Supplement to “A Goodness-of-fit Test on the Number of Biclusters in a Relational Data Matrix”

Chihiro Watanabe · Taiji Suzuki

Appendix A Proof of $\boldsymbol{\lambda_1 \leq \tilde{\lambda}_1 + O_p\left(m^{\frac{1}{3}-\epsilon}\right)}$ for some $\boldsymbol{\epsilon > 0}$ in the null case

Proof By assumption, the background can be divided to H disjoint submatrices, whose row and column sizes are equal to or larger than $n_{\min}$ and $p_{\min}$, respectively. Let $X^{(k)} \in \mathbb{R}^{|I_k| \times |J_k|}$ be a submatrix of matrix $X \in \mathbb{R}^{n \times p}$ corresponding to the row and column indices (I_k, J_k) of the kth bicluster. To distinguish the indices of the biclusters from those of the background submatrices, let $X^{(K+h)} \in \mathbb{R}^{|I_{K+h}| \times |J_{K+h}|}$ be a submatrix of matrix $X \in \mathbb{R}^{n \times p}$ corresponding to the row and column indices (I_{K+h}, J_{K+h}) of the hth background submatrix. We define a bicluster-wise constant matrix $Q^{(k)}$ for each kth bicluster $(k = 1, \ldots, K)$,

$$
\begin{aligned}
Q^{(k)} &\equiv Z^{(k)} - \frac{\tilde{s}_k}{s_k} \tilde{Z}^{(k)} = \frac{1}{s_k} \left(\tilde{P}^{(k)} - P^{(k)} \right) \\
&= \left(\frac{1}{|\mathcal{I}_k|} \sum_{(i,j) \in \mathcal{I}_k} Z_{ij} \right) \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{|I_k| \times |J_k|},
\end{aligned} \tag{1}
$$

and a submatrix-wise constant matrix $Q^{(K+h)}$ for each hth background submatrix $(h = 1, \ldots, H)$,

$$
Q^{(K+h)} \equiv Z^{(K+h)} - \frac{\tilde{s}_0}{s_0} \tilde{Z}^{(K+h)} = \frac{1}{s_0} \left(\tilde{P}^{(K+h)} - P^{(K+h)} \right)
$$

C. Watanabe (corresponding author)
Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
E-mail: watanabe-chihiro763@g.ecc.u-tokyo.ac.jp

T. Suzuki
Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo, Japan

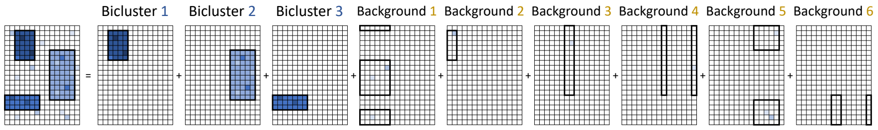

Fig. 1 Decomposition of the noise matrix Z to the K biclusters $\{\underline{Z}^{(k)}\}$, $k=1,\dots,K$ and the H background submatrices $\{\underline{Z}^{(K+h)}\}$, $h=1,\dots,H$. In the case of this figure, $K=3$ and $H=6$.

$$= \left(\frac{1}{|\mathcal{I}_0|} \sum_{(i,j)\in\mathcal{I}_0} Z_{ij} \right) \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{|I_{K+h}|\times|J_{K+h}|}. \tag{2}$$

Based on these matrices, let $\underline{Z}^{(k)}$, $\underline{\tilde{Z}}^{(k)}$, and $\underline{Q}^{(k)}$, respectively, be $n\times p$ matrices whose entries in the kth bicluster (i.e., $\{(i,j): i\in I_k, j\in J_k\}$) are $Z^{(k)}$, $\tilde{Z}^{(k)}$ and $Q^{(k)}$ and whose all the other entries are zero. Similarly, let $\underline{Z}^{(K+h)}$, $\underline{\tilde{Z}}^{(K+h)}$, and $\underline{Q}^{(K+h)}$, respectively, be $n\times p$ matrices whose entries in the hth background submatrix (i.e., $\{(i,j): i\in I_{K+h}, j\in J_{K+h}\}$) are $Z^{(K+h)}$, $\tilde{Z}^{(K+h)}$ and $Q^{(K+h)}$ and whose all the other entries are zero. Figure 1 shows an example of $\{\underline{Z}^{(k)}\}$, where $k=1,\dots,K+H$. Finally, we define matrix Q by $Q\equiv\sum_{k=1}^{K+H}\underline{Q}^{(k)}$.

Let $\boldsymbol{v}_1$ be the normalized eigenvector of matrix $Z^\top Z$ corresponding to the maximum eigenvalue λ_1. Since the largest singular value $\sqrt{\tilde{\lambda}_1}$ of matrix $\tilde{Z}$ is equal to the operator norm of $\tilde{Z}$, we have

$$\begin{aligned}
\tilde{\lambda}_1 &= \left(\sup_{\boldsymbol{u}} \frac{\|\tilde{Z}\boldsymbol{u}\|}{\|\boldsymbol{u}\|}\right)^2 \geq \left(\frac{\|\tilde{Z}\boldsymbol{v}_1\|}{\|\boldsymbol{v}_1\|}\right)^2 = \|\tilde{Z}\boldsymbol{v}_1\|^2 = \left\|\sum_{k=1}^{K+H} \underline{\tilde{Z}}^{(k)}\boldsymbol{v}_1\right\|^2 \\
&= \left\| \left[\sum_{k=1}^{K}\left(\frac{s_k}{\tilde{s}_k}\right)(\underline{Z}^{(k)}-\underline{Q}^{(k)})\boldsymbol{v}_1\right] + \left[\sum_{h=1}^{H}\left(\frac{s_0}{\tilde{s}_0}\right)(\underline{Z}^{(K+h)}-\underline{Q}^{(K+h)})\boldsymbol{v}_1\right] \right\|^2 \\
&= \left\| \left[\sum_{k=1}^{K+H}(\underline{Z}^{(k)}-\underline{Q}^{(k)})\boldsymbol{v}_1\right] + \left[\sum_{k=1}^{K+H}\left(\frac{s_k}{\tilde{s}_k}-1\right)(\underline{Z}^{(k)}-\underline{Q}^{(k)})\boldsymbol{v}_1\right] \right\|^2 \\
&\geq \left[\left\|\sum_{k=1}^{K+H}(\underline{Z}^{(k)}-\underline{Q}^{(k)})\boldsymbol{v}_1\right\| - \left\|\sum_{k=1}^{K+H}\left(1-\frac{s_k}{\tilde{s}_k}\right)(\underline{Z}^{(k)}-\underline{Q}^{(k)})\boldsymbol{v}_1\right\| \right]^2 \\
&= \left[\|(Z-Q)\boldsymbol{v}_1\| - \left\|\sum_{k=1}^{K+H}\left(1-\frac{s_k}{\tilde{s}_k}\right)(\underline{Z}^{(k)}-\underline{Q}^{(k)})\boldsymbol{v}_1\right\| \right]^2 \\
&\geq \|(Z-Q)\boldsymbol{v}_1\|^2 - 2\|(Z-Q)\boldsymbol{v}_1\| \left\|\sum_{k=1}^{K+H}\left(1-\frac{s_k}{\tilde{s}_k}\right)(\underline{Z}^{(k)}-\underline{Q}^{(k)})\boldsymbol{v}_1\right\| \\
&\geq \|(Z-Q)\boldsymbol{v}_1\|^2 - 2\|(Z-Q)\boldsymbol{v}_1\| \left[\sum_{k=1}^{K+H}\left|1-\frac{s_k}{\tilde{s}_k}\right| \|(\underline{Z}^{(k)}-\underline{Q}^{(k)})\boldsymbol{v}_1\|\right]
\end{aligned}$$

$$
\begin{aligned}
&\geq \|(Z-Q)\boldsymbol{v}_1\|^2 - 2\|(Z-Q)\boldsymbol{v}_1\| \left[\sum_{k=1}^{K+H} \left|1-\frac{s_k}{\tilde{s}_k}\right| \left(\|\underline{Z}^{(k)}\boldsymbol{v}_1\| + \|\underline{Q}^{(k)}\boldsymbol{v}_1\|\right)\right] \\
&\geq \|(Z-Q)\boldsymbol{v}_1\|^2 - 2\|(Z-Q)\boldsymbol{v}_1\| \left[\sum_{k=1}^{K+H} \left|1-\frac{s_k}{\tilde{s}_k}\right| \left(\sqrt{\lambda_1^{(k)}} + \|\underline{Q}^{(k)}\boldsymbol{v}_1\|\right)\right] \\
&\geq \lambda_1 - 2\sqrt{\lambda_1}\|Q\boldsymbol{v}_1\| \\
&- 2(\sqrt{\lambda_1} + \|Q\boldsymbol{v}_1\|) \left[\sum_{k=1}^{K+H} \left|1-\frac{s_k}{\tilde{s}_k}\right| \left(\sqrt{\lambda_1^{(k)}} + \|\underline{Q}^{(k)}\boldsymbol{v}_1\|\right)\right]. \qquad (3)
\end{aligned}
$$

where $\lambda_1^{(k)}$ is the maximum eigenvalue of matrix $(\underline{Z}^{(k)})^\top \underline{Z}^{(k)}$ (which is equal to that of matrix $(Z^{(k)})^\top Z^{(k)}$). From the third line in (3), we used the notation that $s_{K+1} = \cdots = s_{K+H} = s_0$ for simplicity.

Subsequently, we show the probabilistic orders of $\|\underline{Q}^{(k)}\boldsymbol{v}_1\|$ and $\|Q\boldsymbol{v}_1\|$. The non-zero entries in matrix $(\underline{Q}^{(k)})^\top \underline{Q}^{(k)}$ is only located in a submatrix $\{(i,j) : i \in J_k, j \in J_k\}$, and all of their values are $|I_k|\eta_k^2$ by (1) and (2), where

$$\eta_k \equiv \frac{1}{|\mathcal{I}_k|} \sum_{(i,j)\in\mathcal{I}_k} Z_{ij} = O_p\left(\frac{1}{\sqrt{|\mathcal{I}_k|}}\right). \qquad (4)$$

Therefore, we have

$$(\underline{Q}^{(k)})^\top \underline{Q}^{(k)} \boldsymbol{v}_1 = |I_k||J_k|\eta_k^2 (\boldsymbol{v}_1^\top \boldsymbol{u}^{(k)})\boldsymbol{u}^{(k)}, \qquad (5)$$

where $\boldsymbol{u}^{(k)} \in \mathbb{R}^p$ is a vector whose entries are defined by $\boldsymbol{u}_j^{(k)} = \frac{1}{\sqrt{|J_k|}}$ if $j \in J_k$ and $\boldsymbol{u}_j^{(k)} = 0$ otherwise. Note that this vector satisfies $\|\boldsymbol{u}^{(k)}\| = 1$. From (5), we have

$$\|\underline{Q}^{(k)}\boldsymbol{v}_1\| = \sqrt{\boldsymbol{v}_1^\top (\underline{Q}^{(k)})^\top \underline{Q}^{(k)} \boldsymbol{v}_1} = \sqrt{|I_k||J_k|\eta_k^2 (\boldsymbol{v}_1^\top \boldsymbol{u}^{(k)})^2}. \qquad (6)$$

To upper bound the right side of (6), we refer to the following important property of each jth eigenvector $\boldsymbol{v}_j$ of matrix $Z^\top Z$, which has been proven in (Bloemendal et al. 2016).

Theorem 1 (Delocalization property of an eigenvector of a sample covariance matrix (Bloemendal et al. 2016)) *Under the assumptions in Sect. 2, from Theorem 2.17 in (Bloemendal et al. 2016), a normalized eigenvector $\boldsymbol{v}_j$ of matrix $Z^\top Z$ (i.e., $\|\boldsymbol{v}_j\| = 1$) has a delocalization property, that is, for all $\tilde{d} \in \mathbb{N}$, for any deterministic vectors $\{\boldsymbol{w}^{(i)}\}$ that satisfies $\|\boldsymbol{w}^{(i)}\| = 1$ for $i = 1, \ldots, m^{\tilde{d}}$, for all $\epsilon > 0$,*

$$\max_{i\in 1,\ldots,m^{\tilde{d}}} \max_{j=1,\ldots,p} |\boldsymbol{v}_j^\top \boldsymbol{w}^{(i)}| = O_p\left(m^{-\frac{1}{2}+\epsilon}\right). \qquad (7)$$

Based on the above delocalization property of vector $\boldsymbol{v}_1$ and (6), we have

$$\|\underline{Q}^{(k)}\boldsymbol{v}_1\| = \sqrt{|\mathcal{I}_k| O_p\left(\frac{1}{|\mathcal{I}_k|}\right) O_p\left(m^{-1+2\epsilon}\right)} = O_p\left(m^{-\frac{1}{2}+\epsilon}\right), \text{ for all } \epsilon > 0. \tag{8}$$

As for $\|Q\boldsymbol{v}_1\|$, we can derive its upper bound by

$$\begin{aligned} \|Q\boldsymbol{v}_1\| &= \left\| \sum_{k=1}^{K+H} \underline{Q}^{(k)}\boldsymbol{v}_1 \right\| \leq \sum_{k=1}^{K+H} \|\underline{Q}^{(k)}\boldsymbol{v}_1\| = \sum_{k=1}^{K+H} \sqrt{|\mathcal{I}_k|\eta_k^2(\boldsymbol{v}_1^\top \boldsymbol{u}^{(k)})^2} \\ &= \sum_{k=1}^{K+H} \sqrt{|\mathcal{I}_k| O_p\left(\frac{1}{|\mathcal{I}_k|}\right)} |\boldsymbol{v}_1^\top \boldsymbol{u}^{(k)}| = (K+H) O_p\left(m^{-\frac{1}{2}+\epsilon}\right). \end{aligned} \tag{9}$$

Here, we used the fact that (7) holds from (Bloemendal et al. 2016).

From Lemma B1 in Appendix B, $|\tilde{s}_k - s_k| = O_p\left(\frac{1}{\sqrt{|\mathcal{I}_k|}}\right)$ holds, which results in

$$\left|1 - \frac{s_k}{\tilde{s}_k}\right| = O_p\left(\frac{1}{\sqrt{|\mathcal{I}_k|}}\right). \tag{10}$$

By substituting (8), (9), (10), and the fact that $\sqrt{\lambda_1^{(k)}} = O_p\left(|\mathcal{I}_k|^{\frac{1}{4}}\right)$ from (Pillai and Yin 2014), into (3), we obtain

$$\begin{aligned} \tilde{\lambda}_1 &\geq \lambda_1 - 2(K+H)O_p\left(m^{\epsilon}\right) - 2\left[O_p\left(m^{\frac{1}{2}}\right) + (K+H)O_p\left(m^{-\frac{1}{2}+\epsilon}\right)\right] \\ &\left\{\sum_{k=1}^{K+H} O_p\left(|\mathcal{I}_k|^{-\frac{1}{2}}\right)\left[O_p\left(|\mathcal{I}_k|^{\frac{1}{4}}\right) + O_p\left(m^{-\frac{1}{2}+\epsilon}\right)\right]\right\} \\ &= \lambda_1 - 2(K+H)O_p\left(m^{\epsilon}\right) - 2\left[O_p\left(m^{\frac{1}{2}}\right) + (K+H)O_p\left(m^{-\frac{1}{2}+\epsilon}\right)\right] \\ &\left\{\sum_{k=1}^{K+H} \left[O_p\left(|\mathcal{I}_k|^{-\frac{1}{4}}\right) + O_p\left(|\mathcal{I}_k|^{-\frac{1}{2}} m^{-\frac{1}{2}+\epsilon}\right)\right]\right\}. \end{aligned} \tag{11}$$

By taking $\epsilon < \frac{1}{2}$, the lower bound in (11) can be simplified as follows:

$$\begin{aligned} \tilde{\lambda}_1 &\geq \lambda_1 - 2(K+H)O_p\left(m^{\epsilon}\right) - 2\left[O_p\left(m^{\frac{1}{2}}\right) + (K+H)O_p\left(m^{-\frac{1}{2}+\epsilon}\right)\right] \\ &\left\{\sum_{k=1}^{K+H} \left[O_p\left(|\mathcal{I}_k|^{-\frac{1}{4}}\right) + O_p\left(|\mathcal{I}_k|^{-\frac{1}{2}} |\mathcal{I}_k|^{-\frac{1}{4}+\frac{1}{2}\epsilon}\right)\right]\right\} \\ &= \lambda_1 - 2(K+H)O_p\left(m^{\epsilon}\right) - 2\left[O_p\left(m^{\frac{1}{2}}\right) + (K+H)O_p\left(m^{-\frac{1}{2}+\epsilon}\right)\right] \\ &\left[\sum_{k=1}^{K+H} O_p\left(|\mathcal{I}_k|^{-\frac{1}{4}}\right)\right] \end{aligned}$$

$$\begin{aligned}
&\geq \lambda_1 - 2(K+H)O_p\left(m^{\epsilon}\right) - 2\left[O_p\left(m^{\frac{1}{2}}\right) + (K+H)O_p\left(m^{-\frac{1}{2}+\epsilon}\right)\right] \\
&(K+H)O_p\left[\left(\min_{k=1,\ldots,K+H}|\mathcal{I}_k|\right)^{-\frac{1}{4}}\right] \\
&= \lambda_1 - 2(K+H)O_p\left(m^{\epsilon}\right)\left\{O_p(1) + O_p\left[m^{\frac{1}{2}}\left(\min_{k=1,\ldots,K+H}|\mathcal{I}_k|\right)^{-\frac{1}{4}}\right]\right\} \\
&= \lambda_1 - 2(K+H)O_p\left(m^{\epsilon}\right)O_p\left[m^{\frac{1}{2}}\left(\min_{k=1,\ldots,K+H}|\mathcal{I}_k|\right)^{-\frac{1}{4}}\right] \\
&= \lambda_1 - 2(K+H)O_p\left[m^{\frac{1}{2}+\epsilon}\left(\min_{k=1,\ldots,K+H}|\mathcal{I}_k|\right)^{-\frac{1}{4}}\right]. \qquad (12)
\end{aligned}$$

From the assumption (iv) that $(K+H)\left(\min_{k=1,\ldots,K+H}|\mathcal{I}_k|\right)^{-\frac{1}{4}} = O\left(m^{-\frac{1}{6}-\epsilon_1}\right)$ for some $\epsilon_1 > 0$, by taking $\epsilon < \epsilon_1$, we have

$$\lambda_1 \leq \tilde{\lambda}_1 + O_p\left(m^{\frac{1}{3}-(\epsilon_1-\epsilon)}\right), \qquad (13)$$

which concludes the proof. □

Appendix B Proof of $|\tilde{s}_k - s_k| = O_p\left(\frac{1}{\sqrt{|\mathcal{I}_k|}}\right)$.

Let $A^{(k)}$, $P^{(k)}$, and $\tilde{P}^{(k)}$, respectively, be the kth **null** bicluster ($k = 1, \ldots, K$) or background ($k = 0$) of matrices A, P, and $\tilde{P}$.

Lemma B1 *Under the assumption that* $\mathbb{E}[Z_{ij}^4] < \infty$*,*

$$|\tilde{s}_k - s_k| = O_p\left(\frac{1}{\sqrt{|\mathcal{I}_k|}}\right), \qquad (14)$$

where $\mathcal{I}_k \equiv \{(i,j) : g_{ij} = k\}$ *(i.e., the set of entries in the* k*th group).*

Proof By definition, we have

$$\begin{aligned}
\tilde{s}_k^2 &\equiv \frac{1}{|\mathcal{I}_k|}\sum_{(i,j)\in\mathcal{I}_k}\left(A_{ij}^{(k)} - \tilde{b}_k\right)^2 = \frac{1}{|\mathcal{I}_k|}\sum_{(i,j)\in\mathcal{I}_k}\left[\left(A_{ij}^{(k)}\right)^2 - \tilde{b}_k^2\right] \\
&= \frac{1}{|\mathcal{I}_k|}\sum_{(i,j)\in\mathcal{I}_k}\left[\left(A_{ij}^{(k)}\right)^2 - \tilde{b}_k^2\right] - \frac{1}{|\mathcal{I}_k|}2b_k\sum_{(i,j)\in\mathcal{I}_k}\left(A_{ij}^{(k)} - \tilde{b}_k\right) \\
&= \frac{1}{|\mathcal{I}_k|}\sum_{(i,j)\in\mathcal{I}_k}\left(A_{ij}^{(k)} - b_k\right)^2 - \left(b_k - \tilde{b}_k\right)^2, \qquad (15)
\end{aligned}$$

where $\tilde{b}_k \equiv \frac{1}{|\mathcal{I}_k|}\sum_{(i,j)\in\mathcal{I}_k} A_{ij}^{(k)}$.

From (15), we have

$$\tilde{s}_k^2 - s_k^2 = \frac{1}{|\mathcal{I}_k|} \sum_{(i,j)\in\mathcal{I}_k} \left(A_{ij}^{(k)} - b_k\right)^2 - s_k^2 - \left(b_k - \tilde{b}_k\right)^2$$
$$= \frac{1}{|\mathcal{I}_k|} \sum_{(i,j)\in\mathcal{I}_k} Y_{ij}^{(k)} - \left(b_k - \tilde{b}_k\right)^2, \quad (16)$$

where used the notation that $Y_{ij}^{(k)} \equiv \left(A_{ij}^{(k)} - b_k\right)^2 - s_k^2$. Based on the assumption that the entries $\left(A_{ij}^{(k)}\right)_{(i,j)\in\mathcal{I}_k}$ are generated in the i.i.d. sense in each kth group, the random variables $\left(Y_{ij}^{(k)}\right)_{(i,j)\in\mathcal{I}_k}$ are also independent, and their expectations and variances satisfy

$$\mathbb{E}\left[Y_{ij}^{(k)}\right] = \mathbb{E}\left[\left(A_{ij}^{(k)} - b_k\right)^2\right] - s_k^2 = 0,$$
$$\mathbb{V}\left[Y_{ij}^{(k)}\right] = \mathbb{E}\left[\left(Y_{ij}^{(k)}\right)^2\right] = s_k^4\left(\mathbb{E}\left[\left(Z_{ij}^{(k)}\right)^4\right] - 1\right), \quad (17)$$

which results in

$$\mathbb{E}\left[\frac{1}{|\mathcal{I}_k|} \sum_{(i,j)\in\mathcal{I}_k} Y_{ij}^{(k)}\right] = 0,$$
$$\mathbb{V}\left[\frac{1}{|\mathcal{I}_k|} \sum_{(i,j)\in\mathcal{I}_k} Y_{ij}^{(k)}\right] = \frac{1}{|\mathcal{I}_k|} s_k^4\left(\mathbb{E}\left[\left(Z_{ij}^{(k)}\right)^4\right] - 1\right). \quad (18)$$

From (18) and Chebyshev's inequality, for all $t > 0$, we have

$$\Pr\left[\left|\frac{1}{|\mathcal{I}_k|} \sum_{(i,j)\in\mathcal{I}_k} Y_{ij}^{(k)}\right| \geq t\sqrt{\frac{1}{|\mathcal{I}_k|} s_k^4\left(\mathbb{E}\left[\left(Z_{ij}^{(k)}\right)^4\right] - 1\right)}\right] \leq \frac{1}{t^2}, \quad (19)$$

which results in

$$\left|\frac{1}{|\mathcal{I}_k|} \sum_{(i,j)\in\mathcal{I}_k} Y_{ij}^{(k)}\right| = O_p\left(\frac{1}{\sqrt{|\mathcal{I}_k|}}\right). \quad (20)$$

from the assumption of $\mathbb{E}\left[\left(Z_{ij}^{(k)}\right)^4\right] < \infty$.

As for the second term in (16), we have

$$\left(b_k - \tilde{b}_k\right)^2 = \left[\frac{1}{|\mathcal{I}_k|} \sum_{(i,j)\in\mathcal{I}_k} \left(P_{ij}^{(k)} - A_{ij}^{(k)}\right)\right]^2 = \frac{s_k^2}{|\mathcal{I}_k|^2}\left(\sum_{(i,j)\in\mathcal{I}_k} Z_{ij}^{(k)}\right)^2. \quad (21)$$

From (21), we have

$$\mathbb{E}\left[\left(b_k - \tilde{b}_k\right)^2\right] = \frac{s_k^2}{|\mathcal{I}_k|^2}\mathbb{V}\left[\sum_{(i,j)\in\mathcal{I}_k} Z_{ij}^{(k)}\right] = \frac{s_k^2}{|\mathcal{I}_k|}, \tag{22}$$

since $Z_{ij}^{(k)}$ has a unit variance.

From (22) and Markov's inequality, we have

$$\begin{aligned} &\forall t > 0,\ \Pr\left[\left(b_k - \tilde{b}_k\right)^2 \geq t\right] \leq \frac{s_k^2}{|\mathcal{I}_k|}\frac{1}{t} \\ \iff &\forall t' > 0,\ \Pr\left[\left(b_k - \tilde{b}_k\right)^2 \geq \frac{s_k^2}{|\mathcal{I}_k|}t'\right] \leq \frac{1}{t'}, \end{aligned} \tag{23}$$

which results in

$$\left(b_k - \tilde{b}_k\right)^2 = O_p\left(\frac{1}{|\mathcal{I}_k|}\right). \tag{24}$$

Using (20), (24), and (16), we finally obtain

$$|\tilde{s}_k^2 - s_k^2| \leq |\frac{1}{|\mathcal{I}_k|}\sum_{(i,j)\in\mathcal{I}_k} Y_{ij}^{(k)}| + |\left(b_k - \tilde{b}_k\right)^2| = O_p\left(\frac{1}{\sqrt{|\mathcal{I}_k|}}\right), \tag{25}$$

which results in

$$|\tilde{s}_k - s_k| = \frac{|\tilde{s}_k^2 - s_k^2|}{|\tilde{s}_k + s_k|}. \tag{26}$$

From (25), we see that $\tilde{s}_k$ converges in probability to s_k, and thus $\frac{1}{|\tilde{s}_k + s_k|}$ converges in probability to $\frac{1}{2s_k} > 0$. Therefore, we have

$$|\tilde{s}_k - s_k| = O_p\left(\frac{1}{\sqrt{|\mathcal{I}_k|}}\right), \tag{27}$$

which concludes the proof. □

Appendix C Proof of $\tilde{\lambda}_1 \leq \lambda_1 + O_p\left(m^{\frac{1}{3}-\epsilon}\right)$ for some $\epsilon > 0$ in the null case

Proof Let $\tilde{\boldsymbol{v}}_1^{(k)} \in \mathbb{R}^{|J_k|}$ be a subvector of $\tilde{\boldsymbol{v}}_1$ corresponding to the columns of the kth submatrix in observed matrix A, and let $\tau_k \equiv \frac{s_k}{\tilde{s}_k}$. In (10), we have already shown that $|1 - \tau_k| = O_p\left(\frac{1}{\sqrt{|\mathcal{I}_k|}}\right)$. The maximum eigenvalue $\tilde{\lambda}_1$ of matrix $\tilde{Z}^\top\tilde{Z}$ can be upper bounded as follows:

$$\tilde{\lambda}_1 = \|\tilde{Z}\tilde{\boldsymbol{v}}_1\|^2 = \left\|\sum_{k=1}^{K+H} \tau_k\left(\underline{Z}^{(k)} - \underline{Q}^{(k)}\right)\tilde{\boldsymbol{v}}_1\right\|^2 \quad (\because (1))$$

$$
\begin{aligned}
&= \left\| \left\{ Z + \sum_{k=1}^{K+H} \left[(\tau_k - 1)\underline{Z}^{(k)} - \tau_k \underline{Q}^{(k)} \right] \right\} \tilde{\boldsymbol{v}}_1 \right\|^2 \\
&= \|Z\tilde{\boldsymbol{v}}_1\|^2 + 2\tilde{\boldsymbol{v}}_1^\top Z^\top \sum_{k=1}^{K+H} \left[(\tau_k - 1)\underline{Z}^{(k)} - \tau_k \underline{Q}^{(k)} \right] \tilde{\boldsymbol{v}}_1 \\
&+ \left\| \sum_{k=1}^{K+H} \left[(\tau_k - 1)\underline{Z}^{(k)} - \tau_k \underline{Q}^{(k)} \right] \tilde{\boldsymbol{v}}_1 \right\|^2 \\
&\leq \|Z\tilde{\boldsymbol{v}}_1\|^2 + 2\sqrt{\lambda_1} \sum_{k=1}^{K+H} |\tau_k - 1| \|\underline{Z}^{(k)}\tilde{\boldsymbol{v}}_1\| - 2\tilde{\boldsymbol{v}}_1^\top Z^\top \sum_{k=1}^{K+H} \tau_k \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1 \\
&+ \left\| \sum_{k=1}^{K+H} \left[(\tau_k - 1)\underline{Z}^{(k)} - \tau_k \underline{Q}^{(k)} \right] \tilde{\boldsymbol{v}}_1 \right\|^2 \\
&= \|Z\tilde{\boldsymbol{v}}_1\|^2 + 2\sqrt{\lambda_1} \sum_{k=1}^{K+H} |\tau_k - 1| \|Z^{(k)}\tilde{\boldsymbol{v}}_1^{(k)}\| - 2\tilde{\boldsymbol{v}}_1^\top Z^\top \sum_{k=1}^{K+H} \tau_k \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1 \\
&+ \left\| \sum_{k=1}^{K+H} \left[(\tau_k - 1)\underline{Z}^{(k)} - \tau_k \underline{Q}^{(k)} \right] \tilde{\boldsymbol{v}}_1 \right\|^2 \\
&\leq \|Z\tilde{\boldsymbol{v}}_1\|^2 + 2\sqrt{\lambda_1} \sum_{k=1}^{K+H} |\tau_k - 1| \sqrt{\lambda_1^{(k)}} - 2\tilde{\boldsymbol{v}}_1^\top Z^\top \sum_{k=1}^{K+H} \tau_k \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1 \\
&+ \left\| \sum_{k=1}^{K+H} \left[(\tau_k - 1)\underline{Z}^{(k)} - \tau_k \underline{Q}^{(k)} \right] \tilde{\boldsymbol{v}}_1 \right\|^2 \\
&= \|Z\tilde{\boldsymbol{v}}_1\|^2 + 2O_p\left(m^{\frac{1}{2}}\right) \sum_{k=1}^{K+H} O_p\left(|\mathcal{I}_k|^{-\frac{1}{4}}\right) - 2\tilde{\boldsymbol{v}}_1^\top Z^\top \sum_{k=1}^{K+H} \tau_k \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1 \\
&+ \left\| \sum_{k=1}^{K+H} \left[(\tau_k - 1)\underline{Z}^{(k)} - \tau_k \underline{Q}^{(k)} \right] \tilde{\boldsymbol{v}}_1 \right\|^2 \\
&= \|Z\tilde{\boldsymbol{v}}_1\|^2 + 2(K+H)O_p\left[m^{\frac{1}{2}} \left(\min_{k=1,\ldots,K+H} |\mathcal{I}_k| \right)^{-\frac{1}{4}} \right] \\
&- 2\tilde{\boldsymbol{v}}_1^\top Z^\top \sum_{k=1}^{K+H} \tau_k \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1 + \left\| \sum_{k=1}^{K+H} \left[(\tau_k - 1)\underline{Z}^{(k)} - \tau_k \underline{Q}^{(k)} \right] \tilde{\boldsymbol{v}}_1 \right\|^2 \\
&\leq \|Z\tilde{\boldsymbol{v}}_1\|^2 + 2(K+H)O_p\left[m^{\frac{1}{2}} \left(\min_{k=1,\ldots,K+H} |\mathcal{I}_k| \right)^{-\frac{1}{4}} \right] \\
&- 2\tilde{\boldsymbol{v}}_1^\top Z^\top \sum_{k=1}^{K+H} \tau_k \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1 + \left[\sum_{k=1}^{K+H} |\tau_k - 1| \|\underline{Z}^{(k)}\tilde{\boldsymbol{v}}_1\| + \sum_{k=1}^{K+H} \tau_k \|\underline{Q}^{(k)}\tilde{\boldsymbol{v}}_1\| \right]^2
\end{aligned}
$$

$$
\begin{aligned}
&\le \|Z\tilde{\boldsymbol{v}}_1\|^2 + 2(K+H) O_p\left[m^{\frac{1}{2}}\left(\min_{k=1,\dots,K+H}|\mathcal{I}_k|\right)^{-\frac{1}{4}}\right] \\
&- 2\sum_{k=1}^{K+H}\tau_k \tilde{\boldsymbol{v}}_1^\top Z^\top \underline{Q}^{(k)}\tilde{\boldsymbol{v}}_1 + \left[\sum_{k=1}^{K+H} O_p\left(|\mathcal{I}_k|^{-\frac{1}{4}}\right) + \sum_{k=1}^{K+H}\tau_k\|\underline{Q}^{(k)}\tilde{\boldsymbol{v}}_1\|\right]^2 \\
&\left(\because \|\underline{Z}^{(k)}\tilde{\boldsymbol{v}}_1\| \le \sqrt{\lambda_1^{(k)}} = O_p\left(|\mathcal{I}_k|^{\frac{1}{4}}\right)\right) \\
&= \|Z\tilde{\boldsymbol{v}}_1\|^2 + O_p\left[m^{\frac{1}{2}} C^{(K,H)}\right] - 2\sum_{k=1}^{K+H}\tau_k \tilde{\boldsymbol{v}}_1^\top Z^\top \underline{Q}^{(k)}\tilde{\boldsymbol{v}}_1 \\
&+ \left[O_p\left(C^{(K,H)}\right) + \sum_{k=1}^{K+H}\tau_k\|\underline{Q}^{(k)}\tilde{\boldsymbol{v}}_1\|\right]^2, \qquad (28)
\end{aligned}
$$

where we denote $C^{(K,H)} \equiv (K+H)\left(\min_{k=1,\dots,K+H}|\mathcal{I}_k|\right)^{-\frac{1}{4}}$.

The eigenvectors $\{\boldsymbol{v}_j\}$ of symmetric matrix $Z^\top Z$ form an orthonormal system, and thus there exists a unique set of coefficients $\{c_j\}$ such that

$$
\tilde{\boldsymbol{v}}_1 = \sum_{j=1}^{p} c_j \boldsymbol{v}_j = \tilde{\boldsymbol{v}}^{\mathrm{L}} + \tilde{\boldsymbol{v}}^{\mathrm{S}}, \qquad (29)
$$

where

$$
\begin{aligned}
&\tilde{\boldsymbol{v}}^{\mathrm{L}} \equiv \sum_{j=1}^{t} c_j \boldsymbol{v}_j, \quad \tilde{\boldsymbol{v}}^{\mathrm{S}} \equiv \sum_{j=t+1}^{p} c_j \boldsymbol{v}_j, \\
&\lambda_t \ge \lambda_1 - n^d, \quad \lambda_{t+1} < \lambda_1 - n^d, \quad d = \frac{5}{7}. \qquad (30)
\end{aligned}
$$

By substituting (29) into the last term in (28) and from the similar discussion as in (5),

$$
\begin{aligned}
&\|\underline{Q}^{(k)}\tilde{\boldsymbol{v}}_1\|^2 = \tilde{\boldsymbol{v}}_1^\top (\underline{Q}^{(k)})^\top \underline{Q}^{(k)}\tilde{\boldsymbol{v}}_1 = \sum_{j=1}^{p}\sum_{j'=1}^{p} c_j c_{j'} \boldsymbol{v}_j^\top (\underline{Q}^{(k)})^\top \underline{Q}^{(k)} \boldsymbol{v}_{j'} \\
&= \sum_{j=1}^{p}\sum_{j'=1}^{p} c_j c_{j'} |I_k||J_k|\eta_k^2 (\boldsymbol{v}_j^\top \boldsymbol{u}^{(k)})(\boldsymbol{v}_{j'}^\top \boldsymbol{u}^{(k)}) = |I_k||J_k|\eta_k^2 \left[\sum_{j=1}^{p} c_j (\boldsymbol{v}_j^\top \boldsymbol{u}^{(k)})\right]^2 \\
&\le |I_k||J_k|\eta_k^2 \left[\sqrt{\sum_{j=1}^{p} c_j^2}\sqrt{\sum_{j=1}^{p}(\boldsymbol{v}_j^\top \boldsymbol{u}^{(k)})^2}\right]^2 = |I_k||J_k|\eta_k^2 \|\tilde{\boldsymbol{v}}_1\|^2 \left[\sum_{j=1}^{p}(\boldsymbol{v}_j^\top \boldsymbol{u}^{(k)})^2\right] \\
&= |I_k||J_k|\eta_k^2 \left[\sum_{j=1}^{p}(\boldsymbol{v}_j^\top \boldsymbol{u}^{(k)})^2\right] \le |I_k||J_k|\eta_k^2\, p \max_{j=1,\dots,p}(\boldsymbol{v}_j^\top \boldsymbol{u}^{(k)})^2 \\
&= |\mathcal{I}_k| O_p\left(|\mathcal{I}_k|^{-1}\right) p\, O_p\left(m^{-1+2\epsilon}\right) = O_p\left(m^{2\epsilon}\right). \qquad (31)
\end{aligned}
$$

Here, we used the fact that (7) holds from (Bloemendal et al. 2016).

By combining (28) and (31),

$$\begin{aligned}
&\tilde{\lambda}_1 \leq \|Z\tilde{\boldsymbol{v}}_1\|^2 - 2\sum_{k=1}^{K+H} \tau_k \tilde{\boldsymbol{v}}_1^\top Z^\top \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1 + O_p\left[m^{\frac{1}{2}} C^{(K,H)}\right] \\
&+ \left[O_p\left(C^{(K,H)}\right) + \sum_{k=1}^{K+H} \tau_k O_p\left(m^\epsilon\right)\right]^2 \\
&= \|Z\tilde{\boldsymbol{v}}_1\|^2 - 2\sum_{k=1}^{K+H} \tau_k \tilde{\boldsymbol{v}}_1^\top Z^\top \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1 + O_p\left[m^{\frac{1}{2}} C^{(K,H)}\right] \\
&+ \left[O_p\left(C^{(K,H)}\right) + \sum_{k=1}^{K+H} \left(1 + O_p\left(|\mathcal{I}_k|^{-\frac{1}{2}}\right)\right) O_p\left(m^\epsilon\right)\right]^2 \\
&= \|Z\tilde{\boldsymbol{v}}_1\|^2 - 2\sum_{k=1}^{K+H} \tau_k \tilde{\boldsymbol{v}}_1^\top Z^\top \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1 + O_p\left[m^{\frac{1}{2}} C^{(K,H)}\right] \\
&+ \left[O_p\left(C^{(K,H)}\right) + (K+H) O_p\left(m^\epsilon\right)\right]^2 \\
&= \|Z\tilde{\boldsymbol{v}}_1\|^2 - 2\sum_{k=1}^{K+H} \tau_k \tilde{\boldsymbol{v}}_1^\top Z^\top \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1 + O_p\left[(K+H) m^{\frac{1}{2}} \left(\min_{k=1,\dots,K+H} |\mathcal{I}_k|\right)^{-\frac{1}{4}}\right] \\
&+ O_p\left[(K+H)^2 m^{2\epsilon}\right]. \qquad (32)
\end{aligned}$$

As for the third term in (32), based on the assumption (iv),

$$O_p\left[(K+H) m^{\frac{1}{2}} \left(\min_{k=1,\dots,K+H} |\mathcal{I}_k|\right)^{-\frac{1}{4}}\right] = O_p\left(m^{\frac{1}{3}-\epsilon_1}\right). \qquad (33)$$

With regard to the fourth term in (32), based on the assumption (iv) that $K + H = O\left(m^{\frac{1}{42}-\epsilon_1}\right)$ for some $\epsilon_1 > 0$, by taking $\epsilon < \epsilon_1$,

$$O_p\left[(K+H)^2 m^{2\epsilon}\right] = O_p\left(m^{\frac{1}{21}-2(\epsilon_1-\epsilon)}\right). \qquad (34)$$

An upper bound of the first and second terms in (32) is given by

$$\begin{aligned}
&\|Z\tilde{\boldsymbol{v}}_1\|^2 - 2\sum_{k=1}^{K+H} \tau_k \tilde{\boldsymbol{v}}_1^\top Z^\top \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1 \\
&= (\tilde{\boldsymbol{v}}^{\mathrm{L}} + \tilde{\boldsymbol{v}}^{\mathrm{S}})^\top Z^\top Z (\tilde{\boldsymbol{v}}^{\mathrm{L}} + \tilde{\boldsymbol{v}}^{\mathrm{S}}) - 2\sum_{k=1}^{K+H} \tau_k \tilde{\boldsymbol{v}}_1^\top Z^\top \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1 \\
&= \left(\tilde{\boldsymbol{v}}^{\mathrm{L}}\right)^\top Z^\top Z \tilde{\boldsymbol{v}}^{\mathrm{L}} + \left(\tilde{\boldsymbol{v}}^{\mathrm{S}}\right)^\top Z^\top Z \tilde{\boldsymbol{v}}^{\mathrm{S}} - 2\sum_{k=1}^{K+H} \tau_k \tilde{\boldsymbol{v}}_1^\top Z^\top \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1
\end{aligned}$$

$$
\begin{aligned}
&= \left(\sum_{j=1}^{t} c_j \boldsymbol{v}_j\right)^{\top} \left(\sum_{j=1}^{t} c_j Z^{\top} Z \boldsymbol{v}_j\right) + \left(\sum_{j=t+1}^{p} c_j \boldsymbol{v}_j\right)^{\top} \left(\sum_{j=t+1}^{p} c_j Z^{\top} Z \boldsymbol{v}_j\right) \\
&- 2 \sum_{k=1}^{K+H} \tau_k \tilde{\boldsymbol{v}}_1^{\top} Z^{\top} \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1 \\
&= \left(\sum_{j=1}^{t} c_j \boldsymbol{v}_j\right)^{\top} \left(\sum_{j=1}^{t} c_j \lambda_j \boldsymbol{v}_j\right) + \left(\sum_{j=t+1}^{p} c_j \boldsymbol{v}_j\right)^{\top} \left(\sum_{j=t+1}^{p} c_j \lambda_j \boldsymbol{v}_j\right) \\
&- 2 \sum_{k=1}^{K+H} \tau_k \tilde{\boldsymbol{v}}_1^{\top} Z^{\top} \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1 \\
&= \sum_{j=1}^{t} c_j^2 \lambda_j \|\boldsymbol{v}_j\|^2 + \sum_{j=t+1}^{p} c_j^2 \lambda_j \|\boldsymbol{v}_j\|^2 - 2 \sum_{k=1}^{K+H} \tau_k \tilde{\boldsymbol{v}}_1^{\top} Z^{\top} \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1 \\
&= \sum_{j=1}^{t} c_j^2 \lambda_j + \sum_{j=t+1}^{p} c_j^2 \lambda_j - 2 \sum_{k=1}^{K+H} \tau_k \tilde{\boldsymbol{v}}_1^{\top} Z^{\top} \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1 \\
&\leq \lambda_1 \sum_{j=1}^{t} c_j^2 + \lambda_{t+1} \sum_{j=t+1}^{p} c_j^2 - 2 \sum_{k=1}^{K+H} \tau_k \tilde{\boldsymbol{v}}_1^{\top} Z^{\top} \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1 \\
&\leq \lambda_1 \sum_{j=1}^{t} c_j^2 + (\lambda_1 - n^d) \sum_{j=t+1}^{p} c_j^2 - 2 \sum_{k=1}^{K+H} \tau_k \tilde{\boldsymbol{v}}_1^{\top} Z^{\top} \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1 \quad (\because (30)) \\
&= \lambda_1 \sum_{j=1}^{p} c_j^2 - n^d \sum_{j=t+1}^{p} c_j^2 - 2 \sum_{k=1}^{K+H} \tau_k \tilde{\boldsymbol{v}}_1^{\top} Z^{\top} \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1 \\
&= \lambda_1 \|\tilde{\boldsymbol{v}}_1\|^2 - n^d \|\tilde{\boldsymbol{v}}^{\mathrm{S}}\|^2 - 2 \sum_{k=1}^{K+H} \tau_k \tilde{\boldsymbol{v}}_1^{\top} Z^{\top} \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1 \\
&= \lambda_1 - n^d \|\tilde{\boldsymbol{v}}^{\mathrm{S}}\|^2 - 2 \sum_{k=1}^{K+H} \tau_k \tilde{\boldsymbol{v}}_1^{\top} Z^{\top} \underline{Q}^{(k)} \tilde{\boldsymbol{v}}^{\mathrm{L}} - 2 \sum_{k=1}^{K+H} \tau_k \tilde{\boldsymbol{v}}_1^{\top} Z^{\top} \underline{Q}^{(k)} \tilde{\boldsymbol{v}}^{\mathrm{S}}. \qquad (35)
\end{aligned}
$$

Let $\boldsymbol{u}^{(k)} \in \mathbb{R}^p$ be a vector whose entries are defined by $\boldsymbol{u}_j^{(k)} = \frac{1}{\sqrt{|J_k|}}$ if $j \in J_k$ and $\boldsymbol{u}_j^{(k)} = 0$ otherwise. As for the third term in (35), using the fact that $\underline{Q}^{(k)} \boldsymbol{v}_j = \eta_k |J_k| (\boldsymbol{v}_j^{\top} \boldsymbol{u}^{(k)}) \boldsymbol{u}^{(k)}$, for all $\epsilon > 0$,

$$
\begin{aligned}
&- \tilde{\boldsymbol{v}}_1^{\top} Z^{\top} \underline{Q}^{(k)} \tilde{\boldsymbol{v}}^{\mathrm{L}} \leq |\tilde{\boldsymbol{v}}_1^{\top} Z^{\top} \underline{Q}^{(k)} \tilde{\boldsymbol{v}}^{\mathrm{L}}| = \left| \sum_{j=1}^{t} c_j \tilde{\boldsymbol{v}}_1^{\top} Z^{\top} \underline{Q}^{(k)} \boldsymbol{v}_j \right| \\
&= \left| \eta_k |J_k| \tilde{\boldsymbol{v}}_1^{\top} Z^{\top} \boldsymbol{u}^{(k)} \sum_{j=1}^{t} c_j (\boldsymbol{v}_j^{\top} \boldsymbol{u}^{(k)}) \right|
\end{aligned}
$$

$$
\begin{aligned}
&= O_p\left(|\mathcal{I}_k|^{-\frac{1}{2}}\right)|J_k|\left|\sum_{j=1}^{t} c_j(\boldsymbol{v}_j^\top \boldsymbol{u}^{(k)})\right|\left|\tilde{\boldsymbol{v}}_1^\top Z^\top \boldsymbol{u}^{(k)}\right| \\
&= O_p(1)\left|\sum_{j=1}^{t} c_j(\boldsymbol{v}_j^\top \boldsymbol{u}^{(k)})\right|\left|\tilde{\boldsymbol{v}}_1^\top Z^\top \boldsymbol{u}^{(k)}\right| \\
&\le O_p(1)\sqrt{\sum_{j=1}^{t} c_j^2}\sqrt{\sum_{j=1}^{t} |\boldsymbol{v}_j^\top \boldsymbol{u}^{(k)}|^2}\left|\tilde{\boldsymbol{v}}_1^\top Z^\top \boldsymbol{u}^{(k)}\right| \\
&\le O_p(1)\|\tilde{\boldsymbol{v}}_1\|\sqrt{t}\, O_p\left(m^{-\frac{1}{2}+\epsilon}\right)\left|\tilde{\boldsymbol{v}}_1^\top Z^\top \boldsymbol{u}^{(k)}\right| \\
&= \sqrt{t}\, O_p\left(m^{-\frac{1}{2}+\epsilon}\right)\left|\tilde{\boldsymbol{v}}_1^\top Z^\top \boldsymbol{u}^{(k)}\right| \le \sqrt{t}\, O_p\left(m^{-\frac{1}{2}+\epsilon}\right)\|\tilde{\boldsymbol{v}}_1^\top Z^\top\|\|\boldsymbol{u}^{(k)}\| \\
&= \sqrt{t}\, O_p\left(m^{-\frac{1}{2}+\epsilon}\right)\|\tilde{\boldsymbol{v}}_1^\top Z^\top\| \le \sqrt{t}\, O_p\left(m^{-\frac{1}{2}+\epsilon}\right)\sqrt{\lambda_1} \\
&= \sqrt{t}\, O_p\left(m^{-\frac{1}{2}+\epsilon}\right)O_p\left(m^{\frac{1}{2}}\right) = \sqrt{t}\, O_p\left(m^{\epsilon}\right). \qquad (36)
\end{aligned}
$$

With regard to the fourth term in (35), we have

$$
\begin{aligned}
&-\tilde{\boldsymbol{v}}_1^\top Z^\top \underline{Q}^{(k)}\tilde{\boldsymbol{v}}^{\mathrm{S}} \le |\tilde{\boldsymbol{v}}_1^\top Z^\top \underline{Q}^{(k)}\tilde{\boldsymbol{v}}^{\mathrm{S}}| \le \|\tilde{\boldsymbol{v}}_1\|\|Z^\top \underline{Q}^{(k)}\tilde{\boldsymbol{v}}^{\mathrm{S}}\| = \|Z^\top \underline{Q}^{(k)}\tilde{\boldsymbol{v}}^{\mathrm{S}}\| \\
&\le \|Z^\top \underline{Q}^{(k)}\|_{\mathrm{op}}\|\tilde{\boldsymbol{v}}^{\mathrm{S}}\| \le \|Z\|_{\mathrm{op}}\|\underline{Q}^{(k)}\|_{\mathrm{F}}\|\tilde{\boldsymbol{v}}^{\mathrm{S}}\| = \sqrt{\lambda_1|\mathcal{I}_k|}\;|\eta_k|\;\|\tilde{\boldsymbol{v}}^{\mathrm{S}}\|. \qquad (37)
\end{aligned}
$$

By substituting (36) and (37) into (35),

$$
\begin{aligned}
&\|Z\tilde{\boldsymbol{v}}_1\|^2 - 2\sum_{k=1}^{K+H}\tau_k\tilde{\boldsymbol{v}}_1^\top Z^\top \underline{Q}^{(k)}\tilde{\boldsymbol{v}}_1 \\
&\le \lambda_1 - n^d\|\tilde{\boldsymbol{v}}^{\mathrm{S}}\|^2 + 2\sum_{k=1}^{K+H}\tau_k\sqrt{t}\, O_p\left(m^{\epsilon}\right) + 2\sum_{k=1}^{K+H}\tau_k\sqrt{\lambda_1|\mathcal{I}_k|}\;|\eta_k|\;\|\tilde{\boldsymbol{v}}^{\mathrm{S}}\| \\
&= \lambda_1 - n^d\|\tilde{\boldsymbol{v}}^{\mathrm{S}}\|^2 + 2\sum_{k=1}^{K+H}\sqrt{t}\, O_p\left(m^{\epsilon}\right) + 2\sum_{k=1}^{K+H}\tau_k\sqrt{\lambda_1|\mathcal{I}_k|}\;|\eta_k|\;\|\tilde{\boldsymbol{v}}^{\mathrm{S}}\| \\
&= \lambda_1 - n^d\|\tilde{\boldsymbol{v}}^{\mathrm{S}}\|^2 + \sqrt{t}\, O_p\left[(K+H)m^{\epsilon}\right] + 2\sum_{k=1}^{K+H}\tau_k\sqrt{\lambda_1|\mathcal{I}_k|}\;|\eta_k|\;\|\tilde{\boldsymbol{v}}^{\mathrm{S}}\|. \qquad (38)
\end{aligned}
$$

From now on, we derive the probabilistic order of t. We denote the jth normalized eigenvalue of matrix $Z^\top Z$ as $\nu_j \equiv \frac{1}{n}\lambda_j$, and define the following variables:

$$
\nu_+ \equiv \left(1+\sqrt{\frac{p}{n}}\right)^2, \quad \nu_- \equiv \left(1-\sqrt{\frac{p}{n}}\right)^2, \quad \epsilon_3 \equiv \nu_+ - \nu_1. \qquad (39)
$$

Note that $|\epsilon_3| = O_p\left(\phi^C m^{-\frac{2}{3}}\right)$ holds for some constant $C > 0$ and $\phi \equiv (\log p)^{\log\log p}$ from (4.1) of (Pillai and Yin 2014). Since $\phi = o(m^{\epsilon_4})$ holds for any $\epsilon_4 > 0$, by taking $\epsilon_5 \equiv C\epsilon_4$, we have

$$|\epsilon_3| = O_p\left(m^{-\frac{2}{3}+\epsilon_5}\right), \text{ for any } \epsilon_5 > 0. \tag{40}$$

From (3.7) of (Pillai and Yin 2014), we have

$$\left|\bar{n} - \frac{t}{p}\right| = O_p\left(m^{-1+\epsilon_6}\right), \text{ for all } \epsilon_6 > 0, \tag{41}$$

where $\bar{n} \equiv \int_{\nu_1 - n^{d-1}}^{\infty} q(x)\mathrm{d}x$ and

$$q(x) = \frac{1}{2\pi}\frac{n}{p}\frac{\sqrt{\max\{(\nu_+ - x)(x - \nu_-), 0\}}}{x}. \tag{42}$$

From (42), by taking $\epsilon_5 < d - \frac{1}{3} = \frac{8}{21}$, we have

$$\begin{aligned} &q(\nu_1 - n^{d-1}) = q(\nu_+ - n^{d-1} - \epsilon_3) \\ &= \frac{\sqrt{\nu_+ - \nu_-}}{\nu_+}\left[n^{\frac{d-1}{2}} + O_p\left(m^{-\frac{1}{3}+\frac{\epsilon_5}{2}}\right)\right]\left[1 + O\left(m^{\frac{d-1}{2}}\right) + O_p\left(m^{-\frac{1}{3}+\frac{\epsilon_5}{2}}\right)\right] \\ &= \frac{\sqrt{\nu_+ - \nu_-}}{\nu_+} n^{\frac{d-1}{2}} + O_p\left(m^{\frac{d-1}{2}}\right). \end{aligned} \tag{43}$$

From (40) and (43), by setting $\epsilon_5 < d - \frac{1}{3}$,

$$\begin{aligned} \bar{n} &= \int_{\nu_1 - n^{d-1}}^{\infty} q(x)\mathrm{d}x \le \left|\int_{\nu_1 - n^{d-1}}^{\nu_+} q(x)\mathrm{d}x\right| + \left|\int_{\nu_+}^{\infty} q(x)\mathrm{d}x\right| = \left|\int_{\nu_1 - n^{d-1}}^{\nu_+} q(x)\mathrm{d}x\right| \\ &\le |\epsilon_3 + n^{d-1}|\; q(\nu_1 - n^{d-1}) = O_p\left(m^{d-1}\right) O_p\left(m^{\frac{d-1}{2}}\right) = O_p\left(m^{\frac{3(d-1)}{2}}\right). \end{aligned} \tag{44}$$

From (44) and (41), by setting $\epsilon_6 < \frac{3}{2}d - \frac{1}{2}$,

$$t = O_p\left(m^{\frac{3}{2}d - \frac{1}{2}}\right). \tag{45}$$

By substituting (45) into (38) and from the assumption in (30) that $d = \frac{5}{7}$, for all $\epsilon > 0$,

$$\begin{aligned} &\|Z\tilde{\boldsymbol{v}}_1\|^2 - 2\sum_{k=1}^{K+H} \tau_k \tilde{\boldsymbol{v}}_1^\top Z^\top \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1 \\ &\le \lambda_1 + O_p\left[(K+H)m^{\frac{2}{7}+\epsilon}\right] + \|\tilde{\boldsymbol{v}}^{\mathrm{S}}\|\left(2\sqrt{\lambda_1}\sum_{k=1}^{K+H}\tau_k\sqrt{|\mathcal{I}_k|}\;|\eta_k| - n^d\|\tilde{\boldsymbol{v}}^{\mathrm{S}}\|\right) \\ &= \lambda_1 + O_p\left[(K+H)m^{\frac{2}{7}+\epsilon}\right] + \|\tilde{\boldsymbol{v}}^{\mathrm{S}}\|\left(2\sqrt{\lambda_1}\;\varpi - n^d\|\tilde{\boldsymbol{v}}^{\mathrm{S}}\|\right), \end{aligned} \tag{46}$$

where

$$\varpi \equiv \sum_{k=1}^{K+H} \tau_k \sqrt{|\mathcal{I}_k|}\ |\eta_k| = \sum_{k=1}^{K+H} \left[1 + O_p\left(\frac{1}{\sqrt{|\mathcal{I}_k|}} \right) \right] \sqrt{|\mathcal{I}_k|}\ O_p\left(\frac{1}{\sqrt{|\mathcal{I}_k|}} \right)$$
$$= O_p(K+H). \tag{47}$$

By using $d = \frac{5}{7}$ and (47), the third term in the right side of (46) can be upper bounded by

$$\|\tilde{\boldsymbol{v}}^{\mathrm{S}}\| \left(2\sqrt{\lambda_1}\ \varpi - n^d \|\tilde{\boldsymbol{v}}^{\mathrm{S}}\| \right) = 2\|\tilde{\boldsymbol{v}}^{\mathrm{S}}\| \sqrt{\lambda_1}\ \varpi - n^d \|\tilde{\boldsymbol{v}}^{\mathrm{S}}\|^2 - \frac{\lambda_1 \varpi^2}{n^d} + \frac{\lambda_1 \varpi^2}{n^d}$$
$$= -\frac{(\sqrt{\lambda_1}\ \varpi - n^d \|\tilde{\boldsymbol{v}}^{\mathrm{S}}\|)^2}{n^d} + \frac{\lambda_1 \varpi^2}{n^d} \le \frac{\lambda_1 \varpi^2}{n^d} = O_p\left[(K+H)^2 m^{\frac{2}{7}} \right], \tag{48}$$

which results in that

$$\|Z\tilde{\boldsymbol{v}}_1\|^2 - 2\sum_{k=1}^{K+H} \tau_k \tilde{\boldsymbol{v}}_1^\top Z^\top \underline{Q}^{(k)} \tilde{\boldsymbol{v}}_1$$
$$\le \lambda_1 + O_p\left[(K+H) m^{\frac{2}{7}+\epsilon} \right] + O_p\left[(K+H)^2 m^{\frac{2}{7}} \right]$$
$$\le \lambda_1 + O_p\left[(K+H)^2 m^{\frac{2}{7}+\epsilon} \right], \text{ for all } \epsilon > 0. \tag{49}$$

Therefore, from (32), (33), and (34), for all $\epsilon > 0$,

$$\tilde{\lambda}_1 \le \lambda_1 + O_p\left[(K+H)^2 m^{\frac{2}{7}+\epsilon} \right] + O_p\left(m^{\frac{1}{3}-\epsilon_1} \right) + O_p\left(m^{\frac{1}{21}-2(\epsilon_1-\epsilon)} \right). \tag{50}$$

From the assumption (iv) that $K + H = O\left(m^{\frac{1}{42}-\epsilon_1} \right)$ for some $\epsilon_1 > 0$, by taking $\epsilon < \epsilon_1$, we finally obtain

$$\tilde{\lambda}_1 \le \lambda_1 + O_p\left(m^{\frac{1}{3}-\tilde{\epsilon}} \right), \text{ for some } \tilde{\epsilon} > 0, \tag{51}$$

which concludes the proof. □

Appendix D Disjoint submatrix localization algorithm based on simulated annealing

In this section, we develop a simulated annealing (SA) algorithm to find the bicluster structure of a given observed matrix. As in (Flynn and Perry 2020), the proposed algorithm is based on the (generalized) profile likelihood (Murphy and Vaart 2000). Given an estimated bicluster assignment $\hat{g}$, the generalized profile-likelihood criterion for an exponential family model is given by

$$F(\hat{g}) \equiv \sum_{k=0}^{K} \hat{p}_k f\left(\frac{1}{|\hat{\mathcal{I}}_k|} \sum_{(i,j)\in\hat{\mathcal{I}}_k} A_{ij} \right), \tag{52}$$

where $\hat{p}_k \in \mathbb{R}$ is the proportion of entries in the kth group ($k \in \{0, 1, \ldots, K\}$) in the estimated bicluster structures to all the np entries and $f : \mathbb{R} \mapsto \mathbb{R}$ is a given function. The specific definition of function f for each experiment is given in Sect. 4. These settings in Sect. 4 is based on the following derivation.

Gaussian LBM (G-LBM) We assume that each entry A_{ij} in the kth group of the data matrix independently follows the Gaussian distribution $\mathcal{N}(b_k, \sigma)$, where σ is a known standard deviation, which is common to all the groups. This assumption is necessary for deriving function f in the framework of profile likelihood maximization, and to derive another submatrix localization algorithm that does not require such assumption is beyond the scope of this paper. In this case, the log likelihood is given by

$$\begin{aligned}\mathcal{L}(\hat{g}) &= \sum_{k=0}^{K} \sum_{(i,j)\in\hat{\mathcal{I}}_k} \left(-\log\sqrt{2\pi\sigma^2} - \frac{(A_{ij} - b_k)^2}{2\sigma^2} \right) \\ &= -np\log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2} \sum_{k=0}^{K} \sum_{(i,j)\in\hat{\mathcal{I}}_k} (A_{ij} - b_k)^2. \end{aligned} \tag{53}$$

By replacing b_k with the maximum likelihood estimator $\hat{b}_k = \frac{1}{|\hat{\mathcal{I}}_k|} \sum_{(i,j)\in\hat{\mathcal{I}}_k} A_{ij}$, we obtain

$$\begin{aligned}F^{(0)}(\hat{g}) &\equiv -np\log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2} \sum_{k=0}^{K} \sum_{(i,j)\in\hat{\mathcal{I}}_k} (A_{ij} - \hat{b}_k)^2 \\ &= -np\log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}\|A\|_{\mathrm{F}}^2 + \sum_{k=0}^{K} |\hat{\mathcal{I}}_k| \frac{\hat{b}_k^2}{2\sigma^2} \\ &= -np\log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}\|A\|_{\mathrm{F}}^2 + \frac{np}{\sigma^2} \sum_{k=0}^{K} \hat{p}_k \frac{\hat{b}_k^2}{2}. \end{aligned} \tag{54}$$

Since maximization of $F^{(0)}(\hat{g})$ is equivalent to that of $F(\hat{g}) \equiv \sum_{k=0}^{K} \hat{p}_k \frac{\hat{b}_k^2}{2}$, we define $F(\hat{g})$ as the profile likelihood. This corresponds to the definition of $f(x) \equiv x^2/2$ in (52).

Bernoulli LBM (B-LBM) We assume that each entry A_{ij} in the kth group of the data matrix independently follows the Bernoulli distribution Bernoulli(b_k). In this case, the log likelihood is given by

$$\begin{aligned}\mathcal{L}(\hat{g}) &= \sum_{k=0}^{K} \sum_{(i,j)\in\hat{\mathcal{I}}_k} \log\left[b_k^{A_{ij}} (1 - b_k)^{1-A_{ij}} \right] \\ &= \sum_{k=0}^{K} \left[|\hat{\mathcal{I}}_k| \hat{b}_k \log b_k + |\hat{\mathcal{I}}_k| (1 - \hat{b}_k) \log(1 - b_k) \right]. \end{aligned} \tag{55}$$

By replacing b_k with $\hat{b}_k$, we obtain

$$\begin{aligned} F^{(0)}(\hat{g}) &\equiv \sum_{k=0}^{K} \left[|\hat{\mathcal{I}}_k| \hat{b}_k \log \hat{b}_k + |\hat{\mathcal{I}}_k| (1 - \hat{b}_k) \log(1 - \hat{b}_k) \right] \\ &= np \sum_{k=0}^{K} \hat{p}_k \left[\hat{b}_k \log \hat{b}_k + (1 - \hat{b}_k) \log(1 - \hat{b}_k) \right]. \end{aligned} \tag{56}$$

Since maximization of $F^{(0)}(\hat{g})$ is equivalent to that of $F(\hat{g}) \equiv \frac{1}{np} F^{(0)}(\hat{g})$, we define $F(\hat{g})$ as the profile likelihood. This corresponds to the definition of $f(x) \equiv x \log x + (1 - x) \log(1 - x)$ in (52).

Poisson LBM (P-LBM) We assume that each entry A_{ij} in the kth group of the data matrix independently follows the Poisson distribution $\mathrm{Pois}(b_k)$. In this case, the log likelihood is given by

$$\begin{aligned} \mathcal{L}(\hat{g}) &= \sum_{k=0}^{K} \sum_{(i,j) \in \hat{\mathcal{I}}_k} \log \left(\frac{b_k^{A_{ij}} \exp(-b_k)}{A_{ij}!} \right) \\ &= \sum_{k=0}^{K} |\hat{\mathcal{I}}_k| \hat{b}_k \log b_k - \sum_{k=0}^{K} |\hat{\mathcal{I}}_k| b_k - \sum_{i=1}^{n} \sum_{j=1}^{p} \log(A_{ij}!). \end{aligned} \tag{57}$$

By replacing b_k with $\hat{b}_k$, we obtain

$$\begin{aligned} F^{(0)}(\hat{g}) &\equiv \sum_{k=0}^{K} |\hat{\mathcal{I}}_k| (\hat{b}_k \log \hat{b}_k - \hat{b}_k) - \sum_{i=1}^{n} \sum_{j=1}^{p} \log(A_{ij}!) \\ &= np \sum_{k=0}^{K} \hat{p}_k (\hat{b}_k \log \hat{b}_k - \hat{b}_k) - \sum_{i=1}^{n} \sum_{j=1}^{p} \log(A_{ij}!). \end{aligned} \tag{58}$$

Since maximization of $F^{(0)}(\hat{g})$ is equivalent to that of $F(\hat{g}) \equiv \sum_{k=0}^{K} \hat{p}_k (\hat{b}_k \log \hat{b}_k - \hat{b}_k)$, we define $F(\hat{g})$ as the profile likelihood. This corresponds to the definition of $f(x) \equiv x \log x - x$ in (52).

D.0.1 The naive implementation of SA-based submatrix localization

Let $\mathcal{G}_K$ be a set of all bicluster structures with (non-empty) K biclusters, which are disjoint, but which are not necessarily bi-disjoint. In SA, we first define a sequence of temperatures $\{T_t\}_{t=0}^{\infty}$, a threshold ϵ^{SA}, and the initial state (i.e., bicluster assignment) $\hat{g}^{(0)} \in \mathcal{G}_K$. For each state $g \in \mathcal{G}_K$, we also define a set of its neighbors $N(g) \subseteq \mathcal{G}_K$ and a transition probability $R(g, g') \in [0, 1]$ to a given state $g' \in \mathcal{G}_K$. Here, we set $R(g, g') = 0$ iff $g' \notin N(g)$.

For each step $t = 0, 1, 2, \ldots$, if $T_t < \epsilon$, we stop the algorithm and output the final state $\hat{g}^{(t)}$. If $T_t \geq \epsilon$, we randomly choose a candidate for the next state $\tilde{g} \in N(\hat{g}^{(t)})$ with probability $R(\hat{g}^{(t)}, \tilde{g})$, and compute the difference of the objective function value $\Delta F \equiv F(\tilde{g}) - F(\hat{g}^{(t)})$. If $\Delta F > 0$, we set the next state at $\hat{g}^{(t+1)} = \tilde{g}$. If $\Delta F \leq 0$, we set the next state at $\hat{g}^{(t+1)} = \tilde{g}$ with probability $\exp\left(\frac{\Delta F}{T_t}\right)$, and set it at the current state $\hat{g}^{(t+1)} = \hat{g}^{(t)}$ with probability $1 - \exp\left(\frac{\Delta F}{T_t}\right)$.

Specifically, we propose Algorithm 1 as an example of SA for approximately maximizing the generalized profile likelihood F. In Algorithm 1, we define that the neighbors $N(g)$ of a state g is a set of all possible bicluster assignments that can be obtained by adding/removing one row or column to/from one bicluster in g. As for the transition probability, we define that one of the elements in $N(g)$ is chosen from the uniform distribution on $N(g)$ (i.e., $R(g, g') = 1/|N(g)|$ for $g' \in N(g)$).

We can easily check that the above settings satisfy the following *irreducibility* and *weak reversibility*:

- Irreducibility: for any pair $g, g' \in \mathcal{G}_K$, there exists some sequence of transitions from g to g' with non-zero probability.
- Weak reversibility: for any pair $g, g' \in \mathcal{G}_K$ and $\tilde{F} \in \mathbb{R}$, the following two propositions (P1) and (P2) are mutually equivalent:
  - (P1) there exists some sequence of transitions $g_1 = g, g_2, \ldots, g_p = g'$ with non-zero probability that satisfies $F(g_t) \geq \tilde{F}$ for all $t \in \{1, \ldots, p\}$.
  - (P2) there exists some sequence of transitions $g_1 = g', g_2, \ldots, g_p = g$ with non-zero probability that satisfies $F(g_t) \geq \tilde{F}$ for all $t \in \{1, \ldots, p\}$.

We define that a state g is *locally optimal* if there is no state $g' \in \mathcal{G}_K$ that satisfies the following two conditions simultaneously: $F(g') > F(g)$, and there exists some sequence of transitions $g_1 = g, g_2, \ldots, g_p = g'$ with non-zero probability that satisfies $F(g_t) \geq F(g)$ for all $t \in \{1, \ldots, p\}$. For a locally but not globally optimal solution g, we define its *depth* as the minimum r that satisfies the following condition: there exists some g' such that $F(g') > F(g)$ and there exists some sequence of transitions $g_1 = g, g_2, \ldots, g_p = g'$ with non-zero probability that satisfies $F(g_t) \geq F(g) - r$ for all $t \in \{1, \ldots, p\}$. By setting the sequence of temperatures at $T_t = [\max_{g \in \mathcal{G}_K} F(g) - \min_{g \in \mathcal{G}_K} F(g)] / \log(t+2)$ for all $t \geq 0$ (Hajek 1988), for example, the proposed Algorithm 1 also satisfies the following conditions:

- $T_t \geq T_{t+1}$ holds for all $t \geq 0$, and $\lim_{t \to \infty} T_t = 0$.
- $\sum_{t=0}^{\infty} \exp\left(-\frac{r^*}{T_t}\right) = +\infty$, where r^* is the maximum depth of all the locally but not globally optimal solutions.

It has been proven that under the above conditions, the probability that an SA algorithm outputs the global optimal solution converges to one in the limit of $t \to \infty$ (Hajek 1988).

D.0.2 A further approximated version of SA-based submatrix localization algorithm

Although the naive SA algorithm in Sect. D.0.1 is tractable compared to the exhaustive search, it still requires too many steps for the algorithm to converge.

Algorithm 1 A naive SA algorithm for finding the maximum profile likelihood solution $\hat{g}$.

```
Require: A cooling schedule of temperature $\{T_t\}_{t=0}^{\infty}$ and a threshold $\epsilon^{\mathrm{SA}}$.
Ensure: Approximated optimal bicluster assignment $\hat{g}$.
 1: $t \leftarrow 0$.
 2: Randomly generate an initial bicluster assignment $\hat{g}$, which is disjoint but not necessarily
    bi-disjoint.
 3: while $T_t \geq \epsilon^{\mathrm{SA}}$ do
 4:     Set $\tilde{g} \leftarrow \hat{g}$ and randomly choose an index $k_0$ from the uniform distribution on
        $\{1, \dots, 2K\}$.
 5:     if $k_0 \leq K$ then
 6:         Set bicluster index $k \leftarrow k_0$.
 7:         Let $I_k$ and $J_k = \{j_1, \dots, j_{|J_k|}\}$, respectively, be the sets of row and column indices
            in the $k$th bicluster. We define add and remove lists as follows.
 8:         For $i \in I_k$, let $\mathcal{I}_{ki}^{\mathrm{rem}}$ be the set of entries in the $i$th row of the $k$th bicluster (i.e.,
            $\{(i, j_1), (i, j_2), \dots, (i, j_{|J_k|})\}$). We define the remove list as $\mathcal{I}_k^{\mathrm{rem}} = \{\mathcal{I}_{ki}^{\mathrm{rem}}\}_{i \in I_k}$.
 9:         Let $\bar{I}_k$ be the set of row indices $i$ that satisfies $\bigcap_{s=1}^{|J_k|} \bigcap_{k'=1}^{K} [(i, j_s) \notin \mathcal{I}_{k'}]$. For $i \in \bar{I}_k$,
            let $\mathcal{I}_{ki}^{\mathrm{add}}$ be the set of entries $\{(i, j_1), (i, j_2), \dots, (i, j_{|J_k|})\}$. We define the add list
            as $\mathcal{I}_k^{\mathrm{add}} = \{\mathcal{I}_{ki}^{\mathrm{add}}\}_{i \in \bar{I}_k}$.
10:         Let $\mathcal{I}_0$ be the set of background entries in $\tilde{g}$. Set $y^{\mathrm{add}} \leftarrow (|\bar{I}_k| \geq 2) \cup [(|\bar{I}_k| =
            1) \cap (\mathcal{I}_0 \neq \mathcal{I}_k^{\mathrm{add}})]$, which is a flag of whether or not we can execute "add" operation.
            This guarantees that the set of background entries is not null.
11:         if $|I_k| \geq 2$ and $y^{\mathrm{add}} = \mathrm{True}$ then
12:             Randomly choose $i$ from the uniform distribution on $\{1, \dots, |I_k| + |\bar{I}_k|\}$. If $i \leq
                |I_k|$, remove $\mathcal{I}_{ki}^{\mathrm{rem}}$ from the $k$th bicluster and add it to the background in $\tilde{g}$. If
                $i > |I_k|$, remove $\mathcal{I}_{k(i-|I_k|)}^{\mathrm{add}}$ from the background and add it to the $k$th bicluster
                in $\tilde{g}$.
13:         else if $|I_k| \geq 2$ then
14:             Randomly choose $i$ from the uniform distribution on $\{1, \dots, |I_k|\}$. Remove $\mathcal{I}_{ki}^{\mathrm{rem}}$
                from the $k$th bicluster and add it to the background in $\tilde{g}$.
15:         else if $y^{\mathrm{add}} = \mathrm{True}$ then
16:             Randomly choose $i$ from the uniform distribution on $\{1, \dots, |\bar{I}_k|\}$. Remove $\mathcal{I}_{ki}^{\mathrm{add}}$
                from the background and add it to the $k$th bicluster in $\tilde{g}$.
17:         end if
18:     else
19:         Set the bicluster index $k \leftarrow k_0 - K$.
20:         Execute lines 7 to 17 by swapping the rows and columns in all the operations.
21:     end if
22:     if $F(\tilde{g}) - F(\hat{g}) > 0$ then
23:         $\hat{g} \leftarrow \tilde{g}$.
24:     else
25:         With probability $\exp\left(\frac{F(\tilde{g}) - F(\hat{g})}{T_t}\right)$, $\hat{g} \leftarrow \tilde{g}$.
26:     end if
27:     $t \leftarrow t + 1$.
28: end while
```

Therefore, in this subsection, we propose a further approximation of Algorithm 1. The main idea here is to first compress an observed data matrix A by using row-column clustering, and then apply an SA algorithm on the compressed data matrix.

Remark that the **null** group-wise mean matrix P with K biclusters has at most 2^K distinct rows, depending on whether or not it includes each kth bicluster ($k = 1, \dots, K$). Based on this fact, we first apply a clustering method (e.g., hierarchical clustering) to the rows of matrix A, by setting the number of clusters at $L_1 \in \mathbb{N}$, which satisfies $\min\{2^K, n\} \leq L_1 \leq n$. Based on a similar discussion, we also perform column clustering with number of clusters L_2 that satisfies $\min\{2^K, p\} \leq L_2 \leq p$. Then, we define the compressed observed matrix $A^{\mathrm{comp}} \in \mathbb{R}^{L_1 \times L_2}$ and matrix $M \in \mathbb{N}^{L_1 \times L_2}$ as follows:

$$\begin{aligned} A^{\mathrm{comp}} &= (A^{\mathrm{comp}}_{hh'})_{1 \leq h \leq L_1, 1 \leq h' \leq L_2}, \qquad A^{\mathrm{comp}}_{hh'} = \frac{1}{|\mathcal{I}^{\mathrm{comp}}_{hh'}|} \sum_{(i,j) \in \mathcal{I}^{\mathrm{comp}}_{hh'}} A_{ij}, \\ M &= (M_{hh'})_{1 \leq h \leq L_1, 1 \leq h' \leq L_2}, \qquad M_{hh'} = |\mathcal{I}^{\mathrm{comp}}_{hh'}|, \end{aligned} \tag{59}$$

where $\mathcal{I}^{\mathrm{comp}}_{hh'}$ is the set of entries of matrix A in the hth row cluster and the h'th column cluster.

Next, we apply an SA algorithm to the compressed observed matrix A^{comp}. Let $\hat{g}^{\mathrm{comp}}_{hh'} \in \{0, 1, \dots, K\}$ be the estimated group index of the (h, h')th entry of matrix A^{comp}, and let $\mathcal{J}^{\mathrm{comp}}_k \subseteq \{(1,1), \dots, (L_1, L_2)\}$ be the set of entries in the kth estimated group ($k = 0, 1, \dots, K$) of matrix A^{comp}. Note that we have $\mathcal{J}^{\mathrm{comp}}_k = \{(h, h') : \hat{g}^{\mathrm{comp}}_{hh'} = k\}$.

The key insight is that we have

$$\begin{aligned} \hat{p}_k &= \frac{1}{np} \sum_{(h,h') \in \mathcal{J}^{\mathrm{comp}}_k} M_{hh'}, \\ \frac{1}{|\hat{\mathcal{I}}_k|} \sum_{(i,j) \in \hat{\mathcal{I}}_k} A_{ij} &= \frac{1}{np\hat{p}_k} \sum_{(h,h') \in \mathcal{J}^{\mathrm{comp}}_k} M_{hh'} A^{\mathrm{comp}}_{hh'}. \end{aligned} \tag{60}$$

Based on the above fact, we can compute the objective function value (i.e., profile likelihood) based on the matrices A^{comp} and M, and the bicluster assignment $\hat{g}^{\mathrm{comp}} = (\hat{g}^{\mathrm{comp}}_{hh'})_{1 \leq h \leq L_1, 1 \leq h' \leq L_2}$ by

$$F(\hat{g}^{\mathrm{comp}}) \equiv \sum_{k=0}^{K} \left(\frac{1}{np} \sum_{(h,h') \in \mathcal{J}^{\mathrm{comp}}_k} M_{hh'} \right) f \left(\frac{1}{np\hat{p}_k} \sum_{(h,h') \in \mathcal{J}^{\mathrm{comp}}_k} M_{hh'} A^{\mathrm{comp}}_{hh'} \right). \tag{61}$$

From these observations, Algorithm 2 provides an approximated solution of Algorithm 1.

Appendix E Greedy submatrix localization algorithm

To check the sensitivity of the proposed test with regard to the biclustering method, we also try Algorithm 3, which is an extension of the greedy biclustering method in (Flynn and Perry 2020) to general disjoint bicluster structure. We first use the same approximation method as in Algorithm 2 to compress a data matrix and then apply Algorithm 3 to the compressed data matrix.

Algorithm 2 Approximated SA algorithm for finding the maximum profile likelihood solution $\hat{g}$.

```
Require: A set of row and column cluster numbers $(L_1, L_2)$ that satisfies $L_1 \geq 2^K$ and
    $L_2 \geq 2^K$, a cooling schedule of temperature $\{T_t\}_{t=0}^{\infty}$ and a threshold $\epsilon^{\mathrm{SA}}$.
Ensure: An approximated optimal bicluster assignment $\hat{g}$.
  Apply a clustering algorithm to the rows of observed matrix $A$ with the number of
    clusters $L_1$.
  Apply a clustering algorithm to the columns of observed matrix $A$ with the number of
    clusters $L_2$.
  Let $\mathcal{I}_{hh'}^{\mathrm{comp}}$ be the set of entries of matrix $A$ in the $h$th row cluster and the $h'$th column
    cluster, and let $\mathcal{I}^{\mathrm{comp}} = (\mathcal{I}_{hh'}^{\mathrm{comp}})_{1 \leq h \leq L_1, 1 \leq h' \leq L_2}$. Based on the clustering result $\mathcal{I}^{\mathrm{comp}}$,
    define the matrices $A^{\mathrm{comp}}$ and $M$ by (59).
  $t \leftarrow 0$.
  Randomly generate initial (compressed) bicluster assignment $\hat{g}^{\mathrm{comp}}$, which is disjoint
    but not necessarily bi-disjoint.
  Execute lines 3 to 28 in Algorithm 1 by replacing $A$ and $\hat{g}$ with $A^{\mathrm{comp}}$ and $\hat{g}^{\mathrm{comp}}$,
    respectively. As for the objective function value, we can compute it by using (61).
  Convert the set of results $\mathcal{I}^{\mathrm{comp}}$ and $\hat{g}^{\mathrm{comp}}$ into the bicluster assignment $\hat{g}$ of the original
    observed matrix $A$.
```

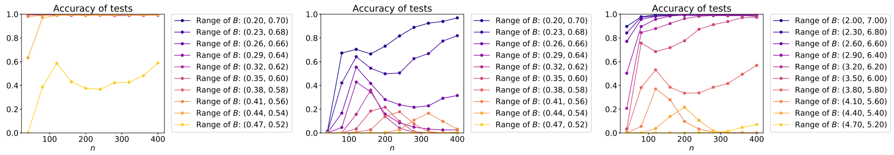

Fig. 2 The accuracy of the proposed test in selecting the number of biclusters K when using the greedy submatrix localization algorithm. The left, center, and right figures, respectively, illustrate the results where each entry of observed matrix A was generated using Gaussian, Bernoulli, and Poisson distributions.

We checked the accuracy of the proposed test by using Algorithm 3 for submatrix localization. Aside from the submatrix localization algorithm, we used the same experimental settings as in Sect. 4.3. We applied Algorithm 3 to each observed matrix $1,000$ times and adopted the best solution that achieved the maximum profile likelihood in the last step of the algorithm.

Figure 2 shows the accuracy of the proposed test. Although Algorithm 3 has no theoretical guarantee for obtaining the global optimal bicluster structure, we see that the proposed test with Algorithm 3 could achieve higher accuracy than that with the SA-based algorithm in most cases under this experimental setting.

Appendix F Attributes of the Divorce Predictors data set

Table 1 indicates the meaning of each attribute index of the Divorce Predictors data set (Yöntem et al. 2019), which we used in the experiment.

Algorithm 3 Greedy algorithm for finding the maximum profile likelihood solution $\hat{g}$.

```
**Ensure:** Approximated optimal bicluster assignment $\hat{g}$.
 1: Randomly generate an initial bicluster assignment $\hat{g}$, which is disjoint but not necessarily
    bi-disjoint. Let $I_k$ and $J_k$, respectively, be the sets of row and column indices in the $k$th
    bicluster ($k = 1, \dots, K$) in $\hat{g}$.
 2: while True do
 3:   $\hat{g}^{(0)} \leftarrow \hat{g}$.
 4:   for $k = 1, \dots, K$ do
 5:     $\mathit{\Delta} F \leftarrow \begin{bmatrix} 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{n+p}$.
 6:     for $i = 1, \dots, n$ do
 7:       $\tilde{g} \leftarrow \hat{g}$.
 8:       if $i \in I_k$ then
 9:         if $|I_k| > 1$ then
10:           $\tilde{g}_{ij} \leftarrow 0$ for $j \in J_k$.
11:         end if
12:       else
13:         Let $y =$ True if there is at least one background entry when the $i$th row is
            added to the $k$th bicluster in $\tilde{g}$, and let $y =$ False otherwise. Let $y' =$ True
            if no entry in the $i$th row with column indices $J_k$ are included in any other
            bicluster than the $k$th one in $\hat{g}$ (i.e., the disjoint condition is satisfied).
14:         if $y \cap y'$ then
15:           $\tilde{g}_{ij} \leftarrow k$ for $j \in J_k$.
16:         end if
17:       end if
18:       $\tilde{F} \leftarrow F(\tilde{g})$. $\mathit{\Delta} F_i \leftarrow \tilde{F} - F(\hat{g})$.
19:     end for
20:     for $j = 1, \dots, p$ do
21:       Execute lines 7 to 18 by swapping the rows and columns in all the operations.
22:     end for
23:     Let $n^{\mathrm{cand}}$ be the number of entries in $\mathit{\Delta} F$ that satisfy $\mathit{\Delta} F_i > 0$ and let $\boldsymbol{\theta} =$
        $(\theta_i)_{1 \le i \le n^{\mathrm{cand}}}$ be the descending order of the indices of $\mathit{\Delta} F$ (i.e., $\mathit{\Delta} F_{\theta_1} \ge \cdots \ge$
        $\mathit{\Delta} F_{\theta_{n^{\mathrm{cand}}}} > 0$). $\tilde{g} \leftarrow \hat{g}$. $g^{\mathrm{opt}} \leftarrow \hat{g}$. $\mathit{\Delta} F^{\mathrm{opt}} \leftarrow -\infty$. $F^{\mathrm{opt}} \leftarrow F(\hat{g})$.
24:     for $i = 1, \dots, n^{\mathrm{cand}}$ do
25:       $t \leftarrow \theta_i$. Let $\tilde{I}_k$ and $\tilde{J}_k$, respectively, be the sets of row and column indices in
          the $k$th bicluster ($k = 1, \dots, K$) in $\tilde{g}$.
26:       if $t \le n$ then
27:         if $t \in \tilde{I}_k$ then
28:           if $|\tilde{I}_k| > 1$ then
29:             $\tilde{g}_{tj} \leftarrow 0$ for $j \in \tilde{J}_k$.
30:           end if
31:         else
32:           Let $y =$ True if there is at least one background entry when the $t$th row is
              added to the $k$th bicluster in $\tilde{g}$, and let $y =$ False otherwise. Let $y' =$ True
              if no entry in the $t$th row with column indices $\tilde{J}_k$ are included in any other
              bicluster than the $k$th one in $\tilde{g}$.
33:           if $y \cap y'$ then
34:             $\tilde{g}_{tj} \leftarrow k$ for $j \in \tilde{J}_k$.
35:           end if
36:         end if
37:         $\tilde{F} \leftarrow F(\tilde{g})$. $\mathit{\Delta} F \leftarrow \tilde{F} - F(\hat{g})$.
38:       else
39:         Execute lines 27 to 37 by swapping the rows and columns in all the operations.
40:       end if
41:       if $\mathit{\Delta} F > \mathit{\Delta} F^{\mathrm{opt}}$ then
42:         $g^{\mathrm{opt}} \leftarrow \tilde{g}$. $F^{\mathrm{opt}} \leftarrow \tilde{F}$. $\mathit{\Delta} F^{\mathrm{opt}} \leftarrow \mathit{\Delta} F$.
43:       end if
44:     end for
45:     $\hat{g} \leftarrow g^{\mathrm{opt}}$.
46:   end for
47:   if $\hat{g}^{(0)} = \hat{g}$ then
48:     break
49:   end if
50: end while
```

Table 1 Attributes of the Divorce Predictors data set (Yöntem et al. 2019).

1	If one of us apologizes when our discussion deteriorates, the discussion ends.
2	I know we can ignore our differences, even if things get hard sometimes.
3	When we need it, we can take our discussions with my spouse from the beginning and correct it.
4	When I discuss with my spouse, to contact him will eventually work.
5	The time I spent with my wife is special for us.
6	We don't have time at home as partners.
7	We are like two strangers who share the same environment at home rather than family.
8	I enjoy our holidays with my wife.
9	I enjoy traveling with my wife.
10	Most of our goals are common to my spouse.
11	I think that one day in the future, when I look back, I see that my spouse and I have been in harmony with each other.
12	My spouse and I have similar values in terms of personal freedom.
13	My spouse and I have similar sense of entertainment.
14	Most of our goals for people (children, friends, etc.) are the same.
15	Our dreams with my spouse are similar and harmonious.
16	We're compatible with my spouse about what love should be.
17	We share the same views about being happy in our life with my spouse.
18	My spouse and I have similar ideas about how marriage should be.
19	My spouse and I have similar ideas about how roles should be in marriage.
20	My spouse and I have similar values in trust.
21	I know exactly what my wife likes.
22	I know how my spouse wants to be taken care of when she/he sick.
23	I know my spouse's favorite food.
24	I can tell you what kind of stress my spouse is facing in her/his life.
25	I have knowledge of my spouse's inner world.
26	I know my spouse's basic anxieties.
27	I know what my spouse's current sources of stress are.
28	I know my spouse's hopes and wishes.
29	I know my spouse very well.
30	I know my spouse's friends and their social relationships.
31	I feel aggressive when I argue with my spouse.
32	When discussing with my spouse, I usually use expressions such as 'you always' or 'you never.'
33	I can use negative statements about my spouse's personality during our discussions.
34	I can use offensive expressions during our discussions.
35	I can insult my spouse during our discussions.
36	I can be humiliating when we discussions.
37	My discussion with my spouse is not calm.
38	I hate my spouse's way of open a subject.
39	Our discussions often occur suddenly.
40	We're just starting a discussion before I know what's going on.
41	When I talk to my spouse about something, my calm suddenly breaks.
42	When I argue with my spouse, I only go out and I don't say a word.
43	I mostly stay silent to calm the environment a little bit.
44	Sometimes I think it's good for me to leave home for a while.
45	I'd rather stay silent than discuss with my spouse.
46	Even if I'm right in the discussion, I stay silent to hurt my spouse.
47	When I discuss with my spouse, I stay silent because I am afraid of not being able to control my anger.
48	I feel right in our discussions.
49	I have nothing to do with what I've been accused of.
50	I'm not actually the one who's guilty about what I'm accused of.
51	I'm not the one who's wrong about problems at home.
52	I wouldn't hesitate to tell my spouse about her/his inadequacy.
53	When I discuss, I remind my spouse of her/his inadequacy.
54	I'm not afraid to tell my spouse about her/his incompetence.

References

Bloemendal A, Knowles A, Yau HT, Yin J (2016) On the principal components of sample covariance matrices. *Probability Theory and Related Fields* 164:459–552

Flynn CJ, Perry PO (2020) Profile likelihood biclustering. *Electronic Journal of Statistics* 14(1):731–768

Hajek B (1988) Cooling schedules for optimal annealing. *Mathematics of Operations Research* 13(2):311–329

Murphy SA, Vaart AWVD (2000) On profile likelihood. *Journal of the American Statistical Association* 95(450):449–465

Pillai NS, Yin J (2014) Universality of covariance matrices. *Annals of Applied Probability* 24(3):935–1001

Yöntem MK, Adem K, Ilhan T, Kılıçarslan S (2019) Divorce prediction using correlation based feature selection and artificial neural networks. *Nevşehir HacıBektaş Veli Üniversitesi SBE Dergisi* 9:259–273