RESEARCH ARTICLE

# Directionally dependent multi-view clustering using copula model

**Kahkashan Afrin**[ID]$^{1}$, **Ashif S. Iquebal**$^{1}$, **Mostafa Karimi**$^{2}$, **Allyson Souris**$^{3}$, **Se Yoon Lee**$^{3}$, **Bani K. Mallick**$^{3}$*

**1** Department of Industrial & Systems Engineering, Texas A&M University, College Station, TX, United States of America, **2** Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, United States of America, **3** Department of Statistics, Texas A&M University, College Station, TX, United States of America

☉ These authors contributed equally to this work.
* bmallick@stat.tamu.edu

## Abstract

Recent developments in high-throughput methods have resulted in the collection of high-dimensional data types from multiple sources and technologies that measure distinct yet complementary information. Integrated clustering of such multiple data types or multi-view clustering is critical for revealing pathological insights. However, multi-view clustering is challenging due to the complex dependence structure between multiple data types, including directional dependency. Specifically, genomics data types have pre-specified directional dependencies known as the central dogma that describes the process of information flow from DNA to messenger RNA (mRNA) and then from mRNA to protein. Most of the existing multi-view clustering approaches assume an independent structure or pair-wise (non-directional) dependence between data types, thereby ignoring their directional relationship. Motivated by this, we propose a biology-inspired Bayesian integrated multi-view clustering model that uses an asymmetric copula to accommodate the directional dependencies between the data types. Via extensive simulation experiments, we demonstrate the negative impact of ignoring directional dependency on clustering performance. We also present an application of our model to a real-world dataset of breast cancer tumor samples collected from The Cancer Genome Altas program and provide comparative results.

## Introduction

The advancements in high throughput technologies and the emergence of several supportive programs such as The Genome Technology Program at the National Human Genome Research Institute and The Cancer Genome Atlas program have enabled the capabilities for rapid, high-quality, and low-cost collection of genomics data from multiple sources [1, 2]. These data types, collected from several heterogeneous sources for the same set of objects or patients, often provide unique but complementary information. They can be thought of as providing different views for the same underlying phenomenon (with each data type representing a particular view) and thus are referred to as "multi-view" datasets.

With this explosion of data, there is a strong need for integrated analysis of multi-view data to not only provide an immense amount of added information for making inference about the objects but also to explore and utilize the complex associations between multiple data types [3]. Hence, an integrated analysis of multi-view data has emerged as one of the promising areas of research. In several studies, integration of multiple data types has been shown to provide more comprehensive and radically new perspectives for understanding molecular pathways and the progression of diseases such as cancer as compared to analyzing individual data types separately [3–5].

In this research, we are concerned with proposing an effective approach for an integrated multi-view clustering where the objective is to group a set of patients, based on different genomic data types—an emerging field with only a few publications so far [6]. Following the existing literature in this domain, we refer to this integrative clustering as vertical multi-view clustering or consensus clustering [3, 6, 7].

Existing studies on vertical multi-view clustering can be primarily categorized into the following two groups [3]:

i.  Separate or source-specific clustering of each data type followed by post-hoc integration of the clustering outcomes often without incorporating any association between the data types [8, 9]. This is also referred to as late integration or ensemble clustering [10].

ii.  Concatenating the data types prior to clustering (or early integration) to obtain a single or unified model using the concatenated/joint data [11].

The two-stage late integration approach often fails to explore and exploit the association between different data types by assuming no dependence structure. On the other hand, early integration with concatenated data can have scaling and high-dimensionality issues and fail to recognize the individual contribution of each data type [3].

Hence, effective multi-view data integration methods are required that accommodates the dependence structure across the data types. Incorporating such dependencies between the multiple data types has been shown to encapsulate comprehensive information and deep understanding [5, 12]. However, effective multi-view data integration to capture the dependence between multiple data types remains a key challenge [13].

Recently, authors in [7] proposed an approach for integrative analysis via multiple dataset integration (MDI) by modeling each dataset (or data types) using a Dirichlet-multinomial mixture model and the association between the data types was captured by using the pairwise dependence between the clusters. Their method allowed for the identification of groups of genes that often fell together in one cluster. However, it did not provide a direct route to obtain the overall clustering, which is often of interest in practical applications. Along the similar lines, authors in [3] proposed a statistical model, called Bayesian consensus clustering (BCC), for integrating two or more data types. Similar to MDI, BCC also assumes a Dirichlet-multinomial mixture model for the data types. However, their approach was based on defining a source-specific cluster (i.e., separately clustering the objects for each of the datasets) as well as a consensus (i.e., overall) clustering. The dependence between the data types was captured by defining a parameter that controls the adherence of the source-specific clustering to the consensus clustering. They also emphasized on the computational scalability and robustness of the Bayesian framework for simultaneously estimating the consensus clustering as well as the source-specific clustering as compared to both late and early integration [3].

These recent studies on integrative analysis have focused both on the issue of source-specific and provides an effective way of consensus clustering while accounting for the dependence between multiple data types. However, the dependence between different data types is

extremely complex and is governed by the underlying molecular biology (in the case of genomics data). Specifically, genomics data types have dependencies that are often directional. For example, the central dogma of molecular biology [14, 15] describes the flow of information from DNA to messenger RNA (mRNA) through transcription and then from mRNA to protein through translation [16, 17], making these data types directionally dependent. Further, the central dogma explains that this transfer of information is acyclic (with a pre-specified direction), for example, the transfer of information from gene to protein and not from protein to protein or protein to gene [18].

This directional relationship between the data types is crucial for explaining physiological traits and clinical outcomes [19]. Furthermore, it has been analytically proven that the more remote an omics level or data source is from a physiological trait, the smaller the magnitude of their correlation is. For instance, the proteome-trait (protein level) correlation test is more powerful than the transcriptome-trait (RNA level) correlation test, which in turn is more powerful than the genotype-trait (DNA level) correlation test [19]. Nonetheless, the current multi-view clustering studies in the literature do not incorporate this directional information into their modeling. To address this gap, we propose a biology-inspired integrated multi-view clustering model called Bayesian directional multi-view clustering that incorporates the directional dependence between the data types using a copula function.

Copulas are multivariate distribution functions that allow us to model the dependence structure by considering the marginals [20]. Owing to the modeling flexibility provided by copulas, they have been used extensively in the literature for obtaining the dependencies between the data types. For instance, [21] used a Gaussian copula to construct a Dirichlet prior mixture of multivariate distributions to perform dependency-seeking clustering and showed significant improvement in the clustering results. Nonetheless, to the best of our knowledge, no existing work in the multi-view clustering has employed directional dependencies.

In this work, we obtain the directional dependence between the data types using an asymmetric copula regression. The use of asymmetric copula is crucial for modeling the directionality as symmetric copulas can only provide the directional dependence in the marginal behaviour but not in the joint behavior, as pointed out by [22]. Therefore, symmetric copulas may not be used to capture directional dependence. Here, we used the Rodriguez-Lallena and Ubeda-Flores [23] family of asymmetric copulas to capture the directional dependence. Further, we analyzed the asymmetric copulas from a regression perspective that allows us to obtain not only the direction of dependence between the data types but also to quantify this directional dependence [22].

To evaluate the efficacy, we applied our model to both synthetic as well as a benchmark real-world dataset. Using the results, we demonstrate that modeling the directional dependence between different datasets improves the clustering performance. For the real-world application, we used the dataset of breast cancer tumor samples that is publicly available from The Cancer Genome Atlas (TCGA) program (https://www.cancer.gov/tcga) [2, 3].

The rest of the article is organized as follows: In the **Methods** section, we describe the background on the Dirichlet mixture model for the integrative analysis and copula to capture directional dependence. The **Copula-based multi-view clustering** section presents our copula-based multi-view clustering approach and the posterior inference. We describe the simulation and case study examples along with the results and comparative analysis in the **Results** section. Finally, we conclude the paper with a brief discussion and future works in the **Conclusion** section.

## Methods

In this section, we briefly introduce the Dirichlet mixture model and extend it to a multi-view setting. Next, we present an overview of copulas and discuss a copula framework for capturing the directional dependence between multiple data types.

### Dirichlet mixture models

Let $\{\mathbf{X}_m\}_{m=1}^M$ denote a collection of $M$ distinct data types for $N$ objects (e.g., patients with different cancer tumor), and for each data type $m$, the notation $X_{mi}$ represents the data corresponding to the $i$-th object. For example, if the first data source (indexed by $m = 1$) is RNA gene expression, then $X_{1i}$ denotes the RNA gene expression profile for the $i$-th patient. In the context of this work, we assume that each data type is available for all $N$ objects. For each data type, $m$, let $L_{mi} = \{1, 2, \ldots, K\}$ denote a latent variable corresponding to the data $X_{mi}$ such that $L_{mi} = k$ implies that the $i$-th object belongs to the $k$-th cluster.

Since the objective of multi-view clustering is to partition the $N$ objects into $K$ clusters using the integrated data, it is intuitive to consider that the data $X_{mi}$ for each data type $m$ is generated from a mixture density given as [24, 25]:

$$p_m(X_{mi}) = \sum_{k=1}^K \pi_{mk} f(X_{mi}|\theta_{mk}) \tag{1}$$

where $\pi_{mk} = \Pr[L_{mi} = k] \in [0, 1]$ represents the probability of $X_{mi}$ belonging to the $k$-th cluster, and $f(X_{mi}|\theta_{mk})$ is a probability density function for the data $X_{mi}$ indexed by the parameters $\theta_{mk}$. If $f(\cdot)$ is chosen to be a Gaussian density $N(\mu_{mk}, \sigma_{mk}^2)$ with mean $\mu_{mk}$ and variance $\sigma_{mk}^2$, then we have $\theta_{mk} = (\mu_{mk}, \sigma_{mk})$, leading to a Gaussian mixture model [26]. For a detailed explanation for the mixture model, refer to [25]. With this, a hierarchical structure of Dirichlet mixture model for each data source $m$ may be obtained as [27–29]:

$$X_{mi}|L_{mi}, \theta_m \sim F(\theta_{mL_{mi}}); \quad \theta_{mk} \sim G^{(0)} \tag{2}$$

$$L_{mi}|\pi_m \sim \text{Multinomial}(\pi_{m1}, \pi_{m2}, \ldots \pi_{mK}) \tag{3}$$

$$\pi_{m1}, \pi_{m2}, \ldots \pi_{mK} \sim \text{Dirichlet}(\alpha/K, \ldots. \alpha/K) \tag{4}$$

where $F$ is the (cumulative) distribution function for the density $f(\cdot)$ participating in the mixture density in Eq (1), $G^{(0)}$ denotes a base measure for the parameter $\theta_{mk}$, and $\alpha > 0$ denotes the scaling parameter for the Dirichlet distribution in Eq (4). For a detailed description for the Dirichlet mixture model, and its practical implementation for clustering, refer to [30].

### Directional dependency and copula

One of the fundamental ideas of molecular biology is the central dogma [14, 15], which describes the process of information flow via a two-step process of transcription and translation, where the information in genes flow to proteins: DNA to mRNA, and mRNA to protein. Further, the central dogma explains that this transfer of information is acyclic (with a pre-specified direction), for example, the transfer of information is from gene to protein and not from protein to protein or protein to gene [18].

The key motivation of our research is to subsume this *directional dependency* into a multi-view clustering problem [3, 17, 31], leading to a unified and directionally dependent clustering model.

Directional dependence was first presented by authors in [32] using an ordinary linear regression model. Under this framework, a random variable $X$ is directionally dependent on $Y$, if the square of sample skewness of $Y$, denoted by $\hat{\gamma}_Y^2$, is less than that of $X$ when we regress $X$ on $Y$. That is, $\hat{\gamma}_Y^2 < \hat{\gamma}_X^2$ when $X = \beta Y + \epsilon$. The fundamental notion here is that the square of the skewness of the response variable in a linear regression setting is always less than equal to the square of the skewness of the explanatory variable. More recently, authors in [22, 33–35] argued that copula regression models might offer a possibility to capture the directional dependence between variables. This is mostly because the copula regression approach can model the joint dependence structure between the random variables, independently from the choice of the marginal distributions.

In this paper, we make use of copulas [36, 37] to accommodate directional dependency into a multi-view clustering framework. Copulas are widely used to model the dependence structure between random variables by decoupling the dependence structure from the marginal distribution [38]. Specifically, an $n$-dimensional copula function $C = C(u_1, u_2, \ldots, u_n)$ is a multivariate distribution function defined on a unit hypercube with uniform marginals given as:

$$C(u_1, u_2, \ldots, u_n) = P(U_1 \leq u_1, \ldots, U_n \leq u_n) \tag{5}$$

where $U_i \sim \text{uniform}(0, 1)$, $i = 1, 2, \ldots, n$. Given a vector of random variables denoted by $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ with joint distribution function given as:

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n),$$

the uniqueness of the copula associated with $F$ was initially observed in *Sklar's theorem* [39].

**Theorem 1** (Sklar's theorem). *Let F be an n-dimensional joint distribution function with margins $F_1, \ldots, F_n$. Then there exists an n-dimensional copula C that satisfies the following equality for all $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$:*

$$F(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n)) \tag{6}$$

*Additionally, if all the marginals are continuous, then C is unique. Conversely, if C is a copula, and all the margins $F_1, \ldots, F_n$ are univariate distribution functions, then the function F that satisfies the above equation is a joint distribution function with margins $F_1, \ldots, F_n$.*

Using Eq (6), we can now represent the joint density of multivariate random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ in terms of a copula function and the marginals. Thus, copula offers a simple, yet powerful approach to sample from the joint distribution of random variables with known marginals.

## Copula for directional dependency

The key idea behind accommodating directional dependency between two random variables, say $U$ and $V$, using copula is to construct an *asymmetric* copula [40]. Formally, a bivariate asymmetric copula $C(u, v) : [0, 1]^2 \rightarrow [0, 1]$ is defined by any copula function that satisfies $C_{U|V}(u, v) \neq C_{V|U}(v, u)$, where $C_{U|V}(u, v) = \partial C(u, v)/\partial v$ and $C_{V|U}(v, u) = \partial C(v, u)/\partial u$. In this paper, we consider an asymmetric copula from the Rodriguez-Lallena and Ubeda-Flores family and is given as [23]:

$$C(u, v, \phi) = uv + f(u)g(u) = uv + \vartheta uv(1 - u)^\alpha (1 - v)^\beta \tag{7}$$

where $u, v \in [0, 1]$, and $\phi = (\vartheta, \alpha, \beta)$, $\alpha > 1, \beta > 1$. The association parameter $\vartheta \in [-1, 1]$ measures the dependence between $u$ and $v$, while the asymmetry in the copula is captured by the parameters $\alpha$ and $\beta$. Given $n$ observations $\{(u_i, v_i)\}_{i=1}^n$, the maximum likelihood estimates

(MLE) of $\alpha$ and $\beta$ are given as:

$$\begin{aligned}
\hat{\alpha} &= \frac{\sum_{i=1}^{n}((1-u_i)\log(1-u_i)-u_i)}{\sum_{i=1}^{n}u_i\log(1-u_i)} \\
\hat{\beta} &= \frac{\sum_{i=1}^{n}((1-v_i)\log(1-v_i)-v_i)}{\sum_{i=1}^{n}v_i\log(1-v_i)}
\end{aligned} \tag{8}$$

Given $\alpha$ and $\beta$, the admissibility bound for $\vartheta$ as shown by [41] is given as:

$$-\min\left\{\left(\frac{\alpha+1}{\alpha-1}\right)^{\alpha-1},\left(\frac{\beta+1}{\beta-1}\right)^{\beta-1}\right\}\leq\vartheta\leq\min\left\{\left(\frac{\alpha+1}{\alpha-1}\right)^{\alpha-1},\left(\frac{\beta+1}{\beta-1}\right)^{\beta-1}\right\} \tag{9}$$

Let $\rho_{v\to u}$ denote the degree of directional dependency of $U$ on $V$, with a higher value indicates a stronger directional dependency. Adopting the idea of [22], $\rho_{v\to u}$ for the copula in Eq (7) can be expressed as:

$$\rho_{v\to u}=12\cdot E[(r_{U|V}(u))^2]-3 \tag{10}$$

where the expectation $E[\cdot]$ is taken with respect to the copula regression function $r_{U|V}(u)$ defined by $r_{U|V}(u)=1-\int_0^1 C_{U|V}(u,v)dv$. For the Rodriguez-Lallena and Ubeda-Flores copula, we can express Eq (10) in a closed-form [22, 33]:

$$\rho_{v\to u}=\frac{3\vartheta^2\alpha^2\beta^2}{(2+\beta)^2(1+2\alpha)}\quad\text{and}\quad\rho_{u\to v}=\frac{3\vartheta^2\alpha^2\beta^2}{(2+\alpha)^2(1+2\beta)} \tag{11}$$

With this copula approach to model directional dependency, we now extend the Bayesian framework presented in Eqs (2)–(4) to incorporate directional dependency for multi-view clustering.

## Copula-based multi-view clustering

In this section, we present our Dirichlet mixture model to incorporate directional dependence for multi-view clustering followed by the details of posterior inference for our model.

### Copula-based dirichlet mixture model

The Dirichlet mixture model as presented in Eqs (2)–(4) has been constructed for each data source, $m$, individually. Hence, there is no information borrowing between datasets. In what follows, we make use of the formula of directional dependency between a pair of data sets (see Eq (11)), and subsequently utilize these quantities in clustering through the Dirichlet mixture model.

A hierarchical formulation for a copula-based Dirichlet mixture model based on the Rodri-guez-Lallena and Ubeda-Flores copula is given as:

$$X_{mi}|L_{mi}, \theta \sim F(\theta_{mL_{mi}}); \quad \theta_{mk} \sim G^{(0)} \tag{12}$$

$$p(L_{1i}, \ldots L_{Mi}|\gamma, \rho) \propto \prod_{m=1}^{M} \gamma_{mL_{mi}} \prod_{m=1}^{M} \prod_{k=1, k \neq m}^{M} \left(1 + \rho_{m \to k} \mathbb{1}_{\{m \to k\}} \mathbb{1}_{\{L_{mi} = L_{ki}\}}\right) \tag{13}$$

$$\gamma_{m1}, \gamma_{m2}, \ldots, \gamma_{mK} \sim \text{Gamma}(\alpha_m/K, 1) \tag{14}$$

$$\rho_{m \to k} \sim H_{m,k} \tag{15}$$

where Gamma($a$, $b$) denotes the gamma distribution with shape parameter $a$ and rate parameter $b$, and $\mathbb{1}_{\{m \to k\}}$ is an indicator function such that $\mathbb{1}_{\{m \to k\}} = 1$ (otherwise zero) if there exists a known directional dependency from the $m$-th dataset towards the $k$-th dataset, while $\rho_{m \to k}$ is the degree of directional dependency associated with the two data types (see Eq (11). Notations $F$, $G^{(0)}$, and $\theta_{mk}$ are used in the same way as used in the Dirichlet mixture model (Eqs (2)–(4)). The weights $\gamma_{mL_{mi}}$ are derived from $\pi_{mk}$ as shown in Eq (3) such that $\pi_{mj} = \gamma_{mj} / \sum_{k=1}^{K} \gamma_{mk}$. Finally, $H_{m,k}$ is the prior on the parameter $\rho_{m \to k}$ and it will be discussed in the following sections.

From the foregoing model, we note that: (a) the major distinction between the proposed model (Eqs (12)–(15)) and the standard Dirichlet mixture model (Eqs (2)–(4)) is observed by the clustering allocation procedure induced by the latent variables $L_{1i}, L_{1n}, \ldots, L_{Mi}$ and (b) the integration of directional dependency via $\rho_{u \to v}$ provides an advancement over the existing integrative models such as [3, 7].

## Directional dependence prior

For each $i$, given two data types $u$ and $v$, let the notations $\hat{\alpha}_{uv}^{i}$ and $\hat{\beta}_{uv}^{i}$ denote the MLEs of the two parameters, $\alpha$ and $\beta$, of the Rodriguez-Lallena and Ubeda-Flores copula, given by Eq (8). Then after averaging each of the quantities over the $n$ observations, that is, $\hat{\alpha}_{uv} = \frac{1}{n} \sum_{i=1}^{n} \hat{\alpha}_{uv}^{i}$ and $\hat{\beta}_{uv} = \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}_{uv}^{i}$, we can express the directional dependence of the $v$-th dataset on the $u$-th dataset as follows [33]:

$$\rho_{u \to v} = \frac{3\vartheta_{uv}^2 \hat{\alpha}_{uv}^2 \hat{\beta}_{uv}^2}{(2 + \hat{\alpha}_{uv})^2 (1 + 2\hat{\beta}_{uv})} \tag{16}$$

where $\vartheta_{uv}$ is a measure of association between the random variables $U$ and $V$. Note that $\rho_{u \to v}$ in Eq (16) is a *copula-based random variable*: first, two quantities $\hat{\alpha}_{uv}$ and $\hat{\beta}_{uv}$ are driven from a directional copula; and second, $\vartheta_{uv}$ is the stochastic part, rendering $\rho_{u \to v}$ as a random quantity.

In this work, we consider a Gaussian distribution for the random variable $\vartheta_{uv}$, given by $\vartheta_{uv} \sim N(0, b_{uv}^2/9)$, where the $b_{uv}$ is obtained by the admissibility bound presented in Eq (9) as:

$$b_{uv} = \min_{i \in \{1, \ldots, n\}} \left\{ \left(\frac{\hat{\alpha}_{uv}^i + 1}{\hat{\alpha}_{uv}^i - 1}\right)^{\hat{\alpha}_{uv}^i - 1}, \left(\frac{\hat{\beta}_{uv}^i + 1}{\hat{\beta}_{uv}^i - 1}\right)^{\hat{\beta}_{uv}^i - 1} \right\} \tag{17}$$

The motivation behind this choice of $b_{uv}$ is as follows: as a Gaussian random variable, $\vartheta_{uv}$ has a variance of $b_{uv}^2/9$ such that $P(|\vartheta_{uv}| \leq b_{uv}) = 0.997$. This implies that $-b_{uv} \leq \vartheta_{uv} \leq b_{uv}$

holds with a very high probability. Since $-b_{uv} \leq \vartheta_{uv} \leq b_{uv}$ holds with a very high probability, the posterior updates are all conjugate updates without the added computational burdens of truncated distributions. Moreover, as $\vartheta_{uv}$ is normally distributed, it is easy to see that $\rho_{u \to v} \propto \vartheta_{uv}^2$ and follows a gamma distribution. In particular,

$$\rho_{u \to v} \sim \text{Gamma}\left(\frac{1}{2}, \frac{4(2 + \hat{\alpha}_{uv})^2(1 + 2\hat{\beta}_{uv})}{2\hat{\alpha}_{uv}^2 \hat{\beta}_{uv}^2 (b_{uv}/3)^2}\right) \tag{18}$$

This defines our prior $H_{u,v}$ on $\rho_{u \to v}$.

## Posterior updates

In this section, we present a general Bayesian framework to estimate the posterior updates using a Gibbs sampling approach. For this, we follow the posterior inference, as presented in [7]. We begin by referring to our general model in Eq (13). We first obtain the joint density of the latent allocation variable $L_{mi}$'s by defining a normalizing constant $Z$ as:

$$Z = \sum_{j_1=1}^{K} \cdots \sum_{j_M=1}^{K} \left[\prod_{k=1}^{M} \gamma_{kj_k} \prod_{k=1}^{M} \prod_{\ell=1,\ell\neq k}^{M} (1 + \rho_{k \to \ell} \mathbb{1}_{\{k \to \ell\}} \mathbb{1}_{\{j_k = j_\ell\}})\right] \tag{19}$$

such that the joint density for $N$ objects is given as:

$$p(\{L_{1i}, \ldots L_{Mi}\}_{i=1}^{N}) = \frac{1}{Z^N} \prod_{i=1}^{N} \left[\prod_{k=1}^{M} \gamma_{kj_k} \prod_{k=1}^{M} \prod_{\ell=1,\ell\neq k}^{M} (1 + \rho_{k \to \ell} \mathbb{1}_{\{k \to \ell\}} \mathbb{1}_{\{j_k = j_\ell\}})\right] \tag{20}$$

Following [42], we define the following joint density function using a strategic latent variable $\xi$ to provide the basis for a fully Bayesian framework:

$$p(\{L_{1i}, \ldots L_{Mi}\}_{i=1}^{N}, \xi) = \frac{\xi^{N-1} \exp(\xi Z)}{(N-1)!} \times p(\{L_{1i}, \ldots L_{Mi}\}_{i=1}^{N}) \tag{21}$$

such that the conditional distribution for the latent variable is given as:

$$\xi | - \sim \text{Gamma}(N, Z) \tag{22}$$

Subsequently, we note that the conditional on $\gamma_{mj_m}$ is given as:

$$\gamma_{mj_m} | - \sim \text{Gamma}(a_\gamma, b_\gamma) \tag{23}$$

$$a_\gamma = \sum_{i=1}^{N} \mathbb{1}\{L_{mi} = j_m\} + \frac{\alpha_M}{K} \tag{24}$$

$$b_\gamma = \xi \sum_{j_1=1}^{K} \cdots \sum_{j_{m-1}=1}^{K} \sum_{j_{m+1}=1}^{K} \cdots \sum_{j_M=1}^{K} \left[\prod_{k=1,k\neq m}^{M} \gamma_{kj_k} \prod_{k=1}^{M} \prod_{\ell=k,\ell\neq k}^{M} (1 + \rho_{k \to \ell} \mathbb{1}_{\{k \to \ell\}} \mathbb{1}_{\{j_k = j_\ell\}})\right] + 1 \tag{25}$$

In a similar fashion, the conditional on $\rho_{m \to p}$ may be deduced as:

$$\rho_{m \to p}|- \sim \text{Gamma}(a_\rho, b_\rho) \tag{26}$$

$$a_\rho = \sum_{i=1}^{N} \mathbb{1}\{L_{mi} = L_{pi}\} + \frac{1}{2} \tag{27}$$

$$b_\rho = \xi \sum_{j_1=1}^{K} \cdots \sum_{\substack{j_{m-1}=1 \\ j_{m+1}=1}}^{K} \cdots \sum_{\substack{j_{p-1}=1 \\ j_{p+1}=1}}^{K} \cdots \sum_{j_M=1}^{K} \left[ \prod_{k=1}^{M} \gamma_{kj_k} \prod_{k=1}^{M} \prod_{\substack{\ell=1 \\ \ell \neq k, \ell \neq p}}^{M} (1 + \rho_{k \to \ell} \mathbb{1}_{\{k \to \ell\}} \mathbb{1}_{\{L_{ki}=L_{\ell i}\}}) \times \right.$$
$$\left. \prod_{\substack{k=1 \\ k \neq m, k \neq p}}^{M} (1 + \rho_{k \to p} \mathbb{1}_{\{k \to p\}} \mathbb{1}_{\{L_{ki}=L_{pi}\}}) \right] + \frac{4(2 + \hat{\alpha}_{mp})^2(1 + 2\hat{\beta}_{mp})}{2\hat{\alpha}_{mp}^2 \hat{\beta}_{mp}^2 (b_{mp}/3)^2} \tag{28}$$

Finally, the conditional distribution for the latent variables $L_{mi}$ is given as:

$$p(L_{mi} = c|\gamma, \quad \rho, x_{mi}, x_{m,-i}^c, L_{-m,i}, L_{m,-i}) = b_L \gamma_{mc} \prod_{k=1}^{m-1} (1 + \rho_{k \to m} \mathbb{1}_{\{k \to m\}} \mathbb{1}_{\{L_{ki}=L_{mi}\}}) \times$$
$$\prod_{k=m+1}^{M} (1 + \rho_{m \to k} \mathbb{1}_{\{m \to k\}} \mathbb{1}_{\{L_{mi}=L_{ki}\}}) \times \int f_m(x_{mi}, x_{m,-i}^c|\theta_m) g_m^{(0)}(\theta_m) d\theta_m \tag{29}$$

where $x_{mi}$ is observation $i$ for data type $m$, $x_{m,-i}^c$ are all the observations not including $x_{mi}$ in data type $m$ associated with component $c$, $L_{m,-i}$ are all the $L_{mj}$ such that $i \neq j$, $L_{-m,i}$ are all the $L_{ki}$ such that $k \neq m$, and $b_L$ is a normalizing constant that ensures $\sum_{c=1}^{K} p(L_{mi} = c|\gamma, \rho, x_{mi}, x_{m,-i}^c, L_{-m,i}, L_{m,-i}) = 1$. Please note that a latent variable $\xi$ has been added to help with the computational efficiency and that the updates on $\theta_m$ will depend on the choice of $f_m(\cdot)$ and $G^{(0)}$. The posterior updates are based on pairwise directional dependence. To obtain the consensus/global clustering, we leverage the fundamental idea of the central dogma, i.e., the directional dependence in genomic data. Based on prior studies [19], we note that the more remote an omics level or data source is from a physiological trait, the smaller the magnitude of their correlation is. Since protein data is closest to the physiological trait (i.e., cancer type), we deem the clustering result of the protein as the final or global clustering.

## Results

To demonstrate the effectiveness of incorporating directional dependencies in multi-view clustering, we consider both simulated and real-world examples. We also compare the performance of our approach with competing methods such as [3] and [7].

### Simulated data

For the simulation experiment, we consider two data types, $U$ and $V$, each generated from a univariate Gaussian mixture model with two components. The corresponding means and standard deviations of the mixture model are 0, 3 and 1, 0.5, respectively, and the directional dependency is captured by an asymmetric Tawn copula given as:

$$C(u, v) = \exp\left( \log(u) + \log(v) A\left( \frac{\log(v)}{\log(u)\log(v)} \right) \right) \tag{30}$$

where $A(.)$ is called the Pickands dependence. For the Tawn copula, the Pickands function is given as:

$$A(t) = (1 - \psi_1)(1 - t) + (1 - \psi_2)t + [(\psi_1(1 - t))^\theta + (\psi_2 t)^\theta]^{1/\theta} \tag{31}$$

In this case, we consider the Tawn copula of Type 1 for which $\psi_2 = 1$. For the current simulation study, we set the values of $\psi_1$ and $\theta$ to 0.5 and 30, respectively, and the direction of dependence is set from $V$ to $U$. Details on the Tawn copula can be found in [43]. For each data type, we generate 500 data points, and the measure of directional dependence (from $V$ to $U$) is estimated from Eq (11). For the Tawn copula constructed above, the directional dependence is equal to 1.73. Note that the dependence in the opposite direction was obtained to be -1.02, indicating no notable dependence.

We test the performance of the proposed methodology using three different scenarios:

i.  when the true directionality is used, i.e., $\mathbb{1}_{\{V \to U\}}$,

ii.  when there is no direction of dependence, and

iii.  when the direction of dependence is reversed, i.e., $\mathbb{1}_{\{U \to V\}}$.

In the first case, the method is able to correctly predict the 2 clusters and is shown in the joint similarity matrix in Fig 1(a). The similarity matrix displays the posterior probability of samples $i$ and $j$ to belong to the same cluster (see [7] for details). The overall accuracy of clustering for this case is 97.8%. The clustering results corresponding to cases (ii) and (iii) are presented in the joint similarity matrix shown in Fig 1(b) and 1(c). We note that the clustering performance is affected in both the cases, but more in case (iii), where we observe three different clusters as opposed to two clusters as in cases (i) and (ii). The intuition behind the significantly worse performance of case (iii) as compared to cases (i) and (ii) is that in case (iii), we wrongly reverse the directional dependence between the data types. In contrast, case (ii) is more reminiscent of the integrative analysis presented by [3, 7]. The clustering outcome is affected since the underlying models consider dependence without any directionality. Additionally, in case (ii), where no directionality is considered, we do get two clusters. However, the clustering accuracy is 97.2%, lower than that of the first case.

In addition to studying the effect of causal relations on the clustering performance, we also investigate the effect of different copulas and sample size. We first look at the effect of different copulas. In this simulation study, we refer to three different copulas: Tawn Type 1 (TT1), Tawn Type 2 (TT2), and BB1 copula. See [43] for the functional form and dependence structure. Note that by varying the copula model, we essentially modify the dependence structure. Results obtained from 10 simulation runs show that clustering accuracy for TT1, TT2, and BB1 copulas for each of the aforementioned cases (case (i), (ii), and (iii)) are (0.97, 0.96, 0.84), (0.95, 0.93, 0.94), and (0.97, 0.92, 0.94), respectively. The average number of clusters were recorded as (2, 2, 3.67), (2, 3, 3.4), and (2, 3, 2). We note that by incorporating the directional dependency, we are able to consistently achieve a higher accuracy as well as identify the two clusters. For case (ii), i.e., integrative clustering without any direction of dependence achieves marginally lower accuracy but fails to identify the two clusters correctly. Case (iii) performs worse both in terms of accuracy and the number of clusters.

Next, we look at the effect of the sample size. For this, we fix our copula to Tawn Type 1 and obtain the results for three different sample sizes: 250, 500, 750. Corresponding to these sample sizes, the clustering accuracy was noted as (0.97, 0.96, 0.84), (0.97, 0.97, 0.87), (0.97, 0.96, 0.64). In addition, the average number of clusters were obtained as (2, 2, 3.67), (2, 2, 3),
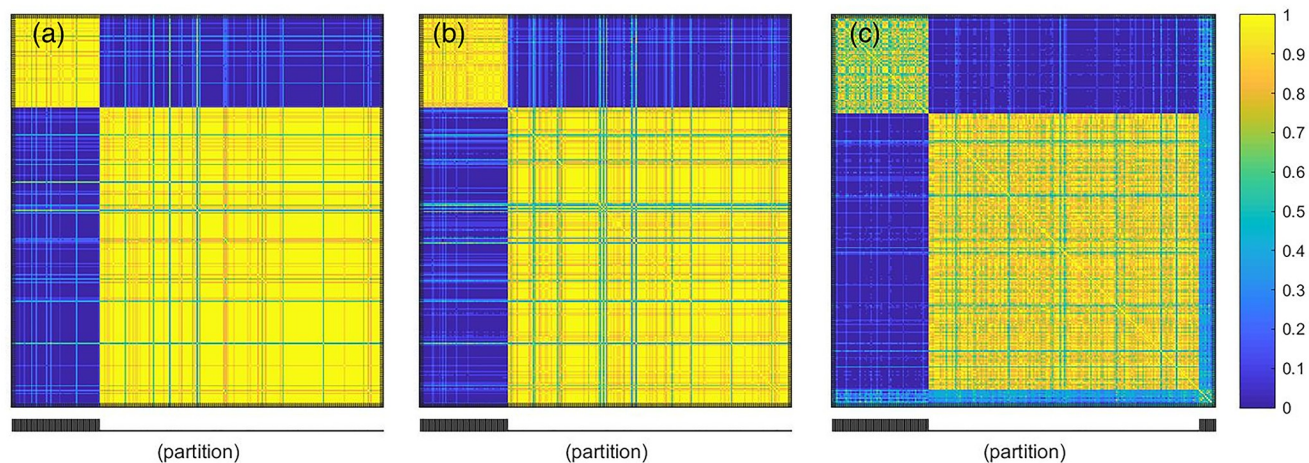
**Fig 1. Similarity matrix for the simulation study: (a) the true direction of dependence is considered, (b) no directional dependence is considered, (c) the direction of dependence is reversed.** The colormaps show the posterior probability of samples $i$ and $j$ to belong to the same cluster.

and (2, 2, 5). We again note that the proposed methodology performs better as compared to the case when no directionality is considered and when the direction is reversed.

We also note that the computational complexity scales linearly with the sample size. In fact, the computational complexity of the present method is directly proportional to the number of data sources $M$, the number of clusters $K$, and the number of genes $N$. So the algorithm scales as $O(NMK)$. For sample sizes of 250, 500, and 750, the algorithm converges in 167.2 sec, 364.5 sec, and 520.16 sec when running in parallel. Convergence of the method can also be argued from the standpoint of clustering results, i.e., as sufficient data is made available, the clusters estimated from the method will closely resemble the true clusters [3].

## TCGA breast cancer data

For the application of our model to a real-world dataset, we considered the breast cancer tumor samples from TCGA program that consists of multi-source genomic data for a common set of patients. Since breast cancer is a heterogeneous disease and can be effectively used in the case studies for clustering models. This dataset has become a benchmark dataset used in several multi-view clustering studies, such as [3]. This dataset is available for download from the web portal of TCGA (https://www.cancer.gov/tcga) and contains a common set of 348 breast cancer tumor samples (i.e., $N = 348$) and four distinct data types (i.e., $M = 4$):

- RNA gene expression (GE) data for 645 genes.

- DNA methylation (ME) data for 574 probes.

- miRNA expression (miRNA) data for 423 miRNAs.

- Reverse phase protein array (RPPA) data for 171 proteins.

We chose the 171 genes, probes, and miRNAs corresponding to the 171 existing proteins in the other three data types. These 171 genes (or their product proteins) are carefully chosen to contain the genes such as PIK3CA, PTEN, AKT1, TP53, GATA3, CDH1, RB1, MLL3, MAP3K1, and CDKN1B that are well-known to be important for classification of breast cancer subtypes [2].

It is known that these four data types manifest differently, but at the same time, are highly related in that they are directionally dependent [14, 15]. This directional dependence can be
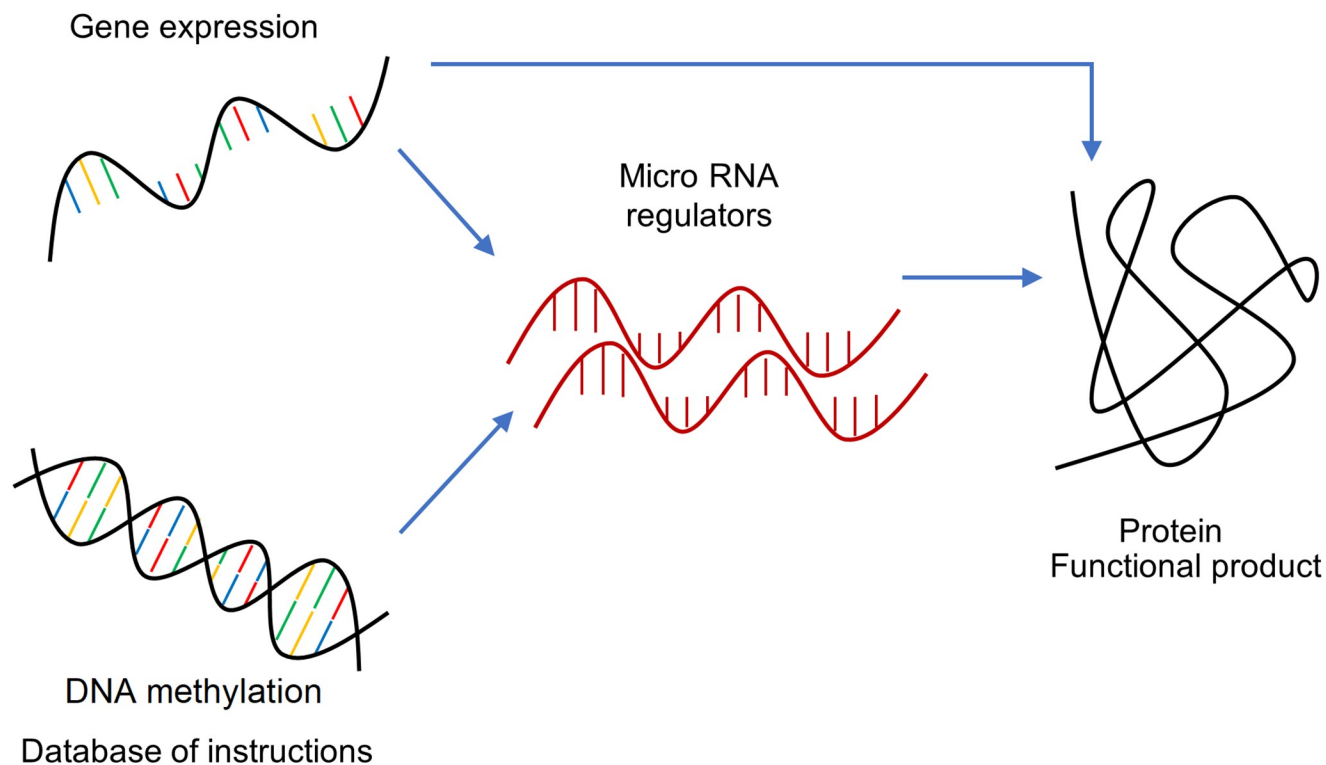
**Fig 2. The central dogma of molecular biology.** Directional dependencies among biological components.

determined by the central dogma of molecular biology, where the transcription and translation process determines the direction of dependence between DNA to RNA gene expression and from RNA gene expression to protein, respectively. Fig 2 shows the four data types used in this study, along with the direction of dependence between them. We consider three pathways for directional dependence. The first dependence pathway is from RNA gene expression to protein. This is the fundamental relationship, as explained by the central dogma. The second dependence pathway is from RNA gene expression to protein via miRNA (or microRNAs). MiRNAs are single-stranded RNAs that exert their regulatory action by binding RNAs gene expression and preventing their translation into proteins. The last pathway is from DNA methylation to protein via miRNAs. This dependence is primarily based on recent studies that have shown the influence of DNA methylation directly on the expression of miRNAs [44]. Note that these relationships are not exhaustive and additional experiments will be needed to fully model the pathways that influence the production of protein from DNA.

As discussed in the foregoing, the central dogma explains that this transfer of information has a pre-specified direction, for example, the transfer of information is from gene to protein and not from protein to protein or protein to gene [18]. From a statistical perspective, both *transcription* and *translation* might be designed in terms of directional dependency in our copula model because opposite dependencies (i.e., Protein to RNA and RNA to DNA) may not exist. Considering the two directions are *a priori* known, we can design them by providing deterministic directional indicators $\{k \rightarrow p\}$ for each process where $k$ and $p$ are indices for the corresponding data types. Also, the corresponding strengths of directional dependencies are quantified by $\rho_{k \rightarrow p}$. From a numerical perspective, we note that providing a deterministic

directional indicator into our model reduces the computational burden in the summation of $b_\rho$, and therefore, contributes to the computation speed.

Since we use the copula at the data level, not at the latent level, matching the data dimensions for each data type is necessary ($D_1 = D_2 = D_3 = D_4 = 171$). This is because we modeled the dependence between the data types through the directional relationship between their features (such as genes, RNAs, proteins, etc.). These four data types are measured on different platforms and represent different biological components. However, they all represent genomic data for the same sample set, and it is reasonable to expect shared structure while considering directional dependencies at hand [45].

As explained in the previous section, we are clustering samples based on four data types: gene expression, DNA methylation, microRNA, and RPPA for breast cancer from the TCGA data. Prior studies using this data have found that the total number of clusters can vary from two [46] to 10 [47]. However, mainly four prominent subtypes have been identified based on multi-source consensus clustering of the TCGA data as Basal, Luminal A, Luminal B, and HER2 [2]. We incorporated our prior biological knowledge, i.e., the directional dependence based on the central dogma [15, 18], into our integrative clustering algorithm. Since proteins are the final outcome, we consider the consensus (or final) clustering to be the protein clusters, which summarize all the information from the other three datasets inside itself.

To initialize the Bayesian posterior update algorithm, we take advantage of our finite Dirichlet mixture model to define the number of clusters ($K$). Although our model considers a finite mixture model, it is equivalent to a Dirichlet process mixture model when $N \rightarrow \infty$ and therefore, $K$ specifies an upper bound on the number of clusters present in the data. Authors in [48] argue that if the number of clusters specified by $K$ is sufficiently large, the posterior updates can automatically determine the true number of clusters present in the data. Based on this analogy, [7] suggest that a pragmatic choice for the upper bound on the number of cluster to avoid the computational burden is $K = \lceil N/2 \rceil$. From our experimental studies, we note that even if this upper bound is as high as 500, our algorithm correctly predicts the number of clusters to be four, similar to the sub-typing in TCGA. This shows that the clustering results are not contingent on the choice of $K$.

Since copy number variation or the focal amplification/deletion of a region of gene, is associated with breast cancer risk and prognosis [49–53], we calculate the fraction of the genome altered (FGA) as a measure of copy number activity as described in the Supplementary Section VII of [2] (with copy number level threshold $T = 0.15$) for each cluster. Our results are summarized in Table 1 and visualized in Fig 3. The TCGA breast cancer subtypes and our clusters have different structures, but they are non-independent based on the Chi-squared test of independence (p-value < 0.0001). Clusters 1 is mostly a combination of Luminal like breast cancer subtypes (Luminal A and Luminal B), which are similar to each other with average FGA of $0.19 \pm 0.15$ and almost 10% of their samples have high FGA of more than 0.4. Cluster 2 is

**Table 1. Confusion matrix for the clustering assignment.**

| Estimated cluster \ True cluster | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Her2 | 5 | 4 | 22 | 8 |
| Basal | 2 | 5 | 48 | 17 |
| Lum A | 41 | 103 | 2 | 15 |
| Lum B | 34 | 22 | 5 | 15 |

The results are shown for the clustering of four cancer data types: Her2, Basal, Lum A, and Lum B.

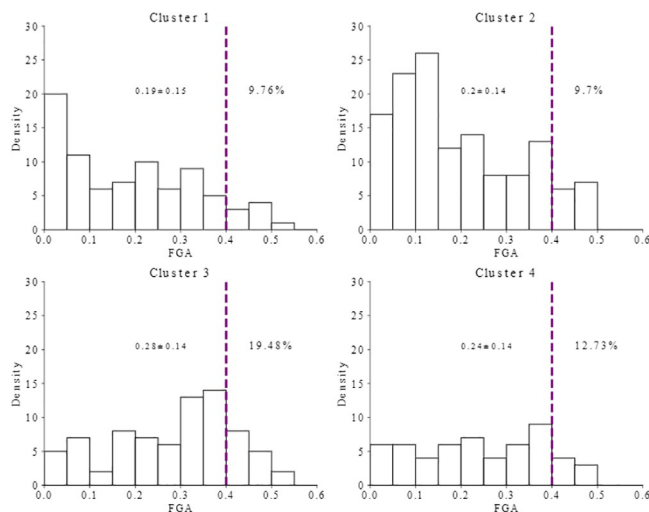https://doi.org/10.1371/journal.pone.0238996.t001

**Fig 3. Distribution of FGA across four clusters of breast cancer.** Dashed line represents the genes with high FGA values.

mostly Luminal like breast cancer B with higher FGA (0.2 ± 0.14). Moreover, cluster 2 contains similar high FGA in comparison to cluster 1 with almost 10%. Cluster 3 contains more of HER2 and Basal subtypes, which are more similar to each other with the highest FGA (0.28 ± 0.14) and very high FGA (approximately 19%). Even though cluster 4 is more spread over the four known subtypes (Her2, Basal, Luminal A and Luminal B), they include samples with high FGA (0.24 ± 0.14) and lower standard deviation compared to cluster 2. Moreover, cluster 4 is second in having high FGA samples with almost 13%.

As mentioned in the foregoing, two other state-of-the-art algorithms for integrative clustering exists: Bayesian Consensus Clustering (BCC) [3] and Multiple Dataset Integration (MDI) [7]. However, it is not feasible to compare our results with MDI as it provides separate clustering for each data type as opposed to a consensus or global clustering. To compare the performance of our method with BCC, we use the Rand index as the number of clusters reported by BCC is three, while four clusters exist in the TCGA dataset. Essentially, the Rand index measures the level of similarity between two clustering methods without employing the data labels [54]. When one of the clustering methods is the ground truth, it essentially measures the proportion of the correct allocations. We note that the Rand index of BCC is 0.68, and for the proposed method, it is 0.70. Not only BCC performs marginally worse in terms of correctly labeling the datasets, but also the algorithm needs to know the true number of clusters beforehand, which is a major limitation. The authors developed a heuristic approach to calculate the number of clusters as a pre-processing step that suggested only three clusters as opposed to four true clusters predefined in the TCGA dataset. In comparison, our integrated analysis method correctly identified the four clusters out of the maximum number 174 potential clusters ($K = \lceil N/2 \rceil$).

## Conclusion

It is known that the genomics data types collected from multiple sources are often related, and their integrated analysis can significantly improve the downstream analysis such as the clustering outcome. Several studies have been proposed in the literature for integrated analysis of multi-view data that attempts to capture the association or dependence between different data types. Nonetheless, the dependence between real-world data types often have many added

level of complexity due to the underlying natural phenomena. This is often true in genomics, where, underlined by the central dogma, the data types are not only dependent but directionally dependent. We utilized this domain knowledge and proposed a novel method for multi-view clustering by incorporating the pre-specified directional dependence between the genomic data types using a copula model. The use of copulas to model the directional dependence provides a robust and versatile tool to capture the directional dependence in joint behavior. The application of the proposed method on synthetic as well as real datasets demonstrates its efficacy. Most importantly, we believe that capturing directional dependence instead of simple dependence can provide an added understanding of the underlying process.

With the groundwork of directionally dependent multi-view clustering in this work, several improvements can be made over the proposed model. Firstly, we can utilize spike-slab priors [55–57] for feature selection while performing directional clustering in high-dimensional TCGA applications. Secondly, an approach to deal with a different number of features in each data type can add more flexibility to the proposed model. Thirdly, we may incorporate hidden Markov models (HMM) in our proposed approach for modeling longitudinal clustering where the class type (or process state) of a patient is a hidden variable, and the multiple data sources are the observed variables. [58–60]. Finally, more rigorous and in-depth comparative analysis between the dependence and directional dependence seeking multi-view clustering with several datasets are needed.

## Author Contributions

**Conceptualization:** Kahkashan Afrin, Ashif S. Iquebal, Mostafa Karimi, Allyson Souris, Se Yoon Lee, Bani K. Mallick.

**Funding acquisition:** Bani K. Mallick.

**Methodology:** Kahkashan Afrin, Ashif S. Iquebal, Mostafa Karimi, Allyson Souris, Se Yoon Lee, Bani K. Mallick.

**Software:** Kahkashan Afrin, Ashif S. Iquebal, Mostafa Karimi, Allyson Souris, Se Yoon Lee.

**Supervision:** Bani K. Mallick.

**Validation:** Kahkashan Afrin, Ashif S. Iquebal, Mostafa Karimi, Se Yoon Lee.

**Writing – original draft:** Kahkashan Afrin, Ashif S. Iquebal, Mostafa Karimi, Allyson Souris, Se Yoon Lee, Bani K. Mallick.

**Writing – review & editing:** Kahkashan Afrin, Ashif S. Iquebal, Mostafa Karimi, Allyson Souris, Se Yoon Lee, Bani K. Mallick.

## References

1. National Human Genome Research Institute. Genome Technology Program [cited 15 Mar 2020]. Available: https://www.genome.gov/Funded-Programs-Projects/Genome-Technology-Program.

2. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490(7418):61. https://doi.org/10.1038/nature11412

3. Lock EF, Dunson DB. Bayesian consensus clustering. Bioinformatics. 2013; 29(20):2610–2616. https://doi.org/10.1093/bioinformatics/btt425 PMID: 23990412

4. Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nature Reviews Genetics. 2018; 19(5):299. https://doi.org/10.1038/nrg.2018.4

5. Reif DM, White BC, Moore JH. Integrated analysis of genetic, genomic and proteomic data. Expert Review of Proteomics. 2004; 1(1):67–75. https://doi.org/10.1586/14789450.1.1.67

6. Ickstadt K, Schäfer M, Zucknick M. Toward integrative Bayesian analysis in molecular biology. Annual Review of Statistics and Its Application. 2018; 5:141–167. https://doi.org/10.1146/annurev-statistics-031017-100438

7. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. Bioinformatics. 2012; 28(24):3290–3297. https://doi.org/10.1093/bioinformatics/bts595 PMID: 23047558

8. Wang H, Shan H, Banerjee A. Bayesian cluster ensembles. Statistical Analysis and Data Mining: The ASA Data Science Journal. 2011; 4(1):54–70. https://doi.org/10.1002/sam.10098

9. Bruno E, Marchand-Maillet S. Multiview clustering: a late fusion approach using latent models. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval; 2009. p. 736–737.

10. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic acids research. 2018; 46(20):10546–10562. https://doi.org/10.1093/nar/gky889 PMID: 30295871

11. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009; 25 (22):2906–2912. https://doi.org/10.1093/bioinformatics/btp543

12. Chaudhuri K, Kakade SM, Livescu K, Sridharan K. Multi-view clustering via canonical correlation analysis. In: Proceedings of the 26th annual international conference on machine learning; 2009. p. 129–136.

13. Zhao J, Xie X, Xu X, Sun S. Multi-view learning overview: Recent progress and new challenges. Information Fusion. 2017; 38:43–54. https://doi.org/10.1016/j.inffus.2017.02.007

14. Crick FH. On protein synthesis. In: Symposia of the Society for Experimental Biology vol. 12; 1958. p. 8.

15. Crick F. Central dogma of molecular biology. Nature. 1970; 227(5258):561–563. https://doi.org/10. 1038/227561a0

16. Alberts B. Molecular biology of the cell. Garland Science, New York; 2017.

17. Kim S, Oesterreich S, Kim S, Park Y, Tseng GC. Integrative clustering of multi-level omics data for disease subtype discovery using sequential double regularization. Biostatistics. 2017; 18(1):165–179. https://doi.org/10.1093/biostatistics/kxw039 PMID: 27549122

18. Weber M. The central dogma as a thesis of causal specificity. History and philosophy of the life sciences. 2006; p. 595–609.

19. Qin H, Niu T, Zhao J. Identifying Multi-Omics Causers and Causal Pathways for Complex Traits. Frontiers in genetics. 2019; 10:110. https://doi.org/10.3389/fgene.2019.00110

20. Nelsen RB. An introduction to copulas. Springer Science & Business Media; 2007.

21. Rey M, Roth V. Copula mixture model for dependency-seeking clustering. arXiv preprint arXiv:12066433. 2012;.

22. Sungur EA. A note on directional dependence in regression setting. Communications in Statistics—Theory and Methods. 2005; 34(9-10):1957–1965. https://doi.org/10.1080/03610920500201228

23. Rodrıguez-Lallena JA, Úbeda-Flores M. A new class of bivariate copulas. Statistics & probability letters. 2004; 66(3):315–325. https://doi.org/10.1016/j.spl.2003.09.010

24. McLachlan GJ, Basford KE. Mixture models: Inference and applications to clustering. vol. 38. M. Dekker New York; 1988.

25. Bishop CM. Pattern recognition and machine learning. Springer-Verlag Berlin Heidelberg; 2006.

26. Rasmussen CE. The infinite Gaussian mixture model. In: Advances in neural information processing systems; 2000. p. 554–560.

27. Walker SG. Sampling the Dirichlet mixture model with slices. Communications in Statistics—Simulation and Computation®. 2007; 36(1):45–54. https://doi.org/10.1080/03610910601096262

28. Hjort NL, Holmes C, Müller P, Walker SG. Bayesian nonparametrics. vol. 28. Cambridge University Press; 2010.

29. Müller P, Quintana FA, Jara A, Hanson T. Bayesian nonparametric data analysis. Springer, Cham; 2015.

30. Görür D, Rasmussen CE. Dirichlet process gaussian mixture models: Choice of the base distribution. Journal of Computer Science and Technology. 2010; 25(4):653–664. https://doi.org/10.1007/s11390-010-9355-8

31. Wang C, Machiraju R, Huang K. Breast cancer patient stratification using a molecular regularized consensus clustering method. Methods. 2014; 67(3):304–312. https://doi.org/10.1016/j.ymeth.2014.03.005

**32.** Dodge Y, Rousson V. Direction dependence in a regression line. Communications in Statistics-Theory and Methods. 2000; 29(9-10):1957–1972. https://doi.org/10.1080/03610920008832589

**33.** Sungur EA. Some observations on copula regression functions. Communications in Statistics—Theory and Methods. 2005; 34(9-10):1967–1978. https://doi.org/10.1080/03610920500201244

**34.** Kim D, Kim JM. Analysis of directional dependence using asymmetric copula-based regression models. Journal of Statistical Computation and Simulation. 2014; 84(9):1990–2010. https://doi.org/10.1080/00949655.2013.779696

**35.** Jung YS, Kim JM, Kim J. New approach of directional dependence in exchange markets using generalized FGM copula function. Communications in Statistics—Simulation and Computation®. 2008; 37 (4):772–788. https://doi.org/10.1080/03610910701711091

**36.** Trivedi PK, Zimmer DM, et al. Copula modeling: an introduction for practitioners. Foundations and Trends® in Econometrics. 2007; 1(1):1–111. https://doi.org/10.1561/0800000005

**37.** Jaworski P, Durante F, Hardle WK, Rychlik T. Copula theory and its applications. vol. 198. Springer-Verlag Berlin Heidelberg; 2010.

**38.** Demarta S, McNeil AJ. The t copula and related copulas. International statistical review. 2005; 73 (1):111–129.

**39.** Sklar M. Fonctions de repartition an dimensions et leurs marges. Publ Inst Statist Univ Paris. 1959; 8:229–231.

**40.** Liebscher E. Construction of asymmetric multivariate copulas. Journal of Multivariate analysis. 2008; 99 (10):2234–2250. https://doi.org/10.1016/j.jmva.2008.02.025

**41.** Bairamov I, Kotz S, Bekci M. New generalized Farlie-Gumbel-Morgenstern distributions and concomitants of order statistics. Journal of Applied Statistics. 2001; 28(5):521–536. https://doi.org/10.1080/02664760120047861

**42.** Nieto-Barajas LE, Prünster I, Walker SG, et al. Normalized random measures driven by increasing additive processes. The Annals of Statistics. 2004; 32(6):2343–2360. https://doi.org/10.1214/009053604000000625

**43.** Kraus D. D-vine copula based quantile regression and the simplifying assumption for vine copulas. 2017;.

**44.** Glaich O, Parikh S, Bell RE, Mekahel K, Donyo M, Leader Y, et al. DNA methylation directs microRNA biogenesis in mammalian cells. Nature communications. 2019; 10(1):1–11. https://doi.org/10.1038/s41467-019-13527-1 PMID: 31827083

**45.** Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. The annals of applied statistics. 2013; 7(1):523. https://doi.org/10.1214/12-AOAS597

**46.** Duan Q, Kou Y, Clark N, Gordonov S, Ma'ayan A. Metasignatures identify two major subtypes of breast cancer. CPT: pharmacometrics & systems pharmacology. 2013; 2(3):1–10.

**47.** Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012; 486(7403):346. https://doi.org/10.1038/nature10983 PMID: 22522925

**48.** Rousseau J, Mengersen K. Asymptotic behaviour of the posterior distribution in overfitted mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2011; 73(5):689–710. https://doi.org/10.1111/j.1467-9868.2011.00781.x

**49.** Kumaran M, Cass CE, Graham K, Mackey JR, Hubaux R, Lam W, et al. Germline copy number variations are associated with breast cancer risk and prognosis. Scientific reports. 2017; 7(1):14621. https://doi.org/10.1038/s41598-017-14799-7 PMID: 29116104

**50.** Fan X, Edrisi M, Navin N, Nakhleh L. Benchmarking tools for copy number aberration detection from single-cell DNA sequencing data. bioRxiv. 2019; p. 696179.

**51.** Mallory XF, Edrisi M, Navin N, Nakhleh L. Assessing the performance of methods for copy number aberration detection from single-cell DNA sequencing data. PLoS computational biology. 2020; 16(7): e1008012. https://doi.org/10.1371/journal.pcbi.1008012

**52.** Knijnenburg TA, Wang L, Zimmermann MT, Chambwe N, Gao GF, Cherniack AD, et al. Genomic and molecular landscape of DNA damage repair deficiency across The Cancer Genome Atlas. Cell reports. 2018; 23(1):239–254. https://doi.org/10.1016/j.celrep.2018.03.076 PMID: 29617664

**53.** Edrisi M, Zafar H, Nakhleh L. A Combinatorial Approach for Single-cell Variant Detection via Phylogenetic Inference. bioRxiv. 2019; p. 693960.

**54.** Rand WM. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association. 1971; 66(336):846–850. https://doi.org/10.1080/01621459.1971.10482356

**55.** Ishwaran H, Rao JS, et al. Spike and slab variable selection: frequentist and Bayesian strategies. The Annals of Statistics. 2005; 33(2):730–773. https://doi.org/10.1214/009053604000001147

**56.** Cui K, Cui W. Spike-and-Slab Dirichlet Process Mixture Models. Open Journal of Statistics. 2012; 2:512–518. https://doi.org/10.4236/ojs.2012.25066

**57.** Rockova V, McAlinn K, et al. Dynamic variable selection with spike-and-slab process priors. Bayesian Analysis. 2020;. https://doi.org/10.1214/20-BA1199

**58.** Helske S, Helske J. Mixture hidden Markov models for sequence data: The seqHMM package in R. arXiv preprint arXiv:170400543. 2017;.

**59.** Altman RM. Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. Journal of the American Statistical Association. 2007; 102(477):201–210. https://doi.org/10.1198/016214506000001086

**60.** Maruotti A. Mixed hidden markov models for longitudinal data: An overview. International Statistical Review. 2011; 79(3):427–454. https://doi.org/10.1111/j.1751-5823.2011.00160.x