(\boldsymbol{M}_k) = k$. Let $\epsilon = \|\boldsymbol{E}_k\|_\infty$ and $\Gamma = \|\boldsymbol{M}_k\|_\infty$. Let $\rho \geq C\log(qp)/q$ for C large enough and $q \leq p$. Then, the HSVT estimate $\widehat{\boldsymbol{M}}$ with parameter k is such that*

$$
\mathbb{E}\Big[\max_{j\in[q]}\frac{1}{p}\|\widehat{\boldsymbol{M}}^T_{j\cdot} - \boldsymbol{M}^T_{j\cdot}\|_2^2\Big] \leq \frac{p(C\sigma^2 + \rho^2\epsilon q)}{\rho^2\sigma_k(\boldsymbol{M}_k)^2}\left(\Gamma^2 + \frac{\sigma^2}{\rho^2}\right) + \frac{C\sigma^2 k\log p}{p\rho^2} + \frac{C(\Gamma+\epsilon)^2}{p} + 2\epsilon^2 + \frac{C}{(pq)^2}.
$$

We start by identifying certain high probability events. Subsequently, using these events and Lemma 7, we shall conclude the proof.

High Probability Events. For some positive absolute constant $C > 0$, define

$$
\begin{aligned}
E_1 &:= \Big\{ |\widehat{\rho} - \rho| \le \rho/20 \Big\}, \\
E_2 &:= \Big\{ \|\boldsymbol{Y} - \rho \boldsymbol{M}\|_2 \le C\sigma\sqrt{p} \Big\}, \qquad (26) \\
E_3 &:= \Big\{ \|\boldsymbol{Y} - \rho \boldsymbol{M}\|_{\infty,2}, \|\boldsymbol{Y} - \rho \boldsymbol{M}\|_{2,\infty} \le C\sigma\sqrt{p} \Big\}, \qquad (27) \\
E_4 &:= \Big\{ \max_{j \in [q]} \|\varphi^{\boldsymbol{B}}_{\sigma_k(\boldsymbol{B})}\Big(\boldsymbol{Y}^T_{j\cdot} - \rho \boldsymbol{M}^T_{j\cdot}\Big)\|_2^2 \le C\sigma^2 k \log(p) \Big\}, \qquad (28) \\
E_5 &:= \left\{ \left(1 - \sqrt{\frac{20\log(qp)}{\rho qp}}\right)\rho \le \widehat{\rho} \le \frac{1}{1 - \sqrt{\frac{20\log(qp)}{\rho qp}}}\rho \right\}.
\end{aligned}
$$

In (28) above, $\boldsymbol{B} \in \mathbb{R}^{q \times p}$ is a deterministic matrix. Let the singular value decomposition of $\boldsymbol{B}$ be given as $\boldsymbol{B} = \sum_{i=1}^{q} \sigma_i(\boldsymbol{B}) x_i y_i^T$, where $\sigma_i(\boldsymbol{B})$ are the singular vectors of $\boldsymbol{B}$ in decreasing order and x_i, y_i are the left and right singular vectors respectively. Recall the definition of $\varphi^{\boldsymbol{B}}_{\lambda}$ in (18). In particular, we choose $\lambda = \sigma_k(\boldsymbol{B})$, the kth singular value of $\boldsymbol{B}$ in (28). As a result, in effect, we are bounding norm of projection of random vector $\boldsymbol{Y}_{j\cdot} - \rho \boldsymbol{M}_{j\cdot}$ for any given deterministic subspace of $\mathbb{R}^p$ of dimension k.

LEMMA 8. *For some positive constant $c_1 > 0$ and $C > 0$ large enough in definitions of $E_1, \ldots, E_5$,*

$$
\begin{aligned}
\mathbb{P}(E_1) &\ge 1 - 2e^{-c_1 pq\rho} - (1-\rho)^{pq}, \\
\mathbb{P}(E_2) &\ge 1 - 2e^{-p}, \\
\mathbb{P}(E_3) &\ge 1 - 2e^{-p}, \qquad (29) \\
\mathbb{P}(E_4) &\ge 1 - \frac{2}{(qp)^{10}}. \\
\mathbb{P}(E_5) &\ge 1 - \frac{2}{(qp)^{10}}.
\end{aligned}
$$

We bound the probability of events $E_1, \ldots, E_5$ in that order.

Bounding E_1. Let

$$
\widehat{\rho}_0 = \Big(\sum_{i=1}^{q} \sum_{j=1}^{p} \mathbf{1}(Y_{ij} \text{ is obs.}) \Big) / (q\ p).
$$

That is, $\widehat{\rho} = \max(\widehat{\rho}_0, 1/(pq))$ and $\mathbb{E}[\widehat{\rho}_0] = \rho$. We define the event $E_6 := \{\widehat{\rho}_0 = \widehat{\rho}\}$. Thus, we have that

$$
\mathbb{P}(E_1^c) = \mathbb{P}(E_1^c \cap E_6) + \mathbb{P}(E_1^c \cap E_6^c)
$$

$$
\begin{aligned}
&= \mathbb{P}(|\widehat{\rho}_0 - \rho| \geq \rho/20) + \mathbb{P}(E_1^c \cap E_6^c) \\
&\leq \mathbb{P}(|\widehat{\rho}_0 - \rho| \geq \rho/20) + \mathbb{P}(E_6^c) \\
&= \mathbb{P}(|\widehat{\rho}_0 - \rho| \geq \rho/20) + (1-\rho)^{qp},
\end{aligned}
$$

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

$$\mathbb{P}(|\widehat{\rho}_0 - \rho| \geq \rho/20) \geq 1 - 2e^{-c_1 \rho q p}.$$

Bounding $\boldsymbol{E}_2$. To start with, $\mathbb{E}[\boldsymbol{Y}] = \rho \boldsymbol{M}$. For any $i \in [q], j \in [p]$, the Y_{ij} are independent, 0 with probability $1-\rho$ and with probability ρ equal to $M_{ij} + \boldsymbol{\varepsilon}_{ij}$ with $\|\boldsymbol{\varepsilon}_{ij}\|_{\psi_2} \leq \sigma$. Therefore, it follows that $\|Y_{ij} - \rho M_{ij}\|_{\psi_2} \leq C'\sigma$ for a constant $C' > 0$. Since $q \leq p$, using Theorem 9 it follows that for an appropriately large constant $C > 0$,

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

Bounding $\boldsymbol{E}_3$. Recall that we assume $q \leq p$. Observe that for any matrix $A \in \mathbb{R}^{q \times p}$, $\|A\|_{\infty,2},\ \|A\|_{2,\infty} \leq \|A\|_2$. Thus using the argument to bound $\boldsymbol{E}_2$, we have (29).

Bounding $\boldsymbol{E}_4$. Consider for $j \in [q]$,

$$\|\varphi^{\boldsymbol{B}}_{\sigma_k(\boldsymbol{B})}\left(\boldsymbol{Y}_{j\cdot}^T - \rho \boldsymbol{M}_{j\cdot}^T\right)\|_2^2 = \sum_{i=1}^k \|y_i y_i^T (\boldsymbol{Y}_{j\cdot}^T - \rho \boldsymbol{M}_{j\cdot}^T)\|_2^2 \qquad \leq \sum_{i=1}^k \left(y_i^T (\boldsymbol{Y}_{j\cdot}^T - \rho \boldsymbol{M}_{j\cdot}^T)\right)_2^2 = \sum_{i=1}^k Z_i^2,$$

where $Z_i = y_i^T (\boldsymbol{Y}_{j\cdot}^T - \rho \boldsymbol{M}_{j\cdot}^T)$. By definition of the ψ_2 norm of a random variable and since y_i is unit norm vector that is deterministic (and hence independent the of random vector $\boldsymbol{Y}_{j\cdot}^T - p\boldsymbol{M}_{j\cdot}^T$), it follows that

$$\|Z_i\|_{\psi_2} = \|y_i^T (\boldsymbol{Y}_{j\cdot} - p\boldsymbol{M}_{j\cdot})\|_{\psi_2} \leq \|(\boldsymbol{Y}_{j\cdot} - p\boldsymbol{M}_{j\cdot})\|_{\psi_2}.$$

Since the coordinates of $\boldsymbol{Y}_{j\cdot}^T - \rho \boldsymbol{M}_{j\cdot}^T$ are mean-zero and independent, with ψ_2 norm bounded by $\sqrt{C}\sigma$ for some absolute constant $C > 0$, using arguments from Agarwal et al. (2019, 2021), it follows that

$$\mathbb{P}\Big(\sum_{i=1}^k Z_i^2 > t\Big) \leq 2k \exp\Big(-\frac{t}{kC\sigma^2}\Big).$$

Therefore, for choice of $t = C\sigma^2 k \log p$ with large enough constant $C > 0$, $q \le p$, and taking a union bound over all $j \in [p]$, we have that

$$\mathbb{P}\left(E_4^c\right) \le \frac{2}{(qp)^{10}}.$$

Bounding E_5. Recall the definition of $\widehat{\rho}$. By the binomial Chernoff bound, for $\varepsilon > 1$,

$$\mathbb{P}\left(\widehat{\rho} > \varepsilon\rho\right) \le \exp\left(-\frac{(\varepsilon-1)^2}{\varepsilon+1} qp\rho\right), \quad \text{and}$$
$$\mathbb{P}\left(\widehat{\rho} < \frac{1}{\varepsilon}\rho\right) \le \exp\left(-\frac{(\varepsilon-1)^2}{2\varepsilon^2} qp\rho\right).$$

By the union bound,

$$\mathbb{P}\left(\frac{1}{\varepsilon}\rho \le \widehat{\rho} \le \rho\varepsilon\right) \ge 1 - \mathbb{P}\left(\widehat{\rho} > \varepsilon\rho\right) - \mathbb{P}\left(\widehat{\rho} < \frac{1}{\varepsilon}\rho\right).$$

Noticing $\varepsilon + 1 < 2\varepsilon < 2\varepsilon^2$ for all $\varepsilon > 1$, and substituting $\varepsilon = \left(1 - \sqrt{\frac{20\log(qp)}{qp\rho}}\right)^{-1}$ completes the proof. The following are immediate corollaries of the above stated bounds.

COROLLARY 2. *Let $E := E_1 \cap E_2$. Then, for $\rho \ge C\log(qp)/q$,*

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

where C_1 and c_2 are positive constants.

COROLLARY 3. *Let $E := E_2 \cap E_3 \cap E_4 \cap E_5$. Then,*

$$\mathbb{P}(E^c) \le \frac{C_1}{(qp)^{10}},$$

where C_1 is an absolute positive constant.

Probabilistic Bound for HSVT based Matrix Estimation. Recall $\epsilon = \|\boldsymbol{E}_k\|_\infty$. Then $\|\boldsymbol{E}_k\|_F^2 \le \epsilon qp$. And $\|\boldsymbol{E}_k\|_2^2 \le \|\boldsymbol{E}_k\|_F^2 \le \epsilon qp$. Let $\rho \ge C\log(qp)/q$ for C large enough and recall $q \le p$. Further, recall $\Gamma = \|\boldsymbol{M}_k\|_\infty$; thus, $\|\boldsymbol{M}\|_\infty \le \Gamma + \epsilon$. Then $\|[\boldsymbol{M}_k]_{j\cdot}^T\|_2 \le \Gamma\sqrt{p}$ and $\|[\boldsymbol{M}]_{j\cdot}^T\|_2 \le (\Gamma+\epsilon)\sqrt{p}$.

Define $E = E_1 \cap E_2 \cap E_3 \cap E_4 \cap E_5$. Then, from Corollaries 2 and 3, we have that $\mathbb{P}(E^c) \le \frac{C_1}{(qp)^{10}}$ for large enough constant $C_1 > 0$.

Under E_5, we have $\varepsilon = \max(\widehat{\rho}/\rho, \rho/\widehat{\rho}) \le \left(1 - \sqrt{\frac{20\log(qp)}{qp\rho}}\right)^{-1}$. Under this choice of ε and using $\rho \ge C\log(qp)/q$, we have that for C large enough, $\varepsilon \le C$ and $(\varepsilon-1)^2 \le C/p$.

Given this setup, under event E, Lemma 7 leads to the following: for all $j \in [q]$ and with appropriately (re-defined) large enough constant $C > 0$,

$$\|\widehat{\boldsymbol{M}}_{j\cdot}^T - \boldsymbol{M}_{j\cdot}^T\|_2^2 \le C\frac{\sigma^2 p + \rho^2\epsilon qp}{\rho^2\sigma_k(\boldsymbol{M}_k)^2}\left(p\Gamma^2 + \frac{\sigma^2 p}{\rho^2}\right) + \frac{C\sigma^2 k\log p}{\rho^2} + C(\Gamma+\epsilon)^2 + 2p\epsilon^2. \tag{30}$$

That is, under event E,

$$\max_{j\in[q]}\frac{1}{p}\|\widehat{\boldsymbol{M}}_{j\cdot}^T - \boldsymbol{M}_{j\cdot}^T\|_2^2 \le C\frac{p(\sigma^2+\rho^2\epsilon q)}{\rho^2\sigma_k(\boldsymbol{M}_k)^2}\left(\Gamma^2 + \frac{\sigma^2}{\rho^2}\right) + \frac{C\sigma^2 k\log p}{p\rho^2} + \frac{C(\Gamma+\epsilon)^2}{p} + 2\epsilon^2. \tag{31}$$

For any random variable X and event A, such that under event A, $X \le B$ and $\mathbb{P}(A^c) \le \delta$, we have

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[X\mathbb{1}(A)] + \mathbb{E}[X\mathbb{1}(A^c)] \\ &\le \mathbb{E}[X\mathbb{1}(A)] + \mathbb{E}[X^2]^{\frac{1}{2}}\mathbb{P}(A^c)^{\frac{1}{2}} \\ &\le B + \mathbb{E}[X^2]^{\frac{1}{2}}\delta^{\frac{1}{2}}.\end{aligned} \tag{32}$$

We shall use this reasoning above to bound $\mathbb{E}\big[\max_{j\in[q]}\frac{1}{p}\|\widehat{\boldsymbol{M}}_{j\cdot}^T - \boldsymbol{M}_{j\cdot}^T\|_2^2\big]$: let $X = \max_{j\in[q]}\frac{1}{p}\|\widehat{\boldsymbol{M}}_{j\cdot}^T - \boldsymbol{M}_{j\cdot}^T\|_2^2$ and $A = E$; B is given by right hand side of (31), $\delta = \frac{C_1}{(qp)^{10}}$; the only missing quantity that remains to be bounded is $\mathbb{E}[X^2]$. We do that next.

To begin with, for any $j \in [q]$,

$$\|\widehat{\boldsymbol{M}}_{j\cdot}^T - \boldsymbol{M}_{j\cdot}^T\|_2 \le \|\widehat{\boldsymbol{M}}_{j\cdot}^T\|_2 + \|\boldsymbol{M}_{j\cdot}^T\|_2 \tag{33}$$

by triangle inequality. As stated earlier, $\|[\boldsymbol{M}]_{j\cdot}^T\|_2 \le (\Gamma+\epsilon)\sqrt{p}$. Next, we bound $\|\widehat{\boldsymbol{M}}_{j\cdot}\|_2^T$. From (17), the fact that $\widehat{\rho} \ge 1/(qp)$, and Lemma 6, we have

$$\begin{aligned}\|\widehat{\boldsymbol{M}}_{j\cdot}^T\|_2 &= \frac{1}{\widehat{\rho}}\|\mathsf{HSVT}_{\lambda_k}(\boldsymbol{Y})_{j\cdot}^T\|_2 \\ &\le q\ p\|\phi_{\lambda_k}^{\boldsymbol{Y}}(\boldsymbol{Y}_{j\cdot}^T)\|_2 \\ &\le q\ p\|\phi_{\lambda_k}^{\boldsymbol{Y}}\|_2\|\boldsymbol{Y}_{j\cdot}^T\|_2 \\ &\le q\ p\|\boldsymbol{Y}_{j\cdot}^T\|_2,\end{aligned} \tag{34}$$

where we used the fact that $\phi_{\lambda_k}^{\boldsymbol{Y}}$ is a projection operator and hence $\|\phi_{\lambda_k}^{\boldsymbol{Y}}\|_2 = 1$. Note that $Y_{ij} = B_{ij}\times(M_{ij}+\boldsymbol{\varepsilon}_{ij})$, where B_{ij} is an independent Bernoulli variable with $\mathbb{P}(B_{ij}=1) = \rho$ representing

whether $(M_{ij}+\boldsymbol{\varepsilon}_{ij}$ is observed or not. Therefore, $|Y_{ij}| = |B_{ij}| \times |M_{ij}+\boldsymbol{\varepsilon}_{ij}| \le (\Gamma+\epsilon)+|\boldsymbol{\varepsilon}_{ij}|$. Therefore, from (33) and (34),

$$\begin{aligned}\max_{j\in[q]} \|\widehat{\boldsymbol{M}}^T_{j\cdot} - \boldsymbol{M}^T_{j\cdot}\|_2 &\le (\Gamma+\epsilon)\sqrt{p} + qp\big(\max_{j\in[q]} \|\boldsymbol{Y}^T_{j\cdot}\|_2\big) \\ &\le (\Gamma+\epsilon)\sqrt{p} + qp \times \sqrt{p}\big(\max_{i\in[p],j\in[q]} |Y_{ij}|\big) \\ &\le 2qp^{\frac{3}{2}}\big(\Gamma+\epsilon+\max_{i\in[p],j\in[q]} |\boldsymbol{\varepsilon}_{ij}|\big). \end{aligned} \tag{35}$$

Using $(a+b)^2 \le 2a^2+2b^2$ twice, we have $(a+b)^4 \le 8(a^4+b^4)$. Therefore, from (35)

$$\max_{j\in[q]} \|\widehat{\boldsymbol{M}}^T_{j\cdot} - \boldsymbol{M}^T_{j\cdot}\|_2^4 \le 16q^4p^6\big((\Gamma+\epsilon)^4 + \max_{i\in[p],j\in[q]} |\boldsymbol{\varepsilon}_{ij}|^4\big). \tag{36}$$

Recall $\mathbb{E}[\boldsymbol{\varepsilon}_{ij}] = 0$, $\|\boldsymbol{\varepsilon}_{ij}\|_{\psi_2} \le \sigma$ and $\boldsymbol{\varepsilon}_{ij}$ are independent across i, j. A property of ψ_2-random variables is that $\big|\eta_{ij}\big|^\theta$ is a $\psi_{2/\theta}$-random variable for $\theta \ge 1$. With choice of $\theta = 4$, we have

$$\mathbb{E}\big[\max_{ij} |\boldsymbol{\varepsilon}_{ij}|^4\big] \le C'\sigma^4 \log^2(qp), \tag{37}$$

for some $C' > 0$ by Lemma 2. From (34), (36), and (37), we have that

$$\Big(\mathbb{E}\big[\max_{j\in[q]} \frac{1}{p^2}\|\widehat{\boldsymbol{M}}^T_{j\cdot} - \boldsymbol{M}^T_{j\cdot}\|_2^4\big]\Big)^{\frac{1}{2}} \le 4q^2p^2\big((\Gamma+\epsilon)^4 + C'\sigma^4\log^2(qp)\big)^{\frac{1}{2}}. \tag{38}$$

Finally, using (31), (32) and (38), we conclude

$$\mathbb{E}\big[\max_{j\in[q]} \frac{1}{p}\|\widehat{\boldsymbol{M}}^T_{j\cdot} - \boldsymbol{M}^T_{j\cdot}\|_2^2\big] \le \frac{p(C\sigma^2+\rho^2\epsilon q)}{\rho^2\sigma_k(\boldsymbol{M}_k)^2}\Big(\Gamma^2+\frac{\sigma^2}{\rho^2}\Big) + \frac{C\sigma^2 k\log p}{p\rho^2} + \frac{C(\Gamma+\epsilon)^2}{p} + 2\epsilon^2 + \frac{C}{(pq)^2}.$$

This completes the proof of Theorem 10.

Appendix H: Proof of Theorem 4

The proof of Theorem 4 will utilize Theorem 10. To begin with, given N time series with observations over $[T]$, the mSSA algorithm as described in Section 1.1 constructs the $L \times (NT/L)$ stacked page matrix $\mathsf{SP}((X_1,\ldots,X_N),T,L)$ with $L = \sqrt{\min(N,T)T}$, i.e. $L \le T$.

As per the model described by (1) and Section 3, it follows that each entry of $\mathsf{SP}((X_1,\ldots,X_N),T,L)$ is an independent random variable; it is observed with probability $\rho \in (0,1]$ independently and when it is observed, its equal to value of the latent time series plus zero-mean sub-Gaussian noise. In particular,

$$\mathbb{E}\big[\mathsf{SP}((X_1,\ldots,X_N),T,L)\big] = \rho\mathsf{SP}((f_1,\ldots,f_N),T,L),$$

where $\mathsf{SP}((f_1,\ldots,f_N),T,L) \in \mathbb{R}^{L\times(NT/L)}$ with entry in row $\ell \in [L]$ and column $(n-1)\times T/L + j$ equal to $f_n(\ell + (j-1)\times L)$. Further, when entry in row $\ell \in [L]$ and column $(n-1)\times T/L + j$ in $\mathsf{SP}((X_1,\ldots,X_N),T,L)$ is observed, i.e. $X_n(\ell + (j-1)\times L) \neq \star$, it is equal to $f_n(\ell + (j-1)\times L) + \eta_n(\ell + (j-1)\times L)$ where $\eta_n(\cdot)$ are independent, zero-mean sub-Gaussian variables with $\|\eta_n(\cdot)\|_{\psi_2} \leq \gamma$ as per the Property 3.

Under Properties 1 and 8, as a direct implication of Proposition 14, $\mathsf{SP}((f_1,\ldots,f_N),T,L)$ has ϵ'-rank at most $R\times G$ with $\epsilon' = R\Gamma_1\epsilon$. That is, there exist rank $k \leq R\times G$ matrix $\boldsymbol{M}_k \in \mathbb{R}^{L\times(NT/L)}$ so that

$$\mathsf{SP}((f_1,\ldots,f_N),T,L) = \boldsymbol{M}_k + \boldsymbol{E}_k,$$

where $\|\boldsymbol{E}_k\|_\infty \leq \epsilon'$. Due to Property 1, it follows that $\|\boldsymbol{M}_k\|_\infty \leq R\Gamma_1\Gamma_2 + \epsilon'$. Under Property 9, we have $\sigma_k(\boldsymbol{M}_k) \geq c\sqrt{NT}/\sqrt{k}$ for some constant $c > 0$.

Define

$$\Gamma = R\Gamma_1\Gamma_2 + \epsilon' \;=\; R\Gamma_1(\Gamma_2 + \epsilon).$$

Recall from Section 1.1, the elements of the imputed multivariate time series are simply the entries of the matrix $\widehat{\mathsf{SP}}((X_1,\ldots,X_N),T,L)$ where $\widehat{\mathsf{SP}}((X_1,\ldots,X_N),T,L) = \frac{1}{\hat{\rho}}\mathsf{HSVT}_k(\mathsf{SP}((X_1,\ldots,X_N),T,L))$. That is, imputation in mSSA is carried out by applying HSVT to the stacked page matrix $\mathsf{SP}((X_1,\ldots,X_N),T,L)$.

All in all, the above description precisely meets the setup of Theorem 10. To apply Theorem 10, we require $\rho \geq C\log(NT)/\sqrt{NT}$ for $C > 0$ large enough. Note that the number of columns in $\widehat{\mathsf{SP}}((X_1,\ldots,X_N),T,L)$ is equal to NT/L for $L = \sqrt{\min(N,T)T}$ – for this choice of L, note that $NT/L \geq L$. Using $\sigma_k^2(\boldsymbol{M}_k) \geq cNT/k$, for some absolute constant $= c \geq 0$, and using Theorem 10, we obtain

$$\begin{aligned}
&\mathbb{E}\Big[\frac{1}{(NT/L)}\|\widehat{\mathsf{SP}}((X_1,\ldots,X_N),T,L) - \mathsf{SP}((f_1,\ldots,f_N),T,L)\|_{2,\infty}^2\Big] \qquad (39)\\
&\leq \frac{k(NT/L)(C\gamma^2+\rho^2\epsilon' L)}{\rho^2 c^2 NT}\Big(\Gamma^2 + \frac{\gamma^2}{\rho^2}\Big) + \frac{C\gamma^2 k\log NT}{(NT/L)\rho^2} + \frac{C(\Gamma+\epsilon')^2}{(NT/L)} + 2(\epsilon')^2 + \frac{C}{(NT)^2}
\end{aligned}$$

Recall that $k \le R \times G$, $\epsilon' = R\Gamma_1\epsilon$, and $\Gamma = R\Gamma_1(\Gamma_2+\epsilon)$. Hence, simplifying (39), we obtain that

$$\begin{aligned}
&\mathbb{E}\Big[\frac{1}{(NT/L)}\|\widehat{\mathsf{SP}}((X_1,\ldots,X_N),T,L) - \mathsf{SP}((f_1,\ldots,f_N),T,L)\|_{2,\infty}^2\Big] \\
&\le \tilde{C}\left(\frac{RG(1+\rho^2 R\epsilon L)}{\rho^2 L}\left(R^2(1+\epsilon^2)+\frac{1}{\rho^2}\right) + \frac{RG\log NT}{(NT/L)\rho^2} + \frac{(R(1+\epsilon))^2}{(NT/L)} + (R\epsilon)^2\right) \\
&\le \tilde{C}\left(\frac{R^3 G\log NT}{\rho^4 L} + \frac{R^4 G(\epsilon+\epsilon^2+\epsilon^3)}{\rho^2}\right),
\end{aligned} \tag{40}$$

where $\tilde{C} = C(c,\Gamma_1,\Gamma_2,\gamma)$ is a positive constant dependent on model parameters including Γ_1,Γ_2,γ.

It can be easily verified that for any matrix, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$,

$$\frac{1}{mn}\|\boldsymbol{A}\|_F^2 \le \frac{1}{n}\|\boldsymbol{A}\|_{\infty,2}^2. \tag{41}$$

Further, there is a one-to-one mapping of $\hat{f}_n(\cdot)$ (resp. $f_n(\cdot)$) to the entries of $\widehat{\mathsf{SP}}((X_1,\ldots,X_N),T,L)$ (resp. $\mathsf{SP}((f_1,\ldots,f_N),T,L)$). Hence,

$$\mathsf{ImpErr}(N,T) = \mathbb{E}\Big[\frac{1}{NT}\|\widehat{\mathsf{SP}}((X_1,\ldots,X_N),T,L) - \mathsf{SP}((f_1,\ldots,f_N),T,L)\|_F^2\Big] \tag{42}$$

Therefore, from (40), (41), and (42) it follows that

$$\mathsf{ImpErr}(N,T) \le C(c,\Gamma_1,\Gamma_2,\gamma)\left(\frac{R^3 G\log NT}{\rho^4 L} + \frac{R^4 G(\epsilon+\epsilon^2+\epsilon^3)}{\rho^2}\right)$$

This completes the proof of Theorem 4.

Appendix I: Proof of Theorem 5

The forecasting algorithm, as described in Section 1.1, computes a linear model between the recent past and immediate future to forecast. We shall bound the forecasting error, $\mathsf{ForErr}(N,T,L)$ as defined in (8). We start with some setup and notations, followed by a key proposition that establishes the existence of a linear model under the setup of Theorem 5, and then conclude with a detailed analysis of noisy, mis-specified least-squares.

Setup, Notations. For $L \ge 1, k \ge 1$, for ease of notations, we define

- $\mathsf{SP}(X) = \mathsf{SP}((X_1,\ldots,X_N),T,L) \in \mathbb{R}^{L\times(NT/L)}$,
- $\mathsf{SP}(f) = \mathsf{SP}((f_1,\ldots,f_N),T,L) \in \mathbb{R}^{L\times(NT/L)}$,
- $\mathsf{SP}'(X) \in \mathbb{R}^{(L-1)\times(NT/L)}$ as the top $L-1$ rows of $\mathsf{SP}((X_1,\ldots,X_N),T,L)$,

- $\mathsf{SP}'(f) \in \mathbb{R}^{(L-1)\times(NT/L)}$ as the top $L-1$ rows of $\mathsf{SP}((f_1,\ldots,f_N),T,L)$.

It is worth noting that $\mathbb{E}[\mathsf{SP}(X)] = \rho\mathsf{SP}(f)$ and hence

$$\mathsf{SP}_{L\cdot}(X)^T = \rho\mathsf{SP}_{L\cdot}(f)^T + \eta, \tag{43}$$

where $\eta \in \mathbb{R}^{(NT)/L}$ is a random vector with each component being independent, zero-mean with its distribution given as: it is 0 with probability $1-\rho$ and with probability ρ, due to Property 3, it equals a zero-mean sub-Gaussian random variable with $\|\cdot\|_{\psi_2} \le \gamma$. Therefore, using arguments in Agarwal et al. (2019, 2021), each component of η is an independent, zero-mean random variable with $\|\cdot\|_{\psi_2}$ bounded above by $C'(\gamma^2 + R\Gamma_1\Gamma_2)$ for some absolute constant $C' > 0$. Let $K = C'(\gamma^2 + R\Gamma_1\Gamma_2)$ and hence each component of η has $\|\cdot\|_{\psi_2}$ bounded by K.

Now, recall that for forecasting, we first apply the imputation algorithm (i.e. HSVT) to $\mathsf{SP}((X_1,\ldots,X_N),T,L)$ by replacing $\star$s, i.e. missing observations by 0 as well as setting all the entries in the last row equal to 0. Equivalently, the imputation algorithm is applied to $\mathsf{SP}'(X)$ after setting all missing values to 0. Let $\widehat{\mathsf{SP}'} \in \mathbb{R}^{L-1\times(NT/L)}$ be the estimate produced from the imputation algorithm applied to $\mathsf{SP}'(X)$. Under the setup of Theorem 4, by following arguments identical to that of Theorems 10 and 4–in particular, refer to (40)–it follows that by selecting the right choice of $k \le R\times G$, we have

$$\mathbb{E}\Big[\frac{1}{(NT/L)}\|\widehat{\mathsf{SP}'} - \mathsf{SP}'(f)\|_{2,\infty}^2\Big] \le \tilde{C}\left(\frac{R^3 G\log NT}{\rho^4 L} + \frac{R^4 G(\epsilon+\epsilon^2+\epsilon^3)}{\rho^2}\right), \tag{44}$$

where $\tilde{C} = C(c,\Gamma_1,\Gamma_2,\gamma) > 0$ is a constant dependent on $c,\Gamma_1,\Gamma_2,\gamma$.

Now, the mSSA forecasting algorithm finds $\widehat{\beta} = \widehat{\beta}((X_1,\ldots,X_N),TL;k)$, by solving the following Ordinary Least Squares (OLS):

$$\widehat{\beta} \in \mathsf{minimize} \quad \|\frac{1}{\widehat{\rho}}\mathsf{SP}(X)_{L\cdot} - \widehat{\mathsf{SP}'}^T\beta\|_2^2 \quad \mathsf{over} \quad \beta \in \mathbb{R}^{L-1}. \tag{45}$$

And subsequently, $\widehat{\mathsf{SP}'}^T\widehat{\beta}$ is used as the estimate for $\mathsf{SP}(f)_{L\cdot} \in \mathbb{R}^{NT/L}$, the Lth row of the latent $\mathsf{SP}(f)$. The goal is to bound the forecasting error $\mathsf{ForErr}(N,T,L)$, which is given by

$$\mathsf{ForErr}(N,T,L) = \mathbb{E}\Big[\frac{1}{(NT/L)}\|\mathsf{SP}(f)_{L\cdot} - \widehat{\mathsf{SP}'}^T\widehat{\beta}\|_2^2\Big].$$

Therefore, our interest is in bounding $\mathbb{E}\big[\|\mathsf{SP}_{L\cdot}(f) - \widehat{\mathsf{SP}'}^T\widehat{\beta}\|_2^2\big]$.

Now, we recall from Proposition 12 that there exists $\beta^* \in \mathbb{R}^{L-1}$, such that

$$\|\mathsf{SP}(f)_{L\cdot}^T - \mathsf{SP}'(f)^T\beta^*\|_\infty \le C_2\epsilon,$$

where $C_2 \coloneqq R\Gamma_1(1+\|\beta^*\|_1)$.

Bounding $\mathbb{E}\big[\|\mathsf{SP}_{L\cdot}(f) - \widehat{\mathsf{SP}'}^T\widehat{\beta}\|_2^2\big]$. By (45) and (43)

$$\begin{aligned}\|\frac{1}{\widehat{\rho}}\mathsf{SP}(X)_{L\cdot} - \widehat{\mathsf{SP}'}^T\widehat{\beta}\|_2^2 &\le \|\frac{1}{\widehat{\rho}}\mathsf{SP}(X)_{L\cdot} - \widehat{\mathsf{SP}'}^T\beta^*\|_2^2\\ &= \|\frac{\rho}{\widehat{\rho}}\mathsf{SP}(f)_{L\cdot} + \eta - \widehat{\mathsf{SP}'}^T\beta^*\|_2^2\\ &= \|\frac{\rho}{\widehat{\rho}}\mathsf{SP}(f)_{L\cdot} - \widehat{\mathsf{SP}'}^T\beta^*\|_2^2 + \|\eta\|_2^2 + 2\eta^T(\frac{\rho}{\widehat{\rho}}\mathsf{SP}(f)_{L\cdot} - \widehat{\mathsf{SP}'}^T\beta^*). \quad (46)\end{aligned}$$

Also,

$$\begin{aligned}\|\frac{1}{\widehat{\rho}}\mathsf{SP}(X)_{L\cdot} - \widehat{\mathsf{SP}'}^T\widehat{\beta}\|_2^2 &= \|\frac{\rho}{\widehat{\rho}}\mathsf{SP}(f)_{L\cdot} + \eta - \widehat{\mathsf{SP}'}^T\widehat{\beta}\|_2^2\\ &= \|\frac{\rho}{\widehat{\rho}}\mathsf{SP}(f)_{L\cdot} - \widehat{\mathsf{SP}'}^T\widehat{\beta}\|_2^2 + \|\eta\|_2^2 + 2\eta^T(\frac{\rho}{\widehat{\rho}}\mathsf{SP}(f)_{L\cdot} - \widehat{\mathsf{SP}'}^T\widehat{\beta}). \quad (47)\end{aligned}$$

From (46) and (47)

$$\begin{aligned}&\mathbb{E}\big[\|\frac{\rho}{\widehat{\rho}}\mathsf{SP}(f)_{L\cdot} - \widehat{\mathsf{SP}'}^T\widehat{\beta}\|_2^2\big] \quad (48)\\ &\le \mathbb{E}\big[\|\frac{\rho}{\widehat{\rho}}\mathsf{SP}(f)_{L\cdot} - \widehat{\mathsf{SP}'}^T\beta^*\|_2^2\big] + 2\mathbb{E}\big[\eta^T\widehat{\mathsf{SP}'}^T(\beta^* - \widehat{\beta})\big]\end{aligned}$$

η is independent of $\widehat{\mathsf{SP}'}$, β^*, and $\widehat{\rho}$; $\mathbb{E}[\eta] = \mathbf{0}$; thus, we have that

$$\mathbb{E}\big[\eta^T\widehat{\mathsf{SP}'}^T\beta^*\big] = 0. \quad (49)$$

By (45), we have $\widehat{\beta} = \widehat{\mathsf{SP}'}^{T,\dagger}\frac{1}{\widehat{\rho}}\mathsf{SP}(X)_{L\cdot}$, where $\widehat{\mathsf{SP}'}^{T,\dagger}$ is pseudo-inverse of $\widehat{\mathsf{SP}'}^T$. That is,

$$\widehat{\beta} = \widehat{\mathsf{SP}'}^{T,\dagger}\frac{\rho}{\widehat{\rho}}\mathsf{SP}(f)_{L\cdot} + \frac{1}{\widehat{\rho}}\widehat{\mathsf{SP}'}^{T,\dagger}\eta. \quad (50)$$

Using cyclic and linearity of Trace operator; the independence properties of η; and (50); we have

$$\begin{aligned}\mathbb{E}[\eta^T\widehat{\mathsf{SP}'}^T\widehat{\beta}] &= \mathbb{E}[\eta^T\widehat{\mathsf{SP}'}^T\widehat{\mathsf{SP}'}^{T,\dagger}\frac{\rho}{\widehat{\rho}}\mathsf{SP}(f)_{L\cdot}] + \mathbb{E}[\frac{1}{\widehat{\rho}}\eta^T\widehat{\mathsf{SP}'}^T\widehat{\mathsf{SP}'}^{T,\dagger}\eta]\\ &= \mathbb{E}[\eta]^T\mathbb{E}[\widehat{\mathsf{SP}'}^T\widehat{\mathsf{SP}'}^{T,\dagger}\frac{\rho}{\widehat{\rho}}]\mathsf{SP}(f)_{L\cdot} + \mathbb{E}[\frac{1}{\widehat{\rho}}\mathsf{Tr}(\eta^T\widehat{\mathsf{SP}'}^T\widehat{\mathsf{SP}'}^{T,\dagger}\eta)]\\ &= \mathbb{E}[\frac{1}{\widehat{\rho}}\mathsf{Tr}(\widehat{\mathsf{SP}'}^T\widehat{\mathsf{SP}'}^{T,\dagger}\eta\eta^T)]\\ &= \mathsf{Tr}\big(\mathbb{E}[\frac{1}{\widehat{\rho}}\widehat{\mathsf{SP}'}^T\widehat{\mathsf{SP}'}^{T,\dagger}]\mathbb{E}[\eta\eta^T]\big)\\ &\le C(\gamma)k/\rho, \quad (51)\end{aligned}$$

where $C(\gamma)$ is a function only of γ. To see the last inequality, we use various facts. First, by the definition of the HSVT algorithm $\widehat{\mathsf{SP}'}^T$ has rank at most k. Second, let $\widehat{\mathsf{SP}'}^T = \boldsymbol{USV}^T$ be the singular value decomposition of $\widehat{\mathsf{SP}'}^T$, we have

$$\begin{aligned}\widehat{\mathsf{SP}'}^T \widehat{\mathsf{SP}'}^{T,\dagger} &= \boldsymbol{USV}^T \boldsymbol{VS}^\dagger \boldsymbol{U}^T \\ &= \boldsymbol{U\tilde{I}U}^T,\end{aligned}$$

That is, $\frac{1}{\widehat{\rho}}\widehat{\mathsf{SP}'}^T \widehat{\mathsf{SP}'}^{T,\dagger}$ is a positive semi-definite matrix and $\mathsf{Tr}(\frac{1}{\widehat{\rho}}\widehat{\mathsf{SP}'}^T \widehat{\mathsf{SP}'}^{T,\dagger}) \le k/\widehat{\rho}$. The matrix $\mathbb{E}[\eta\eta^T]$ is diagonal with all the non-zero entries on diagonal (variance of components of η) bounded above by a constant that depends on γ. For a positive semi-definite matrix A and positive semi-definite diagonal matrix B, $\mathsf{Tr}(AB) \le \|B\|_2 \mathsf{Tr}(A)$. For $\rho \ge C\log(NT)/\sqrt{NT}$ for large enough C, one can verfiy that $\mathbb{E}[1/\widehat{\rho}] \le 2/\rho$. This completes the justification of the last step of (51).

Now consider the term $\|\frac{\rho}{\widehat{\rho}}\mathsf{SP}(f)_{L\cdot} - \widehat{\mathsf{SP}'}^T\beta^*\|_2^2$. Note,

$$\begin{aligned}\|\frac{\rho}{\widehat{\rho}}\mathsf{SP}(f)_{L\cdot} - \widehat{\mathsf{SP}'}^T\beta^*\|_2^2 &= \|\big(\mathsf{SP}(f)_{L\cdot} - \widehat{\mathsf{SP}'}^T\beta^*\big) + \big(\frac{\rho-\widehat{\rho}}{\widehat{\rho}}\big)\mathsf{SP}(f)_{L\cdot}\|_2^2 \\ &\le 2\|\big(\mathsf{SP}(f)_{L\cdot} - \widehat{\mathsf{SP}'}^T\beta^*\big)\|_2^2 + 2\|\frac{\rho-\widehat{\rho}}{\widehat{\rho}}\mathsf{SP}(f)_{L\cdot}\|_2^2. \qquad (52)\end{aligned}$$

We will bound the two terms on the r.h.s of (52) separately. We now consider the first term.

$$\|\mathsf{SP}(f)_{L\cdot} - \widehat{\mathsf{SP}'}^T\beta^*\|_2^2 \le 2\|\mathsf{SP}(f)_{L\cdot} - \mathsf{SP}'(f)^T\beta^*\|_2^2 + 2\|\mathsf{SP}'(f)^T\beta^* - \widehat{\mathsf{SP}'}^T\beta^*\|_2^2. \qquad (53)$$

By Proposition 12

$$\|\mathsf{SP}(f)_{L\cdot} - \mathsf{SP}'(f)^T\beta^*\|_2 \le \|\mathsf{SP}(f)_{L\cdot} - \mathsf{SP}'(f)^T\beta^*\|_\infty\sqrt{NT/L} \;\le\; C_2\epsilon\sqrt{NT/L}, \qquad (54)$$

where we used the fact that for any $v \in \mathbb{R}^p$, $\|v\|_2 \le \|v\|_\infty\sqrt{p}$. And,

$$\|\mathsf{SP}'(f)^T\beta^* - \widehat{\mathsf{SP}'}^T\beta^*\|_2 = \|(\mathsf{SP}'(f) - \widehat{\mathsf{SP}'})^T\beta^*\|_2 \;\le\; \|\mathsf{SP}'(f) - \widehat{\mathsf{SP}'}\|_{2,\infty}\|\beta^*\|_1, \qquad (55)$$

where we used the fact that for any $A \in \mathbb{R}^{q\times p}, v \in \mathbb{R}^p$, $\|Av\|_2 \le \|A^T\|_{2,\infty}\|v\|_1$. Finally, note that

$$\|\mathsf{SP}(f)_{L\cdot} - \widehat{\mathsf{SP}'}^T\widehat{\beta}\|_2^2 \le 2\|\frac{\rho}{\widehat{\rho}}\mathsf{SP}(f)_{L\cdot} - \widehat{\mathsf{SP}'}^T\widehat{\beta}\|_2^2 + 2\|\frac{\rho-\widehat{\rho}}{\widehat{\rho}}\mathsf{SP}(f)_{L\cdot}\|_2^2. \qquad (56)$$

Using (48), (49), (51), (52),(53), (54), (55), and the bound in (56), we obtain

$$\begin{aligned}&\mathbb{E}\big[\|\mathsf{SP}(f)_{L\cdot} - \widehat{\mathsf{SP}'}^T\widehat{\beta}\|_2^2\big] \qquad (57)\\ &\le 4C(\gamma)k/\rho + 6\mathbb{E}\big[\|\frac{\rho-\widehat{\rho}}{\widehat{\rho}}\mathsf{SP}(f)_{L\cdot}\|_2^2\big] + 2C_2\epsilon^2(NT/L) + 2\|\beta^*\|_1^2\|\mathsf{SP}'(f) - \widehat{\mathsf{SP}'}\|_{2,\infty}^2.\end{aligned}$$

Note that $\|\mathsf{SP}(f)\|_\infty \le R\Gamma_1\Gamma_2$. Hence, $\|\mathsf{SP}(f)_{L\cdot}\|_2^2 \le C(\Gamma_1,\Gamma_2)R^2(NT/L)$, for large enough constant $C(\Gamma_1,\Gamma_2)$ that may depend on Γ_1,Γ_2. Using the bounds derived in Lemma 8, one can verify that $\mathbb{E}[(\frac{\rho-\widehat{\rho}}{\widehat{\rho}})^2] \le C/(NT/L)$ for large enough positive constant C. Therefore, we have that

$$6\mathbb{E}\big[\|\frac{\rho-\widehat{\rho}}{\widehat{\rho}}\mathsf{SP}(f)_{L\cdot}\|_2^2\big] \le C(\Gamma_1,\Gamma_2)R^2 \tag{58}$$

Using (44), (58), and the bound in (57); diving by $1/(NT/L)$ on both sides; and noting $k \le R\times G$, we obtain

$$\begin{aligned}
&\mathbb{E}\big[\frac{1}{(NT/L)}\|\mathsf{SP}(f)_{L\cdot} - \widehat{\mathsf{SP}'}^T\widehat{\beta}\|_2^2\big] \\
&\le C(c,\gamma,\Gamma_1,\Gamma_2)\left(\frac{RG}{\rho(NT/L)} + \frac{R^2}{(NT/L)} + R(1+\|\beta^*\|_1)\epsilon^2 + \|\beta^*\|_1^2\left(\frac{R^3G\log NT}{\rho^4 L} + \frac{R^4G(\epsilon+\epsilon^2+\epsilon^3)}{\rho^2}\right)\right) \\
&\le C(c,\gamma,\Gamma_1,\Gamma_2)\left(\max(1,\|\beta^*\|_1,\|\beta^*\|_1^2)\left(\frac{R^3G\log NT}{\rho^4 L} + \frac{R^4G(\epsilon+\epsilon^2+\epsilon^3)}{\rho^2}\right)\right)
\end{aligned} \tag{59}$$

Letting $L = \sqrt{\min(N,T)T}$, using (59), and noting that

$$\mathsf{ForErr}(N,T,L) = \mathbb{E}\big[\frac{1}{(NT/L)}\|\mathsf{SP}(f)_{L\cdot} - \widehat{\mathsf{SP}'}^T\widehat{\beta}\|_2^2\big]$$

completes the proof of Theorem 5.

I.1. Proof of Proposition 12

For this proof, we utilize a modified version of the stacked Hankel matrix defined in Appendix D. Define the modified Hankel matrix for time series f_n, for $n \in [N]$, as $\widetilde{\mathsf{H}}(n) \in \mathbb{R}^{T\times 2T}$, where for $i \in [T], j \in [2T]$, we have

$$\widetilde{\mathsf{H}}(n)_{ij} = f_n(i+j-1-T).$$

Define $\widetilde{\mathsf{SH}} \in \mathbb{R}^{T\times NT}$ as the column wise concatenation of the matrices $\widetilde{\mathsf{H}}(n)$ for $n \in [N]$, i.e., $\widetilde{\mathsf{SH}} := [\widetilde{\mathsf{H}}(1),\ldots,\widetilde{\mathsf{H}}(N)]$. By a straightforward modification of the proof of Proposition 14, we have $\widetilde{\mathsf{SH}}$ has ϵ'-rank bounded by $R\times G$ with $\epsilon' = R\Gamma_1\epsilon$. That is, there exists a matrix $\mathsf{M} \in \mathbb{R}^{T\times NT}$ such that,

$$\mathrm{rank}(\mathsf{M}) \le RG, \quad \|\widetilde{\mathsf{SH}} - \mathsf{M}\|_\infty \le \epsilon'$$

Since $\mathrm{rank}(\mathsf{M}) \le RG$, it must be the case that within the last RG rows of M, there exists at least one row, which we denote as r^*, that can be written as a linear combination of at most RG rows above

it, which we denote as $r_1, \ldots, r_{RG}$. Specifically there exists a vector $\theta := (\theta_1, \ldots, \theta_{RG}) \in \mathbb{R}^{RG}$ such that

$$\mathsf{M}_{r^*,\cdot} = \sum_{\ell=1}^{RG} \theta_\ell \mathsf{M}_{r_\ell,\cdot}$$

Hence for $j \in [2T]$,

$$\begin{aligned}
&\left|\widetilde{\mathsf{SH}}_{r^*,j} - \sum_{\ell=1}^{RG} \theta_\ell \widetilde{\mathsf{SH}}_{r_\ell,j}\right| \\
&= \left|\widetilde{\mathsf{SH}}_{r^*,j} \pm \mathsf{M}_{r^*,j} - \sum_{\ell=1}^{RG} \theta_\ell \widetilde{\mathsf{SH}}_{r_\ell,j} \pm \sum_{\ell=1}^{RG} \theta_\ell \mathsf{M}_{r_\ell,t}\right| \\
&\le \left|\widetilde{\mathsf{SH}}_{r^*,j} - \mathsf{M}_{r^*,j}\right| + \left|\sum_{\ell=1}^{RG} \theta_\ell \widetilde{\mathsf{SH}}_{r_\ell,j} - \sum_{\ell=1}^{RG} \theta_\ell \mathsf{M}_{r_\ell,t}\right| + \left|\mathsf{M}_{r^*,j} - \sum_{\ell=1}^{RG} \theta_\ell \mathsf{M}_{r_\ell,t}\right| \\
&= \left|\widetilde{\mathsf{SH}}_{r^*,j} - \mathsf{M}_{r^*,j}\right| + \left|\sum_{\ell=1}^{RG} \theta_\ell (\widetilde{\mathsf{SH}}_{r_\ell,j} - \mathsf{M}_{r_\ell,t})\right| \\
&\le \epsilon' + \|\theta\|_1 \|\widetilde{\mathsf{SH}}_{r_\ell,j} - \mathsf{M}_{r_\ell,t}\|_\infty \\
&\le R\Gamma_1(1 + \|\theta\|_1)\epsilon. \qquad (60)
\end{aligned}$$

Observe that every entry of $\mathsf{SP}(f)_{L\cdot}$ appears within $\widetilde{\mathsf{SH}}_{r^*,\cdot}$; this can be seen by noting that $\widetilde{\mathsf{SH}}$ is skew-symmetric and thus every entry in the last row of $\widetilde{\mathsf{SH}}$ appears along the appropriate diagonal. Using this skew-symmetric property of $\widetilde{\mathsf{SH}}$ and (60), it implies that by appropriately selecting entries in $\widetilde{\mathsf{SH}}$, there exists $\beta^* \in \mathbb{R}^{L-1}$,

$$\|\mathsf{SP}(f)^T_{L\cdot} - \mathsf{SP}'(f)^T \beta^*\|_\infty \le R\Gamma_1(1 + \|\beta\|_1)\epsilon,$$

where the non-zero entries in β^* correspond to the entries of θ. Noting that $\theta \in \mathbb{R}^{RG}$ implies $\|\beta^*\|_0 \le RG$. This completes the proof.

Appendix J: Proof of Theorem 3

Notation. For integers $t_1 < t_2$ where $t_2 - t_1 + 1 \ge L$, let $\mathsf{SP}((X_1, \ldots, X_N), t_1 : t_2, L)$ represents the stacked page matrix constructed using the contiguous observations $X_n(t_1), \ldots, X_n(t_2),\ \forall n \in [N]$. Throughout, we use the following notations:

- $\mathsf{SP}_0(X) = \mathsf{SP}((X_1, \ldots, X_N), 1 : T, L) \in \mathbb{R}^{L \times (NT/L)}$, with zeros replacing missing values.

- $\mathsf{SP}_1(X) = \mathsf{SP}((X_1, \ldots, X_N), T+1 : T+T_1, L) \in \mathbb{R}^{L \times (NT_1/L)}$, with zeros replacing missing values.
- $\mathsf{SP}_0(f) = \mathsf{SP}((f_1, \ldots, f_N), 1 : T, L) \in \mathbb{R}^{L \times (NT/L)}$.
- $\mathsf{SP}_1(f) = \mathsf{SP}((f_1, \ldots, f_N), T+1 : T+T_1, L) \in \mathbb{R}^{L \times (NT_1/L)}$.
- $\mathsf{SP}_1(\eta) = \mathsf{SP}((\eta_1, \ldots, \eta_N), T+1 : T+T_1, L) \in \mathbb{R}^{L \times (NT_1/L)}$.
- $\mathsf{SP}'_0(X) \in \mathbb{R}^{(L-1) \times (NT/L)}$ as the top $L-1$ rows of $\mathsf{SP}_0(X)$. Let $\mathsf{SP}'_1(X), \mathsf{SP}'_0(f)$, $\mathsf{SP}'_1(f)$ and $\mathsf{SP}'_1(\eta)$ be defined analogously.
- $\widehat{\rho} \coloneqq (\max(1, \sum_{i=1}^{L-1} \sum_{j=1}^{NT/L} \mathbf{1}(\mathsf{SP}_0(X)_{ij} \neq \star)))/(NT - NT/L)$

Recall that we are interested in bounding the following out-of-sample prediction error:

$$\mathsf{TestForErr}(N, T, T_1, L) = \frac{L}{NT_1} \sum_{n=1}^{N} \sum_{m'=1}^{T_1/L} \mathbb{E}\Big[(f_n(T + L \times m') - \bar{f}_n(T + L \times m'))^2\Big].$$

Where the forecasted estimate $\bar{f}_n(\cdot)$, $n \in [N]$ are produced by the algorithm detailed in Section 1.1.

Based on the algorithm, we can write $\mathsf{TestForErr}(N, T, T_1, L)$ as follows:

$$\begin{aligned}\mathsf{TestForErr}(N, T, T_1, L) &= \frac{1}{(NT_1/L)} \mathbb{E}\Big[\| \frac{1}{\widehat{\rho}} \mathsf{SP}'_1(X)^T \widehat{\beta} - \mathsf{SP}_1(f)^T_{L\cdot} \|_2^2\Big] \\ &= \frac{1}{(NT_1/L)} \mathbb{E}\Big[\| \frac{1}{\widehat{\rho}} \mathsf{SP}'_1(X)^T \widehat{\beta} - \mathsf{SP}'_1(f)^T \beta^* \|_2^2\Big].\end{aligned}$$

Before bounding this term, we introduce the following important notation. For $i \in \{0, 1\}$, let $\boldsymbol{U}_i \boldsymbol{\Sigma}_i \boldsymbol{V}_i^T$ denote the Singular Value Decomposition (SVD) of $\mathsf{SP}'_i(f)$. Also, let $\widetilde{\boldsymbol{U}}_i \widetilde{\boldsymbol{\Sigma}}_i \widetilde{\boldsymbol{V}}_i^T$ denote the top k singular components of the SVD of $\mathsf{SP}'_i(X)$, while $\widetilde{\boldsymbol{U}}_i^{\perp} \widetilde{\boldsymbol{\Sigma}}_i^{\perp} (\widetilde{\boldsymbol{V}}_i^{\perp})^T$ denote the remaining $L-k-1$ components such that $\mathsf{SP}'_i(X) = \widetilde{\boldsymbol{U}}_i \widetilde{\boldsymbol{\Sigma}}_i \widetilde{\boldsymbol{V}}_i^T + \widetilde{\boldsymbol{U}}_i^{\perp} \widetilde{\boldsymbol{\Sigma}}_i^{\perp} (\widetilde{\boldsymbol{V}}_i^{\perp})^T$. Finally, let $\boldsymbol{V}_i^{\perp}$ and $\boldsymbol{U}_i^{\perp}$ be matrices of orthornormal basis vectors that span the null space of $\mathsf{SP}'_i(f)$ and $\mathsf{SP}'_i(f)^T$, respectively. Further, let $\widehat{\mathsf{SP}'_i}$ be the HSVT estimate of $\mathsf{SP}'_i(f)$. That is $\widehat{\mathsf{SP}'_i} = \frac{1}{\widehat{\rho}} \widetilde{\boldsymbol{U}}_i \widetilde{\boldsymbol{\Sigma}}_i \widetilde{\boldsymbol{V}}_i^T$. Also, let $\widehat{\mathsf{SP}'_i}^{\perp} = \frac{1}{\widehat{\rho}} \widetilde{\boldsymbol{U}}_i^{\perp} \widetilde{\boldsymbol{\Sigma}}_i^{\perp} (\widetilde{\boldsymbol{V}}_i^{\perp})^T$.

We start the proof by providing a deterministic upper bound for out-of-sample error.

Deterministic Bound. Due to triangle inequality, we have

$$\begin{aligned}\| \frac{1}{\widehat{\rho}} \mathsf{SP}'_1(X)^T \widehat{\beta} - \mathsf{SP}'_1(f)^T \beta^* \|_2^2 &= \| \frac{1}{\widehat{\rho}} \mathsf{SP}'_1(X)^T \widehat{\beta} - \mathsf{SP}'_1(f)^T \beta^* + \widehat{\mathsf{SP}'_1}^T \widehat{\beta} - \widehat{\mathsf{SP}'_1}^T \widehat{\beta} \|_2^2 \\ &\leq 2 \| \frac{1}{\widehat{\rho}} \mathsf{SP}'_1(X)^T \widehat{\beta} - \widehat{\mathsf{SP}'_1}^T \widehat{\beta} \|_2^2 + 2 \| \widehat{\mathsf{SP}'_1}^T \widehat{\beta} - \mathsf{SP}'_1(f)^T \beta^* \|_2^2.\end{aligned}$$

Next, we proceed to bound each of the two terms on the right hand side.

First term: $\|\frac{1}{\widehat{\rho}}\mathsf{SP}_1'(X)^T\widehat{\beta} - \widehat{\mathsf{SP}_1'}^T\widehat{\beta}\|_2^2$.

$$\begin{aligned}\|\frac{1}{\widehat{\rho}}\mathsf{SP}_1'(X)^T\widehat{\beta} - \widehat{\mathsf{SP}_1'}^T\widehat{\beta}\|_2^2 &= \|(\widehat{\mathsf{SP}_1'}^{\perp})^T\widehat{\beta}\|_2^2 \\ &= \|\frac{1}{\widehat{\rho}}\widetilde{\boldsymbol{V}}_1^{\perp}\widetilde{\boldsymbol{\Sigma}}_1^{\perp}(\widetilde{\boldsymbol{U}}_1^{\perp})^T\widehat{\beta}\|_2^2 \\ &\leq \|\frac{1}{\widehat{\rho}}\widetilde{\boldsymbol{\Sigma}}_1^{\perp}\|_2^2\|(\widetilde{\boldsymbol{U}}_1^{\perp})^T\widehat{\beta}\|_2^2.\end{aligned} \tag{61}$$

Note that $\|\widetilde{\boldsymbol{\Sigma}}_1^{\perp}\|_2$ equals the $(k+1)$-th singular value of $\mathsf{SP}_1'(X)$. Recall that $\mathbb{E}[\mathsf{SP}_1'(X)] = \rho\mathsf{SP}_1'(f)$ and hence

$$\mathsf{SP}_1'(X) = \rho\mathsf{SP}_1'(f) + \zeta_1, \tag{62}$$

where $\zeta_1 \in \mathbb{R}^{(L-1)\times(NT_1)/L}$ is a random matrix with zero-mean i.i.d. entries where each entry is 0 with probability $1-\rho$ and equals a zero-mean sub-Gaussian random variable with $\|\cdot\|_{\psi_2} \leq \gamma$ with probability ρ (due to Property 3). Next, we show that each component of ζ_1 is an independent, zero-mean random variable with $\|\cdot\|_{\psi_2}$ bounded above by $C'(\gamma + R\Gamma_1\Gamma_2)$ for some absolute constant $C' > 0$. Let ζ_{ij} for $i \in [L-1]$ and $j \in [NT/L]$ denotes the ij-th entry in ζ_1. Further, let $P_{ij} \in \{0,1\}$ denotes the random mask which takes the value 1 with probability ρ such that $\mathsf{SP}_1'(X)_{ij} = P_{ij}(\mathsf{SP}_1'(f)_{ij} + \mathsf{SP}_1'(\eta)_{ij})$. Then, we have

$$\begin{aligned}\|\zeta_{ij}\|_{\psi_2} &= \|\mathsf{SP}_1'(X)_{ij} - \rho\mathsf{SP}_1'(f)_{ij}\|_{\psi_2} \\ &= \|P_{ij}\mathsf{SP}_1'(f)_{ij} + P_{ij}\mathsf{SP}_1'(\eta)_{ij} - \rho\mathsf{SP}_1'(f)_{ij}\|_{\psi_2} \\ &\leq \|P_{ij}\mathsf{SP}_1'(\eta)_{ij}\|_{\psi_2} + \|P_{ij}\mathsf{SP}_1'(f)_{ij} - \rho\mathsf{SP}_1'(f)_{ij}\|_{\psi_2} \\ &\leq C\gamma + \mathsf{SP}_1'(f)_{ij}\|P_{ij} - \rho\|_{\psi_2} \\ &\leq C'(\gamma + R\Gamma_1\Gamma_2),\end{aligned}$$

where $C, C' > 0$ are absolute constants. The first inequality is due to triangle inequality, and the last follows since $P_{ij} - \rho$ is a random variable bounded between $[-\rho, 1-\rho]$ and $\mathsf{SP}_1'(f)_{ij}$ is bounded by $R\Gamma_1\Gamma_2$. With a similar argument, we can also write

$$\mathsf{SP}_0'(X) = \rho\mathsf{SP}_0'(f) + \zeta_0,$$

where each component of ζ_0 is again an independent, zero-mean random variable with $\|\cdot\|_{\psi_2}$ bounded above by $C'(\gamma + R\Gamma_1\Gamma_2)$. Now, recalling that $\mathsf{SP}_1'(X) = \rho\mathsf{SP}_1'(f) + \zeta_1$ and using Weyl's inequality (see Lemma 5), we can bound the $(k+1)$-th singular value of $\mathsf{SP}_1'(X)$ by the largest singular value of ζ_1. That is,

$$\|\widetilde{\boldsymbol{\Sigma}}_1^{\perp}\|_2^2 \le \|\zeta_1\|_2^2. \tag{63}$$

Next, we bound the term $\|(\widetilde{\boldsymbol{U}}_1^{\perp})^T\widehat{\beta}\|_2^2$.

$$\begin{aligned}
\|(\widetilde{\boldsymbol{U}}_1^{\perp})^T\widehat{\beta}\|_2^2 &= \|\widetilde{\boldsymbol{U}}_1^{\perp}(\widetilde{\boldsymbol{U}}_1^{\perp})^T\widehat{\beta}\,\|_2^2 \\
&= \|\widetilde{\boldsymbol{U}}_1^{\perp}(\widetilde{\boldsymbol{U}}_1^{\perp})^T\beta^* + \widetilde{\boldsymbol{U}}_1^{\perp}(\widetilde{\boldsymbol{U}}_1^{\perp})^T(\widehat{\beta}-\beta^*)\|_2^2 \\
&\le 2\|\widetilde{\boldsymbol{U}}_1^{\perp}(\widetilde{\boldsymbol{U}}_1^{\perp})^T\beta^*\|_2^2 + 2\|\widetilde{\boldsymbol{U}}_1^{\perp}(\widetilde{\boldsymbol{U}}_1^{\perp})^T(\widehat{\beta}-\beta^*)\|_2^2 \\
&\le 2\|\widetilde{\boldsymbol{U}}_1^{\perp}(\widetilde{\boldsymbol{U}}_1^{\perp})^T\beta^*\|_2^2 + 2\|\widehat{\beta}-\beta^*\|_2^2.
\end{aligned} \tag{64}$$

First, consider

$$\begin{aligned}
\|\widetilde{\boldsymbol{U}}_1^{\perp}(\widetilde{\boldsymbol{U}}_1^{\perp})^T\beta^*\|_2 &= \|\widetilde{\boldsymbol{U}}_1^{\perp}(\widetilde{\boldsymbol{U}}_1^{\perp})^T\boldsymbol{U}_1(\boldsymbol{U}_1)^T\beta^*\|_2 \\
&\le \left\|\boldsymbol{U}_1^{\perp}(\boldsymbol{U}_1^{\perp})^T\boldsymbol{U}_1(\boldsymbol{U}_1)^T\beta^*\right\|_2 + \left\|\left(\widetilde{\boldsymbol{U}}_1^{\perp}(\widetilde{\boldsymbol{U}}_1^{\perp})^T\boldsymbol{U}_1(\boldsymbol{U}_1)^T - \boldsymbol{U}_1^{\perp}(\boldsymbol{U}_1^{\perp})^T\boldsymbol{U}_1(\boldsymbol{U}_1)^T\right)\beta^*\right\|_2 \\
&\le \left\|\left(\widetilde{\boldsymbol{U}}_1^{\perp}(\widetilde{\boldsymbol{U}}_1^{\perp})^T - \boldsymbol{U}_1^{\perp}(\boldsymbol{U}_1^{\perp})^T\right)\beta^*\right\|_2 \\
&\le \left\|\widetilde{\boldsymbol{U}}_1^{\perp}(\widetilde{\boldsymbol{U}}_1^{\perp})^T - \boldsymbol{U}_1^{\perp}(\boldsymbol{U}_1^{\perp})^T\right\|_2 \|\beta^*\|_2 \\
&= \left\|\widetilde{\boldsymbol{U}}_1\widetilde{\boldsymbol{U}}_1^T - \boldsymbol{U}_1\boldsymbol{U}_1^T\right\|_2 \|\beta^*\|_2 .
\end{aligned} \tag{65}$$

Where in the first equality we use the fact that $\beta^* = \boldsymbol{U}_1(\boldsymbol{U}_1)^T\beta^*$, i.e., β^* lives in the column space of $\mathsf{SP}_1'(f)$ (Property 6). Next, by Wedin sin Θ Theorem (see Davis and Kahan (1970), Wedin (1972)) we bound $\left\|\widetilde{\boldsymbol{U}}_1\widetilde{\boldsymbol{U}}_1^T - \boldsymbol{U}_1\boldsymbol{U}_1^T\right\|_2$ as follows:

$$\begin{aligned}
\left\|\widetilde{\boldsymbol{U}}_1\widetilde{\boldsymbol{U}}_1^T - \boldsymbol{U}_1\boldsymbol{U}_1^T\right\|_2 \|\beta^*\|_2 &\le \frac{\|\mathsf{SP}_1'(X) - \rho\mathsf{SP}_1'(f)\|_2}{\sigma_k(\rho\mathsf{SP}_1'(f))}\|\beta^*\|_2 \\
&= \frac{\|\zeta_1\|_2}{\sigma_k(\rho\mathsf{SP}_1'(f))}\|\beta^*\|_2 .
\end{aligned} \tag{66}$$

For $\|\widehat{\beta}-\beta^*\|_2$, we have:

$$\begin{aligned}\|\widehat{\beta}-\beta^*\|_2^2 &= \|\widetilde{\boldsymbol{U}}_0^{\perp}(\widetilde{\boldsymbol{U}}_0^{\perp})^T(\widehat{\beta}-\beta^*)+\widetilde{\boldsymbol{U}}_0(\widetilde{\boldsymbol{U}}_0)^T(\widehat{\beta}-\beta^*)\|_2^2\\ &= \|\widetilde{\boldsymbol{U}}_0^{\perp}(\widetilde{\boldsymbol{U}}_0^{\perp})^T(\widehat{\beta}-\beta^*)\|_2^2+\|\widetilde{\boldsymbol{U}}_0(\widetilde{\boldsymbol{U}}_0)^T(\widehat{\beta}-\beta^*)\|_2^2\\ &= \|\widetilde{\boldsymbol{U}}_0^{\perp}(\widetilde{\boldsymbol{U}}_0^{\perp})^T(\widehat{\beta}-\beta^*)\|_2^2+\|\widetilde{\boldsymbol{U}}_0^T(\widehat{\beta}-\beta^*)\|_2^2\\ &= \|\widetilde{\boldsymbol{U}}_0^{\perp}(\widetilde{\boldsymbol{U}}_0^{\perp})^T(\beta^*)\|_2^2+\|\widetilde{\boldsymbol{U}}_0^T(\widehat{\beta}-\beta^*)\|_2^2. \end{aligned} \tag{67}$$

Note that the last equality follow from the fact that $\widehat{\beta}=\widehat{\mathsf{SP}_0'}^{T,\dagger}\frac{1}{\widehat{\rho}}\mathsf{SP}_0(X)_{L\cdot}=\widetilde{\boldsymbol{U}}_0(\widetilde{\boldsymbol{\Sigma}}_0)^{\dagger}\widetilde{\boldsymbol{V}}^T\mathsf{SP}_0(X)_{L\cdot}$, where $\widehat{\mathsf{SP}_0'}^{T,\dagger}$ is the pseudoinverse of $\widehat{\mathsf{SP}_0'}^{T}$, and thus $(\widetilde{\boldsymbol{U}}_0^{\perp})^T\widehat{\beta}=0$. The first term in (67) can be bounded using the same argument in (65) and (66), where we utilize the fact that $\beta^*=\boldsymbol{U}_0(\boldsymbol{U}_0)^T\beta^*$ and Wedin $\sin\Theta$ Theorem to get

$$\|\widetilde{\boldsymbol{U}}_0^{\perp}(\widetilde{\boldsymbol{U}}_0^{\perp})^T\beta^*\|_2 \le \frac{\|\zeta_0\|_2}{\sigma_k(\rho\mathsf{SP}_0'(f))}\|\beta^*\|_2\,. \tag{68}$$

What is left is bounding $\|\widetilde{\boldsymbol{U}}_0^T(\widehat{\beta}-\beta^*)\|_2^2$. To that end, first consider

$$\begin{aligned}\|\widehat{\mathsf{SP}_0'}^{T}(\widehat{\beta}-\beta^*)\|_2^2 &\le 2\|\widehat{\mathsf{SP}_0'}^{T}\widehat{\beta}-\mathsf{SP}_0'(f)^T\beta^*\|_2^2+2\|\mathsf{SP}_0'(f)^T\beta^*-\widehat{\mathsf{SP}_0'}^{T}\beta^*\|_2^2\\ &\le 2\|\widehat{\mathsf{SP}_0'}^{T}\widehat{\beta}-\mathsf{SP}_0'(f)^T\beta^*\|_2^2+2\|\mathsf{SP}_0'(f)-\widehat{\mathsf{SP}_0'}\|_{2,\infty}^2\|\beta^*\|_1^2. \end{aligned} \tag{69}$$

Also, consider

$$\begin{aligned}\|\widehat{\mathsf{SP}_0'}^{T}(\widehat{\beta}-\beta^*)\|_2^2 &= (\widehat{\beta}-\beta^*)^T\frac{1}{\widehat{\rho}^2}\widetilde{\boldsymbol{U}}_0\widetilde{\boldsymbol{\Sigma}}_0^2\widetilde{\boldsymbol{U}}^T(\widehat{\beta}-\beta^*)\\ &\ge \sigma_k(\widehat{\mathsf{SP}_0'})^2\|\widetilde{\boldsymbol{U}}_0^T(\widehat{\beta}-\beta^*)\|_2^2. \end{aligned} \tag{70}$$

From (70) and (69) we get,

$$\|\widetilde{\boldsymbol{U}}_0^T(\widehat{\beta}-\beta^*)\|_2^2 \le \frac{2}{\sigma_k(\widehat{\mathsf{SP}_0'})^2}(\|\widehat{\mathsf{SP}_0'}^{T}\widehat{\beta}-\mathsf{SP}_0'(f)^T\beta^*\|_2^2+\|\mathsf{SP}_0'(f)-\widehat{\mathsf{SP}_0'}\|_{2,\infty}^2\|\beta^*\|_1^2). \tag{71}$$

Note that, similar to argument in (62), $\mathsf{SP}_0(X)_{L\cdot}=\rho\mathsf{SP}_0(f)_{L\cdot}+\zeta_0^L$, where ζ_0^L is a vector of i.i.d. entries with $\|\cdot\|_{\psi_2}\le C'(\gamma+R\Gamma_1\Gamma_2)$. Then the term $\|\widehat{\mathsf{SP}_0'}^{T}\widehat{\beta}-\mathsf{SP}_0'(f)^T\beta^*\|_2^2$ can be bounded as follows

$$\|\widehat{\mathsf{SP}_0'}^{T}\widehat{\beta}-\frac{1}{\rho}\mathsf{SP}_0(X)_{L\cdot}\|_2^2$$

$$
\begin{aligned}
&= \|\widehat{\mathsf{SP}_0'}^T \widehat{\beta} - \mathsf{SP}_0(f)_{L\cdot} - \frac{1}{\rho}\zeta_0^L\|_2^2 \\
&= \|\widehat{\mathsf{SP}_0'}^T \widehat{\beta} - \mathsf{SP}_0'(f)^T\beta^*\|_2^2 + \|\frac{1}{\rho}\zeta_0^L\|_2^2 - \frac{2}{\rho}(\widehat{\mathsf{SP}_0'}^T \widehat{\beta} - \mathsf{SP}_0'(f)^T\beta^*)^T\zeta_0^L. \qquad (72)
\end{aligned}
$$

Also, we have

$$
\begin{aligned}
&\|\widehat{\mathsf{SP}_0'}^T \widehat{\beta} - \frac{1}{\rho}\mathsf{SP}_0(X)_{L\cdot}\|_2^2 \\
\leq &\|\widehat{\mathsf{SP}_0'}^T \beta^* - \frac{1}{\rho}\mathsf{SP}_0(X)_{L\cdot}\|_2^2 \\
= &\|(\widehat{\mathsf{SP}_0'}^T - \mathsf{SP}_0'(f)^T)\beta^* - \frac{1}{\rho}\zeta_0^L\|_2^2 \\
= &\|(\widehat{\mathsf{SP}_0'}^T - \mathsf{SP}_0'(f)^T)\beta^*\|_2^2 + \|\frac{1}{\rho}\zeta_0^L\|_2^2 - \frac{2}{\rho}\left((\widehat{\mathsf{SP}_0'}^T - \mathsf{SP}_0'(f)^T)\beta^*\right)^T\zeta_0^L. \qquad (73)
\end{aligned}
$$

From (72) and (73) we have,

$$
\begin{aligned}
\|\widehat{\mathsf{SP}_0'}^T \widehat{\beta} - \mathsf{SP}_0'(f)^T\beta^*\|_2^2 &\leq \|(\widehat{\mathsf{SP}_0'}^T - \mathsf{SP}_0'(f)^T)\beta^*\|_2^2 + \frac{2}{\rho}\left((\widehat{\mathsf{SP}_0'}^T)(\widehat{\beta} - \beta^*)\right)^T\zeta_0^L \qquad (74) \\
&\leq \|\widehat{\mathsf{SP}_0'} - \mathsf{SP}_0'(f)\|_{2,\infty}^2\|\beta^*\|_1^2 + \frac{2}{\rho}\left((\widehat{\mathsf{SP}_0'}^T)(\widehat{\beta} - \beta^*)\right)^T\zeta_0^L.
\end{aligned}
$$

Finally, from (71) and (74) we get

$$
\|\widetilde{\boldsymbol{U}}_0^T(\widehat{\beta} - \beta^*)\|_2^2 \leq \frac{4}{\sigma_k(\widehat{\mathsf{SP}_0'})^2}\left(\|\mathsf{SP}_0'(f) - \widehat{\mathsf{SP}_0'}\|_{2,\infty}^2\|\beta^*\|_1^2 + \frac{1}{\rho}\left((\widehat{\mathsf{SP}_0'}^T)(\widehat{\beta} - \beta^*)\right)^T\zeta_0^L\right). \qquad (75)
$$

From (67), (68), and (75) we have

$$
\begin{aligned}
\|\widehat{\beta} - \beta^*\|_2^2 \leq &\frac{\|\zeta_0\|_2^2}{\sigma_k(\rho\mathsf{SP}_0'(f))^2}\|\beta^*\|_2^2 \qquad (76) \\
&+ \frac{4}{\sigma_k(\widehat{\mathsf{SP}_0'})^2}\left(\|\mathsf{SP}_0'(f) - \widehat{\mathsf{SP}_0'}\|_{2,\infty}^2\|\beta^*\|_1^2 + \frac{1}{\rho}\left((\widehat{\mathsf{SP}_0'}^T)(\widehat{\beta} - \beta^*)\right)^T\zeta_0^L\right).
\end{aligned}
$$

For ease of exposition, let

$$
\begin{aligned}
\Delta_1 &\coloneqq \|\mathsf{SP}_0'(f) - \widehat{\mathsf{SP}_0'}\|_{2,\infty}^2\|\beta^*\|_1^2 + \frac{1}{\rho}\left(\widehat{\mathsf{SP}_0'}^T(\widehat{\beta} - \beta^*)\right)^T\zeta_0^L \\
\Delta_2 &\coloneqq \frac{\|\zeta_0\|_2^2}{\sigma_k(\rho\mathsf{SP}_0'(f))^2}\|\beta^*\|_2^2 + \frac{4}{\sigma_k(\widehat{\mathsf{SP}_0'})^2}(\Delta_1). \qquad (77)
\end{aligned}
$$

Using this definition, (61), (63), (64), (66), and (76), we have

$$\|\frac{1}{\widehat{\rho}}\mathsf{SP}_1'(X)^T\widehat{\beta} - \widehat{\mathsf{SP}_1'}^T\widehat{\beta}\|_2^2 \le \|\frac{1}{\widehat{\rho}}\zeta_1\|_2^2\left(\frac{2\|\zeta_1\|_2^2\|\beta^*\|_2^2}{\sigma_k(\rho\mathsf{SP}_1'(f))^2} + 2\Delta_2\right). \tag{78}$$

Second term: $\|\mathsf{SP}_1'(f)^T\beta^* - \widehat{\mathsf{SP}_1'}^T\widehat{\beta}\|_2^2$. To bound the second term, we follow a similar proof to that shown in Agarwal et al. (2020).

$$\begin{aligned}\|\mathsf{SP}_1'(f)^T\beta^* - \widehat{\mathsf{SP}_1'}^T\widehat{\beta}\|_2^2 &= \|\mathsf{SP}_1'(f)^T\beta^* + \widehat{\mathsf{SP}_1'}^T\beta^* - \widehat{\mathsf{SP}_1'}^T\beta^* - \widehat{\mathsf{SP}_1'}^T\widehat{\beta}\|_2^2 \\ &\le 2\|(\mathsf{SP}_1'(f) - \widehat{\mathsf{SP}_1'})^T\beta^*\|_2^2 + 2\|\widehat{\mathsf{SP}_1'}^T(\beta^* - \widehat{\beta})\|_2^2.\end{aligned} \tag{79}$$

Next, we bound the two terms on the right hand side. First, we bound $\|(\mathsf{SP}_1'(f) - \widehat{\mathsf{SP}_1'})^T\beta^*\|_2^2$ as follows.

$$\|(\mathsf{SP}_1'(f) - \widehat{\mathsf{SP}_1'})^T\beta^*\|_2^2 \le \|\mathsf{SP}_1'(f) - \widehat{\mathsf{SP}_1'}\|_{2,\infty}^2\|\beta^*\|_1^2. \tag{80}$$

Next, we bound the second term $\|\widehat{\mathsf{SP}_1'}^T(\beta^* - \widehat{\beta})\|_2^2$.

$$\begin{aligned}\|\widehat{\mathsf{SP}_1'}^T(\beta^* - \widehat{\beta})\|_2^2 &\le \frac{1}{\widehat{\rho}^2}\|(\widetilde{\boldsymbol{V}}_1\widetilde{\boldsymbol{\Sigma}}_1\widetilde{\boldsymbol{U}}_1^T + \rho\mathsf{SP}_1'(f)^T - \rho\mathsf{SP}_1'(f)^T)(\beta^* - \widehat{\beta})\|_2^2 \\ &\le \frac{2}{\widehat{\rho}^2}\|(\widetilde{\boldsymbol{V}}_1\widetilde{\boldsymbol{\Sigma}}_1\widetilde{\boldsymbol{U}}_1^T - \rho\mathsf{SP}_1'(f)^T)(\beta^* - \widehat{\beta})\|_2^2 + \frac{2\rho^2}{\widehat{\rho}^2}\|\mathsf{SP}_1'(f)^T(\beta^* - \widehat{\beta})\|_2^2 \\ &\le \frac{2}{\widehat{\rho}^2}\|\widetilde{\boldsymbol{V}}_1\widetilde{\boldsymbol{\Sigma}}_1\widetilde{\boldsymbol{U}}_1^T - \rho\mathsf{SP}_1'(f)^T\|_2^2\|(\beta^* - \widehat{\beta})\|_2^2 + \frac{2\rho^2}{\widehat{\rho}^2}\|\mathsf{SP}_1'(f)^T(\beta^* - \widehat{\beta})\|_2^2.\end{aligned}$$

Further, note that

$$\begin{aligned}\|\widetilde{\boldsymbol{V}}_1\widetilde{\boldsymbol{\Sigma}}_1\widetilde{\boldsymbol{U}}_1^T - \rho\mathsf{SP}_1'(f)^T\|_2^2 &\le 2\|\widetilde{\boldsymbol{V}}_1\widetilde{\boldsymbol{\Sigma}}_1\widetilde{\boldsymbol{U}}_1^T - \mathsf{SP}_1'(X)^T\|_2^2 + 2\|\mathsf{SP}_1'(X)^T - \rho\mathsf{SP}_1'(f)^T\|_2^2 \\ &\le 4\|\mathsf{SP}_1'(X)^T - \rho\mathsf{SP}_1'(f)^T\|_2^2 = 4\|\zeta_1\|_2^2.\end{aligned}$$

Where the last inequality follows from the fact that $\|\widetilde{\boldsymbol{V}}_1\widetilde{\boldsymbol{\Sigma}}_1\widetilde{\boldsymbol{U}}_1^T - \mathsf{SP}_1'(X)^T\|_2$ is the $k+1$-th singular value of $\mathsf{SP}_1'(X)$ and hence is bounded by $\|\mathsf{SP}_1'(X)^T - \rho\mathsf{SP}_1'(f)^T\|_2$ using Weyl's inequality. Therefore,

$$\|\widehat{\mathsf{SP}_1'}^T(\beta^* - \widehat{\beta})\|_2^2 \le \frac{8}{\widehat{\rho}^2}\|\zeta_1\|_2^2\|\beta^* - \widehat{\beta}\|_2^2 + \frac{2\rho^2}{\widehat{\rho}^2}\|\mathsf{SP}_1'(f)^T(\beta^* - \widehat{\beta})\|_2^2. \tag{81}$$

Next, we bound $\|\mathsf{SP}_1'(f)^T(\beta^* - \widehat{\beta})\|_2^2$. Recall that $\boldsymbol{U}_0$ span the column space of $\mathsf{SP}_1'(f)$. Thus $\mathsf{SP}_1'(f)^T = \mathsf{SP}_1'(f)^T \boldsymbol{U}_0 \boldsymbol{U}_0^T$, therefore,

$$\begin{aligned}\|\mathsf{SP}_1'(f)^T(\beta^* - \widehat{\beta})\|_2^2 &= \|\mathsf{SP}_1'(f)^T \boldsymbol{U}_0 \boldsymbol{U}_0^T(\beta^* - \widehat{\beta})\|_2^2 \\ &\le \|\mathsf{SP}_1'(f)\|_2^2 \|\boldsymbol{U}_0 \boldsymbol{U}_0^T(\beta^* - \widehat{\beta})\|_2^2.\end{aligned} \tag{82}$$

Recall that $\widetilde{\boldsymbol{U}}_0$ denote the top k left singular vectors of $\mathsf{SP}_0'(x)$, and consider

$$\begin{aligned}\|\boldsymbol{U}_0 \boldsymbol{U}_0^T(\beta^* - \widehat{\beta})\|_2^2 &= \|(\boldsymbol{U}_0 \boldsymbol{U}_0^T + \widetilde{\boldsymbol{U}}_0 \widetilde{\boldsymbol{U}}_0^T - \widetilde{\boldsymbol{U}}_0 \widetilde{\boldsymbol{U}}_0^T)(\beta^* - \widehat{\beta})\|_2^2 \\ &\le 2\|\boldsymbol{U}_0 \boldsymbol{U}_0^T - \widetilde{\boldsymbol{U}}_0 \widetilde{\boldsymbol{U}}_0^T\|_2^2 \|\beta^* - \widehat{\beta}\|_2^2 + 2\|\widetilde{\boldsymbol{U}}_0 \widetilde{\boldsymbol{U}}_0^T(\beta^* - \widehat{\beta})\|_2^2.\end{aligned} \tag{83}$$

Using (83), (75) and Wedin $\sin\Theta$ Theorem, we obtain,

$$\begin{aligned}\|\boldsymbol{U}_0 \boldsymbol{U}_0^T(\beta^* - \widehat{\beta})\|_2^2 \le{}& \frac{2\|\zeta_0\|_2^2}{\sigma_k(\rho \mathsf{SP}_0'(f))^2}\|\beta^* - \widehat{\beta}\|_2^2 \\ &+ \frac{8}{\sigma_k(\widehat{\mathsf{SP}_0'})^2}\left(\|\mathsf{SP}_0'(f) - \widehat{\mathsf{SP}_0'}\|_{2,\infty}^2 \|\beta^*\|_1^2 + \frac{1}{\rho}\left(\widehat{\mathsf{SP}_0'}^T(\widehat{\beta} - \beta^*)\right)^T \zeta_0^L\right).\end{aligned} \tag{84}$$

Using (82) and (84), we have

$$\begin{aligned}\|\mathsf{SP}_1'(f)^T(\beta^* - \widehat{\beta})\|_2^2 \le{}& \|\mathsf{SP}_1'(f)\|_2^2 \frac{2\|\zeta_0\|_2^2}{\sigma_k(\rho \mathsf{SP}_0'(f))^2}\|\beta^* - \widehat{\beta}\|_2^2 \\ &+ \frac{8\|\mathsf{SP}_1'(f)\|_2^2}{\sigma_k(\widehat{\mathsf{SP}_0'})^2}\left(\|\mathsf{SP}_0'(f) - \widehat{\mathsf{SP}_0'}\|_{2,\infty}^2 \|\beta^*\|_1^2 + \frac{1}{\rho}\left(\widehat{\mathsf{SP}_0'}^T(\widehat{\beta} - \beta^*)\right)^T \zeta_0^L\right).\end{aligned} \tag{85}$$

Finally, using (85) and (81), we have

$$\begin{aligned}\|\widehat{\mathsf{SP}_1'}^T(\beta^* - \widehat{\beta})\|_2^2 \le{}& \frac{8}{\widehat{\rho}^2}\|\zeta_1\|_2^2 \|\beta^* - \widehat{\beta}\|_2^2 \\ &+ \frac{4}{\widehat{\rho}^2}\frac{\|\zeta_0\|_2^2 \|\mathsf{SP}_1'(f)\|_2^2}{\sigma_k(\mathsf{SP}_0'(f))^2}\|\beta^* - \widehat{\beta}\|_2^2 \\ &+ \frac{16\rho^2}{\widehat{\rho}^2}\frac{\|\mathsf{SP}_1'(f)\|_2^2}{\sigma_k(\widehat{\mathsf{SP}_0'})^2}\left(\|\mathsf{SP}_0'(f) - \widehat{\mathsf{SP}_0'}\|_{2,\infty}^2 \|\beta^*\|_1^2 + \frac{1}{\rho}\left(\widehat{\mathsf{SP}_0'}^T(\widehat{\beta} - \beta^*)\right)^T \zeta_0^L\right).\end{aligned} \tag{86}$$

Finally, combining (86), (80), (79), and (77) yields,

$$\|\mathsf{SP}_1'(f)^T\beta^* - \widehat{\mathsf{SP}_1'}^T\widehat{\beta}\|_2^2 \le C\|\mathsf{SP}_1'(f) - \widehat{\mathsf{SP}_1'}\|_{2,\infty}^2\|\beta^*\|_1^2 + \frac{C}{\widehat{\rho}^2}\|\zeta_1\|_2^2\Delta_2 + \frac{C}{\widehat{\rho}^2}\frac{\|\zeta_0\|_2^2\|\mathsf{SP}_1'(f)\|_2^2}{\sigma_k(\mathsf{SP}_0'(f))^2}\Delta_2 + \frac{C\rho^2}{\widehat{\rho}^2}\frac{\|\mathsf{SP}_1'(f)\|_2^2\Delta_1}{\sigma_k(\widehat{\mathsf{SP}_0'})^2}. \tag{87}$$

Combining. Incorporating the two bounds in (78) and (87) yields,

$$\begin{aligned}\|\frac{1}{\widehat{\rho}}\mathsf{SP}_1'(X)^T\widehat{\beta} - \mathsf{SP}_1'(f)^T\beta^*\|_2^2 \le C\|\frac{1}{\widehat{\rho}}\zeta_1\|_2^2\left(\frac{\|\zeta_1\|_2^2\|\beta^*\|_2^2}{\sigma_k(\rho\mathsf{SP}_1'(f))^2} + \Delta_2\right) \\ + C\|\mathsf{SP}_1'(f) - \widehat{\mathsf{SP}_1'}\|_{2,\infty}^2\|\beta^*\|_1^2 \\ + \frac{C}{\widehat{\rho}^2}\frac{\|\zeta_0\|_2^2\|\mathsf{SP}_1'(f)\|_2^2}{\sigma_k(\mathsf{SP}_0'(f))^2}\Delta_2 + \frac{C\rho^2}{\widehat{\rho}^2}\frac{\|\mathsf{SP}_1'(f)\|_2^2\Delta_1}{\sigma_k(\widehat{\mathsf{SP}_0'})^2}.\end{aligned} \tag{88}$$

For some absolute constant $C > 0$.

High Probability Bound. We start by defining the following high probability events. Let $C(\Gamma_1, \Gamma_2, \gamma)$ be a positive constant dependent on model parameters $\Gamma_1, \Gamma_2, \gamma$, and let $C > 0$ be some positive absolute constant, define

$$\begin{aligned}\bar{E}_1 &:= \left\{\|\zeta_0\|_2 \le C(\gamma + R\Gamma_1\Gamma_2)\sqrt{NT/L}\right\}, \\ \bar{E}_2 &:= \left\{\|\zeta_1\|_2 \le C(\gamma + R\Gamma_1\Gamma_2)\sqrt{NT_1/L}\right\}, \\ \bar{E}_3 &:= \left\{\left(1 - \sqrt{\frac{20\log(NT)}{\rho NT}}\right)\rho \le \widehat{\rho} \le \frac{1}{1 - \sqrt{\frac{20\log(NT)}{\rho NT}}}\rho\right\}, \\ \bar{E}_4 &:= \left\{\|\mathsf{SP}_0'(f) - \widehat{\mathsf{SP}_0'}\|_{2,\infty}^2 \le C(\gamma, \Gamma_1, \Gamma_2)\left(\frac{(NT)^2R^2}{\rho^4\sigma_k(\mathsf{SP}_0'(f))^2L^2} + \frac{kR^2\log NT/L}{\rho^2}\right)\right\}, \qquad (89) \\ \bar{E}_5 &:= \left\{\|\mathsf{SP}_1'(f) - \widehat{\mathsf{SP}_1'}\|_{2,\infty}^2 \le C(\gamma, \Gamma_1, \Gamma_2)\left(\frac{(NT_1)^2R^2}{\rho^4\sigma_k(\mathsf{SP}_1'(f))^2L^2} + \frac{kR^2\log NT_1/L}{\rho^2} + \frac{R^2T_1}{T}\right)\right\} \qquad (90)\end{aligned}$$

Using Theorem 9, we have the following,

$$\mathbb{P}(\bar{E}_1) \ge 1 - 2\exp\left(\frac{-NT}{L}\right),$$
$$\mathbb{P}(\bar{E}_2) \ge 1 - 2\exp\left(\frac{-NT_1}{L}\right).$$

Further by Lemma 8, $\mathbb{P}(\bar{E}_3) \ge 1 - \frac{2}{(NT)^{10}}$. Finally, the probabilities of $\bar{E}_4$ and $\bar{E}_5$ are bounded as we show next.

LEMMA 9. *Let $\bar{E}_4$ and $\bar{E}_5$ be defined as in (89) and (90). Then, for a constant $C > 0$,*

$$\mathbb{P}(\bar{E}_4) \geq 1 - \frac{C}{(NT)^{10}},$$
$$\mathbb{P}(\bar{E}_5) \geq 1 - \frac{C}{(NT_1)^{10}} - \frac{C}{(NT)^{10}}.$$

Bounding $\bar{E}_4$ and $\bar{E}_5$. $\mathbb{P}(\bar{E}_4)$ and $\mathbb{P}(\bar{E}_5)$ can be bounded using a direct utilization of Lemma 7 and the high probability events defined in Appendix G.5. Starting with $\bar{E}_4$, using (30), and recalling that in this theorem setup $\epsilon = 0, \Gamma = R\Gamma_1\Gamma_2$ (Property 1 and Property 2) and $\sigma = \gamma$ (Property 3), we have that with probability $1 - \frac{C}{(NT)^{10}}$,

$$\begin{aligned}\|\mathsf{SP}_0'(f) - \widehat{\mathsf{SP}_0'}\|_{2,\infty}^2 &\leq C\frac{\gamma^2(NT)^2}{\rho^2\sigma_k(\mathsf{SP}_0'(f))^2L^2}\left((R\Gamma_1\Gamma_2)^2 + \frac{\gamma^2}{\rho^2}\right) + \frac{C\gamma^2 k\log NT/L}{\rho^2} + C(R\Gamma_1\Gamma_2)^2 \\ &\leq C(\gamma,\Gamma_1,\Gamma_2)\left(\frac{(NT)^2R^2}{\rho^4\sigma_k(\mathsf{SP}_0'(f))^2L^2} + \frac{kR^2\log NT/L}{\rho^2}\right).\end{aligned}$$

A similar argument can be used for $\bar{E}_5$, while noting that the term $\frac{C}{(NT)^{10}}$ shows up due to utilizing the estimate $\widehat{\rho}$, which is estimated from the first T observations. Precisely, we get the following,

$$\|\mathsf{SP}_1'(f) - \widehat{\mathsf{SP}_1'}\|_{2,\infty}^2 \leq C\frac{\gamma^2(NT_1)^2}{\rho^2\sigma_k(\mathsf{SP}_1'(f))^2L^2}\left((R\Gamma_1\Gamma_2)^2 + \frac{\gamma^2}{\rho^2}\right) + \frac{C\gamma^2 k\log(NT_1/L)}{\rho^2} + C\frac{(R\Gamma_1\Gamma_2)^2T_1}{T}$$

$$\leq C(\gamma,\Gamma_1,\Gamma_2)\left(\frac{(NT_1)^2R^2}{\rho^4\sigma_k(\mathsf{SP}_1'(f))^2L^2} + \frac{R^2k\log(NT_1/L)}{\rho^2} + \frac{R^2T_1}{T}\right).$$

Now, given these events, we will provide the high probability bound. Let $\bar{E} := \bar{E}_1 \cap \bar{E}_2 \cap \bar{E}_3 \cap \bar{E}_4 \cap \bar{E}_5$.

$$\mathbb{P}(\bar{E}^c) \leq \frac{C_0}{(NT)^{10}} + \frac{C_1}{(NT_1)^{10}}, \tag{91}$$

for some absolute constants $C_0, C_1 > 0$. Note that under event $\bar{E}_3$, we have that $\widehat{\rho} \geq \rho\left(1 - \sqrt{\frac{20\log(NT)}{\rho NT}}\right)$. By further using the assumption $\rho \geq C\log(NT)/\sqrt{NT}$ for a sufficiently large C we have that $\widehat{\rho} \geq C'\rho$ and $\frac{(\widehat{\rho}-\rho)^2}{\widehat{\rho}^2} \leq \frac{C}{\sqrt{NT}}$. Now, recall Δ_1 and Δ_2 definition in (77). Under event $\bar{E}$, we can bound Δ_1 as follows,

$$\begin{aligned}\Delta_1 &= \|\widehat{\mathsf{SP}_0'} - \mathsf{SP}_0'(f)\|_{2,\infty}^2\|\beta^*\|_1^2 + \frac{1}{\rho}\left(\widehat{\mathsf{SP}_0'}^T(\widehat{\beta} - \beta^*)\right)^T\zeta_0^L \\ &\leq C(\gamma,\Gamma_1,\Gamma_2)\|\beta^*\|_1^2\left(\frac{(NT)^2R^2}{\rho^4\sigma_k(\mathsf{SP}_0'(f))^2L^2} + \frac{kR^2\log(NT/L)}{\rho^2}\right) + \frac{1}{\rho}\left(\widehat{\mathsf{SP}_0'}^T(\widehat{\beta} - \beta^*)\right)^T\zeta_0^L.\end{aligned}$$

Similarly, under event $\bar{E}$, we can bound Δ_2 as follows,

$$\Delta_2 \le C(\gamma,\Gamma_1,\Gamma_2)\|\beta^*\|_1^2\left(\frac{NTR^2}{L\sigma_k(\rho \mathsf{SP}_0'(f))^2}+\frac{1}{\sigma_k(\widehat{\mathsf{SP}_0'})^2}\left(\frac{(NT)^2R^2}{\rho^4\sigma_k(\mathsf{SP}_0'(f))^2L^2}+\frac{kR^2\log NT/L}{\rho^2}\right)\right)$$
$$+\frac{C}{\rho\sigma_k(\widehat{\mathsf{SP}_0'})^2}\left(\left(\widehat{\mathsf{SP}_0'}^T(\widehat{\beta}-\beta^*)\right)^T\zeta_0^L\right).$$

Further, using Weyl's inequality (see Lemma 5), we can bound $|\sigma_k(\widehat{\mathsf{SP}_0'})-\sigma_k(\mathsf{SP}_0'(f))|$ as follows,

$$\begin{aligned}|\sigma_k(\widehat{\mathsf{SP}_0'})-\sigma_k(\rho\mathsf{SP}_0'(f))| &= \frac{1}{\widehat{\rho}}|\sigma_k(\widetilde{\mathbf{\Sigma}}_0)-\widehat{\rho}\sigma_k(\mathsf{SP}_0'(f))| \\ &\le \frac{1}{\widehat{\rho}}|\sigma_k(\widetilde{\mathbf{\Sigma}}_0)-\rho\sigma_k(\mathsf{SP}_0'(f))|+\frac{|\widehat{\rho}-\rho|}{\widehat{\rho}}\sigma_k(\mathsf{SP}_0'(f)) \\ &\le \frac{\|\zeta_0\|_2}{\widehat{\rho}}+\frac{|\widehat{\rho}-\rho|}{\widehat{\rho}}\sigma_k(\mathsf{SP}_0'(f))\end{aligned}$$

Under $\bar{E}$, and using property 4, we have that with probability of at least $1-\frac{1}{(NT)^{10}}$,

$$\begin{aligned}\frac{|\sigma_k(\widehat{\mathsf{SP}_0'})-\sigma_k(\mathsf{SP}_0'(f))|}{\sigma_k(\mathsf{SP}_0'(f))} &\le \frac{C(\gamma+R\Gamma_1\Gamma_2)\sqrt{NT/L}}{\rho\sigma_k(\mathsf{SP}_0'(f))}+\frac{|\widehat{\rho}-\rho|}{\widehat{\rho}} \\ &\le \frac{C(\gamma+R\Gamma_1\Gamma_2)\sqrt{k}}{\rho\sqrt{L}}+\frac{C}{\sqrt{NT}}.\end{aligned}$$

Using $\rho \ge C(\gamma+R\Gamma_1\Gamma_2)\sqrt{\frac{k}{L}}$ we get $\frac{1}{\sigma_k(\widehat{\mathsf{SP}_0'})^2} \le \frac{C}{\sigma_k(\mathsf{SP}_0'(f))^2}$. Using property 4, we get the following bounds for Δ_1 and Δ_2,

$$\Delta_1 \le C(\gamma,\Gamma_1,\Gamma_2,c)\|\beta^*\|_1^2kR^2\left(\frac{NT}{L^2\rho^4}+\frac{\log(NT/L)}{\rho^2}\right)+\frac{1}{\rho}\left(\widehat{\mathsf{SP}_0'}^T(\widehat{\beta}-\beta^*)\right)^T\zeta_0^L. \tag{92}$$

$$\begin{aligned}\Delta_2 &\le C(\gamma,\Gamma_1,\Gamma_2,c)\|\beta^*\|_1^2\left(\frac{kR^2}{L\rho^2}+\frac{k^2R^2}{NT}\left(\frac{NT}{L^2\rho^4}+\frac{\log(NT/L)}{\rho^2}\right)\right) \\ &+\frac{Ck}{\rho NT}\left(\left(\widehat{\mathsf{SP}_0'}^T(\widehat{\beta}-\beta^*)\right)^T\zeta_0^L\right) \\ &\le C(\gamma,\Gamma_1,\Gamma_2,c)\|\beta^*\|_1^2k^2R^2\left(\frac{1}{L\rho^2}+\frac{\log(NT/L)}{L}\right) \\ &+\frac{Ck}{\rho NT}\left(\left(\widehat{\mathsf{SP}_0'}^T(\widehat{\beta}-\beta^*)\right)^T\zeta_0^L\right),\end{aligned} \tag{93}$$

where $\rho \geq C(\gamma + R\Gamma_1\Gamma_2)\sqrt{\frac{k}{L}}$ is used to obtain the last inequality. Finally, using properties 4 and 5, $\widehat{\rho} \geq C'\rho$, and (88), (92), and (93), we have under event $\bar{E}$,

$$
\begin{aligned}
&\|\frac{1}{\widehat{\rho}}\mathsf{SP}_1'(X)^T\widehat{\beta} - \mathsf{SP}_1'(f)^T\beta^*\|_2^2 \qquad (94)\\
&\leq C(\gamma,\Gamma_1,\Gamma_2,c)\left(\frac{k^3NT_1R^6}{L^2\rho^4} + \frac{RT_1}{T}\right)\|\beta^*\|_1^2\\
&+ C(\gamma,\Gamma_1,\Gamma_2,c)\left(\frac{k^3R^6\log(NT/L)}{\rho^2}(\frac{NT_1}{L^2} + \frac{T_1}{T}) + \frac{kR^2\log(NT_1/L)}{\rho^2}\right)\|\beta^*\|_1^2\\
&+ C(\gamma,\Gamma_1,\Gamma_2,c)\frac{R^4k^2T_1}{T\rho^3}\left(\widehat{\mathsf{SP}_0'}^T(\widehat{\beta}-\beta^*)\right)^T\zeta_0^L.
\end{aligned}
$$

Expectation Bound. We get the bound in expectation using the high probability bound above, and by assuming that our forecast is bounded such that $|\bar{f}_n(T + L\times m')| \leq R\Gamma_1\Gamma_2$ for $m' \in [T_1/L]$. Specifically, we have using (94) and (91),

$$
\begin{aligned}
&\mathsf{TestForErr}(N,T,T_1,L) = \frac{1}{(NT_1/L)}\mathbb{E}\left[\left\|\frac{1}{\widehat{\rho}}\mathsf{SP}_1'(X)^T\widehat{\beta} - \mathsf{SP}_1'(f)^T\beta^*\right\|_2^2\right]\\
&\leq \frac{1}{(NT_1/L)}\mathbb{E}\left[\left\|\frac{1}{\widehat{\rho}}\mathsf{SP}_1'(X)^T\widehat{\beta} - \mathsf{SP}_1'(f)^T\beta^*\right\|_2^2\middle|\bar{E}\right] + \frac{CR^2\Gamma_1^2\Gamma_2^2}{(N\min(T,T_1))^{10}}\\
&\leq \frac{L}{NT_1}C(\gamma,\Gamma_1,\Gamma_2,c)\Bigg(\left(\frac{k^3NT_1R^6}{L^2\rho^4} + \frac{RT_1}{T}\right)\|\beta^*\|_1^2\\
&+ \left(\frac{k^3R^6\log(NT/L)}{\rho^2}\left(\frac{NT_1}{L^2} + \frac{T_1}{T}\right) + \frac{kR^2\log(NT_1/L)}{\rho^2}\right)\|\beta^*\|_1^2\\
&+ \frac{R^4k^2T_1}{T\rho^3}\mathbb{E}\left[\left(\widehat{\mathsf{SP}_0'}^T(\widehat{\beta}-\beta^*)\right)^T\zeta_0^L\middle|\bar{E}\right]\Bigg)\\
&+ \frac{CR^2\Gamma_1^2\Gamma_2^2}{(N\min(T,T_1))^{10}}.
\end{aligned}
$$

Noting that the $\mathbb{E}[\zeta_0^L|\bar{E}] = \mathbf{0}$, and ζ_0^L is independent of $\widehat{\mathsf{SP}_0'}$, $\widehat{\rho}$, β^* and the event $\bar{E}$; we have

$$\mathbb{E}\left[\left(\widehat{\mathsf{SP}_0'}^T\beta^*\right)^T\zeta_0^L\right] = 0.$$

By (3), we have $\widehat{\beta} = \widetilde{\boldsymbol{U}}_0(\widetilde{\boldsymbol{\Sigma}}_0)^\dagger\widetilde{\boldsymbol{V}}^T\mathsf{SP}_0(X)_{L\cdot}$. That is,

$$\widehat{\beta} = \widetilde{\boldsymbol{U}}_0(\widetilde{\boldsymbol{\Sigma}}_0)^\dagger\widetilde{\boldsymbol{V}}^T\rho\mathsf{SP}_0(f)_{L\cdot} + \widetilde{\boldsymbol{U}}_0(\widetilde{\boldsymbol{\Sigma}}_0)^\dagger\widetilde{\boldsymbol{V}}^T\zeta_0^L. \qquad (95)$$

Using cyclic and linearity of Trace operator; the independence properties of ζ_0^L; and (95); we have

$$\begin{aligned}
&\mathbb{E}\Big[\left(\widehat{\mathsf{SP}_0'}^T\widehat{\beta}\right)^T\zeta_0^L\Big] \qquad (96)\\
&=\mathbb{E}\Big[\left(\widehat{\mathsf{SP}_0'}^T\widetilde{\boldsymbol{U}}_0(\widetilde{\boldsymbol{\Sigma}}_0)^\dagger\widetilde{\boldsymbol{V}}^T\rho\mathsf{SP}_0(f)_{L\cdot}\right)^T\zeta_0^L\Big]+\mathbb{E}\Big[\left(\widetilde{\boldsymbol{V}}_0\widetilde{\boldsymbol{V}}^T\zeta_0^L\right)^T\zeta_0^L\Big]\\
&=\mathbb{E}[\mathsf{Tr}((\zeta_0^L)^T\widetilde{\boldsymbol{V}}_0\widetilde{\boldsymbol{V}}^T\zeta_0^L)]\\
&=\mathbb{E}[\mathsf{Tr}(\widetilde{\boldsymbol{V}}_0\widetilde{\boldsymbol{V}}^T\zeta_0^L(\zeta_0^L)^T]\\
&=\mathsf{Tr}\big(\mathbb{E}[\widetilde{\boldsymbol{V}}_0\widetilde{\boldsymbol{V}}^T]\mathbb{E}[\zeta_0^L(\zeta_0^L)^T]\big)\\
&\le C(\gamma+\Gamma_1\Gamma_2R)^2k.
\end{aligned}$$

Where to obtain the last inequality we use the trace property $\mathsf{Tr}(AB)\le\|B\|_2\mathsf{Tr}(A)$ for positive semi-definite matrices A, B, and that rank of $\widehat{\mathsf{SP}_0'}$ is k. Finally, using (96), and recalling that $T_1\ge L$ and $L\le T$ we get,

$$\begin{aligned}
&\mathsf{TestForErr}(N,T,T_1,L)\\
&\le\frac{L}{NT_1}C(\gamma,\Gamma_1,\Gamma_2,c)\left(\left(\frac{R^6k^3NT_1}{L^2\rho^4}+\frac{RT_1}{T}\right)\|\beta^*\|_1^2\right.\\
&\left.+\left(\frac{R^6k^3\log(NT/L)}{\rho^2}\left(\frac{NT_1}{L^2}+\frac{T_1}{T}\right)+\frac{R^2k\log(NT_1/L)}{\rho^2}\right)\|\beta^*\|_1^2+\frac{R^6k^3T_1}{T\rho^3}\right)\\
&+\frac{CR^2\Gamma_1^2\Gamma_2^2}{(NL)^{10}}\\
&\le\frac{L}{NT_1}C(\gamma,\Gamma_1,\Gamma_2,c)\max(1,\|\beta^*\|_1^2)\left(\frac{R^6k^3NT_1}{L^2\rho^4}+\frac{R^6k^3T_1}{T\rho^3}\right.\\
&\left.+\frac{R^6k^3\log(NT)}{\rho^2}\left(\frac{NT_1}{L^2}+\frac{T_1}{T}\right)+\frac{R^2k\log(NT_1)}{\rho^2}+\frac{R^2}{(NL)^{10}}\right)\\
&\le\frac{L}{NT_1}C(\gamma,\Gamma_1,\Gamma_2,c)\max(1,\|\beta^*\|_1^2)\left(\frac{R^6k^3\log(NT)}{\rho^4}\left(\frac{NT_1}{L^2}+\frac{T_1}{T}\right)+\frac{R^2k\log(NT_1)}{\rho^2}\right).
\end{aligned}$$

Then, with $L=\sqrt{\min(N,T)T}$, we get,

$$\begin{aligned}
&\mathsf{TestForErr}(N,T,T_1,L)\\
&\le\frac{\sqrt{\min(N,T)T}}{NT_1}C(\gamma,\Gamma_1,\Gamma_2,c)\max(1,\|\beta^*\|_1^2)\left(\frac{R^6k^3\log(NT)}{\rho^4}\left(\frac{NT_1}{T\min(N,T)}+\frac{T_1}{T}\right)+\frac{R^2k\log(NT_1)}{\rho^2}\right)
\end{aligned}$$

$$\leq \frac{T}{T_1}\frac{\sqrt{\min(N,T)T}}{NT}C(\gamma,\Gamma_1,\Gamma_2,c)\max(1,\|\beta^*\|_1^2)\left(\frac{R^6k^3\log(NT)}{\rho^4}\left(\frac{NT_1}{T\min(N,T)}+\frac{T_1}{T}\right)+\frac{R^2k\log(NT_1)}{\rho^2}\right)$$

$$\leq \frac{\sqrt{\min(N,T)T}}{NT}C(\gamma,\Gamma_1,\Gamma_2,c)\max(1,\|\beta^*\|_1^2)\left(\frac{R^6k^3\log(NT)}{\rho^4}\left(\frac{N}{\min(N,T)}+1\right)+\frac{TR^2k\log(NT_1)}{T_1\rho^2}\right)$$

$$\leq C(\gamma,\Gamma_1,\Gamma_2,c)\max(1,\|\beta^*\|_1^2)\left(\frac{R^6k^3\log(N\max(T,T_1))}{\rho^4\sqrt{\min(N,T)T}}\left(\max(1,\frac{N}{T})+\frac{T}{T_1}\right)\right).$$

Choosing $k = RG$ completes the proof.

Appendix K: Proof of Theorem 6

Setup, Notations. For $L \geq 1, k \geq 1$, for ease of notations, we define

- $\mathsf{SP}(X) = \mathsf{SP}((X_1,\ldots,X_N),T,L) \in \mathbb{R}^{L\times(NT/L)}$,
- $\mathsf{SP}(X^2) = \mathsf{SP}((X_1^2,\ldots,X_N^2),T,L) \in \mathbb{R}^{L\times(NT/L)}$,
- $\mathsf{SP}(f) = \mathsf{SP}((f_1,\ldots,f_N),T,L) \in \mathbb{R}^{L\times(NT/L)}$,
- $\mathsf{SP}(f^2) = \mathsf{SP}((f_1^2,\ldots,f_N^2),T,L) \in \mathbb{R}^{L\times(NT/L)}$,
- $\mathsf{SP}(\sigma^2) = \mathsf{SP}((\sigma_1^2,\ldots,\sigma_N^2),T,L) \in \mathbb{R}^{L\times(NT/L)}$,
- $\mathsf{SP}(f^2+\sigma^2) = \mathsf{SP}(f^2)+\mathsf{SP}(\sigma^2)$.

Recalling that $\rho = 1$, we note that

$$\mathbb{E}[\mathsf{SP}(X)] = \mathsf{SP}(f), \quad \mathbb{E}[\mathsf{SP}(X^2)] = \mathsf{SP}(f^2+\sigma^2).$$

Further, from the definition of the variance estimation algorithm, we recall

$$\widehat{\mathsf{SP}}(f) \coloneqq \widehat{\mathsf{SP}}((X_1,\ldots,X_N),T,L) = \frac{1}{\widehat{\rho}}\mathsf{HSVT}_k(\mathsf{SP}((X_1,\ldots,X_N),T,L))$$
$$\widehat{\mathsf{SP}}(f^2+\sigma^2) \coloneqq \widehat{\mathsf{SP}}((X_1^2,\ldots,X_N^2),T,L) = \frac{1}{\widehat{\rho}}\mathsf{HSVT}_k(\mathsf{SP}((X_1^2,\ldots,X_N^2),T,L))$$

We denote

- $\widehat{\mathsf{SP}}(f^2) = \widehat{\mathsf{SP}}(f)\circ\widehat{\mathsf{SP}}(f)$
- $\widehat{\mathsf{SP}}(\sigma^2) = \max\left(\widehat{\mathsf{SP}}(f^2+\sigma^2)-\widehat{\mathsf{SP}}(f^2),\mathbf{0}\right)$,

where $\mathbf{0} \in \mathbb{R}^{L\times(NT/L)}$ is a matrix of all zeroes, and we apply the $\max(\cdot)$ above entry-wise. We remind the reader the output of the variance estimation algorithm is $\widehat{\mathsf{SP}}(\sigma^2)$. Thus, we have

$$\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\left(\sigma_n(t)^2-\hat{\sigma}_n^2(t)\right)^2 = \frac{1}{NT}\|\mathsf{SP}(\sigma^2)-\widehat{\mathsf{SP}}(\sigma^2)\|_F^2.$$

Initial Decomposition. Note that since $\sigma_n^2(t) \geq 0$ for $n \in [N]$ and $t \in [T]$, we have that

$$\begin{aligned}
&\frac{1}{NT}\|\mathsf{SP}(\sigma^2) - \widehat{\mathsf{SP}}(\sigma^2)\|_F^2 \\
&\leq \frac{1}{NT}\|\mathsf{SP}(\sigma^2) - (\widehat{\mathsf{SP}}(f^2+\sigma^2) - \widehat{\mathsf{SP}}(f^2))\|_F^2 \\
&= \frac{1}{NT}\|\mathsf{SP}(f^2+\sigma^2) - \mathsf{SP}(f^2) - (\widehat{\mathsf{SP}}(f^2+\sigma^2) - \widehat{\mathsf{SP}}(f^2)\|_F^2 \\
&\leq \frac{2}{NT}\|\mathsf{SP}(f^2+\sigma^2) - \widehat{\mathsf{SP}}(f^2+\sigma^2)\|_F^2 + \frac{2}{NT}\|\mathsf{SP}(f^2) - \widehat{\mathsf{SP}}(f^2)\|_F^2 \qquad (97)
\end{aligned}$$

We bound the two terms on the r.h.s of (97) separately.

Bounding $\mathbb{E}[\|\mathsf{SP}(f^2) - \widehat{\mathsf{SP}}(f^2)\|_F^2]$.

$$\begin{aligned}
\|\mathsf{SP}(f^2) - \widehat{\mathsf{SP}}(f^2)\|_F^2 &= \sum_{n=1}^{N}\sum_{t=1}^{T}\left(f_n^2(t) - \hat{f}_n^2(t)\right)^2 \\
&= \sum_{n=1}^{N}\sum_{t=1}^{T}\left(f_n(t) - \hat{f}_n(t)\right)^2\left(f_n(t) + \hat{f}_n(t)\right)^2 \\
&\leq \left[\max_{n\in[N],t\in[T]}\left(f_n(t) + \hat{f}_n(t)\right)^2\right]\left[\sum_{n=1}^{N}\sum_{t=1}^{T}\left(f_n(t) - \hat{f}_n(t)\right)^2\right] \\
&\overset{(a)}{\leq} C(\Gamma_1,\Gamma_2,\Gamma_3)R^2\left[\sum_{n=1}^{N}\sum_{t=1}^{T}\left(f_n(t) - \hat{f}_n(t)\right)^2\right] \\
&= C(\Gamma_1,\Gamma_2,\Gamma_3)R^2\|\mathsf{SP}(f) - \widehat{\mathsf{SP}}(f)\|_F^2 \qquad (98)
\end{aligned}$$

Bounding $\|\mathsf{SP}(f^2+\sigma^2) - \widehat{\mathsf{SP}}(f^2+\sigma^2)\|_F^2$. To bound $\|\mathsf{SP}(f^2+\sigma^2) - \widehat{\mathsf{SP}}(f^2+\sigma^2)\|_F^2$, we modify the proof of Theorem 4 in a straightforward manner. The need for the modification is that Theorem 4 was proven for the case where the coordinate wise noise, $\eta_n(t) = X_n(t) - f_n(t)$ are independent sub-gaussian random variables, and $\|\eta\|_{\psi_2} \leq \gamma$. However, one can verify that $X_n^2(t) - f_n^2(t) - \sigma_n^2(t)$ is a sub-exponential random variable with $\|\cdot\|_{\psi_1}$ norm bounded as

$$\begin{aligned}
\|X_n^2(t) - f_n^2(t) - \sigma_n^2(t)\|_{\psi_1} &\leq \|X_n^2(t)\|_{\psi_1} \\
&= \|f_n^2(t) + 2f_n(t)\eta_n(t) + \eta_n^2(t)\|_{\psi_1} \\
&\leq 2\|f_n^2(t)\|_{\psi_1} + 2\|\eta_n^2(t)\|_{\psi_1} \\
&= 2\|f_n(t)\|_{\psi_2}^2 + 2\|\eta_n(t)\|_{\psi_2}^2 \\
&\leq C(\Gamma_1,\Gamma_2)R^2 + 2\gamma^2 \\
&\leq C(\Gamma_1,\Gamma_2,\gamma)R^2,
\end{aligned}$$

where we have use the standard facts that for a random variable A, $\|A - \mathbb{E}[A]\|_{\psi_1} \le \|A\|_{\psi_1}$ and $\|A^2\|_{\psi_1} = \|A\|_{\psi_2}^2$.

Further, note that by using Properties 1, 2, 10, and 11, and a straightforward modification of Proposition 14, we have

$$\begin{aligned}\operatorname{rank}(\mathsf{SP}(f^2+\sigma^2)) &\le \operatorname{rank}(\mathsf{SP}(f^2)) + \operatorname{rank}(\mathsf{SP}(\sigma^2)) \\ &\le (RG)^2 + (R'G'),\end{aligned}$$

where we have used that for any two matrices $\boldsymbol{A}, \boldsymbol{B}$, we have $\operatorname{rank}(\boldsymbol{A} \circ \boldsymbol{A}) \le \operatorname{rank}(\boldsymbol{A})^2$, where $\circ$ denotes Hadamard product, and $\operatorname{rank}(\boldsymbol{A}+\boldsymbol{B}) \le \operatorname{rank}(\boldsymbol{A}) + \operatorname{rank}(\boldsymbol{B})$. We define $\tilde{k} := (RG)^2 + (R'G')$.

Modified Theorem 4. Below, we state the modified version of Theorem 4 to get our desired result.

LEMMA 10 **(Imputation Error)**. *Let the conditions of Theorem 6 hold. Then,*

$$\begin{aligned}&\mathbb{E}\Big[\max_{j\in[L]} \frac{1}{(NT/L)} \|\mathsf{SP}(f^2+\sigma^2)^T_{L,\cdot} - \widehat{\mathsf{SP}}(f^2+\sigma^2)^T_{L,\cdot}\|_2^2\Big] \\ &\le C(\Gamma_1,\Gamma_2,\Gamma_1',\Gamma_2',\gamma,R,R')\left(\frac{(G^2+G')\log^2 NT}{L}.\right),\end{aligned}$$

where $C(\Gamma_1,\Gamma_2,\Gamma_1',\Gamma_2',\gamma,R,R')$ is a term that depends only polynomially on Γ_1, Γ_2, Γ_1', Γ_2', γ, R, R'.

T o reduce redundancy, we provide an overview of the argument needed for this proof, focusing only the parts of the arguments made in Theorem 4 that need to be modified. For ease of exposition, we let $\tilde{C} = C(\Gamma_1,\Gamma_2,\Gamma_1',\Gamma_2',\gamma,R,R')$. We being by matching notation with that used in Theorem 4; in particular with respect to $\rho, k, \epsilon, \Gamma$. Under the setup of Theorem 6, we have $\rho = 1$, $k = \tilde{k}$, $\epsilon = 0$, $\Gamma \le \tilde{C}$ Further, recall the definition of $\boldsymbol{Y}, \boldsymbol{M}, p, q, \sigma$ from Appendix G.1. We will now use $\boldsymbol{Y} = \mathsf{SP}(X^2)$, and $\boldsymbol{M} = \mathsf{SP}(f^2+\sigma^2))$, $\sigma = \gamma$, $p = (NT/L)$, $q = L$. One can verify that there is only required change to the proof of Theorem 4; in particular, in the argument made to prove Theorem 10, we need to re-define events E_2, E_3, E_4 in (26), (27), (28) for the case where $(\boldsymbol{Y} - \boldsymbol{M})_{ij}$ is mean-zero sub-exponential. Using the result from Agarwal et al. (2019, 2021), which bounds the operator norm of a matrix with sub-exponential mean-zero entries, we have with probability at least $1 - 1/((NT)^{10})$

$$\|\boldsymbol{Y} - \boldsymbol{M}\|_2 \le \tilde{C}\sqrt{(NT/L)}\log^2 NT \tag{99}$$

As a result (99), and standard concentration inequalities for sub-exponential random variables, we have the modified events, $\tilde{E}_2, \tilde{E}_3, \tilde{E}_4$.

$$\tilde{E}_2 := \Big\{ \|\boldsymbol{Y} - \rho \boldsymbol{M}\|_2 \le \tilde{C}\sqrt{(NT/L)}\log^2 NT \Big\},$$
$$\tilde{E}_3 := \Big\{ \|\boldsymbol{Y} - \rho \boldsymbol{M}\|_{\infty,2}, \|\boldsymbol{Y} - \rho \boldsymbol{M}\|_{2,\infty} \le \tilde{C}\sqrt{(NT/L)}\log^2 NT \Big\},$$
$$\tilde{E}_4 := \Big\{ \max_{j\in[q]} \|\varphi^{\boldsymbol{B}}_{\sigma_k(\boldsymbol{B})}\Big(\boldsymbol{Y}^T_{j\cdot} - \rho \boldsymbol{M}^T_{j\cdot}\Big)\|_2^2 \le \tilde{C}\tilde{k}\log^2(NT/L) \Big\},$$

Using these modified events in the proofs of Theorem 10 and Theorem 4, and appropriately simplifying leads to the desired result.

By Lemma 10 and (41), we have that

$$\frac{1}{NT}\mathbb{E}[\|\mathsf{SP}(f^2+\sigma^2) - \widehat{\mathsf{SP}}(f^2+\sigma^2)\|_F^2 \le \mathbb{E}\big[\max_{j\in[L]} \frac{1}{(NT/L)}\|\mathsf{SP}(f^2+\sigma^2)^T_{L,\cdot} - \widehat{\mathsf{SP}}(f^2+\sigma^2)^T_{L,\cdot}\|_2^2\big]$$
$$\le C(\Gamma_1,\Gamma_2,\Gamma_1',\Gamma_2',\gamma,R,R')\left(\frac{(G^2+G')\log^2 NT}{L}.\right). \quad (100)$$

Completing proof. Substituting (98) and (100) into (97) and letting $L = \sqrt{\min(N,T)T}$

$$\frac{1}{NT}\|\mathsf{SP}(\sigma^2) - \widehat{\mathsf{SP}}(\sigma^2)\|_F^2 \le C(\Gamma_1,\Gamma_2,\Gamma_3,\Gamma_1',\Gamma_2',\gamma,R,R')\left(\frac{(G^2+G')\log^2 NT}{\sqrt{\min(N,T)T}}.\right).$$

This completes the proof.

Appendix L: tSSA Proofs

L.1. Proof of Propositions 3 and 5

Consider $n \in [N],\ \ell \in [L],\ s \in [T/L]$. By Property 1,

$$\mathbf{T}_{n\ell s} = f_n((s-1)\times L + \ell)$$
$$= \sum_{r=1}^{R} U_{nr} W_{r((s-1)\times L+\ell)}. \quad (101)$$

The Hankel matrix induced by time series $W_{r\cdot}$ has rank at most G as per Property 2. The Page matrix associated with it is of dimension $L \times T/L$ with entry in its ℓ-th row and s-th column equal to $W_{r((s-1)\times L+\ell)}$. Since this Page matrix can be viewed as a sub-matrix of the Hankel matrix, it has rank at most G as well. That is, there exists vectors $w^r_{\ell\cdot}, v^r_{s\cdot} \in \mathbb{R}^G$ such that

$$W_{r((s-1)\times L+\ell)} = \sum_{g=1}^{G} w^r_{\ell g} v^r_{sg}. \quad (102)$$

From (101) and (102), it follows that

$$\begin{aligned}\mathbf{T}_{n\ell s} &= \sum_{r=1}^{R} U_{nr}\left(\sum_{g=1}^{G} w^r_{\ell g} v^r_{sg}\right) \\ &= \sum_{r\in[R],g\in[G]} U_{nr} w^r_{\ell g} v^r_{sg} \\ &= \sum_{r\in[R],g\in[G]} a_{n\ (r,g)} b_{\ell\ (r,g)} c_{s\ (r,g)}, \end{aligned} \tag{103}$$

where $a_{n\ (r,g)} = U_{nr}$, $b_{\ell\ (r,g)} = w^r_{\ell g}$ and $c_{s\ (r,g)} = v^r_{sg}$. Thus (103) implies that $\mathbf{T}$ has CP-rank at most $R \times G$, which completes the proof for Propositions 3.

For Proposition 5 , by the setup and model definition, it follows $\mathbb{T}_{n\ell s} = X_n((s-1)\times L + \ell)$. And $X_n((s-1)\times L+\ell) = \star$ with probability $1-\rho$ and $f_n((s-1)\times L+\ell) + \eta_n((s-1)\times L+\ell)$ with probability ρ, where $\eta_n((s-1)\times L+\ell)$ are independent and zero-mean. Therefore, it follows that the entries of $\mathbb{T}$ are independent and

$$\begin{aligned}\mathbb{E}[\mathbb{T}_{n\ell s}] &= \mathbb{E}[X_n((s-1)\times L+\ell)] \\ &= \rho f_n((s-1)\times L+\ell) \\ &= \rho \mathbf{T}_{n\ell s}.\end{aligned}$$

That is, $\mathbb{E}[\mathbb{T}] = \rho\mathbf{T}$. This concludes the proof.

L.2. Proof of Proposition 6

From Property 7, and our choice of parameter L for mSSA ($L = \sqrt{\min(N,T)T}$) and tSSA ($L = \sqrt{T}$), we have that

$$\mathsf{ImpErr}(N,T;\mathrm{tSSA}) = \tilde{\Theta}\left(\frac{1}{\min\left(N,\sqrt{T}\right)^2}\right) = \tilde{\Theta}\left(\frac{1}{\min\left(N^2,T\right)}\right), \tag{104}$$

$$\mathsf{ImpErr}(N,T;\mathrm{mSSA}) = \tilde{\Theta}\left(\frac{1}{\sqrt{\min(N,T)T}}\right), \tag{105}$$

$$\mathsf{ImpErr}(N,T;\mathrm{ME}) = \tilde{\Theta}\left(\frac{1}{\min\left(N,T\right)}\right). \tag{106}$$

We proceed in cases.

Case 1: $T = o(N)$**.** In this case, from (104), (105), and (106), we have

$$\mathsf{ImpErr}(N,T;\mathrm{tSSA}),\ \mathsf{ImpErr}(N,T;\mathrm{mSSA}),\ \mathsf{ImpErr}(N,T;\mathrm{ME}) = \tilde{\Theta}\left(\frac{1}{T}\right)$$

Case 2: $N = o(T)$**.** In this case, from (104), (105), and (106), we have

$$\mathsf{ImpErr}(N,T;\mathrm{tSSA}) = \tilde{\Theta}\left(\frac{1}{N^2}\right), \tag{107}$$

$$\mathsf{ImpErr}(N,T;\mathrm{mSSA}) = \tilde{\Theta}\left(\frac{1}{\sqrt{NT}}\right), \tag{108}$$

$$\mathsf{ImpErr}(N,T;\mathrm{ME}) = \tilde{\Theta}\left(\frac{1}{N}\right).$$

In this case, we have

$$\mathsf{ImpErr}(N,T;\mathrm{tSSA}), \mathsf{ImpErr}(N,T;\mathrm{mSSA}) = \tilde{o}(\mathsf{ImpErr}(N,T;\mathrm{ME})).$$

It remains to compare the relative performance of tSSA and mSSA for the regime $N = o(T)$. Towards this, note from (107) and (108) that

$$\begin{aligned}
&\mathsf{ImpErr}(N,T;\mathrm{tSSA}) = \tilde{o}(\mathsf{ImpErr}(N,T;\mathrm{mSSA})) \\
\iff &\frac{1}{N^2} = \tilde{o}\left(\frac{1}{\sqrt{NT}}\right) \\
\iff &T^{1/3} = o(N)
\end{aligned}$$

This completes the proof.

L.3. Proof of Proposition 13

PROPOSITION 15. *Let Properties 13, 2, and 3 hold. Then, for any* $1 \le L \le \sqrt{T}$, $\mathbf{HT}$ *has CP-rank at most* $R \times G$*. Further, all entries of* $\mathbb{HT}$ *are independent random variables with each entry observed with probability* $\rho \in (0,1]$*, and* $\mathbb{E}[\mathbb{HT}] = \rho\mathbf{HT}$.

Consider $n_1, \ldots, n_d \in [N_1] \times \cdots \times [N_d]$, $\ell \in [L]$, $s \in [T/L]$. By Property 13,

$$\begin{aligned}
\mathbf{HT}_{n_1,\ldots,n_d,\ell,s} &= f_{n_1,\ldots,n_d}((s-1) \times L + \ell) \\
&= \sum_{r=1}^{R} U_{n_1,r} \ldots U_{n_d,r}\, W_{r,((s-1)\times L+\ell)},
\end{aligned}$$

The rest of the proof follows in a similar fashion to that of Proposition 3.

Appendix M: Additional Figures

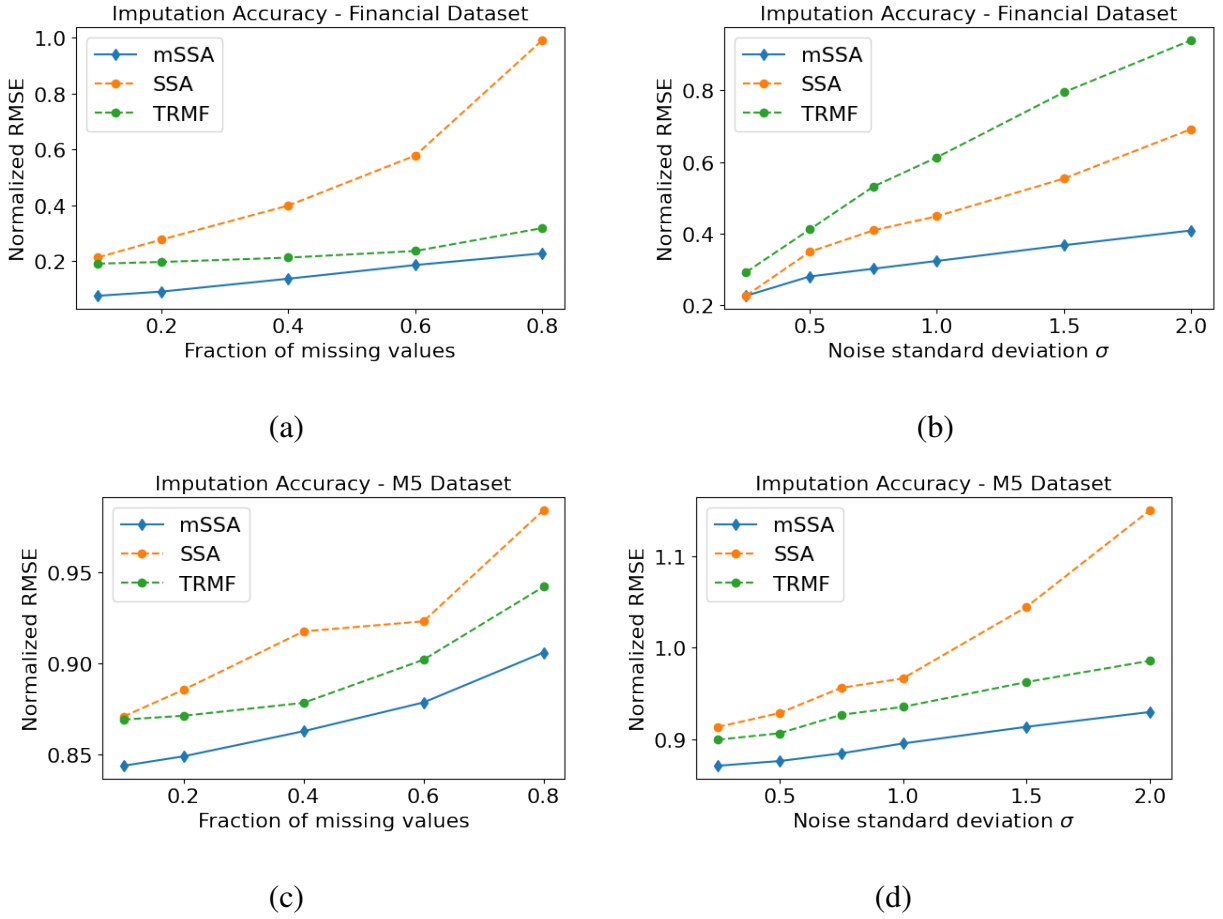

Figure 6 mSSA vs. TRMF vs. SSA - imputation performance on the Financial and M5 datasets. Figures 6a, and 6c show imputation accuracy of mSSA, TRMF and SSA as we vary the fraction of missing values; Figures 6b, and 6d show imputation accuracy as we vary the noise level (and with 50% of values missing).

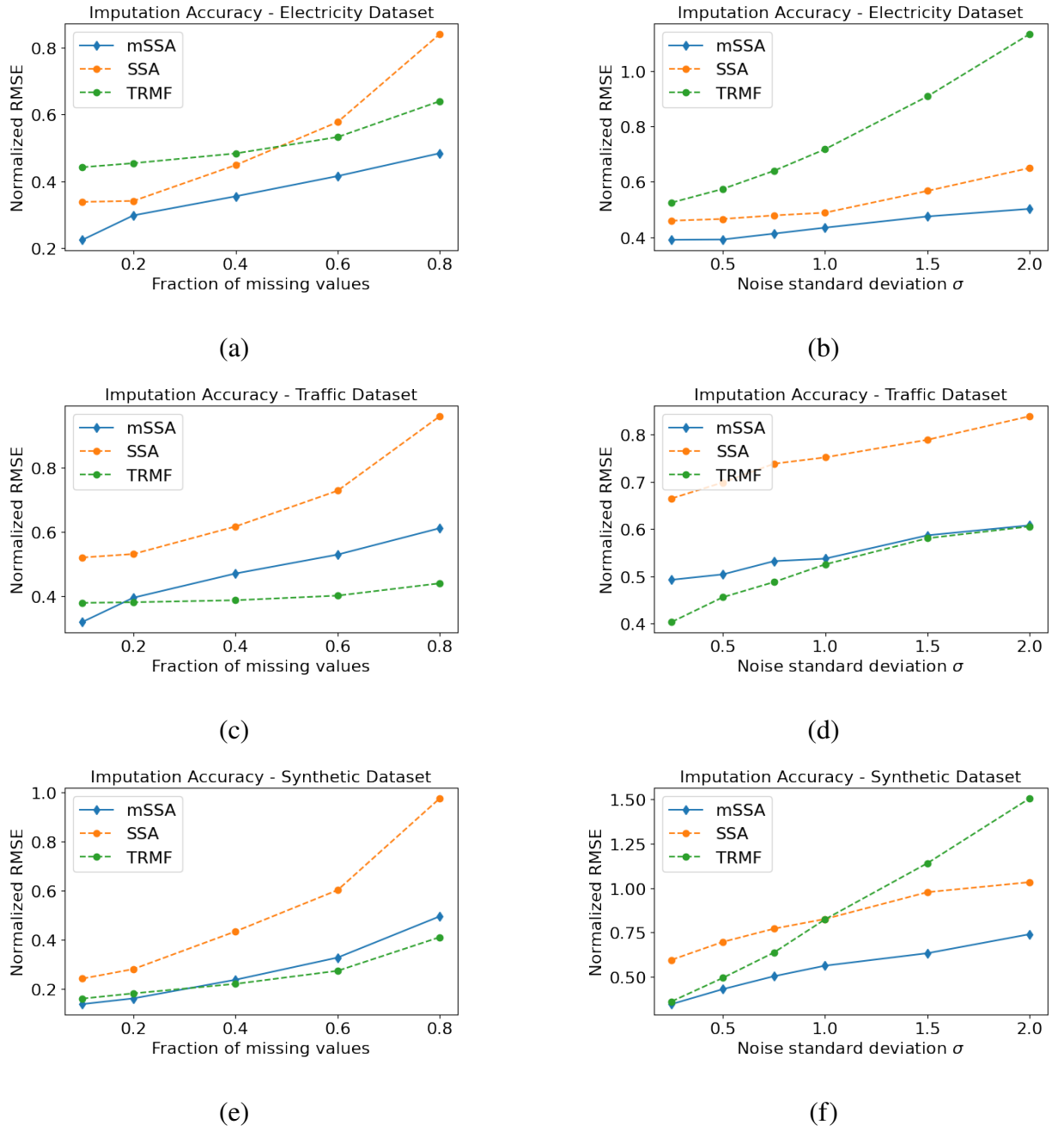

Figure 7 mSSA vs. TRMF vs. SSA - imputation performance on the Electricity, Traffic and Synthetic datasets. Figures 7a, 7c, and 7e, show imputation accuracy of mSSA, TRMF and SSA as we vary the fraction of missing values; Figures 7b, 7d, and 7f show imputation accuracy as we vary the noise level (and with 50% of values missing).

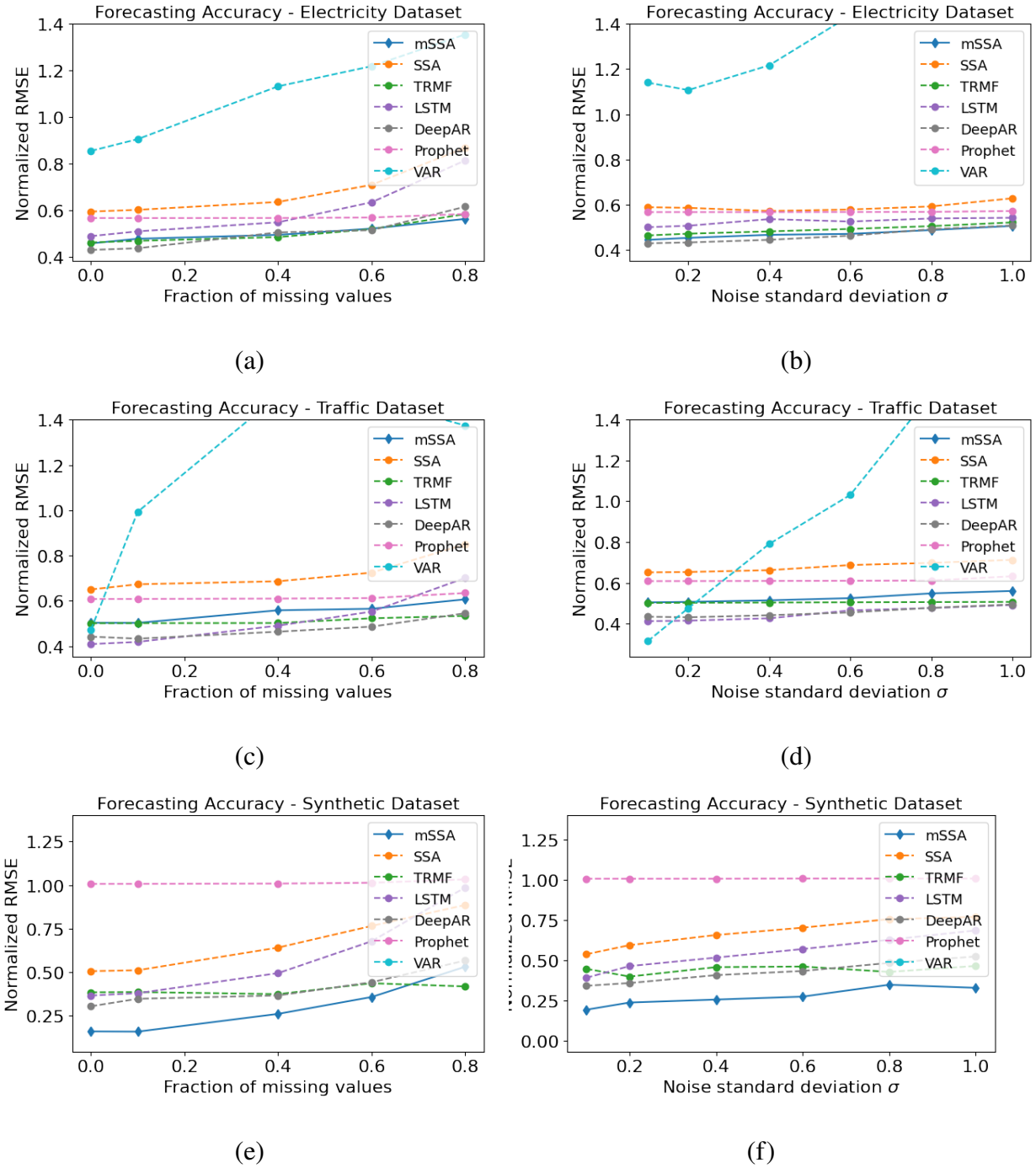

Figure 8 mSSA forecasting performance on standard multivariate time series benchmark is competitive with/outperforming industry standard methods as we vary the number of missing data and noise level. Figures 8a, 8c, and 8e show the forecasting accuracy of all methods (some of VAR results are not shown due to its relatively high error) on the Electricity, Traffic and Synthetic datasets with varying fraction of missing values; Figures 8b, 8d, and 8f shows the forecasting accuracy on the same datasets with varying noise level.

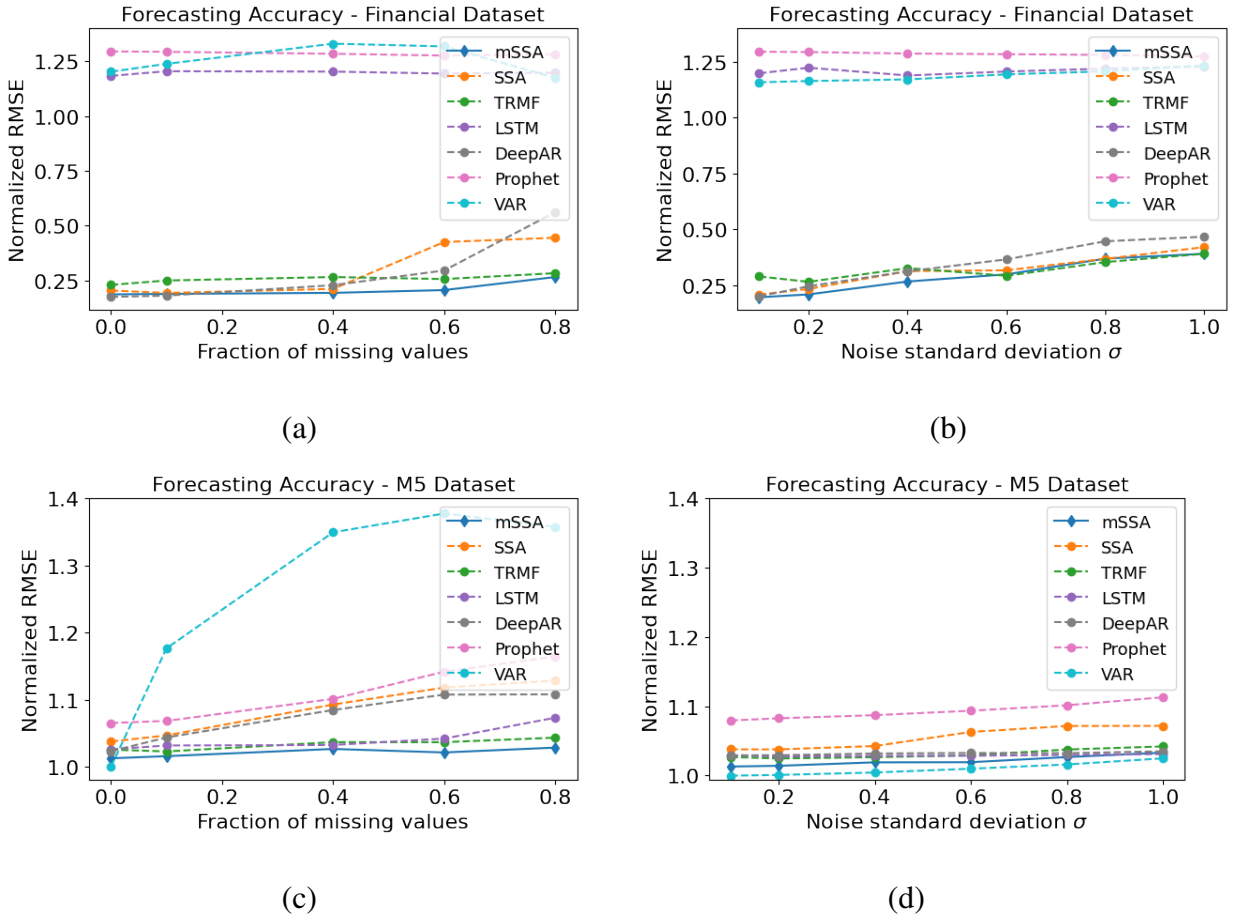

Figure 9 Figures 9a, and 9c show the forecasting accuracy of all methods (some of VAR results are not shown due to its relatively high error) on the financial and M5 datasets with varying fraction of missing values; Figures 9b, and 9d show the forecasting accuracy on the same datasets with varying noise levels.

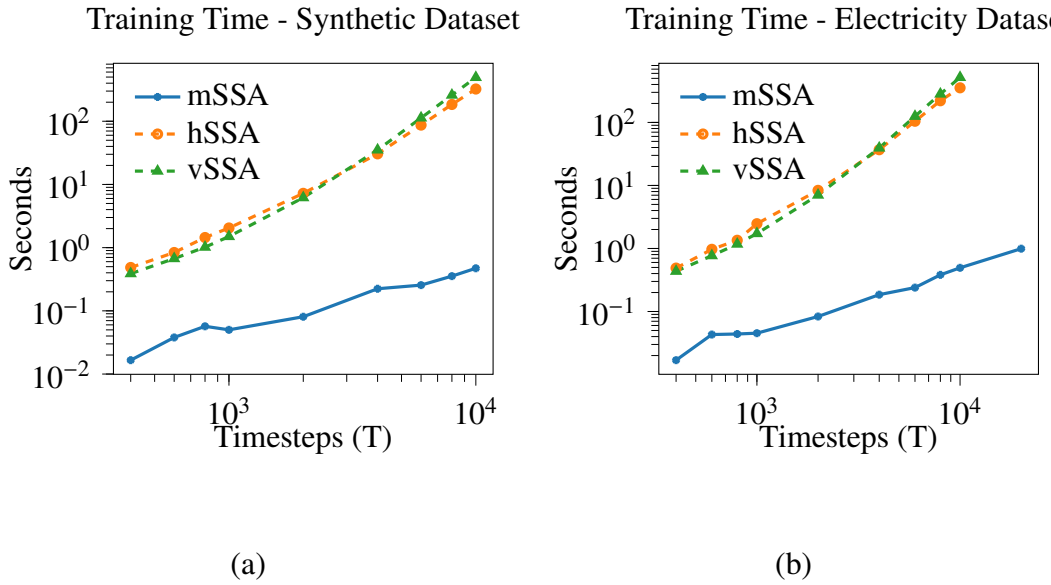

Figure 10 The training time of the original mSSA variants (hSSA in the orange dotted line and vSSA in the green dotted line) are orders of magnitude higher than that of the mSSA variant we propose (blue solid line).

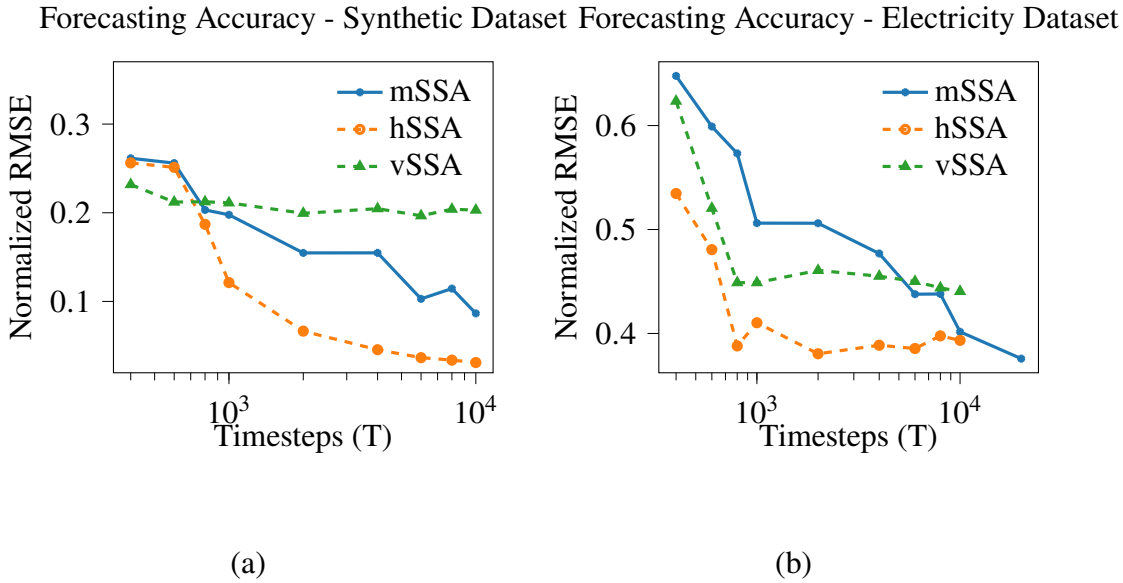

Figure 11 The forecasting error of the original mSSA variants (hSSA in the orange dotted line and vSSA in the green dotted line) and the proposed mSSA variant (blue solid line) as we increase T.

References

Anish Agarwal, Abdullah Alomar, and Devavrat Shah. 2022. On multivariate singular spectrum analysis and its variants. *ACM SIGMETRICS Performance Evaluation Review* 50, 1 (2022), 79–80.

Anish Agarwal, Muhammad Jehangir Amjad, Devavrat Shah, and Dennis Shen. 2018. Model Agnostic Time Series Analysis via Matrix Estimation. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 2, 3 (2018), 40.

Anish Agarwal, Devavrat Shah, and Dennis Shen. 2020. On Principal Component Regression in a High-Dimensional Error-in-Variables Setting. *arXiv preprint arXiv:2010.14449* (2020).

Anish Agarwal, Devavrat Shah, Dennis Shen, and Dogyoon Song. 2019. On robustness of principal component regression. In *Advances in Neural Information Processing Systems*. 9889–9900.

Anish Agarwal, Devavrat Shah, Dennis Shen, and Dogyoon Song. 2021. On Robustness of Principal Component Regression. *Accepted to appear in Journal of the American Statistical Association* (2021).

Abdullah Alomar, Munther Dahleh, Sean Mann, and Devavrat Shah. 2023. Samossa: Multivariate singular spectrum analysis with stochastic autoregressive noise. *Advances in Neural Information Processing Systems* 36 (2023), 8442–8486.

Brian DO Anderson, Manfred Deistler, Weitian Chen, and Alexander Filler. 2012. Autoregressive models of singular spectral matrices. *Automatica* 48, 11 (2012), 2843–2849.

Marta Banbura and Michele Modugno. 2014. Maximum Likelihood Estimation of Factor Models on Datasets with Arbitrary Pattern of Missing Data. *Journal of Applied Econometrics* 29, 1 (2014), 133–160. https://EconPapers.repec.org/RePEc:wly:japmet:v:29:y:2014:i:1:p:133-160

Boaz Barak and Ankur Moitra. 2016. Noisy Tensor Completion via the Sum-of-Squares Hierarchy. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016 (JMLR Workshop and Conference Proceedings, Vol. 49)*, Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir (Eds.). JMLR.org, 417–445. http://proceedings.mlr.press/v49/barak16.html

Matteo Barigozzi and Matteo Luciani. 2019. Quasi maximum likelihood estimation of non-stationary large approximate dynamic factor models. *arXiv preprint arXiv:1910.09841* (2019).

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

Juan Bógalo, Pilar Poncela, and Eva Senra. 2020. Understanding fluctuations through Multivariate Circulant Singular Spectrum Analysis. *arXiv preprint arXiv:2007.07561* (2020).

David Broomhead and Gregory King. 1986. *On the Qualitative Analysis of Experimental Dynamical Systems*. Vol. 11. Adam Hilger Bristol, Bristol (UK).

Changxiao Cai, Gen Li, H. Vincent Poor, and Yuxin Chen. 2019. Nonconvex Low-Rank Tensor Completion from Noisy Data. *Advances in neural information processing systems* 32 (2019), 1863–1874. https://proceedings.neurips.cc/paper/2019/file/a1519de5b5d44b31a01de013b9b51a80-Paper.pdf

François Chollet. 2015. keras. https://github.com/fchollet/keras.

Chandler Davis and William Morton Kahan. 1970. The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* 7, 1 (1970), 1–46.

Manfred Deistler and Brian Anderson. 2008. GENERALIZED LINEAR DYNAMIC FACTOR MODELS-ASTRUCTURE THEORY. In *Workshop on Inverse and Partial Information Problems*, Vol. 1. 50.

Catherine Doz, Domenico Giannone, and Lucrezia Reichlin. 2012. A Quasi–Maximum Likelihood Approach for Large, Approximate Dynamic Factor Models. *The Review of Economics and Statistics* 94, 4 (11 2012), 1014–1024.

Facebook. 2020. Prophet. https://facebook.github.io/prophet/. Online; accessed 25 February 2020.

Mario Forni, Marc Hallin, Marco Lippi, and Lucrezia Reichlin. 2000. The Generalized Dynamic-Factor Model: Identification and Estimation. *The Review of Economics and Statistics* 82, 4 (2000), 540–554. http://www.jstor.org/stable/2646650

Matan Gavish and David L Donoho. 2014. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory* 60, 8 (2014), 5040–5053.

M. Ghil, M. R. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, M. E. Mann, A. W. Robertson, A. Saunders, Y. Tian, F. Varadi, and P. Yiou. 2002. Advanced Spectral Method for Climatic Time Series. *Reviews of Geophysics* 40, 1 (2002), 3–1–3–41.

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

Loukas Grafakos. 2008. *Classical fourier analysis*. Vol. 2. Springer.

Marc Hallin and Roman Liška. 2007. Determining the Number of Factors in the General Dynamic Factor Model. *J. Amer. Statist. Assoc.* 102, 478 (2007), 603–617. http://www.jstor.org/stable/27639890

Hossein Hassani, Saeed Heravi, and Anatoly Zhigljavsky. 2013. Forecasting UK industrial production with multivariate singular spectrum analysis. *Journal of Forecasting* 32, 5 (2013), 395–408.

Hossein Hassani and Rahim Mahmoudvand. 2013. Multivariate singular spectrum analysis: A general view and new vector forecasting approach. *International Journal of Energy and Statistics* 1, 01 (2013), 55–83.

Hossein Hassani and Rahim Mahmoudvand. 2018. *Singular spectrum analysis: Using R*. Springer.

Rob J Hyndman and George Athanasopoulos. 2018. *Forecasting: principles and practice*. OTexts.

S Makridakis, E Spiliotis, and V Assimakopoulos. 2020. The M5 accuracy competition: Results, findings and conclusions. *Int J Forecast* (2020).

Vicente Oropeza and Mauricio Sacchi. 2011. Simultaneous seismic data denoising and reconstruction via multichannel singular spectrum analysis. *Geophysics* 76, 3 (2011), V25–V32.

Guy Plaut and Robert Vautard. 1994. Spells of Low-Frequency Oscillations and Weather Regimes in the Northern Hemisphere. *Journal of Atmospheric Sciences* 51, 2 (1994), 210 – 236.

Nikhil Rao, Hsiang-Fu Yu, Pradeep K Ravikumar, and Inderjit S Dhillon. 2015. Collaborative Filtering with Graph Information: Consistency and Scalable Methods. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). Curran Associates, Inc., 2107–2115.

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

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2019. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting* (2019).

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Rajat Sen, Hsiang-Fu Yu, and Inderjit S Dhillon. 2019. Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting. In *Advances in Neural Information Processing Systems*. 4838–4847.

Devavrat Shah and Christina Lee Yu. 2019. Iterative Collaborative Filtering for Sparse Noisy Tensor Estimation. In *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 41–45.

James H. Stock and Mark W. Watson. 2002. Forecasting Using Principal Components from a Large Number of Predictors. *J. Amer. Statist. Assoc.* 97, 460 (2002), 1167–1179. http://www.jstor.org/stable/3085839

Artur Trindade. 2014. UCI Machine Learning Repository - Individual Household Electric Power Consumption Data Set. (2014). https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014

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

Larry Wasserman. 2006. *All of nonparametric statistics*. Springer.

Per-Åke Wedin. 1972. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics* 12, 1 (1972), 99–111.

Kevin W Wilson, Bhiksha Raj, and Paris Smaragdis. 2008. Regularized non-negative matrix factorization with temporal dependencies for speech denoising. In *Ninth Annual Conference of the International Speech Communication Association*.

WRDS. 2021. The Trade and Quote (TAQ) database. (2021). https://wrds-www.wharton.upenn.edu/pages/support/data-overview/wrds-overview-taq/

Dong Xia, Ming Yuan, and Cun-Hui Zhang. 2018. Statistically Optimal and Computationally Efficient Low Rank Tensor Completion from Noisy Entries. arXiv:1711.04934 [stat.ML]

Jiaming Xu. 2017. Rates of convergence of spectral methods for graphon estimation. *arXiv preprint arXiv:1709.03183* (2017).

Christina Lee Yu. 2020. Tensor Estimation with Nearly Linear Samples. *arXiv preprint arXiv:2007.00736* (2020).

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.