Journal of Machine Learning Research 24 (2023) 1-43 Submitted 6/22; Published 8/23

Multi-source Learning via Completion of Block-wise Overlapping Noisy Matrices

Doudou Zhou DOUDOUZHOU@HSPH.HARVARD.EDU
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Boston, Massachusetts 02115, USA

Tianxi Cai TCAI@HSPH.HARVARD.EDU
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Boston, Massachusetts 02115, USA

Junwei Lu JUNWEILU@HSPH.HARVARD.EDU
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Boston, Massachusetts 02115, USA

Editor: Ali Shojaie

Abstract

Electronic healthcare records (EHR) provide a rich resource for healthcare research. An important problem for the efficient utilization of the EHR data is the representation of the EHR features, which include the unstructured clinical narratives and the structured codified data. Matrix factorization-based embeddings trained using the summary-level co-occurrence statistics of EHR data have provided a promising solution for feature representation while preserving patients' privacy. However, such methods do not work well with multi-source data when these sources have overlapping but non-identical features. To accommodate multi-sources learning, we propose a novel word embedding generative model. To obtain multi-source embeddings, we design an efficient **B**lock-wise **O**verlapping **N**oisy **M**atrix **I**ntegration (BONMI) algorithm to aggregate the multi-source pointwise mutual information matrices optimally with a theoretical guarantee. Our algorithm can also be applied to other multi-source data integration problems with a similar data structure. A by-product of BONMI is the contribution to the field of matrix completion by considering the missing mechanism other than the entry-wise independent missing. We show that the entry-wise missing assumption, despite its prevalence in the works of matrix completion, is not necessary to guarantee recovery. We prove the statistical rate of our estimator, which is comparable to the rate under independent missingness. Simulation studies show that BONMI performs well under a variety of configurations. We further illustrate the utility of BONMI by integrating multi-lingual multi-source medical text and EHR data to perform two tasks: (i) co-training semantic embeddings for medical concepts in both English and Chinese and (ii) the translation between English and Chinese medical concepts. Our method shows an advantage over existing methods.

Keywords: Word embedding, data integration, singular value decomposition, transfer learning.

1. Introduction

1.1 Background

Electronic health records (EHR) have been playing a more and more important role in healthcare research ranging from disease phenotyping (Ahuja et al., 2020; De Freitas et al., 2021; Liu et al., 2021) to precision medicine (Raghu et al., 2017; Parbhoo et al., 2017; Sonabend et al., 2020; Zhou et al., 2022b). Despite the translational potential of EHR data, generating reliable real-world evidence from EHR data has been highly challenging, in part due to the significant heterogeneity across multiple healthcare centers. One approach to generalizability and reproducibility is through consensus learning with multiple EHR (multi-EHR). However, harmonizing data from multi-EHR for consensus learning is a major roadblock due to the lack of interoperability across healthcare systems (Rajkomar et al., 2018). The same clinical concepts can be represented by distinct codes or even different languages such as English and Chinese at different healthcare systems (Hernandez et al., 2009; Abhyankar et al., 2012). As a result, it is common for multiple data sources to have overlapping but non-identical clinical codes.

Although common data model has been increasingly adopted to improve interoperability across healthcare systems, significant discrepancies remain since most healthcare systems have adopted some but not all existing ontologies such as the International Classification of Disease (ICD) codes for diseases, Current Procedural Terminology (CPT) for procedures (Hirsch et al., 2015), and Logical Observation Identifiers Names and Codes (LOINC) for laboratory tests (Weeks and Pardee, 2019). One approach to harmonize data is to map local EHR codes to common ontologies (Kume et al., 2019; Tournavitis et al., 2009; Baloukas et al., 2010). Such an approach, generally requiring some level of manual effort and domain knowledge is thus time and resource-intensive and not scalable (Baorto et al., 1998; Kume et al., 2019). An alternative approach to data harmonization is through representation learning, which has been highly successful in natural language processing (Mikolov et al., 2013a). If a unified set of embeddings can be trained for all EHR codes across multiple systems, these embedding vectors can bridge data from different sources and hence achieve harmonization.

For training unified embeddings for all clinical concepts, the traditional one-hot approach will suffer from the curse of dimension. To train embeddings for a large number of concepts with observed relationship pairs, knowledge graph-based approaches (Wang et al., 2014; Yao et al., 2019) have been shown as highly effective. However, large-scale EHR data are typically represented as sequences of encounters over time. Based on co-occurrence patterns of EHR codes, neural network-based methods such as the skip-gram algorithm (Mikolov et al., 2013a; Pennington et al., 2014; Lin et al., 2019; Boag and Kané, 2017) have been proposed. Training of the skip-gram algorithm can also be equivalently achieved via performing singular value decomposition (SVD) on a pointwise mutual information (PMI) matrix derived from co-occurrence summaries of EHR data (Levy and Goldberg, 2014; Beam et al., 2019; Hong et al., 2021). The SVD-PMI algorithm is particularly appealing due to its scalability and privacy-preserving since it only requires simple summary-level EHR data, which can be shared across multiple systems.

However, these existing methods are ineffective in co-training embeddings for multi-EHR data when the EHR codes from multiple sources overlap but are not identical. Neural

network methods that require patient-level data are not feasible due to data-sharing constraints. The simple pre-training approach effectively assumes that code pairs that do not appear in the same health system have low similarity, which is a poor assumption since two distinct codes can represent the same clinical concepts in two systems due to coding heterogeneity. To better accommodate coding heterogeneity, one may separately train embeddings within each EHR source and then follow up with an alignment step such that codes shared at multiple systems have similar embeddings after alignment (Smith et al., 2017; Kementchedjhieva et al., 2018; Conneau et al., 2018). However, such two-step methods are not efficient both computationally and statistically and are less generalizable to settings with more than two sources.

In this paper, we propose the **B**lock-wise **O**verlapping **N**oisy **M**atrix **I**ntegration (BONMI) algorithm to co-train embeddings for clinical concepts from multiple sources. The BONMI algorithm is built on top of a novel generative model similar to Arora et al. (2016, 2018) but allows EHR codes or concepts to belong to multiple sources. Specifically, let $w_t^{(s,i)}$ denote the code the ith patient from the sth source receives at time t. We assume that the probability of the code $w_t^{(s,i)}$ taking value w is $\mathbb{P}(w_t^{(s,i)} = w \mid \mathbf{c}_t^{(s,i)}) \propto \exp(\mathbf{x}_w^\top \mathbf{c}_t^{(s,i)})$, where $\mathbf{x}_w$ is the underlying embedding representation of the specific code w assumed to be the same for all sources w occurs, and $\mathbf{c}_t^{(s,i)}$ is some latent vector with a random walk on the sphere associated with each patient in each source. Our model allows the multiple sources to have overlapping but not identical corpora.

Under the proposed generative model, the SVD-PMI estimator consistently estimates the underlying embeddings when there is a single data source (Arora et al., 2018; Lu et al., 2023). With multiple sources, the PMI of code pairs that do not co-occur in the same system is not observable. Under the proposed generative model, the underlying embeddings for each code do not depend on the source and can be recovered based on data from the source that contains the code. However, it is no longer feasible to estimate the embeddings through simple decomposition of PMI matrices due to the missingness of co-pairs that do not co-occur. We propose an efficient estimation procedure based on the low-rank nature of the PMI matrices and the low-dimensional embedding vectors. Our idea connects to the orthogonal Procrustes problem (Gower and Dijksterhuis, 2004; Schönemann, 1966; Gower, 1975), which has been widely used to align embeddings across languages in the machine translation (Kementchedjhieva et al., 2018; Smith et al., 2017; Conneau et al., 2018; Søgaard et al., 2018; Xing et al., 2015). We use an orthogonal transformation to align the eigenspace of the two sub-matrices through their overlap, then complete the missing entries by the inner products of the two low-rank components. Moreover, we generalize our method to the multiple sources scenario by applying the method to each pair of the sub-matrices. Since BONMI operates on matrices from any two sources, it is suitable for parallel computing.

1.2 Related Literature

Related works to BONMI can be classified into two categories: (i) matrix completion and (ii) multi-source data integration. Matrix completion aims to recover a low-rank matrix given a subset of its entries which may be corrupted by noise (Keshavan et al., 2010; Candès and Recht, 2009). It has received considerable attention due to the diverse applications such as collaborative filtering (Hu et al., 2008; Rennie and Srebro, 2005) and recommenda-

tion systems. As reviewed in Nguyen et al. (2019), diverse algorithms have been proposed including Frobenius norm minimization (Lee and Bresler, 2010), alternative minimization (Haldar and Hernando, 2009; Tanner and Wei, 2016; Wen et al., 2012), optimization over smooth Riemannian manifold (Vandereycken, 2013) and stochastic gradient descent (Koren et al., 2009; Takács et al., 2007; Paterek, 2007; Sun and Luo, 2016; Ge et al., 2016, 2017; Du et al., 2017; Ma et al., 2018). However, most existing literature on matrix completion assumed that the observed entries are independently sampled (Keshavan et al., 2010; Chen and Wainwright, 2015; Candes and Plan, 2010; Candès and Tao, 2010; Mazumder et al., 2010; Chen, 2015; Keshavan et al., 2010; Chen and Wainwright, 2015; Zheng and Lafferty, 2016; Fazel, 2002; Candès and Recht, 2009; Cai et al., 2010; Tanner and Wei, 2013; Combettes and Pesquet, 2011; Jain et al., 2010, e.g.), which does not hold in the current setting as the missingness will always be block-wise.

On the other hand, many works on multi-source data integration analysis needed to deal with the block-wise missingness for downstream analyses, such as model selection (Xue and Qu, 2020), principal component analysis (PCA) (Cai et al., 2016; Zhu et al., 2018), classification (Yuan et al., 2012; Xiang et al., 2014) and prediction (Yu et al., 2020). However, these methods do not apply to the current problem since they need to use the patient-level data and have additional model assumptions such as a classification model (Yuan et al., 2012; Xiang et al., 2014) or a regression model (Yu et al., 2020; Xue and Qu, 2020). For example, Xue and Qu (2020) focused on the model selection when the covariates were block-wise missing due to incomplete observations. They assumed a linear model between the response and the covariates and showed the consistency of the estimation of the linear coefficients. Although our problem can also be modeled through a regression framework by using the observed entries to predict the missing entries, the independent assumption required by Xue and Qu (2020) will be violated. Specifically, Xue and Qu (2020) assumed not only that the covariates are independently sampled but also that the observation errors are independent and normally distributed. If we fit Xue and Qu (2020) to the current setting, both assumptions would be violated since the "covariates" and the errors are not independently sampled. Cai et al. (2016) proposed a structured matrix completion (SMC) algorithm that leverages the approximate low-rank structure to recover the missing off-diagonal sub-matrix efficiently. However, the SMC algorithm considers a noiseless scenario and does not allow for a multi-block missingness structure, ubiquitous in the integrative analysis of multi-source or multi-view data. Since SMC operates on a 2×2 block matrix with a missing block in the off-diagonal sub-matrices, it cannot fully utilize the observed information when applied to our problem. Approximation errors can also make SMC fail to perform well with a lack of theoretical guarantee. As demonstrated by our numerical studies, SMC performs poorly compared to BONMI in the presence of noise in the case of two sources.

1.3 Our Contribution

Our paper extends the word vector generative model in Arora et al. (2016) to accommodate multi-sources learning, which allows these sources to have overlapping but not identical entities. We design an efficient algorithm BONMI to aggregate the multi-source PMI matrices optimally with a theoretical guarantee to obtain multi-source embeddings. BONMI can also

be applied to other multi-source data integration problems with a similar data structure. A by-product of BONMI is the contribution to the field of matrix completion by considering the missing mechanism other than the entry-wise independent missing. We show that the entry-wise missing assumption, despite its prevalence in the works of matrix completion, is not necessary to guarantee recovery. We prove the statistical rate of our estimator, which is comparable to the rate under the independently missing assumption (Ma et al., 2018; Chen and Wainwright, 2015; Negahban and Wainwright, 2012; Koltchinskii et al., 2011).

The rest of the paper is organized as follows. In Section 2, we introduce in detail the proposed BONMI method. The theoretical properties of BONMI are analyzed in Section 3. Simulation results are shown in Section 4 to investigate the numerical performance of the proposed method. A real data application is given in Section 5. Section 6 extends the model to asymmetric matrices and concludes the paper. For space reasons, the proofs of the main results are given in the supplement. In addition, some key technical tools used in the proof of the main theorems are also developed and proved in the supplement.

2. Methodology

2.1 Notations

We first introduce some notations. We use bold-faced symbols to represent vectors and matrices. For any vector $\boldsymbol{v}$, $\|\boldsymbol{v}\|$ denotes its Euclidean norm. For any matrix $\mathbf{A} \in \mathbb{R}^{d\times q}$, we let $\sigma_j(\mathbf{A})$ and $\lambda_j(\mathbf{A})$ (if $d = q$) denote its respective jth largest singular value and eigenvalue. The smallest singular value $\sigma_{\min(m,n)}(\mathbf{A})$ will be denoted by $\sigma_{\min}(\mathbf{A})$. We let $\|\mathbf{A}\|$, $\|\mathbf{A}\|_{\mathrm{F}}$, $\|\mathbf{A}\|_{2,\infty}$ and $\|\mathbf{A}\|_{\infty}$ respectively denote the spectral norm (i.e., the largest singular value), the Frobenius norm, the ℓ_2/ℓ_∞ norm (i.e., the largest ℓ_2 norm of the rows), and the entry-wise ℓ_∞ norm (the largest magnitude of all entries) of $\mathbf{A}$. We let $\mathbf{A}_{j,\cdot}$ and $\mathbf{A}_{\cdot,j}$ denote the jth row and jth column of $\mathbf{A}$, and let $\mathbf{A}(i,j)$ denote the (i,j) entry of $\mathbf{A}$. Besides, we use the symbol $\equiv$ to denote 'defined to be.' For any integer $d \geq 1$, we let $[d] \equiv \{1, ..., d\}$. For indices sets $\Omega_1 \subseteq [d]$ and $\Omega_2 \subseteq [q]$, we use $\mathbf{A}_{\Omega_1,\Omega_2}$ to represent its sub-matrix with row indices Ω_1 and column indices Ω_2.

We let $\mathscr{O}^{n\times r}$ represent the set of all $n \times r$ orthonormal matrices. For a sub-Gaussian random variable Y, its sub-Gaussian norm is defined as $\|Y\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}e^{-Y^2/t^2} \leq 2\}$. We use the standard notation $f(n) = O(g(n))$ or $f(n) \lesssim g(n)$ to represent $|f(n)| \leq c|g(n)|$ for some constant $c > 0$.

2.2 Model

We extend the log-linear model word production model proposed by Arora et al. (2016) to multiple sources. Assume that we have m sources, and in the sth source, we have n_s independent sequences, which may be referred to as n_s patients for EHR data. Let $\mathcal{V}_s$ be the word set of the sth source with size $N_s = |\mathcal{V}_s|$. For simplicity, we assume that each sequence has length T. For the ith sequence from the sth source, the code sequence is $\{w_1^{(s,i)}, \ldots, w_T^{(s,i)}\}$. In short, for each sequence in the sth source, the occurrence probability of a code w at time t is determined by its latent vector $\mathbf{x}_w \in \mathbb{R}^r$ and a discourse vector

$\mathbf{c}_t^{(s,i)} \in \mathbb{R}^r$ with the random walk on the sphere. Specifically,

$$\mathbb{P}(w_t^{(s,i)} = w \mid \mathbf{c}_t^{(s,i)}) = \frac{\exp(\mathbf{x}_w^\top \mathbf{c}_t^{(s,i)})}{\sum_{w' \in \mathcal{V}_s} \exp(\mathbf{x}_{w'}^\top \mathbf{c}_t^{(s,i)})}\,.$$

Under this generative model, we are interested in estimating the clinical codes' embeddings $\{\mathbf{x}_w\}_{w \in \mathcal{V}^*}$ based on the summary statistics only such as the co-occurrence matrices to preserve the privacy where $\mathcal{V}^* = \cup_{s=1}^m \mathcal{V}_s$ is the corpus of the clinical features from the m sources. We further denote the corpus for all clinical features as $\mathcal{V}$, where $\mathcal{V}_s \subset \mathcal{V}$ is generated by a binomial model. For the sth source, the corpus $\mathcal{V}_s$ is a random subset of $\mathcal{V}$ sampled as

$$\mathbb{P}(w \in \mathcal{V}_s) = p_s \in (0,1), \text{ for } w \in \mathcal{V} \text{ and } s \in [m] \text{ independently.} \tag{1}$$

Notice that it is not necessary to have $\mathcal{V}^* = \mathcal{V}$. This model for $\{\mathcal{V}_s\}_{s=1}^m$ allows the emergence of new features which may have not been included by the current sources, such as COVID-19 (Zhou et al., 2022a). When a new source is incorporated, some new features in $\mathcal{V} \backslash \mathcal{V}^*$ can occur in the new source. The PMI matrix $\mathbf{PMI} = \{\mathrm{PMI}(w,w')\}_{w,w' \in \mathcal{V}}$ of the population corpus $\mathcal{V}$ is defined as

$$\mathrm{PMI}(w,w') = \log \frac{p(w,w')}{p(w)p(w')}, \text{ for } w, w' \in \mathcal{V}\,,$$

where $p(w)$ is the occurrence probability of the word w and $p(w,w')$ is the co-occurrence probability of the words w and w'. Arora et al. (2016) showed that $\log p(w,w') = \|\mathbf{x}_w + \mathbf{x}_{w'}\|^2/(2r) - 2\log Z + o(1)$ and $\log p(w) = \|\mathbf{x}_w\|^2/(2r) - \log Z + o(1)$ for some constant Z, then derived that $\mathrm{PMI}(w,w') \approx \mathbf{x}_w^\top \mathbf{x}_{w'}/r$ (see, e.g., Theorem 2.2 of Arora et al. (2016) and Proposition 4.4 of Lu et al. (2023)), which implies the rationale of recovering word embeddings from the PMI matrices. With a bit of abuse of notation, we also define $\mathbf{x}_w = \mathbf{x}_w/\sqrt{r}$ as the word embedding. With a single hospital, **PMI**, while not directly observable, can be estimated empirically, and hence performing an SVD of the empirical **PMI** can lead to consistent estimators for $\mathbf{x}_w$. However, in the settings where different hospitals have overlapping but non-identical codes, entries of **PMI** can not be directly estimated for code pairs that do not belong to the same hospital.

On the other hand, the principal sub-matrices of **PMI** can be estimated from each source. Let $p_s(w,w')$ be the co-occurrence probability of codes w and w' in windows of size q in the sth source, $p_s(w) = \sum_{w' \in \mathcal{V}_s} p_s(w,w')$ and the population PMI matrix for the sth source $\mathbf{PMI}_s \in \mathbb{R}^{N_s \times N_s}$ as

$$\mathrm{PMI}_s(w,w') = \log \frac{p_s(w,w')}{p_s(w)p_s(w')} \quad \text{for } w, w' \in \mathcal{V}_s\,.$$

The estimates of the PMI matrices using the co-occurrence statistics are

$$\widehat{\mathrm{PMI}}_s(w,w') = \log \frac{\mathcal{C}_s(w,w')}{\mathcal{C}_s(w,\cdot)\mathcal{C}_s(w',\cdot)}\,, \text{ for } s \in [m]\,,$$

where $\mathcal{C}_s = \left[\mathcal{C}_s(w, w')\right]_{w,w' \in \mathcal{V}_s}$ is the observed co-occurrence of word w with word w' in the window size q across all sequences of the sth source defined similar to Beam et al. (2019); Lu et al. (2023):

$$\mathcal{C}_s(w, w') = |\{(t, h) : |t - h| \leq q \text{ and } w_t^{(s,i)} = w, w_h^{(s,i)} = w' | t, h \in [T], i \in [n_s]\}|$$

and $\mathcal{C}_s(w, \cdot) = \sum_{w' \in \mathcal{V}_s} \mathcal{C}_s(w, w')$. We then define the shifted positive PMI (SPPMI) matrix estimator as

$$\widehat{\mathrm{SPPMI}}_s(w, w') = \max\{\widehat{\mathrm{PMI}}_s(w, w'), \eta\}, \text{ for } s \in [m]\,,$$

where $\eta > -\infty$ is a given threshold value. In the theoretical analysis, one may show that $\mathrm{PMI}_s(w, w')$ is lower bounded by some constant with high probability under appropriate assumptions (Lu et al., 2023). So the shifted positive PMI would be closer to the truth than the original PMI with a high probability if η is chosen properly. According to Levy and Goldberg (2014), the shifted positive PMI (SPPMI) can perform better than the PMI matrix, with the reasoning being that humans tend to more easily associate positive values (e.g. 'Canada' and 'snow') rather than negative ones ('Canada' and 'desert'). The default choice for η is set as 0, meaning no shift and setting negative PMI values as 0. Empirically, we find that $\eta = 0$ works well. For ease of presentation, we will use the positive PMI matrix (PPMI) matrix estimator as

$$\widehat{\mathrm{PPMI}}_s(w, w') = \max\{\widehat{\mathrm{PMI}}_s(w, w'), 0\}, \text{ for } s \in [m]$$

throughout the paper, while our theorems still hold for other choices of η. Let $\mathbf{X}_s \in \mathbb{R}^{N_s \times r}$ be the matrix whose rows are composed of the word embeddings in the sth source. Define the error matrix $\mathbf{E}^s = \widehat{\mathbf{PPMI}}_s - \mathbf{X}_s\mathbf{X}_s^\top$. The estimated PPMI matrices approximate the population PMI matrices thus approximating $\mathbf{X}_s\mathbf{X}_s^\top$ such that

$$\|\mathbf{E}^s\|_\infty \lesssim n_s^{-\frac{1}{4}} T^{-\frac{1}{2}} \quad \text{and} \quad \|\mathbf{E}^s\| \lesssim N_s n_s^{-\frac{1}{4}} T^{-\frac{1}{2}}\,,$$

which follows straightforwardly from Lu et al. (2023) under some mild assumptions. In reality, n_s can be sufficiently larger than N_s. For example, in the EHR system, the number of clinical codes is smaller than $10,000$ while the number of patients can be 23 million (Zhou et al., 2022a).

Without loss of generality, we assume that $\mathcal{V} = [N]$ and $\mathcal{V}^* = [N_0]$, where $N_0 = |\mathcal{V}^*|$, otherwise we can rearrange the orders of codes. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\top \in \mathbb{R}^{N \times r}$ be the population embedding matrix and $\mathbf{W}^* = \mathbf{X}\mathbf{X}^\top$. We then have

$$\mathbf{W}^s \equiv \widehat{\mathbf{PPMI}}_s = \mathbf{W}_s^* + \mathbf{E}^s = \left[\mathbf{W}^*(i, j)\right]_{j \in \mathcal{V}_s}^{i \in \mathcal{V}_s} + \mathbf{E}^s, \text{ for } s \in [m] \tag{2}$$

by the definition of $\mathbf{E}^s$. Define $\sigma_s = \|\mathbf{E}^s\|/\sqrt{N_s}$. In our theoretical analysis, we only use the operator norm of the error matrices $\mathbf{E}^s$. Therefore, our results can be applied to general matrix integration problems.

Our task is to first recover $\mathbf{W}_0^* = \mathbf{W}_{\mathcal{V}^*,\mathcal{V}^*}^* = \left[\mathbf{W}^*(i, j)\right]_{j \in \mathcal{V}^*}^{i \in \mathcal{V}^*} \in \mathbb{R}^{N_0 \times N_0}$, and then obtain the embeddings $\mathbf{X}^* = (\mathbf{x}_1, \ldots, \mathbf{x}_{N_0})^\top$ by performing SVD on the estimate of $\mathbf{W}_0^*$. Let the eigendecomposition of $\mathbf{W}^*$ be

$$\mathbf{W}^* = \mathbf{U}^*\mathbf{\Sigma}^*(\mathbf{U}^*)^\top, \tag{3}$$

where $\mathbf{U}^* \in \mathbb{R}^{N\times r}$ consists of orthonormal columns, and $\mathbf{\Sigma}^*$ is an $r \times r$ diagonal matrix with eigenvalues in a descending order, i.e., $\lambda_{\max} = \lambda_1 \geq \cdots \geq \lambda_r = \lambda_{\min} > 0$.

We first state two assumptions that are standard in existing literature (Candès and Recht, 2009; Ma et al., 2018) for sample complexity and the incoherence condition which basically assumes information is distributed uniformly among entries.

Assumption 1 (Incoherence condition) *The coherence coefficient of* $\mathbf{U}^*$ *satisfies* $\mu_0 = O(1)$*, where* $\mu_0 = \mu(\mathbf{U}^*) = \frac{N}{r}\max_{i\in[N]}\sum_{j=1}^{r}\mathbf{U}^*(i,j)^2$.

Assumption 2 (Sample complexity) *The sampling probability* $p_0 = \min_{s\in[m]} p_s$ *satisfies* $p_0 \geq C\sqrt{\mu_0 r \log N/N}$ *for some sufficiently large constant* C*. Besides,* $\max_{s\in[m]} p_s/p_0 = O(1)$.

Remark 1 *Based on the observed data, we can estimate* $\mathbf{W}_0^*$ *but not* $\mathbf{W}^*$*, as we have no information on* $\{\mathbf{x}_w\}_{w\in\mathcal{V}\setminus\mathcal{V}^*}$*. The inclusion of* $\mathbf{W}^*$ *and the sampling model* (1) *of* $\{\mathcal{V}_s\}_{s=1}^m$ *serves as a convenient random setup and links to the matrix completion literature. In reality, the overlapping matrices are determined by dictionaries linking multiple sources or ontologies such as the ICD and CPT codes commonly adopted in the EHR. Under the sampling model* (1)*,* $\mathbf{W}_0^*$ *is a random sub-matrix of* $\mathbf{W}^*$*. Instead of making assumptions on* $\mathbf{W}_0^*$*, which is a random object, we impose assumptions on* $\mathbf{W}^*$ *(i.e., Assumption 1 above and Assumption 4 in Section 3).*

2.3 An Ideal Case

To illustrate the BONMI algorithm, we first consider an ideal case that the error matrices $\{\mathbf{E}^s\}_{s=1}^m$ are zero and we observe the truth $\{\mathbf{W}_s^*\}_{s=1}^m$ when $m = 2$. To simplify the notations, we denote $s\backslash k \equiv \mathcal{V}_s\backslash\mathcal{V}_k$ and $s\cap k \equiv \mathcal{V}_s\cap\mathcal{V}_k$ when they are used as the subscripts of a matrix, and recall that $\mathbf{W}_s^* \equiv \mathbf{W}^*_{\mathcal{V}_s,\mathcal{V}_s}$. Assume the two sampled sub-matrices are $\mathbf{W}_s^*$ and $\mathbf{W}_k^*$. Since the singular values are invariant under row/column permutations, without loss of generality, we can rearrange our data matrices such that

$$\mathbf{W}_s^* = \begin{bmatrix} \mathbf{W}^*_{s\backslash k,s\backslash k} & \mathbf{W}^*_{s\backslash k,s\cap k} \\ \mathbf{W}^*_{s\cap k,s\backslash k} & \mathbf{W}^*_{s\cap k,s\cap k} \end{bmatrix}; \quad \mathbf{W}_k^* = \begin{bmatrix} \mathbf{W}^*_{s\cap k,s\cap k} & \mathbf{W}^*_{s\cap k,k\backslash s} \\ \mathbf{W}^*_{k\backslash s,s\cap k} & \mathbf{W}^*_{k\backslash s,k\backslash s} \end{bmatrix} \tag{4}$$

and

$$\mathbf{W}_0^* = \begin{bmatrix} \mathbf{W}^*_{s\backslash k,s\backslash k} & \mathbf{W}^*_{s\backslash k,s\cap k} & \mathbf{W}^*_{s\backslash k,k\backslash s} \\ \mathbf{W}^*_{s\cap k,s\backslash k} & \mathbf{W}^*_{s\cap k,s\cap k} & \mathbf{W}^*_{s\cap k,k\backslash s} \\ \mathbf{W}^*_{k\backslash s,s\backslash k} & \mathbf{W}^*_{k\backslash s,s\cap k} & \mathbf{W}^*_{k\backslash s,k\backslash s} \end{bmatrix}. \tag{5}$$

Recall that our goal is to recover $\mathbf{W}_0^*$ based on the observed $\mathbf{W}_s^*$ and $\mathbf{W}_k^*$. This can be achieved by estimating the missing blocks $\mathbf{W}^*_{s\backslash k,k\backslash s}$ and $\mathbf{W}^*_{k\backslash s,s\backslash k} = \mathbf{W}^{*\top}_{s\backslash k,k\backslash s}$ by the symmetry of $\mathbf{W}_0^*$. As the missing entries are block-wise, a theoretical guarantee based on the assumption of independent missing will fail in the current case. Instead, we propose a method based on the orthogonal transformation, which exploits the following proposition.

Proposition 2 *Suppose* $\mathbf{W}^*$ *has eigendecomposition* (3) *and satisfies Assumptions 1 and 2. Since* $\max\left\{\text{rank}(\mathbf{W}_s^*), \text{rank}(\mathbf{W}_k^*)\right\} \leq \text{rank}(\mathbf{W}^*) = r$*, we suppose the eigendecompositions of* $\mathbf{W}_s^*$ *and* $\mathbf{W}_k^*$ *are*

$$\mathbf{W}_s^* = \mathbf{V}_s^*\mathbf{\Sigma}_s^*(\mathbf{V}_s^*)^\top \text{ and } \mathbf{W}_k^* = \mathbf{V}_k^*\mathbf{\Sigma}_k^*(\mathbf{V}_k^*)^\top,$$

where $\mathbf{V}_s^*$ *and* $\mathbf{V}_k^*$ *are the eigenvectors of* $\mathbf{W}_s^*$ *and* $\mathbf{W}_k^*$*, respectively. We further decompose* $\mathbf{V}_s^*$ *and* $\mathbf{V}_k^*$ *as* $\mathbf{V}_s^* = ((\mathbf{V}_{s1}^*)^\top, (\mathbf{V}_{s2}^*)^\top)^\top$, $\mathbf{V}_k^* = ((\mathbf{V}_{k1}^*)^\top, (\mathbf{V}_{k2}^*)^\top)^\top$ *with* $\mathbf{V}_{s2}^*, \mathbf{V}_{k1}^* \in \mathbb{R}^{|\mathcal{V}_s \cap \mathcal{V}_k| \times r}$*. Then with probability at least* $1 - O(1/N^3)$, $\mathbf{W}_{s\setminus k, k\setminus s}^*$ *in* (5) *can be exactly given by*

$$\mathbf{W}_{s\setminus k,k\setminus s}^* = \mathbf{V}_{s1}^*(\mathbf{\Sigma}_s^*)^{1/2}\mathbf{G}((\mathbf{\Sigma}_s^*)^{1/2}(\mathbf{V}_{s2}^*)^\top \mathbf{V}_{k1}^*(\mathbf{\Sigma}_k^*)^{1/2})(\mathbf{\Sigma}_k^*)^{1/2}(\mathbf{V}_{k2}^*)^\top, \tag{6}$$

where $\mathbf{G}(\cdot)$ *is a matrix value function defined as:* $\mathbf{G}(\mathbf{C}) = \mathbf{H}\mathbf{Z}^\top$ *with* $\mathbf{H}\mathbf{\Omega}\mathbf{Z}^\top$ *the SVD of* $\mathbf{C} \in \mathbb{R}^{r\times r}$.

Proposition 2 shows that, when there is no error, $\mathbf{W}_{s\setminus k,k\setminus s}^*$ can be recovered precisely based on $\mathbf{W}_s^*$ and $\mathbf{W}_k^*$ with high probability. The proposition can be easily extended to the case when $m > 2$.

2.4 BONMI Algorithm for Two Sources

When noise exists, we use the above idea but add an additional step of weighted average. Since it is possible to observe the entries of $\mathbf{W}^*$ more than once due to multiple sources, the weighted average is a natural idea to reduce the variance of estimation in the existence of noise. In reality, heterogeneity always exists which means the errors of different sources may vary. As a result, we decide to use the weights inversely proportional to the error levels. We start with the case $m = 2$ again. Currently, we decompose two overlapping matrices $\mathbf{W}^s = \mathbf{W}_s^* + \mathbf{E}^s$ and $\mathbf{W}^k = \mathbf{W}_k^* + \mathbf{E}^k$ as follows

$$\mathbf{W}^s = \begin{bmatrix} \mathbf{W}^s_{s\setminus k,s\setminus k} & \mathbf{W}^s_{s\setminus k,s\cap k} \\ \mathbf{W}^s_{s\cap k,s\setminus k} & \mathbf{W}^s_{s\cap k,s\cap k} \end{bmatrix}, \quad \mathbf{W}^k = \begin{bmatrix} \mathbf{W}^k_{s\cap k,s\cap k} & \mathbf{W}^k_{s\cap k,k\setminus s} \\ \mathbf{W}^k_{k\setminus s,s\cap k} & \mathbf{W}^k_{k\setminus s,k\setminus s} \end{bmatrix}, \quad \text{for } 1 \le s < k \le m.$$

Then we can combine $\mathbf{W}_s$ and $\mathbf{W}_k$ to obtain

$$\widetilde{\mathbf{W}} = \begin{bmatrix} \mathbf{W}^s_{s\setminus k,s\setminus k} & \mathbf{W}^s_{s\setminus k,s\cap k} & \mathbf{0} \\ \mathbf{W}^s_{s\cap k,s\setminus k} & \mathbf{W}^a_{s\cap k,s\cap k} & \mathbf{W}^k_{s\cap k,k\setminus s} \\ \mathbf{0} & \mathbf{W}^k_{k\setminus s,s\cap k} & \mathbf{W}^k_{k\setminus s,k\setminus s} \end{bmatrix}, \tag{7}$$

where $\mathbf{W}^a_{s\cap k,s\cap k} \equiv \alpha_s \mathbf{W}^s_{s\cap k,s\cap k} + \alpha_k \mathbf{W}^k_{s\cap k,s\cap k}$ is the weighted average of the overlapping part with $\alpha_i > 0, i = s, k$ and $\alpha_s + \alpha_k = 1$. The weights should ideally depend on the strength of the error matrices, $\mathbf{E}^s$ and $\mathbf{E}^k$, to optimize estimation. We detail the estimation of the weights in Section 2.5. To estimate $\mathbf{W}^*_{s\setminus k,k\setminus s}$, let

$$\widetilde{\mathbf{W}}_s = \begin{bmatrix} \mathbf{W}^s_{s\setminus k,s\setminus k} & \mathbf{W}^s_{s\setminus k,s\cap k} \\ \mathbf{W}^s_{s\cap k,s\setminus k} & \mathbf{W}^a_{s\cap k,s\cap k} \end{bmatrix} \quad \text{and} \quad \widetilde{\mathbf{W}}_k = \begin{bmatrix} \mathbf{W}^a_{s\cap k,s\cap k} & \mathbf{W}^k_{s\cap k,k\setminus s} \\ \mathbf{W}^k_{k\setminus s,s\cap k} & \mathbf{W}^k_{k\setminus s,k\setminus s} \end{bmatrix},$$

and the rank-r eigendecompositions of $\widetilde{\mathbf{W}}_s$ and $\widetilde{\mathbf{W}}_k$ be $\widetilde{\mathbf{V}}_s\widetilde{\mathbf{\Sigma}}_s\widetilde{\mathbf{V}}_s^\top$ and $\widetilde{\mathbf{V}}_k\widetilde{\mathbf{\Sigma}}_k\widetilde{\mathbf{V}}_k^\top$, respectively. Specifically, $\widetilde{\mathbf{V}}_s$ and $\widetilde{\mathbf{V}}_k$ can be decomposed block-wise such that $\widetilde{\mathbf{V}}_s = (\widetilde{\mathbf{V}}_{s1}^\top, \widetilde{\mathbf{V}}_{s2}^\top)^\top$ and $\widetilde{\mathbf{V}}_k = (\widetilde{\mathbf{V}}_{k1}^\top, \widetilde{\mathbf{V}}_{k2}^\top)^\top$ where $\widetilde{\mathbf{V}}_{s2}, \widetilde{\mathbf{V}}_{k1} \in \mathbb{R}^{|\mathcal{V}_s\cap\mathcal{V}_k|\times r}$. So the estimator of $\mathbf{W}^*_{s\setminus k,k\setminus s}$ is

$$\widetilde{\mathbf{W}}_{sk} = \widetilde{\mathbf{V}}_{s1}\widetilde{\mathbf{\Sigma}}_s^{1/2}\mathbf{G}(\widetilde{\mathbf{\Sigma}}_s^{1/2}\widetilde{\mathbf{V}}_{s2}^\top\widetilde{\mathbf{V}}_{k1}\widetilde{\mathbf{\Sigma}}_k^{1/2})\widetilde{\mathbf{\Sigma}}_k^{1/2}\widetilde{\mathbf{V}}_{k2}^\top, \tag{8}$$

according to the Proposition 2. After getting $\widetilde{\mathbf{W}}_{sk}$, we impute it back to $\widetilde{\mathbf{W}}$ to obtain

$$\widehat{\mathbf{W}} = \begin{bmatrix} \mathbf{W}^s_{s\setminus k,s\setminus k} & \mathbf{W}^s_{s\setminus k,s\cap k} & \widetilde{\mathbf{W}}_{sk} \\ \mathbf{W}^s_{s\cap k,s\setminus k} & \mathbf{W}^a_{s\cap k,s\cap k} & \mathbf{W}^k_{s\cap k,k\setminus s} \\ \widetilde{\mathbf{W}}_{sk}^\top & \mathbf{W}^k_{k\setminus s,s\cap k} & \mathbf{W}^k_{k\setminus s,k\setminus s} \end{bmatrix}. \tag{9}$$

Then we can obtain the rank-r eigendecomposition of $\widehat{\mathbf{W}}$, denoted as $\widehat{\mathbf{W}}_r = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{U}}^\top$, as an estimate of $\mathbf{W}_0^*$.

2.5 BONMI Algorithm

We next introduce the BONMI algorithm for recovering $\mathbf{W}_0^*$, based on $m \geq 2$ PPMI matrices $\{\mathbf{W}^s\}_{s\in[m]}$. Our algorithm consists of three main steps: (I) aggregation of the m matrices, (II) estimation of missing parts, and (III) low-rank approximation, as summarized in Algorithm 1.

Algorithm 1: **B**lock-wise **O**verlapping **N**oisy **M**atrix **I**ntegration (BONMI).

Input: m symmetric matrices $\{\mathbf{W}^s\}_{s\in[m]}$ and the corresponding index sets $\{\mathcal{V}_s\}_{s\in[m]}$; the rank r; $N_0 = |\cup_{s=1}^m \mathcal{V}_s|$;

Step I (a) Estimation of weights: for $1 \leq s \leq m$ **do**

Let $\widehat{\mathbf{U}}_s\widehat{\mathbf{\Sigma}}_s(\widehat{\mathbf{U}}_s)^\top$ be the rank-r eigendecomposition of $\mathbf{W}^s$. Estimate σ_s by

$$\widehat{\sigma}_s = \|\mathbf{W}^s - \widehat{\mathbf{U}}_s\widehat{\mathbf{\Sigma}}_s(\widehat{\mathbf{U}}_s)^\top\|/\sqrt{N_s}; \tag{10}$$

end

Step I (b) Aggregation: Create $\widetilde{\mathbf{W}} \in \mathbb{R}^{N_0\times N_0}$ by (11).

Step II (a) Spectral initialization: for $1 \leq s \leq m$ **do**

Let $\widetilde{\mathbf{V}}_s\widetilde{\mathbf{\Sigma}}_s\widetilde{\mathbf{V}}_s^\top$ be the rank-r eigendecomposition of $\widetilde{\mathbf{W}}_{\mathcal{V}_s\mathcal{V}_s}$.

end

Step II (b) Estimation of missing parts: for $1 \leq s < k \leq m$ **do**

Obtain $\widetilde{\mathbf{W}}_{sk}$ using $\widetilde{\mathbf{V}}_s, \widetilde{\mathbf{\Sigma}}_s, \widetilde{\mathbf{V}}_k, \widetilde{\mathbf{\Sigma}}_k$ by (8). If a missing entry (i,j) is estimated by multiple pairs of sources (s,k), choose the one estimated by the pair with the smallest $\widehat{\sigma}_s^2 + \widehat{\sigma}_k^2$. Denote the imputed matrix as $\widehat{\mathbf{W}}$.

end

Step III Low rank approximation: Obtain the rank-r eigendecomposition of $\widehat{\mathbf{W}}$: $\widehat{\mathbf{W}}_r = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{U}}^\top$.

Output: $\widehat{\mathbf{X}} = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}$.

Step I: Aggregation. We first aggregate $\{\mathbf{W}^s\}_{s\in[m]}$ to obtain $\widetilde{\mathbf{W}}$ similar to the $m = 2$ case, which requires an estimation for the weights $\{\alpha_s\}_{s\in[m]}$. Similar to standard meta-analysis, the optimal weight for the sth source can be chosen as σ_s^{-2}. We estimate σ_s as by $\widehat{\sigma}_s = \|\mathbf{W}^s - \widehat{\mathbf{U}}_s\widehat{\mathbf{\Sigma}}_s\widehat{\mathbf{U}}_s^\top\|/\sqrt{N_s}$, where $\widehat{\mathbf{U}}_s\widehat{\mathbf{\Sigma}}_s\widehat{\mathbf{U}}_s^\top$ is the rank-r eigendecomposition of $\mathbf{W}^s$. We then create the matrix $\widetilde{\mathbf{W}} \in \mathbb{R}^{n\times n}$ as follows

$$\widetilde{\mathbf{W}}(i,j) = \sum_{s=1}^m \alpha_{ij}^s \mathbf{W}^s(v_i^s, v_j^s)\mathbb{1}(i,j \in \mathcal{V}_s), \tag{11}$$

for all pairs of (i,j) such that $\mathcal{S}_{ij} \equiv \sum_{s=1}^m \mathbb{1}(i,j \in \mathcal{V}_s) > 0$, where v_i^s denotes the row(column) index in $\mathbf{W}^s$ corresponding to the ith row(column) of $\mathbf{W}_0^*$, and $\alpha_{ij}^s = \frac{1}{\widehat{\sigma}_s^2}\big(\sum_{k=1}^m \mathbb{1}(i,j \in \mathcal{V}_k)\widehat{\sigma}_k^{-2}\big)^{-1}$. The entries in the missing blocks with $\mathcal{S}_{ij} = 0$ are initialized as zero.

Remark 3 *It is natural to use the inverse of noise variances as the weights to aggregate multiple observations, for instance, weighted least squares (Ruppert and Wand, 1994). Here we follow the same routine, and the choice is direct since, intuitively, in this way, we can minimize the variance of the noise of the overlapping matrices. A formal analysis is provided in Section S.2.3.*

Step II: Imputation. We next impute the missing entries with $\mathcal{S}_{ij} = 0$. For $1 \le s \le k \le m$, we impute the entries of $\widetilde{\mathbf{W}}$ corresponding to $(\mathcal{V}_s \backslash \mathcal{V}_k) \times (\mathcal{V}_k \backslash \mathcal{V}_s)$ using $\widetilde{\mathbf{W}}_s \equiv \widetilde{\mathbf{W}}_{\mathcal{V}_s, \mathcal{V}_s}$ and $\widetilde{\mathbf{W}}_k$ the same way as (8). If a missing entry (i, j) can be estimated by multiple pairs of sources (s, k), we choose the one estimated by the pair with the smallest $\widehat{\sigma}_s^2 + \widehat{\sigma}_k^2$. After Steps I and II, all missing entries of $\widetilde{\mathbf{W}}$ are imputed, and we denote the imputed matrix as $\widehat{\mathbf{W}}$.

Remark 4 *When a missing entry can be estimated from multiple source pairs, it may be desirable to use a weighted average of these estimates. However, determining an optimal set of weights is more challenging in this case compared to Step I because the variances of these estimates are difficult to estimate. We instead choose the estimate based on the source pairs with the highest overall quality as measured by $\widehat{\sigma}_s^2 + \widehat{\sigma}_k^2$.*

Step III: Low-raxnk approximation. Finally, we factorize $\widehat{\mathbf{W}}$ by rank-r eigendecomposition to obtain the final estimator: $\widehat{\mathbf{W}}_r \equiv \widehat{\mathbf{U}} \widehat{\mathbf{\Sigma}} \widehat{\mathbf{U}}^\top$.

Remark 5 (Computational complexity) *The main computational cost of the BONMI algorithm comes from the eigendecomposition, which is of order $O(\sum_{s=1}^m |\mathcal{V}_s|^2 r)$. The estimation step involves matrix multiplication and a singular value decomposition of a $r \times r$ matrix for each source pair s and k, resulting in a computational cost of order $O(\sum_{1 \le s < k \le m} |\mathcal{V}_s||\mathcal{V}_k| r)$. Thus, the overall computational cost is of order*

$$O\big(\textstyle\sum_{s=1}^m |\mathcal{V}_s|^2 r + \sum_{1 \le s < k \le m} |\mathcal{V}_s||\mathcal{V}_k| r\big) = O\big((\textstyle\sum_{s=1}^m |\mathcal{V}_s|)^2 r\big)\,.$$

In comparison, the computational complexity of the gradient descent-based algorithms is $O\big(T(\sum_{s=1}^m |\mathcal{V}_s|)^2 r\big)$, where T is the iteration complexity dependent on the pre-set precision ϵ. Existing algorithms set $T = n/r \log(1/\epsilon)$ (Sun and Luo, 2016), $T = r^2 \log(1/\epsilon)$ (Chen and Wainwright, 2015), and $T = \log(1/\epsilon)$ (Ma et al., 2018). Thus, the BONMI algorithm is more computationally efficient compared to these algorithms.

3. Theoretical Analysis

In this section, we investigate the theoretical properties of the algorithm. We first present some general assumptions required by our theorems. To this end, we define the condition number $\tau \equiv \lambda_1(\mathbf{W}^*)/\lambda_r(\mathbf{W}^*) = \lambda_{\max}/\lambda_{\min}$. Besides, we need conditions to bound the noise strength and the condition number.

Assumption 3 *Let $\sigma \equiv \max_{s \in [m]} \sigma_s$. Then σ satisfies $\sigma \ll \lambda_{\min} \sqrt{p_0/N}$.*

Assumption 4 *$\tau \equiv \lambda_1(\mathbf{W}^*)/\lambda_r(\mathbf{W}^*) = \lambda_{\max}/\lambda_{\min} = O(1)$. Throughout this paper, we assume the condition number is bounded by a fixed constant, independent of the problem size (i.e., N and r).*

Remark 6 *Assumptions 1-4 are standard assumptions in many existing literature (Ma et al., 2018; Chen and Wainwright, 2015; Negahban and Wainwright, 2012; Koltchinskii et al., 2011), whereas different rates are required. Specifically, in Assumption 2, we only require the sampling probability to be of the order $O(\sqrt{\log N/N})$, which can tend to zero when the population size tends to infinity. In our setting, the sample size of each source is about $N^2p_0^2$. Then we have $N^2p_0^2 \geq C^2\mu_0 rN\log N$. Relatively, Ma et al. (2018) requires that the sample size satisfies $N^2p \geq C\mu_0^3 r^3 N\log^3 N$ for some sufficiently large constant $C>0$ where p is the entrywise sampling probability. In Assumption 3, the sampling probability p_0 and the eigenvalue $\lambda_{\min}$ can vary with N. Compared to Ma et al. (2018), they require the noise satisfies $\sigma \ll \lambda_{\min}\sqrt{\frac{p}{N\kappa^3\mu_0 r\log^3 N}}$. Our signal-to-noise ratio assumption has the same order as theirs up to some constants and log factors since μ_0, r, and κ are assumed to be constants.*

The parameter of interest is $\mathbf{W}_0^*$ with eigendecomposition $\mathbf{W}_0^* = \mathbf{X}^*(\mathbf{X}^*)^\top = \mathbf{U}_0^*\mathbf{\Sigma}_0^*(\mathbf{U}_0^*)^\top$. Let $\widehat{\mathbf{X}} = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}^{1/2}$ be the output of Algorithm 1 and define $K = r\mu_0\tau$. The upper bound for the estimation errors of $\mathbf{X}^*$ (which is identifiable up to an orthogonal transformation) and hence $\mathbf{W}_0^*$ under the special case of $m=2$ is presented in Theorem 7.

Theorem 7 *Under Assumptions 1, 2, 3, and 4, when $m=2$, with probability at least $1-O(N^{-3})$, there exists $\mathbf{O}_X \in \mathscr{O}^{r\times r}$ such that*

- *if $p_0 = o(1/\log N)$ or p_0 is bounded away from 0, we have*

$$\|\widehat{\mathbf{X}}\mathbf{O}_X - \mathbf{X}^*\| \lesssim \frac{\{(1-p_0)K^2+1\}K}{\sqrt{\lambda_{\min}}}\sqrt{N}\sigma; \tag{12}$$

- *otherwise,*

$$\|\widehat{\mathbf{X}}\mathbf{O}_X - \mathbf{X}^*\| \lesssim \frac{\{(1-p_0)K^2(p_0\log N)+1\}K}{\sqrt{\lambda_{\min}}}\sqrt{N}\sigma. \tag{13}$$

Remark 8 *Besides the spectral norm upper bound for $\widehat{\mathbf{X}}$, we can also obtain upper bound of other metrics such as $\|\widehat{\mathbf{X}}\widehat{\mathbf{X}}^\top - \mathbf{X}^*(\mathbf{X}^*)^\top\|$, $\|\widehat{\mathbf{X}}\mathbf{O}_X - \mathbf{X}^*\|_{\mathrm{F}}$ and $\|\widehat{\mathbf{X}}\widehat{\mathbf{X}}^\top - \mathbf{X}^*(\mathbf{X}^*)^\top\|_{\mathrm{F}}$ using the following inequalities $\|\widehat{\mathbf{X}}\mathbf{O}_X - \mathbf{X}^*\|_{\mathrm{F}} \leq \sqrt{2r}\|\widehat{\mathbf{X}}\mathbf{O}_X - \mathbf{X}^*\|$ and $\|\widehat{\mathbf{X}}\widehat{\mathbf{X}}^\top - \mathbf{X}^*(\mathbf{X}^*)^\top\|_{\mathrm{F}} \leq \sqrt{2r}\|\widehat{\mathbf{X}}\widehat{\mathbf{X}}^\top - \mathbf{X}^*(\mathbf{X}^*)^\top\| \leq 2\sqrt{2r}\|\widehat{\mathbf{W}} - \mathbf{X}^*(\mathbf{X}^*)^\top\|$, where the bound of $\|\widehat{\mathbf{W}} - \mathbf{X}^*(\mathbf{X}^*)^\top\|$ is derived in the proof of Theorem 7.*

Remark 9 *Here we compare our result with the state of art result in matrix completion literature (Ma et al., 2018) under the random missing condition. However, we should notice that their theorems don't hold under the current missing pattern since their entrywise independent sampling assumption is violated. Their operator norm error converges to*

$$\|\widehat{\mathbf{X}}\mathbf{O}_X - \mathbf{X}^*\| \lesssim \frac{\sigma}{\lambda_{\min}(\mathbf{W}_0^*)}\sqrt{\frac{N_0}{p}}\|\mathbf{X}^*\|, \tag{14}$$

where p is the entrywise sampling probability under their setting. We can show that $p \approx 1-2(p_0-p_0^2)^2/(2p_0-p_0^2)^2 = (2-p_0^2)/(2-p_0)^2$, $N_0 \approx N(2p_0-p_0^2) \approx Np_0$, $\lambda_{\min}(\mathbf{W}_0^*) \approx p_0\lambda_{\min}$

and $\|\mathbf{X}^*\| \approx \sqrt{p_0 r \mu_0 \lambda_{\max}}$ *(see the proof of Theorem 7). As a result, their error bound* (14) *reduces to*

$$\|\widehat{\mathbf{X}}\mathbf{O}_X - \mathbf{X}^*\| \lesssim \frac{(2-p_0)\sqrt{K}}{\sqrt{\lambda_{\min}}}\sqrt{N}\sigma. \tag{15}$$

When $p_0 \to 1$*, our rate is* (12)*, which has a difference with* (15) *in the order of* $\sqrt{K}$*; when* $p_0 \to 0$*, our rate is* (12) *or* (13)*, which has the difference with* (15) *in the order of* $K^{5/2}\max\{1, p_0 \log N\}$*. It means that our rate is the same as theirs up to some constants or log factors, which means that the error bound can be similar even under different sampling scenarios. The additional factor* K *may be caused by the dependence of the sampling pattern.*

Remark 10 *The minimax lower bound for matrix completion has been established under the random missing setting (Candès and Tao, 2010; Koltchinskii et al., 2011; Lounici, 2011; Negahban and Wainwright, 2012). For example, the lower bound for* $\inf_{\widehat{\mathbf{W}}_0}\sup_{\mathbf{W}_0^*\in\mathcal{W}}\|\widehat{\mathbf{W}}_0 - \mathbf{W}_0^*\|_2$ *is* $\sigma\sqrt{N_0 r/p}$ *(Lounici, 2011, Theorem 3), where* $\mathcal{W} = \{\mathbf{W}_0^* \in \mathbb{R}^{N_0\times N_0} : \|\mathbf{W}_0\|_\infty \le a, \mathrm{rank}(\mathbf{W}_0^*) \le r\}$*,* σ^2 *is the variance of the Gaussian noise in the observations and* p *is their entry-wise sampling probability. If the rate is adapted to our setting, it can be rewritten as* $\sqrt{Np_0 r}(2-p_0)\sigma$ *following the analysis of Remark 9. While in the proof of Theorem 7, we also show that* $\widehat{\mathbf{W}}$ *defined in* (9) *satisfies* $\|\widehat{\mathbf{W}} - \mathbf{W}_0^*\|_2 \lesssim \sqrt{Np_0}(2-p_0)K^2\sigma$*. Thus, our upper bound matches the minimax rate with regard to the sample size* N *and the sampling probability* p_0*, with a difference of the factor* $K^2/\sqrt{r} = r^{3/2}\mu_0^2\tau^2$*.*

Remark 11 *The sampling model* (1) *assumes independence within each source. Under certain models of dependence, our theoretical results remain valid, as discussed in Supplementary S.5.*

Based on Theorem 7, we generalize it to $m > 2$ sources and derive the following theorem.

Theorem 12 *Given* $0 < \epsilon < 1$*, let* $m = \lceil \log\epsilon / \log(1-p_0)\rceil$*. Under Assumptions 1, 2, 3, and 4, with probability at least* $1 - O\big(\frac{\log^2\epsilon}{\log^2(1-p_0)N^3}\big)$*, we have* $n \ge (1-\epsilon)N$ *and there exists* $\mathbf{O}_X \in \mathscr{O}^{r\times r}$ *such that*

- *if* $p_0 = o(1/\log N)$ *or* p_0 *is bounded away from* 0*, we have*

$$\|\widehat{\mathbf{X}}\mathbf{O}_X - \mathbf{X}^*\| \lesssim \Big\{1 + \frac{(1-p_0)K^2\log^2\epsilon}{\log^2(1-p_0)}\sqrt{\frac{p_0}{1-(1-p_0)^m}}\Big\}K\sqrt{\frac{N_0}{\lambda_{\min}}}\sigma; \tag{16}$$

- *otherwise,*

$$\|\widehat{\mathbf{X}}\mathbf{O}_X - \mathbf{X}^*\| \lesssim \Big\{1 + \frac{(1-p_0)K^2(p_0\log N)\log^2\epsilon}{\log^2(1-p_0)}\sqrt{\frac{p_0}{1-(1-p_0)^m}}\Big\}K\sqrt{\frac{N_0}{\lambda_{\min}}}\sigma. \tag{17}$$

Remark 13 *The above theorem gives us guidance on how many sources we need to recover enough parts of* $\mathbf{W}^*$*. The order of* m *can be* $|1/\log(1-p_0)| \approx 1/p_0$ *when* p_0 *is small. Besides, compared to* (12) *and* (13)*, the rates of* (16) *and* (17) *have only difference in the log terms, which means that even we choose* m *of the maximum order above, the rate of our error bounds will not change too much.*

Remark 14 *The multi-source embeddings have diverse applications, for example, the machine translation (Mikolov et al., 2013b; Xing et al., 2015; Shi et al., 2020) or code mapping (Hernandez et al., 2009; Zhou et al., 2012; Fidahussein and Vreeman, 2014). To be specific, for each pair of sources (s, k), we can match the entity $i \in \mathcal{V}_s \backslash \mathcal{V}_k$ to some entity $j \in \mathcal{V}_k$ such that*

$$j = \arg\max_{l \in \mathcal{V}_k} \cos(\widehat{\mathbf{X}}_i, \widehat{\mathbf{X}}_l) \text{ where } \cos(\widehat{\mathbf{X}}_i, \widehat{\mathbf{X}}_l) = (\widehat{\mathbf{X}}_i)^\top \widehat{\mathbf{X}}_l / \big(\|\widehat{\mathbf{X}}_i\| \|\widehat{\mathbf{X}}_l\|\big).$$

If $\cos(\widehat{\mathbf{X}}_i, \widehat{\mathbf{X}}_j)$ is larger than a threshold c, we can translate the entity i from the sth source (language) to the entity j in the kth source (language). We can determine c by either setting a desired sensitivity using test data or through cross-validation with translated pairs or a specificity that can be approximated by the distribution of cosine similarity of related but not synonymous pairs. Once we obtain the spectral error bound of $\|\widehat{\mathbf{X}}\mathbf{O}_X - \mathbf{X}^*\|$, we can utilize it to construct the bound of the translation accuracy. For example, we can bound $\|\widehat{\mathbf{X}}_i - \widehat{\mathbf{X}}_j\|$ when $\mathbf{X}_i^* = \mathbf{X}_j^*$ or $\mathbb{P}\big(\cos(\widehat{\mathbf{X}}_i, \widehat{\mathbf{X}}_j) \geq c\big)$ when $\cos(\mathbf{X}_i^*, \mathbf{X}_j^*) \leq c_0$ for some c_0. According to the translation procedure, the translation accuracy can depend on the bound $\sup_{l \in \mathcal{V}_k} |\cos(\widehat{\mathbf{X}}_i, \widehat{\mathbf{X}}_l) - \cos(\mathbf{X}_i^*, \mathbf{X}_l^*)|$ where $i \in \mathcal{V}_s \backslash \mathcal{V}_k$. To bound the quantity, we need additional assumptions on the structures of the underlying matrix $\mathbf{W}^*$. For instance, if the entities i and j are synonym or translated pairs in different languages, then $\cos(\mathbf{X}_j^*, \mathbf{X}_i^*) > c_1$ for some constant c_1, otherwise, $\cos(\mathbf{X}_j^*, \mathbf{X}_i^*) \leq c_0$ for some $c_0 < c_1$.

4. Simulation

In this section, we examine the performance of Algorithm 1 from extensive simulation studies for various values of p_0, m, and σ.

4.1 Comparable Methods

We compare with SMC (Cai et al., 2016) and a state-of-the-art matrix completion algorithm under the uniform sampling assumption, vanilla gradient descent (VGD) (Ma et al., 2018). Since SMC can only be applied to complete a single missing block, we use it to complete the missing blocks of each pair of sources. After all missing blocks are imputed, we use the rank-r SVD to obtain the low-rank estimator for SMC. For VGD, the input is the partially observed matrix $\widetilde{\mathbf{W}}$ created in Step I (b) of Algorithm 1, where the unobserved entries are treated as missing values. They will also produce an estimator for $\mathbf{X}^*$. Another standard approach is to use one data source as pre-training and the new data sources to continue training. This effectively corresponds to imputing the missing blocks of the PMI matrix as zero. We call the method 'Pre-trained'.

Besides, a potential application of BONMI is machine translation. To be specific, in reality, the overlapping parts may not be known fully. For instance, $\{\mathbf{W}^s\}_{s \in [m]}$ are multilingual co-occurrence matrices or PMI matrices (Levy and Goldberg, 2014), then each vertex is a word and the overlapping parts are created by bilingual dictionaries, which are limited in some low-resource languages and always cover only a small proportion of the corpora. In this case, BONMI can utilize these matrices and their known overlap to train multilingual word embeddings (i.e., $\widehat{\mathbf{X}}$). For the words not known in the overlapping set, if their embeddings (i.e., rows of $\widehat{\mathbf{X}}$) are close enough, it means that they have a similar

meaning and should be translated to each other. We evaluate the translation precision in the simulation setting (iii). As a baseline, we also compare BONMI to the popular orthogonal transformation method (Smith et al., 2017) which uses the single-source embeddings $\widehat{\mathbf{X}}_s = \widehat{\mathbf{U}}_s\widehat{\mathbf{\Sigma}}_s^{1/2}$ for $s \in [m]$. We denote the method as 'Orth'.

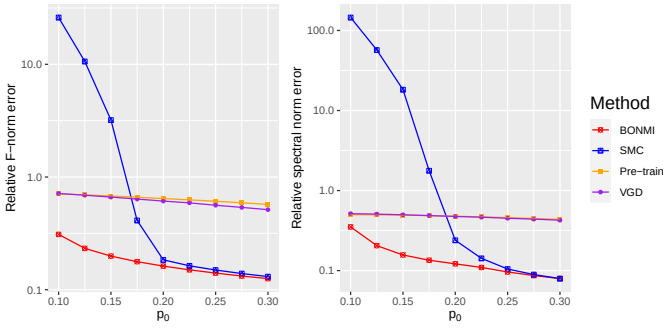
For all methods, we use the true rank r for simplicity in the following numeric experiments. We propose a data-driven method for choosing r for real-world problems in Section 5. To validate this proposed tuning strategy for r, we evaluate the performance of all algorithms using the estimated r in simulations. As summarized in Supplementary S.6, the relative performance of different algorithms shares a similar pattern as those given in Section 4.3 with BONMI outperforming other competing methods.

4.2 Data Generation Mechanisms and Evaluation Metrics

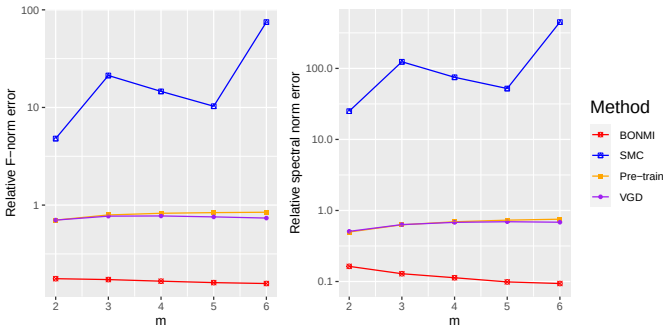
Throughout, we fix $N = 25,000$ and $r = 200$ which are comparable to our real data. We then generate the word embedding matrix $\mathbf{X} = \mathbf{U}^*(\mathbf{\Sigma}^*)^{\frac{1}{2}}$ and therefore $\mathbf{W}^* = \mathbf{U}^*\mathbf{\Sigma}^*(\mathbf{U}^*)^\top$, where $\mathbf{\Sigma}^*$ is a diagonal matrix whose diagonal elements are generated independently from the uniform distribution $\mathrm{U}(\sqrt{N}, 4\sqrt{N})$. The matrix $\mathbf{U}^*$ is drawn randomly from the Haar measure. Specifically, we generate a matrix $\mathbf{H} \in \mathbb{R}^{N\times r}$ with i.i.d. standard Gaussian entries, then apply the QR decomposition to $\mathbf{H}$ and assign $\mathbf{U}^*$ with the Q part of the result. To generate data for the sth source, we generate a sequence of independent Bernoulli random variables with success rate p_0: $\delta^s = (\delta_1^s, \ldots, \delta_N^s)$ to form the index set $\mathcal{V}_s = \{i : \delta_i^s = 1, i \in [N]\}$, for $s \in [m]$. We then generate the error matrix $\mathbf{E}^s \in \mathbb{R}^{|\mathcal{V}_s|\times|\mathcal{V}_s|}$ with its upper triangular block including the diagonal elements from the normal distribution $N(0, \sigma_s^2)$ and lower triangular block decided by symmetry, where we let the noise level σ_s vary across the m sources.

We consider three settings with the first two focusing on the task of matrix completion and setting (iii) focusing on the downstream task of machine translation. For the matrix completion task, we consider two settings: (i) $m = 2$, $\sigma_s = 0.1s$, and let p_0 vary from 0.1 to 0.3; (ii) $p_0 = 0.1$, $\sigma_s = 0.1$, and let m vary from 2 to 6. For the machine translation task, we consider the setting (iii) where we let $m = 2, 3$, $p_0 = 0.1$, $\sigma_s = s\sigma$ and let a noise level σ vary from 0.3 to 0.5.

To evaluate the performance of matrix completion, we use the relative F-norm and spectral norm errors of the estimation of $\mathbf{W}_0^*$ defined as $\mathrm{err}_{\mathrm{F}}(\widehat{\mathbf{W}}, \mathbf{W}_0^*) = \frac{\|\widehat{\mathbf{W}}-\mathbf{W}_0^*\|_{\mathrm{F}}}{\|\mathbf{W}_0^*\|_{\mathrm{F}}}$ and $\mathrm{err}_2(\widehat{\mathbf{W}}, \mathbf{W}_0^*) = \frac{\|\widehat{\mathbf{W}}-\mathbf{W}_0^*\|}{\|\mathbf{W}_0^*\|}$. To evaluate the overall performance of machine translation in setting (iii), we additionally generate test data for evaluation. Specifically, we additionally sample $n_{\mathrm{test}} = 2000$ vertices from $\mathcal{V}\backslash\mathcal{V}^*$ where $\mathcal{V}^* = \cup_{s=1}^m \mathcal{V}_s$, denoted as $\mathcal{V}^{\mathrm{test}}$, and combine $\mathcal{V}^{\mathrm{test}}$ and $\mathcal{V}_s$ to get $\mathcal{V}_s' = \mathcal{V}^{\mathrm{test}} \cup \mathcal{V}_s$ as the final vertex set of the sth source. We then use $\mathcal{V}_s'$ to generate $\mathbf{W}^s$. Notice now $\mathbf{E}^s \in \mathbb{R}^{|\mathcal{V}_s'|\times|\mathcal{V}_s'|}$. However, we treat elements of $\mathcal{V}^{\mathrm{test}}$ as unique across the m sources, which means that we will not combine elements of $\mathcal{V}^{\mathrm{test}}$ in Algorithm 1. The role of $\mathcal{V}^{\mathrm{test}}$ is exactly the testing set in machine translation. We average the $m-1$ translation precision from the sth source to the 1st source, $s = 2, \ldots, m$. The translation precision is defined as follows: for $i \in \mathcal{V}^{\mathrm{test}}$, we can get its embedding in the sth source corresponding to one row in $\widehat{\mathbf{X}}$, denoted as $\widehat{\mathbf{X}}_i$. Then we find its closest vector $\widehat{\mathbf{X}}_j$ for $j \in \mathcal{V}_1'$ with the largest cosine similarity as illustrated in Remark 14. If the jth element from the



(a) setting (i): fix $m = 2$ and range p_0 from 0.1 to 0.3.



(b) setting (ii): fix $p_0 = 0.1$ and range m from 2 to 6.

Figure 1: Simulation results of settings (i) and (ii). The relative estimation errors of $\mathbf{W}_0^*$ are presented.

1st source and the ith element from the sth source are the same element in $\mathcal{V}^*$, we treat it as a correct translation. The precision of the sth source is the ratio of correct translations among the test set $\mathcal{V}^{\text{test}}$ in the sth source.

4.3 Results

We summarize simulation results averaged over 50 replications for settings (i)-(ii) in Figure 1 and setting (iii) in Figure 2. BONMI outperforms all competing methods across the three settings. In settings (i) and (ii), the results of the F-norm and spectral norm errors are consistent. In setting (i), we can see that the relative errors of all methods decrease when the observation rate p_0 increases as expected. The advantage of BONMI in the accuracy of matrix completion is more pronounced when the observation rate p_0 is low. When p_0 is very small, SMC tends to fail. In setting (ii), the error of BONMI decreases as m increases, which is due to the information gained from multiple sources. However, both the naive pre-training method and VGD do not always perform better as m increases. Overall, BONMI dominates all competing methods across different choices of m.

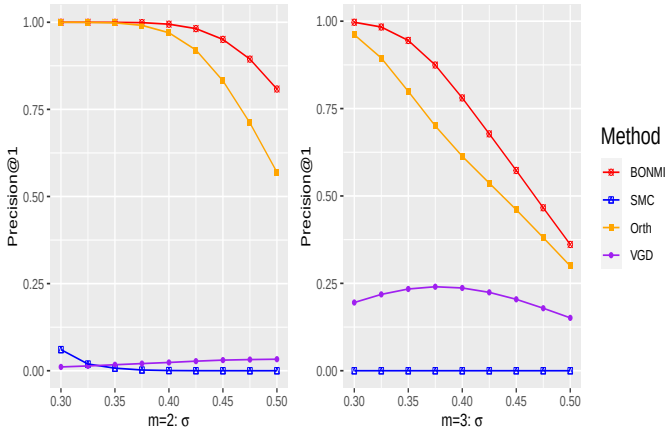


Figure 2: setting (iii): fix $p_0 = 0.1$ and range σ from 0.3 to 0.5.

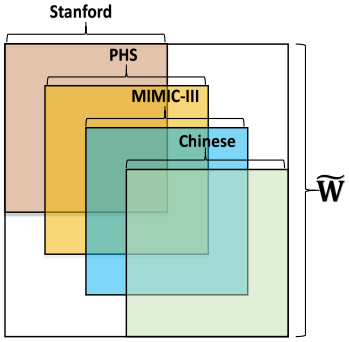


Figure 3: The aggregation of the four PPMI matrices. The four sources all have an overlapping with each other source.

5. Real Data Analysis

In this section, we apply BONMI to obtain clinical concept embeddings using multiple PPMI matrices in two different languages, English and Chinese. The clinical concepts in English have been mapped to *Concept Unique Identifiers* (CUIs) in the Unified Medical Language System (UMLS) (Humphreys and Lindberg, 1993). Our goal is to enable the integration of multiple PPMI matrices to co-train clinical concept embeddings for both CUIs and Chinese clinical terms.

The input data ensemble consists of three CUI PPMI matrices and one Chinese PPMI matrix. The three CUI PPMI matrices are independently derived from three data sources (i) 20 million clinical notes at Stanford (Finlayson et al., 2014); (ii) 10 million notes of 62K patients at Partners Healthcare System (PHS) (Beam et al., 2019); and (iii) health records from MIMIC-III, a freely accessible critical care database (Johnson et al., 2016). The multi-source raw data consist of the text data (i) and the EHR data (ii) and (ii), which are prepossessed in the same way as Beam et al. (2019) and Hong et al. (2021) to generate

the CUI-level co-occurrence matrices. The PPMI matrices are then obtained following Section 2.2. We choose sub-matrices from these sources by thresholding the frequency of these CUI and keeping those with semantic types related to medical concepts. Finally, we obtain the Stanford, PHS, and MIMIC PPMI matrices with 8922, 10964, and 8524 CUIs respectively. The mean overlapping CUIs of any two sources is 4480 and the total number of the unique CUIs of the three sources is 17963.

Multiple Chinese medical text data sources, such as medical textbooks and Wikipedia, are also collected. We then build a PPMI matrix of 8628 Chinese medical terms. A Chinese-English medical dictionary is used to translate these Chinese medical terms to English, which are further mapped to CUI. Finally, we obtain 4201 Chinese-CUI pairs, and we use 2000 pairs as the training set (the known overlapping set) and the other 2201 pairs as the test set to evaluate the translation precision. An illustration of the aggregation of the three CUI PPMI matrices and one Chinese PPMI matrix is presented in Figure 3.

From each method, we obtain the embedding vectors for all entities by performing an SVD on the imputed PPMI matrix obtained from each method, respectively. To evaluate the quality of the obtained embedding, we compare the cosine similarity of trained embeddings against the gold standard human annotations of the concept similarity and relatedness. We considered two sets of human annotated relatedness and similarity: (I) 200 pairs of Chinese medical terms randomly selected and annotated by four clinical experts; and (II) 566 pairs of UMLS concepts in English previously annotated by eight researchers in Pakhomov et al. (2010). The Chinese medical terms were also translated into English and mapped to the UMLS CUI while the 566 UMLS concepts in English were also translated into Chinese. Each concept pair thus can be viewed as CUI-CUI pair (CUI), Chinese-Chinese pair (Chinese), and Chinese-CUI pair (Cross). The gold standard human annotation assigns each concept pair a relatedness and similarity score, defined as the average score from all reviewers. For each concept pair, we compare the cosine similarity of their associated embeddings against the human annotations of their similarity and relatedness. We evaluate the quality of the embeddings based on (a) the rank correlation between the cosine similarity and human annotation; and (b) the accuracy in translating Chinese medical terms to CUIs in English. The Precision@k is defined similarly to the translation accuracy in Section 4. The difference is that when the truth CUI is among the CUIs with the top k largest cosine similarity to the Chinese term, then it is treated as a correct translation. Precision@k is the ratio of the correct translations given a k. Here we choose k as 5, 10 and 20.

To choose the rank of the matrix, we analyze the eigen decay of the matrices. The eigen decay has been widely used to determine the rank of low-rank matrices, for example, in principal component analysis (Jolliffe, 2005), word embedding (Hong et al., 2021) and network analysis (Arroyo et al., 2021). We calculate the eigen decay of the overlapping sub-matrices of each pair of sources and choose the rank r that makes the cumulative eigenvalue percentage of at least one of the matrices more than 95%, which is 300. We then use $r = 300$ for all methods.

Finally, to compare with the neural-based embeddings, we also include a BERT-based algorithm CODER (mediCal knOwledge embeDded tErm Representation) proposed by Yuan et al. (2022). The CODER algorithm is trained on top of BERT (Devlin et al., 2019) with contrastive learning on the multi-lingual relation pairs from UMLS (Bodenreider, 2004) to improve medical term embedding. We use the pre-trained model of Yuan et al. (2022),

whose input is the code descriptions, where we use the preferred English terms for CUIs and Chinese terms, and the output of CODER is the CUI and Chinese embeddings of dimension 768.

Besides, the machine translation accuracy can suffer from the limited size of the Chinese corpus. On the other hand, CODER embeddings utilize the multi-lingual semantic information from code descriptions, which serves as complementary information to the PMI-based embeddings. As a result, we use the CODER embeddings to assist the machine translation task by concatenating their embeddings with the EHR-based embeddings (denoted by 'method+', e.g., 'BONMI+').

5.1 Results

We present the results of BONMI and other competing methods in Table 1. We can observe that all methods other than SMC perform similarly when assessing the relatedness and similarity of Chinese-Chinese and CUI-CUI since these pairs belong to the same corpus. However, BONMI outperforms other methods when evaluating relatedness and similarity between Chinese-CUI pairs that belong to the missing blocks. BONMI also attains higher accuracy in the translation task. This suggests that BONMI has the advantage over existing methods in providing embedding vectors that enable an accurate assessment of relatedness between entity pairs that do not belong to the same corpus.

6. Discussion and Conclusion

6.1 Generalization

We consider the completion of the PMI matrices and the estimation of multi-source embeddings in this paper. However, we also notice that there exist many applications involving the completion of missing blocks for asymmetric matrices, for example, genomic data integration (Cai et al., 2016), multimodality data analysis (Xue and Qu, 2020), and other applications mentioned in the introduction. Hence, in this section, we introduce an algorithm designed for asymmetric matrices without repeated observations. Now assume that $\mathbf{W}^* \in \mathbb{R}^{N\times D}$ of rank r is asymmetric with $\lambda_{\max} = \sigma_1(\mathbf{W}^*) > \lambda_{\min} = \sigma_r(\mathbf{W}^*) > 0$. Here $\mathbf{W}^*$ does not have to come from the inner product of word embeddings but can be any low-rank matrix. Furthermore, we have a noisy-corrupted matrix $\mathbf{W}$ such that $\mathbf{W} = \mathbf{W}^* + \mathbf{E}$. For the sth source, we sample two index sets $\mathcal{V}_{s1} \subseteq [N]$ and $\mathcal{V}_{s2} \subseteq [D]$ independently such that for each $i \in [N]$ and $j \in [D]$, we assign i to $\mathcal{V}_{s1}$ with probability p_{s1} and j to $\mathcal{V}_{s2}$ with probability p_{s2} independently:

$$\mathbb{P}(i \in \mathcal{V}_{s1}) = p_{s1} \in (0,1), \text{ for } i \in [N], \quad \mathbb{P}(j \in \mathcal{V}_{s2}) = p_{s2} \in (0,1), \text{ for } j \in [D], s \in [m]\,.$$

With the index sets $\mathcal{V}_{s1}$ and $\mathcal{V}_{s2}$, a matrix $\mathbf{W}^s$ is observed

$$\mathbf{W}^s = \mathbf{W}_{\mathcal{V}_{s1},\mathcal{V}_{s2}}, \text{ for } s \in [m]\,. \tag{18}$$

Let $\mathcal{V}_1^* = \cup_{s=1}^m \mathcal{V}_{s1}$ and $\mathcal{V}_2^* = \cup_{s=1}^m \mathcal{V}_{s2}$, then our task is to recover

$$\mathbf{W}_0^* \equiv \mathbf{W}^*_{\mathcal{V}_1^*,\mathcal{V}_2^*} = \left[\mathbf{W}^*(i,j)\right]_{j\in\mathcal{V}_2^*}^{i\in\mathcal{V}_1^*} \in \mathbb{R}^{n_1\times n_2}, \quad \text{where } N_0 = |\mathcal{V}_1^*| \text{ and } D_0 = |\mathcal{V}_2^*|.$$

Table 1: Results of the integration of four PPMI matrices: (a) rank correlation between the pairwise cosine similarity of estimated embedding vectors from the completed matrix and the similarity or relatedness from human annotation; (b) accuracy in translation based on the estimated embedding vectors.

(a) Rank correlation with human annotations

Source	Type	Set	BONMI	Pre-train	SMC	VGD	CODER
Chinese	Rel	I	0.741	0.756	0.066	0.761	0.519
	Rel	II	0.661	0.659	0.327	0.663	0.482
	Sim	I	0.707	0.724	0.105	0.731	0.715
	Sim	II	0.716	0.728	0.271	0.726	0.469
CUI	Rel	I	0.678	0.639	0.369	0.643	0.398
	Rel	II	0.604	0.598	0.141	0.592	0.351
	Sim	I	0.615	0.601	0.243	0.582	0.741
	Sim	II	0.634	0.635	0.171	0.622	0.451
Cross	Rel	I	0.671	0.408	0.321	0.418	0.502
	Rel	II	0.655	0.424	0.301	0.358	0.424
	Sim	I	0.607	0.322	0.369	0.339	0.724
	Sim	II	0.699	0.445	0.335	0.399	0.428

(b) Translation accuracy

		BONMI+	Pre-train+	SMC+	VGD+	CODER
Precision	@5	0.708	0.617	0.569	0.614	0.553
	@10	0.766	0.683	0.633	0.677	0.621
	@20	0.812	0.728	0.691	0.719	0.683

Without loss of generality, we assume $\mathcal{V}_1^* = [N_0]$ and $\mathcal{V}_2^* = [D_0]$. The estimation procedure is summarized in Algorithm 2.

The model (18) is more general than (2) in two ways: (i) (18) considers the asymmetric matrix and (ii) (18) does not assume repeated observations. To be specific, the overlapping parts of $\mathbf{W}^s$ are the same for different sources, which means that they are only observed once, and Step I Aggregation in Algorithm 1 is not applicable now. The two relaxations make the model (18) more flexible and realistic for the applications mentioned above. However, if one does have repeated observations from each source with the heterogeneous noise level, a weighted aggregation procedure similar to Step I (b) of Algorithm 1 can be applied and will not affect the theoretical guarantee of the estimator.

Assume that $\mathbf{W}_0^*$ has SVD $\mathbf{W}_0^* = \mathbf{U}_0^*\mathbf{\Sigma}_0^*(\mathbf{V}_0^*)^\top = \mathbf{X}^*(\mathbf{Y}^*)^\top$, where $\mathbf{U}_0^* \in \mathbb{R}^{N_0\times r}$ are the left-singular vectors, $\mathbf{V}_0^* \in \mathbb{R}^{D_0\times r}$ are the right-singular vectors, and $\mathbf{\Sigma}_0^*$ is an $r\times r$ diagonal matrix with singular values in a descending order. In addition, $\mathbf{X}^* = \mathbf{U}_0^*(\mathbf{\Sigma}_0^*)^{1/2} \in \mathbb{R}^{N_0\times r}$ and $\mathbf{Y}^* = \mathbf{V}_0^*(\mathbf{\Sigma}_0^*)^{1/2} \in \mathbb{R}^{D_0\times r}$. Let $\widehat{\mathbf{X}} = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}^{1/2}$ and $\widehat{\mathbf{Y}} = \widehat{\mathbf{V}}\widehat{\mathbf{\Sigma}}^{1/2}$ be the output of Algorithm 2. Without loss of generality, assume $N = \max\{N, D\}$ and $p_0 = \min_{s\in[m]}\{p_{s1}, p_{s2}\}$. Similar results to Theorem 7 and Theorem 12 can be provided. For example, when $m = 2$, if $p_0 = o(1/\log N)$ or p_0 is bounded away from zero, we can prove that under similar

Algorithm 2: BONMI for asymmetric matrices.

Input: m matrices $\{\mathbf{W}^s\}_{s\in[m]}$ and the corresponding index sets $\left\{\mathcal{V}_{s1}, \mathcal{V}_{s2}\right\}_{s\in[m]}$; the rank r; $N_0 = |\cup_{s=1}^m \mathcal{V}_{s1}|$ and $D_0 = |\cup_{s=1}^m \mathcal{V}_{s2}|$;

Step I Aggregation: Create $\widetilde{\mathbf{W}} \in \mathbb{R}^{N_0\times D_0}$ as follows:

$$\widetilde{\mathbf{W}}(i,j) = \mathbf{W}^s(v_i^{s1}, v_j^{s2}) \text{ if } i \in \mathcal{V}_{s1} \text{ and } j \in \mathcal{V}_{s2} \text{ for some } s \in [m]$$

for all pairs of (i,j) such that $\mathcal{S}_{ij} \equiv \sum_{s=1}^m \mathbb{1}(i \in \mathcal{V}_{s1}, j \in \mathcal{V}_{s2}) > 0$, where $v_i^{s1}(v_j^{s2})$ denotes the row(column) index in $\mathbf{W}^s$ corresponding to the ith row(jth column) of $\mathbf{W}_0^*$, and the entries in the missing blocks with $\mathcal{S}_{ij} = 0$ are initialized as zero.

Step II (a) Spectral initialization: for $1 \le s \le m$ **do**

Let $\widetilde{\mathbf{U}}_s\widetilde{\mathbf{\Sigma}}_s\widetilde{\mathbf{V}}_s^\top$ be the rank-r SVD of $\mathbf{W}^s$.

end

Step II (b) Estimation of missing parts: for $1 \le s < k \le m$ **do**

Obtain $\widetilde{\mathbf{W}}_{sk}$ and $\widetilde{\mathbf{W}}_{ks}$ using $\widetilde{\mathbf{U}}_s, \widetilde{\mathbf{\Sigma}}_s, \widetilde{\mathbf{V}}_s, \widetilde{\mathbf{U}}_k, \widetilde{\mathbf{\Sigma}}_k, \widetilde{\mathbf{V}}_k$:

$$\widetilde{\mathbf{W}}_{sk} \equiv \widetilde{\mathbf{U}}_{s1}\widetilde{\mathbf{\Sigma}}_s^{1/2}\mathbf{G}(\widetilde{\mathbf{\Sigma}}_s^{1/2}\widetilde{\mathbf{U}}_{s2}^\top\widetilde{\mathbf{U}}_{k1}\widetilde{\mathbf{\Sigma}}_k^{1/2})\widetilde{\mathbf{\Sigma}}_k^{1/2}\widetilde{\mathbf{V}}_{k2}^\top,$$

$$\widetilde{\mathbf{W}}_{ks} \equiv \widetilde{\mathbf{U}}_{k2}\widetilde{\mathbf{\Sigma}}_k^{1/2}\mathbf{G}(\widetilde{\mathbf{\Sigma}}_k^{1/2}\widetilde{\mathbf{V}}_{k1}^\top\widetilde{\mathbf{V}}_{s2}\widetilde{\mathbf{\Sigma}}_s^{1/2})\widetilde{\mathbf{\Sigma}}_s^{1/2}\widetilde{\mathbf{V}}_{s1}^\top.$$

Here $\widetilde{\mathbf{U}}_s = (\widetilde{\mathbf{U}}_{s1}^\top, \widetilde{\mathbf{U}}_{s2}^\top)^\top$, $\widetilde{\mathbf{U}}_s = (\widetilde{\mathbf{V}}_{s1}^\top, \widetilde{\mathbf{V}}_{s2}^\top)^\top$, $\widetilde{\mathbf{U}}_k = (\widetilde{\mathbf{U}}_{k1}^\top, \widetilde{\mathbf{U}}_{k2}^\top)^\top$, and $\widetilde{\mathbf{V}}_k = (\widetilde{\mathbf{V}}_{k1}^\top, \widetilde{\mathbf{U}}_{k2}^\top)^\top$ are decomposed similarly to (8). $\widetilde{\mathbf{W}}_{sk}$ and $\widetilde{\mathbf{W}}_{sk}$ are used like in (9). If a missing entry (i,j) is estimated by multiple pairs of sources (s,k), choose the one estimated by the first pair. Denote the imputed matrix as $\widehat{\mathbf{W}}$.

end

Step III Low rank approximation: Obtain the rank-r SVD of $\widehat{\mathbf{W}}$: $\widehat{\mathbf{W}}_r = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^\top$.

Output: $\widehat{\mathbf{U}}$, $\widehat{\mathbf{\Sigma}}$, $\widehat{\mathbf{V}}$.

assumptions,

$$\|\widehat{\mathbf{X}}\mathbf{O}_X - \mathbf{X}^*\| \lesssim \frac{\{(1-p_0)K^2+1\}K}{\sqrt{\lambda_{\min}}}\sqrt{N}\sigma, \quad \|\widehat{\mathbf{Y}}\mathbf{O}_Y - \mathbf{Y}^*\| \lesssim \frac{\{(1-p_0)K^2+1\}K}{\sqrt{\lambda_{\min}}}\sqrt{N}\sigma,$$

and otherwise,

$$\|\widehat{\mathbf{X}}\mathbf{O}_X - \mathbf{X}^*\| \lesssim \frac{\{(1-p_0)K^2(p_0\log N)+1\}K}{\sqrt{\lambda_{\min}}}\sqrt{N}\sigma,$$

$$\|\widehat{\mathbf{Y}}\mathbf{O}_Y - \mathbf{Y}^*\| \lesssim \frac{\{(1-p_0)K^2(p_0\log N)+1\}K}{\sqrt{\lambda_{\min}}}\sqrt{N}\sigma,$$

for some rotation matrices $\mathbf{O}_X, \mathbf{O}_Y \in \mathscr{O}^{r\times r}$. The proofs are the simple extensions of the proof of Theorem 7 and Theorem 12.

6.2 Conclusion

This paper proposes BONMI, which aims at multi-source learning. Our method is computationally efficient with a theoretical guarantee. The performance of our algorithm is verified by simulation and real data analysis. For theoretical guarantee, we require the sampling probability of each source to have the order of $\sqrt{\log N/N}$, which is a small order of N and can be satisfied easily in many applications. Besides, we extend BONMI to asymmetric matrices without repeated observations for other potential applications such as genomic data integration.

References

Swapna Abhyankar, Dina Demner-Fushman, and Clement J McDonald. Standardizing clinical laboratory data for secondary use. *Journal of biomedical informatics*, 45(4):642–650, 2012.

Yuri Ahuja, Doudou Zhou, Zeling He, Jiehuan Sun, Victor M Castro, Vivian Gainer, Shawn N Murphy, Chuan Hong, and Tianxi Cai. surelda: A multidisease automated phenotyping method for the electronic health record. *Journal of the American Medical Informatics Association*, 27(8):1235–1243, 2020.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, 2018.

Jesús Arroyo, Avanti Athreya, Joshua Cape, Guodong Chen, Carey E Priebe, and Joshua T Vogelstein. Inference for multiple heterogeneous networks with a common invariant subspace. *Journal of Machine Learning Research*, 22(142):1–49, 2021.

Christos Baloukas, Lazaros Papadopoulos, Dimitrios Soudris, Sander Stuijk, Olivera Jovanovic, Florian Schmoll, Daniel Cordes, Robert Pyka, Arindam Mallik, Stylianos Mamagkakis, et al. Mapping embedded applications on mpsocs: the mnemee approach. In *2010 IEEE Computer Society Annual Symposium on VLSI*, pages 512–517. IEEE, 2010.

Afonso S. Bandeira and Ramon van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability*, 44(4):2479 – 2506, 2016.

David M Baorto, James J Cimino, Curtis A Parvin, and Michael G Kahn. Combining laboratory data sets from multiple institutions using the logical observation identifier names and codes (LOINC). *International Journal of Medical Informatics*, 51(1):29–37, 1998.

Andrew L. Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin Weber, Nathan Palmer, Xu Shi, Tianxi Cai, and Isaac S. Kohane. Clinical concept embeddings learned

from massive sources of multimodal medical data. In *Biocomputing 2020*. WORLD SCIENTIFIC, nov 2019.

Willie Boag and Hassan Kané. Awe-cm vectors: Augmenting word embeddings with a clinical metathesaurus. *arXiv preprint arXiv:1712.01460*, 2017.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

Tianxi Cai, T Tony Cai, and Anru Zhang. Structured matrix completion with applications to genomic data integration. *Journal of the American Statistical Association*, 111(514): 621–633, 2016.

Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717, 2009.

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

Yudong Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, 2015.

Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. arXiv preprint arXiv:1509.03025, 2015.

Patrick L Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*. Springer, 2011.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *Proceedings of ICLR*, 2018.

Jessica K De Freitas, Kipp W Johnson, Eddye Golden, Girish N Nadkarni, Joel T Dudley, Erwin P Bottinger, Benjamin S Glicksberg, and Riccardo Miotto. Phe2vec: Automated disease phenotyping based on unsupervised embeddings from electronic health records. *Patterns*, 2(9):100337, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Maryam Fazel. *Matrix rank minimization with applications*. PhD thesis, Stanford University, 2002.

Mustafa Fidahussein and Daniel J Vreeman. A corpus-based approach for automated loinc mapping. *Journal of the American Medical Informatics Association*, 21(1):64–72, 2014.

Samuel G Finlayson, Paea LePendu, and Nigam H Shah. Building the graph of medicine from millions of clinical narratives. *Scientific data*, 1(1):1–9, 2014.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*. PMLR, 2017.

John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

John C Gower and Garmt B Dijksterhuis. *Procrustes problems*, volume 30. Oxford University Press on Demand, 2004.

Justin P Haldar and Diego Hernando. Rank-constrained solutions to linear matrix equations using powerfactorization. *IEEE Signal Processing Letters*, 16(7):584–587, 2009.

Penni Hernandez, Tanya Podchiyska, Susan Weber, Todd Ferris, and Henry Lowe. Automated mapping of pharmacy orders from two electronic health record systems to rxnorm within the stride clinical data warehouse. In *AMIA Annual Symposium Proceedings*, volume 2009, page 244. American Medical Informatics Association, 2009.

Joshua A Hirsch, Thabele M Leslie-Mazwi, Gregory N Nicola, Robert M Barr, Jacqueline A Bello, William D Donovan, Raymond Tu, Mark D Alson, and Laxmaiah Manchikanti. Current procedural terminology; a primer. *Journal of Neurointerventional Surgery*, 7(4): 309–312, 2015.

Chuan Hong, Everett Rush, Molei Liu, Doudou Zhou, Jiehuan Sun, Aaron Sonabend, Victor M Castro, Petra Schubert, Vidul A Panickan, Tianrun Cai, et al. Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data. *NPJ digital medicine*, 4(1):1–11, 2021.

Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*, 2008. doi: 10.1109/ICDM.2008.22.

Betsy L Humphreys and DA Lindberg. The UMLS project: making the conceptual connection between users and the information they need. *Bulletin of the Medical Library Association*, 81(2):170, 1993.

Prateek Jain, Raghu Meka, and Inderjit Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9, 2016.

Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2005.

Yova Kementchedjhieva, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard. Generalizing Procrustes analysis for better bilingual dictionary induction. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 211–220, Brussels, Belgium, October 2018. Association for Computational Linguistics.

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

Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302 – 2329, 2011.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

Naoto Kume, Kenji Suzuki, Shinji Kobayashi, Hiroyuki Yoshihara, and Kenji Araki. Original laboratory test code mapping system using test result data on electronic health record. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, pages 1518–1519. IOS Press, 2019.

Kiryung Lee and Yoram Bresler. Admira: Atomic decomposition for minimum rank approximation. *IEEE Transactions on Information Theory*, 56(9):4402–4416, 2010.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Chin Lin, Yu-Sheng Lou, Dung-Jang Tsai, Chia-Cheng Lee, Chia-Jung Hsu, Ding-Chung Wu, Mei-Chuen Wang, and Wen-Hui Fang. Projection word embedding model with hybrid sampling training for classifying icd-10-cm codes: Longitudinal observational study. *JMIR medical informatics*, 7(3):e14499, 2019.

Xiaokang Liu, Jessica Chubak, Rebecca A Hubbard, and Yong Chen. Sat: a surrogate-assisted two-wave case boosting sampling method, with application to ehr-based association studies. *Journal of the American Medical Informatics Association*, 2021.

Karim Lounici. Optimal spectral norm rates for noisy low-rank matrix completion. *arXiv preprint arXiv:1110.5346*, 2011.

Junwei Lu, Jin Yin, and Tianxi Cai. Knowledge graph embedding with electronic health records data via latent graphical block model. *arXiv preprint arXiv:2305.19997*, 2023.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*. PMLR, 2018.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11: 2287–2322, 2010.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Quoc V Le, and Ilya+ Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013b.

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13:1665–1697, 2012.

Luong Trung Nguyen, Junhan Kim, and Byonghyo Shim. Low-rank matrix completion: A contemporary survey. *IEEE Access*, 7:94215–94237, 2019.

Serguei Pakhomov, Bridget McInnes, Terrence Adam, Ying Liu, Ted Pedersen, and Genevieve B Melton. Semantic similarity and relatedness between clinical terms: an experimental study. In *AMIA annual symposium proceedings*, volume 2010. American Medical Informatics Association, 2010.

Sonali Parbhoo, Jasmina Bogojeska, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. Combining kernel and model based learning for hiv therapy selection. *AMIA Summits on Translational Science Proceedings*, 2017:239, 2017.

Arkadiusz Paterek. Improving regularized singular value decomposition for collaborative filtering. In *Proceedings of KDD cup and workshop*, volume 2007, 2007.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Machine Learning for Healthcare Conference*, pages 147–163. PMLR, 2017.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):18, 2018.

Jasson DM Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, 2005.

David Ruppert and Matthew P Wand. Multivariate locally weighted least squares regression. *The Annals of Statistics*, 22:1346–1370, 1994.

Peter H Schönemann. Varisim: A new machine method for orthogonal rotation. *Psychometrika*, 31(2):235–248, 1966.

Xu Shi, Xiaoou Li, and Tianxi Cai. Spherical regression under mismatch corruption with application to automated knowledge translation. *Journal of the American Statistical Association*, pages 1–12, 2020.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. arXiv preprint arXiv:1702.03859, 2017.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Aaron Sonabend, Nilanjana Laha, Ashwin N Ananthakrishnan, Tianxi Cai, and Rajarshi Mukherjee. Semi-supervised off policy reinforcement learning. *arXiv preprint arXiv:2012.04809*, 2020.

Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

Gábor Takács, István Pilászy, Bottyán Németh, and Domonkos Tikk. Major components of the gravity recommendation system. *Acm Sigkdd Explorations Newsletter*, 9(2):80–83, 2007.

Jared Tanner and Ke Wei. Normalized iterative hard thresholding for matrix completion. *SIAM Journal on Scientific Computing*, 35(5):S104–S125, 2013.

Jared Tanner and Ke Wei. Low rank matrix completion by alternating steepest descent methods. *Applied and Computational Harmonic Analysis*, 40(2):417–429, 2016.

Georgios Tournavitis, Zheng Wang, Björn Franke, and Michael FP O'Boyle. Towards a holistic approach to auto-parallelization: integrating profile-driven parallelism detection and machine-learning based mapping. *ACM Sigplan notices*, 44(6):177–187, 2009.

Joel A Tropp. Norms of random submatrices and sparse approximation. *Comptes Rendus Mathematique*, 346(23-24):1271–1274, 2008.

Bart Vandereycken. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

John Weeks and Roy Pardee. Learning to share health care data: a brief timeline of influential common data models and distributed health data networks in us health care research. *EGEMs*, 7(1), 2019.

Zaiwen Wen, Wotao Yin, and Yin Zhang. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation*, 4(4):333–361, 2012.

Shuo Xiang, Lei Yuan, Wei Fan, Yalin Wang, Paul M Thompson, Jieping Ye, and the Alzheimer's Disease Neuroimaging Initiative. Bi-level multi-source learning for heterogeneous block-wise missing data. *NeuroImage*, 102:192–206, 2014.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

Fei Xue and Annie Qu. Integrating multisource block-wise missing data in model selection. *Journal of the American Statistical Association*, 0(0):1–14, 2020.

Liang Yao, Chengsheng Mao, and Yuan Luo. KG-BERT: BERT for knowledge graph completion. *arXiv preprint arXiv:1909.03193*, 2019.

Guan Yu, Quefeng Li, Dinggang Shen, and Yufeng Liu. Optimal sparse linear prediction for block-missing multi-modality data without imputation. *Journal of the American Statistical Association*, 115(531):1406–1419, 2020.

Lei Yuan, Yalin Wang, Paul M Thompson, Vaibhav A Narayan, Jieping Ye, and the Alzheimer's Disease Neuroimaging Initiative. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *NeuroImage*, 61(3):622–632, 2012.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of Biomedical Informatics*, page 103983, 2022.

Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. arXiv preprint arXiv:1605.07051, 2016.

Doudou Zhou, Ziming Gan, Xu Shi, Alina Patwari, Everett Rush, Clara-Lea Bonzel, Vidul A Panickan, Chuan Hong, Yuk-Lam Ho, Tianrun Cai, et al. Multiview incomplete knowledge graph integration with application to cross-institutional ehr data harmonization. *Journal of Biomedical Informatics*, 133:104147, 2022a.

Doudou Zhou, Yufeng Zhang, Aaron Sonabend-W, Zhaoran Wang, Junwei Lu, and Tianxi Cai. Federated offline reinforcement learning. *arXiv preprint arXiv:2206.05581*, 2022b.

Li Zhou, Joseph M Plasek, Lisa M Mahoney, Frank Y Chang, Dana DiMaggio, and Roberto A Rocha. Mapping partners master drug dictionary to rxnorm using an nlp-based approach. *Journal of Biomedical Informatics*, 45(4):626–633, 2012.

Huichen Zhu, Gen Li, and Eric F Lock. Generalized integrative principal component analysis for multi-type data with block-wise missing structure. *Biostatistics*, 21(2):302–318, 09 2018.

Technical supplementary material to

Multi-source Learning via Completion of Block-wise Overlapping Noisy Matrices

This document contains the supplementary material to the paper "Multi-source Learning via Completion of Block-wise Overlapping Noisy Matrices". We mainly provide technical details of proving Proposition 2, Theorems 7, and 12 here.

S.1. Proof of Proposition 2

Proof First, we prove $\text{rank}(\mathbf{W}^*_{s\cap k,s\cap k}) = r$. Let $N_{sk} \equiv |\mathcal{V}_s \cap \mathcal{V}_k|$, then by Lemma S.1, we have

$$N_{sk} \geq \frac{3p_s p_k N}{2} \geq C\mu_0 r \log N$$

holds with probability $1-O(1/N^3)$ for some sufficiently large constant C, where the second inequality comes from Assumption 2. Then, by Lemma S.2, we have

$$\lambda_r(\mathbf{W}^*_{s\cap k,s\cap k}) \geq \sigma_{\min}(\mathbf{U}^*_{s\cap k})\lambda_r(\mathbf{\Sigma}^*)\sigma_{\min}(\mathbf{U}^*_{s\cap k}) \geq \frac{n_{sk}\lambda_{\min}}{2N} > 0$$

with probability $1-1/N^3$. As a result, $\text{rank}(\mathbf{W}^*_{s\cap k,s\cap k}) = r$ with probability as least $1-O(1/N^3)$. Under this event, since $\mathbf{W}^*_{s\cap k,s\cap k}$ is a principal sub-matrix of $\mathbf{W}^*_s$, we have $\text{rank}(\mathbf{W}^*_s) \geq \text{rank}(\mathbf{W}^*_{s\cap k,s\cap k}) = r$. Besides, $\mathbf{W}^*_s$ is a principal sub-matrix of $\mathbf{W}^*$, we have $\text{rank}(\mathbf{W}^*_s) \leq \text{rank}(\mathbf{W}^*) = r$. Combining the two inequalities, we have $\text{rank}(\mathbf{W}^*_s) = r$. The same conclusion holds for $\mathbf{W}^*_k$. We then prove that $\mathbf{W}^*_{s\setminus k,k\setminus s}$ has the representation of (6). Recall the eigen-decomposition of $\mathbf{W}^* = \mathbf{U}^*\mathbf{\Sigma}^*(\mathbf{U}^*)^\top$ and by definition we will have

$$\mathbf{W}^*_{s\cap k,s\cap k} = \mathbf{U}^*_{s\cap k}\mathbf{\Sigma}^*(\mathbf{U}^*_{s\cap k})^\top = \mathbf{V}^*_{s2}\mathbf{\Sigma}^*_s(\mathbf{V}^*_{s2})^\top = \mathbf{V}^*_{k1}\mathbf{\Sigma}^*_k(\mathbf{V}^*_{k1})^\top, \tag{S.1}$$

which also implies that $\text{rank}(\mathbf{V}^*_{s2}) = \text{rank}(\mathbf{V}^*_{k1}) = r$. Multiplying $\mathbf{V}^*_{k1}$ on the both sides of the last equation, we obtain

$$\mathbf{V}^*_{s2}(\mathbf{\Sigma}^*_s)^{1/2}(\mathbf{\Sigma}^*_s)^{1/2}(\mathbf{V}^*_{s2})^\top\mathbf{V}^*_{k1} = \mathbf{V}^*_{k1}(\mathbf{\Sigma}^*_k)^{1/2}(\mathbf{\Sigma}^*_k)^{1/2}(\mathbf{V}^*_{k1})^\top\mathbf{V}^*_{k1}$$

and the following equation

$$\mathbf{V}^*_{k1}(\mathbf{\Sigma}^*_k)^{1/2} = \mathbf{V}^*_{s2}(\mathbf{\Sigma}^*_s)^{1/2}\widehat{\mathbf{R}},$$

where $\widehat{\mathbf{R}} = (\mathbf{\Sigma}^*_s)^{1/2}(\mathbf{V}^*_{s2})^\top\mathbf{V}^*_{k1}\big((\mathbf{V}^*_{k1})^\top\mathbf{V}^*_{k1}\big)^{-1}(\mathbf{\Sigma}^*_k)^{-1/2}$. It is easy to verify that $\widehat{\mathbf{R}}^\top\widehat{\mathbf{R}} = \mathbf{I}_r$ and then it is obvious that

$$\widehat{\mathbf{R}} = \arg\min_{\mathbf{R}\in\mathscr{O}^{r\times r}} \|\mathbf{V}^*_{s2}(\mathbf{\Sigma}^*_s)^{1/2}\mathbf{R} - \mathbf{V}^*_{k1}(\mathbf{\Sigma}^*_k)^{1/2}\|_{\text{F}}.$$

Then by Lemma 22 of Ma et al. (2018), we prove that $\widehat{\mathbf{R}} = \mathbf{G}\big((\mathbf{V}^*_{s2}(\mathbf{\Sigma}^*_s)^{1/2})^\top\mathbf{V}^*_{k1}(\mathbf{\Sigma}^*_k)^{1/2}\big)$.

Again by (S.1), we have

$$(\mathbf{U}^*_{s\cap k})^\top = \mathbf{\Sigma}^{*-1}\big((\mathbf{U}^*_{s\cap k})^\top\mathbf{U}^*_{s\cap k}\big)^{-1}(\mathbf{U}^*_{s\cap k})^\top\mathbf{V}^*_{k1}\mathbf{\Sigma}^*_k(\mathbf{V}^*_{k1})^\top. \tag{S.2}$$

In addition, we have

$$\mathbf{U}^*_{s\cap k}\mathbf{\Sigma}^*(\mathbf{U}_{s\setminus k})^\top = \mathbf{W}^*_{s\cap k,k\setminus s} = \mathbf{V}^*_{k1}\mathbf{\Sigma}^*_k(\mathbf{V}^*_{k2})^\top. \tag{S.3}$$

Combining (S.2) and (S.3), we have

$$\begin{aligned}
&(\mathbf{U}^*_{s\cap k})^\top\mathbf{V}^*_{k1}\{(\mathbf{V}^*_{k1})^\top\mathbf{V}^*_{k1}\}^{-1}(\mathbf{V}^*_{k2})^\top\\
&= \mathbf{\Sigma}^{*-1}\big((\mathbf{U}^*_{s\cap k})^\top\mathbf{U}^*_{s\cap k}\big)^{-1}(\mathbf{U}^*_{s\cap k})^\top\mathbf{V}^*_{k1}\mathbf{\Sigma}^*_k(\mathbf{V}^*_{k1})^\top\mathbf{V}^*_{k1}\{(\mathbf{V}^*_{k1})^\top\mathbf{V}^*_{k1}\}^{-1}(\mathbf{V}^*_{k2})^\top\\
&= \mathbf{\Sigma}^{*-1}\big((\mathbf{U}^*_{s\cap k})^\top\mathbf{U}^*_{s\cap k}\big)^{-1}(\mathbf{U}^*_{s\cap k})^\top\mathbf{V}^*_{k1}\mathbf{\Sigma}^*_k(\mathbf{V}^*_{k2})^\top\\
&= \mathbf{\Sigma}^{*-1}\big((\mathbf{U}^*_{s\cap k})^\top\mathbf{U}^*_{s\cap k}\big)^{-1}(\mathbf{U}^*_{s\cap k})^\top\mathbf{U}^*_{s\cap k}\mathbf{\Sigma}^*(\mathbf{U}_{s\setminus k})^\top = (\mathbf{U}_{s\setminus k})^\top,
\end{aligned} \tag{S.4}$$

where the first equation comes from (S.2) and the second equation comes from (S.3). Finally,

$$\begin{aligned}
&\mathbf{V}^*_{s1}(\mathbf{\Sigma}^*_s)^{1/2}\mathbf{G}((\mathbf{\Sigma}^*_s)^{1/2}(\mathbf{V}^*_{s2})^\top\mathbf{V}^*_{k1}(\mathbf{\Sigma}^*_k)^{1/2})(\mathbf{\Sigma}^*_k)^{1/2}(\mathbf{V}^*_{k2})^\top\\
&= \mathbf{V}^*_{s1}(\mathbf{\Sigma}^*_s)^{1/2}(\mathbf{\Sigma}^*_s)^{1/2}(\mathbf{V}^*_{s2})^\top\mathbf{V}^*_{k1}\{(\mathbf{V}^*_{k1})^\top\mathbf{V}^*_{k1}\}^{-1}(\mathbf{\Sigma}^*_k)^{-1/2}(\mathbf{\Sigma}^*_k)^{1/2}(\mathbf{V}^*_{k2})^\top\\
&= \mathbf{U}^*_{s\setminus k}\mathbf{\Sigma}^*(\mathbf{U}^*_{s\cap k})^\top\mathbf{V}^*_{k1}\{(\mathbf{V}^*_{k1})^\top\mathbf{V}^*_{k1}\}^{-1}(\mathbf{V}^*_{k2})^\top\\
&= \mathbf{U}^*_{s\setminus k}\mathbf{\Sigma}^*(\mathbf{U}^*_{k\setminus s})^\top = \mathbf{W}^*_{s\setminus k,k\setminus s},
\end{aligned}$$

where the first equation comes from $\widehat{\mathbf{R}} = \mathbf{G}\big((\mathbf{V}^*_{s2}(\mathbf{\Sigma}^*_s)^{1/2})^\top\mathbf{V}^*_{k1}(\mathbf{\Sigma}^*_k)^{1/2}\big)$, the second equation comes from $\mathbf{V}^*_{s1}\mathbf{\Sigma}^*_s(\mathbf{V}^*_{s2})^\top = \mathbf{W}^*_{s\setminus k,s\cap k} = \mathbf{U}^*_{s\setminus k}\mathbf{\Sigma}^*(\mathbf{U}^*_{s\cap k})^\top$, and the third equation comes from (S.4). Then we finish the proof. ■

S.2. Proof of Theorem 7

When $m = 2$, we adopt the notations of Section 2.4 by assuming the two observed submatrices are $\mathbf{W}^s$ and $\mathbf{W}^k$. To prove the theorem, recall that $\widetilde{\mathbf{W}}_{sk}$ defined by (8) is the estimate of $\mathbf{W}^*_{s\setminus k,k\setminus s}$, the main effort lies on the perturbation bound of $\|\widetilde{\mathbf{W}}_{sk} - \mathbf{W}^*_{s\setminus k,k\setminus s}\|$. After we obtain it, the perturbation bound of $\|\widehat{\mathbf{W}} - \mathbf{W}^*\|$ can also be figured out. As a result, the error of the rank r factorization of $\widehat{\mathbf{W}}$ can also be bounded, which leads to Theorem 7. Before we derive $\|\widetilde{\mathbf{W}}_{sk} - \mathbf{W}^*_{s\setminus k,k\setminus s}\|$, we need the basic spectral properties of $\mathbf{W}^*_0$ defined in (5), $\mathbf{W}^*_s$, $\mathbf{W}^*_k$ defined in (4), which are presented in the Section S.2.1.

S.2.1 Characterization of The Underlying Matrix

Recall that $N_s = |\mathcal{V}_s|$, $N_k = |\mathcal{V}_k|$ and $N_{sk} = |\mathcal{V}_s \cap \mathcal{V}_k|$. First, by Lemma S.1, we have

$$\frac{p_l N}{2} \le N_l \le \frac{3p_l N}{2}, \text{ for } l = s,k \text{ and } \frac{p_s p_k N}{2} \le N_{sk} \le \frac{3p_s p_k N}{2} \tag{S.5}$$

hold simultaneously with probability $1 - O(1/N^3)$. Throughout, our analysis is conditional on (S.5). By Proposition S.3, we have

$$\lambda_r(\mathbf{W}^*_l) \ge \frac{N_l\lambda_{\min}}{2N} \ge \frac{p_l\lambda_{\min}}{4}, \text{ for } l = s,k$$

hold simultaneously with probability $1-2/N^3$ since by the Assumption 2, we have $N_l \geq Np_0/2 \geq 16\mu_0 r(\log r + \log N^3), l = s, k$. Then by Lemma S.4, we will have

$$\mu_l = \mu(\mathbf{V}_l^*) = \frac{N_l}{r} \max_{i=1,\ldots,n_l} \sum_{j=1}^{r} \mathbf{V}_l^*(i,j)^2 \leq 2\tau\mu_0, l = s, k.$$

In addition, by Proposition S.5, we have

$$\lambda_1(\mathbf{W}_l^*) \leq \frac{n_l r\mu_0}{N}\lambda_{\max} \leq \frac{3p_l r\mu_0}{2}\lambda_{\max}, \text{ for } l = s, k.$$

As a result, we have the condition number of $\mathbf{W}_l^*$:

$$\tau_l = \lambda_1(\mathbf{W}_l^*)/\lambda_r(\mathbf{W}_l^*) \leq 6r\mu_0\tau, \text{ for } l = s, k.$$

S.2.2 Imputation Error

After we characterize the spectral properties of $\mathbf{W}_0^*$ defined in (5), $\mathbf{W}_l^*, l = s, k$, we begin to control $\|\widetilde{\mathbf{W}}_{sk} - \mathbf{W}^*_{s\setminus k, k\setminus s}\|$. Using the notations of Proposition 2 and Section 2.4, we define

$$\begin{aligned}
&\mathbf{A} = \mathbf{V}_s^*(\mathbf{\Sigma}_s^*)^{1/2};\ \mathbf{B} = \mathbf{V}_k^*(\mathbf{\Sigma}_k^*)^{1/2};\ \widetilde{\mathbf{A}} = \widetilde{\mathbf{V}}_s(\widetilde{\mathbf{\Sigma}}_s)^{1/2};\ \widetilde{\mathbf{B}} = \widetilde{\mathbf{V}}_k(\widetilde{\mathbf{\Sigma}}_k)^{1/2};\\
&\mathbf{A}_1 = \mathbf{V}_{11}^*(\mathbf{\Sigma}_1^*)^{1/2};\ \mathbf{A}_2 = \mathbf{V}_{12}^*(\mathbf{\Sigma}_1^*)^{1/2};\ \mathbf{B}_1 = \mathbf{V}_{21}^*(\mathbf{\Sigma}_2^*)^{1/2};\ \mathbf{B}_2 = \mathbf{V}_{22}^*(\mathbf{\Sigma}_1^*)^{1/2};\\
&\widetilde{\mathbf{A}}_1 = \widetilde{\mathbf{V}}_{11}(\widetilde{\mathbf{\Sigma}}_1)^{1/2};\ \widetilde{\mathbf{A}}_2 = \widetilde{\mathbf{V}}_{12}(\widetilde{\mathbf{\Sigma}}_1)^{1/2};\ \widetilde{\mathbf{B}}_1 = \widetilde{\mathbf{V}}_{21}(\widetilde{\mathbf{\Sigma}}_2)^{1/2};\ \widetilde{\mathbf{B}}_2 = \widetilde{\mathbf{V}}_{22}(\widetilde{\mathbf{\Sigma}}_2)^{1/2}
\end{aligned} \tag{S.6}$$

and $\mathbf{Q}_A = \mathbf{G}(\widetilde{\mathbf{A}}^\top\mathbf{A})$, $\mathbf{Q}_B = \mathbf{G}(\widetilde{\mathbf{B}}^\top\mathbf{B})$, $\widetilde{\mathbf{O}} = \mathbf{G}(\widetilde{\mathbf{A}}_2^\top\widetilde{\mathbf{B}}_1)$. It is easy to see that

$$\widetilde{\mathbf{W}}_{sk} = \widetilde{\mathbf{A}}_1\widetilde{\mathbf{O}}^\top\widetilde{\mathbf{B}}_2^\top = \widetilde{\mathbf{A}}_1\mathbf{Q}_A(\mathbf{Q}_A^\top\widetilde{\mathbf{O}}^\top\mathbf{Q}_B)\mathbf{Q}_B^\top\widetilde{\mathbf{B}}_2^\top = \widetilde{\mathbf{A}}_1\mathbf{Q}_A\mathbf{G}(\mathbf{Q}_B^\top\widetilde{\mathbf{B}}_1^\top\widetilde{\mathbf{A}}_2\mathbf{Q}_A)\mathbf{Q}_B^\top\widetilde{\mathbf{B}}_2^\top. \tag{S.7}$$

Then by Proposition 2, we have

$$\begin{aligned}
&\|\widetilde{\mathbf{W}}_{sk} - \mathbf{W}^*_{s\setminus k, k\setminus s}\| = \|(\widetilde{\mathbf{A}}_1\mathbf{Q}_A)(\mathbf{Q}_A^\top\widetilde{\mathbf{O}}^\top\mathbf{Q}_B)(\mathbf{Q}_B^\top\widetilde{\mathbf{B}}_2^\top) - \mathbf{A}_1\mathbf{O}^\top\mathbf{B}_2^\top\|\\
=&\|(\widetilde{\mathbf{A}}_1\mathbf{Q}_A)(\mathbf{Q}_A^\top\widetilde{\mathbf{O}}^\top\mathbf{Q}_B)(\mathbf{Q}_B^\top\widetilde{\mathbf{B}}_2^\top) - \mathbf{A}_1(\mathbf{Q}_A^\top\widetilde{\mathbf{O}}^\top\mathbf{Q}_B)(\mathbf{Q}_B^\top\widetilde{\mathbf{B}}_2^\top)\\
&+ \mathbf{A}_1(\mathbf{Q}_A^\top\widetilde{\mathbf{O}}^\top\mathbf{Q}_B)(\mathbf{Q}_B^\top\widetilde{\mathbf{B}}_2^\top) - \mathbf{A}_1(\mathbf{Q}_A^\top\widetilde{\mathbf{O}}^\top\mathbf{Q}_B)\mathbf{B}_2^\top\\
&+ \mathbf{A}_1(\mathbf{Q}_A^\top\widetilde{\mathbf{O}}^\top\mathbf{Q}_B)\mathbf{B}_2^\top - \mathbf{A}_1\mathbf{O}^\top\mathbf{B}_2^\top\|\\
\leq&\|\widetilde{\mathbf{B}}_2\|\|\widetilde{\mathbf{A}}_1\mathbf{Q}_A - \mathbf{A}_1\| + \|\widetilde{\mathbf{A}}_1\|\|\widetilde{\mathbf{B}}_2\mathbf{Q}_B - \mathbf{B}_2\| + \|\mathbf{A}_1\|\|\mathbf{B}_2\|\|\mathbf{Q}_A^\top\widetilde{\mathbf{O}}^\top\mathbf{Q}_B - \mathbf{O}\|.
\end{aligned}$$

Applying Proposition S.9, Lemma S.6, Lemma S.7, Lemma S.10, with $f(p_0, N)$ defined in (S.18), we have

$$\begin{aligned}
&\|\widetilde{\mathbf{W}}_{sk} - \mathbf{W}^*_{s\setminus k, k\setminus s}\|\\
&\lesssim (1-p_0)(\|\mathbf{B}\|\|\widetilde{\mathbf{A}}\mathbf{Q}_A - \mathbf{A}\| + \|\mathbf{A}\|\|\widetilde{\mathbf{B}}\mathbf{Q}_B - \mathbf{B}\| + \|\mathbf{A}\|\|\mathbf{B}\|\|\mathbf{Q}_A^\top\widetilde{\mathbf{O}}^\top\mathbf{Q}_B - \mathbf{O}\|)\\
&\lesssim (1-p_0)\left\{r\mu_0\tau + f(p_0,N)^2(r\mu_0\tau)^2\right\}(\|\widetilde{\mathbf{E}}_1\| + \|\widetilde{\mathbf{E}}_2\|)\\
&\lesssim (1-p_0)(r\mu_0\tau)^2 f(p_0,N)^2\sqrt{Np_0}\sigma
\end{aligned}$$

with probability $1 - O(1/N^3)$.

S.2.3 Completion Error

After we impute the missing blocks, we can bound $\|\widehat{\mathbf{W}} - \mathbf{W}^*\|$ where $\widehat{\mathbf{W}}$ is defined as (9). Notice that

$$\widehat{\mathbf{W}} = \mathbf{W}^* + \widetilde{\mathbf{E}} + \widetilde{\mathbf{F}}, \tag{S.8}$$

where

$$\widetilde{\mathbf{E}} = \begin{bmatrix} \mathbf{E}^s_{s\setminus k, s\setminus k} & & \mathbf{E}^s_{s\setminus k, s\cap k} & \mathbf{O} \\ \mathbf{E}^s_{s\cap k, s\setminus k} & \alpha_s \mathbf{E}^s_{s\cap k, s\cap k} + \alpha_k \mathbf{E}^k_{s\cap k, s\cap k} & & \mathbf{E}^k_{s\cap k, k\setminus s} \\ \mathbf{O} & & \mathbf{E}^k_{k\setminus s, s\cap k} & \mathbf{E}^k_{k\setminus s, k\setminus s} \end{bmatrix},$$

and

$$\widetilde{\mathbf{F}} = \begin{bmatrix} \mathbf{O} & \mathbf{O} & \widetilde{\mathbf{W}}_{sk} - \mathbf{W}^*_{s\setminus k, k\setminus s} \\ \mathbf{O} & \mathbf{O} & \mathbf{O} \\ \widetilde{\mathbf{W}}^\top_{sk} - \mathbf{W}^*_{k\setminus s, s\setminus k} & \mathbf{O} & \mathbf{O} \end{bmatrix}.$$

Then we only need to bound $\|\widetilde{\mathbf{E}}\|$ and $\|\widetilde{\mathbf{F}}\|$. It is easy to see that $\|\widetilde{\mathbf{F}}\| = \|\widetilde{\mathbf{W}}_{sk} - \mathbf{W}^*_{s\setminus k, k\setminus s}\|$, then we only need to bound $\|\widetilde{\mathbf{E}}\|$. It is easy to see that $\|\widetilde{\mathbf{E}}\| \le \|\mathbf{E}^s\| + \|\mathbf{E}^k\| \lesssim \sqrt{Np_0}\sigma$. However, to give some intuition on the choice of α_s and α_k, we consider a special case that the entries of $\mathbf{E}^s$ and $\mathbf{E}^k$ are independent mean zero sub-Gaussian random variables with sub-Gaussian norm σ. Then by Corollary 3.3 of Bandeira and van Handel (2016), we have

$$\mathbb{E}\|\widetilde{\mathbf{E}}\| \lesssim \sigma^* + \sigma\sqrt{\log N_0},$$

where $\sigma = \max\{\sigma_s, \sigma_k\}$ and $\sigma^* = \max_i \sqrt{\sum_j \mathbb{E}\widetilde{\mathbf{E}}^2_{ij}}$. It is easy to see that

$$\sigma^* = \max\{\sqrt{N_s}\sigma_s, \sqrt{N_k}\sigma_k, \sqrt{(N_s - N_{sk})\sigma_s^2 + (N_k - N_{sk})\sigma_k^2 + N_{sk}(\alpha_s^2\sigma_s^2 + \alpha_k^2\sigma_k^2)}\}.$$

In addition, by Lemma 11 and Proposition 1 of Chen and Wainwright (2015), there exists a universal constant $c > 0$ such that

$$\mathbb{P}\{\|\widetilde{\mathbf{E}}\| \ge c(\sigma^* + \sigma\log N_0)\} \le N_0^{-12}.$$

In order to minimize $\|\widetilde{\mathbf{E}}\|$ with regard to α_s and α_k, the best we can do is to minimize its upper bound. It is easy to see that

$$(\alpha_1^*, \alpha_2^*) = (\sigma_2^2/(\sigma_1^2 + \sigma_2^2), \sigma_1^2/(\sigma_1^2 + \sigma_2^2)) = \arg\min_{\alpha_1+\alpha_2=1, \alpha_1>0, \alpha_2>0} \alpha_1^2\sigma_1^2 + \alpha_2^2\sigma_2^2.$$

In reality, we do not know σ_s and σ_k, but we can estimate them by (10). Since

$$\alpha_1^2\sigma_1^2 + \alpha_2^2\sigma_2^2 \le (\alpha_1^2 + \alpha_2^2)\sigma^2 \le (\alpha_1 + \alpha_2)^2\sigma^2 = \sigma^2,$$

we have $\sigma^* \le \sqrt{N_0}\sigma$. So $\|\widetilde{\mathbf{E}}\| \lesssim \sigma^* \le \sqrt{N_0}\sigma$ with probability at least $1 - N_0^{-12} \ge 1 - O(1/N^3)$. By $N_0 = N_s + N_k - N_{sk} \le 3Np_s/2 + 3Np_k/2 - Np_sp_k/2 \lesssim Np_0$, we get $\sigma^* \lesssim \sqrt{Np_0}\sigma$. Finally, we have

$$\|\widehat{\mathbf{W}} - \mathbf{W}^*\| \le \|\widetilde{\mathbf{E}}\| + \|\widetilde{\mathbf{F}}\| \lesssim \sqrt{Np_0}\sigma + (1 - p_0)(r\mu_0\tau)^2 f(p_0, N)^2\sqrt{Np_0}\sigma.$$

The inequality still holds if the sub-Gaussian assumption does not hold.

S.2.4 Low-rank Approximation

The last step is to do rank-r eigendecomposition on $\widehat{\mathbf{W}}$ to obtain $\widehat{\mathbf{W}}_r = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{U}}^\top = \widehat{\mathbf{X}}\widehat{\mathbf{X}}^\top$ where $\widehat{\mathbf{X}} = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}^{1/2}$. Then there exists an orthogonal matrix $\mathbf{O}_X$ such that

$$\begin{aligned}\|\widehat{\mathbf{X}}\mathbf{O}_X - \mathbf{X}^*\| &\lesssim \frac{\|\widehat{\mathbf{W}} - \mathbf{W}^*\| r\mu_0\tau}{\sqrt{\lambda_r(\mathbf{W}_0^*)}} \lesssim \frac{\|\widehat{\mathbf{W}} - \mathbf{W}^*\| r\mu_0\tau}{\sqrt{\lambda_{\min} p_0}} \\ &\lesssim \{(1-p_0)(r\mu_0\tau)^2 f(p_0,N)^2 + 1\} r\mu_0\tau \sqrt{\frac{N}{\lambda_{\min}}}\sigma.\end{aligned}$$

by a similar proof as Lemma S.7 and the fact that $\lambda_r(\mathbf{W}_0^*) \geq \lambda_r(\mathbf{W}_s^*) \geq p_0\lambda_{\min}/4$. Finally, this upper bound holds with probability at least $1 - O(1/N^3)$ by the probability union bound.

S.3. Proof of Theorem 12

We know that $N_0 \sim \text{Binomial}(N, 1 - \prod_{s=1}^m (1-p_s))$, so by the same argument to Lemma S.1, we have

$$N\{1 - \prod_{s=1}^m (1-p_s)\}/2 \leq N_0 \leq 3N\{1 - \prod_{s=1}^m (1-p_s)\}/2$$

with probability $1 - O(1/N^3)$. As a result,

$$\{1-(1-p_0)^m\}\lambda_{\min} \lesssim \lambda_r(\mathbf{W}_0^*) \leq \lambda_1(\mathbf{W}_0^*) \lesssim \{1-(1-p_0)^m\} r\mu_0\lambda_{\max} \tag{S.9}$$

by a similar argument as in the proof of Theorem 7 and the Assumption 2 that $p_s/p_0 = O(1)$. In addition, let $\mathbf{E} = \widehat{\mathbf{W}} - \mathbf{W}_0^*$, then by a similar decomposition as in (S.8), we will have

$$\|\mathbf{E}\| \leq \|\widetilde{\mathbf{E}}\| + \sum_{s=1}^{m-1}\sum_{k=s+1}^{m} \|\mathbf{T}^{sk} \circ (\widetilde{\mathbf{W}}_{sk} - \mathbf{W}^*_{s\backslash k, k\backslash s})\|$$

where $\widetilde{\mathbf{E}} \in R^{n\times n}$ with

$$\widetilde{\mathbf{E}}(i,j) = \sum_{s=1}^m \alpha_{ij}^s \mathbf{E}^s(v_i^s, v_j^s)\mathbb{1}(i,j \in \mathcal{V}_s), \text{ for } \mathcal{S}_{ij} > 0$$

and $\widetilde{\mathbf{E}}(i,j) = 0$ for $\mathcal{S}_{ij} = 0$. Here we denote $\circ$ as the Hadamard product operator and $\mathbf{T}^{sk}, s \neq k \in [m]$ are 0/1 matrices decided by the Algorithm 1. According to the Algorithm 1, the nonzero entries of $\mathbf{T}^{sk}, s \neq k \in [m]$ are block-wise, which implies that

$$\|\mathbf{T}^{sk} \circ (\widetilde{\mathbf{W}}_{sk} - \mathbf{W}^*_{s\backslash k, k\backslash s})\| \leq \|\widetilde{\mathbf{W}}_{sk} - \mathbf{W}^*_{s\backslash k, k\backslash s}\|.$$

Then, by the proof of Theorem 7, we have $\|\widetilde{\mathbf{E}}\| \lesssim \sqrt{Np_0}\sigma$ and

$$\|\widetilde{\mathbf{W}}_{sk} - \mathbf{W}^*_{s\backslash k, k\backslash s}\| \lesssim (1-p_0)(r\mu_0\tau)^2 f(p_0,N)^2\sqrt{Np_0}\sigma$$

hold simultaneously with probability $1 - O(m^2/N^3)$ for $1 \leq s < k \leq m$. As a result,

$$\|\widehat{\mathbf{W}} - \mathbf{W}_0^*\| \lesssim m(m-1)(1-p_0)(r\mu_0\tau)^2 f(p_0,N)^2\sqrt{Np_0}\sigma + \sqrt{Np_0}\sigma$$

and

$$\|\widehat{\mathbf{X}}\mathbf{O}_X - \mathbf{X}^*\| \lesssim \frac{\|\widehat{\mathbf{W}} - \mathbf{W}_0^*\| r\mu_0\tau}{\sqrt{\lambda_r(\mathbf{W}_0^*)}}.$$

By (S.9), we have

$$\|\widehat{\mathbf{X}}\mathbf{O}_X - \mathbf{X}^*\| \lesssim \Big\{1 + m^2(1-p_0)(r\mu_0\tau)^2 f(p_0,N)^2 \sqrt{\frac{p_0}{1-(1-p_0)^m}}\Big\} r\mu_0\tau\sqrt{\frac{N}{\lambda_{\min}}}\sigma$$

with probability $1 - O(m^2/N^3)$. Given $0 < \epsilon < 1$, we have

$$\mathbb{P}\{N_0 < (1-\epsilon)N\} = O\big(\frac{1}{N^3}\big)$$

when $m \approx \log(\epsilon - \sqrt{\frac{3\log N}{2N}})/\log(1-p_0)$ by the fact that $N_0 \sim \text{Binomial}(N, 1-\prod_{s=1}^m(1-p_s))$ and the Bernstein inequality. Since $\lim_{N\to\infty}\sqrt{\log N/N} = 0$ we have $m \approx \log\epsilon/\log(1-p_0)$. Finally, we have

$$\|\widehat{\mathbf{X}}\mathbf{O}_X - \mathbf{X}^*\| \lesssim \Big\{1 + \frac{\log^2\epsilon}{\log^2(1-p_0)}(1-p_0)(r\mu_0\tau)^2 f(p_0,N)^2 \sqrt{\frac{p_0}{1-(1-p_0)^m}}\Big\} r\mu_0\tau\sqrt{\frac{N}{\lambda_{\min}}}\sigma$$

with probability $1 - O(m^2/N^3)$.

S.4. Details of the Proof of Theorem 7

Here we present some key lemmas and propositions needed for our proof of Theorem 7.

Lemma S.1 (The dimension of the sub-matrices) *Under the assumption that*

$$p_s \geq p_0 \geq C\sqrt{\mu_0 r\tau\log N/N},$$

for some sufficiently large constant C, we have

$$\frac{p_s N}{2} \leq N_s \leq \frac{3p_s N}{2} \text{ and } \frac{p_s p_k N}{2} \leq N_{sk} \leq \frac{3p_s p_k N}{2}, s \neq k, s, k \in [m] \tag{S.10}$$

with probabilities $1 - O(m^2/N^3)$.

Proof By the Bernstein inequality, we have

$$\mathbb{P}\{Y \leq pn - t\} \leq \exp\{-\frac{\frac{1}{2}t^2}{np(1-p)+\frac{1}{3}t}\} \text{ and } \mathbb{P}\{Y \geq pn + t\} \leq \exp\{-\frac{\frac{1}{2}t^2}{np(1-p)+\frac{1}{3}t}\}$$

if $Y \sim \text{Binomial}(n,p)$. Since $N_s \sim \text{Binomial}(N, p_s)$ and $N_{sk} \sim \text{Binomial}(N, p_s p_k)$, let $t = \frac{p_s}{2}$, we have

$$\mathbb{P}\{\frac{p_s N}{2} \leq N_s \leq \frac{3p_s N}{2}\} \geq 1 - 2\exp\{-\frac{3p_s N}{28}\}.$$

Similarly, we have

$$\mathbb{P}\{\frac{p_s p_k N}{2} \leq N_{sk} \leq \frac{3p_s p_k N}{2}\} \geq 1 - 2\exp\{-\frac{3p_s p_k N}{28}\}.$$

In addition, by $p_s \geq p_0 \geq C\sqrt{\mu_0 r\tau \log N/N}$, we have $\exp\{-3p_sN/28\} = O(1/N^3)$ and $\exp\{-3p_sp_kN/28\} = O(1/N^3)$. Finally, by the probability union bound, (S.10) holds with probability $1 - O(m^2/N^3)$. ■

Lemma S.2 (Lemma 5, Cai et al. (2016)) *Suppose $\mathbf{U} \in \mathbb{R}^{N\times r}$ ($N \geq r$) is a fixed matrix with orthonormal columns. Denote $\mu = \max_{1\leq i\leq N} \frac{N}{r}\sum_{j=1}^r u_{ij}^2$. Suppose we uniform randomly draw n rows (with or without replacement) from $\mathbf{U}$ and denote it as $\mathbf{U}_\Omega$, where Ω is the index set. When $n \geq 4\mu r(\log r + c)/(1-\alpha)^2$ for some $0 < \alpha < 1$ and $c > 1$, we have*

$$\sigma_{\min}(\mathbf{U}_\Omega) \geq \sqrt{\frac{\alpha n}{N}}$$

with probability $1 - 2e^{-c}$.

By Lemma S.2, we will directly have the following proposition.

Proposition S.3 *Let $\alpha = \frac{1}{2}$ and $c = \log 2N^3$ in Lemma S.2, then when*

$$N_s \geq 16\mu_0 r(\log r + \log 2N^3),$$

we have $\sigma_{\min}(\mathbf{U}^*_{\mathcal{V}_s}) \geq \sqrt{\frac{N_s}{2N}}$ with probability $1 - 1/N^3$. In addition, under the event, we have

$$\lambda_r(\mathbf{W}_s^*) = \lambda_r(\mathbf{U}^*_{\mathcal{V}_s}\mathbf{\Sigma}^*(\mathbf{U}^*_{\mathcal{V}_s})^\top) \geq \sigma_{\min}(\mathbf{U}^*_{\mathcal{V}_s})\lambda_r(\mathbf{\Sigma}^*)\sigma_{\min}(\mathbf{U}^*_{\mathcal{V}_s}) \geq \frac{N_s\lambda_{\min}}{2N}.$$

Lemma S.4 (Incoherence condition of the sub-matrices) *Recall that $\mathbf{V}_s^*\mathbf{\Sigma}_s^*(\mathbf{V}_s^*)^\top$ is the rank-r eigendecomposition of $\mathbf{W}_s^*$. Assume that $\lambda_r(\mathbf{W}_s^*) \geq \frac{N_s\lambda_{\min}}{2N}$. Then the incoherence of $\mathbf{V}_s^*$ satisfies*

$$\mu_s \equiv \mu(\mathbf{V}_s^*) = \frac{N_s}{r}\max_{i=1,\ldots,n_s}\sum_{j=1}^r \mathbf{V}_s^*(i,j)^2 \leq 2\tau\mu_0.$$

Proof Since $\mathbf{W}_s^* = \mathbf{U}^*_{\mathcal{V}_s}\mathbf{\Sigma}^*(\mathbf{U}^*_{\mathcal{V}_s})^\top = \mathbf{V}_s^*\mathbf{\Sigma}_s^*(\mathbf{V}_s^*)^\top$, we have

$$\mathbf{V}_s^* = \mathbf{U}^*_{\mathcal{V}_s}(\mathbf{\Sigma}^*)^{\frac{1}{2}}\mathbf{O}_s^\top(\mathbf{\Sigma}_s^*)^{-\frac{1}{2}}$$

where $\mathbf{O}_s = (\mathbf{\Sigma}_s^*)^{-\frac{1}{2}}(\mathbf{V}_s^*)^\top\mathbf{U}^*_{\mathcal{V}_s}(\mathbf{\Sigma}^*)^{\frac{1}{2}} \in \mathscr{O}^{r\times r}$. Then

$$\sum_{j=1}^r \mathbf{V}_s^*(i,j)^2 \leq \sum_{j=1}^r \mathbf{U}^*_{\mathcal{V}_s}(i,j)^2\|(\mathbf{\Sigma}_s^*)^{-\frac{1}{2}}\|^2\|(\mathbf{\Sigma}^*)^{\frac{1}{2}}\|^2 \leq \frac{r\mu_0}{N}\frac{\lambda_{\max}}{\lambda_r(\mathbf{W}_s^*)}$$

As a result,

$$\mu_s = \frac{N_s}{r}\max_{i=1,\ldots,n_s}\sum_{j=1}^r \mathbf{V}_s^*(i,j)^2 \leq \frac{N_s\mu_0}{N}\frac{\sigma_{\max}}{\sigma_r(\mathbf{W}_s^*)} \leq \frac{2\lambda_{\max}}{\lambda_{\min}}\mu_0 = 2\tau\mu_0.$$

■

Proposition S.5 (Upper bound of the operator norm of the sub-matrices) *We have*

$$\lambda_1(\mathbf{W}_s^*) \leq \min\{1, \frac{N_s r\mu_0}{N}\}\lambda_{\max} \text{ for } s \in [m].$$

Proof It is obviously that $\lambda_1(\mathbf{W}_s^*) = \lambda_1(\mathbf{U}_{\mathcal{V}_s}^*\mathbf{\Sigma}^*(\mathbf{U}_{\mathcal{V}_s}^*)^\top) \leq \sigma_{\max}(\mathbf{U}_{\mathcal{V}_s}^*)^2\lambda_{\max}(\mathbf{\Sigma}^*) \leq \lambda_{\max}$ because $\sigma_{\max}(\mathbf{U}_{\mathcal{V}_s}^*) \leq 1$. Besides, we have $\|\mathbf{U}_{\mathcal{V}_s}^*\|^2 \leq N_s\|\mathbf{U}_{\mathcal{V}_s}^*\|_{2,\infty}^2 \leq N_s r\mu_0/N$ where the first inequality comes from the property of ℓ_2/ℓ_∞ norm and the second inequality comes from $\mu_0 = \mu(\mathbf{U}^*)$ and the definition of incoherence. ■

S.4.1 Error Matrix

Recalling that $\widetilde{\mathbf{W}}_s \equiv \widetilde{\mathbf{W}}_{\mathcal{V}_s,\mathcal{V}_s}$, we characterize the operator norm of $\widetilde{\mathbf{W}}_s - \mathbf{W}_s^*, s \in [m]$ in the Lemma S.6.

Lemma S.6 *Let $\widetilde{\mathbf{E}}_s := \widetilde{\mathbf{W}}_s - \mathbf{W}_s^*, s \in [m]$. Under Assumptions 2, 3, and the condition $p_s N/2 \leq n_s \leq 3p_s N/2, s \in [m]$, we have*

$$\|\widetilde{\mathbf{E}}_s\| \lesssim \sqrt{Np_0}\sigma \ll \frac{p_0\lambda_{\min}}{4} \leq \lambda_r(\mathbf{W}_s^*), s \in [m]$$

with probability $1 - O(m/N^3)$.

Proof Recall that

$$\widetilde{\mathbf{W}}_s(v_i^s, v_j^s) = \widetilde{\mathbf{W}}(i,j) = \sum_{k=1}^m \alpha_{ij}^k \mathbf{W}^k(v_i^k, v_j^k)\mathbb{1}(i,j \in \mathcal{V}_k), i,j \in \mathcal{V}_s.$$

We then have

$$\widetilde{\mathbf{E}}_s(v_i^s, v_j^s) = \sum_{k=1}^m \alpha_{ij}^k \mathbf{E}^k(v_i^k, v_j^k)\mathbb{1}(i,j \in \mathcal{V}_s), i,j \in \mathcal{V}_s$$

and it is easy to see $\|\widetilde{\mathbf{E}}_s\| \leq \max_{k\in[m]}\|\widehat{\mathbf{E}}^k\| \lesssim \sqrt{N_s}\sigma, s \in [m]$. In addition, $N_s \leq 3p_s N/2$ leads to

$$\|\widetilde{\mathbf{E}}_s\| \lesssim \sqrt{Np_0}\sigma, s \in [m]$$

with probability at least $1 - O(m/N^3)$, and based on Assumption 3, we have

$$\|\widetilde{\mathbf{E}}_s\| \ll \frac{p_0\lambda_{\min}}{4} \leq \lambda_r(\mathbf{W}_s^*),\ s \in [m].$$

■

We then bound $\|\widetilde{\mathbf{A}}\mathbf{Q}_A - \mathbf{A}\|$ and $\|\widetilde{\mathbf{B}}\mathbf{Q}_B - \mathbf{B}\|$ for the case $m = 2$ in the following lemma.

Lemma S.7 *Based on the notation on Section S.2.2 with the assumptions that $\|\widetilde{\mathbf{E}}_l\| \ll \lambda_r(\mathbf{W}_l^*)$ and $\tau_l = \lambda_1(\mathbf{W}_l^*)/\lambda_r(\mathbf{W}_l^*), l = s,k$ are bounded, we have*

$$\|\widetilde{\mathbf{A}}\mathbf{Q}_A - \mathbf{A}\| \lesssim \frac{\tau_s}{\sqrt{\lambda_r(\mathbf{W}_s^*)}}\|\widetilde{\mathbf{E}}_s\| \quad \text{and} \quad \|\widetilde{\mathbf{B}}\mathbf{Q}_B - \mathbf{B}\| \lesssim \frac{\tau_k}{\sqrt{\lambda_r(\mathbf{W}_k^*)}}\|\widetilde{\mathbf{E}}_k\|.$$

Proof Define $\mathbf{Q}_s = \mathbf{G}(\widetilde{\mathbf{V}}_s^\top \mathbf{V}_s^*)$, $\mathbf{Q}_k = \mathbf{G}(\widetilde{\mathbf{V}}_k^\top \mathbf{V}_k^*)$ and recall that $\mathbf{Q}_A = \mathbf{G}(\widetilde{\mathbf{A}}^\top \mathbf{A})$ and $\mathbf{Q}_B = \mathbf{G}(\widetilde{\mathbf{B}}^\top \mathbf{B})$. The key decomposition we need is the following:

$$\widetilde{\mathbf{A}}\mathbf{Q}_A - \mathbf{A} = \widetilde{\mathbf{A}}(\mathbf{Q}_A - \mathbf{Q}_s) + \widetilde{\mathbf{V}}_s[\widetilde{\mathbf{\Sigma}}_s^{\frac{1}{2}}\mathbf{Q}_s - \mathbf{Q}_s(\mathbf{\Sigma}_s^*)^{\frac{1}{2}}] + (\widetilde{\mathbf{V}}_s\mathbf{Q}_s - \mathbf{V}_s^*)(\mathbf{\Sigma}_s^*)^{\frac{1}{2}}. \tag{S.11}$$

For the spectral norm error bound, the triangle inequality together with (S.11) yields

$$\|\widetilde{\mathbf{A}}\mathbf{Q}_A - \mathbf{A}\| \le \|\widetilde{\mathbf{\Sigma}}_s^{\frac{1}{2}}\|\|\mathbf{Q}_A - \mathbf{Q}_s\| + \|\widetilde{\mathbf{\Sigma}}_s^{\frac{1}{2}}\mathbf{Q}_s - \mathbf{Q}_s(\mathbf{\Sigma}_s^*)^{\frac{1}{2}}\| + \sqrt{\lambda_1(\mathbf{\Sigma}_s^*)}\|\widetilde{\mathbf{V}}_s\mathbf{Q}_s - \mathbf{V}_s^*\|,$$

where we have also used the fact that $\|\widetilde{\mathbf{V}}_s\| = 1$. Recognizing that $\|\widetilde{\mathbf{W}}_s - \mathbf{W}_s^*\| = \|\widetilde{\mathbf{E}}_s\| \ll \lambda_r(\mathbf{W}_s^*)$ and the assumption that $\lambda_1(\mathbf{W}_s^*)/\lambda_r(\mathbf{W}_s^*)$ is bounded, we can apply Lemmas 47, 46, 45 of Ma et al. (2018) to obtain

$$\|\mathbf{Q}_A - \mathbf{Q}_s\| \lesssim \frac{1}{\lambda_r(\mathbf{W}_s^*)}\|\widetilde{\mathbf{E}}_s\|,$$

$$\|\widetilde{\mathbf{\Sigma}}_s^{\frac{1}{2}}\mathbf{Q}_s - \mathbf{Q}_s(\mathbf{\Sigma}_s^*)^{\frac{1}{2}}\| \lesssim \frac{1}{\sqrt{\lambda_r(\mathbf{W}_s^*)}}\|\widetilde{\mathbf{E}}_s\|,$$

$$\|\widetilde{\mathbf{V}}_s\mathbf{Q}_s - \mathbf{V}_s^*\| \lesssim \frac{1}{\lambda_r(\mathbf{W}_s^*)}\|\widetilde{\mathbf{E}}_s\|.$$

These taken collectively imply the advertised upper bound

$$\|\widetilde{\mathbf{A}}\mathbf{Q}_A - \mathbf{A}\| \lesssim \frac{\sqrt{\lambda_1(\mathbf{W}_s^*)}}{\lambda_r(\mathbf{W}_s^*)}\|\widetilde{\mathbf{E}}_s\| + \frac{1}{\sqrt{\lambda_r(\mathbf{W}_s^*)}}\|\widetilde{\mathbf{E}}_s\| \lesssim \frac{\sqrt{\tau_s}}{\sqrt{\lambda_r(\mathbf{W}_s^*)}}\|\widetilde{\mathbf{E}}_s\|,$$

where we also utilize the fact that $\|\widetilde{\mathbf{\Sigma}}_s\| \le \|\mathbf{\Sigma}_s^*\| + \|\widetilde{\mathbf{E}}_s\| \le 2\|\mathbf{\Sigma}_s^*\| = 2\|\mathbf{W}_s^*\|$ and $\lambda_1(\mathbf{W}_s^*)/\lambda_r(\mathbf{W}_s^*)$ is bounded. Similarly, we have

$$\|\widetilde{\mathbf{B}}\mathbf{Q}_B - \mathbf{B}\| \lesssim \frac{\sqrt{\tau_k}}{\sqrt{\lambda_r(\mathbf{W}_k^*)}}\|\widetilde{\mathbf{E}}_k\|.$$

Combined with the fact that $\tau_l = \lambda_1(\mathbf{W}_l^*)/\lambda_r(\mathbf{W}_l^*) \le 6r\mu_0\tau, l = s, k$, we have

$$\|\widetilde{\mathbf{A}}\mathbf{Q}_A - \mathbf{A}\| \lesssim \frac{\sqrt{r\mu_0\tau}}{\sqrt{\lambda_r(\mathbf{W}_s^*)}}\|\widetilde{\mathbf{E}}_s\| \quad \text{and} \quad \|\widetilde{\mathbf{B}}\mathbf{Q}_B - \mathbf{B}\| \lesssim \frac{\sqrt{r\mu_0\tau}}{\sqrt{\lambda_r(\mathbf{W}_k^*)}}\|\widetilde{\mathbf{E}}_k\|.$$

■

S.4.2 Probability Bound for Submatrix

Lemma S.8 *Denote $\mathbf{R} \in \mathbb{R}^{d\times d}$ for the square diagonal matrix whose jth diagonal entry is y_j, where $\{y_j\}_{j=1}^n$ is a sequence of independent $0-1$ random variables with common mean p. Let $\mathbf{B} \in \mathbb{R}^{q\times d}$ with rank r and $d > \max\{e^2, r^2\}$.*

- *If $p = o(1/\log d)$ or p is bounded away from 0 for all d, we have*

$$\mathbb{P}\{\|\mathbf{BR}\| \ge Cp^{\frac{1}{2}}\|\mathbf{B}\|\} \le \delta \tag{S.12}$$

- *else,*

$$\mathbb{P}\{\|\mathbf{BR}\| \geq Cp^{\frac{1}{2}}\sqrt{p\log d}\|\mathbf{B}\|\} \leq \delta \tag{S.13}$$

for some universal positive constant C and $\delta = 1/d^3$.

Proof By Theorems 3.1 and 4.1 of Tropp (2008), we have

$$\mathbb{E}_k\|\mathbf{BR}\| \leq 6\sqrt{\max\{k, 2\log r\}}\frac{p}{1-p}\max_{|T|\leq p^{-1}}[\sum_{j\in T}\|\boldsymbol{b}_j\|_2^k]^{1/k} + \sqrt{p}\|\mathbf{B}\|. \tag{S.14}$$

for $k \in [2, \infty)$ where $\mathbb{E}_k\mathbf{X} = (\mathbb{E}|\mathbf{X}|^k)^{1/k}$ and the ℓ_1 to ℓ_2 operator norm $\|\cdot\|_{1\to 2}$ computes the maximum ℓ_2 norm of a column. In addition, $\boldsymbol{b}_j$ is the jth column of $\mathbf{B}$ and $T \subset [d]$. Since $\|\boldsymbol{b}_j\|_2 \leq \|\mathbf{B}\|$, we have

$$\max_{|T|\leq p^{-1}}[\sum_{j\in T}\|\boldsymbol{b}_j\|_2^k]^{1/k} \leq (p^{-1}\|\mathbf{B}\|^k)^{1/k} = p^{-1/k}\|\mathbf{B}\|.$$

As a result,

$$\mathbb{E}_k\|\mathbf{BR}\| \leq p^{\frac{1}{2}}\{\frac{6\sqrt{\max\{k, 2\log r\}}p^{\frac{1}{2}-\frac{1}{k}}}{1-p} + 1\}\|\mathbf{B}\| \tag{S.15}$$

for $k \in [2, \infty)$. In addition, it is obviously that $\mathbb{E}_k\|\mathbf{BR}\| \leq \|\mathbf{B}\|$. When $p \geq \frac{1}{2}$, we have

$$p^{\frac{1}{2}}\{\frac{6\sqrt{\max\{k, 2\log r\}}p^{\frac{1}{2}-\frac{1}{k}}}{1-p} + 1\} \geq p^{\frac{1}{2}}\{12\sqrt{2\log r}p^{\frac{1}{2}-\frac{1}{k}} + 1\} \geq \frac{1}{\sqrt{2}}\{12\sqrt{\log r} + 1\} > 1$$

and when $p < \frac{1}{2}$ we have

$$p^{\frac{1}{2}}\{\frac{6\sqrt{\max\{k, 2\log r\}}p^{\frac{1}{2}-\frac{1}{k}}}{1-p} + 1\} < p^{\frac{1}{2}}\{12\max\sqrt{\{k, 2\log r\}}p^{\frac{1}{2}-\frac{1}{k}} + 1\}.$$

As a result, we have

$$\mathbb{E}_k\|\mathbf{BR}\| \leq c_1(p, r, k)\|\mathbf{B}\|$$

where $c_1(p, r, k) = \min\{1, p^{\frac{1}{2}}\{12\sqrt{\max\{k, 2\log r\}}p^{\frac{1}{2}-\frac{1}{k}}+1\}\}$. Let $k_0 = \log d \geq 2\log r$. Then by Markov inequality, we have

$$\mathbb{P}\{\|\mathbf{BR}\| \geq p^{\frac{1}{2}}\{\delta^{-1/k_0}c_1(p, r, k_0)/\sqrt{p}\}\|\mathbf{B}\|\} \leq \delta. \tag{S.16}$$

We discuss the (S.16) dependent on the conditions of p.

Case 1: $0 < p < c_3/\log d$ for all $d > 0$ and some fixed constant $c_3 > 0$. Then $\delta^{-1/q_0} = e^3$ is a constant. In addition, $\sqrt{k_0}p^{\frac{1}{2}-\frac{1}{k_0}} \leq \sqrt{c_3}\{c_3/\log d\}^{-1/\log d} < c_4$ for some constant c_4 since $\lim_{x\to\infty} x^{1/x} = 1$ is bounded. As a result, $c_1(p, r, k_0)/\sqrt{p} \leq 12c_4 + 1$ is also bounded.

Case 2: $p \geq c_5$ for all $d > 0$ and some fixed constant $0 < c_5 < 1$. Then let $c_6 = 1/\sqrt{c_5}$ and we have

$$\mathbb{P}\{\|\mathbf{BR}\| > p^{\frac{1}{2}}c_6\|\mathbf{B}\|\} \leq \delta \tag{S.17}$$

since $\|\mathbf{BR}\| \leq \|\mathbf{B}\|$ almost surely.

Case 3: $p = g(d)/\log d$ for some function $g(d) > 0$ which satisfies $\lim_{d\to\infty} g(d) = \infty$ and $\lim_{d\to\infty} g(d)/\log d = 0$. We still have $\delta^{-1/k_0} = e^3$. In addition, $c_1(p, r, k_0)/\sqrt{p} \le 12\sqrt{k_0}p^{\frac{1}{2}-\frac{1}{k_0}} + 1 \le 12\sqrt{g(d)}(\frac{\log d}{g(d)})^{1/\log d} + 1 \le c_7\sqrt{g(d)} = c_7\sqrt{p\log d}$ for some constant c_7 since $(\log d/g(d))^{1/\log d}$ is bounded.

Based on Case 1, 2 and 3, letting $C = \max\{e^3(12c_4+1), c_6, e^3c_7\}$, we will get the result. ■

Let $c_1 = \lim_{N\to\infty} p_0$ and $c_2 = \lim_{N\to\infty} p_0 \log N$. Define

$$f(p_0, N) = \mathbb{1}(c_1 > 0 \text{ or } c_2 = 0) + \{1 - \mathbb{1}(c_1 > 0 \text{ or } c_2 = 0)\}\sqrt{p_0 \log N}. \tag{S.18}$$

Then we have the following proposition.

Proposition S.9 *Based on the definition of* (S.7)*, under the assumption that p_0 is bounded away from 1, e.g., $\lim_{N_0\to\infty} p_0 < 1$, directly apply Lemma S.8, we will get*

$$\begin{aligned}
\|\widetilde{\mathbf{A}}_1\mathbf{Q}_A - \mathbf{A}_1\| &\lesssim \sqrt{1-p_0}\|\widetilde{\mathbf{A}}\mathbf{Q}_A - \mathbf{A}\|; \quad \|\widetilde{\mathbf{A}}_2\mathbf{Q}_A - \mathbf{A}_2\| \lesssim \sqrt{p_0}f(p_0,N)\|\widetilde{\mathbf{A}}\mathbf{Q}_A - \mathbf{A}\|;\\
\|\widetilde{\mathbf{B}}_2\mathbf{Q}_B - \mathbf{B}_2\| &\lesssim \sqrt{1-p_0}\|\widetilde{\mathbf{B}}\mathbf{Q}_B - \mathbf{B}\|; \quad \|\widetilde{\mathbf{B}}_1\mathbf{Q}_A - \mathbf{B}_1\| \lesssim \sqrt{p_0}f(p_0,N)\|\widetilde{\mathbf{B}}\mathbf{Q}_B - \mathbf{B}\|;\\
\|\widetilde{\mathbf{A}}_1\| &\lesssim \sqrt{1-p_0}\|\widetilde{\mathbf{A}}\|; \quad \|\mathbf{A}_1\| \lesssim \sqrt{1-p_0}\|\mathbf{A}\|;\\
\|\mathbf{A}_2\| &\lesssim \sqrt{p_0}f(p_0,N)\|\mathbf{A}\|; \quad \|\widetilde{\mathbf{B}}_1\| \lesssim \sqrt{p_0}f(p_0,N)\|\widetilde{\mathbf{B}}\|;\\
\|\widetilde{\mathbf{B}}_2\| &\lesssim \sqrt{1-p_0}\|\widetilde{\mathbf{B}}\|; \quad \|\mathbf{B}_2\| \lesssim \sqrt{1-p_0}\|\mathbf{B}\|;
\end{aligned} \tag{S.19}$$

with probability $1 - 10/N^3$.

S.4.3 Orthogonal Procrustes Problem

Lemma S.10 (Orthogonal Procrustes problem) *Based on the definition of* (S.7)*, the condition of* (S.19)*, Assumption 2, $\lambda_1(\mathbf{W}_l^*) \le 3p_0r\mu_0/2\lambda_{\max}$, $\lambda_r(\mathbf{W}_l^*) \ge p_l\lambda_{\min}/4$, $\|\widetilde{\mathbf{E}}_l\| \ll \lambda_r(\mathbf{W}_l^*), l = s, k$, and $N_{sk} \ge 64r\mu_0\tau(\log r + \log 2N^3)$, we have*

$$\|\mathbf{Q}_B^\top\widetilde{\mathbf{O}}\mathbf{Q}_A - \mathbf{O}\| \lesssim \frac{f(p_0,N)^2r\mu_0\tau}{p_0\lambda_{\min}}\{\|\widetilde{\mathbf{E}}_s\| + \|\widetilde{\mathbf{E}}_k\|\} \tag{S.20}$$

with probability $1 - 2/N^3$.

Proof First,

$$\begin{aligned}
&\|\mathbf{A}_2^\top\mathbf{B}_1 - \mathbf{Q}_A^\top\widetilde{\mathbf{A}}_2^\top\widetilde{\mathbf{B}}_1\mathbf{Q}_B\| \le \|\mathbf{A}_2\|\|\widetilde{\mathbf{B}}_1\mathbf{Q}_B - \mathbf{B}_1\| + \|\widetilde{\mathbf{B}}_1\|\|\widetilde{\mathbf{A}}_2\mathbf{Q}_A - \mathbf{A}_2\|\\
&\le p_0f(p_0,N)^2\{\|\mathbf{A}\|\|\widetilde{\mathbf{B}}\mathbf{Q}_B - \mathbf{B}\| + \|\widetilde{\mathbf{B}}\|\|\widetilde{\mathbf{A}}\mathbf{Q}_A - \mathbf{A}\|\}\\
&\le 2p_0f(p_0,N)^2\{\|\mathbf{A}\|\|\widetilde{\mathbf{B}}\mathbf{Q}_B - \mathbf{B}\| + \|\mathbf{B}\|\|\widetilde{\mathbf{A}}\mathbf{Q}_A - \mathbf{A}\|\}\\
&\lesssim p_0f(p_0,N)^2\{\sqrt{\frac{r\mu_0\tau\lambda_1(\mathbf{W}_s^*)}{\lambda_r(\mathbf{W}_k^*)}}\|\widetilde{\mathbf{E}}_k\| + \sqrt{\frac{r\mu_0\tau\lambda_1(\mathbf{W}_k^*)}{\lambda_r(\mathbf{W}_s^*)}}\|\widetilde{\mathbf{E}}_s\|\}\\
&\le p_0f(p_0,N)^2r\mu_0\tau\{\|\widetilde{\mathbf{E}}_s\| + \|\widetilde{\mathbf{E}}_k\|\}
\end{aligned}$$

where the second inequality comes from (S.19), the third inequality comes from $\|\widetilde{\mathbf{B}}\| \leq \sqrt{\|\mathbf{W}_k^*\| + \|\widetilde{\mathbf{E}}_k\|} \leq \sqrt{2\|\mathbf{W}_k^*\|} \leq 2\|\mathbf{B}\|$ and the last inequality comes from Lemma S.7. In addition, since

$$\begin{aligned}\sigma_{r-1}(\mathbf{A}_2^\top \mathbf{B}_1) \geq \sigma_r(\mathbf{A}_2^\top \mathbf{B}_1) &= \sigma_r((\mathbf{V}_{s2}^*)^\top (\mathbf{\Sigma}_s^*)^{1/2}(\mathbf{\Sigma}_k^*)^{1/2}\mathbf{V}_{k1}^*) \\ &\geq \sigma_{\min}(\mathbf{V}_{s2}^*)\sqrt{\lambda_r(\mathbf{\Sigma}_s^*)\lambda_r(\mathbf{\Sigma}_k^*)}\sigma_{\min}(\mathbf{V}_{k1}^*)\end{aligned}$$

and again by $p_0 \geq C\sqrt{\mu_0 r\tau \log N/N}$, we will have $p_0 \geq \sqrt{64r\mu_0\tau(\log r + \log 2N^3)/N}$. Then by Lemma S.2, $\sigma_{\min}(\mathbf{V}_{s2}^*)\sigma_{\min}(\mathbf{V}_{k1}^*) \geq p_0/6$ holds with probability $1 - 2/N^3$. Then

$$\sigma_{r-1}(\mathbf{A}_2^\top \mathbf{B}_1) \geq \sigma_r(\mathbf{A}_2^\top \mathbf{B}_1) \geq p_0^2\lambda_{\min}/24.$$

So we can apply Lemma 23 of Ma et al. (2018) to get

$$\begin{aligned}\|\mathbf{Q}_A^\top \widetilde{\mathbf{O}}\mathbf{Q}_B - \mathbf{O}\| &\leq \frac{\|\mathbf{A}_2^\top \mathbf{B}_1 - \mathbf{Q}_A^\top \widetilde{\mathbf{A}}_2^\top \widetilde{\mathbf{B}}_1 \mathbf{Q}_B\|}{\sigma_{r-1}(\mathbf{A}_2^\top \mathbf{B}_1) + \sigma_r(\mathbf{A}_2^\top \mathbf{B}_1)} \\ &\leq \frac{p_0 f(p_0, N)^2 r\mu_0\tau}{2p_0^2\lambda_{\min}/24}\{\|\widetilde{\mathbf{E}}_s\| + \|\widetilde{\mathbf{E}}_k\|\} \lesssim \frac{f(p_0, N)^2 r\mu_0\tau}{p_0\lambda_{\min}}\{\|\widetilde{\mathbf{E}}_s\| + \|\widetilde{\mathbf{E}}_k\|\}.\end{aligned} \tag{S.21}$$

■

S.5. Discussion about Dependent Sampling

In this section, we explore the relaxation of the independence model (1) to accommodate a wider range of scenarios. To do this, we introduce a dependent model

$$\{p_s\}_{s=1}^m \sim P^m, \tag{S.22}$$

where the probabilities $\{p_s\}_{s=1}^m$ are drawn from an m-dimension distribution P^m over the sample space $(C\sqrt{\mu_0 r \log N/N}, 1)^m$ with a sufficiently large constant C. There are no other restrictions on the distribution P^m, so $\{p_s\}_{s=1}^m$ can be highly dependent. Conditioning on $\{p_s\}_{s=1}^m$, we assume that

$$\mathbb{I}(w \in \mathcal{V}_s \mid p_s) \text{ for } w \in \mathcal{V} \text{ and } s \in [m] \text{ independently, and } \mathbb{P}(w \in \mathcal{V}_s \mid p_s) = p_s. \tag{S.23}$$

This model allows the dependence for the emergence of corpora which is decided by P^m. Specifically, we have

$$\mathbb{P}(w_1 \in \mathcal{V}_s, w_2 \in \mathcal{V}_s) \neq \mathbb{P}(w_1 \in \mathcal{V}_s)\mathbb{P}(w_2 \in \mathcal{V}_s)$$

when $\mathrm{Var}(p_s) > 0$ and

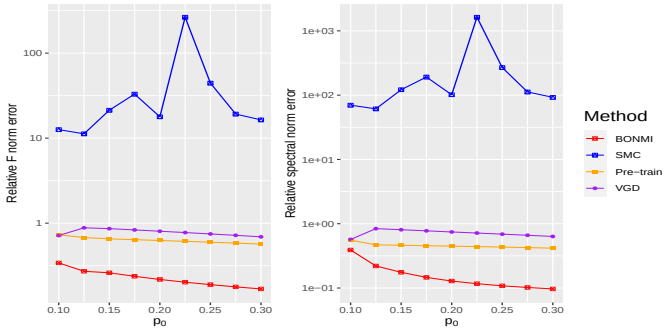
$$\mathbb{P}(w_1 \in \mathcal{V}_s, w_2 \in \mathcal{V}_k) \neq \mathbb{P}(w_1 \in \mathcal{V}_s)\mathbb{P}(w_2 \in \mathcal{V}_k)$$

when $\mathrm{Cov}(p_s, p_k) \neq 0$ for $s \neq k$. As a result, if w_1 belongs to $\mathcal{V}_s$, it may influence the occurrence probability of other codes for $\mathcal{V}_s$ and $\mathcal{V}_k$ for $k \neq s$.

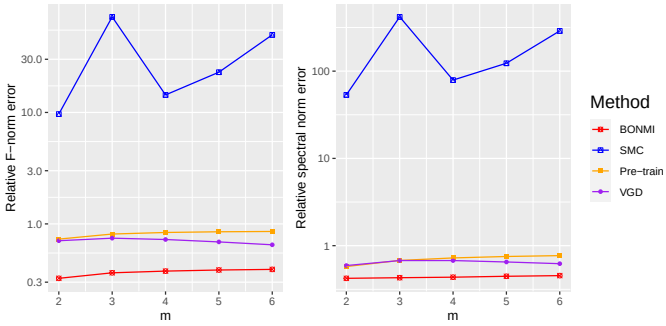
Despite these changes, our theorems still hold under the model defined by (S.22) and (S.23). Specifically, conditioning on $\{p_s\}_{s=1}^m$, we can still guarantee the overlapping matrices are of rank r and apply Lemma S.8. Following the same analysis, we can prove the same bounds as Theorems 7 and 12.

S.6. Additional Simulation Results

To validate the data-driven method for choosing r in Section 5, we rerun all simulations in Section 4 while only replacing the true rank with the estimated one for all methods. We observe a similar pattern to the results in Section 4.3, and BONMI still performs the best. The results are presented in Figures 4.



(a) setting (i): fix $m = 2$ and range p_0 from 0.1 to 0.3.



(b) setting (ii): fix $p_0 = 0.1$ and range m from 2 to 6.

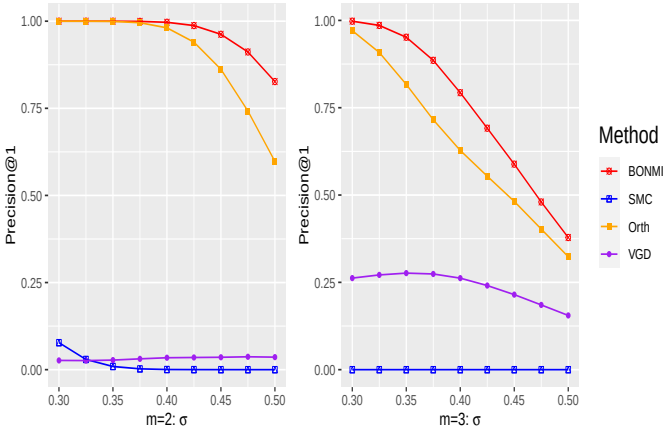


Figure 4: Simulation results of settings (i) and (ii). The relative estimation errors of $\mathbf{W}_0^*$ are presented. setting (iii): fix $p_0 = 0.1$ and range σ from 0.3 to 0.5.