Running in circles: practical limitations for real-life application of data fission and data thinning in post-clustering differential analysis

Benjamin Hivert
Univ. Bordeaux, INSERM, INRIA, SISTM team, BPH,
Vaccine Research Institute, VRI, Hôpital Henri Mondor,
Denis Agniel
RAND Corporation,
Rodolphe Thiébaut
Univ. Bordeaux, INSERM, INRIA, SISTM team, BPH,
CHU Bordeaux, Service d'information médicale,
Vaccine Research Institute, VRI, Hôpital Henri Mondor,
and
Boris P. Hejblum
Univ. Bordeaux, INSERM, INRIA, SISTM team, BPH,
Vaccine Research Institute, VRI, Hôpital Henri Mondor

October 7, 2024

Abstract

Post-clustering inference in scRNA-seq analysis presents significant challenges in controlling Type I error during Differential Expression Analysis. Data fission, a promising approach, aims to split the data into two new independent parts, but relies on strong parametric assumptions of non-mixture distributions, which are violated in clustered data. We show that applying data fission to these mixtures requires knowledge of the clustering structure to accurately estimate component-specific scale parameters. These estimates are critical for ensuring decomposition and independence. We theoretically quantify the direct impact of the bias in estimating this scales parameters on the inflation of the Type I error rate, caused by a deviation from the independence. Since component structures are unknown in practice, we propose a heteroscedastic model with non-parametric estimators for individual scale parameters. This model uses proximity between observations to capture the effect of the underlying mixture on data dispersion. While this approach works well when

clusters are well-separated, it introduces bias when separation is weak, highlighting the difficulty of applying data fission in real-world scenarios with unknown degrees of separation.

Keywords: Unsupervised learning, Mixture Model, Post-clustering inference, Type I error, Non-parametric estimation, local variance

1 Introduction

Clustering encompasses all unsupervised statistical methods that group observations into homogeneous and separated clusters. Widely used in various application fields such as biology or genomics (Jaeger and Banks, 2023), clustering plays a significant role to uncover or summarize signals contained in different kind of multivariate data. In the context of single-cell RNA-seq (scRNA-seq) technologies which are high-throughput genomics techniques that measure gene expression at the single-cell level, providing insights into cellular heterogeneity and functional diversity within complex biological tissues, cluster analysis is the first step of the traditional pipeline for data analysis (Amezquita et al., 2020). The clustering groups cells based on their gene expression/abundance. Afterwards, the differential expression analysis, comparing gene abundance between groups, allows to identify and annotate cellular sub-populations. This leads to the identification of marker genes that could potentially serve as cell-type marker genes (Pullin and McCarthy, 2024).

However, such a two-step pipeline for post-clustering differential analysis requires using the same data twice: first to estimate the clusters, and then again to estimate the differences between them and perform significance testing – a procedure sometimes referred to as "double-dipping" (Kriegeskorte et al., 2009). In post-clustering differential analysis, it has been shown that the primary risk of double dipping is to compromise the control of the Type I error rate of otherwise well-calibrated testing procedure for traditional differential analysis, leading to false positives (Zhang et al., 2019). Fundamentally, uncertainties stemming from cluster analysis outcomes, particularly related with determining the optimal number of clusters, could create artificial differences between homogeneous groups of observations. Failure to account for the double use of the data during the analysis may lead differential analysis tools to erroneously identify those artificial differences (Hivert et al.,

2024a). Although challenges related to double dipping are increasingly studied, they remain unresolved in the context of scRNA-seq data analysis (Lähnemann et al., 2020). We described below the proposed approaches and the remaining challenges.

Leveraging the selective inference framework (Fithian et al., 2014), which involves conditioning on the clustering event during the derivation of the test statistic and the associated p-value, various methodological efforts have been proposed to address this double dipping issue in post-clustering inference (Gao et al., 2022). Zhang et al. (2019) introduced the Truncated-Normal test (TN-test), a four-step post-clustering differential analysis procedure that involves splitting the data into two parts, followed by cluster analysis on the first part of the data. Subsequently, a support vector machine (SVM) classifier is applied to the clustered data to learn the clustering structure and predict the cluster labels on the remaining data. Finally, a differential analysis is conducted between the predicted clusters using a truncated-normal test. This truncation, on either side of the hyperplane fitted by the SVM, allows to correct for the double dipping issue. However, information loss may occur as an inherent consequence of the data-splitting process. Moreover, the method involves multiple steps, introducing complexity into the overall analytical framework, and Song et al. (2023) highlighted the poor performances of the TN-test in their numerical simulations. More recently, Hivert et al. (2024a) and Chen and Gao (2023) have introduced univariate selective tests tailored to detect mean differences between two (multivariate) clusters. These methods explicitly condition on the clustering event to derive their p-value, relying on the set of all perturbed data sets that would yield the same partition when subjected to the same clustering algorithm. Hivert et al. (2024a) rely on a Monte Carlo approach to suit any clustering algorithm, which comes at the expense of extensive computational times. On the other hand, Chen and Gao (2023) explicitly describe this set specifically for $k-$means

and hierarchical clustering applied to the squared distance matrix. Bachoc et al. (2023) also introduced a selective test designed for convex clustering. All these proposed selective methods designed for post-clustering inference either require specific clustering algorithms or introduce new specific test statistics. This consequently increases their complexity and makes their application less straightforward in the context of scRNA-seq data analysis.

Leiner et al. (2023) have drawn inspiration from data splitting (which effectively addresses overfitting issues in machine learning) to break free from the selective inference framework. They propose a method called "data fission", wherein the information within each individual observation is split into two independent parts. The first part could be used for cluster analysis, and labeling observations on the second part simply by matching them to their first counterpart. Differential analysis could then be conducted on the remaining information, *i.e.* the second part. However, in the context of post-clustering differential analysis, it is imperative that the two parts be independent, as each analysis (cluster analysis and differential analysis) must be performed independently to effectively prevent double dipping and the associated inflation of Type I error. Data fission requires strong parametric distributional assumptions, with only the Gaussian and Poisson (Neufeld et al., 2024) distributions ensuring independence between the two fissioned parts. Expanding upon this concept, Neufeld et al. (2023a) have generalized the fission process by introducing "data thinning". Building on the same foundational idea, they not only succeed in developing a process capable of decomposing data into more than two parts, but also broaden the spectrum of distributions where independence between each part is provided. This includes the negative binomial distribution, widely used when modeling RNA-seq data.

Data fission and data thinning have emerged as promising alternatives to selective inference for post-clustering inference due to their compatibility with various clustering

methods and differential analysis tests. However, they present some inherent limitations that make them difficult to apply for post-clustering inference. First, these methods lack results and justification when applied to mixture distributions, which are commonly used to model data with a clustered structure (Macqueen, 1967). Consequently, the absence of such results inherently assumes a global null hypothesis of no clusters in the data when applying data fission or data thinning. Additionally, these methods assume prior knowledge of parameters (e.g., variance for the Gaussian distribution or overdispersion for the negative binomial distribution). Although robust estimators for these parameters could theoretically ensure the validity of the method, this further adds complexity for clustered data where each cluster has a different parameter value. In the absence of knowledge about the data structure, specifically the clusters, the only justifiable estimator is the full-sample one (*i.e.* computed using all the observations regardless of their mixture component) that fails to correctly estimate the intra-component parameter value.

In this article, we demonstrate that these approaches are not practical for real-world applications. Performing cluster analysis inherently assumes that the data originate from mixture models, contradicting the parametric assumptions of data fission and data thinning, which cannot decompose such distributions. Even if it is possible to move beyond the framework of mixture models by modeling each observation as a realization from distinct random variables with their own parameter values, accurately estimating scale parameters remains challenging without knowledge of the mixture. Specifically, we establish a link between the bias in estimating the variance parameter in the Gaussian distribution and the expected Type I error rate of the two-sample t-test (Welch, 1947), underscoring the critical importance of a robust estimator. We employ a non-parametric method for estimating local variance in the Gaussian setting to try to use proximity between observations

as a proxy of the unknown underlying mixture. However the poor performances of this approach demonstrates that accurate variance estimation, and hence control of the Type I error rate, remains challenging without knowing the mixture, *i.e.* the clustering, prior to decomposition.

2 Methods

In the following, capital letters represent random variables with a probability distribution function denoted as p, x_i denotes a set of n realizations of X, and matrices and vectors are bolden.

2.1 Data fission and data thinning

Data fission and data thinning are methods designed to decompose a random variable into two new independent random variables that can ensure the post-clustering inference validity.

Let $\boldsymbol{X}$ be a random variable with a known distribution. Both data fission (Leiner et al., 2023) and its generalization, data thinning (Neufeld et al., 2023a), aim to decompose the random variable $\boldsymbol{X}$ into two (or more in data thinning) new random variables $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$. These new variables are designed to i) retain information from the original variable $\boldsymbol{X}$, and ii) be independent. Of note, data fission can generate pairs of $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ that are not independent, but her we will focus on the independent cases only. The balance between the amount of information from $\boldsymbol{X}$ kept in either $\boldsymbol{X}^{(1)}$ or $\boldsymbol{X}^{(2)}$ is tuned by a hyperparameter τ. Such a decomposition can be performed for various probability distributions of $\boldsymbol{X}$. For data fission, Leiner et al. (2023) identified only two distributions, the Gaussian and the Poisson, that satisfy the independence requirement. In contrast, Neufeld

et al. (2023a) established a comprehensive decomposition that ensures independence for all convolution-closed distributions. This encompasses Gaussian, Poisson (Neufeld et al., 2024) and Negative Binomial (Neufeld et al., 2023b) distributions. Decompositions for the

Distribution of $\boldsymbol{X}$	$\boldsymbol{\tau}$	**Data fission**	**Data thinning**
$\mathcal{P}(\lambda)$	$\tau \in [0,1]$	$Z \sim \text{Binom}(X, \tau)$ $X^{(1)} = Z$ $X^{(2)} = X - Z$	$X^{(1)} \,\vert\, X = x \sim \text{Binom}(x, \tau)$ $X^{(2)} = X - X^{(1)}$
$\mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_{p\times p})$	$\tau \in]0, +\infty)$ $\tau_2 \in]0,1[$	$\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}_p, \boldsymbol{\Sigma}_{p\times p})$ $\boldsymbol{X}^{(1)} = \boldsymbol{X} + \tau \boldsymbol{Z}$ $\boldsymbol{X}^{(2)} = \boldsymbol{X} - \frac{1}{\tau}\boldsymbol{Z}$	$\boldsymbol{X}^{(1)} \,\vert\, \boldsymbol{X} = \boldsymbol{x} \sim \mathcal{N}(\tau_2 \boldsymbol{x}, \tau_2(1-\tau_2)\boldsymbol{\Sigma}_{p\times p})$ $\boldsymbol{X}^{(2)} = \boldsymbol{X} - \boldsymbol{X}^{(1)}$
$\text{NegBin}(\mu, \theta)$	$\tau \in [0,1]$	No fission	$X^{(1)} \,\vert\, X = x \sim \text{BetaBin}(x, \tau\theta, (1-\tau)\theta)$ $X^{(2)} = X - X^{(1)}$

Table 1: Data fission and data thinning decompositions for three usual distributions: Poisson, Gaussian and Negative Binomial.

three distributions into two independent random variables $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, are detailed in Table 1. For proofs of independence, please refer to the Section 1 of the Supplementary Materials for the Gaussian case, and to Neufeld et al. (2024) and Neufeld et al. (2023b) for the Poisson and negative binomial distribution, respectively. Both Gaussian and negative binomial data fissions/thinnings require knowledge of scale parameters (namely $\boldsymbol{\Sigma}$ or θ) for practical feasibility, as highlighted in Table 1. Theoretical guarantees, and especially the independence between $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, are based on using the true values of these parameters. Yet, in real-life applications these are unknown and need to be estimated, most likely from the data as described in Neufeld et al. (2023b).

2.2 Limits in practical application of data fission and data thinning

Data fission and data thinning methods face a circular challenge when applied for post-clustering inference to mixture distributions used to describe clustered data. They require accurate estimates of intra-component parameters (like the variance), which depend on knowing the true cluster – but estimating the clusters is the whole point of the analysis in the first place, and true clusters are never known a priori.

2.2.1 Mixture distributions

Neufeld et al. (2024) and Neufeld et al. (2023b) have proposed the application of data thinning for post-clustering inference. Existing decompositions are currently limited to non-mixture distributions. Unfortunately, these distributions can only describe a global null hypothesis, assuming the complete absence of separated clusters within the data. In scenarios with true clusters, a more appropriate modeling approach involves the use of mixture models (Bouveyron et al., 2019) – each component of the mixture representing a distinct cluster. Given n i.i.d. realizations of random variables following a mixture of distributions, the density function for an observation $\mathbf{x}_i$ is: $p(\mathbf{x}_i) = \sum_{k=1}^{K} \pi_k f(\mathbf{x}_i|\boldsymbol{\theta}_k)$, where π_k is the probability that an observation was generated by the k^{th} component and $f(.|\boldsymbol{\theta}_k)$ is the density of the k^{th} component with its specific parameters $\boldsymbol{\theta}_k$ (for simplicity, we only consider the case where all components belong to the same parametric distribution with density f). In this mixture setting, where homogeneity holds solely at the component level, data fission and data thinning can only be applied within each individual component.

Data fission and data thinning therefore create a circular situation, as illustrated in the schematic Figure 1. To be applied for post-clustering inference, these methods require

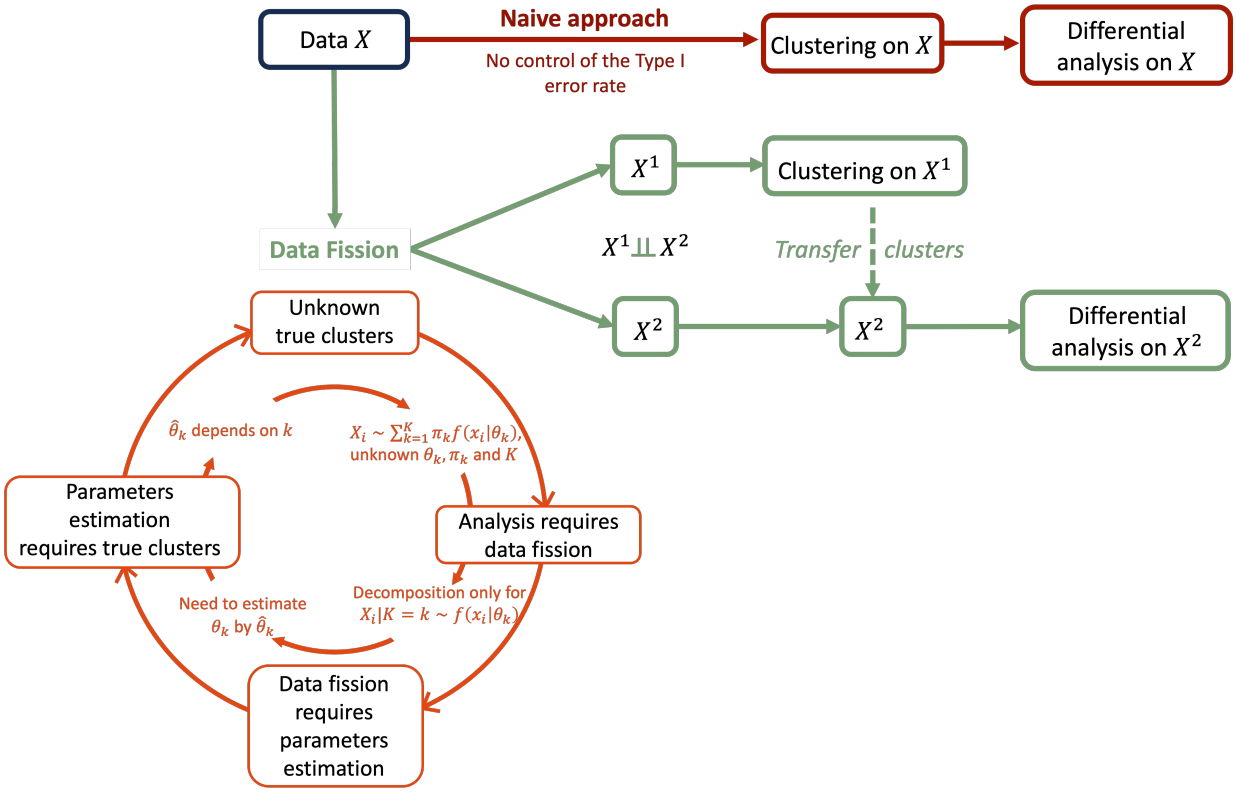

Figure 1: Schematic view illustrating the circularity induced by data fission for post-clustering differential analysis

knowledge of the intra-component parameters $\boldsymbol{\theta}_k$ which, in turn, depend themselves on the components that we would estimate. An effective approach could be to estimate a global parameter $\hat{\boldsymbol{\theta}}$ based on all observations. However, this would make the critical assumption that the parameter value is the same across all components, *i.e.* $\boldsymbol{\theta} = \boldsymbol{\theta}_k$; this means that the data are distributed according to the same distribution globally, regardless of their component:

$$p(\mathbf{x}_i) = f(\mathbf{x}_i|\boldsymbol{\theta}_k) = f(\mathbf{x}_i|\boldsymbol{\theta}). \tag{1}$$

This corresponds to the global null of no cluster. Yun and Foygel Barber (2023) highlighted similar challenges for existing post-clustering selective tests.

2.2.2 Scale parameter prior knowledge and estimation

We will restrict our analysis to the Gaussian setting, that is considering:

$$f\left(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_g|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\right)$$

with $k = 1, \ldots, K$, $K \geq 1$ the number of components in the mixture, and $\boldsymbol{\mu}_k \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p\times p}$ respectively the mean vector and the covariance matrix of the k^{th} components. Much of the results apply to the negative binomial case, with the overdispersion parameter being analogous to the variance parameter in the Gaussian case. In this setting, the challenge lies in estimating $\boldsymbol{\Sigma}_k$ for data fission or data thinning.

Figure 2 provides a comprehensive overview of the challenges associated with variance estimation in data fission. Panel **A** presents an illustrative example involving 300 realizations of a multivariate Gaussian distribution ($K = 1$) with a mean vector $\boldsymbol{\mu} = \left(\begin{smallmatrix}0\\0\end{smallmatrix}\right)$ and a covariance matrix $\boldsymbol{\Sigma} = \left(\begin{smallmatrix}4 & 0\\0 & 1\end{smallmatrix}\right)$. The $k-$means algorithm was applied to this dataset, resulting in the estimation of two clusters, C_1 and C_2. Data fission performance was evaluated when using different estimations of $\boldsymbol{\Sigma}$. First, we considered the true intra-components covariance matrix $\boldsymbol{\Sigma}_k$ (see section 2.3.1 for a proposed approach to data fission assuming several mixture component). We also considered the overall sample covariance matrix $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{x}_i - \overline{\boldsymbol{X}}\right)\left(\boldsymbol{x}_i - \overline{\boldsymbol{X}}\right)^t$ where $\overline{\boldsymbol{X}}$ is the sample mean vector. We finally use the $k-$means results to compute an intra-cluster covariance matrix defined as: $\widehat{\boldsymbol{\Sigma}}_{\hat{k}} = \frac{1}{|C_k|-1}\sum_{i=1\in C_k}\left(\boldsymbol{x}_i - \overline{\boldsymbol{X}}_{C_k}\right)\left(\boldsymbol{x}_i - \overline{\boldsymbol{X}}_{C_k}\right)^t$ where $\overline{\boldsymbol{X}}_{C_k}$ is the sample mean vector of the cluster C_k. Since both clusters originate from the same component, there is no inherent differences between them, implying that the t-test p-values should exhibit a uniform distribution. Panel **B** presents a QQ-plot illustrating the resulting p-values against the Uniform distribution for the test on X_1 in 1,000 replications of the experiment. In this scenario where the mixture has only one component ($K = 1$), the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$

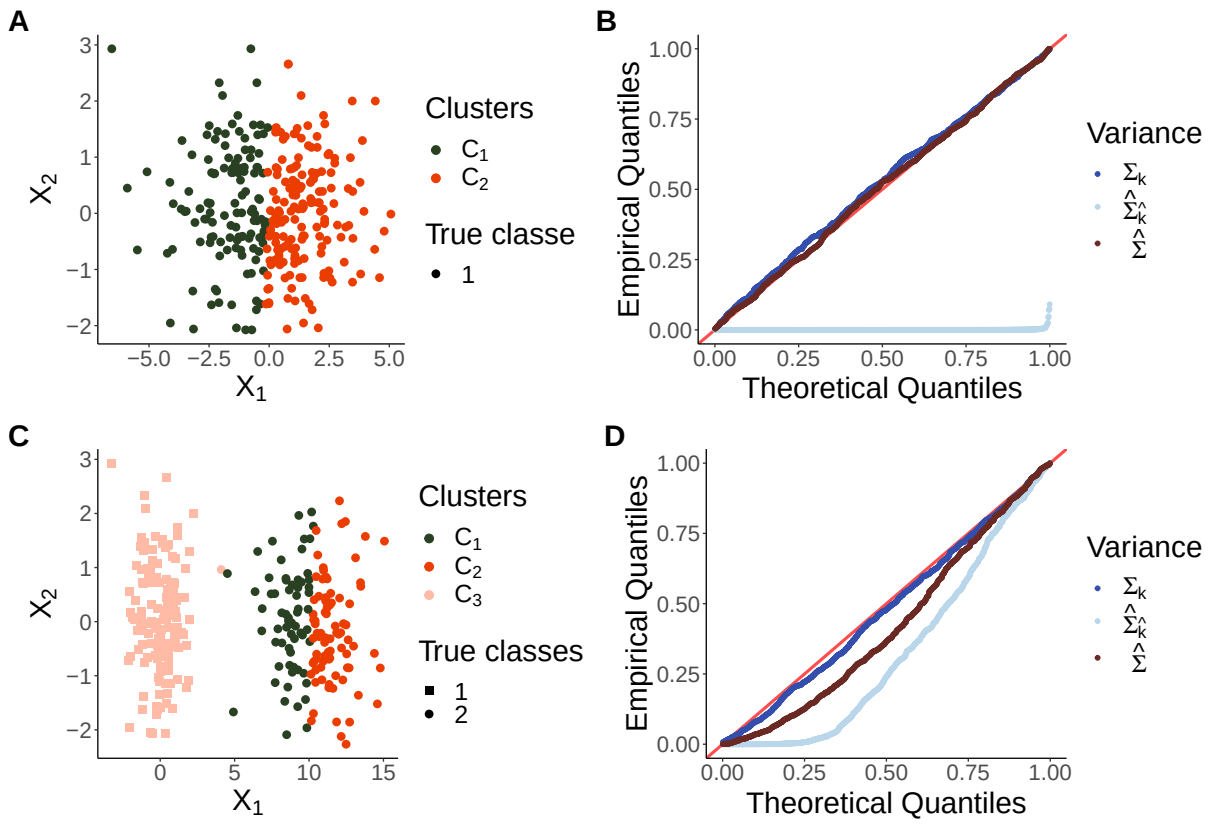

Figure 2: **Toy example illustrating the impact of variance estimation on data fission p-values Panel A:** A two-dimensional Gaussian distribution incorrectly clustered into 2 clusters. **Panel B:** QQ-plot of the t-test p-values for the comparison between the two estimated clusters across 1,000 simulations when data fission is performed with 3 different variance estimators. **Panel C:** Extension of the problem to a two-component, two-dimensional Gaussian mixture incorrectly clustered into 3 clusters. C_1 and C_2 originate from the same component, which is erroneously split into two. Panel **D**: t-test p-values for the comparison between C_1 and C_2 over 1,000 data simulations using the same three variance estimators for the data fission.

provides an unbiased estimate of the true $\mathbf{\Sigma}$, resulting in uniformly distributed p-values. However, when considering the intra-cluster covariance matrices $\widehat{\mathbf{\Sigma}}_{\hat{k}}$ (for $k = 1, 2$) derived from the $k-$means results, the p-values no longer exhibit a uniform distribution. In this case, the estimated matrices $\widehat{\mathbf{\Sigma}}_{\hat{k}}$ for each cluster drastically underestimate the true covariance matrix $\mathbf{\Sigma}$, compromising the independence between $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$. This deviation from independence leads to spuriously concluding the estimated clusters are truly different from one another.

Panel **C** introduces a scenario with two true clusters generated using a mixture of two Gaussian distributions ($K = 2$). The $k-$means identifies 3 clusters (for illustration purposes), incorrectly splitting one mixture component into clusters C_1 and C_2. Data fission performance was again compared when the decomposition is performed using the same 3 covariance estimators as in the previous scenario. Panel **D** presents the resulting p-values for the t-test on C_1 and C_2 on X_1 over 1,000 replications of the analysis. Efficient Type I error control is achievable only by considering the true intra-components covariance matrix. Both global and estimated intra-cluster covariances are biased estimators, leading to correlations between $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, compromising control of the Type I error rate. This underscores the intricate challenges associated with covariance estimation for data fission in practical scenarios, mainly due to the misspecification of the generative model for data fission arising when genuine clusters exist within the dataset.

2.3 Practical solutions

To address the circular situation of data fission and thinning in Gaussian mixture models, we propose modeling each observation as a realization of its own Gaussian distribution. This approach bypasses the need for prior knowledge of the true data structure, allowing

for individual-level data fission and thinning. Nonetheless, the accurate estimation of individual variance parameters remains a critical challenge for practical feasibility.

2.3.1 Individual fission (or thinning) for mixture models

In Gaussian mixture models, parameters are typically component-specific, meaning that they are assumed to be shared across all individuals within each component. As explained above and in Figure 1, this assumption poses a circular challenge as it requires knowledge of the true data structure to be able to accurately estimate the component-specific covariance matrices that are then needed to perform data fission or data thinning. To address this limitation, we propose an alternative approach where each observation is modeled as a realization of its own Gaussian distribution, *i.e.* $p_i(\boldsymbol{x}_i) = f\left(\boldsymbol{x_i}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)$. Thus, observations are no longer assumed to be identically distributed. Consequently, the covariance matrix $\boldsymbol{\Sigma}_i$ is no longer specific to the components but instead to the individual observations. Despite this individual-level definition, two individuals drawn from the same (unknown) component are expected to have very close variance parameters. By bypassing the components in the definition of $\boldsymbol{\Sigma}_i$, this modeling strategy theoretically encompasses both the global null ($K = 1$) and the mixture ($K \geq 1$) settings, and opens up individual-level data fission and data thinning. As highlighted above, variance estimation remains crucial for the practical feasibility of these methods. This new modeling assumes individual variances that still need to be known (ideally), or precisely estimated in real-life settings.

2.3.2 Non-parametric local variance estimation

To estimate $\boldsymbol{\Sigma}_i$, we propose to use weighted variances where the weights are determined through non-parametric kernel smoothing. The underlying principle behind this approach is that, despite the individual-specific nature of the variance, two observations within the

same component of the mixture (i.e., two neighbors) should display similar variance patterns. The non-parametric weights assigned to each individual reflect their contribution to the estimation of the variance for the i^{th} observation, effectively capturing the proximity between individuals. First, let's assume that we are under the univariate setting, that is: $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, n$. We define $\widehat{\sigma^2}_i$, the resulting estimate of σ_i^2, as:

$$\widehat{\sigma^2}_i = \frac{\sum_{j=1}^{n} w_{ij} \left(x_i - \widehat{m}_i\right)^2}{\left(\sum_{j=1}^{n} w_{ij}\right) - 1} \tag{2}$$

where w_{ij} are individual-weights and $\widehat{m}_i = \sum_{j=1} w_{ij} x_i \Big/ \sum_{j=1}^{n} w_{ij}$ are individual-specific weighted means. Ideally, w_{ij} should be zero (or very small) for all the observations x_j that are not in the same component as x_i.

The catch of the variance computation in (2) lies in determining each individual weight w_{ij}. Since it is essential for w_{ij} to appropriately capture the proximity between observations, we opted for a kernel-based definition: $w_{ij} = K(x_i - x_j)$, where $K(u) = \frac{1}{\sqrt{2\pi}h} e^{-\frac{u^2}{2h^2}}$ represents the Gaussian kernel providing the proximity measurement between x_i and x_j. This ensures that individual weights reflect the local relationships within the data. This kernel choice focuses on nearby points, emphasizing their influence on the weighted variance estimate. The parameter h in the definition of K serves as the bandwidth parameter, controlling the width of the kernel and, consequently, the neighborhood around each observation x_i that contributes to its weighted variance estimate. A smaller h results in more localized estimations, emphasizing nearby points, but can underestimate the variances. Conversely, a larger h includes a broader range of observations, and an excessively large value might consider almost all observations in the variance estimation, yielding an estimator close to the full sample one $\widehat{\mathbf{\Sigma}}$. Therefore, an optimal choice for h would in-

volve considering only the observations in the same components of the underlying mixture. However, achieving this ideal scenario is impractical as it is equivalent to knowing the true components of the mixture.

Bandwidth calibration is a crucial step in any kernel method (Heidenreich et al., 2013). We propose to use individual-specific bandwidth $h_i = h(x_i)$ to reduce the bias in kernel density estimation. To accomplish this, we first estimate the changepoint, i^* in the spread of distances from observation x_i to all other observations. This changepoint delineates the change in components of the mixture: observations with distances preceding this changepoint are deemed part of the same component as x_i, whereas those following it are considered part of different components. Subsequently, we define $h_i = |x_i - x_{i^*}|$ as the distance between x_i and the observation x_{i^*} that is furthest from x_i before the break in the mixture. Thus, h_i is determined in such a way that the individual bandwidth accommodates only the observations in the same component as x_i, ensuring its size is sufficient to encompass all observations of the component (Chacón and Duong, 2020).

3 Results

3.1 Quantification of Type I error rate as a function of the bias in the variance estimation

Independence between $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ is guaranteed solely when the true variance is used for decomposition. Substituting the true variance $\boldsymbol{\Sigma}_i$ with an estimate $\widehat{\boldsymbol{\Sigma}}_i$ introduces correlations between these new random variables as demonstrated by Proposition 1 for data fission (a proof of this proposition could be found in Section 2 of the Supplementary Materials) and by Proposition 10 of Neufeld et al. (2023a).

Proposition 1 *Let* $\boldsymbol{X} \sim \mathcal{N}\left(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_{p\times p}\right)$ *be a Gaussian random variable. Suppose we apply data fission on* $\boldsymbol{X}$ *as described in Table 1 but using* $\boldsymbol{Z} \sim \mathcal{N}\left(\boldsymbol{0}, \widehat{\boldsymbol{\Sigma}}\right)$*, where* $\widehat{\boldsymbol{\Sigma}}$ *is an estimate of the* $\boldsymbol{\Sigma}$*. Then,* $\text{Cov}\left(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}\right) = \boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}$.

Proposition 1 indicates that if $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ are not independent, as with a biased estimator of $\boldsymbol{\Sigma}$ that induces significant covariance, a split of a single component into two estimated clusters in $\boldsymbol{X}^{(1)}$ is easily transferred to $\boldsymbol{X}^{(2)}$ and results in false positives during the inference step (even if the latter is carried out on $\boldsymbol{X}^{(2)}$).

To further describe the repercussions of variance estimation in the context of data fission, we derived an analytical expression for the Type I error rate of the t-test as a function of the bias in estimating this parameter. Let $X_1, \ldots, X_n$ be n independent and identically distributed random variables such that, for all $i = 1, \ldots, n$, $X_i \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$. Here, n represents the sample size. Recall that the sample $X_1, \ldots, X_n$ is normally distributed: it contains no real clusters, and therefore, no actual difference in means exists between subgroups of observations that is not a consequence of the clustering. Let $Z_1, \ldots, Z_n \overset{i.i.d}{\sim} \mathcal{N}(0, b^2)$ and $\tau \in,]0, +\infty[$. Our aim is to perform data fission of each X_i, using the $X_i^{(1)}$ for $k-$means clustering (with $K = 2$) and the $X_i^{(2)}$ for differential testing between the two inferred clusters. Here, b^2 represents any value used as a plug-in for the variance of the X_i, and in particular, b^2 may be an estimate obtained for σ^2. Given the generation process of the X_i, it is established that regardless of the clustering on the $X_i^{(1)}$, there should be no mean difference between the estimated clusters on $X_i^{(2)}$ as long as independence is achieved. Let C_1 and C_2 be the two estimated clusters on the $X_i^{(1)}$ with the same intra-cluster variance, which is a reasonable hypothesis with $k-$means clustering as explained in Section 3 of the Supplementary Materials. Since we are under the null hypothesis of no mean difference between the clusters, the T statistic for the t-test between C_1 and C_2 using the $X_i^{(2)}$ is

given by:

$$\frac{\overline{X^{(2)}_{C_1}} - \overline{X^{(2)}_{C_2}}}{\sqrt{\frac{4s^2\left(X^{(2)}\right)}{n}}} \qquad \text{where} \qquad \overline{X^{(2)}_{C_j}} = \frac{1}{|C_k|}\sum_{i \in C_k} X_i^{(2)} \tag{3}$$

Here, $s^2\left(X^{(2)}\right)$ is their shared intra-cluster variance computed using the $X_i^{(2)}$. It can be demonstrated that:

$$T \overset{\mathcal{L}}{\sim} \mathcal{N}\left(\frac{\rho\sqrt{n}}{\sqrt{\frac{\pi}{2} - \rho^2}}, 1\right)$$

where $\rho = \text{Cor}\left(X_i^{(1)}, X_i^{(2)}\right) = \frac{\left(\sigma^2 - b^2\right)}{\sqrt{(\sigma^2 + \tau^2 b^2) \times \left(\sigma^2 + \frac{1}{\tau^2} b^2\right)}}$. Section 3 of the Supplementary Materials gives details on the derivation of this test statistic and its distribution. The associated Type I error rate for this test is given by $1 - F(q_{\alpha/2}) + F(-q_{\alpha/2})$, where F is the cumulative distribution function of $\mathcal{N}\left(\frac{\rho\sqrt{n}}{\sqrt{\frac{\pi}{2} - \rho^2}}, 1\right)$ and $q_{\alpha/2}$ is the quantile of a standard Gaussian distribution $\mathcal{N}(0, 1)$.

We validated this theoretical result through numerical simulations, and conducted a detailed exploration of the influence of variance values and sample size on the resulting Type I error rate. We generated n realizations of a Gaussian random variable with a mean $\mu = 0$ and variances σ^2, with 1,000 Monte Carlo repetitions. Subsequently, we used data fission with varying values of $b^2 = \widehat{\sigma^2}$, obtaining $X^{(1)}$ for $k-$means clustering with $K = 2$ and $X^{(2)}$ for testing mean differences between the two estimated clusters. Initially, we examine the impact of the original true variance σ^2 by considering σ^2 values of $\{0.01, 0.25, 1, 4\}$ for a fixed sample size of $n = 100$. Figure 3A, illustrates the relationship between the bias in estimating σ^2 and the Type I error rate, demonstrating a consistent agreement with the derived theoretical error rate.

We further documented the behavior of the Type I error for a fixed $\sigma^2 = 1$ with varying sample sizes $n \in \{50, 100, 200, 100, 500, 1{,}000\}$ in Figure 3B, which shows the expected impact of the sample size on the Type I error rate. These findings collectively underscore the critical importance of accurately estimating the variance to achieve a well-calibrated

Type I error rate with data fission.

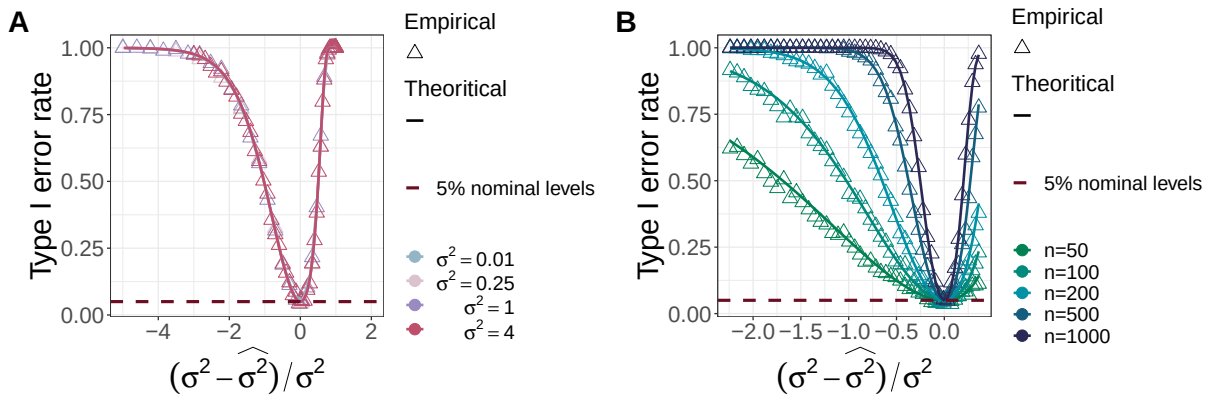

Figure 3: **Impact of variance estimation on Type I error rate in data fission**. Panel **A**: Evolution of the estimated Type I error rate in data fission as a function of the relative bias and the original variance of the data. Panel **B**: Evolution of the estimated Type I error rate in data fission as a function of the relative bias and the sample size.

3.2 Performances of the local variance estimator

We conducted simulation studies to assess the performance of the non-parametric variance estimator defined in (2). Generating univariate data akin to the motivating example in Figure 2 with $n = 100$ realizations from a two-component univariate Gaussian mixture: $0.5\mathcal{N}(0, \sigma^2) + 0.5\mathcal{N}(\delta, \sigma^2)$. We explored a range of ratio δ/σ values from 0 (*i.e.* no separation between the components, representing the global null of no clusters in the data) to 100 (indicating an extreme separation). Different values of σ^2 were considered: $\sigma^2 \in \{0.01, 1, 4\}$. For each pair (δ, σ^2) defining the mixture, variance was estimated from the data using our proposed weighted local variance, and the result was used for individual data fission. We then considered only the observations coming from the first component of the mixture for $k-$means clustering on the corresponding $X^{(1)}$ with $K = 2$. This resulted

in one true component splitting into two incorrect clusters. We then tested the mean differences between those two clusters using the t-test on $X^{(2)}$. This scenario was replicated 1,000 times, and we computed the empirical Type I error rate at the $\alpha = 5\%$ level.

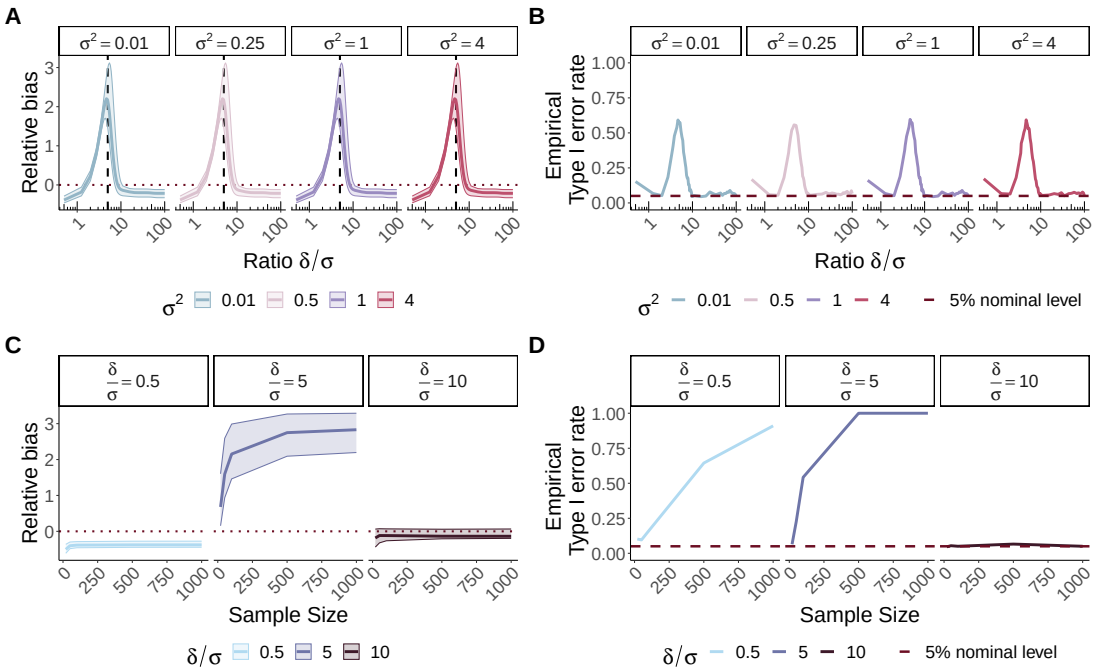

Figure 4: **Performance evaluation of the non-parametric variance estimator in a simulated univariate setting.** All empirical results were obtained through 1,000 simulations of the data. Panel **A**: Median relative bias, defined as $\left(\widehat{\sigma^2} - \sigma^2\right)/\sigma^2$ and its associated inter-quantile range, against the signal-to-noise ratio δ/σ. Panel **B**: Type I error rate at the $\alpha = 5\%$ level against δ/σ. Panel **C**: Median relative bias and its associated inter-quantile range as a function of the sample size for three degrees of separation between the two components (informed by the ratio δ/σ). Panel **D**: Type I error rate at the $\alpha = 5\%$ level as a function of sample size.

Figure 4A shows that local variances are underestimated (manifested as a negative relative bias) until the signal vs. noise ratio δ/σ reaches approximately 3 – a threshold value for separation in a Gaussian mixture model previously reported in the literature (Siffer et al.,

2018; Hivert et al., 2024a). Consequently, there are no clear change-points in the spread of distances for individual observations and all observations should contribute significantly to the variance calculation. However, this is not the case here (with a bandwidth h_i that is too small), leading to underestimated local variances. As the ratio increases within the range $3.5 \leq \delta/\sigma \leq 10$, component separation becomes clearer but the identification of change-points remains challenging. Change-points detection is intricate for observations in the tails in-between the components, leading to non-zero weight for observations in both components. Consequently, an overestimation of the local variances (positive relative biases) is observed together with an increase in their associated inter-quantile range. Finally, for $\delta/\sigma > 10$, sufficient component separation ensures consistent local variance estimation, highlighting the critical importance of clear-cut component separation for accurate variance estimation. So, as long as observations are well separated, our methodology outlined in 2.3.2 is able to provide accurate variance estimates. In Figure 4B, the Type I error rate at 5% remains well calibrated for those values of δ/σ that ensure robust variance estimation (i.e. $\delta/\sigma > 10$), illustrating the reliable testing performance with good component separations. Figure 4C demonstrates how increasing the sample size does not impact the performance of the non-parametric local variance estimations for three representative values of the ratio $\delta/\sigma \in \{0.5, 5, 10\}$, pointing towards component separation as the main driver of testing performance. The corresponding Type I error rates depicted in Figure 4D thus align with the observed relative bias results, underscoring the critical role of accurate estimation of local variances to ensure Type I error control.

3.3 Application to single-cell RNA-seq data analysis

Single-cell RNA-seq (scRNA-seq) data analysis pipelines often involve an initial cluster analysis followed by differential analysis to identify marker genes and annotate cell populations based on gene expression. Given their overdispersed count nature, the negative binomial distribution is favored over the Gaussian distribution to model scRNA-seq data. Unfortunately, the negative binomial distribution also raises challenges for data thinning. The overdispersion parameter θ plays a role similar to the variance parameter in the Gaussian distribution, as it is considered known. Therefore, the quality of its estimation is directly impacts $\text{Cov}\left(X^{(1)}, X^{(2)}\right) = \tau(1-\tau)\frac{\mu^2}{\theta}\left(1 - \frac{\theta+1}{\widehat{\theta}+1}\right)$, where $\widehat{\theta}$ is an estimated of θ (Neufeld et al., 2023a). Also, for negative binomial mixtures, data thinning decomposition is once again feasible only at the component level. As the overdispersion parameter is component-specific (Li et al., 2018), providing an estimator that ensures independence is here also a harduous and circular task, given that the components themselves are again unknown and thus require estimation through data thinning.

We further illustrate the necessity of applying intra-component data thinning (with the associated intra-components overdispersion) to ensure Type I error control in the post-clustering inference setting with numerical simulations. We generated $n = 100$ observations from a two-component negative binomial mixture: $0.5\text{NegBin}(\mu_1, \theta_1) + 0.5\text{NegBin}(\mu_2, \theta_2)$ with component parameters $(\mu_1, \theta_1) = (5, 5)$ and $(\mu_2, \theta_2) = (60, 40)$. We conducted similar post-clustering inference as in Figure 2. In Figure 5A, applying the $k-$means with $K = 3$ clusters reveals that the first mixture component is erroneously split into 2 clusters (labeled C_1 and C_3). We then assessed the Type I error rate associated with the Wilcoxon test between these incorrect clusters when applying data thinning with various overdispersion estimators. First, we applied intra-component data thinning using oracle estimates $\widetilde{\theta}_k$,

$k = 1, 2$, representing true intra-component overdispersion parameters (infeasible in real-life application where the true cluster structure is unknown). We compared those results with two alternatives that are feasible in practice: i) applying intra-cluster data thinning based on the $k-$means results from Figure 5A with their associated $\widehat{\theta}_{\hat{k}}$, $\hat{k} = 1, 2, 3$, and global data thinning with its associated $\widehat{\theta}$. Of note, all the overdispersion estimations were performed using Maximum Likelihood. Figure 5B presents the QQ-plot against the Uniform distribution of the associated Wilcoxon p-values across 1,000 simulations. Similarly to the Gaussian setting, achieving uniformly distributed p-values is only possible through intra-component data thinning using their oracle overdispersion estimates $\widetilde{\theta}_k$. All other data thinning approaches are performed with biased estimators of the overdispersion compromising the independence between $X^{(1)}$ and $X^{(2)}$, and therefore leading to failure in the control of the post-clustering Type I error rate.

Extending our investigation from simulated scenarios to real-life applications, we used a single-cell RNA-seq dataset from the Tabula Sapiens Consortium (Consortium* et al., 2022) to delve into the practical challenges of estimating overdispersion. Our analysis focused on five distinct cell populations: 2,560 neutrophils, 105 macrophages, 386 monocytes, 454 granulocytes, and 833 CD4 T cells – all collected from a single donor. In this controlled setting, where cell types were known, we succeeded in estimating the overdispersion of 8,333 genes for each cell type using the variance stabilizing transformation implemented in the `sctransform` package (Choudhary and Satija, 2022). Figure 5C illustrates the comparison of gene overdispersion when estimated solely in neutrophils versus the overdispersion estimated for the same genes in the other four cellular populations. Root Mean Squared Error (RMSE) values were computed to quantify the agreement between estimations. Our findings reveal that overdispersion is specific to each cell population, as evidenced by a notable

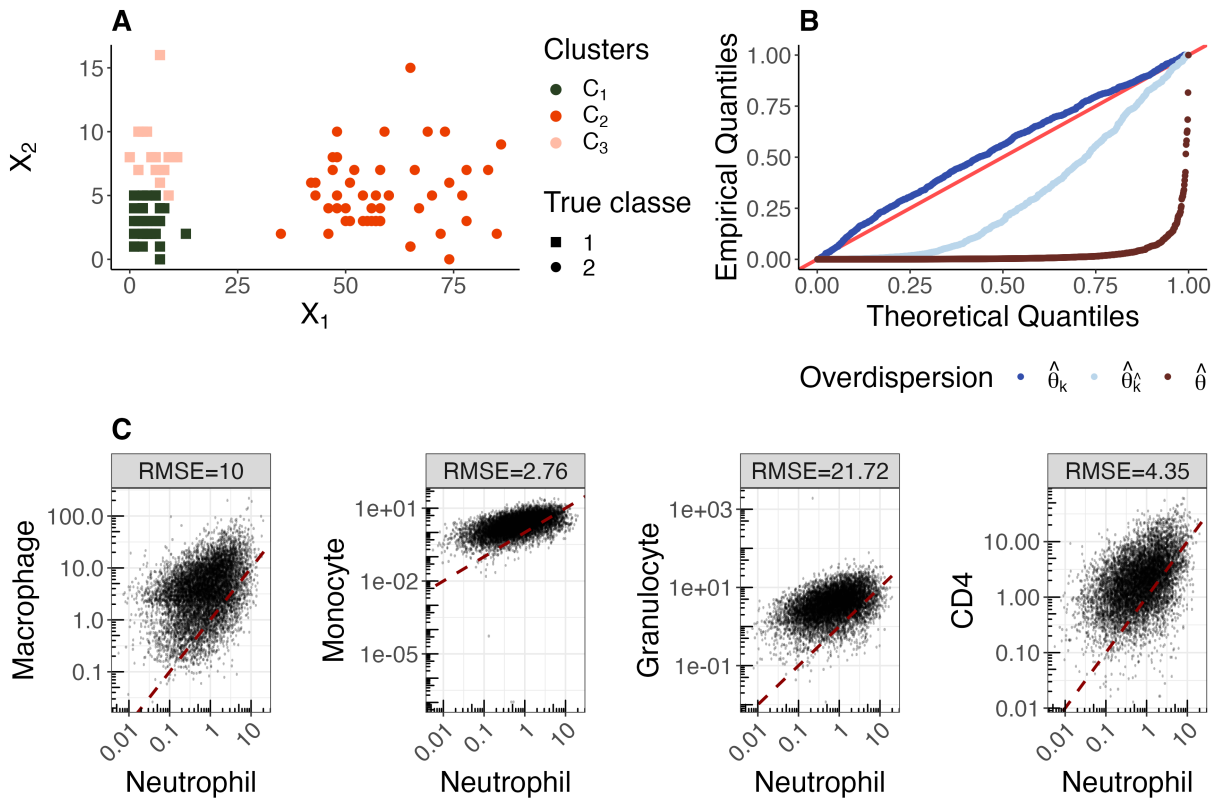

Figure 5: **Challenges in estimating overdispersion, a gene-specific parameter, for negative binomial data thinning.** Panel **A** showcases erroneous clustering results on a simulated dataset, while Panel **B** presents a QQ-plot against the Uniform distribution, displaying Wilcoxon p-values from various data thinning approaches over 1,000 simulations when testing between the two erroneous clusters. In Panel **C**, the estimated overdispersions of genes within different cell populations (macrophages, monocytes, granulocytes, and CD4) are plotted against those in neutrophils, along with their associated Root Mean Square Error (RMSE) for comparison.

deviation from the diagonal and relatively high RMSE values. This underscores the challenge of accurately estimating this parameter without prior knowledge of the true mixture underlying the data. Combined with our simulation studies, these results demonstrate how Type I error can easily be inflated in real-life applications of data thining for scRNA-seq

data analysis, particularly due to the difficulty in providing an unbiased estimate of gene overdispersion.

4 Discussion

We highlight here the practical limitations inherent in data fission and its extension, data thinning, for post-clustering inference challenges. A crucial issue is the assumption of a homogeneous data distribution, which implies an absence of true clusters in the data. To address this limitation and adapt to scenarios with true classes, a shift towards mixture models becomes imperative. However, these models lack a predefined decomposition through data fission or thinning.

We have proposed an intra-component decomposition for data fission and data thining, and demonstrated its theoretical validity. It relies on a priori knowledge of the mixture components scale parameters, such as variances in the Gaussian distribution or the overdispersion in the negative binomial distribution. However in real-life applications, these parameters are unknown. Adequately estimating these parameters becomes intricate in the presence of true clusters, given their component-specific nature; meanwhile the quality of the estimation of those parameters is directly linked with the covariance between the new random variables, $X^{(1)}$ and $X^{(2)}$, decomposing the original data. Only unbiased estimation of these parameters ensures the independence between $X^{(1)}$ and $X^{(2)}$. That independence is paramount for post-clustering inference to adequately control the Type I error rate.

In the Gaussian framework, we theoretically quantify the relationship between the relative bias in variance estimation and the associated Type I error in post-clustering t-tests. In practice, our simulation results suggest that a small relative bias can be acceptable while still achieving effective Type I error control. These first results pave the way for defining

a principled approach to tuning the hyperparmater τ in data fission and data thining for post-clustering inference to optimize statistical power.

As a solution to avoid the need for prior knowledge of component-specific variance parameters in the Gaussian mixture setting, we propose a heteroscedastic model with individual variances, that can be replaced by a plug-in estimate (such as a non-parametric estimator of the local variance). This approach aligns more closely with the distributional assumptions made by data fission and thinning. However, the performance of our non-parametric approach relies heavily on the choice of its bandwidth. The best bandwidth would be the one capturing only the observations originating from the same component, but it would again require knowledge of the true mixture components in spite of their estimation being part of the first clustering step of the method. Consequently, while this heteroscedastic model better fits the parametric assumptions of data fission and thinning, leveraging an accurate enough plug-in estimator remains challenging. This underscores the difficulty of adapting these methods to mixture distributions. We show that when the signal to noise ratio is extremely favorable for the clustering, with very well-separated components, this approach unlocks the use of data-fission for post-clustering differential testing of clustered data. In practice though, one can question the need for data fission in such cases as the uncertainty regarding the clustering is actually very small when $\delta/\sigma \geq 10$. We have also investigated iteratively fissioning the data and updating the variance plug-in estimate, but this heuristic strategy requires even larger separation of the components to effectively work.

Finally, we demonstrate that the results derived from the Gaussian distribution context are readily applicable to the negative binomial distribution commonly employed in modeling RNA-seq data. Specifically, we illustrate on real data that overdispersion is also component-

specific. Therefore, without knowing the true cluster structure in the data, data thinning cannot ensure the needed independence between clustering and differential testing.

In practice, the application of data fission or data thinning for post-clustering inference appears to be akin to a self-referential loop generating circular reasoning. While introduced as a solution for addressing challenges in post-clustering inference, all strategies that could theoretically ensure independence between the two stages of the analysis ultimately rely on knowledge of the true, but unknown, clusters. Despite its conceptual appeal, the practical utility of these methods for post-clustering inference remains limited to extreme cases with extremely high signal vs. noise ratios, emphasizing the need for alternative methodologies that can navigate the complexities of unknown class structures more effectively.

All codes and data needed to reproduce the results presented here are openly accessible from Zenodo with DOI 10.5281/zenodo.11207777 (Hivert et al., 2024b).

5 Acknowledgements

BH is supported partly by the Digital Public Health Graduate's school, funded by the PIA 3 (Investments for the Future – Project reference: 17-EURE-0019). The work was supported through the DESTRIER Inria Associate-Team from the Inria@SiliconValley program (analytical code: DRI-012215), and by the CARE project funded from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement n° IMI2-101005077. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA and Bill & Melinda Gates Foundation, Global Health Drug Discovery Institute, University of Dundee. This work has received funding from the European Union's Horizon 2020 research and innovation programme under EHVA grant agreement n° H2020-681032. This study was carried out in the framework of the University of Bordeaux's

France 2030 program / RRI PHDS. This work benefited from State aid managed by the Agence Nationale de la Recherche under the France 2030 program, reference AI4scMED ANR-22-PESN-0002. Computer time for this study was provided by the computing facilities MCIA (Mésocentre de Calcul Intensif Aquitain) of the Université de Bordeaux and of the Université de Pau et des Pays de l'Adour.

References

Amezquita, R. A., Lun, A. T., Becht, E., Carey, V. J., Carpp, L. N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., et al. (2020). Orchestrating single-cell analysis with bioconductor. *Nature methods*, 17(2):137–145.

Bachoc, F., Maugis-Rabusseau, C., and Neuvial, P. (2023). Selective inference after convex clustering with ℓ_1 penalization. *arXiv preprint arXiv:2309.01492*.

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-based clustering and classification for data science: with applications in R*, volume 50. Cambridge University Press.

Chacón, J. E. and Duong, T. (2020). *Multivariate Kernel Smoothing and Its Applications SMOOTHING AND ITS APPLICATIONS*. CRC PRESS.

Chen, Y. T. and Gao, L. L. (2023). Testing for a difference in means of a single feature after clustering. *arXiv preprint arXiv:2311.16375*.

Choudhary, S. and Satija, R. (2022). Comparison and evaluation of statistical error models for scrna-seq. *Genome Biology*, 23:20.

Consortium*, T. T. S., Jones, R. C., Karkanias, J., Krasnow, M. A., Pisco, A. O.,

Quake, S. R., Salzman, J., Yosef, N., Bulthaup, B., Brown, P., et al. (2022). The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896.

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

Gao, L. L., Bien, J., and Witten, D. (2022). Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, pages 1–11.

Heidenreich, N.-B., Schindler, A., and Sperlich, S. (2013). Bandwidth selection for kernel density estimation: a review of fully automatic selectors. *AStA Advances in Statistical Analysis*, 97:403–433.

Hivert, B., Agniel, D., Thiébaut, R., and Hejblum, B. P. (2024a). Post-clustering difference testing: Valid inference and practical considerations with applications to ecological and biological data. *Computational Statistics & Data Analysis*, 193:107916.

Hivert, B., Agniel, D., Thiébaut, R., and Hejblum, B. P. (2024b). Reproducible codes and results for Running in circles: is practical application feasible for data fission and data thinning in post-clustering differential analysis? (Version v1) [Data set]. *Zenodo*. DOI: 10.5281/zenodo.11207777.

Jaeger, A. and Banks, D. (2023). Cluster analysis: A modern statistical review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(3):e1597.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535–540.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., et al. (2020). Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35.

Leiner, J., Duan, B., Wasserman, L., and Ramdas, A. (2023). Data fission: splitting a single data point. *Journal of the American Statistical Association*, pages 1–12.

Li, Q., Noel-MacDonnell, J. R., Koestler, D. C., Goode, E. L., and Fridley, B. L. (2018). Subject level clustering using a negative binomial model for small transcriptomic studies. *BMC bioinformatics*, 19(1):1–10.

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*.

Neufeld, A., Dharamshi, A., Gao, L. L., and Witten, D. (2023a). Data thinning for convolution-closed distributions. *arXiv preprint arXiv:2301.07276*.

Neufeld, A., Gao, L. L., Popp, J., Battle, A., and Witten, D. (2024). Inference after latent variable estimation for single-cell rna sequencing data. *Biostatistics*, 25(1):270–287.

Neufeld, A., Popp, J., Gao, L. L., Battle, A., and Witten, D. (2023b). Negative binomial count splitting for single-cell rna sequencing data. *arXiv preprint arXiv:2307.12985*.

Pullin, J. M. and McCarthy, D. J. (2024). A comparison of marker gene selection methods for single-cell rna sequencing data. *Genome Biology*, 25(1):56.

Siffer, A., Fouque, P.-A., Termier, A., and Largouët, C. (2018). Are your data gathered? In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 2210–2218.

Song, D., Li, K., Ge, X., and Li, J. J. (2023). Clusterde: a post-clustering differential expression (de) method robust to false-positive inflation caused by double dipping. *Research Square*.

Welch, B. L. (1947). The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35.

Yun, Y.-J. and Foygel Barber, R. (2023). Selective inference for clustering with unknown variance. *Electronic Journal of Statistics*, 17(2):1923–1946.

Zhang, J. M., Kamath, G. M., and David, N. T. (2019). Valid post-clustering differential analysis for single-cell rna-seq. *Cell systems*, 9(4):383–392.

.

Supplementary Materials to "Running in circles: practical limitations for real-life application of data fission and data thinning in post-clustering differential analysis"

S1 Independence proof of the Gaussian process

Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{p\times p})$. Considering the fission process for Gaussian data described in Table 1, we can decompose $\boldsymbol{X}$ into two new random variables $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ using a new random variable $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. It follows form this decomposition that:

$$\boldsymbol{X}^{(1)} \sim \mathcal{N}\left(\boldsymbol{\mu}, \left(1+\tau^2\right)\boldsymbol{\Sigma}\right) \quad \text{and} \quad \boldsymbol{X}^{(2)} \sim \mathcal{N}\left(\boldsymbol{\mu}, \left(1+\frac{1}{\tau^2}\right)\boldsymbol{\Sigma}\right)$$

Moreover, we have $\boldsymbol{X}^{(1)} \perp\!\!\!\perp \boldsymbol{X}^{(2)}$. Indeed, we have:

$$\begin{aligned}
\text{Cov}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}) &= \mathbb{E}\left[\left(\boldsymbol{X}^{(1)} - \mathbb{E}[\boldsymbol{X}^{(1)}]\right)\left(\boldsymbol{X}^{(2)} - \mathbb{E}[\boldsymbol{X}^{(2)}]\right)^t\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{X}^{(1)} - \boldsymbol{\mu}\right)\left(\boldsymbol{X}^{(2)} - \boldsymbol{\mu}\right)^t\right] \qquad \text{since} \qquad \mathbb{E}[\boldsymbol{X}^{(1)}] = \mathbb{E}[\boldsymbol{X}^{(2)}] = \boldsymbol{\mu} \\
&= \mathbb{E}\left[\boldsymbol{X}^{(1)}\boldsymbol{X}^{(2)^t} - \boldsymbol{X}^{(1)}\boldsymbol{\mu}^t - \boldsymbol{\mu}\boldsymbol{X}^{(2)^t} + \boldsymbol{\mu}\boldsymbol{\mu}^t\right] \\
&= \mathbb{E}[\boldsymbol{X}^{(1)}\boldsymbol{X}^{(2)^t}] - \mathbb{E}[\boldsymbol{X}^{(1)}]\boldsymbol{\mu}^t - \boldsymbol{\mu}\mathbb{E}[\boldsymbol{X}^{(2)^t}] + \boldsymbol{\mu}\boldsymbol{\mu}^t \\
&= \mathbb{E}[\boldsymbol{X}^{(1)}\boldsymbol{X}^{(2)^t}] - \boldsymbol{\mu}\boldsymbol{\mu}^t - \boldsymbol{\mu}\boldsymbol{\mu}^t + \boldsymbol{\mu}\boldsymbol{\mu}^t \\
&= \mathbb{E}[\boldsymbol{X}^{(1)}\boldsymbol{X}^{(2)^t}] - \boldsymbol{\mu}\boldsymbol{\mu}^t
\end{aligned}$$

Moreover, we have:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{X}^{(1)}\boldsymbol{X}^{(2)^t}] &= \mathbb{E}\left[(\boldsymbol{X}+\tau\boldsymbol{Z})\left(\boldsymbol{X} - \frac{1}{\tau}\boldsymbol{Z}\right)^t\right] \\
&= \mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^t - \frac{1}{\tau}\boldsymbol{X}\boldsymbol{Z}^t + \tau\boldsymbol{Z}\boldsymbol{X}^t - \boldsymbol{Z}\boldsymbol{Z}^t\right] \\
&= \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^t] - \frac{1}{\tau}\mathbb{E}[\boldsymbol{X}\boldsymbol{Z}^t] + \tau\mathbb{E}[\boldsymbol{Z}\boldsymbol{X}^t] - \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^t] \\
&= \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^t] - \frac{1}{\tau}\mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{Z}^t] + \tau\mathbb{E}[\boldsymbol{Z}]\mathbb{E}[\boldsymbol{X}^t] - \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^t] \qquad \text{since} \qquad \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Z} \\
&= \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^t] - \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^t] \qquad \text{since} \qquad \mathbb{E}[\boldsymbol{Z}] = 0
\end{aligned}$$

But we also have:

$$\begin{aligned}
&\mathbb{V}\text{ar}(\boldsymbol{X}) = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^t] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}^t] = \boldsymbol{\Sigma} \\
&\iff \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^t] = \boldsymbol{\Sigma} + \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}^t] \\
&\iff \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^t] = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^t \qquad \text{since} \qquad \mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu}
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{V}\text{ar}(\boldsymbol{Z}) = \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^t] - \mathbb{E}[\boldsymbol{Z}]\mathbb{E}[\boldsymbol{Z}^t] = \boldsymbol{\Sigma} \\
&\iff \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^t] = \boldsymbol{\Sigma} + \mathbb{E}[\boldsymbol{Z}]\mathbb{E}[\boldsymbol{Z}^t] \\
&\iff \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^t] = \boldsymbol{\Sigma} \qquad \text{since} \qquad \mathbb{E}[\boldsymbol{Z}] = 0
\end{aligned}$$

So finally,

$$\text{Cov}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}) = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^t] - \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^t] - \boldsymbol{\mu}\boldsymbol{\mu}^t = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^t - \boldsymbol{\Sigma} - \boldsymbol{\mu}\boldsymbol{\mu}^t = 0 \tag{4}$$

S2 Impact of covariance estimation under the independence between $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$

Now, let suppose that we use $\boldsymbol{Z} \sim \mathcal{N}\left(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}\right)$ to perform data fission. It follows from equation (4) that, since $\mathbb{V}\text{ar}\left(\boldsymbol{Z}\right) = \hat{\boldsymbol{\Sigma}}$,

$$\text{Cov}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}) = \boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}$$

S3 Derivation of t-test statistic under the null of no-cluster in the univariate data fission post-clustering setting

Let $X_1, \ldots, X_n$ be n independent and identically distributed random variables such that, for all $i = 1, \ldots, n$, $X_i \overset{\text{i.i.d}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$. Here, n represents the sample size. Let $Z_1, \ldots, Z_n \overset{\text{i.i.d}}{\sim} \mathcal{N}\left(0, b^2\right)$ and $\tau \in (0, +\infty)$. Here, b^2 represents any value used as a plug-in for the variance of X, and in particular, b^2 can be an estimate obtained for σ^2. For all $i = 1, \ldots, n$, the splitting process of X_i is given by:

$$X_i^{(1)} = X_i + \tau Z_i \qquad \text{and} \qquad X_i^{(2)} = X_i - \frac{1}{\tau} Z_i$$

We can immediately deduce the marginal distributions of $X_i^{(1)}$ and $X_i^{(2)}$ for all $i = 1, \ldots, n$, thanks to the independence between X_i and Z_i:

$$X_i^{(1)} \sim \mathcal{N}\left(\mu, \sigma^2 + \tau^2 b^2\right) \qquad \text{and} \qquad X_i^{(2)} \sim \mathcal{N}\left(\mu, \sigma^2 + \frac{1}{\tau^2} b^2\right) \tag{5}$$

We denote $\sigma^2_{X^{(1)}} = \sigma^2 + \tau^2 b^2$ and $\sigma^2_{X^{(2)}} = \sigma^2 + \frac{1}{\tau^2} b^2$ as the respective variances of $X_i^{(1)}$ and $X_i^{(2)}$ above.

In the context of data fission to address the challenges of post-clustering inference, a clustering algorithm is applied to the observations of $\boldsymbol{X}^{(1)}$. Without loss of generality, we assume that the clustering algorithm applied to the realizations separates $X_1^{(1)}, \ldots, X_n^{(1)}$ into two clusters C_1 and C_2 around μ (which is typically the case with the k-means algorithm

or with a two-component Gaussian mixture model with homogeneous variance when n is sufficiently large). We also assume that these two clusters have the same size and the same variance.
Thus, the clusters C_1 and C_2 can be expressed as:

$$C_1 = \left\{ i = 1, \dots, n : X_i^{(1)} > \mu \right\} \qquad \text{and} \qquad C_2 = \left\{ i = 1, \dots, n : X_i^{(1)} \leq \mu \right\}$$

We can then derive the conditional distributions of $X_i^{(1)}|C_1$ and $X_i^{(1)}|C_2$. Indeed, $\mathbb{P}(X_i^{(1)} = x|C_1) = \mathbb{P}(X_i^{(1)} = x|X_i^{(1)} > \mu)$ with $X_i^{(1)} \sim \mathcal{N}\left(\mu, \sigma^2_{X^{(1)}}\right)$. Therefore, $X_i^{(1)}|X_i^{(1)} > \mu$ follows a half-normal distribution, and for all $i = 1, \dots, n$:

$$\mathbb{E}\left[X_i^{(1)}|X_i^{(1)} > \mu\right] = \mu + \sqrt{\frac{2\sigma^2_{X^{(1)}}}{\pi}} \quad \text{and} \quad \text{Var}\left(X_i^{(1)}|X_i^{(1)} > \mu\right) = \left(1 - \frac{2}{\pi}\right)\sigma^2_{X^{(1)}}$$

Since the cluster $C_2 = \left\{ i = 1, \dots, n : X_i^{(1)} \leq \mu \right\}$ simply represents the cluster on the other side of the mean μ, we similarly have:

$$\mathbb{E}\left[X_i^{(1)}|X_i^{(1)} \leq \mu\right] = \mu - \sqrt{\frac{2\sigma^2_{X^{(1)}}}{\pi}} \quad \text{and} \quad \text{Var}\left(X_i^{(1)}|X_i^{(1)} \leq \mu\right) = \left(1 - \frac{2}{\pi}\right)\sigma^2_{X^{(1)}}$$

In the context of post-clustering inference, hypothesis tests are performed on the other part of the information, contained in this case in $\boldsymbol{X}^{(2)}$. We are interested in performing a two-sample t-test to evaluate a potential difference in means on $\boldsymbol{X}^{(2)}$ according to the groups defined by the two clusters $C_1 = \left\{ i = 1, \dots, n : X_i^{(1)} > \mu \right\}$ and $C_2 = \left\{ i = 1, \dots, n : X_i^{(1)} \leq \mu \right\}$. Thus, we focus on the following hypotheses:

$$\mathcal{H}_0 : \mu_{C_1} = \mu_{C_2} \qquad \text{vs} \qquad \mathcal{H}_1 : \mu_{C_1} \neq \mu_{C_2}$$

where $\mu_{C_1} = \mathbb{E}\left[X_i^{(2)}|X_i^{(1)} > \mu\right]$ and $\mu_{C_2} = \mathbb{E}\left[X_i^{(2)}|X_i^{(1)} \leq \mu\right]$ are the means of $X_i^{(2)}$ in cluster C_1 and cluster C_2, respectively.

Since we have assumed that the two resulting clusters have equal variances (and the same size $n/2$), we denote the common variance as $s^2(X^{(2)}) = \text{Var}\left(X_i^{(2)}|X_i^{(1)} > \mu\right) = \text{Var}\left(X_i^{(2)}|X_i^{(1)} \leq \mu\right)$. The corresponding test statistic is then given by:

$$T = \frac{\overline{X^{(2)}_{C_1}} - \overline{X^{(2)}_{C_2}}}{\sqrt{\frac{4s^2\left(X^{(2)}\right)}{n}}} \quad \text{where} \quad \overline{X^{(2)}_{C_k}} = \frac{1}{n/2}\sum_{i \in C_k} X_i^{(2)} \quad \text{for} \quad k = 1, 2$$

Although each $X_i^{(2)}$ is Gaussian, this is no longer true conditionally on the clusters, i.e., on $X_i^{(1)} > \mu$ for C_1 (or on $X_i^{(1)} \leq \mu$ for C_2). However, when n is sufficiently large, we can apply the Central Limit Theorem, which gives us:

$$\overline{X^{(2)}_{C_1}} - \overline{X^{(2)}_{C_2}} \overset{\mathcal{L}}{\sim} \mathcal{N}\left(\mu_{C_1} - \mu_{C_2}, \frac{4}{n}s^2\left(X^{(2)}\right)\right)$$

The asymptotic distribution of our test statistic T is therefore:

$$T = \frac{\overline{X_{C_1}^{(2)}} - \overline{X_{C_2}^{(2)}}}{\sqrt{\frac{4s^2\left(X^{(2)}\right)}{n}}} \overset{\mathcal{L}}{\sim} \mathcal{N}\left(\frac{\mu_{C_1} - \mu_{C_2}}{\sqrt{\frac{4s^2\left(X^{(2)}\right)}{n}}}, 1\right) \tag{6}$$

This asymptotic distribution therefore depends on three quantities: μ_{C_1}, μ_{C_2}, and $s^2(X^{(2)})$, which can be computed. By the law of total expectation, we observe that:

$$\mu_{C_1} = \mathbb{E}\left[X_i^{(2)} | X_i^{(1)} > \mu\right] = \mathbb{E}\left[\mathbb{E}\left[X_i^{(2)} | X_i^{(1)}\right] | X_i^{(1)} > \mu\right] \tag{7}$$

Since $X_i^{(1)}$ and $X_i^{(2)}$ are two Gaussian random variables, for all $i = 1, \dots, n$, we have the following bivariate Gaussian vector:

$$\begin{pmatrix} X_i^{(1)} \\ X_i^{(2)} \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma^2_{X^{(1)}} & \rho\sigma_{X^{(1)}}\sigma_{X^{(2)}} \\ \rho\sigma_{X^{(1)}}\sigma_{X^{(2)}} & \sigma^2_{X^{(2)}} \end{pmatrix}\right)$$

where $\rho = \text{Cor}\left(X_i^{(1)}, X_i^{(2)}\right)$. Using the properties of multivariate Gaussian distributions, we can deduce the conditional distribution of $X_i^{(2)}|X_i^{(1)}$, which for all $i = 1, \dots, n$, is:

$$X_i^{(2)} | X_i^{(1)} \sim \mathcal{N}\left(\mu + \frac{\rho\sigma_{X^{(1)}}\sigma_{X^{(2)}}}{\sigma^2_{X^{(1)}}}\left(X_i^{(1)} - \mu\right), \sigma^2_{X^{(2)}} - \frac{\rho^2\sigma^2_{X^{(1)}}\sigma^2_{X^{(2)}}}{\sigma^2_{X^{(1)}}}\right) \tag{8}$$

which simplifies to:

$$X_i^{(2)} | X_i^{(1)} \sim \mathcal{N}\left(\mu + \rho\frac{\sigma_{X^{(2)}}}{\sigma_{X^{(1)}}}\left(X_i^{(1)} - \mu\right), \sigma^2_{X^{(2)}}\left(1 - \rho^2\right)\right) \tag{9}$$

Substituting the expectation of $X_i^{(2)}|X_i^{(1)}$ from above into equation (7), we obtain:

$$\begin{aligned}
\mu_{C_1} &= \mathbb{E}\left[\mu + \rho\frac{\sigma_{X^{(2)}}}{\sigma_{X^{(1)}}}\left(X_i^{(1)} - \mu\right) | X_i^{(1)} > \mu\right] \\
&= \mu + \rho\frac{\sigma_{X^{(2)}}}{\sigma_{X^{(1)}}}\left(\mathbb{E}\left[X_i^{(1)} | X_i^{(1)} > \mu\right] - \mu\right) \\
&= \mu + \rho\frac{\sigma_{X^{(2)}}}{\sigma_{X^{(1)}}}\left(\mu + \sqrt{\frac{2\sigma^2_{X^{(1)}}}{\pi}} - \mu\right) \\
&= \mu + \rho\sqrt{\frac{2}{\pi}\sigma^2_{X^{(2)}}}
\end{aligned}$$

By an identical reasoning, we find:

$$\mu_{C_2} = \mathbb{E}\left[X_i^{(2)} | X_i^{(1)} \leq \mu\right] = \mu - \rho\sqrt{\frac{2}{\pi}\sigma^2_{X^{(2)}}}$$

Finally, using the law of total variance, we have:

$$\text{Var}\left(X_i^{(2)} | X_i^{(1)} > \mu\right) = \mathbb{E}\left[\text{Var}\left(X_i^{(2)} | X_i^{(1)}\right) | X_i^{(1)} > \mu\right] + \text{Var}\left(\mathbb{E}\left[X_i^{(2)} | X_i^{(1)}\right] | X_i^{(1)} > \mu\right)$$

From equation (9), we first observe that:

$$\begin{aligned}\mathbb{E}\left[\mathrm{Var}\left(X_i^{(2)}|X_i^{(1)}\right)\middle| X_i^{(1)} > \mu\right] &= \mathbb{E}\left[\sigma^2_{X^{(2)}}\left(1-\rho^2\right)|X_i^{(1)} > \mu\right] \\ &= \sigma^2_{X^{(2)}}\left(1-\rho^2\right)\end{aligned}$$

and then that:

$$\begin{aligned}\mathrm{Var}\left(\mathbb{E}\left[X_i^{(2)}|X_i^{(1)}\right]|X_i^{(1)} > \mu\right) &= \mathrm{Var}\left(\mu + \rho\frac{\sigma_{X^{(2)}}}{\sigma_{X^{(1)}}}\left(X_i^{(1)}-\mu\right)|X_i^{(1)} > \mu\right) \\ &= \rho^2\frac{\sigma^2_{X^{(2)}}}{\sigma^2_{X^{(1)}}}\mathrm{Var}\left(X_i^{(1)}|X_i^{(1)} > \mu\right) \\ &= \rho^2\frac{\sigma^2_{X^{(2)}}}{\sigma^2_{X^{(1)}}}\left(1-\frac{2}{\pi}\right)\sigma^2_{X^{(1)}} \\ &= \rho^2\left(1-\frac{2}{\pi}\right)\sigma^2_{X^{(2)}}\end{aligned}$$

Finally, we obtain:

$$\begin{aligned}s^2\left(X^{(2)}\right) &= \mathrm{Var}\left(X_i^{(2)}|X_i^{(1)} > \mu\right) \\ &= \sigma^2_{X^{(2)}}\left(1-\rho^2\right) + \rho^2\left(1-\frac{2}{\pi}\right)\sigma^2_{X^{(2)}} \\ &= \sigma^2_{X^{(2)}}\left(1-\frac{2}{\pi}\rho^2\right)\end{aligned}$$

By a similar reasoning, we find:

$$\mathrm{Var}\left(X_i^{(2)}|X_i^{(1)} \leq \mu\right) = \sigma^2_{X^{(2)}}\left(1-\frac{2}{\pi}\rho^2\right),$$

which, fortunately, confirms our initial assumption of equal intra-cluster variances. Thus, we can finally compute:

$$\mathbb{E}\left[T\right] = \frac{\mu_{C_1}-\mu_{C_2}}{\sqrt{\frac{4s^2\left(X^{(2)}\right)}{n}}} = \frac{\mu+\rho\sqrt{\frac{2}{\pi}\sigma^2_{X^{(2)}}} - \left(\mu-\rho\sqrt{\frac{2}{\pi}\sigma^2_{X^{(2)}}}\right)}{\sqrt{\frac{4\sigma^2_{X^{(2)}}\left(1-\frac{2}{\pi}\rho^2\right)}{n}}} = \frac{\rho\sqrt{n}\sqrt{\frac{2}{\pi}}}{\sqrt{1-\frac{2}{\pi}\rho^2}},$$

and we ultimately find:

$$T \overset{\mathcal{L}}{\sim} \mathcal{N}\left(\frac{\rho\sqrt{n}}{\sqrt{\frac{\pi}{2}-\rho^2}}, 1\right). \tag{10}$$

Recall that the sample $X_1, \ldots, X_n$ is normally distributed: it does not contain any true clusters, and thus no real difference in means exists between subgroups of observations other than what is due to clustering. Therefore, the test statistic T should be under $\mathcal{H}_0$, and thus centered around 0. In our result in (10), we observe a deviation of the distribution of T from 0, quantified by $\frac{\rho\sqrt{n}}{\sqrt{\frac{\pi}{2}-\rho^2}}$. The Type I error at level α associated with this test is

then given by $1 - F(q_{\alpha/2}) + F(-q_{\alpha/2})$, where F is the cumulative distribution function of the normal distribution $\mathcal{N}\left(\frac{\rho\sqrt{n}}{\sqrt{\frac{\pi}{2}-\rho^2}}, 1\right)$ and $q_{\alpha/2}$ is the quantile of order $\alpha/2$ of the standard normal distribution $\mathcal{N}(0,1)$.

Here we assumed that all variances were known, and thus $s^2\left(X^{(2)}\right)$ was known as well. Therefore, it was possible to calculate the distribution of the test statistic for the Z test to compare means between two samples. However, this result extends easily to the more practical case where $s^2\left(X^{(2)}\right)$ is unknown and an estimate $\widehat{s^2}\left(X^{(2)}\right)$ is used instead. In this case, still assuming that variances are equal between the two clusters, the distribution of the test statistic T follows a Student's t distribution $\mathcal{T}(n-2)$ (due to the uncertainties associated with estimating this common variance). The associated Type I error then becomes: $1 - F_{\mathcal{T}}(q_{\alpha/2}) + F_{\mathcal{T}}(-q_{\alpha/2})$, where $F_{\mathcal{T}}$ is the cumulative distribution function of the non-central Student's t distribution with mean $\frac{\rho\sqrt{n}}{\sqrt{\frac{\pi}{2}-\rho^2}}$ and $n-2$ degrees of freedom, and $q_{\alpha/2}$ is the quantile of order $\alpha/2$ of the Student's t distribution with $n-2$ degrees of freedom.

This result underscores the crucial importance of precise variance estimation for the application of data fission. Indeed, for the test to be valid, that is, for the Type I error to be controlled at level α, the distribution of the test statistic must be centered at 0. This implies that:

$$
\begin{aligned}
\rho\sqrt{n} = 0 &\iff \rho = 0 \\
&\iff \operatorname{Cor}\left(X_i^{(1)}, X_i^{(2)}\right) = 0 \quad \forall i = 1, \ldots, n \\
&\iff \frac{\operatorname{Cov}\left(X_i^{(1)}, X_i^{(2)}\right)}{\sigma_{X^{(1)}}\sigma_{X^{(2)}}} = 0 \quad \forall i = 1, \ldots, n \\
&\iff \operatorname{Cov}\left(X_i^{(1)}, X_i^{(2)}\right) = 0 \quad \forall i = 1, \ldots, n \\
&\iff \sigma^2 - b^2 = 0
\end{aligned}
$$

Thus, only a data fission procedure performed with the true variance parameter (or at least an unbiased estimator of it) can ensure effective control of the Type I error.