A Diagnostic Analysis for Gaussian Mixture Models in Single-Cell Population Analysis

Quanjin (Terry) Su
UC Berkeley, Purdom Group

September 25, 2025

1 Introduction

Single-cell RNA sequencing allows for the analysis of gene expression profiles at the individual cell level, which reveals inter-cellular heterogeneity. In population studies, comparing single-cell data from different samples or conditions (e.g., diseased vs. healthy) is used to identify key cell subpopulations and molecular mechanisms. Although single-cell data is often analyzed at the cell level, sample-level comparisons (e.g., across cohorts or time points) is equally important. The GloScope method [1] addresses this by modeling the entire cell population of each sample as a single probability distribution. Specifically, GloScope fits a Gaussian Mixture Model (GMM) to the cell distribution for each sample in a low-dimensional embedding space. It then quantifies the differences between samples by calculating the KL-divergence between these probability densities. This makes the GloScope method well suited for sample-level comparisons and batch effect evaluation. It is important to note that the sample divergence calculated by GloScope naturally includes differences in both cell expression, which determines a cell's location in the low-dimensional embedding space, as well as differences in cell composition (i.e., different proportions of cell types).

However, a central challenge in interpreting GloScope's results arises from technical batch effects, which are common in single-cell RNA sequencing. These effects can introduce bias into GloScope analysis, which focuses on the overall distribution of samples. Traditional batch correction methods, like Harmony[2], effectively correct for differences in expression and position caused by batch effects by aligning cells from different batches in a low-dimensional space. However, methods like Harmony were not designed to adjust cell type proportions and therefore cannot eliminate inherent differences in cell composition between samples. This leads to a central challenge in interpreting GloScope's results after batch correction. While methods like Harmony can align cell populations to correct for technical differences in expression, they are not designed to alter the underlying cell type proportions. Consequently, a large post-correction divergence measured by GloScope remains ambiguous. For instance, if we compare a patient's sample before and after treatment, a significant divergence could signal a desired therapeutic shift in cell states. However, it could also simply reflect a change in the proportion of immune cell types, with no actual change in their individual expression profiles. This ambiguity makes it difficult to isolate true biological state changes from simple compositional shifts.

To address this ambiguity, this project aims to systematically investigate whether the Gaussian Mixture Model (GMM) itself offers a way to deconstruct these sources of divergence. In theory, if its mixture components are interpreted as distinct cell types, the model parameter

π would correspond to the cell composition, while μ (the mean) would represent the location of the cell types and Σ (the covariance) would capture their expression variability and shape. This project aims to systematically investigate the feasibility of this interpretation in real data. Specifically, we seek to determine if a GMM can stably and reliably decompose single-cell data into its cell type and expression state components, which would potentially allow for the quantification of these two sources of difference.

To this end, we conducted an in-depth diagnostic analysis of the GMM components based on the GloScope method using an HIV-PBMC (peripheral blood mononuclear cell) dataset. However, our analysis ultimately shows that the GMM components prioritize fitting the overall density of the data. As a result, their correspondence with biologically defined cell types is weak, and the component-to-cell-type mapping is not consistent across different samples.

2 Methodology

2.1 Modeling Cell Distributions with Gaussian Mixture Models (GMMs)

To represent the cell distribution within each sample, we model its low-dimensional embedding with a Gaussian Mixture Model (GMM), which assumes the data is generated from a weighted sum of k multivariate Gaussian distributions. For a single sample, this allows its probability density function $f(x)$ to be expressed as a weighted sum of k Gaussian components:

$$f(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \tag{1}$$

where the model parameters are:

- k: The number of Gaussian components (i.e., clusters) in the model.
- π_k: The mixture proportion, representing the fraction of cells belonging to component k.
- μ_k: The mean vector, representing the center of component k in the low-dimensional space.
- Σ_k: The covariance matrix, which describes the shape and spread of component k.

In our analysis, a separate GMM is fitted independently to each sample j. Therefore, all parameters are estimated for each sample, denoted by k_j, π_{jk}, μ_{jk}, and Σ_{jk}.

2.2 Parameter Estimation and Model Selection

A critical step in the GloScope framework is to establish a common coordinate system for all samples. This is achieved by first combining the normalized expression data from all cells across all samples and then performing Principal Component Analysis (PCA) on this combined dataset. The GMM for each sample is subsequently fitted to the coordinates of its cells within this shared, low-dimensional PCA space. The model parameters π_k, μ_k, and Σ_k are unknown and must be estimated from the data. We use the Expectation-Maximization (EM) algorithm to find the parameter estimates that maximize the data log-likelihood. This process automates model selection using the R package `mclust`[3]. It iterates through a pre-set range of values for k (e.g., 1-9) and tests various covariance structure models. It ultimately selects the optimal combination of k and model type based on the highest Bayesian Information Criterion (BIC). Common covariance models used for this sample include:

- EVE: Components have the same volume and orientation, but their shapes can vary.
- VVE: Components have the same orientation, but their volumes and shapes can vary.
- VVV: Components can vary in volume, shape, and orientation, representing the most flexible model.

These models impose constraints on the geometric properties of the covariance matrix Σ_k, which can be understood through its spectral (eigen) decomposition. The **orientation** of the Gaussian ellipsoid is defined by the eigenvectors, which determine the directions of its principal axes. Its **shape** is determined by the relative sizes of the eigenvalues, controlling the ellipsoid's eccentricity (i.e., whether it is spherical or stretched). Finally, its **volume** is proportional to the product of these eigenvalues. For example, the **VVE** model, which was most frequently selected in our analysis, assumes that all components have **V**arying volume and **V**arying shape, but share an **E**qual orientation. This means each component k has its own unique eigenvalues but shares the same set of eigenvectors with other components in the mixture.

2.3 Quantifying Sample Divergence with KL-Divergence

After fitting a GMM for each sample, GloScope quantifies the divergence $d(f_j, f_l)$ between any two samples, j and l, by calculating the symmetrized Kullback-Leibler (KL) divergence:

$$d(f_j, f_l) = KL(f_j||f_l) + KL(f_l||f_j) \tag{2}$$

where,

$$KL(f_j||f_l) = \int f_j(x) \log \frac{f_j(x)}{f_l(x)} dx \tag{3}$$

The KL-divergence measures the information loss when one distribution is used to approximate another. Because the GMM density is determined by π, μ, and Σ, the KL distance is a composite measure that incorporates differences in both cell composition and expression/position. Finally, this process yields an $n \times n$ distance matrix, which can be visualized using heatmaps and multidimensional scaling (MDS). Biologically, a larger KL-divergence indicates that the overall cell distributions of two samples have undergone a substantial change in their composition and/or expression patterns. For example, when comparing the early acute and chronic stages of HIV infection, an increase in the KL value could reflect a significant change in immune cell proportions or a major shift in the transcriptomic state of the same cell types. This aligns with the research motivation discussed previously—the divergence captured by GloScope is comprehensive, reflecting combined changes in composition and expression rather than either one alone.

2.4 Core Methodological Challenge: The "Matching Problem"

An idealized assumption is that a GMM component could directly correspond to a specific biological cell type. If this assumption holds, the parameters would neatly separate cell composition (π) from cell state (μ, Σ). The goal of batch correction would be to align the cell states, ensuring that for corresponding clusters across samples A and B, $\mu_{kA} \approx \mu_{kB}$ and $\Sigma_{kA} \approx \Sigma_{kB}$. This would enable a powerful, disentangled analysis: one could, for instance, create a divergence metric that compares only the cell state parameters (μ, Σ) to measure shifts in expression, while ignoring differences in cell composition (π).

However, a key difficulty arises from this naive approach because the GMM is fitted independently for each sample. This independence means there is no enforced correspondence between the models from different samples. For instance, the optimal number of components, k, may differ from one sample to the next. Even if k happens to be the same, there is no guarantee that component 1 from sample A corresponds to the same cell type as component 1 from sample B. Furthermore, the relationship between statistical components and biological cell types is often complex: a single biological cell type (e.g., T cells) might be fitted as a single component in one sample but be split into multiple components in another, or vice versa.

This uncertainty leads to the "matching problem": we cannot know if or how the GMM components from different samples correspond to one another, making it difficult to separate compositional differences from expression differences by directly comparing GMM parameters. Therefore, exploring the feasibility of solving this matching problem became a primary objective of this project. A positive answer would enable the decomposition of GloScope's divergence, whereas a negative answer would highlight a fundamental limitation of this post-hoc analysis strategy.

2.5 Data and Implementation

To explore how to decompose the differences captured by KL-divergence, we used an HIV-PBMC dataset for our analysis. The dataset contains peripheral blood mononuclear cell sequencing data from four patients at eight key time points, ranging from pre-infection to one year post-infection. The raw data underwent a standard preprocessing pipeline, including cell and gene filtering for quality control, log-normalization to stabilize variance, and Principal Component Analysis (PCA) for dimensionality reduction. Subsequently, Uniform Manifold Approximation and Projection (UMAP) was applied to the principal components for two-dimensional visualization.

3 Result

3.1 Baseline Results Using GloScope's Default Parameters

We first ran the GloScope analysis on the HIV-PBMC dataset using the default parameters (k-range of 1-9 for the GMM) to establish a baseline result. This provides a standard application of the method, yielding a sample-level distance matrix that quantifies the overall distributional differences between each pair of samples.

The results are visualized as a heatmap of the pairwise KL-divergences (Figure 1) and a corresponding two-dimensional projection using multidimensional scaling (MDS) (Figure 2). To investigate potential patterns, we encoded the clinical metadata directly into the visualization: each point is colored by patient and given a unique shape corresponding to its time point. Despite this detailed mapping, a key observation is the lack of a interpretable structure. Samples from the same patient (i.e., same color) do not systematically cluster together, nor do samples from the same time point (i.e., same shape) form distinct groups. This ambiguity in the baseline results motivates our deeper diagnostic analysis of the underlying GMMs.

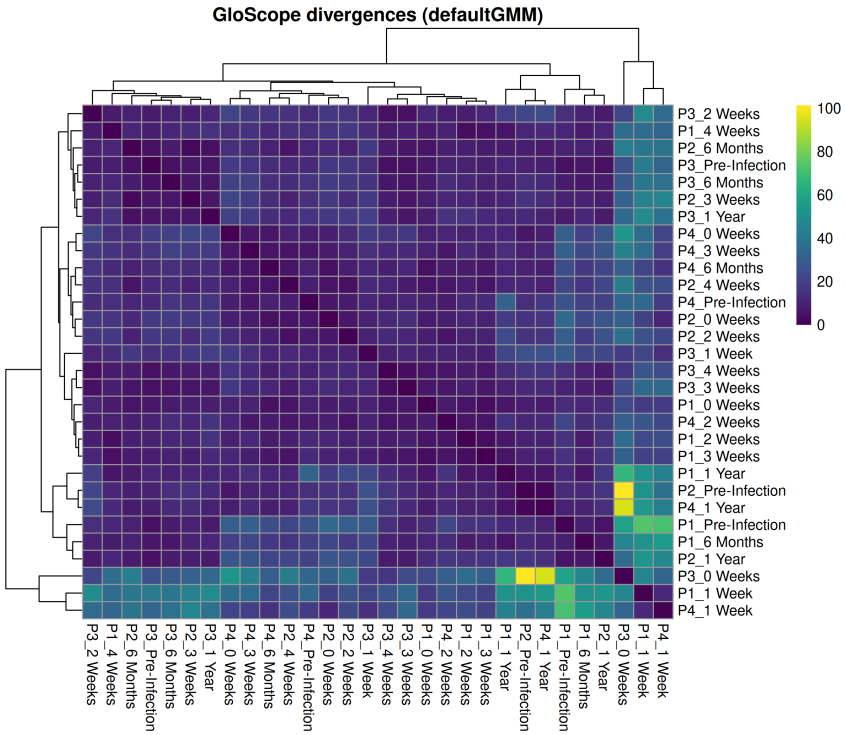

Figure 1: GloScope Divergence Heatmap. The heatmap displays the symmetrized KL-divergence matrix between all pairs of samples using default GMM parameters. Darker colors indicate greater dissimilarity between sample distributions.

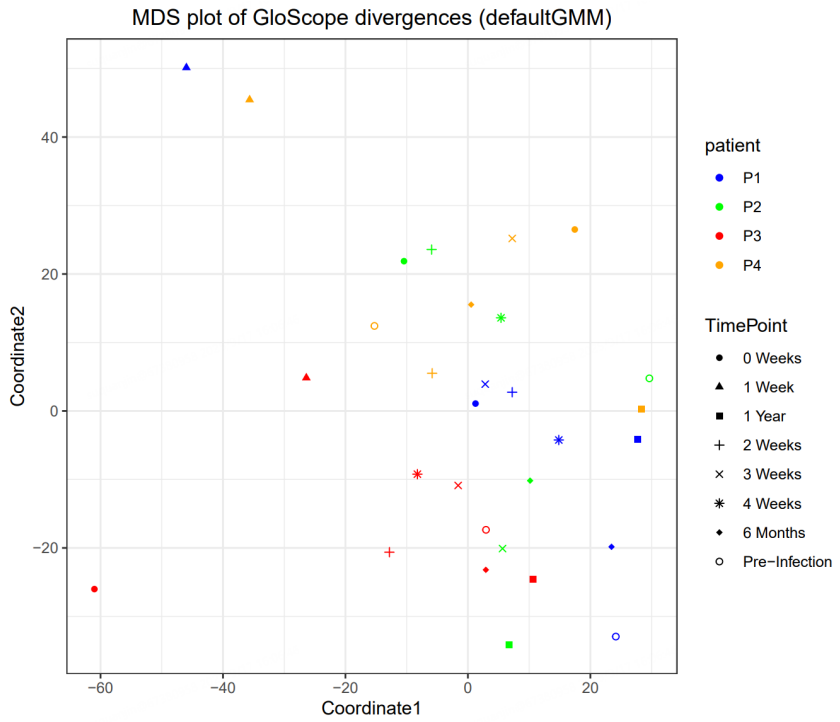

Figure 2: MDS Plot of GloScope Divergences. MDS visualization of the divergence matrix from Figure 1. Each point represents a sample, with color indicating the patient and shape representing the time point. The plot shows the global relationship between samples, but clear groupings by clinical features like infection time point are not immediately apparent.

3.2 Initial Diagnostics Reveal GMM Inconsistencies

Motivated by our goal to decompose these overall divergences, we next performed an in-depth diagnostic of the underlying GMMs that generate these distances. As a first step toward investigating the "matching problem," we diagnosed the underlying GMMs fitted with the default parameters. This initial summary, detailed in Table 1, revealed two issues that complicate cross-sample comparisons: First, we observed evidence of potential model underfitting, highlighted by a "ceiling effect" where 60% of the samples (18 out of 30) selected k=9, the maximum value in the allowed search range. This strongly suggests that the preset k-range was too restrictive to capture the true complexity of the data. Compounding this problem, the analysis showed a lack of model consistency: while 63.3% of samples selected the VVE covariance model as optimal, the remaining samples converged on other structures (EVE or VVV). This heterogeneity in model choice introduces a significant confounding variable that undermines the direct comparison of parameters across samples.

These findings motivated a more systematic parameter search to establish a consistent and adequate modeling framework before attempting to interpret the GMM components.

Table 1: GMM Parameter Selection under Default Settings (k from 1 to 9)

Optimal Covariance Model Distribution	
VVE	19 samples (63.3%)
VVV	7 samples (23.3%)
EVE	4 samples (13.3%)
Optimal k Value Distribution	
k=6	1 samples (3.3%)
k=7	4 samples (13.3%)
k=8	7 samples (23.3%)
k=9	**18 samples (60.0%)**

3.3 Expanding the Parameter Search Leads to Model Convergence

Table 2: GMM Parameter Selection under Expanded Settings (k from 6 to 20)

Optimal Covariance Model Distribution	
VVE	**30 samples (100%)**
VVV	0 samples (0%)
EVE	0 samples (0%)
Optimal k Value Distribution (Median = 14)	
k=6	1 sample
k=8	2 samples
k=9	1 sample
k=10	1 sample
k=11	1 sample
k=12	4 samples
k=13	3 samples
k=14	5 samples
k=15	5 samples
k=16	1 sample
k=18	3 samples
k=19	1 samples
k=20	2 samples

To address the issues of underfitting and inconsistency identified in our initial diagnostics, we expanded the search range for the number of components from 6 to 20. The lower bound was set to 6 because our initial analysis showed that no sample selected an optimal k value less than 6 under the default settings (Table 1). As shown in Table 2, this adjustment had two significant effects. First, the median optimal k across all samples increased from 9 to 14, confirming that the initial range was too restrictive. Second, and more importantly, this change led to a consistent choice of covariance structure: all 30 samples (100%) selected the VVE model as optimal.

The reason for this convergence is illustrated by the Bayesian Information Criterion (BIC) curves shown in Figure 3. When the search space for k is sufficiently large, the BIC value for the VVE model (purple curve) consistently surpasses those of the other models. In contrast, within the original, lower k-range, the optimal model choice was often ambiguous, leading to the inconsistency we previously observed. Achieving a uniform model choice is a critical step, as it removes a key technical confounder and provides a consistent framework for comparing GMM components across all samples.

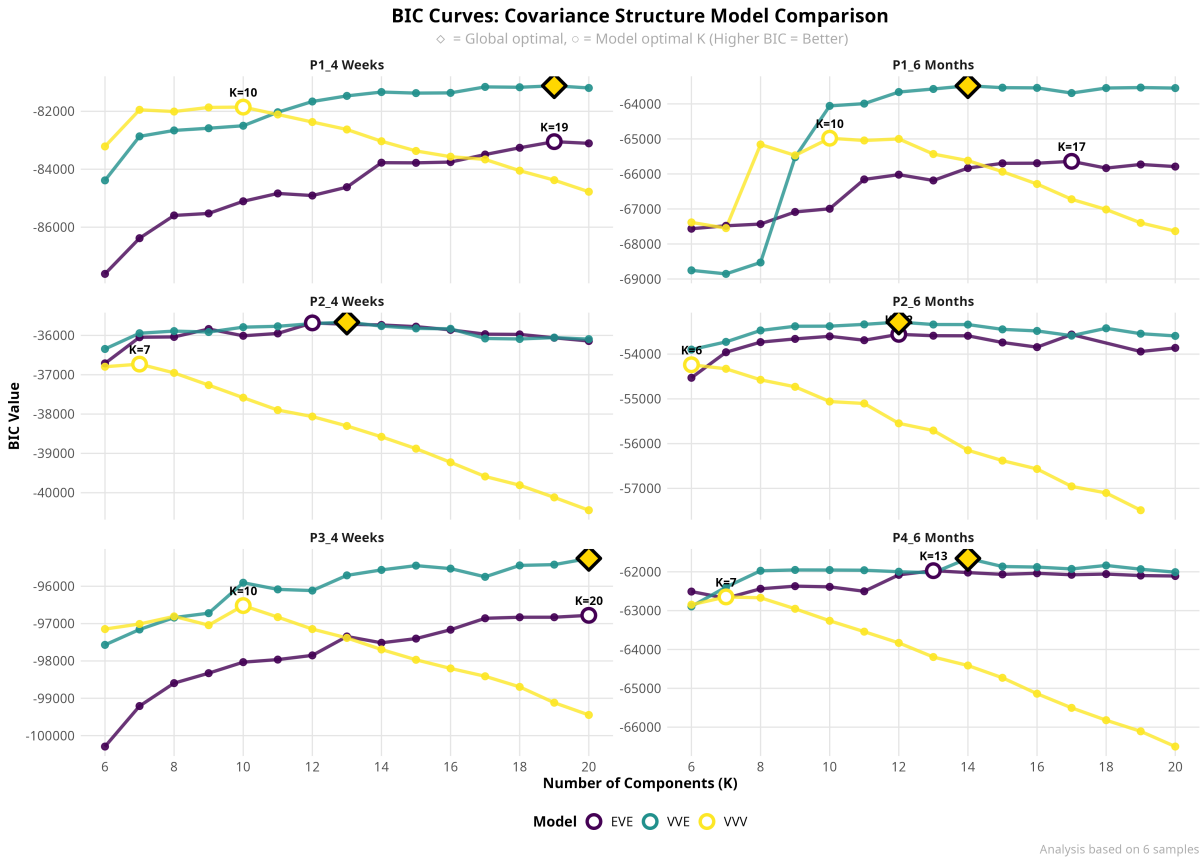

Figure 3: BIC Curves for Model Selection Across an Expanded k-Range. This plot shows the BIC values (Y-axis, higher is better) for different covariance models across the expanded range of component numbers (k, on the X-axis) for six representative samples. The symbols highlight the optimal choices: the yellow diamond identifies the global optimum (the model with the highest BIC score overall for that sample), while the white circles mark the local optima (the peak BIC score for each of the other individual model curves). The VVE model (purple line) consistently emerges as the optimal choice in the higher k-range for all samples.

3.4 GloScope's Global Divergence is Robust to GMM Parameterization

Given the significant changes in the underlying GMMs described above—namely, a higher number of components and a uniform covariance structure—we next assessed the impact of these changes on the final GloScope output. To do this, we compared the MDS plots of the sample divergence matrices generated with the default (k from 1 to 9) and the expanded (k from 6 to 20) parameter ranges (Figure 4).

Unexpectedly, the global arrangement of samples in the MDS plots remained highly consistent despite the substantial refitting of the underlying components. This finding of robustness suggests that GloScope's KL-divergence, which integrates over the entire probability density, is more sensitive to the overall distributional shape of a sample than to the specific number or configuration of its underlying GMM components. This supports our central thesis from another perspective: individual GMM components should not be over-interpreted as consistent, directly comparable biological entities. Their internal structure can be substantially altered without affecting the macroscopic comparison of distributions, suggesting they are flexible statistical constructs rather than rigid biological units.

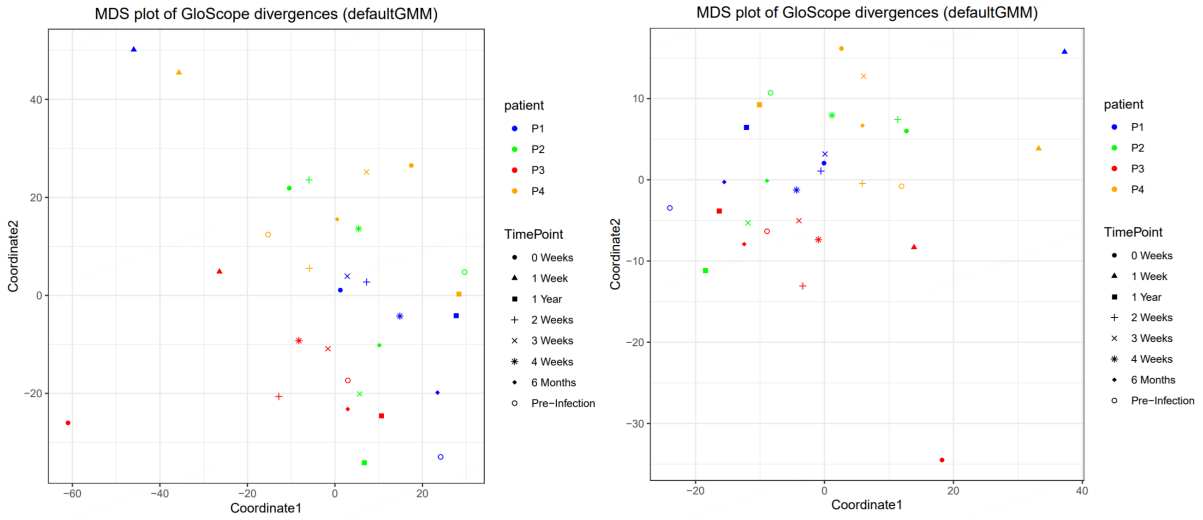

Figure 4: Comparison of MDS Plots Before and After GMM Refitting. The side-by-side plots show the MDS representation of the GloScope divergence matrix using the default k-range of 1 to 9 (left) and the expanded k-range of 6 to 20 (right). Despite a significant increase in the median number of GMM components (from 9 to 14), the overall spatial arrangement of the samples is remarkably consistent. Visual encoding is consistent with Figure 2.

3.5 Visualization Confirms the Complexity of the "Matching Problem"

To visually interrogate the severity of the "matching problem," we analyzed the GMM components from the 6-month time point across all four patients. We first visualized the GMM component centers ($\hat{\mu}_{jk}$), overlaid as stars on the cell distributions in the shared PCA space (Figure 5). To create a biological reference for comparison, we also established a set of global "cell type anchors." These anchors are defined as the centroid for each major annotated cell type, calculated by taking the mean of the PCA coordinates of all cells of that type from all samples combined. A heatmap of the Euclidean distances between all component centers and these anchors reveals the complexity of their relationships (Figure 6).

To first evaluate the best-case scenario for the matching problem, we identified a block of components-specifically `P1_C4`, `P2_C1`, and `P4_C1`-in the heatmap (Figure 7) that showed strong correspondence to the B cell anchor across multiple patients. In the PCA visualization (Figure 5), we highlighted these components as red stars ($\star$). The highlighted components from Patient 1, 2, and 4 are all precisely located within the center of their respective B cell populations (yellow dots). This demonstrates an ideal one-to-one mapping, where a single GMM component successfully represents the same biological cell type across different samples. It is also worth noting that Patient 3, which has a substantially lower cell count than the other samples, does not yield a similarly distinct B cell component, highlighting the impact of sample-specific cell composition on the GMM.

However, this ideal one-to-one mapping is the rare exception. The more common "many-to-many" relationship is clearly revealed when we examine components with an ambiguous biological identity. For instance, an examination of the full distance heatmap (Figure 6) reveals a block of components—namely `P1_C14`, `P2_C3`, and `P4_C6`—that are all relatively distant from any single cell type anchor. Their ambiguous nature is confirmed in the PCA plot (Figure 5), where they are located in regions where distinct cell types, such as B cells and T cells, overlap.

This complexity is further illustrated by the GMM's tendency to fracture a single, homo-

geneous cell population into multiple components. A clear illustration is found in Patient 4, where the dense population of monocytes (green dots) is modeled by no fewer than seven separate GMM components (black stars, Figure 5). This placement of numerous components reveals the GMM's primary goal: it is not to identify discrete biological clusters, but rather to achieve a better statistical fit to the data's overall density. This fundamental conflict between optimizing for statistical fit versus achieving a meaningful biological partitioning is the source of the "matching problem" and makes any post-hoc mapping of components infeasible.

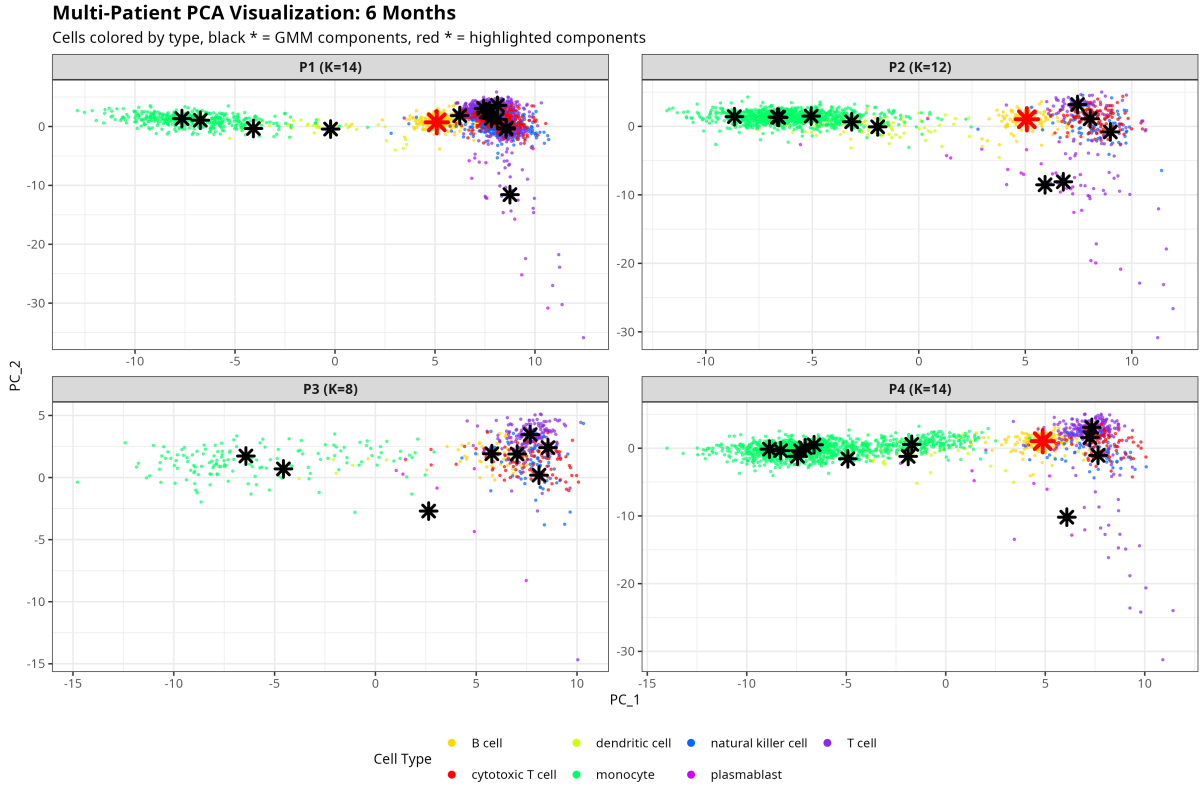

Figure 5: PCA Visualization of GMM Components and Cell Types. The plots show cells from four patients at the six-months time point. In each panel, each dot is a cell, colored by its annotated cell type. Black stars ($\star$) represent the centers (μ vectors) of the GMM components for that patient, projected onto the first two principal components. Red stars ($\star$) highlight the GMM components corresponding to the B cell population, selected to represent an **ideal case** of one-to-one mapping between a statistical component and a biological cell type across different samples.

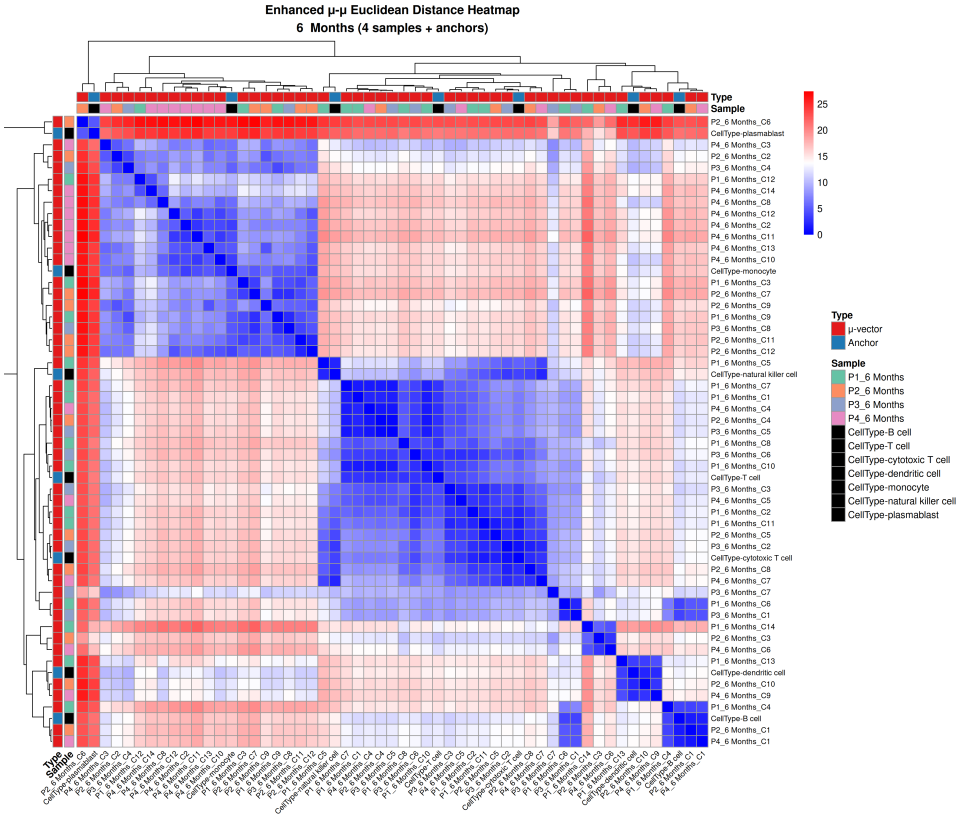

Figure 6: Heatmap of Distances Between GMM Component Centers. The heatmap shows the Euclidean distances in the full PCA space between GMM components from all four patients at six months timepoint, as well as their distances to global cell type anchors (denoted as CellType-*).

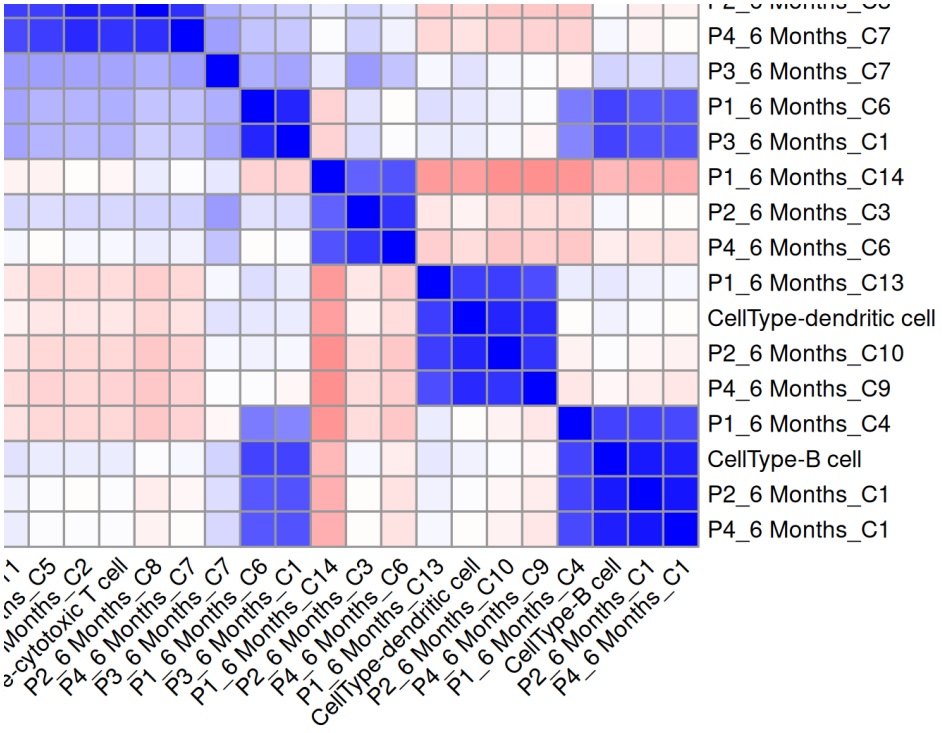

Figure 7: Magnified View Highlighting an Ideal Component Cluster. This magnified view of Figure 6 highlights a block of components corresponding to B cells from multiple patients. These components are relatively close to the global B cell anchor, representing a case of successful cross-sample matching. Their corresponding locations in PCA space are shown as red stars in Figure 5.

4 Conclusion

Our diagnostic analysis yielded two primary findings. First, GloScope's method of comparing samples via the KL-divergence of their GMM densities is remarkably robust. The final sample-level comparisons are largely insensitive to significant changes in the underlying GMM parameterization. Second, this robustness arises precisely because the individual GMM components do not function as consistent proxies for discrete biological cell types. Our investigation into the "matching problem" demonstrated that when fitted independently, GMM components exhibit a complex, many-to-many relationship with annotated cell types, and their number and location are inconsistent across samples. This confirms they are flexible statistical constructs whose primary goal is to model the overall density, not to act as consistent biological entities.

Consequently, within the current framework, using post-hoc mapping of GMM components to perform a comparison based solely on expression is not feasible. The analysis in this project not only validates the robustness of GloScope for detecting composite differences—showing its results are not overly dependent on the specific decomposition of the underlying GMM components—but also reveals, from a new perspective, its limitations in disentangling compositional and expression differences. Because the GMM components are not consistent across samples, we cannot reliably use their parameters to disentangle changes in cell composition from changes in expression. Furthermore, we found that this issue persists even when forcing all samples to share the same number of components (k), confirming that the lack of correspondence is a fundamental result of the independent fitting process rather than a simple parameter choice. This provides a clear direction for future methodological improvements. Instead of

attempting to post-process inconsistent GMM components, future models should likely aim to develop novel joint-estimation methods that can enforce component correspondence across samples directly at the modeling stage.

References

[1] Hao Wang et al. "Visualizing scRNA-Seq data at population scale with GloScope". In: Genome Biology 25.1 (2024), p. 259. DOI: 10.1186/s13059-024-03398-1. URL: https://doi.org/10.1186/s13059-024-03398-1.

[2] Ilya Korsunsky et al. "Fast, sensitive and accurate integration of single-cell data with Harmony". In: Nature Methods 16.12 (2019), pp. 1289–1296. DOI: 10.1038/s41592-019-0619-0. URL: https://doi.org/10.1038/s41592-019-0619-0.

[3] Luca Scrucca et al. "mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models". In: The R Journal 8.1 (2016), pp. 289–317. DOI: 10.32614/RJ-2016-021.