genes

MDPI

Article

Clustering Gene Expressions Using the Table Invitation Prior

Charles W. Harrison †, Qing He † and Hsin-Hsiung Huang *

Department of Statistics and Data Science, University of Central Florida, 4000 Central Florida Blvd, Orlando, FL 32816, USA

* Correspondence: hsin.huang@ucf.edu

† These authors contributed equally to this work.

Abstract: A prior for Bayesian nonparametric clustering called the Table Invitation Prior (TIP) is used to cluster gene expression data. TIP uses information concerning the pairwise distances between subjects (e.g., gene expression samples) and automatically estimates the number of clusters. TIP's hyperparameters are estimated using a univariate multiple change point detection algorithm with respect to the subject distances, and thus TIP does not require an analyst's intervention for estimating hyperparameters. A Gibbs sampling algorithm is provided, and TIP is used in conjunction with a Normal-Inverse-Wishart likelihood to cluster 801 gene expression samples, each of which belongs to one of five different types of cancer.

Keywords: Bayesian clustering; distance-dependent clustering; genome-wide association studies; gene expression

Citation: Harrison, C.W.; He, Q.; Huang, H.-H. Clustering Gene Expressions Using the Table Invitation Prior. *Genes* **2022**, *13*, 2036. https://doi.org/10.3390/genes13112036

Academic Editor: George C. Tseng

Received: 14 September 2022
Accepted: 1 November 2022
Published: 4 November 2022

1. Introduction

The goal of clustering is to provide informative groupings (i.e., clusters) of similar objects. The objects are referred to in this paper as "subjects", and an example of an individual subject is a single gene expression from one person. The term "subjects" will be used throughout this paper in order to avoid confusion associated with the term "samples" in a statistical context. Note that, in practice, an individual subject may correspond to an individual vector, matrix, or higher-order tensor. In this work, vectors are considered for the sake of simplicity. In contrast, a "subject index" refers to an identifier for a subject. For example, in vector-variate data, a subject index $i \in \{1, 2, \ldots, n\}$ refers to the ith row in the provided dataset and n is the total number of subjects. The notation $\mathbf{x}_i$ is used to refer to the subject (object) itself (i.e., a vector, matrix, tensor, etc.). The goal of Bayesian clustering is to produce a set of cluster assignments for each subject while also calculating the probability that two subjects are clustered together given the observed data and prior assumptions. Mathematically, this is represented as

$$P(\mathbf{c} \mid \mathbf{X}) \propto P(\mathbf{X} \mid \mathbf{c})P(\mathbf{c}) \tag{1}$$

where $\mathbf{X}$ refers to the data, $\mathbf{c}$ is a vector of n cluster assignments (e.g., the cluster assignments for each of the n gene expression samples), and $P(\mathbf{c} \mid \mathbf{X})$ represents the posterior probability of a cluster configuration, $P(\mathbf{X} \mid \mathbf{c})$ is the likelihood, and $P(\mathbf{c})$ is the cluster prior which is the focus of this paper.

A well-known challenge in clustering is to estimate the unknown number of clusters K^* [1,2]. Some clustering methods, such as MCLUST, involve an analyst fitting several models with varying degrees of complexity and then choosing the desired model based on a chosen clustering metric [3,4]. A distinct, though similar approach, uses the gap statistic [2] in conjunction with another clustering algorithm (e.g., hierarchical clustering) to estimate the number of clusters.

Bayesian nonparametric models refer to a flexible class of prior distributions that may be used in a variety of settings including clustering. In the context of clustering, the use of

such priors means that an analyst does not need to specify an estimate for the number of clusters since the number of clusters is modeled as a random variable. A variety of methods have been proposed to accomplish this task, and the relevant methods are reviewed in the sections that follow. Two methods include the Ewens-Pitman Attraction (EPA) prior [5] and a well-known special case of EPA called the Chinese Restaurant Process (CRP) [6,7].

The CRP, a variant of the Dirichlet process, is a well-known prior used in Bayesian clustering. One drawback of CRP is that it does not utilize information pertaining to the similarity between subjects (e.g., gene expression samples). A natural extension of the CRP is one that includes the aforementioned similarity information, and one such extension is the EPA prior. Although EPA utilizes similarity information, the primary drawback of EPA is that it relies on the choice of a hyperparameter (this is α in Section 1.2 below). Consequently, an analyst using EPA must either choose a fixed value for the EPA hyperparameter α or rely on an approximate posterior distribution for α [5,8]. In the context of Bayesian clustering, a number of samples are taken from a posterior distribution which can be time consuming, so manually tuning a hyperparameter is not desirable.

The focus of this paper is the Table Invitation Prior (TIP) which is an attempt to maintain the advantage of Bayesian clustering (i.e., the analyst does not have to specify the number of clusters) while removing the need for an analyst to carefully tune a hyperparameter. Although the approximate posterior distribution for α used in EPA removes the need of the analyst to tune the hyperparameter, empirical results show that TIP gives superior results and is less susceptible to splitting clusters as compared to EPA. Bayesian clustering methods often rely on the use of similarity functions to capture the relationships between subjects (e.g., gene expression samples), and thus a brief review of pairwise similarity functions is provided.

1.1. Pairwise Similarity Functions

Some Bayesian clustering priors use the similarity between subjects in order to obtain clusters that contain subjects that are similar to each other [5,9]. Let the similarity between two subjects with indices i and j be given by $\lambda(i,j)$ for $i = 1,2,\ldots,n$ and $j = 1,2,\ldots,n$ where n is the number of subjects. The similarity function λ may take a variety of forms, and in this paper the similarity function used is the exponential decay function [5,9]:

$$\lambda(i,j) = \exp(-\tau d_{ij}) \tag{2}$$

where $\tau > 0$ is a hyperparameter and d_{ij} is the distance between the ith and jth subjects. Following the approach taken in [10], the hyperparameter τ is set to the following:

$$\hat{\tau} = \frac{1}{\tilde{d}} \tag{3}$$

where $\tilde{d}$ is the median of the pairwise distances of the strictly upper triangular portion of the distance matrix:

$$\tilde{d} = \text{median}\{d_{ij} : i > j, i, j \in \{1,2,\ldots,n\}\}. \tag{4}$$

The choice of the median is heuristic, but there is a justification. Equation (3) implies that

$$\lim_{d_{ij}\to\infty} \exp\left(\frac{-d_{ij}}{\tilde{d}}\right) = 0,$$

$$\lim_{d_{ij}\to 0} \exp\left(\frac{-d_{ij}}{\tilde{d}}\right) = 1,$$

and

$$\lim_{d_{ij}\to\tilde{d}} \exp\left(\frac{-d_{ij}}{\tilde{d}}\right) = \exp(-1).$$

Consequently, similarity values corresponding to subject pairs whose distances are significantly larger than the overall median distance go to zero whereas subject pairs that are very close to each other have a similarity value that is closer to 1. Subject pairs whose distance from each other is close to the overall median distance have a similarity value that is between 0 and 1.

1.2. Ewens-Pitman Attraction Prior

The EPA distribution uses the pairwise similarity between subjects and a sequential sampling scheme to induce a partition of n subjects [5,11]. Let $\sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}$ be a random permutation of the subject indices $\{1, 2, \ldots, n\}$. Then the conditional probability of a subject with index i joining cluster k is given by the following:

$$P(c_{\sigma_i} = k | \alpha, \delta, \lambda, c(\sigma_1, \ldots, \sigma_{i-1})) = \begin{cases} \frac{i-1-\delta q_{i-1}}{\alpha+i-1} \frac{\sum_{\sigma_S \in S} \lambda(\sigma_i, \sigma_s)}{\sum_{s=1}^{i-1} \lambda(\sigma_i, \sigma_s)} & \text{if } S \in c(\sigma_1, \ldots, \sigma_{i-1}) \\ \frac{\alpha+\delta q_{i-1}}{\alpha+i-1} & \text{if } S \text{ is a new cluster} \end{cases} \quad (5)$$

where $\alpha > 0$ is a hyperparameter that controls the extent to which a new cluster is created, q_{i-1} is the number of clusters that are assigned among the first $i-1$ subjects, $\delta \in [0, 1)$ is a "discount" hyperparameter, λ is a similarity function, and $c(\sigma_1, \sigma_2, \ldots, \sigma_{i-1})$ are the part assignments for the first $i-1$ permuted subjects $\sigma_1, \ldots, \sigma_{i-1}$. The discount parameter is specific to EPA and its purpose is to incorporate information about the number of clusters in a previous iteration when computing the probability of a new cluster in the current iteration.

The value for α may be treated as a constant or it can be sampled from a distribution as described in West [8]. Specifically, West's approximate posterior distribution for α, given the number of clusters n_k, is:

$$\alpha \mid n_k \sim \Gamma(a + n_k - 1, b + \gamma + log(n)) \quad (6)$$

where Γ denotes the gamma distribution, γ is Euler's constant, and the prior parameters are a and b. In this work, $a = b = 1$ so that the prior for α has exponential distribution with a scale parameter of 1.

Chinese Restaurant Process

The CRP is a special case of EPA that occurs when the discount parameter $\delta = 0$ and $\lambda(i, j)$ is constant for all subject indices i and j [5–7]. The conditional probability of a subject $\mathbf{x}_i$ joining cluster k is given by the following:

$$P(c_{\sigma_i} = k \mid \alpha, c(\sigma_1, \sigma_2, \ldots, \sigma_{i-1})) = \begin{cases} \frac{|S|}{\alpha+i-1} & \text{if } S \in c(\sigma_1, \sigma_2, \ldots, \sigma_{i-1}) \\ \frac{\alpha}{\alpha+i-1} & \text{if } S \text{ is a new subset} \end{cases} \quad (7)$$

The CRP is obtained by taking the product of (7) over all possible partitions.

2. Table Invitation Prior

In this section, the Table Invitation Prior (TIP) is presented in the context of a Gibbs sampler in iteration $t = 1, 2, \ldots, T$. An analogy is now provided to illustrate the prior's mechanics. Suppose that n subjects $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ (i.e., vectors, matrices, tensors, etc.) are sitting in a restaurant with $k = 1, 2, \ldots, K^{(t)}$ tables (clusters). A randomly selected subject with index r is chosen and then the $\hat{n}_{\tau_r}$ subjects that are most similar to the subject with index r are "invited" to a new table (cluster) $K^{(t)} + 1$ (in this paper, all the $\hat{n}_{\tau_r}$ subjects accept the invitation with probability 1). The posterior probability of every subject belonging to a table (cluster) is computed for tables (clusters) $1, 2, \ldots, K^{(t)}, K^{(t)} + 1$ and, using the probabilities, the subjects are randomly assigned to a table (i.e., sample the posterior cluster assignment for every subject). The variable t is incremented by 1 and the this process continues; here t is the number of times the above process has occurred so far.

A more formal description of the Table Invitation Prior now follows. For the iteration t in a Gibbs sampler, let the random variable r be a randomly selected subject index (e.g., a randomly selected index corresponding to an individual gene expression) from a discrete uniform distribution

$$r \sim \mathcal{U}\{0, n\} \tag{8}$$

so that $r \in \{1, 2, \ldots, n\}$. Suppose a random subject $\mathbf{x}_r$ is selected (i.e., $\mathbf{x}_r$ can be a vector, matrix, higher-order tensor, etc.). The set of similarity values between subject $\mathbf{x}_r$, itself, and the other $n-1$ subjects is

$$\Lambda_r = \{\lambda(r, i) : i \in \{1, 2, \ldots, n\}\} \tag{9}$$

where $\lambda(r, i)$ is the similarity between the rth subject and the ith subject. Let the jth ordered similarity value in the set Λ_r be $\Lambda_{r^{(j)}}$ for $j = 1, 2, \ldots, n$ and let

$$\Lambda_{r^{(n)}} = \lambda(r, r) > \Lambda_{r^{(n-1)}} > \Lambda_{r^{(n-2)}} \ldots > \Lambda_{r^{(1)}}. \tag{10}$$

The set of indices of the n_{τ_r} subjects that are most similar to subject $\mathbf{x}_r$ is given by

$$S_r = \{r = r^{(n)}, r^{(n-1)}, r^{(n-2)}, \ldots, r^{(n-n_{\tau_r}+1)}\} \tag{11}$$

where $n_{\tau_r} \in \{1, 2, \ldots, n-1\}$ is a hyperparameter. The estimation of hyperparameter n_{τ_r} proceeds in the following manner. First, recall that r is a randomly selected subject index so that $r \in \{1, 2, \ldots, n\}$. The pairwise distances with respect to subject r are extracted and the distance from subject r to itself is removed:

$$d_r = \{d_{r,j} : j \in \{1, 2, \ldots, r-1, r+1 \ldots, n\}\},$$

The distances are then sorted in ascending order:

$$d_r^* = \{d_{r,j^*} : d_{r,j^*} < d_{r,j^*+1}, j^* \in \{1, 2, \ldots, r-1, r+1, \ldots, n\}\}. \tag{12}$$

Next, a univariate multiple change point detection algorithm is applied to the sorted distances. The change point detection algorithm used in this paper is binary segmentation from the changepoint library in R [12]. The binary segmentation function in R takes a hyperparameter, denoted by Q, that is the maximum number of changepoints (13). In this paper, the value is set to $\lfloor \frac{n}{2} + 1 \rfloor$ since the changepoint library will throw an error if $Q > \frac{n}{2} + 1$; also this allows the change point method to have the maximum amount of flexibility in detecting changes in the subject distances. Let the set of change points be given by

$$\hat{\boldsymbol{\tau}}_r = \{\hat{\tau}_{r,1}, \hat{\tau}_{r,2}, \ldots\}, \tag{13}$$

then the number of subjects that are similar to the subject with index r is taken to follow a Poisson distribution:

$$\hat{n}_{\tau_r} \sim Poi(\hat{\tau}_{r,1}).$$

Consequently, the set $\hat{S}_r$ is given by:

$$\hat{S}_r = \{r = r^{(n)}, r^{(n-1)}, r^{(n-2)}, \ldots, r^{(n-\hat{n}_{\tau_r}+1)}\}. \tag{14}$$

Let the vector containing the cluster assignments be denoted by $\mathbf{c}_t$ so that the ith element contains the cluster assignment for the ith subject and

$$c_{t,i} \in \{1, 2, \ldots, K^{(t)}\} \text{ for } i = 1, 2 \ldots, n$$

where $K^{(t)}$ is the total number of clusters after posterior sampling in iteration t. The Table Invitation Prior is based on selecting a random subject index r and forming a new cluster

with $\hat{n}_{\tau_r}$ subjects that are most similar to subject $\mathbf{x}_r$. That is, the new cluster is formed using the subjects whose indices are in the set $\hat{S}_r$. Consider a modified cluster vector $\tilde{\mathbf{c}}_t$

$$\tilde{c}_{t,i} = \begin{cases} c_{t,i} & i \notin \hat{S}_r \\ K^{(t)} + 1 & i \in \hat{S}_r, \end{cases} \tag{15}$$

then the Table Invitation Prior (TIP) is given by

$$P(c_{t+1,i} = k | \mathbf{X}, p, \lambda) \propto \sum_{\tilde{c}_{t,j}=k} \lambda(i,j) \text{ for } k \in \{1, 2, \ldots, K^{(t)}, K^{(t)} + 1\}. \tag{16}$$

3. TIP Gibbs Sampler

An implementation of a Gibbs sampler with a Table Invitation Prior for clustering corresponds to the following steps, and it is summarized in Algorithm 1. Initially, all subjects are sitting at $K^{(0)}$ tables (clusters). For Gibbs sampling iteration $t = 1$, a random subject with index r is chosen and the $\hat{n}_{\tau_r}$ most similar subjects with indices $r_{(n-1)}, r_{(n-2)}, \ldots, r_{(n-\hat{n}_{\tau_r}+1)}$ are assigned to table $K^{(0)} + 1$ with subject $\mathbf{x}_r$. The conditional probabilities for all subjects $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ (i.e. vectors, matrices, tensors, etc.) are computed using Equation (16) for all clusters $k = 1, 2, \ldots, K^{(0)} + 1$; if desired, a likelihood value may be computed for each table (cluster). Next, the posterior probability is computed and the subject's posterior cluster assignment is sampled (i.e., sampled from the set $\{1, 2, \ldots, K^{(0)}, K^{(0)} + 1\}$). This gives a partition with $K^{(1)}$ clusters (tables). In the second Gibbs sampling iteration $t = 2$, a random subject with index $r \sim \mathcal{U}\{0, n\}$ is chosen and the $\hat{n}_{\tau_r}$ most similar subjects with indices $r_{(n-1)}, r_{(n-2)}, \ldots, r_{(n-\hat{n}_{\tau_r}+1)}$ are assigned to table $K^{(1)} + 1$ with subject $\mathbf{x}_r$. The conditional probabilities for all subjects $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are computed using Equation (16) for all clusters $k = 1, 2, \ldots, K^{(1)} + 1$; again, a likelihood value may be computed for each table (cluster), and each subject's posterior cluster assignment is sampled (i.e., sampled from the set $\{1, 2, \ldots, K^{(1)} + 1\}$). This process continues for $t \in \{3, \ldots, T\}$.

Algorithm 1: Table Invitation Prior Clustering

Inputs: n subjects $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$, number of Gibbs sampling iterations T, initial cluster assignments $\mathbf{c}_0$, similarity function λ with hyperparameters $\mathbf{\Theta}$, and a distance matrix $\mathbf{D}$.
Output: posterior cluster assignments $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_T$.
Sort the ith row of the distance matrix to obtain d_i^*.
Compute $\hat{n}_{\tau_i}$ for each subject $\mathbf{x}_i$ for $i \in \{1, 2, \ldots, n\}$ by applying a univariate multiple change point detection algorithm to the set of sorted distances d_i^*.
Compute the similarity sets $\hat{S}_i$ using Equation (14) for $i \in \{1, 2, \ldots, n\}$.
for *t in 1:T* **do**
    Draw a random subject index $r \sim \mathcal{U}\{0, n\}$.
    Using the set $\hat{S}_r$, compute the modified cluster vector $\tilde{\mathbf{c}}_t$ via (15) given $\mathbf{c}_{t-1}$.
    parallel for *i in 1:n* **do**
        for *k in 1:($K^{(t)}$ + 1)* **do**
            Compute the log-prior $\log(P(c_i = k \mid \mathbf{X}))$ via (16).
            Compute the log-likelihood (if desired; i.e., Normal-Inverse-Wishart etc.).
            Compute the log-posterior = log-prior + log-likelihood.
            Convert the log-posterior to a probability.
            Sample the posterior cluster assignment $c_{i,t}$.
        end
    end
end

3.1. Posterior Cluster Assignments

The TIP Gibbs sampler produces a set of posterior cluster vectors $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_T$. However, it is necessary to produce a single clustering from this set of posterior cluster assignments, and this section describes the methodology used in this paper to accomplish this task (other methods may be used for this task). Each posterior cluster vector is transformed into an $n \times n$ posterior proximity matrix $\mathbf{B}^{(t)}$:

$$B_{i,j}^{(t)} = \begin{cases} 0 & c_{t,i} \neq c_{t,i} \\ 1 & c_{t,i} = c_{t,j} \end{cases} \tag{17}$$

where $c_{t,i}$ and $c_{t,j}$ are the posterior cluster assignments for subject $\mathbf{x}_i$ and $\mathbf{x}_j$ after Gibbs sampling iteration $t \in \{1, 2, \ldots, T\}$. The posterior similarity matrix is given by

$$\bar{\mathbf{B}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{B}^{(t)}. \tag{18}$$

A vector of posterior cluster assignments is computed using the Posterior Expected Adjusted Rand (PEAR) index, and technical details as well as an application of PEAR to gene expression data may be found in Fritsch and Ickstadt [13]. Let ρ denote the PEAR index function. Using the posterior similarity matrix $\bar{\mathbf{B}}$, the cluster vector that maximizes the PEAR index is taken to be the posterior cluster assignment vector:

$$\mathbf{c}^* = \underset{\mathbf{c} \in \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_T\}}{\arg\max} \; \rho(\mathbf{c} | \bar{\mathbf{B}}). \tag{19}$$

The computation of the PEAR index is accomplished using the mcclust package in R [13,14].

3.2. Likelihood Function

The Table Invitation Prior may be used for clustering vectors, matrices, and higher-order tensors, assuming that a suitable distance metric is available. One component of a Gibbs sampler utilizing TIP that may change depending on the dataset is the likelihood function. In this paper vector-variate data are considered, and the conjugate Normal-Inverse-Wishart (NIW) prior for the mean and covariance is utilized. Let $\mathbf{x}_i \in \mathbb{R}^p$ be a $p \times 1$ vector and represent the ith subject for $i = 1, 2, \ldots, n$. Let c_i be the cluster assignment for subject i. Assume that

$$\mathbf{x}_i | c_i = k \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{20}$$

$$(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \sim NIW(\boldsymbol{\mu}_0, \lambda_0, \boldsymbol{\Psi}_0, \nu_0), \tag{21}$$

then the joint posterior for $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k | \mathbf{x}_i)$ is given by

$$(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k | y_i) \sim NIW(\boldsymbol{\mu}_{1k}, \lambda_{1k}, \boldsymbol{\Psi}_{1k}, \nu_{1k}) \tag{22}$$

where NIW denotes the Normal-Inverse-Wishart distribution. The posterior arguments are given by

$$\boldsymbol{\mu}_{1k} = \frac{\lambda_0 \boldsymbol{\mu}_0 + n_k \bar{\mathbf{x}}}{\lambda_0 + n_k},$$

$$\lambda_{1k} = \lambda_0 + n_k,$$

$$\nu_{1k} = \nu_0 + n_k,$$

and

$$\boldsymbol{\Psi}_{1k} = \boldsymbol{\Psi}_0 + \sum_{i=1}^{n_k} (\mathbf{x}_i - \bar{x}_k)'(\mathbf{x}_i - \bar{x}_k) + \frac{\lambda_0 n_k}{\lambda_0 + n_k} (\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_k - \mu_0)'.$$

There are four hyperparameters and the following values are used:

$$\hat{\mu}_0 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i,$$

$$\hat{\lambda}_0 = 1,$$

$$\hat{\nu}_0 = p,$$

and

$$\hat{\mathbf{\Psi}}_0 = \left((p-1)\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_i - \bar{\mathbf{x}})\right)^{-1}.$$

3.3. Visualizing The Posterior Similarity Matrix

The $n \times n$ symmetric matrix $\bar{\mathbf{B}}$ can be viewed an undirected weighted graph with n vertices where the edge between the ith and jth subjects (vertices) represents the posterior probability that subject $\mathbf{x}_i$ and subject $\mathbf{x}_j$ are in the same cluster. A network plot may be used to view $\bar{\mathbf{B}}$ directly, but the number of edges corresponding to small posterior probabilities may unnecessarily complicate the plot. Consequently, we show the plot of the graph $\bar{\mathbf{B}}_1$ which is the result of removing the maximum number of edges in the graph $\bar{\mathbf{B}}$ such that the number of components in the graph is 1. That is, the graph $\bar{\mathbf{B}}_1$ has the minimum entropy among all subgraphs with one component and we call its corresponding network plot the "one-cluster plot". The idea is to remove as many connections as possible while still maintaining one component so that the clusters' relationships with each other are revealed. The network plots are used in Section 5 to visualize the cluster results.

4. Simulation Data

In this section, a clustering simulation is presented to compare TIP with various clustering algorithms including EPA, MCLUST, and linkage-based methods. For EPA, $\delta = 0$ and α follows West's posterior given by Equation (6).

4.1. Simulation Description

The simulation is given by the following. A dataset $\mathbf{X}$ with n subjects $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ is generated where $\mathbf{x}_n \in R^p$. Each subject $\mathbf{x}_i$ for $i = 1, 2, \ldots, n$ is generated according to its true cluster assignment k so that

$$\mathbf{x}_i \mid k \sim \mathrm{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \text{ for } i = 1, 2, \ldots, n,$$

$$\boldsymbol{\mu}_k \sim \mathrm{N}_p(\mathbf{0}, 10\mathbf{I}_p),$$

and

$$\Sigma_k \sim \mathrm{W}^{-1}(\mathbf{I}_p, p+1).$$

Here N_p denotes the p-variate multivariate normal distribution and W_p^{-1} denotes the inverse Wishart distribution. The number of burn-in iterations for both TIP and EPA is set to 1000, and the number of sampling iterations is set to 1000.

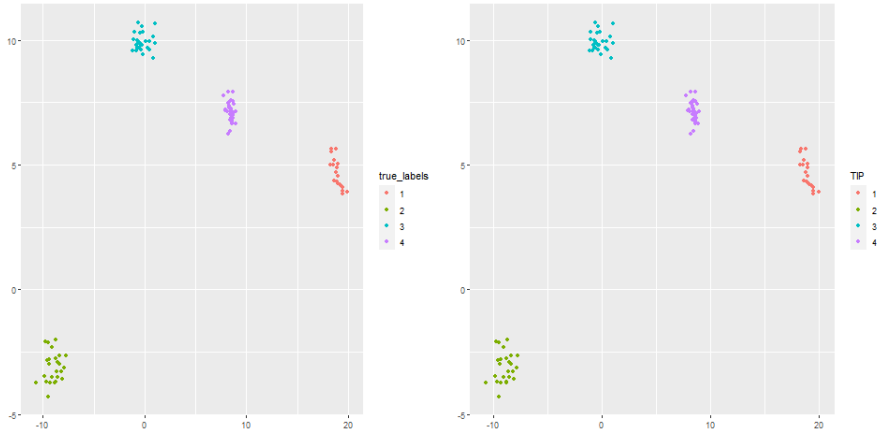
4.2. Simulation Results: Normal-Inverse-Wishart Likelihood Function

In this section, simulation results for TIP and EPA in conjunction with a Normal-Inverse-Wishart likelihood function are presented. TIP and EPA are compared with the MCLUST algorithm, k-means clustering, and hierarchical clustering [4,15]. For k-means and hierarhcial clustering, the number of clusters is estimated using the gap statistic [2].

4.2.1. *Well Separated Clusters* : $p = 2, K^* = 4$ and $n = 110$

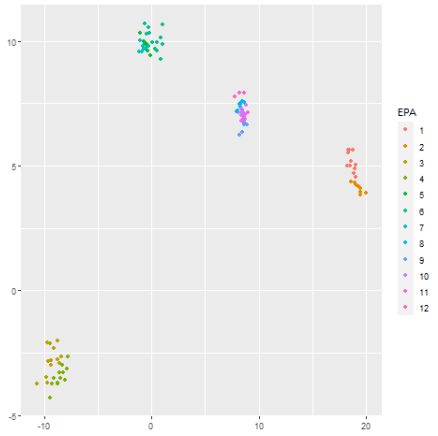
In this simulation there are $K^* = 4$ well separated clusters. Each cluster is composed of $n_1 = 20$, $n_2 = 25$, $n_3 = 30$, and $n_4 = 35$ vectors in $p = 2$ dimensional space. The results

are shown in Figure 1. Both TIP and MCLUST cluster the datasets perfectly while EPA with West's posterior is too aggressive and results in 12 clusters. Hierarchical clustering (complete linkage) is utilized in conjunction with the gap statistic. The optimal number of clusters given by the gap statistic is 4, and hierarchical clustering using complete linkage with exactly 4 clusters perfectly separates the data. The gap statistic is also used with k-means and the optimal number of clusters given by the gap statistic is 4, and k-means perfectly separates the data.

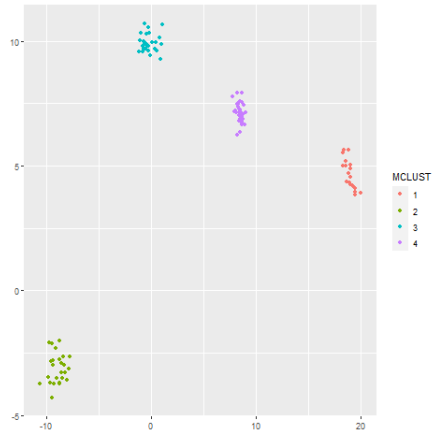


(**a**) True Cluster

(**b**) TIP

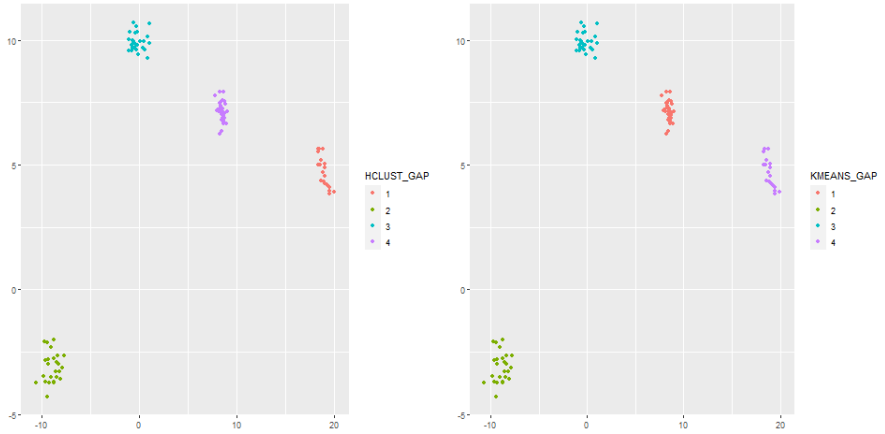


(**c**) EPA



(**d**) MCLUST

Figure 1. *Cont.*

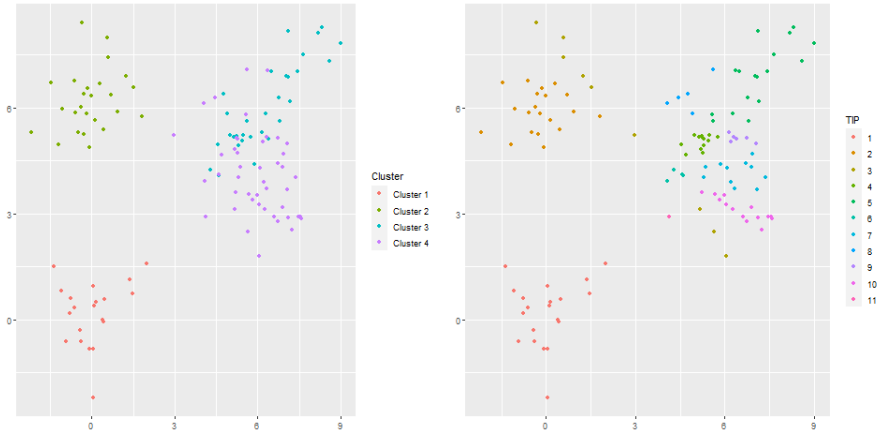


(**e**) HCLUST-GAP (**f**) KMEANS-GAP

Figure 1. Panel (**a**) shows the true cluster assignments when $p = 2$, $K^* = 4$, and $n = 110$. Panel (**b**) shows the posterior TIP cluster assignments, Panel (**c**) shows the posterior EPA assignments, Panel (**d**) shows the MCLUST assignments, Panel (**e**) shows the hierarchical clustering assignments (complete linkage) where the number of clusters is determined via the gap statistic, and Panel (**f**) shows the k-means clustering assignments where the number of clusters is determined via the gap statistic.

4.2.2. Overlapped Clusters: $p = 2$, $K^* = 4$ and $n = 120$

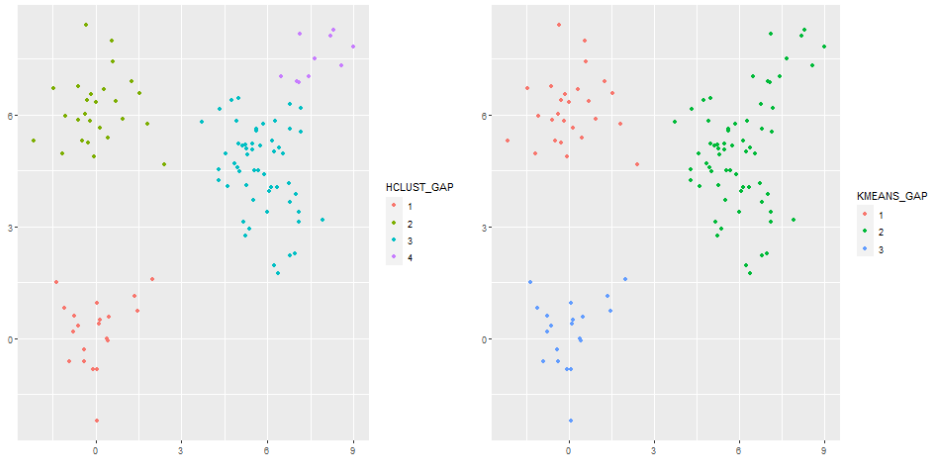
In this simulation there are $K^* = 4$ clusters, but two of the clusters are overlapped. In this case, the cluster sizes are given by $n_1 = 20$, $n_2 = 25$, $n_3 = 30$, and $n_4 = 45$ vectors in $p = 2$ dimensional space. The results are shown in Figure 2. EPA gives 15 clusters, TIP gives 11 clusters, and MCLUST gives 5 clusters. MCLUST divides the two overlapped clusters into 3 clusters whereas TIP is too aggressive and divides the overlapped clusters into 8 clusters. Similarly, EPA divides the overlapped clusters into 8 clusters, but, unlike TIP, EPA also divides Cluster 1 into 2 clusters. Hierarchical clustering (complete linkage) is used in conjunction with the gap statistic and gives 4 clusters, though the resulting cluster assignments are not necessarily accurate since part of Cluster 3 is clustered with part of Cluster 4. K-means is also applied to the dataset in conjunction with the gap statistic; the optimal number of clusters according to the gap statistic is 3 which fuses two of the true clusters (i.e., Cluster 3 and Cluster 4) together.



(**a**) True Cluster (**b**) TIP

Figure 2. *Cont.*

(**c**) EPA (**d**) MCLUST



(**e**) HCLUST-GAP (**f**) KMEANS-GAP

Figure 2. Panel (**a**) shows the true cluster assignments when $p = 2, K^* = 4$, and $n = 110$. Panel (**b**) shows the posterior TIP cluster assignments, Panel (**c**) shows the posterior EPA assignments, Panel (**d**) shows the MCLUST assignments, Panel (**e**) shows the hierarchical clustering assignments where the number of clusters is determined via the gap statistic, and Panel (**f**) shows the k-means clustering assignments where the number of clusters is determined via the gap statistic.

5. Application: Clustering Gene Expression Data

In this section TIP is applied to a dataset pertaining to RNA-Seq gene expression levels as measured by an Illumina HiSeq platform [16]. The data were accessed from the UCI Machine Learning Repository [17] and were collected as a part of the Cancer Genome Atlas Pan-Cancer analysis project [18]. There are $n = 801$ gene expression samples (i.e., $n = 801$ subjects) and $p = 20,531$ gene expression levels. The 801 gene expression samples can be classified into one of 5 classes, and each class corresponds to a different type of cancer: BRCA, COAD, KIRC, LUAD, and PRAD.

Principal components analysis was applied to the data, and 7 principal components were used so that $p = 7$. A plot showing the cumulative variance explained by a given number of principal components is shown in Figure 3. The reason that 7 principal components were used is that it takes a relatively large number of dimensions to explain percentages of the variance greater than 80%. The first 7 principal components explain about 80% of the variance, but it takes 22 principal components to explain 85% of the variance, 82 principal components to explain 90% of the variance, 269 principal components to explain 95% of the variance, and 603 principal components to explain 99% of the variance.

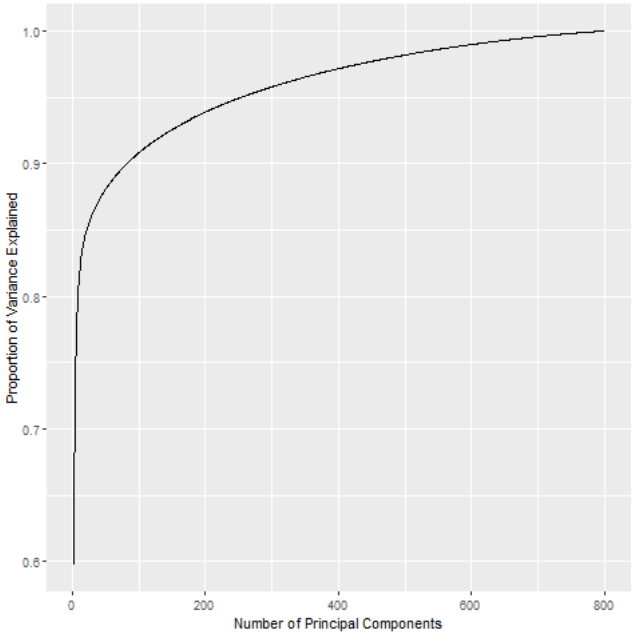


Figure 3. The cumulative proportion of variance explained by the principal component analysis.

The clustering methods are applied to the principal components so that $p = 7$ and $n = 801$. The TIP posterior cluster assignments are shown in Table 1. There is a small overlap of classes in cluster 3 where there are 270 BRCA gene expression samples and 1 LUAD gene expression sample. Also, 30 BRCA gene expression samples form a distinct cluster (see cluster 5). The one-cluster plot is shown in Figure 4 and shows a small amount of overlap between LUAD and BRCA which is consistent with the posterior cluster assignments in Table 1.

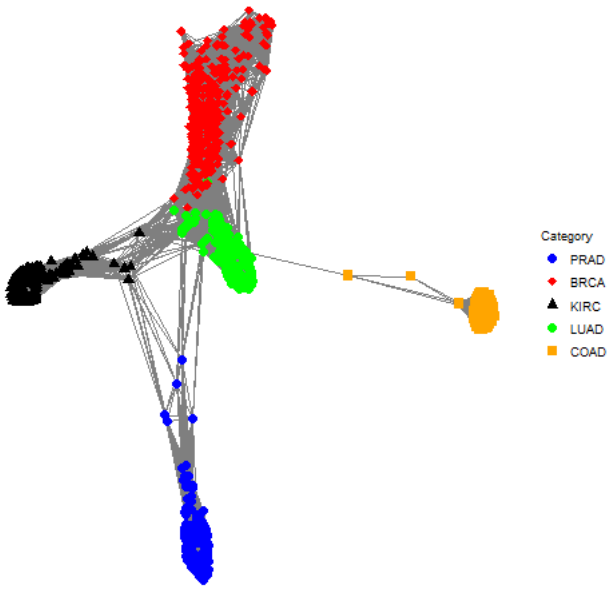


Figure 4. A one-cluster plot with respect to TIP where each graph vertex (i.e., a colored dot) corresponds to a subject (e.g., an individual gene expression sample) and the edge weights (i.e., the lines) correspond to the elements in the matrix $\bar{\mathbf{B}}_1$. Specifically, the edge between subject i and subject j is the posterior probability that subject i and subject j are in the same cluster. Shorter lines correspond to larger posterior probabilities, so pairs of graph vertices that are closer to each other in the plot are more likely to be assigned to the same cluster. The plot shows a minor overlap between BRCA (red diamond) and LUAD (green circle).

Table 1. The distribution of the posterior TIP cluster assignments. The number in parenthesis is the number of subjects (e.g., gene expression samples) for one of the five cancer types.

Cluster ID	Distribution
1	PRAD (136)
2	LUAD (140)
3	BRCA (270) LUAD (1)
4	KIRC (146)
5	BRCA (30)
6	COAD (78)

The posterior cluster assignments for EPA are shown in Table 2. EPA is able to separate the classes quite well, but there is one cluster where there is substantial overlap between classes. Cluster 10 is comprised of samples from BRCA, COAD, KIRC, LUAD, and PRAD whereas this does not occur for TIP and MCLUST. Furthermore, Cluster 6 contains samples from both BRCA and LUAD; this is true for TIP and EPA. The one-cluster plot for EPA is shown in Figure 5, and it shows that there is overlap between LUAD and BRCA as well as BRCA and COAD.

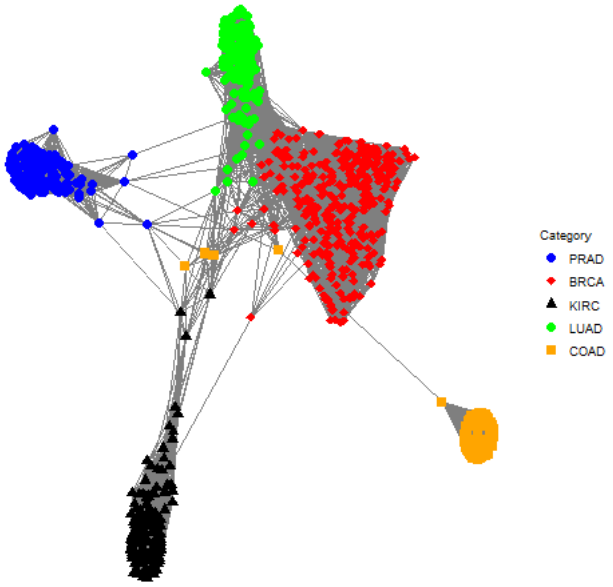


Figure 5. A one-cluster plot with respect to EPA where each graph vertex (i.e., a colored dot) corresponds to a subject (e.g., an individual gene expression sample) and the edge weights (i.e., the lines) correspond to the elements in the matrix $\bar{\mathbf{B}}_1$. Specifically, the edge between subject i and subject j is the posterior probability that subject i and subject j are in the same cluster. Shorter lines correspond to larger posterior probabilities, so pairs of graph vertices that are closer to each other in the plot are more likely to be assigned to the same cluster. There is an overlap between LUAD (green circle) and BRCA (red diamond) as well as BRCA (red diamond) and COAD (orange square).

The cluster assignments for MCLUST are shown in Table 3. MCLUST, like TIP, performs well. MCLUST produces one cluster with a minor amount of overlap: cluster 1 features 57 BRCA samples and 2 LUAD samples. Furthermore, BRCA is split between two clusters: one with 57 BRCA samples and another with 243 BRCA samples. This is similar to the TIP results.

Hierarchical clustering is applied in conjunction with the gap statistic to choose the number of clusters, and the R package cluster is used to compute the gap statistic [19]. The settings used for hierarchical clustering and k-means are the default settings in the stats library in R [20]. The results for hierarchical clustering using complete linkage are shown in Table 4. The optimal number of clusters estimated via the gap statistic is 7, but complete

linkage clustering is unable to separate the classes. The results for hierarchical clustering using single linkage are shown in Table 5. The optimal number of clusters estimated by the gap statistic is 1, and thus single linkage clustering is unable to separate the classes. The results for hierarchical clustering using median linkage are shown in Table 6. The optimal number of clusters given by the gap statistic is 1, and thus median linkage clustering is unable to separate the classes. K-means clustering is also used in conjunction with the gap statistic. The optimal number of clusters according to the gap statistic is 5, but the resulting clusters, which are provided in Table 7, do not separate the data well.

Table 2. The distribution of the EPA cluster assignments. The number in parenthesis is the number of subjects for one of the five cancer types.

Cluster ID	Distribution
1	PRAD (124)
2	LUAD (124)
3	PRAD (11)
4	BRCA (100)
5	KIRC (128)
6	BRCA (104) LUAD (2)
7	BRCA (32)
8	KIRC (16)
9	COAD (74)
10	BRCA (3) COAD (4) KIRC (2) LUAD (15) PRAD (1)
11	BRCA (61)

Table 3. The distribution of the MCLUST cluster assignments. The number in parenthesis is the number of subjects for one of the five cancer types.

Cluster ID	Distribution
1	BRCA (57) LUAD (2)
2	LUAD (139)
3	PRAD (136)
4	BRCA (243)
5	KIRC (146)
6	COAD (78)

Table 4. The distribution of the hierarchical cluster assignments using complete linkage (default settings in R are used) and the gap statistic to select the number of clusters. The gap statistic suggests 7 clusters.

Cluster ID	Distribution
1	BRCA (57) COAD (7) KIRC (42) LUAD (32) PRAD (25)
2	BRCA (52) COAD (8) LUAD (18) PRAD (22)
3	BRCA (84) COAD (17) LUAD (17)
4	COAD (17) LUAD (34) PRAD (40)
5	BRCA (68) KIRC (23) LUAD (18)
6	BRCA (70) COAD (8) KIRC (29) LUAD (22) PRAD (49)
7	KIRC (52) COAD (21)

Table 5. The distribution of the hierarchical cluster assignments using single linkage (default settings in R are used) and the gap statistic to select the number of clusters. The gap statistic suggests exactly 1 cluster.

Cluster ID	Distribution
1	BRCA (300) COAD (78) KIRC (146) LUAD (141) PRAD (136)

Table 6. The distribution of the hierarchical cluster assignments using median linkage (default settings in R are used) and the gap statistic to select the number of clusters. The gap statistic suggests exactly 1 cluster.

Cluster ID	Distribution
1	BRCA (300) COAD (78) KIRC (146) LUAD (141) PRAD (136)

Table 7. The distribution of the hierarchical cluster assignments using k-means (default settings in R are used) and the gap statistic to select the number of clusters. The gap statistic suggests 5 clusters.

Cluster ID	Distribution
1	BRCA (68) COAD (27) KIRC (39) LUAD (42) PRAD (28)
2	BRCA (77) COAD (14) KIRC (31) LUAD (26) PRAD (41)
3	COAD (34) LUAD (1)
4	BRCA (57) COAD (3) KIRC (40) LUAD (46) PRAD (37)
5	BRCA (98) KIRC (36) LUAD (26) PRAD (30)

6. Conclusions and Discussion

In this work, a Bayesian nonparametric clustering prior called the Table Invitation Prior (TIP) was introduced. TIP does not require the analyst to specify the number of clusters, and its hyperparameters are automatically estimated via univariate multiple change point detection. EPA is a prior on partitions and is used for Bayesian clustering. Unlike TIP, the probability of a new cluster in EPA depends on preset hyperparameters (i.e., δ and $\alpha > -\delta$), which is not data-driven, and it may lead to a bias of the number of clusters due to improper hyperparameter values. The main difference between TIP and MCLUST is that TIP is a Bayesian cluster prior which can be incorporated with various types of likelihoods and priors for the parameters in the likelihood. For example, TIP can work with a conjugate using the Normal-Inverse-Wishart prior of for unknown mean and covariance matrix. MCLUST is based on a mixture model of finite Gaussian likelihoods and uses an expectation–maximization (EM) algorithm [21] for the Gaussian mixture parameter estimation with a preset covariance structure. TIP was used in conjunction with a Normal-Inverse-Wishart conjugate prior to cluster gene expression data, and it was compared with a variety of other clustering methodologies, including another Bayesian nonparametric clustering method called EPA, MCLUST, hierarchical clustering in conjunction with the gap statistic, and k-means clustering in conjunction with the gap statistic.

Author Contributions: Conceptualization, H.-H.H. and C.W.H.; formal analysis, C.W.H. and Q.H.; data collection, C.W.H.; writing—original draft preparation, H.-H.H., C.W.H. and Q.H.; writing—review and editing, H.-H.H., C.W.H. and Q.H.; project administration, H.-H.H.; funding acquisition, H.-H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the National Science Foundation grant to H.-H.H. (DMS-1924792).

Institutional Review Board Statement: Ethical review and approval are not required for this study due to using publicly available data.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are publicly available from the UCI data repository mentioned in the manuscript.

Acknowledgments: We would like to thank the reviewers for providing valuable comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest. Charles W. Harrison recently joined Amazon.com, and this work was completed prior to Charles' employment.

References

1. McCullagh, P.; Yang, J. How many clusters? *Bayesian Anal.* **2008**, *3*, 101–120. [CrossRef]
2. Tibshirani, R.; Walther, G.; Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. (Stat. Methodol.)* **2001**, *63*, 411–423. [CrossRef]
3. Fraley, C.; Raftery, A.E. Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.* **2002**, *97*, 611–631. [CrossRef]
4. Scrucca, L.; Fop, M.; Murphy, T.B.; Raftery, A.E. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R J.* **2016**, *8*, 289. [CrossRef] [PubMed]
5. Dahl, D.B.; Day, R.; Tsai, J.W. Random partition distribution indexed by pairwise information. *J. Am. Stat. Assoc.* **2017**, *112*, 721–732. [CrossRef] [PubMed]
6. Aldous, D.J. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*; Springer: Berlin/Heidelberg, Germany, 1985; pp. 1–198.
7. Pitman, J. Exchangeable and partially exchangeable random partitions. *Probab. Theory Relat. Fields* **1995**, *102*, 145–158. [CrossRef]
8. West, M. *Hyperparameter Estimation in Dirichlet Process Mixture Models*; Technical Report; Institute of Statistics and Decision Sciences, Duke University: Durham, NC, USA, 1992.
9. Blei, D.M.; Frazier, P.I. Distance dependent Chinese restaurant processes. *J. Mach. Learn. Res.* **2011**, *12*, 2461–2488.
10. Gretton, A.; Borgwardt, K.M.; Rasch, M.J.; Schölkopf, B.; Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.* **2012**, *13*, 723–773.
11. Dahl, D.B.; Andros, J.; Carter, J.B. Cluster Analysis via Random Partition Distributions. *arXiv* **2021**, arXiv:2106.02760.
12. Killick, R.; Eckley, I. Changepoint: An R package for changepoint analysis. *J. Stat. Softw.* **2014**, *58*, 1–19. [CrossRef]
13. Fritsch, A.; Ickstadt, K. Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Anal.* **2009**, *4*, 367–391. [CrossRef]
14. Fritsch, A.; Fritsch, M.A. *Package 'Mcclust'*. 2009. Available online: http://www2.uaem.mx/r-mirror/web/packages/mcclust/mcclust.pdf (accessed on 14 August 2022).
15. Lawlor, N.; Fabbri, A.; Guan, P.; George, J.; Karuturi, R.K.M. MultiClust: An R-package for identifying biologically relevant clusters in cancer transcriptome profiles. *Cancer Inform.* **2016**, *15*, CIN-S38000. [CrossRef] [PubMed]
16. Fiorini, S. Gene Expression Cancer RNA-Seq Data Set. 2016. Available online: https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq (accessed on 14 August 2022).
17. Lichman, M. *UCI Machine Learning Repository*. 2013 . Available online: https://archive.ics.uci.edu/ml/index.php (accessed on 14 August 2022).
18. Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **2013**, *45*, 1113–1120. [CrossRef] [PubMed]
19. Maechler, M.; Rousseeuw, P.; Struyf, A.; Hubert, M.; Hornik, K. *Cluster: Cluster Analysis Basics and Extensions*; R Package Version 2.1.4—For New Features, See the 'Changelog' File (in the Package Source). 2022. Available online: https://cran.r-project.org/web/packages/cluster/index.html (accessed on 14 August 2022).
20. R CoreTeam. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing, Vienna, Austria, 2013; p. 201.
21. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. (Methodol.)* **1977**, *39*, 1–22.