Annals of the Institute of Statistical Mathematics (2023) 75:979–1009
https://doi.org/10.1007/s10463-023-00869-3

A goodness-of-fit test on the number of biclusters in a relational data matrix

Chihiro Watanabe[1] · **Taiji Suzuki**[1,2]

Received: 1 June 2022 / Accepted: 28 February 2023 / Published online: 17 April 2023

Abstract
Biclustering is a method for detecting homogeneous submatrices in a given matrix. Although there are many studies that estimate the underlying bicluster structure of a matrix, few have enabled us to determine the appropriate number of biclusters. Recently, a statistical test on the number of biclusters has been proposed for a regular-grid bicluster structure. However, when the latent bicluster structure does not satisfy such regular-grid assumption, the previous test requires a larger number of biclusters than necessary for the null hypothesis to be accepted, which is not desirable in terms of interpreting the accepted structure. In this study, we propose a new statistical test on the number of biclusters that does not require the regular-grid assumption and derive the asymptotic behavior of the proposed test statistic in both null and alternative cases. We illustrate the effectiveness of the proposed method by applying it to both synthetic and practical data matrices.

Keywords Biclustering · Submatrix detection · Goodness-of-fit test · Random matrix theory

1 Introduction

Relational data are a kind of matrix data, the entries of which reflect some kind of relationship between two (generally different) objects. For example, the rows and columns, respectively, of an observed matrix $A \in \mathbb{R}^{n \times p}$ represent customers and products, and each entry A_{ij} is a number of times for which the ith customer purchased the jth product. It has been shown that we can successfully model various kinds of relational data matrices, including customer-item transaction/rating data (Shan and Banerjee 2008; Symeonidis et al. 2007), document-word co-occurrence

✉ Chihiro Watanabe
chihiro.watanabe.xz@hco.ntt.co.jp

[1] Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

[2] Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo, Japan

data (Dhillon 2001; França 2012), and gene expression data (Madeira and Oliveira 2004; Oghabian et al. 2014; Prelić et al. 2006; Tanay et al. 2002), by assuming the existence of a latent *bicluster* or "homogeneous" submatrix (e.g., the entries in each bicluster are identically distributed). In the example above, concerning the customer-item relational data matrix, this assumption corresponds to that there are some groups of customers $I \subseteq \{1, \dots, n\}$ and some groups of items $J \subseteq \{1, \dots, p\}$, and that the customers in I tend to purchase the items in J at a similar frequency.

Regarding the bicluster structure of a relational data matrix, the following two problems have been extensively studied in the literature: *submatrix detection* and *localization*. Submatrix detection serves to detect the existence of such biclusters in a given observed matrix A (i.e., whether or not matrix A contains at least one bicluster) (Butucea and Ingster 2013; Hartigan 1972; Ma and Wu 2015; Shabalin et al. 2009). In this paper, as in a number of previous studies (Cai et al. 2017; Chen and Xu 2016), we distinguish such a task from submatrix localization (which is also known as *biclustering*), the purpose of which is to recover the exact position of such biclusters. So far, many biclustering methods have been proposed for a fixed number of biclusters K (Cai et al. 2017; Chen and Xu 2016; Hajek et al. 2018; Hochreiter et al. 2010; Shabalin et al. 2009). In most practical cases, however, there would not be any prior knowledge about K in a given data matrix. Therefore, it is an important task to develop some method to appropriately determine K from the observed data A. In the next two paragraphs, we outline some related studies that propose methods for choosing K.

A statistical test on the number of biclusters K. Although many studies have tested whether an observed matrix A contains any large average submatrix (Brennan et al. 2019; Butucea and Ingster 2013; Cai and Wu 2020; Liu and Guo 2018; Ma and Wu 2015), few statistical test methods have been proposed for ascertaining the number of biclusters K in a given matrix A. Recently, statistical tests on K have been proposed in (Bickel and Sarkar 2016; Hu et al. 2020; Lei 2016; Watanabe and Suzuki 2021) with the constraint that the underlying bicluster structure should be represented by a regular grid (as shown in Fig. 1b-2). Particularly, in (Bickel and Sarkar 2016; Hu et al. 2020; Lei 2016), the observed matrix A (and thus its bicluster structure) is assumed to be square symmetric. However, if the latent bicluster structure does not satisfy the regular-grid constraint (as shown in Fig. 1b-1), such a test needs a larger hypothetical number of biclusters K_0 than necessary (i.e., a finer bicluster structure than necessary) to accept the null hypothesis $K = K_0$, which is not desirable from the perspective of interpreting the accepted bicluster structure.[1] To cope with such a problem, a more flexible model is required, one which can represent the existence of local biclusters (Shabalin et al. 2009). For a singular value decomposition-based biclustering, a stopping criterion has been proposed for detecting multiple biclusters (which determines K) based on stability selection in (Sill

[1] By its nature, non-rejection of a null hypothesis in a statistical test does not mean the positive acceptance of it, and this is also the case when selecting the number of biclusters with the proposed test. However, we prioritize simplicity and use the term "accepted" to mean "not rejected" throughout this paper.

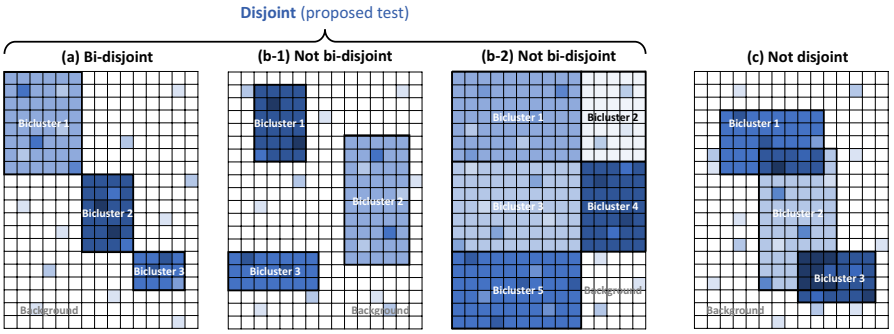

Fig. 1 (**a**) Bi-disjoint, (**b**) disjoint but not bi-disjoint, and (**c**) not disjoint bicluster structures. In the proposed method, we assume that the underlying bicluster structure is disjoint, but not necessarily bi-disjoint. We assume that the observed matrix consists of one or multiple biclusters, in each of which the entries are generated in the i.i.d. sense. Note that in the cases of (b-1) and (**c**), it is not always possible to make all the rows and columns within the biclusters contiguous by sorting the rows and columns

et al. 2011). This method has made it possible to detect a bicluster structure with Type I error control and without the regular-grid constraint. However, unlike the method we propose in this paper, its Type I error has been guaranteed only in terms of an upper bound, not the null distribution of a test statistic. Therefore, in this previous study (Sill et al. 2011), no means is provided to perform a statistical test on the number of biclusters. Moreover, this method has no theoretical guarantee for the alternative cases (i.e., statistical power). In this study, we address these problems by developing a new statistical test on K, which does not require the regular-grid constraint and whose test statistic T shows a good property in an alternative case (i.e., with high probability, T asymptotically increases with the matrix size, and thus, the Type II error converges in probability to zero), as shown in Theorem 2. It must be noted that, unlike information criteria, the general purpose of statistical tests is to determine whether a certain null hypothesis should be rejected or not while allowing a non-zero Type I error, not to guarantee the consistency on the model selection result.

An information criterion on the number of biclusters K. Some studies have proposed that K can be determined based on the minimum description length (Sakai and Yamanishi 2013; Tepper and Sapiro 2016; Yamanishi et al. 2019) and modified DIC for the biclustering problem (Chekouo and Murua 2015; Chekouo et al. 2015). Particularly, under the regular-grid constraint of the bicluster structure, an information criterion called integrated completed likelihood (ICL) has been proposed for determining K (Corneli et al. 2015; Lomet et al. 2012; Wyse et al. 2017), which approximates the maximum marginal likelihood of a given K. These methods aim to select the optimal number of biclusters K from a given set of candidates, in terms of some criterion (e.g., marginal likelihood). This purpose is different from that of

a statistical test, which aims to judge whether we accept a hypothetical number of biclusters K_0 with a specific significance level given by a user.

Other approaches for determining the number of biclusters K. Aside from the above information criterion-based and statistical test-based methods, some studies have proposed the construction of a generative model of the bicluster structure including the number of biclusters K and the subsequent selection of an optimal model in terms of some measure (e.g., choosing a MAP estimator) (Moran 2019; Raff et al. 2020). There have also been some heuristic criteria for determining the number of biclusters in an observed matrix, which have been proposed as stopping rules for top-down division-based biclustering algorithms (Duffy and Quiroz 1991; Hartigan 1972; Tibshirani et al. 1999) or bottom-up merging-based one (Pio et al. 2013).

In this study, we consider the notions of *disjointness* and *bi-disjointness*. A bicluster structure can be called a *disjoint* structure iff each entry belongs to at most one bicluster (as shown in Fig. 1a and b). *Bi-disjointness* has a stricter condition: we call a bicluster structure a *bi-disjoint* structure iff each row or column belongs to at most one bicluster (as shown in Fig. 1a). We develop for the first time a statistical test on K under the assumptions that the underlying bicluster structure is disjoint (but not necessarily bi-disjoint) and that the submatrix localization algorithm is consistent.

To guarantee the asymptotic behavior of the proposed test statistic in the null case (i.e., $K = K_0$), which is given in Theorem 1, we use the properties of a random matrix with a sub-exponential decay (Bloemendal et al. 2016; Pillai and Yin 2014). Moreover, we derive its behavior in the alternative case (i.e., $K > K_0$) such that it increases with the matrix size in high probability, as given in Theorem 2. Unlike a previous study (Watanabe and Suzuki 2021), wherein the number of biclusters K is assumed to be a fixed constant that does not depend on the matrix size m, we consider a case in which K might increase with m (the precise description of this assumption is given in (9)). Additionally, since we consider more general bicluster structures (i.e., without the regular-grid assumption) than in the previous study (Watanabe and Suzuki 2021), we use a different approach to complete the proof in the alternative case. Based on these results, in Sect. 2, we explain a method for estimating K from an observed matrix A, by sequentially testing the hypothetical numbers of biclusters in an ascending order (i.e., $K_0 = 0, 1, 2, \ldots$) until the null hypothesis is accepted.

This paper is organized as follows: in Sect. 2, we describe the problem settings and the model of the underlying bicluster structure in a data matrix. Next, in Sect. 3, we propose a statistical test on the number of biclusters K in a data matrix and derive its theoretical guarantee in both null and alternative cases. In Sect. 4, we provide some experimental results that demonstrate the effectiveness of the proposed test. Finally, in Sect. 5, we discuss the obtained results and limitations of the proposed method. In the supplementary material (Watanabe and Suzuki 2023), we provide the proofs of Theorem 1 and the disjoint submatrix localization algorithms that we use in the experiments.

2 Problem setting and statistical model for goodness-of-fit test for submatrix detection problem

Let $A \in \mathbb{R}^{n\times p}$ be an $n \times p$ observed matrix. Given such an observed matrix A, the goal of submatrix detection problem is to determine whether it contains one or multiple disjoint submatrices, say *biclusters*, in each of which the entries are generated in the i.i.d. sense (Fig. 1a and b). As in the previous studies (Cai et al. 2017; Liu and Guo 2018), we distinguish the submatrix detection from *localization* problems in that the goal of the latter is not only to detect the existence of biclusters in an observed matrix, but to estimate their precise locations.

Let K be the minimum number of such biclusters to represent the matrix A, which is unknown beforehand. Aside from the K biclusters, we assume that there are "background" entries in matrix A that do not belong to any bicluster. Note that the difference between a bicluster and the background is in that the former can be represented as a submatrix (i.e., $\{(i,j) : i \in I_k, j \in J_k\}$ for some sets $I_k \subset \{1,\dots,n\}$ and $J_k \subset \{1,\dots,p\}$), while the latter does not necessarily have such a submatrix structure, as shown in Fig. 1. It must also be noted that multiple bicluster assignments may exist that represent an equivalent bicluster structure. For instance, in a regular-grid bicluster structure (as shown in Fig. 1b-2), any block can be defined as the background. In such cases, the consistency condition that we give later in (v) requires that the probability converges to one with increasing matrix size that an estimated bicluster assignment is included in the set of correct bicluster assignments.

We denote the bicluster index of the (i,j)th entry of matrix A as $g_{ij} \in \{0,1,\dots,K\}$, where $g_{ij} = k$ if the (i,j)th entry belongs to the kth bicluster for some $k \in \{1,\dots,K\}$ and $g_{ij} = 0$ if it belongs to the background. We define the set of group indices of all the entries as $g \equiv (g_{ij})_{1\le i\le n, 1\le j\le p}$. We also define that $\mathcal{I}_k \equiv \{(i,j) : g_{ij} = k\}$, which represents the set of entries in the kth group. Specifically, we consider the following model:

$$\begin{aligned} &P = (P_{ij})_{1\le i\le n,1\le j\le p}, \quad P_{ij} = b_{g_{ij}}. \\ &\sigma = (\sigma_{ij})_{1\le i\le n,1\le j\le p}, \quad \sigma_{ij} = s_{g_{ij}}. \\ &A = (A_{ij})_{1\le i\le n,1\le j\le p}, \quad \mathbb{E}[A_{ij}] = P_{ij}, \quad \mathbb{E}[(A_{ij}-P_{ij})^2] = \sigma_{ij}^2, \end{aligned} \tag{1}$$

where b_k and $s_k > 0$, respectively, are the mean and standard deviation of the kth bicluster ($k = 1,\dots,K$) or background ($k = 0$). This model is a generalized version of well-studied submatrix detection models, in which we assume that the mean of the background noise is zero (i.e., $b_0 = 0$) (Butucea and Ingster 2013; Ma and Wu 2015; Shabalin et al. 2009). Let $Z \in \mathbb{R}^{n\times p}$ be a standardized noise matrix, which is given by

$$Z = (Z_{ij})_{1\le i\le n,1\le j\le p}, \qquad Z_{ij} = \frac{A_{ij}-P_{ij}}{\sigma_{ij}}. \tag{2}$$

In most cases, the number of biclusters K is unknown in advance. This study aims to develop a statistical test on K, which is based on the following null (N) and alternative (A) hypotheses:

$$(N) : K = K_0, \qquad (\mathrm{A}) : K > K_0, \tag{3}$$

where K_0 is a given hypothetical number of biclusters. In this study, we only consider the cases where $K_0 \leq K$. If $K = K_0$ (i.e., the null case), then we call it a *realizable* case. Otherwise (i.e., in the alternative case), we call it an *unrealizable* case. To select the number of biclusters for a given observed matrix A, we propose the sequential testing of the bicluster numbers $K_0 = 0, 1, 2, \ldots$ until the null hypothesis (N) is accepted. Let $\hat{K}$ be the hypothetical number of biclusters when (N) is accepted. The proposed method outputs $\hat{K}$ as the selected number of biclusters in matrix A.

Notations. Throughout this paper, we use the following notations:

$$\begin{aligned}
X_m = O_p\big[f(m)\big] &\Leftrightarrow \forall \epsilon > 0,\ \exists C > 0, M > 0,\ \forall m \geq M, \\
&\quad \Pr\big[Cf(m) \geq |X_m|\big] \geq 1 - \epsilon. \\
X_m = \Omega_p\big[f(m)\big] &\Leftrightarrow \forall \epsilon > 0,\ \exists C > 0, M > 0,\ \forall m \geq M, \\
&\quad \Pr\big[Cf(m) \leq |X_m|\big] \geq 1 - \epsilon. \\
X_m = \Theta_p\big[f(m)\big] &\Leftrightarrow \forall \epsilon > 0,\ \exists C_1, C_2 > 0, M > 0,\ \forall m \geq M, \\
&\quad \Pr\big[C_1 f(m) \leq |X_m| \leq C_2 f(m)\big] \geq 1 - \epsilon.
\end{aligned} \tag{4}$$

$$\|A\|_{\mathrm{op}} = \sup_{\boldsymbol{u} \in \mathbb{R}^p \setminus \mathbf{0}} \frac{\|A\boldsymbol{u}\|}{\|\boldsymbol{u}\|}, \qquad \|A\|_{\mathrm{F}} = \sqrt{\sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{p} A_{ij}^2}. \tag{5}$$

In the proofs in Sect. 3, we use the following **sample** mean matrix $\tilde{P}$ and standard deviation matrix $\tilde{\sigma}$ for the **correct** block structure, and matrix $\tilde{Z}$:

$$\begin{aligned}
\tilde{P} &= (\tilde{P}_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}, & \tilde{P}_{ij} &= \tilde{b}_{g_{ij}}, \\
\tilde{\sigma} &= (\tilde{\sigma}_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}, & \tilde{\sigma}_{ij} &= \tilde{s}_{g_{ij}}, \\
\tilde{Z} &= (\tilde{Z}_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}, & \tilde{Z}_{ij} &= \frac{A_{ij} - \tilde{P}_{ij}}{\tilde{\sigma}_{ij}}.
\end{aligned} \tag{6}$$

where $\tilde{b}_k$ and $\tilde{s}_k$, respectively, are the **sample** mean and standard deviation of the entries in the kth **null** group in observed matrix A. Let $\tilde{\lambda}_1$ and $\tilde{\boldsymbol{v}}_1$, respectively, be the maximum eigenvalue of matrix $\tilde{Z}^\top \tilde{Z}$ and the corresponding eigenvector whose Euclid norm is constrained to be one:

$$\tilde{Z}^\top \tilde{Z} \tilde{\boldsymbol{v}}_1 = \tilde{\lambda}_1 \tilde{\boldsymbol{v}}_1, \quad \|\tilde{\boldsymbol{v}}_1\| = 1. \tag{7}$$

Similarly, we denote the eigenvalues (in descending order) and the corresponding normalized eigenvectors of matrix $Z^\top Z$ as $\{\lambda_j\}$ and $\{\boldsymbol{v}_j\}$, respectively:

$$\begin{aligned} &Z^\top Z \boldsymbol{v}_j = \lambda_j \boldsymbol{v}_j, \quad \|\boldsymbol{v}_j\| = 1, \quad j = 1, \dots, p. \\ &\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p. \end{aligned} \tag{8}$$

Assumptions. In Sect. 3, we develop a statistical test for the number of biclusters in an observed matrix based on the following assumptions:

(i) We assume that the biclusters (and the background) are *disjoint*, that is, there is no entry (i, j) that is assigned to two or more biclusters (i.e., $\mathcal{I}_k \cap \mathcal{I}_{k'} = \emptyset$ if $k \neq k', k, k' \in \{0, 1, \dots, K\}$). Moreover, we assume that each (i, j)th entry belongs to exactly one group.

(ii) We assume that each entry Z_{ij} of the standardized noise matrix has a sub-exponential decay (i.e., there exists some $\vartheta > 0$ such that for $x > 1$, $\Pr\left(\left|Z_{ij}\right| > x\right) \leq \vartheta^{-1} \exp(-x^{\vartheta})$). Furthermore, we assume the following conditions:

- $\max_{k=0,1,\dots,K} s_k = O(1)$, and $\min_{k=0,1,\dots,K} s_k = \Omega(1)$.
- $\max_{k=1,\dots,K,k'=1,\dots,K} \left|b_k - b_{k'}\right| = O(1)$, and $\max_{k=0,1,\dots,K} |b_k| = O(K)$. We also assume that the minimum difference between a pair of different biclusters (including background) is lower bounded by some constant that does not depend on the matrix size: $\min_{k \neq k'} |b_k - b_{k'}| \geq C^{\boldsymbol{b}} > 0$.
- $\check{C} \equiv \max_{i=1,\dots,n, j=1,\dots,p} \mathbb{E}[Z_{ij}^4] = O(1)$.

(iii) We assume that both the row and column sizes n and p of the observed matrix increase in proportion to an integer m (i.e., $n, p \propto m$) and we consider an asymptotics of $m \to \infty$.

(iv) As for the background, we assume that it can be divided into H disjoint submatrices. Regarding the row and column sizes of each kth submatrix, ($|I_k|$ and $|J_k|$, respectively), we assume that they monotonically increase with m, where $k = 1, \dots, K + H$.

- In Theorem 1 for a **realizable** case, we assume that the minimum number of biclusters K and that of background submatrices H to represent the observed matrix A satisfy the following conditions:

$$K + H = O\left(m^{\frac{1}{42} - \epsilon_1}\right), \text{ for some } \epsilon_1 > 0. \tag{9}$$

$$\begin{aligned} n_{\min} &\equiv \min_{k=1,\dots,K+H} |I_k| = \Omega\left(m^{\frac{8}{21}}\right), \\ p_{\min} &\equiv \min_{k=1,\dots,K+H} |J_k| = \Omega\left(m^{\frac{8}{21}}\right). \end{aligned} \tag{10}$$

Note that from these conditions, for some $\epsilon_1 > 0$,

$$(K+H)\left(\min_{k=1,\dots,K+H}|\mathcal{I}_k|\right)^{-\frac{1}{4}} \leq \frac{K+H}{n_{\min}^{\frac{1}{4}}p_{\min}^{\frac{1}{4}}} = O\left(m^{-\frac{1}{6}-\epsilon_1}\right). \tag{11}$$

- In Theorem 2 for an **unrealizable** case, we assume the following stricter condition:

$$\frac{K+H}{\sqrt{n_{\min}p_{\min}}} = O\left(m^{-\frac{3}{4}-\epsilon_2}\right), \text{ for some } \epsilon_2 > 0. \tag{12}$$

(v) In the realizable case, we assume that a submatrix localization algorithm for estimating the bicluster structure g is *consistent*, that is, $\Pr(\hat{g} = g) \to 1$ in the limit of $m \to \infty$, where g and $\hat{g}$, respectively, are the null and estimated bicluster structures of the observed matrix A (the precise definition of $\hat{g}$ is given in Sect. 3).

Assumption (i) indicates the basic problem setting that we have already explained in Sect. 1. In assumption (ii), the sub-exponential decay condition is required for the result (15) from (Pillai and Yin 2014) to hold. From a practical perspective, we can show that various distributions corresponding to different data types (e.g., Gaussian, Bernoulli, and Poisson distributions) satisfy such a condition. The upper and lower bounds regarding the mean and standard deviation parameters and the assumption on $\check{C}$ are mainly used to guarantee the asymptotic power of the proposed test, which is given in Theorem 2. It must be noted that under the assumption of some distributions (e.g., Bernoulli distribution), most conditions [i.e., $\max_{k=0,1,\dots,K} s_k = O(1)$, $\max_{k=1,\dots,K,k'=1,\dots,K}|b_k - b_{k'}| = O(1)$, $\max_{k=0,1,\dots,K}|b_k| = O(K)$, and $\check{C} = O(1)$] are always satisfied. Assumption (iii) indicates the asymptotic framework based on which we derive the behavior of the proposed test statistic in Theorems 1 and 2. In Sect. 5, we discuss the possibility of extension to the non-asymptotic setting. Assumption (iv) regarding the number of submatrices and the minimum submatrix size is used to prove Theorems 1 and 2, and the specific number in the order of each variable [e.g., $\frac{1}{42}$ in (9)] is determined by first defining it as some variable and then deriving its specific value so that all the inequalities in the proofs hold without contradiction. The consistency condition in submatrix localization, which is defined in assumption (v), has been considered by several previous studies, although most of them have assumed the existence of at most one bicluster in a given observed matrix (Balakrishnan et al. 2011; Brennan et al. 2018; Butucea et al. 2015; Hajek et al. 2017, 2018; Kolar et al. 2011; Luo and Zhang 2020). Multiple biclusters may be localized by applying these methods to a given observed matrix multiple times; however, there is no guarantee for the consistency of such a heuristic approach. For bi-disjoint bicluster structures, some submatrix localization algorithms based on the maximum likelihood estimator for a known model parameter (Chen and Xu 2016) and singular value decomposition (Cai et al. 2017) have been shown to be consistent. As another method, by imposing the regular-grid constraint to the underlying bicluster structure, we can consider a special case of non-bi-disjoint bicluster structures,

which can be represented as a result of row-column clustering (as shown in Fig. 1b-2). As for the biclustering problem with such a regular-grid structure, Flynn and Perry (Flynn and Perry 2020) have proposed a consistent algorithm based on the criterion of (generalized) profile likelihood. However, these algorithms cannot be directly applied to our case, where the localization problem cannot be formulated as row-column (hard) clustering (Shabalin et al. 2009). Although the proposed test itself can be applied without the bi-disjoint assumption, currently there is no way to consistently estimate non-bi-disjoint bicluster structures. Instead, we propose a heuristic submatrix localization algorithm in Appendix D in the supplementary material (Watanabe and Suzuki 2023) and use it the experiments. To develop a consistent submatrix localization algorithm that can be applied without the bi-disjoint assumption is beyond the scope of this paper.

3 A test statistic for determining the number of biclusters

We develop the test statistic T of the proposed test based on the estimated version of the standardized noise matrix Z in (2), given a hypothetical number of biclusters K_0. We denote the **estimated** group index of the (i, j)th entry of matrix A as $\hat{g}_{ij} \in \{0, 1, \dots, K_0\}$, where $\hat{g}_{ij} = k$ if the (i, j)th entry is estimated to be a member of the kth bicluster for some k and $\hat{g}_{ij} = 0$ otherwise (i.e., the (i, j)th entry is estimated to be a member of background). We define the set of **estimated** group indices of all the entries as $\hat{g} \equiv (\hat{g}_{ij})_{1\le i\le n, 1\le j\le p}$. We also define that $\hat{\mathcal{I}}_k \equiv \{(i,j) : \hat{g}_{ij} = k\}$, which represents the **estimated** set of entries in the kth group.

Based on the above notations, the **estimated** mean, standard deviation, and noise matrices $\hat{P}$, $\hat{\sigma}$ and $\hat{Z}$ are given by

$$\begin{aligned}
&\hat{\boldsymbol{b}} = (\hat{b}_k)_{1\le k\le K_0}, && \hat{b}_k = \frac{1}{|\hat{\mathcal{I}}_k|}\sum_{(i,j)\in\hat{\mathcal{I}}_k} A_{ij},\\
&\hat{P} = (\hat{P}_{ij})_{1\le i\le n, 1\le j\le p}, && \hat{P}_{ij} = \hat{b}_{\hat{g}_{ij}},\\
&\hat{\boldsymbol{s}} = (\hat{s}_k)_{1\le k\le K_0}, && \hat{s}_k = \sqrt{\frac{1}{|\hat{\mathcal{I}}_k|}\sum_{(i,j)\in\hat{\mathcal{I}}_k}\left(A_{ij} - \hat{P}_{ij}\right)^2},\\
&\hat{\sigma} = (\hat{\sigma}_{ij})_{1\le i\le n, 1\le j\le p}, && \hat{\sigma}_{ij} = \hat{s}_{\hat{g}_{ij}}.
\end{aligned} \tag{13}$$

$$\hat{Z} = (\hat{Z}_{ij})_{1\le i\le n, 1\le j\le p}, \qquad \hat{Z}_{ij} = \frac{A_{ij} - \hat{P}_{ij}}{\hat{\sigma}_{ij}}. \tag{14}$$

To construct a statistical test on the number of biclusters K, we use the following result from (Pillai and Yin 2014), which shows that the scaled maximum eigenvalue T^* of sample covariance matrix $Z^\top Z$ converges in law to the Tracy–Widom distribution with index 1 (TW_1) in the limit of $m \to \infty$:

$$T^* = \frac{\lambda_1 - a^{\mathrm{TW}}}{b^{\mathrm{TW}}}, \quad T^* \rightsquigarrow TW_1 \text{ (Convergence in law)}, \tag{15}$$

where λ_1 is the maximum eigenvalue of matrix $Z^\top Z$ and

$$a^{\mathrm{TW}} = \left(\sqrt{n} + \sqrt{p}\right)^2, \quad b^{\mathrm{TW}} = \left(\sqrt{n} + \sqrt{p}\right)\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{p}}\right)^{\frac{1}{3}}. \tag{16}$$

Based on the above fact, we define the test statistic T, which is an estimator of T^* in (15), from the maximum eigenvalue $\hat{\lambda}_1$ of matrix $\hat{Z}^\top \hat{Z}$:

$$T = \frac{\hat{\lambda}_1 - a^{\mathrm{TW}}}{b^{\mathrm{TW}}}. \tag{17}$$

By using the test statistic T, we define the rule of the proposed test at the significance level of α as follows:

$$\text{Reject null hypothesis } (K = K_0), \quad \text{if } T \geq t(\alpha), \tag{18}$$

where $t(\alpha)$ is the α upper quantile of the TW_1 distribution. We give the theoretical guarantees for the above test in both the null and alternative cases in Theorems 1 and 2, respectively.

Theorem 1 (Realizable case) *Under the assumptions in Sect. 2, if* $K = K_0$,

$$T \rightsquigarrow TW_1 \text{ (Convergence in law)}, \tag{19}$$

in the limit of $m \to \infty$, *where T is defined as in* (17).

Proof To apply the result reported in (Pillai and Yin 2014), we consider the difference between T^* and T. By definitions of (15) and (17), we have

$$|T - T^*| = \frac{|\hat{\lambda}_1 - \lambda_1|}{b^{\mathrm{TW}}}. \tag{20}$$

Next, we prove that the right side of (20) can be bounded by $\frac{|\lambda_1 - \hat{\lambda}_1|}{b^{\mathrm{TW}}} = O_p(m^{-\epsilon})$ for some $\epsilon > 0$, which is given in Lemma 3. If this bound holds, from Slutsky's theorem, (19) also holds. To show Lemma 3, we first state the following Lemmas 1 and 2, which give the lower and upper bounds for the maximum eigenvalue $\tilde{\lambda}_1$ of matrix $\tilde{Z}^\top \tilde{Z}$. The proofs of the following lemmas are mainly based on those given in (Watanabe and Suzuki 2021). The main differences between them are as follows: first, we assume a regular-grid bicluster structure in the previous study, whereas we consider a more generalized disjoint one. Second, unlike the previous study, where we assume that the null number of biclusters K is a fixed constant that does not depend on the matrix size m, we consider a case in which K might increase with m. □

Lemma 1 *Under the assumptions noted in Sect. 2, if* $K = K_0$,

$$\lambda_1 \leq \tilde{\lambda}_1 + O_p\left(m^{\frac{1}{3}-\epsilon}\right), \quad \text{for some } \epsilon > 0. \tag{21}$$

Proof A proof is given in Appendix A in the supplementary material (Watanabe and Suzuki 2023). □

Lemma 2 *Under the assumptions in Sect. 2, if* $K = K_0$,

$$\tilde{\lambda}_1 \leq \lambda_1 + O_p\left(m^{\frac{1}{3}-\epsilon}\right), \quad \text{for some } \epsilon > 0. \tag{22}$$

Proof A proof is given in Appendix C in the supplementary material (Watanabe and Suzuki 2023). □

Lemma 3 *Under the assumptions in Sect. 2, if* $K = K_0$,

$$\frac{|\lambda_1 - \hat{\lambda}_1|}{b^{\mathrm{TW}}} = O_p(m^{-\epsilon}), \quad \text{for some } \epsilon > 0. \tag{23}$$

Proof By combining Lemmas 1, 2, and the definition of b^{TW} in (16), we have

$$\frac{|\lambda_1 - \tilde{\lambda}_1|}{b^{\mathrm{TW}}} = O_p(m^{-\epsilon}), \quad \text{for some } \epsilon > 0. \tag{24}$$

We consider the following three events:

- $\mathcal{E}_m^{(1)}$ represents the event that $\tilde{Z} = \hat{Z}$ holds.
- $\mathcal{E}_m^{(2)}$ represents the event that the solution given by the submatrix localization algorithm is correct (i.e., $\hat{g} = g$).
- $\mathcal{E}_{m,C}^{(3)}$ represents the event that $\frac{|\lambda_1 - \tilde{\lambda}_1|}{b^{\mathrm{TW}}} \leq Cm^{-\epsilon}$ holds.

The joint probability of events $\mathcal{E}_m^{(1)}$ and $\mathcal{E}_{m,C}^{(3)}$ can be lower bounded by

$$\begin{aligned} \Pr\left(\mathcal{E}_m^{(1)} \cap \mathcal{E}_{m,C}^{(3)}\right) &\geq \Pr\left(\mathcal{E}_m^{(2)} \cap \mathcal{E}_{m,C}^{(3)}\right) \\ &\geq 1 - \Pr\left[\left(\mathcal{E}_m^{(2)}\right)^{\mathrm{C}}\right] - \Pr\left[\left(\mathcal{E}_{m,C}^{(3)}\right)^{\mathrm{C}}\right], \end{aligned} \tag{25}$$

where $\mathcal{E}^{\mathrm{C}}$ is the complement of event $\mathcal{E}$. From the consistency assumption (v) in Sect. 2, if $K = K_0$, the second term on the right side of (25) satisfies that $\Pr\left[\left(\mathcal{E}_m^{(2)}\right)^{\mathrm{C}}\right] \to 0$ in the limit of $m \to \infty$. As for the third term $\Pr\left[\left(\mathcal{E}_{m,C}^{(3)}\right)^{\mathrm{C}}\right]$, we already have (24). By combining these facts, we have

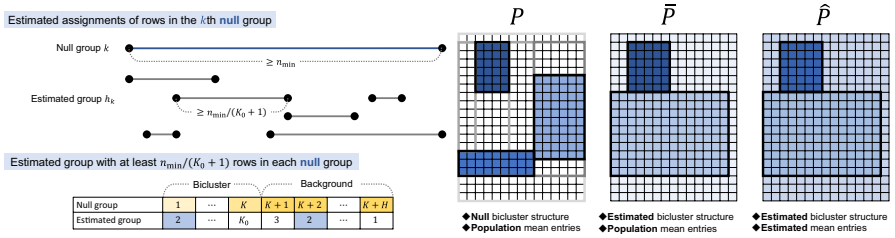

Fig. 2 Definition of matrices P, $\bar{P}$, and $\hat{P}$ in an unrealizable case

$$\forall \tilde{\epsilon} > 0,\ \exists C > 0, M > 0,\ \forall m \geq M,\ \Pr\left(\mathcal{E}_m^{(1)} \cap \mathcal{E}_{m,C}^{(3)}\right) \geq 1 - \tilde{\epsilon}, \tag{26}$$

which results in (23). □

From Lemma 3, we finally obtain

$$|T - T^*| = \frac{|\hat{\lambda}_1 - \lambda_1|}{b^{\mathrm{TW}}} = O_p(m^{-\epsilon}),\ \text{for some } \epsilon > 0. \tag{27}$$

By combining this fact with Slutsky's theorem, the convergence of the test statistic T in law to TW_1 distribution in (19) holds. □

Theorem 2 (Unrealizable case) *Under the assumptions in Sect. 2, if* $K > K_0$,

$$T = O_p\left(m^{\frac{5}{3}}\right), \tag{28}$$

and

$$T = \Omega_p\left(m^{\frac{2}{3}}\right), \tag{29}$$

where T is defined as in (17).

Proof We first prove the upper bound in (28). Let $\underline{X}^{\mathrm{E}(k)}$ be an $n \times p$ matrix whose entries in the kth **estimated** bicluster (including background) are the same as matrix X and all the other entries are zero. Since the Frobenius norm upper bounds the operator norm,

$$\begin{aligned} \|\hat{Z}\|_{\mathrm{op}} \leq \|\hat{Z}\|_{\mathrm{F}} &= \sqrt{\sum_{k=0}^{K_0} \|\underline{\hat{Z}}^{\mathrm{E}(k)}\|_{\mathrm{F}}^2} = \sqrt{\sum_{k=0}^{K_0} \frac{1}{\hat{s}_k^2} \|\underline{A}^{\mathrm{E}(k)} - \underline{\hat{P}}^{\mathrm{E}(k)}\|_{\mathrm{F}}^2} \\ &= \sqrt{\sum_{k=0}^{K_0} \frac{|\hat{\mathcal{I}}_k|}{\|\underline{A}^{\mathrm{E}(k)} - \underline{\hat{P}}^{\mathrm{E}(k)}\|_{\mathrm{F}}^2} \|\underline{A}^{\mathrm{E}(k)} - \underline{\hat{P}}^{\mathrm{E}(k)}\|_{\mathrm{F}}^2} = \sqrt{\sum_{k=0}^{K_0} |\hat{\mathcal{I}}_k|} = \sqrt{np}. \end{aligned} \tag{30}$$

From the assumption (iii), let $n = C_n m$ and $p = C_p m$, where C_n and C_p are positive constants. According to the definition in (16), we have

$$a^{\mathrm{TW}} = \left(\sqrt{C_n} + \sqrt{C_p}\right)^2 m, \quad b^{\mathrm{TW}} = \left(\sqrt{C_n} + \sqrt{C_p}\right)\left(\frac{1}{\sqrt{C_n}} + \frac{1}{\sqrt{C_p}}\right)^{\frac{1}{3}} m^{\frac{1}{3}}. \tag{31}$$

By substituting (30) and (31) into (17), we have

$$T = \frac{\|\hat{Z}\|_{\mathrm{op}}^2 - a^{\mathrm{TW}}}{b^{\mathrm{TW}}} \leq \frac{np - a^{\mathrm{TW}}}{b^{\mathrm{TW}}} = O_p\left(m^{\frac{5}{3}}\right), \tag{32}$$

which concludes the proof of (28).

We next show the lower bound in (29). As shown in Fig. 2, we define that $\bar{P}$ is a matrix that consists of the **estimated** bicluster structure and entries of the **population** means. For instance, if the (i, j)th entry of observed matrix A belongs to an estimated bicluster that consists of $n^{(1)}$ entries of the k_1th null bicluster and $n^{(2)}$ entries of the k_2th null bicluster, its value in matrix $\bar{P}$ is given by $\bar{P}_{ij} = \left(n^{(1)} b_{k_1} + n^{(2)} b_{k_2}\right) / \left(n^{(1)} + n^{(2)}\right)$.

From the assumption (iv), for all $k \in \{1, \ldots, K + H\}$, the kth **null** submatrix (i.e., bicluster or background submatrix) has a size of at least $n_{\min} \times p_{\min}$. Therefore, for all **null** group index $k \in \{1, \ldots, K + H\}$, there exists at least one **estimated** group $h_k \in \{0, 1, \ldots, K_0\}$ that contains a submatrix of the kth **null** submatrix with the size of $\frac{n_{\min}}{K_0+1} \times \frac{p_{\min}}{K_0+1}$ or more (Fig. 2). Since $K_0 < K$, there exists at least one set of **null** group indices (k_1, k_2), $k_1, k_2 \in \{1, \ldots, K + H\}$ that satisfies $h_{k_1} = h_{k_2}$, $k_1 \neq k_2$, and $k_1 \in \{1, \ldots, K\}$ (i.e., a pair of mutually different **null** groups that belong to the same group in the **estimated** bicluster structure). In other words, there exists at least one **estimated** group $\bar{k}$ $(= h_{k_1} = h_{k_2})$ that satisfies the following conditions:

- The $\bar{k}$th **estimated** group contains two submatrices. We denote the sets of entries in these two submatrices as $\mathcal{I}^{(1)}$ and $\mathcal{I}^{(2)}$.
- The set of entries $\mathcal{I}^{(1)}$ forms a submatrix of the k_1th **null** group $(k_1 \in \{1, \ldots, K\})$. The size of submatrix $\mathcal{I}^{(1)}$ is at least $\frac{n_{\min}}{K_0+1} \times \frac{p_{\min}}{K_0+1}$.
- The set of entries $\mathcal{I}^{(2)}$ forms a submatrix of the k_2th **null** group $(k_2 \in \{0, 1, \ldots, K\})$. The size of submatrix $\mathcal{I}^{(2)}$ is at least $\frac{n_{\min}}{K_0+1} \times \frac{p_{\min}}{K_0+1}$.
- The **null** groups of $\mathcal{I}^{(1)}$ and $\mathcal{I}^{(2)}$ are mutually different (i.e., $k_1 \neq k_2$).

The population means of submatrices $\mathcal{I}^{(1)}$ and $\mathcal{I}^{(2)}$, respectively, are b_{k_1} and b_{k_2}. We assume $b_{k_1} > b_{k_2}$ without loss of generality. Let $\bar{b}$ be the constant value of the submatrices $\mathcal{I}^{(1)}$ and $\mathcal{I}^{(2)}$ in matrix $\bar{P}$. Here, we consider the following two patterns:

- If $\bar{b} \geq (b_{k_1} + b_{k_2})/2$, we have $|\bar{b} - b_{k_2}| \geq (b_{k_1} - b_{k_2})/2$.
- If $\bar{b} < (b_{k_1} + b_{k_2})/2$, we have $|\bar{b} - b_{k_1}| > (b_{k_1} - b_{k_2})/2$.

Therefore, for any case, there exists a submatrix $\mathcal{I}$ that satisfies the following two conditions (note that $\mathcal{I} = \mathcal{I}^{(2)}$ in the former case and $\mathcal{I} = \mathcal{I}^{(1)}$ in the latter case):

- The size of submatrix $\mathcal{I}$ is at least $\frac{n_{\min}}{K_0+1} \times \frac{p_{\min}}{K_0+1}$.
- Let $b_{\mathcal{I}}$ and $\bar{b}_{\mathcal{I}}$, respectively, be the constant values of submatrix $\mathcal{I}$ in matrices P and $\bar{P}$. Note that we have $\bar{b}_{\mathcal{I}} = \bar{b}_{\bar{k}}$. From the assumption (ii), the following inequality holds:

$$|\bar{b}_{\mathcal{I}} - b_{\mathcal{I}}| \geq \min_{k \neq k'} |b_k - b_{k'}|/2 \geq \frac{C^{\boldsymbol{b}}}{2}. \tag{33}$$

The difference between $\hat{b}_{\bar{k}}$ and $\bar{b}_{\bar{k}}$ is given by

$$\begin{aligned} |\hat{b}_{\bar{k}} - \bar{b}_{\bar{k}}| &= \frac{1}{|\hat{\mathcal{I}}_{\bar{k}}|} \left| \sum_{(i,j) \in \hat{\mathcal{I}}_{\bar{k}}} \left(\hat{P}_{ij} - \bar{P}_{ij} \right) \right| = \frac{1}{|\hat{\mathcal{I}}_{\bar{k}}|} \left| \sum_{(i,j) \in \hat{\mathcal{I}}_{\bar{k}}} \left(A_{ij} - P_{ij} \right) \right| \\ &\leq \frac{\max_{k=0,1,\ldots,K} s_k}{|\hat{\mathcal{I}}_{\bar{k}}|} \left| \sum_{(i,j) \in \hat{\mathcal{I}}_{\bar{k}}} Z_{ij} \right|. \end{aligned} \tag{34}$$

To derive the upper bound of the right side of (34), we cannot take the same strategy as in the previous study (Watanabe and Suzuki 2021), since it uses the assumption that the $\bar{k}$th estimated group can always be represented as a submatrix and it does not consider a general background structure. Therefore, we adopt an alternative approach to use the Lyapunov variant of the central limit theorem. Here, Z_{ij} independently follows a distribution with zero mean and unit variance. From the sub-exponential condition (ii), for any $\check{n} \in \mathbb{N}$, we have $\mathbb{E}[Z_{ij}^{\check{n}}] < \infty$. Let $\mathcal{I}^{\mathrm{CLT}}$ be a subset of entries in a $n \times p$ matrix. By defining $\delta \equiv 2$ and from the assumption (ii), the following Lyapunov's condition holds:

$$\begin{aligned} &\lim_{|\mathcal{I}^{\mathrm{CLT}}| \to \infty} \frac{1}{|\mathcal{I}^{\mathrm{CLT}}|^{1+\frac{1}{2}\delta}} \sum_{(i,j) \in \mathcal{I}^{\mathrm{CLT}}} \mathbb{E}\left[|Z_{ij}|^{2+\delta} \right] \leq \lim_{|\mathcal{I}^{\mathrm{CLT}}| \to \infty} \frac{\check{C} |\mathcal{I}^{\mathrm{CLT}}|}{|\mathcal{I}^{\mathrm{CLT}}|^2} \\ &= \lim_{|\mathcal{I}^{\mathrm{CLT}}| \to \infty} \frac{\check{C}}{|\mathcal{I}^{\mathrm{CLT}}|} = 0. \end{aligned} \tag{35}$$

Therefore, from the Lyapunov variant of the central limit theorem,

$$\frac{1}{\sqrt{|\mathcal{I}^{\mathrm{CLT}}|}} \sum_{(i,j) \in \mathcal{I}^{\mathrm{CLT}}} Z_{ij} \rightsquigarrow N(0,1). \tag{36}$$

From (34), Prokhorov's theorem (van der Vaart 1998), and the fact that $|\hat{\mathcal{I}}_{\bar{k}}| \geq |\mathcal{I}| \geq n_{\min} p_{\min}/(K_0+1)^2$ (note that the right side monotonically increases with the matrix size m from the assumption (iv)), we have

$$\begin{aligned} |\hat{b}_{\bar{k}} - \bar{b}_{\bar{k}}| &\leq \frac{\max_{k=0,1,\ldots,K} s_k}{\sqrt{|\hat{\mathcal{I}}_{\bar{k}}|}} O_p(1) \leq \left(\max_{k=0,1,\ldots,K} s_k \right) \frac{K_0+1}{\sqrt{n_{\min} p_{\min}}} O_p(1) \\ &\leq O_p\left(\frac{K+H}{\sqrt{n_{\min} p_{\min}}} \right) \quad (\because K_0 + 1 < K + H \text{ and assumption (ii)}). \end{aligned} \tag{37}$$

From (37), we have

$$\left| |b_{\mathcal{I}} - \bar{b}_{\mathcal{I}}| - |b_{\mathcal{I}} - \hat{b}_{\bar{k}}| \right| \leq |\hat{b}_{\bar{k}} - \bar{b}_{\bar{k}}| \leq O_p\left(\frac{K+H}{\sqrt{n_{\min} p_{\min}}} \right). \tag{38}$$

By combining this result with (33),

$$\frac{C^{\boldsymbol{b}}}{2} \leq |b_{\mathcal{I}} - \hat{b}_{\bar{k}}| + O_p\left(\frac{K+H}{\sqrt{n_{\min} p_{\min}}} \right). \tag{39}$$

Therefore, from (12) in the assumption (iv), we have

$$|b_{\mathcal{I}} - \hat{b}_{\bar{k}}| = \Omega_p(1). \tag{40}$$

Let $X^{\mathcal{I}}$ be a submatrix of X with the set of entries $\mathcal{I}$. Since the operator norm of a submatrix is not larger than that of the original matrix,

$$\|\hat{Z}\|_{\mathrm{op}} \geq \|\hat{Z}^{\mathcal{I}}\|_{\mathrm{op}} \geq \frac{1}{\hat{s}_{\bar{k}}} \left| \|A^{\mathcal{I}} - P^{\mathcal{I}}\|_{\mathrm{op}} - \|P^{\mathcal{I}} - \hat{P}^{\mathcal{I}}\|_{\mathrm{op}} \right|. \tag{41}$$

Let $\bar{k}^{\mathrm{N}}$ be the **null** bicluster index (including background) of submatrix $\mathcal{I}$. Note that $b_{\mathcal{I}} = b_{\bar{k}^{\mathrm{N}}}$. As for the first term in (41), from the assumption (ii), we have

$$\begin{aligned} \|A^{\mathcal{I}} - P^{\mathcal{I}}\|_{\mathrm{op}} = s_{\bar{k}^{\mathrm{N}}} \|Z^{\mathcal{I}}\|_{\mathrm{op}} \leq s_{\bar{k}^{\mathrm{N}}} \|Z\|_{\mathrm{op}} &\leq \left(\max_{k=0,1,\ldots,K} s_k \right) O_p(\sqrt{m}) \\ &= O_p(\sqrt{m}). \end{aligned} \tag{42}$$

In regard to the second term in (41), since all the entries in matrix $(P^{\mathcal{I}} - \hat{P}^{\mathcal{I}})$ is $(b_{\mathcal{I}} - \hat{b}_{\bar{k}})$ and thus its rank is one, we have

$$\begin{aligned} \|P^{\mathcal{I}} - \hat{P}^{\mathcal{I}}\|_{\mathrm{op}} = \|P^{\mathcal{I}} - \hat{P}^{\mathcal{I}}\|_{\mathrm{F}} = \sqrt{|\mathcal{I}|(b_{\mathcal{I}} - \hat{b}_{\bar{k}})^2} &\geq \frac{\sqrt{n_{\min} p_{\min}}}{K_0+1} |b_{\mathcal{I}} - \hat{b}_{\bar{k}}| \\ &\geq \frac{\sqrt{n_{\min} p_{\min}}}{K+H} |b_{\mathcal{I}} - \hat{b}_{\bar{k}}| = \Omega_p\left(\frac{\sqrt{n_{\min} p_{\min}}}{K+H} \right). \end{aligned} \tag{43}$$

To derive the last equation, we used the assumption (iv) and (40).

Finally, we can derive an upper bound of $\hat{s}_{\bar{k}}$ by

$$\begin{aligned}
\hat{s}_{\bar{k}} &= \frac{1}{\sqrt{|\hat{\mathcal{I}}_{\bar{k}}|}} \|\underline{A}^{\mathrm{E}(\bar{k})} - \hat{\underline{P}}^{\mathrm{E}(\bar{k})}\|_{\mathrm{F}} \leq \frac{1}{\sqrt{|\hat{\mathcal{I}}_{\bar{k}}|}} \left(\|A - P\|_{\mathrm{F}} + \|\underline{P}^{\mathrm{E}(\bar{k})} - \hat{\underline{P}}^{\mathrm{E}(\bar{k})}\|_{\mathrm{F}} \right) \\
&= \frac{1}{\sqrt{|\hat{\mathcal{I}}_{\bar{k}}|}} \left(\sqrt{\sum_{k=1}^{K+H} s_k^2 \|Z^{\mathrm{N}(k)}\|_{\mathrm{F}}^2} + \|\underline{P}^{\mathrm{E}(\bar{k})} - \hat{\underline{P}}^{\mathrm{E}(\bar{k})}\|_{\mathrm{F}} \right) \\
&\leq \frac{1}{\sqrt{|\hat{\mathcal{I}}_{\bar{k}}|}} \left[\left(\max_{k=0,1,\ldots,K} s_k \right) \|Z\|_{\mathrm{F}} + \|\underline{P}^{\mathrm{E}(\bar{k})} - \hat{\underline{P}}^{\mathrm{E}(\bar{k})}\|_{\mathrm{F}} \right] \\
&= \frac{1}{\sqrt{|\hat{\mathcal{I}}_{\bar{k}}|}} \left[\left(\max_{k=0,1,\ldots,K} s_k \right) \|Z\|_{\mathrm{F}} + \sqrt{\sum_{(i,j)\in\hat{\mathcal{I}}_{\bar{k}}} \left(P_{ij} - \hat{b}_{\bar{k}}\right)^2} \right] \\
&\leq \frac{K+H}{\sqrt{n_{\min} p_{\min}}} \left(\max_{k=0,1,\ldots,K} s_k \right) \|Z\|_{\mathrm{F}} + \max_{k=0,1,\ldots,K} \left| b_k - \hat{b}_{\bar{k}} \right|.
\end{aligned} \tag{44}$$

The second term in (44) can be upper bounded as follows:

$$\begin{aligned}
&\max_{k=0,1,\ldots,K} \left| b_k - \hat{b}_{\bar{k}} \right| \leq |\hat{b}_{\bar{k}}| + \max_{k=0,1,\ldots,K} |b_k| \\
&= \frac{1}{|\hat{\mathcal{I}}_{\bar{k}}|} \left| \sum_{(i,j)\in\hat{\mathcal{I}}_{\bar{k}}} (\sigma_{ij} Z_{ij} + P_{ij}) \right| + \max_{k=0,1,\ldots,K} |b_k| \\
&\leq \frac{1}{|\hat{\mathcal{I}}_{\bar{k}}|} \left(\max_{k=0,1,\ldots,K} s_k \right) \sum_{(i,j)\in\hat{\mathcal{I}}_{\bar{k}}} \left| Z_{ij} \right| + 2 \max_{k=0,1,\ldots,K} |b_k| \\
&\leq \frac{1}{|\hat{\mathcal{I}}_{\bar{k}}|} \left(\max_{k=0,1,\ldots,K} s_k \right) \sqrt{\sum_{(i,j)\in\hat{\mathcal{I}}_{\bar{k}}} \left| Z_{ij} \right|^2} \sqrt{|\hat{\mathcal{I}}_{\bar{k}}|} + 2 \max_{k=0,1,\ldots,K} |b_k| \\
&\leq \frac{1}{\sqrt{|\hat{\mathcal{I}}_{\bar{k}}|}} \left(\max_{k=0,1,\ldots,K} s_k \right) \|Z\|_{\mathrm{F}} + 2 \max_{k=0,1,\ldots,K} |b_k| \\
&\leq \frac{K+H}{\sqrt{n_{\min} p_{\min}}} \left(\max_{k=0,1,\ldots,K} s_k \right) \|Z\|_{\mathrm{F}} + 2 \max_{k=0,1,\ldots,K} |b_k|.
\end{aligned} \tag{45}$$

Since $\mathbb{E}[Z_{ij}^2] = \mathbb{V}[Z_{ij}] + \mathbb{E}[Z_{ij}]^2 = 1$ and $\mathbb{V}[Z_{ij}^2] = \mathbb{E}[Z_{ij}^4] - \mathbb{E}[Z_{ij}^2]^2 = \mathbb{E}[Z_{ij}^4] - 1 < \infty$ from the assumption (ii), from the central limit theorem and Prokhorov's theorem (van der Vaart 1998), we have

$$\frac{1}{\sqrt{np}} \sum_{i=1}^{n} \sum_{j=1}^{p} (Z_{ij}^2 - 1) = O_p(1) \iff \sum_{i=1}^{n} \sum_{j=1}^{p} Z_{ij}^2 = \|Z\|_{\mathrm{F}}^2 = np + O_p(m), \tag{46}$$

which results in that

$$\|Z\|_{\mathrm{F}} = O_p(m). \tag{47}$$

By substituting (45) and (47) into (44), and using the assumption (ii),

$$\hat{s}_{\bar{k}} \leq \frac{K+H}{\sqrt{n_{\min}p_{\min}}} O_p(m) + O(K) = O_p\left[\frac{(K+H)m}{\sqrt{n_{\min}p_{\min}}}\right]. \tag{48}$$

By substituting (42), (43), and (48) into (41), and using (12) in the assumption (iv), we finally have

$$\begin{aligned} \|\hat{Z}\|_{\mathrm{op}} &\geq \Omega_p\left[\frac{\sqrt{n_{\min}p_{\min}}}{(K+H)m}\right] \left| \Omega_p\left(\frac{\sqrt{n_{\min}p_{\min}}}{K+H}\right) - O_p(\sqrt{m}) \right| \\ &= \Omega_p\left[\frac{n_{\min}p_{\min}}{(K+H)^2 m}\right] = \Omega_p\left(m^{\frac{1}{2}+2\epsilon_2}\right) \\ \iff \|\hat{Z}\|_{\mathrm{op}}^2 &= \Omega_p\left(m^{1+4\epsilon_2}\right), \text{ for some } \epsilon_2 > 0. \end{aligned} \tag{49}$$

By substituting (49) and (31) into (17), we have

$$T = \frac{\|\hat{Z}\|_{\mathrm{op}}^2 - a^{\mathrm{TW}}}{b^{\mathrm{TW}}} = \Omega_p\left(m^{\frac{2}{3}+4\epsilon_2}\right) \geq \Omega_p\left(m^{\frac{2}{3}}\right), \tag{50}$$

which concludes the proof. □

4 Experiments

As we explained in Sect. 2, currently, we do not have any consistent submatrix localization method that can be applied to general disjoint block structure. Instead, in all the following experiments, we used Algorithm 2 in Appendix D in the supplementary material (Watanabe and Suzuki 2023) for estimating the bicluster structure of a given matrix, although it is not guaranteed to be consistent, and thus, Theorem 1 does not necessarily hold with it.

4.1 The convergence of test statistic T in law to TW_1 distribution in the realizable case

We first checked the asymptotic behavior of the proposed test statistic T in the null case (i.e., $K = K_0$) by using synthetic data matrices, which were generated from the Gaussian, Bernoulli, and Poisson distributions. In this case, from Theorem 1, T converges in law to the TW_1 distribution in the limit of $m \to \infty$.

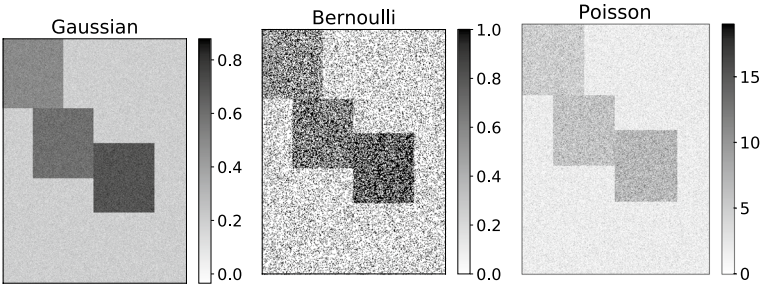

Fig. 3 Examples of the observed data matrices. The left, center, and right figures show the Gaussian, Bernoulli, and Poisson cases, respectively

We set the null number of biclusters at $K = 3$ in all the settings of distributions, and tried 10 sets of matrix sizes: $(n, p) = (500 \times i, 375 \times i)$, $i = 1, \ldots, 10$. For each distribution, we defined the null set of parameters and the relative entropy function f of the generalized profile likelihood as follows. These experimental settings of the relative entropy function f follow those in (Flynn and Perry 2020), and they are based on the framework of profile-likelihood maximization in Gaussian, Bernoulli, and Poisson models, as shown in Appendix D in the supplementary material (Watanabe and Suzuki 2023).

- **Gaussian LBM (G-LBM)**: Each entry in the kth group $(k = 0, 1, \ldots, K)$ of observed matrix A was generated independently from the Gaussian distribution $\mathcal{N}(b_k, s_k)$, where

$$\boldsymbol{b} = \begin{pmatrix} 0.2 & 0.5 & 0.6 & 0.7 \end{pmatrix}^\top, \ \boldsymbol{s} = 0.05 \times \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix}^\top. \tag{51}$$

$$f(x) \equiv x^2/2. \tag{52}$$

- **Bernoulli LBM (B-LBM)**: Each entry in the kth group of observed matrix A was generated independently from the Bernoulli distribution Bernoulli(b_k), where

$$\boldsymbol{b} = \begin{pmatrix} 0.2 & 0.5 & 0.6 & 0.7 \end{pmatrix}^\top. \tag{53}$$

$$f(x) \equiv x \log\left(\max\{x, 10^{-5}\}\right) + (1 - x) \log\left(\max\{1 - x, 10^{-5}\}\right). \tag{54}$$

- **Poisson LBM (P-LBM)**: Each entry in the kth group of observed matrix A was generated independently from the Poisson distribution Pois(b_k), where

$$\boldsymbol{b} = \begin{pmatrix} 2 & 5 & 6 & 7 \end{pmatrix}^\top. \tag{55}$$

$$f(x) \equiv x \log\left(\max\{x, 10^{-5}\}\right) - x. \tag{56}$$

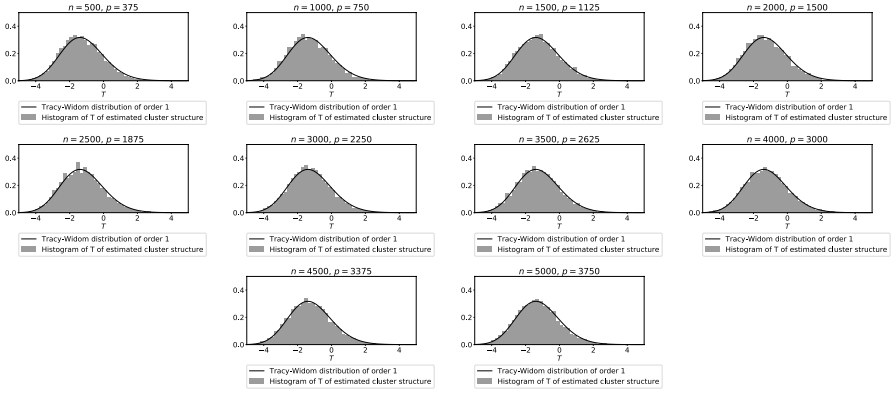

Fig. 4 Histogram of the proposed test statistic T, which was computed with **estimated** bicluster structure (**G-LBM**). The titles of the figures show different matrix sizes

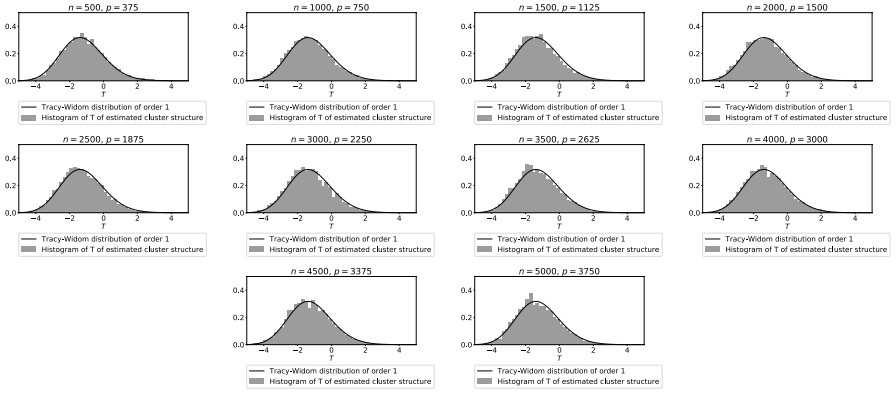

Fig. 5 Histogram of the proposed test statistic T, which was computed with **estimated** bicluster structure (**B-LBM**). The titles of the figures show different matrix sizes

For each combination of the distribution and matrix size settings, we randomly generated 5, 000 data matrices A based on the **null** (non-bi-disjoint) bicluster structure, which was defined as follows. Let $K_1 \equiv (3K + 4 + K \bmod 2)/2$, $K_2 \equiv (3K + 4 - K \bmod 2)/2$, $n_1 \equiv \lfloor n/K_1 \rfloor$, and $p_1 \equiv \lfloor p/K_2 \rfloor$. For each kth bicluster $(k = 1, \ldots, K)$, we also define $k_1 \equiv (3k - 2 - k \bmod 2)/2$ and $k_2 \equiv (3k - 4 + k \bmod 2)/2$. Based on these variables, the set of rows and columns of matrix A belonging to the kth bicluster is given by $I_k = \{k_1 n_1 + 1, \ldots, (k_1 + 2)n_1\}$ and $J_k = \{k_2 p_1 + 1, \ldots, (k_2 + 2)p_1\}$, respectively. Figure 3 shows the examples of Gaussian, Bernoulli, and Poisson data matrices.

After generating the observed data matrices, we estimated their bicluster structures by the proposed submatrix localization algorithm (i.e., Algorithm 2 in Appendix D in the supplementary material (Watanabe and Suzuki 2023)). To compress

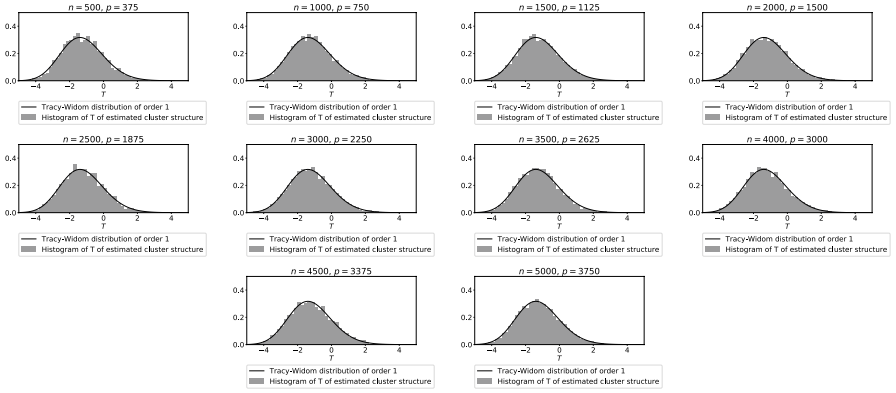

Fig. 6 Histogram of the proposed test statistic T, which was computed with **estimated** bicluster structure (**P-LBM**). The titles of the figures show different matrix sizes

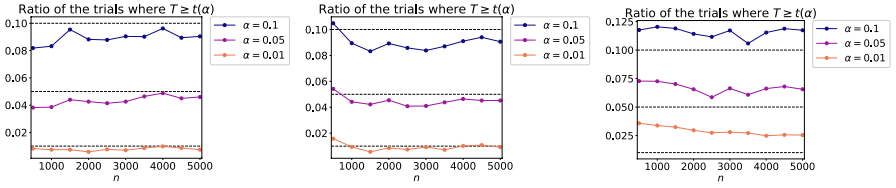

Fig. 7 Empirical tail probabilities of the proposed test statistic T under the three settings of distributions, which was computed with **estimated bicluster structure**. The left, center, and right figures, respectively, show the results where each entry of the observed matrix A was generated using Gaussian, Bernoulli, and Poisson distributions. The horizontal line indicates the row size n of matrix A, and the dashed lines indicate the three significance levels

the original data matrix A, we applied Ward's hierarchical clustering method (Ward 1963) to the rows and columns of matrix A with the number of clusters $L_1 = \min\{2^K, n\}$ and $L_2 = \min\{2^K, p\}$, respectively. Initial bicluster structures in the submatrix localization algorithm were given as follows: for each kth bicluster, it contains a (uniformly randomly chosen) single entry $A_{i_k j_k}$ in A, where $(i_k, j_k) \neq (i_{k'}, j_{k'})$ for $k \neq k'$. In Appendix D in the supplementary material (Watanabe and Suzuki 2023), we describe a sufficient condition regarding the cooling schedule $\{T_t\}$ for the SA algorithm to converge in probability to the global optimal solution. However, such a setting requires too many iterations to converge. In our experiments, we used the following cooling schedule instead: $T_t = 0.999^t$ for all $t \geq 0$. We also defined a threshold of temperature at $\epsilon^{\mathrm{SA}} = 10^{-5}$. Since this setting no longer guaranteed a convergence in probability to the global optimal solution, we applied the submatrix localization algorithm to each observed matrix five times and adopted the best solution that achieved the maximum profile likelihood in the last step of the algorithm (this procedure was also performed in all the subsequent experiments in Sects. 4.2,

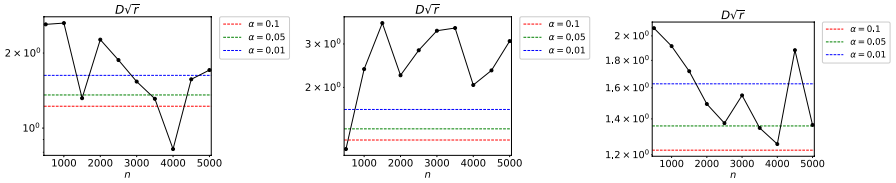

Fig. 8 Test statistics $D\sqrt{r}$ of the KS test (Conover 1999), which was computed using an **estimated bicluster structure**. The left, center, and right figures, respectively, depict the results where each entry of the observed matrix A was generated using Gaussian, Bernoulli, and Poisson distributions. Given a significance level α^{KS} for the KS test, iff $D\sqrt{r} > \alpha^{\mathrm{KS}}$, then the null hypothesis that T follows the TW_1 distribution is rejected

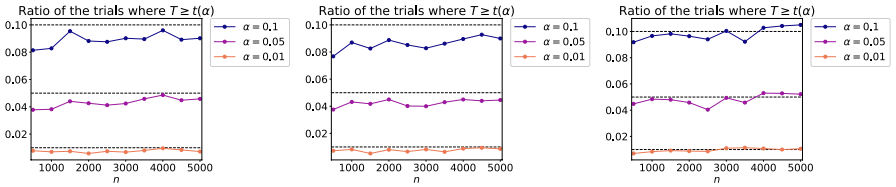

Fig. 9 Empirical tail probabilities of the proposed test statistic T under the three settings of distributions, which was computed with **null bicluster structure**. The left, center, and right figures, respectively, represent the results where each entry of the observed matrix A was generated using Gaussian, Bernoulli, and Poisson distributions

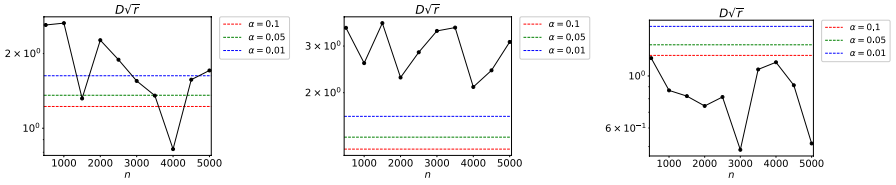

Fig. 10 Test statistics $D\sqrt{r}$ of the KS test (Conover 1999), which was computed with **null bicluster structure**. The left, center, and right figures, respectively, show the results where each entry of the observed matrix A was generated using Gaussian, Bernoulli, and Poisson distributions

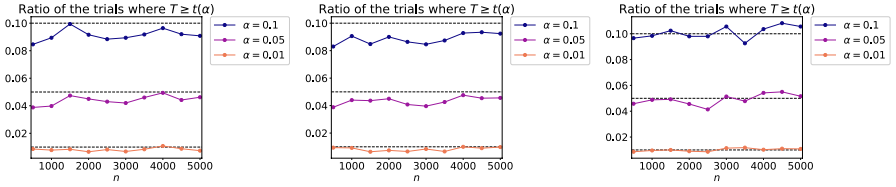

Fig. 11 Empirical tail probabilities of the statistic T^* under the three settings of distributions. The left, center, and right figures, respectively, represent the results where each entry of the observed matrix A was generated using Gaussian, Bernoulli, and Poisson distributions

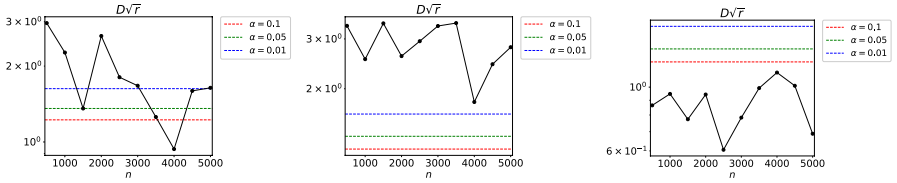

Fig. 12 Test statistics $D\sqrt{r}$ of the KS test (Conover 1999), which was computed with statistic T^*. The left, center, and right figures, respectively, show the results where each entry of the observed matrix A was generated using Gaussian, Bernoulli, and Poisson distributions

4.3, and 4.4). Based on the estimated bicluster structure, we finally applied the proposed statistical test by setting the hypothetical number of biclusters at K.

Figures 4, 5, and 6, respectively, show the histograms of the proposed test statistic T with different matrix sizes under the Gaussian, Bernoulli, and Poisson settings. Figure 7 illustrates the empirical tail probabilities of the proposed test statistic T (i.e., the ratios of the trials where $T \geq t(0.01)$, $T \geq t(0.05)$, and $T \geq t(0.1)$, where $t(\alpha)$ is the α upper quantile of the TW_1 distribution) under the three settings of distributions. As in the previous study (Watanabe and Suzuki 2021), we used the approximated values $t(0.01) \approx 2.02345$, $t(0.05) \approx 0.97931$, and $t(0.1) \approx 0.45014$, based on Table 2 in (Tracy and Widom 2009). To check the convergence of T to the TW_1 distribution, we also applied the Kolmogorov-Smirnov (KS) test (Conover 1999) to the test statistics T of the 5, 000 trials, and plotted the results in Fig. 8. Let D be the maximum absolute difference between the empirical distribution function of T and the cumulative distribution function of the TW_1 distribution. The test statistic of the KS test is $D\sqrt{r}$, where r is the number of trials (i.e., 5, 000 in this case).

From Figs. 4, 5, 6, 7 and 8, we see that the proposed test statistic T converges in law to the TW_1 distribution in each setting of distributions. In the Bernoulli case, however, the convergence of T in law to the TW_1 distribution is slow, compared to the other two cases (i.e., Gaussian and Poisson). To investigate the cause of this, we also computed $\tilde{T}$ with the **null** bicluster structure [i.e., $\tilde{T} \equiv (\tilde{\lambda}_1 - a^{\mathrm{TW}})/b^{\mathrm{TW}}$] and T^* in (15) and plotted their empirical tail probabilities and the test statistics of the KS test in Figs. 9, 10, 11, and 12. From these figures, we see that the convergence of $\tilde{T}$ and T^* in law to the TW_1 distribution is still slow in the Bernoulli case. Therefore, the slow convergence of T would not have been caused by the low accuracy in submatrix localization or in parameter estimation regarding $\boldsymbol{b}$, but it would have been a problem specific to a Bernoulli random matrix.

4.2 The asymptotic behavior of test statistic *T* in the unrealizable case

Second, we consider the unrealizable cases (i.e., $K > K_0$). Specifically, under the assumptions firstly that the total number of biclusters and background submatrices $(K + H)$ is a fixed constant that does not depend on the matrix size and secondly that the minimum row and column sizes of these submatrices ($n_{\min}$ and $p_{\min}$,

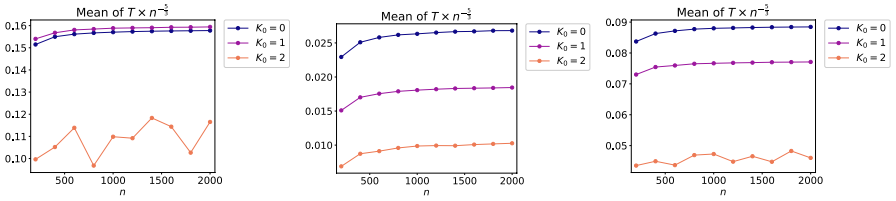

Fig. 13 Mean of the proposed test statistics T divided by $n^{\frac{5}{3}}$ in the unrealizable case for 100 trials. The null number of biclusters was set at $K = 3$. The left, center, and right figures, respectively, represent the results where each entry of observed matrix A was generated using Gaussian, Bernoulli, and Poisson distributions. The horizontal line represents the row size n of the observed matrix. Each plotted line represents a result for a given hypothetical number of biclusters K_0

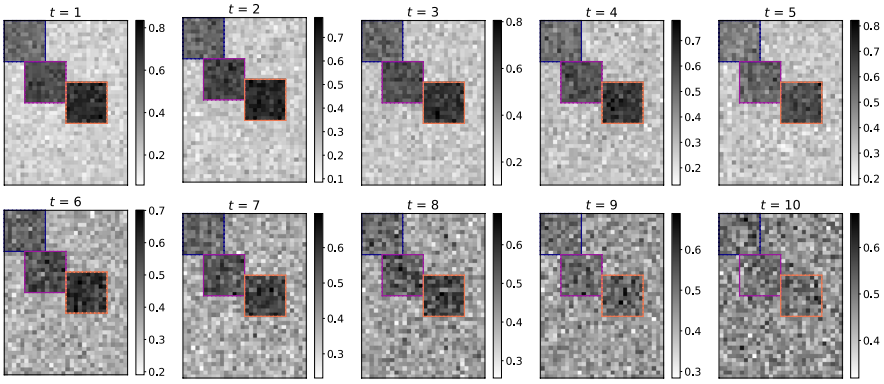

Fig. 14 Examples of the observed data matrices for $t = 1, \ldots, 10$ (**G-LBM**). The colored boxes show the null bicluster structure

respectively) satisfy $n_{\min} = \Omega_p(m)$ and $p_{\min} = \Omega_p(m)$, from (49) and (28), we have $T = \Theta_p\left(m^{\frac{5}{3}}\right)$.

Based on the same procedure outlined in Sect. 4.1, we generated Gaussian, Bernoulli, and Poisson random matrices with three biclusters (i.e., $K = 3$), estimated their bicluster structures, and computed the test statistics. We used the same settings as in Sect. 4.1 for (1) the null parameters of three distributions (51), (53), and (55), (2) the procedure to generate the observed matrices, and (3) the SA-based submatrix localization algorithm. In this experiment, we tried the following 10 sets of matrix sizes: $(n, p) = (200 \times i, 150 \times i)$, $i = 1, \ldots, 10$. For each combination of the distribution and matrix size settings, we randomly generated 100 data matrices A, estimated their bicluster structures with $K_0 = 0, 1, \ldots, K - 1$, and checked the average behavior of test statistic T.

Figure 13 represents the asymptotic behavior of the mean of the proposed test statistic T divided by $n^{\frac{5}{3}}$ under unrealizable settings. This figure illustrates that T increases in proportion to $m^{\frac{5}{3}}$ in all the settings of distributions, as shown in the first paragraph of this section.

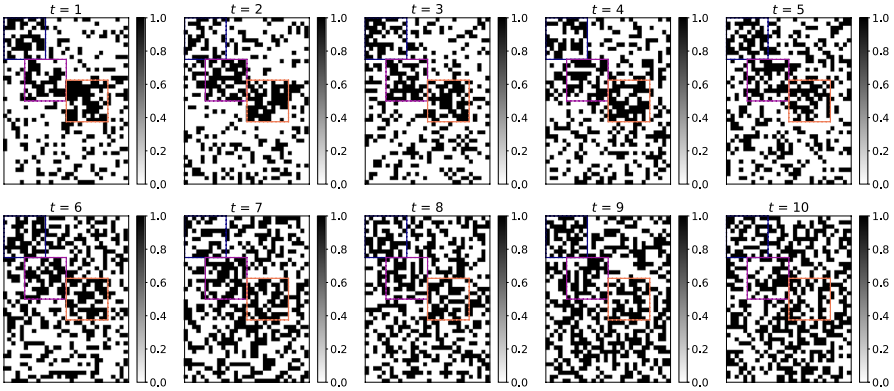

Fig. 15 Examples of the observed data matrices for $t = 1, \ldots, 10$ (**B-LBM**)

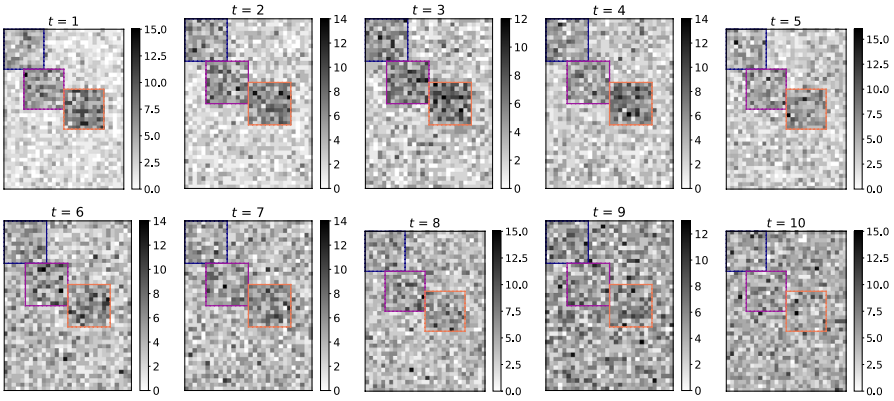

Fig. 16 Examples of the observed data matrices for $t = 1, \ldots, 10$ (**P-LBM**)

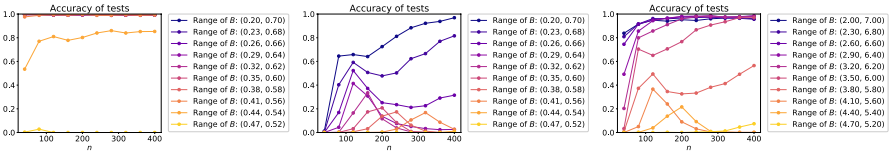

Fig. 17 Accuracy of the proposed test in selecting the number of biclusters K, under 10 different mean parameter settings $\{\boldsymbol{b}^{(1)}, \ldots, \boldsymbol{b}^{(10)}\}$. The left, center, and right figures, respectively, illustrate the results where each entry of observed matrix A was generated using Gaussian, Bernoulli, and Poisson distributions

4.3 The accuracy of the proposed test in selecting the number of biclusters *K*

Third, we checked the accuracy of the proposed test in selecting the number of biclusters K, by using the synthetic Gaussian, Bernoulli, and Poisson data matrices

that were generated using the procedure outlined in Sect. 4.1. We set the null number of biclusters at $K = 3$. As for the null mean parameters of the three distributions, we tried the ten settings $\{\boldsymbol{b}^{(1)}, \ldots, \boldsymbol{b}^{(10)}\}$, where

$$\boldsymbol{b}^{(t)} \equiv \left(1 - \frac{t}{10}\right)\left(\boldsymbol{b} - 0.5\begin{bmatrix}1 & \ldots & 1\end{bmatrix}^\top\right) + 0.5\begin{bmatrix}1 & \ldots & 1\end{bmatrix}^\top \text{ (G} - \text{ and B} - \text{LBMs)}, \tag{57}$$

$$\boldsymbol{b}^{(t)} \equiv \left(1 - \frac{t}{10}\right)\left(\boldsymbol{b} - 5\begin{bmatrix}1 & \ldots & 1\end{bmatrix}^\top\right) + 5\begin{bmatrix}1 & \ldots & 1\end{bmatrix}^\top \text{ (P} - \text{LBM)}, \tag{58}$$

for all $t = 1, \ldots, 10$. In the above settings, we set $\boldsymbol{b}$ at the same vector as given in (51), (53), and (55) for each setting of distributions. For the standard deviation parameter $\boldsymbol{s}$, we used the same setting as in (51). Aside from these model parameters, we used the same settings as in Sect. 4.1 for (1) the procedure to generate the observed matrices, and (2) the SA-based submatrix localization algorithm. Figures 14, 15, and 16, respectively, depict the examples of generated data matrices in Gaussian, Bernoulli, and Poisson cases. In this experiment, we tried the following 10 sets of matrix sizes: $(n, p) = (40 \times i, 30 \times i)$, $i = 1, \ldots, 10$. For each combination of the distribution and matrix size settings, we randomly generated 1, 000 data matrices A and applied the proposed sequential test with a significance level of $\alpha = 0.01$ and the hypothetical number of biclusters $K_0 = 0, 1, 2, \ldots$.

Figure 17 shows the accuracy of the proposed test, that is, the ratio of trials where the selected number of biclusters $\hat{K}$ was equal to the null one K. From Fig. 17, it is clear that the proposed test achieved higher accuracy with the larger matrix sizes and with the smaller differences between the group-wise means. This result is consistent with our intuition, since larger matrix sizes and smaller differences between the elements in mean vector tend to make it more difficult to correctly estimate the underlying bicluster structure of matrix A, based on which we computed the test statistic T.

4.4 Goodness-of-fit test and model selection with practical data set

Finally, we applied the proposed test and the conventional LBM-based one (Watanabe and Suzuki 2021) to the Divorce Predictors data set (Yöntem et al. 2019) from the UCI Machine Learning Repository (Dua and Graff 2017), and compared the results. The rows and columns of the original observed matrix $\breve{A} \in \mathbb{R}^{170\times 54}$ represent the 170 participants and 54 attributes, respectively, and each (i, j)th entry shows the Divorce Predictors Scale (DPS), which takes values of $0, 1, \ldots, 4$. According to (Yöntem 2017), the original questionnaire was done based on the following five-factor scale: 0: "Never," 1: "Rarely," 2: "Occasionally," 3: "Often," and 4: "Always," which was used as a score for Attributes 31 to 54. As for Attributes 1 to 30, this scale was reversed (i.e., 0 meant "Always" and 4 meant "Never") so that higher values indicated a higher divorce risk in all the attributes. Based on the original matrix $\breve{A}$, we defined a binary data matrix A by setting $A_{ij} = 1$ if $\breve{A}_{ij} \geq 2$ for the pair of ith participant and the jth attribute, and $A_{ij} = 0$ otherwise. The upper left section of Fig. 18 depicts the observed data matrix.

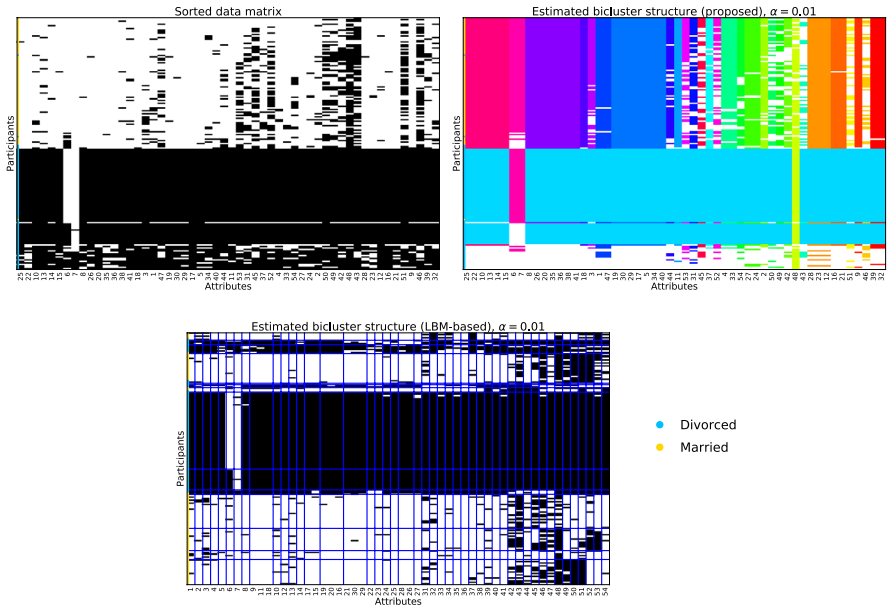

Fig. 18 Sorted observed data matrix of the Divorce Predictors data set (Yöntem et al. 2019) (upper left) and its estimated bicluster structures when the null hypotheses were accepted by the proposed test (upper right) and the previous one (bottom). In the upper left and bottom figures, the black and white elements represent one and zero, respectively. In the upper right figure, the sorting orders of the rows and columns are the same as in the upper left figure, and the color of each element indicates its group index (the white elements were estimated as background), regardless of its value. In the bottom figure, the blue lines represent the regular-grid bicluster structure (note that in this figure, the sorting orders of the rows and columns are different from those of the upper left figure. The meaning of each attribute index is shown in Appendix F in the supplementary material (Watanabe and Suzuki 2023)

As for the proposed test, we applied it sequentially as in Sect. 4.3 with a significance level of $\alpha = 0.01$ until some hypothetical number of biclusters was accepted. In the SA algorithm, we used the relative entropy function f in (54) and the cooling schedule of $T_t = 0.9999^t$ for all $t \geq 0$. For each hypothetical number of biclusters K_0, we set the threshold at $\epsilon^{\mathrm{SA}} = 10^{-K_0/2.5-2}$. Based on these settings, we applied the submatrix localization algorithm 30 times and adopted the best solution that achieved the maximum profile likelihood in the last step. Based on the estimated bicluster structure, we applied the proposed statistical test.

Regarding the conventional LBM-based test, we used the same settings as those employed by Watanabe and Suzuki (Watanabe and Suzuki 2021). That is, for each hypothetical set of row and column cluster numbers (K_0, H_0), we estimated the regular-grid bicluster structure by applying Ward's hierarchical clustering method (Ward 1963) to the rows and columns of observed matrix. Based on the estimated row and column cluster assignments, we applied the test in (Watanabe and Suzuki 2021) with a significance level of $\alpha = 0.01$. We tried multiple combinations of K_0 and H_0 in the following order:

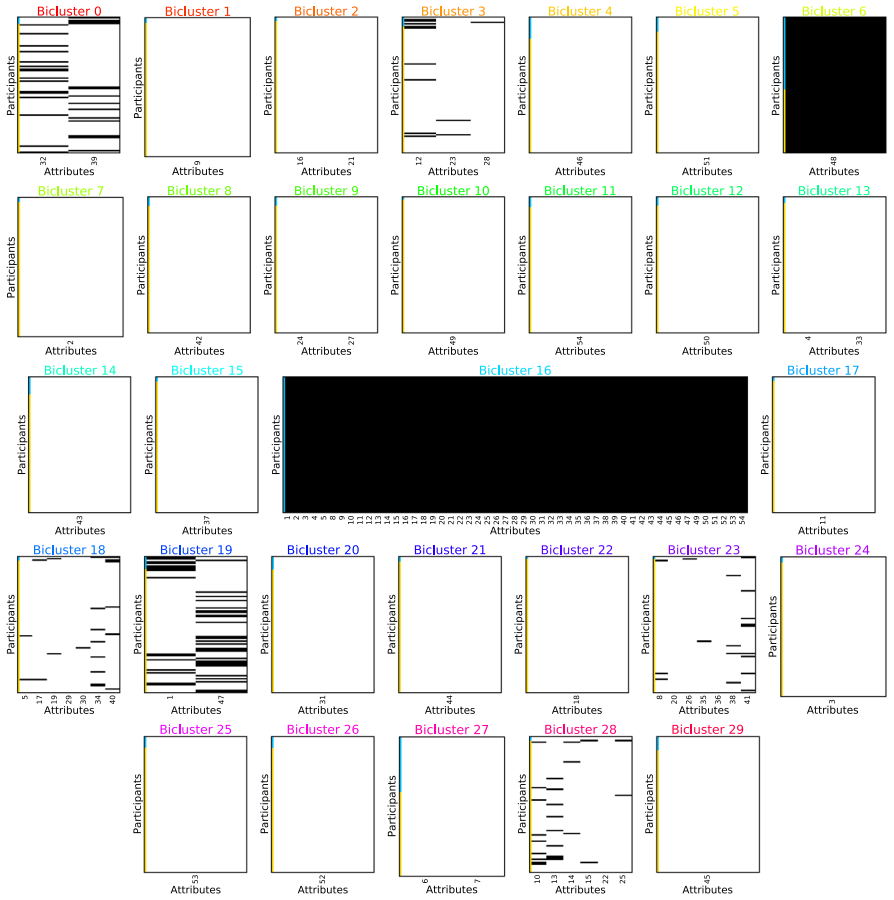

Fig. 19 Estimated biclusters of the observed matrix when the proposed test accepted the null hypothesis. The font color of the title corresponds to the bicluster color in Fig. 18, and the left side color for each row indicates the class of each participant in the bicluster, as in the legend of Fig. 18

$$(K_0, H_0) = (1,1), (1,2), (2,1), (1,3), (2,2), (3,1), \ldots, \tag{59}$$

until the null hypothesis was accepted.

Based on the above settings, the estimated number of biclusters by the proposed test was 30, while the estimated set of row and column cluster numbers by the conventional one (Watanabe and Suzuki 2021) was (14, 46) (i.e., the estimated number of biclusters was 644, aside from the background). The upper right and bottom sections of Fig. 18 show, respectively, the estimated bicluster structure when the null hypotheses were accepted by the proposed and previous LBM-based tests. From these results, we see that that the proposed test could capture the bicluster structure more flexibly than the previous regular-grid-based one, and thereby accepted the smaller hypothetical number of biclusters.

More specifically, Fig. 19 shows each estimated bicluster when the null hypothesis was accepted by the proposed test. Most of the estimated biclusters contained constant values (i.e., 0 or 1), except for Biclusters 0, 3, 18, 19, 23, and 28, and many biclusters consisted of a small number of attributes (i.e., one or two). Except Biclusters 6 and 27, most of the rows (i.e., participants) in each bicluster belonged to the same class (i.e., divorced or married). Each estimated bicluster was composed of some homogeneous sets of rows and columns: for example, Bicluster 9 shows that there exists a (mostly) married group of participants who gave small DPS (i.e., $\breve{A}_{ij} \leq 1$) to the attributes 24 and 27, both of which were related to knowledge about the stress of their spouse. From Bicluster 16, we also see that there existed a divorced group of participants who gave large DPS (i.e., $\breve{A}_{ij} \geq 2$) to many attributes, including their similarity to the spouse (e.g., attributes from 12 to 20) and their awareness of the spouse (e.g., attributes from 21 to 30).

5 Discussion

In this study, we derived the asymptotic behavior of the proposed test statistic T in both the null and alternative cases, where the null number of biclusters might increase with the matrix size (as the condition given in (iv)). Unlike the previous study (Watanabe and Suzuki 2021), we can apply the proposed method when the underlying bicluster structure is not necessarily represented by a regular grid. By sequentially testing the hypothetical numbers of biclusters in an ascending order, we can select an appropriate number of biclusters in a given observed matrix. We experimentally showed the asymptotic behavior of the proposed test statistic T and its accuracy in selecting the correct number of biclusters with synthetic data matrices. Moreover, we analyzed the test result with a practical data set.

Although there is currently no other test statistic that can be used in this problem setting, as also pointed out in (Lei 2016), it can be expected that the power of the proposed test based on the operator norm of $\hat{Z}$ (i.e., the largest singular value of matrix $\hat{Z}$) would be higher than that of a test based on the Frobenius norm of $\hat{Z}$ (i.e., the square root of the sum of all the eigenvalues of $\hat{Z}^\top \hat{Z}$), since the latter test does not take into account information on structural (i.e., biclusterwise) deviations of element values.

As in the previous regular-grid-based test (Watanabe and Suzuki 2021), it is an important future work to reveal the non-asymptotic property of the test statistic T, that is, its convergence rate to the TW_1 distribution. To solve this problem, we need to derive the behavior of T in case that the submatrix localization algorithm does *not* output the correct bicluster structure, which requires more careful analysis.

The main theorem 1 of this paper is based on the consistency assumption of the submatrix localization algorithm. As we described in Sect. 2, we do not have such an algorithm for a general disjoint block structure, and we used a heuristic simulated annealing approach in the experiments. It would be another important direction to

develop a specific submatrix localization algorithm that is theoretically guaranteed to be consistent.

From a practical perspective, some studies (Ben-Dor et al. 2002; Liu et al. 2004) demonstrate the effectiveness of analyzing a gene expression data matrix by assuming the existence of *order-preserving biclusters*, in which a set of rows (i.e., genes) has a similar linear ordering of columns (i.e., conditions). Such a definition of homogeneousness is different from ours, whereby we assume that each entry in a bicluster is generated in the i.i.d. sense. Additionally, some practical relational data matrices (e.g., MovieLens (Harper and Konstan 2015) and Jester (Goldberg et al. 2001) data sets) contain missing entries. It is an important topic in future research to construct a statistical test on K for such cases deriving its theoretical guarantee.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10463-023-00869-3.

Acknowledgements We would like to thank Editage (www.editage.com) for English language editing.

Funding This work was supported by JSPS KAKENHI (18H03201 and 20H00576), Fujitsu Laboratories Ltd., and JST CREST.

References

Balakrishnan, S., Kolar, M., Rinaldo, A., Singh, A., Wasserman, L. (2011). Statistical and computational tradeoffs in biclustering. In: *NIPS 2011 workshop on computational trade-offs in statistical learning*.

Ben-Dor, A., Chor, B., Karp, R., Yakhini, Z. (2002). Discovering local structure in gene expression data: The order-preserving submatrix problem. In: *Proceedings of the Sixth Annual International Conference on Computational Biology* (pp 49–57).

Bickel, P. J., Sarkar, P. (2016). Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), *78*(1), 253–273.

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

Brennan, M., Bresler, G., Huleihel, W. (2018). Reducibility and computational lower bounds for problems with planted sparse structure. In: *Proceedings of the 31st Conference On Learning Theory* (vol 75, pp. 48–166). Proceedings of Machine Learning Research.

Brennan, M., Bresler, G., Huleihel, W. (2019). Universality of computational lower bounds for submatrix detection. In: *Proceedings of the 32nd Conference On Learning Theory* (vol 99, pp. 417–468). Proceedings of Machine Learning Research.

Butucea, C., Ingster, Y. I. (2013). Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli, 19*(5B), 2652–2688.

Butucea, C., Ingster, Y.I., Suslina, I.A. (2015). Sharp variable selection of a sparse submatrix in a high-dimensional noisy matrix. *ESAIM*: *Probability and Statistics* 19:115–134.

Cai, T. T., Wu, Y. (2020). Statistical and computational limits for sparse matrix detection. *Annals of Statistics, 48*(3), 1593–1614.

Cai, T. T., Liang, T., Rakhlin, A. (2017). Computational and statistical boundaries for submatrix localization in a large noisy matrix. *Annals of Statistics, 45*(4), 1403–1430.

Chekouo, T., Murua, A. (2015). The penalized biclustering model and related algorithms. *Journal of Applied Statistics, 42*(6), 1255–1277.

Chekouo, T., Murua, A., Raffelsberger, W. (2015). The Gibbs-plaid biclustering model. *Annals of Applied Statistics, 9*(3), 1643–1670.

Chen, Y., Xu, J. (2016). Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *Journal of Machine Learning Research, 17*(27), 1–57.

Conover, W. J. (1999). *Practical nonparametric statistics*. New York: John Wiley & Sons.
Corneli, M., Latouche, P., Rossi, F. (2015). Exact ICL maximization in a non-stationary time extension of the latent block model for dynamic networks. In: *Proceedings of the 23-th European Symposium on Artificial Neural Networks* (pp. 225–230). Computational Intelligence and Machine Learning.
Dhillon, I.S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. In: *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 269–274).
Dua, D., Graff, C. (2017). UCI machine learning repository. http://archive.ics.uci.edu/ml, University of California, Irvine, School of Information and Computer Sciences.
Duffy, D. E., Quiroz, A. J. (1991). A permutation-based algorithm for block clustering. *Journal of Classification, 8*, 65–91.
Flynn, C. J., Perry, P. O. (2020). Profile likelihood biclustering. *Electronic Journal of Statistics, 14*(1), 731–768.
França, FOD. (2012). Scalable overlapping co-clustering of word-document data. In: *2012 11th International Conference on Machine Learning and Applications* (pp. 464–467).
Goldberg, K., Roeder, T., Gupta, D., Perkins, C. (2001). Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval, 4*(2), 133–151.
Hajek, B., Wu, Y., Xu, J. (2017). Information limits for recovering a hidden community. *IEEE Transactions on Information Theory, 63*(8), 4729–4745.
Hajek, B., Wu, Y., Xu, J. (2018). Submatrix localization via message passing. *Journal of Machine Learning Research, 18*(186), 1–52.
Harper, F. M., Konstan, J. A. (2015). The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems, 5*(4), 1–19.
Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association, 67*(337), 123–129.
Hochreiter, S., Bodenhofer, U., Heusel, M., Mayr, A., Mitterecker, A., Kasim, A., Khamiakova, T., Sanden, S. V., Lin, D., Talloen, W., Bijnens, L., Göhlmann, H. W. H., Shkedy, Z., Clevert, D. A. (2010). FABIA: Factor analysis for bicluster acquisition. *Bioinformatics, 26*(12), 1520–1527.
Hu, J., Zhang, J., Qin, H., Yan, T., Zhu, J. (2020). Using maximum entry-wise deviation to test the goodness of fit for stochastic block models. *Journal of the American Statistical Association* 0(0):1–10.
Kolar, M., Balakrishnan, S., Rinaldo, A., Singh, A. (2011). Minimax localization of structural information in large noisy matrices. *Advances in Neural Information Processing Systems, 24*, 909–917.
Lei, J. (2016). A goodness-of-fit test for stochastic block models. *The Annals of Statistics, 44*(1), 401–424.
Liu, J., Yang, J., Wang, W. (2004). Biclustering in gene expression data by tendency. In: *Proceedings of 2004 IEEE Computational Systems Bioinformatics Conference* (pp. 182–193).
Liu, Y., Guo, J. (2018). Distribution-free, size adaptive submatrix detection with acceleration. arXiv: 1804.10887.
Lomet, A., Govaert, G., Grandvalet, Y. (2012). Model selection in block clustering by the integrated classification likelihood. In: *Proceedings of 20th International Conference on Computational Statistics* (pp. 519–530).
Luo, Y., Zhang, A. (2020). Tensor clustering with planted structures: Statistical optimality and computational limits. In: *2020 Joint Statistical Meetings*.
Ma, Z., Wu, Y. (2015). Computational barriers in minimax submatrix detection. *Annals of Statistics, 43*(3), 1089–1116.
Madeira, S. C., Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1*(1), 24–45.
Moran, G.E. (2019). Bayesian approaches for modeling variation. PhD thesis, University of Pennsylvania, Pennsylvania, United States.
Oghabian, A., Kilpinen, S., Hautaniemi, S., Czeizler, E. (2014). Biclustering methods: Biological relevance and application in gene expression analysis. *PLOS ONE, 9*(3), e90801.
Pillai, N. S., Yin, J. (2014). Universality of covariance matrices. *Annals of Applied Probability, 24*(3), 935–1001.
Pio, G., Ceci, M., D'Elia, D., Loglisci, C., Malerba, D. (2013). A novel biclustering algorithm for the discovery of meaningful biological correlations between microRNAs and their target genes. *BMC Bioinformatics, 14*(7), S8.
Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L., Zitzler, E. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics, 22*(9), 1122–1129.

Raff, E., Zak, R., Munoz, G.L., Fleming, W., Anderson, H.S., Filar, B., Nicholas, C., Holt, J. (2020). Automatic Yara rule generation using biclustering. In: *Proceedings of the 13th ACM Workshop on Artificial Intelligence and Security* (pp 71–82).

Sakai, Y., Yamanishi, K. (2013). An NML-based model selection criterion for general relational data modeling. In: *Proceedings of 2013 IEEE International Conference on Big Data* (pp. 421–429).

Shabalin, A. A., Weigman, V. J., Perou, C. M., Nobel, A. B. (2009). Finding large average submatrices in high dimensional data. *Annals of Applied Statistics, 3*(3), 985–1012.

Shan, H., Banerjee, A. (2008). Bayesian co-clustering. In: *Proceedings of the 8th IEEE International Conference on Data Mining* (pp. 530–539).

Sill, M., Kaiser, S., Benner, A., Kopp-Schneider, A. (2011). Robust biclustering by sparse singular value decomposition incorporating stability selection. *Bioinformatics, 27*(15), 2089–2097.

Symeonidis, P., Nanopoulos, A., Papadopoulos, A., Manolopoulos, Y. (2007). Nearest-biclusters collaborative filtering with constant values. In: *Advances in Web Mining and Web Usage Analysis, WebKDD 2006, Lecture Notes in Computer Science* (vol. 4811, pp. 36–55).

Tanay, A., Sharan, R., Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics, 18*(1–Suppl), S136–S144.

Tepper, M., Sapiro, G. (2016). Fast L1-NMF for multiple parametric model estimation. arXiv:1610.05712.

Tibshirani, R., Hastie, T., Eisen, M., Ross, D., Botstein, D., Brown, P. (1999). Clustering methods for the analysis of DNA microarray data. Tech. rep., Department of health research and policy, department of statistics, department of genetics and department of biochemistry, Stanford University.

Tracy, C.A., Widom, H. (2009). The distributions of random matrix theory and their applications. In: *New Trends in Mathematical Physics* (pp. 753–765), Springer.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge, England: Cambridge University Press.

Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236–244.

Watanabe, C., Suzuki, T. (2021). Goodness-of-fit test for latent block models. *Computational Statistics & Data Analysis, 154*, 107090.

Watanabe, C., Suzuki, T. (2023). Supplement to "a goodness-of-fit test on the number of biclusters in a relational data matrix". *Annals of the Institute of Statistical Mathematics*.

Wyse, J., Friel, N., Latouche, P. (2017). Inferring structure in bipartite networks using the latent blockmodel and exact ICL. *Network Science, 5*(1), 45–69.

Yamanishi, K., Wu, T., Sugawara, S., Okada, M. (2019). The decomposed normalized maximum likelihood code-length criterion for selecting hierarchical latent variable models. *Data Mining and Knowledge Discovery, 33*, 1017–1058.

Yöntem, M.K. (2017). The predictive role of the styles of parenthood origin on divorce predictors. PhD thesis, Gaziosmanpasa University, Tokat, Turkey.

Yöntem, M. K., 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.