Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Econometrics & MS

Clustering the subspace of high-dimensional omics data

A comparison study of tandem and joint matrix factorisation and clustering techniques on simulated and empirical RNA-seq cancer data

Nikki van den Berg (472123)

Supervisor:	Jeffrey Durieux
Second assessor:	Eoghan O'Neill
Date final version:	31st August 2023

Abstract

Novel biomarkers and tumour subtypes can be discovered by the clustering of high-dimensional cancer omics data. However, clustering algorithms do not work efficiently because of the "curse of dimensionality". Traditionally, matrix factorisation methods such as Principal Component Analysis (PCA), Independent Component Analysis (ICA) or Non-negative Matrix Factorisation (NMF) are applied before clustering to reduce dimensions. This tandem approach is found to be suboptimal because the matrix factorisation and clustering step do not have the same optimisation criteria. Consequently, if clusters are formed in the higher dimensions that are not included in the dimensions constructed by the matrix factorisation step, the cluster structure will be lost. Reduced K-means (RKM) and Factorial K-means (FKM) integrate the matrix factorisation and clustering steps in one optimisation criterion. The joint approach aims to retain cluster structure in a lower dimensional subspace. With extensive simulation studies and an empirical study on a cancer RNA-seq dataset, we compare the performance of the tandem approach, being PCA, ICA and NMF with K-means to the joint approach, RKM and FKM. By testing the methods under different residual structures and in combination with feature selection and component specifications, we found that RKM and ICA with K-means outperformed the other methods regarding cluster membership identification. NMF had a superior subspace recovery performance. High-performing combinations include the feature selection method interquartile range (IQR) with RKM and coexpression with ICA and K-means. Lastly, we found that selecting components for a relatively low dimensionality (i.e. five) rather than a large dimensionality (i.e. twenty or thirty) yields higher clustering accuracies.

Keywords — *Cancer, Cluster analysis, Matrix factorisation, RNA-seq, Simulation study, Simultaneous clustering*

Contents

List of Tables

List of Figures

Table 1: Glossary of key terms.

Cancer omics data	
Omics data	High-dimensional data resulting from studies in genomics, transcriptomics, proteomics, metabolomics, etc. Can be analysed to reveal cellular activities and sample characteristics (Stein-O'Brien et al., 2018).
RNA-seq	High-throughput sequencing technique that measures the number of short reads from each gene and summarises this into gene counts (Stein-O'Brien et al., 2018).
Phenotype	Observable features of a sample that result from the corresponding genotype (Wojczynski & Tiwari, 2008).
Biomarkers	A feature, gene or molecule that can identify pathological processes (Stein-O'Brien et al., 2018).
Matrix Factorisation	
Matrix Factorisation (MF)	A method to approximate an observed data matrix using the product of a signal matrix and a loading matrix (Stein-O'Brien et al., 2018).
Signal matrix	By MF constructed matrix with components as rows and genes as columns. The contributions of the genes to the components can be inferred and analysed to define molecular signatures for a phenotype (Stein-O'Brien et al., 2018).
Loading matrix	By MF constructed matrix with samples as rows and components as columns. The activities of the samples in the components can be analysed to associate phenotypes with the samples (Stein-O'Brien et al., 2018).
Feature Selection	A method to exclude genes that do not contain information that can be used for tumour (sub)types partitioning (Källberg et al., 2021).
Principal Component Analysis (PCA)	A MF technique that constructs orthogonal components that can be ranked by their explained variance in the observed data (Stein-O'Brien et al., 2018).
Independent Component Analysis (ICA)	A MF technique that constructs statistically independent non-Gaussian components (Sompairac et al., 2019).
Non-negative Matrix Factorisation (NMF)	A MF technique that constructs components that contain elements that are equal or greater than zero (Stein-O'Brien et al., 2018).
Clustering	
Cluster subspace	The subspace of the variables where the centroids of the clusters reside (Timmerman et al., 2010).
K-means	Partitions samples into clusters and allocates samples to clusters with the nearest centroid (MacQueen, 1967).
Reduced K-means (RKM)	A joint cluster allocation and dimension reduction that maximises the between variance of clusters in the subspace (De Soete & Carroll, 1994).
Factorial K-means (FKM)	A joint cluster allocation and dimension reduction technique that minimises the within variance of clusters in the subspace (Vichi & Kiers, 2001).
Adjusted Rand Index (ARI)	Measure for clustering accuracy. Takes a value between 0 (no cluster recovery) and 1 (perfect cluster recovery) (Hubert & Arabie, 1985).
Subspace recovery (Phi)	Measure that represents the proportionality between the columns of the estimated and the simulated loading matrices (Kuhn & Tucker, 1951).

Chapter 1

Introduction

Cancer therapy is improved by the discovery of novel biomarkers, tumour subtypes, or signature expression patterns (Tsimberidou et al., 2020). Such findings are made by analysing cancer omics data. Omics data is a field of research which includes subfields such as genomics, transcriptomics and proteomics (Stein-O'Brien et al., 2018). The experimental techniques involved in omics research create high-dimensional datasets. One of the most common experimental techniques in transcriptomics is RNA sequencing (RNA-seq). RNA-seq measures the number of short reads from each gene and summarises this into gene counts (Stein-O'Brien et al., 2018). The distribution of RNA-seq data is non-Gaussian and is count-based (Yu et al., 2021). Furthermore, the number of variables, the genes, greatly outnumber the number of observations, the tumoral samples, yielding a high-dimensional dataset.

Commonly, high-dimensional data is transformed into a low-dimensional structure with Matrix Factorisation (MF) techniques to preserve as much information as possible (Stein-O'Brien et al., 2018). Common MF techniques used in the analysis of cancer omics data are Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Non-Negative Matrix Factorisation (NMF) (Jolliffe, 2002; Herault & Jutten, 1986; Paatero & Tapper, 1994). These methods differ in the constraints introduced to identify the signal and loading matrices. In short, PCA constructs orthogonal components, ICA constructs non-normally distributed components that are as mutually independent as possible, and NMF constructs components that are non-negative (Jolliffe, 2002; Sompairac et al., 2019; Gaujoux & Seoighe, 2010). The resulting components simplify the interpretation and the inference of the omics data. Clustering algorithms such as hierarchical clustering or K-means are applied to the components to get further insight, for example, to identify or classify tumour subtypes (Stein-O'Brien et al., 2018; Sompairac et al., 2019). Accurate clustering could identify novel tumour subtypes that may have to be treated differently, which discovery can improve patient survival (Källberg et al., 2021).

However, performing cluster analysis after applying a MF technique, known as the tandem approach, is often found to be suboptimal. MF and clustering algorithms have different optimisation criteria, which can lead to loss of cluster structure (Iodice D'Enza et al., 2014). For example, PCA aims to find a linear combination of variables that maximise the explained variance (Jolliffe, 2002), whereas K-means aims to allocate observations to clusters based on similarity (MacQueen, 1967). Hence, if clusters are formed in higher dimensions that are not included in the latent dimensions found by PCA, the cluster structure will be lost (Timmerman

et al., 2010). Accurate subspace recovery is therefore important to preserve cluster structure.

As a solution to the shortcomings of tandem approaches, De Soete & Carroll (1994) proposed a joint cluster allocation and dimension reduction technique called Reduced K-means (RKM) (Arabie & Hubert, 1996). The loss function of RKM is specified to maximise the *between* variance of clusters in the subspace (De Soete & Carroll, 1994). Vichi & Kiers (2001) also proposed a joint technique called Factorial K-means (FKM), which minimises the *within* variance of clusters in the subspace. Alternatively, one could use a compromise model between FKM, RKM and PCA. Yamamoto & Hwang (2014) introduced a decomposition of the objective function of RKM, which Markos, D'enza & van de Velden (2019) used to develop the compromise model. In this model, the loss functions of the joint methods FKM, RKM and tandem method PCA with K-means are combined. Researchers can use this to tune the model and suit it to the data at hand.

There is a knowledge gap in the application of joint methods on bulk RNA-seq cancer omics data. Similar joint methods have only been evaluated on single-cell RNA-seq data (Wu & Ma, 2020; W. Liu et al., 2022). Single-cell RNA-seq data is more sparse than bulk RNA-seq data, which makes the clustering performances difficult to translate to bulk RNA-seq data (Jiang et al., 2022). Hence, we aim to contribute to the research by benchmarking the tandem and joint methods applied to (simulated) cancer omics data.

We have identified another gap in the research: it is not clear how the clustering performances depend on residual structures. Many studies are empirical, i.e. applied to real cancer data. Because we do not know the residual structure of empirical data, we cannot infer how clustering algorithms interact with different types of noise. Yet, this is important as findings by De Soete & Carroll (1994); Vichi & Kiers (2001); Timmerman et al. (2010); Yamamoto & Hwang (2014) indicate that the suitability of clustering models depends on the type of residuals. The authors show that residuals can lie in the subspace of the clusters, or they can exist in the complement space. Masking variables, a source of complement noise, do not reflect the cluster structure but are correlated to each other. When there are masking variables, FKM and RKM could fail to identify the cluster structure and optimal subspace (Yamamoto & Hwang, 2014). These findings show that an extensive simulation study on different structures of noise is necessary to properly evaluate the clustering algorithms.

We investigate the following main research question: *"Do joint MF and clustering algorithms outperform benchmark tandem techniques in preserving cluster structure in (simulated) cancer omics data?"*. We use subspace recovery and cluster membership identification as quality criteria. We test the methods in a simulation study and an empirical analysis to find relationships between the clustering algorithms, data characteristics and residual structures.

In the simulation study, we will analyse the effect of residual structures on the performance of joint and tandem methods when they are applied to simulated cancer omics data. Specifically, we aim to answer the research subquestions: *"How does the performance of joint and tandem methods depend on the level of random noise, subspace noise and masking variables?"* and *"Which joint, compromise, or tandem approach is the most suitable in the presence of random noise,*

subspace noise and masking variables?"

We hypothesise that overall, the joint MF and clustering algorithms will outperform the benchmark tandem approach because the optimisation criteria in the joint algorithms are designed in such a way that cluster structure is optimally preserved in the subspace. Based on the study of Timmerman et al. (2010), we expect that RKM is suitable in the presence of subspace noise and that FKM is suitable for data with a large fraction of complement noise. However, when masking variables are the source of the complement noise, we think that both RKM and FKM will fail (Yamamoto & Hwang, 2014). The independence criterion in ICA might separate the masking signal from the cluster structure, resulting in higher clustering accuracies in combination with K-means (Sompairac et al., 2019).

In the empirical analysis, we will compare the clustering accuracy of joint and tandem techniques considering multiple feature selection and latent dimension options. For this experiment, we use a pan-cancer RNA-seq dataset from which we know the ground truth cluster labels. We aim to answer the research subquestions: *"Do joint methods outperform tandem techniques in clustering empirical cancer omics data, and how does their performance depend on feature selection and latent dimension options?"* and *"Can we interpret the signal and loading matrix of the ICA components that are computed from the pan-cancer dataset?"*.

We hypothesise that RKM performs well in the empirical setting. Feature selection methods will likely select informative genes and sort out the non-informative genes, i.e. variables that can be considered as noise. Hence, there is little chance that the gene expression data will consist of a large proportion of complement residuals, which facilitates a suitable environment for the RKM objective function. We also expect ICA with K-means to perform well. Cancer omics data is not normally distributed, hence an algorithm such as ICA that captures non-Gaussian signals is suitable (Sompairac et al., 2019). This will also make the functional annotation of the ICA components possible. Furthermore, NMF estimates non-negative signal and loading matrices, corresponding to the non-negative nature of omics data (Stein-O'Brien et al., 2018). This could mean that NMF will also yield high clustering accuracies. We expect that choices in feature selection are universal to all methods. However, we think that the choice of the number of latent dimensions is method-specific as the constraints in the algorithms could require different numbers of dimensions in the most optimal subspace.

This thesis uses the following structure: in Section 2, we discuss findings of studies on analysing cancer omics data and findings of studies that applied MF and clustering algorithms; in Section 3, we describe the MF and clustering algorithms; in Section 4, we describe the approach and results of the simulation study; in Section 5, we describe the approach and results of the empirical analysis. In Section 6 we discuss the implications of the findings, report the limitations of the study and put the findings into perspective.

Chapter 2

Literature

In this section, we first review how MF techniques are used to analyse cancer omics data. After this, we discuss studies that cluster cancer omics data after applying MF techniques, also known as the tandem approach. Next, we cover the findings of studies on joint dimension reduction techniques and how this research contributes to the field.

2.1 Analysing cancer omics data with MF techniques

Developments in precision medicine, which involves the analysis of the genetic and clinical profile of an individual, improve cancer diagnosis and treatment (Kamat & Matulay, 2018). The genetic profile of an individual is characterised by gene expression patterns, which are collected using high-throughput techniques such as RNA-seq. RNA-seq experiments result in high-dimensional datasets, categorised as omics data. When cancer omics data is collected using RNA-seq, it consists of gene-level counts of tumour samples. These gene-level counts are dependent on the state of the biological system and thus carry tumour-specific information (Stein-O'Brien et al., 2018).

One can interpret omics data by clustering. However, due to the "curse of dimensionality", clustering algorithms do not work efficiently (Hinneburg & Keim, 1999). In higher dimensions, data becomes more sparse and distances become harder to distinguish (Tomašev et al., 2011). Therefore, before analysis, MF techniques are applied. MF transforms the data into a lower dimensional structure while retaining as much information as possible. MF decomposes the omics data matrix into the signal matrix and the loading matrix, containing information about the molecular and sample relationships, respectively (Stein-O'Brien et al., 2018; Engreitz et al., 2010; Biton et al., 2014). The columns in the signal matrix contain the contributions of genes to a phenotype, which is suitable for biomarker discovery analysis (Stein-O'Brien et al., 2018). Tumour subtypes can be identified by performing clustering analysis on the rows of the loading matrix, which contain the contributions of samples to factors (Sompairac et al., 2019; Stein-O'Brien et al., 2018). When the clustering analysis of the loading matrix is accurate, it might discover novel tumour subtypes that could be treated differently, improving cancer therapy (Källberg et al., 2021).

Commonly used MF techniques include PCA, ICA and NMF (Sompairac et al., 2019). These methods differ in their underlying statistical assumptions and capture genetic signals differently.

The components of PCA maximise the captured variance in the original data and are ranked by how much they explain the variance in the original data (Jolliffe, 2002). Stein-O'Brien et al. (2018) found that PCA captures dominant signals but can mix multiple biological processes in one component, which makes it difficult to interpret. ICA captures statistically independent sources of variation in the data, which are associated with gene sets and are thus easier to interpret (Sompairac et al., 2019). Engreitz et al. (2010) and Biton et al. (2014) used ICA to functionally annotate the components computed from high-dimensional RNA-seq cancer data and were able to annotate associated sets of coexpressed genes. The components of NMF are non-negative, similar to transcriptional data. Stein-O'Brien et al. (2018) found that the components of NMF contain information on overexpressed genes in a single phenotype because NMF additively adds signals to components.

2.2 Clustering cancer omics data using tandem techniques

The tandem approach includes clustering the omics data after applying MF. Clustering algorithms such as K-means or hierarchical clustering cluster the latent dimensions of the omics data, with the goal of discovering new tumour (sub)types (Stein-O'Brien et al., 2018). This is only possible if MF techniques preserve cluster structure accurately (Timmerman et al., 2010).

Vidman et al. (2019) analysed how sample size, heterogeneity, and the distribution of high-dimensional RNA-seq data affect the performances of K-means and hierarchical clustering after reducing the dimensions with PCA. The authors found that sample size did not affect the performance, but that cluster distribution was important. The authors also found that clustering homogeneous data is preferred, thus analysing female and male data separately.

Fonseca et al. (2017) performed a K-means clustering analysis on temporal RNA-seq data after ICA MF. The authors could identify clusters with distinct expression patterns, which is in line with the multi-modal character of the mixing matrix constructed with ICA (Sompairac et al., 2019).

Feature selection is often applied to RNA-seq data before analysing the data with MF and clustering techniques (Källberg et al., 2021; Freyhult et al., 2010). The difference between feature selection and MF is that feature selection methods exclude non-informative genes, while MF techniques explain a higher-dimensional structure in a lower-dimensional subspace (Källberg et al., 2021; Stein-O'Brien et al., 2018). Feature selection can be done by analysing for example the level, variance, similarity, or modality (Källberg et al., 2021). Each method captures different information about the genetic profile and tumoural activity, hence selecting different sets of genes.

Selecting genes based on the level of coexpression, that is, the simultaneous expression of two or more genes, is useful when the data is very noisy. When one selects genes with a higher-than-average expression level, it is easier to identify differentially expressed genes (Källberg et al., 2021).

Selecting genes with high variability is the most common method and is motivated by the discovery of intratumorally differentiated genes (Källberg et al., 2021; Freyhult et al., 2010). When genes have highly variable gene expression patterns, they could contain tumour-specific

information (Källberg et al., 2021). However, there is a chance of including highly variable genes that are not intratumorally differentially expressed (Freyhult et al., 2010). Freyhult et al. (2010) compared selecting genes based on variance and level, and found that selection based on variance performed better.

Selecting genes based on similarity is based on the assumption that when genes have similar expression patterns, they assemble into gene modules (Z. Wang et al., 2014). Z. Wang et al. (2014) found that selecting genes based on coexpression can identify similar genes, resulting in accurate cluster partitions.

Selecting genes based on modality comes from the idea that gene expression distributions have multiple modes if genes are differentially expressed. Hence, testing for multimodality can identify informative genes. Källberg et al. (2021) compared selecting genes based on variance, level, similarity, and modality and found that the best clustering performances were in combination with selection methods based on modality.

Feng et al. (2020) reviewed feature selection in combination with MF and clustering algorithms applied to single-cell RNA-seq data. The authors found that selecting high-variable genes before applying MF and clustering techniques improved performance. Furthermore, Feng et al. (2020) found that applying MF algorithms improves the performance of clustering algorithms. ICA performed well in compressed feature spaces, and PCA was more stable than ICA and NMF. When performing clustering analysis after MF, K-means performed better than other clustering algorithms such as hierarchical clustering.

2.3 Clustering data using joint techniques

The joint approach integrates the optimisation criteria of dimension reduction and clustering algorithms. It aims to preserve cluster structure while effectively reducing the number of variables (Timmerman et al., 2010). This study will use RKM and FKM as joint MF and clustering techniques because these methods are based on commonly used clustering algorithm K-means and have not yet been extensively studied on (simulated) cancer omics data.

Timmerman et al. (2010) use a simulation study that generates correlated normally distributed data to compare RKM and FKM. The authors find that RKM and FKM complement each other, such that when RKM fails, FKM performs well and vice versa. The authors find that the choice between RKM and FKM depends on the proportion of residuals that exist in the subspace compared to residuals that lie in the complement of that subspace. Specifically, Timmerman et al. (2010) find that when the size of subspace residual variance compared to the complement residual variance increases, the subspace recovery of FKM decreases. Contrarily, RKM increases when the size of subspace residual variance gets larger. As FKM and RKM complement each other, Timmerman et al. (2010) recommend considering both models when the residual structure is not clear.

Yamamoto & Hwang (2014) propose an extension to FKM and RKM, called Generalised Reduced Clustering (GRC). With the use of simulation studies, Yamamoto & Hwang (2014) find that GRC is suitable when the data consists of masking variables. These variables are irrelevant to the cluster structure but are correlated between themselves. The authors find that FKM and

RKM do not perform well in these circumstances. However, this study is particularly useful as it introduced a decomposition of the objective function of RKM, leading to a compromise model between PCA and FKM. Vichi et al. (2019) introduced a convex combination that includes parameter *alpha* and Markos, D'enza & van de Velden (2019) designed function `cluspca` in R package **clustrd**. This created an easy-to-use compromise model of FKM, RKM and PCA (Markos, Iodice D'Enza & van de Velden, 2019).

Chapter 3

Models

3.1 Tandem techniques

We construct the tandem approach using PCA, ICA, and NMF as MF techniques, coupled with K-means as a clustering algorithm. We choose to use PCA because it is a benchmark MF approach that efficiently captures variance in the components, and has a fast computation time. We choose to use ICA because it is commonly used in gene expression analyses as it captures statistically independent signals, which are often associated with gene sets that can be traced back to biological processes (Sompairac et al., 2019). Furthermore, ICA is considered as a suitable method for gene expression analysis as RNA-seq data is not normally distributed, which is a requirement for ICA (Tharwat, 2018). NMF is considered as a suitable method for the analysis of RNA-seq data because the components of NMF are non-negative, similar to transcriptional data (Stein-O'Brien et al., 2018).

MF techniques take data matrix $X \in \mathbb{R}_m^N$ as their input matrix with N as the observed samples and m as the observed features. Matrix X is approximated as a sum of products of p pairs of vectors with size N and m. The fundamental equation in MF is:

$$\mathbf{X} \approx \mathbf{AS} = \sum_{k=1}^{p} \mathbf{a_k} \otimes \mathbf{s_k}, \tag{3.1}$$

where $\mathbf{a_k}$ are columns of $\mathbf{A}_{N\times p}$, $\mathbf{s_k}$ are rows of $\mathbf{S}_{p\times m}$ and sets of $\mathbf{a_k}$ and $\mathbf{s_k}$ are called components. The vectors $\mathbf{a_k}$ and $\mathbf{s_k}$ are often defined as the loading and signal vectors, respectively. The objective in MF algorithms is to find the set of components that solve

$$\min(\mathbf{a_k}, \mathbf{s_k}) = ||\mathbf{X} - \sum_{k=1}^{p} \mathbf{a_k} \otimes \mathbf{s_k}||^2 \tag{3.2}$$

where $||\dots||$ defines the sum of the Euclidean norms of the columns in a matrix.

This problem is underdetermined and constraints need to be introduced because only $\mathbf{X}$ is known. Each MF technique uses specific constraints, which lead to different resulting sets of $\mathbf{a_k}$ and $\mathbf{s_k}$ (Sompairac et al., 2019).

3.1.1 Principal Component Analysis (PCA)

PCA constructs orthogonal components by introducing the constraint that the vectors $\mathbf{a_k}$ are orthogonal such that $(\mathbf{a_i}, \mathbf{a_j}) = 0$ for $i \neq j$. Furthermore, Equation 3.2 must give the same components as the solution for different orders of matrix decomposition p. PCA decomposes the data matrix $\mathbf{X}_{m\times N}$ into a score matrix $\mathbf{S}_{m\times p}$, loading matrix $\mathbf{A}_{N\times p}$ and residual matrix $\mathbf{E}_{m\times N}$[1]:

$$\mathbf{X} = \mathbf{S}\mathbf{A}^T + \mathbf{E} \tag{3.3}$$

The minimisation problem is convex and results in a unique global minimum. Hence, the resulting orthogonal components can naturally be ranked by their explained variance of the original data (Jolliffe, 2002; Sompairac et al., 2019). Moreover, data should be scaled before applying PCA. In this research, we perform PCA using the R package **stats** (R-CoreTeam, 2023).

3.1.2 Independent Component Analysis (ICA)

The goal of ICA is to separate mixed signals based on independence (Tharwat, 2018; Sompairac et al., 2019). The constraint in Equation 3.1 is therefore that all sets of $\mathbf{a_k}$ and $\mathbf{s_k}$ must be as mutually independent as possible (Sompairac et al., 2019). ICA decomposes the observed data matrix $\mathbf{X}_{N\times m}$ into source signal matrix $\mathbf{S}_{p\times m}$ and mixing matrix $\mathbf{A}_{N\times p}$, i.e. the loading matrix, with p independent components such that $\mathbf{X} = \mathbf{AS}$. This equation is solved by finding unmixing matrix $\mathbf{W}_{N\times p}$, which transforms the observed data matrix $\mathbf{X}_{N\times m}$ into a set of independent signals $\mathbf{Y}_{p\times m}$, that is, $\mathbf{Y} = \mathbf{W}^T\mathbf{X}$. A fundamental restriction in ICA is that source signals must be non-Gaussian and independent. ICA methods extract source signals by searching for non-Gaussian signals in the observed data matrix (Tharwat, 2018). ICA components are uniquely defined if the components are independent and if there is at most one Gaussian component (Mesters & Zwiernik, 2022). There are multiple approaches to finding unmixing matrix $\mathbf{W}$. The most common are based on non-Gaussianity, minimising the mutual information or estimation by using maximum likelihood (Tharwat, 2018). The methods are all based on independence but result in slightly different unmixing matrices.

In this research, we use the FastICA algorithm from R package **fastICA** which maximises non-Gaussianity using a fixed-point iteration scheme (Marchini et al., 2021; Hyvärinen & Oja, 2000). We use the FastICA algorithm because it has a cubic convergence speed and does not have parameters that need to be tuned. The algorithm first preprocesses the data with centring and whitening, i.e. projecting the data onto its principal components such that the components are uncorrelated and have unit variance (Hyvärinen & Oja, 2000). The algorithm then finds an un-mixing matrix $\mathbf{W}$ that maximises the non-Gaussianity. For this solution, matrix $\mathbf{W}$ is constrained to be orthonormal such that the estimated components are uncorrelated (Marchini et al., 2021).

FastICA approximates the negative entropy (J), termed neg-entropy, of $\mathbf{W}^T\mathbf{X}$ to measure the

[1]The decomposition of PCA is often in literature depicted as the transposed version of the decomposition used in ICA and NMF.

non-Gaussianity. Negative entropy is defined as $J(y) = H(y_{Gaussian}) - H(y)$, which contains Gaussian random variable $H(y_{Gaussian})$ that has the same covariance matrix as y (Tharwat, 2018). The algorithm calculates the entropy of a random variable Z with N possible outcomes as:

$$H(Z) = -E[log(p_z(z))] = -\frac{1}{N}\sum_{t}^{N} log(p_z(z^t)), \tag{3.4}$$

where it uses $p_z(z^t)$ as the probability of event z^t, with $t = 1, 2, \ldots, N$ (Tharwat, 2018). When all variables are Gaussian, the negative entropy J is zero. The FastICA paper uses an approximation of calculating the neg-entropy, based on the maximum entropy principle:

$$J(y) \approx \sum_{i=1}^{p} k_i (E[G_i(y)] - E[G_i(v)])^2, \tag{3.5}$$

where k_i are positive constants and random variable $v \sim \mathcal{N}(0, 1)$. $E[G_i(y)]$ is the entropy of variable G and has different choices such as $G(y) = \frac{1}{\alpha}\log\cosh(\alpha y)$ where $1 \leq \alpha \leq 2$ or $G(y) = -\exp(y^2/2)$ (Tharwat, 2018; Marchini et al., 2021). In this study, we use the default setting, that is, the *logcosh* function with $\alpha = 1$.

FastICA computes independent components by finding independent sources of variation. The problem is not convex, hence solutions to the optimisation problem depend on the initialisation of the components and are not naturally ranked, contrary to PCA (Herault & Jutten, 1986; Sompairac et al., 2019). Furthermore, the sign of the independent components can be changed and does not influence the ICA model (Tharwat, 2018).

3.1.3 Non-negative Matrix Factorisation (NMF)

NMF decomposes $\mathbf{X}_{N\times m}$ into matrices $\mathbf{A}_{N\times p}$ and $\mathbf{S}_{p\times m}$ such that $\mathbf{X} \approx \mathbf{AS}$ with the constraint that all elements in matrices $\mathbf{X}$, $\mathbf{A}$ and $\mathbf{S}$ must be non-negative. The p non-negative components do not have to be orthogonal or independent like PCA or ICA, and they may overlap (Gaujoux & Seoighe, 2010). NMF estimates $\mathbf{A}$ and $\mathbf{S}$ by finding the (local) minimum of the following problem:

$$\min_{\mathbf{A},\mathbf{S}\geq 0}[D(\mathbf{X}, \mathbf{AS}) + R(\mathbf{A}, \mathbf{S})], \tag{3.6}$$

where D is a loss function that evaluates the approximation. R is optional, defining a regularisation function that constraints desirable properties such as sparsity or smoothness to matrices $\mathbf{A}$ and $\mathbf{S}$ (Gaujoux & Seoighe, 2010). The optimisation problem is not convex, similar to ICA. Hence, the solution is dependent on the initialisation of the components, and the components cannot be naturally ranked (Paatero & Tapper, 1994; Sompairac et al., 2019).

In this research, we use the R package **NMFN** to analyse the data (S. Liu, 2022). Multiple algorithms are available in this package, from which we use the multiplicative updating approach which is the default.

3.1.4 K-means

K-means partitions N samples into C clusters and allocates samples to clusters with the nearest centroid. Distances are calculated based on a chosen dissimilarity measure for which we use Euclidian distance. The number of clusters C needs to be determined beforehand. The algorithm minimises the within-cluster sum of squares:

$$\arg\min_{B} \sum_{i}^{C} \sum_{\mathbf{x} \in B_i} ||\mathbf{x} - \mu_{\mathbf{i}}||^2, \tag{3.7}$$

where the k clusters are defined as $\mathbf{B} = \{B_1, \ldots, B_C\}$, the d-dimensional observations as $\{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$ and the associated centroids as $\{\mu_\mathbf{1}, \ldots, \mu_\mathbf{C}\}$ (MacQueen, 1967).

We use the function `kmeans` in R package **stats** to analyse the samples (R-CoreTeam, 2023). We use the default algorithm, which is from Hartigan & Wong (1979). We use 100 random starts.

3.2 Joint techniques

Joint techniques integrate the optimisation criteria of dimension reduction and clustering algorithms. They aim to preserve cluster structure while effectively reducing the number of variables (Timmerman et al., 2010). In this research, we study RKM and FKM. We use the following notation: we denote $\mathbf{X}_{N \times m}$ as the data matrix and $\mathbf{A}_{m \times p}$ is the orthogonal loading matrix such that $\mathbf{A}^T\mathbf{A} = \mathbf{I}_p$. Let $\mathbf{U}_C$ be the $N \times C$ binary cluster membership matrix and $\mathbf{F}_{C \times p}$ be the cluster centroid matrix, where the rows denote the positions of the clusters in the reduced p-dimensional space. As we recall from Section 3.1, there are N observations, m features, p components and C clusters.

3.2.1 Reduced K-means (RKM)

RKM is a joint cluster allocation and dimension reduction technique that maximises the *between* variance of clusters in the subspace (De Soete & Carroll, 1994). RKM is suitable for data where there are no subspace and complement residuals present (Timmerman et al., 2010). The objective function of RKM is:

$$\min \phi_{RKM}(\mathbf{A}, \mathbf{U}_C, \mathbf{F}) = ||\mathbf{X} - \mathbf{U}_C\mathbf{F}\mathbf{A}'||^2, \tag{3.8}$$

where $||\ldots||$ is the Frobenius norm (Markos, D'enza & van de Velden, 2019). From this loss function, we see that RKM minimises the sum of squared distances between the observed data and the centroids in the subspace projected by loading matrix $\mathbf{A}$ (Timmerman et al., 2010).

Yamamoto & Hwang (2014) suggested to insert the solution of the cluster means, being $\mathbf{F} = (\mathbf{U}_C'\mathbf{U}_C)^{-1}\mathbf{U}_C'\mathbf{X}\mathbf{A}$. $\mathbf{F}$ is derived from the optimal solution of the RKM model $\mathbf{X} = \mathbf{U}\mathbf{F}\mathbf{A}' + \mathbf{E}_R$, namely:

$$\mathbf{X} = \mathbf{U}\mathbf{F}\mathbf{A}' + \underbrace{\mathbf{E}_R}_{0}, \tag{3.9}$$

which we rewrite to:

$$\mathbf{XA} = \mathbf{UF}\underbrace{\mathbf{A}'\mathbf{A}}_{\mathbf{I}_p}, \tag{3.10}$$

which we can rewrite to:

$$\mathbf{U}'\mathbf{XA} = \mathbf{U}'\mathbf{UF} \tag{3.11}$$

that leads to:

$$\mathbf{F} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{XA} \tag{3.12}$$

Hence, we can insert $\mathbf{F}$ in the objective function of RKM (Equation 3.8), resulting in the following notation:

$$\min \phi_{RKM}(\mathbf{A}, \mathbf{U}_C) = ||\mathbf{X} - \mathbf{PXAA}'||^2, \tag{3.13}$$

where $\mathbf{P} = \mathbf{U}_C(\mathbf{U}_C'\mathbf{U}_C)^{-1}\mathbf{U}_C'$. Using $\mathbf{P}$ and the trace operator, we have

$$||\mathbf{X} - \mathbf{PXAA}'||^2 = \text{Tr}(\mathbf{X}'\mathbf{X}) - \text{Tr}(\mathbf{A}'\mathbf{X}'\mathbf{PXA}). \tag{3.14}$$

From this expression, we can see that the between cluster variance in the reduced space, denoted in the second term on the right-hand side, is maximised when ϕ_{RKM} is minimised.

3.2.2 Factorial K-means (FKM)

FKM is a joint dimension reduction and clustering technique that minimises the *within* variance of clusters in the subspace (Vichi & Kiers, 2001). The objective function of FKM is:

$$\min \phi_{FKM}(\mathbf{A}, \mathbf{U}_C, \mathbf{F}) = ||\mathbf{XA} - \mathbf{U}_C\mathbf{F}||^2. \tag{3.15}$$

Hence, FKM minimises the sum of squared distances between the projected observed datapoints and the cluster centroids in the projected space (Timmerman et al., 2010). We can insert the solution of $\mathbf{F}$ to rewrite Equation 3.15 to

$$\min \phi_{FKM}(\mathbf{A}, \mathbf{U}_C) = ||\mathbf{XA} - \mathbf{PXA}||^2. \tag{3.16}$$

3.2.3 The `cluspca` algorithm

Yamamoto & Hwang (2014) propose the following decomposition of the RKM objective function in Equation 3.13:

$$||\mathbf{X} - \mathbf{PXAA}'||^2 = ||\mathbf{X} - \mathbf{XAA}'||^2 + ||\mathbf{XA} - \mathbf{PXA}||^2, \tag{3.17}$$

which shows that the objective function of RKM can be decomposed into a compromise of PCA and FKM (Markos, D'enza & van de Velden, 2019). Vichi et al. (2019) introduce a convex combination, which leads to the following objective function:

$$\min \phi_{ClusPCA}(\mathbf{A}, \mathbf{U}_C) = \alpha||\mathbf{X} - \mathbf{XAA}'||^2 + (1-\alpha)||\mathbf{XA} - \mathbf{PXA}||^2. \tag{3.18}$$

Minimising $\phi_{ClusPCA}$ is equal to maximising

$$\text{Tr}(\mathbf{A}'\mathbf{X}'((1-\alpha)\mathbf{P} - (1-2\alpha)\mathbf{I})\mathbf{XA}). \tag{3.19}$$

We compute the solutions of the RKM and FKM minimisation problems using the function `cluspca` from R package **clustrd** (Markos, Iodice D'Enza & van de Velden, 2019). The function consists of the following alternating least-squares algorithm (Markos, D'enza & van de Velden, 2019):

1. Initialise cluster membership matrix $\mathbf{U}_C$.
2. Find loading matrix $\mathbf{A}$ by taking the eigendecomposition of $\mathbf{X}'((1-\alpha)\mathbf{P} - (1-2\alpha)\mathbf{I})\mathbf{X}$.
3. Define new cluster membership matrix $\mathbf{U}_C$ by performing K-means to the subspace sample coordinates $\mathbf{XA}$.
4. Repeat steps 2-4 until convergence, i.e. until $\mathbf{U}_C$ remains constant.

The choice of α determines which method will be computed. When $\alpha = 0.5$, the algorithm computes the RKM solution, and for $\alpha = 0$ the FKM solution. If $\alpha = 1$, the problem reduces to the tandem approach, which is PCA followed by K-means.

Hence, we use $\alpha = 0.5$ for the RKM solution, and $\alpha = 0$ for the FKM solution. We use 100 random starts, similar to the K-means algorithm.

3.2.4 Subspace and complement residuals

To illustrate the differences between RKM and FKM, it is useful to describe the residual structure of both models, similar to Timmerman et al. (2010). In this section, we denote $\mathbf{U}_C$ as $\mathbf{U}$ for simplification, yet these are equal. We describe the residual structure by fitting the RKM model:

$$\mathbf{X} = \mathbf{UFA}' + \mathbf{E}_R, \tag{3.20}$$

where we define $\mathbf{E}_R$ as the $(N \times m)$ residual matrix. From this equation, we can derive the optimal centroid matrix $\mathbf{F} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{XA}$. Thus, we can rewrite the RKM model 3.20 as:

$$\mathbf{X} = \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{XAA}' + \mathbf{E}_R \tag{3.21}$$

For the FKM model, we define it as specified by Vichi & Kiers (2001):

$$\mathbf{XAA}' = \mathbf{UFA}' + \mathbf{E}_F, \tag{3.22}$$

with $\mathbf{E}_F$ as the $(N \times m)$ residual matrix. The optimal centroid matrix of the FKM model is derived as:

$$\mathbf{F} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{XA}\underbrace{\mathbf{A}'\mathbf{A}}_{\mathbf{I}_Q} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{XA}, \tag{3.23}$$

and so we can rewrite model 3.22 as:

$$\mathbf{XAA'} = \mathbf{U}(\mathbf{U'U})^{-1}\mathbf{U'XAA'} + \mathbf{E}_F, \tag{3.24}$$

When we compare model 3.24 with model 3.21, we can see that the FKM model uses centroids and observations that both lie in the reduced space, while RKM assumes that only the centroids lie in the reduced space.

When we rewrite model 3.22 to $\mathbf{E}_F = \mathbf{XAA'} - \mathbf{UFA'}$, we see that $\mathbf{E}_F$ lies in the row space of $\mathbf{A'}$. This expression suggests that $\mathbf{E}_F = \mathbf{EA'}$ exists with a residual matrix $\mathbf{E}$ that lies in the same column space as $\mathbf{E}_F$. From this expression, it shows that $\mathbf{E}_F$ consists of the subspace residuals ($\mathbf{E}$) that are projected back to the observed space of $\mathbf{X}$ by post-multiplication of $\mathbf{A'}$. With this expression, we can rewrite Equation 3.22 as:

$$\mathbf{XAA'} = \mathbf{UFA'} + \mathbf{E}_F = \mathbf{UFA'} + \mathbf{EA'} = (\mathbf{UF} + \mathbf{E})\mathbf{A'} \tag{3.25}$$

We can derive this further to:

$$\mathbf{XA} = \mathbf{UF} + \mathbf{E} \tag{3.26}$$

When we want to write the model equation as a function of solely the observed data $\mathbf{X}$, we need to define the complement residuals $\mathbf{X} - \mathbf{XAA'} = \mathbf{E}^{\perp}\mathbf{A}^{\perp}$ for which it holds that $\mathbf{A'A}^{\perp} = 0$. Using this expression in Equation 3.25, we write:

$$\mathbf{X} = \mathbf{UFA'} + \mathbf{EA'} + \mathbf{X} - \mathbf{XAA'} = \mathbf{UFA'} + \mathbf{EA'} + \mathbf{E}^{\perp}\mathbf{A}^{\perp} \tag{3.27}$$

In Equation 3.27, the full residual model, we see the contribution of the subspace residuals $\mathbf{EA'}$ and the complement residuals $\mathbf{E}^{\perp}\mathbf{A}^{\perp}$.

From the objective functions specified in Section 3.2, we can derive the most optimal data for RKM and FKM. We can see that the FKM loss function (3.15) is 0 if and only if $\mathbf{XA} = \mathbf{UF}$. This is only the case when the subspace residuals $\mathbf{E}$ are 0. The RKM loss function (3.8) is only 0 if and only if $\mathbf{X} = \mathbf{UFA'}$, which is the case when $\mathbf{E} = 0$ and $\mathbf{E}^{\perp} = 0$, *i.e.* no subspace and no complement residuals (Timmerman et al., 2010).

Chapter 4

Simulation study

In the simulation study, we assess the performance of tandem and joint methods on simulated cancer omics data in the presence of three types of noise: random noise, masking variables and subspace and complement noise (Figure 4.1).

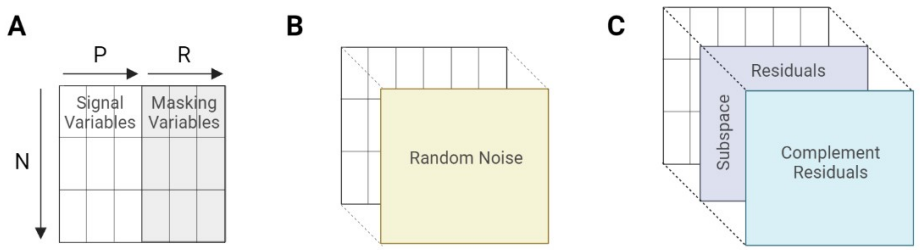

Figure 4.1: Graphic depiction of the three types of noise used in this study. A) Masking variables are added to the column space. B) Random Noise is generated and added to all elements. C) Subspace and complement residuals are computed and added to all elements.

We add random noise to mimic the sampling noise found in empirical data. Masking variables are a source of complement noise and have a deteriorating impact on RKM and FKM (Yamamoto & Hwang, 2014). As we recall from the Literature, the performance of RKM and FKM depends on the residual structure. RKM minimises the distance between the observed data and the centroids in the subspace projected by the loading matrix (see Section 3.2.1) and its perfect data has no subspace and no complement residuals (see Section 3.2.4) (Timmerman et al., 2010). FKM minimises the projected observed datapoints and the projected cluster centroids (see Section 3.2.2) and its perfect data has no subspace residuals (see Section 3.2.4). Hence, their objective functions depend on the Proportion of Subspace Residual variance (PSR), which is a measure for the size of the subspace residuals compared to the complement residuals.

The simulation study includes two experiments that use the same data-generating method to create the signal and masking variables, the random noise and the subspace and complement residuals. We first explain how the data is generated, whereafter we explain the approach and details of each experiment.

4.1 Data

We introduce the notation used in this section:

N	Number of observations ($i = 1, \ldots, N$)
C	Number of clusters ($c = 1, \ldots, C$)
Q	Number of components ($q = 1, \ldots, Q$)
P	Number of signal variables ($p = 1, \ldots, P$)
R	Number of masking variables ($r = 1, \ldots, R$)
$\mathbf{X}$	$N \times (P + R)$ data matrix with $\mathbf{X} = (\mathbf{X_1}\|\mathbf{X_2})$
$\mathbf{U}$	$N \times C$ binary membership matrix
$\mathbf{F}$	$N \times Q$ centroid matrix
$\mathbf{A}$	$(P + R) \times Q$ loading matrix with $\mathbf{A}' = (\mathbf{A_1'}\|\mathbf{A_2'})$
$\mathbf{N}$	$N \times (P + R)$ noise matrix
$\mathbf{EA}'$	$N \times (P + R)$ subspace residual matrix
$\mathbf{E}^{\perp}\mathbf{A}^{\perp\prime}$	$N \times (P + R)$ complement residual matrix

4.1.1 Signal and masking variables

The base matrix $\mathbf{X}$ is generated in two parts, such that $\mathbf{X} = (\mathbf{X_1}|\mathbf{X_2})$ (Figure 4.1A). $\mathbf{X_1}$ denotes the signal component with $P = 10$ variables related to the cluster structure. $\mathbf{X_2}$ is composed of R masking variables. These variables are correlated to each other but are not related to the cluster structure. $\mathbf{X_1}$ and $\mathbf{X_2}$ are generated as follows:

$$\mathbf{X_1} = \mathbf{F_1}\mathbf{A_1'} \tag{4.1}$$

$$\mathbf{X_2} = \mathbf{F_2}\mathbf{A_2'}, \tag{4.2}$$

where we generate true component score matrix $\mathbf{F_1}$ using:

$$\mathbf{F_1} = \mathbf{U}\begin{bmatrix} c_1 & c_2 & c_1 & c_2 \\ c_1 & c_2 & c_2 & c_1 \end{bmatrix}' \tag{4.3}$$

We simulate the cluster membership matrix $\mathbf{U}$ to have an equal proportion of cluster observations, that is, 15 observations in each cluster. We manipulate the distances between the clusters with parameters c_1 and c_2, analysing how cluster distances affect cluster algorithm performances.

We independently sample the elements in $\mathbf{F_2} \sim \mathcal{N}(\frac{c_1+c_2}{2}, \frac{|c_1-c_2|}{\sqrt{2}})$ to simulate a random component score matrix such that the R masking variables do not reflect the cluster structure. Furthermore, we sample the elements of the loading matrix $\mathbf{A_1}$, $\mathbf{A_2} \sim \mathcal{P}(\lambda = 3)$. We choose $\lambda = 3$ because this is a not heavy-tailed distribution as compared to higher lambdas, which results in a non-sparse matrix with no extreme outliers. We choose a Poisson distribution because of the count-based non-Gaussian nature of omics data (Stein-O'Brien et al., 2018).

We fix the number of observations to $N = 60$, which is similar to the smallest level of observations used in Timmerman et al. (2010). We consider 60 observations because the number of observations in RNA-seq omics datasets is little compared to the number of variables (Grossman et al., 2016). Furthermore, we fix the simulation study to four clusters ($C = 4$) with two components ($Q = 2$), similar to Yamamoto & Hwang (2014).

4.1.2 Random noise

We generate random noise as follows (Figure 4.1B):

$$\frac{\mathbf{N}}{||\mathbf{N}||}||\mathbf{N}||\sqrt{\frac{Noise}{1-Noise}}, \tag{4.4}$$

which we add to the observed generated signal and masking variables:

$$(\mathbf{X_1}|\mathbf{X_2}) + \frac{\mathbf{N}}{||\mathbf{N}||}||\mathbf{N}||\sqrt{\frac{Noise}{1-Noise}}, \tag{4.5}$$

where we sample $\mathbf{N} \sim \mathcal{N}(0, 1)$ and manipulate the factor $Noise$ relative to the Sum of Squares of the signal matrix $\mathbf{X}$.

4.1.3 Subspace and complement residuals

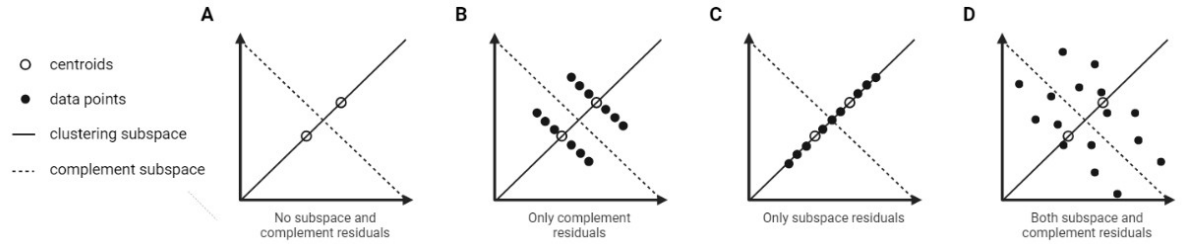

Figure 4.2: Graphic simplified depiction of the space where the centroids exist, and how the subspace and complement residuals are related to that subspace. Based on Timmerman et al. (2010).

We add subspace noise as follows, similar to Timmerman et al. (2010):

$$(\mathbf{X_1}|\mathbf{X_2}) + \mathbf{E_s}(\mathbf{A_1'}|\mathbf{A_2'}) + \mathbf{E_c^{\perp}}(\mathbf{A_1}|\mathbf{A_2})^{\perp'}, \tag{4.6}$$

where the second and third terms compose the subspace and complement residuals, respectively (Figure 4.1C). The decomposition of the residuals into subspace and complement residuals is described in detail in Section 3.2.4. The complement loading matrix $(\mathbf{A_1}|\mathbf{A_2})^{\perp'}$ is computed with the criterion that $(\mathbf{A_1'}|\mathbf{A_2'})(\mathbf{A_1}|\mathbf{A_2})^{\perp} = \mathbf{0}$ (Timmerman et al., 2010). Figure 4.2 shows a simplified version of how subspace and complement residuals reside in the cluster subspaces. We sample subspace residual matrix $\mathbf{E_s} \sim \mathcal{N}(0, \sigma^2_{\mathbf{E_s}} = 1)$ and complement residual matrix $\mathbf{E_c^{\perp}} \sim \mathcal{N}(0, \sigma^2_{\mathbf{E_c^{\perp}}})$. We tune the term $\sigma^2_{\mathbf{E_c^{\perp}}}$ with the Proportion of Subspace Residual variance (PSR) (Timmerman et al., 2010):

$$PSR = \frac{\sigma^2_{\mathbf{E_s}}}{\sigma^2_{\mathbf{E_s}} + \sigma^2_{\mathbf{E}^{\perp}_{\mathbf{c}}}} = \frac{1}{1 + \sigma^2_{\mathbf{E}^{\perp}_{\mathbf{c}}}} \tag{4.7}$$

4.2 Approach

4.2.1 Experiment 1. Random noise and subspace residuals

This experiment aims to assess how well the MF and clustering algorithms can recover the cluster subspaces and cluster memberships when the level of random noise and subspace residuals is increased (Figure 4.3), and how it depends on cluster overlap and the number of masking variables. We generate the data as previously described, and construct $\mathbf{X_2}$ with $R \in \{0, 5, 10\}$ masking variables. We manipulate the distances between the clusters by defining $c_1 = 10$ and $c_2 \in \{13, 15, 17\}$. We set $Noise \in \{0.01, 0.05, 0.10, 0.15\}$ for generating the random noise. For the subspace and complement residuals, we vary the PSR with four levels: $PSR \in \{0.01, 0.05, 0.10, 0.15\}$, which captures a dynamic range in the cluster performances similar to Timmerman et al. (2010). As NMF requires non-negative elements, we permute all negative generated elements with $\sim \mathcal{B}(p = 0.5)$ to set them to 0 or 1 with a chance of 50%. As this introduces a bias in the generated data, we report this limitation as the Permutation Fraction (PF), which depicts the fraction of the elements that have been permuted.

We fully cross the model specification of both approaches and replicate each combination 10 times (Table 4.1). This results in $3 \times 3 \times 4 \times 10 + 3 \times 3 \times 4 \times 10 = 720$ simulated data sets. We evaluate all joint and tandem clustering algorithms with the quality criteria (Section 4.3) and compute the partial influences of each parameter using a full-factorial Repeated Measures ANalysis Of VAriance (RMANOVA) (Section 4.3.2). The approach is summarised in Figure 4.3.

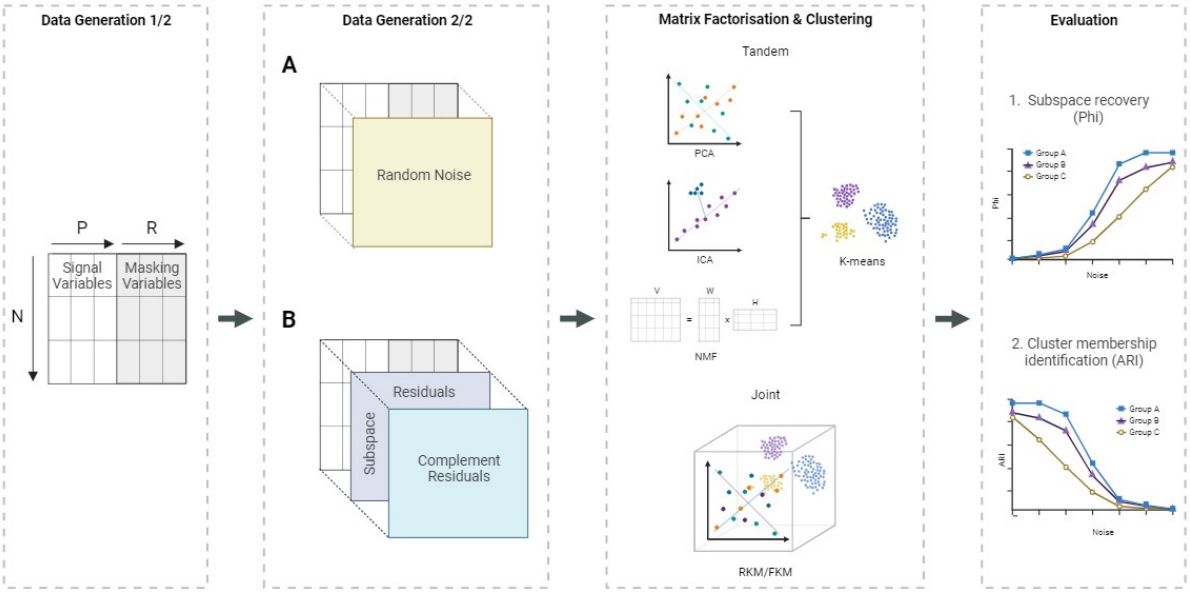

Figure 4.3: Flowchart of the approach used to analyse the performance of the joint and tandem approaches. First, the signal and masking variables are generated. Secondly, either A) random noise is added as described in Section 4.1.2, or B) subspace and complement residuals are added as described in Section 4.1.3. After this, MF and clustering algorithms are performed and evaluated regarding subspace recovery and cluster membership identification.

Table 4.1: Parameter initialisation settings simulation study in Experiment 1.

Parameter	**Initialisation settings**
Centroid 1 (c_1)	10
Centroid 2 (c_2)	{13, 15, 17}
Number of observations (N)	60
Number of clusters (C)	4
Number of components (Q)	2
Number of cluster variables (P)	10
Number of masking variables (R)	{0, 5, 10}
Loading matrix mean and variance (λ)	3
A. Level of noise ($Noise$)	{0.01, 0.05, 0.10, 0.15}
B. Proportion of Subspace Residual variance (PSR)	{0.01, 0.05, 0.10, 0.15}

4.2.2 Experiment 2. The alpha parameter

In this experiment, we aim to analyse how compromises between FKM, RKM and PCA perform in the presence of random noise, subspace noise and masking variables (Figure 4.4). As we recall from the Literature, the loss functions of FKM and RKM can be decomposed and integrated into one function, developed for R as function `cluspca` (Markos, D'enza & van de Velden, 2019). This decomposition is described in detail in Section 3.2.3. By changing the Alpha parameter in the algorithm, we can assess the performance of FKM ($\alpha = 0$), RKM ($\alpha = 0.5$), PCA ($\alpha = 1$) and a mixture of these models ($0 < \alpha < 0.5$ || $0.5 < \alpha < 1$). Hence, can find the optimal Alpha for each situation.

We generate the data as described in Section 4.1.1. First, we assess the cluster algorithms in the case of random noise, varying $Noise \in \{0.01, 0.03, 0.05, 0.07, 0.09, 0.11, 0.13, 0.15\}$ and setting $R = 0$. Secondly, we assess the Alpha parameter in the presence of subspace residuals, setting $R = 0$ and $PSR \in \{0.01, 0.03, 0.05, 0.07, 0.09, 0.11, 0.13, 0.15\}$. Again, as NMF requires non-negative elements, we permute all negative generated elements with $\sim \mathcal{B}(p = 0.5)$ to 0 or 1 with a chance of 50%. As this introduces a bias in the generated data, we report this limitation as the PF, which depicts the fraction of the elements that have been permuted. Lastly, we analyse the clustering performance in the presence of masking variables, setting $Noise = 0$, $PSR = 0$ and varying $R \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ (Table 4.2). We evaluate the performances according to the subspace recovery and the cluster membership identification (Section 4.3).

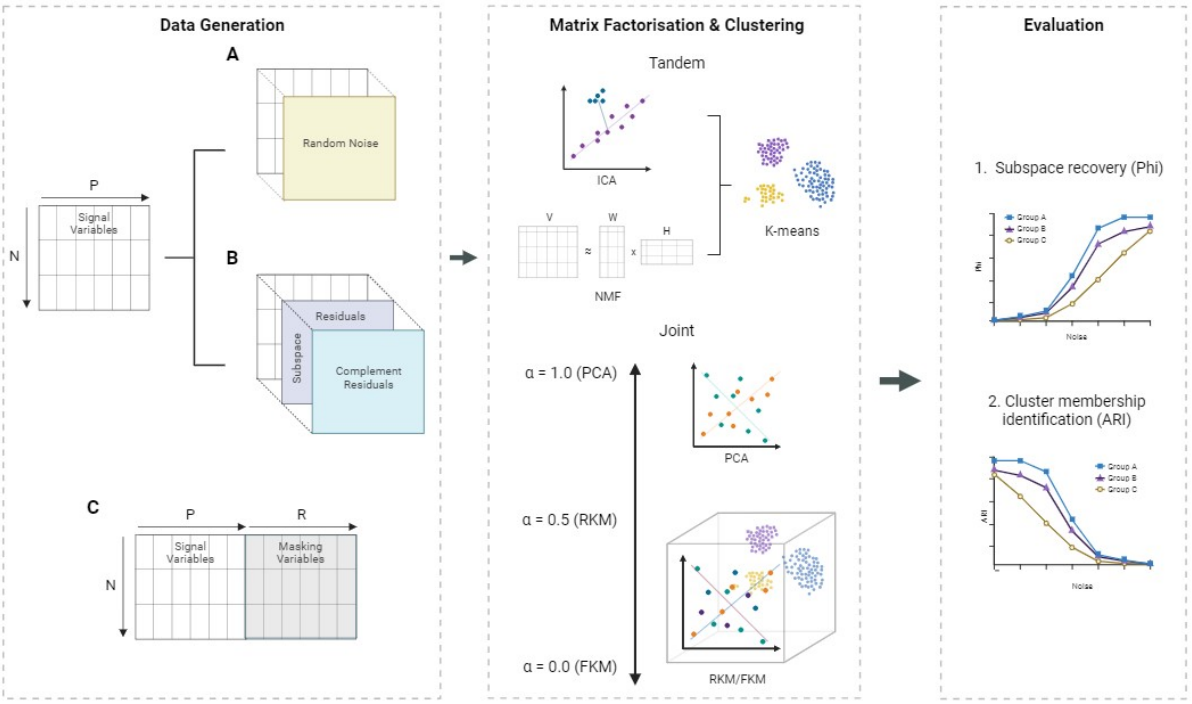

Figure 4.4: Flowchart of the methods used to analyse the performance of compromises between FKM, RKM and PCA. First, the data is generated, either with A) random noise, B) subspace and complement residuals, or C) masking variables. After this, the *cluspca* algorithm is performed with α ranging from 0.0 (FKM) to 1.0 (PCA), along with ICA and NMF. Performance is evaluated using subspace recovery and cluster membership identification.

Table 4.2: Parameter initialisation settings simulation study in Experiment 2.

Parameter	**Initialisation settings**
Alpha (α)	{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1}
Centroid 1 (c_1)	10
Centroid 2 (c_2)	17
Number of observations (N)	60
Number of clusters (C)	4
Number of components (Q)	2
Number of cluster variables (P)	10
Loading matrix mean and variance (λ)	3
A. Level of noise ($Noise$)	{0.01, 0.03, 0.05, 0.07, 0.09, 0.11, 0.13, 0.15}
B. Proportion of Subspace Residual variance (PSR)	{0.01, 0.03, 0.05, 0.07, 0.09, 0.11, 0.13, 0.15}
C. Number of masking variables (R)	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}

4.3 Performance evaluation

4.3.1 Quality criteria

Similar to Timmerman et al. (2010), we use subspace recovery and cluster membership identification as quality criteria. The mean of the Tucker congruence (Phi) coefficients is used to measure how well the subspace is recovered, where Phi represents the proportionality between the columns of the estimated and the simulated loading matrices (Kuhn & Tucker, 1951).

The Adjusted Rand Index (ARI) measures the recovery of the underlying cluster structure. ARI is the corrected version of the Rand index, i.e. the corrected-for-chance version. It measures the similarity between the two partitions. It takes a value of (-)1 if the partitions are perfectly proportional, and a value of 0 when the cluster membership matrices $\mathbf{U}$ and $\hat{\mathbf{U}}$ are not more proportional than that was expected by chance (Hubert & Arabie, 1985). Specifically, following a similar notation as Jaskowiak et al. (2018), it calculates for resulting partition $\mathcal{Z}_1$ and *a priori* defined cluster partition $\mathcal{Z}_2$:

$$ARI(\mathcal{Z}_1, \mathcal{Z}_2) = \frac{a - \frac{(a+c)(a+b)}{(a+b+c+d)}}{\frac{(a+c)(a+b)}{2} - \frac{(a+c)(a+b)}{(a+b+c+d)}}, \tag{4.8}$$

where a denotes the number of pairs of observations that are in the same clusters in partitions $\mathcal{Z}_1$ and $\mathcal{Z}_2$, b denotes the pairs of observations that are in the same cluster in $\mathcal{Z}_1$ and in different clusters in $\mathcal{Z}_2$, c denotes the pair of observations that are in different clusters in $\mathcal{Z}_1$ and in the same cluster in $\mathcal{Z}_2$, and d denotes the pair of observations that are in different clusters in partitions $\mathcal{Z}_1$ and $\mathcal{Z}_2$.

4.3.2 Statistical tests

We measure the partial effect of the parameters on the subspace recovery (Phi) and the ARI with a full-factorial RMANOVA, similar to Timmerman et al. (2010). We use the function `ezANOVA` from the R package **rstatix** (Kassambara, 2023). This function computes the generalised η^2 measure of the effect size of each factor (Bakeman, 2005). We follow the rule-of-thumb measures for magnitudes of effect sizes by Miles & Shevlin (2001) that states $\eta^2 = 0.01$ is a small effect, $\eta^2 = 0.06$ is a medium effect and $\eta^2 = 0.14$ is a large effect.

4.4 Results

4.4.1 Experiment 1. Random noise and subspace residuals

In this experiment, we aim to investigate how the performance of joint and tandem methods depends on the factors random noise ($Noise$), the proportion of subspace variance (PSR), the number of masking variables (R) and distance between centroids (c_2) (see Section 4.2.1). The performances are evaluated with the subspace recovery using Phi and the clustering accuracy using ARI (see Section 4.3). We further compute the partial effects of the factors using the RMANOVA statistical test (see Section 4.3.2). Firstly, we discuss the performances when we increase the level of random noise. Secondly, we discuss the performances when we increase the size of the subspace variance with PSR.

Random noise

We observe in Table 4.3 that RKM has the highest ARI when we average over all factors. We also observe that NMF has the highest mean subspace recovery.

The accuracy in the partitioning of the clusters decreases when the distances between the centroids become smaller and the number of masking variables becomes larger (Table 4.3, Figure

4.5A). Table 4.3 and Figure 4.5B show that there is no significant change in subspace recovery when we change the number of masking variables. Subspace recovery slightly increases when the distances between the centroids get larger (for all results, see Appendix Table A.1 and Figure A.1).

When plotting the clustering accuracy and against the level of noise, we see that RKM has the highest clustering accuracy when the level of random noise is increased, starting from a perfect recovery ($ARI = 1.00$) and deteriorating to $ARI = 0.25$ (Figure 4.5A). NMF is superior in recovering the subspace, staying stable around $Phi = 0.9$ (Figure 4.5B). PCA and ICA follow similar performances as RKM in recovering the subspace and cluster memberships but are slightly worse when there are no masking variables ($R = 0$). While NMF can almost perfectly capture the subspace ($Phi = 0.9$) even when there are masking variables present, it is worse at recovering cluster memberships compared to PCA, ICA and RKM. We also observe that FKM has the worst performance in both quality criteria ($ARI = 0.05, Phi = 0.2$).

We further investigate the complex interactions of the factors using a RMANOVA test, measuring how factors partially affect the variance in the results. We observe that the choice of method has a large partial effect on the performance ($\eta^2 = 0.87$), yet this is not significant (Table 4.4). Centroid distance, Noise, and masking variables are all significant factors, from which only $Method \times Noise$ has a large effect ($\eta^2 = 0.17^{***}$) on subspace recovery and clustering accuracy.

Table 4.3: Summary of results, averaged over ten replications and the factor $Noise$.

		PCA		**ICA**		**NMF**		**RKM**		**FKM**	
		ARI	Phi	ARI	Phi	ARI	Phi	ARI	Phi	ARI	Phi
$R = 0$	$c_2 = 13$	0.24	0.54	0.22	0.55	0.06	0.93	0.25	0.55	0.03	0.25
	$c_2 = 15$	0.39	0.58	0.38	0.62	0.23	0.94	0.41	0.59	0.05	0.21
	$c_2 = 17$	0.55	0.60	0.48	0.62	0.34	0.95	0.56	0.60	0.08	0.19
$R = 5$	$c_2 = 13$	0.20	0.53	0.15	0.54	0.06	0.91	0.19	0.52	0.03	0.24
	$c_2 = 15$	0.28	0.60	0.23	0.62	0.10	0.91	0.28	0.59	0.03	0.16
	$c_2 = 17$	0.32	0.62	0.26	0.62	0.14	0.89	0.34	0.62	0.02	0.14
$R = 10$	$c_2 = 13$	0.13	0.55	0.14	0.56	0.05	0.91	0.12	0.53	0.02	0.18
	$c_2 = 15$	0.19	0.59	0.21	0.63	0.07	0.90	0.19	0.59	0.01	0.14
	$c_2 = 17$	0.23	0.62	0.28	0.66	0.15	0.88	0.24	0.61	0.02	0.12
	Mean	0.28	0.58	0.26	0.60	0.13	0.91	0.29	0.58	0.03	0.18

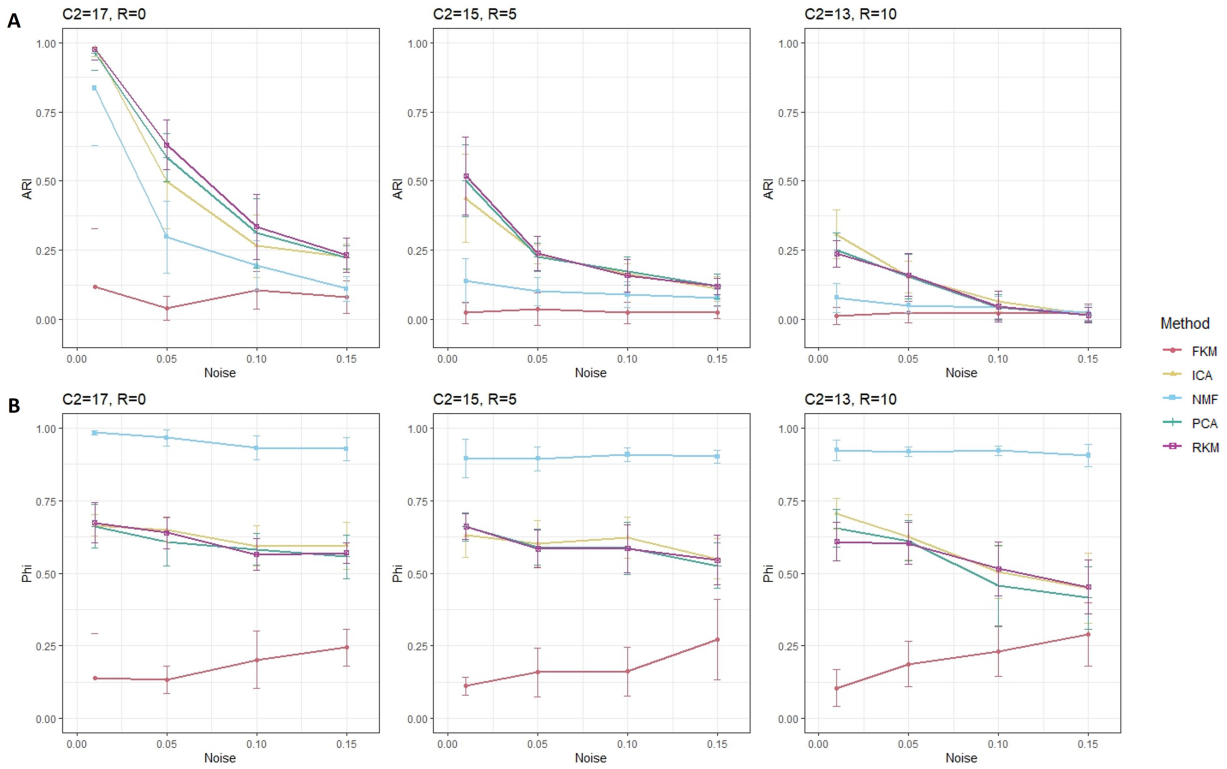

Figure 4.5: A) Clustering accuracy results are measured with the ARI. B) Subspace recovery results are measured with the Tucker congruence coefficient (Phi). *Note: C2: initialisation setting centroid 2, R: number of masking variables (see Section 4.1)*

Table 4.4: Full-factorial RMANOVA test for Phi and the ARI.

Factors	**Phi** $[\eta^2]$	**ARI** $[\eta^2]$
Method	0.87	0.87
Method x c	0.09***	0.02***
Method x Noise	0.17***	0.17***
Method x R	0.02***	0.02***
Method x c x Noise	0.06***	0.01
Method x c x R	0.01	0.00
Method x Noise x R	0.03***	0.08***
Method x c x Noise x R	0.03	0.08***

Note: (i) (***): $p \leq 0.001$, (**): $p \leq 0.01$, (*): $p \leq 0.05$. All p-values are corrected for multiple hypothesis testing using the Bonferroni method. (ii) $\eta^2 = 0.01$: small effect, $\eta^2 = 0.06$: medium effect and $\eta^2 = 0.14$: large effect.

Subspace residuals

We observe in Table 4.5 that similar to the previous results, RKM has the highest *ARI* and NMF has the highest mean subspace recovery when we average the results over all factors.

When we decrease the distance between the centroids and increase the number of masking variables, we see that the clustering accuracy decreases (Table 4.5, Figure 4.6). The subspace recovery slightly increases when the *PSR* increases. The subspace recovery stays relatively stable when the number of masking variables changes (for all results, see Appendix Table A.2 and Figure A.2).

When we plot the clustering accuracy against the level of PSR, we see that RKM has the largest clustering accuracy (Figure 4.6A). In the case of no masking variables and a large centroid distance, all methods except for FKM plateau around ARI of 0.95 when the PSR is increased. However, PCA and RKM reach this plateau when $PSR = 0.05$, and ICA and PCA reach $ARI = 0.95$ when $PSR = 0.10$. If we look at the other quality criteria, the subspace recovery depicted in Figure 4.6B, we see that NMF has an almost perfect subspace recovery and that FKM can barely recover elements of the loading matrix. We also observe that ICA starts at a Phi of 0 but increases to a stable Phi around 0.70 when $PSR > 0.05$.

The RMANOVA test, depicted in Table 4.6, shows that the choice of method has the largest effect ($\eta^2 = 0.81$), although this effect is not significant. Regarding subspace recovery, all other grouped factors have a small effect, with the exception of $Method \times PSR$ ($\eta^2 = 0.60^{***}$). This suggests that the choice of Method and PSR is important to capture the loading matrices of the subspace well. For the clustering accuracy, we see that next to $Method \times PSR$, $Method \times PSR \times R$ and $Method \times c \times PSR \times R$ have medium partial effects on the variance.

We explained in Section 4.2.1 that we permuted the generated data if the elements are negative, because NMF requires non-negative elements. The maximum fraction of permuted elements in this experiment is 6.75% when $PSR = 0.01$, $R = 0$ and $c_2 = 13$ (Table A.3). This suggests that when there are more complement residuals generated, 6.75% of the elements are changed to 0 or 1. In this case, methods that are designed to be suitable for complementary residuals, like FKM, are less likely to perform well (Timmerman et al., 2010). This is in line with the result we see in Figure 4.6.

Table 4.5: Summary of results, averaged over ten replications and the factor PSR.

		PCA		**ICA**		**NMF**		**RKM**		**FKM**	
		ARI	Phi	ARI	Phi	ARI	Phi	ARI	Phi	ARI	Phi
$R = 0$	$c_2 = 13$	0.28	0.56	0.27	0.46	0.16	0.91	0.30	0.57	0.10	0.10
	$c_2 = 15$	0.60	0.61	0.58	0.50	0.42	0.91	0.65	0.62	0.18	0.11
	$c_2 = 17$	0.75	0.62	0.66	0.52	0.61	0.93	0.80	0.65	0.33	0.10
$R = 5$	$c_2 = 13$	0.18	0.58	0.14	0.48	0.08	0.90	0.19	0.58	0.06	0.10
	$c_2 = 15$	0.34	0.61	0.22	0.51	0.10	0.89	0.38	0.62	0.13	0.10
	$c_2 = 17$	0.44	0.62	0.24	0.52	0.14	0.88	0.48	0.61	0.20	0.08
$R = 10$	$c_2 = 13$	0.16	0.57	0.15	0.53	0.08	0.91	0.16	0.58	0.05	0.08
	$c_2 = 15$	0.23	0.57	0.22	0.55	0.09	0.90	0.24	0.56	0.08	0.08
	$c_2 = 17$	0.29	0.57	0.28	0.59	0.13	0.87	0.29	0.57	0.10	0.08
	Mean	0.36	0.59	0.31	0.52	0.20	0.90	0.39	0.59	0.14	0.09

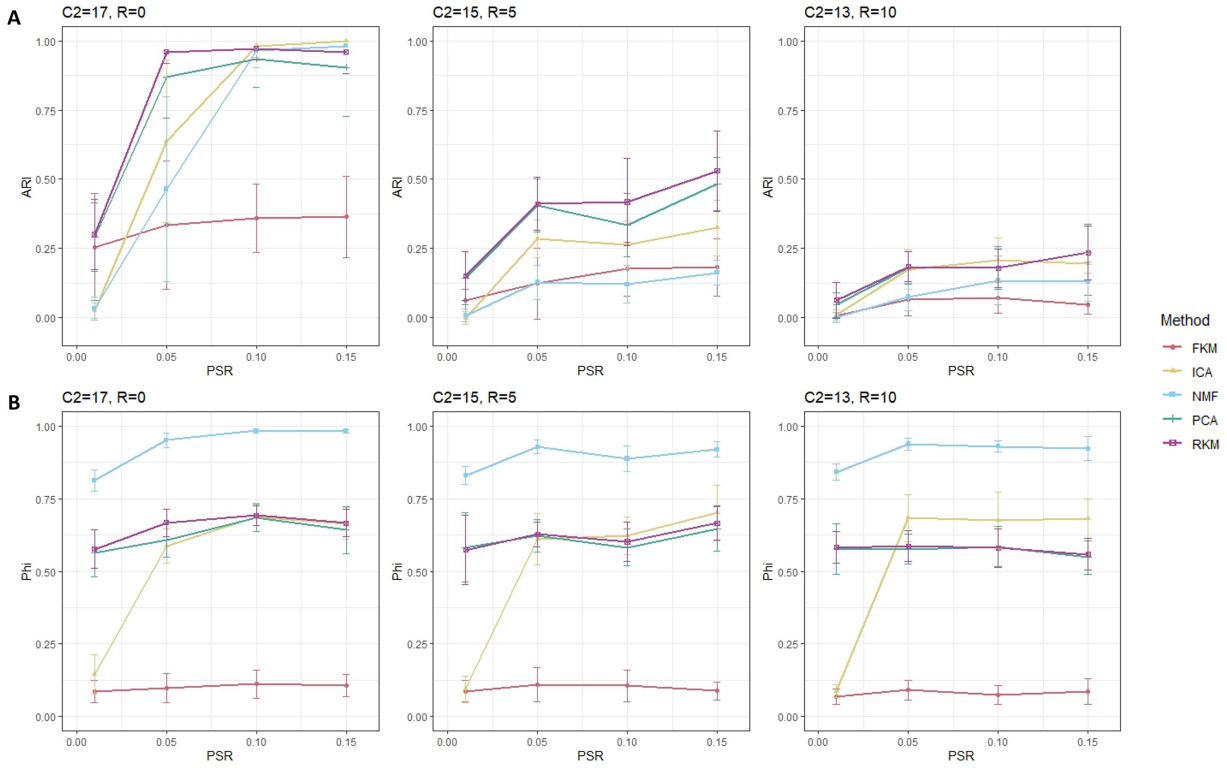

Figure 4.6: A) Clustering accuracy results are measured with the ARI. B) Subspace recovery results are measured with the Tucker congruence coefficient (Phi). *Note: C2: initialisation setting centroid 2, R: number of masking variables (see Section 4.1)*

Table 4.6: Full-factorial RMANOVA test for the Phi and the ARI.

Factors	**Phi** $[\eta^2]$	**ARI** $[\eta^2]$
Method	0.81	0.81
Method x c	0.01*	0.01***
Method x PSR	0.60***	0.05***
Method x R	0.02***	0.03***
Method x c x PSR	0.03***	0.01**
Method x c x R	0.00	0.01***
Method x PSR x R	0.04***	0.08***
Method x c x PSR x R	0.04*	0.09***

Note: (i) (***): $p \leq 0.001$, (**): $p \leq 0.01$, (*): $p \leq 0.05$. All p-values are corrected for multiple hypothesis testing using the Bonferroni method. (ii) $\eta^2 = 0.01$: small effect, $\eta^2 = 0.06$: medium effect and $\eta^2 = 0.14$: large effect.

4.4.2 Experiment 2. The alpha parameter

In the second experiment, we aim to analyse which joint, compromise, or tandem approach is the most suitable in the presence of random noise, subspace noise and masking variables (see Section 4.2.2). By changing the alpha parameter, we tested compromises between FKM ($\alpha = 0$), RKM ($\alpha = 0.5$) and PCA ($\alpha = 1$) in the presence of only random noise, subspace and complement residuals or masking variables. With the heatmap visualisation of the results, we can compare which method is the most suitable in each situation (Figure 4.7, full results in Table A.5-A.10).

We see in Figure 4.7A that when we increase the random noise, compromises between FKM and RKM perform well compared to the other methods regarding clustering accuracy ($N = 0.01$, $ARI = 0.98$; $N = 0.13$, $ARI = 28$) (Table A.5). When the loss function is fully specified to FKM, it has the worst performance ($ARI \simeq 0$). Similar to the first experiment, we see that NMF can capture the subspace almost perfectly ($N = 0.01, Phi = 0.99; N = 0.15, Phi = 0.92$), whereas other methods do not surpass $Phi = 0.70$ (Table A.6).

We see in Figure 4.7B that when $0.03 < PSR \leq 0.09$, RKM has the highest clustering accuracy ($0.93 < ARI < 0.99$) (Table A.7). When $PSR \geq 0.11$, ICA has the highest clustering accuracy ($ARI = 1.00$). Compromises between RKM and PCA also perform well. FKM has the worst performance, having a maximum of $ARI = 0.43$ when $PSR = 0.15$. NMF shows the best subspace recovery ($0.82 < Phi < 0.98$), and compromises between RKM and PCA also show considerable results ($0.55 < Phi < 0.68$) (Table A.8). We also observe that for low PSR values, compromises between $0.5 < \alpha < 1$ perform considerably better than ICA regarding clustering accuracy and subspace recovery.

Similar to the first experiment, we check whether the generated data for the PSR analysis is non-negative. We measured that the maximum fraction of permuted elements was 7.85% when $PSR = 0.01$. Although there could be a bias in the results, the performance of FKM is substantially worse than other methods. This suggests that without the bias, FKM would still perform worse.

When we add masking variables to the data, we see that compromises between FKM and RKM have the highest clustering accuracy ($0.75 < ARI < 1.00$) (Figure 4.7C, Table A.9). We see that when the fraction of masking variables is increased, the optimal value of α decreases from $0.1 < \alpha < 0.7$ to $\alpha = 0.1$. While for the joint methods perfect clustering accuracy is still possible when $R = 10$, the tandem methods PCA, ICA and NMF have an ARI of 0.43, 0.50 and 0.16, respectively.

NMF shows an almost perfect subspace recovery ($0.87 < Phi < 0.95$) (Figure 4.7C, Table A.10). Interestingly, we see that the ARI and Phi results contrast each other, as compromises between FKM and RKM do well for clustering accuracy, while for subspace recovery the compromises between RKM and PCA are better ($0.58 < Phi < 0.67$). Nonetheless, the ICA ($0.64 < Phi < 0.70$) and NMF ($0.87 < Phi < 0.95$) tandem methods outperform the compromises in recovering the subspace.

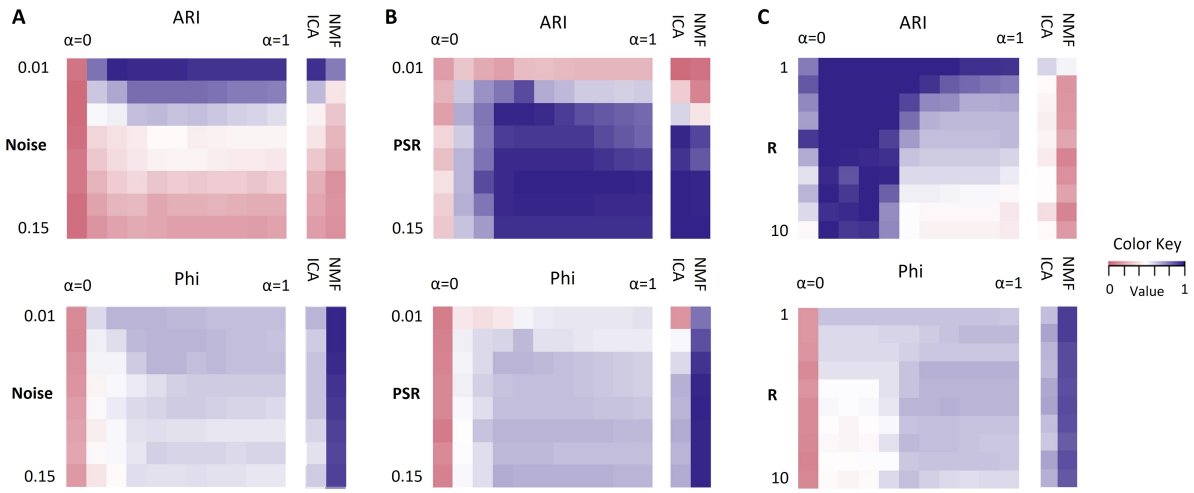

Figure 4.7: Heatmaps of clustering accuracy and subspace recovery. By changing the *alpha* parameter in the algorithm, we can assess the performance of FKM ($\alpha = 0$), RKM ($\alpha = 0.5$), PCA ($\alpha = 1$) and a mixture of these models ($0 < \alpha < 0.5$ || $0.5 < \alpha < 1$). A) Results in the case of adding random noise ($Noise$). B) Results in the case of increasing the size of the variance of the residuals that lie in the subspace (PSR). C) Results in the case of adding masking variables (R).

Chapter 5

Empirical analysis

In this chapter, we expand our research to the empirical setting. In Section 5.1, we describe the empirical data used for this analysis. Next, we explain the approach in Section 5.2. The performance evaluation is described in Section 5.3. The results of the empirical analysis are depicted in Section 5.4.

5.1 Data

The empirical data is obtained from The Cancer Genome Atlas (TCGA) databank through Broad institute GDAC Firehose[1] (Grossman et al., 2016). The experimental RNA-seq raw count data was collected using Illumina HiSeq 2000 RNA Sequencing Version 2. The cancer types brain (Lower Grade Glioma (LGG)), breast (Breast Invasive Carcinoma (BRCA)), kidney type 1 (Kidney Renal Papillary Cell Carcinoma (KIRP)), kidney type 2 (Kidney Renal Clear Cell Carcinoma (KIRC)), stomach type 1 (Somach Adenocarcinoma (STAD)), stomach type 2 (Stomach and Esophageal Carcinoma (STES)), uterine type 1 (Uterine Corpus Endometrial Carcinoma (UCEC)) and uterine type 2 (Uterine Carcinosarcoma (UCS)) are included in the analysis (Table 5.1). Each cancer dataset consists of the same $m = 20{,}531$ genes. Only primary solid tumours were included. We choose to include four cancer types whose gene expression patterns differ such that the clusters are distinguishable. We include subtypes for kidney, stomach and uterine cancer to create overlapping clusters, making the clustering task less trivial.

We want a dataset that consists of clusters with ground-truth labels for the empirical analysis and a feasible number of samples and features. Hence, we combine 25 randomly sampled observations from each cancer-specific dataset in a pan-cancer dataset. In this way, the pan-cancer dataset has 8×25 samples, with eight equally distributed clusters (Table 5.2). We choose to randomly sample observations as we want to limit possible introduced bias from clinicopathological features of the samples. Furthermore, 8×25 samples ensure that there are enough observations while retaining feasibility regarding computer memory.

We see in Figure 5.1A that the distribution of the logarithm of the mean and standard deviation of the genes in the cancer types are similar. We see in Figure 5.1B that some cancer (sub)types

[1] https://gdac.broadinstitute.org/

have overlapping cluster structures (i.e. STAD and STES), while other clusters do not have overlapping local cluster structures (i.e. LGG).

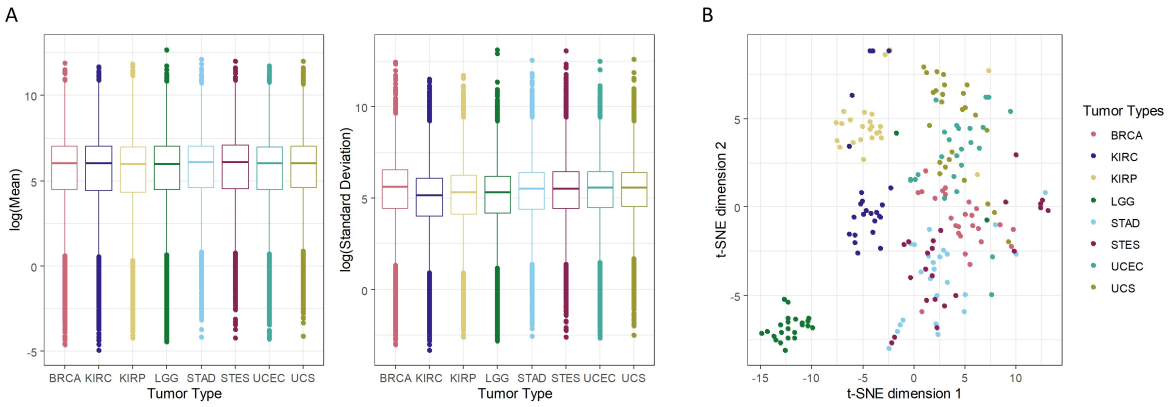

Figure 5.1: Characteristics of the pan-cancer dataset. A) Distribution of the mean and standard deviation of the gene counts across the samples. B) t-SNE representation.

Table 5.1: Collected data from The Cancer Genome Atlas (TCGA) databank.

Cancer type	***N***	**C**	***m***
Brain (LGG)	530	1	20,531
Breast (BRCA)	1212	1	20,531
Kidney type 1 (KIRP)	323	1	20,531
Kidney type 2 (KIRC)	606	1	20,531
Stomach type 1 (STAD)	450	1	20,531
Stomach type 2 (STES)	646	1	20,531
Uterine type 1 (UCEC)	201	1	20,531
Uterine type 2 (UCS)	57	1	20,531

Note: N: number of observations, C: number of clusters, m: number of genes.

Table 5.2: Empirical pan-cancer data set description

Cancer type	$\boldsymbol{N}$	**C**	$\boldsymbol{m}$	$\boldsymbol{d}$
Pan-cancer	200	8	20,531	17,221
- Brain (LGG)	25	1	20,531	NA
- Breast (BRCA)	25	1	20,531	NA
- Kidney type 1 (KIRP)	25	1	20,531	NA
- Kidney type 2 (KIRC)	25	1	20,531	NA
- Stomach type 1 (STAD)	25	1	20,531	NA
- Stomach type 2 (STES)	25	1	20,531	NA
- Uterine type 1 (UCEC)	25	1	20,531	NA
- Uterine type 2 (UCS)	25	1	20,531	NA

Note: N: number of observations, C: number of clusters, m: number of genes, d: number of genes after pre-processing (removing genes with $sd = 0$ and $quartile < 0.15$ (Section 5.2.1) (separate cancer types have not been pre-processed, hence denoted with NA).

5.2 Approach

The goal of the empirical analysis is to investigate which clustering method yields the highest clustering accuracy when applied to empirical pan-cancer data. The cancer omics data consists of eight equally distributed clusters, as described in Section 5.1. We furthermore aim to investigate whether choices in feature selection and latent dimensions are universal to MF and clustering techniques or if they are method-specific. We first pre-process the data, as described in Section 5.2.1. After this, we follow a similar comparative framework as Feng et al. (2020) to measure the clustering quality w.r.t. the choices of (1) the feature selection method, (2) the number of genes selected in the previous step, (3) the dimension reduction and clustering algorithm, and (4) the number of components (see Table 5.3 and Figure 5.2).

Table 5.3: Empirical analysis setup.

Step 1		*Step 2*	
Feature selection	**Number of genes**	**Clustering algorithm**	**Number of components**
IQR	100, 1,000, 3,000	PCA + K-means	5, 20, 30
SD	100, 1,000, 3,000	ICA + K-means	5, 20, 30
M	100, 1,000, 3,000	NMF + K-means	5, 20, 30
SIM	100, 1,000, 3,000	RKM	5
DIP	100, 1,000, 3,000	FKM	5
NFS	17,221		

Note: We do not compute components with RKM and FKM in combination with NFS due to computer memory limitations.

Section 5.2.2 describes the methods used for feature selection and the motivation behind it. The number of components is based on research conducted by Vidman et al. (2019) and Feng et al. (2020). As the function `cluspca` does not allow for more components than clusters, we only analyse five components for RKM and FKM. We do not evaluate compromises between FKM, RKM and PCA with the alpha-parameter because of the computation time and memory requirements. Furthermore, we do not compute components with RKM and FKM in combination with No Feature Selection (NFS) due to computer memory limitations.[2].

[2]After running the interaction of NFS with RKM and FKM on a supercomputer for a week with no results, the job was cancelled, suggesting that this combination is not feasible to include in the comparison study.

Figure 5.2: Experimental setup. First, data is collected and pre-processed as described in Section 5.2.1. Genes are selected using the feature selection methods described in Section 5.2.2 and then analysed with one of the models described in Section 3.1 and Section 3.2. For each combination, the *ARI* is computed to compare the performances of the models given the choice of feature selection method and latent dimensions. Additional analyses include t-SNE visualisations and functional annotation of the ICA components, visualised with heatmaps.

5.2.1 Data pre-processing

We first $log(x+1)$ transform the data, which is found to enhance stability and component interpretation (Sompairac et al., 2019). We remove genes with no variance across the samples and remove genes below the 15^{th} percentile (Vidman et al., 2019). We execute these steps to reduce dimensions while retaining enough information.

5.2.2 Feature selection

After pre-processing the RNA-seq data, it is common to select the N most informative genes to reduce the dimensions and to improve cluster performances (Freyhult et al., 2010). In this study, we select $m = 100, 1{,}000$, and $3{,}000$ genes either based on their variance, level, similarity, or modality. The number of genes is determined by findings of Jaskowiak et al. (2018); Vidman et al. (2019); Källberg et al. (2021); Freyhult et al. (2010). We measure these characteristics using the interquartile range and commonly used standard deviation, mean expression, level of coexpression, and the diptest respectively. We choose these methods as each method captures a different characteristic of gene expression, which results in different selected genes. We compare the selection methods to the baseline, which is NFS (including all genes).

Selection based on variance (IQR, SD)

The selection of genes based on variance can be done either using the Interquartile Range (IQR) or Standard Deviation (SD). For the selection with the IQR, we calculate the distance between Q1 and Q3 (the IQR) and select the genes with the largest distance. Selection based on SD is done by calculating the variance in the expression of each gene and selecting the genes with the highest SD (Freyhult et al., 2010; Källberg et al., 2021).

Selection based on level of expression (M)

We determine the gene expression level by calculating the Mean (M) expression value for each gene and selecting the genes with the highest level of gene expression (Källberg et al., 2021).

Selection based on similarity (SIM)

We calculate the co-expression of each gene to all other genes using the Spearman correlation (Z. Wang et al., 2014; Källberg et al., 2021). Denote $s_{ij} = |\rho_{ij}|$ as the absolute Spearman rank correlation between the expression of gene i and j. The medians of elements s_{ij}, part of similarity matrix S, are ranked to calculate the similarity score:

$$SIM_i = median_{j,j \neq i}(s_{ij}). \tag{5.1}$$

We include the m genes with the highest Similarity (SIM) scores, i.e. the highest median correlations, for further analysis.

Selection based on the dip test (DIP)

We use the Dip Test (DIP) to test for modality (Hartigan & Hartigan, 1985). The dip test measures the modality of genes, for which a smaller p-value suggests a non-unimodal distribution. The test was performed on each gene, whereafter the genes with the smallest p-values were selected. The dip test was calculated using the R package **diptest** (Maechler, 2021).

No feature selection (NFS)

RNA-seq data is commonly analysed without feature selection. Although NFS circumvents the possibility that informative genes are removed from the dataset, it could lead to the inclusion of genes that are not related to the cluster structure (Freyhult et al., 2010). As discussed before, we do not evaluate the combination of NFS with FKM and RKM due to limitations in computer memory.

5.3 Performance evaluation

5.3.1 Quality criteria

Clustering accuracy

We evaluate the performances of the models in combination with each feature selection, number of genes and dimension option using the Adjusted Rand Index (ARI), as described in 4.3.

Dimension analysis

In the dimension analysis, we use t-distributed Stochastic Neighbour Embedding (t-SNE) to visualise higher-dimensional subspaces found by the MF methods. Specifically, we analyse how the choice of the number of components affects clustering performances by visualising different choices of dimensions using t-SNE representations similar to Feng et al. (2020). Furthermore, we use t-SNE to demonstrate intuitively that the feature selection step can improve clustering performances. It should be noted that t-SNE cannot be used to identify outliers as the empty space in t-SNE maps carries no meaning (Van der Maaten & Hinton, 2008).

t-SNE is a technique that is used for the visualisation of high-dimensional data by assigning data points locations in a two or three-dimensional map (Van der Maaten & Hinton, 2008). t-SNE improves the original algorithm SNE by reducing the number of data points crowded in the centre of the map (Van der Maaten & Hinton, 2008). The algorithm starts by transforming the high-dimensional Euclidian distances between data points into similarities, given by conditional probabilities (Van der Maaten & Hinton, 2008). Following the notation of Van der Maaten & Hinton (2008), the pairwise similarity $p_{j|i}$ in the high-dimensional space of datapoint x_j to datapoint x_i is calculated as follows:

$$p_{j|i} = \frac{exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-||x_i - x_k||^2/2\sigma_i^2)}, \tag{5.2}$$

where we denote the variance of the Gaussian that is centred on data point x_i as σ_i. The joint probabilities p_{ij} in the high-dimensional space are set to symmetrised conditional probabilities, $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$, to circumvent outliers from being not well determined (Van der Maaten & Hinton, 2008). Hence, all data points will contribute significantly to the cost function that is minimised to find the solution. In t-SNE, the low-dimensional map uses the Student t-distribution with one degree of freedom as the heavy-tailed distribution (Van der Maaten & Hinton, 2008). We define the joint probabilities q_{ij} as:

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||y_k - y_l||^2)^{-1}}. \tag{5.3}$$

Lastly, t-SNE minimises the Kullback-Leibler divergence between joint probabilities p_{ij} and q_{ij} in high and low-dimensional space, respectively (Van der Maaten & Hinton, 2008):

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + ||y_i - y_j||^2)^{-1}. \tag{5.4}$$

In this research, we use R package **Rtsne** to make the t-SNE calculations (Krijthe, 2015).

Functional annotation

The ability to interpret latent dimensions can be important for researchers, for example, to discover biomarkers. To address this field of research, we perform gene enrichment analysis on the estimated subspace of ICA, explained in Section 5.3.1. We limit this analysis to the ICA components as ICA is known for its interpretability (Engreitz et al., 2010; Sompairac et al., 2019). We investigate which genes contribute the most to the components, and functionally annotate these genes using a Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis (Ashburner et al., 2000; Kanehisa & Goto, 2000). Lastly, we study the activity of the samples in the components, from which we can infer which biological processes are most associated with the tumour (sub)types.

As we recall from the Literature, ICA is used for functional annotation because it captures statistically independent signals that can be traced back to biological processes (Sompairac et al., 2019). This is important for biomarker discovery and tumour subtyping (Stein-O'Brien et al., 2018). We include this in our analysis as this application could be helpful for researchers to decide whether ICA is suitable for their research.

We can examine the biological processes related to the ICA components by investigating the two sources of information from the ICA decomposition: the source signal matrix **S** and the mixing matrix **A** (Figure 5.3).

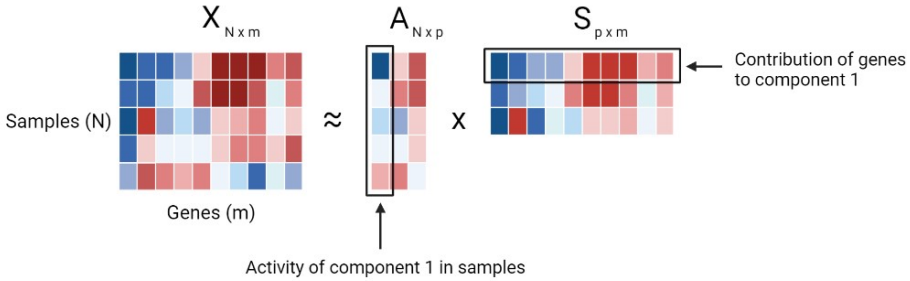

Figure 5.3: Matrix factorisation methods decompose a data matrix $\mathbf{X}_{N\times m}$ into a mixing matrix $\mathbf{A}_{N\times p}$ and a source signal matrix $\mathbf{S}_{p\times m}$.

In this research, we compute five ICA components, reduced from 3,000 genes selected with IQR. We choose to analyse five components as this is a feasible number to interpret compared to 20 or 30 components. Furthermore, we use this feature selection method and the number of genes because this combination performed well in the clustering analysis (see Section 5.4).

Annotating components based on source signal matrix S

The rows of the source signal matrix describe the genes' contributions to the ICA components. The ICA algorithm constructs normalised components, such that the components have unit variance and mean zero. Furthermore, the signs of the elements in the components are arbitrarily chosen and can be interchanged. For each component, we aim to functionally annotate the genes that contribute the most, which we call the *active genes* (Engreitz et al., 2010). We identify the active genes by setting a threshold of ± 3 standard deviations from the mean (Teschendorff et al., 2007; Biton et al., 2014; Engreitz et al., 2010). In this way, we select the sets of active genes with positive and negative loadings (Lee & Batzoglou, 2003).

Next, we functionally annotate the gene sets by performing gene enrichment analysis with GOstats using the KEGG and GO categories (Falcon & Gentleman, 2007; Kanehisa & Goto, 2000; Ashburner et al., 2000). In GOstats, we test with a hypergeometric probability if the number of selected genes associated with the biological process is more than expected. This test is called the hypergeometric test for over-representation, which is the one-tailed variant of Fisher's exact test (Falcon & Gentleman, 2007). We set the p-value of the GO and KEGG analyses to 0.001 and 0.05, respectively. We select terms that have a $CategorySize > 10$, i.e. terms that have more than ten genes annotated to them. We choose these parameters as this results in a high enough threshold that selects a small group of GO and KEGG terms that we can associate with the components. We report all annotated biological processes for each component for the positive and negative loadings, along with the p-value. We cannot correct for multiple testing, as there is implicit interdependence between parent-child GO terms (Alexa et al., 2006).

Analysing components based on mixing matrix A

The mixing matrix, i.e. the loading matrix, contains information on how much the components are active in the samples. If tumour samples show different component activities, they also have different associated gene expression patterns (Engreitz et al., 2010; Biton et al., 2014). Similar to Biton et al. (2014), we group tumour samples based on activity and characterise these tumour (sub)types. Next, we link this tumour characteristic to the associated active genes of the components. We extend this annotation to the associated biological processes annotated with GOstats in the previous section.

5.3.2 Statistical tests

We provide statistical validity to the empirical analysis using the Kruskal-Wallis test to test whether more than two groups are significantly different. When this is true, the Wilcoxon signed-rank test is used for pair-wise comparisons, corrected for multiple testing using the Bonferroni method.

5.4 Results

In the empirical analysis, we aim to analyse whether joint methods outperform tandem techniques in clustering empirical cancer omics data, and how their performance depends on feature selection and latent dimension options. In Section 5.4.1, we will compare the clustering accuracy corresponding to each method and in Section 5.4.2 we will perform an intuitive dimension analysis. We furthermore extend this research with a functional annotation analysis in Section 5.4.3.

5.4.1 Clustering accuracy

As we recall from Section 5.2, the goal of the empirical analysis is to investigate which clustering method yields the highest clustering accuracy when applied to empirical pan-cancer data. We furthermore aim to investigate whether choices in feature selection and latent dimensions are

universal to MF and clustering techniques or if they are method-specific.

We start by looking at the top-performing results depicted in Table 5.4. We observe that some tandem specification runs outperform the joint specification results. The top-three clustering accuracies are obtained by NMF ($ARI = 0.60, 0.59, 0.59$), with the fourth and fifth places achieved by ICA ($ARI = 0.59, 0.58$). The full ranking of the results is depicted in Table B.1-B.3.

Table 5.4: Model specification top-5 clustering results.

Ranking	**Feature Selection**	**N. Genes**	**Method**	**N. Components**	**ARI**
1	IQR	1,000	NMF	5	0.60
2	SIM	3,000	NMF	30	0.59
3	M	1,000	NMF	5	0.59
4	IQR	1,000	ICA	5	0.59
5	SIM	3,000	ICA	30	0.58

Although this ranking could give an answer to the research question, it would not provide a robust overview of which methods to use on which occasion. We seek to find method specifications with the highest chance of performing well, which can be inferred more robustly from the distributions of the results and comparing medians instead from the maxima. In Figure 5.4, all aggregated results are depicted. Looking at Figure 5.4C, we see that RKM has the highest clustering accuracy, with a median of $ARI = 0.4$. FKM has the worst performance (median: $ARI = 0.1$). Comparing the tandem results, we see that PCA has the worst performance, while ICA and NMF have similar performances.

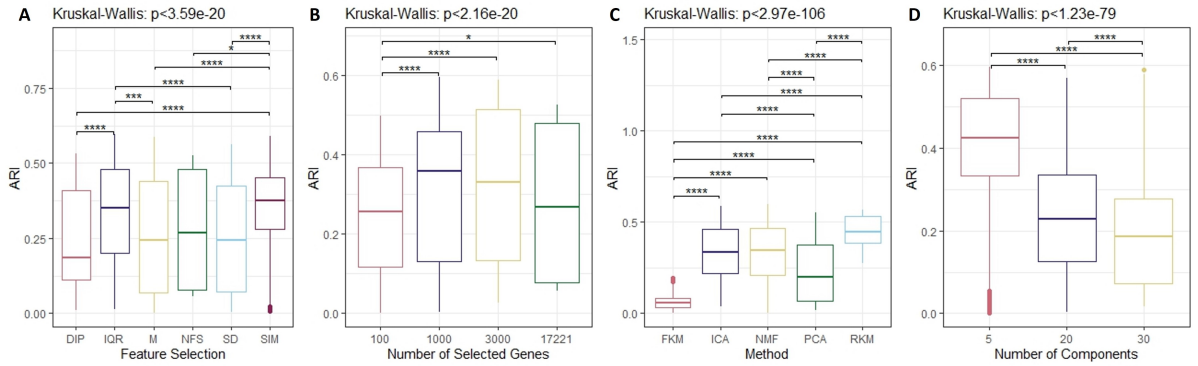

Figure 5.4: Aggregated results of the empirical analysis. A) Results grouped by feature selection, B) Results grouped by number of selected genes, C) Results grouped by clustering algorithm and D) Results grouped by the number of computed components. Pairwise comparisons are computed with the Wilcoxon signed rank test (Section 4.3.2), non-significant pairs are not shown.

Next, we investigate the specific effects of feature selection and MF options and how they interact with the clustering algorithms. We plotted the results for each decision in the process grouped by the clustering algorithms in Figure 5.5.

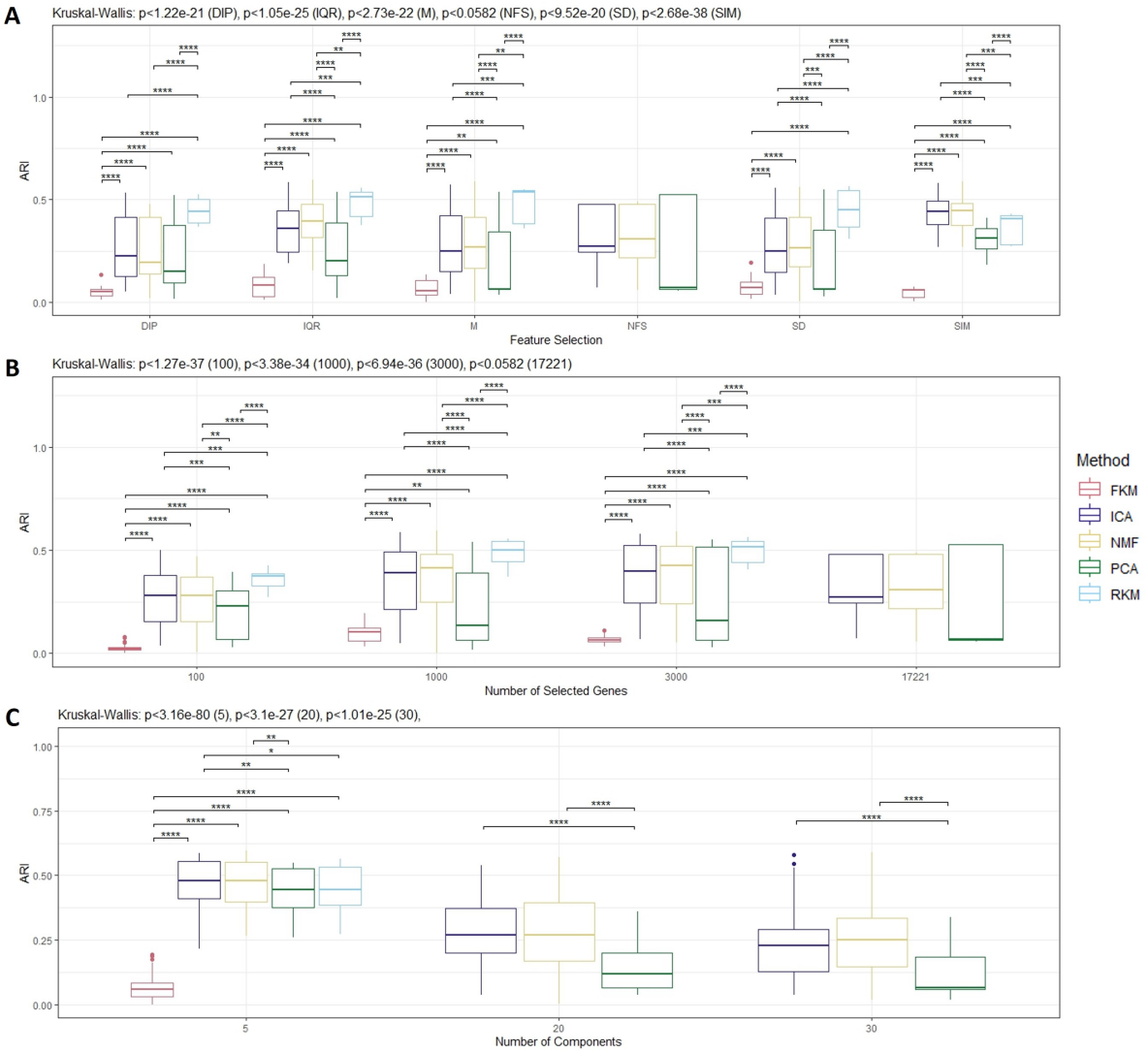

Figure 5.5: Results of the empirical analysis, grouped by method. A) Results grouped by feature selection, B) Results grouped by number of selected genes, and D) Results grouped by the number of computed components. Pairwise comparisons are computed with the Wilcoxon signed rank test (Section 4.3.2), non-significant pairs are not shown.

Starting with the feature selection choice, we see in Figure 5.4A that IQR and SIM have the highest median clustering accuracy. Figure 5.5A shows that IQR interacts well with RKM, while SIM interacts well with ICA and NMF. We further investigated the difference between IQR and SIM in Figure 5.6, and we observe that IQR and SIM select a small percentage of genes similarly: only 0%, 2% and 13% for 100, 1,000 and 3,000 genes, respectively. This tells us that different sets of genes can equally contribute to high clustering accuracy. We further see in Figure 5.4A that feature selection methods M and SD perform equally well as compared to NFS.

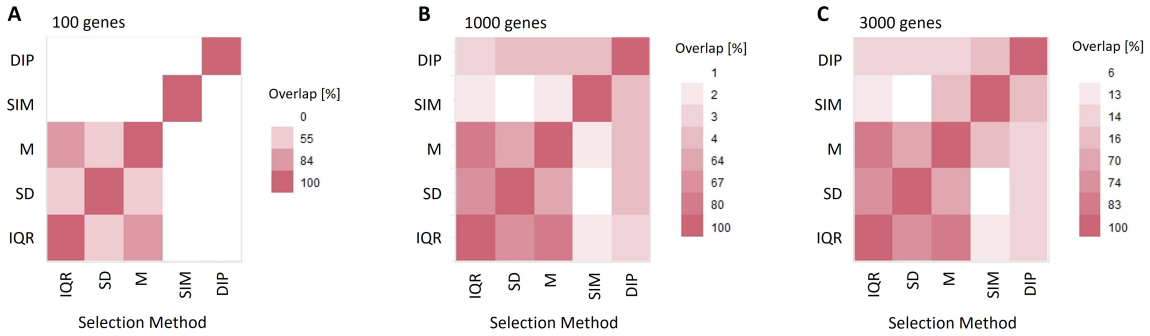

Figure 5.6: Heatmap visualisation of the fraction of overlap in the number of selected genes, for 100 genes (A), 1,000 genes (B), and 3,000 genes (C).

We see in Figure 5.4B that selecting all genes (17,221 genes), 3,000 and 1,000 genes have better performances than reducing the number of genes to 100. This could be explained by the fact that there are non-informative genes present in the dataset that are removed using the feature selection methods. However, if there are too many genes removed, information is lost, resulting in worse performances. We observe in Figure 5.5B that the clustering algorithms interact in a similar fashion to 1,000 or 3,000 genes: RKM performs best, ICA and NMF have slightly worse performances, PCA has a large variance in the result with a low median clustering accuracy ($ARI \simeq 0.2$) and FKM has the lowest accuracy.

In Figure 5.4D, we see that the best number of components is five. When grouping the results by method, as done in Figure 5.5C, we see that for five components, RKM does not have the highest median clustering accuracy, but ICA and NMF do. Further investigation of the results with only five components shows that SIM has a better performance than IQR and that selecting 3,000 genes is a good choice (Figure B.1).

5.4.2 Dimension analysis

We can use t-SNE visualisations to intuitively assess whether MF in the tandem approach is effective after applying feature selection. As an example, we select 3,000 genes with IQR. We compute the t-SNE representation of this dataset, and we apply MF to the dataset with ICA, constructing five and 30 components. We then compute the t-SNE maps of these five and 30 components and compare these maps to the original map that did not include MF (Figure 5.7).

From Figure 5.7, we see that the local cluster structure has less overlap after MF. This suggests that only applying feature selection is not sufficient when clustering is the goal of the study. However, researchers should investigate multiple component choices as Figure 5.7 also shows that 30 ICA components did not improve local cluster separation. This could be a result of overfitting and can be addressed by performing a cross-validation analysis or using a selection heuristic for the number of components.

When we analyse the subspace captured by the joint methods on the same 3,000 genes selected with IQR, we see that while FKM can define separated clusters, RKM has many overlapping local cluster structures (Figure 5.8). This contradicts the fact that RKM has higher clustering accuracies than FKM, and suggests that one should be careful using t-SNE plots as an indicator

for clustering accuracy. All t-SNE visualisations of the clustering algorithms and their latent dimension options are shown in Figure B.2.

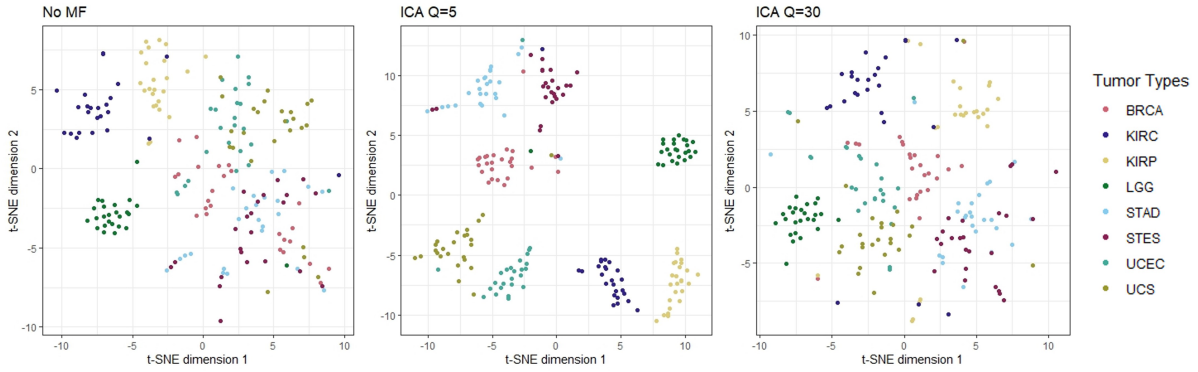

Figure 5.7: t-SNE visualisation of the IQR 3,000 genes data with no matrix factorisation and with ICA matrix factorisation to 5 and 30 components.

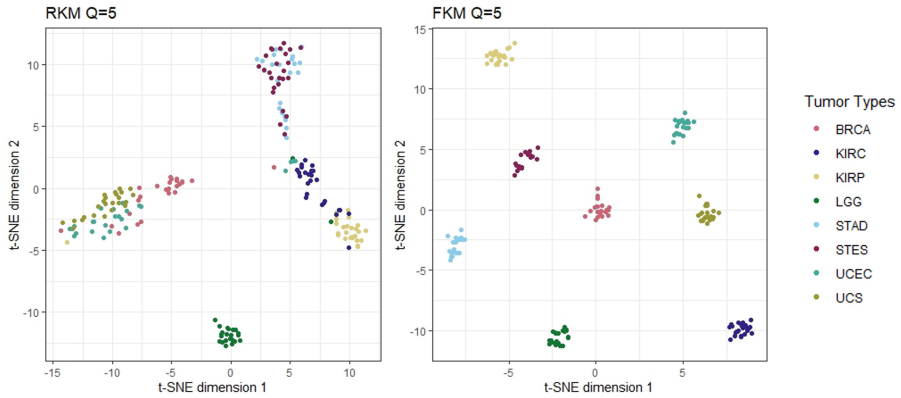

Figure 5.8: t-SNE visualisation of the IQR 3,000 genes data when RKM and FKM are applied to construct five components.

5.4.3 Functional annotation

Another important aspect of clustering empirical cancer omics data is the interpretation of the constructed components by the MF techniques (Stein-O'Brien et al., 2018; Sompairac et al., 2019) (see Section 5.3.1). Hence, we aim to examine the biological processes related to the ICA components by investigating the two sources of information from the ICA decomposition: the source signal matrix $\mathbf{S}$ and the mixing matrix $\mathbf{A}$ (Figure 5.3).

We observe that the ICA algorithm creates a subspace in the source signal matrix that can be easily functionally annotated, distinguishing multiple important cellular activities. In Table 5.5, we summarised the biological processes and metabolic pathways assigned to the components. The result shows that the independence criterion in the ICA algorithm can separate biological signals in the pan-cancer dataset. For a detailed overview of all GO and KEGG annotations, see Tables B in the Appendix.

Table 5.5: Summary of the functional annotation of the ICA components. Note: signs in ICA components are arbitrary.

	GO/KEGG Annotation	**Positive Activity**	**Negative Activity**
C1	Response to protein signal	STES	KIRP
C2	Muscle contraction	UCEC	LGG
C3	Immune response	-	STAD
C4	Protein formation	UCS	KIRP
C5	Metabolic activity	KIRP	BRCA

Figure 5.9A shows that the tumours have different component activities in the mixing matrix, meaning that the tumours have different gene expression patterns. By grouping the clusters based on similar component activity, we see that tumour subtypes have similar gene expression patterns. This shows that the biological theory, i.e. that tumour subtypes are related to the same cancer type, translates into the signals captured with the ICA algorithm. We can also observe that subtypes UCS and UCEC are not similar to the rest of the tumour clusters.
When we look into the activities within the components, we see that UCS shows opposing activity with KIRP in the signals that are annotated as *Protein formation* (Figure 5.9A). Contrarily, we see no opposing activity in component three, which is defined as *Immune response*, which could mean that this biological process is active in all tumour types.

Furthermore, we see in Figure 5.9B that there is heterogeneity in the activity of the components when comparing samples within a tumour type. This means that even though samples are labelled as the same tumour subtype, they could have different gene expression patterns. This could indicate for example that there might be more molecular subtypes within these specific tumour types. Another explanation could be that the samples originate from individuals with different clinicopathological features such as gender or age and hence might have different gene expression patterns.

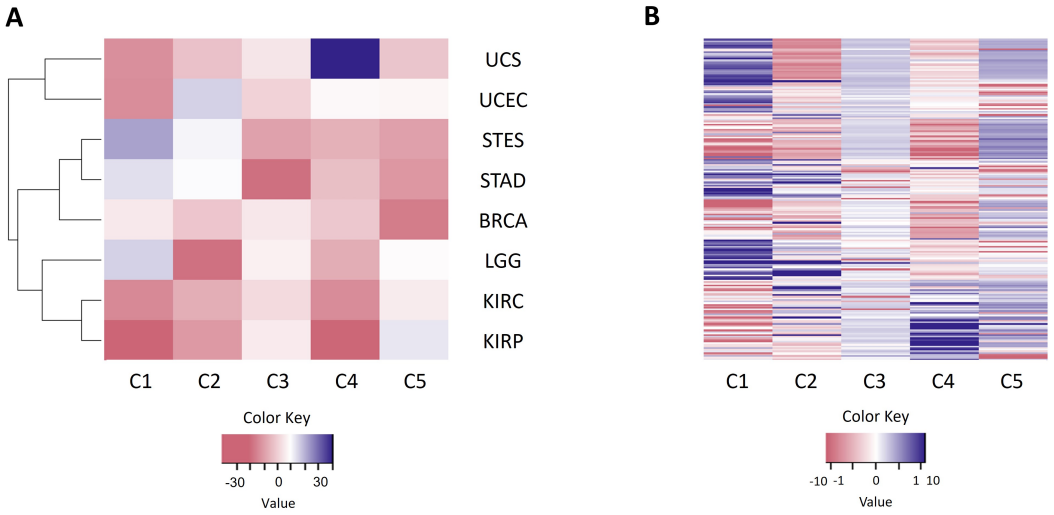

Figure 5.9: Heatmaps of the activity of the clusters (A) and samples (B) in the ICA components.

Chapter 6

Discussion

Cancer omics data is high-dimensional and difficult to interpret. MF techniques such as PCA, ICA or NMF are applied to reduce dimensions, whereafter clustering algorithms are performed to discover biomarkers or new tumour subtypes. This traditional tandem approach is found to be suboptimal because the MF and clustering algorithms do not have the same optimisation criteria. Consequently, the cluster structure can be lost in the latent dimensions that are not captured by MF algorithms. Joint MF and clustering algorithms such as RKM and FKM address this limitation by integrating the optimisation criteria of reducing dimensions and clustering.

We investigated whether RKM and FKM outperformed PCA, ICA and NMF with K-means in recovering the cluster subspace and identifying the cluster memberships applied to cancer omics data. We addressed our main research question, *"Do joint MF and clustering algorithms outperform benchmark tandem techniques in preserving cluster structure in (simulated) cancer omics data?"*, with a simulation study and an empirical analysis. In the simulation study, we analysed the methods in the environment of different structures of noise and measured the interactions between data characteristics and the clustering algorithms. In the empirical analysis, we compared multiple feature selection and component options, investigating which specifications yield the highest clustering accuracies on empirical pan-cancer RNA-seq data. In this section, we recall the research subquestions and report the conclusions. We reflect on our hypotheses, and state limitations concerning our research. We will also provide suggestions for future research and give our recommendations.

6.1 Concluding remarks

6.1.1 Simulation study

In the simulation study, we investigated how different structures of noise affected the performances of the methods. We recall our first research subquestion: *"How does the performance of joint and tandem methods depend on the level of random noise, subspace noise and masking variables?"*.

We observed that the clustering accuracy (measured with the ARI) of all methods decreased

when we increased the level of random noise, which was in line with our expectations. When we increased the subspace residual variance, the ARI increased. However, we saw that subspace recovery was less affected by the random and subspace noise, as we observed a small decrease in subspace recovery. This can be explained by the fact that the noise is added after the projection of the cluster structure by the loading matrix. The projection could have a large enough effect on the sign and magnitude of the elements in the observed data matrix that the cluster subspace is relatively easy to recover.

Comparing the performances of the algorithms led to the same ranking in the case of random noise and subspace residuals. RKM yielded the highest clustering accuracies, which was in line with our expectations. ICA, NMF and PCA had similar performances. We first conjectured that ICA might separate the signal of masking variables from the signal variables, resulting in high ARI-values, but we saw that ICA only outperformed RKM in the case of no masking variables. Interestingly, FKM yielded the lowest clustering accuracies, which we did not expect for large complement residual fractions based on the study by Timmerman et al. (2010). For both residual structures, NMF had superior results in recovering the subspace. This was as expected as NMF estimates non-negative loading matrices, corresponding to the generated non-negative loading matrices.

When we investigated the partial effects of the parameters on the ARI and Phi with the RMANOVA test, we saw that the factor $Method$ had a large effect, but that it was not significant. Additionally, for the random and subspace noise, $Method \times Noise$ and $Method \times PSR$ had a large significant effect, respectively. This is in line with Timmerman et al. (2010) that showed large partial effects of subspace variance, but in contrast to their found large effect of cluster overlap which is similar to centroid distance.

The second research subquestion was: *"Which joint, compromise, or tandem approach is the most suitable in the presence of random noise, subspace noise and masking variables?"*.

We saw that for random noise, compromises between FKM and RKM performed best. When the proportion of subspace residuals was smaller than 9%, RKM performed the best. ICA performed the best when the PSR crossed the 0.09 mark, which corresponds to the previous experiment in the case of no masking variables. When we analysed the effect of masking variables, we saw that the effect on the cluster membership identification was the opposite of on the subspace recovery. Namely, when the fraction of masking variables was increased, the optimal alpha decreased, but for the subspace recovery, the optimal alpha increased. Nonetheless, NMF had the highest performance in subspace recovery.

6.1.2 Empirical analysis

For the empirical analysis, we recall our first research subquestion: *"Do joint methods outperform tandem techniques in clustering empirical cancer omics data, and how does their performance depend on feature selection and latent dimension options?"*.

We observed that similar to the simulation study, RKM and FKM had the best and worst

performance, respectively. Our expectations were that the non-Gaussian nature of ICA and the non-negative nature of NMF were suitable for omics data. This was confirmed: from the tandem methods, ICA and NMF had similar performances and PCA performed the worst. Even though PCA is a widely-used method to compress dimensions before clustering, these results indicate that this is not a suitable method for RNA-seq data.

The best feature selection option was IQR and SIM. DIP performed the worst, which was in contrast to the findings of Källberg et al. (2021) who found DIP to be the top-performing feature selection method. Although we thought that the feature selection options would yield similar clustering accuracies with the clustering algorithms, we saw that RKM performed well with the IQR selection, while for the SIM approach, ICA and NMF performed well. From the number of selected genes, we deducted that one should not select 100 genes, but should rather select 1,000 genes. When grouping on 1,000 and 3,000 genes, we saw that RKM performed the best. Five components performed significantly better than 20 and 30, suggesting that more than five components resulted in overfitting (Stein-O'Brien et al., 2018). Because five components yielded significantly better results, we filtered the data to five components. We saw that ICA and NMF performed better than RKM and we found that selecting 3,000 genes with SIM had a high performance as well.

Our second research subquestion was: *"Can we interpret the signal and loading matrix of the ICA components that are computed from the pan-cancer dataset?"*.

We were able to interpret the ICA signal and mixing matrices, distinguishing biological processes and metabolic pathways and we could annotate these to the components, similar to Engreitz et al. (2010); Biton et al. (2014). We saw differential tumoral activity across the samples in the components, suggesting that the cancer types have different gene expression patterns, corresponding to the biological theory. This result could be a reason that researchers may prefer using ICA in analysing and clustering cancer omics data.

6.2 Limitations

Several aspects of our study leave room for improvement. Firstly, we remark that in the simulation study, the performances of the clustering algorithms depend on the parameter settings and the data generation model. For example, there was a bias towards NMF, as the loading matrix was generated to be non-negative. Also, as a consequence, the elements corresponding to the complement residuals were altered, which may have had a particular effect on FKM because that is a method suitable for data with complement residuals (Timmerman et al., 2010). Thus, future research could investigate other parameter settings and loading matrix generation techniques.

Secondly, a limitation of the empirical study is that the results cannot be directly translated to other RNA-seq datasets, as each RNA-seq dataset is different (Källberg et al., 2021). An improvement would be to test the same methods, feature selection methods and dimension options on multiple empirical pan-cancer RNA-seq datasets to validate the results.

Thirdly, we did not know the clinicopathological background of the tumoural samples. Although we limited this bias by randomly sampling the observations from the collected datasets, it would be beneficial to have incorporated such information. Vidman et al. (2019) found that class balance is essential for accurate clustering, so future research could give more insight into whether class (in)balance affects the clustering performance of similar pan-cancer empirical RNA-seq datasets as well.

Lastly, there are more preprocessing steps, feature selection methods, MF and clustering algorithms and distance metrics available that can be tested and compared on cancer omics data. We did not consider untransformed data or other gene types such as isoforms or exons (Jaskowiak et al., 2018). We did not test feature selection methods based on the third quartile, entropy or (non-)parametric bimodality indices (Källberg et al., 2021). Clustering algorithms such as k-medoids or variants on hierarchical clustering could be tested as well (Jaskowiak et al., 2018). Moreover, we used the classical Euclidian distance metric in K-means, but there are more distance measures that could be suitable for sequencing data. These include for example other classical distances like Manhattan or Supreme distance, or coefficients such as the Pearson or Jackknife correlation coefficients (Jaskowiak et al., 2018).

6.3 Future research

Considering the existing literature and the conducted study, we have identified three fields of research that play a crucial role in improving the use of clustering techniques on cancer data.

We recognise that model selection is a prevalent problem. In empirical practice, one has to choose the MF and clustering algorithm and the desired number of components and clusters. For the MF and clustering algorithm, one could compare methods or make an educational guess based on the statistics of the data at hand (Timmerman et al., 2010). For the number of components and clusters, the general approach is to use quality criteria to determine which number of components and clusters is optimal. Commonly used methods for estimating the optimal number of components are Cattell's scree test for PCA (Cattell, 1966), the Bayesian information criterion (BIC) for ICA (Schwarz, 1978), and the cophenetic correlation coefficient (CCC) for NMF (Brunet et al., 2004). These quality criteria are not perfect, hence newer methods have been proposed such as the Velicer's Minimum Average Partial (MAP) for PCA (Velicer, 1976), or the Maximally Stable Transcriptome Dimension (MSTD) for ICA (Kairov et al., 2017). In the case of estimating the optimal number of clusters, traditional approaches include the elbow method (Thorndike, 1953), using information criteria such as AIC or BIC (Schwarz, 1978), or cross-validation. Other popular heuristics include the Silhouette index (Kaufman & Rousseeuw, 1990), or the Calinski-Harabasz index (Calinski & Harabasz, 1974). These methods have not been extensively compared on cancer omics data, hence indicating a future field of research.

Additionally, we see that extending MF and clustering algorithms to multimodal and temporal analysis is the next step in cancer research. The sizes of cancer omics data continue to grow, and with the increase in information, it is necessary to develop techniques that can capture data from multiple sources. Examples of multimodal learning include tensor decompositions (Hore et al., 2016; Durham et al., 2018; Zhu et al., 2016; M. Wang et al., 2019). An example of a

temporal-based clustering is "temp-ICA" (Fonseca et al., 2017). Future studies could investigate whether this could be extended to other MF techniques as well.

The promising results of joint algorithm RKM, which is based on PCA and K-means, suggest that joint algorithms based on NMF and ICA could yield high clustering accuracies as well. Some algorithms are proposed, for example, "DRjCC", a joint dimension reduction and clustering algorithm based on NMF for the analysis of single-cell RNA-seq data (Wu & Ma, 2020). These advances are important for developments in joint algorithms for cancer omics data. Nonetheless, we also acknowledge that joint MF and clustering algorithms require more computation power, memory and time. Cancer omics data is high-dimensional, and even if joint algorithms outperform tandem techniques, the computation time involved could discourage researchers from using it. Hence, future research could investigate how to include approximations in such algorithms, similar to the FastICA algorithm (Hyvärinen & Oja, 2000).

6.4 Recommendations

Concluding, we recommend using RKM and/or ICA with K-means. These methods yielded high clustering accuracies in the simulation and empirical analysis. However, we do not recommend using RKM on large datasets, that is, bigger than 3,000 genes. The required computational memory impedes the use of joint MF and clustering algorithms in combination with larger datasets. Moreover, NMF had results close to RKM and ICA, with particularly good subspace recovery results in the simulation study. Hence, NMF can be considered as well. FKM should be avoided in all situations, as it could barely recover any clusters and could not approximate the cluster subspace. Our functional annotation analysis underlines the findings that ICA is suitable for inferring molecular and sample relationships.

We furthermore recommend combining the feature selection method IQR with RKM and SIM with ICA and K-means. We discourage selecting less than 1,000 genes, as this excluded informative genes. Lastly, we recommend selecting components for a relatively low dimensionality rather than a large dimensionality, as selecting five components outperformed either twenty or thirty components.

References

Alexa, A., Rahnenführer, J. & Lengauer, T. (2006, 7). Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, *22*, 1600-1607. doi: 10.1093/bioinformatics/btl140

Arabie, P. & Hubert, L. (1996). Advances in cluster analysis relevant to marketing research. In W. Gaul & D. Pfeifer (Eds.), (p. 3-19). Springer Berlin Heidelberg.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, *25*, 25-29. doi: 10.1038/75556

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*, 379-384. doi: 10.3758/BF03192707

Biton, A., Bernard-Pierrot, I., Lou, Y., Krucker, C., Chapeaublanc, E., Rubio-Pérez, C., ... Radvanyi, F. (2014). Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. *Cell Reports*, *9*(4), 1235-1245. doi: https://doi.org/10.1016/j.celrep.2014.10.035

Brunet, J., Tamayo, P., Golub, T. R. & Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, *101*(12), 4164-4169. doi: 10.1073/pnas.0308531101

Calinski, T. & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *3*(1), 1–27.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245-276. doi: 10.1207/s15327906mbr0102_10

De Soete, G. & Carroll, J. D. (1994). *K-means clustering in a low-dimensional euclidean space.*

Durham, T. J., Libbrecht, M. W., Howbert, J. J., Bilmes, J. & Noble, W. S. (2018). Predictd parallel epigenomics data imputation with cloud-based tensor decomposition. *Nature communications*, *9*(1), 1402.

Engreitz, J. M., Daigle, B. J., Marshall, J. J. & Altman, R. B. (2010, 12). Independent component analysis: Mining microarray data for fundamental human gene expression modules. *Journal of Biomedical Informatics*, *43*, 932-944. doi: 10.1016/j.jbi.2010.07.001

Falcon, S. & Gentleman, R. (2007, 1). Using gostats to test gene lists for go term association. *Bioinformatics*, *23*, 257-258. doi: 10.1093/bioinformatics/btl567

Feng, C., Liu, S., Zhang, H., Guan, R., Li, D., Zhou, F., ... Feng, X. (2020, 3). Dimension reduction and clustering models for single-cell rna sequencing data: A comparative study. *International Journal of Molecular Sciences*, *21*. doi: 10.3390/ijms21062181

Fonseca, F., Thelma, S., Campana, N. A. C., Maciel, F. T. E., Azevedo, B. L. M., Camila, F. A., ... Silva (2017, 4). Independent component analysis (ica) based-clustering of temporal rna-seq data. *PLOS ONE*, *12*, 1-12. doi: 10.1371/journal.pone.0181195

Freyhult, E., Landfors, M., Önskog, J., Hvidsten, T. R. & Rydén, P. (2010, 10). Challenges in microarray class discovery: A comprehensive examination of normalization, gene selection and clustering. *BMC Bioinformatics*, *11*. doi: 10.1186/1471-2105-11-503

Gaujoux, R. & Seoighe, C. (2010). A flexible r package for nonnegative matrix factorization. *BMC bioinformatics*, *11*(1), 1–9.

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A. & Staudt, L. M. (2016). Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, *375*(12), 1109-12. doi: 10.1056/NEJMp1607591

Hartigan, J. A. & Hartigan, P. M. (1985). The dip test of unimodality. *The Annals of Statistics*, *13*(1), 70 – 84. doi: 10.1214/aos/1176346577

Hartigan, J. A. & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108.

Herault, J. & Jutten, C. (1986). Space or time adaptive signal processing by neural network models. *AIP Conference Proceedings*, *151*(1), 206-211. doi: 10.1063/1.36258

Hinneburg, A. & Keim, D. A. (1999). Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. In *Proceedings of the 25th international conference on very large databases, 1999* (pp. 506–517).

Hore, V., Vinuela, A., Buil, A., Knight, J., McCarthy, M. I., Small, K. & Marchini, J. (2016). Tensor decomposition for multiple-tissue gene expression experiments. *Nature genetics*, *48*(9), 1094–1100.

Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193-218.

Hyvärinen, A. & Oja, E. (2000). Independent component analysis: algorithms and applications.

Iodice D'Enza, A., Van de Velden, M. & Palumbo, F. (2014). On joint dimension reduction and clustering of categorical data. *Studies in Classification, Data Analysis, and Knowledge Organization*, *49*, 161-169. doi: 10.1007/978-3-319-06692-9_18

Jaskowiak, P. A., Costa, I. G. & Campello, R. J. G. B. (2018, 1). Clustering of rna-seq samples: Comparison study on cancer data. *Methods*, *132*, 42-49. doi: 10.1016/j.ymeth.2017.07.023

Jiang, R., Sun, T., Song, D. & Li, J. J. (2022). Statistics or biology: the zero-inflation controversy about scrna-seq data. *Genome Biology*, *23*, 31. doi: 10.1186/s13059-022-02601-5

Jolliffe, I. (2002). Principal component analysis. In *Wiley statsref: Statistics reference online.* doi: 10.1002/9781118445112.stat06472

Kairov, U., Cantini, L., Greco, A., Molkenov, A., Czerwinska, U., Barillot, E. & Zinovyev, A. (2017, 9). Determining the optimal number of independent components for reproducible transcriptomic data analysis. *BMC Genomics*, *18*. doi: 10.1186/s12864-017-4112-9

Kamat, A. M. & Matulay, J. T. (2018). Advances in risk stratification of bladder cancer to guide personalized medicine. *F1000Research*, *7*. doi: 10.12688/f1000research.14903.1

Kanehisa, M. & Goto, S. (2000). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*, 27-30.

Kassambara, A. (2023). rstatix: Pipe-friendly framework for basic statistical tests [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=rstatix` (R package version 0.7.2)

Kaufman, L. & Rousseeuw, P. (1990). *Finding groups in data: An introduction to cluster analysis.* doi: 10.2307/2532178

Krijthe, J. H. (2015). Rtsne: T-distributed stochastic neighbor embedding using barnes-hut implementation [Computer software manual]. (R package version 0.16)

Kuhn, H. W. & Tucker, A. W. (1951). Nonlinear programming. In *Proceedings of the second berkeley symposium on mathematical statistics and probability* (pp. 481–492). Berkeley, CA, USA: University of California Press.

Källberg, D., Vidman, L. & Rydén, P. (2021, 2). Comparison of methods for feature selection in clustering of high-dimensional rna-sequencing data to identify cancer subtypes. *Frontiers in Genetics*, *12*. doi: 10.3389/fgene.2021.632620

Lee, S.-I. & Batzoglou, S. (2003). Application of independent component analysis to microarrays. *Genome Biology*, *4*, R76. Retrieved from `https://doi.org/10.1186/gb-2003-4-11-r76` doi: 10.1186/gb-2003-4-11-r76

Liu, S. (2022). Nmfn: Non-negative matrix factorization [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=NMFN` (R package version 2.0.1)

Liu, W., Liao, X., Yang, Y., Lin, H., Yeong, J., Zhou, X., ... Liu, J. (2022, 7). Joint dimension reduction and clustering analysis of single-cell rna-seq and spatial transcriptomics data. *Nucleic Acids Research*, *50*. doi: 10.1093/nar/gkac219

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. of the fifth berkeley symposium on mathematical statistics and probability* (Vol. 1, p. 281-297). University of California Press.

Maechler, M. (2021). diptest: Hartigan's dip test statistic for unimodality - corrected [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=diptest` (R package version 0.76-0)

Marchini, J. L., Heaton, C. & Ripley, B. D. (2021). fastica: Fastica algorithms to perform ica and projection pursuit [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=fastICA` (R package version 1.2-3)

Markos, A., D'enza, A. I. & van de Velden, M. (2019). Beyond tandem analysis: Joint dimension reduction and clustering in r. *Journal of Statistical Software*, *91*. doi: 10.18637/jss.v091.i10

Markos, A., Iodice D'Enza, A. & van de Velden, M. (2019). Beyond tandem analysis: Joint dimension reduction and clustering in R. *Journal of Statistical Software*, *91*(10), 1–24. doi: 10.18637/jss.v091.i10

Mesters, G. & Zwiernik, P. (2022). Non-independent components analysis. *ArXiv preprint*.

Miles, J. & Shevlin, M. (2001). *Applying regression and correlation: A guide for students and researchers*. Sage.

Paatero, P. & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, *5*(2), 111-126. doi: https://doi.org/10.1002/env.3170050203

R-CoreTeam. (2023). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461 – 464. doi: 10.1214/aos/1176344136

Sompairac, N., Nazarov, P. V., Czerwinska, U., Cantini, L., Biton, A., Molkenov, A., ... Zinovyev, A. (2019, 9). Independent component analysis for unraveling the complexity of cancer omics datasets. *International Journal of Molecular Sciences*, *20*. doi: 10.3390/ijms20184414

Stein-O'Brien, G. L., Arora, R., Culhane, A. C., Favorov, A. V., Garmire, L. X., Greene, C. S., ... Fertig, E. J. (2018). Enter the matrix: Factorization uncovers knowledge from omics. *Trends in Genetics*, *34*(10), 790-805. doi: 10.1016/j.tig.2018.07.003

Teschendorff, A. E., Miremadi, A., Pinder, S. E., Ellis, I. O. & Caldas, C. (2007, 8). An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. *Genome Biology*, *8*. doi: 10.1186/gb-2007-8-8-r157

Tharwat, A. (2018). Independent component analysis: An introduction. *Applied Computing and Informatics*, *17*, 222-249. doi: 10.1016/j.aci.2018.08.006

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, *18*(4), 267–276.

Timmerman, M. E., Ceulemans, E., Kiers, H. A. L. & Vichi, M. (2010). Factorial and reduced k-means reconsidered. *Computational Statistics & Data Analysis*, *54*(7), 1858-1871. doi: 10.1016/j.csda.2010.02.009

Tomašev, N., Radovanović, M., Mladenić, D. & Ivanović, M. (2011). The role of hubness in clustering high-dimensional data. In J. Z. Huang, L. Cao & J. Srivastava (Eds.), (p. 183-195). Springer Berlin Heidelberg.

Tsimberidou, A. M., Fountzilas, E., Nikanjam, M. & Kurzrock, R. (2020, 6). Review of precision cancer medicine: Evolution of the treatment paradigm. *Cancer Treatment Reviews*, *86*. doi: 10.1016/j.ctrv.2020.102019

Van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*, 2579-2605.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*, 321–327.

Vichi, M. & Kiers, H. A. L. (2001). Factorial k-means analysis for two-way data. *Computational Statistics and Data Analysis*, *37*(1), 49–64. doi: 10.1016/S0167-9473(00)00064-5

Vichi, M., Vicari, D. & Kiers, H. A. L. (2019). Clustering and dimension reduction for mixed variables. *Behaviormetrika*, *46*, 243-269. doi: 10.1007/s41237-018-0068-6

Vidman, L., Källberg, D. & Rydén, P. (2019, 12). Cluster analysis on high dimensional rna-seq data with applications to cancer research - an evaluation study. *PLoS ONE*, *14*. doi: 10.1371/journal.pone.0219102

Wang, M., Fischer, J. & Song, Y. S. (2019). Three-way clustering of multi-tissue multi-individual gene expression data using semi-nonnegative tensor decomposition. *The annals of applied statistics*, *13*(2), 1103.

Wang, Z., Lucas, F. A. S., Qiu, P. & Liu, Y. (2014). Improving the sensitivity of sample clustering by leveraging gene co-expression networks in variable selection. *BMC Bioinformatics*, *15*, 153. doi: 10.1186/1471-2105-15-153

Wojczynski, M. K. & Tiwari, H. K. (2008). Definition of phenotype. In *Genetic dissection of complex traits* (Vol. 60, p. 75-105). Academic Press. doi: https://doi.org/10.1016/S0065-2660(07)00404-X

Wu, W. & Ma, X. (2020, 3). Joint learning dimension reduction and clustering of single-cell rna-sequencing data. *Bioinformatics*, *36*, 3825-3832. doi: 10.1093/bioinformatics/btaa231

Yamamoto, M. & Hwang, H. (2014). A general formulation of cluster analysis with dimension reduction and subspace separation. *Behaviormetrika*, *41*, 115-129.

Yu, X., Abbas-Aghababazadeh, F., Chen, Y. A. & Fridley, B. L. (2021). *Statistical and bioinformatics analysis of data from bulk and single-cell rna sequencing experiments* (Vol. 2194). Humana Press Inc. doi: 10.1007/978-1-0716-0849-4_9

Zhu, Y., Chen, Z., Zhang, K., Wang, M., Medovoy, D., Whitaker, J. W., ... Wang, W. (2016). Constructing 3d interaction maps from 1d epigenomes. *Nature communications*, *7*(1), 10812.

Figures 4.1, 4.3, 4.4, 5.2 and 5.3 were created with Biorender.com.

Acronyms

ARI Adjusted Rand Index.

BRCA Breast Invasive Carcinoma.

DIP Dip Test.

FKM Factorial K-means.

GO Gene Ontology.

GRC Generalised Reduced Clustering.

ICA Independent Component Analysis.

IQR Interquartile Range.

KEGG Kyoto Encyclopedia of Genes and Genomes.

KIRC Kidney Renal Clear Cell Carcinoma.

KIRP Kidney Renal Papillary Cell Carcinoma.

LGG Lower Grade Glioma.

M Mean.

MF Matrix Factorisation.

NFS No Feature Selection.

NMF Non-Negative Matrix Factorisation.

PCA Principal Component Analysis.

PF Permutation Fraction.

PSR Proportion of Subspace Residual variance.

RKM Reduced K-means.

RMANOVA Repeated Measures ANalysis Of VAriance.

RNA-seq RNA sequencing.

SD Standard Deviation.

SIM Similarity.

STAD Somach Adenocarcinoma.

STES Stomach and Esophageal Carcinoma.

t-SNE *t*-distributed Stochastic Neighbour Embedding.

UCEC Uterine Corpus Endometrial Carcinoma.

UCS Uterine Carcinosarcoma.

Appendix A

Simulation study

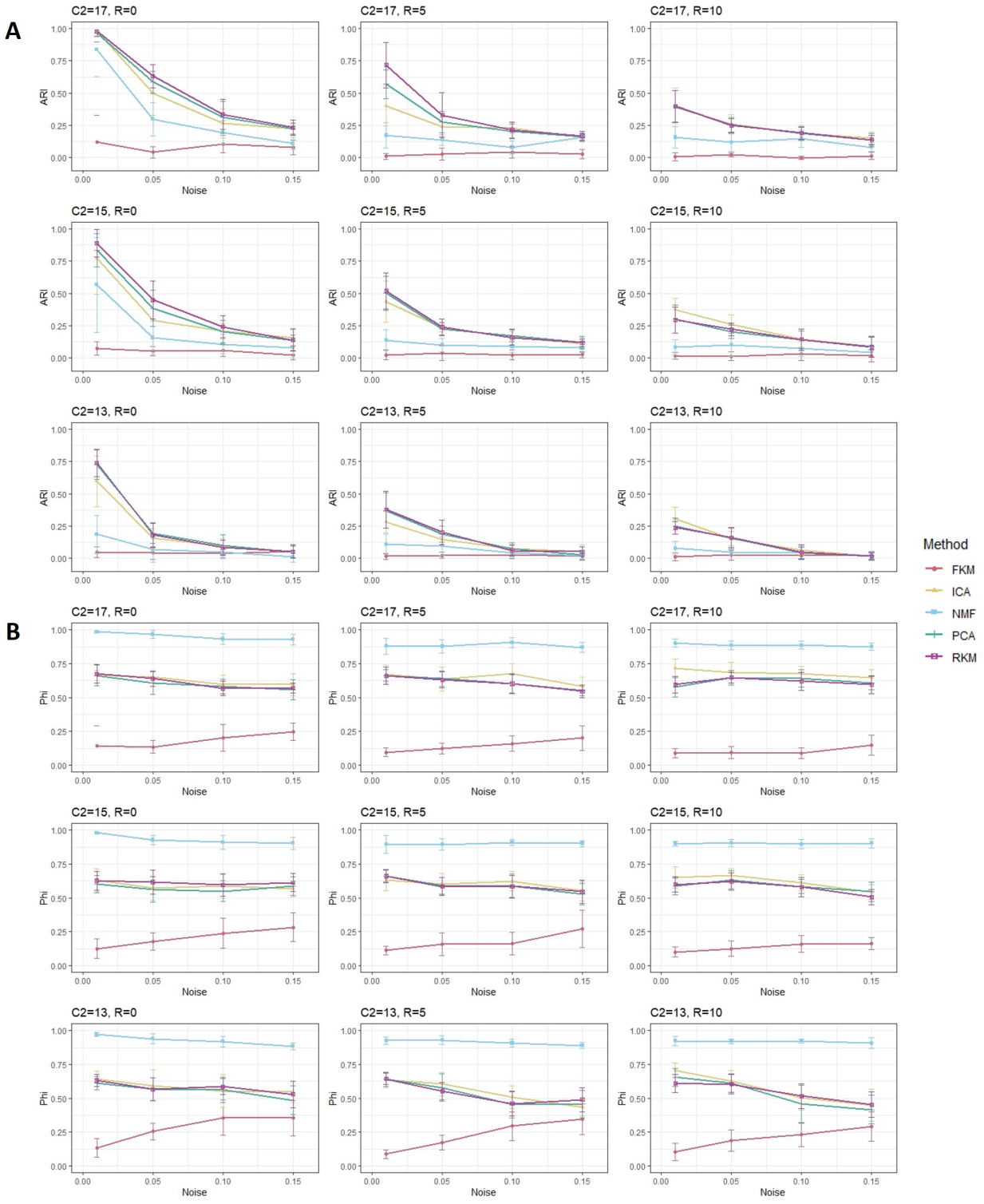

Figure A.1: A) Clustering accuracy results are measured with the Adjusted Rand Index (ARI). B) Subspace recovery results are measured with the Tucker congruence coefficient (Phi). *Note: C2: initialisation setting centroid 2, R: number of masking variables (see Section 4.1)*

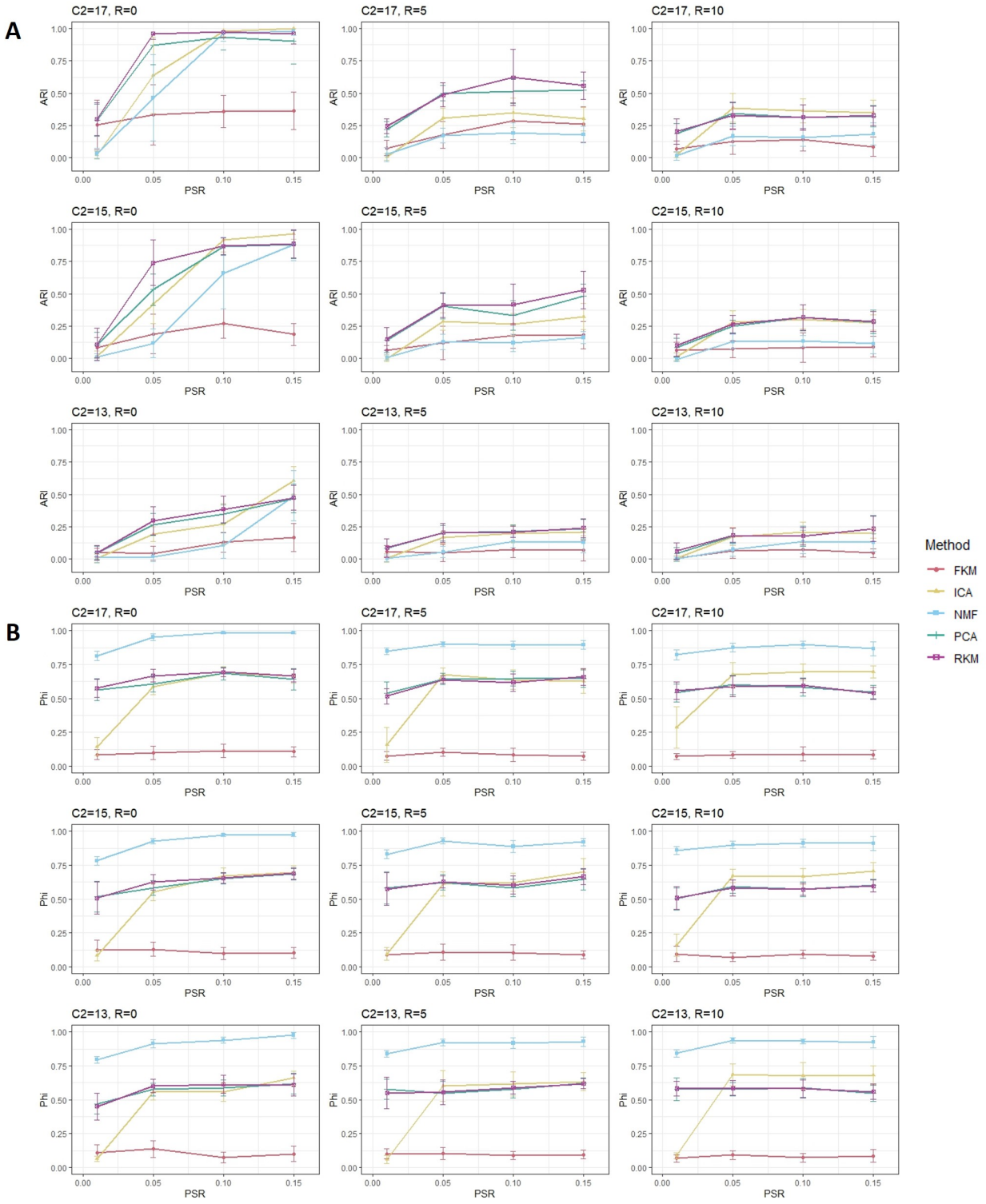

Figure A.2: A) Clustering accuracy results are measured with the Adjusted Rand Index (ARI). B) Subspace recovery results are measured with the Tucker congruence coefficient (Phi). *Note: C2: initialisation setting centroid 2, R: number of masking variables (see Section 4.1)*

Table A.1: Total results when increasing factor *Noise*, averaged over ten replications.

			PCA		**ICA**		**NMF**		**RKM**		**FKM**	
			ARI	Phi	ARI	Phi	ARI	Phi	ARI	Phi	ARI	Phi
$R = 0$	$c_2 = 13$	$N = 0.01$	0.63	0.58	0.56	0.61	0.13	0.97	0.67	0.59	0.03	0.13
		$N = 0.05$	0.19	0.56	0.19	0.58	0.07	0.94	0.19	0.55	0.04	0.24
		$N = 0.10$	0.10	0.57	0.09	0.56	0.03	0.92	0.10	0.56	0.04	0.27
		$N = 0.15$	0.05	0.45	0.04	0.47	0.00	0.88	0.05	0.49	0.03	0.35
	$c_2 = 15$	$N = 0.01$	0.89	0.66	0.88	0.70	0.62	0.98	0.94	0.67	0.05	0.13
		$N = 0.05$	0.33	0.57	0.36	0.60	0.15	0.95	0.33	0.58	0.06	0.16
		$N = 0.10$	0.23	0.58	0.19	0.60	0.07	0.93	0.22	0.56	0.04	0.20
		$N = 0.15$	0.12	0.53	0.11	0.57	0.07	0.89	0.14	0.55	0.06	0.36
	$c_2 = 17$	$N = 0.01$	0.99	0.66	0.98	0.69	0.83	0.98	0.99	0.67	0.17	0.17
		$N = 0.05$	0.60	0.60	0.49	0.65	0.27	0.97	0.65	0.62	0.02	0.11
		$N = 0.10$	0.39	0.58	0.30	0.61	0.16	0.95	0.38	0.57	0.07	0.24
		$N = 0.15$	0.21	0.57	0.17	0.54	0.12	0.91	0.23	0.56	0.04	0.23
$R = 5$	$c_2 = 13$	$N = 0.01$	0.45	0.67	0.31	0.64	0.10	0.92	0.48	0.67	0.03	0.11
		$N = 0.05$	0.19	0.53	0.15	0.55	0.07	0.90	0.15	0.55	0.04	0.26
		$N = 0.10$	0.14	0.50	0.10	0.51	0.05	0.91	0.12	0.45	0.03	0.27
		$N = 0.15$	0.01	0.43	0.02	0.46	0.01	0.92	0.01	0.43	0.03	0.31
	$c_2 = 15$	$N = 0.01$	0.59	0.64	0.43	0.66	0.17	0.91	0.61	0.65	0.03	0.11
		$N = 0.05$	0.23	0.65	0.20	0.66	0.10	0.90	0.24	0.63	0.00	0.14
		$N = 0.10$	0.18	0.57	0.17	0.63	0.06	0.91	0.16	0.54	0.04	0.20
		$N = 0.15$	0.12	0.52	0.12	0.54	0.06	0.90	0.13	0.53	0.03	0.21
	$c_2 = 17$	$N = 0.01$	0.56	0.68	0.40	0.65	0.19	0.91	0.62	0.69	0.01	0.07
		$N = 0.05$	0.35	0.61	0.26	0.60	0.14	0.89	0.36	0.60	0.03	0.12
		$N = 0.10$	0.19	0.61	0.16	0.65	0.12	0.88	0.18	0.60	0.03	0.16
		$N = 0.15$	0.20	0.57	0.20	0.59	0.11	0.89	0.20	0.59	0.03	0.22
$R = 10$	$c_2 = 13$	$N = 0.01$	0.25	0.64	0.30	0.68	0.08	0.92	0.26	0.60	0.02	0.08
		$N = 0.05$	0.15	0.60	0.14	0.60	0.05	0.91	0.14	0.57	0.02	0.14
		$N = 0.10$	0.07	0.49	0.06	0.52	0.03	0.89	0.05	0.51	0.03	0.23
		$N = 0.15$	0.03	0.45	0.05	0.44	0.04	0.90	0.04	0.43	0.02	0.26
	$c_2 = 15$	$N = 0.01$	0.32	0.61	0.35	0.68	0.12	0.90	0.34	0.60	0.01	0.07
		$N = 0.05$	0.23	0.62	0.27	0.66	0.06	0.92	0.23	0.63	-0.00	0.10
		$N = 0.10$	0.11	0.59	0.11	0.62	0.06	0.89	0.11	0.59	0.01	0.15
		$N = 0.15$	0.10	0.55	0.12	0.57	0.04	0.88	0.09	0.55	0.02	0.22
	$c_2 = 17$	$N = 0.01$	0.38	0.58	0.44	0.68	0.17	0.88	0.39	0.59	0.02	0.10
		$N = 0.05$	0.21	0.63	0.28	0.68	0.15	0.87	0.21	0.63	0.01	0.09
		$N = 0.10$	0.19	0.63	0.23	0.65	0.14	0.88	0.19	0.63	0.02	0.12
		$N = 0.15$	0.16	0.62	0.18	0.64	0.14	0.89	0.15	0.59	0.02	0.15
		Mean	0.28	0.58	0.26	0.60	0.13	0.91	0.29	0.58	0.03	0.18

Table A.2: Total results when increasing factor PSR, averaged over ten replications.

			PCA		**ICA**		**NMF**		**RKM**		**FKM**	
			ARI	Phi	ARI	Phi	ARI	Phi	ARI	Phi	ARI	Phi
$R = 0$	$c_2 = 13$	$PSR = 0.01$	0.05	0.47	0.01	0.06	0.01	0.80	0.05	0.45	0.05	0.11
		$PSR = 0.05$	0.27	0.58	0.19	0.56	0.02	0.91	0.30	0.60	0.05	0.14
		$PSR = 0.10$	0.35	0.59	0.27	0.55	0.11	0.94	0.39	0.61	0.13	0.07
		$PSR = 0.15$	0.47	0.62	0.61	0.66	0.49	0.98	0.47	0.61	0.17	0.10
	$c_2 = 15$	$PSR = 0.01$	0.10	0.52	0.02	0.09	0.01	0.78	0.11	0.51	0.09	0.12
		$PSR = 0.05$	0.53	0.58	0.42	0.55	0.12	0.93	0.74	0.62	0.19	0.13
		$PSR = 0.10$	0.87	0.65	0.92	0.67	0.66	0.97	0.87	0.65	0.27	0.10
		$PSR = 0.15$	0.88	0.68	0.96	0.69	0.88	0.98	0.89	0.69	0.19	0.10
	$c_2 = 17$	$PSR = 0.01$	0.29	0.56	0.02	0.14	0.03	0.81	0.30	0.58	0.26	0.09
		$PSR = 0.05$	0.87	0.61	0.64	0.59	0.46	0.95	0.96	0.67	0.34	0.10
		$PSR = 0.10$	0.94	0.68	0.98	0.69	0.97	0.98	0.97	0.69	0.36	0.11
		$PSR = 0.15$	0.90	0.64	1.00	0.67	0.98	0.98	0.96	0.67	0.36	0.11
$R = 5$	$c_2 = 13$	$PSR = 0.01$	0.09	0.58	0.01	0.06	0.00	0.84	0.09	0.55	0.06	0.10
		$PSR = 0.05$	0.20	0.55	0.17	0.60	0.05	0.92	0.21	0.55	0.05	0.10
		$PSR = 0.10$	0.21	0.58	0.20	0.62	0.14	0.92	0.21	0.59	0.07	0.09
		$PSR = 0.15$	0.23	0.62	0.21	0.63	0.13	0.93	0.24	0.62	0.07	0.09
	$c_2 = 15$	$PSR = 0.01$	0.14	0.58	-0.00	0.09	0.01	0.83	0.15	0.57	0.06	0.09
		$PSR = 0.05$	0.41	0.62	0.28	0.61	0.13	0.93	0.41	0.63	0.12	0.11
		$PSR = 0.10$	0.33	0.58	0.26	0.62	0.12	0.89	0.42	0.60	0.17	0.11
		$PSR = 0.15$	0.48	0.65	0.32	0.70	0.16	0.92	0.53	0.67	0.18	0.09
	$c_2 = 17$	$PSR = 0.01$	0.22	0.54	0.01	0.16	0.03	0.85	0.24	0.52	0.08	0.08
		$PSR = 0.05$	0.50	0.64	0.31	0.67	0.17	0.90	0.49	0.64	0.18	0.10
		$PSR = 0.10$	0.52	0.65	0.35	0.64	0.19	0.89	0.62	0.62	0.28	0.08
		$PSR = 0.15$	0.52	0.65	0.30	0.63	0.18	0.89	0.56	0.66	0.26	0.07
$R = 10$	$c_2 = 13$	$PSR = 0.01$	0.04	0.58	0.01	0.09	-0.00	0.84	0.06	0.58	0.01	0.07
		$PSR = 0.05$	0.18	0.58	0.17	0.68	0.07	0.94	0.18	0.59	0.07	0.09
		$PSR = 0.10$	0.18	0.59	0.21	0.68	0.13	0.93	0.18	0.58	0.07	0.07
		$PSR = 0.15$	0.24	0.55	0.20	0.68	0.13	0.92	0.23	0.56	0.05	0.08
	$c_2 = 15$	$PSR = 0.01$	0.08	0.50	0.01	0.16	-0.01	0.86	0.10	0.51	0.07	0.09
		$PSR = 0.05$	0.25	0.59	0.28	0.67	0.13	0.90	0.26	0.58	0.07	0.07
		$PSR = 0.10$	0.32	0.57	0.30	0.66	0.13	0.91	0.32	0.57	0.08	0.09
		$PSR = 0.15$	0.28	0.60	0.28	0.71	0.11	0.91	0.29	0.59	0.09	0.08
	$c_2 = 17$	$PSR = 0.01$	0.19	0.54	0.02	0.29	0.01	0.82	0.20	0.56	0.07	0.07
		$PSR = 0.05$	0.34	0.60	0.39	0.68	0.17	0.88	0.33	0.59	0.12	0.08
		$PSR = 0.10$	0.31	0.58	0.36	0.69	0.16	0.90	0.31	0.59	0.14	0.09
		$PSR = 0.15$	0.32	0.55	0.35	0.69	0.18	0.87	0.33	0.54	0.09	0.08
		Mean	0.36	0.59	0.31	0.52	0.20	0.90	0.39	0.59	0.14	0.09

Table A.3: Summary of the PF [%] in Experiment 1.

	PSR:	**0.01**	**0.05**	**0.10**	**0.15**
R = 0	$c_2 = 13$	6.75 (1.77)	0.60 (1.46)	0 (0)	0 (0)
	$c_2 = 15$	6.35 (1.78)	1.23 (3.61)	0 (0)	0 (0)
	$c_2 = 17$	5.13 (1.22)	0.17 (0.19)	1.00 (3.16)	1.00 (3.16)
R = 5	$c_2 = 13$	4.52 (0.93)	0.24 (0.33)	0.28 (0.84)	0 (0)
	$c_2 = 15$	3.47 (1.17)	0.04 (0.08)	0.08 (0.15)	0 (0)
	$c_2 = 17$	3.51 (1.23)	0.27 (0.3)	0.68 (2.1)	0.02 (0.05)
R = 10	$c_2 = 13$	2.72 (1.05)	0.55 (1.57)	0.51 (1.58)	0.03 (0.08)
	$c_2 = 15$	3.07 (0.76)	0.39 (0.56)	0.5 (1.58)	0.01 (0.03)
	$c_2 = 17$	2.21 (0.68)	0.89 (1.57)	0.03 (0.06)	0.52 (1.58)

Table A.4: PF [%] of the datasets generated for each PSR in Experiment 2.

PSR	**PF**
0.01	7.85 (2.82)
0.03	2.35 (3.35)
0.05	0.23 (0.45)
0.07	0 (0)
0.09	0.02 (0.05)
0.11	0.18 (0.39)
0.13	0 (0)
0.15	0 (0)

Table A.5: Total results Experiment 2 - Random noise [ARI].

	Joint										Tandem		
	FKM	RKM									PCA	ICA	NMF
	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$	$\alpha = 1$		
$N = 0.01$	0.06 (0.05)	0.82 (0.15)	**0.98** (0.02)	**0.98** (0.03)	**0.98** (0.03)	0.97 (0.03)	0.96 (0.06)	0.95 (0.07)	0.95 (0.07)	0.95 (0.07)	0.95 (0.07)	0.96 (0.03)	0.72 (0.13)
$N = 0.03$	0.02 (0.03)	0.62 (0.09)	0.70 (0.17)	0.82 (0.17)	**0.83** (0.16)	**0.83** (0.16)	0.82 (0.15)	0.82 (0.16)	0.79 (0.18)	0.79 (0.18)	0.79 (0.18)	0.72 (0.16)	0.43 (0.18)
$N = 0.05$	0.03 (0.05)	0.52 (0.1)	0.54 (0.07)	0.64 (0.15)	**0.65** (0.17)	0.63 (0.19)	**0.65** (0.15)	0.63 (0.15)	0.6 (0.16)	0.6 (0.16)	0.57 (0.18)	0.56 (0.16)	0.42 (0.15)
$N = 0.07$	0.04 (0.03)	0.37 (0.1)	0.41 (0.09)	0.44 (0.07)	**0.49** (0.11)	**0.49** (0.1)	0.45 (0.11)	0.45 (0.11)	0.47 (0.1)	0.46 (0.11)	0.46 (0.09)	0.33 (0.09)	0.21 (0.09)
$N = 0.09$	0.07 (0.06)	0.33 (0.14)	0.38 (0.13)	0.42 (0.14)	0.45 (0.14)	**0.47** (0.14)	**0.47** (0.15)	0.44 (0.17)	0.44 (0.17)	0.44 (0.17)	0.42 (0.18)	0.30 (0.13)	0.20 (0.09)
$N = 0.11$	0.07 (0.08)	0.26 (0.1)	0.34 (0.08)	**0.37** (0.1)	0.31 (0.08)	0.33 (0.13)	0.33 (0.13)	0.33 (0.13)	0.33 (0.11)	0.32 (0.09)	0.34 (0.12)	0.32 (0.17)	0.23 (0.11)
$N = 0.13$	0.03 (0.06)	0.21 (0.1)	0.26 (0.07)	**0.28** (0.08)	0.22 (0.05)	0.24 (0.06)	0.24 (0.05)	0.25 (0.05)	0.25 (0.05)	0.23 (0.03)	0.24 (0.07)	0.19 (0.07)	0.14 (0.08)
$N = 0.15$	0.03 (0.05)	0.15 (0.05)	0.21 (0.07)	0.22 (0.1)	0.21 (0.09)	0.19 (0.06)	0.19 (0.1)	0.19 (0.1)	0.19 (0.1)	0.19 (0.1)	0.20 (0.11)	**0.24** (0.07)	0.16 (0.08)

Table A.6: Total results Experiment 2 - Random noise [Phi].

	Joint										Tandem		
	FKM	RKM									PCA	ICA	NMF
	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$	$\alpha = 1$		
$N = 0.01$	0.13 (0.04)	0.59 (0.07)	0.66 (0.05)	0.67 (0.04)	0.66 (0.04)	0.66 (0.05)	0.65 (0.05)	0.65 (0.05)	0.65 (0.05)	0.64 (0.05)	0.64 (0.05)	0.66 (0.06)	**0.99** (0.01)
$N = 0.03$	0.12 (0.05)	0.54 (0.04)	0.57 (0.08)	0.65 (0.06)	0.66 (0.05)	0.67 (0.06)	0.66 (0.07)	0.65 (0.07)	0.65 (0.07)	0.64 (0.07)	0.64 (0.08)	0.65 (0.03)	**0.98** (0.00)
$N = 0.05$	0.15 (0.05)	0.52 (0.04)	0.53 (0.04)	0.62 (0.09)	0.66 (0.07)	0.66 (0.08)	0.65 (0.09)	0.65 (0.09)	0.65 (0.09)	0.64 (0.09)	0.64 (0.09)	0.68 (0.07)	**0.98** (0.01)
$N = 0.07$	0.15 (0.07)	0.47 (0.04)	0.52 (0.05)	0.54 (0.07)	0.58 (0.06)	0.61 (0.06)	0.61 (0.06)	0.61 (0.07)	0.62 (0.06)	0.62 (0.06)	0.61 (0.06)	0.61 (0.06)	**0.95** (0.02)
$N = 0.09$	0.18 (0.08)	0.48 (0.09)	0.54 (0.05)	0.58 (0.06)	0.60 (0.07)	0.61 (0.05)	0.61 (0.05)	0.61 (0.06)	0.59 (0.07)	0.59 (0.07)	0.59 (0.07)	0.63 (0.09)	**0.96** (0.02)
$N = 0.11$	0.21 (0.09)	0.45 (0.08)	0.52 (0.04)	0.56 (0.06)	0.57 (0.07)	0.56 (0.06)	0.57 (0.05)	0.55 (0.06)	0.55 (0.05)	0.55 (0.05)	0.55 (0.05)	0.62 (0.09)	**0.93** (0.04)
$N = 0.13$	0.22 (0.09)	0.48 (0.11)	0.52 (0.05)	0.56 (0.06)	0.61 (0.07)	0.6 (0.06)	0.6 (0.06)	0.59 (0.06)	0.59 (0.07)	0.59 (0.06)	0.60 (0.06)	0.58 (0.09)	**0.92** (0.04)
$N = 0.15$	0.18 (0.07)	0.42 (0.06)	0.48 (0.05)	0.57 (0.08)	0.56 (0.07)	0.57 (0.05)	0.58 (0.05)	0.57 (0.04)	0.56 (0.05)	0.56 (0.05)	0.56 (0.05)	0.59 (0.06)	**0.92** (0.04)

Table A.7: Total results Experiment 2 - Subspace residuals [ARI].

	Joint										Tandem		
	FKM	RKM									PCA	ICA	NMF
	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$	$\alpha = 1$		
$PSR = 0.01$	0.16 (0.17)	**0.35** (0.32)	0.25 (0.23)	0.25 (0.28)	0.28 (0.28)	0.25 (0.24)	0.25 (0.24)	0.24 (0.22)	0.24 (0.22)	0.24 (0.22)	0.25 (0.21)	0.02 (0.03)	0.01 (0.04)
$PSR = 0.03$	0.23 (0.19)	0.63 (0.08)	0.74 (0.23)	0.88 (0.15)	**0.90** (0.12)	0.84 (0.15)	0.81 (0.15)	0.76 (0.16)	0.73 (0.14)	0.71 (0.13)	0.71 (0.12)	0.30 (0.07)	0.07 (0.06)
$PSR = 0.05$	0.36 (0.19)	0.64 (0.06)	0.76 (0.20)	0.88 (0.17)	**0.93** (0.11)	**0.93** (0.1)	0.91 (0.14)	0.82 (0.16)	0.82 (0.16)	0.80 (0.15)	0.81 (0.16)	0.46 (0.25)	0.3 (0.19)
$PSR = 0.07$	0.29 (0.16)	0.64 (0.12)	0.89 (0.2)	0.94 (0.17)	**0.99** (0.02)	**0.99** (0.02)	0.94 (0.15)	0.94 (0.15)	0.94 (0.15)	0.92 (0.15)	0.91 (0.16)	0.80 (0.30)	0.59 (0.44)
$PSR = 0.09$	0.40 (0.16)	0.64 (0.02)	0.85 (0.19)	**0.99** (0.02)	**0.99** (0.02)	**0.99** (0.02)	**0.99** (0.02)	**0.99** (0.02)	0.98 (0.03)	0.98 (0.03)	0.97 (0.04)	0.93 (0.22)	0.94 (0.13)
$PSR = 0.11$	0.28 (0.18)	0.65 (0.07)	0.77 (0.19)	0.98 (0.02)	0.98 (0.02)	0.97 (0.02)	0.97 (0.02)	0.97 (0.02)	0.97 (0.02)	0.97 (0.02)	0.97 (0.02)	**1** (0.01)	0.99 (0.03)
$PSR = 0.13$	0.42 (0.12)	0.69 (0.12)	0.87 (0.19)	0.91 (0.17)	0.99 (0.03)	0.99 (0.03)	0.99 (0.03)	0.99 (0.03)	0.99 (0.03)	0.99 (0.03)	0.95 (0.12)	**1** (0)	**1** (0)
$PSR = 0.15$	0.43 (0.13)	0.61 (0.09)	0.77 (0.21)	0.98 (0.03)	0.98 (0.03)	0.98 (0.03)	0.98 (0.03)	0.98 (0.03)	0.98 (0.03)	0.98 (0.03)	0.98 (0.03)	**1** (0)	0.99 (0.03)

Table A.8: Total results Experiment 2 - Subspace residuals [Phi].

	Joint										Tandem		
	FKM	RKM									PCA	ICA	NMF
	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$	$\alpha = 1$		
$PSR = 0.01$	0.09 (0.04)	0.43 (0.11)	0.41 (0.07)	0.42 (0.13)	0.51 (0.13)	0.57 (0.1)	0.56 (0.09)	0.55 (0.09)	0.55 (0.08)	0.55 (0.08)	0.55 (0.08)	0.15 (0.07)	**0.82** (0.03)
$PSR = 0.03$	0.11 (0.06)	0.52 (0.06)	0.60 (0.11)	0.63 (0.1)	0.65 (0.06)	0.63 (0.07)	0.61 (0.07)	0.59 (0.08)	0.58 (0.08)	0.58 (0.08)	0.57 (0.08)	0.51 (0.02)	**0.91** (0.03)
$PSR = 0.05$	0.10 (0.03)	0.52 (0.04)	0.59 (0.09)	0.61 (0.08)	0.65 (0.05)	0.64 (0.06)	0.63 (0.06)	0.59 (0.07)	0.59 (0.06)	0.58 (0.06)	0.58 (0.06)	0.55 (0.06)	**0.93** (0.04)
$PSR = 0.07$	0.12 (0.05)	0.51 (0.04)	0.61 (0.09)	0.64 (0.06)	0.65 (0.04)	0.65 (0.04)	0.64 (0.06)	0.63 (0.06)	0.63 (0.06)	0.63 (0.06)	0.63 (0.06)	0.63 (0.08)	**0.97** (0.01)
$PSR = 0.09$	0.08 (0.04)	0.51 (0.03)	0.60 (0.09)	0.66 (0.03)	0.66 (0.03)	0.66 (0.03)	0.66 (0.03)	0.66 (0.03)	0.66 (0.03)	0.66 (0.03)	0.66 (0.03)	0.68 (0.08)	0.97 (0.02)
$PSR = 0.11$	0.14 (0.06)	0.50 (0.04)	0.57 (0.11)	0.64 (0.06)	0.63 (0.06)	0.63 (0.06)	0.63 (0.06)	0.63 (0.06)	0.63 (0.06)	0.63 (0.06)	0.63 (0.06)	0.68 (0.05)	**0.98** (0.01)
$PSR = 0.13$	0.11 (0.06)	0.51 (0.05)	0.62 (0.08)	0.63 (0.08)	0.65 (0.06)	0.65 (0.06)	0.65 (0.06)	0.65 (0.06)	0.65 (0.06)	0.65 (0.06)	0.65 (0.06)	0.72 (0.06)	**0.98** (0.01)
$PSR = 0.15$	0.10 (0.04)	0.50 (0.04)	0.57 (0.09)	0.68 (0.05)	0.68 (0.05)	0.68 (0.05)	0.68 (0.06)	0.67 (0.06)	0.67 (0.06)	0.67 (0.06)	0.67 (0.06)	0.7 (0.04)	**0.98** (0.01)

Table A.9: Total results Experiment 2 - Masking variables [ARI].

	Joint										Tandem		
	FKM	RKM									PCA	ICA	NMF
	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$	$\alpha = 1$		
$R = 1$	0.8 (0.19)	**1** (0)	**1** (0)	**1** (0)	**1** (0)	**1** (0)	**1** (0)	**1** (0)	0.97 (0.1)	0.97 (0.1)	0.96 (0.1)	0.60 (0.24)	0.39 (0.3)
$R = 2$	0.87 (0.15)	**1** (0)	**1** (0)	**1** (0)	**1** (0)	**1** (0)	0.95 (0.11)	0.87 (0.15)	0.84 (0.15)	0.81 (0.15)	0.79 (0.13)	0.50 (0.1)	0.19 (0.07)
$R = 3$	0.77 (0.21)	**1** (0)	**1** (0)	**1** (0)	**1** (0)	0.91 (0.14)	0.76 (0.16)	0.75 (0.14)	0.69 (0.05)	0.69 (0.05)	0.69 (0.04)	0.54 (0.07)	0.19 (0.09)
$R = 4$	0.72 (0.13)	**1** (0)	**1** (0)	**1** (0)	**1** (0)	0.82 (0.19)	0.68 (0.1)	0.66 (0.05)	0.66 (0.05)	0.66 (0.05)	0.65 (0.04)	0.44 (0.1)	0.19 (0.09)
$R = 5$	0.94 (0.11)	**1** (0)	**1** (0)	**1** (0)	0.93 (0.15)	0.72 (0.16)	0.65 (0.05)	0.65 (0.04)	0.65 (0.04)	0.65 (0.04)	0.66 (0.02)	0.48 (0.14)	0.17 (0.09)
$R = 6$	0.69 (0.19)	**0.99** (0.03)	0.98 (0.06)	0.97 (0.09)	0.97 (0.11)	0.63 (0.1)	0.62 (0.07)	0.62 (0.06)	0.61 (0.06)	0.61 (0.06)	0.60 (0.06)	0.40 (0.09)	0.13 (0.07)
$R = 7$	0.56 (0.3)	0.98 (0.07)	0.89 (0.15)	**1** (0)	0.99 (0.03)	0.58 (0.05)	0.58 (0.05)	0.58 (0.05)	0.58 (0.05)	0.58 (0.05)	0.54 (0.1)	0.46 (0.07)	0.13 (0.07)
$R = 8$	0.67 (0.31)	**1** (0)	0.94 (0.12)	**1** (0)	0.85 (0.24)	0.54 (0.08)	0.53 (0.08)	0.52 (0.08)	0.52 (0.08)	0.52 (0.08)	0.48 (0.15)	0.4 (0.08)	0.16 (0.11)
$R = 9$	0.58 (0.29)	0.96 (0.12)	0.97 (0.08)	**1** (0)	0.75 (0.32)	0.5 (0.11)	0.47 (0.09)	0.47 (0.09)	0.47 (0.09)	0.47 (0.09)	0.43 (0.11)	0.40 (0.06)	0.13 (0.07)
$R = 10$	0.48 (0.33)	**1** (0)	0.95 (0.11)	0.95 (0.15)	0.77 (0.26)	0.51 (0.09)	0.46 (0.09)	0.46 (0.09)	0.46 (0.08)	0.46 (0.09)	0.43 (0.07)	0.50 (0.08)	0.16 (0.10)

Table A.10: Total results Experiment 2 - Masking variables [Phi].

	Joint										Tandem		
	FKM	RKM									PCA	ICA	NMF
	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$	$\alpha = 1$		
$R = 1$	0.17 (0.07)	**0.64** (0.03)	0.64 (0.03)	0.64 (0.03)	0.64 (0.02)	0.64 (0.02)	0.63 (0.02)	0.63 (0.02)	0.63 (0.03)	0.63 (0.04)	0.63 (0.05)	0.64 (0.05)	**0.95** (0.02)
$R = 2$	0.17 (0.09)	0.59 (0.05)	0.59 (0.06)	0.59 (0.06)	0.59 (0.06)	0.60 (0.07)	0.61 (0.09)	0.64 (0.1)	0.65 (0.08)	0.65 (0.09)	0.66 (0.09)	0.67 (0.05)	**0.93** (0.02)
$R = 3$	0.16 (0.06)	0.59 (0.05)	0.59 (0.05)	0.59 (0.05)	0.59 (0.05)	0.61 (0.04)	0.62 (0.05)	0.62 (0.05)	0.63 (0.04)	0.63 (0.04)	0.62 (0.04)	0.66 (0.06)	**0.91** (0.03)
$R = 4$	0.13 (0.05)	0.56 (0.06)	0.56 (0.06)	0.56 (0.06)	0.56 (0.05)	0.63 (0.09)	0.67 (0.06)	0.67 (0.06)	0.67 (0.06)	0.67 (0.07)	0.67 (0.07)	0.63 (0.09)	**0.90** (0.03)
$R = 5$	0.15 (0.05)	0.50 (0.05)	0.51 (0.05)	0.51 (0.05)	0.56 (0.09)	0.63 (0.08)	0.66 (0.05)	0.66 (0.05)	0.67 (0.05)	0.66 (0.05)	0.66 (0.05)	0.7 (0.07)	**0.90** (0.04)
$R = 6$	0.12 (0.07)	0.51 (0.06)	0.51 (0.04)	0.52 (0.06)	0.56 (0.07)	0.66 (0.03)	0.67 (0.04)	0.67 (0.04)	0.67 (0.04)	0.67 (0.04)	0.67 (0.04)	0.71 (0.07)	**0.91** (0.04)
$R = 7$	0.12 (0.08)	0.5 (0.07)	0.47 (0.09)	0.5 (0.06)	0.51 (0.05)	0.64 (0.06)	0.64 (0.06)	0.64 (0.06)	0.64 (0.06)	0.64 (0.07)	0.63 (0.07)	0.68 (0.07)	**0.89** (0.02)
$R = 8$	0.11 (0.07)	0.48 (0.06)	0.47 (0.06)	0.48 (0.06)	0.56 (0.09)	0.65 (0.04)	0.65 (0.04)	0.64 (0.05)	0.64 (0.05)	0.63 (0.06)	0.62 (0.06)	0.67 (0.08)	**0.88** (0.05)
$R = 9$	0.11 (0.04)	0.49 (0.06)	0.49 (0.08)	0.50 (0.06)	0.50 (0.09)	0.62 (0.07)	0.64 (0.04)	0.64 (0.04)	0.63 (0.04)	0.63 (0.05)	0.63 (0.05)	0.7 (0.09)	**0.91** (0.03)
$R = 10$	0.11 (0.03)	0.49 (0.05)	0.46 (0.07)	0.49 (0.05)	0.58 (0.10)	0.65 (0.04)	0.60 (0.06)	0.59 (0.07)	0.58 (0.08)	0.58 (0.07)	0.58 (0.07)	0.70 (0.04)	**0.87** (0.03)

Appendix B

Empirical analysis

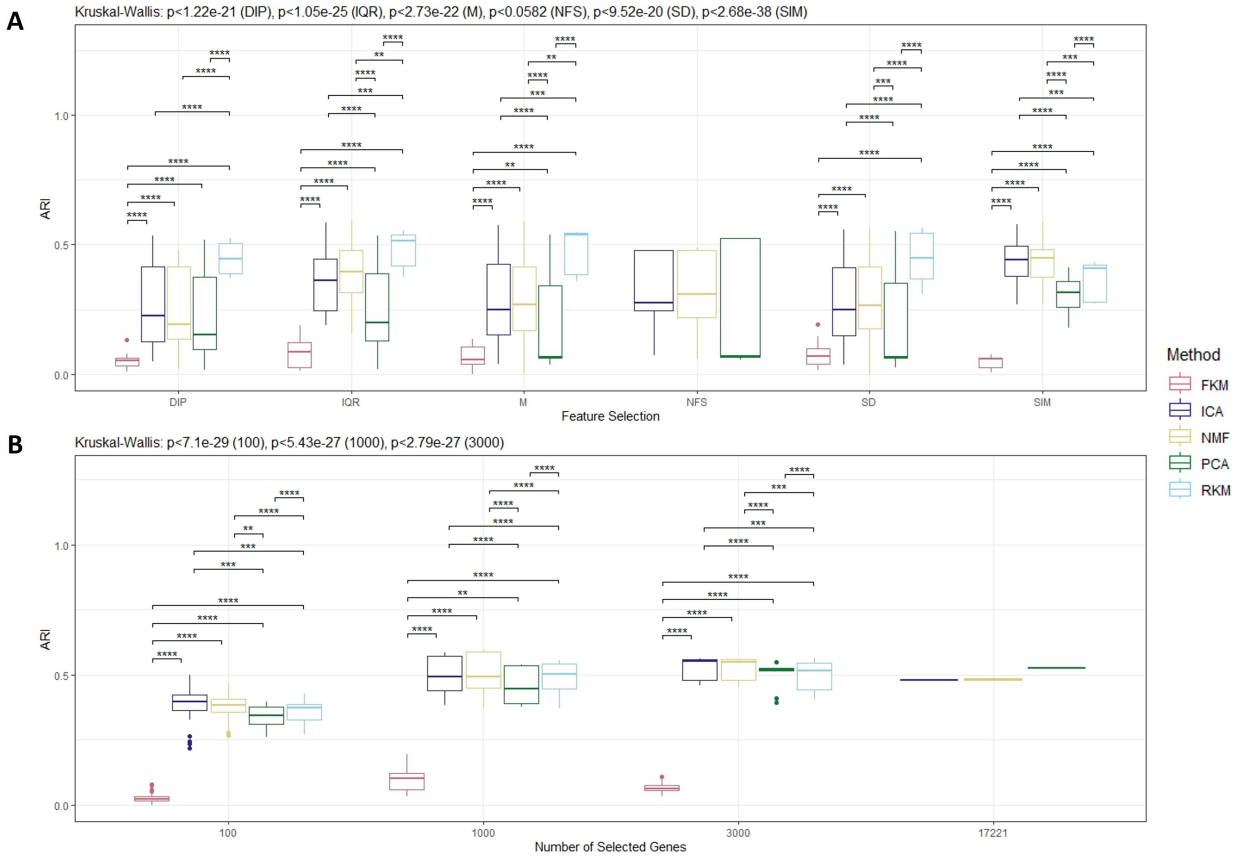

Figure B.1: Results clustering analysis, specified to 5 number of components. A) Results depicted for Feature Selection, B) Results depicted for Number of Selected Genes.

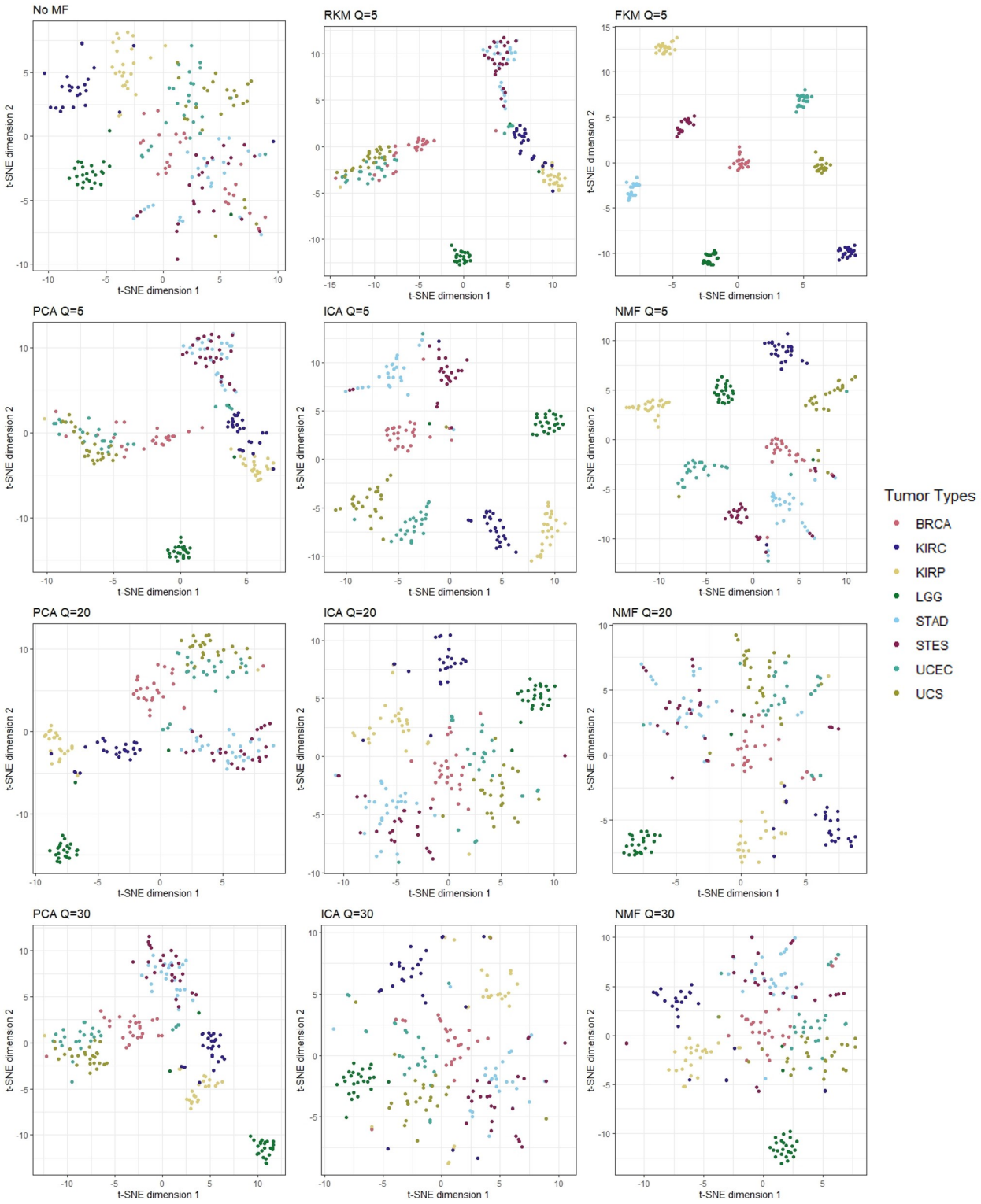

Figure B.2: t-SNE comparison of the interaction between clustering algorithms with the number of components.

Table B.1: Results top-performing specifications for clustering analysis, rank 1-59.

Ranking	**Feature Selection**	**N. Genes**	**Method**	**N. Components**	**ARI**
1	IQR	1000	NMF	5	0.60
2	SIM	3000	NMF	30	0.59
3	M	1000	NMF	5	0.59
4	IQR	1000	ICA	5	0.59
5	SIM	3000	ICA	30	0.58
6	M	1000	ICA	5	0.57
7	SIM	3000	NMF	20	0.57
8	SD	3000	RKM	5	0.56
9	IQR	3000	ICA	5	0.56
10	SD	3000	NMF	5	0.56
11	IQR	3000	NMF	5	0.56
12	SD	3000	ICA	5	0.56
13	IQR	1000	RKM	5	0.56
14	M	3000	ICA	5	0.55
15	M	3000	NMF	5	0.55
16	SD	3000	PCA	5	0.55
17	M	1000	RKM	5	0.55
18	M	3000	RKM	5	0.55
19	SIM	1000	NMF	30	0.54
20	M	1000	PCA	5	0.54
21	SIM	3000	ICA	20	0.54
22	SIM	1000	ICA	20	0.54
23	IQR	1000	PCA	5	0.54
24	DIP	3000	ICA	5	0.53
25	SIM	1000	ICA	30	0.53
26	NFS	17221	PCA	5	0.53
27	DIP	1000	RKM	5	0.53
28	M	3000	PCA	5	0.53
29	DIP	3000	RKM	5	0.53
30	IQR	3000	RKM	5	0.52
31	DIP	3000	PCA	5	0.52
32	SD	1000	ICA	5	0.52
33	IQR	3000	PCA	5	0.52
34	SIM	1000	NMF	20	0.50
35	SD	1000	NMF	5	0.50
36	M	100	ICA	5	0.50
37	IQR	3000	NMF	20	0.49
38	NFS	17221	NMF	20	0.49
39	DIP	100	ICA	5	0.48
40	NFS	17221	NMF	5	0.48
41	SIM	3000	ICA	5	0.48
42	SIM	3000	NMF	5	0.48
43	IQR	1000	NMF	20	0.48
44	NFS	17221	ICA	5	0.48
45	IQR	1000	NMF	30	0.48
46	DIP	3000	NMF	5	0.48
47	DIP	100	NMF	5	0.47
48	IQR	3000	NMF	30	0.46
49	IQR	1000	ICA	20	0.46
50	DIP	1000	PCA	5	0.46
51	IQR	3000	ICA	20	0.46
52	NFS	17221	NMF	30	0.46
53	M	3000	NMF	20	0.45
54	SD	1000	RKM	5	0.45
55	SD	1000	ICA	20	0.45
56	SIM	1000	NMF	5	0.45
57	DIP	1000	ICA	5	0.45
58	SD	1000	PCA	5	0.45
59	SIM	100	ICA	20	0.44

Table B.2: Results top-performing specifications for clustering analysis, rank 60-119.

Ranking	**Feature Selection**	**N. Genes**	**Method**	**N. Components**	**ARI**
60	M	1000	ICA	30	0.44
61	SIM	1000	ICA	5	0.44
62	DIP	1000	NMF	5	0.44
63	IQR	100	ICA	30	0.44
64	M	1000	NMF	30	0.43
65	SIM	1000	RKM	5	0.43
66	SIM	3000	RKM	5	0.43
67	SD	3000	NMF	20	0.43
68	IQR	100	RKM	5	0.43
69	SD	1000	NMF	30	0.43
70	SD	3000	ICA	20	0.42
71	SIM	100	NMF	20	0.42
72	IQR	100	NMF	30	0.42
73	DIP	1000	ICA	20	0.42
74	M	1000	NMF	20	0.42
75	SD	100	ICA	5	0.42
76	DIP	100	RKM	5	0.41
77	IQR	100	ICA	5	0.41
78	M	100	NMF	5	0.41
79	DIP	3000	NMF	20	0.41
80	SIM	3000	PCA	5	0.41
81	IQR	100	NMF	5	0.41
82	M	100	RKM	5	0.41
83	IQR	100	NMF	20	0.40
84	IQR	100	PCA	5	0.40
85	SD	100	NMF	5	0.40
86	SIM	100	ICA	30	0.40
87	SIM	100	NMF	5	0.39
88	IQR	1000	ICA	30	0.39
89	DIP	3000	ICA	20	0.39
90	SIM	100	ICA	5	0.38
91	M	100	NMF	20	0.38
92	DIP	100	PCA	5	0.38
93	SIM	1000	PCA	5	0.38
94	IQR	100	ICA	20	0.37
95	SD	100	RKM	5	0.37
96	SD	1000	NMF	20	0.37
97	SIM	1000	PCA	20	0.36
98	SD	100	PCA	5	0.36
99	SIM	100	NMF	30	0.35
100	M	100	PCA	5	0.34
101	NFS	17221	ICA	20	0.34
102	M	3000	ICA	20	0.34
103	SIM	3000	PCA	30	0.34
104	M	1000	ICA	20	0.33
105	SIM	3000	PCA	20	0.33
106	M	3000	ICA	30	0.32
107	SIM	1000	PCA	30	0.32
108	IQR	3000	ICA	30	0.31
109	SD	3000	ICA	30	0.31
110	SIM	100	PCA	20	0.28
111	SIM	100	RKM	5	0.28
112	SIM	100	PCA	30	0.28
113	SD	3000	NMF	30	0.28
114	NFS	17221	ICA	30	0.28
115	M	3000	NMF	30	0.27
116	IQR	100	PCA	20	0.26
117	IQR	100	PCA	30	0.26
118	SIM	100	PCA	5	0.26
119	SD	1000	ICA	30	0.26

Table B.3: Results top-performing specifications for clustering analysis, rank 120-174.

Ranking	**Feature Selection**	**N. Genes**	**Method**	**N. Components**	**ARI**
120	DIP	3000	NMF	30	0.26
121	DIP	100	ICA	20	0.25
122	DIP	3000	ICA	30	0.25
123	M	100	ICA	20	0.25
124	M	100	ICA	30	0.25
125	IQR	3000	PCA	20	0.24
126	SD	100	ICA	20	0.24
127	DIP	100	PCA	30	0.24
128	DIP	100	NMF	20	0.22
129	IQR	3000	PCA	30	0.22
130	DIP	1000	NMF	20	0.22
131	SD	100	NMF	20	0.21
132	SD	100	NMF	30	0.20
133	DIP	100	NMF	30	0.20
134	IQR	1000	PCA	30	0.20
135	SD	100	ICA	30	0.20
136	SD	1000	FKM	5	0.19
137	M	100	NMF	30	0.19
138	IQR	1000	FKM	5	0.19
139	DIP	1000	NMF	30	0.18
140	DIP	100	ICA	30	0.17
141	DIP	1000	ICA	30	0.17
142	DIP	100	PCA	20	0.16
143	DIP	1000	PCA	20	0.14
144	M	1000	FKM	5	0.14
145	IQR	1000	PCA	20	0.13
146	DIP	1000	FKM	5	0.13
147	DIP	3000	PCA	20	0.12
148	M	3000	FKM	5	0.11
149	IQR	3000	FKM	5	0.11
150	SD	3000	FKM	5	0.11
151	M	3000	PCA	30	0.10
152	DIP	1000	PCA	30	0.09
153	SD	3000	PCA	30	0.09
154	NFS	17221	PCA	30	0.08
155	SD	100	FKM	5	0.08
156	DIP	100	FKM	5	0.08
157	SD	1000	PCA	30	0.08
158	SIM	3000	FKM	5	0.07
159	SD	100	PCA	30	0.07
160	DIP	3000	PCA	30	0.07
161	M	1000	PCA	30	0.07
162	DIP	3000	FKM	5	0.07
163	M	100	PCA	20	0.07
164	SIM	1000	FKM	5	0.07
165	NFS	17221	PCA	20	0.06
166	SD	1000	PCA	20	0.06
167	SD	3000	PCA	20	0.06
168	M	1000	PCA	20	0.06
169	M	3000	PCA	20	0.06
170	M	100	PCA	30	0.06
171	M	100	FKM	5	0.06
172	SD	100	PCA	20	0.05
173	IQR	100	FKM	5	0.05
174	SIM	100	FKM	5	0.03

Component	GO Biological Process	p-value
1	**Positive**	
	organic hydroxy compound transport	3.0e-04
	lipid transport	9.6e-04
	Negative	
	cellular response to amino acid stimulus	5.0e-06
	peptide cross-linking	6.4e-06
	cellular response to acid chemical	7.4e-06
	response to amino acid	1.4e-05
	response to acid chemical	2.1e-05
	tissue development	2.4e-05
	cellular response to organonitrogen compound	5.2e-05
	response to organonitrogen compound	6.7e-05
	cellular response to nitrogen compound	7.1e-05
	cell adhesion	7.6e-05
	endodermal cell differentiation	8.0e-05
	response to oxygen-containing compound	8.4e-05
	response to nitrogen compound	9.2e-05
	endoderm formation	9.2e-05
	skeletal system development	1.1e-04
	ossification	1.1e-04
	cellular response to oxygen-containing compound	1.1e-04
	collagen fibril organization	1.5e-04
	endoderm development	1.7e-04
	platelet activation	3.0e-04
	anatomical structure morphogenesis	3.4e-04
	embryo development	3.6e-04
	formation of primary germ layer	4.2e-04
	collagen metabolic process	4.6e-04
	extracellular matrix organization	8.8e-04
	extracellular structure organization	8.8e-04
	external encapsulating structure organization	8.8e-04
	anatomical structure development	9.8e-04
2	**Positive**	
	muscle contraction	2.8e-09
	muscle system process	2.5e-08
	platelet aggregation	8.1e-06
	wound healing	8.7e-06
	platelet activation	3.4e-05
	homotypic cell-cell adhesion	3.8e-05
	response to wounding	4.9e-05
	actin filament-based process	5.2e-05
	muscle structure development	5.4e-05
	myofibril assembly	1.5e-04
	cytoskeleton organization	1.6e-04
	striated muscle cell development	1.8e-04
	smooth muscle contraction	2.4e-04
	actin cytoskeleton organization	2.6e-04
	blood coagulation	3.9e-04
	coagulation	3.9e-04
	supramolecular fiber organization	4.1e-04
	hemostasis	4.2e-04
	system process	4.4e-04
	cellular component assembly involved in morphogenesis	6.2e-04
	Negative	
	ossification	1.3e-04
	osteoblast differentiation	9.3e-04
	response to vitamin	9.3e-04
	positive regulation of cell communication	9.9e-04
3	**Positive**	
	regulation of endothelial cell proliferation	4.7e-04
	endothelial cell proliferation	6.1e-04
	Negative	
	antigen processing and presentation of endogenous antigen	1.1e-06
	cell killing	2.0e-06
	positive regulation of immune effector process	8.4e-06
	regulation of T cell mediated cytotoxicity	2.2e-05
	positive regulation of T cell mediated cytotoxicity	2.2e-05

	antigen processing and presentation of peptide antigen	2.3e-05
	antigen processing and presentation of endogenous peptide antigen via MHC class I	3.8e-05
	T cell mediated cytotoxicity	4.9e-05
	regulation of immune effector process	5.5e-05
	antigen processing and presentation of endogenous peptide antigen	6.1e-05
	positive regulation of T cell mediated immunity	6.1e-05
	positive regulation of leukocyte mediated cytotoxicity	7.4e-05
	iron ion transport	7.4e-05
	positive regulation of cell killing	7.4e-05
	antigen processing and presentation	8.1e-05
	hemopoiesis	8.7e-05
	antigen processing and presentation of peptide antigen via MHC class I	9.0e-05
	regulation of leukocyte mediated cytotoxicity	1.1e-04
	antimicrobial humoral response	1.1e-04
	cation transport	1.2e-04
	hematopoietic or lymphoid organ development	1.2e-04
	regulation of T cell mediated immunity	1.3e-04
	immune system development	1.4e-04
	positive regulation of cytokine production involved in immune response	1.5e-04
	positive regulation of cell activation	1.8e-04
	positive regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	2.0e-04
	negative regulation of innate immune response	2.0e-04
	positive regulation of immune response	2.3e-04
	positive regulation of adaptive immune response	2.3e-04
	regulation of cell killing	2.3e-04
	metal ion homeostasis	2.4e-04
	transition metal ion transport	2.6e-04
	antigen processing and presentation of exogenous peptide antigen	2.6e-04
	positive regulation of lymphocyte mediated immunity	2.6e-04
	regulation of response to stress	2.7e-04
	myeloid cell differentiation	2.8e-04
	T cell mediated immunity	3.0e-04
	immune system process	3.2e-04
	lymphocyte mediated immunity	3.2e-04
	antigen processing and presentation of exogenous antigen	3.3e-04
	leukocyte mediated cytotoxicity	3.7e-04
	adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	3.9e-04
	negative regulation of immune effector process	4.2e-04
	ion transport	4.4e-04
	maintenance of location in cell	4.6e-04
	positive regulation of leukocyte mediated immunity	4.6e-04
	cytokine production involved in immune response	5.1e-04
	positive regulation of production of molecular mediator of immune response	5.1e-04
	regulation of cytokine production involved in immune response	5.1e-04
	iron ion homeostasis	5.1e-04
	positive regulation of T cell activation	5.4e-04
	homeostatic process	5.5e-04
	cation homeostasis	5.9e-04
	positive regulation of immune system process	6.2e-04
	regulation of lymphocyte mediated immunity	6.2e-04
	positive regulation of response to stimulus	6.7e-04
	immune effector process	6.8e-04
	negative regulation of response to biotic stimulus	6.8e-04
	positive regulation of leukocyte cell-cell adhesion	6.9e-04
	inorganic ion homeostasis	7.2e-04
	leukocyte mediated immunity	7.6e-04
	ion homeostasis	7.7e-04
	leukocyte differentiation	8.0e-04
	regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	8.1e-04
	metal ion transport	9.6e-04
	adaptive immune response	9.6e-04
4	**Positive**	
	positive regulation of multicellular organismal process	2.6e-05
	regulation of response to stimulus	3.0e-04
	sensory perception	4.2e-04
	system process	4.7e-04
	tissue homeostasis	5.6e-04
	retina homeostasis	5.9e-04
	amyloid fibril formation	5.9e-04
	positive regulation of receptor-mediated endocytosis	6.7e-04

	regulation of multicellular organismal process	8.1e-04
	response to vitamin	9.3e-04
	anatomical structure homeostasis	9.6e-04
	Negative	
	cytoplasmic translation	2.4e-26
	translation	6.1e-22
	peptide biosynthetic process	9.3e-22
	cellular macromolecule biosynthetic process	1.2e-20
	amide biosynthetic process	1.5e-20
	peptide metabolic process	1.5e-19
	cellular amide metabolic process	6.0e-18
	organonitrogen compound biosynthetic process	5.2e-15
	cellular macromolecule metabolic process	1.6e-11
	macromolecule biosynthetic process	2.4e-10
	cellular nitrogen compound biosynthetic process	8.7e-10
	cellular biosynthetic process	3.8e-08
	organic substance biosynthetic process	5.7e-08
	ribosome biogenesis	7.6e-08
	rRNA processing	9.0e-08
	biosynthetic process	9.6e-08
	protein metabolic process	1.8e-07
	gene expression	2.1e-07
	cellular nitrogen compound metabolic process	4.7e-07
	rRNA metabolic process	5.2e-07
	ribosomal small subunit biogenesis	6.6e-07
	ribonucleoprotein complex biogenesis	1.2e-06
	ncRNA processing	3.5e-06
	organonitrogen compound metabolic process	8.0e-06
	ribosome assembly	7.2e-05
	ncRNA metabolic process	9.3e-05
	non-membrane-bounded organelle assembly	1.0e-04
	ribosomal large subunit assembly	1.1e-04
	regulation of ubiquitin protein ligase activity	1.6e-04
	negative regulation of ubiquitin-protein transferase activity	2.1e-04
	translational elongation	3.4e-04
	negative regulation of protein ubiquitination	4.9e-04
	regulation of cellular macromolecule biosynthetic process	6.0e-04
	negative regulation of protein modification by small protein conjugation or removal	6.0e-04
	ribonucleoprotein complex assembly	8.2e-04
	regulation of ubiquitin-protein transferase activity	8.7e-04
	ribonucleoprotein complex subunit organization	9.9e-04
5	**Positive**	
	regulation of proteolysis	1.5e-04
	receptor clustering	1.5e-04
	maintenance of location	2.0e-04
	glycolytic process	2.6e-04
	ATP generation from ADP	2.6e-04
	ADP metabolic process	2.9e-04
	hexose metabolic process	3.1e-04
	amyloid fibril formation	3.7e-04
	monosaccharide metabolic process	3.8e-04
	maintenance of location in cell	3.8e-04
	regulation of protein metabolic process	3.9e-04
	nucleoside diphosphate phosphorylation	4.0e-04
	purine nucleoside diphosphate metabolic process	4.0e-04
	purine ribonucleoside diphosphate metabolic process	4.0e-04
	negative regulation of neuron projection development	4.0e-04
	nucleotide phosphorylation	4.0e-04
	catabolic process	4.2e-04
	pyruvate metabolic process	4.4e-04
	negative regulation of protein metabolic process	4.6e-04
	negative regulation of endopeptidase activity	5.0e-04
	muscle contraction	5.3e-04
	ribonucleoside diphosphate metabolic process	5.3e-04
	negative regulation of peptidase activity	6.0e-04
	amyloid-beta clearance	7.4e-04
	nucleoside diphosphate metabolic process	7.5e-04
	humoral immune response	8.8e-04
	negative regulation of intrinsic apoptotic signaling pathway	8.8e-04
	positive regulation of lipid localization	9.9e-04
	regulation of endopeptidase activity	9.9e-04

Negative

cardiac muscle cell action potential involved in contraction	3.6e-04
regulation of cardiac muscle cell action potential	3.6e-04
regulation of cardiac muscle cell contraction	4.4e-04
regulation of actin filament-based movement	5.1e-04
cell communication involved in cardiac conduction	6.0e-04
regulation of action potential	6.0e-04
cardiac muscle cell action potential	8.9e-04

Component	KEGG Pathway	p-value
1	**Positive**	
	Toll-like receptor signaling pathway	1.4e-02
	ECM-receptor interaction	3.1e-02
	Negative	
	ECM-receptor interaction	5.5e-10
	Protein digestion and absorption	1.8e-09
	Focal adhesion	4.1e-09
	Amoebiasis	4.8e-05
	Bacterial invasion of epithelial cells	2.5e-02
	Leukocyte transendothelial migration	4.3e-02
2	**Positive**	
	Vascular smooth muscle contraction	7.1e-09
	Focal adhesion	2.3e-03
	Tight junction	3.7e-03
	Gastric acid secretion	6.6e-03
	Hypertrophic cardiomyopathy	8.7e-03
	Dilated cardiomyopathy	9.5e-03
	Arrhythmogenic right ventricular cardiomyopathy	1.1e-02
	Regulation of actin cytoskeleton	1.5e-02
	Viral myocarditis	1.8e-02
	Leukocyte transendothelial migration	4.3e-02
	Negative	
	ECM-receptor interaction	1.3e-02
	Porphyrin metabolism	2.6e-02
3	**Positive**	
	NA	*NA*
	Negative	
	Viral myocarditis	2.9e-04
	Antigen processing and presentation	5.1e-04
	Porphyrin metabolism	1.6e-03
	Autoimmune thyroid disease	9.1e-03
	Allograft rejection	9.1e-03
	Graft-versus-host disease	9.1e-03
	Type I diabetes mellitus	1.2e-02
	Natural killer cell mediated cytotoxicity	3.5e-02
	Phagosome	3.8e-02
	Ribosome	4.1e-02
4	**Positive**	
	ECM-receptor interaction	5.10e-04
	Focal adhesion	5.92e-03
	Amoebiasis	9.59e-03
	Negative	
	Ribosome	1.9e-23
5	**Positive**	
	Glycolysis / Gluconeogenesis	6.40e-04
	Porphyrin metabolism	3.42e-03
	Negative	
	Arrhythmogenic right ventricular cardiomyopathy	3.20e-03
	RNA degradation	4.58e-02
	Pathways in cancer	4.99e-02