STATISTICAL METHODS FOR DECONVOLUTION IN CANCER GENOMICS

Liuqing (Jasmine) Yang

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2019

Approved by:

J. S. Marron

Hongtu Zhu

Shu Lu

Quoc Tran-Dinh

Kai Zhang

ABSTRACT

LIUQING (JASMINE) YANG: Statistical Methods for Deconvolution in Cancer Genomics
(Under the direction of J. S. Marron and Hongtu Zhu)

With the advance of deep sequencing techniques, intratumor heterogeneity becomes a prevalent confounding factor to tumor genomic profiling studies. The heterogeneous composition of a tumor tissue can potentially lead to false positive differential expression conclusions and influence patients' clinical outcomes and therapeutic responses. Many deconvolution methods aiming to separate the subcomponent signals have been developed in the past decades, modeling the tumor genomic profiling as a linear combination of the abundance of the mixing components. In this dissertation, we characterize a two-components (tumor versus non-tumor) model and develop a Fast Tumor Deconvolution (FasTD) pipeline to address the heterogeneity issue. We build a semi-parametric regression-based framework utilizing raw measured gene expression values, and provide mixing proportions and individual component genomic profiles as outputs. We demonstrate our method and show it is more than a thousand times faster than several current probabilistic models. Both simulated data and real data applications are provided to demonstrate the effectiveness of our proposed method. Our method is then extended to deconvolve heterogeneous tumor samples with more than two subcomponents. The extended pipeline (FasTDK) can effectively deconvolve an unknown component in K-subcomponent mixtures provided with $K-1$ reference profiles.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the help of many individuals who have offered guidance and support in completion of this work.

I would like to offer my utmost gratitude to my advisors Drs. J. S. Marron and Hongtu Zhu for their mentoring and guidance throughout my Ph.D. study. They taught me how to think statistically and face every challenge as an opportunity. Without their support, this thesis would not have been completed or written.

I gratefully acknowledge my dissertation committee Drs. Shu Lu, Quoc Tran-Dinh and Kai Zhang for their valuable advice and insights contributed to this study. Many thanks go in particular to Dr. Wenyi Wang from the Department of Bioinformatics and Computational Biology at MD Anderson Cancer Center. Without the collaboration of her lab this work would not have been started and the biological insights of the work would not have been developed.

I give further thanks to my fellow Ph.D. students of the Department of Statistics and Operations Research for building up a creative and lively learning environment during my research work.

Most of all I owe my deepest gratitude to my husband Mr. Yang Chen and my parents Mr. Zhongmao Yang and Mrs. Min Wang for their unconditional love, unending support and uplifting encouragement that helped me to overcome many challenges in pursuing this doctoral degree. Without their support none of this would be possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

1.1 Background: Intratumor Heterogeneity

Recently there have been extensive studies researching tumor heterogeneity due to its significant impact on tumor genetic analysis and therapeutic decisions, (Burrell et al., 2013). There are two levels of tumor heterogeneity: intertumor and intratumor. The former one refers to the variation between individuals with the same tumor type, while the latter one is observed within one tumor tissue sample. Intratumor heterogeneity can be caused by subclone events or mixture of other cell/tissue types in tumor tissues, such as blood vessels, immune cells, stromal cells, etc. This variation within a tumor tissue sample is what we will focus and build a model on in this dissertation work. Because of recent advances in deep sequencing techniques, intratumor heterogeneity is observed more and more at the molecular level, (Shibata, 2012). Gene expression profiles extracted from these mixed tumor samples for studying tumor/normal differences can be strongly confounded by the composition of cell types in the mixture samples (Figure 1.1). Meanwhile, extensive studies have demonstrated this intratumor heterogeneity is closely related to therapeutic responses. Shibata (2012) concludes that more heterogeneity in tumors with many subtype-specific mutations is positively correlated with the failure rate of chemotherapy. In addition, Moffitt et al. (2015) report that patients with samples with an 'activated' stromal subtype had a much worse survival rate than patients with another normal-like stroma subtype. This result suggests the importance of identifying individual sub-components in a bulk tissue, which can lead to alternative therapeutic choices.

The importance of *quantifying* the sub-components in tumor samples is further supported by studies of tumor micro-environment. Junttila and de Sauvage (2013) point out that tumors should be evaluated as complete organs (Figure 1.2) instead of simple masses of epithelial cells. DeNardo et al. (2011), show that the amount of macrophages, one kind of immune cell, in the

tumor tissue was correlated with clinical outcomes in breast cancer. Considering the expensive manual microdissection of different cell types, it is important to develop *in silico* modeling methods to accurately measure the composition of the tumor. The mathematical process of identifying and separating constituent components is called *deconvolution*. One recent deconvolution method developed by Quon et al. (2013) shows significant improvements in prognostic prediction and other clinical variables for lung and prostate cancer after using deconvolved tumor gene expression profiles. The promising outcomes of these computational methods motivate more and more researchers to develop effective tumor deconvolution methods to address the intratumor heterogeneity issue, including the work introduced in this dissertation.

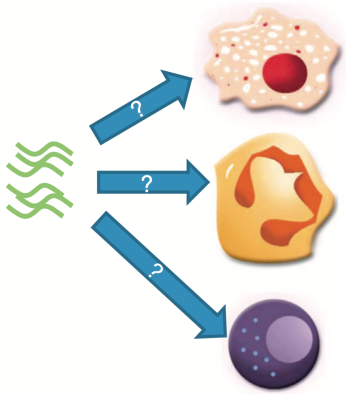


Figure 1.1: Tumor sample transciptome can be heterogeneous due to an underlying mixture of cell types. Figure from Shen-Orr and Gaujoux (2013)

1.2 Objectives and Problem Statement

Solid tumor tissues have demonstrated extensive genetic heterogeneity even within individual samples, which is often referred as intratumor heterogeneity. Tumor samples collected by physicians may be a mixture of cancer subtypes and/or normal cells such as stromal, immune and blood cells. It is usually of great interest for both clinicians and genetic researchers to obtain an explicit expression value for each of the sub-components in the mixture, as well as their corresponding proportions.

Statistically, if we model each of these mixture components' genetic expression levels as independent random variables, the observed genomic profile for a mixed tumor sample can be formulated as the weighted sum of these variables.

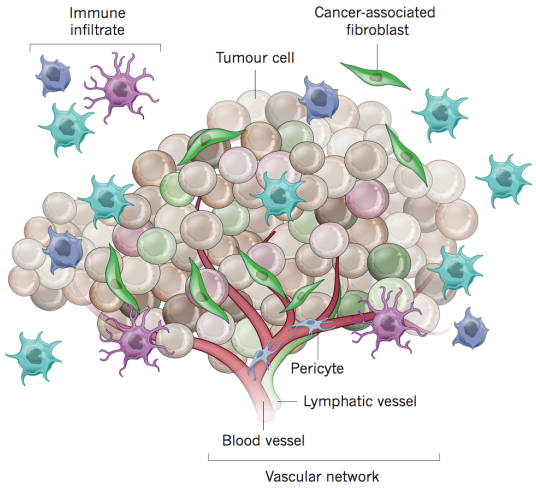


Figure 1.2: Picturing tumor tissue as an organ to demonstrate tumor heterogeneity. Tumor formation involves the co-evolution of tumor cells, stromal cells, immune cells, extracellular matrix and vascular network. Figure from Junttila and de Sauvage (2013)

Let the *observed genomic profiles* for S, the number of tumor samples, and G, the number of genes be denoted as the matrix $\mathbf{Y} \in \mathbb{R}^{G \times S}$. Each matrix element $y_{gi} \in \mathbb{R}$ is the observed expression value of gene $g \in \{1, \ldots, G\}$ from sample $i \in \{1, \ldots, S\}$. We assume the entries of $\mathbf{Y}$ are gene expression values after using some standard preprocessing procedures. Let K denote the number of subtypes present in the mixture sample. *Subtype matrix* $\mathbf{X} \in \mathbb{R}^{G \times K}$ contain the subtype mean specific expression values for G genes in each column. Then the *proportion matrix* is defined as $\mathbf{\Theta} \in \mathbb{R}^{K \times S}$, whose i^{th} column contains the proportions of each subtype in sample i. For each i, the column values should be non-negative and sum up to one. Lastly, the matrix $\mathbf{E}$ will store all the error terms. We assume the convolution of cell type expression values takes place in linear space. Hence the observed genomic profiles $\mathbf{Y}$ can be formulated as the multiplication of $\mathbf{X}$ and $\mathbf{\Theta}$ plus the error term:

$$\mathbf{Y} = \mathbf{X}\mathbf{\Theta} + \mathbf{E}$$

We emphasize the linearity assumption of the model, which means each subpopulation's contribution to the observed expression value of a gene is linearly proportional to the abundance of that subpopulation in the mixture. This is different from the previous common practice of modeling convolution after applying the log2-transformation to the expression data. Zhong and Liu (2012)

point out the convolution of tumor genomic profiles should be modeled in the linear space. They proved that if the log-transformed data is used as input, the output subtype specific expression profile $\mathbf{X}$ will be underestimated.

Provided with the genomic information of the mixed tumor samples, i.e. the matrix $\mathbf{Y}$, our target is to efficiently and accurately estimate the elements of $\mathbf{\Theta}$, as well as the tumor-specific expression pattern in $\mathbf{X}$.

1.3 Methods Overview

Many algorithms and methods have been developed to tackle the intratumor heterogeneity issue. As noted in Carroll et al. (2006), many review papers have summarized current deconvolution methods from different perspectives. Mohammadi et al. (2017) emphasize the choices of loss function, optimization constraints and regularization techniques involved in different methods. Shen-Orr and Gaujoux (2013) divide the computational methodologies into five classes, according to what input data is required for the method and the resolution of the output data, i,e. output $\mathbf{X}$ or $\mathbf{\Theta}$, or both. Wang et al. (2016a) summarize tumor purity estimation tools from the perspective of the data platform on which the model is built: based on genomic or epigenetic data. In this dissertation, we group current methods in three different ways.

The first grouping perspective is *regression-based* frameworks versus *probabilistic* models, Mohammadi et al. (2017). A regression-based model often involves techniques such as ordinary least squares (OLS), as proposed in our method, or quadratic programming (QP) frameworks as demonstrated in Zhong et al. (2013), and a series of other advanced optimization algorithms. For example, given the reference cell type expression profile, for a single gene g, the problem can be formatted as:

$$\underset{\mathbf{x}_g, \boldsymbol{\theta}_i \in \mathbb{R}^K}{\operatorname{argmin}} \sum_{i=1}^{S} \mathcal{L}(y_{gi} - \mathbf{x}_g^T \boldsymbol{\theta}_i) + \lambda_1 \mathcal{R}_1(\boldsymbol{\theta}_i) + \lambda_2 \mathcal{R}_2(\mathbf{x}_g)$$

where $\mathcal{L}$ and R_1, R_2 are loss function and regularization functions respectively.

On the other hand, the probabilistic models often have distributional assumptions for matrices $\mathbf{Y}, \mathbf{X\Theta}$ or $\mathbf{E}$, such as:

$$x_g \sim \mathcal{LN}(\mu_g, \sigma_g^2)$$

$$\epsilon_{gi} \sim \mathcal{N}(0, \sigma_g^2)$$

$$y_{gi} \sim \text{multinomial distribution}$$

$$\boldsymbol{\theta}_i \sim \text{Dirichlet}(\boldsymbol{\alpha}) \quad \text{where} \quad \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K).$$

This framework usually involves optimizing a likelihood function by Bayesian approach. For example, methods such as DSection (Erkkilä et al., 2010), GLAD (Saddiki et al., 2015), ISOLATE (Quon and Morris, 2009) and ISOpure (Quon et al., 2013) are developed based on the Latent Dirichlet Allocation work developed by Blei et al. (2003). In particular, the proportion coefficients use the Dirichlet distribution as a prior, for its convenient properties such as being conjugate to the multinomial distribution, in the exponential family. One problem with methods of this kind is the intractability of the posterior distribution. Approximation algorithms such as Laplace approximation, variational EM and MCMC are often used to conduct inference for the hidden variables. However, the computation time might suffer greatly from this type of approximation. Most of these papers do not report the computation time, but this can be a big drawback for methods of this type. For example, both GLAD and the most updated version of DeMixT (Wang et al., 2017), took several hours to process 100 samples, not including the pre-processing time. In contrast, our method, which will be discussed in detail in Chapter 2, uses less than one second processing time.

The second categorization of existing methods is *supervised* deconvolution versus *unsupervised*. In the field of signal processing, this categorization is often named as semi-/guided BSS versus Blind Source Separation (BSS) in Hesse and James (2006). BSS refers to the process of separating source signals from a linear combination of these signals, either without the aid of information about the source signals or using some prior knowledge. Various supervised/semi-BSS approaches may require different types of prior information, e.g.,

(a) Some approaches depend on having expression data from a purified reference sample for each cell or tissue type, i.e. $\mathbf{X}$ is provided. Abbas et al. (2009) developed one of the early regression based methods, assuming a pure expression matrix of each of the different cell types is provided.

(b) Some require cell-type specific marker genes, which are highly expressed in one cell type but not expressed in other cell types, to be available, such as the Digital Sorting Algorithm proposed in Zhong et al. (2013).

(c) Some other approaches assume availability of the proportions of each sample or cell type, i.e. the matrix $\mathbf{\Theta}$. These measures are usually evaluated by pathologists as in Stuart et al. (2004).

For most of the supervised methods, the number of sub components K is also assumed to be known.

Unsupervised approaches usually involve clustering, Independent Component Analysis (ICA), Principle Component Analysis (PCA), or Nonnegative Matrix Factorization (NMF). Among these tools, NMF is considered to be especially suitable for biological data, as it constrains all input and output data to be positive. A study in Moffitt et al. (2015) uses unsupervised NMF to identify exemplar genes, which are defined as genes with distinctly large weights in a single column of the matrix X. They demonstrate that these exemplar genes have subtype related meanings which can be confirmed from other studies. Moreover, combining these techniques with prior information has demonstrated better deconvolution performance. In Gaujoux and Seoighe (2012), where prior knowledge in the form of a set of marker genes is used, the accuracy of gene expression deconvolution is improved with the marker-guided NMF algorithm. Another rising class of unsupervised methods is called Convex Analysis of Mixtures (CAM), which utilizes a geometric approach to identify the vertices (marker genes) of the most tightly fitting scatter simplex Wang et al. (2016b), that encloses the observed data points. When the algorithm obtains the best simplex (convex hull) to fit the data, it is supposed to simultaneously obtain the vertices corresponding to K subpopulation marker genes, and give rise to the mixing proportions for each subpopulation. The highlight of this method is that it has no requirement on prior information such as the number, identity, or composition of the subpopulations present in mixed samples. However, some basic assumptions, such as a good uniform sampling over the space spanned by the subtype profile matrices and the intrinsic mixture diversity, still need to hold to ensure the model being identifiable.

The third categorization of current methods is based on the source of data: genetic data (such as DNA copy number, methylation data) versus transcriptome expression data (such as microarray, RNAseq data). For a common set of mixed tumor samples, one would expect the tumor purity estimates obtained from these two sources to be similar to each other. However, based on our

results of comparing estimates from a DNA copy number based method, and estimates from a transcriptome data based method, we did not observe a relatively high concordance of estimates (Table 2.4). Similar results were also reported in Wang and Wang (2015), where the highest correlation was observed as 0.57 in ovarian cancer.

The deconvolution method we propose here is a transcript based semi-BSS method, which assumes some reference expression matrix to be available, whose output includes both the coefficient matrix $\mathbf{\Theta}$ and the subcomponent specific expression profile matrix $\mathbf{X}$. With a better understanding of the intratumor heterogeneity issue and the development of deconvolution methodologies, we foresee the incorporation of the deconvolution step as an important part of genome profiling study pipelines in the future.

1.4 Dissertation Organization

To address the intratumor heterogeneity issue we start with a simple case in Chapter 2: assuming a tumor-normal two-component structure for the mixed tumor sample, i.e. a single cancer type and one type of normal tissue contamination. Details of developing and testing such two-component method are discussed in Chapter 2. We will extend these ideas and develop a multiple-component tumor deconvolution tool in Chapter 3.

At the end of this dissertation, we will discuss some potential work beyond the dissertation, such as applying the method to multiple platform data, and exploring the effect of data integration on tumor deconvolution.

CHAPTER 2
Two-component Source Separation Analysis

2.1 Introduction

Most deconvolution methods for identifying cell-type specific signals from heterogeneous tumor samples operate under a linearity assumption, Zhong and Liu (2012), in which the expression value of the mixture is a weighted sum of signals of its constituent cell types. Following this assumption, we propose a Fast Tumor Deconvolution (FasTD) pipeline, which deconvolves mixed genomic profiles into tumor/non-tumor profiles, as well as an estimation of the mixing proportions for each component. The method is a semi-Blind Source Separation (BSS) procedure as it requires reference gene expression profiles from non-tumor samples. But these reference samples do not have to be matched to the tumor samples. In addition, FasTD can be easily extended to more sub-components if the reference expression profiles for all but one subcomponents are available.

Another highlight of our method is that no distributional assumption is imposed on the expression subgroups, only mean and variance parameters are assumed. Our semi-parametric method has been tested and compared with other probabilistic methods in simulated data and real applications. The results show FasTD is both accurate and highly efficient, with biologically meaningful outputs for a TCGA prostate cancer case study.

As summarized in Figure 2.1, there are six key components of the FasTD pipeline. Firstly, the expression matrix $\mathbf{Y}$ of the mixing samples should be provided. Reference profiles for one of the constituent component in the mixtures are also given (Section 2.2.1). Secondly, some simple gene filtering is conducted to select genes differentiating the observed mixing expressions from the reference population. Procedures that remove outliers or influential points are recommended in this step. Then the core purity estimation part (Section 2.2.2) is the regression and the optimization step, which utilizes the first and the second moment information of the data sets, respectively.

After obtaining the purity estimates, individual deconvolution in Section 2.3 using our designed estimator is performed and concluded by the output of the tumor-specific expression matrix $\mathbf{T}$.

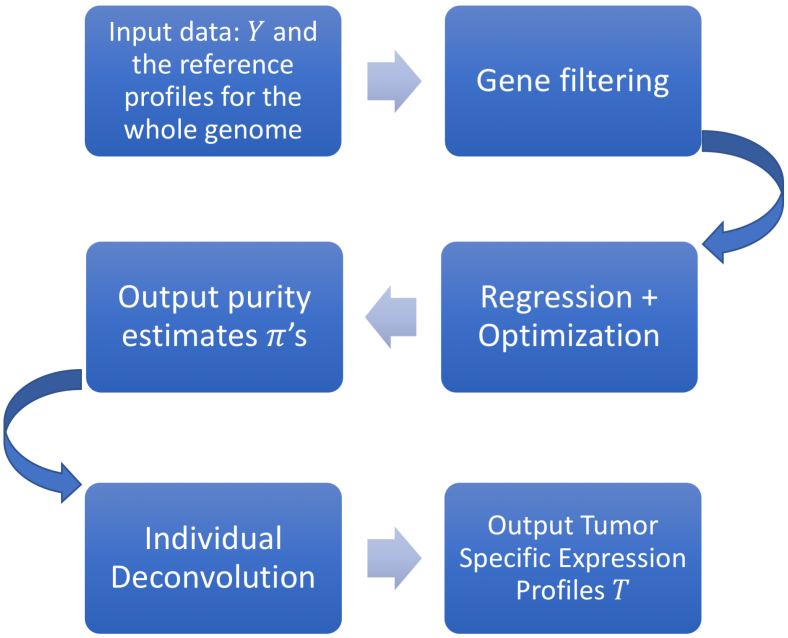


Figure 2.1: Flowchart of the FasTD pipeline.

2.2 Tumor Purity Proportion Estimation

We present a mixing proportion estimation procedure, which outputs the first set of estimates of our Fast Tumor Deconvolution pipeline 'FasTD'. For this procedure to work, we need to assume at most two major components are present in the tumor mixture tissue samples: a pure homogeneous tumor component and a homogeneous normal/non-tumor component. When normal control samples are available to us, we are able to set up a regression model, whose regression coefficients are directly related to the proportions of the tumor component in the mixture samples. Our goal is to fully recover these proportion coefficients which describe the purity of tumor samples. After running through the regression analyses and an optimization step, users can obtain the mixing proportion estimate for each tumor sample, as well as the mean and variance estimates for both components, within a short computational time. In Section 2.4.1, we will compare the simulation

results of our semi-parametric method with two other parametric methods to show the effectiveness of our method.

2.2.1 Problem Formulation and Model Assumptions

Let Y_{ig}, be the observed gene expression value for mixed tumor sample i, $i = 1\ , \ldots,\ S$, gene g, $g = 1, \ldots, G$. We assume each observed value Y_{ig} is a proportional sum of two quantities: the gene expression value for sample i, gene g from tumor cells, denoted as random variables T_{ig}, and that from normal cells, denoted as random variable N_{ig}. The proportional coefficient, denoted as $\pi_i \in [0, 1]$, captures the percentage of tumor cells in that sample:

$$Y_{ig} = \pi_i T_{ig} + (1 - \pi_i) N_{ig} \quad \forall i, g. \tag{2.1}$$

Random variables T_{ig} and N_{ig} are assumed to be independent of each other across samples and across genes. We also assume T_{ig} follows mean μ_{Tg} and variance σ^2_{Tg} and N_{ig} follows mean μ_{Ng} and variance σ^2_{Ng}. Note that there is no assumption on the distribution of these variables:

$$\begin{cases} T_{ig} \sim (\mu_{Tg}, \sigma^2_{Tg}) \\ N_{ig} \sim (\mu_{Ng}, \sigma^2_{Ng}) \end{cases}$$

Another important assumption is the gene expression profiles for N_0 number of normal control samples are available to us. Hence the parameters μ_{Ng} and σ^2_{Ng} assumed for N_{ig} can be easily estimated. Based on these assumptions, π_i, μ_{Tg}, σ^2_{Tg} for S samples and G genes are quantities remained to be estimated.

2.2.2 Purity Estimation using Moments Information

To show that a regression model can be derived based on assumptions in subsection 2.2.1, we will rearrange the problem formulated in (2.1). Let $\boldsymbol{\mu}_T \in \mathbb{R}^G$ and $\boldsymbol{\Sigma}_T \in \mathbb{R}^{G \times G}$ denote the mean vector and covariance matrix of gene expression for G genes in tumor cells, and $\boldsymbol{\mu}_N \in \mathbb{R}^G$ and $\boldsymbol{\Sigma}_N \in \mathbb{R}^{G \times G}$ for G genes in normal cells. Then the observed gene expression vector $\mathbf{y}_i \in \mathbb{R}^G$ for

mixed tumor sample i is the proportional sum of $\boldsymbol{\mu}_T$ and $\boldsymbol{\mu}_N$, plus some noise $\boldsymbol{\epsilon}_i$:

$$\mathbf{y}_i = \pi_i \boldsymbol{\mu}_T + (1 - \pi_i)\boldsymbol{\mu}_N + \boldsymbol{\epsilon}_i \tag{2.2}$$

where

$$\boldsymbol{\epsilon}_i \sim \Big(\mathbf{0}, \quad \pi_i^2 \boldsymbol{\Sigma}_T + (1 - \pi_i)^2 \boldsymbol{\Sigma}_N\Big).$$

Subtracting $\boldsymbol{\mu}_N$ from both sides of (2.2) we get:

$$\mathbf{y}_i - \boldsymbol{\mu}_N = \pi_i(\boldsymbol{\mu}_T - \boldsymbol{\mu}_N) + \boldsymbol{\epsilon}_i. \tag{2.3}$$

If $\boldsymbol{\mu}_N$ can be estimated from N_0 normal control samples and $\boldsymbol{\mu}_T$ is known, the expression in (2.3) very much resembles a simple linear regression without intercept, with π_i being the coefficient, $(\boldsymbol{\mu}_T - \boldsymbol{\mu}_N)$ the data matrix and $(\mathbf{y}_i - \boldsymbol{\mu}_N)$ the response. When $\boldsymbol{\mu}_T$ is unknown, we present a method to estimate the quantity $(\boldsymbol{\mu}_T - \boldsymbol{\mu}_N)$ in Step 3 of this subsection, hence estimating π_i by setting up a regression analysis.

The other issue that needs to be addressed is the selection of input genes. If we include all G genes whose expression values are available, some genes with the same expression levels in normal cells and tumor cells, i.e., $\mu_{Tg} = \mu_{Ng}$, will not be helpful in estimating π_i but only introducing noise to the regression model (2.3). Therefore, the ideal case is to select a set of feature genes, which is often referred as Differentially Expressed (DE) genes (in tumor cells versus non-tumor cells), as our input genes. For now, we simply use a two sample t-test to detect whether a gene is differentially expressed or not.

- **Step 1. Select Feature Genes**

  Given the observed gene expression profiles for tumor mixture samples and normal samples, we would like to select $\tilde{G}$ feature genes/DE genes that best separate the tumor mixture group and normal group, by evaluating each gene's t-statistics. For example, a threshold of ≥ 5 for the absolute t-statistic value can be used. Five is chosen based on the critical value corresponding to a 0.05 significant p-value adjusted for 20,000 genes of multiple t-tests.

- **Step 2. Estimate $\boldsymbol{\mu}_N$ and Compute Gene Weighting Matrix K**

The mean vector for normal gene expression $\boldsymbol{\mu}_N$ is estimated from N_0 normal control samples and then plugged into (2.3):

$$\hat{\boldsymbol{\mu}}_N = \frac{1}{N_0}\sum_{i=1}^{N_0}\mathbf{N}_i$$

$$\mathbf{y}_i - \hat{\boldsymbol{\mu}}_N = \pi_i(\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}_N) + (1-\pi_i)(\boldsymbol{\mu}_N - \hat{\boldsymbol{\mu}}_N) + \boldsymbol{\epsilon}_i. \tag{2.4}$$

To prevent genes with high expression values dominating the analysis, we will normalize genes across subjects. Each selected gene is treated equally by the following normalization. We introduce the gene weighting diagonal matrix $\mathbf{K}$, with positive elements on the diagonal. To start with, one can set the weights of each gene as the reciprocal of the sample standard deviation computed from $\mathbf{Y}_{S\times\tilde{G}}$:

$$\hat{\sigma}_g^2 = \frac{1}{(S-1)}\sum_{i=1}^{S}(Y_{ig} - \bar{Y}_{\cdot g})^2 \quad \forall g, \qquad \mathbf{K} = \begin{pmatrix} \frac{1}{\hat{\sigma}_1} & & \\ & \ddots & \\ & & \frac{1}{\hat{\sigma}_{\tilde{G}}} \end{pmatrix}_{\tilde{G}\times\tilde{G}}. \tag{2.5}$$

- **Step 3. Regression Analysis**

Equation (2.4) is still an under-determined system as $\boldsymbol{\mu}_T$ is unknown, but we can estimate $\mathbf{K}(\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}_N)$ for a corresponding π_i by method of moments.

$$\begin{aligned} &\frac{1}{S}\sum_{i=1}^{S}\mathbf{K}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_N) \\ &= \frac{1}{S}\sum_{i=1}^{S}\pi_i\cdot\mathbf{K}(\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}_N) + \frac{1}{S}\sum_{i=1}^{S}(1-\pi_i)\cdot\mathbf{K}(\boldsymbol{\mu}_N - \hat{\boldsymbol{\mu}}_N) + \frac{1}{S}\sum_{i=1}^{S}\mathbf{K}\boldsymbol{\epsilon}_i. \end{aligned} \tag{2.6}$$

Letting

$$\bar{\mathbf{y}} = \frac{1}{S}\sum_{i=1}^{S}\mathbf{y}_i, \quad \bar{\pi} = \frac{1}{S}\sum_{i=1}^{S}\pi_i, \quad \bar{\boldsymbol{\epsilon}} = \frac{1}{S}\sum_{i=1}^{S}\boldsymbol{\epsilon}_i$$

(2.6) can be rewritten as:

$$\mathbf{K}(\bar{\mathbf{y}} - \hat{\boldsymbol{\mu}}_N) = \bar{\pi}\cdot\mathbf{K}(\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}_N) + (1-\bar{\pi})\cdot\mathbf{K}(\boldsymbol{\mu}_N - \hat{\boldsymbol{\mu}}_N) + \mathbf{K}\bar{\boldsymbol{\epsilon}}. \tag{2.7}$$

$\boldsymbol{\epsilon}_i$ was assumed to have mean $\mathbf{0}$, $\hat{\boldsymbol{\mu}}_N$ with mean $\boldsymbol{\mu}_N$. By law of large number, when S and N_0 are large enough, term $\bar{\boldsymbol{\epsilon}}$ and $(\boldsymbol{\mu}_N - \hat{\boldsymbol{\mu}}_N)$ will go to $\mathbf{0}$ with probability 1. Hence the left hand side of (2.7) can be approximated by $\bar{\pi} \cdot \mathbf{K}(\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}_N)$. Then an estimate of $\mathbf{K}(\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}_N)$ is obtained by:

$$\widehat{\mathbf{K}(\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}_N)} = \frac{\mathbf{K}(\bar{\mathbf{y}} - \hat{\boldsymbol{\mu}}_N)}{\bar{\pi}} + \mathcal{O}(\frac{1}{\sqrt{N_0}}) + \mathcal{O}(\frac{1}{\sqrt{S}}). \tag{2.8}$$

From (2.4) we know for sample i it satisfies:

$$\mathbf{K}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_N) = \pi_i \cdot \mathbf{K}(\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}_N) + (1 - \pi_i)\mathbf{K}(\boldsymbol{\mu}_N - \hat{\boldsymbol{\mu}}_N) + \mathbf{K}\boldsymbol{\epsilon}_i$$

Letting

$$\mathbf{K}\boldsymbol{\delta}_i = (1 - \pi_i)\mathbf{K}(\boldsymbol{\mu}_N - \hat{\boldsymbol{\mu}}_N) + \mathbf{K}\boldsymbol{\epsilon}_i$$

be the new error term, whose expectation is still $\mathbf{0}$.

Then the expression in (2.8) is the key to remove the unknown $\boldsymbol{\mu}_T$ in the regression set-up:

$$\begin{aligned}
\mathbf{K}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_N) &= \pi_i \cdot \mathbf{K}(\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}_N) + \mathbf{K}\boldsymbol{\delta}_i \\
&= \pi_i \cdot \widehat{\mathbf{K}(\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}_N)} + \mathbf{K}\boldsymbol{\delta}_i \\
\mathbf{K}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_N) &= \frac{\pi_i}{\bar{\pi}} \cdot \mathbf{K}(\bar{\mathbf{y}} - \hat{\boldsymbol{\mu}}_N) + \mathbf{K}\boldsymbol{\delta}_i + \mathcal{O}(\frac{1}{\sqrt{N_0}}) + \mathcal{O}(\frac{1}{\sqrt{S}}).
\end{aligned} \tag{2.9}$$

Based on the relations in (2.9), one can obtain the least square estimator of $\beta_i = \frac{\pi_i}{\bar{\pi}}$ by:

$$\hat{\beta}_i = \Big(\widehat{\frac{\pi_i}{\bar{\pi}}}\Big)_{LS} = \max\Big(\frac{(\bar{\mathbf{y}} - \hat{\boldsymbol{\mu}}_N)^T\mathbf{K}^2(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_N)}{(\bar{\mathbf{y}} - \hat{\boldsymbol{\mu}}_N)^T\mathbf{K}^2(\bar{\mathbf{y}} - \hat{\boldsymbol{\mu}}_N)}, \quad 0\Big). \tag{2.10}$$

One would expect a negative slope, i.e. $\hat{\beta}_i$, if there are a lot of noise present in the data. But realistically a negative scaled version of π_i is not meaningful. Therefore, we put a non-negative constraint for β_i in (2.10).

- **Step 4. Second Moment Optimization to Fully Recover π_i**

  So far, we have obtained $\hat{\beta}_i$'s, which are the scaled values of $\hat{\pi}_i$'s. But the proportions are not yet fully recovered as $\bar{\pi}$ is unknown. The remaining question is how to provide a good estimation for $\bar{\pi}$.

Continually motivated by the method of moment, we decide to utilize the second moment information from observations Y_{ig}. According to our model's parameter assumption (2.3), the error term $\boldsymbol{\epsilon}_i$ follows:

$$\boldsymbol{\epsilon}_i = \mathbf{y}_i - \boldsymbol{\mu}_N - \pi_i(\boldsymbol{\mu}_T - \boldsymbol{\mu}_N) \sim \Big(\mathbf{0}, \quad \pi_i^2\boldsymbol{\Sigma}_T + (1-\pi_i)^2\boldsymbol{\Sigma}_N\Big) \quad \forall i.$$

After estimating $\hat{\boldsymbol{\mu}}_N$ and introducing the weighting matrix $\mathbf{K}$, the error term follows:

$$\begin{aligned}\mathbf{K}\boldsymbol{\epsilon}_i &= \mathbf{K}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_N) - \pi_i\mathbf{K}(\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}_N) - (1-\pi_i)\mathbf{K}(\boldsymbol{\mu}_N - \hat{\boldsymbol{\mu}}_N) \\ &\sim \Big(\mathbf{0}, \quad \pi_i^2\mathbf{K}\boldsymbol{\Sigma}_T\mathbf{K}' + (1-\pi_i)^2\mathbf{K}\boldsymbol{\Sigma}_N\mathbf{K}'\Big) \quad \forall i.\end{aligned}$$

Set the second moment information of $\mathbf{K}\boldsymbol{\epsilon}_i$ as $\mathbf{Z}_i = \mathbf{K}\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i'\mathbf{K}'$, then the mean of $\mathbf{Z}_i$ becomes:

$$\mathbb{E}(\mathbf{Z}_i) = \pi_i^2\mathbf{K}\boldsymbol{\Sigma}_T\mathbf{K}' + (1-\pi_i)^2\mathbf{K}\boldsymbol{\Sigma}_N\mathbf{K}'. \tag{2.11}$$

The exact values of $\mathbf{Z}_i$ are not observed. But we have S approximated independent observations of $\mathbf{K}\boldsymbol{\epsilon}_i$'s, according to (2.9):

$$\begin{aligned}\mathbf{Z}_i &= (\mathbf{K}\boldsymbol{\epsilon}_i)^{\otimes 2} \\ &= \Big(\mathbf{K}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_N) - \pi_i\mathbf{K}(\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}_N) - (1-\pi_i)\mathbf{K}(\boldsymbol{\mu}_N - \hat{\boldsymbol{\mu}}_N)\Big)^{\otimes 2} \\ &\approx \Big(\mathbf{K}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_N) - \pi_i\mathbf{K}(\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}_N)\Big)^{\otimes 2} \\ &= \Big(\mathbf{K}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_N) - \hat{\beta}_i \cdot \mathbf{K}(\bar{\mathbf{y}} - \hat{\boldsymbol{\mu}}_N) + \mathcal{O}(\frac{1}{\sqrt{N_0}}) + \mathcal{O}(\frac{1}{\sqrt{S}})\Big)^{\otimes 2} \\ &i = 1, \ldots, S.\end{aligned} \tag{2.12}$$

for any matrix X, $X^{\otimes 2} = XX^T$.

We introduce an objective function, which is to minimize the sum of squared deviation of each observation from the mean:

$$\begin{aligned}&\sum_{i=1}^{S} \|\text{diag}(\mathbf{Z}_i - \mathbb{E}(\mathbf{Z}_i))\|_2^2 \\ =&\sum_{i=1}^{S} \left\|\text{diag}(\left(\mathbf{K}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_N) - \hat{\beta}_i \cdot \mathbf{K}(\bar{\mathbf{y}} - \hat{\boldsymbol{\mu}}_N)\right)^{\otimes 2} - \pi_i^2 \mathbf{K}\boldsymbol{\Sigma}_T\mathbf{K}' - (1-\pi_i)^2\mathbf{K}\boldsymbol{\Sigma}_N\mathbf{K}')\right\|_2^2.\end{aligned} \tag{2.13}$$

Note that the diagonal elements of $\boldsymbol{\Sigma}_T$ is unknown but can be estimated simultaneously by solving the optimization problem. The diagonal elements of $\boldsymbol{\Sigma}_N$ can be estimated by the normal control sample variance $\frac{1}{(N_0-1)}\sum_{i=1}^{N_0}(\mathbf{N}_i - \hat{\boldsymbol{\mu}}_N)^2$. In addition, we can utilize the information obtained earlier in (2.10) from the least square estimator, to replace $\hat{\pi}_i$ with $\hat{\beta}_i\bar{\pi}$ due to the fact that $\hat{\beta}_i = \frac{\hat{\pi}_i}{\bar{\pi}}$.

Now plug into (2.13) all the estimated quantities we have so far, and formulate the optimization problem as:

$$\begin{aligned}\min_{\bar{\pi},\text{diag}(\boldsymbol{\Sigma}_T)} &\sum_{i=1}^{S} \left\|\text{diag}(\left(\mathbf{K}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_N) - \hat{\beta}_i \cdot \mathbf{K}(\bar{\mathbf{y}} - \hat{\boldsymbol{\mu}}_N)\right)^{\otimes 2}\right. \\ &\left. - (\hat{\beta}_i\bar{\pi})^2\mathbf{K}\boldsymbol{\Sigma}_T\mathbf{K}' - (1-\hat{\beta}_i\bar{\pi})^2\mathbf{K}\hat{\boldsymbol{\Sigma}}_N\mathbf{K}')\right\|_2^2 \\ s.t.:\quad &\min(\frac{1}{\max(\hat{\beta}_i)}, 1) \geq \bar{\pi} \geq 0 \\ &\text{diag}(\boldsymbol{\Sigma}_T) \geq \mathbf{0}.\end{aligned} \tag{2.14}$$

Problem (2.14) is a quadratic programming problem with one unknown scalar $\bar{\pi}$ and a vector of gene variances. All variables should be non-negative and the upper bound for $\bar{\pi}$ is restrained so that the final estimates for π_i's are no larger than 1.

Note that when $\bar{\pi}$ is given, (2.14) is reduced to:

for all $g = 1, \ldots, \tilde{G}$,

$$\begin{aligned}\min_{\sigma_{Tg}^2}\quad &\sum_{i=1}^{S} \left\|\left(K_g(y_{ig} - \hat{\mu}_{Ng}) - \hat{\beta}_i \cdot K_g(\bar{y}_{\cdot g} - \hat{\mu}_{Ng})\right)^2 - (\hat{\beta}_i\bar{\pi})^2 K_g^2\sigma_{Tg}^2 - (1-\hat{\beta}_i\bar{\pi})^2 K_g^2\hat{\sigma}_{Ng}^2\right\|_2^2 \\ s.t.:\quad &\sigma_{Tg}^2 \geq 0.\end{aligned} \tag{2.15}$$

where K_g is the g^{th} diagonal element of $\mathbf{K}$. The elements of diag($\mathbf{\Sigma}_T$) can be evaluated independently here because we have assumed the expression level for each gene is independent of each other. In addition, the optimal value of σ^2_{Tg} that minimizes (2.15) for gene g can be explicitly expressed and simplified as:

$$\hat{\sigma}^2_{Tg} = \max\left(\frac{\sum_{i=1}^{S}\left[\left((y_{ig}-\hat{\mu}_{Ng})-\hat{\beta}_i(\bar{y}_{\cdot g}-\hat{\mu}_{Ng})\right)^2-(1-\hat{\beta}_i\bar{\pi})^2\hat{\sigma}^2_{Ng}\right]\cdot\hat{\beta}_i^2}{\bar{\pi}^2\sum_{i=1}^{S}\hat{\beta}_i^4}, 0\right). \tag{2.16}$$

We see from (2.16) that the optimal solution $\hat{\sigma}^2_{Tg}$ is a function of $\bar{\pi}$. Therefore if we plug (2.16) back into (2.14), the original optimization problem is rewritten as below with only one scalar unknown $\bar{\pi}$:

$$\begin{aligned}
\min_{\bar{\pi}} \quad & \sum_{i=1}^{S}\sum_{g=1}^{\tilde{G}}\Bigg[\left(K_g(y_{ig}-\hat{\mu}_{Ng})-\hat{\beta}_i\cdot K_g(\bar{y}_{\cdot g}-\hat{\mu}_{Ng})\right)^2-(1-\hat{\beta}_i\bar{\pi})^2K_g^2\hat{\sigma}^2_{Ng} \\
& -\hat{\beta}_i^2K_g^2\frac{\sum_{i=1}^{S}\left[\left((y_{ig}-\hat{\mu}_{Ng})-\hat{\beta}_i(\bar{y}_{\cdot g}-\hat{\mu}_{Ng})\right)^2-(1-\hat{\beta}_i\bar{\pi})^2\hat{\sigma}^2_{Ng}\right]_+\cdot\hat{\beta}_i^2}{\sum_{i=1}^{S}\hat{\beta}_i^4}\Bigg]^2 \qquad (2.17) \\
\textit{s.t.:} \quad & \min\left(\frac{1}{\max(\hat{\beta}_i)}, 1\right)\geq\bar{\pi}\geq 0.
\end{aligned}$$

The optimal value of $\bar{\pi}$ in (2.17) is determined using the Matlab *fminunc* routine. By obtaining $\bar{\pi}$, we have fully recovered the mixing proportions for the tumor component in each mixture samples as $(\hat{\beta}_i\times\bar{\pi})$. The final step is to use this newly estimated $\hat{\pi}_i$'s to obtain the mean and variance estimates for the tumor component over the whole genome.

- **Step 5. Obtain Mean and Variance Estimates for All G Genes**

  Only a subset of feature genes, $\tilde{G}$ genes, were used from Step 2 to 4. With the purity estimates obtained in Step 4, we are able to provide mean and variance estimates for all G genes specific to the tumor component.

For gene g, $g = 1, \ldots, G$, treating mean μ_{Tg} in (2.3) as the only unknown, a least square estimate is obtained as:

$$\hat{\mu}_{Tg} = \max\big(\frac{\sum_{i=1}^{S}(Y_{ig} - \hat{\mu}_{Ng})}{\sum_{i=1}^{S}\hat{\pi}_i} + \hat{\mu}_{Ng}, 0\big). \tag{2.18}$$

Variance σ_{Tg}, for $g = 1, \ldots, G$, is obtained by the same expression as in (2.16), which is the optimal value that minimizes (2.15) when $\bar{\pi}$ is known.

2.3 Individual Deconvolution

Section 2.2 provides us with an efficient procedure to estimate the tumor proportion for each mixture sample, as well as the estimated mean and variance parameters for each gene. With these estimates, our next goal is to recover the real tumor expression value from the observed mixed expression value, for each gene in each mixture sample. This step is referred as individual deconvolution.

2.3.1 Estimators of the Explicit Tumor Expression Value

Using the same notations and assumptions as in section 2.2.1, we continue to assume the observed mixed sample expression value Y_{ig} is a proportional sum of two independent random variables:

$$Y_{ig} = \pi_i T_{ig} + (1 - \pi_i) N_{ig}.$$

We also assume at this step, the estimates of π_i and those of the mean and variance parameters, μ_{Tg} and σ^2_{Tg}, μ_{Ng} and σ^2_{Ng}, for $T_{\cdot g}$ and $N_{\cdot g}$ respectively, can be obtained from section 2.2. Hence these parameters are all treated as known quantities from now on. Constructing a good estimator for T_{ig} is our next target. To illustrate the main idea, we start with assuming T_{ig} and N_{ig} to be two normally distributed random variables.

$$\begin{cases} T_{ig} \sim \mathcal{N}(\mu_{Tg}, \sigma^2_{Tg}) \\ N_{ig} \sim \mathcal{N}(\mu_{Ng}, \sigma^2_{Ng}) \end{cases}$$

Then, the probability density function of Y_{ig} is the convolution of $f_{N_{ig}}$ and $f_{T_{ig}}$:

$$\begin{aligned} f_{Y_{ig}}(y_{ig}) &= \int_{-\infty}^{\infty} f_{N_{ig}}(\frac{y_{ig}}{1-\pi_i} - \frac{\pi_i}{1-\pi_i}t_{ig}) f_{T_{ig}}(t_{ig}) dt_{ig} \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_{Ng}} \exp\Big[- \frac{(\frac{y_{ig}}{1-\pi_i} - \frac{\pi_i}{1-\pi_i}t_{ig} - \mu_{Ng})^2}{2\sigma_{Ng}^2}\Big] \times \\ &\quad \frac{1}{\sqrt{2\pi}\sigma_{Tg}} \exp\Big[- \frac{(t_{ig}-\mu_{Tg})^2}{2\sigma_{Tg}^2}\Big] dt_{ig}. \end{aligned}$$

There exists a closed expression for the above integral, which shows Y_{ig} is also normally distributed:

$$f_{Y_{ig}}(y_{ig}) = \frac{1}{\sqrt{2\pi(\pi_i^2\sigma_{Tg}^2 + (1-\pi_i)^2\sigma_{Ng}^2)}} \exp\Big[- \frac{(y_{ig} - \pi_i\mu_{Tg} - (1-\pi_i)\mu_{Ng})^2}{2(\pi_i^2\sigma_{Tg}^2 + (1-\pi_i)^2\sigma_{Ng}^2)}\Big]$$

or

$$Y_{ig} \sim \mathcal{N}(\pi_i\mu_{Tg} + (1-\pi_i)\mu_{Ng}, \quad \pi_i^2\sigma_{Tg}^2 + (1-\pi_i)^2\sigma_{Ng}^2).$$

Next, we can obtain the conditional distribution of T_{ig}, given the observed data y_{ig}:

$$\begin{aligned} f(t_{ig}|y_{ig}) &= \frac{f(y_{ig}|t_{ig}) f_{T_{ig}}(t_{ig})}{f_{Y_{ig}}(y_{ig})} = \frac{f_{N_{ig}}(\frac{y_{ig}-\pi_i t_{ig}}{1-\pi_i}) f_{T_{ig}}(t_{ig})}{f_{Y_{ig}}(y_{ig})} \\ &\propto \phi\Big(\frac{y_{ig}-\pi_i t_{ig}}{1-\pi_i}\Big|\mu_{Ng}, \sigma_{Ng}^2\Big) \cdot \phi(t_{ig}|\mu_{Tg}, \sigma_{Tg}^2) \end{aligned} \tag{2.19}$$

where $\phi(\cdot|\mu, \sigma^2)$ is a normal density with mean μ and variance σ^2. The probability density function of T_{ig} given Y_{ig} is also following a normal distribution:

$$T_{ig}|Y_{ig} \sim \mathcal{N}(\pi_i\mu_{Tg} + (1-\pi_i)\mu_{Ng}, \quad \pi_i^2\sigma_{Tg}^2 + (1-\pi_i)^2\sigma_{Ng}^2).$$

- **The likelihood estimator, $\hat{T}_{ig}$**

  Since we have obtained the expression for the conditional density function of T_{ig} in (2.19), one strategy is to design a T_{ig} estimator that maximizes this probability:

$$\underset{t_{ig}}{\text{argmax}} \quad \phi(t_{ig}|\mu_{Tg}, \sigma_{Tg}^2) \cdot \phi\Big(\frac{y_{ig}-\pi_i t_{ig}}{1-\pi_i}\Big|\mu_{Ng}, \sigma_{Ng}^2\Big). \tag{2.20}$$

If we take the first derivative of the log likelihood function of (2.20) and set it to zero, we get a likelihood estimator, denoted as $\hat{T}_{ig}$:

$$\hat{T}_{ig} = \frac{(1-\pi_i)^2\sigma_{Ng}^2 \cdot \mu_{Tg} + \pi_i^2\sigma_{Tg}^2 \cdot \tilde{T}_{ig}}{(1-\pi_i)^2\sigma_{Ng}^2 + \pi_i^2\sigma_{Tg}^2} \tag{2.21}$$

where $\tilde{T}_{ig} = \frac{y_{ig}}{\pi_i} - \frac{1-\pi_i}{\pi_i}\mu_{Ng}$.

It is interesting to see from (2.21) that $\hat{T}_{ig}$ is a weighted sum of μ_{Tg} and $\tilde{T}_{ig}$. Both are two important quantities containing information about the true T_{ig}. Take a closer look at $\tilde{T}_{ig}$. Given the true value T_{ig},

$$\begin{aligned} \tilde{T}_{ig}|T_{ig} &= \frac{\pi_i T_{ig} + (1-\pi_i)N_{ig}}{\pi_i} - \frac{1-\pi_i}{\pi_i}\mu_{Ng} \\ &= T_{ig} + \frac{1-\pi_i}{\pi_i}(N_{ig} - \mu_{Ng}). \end{aligned}$$

Based on the independence and normality assumption for T_{ig} and N_{ig}, we get:

$$\tilde{T}_{ig}|T_{ig} \sim (T_{ig}, \quad (\frac{1-\pi_i}{\pi_i})^2\sigma_{Ng}^2). \tag{2.22}$$

or

$$\begin{aligned} \mathrm{E}(\tilde{T}_{ig}|T_{ig}) &= T_{ig} \\ \mathrm{Var}(\tilde{T}_{ig}|T_{ig}) &= (\frac{1-\pi_i}{\pi_i})^2\sigma_{Ng}^2 \end{aligned}$$

Thus, $\tilde{T}_{ig}$ is important because its conditional distribution centers around the true T_{ig}. On the other hand, μ_{Tg} is also important because it is the mean value of T_{ig} when T_{ig} is treated as a random variable.

Inspired by the expression of $\hat{T}_{ig}$, we propose a general format for estimators of T_{ig}, which is a weighted sum between μ_{Tg} and $\tilde{T}_{ig}$:

$$T_{ig}^G = \frac{a \cdot \mu_{Tg} + b \cdot \tilde{T}_{ig}}{a+b}. \tag{2.23}$$

We regard T_{ig}^G as an efficient way to combine all available information about T_{ig}. Next we investigate choice of the coefficients a and b.

- **The oracle estimator,** T_{ig}^*

  A useful criterion is the smallest conditional Mean Square Error (cMSE) given T_{ig}. The cMSE for the general form is written as:

$$MSE(T_{ig}^G|T_{ig}) = \left[\mathrm{E}(\frac{a \cdot \mu_{Tg} + b \cdot \tilde{T}_{ig}}{a+b}|T_{ig}) - T_{ig}\right]^2 + \mathrm{Var}(\frac{a \cdot \mu_{Tg} + b \cdot \tilde{T}_{ig}}{a+b}|T_{ig}).$$

  With the conditional mean and variance of $\tilde{T}_{ig}$ as shown in (2.22), the above cMSE is simplified as:

$$MSE(T_{ig}^G|T_{ig}) = \left[\frac{a \cdot (T_{ig} - \mu_{Tg})}{a+b}\right]^2 + \frac{b^2 \cdot (\frac{1-\pi_i}{\pi_i})^2 \sigma_{Ng}^2}{(a+b)^2}. \tag{2.24}$$

  To minimize (2.24), we will take the first derivative of the function with respect to a and b, and set them to zero. Then we can get the optimal solutions for a and b:

$$\begin{aligned} a^* &= (1-\pi_i)^2 \sigma_{Ng}^2 \\ b^* &= \pi_i^2 (T_{ig} - \mu_{Tg})^2. \end{aligned} \tag{2.25}$$

  Plugging (2.25) into the general form T_{ig}^G in (2.23), we get the optimal estimator T_{ig}^*. It is an estimator that minimizes the conditiontal MSE when T_{ig} is given:

$$T_{ig}^* = \frac{(1-\pi_i)^2 \sigma_{Ng}^2 \cdot \mu_{Tg} + \pi_i^2 (T_{ig} - \mu_{Tg})^2 \cdot \tilde{T}_{ig}}{(1-\pi_i)^2 \sigma_{Ng}^2 + \pi_i^2 (T_{ig} - \mu_{Tg})^2}. \tag{2.26}$$

  Or it can be written as:

$$T_{ig}^* = \frac{(\frac{1-\pi_i}{\pi_i} \cdot \frac{\sigma_{Ng}}{\sigma_{Tg}})^2 \cdot \mu_{Tg} + (\frac{T_{ig} - \mu_{Tg}}{\sigma_{Tg}})^2 \cdot \tilde{T}_{ig}}{(\frac{1-\pi_i}{\pi_i} \cdot \frac{\sigma_{Ng}}{\sigma_{Tg}})^2 + (\frac{T_{ig} - \mu_{Tg}}{\sigma_{Tg}})^2}. \tag{2.27}$$

  In real situations, we do not know T_{ig}, hence we can not obtain T_{ig}^*. But in simulation studies, (2.26) still provides a useful basis for comparison, as this is the lower bound on what we can

achieve in minimizing the cMSE. If we compare the expressions between $\hat{T}_{ig}$ and T^*_{ig}, we see that $\hat{T}_{ig}$ just uses σ^2_{Tg} to replace the unknown part of T^*_{ig}: $(T_{ig} - \mu_{Tg})^2$. This makes sense if T_{ig} is treated as a random variable, because σ^2_{Tg} is the expected value of $(T_{ig} - \mu_{Tg})^2$.

- **The plug-in estimator,** $\check{T}_{ig}$

  Motivated by the goal of minimizing the conditional MSE, and with the fact that the conditional mean of $\tilde{T}_{ig}$ is T_{ig}, we propose another estimator $\check{T}_{ig}$, which plugs $\tilde{T}_{ig}$ into the unknown part of T^*_{ig}:

$$\check{T}_{ig} = \frac{(1-\pi_i)^2\sigma^2_{Ng} \cdot \mu_{Tg} + \pi_i^2(\tilde{T}_{ig} - \mu_{Tg})^2 \cdot \tilde{T}_{ig}}{(1-\pi_i)^2\sigma^2_{Ng} + \pi_i^2(\tilde{T}_{ig} - \mu_{Tg})^2}. \tag{2.28}$$

  This estimator follows the general format we proposed earlier as a weighted sum of μ_{Tg} and $\tilde{T}_{ig}$. The only random variable present in this estimator is $\tilde{T}_{ig}$.

Next we will rearrange the estimators to show that the performance of these estimators depends on two important quantities. First we will divide both the numerator and denominator of T^*_{ig} and $\tilde{T}_{ig}$ by σ^2_{Tg}. Then let $X = \frac{T_{ig}-\mu_{Tg}}{\sigma_{Tg}}$, $\tilde{X} = \frac{\tilde{T}_{ig}-\mu_{Tg}}{\sigma_{Tg}}$, $A = \frac{1-\pi_i}{\pi_i}$, and $B = \frac{\sigma_{Ng}}{\sigma_{Tg}}$. Note that under the normality assumption, $X \sim \mathcal{N}(0,1)$, and the conditional distribution of $\tilde{X}|T_{ig} \sim \mathcal{N}(X, (AB)^2)$. Finally the estimators are rewritten as:

$$T^*_{ig} = \frac{\left(\frac{1-\pi_i}{\pi_i} \cdot \frac{\sigma_{Ng}}{\sigma_{Tg}}\right)^2 \mu_{Tg} + \left(\frac{T_{ig}-\mu_{Tg}}{\sigma_{Tg}}\right)^2 \tilde{T}_{ig}}{\left(\frac{1-\pi_i}{\pi_i} \cdot \frac{\sigma_{Ng}}{\sigma_{Tg}}\right)^2 + \left(\frac{T_{ig}-\mu_{Tg}}{\sigma_{Tg}}\right)^2} = \frac{(AB)^2\mu_{Tg} + X^2\tilde{T}_{ig}}{(AB)^2 + X^2}. \tag{2.29}$$

$$\hat{T}_{ig} = \frac{\left(\frac{1-\pi_i}{\pi_i} \cdot \frac{\sigma_{Ng}}{\sigma_{Tg}}\right)^2 \mu_{Tg} + 1 \cdot \tilde{T}_{ig}}{\left(\frac{1-\pi_i}{\pi_i} \cdot \frac{\sigma_{Ng}}{\sigma_{Tg}}\right)^2 + 1} = \frac{(AB)^2\mu_{Tg} + 1 \cdot \tilde{T}_{ig}}{(AB)^2 + 1}. \tag{2.30}$$

$$\check{T}_{ig} = \frac{\left(\frac{\sigma_{Ng}}{\sigma_{Tg}} \cdot \frac{1-\pi_i}{\pi_i}\right)^2 \mu_{Tg} + \left(\frac{\tilde{T}_{ig}-\mu_{Tg}}{\sigma_{Tg}}\right)^2 \tilde{T}_{ig}}{\left(\frac{1-\pi_i}{\pi_i} \cdot \frac{\sigma_{Ng}}{\sigma_{Tg}}\right)^2 + \left(\frac{\tilde{T}_{ig}-\mu_{Tg}}{\sigma_{Tg}}\right)^2} = \frac{(AB)^2\mu_{Tg} + \tilde{X}^2\tilde{T}_{ig}}{(AB)^2 + \tilde{X}^2}. \tag{2.31}$$

The rearrangement displays two sets of quantities: AB and X. Note that AB captures the odds ratio of tumor/non-tumor purity(A) and the variance(B). In our analysis, we will focus on samples which are not pure tumor or non-tumor tissues. This means the mixing proportions are

away from zero or one. Further more, we decide to focus on genes whose expression values are more heterogeneous in tumor tissue than in non-tumor tissues, i.e. $\sigma_{Tg} > \sigma_{Ng}$, which are more biologically interesting. Thus we expect to work on cases where AB values are small non-negative numbers.

The other important quantity X is unknown to us in real situations. However if T_{ig} follows the normal distribution, X will follow a standard normal distribution. Therefore, for our future analysis on the performance of these estimators, if we concentrate on AB and X values within certain range, i.e. [0, 3] and [-3, 3] respectively, the results should still be representative for a large number of the population. The performance of these estimators in certain ranges of AB and X will be shown in Section 2.4.2.

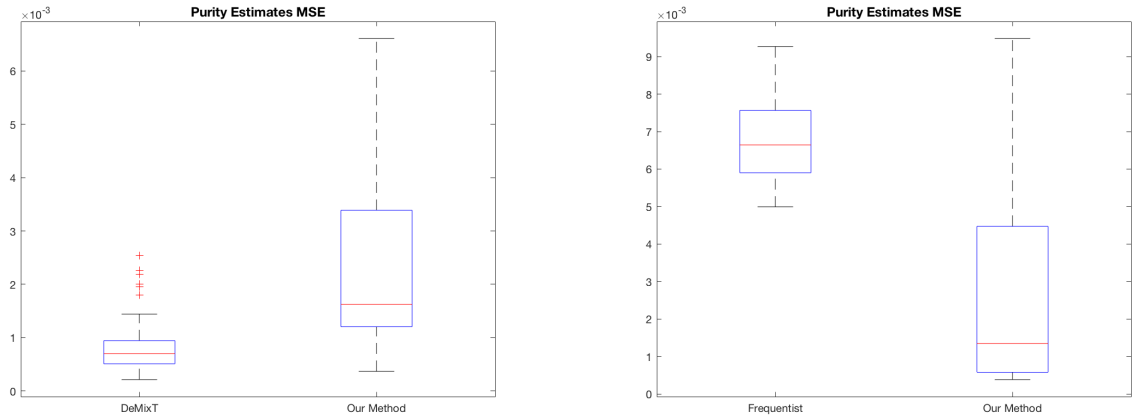
2.4 Simulation Study

2.4.1 Comparing Proportion Estimates with Two Parametric Models

The mixing proportion estimation is an important indicator of a successful deconvolution event. To validate our π_i estimates, we will test it with simulated mixture profiles and compare the results with two other parametric deconvolution methods: DeMixT (Wang et al., 2017) and a Frequentist method developed by Dr. Rongjie Liu in Dr. Zhu's group. Similar to our method, both of these methods have the tumor proportion estimation as their first step and assume the availability of normal control samples. We will particularly focus on comparison with DeMixT, which is a recently developed method based on DeMix published in Ahn et al. (2013). DeMixT expanded the deconvolution capacity to three components, with the requirement of pure expression profiles available for the other two non-tumor components. DeMixT assumes gene expression follows a log2-normal distribution. This means the observed data Y_{ig} is the convolution of the density function for two log2-normal distributions, under the same parameter formulation of the problem we presented in section 2.2.1. As a closed form of the complete likelihood function for Y_{ig} cannot be obtained, numerical integration is used to estimate the parameters. More specifically, Iterated Conditional Modes algorithm is used to search for the π_i that locally maximizes the joint probability conditioning on the rest of parameters. Whereas the Frequentist method assumes the observed data Y_{ig} is the

convolution of two Negative Binomial (NB) gene expression densities. Still a closed form of the density function of Y_{ig} is not available, hence the Monte Caro EM algorithm is used to search for the local optimal π_i that maximizes the complete likelihood. Both of these methods are developed based on raw counts data, instead of log2-transformed according to Zhong and Liu (2012). And both are parametric probabilistic models, which we discussed in Section 1.3 that are likely to suffer from approximation error and lengthy computation time.

To compare our semi-parametric method with these two parametric methods, we simulated two separate datasets of 500 genes and 100 sample observations, whose entries are the sum of either two log2 normal distributed random variables or two NB distributed random variables. Each dataset has 100 simulation replicates. The true tumor proportions π_i were designed to be uniformly distributed over the [0,1] interval. Performance of these methods are compared according to their Mean Square Errors (MSE). Results are shown in Figure 2.2.



(a) DeMixT vs. FasTD **(b)** Frequentist vs. FasTD

Figure 2.2: Purity estimation performance comparison between FasTD and other methods.

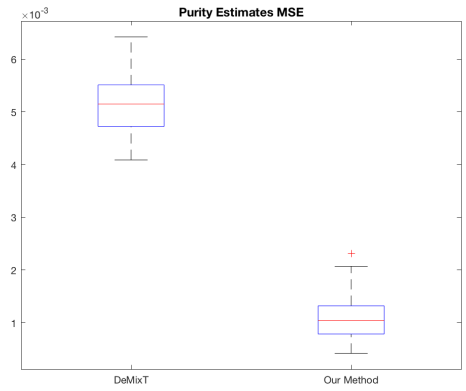
Input data size	Method	Platform	CPUtime (in sec)
500 x 100	DeMixT	R	3,047
	Frequentist	Python	72,000
	Our Method	Matlab	less than 1

Table 2.1: Running time comparison in seconds over 100 replications.

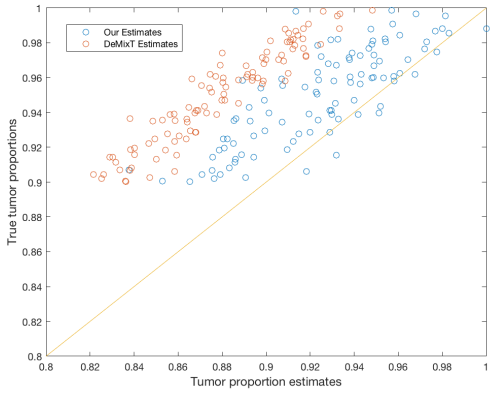
Recall that we simulated two datasets according to the distributional assumptions imposed by DeMixT and the Frequentist method. As shown in Figure 2.2, the DeMixT method slightly outperforms ours (at a scale of 10^{-3}), which is not surprised as the simulated dataset is generated exactly according to a log2 normal distribution. Our method uses the squared loss function during the π_i estimation, which is known to be an asymptotically optimal loss function when the underlying model is perturbed by Gaussian noise. However, as shown in Table 2.1, our method only takes 0.74 second per dataset to get the mixing proportions. This is more than 4,000 times faster than DeMixT (3047 seconds), with a negligible compensation of Mean Square Error (MSE). Comparing with the Frequentist method, our method performs better both in accuracy, as shown in Figure 2.2 (b), and in time, more than 72,000 times faster (Table 2.1). This can be explained by the approximation errors generated by the Monte Caro EM algorithm, when the complete joint likelihood function is complicated.

One limitation of the DeMixT method we observed when running the simulation study is that it restricts purity proportions to a range of [0.05, 0.95]. This means the DeMixT method will not be able to handle the cases when the tumor sample is pure tumor or pure non-tumor, while these samples are very likely to exist in real scenarios. To further investigate this issue and confirm our method does not have this constrained window for π, we simulated a group of datasets with tumor proportion ranges from 0.9 to 1 (since it is of more interest for researchers to investigate on tumor with high purity instead of almost-normal cases). As we can see from Figure 2.3 (a), for samples with high tumor proportions, our method performs much better than the DeMixT method. Figure 2.3 (b) is a scatter plot of two sets of 100 sample estimates generated by DeMixT (orange points) and by our method (blue points). For each sample point, the x-axis value is the estimated proportion and the y-axis value is the truth. The better the sample points aligning with the yellow y=x line, the better estimation results from that method. From Figure 2.3 (b) we see that estimates from the DeMixT are greatly biased, compared with ours in the cases when most samples have 0.9 to 1 tumor proportions.

Yet another limitation we observed about the DeMixT method, and possibly other methods which utilizes numerical integration algorithms, is the issue of approximation error. Since the complete likelihood for two convoluted log-2 normal random variables is not a closed form, the



(a) Purity estimation performance comparison between DeMixT and our method over the (0.9, 1) interval



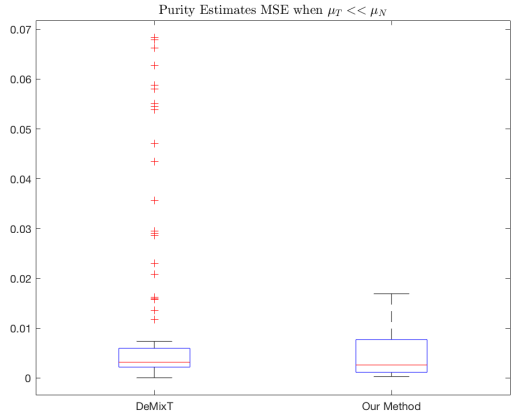
(b) Scatterplot of DeMixT Estimates/Our Estimates vs. Real Values for 100 Mixed Tumor Samples

Figure 2.3: Purity estimation performance comparison between DeMixT and our method, when true π_i's spread over [0.9, 1] interval. In this interval, the DeMixT method has poor performance and more bias.

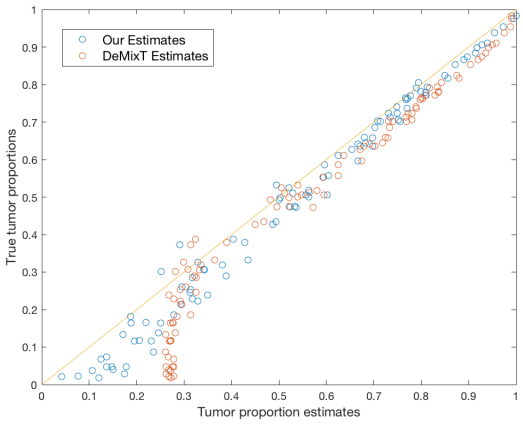
authors use numerical integration to approximate it. The problem is when the integrand function inside the complete likelihood function is steeply peaked, shown in Figure 2.5 (b), it needs a very large bin number to capture the peak. Increasing the bin number will either greatly prolong the estimation time, or still fail to return an accurate estimation when the peak is extremely steep. Because this case is frequently observed when the tumor mean is much smaller than normal mean, we simulated a convoluted dataset with two log2 normal random variables, whose mean parameters μ_N and μ_T is at least 4 units apart. As shown in Figure 2.4 (a), our method again outperforms DeMixT in the $\mu_{Ng} >> \mu_{Tg}$ scenario, with smaller and less variant MSE distribution for 100 simulation runs. Figure 2.4 (b) is an illustration of a single simulation result for 100 sample points. We observe that DeMixT (in orange) performs poorly especially when the true tumor purity is smaller than 0.3. The DeMixT estimates deviate from the y=x line, whose x value is the estimated value and y the true proportion. But this deviation is not observed in samples estimated by our method (blue in Figure 2.4 b).

Figure 2.5 (a) and (b) illustrate the idea why DeMixT works well in Figure 2.4 (b) when the true proportions are large. We plotted the integrand function in Figure 2.5 of the complete likelihood function for an observed $y_{ig} = 4 \times 10^4$, with $\mu_N = 15$, $\mu_{Tg} = 10$ and $\sigma_T = \sigma_N = 1$, with two different tumor proportions. Figure 2.5 (a) shows that a larger tumor purity, such as π=0.8, would mitigate the peaking effect, thus enable the numeric integration algorithm to capture the optimal

n. But the algorithm is likely to fail when π is as low as 0.2, for the integrand function in Figure 2.5 (b).

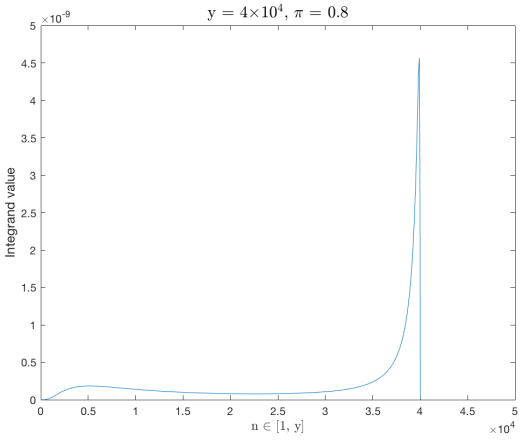


(a) Purity estimation performance comparison between DeMixT and our method when μ_{Ng} and μ_{Tg} is distantly apart

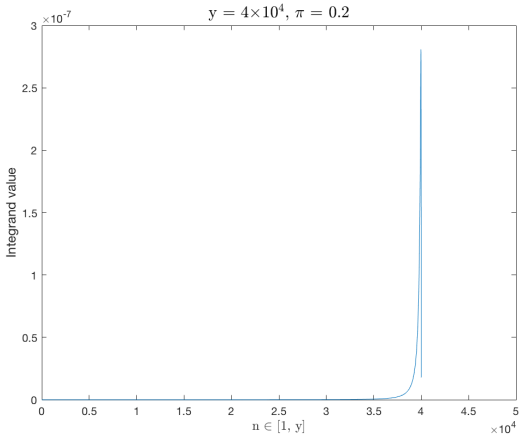


(b) Scatterplot of DeMixT Estimates/Our Estimates vs. Real Values for 100 Mixed Tumor Samples

Figure 2.4: Purity estimation performance comparison between DeMixT and our method, in cases when $\mu_{Ng} >> \mu_{Tg}$. Our outperforms DeMixT, especially when π_i is small



(a) Integrand function plot with a relatively large purity proportion



(b) Integrand function plot with a relatively small purity proportion

Figure 2.5: Illustration of the peaking effect for the likelihood functions that DeMixT trying to numerically approximate. When the observed y is large, and μ_N and μ_T are apart from each other, the integrand function tends to steeply peak within a small region. However, this effect can be alleviated by a relatively large π.

2.4.2 Performance Evaluation of Individual Deconvolution Estimator

Since our overall strategy is to construct estimators with small cMSE, it is desirable to have explicit expressions for all corresponding cMSEs. The closed cMSE expression for T^*_{ig} and $\hat{T}_{ig}$ can be derived as:

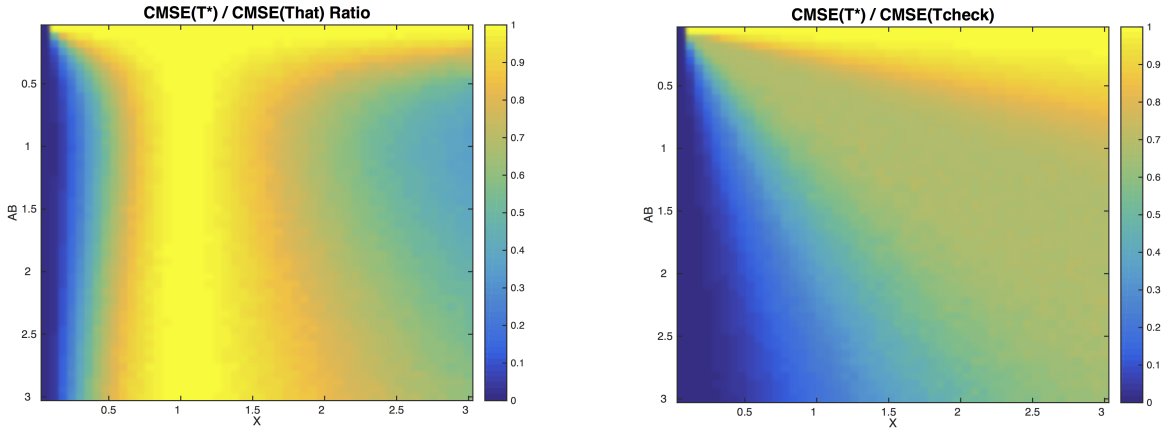
$$MSE(T^*_{ig}|T_{ig}) = \Big[\frac{(AB)^2X}{(AB)^2+X^2}\Big]^2 + \Big[\frac{(AB)X^2}{(AB)^2+X^2}\Big]^2 \tag{2.32}$$

$$MSE(\hat{T}_{ig}|T_{ig}) = \Big[\frac{(AB)^2X}{(AB)^2+1}\Big]^2 + \Big[\frac{(AB)}{(AB)^2+1}\Big]^2. \tag{2.33}$$

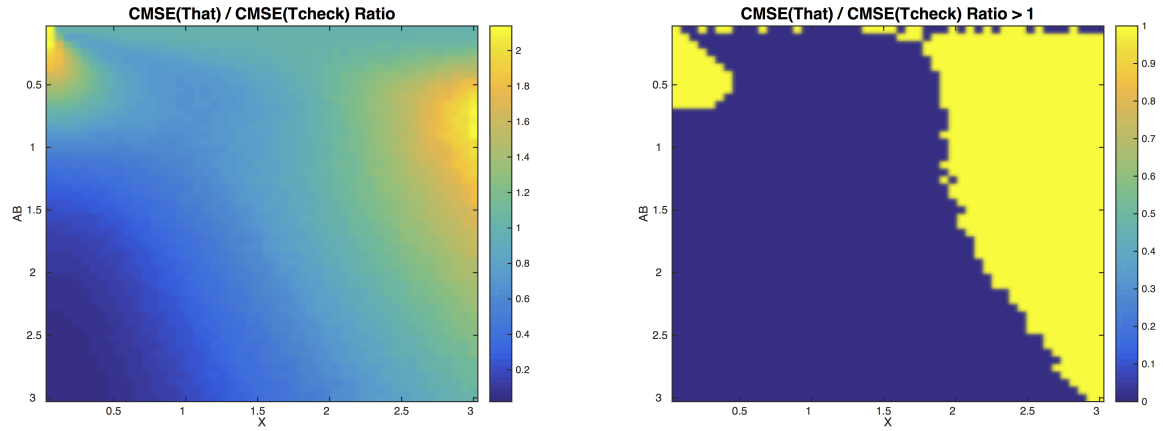
The difficult part is to have a closed form for the cMSE of $\check{T}_{ig}$. Hence we will run simulation studies to evaluate the estimators' performance. For all the simulations, we set $\mu_{Tg} = 1$ and $\sigma^2_{Tg} = 1$. Then for a fixed AB and X value, true T_{ig} is known and $\tilde{T}_{ig}$ can be simulated according to the normal distribution (2.22). 5000 runs of simulation were performed for each set of AB, X value. The cMSE of each estimator was then computed and compared with other estimators as shown in Figure 2.6.

Neither the ratio $\frac{cMSE(T^*_{ig})}{cMSE(\hat{T}_{ig})}$, nor $\frac{cMSE(T^*_{ig})}{cMSE(\check{T}_{ig})}$ exceeds the value 1, as shown in Figure 2.6(a) and (b) respectively, which demonstrates T^*_{ig} is the optimal estimator with the minimum cMSE. Around the region $X = 1$, the estimator $\hat{T}_{ig}$ performs as well as T^*_{ig}. This is as expected because when $X = 1$, $\hat{T}_{ig}$ is identical to T^*_{ig} (2.29 and 2.30). Then we see from Figure 2.6 (c) and (d) that $\check{T}_{ig}$ and $\hat{T}_{ig}$ outperform each other in different AB, X regions. $\check{T}_{ig}$ outperforms $\hat{T}_{ig}$ when both X and AB are small and X deviates from 1, and when X is larger, e.g. $X > 2$. However, we also notice from Figure 2.6 (b) and (c), that in general when AB is getting large (but not X), the performance of $\check{T}_{ig}$ is getting worse (i.e. getting bluer). This seems to be because:

$$\begin{aligned}\check{T}_{ig} &= \frac{\left(\frac{\sigma_{Ng}}{\sigma_{Tg}}\cdot\frac{1-\pi_i}{\pi_i}\right)^2\mu_{Tg} + \left(\frac{\tilde{T}_{ig}-\mu_{Tg}}{\sigma_{Tg}}\right)^2\tilde{T}_{ig}}{\left(\frac{1-\pi_i}{\pi_i}\cdot\frac{\sigma_{Ng}}{\sigma_{Tg}}\right)^2 + \left(\frac{\tilde{T}_{ig}-\mu_{Tg}}{\sigma_{Tg}}\right)^2} \\ &= \frac{(AB)^2\mu_{Tg} + \tilde{X}^2\tilde{T}_{ig}}{(AB)^2+\tilde{X}^2}\end{aligned}$$



(a) Performance of $\hat{T}_{ig}$ comparing with T^*_{ig} **(b)** Performance of $\check{T}_{ig}$ comparing with T^*_{ig}



(c) Performance of $\check{T}_{ig}$ comparing with $\hat{T}_{ig}$ **(d)** Performance of $\check{T}_{ig}$ comparing with $\hat{T}_{ig}$

Figure 2.6: Performance evaluation of estimators under different AB and X values. (a) and (b) confirms the optimal performance of T^*_{ig} and shows the regions where $\hat{T}_{ig}$ and $\check{T}_{ig}$ perform as well as T^*_{ig} (the yellow region). (c) and (d) compares the performance between $\check{T}_{ig}$ and $\hat{T}_{ig}$. Yellow region in (d) shows when cMSE($\check{T}_{ig}$) is smaller than cMSE($\hat{T}_{ig}$).

and $\tilde{X}|T_{ig} \sim \mathcal{N}(X, (AB)^2)$. When the conditional variance term AB gets larger, $\tilde{X}$ becomes more variant hence the estimator $\check{T}_{ig}$ is more unstable. Through this simulation study, we observe the large variance of the estimator $\check{T}_{ig}$. Therefore in real data applications, we recommend to use $\hat{T}_{ig}$ to estimate the tumor-specific expressions in the mixtures.

2.5 Real Data Applications

In this section, we apply the two-component deconvolution method FasTD to several real datasets. First a validation dataset with known mixing proportions is introduced in Section 2.5.1.

The FasTD estimates for the proportions are compared with the ground truths to demonstrate the effectiveness of our method. Then FasTD is applied to a larger cohort study, The Cancer Genome Atlas (TCGA) in Section 2.5.2. More than six thousand tumor purity estimates are obtained for the fourteen cancer types in this dataset. A case study of the TCGA prostate cancer in Section 2.5.3 demonstrates the value added from performing the deconvolution step before downstream genomic analysis.

2.5.1 Benchmark Dataset Validation

Dataset GSE33076 (Siegert et al., 2012) is designed to evaluate the linearity of amplification between gene expression values and the amounts of RNA in adult retina cells in mouse models. In the original experimental design, two distinct retina cell types: cone cells and starburst cells, were fluorescence marked and sorted using the Fluorescence Activated Cell Sorting (FACS) technique and then mixed in compositions summarized in Table 2.2. There are eight different designs with a total number of 200 cells in each design. Both pure and mixed cone and starburst cell groups are available.

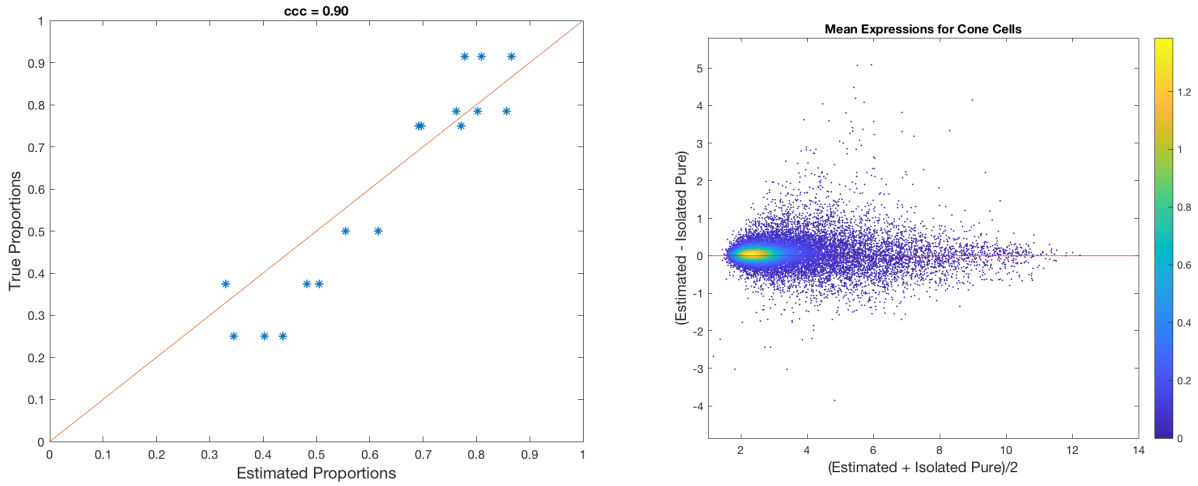
Design #	Number of Replicates	Tissue Type	Cone Cell #	Starburst Cell #
1	3	Pure	0	200
2	3	Pure	200	0
3	3	Mixed	50	150
4	3	Mixed	75	125
5	3	Mixed	100	100
6	3	Mixed	150	50
7	3	Mixed	157	43
8	3	Mixed	183	17

Table 2.2: Experimental design of GSE33076 and the compositions for the mixtures. Each design consists of 200 mouse retina cells in three replicates. Cone cells and starburst cells are either isolated as a pure group or mixed in different cell numbers.

The gene expression of each cell group was measured on the Affymetrix Mouse Gene 1.0 ST Array platform and was processed by the authors to remove contamination from rod cells, producing a working size of 22347 genes for each sample. Each design has three replicates thus there are twenty-four pure/mixed samples in total. In the original study of Siegert et al. (2012), ten cell-type

specific genes were used to test the linearity of the RNA amplification, which was demonstrated to be proportional to the amount of specific cells.

In our validation study, we allow the information for the pure cone cell groups (Design #2 in Table 2.2) to be covered, and take use of the expression of the whole genome to estimate the mixing proportions through FasTD, without the help of domain-knowledge about the cell-type specific genes. Figure 2.7 shows the results of our analysis, where the cone component is regarded as the unknown component in mixed Designs # 3-8, and the expression profiles of the pure starburst samples in Design #1 are used as the reference profiles. The convolution is modeled in the raw expression space instead of using the log2-transformed values. Genes with large difference between the mean expression of the mixtures and that of the reference profiles, as well as with low leverage in the regression step are used as the input for the FasTD pipeline. The FasTD outputs include both the proportion estimates and the mean expression estimates for the cone component are presented in Figure 2.7.



(a) Scatterplot of truth vs. estimates

(b) Mean expression values comparison

Figure 2.7: FasTD performance in estimating proportional coefficients and mean expression values for the cone component for dataset GSE33076. The scatter plot (a) demonstrates a good correlation (CCC = 0.9) between the FasTD estimates (x-axis) and the true proportions (y-axis), where each point is a sample. Plot (b) shows the differences (in y-axis) between the estimated mean expression value for the cone component in the mixtures, and that computed in the pure isolated cone cells. The x-axis shows the spread of these two quantities. All values are transformed into log2-scale. Genes are presented as dots and colored by the density value. A large density of genes clustered around the reference line suggesting a good estimation of the mean expression values.

Results in Figure 2.7 demonstrate our method to be effective in deconvoluting the two-component mixtures with reference profiles available for one component. The Concordance Correlation Coefficient (Lawrence and Lin, 1989) of 0.9 shows a good correlation between our estimated proportions and the ground truth. The other deconvolution product from the FasTD pipeline is the mean expression estimates for the cone component. As shown in Figure 2.7 (b) where genes are represented in points and colored by density, the differences for most genes are small, between our mean expression estimates and the ground truth (computed from the isolated pure cone cell samples in Design #2). These results show that the FasTD method is effective in deconvoluting two-component mixtures in this real data application.

2.5.2 The Cancer Genome Atlas Pan-can Deconvolution Analysis

In this section, we aim to apply the tumor purity estimation and individual deconvolution methods we developed in Sections 2.2 and 2.3, to 14 of The Cancer Genome Atlas (TCGA) cancer types. These cancer types were selected because they are all solid tumor primary cancers, also for each cancer type there are at least 19 normal control samples available. The number of original normal and tumor samples for each cancer type is summarized in the second and third columns of Table 2.3. The full names of each cancer type can be found online (TCGA Study Abbreviations, 2019).

The TCGA data represent a continuum between samples with essentially all normal cells and those with a high proportion of tumor cells. Since there is no sharp boundary, we use preprocessing to eliminate uncertain samples. We developed a procedure to detect these samples as a preprocessing step of the deconvolution pipeline. All uncertain samples are discarded before running FasTD. Our strategy is to use a Rank based Mann-Whitney U test to estimate the power of each gene in differentiating the normal and mixed tumor populations. The top one thousand genes with the smallest p-values are selected as the feature genes, whose first two principal component(PC) scores are then extracted for hierarchical clustering. Samples clustered with a majority of its opposite group are identified as uncertain samples. These samples are excluded in our deconvolution analysis. Figure 2.8 shows an example of this preprocessing step in the Lung Squamous Cell Carcinoma (LUSC) cancer in TCGA. Two LUSC normal samples are clustered with

14 TCGA CancerTypes	Original Normal Sample Counts	Original Tumor Sample Counts	After Preprocessing Normal Counts	After Preprocessing Tumor Counts
BLCA	19	414	17	385
BRCA	113	1101	98	1032
COAD	41	477	33	442
HNSC	44	500	31	494
KICH	24	65	23	64
KIRC	72	538	66	495
KIRP	32	288	26	276
LIHC	50	371	50	362
LUAD	59	532	57	446
LUSC	49	502	48	486
PRAD	52	333	47	295
STAD	32	375	32	299
THCA	57	502	55	418
UCEC	35	549	26	524

Table 2.3: Summary of normal and tumor sample counts for 14 TCGA cancer types before and after preprocessing.

other tumor samples hence regarded as uncertain. Columns 4 and 5 of Table 2.3 summarize the finalized working sample numbers for the fourteen cancer types after this preprocessing step.

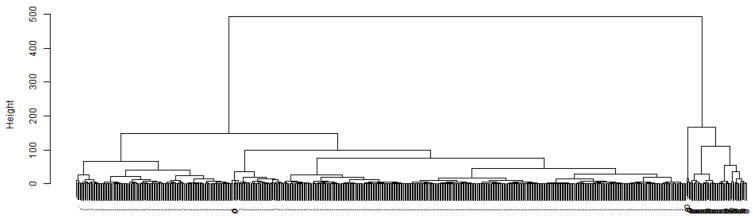


Figure 2.8: Illustration of the preprocessing step to identify uncertain samples in the TCGA LUSC dataset. The height in the y-axis is the euclidean distance between two clusters. Each leaf node represents a sample. The original tumor samples are labeled as dots and the original normal samples are labeled as circles. Two LUSC normal samples, shown as two overlapping circles on the lower-left part of the plot clustered with other tumor samples are identified as uncertain samples and disregarded for future analysis.

Due to the absence of ground truth for TCGA tumor purities, we evaluate the purity estimates for 14 cancer types by comparing them with the results obtained by ABSOLUTE (Carter et al., 2012), a DNA-based method that has been regarded within the research community as a gold standard for purity estimates. In Table 2.4 (with cancer types sorted by correlation for easy reading), we observe good concordance between these methods for several cancer types

such as Kidney Chromophobe (KICH), Prostate Adenocarcinoma (PRAD), Lung Squamous Cell Carcinoma (LUSC) and Colon Adenocarcinoma (COAD). However, we do also observe that some purity estimates obtained from FasTD and those from ABSOLUTE are not consistent, this discrepancy may arise from not only the difference between the two methodologies, but also the biological difference between DNA copy number datasets and RNA expression datasets. What's more, the presence of only two constituent subcomponents in a tumor sample is a strong assumption for complex cancer types such as the Head-Neck Squamous Cell Carcinoma (HNSC).

TCGA Cancer Type	RMSE	Sample #	Pearson Corr.	95% C.I.
KICH	0.11	64	0.67	(0.51, 0.79)
PRAD	0.16	287	0.65	(0.58, 0.72)
LUSC	0.23	478	0.55	(0.48, 0.61)
COAD	0.14	427	0.54	(0.47, 0.61)
BLCA	0.23	376	0.29	(0.19, 0.38)
STAD	0.33	293	0.28	(0.18, 0.39)
LUAD	0.19	418	0.22	(0.13, 0.31)
BRCA	0.21	989	0.21	(0.15, 0.27)
HNSC	0.19	480	0.11	(0.02, 0.2)
KIRP	0.30	262	0.05	(-0.07, 0.17)
UCEC	0.21	505	0.04	(-0.05, 0.13)
LIHC	0.43	339	-0.05	(-0.16, 0.06)
THCA	0.29	377	-0.11	(-0.21, -0.01)
KIRC	0.22	463	-0.13	(-0.22, -0.04)

Table 2.4: FasTD purity estimates comparison with ABSOLUTE. Two sets of estimates for the same collection of overlapping samples are evaluated by Root Mean Squared Error (RMSE) and Pearson correlation with 95% Confidence Interval. The correlation coefficients differ by cancer types.

Prostate cancer purity estimates obtained by RNA-based methods such as FasTD yield high concordance with estimates obtained by DNA-based methods. (This is also the case when DeMixT estimates are compared with ABSOLUTE estimates). In light of this high concordance and the relatively large sample numbers, we decide to conduct more analysis on this cancer type in the following section.

2.5.3 A Case Study: Prostate Cancer of The Cancer Genome Atlas

The high consistency in purity estimates between the two-component deconvolution methods and the ABSOLUTE estimates, suggests the PRAD mixture samples may fit well to the two constituent subcomponents assumption. Hence we conduct more in-depth investigation on this

cancer type, trying to understand the effect of deconvolution on downstream genomic analysis results.

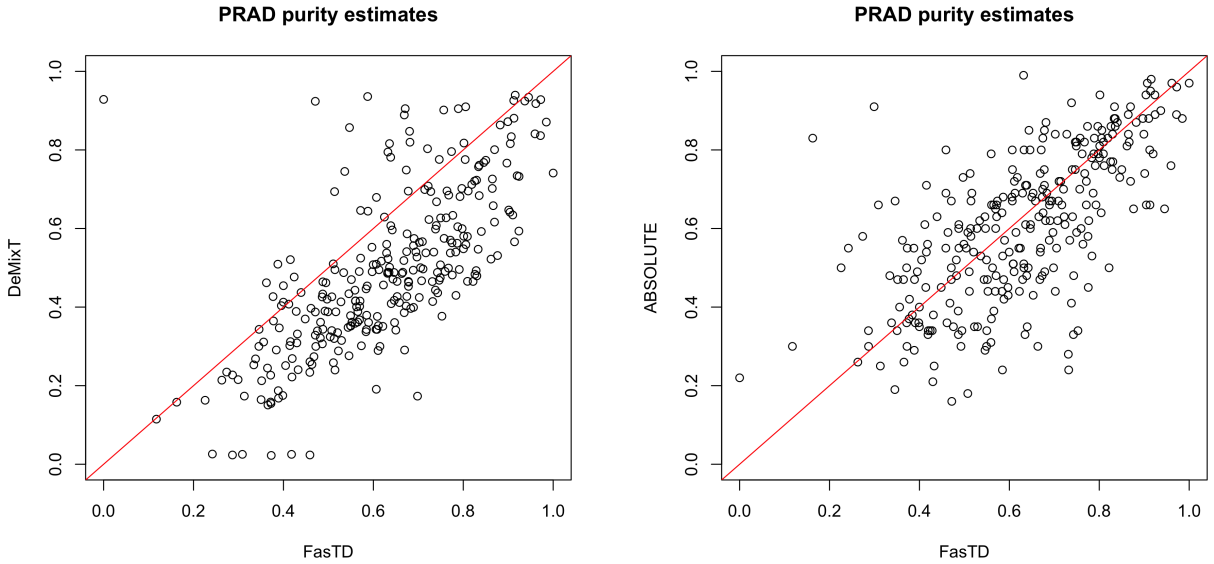
2.5.3.1 Tumor Purity Estimation

RNASeq sequencing data for prostate cancer samples collected from the Illumina HiSeq platform Version 2 Level 3 files are used. There are 47 normal control samples and 295 mixed tumor samples after the preprocessing step. Not all of the genes will be used as an input for our method. In fact, only genes that are differentially expressed in tumor and normal cells will serve as useful signals in the deconvolution steps. As described in Section 2.2.2, we start the purity estimation procedure with selecting the feature genes.

After performing a simple two sample t-test for each gene, the first set of feature genes are selected whose absolute t-statistics are larger than five. As the true purity information is not known to us, we evaluate these estimates by comparing them with the results obtained by the ABSOLUTE method. ABSOLUTE is a DNA-based method which has been applied to a large number of TCGA samples to estimate the tumor purity, and regarded as a *gold* standard for purity estimates in the community. (One should keep in mind that if the purity estimates obtained from our method and those from ABSOLUTE are not consistent, it may not only be due to the difference between the two methodologies, but can be due to the biological difference between DNA copy number and RNA expression data.) In addition, we compare our purity estimates with those obtained by DeMixT. DeMixT uses the same RNAseq dataset and we observe a high correlation between DeMixT estimates and ours: 0.92 (Figure 2.9 a). But for this set of real data, DeMixT takes almost five hours to get the purity estimates, whereas our method takes less than 5 minutes. Therefore, for two-source signals deconvolution on RNAseq data, comparing with DeMixT, FasTD is able to produce real data results in a much shorter time period.

We present the estimates for PRAD tumor samples obtained by three methods in Figure 2.9. Each point represents a mixed PRAD tumor sample. The x-axis value is the tumor proportions obtained from our method (Section 2.2.2). The y-axis represents proportion estimates either from DeMixT (Figure 2.9 a) or ABSOLUTE (Figure 2.9 b). We observe in Figure 2.9 (b) a good concordance between ABSOLUTE and FasTD tumor proportion estimates, with a Pearson

correlation equal to 0.65. Figure 2.9 (a) suggests a strong correlation, however either our proportions are under-estimated, or the DeMixT results are over-estimated. The consistency of the purities across different methods adds to the credibility of FasTD purity estimates, which will be further used for the individual deconvolution part of the pipeline.



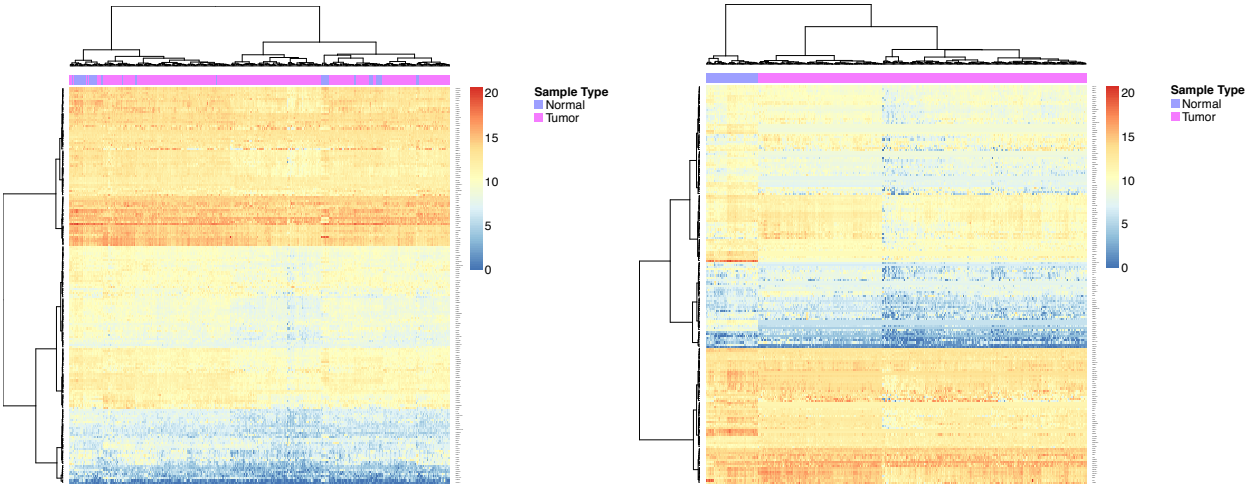
(a) Our method vs. DeMixT, cor = 0.71 **(b)** Our method vs. ABSOLUTE, cor = 0.65

Figure 2.9: Purity estimation performance comparison between our method and the other two methods for PRAD tumor samples. In both plots the x-axis shows the FasTD estimate value and each point is a tumor sample. The y-axis in (a) is the DeMixT estimate and that in (b) is the ABSOLUTE estimate for the same set of PRAD tumor samples. Overall, the Pearson correlation coefficients between different methods are high in this cancer type.

2.5.3.2 Biological Implications using Deconvolved Profiles

With the tumor purity estimates, one can get the mean and variance estimates for the tumor component according to (2.18) and (2.16), respectively. Then the tumor specific expression profiles can be obtained for each PRAD sample using the likelihood estimator $\hat{T}_{ig}$ introduced in Section 2.3. There is no gold standard for purified tumor expression profiles in the TCGA community, so it is not feasible to directly validate these individual deconvolution results. However, we demonstrate the statistical analysis results using the purified tumor expression profiles are consistent with current understanding of the prostate cancer, which indirectly validate and show the value of using deconvoluted profiles.

Using the FasTD individual deconvolution profiles, we observe sharper contrast between tumor and normal populations for tumorigenesis related pathways. Take the Epithelial Mesenchymal Transition (EMT) pathway for example, which has been reported to be associated with cancer metastasis (Ye and Weinberg, 2015). When we use clustering analysis to examine the tumor-normal differences of the EMT genes, we observe a better separation of the tumor versus normal samples using deconvoluted tumor profiles in Figure 2.10 (b), which is not observed using the original tumor profiles in Figure 2.10 (a). Hence by conducting tumor deconvolution, we are able to obtain 'cleaner' tumor signals better separated from the adjacent normal cells, which may provide new insights on tumor specific gene expression changes.



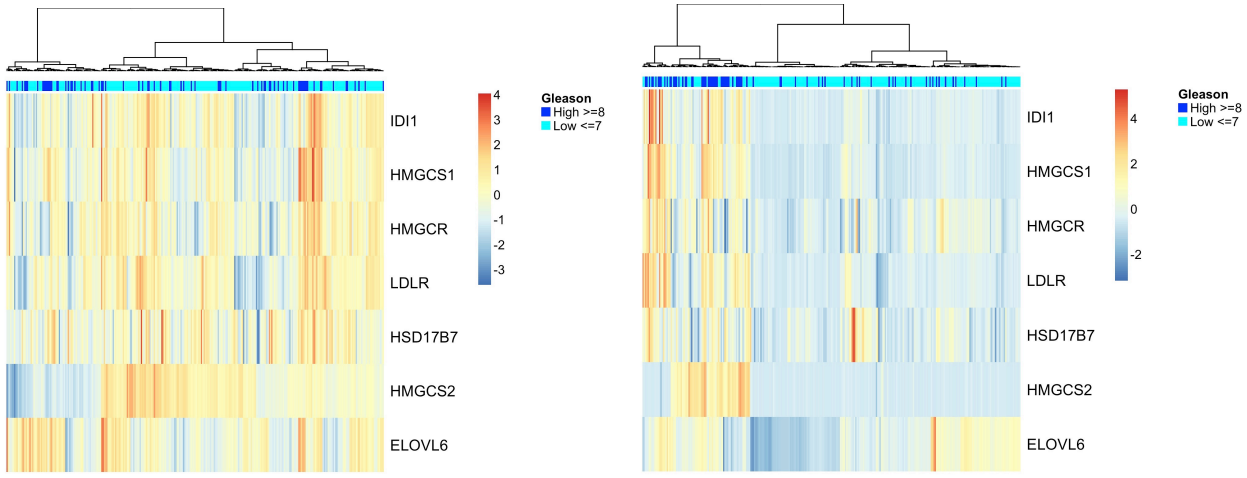
(a) Before deconvolution **(b)** After deconvolution

Figure 2.10: Two-way clustering results of PRAD samples and genes in the epithelial mesenchymal transition (EMT) pathway before and after DeMixT deconvolution. Each row represents a gene in the pathway and each column represent a prostate sample. Top color bar highlights the normal samples (purple) versus tumor samples (pink). A better separation of the tumor versus normal samples is observed only after deconvolution.

Furthermore, as a result of the collaboration work with the DeMixT group (Wang et al., 2017) and scientists at the MD Anderson Cancer Center, we observe more biological insights using the deconvoluted profiles. A recent study (Chen et al., 2018) uncovered that the SREBP-dependent lipogenic program is hyperactivated in prostate cancer samples, which motivates us to examine the expression level of these signature genes in the TCGA prostate samples. Using the seven SREBP mediated lipogenesis signature genes reported in Chen et al. (2018), we perform a one-way clustering analysis for the prostate dataset. We also label the tumor samples by Gleason Score, an

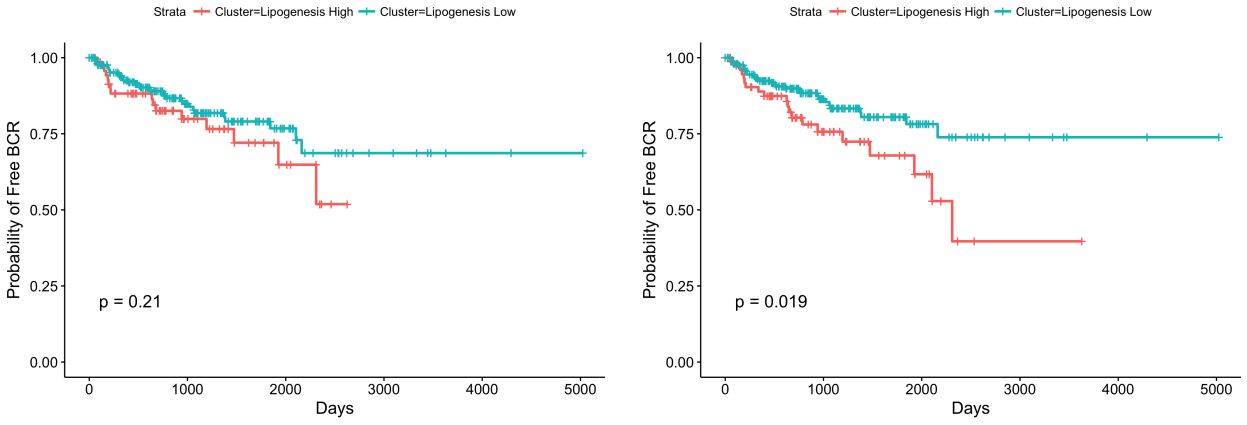
independent characterization for prostate cancer cells that measures how different they are from normal cells and how likely they are to spread. Consistently, the clustering result in Figure 2.11 shows the increased expression in the lipogenesis signature genes is associated with higher Gleason scores of PRAD tumor samples, which is in line with the findings in Chen et al. (2018) but only observed using the deconvoluted expression matrix. Moreover, when we separate tumor samples into the 'Lipogenesis High' and the 'Lipogenesis Low' groups according to the expression levels of these seven lipogenesis genes, we observe a significant decrease ($p = 0.019$) in the probability of Free Biochemical Recurrence (BCR) event for the Lipogenesis High group (Figure 2.11.d). This difference is only observed using the deconvoluted profiles and is consistent with the previous finding that the up-regulated lipogenesis genes promote prostate cancer metastasis (Chen et al., 2018).

The consistency between current biological findings in prostate cancer and the statistical conclusions derived from deconvoluted gene expression profiles demonstrates the value added from performing the deconvolution step before downstream genomic analysis. The deconvoluted profiles for each cellular component at the gene level help us to move forward from working with fractions of the subcomponents to performing more systematic analyses with higher tumor to normal contrast. These analyses may lead to new insights/discoveries in tumorigenesis that will not be observable using the original profiles (Figure 2.10 & 2.11).



(a) Before deconvolution

(b) After deconvolution



(c) Before deconvolution

(d) After deconvolution

Figure 2.11: (a) and (b) One-way clustering for PRAD tumor samples with 7 SREBP mediated lipogenesis signature genes before and after deconvolution. Each row represents a gene and each column a prostate tumor sample. All tumor samples have a categorized Gleason score indicated by the color bar. Tumor samples with higher Gleason scores ($\leq$8) are clustered for higher expression values in these SREBP mediated lipogenesis signature genes after deconvolution. (c) and (d) shows the Kaplan-Meier plot for BCR events, stratified for different tumor clusters grouped by the expression levels of 7 lipogenesis genes in (a) and (b), respectively. Vertical lines indicate the time at which censoring occurred. The log-rank test is used to compare survival curves of two groups, whose p-value is shown on the plot. Statistically significant difference in the probability of BCR event for two clusters can only be observed using the deconvolved expression values.

CHAPTER 3
Multiple Components Separation Analysis

3.1 Introduction

The tumor tissue can be evaluated as a complete organ with its own micro-environment as illustrated in Figure 1.2, suggesting complex constitute components. In addition, it has been reported in many cancer types, tumor cells can be further separated into distinctive subtypes according to their genomic profile. For example, a well-established classifier, PAM50, is often used to categorize breast tumors into five intrinsic subtypes based on their expression profiles (Bernard et al., 2009). The results of our pan-can deconvolution analysis for the TCGA study (Table 2.4) also suggest that the assumption of only two constituent subcomponents in a tumor sample might be too strong for complex cancer types. Therefore, it is desired to develop a deconvolution tool applicable for bulk tumor tissue samples containing more than two subcomponents. In this chapter, we will demonstrate our work of extending the tool developed in Chapter 2 to account for multiple-subcomponent scenarios.

The development of the Fast Tumor Deconvolution for K components (FasTDK) pipeline is explained in details in Section 3.2. We demonstrate that FasTDK is effective in estimating the mixing proportions for K constituent subcomponents and estimating the tumor specific expression profiles through simulation studies (Section 3.3.1) and a real data application (Section 3.4). To address the challenge of increasing noise brought by additional subcomponents, we introduce a new gene filtering scheme in Section 3.2.4. Through the sensitivity analysis in Section 3.3.1, we find the estimation accuracy for mixing proportions gets better when the number of mixing samples increases. The real data analysis in Section 3.4.3 also helps us to discover a new cut-off criterion, which improves the estimation accuracy by selecting genes with strong signals for subcomponents in small proportions.

3.2 K Components with K-1 Reference Profiles

In situations when multiple subcomponents are present in a tumor mixture sample, a deconvolution tool capable of separating more than two subcomponents is desirable. We present a deconvolution pipeline 'FasTDK', to estimate the proportion and type-specific expression value of the unknown subcomponent in a K-component mixture, with the assumption that the total number of subcomponents K is known, and $K-1$ expression reference profiles are available. In the case of mixed gene expression data derived from bulk tumor tissues, tumor is often regarded as the unknown subcomponent, and we assume some reference expression profiles for other cell types in the mixtures such as the immune cells, blood cells and stromal cells, are available. In this section, we will first set up the problem mathematically and then describe in details how to deconvolute K-components mixing samples using the strategy extended from the FasTD pipeline. We will also introduce a new gene filtering scheme in subsection 3.2.4 to facilitate the more complex multi-component deconvolution.

3.2.1 Problem Formulation and Model Assumptions

Given the expression data obtained from S bulk tissue tumor samples consists of K subcomponents, our goal is to estimate the proportions of the tumor component, and the tumor-specific expression values in each of the mixture samples.

Let Y_{ig}, be the observed gene expression value for mixed tumor sample i, $i = 1\ , \ldots,\ S$, gene g, $g = 1, \ldots, G$. We assume each observed value Y_{ig} is a proportional sum of K independent components, $C_{ig}^1, \ldots, C_{ig}^K$, where C_{ig}^j, $j = 1, \ldots, K$ is a gene expression random variable for sample i, gene g in subcomponent j. The proportional weight for subcomponent j in sample i is denoted as π_{ij}. For each sample i, the proportional weights are constrained in a $K-1$ standard simplex, *i.e.*, $\sum_{j=1}^{K} \pi_{ij} = 1, \pi_{ij} \in [0,1], \forall j = 1, \ldots, K$. Hence, the observed mixing expression value Y_{ig} can be written as:

$$Y_{ig} = \pi_{i1} C_{ig}^1 + \pi_{i2} C_{ig}^2 + \cdots + \pi_{iK} C_{ig}^K \quad \forall i, g. \tag{3.1}$$

Without loss of generality, we assume the K^{th} subcomponent is the unknown in the mixture and reference profiles for the first $K-1$ subgroups are available. This means the mean and variance information for random variables $C_{ig}^1, \ldots, C_{ig}^{K-1}$ can be estimated from the reference profiles, i.e. $\hat{\mu}_{jg}$ and $\hat{\sigma}_{jg}^2$ for gene g in subcomponent $j = 1, \ldots, K$, are treated as known quantities. Again, there is no assumption on the distribution of these variables. The goal of the deconvolution problem, is to estimate the remaining unknown parameters in model (3.1): the mixing proportions $\pi_{ij}, j = 1, \ldots, K$, the mean and variance for the expression of gene g in the unknown component, $\hat{\mu}_{Kg}$ and $\hat{\sigma}_{Kg}^2$, and the explicit expression values of C_{ig}^K.

3.2.2 Proportion Estimation using Moments Information

Estimation procedures for the mixing proportions, and the mean and variance parameters for the unknown component are explained in this subsection using the first and second moments information of the dataset.

3.2.2.1 Estimating Regression Coefficients

First, we rewrite model (3.1) in a vector form. Let $\boldsymbol{\mu}_j \in \mathbb{R}^{G\times 1}$ and $\boldsymbol{\Sigma}_j \in \mathbb{R}^{G\times G}$ denote the mean vector and covariance matrix of gene expression for G genes in the j^{th} subcomponent. The observed gene expression vector $\mathbf{y}_i \in \mathbb{R}^{G\times 1}$ for mixed tumor sample i is the proportional sum of the mean expression vectors of all subcomponents, plus some noise $\boldsymbol{\epsilon}_i$:

$$\mathbf{y}_i = \sum_{j=1}^{K} \pi_{ij}\boldsymbol{\mu}_j + \boldsymbol{\epsilon}_i \tag{3.2}$$

where

$$\boldsymbol{\epsilon}_i \sim \left(\mathbf{0}, \quad \sum_{j=1}^{K} \pi_{ij}^2 \boldsymbol{\Sigma}_j\right) \quad \forall i.$$

Subtracting $\boldsymbol{\mu}_1$ from both sides of (3.2) we get:

$$\mathbf{y}_i - \boldsymbol{\mu}_1 = \sum_{j=1}^{K} \pi_{ij}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_1) + \boldsymbol{\epsilon}_i, \tag{3.3}$$

which resembles a multiple linear regression (MLR) without an intercept. However, one of the predictors on the right side of (3.3) is unknown: $(\boldsymbol{\mu}_K - \boldsymbol{\mu}_1)$. We strategically use the mean expression value of the observed data to estimate this quantity. Since:

$$\begin{aligned}\frac{1}{S}\sum_{i=1}^{S}(\mathbf{y}_i - \boldsymbol{\mu}_1) &= \frac{1}{S}\sum_{i=1}^{S}\sum_{j=1}^{K}\pi_{ij}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_1) + \frac{1}{S}\sum_{i=1}^{S}\boldsymbol{\epsilon}_i \\ &\approx \frac{1}{S}\sum_{i=1}^{S}\sum_{j=1}^{K}\pi_{ij}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_1) \\ &= \sum_{j=1}^{K-1}\bar{\pi}_j(\boldsymbol{\mu}_j - \boldsymbol{\mu}_1) + \bar{\pi}_K(\boldsymbol{\mu}_K - \boldsymbol{\mu}_1)\end{aligned} \tag{3.4}$$

where $\bar{\pi}_j = \frac{\sum_{i=1}^{S}\bar{\pi}_{ij}}{S}$, for $j = 1, \ldots, K$, we have an estimate of:

$$\widehat{\boldsymbol{\mu}_K - \boldsymbol{\mu}_1} = \frac{1}{\bar{\pi}_K}[(\bar{\mathbf{y}} - \boldsymbol{\mu}_1) - \sum_{j=1}^{K-1}\bar{\pi}_j(\boldsymbol{\mu}_j - \boldsymbol{\mu}_1)] \tag{3.5}$$

where $\bar{\mathbf{y}} = \frac{1}{S}\sum_{i=1}^{S}\mathbf{y}_i$.

Next, replace the unknown predictor in (3.3) with (3.5) and the multiple linear regression problem becomes:

$$\begin{aligned}\mathbf{y}_i - \hat{\boldsymbol{\mu}}_1 &= \sum_{j=1}^{K-1}\pi_{ij}\cdot(\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_1) + \pi_{iK}\cdot(\boldsymbol{\mu}_K - \hat{\boldsymbol{\mu}}_1) + \boldsymbol{\epsilon}_i \\ &= \sum_{j=1}^{K-1}\pi_{ij}\cdot(\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_1) + \pi_{iK}\cdot(\widehat{\boldsymbol{\mu}_K - \boldsymbol{\mu}_1}) + \boldsymbol{\delta}_i \\ &= \sum_{j=1}^{K-1}\pi_{ij}\cdot(\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_1) + \frac{\pi_{iK}}{\bar{\pi}_K}\cdot\{(\bar{\mathbf{y}} - \hat{\boldsymbol{\mu}}_1) - \sum_{j=1}^{K-1}\bar{\pi}_j(\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_1)\} + \boldsymbol{\delta}_i \\ &= \sum_{j=1}^{K-1}(\pi_{ij} - \frac{\pi_{iK}}{\bar{\pi}_K}\bar{\pi}_j)(\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_1) + \frac{\pi_{iK}}{\bar{\pi}_K}(\bar{\mathbf{y}} - \hat{\boldsymbol{\mu}}_1) + \boldsymbol{\delta}_i.\end{aligned} \tag{3.6}$$

To ensure homoscedasticity for the above regression problem, we introduce the gene weighting diagonal matrix $\mathbf{Q}$, with positive elements on the diagonal. To start with, one can set the weights of each gene as the reciprocal of the sample standard deviation computed from $\mathbf{Y} \in \mathbb{R}^{S\times G}$:

$$\hat{\sigma}_g^2 = \frac{1}{(S-1)} \sum_{i=1}^{S} (y_{ig} - \bar{y}_{\cdot g})^2 \quad \forall g, \qquad \mathbf{Q} = \begin{pmatrix} \frac{1}{\hat{\sigma}_1} & & \\ & \ddots & \\ & & \frac{1}{\hat{\sigma}_G} \end{pmatrix}_{\tilde{G} \times G}.$$

Then the final expression of the multiple linear regression problem becomes:

$$\mathbf{Q}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_1) = \sum_{j=1}^{K-1} (\pi_{ij} - \frac{\pi_{iK}}{\bar{\pi}_K} \bar{\pi}_j) \mathbf{Q}(\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_1) + \frac{\pi_{iK}}{\bar{\pi}_K} \mathbf{Q}(\bar{\mathbf{y}} - \hat{\boldsymbol{\mu}}_1) + \mathbf{Q}\boldsymbol{\delta}_i. \tag{3.7}$$

For each sample i, denote $\boldsymbol{\beta}_i = (\hat{\beta}_{i2}, \hat{\beta}_{i3}, \ldots, \hat{\beta}_{iK})$ as the coefficients obtained from (3.7). There exists such relations:

$$\begin{aligned} \hat{\beta}_{ij} &= \pi_{ij} - \hat{\beta}_{iK} \bar{\pi}_j \quad \forall j = 2, \ldots, K-1 \\ \hat{\beta}_{iK} &= \frac{\pi_{iK}}{\bar{\pi}_K}. \end{aligned} \tag{3.8}$$

It is worth-noting that for a given sample i, the π_{ij}'s are subject to the $K-1$ standard simplex constraint ($\sum_{j=1}^{j=K} \pi_{ij} = 1; 0 \leq \pi_{ij} \leq 1, \forall j = 1, \ldots, K$). Hence, taking this constraint into consideration, the final coefficient estimates should have constraints:

$$\begin{aligned} &\hat{\beta}_{iK} \geq 0 \\ &-\hat{\beta}_{iK} \leq \hat{\beta}_{ij} \leq 1, \quad \forall j = 2, \ldots, K-1. \end{aligned}$$

3.2.2.2 Estimating $\bar{\pi}$'s for each subcomponent

We have obtained the regression coefficients β_{ij}'s, but to fully recover π_{ij} through (3.8), we need estimates of $\bar{\pi}_2, \ldots, \bar{\pi}_K$. To achieve this goal, the second moment information of the dataset is utilized.

We have assumed the error term $\boldsymbol{\epsilon}_i$ follows:

$$\boldsymbol{\epsilon}_i = \mathbf{y}_i - \boldsymbol{\mu}_1 - \sum_{j=1}^{K} \pi_{ij} (\boldsymbol{\mu}_j - \boldsymbol{\mu}_1) \sim \Big(\mathbf{0}, \quad \sum_{j=1}^{K} \pi_{ij}^2 \boldsymbol{\Sigma}_j \Big) \quad \forall i.$$

Similar to the derivations in (2.12) and (2.11), for each sample i, the weighted second moment information of $\boldsymbol{\epsilon}_i$ and its expectation can be expressed as:

$$\mathbf{Z}_i \approx \big(\mathbf{Q}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_1) - \sum_{j=1}^{K-1} \hat{\beta}_{ij} \cdot \mathbf{Q}(\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_1) - \hat{\beta}_{iK}\mathbf{Q}(\bar{\mathbf{y}} - \hat{\boldsymbol{\mu}}_1)\big)^{\otimes 2}$$

and

$$\mathbb{E}(\mathbf{Z}_i) = \sum_{j=1}^{K} \pi_{ij}^2 \mathbf{Q}\boldsymbol{\Sigma}_j\mathbf{Q}'$$

respectively. Based on these two quantities, we construct an objective function, which minimizes the sum of squared deviation of the weighted second moment of the error term from its expectation. In other words, this objective function reduces the deviation of each of these moment conditions from that would be theoretically obtained.

The mathematically expression of the objective function is written as:

$$\begin{aligned} &\sum_{i=1}^{S} \|\mathrm{diag}[\mathbf{Z}_i - \mathbb{E}(\mathbf{Z}_i)]\|_2^2 \\ =&\sum_{i=1}^{S} \Big\|\mathrm{diag}[\big(\mathbf{Q}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_1) - \sum_{j=1}^{K-1} \hat{\beta}_{ij} \cdot \mathbf{Q}(\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_1) - \hat{\beta}_{iK}\mathbf{Q}(\bar{\mathbf{y}} - \hat{\boldsymbol{\mu}}_1)\big)^{\otimes 2} - \sum_{j=1}^{K} \pi_{ij}^2 \mathbf{Q}\boldsymbol{\Sigma}_j\mathbf{Q}']\Big\|_2^2. \end{aligned} \tag{3.9}$$

The proportion parameters π_{ij}'s in (3.9) are still unknown. But we can replace them with the regression coefficients we obtained in (3.8):

$$\begin{aligned} \pi_{ij} &= \hat{\beta}_{ij} + \hat{\beta}_{iK}\bar{\pi}_j \quad \forall j = 2, \ldots, K-1 \\ \pi_{iK} &= \hat{\beta}_{iK}\bar{\pi}_K \\ \pi_{i1} &= 1 - \sum_{j=2}^{K-1} (\hat{\beta}_{ij} + \hat{\beta}_{iK}\bar{\pi}_j) - \hat{\beta}_{iK}\bar{\pi}_K \end{aligned} \tag{3.10}$$

where $\forall j = 1, \ldots, K, \quad \bar{\pi}_j = \frac{1}{S}\sum_{i=1}^{S} \pi_{ij}$.

Then the remaining unknowns in (3.9) are the averages of the proportions parameters $\bar{\pi}_j$, $j = 1, \ldots, K$, and the variance parameters of the unknown component σ_{Kg}^2, $g = 1, \ldots, G$. Note that the proportion parameters π_{ij}'s are subject to the $K-1$ standard simplex constraint. Therefore,

we need to add the derived constraints for the corresponding $\bar{\pi}_j$'s:

$$
\begin{aligned}
&0 \le \bar{\pi}_K \le \min(\frac{1}{\max_i(\hat{\beta}_{iK})}, 1) \\
&\min(\max_i(-\frac{\hat{\beta}_{ij}}{\hat{\beta}_{iK}}, 0), 1) \le \bar{\pi}_j \le \max(\min_i(\frac{1-\hat{\beta}_{ij}}{\hat{\beta}_{iK}}, 1), 0) \quad \forall j = 2, \ldots, K-1 \\
&\min(\max_i(-\frac{\sum_{j=2}^{K-1}\hat{\beta}_{ij}}{\hat{\beta}_{iK}}, 0), 1) \le \sum_{j=2}^{K}\bar{\pi}_j \le \max(\min_i(\sum_{j=2}^{K}\frac{1-\sum_{j=2}^{K-1}\hat{\beta}_{ij}}{\hat{\beta}_{iK}}, 1), 0).
\end{aligned}
\tag{3.11}
$$

With all the constraints in (3.11) and the objective function in (3.9), we first formulated an optimization problem solving for the average proportions and the variance parameters simultaneously:

$$
\begin{aligned}
\min_{\bar{\pi}_2,\ldots,\bar{\pi}_K,\mathrm{diag}(\mathbf{\Sigma}_K)} & \sum_{i=1}^{S}\Big\|\mathrm{diag}[\big(\mathbf{Q}(\mathbf{y}_i-\hat{\boldsymbol{\mu}}_1)-\sum_{j=1}^{K-1}\hat{\beta}_{ij}\cdot\mathbf{Q}(\hat{\boldsymbol{\mu}}_j-\hat{\boldsymbol{\mu}}_1)-\hat{\beta}_{iK}\mathbf{Q}(\bar{\mathbf{y}}-\hat{\boldsymbol{\mu}}_1)\big)^{\otimes 2} \\
& -\sum_{j=2}^{K-1}(\hat{\beta}_{ij}+\hat{\beta}_{iK}\bar{\pi}_j)^2\mathbf{Q}\mathbf{\Sigma}_j\mathbf{Q}'-(\hat{\beta}_{iK}\bar{\pi}_K)^2\mathbf{Q}\mathbf{\Sigma}_K\mathbf{Q}' \\
& -(1-\sum_{j=2}^{K-1}(\hat{\beta}_{ij}+\hat{\beta}_{iK}\bar{\pi}_j)-\hat{\beta}_{iK}\bar{\pi}_K)^2\mathbf{Q}\mathbf{\Sigma}_1\mathbf{Q}']\Big\|_2^2 \\
\textit{s.t.:}\quad & 0 \le \bar{\pi}_K \le \min(\frac{1}{\max_i(\hat{\beta}_{iK})}, 1) \\
& \min(\max_i(-\frac{\hat{\beta}_{ij}}{\hat{\beta}_{iK}}, 0), 1) \le \bar{\pi}_j \le \max(\min_i(\frac{1-\hat{\beta}_{ij}}{\hat{\beta}_{iK}}, 1), 0) \quad \forall j = 2, \ldots, K-1 \\
& \min(\max_i(-\frac{\sum_{j=2}^{K-1}\hat{\beta}_{ij}}{\hat{\beta}_{iK}}, 0), 1) \le \sum_{j=2}^{K}\bar{\pi}_j \le \max(\min_i(\sum_{j=2}^{K}\frac{1-\sum_{j=2}^{K-1}\hat{\beta}_{ij}}{\hat{\beta}_{iK}}, 1), 0) \\
& \mathrm{diag}(\mathbf{\Sigma}_K) \ge \mathbf{0}.
\end{aligned}
\tag{3.12}
$$

Above is a quadratic programming problem with $K-1$ unknown scalars $\bar{\pi}_2, \ldots, \bar{\pi}_K$ and a vector of gene variances for the K^{th} tumor component. When the scalars are given, with the assumption of independence between genes, we have an explicit expression for the optimal solution of σ^2_{Kg}, as

a function of $\bar{\pi}_j$, $j = 2, \ldots, K$:

$$\hat{\sigma}^2_{Kg} = \Big\{ \sum_{i=1}^{S} \Big[\big(y_{ig} - \hat{\mu}_{1g} - \sum_{j=1}^{K-1} \hat{\beta}_{ij} \cdot (\hat{\mu}_{jg} - \hat{\mu}_{1g}) - \hat{\beta}_{iK}(\bar{y}_{\cdot g} - \hat{\mu}_{1g}) \big)^2 \\ - \sum_{j=2}^{K-1} (\hat{\beta}_{ij} + \hat{\beta}_{iK}\bar{\pi}_j)^2 \hat{\sigma}^2_{jg} - (1 - \sum_{j=2}^{K-1} (\hat{\beta}_{ij} + \hat{\beta}_{iK}\bar{\pi}_j) - \hat{\beta}_{iK}\bar{\pi}_K)^2 \hat{\sigma}^2_{1g} \Big] \cdot \hat{\beta}^2_{iK} / (\bar{\pi}^2_K \sum_{i=1}^{S} \hat{\beta}^4_{iK}) \Big\}_+ . \tag{3.13}$$

This expression for optimal $\hat{\sigma}^2_{Kg}$ is plugged back into (3.12). Then the final version of the optimization problem becomes:

$$\min_{\bar{\pi}_2, \ldots, \bar{\pi}_K} \sum_{i=1}^{S} \sum_{g=1}^{\tilde{G}} \Big\{ \big[Q_g(y_{ig} - \hat{\mu}_{1g}) - \sum_{j=1}^{K-1} \hat{\beta}_{ij} Q_g(\hat{\mu}_{jg} - \hat{\mu}_{1g}) - \hat{\beta}_{iK} Q_g(\bar{y}_{\cdot g} - \hat{\mu}_{1g}) \big]^2 \\ - \sum_{j=2}^{K-1} (\hat{\beta}_{ij} + \hat{\beta}_{iK}\bar{\pi}_j)^2 Q^2_g \hat{\sigma}^2_{jg} - \big[1 - \sum_{j=2}^{K-1} (\hat{\beta}_{ij} + \hat{\beta}_{iK}\bar{\pi}_j) - \hat{\beta}_{iK}\bar{\pi}_K \big]^2 Q^2_g \hat{\sigma}^2_{1g} \\ - \frac{\hat{\beta}^2_{iK}}{\sum_{i=1}^{S} \hat{\beta}^4_{iK}} \Big\{ \sum_{i=1}^{S} \Big[\big[Q_g(y_{ig} - \hat{\mu}_{1g}) - \sum_{j=1}^{K-1} \hat{\beta}_{ij} Q_g(\hat{\mu}_{jg} - \hat{\mu}_{1g}) - \hat{\beta}_{iK} Q_g(\bar{y}_{\cdot g} - \hat{\mu}_{1g}) \big]^2 \\ - \sum_{j=2}^{K-1} (\hat{\beta}_{ij} + \hat{\beta}_{iK}\bar{\pi}_j)^2 Q^2_g \hat{\sigma}^2_{jg} - (1 - \sum_{j=2}^{K-1} (\hat{\beta}_{ij} + \hat{\beta}_{iK}\bar{\pi}_j) - \hat{\beta}_{iK}\bar{\pi}_K)^2 Q^2_g \hat{\sigma}^2_{1g} \Big] \cdot \hat{\beta}^2_{iK} \Big\}_+ \Big\}^2 \tag{3.14}$$

$$\textit{s.t.:} \quad 0 \le \bar{\pi}_K \le \min\Big(\frac{1}{\max_i(\hat{\beta}_{iK})}, 1\Big)$$

$$\min\Big(\max_i\Big(-\frac{\hat{\beta}_{ij}}{\hat{\beta}_{iK}}, 0\Big), 1\Big) \le \bar{\pi}_j \le \max\Big(\min_i\Big(\frac{1 - \hat{\beta}_{ij}}{\hat{\beta}_{iK}}, 1\Big), 0\Big) \quad \forall j = 2, \ldots, K-1$$

$$\min\Big(\max_i\Big(-\frac{\sum_{j=2}^{K-1} \hat{\beta}_{ij}}{\hat{\beta}_{iK}}, 0\Big), 1\Big) \le \sum_{j=2}^{K} \bar{\pi}_j \le \max\Big(\min_i\Big(\sum_{j=2}^{K} \frac{1 - \sum_{j=2}^{K-1} \hat{\beta}_{ij}}{\hat{\beta}_{iK}}, 1\Big), 0\Big)$$

$$\mathrm{diag}(\mathbf{\Sigma}_K) \ge \mathbf{0}.$$

3.2.2.3 Obtaining Constituent Subcomponent Proportions

Plug the optimal values $\hat{\bar{\pi}}_2, \ldots, \hat{\bar{\pi}}_K$ obtained from the optimization problem (3.14) into (3.10). We can get the final estimates of the sample-wise proportion coefficients for each subcomponent in the mixtures. For sample i, $i = 1, \ldots, S$:

$$\begin{aligned}
\hat{\pi}_{ij} &= \hat{\beta}_{ij} + \hat{\beta}_{iK}\hat{\bar{\pi}}_j \quad \forall j = 2, \ldots, K-1 \\
\hat{\pi}_{iK} &= \hat{\beta}_{iK}\hat{\bar{\pi}}_K \\
\hat{\pi}_{i1} &= 1 - \sum_{j=2}^{K-1}(\hat{\beta}_{ij} + \hat{\beta}_{iK}\hat{\bar{\pi}}_j) - \hat{\beta}_{iK}\hat{\bar{\pi}}_K.
\end{aligned} \tag{3.15}$$

3.2.2.4 Estimating Expression Mean and Variance for the Unknown

Estimates of the expression variance $\hat{\sigma}^2_{Kg}$ for gene g in subcomponent K can be obtained by plugging $\hat{\bar{\pi}}_1, \ldots, \hat{\bar{\pi}}_K$ into expression (3.13). The corresponding mean expression estimate is obtained by:

$$\hat{\mu}_{Kg} = (\frac{1}{S}\sum_{i=1}^{S} y_{ig} - \sum_{j=1}^{K-1} \hat{\bar{\pi}}_j \mu_{jg})_+ / \hat{\bar{\pi}}_K. \tag{3.16}$$

With the estimates for the mixing proportions, the expression mean and variance for genes in the unknown component, one can proceed to the individual deconvolution part of the pipeline, to estimate the explicit expression values for the unknown component in the mixtures.

3.2.3 Individual Deconvolution

As a result of the steps in Section 3.2.2, many unknown parameters are estimated. From here and onward, we assume the mean and variance parameters for each gene expression random variable $C^1_g, \ldots, C^K_g, \forall g = 1, \ldots, G$, are known to us, as well as the subcomponent level proportions $\pi_{i1}, \ldots, \pi_{iK}$, for all samples $1, \ldots, S$.

The next goal is to estimate the explicit expression value C^K_{ig} for the unknown subcomponent K. In the individual deconvolution section of Chapter 2 (Section 2.3), we introduced an estimator that combines two important quantities associated with the target value T_{ig}. Likewise, we also identified two important quantities capturing the information of target value C^K_{ig} in the multiple component case: μ_{Kg} and $\tilde{C}^K_{ig}$, where

$$\tilde{C}^K_{ig} = \frac{Y_{ig} - \sum_{j=1}^{K-1} \pi_{ij}\mu_{jg}}{\pi_{iK}},$$

whose conditional mean and variance is written as:

$$\begin{aligned} \mathrm{E}(\tilde{C}_{ig}^K | C_{ig}^K) &= C_{ig}^K \\ \mathrm{Var}(\tilde{C}_{ig}^K | C_{ig}^K) &= \sum_{j=1}^{K-1} (\frac{\pi_{ij}}{\pi_{iK}})^2 \sigma_{jg}^2. \end{aligned}$$

With a strategy similar to that of Section 2.3.1, we first propose a general estimator of C_{ig}^K to be a weighted sum between μ_{Kg} and $\tilde{C}_{ig}^K$ of the form:

$$C_{ig}^{K_G} = \frac{a \cdot \mu_{Kg} + b \cdot \tilde{C}_{ig}^K}{a+b} \tag{3.17}$$

where a and b are two arbitrary numbers. Then the optimal solutions for a and b are obtained by minimizing the conditional MSE of $C_{ig}^{K_G}$:

$$\begin{aligned} a^* &= \sum_{j=1}^{K-1} \pi_{ij}^2 \sigma_{jg}^2 \\ b^* &= \pi_{iK}^2 (\tilde{C}_{ig}^K - C_{ig}^K)^2 \end{aligned} \tag{3.18}$$

Plug the optimal a^* and b^* back to $C_{ig}^{K_G}$ into obtain the oracle estimator, $\dot{C}_{ig}^K$:

$$\dot{C}_{ig}^K = \frac{\sum_{j=1}^{K-1} \pi_{ij}^2 \sigma_{jg}^2 \cdot \mu_{Kg} + \pi_{iK}^2 (\tilde{C}_{ig}^K - C_{ig}^K)^2 \cdot \tilde{C}_{ig}^K}{\sum_{j=1}^{K-1} \pi_{ij}^2 \sigma_{jg}^2 + \pi_{iK}^2 (\tilde{C}_{ig}^K - C_{ig}^K)^2} \tag{3.19}$$

If we carefully examine the expressions of the weight coefficients in (3.19), we will have a better understanding of the oracle estimator $\dot{C}_{ig}^K$. The weight a^* contains the variance part of the reference subcomponents (from 1 to K-1), as well as their proportions in the mixture sample i. The weight b^* captures the deviation of the observed value $\tilde{C}_{ig}^K$ from its conditional mean C_{ig}^K, and the proportion of the unknown component π_{iK}. In cases when the variance(s) of the reference profiles are large, or when the proportions of the reference components are relatively large, it will give more weight to μ_{Kg}, the expression mean of C_{ig}^K, because the observed quantity $\tilde{C}_{ig}^K$ is less reliable. In contrast, for cases when the proportion of the unknown component is large, or when the deviation of the true expression value from its mean is large, the observed quantity $\tilde{C}_{ig}^K$ will be given more weights.

Though being a reasonable estimator, it is easy to realize that in real cases $\dot{C}_{ig}^K$ is not achievable due to the unknown part C_{ig}^K in b^*. After reorganizing $\dot{C}_{ig}^K$:

$$\dot{C}_{ig}^{K} = \frac{\sum_{j=1}^{K-1} \pi_{ij}^2 (\frac{\sigma_{jg}}{\sigma_{Kg}})^2 \cdot \mu_{Kg} + \pi_{iK}^2 (\frac{\tilde{C}_{ig}^{K} - C_{ig}^{K}}{\sigma_{Kg}})^2 \cdot \tilde{C}_{ig}^{K}}{\sum_{j=1}^{K-1} \pi_{ij}^2 (\frac{\sigma_{jg}}{\sigma_{Kg}})^2 + \pi_{iK}^2 (\frac{\tilde{C}_{ig}^{K} - C_{ig}^{K}}{\sigma_{Kg}})^2}$$

we recognize the unknown part $(\frac{\tilde{C}_{ig}^{K} - C_{ig}^{K}}{\sigma_{Kg}})^2$ can be replaced by its expected value of 1. Hence we arrive at our final estimator $\hat{C}_{ig}^{K}$:

$$\hat{C}_{ig}^{K} = \frac{\sum_{j=1}^{K-1} \pi_{ij}^2 (\frac{\sigma_{jg}}{\sigma_{Kg}})^2 \cdot \mu_{Kg} + \pi_{iK}^2 \cdot \tilde{C}_{ig}^{K}}{\sum_{j=1}^{K-1} \pi_{ij}^2 (\frac{\sigma_{jg}}{\sigma_{Kg}})^2 + \pi_{iK}^2}. \tag{3.20}$$

Since the expression value of the unknown subcomponent (instead of the other K-1 components) is usually of most interest, this concludes our individual deconvolution part for the multi-component mixture separation problem.

3.2.4 A New Gene Filtering Scheme

In the two-component setting of Section 2.2.2, we mentioned that the gene filtering step can be easily computed by a two sample t-test. For the multiple component case, each subcomponent will contribute some noise to the mixture expression, which adds to the difficulty of detecting the signals from the proportion coefficients π_{ik}'s. Therefore, we developed a more sophisticated gene selection scheme for the multiple component case. In summary, the strategy is to carefully select a set of representative genes/observations from the regression step to serve as the input for the whole FasTDK pipeline.

As we have seen from Section 3.2.2, the multiple linear regression is the first key to the success of overall estimation accuracy. To illustrate the idea, we rewrite the MLR problem in equation 3.7 as follows:

$$\mathbf{y}_i = \beta_{i1}\mathbf{x}_1 + \beta_{i2}\mathbf{x}_2 + \cdots + \beta_{ik}\mathbf{x}_k + \boldsymbol{\epsilon}_i \tag{3.21}$$

where $\mathbf{y}_i \in \mathbb{R}^{G\times 1}$, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_k] \in \mathbb{R}^{G\times k}$ and $i = 1, \ldots, S$. For each mixing sample i, the genes are the observations. Note that for each individual regression i, the input feature space $\mathbf{X} \in \mathbb{R}^{G\times k}$ does not change. So it is critical to ensure a good set of observations are selected for coefficient estimation.

Our first step is to compute the leverage score of each row/observation of $\mathbf{X}$ and compare it to $2p/n$, a general guideline to evaluate the leverage for a linear additive regression model (Goodall, 1993), where n is the number of observations and p is the number of parameters in the model. This step is expected to remove any influential observations which might be outliers, making the regression results more robust. Figure 3.1 illustrates the effect of removing the genes with high leverage scores in a real data set introduced in Section 3.4.1, treating the liver component as the unknown.

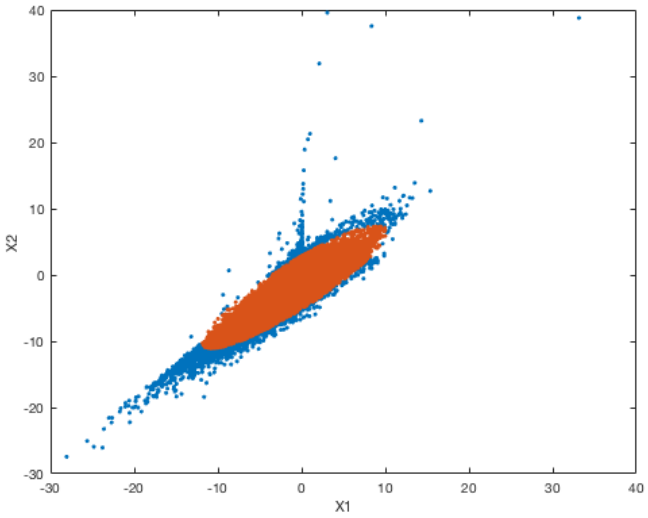


Figure 3.1: Visualization of low leverage observations in a 2-dimensional feature space. $X1$ and $X2$ are values for the two predictors in the multiple linear regression step of a FasTDK application. Each point is a gene. The whole genome is colored in blue while the genes with low leverage are colored in orange. This filtering step is designed to remove potential outliers.

Next, we carry over the concept of optimal experimental design to further refine the geneset. After removing the high leverage points, we are left with a subset of G' observations. Using this subset of observations, the linear model (3.2.4) for each sample i ($i = 1, \ldots, S$) becomes $\mathbf{y}_i' = \mathbf{X}'\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i$, where $\mathbf{y}_i' \in \mathbb{R}^{G' \times 1}$, $\mathbf{X}' \in \mathbb{R}^{G' \times k}$ and $\boldsymbol{\epsilon}_i \in \mathbb{R}^{G' \times 1}$ is an i.i.d Guassian noise vector with zero mean and finite variance. In typical experimental design problems, a small subset of observations $M \subset \{1, \ldots, G'\}$ of r rows from $\mathbf{X}'$ is selected, so that the regression efficiency is maximized and the total experimental cost is minimized. This new design $\mathbf{X}_M$ is often selected by maximizing the precision of estimating $\boldsymbol{\beta}$ and referred as an *optimal design*. Though in our

problem the experimental cost is not a major concern, the same goal is pursued: to minimize the variance of the estimated coefficients.

To minimize the variance of the coefficients is to minimize the covariance matrix of $\Sigma^{-1} = (\mathbf{X}_M^T \mathbf{X}_M)^{-1}$. As discussed in Pukelsheim (2006), various optimality-criteria have been developed as the invariants of this matrix to evalute how well Σ^{-1} is minimized on a selected design. What we have chosen is a popular criterion called 'D-optimality', which minimizes the determinant of Σ^{-1}. Figure 3.2 illustrates the effect of the D-optimal criterion and how the points are selected according to this criterion for the leverage filtered genes in Figure 3.1. We can observe genes surrounding the volume are selected rather than those around the origin. This is a desirable feature in our situation, because the genes around the origin not only contribute less signal to estimating the regression coefficients, but also bring more noise to the following optimization step, where we mainly focus on the second moment information. An additional benefit from this selection of a much smaller subset of observations is the reduced computational burden for the optimization step.

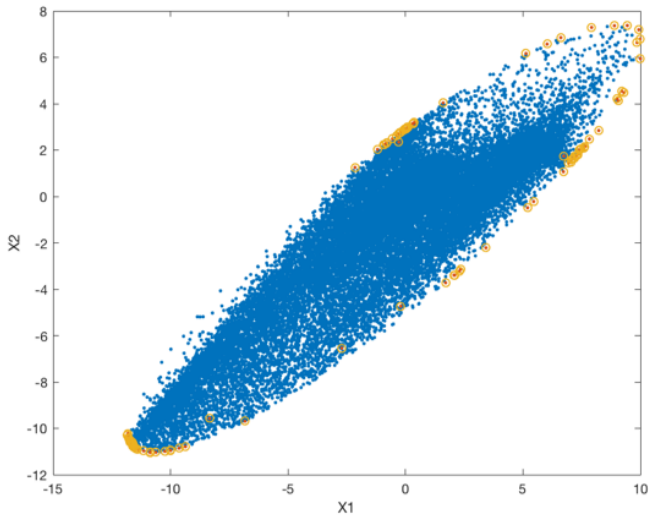


Figure 3.2: One hundred D-optimality genes (yellow circled) are selected in a real data application in a 2-dimensional feature space. The x- and y-axes are the row values for $\mathbf{X}' \in \mathbb{R}^{G' \times 2}$ and each point is a gene. The D-optimality criterion minimizes the determinant of Σ^{-1}, which tends to select points that are representatives in the sense of lying near the edge of the data set.

We utilize an R package 'AlgDesign' (Wheeler, 2004) to generate the D-optimal designs. This package creates exact D-optimal designs when provided with the number of trials needed, in which the Fedorov (1972) row exchange algorithm is implemented to select the optimal subset. In real

data applications, it is recommended to run the gene filtering steps we presented here and use this subset as the input for FasTDK.

3.3 Simulation Study

3.3.1 Performance Evaluation of Estimated Proportions

To simulate data that can better present biological scenarios, we utilize the sample means and standard deviations from real expression data for simulation. The dataset should be gene expression data collected in a pure cell type, ideally originally from a mixture. The number of samples might be limited due to this requirement, still the sample mean and standard deviation are useful to serve as parameters for our simulation purpose.

The real expression data we use here was downloaded online from the NIH GEO site with submission number GSE24223. The original study is a benchmark dataset in Grigoryev et al. (2010). This dataset includes transcriptome mRNA profiling of whole blood and purified CD4, CD8 T cells, B cells and monocytes in tandem with high-throughput flow cytometry in 10 kidney transplant patients sampled serially pre-transplant, 1, 2, 4, 8 and 12 weeks. Four cell type specific RNA probe expression values from the week 2 time point are selected, which has relatively more samples. The sample mean and SD parameters are generated from 7 CD4, 7 CD8, 5 CD19 and 3 CD56 samples.

The simulation design is as follows. Cell type CD4, CD8 and CD19 are treated as the known subcomponents in the mixture, whose sample mean and standard deviation are used to generate the reference profiles for each cell type (only the positive part from a Gaussian distribution). Cell type CD56 is assumed to be the unknown subcomponent in the mixture without reference profiles. In order to generate S mixing samples, S samples are first simulated independently for each of the four cell types according to a normal distribution. Then the four populations are combined according to a proportion matrix $\mathbf{\Pi} \in \mathbb{R}^{4\times S}$, whose columns are randomly and uniformly distributed in the standard 3-simplex space.

Our goal is to evaluate the proportion estimates obtained by our procedure, given that the true π_{ij}'s, $j = 1, \ldots, 4$, $i = 1, \ldots, S$ are known to us. In Figure 3.3 we show the estimation results for 100 mixture samples in one simulation run. Each point represents a mixing sample

and there are four subplots corresponding to four subcomponents. The x-axis values of the scatter plots are the proportions estimates acquired by our method and the y-axis values are the truth. A red $y = x$ reference line is added in each subplot to clearly highlight the approximation of equality. All proportion estimates align well with the red line, from Figure 3.3 (a) to (d). This demonstrates our method to be effective in estimating the mixing proportions not only for the unknown subcomponent, but also for the subcomponents with reference profiles.

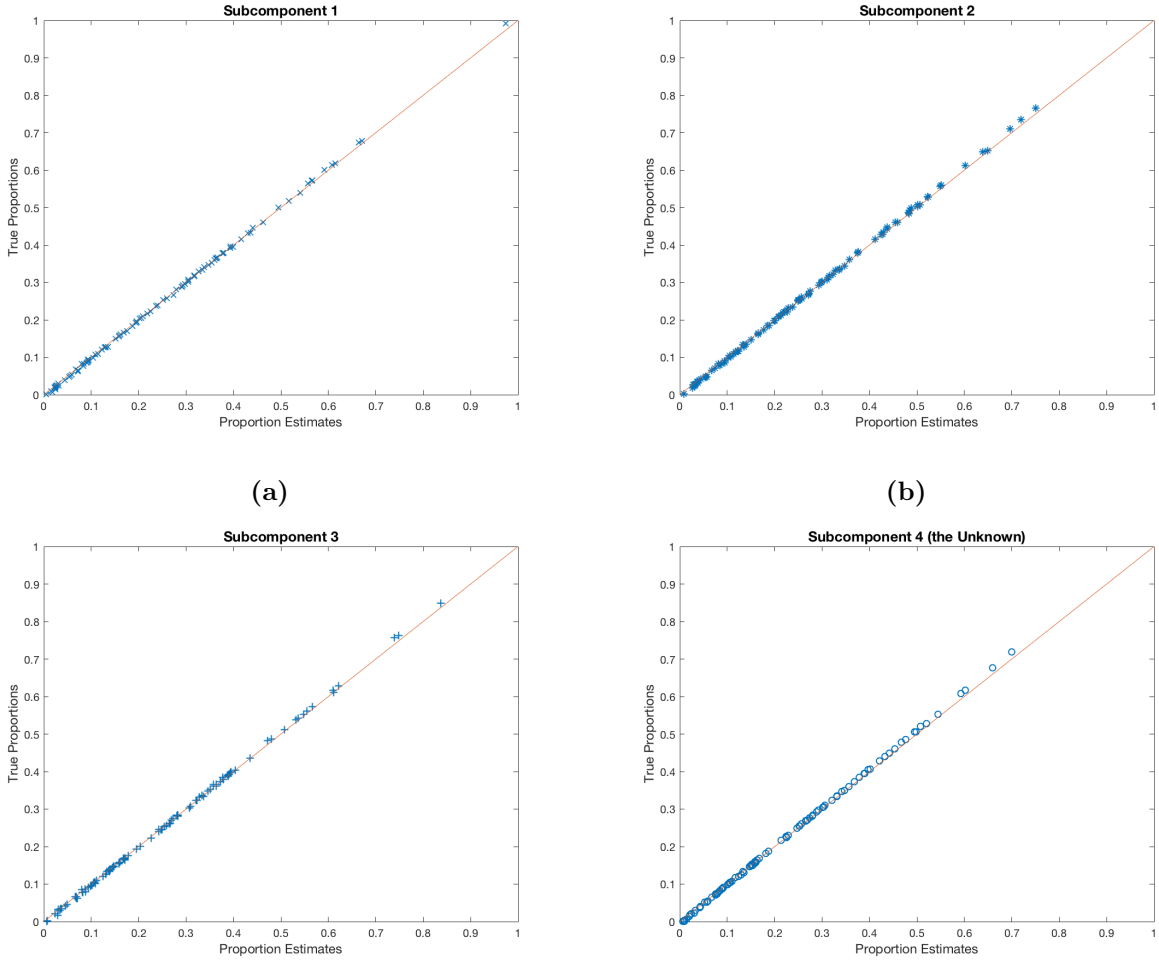


Figure 3.3: Scatterplot of 100 mixing samples' proportion estimates versus the truth for 4 subcomponents. Each point represents a mixture sample. The x-axis value captures the proportion estimate acquired by our method and the y-axis values are the ground truth. For all subcomponents (a)-(d), the 100 samples align well around the red $y = x$ line, which indicates our procedure is very effective.

3.3.1.1 Sensitivity Analysis

A key step in our procedure which can greatly affect the overall estimation accuracy is the performance of $\bar{\pi}_K$ estimation. Therefore, we want to investigate how its performance changes with different parameters inputs.

In the following simulation setting, we are particularly interested in how $\bar{\pi}_K$ changes with the number of mixing samples, S, and the standard deviation of the unknown component σ_{Kg}. We continue to utilize the sample means from the dataset GSE24223 and treat CD4, CD8 and CD19 as the known/reference subcomponents, and CD56 as the unknown. However, to simplify the analysis, we set expression the Standard Deviation (SD) to be the same for all genes. More specifically, SD = 0.3 for all genes from the reference subcomponents, which is about the median number of all SDs in the real data. But σ_K for the unknown components will be tested with different values in the experiment. We also keep the number of reference samples to be the same, 100, for the known subcomponents, but test the effect of different numbers of mixture samples on the $\bar{\pi}_K$ estimation.

The performance of $\bar{\pi}_K$ estimation is evaluated by the Mean Absolute Error (MAE), which measures the absolute deviation of the estimates from the truth. Then we designed nine simulation settings given 3 different number of mixing samples (N = 50, 100 or 200), and 3 different SDs ($\sigma_K = 0.1, 0.3$ or 0.6). There are 100 Monte Carlo runs for each setting. Figure 3.4 shows us that there is a general trend that the more mixture observations, the better performance we will gain for estimating the unknown proportion average $\bar{\pi}_K$. This makes sense because given the same amount of information about the references, we are provided with more information for the unknown part. Perhaps surprising, the performance does not always improve for smaller sigma because the blue curve ($\sigma_K = 0.1$) is above the red one ($\sigma_K = 0.3$). This result motivates us to investigate more on the effect of σ_K on $\bar{\pi}_K$ estimation.

In the new settings, σ_K changes from 0 to 0.9 with an increase of 0.1, given the number of mixture samples is 100 or 500 (Figure 3.5). Consistent with the previous observations in Figure 3.4, more mixture samples give better accuracy as the blue curve is generally under the red curve. However, we observe a parabola shape for $S = 100$ but less so for $S = 500$. This suggests there exists some bias when σ_K is very small, but this bias decreases as sample number increases.

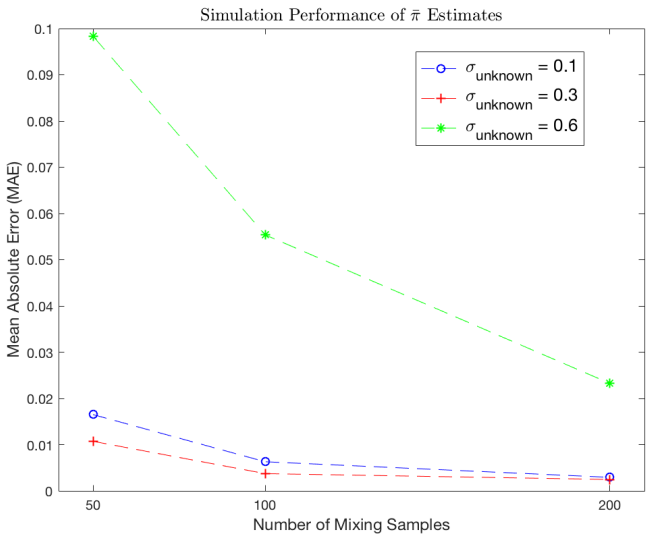


Figure 3.4: Simulation Results for 100 runs studies how proportion estimation accuracy is driven by the number of mixing samples and the variance of the unknown component. The decreasing trend towards the right side shows more mixing sample improves accuracy of the unknown subcomponent proportion estimation. But the estimation performance is best when the unknown SD is at a middle value (= 0.3 in red).

3.4 Empirical Dataset Validation

To test our method with real world data, we look for a gene expression dataset with mixture samples consisting of multiple homogeneous tissue/cell-type subcomponents, and with reference profilings available.

3.4.1 Dataset GSE19830

The microarray expression dataset GSE19830 of Shen-Orr et al. (2010) serves well for our validation purpose. This dataset was designed to test the feasability and sensitivity of statistical deconvolution. It not only has mixture samples and reference profiles, but also has ground truth for the mixing proportions. In the original experimental design, tissue samples from the brain, liver and lung of a single rat were mixed at the cRNA homogenate level in different proportions (shown in Table 3.1 as different Experiments.) Each experiment design, including the isolated pure subsets, were analyzed using expression arrays (Affymetrix) in triplicate. The overall dimension of the dataset we have been working with is 31099 probes and 36 samples.

Dataset GSE19830 was originally generated to compare the measured gene expression pattern of each mixed sample with the reconstituted pattern, simulated from pure tissue samples multiplied

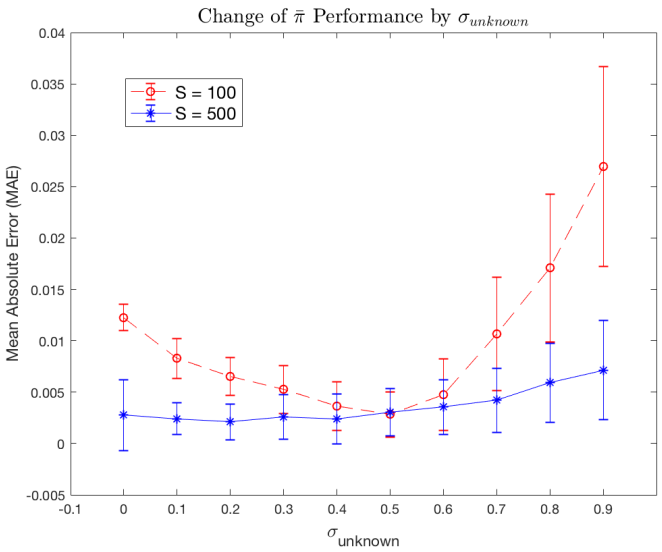


Figure 3.5: Simulation Results for 100 Monte Carlo runs studies how proprotion estimation accuracy is driven by the variance of the unknown component. The bias generated by small σ_K can be reduced when the mixing sample number increases.

by the proportions of the pure tissue in the sample. Figure 3.6 demonstrates in this dataset, the measured expression pattern can be effectively reconstituted *in silico*. The y-axis of Figure 3.6 is the measured mean expression pattern of mixing samples in Experiment #4 with 70% Liver, 5% Brain and 25% Lung tissues. It is plotted against the mean expression pattern reconstituted proportionally from the expression values obtained from the pure tissue samples. In both subplots, we see a large density of probes (represented as points and colored as yellow) are aligning well with the y=x diagonal line with high Pearson correlation coefficients.

Zhong and Liu (2012) states that the convolution of subcomponents is more effective when done in the linear space before applying the log2-transformation to the raw expression data. This is illustrated in Figure 3.6, where (a) shows the disruption of the log2-transformation applied before reconstitution, where a fraction of probes are underestimated using the reconstituted values. As expected from their work, this disruption is not observed when the convolution process is instead performed before log-2 transformation (Figure 3.6 b). Hence in our analysis, the raw expression data is the input for modeling and log-2 transformation is only used for visualization purposes.

Experiment #	Number of Replicates	Tissue Type	Liver	Brain	Lung
1	3	Pure	100	0	0
2	3	Pure	0	100	0
3	3	Pure	0	0	100
4	3	Mixed	70	5	25
5	3	Mixed	25	70	5
6	3	Mixed	70	25	5
7	3	Mixed	45	45	10
8	3	Mixed	55	20	25
9	3	Mixed	50	30	20
10	3	Mixed	55	30	15
11	3	Mixed	50	40	10
12	3	Mixed	60	35	5

Table 3.1: Summary of dataset GSE19830. For each experiment design, cRNA from liver, brain and lung tissue samples derived from a single rat were extracted and mixed in different proportions(%) with 3 replicates.

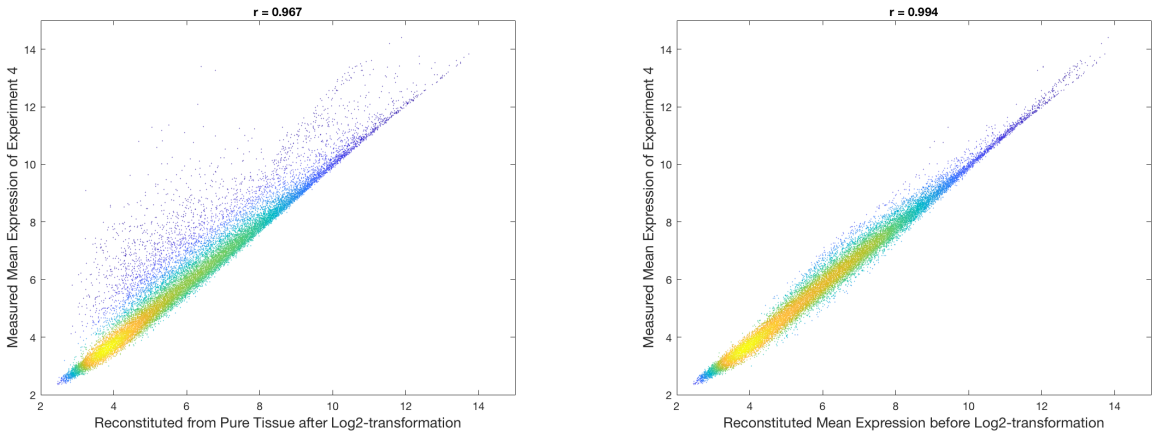
3.4.2 Solving Mixing Proportions using Simple Linear Regression

In Shen-Orr et al. (2010) a simple linear regression model is applied to deconvolve the expression profile of each subcomponent in the mixture samples, where mixture ratios are known parameters. Since our interests are more on the mixing proportions, we want to test how well this model performs if the tissue-type specific expression profiles are provided to estimate the proportions. Using notation consistent with ours, the model is written as:

$$\mathbf{y}_i = \pi_i^{liver}\boldsymbol{\mu}^{liver} + \pi_i^{brain}\boldsymbol{\mu}^{brain} + \pi_i^{lung}\boldsymbol{\mu}^{lung} + \boldsymbol{\epsilon}_i \tag{3.22}$$

where $\mathbf{y}_i$ is the response vector for sample $i = 1, \ldots, 27$, $\boldsymbol{\epsilon}_i$ is the random error vector, and $\boldsymbol{\mu}^{liver}, \boldsymbol{\mu}^{brain}, \boldsymbol{\mu}^{lung}$ are the predictors. The predictors are assumed to be known, which can be estimated by the mean expression value of the pure tissue samples.

The goal is to use simple linear regression to estimate the mixing proportions π_i^{liver},π_i^{brain}, π_i^{lung} for each mixture sample. Note that to obtain these mixing proportions, the least-square regression coefficients are first obtained. Any negative coefficients are set to zero. Then the non-negative coefficients are normalized so that they sum up to one for each sample. These normalized



(a) Convolution modeled at log-2 data level **(b)** Convolution modeled at raw data level

Figure 3.6: Measured gene expression pattern in a heterogeneous mixing sample can be modeled as the weighted sum of gene expression derived from pure tissue samples. The y-axis is the measured expression pattern/mean of mixing samples with 70% Liver, 5% Brain and 25% Lung tissues. The x-axis is the expression pattern reconstituted proportionally from the mean expression values obtained from the pure tissue samples. Each point represents a probe. Color represents point density from a single probe (purple) to lots of probes (yellow). This plot shows that the expression pattern/mean in the mixtures behaves similarly as the values reconstituted from pure tissue samples. The fraction of probes that deviate from the diagonal line suggests reconstitution is better done at the raw data level, *i.e.* before log2-transformation.

coefficients are the final version of estimates for the mixing proportions, which are plotted in Figure 3.7.

Figure 3.7 shows the effectiveness of model (3.22) in estimating the mixing proportions. The predictors in the regression model are estimated from Experiment 1-3 in Table 3.1. The fact that model (3.22) is returning a good estimation of the mixing proportions, suggests it should be valid to assume the mean expression values obtained from the pure tissue can serve as the reference profiles to infer the mean values for the subcomponents in the mixtures. This is an important assumption to hold in order to test our method on this validation dataset.

However, the pure tissue mean expression values for each subcomponent are not always available. It can be very costly or technically infeasible to obtain isolated pure tissues for each individual subcomponents. The FasTDK method we developed in Section 3.2.1 allows one of the subcomponents to be unknown, i.e. without reference profiles from the pure tissues. Next we will test our FasTDK pipeline on the same dataset, assuming one of predictors in model (3.22) is unknown.

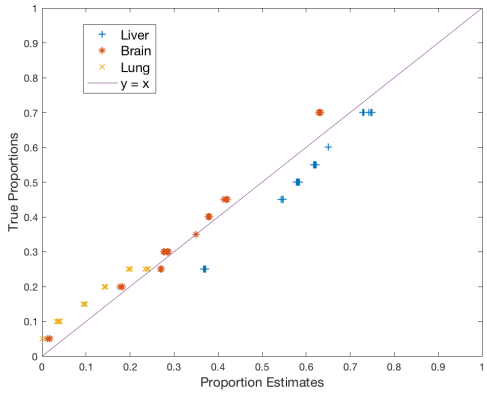


Figure 3.7: Deconvolved proportion estimates for Benchmark Dataset GSE19830 using simple linear regression in model (3.22). The x values are estimated while the y values are the true proportions. Each data point represents a mixture sample. All data points aligning well around the y=x reference line indicates the effectiveness of the model.

3.4.3 Estimating Mixing Proportions using FasTDK

Since all of the three mixing components in the validation dataset GSE19830 have pure expression profiles, we can treat each of the liver, brain and lung component as unknown in turn. We summarize the application of the FasTDK tool on this dataset in two ways, depending on whether the unknown component is among the larger proportions in the mixtures.

3.4.3.1 When the major component is the unknown

Based on the mixing proportions in Table 3.1, for most of the 27 samples, liver is the major component in the mixtures. The average proportion of the liver component is 0.53, while it is 0.33 for brain and 0.13 for lung. In the setting of tumor genomic deconvolution, one would expect the tumor component to be both the unknown and the major component (tumor samples are contaminated by other cell types). Hence, we start with validating FasTDK on the dataset GSE19830 by treating the liver component as the unknown.

As suggested in Section 3.2.4, a subset of the D-optimal genes is first selected before running FasTDK. Then FasTDK outputs the estimated proportions for all three subcomponents in the mixtures. These estimates are plotted against the ground truth (Figure 3.8). We observe a good correlation between the FasTDK estimates and the ground truth with a concordance correlation coefficient greater than 0.9. This demonstrates the FasTDK procedure is effective in estimating the

proportions for multiple subcomponent mixtures, in the case where the unknown component is the major component in the mixtures. We expect the constraint of 'major component' can be relaxed when the number of mixing samples increases.

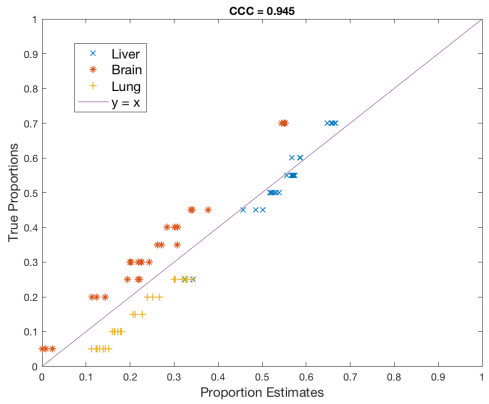


Figure 3.8: Deconvolved proportion estimates for the Benchmark Dataset GSE19830 when liver is the unknown component. The x values are proportion estimates while the y values are the ground truth. Each data point represents a mixing sample. All data points are aligning well around the $y = x$ reference line. This alignment is quantitatively evaluated by the Concordance Correlation Coefficient (Lawrence and Lin, 1989): 0.95 , which indicates the effectiveness of our method.

3.4.3.2 When the unknown component is among the smaller proportions

We then apply the FasTDK pipeline to the other two scenarios: treating in turn each of the brain and the lung component as the unknown. According to Table 3.1, the proportions of these two components in most of the mixing samples are relatively small. To our surprise, the FasTDK estimates for both scenarios are poor, with ccc values around 0. Figure 3.9 investigates this by studying the intermediate estimates. The horizontal axis are the proportion estimates using the $\boldsymbol{\beta}_{ik}$'s in (3.8) from the regression step of FasTDK multiplied by the true $\bar{\pi}_k$'s. When we examine the performance of the intermediate estimates $\boldsymbol{\beta}_{ik}$'s in the brain and the lung cases, we find out many of these regression coefficients are far off from the truth (Figure 3.9 b & c). This is particularly true when we compare these intermediate estimates in these two scenarios with that in the liver unknown case.

We suspect the performance of Figure 3.9 (b) and (c) is affected by the small proportions of the unknown component in the mixtures. When a subcomponent is relatively small in proportions, as a whole it contribute less signals to the observed mixed sample expression. However, based on the

linear additive assumption in (3.22), we expect the genes with relatively large mean expression value can offset this small proportion effect. For these types of genes, signals from the small proportion subcomponent are more detectable in the mixtures. Hence, we develop a simple but effective cut-off to select such genes. The effectiveness of this cut-off is not only supported by the improvement of our FasTDK estmation in Figure 3.10, but also supported by the goodness of fit between this dataset and its linear model, which we explore in more details in Section 3.4.3.3.

According to model (3.22), the observed $\mathbf{y}_i$ is a convex combination of the mean expression values of the three subcomponents. If the unknown component is among the small proportions and the mean expression values for the major component are available, only genes whose observed mean expression is larger than the mean expression of the major component will be selected. This cut-off helps to separate genes with strong signals from the minority groups.

We apply this cut-off to this dataset and separate the genes into two subsets. Let the set $A_g \subset \{1, \ldots, G\}$ denote the genes whose $\bar{y}_{\cdot g} > \mu_g^{liver}$, and the set $B_g \subset \{1, \ldots, G\}$ denote the genes whose $\bar{y}_{\cdot g} \leq \mu_g^{liver}$, where $\bar{y}_{\cdot g}$ is the sample mean expression of all mixing data for gene g and μ_g^{liver} is the mean expression for that gene in the pure liver component. We observed that use of set A_g as the initial input for the FasTDK pipeline greatly improves the performance of the proportional estimates (Figure 3.10).

3.4.3.3 Evidence supporting the cut-off criterion

Furthermore, we find that these genes in the mixing experiment are better fitted to the additive linear model (3.22) originally proposed for this dataset in Shen-Orr and Gaujoux (2013). Based on the model and the law of large numbers, for gene g, the observed sample mean of the mixtures can be approximated as:

$$\begin{aligned} \bar{y}_{\cdot g} &= \frac{1}{S}\sum_{i=1}^{S} y_{ig} \\ &\approx \bar{\pi}^{liver}\mu_g^{liver} + \bar{\pi}^{brain}\mu_g^{brain} + \bar{\pi}^{lung}\mu_g^{lung} \\ &\approx 0.53\,\mu_g^{liver} + 0.33\,\mu_g^{brain} + 0.13\,\mu_g^{lung}. \end{aligned} \tag{3.23}$$

Then let Δ_g denote the residual value of this approximation:

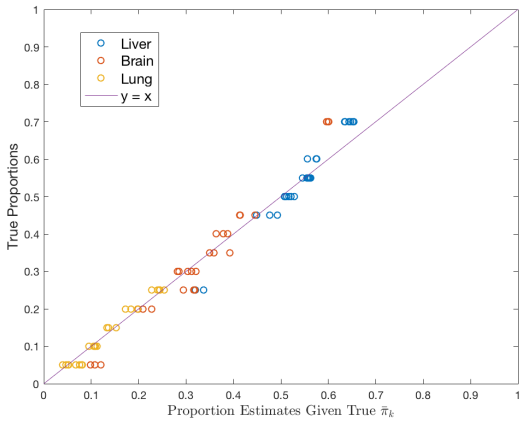
$$\Delta_g = \bar{y}_{\cdot g} - 0.53\mu_g^{liver} - 0.33\mu_g^{brain} - 0.13\mu_g^{lung}.$$

We expect the residual values for the whole genome to have mean zero. However, when we compare the residual values between gene subsets A_g and B_g, we observe great differences both in the overall ditribution (Figure 3.11) and in the mean values: 0.0055 for set A_g versus -0.329 for B_g.

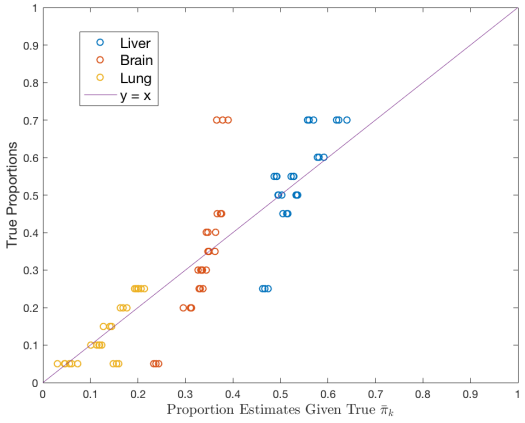
3.4.4 Conclusions

GSE19830 serves well as a validation dataset because it not only has mixed samples with known mixing proportions, but also includes expression profiles of pure tissue types. What's more, the linear additive model for the subcomponents has been tested in publication (Shen-Orr et al., 2010). The analyses on the validation dataset GSE19830 demonstrate that our method is effective in estimating the proportion coefficients for the constituent subcomponents in a mixture, especially when the unknown component is the majority subcomponent. Even if the unknown is in small proportions among the mixtures, a simple cut-off criterion introduced in section 3.4.3.2 can help to select a subset of genes with strong signals for the minority subcomponents to improve the estimation accuracy.

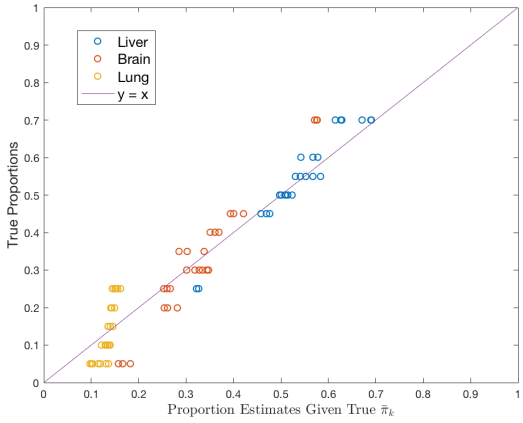
Based on the simulation study in Section 3.4, we observe that the larger the number of mixing samples is, the better the estimation accuracy will be. Considering the relative small sample size for GSE19830, we anticipate the performance of our FasTDK method to improve more once a larger size validation dataset is available.



(a) Liver Unknown

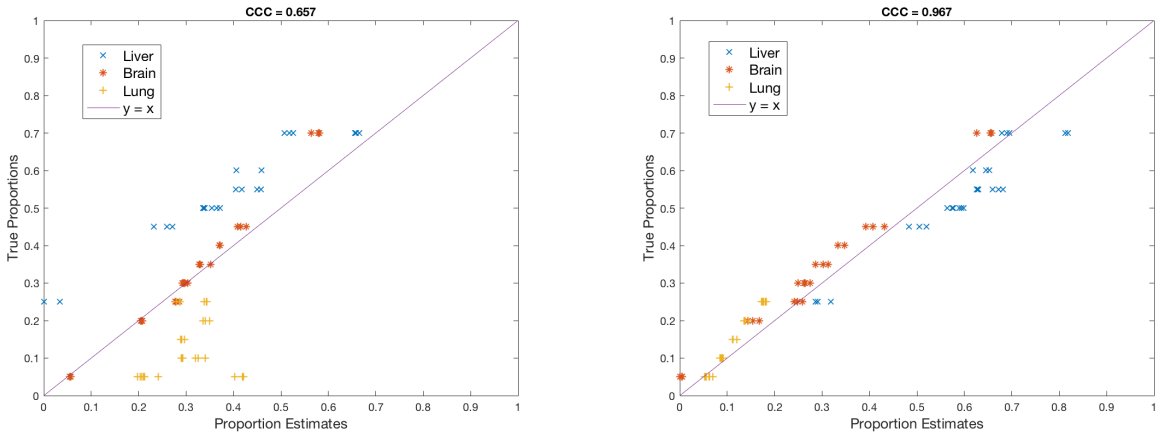


(b) Brain Unknown



(c) Lung Unknown

Figure 3.9: Quality check of GSE19830 intermediate estimates. Three scenarios are studied when the liver, brain and lung subcomponents are assumed to be unknown in turn. Each point is a mixing sample. The x-values are the FasTDK estimates computed by the intermediate regression coefficients $\boldsymbol{\beta}_{ik}$ and the true $\bar{\pi}_k$'s, according to (3.8), which are plotted against the true proportions on the y-axis. Comparing with the liver unknown case in (a), plots (b) and (c) suggest the final proportion estimates for both the brain and the lung cases are confounded by the biased intermediate results in the regression step.



(a) Brain Unknown **(b)** Lung Unknown

Figure 3.10: Performances of proprotion estimates using gene subset A_g when either the brain or the lung tissue is the unknown component. Each point is a mixing sample whose FasTDK estimates are the x-values and ground truths are the y-values. The performances are greatly improved using geneset A_g: CCC values increase from around 0 to 0.657 and 0.967 for the brain and the lung cases, respectively. This result suggests when the unknown component only occupies a small proportion in the mixtures, genes with stronger signals from the minority groups tend to give better coefficient estimation.

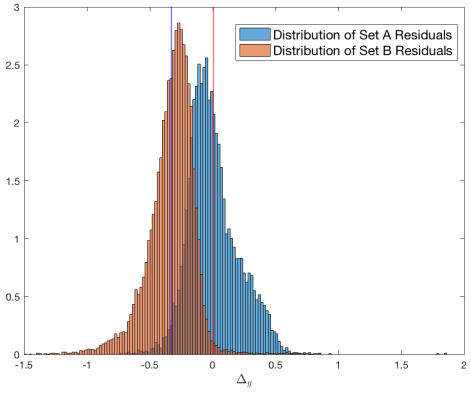


Figure 3.11: Normalized histograms of the residual values Δ_g for gene subsets A and B. The red vertical line indicates the mean value of Δ_g for the set A ($\bar{\Delta}_A = 0.0055$), while the blue line is that for the set B ($\bar{\Delta}_B = -0.329$). This suggests gene subset A is better fitted to the linear additive model (3.22) originally proposed for dataset GSE19830.

CHAPTER 4
Potential Work Beyond the Dissertation

In this dissertation work, we have presented two deconvolution tools that both produce estimates of the mixing proportions and the tumor-specific expression profiles for bulk tumor tissue samples. The variance and consistency of these estimators are yet to be derived and demonstrated as potential work beyond the dissertation.

The RNAseq gene expression data we used in testing our method is the count data. However, the RNAseq process pipeline has several choices of gene expression quantification measure, which includes counts, FPKM (Fragments Per Kilobase of transcript per Million mapped reads), RPKM (Reads Per Kilobase of transcript per Million mapped reads), and TPM (Transcripts Per Million), Conesa et al. (2016). A recent published paper suggests TPM to be the best value to use under the linearity assumption of tumor deconvolution analysis according to Jin et al. (2017). Thus it will be interesting to investigate the effect of using an alternative abundance measure in testing our deconvolution tools.

Our methods have currently only been tested on either the RNAseq expression datasets or the microarray datasets. There are many more types of molecular datasets whose interpretation might be altered by intratumor heterogeneity. For example, studies have shown tumor heterogeneity is associated with DNA methylation level in Varley et al. (2009). Thus it would be interesting to test our methods on other data types when the linearity assumption and the reference profile availability can be satisfied.

After individual datatypes are tested, some potential work beyond the dissertation may include investigating the effect of data integration, and whether the integration improves the accuracy of tumor deconvolution or clinical outcomes. In addition, how to apply tools developed in the context of intra-tumor heterogeneity to other mixture signals deconvolution process, is another interesting problem to explore.

BIBLIOGRAPHY

Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z., and Clark, H. F. (2009). Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS ONE*, 4(7).

Ahn, J., Yuan, Y., Parmigiani, G., Suraokar, M. B., Diao, L., Wistuba, I. I., and Wang, W. (2013). DeMix: Deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics*, 29(15):1865–1871.

Bernard, P. S., Parker, J. S., Mullins, M., Cheung, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenhush, J. F., Stijleman, I. J., Palazzo, J., Matron, J. S., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., and Perou, C. M. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of machine Mearning Research*, 3:993–1022.

Burrell, R. A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338–345.

Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2006). Measurement error in nonlinear models: a modern perspective.

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., Beroukhim, R., Pellman, D., Levine, D. A., Lander, E. S., Meyerson, M., and Getz, G. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, 30(5):413–421.

Chen, M., Zhang, J., Sampieri, K., Clohessy, J. G., Mendez, L., Gonzalez-Billalabeitia, E., Liu, X.-S., Lee, Y.-R., Fung, J., Katon, J. M., et al. (2018). An aberrant srebp-dependent lipogenic program promotes metastatic prostate cancer. *Nature genetics*, 50(2):206.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., and Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1):13.

DeNardo, D. G., Brennan, D. J., Rexhepaj, E., Ruffell, B., Shiao, S. L., Madden, S. F., Gallagher, W. M., Wadhwani, N., Keil, S. D., Junaid, S. A., Rugo, H. S., Shelley Hwang, E., Jirstrm, K., West, B. L., and Coussens, L. M. (2011). Leukocyte complexity predicts breast cancer survival and functionally regulates response to chemotherapy. *Cancer Discovery*, 1(1):54–67.

Erkkilä, T., Lehmusvaara, S., Ruusuvuori, P., Visakorpi, T., Shmulevich, I., and Lähdesmäki, H. (2010). Probabilistic analysis of gene expression measurements from heterogeneous tissues. *Bioinformatics*, 26(20):2571–2577.

Fedorov, V. (1972). Theory of optimal experiments.

Gaujoux, R. and Seoighe, C. (2012). Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: A case study. *Infection, Genetics and Evolution*, 12(5):913–921.

Goodall, C. R. (1993). 13 computation using the qr decomposition.

Grigoryev, Y. A., Kurian, S. M., Avnur, Z., Borie, D., Deng, J., Campbell, D., Sung, J., Nikolcheva, T., Quinn, A., Schulman, H., et al. (2010). Deconvoluting post-transplant immunity: cell subset-specific mapping reveals pathways for activation and expansion of memory t, monocytes and b cells. *PloS one*, 5(10):e13358.

Hesse, C. W. and James, C. J. (2006). On semi-blind source separation using spatial constraints with applications in EEG analysis. *IEEE Transactions on Biomedical Engineering*, 53:2525—-2534.

Jin, H., Wan, Y.-W., and Liu, Z. (2017). Comprehensive evaluation of RNA-seq quantification methods for linearity. *BMC Bioinformatics*, 18(S4):117.

Junttila, M. R. and de Sauvage, F. J. (2013). Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature*, 501(7467):346–354.

Lawrence, I. and Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268.

Moffitt, R. A., Marayati, R., Flate, E. L., Volmar, K. E., Loeza, S. G. H., Hoadley, K. A., Rashid, N. U., Williams, L. A., Eaton, S. C., Chung, A. H., Smyla, J. K., Anderson, J. M., Kim, H. J., Bentrem, D. J., Talamonti, M. S., Iacobuzio-Donahue, C. A., Hollingsworth, M. A., and Yeh, J. J. (2015). Virtual microdissection identifies distinct tumor- and stroma- specific subtypes of pancreatic ductal adenocarcinoma. *Nature genetics*, 47(10):1168–1178.

Mohammadi, S., Zuckerman, N., Goldsmith, A., and Grama, A. (2017). A Critical Survey of Deconvolution Methods for Separating Cell Types in Complex Tissues. *Proceedings of the IEEE*, 105(2):340–366.

Pukelsheim, F. (2006). *Optimal design of experiments*. SIAM.

Quon, G., Haider, S., Deshwar, A. G., Cui, A., Boutros, P. C., and Morris, Q. (2013). Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome Medicine*, 5(3):29.

Quon, G. and Morris, Q. (2009). ISOLATE: A computational strategy for identifying the primary origin of cancers using high-throughput sequencing. *Bioinformatics*, 25(21):2882–2889.

Saddiki, H., McAuliffe, J., and Flaherty, P. (2015). GLAD: A mixed-membership model for heterogeneous tumor subtype classification. *Bioinformatics*, 31(2):225–232.

Shen-Orr, S. S. and Gaujoux, R. (2013). Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Current Opinion in Immunology*, 25(5):571–578.

Shen-Orr, S. S., Tibshirani, R., Khatri, P., Bodian, D. L., Staedtler, F., Perry, N. M., Hastie, T., Sarwal, M. M., Davis, M. M., and Butte, A. J. (2010). Cell type–specific gene expression differences in complex tissues. *Nature Methods*, 7(4):287–289.

Shibata, D. (2012). Heterogeneity and Tumor History. *Science*, 336(6079):304–305.

Siegert, S., Cabuy, E., Scherf, B. G., Kohler, H., Panda, S., Le, Y.-Z., Fehling, H. J., Gaidatzis, D., Stadler, M. B., and Roska, B. (2012). Transcriptional code and disease map for adult retinal cell types. *Nature neuroscience*, 15(3):487.

Stuart, R. O., Wachsman, W., Berry, C. C., Wang-Rodriguez, J., Wasserman, L., Klacansky, I., Masys, D., Arden, K., Goodison, S., McClelland, M., Wang, Y., Sawyers, A., Kalcheva, I., Tarin, D., and Mercola, D. (2004). In silico dissection of cell-type-associated patterns of gene expression in prostate cancer. *Proceedings of the National Academy of Sciences*, 101(2):615–620.

TCGA Study Abbreviations (2019). Tcga study abbreviations. Accessed: 2019-04-11.

Varley, K. E., Mutch, D. G., Edmonston, T. B., Goodfellow, P. J., and Mitra, R. D. (2009). Intra-tumor heterogeneity of MLH1 promoter methylation revealed by deep single molecule bisulfite sequencing. *Nucleic Acids Research*, 37(14):4603–4612.

Wang, F., Zhang, N., Wang, J., Wu, H., and Zheng, X. (2016a). Tumor purity and differential methylation in cancer epigenomics. *Briefings in Functional Genomics*.

Wang, N., Hoffman, E. P., Chen, L., Chen, L., Zhang, Z., Liu, C., Yu, G., Herrington, D. M., Clarke, R., and Wang, Y. (2016b). Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Scientific Reports*, 6(November 2015):18909.

Wang, N. and Wang, Y. (2015). Unsupervised Signal Deconvolution for Multi-Scale Characterization of Tissue Heterogeneity. *Dissertation*.

Wang, Z., Morris, J. S., Cao, S., Ahn, J., Liu, R., Tyekucheva, S., Li, B., Lu, W., Tang, X., Wistuba, I. I., Bowden, M., Mucci, L., Loda, M., Parmigiani, G., Holmes, C. C., and Wang, W. (2017). Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration. *bioRxiv*.

Wheeler, R. (2004). Algdesign. the r project for statistical computing.

Ye, X. and Weinberg, R. A. (2015). Epithelial–mesenchymal plasticity: a central regulator of cancer progression. *Trends in cell biology*, 25(11):675–686.

Zhong, Y. and Liu, Z. (2012). Gene expression deconvolution in linear space. *Nature Methods*, 9(1):8–9.

Zhong, Y., Wan, Y.-W., Pang, K., Chow, L. M., and Liu, Z. (2013). Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics*, 14(1):89.