BGMM Clustering Analysis

Jonathan Ma

2025-12-01

Table of Contents

1 Bayesian-Gaussian Mixture Model. 2
1.1 PCA 2
2 Patient Clustering 3
2.1 Clustering Patient Random EFfects 3
2.2 Summary by cluster 4
2.3 Age Distribution by Cluster 4
2.4 Race Distribution by Cluster 5
2.5 Cluster Size 6
2.6 Tumor Size by cluster 7

1 Bayesian-Gaussian Mixture Model.

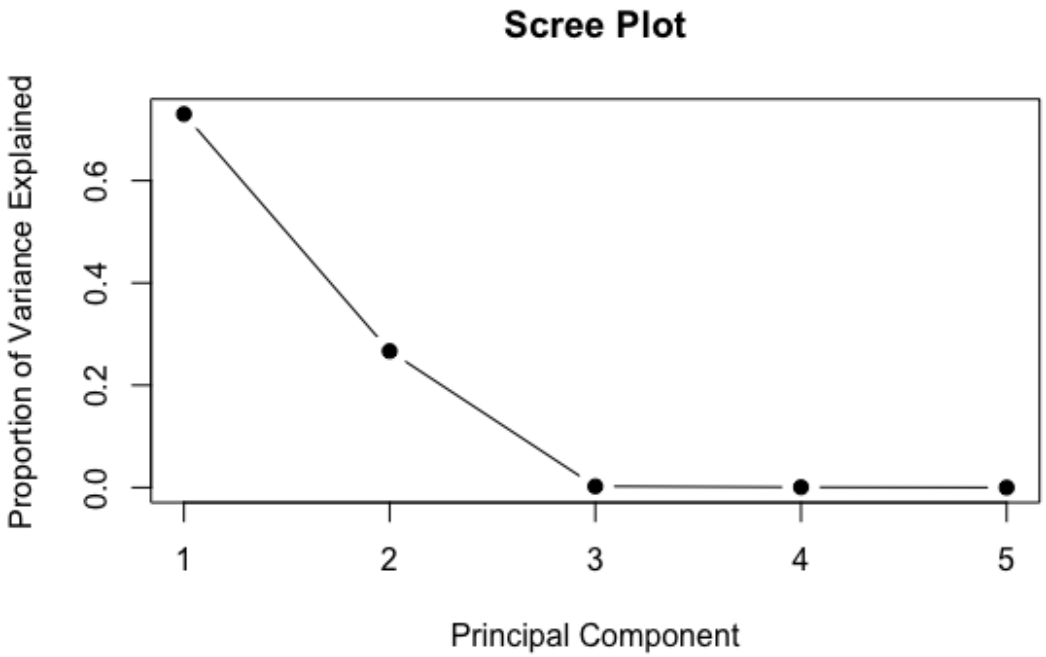
GMM is a probabilistic clustering method assuming data is generated from multiple gaussian distributions, and the Bayesian version extends this by incorporating prior distributions, which makes inference more reliable when data is noisy and small. BGMM also uses posterior probabilities for a softer clustering approach that is useful when the group structure is nested, as is patient-level health data

1.1 PCA

High dimensional datasets often contain collinear features that may lead mixtures to overfit, so we employ PCA so the original variables into a smaller set of uncorrelated components capture the majority of the variance in the data. This not only speeds up the BGMM but also improves stability of estimates, maximizing our signal-to-noise ratio

PCA Summary: Proportion of Variance Explained

Principal Component	Standard deviation	Proportion of Variance	Cumulative Proportion
PC1	1.9106	0.7301	0.7301
PC2	1.1550	0.2668	0.9969
PC3	0.1100	0.0024	0.9993
PC4	0.0582	0.0007	1.0000



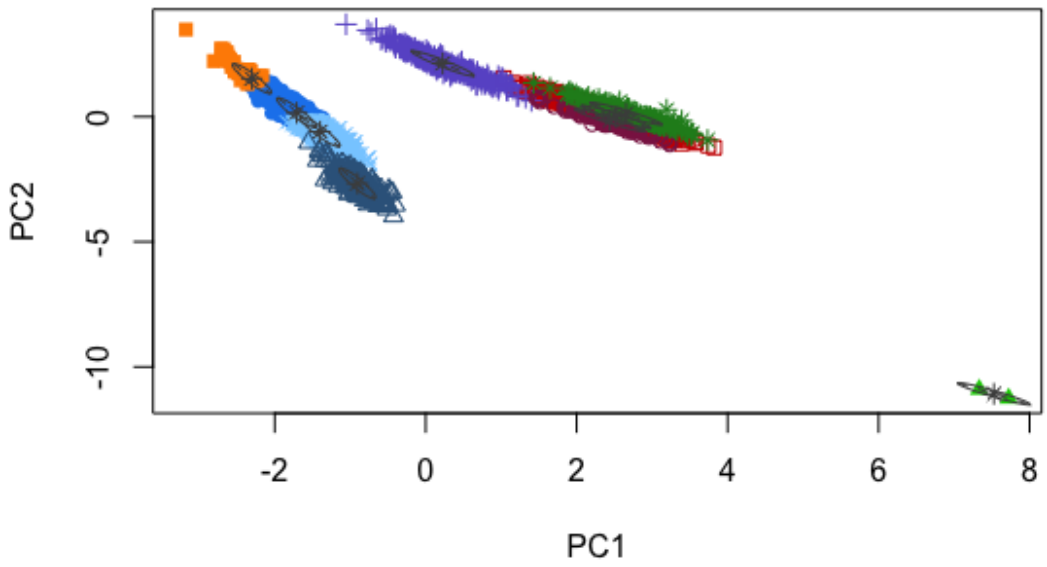
PC1 explained 73.0% and PC2 explained 26.7%, for a cumulative variance of 99.7%. This strong two-dimensional representation justifies the use of only the first two principal components for downstream clustering. The scree plot visually supports this choice, showing a sharp drop in variance after PC2, with subsequent components contributing negligible information.

2 Patient Clustering

After reducing the dimensionality with PCA, the next step is to cluster patients based on their posterior patient-level effects. These effects summarize each patient's underlying risk profile after accounting for tumor-level covariates and regional variation. The resulting clusters help highlight clinically meaningful patterns such as age, race, or tumor size distributions, that may signal differential risk pathways within the population.

2.1 Clustering Patient Random EFfects

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EEV (ellipsoidal, equal volume and shape) model with 9 components:
##
##  log-likelihood    n df       BIC       ICL
##       -4111.812 3026 37 -8520.179 -10111.73
##
## Clustering table:
##   1   2   3   4   5   6   7   8   9
## 697 777   2 314  56 816 101 119 144
```



2.2 Summary by cluster

We now merge the clusters back to the metadata, and summarize by cluster.

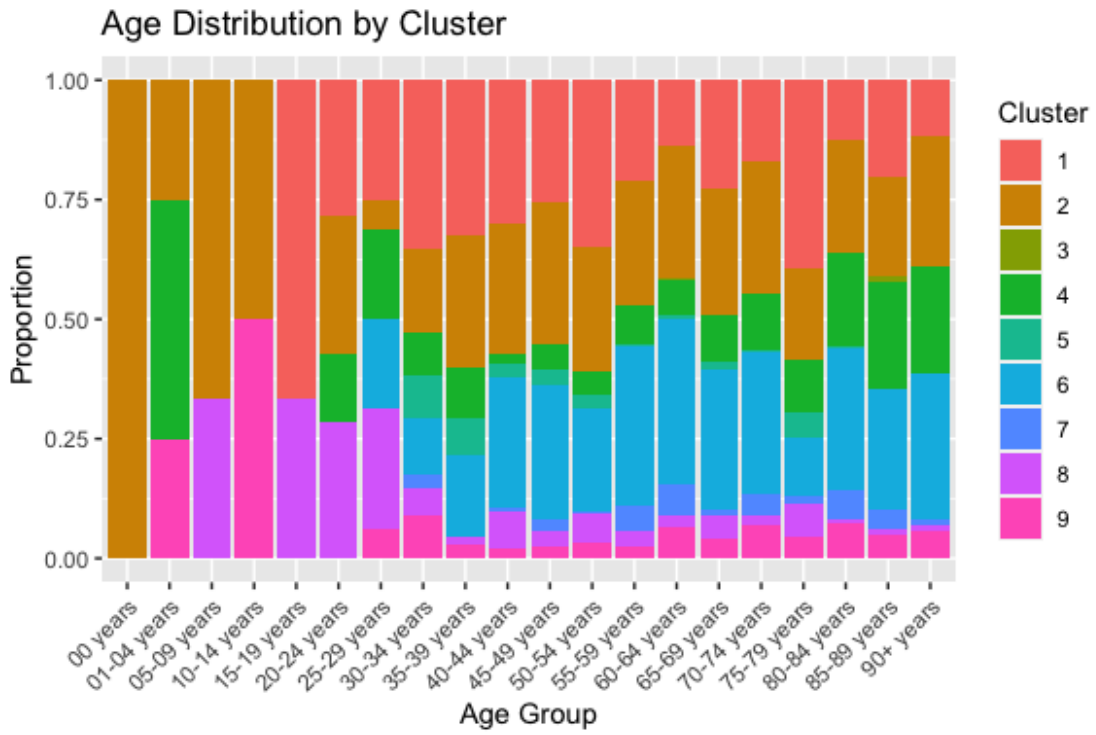
Table: Demographic Summary by Cluster

Cluster	Count	Most Common Age	Most Common Race	Most Common Sex	Most Common Region
1	697	75-79 years	W	Female	1
2	777	60-64 years	W	Female	1
3	4	60-64 years	W	Female	1
4	314	70-74 years	W	Male	1
5	56	75-79 years	O	Female	1
6	817	60-64 years	W	Female	1
7	101	60-64 years	W	Female	1
8	120	75-79 years	W	Female	1
9	145	70-74 years	W	Female	1

2.3 Age Distribution by Cluster

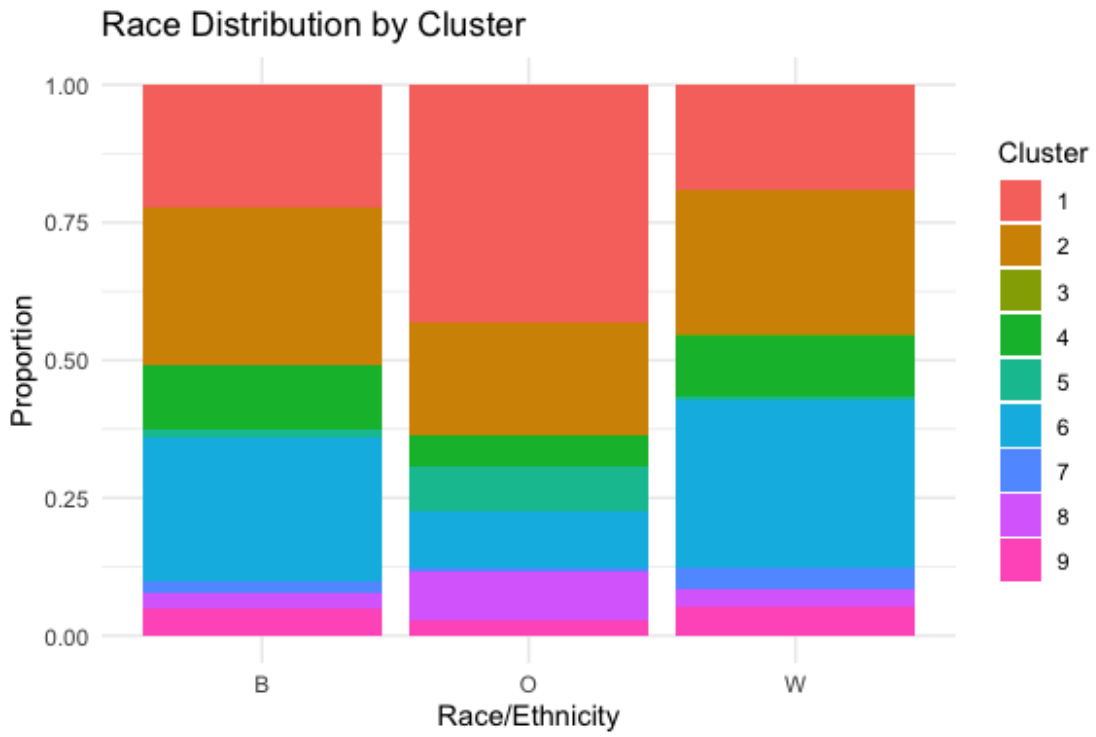
The age–by–cluster distribution shows that the patient groups differ substantially in their age composition. The largest cluster, representing patients in their mid-70s to late-70s, dominates the older age categories, indicating that one major group is composed primarily

of elderly individuals. In contrast, several other clusters including those corresponding to the 60–64 and 65–69 age ranges are concentrated among younger seniors, suggesting a distinct, somewhat healthier or earlier-stage demographic. Smaller clusters begin to appear gradually through middle and later adulthood, reflecting more specialized or less common patient profiles. Overall, the visualization highlights age as a key driver of cluster separation, with some groups representing the oldest and potentially highest-risk patients and others representing comparatively younger subsets within the older adult population.



2.4 Race Distribution by Cluster

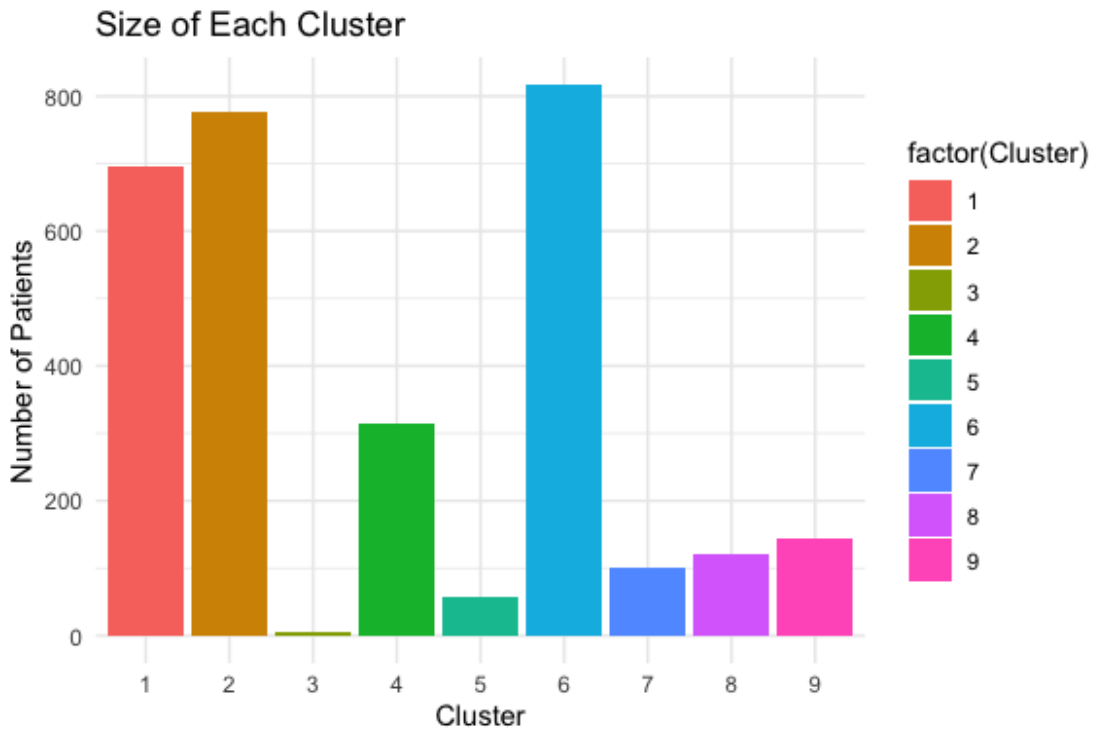
The race distribution reveals that some clusters are broadly represented across all racial groups, while others are more concentrated within specific categories. Clusters 4 and 5 appear consistently across Black, White, and “Other” race groups, suggesting that these clusters capture demographic or clinical patterns not strongly tied to race. Cluster 1, which contains many of the older and higher-risk patients, makes up a relatively larger share of White and Black patients compared to the “Other” group. Conversely, the smaller clusters particularly the ones associated with clusters 8 and 9 are more prominent within the “Other” race category, indicating that these patients may have distinct characteristics or risk factors. Overall, the racial patterns suggest moderate heterogeneity, with some clusters distributed broadly across groups and others more race-specific.



2.5 Cluster Size

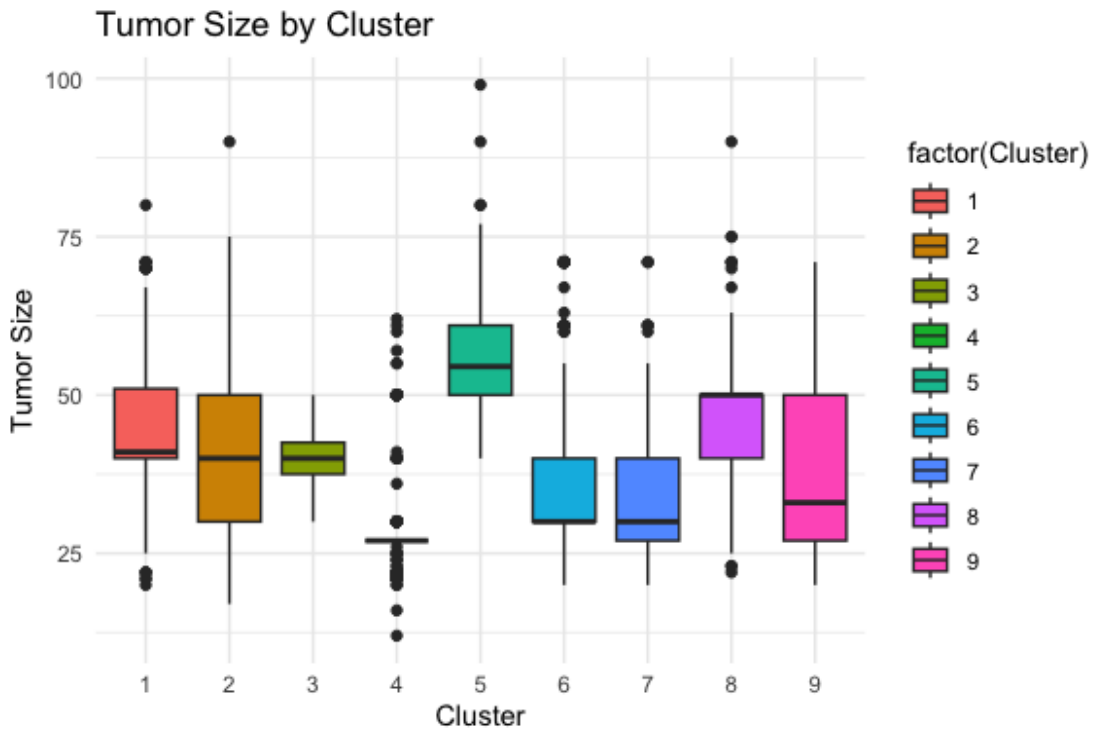
The two largest groups are composed of White females aged 60–64 years, collectively representing over 1,600 patients. This suggests that this demographic is highly prevalent in the dataset and may define the “typical” patient profile. Another large group consists of White females aged 75–79 years, with over 500 patients, making it a substantial older-age subgroup.

A moderately sized group includes White males aged 70–74 years, suggesting a distinct older male segment. Smaller profile groups include patients identified as ‘Other’ race and aged 75–79, as well as White females aged 70–74, indicating some demographic diversity among mid-sized clusters.



2.6 Tumor Size by cluster

The group of White females aged 60–64 and the group of White females aged 75–79 exhibit the largest median tumor sizes, along with wide variability and several extreme outliers, suggesting delayed diagnoses or more aggressive disease in these subpopulations. In contrast, the group of White females aged 70–74 has the smallest tumor sizes with a tighter distribution, potentially indicating earlier detection or lower underlying risk. Clusters dominated by non-White or “Other” race patients (e.g., females aged 75–79) display moderate to high tumor sizes, but with more consistent ranges, possibly reflecting systemic barriers to timely diagnosis. Interestingly, younger groups or smaller clusters do not always correspond to lower tumor burden, suggesting that age alone does not explain variation in tumor size.


Tumor Size by Cluster
Tumor Size
100
75
50
25
1
2
3
4
5
6
7
8
9
Cluster
factor(Cluster)
1
2
3
4
5
6
7
8
9