Dynamic Nonlinear Matrix Completion for Time-Varying Data Imputation Supplementary Material

Anonymous

A Proof for Theorem 1

Theorem 1. *Suppose* $\boldsymbol{X}_t = [\boldsymbol{x}_{t-w+1}, \boldsymbol{x}_{t-w+2}, \ldots, \boldsymbol{x}_t]$ *is given by Assumption 1. Let* $\phi : \mathbb{R}^d \mapsto \mathbb{R}^{\binom{d+q}{q}}$ *be a q-order polynomial feature map. Let* $c_t = \max(\|\boldsymbol{z}_{t-w+1}\|, \ldots, \|\boldsymbol{z}_t\|)$. *Then with probability 1, there exists a matrix* $\hat{\boldsymbol{X}}_t$ *with rank at most* $\min\left\{\binom{r+\theta}{\theta}, d, w\right\}$ *such that* $\|\boldsymbol{X}_t - \hat{\boldsymbol{X}}_t\|_F \leq \frac{\gamma c_t (w-2)^{1.5}}{3}$ *and* $\operatorname{rank}(\phi(\hat{\boldsymbol{X}}_t)) \leq \min\left\{\binom{r+\theta q}{\theta q}, \binom{d+q}{q}, w\right\}$.

Proof. Without loss of generality, we assume that w is an odd number.

$$\begin{aligned}
&\|g_t(\boldsymbol{z}_t) - g_{t-\frac{w-1}{2}}(\boldsymbol{z}_t)\| \\
\leq&\|g_t(\boldsymbol{z}_t) - g_{t-1}(\boldsymbol{z}_t)\| + \|g_{t-1}(\boldsymbol{z}_t) - g_{t-2}(\boldsymbol{z}_t)\| + \cdots \\
&+ \|g_{t-\frac{w-1}{2}+1}(\boldsymbol{z}_t) - g_{t-\frac{w-1}{2}}(\boldsymbol{z}_t)\| \\
\leq&\tfrac{w-1}{2}\gamma\|\boldsymbol{z}_t\|.
\end{aligned} \tag{1}$$

Similarly, we

$$\|g_s(\boldsymbol{z}_s) - g_{t-\frac{w-1}{2}}(\boldsymbol{z}_s)\| \leq (\tfrac{w-1}{2} + s - t)\gamma\|\boldsymbol{z}_s\|, \tag{2}$$

where $s = t - \frac{w-1}{2}, \ldots, t$. We also have

$$\|g_s(\boldsymbol{z}_s) - g_{t-\frac{w-1}{2}}(\boldsymbol{z}_s)\| \leq (t - s - \tfrac{w-1}{2})\gamma\|\boldsymbol{z}_s\|, \tag{3}$$

where $s = t - w + 1, \ldots, t - \frac{w-1}{2} - 1$. Putting (2) and (3) together, we get

$$\begin{aligned}
&\sum_{s=t-w+1}^{t} \left\|g_s(\boldsymbol{z}_s) - g_{t-\frac{w-1}{2}}(\boldsymbol{z}_s)\right\|^2 \\
\leq&2\sum_{v=1}^{(w-3)/2} v^2\gamma^2 c_t^2 \\
=&\gamma^2 c_t^2 (w-1)(w-2)(w-3)/12 \\
\leq&\gamma^2 c_t^2 (w-2)^3/12,
\end{aligned} \tag{4}$$

where $c_t = \max(\|\boldsymbol{z}_{t-w+1}\|, \ldots, \|\boldsymbol{z}_t\|)$. Let

$$\hat{\boldsymbol{X}}_t = (\hat{\boldsymbol{x}}_{t-w+1}, \hat{\boldsymbol{x}}_{t-w+2}, \ldots, \hat{\boldsymbol{x}}_t),$$

where $\hat{\boldsymbol{x}}_s = g_{t-\frac{w-1}{2}}(\boldsymbol{z}_s)$, $s = t-w+1, \ldots, t$. According to Lemma 1 of (Fan, Zhang, and Udell 2020), with probability 1, we have

$$\operatorname{rank}(\hat{\boldsymbol{X}}_t) \leq \min\left\{\binom{r+\theta}{\theta}, d, w\right\}. \tag{5}$$

On the other hand, according to (4) and the definition of $\hat{\boldsymbol{X}}_t$, we have

$$\|\boldsymbol{X}_t - \hat{\boldsymbol{X}}_t\|_F \leq \frac{\gamma c_t (w-2)^{1.5}}{3}. \tag{6}$$

Now combining (5) and (6), we conclude that $\boldsymbol{X}_t$ can be approximated by a matrix $\hat{\boldsymbol{X}}_t$ with rank at most $\min\left\{\binom{r+\theta}{\theta}, d, w\right\}$ and the approximation error is at most $\gamma c_t (w-2)^{1.5}/3$. This finished the proof for the first part of the theorem.

Let ϕ be a q-order polynomial feature map. According to Lemma 1 of (Fan, Zhang, and Udell 2020), we have

$$\operatorname{rank}(\phi(\hat{\boldsymbol{X}}_t)) \leq \min\left\{\binom{r+\theta q}{\theta q}, \binom{d+q}{q}, w\right\}. \tag{7}$$

Then we conclude than $\boldsymbol{X}_t$ can be approximated by a matrix $\hat{\boldsymbol{X}}_t$ satisfying $\operatorname{rank}(\phi(\hat{\boldsymbol{X}}_t)) \leq \min\left\{\binom{r+\theta q}{\theta q}, \binom{d+q}{q}, w\right\}$. Then we finish the proof. □

B Gradient related to polynomial kernels

Denote by $\mathcal{L}_t$ the objective function in (5) of the main paper. We have

$$\frac{\partial \mathcal{L}_t}{\partial \boldsymbol{K}_t} = \frac{p}{2}\boldsymbol{K}_t^{\frac{p}{2}-1} = \frac{p}{2}\boldsymbol{V}_t \boldsymbol{\Lambda}_t^{\frac{p}{2}-1}\boldsymbol{V}_t^\top, \tag{8}$$

where $\boldsymbol{V}_t$ and $\operatorname{diag}(\boldsymbol{\Lambda}_t)$ are the eigenvectors and eigenvalues of $\boldsymbol{K}_t$ respectively. When $\boldsymbol{K}_t$ is computed by a polynomial kernel $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^\top \boldsymbol{x}_j + a)^q$, we have

$$\begin{aligned}
\frac{\partial \mathcal{L}_t}{\partial [\boldsymbol{x}_t]_{\bar{\omega}}} =&\sum_{i=1}^{w}\sum_{j=1}^{w} \frac{\partial \mathcal{L}_t}{\partial [\boldsymbol{K}_t]_{ij}} \frac{\partial [\boldsymbol{K}_t]_{ij}}{\partial [\boldsymbol{x}_t]_{\bar{\omega}}} \\
=&\left[2q\boldsymbol{X}_t \left(\boldsymbol{\alpha} \odot \left(\boldsymbol{X}_t^\top \boldsymbol{x}_t + a\right)^{\odot(q-1)}\right)\right]_{\bar{\omega}},
\end{aligned} \tag{9}$$

where $\boldsymbol{\alpha} = \left[\frac{\partial \mathcal{L}_t}{\partial \boldsymbol{K}_t}\right]_{:w}$. Invoking (8) into (9), we arrive at

$$\frac{\partial \mathcal{L}_t}{\partial [\boldsymbol{x}_t]_{\bar{\omega}}} \left[2q\boldsymbol{X}_t \left(\left(\frac{p}{2}\boldsymbol{V}_t \boldsymbol{\Lambda}_t^{\frac{p}{2}-1}\boldsymbol{v}_t\right) \odot \left(\boldsymbol{X}_t^\top \boldsymbol{x}_t + a\right)^{\odot(q-1)}\right)\right]_{\bar{\omega}}, \tag{10}$$

where $\boldsymbol{v}_t$ denotes the last columns of $\boldsymbol{V}_t^\top$.

C Proof for Theorem 2

Theorem 2. *Let $\boldsymbol{K}_t$ be the Gaussian kernel matrix with parameter σ. There exists a matrix $\tilde{\boldsymbol{K}}_t$ with rank at most $min\left\{\binom{r+\theta q}{\theta q}, \binom{d+q}{q}, w\right\}$ such that*

$$\|\boldsymbol{K}_t - \tilde{\boldsymbol{K}}_t\|_F \leq \frac{C_t \gamma w^2}{2\sigma^2} + \frac{C_t' w^2}{\sigma^{2(q+1)}(q+1)!}, \tag{11}$$

where C_t and C_t' are positive values relying on θ, q, and $\max(\|\boldsymbol{z}_{t-w+1}\|, \ldots, \|\boldsymbol{z}_t\|)$.

Proof. Let $\tilde{\boldsymbol{K}} = \boldsymbol{\Gamma} \odot \sum_{u=1}^{q} \sigma^{2u} u! \hat{\boldsymbol{K}}_j$, where $[\boldsymbol{\Gamma}]_{ij} = \exp\left(-\frac{\|\boldsymbol{x}_i\|^2 + \|\boldsymbol{x}_j\|^2 + 2a}{2\sigma^2}\right)$. According to Corollary 1 of (Fan, Zhang, and Udell 2020), we have

$$\left\|\hat{\boldsymbol{K}}_\sigma - \tilde{\boldsymbol{K}}\right\|_F \leq C_1, \tag{12}$$

where $C_1 = w^2 \exp\big(-\frac{\min_i \|\hat{\boldsymbol{x}}_i\|^2}{\sigma^2}\big) \frac{\max_i \|\hat{\boldsymbol{x}}_i\|^q}{\sigma^{2(q+1)}(q+1)!}$ and $\text{rank}(\tilde{\boldsymbol{K}}) \leq \binom{r+\theta q}{\theta q}$ provided that w/r is large enough. On the other hand, we have

$$\begin{aligned}
&\|\boldsymbol{K}_\sigma - \hat{\boldsymbol{K}}_\sigma\|_F^2 \\
=&\frac{1}{4\sigma^4} \sum_{ij} \left(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 - \|\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{x}}_j\|^2\right)^2 \\
\leq&\frac{1}{4\sigma^4} \sum_{ij} C_{ij} \left(\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\| + \|\boldsymbol{x}_j - \hat{\boldsymbol{x}}_j\|\right)^2 \\
\leq&\frac{\max_{ij} C_{ij}}{4\sigma^4} \sum_{ij} \left(\frac{w-1}{2}\gamma\|\boldsymbol{z}_i\| + \frac{w-1}{2}\gamma\|\boldsymbol{z}_j\|\right)^2 \\
\leq&\frac{\gamma^2 w^2 (w-1)^2 \max_{ij} C_{ij} \max_i \|\boldsymbol{z}_i\|^2}{4\sigma^4},
\end{aligned} \tag{13}$$

where $C_{ij} = 2\max(\|\boldsymbol{x}_i\|, \|\boldsymbol{x}_j\|, \|\hat{\boldsymbol{x}}_i\|, \|\hat{\boldsymbol{x}}_j\|)$. Combining (12) with (13), we obtain

$$\begin{aligned}
&\|\boldsymbol{K}_\sigma - \tilde{\boldsymbol{K}}\|_F \\
\leq&\|\boldsymbol{K}_\sigma - \hat{\boldsymbol{K}}_\sigma\|_F + \left\|\hat{\boldsymbol{K}}_\sigma - \tilde{\boldsymbol{K}}\right\|_F \\
\leq&\frac{\gamma w^2 C_x C_z}{2\sigma^2} + \frac{w^2 C_x' C_x^q}{\sigma^{2(q+1)}(q+1)!},
\end{aligned} \tag{14}$$

where $C_x' = \exp\big(-\frac{\min_i \|\hat{\boldsymbol{x}}_i\|^2}{\sigma^2}\big)$, $C_x = \sqrt{2\max(\|\boldsymbol{x}_i\|, \|\boldsymbol{x}_j\|, \|\hat{\boldsymbol{x}}_i\|, \|\hat{\boldsymbol{x}}_j\|)}$, and $C_z = \max_i \|\boldsymbol{z}_i\|$. SInce g_t is polynomial, there exists a constant C_θ large enough such that $\max_i \|\hat{\boldsymbol{x}}_i\| \leq C_\theta \max_i \|\boldsymbol{z}_i\|$, where $i = t-w+1, \ldots, t$. Letting $C_t = \sqrt{2C_\theta}\left(\max_i \|\boldsymbol{z}_i\|\right)^{3/2}$ and $C_t' = \exp(-\frac{C_\theta^2 (\max_i \|\boldsymbol{z}_i\|)^2}{\sigma^2})\left(2C_\theta \max_i \|\boldsymbol{z}_i\|\right)^{q/2}$. It follows from (15) that

$$\|\boldsymbol{K}_\sigma - \tilde{\boldsymbol{K}}\|_F \leq \frac{C_t \gamma w^2}{2\sigma^2} + \frac{C_t' w^2}{\sigma^{2(q+1)}(q+1)!}. \tag{15}$$

This finished the proof. □

D Rank-one modification for fast EVD

Here we show how to perform rank-one modification (Brand 2006) twice to compute the eigenvalue decomposition of $\boldsymbol{K}_t$. Let $\boldsymbol{e}_w = [0, 0, \ldots, 0, 1]^\top$ and $\tilde{\boldsymbol{k}}' = [\boldsymbol{k}'^\top\ k(\boldsymbol{x}_t, \boldsymbol{x}_t)]^\top$. The method is detailed in Algorithm 1.

Algorithm 1: Rank-one modification for fast EVD of $\boldsymbol{K}_t$

Input: $\boldsymbol{V}_{t-1}', \boldsymbol{\Lambda}_{t-1}', \boldsymbol{e}_w, \boldsymbol{k}', \tilde{\boldsymbol{k}}'$

1: $\boldsymbol{U} \leftarrow \boldsymbol{V}_{t-1}'$, $\boldsymbol{V} \leftarrow [\boldsymbol{V}_{t-1}'^\top\ \boldsymbol{0}]^\top$, $\boldsymbol{a} \leftarrow \bar{\boldsymbol{k}}'$, $\boldsymbol{b} \leftarrow \boldsymbol{e}_w$

2: $\boldsymbol{m} = \boldsymbol{U}^\top \boldsymbol{a}$, $\boldsymbol{p} = \boldsymbol{a} - \boldsymbol{U}\boldsymbol{m}$, $\bar{\boldsymbol{p}} = \boldsymbol{p}/\|\boldsymbol{p}\|$.

3: $\boldsymbol{n} = \boldsymbol{V}^\top \boldsymbol{b}$, $\boldsymbol{q} = \boldsymbol{b} - \boldsymbol{V}\boldsymbol{n}$, $\bar{\boldsymbol{q}} = \boldsymbol{q}/\|\boldsymbol{q}\|$.

4: $\boldsymbol{W} := \begin{bmatrix} \boldsymbol{\Lambda}_{t-1}' & \boldsymbol{0} \\ \boldsymbol{0} & 0 \end{bmatrix} + \begin{bmatrix} \boldsymbol{m} \\ \|\boldsymbol{p}\| \end{bmatrix} \begin{bmatrix} \boldsymbol{n} \\ \|\boldsymbol{q}\| \end{bmatrix}^\top$.

5: $\boldsymbol{W} = \boldsymbol{U}'\boldsymbol{\Sigma}'\boldsymbol{V}'^\top$.

6: $\bar{\boldsymbol{U}} \leftarrow \boldsymbol{U}\ \bar{\boldsymbol{p}}]\boldsymbol{U}'$, $\bar{\boldsymbol{V}} \leftarrow \boldsymbol{V}\ \bar{\boldsymbol{q}}]\boldsymbol{V}'$.

7: $\boldsymbol{U} \leftarrow [\bar{\boldsymbol{U}}\top\ \boldsymbol{0}]^\top$, $\boldsymbol{V} \leftarrow \bar{\boldsymbol{V}}$, $\boldsymbol{a} \leftarrow \boldsymbol{e}_w$, $\boldsymbol{b} \leftarrow \tilde{\boldsymbol{k}}'$

8: $\boldsymbol{m} = \boldsymbol{U}^\top \boldsymbol{a}$, $\boldsymbol{p} = \boldsymbol{a} - \boldsymbol{U}\boldsymbol{m}$, $\bar{\boldsymbol{p}} = \boldsymbol{p}/\|\boldsymbol{p}\|$.

9: $\boldsymbol{n} = \boldsymbol{V}^\top \boldsymbol{b}$, $\boldsymbol{q} = \boldsymbol{b} - \boldsymbol{V}\boldsymbol{n}$, $\bar{\boldsymbol{q}} = \boldsymbol{q}/\|\boldsymbol{q}\|$.

10: $\boldsymbol{W} := \begin{bmatrix} \boldsymbol{\Sigma}' & \boldsymbol{0} \\ \boldsymbol{0} & 0 \end{bmatrix} + \begin{bmatrix} \boldsymbol{m} \\ \|\boldsymbol{p}\| \end{bmatrix} \begin{bmatrix} \boldsymbol{n} \\ \|\boldsymbol{q}\| \end{bmatrix}^\top$.

11: $\boldsymbol{W} = \boldsymbol{U}'\boldsymbol{\Sigma}'\boldsymbol{V}'^\top$.

12: $\boldsymbol{U}_t \leftarrow [\boldsymbol{U}\ \bar{\boldsymbol{p}}]\boldsymbol{U}'$, $\boldsymbol{\Lambda}_t \leftarrow \boldsymbol{\Sigma}'$, $\boldsymbol{V}_t \leftarrow [\boldsymbol{V}\ \bar{\boldsymbol{q}}]\boldsymbol{V}'$.

Output: $\boldsymbol{K}_t \approx \boldsymbol{V}_t \boldsymbol{\Lambda}_t \boldsymbol{V}_t^\top$.

E Proof for Theorem 3

Lemma 1. *Suppose $\boldsymbol{X}^*$ is partially observed (uniformly at random) and the number of observed entries in each column of $\boldsymbol{X}^*$ is at least r. Suppose the number of observed entries (denoted by Ω) of $\boldsymbol{X}^*$ is sufficiently large such that the observed entries can only be fitted by a $\hat{\theta}$-order polynomial function on some latent variable $\hat{\boldsymbol{z}} \in \mathbb{R}^{\hat{r}}$. Let $\mathcal{S} = \{\boldsymbol{X} \in \mathbb{R}^{d\times n} : [\boldsymbol{X}]_\Omega = [\boldsymbol{X}^*]_\Omega,\ \text{rank}(\boldsymbol{K}) \leq R < n,\ [\boldsymbol{K}]_{ij} = (\boldsymbol{x}_i^\top \boldsymbol{x}_j + b)^q,\ i,j = 1, \ldots, n,\ b \in \mathbb{R}^+,\ q \in \mathbb{Z}^+\}$, where Then the covering numbers of $\mathcal{S}$ satisfy*

$$\mathcal{N}(\mathcal{S}, \|\cdot\|_F, \epsilon) \leq \left(\frac{3\beta}{\epsilon}\right)^{ab}$$

Proof. The assumption indicates that there exist r_j, θ_j, and s such that

$$\sum_{j=1}^{s} \binom{r_j + \theta_j q}{\theta_j q} \leq R, \tag{16}$$

and the columns of $\boldsymbol{X}$ can be fitted by s polynomial functions

$$f_j : \mathbb{R}^{r_j} \to \mathbb{R}^d,\ \ j = 1, 2, \ldots, s.$$

The difficulty is that we do not know what order and how many polynomials are fitted by the columns of $\boldsymbol{X}$. We consider the following special cases.

Case 1: highest-order polynomials. The columns of $\boldsymbol{X}$ lie on polynomials with the possibly highest order., which means $r_1 = \cdots = r_s = 1$. Without loss of generality, let $\theta_1 = \cdots = \theta_s = \theta^+$. We have

$$\theta^+ = \max\left\{\hat{\theta} \in \mathbb{Z}^+ : s\binom{1+\hat{\theta}q}{\hat{\theta}q} \le R\right\} = \frac{R}{s} - 1.$$

Then the number of parameters (polynomial coefficients and latent variables) required to determine $\boldsymbol{X}$ is

$$\pi_1 = n + s\binom{1+\theta^+}{\theta^+}d = n + Rd.$$

Case 2: linear functions. The columns of $\boldsymbol{X}$ lie on lines, which means $\theta_1 = \cdots = \theta_s = 1$. Without loss of generality, let $r_1 = \cdots = r_s = r^+$. We have

$$r^+ = \max\left\{\hat{r} \in \mathbb{Z}^+ : s\binom{\hat{r}+q}{q} \le R\right\}.$$

Since $\binom{\hat{r}+q}{q} \approx \frac{(\hat{r}+q)^q}{q!}$, we get

$$r^+ \approx \left\lceil \left(\frac{Rq!}{s}\right)^{1/q} - q \right\rceil. \tag{17}$$

Here the minimum s is 1 and the maximum s is $R/(q+1)$. If R is sufficiently small, we obtain

$$r^+ < \binom{r+\theta}{\theta},$$

which contradicts with the fact $d \ge \operatorname{rank}(\boldsymbol{X}) \ge \binom{r+\theta}{\theta}$. Therefore, Case 2 will not happen if R is sufficiently small, i.e.,

$$R < s\binom{\binom{r+\theta}{\theta}+q}{q}, \tag{18}$$

or if $R < \binom{\binom{r+\theta}{\theta}+q}{q}$ more strictly.

Case 3: low-order polynomials. The columns of $\boldsymbol{X}$ lie on polynomials with order at least 2. Without loss of generality, we assume $\theta_1 = \cdots = \theta_s = \theta^\star \ge 2$ and $r_1 = \cdots = r_s = r^\star$. To ensure that (16) and $s\binom{r^\star+\theta^\star}{\theta^\star} \ge \operatorname{rank}(\boldsymbol{X})$ hold simultaneously and $r^\star$ is sufficiently large, we get

$$\begin{aligned} r^\star = \max\Big\{\hat{r} \in \mathbb{Z}^+ :\ & s\binom{\hat{r}+\theta^\star q}{\theta^\star q} \le R,\ \theta^\star \in \mathbb{Z}^+/\{1\}, \\ & s\binom{\hat{r}+\theta^\star}{\theta^\star} \ge \binom{r+\theta}{\theta}\Big\}. \end{aligned} \tag{19}$$

Let $\psi(v, C)$ be the root of equation $\binom{u+v}{v} = C$ with variable u. We have

$$\begin{aligned} r^\star = \max\Big\{\hat{r} \in \mathbb{Z}^+ :\ & \psi(\theta^\star, \tbinom{r+\theta}{\theta}/s) \le \hat{r} \le \psi(\theta^\star q, R/s), \\ & \theta^\star \in \mathbb{Z}^+/\{1\}\Big\}. \end{aligned} \tag{20}$$

Then the number of parameters required to determine $\boldsymbol{X}$ is

$$\pi_3 = \max_{s \in \mathbb{Z}^+} nr^\star + s\binom{r^\star+\theta^\star}{\theta^\star}d.$$

Since $n \gg d$ and $\binom{r^\star+\theta^\star}{\theta^\star} \ll R$, it suffices to let $s = 1$ and we arrive at

$$\pi_3 = nr^\star + \binom{r^\star+\theta^\star}{\theta^\star}d.$$

It is obvious that

$$\pi_1 < \pi_3.$$

□

Lemma 2. *Let $\mathcal{S}$ be the set of matrices $\boldsymbol{X} \in \mathbb{R}^{d\times n}$ whose columns are given by a polynomial function of order at most θ on a latent variable $\boldsymbol{z} \in \mathbb{R}^r$, where $\|\boldsymbol{X}\|_F \le \delta$. Then there exists a constant c such that the covering numbers of $\mathcal{S}$ with respect to Frobenius norm satisfy*

$$\mathcal{N}(\mathcal{S}, \|\cdot\|_F, \epsilon) \le \left(\frac{c\delta}{\epsilon}\right)^{rn+d\binom{r+\theta}{\theta}}.$$

Proof. Suppose $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathcal{S}$. It means $\boldsymbol{X}_j = g_j(\boldsymbol{Z}_j) = \boldsymbol{P}_j\tilde{\boldsymbol{Z}}_j$, where $\tilde{\boldsymbol{Z}}_j \in \mathbb{R}^{\binom{r^\star+\theta^\star}{\theta^\star}\times n}$ and $\boldsymbol{P}_j \in \mathbb{R}^{d\times\binom{r^\star+\theta^\star}{\theta^\star}}$ denote the binomial terms and the coefficients respectively, $j = 1, 2$. Suppose g_j is L-Lipschitz continuous, $\|\boldsymbol{Z}_j\|_F \le \delta_1$, $\|\tilde{\boldsymbol{Z}}_j\|_F \le \delta_2$, and $\|\boldsymbol{P}_j\|_F \le \delta_3$, $j = 1, 2$. We have

$$\begin{aligned} &\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_F = \|g_1(\boldsymbol{Z}_1) - g_2(\boldsymbol{Z}_2)\|_F \\ \le&\|g_1(\boldsymbol{Z}_1) - g_1(\boldsymbol{Z}_2)\|_F + \|g_1(\boldsymbol{Z}_2) - g_2(\boldsymbol{Z}_2)\|_F \\ \le& L\|\boldsymbol{Z}_1 - \boldsymbol{Z}_2\| + \|\tilde{\boldsymbol{Z}}_2\|_F\|\boldsymbol{P}_1 - \boldsymbol{P}_2\|_F. \end{aligned} \tag{21}$$

Suppose $\|\boldsymbol{Z}_1 - \boldsymbol{Z}_2\| \le \frac{\epsilon}{2L}$ and $\|\boldsymbol{P}_1 - \boldsymbol{P}_2\|_F \le \frac{\epsilon}{2\|\tilde{\boldsymbol{Z}}_2\|_F}$. It follows that

$$\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_F \le \epsilon. \tag{22}$$

Then we can bound the ϵ-covering number of $\mathcal{S}$ as

$$\begin{aligned} \mathcal{N}(\mathcal{S}_{ab}, \|\cdot\|_F, \epsilon) &\le \left(\frac{6L\delta_1}{\epsilon}\right)^{r^\star n}\left(\frac{6\delta_2\delta_3}{\epsilon}\right)^{d\binom{r^\star+\theta^\star}{\theta^\star}} \\ &\le \left(\frac{6\max(L\delta_1, \delta_2\delta_3)}{\epsilon}\right)^{r^\star n+d\binom{r^\star+\theta^\star}{\theta^\star}}. \end{aligned} \tag{23}$$

Although L and $\{\delta_i\}_{i=1}^3$ are unknown, they are related to $\|\boldsymbol{X}\|_F$. We can bound $6\max(L\delta_1, \delta_2\delta_3)$ by $c\|\boldsymbol{X}\|_F$, where c is a sufficiently large constant. Now we get

$$\mathcal{N}(\mathcal{S}, \|\cdot\|_F, \epsilon) \le \left(\frac{c\delta}{\epsilon}\right)^{rn+d\binom{r+\theta}{\theta}}.$$

□

We give the following lemma.

Lemma 3 (Hoeffding inequality for sampling without replacement (Serfling 1974))**.** *Let $X_1, X_2, \ldots, Xs$ be a set of samples taken without replacement from a distribution $\{x_1, x_2, \ldots, x_N\}$ of mean u and variance σ^2. Denote $a = \min_i x_i$ and $b = \max_i x_i$. Then*

$$P\left[\Big|\frac{1}{s}\sum_{i=1}^{s} X_i - u\Big| \ge t\right] \le 2\exp\left(-\frac{2st^2}{(1-(s-1)/N)(b-a)^2}\right).$$

Let $\hat{\mathcal{L}}(\boldsymbol{X}) := \frac{1}{|\Omega|}\|\mathcal{P}_\Omega(\boldsymbol{Y}-\boldsymbol{X})\|_F^2$ and $\mathcal{L}(\boldsymbol{X}) := \frac{1}{N}\|\boldsymbol{Y}-\boldsymbol{X}\|_F^2$. where $N = dn$. Suppose $\max\{\|\boldsymbol{Y}\|_\infty, \|\boldsymbol{X}\|_\infty\} \leq \beta$. According to Lemma 3, we have

$$P\left[|\hat{\mathcal{L}} - \mathcal{L}| \geq t\right] \leq 2\exp\left(-\frac{2|\Omega|t^2}{(1-(|\Omega|-1)/n^d)\eta^2}\right),$$

where $\eta = 4\beta^2$. Using union bound for all $\bar{\boldsymbol{X}} \in \mathcal{S}$ (defined in Lemma 2), we obtain

$$\begin{aligned}&P\left[\sup_{\bar{\boldsymbol{X}}\in\mathcal{S}} |\hat{\mathcal{L}}(\bar{\boldsymbol{X}}) - \mathcal{L}(\bar{\boldsymbol{X}})| \geq t\right]\\ \leq& 2|\mathcal{S}|\exp\left(-\frac{2|\Omega|t^2}{(1-(|\Omega|-1)/N)\eta^2}\right).\end{aligned}$$

Equivalently, with probability at least $1-2N^{-1}$, we have

$$\begin{aligned}\sup_{\bar{\boldsymbol{X}}\in\mathcal{S}} |\hat{\mathcal{L}}(\bar{\boldsymbol{X}}) - \mathcal{L}(\bar{\boldsymbol{X}})| \leq& \sqrt{\frac{\eta^2\log(|\mathcal{S}|N)}{2}\left(\frac{1}{|\Omega|} - \frac{1}{N} + \frac{1}{N|\Omega|}\right)}\\ \leq& \sqrt{\frac{\eta^2\log(|\mathcal{S}|N)}{2|\Omega|}} \triangleq \Upsilon.\end{aligned}$$

Since $|\sqrt{u}-\sqrt{v}| \leq \sqrt{|u-v|}$ holds for any non-negative u and v, we have

$$\sup_{\bar{\boldsymbol{X}}\in\mathcal{S}}\left|\sqrt{\hat{\mathcal{L}}(\bar{\boldsymbol{X}})} - \sqrt{\mathcal{L}(\bar{\boldsymbol{X}})}\right| \leq \sqrt{\Upsilon}.$$

As $\epsilon \geq \|\boldsymbol{X}-\bar{\boldsymbol{X}}\|_F \geq \|\mathcal{P}(\boldsymbol{X}-\bar{\boldsymbol{X}})\|_F$, we have

$$\begin{aligned}&\left|\sqrt{\mathcal{L}(\boldsymbol{X})} - \sqrt{\mathcal{L}(\bar{\boldsymbol{X}})}\right|\\ =& \frac{1}{\sqrt{N}}\left|\|\boldsymbol{Y}-\boldsymbol{X}\|_F - \|\boldsymbol{Y}-\bar{\boldsymbol{X}}\|_F\right| \leq \frac{\epsilon}{\sqrt{N}}\end{aligned}$$

and

$$\begin{aligned}&\left|\sqrt{\hat{\mathcal{L}}(\boldsymbol{X})} - \sqrt{\hat{\mathcal{L}}(\bar{\boldsymbol{X}})}\right|\\ =& \frac{1}{\sqrt{|\Omega|}}\left|\|\mathcal{P}_\Omega(\boldsymbol{Y}-\boldsymbol{X})\|_F - \|\mathcal{P}_\Omega(\boldsymbol{Y}-\bar{\boldsymbol{X}})\|_F\right| \leq \frac{\epsilon}{\sqrt{|\Omega|}}.\end{aligned}$$

It follows that

$$\begin{aligned}&\sup_{\boldsymbol{X}\in\mathcal{S}}\left|\sqrt{\hat{\mathcal{L}}(\boldsymbol{X})} - \sqrt{\mathcal{L}(\boldsymbol{X})}\right|\\ \leq& \sup_{\boldsymbol{X}\in\mathcal{S}}\left|\sqrt{\hat{\mathcal{L}}(\boldsymbol{X})} - \sqrt{\hat{\mathcal{L}}(\bar{\boldsymbol{X}})}\right| + \left|\sqrt{\hat{\mathcal{L}}(\bar{\boldsymbol{X}})} - \sqrt{\mathcal{L}(\boldsymbol{l}\bar{\boldsymbol{X}})}\right|\\ &+ \left|\sqrt{\mathcal{L}(\bar{\boldsymbol{X}})} - \sqrt{\mathcal{L}(\boldsymbol{X})}\right|\\ \leq& \frac{\epsilon}{\sqrt{|\Omega|}} + \sqrt{\Upsilon} + \frac{\epsilon}{\sqrt{N}} \leq \frac{2\epsilon}{\sqrt{|\Omega|}} + \sqrt{\Upsilon}.\end{aligned}$$

Now let $\epsilon = \beta$, we arrive at

$$\begin{aligned}&\left|\sqrt{\hat{\mathcal{L}}(\boldsymbol{X})} - \sqrt{\mathcal{L}(\boldsymbol{X})}\right|\\ \leq& \frac{2\beta}{\sqrt{|\Omega|}} + \beta\left(\frac{8\log N + 8\left(r^\star n + d\binom{r^\star+\theta^\star}{\theta^\star}\right)\log\frac{c\delta}{\beta}}{|\Omega|}\right)^{1/4}\\ \leq& c'\beta\left(\frac{\left(r^\star n + d\binom{r^\star+\theta^\star}{\theta^\star}\right)\log\frac{\delta}{\beta}}{|\Omega|}\right)^{1/4},\end{aligned} \tag{24}$$

where c' is a constant. This finished the proof.

It follows that

$$r^\star = \max\left\{\hat{r} \in \mathbb{Z}^+ :\ r_l \leq \hat{r} \leq r_u\right\}, \tag{25}$$

where

$$\begin{aligned}r_l &= \left(\frac{\binom{r+\theta}{\theta}\theta^\star!}{s}\right)^{1/\theta^\star} - \theta^\star,\\ r_u &= \left(\frac{R(\theta^\star q)!}{s}\right)^{1/(\theta^\star q)} - \theta^\star q.\end{aligned}$$

F Proof for Corollary 1

Proof.

$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\| \leq \|\boldsymbol{X} - \bar{\boldsymbol{X}}\| + \|\boldsymbol{E}\| \tag{26}$$

Denote $\Delta = c\beta\left(\frac{\left(r^\star n + d\binom{r^\star+\theta^\star}{\theta^\star}\right)\log\frac{\delta}{\beta}}{|\Omega|}\right)^{1/4}$.

$$\begin{aligned}&\frac{1}{\sqrt{dn}}\|\boldsymbol{X}_t - \bar{\boldsymbol{X}}_t\|_F\\ \leq& \frac{1}{\sqrt{|\Omega|}}\|\mathcal{P}_\Omega(\boldsymbol{X}_t - \bar{\boldsymbol{X}}_t)\|_F + \Delta\\ \leq& \frac{1}{\sqrt{|\Omega|}}\|\mathcal{P}_\Omega(\boldsymbol{X}_t - \hat{\boldsymbol{X}}_t)\|_F + \frac{1}{\sqrt{|\Omega|}}\|\mathcal{P}_\Omega(\boldsymbol{E}_t)\|_F + \Delta\\ \leq& \frac{1}{\sqrt{|\Omega|}}\|\mathcal{P}_\Omega(\boldsymbol{E}_t)\|_F + \Delta.\end{aligned} \tag{27}$$

It follows that

$$\begin{aligned}&\frac{1}{\sqrt{dn}}\|\boldsymbol{X}_t - \hat{\boldsymbol{X}}_t\| \leq \frac{1}{\sqrt{dn}}\|\boldsymbol{X}_t - \bar{\boldsymbol{X}}_t\| + \frac{1}{\sqrt{dn}}\|\boldsymbol{E}_t\|\\ \leq& \frac{1}{\sqrt{|\Omega|}}\|\mathcal{P}_\Omega(\boldsymbol{E}_t)\|_F + \Delta + \frac{1}{\sqrt{dn}}\|\boldsymbol{E}_t\|\\ \leq& \frac{2\varepsilon_t}{\sqrt{|\Omega|}} + \Delta.\end{aligned} \tag{28}$$

Or equivalently, we have

$$
\begin{aligned}
&\frac{1}{|\bar{\Omega}|}\sum_{(i,j)\in\bar{\Omega}}\left([\boldsymbol{X}_t]_{ij}-[\hat{\boldsymbol{X}}_t]_{ij}\right)^2\\
\leq&\frac{dn}{dn-|\Omega|}\left(\frac{2\varepsilon_t}{\sqrt{|\Omega|}}+\Delta\right)^2\\
=&\frac{dn}{dn-|\Omega|}\left(\frac{2\varepsilon_t}{\sqrt{|\Omega|}}+c\beta\left(\frac{\left(r^\star n+d\binom{r^\star+\theta^\star}{\theta^\star}\right)\log\frac{\delta}{\beta}}{|\Omega|}\right)^{1/4}\right)^2\\
\leq&\frac{dn}{dn-|\Omega|}\left(\frac{8\varepsilon_t^2}{|\Omega|}+c'\beta^2\left(\frac{\left(r^\star n+d\binom{r^\star+\theta^\star}{\theta^\star}\right)\log\frac{\delta}{\beta}}{|\Omega|}\right)^{1/2}\right),
\end{aligned}
$$

where $c'=2c$ is a constant. This finished the proof.

$\square$

G More about the experiments

Data preprocessing Since the variables in the SML2010 indoor temperature dataset and Air Quality dataset have very different scales, we rescale all variables by their standard deviations.

Parameter setting of D-NLMC For the synthetic data, we set $w=20$, $R=15$, and $\mu=1$. For the Chlorine level dataset, we set $w=100$, $R=50$, and $\mu=1$. For the SML2010 indoor temperature dataset, we set $w=50$, $R=25$, and $\mu=1$. For the Air Quality dataset, we set $w=50$, $R=25$, and $\mu=3$. Note that in OL-LRMC, we used the same w as D-NLMC.