

# Are Clusterings of Multiple Data Views Independent?

Lucy L. Gao <sup>\*1</sup>, Jacob Bien<sup>2</sup>, and Daniela Witten<sup>3</sup>

<sup>1</sup>*Department of Biostatistics, University of Washington*

<sup>2</sup>*Department of Data Sciences and Operations, University of Southern California*

<sup>3</sup>*Departments of Statistics and Biostatistics, University of Washington*

January 15, 2019

## Abstract

In the Pioneer 100 (P100) Wellness Project (Price *and others*, 2017), multiple types of data are collected on a single set of healthy participants at multiple timepoints in order to characterize and optimize wellness. One way to do this is to identify clusters, or subgroups, among the participants, and then to tailor personalized health recommendations to each subgroup. It is tempting to cluster the participants using all of the data types and timepoints, in order to fully exploit the available information. However, clustering the participants based on multiple *data views* implicitly assumes that a single underlying clustering of the participants is shared across all data views. If this assumption does not hold, then clustering the participants using multiple data views may lead to spurious results. In this paper, we seek to evaluate the assumption that there is some underlying relationship among the clusterings from the different data views, by asking the question: are the clusters within each data view *dependent* or *independent*? We develop a new test for answering this question, which we then apply to clinical, proteomic, and metabolomic data, across two distinct timepoints, from the P100 study. We find that while the subgroups of the participants defined with respect to any single data type seem to be dependent across time, the clustering among the participants based on one data type (e.g. proteomic data) appears not to be associated with the clustering based on another data type (e.g. clinical data). Data integration; Hypothesis testing; Model-based clustering; Multiple-view data.

## 1 Introduction

Complex biological systems consist of diverse components with dynamics that may vary over time, and so these systems often cannot be fully characterized by any single type of data, or at any single snapshot in time. Consequently, it has become increasingly common for researchers to collect multiple data sets, or *views*, for a single set of observations. In the machine learning literature, this is known as the *multiple-view* or *multi-view* data setting.

---

<sup>\*</sup>lucygao@uw.edu

Multiple-view data has been applied extensively to characterize disease, such as in The Cancer Genome Atlas Project (Cancer Genome Atlas Research Network, 2008). In contrast, The Pioneer 100 (P100) Wellness Project (Price *and others*, 2017) collected multiple-view data from healthy participants to characterize wellness, and to optimize wellness of the participants through personalized healthcare recommendations. One way to do this is to identify subgroups of similar participants using cluster analysis, and then tailor recommendations to each subgroup.

In recent years, many papers have proposed clustering methods in the multiple-view data setting (Bickel and Scheffer, 2004; Shen *and others*, 2009; Kumar *and others*, 2011; Kirk *and others*, 2012; Lock and Dunson, 2013; Gabasova *and others*, 2017). The vast majority of these methods “borrow strength” across the data views to obtain a more accurate clustering of the observations than would be possible based on a single data view. Implicitly, these methods assume that there is a single *consensus* clustering shared by all data views.

The P100 data contains many data views; multiple data types (e.g. clinical data and proteomic data) are available at multiple timepoints. Thus, it is tempting to apply consensus clustering methods to identify subgroups of the P100 participants. However, before doing so, it is important to check the assumption that there exists a single consensus clustering. If instead different views reflect unrelated aspects of the participants, then there is no “strength to be borrowed” across the views, and it would be better to perform a separate clustering of the observations in each view. Before attempting cluster analysis of the P100 data, it is critical that we determine which combinations of views have “strength to be borrowed”, and which combinations do not.

This raises the natural question of how associated the underlying clusterings are in each view. Suppose we cluster the P100 participants twice, once using their baseline clinical data, and once using their baseline proteomic data. *Can we tell from the data whether the two views’ underlying clusterings are related or unrelated?* Answering this question provides useful information:

*Case 1:* If the underlying clusterings appear related, then this increases confidence that the clusterings are scientifically meaningful, and offers some support for performing a consensus clustering of the P100 participants that integrates baseline clinical and proteomic views.

*Case 2:* If the underlying clusterings appear unrelated, we must consider two explanations.

1. Perhaps clinical and proteomic views measure different properties about the participants, and therefore identify complementary (or “orthogonal”) clusterings. If so, then a consensus clustering is unlikely to provide meaningful results, and may cause us to lose valuable information about the subgroups underlying the individual data views.
2. Perhaps the subgroups underlying the data views are indeed related, but they appear unrelated due to noise. If so, then we might be skeptical of any results obtained on these very noisy data, whether from consensus clustering or another approach.

In Case 2, it would not be appropriate to perform consensus clustering.

To determine from the data whether the two views' clusterings are related or unrelated, it is tempting to apply a clustering procedure (e.g. k-means) to each view, then apply well-studied tests of independence of categorical variables (e.g. the  $\chi^2$ -test for independence, the  $G$ -test for independence, or Fisher's exact test) to the estimated cluster assignments. However, such an approach relies on an assumption that the estimated cluster assignments are independent and identically distributed samples from the joint distribution of the cluster membership variables, which is not satisfied in practice. Thus, there is a need for an approach which takes into account the fact that the clusterings are estimated from the data.

The rest of this paper is organized as follows. In Section 2, we propose a mixture model for two-view data. In Section 3, we use this model to develop a test of the null hypothesis that clusterings on two views of a single set of observations are independent. We explore the performance of our proposed hypothesis test via numerical simulation in Section 4. In Section 5, we connect and compare our proposed hypothesis test to the aforementioned approach of applying the  $G$ -test for independence to the estimated cluster assignments, and draw connections between this approach and the mutual information statistic (Meilă, 2007). In Section 6, we apply our method to the clinical, proteomic, and metabolomic datasets from the P100 study. In Section 7, we provide a discussion, which includes the extension to more than two views.

## 2 A Mixture Model for Multiple-View Data

### 2.1 Model Specification

In what follows, we consider the case of two data views. We will discuss the extension to more than two views in Section 7.

Suppose we have  $p_1$  and  $p_2$  features in the first and second data view, respectively. For a single observation, let  $X^{(1)} \in \mathbb{R}^{p_1}$  and  $X^{(2)} \in \mathbb{R}^{p_2}$  denote the random vectors corresponding to the two data views and let  $Z^{(1)} \in \{1, \dots, K^{(1)}\}$  and  $Z^{(2)} \in \{1, \dots, K^{(2)}\}$  be unobserved random variables, indicating the latent group memberships of this observation in the two data views. Here,  $K^{(1)}$  and  $K^{(2)}$  represent the number of clusters in the two data views, which we assume for now to be known (we will consider the case in which they are unknown in Section 2.4). We assume that  $X^{(1)}$  and  $X^{(2)}$  are conditionally independent given the pair of cluster memberships,  $(Z^{(1)}, Z^{(2)})$ ; this assumption is common in the multi-view clustering literature (see e.g. Bickel and Scheffer 2004, Rogers *and others* 2008, Kumar *and others* 2011, Lock and Dunson 2013, Gabasova *and others* 2017). Further, suppose that

$$f(X^{(l)} | Z^{(l)} = k) = \phi^{(l)}(X^{(l)}; \theta_k^{(l)}) \quad \text{for } 1 \leq k \leq K^{(l)}, 1 \leq l \leq 2, \quad (2.1)$$

$$P(Z^{(1)} = k, Z^{(2)} = k') = \Pi_{kk'} \quad \text{for } 1 \leq k \leq K^{(1)}, 1 \leq k' \leq K^{(2)}, \quad (2.2)$$

where  $\phi^{(l)}(\cdot; \theta)$  denotes a density function with parameter  $\theta$ , and  $\Pi \in \Delta^{K^{(1)} \times K^{(2)}} \equiv \{S \in \mathbb{R}^{K^{(1)} \times K^{(2)}} : S_{kk'} \geq 0, \sum_{k=1}^{K^{(1)}} \sum_{k'=1}^{K^{(2)}} S_{kk'} = 1\}$ . Equations (2.1)–(2.2) are an extension of the finite mixture model (McLachlan and Peel, 2000) to the case of two data views. We further assume

that each cluster has positive probability, i.e.  $P(Z^{(1)} = k) > 0$  and  $P(Z^{(2)} = k') > 0$ , and so  $\Pi 1_{K^{(2)}} \in \Delta_+^{K^{(1)}}$  and  $\Pi^T 1_{K^{(1)}} \in \Delta_+^{K^{(2)}}$ , where  $\Delta_+^K \equiv \{s \in \mathbb{R}^K : s_k > 0, \sum_{k=1}^K s_k = 1\}$ .

Let  $\theta^{(1)} \equiv (\theta_1^{(1)}, \dots, \theta_{K^{(1)}}^{(1)})$  and  $\theta^{(2)} \equiv (\theta_1^{(2)}, \dots, \theta_{K^{(2)}}^{(2)})$ . The joint density of  $X^{(1)}$  and  $X^{(2)}$  is

$$\begin{aligned} f(X^{(1)}, X^{(2)}; \theta^{(1)}, \theta^{(2)}, \Pi) &= \sum_{k, k'} \Pi_{kk'} f(X^{(1)}, X^{(2)} \mid Z^{(1)} = k, Z^{(2)} = k') \\ &= \sum_{k, k'} \Pi_{kk'} f(X^{(1)} \mid Z^{(1)} = k) f(X^{(2)} \mid Z^{(2)} = k') \\ &= \sum_{k, k'} \Pi_{kk'} \phi^{(1)}(X^{(1)}; \theta_k^{(1)}) \phi^{(2)}(X^{(2)}; \theta_{k'}^{(2)}), \end{aligned} \quad (2.3)$$

where the second equality follows from conditional independence of  $X^{(1)}$  and  $X^{(2)}$  given  $Z^{(1)}$  and  $Z^{(2)}$ , and the last equality follows from (2.1).

The matrix  $\Pi$  governs the statistical dependence between the two data views. It will be useful for us to parameterize  $\Pi$  in terms of a triplet  $(\pi^{(1)}, \pi^{(2)}, C)$  that separates the single-view information from the cross-view information.

**Proposition 1.** *Suppose  $\pi^{(1)} \in \Delta_+^{K^{(1)}}$  and  $\pi^{(2)} \in \Delta_+^{K^{(2)}}$ . Then,*

$$\left\{ \Pi \in \Delta^{K^{(1)} \times K^{(2)}} : \Pi 1_{K^{(2)}} = \pi^{(1)}, \Pi^T 1_{K^{(1)}} = \pi^{(2)} \right\} = \left\{ \text{diag}(\pi^{(1)}) C \text{diag}(\pi^{(2)}) : C \in \mathcal{C}_{\pi^{(1)}, \pi^{(2)}} \right\},$$

where  $\mathcal{C}_{\pi^{(1)}, \pi^{(2)}} = \{C \in \mathbb{R}^{K^{(1)} \times K^{(2)}} : C_{kk'} \geq 0, C\pi^{(2)} = 1_{K^{(1)}}, C^T \pi^{(1)} = 1_{K^{(2)}}\}$ .

A proof of Proposition 1 is given in Appendix A.1 of the Supplementary Materials.

Proposition 1 indicates that any matrix  $\Pi \in \Delta^{K^{(1)} \times K^{(2)}}$  with  $\Pi 1_{K^{(2)}} \in \Delta_+^{K^{(1)}}$  and  $\Pi^T 1_{K^{(1)}} \in \Delta_+^{K^{(2)}}$  can be written as the product of its row sums  $\pi^{(1)}$ , its column sums  $\pi^{(2)}$ , and a matrix  $C$ . Therefore, we can rewrite the joint probability density (2.3) as follows:

$$\begin{aligned} f(X^{(1)}, X^{(2)}; \theta^{(1)}, \theta^{(2)}, \Pi) &= \sum_{k, k'} \pi_k^{(1)} C_{kk'} \pi_{k'}^{(2)} \phi^{(1)}(X^{(1)}; \theta_k^{(1)}) \phi^{(2)}(X^{(2)}; \theta_{k'}^{(2)}) \\ &\equiv f(X^{(1)}, X^{(2)}; \theta^{(1)}, \theta^{(2)}, \pi^{(1)}, \pi^{(2)}, C). \end{aligned} \quad (2.4)$$

In what follows, we will parametrize the density of  $X^{(1)}$  and  $X^{(2)}$  in terms of  $\theta^{(1)}, \theta^{(2)}, \pi^{(1)}, \pi^{(2)}$ , and  $C$ , rather than in terms of  $\theta^{(1)}, \theta^{(2)}$ , and  $\Pi$ .

The following proposition characterizes the marginal distributions of  $X^{(1)}$  and  $X^{(2)}$ .

**Proposition 2.** *Suppose  $X^{(1)}$  and  $X^{(2)}$  have joint distribution (2.4). Then for  $l = 1, 2$ ,  $X^{(l)}$  has marginal density given by*

$$f(X^{(l)}; \theta^{(l)}, \pi^{(l)}) = \sum_{k=1}^{K^{(l)}} \pi_k^{(l)} \phi_k^{(l)}(X^{(l)}; \theta_k^{(l)}). \quad (2.5)$$

Proposition 2 follows from (2.1) – (2.2). Proposition 2 shows that for  $l = 1, 2$ ,  $X^{(l)}$  marginally follows a mixture model with parameters  $\theta^{(l)}$  and cluster membership probabilities  $\pi^{(l)}$ . Note that the marginal density of  $X^{(1)}$  does not depend on  $\theta^{(2)}, \pi^{(2)}$ , and  $C$ , and similarly, the marginal density of  $X^{(2)}$  does not depend on  $\theta^{(1)}, \pi^{(1)}$ , and  $C$ ; this fact will be critical to our approach to parameter estimation in Section 2.3.

The model described in this section is closely related to several multiple-view mixture models proposed in the literature: see e.g. Rogers *and others* (2008), Kirk *and others* (2012), Lock and Dunson (2013), and Gabasova *and others* (2017). However, the focus of those papers is cluster estimation: they do not provide a statistical test of association, and for the most part, impose additional structure on the probability matrix  $\Pi$  in order to encourage similarity between the clusters estimated in each data view. By contrast, the focus of this paper is inference: testing for dependence between the clusterings in different data views. The model described in this section is a step towards that goal.

## 2.2 Interpreting $\Pi$

In Figures 1(i)–(iii),  $n = 15$  independent pairs  $\{(X_i^{(1)}, X_i^{(2)})\}_{i=1}^n$  are drawn from the model (2.1)–(2.2), for three choices of  $\Pi$ . The left-hand panel represents the  $p_1 = 2$  features in the first data view, and the right-hand panel represents the  $p_2 = 2$  features in the second data view. For  $l = 1, 2$ , the observations  $\{X_i^{(l)}\}_{i=1}^n$  in the  $l$ th data view belong to two clusters, where the latent variables  $\{Z_i^{(l)}\}_{i=1}^n$  characterize cluster membership in the  $l$ th data view. Light and dark gray represent the clusters in the first view, and circles and triangles represent the clusters in the second view.

Figures 1(i)–(ii) correspond to two special cases of  $\Pi$  that are easily interpretable. In Figure 1(i),  $\Pi$  has rank one, i.e.  $\Pi = \pi^{(1)}[\pi^{(2)}]^T$ , so that the clusterings in the two data views are independent. Thus, whether an observation is light or dark appears to be roughly independent of whether it is a circle or a triangle. In Figure 1(ii),  $K^{(1)} = K^{(2)}$  and  $\Pi$  is diagonal (up to a permutation of the rows), so that the clusterings in the two data views are identical. Thus, all of the circles are light and all of the triangles are dark. Another special case is when  $\Pi$  is block diagonal (up to a permutation) with  $K_B$  blocks. Then, the clusterings of the two data views agree about the presence of  $K_B$  “meta-clusters” in the data. For example, one clustering might be a refinement of the other, or if one view has clusters  $A, B, C, D$ , and the other has clusters  $a, b, c, d$ , it could be that  $A \cup B = a$  and  $C = b \cup c$  and  $D = d$ .

In general,  $\Pi$  will be neither exactly rank one nor exactly (block) diagonal; Figure 1(iii) provides such an example. Furthermore,  $\hat{\Pi}$  (an estimator for  $\Pi$ ) almost certainly will be neither. Nonetheless, examination of  $\hat{\Pi}$  can provide insight into the relationships between the two clusterings. For example, if  $\hat{\Pi}$  is far from rank one, then this suggests that the clusterings in the two data views may be dependent. We will formalize this intuition in Section 3.

## 2.3 Estimation

### 2.3.1 Estimation Procedure and Algorithm

Given  $n$  independent pairs  $(X_1^{(1)}, X_1^{(2)}), \dots, (X_n^{(1)}, X_n^{(2)})$  drawn from the model (2.1)–(2.2), the log-likelihood takes the form

$$\ell(\theta^{(1)}, \theta^{(2)}, \pi^{(1)}, \pi^{(2)}, C) = \sum_{i=1}^n \log f(X_i^{(1)}, X_i^{(2)}; \theta^{(1)}, \theta^{(2)}, \pi^{(1)}, \pi^{(2)}, C), \quad (2.6)$$

where  $f(\cdot, \cdot; \theta^{(1)}, \theta^{(2)}, \pi^{(1)}, \pi^{(2)}, C)$  is defined in (2.4). A custom expectation-maximization (EM; Dempster *and others* 1977; McLachlan and Krishnan 2007) algorithm could be developed to solve (2.6) for a local optimum (a global optimum is typically unattainable, as (2.6) is non-concave). We instead take a simpler approach. Proposition 2 implies that for  $l = 1, 2$ , we can estimate  $\theta^{(l)}$  and  $\pi^{(l)}$  by maximizing the marginal likelihood for the  $l$ th data view, given by

$$\ell(\theta^{(l)}, \pi^{(l)}) = \sum_{i=1}^n \log f(X_i^{(l)}; \theta^{(l)}, \pi^{(l)}), \quad (2.7)$$

where  $f(\cdot; \theta^{(l)}, \pi^{(l)})$  is defined in (2.5). Each of these maximizations can be performed using standard EM-based software for model-based clustering of a single data view. Let  $\hat{\theta}^{(1)}, \hat{\pi}^{(1)}, \hat{\theta}^{(2)}$ , and  $\hat{\pi}^{(2)}$  denote the maximizers of (2.7). Next, to estimate  $C$ , we maximize the joint log-likelihood (2.6) evaluated at  $\hat{\theta}^{(1)}, \hat{\pi}^{(1)}, \hat{\theta}^{(2)}$ , and  $\hat{\pi}^{(2)}$ , subject to the constraints imposed by Proposition 1:

$$\hat{C} \equiv \arg \min_{C \in \mathcal{C}_{\hat{\pi}^{(1)}, \hat{\pi}^{(2)}}} \left[ -\ell(\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \hat{\pi}^{(1)}, \hat{\pi}^{(2)}, C) \right], \quad (2.8)$$

where  $\mathcal{C}_{\hat{\pi}^{(1)}, \hat{\pi}^{(2)}} = \{C \in \mathbb{R}^{K^{(1)} \times K^{(2)}} : C_{kk'} \geq 0, C\hat{\pi}^{(2)} = 1_{K^{(1)}}, C^T \hat{\pi}^{(1)} = 1_{K^{(2)}}\}$ . Equation 2.8 is a convex optimization problem, which we solve using a combination of exponentiated gradient descent (Kivinen and Warmuth, 1997) and the Sinkhorn-Knopp algorithm (Franklin and Lorenz, 1989), as detailed in Appendix B of the Supplementary Materials. Details of our approach for fitting the model (2.1)–(2.2) are given in Algorithm 1.

### 2.3.2 Justification of Estimation Procedure

The estimation procedure in Section 2.3.1 does not maximize the joint likelihood (2.6); nonetheless, we will argue that it is an attractive approach.

To begin, in Step 1 of Algorithm 1, we estimate  $\theta^{(1)}$  and  $\pi^{(1)}$  by maximizing the marginal likelihood (2.7). This decision leads to computational advantages, as it enables us to make use of efficient software for clustering a single data view, such as the `mclust` package (Scrucca *and others*, 2016) in R. We can further justify this decision using conditional inference theory. Equation 3.6 in Reid (1995) extends the definition of ancillary statistics to a setting with nuisance parameters. We show that  $X^{(2)}$  is ancillary (in the extended sense of Reid 1995)

---

**Algorithm 1** Procedure for fitting the model (2.1)–(2.2)

---

1. Maximize the marginal likelihoods (2.7) in order to obtain the marginal MLEs  $\hat{\theta}^{(1)}$ ,  $\hat{\pi}^{(1)}$  and  $\hat{\theta}^{(2)}$ ,  $\hat{\pi}^{(2)}$ . This can be done using standard software for model-based clustering.

2. Define matrices  $\hat{\phi}^{(1)} \in \mathbb{R}^{n \times K^{(1)}}$  and  $\hat{\phi}^{(2)} \in \mathbb{R}^{n \times K^{(2)}}$  with elements

$$\hat{\phi}_{ik}^{(1)} = \phi^{(1)} \left( X_i^{(1)}; \hat{\theta}_k^{(1)} \right) \quad \text{and} \quad \hat{\phi}_{ik'}^{(2)} = \phi^{(2)} \left( X_i^{(2)}; \hat{\theta}_{k'}^{(2)} \right). \quad (2.9)$$

3. Fix a step size  $s > 0$ . Theorem 5.3 from Kivinen and Warmuth (1997) gives conditions on  $s$  that guarantee convergence.

4. Let  $\hat{C}^1 = 1_{K^{(1)}} 1_{K^{(2)}}^T$ . For  $t = 1, 2, \dots$  until convergence:

- (a) Define  $M_{kk'} = \hat{C}_{kk'}^t \exp\{sG_{kk'} - 1\}$ , where  $G_{kk'} = \sum_{i=1}^n \frac{\hat{\phi}_{ik}^{(1)} \hat{\phi}_{ik'}^{(2)}}{[\hat{\phi}_i^{(1)}]^T \text{diag}(\hat{\pi}^{(1)}) \hat{C}^t \text{diag}(\hat{\pi}^{(2)}) \hat{\phi}_i^{(2)}}$ .

- (b) Let  $u^0 = 1_{K^{(2)}}$  and  $v^0 = 1_{K^{(1)}}$ . For  $t' = 1, 2, \dots$ , until convergence:

- i.  $u^{t'} = \frac{1_{K^{(2)}}}{M^T \text{diag}(\hat{\pi}^{(1)}) v^{t'-1}}, \quad v^{t'} = \frac{1_{K^{(1)}}}{M \text{diag}(\hat{\pi}^{(2)}) u^{t'}}$ ,

where the fractions denote element-wise vector division.

- (c) Let  $u$  and  $v$  be the vectors to which  $u^{t'}$  and  $v^{t'}$  converge. Let  $\hat{C}_{kk'}^{t+1} = u_k M_{kk'} v_{k'}$ .

5. Let  $\hat{C}$  denote the matrix to which  $\hat{C}^t$  converges, and let  $\hat{\Pi} = \text{diag}(\hat{\pi}^{(1)}) \hat{C} \text{diag}(\hat{\pi}^{(2)})$ .
-

for  $\theta^{(1)}$ ,  $\pi^{(1)}$ , and  $C$  by using the definition of conditional densities, and Proposition 2, to rewrite (2.4) as

$$f(X^{(1)}, X^{(2)}; \theta^{(1)}, \theta^{(2)}, \pi^{(1)}, \pi^{(2)}, C) = f(X^{(1)} | X^{(2)}; \theta^{(1)}, \theta^{(2)}, \pi^{(1)}, \pi^{(2)}, C) f(X^{(2)}; \theta^{(2)}, \pi^{(2)}).$$

Thus, Reid (1995) argues that we should use only  $X^{(2)}$ , and not  $X^{(1)}$ , to estimate  $\theta^{(2)}$  and  $\pi^{(2)}$ . In Step 1 of Algorithm 1, we are doing exactly this.

In Steps 3–5 of Algorithm 1, we maximize  $\ell(\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \hat{\pi}^{(1)}, \hat{\pi}^{(2)}, \cdot)$ , giving  $\hat{C}$ , which is a pseudo maximum likelihood estimator for  $C$  in the sense of Gong and Samaniego (1981). This decision also leads to computational advantages, as it enables us to make use of efficient convex optimization algorithms in estimating  $C$ . Results in Gong and Samaniego (1981) suggest that when  $\hat{\theta}^{(1)}$ ,  $\hat{\theta}^{(2)}$ ,  $\hat{\pi}^{(1)}$ , and  $\hat{\pi}^{(2)}$  are good estimates,  $\hat{C}$  is so as well.

## 2.4 Selection of the Number of Clusters

In Sections 2 and 3, our discussion assumed that  $K^{(1)}$  and  $K^{(2)}$  are known. However, this is rarely the case in practice. Recall that we estimate  $\theta^{(1)}$  and  $\pi^{(1)}$  by maximizing the marginal likelihood (2.7), which amounts to performing model-based clustering of  $X^{(1)}$  only. Thus, to select the number of clusters  $K^{(1)}$ , we can make use of an extensive literature (reviewed in e.g. Mirkin 2011) on choosing the number of clusters when clustering a single data view. For example, we can use AIC or BIC to select  $K^{(1)}$  and  $K^{(2)}$ .

# 3 Testing Whether Two Clusterings are Independent

## 3.1 A Brief Review of Pseudo Likelihood Ratio Tests

Let  $\ell(\alpha, \beta, \gamma)$  be the log-likelihood function for a random sample, where  $\mathcal{A}$  is the parameter space of  $\alpha$ . Given a null hypothesis  $H_0 : \alpha = \alpha_0$  for some  $\alpha_0 \in \mathcal{A}$ , an alternative hypothesis  $H_1 : \alpha \neq \alpha_0$ , and an estimator  $\hat{\gamma}$ , the pseudo likelihood ratio statistic (Self and Liang, 1987) is defined to be  $\log \tilde{\Lambda} \equiv \sup_{\alpha, \beta} \ell(\alpha, \beta, \hat{\gamma}) - \sup_{\beta} \ell(\alpha_0, \beta, \hat{\gamma})$ . Let  $\alpha^*$  be the true parameter value for  $\alpha$ . If  $\alpha^*$  is an interior point of  $\mathcal{A}$ , then under some regularity conditions, if  $H_0$  holds, then  $2 \log \tilde{\Lambda} \xrightarrow{d} \chi_r^2$ , where  $r$  is the dimension of  $\mathcal{A}$  (Chen and Liang, 2010).

## 3.2 A Pseudo Likelihood Ratio Test for Independence

In this subsection, we develop a test for the null hypothesis that  $H_0 : C = 1_{K^{(1)}} 1_{K^{(2)}}^T$ , or equivalently, that  $H_0 : \Pi = \pi^{(1)}(\pi^{(2)})^T$ : that is, we test whether  $Z^{(1)}$  and  $Z^{(2)}$  are independent, i.e. whether the cluster memberships in the two data views are independent. We could use a likelihood ratio test statistic to test  $H_0$ ,

$$\begin{aligned} \log \Lambda &\equiv \sup_{\theta^{(1)}, \theta^{(2)}, \pi^{(1)}, \pi^{(2)}, C} \ell(\theta^{(1)}, \theta^{(2)}, \pi^{(1)}, \pi^{(2)}, C) - \sup_{\theta^{(1)}, \theta^{(2)}, \pi^{(1)}, \pi^{(2)}} \ell(\theta^{(1)}, \theta^{(2)}, \pi^{(1)}, \pi^{(2)}, 1_{K^{(1)}} 1_{K^{(2)}}^T) \\ &= \sup_{\theta^{(1)}, \theta^{(2)}, \pi^{(1)}, \pi^{(2)}, C} \ell(\theta^{(1)}, \theta^{(2)}, \pi^{(1)}, \pi^{(2)}, C) - [\ell(\hat{\theta}^{(1)}, \hat{\pi}^{(1)}) + \ell(\hat{\theta}^{(2)}, \hat{\pi}^{(2)})], \end{aligned} \quad (3.10)$$



where the second equality follows from noticing that substituting  $C = 1_{K^{(1)}} 1_{K^{(2)}}^T$  into (2.6) yields

$$\ell(\theta^{(1)}, \theta^{(2)}, \pi^{(1)}, \pi^{(2)}, C) = \ell(\theta^{(1)}, \pi^{(1)}) + \ell(\theta^{(2)}, \pi^{(2)}), \quad (3.11)$$

where  $\ell(\theta^{(l)}, \pi^{(l)})$  for  $l = 1, 2$  are defined in (2.7), and recalling the definition of  $\hat{\theta}^{(1)}, \hat{\pi}^{(1)}, \hat{\theta}^{(2)}$ , and  $\hat{\pi}^{(2)}$  as the maximizers of (2.7). However, (3.10) requires maximizing  $\ell(\theta^{(1)}, \theta^{(2)}, \pi^{(1)}, \pi^{(2)}, C)$ , which would require a custom EM algorithm; furthermore, the resulting test statistic will typically involve the difference between two local maxima (since each term in (3.10) requires fitting an EM algorithm). This leads to erratic behavior, such as negative values of  $\log \Lambda$ .

Therefore, instead of taking the approach in (3.10), we develop a pseudo likelihood ratio test, as in Section 3.1. We use the marginal MLEs,  $\hat{\theta}^{(1)}, \hat{\pi}^{(1)}$ , and  $\hat{\theta}^{(2)}, \hat{\pi}^{(2)}$ , instead of performing the joint optimization in (3.10). This leads to the test statistic

$$\begin{aligned} \log \tilde{\Lambda} &\equiv \sup_{C \in \mathcal{C}_{\hat{\pi}^{(1)}, \hat{\pi}^{(2)}}} \ell(\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \hat{\pi}^{(1)}, \hat{\pi}^{(2)}, C) - \ell(\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \hat{\pi}^{(1)}, \hat{\pi}^{(2)}, 1_{K^{(1)}} 1_{K^{(2)}}^T) \\ &= \ell(\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \hat{\pi}^{(1)}, \hat{\pi}^{(2)}, \hat{C}) - [\ell(\hat{\theta}^{(1)}, \hat{\pi}^{(1)}) + \ell(\hat{\theta}^{(2)}, \hat{\pi}^{(2)})] \end{aligned} \quad (3.12)$$

$$= \sum_{i=1}^n \log \left[ \frac{(\hat{\phi}_i^{(1)})^T \text{diag}(\hat{\pi}^{(1)}) \hat{C} \text{diag}(\hat{\pi}^{(2)}) \hat{\phi}_i^{(2)}}{(\hat{\phi}_i^{(1)})^T \hat{\pi}^{(1)} (\hat{\pi}^{(2)})^T \hat{\phi}_i^{(2)}} \right]. \quad (3.13)$$

where  $\hat{C}$  in (3.12) is defined in (2.8),  $\mathcal{C}_{\hat{\pi}^{(1)}, \hat{\pi}^{(2)}}$  is defined in Proposition 1,  $\hat{\phi}^{(1)}$  and  $\hat{\phi}^{(2)}$  are defined in (2.9), and the last equality follows from (2.6), (2.7), and (2.9). In addition to taking advantage of the computationally efficient estimation procedure described in Section 2.3.1, the pseudo likelihood ratio test statistic does not exhibit the erratic behavior exhibited by the likelihood ratio test statistic. This stability comes from all three terms in (3.12) involving the same local maxima (as opposed to different local maxima).

### 3.3 Approximating the Null Distribution of $\log \tilde{\Lambda}$

The discussion in Section 3.1 suggests that under  $H_0 : C = 1_{K^{(1)}} 1_{K^{(2)}}^T$ , one might expect that  $2 \log \tilde{\Lambda} \xrightarrow{d} \chi_r^2$ , where  $r = (K^{(1)} - 1)(K^{(2)} - 1)$  is the dimension of  $\mathcal{C}_{\pi^{(1)}, \pi^{(2)}}$ . However, this approximation performs poorly in practice, due to violations of the regularity conditions in Chen and Liang (2010). Furthermore, we will often be interested in data applications in which  $n$  is relatively small. Hence, we propose a permutation approach. We observe from (3.11) that under  $H_0$ , the log-likelihood is identical under any permutation of the order of the samples in each view. Hence, we take  $B$  random permutations of the samples  $X_i^{(2)}$  from the second view, and compare the observed value of  $\log \tilde{\Lambda}$  to its empirical distribution in these permutation samples. Details are given in Algorithm 2. Since  $\hat{\phi}^{(1)}, \hat{\phi}^{(2)}, \hat{\pi}^{(1)}$ , and  $\hat{\pi}^{(2)}$  are invariant to permutation, for each permutation we need only to estimate  $C$ . This is another advantage of our test over the likelihood ratio test discussed in Section 3.2, which would require repeating the EM algorithm in every permutation. Even when we reject the null hypothesis, the clusters could be only weakly dependent; thus, it is helpful to measure the strength of association between the views. Recalling from Section 2.2 that  $\text{rank}(\Pi) = 1$  implies independence of the clusterings in the two data views, we propose to calculate the

---

**Algorithm 2** A Permutation Approach for Testing  $H_0 : C = 1_{K^{(1)}} 1_{K^{(2)}}^T$

---

1. Compute  $\log \tilde{\Lambda}$  according to (3.13) using the original data,  $X^{(1)}$  and  $X^{(2)}$ .
  2. For  $b = 1, \dots, B$ , where  $B$  is the number of permutations:
    - (a) Permute the observations in  $X^{(2)}$  to obtain  $X^{(2,*b)}$ .
    - (b) Compute  $\log \tilde{\Lambda}^{*b}$  according to (3.13) based on  $X^{(1)}$  and  $X^{(2,*b)}$ .
  3. The p-value for testing  $H_0 : C = 1_{K^{(1)}} 1_{K^{(2)}}^T$  is given by  $\frac{1}{B} \sum_{b=1}^B 1_{\{\log \tilde{\Lambda} \leq \log \tilde{\Lambda}^{*b}\}}$ .
- 

effective rank (Vershynin, 2012) of  $\hat{\Pi}$ , defined in Algorithm 1 – the ratio of the sum of the singular values of  $\hat{\Pi}$ , and the largest singular value of  $\hat{\Pi}$ . The effective rank of a matrix is bounded between 1 and its rank, and the matrix is far from rank-1 when its effective rank is far from 1. For example, in Figure 1(iii), the effective rank of  $\Pi$  is 1.5, and is upper bounded by 2. Thus, the effective rank of  $\hat{\Pi}$  is bounded between 1 and  $\min\{K^{(1)}, K^{(2)}\}$ , and  $\hat{\Pi}$  is far from rank-1 when its effective rank is far from 1.

## 4 Simulation Results

To investigate the Type I error and power of our test, we generate data from (2.1)–(2.2), with

$$\Pi = \frac{1 - \delta}{K^2} 1_K 1_K^T + \frac{\delta}{K} I_K, \quad (4.14)$$

for  $K = 6$  and for a range of values of  $\delta \in [0, 1]$ , where  $\delta = 0$  corresponds to independent clusterings, and  $\delta = 1$  corresponds to identical clusterings. We draw the observations in the  $l$ th data view from a Gaussian mixture model, for which the  $k$ th mixture component is a  $N_p(\mu_k^{(l)}, \Sigma^{(l)})$  distribution, with  $p = 10$ , and with  $\mu_k^{(l)}$  given in Appendix C.1 of the Supplementary Materials.

We simulate 2000 data sets for  $\Sigma^{(1)} = \Sigma^{(2)} = \sigma^2 I_p$  for a range of values of  $\sigma$  and  $n$ , and evaluate the power of the pseudo likelihood ratio test of  $H_0 : C = 1_{K_1} 1_{K_2}^T$  described in Section 3.2 at nominal significance level  $\alpha = 0.05$ , when the number of clusters is correctly and incorrectly specified. To perform Step 1 of Algorithm 1, we use the package `mclust` in R to fit Gaussian mixture models with a common  $\sigma^2 I_p$  covariance matrix (the “EII” covariance structure in `mclust`). We use  $B = 200$  permutation samples in Step 2 of Algorithm 2. Simulations in this paper were conducted using the `simulator` package (Bien, 2016) in R. Results are shown in Figure 2.

The pseudo likelihood ratio test controls the Type I error close to the nominal  $\alpha = 0.05$  level, even when the number of clusters is misspecified. Power tends to increase as  $\delta$  (defined in (4.14)) increases, and tends to decrease as  $\sigma$  increases. Compared to using the correct number of clusters, using too many clusters yields lower power, but using too few clusters can sometimes yield higher power (e.g. in the middle panel of Figure 2). This is because,

when the signal-to-noise ratio is low, the true clusters are not accurately estimated; thus, combining several true clusters into a single “meta-cluster” can sometimes, but not always, lead to improved agreement between clusterings across the two data views. We explore the impact of the choice of  $K$  on the performance of the pseudo likelihood ratio test in Appendix C.2.1 of the Supplementary Materials.

Additional values of  $K$  and  $p$  are investigated in Appendix C.2.2.

## 5 Connection to the G-test for Independence and Mutual Information

Let  $\hat{M}^{(1)} = (\hat{M}_1^{(1)}, \dots, \hat{M}_n^{(1)})$  and  $\hat{M}^{(2)} = (\hat{M}_1^{(2)}, \dots, \hat{M}_n^{(2)})$  denote the results of applying a clustering procedure to  $X^{(1)}$  and  $X^{(2)}$  respectively. In this notation,  $\hat{M}_i^{(1)} \in \{1, \dots, K^{(1)}\}$  and  $\hat{M}_i^{(2)} \in \{1, \dots, K^{(2)}\}$  denote the estimated cluster assignment for the  $i$ th observation in the two views. To test whether  $Z^{(1)}$  and  $Z^{(2)}$  are independent, we could naively apply tests on  $\hat{M}^{(1)}$  and  $\hat{M}^{(2)}$  for whether two categorical variables are independent. For instance, we could use the  $G$ -test statistic for independence (Chapter 3.2, Agresti 2003), given by

$$G^2(\hat{M}^{(1)}, \hat{M}^{(2)}) = 2 \sum_{k=1}^{K^{(1)}} \sum_{k'=1}^{K^{(2)}} \hat{N}_{kk'} \log \left[ (n \hat{N}_{kk'}) / (\hat{N}_{k.} \hat{N}_{.k'}) \right], \quad (5.15)$$

where  $\hat{N}_{kk'} = |\{i \in \{1, \dots, n\} : \hat{M}_i^{(1)} = k, \hat{M}_i^{(2)} = k'\}|$ ,  $\hat{N}_{.k'} = \sum_k \hat{N}_{kk'}$ , and  $\hat{N}_{k.} = \sum_{k'} \hat{N}_{kk'}$ .

Under the model  $\hat{N}_{kk'} \stackrel{\text{ind}}{\sim} \text{Poisson}(\mu_{kk'})$ ,  $\log \mu_{kk'} = \alpha_k + \beta_{k'} + \gamma_{kk'}$ , the  $G$ -test statistic for independence (5.15) is a likelihood ratio test statistic for testing the null hypothesis of independence, i.e. for testing  $H_0: \gamma_{kk'} = 0$  for all  $k, k'$ . Thus, under  $H_0: \gamma_{kk'} = 0$ ,

$$G^2(\hat{M}^{(1)}, \hat{M}^{(2)}) \xrightarrow{d} \chi^2_{(K^{(1)}-1)(K^{(2)}-1)}. \quad (5.16)$$

The  $G$ -test statistic for independence (5.15) relies on an assumption which is violated in our setting, namely that  $\{(\hat{M}_i^{(1)}, \hat{M}_i^{(2)})\}_{i=1}^n$  are independent and identically distributed samples from the distribution of  $(Z^{(1)}, Z^{(2)})$ . It is nonetheless a natural approach to the problem of comparing two views' clusterings. In fact, the mutual information of Meilă (2007) for measuring the similarity between two clusterings of a single  $n \times p$  dataset can be written as a scaled version of the  $G$ -test statistic; when applied to instead measure the similarity between  $\hat{M}^{(1)}$  and  $\hat{M}^{(2)}$ , the mutual information  $I(\hat{M}^{(1)}, \hat{M}^{(2)})$  is given by

$$I(\hat{M}^{(1)}, \hat{M}^{(2)}) = G^2(\hat{M}^{(1)}, \hat{M}^{(2)}) / 2n. \quad (5.17)$$

While the proposed pseudo likelihood ratio test statistic (3.13) for testing independence of  $Z^{(1)}$  and  $Z^{(2)}$  does not resemble the simple  $G$ -test statistic for independence in (5.15), we show here that they are in fact quite related.

Let  $\hat{r}_i^{(1)} = \frac{\hat{\phi}_i^{(1)}}{1_{K^{(1)}}^T \hat{\phi}_i^{(1)}}$  and  $\hat{r}_i^{(2)} = \frac{\hat{\phi}_i^{(2)}}{1_{K^{(2)}}^T \hat{\phi}_i^{(2)}}$  be the vectors giving the soft-clustering assignment weights (or “responsibilities”) for the  $i$ th observation in the two views, where  $\hat{\phi}_i$  is defined

in (2.9). We rewrite the pseudo likelihood ratio test statistic (3.13) as

$$\log \tilde{\Lambda} \left( \hat{\Pi}, \{\hat{r}_i^{(1)}, \hat{r}_i^{(2)}\}_{i=1}^n \right) = \sum_{i=1}^n \log \left[ \frac{(\hat{r}_i^{(1)})^T \hat{\Pi} \hat{r}_i^{(2)}}{(\hat{r}_i^{(1)})^T (\hat{\Pi} \mathbf{1}_{K^{(2)}}) (\mathbf{1}_{K^{(1)}}^T \hat{\Pi}) \hat{r}_i^{(2)}} \right], \quad (5.18)$$

where  $\hat{\Pi}$  is defined in Algorithm 1. In the following proposition, we consider replacing the “soft” cluster assignments  $\hat{r}_i^{(1)}$  and  $\hat{r}_i^{(2)}$  with “hard” cluster assignments, and replacing the estimate  $\hat{\Pi}$  derived from the “soft” cluster assignments with an estimate derived from “hard” cluster assignments, in (5.18). In what follows,

$$\hat{M}_i^{(1)} \equiv \arg \max_{k \in \{1, 2, \dots, K^{(1)}\}} \hat{r}_{ik}^{(1)}, \quad \hat{M}_i^{(2)} \equiv \arg \max_{k' \in \{1, 2, \dots, K^{(2)}\}} \hat{r}_{ik'}^{(2)}. \quad (5.19)$$

**Proposition 3.** *Let  $\hat{M}^{(1)}$  and  $\hat{M}^{(2)}$  be the estimated model-based cluster assignments in each data view defined by (5.19). Let  $\hat{N}$  be the matrix with entries  $\hat{N}_{kk'}$  containing the number of observations assigned to cluster  $k$  in view 1 and cluster  $k'$  in view 2. Then,*

$$\log \tilde{\Lambda}(\hat{N}/n, \{e_{\hat{M}_i^{(1)}}, e_{\hat{M}_i^{(2)}}\}_{i=1}^n) = nI(\hat{M}^{(1)}, \hat{M}^{(2)}) = G^2(\hat{M}^{(1)}, \hat{M}^{(2)})/2, \quad (5.20)$$

where  $\log \tilde{\Lambda}(\cdot, \cdot)$  is defined in (5.18), and  $e_t$  is the unit vector that contains a 1 in the  $t$ th element.

Proposition 3 follows by algebra, and says that replacing the soft cluster assignments in the pseudo likelihood ratio test statistic of Section 3 with hard cluster assignments yields *exactly* the  $G$ -test statistic for independence (5.15) (and the mutual information given in (5.17))! In fact, in the special case of fitting multiple-view Gaussian mixtures with common covariance matrix  $\sigma^2 I_{p_1}$  in the first view and  $\sigma^2 I_{p_2}$  in the second view, we will show that as  $\sigma \rightarrow 0$ , and the soft cluster assignments converge to hard cluster assignments, the pseudo likelihood ratio test statistic converges to the  $G$ -test for independence. In what follows,  $\log \tilde{\Lambda} \equiv \log \tilde{\Lambda} \left( \hat{\Pi}, \{\hat{r}_i^{(1)}, \hat{r}_i^{(2)}\}_{i=1}^n \right)$ , as in (3.13) and (5.18).

**Proposition 4.** *Let  $\sigma^2 > 0$ . Suppose that to compute  $\log \tilde{\Lambda}$ , we fit the model (2.1)–(2.2), for  $\phi^{(1)}$  and  $\phi^{(2)}$  densities of Gaussian distributions with covariance matrices  $\sigma^2 I_{p_1}$  and  $\sigma^2 I_{p_2}$  respectively. Let  $\tilde{M}^{(1)}$  and  $\tilde{M}^{(2)}$  denote the results of applying  $k$ -means clustering on the two data views. Then, as  $\sigma^2 \rightarrow 0$ ,  $\log \tilde{\Lambda} \rightarrow nI(\tilde{M}^{(1)}, \tilde{M}^{(2)}) = G^2(\tilde{M}^{(1)}, \tilde{M}^{(2)})/2$ .*

Proposition 4 is proven in Appendix A.2 of the Supplementary Materials. When  $\sigma^2 > 0$ , the pseudo likelihood ratio test statistic, the  $G$ -test statistic, and the mutual information are not equivalent. We can thus think of the pseudo likelihood ratio test statistic as reflecting the uncertainty associated with the clusterings obtained on the two views, and the  $G$ -test statistic and the mutual information as ignoring the uncertainty associated with the clusterings. This suggests that the pseudo likelihood ratio test of Section 3.2 outperforms the  $G$ -test for independence when the sample size is small and/or there is little separation between the clusters.

To confirm this intuition, we return to the simulation set-up described in Section 4, and compare the performances of the pseudo likelihood ratio test (3.13) and the  $G$ -test

for independence (5.15) for testing  $H_0 : C = 1_{K^{(1)}} 1_{K^{(2)}}^T$ . We obtain p-values for (5.15) using the  $\chi^2$  approximation from (5.16), and using a permutation approach, where we take  $B$  permutations of the elements of  $\hat{M}^{(2)}$ , and compare the observed value of (5.15) to its empirical distribution in these permutation samples. The results are shown in Figure 3; we see that the two tests yield similar power when the sample size is larger and/or the value of  $\sigma$  is smaller, and that the pseudo likelihood ratio test yields higher power than the G-test for independence when the sample size is smaller and/or the value of  $\sigma$  is larger. We note that the  $\chi^2$  approximation for the G-test from (5.16) does not control the Type I error. Additional values of  $p$  and  $K$ , additional values of  $\Sigma^{(l)}$ , and non-Gaussian finite mixture models are investigated in Appendices C.3.1, C.3.2, and C.3.3 of the Supplementary Materials, respectively; the results are similar to those described in this section.

## 6 Application to the Pioneer 100 Wellness Project (Price *and others*, 2017)

### 6.1 Introduction to the Scientific Problem

In the P100 Wellness Project (Price *and others*, 2017), multiple biological data types were collected at multiple timepoints for 108 healthy participants. For each participant, whole genome sequences were measured, activity tracking data were collected daily over nine months, and clinical laboratory tests, metabolomes, proteomes, and microbiomes were measured at three-month, six-month, and nine-month timepoints. The P100 study aims to optimize wellness of the participants through personalized healthcare recommendations. In particular, clinical biomarkers measured at baseline were used to make personalized health recommendations.

As an alternative approach, we could identify subgroups of individuals with similar clinical profiles using cluster analysis, and then develop interventions tailored to each subgroup. It is tempting to identify these subgroups using not just clinical data at baseline, but also other types of data (e.g. proteomic data) at other timepoints. We could do this by applying a multi-view consensus clustering method (e.g. Shen *and others* 2009). However, such an approach assumes that there is a single true clustering underlying all data types at all timepoints. Therefore, before applying a consensus clustering approach, we should determine whether there is any evidence that the clusterings underlying the data types and/or timepoints are at all related (in which case consensus clustering may lead to improved estimation of the clusters) or whether the clusterings are completely unrelated (in which case one would be better off simply performing a separate clustering of the observations in each view). In what follows, we will use the hypothesis test developed in Section 3 to determine whether clusterings of P100 participants based on clinical, proteomic, and genomic data are dependent across timepoints, and across data types.

### 6.2 Data Analysis

At each of the three timepoints, 207 clinical measurements, 268 proteomic measurements, and 642 metabolomic measurements were available for  $n = 108$  observations. In the following, we

define a data view to be a single data type at a single timepoint. In each view, we removed features missing in more than 25% of participants, and removed participants missing more than 25% of features. Next, features in each view with standard deviation 0 were removed. The remaining missing data were imputed using nearest neighbors imputation in the `impute` package in R (Hastie *and others*, 2017). Features in each view were then adjusted for gender using linear regression. Finally, the remaining features were scaled to have standard deviation 1. As in Section 4, we consider the model (2.1)–(2.2) under the assumption that each component in the mixture is drawn from a Gaussian distribution. For each data view, we fit the model using the `mclust` package in R, with a common  $\sigma^2 I$  covariance matrix (the “EII” covariance structure in `mclust`). To test  $H_0 : C = 1_{K(1)} 1_{K(2)}^T$ , we compute p-values using the permutation approximation discussed in Section 3.3 with  $B = 10^5$ . Based on the results in Appendix C.2.1 of the Supplementary Materials, we choose the number of clusters in each view by BIC under the constraint that the number of clusters is greater than one.

We now compare the clusterings in the clinical data at the first and third timepoints, the clustering in the proteomic data at the first and third timepoints, and the clusterings in the metabolomic data at the first and third timepoints. The sample sizes and results are reported in Table 1. For each data type, the clusters found at each timepoint are displayed in Figure 4.

We find strong evidence that for each data type, the clusterings at the first and third timepoints are not independent. We further measure the strength of dependence through the effective rank of  $\hat{\Pi}$ , as described in Section 3.3. For the clusterings in the clinical data, the effective rank of  $\hat{\Pi}$  is 1.63, and is upper bounded by 2. For the clusterings in the proteomic data, the effective rank of  $\hat{\Pi}$  is 1.90, and is upper bounded by 5. For the clusterings in the metabolomic data, the effective rank of  $\hat{\Pi}$  is 1.2, and is upper bounded by 3. These results suggest that the strengths of association for the clusterings estimated on the clinical data, the proteomic data, and the metabolomic data, are strong, moderate, and weak respectively. The fact that the clusterings estimated on some data types are strongly dependent over time provides evidence that they are scientifically meaningful. Furthermore, it suggests that performing consensus clustering on some data types (e.g. clinical data and proteomic data) across timepoints may be reasonable.

We now focus on comparing clusterings in the clinical, proteomic, and metabolomic data at a single timepoint. The sample sizes and results are reported in Table 1.

The results provide modest evidence that proteomic and metabolomic data at a given timepoint are dependent, and provide weak evidence that clinical and metabolomic data are dependent. However, on balance, the evidence that the clusterings are dependent across data types is weaker than we might expect. This suggests to us that the underlying subgroups defined by the three data types are in fact quite different, and that we should be very wary of performing a consensus clustering type approach across data types, or any analysis strategy that assumes that all three data types are getting at the same set of underlying clusters.

## 7 Discussion

Most existing work on multiple-view clustering has focused on the problem of *estimation*: namely, on exploiting the availability of multiple data views in order to cluster the observa-

tions more accurately. In this paper, we have instead focused on the relatively unexplored problem of *inference*: we have proposed a hypothesis test to determine whether clusterings based on multiple data views are independent or associated.

In Section 6, we applied our test to the P100 Wellness Study (Price *and others*, 2017). We found strong evidence that clusterings based on clinical data and proteomic data persist over time, i.e. that the subgroups defined by the clinical data and the proteomic data are similar at different timepoints. This suggests that if we wish to identify participant subgroups based on (say) clinical data, then it may be worthwhile to apply a consensus clustering approach to the clinical data from multiple timepoints. However, we found only modest evidence that clusterings based on different data types are dependent! This suggests that we should be cautious about identifying participant subgroups by applying consensus clustering across multiple data types, as the clusterings underlying the distinct data types may be quite different.

Throughout this paper, we compared clusterings on  $L = 2$  data views. We may also wish to compare clusterings across  $L > 2$  views. Let  $X^{(l)} \in \mathbb{R}^{p_l}$  for  $1 \leq l \leq L$  be the random vectors corresponding to the  $L$  views. Suppose  $X^{(l)}$  are generated according to (2.1) for  $1 \leq l \leq L$ , where  $(Z^{(1)}, \dots, Z^{(L)})$  are unobserved multinomial random variables with probabilities given by  $P(Z^{(1)} = k_1, \dots, Z^{(L)} = k_L) = \Pi_{k_1 \dots k_L}$ , for  $1 \leq k_l \leq K^{(l)}$  and  $1 \leq l \leq L$ , where the sum of  $\Pi$  over all indices is 1 and  $\Pi_{k_1 \dots k_L} \geq 0$ . Results analogous to Propositions 1 and 2 hold in this setting. Thus, we can estimate the parameters in the extended model much as we did in Section 2.3.1, replacing the Sinkhorn-Knopp algorithm for matrix balancing with a tensor balancing algorithm (see e.g. Sugiyama *and others* 2017). To test the null hypothesis that  $Z^{(1)}, \dots, Z^{(L)}$  are mutually independent, we can develop a pseudo likelihood ratio test much as we did in Section 3, where instead of permuting the observations in  $X^{(2)}$  in Step 2(a) of Algorithm 2, we permute the observations in  $X^{(2)}, \dots, X^{(L)}$ . Alternatively, one can simply test for pairwise independence between clusterings, instead of testing for mutual independence between clusterings on all views, as we did in Section 6.

An R package titled `multiviewtest` is available online at <https://github.com/lucylgao/multiviewtest> and is forthcoming on CRAN. Code to reproduce the data analysis in Section 6, and to reproduce the simulations in Sections 4 and 5 and in Appendix C, are available online at <https://github.com/lucylgao/independent-clusterings-code>.

## Acknowledgments

We thank Nathan Price and John Earls for responding to inquiries about the P100 data, and Will Fithian for a useful conversation. Lucy L. Gao received funding from the Natural Sciences and Engineering Research Council of Canada. Daniela Witten and Jacob Bien were supported by NIH Grant R01GM123993. Jacob Bien was supported by NSF CAREER Award DMS-1653017. Daniela Witten was supported by NIH Grant DP5OD009145, NSF CAREER Award DMS-1252624, and Simons Investigator Award No. 560585. *Conflict of Interest*: None declared.

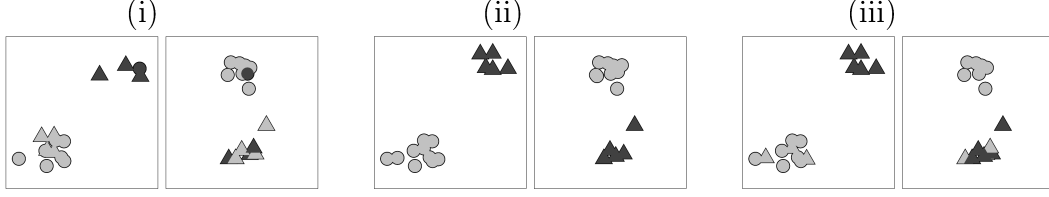


Figure 1: Clusters in the first view are represented with dark and light shades of gray, and clusters in the second view are represented with circles and triangles. (i) The clusterings in the two views are independent, i.e.  $\Pi$  has rank one, so the shade of gray (dark or light) and shape (circle or triangle) are unassociated. (ii) The clusterings in the two views are the same, i.e.  $\Pi$  is diagonal (up to permutation of rows), so the shade of gray (dark or light) and shape (circle or triangle) are perfectly correlated. (iii) The clusterings in the two view are somewhat dependent, i.e.  $\Pi$  is neither diagonal nor rank one.

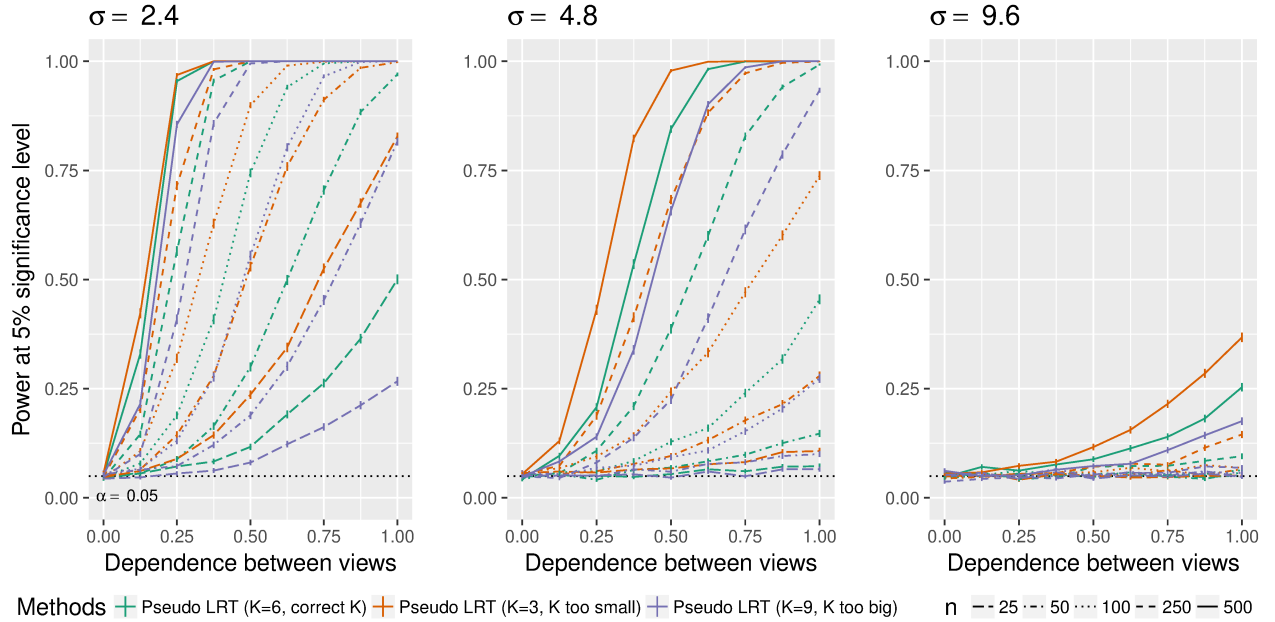


Figure 2: Power of the pseudo likelihood ratio test of  $H_0 : C = 1_{K_1} 1_{K_2}^T$  with  $p = 10$ ,  $K = 6$  and  $\sigma \in \{2.4, 4.8, 9.6\}$  in the simulation setting described in Section 4. The  $x$ -axis displays  $\delta$ , defined in (4.14), and the  $y$ -axis displays the power.



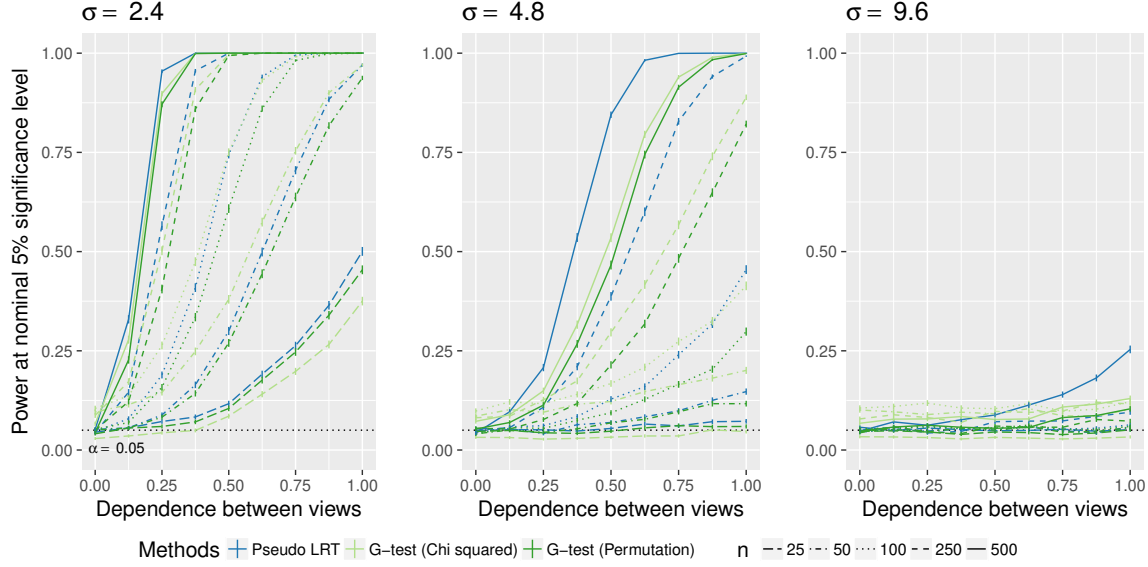


Figure 3: For the simulation study described in Section 5, power of the pseudo likelihood ratio test and the  $G$ -test of independence for  $p = 10$ ,  $K = 6$  and  $\sigma \in \{2.4, 4.8, 9.6\}$ , with  $\delta$ , defined in (4.14), on the  $x$ -axis and power on the  $y$ -axis.

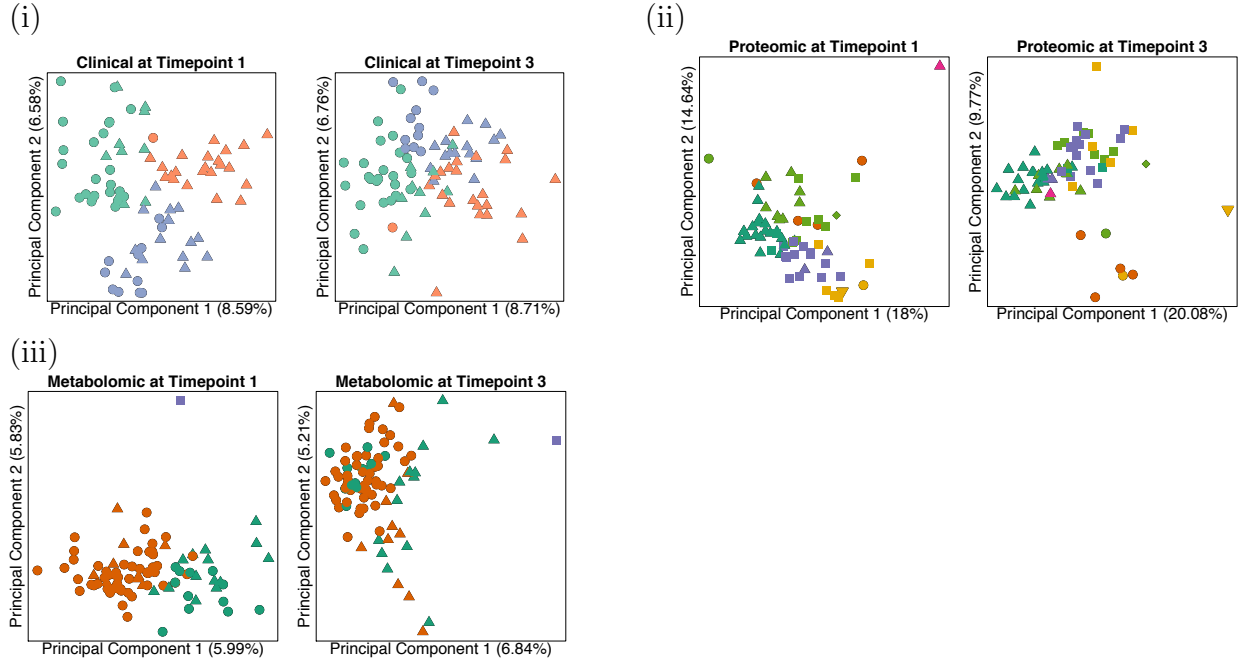


Figure 4: For three different data types, a comparison of the clustering at the first timepoint (represented with colors) with the clustering at the third timepoint (represented with shapes). In each data type, there is strong evidence of dependence ( $p$ -value  $< 0.0001$ ). The data types are (i) clinical measurements, (ii) proteomic measurements, and (iii) metabolomic measurements.

View 1	View 2	$n$	$p_1$	$p_2$	p-value
Clinical at Timepoint 1	Clinical at Timepoint 3	83	204	198	< 0.0001
Proteomic at Timepoint 1	Proteomic at Timepoint 3	66	249	257	< 0.0001
Metabolomic at Timepoint 1	Metabolomic at Timepoint 3	88	641	640	< 0.0001
Clinical at Timepoint 1	Proteomic at Timepoint 1	70	204	249	0.236
Clinical at Timepoint 2	Proteomic at Timepoint 2	60	205	254	0.091
Clinical at Timepoint 3	Proteomic at Timepoint 3	66	198	257	0.950
Clinical at Timepoint 1	Metabolomic at Timepoint 1	98	204	641	0.034
Clinical at Timepoint 2	Metabolomic at Timepoint 2	89	205	641	0.073
Clinical at Timepoint 3	Metabolomic at Timepoint 3	81	198	640	0.328
Proteomic at Timepoint 1	Metabolomic at Timepoint 1	72	249	641	0.402
Proteomic at Timepoint 2	Metabolomic at Timepoint 2	67	254	641	0.004
Proteomic at Timepoint 3	Metabolomic at Timepoint 3	73	257	640	0.020

Table 1: Results from the test of  $H_0 : C = 1_{K(1)}1_{K(2)}^T$  developed in Section 3.1 applied to clinical, proteomic, and metabolomic data at the first and third timepoints, and applied to pairs of data views defined by different data types. Sample sizes  $n$ , dimensions in each view  $p_1$  and  $p_2$ , and p-values obtained using the permutation approximation from Section 3.3 are reported.

## References

- AGRESTI, ALAN. (2003). *Categorical data analysis*, Volume 482. John Wiley & Sons.
- BECK, AMIR AND TEBoulLE, MARC. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters* **31**(3), 167–175.
- BICKEL, STEFFEN AND SCHEFFER, TOBIAS. (2004). Multi-view clustering. In: *ICDM*, Volume 4. pp. 19–26.
- BIEN, JACOB. (2016). The simulator: An engine to streamline simulations. *arXiv preprint arXiv:1607.00021*.
- CANCER GENOME ATLAS RESEARCH NETWORK. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**(7216), 1061–1068.
- CHEN, YONG AND LIANG, KUNG-YEE. (2010). On the asymptotic behaviour of the pseudolikelihood ratio test statistic with boundary problems. *Biometrika* **97**(3), 603–620.
- CUTURI, MARCO. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In: *Advances in Neural Information Systems*. pp. 2292–2300.
- DEMPSTER, ARTHUR P., LAIRD, NAN M. AND RUBIN, DONALD B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1), 1–38.

- FRANKLIN, JOEL AND LORENZ, JENS. (1989). On the scaling of multidimensional matrices. *Linear Algebra and its Applications* **114/115**, 717–735.
- FRIEDMAN, JEROME, HASTIE, TREVOR AND TIBSHIRANI, ROBERT. (2001). *The elements of statistical learning*, Volume 1, Springer Series in Statistics. Springer-Verlag New York.
- GABASOVA, EVELINA, REID, JOHN AND WERNISCH, LORENZ. (2017). Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. *PLoS Computational Biology* **13**(10), e1005781.
- GONG, GAIL AND SAMANIEGO, FRANCISCO J. (1981). Pseudo maximum likelihood estimation: theory and applications. *The Annals of Statistics* **9**(4), 861–869.
- HASTIE, TREVOR, TIBSHIRANI, ROBERT, NARASIMHAN, BALASUBRAMANIAN AND CHU, GILBERT. (2017). *impute: Imputation for microarray data*,. R package version 1.50.1.
- HUBERT, LAWRENCE AND ARABIE, PHIPPS. (1985). Comparing partitions. *Journal of Classification* **2**(1), 193–218.
- KIRK, PAUL, GRIFFIN, JIM E., SAVAGE, RICHARD S., GHAHRAMANI, ZOUBIN AND WILD, DAVID L. (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* **28**(24), 3290–3297.
- KIVINEN, JYRKI AND WARMUTH, MANFRED K. (1997). Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation* **132**(1), 1–63.
- KUMAR, ABHISHEK, RAI, PIYUSH AND DAUME, HAL. (2011). Co-regularized multi-view spectral clustering. In: *Advances in Neural Information Processing Systems*. pp. 1413–1421.
- LOCK, ERIC F. AND DUNSON, DAVID B. (2013). Bayesian consensus clustering. *Bioinformatics* **29**(20), 2610–2616.
- MCLACHLAN, GEOFFREY AND KRISHNAN, THRIYAMBAM. (2007). *The EM algorithm and extensions*, Volume 382. John Wiley & Sons.
- MCLACHLAN, GEOFFREY AND PEEL, DAVID. (2000). *Finite mixture models*. John Wiley & Sons.
- MEILÄ, MARINA. (2007). Comparing clusterings – an information based distance. *Journal of Multivariate Analysis* **98**(5), 873–895.
- MIRKIN, BORIS. (2011). Choosing the number of clusters. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **1**(3), 252–260.
- PRICE, NATHAN D., MAGIS, ANDREW T., EARLS, JOHN C., GLUSMAN, GUSTAVO, LEVY, ROIE, LAUSTED, CHRISTOPHER, McDONALD, DANIEL T., KUSEBAUCH, ULRIKE, MOSS, CHRISTOPHER L., ZHOU, YONG, QIN, SHIZHEN, MORITZ, ROBERT L., BROGAARD, KRISTIN, OMENN, GILBERT S., LOVEJOY, JENNIFER C. and others.

- (2017). A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nature Biotechnology* **35**(8), 747–756.
- REID, NANCY. (1995). The roles of conditioning in inference. *Statistical Science* **10**(2), 138–157.
- ROCKAFELLAR, R. TYRRELL AND WETS, ROGER J.-B. (1998). *Variational analysis*, Volume 317, Grundlehren der mathematischen Wissenschaften. Springer-Verlag Berlin Heidelberg.
- ROGERS, SIMON, GIROLAMI, MARK, KOLCH, WALTER, WATERS, KATRINA M., LIU, TAO, THRALL, BRIAN AND WILEY, H. STEVEN. (2008). Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics* **24**(24), 2894–2900.
- SCRUCCA, LUCA, FOP, MICHAEL, MURPHY, T. BRENDAN AND RAFTERY, ADRIAN E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal* **8**(1), 289.
- SELF, STEVEN G. AND LIANG, KUNG-YEE. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**(398), 605–610.
- SHEN, RONGLAI, OLSHEN, ADAM B. AND LADANYI, MARC. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25**(22), 2906–2912.
- SUGIYAMA, MAHITO, NAKAHARA, HIROYUKI AND TSUDA, KOJI. (2017). Tensor balancing on statistical manifold. *arXiv preprint arXiv:1702.08142*.
- VERSHYNIN, ROMAN. (2012). Introduction to the non-asymptotic analysis of random matrices. In: Eldar, Yonina C. and Kutyniok, Gitta (editors), *Compressed Sensing: Theory and Applications*, Chapter 5. Cambridge: Cambridge University Press, pp. 210–268.

# Are Clusterings of Multiple Data Views Independent?

## Supplementary Materials

### A Proofs

#### A.1 Proof of Proposition 1

Suppose  $\pi^{(1)} \in \Delta_+^{K^{(1)}}$  and  $\pi^{(2)} \in \Delta_+^{K^{(2)}}$ , where  $\Delta_+^K \equiv \{s \in \mathbb{R}^K : s_k > 0, \sum_{k=1}^K s_k = 1\}$ . Let  $\mathcal{C}_{\pi^{(1)}, \pi^{(2)}} = \{C \in \mathbb{R}^{K^{(1)} \times K^{(2)}} : C_{kk'} \geq 0, C\pi^{(2)} = 1_{K^{(1)}}, C^T\pi^{(1)} = 1_{K^{(2)}}\}$ . First, we show that  $\{\Pi \in \Delta^{K^{(1)} \times K^{(2)}} : \Pi 1_{K^{(2)}} = \pi^{(1)}, \Pi^T 1_{K^{(1)}} = \pi^{(2)}\} \subseteq \{\text{diag}(\pi^{(1)})C\text{diag}(\pi^{(2)}) : C \in \mathcal{C}_{\pi^{(1)}, \pi^{(2)}}\}$ .

Suppose  $\Pi \in \Delta^{K^{(1)} \times K^{(2)}} = \{S \in \mathbb{R}^{K^{(1)} \times K^{(2)}} : S_{kk'} \geq 0, \sum_{k=1}^{K^{(1)}} \sum_{k'=1}^{K^{(2)}} S_{kk'} = 1\}$  such that  $\Pi \cdot 1_{K^{(2)}} = \pi^{(1)}$  and  $\Pi^T \cdot 1_{K^{(1)}} = \pi^{(2)}$ . Define  $C \in \mathbb{R}_+^{K^{(1)} \times K^{(2)}}$  by  $C_{kk'} = \frac{\Pi_{kk'}}{\pi_k^{(1)} \pi_{k'}^{(2)}}$ . The denominator is nonzero because  $\pi^{(1)} \in \Delta_+^{K^{(1)}}$  and  $\pi^{(2)} \in \Delta_+^{K^{(2)}}$ . Further,  $C_{kk'} \geq 0$  and  $\Pi = \text{diag}(\pi^{(1)})C\text{diag}(\pi^{(2)})$ . Now,

$$C\pi^{(2)} = \begin{bmatrix} \sum_{k'} C_{1k'} \pi_{k'}^{(2)} \\ \vdots \\ \sum_{k'} C_{K^{(1)}k'} \pi_{k'}^{(2)} \end{bmatrix} = \begin{bmatrix} \sum_{k'} \frac{\Pi_{1k'}}{\pi_1^{(1)}} \\ \vdots \\ \sum_{k'} \frac{\Pi_{K^{(1)}k'}}{\pi_{K^{(1)}}^{(1)}} \end{bmatrix} = 1_{K^{(1)}}$$

since  $\Pi 1_{K^{(2)}} = \pi^{(1)}$  implies that  $\sum_{k'} \Pi_{kk'} = \pi_k^{(1)}$ . Similarly, we can show  $C^T\pi^{(1)} = 1_{K^{(2)}}$ .

Next, we show that

$$\{\Pi \in \Delta^{K^{(1)} \times K^{(2)}} : \Pi 1_{K^{(2)}} = \pi^{(1)}, \Pi^T 1_{K^{(1)}} = \pi^{(2)}\} \supseteq \{\text{diag}(\pi^{(1)})C\text{diag}(\pi^{(2)}) : C \in \mathcal{C}_{\pi^{(1)}, \pi^{(2)}}\}.$$

Suppose  $C \in \mathbb{R}^{K^{(1)} \times K^{(2)}}$  such that  $C_{kk'} \geq 0, C\pi^{(2)} = 1_{K^{(1)}}$  and  $C^T\pi^{(1)} = 1_{K^{(2)}}$ . Define

$$\Pi = \text{diag}(\pi^{(1)})C\text{diag}(\pi^{(2)}).$$

Then  $\Pi_{kk'} = \pi_k^{(1)} C_{kk'} \pi_{k'}^{(2)}$ . Since  $\pi_k^{(1)} > 0, C_{kk'} \geq 0$  and  $\pi_{k'}^{(2)} > 0$ , it follows that  $\Pi_{kk'} \geq 0$ . Now,

$$\begin{aligned} \Pi 1_{K^{(2)}} &= \text{diag}(\pi^{(1)})C\text{diag}(\pi^{(2)})1_{K^{(2)}} \\ &= \text{diag}(\pi^{(1)})C\pi^{(2)} \\ &= \text{diag}(\pi^{(1)})1_{K^{(1)}} = \pi^{(1)}. \end{aligned}$$

Similarly, we can show  $\Pi^T 1_{K^{(1)}} = \pi^{(2)}$ . Further, since  $\pi^{(1)} \in \Delta_+^{K^{(1)}}$ ,  $\sum_{kk'} \Pi_{kk'} = 1_{K^{(1)}}^T \Pi 1_{K^{(2)}} = 1_{K^{(1)}}^T \pi^{(1)} = 1$ , so  $\Pi \in \Delta^{K^{(1)} \times K^{(2)}}$ .

Hence, we have proved Proposition 1.

## A.2 Proof of Proposition 4

Let  $\sigma^2 > 0$ . Suppose throughout that we fit the model (2.1)–(2.2), for  $\phi^{(1)}(\cdot; \theta)$  the density of a  $N_{p_1}(\theta, \sigma^2 I_{p_1})$  random variable and  $\phi^{(2)}(\cdot; \theta)$  the density of a  $N_{p_2}(\theta, \sigma^2 I_{p_2})$  random variable. This amounts to applying Gaussian mixture model-based clustering with common covariance matrix  $\sigma^2 I$  to each view.

Let  $\hat{\pi}^{(1)}$ ,  $\hat{\pi}^{(2)}$ ,  $\hat{\theta}^{(1)}$ , and  $\hat{\theta}^{(2)}$  denote the maximizers of (2.7), and let

$$\hat{r}_i^{(1)} = \frac{\hat{\phi}_i^{(1)}}{1^T \hat{\phi}_i^{(1)}}, \quad \hat{r}_i^{(2)} = \frac{\hat{\phi}_i^{(2)}}{1^T \hat{\phi}_i^{(2)}}, \quad (\text{A.1})$$

where  $\hat{\phi}^{(1)}$  and  $\hat{\phi}^{(2)}$  are defined in (2.9).

Since Gaussian mixture model-based clustering with common covariance matrix  $\nu^2 I$  converges to k-means clustering in each view as  $\nu^2 \rightarrow 0$  (see Section 14.3.7 in Friedman *and others* (2001) for details), as  $\sigma^2 \rightarrow 0$ ,

$$\hat{r}_{ik}^{(1)} \rightarrow \mathbb{1}\{\tilde{M}_i^{(1)} = k\}, \quad \hat{r}_{ik'}^{(2)} \rightarrow \mathbb{1}\{\tilde{M}_i^{(2)} = k'\}, \quad \hat{\pi}_k^{(1)} \rightarrow \frac{\tilde{N}_{k\cdot}}{n}, \quad \hat{\pi}_{k'}^{(2)} \rightarrow \frac{\tilde{N}_{\cdot k'}}{n}, \quad (\text{A.2})$$

where  $\tilde{M}_i^{(1)}$  and  $\tilde{M}_i^{(2)}$  are the estimated k-means cluster assignments of the  $i$ th observation in each view,  $\tilde{N}_{k\cdot} = \sum_{k'} \tilde{N}_{kk'}$ , and  $\tilde{N}_{\cdot k'} = \sum_k \tilde{N}_{kk'}$ , for  $\tilde{N}_{kk'} = |\{i \in \{1, \dots, n\} : \tilde{M}_i^{(1)} = k, \tilde{M}_i^{(2)} = k'\}|$ .

We now rewrite (2.8) as

$$\begin{aligned} \hat{C} &= \arg \min_{C \in \mathcal{C}_{\hat{\pi}^{(1)}, \hat{\pi}^{(2)}}} \left[ - \sum_{i=1}^n \log \left( \sum_{k=1}^{K^{(1)}} \sum_{k'=1}^{K^{(2)}} \hat{\pi}_k^{(1)} C_{kk'} \hat{\pi}_{k'}^{(2)} \hat{\phi}_{ik}^{(1)} \hat{\phi}_{ik'}^{(2)} \right) \right] \\ &= \arg \min_{C \in \mathcal{C}_{\hat{\pi}^{(1)}, \hat{\pi}^{(2)}}} \left[ - \sum_{i=1}^n \log \left( \sum_{k=1}^{K^{(1)}} \sum_{k'=1}^{K^{(2)}} \hat{\pi}_k^{(1)} C_{kk'} \hat{\pi}_{k'}^{(2)} \hat{\phi}_{ik}^{(1)} \hat{\phi}_{ik'}^{(2)} \right) - \sum_{i=1}^n \log \left( 1^T \hat{\phi}_i^{(1)} 1^T \hat{\phi}_i^{(2)} \right) \right] \\ &= \arg \min_{C \in \mathcal{C}_{\hat{\pi}^{(1)}, \hat{\pi}^{(2)}}} \left[ - \sum_{i=1}^n \log \left( \sum_{k=1}^{K^{(1)}} \sum_{k'=1}^{K^{(2)}} \hat{\pi}_k^{(1)} C_{kk'} \hat{\pi}_{k'}^{(2)} \hat{r}_{ik}^{(1)} \hat{r}_{ik'}^{(2)} \right) \right], \end{aligned}$$

where the second equality holds because the quantities  $\hat{\phi}_i^{(1)}$  and  $\hat{\phi}_i^{(2)}$  defined in (2.9) do not depend on  $C$ , and the third equality follows from the definition of  $\hat{r}_i^{(1)}$  and  $\hat{r}_i^{(2)}$  in (A.1). It follows that

$$\hat{C} = \arg \min_{C \in \mathcal{C}_{\hat{\pi}^{(1)}, \hat{\pi}^{(2)}}} g^\sigma(C), \quad (\text{A.3})$$

where

$$g^\sigma(C) \equiv - \sum_{i=1}^n \log \left( \sum_{k=1}^{K^{(1)}} \sum_{k'=1}^{K^{(2)}} \hat{\pi}_k^{(1)} C_{kk'} \hat{\pi}_{k'}^{(2)} \hat{r}_{ik}^{(1)} \hat{r}_{ik'}^{(2)} \right).$$

By (A.2), as  $\sigma^2 \rightarrow 0$ ,  $g^\sigma$  converges pointwise to  $g$ , where

$$\begin{aligned}
g(C) &\equiv - \sum_{i=1}^n \log \left( \sum_{k=1}^{K^{(1)}} \sum_{k'=1}^{K^{(2)}} \frac{\tilde{N}_{k.}}{n} C_{kk'} \frac{\tilde{N}_{.k'}}{n} \mathbb{1}\{\tilde{M}_i^{(1)} = k\} \mathbb{1}\{\tilde{M}_i^{(2)} = k'\} \right) \\
&= - \sum_{i=1}^n \sum_{k=1}^{K^{(1)}} \sum_{k'=1}^{K^{(2)}} \mathbb{1}\{\tilde{M}_i^{(1)} = k\} \mathbb{1}\{\tilde{M}_i^{(2)} = k'\} \log \left( \frac{\tilde{N}_{k.}}{n} C_{kk'} \frac{\tilde{N}_{.k'}}{n} \right) \\
&= - \sum_{k=1}^{K^{(1)}} \sum_{k'=1}^{K^{(2)}} \tilde{N}_{kk'} \log \left( \frac{\tilde{N}_{k.}}{n} C_{kk'} \frac{\tilde{N}_{.k'}}{n} \right).
\end{aligned}$$

Applying the method of Lagrange multipliers, we find that  $\tilde{C} \equiv \arg \min_{C \in \mathcal{C}_{\hat{\pi}^{(1)}, \hat{\pi}^{(2)}}} g(C)$  satisfies

$$\tilde{C}_{kk'} = \frac{n \tilde{N}_{kk'}}{\tilde{N}_{k.} \tilde{N}_{.k'}}. \quad (\text{A.4})$$

By Exercise 7.23(c) in Rockafellar and Wets (1998),  $\{g^\sigma(C)\}_{\sigma>0}$  is essentially bounded. By Theorem 7.17 in Rockafellar and Wets (1998), the epigraphical limit of  $g^\sigma$  is  $g$ . Finally,  $g^\sigma$  and  $g$  are continuous and proper. Hence, by Theorem 7.33 in Rockafellar and Wets (1998),

$$\tilde{C} = \arg \min_{C \in \mathcal{C}_{\hat{\pi}^{(1)}, \hat{\pi}^{(2)}}} g(C) = \lim_{\sigma \rightarrow 0} \arg \min_{C \in \mathcal{C}_{\hat{\pi}^{(1)}, \hat{\pi}^{(2)}}} g^\sigma(C). \quad (\text{A.5})$$

By (A.3), (A.4) and (A.5), as  $\sigma^2 \rightarrow 0$  we have

$$\hat{C}_{kk'} \rightarrow \tilde{C}_{kk'} = \frac{n \tilde{N}_{kk'}}{\tilde{N}_{k.} \tilde{N}_{.k'}}. \quad (\text{A.6})$$

By (A.2) and (A.6), and the definition of  $\hat{\Pi}$  in Algorithm 1,

$$\hat{\Pi}_{kk'} \rightarrow \frac{\tilde{N}_{kk'}}{n}. \quad (\text{A.7})$$

Applying (A.2) and (A.7) to (5.18) yields the result.

## B Exponentiated gradient descent for solving (2.8)

After a transformation of the optimization problem, (2.8) can be efficiently solved using exponentiated gradient descent (Kivinen and Warmuth, 1997), a first-order method specially designed for optimization over the simplex. This is a form of mirror descent, with provable convergence results (Beck and Teboulle, 2003). While there is no analytic solution for the update performed at each iteration of the exponentiated gradient descent, each update can be performed by applying the Sinkhorn-Knopp algorithm, a matrix balancing algorithm with provable convergence results and linear convergence rates (Franklin and Lorenz, 1989).

Define

$$\ell(\theta^{(1)}, \theta^{(2)}, \Pi) = \sum_{i=1}^n \log f(X_i^{(1)}, X_i^{(2)}; \theta^{(1)}, \theta^{(2)}, \Pi)$$

where  $f(\cdot, \cdot; \theta^{(1)}, \theta^{(2)}, \Pi)$  is defined in (2.3). Consider the following optimization problem:

$$\begin{aligned} & \underset{\Pi}{\text{minimize}} && -\ell(\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \Pi) \\ & \text{subject to} && \Pi 1_{K^{(2)}} = \hat{\pi}^{(1)} \\ & && \Pi^T 1_{K^{(1)}} = \hat{\pi}^{(2)} \\ & && \Pi_{kk'} \geq 0. \end{aligned} \tag{B.1}$$

By Proposition 1, we can equivalently write (B.1) as follows:

$$\begin{aligned} & \underset{C}{\text{minimize}} && -\ell(\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \hat{\pi}^{(1)}, \hat{\pi}^{(2)}, C) \\ & \text{subject to} && \text{diag}(\hat{\pi}^{(1)}) C \text{diag}(\hat{\pi}^{(2)}) 1_{K^{(2)}} = \hat{\pi}^{(1)} \\ & && \text{diag}(\hat{\pi}^{(2)}) C^T \text{diag}(\hat{\pi}^{(1)}) 1_{K^{(1)}} = \hat{\pi}^{(2)} \\ & && C_{kk'} \geq 0 \end{aligned}$$

which is equivalent to

$$\begin{aligned} & \underset{C}{\text{minimize}} && -\ell(\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \hat{\pi}^{(1)}, \hat{\pi}^{(2)}, C) \\ & \text{subject to} && \text{diag}(\hat{\pi}^{(1)}) C \hat{\pi}^{(2)} = \hat{\pi}^{(1)} \\ & && \text{diag}(\hat{\pi}^{(2)}) C^T \hat{\pi}^{(1)} = \hat{\pi}^{(2)} \\ & && C_{kk'} \geq 0 \end{aligned}$$

which is in turn equivalent to

$$\begin{aligned} & \underset{C}{\text{minimize}} && -l(\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \hat{\pi}^{(1)}, \hat{\pi}^{(2)}, C) \\ & \text{subject to} && C \hat{\pi}^{(2)} = 1_{K^{(1)}} \\ & && C^T \hat{\pi}^{(1)} = 1_{K^{(2)}} \\ & && C_{kk'} \geq 0, \end{aligned}$$

which is (2.8), the optimization problem we must solve to estimate  $\hat{C}$ . Hence, to find  $\hat{C}$ , we can solve (B.1); let  $\hat{\Pi}$  be the minimizer of (B.1). Then,  $\hat{C}$  can be found by

$$\hat{C}_{kk'} = \frac{\hat{\Pi}_{kk'}}{\hat{\pi}_k^{(1)} \hat{\pi}_{k'}^{(2)}}. \tag{B.2}$$

The motivation for this transformation of (2.8) is that the transformed problem (B.1) can be efficiently solved using an algorithm described in Cuturi (2013).



We will describe the exponentiated gradient algorithm to solve the general problem

$$\begin{aligned}
& \underset{\Pi}{\text{minimize}} && g(\Pi) \\
& \text{subject to} && \sum_{k'} \Pi_{kk'} = \hat{\pi}_k^{(1)} \\
& && \sum_k \Pi_{kk'} = \hat{\pi}_{k'}^{(2)} \\
& && \Pi_{kk'} \geq 0
\end{aligned} \tag{B.3}$$

To solve (B.3), we apply the update

$$\hat{\Pi}^{t+1} = \begin{cases} \arg \min_{\Pi} & g(\hat{\Pi}^t) + \langle \nabla g(\hat{\Pi}^t), \Pi - \hat{\Pi}^t \rangle + \frac{1}{s} \sum_{kk'} \Pi_{kk'} \log(\Pi_{kk'} / \hat{\Pi}_{kk'}^t) \\ \text{subject to} & \sum_{k'} \Pi_{kk'} = \hat{\pi}_k^{(1)} \\ & \sum_k \Pi_{kk'} = \hat{\pi}_{k'}^{(2)} \\ & \Pi_{kk'} \geq 0. \end{cases} \tag{B.4}$$

This is similar to the proximal gradient method, but instead of  $\|\Pi - \hat{\Pi}^t\|_F^2 / (2s)$  we use the Bregman divergence,

$$\frac{1}{s} \sum_{kk'} \Pi_{kk'} \log(\Pi_{kk'} / \hat{\Pi}_{kk'}^t).$$

An advantage of this choice is that the positivity constraint is automatically enforced. For more on this, see Beck and Teboulle (2003). The optimality conditions for the problem (B.4) are

$$[\nabla g(\hat{\Pi}^t)]_{kk'} + [1 + \log(\hat{\Pi}_{kk'}^{t+1} / \hat{\Pi}_{kk'}^t)] / s + \lambda_k + \eta_{k'} = 0, \tag{B.5}$$

for  $1 \leq k \leq K^{(1)}$ ,  $1 \leq k' \leq K^{(2)}$ , where  $\lambda_k$  and  $\eta_{k'}$  are Lagrange multipliers for the row sum and column sum constraints, respectively. This implies that

$$\hat{\Pi}_{kk'}^{t+1} = \hat{\Pi}_{kk'}^t \exp\{-s\lambda_k - s\eta_{k'} - s[\nabla g(\hat{\Pi}^t)]_{kk'} - 1\}.$$

The gradient is given by

$$\nabla g(\Pi) = - \sum_{i=1}^n \frac{\hat{\phi}_i^{(2)} [\hat{\phi}_i^{(1)}]^T}{[\hat{\phi}_i^{(1)}]^T \Pi \hat{\phi}_i^{(2)}}. \tag{B.6}$$

In the special case of (B.1), writing  $G_{kk'} = [\nabla g(\hat{\Pi}^t)]_{kk'}$ , the update

$$\hat{\Pi}_{kk'}^{t+1} = \hat{\Pi}_{kk'}^t \exp\{sG_{kk'} - 1\} \exp\{-s\lambda_k\} \exp\{-s\eta_{k'}\}$$

can be written as

$$\hat{\Pi}^{t+1} = \text{diag}(v) M \text{diag}(u),$$

---

**Algorithm S1** Exponentiated Gradient Descent for Solving (B.1)

---

1. Choose a fixed step size  $s$  and compute  $\hat{\phi}^{(1)}$  and  $\hat{\phi}^{(2)}$  according to (2.9).
2. For  $t = 1, 2, \dots$ , until convergence,

(a) Define

$$M_{kk'} = \hat{\Pi}_{kk'}^t \exp\{sG_{kk'} - 1\}$$

where

$$G_{kk'} = \sum_{i=1}^n \frac{\hat{\phi}_{ik}^{(1)} \hat{\phi}_{ik'}^{(2)}}{[\hat{\phi}_i^{(1)}]^T \hat{\Pi}^t \hat{\phi}_i^{(2)}}.$$

(b) (Sinkhorn-Knopp algorithm) Let  $u^0 = 1_{K^{(2)}}$  and  $v^0 = 1_{K^{(1)}}$ . For  $t' = 1, 2, \dots$ , until convergence,

- $u^{t'} = \frac{\hat{\pi}^{(2)}}{(\hat{\Pi}^{t+1})^T v^{t'-1}}$ , where the fraction denotes element-wise vector division
- $v^{t'} = \frac{\hat{\pi}^{(1)}}{\hat{\Pi}^{t+1} u^{t'}}$ , where the fraction denotes element-wise vector division

(c) Let  $u$  and  $v$  denote the vectors to which  $u^{t'}$  and  $v^{t'}$  converge. Update

$$\hat{\Pi}_{kk'}^{t+1} = u_k M_{kk'} v_{k'}.$$


---

where  $u_k = \exp\{-s\lambda_k\}$ ,  $v_{k'} = \exp\{-s\eta_{k'}\}$ , and  $M_{kk'} = \hat{\Pi}_{kk'}^t \exp\{sG_{kk'} - 1\}$ .

Since  $\hat{\Pi}^{t+1}$  must satisfy the row and column sum constraints,  $u$  and  $v$  must be chosen accordingly. As in Cuturi (2013), we can apply the Sinkhorn Theorem and Sinkhorn-Knopp algorithm to find  $u$  and  $v$ . By the Sinkhorn Theorem,  $u$  and  $v$  are unique modulo scalar multiplication of  $u$  with a positive number and scalar division of  $v$  by that same positive number. The Sinkhorn-Knopp algorithm (alternatively rescaling the columns so that the rows sum to  $\hat{\pi}^{(1)}$  then rescaling the rows so that the columns sum to  $\hat{\pi}^{(2)}$ ) can be applied to  $M_{kk'}$  to find  $u$  and  $v$ . Hence, to perform the update, we simply multiply  $\hat{\Pi}_{kk'}^t$  by  $\exp\{sG_{kk'} - 1\}$  and then apply the Sinkhorn-Knopp algorithm to the updated matrix so that the row and column sum constraints are satisfied. Algorithm S1 provides the details. Using Proposition (1), we can then use our bijection between  $\Pi$  to  $C$  (B.2) to obtain Algorithm 1 from Algorithm S1.

## C Simulations

### C.1 Mean matrices for simulations in Section 4

In the simulations described in Section 4, data are generated from (2.1)–(2.2) with

$$\Pi = \frac{1-\delta}{K^2} 1_K 1_K^T + \frac{\delta}{K} I_K \quad (\text{C.1})$$

for  $K = 6$ . In the  $l$ th data view, the observations are drawn from a multivariate Gaussian mixture model, for which the  $k$ th component in the mixture (corresponding to the  $k$ th cluster) is a  $N_p(\mu_k^{(l)}, \sigma^2 I_p)$  distribution, with  $p = 10$ . The  $p \times K$  mean matrices for the two data views are of the form

$$\begin{aligned}\mu^{(1)} &= \begin{bmatrix} 2 \cdot 1_5 & 0_5 & 2 \cdot 1_5 & -2 \cdot 1_5 & 0_5 & -2 \cdot 1_5 \\ 0_5 & 2 \cdot 1_5 & -2 \cdot 1_5 & 0_5 & -2 \cdot 1_5 & 2 \cdot 1_5 \end{bmatrix}, \\ \mu^{(2)} &= \begin{bmatrix} -2 \cdot 1_6 & 0_6 & -2 \cdot 1_6 & 2 \cdot 1_6 & 0_4 & 2 \cdot 1_4 \\ 0_4 & -2 \cdot 1_4 & 2 \cdot 1_4 & 0_4 & 2 \cdot 1_6 & -2 \cdot 1_6 \end{bmatrix}.\end{aligned}$$

## C.2 Supplementary simulations to Section 4

### C.2.1 Selection of the number of clusters

Recall from Section 4 that using too few clusters in the pseudo likelihood ratio test sometimes yields better power than using the correct number of clusters. In this section, we demonstrate that using too few clusters can either increase or decrease the power, depending on the situation.

We generate data from (2.1)–(2.2) with  $\Pi = \frac{1-\delta}{K^2} 1_K 1_K^T + \frac{\delta}{K} I_K$  for  $K = 6$ . In the  $l$ th data view, the observations are drawn from a bivariate Gaussian mixture model, for which the  $k$ th component in the mixture (corresponding to the  $k$ th cluster) is a  $N_2(\mu_k^{(l)}, 0.4^2 I_p)$  distribution. We simulate 2000 datasets for a range of values of  $n$ , and for two choices of  $\mu^{(l)}$ :

*Choice 1:*

$$\mu^{(1)} = \begin{bmatrix} 2 & 2 & -2 & -2 \\ -2 & -1 & 1 & 2 \end{bmatrix}, \quad \mu^{(2)} = \begin{bmatrix} -2 & -2 & 2 & 2 \\ -2 & -1 & 1 & 2 \end{bmatrix}, \quad (\text{C.2})$$

*Choice 2:*

$$\mu^{(1)} = \begin{bmatrix} 2 & 2 & -2 & -2 \\ -2 & -1 & 1 & 2 \end{bmatrix}, \quad \mu^{(2)} = \begin{bmatrix} 2 & -2 & -2 & 2 \\ 2 & -2 & -1 & 1 \end{bmatrix}. \quad (\text{C.3})$$

We evaluate the power of the pseudo likelihood ratio test of  $H_0 : C = 1_{K_1} 1_{K_2}^T$  at nominal significance level  $\alpha = 0.05$ , when the number of clusters is incorrectly and correctly specified. Results are displayed in Figure S1.

Observe from the left panel of Figure S1 that when  $\mu^{(l)}$  is given in (C.2), using too few clusters ( $K = 2$ ) yields higher power than the correct number of clusters ( $K = 4$ ). This is because under (C.2), the clusterings on each view contain two natural “meta-clusters”, formed by combining clusters whose means are close; see Figure S2(i) for an illustration. Because the clusters on each view are not well-separated, it is easier to instead cluster the data into the two “meta-clusters”, which are highly in agreement when the four clusters are in agreement. For example, when  $\Pi = I_4/4$ , the  $\Pi$  matrix corresponding to the “meta-clustering” is given by  $I_2/2$ . Thus, testing for independence assuming just two clusters yields better power than correctly assuming four clusters.

By contrast, observe from the right panel of Figure S1 that when  $\mu^{(l)}$  is given in (C.3), using too few clusters ( $K = 2$ ) yields much lower power than the correct number of clusters ( $K = 4$ ). Again, under (C.3), the clusterings on each view contain two natural “meta-clusters” (see Figure S2(ii)), and it is easier to cluster the data into the two “meta-clusters”. However, under (C.3), even when the four clusters are highly in agreement, the meta-clusters are not highly in agreement. For example, when  $\Pi = I_4/4$ , the  $\Pi$  matrix corresponding to the “meta-clustering” is given by  $1_2 1_2^T/4$ . Thus, testing for independence assuming just two clusters yields worse power than correctly assuming four clusters.

We also observe in both panels of Figure S1 that using too many clusters yields slightly lower power than using the correct number of clusters under both (C.2) and (C.3); this result is similar to the results from the simulation setting described in Section 4 and Appendix C.2.2. Furthermore, in both panels of Figure S1, corresponding to the two choices of  $\mu^{(l)}$ , choosing the clusters using BIC on each view yields slightly lower power than using the correct number of clusters in small and moderate sample sizes, and performs as well as using the correct number of clusters when the sample size is large.

### C.2.2 Additional values of $K$ and $p$

We simulate two data views with  $p_1 = p_2 = p = 10$ , and  $K^{(1)} = K^{(2)} = K = 3$ . In the  $l$ th data view, the observations are drawn from a multivariate Gaussian mixture model, for which the  $k$ th component in the mixture (corresponding to the  $k$ th cluster) is a  $N_p(\mu_k^{(l)}, \sigma^2 I_p)$  distribution.

The means of the components in the mixture model, written as a  $p \times K$  matrix, are

$$\mu^{(1)} = \begin{bmatrix} 2 \cdot 1_5 & 0_5 & 2 \cdot 1_5 \\ 0_5 & 2 \cdot 1_5 & -2 \cdot 1_5 \end{bmatrix}, \quad \mu^{(2)} = \begin{bmatrix} -2 \cdot 1_6 & 0_6 & -2 \cdot 1_6 \\ 0_4 & -2 \cdot 1_4 & 2 \cdot 1_4 \end{bmatrix}.$$

Additionally, we simulate two data views with  $p_1 = p_2 = p = 100$ , and  $K^{(1)} = K^{(2)} = K$ , for  $K = 3$  and  $K = 6$ . In the  $l$ th data view, the observations are drawn from a multivariate Gaussian mixture model, for which the  $k$ th component in the mixture (corresponding to the  $k$ th cluster) is a  $N_p(\mu_k^{(l)}, \sigma^2 I_p)$  distribution. For  $K = 3$ , the means of the components in the mixture model, written as a  $p \times K$  matrix, are

$$\mu^{(1)} = \begin{bmatrix} 2 \cdot 1_{50} & 0_{50} & 2 \cdot 1_{50} \\ 0_{50} & 2 \cdot 1_{50} & -2 \cdot 1_{50} \end{bmatrix},$$

$$\mu^{(2)} = \begin{bmatrix} -2 \cdot 1_{60} & 0_{60} & -2 \cdot 1_{60} \\ 0_{40} & -2 \cdot 1_{40} & 2 \cdot 1_{40} \end{bmatrix}.$$

For  $K = 6$ , the means are

$$\mu^{(1)} = \begin{bmatrix} 2 \cdot 1_{50} & 0_{50} & 2 \cdot 1_{50} & -2 \cdot 1_{50} & 0_{50} & -2 \cdot 1_{50} \\ 0_{50} & 2 \cdot 1_{50} & -2 \cdot 1_{50} & 0_{50} & -2 \cdot 1_{50} & 2 \cdot 1_{50} \end{bmatrix},$$

$$\mu^{(2)} = \begin{bmatrix} -2 \cdot 1_{60} & 0_{60} & -2 \cdot 1_{60} & 2 \cdot 1_{60} & 0_{40} & 2 \cdot 1_{40} \\ 0_{40} & -2 \cdot 1_{40} & 2 \cdot 1_{40} & 0_{40} & 2 \cdot 1_{60} & -2 \cdot 1_{60} \end{bmatrix}.$$

To investigate the type I error and power of our test, we generate data according to (2.1)–(2.2), with a range of  $\Pi$  defined in (C.1).

We simulate 2000 datasets for a range of values of  $n$  and  $\sigma$ , and evaluate the power of the pseudo likelihood ratio test of  $H_0 : C = 1_{K_1} 1_{K_2}^T$  at nominal significance level  $\alpha = 0.05$ , when the number of clusters is correctly and incorrectly specified. Results are shown in Figures S3, S4, and S5.

Results are similar to the simulation study described in Section 4.

### C.3 Supplementary simulations to Section 5

In this section, we consider the  $G$ -test for independence in Section 5 using  $\hat{M}^{(1)}$  and  $\hat{M}^{(2)}$  defined in (5.19). As in Section 5, we perform a simulation study in order to compare the performances of the pseudo likelihood ratio test for testing the null hypothesis  $H_0 : C = 1_{K^{(1)}} 1_{K^{(2)}}^T$  and the  $G$ -test for independence. We obtain p-values for  $G^2(\hat{M}^{(1)}, \hat{M}^{(2)})$  (5.15) using the  $\chi^2$  approximation from (5.16), as well as a permutation approach, where we take  $B$  permutations of the elements of  $\hat{M}^{(2)}$ , and compare the observed value of  $G^2(\hat{M}^{(1)}, \hat{M}^{(2)})$  to its empirical distribution in these permutation samples.

#### C.3.1 Additional values of $p$ and $K$

We return to the simulation set-up described in Section 4 and Appendix C.1, with  $\Sigma^{(1)} = \Sigma^{(2)} = \sigma^2 I_p$ , and investigate a range of values of  $p$  and  $K$ . In addition to the  $G$ -test for independence and the pseudo likelihood ratio test, we also compute the adjusted Rand Index (ARI) of Hubert and Arabie (1985) in order to compare the results of model-based clustering (implemented as in Sections 2.3 and 3) on each view; p-values for the ARI are obtained using a permutation approach. We compare the performance of the pseudo likelihood ratio test, the  $G$ -test for independence, and the ARI for testing the null hypothesis  $H_0 : C = 1_{K^{(1)}} 1_{K^{(2)}}^T$ . The results are in Figures S6, S7, and S8. Results are similar to results from the  $K = 6$  and  $p = 10$  setting in Figure 3; we note that the ARI performs similarly to the  $G$ -test for independence in all cases.

#### C.3.2 Additional values of $\Sigma^{(1)}$ and $\Sigma^{(2)}$

We return to the simulation set-up described in Section 4 with  $K = 3$ ,  $\mu^{(l)}$  given by

$$\mu^{(1)} = \begin{bmatrix} 0 & 0 & \sqrt{12} \\ 2 & -2 & 0 \end{bmatrix}, \quad \mu^{(2)} = \begin{bmatrix} -2 & 0 & 2 \\ 0 & \sqrt{12} & 0 \end{bmatrix}, \quad (\text{C.4})$$

and for two choices of  $\Sigma^{(l)}$ :

$$\text{Choice 1: } \Sigma^{(1)} = \Sigma^{(2)} = \begin{pmatrix} 2.25 & 0.5 \\ 0.5 & 2.25 \end{pmatrix},$$

$$\text{Choice 2: } \Sigma^{(1)} = \begin{pmatrix} 2.25 & 0.5 \\ 0.5 & 2.25 \end{pmatrix} \text{ and } \Sigma^{(2)} = \text{diag}(2.25, 4).$$

To perform Step 1 of Algorithm 1, we use the package `mclust` (Scrucca *and others*, 2016) to fit Gaussian mixture models with a common dense covariance matrix (the “EEE” covariance structure in `mclust`) for  $\Sigma^{(l)}$  given by choice 1 above, and to fit Gaussian mixture models

with a common diagonal covariance matrix (the “EEI” covariance structure in `mclust`) for  $\Sigma^{(l)}$  given by choice 2 above. We compare the performance of the pseudo likelihood ratio test of  $H_0 : C = 1_{K^{(1)}} 1_{K^{(2)}}^T$  at nominal significance level  $\alpha = 0.05$  to the  $G$ -test for independence for testing the null hypothesis  $H_0 : C = 1_{K^{(1)}} 1_{K^{(2)}}^T$ . The results are in Figure S9(i) and Figure S9(ii). Results are similar to results from the  $K = 6$  and  $p = 10$  setting with  $\sigma = 4.8$  in Figure 3.

### C.3.3 Model misspecification

We consider a simulation set-up which compares the performance of the pseudo likelihood ratio test under model misspecification. We generate data from model (2.1) – (2.2), with  $\Pi = \frac{1-\delta}{K^2} 1_K 1_K^T + \frac{\delta}{K} I_K$  for  $K = 3$ . In the  $l$ th data view, the observations are drawn from a finite mixture model for which the  $k$ th component in the mixture (corresponding to the  $k$ th cluster) is a bivariate Student’s  $t$ -distribution with location parameter  $\mu_k^{(l)}$  and scale matrix  $\Sigma = \begin{pmatrix} 2.25 & 0.5 \\ 0.5 & 2.25 \end{pmatrix}$ , where the mean matrices for the two data views are of the form (C.4). We fit Gaussian mixture models with a common covariance matrix (the “EEE” covariance structure in `mclust`), and again use  $B = 200$  permutation samples. The results are in Figure S9(iii). We compare the performance of the pseudo likelihood ratio test of  $H_0 : C = 1_{K^{(1)}} 1_{K^{(2)}}^T$  at nominal significance level  $\alpha = 0.05$  to the  $G$ -test for independence for testing the null hypothesis  $H_0 : C = 1_{K^{(1)}} 1_{K^{(2)}}^T$ , with p-values obtained with a permutation approach. Results remain similar to results from the  $K = 6$  and  $p = 10$  setting with  $\sigma = 4.8$  in Figure 3.

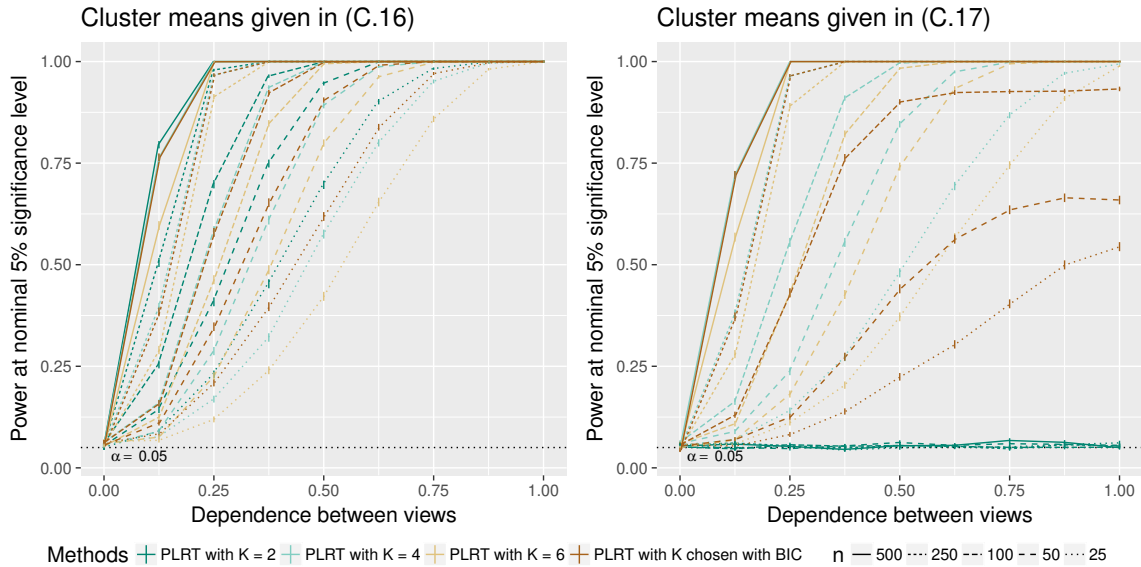


Figure S1: Power of the pseudo likelihood ratio test of  $H_0 : C = 1_{K_1} 1_{K_2}^T$  with  $p = 2$ ,  $K = 4$ , and  $\sigma = 0.4$  in the two simulation settings described in Appendix C.2.1. The  $x$ -axis displays  $\delta$ , defined in (C.1), and the  $y$ -axis displays the power.

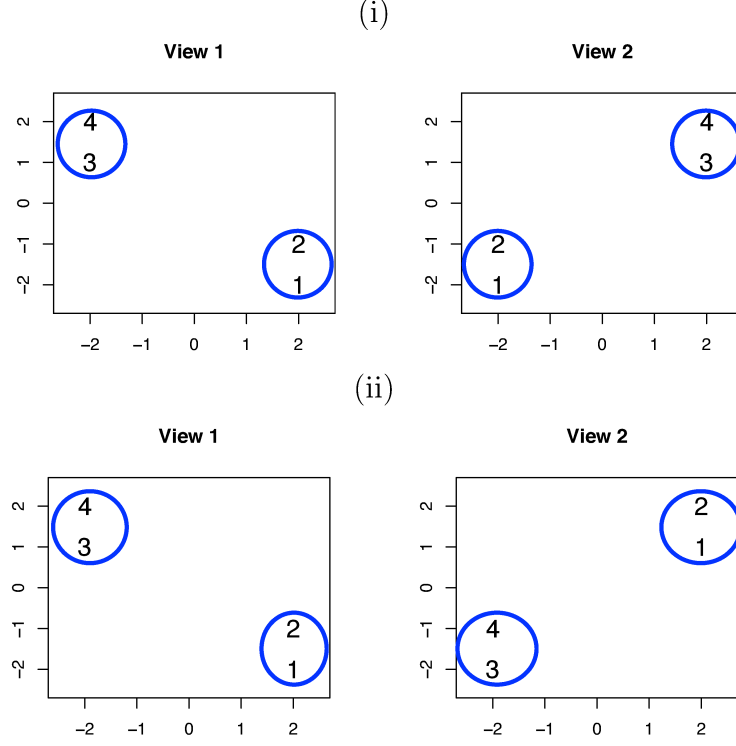


Figure S2: (i) For the simulation study described in Appendix C.2.1, the cluster means and “meta-clusters” under (C.2), where the  $k$ th cluster mean on each view is indicated by the number  $k$ , and the two “meta-clusters” on each view are circled in blue. (ii) For the simulation study described in Appendix C.2.1, the cluster means and “meta-clusters” under (C.3), where the  $k$ th cluster mean on each view is indicated by the number  $k$ , and the two “meta-clusters” on each view are circled in blue.

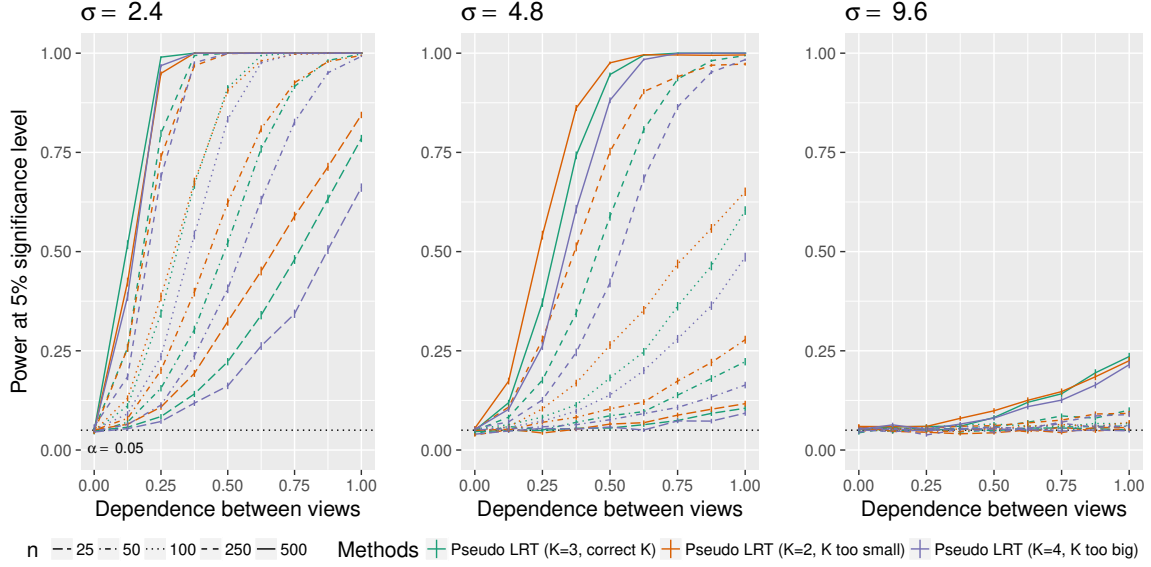


Figure S3: Power of the pseudo likelihood ratio test of  $H_0 : C = 1_{K_1} 1_{K_2}^T$  with  $p = 10$ ,  $K = 3$  and  $\sigma \in \{2.4, 4.8, 9.6\}$  in the simulation setting described in Appendix C.2.2. The  $x$ -axis displays  $\delta$ , defined in (C.1), and the  $y$ -axis displays the power.

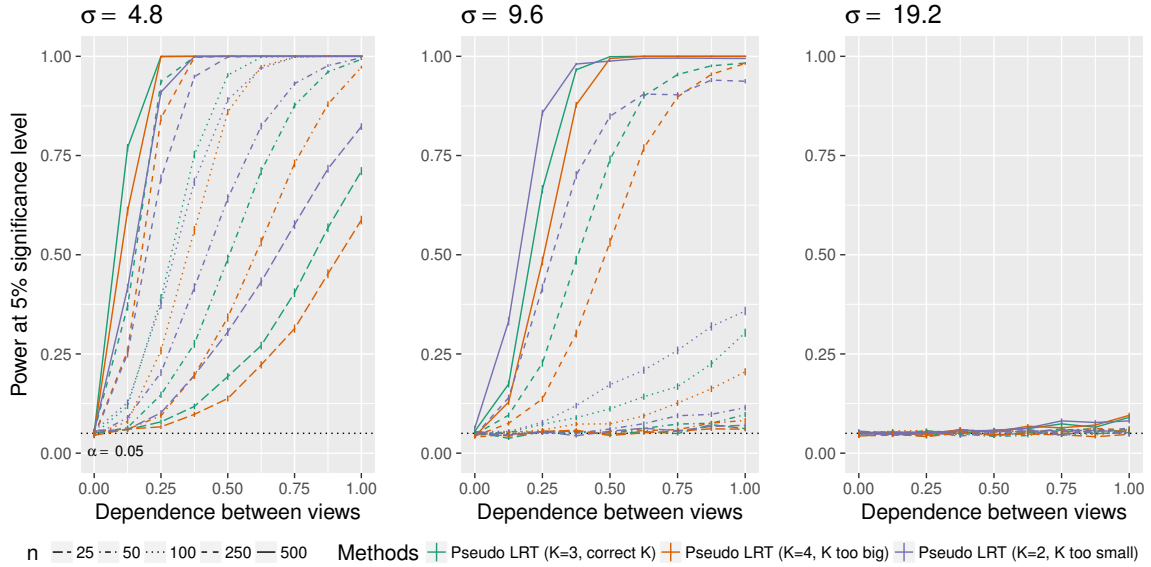


Figure S4: Power of the pseudo likelihood ratio test of  $H_0 : C = 1_{K_1} 1_{K_2}^T$  with  $p = 100$ ,  $K = 3$  and  $\sigma \in \{4.8, 9.6, 19.2\}$  in the simulation setting described in Appendix C.2.2. The  $x$ -axis displays  $\delta$ , defined in (C.1), and the  $y$ -axis displays the power.



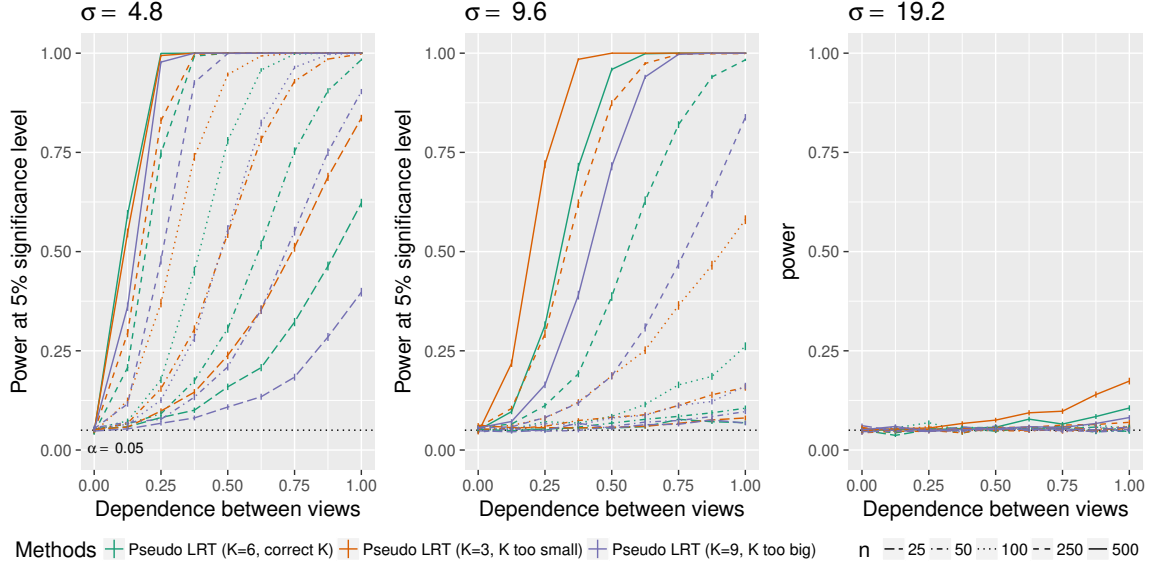


Figure S5: Power of the pseudo likelihood ratio test of  $H_0 : C = 1_{K_1}1_{K_2}^T$  with  $p = 100$ ,  $K = 6$  and  $\sigma \in \{4.8, 9.6, 19.2\}$  in the simulation setting described in Appendix C.2.2. The  $x$ -axis displays  $\delta$ , defined in (C.1), and the  $y$ -axis displays the power.

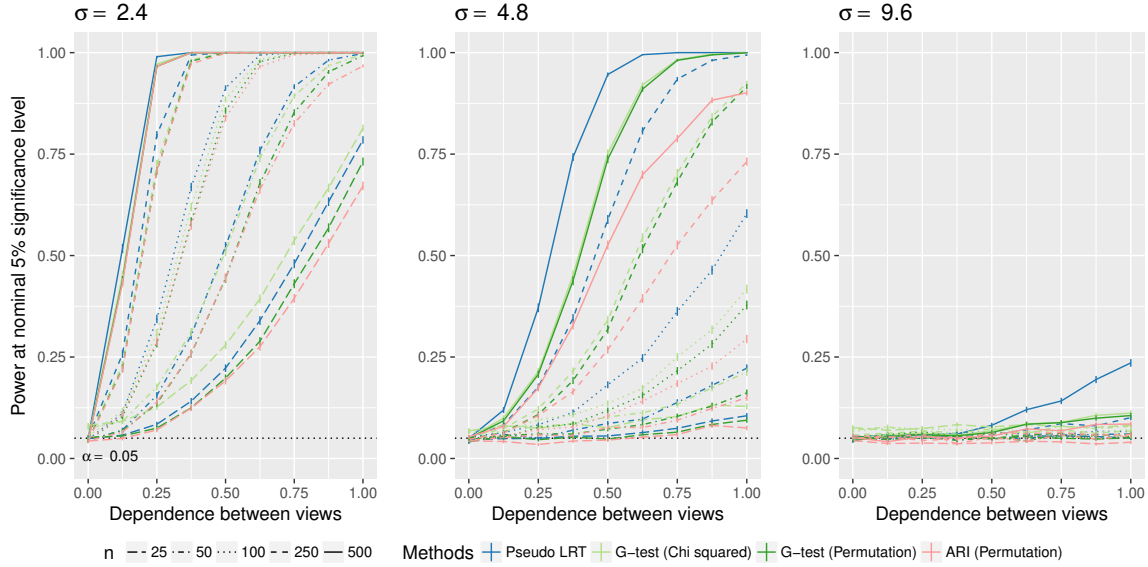


Figure S6: For the simulation study described in Appendix C.3.1, power of the pseudo likelihood ratio test, the  $G$ -test for independence, and the adjusted Rand Index (ARI) for  $p = 10$ ,  $K = 3$  and  $\sigma \in \{2.4, 4.8, 9.6\}$ , with  $\delta$ , defined in (C.1) on the  $x$ -axis and power on the  $y$ -axis.

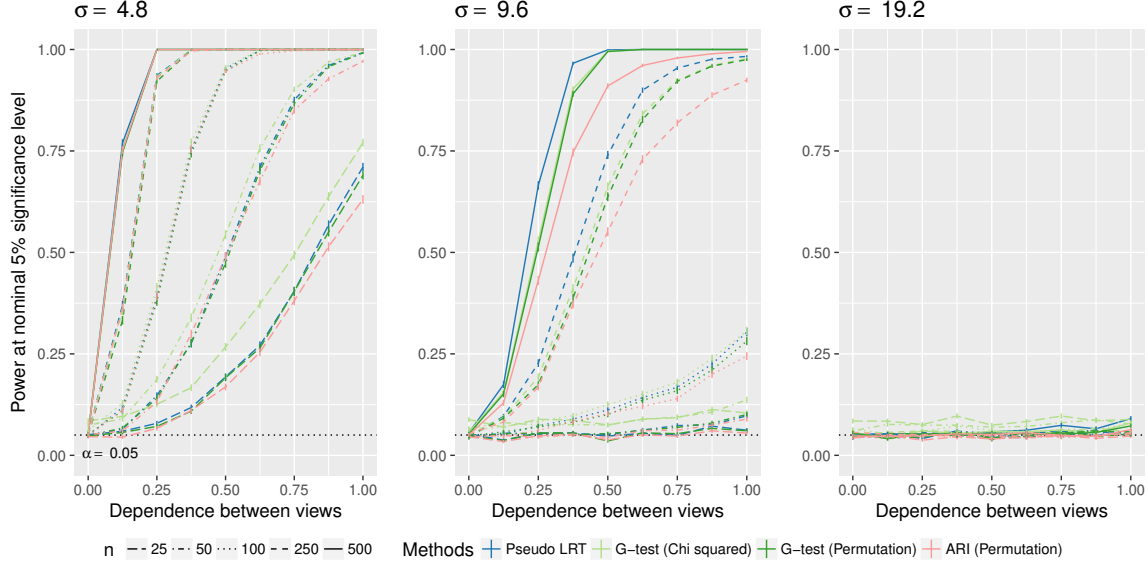


Figure S7: For the simulation study described in Appendix C.3.1, power of the pseudo likelihood ratio test, the  $G$ -test for independence, and the adjusted Rand Index (ARI) for  $p = 100$ ,  $K = 3$  and  $\sigma \in \{4.8, 9.6, 19.2\}$ , with  $\delta$ , defined in (C.1), on the  $x$ -axis, and power on the  $y$ -axis.

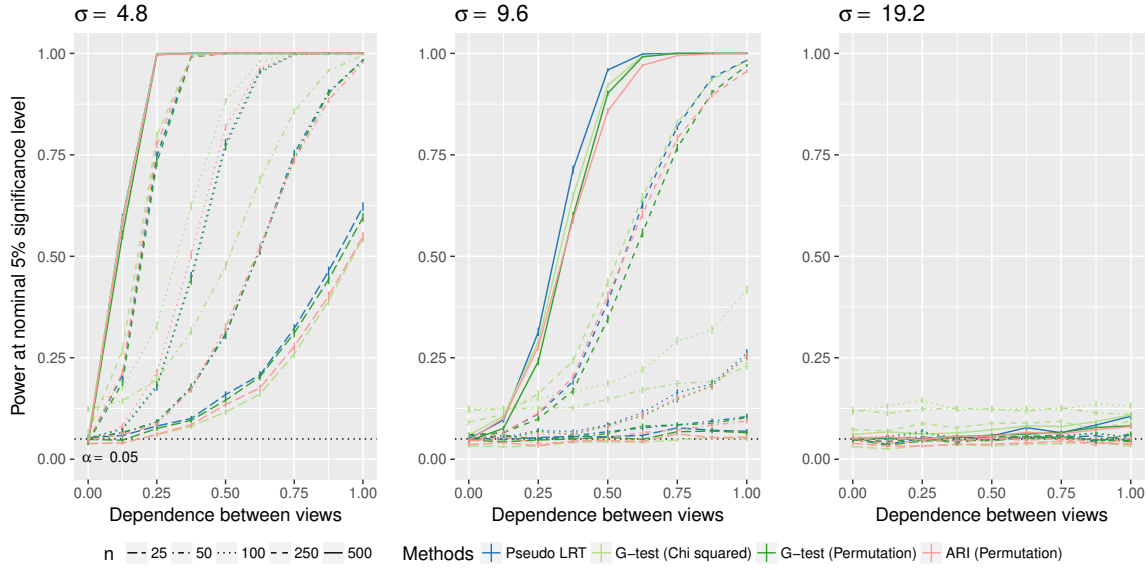


Figure S8: For the simulation study described in Appendix C.3.1, power of the pseudo likelihood ratio test, the  $G$ -test for independence, and the adjusted Rand Index (ARI) for  $p = 100$ ,  $K = 6$  and  $\sigma \in \{4.8, 9.6, 19.2\}$ , with  $\delta$ , defined in (C.1), on the  $x$ -axis, and power on the  $y$ -axis.

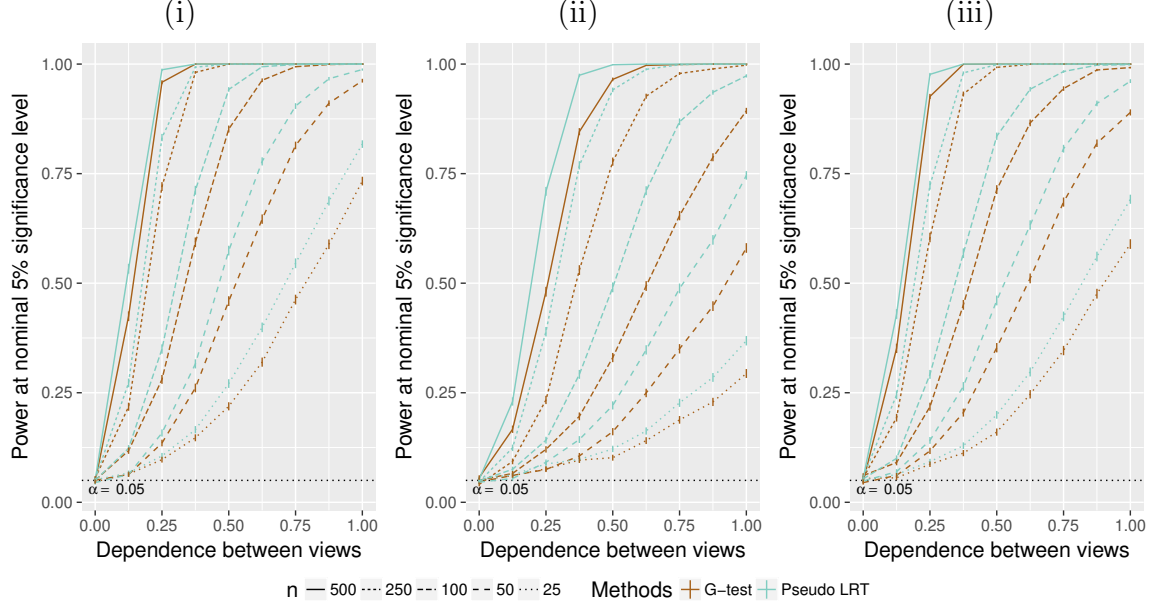


Figure S9: The x-axis displays  $\delta$ , defined in (C.1), and the y-axis displays power. Power of the pseudo likelihood ratio test and the  $G$ -test for independence for (i)  $\Sigma^{(1)} = \Sigma^{(2)} = \begin{pmatrix} 2.25 & 0.5 \\ 0.5 & 2.25 \end{pmatrix}$ , and Gaussian mixture components, (ii)  $\Sigma^{(1)} = \begin{pmatrix} 2.25 & 0.5 \\ 0.5 & 2.25 \end{pmatrix}$ ,  $\Sigma^{(2)} = \text{diag}(2.25, 4)$ , and Gaussian mixture components, (iii) bivariate Student's  $t$ -distributions as mixture components. Details for (i) and (ii) are in Appendix C.3.2, and details for (iii) are in Appendix C.3.3.