Affine-transformation invariant clustering models

Hsin-Hsiung Huang1* and Jie Yang2

1 Department of Statistics and Data Science, University of Central Florida, Orlando, USA

Full list of author information is available at the end of the article

Abstract

We develop a cluster process which is invariant with respect to unknown affine transformations of the feature space without knowing the number of clusters in advance. Specifically, our proposed method can identify clusters invariant under (I) orthogonal transformations, (II) scaling-coordinate orthogonal transformations, and (III) arbitrary nonsingular linear transformations corresponding to models I, II, and III, respectively and represent clusters with the proposed heatmap of the similarity matrix. The proposed Metropolis-Hasting algorithm leads to an irreducible and aperiodic Markov chain, which is also efficient at identifying clusters reasonably well for various applications. Both the synthetic and real data examples show that the proposed method could be widely applied in many fields, especially for finding the number of clusters and identifying clusters of samples of interest in aerial photography and genomics data.

Keywords: Dirichlet process, Ewens process, Metropolis-Hastings algorithm, Markov chain Monte Carlo sampling, Unsupervised learning

1 Introduction

Clustering of objects invariant with respect to affine transformations of feature vectors is an important research topic since objects may be recorded via different angles and positions so that their coordinates may vary and their nearest neighbors may belong to other clusters. For example, the longitude, latitude, and altitude coordinates of an object which are recorded by devices equipped in aircrafts or satellites change across different observational time. In this situation, distance-based clustering method including $k$-means (MacQueen 1967), hierarchical clustering (Ward 1963), clustering based on principal components, spectral clustering (Ng et al. 2001), and others (Jain and Dubes 1988; Ozawa 1985) may fail to identify the correct clusters by grouping nearest points. Another category is distribution-based clustering methods (Banfield and Raftery 1993; Fraley and Raftery 1998; Fraley and Raftery 2002; Fraley and Raftery 2007; McCullagh and Yang 2008; Vogt et al. 2010) which may specify a partition as a parameter in a likelihood function and estimate it under a Bayesian framework.

In certain areas of application, the goal is to cluster objects $i = 1, \ldots, n$ into disjoint subsets based on their feature vectors $Y_i \in \mathbb{R}^d$. In this paper, we propose group
invariance by considering three cases of a cluster process that are invariant with respect to three groups of affine transformations \( g: \mathbb{R}^d \rightarrow \mathbb{R}^d \) acting on the feature space. The group invariance implies that the feature configurations \( Y \) and \( Y' \) in \( \mathbb{R}^{n \times d} \) determine the same clustering, or probability distribution on clusterings, if they belong to the same group orbit that is an equivalence class. For example, if the feature space is Euclidean and \( G \) is the group of Euclidean isometries or congruences, the clustering is a function only of the maximal invariant, which is the array of Euclidean distances \( D_{ij} = \|Y_i - Y_j\| \). For example, image data such as the aerial photography and three-dimensional protein structures are two motivating examples. The shape and relative locations of data may vary due to the change of the viewer's angle and location.

Our goal is to develop a novel clustering method which can identify clusters of \( Y = (Y_1, \ldots, Y_n) \) even when all \( Y_i \)’s are mapped by an unknown affine transformation \( Y'_i = a + AY_i \), where \( a = (a_1, \ldots, a_d) \in \mathbb{R}^d \) and \( A \in \mathbb{R}^{d \times d} \) is nonsingular. Affine-invariant clustering is important when the clusters are not well-separated in the observational space. Although there are previous work on affine-invariant clustering methods (Fitzgibbon and Zisserman 2002; Begelfor and Werman 2006; Shioda R. and Tunçel 2007; Brubaker. S.C. and Vempala 2008; Kumar and Orlin 2008; García-Escudero et al. 2010; Lee et al. 2014), these existing methods handle different problems from ours. These methods aim to cluster the same item observed in different angles or mapped by different unknown affine transformations. Instead, in our problem setting we consider only one unknown affine transformation that is applied to all objects.

The affine transformations consist of three types: (1) index permutations, rotation, one-scaling on all variables, and location-translation transformations that are under the first type of covariance structures and named model I whose transformation and covariance structure \( \sigma^2 I_d \) were also adopted by Vogt et al. (2010); (2) each variable may have different scaling transformations that are under the second type of covariance structures and named model II; (3) the variables are transformed by a nonsingular matrix that is named model III, where the observed variables may be linear combinations of some latent variables in model I. These models cover fairly general situations of clustering in nature.

McCullagh and Yang (2008) constructed a Dirichlet cluster process together with a random partition representing the clustering. In this paper, we follow their setup and extend their framework. We assume that the random partition of objects follows Ewens distribution (Ewens 1972), and we propose a likelihood of the responses which is invariant respect to affine transformations.

## 2 Cluster process and prior distributions

In this paper, an \( \mathbb{R}^d \)-valued cluster process \((Y, B)\) means a random partition \( B \) of the natural numbers, together with an infinite sequence \( Y_1, Y_2, \ldots \) of random vectors in the state space \( \mathbb{R}^d \). The restriction of such a process to a finite sample \([n] = \{1, \ldots, n\}\) of units or specimens consists of the restricted partition \( B[n] \) accompanied by the finite sequence \( Y[n] = (Y_1, \ldots, Y_n) \). A partition \( B[n]: [n] \times [n] \rightarrow \{0, 1\} \) is the partition of the sample units expressed as a binary cluster-factor matrix of \( B_{ij} = 1 \) if \( Y_i \) and \( Y_j \) are of the same cluster (denoted as \( i \sim j \)), and \( B_{ij} = 0 \) otherwise (McCullagh and Yang 2008). For example, when \( n = 3 \), the partition \( \{(1, 2), 3\} \) and the cluster labels 112 correspond to an equivalence relation.
\[ B = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \]

Notice that the elements of \( B \) are transitional. i.e., if individuals \( i,j,k \) belong to the same cluster, then \( B_{ij} = 1 \) and \( B_{jk} = 1 \) imply \( B_{ik} = 1 \).

The term cluster process implies infinite exchangeability, which means that the joint distribution \( p_n \) of \( (Y[n], B[n]) \) is symmetric (McCullagh and Yang 2006) or invariant under permutations of indices (Pitman 2002), and \( p_n \) is the marginal distribution of \( p_{n+1} \) under deletion of the \((n+1)\)th unit from the sample.

Similar to (McCullagh and Yang 2008), we construct an exchangeable Gaussian mixture as a simple example of clustering processes. First, \( B \sim p \) is some infinitely exchangeable random partition. Secondly, the conditional distribution of the samples \( Y \), which is regarded as a matrix \((Y_i)\) of order \( n \times d \) given \( B \) (say the cluster label \( cl(Y_i) = l \)) and \( \theta \), is Gaussian with mean and variance as follows

\[
E(Y_i | B, \mu_l) = \mu_l, \quad \text{Cov}(Y_{i,r}, Y_{j,s} | B, \theta) = (\delta_{ij} + \theta B_{ij}) \Sigma_{rs},
\]

where \( \mu_l = (\mu_{1l}, \ldots, \mu_{dl}) \in \mathbb{R}^d \) is the centroid of cluster \( k \), \( \delta \) is Kronecker’s delta, that is, \( \delta_{ij} = 1 \) if \( i = j \) and \( 0 \) if \( i \neq j \), \( \theta > 0 \) is a ratio parameter connecting the within- and between-cluster covariance matrices, and \( \Sigma = (\Sigma_{rs}) \) is a positive definite matrix of order \( d \times d \), known as the within-cluster covariance matrix. In our settings, the between-cluster covariance matrix is simply \( \theta \Sigma \), the cluster centroids \( \mu_1, \ldots, \mu_k \) are i.i.d from \( N(\mu, \theta \Sigma) \), and the mean of \( Y \) given \( B \) and \( \mu_1, \ldots, \mu_k \) is

\[
E(Y | B, \mu_1, \ldots, \mu_k) = (\mu_{cl(Y_1)}, \ldots, \mu_{cl(Y_n)})
\]

and the covariance of \( Y \) given \( B \) can also be represented by the covariance of its vector form \( \text{Vec}(Y) = (Y_{11}, \ldots, Y_{1d}, \ldots, Y_{nk}, \ldots, Y_{nd})^T \) as

\[
\text{Cov}(\text{Vec}(Y) | B, \theta) = (I_n + \theta B) \otimes \Sigma
\]

which is an \( nd \times nd \) matrix with “\( \otimes \)” indicating the Kronecker product. \( \Sigma \), the column covariance of \( Y \), is assumed identical for all clusters, \( I_n + \theta B \) is assumed an exchangeable structure for the row covariance of \( Y \), and \( \theta \) is the product of the standard deviations of two rows. There exist competing algorithms that are affine-equivariant and do note impose this requirement (Shioda R. and Tunçel 2007; Kumar and Orlin 2008; García-Escudero et al. 2010; Lee et al. 2014). The identity matrix itself is also a partition in which each cluster consists of one element.

Given the number of clusters \( k \), the cluster sizes \((n_1, \ldots, n_k)\) may follow a multinomial distribution with category probabilities \( \pi = (\pi_1, \ldots, \pi_k) \), where \( \pi \) follows an exchangeable Dirichlet distribution \( \text{Dir}(\lambda/k, \ldots, \lambda/k) \). After integrating out \( \pi \), the partition \( B \) follows a Dirichlet-multinomial prior

\[
p_n(B|\lambda, k) = \frac{k!}{(k - \#B)!} \frac{\Gamma(\lambda) \prod_{b \in B} \Gamma(n_b + \lambda/k)}{\Gamma(\mu + \lambda) \prod_{b \in B} \Gamma(\mu + \lambda/k)}^{\#B}
\]

where \( \#B \leq k \) denotes the number of clusters presented in the partition \( B \) and \( n_b \) is the size of cluster \( b \) (MacEachern 1994; Dahl 2005; McCullagh and Yang 2008). The limit as \( k \to \infty \) is well defined and known as the Ewens’s sampling formula (ESF) with parameter \( \lambda > 0 \).
$$p_n(n_1, \ldots, n_k | \lambda) = \frac{\Gamma(\lambda) \lambda^{\#B}}{\Gamma(n+\lambda)} \prod_{b \in B} \Gamma(n_b),$$

which is also known as Chinese restaurant process (CRP) (Ewens 1972; Neal 2000; Blei and Jordan 2006; Crane 2016). McCullagh and Yang (2008) provided a framework with a proportional configurations if there exists a vector \( (n_1, \ldots, n_k) \) of conditional probabilities. Since the conditional distribution of any permutation \( \pi \) implies that any permutation \( \pi \) of the sequence induces a corresponding permutation \( \pi \), where \( \pi(i) \) and \( \pi(j) \) are equal. The configurations are geometrically equivalent in \( \mathbb{R}^d \) should determine the same conditional distribution on sample partitions.

If the state space \( \mathbb{R}^d \) is regarded as a \( d \)-dimensional Euclidean space with the standard Euclidean inner product and Euclidean metric, the configurations \( Y \) and \( Y' \) are congruent if there exists a vector \( a \in \mathbb{R}^d \) and an orthogonal matrix \( A \in \mathbb{R}^{d \times d} \) such that \( Y' = a + AY \) for each \( i \). Equivalently, the \( n \times n \) arrays of squared Euclidean distances \( D_{ij} = \|Y_i - Y_j\|^2 \) and \( D'_{ij} = \|Y'_i - Y'_j\|^2 \) are equal. The configurations are geometrically similar if \( Y'_i = a + bY_i \) for \( b \in \mathbb{R} \) and \( b \neq 0 \), implying that the arrays of distances are proportional \( D' = b^2 D \).

The geometric equivalence is defined by regarding the observation \( Y \) as a group orbit rather than a point. In general, the group is the affine group \( GA(\mathbb{R}^d) \), \( G = \mathbb{R}^d \times L \) and \( L \) is the collection of all \( d \times d \) nonsingular matrices, with the operation \( (a_1, A_1) \circ (a_2, A_2) = (a_1 + A_1 a_2, A_1 A_2) \) for \( a_i \in \mathbb{R}^d, A_i \in L \) with \( i = 1, 2 \), which is consistent with compositions of affine transformations. The orbit of an element \( Y = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^{n \times d} \) is defined as

$$\text{Orb}(Y) = \{ X \in \mathbb{R}^{n \times d} : \exists g \in G \text{ s.t. } X = g \ast Y \}.$$

We choose a proper prior distribution for the variance ratio \( \theta \), the symmetric \( F \)-family

$$p(\theta) \propto \frac{\theta^{\alpha-1}}{(1 + \theta)^{2\alpha}}$$

with \( \alpha > 0 \) allowing a range of reasonable choices (Chaloner 1987).

We propose a sampling procedure to estimate the partition \( B \) and the parameter \( \theta \) from conditional probabilities. Since the conditional distribution of \( \theta \) does not have a recognized form, we propose to use a discrete version \( \{ p(\theta_j) \}_{j=1}^J \), where \( J \) is a predetermined moderately large integer. Based on our experience, \( J = 100 \) works reasonably well for the real data examples that we have examined.
where the group action is that \( G \) acts on \( \mathbb{R}^{n \times d} \) as

\[
(a, A) \ast Y = (a + AY_1, \ldots, a + AY_n)^T = 1_n a^T + YA^T \tag{2}
\]

where \( 1_n \) is a length-\( n \) vector of 1’s. It can be verified that its vector form \( \text{Vec}((a, A) \ast Y) = 1_n \otimes a + (I_n \otimes A) \text{Vec}(Y) \). If

\[
\text{Vec}(Y) \sim N(1_n \otimes \mu, (I_n + \theta B) \otimes \Sigma),
\]

then an element in the same orbit

\[
\text{Vec}((a, A) \ast Y) \sim N(1_n \otimes (a + A\mu), (I_n + \theta B) \otimes (A\Sigma A^T))
\]

More specifically,

\[
\begin{align*}
\text{Vec}((-\mu, I_d) \ast Y) & \sim N(0, (I_n + \theta B) \otimes \Sigma) \\
\text{Vec}((-\theta^{-1}\mu, \theta^{-1}) \ast Y) & \sim N(0, (I_n + \theta B) \otimes I_d)
\end{align*}
\]

where \( T \) is a \( d \times d \) nonsingular matrix satisfying \( \Sigma = TT^T \).

**Theorem 1** If \( n \leq d \), then all \( Y \in \mathbb{R}^{n \times d} \) of full rank \( n \) belong to the same orbit. If \( n = d + 1 \), then all \( Y \in \mathbb{R}^{n \times d} \) satisfying \( \text{rank}(Y) = \text{rank}(Y - 1_n 1_d^T Y/n) = d \) belong to the same orbit.

The proof of Theorem 1 is relegated to the Appendix A. According to the proof, if \( n = d + 1 \), then \( \text{rank}(Y) = d \) implies that \( \text{rank}(Y - 1_n 1_d^T Y/n) \) is either \( d \) or \( d - 1 \). The case of \( d - 1 \) only occupies a lower-dimensional subspace.

According to Theorem 1, for \( n \leq d + 1 \), the action is essentially transitive in the sense that any configuration of \( n \) distinct points in \( \mathbb{R}^d \) belong to the same orbit: all other orbits are negligible in that they have Lebesgue measure zero. As a result, the observation \( Y \) regarded as a group orbit \( GY \) is uninformative for clustering unless \( n > d + 1 \). We name the orbit and group action defined above as model III.

In model I, which is the case considered in Vogt et al. (2010), the covariance between features are proportional to an identity matrix. The group is \( G = \mathbb{R}^d \times \mathbb{R} \setminus \{0\} \) with the operation \((a_1, b_1) \circ (a_2, b_2) = (a_1 + b_1 a_2, b_1 b_2) \) for \( a_i \in \mathbb{R}^d, b_i \in \mathbb{R} \setminus \{0\}, i = 1, 2 \). The orbit of an element \( Y \in \mathbb{R}^{n \times d} \) and the group action are defined similarly as in (1) and (2) with \( A \) replaced by \( b \). Then

\[
(a, b) \ast Y = 1_n a^T + bY \quad \text{and} \quad \text{Vec}((a, b) \ast Y) = 1_n \otimes a + b \text{Vec}(Y).
\]

then \( \text{Vec}((-\mu, 1) \ast Y) \sim N(0, (I_n + \theta B) \otimes \sigma^2 I_d) \) and

\[
\text{Vec}((-\mu/\sigma, 1/\sigma) \ast Y) \sim N(0, (I_n + \theta B) \otimes I_d),
\]

which correspond to elements in \( \text{Orb}(Y) \).

In essence, the observation is not regarded as a point in \( \mathbb{R}^{n \times d} \) but is treated as a group orbit generated by the group of rigid transformations, or similarity transformations if scalar multiples are permitted. In statistical terms, this approach meshes with the submodel in which the matrix \( \Sigma \) in model I is a scaled identity matrix \( I_d \). An equivalent way of saying the same thing for \( n > d \) is that the column-centered sample matrix

\[
\tilde{Y} = Y - 1_n 1_d^T Y/n \quad \text{determines the sample covariance matrix} \quad S = (\tilde{Y}^T \tilde{Y})/(n - 1)
\]

and hence the Mahalanobis metric \( \|x - x^*\|^2 = (x - x^*)^T S^{-1} (x - x^*) \) in the state space (Mahalanobis 1936; Gnanadesikan and Kettenring 1972). One implication is that the \( n \times n \) matrix


\[ D = (D_{ij}) = \left( Y_i - Y_j \right)^2 \]

of standardized inter-point Mahalanobis distances is maximal invariant, and the conditional distribution on sample partitions depends on \( Y \) only through this matrix.

In practice, the \( d \) variables are sometimes measured on scales that are not commensurate with one another, so the state space seldom has a natural metric. In this case, we assume that \( Y \) and \( Y' \) as equivalent configurations for each feature \( Y, j \) if there are \( a_i \in \mathbb{R} \) and \( b_j \in \mathbb{R} \setminus \{0\} \) such that \( Y'_j = a_j + b_j Y_j \). In model II, the group is the affine group \( G A(\mathbb{R})^d, \mathcal{G} = \mathbb{R}^d \times D \) and \( D = \{ \text{diag}(b_1, \ldots, b_d) | b_i \neq 0, i = 1, \ldots, d \} \) with the operation \((a_1, A_1) \circ (a_2, A_2) = (a_1 + A_1 a_2, A_1A_2)\) for \( a_i \in \mathbb{R}^d, A_i \in D \) with \( i = 1, 2 \). The orbit of an element \( Y \in \mathbb{R}^{n \times d} \) and the group action are defined in (1) and (2) with \( A \in D \). If

\[
\text{Vec}(Y) \sim N \left( \mathbf{1}_n \otimes \mu, (I_n + \theta B) \otimes \text{diag} \{ \sigma_1^2, \ldots, \sigma_d^2 \} \right)
\]

then \( \text{Vec}((-\mu, I_d) \ast Y) \sim N(0, (I_n + \theta B) \otimes \text{diag} \{ \sigma_1^2, \ldots, \sigma_d^2 \}) \), and furthermore \( \text{Vec} ((a, A) \ast Y) \sim N(0, (I_n + \theta B) \otimes I_d) \) with \( a = - (\mu_1/\sigma_1, \ldots, \mu_d/\sigma_d)^T \) and \( A = \text{diag} \{ a_1^{-1}, \ldots, a_d^{-1} \} \), which correspond to elements of the group orbit. No linear combinations are permitted here, so that the integrity of the variables is preserved.

Moreover, in some cases, the location information or shapes of objects from aerial photography applications may be distorted by the viewer’s angle or position so that the variables may be strongly correlated. A more extreme approach avoids the metric assumption by regarding \( Y \) and \( Y' \) as equivalent configurations if there exists a vector \( a \in \mathbb{R}^d \) and a non-singular matrix \( A \in \mathbb{R}^{d \times d} \) such that \( Y'_i = a + A Y_i \) with \( A^T A \) is a positive definite matrix for all \( i \). Consequently, models I, II, III specify the structures of the covariance matrix between features, and the partition \( B \) of \( Y \) is affine invariant and the same as the partition \( B \) of the group orbit \( \mathcal{G} Y \subset \mathbb{R}^{n \times d} \), which is independent of the mean.

### 3.1 Gaussian marginal probabilities

The distribution of the column-centered group orbit, \( \mathcal{G} Y \), is assumed to be a Gaussian distribution

\[
N \left( 0, (I_n + \theta B) \otimes A^T A \right)
\]

which depends only on \( I_n + \theta B \) and \( A^T A \). Actually, it can be verified that for any \((a, A) \in \mathcal{G}, \) the two distributions of group orbits induced by \( N (\mathbf{1}_n \otimes \mu, (I_n + \theta B) \otimes \Sigma) \) and \( N (\mathbf{1}_n \otimes (a + A \mu), (I_n + \theta B) \otimes (A \Sigma A^T)) \) respectively are the same.

McCullagh (2008) studied the \( d \) time series with an autocorrelation \( \Gamma \) and \( n \) observations in time or space following three Gaussian distribution models \( N (0, \Gamma \otimes \Sigma) \) under different assumptions of \( \Sigma \) as follows:

\[
\begin{align*}
\text{Model I:} \Sigma &= \sigma^2 I_d \\
\text{Model II:} \Sigma &= \text{diag} \{ \sigma_1^2, \ldots, \sigma_d^2 \} \\
\text{Model III:} \Sigma &\in PD_d
\end{align*}
\]

where \( PD_d \) is the collection of \( d \times d \) symmetric positive definite matrices. These three models correspond to our three models of affine transformed equivalence classes which
we discussed in the previous section. In this paper, we set \((I_n + \theta B)\) as \(\Gamma\) and \(A^T A\) as \(\Sigma\). Following (McCullagh 2008), the log-likelihood based on \(Y\) for all three models is:

\[
I(\Gamma, \Sigma|Y) = -\frac{1}{2} \log \det(\Gamma \otimes \Sigma) - \frac{1}{2} \text{tr} (Y^T \Gamma^{-1} Y \Sigma^{-1})
\]

\[
= -\frac{1}{2} \log \det(\Gamma) - \frac{n}{2} \log \det(\Sigma) - \frac{1}{2} \text{tr} (Y^T \Gamma^{-1} Y \Sigma^{-1}),
\]

**Lemma 1** \((I_n + \theta B)^{-1} = I_n - \theta WB\), where \(W = \text{diag} \{(1 + \theta N_1)^{-1}, \ldots, (1 + \theta N_n)^{-1}\}\) and \(N_i\) is the \(i\)th diagonal element of \(N = B1_n\).

According to Lemma 1 and its proof, which is relegated to the Appendix A, \(\Gamma = I_n + \theta B\) is always nonsingular for \(\theta > 0\) and its inverse \(\Gamma^{-1} = (I_n + \theta B)^{-1} = I_n - \theta WB\) can be obtained explicitly. To ensure that \(Y^T \Gamma^{-1} Y\) is positive definite with probability 1 (McCullagh 2008), as well as informative group orbits (see Theorem 1 and its subsequent discussion), we assume \(n > d + 1\).

After plugging in the maximum likelihood estimator of \(\Sigma\) which for model III is \(\hat{\Sigma}_I = Y^T \Gamma^{-1} Y / n\), for model II is \(\text{diag} \left(\hat{\Sigma}_I\right)\), and for model I is \(\text{tr} \left(\hat{\Sigma}_I\right) I_d / d\) (McCullagh 2008), the profile likelihood of \(\Gamma\) is

\[
L_p (\Gamma^{-1}|Y) = \begin{cases} 
\det (\Gamma^{-1})^{d/2} / \text{tr} (Y^T \Gamma^{-1} Y)^{nd/2} & (I) \\
\det (\Gamma^{-1})^{d/2} / \prod_{r=1}^d \left(\sum_{(r)} \Gamma^{-1} Y_{(r)}\right)^{n/2} & (II) \\
\det (\Gamma^{-1})^{d/2} / \det (Y^T \Gamma^{-1} Y)^{n/2} & (III)
\end{cases}
\]

where \(Y_{(r)} \in \mathbb{R}^n\) is the \(r\)th column of \(Y\), \(r = 1, \ldots, d\).

The conditional distribution on partitions of \([n]\) depends on the group orbit and the assumptions made regarding \(\Sigma\). For group I, with \(\Sigma \propto I_d\) in the Gaussian model, the likelihood depends only on the distance matrix \(D\), so the likelihood is constant on the orbits associated with the larger group of Euclidean similarities. Therefore, for model I, the similarity transformation can be generalized as if \(Y'_I = a + A Y_i\) for \(A^T A = \sigma^2 I_d\) and \(\sigma \neq 0\), implying that the arrays of distances are proportional \(D' = \sigma^2 D\). Consequently, there is a representative element of the group orbit with feature mean vector \(0\), so that \(\text{Vec}(Y) \sim N(0, (I_n + \theta B) \otimes \sigma^2 I_d)\).

For model II, the affine transformation can be generalized as \(Y'_I = a + A Y_i\), where \(a \in \mathbb{R}^d\) and \(A \in \mathbb{R}^{d \times d}\) with \(A^T A\) as a diagonal matrix with positive diagonal entries for all \(i\). As a result, there is a representative element of the group orbit with feature mean vector \(0\), so that \(\text{Vec}(Y) \sim N(0, (I_n + \theta B) \otimes \text{diag} \{\sigma_1^2, \ldots, \sigma_d^2\})\). This is to work with \(GA(\mathbb{R})^d\) which is the general affine group acting independently on the \(d\) columns of \(Y\). For model III, \(\Sigma\) is an arbitrary matrix in \(PD_d\). The group is \(GA(\mathbb{R}^d)\) and \(n > d + 1\). These three models are nested by model I \(\subset\) model II \(\subset\) model III.

Affine invariance in \(\mathbb{R}^d\) is a strong requirement, which comes at a small cost for moderate \(d\) provided that \(d/n\) is small. When \(d/n \leq 1\), \(Y^T \Gamma^{-1} Y\) is positive definite with probability one (McCullagh 2008), then model III works. However, while \(d/n < 1\) is not small, model III may be inefficient due to some eigenvalues of \(Y^T \Gamma^{-1} Y\) and \(\det (Y^T \Gamma^{-1} Y)\) close to zero (Dempster 1972; Stein 1975). As a result, the profile likelihood of \(\Gamma\) becomes unstable. In contrast, model II is less computationally expensive than model III, and model I is the most efficient one.


4 Markov chain Monte Carlo algorithm for sampling partitions

We use the prior and posterior distributions of $\theta$ and $B$ discussed in Section 2 through a Markov chain Monte Carlo (MCMC) algorithm for sampling partitions. The iterative $\theta$ is obtained by Gibbs sampling (Geman and Geman 1984) according to the conditional distribution $p_n(\theta_j | B, \mathcal{G} Y) \propto p(\theta_j) \times L_p (\Gamma^{-1} | \mathcal{G} Y)$, where $p(\theta_j) \propto \theta_j^{a-1} / (1 + \theta_j)^{2a}$ for $j = 1, \ldots, J$. For instance, $\alpha = 1$ and the discrete set $\{2^{-3}, 2^{-2}, \ldots, 2^{10}\}$ for the range of $\theta$ are used as the default setting in our experiments. For updating $B$, the conditional distribution on partitions is

$$p_n(B | \theta, \mathcal{G} Y) \propto p_n(B | \lambda) \times L_p (\Gamma^{-1} | \mathcal{G} Y),$$

where $p_n(B | \lambda)$ is the Ewens distribution, and a Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970) is used to choose the iterative $B$. $\lambda$ is set as 1 in the following applications. After burning in a certain number of the resulting Markov chain, we use the average of the partition matrix as the similarity matrix to make inference on partition.

The proposal distribution $q(B^{(i+1)} | B^{(i)}, \mathcal{G} Y)$ is proportional to $\exp(-a \times d_{xc})$, where $d_{xc}$ is the distance between each point and the corresponding centroid of the clusters and $a$ is a scale hyperparameter which was set as 2 in our experiments. More specifically, a partition candidate $B^*$ is generated by re-assigning the label of each point with the probability proportional to the reciprocal of the distance between each point and the corresponding centroid.

**Algorithm 1** Metropolis-Hastings algorithm

1: Initialize $B$ and $\theta$.
2: for $i = 1 : N$ do N is the number of total iterations. Suppose that the current values are $\theta^{(i)}$ and $B^{(i)}$.
3: Randomly sample $\theta^{(i+1)}$ from $p_n(\theta_j | B^{(i)}, \mathcal{G} Y), j = 1, \ldots, J$.
4: Propose $B^* \sim q(B^{(i+1)} | B^{(i)}, \mathcal{G} Y)$.
5: Calculate

$$R = \frac{p_n(B^* | \lambda) \times L_p (\Gamma(\theta^{(i+1)}, B^*)^{-1} | \mathcal{G} Y)}{p_n(B^{(i)} | \lambda) \times L_p (\Gamma(\theta^{(i+1)}, B^{(i)})^{-1} | \mathcal{G} Y) \times q(B^* | B^{(i)}, \mathcal{G} Y).}$$

6: Accept $B^{(i+1)} = B^*$ with probability $\min(1, R)$
7: Keep $B^{(i+1)} = B^{(i)}$ with probability $1 - \min(1, R)$
8: end for
9: return all the $B^{(i)}$’s and $\theta^{(i)}$’s.

Since Algorithm 1 is a Metropolis-Hastings algorithm, it satisfies the detailed balance condition, and therefore the generated Markov chain has a stationary distribution (Chib and Greenberg 1995; Gamerman 1997; Robert and Casella 2010). Since we leave a small but positive probability that the partition stays the same in the Gibbs sampling and the discrete posterior of $\theta$ stays positive always, then the transition probability

$$p_n(\theta^{(k+1)}, B^{(k+1)} | \theta^{(k)}, B^{(k)}) > 0$$

where $\theta^{(k+1)} = \theta^{(k)}$ and $B^{(k+1)} = B^{(k)}$, and then the $(\theta, B)$-valued Markov chain constructed by Algorithm 1 is aperiodic.
Lemma 2 If \( n > d + 1 \), the \((\theta, B)\)-valued Markov chain constructed by Algorithm 1 is aperiodic.

Since there is always a positive chance that the partition can be split further into the simplest partition in which each element is a cluster, then all possible partitions communicate with each other, so that the \((\theta, B)\)-valued Markov chain constructed by Algorithm 1 is irreducible. Given the sample size \( n \), the size of the state space of \( B \) known as the Bell number (Bell 1934), and the size of the state space of \( \theta \) are all finite, then the irreducibility also implies positive recurrence. Consequently, the \((\theta, B)\)-valued Markov chain constructed by Algorithm 1 is ergodic (Isaacson and Madsen 1976; Gilks et al. 1996). The properties are summarized as the following lemma and theorem, whose proofs are relegated to the Appendix A.

Lemma 3 If \( n > d + 1 \), the \((\theta, B)\)-valued Markov chain constructed by Algorithm 1 is irreducible, and thus is positive recurrent.

Theorem 2 (Ergodic theorem) If \( n > d + 1 \), the \((\theta, B)\)-valued Markov chain constructed by Algorithm 1 converges to its stationary distribution \( p_n(\theta, B|GY) \propto p(\theta) \times p_n(B|\lambda) \times L_p(\Gamma^{-1}|GY) \). More specifically, for any real-valued function \( f \) satisfying \( \sum_{(\theta, B)} f(\theta, B)p_n(\theta, B|GY) < \infty \), we have

\[
\frac{1}{n+1} \sum_{i=0}^{n} f(\theta^{(i)}, B^{(i)}) \longrightarrow \sum_{(\theta, B)} f(\theta, B)p_n(\theta, B|GY)
\]

almost surely for all initial value \((\theta^{(0)}, B^{(0)})\).

5 Analysis of simulated and real data
We test the proposed Bayesian cluster process with Algorithm 1 on both synthetic and real data. Algorithm 1 with model I and point-wise updating is equivalent to the method of (Vogt et al. 2010). If there is no prior information of the number of clusters, users can set the initial partition \( B \) as \( I_n \) in which each observation is a block. In practice, we use a randomly sampled clusters from a discrete uniform distribution of a range chosen by users. The clustering result is represented by the average of the estimated similarity matrix

\[
S = \sum_{k=n_0+1}^{N} \frac{B^{(k)}}{N-n_0},
\]

where \( n_0 \) is the number of burn-in iterations. Furthermore, we also define a dissimilarity matrix \( D \) as \( 1_n 1_n^T - S \). The dissimilarity matrix, \( D \), can be expressed by a heatmap which represents a matrix with grayscale colors with white as 1, black as 0, and the spectrum of gray as values between 0 and 1. The heatmap of the original similarity matrix cannot be recognized with the naked eye and equivalence relation needs to be decoded from the matrix \( B \). However, in practice, users can identify clusters through including the names of rows and columns of the similarity matrix to find which individuals are clustered together. Additionally, the heatmap function of the stats R package can permute the order of individuals to have cluster blocks with hierarchical dendrograms. It is challenging to monitor convergence of the Markov chain because the sampled clusters are random and may vary
in each iteration. To determine convergence, we run Algorithm 1 ten times for each data set and stop the chain when we observe the number of clusters remain the same in the given chain length (Chang and Fisher 2013).

5.1 Illustrative simulated data

**Four clusters on the vertices of a unit square data** Three simulated data sets are generated for illustration. In the simulation study, 1000 initial burn-in iterations were discarded, and 2000 Markov chains of $B$ samples based on each model were used to calculate $D$. We first applied the proposed cluster process with model I on the synthetic data for four clusters centered at the four vertices of a unit square. For each vertex $\mu_k$, we generate 20 points from $N(\mu_k, (1/4)I_2)$ for $k = 1, \ldots, 4$ (see Fig. 1, the left panel). We call the data $X_I$, and then apply model I to cluster $X_I$ with the average within- and between-cluster distances. The resulting heatmap successfully reveals the true clusters for most of the points (not shown here).

Then we transform the data by $X_{II} = X_I \times \begin{pmatrix} 3 & 0 \\ 0 & 1/3 \end{pmatrix}$. The transformed features seem to have two groups (see Fig. 1, the middle panel), clusters (1, 2) and clusters (3, 4). The cluster process with model I does not work well for this case, while the heatmap based on model II without knowing the transformation can reveal the true clusters for most of the points (not shown here).

Furthermore we transform the data by $X_{III} = X_I \times \begin{pmatrix} 4.1 & 2.1 \\ 1.9 & 1.1 \end{pmatrix}$. The transformed features are aligned in a straight line (see Fig. 1, the right panel). The transformed data $X_{III}$ is more difficult to cluster than $X_I$ and $X_{II}$, since the original four clusters are transformed to be not well separated.

The resulting heatmap using model III with the initial clusters assigned randomly and uniformly from $\{1, 2, 3, 4\}$ reveals the true four clusters for most of the points (see Fig. 2).

5.2 Applications to real data

Besides the synthetic data, we also evaluate the performance of the proposed approach by using real data. We run 3000 MCMC iterations and burn in the first 1000 iterations,
and use the heatmap of matrix $S$ to visualize the clusters. The accuracy rate is based on the average proportion of identical elements of matrix $B$ of the cluster and the true matrix $B$, and compared the accuracy rates with $k$-means (MacQueen 1967) and Mclust using R package ‘mclust’ with its default setting (Fraley and Raftery 2002). The reason why we chose R package ‘mclust’ is that Mclust is a model-based clustering approach using the Gaussian mixture model, which assumes a Gaussian distribution for each component under one of the three types covariance structures (the argument of Mclust: modelNames) 1. Spherical (EII), 2. Diagonal (VVI), and 3. General (VVV) for comparing with our proposed model I, II, III, correspondingly. The main difference is that the Mclust obtains clusters with an expectation–maximization (EM) algorithm (Dempster 1972; McLachlan and Peel 2000), but our method uses a Metropolis-Hasting algorithm with the profile likelihood of $\Gamma$ to sample clusters.

**Model I: Gene expression data of Leukemia patients** The gene expression microarray data (Dua and Graff 2019) has been used to study genetic disorder such as identifying diagnostic or prognostic biomarkers or clustering and classifying diseases (Dudoit et al. 2002). For example, (Golub et al. 1999) classified patients of acute leukemia into two sub types, Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML). For illustration purpose, we use the training set of the leukemia data which consists of 3051 genes and 38 tumor mRNA samples. Pretending we do not know the label information, we would like to cluster the 38 samples according to their 3051 features (gene expression levels). The two clusters comprise 27 ALL cases and 11 AML cases. Since the number of features is larger than the sample size, our approach is not applicable to this dataset directly. Therefore, we first reduce the dimension by projecting the data on the subspace which consists of the first twenty principal components (PC) (Jolliffe 1986). Note that these PC scores are orthonormal which satisfies the assumption of model I. The resulting heatmap based on model I (Fig. 3) reveal the cluster of the 11 AML cases. The accuracy rate using the proposed model I with the initial clusters assigned randomly and uniformly from $\{1,2\}$ is 0.9164, while the accuracy rates of $k$-means and Mclust are 0.6994 and 0.5886, respectively. We noticed that Mclust resulted in only one cluster.
Model II: Geographic coordinate data of Denmark’s 3D Road Network

This three-dimensional road network dataset of geographic coordinates includes the altitude, latitude, and longitude degrees of each road segments in North Jutland in northern Denmark, which is publicly available at the UC Irvine Machine Learning Repository (Kaul 2013; Dua and Graff 2019). Since three spatial dimensional features are orthogonal, it satisfies the assumption of model II so that we use this dataset to demonstrate model II. Three subjects with the road maps OSM ID 144552912 (19 observations), 125829151 (13 observations), 145752974 (14 observations) are used for the clustering analysis. Note that each objects may have several observations measured from different angles, and the altitude values are extracted from NASA’s Shuttle Radar Topography Mission (SRTM) data (Jarvis et al. 2008). The accuracy rate using model II with the initial clusters assigned randomly and uniformly from \{1, 2, 3, 4, 5\} is 1, while the accuracy rates of \(k\)-means with \(k = 3\) and Mclust are 0.7486 and 0.9490, respectively. The resulting heatmap using model II (Fig. 4) reveals 3 clusters correctly.
**Model III: Iris data**

This iris dataset (Fisher 1936) contains three species–Setosa, Versicolor, and Virginica with four features which are the measurements of the variables sepal length and width and petal length and width in centimeters, respectively. Each species consists of 50 iris flowers. The data points are clustered by their four features. Here, \( d = 4, n = 150, k = 3 \). The heatmap of the similarity matrix using model III correctly reflects three clusters corresponding to the three species of iris for most points (Fig. 5). The accuracy rate using the proposed model III with the initial clusters assigned randomly and uniformly from \{1, 2, 3\} is 0.9087, while the accuracy rates of \( k \)-means with \( k = 3 \) and Mclust are 0.7740 and 0.7763, respectively. We noticed that both the \( k \)-means and Mclust result in two clusters by grouping Versicolor and Virginica as a cluster.

**6 Concluding discussion**

The proposed clustering method is invariant under different groups of affine transformations and computationally efficient. It identifies clusters for most samples without knowing the number of clusters in advance, and it may group a big cluster as several small clusters. These problems are dealt with an exchangeable partition prior which avoids label-switching problems and the partition valued in the MCMC algorithm is invariant under linear transformations under three types of covariance structures. The advantage of replacing the Dirichlet-multinomial prior with its limiting process is that we do not need to know the number of clusters in advance. The disadvantage is that it may be less efficient computationally if the number of clusters is known. Note that the proposed approach does not target the partition maximizing the posterior distribution. Instead, it estimates the expected partition or the similarity matrix.

The three clustering models are based on the covariance matrix between variables. There are guidelines of telling which model work best in practice by the experimental design or testing its sample covariance matrix. If the features are orthogonal or orthonormal, then model I and model II are applicable, respectively. Models I and II run faster than model III due to the structure of the covariance matrix. Otherwise, model III can be used in general. It works reasonably well across various applications.

![Fig. 5](image.png) The heatmap of the similarity matrix using model III correctly reveals three clusters corresponding to the three species of iris for most points.
Since we use the profile likelihood of $\Gamma$ in our model, we do not sample the covariance matrix directly, and Lemma 1 and Theorem 1 implies as $n > d + 1$, the proposed Metropolis-Hastings algorithm can work. However, the maximum likelihood estimator (MLE) of the general unstructured covariance matrix will be less efficient if the diagonal covariance structure is actually correct because it will tend to have small eigenvalues and a large determinant of the inverse covariance matrix (i.e. $\Gamma^{-1}$). Indeed, when using model III and even if $n > d + 1$, $\Gamma$ may be near singular. This may make the sampling less efficient, i.e. the acceptance rate may become small (Roberts and Rosenthal 2001). Although the stationary distribution of the sampled clusters’ Markov chain using Algorithm 1 is independent of the initial clusters according to Theorem (2), we practically suggest to set the initial clusters sampled from a discrete uniform distribution of a range given by users instead of setting each individual as a cluster in order to obtain convergent sampled clusters without using a long Markov chain. This makes the proposed Algorithm 1 sample more efficiently from a smaller collection of partition candidates.

The proposed clustering algorithm produces the desired clusters with 2000 iterations after 1000 burn-in iterations in our experiments. The main contributions of our work include: 1) The proposed three clustering models with three types of covariance structures can handle general cases of affine transformations. In contrast, (Vogt et al. 2010) only dealt with the case of model I. 2) Algorithm 1 is efficient, since it updates all individual’s clusters instead of a single individual’s cluster per iteration. It also ensures that the resulting partition-valued Markov chain is ergodic and convergent in distribution. 3) The experiments show the advantages of our cluster process which successfully identifies the true clusters using the proposed distance matrix. In particular if the clusters are not well separated, the similarity matrix with probabilistic nature can still reveal the relationships through hierarchical approaches. The proposed method could be used to extract interesting information from aerial photography, genomic data, and data with attributes under different scales, especially when the nearest neighbors may belong to different clusters in the feature space. The proposed method can be improved in the further work by modeling the mean of each cluster with regression on covariates or non-Gaussian distributions.

Appendix A

Proof of Theorem 1: For any $Y \in \mathbb{R}^{n \times d}$, denote $\tilde{Y} = Y - 1_n 1_n^T Y / n$. Let $\tilde{Y}_{(n-1)}$ be the $(n - 1) \times d$ matrix consisting of the first $n - 1$ rows of $\tilde{Y}$. Since $1_n^T \tilde{Y} = 0$, then rank($\tilde{Y}$) = rank($\tilde{Y}_{(n-1)}$).

If $n \leq d + 1$ and rank($\tilde{Y}_{(n-1)}$) = $n - 1$, that is, $\tilde{Y}_{(n-1)}$ is of full row rank, then there exists an orthogonal matrix $O \in \mathbb{R}^{d \times d}$ (column permutations), such that, $\tilde{Y}_{(n-1)} O = (U, V)$, where $U \in \mathbb{R}^{(n-1) \times (d+1-n)}$ is of full rank, and $V \in \mathbb{R}^{(n-1) \times (d+1-n)}$. We let

$$a = \frac{1}{n} Y^T 1_n, \quad A = O \begin{pmatrix} U^T & 0 \\ V^T & I_{d+1-n} \end{pmatrix}, \quad Z = \begin{pmatrix} I_{n-1} & 0 \\ -1_{n-1}^T & 0 \end{pmatrix}.$$  

It can be verified that $Y = (a, A) \ast Z$. That is, $Y \in \text{Orb}(Z)$, where $Z$ is a constant matrix.

If $n \leq d$ and rank($Y$) = $n$, then rank($\tilde{Y}_{(n-1)}$) = $n - 1$, since $\tilde{Y}_{(n-1)} = W Y$, where $W = (I_{n-1} - 1_{n-1} 1_{n-1}^T / n - 1_{n-1} / n)$ is of full row rank $n - 1$.

Suppose $n = d + 1$ and rank($Y$) = $d = n - 1$. Without any loss of generality, we assume rank($Y_{(n-1)}$) = $n - 1$, where $Y_{(n-1)}$ consists of the first $n - 1$ rows of $Y$. Then $Y_n = c_1 Y_1 + \cdots + c_{n-1} Y_{n-1}$ for some $c_1, \ldots, c_{n-1} \in \mathbb{R}$ and $\tilde{Y}_{(n-1)} = D Y_{(n-1)}$, where
$D = I_{n-1} - 1_{n-1}1_{n-1}^T / n - 1_{n-1}e^T / n$, and $e^T = (c_1, \ldots, c_{n-1})$. It can be verified that if $c_1 + \cdots + c_{n-1} \neq 1$, then rank($D$) = $n - 1$ and rank($\tilde{Y}_{(n-1)}$) = $n - 1$; if $c_1 + \cdots + c_{n-1} = 1$, then rank($D$) = $n - 2$ and rank($\tilde{Y}_{(n-1)}$) = $n - 2$. Note that if $n = d + 1$ but rank($\tilde{Y}_{(n-1)}$) = $n - 2$, then $W \notin \text{Orb}(Z)$. □

**Proof of Lemma 1:** Suppose the partition matrix $B$ consists of $k$ blocks with block sizes $n_1, \ldots, n_k$, where $k \geq 1$, $n_i > 0$ for $i = 1, \ldots, k$, and $n_1 + \cdots + n_k = n$.

We first assume that $B = \text{diag} \{1_{n_1}, 1_{n_2}^T, \ldots, 1_{n_k}^T\}$, which is in its standard form. Then $B = LL^T$ with $L = \text{diag}(1_{n_1}, \ldots, 1_{n_k}) \in \mathbb{R}^{n \times k}$ and $L_n + \theta B = I_n + EET$ with $E = \sqrt{\theta} L$.

According to the Sherman-Morrison-Woodbury formula (see, for example, Section 2.1.4 in Golub and Van Loan (2013)), for matrices $A \in \mathbb{R}^{n \times n}$ and $U, V \in \mathbb{R}^{n \times k}$, $(A + UV^T)^{-1} = A^{-1} - A^{-1}U(V^T A^{-1}U)^{-1}V^TA^{-1}$ if both $A$ and $I + V^T A^{-1}U$ are nonsingular. In our case, $A = I_n$ is nonsingular, $U = V = E$, and $I + V^T A^{-1}U = I_k + ETE = \text{diag}[1 + \theta n_1, \ldots, 1 + \theta n_k]$ is also nonsingular. Thus

\[
(I_n + EET)^{-1} = I_n - E(I_k + ETE)^{-1}E^T = I_n - \theta L(I_k + \theta LTY)^{-1} L^T
\]

\[
= I_n - \theta \cdot \text{diag} \left\{ \frac{1}{1 + \theta n_1} 1_{n_1} 1_{n_1}^T, \ldots, \frac{1}{1 + \theta n_k} 1_{n_k} 1_{n_k}^T \right\}
\]

\[
= I_n - \theta \cdot \text{diag} \left\{ \frac{1}{1 + \theta n_1} I_{n_1}, \ldots, \frac{1}{1 + \theta n_k} I_{n_k} \right\} \cdot \text{diag} \left\{ 1_{n_1} 1_{n_1}^T, \ldots, 1_{n_k} 1_{n_k}^T \right\}
\]

\[
= I_n - \theta WB.
\]

In general, by row-switching and column-switching transformations, we can always transform $B$ into its standard form. That is, there exists an orthogonal matrix $O$ such that $B_* = OBOT$ is in standard form. Let $W_r = OWOT$. Then $(I_n + \theta B)^{-1} = OT(I_n + \theta B)^{-1}O = OT(I_n - \theta W_r B_r)O = I_n - \theta \cdot OT W_r O \cdot OT B_r O = I_n - \theta WB$. □

**Proof of Lemma 3:** In our case, the Markov chain built by Algorithm 1 is actually a discrete chain. It is irreducible since $p_n(\theta^{(k+1)}, B^{(k+1)}|\theta^{(k)}, B^{(k)}) > 0$ for each pair of states. As a direct conclusion of Theorem 4.1 in Gilks et al. (1996), our Markov chain is positive recurrent. □

**Proof of Theorem 2:** Algorithm 1 is a Gibbs sampler plus a Metropolis-Hastings component for sampling $B^{(i+1)}$. Given $B^{(i)}$ and $\theta^{(i+1)}$, the Metropolis-Hastings ratio with proposal distribution $q(B|B^{(i)}, GY)$ and target distribution $p_n(\theta, B|GY)$ is

\[
R(B^{(i)}, B^*) = \frac{p_n(\theta^{(i+1)}, B^*|GY) \cdot q(B^{(i)}|B^*, GY)}{p_n(\theta^{(i+1)}, B^{(i)}|GY) \cdot q(B^{(i)}|B^{(i)}, GY)}
\]

\[
= \frac{p_\theta(\theta^{(i+1)}) \cdot p_n(B^*|\lambda) \cdot L_p(\Gamma(\theta^{(i+1)}, B^{*})^{-1}|GY) \cdot q(B^{(i)}|B^*, GY)}{p_\theta(\theta^{(i+1)}) \cdot p_n(B^{(i)}|\lambda) \cdot L_p(\Gamma(\theta^{(i+1)}, B^{(i)})^{-1}|GY) \cdot q(B^{(i)}|B^{(i)}, GY)}
\]

\[
= \frac{p_n(B^*|\lambda) \cdot L_p(\Gamma(\theta^{(i+1)}, B^{*})^{-1}|GY) \cdot q(B^{(i)}|B^*, GY)}{p_n(B^{(i)}|\lambda) \cdot L_p(\Gamma(\theta^{(i+1)}, B^{(i)})^{-1}|GY) \cdot q(B^{(i)}|B^{(i)}, GY)}
\]

which is exactly $R$ in Algorithm 1. Since Metropolis-Hastings algorithms satisfy detailed balance condition, the target distribution $p_n(\theta, B|GY)$ is a stationary distribution. By Lemmas 2 and 3, the convergence statements follow as a direct conclusion of Theorems 4.3 and 4.4 in Gilks et al. (1996). □
### Appendix B journal name abbreviations for use in Boundary-Layer meteorology

| Journal Name                                                                 | Abbreviation used in BLM                        |
|------------------------------------------------------------------------------|------------------------------------------------|
| ACM Transactions of Mathematical Software                                   | ACM Trans Math Soft                             |
| Acoustics Australia                                                         | Acoust Aust                                     |
| Acta Geophysica                                                              | Acta Geophys                                    |
| Acta Mechanica Synica                                                       | Acta Mech Sinica                                |
| Acta Mechanica Supplement                                                   | Acta Mech Suppl                                  |
| Advances in Atmospheric Science                                             | Adv Atmos Sci                                   |
| Advances in Ecological Research                                             | Adv Ecol Res                                    |
| Advances in Meteorology                                                     | Adv Meteorol                                    |
| Advances in Science and Research                                            | Adv Sci Res                                     |
| Advances in Water Resources                                                 | Adv Water Resour                                |
| Aeolian Research                                                             | Aeolian Res                                     |
| Aerospace Science and Technology                                            | Aerosp Sci Technol                              |
| Agricultural Meteorology                                                    | Agric Meteorol                                  |
| Agricultural and Forest Meteorology                                         | Agric For Meteorol                              |
| Agricultural Water Management                                               | Agric Water Manag                               |
| American Institute of Aeronautics and Astronautics                          | Am Inst Aeronaut                                |
| Astronautics                                                                 | Astronaut                                       |
| Annals of Glaciology                                                        | Ann Glaciol                                     |
| Annalen der Meteorologie                                                    | Ann Meteorol                                    |
| Annals of Statistics                                                        | Ann Stat                                        |
| Antarctic Science                                                           | Antarct Sci                                     |
| Annual Review of Fluid Mechanics                                            | Annu Rev Fluid Mech                             |
| Applied Energy                                                               | Appl Energy                                     |
| Applied Mechanics Review                                                    | Appl Mech Rev                                   |
| Applied Numerical Mathematics                                                | Appl Numer Math                                 |
| Applied Physics B                                                           | Appl Phys B                                     |
| Applied Optics                                                               | Appl Opt                                        |
| Aquatic Botany                                                              | Aquat Bot                                       |
| Archiv fur Meteorologie Geophysik und Bioklimatologie Serie A-Meteorologie und Geophysik | Arch Meteorol Geophys Bioklim Ser A |
| Archiv fur Hydrobiologie                                                    | Arch Hydrobiol                                  |
| Artificial Intelligence                                                     | Artif Intell                                    |
| Astronomy & Astrophysics                                                     | Astron Astrophys                                |
| Atmospheric Measurement Techniques                                          | Atmos Meas Tech                                 |
| Atmosphere-Ocean                                                            | Atmos-Ocean                                     |
| Journal Name                                                      | Abbreviation used in BLM          |
|------------------------------------------------------------------|-----------------------------------|
| Atmospheric Research                                             | Atmos Res                         |
| Atmospheric Science Letters                                      | Atmos Sci Lett                     |
| Australian Journal of Physics                                    | Aust J Phys                        |
| Australian Journal of Botany                                     | Aust J Bot                         |
| Beitraege zur Physik der Atmosphaere                            | Beitr Phys Atmos                   |
| Biogeosciences                                                   | Biogeosciences                     |
| Biometrika                                                       | Biometrika                         |
| Biosystems Engineering                                           | Biosyst Eng                        |
| Boreal Environment Research                                     | Boreal Environ Res                 |
| Boundary-Layer Meteorology                                       | Boundary-Layer Meteorol            |
| Building and Environment                                         | Build Environ                      |
| Bulletin of the American Meteorological Society                  | Bull Am Meteorol Soc               |
| Climate Research                                                 | Clim Res                           |
| Cold Regions Science and Technology                              | Cold Reg Sci Technol               |
| Communications in Agricultural and Applied Biological Sciences   | Commun Agric Appl                  |
| Communications in Mathematical Physics                           | Commun Math Phys                   |
| Communications on Pure and Applied Mathematics                   | Commun Pure Appl                   |
| Comptes Rendus Physique                                          | C R Phys                           |
| Computers and Electronics in Agriculture                         | Comput Electron Agric              |
| Computing and Informatics                                        | Comput Inf                         |
| Computer Methods in Applied Mechanical Engineering               | Comput Methods Appl Mech Eng       |
| Computational Statistics and Data Analysis                       | Comput Stat Data Anal              |
| Contributions to Atmospheric Physics                             | Contr Atmos Phys                   |
| Crop Protection                                                  | Crop Prot                          |
| Deep Sea Research Part II                                        | Deep Sea Res II                    |
| Dynamics of Atmospheres and Oceans                               | Dyn Atmos Oceans                   |
| Earth System Science Data Discussions                            | Earth Syst Sci Data Discuss        |
| Earth Surface Processes and Landforms                            | Earth Surf Process Landf           |
| Ecological Applications                                          | Ecol Appl                          |
| Ecological Indicators                                            | Ecol Indic                         |
| Ecological Modelling                                             | Ecol Model                         |
| Ecology                                                          | Ecology                            |
| Journal Name                                                      | Abbreviation used in BLM |
|------------------------------------------------------------------|--------------------------|
| Electronic Journal of Operational Meteorology                    | Electron J Oper Meteorol |
| Energies                                                         | Energies                 |
| Energy and Buildings                                             | Energy Buil              |
| Energy Conversion and Management                                 | Energy Convers Manag     |
| Environmental Fluid Mechanics                                    | Environ Fluid Mech       |
| Environmental Modeling and Software                              | Environ Modell Softw     |
| Environmental Pollution                                         | Environ Pollut           |
| Environmental Research Letters                                   | Environ Res Lett         |
| Environmental Science and Technology                             | Environ Sci Technol      |
| Environmental Software                                          | Environ Softw            |
| Eos, Transactions, American Geophysical Union                    | Eos Trans AGU            |
| European Journal of Forest Research                             | Eur J For Res            |
| Experiments in Fluids                                           | Exp Fluids               |
| Fisheries Research                                              | Fish Res                 |
| Flow Turbulence and Combustion                                   | Flow Turbul Combust      |
| Forestry                                                         | Forestry                 |
| Freshwater Biology                                               | Freshwater Biol          |
| Functional Ecology                                               | Funct Ecol               |
| Acta Geodaetica et Geophysica Hungarica                          | Geod Geophys             |
| Geografiska Annaler Series A                                     | Geogr Ann Ser A          |
| Geography Compass                                                | Geogr Compass             |
| Geomorphology                                                    | Geomorphology            |
| Geophysical Research Letters                                     | Geophys Res Lett         |
| Geoscientific Instrumentation, Methods and Data Systems          | Geosci Instrum           |
| Geoscientific Model Development                                  | Geosci Model Dev         |
| Global Biogeochemical Sciences                                   | Glob Biogeochem Cycles   |
| Glocal Change Biology                                            | Glob Change Biol         |
| Hydrology and Earth System Sciences                              | Hydrol Earth Syst Sci    |
| Hydrological Processes                                           | Hydrol Proc              |
| IEEE Journal of Ocean Engineering                                | IEEE J Ocean Eng         |
| IEEE Transactions on Geoscience and Remote Sensing               | IEEE Trans Geosci Remote |
| International Journal of Climatology                            | Int J Climatol           |
| International Journal of Wildland Fire                          | Int J Wildland Fire      |
| International Journal of Heat and Fluid Flow                     | Int J Heat Fluid Flow     |
| Journal Name                                                                 | Abbreviation used in BLM                        |
|----------------------------------------------------------------------------|-----------------------------------------------|
| International Journal of Numerical Methods for Fluids                     | Int J Numer Methods Fluids                    |
| International Journal of Remote Sensing                                  | Int J Remote Sens                              |
| Izvestiya, Atmospheric and Oceanic Physics                                | Izv Atmos Ocean Phys                          |
| Journal of Advances in Modeling Earth Systems                             | J Adv Model Earth Syst                         |
| Journal of Aerosol Science                                               | J Aerosol Sci                                 |
| Journal of Agricultural Engineering Research                              | J Agric Eng Res                               |
| Journal of the Air Pollution Control Association                          | J Air Pollut Control Assoc                     |
| Journal of Aircraft                                                       | J Aircr                                       |
| Journal of Applied Meteorology and Climatology                            | J Appl Meteorol Clim                           |
| Journal of Applied Meteorology                                            | J Appl Meteorol                               |
| Journal of Aquatic Plant Management                                       | J Aquat Plant Manag                           |
| Journal of Arid Environments                                              | J Arid Environ                                |
| Journal of Atmospheric and Oceanic Technology                             | J Atmos OceanTechnol                          |
| Journal of Atmospheric Science                                           | J Atmos Sci                                   |
| Journal of Climate                                                        | J Clim                                        |
| Journal of Computational Physics                                         | J Comput Phys                                 |
| Journal of Earth Simulation                                               | J Earth Simul                                 |
| Journal of Earth System Science                                          | J Earth Syst Sci                              |
| Journal of Environmental Engineering                                     | J Environ Eng                                 |
| Journal of Experimental Botany                                           | J Exp Bot                                     |
| Journal of the Faculty of Science Hokkaido University                    | J Fac Sci Hokkaido Univ                       |
| Journal of Field Robotics                                                 | J Field Robot                                 |
| Journal of Fluid Mechanics                                               | J Fluid Mech                                  |
| Journal of Geophysical Research                                         | J Geophys Res                                 |
| Journal of Geophysical Research-Atmospheres                               | J Geophys Res Atmos                           |
| Journal of Glaciology                                                     | J Glaciol                                     |
| Journal of Great Lakes Research                                          | J Great Lakes Res                             |
| Journal of Hazardous Materials                                           | J Hazard Mater A                              |
| Journal of Heat Transfer                                                 | J Heat Transf                                 |
| Journal of Hydraulic Engineering                                         | J Hydraul Eng                                 |
| Journal of Hydrology                                                      | J Hydrol                                     |
| Journal of Hydrometeorology                                              | J Hydrometeorol                               |
| Journal of Marine Research                                               | J Mar Res                                     |
| Journal of Marine Systems                                                | J Mar Syst                                    |
| Journal de Mathematiques Pures et Appliquees                             | J Math Pures Appl                             |
| Journal Name                                                                 | Abbreviation used in BLM                  |
|----------------------------------------------------------------------------|------------------------------------------|
| Journal of Meteorology                                                     | J Meteorol                                |
| Journal of the Meteorological Society of Japan                            | J Meteorol Soc Jpn                        |
| Journal of Oceanography                                                    | J Oceanogr                                |
| Journal of Operational Oceanography                                        | J Oper Oceanogr                           |
| Journal of the operational Research Society                                | J Oper Res Soc                            |
| Journal of the Optical Society of America                                  | J Opt Soc Am                              |
| Journal of Plankton Research                                              | J Plankton Res                            |
| Journal of Solar Energy Engineering                                        | J Sol Energy Eng                          |
| Journal of Quantitative Spectroscopy and Radiative Transfer               | J Quant Spectrosc Radiat Transf           |
| Journal of Renewable and Sustainable Energy                               | J Renew Sust Energy                       |
| Journal of Scientific Statistical Computing                                | J Sci Stat Comput                         |
| Journal of Statistical Physics                                            | J Stat Phys                               |
| Journal of Thermophysics and Heat Transfer                                 | J Thermophys Heat Transf                  |
| Journal of Tropical Ecology                                                | J Trop Ecol                               |
| Journal of Turbulence                                                     | J Turbul                                 |
| Journal of Wind Engineering and Industrial Aerodynamics                   | J Wind Eng Ind Aerodyn                    |
| Landscape and Urban Planning                                              | Landsc Urban Plan                         |
| Limnology and Oceanography                                                | Limnol Oceanogr                           |
| Low Temperature Science                                                   | Low Temp Sci                              |
| Machine Learning                                                          | Mach Learn                               |
| Marine Chemistry                                                          | Mar Chem                                 |
| Mathematische Annalen                                                     | Math Ann                                 |
| Meteorological Applications                                               | Meteorol Appl                             |
| Meteorology and Atmospheric Physics                                        | Meteorol Atmos Phys                       |
| Meteorologische Zeitschrift                                               | Meteorol Z                               |
| Monthly Weather Review                                                    | Mon Weather Rev                          |
| Natural hazards and Earth System Sciences                                  | Nat Hazards Earth Syst Sci               |
| Nature Climate Change                                                     | Nat Clim Change                          |
| Nature Letters                                                            | Nat Clim Change                          |
| Nature Geoscience                                                         | Nat Geosci                               |
| Neural Computation                                                        | Neural Comput                            |
| Nonlinear Processes in Geophysics                                         | Nonlin Process Geophys                   |
| New Zealand Journal of Science                                            | N Z J Sci                                |
| Oceanography                                                              | Oceanography                             |
| Ocean Dynamics                                                            | Ocean Dyn                                |
| Ocean Engineering Science                                                 | Ocean Eng Sci                            |
| Ocean Dynamics                                                            | Ocean Dyn                                |
| Journal of the Optical Society of America                                  | J Opt Soc Am                              |
| Journal of Renewable and Sustainable Energy                               | J Renew Sust Energy                       |
| Journal of Scientific Statistical Computing                                | J Sci Stat Comput                         |
| Journal of Statistical Physics                                            | J Stat Phys                               |
| Journal of Thermophysics and Heat Transfer                                 | J Thermophys Heat Transf                  |
| Journal of Tropical Ecology                                                | J Trop Ecol                               |
| Journal of Turbulence                                                     | J Turbul                                 |
| Journal of Wind Engineering and Industrial Aerodynamics                   | J Wind Eng Ind Aerodyn                    |
| Journal Name                                                                 | Abbreviation used in BLM                      |
|----------------------------------------------------------------------------|-----------------------------------------------|
| Ocean Modeling                                                              | Ocean Model                                   |
| Papers in Physical Oceanography and Meteorology                            | Pap Phys Oceanogr Meteoro                  |
| Particle & Particle Systems Characterization                               | Part Syst Charact                            |
| Particuology                                                               | Particuology                                  |
| Philosophical Transactions of the Royal Society of London                  | Philos Trans R Soc                            |
| Photogrammetric Engineering and Remote Sensing                             | Photogramm Eng Remote Sens                   |
| Physical Review Letters                                                    | Phys Rev Lett                                 |
| Physical Review E                                                          | Phys Rev E                                   |
| Physics and Chemistry of the Earth                                         | Phys Chem Earth                               |
| Physics of Fluids                                                          | Phys Fluids                                  |
| Physics A - Statistical Mechanics and its Applications                     | Physica A Stat Mech Appl                     |
| Physica D                                                                  | Physica D                                    |
| Plant Biosystems                                                           | Plant Biosyst                                |
| PLOS One                                                                   | PLOS One                                     |
| Powder technology                                                          | Powder Technol                                |
| Proceedings of the Royal Society                                          | Proc Roy Soc                                  |
| Progress in Aerospace Science                                              | Prog Aerosp Sci                               |
| Progress in Heat and Mass Transfer                                         | Prog Heat Mass Transf                         |
| Progress in Physical Geography                                             | Prog Phys Geogr                               |
| Pure and Applied Geophysics                                                | Pure Appl Geophys                             |
| Quarterly Journal of the Royal Meteorological Society                      | Q J R Meteorol Soc                             |
| Remote Sensing                                                             | Remote Sens                                  |
| Remote Sensing of Environment                                              | Remote Sens Environ                           |
| Renewable Energy                                                           | Renew Energy                                 |
| Reviews of Geophysics                                                      | Rev Geophys                                  |
| Reviews of Geophysics and Space Physics                                    | Rev Geophys Space Phys                       |
| Review of Scientific Instruments                                           | Rev Sci Inst                                 |
| Science                                                                    | Science                                      |
| Science of the Total Environment                                           | Sci Tot Environ                               |
| Sedimentology                                                              | Sedimentol                                   |
| Siam Journal on Applied Mathematics                                        | SIAM J Appl Math                              |
| Tellus                                                                     | Tellus                                       |
| Tellus Series B - Chemical and Physical Meteorology                        | Tellus Ser B Chem Phys Meteorol              |
Abbreviations
ALL: Acutely lymphoblastic leukemia; AML: Acute myeloid leukemia; MCMC: Markov chain Monte Carlo; PC: Principal components PC; SRTM: NASA’s shuttle radar topography mission

Acknowledgements
The authors thank Peter McCullagh for his insightful comments and suggestions on an early version of this paper. The authors are grateful to the Editor-in-Chief, the Associate Editor and anonymous reviewers for their constructive comments and suggestions which led to remarkable improvement of the paper.

Authors’ contributions
Hsin-Hsiung Huang wrote the draft of the manuscript, developed the algorithms, and conducted the experiments. Jie Yang proposed the methods and the initial algorithms. Both authors read and approved the final manuscript.

Authors’ information
Hsin-Hsiung Huang, Ph.D., is an Associate Professor in the Department of Statistics and Data Science at the University of Central Florida. Jie Yang, Ph.D., is an Associate Professor in the Department of Mathematics, Statistics, and Computer Science at the University of Illinois at Chicago.

Funding
National Science Foundation grants (DMS-1924792, DMS-1924859), the LAS Award for Faculty of Science at the University of Illinois at Chicago, and the In-House Award at the University of Central Florida.

Availability of data and materials
The datasets are from simulation and the UCI Machine Learning Repository and are available as per JSDA policy.

Competing interests
The authors declare that they have no competing interests.

Author details
1Department of Statistics and Data Science, University of Central Florida, Orlando, USA. 2Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, USA.

Received: 20 May 2020 Accepted: 5 October 2020
Published online: 28 October 2020
Neal, R. M.: Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Stat. 9, 249–265 (2000)
Ng, A., Jordan, M., Weiss, Y.: On spectral clustering: Analysis and an algorithm. Adv. Neural Inform. Process. Syst. 14, 849–856 (2001)
Ozawa, K.: A stratificational overlapping cluster scheme. Pattern Recognit. 18, 279–286 (1985)
Pitman, J.: Combinatorial Stochastic Processes. (621). In: Ecole d’Ete de Probabilites de Saint-Flour XXXII-2002. Dept. Statistics, U.C. Berkeley, (2002). Lecture notes for St. Flour course
Robert, C. P., Casella, G.: Introducing Monte Carlo Methods with R. Springer, (2010)
Roberts, G., Rosenthal, J.: Optimal Scaling for Various Metropolis-Hastings Algorithms. Stat. Sci. 16(4), 351–367 (2001)
Shioda R., Tuncel, L.: Clustering via minimum volume ellipsoids. Comput. Optim. Appl. 37, 247–295 (2007)
Stein, C.: Estimation of a covariance matrix. Reitz Lecture, IMS-ASA Annual Meeting in 1975 (1975)
Vogt, J. E., Piabnikaran, S., Fuchs, T. J., Roth, V.: The translation-invariant Wishart-Dirichlet process for clustering distance data. Proceedings of the 27th International Conference on Machine Learning, 1111–1118, (2010)
Ward, J. H.: Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. 58, 236–244 (1963)

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.