Model-Based Clustering for High Dimensional Data

Shahina Rahman* Valen E. Johnson† Irina Gaynanova‡ Anirban Bhattacharya§

Abstract

In a high-dimension feature space, $\boldsymbol{X_{n\times p}}$, we focus on clustering n items on the basis of p features when $p \gg n$. In traditional model based C-component clustering problem, with k parameters per distribution for each cluster, the number of parameters is $\mathcal{O}(\mathrm{pCk})$ $(p \ggg C)$. Hence, in genomics or other high-dimensional data applications, the problem is extremely challenging and often computationally infeasible. Instead of clustering on the original feature matrix $\boldsymbol{X}$, our clustering approach is based on a transformed space that can allow us to find the potential clusters in a much lower dimension, reducing the number of parameters to $\mathcal{O}(\mathrm{C^2k})$. A practical framework for Gaussian-based clustering approach is outlined based on the stochastic search algorithm proposed by [Booth et al., 2008]. To enforce and to ensure separation of the clusters, we use a non-local prior on the mean structure of the transformed cluster configuration. The performance of the proposed methods is studied by simulation, with encouraging results.

Key Words: Feature matrix; G-transformation; High-dimensional Data; Model-based clustering; Non-Local Prior; Reparametrization.

*Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143

†Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143

‡Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143

§Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143

1. Introduction

The goal of clustering is to organize data into a small number of homogeneous groups, thus aiding interpretation. Clustering techniques have been employed in a wide range of scientific fileds, including biology, physics, chemistry and psychology. these techniques can be broadly be classified into two categories: hierarchical methods and partition methods [Kaufman and Rousseeuw, 1990]. The former typically start from a dissimilarity matrixthat captures differences between the objects to be clustered and produce a family of cluster solutions, whose main propoerty is that any two clusters in the family are either disjoint or one is a superset of the other. various popular agglomerative algorithms, such as sibgle, complete and average linkage belong to this class. partition algorithms produce nonoverlapping clusters, whose defining characteristic is that distances between objects belonging to the same cluster are in some sense smaller than distances between the objects in different clusters. The popular K-means algorithm [MacQueen et al., 1967] and its variants are member of this class. A statistically motivated partition method is model-based clustering, which models the data as a sample from a Gaussian mixture distribution, with each component corresponding to a cluster [McLachlan and Basford, 1988]. A number of extensions addressing various aspects of this approach have recently appeared in the literature. For example, [Banfield and Raftery, 1993] generalized model-based clustering to the non-Gaussian case, where [Fraley and Raftery, 2002] extended it to incorporate hierarchical clustering techniques. The issue of variable selection in clustering has started receiving increased attention in the literature recently , [Parsons et al., 2004], [Friedman and Meulman, 2004], [Tadesse et al., 2005], [Hoff et al., 2006], [Guo et al., 2010] and many more.

In traditional model based C-component clustering problem, with k parameters per distribution for each cluster, the number of parameters is $\mathcal{O}(\mathrm{pCk})$ $(p \ggg C)$. Hence, in genomics or other high-dimensional data applications, the problem is extremely challenging and often computationally infeasible. To address this problem, this article proposes instead of clustering on the original feature matrix $\boldsymbol{X}$, our clustering approach is based on a transformed space that can allow us to find the potential clusters in a much lower dimension, reducing the number of parameters to $\mathcal{O}(\mathrm{C^2k})$. To address this problem, this article proposes a practical framework for Gaussian-based clustering approach is outlined based on the stochastic search algorithm proposed by [Booth et al., 2008]. To enforce and to ensure separation of the clusters, we use a non-local prior on the mean structure of the transformed cluster configuration.

The remainder of the article is organized as follows: Section 2 introduces the model based clustering

based on Gaussian mixture model. In Section 3, we propose the transformation on the feature matrix and discusses the new parameters of the transformed matrix. Section 4, we discuss about the priors taken on the Bayesian paradigm. The performance of the proposed method is discussed in Section 5 and 6.

2. Mixture Model Formulation

Suppose n samples have been collected on p variables and organized in a feature matrix $\mathbf{X} = ((x_{i,j}))_{n\times p}$. We focus on the problem of clustering the n rows into C non-overlapping homogeneous clusters. In a model-based clustering, a C-mean cluster problem can be described by a C-component Gaussian mixture. Specifically, the observations $\boldsymbol{x_i} = (x_{i,1}, \cdots, x_{i,p})$ are assumed to be independent and generated from the density

$$f(\boldsymbol{x_i}) = \sum_{k=1}^{C} w_k \phi(x_i; \boldsymbol{\eta_k}, \boldsymbol{\zeta_k}) \tag{1}$$

where $\phi(x_i; \boldsymbol{\eta_k}, \boldsymbol{\zeta_k})$ denotes the Gaussian density function with mean vector $\boldsymbol{\eta_k} = (\eta_{k,1}, \cdots, \eta_{C,p})$ and covariance matrix $\boldsymbol{\zeta}_k$,

$$\phi(\boldsymbol{x_i}; \boldsymbol{\eta_k}, \boldsymbol{\zeta_k}) \quad = \quad \frac{1}{(2\pi)^{p/2}\det(\boldsymbol{\zeta}_{\mathrm{k}})^{1/2}} \times \exp\Big\{ -\frac{1}{2}(\mathrm{x_i} - \boldsymbol{\eta}_{\mathrm{k}})\boldsymbol{\zeta}_{\mathrm{k}}{}^{-1}(\mathrm{x_i} - \boldsymbol{\eta}_{\mathrm{k}})^{\mathrm{T}}\Big\} \tag{2}$$

We additionally assume that $\boldsymbol{\zeta}_k$ is a diagonal matrix. The "weights" w_k's ($w_k \geq 0$ for all $1 \leq k \leq C$ and $\sum_{k=1}^{C} w_k = 1$) are the mixing coefficients, capturing the contribution of the k^{th} cluster. However, for our problem, it is convenient to write the likelihood in terms of the missing cluster identifier, $\boldsymbol{\gamma_i} \in \{1, \cdots, C\}$ for each samples, $i = 1, \cdots, n$. We also introduce the following notation: the mean parameters $\boldsymbol{\eta_k}$ can be collected into a $C \times p$ matrix, with rows corresponding to clusters and columns to variables as,

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_{1,1} & \eta_{1,2} & \cdots & \eta_{1,j} & \cdots & \eta_{1,p} \\ \eta_{2,1} & \eta_{2,2} & \cdots & \eta_{2,j} & \cdots & \eta_{2,p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \eta_{k,1} & \eta_{k,2} & \cdots & \eta_{k,j} & \cdots & \eta_{k,p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \eta_{C,1} & \eta_{C,2} & \cdots & \eta_{C,j} & \cdots & \eta_{C,p} \end{bmatrix}$$

and

$$\boldsymbol{\zeta} = \begin{bmatrix} \zeta_{1,1}^2 & \zeta_{1,2}^2 & \cdots & \zeta_{1,j}^2 & \cdots & \zeta_{1,p}^2 \\ \zeta_{2,1}^2 & \zeta_{2,2}^2 & \cdots & \zeta_{2,j}^2 & \cdots & \zeta_{2,p}^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \zeta_{k,1}^2 & \zeta_{k,2}^2 & \cdots & \zeta_{k,j}^2 & \cdots & \zeta_{k,p}^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \zeta_{C,1}^2 & \zeta_{C,2}^2 & \cdots & \zeta_{C,j}^2 & \cdots & \zeta_{C,p}^2 \end{bmatrix}$$

We had earlier assumed that for each k^{th} cluster, $\boldsymbol{\zeta}_k = \mathrm{Diag}(\zeta_{\mathrm{k,1}}^2, \zeta_{\mathrm{k,2}}^2, \cdots, \zeta_{\mathrm{k,p}}^2)$. Hence, for $i = 1, \cdots, n$, and for $j = 1, \cdots, p$, $x_{i,j} \sim \mathrm{Normal}(\eta_{\gamma_i,j},\ \zeta_{\gamma_i,j}^2)$ independently.

3. G-Transformation and the Reparametrization

To reduce the dimension of the problem, we will base the clustering on the following $n \times n$ matrix,

$$\boldsymbol{G} = \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}/p \tag{3}$$

We define the elements of $\boldsymbol{G}$ matrix as $((g_{ij}))_{n\times n}$. And we will formulate our clustering algorithm based on $((g_{ij}))_{n\times n}$. Based on the original C-component clusters on the feature matrix $\mathbf{X}$, $\boldsymbol{G}$ matrix shows C-clusters in its diagonal and off-diagonals. Figure 1 depicts the true cluster configuration in $\boldsymbol{G}$ matrix when the true number of clusters equals to 3. The explicit expression for the mean and the variance of $((g_{ij}))_{n\times n}$ is as follows.

I . On diagonals of $\boldsymbol{G}$, for $i = j$ (same subject and same cluster),

$$\begin{aligned} E(g_{ii}) &= \frac{1}{p}\sum_{k=1}^{p}(\eta_{\gamma_i,k}^2 + \zeta_{\gamma_i,k}^2) = \mu_{\gamma_i\gamma_i} \\ V(g_{ii}) &= \frac{2}{p^2}\sum_{k=1}^{p}\zeta_{\gamma_i,k}^4\big(1 + 2\frac{\mu_{\gamma_i,k}^2}{\zeta_{\gamma_i,k}^2}\big) = \sigma_{\gamma_i\gamma_i}^2 \end{aligned}$$

2 . On homogeneous offdiagonals of $\boldsymbol{G}$, for $i \neq j$ (different subjects) and $\gamma_i = \gamma_j$ (same cluster)

$$\begin{aligned} E(g_{ij}) &= \frac{1}{p}\sum_{k=1}^{p}\eta_{\gamma_i,k}^2 = \mu_{\bullet\gamma_i} \\ V(g_{ij}) &= \frac{1}{p^2}\sum_{k=1}^{p}\zeta_{\gamma_i,k}^4\big(1 + 2\frac{\mu_{\gamma_i,k}^2}{\zeta_{\gamma_i,k}^2}\big) = \sigma_{\gamma_i}^2/2 = \sigma_{\bullet\gamma_i}^2 \end{aligned}$$

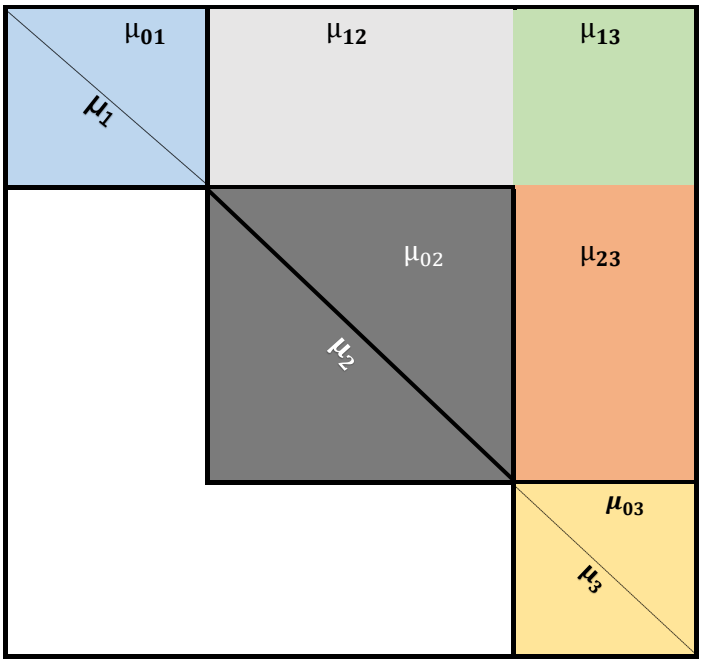

Figure 1: The mean parameters for $\boldsymbol{G}$ matrix when the number of true cluster is 3. It shows $\boldsymbol{\mu_1}, \boldsymbol{\mu_2}$ and $\boldsymbol{\mu_3}$ representing the mean parameter for the diagonals, $\boldsymbol{\mu_{01}}, \boldsymbol{\mu_{02}}$ and $\boldsymbol{\mu_{03}}$ representing the mean parameters in the homogeneous off diagonals and $\boldsymbol{\mu_{12}}, \boldsymbol{\mu_{13}}$ and $\boldsymbol{\mu_{23}}$ representing the mean parameters in the heterogeneous off diagonals.

3 On heterogeneous offdiagonal of $\boldsymbol{G}$, for $i \neq j$ (different subjects) and $\gamma_i \neq \gamma_j$ (different cluster)

$$
\begin{aligned}
E(g_{ij}) &= \frac{1}{p}\sum_{k=1}^{p} \eta_{\gamma_i,k}\eta_{\gamma_j,k} = \mu_{\gamma_i\gamma_j} \\
V(g_{ij}) &= \frac{1}{p^2}\sum_{k=1}^{p} \zeta^2_{\gamma_i,k}\zeta^2_{\gamma_j,k}\big(1 + \frac{\mu^2_{\gamma_i,k}}{\zeta^2_{\gamma_i,k}} + \frac{\mu^2_{\gamma_j,k}}{\zeta^2_{\gamma_j,k}}\big) = \sigma^2_{\gamma_i\gamma_j}.
\end{aligned}
$$

We organize the mean parameters of $\boldsymbol{G}$ matrix in the following matrix,

$$
\boldsymbol{\mu} = \begin{bmatrix}
\mu_{11} & \mu_{\bullet 1} & \mu_{12} & \mu_{13} & \cdots & \mu_{1C} \\
\mu_{22} & & \mu_{\bullet 2} & \mu_{23} & \cdots & \mu_{2C} \\
\mu_{33} & & & \mu_{\bullet 3} & \cdots & \mu_{3C} \\
\cdots & & & & \cdots & \cdots \\
\mu_{CC} & & & & & \mu_{\bullet C}
\end{bmatrix} \tag{4}
$$

Hence, the number of mean parameters that are need to be evaluated for a cluster configuration with C-components is $C(C+3)/2$. This is a drastic reduction of the number of mean parameters in the original

matrix was $C \times p$. Similarly, we organize the variance parameters of $\boldsymbol{G}$ matrix in the following matrix,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \sigma_{13}^2 & \cdots & \sigma_{1C}^2 \\ & \sigma_{22}^2 & \sigma_{23}^2 & \cdots & \sigma_{2C}^2 \\ & & \sigma_{33}^2 & \cdots & \sigma_{3C}^2 \\ & & & \cdots & \cdots \\ & & & & \sigma_{CC}^2 \end{bmatrix} \tag{5}$$

Note that the number of variance parameter for $\boldsymbol{G}$ matrix is $C(C+1)/2$ because for each $k \in \{1, 2, .., C\}$, $\sigma_{\bullet k}^2 = \sigma_{kk}^2/2$.

3.1 Distributional Assumption on $\boldsymbol{G}$

We assumed in the previous section that for $i = 1, \cdots, n$, and for $j = 1, \cdots, p$, $x_{i,j} \sim \text{Normal}(\eta_{\gamma_i,j},\ \zeta_{\gamma_i,j}^2)$ independently. Since, $g_{ij} = \sum_{k=1}^{p} x_{ik}x_{jk}/p$, for $p \ggg 1$, we can assert the CLT on each g_{ij} and assume the following Gaussian distribution for $(i, j) \in \{1, 2, \cdots, n\} \times \{1, 2, \cdots, n\}$,

I. For $i = j$, $g_{ii} \sim \text{Normal}(\mu_{\gamma_i\gamma_i}, \sigma_{\gamma_i\gamma_i}^2)$.

II. For $i \neq j$ and $\gamma_i = \gamma_j$, $g_{ij} \sim \text{Normal}(\mu_{\bullet\gamma_i}, \sigma_{\gamma_i\gamma_i}^2/2)$.

III. For $i \neq j$ and $\gamma_i \neq \gamma_j$, $g_{ij} \sim \text{Normal}(\mu_{\gamma_i\gamma_j}, \sigma_{\gamma_i\gamma_j}^2)$.

4. Prior formulation on G parameters

4.1 Prior on the Cluster Configuration: Stochastic Search Algorithm

The goal of this article is to find the posterior distribution of the cluster id's, $\gamma = \{\gamma_i\}$, $i = 1, ..., n$. Let c denotes the number of clusters in a partition, $n_1, n_2, \cdots, n_c$ number of items in each cluster. The estimation of the cluster ids are elaborated in two steps.

[A] Prior formulation:

Then given, $m = 1/n$, the prior on the cluster configuration is [Crowley, 1997] and [Booth et al., 2008]:

$$\frac{\Gamma(m)m^c}{\Gamma(n+m)} \prod_{k=1}^{c} \Gamma(n_k) \tag{6}$$

Clearly, this prior puts more weight on the set of cluster configurations having lower numbers of clusters. It heavily penalizes those configuration having clusters with fewer numbers of elements.

[B] Stochastic Search: *Biased Random walk and Split-Merge Algorithm*

We use the stochastic search algorithm driven by the mixture of two Metropolis-hastings algorithms proposed by [Booth et al., 2008]. This algorithm will allow us to make either local *Biased Random walk* or large scale *Split-Merge Algorithm* transition at each iteration with probabilities p_b and $1 - p_b$ respectively.

In *biased random walk MH Algorithm*, the are two cases: $c = 1$ and $c \geq 2$. If $c = 1$, choose one of the n objects uniformly at random and move the chosen object to its own cluster. If $c \geq 2$, choose one of the n objects uniformly at random. If the chosen object is a singleton, then move it to one of the other $c - 1$ clusters, each with probability $1/(c - 1)$. If the chosen object is not a singleton, then move it to one of the other $c-1$ clusters, each with probability $1/c$, or make the chosen object its own cluster with probability $1/c$.

In *split and merge* algorithm we randomly decide between a *merge* move with probability $p_m \in (0, 1)$ and a *split* move with probability $1 - p_m$. A merge proposal is constructed by merging two randomly chosen clusters in the current configuration. A split proposal is created by randomly choosing a cluster and then randomly splitting it into two clusters conditionally on neither being empty. A split move is automatically proposed whenever the current state consists of a single cluster, and likewise a merge move is automatically proposed when the current state consists of n clusters.

4.2 Prior on Mean Structure $\boldsymbol{\mu}$

The likelihood of the g-matrix is essentially a mixture of several Gaussian distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. In general, mixtures suffers from a lack of identifiability that plays a fundamental role both in estimation and model selection. This issue can be caused by over-fitting models that could be equivalently defined by less component mixture models. So our aim is to penalize the models or those cluster configurations which are over fitting the underlying true model. That motivates us to use the non-local prior on the mean structure $\boldsymbol{\mu}$ which is defined as follows:

$$\pi(\boldsymbol{\mu}|\delta) = \begin{cases} 0 & \max_k\{|\mu_{ik} - \mu_{jk}|, |\mu_{ii} - \mu_{jj}|, |\mu_{ij} - \mu_{ii}|, |\mu_{ij} - \mu_{jj}|, |\mu_{ii} - \mu_{jj}|\} < \delta \text{ for any } i \neq j \\ \frac{1}{4C} & C = \max_{ij}|g_{ij}| \end{cases} \tag{7}$$

Table 1: Design for Simulation 1

Feature	1-10	11-20	21-220
Cluster 1 (20 items)	$N(2.5, \sigma^2)$	$N(1.5, \sigma^2)$	$N(0, 1)$
Cluster 2 (20 items)	$N(0, \sigma^2)$	$N(1.5, \sigma^2)$	$N(0, 1)$
Cluster 3 (20 items)	$N(0, \sigma^2)$	$N(-1.5, \sigma^2)$	$N(0, 1)$
Cluster 4 (20 items)	$N(-2.5, \sigma^2)$	$N(-1.5, \sigma^2)$	$N(0, 1)$

4.3 Prior on Variance Structure Σ

We put an inverse-gamma prior on $\Sigma = (\sigma)_{ij}$ as follows:

For $(i, j) = \{1, ..., n\} \times \{1, ..., n\}$,

$$\pi(\sigma^2_{\gamma_i\gamma_i}) \sim \text{IG}(\alpha_1, \beta_1)$$

and

$$\pi(\sigma^2_{\gamma_i\gamma_j}) \sim \text{IG}(\alpha_2, \beta_2)$$

We recommend the values of hyper-parameters as $\alpha_1 = \alpha_2 = 6$ and $\beta_1 = 5/2p$ and $\beta_2 = 5/4p$ for our simulation study.

5. Simulation Results

In this section, we illustrate the performance of our proposed algorithm based on two synthetic examples with 4 clusters for simulation 1 and 2 used by [Guo et al., 2010]. In simulation 1, whereas in simulation 2, we generate different numbers of observation for different clusters. There are 4 clusters and $p = 220$, with the first 20 being informative and the remaining ones non-informative. The variables were generated according to the following mechanism: the first 20 are independently distributed Normal$(\boldsymbol{\eta_k}, \boldsymbol{\zeta_k})$ for cluster k, whereas the remaining 200 were generated from Normal$(0, 1)$ for all 4 clusters. Table 1 gives the means for the first 20 variables.

For example, in cluster 1, variables $1 - 10$ all have the same mean value 2.5, and variables $11 - 20$ all have the same mean value 1.5. Notice that variables $1 - 10$ are non-informative for separating clusters 2 and 3, whereas variables $11 - 20$ are non-informative for separating clusters 1 and 2 (as well as clusters 3 and 4). We consider two values of the common variance, $\sigma^2 = 1$ and $\sigma^2 = 4$. The former creates a high

"signal-to-noise ratio (SNR)" scenario, whereas the latter simulates a situation where the "SNR ratio" is low. We repeat the analysis 50 times for each simulation and record the average clustering success rates. In simulation 1, 20 observations are generated from each clusters and $p = 220$. Here the total number of observation is 80. In simulation 2, the clusters have different sample size. The sample size for cluster 3 and 4 has been increased to 200. Therefore, there are two small clusters (1 and 2) with 20 observations each and two large clusters (3 and 4) with 200 observations each.

Table 2: Design for Simulation 2

Feature	1-10	11-20	21-220
Cluster 1(20 items)	$N(2.5, \sigma^2)$	$N(1.5, \sigma^2)$	$N(0, 1)$
Cluster 2 (20 items)	$N(0, \sigma^2)$	$N(1.5, \sigma^2)$	$N(0, 1)$
Cluster 3 (200 items)	$N(0, \sigma^2)$	$N(-1.5, \sigma^2)$	$N(0, 1)$
Cluster 4 (200 items)	$N(-2.5, \sigma^2)$	$N(-1.5, \sigma^2)$	$N(0, 1)$

5.1 Results

Based on 50 simulated datasets, each dataset burned in with 50,000 updates (with annealing), then 100,000 updates. In high SNR, 100% of configurations correctly contained 4 clusters. The accuracy of individual classification was 98%. In low SNR, 78% of configurations contained 4 clusters, 22% 3 clusters classification accuracy was 87%. Performance was comparable to algorithms discussed in Guo et al. (2010).

MCMC algorithm started from random start typically gets stuck in mode with 5 or 6 clusters, with one or two of the large, "true" clusters split approximately evenly into 2 clusters. Small clusters estimated accurately. Post-processing probably useful to investigate number of clusters as δ is changed. 78% of configurations contained 4 clusters, 22% 5 clusters. Classification accuracy was 0.99%.

6. Discussion

Proposed algorithm provides promising results in simple examples. Posterior sampling and/or optimization remains problematic. Hyperparameter selection implicitly "defines" clusters. Computational algorithm is invariant to p, the number of features. Prescreening of features via entropy or variance measures might improve performance in low SNR settings.

References

James G Booth, George Casella, and James P Hobert. Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):119–139, 2008.

Leonard Kaufman and Peter J Rousseeuw. Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, pages 68–125, 1990.

James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.

Geoffrey J McLachlan and Kaye E Basford. *Mixture models: Inference and applications to clustering*, volume 84. Marcel Dekker, 1988.

Jeffrey D Banfield and Adrian E Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821, 1993.

Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631, 2002.

Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *Acm Sigkdd Explorations Newsletter*, 6(1):90–105, 2004.

Jerome H Friedman and Jacqueline J Meulman. Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(4):815–849, 2004.

Mahlet G Tadesse, Naijun Sha, and Marina Vannucci. Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470):602–617, 2005.

Peter D Hoff et al. Model-based subspace clustering. *Bayesian Analysis*, 1(2):321–344, 2006.

Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Pairwise variable selection for high-dimensional model-based clustering. *Biometrics*, 66(3):793–804, 2010.

Evelyn M Crowley. Product partition models for normal means. *Journal of the American Statistical Association*, 92(437):192–198, 1997.