A Technical proofs

A.1 Proof of Lemma 1.

Proof. Since $\widehat{\Theta} = (\widehat{\Theta}_1, ..., \widehat{\Theta}_K)$ are the global minimum of (5), we have

$$0 \geq f(\widehat{\Theta}) - f(\Theta^*) \geq \langle \nabla f(\Theta^*), \widehat{\Theta} - \Theta^* \rangle + \frac{\mu_\Theta}{2} \big\| \widehat{\Theta} - \Theta^* \big\|_F^2.$$

We then have

$$\begin{aligned} \big\| \widehat{\Theta} - \Theta^* \big\|_F^2 &\leq -\frac{2}{\mu_\Theta} \langle \nabla f(\Theta^*), \widehat{\Theta} - \Theta^* \rangle \\ &\leq \frac{2}{\mu_\Theta} \big\| \nabla f(\Theta^*) \big\|_F \cdot \big\| \widehat{\Theta} - \Theta^* \big\|_F. \end{aligned}$$

and hence

$$\big\| \widehat{\Theta} - \Theta^* \big\|_F \leq \frac{2}{\mu_\Theta} \big\| \nabla f(\Theta^*) \big\|_F.$$

□

A.2 Proof of Theorem 2.

Proof. We apply the non-convex optimization result in [37]. Since the initialization condition and (RSC/RSS) are satisfied for our problem according to Lemma 1, we apply Lemma 3 in [37] and obtain

$$\begin{aligned} d^2\Big(B^{(t+1)}, B^*\Big) \leq & \xi^2 \bigg[\Big(1 - \eta \cdot \frac{2}{5} \mu_{\min} \sigma_{\max}\Big) \cdot d^2\Big(B^{(t)}, B^*\Big) \\ & + \eta \cdot \frac{L_\Theta + \mu_\Theta}{L_\Theta \cdot \mu_\Theta} \cdot e^2_{\text{stat},\Theta} \bigg], \end{aligned} \tag{11}$$

where $\xi^2 = 1 + \frac{2}{\sqrt{c-1}}$ and $\sigma_{\max} = \max_k \|\Theta_k^*\|_2$. Define the contraction value

$$\beta = \xi^2 \Big(1 - \eta \cdot \frac{2}{5} \mu_{\min} \sigma_{\max}\Big) < 1,$$

we can iteratively apply (11) for each $t = 1, 2, ..., T$ and obtain

$$d^2\Big(B^{(T)}, B^*\Big) \leq \beta^T d^2\Big(B^{(0)}, B^*\Big) + \frac{\xi^2 \eta}{1 - \beta} \cdot \frac{L_\Theta + \mu_\Theta}{L_\Theta \cdot \mu_\Theta} \cdot e^2_{\text{stat},\Theta}$$

which shows linear convergence up to statistical error. □

A.3 Proof of Lemma 5.

Proof. Since $\widetilde{X}$ is the best rank K approximation for $\overline{X}$, and $\overline{X}^*$ is also rank K, we have $\|\widetilde{X} - \overline{X}\|_F \leq \|\overline{X}^* - \overline{X}\|_F$ and hence

$$\begin{aligned} \|\widetilde{X} - \overline{X}^*\|_F &\leq \|\widetilde{X} - \overline{X}\|_F + \|\overline{X}^* - \overline{X}\|_F \\ &\leq 2\|\overline{X}^* - \overline{X}\|_F = 2\|\overline{E}\|_F. \end{aligned} \tag{12}$$

By definition we have

$$\begin{aligned} \overline{X}^* &= \sum_{k=1}^K \Big(\frac{1}{n} \sum_{i=1}^n m_{ik}^* \Big) \Theta_k^* = B_1^* A^* B_2^{*\top} \\ &= Q_1 R_1 A^* R_2^\top Q_2^\top = Q_1 (A_{\text{diag}} + A_{\text{off}}) Q_2^\top. \end{aligned}$$

Plugging back into (12) we obtain

$$\big\| \widetilde{X} - Q_1 (A_{\text{diag}} + A_{\text{off}}) Q_2^\top \big\|_F \leq 2\|\overline{E}\|_F,$$

and hence

$$\begin{aligned} & \Big\| \sum_{k=1}^K \widetilde{\sigma}_k \widetilde{u}_k \widetilde{v}_k^\top - Q_1 A_{\text{diag}} Q_2^\top \Big\|_F \\ & \leq 2\|\overline{E}\|_F + \|Q_1 A_{\text{off}} Q_2^\top\|_F \\ & \leq 2\|\overline{E}\|_F + \rho_0. \end{aligned} \tag{13}$$

Under mild conditions we have that $\|\overline{E}\|_F \propto n^{-1/2}$ and therefore can be arbitrarily small with large enough n. Moreover, the left hand side of (13) is the difference of two singular value decompositions. According to the matrix perturbation theory, for each k we have (up to permutation)

$$\Big\| \widetilde{\sigma}_k \widetilde{u}_k \widetilde{v}_k^\top - q_{1,k} \cdot a_{\text{diag},k} \cdot q_{2,k}^\top \Big\|_F \leq 2C\rho_0,$$

and hence

$$\Big\| \widetilde{\sigma}_k \widetilde{u}_k \widetilde{v}_k^\top - \frac{1}{n} \sum_{i=1}^n m_{ik}^* \Theta_k^* \Big\|_F \leq 2\widetilde{C}\rho_0.$$

Finally we obtain

$$\begin{aligned} & \Big\| K \cdot \widetilde{\sigma}_k \widetilde{u}_k \widetilde{v}_k^\top - \Theta_k^* \Big\|_F = K \cdot \Big\| \widetilde{\sigma}_k \widetilde{u}_k \widetilde{v}_k^\top - \frac{1}{K} \Theta_k^* \Big\|_F \\ & \leq K \cdot \Big(2\widetilde{C}\rho_0 + \Big| \frac{1}{n} \sum_{i=1}^n m_{ik}^* - \frac{1}{K} \Big| \cdot \|\Theta_k^*\|_F \Big) \\ & \leq 2\widetilde{C} K \rho_0 + (\eta - 1)\sigma_{\max}. \end{aligned}$$

□

A.4 Proof of Theorem 6

We analyze the two estimation step in Algorithm 2.

Update on B_1 and B_2. The update algorithm on B_1 and B_2 is the same with known M. Besides the statistical error defined in (7), we now have an additional error term due to the error in M. Recall that $d^2(M, M^*) = \frac{1}{n} \sum_{i=1}^n \sum_{k_0=1}^K (m_{ik_0} - m_{ik_0}^*)^2$, Lemma 7 quantifies the effect of one estimation step on B.

Lemma 7. *Suppose the conditions in Theorem 2 hold and suppose condition (DC) and (OC) hold, we have*

$$d^2\big(B^{[t]}, B^*\big) \leq C_1 \cdot e^2_{\text{stat},\Theta} + \beta_1 \cdot d^2\big(M^{[t]}, M^*\big),$$

for some constant C_1 and β_1.

Update on M. Lemma 8 quantifies the effect of one estimation step on M.

Lemma 8. *Suppose the condition (TC) holds, we have*

$$d^2\big(M^{[t]}, M^*\big) \le C_2 \cdot e^2_{\mathrm{stat},M} + \beta_2 \cdot d^2\big(B^{[t]}, B^*\big),$$

for some constant C_2 and β_2.

Denote $\beta_0 = \min\{\beta_1, \beta_2\}$, as long as the signal $\sigma_{\max}$ is small and the noise E_i is small enough we can guarantee that $\beta_0 < 1$. Combine Lemma 7 and 8 we complete the proof.

A.5 Proof of Lemma 7.

Proof. The analysis is exactly the same with the case where M is known except that the statistical error is different. Specifically, for each k we have

$$\begin{aligned}
&\nabla_{\Theta_k} f(\Theta^*, M)\\
&= -\frac{1}{n}\sum_{i=1}^n \Big(X_i - \sum_{k_0=1}^K m_{ik_0}\Theta^*_{k_0}\Big)\cdot m_{ik}\\
&= -\frac{1}{n}\sum_{i=1}^n \Big(E_i + \sum_{k_0=1}^K (m^*_{ik_0} - m_{ik_0})\Theta^*_{k_0}\Big)\cdot m_{ik}\\
&= \underbrace{-\frac{1}{n}\sum_{i=1}^n E_i m^*_{ik}}_{R_1} + \underbrace{\frac{1}{n}\sum_{i=1}^n E_i(m^*_{ik} - m_{ik})}_{R_2}\\
&\quad + \underbrace{\frac{1}{n}\sum_{i=1}^n\sum_{k_0=1}^K (m_{ik_0} - m^*_{ik_0})\Theta^*_{k_0}\cdot m_{ik}}_{R_3}
\end{aligned}$$

The first term R_1 is just the usual statistical error term on Θ. For term R_2, denote $e_0 = \frac{1}{n}\sum_{i=1}^n \|E_i\|_F^2$, we have

$$\begin{aligned}
\|R_2\|_F^2 &\le \frac{1}{n^2}\Big(\sum_{i=1}^n \|E_i\|_F^2\Big)\cdot\sum_{i=1}^n (m_{ik} - m^*_{ik})^2\\
&\le \frac{e_0}{n}\sum_{i=1}^n (m_{ik} - m^*_{ik})^2.
\end{aligned}$$

For term R_3, we have

$$\begin{aligned}
&\|R_3\|_F^2\\
&\le \frac{1}{n^2}\Big\|\sum_{i=1}^n\sum_{k_0=1}^K (m_{ik_0} - m^*_{ik_0})\Theta^*_{k_0}\cdot m_{ik}\Big\|_F^2\\
&\le \frac{1}{n^2}\Big(\sum_{i=1}^n\sum_{k_0=1}^K (m_{ik_0} - m^*_{ik_0})^2\Big)\Big(\sum_{i=1}^n\sum_{k_0=1}^K \|\Theta^*_{k_0}\|_F^2\cdot m^2_{ik}\Big)\\
&\le \frac{K\sigma^2_{\max}}{n^2}\Big(\sum_{i=1}^n m^2_{ik}\Big)\cdot\Big(\sum_{i=1}^n\sum_{k_0=1}^K (m_{ik_0} - m^*_{ik_0})^2\Big).
\end{aligned}$$

Taking summation over all k, the first term R_1 gives the statistical error as before, the terms R_2 and R_3 gives

$$\sum_{k=1}^K \|R_2\|_F^2 + \|R_3\|_F^2 \le \frac{e_0 + K\sigma^2_{\max}}{n}\Big(\sum_{i=1}^n\sum_{k=1}^K (m_{ik} - m^*_{ik})^2\Big).$$

□

A.6 Proof of Lemma 8.

Proof. The estimation on M is separable with each m_i. Denote the objective function on observation i as

$$f_i(\Theta, m_i) = \Big\|X_i - \sum_{k=1}^K m_{ik}\cdot\Theta_k\Big\|_F^2. \tag{14}$$

According to condition (DC), the objective function (14) is μ_M-strongly convex in m_i. Similar to the proof of Lemma 1, we obtain

$$\begin{aligned}
\sum_{k=1}^K (m_{ik} - m^*_{ik})^2 &\le \frac{4}{\mu_M^2}\big\|\nabla_{m_i} f_i(\Theta, m^*_i)\big\|_F^2\\
&= \frac{4}{\mu_M^2}\sum_{k=1}^K \Big[\nabla_{m_{ik}} f_i(\Theta, m^*_i)\Big]^2.
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
\nabla_{m_{ik}} f_i(\Theta, m^*_i) &= -\Big\langle X_i - \sum_{k_0=1}^K m^*_{ik_0}\cdot\Theta_{k_0}, \Theta_k\Big\rangle\\
&= \underbrace{-\Big\langle E_i, \Theta^*_k\Big\rangle}_{T_1} + \underbrace{\Big\langle E_i, (\Theta^*_k - \Theta_k)\Big\rangle}_{T_2}\\
&\quad + \underbrace{\Big\langle \sum_{k_0=1}^K m^*_{ik_0}(\Theta_{k_0} - \Theta^*_{k_0}), \Theta_k\Big\rangle}_{T_3}.
\end{aligned}$$

The first term T_1 is just the usual statistical error term on M. For term T_2, we have

$$\begin{aligned}
\sum_{i=1}^n\sum_{k=1}^K (T_2)^2 &\le \sum_{i=1}^n\sum_{k=1}^K \|E_i\|_F^2\cdot\|\Theta^*_k - \Theta_k\|_F^2\\
&= \Big(\sum_{i=1}^n \|E_i\|_F^2\Big)\cdot\Big(\sum_{k=1}^K \|\Theta^*_k - \Theta_k\|_F^2\Big).
\end{aligned} \tag{15}$$

For term T_3 we have

$$\begin{aligned}\sum_{k=1}^{K}(T_3)^2 &\leq \bigg(\sum_{k=1}^{K}\|\Theta_k\|_F^2\bigg)\Big\|\sum_{k_0=1}^{K} m^*_{ik_0}(\Theta_{k_0}-\Theta^*_{k_0})\Big\|_F^2 \\ &\leq K\sigma^2_{\max}\bigg(\sum_{k=1}^{K}\|\Theta^*_k-\Theta_k\|_F^2\bigg)\bigg(\sum_{k=1}^{K}{m^*_{ik}}^2\bigg) \\ &\leq K\sigma^2_{\max}\bigg(\sum_{k=1}^{K}\|\Theta^*_k-\Theta_k\|_F^2\bigg). \end{aligned} \tag{16}$$

Moreover, we have

$$\begin{aligned}\|\Theta^*_k-\Theta_k\|_F &= \|b_k^{1*}b_k^{2*\top}-b_k^1b_k^{2\top}\|_F \\ &\leq \|b_k^{1*}\|_2\|b_k^{2*}-b_k^2\|_2+\|b_k^2\|_2\|b_k^{1*}-b_k^1\|_2 \\ &\leq 2\sigma_{\max}\big(\|b_k^{2*}-b_k^2\|_2+\|b_k^{1*}-b_k^1\|_2\big),\end{aligned}$$

and hence

$$\begin{aligned}&\sum_{k=1}^{K}\|\Theta^*_k-\Theta_k\|_F^2 \\ &\leq 4\sigma^2_{\max}\sum_{k=1}^{K}\big(\|b_k^{2*}-b_k^2\|_2+\|b_k^{1*}-b_k^1\|_2\big)^2 \\ &\leq 8\sigma^2_{\max}d^2(B,B^*).\end{aligned}$$

Combine (15) and (16), taking summation over i, we obtain

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}(T_2)^2+(T_3)^2 \\ &\leq \Big(e_0+K\sigma^2_{\max}\Big)\cdot\bigg(\sum_{k=1}^{K}\|\Theta^*_k-\Theta_k\|_F^2\bigg) \\ &\leq 8\sigma^2_{\max}\Big(e_0+K\sigma^2_{\max}\Big)\cdot d^2(B,B^*).\end{aligned}$$

□

B Detailed rationale on initialization and condition (OC) for jointly learning

Initialization. Define $\overline{X}$, $\overline{X}^*$, $\overline{E}$ as the sample mean of X_i, X_i^*, E_i, respectively. We have

$$\begin{aligned}\overline{X} &= \frac{1}{n}\sum_{i=1}^{n}X_i^*+E_i = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}m^*_{ik}\Theta^*_k+\frac{1}{n}\sum_{i=1}^{n}E_i \\ &= \sum_{k=1}^{K}\Big(\frac{1}{n}\sum_{i=1}^{n}m^*_{ik}\Big)\Theta^*_k+\overline{E} = \overline{X}^*+\overline{E}.\end{aligned}$$

We then do the rank-K SVD on $\overline{X}$ and obtain $[\widetilde{U},\widetilde{S},\widetilde{V}]$ = rank-K SVD of $\overline{X}$. We denote $\widetilde{X} = \widetilde{U}\widetilde{S}\widetilde{V}^\top = \sum_{k=1}^{K}\widetilde{\sigma}_k\widetilde{u}_k\widetilde{v}_k^\top$. It is well known that $\widetilde{X}$ is the best rank K approximation for $\overline{X}$. We propose the following initialization method

$$\Theta_k^{(0)} = K\cdot\widetilde{\sigma}_k\widetilde{u}_k\widetilde{v}_k^\top.$$

To see why this initialization works, we first build intuition for the easiest case, where $E_i=0$ for each i, $\frac{1}{n}\sum_{i=1}^{n}m^*_{ik}=\frac{1}{K}$ for each k, and the columns of B_1^* and B_2^* are orthogonal. In this case it is easy to see that $\overline{X}=\overline{X}^*=\sum_{k=1}^{K}\frac{1}{K}\Theta^*_k$. Note that this expression in a singular value decomposition of $\overline{X}^*$ since we have $\Theta^*_k=b_k^{1*}b_k^{2*\top}$ and the columns $\{b_k^{1*}\}_{k=1}^K$ and columns $\{b_k^{2*}\}_{k=1}^K$ are orthogonal. Now that $\overline{X}$ is exactly rank K, the best rank K approximation would be itself, i.e., $\overline{X}=\widetilde{X}=\sum_{k=1}^{K}\widetilde{\sigma}_k\widetilde{u}_k\widetilde{v}_k^\top$. By the uniqueness of singular value decomposition, as long as the singular values are distinct, we have (up to permutation) $\frac{1}{K}\Theta^*_k=\widetilde{\sigma}_k\widetilde{u}_k\widetilde{v}_k^\top$ and therefore $\Theta^*_k=K\cdot\widetilde{\sigma}_k\widetilde{u}_k\widetilde{v}_k^\top$. This is exactly what we want to estimate.

With this intuition in mind, we relax the restrictions we have and impose the following condition.

Orthogonal Condition (OC). Let $B_1^*=Q_1R_1$ and $B_2^*=Q_2R_2$ be the QR decomposition of B_1^* and B_2^*, respectively. Denote A^* as a diagonal matrix with diagonal elements $\frac{1}{n}\sum_{i=1}^{n}m^*_{ik}$. Denote $R_1A^*R_2^\top = A_{\text{diag}}+A_{\text{off}}$ where A_{diag} captures the diagonal elements and A_{off} captures the off-diagonal elements. We require that $\|A_{\text{off}}\|_F\leq\rho_0$ for some constant ρ_0. Moreover, we require that $\frac{1}{n}\sum_{i=1}^{n}m^*_{ik}\leq\eta/K$ for some η.

This condition requires that B_1^* and B_2^* are not too far away from orthogonal matrix, so that when doing the QR rotation, the off diagonal values of R_1 and R_2 are not too large. The condition $\frac{1}{n}\sum_{i=1}^{n}m^*_{ik}\leq\eta/K$ is trivially satisfied with $\eta=K$. However, in general η is usually a constant that does not scale with K, meaning that the topic distribution among the n observations is more like evenly distributed than several topics dominate.

Finally note that the condition (OC) is for this specific initialization method only. Since we are doing singular value decomposition, we end up with orthogonal vectors so we require that B_1^* and B_2^* are not too far away from orthogonal; since we do not know the value $\frac{1}{n}\sum_{i=1}^{n}m^*_{ik}$ and use $1/K$ to approximate, we require that topics are not far away from evenly distributed so that this approximation is reasonable. In practice we can also initialize using other methods, for example we can do alternating gradient descent on B_1, B_2 and M based on the objective function (10). This method also works reasonably well in practice.

C Detailed node-topic matrices for citation dataset

The detailed two node-topic matrices for citation dataset is given in Table 4 and Table 5.

D Additional figures

Figure 3 and Figure 4 shows the comparison result for binary observation in Section 6, with known topics and unknown topics, respectively.

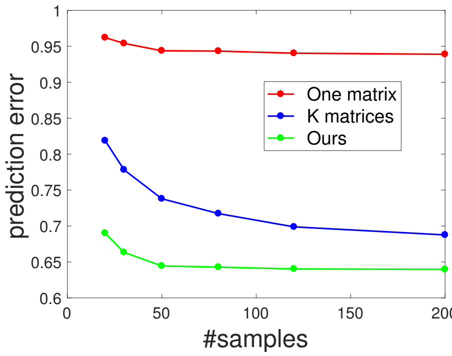

Figure 3: Prediction error for binary observation, with known topics

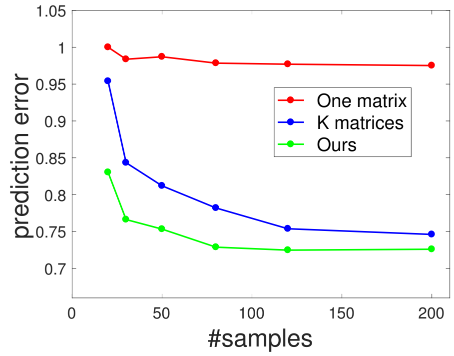

Figure 4: Prediction error for binary observation, with unknown topics

Table 4: The influence matrix B_1 for citation dataset

	black hole energy chains	quantum model field theory	gauge theory field effective	algebra space group structure	states space noncommutative boundary	string theory supergravity supersymmetric
Christopher Pope		0.359		0.468		0.318
Arkady Tseytlin	0.223	0.565		0.25		
Emilio Elizalde		0.109				
Cumrun Vafa			0.85	0.623	0.679	0.513
Edward Witten		0.204		0.795	0.678	1.87
Ashok Das		0.155	0.115	1.07		
Sergei Odintsov						
Sergio Ferrara	0.297	0.889	0.345	0.457	0.453	0.249
Renata Kallosh	0.44	0.512		0.326	0.382	
Mirjam Cvetic		0.339	0.173	0.338		
Burt A. Ovrut	0.265	0.191	0.127	0.328	0.133	
Ergin Sezgin		0.35		0.286		
Ian I. Kogan			0.193			
Gregory Moore		0.323	0.91	0.325	0.536	
I. Antoniadis	0.443	0.485		0.545	0.898	0.342
Mirjam Cvetic	0.152	0.691		0.228	0.187	
Andrew Strominger	0.207	0.374	0.467	1.15		
Barton Zwiebach	0.16			0.222	0.383	0.236
P.K. Townsend		0.629		0.349		0.1
Robert C. Myers		0.439		0.28		
E. Bergshoeff		0.357		0.371		
Amihay Hanany		0.193		0.327		1.09
Ashoke Sen	0.319			0.523		0.571

Table 5: The receptivity matrix B_2 for citation dataset

	black hole energy chains	quantum model field theory	gauge theory field effective	algebra space group structure	states space noncommutative boundary	string theory supergravity supersymmetric
Christopher Pope	0.477	0.794		0.59		
Arkady Tseytlin	0.704	1.16	0.312	0.487		0.119
Emilio Elizalde						
Cumrun Vafa	0.309		0.428	0.844	0.203	0.693
Edward Witten	0.352		0.554	0.585	0.213	0.567
Ashok Das	0.494	0.339		0.172		
Sergei Odintsov		0.472				
Sergio Ferrara	0.423	0.59	0.664	0.776		
Renata Kallosh	0.123	0.625	0.638	0.484		0.347
Mirjam Cvetic	0.47	0.731		0.309		
Burt A. Ovrut	0.314	0.217	0.72	0.409		0.137
Ergin Sezgin		0.108	0.161	0.358		
Ian I. Kogan	0.357	0.382				0.546
Gregory Moore	0.375	0.178	0.721	0.69	0.455	0.517
I. Antoniadis	0.461		0.699	0.532		0.189
Mirjam Cvetic	0.409	1.11	0.173	0.361		
Andrew Strominger		0.718	0.248	0.196	0.133	
Barton Zwiebach			0.308	0.204		0.356
P.K. Townsend	0.337	0.225	0.245	0.522		
Robert C. Myers	0.364	0.956		0.545		0.139
E. Bergshoeff	0.487	0.459	0.174	0.619		
Amihay Hanany	0.282		0.237	0.575		0.732
Ashoke Sen		0.214	0.18	0.37		