Adaptive Bayesian variable clustering via structural learning of breast cancer data

Riddhi Pratim Ghosh1 | Arnab K. Maity2 | Mohsen Pourahmadi3 | Bani K. Mallick3

1Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, Ohio, USA
2Early Clinical Development Oncology Statistics, San Diego, California, USA
3Department of Statistics, Texas A&M University, College Station, Texas, USA

Abstract
The clustering of proteins is of interest in cancer cell biology. This article proposes a hierarchical Bayesian model for protein (variable) clustering hinging on correlation structure. Starting from a multivariate normal likelihood, we enforce the clustering through prior modeling using angle-based unconstrained reparameterization of correlations and assume a truncated Poisson distribution (to penalize a large number of clusters) as prior on the number of clusters. The posterior distributions of the parameters are not in explicit form and we use a reversible jump Markov chain Monte Carlo based technique is used to simulate the parameters from the posteriors. The end products of the proposed method are estimated cluster configuration of the proteins (variables) along with the number of clusters. The Bayesian method is flexible enough to cluster the proteins as well as estimate the number of clusters. The performance of the proposed method has been substantiated with extensive simulation studies and one protein expression data with a hereditary disposition in breast cancer where the proteins are coming from different pathways.

KEYWORDS
angular reparameterization, Bayesian clustering, pathways, reversible jump Markov chain Monte Carlo

1 | INTRODUCTION

In cell biology, different pathways emerge as they play different and critical roles in cell functions. Even though the functionality of a cell is an outcome of all the pathway protein expressions as a whole, the individual analysis of each protein has the potential to unveil the complex characterization of the cell biology which is the key to understanding the proper cell function (Ben-Dor et al., 1999). The goal of the clustering is to distill the data down to a more comprehensible level subdividing the omics data (D’haeseleer, 2005). In this article, we focus on the proteomics data and the interest is to infer the pathways based on the proteins data via the clustering technique. Clustering of proteins is a form of unsupervised learning where the proteins are grouped on the basis of some similarity measures inherent among them. Such clusters can be mapped to find the appropriate pathway based on the available protein expression data. Given the functions of the proteins which are measured via the RPPA technology-based protein expressions, it is of interest to track back the pathways to which the group of proteins belongs, assuming that the pathways do not have an overlap.

A proper clustering method that explicates the pattern involved in the gene expression depending on the
overexpression or under-expression of those uncovers the tumor subtypes. For example, Pollack et al. (2002) showed that profiling of DNA copy number variation has the potential to detect more aggressive breast tumors. Washburn et al. (2003) considered the correlation of mRNA and protein expression of amino acid and nucleotide biosynthetic pathway components for clustering. Ben-Dor et al. (1999) provided an algorithm Cluster Affinity Search Technique which uses an affinity measure between nodes of a graph where the genes are represented as the nodes of a graph. In the absence of genuine variable clustering methods, very often traditional data clustering algorithms have been applied to this setup using brute force (Duda et al., 2001; Vigneau & Quannari, 2003), or ad-hoc algorithms based on aspects of correlation matrices have been proposed. We refer the readers to Jiang et al. (2004) for a detailed discussion of various correlation-based clustering approaches that have been previously used in literature for analyzing differential gene expression data. However, in the current era of next-generation sequencing, the amount of data that one receives and the underlying complexity of the pattern often pose challenges for interpretation and understanding of the results, necessitating a proper and meaningful clustering tool.

In this article, our aim is to cluster the proteins, essentially a variable clustering technique that is drastically different from approaches for clustering observations or subjects. To understand it better, let \( Y \) denote a \( n \times k \) data matrix consisting of \( k \) proteins and \( n \) patients, represented in the matrix form

\[
Y = \begin{pmatrix}
    \text{Protein 1} & \text{Protein 2} & \text{Protein 3} & \text{Protein 4} & \ldots & \text{Protein } k \\
    y_{11} & y_{12} & y_{13} & y_{14} & \ldots & y_{1k} \\
    y_{21} & y_{22} & y_{23} & y_{24} & \ldots & y_{2k} \\
    \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
    y_{n1} & y_{n2} & y_{n3} & y_{n4} & \ldots & y_{nk}
\end{pmatrix}
\]

(1)

From Equation (1), one notes that each of \( n \) rows corresponds to one patient and each of \( k \) columns pertains to one protein. A typical data clustering approach partitions the rows of \( Y \), that is, essentially clustering of the patients. We are interested in partitioning the columns of \( Y \) which is essentially clustering of the proteins, and correlations between the proteins serve as our main building block to implement the algorithm. In a typical data clustering algorithm, we consider how similar the objects are based on a similarity norm (say Euclidean or some other kind of distance). On the contrary, in a variable clustering problem, we are concerned with the correlation among the variables. Hence, highly correlated proteins are more likely to lie in the same cluster. As an example, consider a cluster analysis of a set of proteins that belong to different signaling pathways assuming the pathways are not overlapping. The genetic behaviors control the proliferation of a cell or death of a cell, and depending on signals the proteins receive and send, the cell structure is classified into signaling pathways. In turn, one can assume that the similarly expressed proteins belong to the same pathway which can be recovered via variable cluster analysis. Among the different algorithmic clustering techniques commonly used in practice, hierarchical clustering (agglomerative and divisive approach) and partition methods (K-means clustering) rely on a distance metric (Bibby et al., 1979; Friedman et al., 2001; Rokach & Maimon, 2005) without assuming any underlying probability model for the clusters. In addition, the model-based approach usually assumes a mixture model for the data. Even though there is a vast amount of work in the field of data clustering, the variable clustering problem is in its infancy and has gotten limited attention (Bunea et al., 2020). The literature on Bayesian methods for variable clustering is also sparse with a few notable exceptions (Liechty et al., 2004; Palla et al., 2012). Palla et al. (2012) developed a nonparametric Bayes algorithm based on Chinese restaurant process. On the other hand, our method is in the spirit of Liechty et al. (2004) where a parametric model-based approach has been considered. A key advantage of our approach is that the number of clusters is assumed to be unknown apriori, and is determined using a reversible jump Markov Chain Monte Carlo algorithm (RJMCMC) (Green, 1995). In this article, our contributions can be summarized as, first, developing the model-based variable clustering method with block common correlation structures. Second, we propose a novel variable clustering algorithm using the angular representation of the correlations (Ghosh et al., 2021; Pinheiro & Bates, 1996; Rapisarda et al., 2007; Tsay & Pourahmadi, 2017) and the ensuing angles (hyperspherical coordinates). Third, we elicit substantive prior information on these angles which makes clustering of the variables feasible, a data-driven estimate of the number of clusters that traditional algorithms fail to provide. For the posterior inference, since the angle parameters are badly entangled in the posterior distribution, we resort to the Markov chain Monte Carlo algorithm (Tierney, 1994). For the posterior inference, we resort to the standard RJMCMC techniques as in Fan and Sisson (2011), Green (1995), Green and Hastie (2009), and Robert (2004). The rest of the article is organized as follows. In Section 2, we review angular reparameterization of a correlation matrix and present a clustering model through prior specification on the angles. In Section 3, we describe our posterior computation through RJMCMC. Section 4 presents simulation results and clustering of protein expression data. Finally, Section 5 concludes the article.
2 | REVIEW OF ANGULAR REPARAMETRIZATION (Θ) OF R

This section describes connections between the hyperspherical coordinates (angles) and a correlation matrix $R = (r_{ij})$. For a general $k \times k$ correlation matrix $R$ with 1’s in the diagonal, its Cholesky decomposition is given by $R = BB^\top$ where the Cholesky factor $B$ is a lower triangular matrix. Since the rows of $B$ are vectors of unit length, it turns out that it admits the following representation involving trigonometric functions of some angles (Pinheiro & Bates, 1996; Rapisarda et al., 2007):

$$B = \begin{bmatrix}
1 & 0 & 0 & \cdots & 0 \\
c_{21} & s_{21} & 0 & \cdots & 0 \\
c_{31} & c_{32}s_{31} & s_{31} & \cdots & 0 \\
c_{41} & c_{42}s_{41} & c_{43}s_{24} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
c_{k1} & c_{k2}s_{k1} & c_{k3}s_{2k1} & \cdots & 0 \\
& & & & 1
\end{bmatrix}$$

(2)

with $c_{ij} = \cos(\theta_{ij})$ and $s_{ij} = \sin(\theta_{ij})$, where the angles $\theta_{ij}$’s are measured in radians, $1 \leq j < i \leq k$. Restricting $\theta_{ij} \in [0, \pi]$ makes the diagonal entries of $B$ non-negative, and hence $B$ is unique to which we associate a $(k-1) \times (k-1)$ lower triangular matrix $\Theta$ with $k(k-1)/2$ angles:

$$\Theta = \begin{bmatrix}
\theta_{21} & 0 & 0 & \cdots & 0 \\
\theta_{31} & \theta_{32} & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\theta_{k1} & \theta_{k2} & \theta_{k3} & \cdots & \theta_{k,k-1}
\end{bmatrix}$$

(3)

Note that the $(i,j)$ -th element of $\Theta$ is denoted by $\theta_{i+1,j}$ so that $\theta_{ij}$ corresponds to the $(i,j)$ -th element of $R$, we refer to $\Theta$ as the angular matrix associated to $R$. For further details and applications of these angles, see Creal et al. (2011), Zhang et al. (2015), Tsay and Pourahmadi (2017), and Ghosh et al. (2021). One can characterize block diagonal correlation matrices in terms of structured $\Theta$ matrix, which is completely determined by some (pivotal) angles.

2.1 | Correspondence of clustering between $R$ and $\Theta$

**Proposition 1.** For a block diagonal correlation matrix $R = \text{block diag}(R_1, R_2, \ldots, R_m)$, consisting of $m$ equicorrelated blocks $(r_i$ for block $R_i)$, the corresponding angular matrix $\Theta$ is characterized by only $m$ angles $\theta_1, \theta_2, \ldots, \theta_m$, where $r_i = \cos \theta_i$.

**Proof.** The proof is given in Appendix A.1.

It follows immediately from Proposition 1 that in case of block diagonal correlation matrix, clustering on correlations rendering to $m$ different groups is equivalent to clustering of those $m$ angles by the monotonicity of cosine function. However, this will impose some conditions on the pivotal angles to maintain positive definiteness. Assuming each block has dimension $k_i$ so that $\sum_{i=1}^m k_i = k$, the support of $\theta_i$ is $0 < \theta_i < \arccos(1/(k_i - 1))$ for $i = 1, 2, \ldots, m$.

The following proposition describes the separation of two clusters by showing the equivalence of separation on correlations and angles.

**Proposition 2.** Suppose that $r_1 = \cos \theta_1$, $r_2 = \cos \theta_2$. Then $|\theta_1 - \theta_2| \geq \delta$ if and only if $|r_1 - r_2| \geq \delta \cos \delta$.

**Proof.** The proof is given in Appendix A.2.

2.2 | Likelihood function

In this article, we assume throughout that the data $y_1, y_2, \ldots, y_n$ follow a zero mean normal distribution with covariance assumed to be the correlation matrix $R$. As noted in Proposition 1, $R$ can be written as a function of $m$ pivotal angles, $\theta_{\text{pivot}} = (\theta_1, \theta_2, \ldots, \theta_m)^\top$, and hence denoting the transformation from $\theta_{\text{pivot}}$ to $R$ by $T$, the likelihood is proportional to,

$$L(y_1, y_2, \ldots, y_n | \theta_{\text{pivot}}) \propto \det(T(\theta_{\text{pivot}}))^{-n/2}\exp\left\{\frac{-1}{2}S^{-1}(\theta_{\text{pivot}})\right\},$$

(4)

where $S = \sum_{i=1}^n y_i y_i^\top$.

2.3 | Prior specification on the angles

The number of clusters $m$ can take any value in $1, 2, \ldots, k$. Therefore, we assume a truncated Poisson distribution on $m$. Given $m$, define $k \times m$ matrix $Z$ whose $i$-th row corresponds to the allocation of $i$-th variable in one of the $m$ clusters, that is,

$$Z_{ii} = \begin{cases} 
1 & \text{if } i\text{-th variable belongs to } u\text{-th cluster} \\
0 & \text{otherwise}
\end{cases}$$

Since we are assuming that a variable belongs to exactly one cluster, therefore, each row of $Z$ contains exactly one 1 and the rests are 0s. We assume the following hierarchical prior models for model parameters by
assuming a truncated Poisson distribution on number of clusters \( m \) to penalize a large number of clusters, multinomial distribution on each row of the indicator matrix \( Z \) and a Dirichlet distribution for the multinomial hyper-parameters. The hierarchical prior structure is succinctly described as,

\[
m \sim \text{truncPois}(m; 1, k),
\]

\[
q = (q_1, q_2, ..., q_m) \sim \text{Dirichlet}(\alpha_1, \alpha_2, ..., \alpha_m),
\]

\[
Z_i \sim \text{Multinomial}(1; q_1, q_2, ..., q_m) \quad \text{for} \quad i = 1, 2, ..., k,
\]

where \( \alpha_i \)s are any positive numbers and \( \text{truncPois}(m; 1, k) \) is a truncated Poisson distribution supported on the integers in between 1 and \( k \) (number of proteins or variables) for the number of clusters \( m \). Having sampled \( Z \), the allocations are determined. Let \( k_u \) denote the size of \( u \)-th cluster,

\[
k_u = |\{i : z_{iu} = 1\}|,
\]

for \( u = 1, 2, ..., m \).

Then assume the following prior on \( \theta_{pu} = (\theta_1, \theta_2, ..., \theta_m) \) to shrink them to different values.

\[
\theta_{pu|Z, m, \Lambda} = \prod_{u=1}^{m} Q(\theta_u; 0, \arccos\left(\frac{1}{k_u - 1}\right), \lambda_u)
\]

where \( Q(\theta; 0, a, \lambda) \) is the density of truncated wrapped Exponential distribution (Mardia & Jupp, 2009) between 0 and \( a \) with parameter \( \lambda \). We are clustering the pivotal angles by introducing wrapped exponential distribution distribution with different parameters. Suppose \( \Lambda = (\lambda_1, \lambda_2, ..., \lambda_m) \) and we sample \( \lambda_1, \lambda_2, ..., \lambda_m \) in the following manner,

\[
\begin{align*}
\lambda_1 &\sim N^+(\lambda; 0, 1, 0, \infty) \\
\lambda_2 &\sim N^+(\lambda; 0, 1, \lambda_1, \infty) \\
\lambda_i &\sim N^+(\lambda; 0, 1, \lambda_{i-1}, \infty) \quad \text{for} \quad i = 2, 3, ..., m,
\end{align*}
\]

where \( N^+(a; 0, 1, \infty) \) denotes a truncated normal distribution on \((a, \infty)\) with mean 0 and variance 1 which has the following density,

\[
f(\lambda; \mu = 0, \sigma = 1, a, \infty) = \frac{\phi(\lambda)}{1 - \Phi(a)},
\]

where \( \phi \) and \( \Phi \) are the density and distribution functions of a standard normal distribution respectively.

The salient features of the prior formulation of \( \lambda_i \)s given in (10) are the followings: (1) The prior mean for the \( i \)-th pivotal angles is \( \mathbb{E}(\theta_i) = \arctan(1/\lambda_i) \) for \( i = 1, 2, ..., m \). Since these angles vary in \([0, \pi]\), \( \lambda_i \)s take value on positive real line. (2) Also \( \lambda_i \)'s satisfy \( \lambda_1 < \lambda_2 < \lambda_2 < \cdots < \lambda_m \), which enforces separation of clusters through prior model.

### 3 | POSTERIOR COMPUTATION

With the likelihood function (4) and prior specified in Section 2.3, the posterior distribution is proportional to

\[
\begin{align*}
p(\Theta, Z, \Lambda, m|y_1, y_2, ..., y_n) \\
\propto & L(y_1, y_2, ..., y_n|\Theta, Z, m) \times p(m) \times p(q|m) \\
& \times p(Z|m, q) \\
& \times p(\Theta|\Lambda, Z, m) \times p(\Lambda)
\end{align*}
\]

Our goal in this section is to estimate number of clusters \( m \) and posterior of \( Z \). The algorithm is, thus, accomplished by performing a reversible jump Markov chain Monte Carlo(RJMCMC) algorithm.

From proposed priors, one can note that the clusters are induced by the elements of \( \Lambda \), thus, in the following RJMCMC algorithm (Fan & Sisson, 2011; Green, 1995; Green & Hastie, 2009; Robert, 2004), at each iteration either one element of \( \Lambda \), say \( \lambda_j \) is randomly split into \( (\lambda_{j1}, \lambda_{j2}) \) (Birth step) or two elements of \( \Lambda \) are merged into a single element (Death step). The algorithm is summarized as follows.

**Step 1.** Initialize \( \Theta, \Lambda \). In the initialization step, we assume that there are at least two different clusters to circumvent the possibility of the death step in the beginning. Therefore, \( \Lambda = (\lambda_1, \lambda_2)^T \), where \( \lambda_1 \) and \( \lambda_2 \) are simulated according to (10). Then \( \Theta \) is given by two elements \( \Theta = (\theta_1, \theta_2)^T \), and \( \theta_1 \sim Q(\theta; 0, \arccos(1/k-1), \lambda_1) \), and \( \theta_2 \sim Q(\theta; 0, \arccos(1/k-1), \lambda_2) \), where \( k \) is the nearest integer of \( k/2 \) and \( Q(\cdot) \) is the pdf of a truncated wrapped Exponential distribution defined in Equation (9). It is instructive to note that in the initial stage we take two clusters of almost equal size. This initial choice gets rid of any additional tuning. However, one can also choose other number of balanced clusters with the similar choices of \( \Lambda \) and \( \Theta \).

**Step 2.** A particular iteration, say \( q \)-th iteration consists of a Birth step and a Death step.
Birth Step: Split \( \lambda^{(q)}_{j} \) to \( (\hat{\lambda}^{(q)}_{j}, \lambda^{(q)}_{j})^\top \) by \( \lambda^{(q)}_{j} = \lambda^{(q)}_{j} + \tau, \lambda^{(q)}_{j} = \lambda^{(q)}_{j} - \tau \), where \( \tau \sim \text{Unif}(-\pi/4, \pi/4) \) and dimension of \( \lambda^{(q)} \) is increased by 1 with acceptance probability

\[
\alpha = \min \left\{ 1, \frac{p(\theta^{(q)}, \hat{\lambda}^{(q)}_{n}, \hat{\theta}^{(q)}, \hat{\lambda}^{(q)}_{n}, \lambda^{(q)}_{n}, \theta^{(q)}, \lambda^{(q)}_{n},\lambda^{(q)}_{n}, \theta^{(q)}, \lambda^{(q)}_{n})}{p(\theta^{(q)}, \hat{\lambda}^{(q)}_{n}, \hat{\theta}^{(q)}, \hat{\lambda}^{(q)}_{n}, \lambda^{(q)}_{n}, \theta^{(q)}, \lambda^{(q)}_{n},\lambda^{(q)}_{n}, \theta^{(q)}, \lambda^{(q)}_{n})} \right\}
\]

Death step: Two components \( \lambda^{(q)}_{j} \) and \( \lambda^{(q)}_{j} \) are merged to a single component \( \lambda^{(q)}_{j} = (\hat{\lambda}^{(q)}_{j} - \tau + \lambda^{(q)}_{j} + \tau) / 2 \) with acceptance probability

\[
\alpha = \min \left\{ 1, \frac{p(\theta^{(q)}, \hat{\lambda}^{(q)}_{n}, \hat{\theta}^{(q)}, \hat{\lambda}^{(q)}_{n}, \lambda^{(q)}_{n}, \theta^{(q)}, \lambda^{(q)}_{n},\lambda^{(q)}_{n}, \theta^{(q)}, \lambda^{(q)}_{n})}{p(\theta^{(q)}, \hat{\lambda}^{(q)}_{n}, \hat{\theta}^{(q)}, \hat{\lambda}^{(q)}_{n}, \lambda^{(q)}_{n}, \theta^{(q)}, \lambda^{(q)}_{n},\lambda^{(q)}_{n}, \theta^{(q)}, \lambda^{(q)}_{n})} \right\}
\]

Step 3. Steps 1 and 2 are repeated as many times as required to ensure convergence and the value of \( m \) is determined by which stage is visited maximum number of times, maximum aposteriori estimate (MAP) and posterior estimate of \( Z \) is obtained by averaging over those stages.

### 4 | SIMULATIONS AND DATA ANALYSES

In this section, we compare the numerical performance of our Bayesian variable clustering (BVC) algorithm with a recent method based on covariance difference (COD) of Bunea et al. (2020), partitioning around medoids (PAM) algorithm which minimizes the Manhattan distance of the data points to the medoids (Kaufman & Rousseeuw, 2009) and the classical or standard K-means clustering algorithm. The performance criterion we use is the adjusted rand index (ARI) (Santos & Embrechts, 2009) which is defined as

\[
\text{ARI} = \frac{RI - \text{Expected} \ RI}{\text{Max} \ RI - \text{Expected} \ RI},
\]

where the rand index \( RI = \frac{\text{Number of Agreeing Pairs}}{\text{Number of Pairs}} \). For each instance, we generate 50 replications, and the expected rand index is computed by taking the average RI over 50 replications.

COD and PAM have been implemented using the R packages cord (Luo et al., 2015) and class (Venables & Ripley, 2002) available via CRAN and K-means algorithm has been implemented on the transposed data matrix using \textit{kmeans()} function in R software (R Core Team, 2020). It is instructive to note that the quantity in (12) takes value in the interval \([0,1]\). As the value approaches to 1, the recovery becomes better.

#### 4.1 | Simulation study

We start with an \( m \times m \) matrix \( C = B^TB \) where the entries of the random \((m - 1) \times m \) matrix \( B \) take values -1, 0, 1 with probabilities \( 0.5 \times m^{-1/2}, 1 - m^{-1/2} \) and \( 0.5 \times m^{-1/2} \), respectively, with \( m \) being the number of clusters. Next, we consider a balanced case with each group (cluster) of size \( k/m \). Let \( A = (a_{ij}) \) be the \( k \times m \) membership matrix with \( a_{ij} = 1 \) if the \( i \)-th variable belongs to the \( j \)-th cluster and 0 otherwise. Finally, consider the covariance matrix \( \Sigma = \text{ACA}^\top + \Gamma \) where \( \Gamma \) is a diagonal matrix whose entries are random permutations of \( \{0.5, 0.5 + 1.5/(k - 1), ..., 2\} \) and the corresponding correlation matrix \( R \). With \( k = 200, m = 4 \), we simulate \( n \) independent observations from (I) a multivariate normal distribution, and (II) a multivariate t-distribution with mean zero vector and covariance matrix \( R \), where we vary \( n \) in 100, 300, 600, 900 to compare BVC, COD and K-means algorithms with respect to cluster recovery criterion in (12). Under both (I) and (II), we considered a normal likelihood as in Equation (4). Therefore, (I) renders to the correct model while (II) corresponds to the misspecified model. The results are summarized in Figure 1.

One can note that when the true model is normal, our BVC method outperforms the other three methods for all specified values of the sample size. As \( n \) increases, the performance of the COD becomes better and close to the BVC. In the case of the misspecified model, the BVC method still performs well when \( n \) is small. The ARIs for BVC and COD are 0.70 and 0.33, respectively when \( n = 100 \). When \( n \) is large, the performance of BVC and COD are comparable. The performances of PAM and K-means are inferior to our BVC method irrespective of the sample size and model specification. One point worth mentioning is that BVC is a model-based approach while the other three methods are not. Therefore, under model misspecification, the degradation of the performance of BVC is natural when compared to the correct model. These puts together show the robustness of the BVC method as a better variable clustering algorithm.

#### 4.2 | Application of protein clustering to hereditary breast cancer data

Breast cancer is one of the most common cancers with a massive number of cases reported. For instance, in 2018,
more than 268,000 Americans were estimated to have been diagnosed and 41,000 were estimated to have died from breast cancer-related tumors (Bray et al., 2018). The Cancer Genome Atlas: TCGA is the largest available cancer data consortium consisting of parallel mRNA expressions, DNA copy number, methylation expressions, and protein expressions, along with clinical variables such as survival or the tumor stages for a total of 33 types of tumors. Among them we consider the information of 222 breast tumor samples; we consider 27 different proteins and four different pathways (see Appendix A.3).

Applying our BVC algorithm to this data, the MAP estimate of the number of clusters is 4, which is consistent with the number of pathways. However, applying the COD algorithm in Bunea et al. (2020) the estimated number of clusters is 23, much larger than the known value of 4. In Appendix A.4, we provide the assignments of various proteins in different clusters. Additionally, for the sake of comparison, we have also applied the K-means algorithm to this data for \( k = 4, 23 \), respectively, with results reported in Appendix A.4. The results suggest that our BVC is performing better to cluster the proteins with respect to pathways. Only misclassified proteins are MAPK_pT201_Y204, CD31, CD49b, and CDK1. The COD algorithm reports that the number of clusters is 23 which appears to be too high since the number of proteins is 27. A possible reason could be this algorithm is meant for high dimensional clustering, it fails to detect clustering configuration in small dimensional cases. Comparisons with standard K-means and PAM algorithms also reveal that these two methods result in more disagreement on the cluster configuration of the proteins according to the pathway information. This apart, K-means and PAM algorithm disagree among themselves, for example, ER-alpha, JNK_pT183_pT185, and so on (Table A2). We have also performed hierarchical clustering on this data with various linkages. The results are presented in Figure 2.

For the sake of checking the stability of our proposed BVC method, we compared 95% credible interval of the estimated number of clusters of BVC with the 95% confidence interval obtained via COD. In BVC, 10,000 samples of \( k \) after the convergence of the RJMCMC with a burn-in of 5000 are collected and ordered, and 95% credible interval is reported as the range between 251 and 9750 ordered samples. For COD, we report 95% bootstrap confidence interval with 10000 iterations where at each iteration we sample 222 breast tumor samples with replacement and apply the COD algorithm to estimate the
number of clusters. The 95% credible interval of $k$ via BVC is $[3, 7]$, and confidence interval of $k$ via COD is $[18, 25]$. Clearly, the credible interval obtained via BVC is preferred because it is much narrower and contains the actual number of different pathways. It is instructive to note that $k$ can only take integer values in the above intervals.

5 | DISCUSSION

We have proposed a correlation matrix-based Bayesian clustering technique to recover the protein signaling pathways. This method uses angular reparameterization of the correlation matrix with the specification of wrapped exponential before the angle parameters. Nonetheless, as an alternative, one can use any truncated circular distribution as prior for pivotal angles, for example, von-Mises distribution. However, this particular choice produces a mean which has no closed-form and as a result, our proposed method can not be carried out for posterior analysis. A large amount of recent interest is being channelized to analyze the proteomics data directly because direct analysis of proteins has the potential to uncover the cell functional characteristics. When it is of interest to find the group of proteins having similar functions which may be evident via their expression measurements then our proposed method can be used to bridge that gap. As mentioned earlier, our method is particularly useful when the number of clusters is not known and hence is learned via the posterior MCMC, which is often the case for the real data where determining the number of clusters is itself a tedious job.

ACKNOWLEDGMENTS

The research reported in this paper was supported by grant from the National Institutes of Health (R01CA194391). The authors are indebted to the editor and the referee for their constructive comments which helped to improve this article considerably.

DATA AVAILABILITY STATEMENT

The data is available as an open source or can be requested from the authors. The data that support the findings of this study are available in TCGA at https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga. These data were derived from the following resources available in the public domain: TCPA, https://tcpaportal.org/tcpa/.

ORCID

Riddhi Pratim Ghosh http://orcid.org/0000-0001-8417-0944
Arnab K. Maity http://orcid.org/0000-0002-6692-0155

REFERENCES

Ben-Dor, A., Shamir, R., & Yakhini, Z. (1999). Clustering gene expression patterns. Journal of Computational Biology, 6(3-4), 281–297.
and the fact that Cholesky factor of a block diagonal matrix is also block diagonal and vice versa.
Hence, the lower triangular Cholesky factor $R$ has the form: $B = \text{block diag}(B_1, B_2, \ldots, B_m)$, where $B_i$ is an upper triangular Cholesky factor of $R_i$.

The proof will be complete if we can show that Cholesky factor $B$ of a compound symmetric correlation matrix $R$ can be written in terms of only one angle. From the relationship between angles and Cholesky factor, it follows that $\cos(\theta_{ii}) = b_{ii} = r$ for $j = 2, 3, \ldots, k$. Thus $\theta_{ii}$'s are all equal to a common $\theta$. For any $i > j$, the proof follows by induction. For $i > j$, $r_{ij} = \sum_{l=1}^{i-1} b_{il} b_{jl} + b_{ij} b_{ij}$. By induction hypothesis, all the preceding angles and Cholesky factors are functions of $r$. Thus, the first term is a function of $r$. For the second term, we note $b_{ij} = \prod_{l=1}^{j-1} \sin(\theta_{il})$, which involves all the preceding angles and thus a function of $r$ and $\theta_{ij}$ is a function of $\theta$.

### A.2 Proof of proposition 2

**Proof.** First consider $|\theta_1 - \theta_2| = \delta$. Note that $|r_1 - r_2| = \int_{\theta_1}^{\theta_2} \sin x dx$. Also it is clear that $|r_1 - r_2|$ is an increasing function of $|\theta_1 - \theta_2|$, since $\sin$ is positive in $[0, \pi)$. Now since $\sin$ is increasing in $[0, \pi/2]$ and decreasing in $(\pi/2, \pi)$, $|r_1 - r_2|$ will take minimum value for $\theta_1 - \theta_2 = \delta$ when $\theta_1 = 0$, $\theta_2 = \delta$. Thus the minimum value of $|r_1 - r_2|$ is $1 - \cos \delta$. □

### A.3 Pathway information

See Table A1 for pathway information of the proteins.

### A.4 Cluster assignments of proteins

Table A2 presents the cluster assignments of proteins by BVC, COD, PAM and K-means algorithms.

#### TABLE A1 Pathway protein list

| MAP kinase        | PI3K/AKT/mTOR | JAK-STAT | Wnt     |
|-------------------|---------------|----------|---------|
| ER-alpha          | AKT           | SHC_pY317| CD31    |
| ER-alpha_pS118    | AKT_pS473     | STAT3_pY705| CD49b  |
| ERK2              | AKT_pT308     | STAT5-alpha| CDK1    |
| JNK2              | FOXO3a        | Cyclin_D1|         |
| JNK_pT183_pT185   | PTEN          | Fibronectin|         |
| MAPK_pT202_Y204   | mTOR          | GSK3-alpha-beta|       |
| p38_MAPK          | mTOR_pS2448   | GSK3-alpha-beta_pS21_S9|       |
| p38_pT180_Y182    |               | VEGFR2   | Beta-Catenin|
| Protein                  | Bayesian variable clustering (BVC) | Covariance difference (COD) | K-means \( (m = 4) \) | Partitioning around medoids \( \text{(PAM)} \) \( (m = 4) \) | K-means \( (k = 23) \) |
|--------------------------|-----------------------------------|-----------------------------|------------------------|-------------------------------------------------|------------------------|
| ER-alpha                 | \( C_1 \)                         | \( C_1 \)                   | \( C_3 \)              | \( C_1 \)                                        | \( C_2 \)              |
| ER-alpha_pS118           | \( C_1 \)                         | \( C_1 \)                   | \( C_1 \)              | \( C_2 \)                                        | \( C_7 \)              |
| ERK2                     | \( C_1 \)                         | \( C_2 \)                   | \( C_1 \)              | \( C_2 \)                                        | \( C_{15} \)            |
| JNK2                     | \( C_1 \)                         | \( C_5 \)                   | \( C_1 \)              | \( C_2 \)                                        | \( C_{10} \)            |
| JNK_pT183_pT185          | \( C_1 \)                         | \( C_6 \)                   | \( C_4 \)              | \( C_2 \)                                        | \( C_8 \)              |
| MAPK_pT202_Y204          | \( C_3 \)                         | \( C_7 \)                   | \( C_2 \)              | \( C_3 \)                                        | \( C_{112} \)           |
| p38_MAPK                 | \( C_1 \)                         | \( C_8 \)                   | \( C_4 \)              | \( C_2 \)                                        | \( C_9 \)              |
| p38_pT180_Y182           | \( C_1 \)                         | \( C_9 \)                   | \( C_4 \)              | \( C_2 \)                                        | \( C_{22} \)            |
| AKT                      | \( C_2 \)                         | \( C_{10} \)                | \( C_1 \)              | \( C_2 \)                                        | \( C_{11} \)            |
| AKT_pS473                | \( C_2 \)                         | \( C_{11} \)                | \( C_2 \)              | \( C_4 \)                                        | \( C_1 \)              |
| AKT_pT308                | \( C_2 \)                         | \( C_{12} \)                | \( C_2 \)              | \( C_4 \)                                        | \( C_1 \)              |
| FOXO3a                   | \( C_2 \)                         | \( C_{13} \)                | \( C_4 \)              | \( C_2 \)                                        | \( C_5 \)              |
| PTEN                     | \( C_2 \)                         | \( C_{14} \)                | \( C_1 \)              | \( C_2 \)                                        | \( C_{13} \)            |
| mTOR                     | \( C_2 \)                         | \( C_{15} \)                | \( C_1 \)              | \( C_2 \)                                        | \( C_{21} \)            |
| mTOR_pS2448              | \( C_2 \)                         | \( C_{16} \)                | \( C_1 \)              | \( C_2 \)                                        | \( C_{20} \)            |
| SHC_pY317                | \( C_3 \)                         | \( C_{17} \)                | \( C_4 \)              | \( C_2 \)                                        | \( C_{17} \)            |
| STAT3_pY705              | \( C_3 \)                         | \( C_{18} \)                | \( C_4 \)              | \( C_2 \)                                        | \( C_{23} \)            |
| STAT5-alpha              | \( C_3 \)                         | \( C_2 \)                   | \( C_1 \)              | \( C_2 \)                                        | \( C_{18} \)            |
| CD31                     | \( C_2 \)                         | \( C_3 \)                   | \( C_4 \)              | \( C_2 \)                                        | \( C_6 \)              |
| CD49b                    | \( C_2 \)                         | \( C_{19} \)                | \( C_4 \)              | \( C_2 \)                                        | \( C_3 \)              |
| CDK1                     | \( C_2 \)                         | \( C_3 \)                   | \( C_4 \)              | \( C_2 \)                                        | \( C_5 \)              |
| Cyclin_D1                | \( C_4 \)                         | \( C_{20} \)                | \( C_4 \)              | \( C_2 \)                                        | \( C_{16} \)            |
| Fibronectin              | \( C_4 \)                         | \( C_{21} \)                | \( C_4 \)              | \( C_2 \)                                        | \( C_4 \)              |
| GSK3-alpha-beta          | \( C_4 \)                         | \( C_4 \)                   | \( C_1 \)              | \( C_2 \)                                        | \( C_{21} \)            |
| GSK3-alpha-beta_pS21_S9  | \( C_4 \)                         | \( C_{22} \)                | \( C_2 \)              | \( C_4 \)                                        | \( C_{19} \)            |
| VEGFR2                   | \( C_4 \)                         | \( C_{23} \)                | \( C_1 \)              | \( C_2 \)                                        | \( C_3 \)              |
| beta-Catenin             | \( C_4 \)                         | \( C_4 \)                   | \( C_1 \)              | \( C_2 \)                                        | \( C_{14} \)            |