Bayesian Robust Hankel Matrix Completion with Uncertainty Modeling for Synchrophasor Data Recovery

MING YI and MENG WANG*, Dept. of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, USA
EVANGELOS FARANTATOS and TAPAS BARIK, Electric Power Research Institute, USA

Synchrophasor data suffer from quality issues like missing and bad data. Exploiting the low-rankness of the Hankel matrix of the synchrophasor data, this paper formulates the data recovery problem as a robust low-rank Hankel matrix completion problem and proposes a Bayesian data recovery method that estimates the posterior distribution of synchrophasor data from partial observations. In contrast to the deterministic approaches, our proposed Bayesian method provides an uncertainty index to evaluate the confidence of each estimation. To the best of our knowledge, this is the first method that provides confidence measure for synchrophasor data recovery. Numerical experiments on synthetic data and recorded synchrophasor data demonstrate that our method outperforms existing low-rank matrix completion methods.

Additional Key Words and Phrases: PMU data recovery, robust matrix completion, Bayesian matrix completion, uncertainty modeling, Hankel matrix

ACM Reference Format:
Ming Yi, Meng Wang, Evangelos Farantatos, and Tapas Barik. 2022. Bayesian Robust Hankel Matrix Completion with Uncertainty Modeling for Synchrophasor Data Recovery. 2, 1 (February 2022), 18 pages.

1 INTRODUCTION

PHASOR Measurement Units (PMU) provide synchronized phasor measurements of various locations of the power system, and these data can be used for state estimation [Aminifar et al. 2013; Dobakhshari et al. 2020; Zhao et al. 2015a], post-disturbance analysis [Bhui and Senroy 2016; Guo and Milanović 2015] and system identification [Kamwa and Gerin-Lajoie 2000; Zhou et al. 2006]. Synchrophasor data have quality issues such as missing and bad data, including false data injection attacks from malicious intruders [Liu et al. 2011]. Such quality issues prevent synchrophasor data from being employed for real-time control.

A variety of methods have been developed for PMU missing data recovery such as training deep neural networks [James et al. 2019, 2018; Ren and Xu 2019], designing a dynamic state estimator based on Kalman filter [Jones et al. 2014; Zhou et al. 2014], filling the missing data based on the inference of a dynamic model [Foggo and Yu 2021], formulating it as a low-rank matrix completion problem [Gao et al. 2016b; Genes et al. 2018; Liao et al. 2018; Zhang et al. 2018] and the more general tensor completion problem [Osipov and Chow 2020]. Bad data are corrected by methods like hypothesis testing [Huang et al. 2018; Kosut et al. 2011; Mestav et al. 2019; Mestav and Tong 2020], exploiting spatio-temporal similarities [Wu and Xie 2016], spatial clustering [Wang et al. 2019], independent component analysis [Esmalifalak et al. 2015], and low-rank approaches [Gao et al. 2016a; Hao et al. 2018; Zhang and Wang 2018]. Low-rank methods have the unique advantages among all these efforts: (1) no need of power system topology and line parameters as required by state estimators, (2) no need of training data as required by neural-network-based approaches, and (3) more computationally efficient than tensor approaches. Moreover, synchrophasor data have the special property that not only the data matrix but also the corresponding Hankel matrix is low-rank, and [Hao et al. 2018; Zhang et al. 2018; Zhang and Wang 2019] have leveraged this low-rank Hankel property to enhance the data recovery performance. One major advantage of low-rank Hankel methods is the ability to recover simultaneous and consecutive data issues across all channels, while the conventional low-rank methods fail in this extreme scenario.

The critical limitation of the above methods is the lack of a confidence measure of the returned estimation. Although low-rank methods have theoretical guarantees that the recovery is accurate if the loss/error percentage is less than a threshold, such bound generally underestimates the methods' capabilities and thus is not practical. Only a few works consider the uncertainty modeling for matrix completion problem. Ref. [Zhao and Udell 2020] quantifies the uncertainty based on Gaussian copula. Ref. [Chen et al. 2019] builds a confidence interval for noisy matrix completion. Both works require strong assumptions and consider missing data only. [Babacan et al. 2012] develops a Bayesian approach to recover low-rank matrices. However, [Babacan et al. 2012] develops two separate approaches to handle missing and bad data, respectively, and no confidence measure is provided.

This paper develops a Bayesian low-rank Hankel matrix recovery method to recover missing and bad data. The method also returns an uncertainty index for each recovered value such that the operator can evaluate the confidence of the recovery. Specifically, given the prior distribution of the data, the method computes the posterior distribution using partial observations that contain noise and errors. The mean of the posterior is employed to estimate each data point, and the corresponding variance is viewed as the uncertainty index. The advantage of our method over the existing deterministic approaches [Zhang et al. 2018; Zhang and Wang 2019] on low-rank Hankel matrix recovery are threefold. First, our method provides the uncertainty index to evaluate the confidence of each estimation. Second, our method outperforms the deterministic approaches in handling corrupted data. Third, our method is more robust to parameter selection. For instance, the estimated rank of the Hankel matrix can be set to

*corresponding author. This work was supported in part by the NSF grant # 1932196, AFOSR FA9550-20-1-0122, ARO W911NF-21-1-0255, and the Electric Power Research Institute (EPRI).

Authors' addresses: Ming Yi, yim3@rpi.edu; Meng Wang, wangm7@rpi.edu, Dept. of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA, 12180; Evangelos Farantatos, efarantatos@epri.com; Tapas Barik, TBarik@epri.com, Electric Power Research Institute, Palo Alto, CA, USA, 94304.

XXXX-XXXX/2022/2-ART $15.00
https://doi.org/

be much larger than the actual value initially, and our method can estimate the actual value from the data by iterative pruning. Moreover, our method significantly outperforms conventional Bayesian matrix completion approaches like [Babacan et al. 2012] , because the latter perform poorly on simultaneous data losses or corruptions across all channels. In addition, [Babacan et al. 2012] handles missing and bad data separately, while our method can recover missing data and correct bad data at the same time.

The rest of the paper is organized as follows. The low-rank Hankel property of synchrophasor data and the problem formulation are introduced in Section II. Our proposed approach is presented in Section III. The numerical experiments are reported in Section IV, and Section V concludes the paper. Technical details of our method are described in the supplementary material.

2 PROBLEM FORMULATION

Let a matrix Y contain the ground-truth measurements of m channels in n time instants,

$$Y = [\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_n] \in \mathbb{R}^{m\times n}, \tag{1}$$

where $\boldsymbol{y}_i \in \mathbb{R}^m$ contains the data of m channels at ith time instant. Let $\boldsymbol{E}$ denote the additive bad data and $\boldsymbol{N}$ denote the additive noise data. $\boldsymbol{E}$ is a sparse matrix, and the values in $\boldsymbol{E}$ can be arbitrarily large. $\boldsymbol{N}$ is a dense noise matrix and the values in $\boldsymbol{N}$ are small. Let matrix $\boldsymbol{Y}^o \in \mathbb{R}^{m\times n}$ denote the observed data with each entry satisfying

$$Y^o_{i,j} = Y_{i,j} + E_{i,j} + N_{i,j} \quad (i,j) \in \Omega, \tag{2}$$

where the set Ω contains the indices of the observed entries in $\boldsymbol{Y}^o$.

The objective of robust matrix completion is to recover $\boldsymbol{Y}$ from partial observations $Y^o_{i,j}$ that contain missing data, bad data and noise. Moreover, this paper wants to provide an uncertainty index for the confidence evaluation of each estimation $Y_{i,j}$.

2.1 Low-Rank Hankel Property of PMU Data

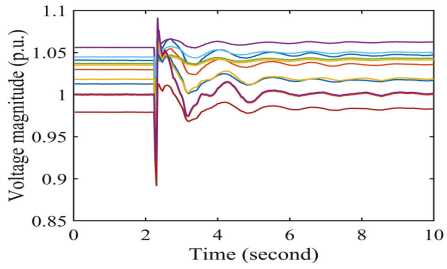


Fig. 1. The measurements of voltage magnitude [Hao et al. 2018]

The Hankel operator $\mathcal{H}$: $\mathbb{R}^{m\times n} \rightarrow \mathbb{R}^{mn_2\times n_1}$ $(n_1 + n_2 = n + 1)$ linearly maps a matrix to its corresponding Hankel matrix, the ith column of which contains the data from all m channels in n_2 consecutive time steps starting from time i, i.e.,

$$X = \mathcal{H}_{n_2}(Y) = \begin{bmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_{n_1} \\ \boldsymbol{y}_2 & \boldsymbol{y}_3 & \cdots & \boldsymbol{y}_{n_1+1} \\ \vdots & \vdots & \cdots & \vdots \\ \boldsymbol{y}_{n_2} & \boldsymbol{y}_{n_2+1} & \cdots & \boldsymbol{y}_n \end{bmatrix} \in \mathbb{R}^{mn_2\times n_1}. \tag{3}$$

Let σ_i denote the ith largest singular value of $\mathcal{H}_{n_2}(Y)$, and let $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ denote the corresponding left and right singular vectors. The rank-r $(r \ll m, n)$ approximation of $\mathcal{H}_{n_2}(Y)$ can be computed from

$$Q^r(\mathcal{H}_{n_2}(Y)) = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T. \tag{4}$$

$Q^r(\mathcal{H}_{n_2}(Y))$ has the smallest normalized approximation error to $\mathcal{H}_{n_2}(Y)$ among all rank r matrices. The normalized approximation error can be computed from $\frac{||Q^r(\mathcal{H}_{n_2}(Y)) - \mathcal{H}_{n_2}(Y)||_F}{||\mathcal{H}_{n_2}(Y)||_F}$.

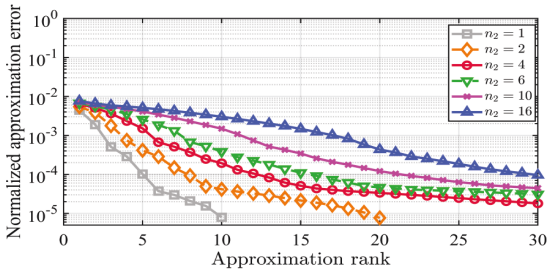


Fig. 2. The normalized approximation errors of different Hankel matrices $\mathcal{H}_{n_2}(Y)$

As discussed in [Hao et al. 2018], the Hankel matrix $\mathcal{H}_{n_2}(Y)$ is often approximately low-rank. That is because for a well-operated power system, some system modes may be highly damped, or not directly measured, or not excited by the input [Hao et al. 2018]. During an event, the observations usually contain at most K modes where K is much less than the system dimension. Then $\mathcal{H}_{n_2}(Y)$ is approximately rank K.

[Hao et al. 2018] provides a formal analysis of the low-rank Hankel property. Here we only show the empirical evaluation on a recorded synchrophasor dataset in Central New York Power System. The dataset in [Hao et al. 2018] contains 11 voltage phasors with 30 samples per second. Fig. 1 shows the voltage magnitude in 10 seconds, and a disturbance occurs at around 2.3 seconds.

Let $Y \in \mathbb{R}^{11\times 300}$ denote measured magnitude of 11 channels in 10 seconds. Fig.2 shows the approximation errors of $\mathcal{H}_{n_2}(Y)$ with varying approximation rank r and the Hankel parameter n_2. All the matrices $\mathcal{H}_{n_2}(Y)$ can be approximated by a rank-6 matrix with a negligible error. For example, when $n_2 = 4$, the rank-6 approximation to $\mathcal{H}_{n_2}(Y)$ has error 0.00067.

3 BAYESIAN ROBUST HANKEL MATRIX COMPLETION

The proposed approach factorizes the Hankel matrix of Y as the product of two factors, the basis matrix $\boldsymbol{D}$ and the coefficient matrix $\boldsymbol{W}$. $\boldsymbol{W}$ is modeled as an element-wise product of two matrices $\boldsymbol{Z}$ and $\boldsymbol{S}$, where the binary matrix $\boldsymbol{Z}$ represents the sparse support, and $\boldsymbol{S}$ represents the non-zero coefficients. Bad data are modeled by a sparse matrix $\boldsymbol{E}$. Each item is modeled by a probability distribution. The algorithm learns the posterior distributions of $\boldsymbol{D}$, $\boldsymbol{Z}$, $\boldsymbol{S}$, and $\boldsymbol{E}$ from obtained partial observations. Our approach then infers the distribution of each entry $Y_{i,j}$. The predictive mean will be calculated

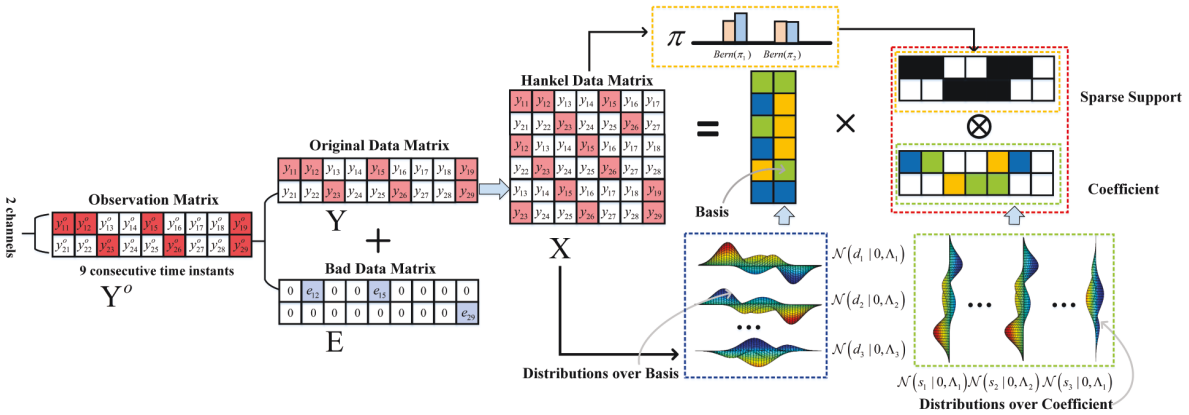


Fig. 3. An overall illustration of the proposed approach. The approach arranges the ground-truth data Y into a Hankel matrix X and then decomposes X in a factorized form with a basis, the sparse support, and the coefficient matrices.

as the estimation of $Y_{i,j}$ and the predictive variance will be computed to measure the uncertainty of the estimation. Fig. 3 shows an overall framework of our proposed algorithm.

This method extends from the conventional Bayesian matrix completion in the following aspects. First, the low-rank Hankel property is exploited to capture the temporal correlations in time series. In conventional low-rank matrix completion, one needs at least r entries in each channel to recover the missing data. The recovery would fail if measurements in all channels were corrupted at the same time instant. Our algorithm additionally considers the temporal correlations and can recover simultaneous missing or bad data. Moreover, the recovery accuracy is enhanced significantly by exploiting the temporal correlations. Second, our method provides the uncertainty measure, which characterizes the confidence of the recovery results. Third, our method can recover both missing and bad data at the same time, as shown in equation (5), while missing and bad data are treated separately in [Babacan et al. 2012]. Lastly, we introduce the additional binary matrix Z to enhance the sparsity of the coefficients W, which in turn leads to a more accurate estimation of the rank and improves the recovery performance.

3.1 Proposed Probabilistic Model

A hierarchical probabilistic model is employed to infer all the latent variables, and (5) to (16) show the model and the prior distribution. (5) is derived from (2), where we use $\boldsymbol{X}$ to denote the low-rank Hankel matrix of the ground-truth data, and the inverse of Hankel matrix $(\mathcal{H}^{\dagger}\boldsymbol{X})_{i,j}$ is employed here to represent $Y_{i,j}$. The formal definition of the inverse Hankel operator $\mathcal{H}^{\dagger}$ can be found in supplementary material. Let $\boldsymbol{X}$ be rank K, then its qth column, denoted by $\boldsymbol{x}_{\cdot q}$, can be written as the product of the basis $\boldsymbol{D} \in \mathbb{R}^{mn_2 \times K}$ with a coefficient vector $\boldsymbol{w}_{\cdot q}$, where $\boldsymbol{w}_{\cdot q}$ is modeled as the element-wise product of two vectors $\boldsymbol{z}_{\cdot q}$ and $\boldsymbol{s}_{\cdot q}$. We introduce the additional binary vector $\boldsymbol{z}_{\cdot q}$ to enhance the sparsity of the coefficients $\boldsymbol{w}_{\cdot q}$. The kth entry of $\boldsymbol{z}_{\cdot q}$, denoted by z_{kq}, is assumed to have a prior Bernoulli distribution with probability π_k. The prior of π_k is a Beta distribution with predefined values a_0 and b_0. Reference [Zhou et al. 2009] shows that data generated from this so-called Beta-Bernoulli process is sparse. Because the actual rank of the Hankel matrix may be unknown, the initial rank K can be set as a large number, and our method can infer the actual rank by gradually pruning the basis using the sparsity of learned Z from data.

The prior distribution $\boldsymbol{s}_{\cdot q}$ is a multivariate Gaussian $\mathcal{N}(\boldsymbol{0}, \gamma_s^{-1}\boldsymbol{I}_K)$, where $\boldsymbol{I}_K$ is a $K \times K$ identity matrix. Each entry of the noise matrix $\boldsymbol{N}$ and the error matrix $\boldsymbol{E}$ is drawn from $\mathcal{N}(0, \gamma_\epsilon^{-1})$ and $\mathcal{N}(0, \beta_{i,j}^{-1})$, respectively. Three gamma priors are incorporated on γ_s, γ_ϵ and $\beta_{i,j}$, following Gamma priors with parameters (c_0, d_0), (e_0, f_0), and (g_0, h_0), respectively. The prior distribution of each row of $\boldsymbol{D}$ is $\mathcal{N}(\boldsymbol{0}, \lambda_d^{-1}\boldsymbol{I}_K)$, where λ_d is a pre-defined value. [Babacan et al. 2012] shows that the Gaussian distribution with Gamma priors models the sparsity of the bad data $E_{i,j}$. The Gaussian assumption for the bad data has been employed in the literature, see, e.g., [Luttinen et al. 2012][Zhao et al. 2015b] and [Babacan et al. 2012]. We employ conjugate priors to simplify calculations and obtain analytical posterior distributions.

For all $p = 1, 2, 3, ..., mn_2$, $q = 1, 2, 3, ..., n_1$, and $k = 1, 2, 3, ..., K$,

$$Y^o_{i,j} = (\mathcal{H}^{\dagger}\boldsymbol{X})_{i,j} + E_{i,j} + N_{i,j} \quad (i,j) \in \Omega, \tag{5}$$

$$\boldsymbol{x}_{\cdot q} = \boldsymbol{D}\boldsymbol{w}_{\cdot q}, \tag{6}$$

$$\boldsymbol{w}_{\cdot q} = (\boldsymbol{z}_{\cdot q} \odot \boldsymbol{s}_{\cdot q}), \tag{7}$$

$$d_{p\cdot} \sim \mathcal{N}(\boldsymbol{0}, \lambda_d^{-1}\boldsymbol{I}_K), \tag{8}$$

$$\boldsymbol{z}_{\cdot q} \sim \prod_{k=1}^{K} \text{Bernoulli}(\pi_k), \tag{9}$$

$$\pi_k \sim \text{Beta}(a_0/K, b_0(K-1)/K), \tag{10}$$

$$\boldsymbol{s}_{\cdot q} \sim \mathcal{N}(\boldsymbol{0}, \gamma_s^{-1}\boldsymbol{I}_K), \tag{11}$$

$$E_{i,j} \sim \mathcal{N}(0, \beta_{i,j}^{-1}) \quad (i,j) \in \Omega, \tag{12}$$

$$N_{i,j} \sim \mathcal{N}(0, \gamma_\epsilon^{-1}), \tag{13}$$

$$\gamma_s \sim \Gamma(c_0, d_0), \tag{14}$$

$$\gamma_\epsilon \sim \Gamma(e_0, f_0), \tag{15}$$

$$\beta_{i,j} \sim \Gamma(g_0, h_0). \tag{16}$$

3.2 Variational Inference for Approximating the Posterior Distributions

Let $\Theta = \{\boldsymbol{d}_{p\cdot}, \boldsymbol{s}_{\cdot q}, \boldsymbol{z}_{\cdot q}, E_{i,j}, \boldsymbol{\pi}_k, \gamma_s, \gamma_\epsilon, \beta_{i,j}, p = 1, 2, 3, ..., mn_2, q = 1, 2, 3, ..., n_1, k = 1, 2, 3, ..., K, (i, j) \in \Omega\}$ denote all the latent variables. Given Y^o_Ω, we aim to compute the posterior $P(\Theta, Y|Y^o_\Omega)$. From the Bayes' theorem,

$$P(\Theta, Y|Y^o_\Omega) = \frac{P(\Theta, Y, Y^o_\Omega)}{P(Y^o_\Omega)}. \tag{17}$$

Because $P(Y^o_\Omega)$ is difficult to calculate by marginalizing all the latent variables, computing (17) is intractable.

The mean field variational inference [Bishop 2006] is employed here to approximate $P(\Theta, Y|Y^o_\Omega)$ by the variational distribution $q(\Theta)$. Mean field approximation assumes that elements in Θ are mutually independent and $q(\Theta)$ is factorized as

$$q(\Theta) = q(D)q(S)q(Z)q(\boldsymbol{\pi})q(E)q(\boldsymbol{\beta})q(\gamma_s)q(\gamma_\epsilon) = \prod_{p=1}^{mn_2} q(\boldsymbol{d}_{p\cdot}) \prod_{q=1}^{n_1} q(\boldsymbol{s}_{\cdot q})q(\boldsymbol{z}_{\cdot q}) \prod_{k=1}^{K} q(\pi_k) \prod_{(i,j)\in\Omega} q(E_{i,j})q(\beta_{i,j})q(\gamma_s)q(\gamma_\epsilon). \tag{18}$$

The Kullback–Leibler (KL) divergence is employed to measure the similarity of two distributions. Variational inference finds the closest approximation $q(\Theta)$ to $P(\Theta, Y_\Omega|Y^o_\Omega)$ by solving the following optimization problem,

$$\begin{aligned} q(\Theta) &= \arg\min_{q(\Theta)} \mathbb{KL}(q(\Theta)||P(\Theta, Y|Y^o_\Omega)) \\ &= \arg\max_{q(\Theta)} \mathbb{E}[\ln P(\Theta, Y, Y^o_\Omega)] - \mathbb{E}[\ln q(\Theta)]. \end{aligned} \tag{19}$$

The above optimization problem is solved approximately by sequentially estimating the approximation distribution of each factor given all the others. Each approximation distribution is obtained through computing the expectations of all the other factors based on learned distributions [Bishop 2006; Blei et al. 2017]. The stationary approximation distribution of the variational inference is a local optimum to the optimization problem (19) [Bishop 2006; Blei et al. 2017]. For example, $q(\boldsymbol{d}_{p\cdot})$ denotes the approximation distribution of $\boldsymbol{d}_{p\cdot}$ while keeping other latent variables fixed. The optimal $q(\boldsymbol{d}_{p\cdot})$ which maximizes the objective function in (19) is

$$\ln q(\boldsymbol{d}_{p\cdot}) = \mathbb{E}_{q(\Theta\setminus\boldsymbol{d}_{p\cdot})}[\ln P(\Theta, Y, Y^o_\Omega)] + \text{constant}, \tag{20}$$

$\mathbb{E}_{q(\Theta\setminus\boldsymbol{d}_{p\cdot})}$ means taking the expectation with respect to all the latent variables except $\boldsymbol{d}_{p\cdot}$.

As all the distributions in the proposed model have conjugate priors, the conditional posterior distributions have explicit forms. We directly present the conditional distribution and expectation of each variable. The details of the derivations are summarized in the supplementary material.

(I) The approximate posterior distribution of $\boldsymbol{d}_{p\cdot}$ (for all $p = 1, ..., mn_2$), the pth row of basis, is a Gaussian distribution with mean $\mathbb{E}[\boldsymbol{d}_{p\cdot}]$, which denotes the expectation of $q(\boldsymbol{d}_{p\cdot})$, and covariance $\Sigma_{\boldsymbol{d}_{p\cdot}}$, i.e.,

$$q(\boldsymbol{d}_{p\cdot}) \sim \mathcal{N}(\mathbb{E}[\boldsymbol{d}_{p\cdot}], \Sigma_{\boldsymbol{d}_{p\cdot}}), \tag{21}$$

where

$$\Sigma_{\boldsymbol{d}_{p\cdot}} = [\mathbb{E}[\gamma_\epsilon] \sum_{q:(p,q)\in\Psi_\Omega} \mathbb{E}[(\boldsymbol{s}_{\cdot q} \odot \boldsymbol{z}_{\cdot q})(\boldsymbol{s}_{\cdot q} \odot \boldsymbol{z}_{\cdot q})^T] + \lambda_d \boldsymbol{I}_K]^{-1}, \tag{22}$$

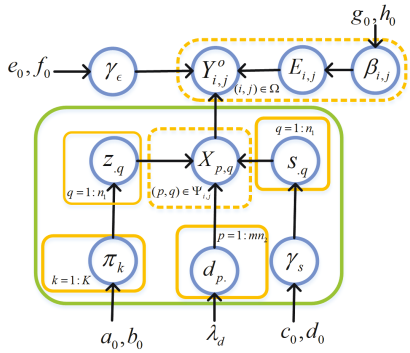


Fig. 4. Graphical representation of the dependence of the random variables in the proposed Bayesian Hankel matrix completion model

$$\mathbb{E}[\boldsymbol{d}_{p\cdot}] = \mathbb{E}[\gamma_\epsilon] \sum_{q:(p,q)\in\Psi_\Omega} \mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})_{p,q}(\mathbb{E}[\boldsymbol{s}_{\cdot q}]^T \odot \mathbb{E}[\boldsymbol{z}_{\cdot q}]^T)\Sigma_{\boldsymbol{d}_{p\cdot}}, \tag{23}$$

Ψ_Ω denotes the set of observed entries in the Hankel matrix of $\boldsymbol{Y}^o$.

(II) The approximate posterior distribution of $\boldsymbol{s}_{\cdot q}$ ($q = 1, ..., n_1$) is a Gaussian distribution.

$$q(\boldsymbol{s}_{\cdot q}) \sim \mathcal{N}(\mathbb{E}[\boldsymbol{s}_{\cdot q}], \Sigma_{\boldsymbol{s}_{\cdot q}}), \tag{24}$$

where

$$\Sigma_{\boldsymbol{s}_{\cdot q}} = [\mathbb{E}[\gamma_\epsilon] \sum_{p:(p,q)\in\Psi_\Omega} \mathbb{E}[\boldsymbol{\phi}^T_{p,q}\boldsymbol{\phi}_{p,q}] + \mathbb{E}[\gamma_s]\boldsymbol{I}_K]^{-1}, \tag{25}$$

$$\mathbb{E}[\boldsymbol{s}_{\cdot q}] = \mathbb{E}[\gamma_\epsilon]\Sigma_{\boldsymbol{s}_{\cdot q}} \sum_{p:(p,q)\in\Psi_\Omega} \mathbb{E}[\boldsymbol{\phi}_{p,q}]^T \mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})_{p,q}, \tag{26}$$

$\boldsymbol{\phi}_{p,q} = \boldsymbol{d}_{p\cdot} \odot \boldsymbol{z}^T_{\cdot q}$.

(III) The approximate posterior distribution of z_{kq} (for all $q = 1, ..., n_1$, and $k = 1, ..., K$) is a Bernoulli distribution.

$$q(z_{kq}) \sim \text{Bernoulli}(\frac{q(z_{kq} = 1)}{q(z_{kq} = 1) + q(z_{kq} = 0)}), \tag{27}$$

with mean and variance

$$\mathbb{E}[z_{kq}] = \frac{q(z_{kq} = 1)}{q(z_{kq} = 1) + q(z_{kq} = 0)}, \tag{28}$$

$$\Sigma_{z_{kq}} = \mathbb{E}[z_{kq}](1 - \mathbb{E}[z_{kq}]), \tag{29}$$

where

$$\begin{aligned} &\ln(q(z_{kq} = 1)) \propto \\ &\frac{-\mathbb{E}[\gamma_\epsilon]}{2} \sum_{p:(p,q)\in\Psi_\Omega} [\text{trace}(\mathbb{E}[\boldsymbol{d}^T_{p\cdot}\boldsymbol{d}_{p\cdot}](\mathbb{E}[\boldsymbol{s}_{\cdot q}\boldsymbol{s}^T_{\cdot q}] \odot \mathbb{E}[\hat{\boldsymbol{z}}_{\cdot q}\hat{\boldsymbol{z}}^T_{\cdot q}]))] \\ &+\mathrm{E}[\gamma_\epsilon] \textstyle\sum_{p:(p,q)\in\Psi_\Omega} \mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})_{p,q}[(\mathbb{E}[\boldsymbol{s}_{\cdot q}] \odot \mathbb{E}[\hat{\boldsymbol{z}}_{\cdot q}])^T \mathbb{E}[\boldsymbol{d}_{p\cdot}]^T] + \mathbb{E}[\ln(\pi_k)], \end{aligned} \tag{30}$$

where $\propto$ denotes "proportional to." $\hat{\boldsymbol{z}}_{\cdot q} = [\hat{z}_{1q}, \hat{z}_{2q}, ..., \hat{z}_{kq}, ..., \hat{z}_{Kq}]^T$. $\hat{z}_{kq} = 1$ and other entries in $\hat{\boldsymbol{z}}_{\cdot q}$ equal to the corresponding entries in $\boldsymbol{z}_{\cdot q}$.

$$\begin{aligned} &\ln(q(z_{kq} = 0)) \propto \\ &\frac{-\mathbb{E}[\gamma_\epsilon]}{2} \sum_{p:(p,q)\in\Psi_\Omega} [\text{trace}(\mathbb{E}[\boldsymbol{d}^T_{p\cdot}\boldsymbol{d}_{p\cdot}](\mathbb{E}[\boldsymbol{s}_{\cdot q}\boldsymbol{s}^T_{\cdot q}] \odot \mathbb{E}[\hat{\boldsymbol{z}}_{\cdot q}\hat{\boldsymbol{z}}^T_{\cdot q}]))] \\ &+\mathrm{E}[\gamma_\epsilon] \textstyle\sum_{p:(p,q)\in\Psi_\Omega} \mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})_{p,q}[(\mathbb{E}[\boldsymbol{s}_{\cdot q}] \odot \mathbb{E}[\hat{\boldsymbol{z}}_{\cdot q}])^T \mathbb{E}[\boldsymbol{d}_{p\cdot}]^T] + \mathbb{E}[\ln(1 - \pi_k)], \end{aligned} \tag{31}$$

$\hat{z}_{kq} = 0$, and other entries in $\hat{z}_{.q}$ equal to the corresponding entries in $z_{.q}$.

(IV) The approximate distribution of π_k ($k = 1, ..., K$) is from a Beta distribution.

$$q(\pi_k) \sim \text{Beta}(a_0/K + \sum_{q=1}^{n_1} \mathbb{E}[z_{kq}], b_0(K-1)/K + n_1 - \sum_{q=1}^{n_1} \mathbb{E}[z_{kq}]). \tag{32}$$

Therefore,

$$\mathbb{E}[\ln(\pi_k)] = \psi(a_0/K + \sum_{q=1}^{n_1} \mathbb{E}[z_{kq}]) - \psi((a_0 + b_0(K-1))/K + n_1), \tag{33}$$

$$\mathbb{E}[\ln(1-\pi_k)] = \psi(b_0(K-1)/K + n_1 - \sum_{q=1}^{n_1} \mathbb{E}[z_{kq}]) - \psi((a_0 + b_0(K-1))/K + n_1), \tag{34}$$

where $\psi(.)$ is the diagamma function.

(V) The approximate posterior distribution of γ_s is a Gamma distribution.

$$q(\gamma_s) \sim \Gamma(\frac{n_1 K}{2} + c_0, \frac{1}{2}\sum_{q=1}^{n_1} \mathbb{E}[s_{.q}^T s_{.q}] + d_0), \tag{35}$$

with mean

$$\mathbb{E}[\gamma_s] = \frac{n_1 K + 2c_0}{\sum_{q=1}^{n_1} \mathbb{E}[\boldsymbol{s}_{.q}^T \boldsymbol{s}_{.q}] + 2d_0}, \tag{36}$$

where $\mathbb{E}[\boldsymbol{s}_{.q}^T \boldsymbol{s}_{.q}] = \mathbb{E}[\boldsymbol{s}_{.q}^T]\mathbb{E}[\boldsymbol{s}_{.q}] + \text{trace}(\Sigma_{\boldsymbol{s}_{.q}})$.

(VI) The approximate posterior distribution of $E_{i,j}$ (for $(i,j) \in \Omega$) is a Gaussian distribution.

$$q(E_{i,j}) \sim \mathcal{N}(\mathbb{E}[E_{i,j}], \Sigma_{E_{i,j}}), \tag{37}$$

with variance and mean

$$\Sigma_{E_{i,j}} = (\mathbb{E}[\gamma_\epsilon] + \mathbb{E}[\beta_{i,j}])^{-1}, \tag{38}$$

$$\mathbb{E}[E_{i,j}] = \mathbb{E}[\gamma_\epsilon]\Sigma_{E_{i,j}}(Y^o_{i,j} - \mathbb{E}[(\mathcal{H}^\dagger \boldsymbol{X})_{ij}]). \tag{39}$$

(VII) The approximate posterior distribution of $\beta_{i,j}$ (for $(i,j) \in \Omega$) is a Gamma distribution.

$$\beta_{i,j} \sim \Gamma(\frac{1}{2} + g_0, \frac{1}{2}\mathbb{E}[E_{i,j}^2] + h_0), \tag{40}$$

with mean

$$\mathbb{E}[\beta_{i,j}] = \frac{1 + 2g_0}{\mathbb{E}[E_{i,j}^2] + 2h_0}. \tag{41}$$

(VI) The approximate posterior distribution of γ_ϵ is a Gamma distribution.

$$q(\gamma_\epsilon) \sim \Gamma(\frac{|\Omega|}{2} + e_0, \frac{1}{2}\mathbb{E}[||\boldsymbol{Y}^o - P_\Omega(\mathcal{H}^\dagger \boldsymbol{X} + \boldsymbol{E})||_F^2] + f_0), \tag{42}$$

with mean

$$\mathbb{E}[\gamma_\epsilon] = \frac{|\Omega| + 2e_0}{\mathbb{E}[||\boldsymbol{Y}^o - P_\Omega(\mathcal{H}^\dagger \boldsymbol{X} + \boldsymbol{E})||_F^2] + 2f_0}. \tag{43}$$

Pruning the basis D and the error matrix E. As $\boldsymbol{D}$ is a redundant basis when K is larger than the ground-truth rank, we propose to prune the basis $\mathbb{E}[\boldsymbol{D}]$ to reduce computation. If $\mathbb{E}[z_{kq}] = 0$ for all q in each iteration, the algorithm removes kth basis atom $\boldsymbol{d}_{.k}$ because $\boldsymbol{d}_{.k}$ does not contribute to the representation of X. Then the algorithm also removes the corresponding $\mathbb{E}[\ln(\pi_k)]$, $\mathbb{E}[\ln(1-\pi_k)]$, $\mathbb{E}[z_{kq}]$, $\mathbb{E}[s_{kq}]$ for all q. Because $\boldsymbol{E}$ is sparse, we also prune $\mathbb{E}[\boldsymbol{E}]$ by thresholding, i.e., if entries in $\mathbb{E}[\boldsymbol{E}]$ are very small (e.g., 10^{-1}), these entries are set as zeroes.

Convergence criteria. Matrix $\bar{\boldsymbol{X}}$ is the estimation for $\boldsymbol{X}$ at the current iteration. Matrix $\bar{\boldsymbol{X}}_{\text{pre}}$ is the estimation for X at the previous iteration The algorithm terminates if $\frac{\|\bar{\boldsymbol{X}} - \bar{\boldsymbol{X}}_{\text{pre}}\|_F}{\|\bar{\boldsymbol{X}}_{\text{pre}}\|_F} < \xi$ for a pre-determined threshold ξ (e.g., 10^{-4}) or if the maximum iterations $T_{\max}$ is reached.

Initialization. After constructing the Hankel matrix $\boldsymbol{X}$ from $\boldsymbol{Y}^o$, where missing entries are filled in zeros, we compute the SVD of $\boldsymbol{X}$ as $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{A}\boldsymbol{V}^T$. $\boldsymbol{D}$ is initialized by $\boldsymbol{U}\boldsymbol{A}^{\frac{1}{2}}$ and $\boldsymbol{S}$ is initialized with $\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{V}^T$. $\boldsymbol{z}_{.q}$ are initialized with all ones. All values in π_k are initialized as 0.5. γ_s, and λ_d are initialized by $\frac{||\boldsymbol{Y}^o||_F}{\sqrt{mn}}$. $1/\gamma_\epsilon$ is initialized by $\frac{||\boldsymbol{Y}^o||_F^2}{mn}$. The initial $\bar{\boldsymbol{X}}^0 = \boldsymbol{D}(\boldsymbol{S} \odot \boldsymbol{Z})$. The initial $\boldsymbol{E}$ is $E_{i,j} = Y^o_{i,j} - (\mathcal{H}^\dagger \bar{\boldsymbol{X}}^0)_{i,j}$ if $(i,j) \in \Omega$ and $E_{i,j} = 0$ otherwise. All the covariance matrices for $\boldsymbol{d}_{p.}$ and $\boldsymbol{s}_{.q}$ are initialized by a $K \times K$ diagonal matrix where the diagonal elements are equal to $\frac{||\boldsymbol{Y}^o||_F}{\sqrt{mn}}$. The covariance matrices for $\boldsymbol{z}_{.q}$ are initialized by a $K \times K$ zero matrix. All the elements in $\boldsymbol{\beta}$ are initialized as $\frac{||\boldsymbol{Y}^o||_F}{\sqrt{mn}}$.

Missing data only. Our algorithm can be simplified if only missing data presents. One can skip steps VI and VII about updating $\mathbb{E}[E_{i,j}]$ and $\mathbb{E}[\beta_{i,j}]$, and other steps remain unchanged.

Computational complexity. The per-iteration computational complexity is $O(\kappa mn_2 n_1 K^4 + mn_2 K^3 + n_1 K^3 + mn_2 n_1 K)$, where κ ($0 < \kappa \leq 1$) is the portion of observed entries. Thus, it is at most linear in the dimension of the Hankel matrix. Derivation of the complexity can be found in Section A.4 in the supplementary materials.

3.3 The uncertainty measure

Let $\boldsymbol{\theta} = \{\boldsymbol{D}, \boldsymbol{Z}, \boldsymbol{S}, \gamma_\epsilon\}$ denote all the latent variables related to $Y_{i,j}$. After computing the posterior distributions, we employ Monte-Carlo integration [Paisley et al. 2012] to estimate the mean and variance of $Y_{i,j}$. Define

$$f^{\boldsymbol{\theta}}(Y_{i,j}) = \mathcal{H}^\dagger(\boldsymbol{D}(\boldsymbol{S} \odot \boldsymbol{Z}))_{i,j}. \tag{44}$$

The predictive mean is computed by

$$\hat{Y}_{i,j} = \mathbb{E}[Y_{i,j}] \approx \frac{1}{L}\sum_{l=1}^{l=L} f^{\boldsymbol{\theta}_l}(Y_{i,j}), \tag{45}$$

where each $\boldsymbol{\theta}^l$ is independently drawn from the learned approximation distributions of $\boldsymbol{D}$, $\boldsymbol{Z}$, $\boldsymbol{S}$, and γ_ϵ. L is the number of Monte-Carlo samples. Similarly, the predictive variance is approximated by

$$\begin{aligned} \text{Var}[Y_{i,j}] &= \mathbb{E}[Y_{i,j}^2] - \mathbb{E}[Y_{i,j}]^2 \\ &\approx \frac{1}{L}\sum_{l=1}^{l=L}\frac{1}{\gamma_\epsilon} + \frac{1}{L}\sum_{l=1}^{l=L} f^{\boldsymbol{\theta}_l}(Y_{i,j})^2 - (\frac{1}{L}\sum_{l=1}^{l=L} f^{\boldsymbol{\theta}_l}(Y_{i,j}))^2. \end{aligned} \tag{46}$$

The derivation details of (45) and (46) are provided in the supplementary materials. A larger L offers a more accurate estimate but also leads to a higher computational cost. In our experiments, $L = 50$ is sufficient to provide a reliable estimate of the mean and the variance.

$\mathbb{E}[Y_{i,j}]$ is used to as an estimate $\hat{Y}_{i,j}$ of $Y_{i,j}$, and $\mathrm{Var}[Y_{i,j}]$ is used as an *uncertainty index* of the estimation, because a larger variance indicates a higher uncertainty in the estimation.

3.4 Parameter Selections of the Algorithm

Several pairs of parameters (a_0, b_0),(c_0, d_0),(e_0, f_0), (g_0, h_0) are needed in the prior distributions (10), (14), (15) and (16) respectively. [Yi and Wang 2021; Zhou et al. 2009] show that c_0 and d_0 are non-informative priors, which have a negligible impact on the results, and can be set as small values (e.g., 10^{-6}). A larger a_0 with fixed b_0 leads to a larger mean of the prior distribution of π_k, which in turn leads to less number of zero entries in $\boldsymbol{z}_{\cdot q}$. Decreasing f_0 with fixed e_0 leads to a larger γ_ϵ, which leads to a smaller variance of the measurement noise N. A larger h_0 with fixed g_0 leads to smaller values of $\boldsymbol{\beta}$, which leads to larger values in $\boldsymbol{E}$. Note that these parameters only have slight impact on the inferred posterior distributions. Section 4.2.3 demonstrates that the proposed method is robust to parameter selections and these parameters can be set in a wide range. A larger n_2 improves the performance of the algorithm but also suffers from higher computational burden. In our experiments, n_2 is set as at most 30, and it is sufficient to obtain reliable recovery results.

3.5 Time Window Selection for Streaming data

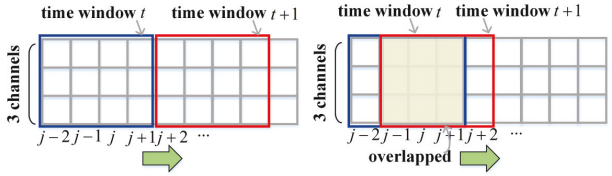


Fig. 5. Non-overlapping and overlapping sliding windows

When handling streaming data in real-time, one needs to truncate the measurements into blocks and process each time block separately. One can use a sliding window with length n and step size s. When the window is non-overlapping ($n = s$), as shown in the left half of Fig. 5, each entry is estimated once in one time window. Otherwise, every entry is estimated $\lfloor n/s \rfloor$ times in different time windows, where $\lfloor x \rfloor$ means the greatest integer less than or equal to x. One can pick the estimate that has the smallest uncertainty index. For example, the right half of Fig. 5 shows overlapping windows with $n = 4$ and $s = 1$.

4 NUMERICAL EXPERIMENTS

4.1 Experimental Setup

Three modes of missing data and bad data are considered, as shown in Fig. 6. For example, M1 means Mode 1 missing data, and B1 means Mode 1 bad data. The value of the additive error is randomly generated from $(2, 4)$ for synthetic data and $(1, 1.5)$ for real data.

- Mode 1: Missing/bad data occur independently and randomly across all the channels and time instants.
- Mode 2: Missing/bad data occur across all the channels at some randomly selected time instants.
- Mode 3: Missing/bad data occur across all the channels at consecutive time instants. The starting time is selected randomly.

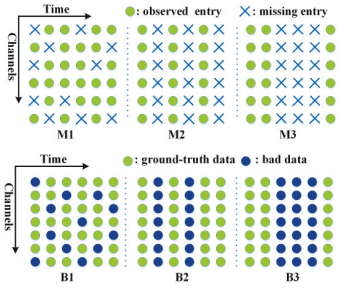


Fig. 6. The missing and bad data generation.

Our proposed Bayesian Robust Hankel matrix completion method, abbreviated by "BRHMC," is compared with the deterministic Hankel robust matrix completion method "SAP" in [Zhang and Wang 2019] and the deterministic robust matrix completion method "R-RMC" in [Cherapanamjeri et al. 2017] for simultaneous recovery of missing and bad data. When the goal is recovering missing data only, we compare a simplified version of our method, abbreviated by "BHMC," with the deterministic Hankel missing data recovery method "AM-FIHT" in [Zhang et al. 2018] and Bayesian missing data recovery method "VSBL" in [Babacan et al. 2012]. Some parameters of BRHMC/BHMC are set as follows for all the experiments: $a_0 = 10^3$, $b_0 = 1$, $c_0 = 10^{-6}$, $d_0 = 10^{-6}$, $e_0 = 10^{-6}$, $h_0 = 10^{-6}$. The experiments are implemented in MATLAB 2019 on a desktop with 3.1 GHz Intel Core i9 and 32 GB RAM.

Evaluation Metrics: Two metrics are used to measure the recovery performance. The Normalized Estimation Error (NEE) is defined as

$$\mathrm{NEE} = \|\hat{Y} - Y\|_F / \|Y\|_F, \tag{47}$$

where $\hat{Y}$ and Y in $\mathbb{R}^{m \times n}$ represent the estimated data and the ground-truth data, respectively. A new metric weighted normalized estimation error (WNEE) is defined as

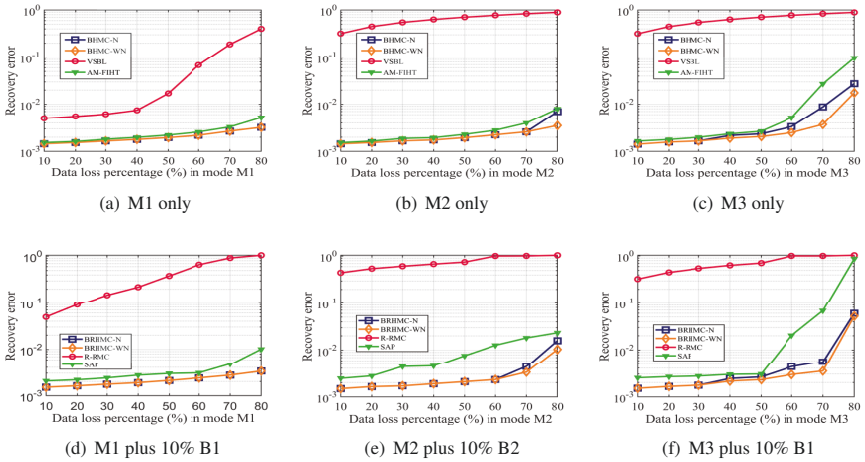
$$\mathrm{WNEE} = \sqrt{\Sigma_{i,j}\left(\frac{(\hat{Y}_{i,j} - Y_{i,j})^2}{\mathrm{Var}[\hat{Y}_{i,j}]}\right) / \left(\Sigma_{i,j}\frac{Y_{i,j}^2}{\mathrm{Var}[\hat{Y}_{i,j}]}\right)}. \tag{48}$$

When $\mathrm{Var}[\hat{Y}_{i,j}]$ is large, there is a higher uncertainty in the estimate $\hat{Y}_{i,j}$. Then from (48), a smaller weight in placed on $\hat{Y}_{i,j}$ when computing the overall performance error. If the variance is the same for all $\hat{Y}_{i,j}$, WNEE is equal to NEE. If WNEE is smaller than NEE, then those estimations with large errors are indeed penalized with a small weight in WNEE and, thus, the corresponding variance is large. Thus, WNEE being smaller than NEE indicates that the uncertainty index indeed represents the accuracy of the estimation.

4.2 Performance on Synthetic Datasets

4.2.1 Dataset generation and parameter setting. We conduct the experiments on synthetic spectrally sparse signals which have the low-rank Hankel property [Zhang et al. 2018; Zhang and Wang 2019]. Each row of Y is a weighted sum of r sinusoids. Specifically, the ground truth $Y_{i,j}$ is generated from

$$Y_{i,j} = \mathrm{Real}\left(\sum_{k=1}^{r} b_{i,k} e^{\iota 2\pi f_k j}\right) \quad i = 1, ..., m, \, j = 1, ..., n, \tag{49}$$



(a) M1 only (b) M2 only (c) M3 only

(d) M1 plus 10% B1 (e) M2 plus 10% B2 (f) M3 plus 10% B1

Fig. 7. The recovery results with different missing/bad data. (a)-(c) show the recovery results with three missing modes. (d)-(f) show the recovery results with three missing and bad modes.

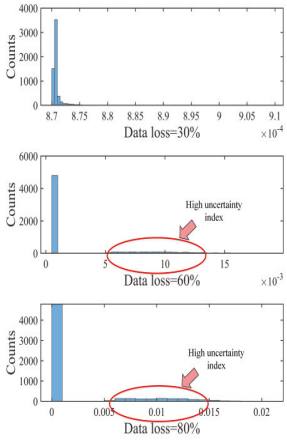


Fig. 8. The histogram of uncertainty index in the M3 mode.

where $\imath$ is the imaginary unit, f_k is the frequency, $b_{i,k}$ is the normalized complex amplitude of the k-th sinusoid, and Real$(\cdot)$ keeps the real part only. We randomly select f_k from $(0, 1)$. The angle of $b_{i,k}$ is randomly selected from $(0, 2\pi)$, and the magnitude is $1 + 10^{0.5a_{i,k}}$, where $a_{i,k}$ is randomly generated from $(0, 1)$. Y is rank $2r$. Here, r is set as 2. $m = 20$, and $n = 300$. Each entry is added with a random Gaussian noise from $\mathcal{N}(0, 0.03^2)$, which is about 1.1% NEE error. Each bad data entry is randomly selected from $(2, 4)$.

Some parameters of BRHMC/BHMC are set as follows: $f_0 = 10^{-6}$, $g_0 = 10^{-6}$, $\xi = 10^{-4}$, $n_2 = 30$. K is 4. $T_{\max}$ is 100. $L = 50$ in (45) and (46). All the results are averaged over 50 independent trials.

4.2.2 Recovery performance. Fig. 7 (a)-(c) compare the missing recovery performance of BHMC with VSBL and AM-FIHT. Fig. 7 (d)-(f) compare the recovery performance of BRHMC with R-RMC and SAP when both missing and bad data exist. BHMC-N denotes NEE error in (47) for BHMC, and BHMC-WN denotes the WNEE error in (48). Because no uncertainty index is provided for all other methods, only NEE error is reported.

The recovery errors of BRHMC/BHMC stay consistently small and outperform all the existing methods. Specifically, the conventional low-rank methods like VSBL and R-RMC perform poorly in Mode 2 and Mode 3, because they cannot handle simultaneous data issues across all channels. Deterministic Hankel-based methods like AM-FIHT and SAP outperform low-rank methods but perform worse than our proposed methods. Moreover, AM-FIHT and SAP are more sensitive to rank selections than our methods. We also tested other distributions of bad data and noise and obtained similar results as those in Fig. 7. Please see Fig. 11 in the supplementary materials.

When the data loss percentage is high, WNEE is less than NEE of our proposed methods. As discussed after (48), this gap indicates that those estimates with larger errors have larger variances. Fig. 8 further shows the histogram of uncertain indices in mode M3. When the data loss percentage increase, the uncertain indices of some entries increase, indicating a less reliable estimation. Our methods can differentiate unreliable estimates from reliable ones.

The average time to compute the posterior distribution is 2-7 seconds, and the Monte-Carlo computation of mean and variance takes around 0.5-1 second. The computational time for AM-FIHT, VSBL, R-RMC and SAP are 0.9-1.3 seconds, 0.1-0.4 seconds, 0.05-0.2 seconds, and 0.2-3 seconds, respectively.

4.2.3 The impact of parameter selections. Numerical experiments are conducted on a dataset with 20% B1 and 20% M2 to test the impact of parameter selections on the performance of BRHMC. As discussed in Section 3.4, we only consider the impact of three pairs $(a_0, b_0), (e_0, f_0), (g_0, h_0)$, and vary one while fixing the other. One can see from Tables 1-3, the recovery errors remain small in a wide range of parameters, and NEE and WNEE are the same.

Table 1. The impact of a_0 (b_0 is fixed and $b_0 = 1$)

a_0	1	10	10^2	10^3	10^4	10^5
(W)NEE	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017

Table 2. The impact of f_0 (e_0 is fixed and $e_0 = 10^{-6}$)

f_0	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
(W)NEE	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017

Table 3. The impact of h_0 (g_0 is fixed and $g_0 = 10^{-6}$)

h_0	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
(W)NEE	0.0024	0.0018	0.0017	0.0017	0.0017	0.0017

Table 4. The impact of the initial rank K

	initial rank K	4	12	20	28	32
Proposed	(W)NEE	0.0017	0.0017	0.0017	0.0017	0.0017
	estimated rank	4	5	5	5	5
SAP	NEE	0.064	0.0040	0.0053	0.0063	0.0067
AM-FIHT	NEE	0.0017	0.0027	0.0035	0.0042	0.0045

Because BHMC prunes the basis during the inference, it is robust to the initial rank K of basis. Table 4 shows that when K is selected

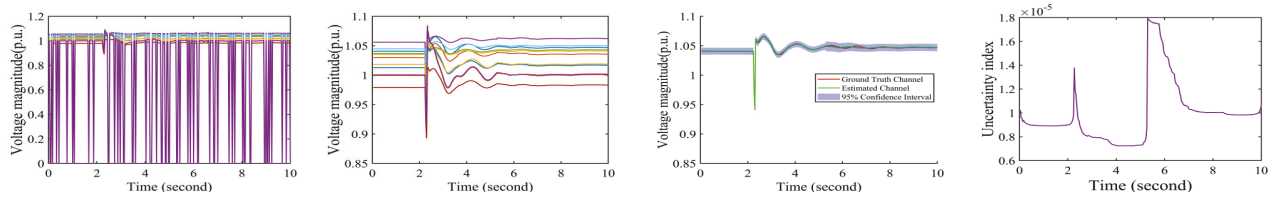


Fig. 9. The recovery performance on 20% M2 missing data and additional noise during 5.6-6.6 seconds. (a) the observed data, (b) the estimated data, (c) the estimated data in one channel with the confidence interval, (d) the corresponding uncertainty index for one channel in (c)

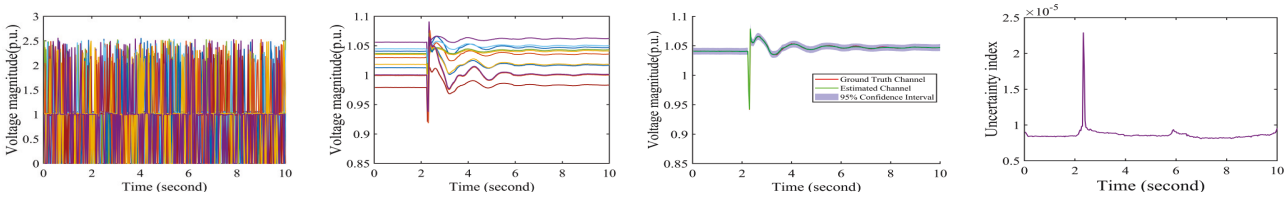


Fig. 10. The recovery performance on 20% M1 missing data and 15% B1 bad data. (a) the observed data, (b) the estimated data, (c) the estimated data in one channel with the confidence interval, (d) the corresponding uncertainty index for one channel in (c)

Table 5. The impact of Hankel parameter n_2

n_2	1	10	20	30	40	50
NEE	0.45	0.0021	0.0018	0.0017	0.0017	0.0017
WNEE	0.014	0.0021	0.0018	0.0017	0.0017	0.0017

from a wide range, the recovery error of the proposed method is always small, and the final estimated rank is close to the ground-truth value 4. In contrast, the performance of SAP and AM-FIHT degrades when the rank is not properly selected. In Table 4, AM-FIHT is tested for 20% M2, while others are tested on 20% B1 and 20% M2.

Table 5 shows the performance when the Hankel block size n_2 increases. When $n_2 = 1$, the method reduces to the conventional Bayesian matrix completion method, which has a large error. Increasing n_2 indeed improves the recovery performance.

4.3 Performance on practical PMU dataset

The recorded synchrophasor dataset in Central New York Power System as shown in Fig. 1 is employed here to evaluate the performance on streaming data. The proposed method is compared with SAP algorithm. $g_0 = 0.2$, $\xi = 10^{-6}$. The window length is set as 50 for our algorithm and 60 for SAP. We use a sliding window with step size 1 for our algorithm. Non-overlapping windows are employed for SAP, because it does not return an uncertainty index to compare the performance of overlapping windows. Two case studies are considered. $n_2 = 20$ for Case 1, and $n_2 = 6$ for Case 2. $f_0 = 10^{-3}$ for Case 1 and Case 2. Besides, the ranks are set as $K = 6$ for two algorithms. The computational time of the non-overlapping windows for Case 1 is 2.6 seconds and is 6.6 seconds for Case 2. Another two case studies for the phasor angle data are included in the supplementary materials.

- Case 1: 20% data are removed following Mode M2. Moreover, additional Gaussian noise from $\mathcal{N}(0, 0.003^2)$ is added to every observation during time 5.6 to 6.6 seconds.
- Case 2: 20% data are removed following Mode M1, and 15% observations contain Mode B1 bad data. Each bad entry is randomly selected from (1,1.5).

Our method can recover the data accurately in both cases. NEE and WNEE for Case 1 are 8.8×10^{-4} and 8.4×10^{-4}, respectively. NEE and WNEE for Case 2 is 2.0×10^{-3} and 1.5×10^{-3}, respectively. In comparison, the NEE of SAP for Case 1 and 2 is 5.9×10^{-3} and 6.0×10^{-3}, respectively, worse than our method. Because SAP does not return the uncertainty index, we do not report the WNEE for SAP. Fig. 9-Fig. 10 show the recovery performance of the cases 1 and 2. We visualize the corrupted data, recovered data, the confidence interval of one channel, and the uncertainty index of the corresponding channel in each subfigure, respectively. The 95% confidence interval for each time instant is the predictive mean plus and minus 1.96 times the predictive standard deviation. In both cases, the ground-truth measurements are located within the confidence interval. At time 2.3 seconds when the event happens, the uncertainty index increases because the method is less confident about the estimation at that time instant. Moreover, in Fig. 9(c), the uncertainty index increases during the time interval 5.6-6.6 seconds, which corresponds to the time when additional noise is introduced. Fig. 9(b) shows that the noise is reduced in the recovery results.

5 CONCLUSIONS

This paper develops a Bayesian low-rank Hankel matrix recovery method to address missing and bad data in synchrophasor measurements. It provides the uncertainty index for the operator to evaluate the estimation accuracy of recovered data in real-time. The method outperforms all the existing methods numerically.

One future direction is to explore the Bayesian tensor matrix completion method by exploiting the Hankel structure. We will also investigate the theoretical guarantee of uncertainty modeling in robust matrix completion problem.

REFERENCES

Farrokh Aminifar, Mohammad Shahidehpour, Mahmud Fotuhi-Firuzabad, and Saeed Kamalinia. 2013. Power system dynamic state estimation with synchronized phasor measurements. *IEEE Transactions on Instrumentation and Measurement* 63, 2 (2013), 352–363.

S Derin Babacan, Martin Luessi, Rafael Molina, and Aggelos K Katsaggelos. 2012. Sparse Bayesian methods for low-rank matrix estimation. *IEEE Transactions on Signal Processing* 60, 8 (2012), 3964–3977.

Pratyasa Bhui and Nilanjan Senroy. 2016. Online Identification of Tripped Line for Transient Stability Assessment. *IEEE Transactions on Power Systems* 31, 3 (2016), 2214–2224.

Christopher M Bishop. 2006. Pattern recognition. *Machine learning* 128, 9 (2006).

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association* 112, 518 (2017), 859–877.

Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. 2019. Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences* 116, 46 (2019), 22931–22937.

Yeshwanth Cherapanamjeri, Kartik Gupta, and Prateek Jain. 2017. Nearly Optimal Robust Matrix Completion. In *International Conference on Machine Learning*. 797–805.

Ahmad Salehi Dobakhshari, Mohammad Abdolmaleki, Vladimir Terzija, and Sadegh Azizi. 2020. Robust Hybrid Linear State Estimator Utilizing SCADA and PMU Measurements. *IEEE Transactions on Power Systems* 36, 2 (2020), 1264–1273.

Mohammad Esmalifalak, Huy Nguyen, Rong Zheng, Le Xie, Lingyang Song, and Zhu Han. 2015. A stealthy attack against electricity market using independent component analysis. *IEEE Systems Journal* 12, 1 (2015), 297–307.

Brandon Foggo and Nanpeng Yu. 2021. Online PMU Missing Value Replacement via Event-Participation Decomposition. *IEEE Transactions on Power Systems* (2021), 1–1.

Pengzhi Gao, Meng Wang, Joe H. Chow, Scott G. Ghiocel, Bruce Fardanesh, George Stefopoulos, and Michael P. Razanousky. 2016a. Identification of Successive "Unobservable" Cyber Data Attacks in Power Systems. *IEEE Transactions on Signal Processing* 64, 21 (Nov. 2016), 5557–5570.

Pengzhi Gao, Meng Wang, Scott G. Ghiocel, Joe H. Chow, Bruce Fardanesh, and George Stefopoulos. March 2016b. Missing Data Recovery by Exploiting Low-dimensionality in Power System Synchrophasor Measurements. *IEEE Transactions on Power Systems* 31, 2 (March 2016), 1006–1013.

Cristian Genes, Inaki Esnaola, Samir M Perlaza, Luis F Ochoa, and Daniel Coca. 2018. Robust recovery of missing data in electricity distribution systems. *IEEE Transactions on Smart Grid* 10, 4 (2018), 4057–4067.

Tingyan Guo and Jovica V Milanović. 2015. Online identification of power system dynamic signature using PMU measurements and data mining. *IEEE Transactions on Power Systems* 31, 3 (2015), 1760–1768.

Yingshuai Hao, Meng Wang, Joe H. Chow, Evangelos Farantatos, and Mahendra Patel. 2018. Model-less Data Quality Improvement of Streaming Synchrophasor Measurements by Exploiting the Low-Rank Hankel Structure. *IEEE Transactions on Power Systems* 33, 6 (June 2018), 6966–6977.

Tong Huang, Bharadwaj Satchidanandan, PR Kumar, and Le Xie. 2018. An online detection framework for cyber attacks on automatic generation control. *IEEE Transactions on Power Systems* 33, 6 (2018), 6816–6827.

JQ James, David J Hill, Victor OK Li, and Yunhe Hou. 2019. Synchrophasor recovery and prediction: A graph-based deep learning approach. *IEEE Internet of Things Journal* 6, 5 (2019), 7348–7359.

JQ James, Albert YS Lam, David J Hill, Yunhe Hou, and Victor OK Li. 2018. Delay aware power system synchrophasor recovery and prediction framework. *IEEE Transactions on Smart Grid* 10, 4 (2018), 3732–3742.

Kevin D Jones, Anamitra Pal, and James S Thorp. 2014. Methodology for performing synchrophasor data conditioning and validation. *IEEE Transactions on Power Systems* 30, 3 (2014), 1121–1130.

Innocent Kamwa and Luc Gerin-Lajoie. 2000. State-space system identification-toward MIMO models for modal analysis and optimization of bulk power systems. *IEEE Transactions on Power Systems* 15, 1 (2000), 326–335.

Oliver Kosut, Liyan Jia, Robert J Thomas, and Lang Tong. 2011. Malicious data attacks on the smart grid. *IEEE Transactions on Smart Grid* 2, 4 (2011), 645–658.

Mang Liao, Di Shi, Zhe Yu, Zhehan Yi, Zhiwei Wang, and Yingmeng Xiang. 2018. An alternating direction method of multipliers based approach for PMU data recovery. *IEEE Transactions on Smart Grid* 10, 4 (2018), 4554–4565.

Yao Liu, Peng Ning, and Michael K Reiter. 2011. False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security (TISSEC)* 14, 1 (2011), 13.

Jaakko Luttinen, Alexander Ilin, and Juha Karhunen. 2012. Bayesian robust PCA of incomplete data. *Neural processing letters* 36, 2 (2012), 189–202.

Kursat Rasim Mestav, Jaime Luengo-Rozas, and Lang Tong. 2019. Bayesian state estimation for unobservable distribution systems via deep learning. *IEEE Transactions on Power Systems* 34, 6 (2019), 4910–4920.

Kursat Rasim Mestav and Lang Tong. 2020. Universal data anomaly detection via inverse generative adversary network. *IEEE Signal Processing Letters* 27 (2020), 511–515.

Denis Osipov and Joe H Chow. 2020. PMU missing data recovery using tensor decomposition. *IEEE Transactions on Power Systems* 35, 6 (2020), 4554–4563.

John Paisley, David M Blei, and Michael I Jordan. 2012. Variational Bayesian inference with stochastic search. In *International Conference on Machine Learning*. 1363–1370.

Chao Ren and Yan Xu. 2019. A fully data-driven method based on generative adversarial networks for power system dynamic security assessment with missing data. *IEEE Transactions on Power Systems* 34, 6 (2019), 5044–5052.

Xinan Wang, Di Shi, Jianhui Wang, Zhe Yu, and Zhiwei Wang. 2019. Online identification and data recovery for PMU data manipulation attack. *IEEE Transactions on Smart Grid* 10, 6 (2019), 5889–5898.

Meng Wu and Le Xie. 2016. Online detection of low-quality synchrophasor measurements: A data-driven approach. *IEEE Transactions on Power Systems* 32, 4 (2016), 2817–2827.

Ming Yi and Meng Wang. 2021. Bayesian Energy Disaggregation At Substations With Uncertainty Modeling. *IEEE Transactions on Power Systems* 37, 1 (2021), 764–775.

Shuai Zhang, Yingshuai Hao, Meng Wang, and Joe H Chow. 2018. Multichannel Hankel matrix completion through nonconvex optimization. *IEEE Journal of Selected Topics in Signal Processing* 12, 4 (2018), 617–632.

Shuai Zhang and Meng Wang. 2018. Correction of simultaneous bad measurements by exploiting the low-rank hankel structure. In *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 646–650.

Shuai Zhang and Meng Wang. 2019. Correction of Corrupted Columns through Fast Robust Hankel Matrix Completion. *IEEE Transactions on Signal Processing* 67, 10 (2019), 2580–2594.

Junbo Zhao, Gexiang Zhang, Kaushik Das, George N Korres, Nikolaos M Manousakis, Avinash K Sinha, and Zhengyou He. 2015a. Power system real-time monitoring by using PMU-based robust state estimation method. *IEEE Transactions on Smart Grid* 7, 1 (2015), 300–309.

Qibin Zhao, Guoxu Zhou, Liqing Zhang, Andrzej Cichocki, and Shun-Ichi Amari. 2015b. Bayesian robust tensor factorization for incomplete multiway data. *IEEE Transactions on Neural Networks and Learning Systems* 27, 4 (2015), 736–748.

Yuxuan Zhao and Madeleine Udell. 2020. Matrix Completion with Quantified Uncertainty through Low Rank Gaussian Copula. *Advances in Neural Information Processing Systems* 33 (2020).

Mingyuan Zhou, Haojun Chen, John Paisley, Lu Ren, Guillermo Sapiro, and Lawrence Carin. 2009. Non-parametric Bayesian dictionary learning for sparse image representations. In *Advances in Neural Information Processing Systems*. 2295–2303.

Ning Zhou, Da Meng, Zhenyu Huang, and Greg Welch. 2014. Dynamic state estimation of a synchronous machine using PMU data: A comparative study. *IEEE Transactions on Smart Grid* 6, 1 (2014), 450–460.

Ning Zhou, John W Pierre, and John F Hauer. 2006. Initial results in power system identification from injected probing signals using a subspace method. *IEEE Transactions on Power Systems* 21, 3 (2006), 1296–1302.

SUPPLEMENTARY MATERIAL

A.1 Beta and Gamma distributions

The Beta and Gamma distributions are introduced here.

The Gamma function is defined as

$$\Gamma(\alpha_1) = \int_0^\infty x^{\alpha_1-1}e^{-x}dx. \tag{50}$$

The Beta distribution is

$$\text{Beta}(\pi_k|\alpha_1,\beta_1) = \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)}(\pi_k)^{\alpha_1-1}(1-\pi_k)^{\beta_1-1}, \tag{51}$$

The "Beta$(\pi_k|\alpha_1,\beta_1)$" denotes that π_k is a Beta distribution with two parameters α_1 and β_1. Other notations have same rule in the following section. The mean of this Beta distribution is $\frac{\alpha_1}{\alpha_1+\beta_1}$ and the variance of this Beta distribution is $\frac{\alpha_1\beta_1}{(\alpha_1+\beta_1)^2(\alpha_1+\beta_1+1)}$.

The Gamma distribution is

$$\Gamma(\gamma_s|c_0,d_0) = \frac{{d_0}^{c_0}(\gamma_s)^{c_0-1}e^{-d_0\gamma_\epsilon}}{\Gamma(c_0)} \propto (\gamma_s)^{c_0-1}e^{-d_0\gamma_s}, \tag{52}$$

where $c_0 > 0, d_0 > 0$. $\propto$ denotes "proportional to". The mean of this Gamma distribution is $\frac{c_0}{d_0}$ and the variance of this Gamma distribution is $\frac{c_0}{{d_0}^2}$.

A.2 The Hankel operator

$\mathcal{H}^\dagger$ is the Moore-Penrose pseudoinverse of $\mathcal{H}$. For any $\boldsymbol{X} \in \mathbb{R}^{mn_2\times n_1}$, $(\mathcal{H}^\dagger\boldsymbol{X})_{i,j} \in \mathbb{R}^{m\times n}$ is defined as

$$\begin{aligned}(\mathcal{H}^\dagger\boldsymbol{X})_{i,j} &= \langle\mathcal{H}^\dagger\boldsymbol{X}, e_ie_j^T\rangle = \frac{1}{\kappa_j}\sum_{\frac{u-i}{m}+v=j} X_{u,v} \\ &= \begin{cases}\frac{1}{\kappa_j}\sum_{j_1=1}^{j} X_{(j_1-1)m+i,j+1-j_1} & j \le n_2 \\ \frac{1}{\kappa_j}\sum_{j_2=j+1-n_2}^{n_j} X_{(j-j_2)m+i,j_2} & j \ge n_2+1\end{cases},\end{aligned} \tag{53}$$

where $\kappa_j = \#\{(j_1,j_2)|j_1+j_2 = j+1 \quad 1 \le j_1 \le n_2, 1 \le j_2 \le n_1\}$ is the number of entries in the jth anti-diagonal of an $n_2 \times n_1$ matrix. $n_j = \min(j, n_1)$.

We employ two sets to define the mapping relationship between the original matrix and the corresponding Hankel matrix. (i,j) denotes one coordinate in the original matrix. $\Psi_{i,j}$ denotes the set of the mapping entries of (i,j) in the corresponding Hankel matrix of $\boldsymbol{Y^o}$. $\Psi_{i,j}$ shows the mapping relationship in (53). One can simply check the corresponding coordinate of (i,j) in $\boldsymbol{X}$ to obtain $\Psi_{i,j}$. $\Psi_{i,j}$ is a subset of Ψ_Ω. $\Psi_{i,j}$ only shows the mapping set of one point (i,j) while Ψ_Ω denotes the set of all the mapping entries of all (i,j) in the corresponding Hankel matrix.

$$\begin{aligned}&\Psi_{i,j} = \{(u,v)|(u,v) = ((j_1-1)m+i, j+1-j_1) \text{ for every} \\ &j_1 = 1,2,...,j, \text{ under the case when } j \le n_2; \\ &(u,v) = ((j-j_2)m+i, j_2) \text{ for every } j_2 = j+1-n_2,...,n_j, \\ &\text{where } n_j = \min(j,n_1), \text{ under the case when } j \ge n_2+1;\}\ (i,j) \in \Omega.\end{aligned} \tag{54}$$

$$\Psi_\Omega = \{(u,v)| \text{ there exists } (i,j) \in \Omega \text{ such that } (u,v) \in \Psi_{i,j}\}. \tag{55}$$

A.3 Updating rule for variational inference

Algorithm 1 Varational Inference for Bayesian Robust Hankel Matrix Completion

Require: The observation matrix $\boldsymbol{Y^o}$. The parameters λ_d, $a_0, b_0, c_0, d_0, e_0, f_0, g_0, h_0$ for prior distributions. The initial basis size K . The maximum iterations $T_{\max}$. The convergence threshold ξ. The Hankel matrix parameter n_2.

1: **Initialization**:Form the Hankel matrix $\boldsymbol{X}$ by $\boldsymbol{Y^o}$. Take the SVD of $\boldsymbol{X}$ by $\boldsymbol{X} = \boldsymbol{UAV}^T$. $\boldsymbol{D}$ is initialized by $\boldsymbol{UA}^{\frac{1}{2}}$. $\boldsymbol{S}$ is initialized with $\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{V}^T$. $\boldsymbol{Z}$ is initialized with all-one matrix. All values in π_k are initialized as 0.5. γ_s and λ_d are initialized by $\frac{||\boldsymbol{Y^o}||_F^2}{mn}$. $1/\gamma_\epsilon$ is initialized by $\frac{||\boldsymbol{Y^o}||_F}{\sqrt{mn}}$. $\bar{\boldsymbol{X}}^0 = \boldsymbol{D}(\boldsymbol{S}\odot\boldsymbol{Z})$. The initial $\boldsymbol{E}$ is $E_{i,j} = Y^o_{i,j} - (\mathcal{H}^\dagger\bar{\boldsymbol{X}}^0)_{i,j}$ if $(i,j) \in \Omega$ and $E_{i,j} = 0$ otherwise. All the elements in $\boldsymbol{\beta}$ are initialized as $\frac{||\boldsymbol{Y^o}||_F}{\sqrt{mn}}$. $\eta = 1, t = 1$.
2: **while** $\eta > \xi$ and $t < T_{\max}$ **do**
3: Compute $\mathbb{E}[\boldsymbol{d}_{p\cdot}]$ from $q(\boldsymbol{d}_{p\cdot})$ by (23) for each $p = 1,2,3,...,mn_2$;
4: Compute $\mathbb{E}[\boldsymbol{s}_{\cdot q}]$ from $q(\boldsymbol{s}_{\cdot q})$ by (26) for all q;
5: **for** $k = 1,2,3,...,K$ **do**
6: Compute $\mathbb{E}[z_{kq}]$ from $q(z_{kq})$ by (28) for all q;
7: Compute $\mathbb{E}[\ln(\pi_k)]$ and $\mathbb{E}[\ln(1-\pi_k)]$ from $q(\pi_k)$ by (33) and (34);
8: **end for**
9: Compute $\mathbb{E}[\gamma_s]$ from $q(\gamma_s)$ by (36);
10: Compute $\mathbb{E}[E_{i,j}]$ from $q(E_{i,j})$ by (39) for all $(i,j) \in \Omega$;
11: Compute $\mathbb{E}[\beta_{i,j}]$ from $q(\beta_{i,j})$ by (41);
12: Compute $\mathbb{E}[\gamma_\epsilon]$ from $q(\gamma_\epsilon)$ by (43);
13: **if** $\mathbb{E}[z_{kq}] = 0$ for all k **then**
14: Remove $\mathbb{E}[\boldsymbol{d}_{\cdot k}]$ in $\mathbb{E}[\boldsymbol{D}]$, $\mathbb{E}[\pi_k]$, and $\mathbb{E}[z_{kq}]$, $\mathbb{E}[s_{kq}]$ for all q;
15: $K = K-1$;
16: **end if**
17: $\bar{\boldsymbol{X}} = \boldsymbol{D}(\boldsymbol{S}\odot\boldsymbol{Z})$;
18: $\eta = \frac{\|\bar{\boldsymbol{X}}-\bar{\boldsymbol{X}}_{\text{pre}}\|_F}{\|\bar{\boldsymbol{X}}_{\text{pre}}\|_F}$;
19: $\bar{\boldsymbol{X}}_{\text{pre}} = \bar{\boldsymbol{X}}$;
20: $t = t+1$;
21: **end while**
22: Estimate the predictive mean $\mathbb{E}[Y_{i,j}]$ and variance $\text{Var}[Y_{i,j}]$ by (45) and (46) for all (i,j).
23: **return** The predictive mean $\mathbb{E}[Y_{i,j}]$ and predictive variance $\text{Var}[Y_{i,j}]$ for each entry.

The details of our proposed approach are summarized in Algorithm 1. Algorithm 1 can be simplified if only missing data presents. One can skip lines 10-11 in Algorithm 1 about updating $\mathbb{E}[E_{i,j}]$ and $\mathbb{E}[\beta_{i,j}]$, and all the other updating rules remain unchanged.

The KL divergence in (19) is difficult to compute because computing $P(\boldsymbol{Y}_\Omega^o)$ is intractable. To see this,

$$\begin{aligned}
&\mathbb{KL}(q(\boldsymbol{\Theta})||P(\boldsymbol{\Theta},\boldsymbol{Y}|\boldsymbol{Y}_\Omega^o)) \\
&= -\int q(\boldsymbol{\Theta})\ln\frac{P(\boldsymbol{\Theta},\boldsymbol{Y}|\boldsymbol{Y}_\Omega^o)}{q(\boldsymbol{\Theta})}d\boldsymbol{\Theta} \\
&= \mathbb{E}[\ln q(\boldsymbol{\Theta})] - \mathbb{E}[\ln P(\boldsymbol{\Theta},\boldsymbol{Y}|\boldsymbol{Y}_\Omega^o)] \\
&= \mathbb{E}[\ln q(\boldsymbol{\Theta})] - \mathbb{E}[\ln P(\boldsymbol{\Theta},\boldsymbol{Y},\boldsymbol{Y}_\Omega^o)] + \ln(P(\boldsymbol{Y}_\Omega^o)) \\
&= -(\mathbb{E}[\ln P(\boldsymbol{\Theta},\boldsymbol{Y},\boldsymbol{Y}_\Omega^o)] - \mathbb{E}[\ln q(\boldsymbol{\Theta})]) + \ln(P(\boldsymbol{Y}_\Omega^o)) \\
&= -\text{ELBO}(q(\boldsymbol{\Theta})) + \ln\left(P(\boldsymbol{Y}_\Omega^o)\right).
\end{aligned} \tag{56}$$

ELBO is evidence lower bound. $\ln\left(P(\boldsymbol{Y}_\Omega^o)\right)$ denotes the natural logarithm of $P(\boldsymbol{Y}_\Omega^o)$. The expectations in (56) are taken with respect to $q(\boldsymbol{\Theta})$. Because $\ln(P(\boldsymbol{Y}_\Omega^o))$ is not related to $q(\boldsymbol{\Theta})$, minimizing the KL divergence is equivalent to maximizing the ELBO. The goal of variational inference is changed to maximizing the ELBO.

The joint probability of observed data and all the parameters is characterized by (57).

$$\begin{aligned}
&P(\boldsymbol{\Theta},\boldsymbol{Y},\boldsymbol{Y}_\Omega^o) \\
=&p(\boldsymbol{Y}_\Omega^o|\boldsymbol{D},\boldsymbol{S},\boldsymbol{Z},\boldsymbol{E},\gamma_\epsilon)p(\boldsymbol{D}|\lambda_d)p(\boldsymbol{S}|\gamma_s)p(\boldsymbol{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi}) \\
&p(\boldsymbol{E}|\boldsymbol{\beta})p(\boldsymbol{\beta})p(\gamma_s)p(\gamma_\epsilon) \\
=&\prod_{(i,j)\in\Omega}\mathcal{N}(Y_{i,j}^o|(\mathcal{H}^\dagger\boldsymbol{X})_{i,j}+E_{i,j},\frac{1}{\gamma_\epsilon})\Gamma(\beta_{i,j}|g_0,h_0) \\
&\prod_{q=1}^{n_1}\mathcal{N}(\boldsymbol{s}_{\cdot q}|0,\frac{1}{\gamma_s}\boldsymbol{I}_K) \\
&\prod_{p=1}^{mn_2}\mathcal{N}(\boldsymbol{d}_{p\cdot}|0,\frac{1}{\lambda_d}\boldsymbol{I}_K)\prod_{k=1}^{K}\text{Beta}(\pi_k|a_0,b_0) \\
&\prod_{q=1}^{n_1}\prod_{k=1}^{K}\text{Bernoulli}(z_{kq}|\pi_k) \\
&\Gamma(\gamma_s|c_0,d_0)\Gamma(\gamma_\epsilon|e_0,f_0).
\end{aligned} \tag{57}$$

$\mathcal{N}(Y_{i,j}^o|(\mathcal{H}^\dagger\boldsymbol{X})_{i,j}+E_{i,j},\frac{1}{\gamma_\epsilon})$ denotes that $Y_{i,j}^o$ follows a Gaussian distribution with mean $(\mathcal{H}^\dagger\boldsymbol{X})_{i,j}+E_{i,j}$ and variance $\frac{1}{\gamma_\epsilon}$ when $\boldsymbol{D}$, $\boldsymbol{Z}$, $\boldsymbol{S}$, $\boldsymbol{E}$ and γ_ϵ are given.

The derivation details of updating rules of variational inference are shown below.

(I) The approximate posterior distribution of $\boldsymbol{d}_{p\cdot}$ is a Gaussian distribution (for all $p = 1, ..., mn_2$).

To see this, note that

$$\begin{aligned}
&\prod_{(i,j)\in\Omega}\mathcal{N}(Y_{i,j}^o|(\mathcal{H}^\dagger\boldsymbol{X})_{i,j}+E_{i,j},\frac{1}{\gamma_\epsilon}) \\
&\propto\prod_{(i,j)\in\Omega}\exp(\frac{-\gamma_\epsilon}{2}(Y_{i,j}^o-E_{i,j}-(\mathcal{H}^\dagger\boldsymbol{X})_{i,j})^2) \\
&\propto\prod_{(i,j)\in\Omega}\exp(\frac{-\gamma_\epsilon}{2}(Y_{i,j}^o-E_{i,j}-\frac{1}{\kappa_j}\sum_{(u,v)\in\Psi_{i,j}}[\boldsymbol{d}_{u\cdot}(\boldsymbol{s}_{\cdot v}\odot\boldsymbol{z}_{\cdot v})])^2) \\
&\overset{\propto}{\sim}\prod_{(i,j)\in\Omega}\exp(\tfrac{-\gamma_\epsilon}{2}(Y_{i,j}^o-E_{i,j}-\tfrac{1}{\kappa_j}\kappa_j[\boldsymbol{d}_{p\cdot}(\boldsymbol{s}_{\cdot q}\odot\boldsymbol{z}_{\cdot q})]_{(p,q)\in\Psi_{i,j}})^2) \\
&\propto\prod_{(i,j)\in\Omega}\exp(\frac{-1}{2}[\boldsymbol{d}_{p\cdot}\gamma_\epsilon(\boldsymbol{s}_{\cdot q}\odot\boldsymbol{z}_{\cdot q})(\boldsymbol{s}_{\cdot q}\odot\boldsymbol{z}_{\cdot q})^T\boldsymbol{d}_{p\cdot}^T]_{(p,q)\in\Psi_{i,j}} \\
&+\gamma_\epsilon(Y_{i,j}^o-E_{i,j})[(\boldsymbol{s}_{\cdot q}\odot\boldsymbol{z}_{\cdot q})^T\boldsymbol{d}_{p\cdot}^T]_{(p,q)\in\Psi_{i,j}}-\frac{\gamma_\epsilon}{2}(Y_{i,j}^o-E_{i,j})^2) \\
&\propto\exp(\boldsymbol{d}_{p\cdot}\frac{-\gamma_\epsilon}{2}\sum_{q:(p,q)\in\Psi_\Omega}(\boldsymbol{s}_{\cdot q}\odot\boldsymbol{z}_{\cdot q})(\boldsymbol{s}_{\cdot q}\odot\boldsymbol{z}_{\cdot q})^T\boldsymbol{d}_{p\cdot}^T \\
&+\gamma_\epsilon\textstyle\sum_{q:(p,q)\in\Psi_\Omega}\mathcal{H}(Y^o-E)_{p,q}(\boldsymbol{s}_{\cdot q}\odot\boldsymbol{z}_{\cdot q})^T\boldsymbol{d}_{p\cdot}^T-\frac{\gamma_\epsilon}{2}\sum_{q:(p,q)\in\Psi_\Omega}\mathcal{H}(Y^o-E)_{p,q}^2).
\end{aligned} \tag{58}$$

where $\mathcal{H}(Y^o-E)_{p,q}$ represents entry (p,q) of the Hankel matrix $\mathcal{H}(Y^o-E)$.

Also note that

$$\mathcal{N}(\boldsymbol{d}_{p\cdot}|0,\frac{1}{\lambda_d}\boldsymbol{I}_K)\propto\exp(-\frac{\lambda_d}{2}\boldsymbol{d}_{p\cdot}\boldsymbol{d}_{p\cdot}^T). \tag{59}$$

Therefore,

$$\begin{aligned}
&\prod_{(i,j)\in\Omega}\mathcal{N}(Y_{i,j}^o|(\mathcal{H}^\dagger\boldsymbol{X})_{i,j}+E_{i,j},\frac{1}{\gamma_\epsilon})\mathcal{N}(\boldsymbol{d}_{p\cdot}|0,\frac{1}{\lambda_d}\boldsymbol{I}_K) \\
&\propto\exp(-\frac{1}{2}\boldsymbol{d}_{p\cdot}(\gamma_\epsilon\sum_{q:(p,q)\in\Psi_\Omega}(\boldsymbol{s}_{\cdot q}\odot\boldsymbol{z}_{\cdot q})(\boldsymbol{s}_{\cdot q}\odot\boldsymbol{z}_{\cdot q})^T+\lambda_d\boldsymbol{I}_K)\boldsymbol{d}_{p\cdot}^T \\
&+\gamma_\epsilon\textstyle\sum_{q:(p,q)\in\Psi_\Omega}\mathcal{H}(Y^o-E)_{p,q}(\boldsymbol{s}_{\cdot q}\odot\boldsymbol{z}_{\cdot q})^T\boldsymbol{d}_{p\cdot}^T-\frac{\gamma_\epsilon}{2}\sum_{q:(p,q)\in\Psi_\Omega}\mathcal{H}(Y^o-E)_{p,q}^2).
\end{aligned} \tag{60}$$

Now it is safe to write that

$$\begin{aligned}
&\ln(\ q(\boldsymbol{d}_{p\cdot})) \\
&=\mathbb{E}_{\boldsymbol{\Theta}\setminus\boldsymbol{d}_{p\cdot}}[\ln p(\boldsymbol{\Theta},\boldsymbol{Y},\boldsymbol{Y}_\Omega^o)]+\text{const.} \\
&=\mathbb{E}_{\boldsymbol{\Theta}\setminus\boldsymbol{d}_{p\cdot}}[\ln p(\boldsymbol{Y}_\Omega^o|\boldsymbol{D},\boldsymbol{S},\boldsymbol{Z},\boldsymbol{E},\gamma_\epsilon)P(\boldsymbol{D}|\lambda_d)]+\text{const.} \\
&=\mathbb{E}[\ln\textstyle\prod_{(i,j)\in\Omega}\mathcal{N}(Y_{i,j}^o|(\mathcal{H}^\dagger\boldsymbol{X})_{i,j}+E_{i,j},\frac{1}{\gamma_\epsilon})\mathcal{N}(\boldsymbol{d}_{p\cdot}|0,\frac{1}{\lambda_d}\boldsymbol{I}_K)]+\text{const.} \\
&=\mathbb{E}[-\frac{1}{2}\boldsymbol{d}_{p\cdot}(\gamma_\epsilon\sum_{q:(p,q)\in\Psi_\Omega}(\boldsymbol{s}_{\cdot q}\odot\boldsymbol{z}_{\cdot q})(\boldsymbol{s}_{\cdot q}\odot\boldsymbol{z}_{\cdot q})^T+\lambda_d\boldsymbol{I}_K)\boldsymbol{d}_{p\cdot}^T \\
&+\gamma_\epsilon\textstyle\sum_{q:(p,q)\in\Psi_\Omega}\mathcal{H}(Y^o-E)_{p,q}(\boldsymbol{s}_{\cdot q}\odot\boldsymbol{z}_{\cdot q})^T\boldsymbol{d}_{p\cdot}^T-\frac{\gamma_\epsilon}{2}\sum_{q:(p,q)\in\Psi_\Omega}\mathcal{H}(Y^o-E)_{p,q}^2]+\text{const.} \\
&=-\frac{1}{2}\boldsymbol{d}_{p\cdot}(\mathbb{E}[\gamma_\epsilon]\sum_{q:(p,q)\in\Psi_\Omega}\mathbb{E}[(\boldsymbol{s}_{\cdot q}\odot\boldsymbol{z}_{\cdot q})(\boldsymbol{s}_{\cdot q}\odot\boldsymbol{z}_{\cdot q})^T]+\lambda_d\boldsymbol{I}_K)\boldsymbol{d}_{p\cdot}^T \\
&+\mathbb{E}[\gamma_\epsilon]\sum_{q:(p,q)\in\Psi_\Omega}\mathcal{H}(Y^o-\mathbb{E}[\boldsymbol{E}])_{p,q}(\mathbb{E}[\boldsymbol{s}_{\cdot q}]\odot\mathbb{E}[\boldsymbol{z}_{\cdot q}])^T\boldsymbol{d}_{p\cdot}^T \\
&-\frac{\mathbb{E}[\gamma_\epsilon]}{2}\sum_{q:(p,q)\in\Psi_\Omega}\mathcal{H}(Y^o-\mathbb{E}[\boldsymbol{E}])_{p,q}^2+\text{const.}.
\end{aligned} \tag{61}$$

The above derivation reveals that $q(\boldsymbol{d}_{p\cdot})$ is a Gaussian distribution with mean $\mathbb{E}[\boldsymbol{d}_{p\cdot}]$ and covariance $\Sigma_{\boldsymbol{d}_{p\cdot}}$, i.e.,

$$q(\boldsymbol{d}_{p\cdot})\sim\mathcal{N}(\mathbb{E}[\boldsymbol{d}_{p\cdot}],\Sigma_{\boldsymbol{d}_{p\cdot}}), \tag{62}$$

where

$$\Sigma_{\boldsymbol{d}_{p\cdot}} = [\mathbb{E}[\gamma_\epsilon] \sum_{q:(p,q)\in\Psi_\Omega} \mathbb{E}[(\boldsymbol{s}_{\cdot q} \odot \boldsymbol{z}_{\cdot q})(\boldsymbol{s}_{\cdot q} \odot \boldsymbol{z}_{\cdot q})^T] + \lambda_d \boldsymbol{I}_K]^{-1}, \quad (63)$$

$$\mathrm{E}[\boldsymbol{d}_{p\cdot}] = \mathbb{E}[\gamma_\epsilon] \sum_{q:(p,q)\in\Psi_\Omega} \mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})_{p,q} (\mathbb{E}[\boldsymbol{s}_{\cdot q}]^T \odot \mathbb{E}[\boldsymbol{z}_{\cdot q}]^T) \Sigma_{\boldsymbol{d}_{p\cdot}}. \quad (64)$$

The required expectation is

$$\begin{aligned} &\mathbb{E}[(\boldsymbol{s}_{\cdot q} \odot \boldsymbol{z}_{\cdot q})(\boldsymbol{s}_{\cdot q} \odot \boldsymbol{z}_{\cdot q})^T] = \mathbb{E}[\boldsymbol{s}_{\cdot q}\boldsymbol{s}_{\cdot q}^T] \odot \mathbb{E}[\boldsymbol{z}_{\cdot q}\boldsymbol{z}_{\cdot q}^T] \\ &= (\mathbb{E}[\boldsymbol{s}_{\cdot q}]\mathbb{E}[\boldsymbol{s}_{\cdot q}]^T + \Sigma_{\boldsymbol{s}_{\cdot q}}) \odot (\mathbb{E}[\boldsymbol{z}_{\cdot q}]\mathbb{E}[\boldsymbol{z}_{\cdot q}]^T + \Sigma_{\boldsymbol{z}_{\cdot q}}), \end{aligned} \quad (65)$$

where $\Sigma_{\boldsymbol{z}_{\cdot q}}$ is

$$\Sigma_{\boldsymbol{z}_{\cdot q}} = \mathrm{diag}[\mathbb{E}[z_{1q}](1 - \mathbb{E}[z_{1q}]), ..., \mathbb{E}[z_{Kq}](1 - \mathbb{E}[z_{Kq}])]. \quad (66)$$

(II) The approximate posterior distribution of $\boldsymbol{s}_{\cdot q}$ ($q = 1, ..., n_1$) is a Gaussian distribution.
To see this, note that

$$\begin{aligned} &\prod_{(i,j)\in\Omega} \mathcal{N}(Y^o_{i,j} | (\mathcal{H}^\dagger \boldsymbol{X})_{i,j}, \frac{1}{\gamma_\epsilon}) \\ &\propto \prod_{(i,j)\in\Omega} \exp(\frac{-\gamma_\epsilon}{2}(Y^o_{i,j} - E_{i,j} - (\mathcal{H}^\dagger \boldsymbol{X})_{i,j})^2) \\ &\propto \prod_{(i,j)\in\Omega} \exp(\frac{-\gamma_\epsilon}{2}(Y^o_{i,j} - E_{i,j} - \frac{1}{\kappa_j} \sum_{(u,v)\in\Psi_{i,j}} [\boldsymbol{d}_{u\cdot}(\boldsymbol{s}_{\cdot v} \odot \boldsymbol{z}_{\cdot v})])^2) \\ &\overset{\propto}{\sim} \prod_{(i,j)\in\Omega} \exp(\frac{-\gamma_\epsilon}{2}(Y^o_{i,j} - E_{i,j} - \frac{1}{\kappa_j}\kappa_j [\boldsymbol{d}_{p\cdot}(\boldsymbol{s}_{\cdot q} \odot \boldsymbol{z}_{\cdot q})]_{(p,q)\in\Psi_{i,j}})^2) \\ &\propto \prod_{(i,j)\in\Omega} \exp(\frac{-\gamma_\epsilon}{2}([\boldsymbol{s}_{\cdot q}^T(\boldsymbol{z}_{\cdot q} \odot \boldsymbol{d}_{p\cdot}^T)(\boldsymbol{d}_{p\cdot} \odot \boldsymbol{z}_{\cdot q}^T)\boldsymbol{s}_{\cdot q}]_{(p,q)\in\Psi_{i,j}} \\ &- 2(Y^o_{i,j} - E_{i,j})[(\boldsymbol{d}_{p\cdot} \odot \boldsymbol{z}_{\cdot q}^T)\boldsymbol{s}_{\cdot q}]_{(p,q)\in\Psi_{i,j}} + (Y^o_{i,j} - E_{i,j})^2)) \\ &\propto \prod_{(i,j)\in\Omega} \exp(\frac{-\gamma_\epsilon}{2}([\boldsymbol{s}_{\cdot q}^T \boldsymbol{\phi}_{p,q}^T \boldsymbol{\phi}_{p,q} \boldsymbol{s}_{\cdot q}]_{(p,q)\in\Psi_{i,j}} \\ &- 2(Y^o_{i,j} - E_{i,j})[\boldsymbol{\phi}_{p,q}\boldsymbol{s}_{\cdot q}]_{(p,q)\in\Psi_{i,j}} + (Y^o_{i,j} - E_{i,j})^2)) \\ &\propto \exp(\frac{-\gamma_\epsilon}{2}(\boldsymbol{s}_{\cdot q}^T \sum_{p:(p,q)\in\Psi_\Omega} \boldsymbol{\phi}_{p,q}^T \boldsymbol{\phi}_{p,q} \boldsymbol{s}_{\cdot q}) \\ &+\gamma_\epsilon \textstyle\sum_{p:(p,q)\in\Psi_\Omega} \mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})_{p,q}[\boldsymbol{\phi}_{p,q}\boldsymbol{s}_{\cdot q}] - \frac{\gamma_\epsilon}{2} \sum_{p:(p,q)\in\Psi_\Omega} \mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})^2_{p,q}), \end{aligned} \quad (67)$$

where we define $\boldsymbol{\phi}_{p,q} = \boldsymbol{d}_{p\cdot} \odot \boldsymbol{z}_{\cdot q}^T$.
Also note that

$$\mathcal{N}(\boldsymbol{s}_{\cdot q} | 0, \frac{1}{\gamma_s}\boldsymbol{I}_K) \propto \exp(\frac{-\gamma_s}{2}(\boldsymbol{s}_{\cdot q}^T \boldsymbol{s}_{\cdot q})). \quad (68)$$

Therefore,

$$\begin{aligned} &\prod_{(i,j)\in\Omega} \mathcal{N}(Y^o_{i,j} | (\mathcal{H}^\dagger \boldsymbol{X})_{i,j} + E_{i,j}, \frac{1}{\gamma_\epsilon}) \mathcal{N}(\boldsymbol{s}_{\cdot q} | 0, \frac{1}{\gamma_s}\boldsymbol{I}_K) \\ &\propto \exp(\frac{-\gamma_\epsilon}{2}(\boldsymbol{s}_{\cdot q}^T \sum_{p:(p,q)\in\Psi_\Omega} \boldsymbol{\phi}_{p,q}^T \boldsymbol{\phi}_{p,q} \boldsymbol{s}_{\cdot q}) \\ &+ \gamma_\epsilon \mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})_{p,q}[\boldsymbol{\phi}_{p,q}\boldsymbol{s}_{\cdot q}]_{(p,q)\in\Psi_{i,j}} - \frac{\gamma_\epsilon}{2}\mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})^2_{p,q}) \\ &\exp(\frac{-\gamma_s}{2}(\boldsymbol{s}_{\cdot q}^T \boldsymbol{s}_{\cdot q})) \\ &\propto \exp(\frac{-1}{2}(\boldsymbol{s}_{\cdot q}^T [\sum_{p:(p,q)\in\Psi_\Omega} \gamma_\epsilon \boldsymbol{\phi}_{p,q}^T \boldsymbol{\phi}_{p,q} + \gamma_s \boldsymbol{I}_K]\boldsymbol{s}_{\cdot q}) \\ &+\gamma_\epsilon \textstyle\sum_{p:(p,q)\in\Psi_\Omega} \mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})_{p,q}\boldsymbol{\phi}_{p,q}\boldsymbol{s}_{\cdot q} - \frac{\gamma_\epsilon}{2} \sum_{p:(p,q)\in\Psi_\Omega} \mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})^2_{p,q}). \end{aligned} \quad (69)$$

Thus,

$$\begin{aligned} &\ln(q(\boldsymbol{s}_{\cdot q})) \\ &= \mathbb{E}_{\Theta\setminus\boldsymbol{s}_{\cdot q}}[\ln p(\Theta, Y, Y^o_\Omega)] + \text{const.} \\ &= \mathbb{E}_{\Theta\setminus\boldsymbol{s}_{\cdot q}}[\ln p(Y^o_\Omega | \boldsymbol{D}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{E}, \gamma_\epsilon) p(\boldsymbol{S}|\gamma_s)] + \text{const.} \\ &= \mathbb{E}[(\frac{-1}{2}(\boldsymbol{s}_{\cdot q}^T [\sum_{p:(p,q)\in\Psi_\Omega} \gamma_\epsilon \boldsymbol{\phi}_{p,q}^T \boldsymbol{\phi}_{p,q} + \gamma_s \boldsymbol{I}_K]\boldsymbol{s}_{\cdot q}) \\ &+\gamma_\epsilon \textstyle\sum_{p:(p,q)\in\Psi_\Omega} \mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})_{p,q}\boldsymbol{\phi}_{p,q}\boldsymbol{s}_{\cdot q} - \frac{\gamma_\epsilon}{2} \sum_{p:(p,q)\in\Psi_\Omega} \mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})^2_{p,q})] + \text{const.} \\ &= \frac{-1}{2}(\boldsymbol{s}_{\cdot q}^T [\sum_{p:(p,q)\in\Psi_\Omega} \mathbb{E}[\gamma_\epsilon]\mathbb{E}[\boldsymbol{\phi}_{p,q}^T \boldsymbol{\phi}_{p,q}] + \mathbb{E}[\gamma_s]\boldsymbol{I}_K]\boldsymbol{s}_{\cdot q}) \\ &+ \mathbb{E}[\gamma_\epsilon] \sum_{p:(p,q)\in\Psi_\Omega} \mathcal{H}(\boldsymbol{Y}^o - \mathbb{E}[\boldsymbol{E}])_{p,q}\mathbb{E}[\boldsymbol{\phi}_{p,q}]\boldsymbol{s}_{\cdot q} \\ &- \frac{\mathbb{E}[\gamma_\epsilon]}{2}\mathbb{E}[\sum_{p:(p,q)\in\Psi_\Omega} \mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})^2_{p,q}] + \text{const.}. \end{aligned} \quad (70)$$

The above derivation reveals that $q(\boldsymbol{s}_{\cdot q})$ is a Gaussian distribution with mean $\mathbb{E}[\boldsymbol{s}_{\cdot q}]$ and covariance $\Sigma_{\boldsymbol{s}_{\cdot q}}$, i.e.,

$$q(\boldsymbol{s}_{\cdot q}) \sim \mathcal{N}(\mathbb{E}[\boldsymbol{s}_{\cdot q}], \Sigma_{\boldsymbol{s}_{\cdot q}}), \quad (71)$$

where

$$\Sigma_{\boldsymbol{s}_{\cdot q}} = [\mathbb{E}[\gamma_\epsilon] \sum_{p:(p,q)\in\Psi_\Omega} \mathbb{E}[\boldsymbol{\phi}_{p,q}^T \boldsymbol{\phi}_{p,q}] + \mathbb{E}[\gamma_s]\boldsymbol{I}_K]^{-1}, \quad (72)$$

$$\mathbb{E}[\boldsymbol{s}_{\cdot q}] = \mathbb{E}[\gamma_\epsilon] \Sigma_{\boldsymbol{s}_{\cdot q}} \sum_{p:(p,q)\in\Psi_\Omega} \mathbb{E}[\boldsymbol{\phi}_{p,q}]^T \mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})_{p,q}. \quad (73)$$

The required expectation is $\mathbb{E}[\boldsymbol{\phi}_{p,q}] = \mathbb{E}[\boldsymbol{d}_{p\cdot}] \odot \mathbb{E}[\boldsymbol{z}_{\cdot q}]^T$ and

$$\begin{aligned} &\mathbb{E}[\boldsymbol{\phi}_{p,q}^T \boldsymbol{\phi}_{p,q}] = (\mathbb{E}[\boldsymbol{d}_{p\cdot}^T \boldsymbol{d}_{p\cdot}]) \odot (\mathbb{E}[\boldsymbol{z}_{\cdot q}^T \boldsymbol{z}_{\cdot q}]) \\ &= (\mathbb{E}[\boldsymbol{d}_{p\cdot}]^T \mathbb{E}[\boldsymbol{d}_{p\cdot}] + \Sigma_{\boldsymbol{d}_{p\cdot}}) \odot (\mathbb{E}[\boldsymbol{z}_{\cdot q}]\mathbb{E}[\boldsymbol{z}_{\cdot q}]^T + \Sigma_{\boldsymbol{z}_{\cdot q}})). \end{aligned} \quad (74)$$

(III) The approximate posterior distribution of z_{kq} (for all $q = 1, ..., n_1$, and $k = 1, ..., K$) is a Bernoulli distribution.
Note that

$$
\begin{aligned}
&\ln(q(z_{kq}))\\
&= \mathbb{E}_{\Theta\setminus z_{kq}}[\ln p(\Theta, Y, Y^o_\Omega)] + \text{const.}\\
&= \mathbb{E}_{\Theta\setminus z_{kq}}[\ln p(Y^o_\Omega|\boldsymbol{D}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{E}, \gamma_\epsilon)p(z_{kq}|\pi_k)] + \text{const.}\\
&= \mathbb{E}[\ln \prod_{(i,j)\in\Omega} \mathcal{N}(Y^o_{i,j}|(\mathcal{H}^\dagger \boldsymbol{X})_{i,j} + E_{i,j}, \frac{1}{\gamma_\epsilon})\\
&\quad \text{Bernoulli}(z_{kq}|\pi_k)] + \text{const.}.
\end{aligned} \tag{75}
$$

Because z_{kq} is binary, $\ln(q(z_{kq}=1))$ can be written as

$$
\begin{aligned}
&\ln(q(z_{kq}=1))\\
&= \mathbb{E}_{\Theta\setminus z_{kq}}[\ln \prod_{(i,j)\in\Omega} \mathcal{N}(Y^o_{i,j}|(\mathcal{H}^\dagger \boldsymbol{X})_{i,j} + E_{i,j}, \frac{1}{\gamma_\epsilon})\pi_k] + \text{const.}\\
&= \mathbb{E}[\ln \prod_{(i,j)\in\Omega} \mathcal{N}(Y^o_{i,j}|(\mathcal{H}^\dagger \boldsymbol{X})_{i,j} + E_{i,j}, \frac{1}{\gamma_\epsilon})] + \mathbb{E}[\ln(\pi_k)] + \text{const.}\\
&= \mathbb{E}[\frac{-\gamma_\epsilon}{2} \sum_{p:(p,q)\in\Psi_\Omega} [\boldsymbol{d}_{p\cdot}(\boldsymbol{s}_{\cdot q} \odot \hat{\boldsymbol{z}}_{\cdot q})(\boldsymbol{s}_{\cdot q} \odot \hat{\boldsymbol{z}}_{\cdot q})^T \boldsymbol{d}^T_{p\cdot}]\\
&+ \gamma_\epsilon \sum_{p:(p,q)\in\Psi_\Omega} \mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})_{p,q}[(\boldsymbol{s}_{\cdot q} \odot \hat{\boldsymbol{z}}_{\cdot q})^T \boldsymbol{d}^T_{p\cdot}] + \mathbb{E}[\ln(\pi_k)] + \text{const.}\\
&= \mathbb{E}[\frac{-\gamma_\epsilon}{2} \sum_{p:(p,q)\in\Psi_\Omega} [\text{trace}(\boldsymbol{d}_{p\cdot}(\boldsymbol{s}_{\cdot q} \odot \hat{\boldsymbol{z}}_{\cdot q})(\boldsymbol{s}_{\cdot q} \odot \hat{\boldsymbol{z}}_{\cdot q})^T \boldsymbol{d}^T_{p\cdot})]\\
&+ \gamma_\epsilon \sum_{p:(p,q)\in\Psi_\Omega} \mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})_{p,q}[(\boldsymbol{s}_{\cdot q} \odot \hat{\boldsymbol{z}}_{\cdot q})^T \boldsymbol{d}^T_{p\cdot}] + \mathbb{E}[\ln(\pi_k)] + \text{const.}\\
&= \mathbb{E}[\frac{-\gamma_\epsilon}{2} \sum_{p:(p,q)\in\Psi_\Omega} [\text{trace}(\boldsymbol{d}^T_{p\cdot}\boldsymbol{d}_{p\cdot}(\boldsymbol{s}_{\cdot q} \odot \hat{\boldsymbol{z}}_{\cdot q})(\boldsymbol{s}_{\cdot q} \odot \hat{\boldsymbol{z}}_{\cdot q})^T)]\\
&+ \gamma_\epsilon \sum_{p:(p,q)\in\Psi_\Omega} \mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})_{p,q}[(\boldsymbol{s}_{\cdot q} \odot \hat{\boldsymbol{z}}_{\cdot q})^T \boldsymbol{d}^T_{p\cdot}] + \mathbb{E}[\ln(\pi_k)] + \text{const.}\\
&= \frac{-\mathbb{E}[\gamma_\epsilon]}{2} \sum_{p:(p,q)\in\Psi_\Omega} [\text{trace}(\mathbb{E}[\boldsymbol{d}^T_{p\cdot}\boldsymbol{d}_{p\cdot}](\mathbb{E}[\boldsymbol{s}_{\cdot q}\boldsymbol{s}^T_{\cdot q}] \odot \mathbb{E}[\hat{\boldsymbol{z}}_{\cdot q}\hat{\boldsymbol{z}}^T_{\cdot q}]))]\\
&+ \mathbb{E}[\gamma_\epsilon] \sum_{p:(p,q)\in\Psi_\Omega} \mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})_{p,q}[(\mathbb{E}[\boldsymbol{s}_{\cdot q}] \odot \mathbb{E}[\hat{\boldsymbol{z}}_{\cdot q}])^T \mathbb{E}[\boldsymbol{d}_{p\cdot}]^T]\\
&+ \mathbb{E}[\ln(\pi_k)] + \text{const.},
\end{aligned} \tag{76}
$$

where $\hat{z}_{kq} = 1$, other entries in $\hat{z}_{\cdot q}$ equal to the corresponding entries in $z_{\cdot q}$. The required expectations are

$$
\mathbb{E}[\boldsymbol{d}^T_{p\cdot}\boldsymbol{d}_{p\cdot}] = \mathbb{E}[\boldsymbol{d}_{p\cdot}]^T \mathbb{E}[\boldsymbol{d}_{p\cdot}] + \Sigma_{\boldsymbol{d}_{p\cdot}}, \tag{77}
$$

$$
\mathbb{E}[\boldsymbol{s}_{\cdot q}\boldsymbol{s}^T_{\cdot q}] = \mathbb{E}[\boldsymbol{s}_{\cdot q}]\mathbb{E}[\boldsymbol{s}_{\cdot q}]^T + \Sigma_{\boldsymbol{s}_{\cdot q}}, \tag{78}
$$

$$
\mathbb{E}[\hat{\boldsymbol{z}}_{\cdot q}\hat{\boldsymbol{z}}^T_{\cdot q}] = \mathbb{E}[\hat{\boldsymbol{z}}_{\cdot q}]\mathbb{E}[\hat{\boldsymbol{z}}_{\cdot q}]^T + \hat{\boldsymbol{B}}_q, \tag{79}
$$

where $\hat{\boldsymbol{B}}_q = \text{diag}[\mathbb{E}[z_{1q}](1-\mathbb{E}[z_{1q}]), ..., \mathbb{E}[z_{Kq}](1-\mathbb{E}[z_{Kq}])]$, $\hat{B}_{kq} = \mathbb{E}[z_{kq}](1-\mathbb{E}[z_{kq}]) = 0$.
$\ln(q(z_{kq}=0))$ can be expressed as

$$
\begin{aligned}
&\ln(q(z_{kq}=0))\\
&= \mathbb{E}_{\Theta\setminus z_{kq}}[\ln \mathcal{N}(Y^o_{i,j}|(\mathcal{H}^\dagger \boldsymbol{X})_{i,j} + E_{i,j}, \frac{1}{\gamma_\epsilon})(1-\pi_k)] + \text{const.}\\
&= \frac{-\mathbb{E}[\gamma_\epsilon]}{2} \sum_{p:(p,q)\in\Psi_\Omega} [\text{trace}(\mathbb{E}[\boldsymbol{d}^T_{p\cdot}\boldsymbol{d}_{p\cdot}](\mathbb{E}[\boldsymbol{s}_{\cdot q}\boldsymbol{s}^T_{\cdot q}] \odot \mathbb{E}[\hat{\boldsymbol{z}}_{\cdot q}\hat{\boldsymbol{z}}^T_{\cdot q}]))]\\
&+ \mathbb{E}[\gamma_\epsilon] \sum_{p:(p,q)\in\Psi_\Omega} \mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})_{p,q}[(\mathbb{E}[\boldsymbol{s}_{\cdot q}] \odot \mathbb{E}[\hat{\boldsymbol{z}}_{\cdot q}])^T \mathbb{E}[\boldsymbol{d}_{p\cdot}]^T]\\
&+ \mathbb{E}[\ln(1-\pi_k)] + \text{const.},
\end{aligned} \tag{80}
$$

where $\hat{z}_{kq} = 0$, other entries in $\hat{z}_{\cdot q}$ equal to the corresponding entries in $z_{\cdot q}$.
Thus, z_{kq} follows a Bernoulli distribution

$$
q(z_{kq}) \sim \text{Bernoulli}(\frac{q(z_{kq}=1)}{q(z_{kq}=1)+q(z_{kq}=0)}), \tag{81}
$$

with mean and variance

$$
\mathbb{E}[z_{kq}] = \frac{q(z_{kq}=1)}{q(z_{kq}=1)+q(z_{kq}=0)}, \tag{82}
$$

$$
\Sigma_{z_{kq}} = \mathbb{E}[z_{kq}](1-\mathbb{E}[z_{kq}]), \tag{83}
$$

where

$$
\begin{aligned}
&\ln(q(z_{kq}=1)) \propto\\
&\frac{-\mathbb{E}[\gamma_\epsilon]}{2} \sum_{p:(p,q)\in\Psi_\Omega} [\text{trace}(\mathbb{E}[\boldsymbol{d}^T_{p\cdot}\boldsymbol{d}_{p\cdot}](\mathbb{E}[\boldsymbol{s}_{\cdot q}\boldsymbol{s}^T_{\cdot q}] \odot \mathbb{E}[\hat{\boldsymbol{z}}_{\cdot q}\hat{\boldsymbol{z}}^T_{\cdot q}]))]\\
&+ \mathbb{E}[\gamma_\epsilon] \sum_{p:(p,q)\in\Psi_\Omega} \mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})_{p,q}[(\mathbb{E}[\boldsymbol{s}_{\cdot q}] \odot \mathbb{E}[\hat{\boldsymbol{z}}_{\cdot q}])^T \mathbb{E}[\boldsymbol{d}_{p\cdot}]^T]\\
&+ \mathbb{E}[\ln(\pi_k)],
\end{aligned} \tag{84}
$$

where $\hat{z}_{kq} = 1$, other entries in $\hat{z}_{\cdot q}$ equal to the corresponding entries in $z_{\cdot q}$.

$$
\begin{aligned}
&\ln(q(z_{kq}=0)) \propto\\
&\frac{-\mathbb{E}[\gamma_\epsilon]}{2} \sum_{p:(p,q)\in\Psi_\Omega} [\text{trace}(\mathbb{E}[\boldsymbol{d}^T_{p\cdot}\boldsymbol{d}_{p\cdot}](\mathbb{E}[\boldsymbol{s}_{\cdot q}\boldsymbol{s}^T_{\cdot q}] \odot \mathbb{E}[\hat{\boldsymbol{z}}_{\cdot q}\hat{\boldsymbol{z}}^T_{\cdot q}]))]\\
&+ \mathbb{E}[\gamma_\epsilon] \sum_{p:(p,q)\in\Psi_\Omega} \mathcal{H}(\boldsymbol{Y}^o - \boldsymbol{E})_{p,q}[(\mathbb{E}[\boldsymbol{s}_{\cdot q}] \odot \mathbb{E}[\hat{\boldsymbol{z}}_{\cdot q}])^T \mathbb{E}[\boldsymbol{d}_{p\cdot}]^T]\\
&+ \mathbb{E}[\ln(1-\pi_k)],
\end{aligned} \tag{85}
$$

where $\hat{z}_{kq} = 0$, other entries in $\hat{z}_{\cdot q}$ equal to the corresponding entries in $z_{\cdot q}$.
(IV) The approximate posterior distribution of π_k ($k = 1, ..., K$) is from a Beta distribution.
Because the prior distribution of π_k is a Beta distribution

$$
\text{Beta}(\pi_k|\frac{a_0}{K}, \frac{b_0(K-1)}{K}) \propto (\pi_k)^{\frac{a_0}{K}-1}(1-\pi_k)^{\frac{b_0(K-1)}{K}-1}. \tag{86}
$$

Given π_k, the likelihood of z_{kq} is a Bernoulli distribution

$$\text{Bernoulli}(z_{kq}|\pi_k) = (\pi_k)^{z_{kq}}(1-\pi_k)^{1-z_{kq}}. \tag{87}$$

Combine (86) and (87) together, we can get

$$\begin{aligned}&\prod_{q=1}^{n_1}\text{Bernoulli}(z_{kq}|\pi_k)\text{Beta}(\pi_k|\frac{a_0}{K},\frac{b_0(K-1)}{K})\\&\propto(\pi_k)^{\frac{a_0}{K}+\sum_{q=1}^{n_1}z_{kq}-1}(1-\pi_k)^{\frac{b_0(K-1)}{K}+n_1-\sum_{q=1}^{n_1}z_{kq}-1}.\end{aligned} \tag{88}$$

Therefore,

$$\begin{aligned}&\ln(q(\pi_k))\\&=\mathbb{E}_{\Theta\setminus\pi_k}[\ln p(\Theta,Y,Y^o_\Omega)]+\text{const.}\\&=\mathbb{E}_{\Theta\setminus\pi_k}[\ln p(Z|\pi_k)p(\pi_k)]+\text{const.}\\&=\mathbb{E}\,[\ln\ \prod_{q=1}^{n_1}\text{Bernoulli}(z_{kq}|\pi_k)\text{Beta}(\pi_k|\tfrac{a_0}{K},\tfrac{b_0(K-1)}{K})]+\text{const.}\\&=\mathbb{E}\,[\ln(\pi_k)^{\frac{a_0}{K}+\sum_{q=1}^{n_1}z_{kq}-1}(1-\pi_k)^{\frac{b_0(K-1)}{K}+n_1-\sum_{q=1}^{n_1}z_{kq}-1}]+\text{const.}\\&=\mathbb{E}[(\frac{a_0}{K}+\sum_{q=1}^{n_1}z_{kq}-1)\ln(\pi_k)+(\frac{b_0(K-1)}{K}+n_1-\sum_{q=1}^{n_1}z_{kq}-1)\\&\ln(1-\pi_k)]+\text{const.}\\&=(\frac{a_0}{K}+\sum_{q=1}^{n_1}\mathbb{E}[z_{kq}]-1)\ln(\pi_k)+(\frac{b_0(K-1)}{K}+n_1\\&-\sum_{q=1}^{n_1}\mathbb{E}[z_{kq}]-1)\ln(1-\pi_k)+\text{const.}.\end{aligned} \tag{89}$$

So $q(\pi_k)$ satisfies a Beta distribution

$$\begin{aligned}&q(\pi_k)\sim\\&\text{Beta}(\frac{a_0}{K}+\sum_{q=1}^{n_1}\mathbb{E}[z_{kq}],\frac{b_0(K-1)}{K}+n_1-\sum_{q=1}^{n_1}\mathbb{E}[z_{kq}]).\end{aligned} \tag{90}$$

The expectation of $\ln(\pi_k)$ is

$$\mathbb{E}[\ln(\pi_k)]=\psi(\frac{a_0}{K}+\sum_{q=1}^{n_1}\mathbb{E}[z_{kq}])-\psi(\frac{a_0+b_0(K-1)}{K}+n_1). \tag{91}$$

The expectation of $\ln(1-\pi_k)$ is

$$\begin{aligned}&\mathbb{E}[\ln(1-\pi_k)]=\\&\psi(\frac{b_0(K-1)}{K}+n_1-\sum_{q=1}^{n_1}\mathbb{E}[z_{kq}])-\psi(\frac{a_0+b_0(K-1)}{K}+n_1).\end{aligned} \tag{92}$$

Note that the equations (91) and (92) are derived based on one property of logarithm Beta function, i.e., if π_k satisfies a Beta distribution $\text{Beta}(\alpha_1,\beta_1)$ with parameters (α_1,β_1), then the expectations of $\ln(\pi_k)$ and $\ln(1-\pi_k)$ are
$\mathbb{E}[\ln(\pi_k)]=\psi(\alpha_1)-\psi(\alpha_1+\beta_1)$
and
$\mathbb{E}[\ln(1-\pi_k)]=\psi(\beta_1)-\psi(\alpha_1+\beta_1)$,
respectively. $\psi(.)$ is the diagamma function and $\psi(\alpha_1)=\frac{\Gamma'(\alpha_1)}{\Gamma(\alpha_1)}$.

(V) The approximate posterior distribution of γ_s is a Gamma distribution.
Given

$$\Gamma(\gamma_s|c_0,d_0)\propto(\gamma_s)^{c_0-1}e^{-d_0\gamma_s}, \tag{93}$$

and

$$\prod_{q=1}^{n_1}\mathcal{N}(\boldsymbol{s}_{\cdot q}|0,\frac{1}{\gamma_s}\boldsymbol{I}_K)\propto(\gamma_s)^{\frac{n_1K}{2}}\exp(-\frac{\sum_{q=1}^{n_1}||\boldsymbol{s}_{\cdot q}||_2^2}{2}\gamma_s). \tag{94}$$

Therefore,

$$\begin{aligned}&\ln(q(\gamma_s))\\&=\mathbb{E}_{\Theta\setminus\gamma_s}[\ln p(\Theta,Y,Y^o_\Omega)]+\text{const.}\\&=\mathbb{E}_{\Theta\setminus\gamma_s}[\ln p(S|\gamma_s)p(\gamma_s)]+\text{const.}\\&=\mathbb{E}[\ln\prod_{q=1}^{n_1}\mathcal{N}(\boldsymbol{s}_{\cdot q}|0,\frac{1}{\gamma_s}I_K)\Gamma(\gamma_s|c_0,d_0)]+\text{const.}\\&=\mathbb{E}[\ln(\gamma_s)^{\frac{n_1K}{2}+c_0-1}\exp[-\gamma_s(\frac{1}{2}\sum_{q=1}^{n_1}||\boldsymbol{s}_{\cdot q}||_2^2+d_0)]]\\&+\text{const.}\\&=\mathbb{E}[(\frac{n_1K}{2}+c_0-1)\ln(\gamma_s)-\gamma_s(\frac{1}{2}\sum_{q=1}^{n_1}||\boldsymbol{s}_{\cdot q}||_2^2+d_0)]\\&+\text{const.}\\&=(\frac{n_1K}{2}+c_0-1)\ln(\gamma_s)-\gamma_s(\frac{1}{2}\sum_{q=1}^{n_1}\mathbb{E}[\boldsymbol{s}_{\cdot q}^T\boldsymbol{s}_{\cdot q}]+d_0)\\&+\text{const.}.\end{aligned} \tag{95}$$

The $q(\gamma_s)$ satisfies a Gamma distribution

$$q(\gamma_s)\sim\Gamma(\frac{n_1K}{2}+c_0,\frac{1}{2}\sum_{q=1}^{n_1}\mathbb{E}[\boldsymbol{s}_{\cdot q}^T\boldsymbol{s}_{\cdot q}]+d_0), \tag{96}$$

with mean

$$\mathbb{E}[\gamma_s]=\frac{\frac{n_1K}{2}+c_0}{\frac{1}{2}\sum_{q=1}^{n_1}\mathbb{E}[\boldsymbol{s}_{\cdot q}^T\boldsymbol{s}_{\cdot q}]+d_0)}, \tag{97}$$

where $\mathbb{E}[\boldsymbol{s}_{\cdot q}^T\boldsymbol{s}_{\cdot q}]=\mathbb{E}[\boldsymbol{s}_{\cdot q}^T]\mathbb{E}[\boldsymbol{s}_{\cdot q}]+\text{trace}(\Sigma_{\boldsymbol{s}_{\cdot q}})$.
(VI) The approximate posterior distribution of $E_{i,j}$ (for $(i,j)\in\Omega$) is a Gaussian distribution.
Because

$$\begin{aligned}&\mathcal{N}(Y^o_{i,j}|(\mathcal{H}^\dagger\boldsymbol{X})_{i,j}+E_{i,j},\frac{1}{\gamma_\epsilon})\mathcal{N}(E_{i,j}|0,\frac{1}{\beta_{i,j}})\\&\propto\exp(\frac{-\gamma_\epsilon}{2}(E_{i,j}^2-2E_{i,j}(Y^o_{i,j}-(\mathcal{H}^\dagger\boldsymbol{X})_{i,j})\\&\quad+(Y^o_{i,j}-(\mathcal{H}^\dagger\boldsymbol{X})_{i,j})^2))\exp(\frac{-\beta_{i,j}}{2}E_{i,j}^2)\\&\propto\exp(\frac{-(\gamma_\epsilon+\beta_{i,j})}{2}E_{i,j}^2+\gamma_\epsilon E_{i,j}(Y^o_{i,j}-(\mathcal{H}^\dagger\boldsymbol{X})_{i,j})-\\&\quad\frac{\gamma_\epsilon}{2}(Y^o_{i,j}-(\mathcal{H}^\dagger\boldsymbol{X})_{i,j})^2),\end{aligned} \tag{98}$$

then we can derive

$$\begin{aligned}
&\ln(\ q(E_{i,j}))\\
&= \mathbb{E}_{\boldsymbol{\Theta}\setminus E_{i,j}}[\ln\, p(\boldsymbol{\Theta}, \boldsymbol{Y}, \boldsymbol{Y}^o_\Omega)] + \text{const.}\\
&= \mathbb{E}_{\boldsymbol{\Theta}\setminus E_{i,j}}[\ln\, p(\boldsymbol{Y}^o_\Omega|\boldsymbol{D}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{E}, \gamma_\epsilon) p(E_{i,j})] + \text{const.}\\
&= \mathbb{E}[\ln\, \mathcal{N}(Y^o_{i,j}|(\mathcal{H}^\dagger \boldsymbol{X})_{i,j} + E_{i,j}, \frac{1}{\gamma_\epsilon}) \mathcal{N}(E_{i,j}|0, \frac{1}{\beta_{i,j}})] + \text{const.}\\
&= \mathbb{E}[\frac{-(\gamma_\epsilon + \beta_{i,j})}{2} E^2_{i,j} + \gamma_\epsilon E_{i,j}(Y^o_{i,j} - (\mathcal{H}^\dagger X)_{i,j})\\
&- \frac{\gamma_\epsilon}{2}(Y^o_{i,j} - (\mathcal{H}^\dagger \boldsymbol{X})_{i,j})^2] + \text{const.}\\
&= \frac{-(\mathbb{E}[\gamma_\epsilon] + \mathbb{E}[\beta_{i,j}])}{2} E^2_{i,j} + \mathbb{E}[\gamma_\epsilon] E_{i,j}(Y^o_{i,j} - \mathbb{E}[(\mathcal{H}^\dagger \boldsymbol{X})_{i,j}])\\
&- \mathbb{E}[\frac{\gamma_\epsilon}{2}(Y^o_{i,j} - (\mathcal{H}^\dagger \boldsymbol{X})_{i,j})^2] + \text{const..}
\end{aligned} \tag{99}$$

The above derivation reveals that $q(E_{i,j})$ is a Gaussian distribution with mean $\mathbb{E}[E_{i,j}]$ and covariance $\Sigma_{E_{i,j}}$, i.e.,

$$q(E_{i,j}) \sim \mathcal{N}(\mathbb{E}[E_{i,j}], \Sigma_{E_{i,j}}), \tag{100}$$

where

$$\mathbb{E}[E_{i,j}] = \mathbb{E}[\gamma_\epsilon]\Sigma_{E_{i,j}}(Y^o_{i,j} - \mathbb{E}[(\mathcal{H}^\dagger \boldsymbol{X})_{i,j}]), \tag{101}$$

$$\Sigma_{E_{i,j}} = \frac{1}{\mathbb{E}[\gamma_\epsilon] + \mathbb{E}[\beta_{i,j}]}. \tag{102}$$

The required expectation is

$$\mathbb{E}[(\mathcal{H}^\dagger \boldsymbol{X})_{i,j}] = \frac{1}{\kappa_j} \sum_{(u,v)\in\Psi_{i,j}} [\mathbb{E}[\boldsymbol{d}_{u.}](\mathbb{E}[\boldsymbol{s}_{.v}] \odot \mathbb{E}[\boldsymbol{z}_{.v}])]. \tag{103}$$

(VII) The approximate posterior distribution of $\beta_{i,j}$ (for $(i,j) \in \Omega$) is a Gamma distribution.
Because

$$\Gamma(\beta_{i,j}|g_0, h_0) \propto (\beta_{i,j})^{g_0-1} e^{-h_0\beta_{i,j}}, \tag{104}$$

and

$$\mathcal{N}(E_{i,j}|0, \frac{1}{\beta_{i,j}}) \propto (\beta_{i,j})^{\frac{1}{2}} \exp(\frac{-\beta_{i,j}}{2} E^2_{i,j}). \tag{105}$$

Combine (104) and (105) together,

$$\begin{aligned}
&\mathcal{N}(E_{i,j}|0, \frac{1}{\beta_{i,j}})\Gamma(\beta_{i,j}|g_0, h_0)\\
&\propto (\beta_{i,j})^{\frac{1}{2}+g_0-1} \exp(-\beta_{i,j}(\frac{1}{2}E^2_{i,j} + h_0)).
\end{aligned} \tag{106}$$

Therefore,

$$\begin{aligned}
&\ln(q(\beta_{i,j}))\\
&= \mathbb{E}_{\boldsymbol{\Theta}\setminus\beta_{i,j}}[\ln\, p(\boldsymbol{\Theta}, \boldsymbol{Y}, \boldsymbol{Y}^o_\Omega)] + \text{const.}\\
&= \mathbb{E}_{\boldsymbol{\Theta}\setminus\beta_{i,j}}[\ln\, p(E_{i,j}|\beta_{i,j}) p(\beta_{i,j})] + \text{const.}\\
&= \mathbb{E}[\ln\, \mathcal{N}(E_{i,j}|0, \frac{1}{\beta_{i,j}})\Gamma(\beta_{i,j}|g_0, h_0)] + \text{const.}\\
&= (\frac{1}{2} + g_0 - 1)\ln(\beta_{i,j}) - \beta_{i,j}(\frac{1}{2}\mathbb{E}[E^2_{i,j}] + h_0) + \text{const.},
\end{aligned} \tag{107}$$

where $\mathbb{E}[E^2_{i,j}] = \mathbb{E}[E_{i,j}]^2 + \Sigma_{E_{i,j}}$. The equation (107) indicates that $\beta_{i,j}$ follows a Gamma distribution

$$q(\beta_{i,j}) \sim \Gamma(\frac{1}{2} + g_0, \frac{1}{2}\mathbb{E}[E^2_{i,j}] + h_0), \tag{108}$$

with mean

$$\mathbb{E}[\beta_{i,j}] = \frac{\frac{1}{2} + g_0}{\frac{1}{2}\mathbb{E}[E^2_{i,j}] + h_0}, \tag{109}$$

where $\mathbb{E}[E^2_{i,j}] = \mathbb{E}[E_{i,j}]^2 + \Sigma_{E_{i,j}}$.
(VI) The approximate posterior distribution of γ_ϵ is a Gamma distribution.
Note that

$$\Gamma(\gamma_\epsilon|e_0, f_0) \propto (\gamma_\epsilon)^{e_0-1} e^{-f_0\gamma_\epsilon}, \tag{110}$$

and

$$\begin{aligned}
&\prod_{(i,j)\in\Omega} \mathcal{N}(Y^o_{i,j}|(\mathcal{H}^\dagger \boldsymbol{X})_{i,j} + E_{i,j}, \frac{1}{\gamma_\epsilon})\\
&\propto (\gamma_\epsilon)^{\frac{|\Omega|}{2}} \exp(\frac{-\gamma_\epsilon}{2}||\boldsymbol{Y}^o - P_\Omega(\mathcal{H}^\dagger \boldsymbol{X} + \boldsymbol{E})||^2_F),
\end{aligned} \tag{111}$$

where $|\Omega|$ is the cadinality of Ω.
Therefore,

$$\begin{aligned}
&\ln(q(\gamma_\epsilon))\\
&= \mathbb{E}_{\boldsymbol{\Theta}\setminus\gamma_\epsilon}[\ln\, p(\boldsymbol{\Theta}, \boldsymbol{Y}, \boldsymbol{Y}^o_\Omega)] + \text{const.}\\
&= \mathbb{E}_{\boldsymbol{\Theta}\setminus\gamma_\epsilon}[\ln\, p(\boldsymbol{Y}^o_\Omega|\boldsymbol{D}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{E}, \gamma_\epsilon) p(\gamma_\epsilon)] + \text{const.}\\
&= \mathbb{E}[\ln \prod_{(i,j)\in\Omega} \mathcal{N}(Y^o_{i,j}|(\mathcal{H}^\dagger \boldsymbol{X})_{i,j} + E_{i,j}, \frac{1}{\gamma_\epsilon})\Gamma(\gamma_\epsilon|e_0, f_0)] + \text{const.}\\
&= (\frac{|\Omega|}{2} + e_0 - 1)\ln(\gamma_\epsilon) + \frac{-\gamma_\epsilon}{2}\mathbb{E}[||\boldsymbol{Y}^o - P_\Omega(\mathcal{H}^\dagger \boldsymbol{X} + \boldsymbol{E})||^2_F] - f_0\gamma_\epsilon + \text{const.}\\
&= (\frac{|\Omega|}{2} + e_0 - 1)\ln(\gamma_\epsilon) + \frac{-\gamma_\epsilon}{2} \sum_{(i,j)\in\Omega} \mathbb{E}[(Y^o_{i,j} - E_{i,j}\\
&- \frac{1}{\kappa_j} \sum_{(u,v)\in\Psi_{i,j}} [\boldsymbol{d}_{u.}(\boldsymbol{s}_{.v} \odot \boldsymbol{z}_{.v})])^2] - f_0\gamma_\epsilon + \text{const.},
\end{aligned} \tag{112}$$

where

$$\begin{aligned}
&\mathbb{E}[(Y^o_{i,j} - E_{i,j} - \frac{1}{\kappa_j} \sum_{(u,v)\in\Psi_{i,j}} [\boldsymbol{d}_{u.}(\boldsymbol{s}_{.v} \odot \boldsymbol{z}_{.v})])^2]\\
&= {Y^o_{i,j}}^2 - 2Y^o_{i,j}\mathbb{E}[E_{i,j}] + \mathbb{E}[E^2_{i,j}] - 2(Y^o_{i,j} - \mathbb{E}[E_{i,j}])\frac{1}{\kappa_j} \sum_{(u,v)\in\Psi_{i,j}}\\
&[\mathbb{E}[\boldsymbol{d}_{u.}](\mathbb{E}[\boldsymbol{s}_{.v}] \odot [\mathbb{E}[\boldsymbol{z}_{.v}])] + \frac{1}{\kappa_j^2}\mathbb{E}[(\sum_{(u,v)\in\Psi_{i,j}} \boldsymbol{d}_{u.}(\boldsymbol{s}_{.v} \odot \boldsymbol{z}_{.v}))^2]
\end{aligned} \tag{113}$$

$$
\begin{aligned}
&= \frac{1}{\kappa_j^2} \sum_{(p,q)\in\Psi_{i,j}} \text{trace}(\mathbb{E}[\boldsymbol{d}_{p.}]^T \mathbb{E}[\boldsymbol{d}_{p.}](\mathbb{E}[\boldsymbol{s}_{.q}]\mathbb{E}[\boldsymbol{s}_{.q}]^T \odot \Sigma_{\boldsymbol{z}_{.q}})) \\
&+ \frac{1}{\kappa_j^2} \sum_{(p,q)\in\Psi_{i,j}} \text{trace}(\mathbb{E}[\boldsymbol{d}_{p.}]^T \mathbb{E}[\boldsymbol{d}_{p.}](\mathbb{E}[\boldsymbol{z}_{.q}]\mathbb{E}[\boldsymbol{z}_{.q}]^T \odot \Sigma_{\boldsymbol{s}_{.q}})) \\
&+ \frac{1}{\kappa_j^2} \sum_{(p,q)\in\Psi_{i,j}} \text{trace}(\mathbb{E}[\boldsymbol{d}_{p.}]^T \mathbb{E}[\boldsymbol{d}_{p.}](\Sigma_{\boldsymbol{z}_{.q}} \odot \Sigma_{\boldsymbol{s}_{.q}})) \\
&+ \frac{1}{\kappa_j^2} \sum_{(p,q)\in\Psi_{i,j}} \text{trace}(\Sigma_{\boldsymbol{d}_{p.}}(\mathbb{E}[\boldsymbol{s}_{.q}]\mathbb{E}[\boldsymbol{s}_{.q}]^T \odot \Sigma_{\boldsymbol{z}_{.q}})) \\
&+ \frac{1}{\kappa_j^2} \sum_{(p,q)\in\Psi_{i,j}} \text{trace}(\Sigma_{\boldsymbol{d}_{p.}}(\mathbb{E}[\boldsymbol{z}_{.q}]\mathbb{E}[\boldsymbol{z}_{.q}]^T \odot \Sigma_{\boldsymbol{s}_{.q}})) \\
&+ \frac{1}{\kappa_j^2} \sum_{(p,q)\in\Psi_{i,j}} \text{trace}(\Sigma_{\boldsymbol{d}_{p.}}(\Sigma_{\boldsymbol{z}_{.q}} \odot \Sigma_{\boldsymbol{s}_{.q}})) + \Sigma_{E_{i,j}} \\
&+ \frac{1}{\kappa_j^2} \sum_{(p,q)\in\Psi_{i,j}} \text{trace}(\Sigma_{\boldsymbol{d}_{p.}}(\mathbb{E}[\boldsymbol{s}_{.q}]\mathbb{E}[\boldsymbol{s}_{.q}]^T \odot \mathbb{E}[\boldsymbol{z}_{.q}]\mathbb{E}[\boldsymbol{z}_{.q}]^T)) \\
&+ (Y^o_{i,j} - \mathbb{E}[E_{i,j}] - \frac{1}{\kappa_j} \sum_{(u,v)\in\Psi_{i,j}} [\mathbb{E}[\boldsymbol{d}_{u.}](\mathbb{E}[\boldsymbol{s}_{.v}] \odot \mathbb{E}[\boldsymbol{z}_{.v}])])^2.
\end{aligned} \tag{114}
$$

The equation (112) indicates that γ_ϵ follows a Gamma distribution

$$
q(\gamma_\epsilon) \sim \Gamma(\frac{|\Omega|}{2} + e_0, \frac{1}{2}\mathbb{E}[||Y^o - P_\Omega(\mathcal{H}^\dagger \boldsymbol{X} + \boldsymbol{E})||_F^2] + f_0), \tag{115}
$$

with mean

$$
\mathbb{E}[\gamma_\epsilon] = \frac{\frac{|\Omega|}{2} + e_0}{\frac{1}{2}\mathbb{E}[||Y^o - P_\Omega(\mathcal{H}^\dagger \boldsymbol{X} + \boldsymbol{E})||_F^2] + f_0}. \tag{116}
$$

A.4 Computational Complexity

The computational complexities for Hankel operation and inverse Hankel operation are $O(mn_2n_1)$. The computational complexity for updating D is $O(\kappa mn_2n_1K^2 + mn_2K^3)$, and the complexity for updating S is $O(\kappa mn_2n_1K^2 + n_1K^3)$. The computational complexity for updating Z is $O(\kappa mn_2n_1K^4 + Kn_1)$ and the computational complexity for updating γ_ϵ is $O(\kappa mn_2n_1K^3 + mn_2n_1K)$. The computational complexities for π_k, γ_s, E, and β are $O(Kn_1)$, $O(K^2n_1)$, $O(mn_2n_1K)$ and $O(mn)$, respectively. The final complexity is $O(\kappa mn_2n_1K^4 + mn_2K^3 + n_1K^3 + mn_2n_1K)$. The complexity scales at most linearly with respective to the dimension of the Hankel matrix.

A.5 Predictive mean and predictive variance

We can derive the predictive mean as follows:

$$
\begin{aligned}
\mathbb{E}[Y_{i,j}] &= \int p(Y_{i,j}|Y^o_\Omega) Y_{i,j} dY_{i,j} \\
&= \int (\int p(Y_{i,j}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|Y^o_\Omega) d\boldsymbol{\theta}) Y_{i,j} dY_{i,j} \\
&= \int (\int p(Y_{i,j}|\boldsymbol{\theta}) Y_{i,j} dY_{i,j}) p(\boldsymbol{\theta}|Y^o_\Omega) d\boldsymbol{\theta} \\
&= \int \mathbb{E}_{p(Y_{i,j}|\boldsymbol{\theta})}[Y_{i,j}] p(\boldsymbol{\theta}|Y^o_\Omega) d\boldsymbol{\theta} \\
&= \int f^{\boldsymbol{\theta}}(Y_{i,j}) p(\boldsymbol{\theta}|Y^o_\Omega) d\boldsymbol{\theta} \\
&\approx \frac{1}{L} \sum_{l=1}^{l=L} f^{\boldsymbol{\theta}_l}(Y_{i,j}) \quad \boldsymbol{\theta}_l \sim q(\boldsymbol{\theta}|Y^o_\Omega).
\end{aligned} \tag{117}
$$

The predictive mean for $Y_{i,j}$ is derived by taking the expectation over the probability $p(Y_{i,j}|Y^o_\Omega)$. $\boldsymbol{\theta} = \{\boldsymbol{D}, \boldsymbol{Z}, \boldsymbol{S}, \gamma_\epsilon\}$. $\mathbb{E}_{p(Y_{i,j}|\boldsymbol{\theta})}[Y_{i,j}]$ is the expectation of $Y_{i,j}$ over $p(Y_{i,j}|\boldsymbol{\theta})$. The integration in last second step of equation (117) is difficult to obtain, thus $\boldsymbol{\theta}_l$ is sampled from $q(\boldsymbol{\theta}|Y^o_\Omega)$ and Monte Carlo integration is employed to approximately compute it.

To derive the predictive variance, we compute $\mathbb{E}[Y^2_{i,j}]$ as follows:

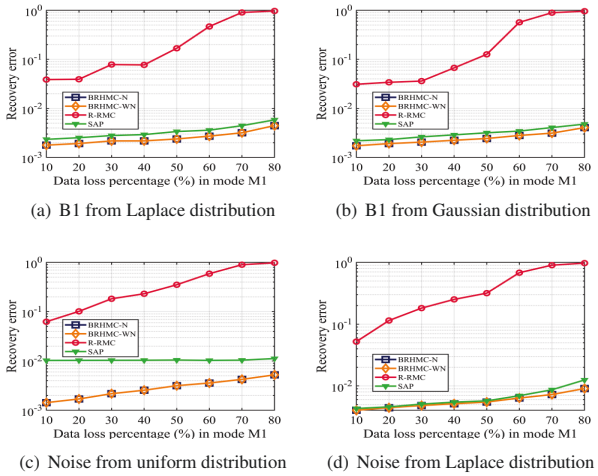
$$
\begin{aligned}
&\mathbb{E}_{p(Y_{i,j}|Y^o_\Omega)}[Y^2_{i,j}] \\
&= \int p(Y_{i,j}|Y^o_\Omega) Y^2_{i,j} dY_{i,j} \\
&= \int (\int p(Y_{i,j}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|Y^o_\Omega) d\boldsymbol{\theta}) Y^2_{i,j} dY_{i,j} \\
&= \int (\int p(Y_{i,j}|\boldsymbol{\theta}) Y^2_{i,j} dY_{i,j}) p(\boldsymbol{\theta}|Y^o_\Omega) d\boldsymbol{\theta} \\
&= \int (\mathbb{E}_{p(Y_{i,j}|\boldsymbol{\theta})}[Y^2_{i,j}]) p(\boldsymbol{\theta}|Y^o_\Omega) d\boldsymbol{\theta} \\
&= \int (\text{Var}_{p(Y_{i,j}|\boldsymbol{\theta})}[Y_{i,j}] + \mathbb{E}^2_{p(Y_{i,j}|\boldsymbol{\theta})}[Y_{i,j}])) p(\boldsymbol{\theta}|Y^o_\Omega) d\boldsymbol{\theta} \\
&= \int (\frac{1}{\gamma_\epsilon} + f^{\boldsymbol{\theta}}(Y_{i,j})^2) p(\boldsymbol{\theta}|Y^o_\Omega) d\boldsymbol{\theta} \\
&\approx \frac{1}{L} \sum_{l=1}^{l=L} \frac{1}{\gamma_\epsilon} + \frac{1}{L} \sum_{l=1}^{l=L} f^{\boldsymbol{\theta}_l}(Y_{i,j})^2 \quad \boldsymbol{\theta}_l \sim q(\boldsymbol{\theta}|Y^o_\Omega).
\end{aligned} \tag{118}
$$

By plugging (117) and (118) into equation (46), the predictive variance can be derived in (46).

A.6 Additional Experiments

A.6.1 The impact of distributions of bad data and noise. In our problem setup, the bad data is generated from uniform distribution and the noise is generated from Gaussian distribution. In this section, we also study the recovery accuracy when the bad data and noise are drawn from different distributions. We consider M1 with 10 % B1 to compare with Fig. 7(d). For the bad data generation, we consider the Laplace distribution with mean 1.5 and standard deviation 0.5. We also consider the Gaussian distribution with mean 1.5 and standard deviation 0.5. For the noise generation, we consider the uniform distribution in the range from 0 to 0.006. We also consider the Laplace distribution with mean 0 and standard deviation 0.08. The recovery

performance is shown in Fig. 11. One can see from Fig. 11 that our proposed method still performs better than the baseline methods. The results are comparable to Fig. 7(d).



(a) B1 from Laplace distribution (b) B1 from Gaussian distribution

(c) Noise from uniform distribution (d) Noise from Laplace distribution

Fig. 11. The recovery results with M1 plus 10% B1 with different bad data or noise distributions. (a)-(b) show the recovery results with bad data generated from different distributions. (c)-(d) show the recovery results with noise generated from different distributions.

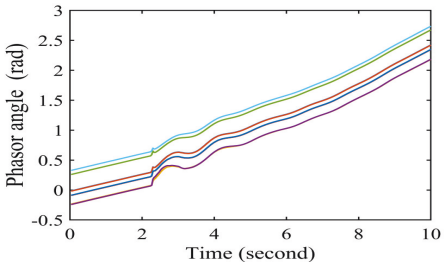


Fig. 12. The measurements of voltage angle [Hao et al. 2018]

A.6.2 Performance on practical PMU phasor angle dataset.

The corresponding PMU angle data of Fig. 1 is shown in Fig. 12. Two extra case studies are considered to verify the effectiveness of our algorithm on the phasor angle dataset. The parameter settings are the same with Case 1 and 2 except that $n_2 = 20$ and $f_0 = 10^{-5}$.

- Case 3: 15% data are removed following Mode M2, and 15% observations contain Mode B2 bad data. Each bad entry is randomly selected from (1,1.5).
- Case 4: 15% data are removed following Mode M3, and 10% observations contain Mode B1 bad data. Each bad entry is randomly selected from (1,1.5).

Our method can also recover the data accurately in both cases for the angle data. The NEE and WNEE for Case 3 are 6.5×10^{-4} and 5.3×10^{-4}, respectively. The NEE and WNEE for Case 4 are 1.3×10^{-3} and 1.1×10^{-3}, respectively. Fig. 13-Fig. 14 show the recovery performance of Case 3 and 4. Similar to the magnitude data, at time 2.3 seconds when the event happens, the uncertainty index increases because the method is less confident about the estimation at that time instant.

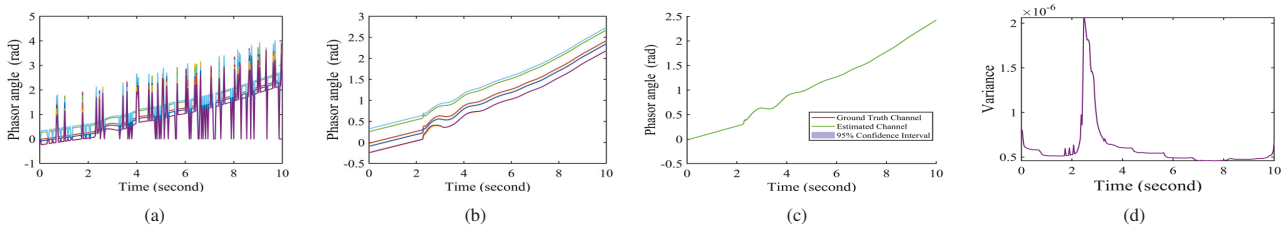


Fig. 13. The recovery performance on 15% M2 missing data and 15% B2 bad data on the angle data. (a) the observed data, (b) the estimated data, (c) the estimated data in one channel with the confidence interval, (d) the corresponding uncertainty index for one channel in (c)

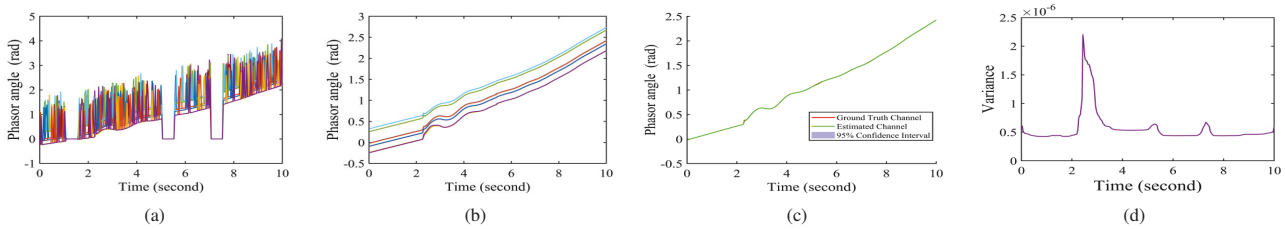


Fig. 14. The recovery performance on 15% M3 missing data and 10% B1 bad data on the angle data. (a) the observed data, (b) the estimated data, (c) the estimated data in one channel with the confidence interval, (d) the corresponding uncertainty index for one channel in (c)