A ROBUST STATISTICAL METHOD FOR GENOME-WIDE ASSOCIATION ANALYSIS OF HUMAN COPY NUMBER VARIATION

HAN WANG, CHANGHU WANG, LINJIE WU, AND RUIBIN XI*

ABSTRACT. Conducting genome-wide association studies (GWAS) in copy number variation (CNV) level is a field where few people involves and little statistical progresses have been achieved, traditional methods suffer from many problems such as batch effects, heterogeneity across genome, leading to low power or high false discovery rate. We develop a new robust method to find disease-risking regions related to CNV’s disproportionately distributed between case and control samples, even if there are batch effects between them, our test formula is robust to such effects. We propose a new empirical Bayes rule to deal with overfitting when estimating parameters during testing, this rule can be extended to the field of model selection, it can be more efficient compared with traditional methods when there are too much potential models to be specified. We also give solid theoretical guarantees for our proposed method, and demonstrate the effectiveness by simulation and real data analysis.

1. INTRODUCTION

Copy number variation (CNV) detection is of great significance in understanding the mechanism of some severe diseases. Many studies have reported the strong correlation between CNV and diseases ranging from obesity to complex neurological diseases such as Parkinson, Alzheimer, mental retardation and schizophrenia to cancer. Zhang et al. [1] gave a comprehensive summary about CNV in human health, disease, and evolution. Although studies investigating accurate copy number states have a rapid growth in the past few years [2, 3], there is little devotion to developing robust statistical methods for identifying significant associations between CNVs and diseases. Meanwhile, the majority of existing genome-wide association studies (GWAS) paid their attentions to associations between single nucleotide polymorphisms (SNPs) and a certain disease, jumping out of the field of SNP-level seldomly. In this paper, we aim to develop an effective and robust statistical method to conduct GWAS in CNV level, extracting meaningful disease-risking CNV regions.

Statistical methods developed to enhance the power of detecting CNVs emerge prominently in recent years, we can roughly classify them into two types. The former is to perform the detection on one chromosome at a time, for the segments possessing CNVs, the corresponding signals appear to present abrupt rise or drop, thus it can be recognized as a change-point problem. Traditional change-point detection methods like maximum likelihood based, sliding window based, \(l_1\)-penalization based [4] or graph based can be applied. Remarkably, one of the most popular maximum likelihood based algorithm is Circular Binary Segmentation (CBS) [5], which splice two ends of the chromosome to make it a circle and then test if there exist a CNV segment with a significantly different mean from the remaining part of the circle. The other statistical methods regarding to CNV detection is to pool information of all samples together. By scanning a shared segment simultaneously across multiple sequences, the power can be improved substantially. Zhang et al. [6] generalized the CBS statistic into the multiple sample case, the corresponding test statistic is
simply a summation of the CBS statistic of a single sample. It’s not hard to imagine that despite the enhanced power in finding rare CNV segments, this generalization inherited some shortcomings of the CBS algorithm. Jeng et al. \cite{7} modelled each CNV region as a 2-class mixture Gaussian distribution, one class with mean zero represents those samples having normal copy number 2, the other represents those CNV-carriers in this region with nonzero mean, then they established a test statistic closely relating to standardized uniform empirical process based statistics to pool information. On account of the heterogeneity of CNVs, different samples may possess different kinds of CNVs on the same region, there is credible doubt about the insufficiency of 2 classes in this model. Moreover, these two multiple-sample approaches described above developed their models based on the assumption that the boundaries of a certain CNV region across samples are the same, when handling real data, this assumption is rarely satisfied.

Researches focusing on establishing statistical methods to investigate associations between CNVs and diseases are not so active in contrast. One natural way to do GWAS in CNV level is to assign a copy number (CN) state to every case or control sample at a specific region, and then conduct a Fisher’s test or chi-square test based on a contingency table. To the best of our knowledge, Barnes’s work \cite{12} was a representative among the few statistical methods devoted to GWAS in CNV level, their work pointed out shortcomings of doing test based on contingency table described above, and they adopted a likelihood-based approach to test whether there is significant correlation between copy number and case-control status. But it’s worth noting that when using generalized linear model (GLM) to describe the connection between signal and copy number, they used the same set of parameters for case and control data, neglecting the possible batch effects, so concise and efficient tools to deal with batch effects are in need under this circumstance. In addition, this method focused on the test of a given CNV region and depended on other methods to pick up suspicious regions.

In this paper, we propose a new testing model to deal with case-control whole genome data, with the aim of finding disease-susceptibility CNV regions and their corresponding genes, handling issues arised from batch effects and heterogeneity across genome simultaneously. When estimating parameters, we develop an empirical Bayes framework to avoid overfitting, this framework can be regarded as a new tool for model selection, we will show the advantage of our method over Bayesian information criterion (BIC) under some conditions and prove consistency of our resulting estimation. Notably, our model can pool the samples’ information together, having the ability to capture short segments without boundaries’ accordance assumption. We also show the effectiveness of our method in both simulation and real data analysis.

2. Methods

2.1. Model formulation. Suppose we have 2 groups of whole genome data, one for case and the other for control. Both the case and control data can be represented in a matrix form: \( \{x_{it}; i = 1, \ldots, N_1, t = 1, \ldots, T\} \) and \( \{y_{jt}; j = 1, \ldots, N_2, t = 1, \ldots, T\} \) denote the case and control information separately, where \( x_{it} \) and \( y_{jt} \) are observations at location \( t \) for the \( i \)th sample of case and \( j \)th sample of control separately, usually they are in the format of the log 2 transform of the copy number ratio (the ratio of the sample and the reference DNA). Here \( N_1 \) is the number of case samples and \( N_2 \) is that of control samples, \( T \) is the number of locations in a sequence. Unlike traditional models paying attention to the distribution of a certain row of the data matrix, which is a particular sample sequence, we are interested in the distribution of a certain column of the data matrix, which is a particular location in the sequence for all samples, if at a location the
distribution of case is significantly different from that of control, this point may be with suspicion of being disease related.

As is commonly recognized, we assume that for each sample, taking case sample for an illustration, the observed values $x_i = \{x_{it}, t = 1, \cdots, T\}$ are distributed as Gaussian and mutually independent. Furthermore, we assume 5 types of CNVs: 2 copies deletion (CN=0), 1 copy deletion (CN=1), normal (CN=2), 1 copy duplication (CN=3), 2 copies duplication (CN=4). Since observations of each CNV type accumulate around a particular value, for each location $t$ in the sequence, we obtain a mixture Gaussian distribution:

$$X_{it} \sim \sum_{k=1}^{5} \alpha_{tk}^d N(\mu_{tk}^d, (\sigma_{tk}^d)^2), \ i = 1, \cdots, N_1 \text{ for case samples at location } t.$$  

$$Y_{jt} \sim \sum_{k=1}^{5} \alpha_{tk}^c N(\mu_{tk}^c, (\sigma_{tk}^c)^2), \ j = 1, \cdots, N_2 \text{ for control samples at location } t.$$  

Where $N(\mu_{tk}^d, (\sigma_{tk}^d)^2)$ ($k = 1, \cdots, 5$) correspond to distribution of cluster CN = 1, \cdots, 5 separately for case sample at location $t$, $\alpha_{tk}$ ($k = 1, \cdots, 5$) are the 5 clusters’ proportion at location $t$, we refer to $\alpha_{tk}^d$, $\mu_{tk}^d$ and $(\sigma_{tk}^d)^2$ ($k = 1, \cdots, 5$) as proportion parameters, location parameters and scale parameters for convenience, illustrations of control model are analogous with that of case.

So far, a natural way of grasping the region we are interested in is to test the equivalence of the two Gaussian mixtures of case and control at a particular point, conventional testing approaches regarding with distribution equivalence can be applied directly. Unfortunately, because of the variety of experimental platforms, unmeasured variables ranging from quality of equipments to technicians and other confounding factors, the Gaussian distributions related to the same cluster for case and control samples may not be the same. Under this circumstance, even for a location without CNV for both case and control samples, the test concerned with distribution equivalence has the potential of claiming a significant outcome. To overcome these effects, we propose a test model as follows:

$$H_0^t: \alpha_{tk}^d = \alpha_{tk}^c \text{ for all } k = 1, \cdots, 5 \text{ at location } t$$  

$$H_1^t: \alpha_{tk}^d \neq \alpha_{tk}^c \text{ for some } k = 1, \cdots, 5 \text{ at location } t \quad (1)$$

Taking notice of the continuity of CNVs, which means that deletion or duplication occurs on a continuous segment of nucleotide bases but not a single base, we have confidence to believe that the distributions between neighboring locations are the same unless there exists a change point of CN state for some samples in these neighboring locations. So we can increase the testing power by performing test bin by bin rather than location by location, to be specific, we split the whole sequence into bins with the same size, then in each bin of length $p$, there are two multivariate Gaussian mixtures for case and control seperately:

$$X_{ib} \sim \sum_{k=1}^{5} \alpha_{bk}^d N(\mu_{bk}^d1_p, (\sigma_{bk}^d)^21_p), \ i = 1, \cdots, N_1 \text{ for case samples at bin } b.$$  

$$Y_{jb} \sim \sum_{k=1}^{5} \alpha_{bk}^c N(\mu_{bk}^c1_p, (\sigma_{bk}^c)^21_p), \ j = 1, \cdots, N_2 \text{ for control samples at bin } b.$$  

3
where $1_p$ denotes a vector of length $p$ with all elements setting to 1, $I_p$ is a $p$-dimension identity matrix, $X_{ib}$ and $Y_{jb}$ are samples of case and control at bin $b$. The distribution related to each cluster is multivariate Gaussian with mean vector of the same value, mutual independence between locations leads to a diagonal covariance matrix, we further assume equivalence of variance for all univariate Gaussian in this bin regarding to the same CN cluster. The testing rule for each bin is the same as that for each location, which merely tests the equivalence of proportion rather than the whole distribution.

A straightforward way of realizing the proportion test is to perform the likelihood ratio test, calculating the maximum likelihood under $H_0$ and the whole feasible region, getting a likelihood ratio test statistic and claiming a p-value taking chi-square distribution as a reference.

2.2. Setting a prior for CN-specific-mean. A new obstacle in front is overfitting. For instance, locations at which all samples have normal copy number or only deletions appear, the 5-cluster mixture Gaussian is obviously an overfitting model. To overcome this issue and make our model applicable and robust to all multi-sample CNV situations, we propose an empirical Bayes method to establish a prior for $\mu^c_{bk}$ and $\mu^d_{bk}$ in the mixture Gaussian distributions, pinning the means on their deserved positions in some sense, by which we can also ensure that $\alpha^c_{bk}$ and $\alpha^d_{bk}$ which we used to test equivalence correspond to the same CN state of case and control samples. From here on, unless otherwise specified, we drop the bin index symbol $b$ for simplicity.

We set $\mu^c_k \sim N(\tau_k, (\sigma^c_{\tau_k})^2)$, $\mu^d_k \sim N(\tau_k, (\sigma^d_{\tau_k})^2)$, $k = 1, \cdots, 5$ for case and control CN-specific-mean at all bins. Where $\tau_k$ is the prior knowledge of the CN-specific-mean, for example, there are CNV regions declared by some other methods, we can set $\tau_k$ as the mean of all observations in CN cluster $k$ found by a certain method. Here we take $\tau_k = -1.3, -0.5, 0, 0.4, 0.73$ for $k = 1, \cdots, 5$ separately according to our experiences. $(\sigma^c_{\tau_k})^2$, $(\sigma^d_{\tau_k})^2$ are the prior variance of $\mu^c_{bk}$ and $\mu^d_{bk}$ separately, controlling the “status” of prior mean $\tau_k$, when $(\sigma^c_{\tau_k})^2$, $(\sigma^d_{\tau_k})^2$ are small, the prior mean displays a dominating status in estimating the CN-specific-mean, whereas weakening the power of sample observations in estimation. When $(\sigma^c_{\tau_k})^2$, $(\sigma^d_{\tau_k})^2$ are large, there is little difference between this setting and no prior setting, leading to unavoidable overfitting and estimation bias. So we have to give a moderate prior variance adaptive to various CNV proportion settings, bin length and conceivable batch effects, taking account of sample size simultaneously. It is well-known that as the sample size goes to infinity, any prior distribution of order $O_p(1)$ will be degenerate compared with the distribution explained by data, so the choice of $(\sigma^c_{\tau_k})^2$ and $(\sigma^d_{\tau_k})^2$ is indeed of great importance, we will show more in detail in the next section and give some theoretical guarantees on its choice.

Incorporating the empirical Bayes formula, we update the likelihood of case and control observations under $H_1$ separately as:

\[
l(\alpha^d_1, \cdots, \alpha^d_5, \mu^d_1, \cdots, \mu^d_5, (\sigma^d_1)^2, \cdots, (\sigma^d_5)^2 | X_1, \cdots, X_{N_1}) \prod_{k=1}^{5} g(\mu^d_k | \tau_k, (\sigma^d_{\tau_k})^2) \\
= \prod_{k=1}^{5} g(\mu^d_k | \tau_k, (\sigma^d_{\tau_k})^2) \prod_{i=1}^{N_1} \sum_{k=1}^{5} \alpha^d_i f(X_i | \mu^d_k, (\sigma^d_k)^2) \triangleq \tilde{f}(X_1, \cdots, X_{N_1} | \theta^d)
\]
$$l(\alpha_1, \ldots, \alpha_5, \mu_1, \ldots, \mu_5, (\sigma_1)^2, \ldots, (\sigma_5)^2 | Y_1, \ldots, Y_{N_2}) \prod_{k=1}^{5} g(\mu_k^c | \tau_k, (\sigma_k^c)^2)$$

$$= \prod_{k=1}^{5} g(\mu_k^c | \tau_k, (\sigma_k^c)^2) \prod_{j=1}^{N_2} \alpha_k^c f(Y_j | \mu_k^c, (\sigma_k^c)^2) \overset{\triangle}{=} \tilde{f}(Y_1, \ldots, Y_{N_2} | \theta^c)$$

Where \( g(\mu_k^d | \tau_k, (\sigma_k^d)^2) \) is the density of distribution \( N(\tau_k, (\sigma_k^d)^2) \) valued on \( \mu_k^d \), \( f(X_i | \mu_k^d, (\sigma_k^d)^2) \) is the density of distribution \( N(\mu_k^d 1_p, (\sigma_k^d)^2 1_p) \) valued on \( X_i \), symbols under control condition are parallel with that of case samples, \( \theta^d \) and \( \theta^c \) denotes the overall parameters to be estimated in a bin.

2.3. EM updating rule. Mixture properties of observations in a specific bin make the solution to maximum likelihood estimation(MLE) by first-order derivation intractable. A general approach that can be applied to this kind of dataset is the EM algorithm. Taking case observations under the alternative hypothesis \( H_1 \) as an instance, there is an unobserved variable \( Z_i^d \) behind every probe intensity observation \( X_i \) for each bin, here \( Z_i^d \) is the true CN state for sample \( i \) in a bin. The E-step of the EM algorithm returns a \( Q \)-function, which is a conditional expectation of the log-likelihood of the complete data(\( X_1, \ldots, X_{N_1}, Z_1^d, \ldots, Z_{N_1}^d \)) given the observed data (\( X_1, \ldots, X_{N_1} \)) and a current estimation of parameters \( \theta^d(t) \), which can also be regarded as a parameter-indexed variational lower bound of the likelihood function, we are going to choose a best lower bound, which is indexed by MLE. The M-step returns estimations of parameters which maximize the \( Q \)-function. Since conclusions are parallel under \( H_1 \) regarding to case and control, unless otherwise mentioned, we only present those of case sample. Concretely, in our setting, the log-likelihood of the complete data is

$$\sum_{i=1}^{N_1} \left\{ \sum_{k=1}^{5} \left[ -\frac{p}{2} \ln(\sigma_k^d)^2 - \frac{1}{2(\sigma_k^d)^2}(X_i - \mu_k^d 1_p)'(X_i - \mu_k^d 1_p) \right] 1_{\{Z_i^d = k\}} \right\}$$

$$+ \sum_{k=1}^{5} 1_{\{Z_i^d = k\}} \ln \alpha_k^d + \sum_{k=1}^{5} \left[ -\frac{p}{2} \ln(\sigma_k^d)^2 - \frac{1}{2\sigma_k^d}(\mu_k^d - \tau_k)^2 \right] + C$$

(4)

Where \( C \) is a constant term which is irrelevant to parameter estimation.

The corresponding \( Q \)-function is simply to replace \( 1_{\{Z_i^d = k\}} \) with \( P(Z_i | X_i, \theta^d(t)) \)

$$Q(\theta^d | \theta^d(t)) = \sum_{i=1}^{N_1} \left\{ \sum_{k=1}^{5} \left[ -\frac{p}{2} \ln(\sigma_k^d)^2 - \frac{1}{2(\sigma_k^d)^2}(X_i - \mu_k^d 1_p)'(X_i - \mu_k^d 1_p) \right] P(Z_i | X_i, \theta^d(t)) \right\}$$

$$+ \sum_{k=1}^{5} P(Z_i | X_i, \theta^d(t)) \ln \alpha_k^d + \sum_{k=1}^{5} \left[ -\frac{p}{2} \ln(\sigma_k^d)^2 - \frac{1}{2(\sigma_k^d)^2}(\mu_k^d - \tau_k)^2 \right]$$

(5)
Hence the EM updating rule is

\[
\begin{align*}
\mu_k^d(t + 1) &= \frac{p(\sigma_{rk}^d)^2 \sum_{i=1}^{N_1} (\bar{X}_i b_{ik}^d(t)) + \tau_k(\sigma_k^d(t))^2}{p(\sigma_{rk}^d)^2 \sum_{i=1}^{N_1} b_{ik}^d(t) + (\sigma_k^d(t))^2} \\
(\sigma_{bk}^d(t + 1))^2 &= \sum_{i=1}^{N_1} (x_i - \mu_k^d(t + 1)1_p)'(x_i - \mu_k^d(t + 1)1_p)b_{ik}^d(t) \\
\alpha_k^d(t + 1) &= \frac{\sum_{i=1}^{N_1} b_{ik}^d(t)}{N_1}
\end{align*}
\]

Where \( p \) is the length of the bin, \( \bar{X}_i \) is the mean of the \( i \)-th case observations in a bin, parameters updated in the \( t \)-th step are denoted by \( (t) \) in their right side. \( b_{ik}^d(t) = P(Z_i = k | X_i, \theta^d(t)) \) is the conditional probability of the latent variable, i.e. the state of the \( i \)-th case sample in a bin belongs to CN cluster \( k \) given observations under the current parameter estimation. The claim of the control sample is similar to that of case sample.

It is noteworthy that the convergent EM may stop at a local maxima but not a global maxima, there are many literatures discussed about it, for example Jin et.al. [14] focused on theoretical guarantees of uniformly weighted mixtures of M isotropic Gaussians, they pointed out that in order to recover a global maximum with at least constant probability, the EM algorithm with random initialization must be executed at least \( e^{\Omega(M)} \) times, so efficient initialization methods are strongly needed such as moment-based initialization. In this paper we adopt outcomes from “mclust” [15] as the initial value, which is a famous R package for classification. The function “Mclust” in the package adopted BIC as the model selection criteria, offering us the parameters \((\alpha_k, \mu_k, \sigma_k^2)\) of each corresponding cluster, assigning each sample a cluster label as well. To accommodate our CNV data environment, we initialize our proportion parameter \( \alpha_k \) according to the classification output of “Mclust”, concretely, initialization of \( \alpha_k \) is proportion of samples whose “mclust” cluster’s mean parameter is located in a specific interval, i.e. we can initialize a CN state of each sample based on outcomes returned by “Mclust”. Once initialization of each sample’s CN state has been ensured, we can calculate the mean and variance of each initialized cluster as our parameter initialization.

In our setting, we initialize a sample to cluster “CN=0” if its cluster’s mean parameter returned by “Mclust” is located in interval \( I_1 \triangleq (-3, -0.9] \) analogously, \( I_2 \triangleq (-0.9, -0.3], I_3 \triangleq (-0.3, 0.25], I_4 \triangleq [0.25, 0.6], I_5 \triangleq (0.6, 3] \).

From the updating rule of \( \mu_k^d \), we can see that the posterior mean of cluster \( k \) is a weighted average of the prior mean and the “data-descriptive” mean of cluster \( k \), with weights \( (\sigma_k^d)^2/N_1 \) and \( p(\sigma_{rk}^d)^2 \alpha_k^d \) separately, here \( \alpha_k^d = \sum_{i=1}^{N_1} b_{ik}^d / N_1 \). So we have confidence to expect that with suitable choice of prior parameters, when the bin size \( p \rightarrow \infty \), the estimated parameters have consistent properties, even if some cluster \( k \) doesn’t exist, we can also expect its corresponding proportion \( \alpha_k \) going to 0.
Before showing our theoretical results, we introduce two important lemmas, the first lemma is from Theorem 3.2 proposed by D. Dacuha-Castelle and E. Gassiat [16], due to space limitation, we only list the most important part of this theorem, detailed explanation and proof are available in the original, proof of the second can be found in the appendix section.

**Lemma 2.1.** Suppose $\mathcal{F} = (f_\xi)_{\xi \in \Gamma}$ is a family of probability densities with respect to $v$, $\Gamma$ is a compact subset of $\mathbb{R}^s$ for some integer $s$, $\mathcal{G}_k$ is the set of all $k$ mixtures of densities of $\mathcal{F}$:

$$
\mathcal{G}_k = \{g_{\alpha,\pi} = \sum_{i=1}^{k} \alpha_i \cdot f_{\xi_i} : \alpha = (\alpha_1, \cdots, \alpha_k), \pi = (\xi_1, \cdots, \xi_k), \forall i = 1, \cdots, k, \xi_i \in \Gamma, 0 \leq \alpha_i \leq 1, \sum_{i=1}^{k} \alpha_i = 1\}
$$

If $X_1, \cdots, X_n$ are $n$ samples from a mixture of $k$ populations; that is $X_1, \cdots, X_n \sim g_0 = \sum_{l=1}^{k} \pi_l^0 f_{\xi_l^0}$. Define for any $g$ in $\mathcal{G}_k$,

$$
l_n(g) = \sum_{i=1}^{n} \log[g(X_i)]
$$

and the statistic

$$
T_n(k) = \sup_{g \in \mathcal{G}_k} l_n(g) - l_n(g_0)
$$

Then under some mild assumptions, $T_n(k)$ converges in distribution to the variable $\frac{1}{2} \sup_{d \in \mathcal{D}} \eta_d^2 \mathbb{1}_{\eta_d \geq 0}$, here $\eta_d$ is the Gaussian process indexed by $\mathcal{D}$ with covariance that is the usual hilbertian product in $H$, $\mathcal{D}$ is the subset of the unit sphere of $H$ of functions of a specific form, $H$ is the Hilbert space $L^2(g_0)$ and does not equal 1 uniformly.

**Lemma 2.2.** We assume $d_k = \tau_{k+1} - \tau_k$ for $k \in \{1, \cdots, 4\}$, and suppose there exists an $\zeta > 0$ s.t. $|\mu_k - \tau_k| < \zeta$ for $k \in \{1, \cdots, 5\}$, $\zeta < \min_{k \in \{1, \cdots, 4\}} d_k/2$, then any nonidentical location permutation rule $\pi$ applied to $\theta^*$ such that $\pi(\theta^*) = \theta^{*\pi}$, we have

$$
G_n(\theta^*) - G_n(\theta^{*\pi}) = O\left(\frac{1}{\sigma_{\tau_k}^2}\right) > 0, \forall \epsilon_1 > 0
$$

(7)

here we denote $\theta^*$ as the overall underlying true parameters $(\mu_k^*, (\sigma_k^*)^2, \alpha_k^*)_{k=1}^5$, $G_n(\theta) = G_n(\mu_1, \cdots, \mu_5) = \sum_{k=1}^{5} \log[g(\mu_k|\tau_k, \sigma_{\tau_k}^2)]$ is the logarithm of the prior distribution.

**Theorem 2.3.** Suppose parameters $(\mu_k^d, (\sigma_k^d)^2, \alpha_k^d)$ for $k = 1, \cdots, 5$ lie in a closed bounded set $\Theta$. When $p \to \infty$, sample size $N_1$ is a fixed constant, $(\sigma_k^d)^2$ are equivalent for $k = 1, \cdots, 5$, the order of $(\sigma_k^d)^2$ is $O(1/m(p))$, where $m(p)$ is any increasing function of $p$ of order less than $O(p)$:

(a) If the 5 clusters all exist in the sample with size $N_1$, the underlying true location and scale parameters are $(\mu_k^d, (\sigma_k^d)^2, \alpha_k^d)$, with $(\mu_k^d)_{k=1}^5$ satisfying $\mu_1^d < \mu_2^d < \cdots < \mu_5^d$, we further assume $d_k = \tau_{k+1} - \tau_k$ for $k \in \{1, \cdots, 4\}$, and suppose there exists an $\zeta > 0$ s.t. $|\mu_k^d - \tau_k| < \zeta$ for $k \in \{1, \cdots, 5\}$, and $\zeta < \min_{k \in \{1, \cdots, 4\}} d_k/2$, then we have estimators $(\hat{\mu}_k^d, (\hat{\sigma}_k^d)^2, \hat{\alpha}_k^d)$ which maximize $\hat{f}(X_1, \cdots, X_{N_1}|\theta^d)$ in (2) converge to $(\mu_k^d, (\sigma_k^d)^2, \alpha_k^d)$ in
probability for all \( k = 1, \ldots, 5 \), where \( \{\alpha_{k}^{d*}\}_{k=1}^{5} \) are the proportion of 5 clusters in the sample of size \( N_1 \):

(b) If some cluster doesn’t exist in the sample with size \( N_1 \), we can get the corresponding estimator \( \hat{\alpha}_{k}^{d*} \overset{P}{\rightarrow} 0 \). Furthermore, if the existing cluster set is denoted as \( K^* \), we have \( (\hat{\mu}_{k}^{d*}, (\hat{\sigma}_{k}^{d*})^2, \hat{\alpha}_{k}^{d*}) \overset{P}{\rightarrow} (\mu_{k}^{d*}, (\sigma_{k}^{d*})^2, \alpha_{k}^{d*}) \) for all \( k \in K^* \), where \( (\mu_{k}^{d*}, (\sigma_{k}^{d*})^2) \) are also the true underlying location and scale parameters, with \( \mu_{k}^{d*} \) satisfying smaller \( \mu_{k}^{d} \) corresponding to smaller index \( k \), \( |\mu_{k}^{d*} - \tau_{k}| < \zeta \) as described in (a), \( \alpha_{k}^{d*} \) are the proportion of existing cluster \( k \in K^* \) in the sample of size \( N_1 \).

We also expect to attain some consistency outcomes as sample size tends to infinity under our empirical Bayes framework, thankfully, the consistency of BIC formula under some regularity conditions inspires much regarding to suitable choice of prior variance \((\sigma_{k}^{d*})^2/(\sigma_{k}^{d})^2\). Keribin [17] discussed the consistent estimation of BIC in selecting the order of mixture models, which implies that the difference between log-likelihood under overfitting setting and log-likelihood under true setting is no more than \( O(\log(n)) \). Furthermore, from a hypothesis testing perspective, suppose \( k_1 > k \), the likelihood ratio test (LRT) statistic for testing \( H_0: k \) mixtures against \( H_1: k_1 \) mixtures can also be expressed as the difference of the 2 log likelihood described above. When \( H_0 \) holds, this LRT statistic tends to a chi-square distribution as \( n \rightarrow \infty \). Intuitively, the order of \( 1/(\sigma_{k}^{d})^2 \) can be any increasing function of \( n \) of order less than \( O(n) \), equipping with this order is sufficient to reverse the tendency of overfitting.

**Theorem 2.4.** Suppose parameters \((\mu_{k}^{d}, (\sigma_{k}^{d})^2, \alpha_{k}^{d})\) for \( k = 1, \ldots, 5 \) lie in a closed bounded set \( \Theta \). When sample size \( N_1 \rightarrow \infty \), bin size \( p \) is a fixed constant, \((\sigma_{\tau k}^{d})^2\) are equivalent for \( k = 1, \ldots, 5 \), the order of \((\sigma_{\tau k}^{d})^2\) is \( O(1/m(N_1)) \), where \( m(N_1) \) is any increasing function of \( N_1 \) with order less than \( O(N_1) \):

(a) If the 5 clusters all exist and the underlying true parameters are \((\mu_{k}^{d*}, (\sigma_{k}^{d*})^2, \alpha_{k}^{d*})\) for all \( k = 1, \ldots, 5 \), and \( (\mu_{k}^{d}) \) satisfying \( \mu_{1}^{d*} < \mu_{2}^{d*} < \cdots < \mu_{5}^{d*} \), we further assume \( d_k = \tau_{k+1} - \tau_{k} \) for \( k \in \{1, \ldots, 4\} \), and suppose there exists an \( \zeta > 0 \) s.t. \(|\mu_{k}^{*} - \tau_{k}| < \zeta \) for all \( k \in \{1, \ldots, 5\} \), \( \zeta < \min_{k \in \{1, \ldots, 4\}} d_k/2 \), then estimators \((\hat{\mu}_{k}^{d}, (\hat{\sigma}_{k}^{d})^2, \hat{\alpha}_{k}^{d})\) maximize \( \hat{f}(X_1, \ldots, X_{N_1'; \theta^d}) \) in (2) converge to true parameters \((\mu_{k}^{d*}, (\sigma_{k}^{d*})^2, \alpha_{k}^{d*})\) in probability for all \( k = 1, \ldots, 5 \);

(b) If some cluster doesn’t exist, the corresponding estimator \( \hat{\alpha}_{k}^{d} \overset{P}{\rightarrow} 0 \), \( \hat{\mu}_{k}^{d} \overset{P}{\rightarrow} \tau_{k} \). Furthermore, if the existing cluster set is denoted as \( K^* \), we have \( (\hat{\mu}_{k}^{d*}, (\hat{\sigma}_{k}^{d*})^2, \hat{\alpha}_{k}^{d*}) \overset{P}{\rightarrow} (\mu_{k}^{d*}, (\sigma_{k}^{d*})^2, \alpha_{k}^{d*}) \) for all \( k \in K^* \), where \((\mu_{k}^{d*}, (\sigma_{k}^{d*})^2)\) are also the true underlying location and scale parameters, with \( \mu_{k}^{d*} \) satisfying smaller \( \mu_{k}^{d} \) corresponding to smaller index \( k \), \( |\mu_{k}^{d*} - \tau_{k}| < \zeta \) as described in (a), \( \alpha_{k}^{d*} \) are the proportion of existing cluster \( k \in K^* \) in the sample.

Moreover, when sample size and bin length both tend to infinity, similar consistency results can also be achieved.

**Theorem 2.5.** Under the same conditions as those in Theorem 2, apart from bin size \( p \rightarrow \infty \), \( N_1/e^p \rightarrow 0 \) and the order of \( \sigma_{\tau k}^{d} \) is \( O(1/m(N_{1; p})) \), where \( m(N_{1; p}) \) is any increasing function of \( N_1 \) and \( p \) with order less than \( O(N_{1; p}) \), the same consistency conclusions can also be obtained as Theorem 2.

From the above 3 theorems we can guarantee consistency properties of our specific empirical Bayes estimators, with the empirical Bayes framework establishing a new prior scheme different from classic prior distributions and robust to overfitting. Furthermore, by arranging the prior
mean in increasing order, letting prior variance vary with sample size and bin length, estimators \( \{\hat{\mu}_k^d\}_{k=1}^5 / \{\hat{\mu}_k^c\}_{k=1}^5 \) can be arranged increasingly as well, ensuring the proportions of case and control are tested correspondingly. Actually if not using this Bayes formula, BIC can also be applied to achieve satisfying estimators when the sample size is large enough, but when using BIC to implement model selection, the procedure needs to be executed at least \( k \) times, where \( k \) is the largest possible model size, for example in our model, we need to apply BIC at least 5 times to ensure the most suitable model. We can expect that when there are tens of thousands of models to specify, as long as we have knowledge about the rough mean, and the means are not too close to each other, doing model selection by our method can be more convenient compared with BIC.

To reduce calculation time, we conduct a conditional test, which means that we suppose the true proportion parameters for case and control both tend to infinity, the power of our likelihood ratio test tends to 1.

By the non-negativity of KL divergence, it’s not hard to deduce that when sample sizes of case and control both tend to infinity, the type II error of likelihood ratio test tends to 0.

**Proposition 2.6.** Suppose the true proportion parameters for case and control are \( \alpha_k^d \) and \( \alpha_k^c \), \( k = 1, \cdots, 5 \). If there exists a constant \( c_b > 0 \), \( \exists k \in \{1, \cdots, 5\} \), s.t. \( |\alpha_k^d - \alpha_k^c| > c_b \), when sample size \( N_1 \to \infty \), \( N_2 \to \infty \), the type II error of likelihood ratio test tends to 0.

### 2.4. Deal with heterogeneity across genome

Let’s revisit our model formulation in section 2.1, where we model 2 isotropous multivariate Gaussian mixtures for case and control samples, which means that for each CN cluster, its mean and variance parameters are the same across locations in the bin. But due to GC-content, sequencing bias or other factors, for each bin of length \( p \), it’s more reasonable to replace \( N(\mu_k^d 1_p, (\sigma_k^d)^2 1_p) \) by \( N(\mu_k^d, \text{Diag}((\sigma_k^d)^2)) \) to account for heterogeneity.
Suppose parameters of prior distribution, we can also obtain consistency conclusions as described in section 2.3. To CN cluster \( k \) length \( p \) multivariate Gaussian mixtures for case and control sample:

\[
X_i \sim \sum_{k=1}^{5} \alpha_k d N(\mu_k d, Diag((\sigma_k d)^2)), \quad i = 1, \ldots, N_1 \text{ for case samples in a bin}
\]

\[
Y_j \sim \sum_{k=1}^{5} \alpha_k c N(\mu_k c, Diag((\sigma_k c)^2)), \quad j = 1, \ldots, N_2 \text{ for control samples in a bin}
\]

Accordingly, we set a Gaussian prior for every element in the mean parameter \( \mu_k d, \mu_k c \) for \( k = 1, \ldots, 5 \), take case samples for an example, the prior density of \( \mu_k d \) is

\[
g(\mu_k d) = \prod_{j=1}^{p} g(\mu_{k j} d | \tau_k, (\sigma_{\tau k})^2)
\]

Under this updated empirical Bayes framework, take case samples as an instance, the likelihood used to conduct parameter estimation in a bin of length \( p \) can be reformulated as:

\[
\tilde{f}(X_1, \ldots, X_{N_1} | \theta d) \triangleq l(\theta d | X_1, \ldots, X_{N_1}) \prod_{k=1}^{5} g(\mu_k d) = \prod_{j=1}^{p} \prod_{k=1}^{5} \alpha_k d f(X_i | \mu_k d, (\sigma_k d)^2) \]

(9)

Analogous to what we have reached in section 2.2 and 2.3, the EM updating rule is

\[
\mu_{k j} d(t + 1) = \frac{(\sigma_{\tau k})^2 \sum_{i=1}^{N_1} (X_{ij} b_{ik} d(t)) + \tau_k (\sigma_{k j} d(t))^2}{(\sigma_{\tau k})^2 \sum_{i=1}^{N_1} b_{ik} d(t) + (\sigma_{k j} d(t))^2}
\]

\[
(\sigma_{k j} d(t + 1))^2 = \frac{\sum_{i=1}^{N_1} (X_{ij} - \mu_{k j} d(t + 1))^2 b_{ik} d(t)}{\sum_{i=1}^{N_1} b_{ik} d(t)}
\]

(10)

for \( k = 1, \ldots, 5, j = 1, \ldots, p \), where \( X_{ij} \) is the \( j \)-th observation of case sample \( i \) in a bin of length \( p \), \( b_{ik} d(t) = P(Z_i = k | X_i, \theta d(t)) \) is denoted as the posterior probability of sample \( i \) belongs to CN cluster \( k \) under the current parameter estimation.

Furthermore, under this anisotropic multivariate Gaussian mixture setting, with suitable choice of prior distribution, we can also obtain consistency conclusions as described in section 2.3.

**Theorem 2.7.** Suppose parameters \( (\mu_k d, (\sigma_k d)^2, \alpha_k d) \) for \( k = 1, \ldots, 5 \) lie in a closed bounded set \( \Theta \), \( (\sigma_{\tau k})^2 \) are equivalent for \( k = 1, \ldots, 5 \), the order of \( (\sigma_{k j} d)^2 \) is \( O(1/m(N_1)) \), where \( m(N_1) \) is any increasing function of \( N_1 \) with order less than \( O(N_1) \), when sample size \( N_1 \to \infty \), for \( \forall p = 1, 2, 3, \ldots \).
(a) if the 5 clusters all exist in the sample, the underlying true location and scale parameters are \((\mu_k^{d*}, (\sigma_k^{d*})^2)^{5}_{k=1}\) with \(\{\mu_k^{d*}\}^{5}_{k=1}\) satisfy \(\mu_1^{d*} < \mu_2^{d*} < \cdots < \mu_5^{d*}\) for \(\forall j = 1, \cdots, p\), we further assume \(d_k = \tau_{k+1} - \tau_k\) for \(k \in \{1, \cdots, 4\}\), and suppose there exists an \(\zeta > 0\) s.t. \(|\mu_k^{d*} - \tau_k| < \zeta\) for \(k \in \{1, \cdots, 5\}, \forall j = 1, \cdots, p\) and \(\zeta < \min_{k \in \{1, \cdots, 4\}} d_k/2\), then we have estimators \((\hat{\mu}_k^{d}, (\hat{\sigma}_k^{d})^2, \hat{\alpha}_k^{d})\) maximize \(\bar{f}(X_1, \cdots, X_N, |\theta|)\) in (9) converge to \((\mu_k^{d*}, (\sigma_k^{d*})^2, \alpha_k^{d*})\) in probability for all \(k = 1, \cdots, 5\), where \(\{\alpha_k^{d*}\}^{5}_{k=1}\) are the true proportions of 5 clusters in the sample;

(b) if some cluster doesn’t exist in the sample, which forms a set \(K^{sc}\), we can get the corresponding \(\hat{\alpha}_k^{d} \xrightarrow{p} 0, \hat{\mu}_k^{d} \xrightarrow{p} \tau_k\) for \(\forall j = 1, \cdots, p\) and \(\forall k \in K^{sc}\). Furthermore, if the existing cluster set is denoted as \(K^*\), we have \((\hat{\mu}_k^{d}, (\hat{\sigma}_k^{d})^2, \hat{\alpha}_k^{d}) \xrightarrow{p} (\mu_k^{d*}, (\sigma_k^{d*})^2, \alpha_k^{d*})\) for \(\forall k \in K^*, j = 1, \cdots, p\), where \((\mu_k^{d*}, (\sigma_k^{d*})^2), k \in K^*\) are also true underlying location and scale parameters with \(\mu_k^{d*}\) satisfy smaller \(\mu_k^{d*}\) corresponding to smaller index \(k\), \(|\mu_k^{d*} - \tau_k| < \zeta\) for \(k \in K^*, \forall j = 1, \cdots, p\) as described in (a), \(\alpha_k^{d*}\) are the proportion of existing cluster \(k \in K^*\) in the sample.

2.5. GWAS accounting for deletion or duplication only. In practice, when encountering significant regions after conducting GWAS in CNV level, we often want to know which type of variation accounts for the significance most. For example, in a certain significant region, if duplication is predominante compared with deletion and duplication carriers in cancer samples are more than those in normal samples, we suspect this region has potential characteristics of proto-oncogene. In this section we propose a new test formula which accounts for differences between case and control samples induced by deletion or duplication only.

If we consider influence induced only by deletion, we can perform a likelihood ratio test of the following:

\[
H_0 : \alpha_1^d = \alpha_1^c, \alpha_2^d = \alpha_2^c \\
H_1 : \alpha_1^d \neq \alpha_1^c \text{ or } \alpha_2^d \neq \alpha_2^c
\]

(11)

where \(H_1\) is the same as before, whereas \(H_0\) only restricts equivalence of proportion on clusters corresponding to deletion, proportion parameters on other 3 clusters are unconstrained, so the likelihood ratio value stands for extent of deletion’s influence. Duplication can be processed similar as deletion:

\[
H_0 : \alpha_4^d = \alpha_4^c, \alpha_5^d = \alpha_5^c \\
H_1 : \alpha_4^d \neq \alpha_4^c \text{ or } \alpha_5^d \neq \alpha_5^c
\]

(12)

where \(\alpha_4^d/\alpha_4^c, \alpha_5^d/\alpha_5^c\) are proportion parameters corresponding to 2 duplication clusters.

Parameter estimation under \(H_1\) can be obtained by iterating according to (10), for ease of calculation, we still conduct a conditional likelihood ratio test, assuming location parameters \(\mu_k^c\), \(\mu_k^d\) and scale parameters \((\sigma_k^c)^2, (\sigma_k^d)^2\) for \(k = 1, \cdots, 5\) are the same under \(H_0\) and \(H_1\). The EM
updating rule of proportion parameters under $H_0$ of (11) is:

$$
\alpha_k^d(t + 1) = \alpha_k^c(t + 1) = \frac{\sum_{i=1}^{N_1} b_{ik}^d(t) + \sum_{j=1}^{N_2} b_{jk}^c(t)}{N_1 + N_2}; k = 1, 2
$$

$$
\alpha_k^d(t + 1) = \frac{\sum_{i=1}^{N_1} b_{ik}^d(t)}{\lambda_1}, \lambda_1 = (N_1 + N_2) \frac{\sum_{i=1}^{N_1} \sum_{k=1}^{5} b_{ik}^d(t) + \sum_{j=1}^{N_2} b_{jk}^c(t)}{\sum_{i=1}^{N_1} \sum_{k=1}^{5} b_{ik}^d(t) + \sum_{j=1}^{N_2} b_{jk}^c(t)}; k = 3, 4, 5
$$

As for duplication, the updating rule is analogous to deletion:

$$
\alpha_k^d(t + 1) = \alpha_k^c(t + 1) = \frac{\sum_{i=1}^{N_1} b_{ik}^d(t) + \sum_{j=1}^{N_2} b_{jk}^c(t)}{N_1 + N_2}; k = 4, 5
$$

$$
\alpha_k^d(t + 1) = \frac{\sum_{i=1}^{N_1} b_{ik}^d(t)}{\lambda_1'}, \lambda_1' = (N_1 + N_2) \frac{\sum_{i=1}^{N_1} \sum_{k=1}^{3} b_{ik}^d(t) + \sum_{j=1}^{N_2} b_{jk}^c(t)}{\sum_{i=1}^{N_1} \sum_{k=1}^{3} b_{ik}^d(t) + \sum_{j=1}^{N_2} b_{jk}^c(t)}; k = 1, 2, 3
$$

$$
\alpha_k^c(t + 1) = \frac{\sum_{j=1}^{N_2} b_{jk}^c(t)}{\lambda_2'}, \lambda_2' = (N_1 + N_2) \frac{\sum_{j=1}^{N_2} \sum_{k=1}^{3} b_{jk}^c(t)}{\sum_{j=1}^{N_2} \sum_{k=1}^{3} b_{jk}^c(t)}; k = 1, 2, 3
$$

where $b_{ik}^d(t) = P(Z_i = k|X_i, \theta^d(t))$ and $b_{jk}^c(t) = P(Z_j = k|Y_j, \theta^c(t))$ are posterior probability of sample $i$ in case and sample $j$ in control belong to CN cluster $k$ under the current parameter estimation correspondingly.

2.6. **Merge.** So far we can implement the “equivalence of proportion between case and control” test bin by bin along the whole chromosome, claiming the significant bins we are interested in. From theorem 1 we can see that in the isotropous multivariate Gaussian mixture case, the longer the bin is, the more accurate parameter estimation we can get, under the premise that each sample don’t have CN state change point at each bin. Specifically, when it comes to the sparse signal detection problem, we can prove the following theorem.

**Theorem 2.8.** When the bin size $p$ and sample size $n$ tend to $\infty$, and assume $\frac{\log(n)}{p} \rightarrow 0$, $\epsilon = n^{-\beta}$ for some $\beta \in (0, 1)$, $A = \sqrt{2r \log(n)}$ for $r \in (0, 1)$, regarding to the signal detection test

$$
H_0 : X_i \sim N(0_1, \sigma_1^2 I_1), \quad 1 \leq i \leq n
$$

$$
H_1 : X_i \sim (1 - \epsilon)N(0_1, \sigma_1^2 I_1) + \epsilon N(A_1, \sigma_2^2 I_1), \quad 1 \leq i \leq n
$$

(15)
the sum of type I and type II error of the likelihood ratio test tends to 0.

It concludes that even if the signal is very sparse, when the bin size tends to ∞, the likelihood under $H_0$ and $H_1$ can be separated completely. The detection boundary of this kind of test has been analysed by Tony et al. [10], whereas they focused on univariate conditions, and the variance under $H_0$ by their assumption is simply set to 1.

Owing to the increased accuracy brought by the longer bin, we want to further enhance our testing performance by merging. The adjacent bins possessing similar distributions can be merged together forming a larger bin, we expect test carried out in the larger bin will be more powerful. The intuition of evaluating whether two adjacent bins can be merged is summarized as following: we consider case samples first, suppose the $b$-th bin of length $p_1$ and the $(b+1)$-th bin of length $p_2$ can be merged together, we can estimate parameters in the newly formed bin $b-(b+1)$ by maximizing $\hat{f}(X_{1,b-(b+1)}; \cdots, X_{N_1,b-(b+1)}|\theta^{d}_{b-(b+1)})$ in (9), here $X_{1,b-(b+1)}; \cdots, X_{N_1,b-(b+1)}$ are $(p_1+p_2)$-dimensional sample, representing observations in the $b$-th and $(b+1)$-th bin. We further suppose the estimators we obtained in the merged bin is $\hat{\theta}^{d}_{b-(b+1)} = (\hat{\alpha}^{d}_{b-(b+1),1}, \cdots, \hat{\alpha}^{d}_{b-(b+1),5}, \hat{\sigma}^{d}_{b-(b+1),1}^2, \cdots, \hat{\sigma}^{d}_{b-(b+1),5}^2)$, and pick up parameters corresponding to the $b$-th bin and $(b+1)$-th bin separately, suppose they are denoted as $\hat{\theta}^{d(1)}_{b-(b+1)}$ and $\hat{\theta}^{d(2)}_{b-(b+1)}$, with $\hat{\theta}^{d(i)}_{b-(b+1)} = (\hat{\alpha}^{d(i)}_{b-(b+1),1}, \cdots, \hat{\sigma}^{d(i)}_{b-(b+1),5})$, $i = 1, 2$,

\[
\hat{\mu}^{d(1)}_{b-(b+1),1} \cdots, \hat{\mu}^{d(1)}_{b-(b+1),1}, \cdots, \hat{\sigma}^{d(1)}_{b-(b+1),1} \cdots \hat{\sigma}^{d(1)}_{b-(b+1),5} \cdots \hat{\sigma}^{d(1)}_{b-(b+1),5}, \hat{\sigma}^{d(2)}_{b-(b+1),1}, \cdots, \hat{\sigma}^{d(2)}_{b-(b+1),5}
\]

are the first $p_1$-dimension of $\hat{\mu}^{d}_{b-(b+1),1} \cdots, \hat{\mu}^{d}_{b-(b+1),1}, \cdots, \hat{\sigma}^{d}_{b-(b+1),5}$ and $\hat{\mu}^{d(2)}_{b-(b+1),1} \cdots, \hat{\sigma}^{d(2)}_{b-(b+1),1}, \cdots, \hat{\sigma}^{d(2)}_{b-(b+1),5}$ are the last $p_2$-dimension part. If not merge, we can also get estimators $\hat{\theta}^{d}_{b} \text{ and } \hat{\theta}^{d}_{b+1}$ by maximizing $\hat{f}(X_{1b}; \cdots, X_{N_1b}|\theta^{d}_{b})$ and $\hat{f}(X_{1,b+1}; \cdots, X_{N_1,b+1}|\theta^{d}_{b+1})$ separately. If for every case sample, their background CN states are the same in these 2 adjacent bins, which is ideal for merging, we can expect the following value can be small:

\[
M^{d}_{b,b+1} = \sum_{i=1}^{N_1} \log[h(X_{i,b}|\theta^{d}_{b})] - \sum_{i=1}^{N_1} \log[h(X_{i,b}|\hat{\theta}^{d(1)}_{b-(b+1)})]
\]

\[
+ \sum_{i=1}^{N_1} \log[h(X_{i,b+1}|\theta^{d}_{b+1})] - \sum_{i=1}^{N_1} \log[h(X_{i,b+1}|\hat{\theta}^{d(2)}_{b-(b+1)})]
\]

as mentioned above, $h(X|\theta)$ denotes mixture Gaussian distribution with parameter $\theta$. Situations in control samples can be analogous, we can obtain a similar $M^{c}_{b,b+1}$ determining whether it’s suitable to merge or not in control samples.

If the background CN states are different in the 2 adjacent bins for some sample, the $(p_1+p_2)$-dimensional mixture Gaussian is the misspecified model, resulting in $\hat{\theta}^{d(1)}_{bm} \text{ and } \hat{\theta}^{d(2)}_{bm}$ are not fine estimators, the value $M^{d}_{b,b+1} \text{ or } M^{c}_{b,b+1}$ described above can be large. Theoretically, take case samples as an instance, we can give the order of $M^{d}_{b,b+1}$ when the $b$-th and $b+1$-th bins can or can’t be merged together.

Theorem 2.9. Suppose proportion parameters $(\alpha^{d}_{b+1}, \cdots, \alpha^{d}_{b+5})$ and $(\alpha^{d}_{b+1}, \cdots, \alpha^{d}_{b+5})$ are the same in the $b$-th and $(b+1)$-th bin if and only if the background CN states are the same in these 2 bins for every sample.
(a) When the proportion parameters are the same in the b-th and (b + 1)-th bin, the order of $M_{b,b+1}^d$ is $O_p(m(N_1p))$, where $p = p_1 + p_2$ is the length of the merged bin $b - (b + 1)$, $m(N_1p)$ is any increasing function of $N_1, p$ with order less than $O(N_1p)$.

(b) If there exists a constant $c_b > 0$, $\exists k \in \{1, \cdots, 5\}$, s.t. $|\alpha_{b,b}^d - \alpha_{(b+1),k}^d| > c_b$, the order of $M_{b,b+1}^d$ is $O_p(N_1)$.

Form the above order estimation, we set the criterion of whether the 2 adjacent bins can or can’t be merged together as: $M_{b,b+1}^d < \lambda_d \log(N_1p)$ and $M_{b,b+1}^c < \lambda_c \log(N_2p)$, then we will merge, else we will not merge. In practice, $\lambda_c$ and $\lambda_d$ can be chosen proportional to sample size, for example, if $N_2 = 2N_1$, we can set $\lambda_c = 2\lambda_d$.

In general, we carry out the merge process by the following iterating steps:

1. Calculate $M_{b,b+1}^d$ and $M_{b,b+1}^c$ for every adjacent bins after the bin-by-bin test.
2. Select adjacent bins with the smallest $M_{b,b+1}^d + M_{b,b+1}^c$ value to merge.
3. Update parameters in the newly formed bin by rule (10).
4. Update $M_{b,b+1}^d$ and $M_{b,b+1}^c$ with respect to the newly formed bin and its adjacent bins.
5. Repeat steps 2-4 until there are no adjacent bins satisfying $M_{b,b+1} < \lambda_d \log(N_1p)$ and $M_{b,b+1} < \lambda_c \log(N_2p)$.

Notably, when there exists samples whose underlying CN states are not the same in the 2 adjacent bins, mixture Gaussian is no longer suitable, under the premise that each Gaussian corresponds with only one CN state. For instance, if in the b-th bin, $X_b \sim N_0(0_{p_1}, I_{p_1})$, while in the $(b + 1)$-th bin, $X_{i,b+1} \sim 0.8 N(0_{p_2}, I_{p_2}) + 0.2 N(0.4I_{p_2}, I_{p_2})$, then $0.8 N(0_{p_1+p_2}, I_{p_1+p_2}) + 0.2 N((0_{p_1}, 0.4I_{p_2}), I_{p_1+p_2})$ is the true distribution for $X_{i,(b-1)}$, but we regard this form as meaningless. When dealing with parameter estimation in merging, we can obtain a meaningful mixture Gaussian by enhancing the power of prior, until location parameter in each CN cluster is located in a reasonable region. According to our experience, location parameters of CN=0 and 1 are restricted to be smaller than -0.9 and -0.4; location parameters of CN=3 and 4 are supposed to be larger than 0.35 and 0.65.

After merging, what we have in hand are longer bins with their corresponding parameters estimated simultaneously in the merging process, which means that parameters under $H_1$ have been well prepared. We only need to estimate proportion parameters under the constraint of $H_0$, that is we do the conditional test again as mentioned in section 2.3—conduct the “equivalence of proportion between case and control” test at each newly formed bin conditional on the known mean and variance parameters of each CN cluster. Estimation of proportion parameters under $H_0$ is analogous with the EM updating rule mentioned in (8). If we focus on single factor—deletion or duplication, rule (13) or (14) can be executed until convergence.

3. SIMULATION

We undertake simulation on 1000 case samples and 2000 control samples on $10^4$ base pairs (bp) sites, after generating 2 matrices of size $1000*10^4$ and $2000*10^4$, with each element sampled from $N(0, 0.3^2)$, we pick up 30 segments to add signals as our interested CNV-GWAS regions. These 30 segments are equally separated into 5 different lengths—10 bp, 30bp, 50bp, 100 bp, 500 bp and 2 types of CNV—deletion and duplication.

We designed 3 different case-control CNV carriers’ proportion contrast settings for both deletion and duplication (table 1-6). Scenario 1 of deletion or duplication stands for a rare CNV setting, with low frequency of observing a CNV carrier, scenario 2 belongs to a middle frequency setting,
while scenario 3 stands for a common CNV setting, with high frequency of observing a CNV carrier. We select carriers with a certain CN state at random and satisfy the proportion settings on table 1-6.

To explore performance affected by batch effects between case and control, we generate some case and control sample with batch effects among them. Specifically, in a CNV-GWAS region, for a certain CN state, case and control are sampled from different Gaussian distribution. Parameters of the Gaussian distribution under different CN state are shown in table 7. If we add no batch effects, parameters under different CN state of case and control are sampled both according to the second row of the table.

We also want to know the robustness of our method under different variance settings, for example, if there’s no CNV carrier in a region, but the variance of this region is very large, there may be a tendency of increased false discovery rate. In the whole chromosome, besides the 30 CNV-GWAS regions, what’s left are 31 normal regions, we select 2 regions of length 20 bp and 200 bp on the 1st and 2nd normal regions separately, they have no CNV carrier both in case and control samples, but we simulate case sample from a Gaussian distribution with larger variance. Concretely, in these 2 regions, case sample are from $N(0, 0.9^2)$, while control sample are from $N(0, 0.6^2)$, we call these 2 regions vary-var1 regions. For the remaining 29 normal regions, we pick up 14 regions to let variance vary not only between case and control, but also among different locations of this region, we call these 14 regions vary-var2 regions. For each vary-var2 region, we split it into equal 10 bp-length bins with index 1, · · · , B if it has 10B bp, bins are drawn from the same distribution if they have the same remainder when their indexes divided by 3. Table 8 shows detailed sampling rules on 14 vary-var2 regions. If we add no different variance rule, elements from both vary-var1 and vary-var2 regions are sampled from $N(0, 0.3^2)$.

In total, we can do simulation under 12 different scenarios (CNV carrier’s frequency low/middle/high × have/not have batch effects × have/not have different variance settings). For each scenario, we randomly simulate 50 case and control data sets, they have the same 30 CNV-GWAS regions.

We compared our method with several CNV calling methods—CBS [5], Median [20], multi-CBS [6]. Circular binary segmentation (CBS) [5] is a very popular and classic method in detecting CNVs on a single chromosome, Cai et al. (2012) [20] performed a median transformation and then use the transformed data to establish a test statistic to call CNV. Zhang et al. [6] generalized CBS into multiple sample setting, aiming at detecting CNV regions shared between multiple samples. We applied these methods on our simulated case and control data sets.

Figure 1 shows the performance in terms of sensitivity and FDR of our method and other 3 methods mentioned above. As is well-known that sensitivity is the proportion of true positives among all real positive incidents, and FDR is the proportion of false positives among all detected incidents. We use the overlap of detected CNVs and designed CNVs to infer true and false discovered CNVs for each sample. For a putative CNV detected by a given method, if the overlapping region is more than half of a designed CNV and more than half of this putative CNV, this designed CNV is claimed to be detected in the current sample and we define it to be a true positive. If the overlapping region with any of the designed CNVs is less than half of the discovered CNV, we define it to be a false positive.

From figure 1 we can see that our method outperforms CBS and Median in all simulation scenarios, it performs better than multiCBS or comparable with multiCBS in terms of sensitivity. Our method is robust to batch effects and variance influences. In terms of sensitivity, multiCBS
is comparable with our method in some scenarios, but it can be severely impacted by varying variance, when the variance of some regions are very large, it misaligns normal samples to other abnormal CN states and results in large FDR value. The power of CBS and Median is lower than multiCBS and our method in every scenario, but these 2 methods seems to be more robust to varying variance than multiCBS as shown in the FDR plot.

As for our aim of finding significant regions in CNV-GWAS, we pick up regions with P value less than 0.05 after merging and regard these as CNV-GWAS significant regions. We compared these discovered regions with 30 designed CNV-GWAS regions, and evaluate our method’s performance in terms of sensitivity and FDR as defined above, figure 2 shows sensitivity of our method with different lengths, batches and variance settings. Figure 3 shows performance in terms of FDR under different settings.

As can be seen from figure 2, apart from some short CNV-GWAS regions in low frequency setting, our method can detect almost all significant regions, when there is no batch effect between case and control sample, no region has varying variance, short regions can be detected with larger power. Figure 3 shows that FDR under all circumstances approaches 0, varying variance seems to have a larger impact on the performance than batch effects. This result verifies effectiveness of our method again.

| CN:0 | CN:1 | CN:2 | CN:3 | CN:4 |
|------|------|------|------|------|
| case | 0    | 0.05 | 0.95 | 0    |
| control | 0 | 0    | 1    | 0    |

**Table 1.** case-control CNV carrier’s proportion contrast, deletion’s scenario 1

| CN:0 | CN:1 | CN:2 | CN:3 | CN:4 |
|------|------|------|------|------|
| case | 0.03 | 0.1  | 0.87 | 0    |
| control | 0.01 | 0.01 | 0.98 | 0    |

**Table 2.** case-control CNV carrier’s proportion contrast, duplication’s scenario 1

| CN:0 | CN:1 | CN:2 | CN:3 | CN:4 |
|------|------|------|------|------|
| case | 0    | 0    | 0.95 | 0.05 |
| control | 0 | 0    | 1    | 0    |

| CN:0 | CN:1 | CN:2 | CN:3 | CN:4 |
|------|------|------|------|------|
| case | 0.17 | 0.46 | 0.37 | 0    |
| control | 0.08 | 0.28 | 0.64 | 0    |

**Table 3.** case-control CNV carrier’s proportion contrast, deletion’s scenario 2

| CN:0 | CN:1 | CN:2 | CN:3 | CN:4 |
|------|------|------|------|------|
| case | 0.03 | 0.1  | 0.87 | 0    |
| control | 0.01 | 0.01 | 0.98 | 0    |

**Table 4.** case-control CNV carrier’s proportion contrast, duplication’s scenario 2

| CN:0 | CN:1 | CN:2 | CN:3 | CN:4 |
|------|------|------|------|------|
| case | 0    | 0    | 0.95 | 0.1  |
| control | 0 | 0    | 0.98 | 0.01 |

| CN:0 | CN:1 | CN:2 | CN:3 | CN:4 |
|------|------|------|------|------|
| case | 0.17 | 0.46 | 0.37 | 0    |
| control | 0.08 | 0.28 | 0.64 | 0    |

**Table 5.** case-control CNV carrier’s proportion contrast, deletion’s scenario 3

| CN:0 | CN:1 | CN:2 | CN:3 | CN:4 |
|------|------|------|------|------|
| case | 0    | 0    | 0.95 | 0.05 |
| control | 0 | 0    | 1    | 0    |

**Table 6.** case-control CNV carrier’s proportion contrast, duplication’s scenario 3
Figure 1. Average sensitivity and FDR with error bar of different method under different settings, height of column denotes mean of sensitivity/FDR, height of error bar denotes mean+standard deviation of sensitivity/FDR. “low”, “mid” and “high” refer to low frequency, middle frequency and high frequency. (A) Sensitivity of different methods under different batches or variance settings of case samples. (B) Sensitivity of different methods under different batch or variance settings of control samples. (C) FDR of different methods under different batch or variance settings of case samples. (D) FDR of different methods under different batch or variance settings of control samples.
**Figure 2.** Average sensitivity with error bar in detecting significant regions in CNV-GWAS under different frequencies, batches and variance settings of our method. Height of column denotes mean of sensitivity, height of error bar denotes mean+standard deviation of sensitivity. x-axis denotes different lengths of CNV, y-axis denotes sensitivity. “low”, “mid”, and “high” refer to low frequency, middle frequency and high frequency.

|        | CN=0 | CN=1 | CN=2 | CN=3 | CN=4 |
|--------|------|------|------|------|------|
| \( \mu \) of case | -1.3 | -0.4 | -0.13 | 0.3 | 0.63 |
| \( \mu \) of control | -1.2 | -0.3 | 0.15 | 0.44 | 0.73 |
| \( \sigma \) of case | 0.5 | 0.13 | 0.3 | 0.18 | 0.4 |
| \( \sigma \) of control | 0.52 | 0.14 | 0.31 | 0.2 | 0.42 |

**Table 7.** Gaussian parameters on CNV-GWAS regions when there’s batch effect

4. **Real Data**

We applied our method to real data with 2042 esophagus cancer (ESCC) samples and 2060 normal (control) samples, these data are from array comparative genomic hybridization (aCGH) and their original form are CEL files. We pool case and control samples together and run the PennCNV-Affy procedure [18] to process raw CEL files, PennCNV-Affy can generate canonical genotype clusters and then convert signal intensity for each sample to log2 ratio values, we also implement a wave adjustment procedure [19] to filter biases caused by GC content. After processed
by PennCNV-Affy, we have case and control matrices with each element a wave-adjusted log2 ratio value, which can be used to do downstream CNV-GWAS analysis.

To increase the reliability of results, we split both case and control samples into 2 sets with equal size—test set and validation set. On the test set we implement the whole procedure of our method and find out suspicious disease-risking regions, for the validation set, we only conduct testing procedure on each segment returned by the merge outcome on test set, if the test also returns a significant p-value, we claim this region a disease-risking region. We perform test returning
FIGURE 4. Scatter plot of log 10 P-value of deletion or duplication on each segment after merging, left plot is deletion, right plot is duplication. Each point refers to a segment with abscissa and ordinate corresponding to log 10 P-value in test and validation set separately.

impact value explained by deletion or duplication only, i.e. we implement test (11) and (12) for every segment in test and validation set.

Figure 4 depicts the log 10 p-value for every segment after merging, the Pearson’s correlation of log 10 p-value returned by test set and validation set are 0.91 and 0.82 for deletion and duplication, indicating the robustness of our method. Figure 5 depicts the signals corresponding to the most significant point in the deletion’s log 10 P-value plot, with (a) and (b) referring to signals of case and control separately. For the overall 2060 control samples, we find 875 copy number deletion carriers, figure 4(b) shows CNV signals of all these 875 samples, the “mid” part is the segment of interest with largest p-value in terms of deletion, we also plot the adjacent “left” and “right” segments to demonstrate reliability of our merge method. For ease of comparison, in figure 5(A) we also plot 875 case sample’s CNV signals, which correspond to the smallest 875 signal intensities in this segment, here we define mean of log 2 ratio value as the signal intensity in a segment. This significant segment is not contained in the well-known RefSeqGene set, and this deletion might play a protective role.

In the real data analysis, we come across many situations that some CNV regions are very long, but due to small proportion of carriers, it’s very hard to detect them, these long regions may have important biological implications. So besides carrying out the likelihood ratio test as mentioned in our method, we propose a new test statistic to deal with long segments, aiming of detecting long CNV regions even if carriers are rare.

For a certain segment, suppose we have estimated parameters for both case and control samples under $H_0$ and $H_1$, all case and control parameters under $H_0$ are denoted as $\hat{\theta}_H$ and $\hat{\theta}_c$, $\hat{\theta}_{H_0}$ and $\hat{\theta}_{H_1}$ are parameters under $H_1$. Suppose there are $B$ bins in this segment, we can
assign each bin \( b \) a group of parameters \( \hat{\theta}_{bH_0}^d, \hat{\theta}_{bH_0}^c, \hat{\theta}_{bH_1}^d \) and \( \hat{\theta}_{bH_1}^c \). For example, if \( \hat{\theta}_{H_0}^d = (\hat{\sigma}^d_{H_0,1}, \ldots, \hat{\sigma}^d_{H_0,5}, \hat{\mu}^d_{H_0,1}, \ldots, \hat{\mu}^d_{H_0,5}, \hat{\sigma}^d_{H_0,1}, \ldots, \hat{\sigma}^d_{H_0,5})^2 \), we can denote \( \hat{\theta}_{bH_0}^d \) as \(( \hat{\sigma}^d_{H_0,1}, \ldots, \hat{\sigma}^d_{H_0,5}, \hat{\mu}^d_{bH_0,1}, \ldots, \hat{\mu}^d_{bH_0,5}, (\hat{\sigma}^d_{bH_0,1})^2, \ldots, (\hat{\sigma}^d_{bH_0,5})^2) \), where \( \hat{\mu}^d_{bH_0,1}, \ldots, \hat{\mu}^d_{bH_0,5}, (\hat{\sigma}^d_{bH_0,1})^2, \ldots, (\hat{\sigma}^d_{bH_0,5})^2 \) are the \((5b - 4)\)-th to \(5b\)-th elements of \( \hat{\mu}^d_{H_0,1}, \ldots, \hat{\mu}^d_{bH_0,5}, (\hat{\sigma}^d_{H_0,1})^2, \ldots, (\hat{\sigma}^d_{H_0,5})^2 \). correspondingly, under this premise that the whole chromosome is split into bins with equal length 5. If case and control samples in this segment are \( X_1, \ldots, X_{N_1}, Y_1, \ldots, Y_{N_2} \) with \( X_i = (X_{i1}, \ldots, X_{ib}) \), \( Y_j = (Y_{j1}, \ldots, Y_{jb}) \) for case sample \( i = 1, \ldots, N_1 \), control sample \( j = 1, \ldots, N_2 \), we define the new test statistic as:

\[
\Lambda_{sum} = \sum_{i=1}^{N_1} \sum_{b=1}^{B} [h(X_{ib}|\hat{\theta}_{bH_1}^d) - h(X_{ib}|\hat{\theta}_{bH_0}^d)] + \sum_{j=1}^{N_2} \sum_{b=1}^{B} [h(Y_{jb}|\hat{\theta}_{bH_1}^c) - h(Y_{jb}|\hat{\theta}_{bH_0}^c)]
\]

here \( h(X_{ib}|\hat{\theta}_{bH_0}^d) \) is the mixture Gaussian distribution with parameters \( \hat{\theta}_{bH_0}^d \), other functions \( h \) are analogous. It’s obvious that this test statistic accumulates case-control difference bin by bin, increasing the power of detecting long CNVs with rare carriers, we call this statistic as sumed likelihood ratio statistic. Here we can also replace proportion parameters under \( H_0 \) by parameters derived from (13) or (14) to explicate the influence of deletion or duplication only.

Figure 6 shows an example of a CNV segment that can’t be detected by the likelihood ratio test but returns a significant value using the sumed likelihood ratio statistic. This region contains 105 probes and overlaps with 116kb length genome region, we detect 43 and 10 carriers among case and control samples separately, after intersecting with RefSeqGene set, it overlaps with 3 genes—PKP1, TNNT2, LAD1. High abundance of LAD1 have been reported to be associated with breast cancer [21], PKP1 is also regarded as a biomarker in several cancers [22, 23]. Both these 2 genes are highly expressed in esophagus, in accordance with cancer type of case samples.

To give an overall picture of the resulting significant regions, we find out genes corresponding to each significant region by intersecting with RefSeqGene set, and reserve those genes whose proportion parameter of deletion or duplication in case is larger than that in control samples. We calculate p-value using both likelihood ratio statistic and sumed likelihood ratio statistic, focusing on influence derived by deletion or duplication only. For each CNV type, 2 significant gene sets \( G_1 \) and \( G_2 \) can be obtained with p-value returned by 2 statistics, we then perform enrichment analysis by computing overlaps of these 2 gene sets with known pathway genes separately, giving us the knowledge that on which pathways gene set are enriched. We select the top 20 pathways with largest gene set enrichment for \( G_1 \) and \( G_2 \), then we pick up pathways emerged in both top 20 pathways of \( G_1 \) and \( G_2 \), table 9, 10 presents the overlap of enriched pathways for \( G_1 \) and \( G_2 \) in each CNV type.

From table 9, we find that genes which have more deletions in cancer samples are enriched in pathways such as cell cycle and immune system, indicating destruction of cell cycle and immune system promotes tumor progression. As for table 10, pathways indicating activities during cancer process have been found, such as cell motility, cell projection organization and cell part morphogenesis, in accordance with the duplication CNV type.

5. DISCUSSION

In this paper, we developed a method aiming at finding disease-risking regions related to CNV’s disproportionately distributed between case and control samples. To overcome the batch effect and heterogeneity, we proposed a test formula only focusing on testing the equivalence of
Figure 5. Signals corresponding to the most significant point in the deletion’s log 10 P-value plot. (A) and (B) refer to signals of case and control, ”mid” part depicts signals of the most significant segment, we also plot its left and right segments for comparison.

Figure 6. A segment contains genes PKP1, TNNT2, LAD1 which can be detected by summed likelihood ratio statistics but can’t be detected by likelihood ratio test.
Gene Set Name | Representative genes
---|---
GO catalytic complex | ERC1, DNAH12, PPP2CB
GO microtubule cytoskeleton | PRKAR2A, CSNK1D, SPAST
GO positive regulation of cellular biosynthetic process | TET1, KLF5, CDH13
GO ribonucleotide binding | ATAD5, LAT52, APC
GO adenyl nucleotide binding | RAD51, ACTR2, HSPD1
GO cell cycle process | BRCA1, KAT2B, DNA2
GO RNA binding | CDC40, XPO1, SMG6
GO chromosome | PRIM1, KAT8, RIF1
GO chromosome organization | KMT2C, TAF9, PCGF6
GO microtubule organizing center | MAPRE1, PPP4R2, CTNNBL1
GO peptidyl amino acid modification | ERBB4, ARID4B, TGFB1
GO protein containing complex assembly | TAF12, AXIN1, JAK2
GO immune system development | FOXO3, RHOA, PPP2R3C

**TABLE 9.** Top pathways in deletion’s significant gene set.

| Gene Set Name | Representative genes |
|---|---|
| GO intrinsic component of plasma membrane | EPHB1, ASIC2, PTPRN2 |
| GO synapse | CHRNA4, MAP1B, GABRB3, |
| GO cell motility | ERBB2, PALLD, BCL2 |
| GO locomotion | PLXNA4, MET, SDC4 |
| GO neuron projection | ANKS1B, STON2, TENM4 |
| GO neurogenesis | CDH4, CAMK1D, SH3GL3 |
| GO cellular component morphogenesis | PALLD, KIT, GRB7 |
| GO neuron differentiation | EYA1, ERCC6, KIRREL3 |
| GO neuron development | NTM, ERCC6, CAMK1D |
| GO cell projection organization | ADGRB1, TMEFF2, ABL2 |
| GO synaptic signaling | PRKCE, ADRA1A, KCNQ2 |
| GO cell part morphogenesis | EIF2AK4, MAP1B, DAB1 |

**TABLE 10.** Top pathways in duplication’s significant gene set.

Proportion of each CN state between case and control. We didn’t make an assumption of accordance of CNV boundaries among all samples, cutting chromosome into bins and then merging bins to enhance power instead.

Furthermore, we raised a new empirical Bayes formula to overcome overfitting, making sure the proportions of CN states between case and control are tested correspondingly. By means of the prior variance varying with sample size and bin length, the prior distribution can still play an important role when the sample size or bin size going to infinity, making it a suitable choice of model selection especially for those situations where there are many models to be specified. We have also proposed the theoretical guarantees of our empirical Bayes estimators under the
circumstance of large sample size or bin size and the prior didn’t require the CN specific prior mean locates quite near by the sample mean.

We demonstrate the effectiveness of our method by simulation and real data analysis, our method behaves well in term of sensitivity and FDR performance in simulation. In real data, besides finding out many famous cancer genes, we pick up some genes that haven’t reported before, which show significant difference between case and control samples. Outcomes in the pathway analysis further validate the meaningfulness of our method.

We have to admit that when sample size is very large, due to heterogeneity across samples, there’s large probability that samples’ CN state are not all the same in the adjacent bins, thereby we can’t merge too much, leading to over-segmentation and low power in detecting large CNVs, under this circumstance, our method needs to accompany with other CNV detection methods which behave well in large CNV detection.

6. APPENDIX

6.1. Proof of theorem 2.3.

Proof. We separate our proof into 2 main parts: existence of all 5 clusters and nonexistence of some clusters, and we consider the first circumstance first. For convinence we abbreviate “with probability tending to 1” as “wpt 1”.

For notation convenience, we drop the symbol of bin location \( b \), case status symbol \( d \) and let \( \theta \) denote the overall parameters \((\mu_k, \sigma_k^2, \alpha_k)_{k=1}^{5}\) to be estimated, let \( \xi \) denote location and scale parameters \((\mu, \sigma^2)\), let \( \alpha \) denote proportion parameters \((\alpha_1, \ldots, \alpha_5)\), simplify \( N_1 \) to \( n \), let \( h(X_i|\theta) \) denote density function \( \sum_{k=1}^{5} \alpha_k f(X_i|\mu_k, \sigma_k^2) \), \( F_{n,p}(X_1, \ldots, X_n|\theta) \triangleq \sum_{i=1}^{n} \log[h(X_i|\theta)] \), \( G_n(\mu_1, \ldots, \mu_5) \triangleq \sum_{k=1}^{5} \log[g(\mu_k|\tau_k, \sigma_k^2)] \), \( L_{n,p}(X_1, \ldots, X_n|\theta) \triangleq F_{n,p}(X_1, \ldots, X_n|\theta) + G_n(\mu_1, \ldots, \mu_5) \), \( F_{n,p}(\theta) \) and \( G_n(\theta) \).

We define a set \( B \) such that \( \forall \theta \in B \), \( h(X|\theta) \) is the “true” underlying distribution, which is equivalent to \( h(X|\{\mu_k^*, (\sigma_k^*)^2, \alpha_k^*\}_{k=1}^{5}) \). If for each element \( \theta \) in \( B \), we drop the proportion part \{\alpha_k\}_{k=2}^{5} \) and denote the left part \((\mu_k, \sigma_k^2)_{k=1}^{5}\) as \( \xi \), all such \( \xi \) formed a new set \( B_\xi \). In the case of existence of all 5 clusters, \( B \) consists of 5! elements obviously, any element in \( B \) is a location permutation of true parameters \((\mu_k^*, (\sigma_k^*)^2, \alpha_k^*)_{k=1}^{5}\) (for example, 0.2N\((-1, 1)\) + 0.8N\((1, 1)\) and 0.8N\((1, 1)\) + 0.2N\((-1, 1)\) denote the same mixture Gaussian distribution). If we define estimators \((\hat{\mu}_k, \hat{\sigma}_k^2, \hat{\alpha}_k)_{k=1}^{5}\) maximize \( \tilde{f}(X_1, \ldots, X_n|\theta) \) in (2) (i.e. maximize \( L_{n,p}(X_1, \ldots, X_n|\theta) \)) as \( \hat{\theta} \triangleq (\hat{\xi}, \hat{\alpha}) \), and denote a set \( \theta = (\xi, \alpha) : |\xi - \xi^*| > \epsilon \text{ or } |\alpha - \alpha^*| > \epsilon \) for some \( \epsilon > 0 \) as \( H \), we want to first claim that

\[
\forall \epsilon > 0, \exists P > 0, \text{ when } p > P, \sup_{\theta \in H} L_{n,p}(\xi, \alpha) < L_{n,p}(\hat{\xi}, \hat{\alpha}) \text{ wpt 1} \tag{16}
\]

To complete (16), we separate \( H \) into 3 sets: \( H = H_1 \cup H_2 \cup H_3 \) with \( H_1 = \{\theta : |\xi - \xi^*| > \epsilon \text{ and } \forall \alpha_k > 0, k = 1, \ldots, 5\} \), \( H_2 = \{\xi - \xi^* > \epsilon \text{ and } \exists \alpha_k = 0, k = 1, \ldots, 5\} \), \( H_3 = \{ |\alpha - \alpha^*| > \epsilon \} \), in the following we will discuss (16) under these 3 circumstances.
With regard to $H_1$, we want to show that for all fixed $\alpha$ with all its $\alpha_k > 0$, $k = 1, \ldots, 5$, we can attain

$$\forall \epsilon > 0, \exists P > 0, \sup_{\xi \mid |\xi - \xi^*| > \epsilon} L_{n,p}(\xi, \alpha) < L_{n,p}(\hat{\xi}_p, \alpha) \text{ wpt } 1$$

(17)

Now we divide $\{\xi: |\xi - \xi^*| > \epsilon\}$ into 2 parts:

$$\left(\bigcup_{\xi \in B_\epsilon \setminus \xi^*} \{\xi: |\xi - \xi^*| < \epsilon\}\right) \bigcup \{\xi: |\xi - \xi^*| > \epsilon\} \bigcup \{\xi: |\xi - \xi^*| < \epsilon\}.$$

For convenience $\bigcup_{\xi \in B_\epsilon \setminus \xi^*} \{\xi: |\xi - \xi^*| < \epsilon\}$ is denoted as $Q$, $Q^c$ is denoted as its complement in set $\{\xi: |\xi - \xi^*| > \epsilon\}$. In order to reach the conclusion described in (17), firstly we show that

$$\sup_{\xi \in Q} L_{n,p}(\xi, \alpha) > \sup_{\xi \in Q^c} L_{n,p}(\xi, \alpha) \text{ wpt } 1$$

(18)

Suppose $\tilde{\xi}_p = \arg\max_{\xi \in Q} F_{n,p}(\xi, \alpha)$, $\hat{\xi}_p = \arg\max_{\xi \in Q^c} L_{n,p}(\xi, \alpha)$, according to the definition of $Q$ and randomicity of $\epsilon$ we can get that there is a location permutation rule $\pi$ such that $\pi(\tilde{\xi}_p) = \tilde{\xi}^\pi_\mu = (\mu^\pi_{pk}, (\sigma^\pi_{pk})^2)_{k=1}^5$ and $\tilde{\xi}_p^\pi \rightarrow \xi^* \text{ a.s.}$, we further assume the number of samples belonging to cluster $k$ is $n_k$ (here we suppose clusters of smaller index $k$ are equipped with smaller location parameter $\mu_k$), and for simplicity, samples $X_1, \ldots, X_{n_1}$ belong to cluster 1, $\ldots$, samples $X_{n_{1}+\cdots+n_{k}+1}, \ldots, X_{n}$, belong to cluster 5. Take samples $X_1, \ldots, X_{n_1}$ for an example, it is obvious that $X_{it} \sim N(\mu^*_i, (\sigma^*_i)^2)$, $\forall i \in 1, \ldots, n_1; \forall t \in 1, \ldots, p$, by law of large numbers and $\tilde{\xi}^\pi_p \rightarrow \xi^* \text{ a.s.}$, we have for $\forall i \in 1, \ldots, n_1$ and $\forall k \neq 1$,

$$\frac{1}{p} \left(\log[f(X_i|\tilde{\mu}^\pi_{p1}, (\tilde{\sigma}^\pi_{p1})^2)] - \log[f(X_i|\hat{\mu}^\pi_{p1}, (\hat{\sigma}^\pi_{p1})^2)]\right)$$

$$= \frac{1}{p} \left(\sum_{t=1}^{p} \log[\phi(X_{it} | \tilde{\mu}^\pi_{p1}, (\tilde{\sigma}^\pi_{p1})^2)] - \sum_{t=1}^{p} \log[\phi(X_{it} | \hat{\mu}^\pi_{p1}, (\hat{\sigma}^\pi_{p1})^2)]\right)$$

(19)

$$\overset{p \to \infty}{\to} \int \log[\frac{\phi(x | \mu^*_i, (\sigma^*_i)^2)}{\phi(x | \mu^\pi_{p1}, (\sigma^\pi_{p1})^2)}] \phi(x | \mu^*_i, (\sigma^*_i)^2) dx = O(1) > 0$$

thereby leading to an order estimation

$$\log[f(X_i|\tilde{\mu}^\pi_{p1}, (\tilde{\sigma}^\pi_{p1})^2)] - \log[f(X_i|\hat{\mu}^\pi_{p1}, (\hat{\sigma}^\pi_{p1})^2)] = O_p(p), \forall k \neq 1$$

(20)

hence $\log[h(X_i|\tilde{\xi}_p, \alpha)] - \log(\alpha^*_1) - \sum_{t=1}^{p} \log[\phi(X_{it} | \tilde{\mu}^\pi_{p1}, (\tilde{\sigma}^\pi_{p1})^2)] = O_p(e^{-p})$ for $\forall i \in 1, \ldots, n_1$.

Here $\alpha^*_i$ represents the location of $\alpha^*_1$ is in accordance with the location of $\{\tilde{\mu}^\pi_{p1}, (\tilde{\sigma}^\pi_{p1})^2\}$ in $\tilde{\xi}_p$.

Samples belonging to clusters 2, $\ldots$, 5 follow similar rules as described above, combining these conclusions we can obtain

$$F_{n,p}(X_1, \ldots, X_n|\tilde{\xi}_p, \alpha) = F_{n,p}(X_1, \ldots, X_n|\tilde{\xi}_p, \alpha^\pi)$$

$$= \sum_{k=1}^{5} n_k \log(\alpha_k) + \sum_{k=1}^{5} \sum_{i \in C_k} \sum_{t=1}^{p} \log[\phi(X_{it} | \tilde{\mu}^\pi_{p1}, (\tilde{\sigma}^\pi_{p1})^2)] + O_p\left(\frac{n}{e^p}\right)$$

(21)

where $C_k$ is a set containing index of samples belonging to cluster $k$. 

25
Next we shift to $\tilde{\xi}_p$, for any realization $x_1, \ldots, x_n$ of sample $X_1, \ldots, X_n$, there is a realization $\tilde{\xi}_p$, by a permutation rule $\pi_1$, we can transform $\tilde{\xi}_p$ into $\tilde{\xi}_p^{\pi_1}$ with location parameters in increasing order, i.e. $\pi_1(\tilde{\xi}_p) = \tilde{\xi}_p^{\pi_1} = (\tilde{\pi}_p^{\pi_1}, (\tilde{\delta}_p^{\pi_1})^{\pi_1})_{k=1}^5$ with $\tilde{\mu}_p^{\pi_1} \leq \cdots \leq \tilde{\mu}_p^{\pi_5}$. By the definition of $Q^c$, it is easy to see that there exists at least one $(\mu_M^i, (\sigma_M^i)^2), k \in \{1, \ldots, 5\}$ such that events $|\tilde{\mu}_p^{\pi_1} - \mu_M^i| > \epsilon$ and $|(\tilde{\pi}_p^{\pi_1})^2 - (\sigma_M^i)^2| > \epsilon$ for $\forall k \in \{1, \ldots, 5\}$ happen at least once. So by a similar argument as (19), samples fall into cluster $k$ must satisfy

$$\log[f(x_i|\tilde{\mu}_p^{\pi_1}, (\tilde{\sigma}_p^{\pi_1})^2)] - \log[f(x_i|\tilde{\mu}_p^{\pi_1}, (\tilde{\sigma}_p^{\pi_1})^2)] = O(p), \ k = 1, \ldots, 5, \ i \in C_k$$

hence we have

$$\log[h(x_i|\tilde{\xi}_p, \alpha)] = \sum_{t=1}^p \log[\phi(x_{it}|\tilde{\mu}_p^{\pi_1}, (\tilde{\sigma}_p^{\pi_1})^2)] - O(p) \ f o r \ \forall i \in C_k $$

Regarding to those samples not belonging to cluster $k$, if $\xi$ doesn’t have any constraints, $h(x_i|\xi, \alpha)$ with fixed $\alpha$ must attains its maximum at $\xi^*$ with $\mu_1 = \cdots = \mu_5 = \mu_M^i, (\sigma_i^2)^2 = \sigma_M^i, (\sigma_M^i)^2$ is the realization of MLE $(\hat{\mu}_M^i, (\hat{\sigma}_M^i)^2)$ of sample $x_1, \ldots, x_i$ following certain normal distribution with $i \in C_k, x_i \neq C_k$. By the properties of MLE, we know that $\hat{\mu}_M^i \xrightarrow{p} \mu_M^i, (\hat{\sigma}_M^i)^2 \xrightarrow{p} (\sigma_M^i)^2$, combining with $\tilde{\xi}_p \xrightarrow{p} \xi^*$ and the continuity of Gaussian density function we can get for $i \in C_k$

$$| \sum_{t=1}^p \log[\phi(x_{it}|\tilde{\mu}_p^{\pi_1}, (\tilde{\sigma}_p^{\pi_1})^2)] - \sum_{t=1}^p \log[\phi(x_{it}|\hat{\mu}_M^i, (\hat{\sigma}_M^i)^2)] | = o(p)$$

$$\log[h(x_i|\tilde{\xi}_p, \alpha)] \leq \log[h(x_i|\xi^*, \alpha)] = \sum_{t=1}^p \log[\phi(x_{it}|\hat{\mu}_M^i, (\hat{\sigma}_M^i)^2)]$$

Combining (23), (24) and (25), a result in analogy with (21) can be achieved

$$F_{n,p}(X_1, \ldots, X_n|\tilde{\xi}_p, \alpha) \leq \sum_{i \in C_k} \sum_{t=1}^p \log[\phi(x_{it}|\tilde{\mu}_p^{\pi_1}, (\tilde{\sigma}_p^{\pi_1})^2)] + \sum_{k \neq k \in C_k} \sum_{t=1}^p \log[\phi(x_{it}|\hat{\mu}_M^i, (\hat{\sigma}_M^i)^2)] - O(p) = (26)$$

$$\sum_{i \in C_k} \sum_{t=1}^p \log[\phi(x_{it}|\tilde{\mu}_p^{\pi_1}, (\tilde{\sigma}_p^{\pi_1})^2)] + \sum_{k \neq k \in C_k} \sum_{t=1}^p \log[\phi(x_{it}|\hat{\mu}_M^i, (\hat{\sigma}_M^i)^2)] - O(p) + o_p(p)$$

where the sign of the last term $o_p(p)$ is uncertain, but no matter what the sign is, it is clear that $o_p(p) - O_p(p) = -O_p(p) \approx 0$. Now compare (26) with (21) leading to $F_{n,p}(X_1, \ldots, X_n|\tilde{\xi}_p, \alpha) = F_{n,p}(X_1, \ldots, X_n|\tilde{\xi}_p, \alpha) - O_p(p)$, adding the fact that $\forall \theta_1, \theta_2, |G_n(\theta_1) - G_n(\theta_2)| < O_p(p)$, we can reach that

$$\sup_{\xi \in Q} L_{n,p}(\xi, \alpha) - \sup_{\xi \in Q^c} L_{n,p}(\xi, \alpha) \geq (F_{n,p}(\tilde{\xi}_p, \alpha) - F_{n,p}(\tilde{\xi}_p, \alpha)) + (G_n(\tilde{\xi}_p) - G_n(\tilde{\xi}_p)) = O_p(p) > 0$$

which means that (18) is achieved, with this in hand, what’s left is to show that for all fixed $\alpha$ with $\alpha_k > 0, k = 1, \ldots, 5$, $\sup_{\xi \in Q} L_{n,p}(\xi, \alpha) < L_{n,p}(\tilde{\xi}_p, \alpha)$ wpt 1 when $p$ is large enough.
We assume \( \hat{\xi}'_p = \arg\max_{\xi \in Q} L_{n,p}(\xi, \alpha) \), similar as described before there exists a location permutation rule \( \pi_2 \) such that \( \pi_2(\hat{\xi}'_p) = \hat{\xi}'_{\pi_2} \Rightarrow \xi^* \), we further assume \( \pi_2(\alpha) = \alpha^{\pi_2} \), noticing that permutating the location of the 5 groups of parameters \( (\mu_k, \sigma_k^2, \alpha_k)_{k=1}^5 \) doesn’t change values of \( F_{n,p}(\xi, \alpha) \), which means that \( F_{n,p}(\hat{\xi}'_p, \alpha) = F_{n,p}(\hat{\xi}'_{\pi_2}, \alpha^{\pi_2}) \). Recall the result presented in (21), we have

\[
F_{n,p}(\hat{\xi}'_p, \alpha) - F_{n,p}(\hat{\xi}'_{\pi_2}, \alpha) \leq \sum_{k=1}^5 n_k | \log(\alpha_k) - \log(\alpha_{\pi_2}^k) | + O_p\left( \frac{n}{p} \right)
\]

In addition, by the definition of \( Q \), we can draw that location parameters \( \{\hat{\mu}_{k1p}\}_{k=1}^5 \) aren’t arranged in increasing order as \( \{\mu_k^{\pi_2}\}_{k=1}^5 \), by lemma 2,

\[
G_n(\hat{\xi}'_{\pi_2}) - G_n(\hat{\xi}') \geq O_p(m(p)) > 0
\]

where \( m(p) \) is any increasing function of order less than \( O(p) \), tending to infinity as long as \( p \to \infty \). Comparing this order estimation with its above inequality, the conclusion \( \sup_{\xi \in Q} L_{n,p}(\xi, \alpha) < L_{n,p}(\hat{\xi}'_p, \alpha) \) wpt 1 is available, meaning that we have completed the proof of (17).

Now let’s turn to \( H_2 \), there exists at least one \( \alpha_k = 0 \) for \( k = 1, \cdots, 5 \), so no matter what values the corresponding \( (\mu_k, \sigma_k^2) \) take, the part \( \alpha_k f(x|\mu_k, \sigma_k^2) \) doesn’t account for \( F_{n,p}(\theta) \) any more, i.e. if there is a \( \alpha_k = 0 \), we only have 4 groups of parameters \( (\alpha_k, \mu_k, \sigma_k^2) \) to fit a 5-cluster \( p \)-dimensional mixture Gaussian model. For example if we assume \( (\hat{\xi}'_p, \hat{\alpha}'_p) = \arg\max_{\alpha \in C_{k1}} L_{n,p}(\xi, \alpha) \), there must exist at least one cluster \( k_1 \) such that

\[
\log[h(X_i|\hat{\xi}'_p, \hat{\alpha}_p)] = \sum_{t=1}^p \log[\phi(X_{it}|\mu_k^{*}, (\sigma_k^{*})^2)] - O_p(p) \text{ for } \forall i \in C_{k1}
\]

For sample \( i \in C_{k2} \neq C_{k1} \), we bound \( \log[h(x|\hat{\xi}'_p, \hat{\alpha}_p)] \) with \( \hat{\xi}_i \) replaced by MLE \( \hat{\xi}_i^* \) of sample \( X_{i1}, \cdots, X_{ip} \) as introduced above

\[
\log[h(X_i|\hat{\xi}_i^*, \hat{\alpha}_p)] \leq \log[f(X_i|\hat{\xi}_i^*, \hat{\alpha}_p)] = \sum_{t=1}^p \log[\phi(X_{it}|\mu_k^{*}, (\sigma_k^{*})^2)],
\]

\[
\sum_{t=1}^p \log[\phi(X_{it}|\mu_k^{*}, (\sigma_k^{*})^2)] = \sum_{t=1}^p \log[\phi(X_{it}|\mu_k^{*}, (\sigma_k^{*})^2)] + o_p(p)
\]

Combining (28) and (29) we can give an upper bound of \( F_{n,p}(\hat{\xi}'_p, \hat{\alpha}_p) \)

\[
F_{n,p}(\hat{\xi}'_p, \hat{\alpha}_p) \leq \sum_{k=1}^5 \sum_{t=1}^p \log[\phi(x_{it}|\mu_k^{*}, (\sigma_k^{*})^2)] - O_p(p) + o_p(p)
\]

\[
= F_{n,p}(\xi^*, \alpha^*) - \sum_{k=1}^5 n_k \log(\alpha_k^*) - O_p(p) + o_p(p) - o_p(n)
\]
In addition to the fact that \( G_n(\tilde{\xi}_n) - G_n(\xi^*) < O_p(p) \), we can obtain

\[
\sup_{\alpha_1=0} L_{n,p}(\xi, \alpha) = L_{n,p}(\xi, \alpha) < L_{n,p}(\xi^*, \alpha^*) \leq L_{n,p}(\hat{\xi}_n, \hat{\alpha}_p)
\]  

(31)

Scenarios belonging to \( H_2 \) apart from only \( \alpha_1 = 0 \) can be discussed in analogy with the case of \( \alpha_1 = 0 \). Furthermore, in the discussion of \( H_2 \), we didn’t make use of the condition \( |\xi - \xi^*| > \epsilon \), so actually we have proved (16) under the constrain of \( \theta \in \{ \theta : \exists \alpha_k = 0, k = 1, \ldots, 5 \} \triangleq H_2' \).

In the following we move to \( H_3 \). Observing that \( H_3 \) has intersection with \( H_1 \) and \( H_2' \), what’s left is to prove (16) with the constrain \( \theta : |\alpha - \alpha^*| > \epsilon, \forall \alpha_k > 0, k = 1, \ldots, 5 \) and \( |\xi - \xi^*| < \epsilon \) \( \triangleq H_3' \).

Suppose \( \hat{\theta}_p = (\tilde{\xi}_p, \tilde{\alpha}_p) = \arg \max_{\theta \in H_3'} L_{n,p}(\xi, \alpha) \), by the randomicity of \( \epsilon, \tilde{\xi}_p \xrightarrow{p} \xi^* \), resulting in similar conclusions as (19), (20) and (21), i.e. we can get

\[
F_{n,p}(X_1, \ldots, X_n|\hat{\theta}_p) = \sum_{i=1}^{5} n_k \log(\tilde{\alpha}_p) + O_p\left(\frac{n}{\epsilon p}\right) + \sum_{k=1}^{p} \sum_{t \in C_k} \log[\phi(X_{it}|\tilde{\alpha}_p, \tilde{\pi}_k) (\tilde{\pi}_k)^2] \)

(32)

If we replace \( \tilde{\alpha}_p \) in \( \hat{\theta}_p\) with \( \alpha^* \), analogous to (32), estimation of \( F_{n,p}(\tilde{\xi}_p, \alpha^*) \) can be obtained. With the addition of \( \alpha^* = \arg \max_{\alpha} \sum_{k=1}^{5} n_k \log(\alpha_k) \) and \( G_n(\tilde{\xi}_p, \alpha^*) = G_n(\tilde{\xi}_p, \alpha^*) \) we have

\[
\begin{align*}
L_{n,p}(\hat{\theta}_p) - L_{n,p}(\xi^*) &= \sum_{k=1}^{5} n_k \log(\tilde{\alpha}_p) - \log(\alpha_k^*)] + O_p\left(\frac{n}{\epsilon p}\right) \\
&= -O_p(n) + o_p(n) < 0
\end{align*}
\]

(33)

thus we have the final conclusion under \( H_3' \): \( \sup_{\theta \in H_3'} L_{n,p}(\xi, \alpha) < L_{n,p}(\xi^*, \alpha^*) \leq L_{n,p}(\hat{\xi}_n, \hat{\alpha}_p) \). By now we have completed the proof of consistency under the condition of existence of all 5 clusters, we will go on to the second main part.

We continue to use the symbol \( B \) to denote a set containing all parameters \( \theta \) such that \( h(x|\theta) \) is the true underlying distribution, in this circumstance, there exist innumerable possible \( \theta \) in \( B \). For example, if the true underlying distribution is \( N(0,1) \), then all distributions \( \sum_{k=1}^{5} \alpha_k N(0,1) \) satisfying \( \sum_{k=1}^{5} \alpha_k = 1 \) is the same as \( N(0,1) \). For notation simplification, we denote parameters only corresponding to set \( K \subseteq \{1, \ldots, 5\} \) in \( \theta \) as \( \theta_K = (\xi_K, \alpha_K) \), for example, paramters only corresponding to clusters 1 and 2 (\( K = \{1, 2\} \)) \( (\mu_k, \sigma_k, \alpha_k)_{k=1}^{2} \) can be denoted as \( \theta_K \). Suppose the existing clusters form a set \( K^* \), maximizer of \( L_{n,p}(X_1, \ldots, X_n|\theta) \) is also denoted as \( \hat{\theta}_p \triangleq (\hat{\xi}_p, \hat{\alpha}_p) \), and we define a set \( S \) as \( \{ \theta : \left| \theta_K^* - \theta_K^* \right| > \epsilon, |\mu_k - \tau_1| > \epsilon \text{ for } \forall k \in K^{sc} \text{ or } |\alpha_k| > \epsilon \text{ for } \forall k \in K^{sc} \} \), under this circumstance we want to first show that

\[
\forall \epsilon > 0, \exists P > 0, \text{ when } p > P, \sup_{\theta \in S} L_{n,p}(\xi, \alpha) < L_{n,p}(\hat{\xi}_n, \hat{\alpha}_p) \text{ wpt 1}
\]

(34)
As in the first main circumstance, we divide \( S \) into 2 parts: 
\[ \bigcup_{\theta \in \mathcal{D}} \{ \theta : |\theta - \hat{\theta}| < \epsilon \} \bigcap S \Delta V \]

and its complement in \( S \) denoted as \( V^c \). Furthermore, \( V^c = V_1^c \bigcup V_2^c \) with \( V_1^c = \{ \theta \in V^c : \exists \hat{\theta} \in B, s.t. |\xi - \hat{\xi}| < \epsilon \text{ for all } \xi > 0, \text{ and } \exists \epsilon > 0, s.t. |\alpha - \hat{\alpha}| > \epsilon \} \), \( V_2^c = \{ \theta \in V^c : \forall \theta \in B, \exists \epsilon > 0, s.t. |\xi - \hat{\xi}| > \epsilon \} \), in the following we will show the correctness of (34) with set \( S \) replaced by \( V_2^c, V, V_1^c \) orderly.

Firstly, we assume \( \hat{\theta}_p = \arg\max_{\theta \in V_2^c} L_{n,p}(\xi, \alpha) \), by the definition of \( V_2^c \) it is not hard to deduce that either of the following 2 circumstances holds:

(a) \( \exists \hat{k} \in K^*, \epsilon > 0, s.t. \forall k \in \{1, \cdots, 5\}, |\hat{\xi}_{p_k} - \xi_k^*| > \epsilon \)

(b) \( \forall k \in K^*, \forall \epsilon > 0, \exists k \in \{1, \cdots, 5\}, s.t. |\hat{\xi}_{p_k} - \xi_k^*| < \epsilon \) and \( \exists k \in \{1, \cdots, 5\}, \exists \epsilon > 0, s.t. \forall k \in K^*, |\hat{\xi}_{p_k} - \xi_k^*| > \epsilon, \hat{\alpha}_{p_k} > \epsilon \)

Circumstance (a) implies any sample \( i \) belonging to cluster \( \hat{k} \) has the property similar with (22) and (23)

\[
\log[f(X_i | \mu_{k}^*, (\sigma_k^*)^2)] - \log[f(X_i | \hat{\mu}_{p_k}, \hat{\sigma}_{p_k}^2)] = O_p(p), \forall k \in \{1, \cdots, 5\},
\]

\[
\log[h(X_i | \hat{\xi}_{p}, \hat{\alpha}_p)] = \sum_{t=1}^{p} \log[\phi(X_{it} | \hat{\mu}_{k}^*, (\sigma_k^*)^2)] - O_p(p)
\]

Therefore we can get

\[
F_{n,p}(X_1, \cdots, X_n | \hat{\xi}_{p}, \hat{\alpha}_p) \leq \sum_{i \in C_k} \sum_{t=1}^{p} \log[\phi(X_{it} | \mu_{k}^*, (\sigma_k^*)^2)] + \sum_{k \neq \hat{k} \in C_{k^*}} \sum_{t=1}^{p} \log[\phi(X_{it} | \mu_{k^*}, (\sigma_k^*)^2)] - O_p(p) = \sum_{i \in C_k} \sum_{t=1}^{p} \log[\phi(X_{it} | \mu_{k}^*, (\sigma_k^*)^2)] + \sum_{k \neq \hat{k} \in C_{k^*}} \sum_{t=1}^{p} \log[\phi(X_{it} | \mu_{k^*}, (\sigma_k^*)^2)] - O_p(p) + O_p(p)
\]

i.e. \( F_{n,p}(\hat{\xi}_{p}, \hat{\alpha}_p) \leq F_{n,p}(\theta^*) - O_p(p) \), adding the fact \( G_n(\hat{\xi}_p) - G_n(\xi^*) < O_p(p) \) we can draw the conclusion

\[
\sup_{\theta \in V_2^c} L_{n,p}(\xi, \alpha) = L_{n,p}(\hat{\xi}_p, \hat{\alpha}_p) < L_{n,p}(\xi^*, \alpha^*) \leq L_{n,p}(\hat{\xi}_p, \hat{\alpha}_p)
\]

In regard to circumstance (b), we can draw a similar conclusion as (32):

\[
F_{n,p}(X_1, \cdots, X_n | \hat{\xi}_{pk}) = \sum_{k \in K^*} n_k \log(\alpha_k) + \sum_{k \in K^*} \sum_{i \in C_k} \sum_{t=1}^{p} \log[\phi(X_{it} | \mu_{k}^*, \hat{\sigma}_{p_k}^2)] + O_p\left(\frac{n}{e^p}\right)
\]

Combining with the fact that \( \alpha^* = \arg\max_{\alpha} \sum_{k \in K^*} n_k \log(\alpha_k) \) and \( \exists k \in K^c, \epsilon > 0, s.t. \hat{\alpha}_{pk} > \epsilon \), an upper bound under circumstance (b) can be obtained:

\[
L_{n,p}(\hat{\theta}_p) < L_{n,p}(\hat{\xi}_p, \alpha^*) \leq L_{n,p}(\hat{\xi}_p, \hat{\alpha}_p) \text{ wpt } 1
\]
Secondly, we assume \( \tilde{\theta}_p = \arg \max L_{n,p}(\xi, \alpha) \), noticing that \( \tilde{\theta}_p \) lies in \( V \), a set satisfying \( \forall \theta \in V \), \( h(x|\theta) \) approximates \( h(x|\theta^*) \) in arbitrary close distance, then \( \forall k \in K^* \), there must exist at least one \( k' \in \{1, \cdots , 5\} \) such that \( \tilde{\mu}_{kp'} \overset{p}{\to} \mu_{k^*} \), based on this fact, we next discuss (34) on set \( V \) under 2 conditions:

(c) \( \exists \epsilon > 0 \), s.t. \( |\tilde{\alpha}_{kp}| > \epsilon \) for \( k \in K_1 \), where \( |K_1| > |K^*| \), here we use "|·|" to denote cardinality of a set.

(d) \( \forall \epsilon > 0 \), \( \exists P > 0 \), when \( p > P \), \( |\tilde{\alpha}_{kp}| < \epsilon \) for \( k \in K_1^\epsilon \), where \( |K_1| = |K^*| \).

With regard to condition (c), \( \forall k \in K_1 \), there exists a \( k^* \in K^* \), s.t. \( \tilde{\xi}_{pk} \overset{p}{\to} \xi_{k^*}^* \), and we introduce a set \( D(k^*) \) to denote those \( k \in K_1 \) with the property \( \tilde{\xi}_{pk} \overset{p}{\to} \xi_{k^*}^* \) holds. Furthermore, if \( c(i) \) represents the cluster to which sample \( i \) belongs to, we have the following approximation of \( F_{n,p}(X_1, \cdots , X_n|\tilde{\theta}_p) \):

\[
F_{n,p}(X_1, \cdots , X_n|\tilde{\theta}_p) = \sum_{k \in K^*} \sum_{c(i)=k^*} \log \left[ \sum_{k \in D(k^*)} \alpha_k f(X_i|\tilde{\xi}_{pk}) + O_p\left(\frac{1}{e^p}\right) \right] \quad (39)
\]

If we pick up a specific term \( \sum_{c(i)=k^*} \log \left[ \sum_{k \in D(k^*)} \alpha_k f(X_i|\tilde{\xi}_{pk}) + O_p\left(\frac{1}{e^p}\right) \right] \overset{\Delta}{=} F_{n,p}^{k^*}(\tilde{\theta}_p) \), by direct calculation it’s not hard to deduce that neglecting terms of smaller order, \( \exp\left[F_{n,p}^{k^*}(\tilde{\theta}_p)\right] \) is a summation of terms like \( \prod_{c(i)=k^*} \alpha_k f(X_i|\tilde{\xi}_{pk}) \), where \( k_i \) traverses any value in \( D(k^*) \) resulting in overall \( |D(k^*)|^{n_{k^*}} \) terms in the summation, here \( n_{k^*} \) is the number of samples belonging to cluster \( k^* \). In the following we want to bound \( F_{n,p}^{k^*}(\tilde{\theta}_p) \) from a hypothesis testing viewpoint. Noticing that

\[
\prod_{c(i)=k^*} \alpha_k f(X_i|\tilde{\xi}_{pk}) = \prod_{k_d \in D(k^*)} \prod_{c(i)=k^*} \prod_{k_i=k_d} \alpha_k f(X_i|\tilde{\xi}_{pk})
\]

and we denote the set \( \{i|k_i = k_d\} \) as \( D_1, \cdots , D_{D(k^*)} \) for \( k_d \in D(k^*) \). Suppose samples from the same set \( D_{k_d} \), \( k_d \in D(k^*) \) follow the same normal distribution with parameters indexed by \( \xi_{k_d} \), consider the following hypothesis testing:

\[
H_0 : \xi_{k_d} \text{ are equivalent for all } k_d \in D(k^*) \\
H_1 : \xi_{k_d} \text{ are not all equivalent for } k_d \in D(k^*)
\]

The likelihood under \( H_1 \) is \( \prod_{k_d \in D(k^*)} \prod_{i \in D_{k_d}} f(X_i|\tilde{\xi}_{Lk_d}) \overset{\Delta}{=} L_{H_1} \), where \( \tilde{\xi}_{Lk_d} \) is the likelihood of sample \( \{X_{i1}, \cdots , X_{ip}\}_{i \in D_{k_d}} \). The likelihood under \( H_0 \) is \( \prod_{k_d \in D(k^*)} \prod_{i \in D_{k_d}} f(X_i|\tilde{\xi}_{Lc(i)}) \overset{\Delta}{=} L_{H_0} \), whereas \( \tilde{\xi}_{Lc(i)} \) is the likelihood of all sample \( \{X_{i1}, \cdots , X_{ip}\} \) with \( i \) satisfying \( c(i) = k^* \). From theory of hypothesis testing we can draw that

\[
\log(L_{H_1}) - \log(L_{H_0}) = O_p(1)
\]
For every $\prod_{k_d \in D(k^*)} \prod_{c_l(i) = k^*} \prod_{k_i = k_d} \alpha_{k_i} f(X_i \xi_{pki}^\ast)$, there is a particular testing formula like (40) corresponding to it, equipping with the property that its value is not larger than the corresponding $\log(L_{H_1})$, i.e. $\log(L_{H_n}) + O_p(1)$. So by now we have attain the result that every term in $\exp[F_{n,p}^\ast(\hat{\theta}_p^\ast)]$ is bounded by $\prod_{c_l(i) = k^*} f(X_i \xi_{Lc(i)}^\ast) \cdot O_p(1)$, therefore we can update the approximation of $F_{n,p}(X_1, \ldots, X_n | \theta_p^\ast)$ in (39):

$$F_{n,p}(X_1, \ldots, X_n | \theta_p^\ast) = \sum_{k^* \in K^*} F_{n,p}^\ast(\xi_{p}^\ast)$$

$$\leq \sum_{k^* \in K^*} \sum_{\xi_{Lc(i)} = k^*} \sum_{t=1}^p \log[\phi(X_{it} | \xi_{Lc(i)})] + O_p(1)$$

(41)

If we denote the set $\{\xi_{Lc(i)} | c(i) = k^*, \forall k^* \in K^*\} \cup \{\xi_{k}, \forall k \in K^c\}$ as $\xi_{Lc}$ and replace $\xi_{k^*}$ in $\theta^*$ with it, we call the new formed $\theta^*$ as $\hat{\theta}_{Lc}$, thus

$$F_{n,p}(X_1, \ldots, X_n | \hat{\theta}_{Lc}) = \sum_{k^* \in K^*} n_{k^*} \log(\alpha_{k^*}) + O_p(\frac{n}{e^p})$$

$$+ \sum_{k^* \in K^*} \sum_{\xi_{Lc(i)} = k^*} \sum_{t=1}^p \log[\phi(X_{it} | \xi_{Lc(i)})]$$

(42)

So we can reach that $F_{n,p}(\hat{\theta}_p^\ast) \leq F_{n,p}(\hat{\theta}_{Lc}) + O_p(1)$.

Taking the between-cluster distance into account, $|\hat{\mu}_{pk} - \tau_k| = O_p(1)$ for $\forall k \in K_1 \setminus K^*$, adding with lemma 2 can lead to

$$G_n(\hat{\xi}_{pLc}^\ast) - G_n(\hat{\xi}_{Lc}^\ast) = -O_p(m(p)) < 0$$

(43)

By now combining the facts we have reached in (41), (42) and (43) leads to $L_{n,p}(\hat{\theta}_p^\ast) < L_{n,p}(\hat{\xi}_{Lc}, \alpha^\ast) \leq L_{n,p}(\hat{\xi}_{pLc}^\ast, \alpha_p^\ast)$ wp 1.

With regard to condition (d), we separate it into 2 smaller cases: (d1) $K_1 = K^*$ and (d2) $K_1 \neq K^*$.

Considering (d1) first, if $\theta_{pK_1}^\ast \overset{p}{\rightarrow} \theta_{K^*}$, since $\theta_p^\ast$ lies in $S$, there exists at least one $k \in K^c$ such that $|\hat{\mu}_{pk} - \tau_k| > \epsilon$ for some $\epsilon > 0$, resulting in $G_n(\hat{\xi}_{pLc}^\ast) - G_n(\xi^\ast) = -O_p(m(p)) < 0$. On the other hand, if we replace $\hat{\xi}_{pK_1}^\ast$ in $\xi^\ast_{K_1}^c$ and denote the new formed $\hat{\theta}_p^\ast$ as $\hat{\theta}_{pK_1}^\ast$, it is obvious that $|G_n(\theta^*) - G_n(\hat{\theta}_{pK_1}^\ast)| \leq O_p(m(p))$, what’s more, by similar arguments as (42) we can show the outcome $F_{n,p}(\hat{\theta}_p^\ast) - F_{n,p}(\hat{\theta}_{pK_1}^\ast) = O_p(\frac{n}{e^p})$, thereby we have $L_{n,p}(\hat{\xi}_{pK_1}^\ast) \leq L_{n,p}(\hat{\theta}_{pK_1}^\ast) = L_{n,p}(\hat{\xi}_{pLc}^\ast, \alpha_p^\ast)$ wp 1; if $\hat{\theta}_{pK_1}$ doesn’t converge in probability to $\theta_{K^*}$, since $\theta_p^\ast$ lies in $V$, by some location permutation $\hat{\theta}_{pK_1}^\ast$ must converge in probability to $\theta_{K^*}$, we call $\hat{\theta}_p^\ast$ after such permutation as $\hat{\theta}_{pK_1}^\ast$. So by lemma 2, $G_n(\hat{\theta}_p^\ast) - G_n(\hat{\theta}_{pK_1}^\ast) = -O_p(m(p))$ can be obtained, again by similar arguments as (42), we can also reach $L_{n,p}(\hat{\theta}_p^\ast) < L_{n,p}(\hat{\theta}_{pK_1}^\ast) \leq L_{n,p}(\hat{\xi}_{pLc}^\ast, \alpha_p^\ast) wp 1.$
Next we consider (d2), only by location permutation can \( \tilde{\theta}_{pK_1}' \) converge in probability to \( \theta_{K^*}^* \), we call the permuted \( \tilde{\theta}_{p}'' \) as \( \tilde{\theta}_{p}'' \) too, thereby following lemma 2, we can attain \( G_{n}(\tilde{\theta}_{p}''\text{-}1) - G_{n}(\tilde{\theta}_{p}''\text{-}1) = -O_{p}(m(p)) \), the claim of \( F_{n,p}(\theta_{p}) - F_{n,p}(\tilde{\theta}_{p}''\text{-}1) = O_{p}(\frac{n}{p}) \) can be achieved similar with (42) as well, i.e. under case (b2), \( L_{n,p}(\xi_{p}) < L_{n,p}(\tilde{\theta}_{p}''\text{-}1) \leq L_{n,p}(\xi_{p},\alpha_{p}) \text{ wpt } 1 \). By now we have completed the proof of (34) with \( S \) replaced by \( \bar{V}_{l}^{e} \).

Last but not least, we consider the set \( V_{l}^{e} \) and assume \( \tilde{\theta}_{p}'' = \arg\max_{\theta \in \bar{V}_{l}^{e}} L_{n,p}(\xi,\alpha) \), we will show \( L_{n,p}(\tilde{\theta}_{p}'') < L_{n,p}(\tilde{\theta}_{p}) \) under the following 2 scenarios:

(e) \( \exists \tilde{\theta} \in B_{\xi} \backslash \xi^* \), s.t. \( \tilde{\xi}'_{p}'' \rightarrow \tilde{\xi} \) and \( \exists \epsilon > 0, \text{ s.t. } |\tilde{\alpha}'_{p}'' - \alpha^*| > \epsilon \)

(f) \( \tilde{\xi}'_{p}'' \rightarrow \xi^* \) and \( \exists \epsilon > 0, \text{ s.t. } |\tilde{\alpha}'_{p}'' - \alpha^*| > \epsilon \)

For scenario (e), it’s natural to arrive that apart from different \( \tilde{\alpha}'_{p}'' \), there is a \( \tilde{\theta}_{p}' \) from set \( V \) in accordance with \( \tilde{\theta}_{p}' \), i.e. \( \tilde{\xi}'_{p}' = \tilde{\xi}'_{p}' \), thereby similar arguments as (41) and (42) can be reached, which leads to \( F_{n,p}(\tilde{\theta}_{p}''\text{-}1) \leq F_{n,p}(\tilde{\theta}_{e}) + O_{p}(1) \). \( G_{n}(\tilde{\xi}'_{p}'') - G_{n}(\tilde{\xi}_{L_e}) = -O_{p}(m(p)) < 0 \) can also be obtained as discussed in the setting on set \( V \), therefore we have \( L_{n,p}(\tilde{\xi}'_{p}'') < L_{n,p}(\tilde{\xi}_{L_e},\alpha^*) \leq L_{n,p}(\tilde{\xi}'_{p}'',\alpha^*) \text{ wpt } 1 \).

For scenario (f), \( \exists \epsilon > 0 \), the location parameters \( \mu_{k}^* \) for \( k \in \{1, \cdots, 5\} \) are separated with each other with distances larger than \( \min_{k \in \{1, \cdots, 5\}} d_{k} > 2\zeta > \epsilon \), therefore we can reach a result similar as (21):

\[
F_{n,p}(\tilde{\theta}_{p}'') = \sum_{i=1}^{n} \log(\tilde{\alpha}'_{pc(i)}') + \sum_{i=1}^{n} \sum_{t=1}^{p} \log[\phi(X_{it}|\tilde{\mu}'_{pc(i)}, (\tilde{\alpha}'_{pc(i)})^{2})] + O_{p}(\frac{n}{e^p})
\]

where \( \sum_{i=1}^{n} \log(\tilde{\alpha}'_{pc(i)}) = \sum_{k' \in K^*} n_{k} \log(\tilde{\alpha}'_{pk^*}) \), which attains its maximum when \( \log(\tilde{\alpha}'_{pk^*}) \) are arranged proportional to \( n_{k} \) for \( k^* \in K^* \), \( n_{k} \) is the number of samples belonging to cluster \( k \). Obviously under condition (f), \( \sum_{k^* \in K^*} n_{k^*} \log(\tilde{\alpha}'_{pk^*}) < \sum_{k^* \in K^*} n_{k^*} \log(\alpha^*_{k^*}) \), furthermore we can get \( F_{n,p}(\tilde{\xi}'_{p}'',\alpha^*) < F_{n,p}(\tilde{\xi}'_{p}',\alpha^*) \), hence \( L_{n,p}(\tilde{\xi}'_{p}'',\alpha^*) < L_{n,p}(\tilde{\xi}'_{p}',\alpha^*) \leq L_{n,p}(\tilde{\theta}_{p}) \text{ wpt } 1 \), which completes the claim under condition (f).

By far we have proved (34), which implies that for \( \forall \epsilon > 0, \forall \theta \in S \), when \( p \) is large enough, there is a \( \eta > 0, \text{ s.t. } L_{n,p}(\theta) < L_{n,p}(\tilde{\theta}_{p}) - \eta \text{ wpt } 1 \), i.e. the event \( \{ \theta \in S \} \) is contained in the event \( \{ L_{n,p}(\theta) < L_{n,p}(\tilde{\theta}_{p}) - \eta \} \) with probability tending to 1, whereas \( P(L_{n,p}(\tilde{\theta}_{p}) < L_{n,p}(\tilde{\theta}_{p}) - \eta) = 0 \), resulting in \( P(\tilde{\theta} \in S^c) \rightarrow 1 \), which coincides with the consistency properties as we claimed in the theorem, so we have completed the full proof.

6.2. Proof of theorem 2.4.

Proof. We consider the case of the existence of all 5 clusters first and adopt symbol rules defined as in theorem 1, since in this setting bin size \( p \) is a fixed constant, for convenience we drop symbol \( p \) in \( F_{n,p}(X_{1}, \cdots, X_{n}) = F_{n,p}(\theta) \), resulting in \( F_{n}(\theta) \). For convenience we abbreviate "with probability tending to 1" as "wpt 1".
As mentioned above, we define a set $B$ such that $\forall \theta \in B$, $h(x|\theta)$ is the true underlying distribution. If we define estimators $(\hat{\mu}_k, \hat{\sigma}_k^2, \hat{\alpha}_k)$ maximize $\hat{f}(x_1, \cdots, x_n|\theta)$ in (2) (i.e. maximize $L_n(x_1, \cdots, x_n|\theta)$) as $\hat{\theta}_n$, we want to first claim that

$$\forall \epsilon > 0, \exists N > 0, \text{when } n > N, \sup_{\{\theta:|\theta - \theta^*| > \epsilon\}} L_n(\theta) < L_n(\hat{\theta}_n), \text{ wpt } 1 \quad (45)$$

To prove (45), we divide the set $\{\theta:|\theta - \theta^*| > \epsilon\}$ into 2 parts:

$$(\bigcup_{\theta \in B \setminus \theta^*} \{\theta:|\theta - \theta^*| < \epsilon\}) \bigcup (\{\theta:|\theta - \theta^*| > \epsilon\} \setminus \bigcup_{\theta \in B \setminus \theta^*} \{\theta:|\theta - \theta| < \epsilon\}).$$

Let $D$ denote the set $\bigcup_{\theta \in B \setminus \theta^*} \{\theta:|\theta - \theta| < \epsilon\}, D^c$ denote its complement in set $\{\theta:|\theta - \theta^*| > \epsilon\}$, then in short we have $\{\theta:|\theta - \theta^*| > \epsilon\} = D \cup D^c$. Next we focus on showing that

$$\sup_{\theta \in D} L_n(\theta) > \sup_{\theta \in D^c} L_n(\theta), \text{ wpt } 1 \quad (46)$$

Suppose $\tilde{\theta}_n = \arg\max_{\theta \in D} F_n(\theta), \tilde{\theta}_n = \arg\max_{\theta \in D^c} L_n(\theta)$, by the randomicity of $\epsilon$ and continuity of $h(x|\theta)$ in $\theta$, $h(\tilde{\theta}_n) \xrightarrow{p} h(\theta)$ for $\theta \in B$; since $\overline{D^c}$ is a closed set, $L_n(x_1, \cdots, x_n|\theta)$ is continuous in $\theta$, $L_n(\theta)$ constrained in $D^c$ must attain its maximum in $\overline{D^c}$ for $\forall \epsilon$ fixed, we must have $h(x|\tilde{\theta}_n) \xrightarrow{wpt} h(x|\theta)$ in probability for $\forall \theta \in B$, so by law of large numbers we can obtain

$$\frac{1}{n}[F_n(\tilde{\theta}_n) - F_n(\tilde{\theta}_n)] = \int \log \frac{h(x|\tilde{\theta}_n)}{h(x|\theta^*)} dx + o_p(1)$$

$$= \int \log \frac{h(x|\theta^*)}{h(x|\tilde{\theta}_n)} dx + o_p(1) + o_p(1) = O_p(1) > 0 \quad (47)$$

Combining with the fact that the order of $1/\sigma_{\pi k}^2$ is $O(m(n))$, which is less than $O(n)$, $\forall k = 1, \cdots, 5$, we can draw the conclusion that

$$\sup_{\theta \in D} L_n(\theta) - \sup_{\theta \in D^c} L_n(\theta) \geq [F_n(\tilde{\theta}_n) - F_n(\tilde{\theta}_n)] + [G_n(\tilde{\theta}_n) - G_n(\tilde{\theta}_n)]$$

$$= O_p(n) - O_p(m(n)) > 0 \text{ wpt } 1 \quad (48)$$

which means that (46) holds, so we can shrink the range of taking supreme of $L_n(\theta)$ from $\{\theta:|\theta - \theta^*| > \epsilon\}$ to set $D$.

Next we aim to show that $\sup_{\theta \in D} L_n(\theta) \leq L_n(\hat{\theta}_n)$ for $\forall \epsilon > 0$ and $n$ sufficiently large. Suppose $\hat{\theta}'_n = \arg\max_{\theta \in D} L_n(\theta)$, by the fact that $\tilde{\theta}'_n$ lies in $D$ we can conclude the location parameters $\hat{\mu}'_{n1}, \cdots, \hat{\mu}'_{n5}$ aren’t arranged in increasing order, hence only by a location permutation rule $\pi$ such that $\pi(\hat{\theta}'_n) \triangleq \hat{\theta}'_n^\pi$ can the location parameters $\{\hat{\mu}'_{nk}\}_{k=1}^5$ be arranged in increasing order, therefore by lemma 2

$$G_n(\hat{\theta}'_n) - G_n(\hat{\theta}'_n^\pi) = -O_p(m(n)) \quad (49)$$
Furthermore, we can see obviously that any location permutation rule of \( \tilde{\theta}' \) doesn’t change the value of \( F_n(\tilde{\theta}') \), i.e. \( F_n(\tilde{\theta}') = F_n(\tilde{\theta}''_{\tilde{\theta}}) \), combining the result we have reached in (49) leads to

\[
\sup_{\theta \in D} L_n(\theta) = F_n(\tilde{\theta}'_{n}) + G_n(\hat{\theta}') = F_n(\tilde{\theta}''_{n}) + G_n(\tilde{\theta}''_{\tilde{\theta}}) - O_p(m(n))
\]

(50)

adding the fact in (48), we complete the proof of (45).

With (45) holds, there exists for \( \forall \epsilon > 0 \), \( \forall \theta \) satisfying \( |\theta - \theta^*| > \epsilon \), there is a \( \eta > 0 \) such that \( L_n(\theta) < L_n(\tilde{\theta}^*) - \eta \). Thus, the event \( \{ |\hat{\theta}_n - \theta^*| > \epsilon \} \) is contained in the event \( \{ L_n(\hat{\theta}_n) < L_n(\tilde{\theta}^*) - \eta \} \) wpt 1, i.e.

\[
P(|\hat{\theta}_n - \theta^*| > \epsilon) \leq P(L_n(\hat{\theta}_n) < L_n(\tilde{\theta}^*) - \eta) = 0
\]

(51)

which implies that \( \hat{\theta}_n \) converges in probability to the true parameter \( \theta^* \) when 5 clusters all exist.

In the following, we concentrate on the proof under the circumstances of nonexistence of some cluster. For notation simplification, we follow some symbols described in theorem 1 and denote parameters only corresponding to set \( K \subseteq \{1, \cdots, 5\} \) in \( \theta \) as \( \theta_K \). \( B \) is also defined as a set such that \( \forall \theta \in B \), \( h(x|\theta) \) is the true underlying distribution. Suppose the true existing clusters form a set as \( K^* \), if we define a set \( J \) as \( \{ \theta : \exists \epsilon > 0, \ s.t. \ |\theta_K^* - \theta_K^*| > \epsilon \ or \ |\alpha_k| > \epsilon \ for \ \exists k \in K^{sc} or \ |\mu_k - \tau_k| > \epsilon \ for \ \exists k \in K^{sc} \} \), to accomplish our consistency’s proof under this case, we want to first claim that

\[
\forall \epsilon > 0, \exists N > 0, \text{when } n > N, \sup_{\theta \in J} L_n(\theta) < L_n(\tilde{\theta}^*) \text{ wpt 1}
\]

(52)

As in the case of existence of all clusters, we separate the set \( J \) into 2 parts: \( \{ \theta : \exists \epsilon > 0, \ s.t. \ |\theta - \tilde{\theta}| > \epsilon \} \cap J \triangleq M \) and its complement in set \( J \) denoted as \( J \setminus M \). As long as we keep in mind that any element \( \tilde{\theta} \) in set \( M \) corresponds to a distribution \( h(x|\tilde{\theta}) \) which has arbitrary close distance to the true one, whereas any element in set \( J \setminus M \) doesn’t. So by similar arguments as in the first circumstance, we can show the conclusion

\[
\sup_{\theta \in M} L_n(\theta) > \sup_{\theta \in J \setminus M} L_n(\theta) \text{ wpt 1}
\]

(53)

In the following we aim to show that \( \sup_{\theta \in M} L_n(\theta) < L_n(\tilde{\theta}^*_{\tilde{\theta}}) \). Suppose \( \tilde{\theta}^* = \arg\max_{\theta \in M} F_n(\theta) \), and we define a set \( \Theta_5 \) from which \( \theta \) can take value as \( \theta \) has no constraints apart from \( h(x|\theta) \) is a mixture of 5 Gaussian distributions. By lemma 1 we know that the log likelihood under the overfitting setting is equal to the true log likelihood plus a random variable of order \( O_p(1) \), so here we can attain the result

\[
\sup_{\theta \in M} F_n(\theta) \leq \sup_{\theta \in \Theta_5} F_n(\theta) = F_n(\theta^*) + O_p(1)
\]

(54)

Next we aim to show \( G_n(\theta^*) - G_n(\tilde{\theta}^*) \gg O_p(1) \), we will discuss it under 2 conditions:

(a) \( \exists \epsilon > 0, \ s.t. \ |\alpha_n| > \epsilon \) for \( k \in K_1 \), where \( |K_1| > |K^*| \), here we use ”|·|” to denote cardinality of a set.
(b) \( \forall \epsilon > 0, \exists N > 0, \text{when } n > N, |\alpha_n| < \epsilon \) for \( k \in K^c_1 \), where \( |K_1| = |K^*| \).
With regard to condition (a), noting that any $\theta \in M$ refers to a distribution close enough to the true one, therefore $\forall k_1 \in K_1 \setminus K^*$, $\exists k^* \in K^*$, s.t. $\bar{\mu}_{nk_1} \xrightarrow{p} \mu_{k^*}$, so $\exists \eta$, s.t. $|\bar{\mu}_{nk_1} - \mu_{k^*}| > \eta$, leading to $G_n(\theta^*) - G_n(\theta) = O_p(m(n))$.

As for condition (b), by the fact that $\theta_n$ is an element of set $J$, it is not hard to deduce that only by permutation can $\mu_k^*$ approximate $\mu_k^*$ in arbitrary close distance for $\forall k \in K^*$, this fact results in $G_n(\theta^*) - G_n(\theta_n) = O_p(m(n))$.

Combining with the result in (54), we can reach the outcome that $\sup L_n(\theta) < L_n(\theta^*) \leq L_n(\theta) \implies 1$, adding the fact we have obtained in (53), we complete the proof of (52). With (52) holds, there exists for $\forall \epsilon > 0$, $\forall \theta \in J$, there is a $\eta > 0$ such that $L_n(\theta) < L_n(\theta_n) - \eta$ $\implies 1$, leading to the fact that the event $\{\theta \in J\}$ is contained in the event $\{L_n(\theta_n) < L_n(\theta) - \eta\} \implies 1$, i.e.

$$P(\theta \in J) \leq P(L_n(\theta) < L_n(\theta_n) - \eta) = 0$$

(55)

$P(\theta \in J^c) = 1$ represents what we stated in theorem 2 holds, i.e. $(\tilde{\mu}_k, \tilde{\sigma}_k, \tilde{\alpha}_k) \xrightarrow{p} (\mu_k^*, (\sigma_k^*)^2, \alpha_k^*)$ for $\forall k \in K^*$, $\hat{\alpha}_k \xrightarrow{p} 0$, $\hat{\mu}_k \xrightarrow{p} \tau_k$ for $\forall k \in K^{sc}$. So by now we have complete the proof of the theorem.

6.3. Proof of theorem 2.5.

*Proof.* We adopt the same proof routine as that of theorem 1. It’s not hard to deduce that apart from 2 circumstances that we’ll discuss in the following, consistency conclusions under other circumstances can be easily derived with the same skills as we described in theorem 1.

Firstly, recall that when the 5 clusters all exist, we separate $H$ into 3 sets $H_1$, $H_2$ and $H_3$ to prove (16). More precisely, we divide $H_1$ into 2 parts $Q$ and $Q^c$, derivation of (18) is the same as before, but derivation of $\sup \limits_{\xi \in Q, \alpha_k} L_{n,p}(\xi, \alpha) < L_{n,p}(\tilde{\xi}_p, \tilde{\alpha}_p)$ has to be changed.

Following the same symbols as theorem 1, we assume $\tilde{\theta}_p^* = (\tilde{\xi}_p^*, \tilde{\alpha}_p^*) = \arg \max \limits_{\xi \in Q, \alpha_k} L_{n,p}(\xi, \alpha)$, by the definition of $Q$, the location parameters $\{\tilde{\xi}_p^{pk}\}_{k=1}^5$ of $\tilde{\xi}_p^*$ aren’t arranged in increasing order, and there exists a location permutation rule $\pi_2$ such that $\pi_2(\tilde{\xi}_p^*) = \tilde{\xi}_p^{\pi_2} \xrightarrow{p} \xi^*$, we further denote $\pi_2(\hat{\alpha}_p^*) = \hat{\alpha}_p^{\pi_2}$, $\pi_2(\tilde{\theta}_p^*) = \tilde{\theta}_p^{\pi_2}$. It is obvious that $F_{n,p}(\tilde{\xi}_p^*) = F_{n,p}(\tilde{\theta}_p^{\pi_2})$, but because of the location parameters’ arrangement in $\tilde{\xi}_p^*$, by lemma 2 we have

$$G_n(\tilde{\xi}_p^{\pi_2}) - G_n(\tilde{\xi}_p^*) \geq O_p(m(np)) > 0$$

i.e. $\sup \limits_{\xi \in Q, \alpha_k} L_{n,p}(\xi, \alpha) < L_{n,p}(\tilde{\theta}_p^{\pi_2}) \leq L_{n,p}(\tilde{\xi}_p^*, \tilde{\alpha}_p)$ can be reached. Proof of scenarios under $H_2$ and $H_3$ is parallel with that in theorem 1.

Secondly, when there’s absence of some cluster, we separate $S$ into 3 sets $V$, $V_1^{c}$ and $V_2^{c}$ to prove the correctness of (34). When discussing (34) on set $V$, we consider 2 sub-cases (c) and (d) successively, now in the setting where $n, p \to \infty$, we derive a new proof procedure under case (c).

We still assume $\tilde{\theta}_p^* = \arg \max \limits_{\theta \in V} L_{n,p}(\theta, \alpha)$, by the definition of $V$ and condition (c), $\forall k^* \in K^*$, there must exist at least one $k \in K_1$ such that $\tilde{\xi}_p^{pk} \xrightarrow{p} \xi_k^*$, and we still use set $D(k^*)$ to denote those $k \in K_1$ with the property $\tilde{\xi}_p^{pk} \xrightarrow{p} \xi_k^*$ holds. Under the setting of this theorem, (39) still
can be reached. Now we focus on those \( k^* \) with \( D(k^*) > 1 \), by the fact that \( |K_1| > |K^*| \), there exists such \( k^* \). We pick up a specific term \( \sum_{c(i)=k^*} \log \left[ \sum_{k \in D(k^*)} \alpha_k f(x_i | \tilde{\xi}_{p,k}^i) + O_p(\epsilon^p) \right] = F_{n,p}^k(\tilde{\theta}_p) \) as described above too. For each sample \( i \) belongs to cluster \( k^* \), i.e. \( c(i) = k^* \), we introduce
\[
k^*_c = \arg \max_{k \in D(k^*)} f(x_i | \tilde{\xi}_{p,k}^i),
\]
then we can attain
\[
\log \left[ \sum_{k \in D(k^*)} \alpha_k f(x_i | \tilde{\xi}_{p,k}^i) \right] \leq \log \left[ \left( \sum_{k \in D(k^*)} \alpha_k \right) f(x_i | \tilde{\xi}_{p,k}^i) \right]
\]
(56)

Now we divide the set \( \{i | c(i) = k^* \} \) by their label \( k^*_c \) and we denote the set \( \{i | k^*_c = k_c, c(i) = k^* \} \) as \( C_{k_c}^{k^*} \) for \( k_c \in D(k^*) \), thus for any realization \( \{x_1, \ldots, x_n\} \) of \( \{X_1, \ldots, X_n\} \), these facts can lead to
\[
F_{n,p}^{k^*}(\tilde{\theta}_p) \leq \sum_{k_c \in D(k^*)} \sum_{i \in C_{k_c}^{k^*}} \log \left[ \left( \sum_{k \in D(k^*)} \alpha_k \right) f(x_i | \tilde{\xi}_{p,k}^i) + O\left( \frac{1}{e^p} \right) \right]
\]
(57)

With this in hand, we can also give a new bound of \( F_{n,p}^{k^*}(\tilde{\theta}_p) \) from a hypothesis testing viewpoint. Suppose samples from the same set \( C_{k_c}^{k^*}, k_c \in D(k^*) \) follow the same normal distribution with parameters indexed by \( \tilde{\xi}_{p,k_c}^i \), consider the following hypothesis testing:
\[
H_0 : \tilde{\xi}_{p,k}^i \text{ are equivalent for all } k_c \in D(k^*)
\]
\[
H_1 : \tilde{\xi}_{p,k}^i \text{ are not all equivalent for } k_c \in D(k^*)
\]
(58)

The likelihood under \( H_1 \) is \( \prod_{k_c \in D(k^*)} \prod_{i \in C_{k_c}^{k^*}} f(X_i | \tilde{\xi}_{L,k_c}) \triangleq L_{H_1} \), where \( \tilde{\xi}_{L,k_c} \) is the likelihood of sample \( \{X_{i1}, \ldots, X_{ip}\} \) in \( C_{k_c}^{k^*} \). The likelihood under \( H_0 \) is \( \prod_{k_c \in D(k^*)} \prod_{i \in C_{k_c}^{k^*}} f(X_i | \hat{\xi}_{Lc(i)}) \triangleq L_{H_0} \), whereas \( \hat{\xi}_{Lc(i)} \) is the likelihood of all sample \( \{X_{i1}, \ldots, X_{ip}\} \) with \( i \) satisfying \( c(i) = k^* \). From theory of hypothesis testing we can draw that
\[
\log(L_{H_1}) - \log(L_{H_0}) = O_p(1)
\]
(59)

Furthermore, by the assumption that \( n/e^p \to 0 \), the right side in (57) can be rewritten as \( \sum_{c(i)=k^*} \log \left[ \sum_{k \in D(k^*)} \alpha_k f(x_i | \tilde{\xi}_{p,k}^i) \right] + o(1) \), which is obviously less than \( \sum_{c(i)=k^*} \log \left( \sum_{k \in D(k^*)} \alpha_k \right) + \log(L_{H_1}) + o(1) \), here we refer to \( L_{H_1} \) as a realization corresponding to those \( \{x_1, \ldots, x_n\} \) we have mentioned in (57). Combining with (59) we can update a new bound of \( F_{n,p}^{k^*}(\tilde{\theta}_p) \):
\[
F_{n,p}^{k^*}(\tilde{\theta}_p) \leq \sum_{c(i)=k^*} \log \left( \sum_{k \in D(k^*)} \alpha_k \right) + \sum_{k_c \in D(k^*)} \sum_{i \in C_{k_c}^{k^*}} \log[f(x_i | \hat{\xi}_{Lc(i)})] + O(1)
\]
If we introduce \( \tilde{\xi}_L \) with the property that \( \forall k^* \in K^* , \xi_{Lk^*} = \tilde{\xi}_{Lk^*} = \hat{\xi}_{L(i)} \) if \( c(i) = k^* , \forall k \notin K^* \), \( \tilde{\xi}_{Lk} = \tau_k \). Then a new approximation of \( F_{n,p}(\tilde{\theta}'_p) \) can be obtained

\[
F_{n,p}(\tilde{\theta}'_p) = \sum_{k^* \in K^*} F_{n,p}(\tilde{\xi}'_p) \leq \sum_{k^* \in K^*} \sum \log \left( \sum_{k \in D(k^*)} \alpha_k \right) + O(1)
+ \sum_{k^* \in K^*} \sum_{k \in D(k^*)} \sum \log[f(x_i,\tilde{\xi}_{Lc(i)})] = F_{n,p}(\tilde{\xi}_L, \tilde{\alpha}'_p) + O(1) \tag{60}
\]

By the property of condition (c), for \( \forall k \in K_1 \setminus K^* , |\hat{\mu}_{pk}^i - \tau_k| = O_p(1) \), adding with lemma 2 can lead to

\[
G_n(\tilde{\xi}'_p) - G_n(\tilde{\xi}_L) = -O_p(m(np)) < 0 \tag{61}
\]

The facts we have reached in (60) and (61) direct to the conclusion we want: \( L_{n,p}(\tilde{\theta}'_p) < L_{n,p}(\tilde{\xi}_L, \tilde{\alpha}'_{n,p}) \leq L_{n,p}(\tilde{\xi}_p, \tilde{\alpha}_p) \) wpt 1

6.4. **Proof of theorem 2.7.**

*Proof.* The whole process in proving theorem 2 can be completely adopted in this theorem, so we omit here. \( \square \)

6.5. **Proof of theorem 2.8.**

*Proof.* We adopt similar proof and parameterization skills with that of theorem 2 by Cai et al. (2011). We show only that the type I error tends to 0, proof of the type II error can proceed similarly.

We introduce a substituted testing problem:

\[
H_0^{(s)} : \bar{X}_i \sim N(0, \frac{\sigma_1^2}{p}), \quad 1 \leq i \leq n
\]

\[
H_1^{(s)} : \bar{X}_i \sim (1 - \epsilon)N(0, \frac{\sigma_1^2}{p}) + \epsilon N(A, \frac{\sigma_2^2}{p}), \quad 1 \leq i \leq n \tag{62}
\]

where \( \bar{X}_i = \frac{1}{p} \sum_{t=1}^p X_{it}^i / p \) is the average of the original \( p \)-dimensional observation \( X_i \). We first conclude that to show the type I error of the likelihood ratio test (LRT) with regard to testing problem (15) tends to 0, it is sufficient to move on to testing problem (62). Concretely, it’s sufficient to show that under the null hypothesis \( H_0^{(s)} \) of (62), the log likelihood ratio \( \log(LR_n) = \log(LR_n(\bar{X}_i, \cdots, \bar{X}_n; \epsilon, \sigma_1^2, \sigma_2^2, A, p)) \) \( \rightarrow -\infty \) in probability as \( n, p \rightarrow \infty, \frac{\log(n)}{p} \rightarrow 0 \).

Suppose the log likelihood ratio of testing problem (15) is \( \log(LR_n) \) for \( \epsilon, \sigma_1^2, \sigma_2^2, A, p \), the corresponding rejection region \( W_1 = \{ \log(LR_n(\bar{X}_i, \cdots, \bar{X}_n; \epsilon, \sigma_1^2, \sigma_2^2, A, p)) > 0 \} \). Analogously, rejection region of the LRT with respect to (62) is \( W_2 = \{ \log(LR_n(\bar{X}_i, \cdots, \bar{X}_n; \epsilon, \sigma_1^2, \sigma_2^2, A, p)) > 0 \} \). So we can have 2 testing rules related to (15):

\[
\varphi_1(\bar{X}_1, \cdots, \bar{X}_n) = \begin{cases} 1 & (\bar{X}_1, \cdots, \bar{X}_n) \in W_1 \\ 0 & (\bar{X}_1, \cdots, \bar{X}_n) \in W_1^c \end{cases}
\]

\[
\varphi_2(\bar{X}_1, \cdots, \bar{X}_n) = \begin{cases} 1 & (\bar{X}_1, \cdots, \bar{X}_n) \in W_2 \\ 0 & (\bar{X}_1, \cdots, \bar{X}_n) \in W_2^c \end{cases}
\]
where $W_1^c$ and $W_2^c$ are the complementary set of $W_1$ and $W_2$ separately.

The Neyman-Pearson lemma tells us the optimality of LRT in this test setting, which implies that the type I error of the testing rule $\varphi_1(\bar{X}_1, \cdots, \bar{X}_n)$ is smaller than that of the testing rule $\varphi_2(\bar{X}_1, \cdots, \bar{X}_n)$. So we have

$$P((X_1, \cdots, X_n) \in W_1|H_0) \leq P((\bar{X}_1, \cdots, \bar{X}_n) \in W_2|H_0) = P((\bar{X}_1, \cdots, \bar{X}_n) \in W_2|H_0^{(s)}) \quad (63)$$

The last term in (63) tends to 0 if the claim "$\log(LR_n) \xrightarrow{p} -\infty$ under $H_0^{(s)}$ of (58) as $n, p \to \infty$, $\frac{\log(n)}{p} \to 0$" holds.

To show $\log(LR_n) \to -\infty$ in probability, we conclude that it is sufficient to show that, as $n \to \infty$, $p \to \infty$ and $\frac{\log(n)}{p} \to 0$

$$E[\log(LR_n)] \to -\infty \quad (64)$$

$$\frac{\text{var}\{\log(LR_n)\}}{(E[\log(LR_n)])^2} \to 0 \quad (65)$$

Suppose (64) and (65) hold as $n \to \infty$, $p \to \infty$ and $\frac{\log(n)}{p} \to 0$, then $\forall \epsilon > 0$, $\forall M > 0$, by Chebyshev’s inequality we have

$$P(\log(LR_n) > -M) = P(\log(LR_n) - E[\log(LR_n)] > -M - E[\log(LR_n)]) \leq P(|\log(LR_n) - E[\log(LR_n)]| > | -M - E[\log(LR_n)]|) \leq \frac{\text{var}\{\log(LR_n)\}}{(E[\log(LR_n)] + M)^2} \leq \frac{1}{2}(E[\log(LR_n)])^2 < \epsilon$$

this coincides with $\log(LR_n) \to -\infty$ in probability.

In the following, we prove (64) and (65) separately. Before that, we introduce some notations first.

Let $\epsilon = \epsilon_n = n^{-\beta}$ for a fixed parameter $0 < \beta < 1$, $A = A_n(r) = \sqrt{2r \log(n)}$ for $0 < r < 1$. Denote the density of $N(\mu, \sigma^2)$ by $\phi_\sigma(x - \mu)$ and let

$$g_n(x) = g_n(x; p, \sigma_1^2, \sigma_2^2) = \frac{\phi_{\frac{\sigma_2}{\sqrt{p}}}}{\phi_{\frac{\sigma_1}{\sqrt{p}}}}(x - A)$$

so we have

$$\log(LR_n(\bar{X}_1, \cdots, \bar{X}_n)) = \sum_{i=1}^{n} \log(LR_n(\bar{X}_i)) = \sum_{i=1}^{n} \log(1 - \epsilon + \epsilon g_n(\bar{X}_i))$$

Noticing the fact $E(\log(g_n(\bar{X}_i))) = 1$, we can draw the following approximation:

$$E[\log(1 - \frac{\epsilon}{1 + \epsilon g_n(\bar{X}_i)}) + \epsilon g_n(\bar{X}_i)] = E[\epsilon g_n(\bar{X}_i) - \frac{\epsilon}{1 + \epsilon g_n(\bar{X}_i)}] + O(\epsilon^2)$$

$$= \epsilon E[1 - \frac{1}{1 + \epsilon g_n(\bar{X}_i)}] + O(\epsilon^2) = \epsilon^2 E[\frac{g_n(\bar{X}_i)}{1 + \epsilon g_n(\bar{X}_i)}] + O(\epsilon^2) = O(\epsilon^2) \quad (68)$$

If we let $f_n(x) = \log\{1 + \epsilon g_n(x)\} - \epsilon g_n(x)$, then $E[\log(LR_n)]$ can be expressed as

$$E[\log(LR_n)] = nE[\log(1 - \epsilon + \epsilon g_n(\bar{X}))]$$

$$= nE[\log\{1 + \epsilon g_n(\bar{X})\} - \epsilon g_n(\bar{X})] + nO(\epsilon^2) \quad (69)$$

$$= nE[f_n(\bar{X})] + o(1)$$
Since there is a constant \( c_1 \in (0, 1) \) and a generic constant \( C > 0 \) s.t. \( \log(1 + x) \leq c_1 x \) for \( x > 1 \) and \( \log(1 + x) - x \leq -Cx^2 \) for \( x \leq 1 \), there exists a constant \( C > 0 \) s.t.

\[
E[f_n(\bar{X})] \leq -C \{ \epsilon E[g_n(\bar{X})1_{\{\epsilon g_n(\bar{X}) > 1\}}] + \epsilon^2 E[g_n^2(\bar{X})1_{\{\epsilon g_n(\bar{X}) \leq 1\}}]\}
\leq -C \epsilon E[g_n(\bar{X})1_{\{\epsilon g_n(\bar{X}) > 1\}}]
\]

(70)

So if we can prove that

\[
n\epsilon E[g_n(\bar{X})1_{\{\epsilon g_n(\bar{X}) > 1\}}] \to \infty
\]

(71)

(64) will hold.

We consider (71) in 2 cases: (a)\( \sigma_2^2 < \sigma_1^2 \) and (b)\( \sigma_2^2 > \sigma_1^2 \). Pay attention to (a) first, by elementary calculation, we can get

\[
g_n(\bar{x}) = \frac{\sigma_1}{\sigma_2} e^{\exp\left\{ \frac{p(\sigma_2^2 - \sigma_1^2)(\bar{x} + \frac{A\sigma_1^2}{\sigma_2^2 - \sigma_1^2})^2}{2\sigma_1^2 \sigma_2^2} \right\} e^{\exp\left\{ -\frac{A^2 p}{2(\sigma_2^2 - \sigma_1^2)} \right\}}
\]

(72)

Recalling notations \( \epsilon = n^{-\beta} \) and \( A = \sqrt{2r \log(n)} \), For simplicity, we introduce \( D = \frac{2\sigma_2^2 \sigma_1^2}{\pi(\sigma_2^2 - \sigma_1^2)} \log(n) + \frac{2\sigma_2^2 \sigma_1^2}{\pi(\sigma_2^2 - \sigma_1^2)} (\beta + \frac{r p}{2\sigma_2^2 - \sigma_1^2}) \log(n) \), then under case (a), \( \bar{x} \) satisfying \( \{\epsilon g_n(\bar{x}) > 1\} \) can be expressed as

\[
\bar{x} \in (-\sqrt{D} - \frac{\sigma_1^2 \sqrt{2r \log(n)}}{\sigma_2^2 - \sigma_1^2}, \sqrt{D} - \frac{\sigma_1^2 \sqrt{2r \log(n)}}{\sigma_2^2 - \sigma_1^2}) \triangleq (A_1, A_2)
\]

\[
\subseteq (-\frac{\sigma_1}{\sigma_1 + \sigma_2} \sqrt{2r \log(n)}, \frac{\sigma_1}{\sigma_1 - \sigma_2} \sqrt{2r \log(n)}) \triangleq (B_1, B_2)
\]

(73)

here \( A_1 = -\sqrt{D} - \frac{\sigma_1^2 \sqrt{2r \log(n)}}{\sigma_2^2 - \sigma_1^2} \), \( A_2 = \sqrt{D} - \frac{\sigma_1^2 \sqrt{2r \log(n)}}{\sigma_2^2 - \sigma_1^2} \), \( B_1 = \frac{\sigma_1}{\sigma_1 + \sigma_2} \sqrt{2r \log(n)} \), \( B_2 = \frac{\sigma_1}{\sigma_1 - \sigma_2} \sqrt{2r \log(n)} \). The “\( \subseteq \)” comes from dropping 2 terms \( \frac{2\sigma_2^2 \sigma_1^2}{\pi(\sigma_2^2 - \sigma_1^2)} \log(n) \) and \( \frac{2\sigma_2^2 \sigma_1^2}{\pi(\sigma_2^2 - \sigma_1^2)} (\beta \log(n)) \) in \( D \), and denote the left term as \( D_1 = \frac{2\sigma_2^2 \sigma_1^2}{\pi(\sigma_2^2 - \sigma_1^2)} \log(n) \), we expect when \( n \) and \( p \to \infty \), \( \frac{\log(n)}{p} \to 0 \), these 2 terms play negligible role compared with the left term \( \frac{2\sigma_2^2 \sigma_1^2}{\pi(\sigma_2^2 - \sigma_1^2)} \log(n) \), i.e.

\[
\frac{E[g_n(\bar{X})1_{\{\bar{x} \in (B_1, A_1)\}}] + E[g_n(\bar{X})1_{\{\bar{x} \in (A_2, B_2)\}}]}{E[g_n(\bar{X})1_{\{\bar{x} \in (B_1, B_2)\}}]} \to 0
\]

(74)

If (74) holds, then

\[
n\epsilon E[g_n(\bar{X})1_{\{\epsilon g_n(\bar{x}) > 1\}}] = n\epsilon E[g_n(\bar{X})1_{\{\bar{x} \in (A_1, A_2)\}}]
\]

\[
= n\epsilon E[g_n(\bar{X})1_{\{\bar{x} \in (B_1, B_2)\}}](1 - o(1))
\]

that is if we relax the interval into a more tractable one, the asymptotic outcome doesn’t change at all. We prove (74) later and now we move our sight back to (71) on the case \( \sigma_2^2 < \sigma_1^2 \). In this setting, we want to obtain

\[
n\epsilon E[g_n(\bar{X})1_{\{\bar{x} \in (B_1, B_2)\}}] \to \infty
\]

(75)

By Mill’s ratio (Wasserman, 2006),

\[
\Phi(\sqrt{2q \log(n)}) = PL(n)n^{-q}
\]
here \( \Phi = 1 - \Phi \) is the survival function of \( N(0, 1) \), \( PL(n) > 0 \) is a generic poly-log-term satisfying
\[
\lim_{n \to \infty} \{PL(n)n^{-\delta}\} = 0 \quad \text{and} \quad \lim_{n \to \infty} \{PL(n)n^{\delta}\} = 0 \quad \text{for any } \delta > 0.
\]
So we can get
\[
n\epsilon E[g_n(\bar{X})1_{\{\bar{x} \in (B_1, B_2)\}}] = n^{-1-\beta} \int_{B_1}^B \phi_{\frac{\bar{x}}{\sqrt{p}}} \left(x - \sqrt{2r \log(n)} \right) dx
= n^{-1-\beta} \left[1 - PL_1(n)n^{-\frac{rp}{(\sigma_1 + \sigma_2)^2}} - PL_2(n)n^{-\frac{rp}{(\sigma_1 - \sigma_2)^2}} \right]
\]
where \( PL_1(n) \) and \( PL_2(n) \) are 2 generic poly-log-terms, it is easy to see that when \( n, p \to \infty, 1 - PL_1(n)n^{-\frac{rp}{(\sigma_1 + \sigma_2)^2}} - PL_2(n)n^{-\frac{rp}{(\sigma_1 - \sigma_2)^2}} \to 1 \), resulting in (75) holds. Furthermore, if (74) holds, we will complete the proof of (71) in the \( \sigma_2 < \sigma_1^2 \) case. Notice that
\[
E[g_n(\bar{X})1_{\{\bar{x} \in (A_2, B_2)\}}] = \Phi(\frac{B_2 - \sqrt{2r \log(n)}}{\sigma_2/\sqrt{p}}) - \Phi(\frac{A_2 - \sqrt{2r \log(n)}}{\sigma_2/\sqrt{p}}),
\]
\[
\frac{B_2 - A_2}{\sigma_2/\sqrt{p}} = \frac{\sqrt{D_1} - \sqrt{D}}{\sigma_2/\sqrt{p}} = \frac{\sqrt{p}}{\sigma_2} \frac{D_1 - D}{\sqrt{D_1} + \sqrt{D}} = \frac{2\sigma_1^2 \sigma_2}{\sqrt{p} (\sigma_1^2 - \sigma_2^2)} (\sqrt{D_1} + \sqrt{D}).
\]
Since \( \sqrt{D} \) and \( \sqrt{D_1} \) are all bounded away from 0,
\[
\frac{2\beta \sigma_1^2 \sigma_2}{\sqrt{p} (\sigma_1^2 - \sigma_2^2)} \log(n) \leq \frac{2\beta \sigma_1^2 \sigma_2}{\sqrt{p} (\sigma_1^2 - \sigma_2^2)} \sqrt{D_1} + \sqrt{D} = C_1 \frac{\log(n)}{p} \to 0
\]
here \( C_1 \) is a constant having no relationship with \( p \) or \( n \). By far we can get \( \frac{B_2 - A_2}{\sigma_2/\sqrt{p}} \to 0 \), i.e. \( E[g_n(\bar{X})1_{\{\bar{x} \in (A_2, B_2)\}}] \to 0 \), proof of the left part \( E[g_n(\bar{X})1_{\{\bar{x} \in (B_1, A_2)\}}] \to 0 \) can be attained in a similar way. Combining the fact we have reached in (76), we have finished the proof of (74).

Next we will focus on (71) under case (b), in this setting, \( \bar{x} \) satisfying \( \{\epsilon g_n(\bar{x}) > 1\} \) can be expressed as
\[
\bar{x} \in (-\infty, -\sqrt{D} - \frac{\sigma_1^2 \sqrt{2r \log(n)}}{\sigma_2^2 - \sigma_1^2}) \cup (\sqrt{D} - \frac{\sigma_1^2 \sqrt{2r \log(n)}}{\sigma_2^2 - \sigma_1^2}, +\infty)
\]
\[
\equiv (-\infty, A_1) \cup (A_2, +\infty)
\]
\[
\subseteq (-\infty, \frac{\sigma_1}{\sigma_1 - \sigma_2} \sqrt{2r \log(n)}) \cup (\frac{\sigma_1}{\sigma_1 + \sigma_2} \sqrt{2r \log(n)}, +\infty)
\]
\[
\equiv (-\infty, B_2) \cup (B_1, +\infty)
\]
where the "\( \subseteq \)" also comes from changing \( D \) into \( D_1 \). With similar proof skills as above, we can also get under case (b),
\[
n\epsilon E[g_n(\bar{X})1_{\{\epsilon g_n(\bar{x}) > 1\}}] = n\epsilon E[g_n(\bar{X})1_{\{\bar{x} \in (-\infty, A_1) \cup (A_2, +\infty)\}}]
= n\epsilon E[g_n(\bar{X})1_{\{\bar{x} \in (-\infty, B_2) \cup (B_1, +\infty)\}}] (1 - o(1))
\]
So when \( \sigma_2^2 > \sigma_1^2 \), if we can attain
\[
n\epsilon E[g_n(\bar{X})1_{\{\bar{x} \in (-\infty, B_2) \cup (B_1, +\infty)\}}] \to \infty
\]
(64) will be fulfilled completely.
By Mill’s ratio and some calculations,
\[ n \epsilon E[g_n(\bar{X})1_{(x \in (-\infty, B_2) \cup (B_1, +\infty))}] = n^{1-\beta} [\int_{-\infty}^{B_2} \phi_{\frac{\mu}{\sigma}}(x - \sqrt{2r \log(n)})dx + \int_{B_1}^{+\infty} \phi_{\frac{\mu}{\sigma}}(x - \sqrt{2r \log(n)})dx] = n^{1-\beta} (1 - PL_3(n)n^{\frac{rp(2\sigma_1 - \sigma_2)}{\sigma_2^2(\sigma_1 + \sigma_2)^2}} - PL_4(n)n^{-\frac{rp}{(\sigma_1 + \sigma_2)^2}}) \] (79)

where \( PL_3(n) \) and \( PL_4(n) \) are also 2 generic poly-log-terms, when \( n, p \to \infty, (1 - PL_3(n)n^{\frac{rp(2\sigma_1 - \sigma_2)}{\sigma_2^2(\sigma_1 + \sigma_2)^2}} - PL_4(n)n^{-\frac{rp}{(\sigma_1 + \sigma_2)^2}}) \to 1, \) resulting in (78) holds. By far we have completed (64), next we show assumption (65).

If we can show that there exists a constant \( C > 0 \) s.t.
\[ E[\log^2(LR_n(\bar{X}))] \leq -CE[\log(LR_n(\bar{X}))] \] (80)
we will have
\[
\frac{\text{var}\{\log(LR_n)\}}{(E[\log(LR_n)])^2} = \frac{n(E[\log^2(LR_n(\bar{X}))] - (E[\log(LR_n(\bar{X}))])^2)}{n^2(E[\log(LR_n(\bar{X}))])^2} \\
\leq -\frac{1}{n} (1 + \frac{C}{E[\log(LR_n(\bar{X}))]}) = -\frac{1}{n} - \frac{C}{E[\log(LR_n)]} \to 0
\]
So it is sufficient to prove (80) holds. Since for all \( x, \) \( \log^2(1 - \frac{\epsilon}{1+\epsilon g_n(x)}) \leq \left[ \frac{\epsilon}{1+\epsilon g_n(x)} \right]^2 \leq \epsilon^2, \) we can get
\[
\log^2\{1 - \epsilon + \epsilon g_n(x)\} = [\log\{1 - \frac{\epsilon}{1 + \epsilon g_n(x)}\} + \log\{1 + \epsilon g_n(x)\}]^2 \\
\leq 2[\log^2\{1 - \frac{\epsilon}{1 + \epsilon g_n(x)}\} + \log^2\{1 + \epsilon g_n(x)\}] \\
\leq 2[\epsilon^2 + \log^2\{1 + \epsilon g_n(x)\}] \\
\]
resulting in
\[ E[\log^2(LR_n(\bar{X}))] \leq E[\log^2\{1 + \epsilon g_n(\bar{X})\}] + o\left(\frac{1}{n}\right) \]
Furthermore, \( \log(1 + x) < C\sqrt{x} \) for \( x > 1 \) and \( \log(1 + x) < x \) for \( x > 0, \)
\[ E[\log^2\{1 + \epsilon g_n(\bar{X})\}] \leq C\epsilon E[g_n(\bar{X})1_{\{g_n(\bar{X}) > 1\}}] + \epsilon^2 E[g_n^2(\bar{X})1_{\{g_n(\bar{X}) \leq 1\}}] \] (81)

From (69) and (70) we have
\[ E[\log(LR_n(\bar{X}))] = E[f_n(\bar{X})] + o\left(\frac{1}{n}\right) \]
\[ \leq -C\epsilon E[g_n(\bar{X})1_{\{g_n(\bar{X}) > 1\}}] + \epsilon^2 E[g_n^2(\bar{X})1_{\{g_n(\bar{X}) \leq 1\}}] \] (82)
Combining (81), (82) and comparing with (80), we now reached assumption (65). So the full proof of the theorem is completed.
6.6. Proof of theorem 2.9.

Proof. (a) We adopt the same symbols as described in section 2.6, for simplicity we drop case’s symbol "b" and suppose sample size of case is $n$.

Firstly, we recall some notations, $\hat{\theta}_{b-(b+1)}$, $\hat{\theta}_b$ and $\hat{\theta}_{b+1}$ are estimators such that $\hat{f}(X_{1,b-(b+1)}, \cdots, X_{n,b-(b+1)}|\theta_{b-(b+1)})$, $\hat{f}(X_{1,b}, \cdots, X_{n,b}|\theta_b)$ and $\hat{f}(X_{1,b+1}, \cdots, X_{n,b+1}|\theta_{b+1})$ attain maximum value, $\hat{\theta}_{b-(b+1)}$ and $\hat{\theta}_{b-(b+1)}^{(2)}$ are separated from $\theta_{b-(b+1)}$, corresponding to parameters in the $b$-th bin and $(b+1)$-th bin separately. Concretely, if bin $b$ and $b+1$ are merged, $X_{i,b-(b+1)} \sim \sum_{k=1}^{5} \alpha_{b-(b+1),k} N(\mu_{b-(b+1),k}, (\sigma_{b-(b+1),k})^2)$, the conditional distribution of $X_{i,b+1}|X_{i,b}$ can be expressed as

$$q(X_{i,b+1}|X_{i,b}, \theta_{b-(b+1)}) = \frac{\sum_{k=1}^{5} \alpha_{b-(b+1),k} \hat{f}(X_{i,b-(b+1)}|\mu_{b-(b+1),k}, (\sigma_{b-(b+1),k})^2))}{\sum_{k=1}^{5} \alpha_{b-(b+1),k} \hat{f}(X_{i,b}|\mu_{b-(b+1),k}, (\sigma_{b-(b+1),k})^2))}$$

where the denominator is the marginal distribution of $X_{i,b}$.

If bin $b$ and $b+1$ can’t be merged, the joint distribution of $X_{i,b}$ and $X_{i,b+1}$ is

$$X_{i,b-(b+1)} \sim \gamma \left[ \sum_{k=1}^{5} \alpha_{b-(b+1),k} N(\mu_{b-(b+1),k}, (\sigma_{b-(b+1),k})^2) \right]$$

$$+ (1 - \gamma) \left[ \sum_{k_2=1}^{5} \sum_{k_1=1}^{5} \alpha_{k_1,k_2} N((\mu_{b,k_1}, \mu_{b+1,k_2}), ((\sigma_{b,k_1})^2, (\sigma_{b+1,k_2})^2)) \right]$$

where $\gamma$ is the proportion of samples which have the same CN state between bin $b$ and $b+1$, whereas $1 - \gamma$ is the proportion of heterogeneous samples between these 2 bins, the 2 bins can be merged if and only if $\lambda = 0$. $\alpha_{k_1,k_2}$ is the proportion of heterogeneous samples which have CN state $k_1$ on bin $b$ and have CN state $k_2$ on bin $b+1$. Since we didn’t assume batch effect within a CN state, $\mu_{b-(b+1),k}^{(1)} = \mu_{b,k_1}$, $\sigma_{b-(b+1),k}^{(1)} = (\sigma_{b,k_1})^2$ for $k_1 = k$; $\mu_{b-(b+1),k}^{(2)} = \mu_{b+1,k_2}$, $\sigma_{b-(b+1),k}^{(2)} = (\sigma_{b+1,k_2})^2$ for $k_2 = k$. Under this situation, we denote $X_{i,b-(b+1)} \sim H(X_{i,b-(b+1)}|\kappa, \lambda)$, $X_{i,b+1}|X_{i,b} \sim \tilde{H}(X_{i,b+1}|X_{i,b}, \kappa, \lambda)$ for simplicity, where $\kappa$ denotes all the parameters in the distribution except $\lambda$.

We further assume $\tilde{\theta}_b = \arg\max_{\theta_b} \sum_{i=1}^{n} \log[h(X_{i,b}|\theta_b)]$, where $h(X_{i,b}|\theta_b) = \sum_{i=1}^{n} \alpha_{b,k} \hat{f}(X_{i,b}|\mu_{b,k}, (\sigma_{b,k})^2)$.

$$(\kappa, \lambda) = \arg\max_{\kappa, \lambda} \sum_{i=1}^{n} \log[\tilde{H}(X_{i,b+1}|X_{i,b}, \kappa, \lambda)], (\kappa, \lambda) = \arg\max_{\kappa, \lambda} \sum_{i=1}^{n} \log[H(X_{i,b-(b+1)}|\kappa, \lambda)],$$

concretely, we suppose $\kappa = (\alpha_{b-(b+1),k}^{(1)} \mu_{b-(b+1),k}^{(1)} (\sigma_{b-(b+1),k}^{(1)})^2, \alpha_{k_1,k_2}^{(2)})$ for $k = 1, \cdots, 5; k_1 = 1, \cdots, 5, k_2 = 1, \cdots, 5, k_1 \neq k_2$. 

42
If we consider the following test:

\[
H_0^{(1)}: \tilde{\alpha}_{b,k} = \alpha^{\tilde{\kappa}}_{b-(b+1),k} + \sum_{k_2 = 1}^{5} \alpha^{\tilde{\kappa}}_{k_2} \mu^{\tilde{\kappa}}_{b,k_2}, \quad \tilde{\mu}^{\tilde{\kappa}}_{b,k} = \mu^{\tilde{\kappa}}_{b-(b+1),k} \quad (\tilde{\sigma}^{\tilde{\kappa}}_{b,k})^2 = (\sigma^{\tilde{\kappa}}_{b-(b+1),k})^2, \quad 1 \leq k \leq 5
\]

\[
H_1^{(2)}: \tilde{\alpha}_{b,k} \neq \alpha^{\tilde{\kappa}}_{b-(b+1),k} + \sum_{k_2 = 1}^{5} \alpha^{\tilde{\kappa}}_{k_2} \mu^{\tilde{\kappa}}_{b,k_2} \quad \text{or} \quad \tilde{\mu}^{\tilde{\kappa}}_{b,k} = \mu^{\tilde{\kappa}}_{b-(b+1),k} \quad (\tilde{\sigma}^{\tilde{\kappa}}_{b,k})^2 = (\sigma^{\tilde{\kappa}}_{b-(b+1),k})^2, \quad \exists k \in \{1, \ldots, 5\}
\]

Under the assumption that \( X_{i,b-(b+1)} \sim H(X_{i,b-(b+1)}|\kappa, \lambda) \), which satisfies conditions in \( H_0 \), the maximum likelihood is \( \sum_{i=1}^{n} \log[H(X_{i,b-(b+1)}|\kappa, \lambda)] \). From properties of likelihood ratio test we can conclude that the difference between \( H_0 \) and \( H_1 \) is \( O_p(1) \), concretely,

\[
\sum_{i=1}^{n} \log[h(X_{i,b} | \tilde{\theta}_b)] + \sum_{i=1}^{n} \log[\tilde{H}(X_{i,b+1} | X_{i,b}, \tilde{\kappa}, \tilde{\lambda})] - \sum_{i=1}^{n} \log[H(X_{i,b-(b+1)} | \kappa, \lambda)] = O_p(1) \quad (83)
\]

Under the new hypothesis framework:

\[
H_0^{(2)}: \lambda = 0, \quad H_1^{(2)}: \lambda \neq 0
\]

Suppose \( \tilde{\theta}_{b-(b+1)} = \sum_{i=1}^{n} \log[h(X_{i,b-(b+1)}|\tilde{\theta}_{b-(b+1)})] \), if bin \( b \) and \( b+1 \) can be merged together, which is equivalent to \( H_0^{(2)} \) is true, the difference of maximum likelihood under \( H_0^{(2)} \) and \( H_1^{(2)} \) is \( O_p(1) \):

\[
\sum_{i=1}^{n} \log[h(X_{i,b-(b+1)}|\tilde{\theta}_{b-(b+1)})] - \sum_{i=1}^{n} \log[H(X_{i,b-(b+1)}|\tilde{\kappa}, \tilde{\lambda})] = O_p(1) \quad (84)
\]

since \( \tilde{\theta}_{b-(b+1)} = \text{argmax}_{\theta_{b-(b+1)}} \tilde{f}(X_{1,b-(b+1)}, \ldots, X_{n,b-(b+1)}) \), we have

\[
\sum_{i=1}^{n} \log[h(X_{i,b-(b+1)}|\tilde{\theta}_{b-(b+1)})] - \sum_{i=1}^{n} \log[h(X_{i,b-(b+1)}|\tilde{\theta}_{b-(b+1)})] < O_p(m(np)) \quad (85)
\]

here \( m(np) \) is any increasing function of \( n, p \) with order less than \( O(np) \), and \( p \) can be length of the merged bin \( b-(b+1) \). Similarly,

\[
0 \leq \sum_{i=1}^{n} \log[h(X_{i,b} | \tilde{\theta}_b)] - \sum_{i=1}^{n} \log[h(X_{i,b} | \tilde{\theta}_b)] < O_p(m(np)) \quad (86)
\]

Combining (83), (84) and (85) we can get

\[
\sum_{i=1}^{n} \log[h(X_{i,b} | \tilde{\theta}_b)] + \sum_{i=1}^{n} \log[\tilde{H}(X_{i,b+1} | X_{i,b}, \tilde{\kappa}, \tilde{\lambda})] < \sum_{i=1}^{n} \log[h(X_{i,b-(b+1)}|\tilde{\theta}_{b-(b+1)})] + O_p(m(np)) \quad (87)
\]

by the fact that

\[
\sum_{i=1}^{n} \log[h(X_{i,b-(b+1)}|\tilde{\theta}_{b-(b+1)})] = \sum_{i=1}^{n} \log[q(X_{i,b+1} | X_{i,b}, \hat{\theta}_{b-(b+1)})] + \sum_{i=1}^{n} \log[h(X_{i,b} | \tilde{\theta}^{(1)}_{b-(b+1)})]
\]
(87) can be reformulated as
\[
\sum_{i=1}^{n} \log [h(X_{i,b} | \hat{\theta}_b)] - \sum_{i=1}^{n} \log [h(X_{i,b} | \hat{\theta}_{b-(b+1)})] < \sum_{i=1}^{n} \log [q(X_{i,b+1} | X_{i,b}, \hat{\theta}_{b-(b+1)})] - \sum_{i=1}^{n} \log [\tilde{H}(X_{i,b+1} | X_{i,b}, \hat{b}, \tilde{\lambda})] + O_p(m(np))
\]
\[
\leq O_p(m(np))
\]
Combining (86) and (88) we can reach the conclusion:
\[
\sum_{i=1}^{n} \log [h(X_{i,b} | \hat{\theta}_b)] - \sum_{i=1}^{n} \log [h(X_{i,b} | \hat{\theta}_{b-(b+1)})] \leq O_p(m(np))
\]

By far we have completed circumstance of case sample, for control sample, the proof skill is completely the same, so the order claimed in the theorem can be reached.

(b) The idea is completely the same as proposition 1. We follow symbols in (a), \(\hat{\theta}_b\) and \(\hat{\theta}_{b+1}\) are estimators of bin \(b\) and \(b+1\) before merging, \(\theta_{b-(b+1)}\) are estimators after merging. So value of \(M_{b,b+1}\) is denoted as:
\[
M_{b,b+1} = \sum_{i=1}^{n} \log [h(X_{i,b} | \hat{\theta}_b)] - \sum_{i=1}^{n} \log [h(X_{i,b} | \hat{\theta}_{b-(b+1)})] + \sum_{i=1}^{n} \log [h(X_{i,b+1} | \hat{\theta}_{b+1})] - \sum_{i=1}^{n} \log [h(X_{i,b+1} | \hat{\theta}_{b-(b+1)})]
\]

By law of large numbers and consistency of \(\hat{\theta}_b, \hat{\theta}_{b+1}\), i.e. \(\hat{\theta}_b \xrightarrow{p} \theta_b, \hat{\theta}_{b+1} \xrightarrow{p} \theta_{b+1}\), we have
\[
\frac{1}{n} \sum_{i=1}^{n} \left\{ \log [h(X_{i,b} | \hat{\theta}_b)] - \log [h(X_{i,b} | \hat{\theta}_{b-(b+1)})] \right\} = \int \log \frac{h(X_b | \theta_b)}{h(X_b | \hat{\theta}_{b-(b+1)})} h(X_b | \theta_b) dX_b + o_p(1)
\]
\[
\frac{1}{n} \sum_{i=1}^{n} \left\{ \log [h(X_{i,b+1} | \hat{\theta}_{b+1})] - \log [h(X_{i,b+1} | \hat{\theta}_{b-(b+1)})] \right\} = \int \log \frac{h(X_{b+1} | \theta_{b+1})}{h(X_{b+1} | \hat{\theta}_{b-(b+1)})} h(X_{b+1} | \theta_{b+1}) dX_{b+1} + o_p(1)
\]
Since \(\exists c_s > 0, \exists k \in \{1, \cdots, 5\}\), s.t. \(|\alpha_{bk} - \alpha_{b-(b+1),k}| > c_s\), there must be \(|\alpha_{bk} - \alpha_{b-(b+1),k}| > \frac{c_s}{2}\), resulting in one of the 2 K-L divergence above attain order \(O_p(1)\), i.e. the order of \(M_{b,b+1}\) is \(O_p(n)\)

6.7. Proof of lemma 2.2.

Proof: For convinence we introduce the symbol \(\pi(\{1, \cdots, 5\}) = \{\pi(1), \cdots, \pi(5)\}\), \(\mu^*_k - \tau_k = \zeta_k\) with \(|\zeta_k| < \zeta\) for \(k = 1, \cdots, 5\), and assume \(e_k = \zeta_{k+1} - \zeta_k\) for \(k \in \{1, \cdots, 4\}\), then by the setting in lemma 2, \(|e_k| < \min_{k \in \{1, \cdots, 4\}} d_k\) for \(k = 1, \cdots, 4\). Since the prior variance \(\sigma^2_{r_k}\) for \(k \in \{1, \cdots, 5\}\),
are the same, (7) is equivalent to
\[ \sum_{k=1}^{5} (\mu_k^* - \tau_k)^2 < \sum_{k=1}^{5} (\mu_{\pi(k)}^* - \tau_k)^2 \]
which is also the same as
\[ \sum_{k=1}^{5} (\tau_k + \zeta_k)\tau_k > \sum_{k=1}^{5} (\tau_{\pi(k)} + \zeta_{\pi(k)})\tau_k \]
i.e.
\[ \sum_{k=1}^{5} (\tau_k - \tau_{\pi(k)})\tau_k > \sum_{k=1}^{5} (\zeta_{\pi(k)} - \zeta_k)\tau_k \]
if we introduce \( d_k \) and \( e_k \) to substitute \( \tau_k - \tau_{\pi(k)} \) and \( \zeta_{\pi(k)} - \zeta_k \), we will have
\[ \sum_{k=1}^{5} \left( \sum_{j=k}^{\pi(k)-1} d_j (-1)^{1_{\{\pi(k) > k\}}} \right) \tau_k > \sum_{k=1}^{5} \left( \sum_{j=k}^{\pi(k)-1} e_j (-1)^{1_{\{\pi(k) > k\}}} \right) \tau_k \]
noticing that \( \tau_k = \tau_1 + \sum_{l=1}^{k-1} d_l \) and \( \sum_{k=1}^{5} \tau_k^2 > \sum_{k=1}^{5} \tau_k \tau_{\pi(k)} \) for any nonidentical location permutation
rule \( \pi \), which is because the arrangement of \( \{\tau_k\}_{k=1}^{5} \) is in increasing order, then \( \sum_{k=1}^{5} (\tau_k - \tau_{\pi(k)})\tau_k \)
is the summation of some \( d_k,d_{k_2} \) for \( k_1,k_2 \in \{1,\ldots,5\} \), from this, there must exist 5 functions
of \( \tau_1,\ldots,\tau_5 \) which can be expressed as \( v_k(\tau_1,\ldots,\tau_5) > 0 \) for \( k = 1,\ldots,5 \) such that
\[ \sum_{k=1}^{5} \left( \sum_{j=k}^{\pi(k)-1} d_j (-1)^{1_{\{\pi(k) > k\}}} \right) \tau_k = \sum_{k=1}^{5} d_k v_k(\tau_1,\ldots,\tau_5) \]
(91)
Since the structures of the two sides of inequality in (90) are the same, we can also get a similar
result as (91)
\[ \sum_{k=1}^{5} \left( \sum_{j=k}^{\pi(k)-1} e_j (-1)^{1_{\{\pi(k) > k\}}} \right) \tau_k = \sum_{k=1}^{5} -e_k v_k(\tau_1,\ldots,\tau_5) \]
(92)
Combining the fact that \( |e_j| < \min_{k \in \{1,\ldots,5\}} d_k \) and \( v_k(\tau_1,\ldots,\tau_5) > 0 \) for \( k = 1,\ldots,5 \),
\[ \sum_{k=1}^{5} -e_k v_k(\tau_1,\ldots,\tau_5) < \sum_{k=1}^{5} d_k v_k(\tau_1,\ldots,\tau_5) \]
which implies (89) and (90) hold, that is what we aim to prove.
\[ \square \]
6.8. Proof of proposition 2.6.

Proof. In our proof we drop the bin’s index \( b \) for simplicity. Suppose case samples \( X_i \sim \sum_{k=1}^{5} \alpha_i \| X_i \| \mu_k^d, (\sigma_k^d)^2 \), control samples \( Y_j \sim \sum_{k=1}^{5} \alpha_c \| Y_j \| \mu_k^c, (\sigma_k^c)^2 \), where \( f(x|\mu_k, (\sigma_k^d)^2) \) is Gaussian
distribution with mean \( \mu_k^d \) and covariance matrix \( Diag((\sigma_k^d)^2) \) valued on \( x \), interpretation of \( f(y|\mu_k^c, (\sigma_k^c)^2) \) is similar. We further suppose estimators of case and control used to calculate
likelihood under \( H_1 \) are \( (\hat{\alpha}_k^d, \hat{\mu}_k^d, (\sigma_k^d)^2)_{k=1}^{5} \triangleq \theta_{H_1}^d \) and \( (\hat{\alpha}_k^c, \hat{\mu}_k^c, (\sigma_k^c)^2)_{k=1}^{5} \triangleq \theta_{H_1}^c \), which can be
obtained by maximizing (9) and control’s analogous form, these parameters have consistency by theorem 4, i.e. \( \hat{\theta}^d_{H_1} \xrightarrow{p} \theta^d_{H_1}, \hat{\theta}^c_{H_1} \xrightarrow{p} \theta^c_{H_1} \). If estimators of proportion parameters under \( H_0 \) are denoted as \( \hat{\omega}_{kH_0} \), \( k = 1, \cdots, 5 \), and the overall parameters of case and control under \( H_0 \) are \( (\hat{\alpha}_{kH_0}, \hat{\mu}_k, (\hat{\sigma}_k^c)^2)_{k=1}^5 \) and \( (\hat{\alpha}_{kH_0}, \hat{\mu}_k, (\hat{\sigma}_k^c)^2)_{k=1}^5 \), we can write the likelihood ratio value as:

\[
\Lambda = \sum_{i=1}^{N_1} \log [h(X_i | \hat{\theta}^d_{H_1})] - \sum_{i=1}^{N_1} \log [h(X_i | \hat{\theta}^d_{H_0})] + \sum_{j=1}^{N_2} \log [h(Y_j | \hat{\theta}^c_{H_1})] - \sum_{j=1}^{N_2} \log [h(Y_j | \hat{\theta}^c_{H_0})]
\]

where \( h(X_i | \hat{\theta}^d_{H_1}) \) is the mixture Gaussian density function \( \sum_{k=1}^5 \hat{\alpha}_k f(X_i | \hat{\mu}_k, (\hat{\sigma}_k^d)^2) \), other 3 functions \( h \) have similar interpretation.

By law of large numbers and consistency of \( \hat{\theta}^d_{H_1}, \hat{\theta}^c_{H_1} \), we have

\[
\frac{1}{N_1} \sum_{i=1}^{N_1} \left\{ \log [h(X_i | \hat{\theta}^d_{H_1})] - \log [h(X_i | \hat{\theta}^d_{H_0})] \right\} = \int \frac{h(X | \hat{\theta}^d_{H_1})}{h(X | \hat{\theta}^d_{H_0})} h(X | \theta^d_{H_1}) dX + o_p(1)
\]

\[
\frac{1}{N_2} \sum_{j=1}^{N_2} \left\{ \log [h(Y_j | \hat{\theta}^c_{H_1})] - \log [h(Y_j | \hat{\theta}^c_{H_0})] \right\} = \int \frac{h(Y | \hat{\theta}^c_{H_1})}{h(Y | \hat{\theta}^c_{H_0})} h(Y | \theta^c_{H_1}) dY + o_p(1)
\] (93)

By the assumption that \( \exists c_b > 0, \exists k \in \{1, \cdots , 5\} \), s.t. \( |\alpha_k^d - \hat{\alpha}_k| < c_b \) or \( |\alpha_k^c - \hat{\alpha}_k| < c_b \), leading to one of the 2 K-L divergences in (87) attain order \( O_p(1) \), so the likelihood ratio value \( \Lambda = O_p(N_1) \) or \( \Lambda = O_p(N_2) \), which means that the type II error of our test tends to 0.

\[
\square
\]

REFERENCES

[1] Zhang F, Gu W, Hurles M E, et al. Copy number variation in human health, disease, and evolution[J]. Annual review of genomics and human genetics, 2009, 10: 451-481.
[2] Carter N P. Methods and strategies for analyzing copy number variation using DNA microarrays[J]. Nature genetics, 2007, 39: S16.
[3] Zhao M, Wang Q, Wang Q, et al. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives[J]. BMC bioinformatics, 2013, 14(11): S1.
[4] Harchaoui Z, Lévy-Leduc C. Multiple change-point estimation with a total variation penalty[J]. Journal of the American Statistical Association, 2010, 105(492): 1480-1493.
[5] Olshen A B, Venkatraman E S, Lucito R, et al. Circular binary segmentation for the analysis of array-based DNA copy number data[J]. Biostatistics, 2004, 5(4): 557-572.
[6] Zhang N R, Siegmund D O, Ji H, et al. Detecting simultaneous changepoints in multiple sequences[J]. Biometrika, 2010, 97(3): 631-645.
[7] Jeng X J, Cai T T, Li H. Simultaneous discovery of rare and common segment variants[J]. Biometrika, 2012, 100(1): 157-172.
[8] Donoho D, Jin J. Higher criticism for detecting sparse heterogeneous mixtures[J]. The Annals of Statistics, 2004, 32(3): 962-994.
[9] Itsara A, Cooper G M, Baker C, et al. Population analysis of large copy number variants and hotspots of human genetic disease[J]. The American Journal of Human Genetics, 2009, 84(2): 148-161.
[10] Tony Cai, T., X. Jessie Jeng, and Jiashun Jin. “Optimal detection of heterogeneous and heteroscedastic mixtures.” Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73.5 (2011): 629-662.
[11] Wasserman, L. (2006) All of Nonparametric statistics. New York: Springer
[12] Barnes C, Plagnol V, Fitzgerald T, et al. A robust statistical method for case-control association testing with copy number variation[J]. Nature genetics, 2008, 40(10): 1245.

46
[13] Daniel Hsu and Sham M Kakade. Learning mixtures of spherical Gaussians: Moment methods and spectral decompositions. In Proceedings of the 4th Conference on Innovations in Theoretical Computer Science, pages 11–20. ACM, 2013

[14] C. Jin, Y. Zhang, S. Balakrishnan, M. J. Wainwright, and M. I. Jordan. On local maxima in the population likelihood of gaussian mixture models: Structural results and algorithmic consequences. In Advances In Neural Information Processing Systems, pages 4116–4124, 2016

[15] Scrucca L., Fop M., Murphy T. B. and Raftery A. E. (2016) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models The R Journal 8/1, pp. 205-233

[16] Dacunha-Castelle D, Gassiat E. Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes[J]. The Annals of Statistics, 1999, 27(4): 1178-1209.

[17] Keribin C. Consistent estimation of the order of mixture models[J]. Sankhyä: The Indian Journal of Statistics, Series A, 2000: 49-66.

[18] Wang K, Li M, Hudley D, Liu R, Glessner J, Grant S, Hakonarson H, Bucan M. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data Genome Research 17:1665-1674, 2007

[19] Diskin, Sharon J., Mingyao Li, Cuiping Hou, Shuzhang Yang, Joseph Glessner, Hakon Hakonarson, Maja Bucan, John M. Maris, and Kai Wang. ”Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms.” Nucleic acids research 36, no. 19 (2008): e126-e126.

[20] Tony Cai, T., X. Jessie Jeng, and Hongzhe Li. "Robust detection and identification of sparse segments in ultrahigh dimensional data analysis.” Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74.5 (2012): 773-797.

[21] Roth, Lee, Swati Srivastava, Mosbit Lindzen, Aldema Sas-Chen, Michal Sheffer, Mattia Lauriola, Yehoshua Enuka et al. "SILAC identifies LAD1 as a filamin-binding regulator of actin dynamics in response to EGF and a marker of aggressive breast tumors.” Science signaling 11, no. 515 (2018).

[22] Haase, Daniela, Tiantian Cui, Linlin Yang, Yunxia Ma, Hongyu Liu, Bernhard Theis, Iver Petersen, and Yuan Chen. "Plakophilin 1 is methylated and has a tumor suppressive activity in human lung cancer.” Experimental and Molecular Pathology 108 (2019): 73-79.

[23] Kaz, A.M., Luo, Y., Dzieciatkowski, S., Chak, A., Willis, J.E., Upton, M.P., Leidner, R.S. and Grady, W.M., 2012. Aberrantly methylated PKP1 in the progression of Barrett’s esophagus to esophageal adenocarcinoma. Genes, Chromosomes and Cancer, 51(4), pp.384-393.

(HAN WANG) SCHOOL OF MATHEMATICAL SCIENCES AND CENTER FOR STATISTICAL SCIENCE, PEKING UNIVERSITY
Email address: wanghanmath@pku.edu.cn

(CHANGHU WANG) SCHOOL OF MATHEMATICAL SCIENCES, PEKING UNIVERSITY
Email address: wangch156@pku.edu.cn

(LINJIE WU) INSTITUTE OF HEMATOLOGY AND BLOOD DISEASE HOSPITAL, CHINESE ACADEMY OF MEDICAL SCIENCES AND PEKING UNION MEDICAL COLLEGE
Email address: wulinjie@ihcams.ac.cn

(RUIBIN XI) SCHOOL OF MATHEMATICAL SCIENCES, CENTER FOR STATISTICAL SCIENCE AND DEPARTMENT OF BIOSTATISTICS, PEKING UNIVERSITY
Email address: ruibinxi@math.pku.edu.cn