Estimation of N^6-*Methyladenosine* Specific Binding mRNA Fragment Proportion

估算N^6-甲基腺苷特异性结合的mRNA片段在混合样本中的比例

Jingxuan Bao
1509561
Supervisor: Dr. Jionglong Su
Co-supervisor: Dr. Jia Meng

Xi'an Jiaotong-Liverpool University
Department of Mathematical Sciences
Report of Final Year Project
April 17, 2019

Abstract

In this project, we are aiming to analyse the RNA methylation sites in a given set of genes in which the two main tasks are RNA methylation site detection and differential methylation analysis. To achieve these goals, the true proportion of m^6A specific binding mRNA in the m^6A-containing sample, *IP sample*, is needed.

To estimate the true proportion of the m^6A specific binding mRNA, we make an comparison between the problem of deconvolution of the tumour samples with respect to the normal samples and the problem of deconvolution of the *IP samples* in terms of the *input samples*. In this research, we examine two methods, ISOpureR and CLARKE respectively. After reviewing the necessary preliminary knowledge of biology and mathematics, we introduced the basic ideas and algorithms of these two method. Generally speaking, ISOpureR focuses on using statistical method whilst the CLARKE concentrates on the mathematical method. In particular, ISOpureR estimate the true proportion by maximising the complete likelihood function with the application of flexible preconditioned conjugate gradient method; CLARKE completes the task by the method of principal components analysis combined with the knowledge in the differential geometry and numerical analysis.

Furthermore, after introducing the rationale and algorithm of both methods, we apply simulated data and real data to assess their performance with the applications of the knowledge of regression analysis. After analysing, we conclude that with respect to the ISOpureR method, second order median regression model performs more efficiently; and with respect to the CLARKE method, median regression model is regarded to be more suitable than the linear regression model.

Regression analysis such as lack-of-fit test, diagnostic test, and F-test are performed to examine the suitability of the regression models. In conclusion, the ISOpureR method is more stable and more efficient when the data has a relatively small number of gene sites. For CLARKE method, although it is able to obtain stable results when applied to the data with a small number of gene sites, the efficiency of estimation is worse than the ISOpureR method; however, it has a better performance when the number of gene sites is relatively large.

keywords: ISOpureR, CLARKE, RNA sequencing, linear regression, quadratic polynomial regression, median regression, significance test, diagnostic test, lack-of-fit test.

摘要

在此项目中，我们专注于研究并分析在给定的一组基因中的RNA甲基化位点。该研究将分为两个模块进行，分别是RNA甲基化位点检测与差异甲基化分析。为了完成上述研究，我们的首要任务是估算N^6-甲基腺苷特异性结合的$mRNA$片段在混合样本中的真实比例。

为了估算该比例，我们利用类推法将估算癌症细胞在癌变组织中的比例的方法应用到估算N^6-甲基腺苷特异性结合的$mRNA$片段在混合样本中的真实比例的问题中。在该项研究中，我们主要探究了$ISOpureR$和$CLARKE$两种方法。$ISOpureR$的方法更加专注于统计学方法，通过最大化完全似然函数来估计真实比例；而$CLARKE$方法更加倾向于利用数学的模型，注重于利用微分几何和数值分析方面的知识进行估计。

此外，我们利用了回归分析的知识，评估两种方法在模拟数据与真实数据下运行的结果。经过分析，我们得到了$ISOpureR$方法在二阶中位数回归模型中所得出的结果更加稳定；而对于应用了$CLARKE$方法的数据，中位数回归模型比线性回归模型能够更加合适地描述模拟结果。

我们利用回归分析中的缺乏测试，诊断测试和F－测试来检验回归模型的适用性。$ISOpureR$方法在当数据样本具有相对较少基因位点信息时，表现得更加稳定且更加有效。对于$CLARKE$方法，当我们将其应用于具有少量基因位点的数据时，虽然它能够获得稳定的结果，但是该方法的估算效率比$ISOpureR$方法差；但是当数据样本的基因位点信息更多时，该方法表现出了更好的性能。

Contents

1 Introduction

RNA sequencing (RNA-Seq), also known as whole transcriptome shotgun sequencing (Morin et al., 2008), is a common method to detect the differences in gene expression in different samples, treatments, as well as different cell populations and experimental conditions (Maher et al., 2009). With the sequencing data, bio-statisticians may carry out more accurate predictions of the RNA methylation sites and differential RNA methylation analysis.

Usually, the prediction of RNA methylation sites is made possible using peak calling (Meng et al., 2017), a computational method applied to identify areas in a genome with aligned reads enrichment after performing a ChIP-sequencing or MeDIP-seq experiment. Moreover, with the improvement of high throughput sequencing data, differential RNA methylation analysis is now available with the help of RNA methylation experiments. Nowadays, N6-methyladenosine (m^6A), an abundant modification in mRNA that is found within some viruses (Beemon and Keith, 1977; Aloni et al., 1979) as well as most eukaryotes (Desrosiers et al., 1974; Perry et al., 1975), becomes increasingly important in various biological processes, such as RNA degradation, cocaine addiction, RNA-protein interaction (Meyer and Jaffrey, 2014).

1.1 Background Information and Previous Work

In this project, there are two main goals to be accomplished, namely, RNA methylation site detection and differential methylation analysis. To achieve these two goals, we need to estimate the real proportion of m^6A specific binding mRNA fragments of m^6A-containing sample (*IP sample*) from the untreated sample (*input sample*) (see Section 2.1.1 for details). However, the proportion of m^6A specific binding mRNA fragments is not known accurately due to the biological noise in the experiments (Wang et al., 2015).

In this paper, in order to be integrated with studies on RNA methylation site detection and differential methylation analysis, two methods of m^6A specific binding mRNA fragments proportion estimation in the *IP sample* are introduced and improved upon. These two methods are called ISOpureR and CLARKE, using totally different ideas in mathematics to achieve the same purpose in biology.

In the following part of this report, we mainly focus on the two methods. In section 2, literature review, we shall review these two methods and some biological and mathematical knowledge that we applied in our report. In section 3, methodology, methods are illustrated in detail. In section 4, results and discussion, we present the results by applying the data generated to test the stability and accuracy of these methods. We compare the two methods and applied both of them into our real data to work out the absolute value of proportion to further complete the task of differential methylation analysis.

1.2 Original Work

Overall, our work is novel in the following way:

- We apply the basic idea of computing the value of proportion by comparison. In this project, we compare estimating the proportion of m^6A specific binding mRNA in the *IP sample* given the condition of *input sample*, with the proportion of cancer cells in the tumour sample given the condition of normal sample. We are able to solve these two problems with the same method because both data share some common features. For example, both data are biological data, which implies we have to handle with the biological noise; both tasks are to estimate the proportion of one pure sample in the contaminated sample.
- we have improved the algorithm of derivative calculation in CLARKE method. When they calculating the first derivative, they applied two-point method to approximate it, which may lead to the results not accurate enough. In our project, we have applied the method of Five-Point Midpoint Formula to reduce the error of approximation. In this way, we have improved the accuracy of the approximation of the first derivative.
- The CLARKE method applied the definition to calculate the curvature, in which we need to reparametrise the curve into unit-speed curve, and then calculate the second derivative with the application of the definition equation. However, in our improved model, we simplify the procedure by using the proposition of the curvature, where rather than reparametrising the curve into unit-speed, we just need to find the derivatives of the parametrised curve and use the obtained derivatives to calculate the proposition equation. In this way, the method would become more efficient in terms of time.
- We compared the ISOpureR method and CLARKE method statistically in terms of all kinds of criteria. In our project, We apply significant test to check whether the regression model can describe the data well; we use the diagnostic test to examine whether the residuals, or errors, are follows the assumptions of the regression model; we study if the linear regression model adequately fits the data by application of lack-of-fit test. Moreover, we apply the higher order polynomial regression and check its availability when the lack-of-fit test fails.

2 Literature Review

This research project is related to both fields of bioinformatics and mathematics. In this section, some important aspects of background knowledge pertaining to biology and mathematics shall be discussed separately in the following sections.

In the next two subsections, biological and mathematical background, each marked subtitle written in bold characters is referred to the name of the related article or book, followed by the main knowledge applied in this project.

2.1 Background of Biology

In this subsection, we shall introduce the biological background of our project, including the overall procedure of RNA sequencing (RNA-Seq), the brief introduction raw data, the two biological methods of analysis, and biological significance of our project.

2.1.1 Introduction to Differential Gene Expression Analysis Using RNA-seq

RNA sequencing was firstly invented to detect the expression of the genomic loci in a cell at a certain period of time over the entire expression range. Later, due to the lifting the restriction of RNA-Seq, independent of expression transcripts counting of known genes, numerous additional application were raised. Among its applications, the detection of gene expression changes between cell populations and experimental conditions is widely used.

According to this article, there are a total of 3 stages to obtain the data we used:

- **RNA Extraction**: First, from a given sample of cells, we use viral, enzymic, or osmotic mechanisms to break down the membrane of a cell to obtain lysed cells for RNA extraction. Next, particular methods such as silica-gel based membrane or the most prevalent methods, liquid-liquid extractions with acidic phenol-chloroform, are used to extract RNA from the lysed cells. After this step, the objects obtained consists mainly of four types of RNA, i.e., the rRNA, tRNA, snRNA, and mRNA. Among all these types of RNA, mRNA will be selected and used to be analysed. (See Figure 1)
- **Library Preparation**: Library refers to, in biological science, a (perfectly random) collection of DNA fragments that are ready for sequencing with a specific protocol. After extracting mRNA, step of mRNA fragmentation is necessary for sequencing as some of mRNA is too long to be sequenced. An average length of 100 base pairs to 300 base pairs (100bp to 300bp) mRNA fragments will be obtained after fragmentation and they will be converted to cDNA to construct a cDNA library waiting for sequencing. (See Figure 1)
- **Alignment**: With the completion of the Human Genome Project, the identification of sequence of human DNA nucleotide base pairs has been completed in 2003, which allows us to align our mRNA sequencing data to the genome to provide us the information about the positions of mRNA fragments in the whole human genome. (See Figure 2)

After these steps, the number of times of a certain gene that is aligned with mRNA fragments would be available, named reads count, which is the data used in this project.

2.1.2 ISOpureR: Computational Purification of Individual Tumour Gene Expression Profiles Leads to Significant Improvements in Prognostic Prediction

Quon et al. (2013) has described a computational purification tool named ISOpureR. Given the tumour cell sample and normal (healthy) cell sample, ISOpureR is able to estimate the proportion of RNA originating from cancer cells, and generate a purified cancer profile for each tumour sample.

ISOpureR employs two main regularisation strategies:

- The first strategy is that, the model assumes that each normal (healthy) profile can be represented by a weighted combination of available healthy tissue profiles. With this assumption, ISOpureR can easily handle the case when tumour data and

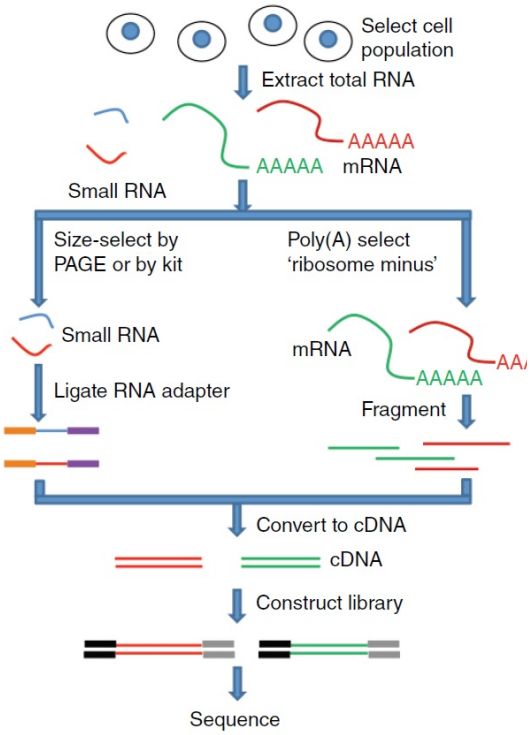

Figure 1: Library Preparation (Dündar et al., 2015)

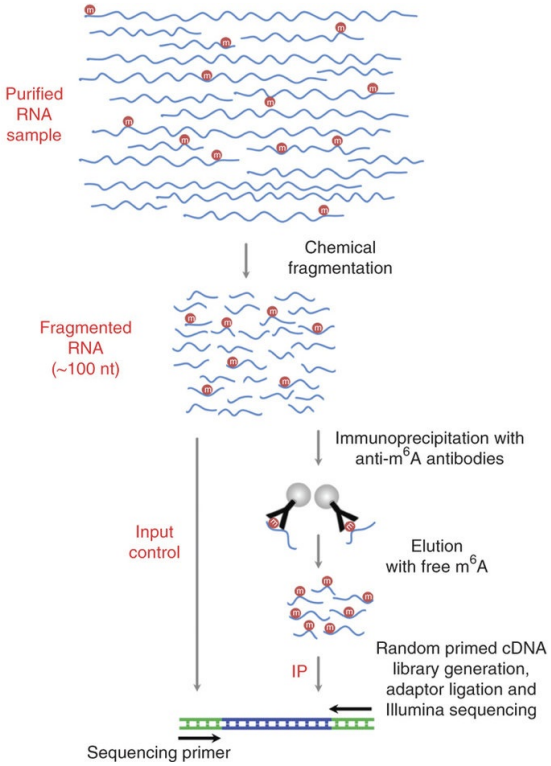

Figure 2: Schema of m^6A-Seq Protocol (Dominissini et al., 2013)

normal data are not matched. Moreover, this model also assumes that both tumour samples and normal samples satisfy Dirichlet distribution.

- The second strategy is that, the model use a two-step approach to maximise the complete likelihood function. The Polack-Ribiere flavour of conjugate gradients Rasmussen and Williams (2006) is used to search directions; and a line search using quadratic and cubic polynomial approximations and the Wolfe-Powell stopping criteria (Sun and Yuan, 2006) is used together with the slope ratio method for guessing initial step sizes.

After these two steps, the actual proportion of cancer cells in the tumour samples can be estimated. Followed by this way, we have done an analogy and improvement to our data so that we were able to predict the actual proportion of m^6A specific binding mRNA in the *IP sample*.

2.1.3 CLARKE: Statistical Expression Deconvolution from Mixed Tissue Samples

In this method, a totally different method of expression deconvolution from mixed tissue samples is discussed, where the proportion of some component cell type remains unknown.

The idea of CLARKE method is described as following: it finds the minimum ratio of mixed sample gene expression and one of the pure sample expression; and calculate the proportion of the pure sample by letting the gene expression of the other pure sample approaching to 0 (Gosink et al., 2008). However, following this idea, the method is likely

to be an underestimate of the true proportion value for both simulated noisy data and observed data. Clarke et al. (2010) have improved this method by transforming the data so that the mean and median of the samples approaching closer together. By plotting the curve of mean and median, applying the method of principal components analysis (Jolliffe, 2002), the author concluded the 'knee' or 'elbow' of the curve give us the most accurate estimation.

2.2 Background of Mathematics

In this subsection, the main knowledge applied to derive the two methods is reviewed. The first two parts of this subsection introduce the knowledge used to derive the method of ISOpureR, and the third part of the subsection is the one for CLARKE.

2.2.1 Statistical Inference

In this book, Casella and Berger (1990) introduce the main topics of statistical science, basics of probability theory, transition between probability and statistics, three statistical principles (sufficiency, likelihood, and invariance), estimation and hypothesis testing. However, in our project, a very small part of these topics are covered such as sampling, Bayes estimators, hypothesis testing.

Here follows the mean concepts and theorems that are used in our project:

- **The Likelihood Function**

  Let $f(\boldsymbol{x}|\theta)$ denote the joint probability density function (pdf) or probability mass function (pmf) of the sample $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$. Then, given that $\boldsymbol{X} = \boldsymbol{x}$ is observed, the function of θ defined by

  $$L(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta) \tag{2.2.1}$$

  is called the likelihood function.

  According to the definition of the likelihood function, it seems that we define the likelihood function to be almost the same as pdf or pmf. The only distinction between these two functions is the difference of the fixed value and the varying value. In the pdf or pmf, the fixed value is parameter θ and the varying value is $\boldsymbol{x}$, whilst in the likelihood function, we consider $\boldsymbol{x}$ as the fixed value and θ is the variable. In other words, when we consider the likelihood function $L(\theta|\boldsymbol{x})$, we regard $\boldsymbol{x}$ to be the observed sample point and θ to be varying over all possible parameter values.

  Moreover, in our project, we use the discrete random vector to describe our data. In this case, if we compare two likelihood function $L(\theta_1|\boldsymbol{x})$ and $L(\theta_2|\boldsymbol{x})$ with the relation

  $$L(\theta_1|\boldsymbol{x}) > L(\theta_2|\boldsymbol{x}), \tag{2.2.2}$$

  then the sample we actually observed is more likely to have occurred when $\theta = \theta_1$ than when $\theta = \theta_2$. This is the main information provided by the likelihood function. We use this knowledge to estimate the parameters in the ISOpureR method (see Equation 3.1.10).

- **Maximum Likelihood Estimators**

  For each sample point $\boldsymbol{x}$, let $\hat{\theta}(\boldsymbol{x})$ be a parameter value at which $L(\theta|\boldsymbol{x})$ attains its maximum as a function of θ, with $\boldsymbol{x}$ held fixed. A maximum likelihood estimator (MLE) of the parameter θ based on a sample $\boldsymbol{X}$ is $\hat{\theta}(\boldsymbol{X})$.

  The MLE is a reasonable choice for an estimator because the MLE is the parameter point which would obtain the most likely observed sample. In such a case, all we need to do is to find the stable (stable with respect to the small change in the data) global maximum of the MLE; and hence the most suitable estimator could be chosen.

2.2.2 The Flexible Preconditioned Conjugate Gradient Method

The conjugate gradient method, in mathematics, is an algorithm for the numerical solution of particular systems with symmetric and positive-definite matrix. The conjugate gradient method is often implemented as an iterative algorithm, applicable to large sparse systems.

In the following part of the illustration, all bold symbols refer to a vector or a matrix. Given the system of equations

$$\boldsymbol{Ax} = \boldsymbol{b} \tag{2.2.3}$$

for the vector $\boldsymbol{x}$ and $\boldsymbol{b}$, and $n \times n$ symmetric, positive-definite, real matrix $\boldsymbol{A}$.

We initially guess $\boldsymbol{x_0} = 0$, considering the system $\boldsymbol{Az} = \boldsymbol{b} - \boldsymbol{Ax_0}$. Assume the solution of the system is $\boldsymbol{x_*}$. From $\boldsymbol{x_0}$, we begin search for the solution and in each iteration, we need a metric to tell us whether our approximation are closer to the solution $\boldsymbol{x_*}$, which is the unique minimiser of the function

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x^T Ax} - \boldsymbol{x^T b} \tag{2.2.4}$$

Next, from the uniqueness of the minimiser, we have that its second derivative is a symmetric positive-definite matrix

$$\nabla^2 f(\boldsymbol{x}) = A, \tag{2.2.5}$$

Then the minimiser solves the initial problems is from its first derivative

$$\nabla f(\boldsymbol{x}) = \boldsymbol{Ax} - \boldsymbol{b}. \tag{2.2.6}$$

From the previous equation, we have the first bases vector $\boldsymbol{p_0}$ to be the negative of the gradient of f, which is $\boldsymbol{Ax} - \boldsymbol{b}$. We take $\boldsymbol{p_0} = \boldsymbol{b} - \boldsymbol{Ax_0}$.

Next, we denote the $\boldsymbol{r_k}$ to be the residual at the kth step,

$$\boldsymbol{r_k} - \boldsymbol{b} - \boldsymbol{Ax_k}. \tag{2.2.7}$$

We require that the next direction search step to be built upon the current residual and all previous search directions, which gives us the expression

$$\boldsymbol{p_k} = \boldsymbol{r_k} - \sum_{i<k} \frac{\boldsymbol{p_i^T A r_k}}{\boldsymbol{p_i^T A p_i}} \boldsymbol{p_i} \tag{2.2.8}$$

From this direction, the next optimal location is

$$\boldsymbol{x_{k+1}} = \boldsymbol{x_k} + \alpha_k \boldsymbol{p_k} \tag{2.2.9}$$

where

$$\alpha_k = \frac{\boldsymbol{p_k^T}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x_k})}{\boldsymbol{p_k^T}\boldsymbol{A}\boldsymbol{p_k}} = \frac{\boldsymbol{p_k^T}\boldsymbol{r_k}}{\boldsymbol{p_k^T}\boldsymbol{A}\boldsymbol{p_k}} \tag{2.2.10}$$

where the last equality follows from the definition of $\boldsymbol{r_k}$. The expression for α_k is derived if one substitutes the expression for x_{k+1} into f and minimising it with respect to α_k,

$$f(\boldsymbol{x_{k+1}}) = f(\boldsymbol{x_k} + \alpha_k \boldsymbol{p_k}) = g(\alpha_k) \tag{2.2.11}$$

and

$$g'(\alpha_k) = 0 \Rightarrow \alpha_k = \frac{\boldsymbol{p_k^T}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x_k})}{\boldsymbol{p_k^T}\boldsymbol{A}\boldsymbol{p_k}} \tag{2.2.12}$$

As Notay (2000) stated, the flexible preconditioned conjugate gradient method converges faster than the ordinary conjugate gradient method. Moreover, Polak-Ribière formula is applied here to improve the convergence dramatically. In our project, we apply the method to minimise the likelihood function (see Section 3.1, CPE and TPE).

2.2.3 Bayesian Data Analysis

This book, as Gelman et al. (2003) stated, is an introductory text on Bayesian inference, a graduate text on effective current approaches to Bayesian modelling, and a handbook of Bayesian methods in applied statistics. A wide range of the knowledge related to Bayesian statistics is introduced but in this project, we mainly focus on the introduction of the multinomial distribution and the Dirichlet distribution.

- **Multinomial Distribution**

  The multinomial distribution is a generalisation of binomial distribution, which allows more than two possible outcomes. The multinomial sampling distribution is applied to describe data where every observation is one of K possible outcomes.

  The probability mass function of the multinomial distribution is

$$Multinomial(\boldsymbol{\gamma} \mid \boldsymbol{\pi}) = \frac{(\sum_{k=1}^{K} \gamma_k)!}{\prod_{k=1}^{K} \gamma_k!} \prod_{k=1}^{K} \pi_k^{\gamma_k} \tag{2.2.13}$$

  where the 'bold' letter represents a vector or matrix, and the 'unbold' letter implies a scalar; $\boldsymbol{\pi}$ refers the parameter vector and $\boldsymbol{\gamma}$ can be seen as the state vector of the random variable. We use Multinomial distribution to describe the tumour profiles in the ISOpureR method (see Section 3.1, $\boldsymbol{x_n}$)

- **Dirichlet Distribution**

  The Dirichlet distribution is the conjugate prior distribution of the parameters of a multivariate generalisation of the beta distribution.

The probability mass function of the Dirichlet distribution is

$$Dirichlet(\boldsymbol{x} \mid \boldsymbol{a}) = \frac{\Gamma(\sum_{k=1}^{K} a_k)}{\prod_{k=1}^{K} \Gamma(a_k)} \prod_{k=1}^{K} x_k^{a_k - 1} \tag{2.2.14}$$

where the 'bold' letter represents a vector or matrix, and the 'unbold' letter implies a scalar; $\boldsymbol{a}$ refers the parameter vector and $\boldsymbol{x}$ can be seen as the state vector of the random variable. We use the Dirichlet distribution to describe the parameters and the cancer profiles in the ISOpureR method (see Section 3.1, $\boldsymbol{\theta_n}, \boldsymbol{c_n}, \boldsymbol{m}$).

2.2.4 Elementary Differential Geometry

In this book, Pressley (2010) introduced the curves and surfaces and applied the knowledge of algebra to solve questions in geometry.

We have used basically the knowledge about the curvature of a curve, which would be illustrated in below:

- **Reparametrization**

  A parametrized curve $\tilde{\gamma} : (\tilde{\alpha}, \tilde{\beta}) \to \mathbb{R}^n$ is a reparametrization of a parametrized curve $\gamma : (\alpha, \beta) \to \mathbb{R}^n$ if there is a smooth bijective map $\phi : (\tilde{\alpha}, \tilde{\beta}) \to (\alpha, \beta)$ (the reparametrization map) such that the inverse map $\phi^{-1} : (\alpha, \beta) \to (\tilde{\alpha}, \tilde{\beta})$ is also smooth and

  $$\tilde{\gamma}(\tilde{t}) = \gamma(\phi(\tilde{t})) \quad for\ all\ \ \tilde{t} \in (\tilde{\alpha}, \tilde{\beta}) \tag{2.2.15}$$

  Reparametrization of a curve is using different equations to represent the same curve. The reason why we do reparametrization is that sometimes we need to find a proper way to express a curve so that we are able to simplify the calculation related the curve. In our project, when studying CLARKE method, we need to reparametrize the curve as unit-speed curve (the Euclidean norm of the first derivative of the curve is equal to 1), so that we can find the curvature of the curve by applying the definition (see Section 3.2), which will be discussed in the next part.

- **Curvature**

  If γ is a unit-speed curve with parameter t, its curvature $\kappa(t)$ at the point $\gamma(t)$ is defined to be $|| \ddot{\gamma}(t) ||$.

  The geometric meaning of curvature is to measure how curved of the curve. If a curve with a larger curvature than the other one at some point, then we say that the curvature is more curved than the other curve at this point. Moreover, the 'elbow' or 'knee' of a curve at some point is the reciprocal of the curvature of the curve at this point. We calculate the curvature in the original CLARKE method by this definition (see Section 3.2, Model Formulation).

  Let $\gamma(t)$ be a (not necessarily unit-speed) parametrised curve in $\mathbb{R}^3$. Then its curvature is

  $$\kappa(t) = \frac{||\gamma' \times \gamma''||}{||\gamma'||^3}. \tag{2.2.16}$$

Without doing unit-speed reparametrisation, this proposition may simplify the procedure of calculating the curvature of a curve. We improve the calculation of the curvature in the CLARKE method by this proposition (see Section 3.2, Model Improvement).

2.2.5 Understanding Analysis

We mainly focus on the Inverse Function Theorem in the book 'Understanding Analysis' and latter, we are going to apply this theorem to prove the existence of the inverse function of arc length (see Equation 3.2.9).

For functions of a single variable, the theorem states that if f is a continuously differentiable function with nonzero derivative at the point a, then f is invertible in a neighbourhood of a, the inverse is continuously differentiable, and the derivative of the inverse function at $b = f(a)$ is the reciprocal of the derivative of f at a:

$$(f^{-1})'(b) = \frac{1}{f'(a)}. \tag{2.2.17}$$

We need to use the theorem to prove the existence and the uniqueness of the inverse function of the arc-length reparametrisation of the curve to convert the parametrised curve into unit-speed curve (see Equation 3.2.9).

2.2.6 Numerical Analysis

In the book 'Numerical Analysis', we review the knowledge about numerical differentiation. Moreover, the Five-Point Midpoint Formula will be examined and applied in our CLARKE method (see Section 3.2, Model Improvement).

For a certain partition, given that h is the fixed difference between two consecutive points. The first derivative of the function $f(x)$ at the point x_0 can be described by

$$f(x_0)' = \frac{f(x_0 - 2h) - 8\,f(x_0 - h) + 8\,f(x_0 + h) - f(x_0 + 2h)}{12h} + \frac{h^4}{30} f^{(5)}(\xi) \tag{2.2.18}$$

where ξ lies between $x_0 - 2h$ and $x_0 + 2h$ (Burden et al., 2016).

2.2.7 Quantile Regression

Quantile functions of a sample is defined by

$$Q(\tau|x) = \beta_0 + \beta_1 x + F_u^{-1}(\tau) \tag{2.2.19}$$

where F_u denotes the common distribution function of the errors; and $F(x) = P(X \leq x)$, while for any $0 < \tau < 1$, we have

$$F^{-1}(\tau) = inf\{x : F(x) \geq \tau\}, \tag{2.2.20}$$

called the τth quantile of X (Koenker, 2005).

Here, additionally, according to Koenker (2005), quantile regression does not make distributional assumptions. In other words, the assumptions about residuals, other than assuming that the response variable is almost continuous. We apply median regression, the quantile regression with $\tau = 0.5$, to both ISOpureR method and CLARKE method (see Section 4.1.2, Median Regression for ISOpureR Method and Analysis, Second Order Median Regression Model for ISOpureR Method, and Median Regression for CLARKE Method and Analysis).

3 Methodology

In this section, both of the statistical model and the mathematical model, ISOpureR and CLARKE, will be discussed and applied to calculate the true proportion of mRNA that is extracted specifically by m^6A.

3.1 ISOpureR

In this part, the full ISOpureR model is defined as follows (note that all the **bold** symbol represents a vector or a matrix):

Assumptions:

- Vector of normal profile $\boldsymbol{h_n}$ is similar to one or more profiles of normal tissue that are input into the algorithm.
- $\boldsymbol{h_n}$ is a convex combination of the normal profiles provided to the algorithm.
- The cancer profiles $\boldsymbol{c_1}$, $\boldsymbol{c_2}$, ..., $\boldsymbol{c_N}$ in the cohort are clustered together around a 'reference cancer profile' $\boldsymbol{m}$.
- Profiles of cancerous tissue are similar (but not identical) to those of the tissue of origin of the tumour type.

Parameters Definition:

From the definition (see Table 1), by Quon et al. (2013) we originally have relation between tumour samples and normal samples shown in the following equations:

$$\boldsymbol{t_n} = \alpha_n \boldsymbol{c_n} + (1 - \alpha_n)\boldsymbol{h_n} + \boldsymbol{e_n} \tag{3.1.1}$$

$$= \alpha_n \boldsymbol{c_n} + \sum_{r=1}^{R} \theta_{n,r} \boldsymbol{b_r} + \boldsymbol{e_n} \tag{3.1.2}$$

where $\boldsymbol{e_n}$ represents the error, α_n represents the proportion of cancer cells in the tumour cells. Here, we use the assumption that $\boldsymbol{h_n}$ is a convex combination of the normal profiles provided to the algorithm.

Data Transformation:

The reason why ISOpureR method needs the data to be transformed is that, in the tumour profiles, the sum of the elements in each of the discretized profiles after robust multi-array

Name	Definition	Remark
$\boldsymbol{t_n}$	Tumour profiles	Input data; Vectors with G elements and $n = 1, 2, \cdots, N$.
$\boldsymbol{b_n}$	Healthy profiles	Input data; Required that $R < N$.
$\boldsymbol{h_n}$	Vector of normal profile	Each normal profile can be represented by a weighted combination of the available healthy tissue profiles $\boldsymbol{b_1}, \boldsymbol{b_2}, \cdots, \boldsymbol{b_R}$.
$\boldsymbol{c_n}$	The cancer profiles	Profiles $\boldsymbol{c_1}, \boldsymbol{c_2}, \cdots, \boldsymbol{c_n}$ in the cohort are clustered together around a 'reference cancer profile' $\boldsymbol{m}$.
$\boldsymbol{m}$	Reference cancer profile	Estimated from the tumour profile data; Has a regularisation applied to it to bias its estimate toward values that are close to the normal profiles.
k_n	The strength parameter	The strength parameter of the Dirichlet distribution over $\boldsymbol{c_n}$ given $\boldsymbol{m}$ $(n = 1, 2, 3, ..., N)$.
k'	The strength parameter	The strength parameter of the Dirichlet distribution over $\boldsymbol{m}$.
$\boldsymbol{\omega}$	Weights	The weights on the normal profiles $\boldsymbol{b_r}$ used to make the weighted combination that forms the mean parameter vector for the Dirichlet distribution over $\boldsymbol{m}$.
$\boldsymbol{v}$	Mean and strength	Represents both the mean and strength of a Dirichlet distribution over $\boldsymbol{\theta_n}$.
$\boldsymbol{x_n}$	Discretized tumour profiles	A count vector derived from discretization of $\boldsymbol{t_n}$.
$\hat{\boldsymbol{x}}$	Normalised reconstruction of the tumour profile	The probability of the discretized tumour profiles under the multinomial distribution, a normalised reconstruction of the tumour profile $\boldsymbol{x_n}$ based on the model parameters.
$\boldsymbol{e_n}$	Error	

Table 1: Notation

average (RMA) normalisation should be on the order of 10^7. To ensure adequate precision in the discretization, if their sum is much less than 10^7, profiles may need to be rescaled.

The discretisation of the tumour profiles $\boldsymbol{t_n}$ is to round each element of $\boldsymbol{t_n}$ to the nearest non-negative integer to obtain our transformed tumour profiles $\hat{\boldsymbol{x}}_{\boldsymbol{n}}$. Discretisation allow us to rescale the tumour profiles so that the total number of observations (the sum of the elements) after discretization is approximately the same across all tumour profiles, and to balance the influence that each tumour profile has on the shared parameters.

In order to allow $\boldsymbol{b_r}$ to be interpreted as a discrete probability distribution over transcripts, we divide each normal profile $\boldsymbol{b_r}$ by the sum of its elements.

After discretization, Equation 3.1.2 becomes

$$\hat{\boldsymbol{x}}_{\boldsymbol{n}} = \alpha_n \boldsymbol{c_n} + \sum_{r=1}^{R} \theta_{n,r} \boldsymbol{b_r} \tag{3.1.3}$$

Till now, all we have to do is to figure out the scalar, or say, the true proportion of cancer cells in the tumour sample, which is α_n.

Model Formulation:

We now formulate our model to solve the absolute proportion of the cancer cells in the tumour sample. Generally speaking, in this model, we use the method of maximising the complete likelihood function to find the most proper estimator, and in each step, we use the flexible preconditioned conjugate gradient method with iteration to find the maximum likelihood function. The details of the method are displayed in the following part of this section (note that all the **bold** symbol represents a vector or a matrix):

First of all, we define our symbol as follows (see Table 2):

Symbol	Definition
$\boldsymbol{B}$	$= [\boldsymbol{b_1}\ \boldsymbol{b_2}\ ...\ \boldsymbol{b_R}]$
$\boldsymbol{\theta_n}$	$= [\theta_{n,1}\ \theta_{n,2}\ ...\ \theta_{n,R}\ \alpha_n]$
$\hat{\boldsymbol{x_n}}$	$= [\boldsymbol{B}\ \boldsymbol{c_n}]\ \boldsymbol{\theta_n} = [\boldsymbol{b_1}\ \boldsymbol{b_2}\ ...\ \boldsymbol{c_n}]\ \boldsymbol{\theta_n}$ $= \theta_{n,1}\boldsymbol{b_1} + \theta_{n,2}\boldsymbol{b_2} + ... + \theta_{n,R}\boldsymbol{b_R} + \alpha_n\boldsymbol{c_n}$
$p\left(\boldsymbol{\theta_n} \mid \boldsymbol{v}\right)$	$= Dirichlet\left(\boldsymbol{\theta_n} \mid \boldsymbol{v}\right)$
$p\left(\boldsymbol{x_n} \mid \boldsymbol{B}, \boldsymbol{\theta_n}, \boldsymbol{c_n}\right)$	$= Multinomial\left(\boldsymbol{x_n} \mid \hat{\boldsymbol{x_n}}\right)$
$p\left(\boldsymbol{c_n} \mid k_n, \boldsymbol{m}\right)$	$= Dirichlet\left(\boldsymbol{c_n} \mid k_n\boldsymbol{m}\right)$
$p\left(\boldsymbol{m} \mid k', \boldsymbol{B}, \boldsymbol{\omega}\right)$	$= Dirichlet\left(\boldsymbol{m} \mid k'\boldsymbol{B\omega}\right)$

Table 2: Notation

Estimate all the parameters $\boldsymbol{\theta_n}$, α_n, $\boldsymbol{c_n}$, $\boldsymbol{v}$, $\boldsymbol{m}$, k_w, k', and $\boldsymbol{\omega}$, applying two-step approach to maximise the complete likelihood function with the flexible preconditioned conjugate gradient method:

$$\mathbb{L} = p(\boldsymbol{m} \mid k', \boldsymbol{B}, \boldsymbol{\omega}) \prod_{n=1}^{N} p(\boldsymbol{c_n} \mid k_n, \boldsymbol{m}) p(\boldsymbol{\theta_n} \mid \boldsymbol{v}) p(\boldsymbol{x_n} \mid \boldsymbol{B}, \boldsymbol{\theta_n}, \boldsymbol{c_n}) \tag{3.1.4}$$

Now, step-by-step illustration of ISOpureR method is displayed in the following part of this section. From Step 1 to Step 4, the purpose of algorithm is to initialise the parameters and we denote these steps as initialisation step; from Step 5 to Step 10, the algorithm is aimed at eliciting the actual cancer profiles, and these steps are denoted as cancer profile estimation step; from Step 11 to Step 15, we are trying to estimat the proportion of tumour profiles, and we denote these steps as tumour profile estimation step.

- **Initialisation Step**:

  Step 1: Denote D as *the number of tumour samples*; K as *the number of cancer profiles*, which is equal to *the number of normal profiles + 1* (+1 is for the reference cancer profile $\boldsymbol{m}$). Let $\boldsymbol{c_n}$ is equal to $\boldsymbol{m}$ (that is, we set $k_n = k' = \infty$ for all n).

  Step 2: Let $\boldsymbol{v}$ be a $K \times 1$ vector, which is generated randomly with the last element having the most weighted proportion.

  Step 3: Initialise the $D \times K$ matrix $\boldsymbol{\theta_n}$ and distribute a higher weight to the cancer component.

Step 4: Initialise $(K-1)\times 1$ vector $\boldsymbol{\omega}$.

- **Cancer Profile Estimation Step (CPE)**:

  In the following steps, the method to find the minimum of log-likelihood function (see Section 2.2.1 for details) is the flexible preconditioned conjugate gradient method (see Section 2.2.2 for details).

  Step 5: To optimise $\boldsymbol{m}$, we first generating a $G\times K$ parameter matrix with the first $K-1$ columns being the logarithm of the normal profiles $log(\boldsymbol{b_n})$ and the last column to be $log(\omega_1\boldsymbol{b_1}+\omega_2\boldsymbol{b_2}+...+\omega_{K-1}\boldsymbol{b_{K-1}})$. Then we initialise the $G\times 1$ vector $\boldsymbol{m}$ such that $\boldsymbol{m} = log(\omega_1\boldsymbol{b_1}+\omega_2\boldsymbol{b_2}+...+\omega_{K-1}\boldsymbol{b_{K-1}})$. Finally, we minimise the log-likelihood function

$$-log\, p(\boldsymbol{m}\mid k', \boldsymbol{B}, \boldsymbol{\omega}) - \sum_{n=1}^{N} log\, p(\boldsymbol{x_n}\mid \boldsymbol{B}, \boldsymbol{\theta_n}, \boldsymbol{m}) \tag{3.1.5}$$

  with respect to $\boldsymbol{m}$ to find the most proper parameter $\boldsymbol{m}$. Where $\boldsymbol{x_n}$ is the n^{th} tumour profile with $1\le n\le D$ and $K-1\le D$.

  Step 6: optimise $\boldsymbol{\theta_n}$ for all n, we minimise the log-likelihood function

$$-log\, p(\boldsymbol{\theta_n}\mid \boldsymbol{v}) - log\, p(\boldsymbol{x_n}\mid \boldsymbol{B}, \boldsymbol{\theta_n}, \boldsymbol{m}) \tag{3.1.6}$$

  with respect to $\boldsymbol{\theta_n}$.

  Step 7: To optimise $\boldsymbol{v}$, we minimise the log-likelihood function

$$-\sum_{n=1}^{N} log\, p(\boldsymbol{\theta_n}\mid \boldsymbol{v}) \tag{3.1.7}$$

  with respect to $\boldsymbol{v}$.

  Step 8:To optimise k', we minimise the function

$$-log\, p(\boldsymbol{m}\mid k', \boldsymbol{B}, \boldsymbol{\omega}) \tag{3.1.8}$$

  with respect to k'.

  Step 9: To optimise $\boldsymbol{\omega}$, we minimise the function

$$-log\, p(\boldsymbol{m}\mid k', \boldsymbol{B}, \boldsymbol{\omega}) \tag{3.1.9}$$

  with respect to $\boldsymbol{\omega}$.

  Step 10: We run Step 5 to Step 9 at least 35 iterations, and check if the change in log likelihood,

$$\mathbb{L} = p\left(\boldsymbol{m}\mid k', \boldsymbol{B}, \boldsymbol{\omega}\right)\prod_{n=1}^{N} p(\boldsymbol{\theta_n}\mid \boldsymbol{v}) p(\boldsymbol{x_n}\mid \boldsymbol{B}, \boldsymbol{\theta_n}, \boldsymbol{m}) \tag{3.1.10}$$

  is smaller than a threshold (the change in log-likelihood function is 10^{-7} or iterations is up to 100 times).

- **Tumour Profile Estimation Step (TPE)**:

  In the following steps, the method to find the minimum of log-likelihood function is still the flexible preconditioned conjugate gradient method (please refer Section 2.2.2 for details).

  Step 11: To optimise $\boldsymbol{c_n}$ for all $1 \leq n \leq N$, we minimise the function

  $$-log\, p(\boldsymbol{c_n} \mid k_n, \boldsymbol{m}) - log\, p(\boldsymbol{x_n} \mid \boldsymbol{B}, \boldsymbol{\theta_n}, \boldsymbol{c_n}) \tag{3.1.11}$$

  with respect to $\boldsymbol{c_n}$.

  Step 12: To optimise $\boldsymbol{\theta_n}$ for all $1 \leq n \leq N$, we minimise the function

  $$-log\, p(\boldsymbol{\theta_n} \mid \boldsymbol{v}) - log\, p(\boldsymbol{x_n} \mid \boldsymbol{B}, \boldsymbol{\theta_n}, \boldsymbol{c_n}) \tag{3.1.12}$$

  with respect to $\boldsymbol{\theta_n}$.

  Step 13: To optimise $\boldsymbol{v}$, we minimise the function

  $$-\sum_{n=1}^{N} log\, p(\boldsymbol{\theta_n} \mid \boldsymbol{v}) \tag{3.1.13}$$

  with respect to $\boldsymbol{v}$.

  Step 14: To optimise $\boldsymbol{k}$, we minimise the function

  $$-\sum_{n=1}^{N} log\, p(\boldsymbol{c_n} \mid k_n, \boldsymbol{m}) \tag{3.1.14}$$

  with respect to $\boldsymbol{k}$.

  Step 15: We run Step 11 to Step 14 at least 35 iterations, and check if the change in complete log likelihood,

  $$\mathbb{L} = \prod_{n=1}^{N} p(\boldsymbol{c_n} \mid k_n, \boldsymbol{m}) p(\boldsymbol{\theta_n} \mid \boldsymbol{v}) p(\boldsymbol{x_n} \mid \boldsymbol{B}, \boldsymbol{\theta_n}, \boldsymbol{c_n}) \tag{3.1.15}$$

  is smaller than a threshold (the change in log-likelihood is 10^{-7} or iterations is up to 100 times).

After all these steps, we are able to estimate the true proportion of the cancer cells in the tumour samples as well as the actual value of each component of the cancer profiles. Moreover, in our project, we just need to replace the tumour samples by *IP samples* and replace the normal samples (healthy samples) by *input sample* to achieve our goal, estimation of true proportion of m^6A binding mRNA in the contaminated mRNA samples.

3.2 CLARKE

In this method, we apply data transformation and plot the graph of mean and median of the ratio of gene expression with respect to the transformed data. After plotting the graph, we use the method of principal components analysis (Jolliffe, 2002) to conclude

the 'elbow', or say, the 'knee' of the curve implies the best estimation of the true value of mixing proportion.

Parameters Definition:

Note that all the **bold** symbols refer to a vector or a matrix. All the parameters we used in this model are shown in the Table 3.

Name	Definition	Remark
A	Pure profile A	
B	Pure profile B	
AB	Mixed profile	
$\boldsymbol{E}(A)$	Gene expression of pure sample A	Input data (if available); $\boldsymbol{E}(A) = (E_1(A), E_2(A), \cdots, E_m(A))$.
$\boldsymbol{E}(B)$	Gene expression of pure sample B	Input data (if available); $\boldsymbol{E}(B) = (E_1(B), E_2(B), \cdots, E_m(B))$.
$\boldsymbol{E}(AB)$	Gene expression of mixed sample	Input data (if available); $\boldsymbol{E}(AB) = (E_1(AB), E_2(AB), \cdots, E_m(AB))$.
$tE_i(A)$	Transformed value of the i^{th} gene expression	$tE_i(A) = log(1 + \alpha E_i(A))$
$tE_i(AB)$	Transformed value of the i^{th} gene expression	$tE_i(AB) = log(1 + \alpha E_i(AB))$
α	Intermediate parameter	Each value of α provides an accurate value of p_A under the situation of each dataset.
$\boldsymbol{R}$	Ratio of gene expression	$\boldsymbol{R} = (R_1, R_2, \cdots, R_m)$; $R_i = \frac{E_i(AB)}{E_i(A)}$
tR_i	The transformed value of gene expression ratio	tR_i is a function of α; $i = 1, 2, \cdots, m$.
$\overline{tR_i}$	The mean value of tR_i	$i = 1, 2, \cdots, m$.
p_A	The true proportion of profile A in profile AB	$E_i(AB) = p_A E_i(A) + (1 - p_A)E_i(B) + \epsilon$; $i = 1, 2, \cdots, m$.
ϵ	Error	

Table 3: Notation

Original Model:

This model has been raised by Gosink et al. (2007), that given the mixed sample AB, one of the pure sample A, we want to estimate the proportion of sample A in the mixed sample AB, which is p_A in the following equation

$$E_i(AB) = p_A E_i(A) + (1 - p_A)E_i(B) + \epsilon. \tag{3.2.1}$$

Let $R_i = \frac{E_i(AB)}{E_i(A)}$. In the noiseless case, we have

$$R_i = p_A \frac{E_i(A)}{E_i(A)} + (1 - p_A)\frac{E_i(B)}{E_i(A)}. \tag{3.2.2}$$

Since the expression value is assumed to be non-negative, we have

$$\lim_{E_i(B)\to 0} R_i = p_A + (1 - p_A)\frac{E_i(B)}{E_i(A)} = p_A. \tag{3.2.3}$$

Therefore, under the assumption that $E_i(B) \to 0$, we have $min_i R_i = p_A$. However, the minimum ratio is likely to be underestimated compared with the true proportion. The problem comes from the noise of the data; and by increasing the small ratio values while shrinking larger ratio values, we may improve the performance of the estimation (Gosink et al., 2007).

Data Transformation:

According to Clarke et al. (2010), we first consider transforming both $\boldsymbol{E}(AB)$ and $\boldsymbol{E}(A)$ into the form

$$tE_i(AB) = log(1 + \alpha E_i(AB)) \tag{3.2.4}$$

$$tE_i(A) = log(1 + \alpha E_i(A)) \tag{3.2.5}$$

for some $\alpha > 0$ for all i. The reason why we transform the data in such a way is that, by Clarke et al. (2010), the underestimation of the minimum ratio compared to the true proportion value is caused by the noise in the observed expression data from mixed samples. After transformation, we are able to reduce the difference between the mean and the median, so that the small ratio value could increase while the larger ratio value could decrease.

Model Formulation:

Given that the minimum value of tR_i is sensitive to the noise in the data, and in particular, to the mean and the median, we apply the information from the mean and median of the tR_i as a function of α to estimate $min_i tR_i$, and hence the true proportion. The mean of tR_i as a function of α is defined as

$$\overline{tR_i(\alpha)} = \frac{1}{m}\sum_{i=1}^{m}\Big[\frac{log(1 + \alpha E_i(AB))}{log(1 + \alpha E_i(A))}\Big] \tag{3.2.6}$$

The value of α plotted on the graph of the function is shown in Figure 3 and Figure 4.

We observe the graphs and apply principal components analysis to conclude that the point gives the minimum ratio, or say, the most accurate proportion, is located at the 'knee' or 'elbow' of the curve (the red points in the Figure 3 and Figure 4).

Now, all we need to do is to find the α, which minimise the radius of curvature.

Step 1: Represent $\overline{tR_i(\alpha)}$ as a curve in the plane.

$$r(\alpha) = (\alpha, \overline{tR_i(\alpha)}) \tag{3.2.7}$$

for any $\alpha_1 \leq \alpha \leq \alpha_2$, where α_1 and α_2 are two real positive numbers.

Step 2: Reparametrise the curve $r(\alpha)$ into unit-speed curve.

Let $s(\alpha)$ be the arc length, defined by

$$s(\alpha) = \int_{\alpha_0}^{\alpha}\sqrt{1 + \Big(\frac{d\,\overline{tR(x)}}{dx}\Big)^2}\,dx \tag{3.2.8}$$

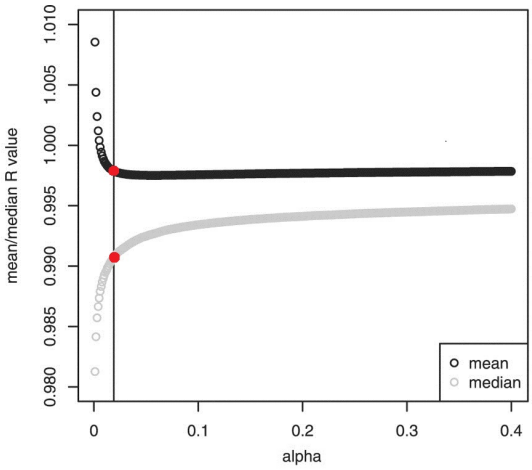

Figure 3: Scree Plot of the Dataset 1 (Clarke et al., 2010)

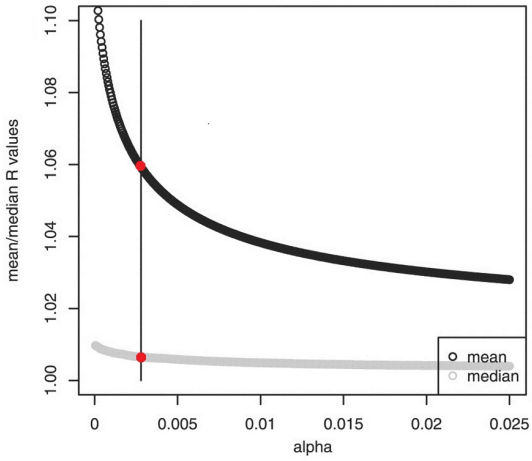

Figure 4: Scree Plot of the Dataset 2 (Clarke et al., 2010)

The function $s(\alpha)$ is a monotonically increasing function because $s(\alpha)$ models the length of a curve; and hence by the Inverse Function Theorem (see 2.2.5), we can express α in term of s, that is

$$\alpha(s) = s^{-1}(\alpha) \tag{3.2.9}$$

In such a way, we are able to find the radius of curvature, which is

$$\rho(s) = \frac{1}{||r''(s)||} \tag{3.2.10}$$

Step 3: Implementation

To find the arc length, we need to find the derivative of $\overline{tR_i(\alpha)}$. In our project, we firstly partition the interval $[\alpha_1, \alpha_2]$ such that

$$\alpha_1 = \beta_0 < \beta_1 < \cdots < \beta_k = \alpha_2,\ \beta_l - \beta_{l-1} = \frac{\alpha_2 - \alpha_1}{k}$$

for $l = 1, 2, \cdots, k$. Then we can obtain the arc length of the curve,

$$s(\beta_j) = \int_{\beta_0}^{\beta_j} \sqrt{1 + \Big(\frac{d\,\overline{tR(x)}}{dx}\Big)^2}\, dx = \frac{1}{\beta_l - \beta_{l-1}} \sum_{l=1}^{j} \sqrt{1 + (\overline{tR_i(\beta_l)}')^2} \tag{3.2.11}$$

where

$$\overline{tR_i(\beta_l)}' = \frac{\overline{tR_i(\beta_l)} - \overline{tR_i(\beta_{l-1})}}{\beta_l - \beta_{l-1}} \tag{3.2.12}$$

for $l, j = 1, 2, \cdots, k$.

Step 4: Find the minimum of radius of curvature

To find the minimum of the radius of curvature is to find the maximum of the second derivative of the corresponding curvature

$$r''(s) = (0, \overline{tR_i(s)}'') \tag{3.2.13}$$

and

$$||r''(s)|| = \sqrt{0^2 + (\overline{tR_i(s)}'')^2} = |\overline{tR_i(s)}''| \tag{3.2.14}$$

where

$$\overline{tR_i(s_l)}'' = \frac{\overline{tR_i(s_{l+1})} - 2\overline{tR_i(s_l)} + \overline{tR_i(s_{l-1})}}{(s_l - s_{l-1})^2} \tag{3.2.15}$$

for $l = 1, 2, \cdots, k-1$.

Once we calculate the minimum of radius of curvature, we are able to find the α, hence the $min_i tR_i$. Therefore, we finally estimate the accurate value of proportion of the pure component A in the mixed sample AB, which, in our project, is the proportion of *input sample* in the *IP sample*.

Model Improvement

In this part, we are going to represent our original work to improve the CLARKE method. We improve the model by changing the method to calculate curvature so that the two-step method of finding curvature is simplified into one step. In the original model, (Clarke et al., 2010) applied the definition of curvature to do calculation. This require us to first reparametrise the curve into a unit-speed curve. In detail, we need to

- find the parametrisation of the curve;
- calculate the first derivative of the parametrised curve;
- calculate the arc-length function of the curve using the first derivative we calculated;
- apply the Inverse Function Theorem to find the inverse of the arc-length function;
- reparametrise the curve into unit-speed curve;
- calculate the second derivative of the reparametrised curve;
- find the Euclidean norm of the second derivative and find its reciprocal.

Whilst in our improved method, we just need to

- find the parametrisation of the curve;
- calculate the first derivative of the parametrised curve;
- calculate the second derivative of the parametrised curve;
- calculate the cross product of the first derivative and second derivative;
- calculate the Euclidean norm of the cross product and the first derivative;
- apply the proposition equation (Equation 2.2.16).

We simplified the model by reducing the process of finding arc-length function, the inverse arc-length function, and the unit-speed reparametrisation of the parametrised curve. In conclusion, we are able to conclude that the efficiency is increased.

Moreover, we also improved the way to calculate the first derivative. We use the Five-Point Midpoint Formula so that the outcomes of the first derivative is likely to be more precise. In detail, the Five-Point Midpoint Formula (see Equation 2.2.18) to approximate the function $f(x)$ has the error term $\frac{h^4}{30} f^{(5)}(\xi)$, where h are the length between a consecutive pair of points. Whilst the error term of the first derivative formula used by

Clarke et al. (2010) is $\frac{h}{2}f''(\xi)$. By comparison, h^4 gives us a much smaller error than h because h is in the order 10^{-1} in our project. Therefore, we are able to conclude that the accuracy of approximation is improved. However, there is no need for us to increase the number of points to approximate the first derivative because h^4, or say, 10^{-4} of error is adequately enough for our project. Although more points of approximation will lead to a more precise result, it will cost much more time for computer to calculate.

The detailed procedure of improved model is described as follows:

Step 1: Represent $\overline{tR_i(\alpha)}$ as a curve in the plane.

$$r(\alpha) = (\alpha, \overline{tR_i(\alpha)}) \tag{3.2.16}$$

for any $\alpha_1 \leq \alpha \leq \alpha_2$, where α_1 and α_2 are two real positive numbers.

Step 2: Calculate the first and second derivatives.

To find the derivatives, we firstly partition the interval $[\alpha_1, \alpha_2]$ such that

$$\alpha_1 = \beta_0 < \beta_1 < \cdots < \beta_k = \alpha_2,\ \beta_j - \beta_{j-1} = \frac{\alpha_2 - \alpha_1}{k}$$

for $j = 1, 2, \cdots, k$. Then, we apply the Five-Point Midpoint Formula to calculate the first derivative,

$$\overline{tR_i(\beta_j)}' = \frac{\overline{tR_i(\beta_{j-2})} - 8\,\overline{tR_i(\beta_{j-1})} + 8\,\overline{tR_i(\beta_{j+1})} - \overline{tR_i(\beta_{j+2})}}{12\,(\beta_j - \beta_{j-1})} \tag{3.2.17}$$

for $j = 2, 3, \cdots, k-2$. Next, we calculate the second derivative,

$$\overline{tR_i(\beta_j)}'' = \frac{\overline{tR_i(\beta_{j+1})} - 2\overline{tR_i(\beta_j)} + \overline{tR_i(\beta_{j-1})}}{(\beta_j - \beta_{j-1})^2} \tag{3.2.18}$$

for $j = 1, 2, \cdots, k-1$.

Step 3: Find the minimum of radius of curvature

To find the minimum of the radius of curvature, we need to apply the proposition. Now, we have

$$r'(\alpha) = (1, \overline{tR_i(\alpha)}') \tag{3.2.19}$$

$$r''(\alpha) = (0, \overline{tR_i(\alpha)}'') \tag{3.2.20}$$

and next, we just need to apply the equation 2.2.16

$$\kappa(\alpha) = \frac{||r(\alpha)' \times r(\alpha)''||}{||r(\alpha)'||^3}. \tag{3.2.21}$$

Though the analysis of the error and the steps of calculation, we may obtain the minimum of the radius of curvature more accurately and more efficiently.

4 Results and Discussions

In this section, we are supposed to compare the two methods, ISOpureR and CLARKE. We firstly generate the test data to examine the stability of both methods; then, we study the statistical characteristics of the results. Finally, we will apply the methods to our real data with analysis.

4.1 Simulated Data

In this part, we have simulated two kinds of data to examine our methods. The first kind of data is simple extreme simulated data, and the other kind of data is simulated data based on the real data.

4.1.1 Simple Extreme Simulated Data

We have generated a simply and extreme simulated dataset, which contains an *input profile* and a *pure* m^6A *binding profile*. There are totally 500 sites and for *input profile*, we set the reads count of the first 250 sites to be 1000 and the rest 250 sites to be 0; for *pure* m^6A *binding profile*, we set the reads count of the first 250 sites to be 0 and the rest 250 sites to be 1000.

Next, we apply binomial distribution to generate the *IP samples* with different mixing proportions with respect to the *input profiles*, $10\%, 20\%, \cdots, 90\%$; and for each proportion, we generate three different samples.

Purity / Methods	Purity 10%			Purity 20%			Purity 30%		
	Set1	Set2	Set3	Set1	Set2	Set3	Set1	Set2	Set3
ISOpureR	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CLARKE	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
	Purity 40%			Purity 50%			Purity 60%		
ISOpureR	0.000	0.000	0.000	0.002	0.003	0.000	0.286	0.286	0.286
CLARKE	$-\infty$	$-\infty$	$-\infty$	0.127	0.127	0.127	0.246	0.244	0.244
	Purity 70%			Purity 80%			Purity 90%		
ISOpureR	0.575	0.570	0.571	1.000	1.000	1.000	1.000	1.000	1.000
CLARKE	0.175	0.180	0.178	0.121	0.118	0.119	0.056	0.058	0.056

Table 4: Results of Simple Extreme Datasets

where the 'Purity' implies the purity we used to generate *IP samples* (real mixing proportions); the numbers in the table represent the proportion of m^6A specific binding mRNA in the *IP samples*. From the results, we are able to observe that both methods are not stable with respect to the extreme cases of dataset. ISOpureR method is likely to approach 0 or 1, and is not sensitive to the purity which are less than 50%; CLARKE method is also insensitive to the purity less than a half, and estimate the corresponding proportion to be $-\infty$. Moreover, the CLARKE method also shows a decreasing tendency when the real mixing proportion keeps increasing after 60% purity.

4.1.2 Simulated Data Based on the Real Data

To figure out whether the problem is caused by the extreme case of the dataset, we next test these two methods by applying a more realistic generated dataset. First of all, we choose to use the dataset gathered from the experiment 'human-A549-C'; in such a

set, there are three *IP samples* and three *input samples*. Next, to keep the generated dataset statistically meaningful, we generate the *pure* m^6A *binding samples* by randomly permutating the *input sample*. Then, we randomly choose 500 gene sites to form our new dataset. Finally, we apply binomial distribution to generate the *IP samples* with different mixing proportions with respect to the *input profiles*, $10\%, 20\%, \cdots, 90\%$ respectively; and for each proportion, we generate 30 different samples.

Here are the results of two methods, ISOpureR and CLARKE respectively, represented by two box plots.

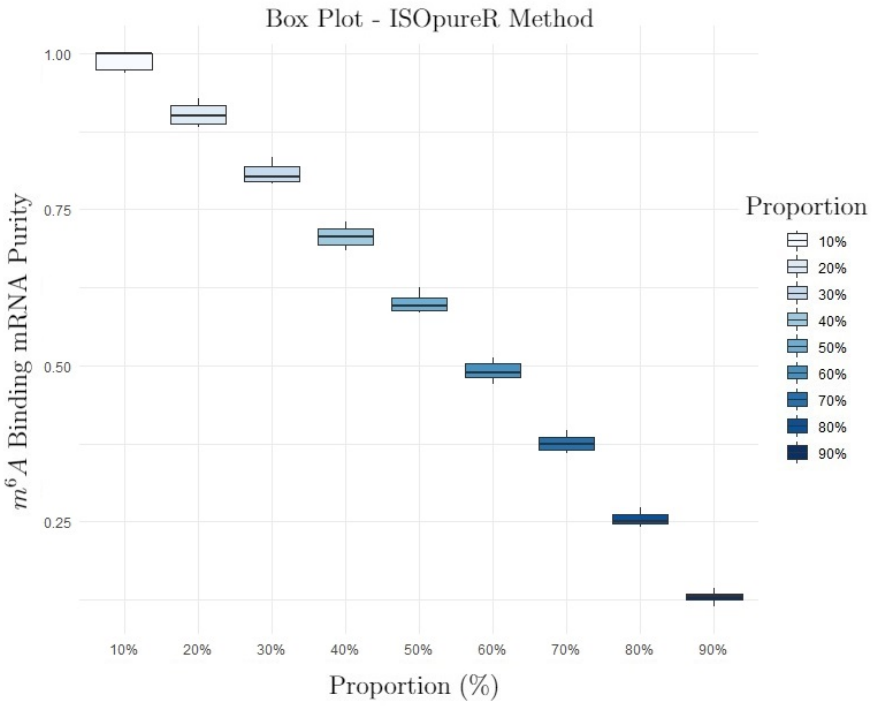

Figure 5: Box Plot of ISOpureR Method

From the box plot of ISOpureR method (see Figure 5), we are able to conclude the ISOpureR method is generally stable. Moreover, the estimated proportion of *pure* m^6A *binding profiles* is likely to be larger than the true mixing proportion, where the overestimation tends to be increasingly higher as the proportion of *input profile* in the *IP sample* decreasing.

From the box plot of CLARKE method (see Figure 6), we may conclude that the method is more likely to be effective when applying to real data instead of extreme data. Although sometimes the outliers may occur, in general, the method is observed to be stable. However, this method is an underestimation of the true proportion.

To explore more information, we plot the scatter plots of two methods, and do the Ordinary Least Squares regression to our results (see Figure 7 and Figure 8).

Moreover, we do the regression analysis to our results, the main features are shown as follows.

- **Regression Analysis to ISOpureR**

  Now, we assume the results gathered from ISOpureR method satisfies the linear

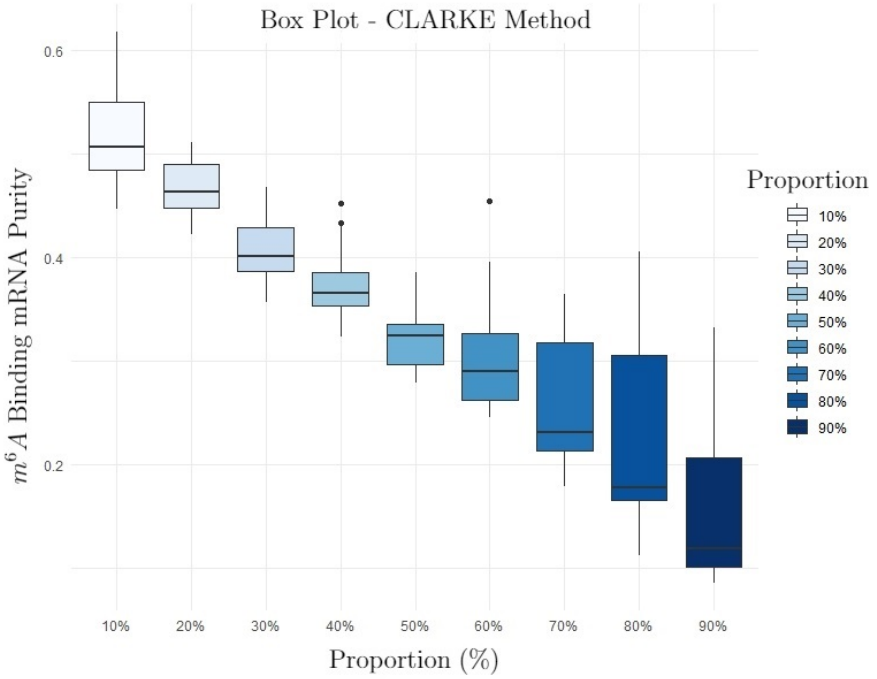

Figure 6: Box Plot of CLARKE Method

regression model, which is

$$y = \beta_0 + \beta_1 x + \epsilon \tag{4.1.1}$$

By doing this, we have assumed that (Prabhakaran, 2016)

- The y-values (or the errors) are independent.
- The y-values can be expressed as a linear function of the x variable.
- Variation of observations around the regression line (the residual SE) is constant (homoscedasticity).
- For given value of x, y values (or the error) are normally distributed.

The fitted model of the method ISOpureR has the coefficients $\beta_0 = 1.123112$, and

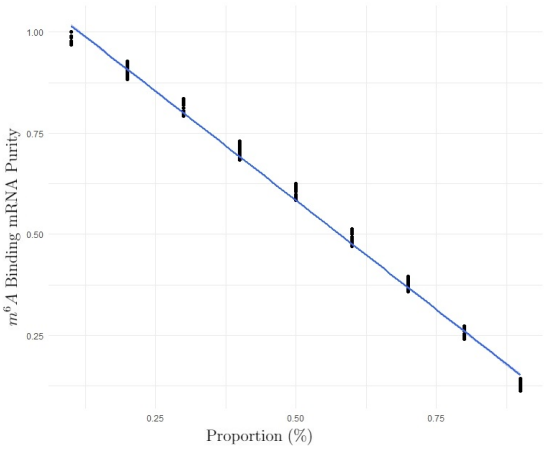

Figure 7: Regression of ISOpureR Methods

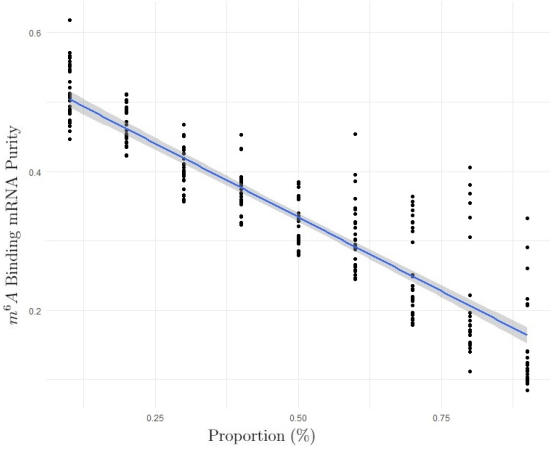

Figure 8: Regression of CLARKE Methods

$\beta_1 = -1.078470$, that is

$$Purity = 1.123112 - 1.078470 * Proportion \tag{4.1.2}$$

Test of Significance:

By Casella and Berger (1990), we know that when analysing sample data, statistical inference allows analysts to assess evidence in favour or some claim about the population from which the sample has been drawn. The methods of inference used to support or reject claims based on sample data are known as tests of significance, with the final conclusion once the test has been carried out is always given by the null hypothesis. Here,

- H_0 : *the coefficient of the linear regression is equal to 0.*
- H_1 : *the coefficient of the linear regression is not equal to 0.*

From the result of regression, the p-value of two coefficients are both less than 2×10^{-16} suggesting that there exist very strong evidence against the H_0 that the coefficients of the equation are equal to 0. Furthermore, the adjusted R^2 is 0.9952, which is very closed to 1; and the p-value of linear regression is less than 2.2×10^{-16}. Both of the two indicators imply that the model fits the data well.

However, a lower p-value and an R^2 value that is closed to 1 cannot represent that the model can adequately describe the data. We need a lack-of-fit test to examine whether the linear model is suitable for the data or not.

Lack-of-Fit Test:

Next, we present a lack-of-fit test to figure out whether the linear regression model is a good choice. By Chatterjee et al. (1987), the lack-of-fit test is used in the numerator to analysis residuals in an analysis of variance in an F-test with the final conclusion given by the null hypothesis, which is,

- H_0: *linear model adequately fits data.*
- H_1: *linear model does not adequately fit data.*

After using R to applying the lack-of-fit test to the linear model, the results are shown in the Table 5. The p-value is 2.2×10^{-16}, which means we have very strong evidence against the H_0. As a results, the linear regression model is concluded to be unsuitable for describing our data.

Model 1: $Purity \sim Proportion$						
Model 2: $Purity \sim as.factor(Proportion)$						
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	268	0.100754				
2	261	0.039327	7	0.061426	58.237	$< 2.2 \times 10^{-16}$

Table 5: R Output of Linear Regression Variance of ISOpureR

Quadratic Polynomial Regression of ISOpureR Method and Analysis

Now, we assume the results gathered from ISOpureR method satisfies the quadratic polynomial regression model, which is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \tag{4.1.3}$$

Now, the results obtained from apply quadratic polynomial to ISOpureR method is shown in Figure 9 with the coefficients $\beta_0 = 1.075930$, $\beta_1 = -0.821117$, and $\beta_2 = -0.257353$. The fitted equation is

$$Purity = 1.075930 - 0.821117 * Proportion - 0.257353 * Proportion^2 \tag{4.1.4}$$

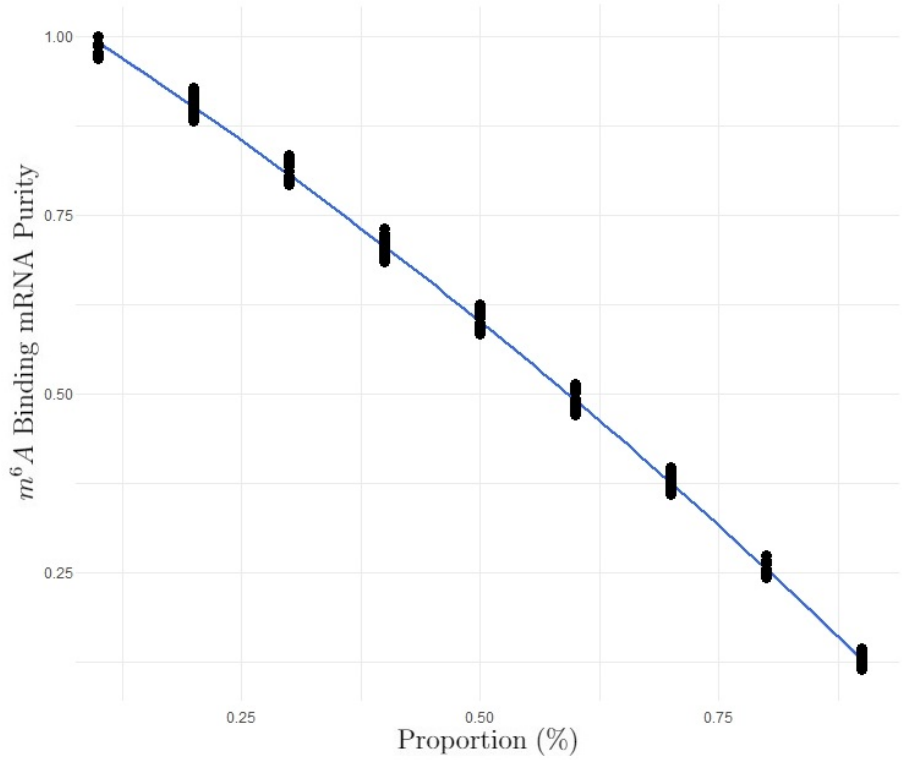

Figure 9: Quadratic Polynomial Regression of ISOpureR Method

1) Test of Significance to the Quadratic Polynomial Regression Model:

The p-value of all the three coefficients are less than 2×10^{16}, which, according to the test of significance, implies that we have very strong evidence against the H_0 that the coefficients of the equation are equal to 0. Moreover, the R^2 of the model is 0.9981 implying that the model fits the data very well.

2) Lack-of-Fit Test of the Quadratic Polynomial Regression Model:

In addition, we further do the lack-of-fit test to our new model, and the result of the test are shown in the Table 6 with the p-value is equal to 0.9577 suggesting we have no evidence against the H_0 that the quadratic polynomial regression model adequately fits our data.

3) Diagnostic Test to the Quadratic Polynomial Regression Model:

Moreover, we do the diagnostic test to the quadratic polynomial regression, the results are shown in the Figure 10.

Model 1: $Purity \sim Proportion + I(Proportion^2)$						
Model 2: $Purity \sim as.factor(Proportion) + I(as.factor(Proportion^2))$						
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	267	0.039556				
2	261	0.039327	6	0.00022897	0.2533	0.9577

Table 6: R Output of Quadratic Polynomial Regression Variance of ISOpureR

In the Figure 10(a), named the residual plot, the x-axis is the predicted or fitted purity values and the y-axis is the residuals, or say, errors. In addition, the red line is fairly horizontal, which means the variance in this model is not a constant. For the second graph of the Figure 10(b), known as the quantile quantile (QQ) plot, the x-axis is the ordered theoretical residuals and the y-axis is the ordered, observed, standardised residuals. As the graph shows, the points do not fit the dash line well. Consequently, we are supposed to have a further discussion that the residual terms, or say, the errors are normally distributed. For the Figure 10(c), the scale-location graph, the x-axis refers to the predicted or fitted purity values and the y-axis is the ordered, observed, the square root of standardised residuals. In this graph, the red line is relatively horizontal and the residuals is observed to be spread equally alone the range of predictors; hence, the assumption of equal variance (homoscedasticity) holds. In terms of the Figure 10(d), this plot is to figure out the influential cases.

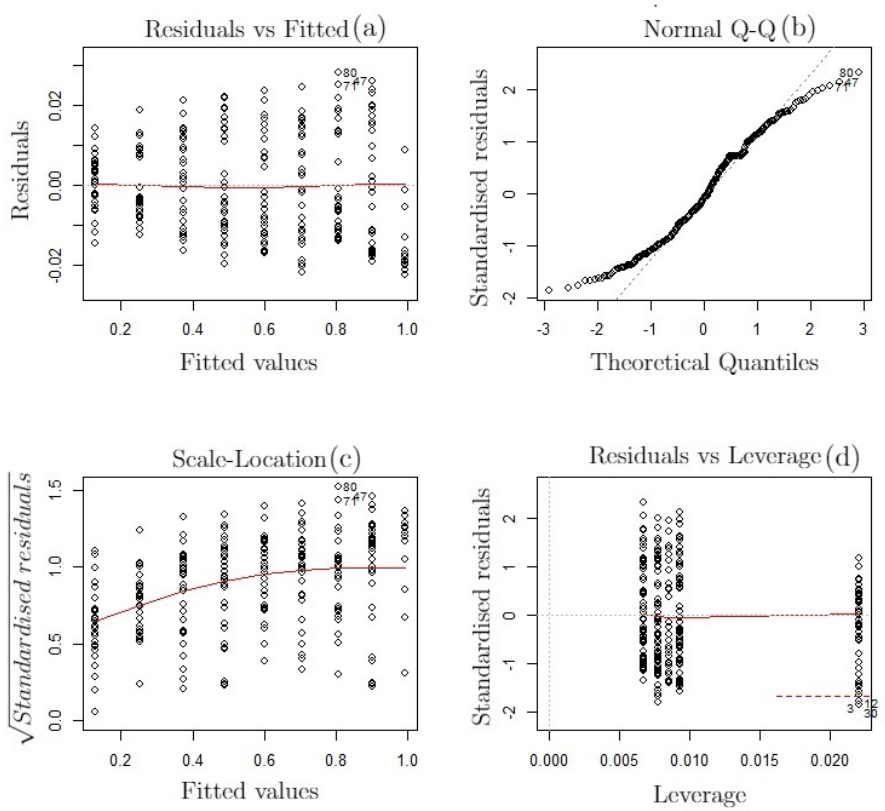

Figure 10: Diagnostic Test for Quadratic Polynomial Regression of ISOpureR Method

To be specific, the dataset may contain some extreme points, but they could be not influential to determine a regression model, which means that the model would not be very different in the situation with or without these points. However, for the outlying points outside of the dashed line, Cook's distance, they are influential to our regression results. According to our Residual vs Leverage plot, there are few points that beyond the Cook's distance line, which are influential and may have altered the regression result. Consequently, the model is generally stable since almost all the points are not influential to our model.

4) Shapiro-Wild Test to the Quadratic Polynomial Regression Model:

To further test for the normality of the residuals, we apply the Shapiro–Wilk test (Wild and Shapiro, 1965) to the residuals of quadratic polynomial regression of ISOpureR method. The Shapiro–Wilk test is a test of normality in frequentist statistics with hypothesis

- H_0: *the sample came from a normally distributed population.*
- H_1: *the sample did not come from a normally distributed population.*

The p-value of the Shapiro–Wilk test is 1.365×10^{-5}, suggesting we have very strong evidence against the H_0. In other words, the residuals do not follow normal distribution. However, according to the (Lumley et al., 2002), the least-squares linear regression do not require any assumption of normal distribution in sufficiently large samples in public health research. Moreover, formal statistical tests for normality are especially undesirable as they will have low power in the small samples where the distribution matters and high power only in large samples where the distribution is unimportant.

Median Regression for ISOpureR Method and Analysis

In the previous part of the section, we have applied the linear regression and quadratic polynomial regression to our results. However, the QQ plot implied that the residuals might not follow normal distribution which undermine our assumption of linear regression and quadratic polynomial regression. Here, we further use the median regression to model our results since according to Furno and Vistocco (2018), the quantile regression does not need the assumption of normality of residuals.

We assume the results gathered from ISOpureR method satisfies the median regression model (see 2.2.19 for details), which is

$$y = \alpha_0 + \alpha_1 x + F^{-1}(0.5) \tag{4.1.5}$$

where α_0 and α_1 are the coefficients of the median regression model; F represents the common distribution function of the errors without any distributional assumptions. After doing regression, we obtain the regression coefficients where $\alpha_0 = 1.12364$, and $\alpha_1 = -1.07808$. In other words,

$$Purity = 1.12364 - 1.07808 * Proportion \tag{4.1.6}$$

where the plot of median regression is shown in the figure 11.

Lack-of-Fit Test for Median Regression of ISOpureR Method

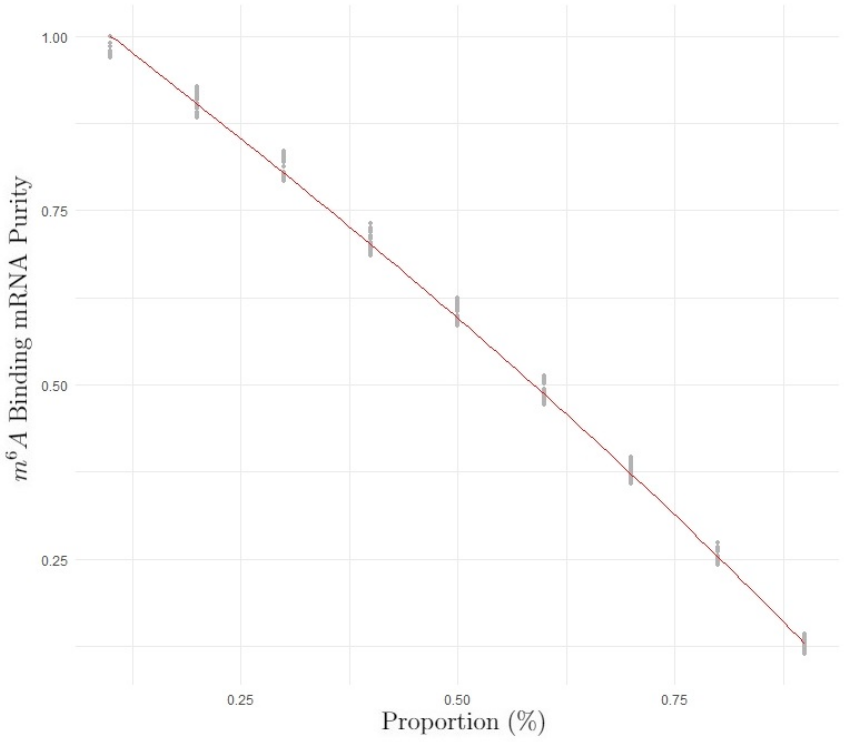

Figure 11: Median Regression of ISOpureR Method

According to the He and Zhu (2003), we apply the lack-of-fit test to the median regression of ISOpureR method, and the results will given with the hypothesis test where

- H_0: *median regression model adequately fits data.*
- H_1: *median regression model does not adequately fit data.*

After application of lack-of-fit test to the ISOpureR model, the p-value is less than 2.2×10^{-16}, implying that we have very strong evidence against the H_0.

Second Order Median Regression Model for ISOpureR Method

Now, we assume the results gathered from ISOpureR method satisfies the generalised median regression model, which is

$$y = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + F^{-1}(0.5) \tag{4.1.7}$$

where α_0, α_1 and α_2 are the coefficients of the median regression model; F represents the common distribution function of the errors without any distributional assumptions.

After calculating the coefficients, we have $\alpha_0 = 1.09092$, $\alpha_1 = -0.88933$, $\alpha_2 = -0.19887$, that is,

$$Purity = 1.09092 - 0.88933 * Proportion - 0.19887 * Proportion^2 \tag{4.1.8}$$

where the plot of median regression model for ISOpureR method is displayed in the Figure 12.

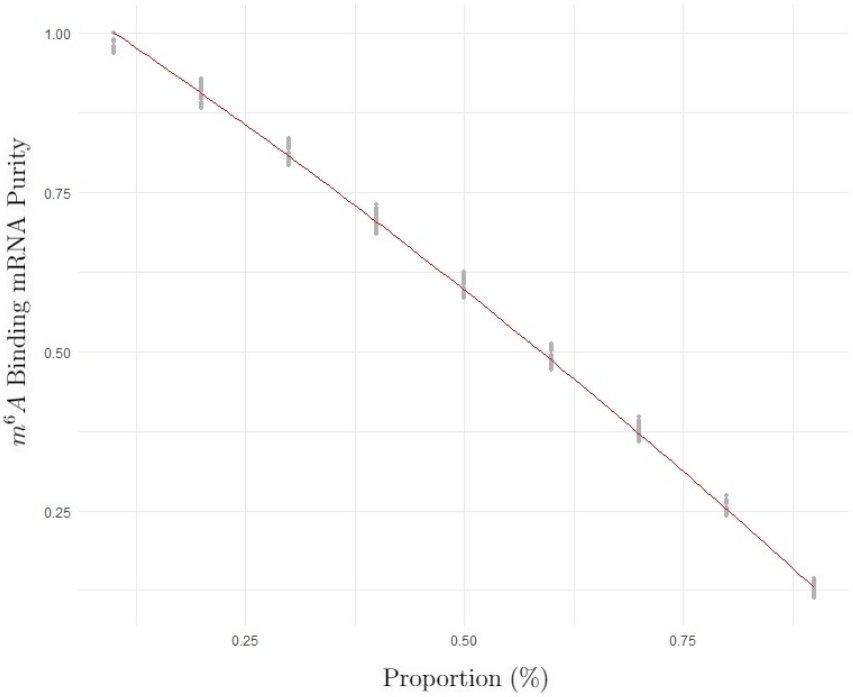

Figure 12: Generalised Median Regression Model of ISOpureR Method

Lack-of-Fit Test for Generalised Median Regression Model of ISOpureR Method

Additionally, we further do the lack-of-fit test using the method introduced by He and Zhu (2003) to our generalised median regression model, and the p-value of the test is equal to 0.21 suggesting we have no evidence against the H_0 that the generalised median regression model adequately fits our data.

- **Regression Analysis to CLARKE**

  Now, we assume the results gathered from CLARKE method satisfies the linear regression model, which is

  $$y = \beta_0 + \beta_1 x + \epsilon \tag{4.1.9}$$

  The fitted model of the method ISOpureR has the coefficients $\beta_0 = 0.546677$, and $\beta_1 = -0.425047$, that is

  $$Purity = 0.546677 - 0.425047 * Proportion \tag{4.1.10}$$

  Test of Significance:

  By Casella and Berger (1990), the final results of significance test is given by the null hypothesis with

  - H_0 : *the coefficient of the linear regression is equal to 0.*
  - H_1 : *the coefficient of the linear regression is not equal to 0.*

  From the result of regression, the p-value of two coefficients are both less than 2×10^{-16} suggesting we have very strong evidence against the H_0 that the coefficients

of the equation are equal to 0. For adjusted R^2, it is equal to 0.8154, implying that the model fits the data well. However, from the plot of linear regression model (see Figure 8), the standard error of each group of proportion is likely to increase when the proportion becomes higher and higher. As a result, this observation may destroy the assumptions of linear regression model, that the variance of observations around the regression line (the residual standard error) is constant (homoscedasticity). Next, we may need further discussion to the noise of the regression model.

Lack-of-Fit Test:

We present a lack-of-fit test to figure out whether the linear regression model is a good choice. By Chatterjee et al. (1987), in the lack-of-fit test hypothesis are listed in the following,

- H_0: *linear model adequately fits data.*
- H_1: *linear model does not adequately fit data.*

After analysing the residuals of the regression results of CLARKE method, we may not be sure about whether the linear regression model fits the data best. Here, we do the lack-of-fit test to examine whether the linear model adequately fits data. The p-value of the linear regression model of the CLARKE method is 0.1153 implying that we have no evidence against the H_0. In such a case, we are able to conclude that the linear model adequately fits our data.

Diagnostic Test:

To further examine the noise of the model, we apply diagnostic test to the CLARKE regression model; and plot four graphs to illustrate the main characteristics.

In Figure 13(a), the residual plot, the red line in this graph is regard to be horizontal, which means the variance in this model may be constant. For the QQ plot (Figure 13(b)), although the points fit the diagonal line well in the middle part of the graph, the points at two sides are higher than the standard line. Consequently, we need further discussion to conclude whether the residuals follow normal distribution. For the Figure 13(c), the red line is almost horizontal and the residuals is observed to be spread equally alone the range of predictors; hence, the assumption of equal variance (homoscedasticity) holds. In terms of the Figure 13(d), Residual vs Leverage plot, there are few points that beyond the Cook's distance line, suggesting the model is generally stable.

Shapiro-Wild Test to the CLARKE Method:

To test the normality of the residuals, we apply the Shapiro–Wilk test (Wild and Shapiro, 1965) to the residuals of linear regression model of CLARKE method. The Shapiro–Wilk test is a test of normality in frequentist statistics with hypothesis

- H_0: *the sample came from a normally distributed population.*
- H_1: *the sample did not come from a normally distributed population.*

The p-value of the Shapiro–Wilk test is 1.136×10^{-11}, suggesting we have very strong evidence against the H_0. In other words, the residuals do not follow normal distribution. However, according to the (Lumley et al., 2002), the test of normality

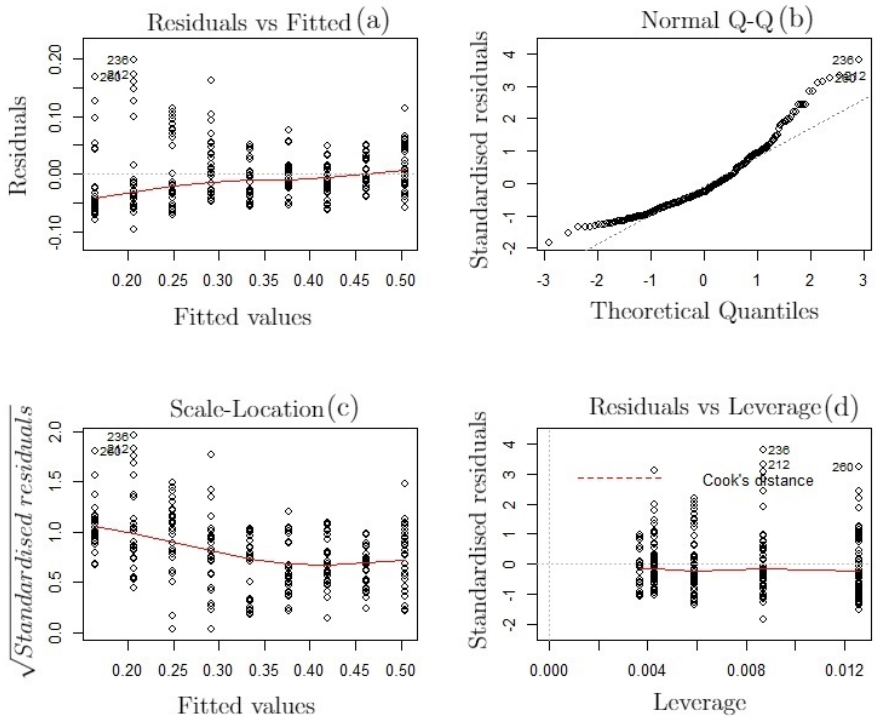

Figure 13: Diagnostic Test of CLARKE Regression Model

is not important in the research related to the public health. In a word, the non-normality will not destroy our model.

Median Regression for CLARKE Method and Analysis:

In the previous part of the section, we have applied the linear regression to the results calculated with application of CLARKE method. However, the QQ plot implied that the residuals might not follow normal distribution which undermine our assumption of linear regression model. Given that the quantile regression does not need the assumption of normality of residuals (Furno and Vistocco, 2018), we further examine whether the median regression model fit the results of CLARKE method.

We assume the results gathered from ISOpureR method satisfies the median regression model, which is

$$y = \alpha_0 + \alpha_1 x + F^{-1}(0.5) \tag{4.1.11}$$

where α_0 and α_1 are the coefficients of the median regression model; F represents the common distribution function of the errors without any distributional assumptions. After doing regression, we obtain the regression coefficients where $\alpha_0 = 0.55341$, and $\alpha_1 = -0.46883$. In other words,

$$Purity = 0.55341 - 0.46883 * Proportion \tag{4.1.12}$$

where the plot of median regression is shown in the figure 14.

Lack-of-Fit Test for Median Regression of CLARKE Method

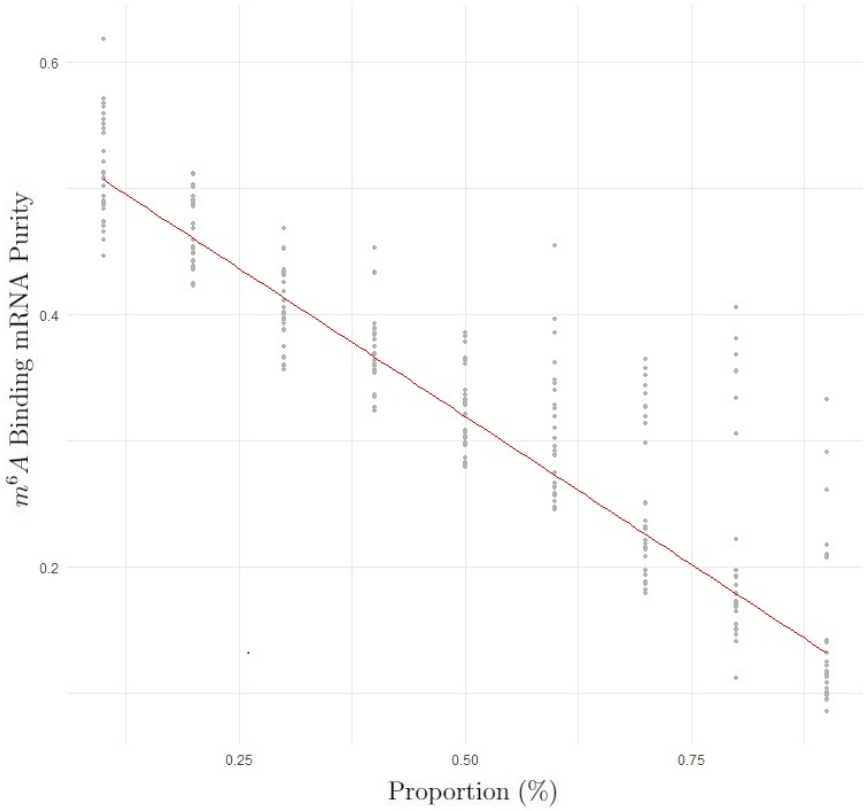

Figure 14: Median Regression of CLARKE Method

According to the He and Zhu (2003), we apply the lack-of-fit test to the median regression of CLARKE method with the hypothesis that

- H_0: *median regression model adequately fits data.*
- H_1: *median regression model does not adequately fit data.*

After application of lack-of-fit test to the ISOpureR model, the p-value is 0.09, implying that we only have weak evidence to against the H_0. In other word, the model is able to fit the data well.

In comparison of both of the methods, we are able to conclude that the ISOpureR method is likely to describe data more adequately when we apply quadratic polynomial regression model to fit the results; whilst for the CLARK method, the linear model is sufficient enough to do so. Moreover, with respect to the performance of two methods, both of them are not stable in terms of the simple extreme data; however, when we apply the methods to the simulated data based on the real data, the ISOpureR may perform better than the CLARKE method because of the underestimation from the CLARKE method. After examining both of the methods' results, we can conclude that both of the methods are stable with respect to the simulated data generated from the real data. In detail, the ISOpureR method is significantly stable with respect to the second order generalised median regression model; whilst the CLARKE method is likely to perform stably in terms of the median regression. Generally speaking, ISOpureR method provides us more accurate, and stable results, than the CLARKE method. Although it might be true that, as Clarke et al. (2010) illustrated, the CLARKE method have increased the efficiency

of estimation compared with the prototype of the CLARKE method created by Gosink et al. (2008) in terms of the deconvolution of tumour sample data; with respect to our deconvolution of m^6A-containing mRNA data, the method is not as efficient as it is stated. Thus, with respect to the results we obtained from simulated data so far, when we analysing the m^6A-containing data, ISOpureR method is likely to perform better than the CLARKE method.

4.2 Real Data

After analysing the test data, we apply the two methods to our real data to estimate the true proportion of m^6A specific binding mRNA in the *IP sample*. Here, we present five results calculated from both of the methods from five real datasets, which are 'human-A549-C', 'human-A549-METTL3-', 'human-A549-METTL14-', 'human-H1ESC-T48', and 'human-hESC-C' respectively.

Experiments / Methods	human-A549-C			
	SRR1182619	SRR1182621	SRR1182623	
ISOpureR	1.000	1.000	1.000	
CLARKE	0.397	0.452	0.450	
	human-A549-METTL3-			
	SRR1182615	SRR1182617	SRR1182629	
ISOpureR	1.000	1.000	1.000	
CLARKE	0.464	0.540	0.375	
	human-A549-METTL14-			
	SRR1182607	SRR1182609	SRR1182611	SRR1182613
ISOpureR	1.000	1.000	1.000	1.000
CLARKE	0.445	0.461	0.430	0.386
	human-H1ESC-T48			
		SRR1035218	SRR1035220	
ISOpureR		1.000	1.000	
CLARKE		0.686	0.698	
	human-hESC-C			
		SRR1035222	SRR1035224	
ISOpureR		1.000	1.000	
CLARKE		0.665	0.619	

Table 7: Results of Real Datasets

In the 'Results of Real Datasets' (Table 7), the numbers represent the proportion of m^6A binding mRNA in the *IP sample*. Although from the previous subsection, we have concluded that the method ISOpureR has a better efficiency than the method CLARKE, it always converge to 1 when we perform the method to the real data. However, the CLARKE method may perform better than the ISOpureR method with respect to the

real m^6A-containing data. This problem may come from the size of the dataset; where when we use the test data, there are totally 500 gene sites, whilst when we refer to the real data, the number of gene sites becomes 69946.

This problem may result from the number of gene sites. In our project, we do an analogy between deconvolution of the tumour samples and the deconvolution of the m^6A-containing samples. In other words, these two methods are not designed to process the m^6A-containing dataset. As a result, when we apply ISOpureR method to out data, the extremely large number of gene sites make the method always converge to 1. Moreover, another difference between the data of m^6A-containing samples and the data of tumour samples is that, our data does not contain many reference profiles. This difference may cause the CLARKE method shows instability when we apply the linear regression model to fit the test data.

In conclusion, the performances of both methods depend on the data we used. In terms of the number of gene sites, when our data has a large number of gene sites, the ISOpureR method will fail since it always converge to 1; however, when the number is relatively small, the ISOpureR method performs better than the CLARKE method no matter what indicators we are used to examine. Moreover, the CLARKE method is likely to be useful when the number of gene sites is large. Although it can also estimate a stable value when the number of gene sites is small, the results we obtained from the CLARKE method is an underestimation and become more and more unstable when the proportion of *input profiles* increase. Next, with respect to the extreme case of data, both of the methods fail to estimate the true proportion of the pure component in the contaminated samples.

5 Conclusions

In this project, we draw an analogy between the method of deconvolution of the tumour samples with respect to the normal samples and the deconvolution of the *IP samples* in terms of the *input samples*.

In order to accomplish the goal of estimation of true proportion of m^6A specific binding mRNA in the *IP sample*, we have examined two methods, ISOpureR and CLARKE respectively. After reviewing the necessary preliminary knowledge of biology and mathematics, we introduced the basic ideas and algorithms of these two method. For the ISOpureR method, the main idea is to maximising the complete likelihood function to estimate the parameters with application of the flexible preconditioned conjugate gradient method; whilst for the CLRAKE method, the model is generated after doing the components analysis and then, we apply the knowledge of differential geometry to find the minimum of radius of curvature.

Furthermore, after having introduced the basic idea of the methods, we applied test data and real data to test the performance of both of the methods with the application of the knowledge of regression analysis. First, we used the simple extreme data to test both two methods and the results showed that both of the methods were unstable to the dataset. Next, we generated a simple simulated dataset using real data, where we generated the *pure* m^6A *binding samples* by randomly permutating the *input sample*; after randomly choosing 500 gene sites to form our new dataset, we apply binomial

distribution to generate the *IP samples* with different mixing proportions with respect to the *input profiles*. Finally, we applied 30 different simulated real data samples to examine the ISOpureR and CLARKE method.

Then, we calculated and analysed the results of both of the methods. For ISOpureR method, we firstly applied the linear regression model to the results and test its rationality. Although the significant test implied that the linear regression model fit the data quite well, the lack-of-fit test suggested that we have very strong evidence against the null hypothesis, that linear model adequately fits data. As a consequence, we used the quadratic polynomial regression to model the ISOpureR method. Similarly, we did the significance test and lack-of-fit test, and the tests have illustrated that the model described the data well and can adequately fit data. However, the diagnostic test implied that the residuals of the fitted results might not follow normal distribution, undermining the assumption of the quadratic polynomial regression. Although evidence have showed that the non-normality was not supposed to be important and would not destroy our regression model due to the research is all about public health, the second order generalised median regression model shows stability and get rid of the assumption requiring normality of the residuals. Therefore, the ISOpureR method was able to be stable and estimated accurately under the second order generalised median regression model. For CLARKE model, we have done the linear regression to the results and the low value of significance test showed that we have very strong evidence against the null hypothesis, the coefficients of the linear regression is equal to 0, and the R^2 value also implied that the model fits the data well. Similar to the ISOpureR model, we also did the lack-of-fit test to the CLARKE method, and the result provided no evidence for us to against the null hypothesis that linear model adequately fits our data. As a result, the linear model was adequately describe the data. Next, we also do the diagnostic test to the CLARKE method and all the results suggested that the regression model met the assumptions of linear regression except the normality of the residuals. Although evidence have showed that the non-normality was not supposed to destroy our regression model because of insignificance of the normality of the public health research, median regression model performed more efficiently because the avoidance of the assumption requiring that the normality of the residuals.

After analysing the behaviours of the two methods in terms of the simulated data, we have also examined the performance of both of the methods when applying to the real data. After computing five different samples of data, we found that the ISOpureR method would always converge to 1 after iteration, whilst the CLARKE method was more stable and accurate. In comparison of both of the methods in terms of both kinds of data, we conclude that the ISOpureR method is more stable and performed more efficiently when the data has a relatively small number of gene sites; otherwise, the results of the ISOpureR method would always converge to 1. For the CLARKE method, it has a better performance when the number of gene sites is relatively large; although it is able to obtain some stable results when applied to the data with a small number of gene sites, the efficiency of estimation is much worse than the ISOpureR method.

In the future, we may improve the model by adjust the parameters of the ISOpureR method such that it can be more efficient when applying to the data with a large number of gene sites; and figure out the critical number of gene sites so that we are able to determine under which conditions of the data, the ISOpureR method will provide us a

better value of proportion, and under what circumstance, the CLARKE will be more stable and accurate than the ISOpureR method.

6 Acknowledgement

I am sincerely grateful for my thesis supervisor, Dr. Jionglong Su and Dr. Jia Meng, for their patient guidance and tutorial. In addition, many thanks to Dr. Zhen Wei for his time to illustrate the preliminary knowledge about biology and statistics in such a detail to me.

References

Y Aloni, R Dhar, and G Khoury. Methylation of nuclear simian virus 40 rnas. *Journal of Virology*, 32(1):52–60, 1979. ISSN 0022-538X. URL https://jvi.asm.org/content/32/1/52.

Karen Beemon and Jerry Keith. Localization of n6-methyladenosine in the rous sarcoma virus genome. *Journal of Molecular Biology*, 113(1):165 – 179, 1977. ISSN 0022-2836. doi: https://doi.org/10.1016/0022-2836(77)90047-X. URL http://www.sciencedirect.com/science/article/pii/002228367790047X.

Annette Burden, Richard L. Burden, and J Douglas Faires. *Numerical Analysis, 10th ed.* Cengage, 01 2016. ISBN 1305253663. doi: 10.13140/2.1.4830.2406.

G. Casella and R.L. Berger. *Statistical Inference*. Duxbury advanced series. Brooks/Cole Publishing Company, 1990. ISBN 9780534119584. URL https://books.google.co.uk/books?id=xA7vAAAAMAAJ.

Samprit Chatterjee, R J. Brock, and G C. Arnold. Applied regression analysis and experimental design. *Journal of Business and Economic Statistics*, 5:160, 01 1987. doi: 10.2307/1391229.

Bertrand Clarke, Jennifer Clarke, and Pearl Seo. Statistical expression deconvolution from mixed tissue samples. *Bioinformatics*, 26(8):1043–1049, 03 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq097. URL https://doi.org/10.1093/bioinformatics/btq097.

Ronald Desrosiers, Karen Friderici, and Fritz Rottman. Identification of methylated nucleosides in messenger rna from novikoff hepatoma cells. *Proceedings of the National Academy of Sciences*, 71(10):3971–3975, 1974. ISSN 0027-8424. doi: 10.1073/pnas.71.10.3971. URL http://www.pnas.org/content/71/10/3971.

Dan Dominissini, Sharon Moshitch-Moshkovitz, Mali Salmon-Divon, Ninette Amariglio, and Gideon Rechavi. Transcriptome-wide mapping of n6-methyladenosine by m 6a-seq based on immunocapturing and massively parallel sequencing. *Nature protocols*, 8: 176–89, 01 2013. doi: 10.1038/nprot.2012.148.

Friederike Dündar, Luce Skrabanek, and Paul Zumbo. *Introduction to Differential Gene Expression Analysis Using RNA-Seq*. Weill Cornell Medical College, 2015. URL http://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf.

Marilena Furno and Domenico Vistocco. Quantile regression: Estimation and simulation. *Quantile Regression: Theory and Applications*, pages 1–287, 01 2018. doi: 10.1002/9781118863718.

A.B. Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. Bayesian data analysis. *The Statistician*, 45, 01 2003. doi: 10.2307/2988417.

Mark Gosink, Howard Petrie, and Nicholas Tsinoremas. Electronically subtracting expression patterns from a mixed cell population. *Bioinformatics (Oxford, England)*, 23: 3328–34, 01 2008. doi: 10.1093/bioinformatics/btm508.

Mark M. Gosink, Howard T. Petrie, and Nicholas F. Tsinoremas. Electronically subtracting expression patterns from a mixed cell population. *Bioinformatics*, 23(24): 3328–3334, 10 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm508. URL https://doi.org/10.1093/bioinformatics/btm508.

Xuming He and Li-Xing Zhu. A lack-of-fit test for quantile regression. *Journal of the American Statistical Association*, 98 (464):1013–1022, 2003. doi: 10.1198/016214503000000963. URL https://doi.org/10.1198/016214503000000963.

Ian Jolliffe. Principal component analysis / i. t. jolliffe. *SERBIULA (sistema Librum 2.0)*, 03 2002.

Roger Koenker. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005. doi: 10.1017/CBO9780511754098.

Thomas Lumley, Paula Diehr, Scott Emerson, and Lu Chen. The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, 23(1):151–169, 2002. doi: 10.1146/annurev.publhealth.23.100901.140546. URL https://doi.org/10.1146/annurev.publhealth.23.100901.140546. PMID: 11910059.

Christopher A. Maher, Chandan Kumar-Sinha, Xuhong Cao, Shanker Kalyana-Sundaram, Bo Han, Xiaojun Jing, Lee Sam, Terrence Barrette, Nallasivam Palanisamy, and Arul M. Chinnaiyan. Transcriptome sequencing to detect gene fusions in cancer. *Nature*, 458:97–101, 2009. doi: https://doi.org/10.1038/nature07638.

Jia Meng, Shao-Wu Zhang, Yufei Huang, and Lian Liu. Qnb: differential rna methylation analysis for count-based small-sample sequencing data with a quad-negative binomial model. *BMC Bioinformatics*, 18(1):387, Aug 2017. ISSN 1471-2105. doi: 10.1186/s12859-017-1808-4. URL https://doi.org/10.1186/s12859-017-1808-4.

Kate Meyer and Samie Jaffrey. The dynamic epitranscriptome: N6-methyladenosine and gene expression control. *Nature Reviews Molecular Cell Biology*, 15, 04 2014. doi: 10.1038/nrm3785.

Ryan D. Morin, Matthew Bainbridge, Anthony Fejes, Martin Hirst, Martin Krzywinski, Trevor J. Pugh, Helen McDonald, Richard Varhol, Steven J.M. Jones, , and Marco A. Marra. Profiling the hela s3 transcriptome using randomly primed cdna and massively parallel short-read sequencing. *BioTechniques*, 45(1):81 – 93, 2008. doi: https://www.future-science.com/doi/pdf/10.2144/000112900.

Y. Notay. Flexible conjugate gradients. *SIAM Journal on Scientific Computing*, 22(4):1444–1460, 2000. doi: 10.1137/S1064827599362314. URL https://doi.org/10.1137/S1064827599362314.

Robert P. Perry, Dawn E. Kelley, Karen Friderici, and Fritz Rottman. The methylated constituents of l cell messenger rna: Evidence for an unusual cluster at the 5' terminus. *Cell*, 4(4):387 – 394, 1975. ISSN 0092-8674. doi: https://doi.org/10.1016/0092-8674(75)90159-2. URL http://www.sciencedirect.com/science/article/pii/0092867475901592.

Selva Prabhakaran. *Assumptions of Linear Regression*. r-statistics.co, 2016. doi: http://r-statistics.co/Assumptions-of-Linear-Regression.html.

A.N. Pressley. *Elementary Differential Geometry*. Springer-Verlag London, 2 edition, 2010. doi: 10.1007/978-1-84882-891-9.

Gerald Quon, Syed Haider, Amit Deshwar, Ang Cui, Paul C Boutros, and Quaid Morris. Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome medicine*, 5:29, 03 2013. doi: 10.1186/gm433.

Carl Edward Rasmussen and Christopher K. I. Williams. Gaussian processes for machine learning. *Matlab Library*, 2006. doi: https://rdrr.io/cran/gpr/src/R/minimize.r.

Wenyu Sun and Ya-Xiang Yuan. *Optimization Theory and Methods: Nonlinear Programming*, volume 1. Springer, 1 edition, 2006. doi: 10.1007/b106451.

Xiao Wang, Boxuan Zhao, Ian Roundtree, Zhike Lu, Dali Han, Honghui Ma, Xiaocheng Weng, Kai Chen, Hailing Shi, and Chuan He. N6-methyladenosine modulates messenger rna translation efficiency. *Cell*, 161:1388–1399, 06 2015. doi: 10.1016/j.cell.2015.05.014.

M. B. Wild and S. S. Shapiro. An analysis of variance test for normality (complete samples)†. *Biometrika*, 52(3-4):591–611, 12 1965. ISSN 0006-3444. doi: 10.1093/biomet/52.3-4.591. URL `https://doi.org/10.1093/biomet/52.3-4.591`.