Semi-Parametric Mixture Gaussian Model to Detect Breast Cancer Intra-Tumor Heterogeneity

BY

DAN ZHAO
B.S., Sun Yat-sen University, Guangzhou, Guangdong, China, 2011
M.S., Yale University, New Haven, CT, US, 2014

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Public Health Sciences (Biostatistics)
in the Graduate College of the
University of Illinois at Chicago, 2019

Chicago, Illinois

Defense Committee:
Dulal Bhaumik, Advisor and Chair
Sanjib Basu, Biostatisics
Runa Bhaumik, Biostatistics
Peter Gann, Pathology
Supriya Mehta, Epidemiology

This thesis is dedicated to my parents,

who have given me the strength to be who I am and to be where I am;

who have made it possible for me to live life as one great adventurer.

ACKNOWLEDGMENTS

Like many great endeavors, the past five years would not have been possible without the love and support of many people. First and foremost, I am forever thankful to my advisor, Dr. Dulal Bhaumik for his guidance and constant encouragement to see me through. He has given me light to find my way, but always be there when I needed him. I have particularly enjoyed his insightful advice that stimulated my intellectual curiosity.

I want to extend grateful thanks to my thesis committee member, Dr. Peter Gann. I first had the chance to do statistical research in his lab; he supervised several research projects and played a significant role in convincing me to complete my Ph.D. dissertation on this topic.

I also would like to thank Dr. Supriya Mehta, for her continuing encouragement and support in every aspect of my academic development. I learned a lot from having discussions with Dr. Sanjib Basu. His knowledge and enthusiasm lead me into the wonderful world of statistics. I also want to thank Runa Bhaumik; the research attitudes I learned from her during our collaboration are priceless to me.

I also thank my fellow students at UIC. Yuanbo Song, Neeraj Kumar, Chengbo Yuan, Julia Xiong, we shared passion and pain. I feel so lucky to be surrounded by such like-minded, caring people.

I am extremely grateful to my parents, they believe in me, many times more than myself, and encouraged me to move forward, even from many thousands of miles away.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF TABLES (Continued)

TABLE **PAGE**

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF ABBREVIATIONS

AHC	Agglomerative Hierarchical Clustering
AIC	Akaike Information Criteria
AJCC	American Joint Committee on Cancer
AR	Androgen Receptor
ARI	Adjusted Rand Index
ASCO	American Society of Clinical Oncology
BCC	Breast Cancer Prognostic Challenge
BCI	Breast Cancer Index
BIC	Bayesian Information Criteria
CAP	College of American Pathologist
CE Mark	Certification Mark
CGH	Comparative Genomic Hybridization
CI	Confidence Interval
CSC	Cancer Stem Cells
EM	Expectation-Maximization
EMT	Epithelial Mesenchymal Transition
ER	Estrogen Receptor

LIST OF ABBREVIATIONS (Continued)

FISH	Fluorescence in Situ Hybridization
GEPs	Gene Expression Panels
HER2	Human Epidermal Growth Receptor 2
HR	Hazard Ratio
ICL	Integrated Completed Likelihood Criterion
IDC NOS	Invasive Ductal Carcinoma, Not Otherwise Specified
IHC	Immunohistochemistry
ISH	In Situ Hybridization
KM	KaplanMeier
KRAS	K-ras or Ki-ras
LEC	Laplace-Empirical Criterion
LumA	Luminal A
LumB	Luminal B
LRTS	Likelihood Ratio Test Statistic
MAP	Maximum A Posteriori
MD	Mahalanobis Distance
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium

LIST OF ABBREVIATIONS (Continued)

MIR	Minimum Information Ratio criterion
MLE	Maximum Likelihood Estimation
MPS	Massively Parallel Sequencing
mRNA	messenger RNA
OS	Overall Survival
PAM50	Prosigna Breast Cancer Prognostic Gene Signature Assay
PCR	Polymerase Chain Reaction
PR	Progesterone Receptor
QA	Quality Assessment
RIN	RNA Integrity Number
RV	Random Variable
SNP	Single Nucleotide Polymorphisms
t-SNE	t-distributed Stochastic Neighbor Embedding
TCGA	The Cancer Genome Atlas
UICC	Union for International Cancer Control
WHO	World Health Organization

SUMMARY

Breast cancer intratumor heterogeneity challenges our ability to predict patients' outcomes or responses to targeted therapy; yet, available methods are limited to measure intratumor heterogeneity quantitatively. The goal of this research is to develop statistical methodologies for high dimensional PAM50 gene expression data to characterize the intratumor heterogeneity for better treatment option. In this dissertation, I propose two approaches for classification of intratumor heterogeneity: non-parametric clustering methods and finite mixture Gaussian method. For non-parametric clustering methods, I use Mahalanobis distance for classification. For finite mixture Gaussian method, as the parameters of these Gaussian mixtures cannot be estimated in closed form, so estimates are typically obtained via an iterative process, e.g. EM algorithm. However, finite mixture modeling can suffer from locally optimal solutions because of poor initial starting values. I improve EM in mixture Gaussian model by applying a simple and efficient initialization strategy based on Mahalanobis distance. This improved method allows the model to borrow information from data without any distributional assumption. The proposed model is illustrated with two real datasets from breast cancer patients, and also evaluated using simulated datasets.

CHAPTER 1

INTRODUCTION

Breast cancer is the first and foremost common cancer in female. Every year, 1.7 million women are newly diagnosed with breast cancer all over the world (80). It can cause huge health and economic burden in developed countries and developing countries (214). The 5-year survival rate for breast cancer has wide variation (185), largely due to the heterogeneity in the nature of breast cancer. In an era of personalized medicine, patients with different types of breast cancer may response differently to certain types of treatments, in addition, within the same type of breast cancer, patients may react differently and have different risks of death or relapse (136).

Currently, breast cancer is recognized as a heterogeneous disease, mainly due to clinical and morphologic variation between patients (intertumor heterogeneity), and also due to the morphologic and genomic variation within a single patient (intratumor heterogeneity) (172; 138). The tumor heterogeneity in breast cancer creates diagnostic and therapeutic challenges, thus becoming an obstacle for the development of precision treatment for women with breast cancer (86).

A classification mechanism for capturing biological features of breast cancer heterogeneity may provide a guidance of treatment administration strategies. Intertumor heterogeneity, defined as heterogeneity between patients, has long been recognized by clinicians and pathologists. The distinct character of breast cancer intertumor heterogeneity can be realized through

various methods. Traditionally, pathological examination at the morphological level can provide a histological grade, which includes the information of differentiation level and nuclear pleomorphism (178; 28). The subtype of breast cancer based on morphological characters can be further divided based on molecular signatures, for example, expressions of protein biomarkers or immunohistochemical classification (232; 151). Differential expressions of these protein biomarkers, i.e. estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth receptor 2(HER2), provide a clinically meaningful classification for breast cancer and direct underlying treatment application strategy (94; 233). Rapid advancement in gene sequencing and microarray analysis lead to gene expression based classification. One of the most widely used approaches in the clinical assessment of breast cancer is PAM50 method from ARUP Laboratories (174; 162). Five distinct subtypes were defined: Basal, Luminal A, Luminal B, human epidermal growth factor receptor 2 related (HER2), and Normal. Each subtype will have distinct prognosis and will receive different treatment.

The aforementioned methods focus on intertumor heterogeneity, assuming that each patient belongs to a single discrete class, even though intratumor heterogeneity is acknowledged with evidences (6; 149; 115). Current method to define intratumor heterogeneity is not completely understood (163; 38; 47). Current methods to measure intratumor heterogeneity include genome-wide measurements of gene expression or chromosome copy number in tumors. These methods can provide quantitative data that is capable of capturing global genomic intratumor heterogeneity. However, these methods require comprehensive sequencing of the tumor genome,

but the costs and time required for whole genome single cell sequencing of tumors restrict its wide clinical application (18; 164; 8; 161).

The objectives of this dissertation are to provide statistical methodologies including analysis, inferences and model selection for intratumor heterogeneity identification based on gene expression data. I would like to develop a statistical model for evaluating the intratumor heterogeneity by the PAM50 gene expression profile. In this work, I first apply a non-parametric clustering method on PAM50 gene expression data. I implement Mahalanobis distance to measure the level of purity in breast cancers designed as Luminal A. The reason for focusing on Luminal A cases is as hypothetically, admixture with any other intrinsic (Luminal B, HER2-enriched, Basal) would be connected to more aggressive disease and worse outcomes. I classify subjects based on their purity and then compare the differences of clinical characteristics and overall survival time.

Besides the non-parametric method, I would like to apply a formal statistical model to study the nature of intratumor heterogeneity, as it can provide a robust statistical procedure that allows the selection of optimum model, thus eliminating the ambiguity associated with non-parametric model. One of the most widely used parametric methods for clusters analysis is finite mixture Gaussian method. For finite mixture Gaussian model, there exists latent variable, thus the parameter can not be estimated numerically, I carry out the Expectation-Maximization (EM) algorithm for parameter estimation. To implement EM algorithm in mixture Gaussian model, one critical issue is: choose initial parameter values to pass on to the EM algorithm. It is critical for EM algorithm, as EM does not necessarily guarantee by convergence to the global

maximum, and the ability of finding global optimum depends on the initialization point of EM algorithm (127; 234; 90). Besides, the initial point of EM will determine how fast EM converges (127; 234; 90). Traditionally, one way to start EM is to use randomly generated value, and search the whole landscape, without using any prior information from data (116; 112). However, this strategy is too time consuming to apply to high-dimensional datasets. Another initialization strategy is to choose the starting value suggested by the data, for example, the agglomerative hierarchical clustering proposed by Fraley and Raftery (81; 190).This initialization strategy in general is more efficient than uninformed technique; unfortunately, this method has been criticized for using just one set of starting value, which restricts the search for global optimum. Furthermore, certain strategy tends to favor specific shapes or patterns. To overcome these limitations, I plan a semi-parametric way to initialize EM algorithm. The method can incorporate the information from data, but not imposing any restrictions on the form of the distribution of data. Moreover, this method is able to provide multiple initial values, such that we will have better chance of finding the global optimum.

Another area of research related to finite mixture Gaussian is model selection to find the number of clusters (components) and the optimum model structure. Selection of optimum number of components is one of the most critical problems of mixture model. One unique feature for finite mixture Gaussian is that we need to select variance covariance structure. To achieve the parsimony for variance covariance matrices, constraints are imposed such that different constraints on the covariance matrix provide distinct models with similar geometric properties. One of the most widely used constraints come from Banfield and Raftery (12); there

are three types of variance covariance structure, namely: spherical family, diagonal family and general family. I use model selection technique to determine the optimal number of components and covariance structure simultaneously, by finding the besting fitting model. I study four most commonly used model selection techniques: Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), Integrated Completed Likelihood criterion (ICL) and likelihood ratio test statistic (LRTS).

The remainder of this thesis is organized as follows: In Chapter 2, I describe the concept of breast cancer heterogeneity and the causes of cancer heterogeneity. I also introduce different methods to classify intertumor heterogeneity and intratumor heterogeneity, in the last section of chapter 2, I provide details about subject recruitment, data collection, data reprocessing as well as the data structure. In Chapter 3, I first present the non-parametric cluster method and its application to real data. In second part of Chapter 3, I describe a procedure for fitting mixture Gaussian model and introduce the initialization strategy I proposed. I also describe the model selection techniques used for finite mixture model. In Chapter 4, the proposed method is evaluated with other methods in simulated data. In Chapter 5, I present the results obtained from real data analysis, the proposed method is evaluated by model fitting and clinical outcomes. Finally, in Chapter 6, I summarize the work and discuss its limitations and future applications.

CHAPTER 2

BACKGROUND

Breast cancer is a major health problem in the United States and worldwide. There are 249260 new cases of invasive breast cancer and 40890 deaths caused by breast cancer in 2016 based on The American Cancer Society estimation (157; 221; 196; 223). Breast cancer is the second cause of cancer-related death in the United States,with approximately 39620 deaths among women, constituting about 14% of all cancer-related deaths among women. It has become clear that breast cancer can display astonishingly distinct morphological, behavioral and genetic variability, based on accumulating evidences of heterogeneity in breast cancers (145; 147; 28).

The heterogeneity in breast cancer can be mainly divided into intertumor and intratumor heterogeneity. Intertumor heterogeneity, by definition, refers to variation between patients with the same histological type. Intertumor heterogeneity has been recognized by physicians and researchers long time ago, the cause of intertumor heterogeneity is believed to be patient-specific characteristics such as genetic mutation, somatic expression profile and environmental factors (137; 49; 206; 155).

Intertumor heterogeneity can be illustrated by clinical disease staging based on physical examination and imaging (178; 228). The classification of malignant tumors (TNM) system proposed by the American Joint Committee on Cancer (AJCC)/Union for International Cancer Control (UICC) incorporates the size of the primary tumor, the nearby regional lymph nodes,

and the spread of cancer from one part to the other part of body (207; 70).Traditional, breast cancer treatment decision is made based on the tumor characteristics, such as histopathologic features and biomarker profile. Treatment decision can be affected by patients age, menopausal status, and general health as well (166). The aforementioned demographic characteristics have a profound impact on patients' response to treatment, and contribute significantly to clinical outcome differences (73).

2.1 Breast cancer intratumor heterogeneity

Intratumor heterogeneity, referring to variations within a single patient, was first observed by histopathologists as sections of different morphology or staining behavior (79; 220). However, intratumor heterogeneity is relatively unexplored compared to intertumor heterogeneity, partially due to the challenges in quantitatively characterizing and measuring intratumor heterogeneity.

In the past decades, there have been increasing reports of intratumor heterogeneity of gene expression and DNA mutations (24; 187). In 1800s, Rudolf Virchow reported the morphoogical heterogeneity of malignant cells within individual tumor (36). The development and progress of model techniques such as cell-staining and gene sequencing methods, subsequently made it possible for scientists to characterize intratumor heterogeneity (77; 215). Currently, intratumor heterogeneity is defined at the molecular level by the genetic profile differences observed among individual malignant cells. Cancer cells within a given patient may differ in phenotype, as well as in genotype (205).

Breast caner intratumor heterogeneity can materialize as morphologic and biochemical variation. This variation in tumor phenotype provides the evidence for tumor evolution. The evolution of tumor was shown to play a significant role in disease progression and resistance to therapy (11; 71).

Intratumor heterogeneity can exist either between different geographical regions of a tumor (spatial intratumor heterogeneity), or as the evolution of a tumor over period of time (temporal intratumor heterogeneity). See in Figure 1.

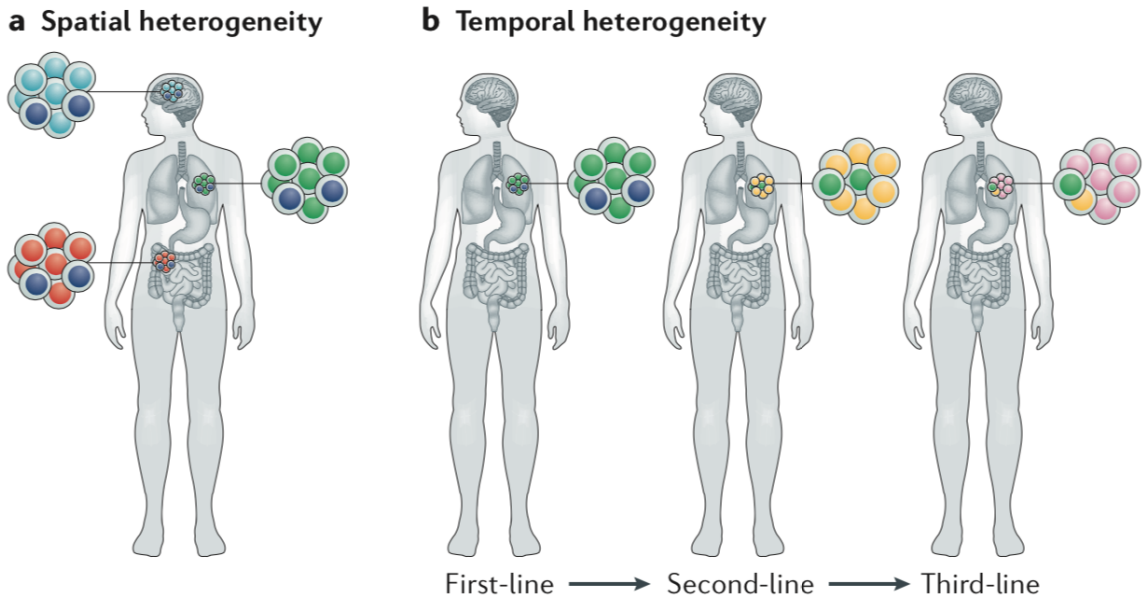

Figure 1. Spatial and Temporal intratumour heterogeneity.

2.1.1 Spatial heterogeneity

Spatial heterogeneity is the variation across different regions within a patient. In addition to the spatial heterogeneity with respect to histological features, genetic spatial heterogeneity has also been frequently observed. Comparative genomic hybridization (CGH) analysis of geographically separate regions of primary breast cancers has revealed various levels of genetic heterogeneity within a single tumor (144). Monogenomic tumor means the presence of one major subpopulation with a stable genome, while polygenomic cancer are characterized by presence of multiple genetically distinct subpopulations at the same or different locations (88). Another evidence for the existence of spatial heterogeneity comes from single cell sequencing analysis (143). The study involves multi-region sampling of tumors and indicated that many irregularly distributed passenger mutations are not expressed. This expression pattern affects the use of gene expression signatures, as biomarkers associated with better prognosis and other biomarkers associated with worse prognosis may be present in geographically distinct sub-regions within the same tumor (153; 144).

The characteristics of cancer cells at different sites within a tumor can be distinct, owing to the influences of micro-environment factors and site specific characteristics. Selective pressures lead to genetic evolution; thus spatial heterogeneity may be observed among the cells present within a tumor in a single anatomical region.

2.1.2 Temporal heterogeneity

Temporal heterogeneity is a term that refers to the variation in genetic diversity of a tumor over a period of time. Data from numerous studies that investigated biopsy sampling to

characterize tumors evolution has demonstrated that chemotherapy treatment can alter the molecular characteristics of tumor over time (146; 78). In particular, gene mutations are the foundation of cell replication and cycle regulation, and can contribute significantly to temporal heterogeneity. In one study, researchers found that treating glioblastomas using temozolomide can turn on transition mutations in MMR genes and lead to the development of a hypermutated phenotype (146). Precise medicine may bring more selective pressures on cancer cells than non-specific therapies such as chemotherapy. Because of that, many of the temporal heterogeneity are observed in the context of targeted therapies.

Tumors evolve over time between the primary tumor before a treatment and local or distant recurrences after treatment. Genomic analysis of the differences between primary breast tumors and their metastases have been systematically documented (215; 158; 31). While metastases tend to be substantially similar to their primaries in terms of genetic alterations, 31% of primary breast cancers and their metachronous metastases displayed significant differences in gene copy number, as revealed in the study by CGH and FISH (111; 158).

A recent study suggested that the primary tumor appeared to have more clonal diversity than the distant metastasis regarding the mutations frequencies and gene expression structures, suggesting chemotherapy treatment and micro-environment may lead to the temporal heterogeneity (62). Several studies suggest that metastases may evolve from a small number of clones at primary tumor (86; 193). These studies also suggest that the abnormal growth of primary tumors and their metastases may go through parallel and independent evolution (126).

2.2 Causations of intertumor and intratumor heterogeneity

Two major theories have been proposed to explain the intratumor heterogeneity: clonal evolution and cancer stem cells (CSC) (131; 89). These two theories were originally thought to be mutually exclusive, but now they are considered complementary (192). Both theories acknowledge that tumor is initiated from single cell with multiple molecular alterations that allow potentially unlimited proliferation; these two theories also assume that the micro-environment has a significant contribution to tumor evolution (126). Despite that, these theories have important differences and will be discussed below.

2.2.1 Clonal evolution theory

The clonal evolution theory was proposed by Nowell in 1976 (154), stating that a single cell can lead to tumor growth through an accumulation of mutations. The continuous change in the genetic instability results in a sequential selection of diverse subpopulations with the presence of more aggressive phenotypes (131). Further more, each subpopulation can mutate separately, thus contributing to intratumor heterogeneity (37; 154; 126). Two types of clonal evolution have been established: linear evolution and branched evolution; see in Figure 2. In linear evolution, a sequential acquisition of mutations leads to clones that are more adjusted to the environment than their predecessors (225; 30). The heterogeneity level through linear evolution is relatively low: heterogeneity happens only when a new clone has not completely outgrown its predecessor. Contrary to linear evolution, the branched evolution assumes that different sub-clones coexist and evolve simultaneously. This evolution can be related to a tree growing branch, where the root is the original clone and various tree branches represent

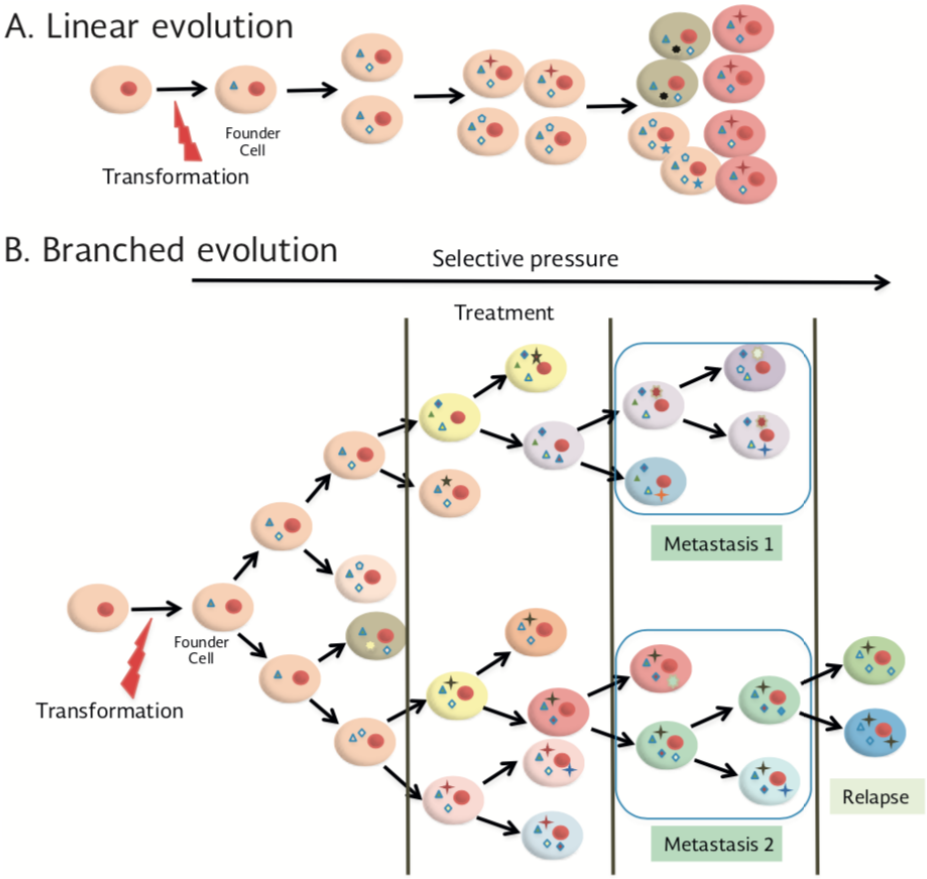

Figure 2. Clonal evolution theory

different sub-clones with diverse accumulated mutations that are separated geographically. The newly arising subpopulations conserve the founder cells' nature and acquire new phenotypic and genetic features that help them to survive better under selective pressures. One thing worth mentioning is that mutational, environmental, and selective pressures can affect the clonal evolution by selecting the fittest clones, thus leading to clonal expansions. Since branching evolution generates greater diversity, hence it may have more contribution to heterogeneity compared to linear evolution. Breast cancer intratumor heterogeneity through clonal evolution has been observed, for HER2-positive tumors; different patterns of HER2 gene amplification have been observed in different regions of a tumor (88).

Furthermore, following Darwinian rules that the most fit clones will progress, tumor progression depends on population size, mutation rate, and selective pressures from micro-environment and/or external factors (217). During tumor growth, subpopulations evolve and acquire mutations that may increase drug resistance to specific therapies. Once this happens, specific subpopulations become the prevalent proportion of the tumor since they are better fitted to the environment (87; 237). In breast cancer, mutations in gene expression due to neoadjuvant treatment with the aromatase inhibitor have been explored (140). This study illustrates that neoadjuvant affects proliferation and expression patterns of ER-related genes (139). In another study, Miller sampled from two time points: before and after treatment with letrozole (171; 170). From the results of whole-genome sequencing, Miller demonstrated that the inhibitor treatment induces evolving of the clonal populations with the acquisition of new enrichments compared with the pre-existing ones.

2.2.2 Cancer stem cell theory

The assumption of clonal evolution model is that all cells within the tumor have the same potential to promote tumor progression. On the contrast, the assumption for Cancer stem cell (CSC) theory is that a tumor evolves from a rare small population of CSCs that are capable of self-renewal, and the CSCs lose stemness ability by differentiating into non-CSC, thus ending with several subpopulations with new characteristics as shown in Figure 3.

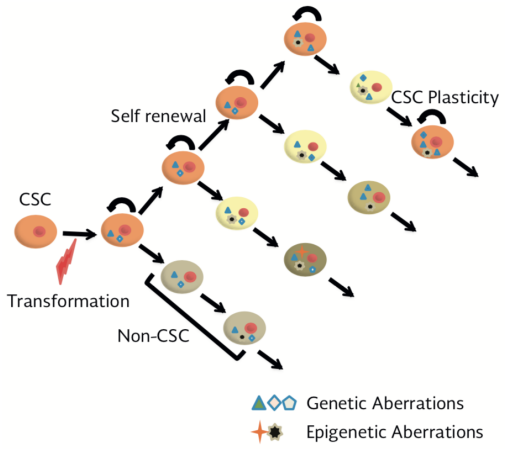

Figure 3. Cancer stem cell theory

This abnormal differentiation capacity of CSCs is considered to be responsible for intratumor heterogeneity (171). The CSC theory was first observed in hematopoietic tumors; later, it was identified in solid tumors such as breast and brain cancers. The combination of cell-surface

markers (CD44+/CD24-/low) was used in the identification of CSCs in several researches. The existence of CSCs in breast cancer was supported by a study, after injection CD44+/CD24-/low cells in a xenograft; this small amount of stem cells have been isolated from cultured breast cancer cell lines and were able to form a tumor (4).

In the CSC theory, the tumor is hierarchically structured; cells with high capacity of proliferation and self-renewal are placed in the highest order and named as CSCs (132). The CSCs have the ability to differentiate into a non-CSC cell, to promote uncontrollable tumor growth. Non-CSCs represent the majority of the tumor, but have less contribution to tumor growth (68). Drug resistance has been explained by CSCs theory: the CSCs provide the ability to resist cancer therapy, targeted therapies in turn, will induce the transformation of cancer cells into CSCs. It is supposed to have a balance between the CSCs and Non-CSCs, any imbalance may cause a shift to an enriched CSC that will likely result in an aggressive phenotype and poor prognosis (5).

Furthermore, several researches have demonstrated that epithelial-mesenchymal transition (EMT) contribute to the development of CSCs (107). The EMT has been comprehensively studied since EMT is one of the critical steps during tumor metastasis. During the process of EMT, the epithelial cells present reorganization of their cytoskeleton, losing their tight junction proteins and apicobasal polarity. The loss of epithelial markers is one of the main milestone of EMT: decreased expression of E-cadherin, claudins and occludins, and concomitantly increased expression of mesenchymal markers like N-cadherin, small muscle actin, and vimentin (93). Several studies have demonstrated the participation of these transcription factors in the process

of changes in cell migration and stem cell formation (123; 43). There are large number of evidences showing that the EMT is connected with the CSC phenotype, and EMT together with CSCs contribute to an aggressive phenotype (45; 141).

2.3 Methods to classify breast cancer intertumor heterogeneity

For many decades, breast cancer heterogeneity has been recognized by pathologists, who have classified tumors into variant histological subtypes (72).The heterogeneity observed from breast cancers introduces the concept that there is not just one disease, but a collection of distinct diseases of the breast and the cells composing the breast. The distinct nature and characteristics of these diseases can be identified and classified through traditional pathological examination at the morphological level, biomarker level as well as gene expression level.

2.3.1 Subtypes of breast cancer by histopathology

The morphologic heterogeneity of breast cancer constitutes the fundamental elements of the histopathologic classification of breast cancer. The most common subtype is invasive ductal carcinoma, not otherwise specified (IDC NOS), representing 80% of invasive breast cancers. However, IDC NOS is not well defined; the 2012 World Health Organization (WHO) classification defines IDC NOS by exclusion, as the heterogeneous group of tumors that fail to exhibit sufficient characteristics to achieve classification as a specific histological type (198). Invasive lobular carcinoma (ILC) is the second most common subtype, representing approximately 10% of invasive breast cancers. The less common subtypes include mucinous, cribriform, micropapillary, papillary, tubular, medullary, metaplastic, and inflammatory carcinomas (156; 48). Tubular, mucinous, and papillary carcinomas usually have superior clinical outcome compared

to IDC and ILC (183; 177). In contrast. metaplastic carcinoma and poorly differentiated IDC NOS have significantly worse clinical outcome (198).

2.3.2 Subtypes of breast cancer by IHC markers

These subtypes of breast cancer can be further classified based on their molecular signatures, i.e, expression of protein biomarkers or immunohistochemical classification. The expression of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) is assessed routinely in all invasive breast carcinomas by immunohistochemistry (IHC) according to the recommendations by American Society of Clinical Oncology/College of American Pathologist (ASCO/ CAP) (95; 233). The biomarkers provide a practical classification for breast cancer with clinical meaning, and there are well established prognostic and predictive factors; their expression in breast carcinomas is critical in guiding patient treatment (91; 97).

The HER2 biomarker is overly expressed in about 15-20% of primary breast cancer detected by IHC staining(198). HER2-positive breast cancers have the worst prognosis among all subtypes of breast cancers, however, HER2-positive patients show well response to anti-HER2 therapy (e.g., trastuzumab, lapatinib) (59; 50).

Breast cancers that have no PR, ER or HER2 expression, are defined as triple-negative breast cancers, usually represent the extremely heterogeneous groups histologically, genetically and prognostically. Arising evidences suggest that nuclear expression of the androgen receptor (AR) can be detected in approximately 12-55% of triple-negative breast cancer (175; 92; 16).

The prognostic significance of AR expression in triple-negative cancers is contentious, but it was shown to be associated with improved survival in other tumor subtypes (159).

Hundreds of newly developed biomarkers have been introduced into breast cancer for potential diagnostic, prognostic, and therapeutic implications. However, the association between tumor heterogeneity and biomarker expression remains unclear. A systematic approach and standardized biomarkers are required to better guide therapeutic decisions.

2.3.3 Subtypes of breast cancer by gene expression

The rapid developments in technology and data processing have yielded increasing knowledge about the molecular heterogeneity in cancer cells. The most highly cited gene expression studies targeted at identifying the mutations that are capable of prognosis prediction (221) or propensity to metastasize (226). Other studies, such as Sorlie (200), provide tumor classification based on expression of a group of "predefined intrinsic genes" (168). In these kind of studies, five distinct subtypes were identified: Luminal A (LumA), Luminal B (LumB), Basal, Human epidermal growth factor receptor 2-related (HER2-related), and Normal, each with distinct prognosis and response differently to certain type of therapies. This classification system is now introduced to clinical use, as there are well developed sequencing techniques (167; 199; 229).

These classification methods are not mutually exclusive (232). The biomarkers classification are frequently used as surrogates indicator for the intrinsic subtypes based on gene expression (151). The Luminal A and Luminal B subtypes demonstrate tumor heterogeneity within ER-positive breast cancers and usually have more positive survival outcome than HER2-enriched

and basal-like subtypes. Luminal A and Luminal A subtype both express ER, but Luminal B are characterized by increased expression of proliferation-associated genes and have worse prognosis than Luminal A (201). The HER2-enriched subtype has increased expression of HER2 and proliferation genes. The Basal subtype is characterized by increased genes expressed in Basal epithelial cells, and has triple-negative in 70% of its cases (200). Additional proposed subtypes include claudin-low tumors with stem-like signature (173) and AR-positive molecular apocrine tumors(75).

A variety of gene expression classification systems have been developed for the past several years: Oncotype Dx (52; 106), MammaPrint(5), PAM50 (Prosigna) (162; 152; 150), EndoPredict (64), and Breast Cancer Index (BCI) (191). The differences come from different gene expression platforms. Among all commercially available gene expression panels (GEPs), the PAM50 is so far the most widely used GEP in clinical practices, as it has clinical validation (46) and received Europe's approval in 2012, and it was approved by U.S. Food and Drug Administration one year later in 2013.The PAM50 derived its classification algorithm using an independent training set and was aimed at classification for independent single-sample (162). The validation of PAM50 is still ongoing, and a gold standard for subgroup classification by gene expression is yet to come (150).

Among all the aforementioned methods, breast cancer subtypes share some overlapping characteristics. As displayed in Figure 4, histopathological types, for example, medullary cancers, are correlated with ER/HER2 receptor status and also with Luminal A and Basal subtypes.

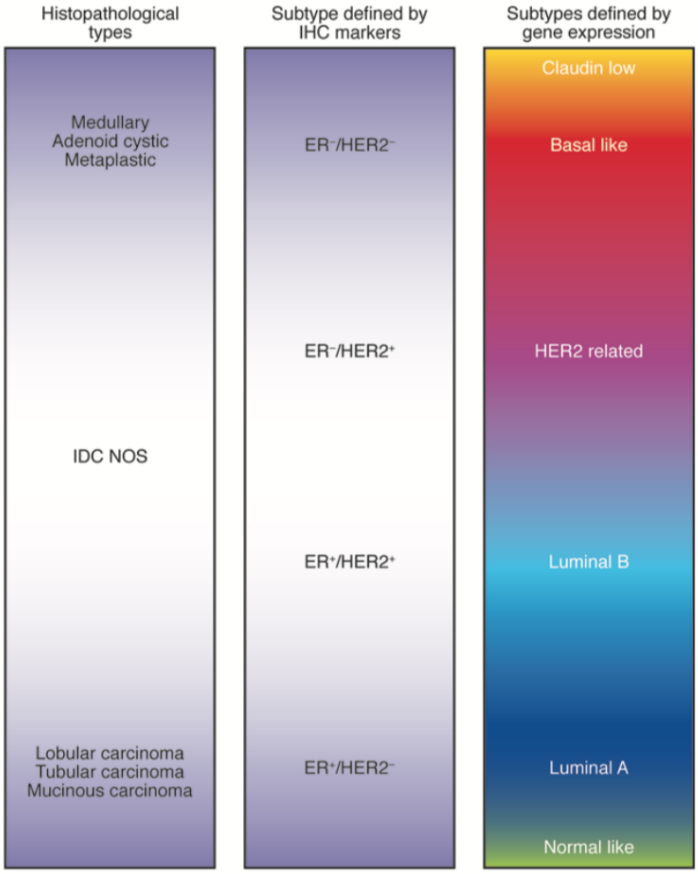

Figure 4. Subtype of breast cancer

2.4 Methods to classify breast cancer intratumor heterogeneity

The aforementioned breast cancer classification methods, by IHC or by gene expression profiling, focuses on intertumor heterogeneity. To characterize intratumor genetic heterogeneity, microarray technologies and massive gene sequencing have made it possible to generate genome-wide gene expression profile and chromosome copy number, offering quantitative information that can capture global intratumor heterogeneity.

2.4.1 Tumor bulk sequencing

Tumor bulk sequencing approach provides an integrated measurement of the underlying clonal complexity and copy number alterations through statistical inference, using certain algorithms such as ABSOLUTE (41) or PyClone (194). However, both methods require prior information of the specific features of the tumors to obtain optimal results. For example, ABSOLUTE method can integrate recurrent cancer karyotype models to identify the most common karyotype that would fit the data. However, more meaningful results will be attained by setting up appropriate prior knowledge on either tumor purity or on the number of complete sets of chromosomes in a cell. Furthermore, to draw inferences on the architecture of cancers based on tumor bulk sequencing, it requires comprehensive sequencing of whole genome, which is not practical at present due to the cost and time required for comprehensive sequencing.

2.4.2 Single-molecule sequencing

Single-molecule sequencing technology is another way to perform bulk sequencing. Compared to massively parallel sequencing (MPS), single-molecule sequencing no longer requires PCR amplification, thus eliminating biases coming with PCR amplification.

Single-molecule sequencing requires less starting information, has faster turnaround time, and can better identify genomic rearrangements (180). However, to capture intratumor heterogeneity, single-molecule sequencing has the similar challenges as bulk sequencing. As it is not a direct measure of intratumor heterogeneity, the subclonal architecture can only be statistically validated. Finally, the current method of single-molecule sequencing is not large enough for heterogeneity studies. Advanced technologies are needed for the application of single-molecule sequencing in intratumor genetic heterogeneity.

2.4.3 Single-cell sequencing

Single-cell sequencing is a relatively objective method to measure intratumor heterogeneity. Compared to bulk sequencing such as MPS or single-molecule sequencing, single-cell sequencing provide direct information of clonal genotypes by identifying lists of genetic alterations in each tumor cell that composes a tumor (9). One example comes from Navin in 2011 (143), they have successfully developed and applied MPS to single nucleus sequencing, and indicated that many breast cancers are constituted of multiple subclones with different genetic characteristics. However, the cost and time required for whole exome or whole genome single cell sequencing of tumors are prohibitive for clinical use. Furthermore, this method collects data from single cell; it cannot provide any comprehensive information on whole tumor cell population. Furthermore, although single-cell sequencing can correctly identify architecture variations, the genome-wide assessment of mutations in single cells is still challenging due to whole genome amplification (222).

2.5 Clinical implications of intratumor heterogeneity

Despite better understanding of phenotypic and genetic aspects of tumor heterogeneity, no significant clinical improvement has been made with respect to effective diagnostic, prognostic, or therapeutic strategies for breast cancer (18).

Heterogeneity adds the energy for resistance; thus, an accurate and meaningful measurement of tumor heterogeneity is critical for the development of effective breast cancer therapies. Breast cancer heterogeneity impinges on patients' prognosis and response to target therapy. Because of this, heterogeneity is one of the most fundamental and clinically relevant topics in cancer research.

Intratumor heterogeneity adds another layer of complexity to the problem. Intratumor heterogeneity is potentially to play an important role in responsiveness to chemotherapy (155; 87; 135). It may explain why some patients who initially respond well to certain cancer treatment but eventually relapse, as the new tumors will no longer response to the therapy. The greater diversity of tumor cells, the more likely that an occasional cell may develop to adapt to the stress imposed by drug.

Some subtypes seem to have greater intratumor heterogeneity than others (160). An optimal diagnostic test will require the identification of even minor subpopulations of cells with alterations related to increased aggressiveness or therapy resistance.

Currently, patients are treated based on the ER/PR/HER2 status of the primary tumor, and metastatic sites may not always biopsied for histologic confirmation due to the cost (7). However, mutations in the initial tumor may not be responsible for tumor progression, it is

critical to identify the dominant local clones driving metastatic disease (9). In ideal situations, sequencing technologies should be used to assess intratumor heterogeneity for each patient at diagnose stage, then monitor clonal dynamics during the whole period of disease progression and treatment. This will allow the identification of genetic changes driving resistance and can provide insights about therapy adjustments as disease progresses (200; 109; 19). Single cells sequencing has presented promising insight into breast cancer intratumor heterogeneity. However, the technique is still impractical because it requires large numbers of genome-wide amplification. In addition, mapping and exclusion is of challenging nature; the development of methods for data analysis and interpretation is still in an early phase. In contrast to these aforementioned efforts at characterizing global genomic intratumor heterogeneity, in this thesis, I will focuses on subtype heterogeneity and the potential coexistence of multiple subtypes within a patient.

Our purpose in this thesis is to develop a quantitative measure to evaluate the purity of certain subgroup in breast cancers Luminal A patients based on PAM50 algorithm. We initially focus on Luminal A cases because hypothetically, Luminal A admixture with any other intrinsic subtypes (Luminal B, HER2-enriched, Basal) would be connected to more aggressive disease and worse survival outcome.

2.6 Dataset

We developed and validated our proposed intratumor heterogeneity quantification methods, using two independent and publicly available cohorts: the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort containing 17814 gene expression pro-

files from 1980 patients, and The Cancer Genome Atlas (TCGA) BRCA provisional cohort containing 20532 gene expression profiles from 1081 patients.

2.6.1 METABRIC clinical and histopathological data

In 2012, the Breast Cancer Prognostic Challenge (BCC) organizers released the METABRIC dataset (53). The study is the largest global study of breast cancer; it integrated genomic analysis of breast cancers as well as long-term clinical information.

The tumors in the initial METABRIC cohort were collected from 1977 to 2005 based on five centers located in UK and Canada. METABRIC assembled a collection of over 2000 clinically annotated fresh frozen breast cancer specimens and a subset of normals specimens. The nucleic acids were isolated from frozen tissue and were reviewed to assess the presence of invasive tumour, pre-malignant or benign changes, tumour cellularity, and lymphocytic infiltration in specific subgroups. Cases were included from the METABRIC cohort if they met the criteria: the histology slides were available from central pathology review together with corresponding clinical and gene expression analyses data. After applying this criteria, there are 1980 subjects in METABRIC cohort.

All primary data were deposited at the Genome-phenome Archive(EGAS00000000083) and could be downloaded after the request was approved by METABRIC Data Access Committee. Gene expression data, copy number data from the original METABRIC publication can also be found on the freely available cBioPortal.

Tumors in METABRIC were primary invasive breast cancers for which clinical information could be linked to DNA and RNA specimens. Tumor clinical related information was included

in METABRIC, such as overall survival, grade, tumor size, age at diagnosis, number of lymph nodes positive, ER status, PR status, HER2 status and PAM50 subtypes.

2.6.2 METABRIC gene expression data

Messenger RNA (mRNA) were extracted from each specimen and subject to copy number and genotype analysis on the Affymetrix SNP 6.0 platform and transcriptional profiling on the Illumina HT-12 v3 platform(Illumina_Human_WG-v3).

DNA and RNA samples from UK were extracted from fresh frozen tumors using the DNeasy Blood and Tissue Kit and the miRNeasy Kit (Qiagen, Crawley, UK) on the QIAcube (Qiagen) following the manufacturers instructions. Samples from Canadian site were extracted from 10 um sections from fresh frozen tumors using the MagAttract DNA Mini M48 Kit and miRNeasy 96 Kit (Qiagen) following the manufacturers instructions. Nucleic acids were quantified assessed with NanoDrop ND-8000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA).

RNA quality was examined using the Agilent 2100 Bioanalyser Nanochip (Agilent Technologies, Wokingham, UK). Tumour samples for which the RNA had an RNA Integrity Number (RIN) > 7 were hybridized to expression arrays, a less stringent RIN > 5 was required for normal RNA. Random DNA and RNA samples were kept for genotyping purposes to ensure sample uniqueness using the AmpFlSTRIdentifiler PCR Amplification Kit (Applied Biosystems, Foster City, CA, USA).

Total RNA was used to generate cRNA using the Illumina Totalprep RNA amplification kit (Ambion, Warrington, UK) and hybridized onto Illumina Human HT-12 v3 Expression

Beadchips per the manufacturers instructions and was scanned on the Illumina BeadArray Reader.

2.6.3 METABRIC data pre-processing and normalisation

A custom R package was used to process each BeadChip once scanning was complete and raw data were available (67). The process included the generation of quality assessment (QA) information and adjustment for spatial artifacts with the BASH (40). Once all arrays on the chip had been processed, the bead-level data were summarized, yielding a series of 48803×12 matrices of log2 intensities, standard errors, and number of observations. After all chips had been processed, the summarized matrices were combined.

Potential outlier arrays were removed by comparing with the bead-level QA scores derived using the control probes on each array. Arrays with P95 scan metric less than 200 were excluded as they were considered to have failed hybridisation. These arrays then were removed from the dataset so as not to influence the following QA. A testing procedure that detects multivariate outlier in the arrayMvout package was used to identify poor quality arrays based on the QA scores (10). Additionally, this approach was applied to bead-level QA information for each location separately. All arrays that remained after this three-step procedure were retained in the subsequent analysis.

In order to avoid the influence of certain specific probes on the normalization, the comprehensive re-annotation of the Illumina HT-12 v3 platform is applied here (13). In the final data, a single target distribution was generated using the list of suitable probes. The ER-positive and ER-negative samples were normalized separately and the final target distribution comes

from average. Each array was then normalized to the target by quantile normalizing probes belonging to the target distribution, while the values for the remaining probes were obtained using the weighted normalized probability of the target distribution probes with most similar intensities prior to normalization. A linear model was then fit using the limma Bioconductor package to remove any batch effects associated with the position of an array on the Illumina BeadChip (181). In METABRIC mRNA data,I impute missing, zero or negative values using real numbers randomly sampled from a uniform distribution in the range [.05,.95].

2.6.4 METABRIC PAM50 classification

Samples were classified into the five intrinsic subtypes based on PAM50 (162; 200). As probe annotation is an important consideration in sample classification using microarray data (66), the PAM50 gene-list was refined such that genes with perfect annotation on the Illumina HT-12 v3 BeadChip (40) were kept for classification. As a result, 3 genes were not included in the classification. For genes with more than one probe, probes were selected on the basis of their annotation. As previously recommended (199), all probes were median centered prior to classification. Due to the imbalance in ER status, 100 random reference distributions consisting of all ER-negative samples were defined, ER-positive samples were randomly selected during the median centering step.This resulted in 100 different classifications and the final subtype calls were derived by taking a consensus across all 100 trials. Samples were then assigned to one of the five intrinsic subtypes using the Spearman correlation to the published centroids and the transformed intensities, where samples with correlations < 0.1 for all subtypes were not classified (NC) (0 samples in the discovery set, 6 samples in the validation set).

Zhao indiated that PAM50 classification accuracy may be affected by extreme differences in the prevalence of ER-positive cases between a study cohort and the benchmark training cohort (195). However, their simulation results indicate acceptably low error rates throughout an ER-positive prevalence range of about 60-80%, and in METABRIC, the ER-positive prevalence was 75.6% . Our re-generated PAM50 classifications labels were identical to those recorded in the downloaded datasets (20). Figure 6 is a heat map displaying the hierarchical clustering, based on expression of individual PAM50 genes for METABRIC data.

2.6.5 TCGA data

Another large genomics data consortia is the Cancer Genome Atlas (TCGA). The TCGA BRCA provisional cohort we used in this thesis contains 20532 unique genes from 1081 patients. The datasets include annotated somatic mutation, raw simple somatic mutation, gene expression quantification, copy number segment, masked copy number segment, isoform expression quantification, and microRNA (miRNA) expression quantification. The mRND data in TCGA was obtained using the Illumina HiSeq platform. The clinical information in TCGA includes the case identifier, disease type, gender, age at diagnosis, overall survival, grade, number of lymph nodes positive, ER status, PR status, HER2 status and PAM50 subtypes, except tumor size.

All clinical and genomic data for the TCGA cohorts can be downloaded from cBioportal. The mRNDA data from TCGA was log2(X + 1) transformed to standardize the data before analysis.

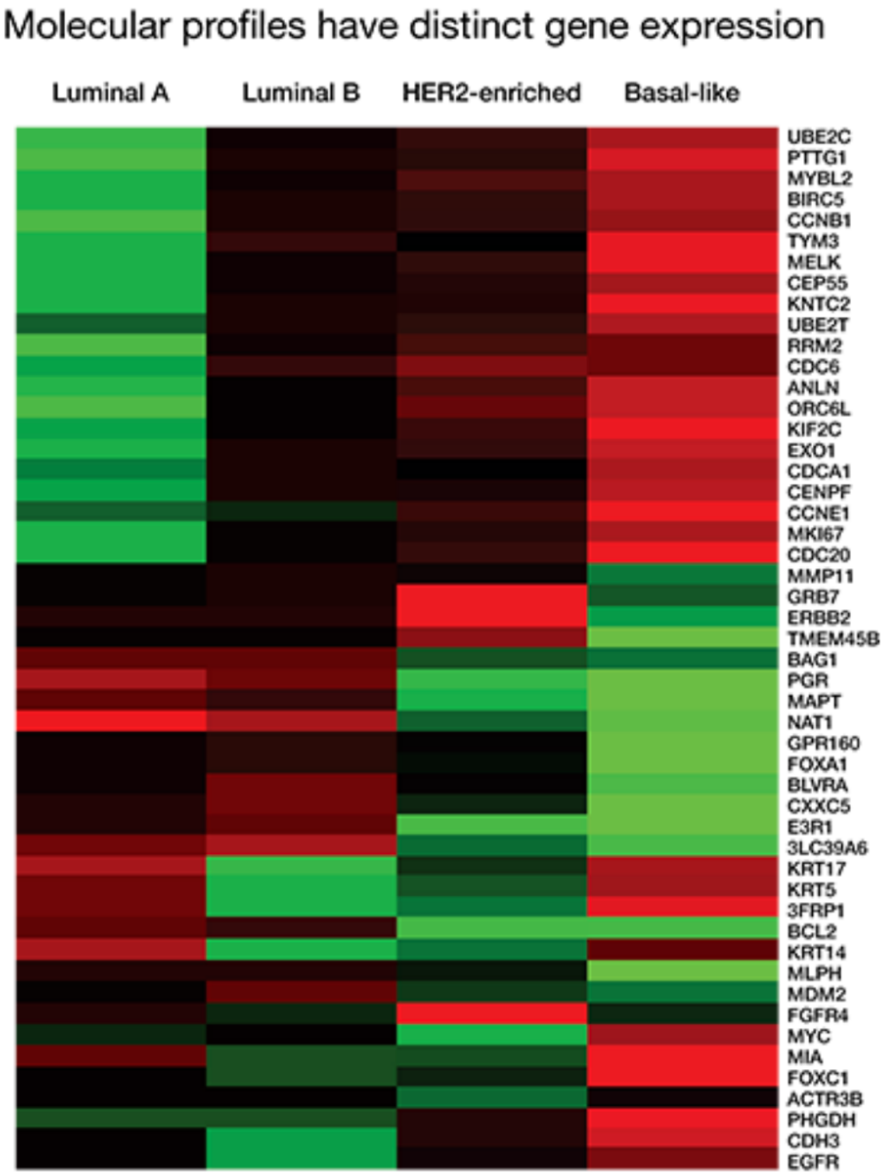

Figure 5. PAM50 gene heatmap for METABRIC data

CHAPTER 3

METHODOLOGY

Cluster Analysis attempts to solve the problem of finding meaningful subgroups of interest within heterogeneous data. In general, cluster analysis seeks to identify homogeneous subgroups such that each subgroup corresponds to a distinct set of characteristics and can be well separated from others (96; 216). Homogeneity means that subjects within the same cluster should resemble one another, Clustering is the procedure of grouping a set of data points such that subjects belonging to the same cluster share some similarities and subjects belonging to different clusters are different in the same sense (104).

3.1 Non-parametric cluster method

In this section, we focus on using non-parametric unsupervised cluster method to identify the homogeneous clusters within the heterogeneous data sets. These non-parametric approaches do not require any prior information about the datasets; these approaches consider the data as forming a static distribution and determines the most remote points, which are used as criteria to assign a point to cluster. Various methods for unsupervised clustering have been studied (184; 15; 114; 125; 14).

3.1.1 Mahalanobis distance method

One straightforward method to identify a cluster is to calculate a 'distance' of each point from a "center" of the data points and define the clusters accordingly. Next, we will describe a distance based cluster method.

To answer this kind of question, a measure of distance between clusters in terms of multiple characteristics is used. The most often used measure is the Mahalanobis distance; Mahalanobis proposed this distance measure in 1930 (122) in his studies on racial likeness. After that, Mahalanobis distance has been used as an important tool in data analysis with multiple dimensions; it has found applications in many fields, such as classification, statistical pattern recognition, medical diagnosis or remote sensing.

In practice, we are primarily interested in measuring how "distant" an observation is from the center of a distribution. For example, given a data matrix A with n rows of observations and each observation has m measured features.

$$\mathbf{A} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}_{n \times m} \tag{3.1}$$

The Mahalanobis distance D^2 for each observation vector $\mathbf{x_i} = [x_{i1}, x_{i2}, \dots, x_{im}]$ is calculated as a function of a m-dimensional vector $\bar{\mathbf{x}}$ containing the means for each column, and

a (nonsingular) variance covariance matrix Σ of dimensions $m \times m$ that the diagonal element contains variances and pair-wise column covariance values elsewhere.

$$D^2(\mathbf{x_i}) = (\mathbf{x_i} - \bar{\mathbf{x}})^T \Sigma^{-1} (\mathbf{x_i} - \bar{\mathbf{x}}) \tag{3.2}$$

When applied to n=47 points in m=2 dimensions, the calculated D^2 values follow a characteristic elliptical pattern with D^2 radiating out from the central location of the distribution. In the following Figure 6, given a set of points distributed in two-dimensional space, the Mahalanobis distance D^2 for each point has been calculated. The D^2 values asymptotically follow chi-squared distribution (**?**; 39). In this example, there are two points that are highly possible of not belonging to the distribution and can be classified as outliers when $\alpha = 0.975$ as they are above that probability threshold.

Under the assumption of multivariate normality of the underlying quantitative measurements, Mahalanobis distance measure D^2 may be interpreted as the sum of m independent standard normal variables, thus following a chi-squared distribution with degrees of freedom equal to the number of dimensions i.e. m (179; 44). The centering of $\mathbf{x}$ and scaling of Σ are needed to arrive at the above distributional property of D^2.

Although a large number of distance measures of similarity between groups have been proposed, the Mahalanobis distance D^2 has been found to be the most suitable in a majority of applications. It is now known that many standard distance measures, such as Kolmogorov's

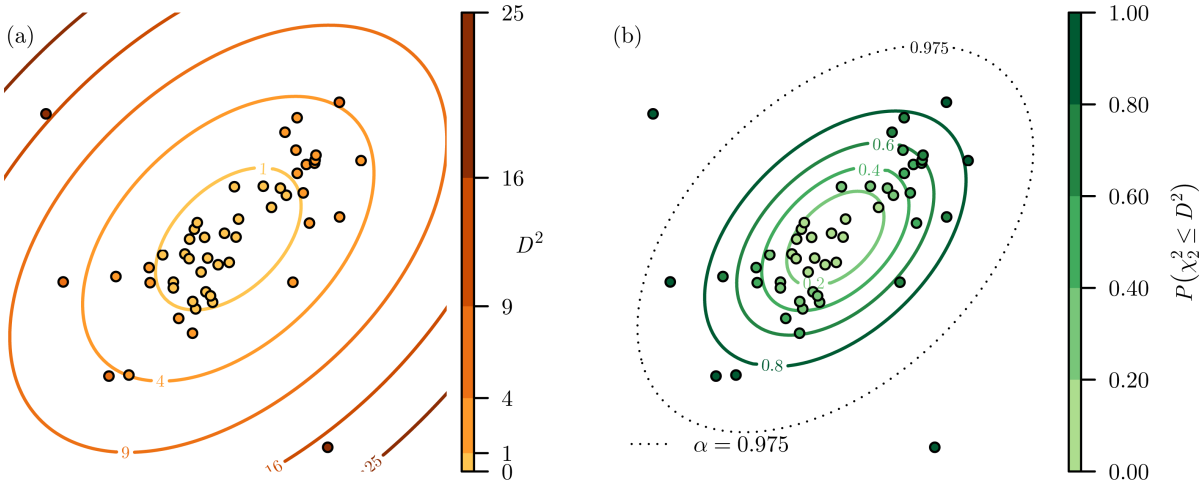

Figure 6. Illustrative example of Mahalanobis distance in two dimensions

variational distance, the Hellinger distance, Rao's distance, are extended functions of Mahalanobis distance under assumptions of normality (230; 63; 17).

To capture the intratumor heterogeneity based on gene expression, we use Mahalanobis distance to measure how far away an individual subject lies with respect to the centroid of a specific subgroup. For a sample of 47 genes, we first calculate the centroid of the subgroup, which is the mean of 47 genes; then we calculate the covariance matrix of 47 genes within this specific subgroup. For each patient, we calculate five distances, the distance of this subject to the centroid of its own group (LumA), and the distances of this subject to the centroids of other four subgroups (LumB, HER2, Basal, Normal). Then, we use 10 fold cross validation to find the optimum threshold α to assign subject to cluster. In this thesis. we assume there may be 2 or 3 clusters within the LumA group. When there are 2 clusters, one cluster contains the

"pure" LumA patients, another cluster contains the "admixed" LumA patients. When there are 3 clusters, similarly, one cluster contains the "pure" patients, one cluster contains the "neither" patients and the last cluster contains the "admixed" patients. The definition of "pure" patient is as follows:

1. Mahalanobis distance to centroid of its own group $\leq \alpha$ % of Mahalanobis distance of its own group **And**
2. Mahalanobis distance to centroid of other group A $\geq \alpha$ % of Mahalanobis distance of subgroup A **And**
3. Mahalanobis distance to centroid of other group B $\geq \alpha$ % of Mahalanobis distance of subgroup B **And**
4. Mahalanobis distance to centroid of other group C $\geq \alpha$ % of Mahalanobis distance of subgroup C **And**
5. Mahalanobis distance to centroid of other group D $\geq \alpha$ % of Mahalanobis distance of subgroup D

3.1.2 Distance ratio method

Next, we describe another non-parametric method for cluster analysis. For each subject, we have five distances, the distances of this subject to the centroid of its own group, and the distances of this subject to the centroids of other four subgroups, among the four non-assigned distance, we pick up the smallest distance, and then calculate a score, which is the ratio of

the Mahalanobis distance from its own group to the Mahalanobis distance to the nearest non-assigned group. Then a subject is assigned to a cluster by tertiling this ratio into three clusters.

3.1.3 Overall survival based on Mahalanobis distance method

Figure 7 and Figure 8 show the Kaplan-Meier plots of overall survival (OS) for METABRIC Luminal A cases stratified according to Mahalanobis distance assuming two clusters and three clusters separately. There is a statistically significant difference between different clusters at both scenarios. When there are three clusters (Figure 8), the "pure" subjects have much better long term chance of survival compared with "admixed" subjects.

3.2 Finite mixture model cluster

Non-parametric cluster algorithm is limited by lack of statistical inference to determine the goodness of fit with respect to the number of clusters (65; 213). In addition, the results of non-parametric cluster algorithm can be influenced by the order of data input (202; 219). There is a number of concerns about the validity of non-parametric cluster algorithm as it is capable of producing clusters even when there are no true clusters (224). A number of methods have been proposed to resolve this weakness, such as re-sampling technique (213). However, these methods are ad-hoc, time consuming and would be impractical to implement in high dimensional datasets.

Unlike non-parametric cluster, model-based cluster analysis provides classical statistical inference naturally, it is a sound statistical procedure that allows the comparison of non-nested models. This approach can penalize model complexity and reward for parsimony when comparing different models (116). Model-based cluster analysis would eliminate the ambiguity

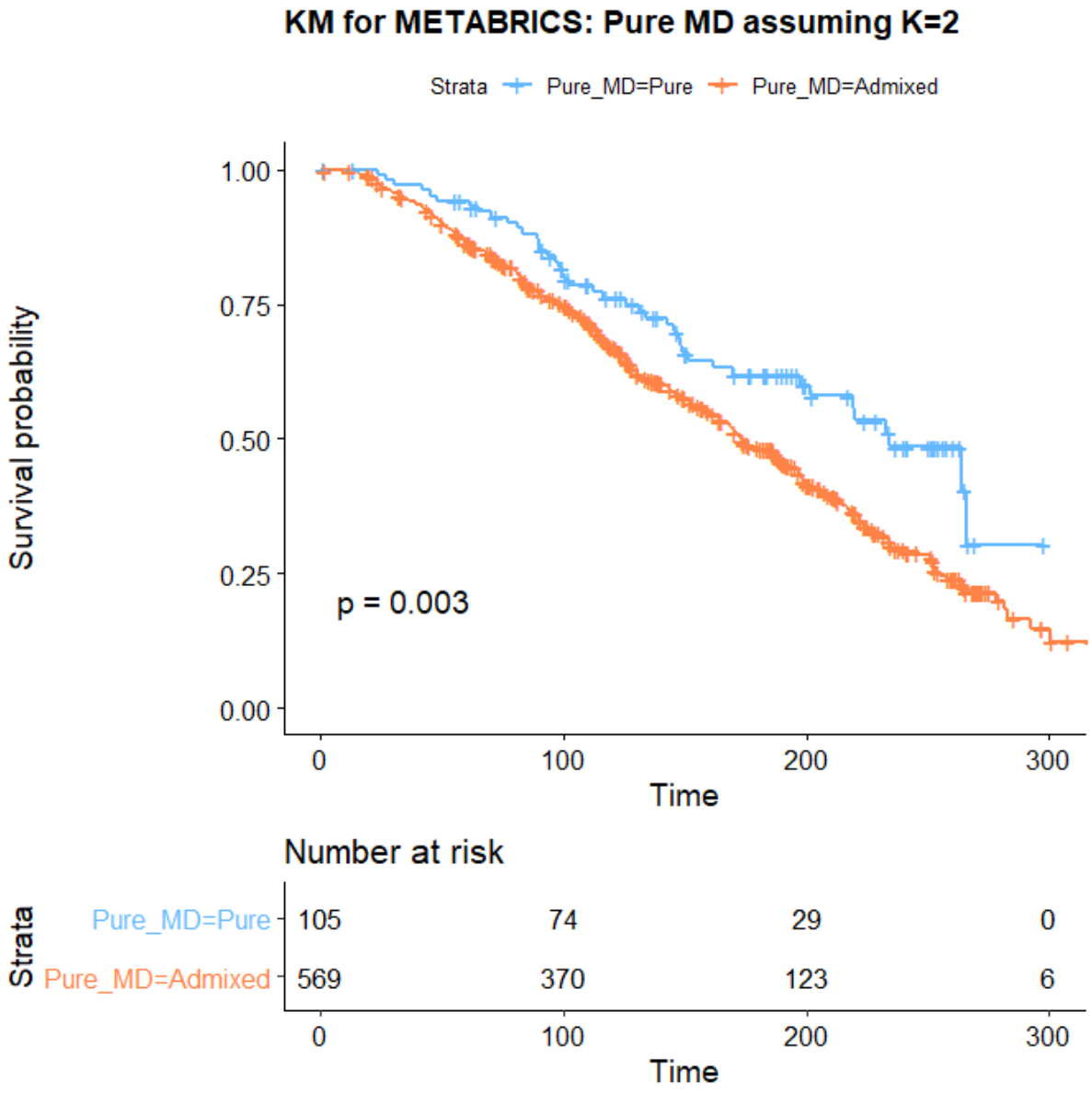

Figure 7. KM curve based on Mahalanobis distance assuming 2 components

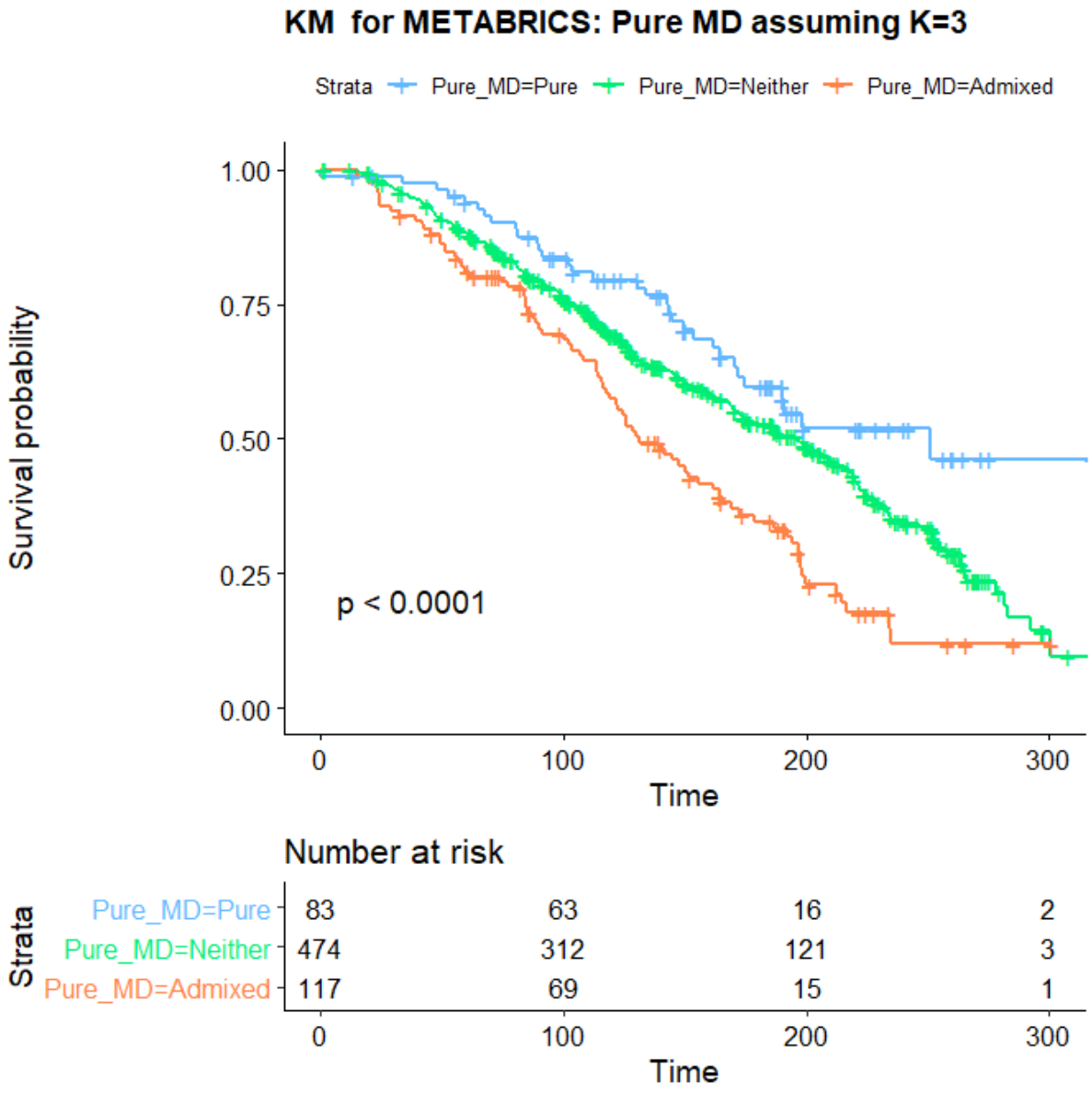

Figure 8. KM curve based on Mahalanobis distance assuming 3 components

associated with subjective criteria (82; 83; 238); thus it is an important tool for cluster analysis (12; 58; 176).

3.2.1 Motivational example

We frequently encounter data from heterogeneous sources. For example, while measuring weights of all people in a town, we can use a mixture of two Gaussian distributions for the data, since by common sense, men and women have quite different weights genetically. Therefore, it can be considered as two different sources of data. Another example comes from mixture of Poisson distribution. Insurance companies receive tremendous number and amounts of claims each year. Studies indicated that a mixture of Poisson distribution is appropriate for the number of claims made by each individual (197). Majority of people have very few claims each year; which can be represented by a Poisson distribution with a small rate, while minority of people make a lot of claims each year, for them a Poisson distribution with a high rate can be used. In the insurance example, we have a sample consisting of heterogeneous subjects, and a mixture of Poisson model is appropriate here. Mixture model has another application in image processing. An image can be seen as a composition of different textures, brightness and colors. Mixture model has shown to be a powerful tool for image reconstruction, classification and segmentation (118; 189).

Mixture model is commonly used to model data generated by mixed sources in many areas. Hartigan (98) gives a nice summary of mixture model theory as well as many possible applications of mixture in real life; see Table I.

TABLE I

DIFFERENT TYPES OF MIXTURE DATA

Data type	Distribution family	Example
Discrete	Possion	Insurance claim
Discrete	Binomial	Patients responding to a treatment
Continuous	Normal	Genome data; Image data
Heavy tail	Cauchy	Stock

3.2.2 Finite mixture model based cluster

Under the assumption that observed data comes from a population consisting of several sub-populations, finite mixture model fits each sub-populations separately, and the overall population is a weighted sum of these sub-populations (176; 82). A mixture model can be written as a convex combination of multiple distribution functions; The pioneering work on finite mixture model can be found in (76; 209). Recent comprehensive reviews of mixture model can be found in (119; 130; 128; 186; 134). Mixture models are widely used in statistical modeling because of the flexibility and the potential interpretation of the mixing process (74; 85).

3.2.2.1 Set-up of finite mixture model

Define χ as a measurable space, more specifically, χ is a subset of R^r equipped with Borel set all through this thesis. Φ_θ is the probability measures on χ indexed by parameter θ, $\theta \in R^m$. The parametric family is denoted by:

$$G = \Phi_\theta, \theta \in R^m, \tag{3.3}$$

In addition, we assume that each Φ_θ has a density function $\phi_\theta(x)$,$x \in \chi$ with respect to a common dominating measure λ. $\phi(x, \theta)$ is denoted as a density function or a kernel function. A finite mixture model with K components is defined as:

$$f(x; \phi) = \sum_{k=1}^{K} \tau_k \phi(x, \theta_k), \tau_k > 0, \sum_{k=1}^{K} \tau_k = 1, \tag{3.4}$$

In finite mixture model, we are particularly interested in three sets of parameters. They are:

1 K: the number of components;

2 τ_k,k = 1,2,...,K: weight of each component;

3 θ_k,k = 1,2,...,K: parameter vectors of each component.

In most applications of finite mixture model, the value of K is unknown, but considered to be a fixed value, and has to be determined from observed data, along with the mixing proportions and the component parameters.

An important issue arising from a finite mixture model is identifiability, which can be seen as the foundation for parameter estimation in finite mixture model. The estimation of ϕ will become pointless if the parameters in ϕ are not distinguishable. By definition, a parametric distribution is indicated to be identifiable if different parametric values lead to different members

of the family. The identifiability for finite mixture model is similarly defined (208; 211). The finite mixture model is said to be identifiable if for any two members $f(x;\phi)$ and $f(x;\phi^*)$,

$$\sum_{k=1}^{K} \tau_k \phi(x, \theta_k) = \sum_{k=1}^{K^*} \tau_k^* \phi(x, \theta_k^*), \tag{3.5}$$

if and only if $K = K^*$, $(\tau_1, \tau_2, ..., \tau_K)$=$(\tau_1^*, \tau_2^*, ..., \tau_K^*)$ and $(\theta_1, \theta_2, ..., \theta_K)$=$(\theta_1^*, \theta_2^*, ..., \theta_K^*)$(236; 208).

3.2.2.2 The Gaussian mixture model

Cluster analysis using finite Gaussian mixtures model has a long history dating back to the detection of outliers by Newcomb (148) and modeling unobserved classes in biological specimens from Pearson (165). Finite Gaussian mixture model now has been widely used in many areas, such as marketing, medicine, physics, and economics. More recently, with the advancement in computer processing power, Gaussian mixture model has been commonly used in genetics (60) and imaging processing (110).

The Gaussian mixture model is a powerful tool for mixed sources of continuous data by modeling a mixture normal probability density function, thus allowing use of formal statistical procedures for parameter estimation and model optimization. When datasets are high dimensional, i.e., each observation has various features, mixture Gaussian can been naturally extended to deal with multivariate continuous observations, $\mathbf{x_1}, ..., \mathbf{x_n}$, where $\mathbf{x_i}$ is an r-dimensional vector, and r represents the number of features. Typically, the elements $x_{i1}, ..., x_{ir}$ of $\mathbf{x_i}$ measure r features for a subject labeled i sampled from a population. A common assumption for finite mixture model is that the observations, i.e., $\mathbf{x} = (\mathbf{x_1}, ..., \mathbf{x_n})$, are i.i.d. realizations from a mul-

tivariate random variable drawn from a mixture with K components. The probability density of $\mathbf{x}$ is:

$$f(x;\theta) = \tau_1\phi(\mathbf{x};\mathbf{u_1},\mathbf{\Sigma_1}) + ... + \tau_K\phi(\mathbf{x};\mathbf{u_K},\mathbf{\Sigma_K}), \tag{3.6}$$

The model is parameterized in terms of distinct model parameters θ and τ. θ=($\mathbf{u},\mathbf{\Sigma}$), is the parameter of a multivariate normal distribution with mean $\mathbf{u}$ and variance-covariance matrix $\mathbf{\Sigma}$. $\boldsymbol{\tau}$ is the proportion of each component. Yakowitz proved that finite mixtures of multivariate normal distributions is generically identifiable (236).

3.2.2.3 Variance Covariance matrix decomposition

A multivariate mixture Gaussian model with unconstrained variance covariance matrices $\mathbf{\Sigma}$ is highly parameterized, the total number of parameters to be estimated is $K(r+r(r+1)/2+1)-1$ (12). For example, when modeling a mixture of 5 multivariate Gaussian distributions with 10-dimensional observations, one has to estimate as many as 329 parameters. When it comes to high dimensional datasets, the number of parameters increases exponentially in the general variance-covariance setting.

With general variance covariance matrices model, one disadvantage is that, there will be a large number of parameters to be estimated, each additional parameter means longer computational time. Another disadvantage is that lack of parsimony will cause the difficulty of interpretation. This causes more severe concern, since a statistical model always requires easy and meaningful interpretation.

An unconstrained multivariate mixture may be too general; to achieve parsimony, we can put certain constraints on the variance–covariance matrices. Various authors inspected different

constrained models, one of the most widely used model comes from Banfield and Raftery (12). They suggested using the eigenvalue decomposition of the covariance matrix $\boldsymbol{\Sigma_k}$:

$$\boldsymbol{\Sigma_k} = \lambda_k \mathbf{D_k} \boldsymbol{A_k} \mathbf{D'_k}, \tag{3.7}$$

where $\boldsymbol{\lambda_k}$ is the largest eigenvalue of $\boldsymbol{\Sigma_k}$. It is a scale parameter which can control the volume of cluster in the sample space, the volume is proportional to the standard deviation ellipsoid. D_k is an orthogonal matrix with each column corresponding to the normalized eigenvectors of $\boldsymbol{\Sigma_k}$. D_k determines the orientation of each cluster with respect to the coordinate axes. $D_k = I_r$ representing the case where principal components are aligned with the coordinate axes. A_k is a diagonal matrix with decreasing eigenvalues of $\boldsymbol{\Sigma_k}$, divided by the maximum eigenvalue. The elements of diagonal matrix A_k control the shape of variance covariance structure, when $A_k =$ Diag (1, ..., 1), it corresponds to a spherical structure, whereas a cluster with $A_k =$ Diag (1, A_{k2}, ..., A_{kr}), where $A_{kj} \ll 1$, is concentrated around a line in the sample space.

Therefore when clusters share similar parameters $\boldsymbol{\lambda_k}$, D_k, and A_k, they share similar geometric properties. By imposing constraints on $\boldsymbol{\lambda_k}$, D_k, and A_k, the parsimony for variance-covariance matrix is achieved. Different constraints on the variance covariance matrix provide distinct models that are capable to capture the structure of datasets. Allowing the parameters in (5.5) to vary between clusters provides easily interpretable models that can be used to infer the structure of different clusters. The most important advantage for constrained variance-covariance structure is the smaller number of unknown parameters. For example, in the most

general model ([VVV]), the parameters to be estimated is $\alpha + K\beta$, where $\alpha = Kr + K - 1$, contains the information of number of components and the number of dimension, $\beta = r(r+1)/2$, contains the information of number of dimension.

Celeux (42) and Bensmail (21; 22) defined three major variance covariance structures: spherical, diagonal and general families. Spherical family is the simplest structure, in which each variable has the same variance such that the distribution is spherical. Diagonal family leads to axis-paralleled elliptical components as the variances in each dimension may vary. General family has the least constrained structure, in which variances covariance matrices are not required to be diagonal. The table below summarizes the characteristics of nine commonly used structures.

TABLE II

PARAMETERIZATIONS OF THE COVARIANCE MATRIX $\boldsymbol{\Sigma}_K$

Model	Covariances	Family	Volume	Shape	Orientation
EII	$\boldsymbol{\lambda I}$	Spherical	Equal	Equal	NA
VII	$\boldsymbol{\lambda_k I}$	Spherical	Variable	Equal	NA
EEI	$\boldsymbol{\lambda B}$	Diagonal	Equal	Equal	Axes
EVI	$\boldsymbol{\lambda B_k}$	Diagonal	Equal	Variable	Axes
VVI	$\boldsymbol{\lambda_k B_k}$	Diagonal	Variable	Variable	Axes
EEE	$\boldsymbol{\lambda DAD'}$	General	Equal	Equal	Equal
EEV	$\boldsymbol{\lambda D_k A D_k'}$	General	Equal	Equal	Variable
EVV	$\boldsymbol{\lambda D_k A_k D_k'}$	General	Equal	Variable	Variable
VVV	$\boldsymbol{\lambda_k D_k A_k D_k'}$	General	Variable	Variable	Variable

3.2.2.3.1 Spherical Family Variance Covariance Structure

Spherical family is the most constrained model where variables of all components have the same variance. Spherical family corresponds to the diagonal matrices with same diagonal elements. There are two structures in this family.

1 Model EII ($\lambda\mathbf{I}$). This is the most parsimonious model with all components having same volume and same shape. The covariance matrix is a diagonal matrix, where $\mathbf{I}$ denotes the $r \times r$ identity matrix. The number of parameters needs to be estimated is $\alpha + 1$.

2 Model VII ($\lambda_K\mathbf{I}$). The volume of spherical model is not equal, allowing the volume to vary while constraining the shape to be the same. In this case, the covariance are diagonal matrices with equal diagonal elements, but λ is allowed to change for each component. The number of parameters needs to be estimated is $\alpha + K$.

3.2.2.3.2 Diagonal Family Variance Covariance Structure

The diagonal family allows the variance of each variable to vary within clusters. This structure is obtained by constraining $\mathbf{B} = \lambda\mathbf{DAD'}$, where $\mathbf{B}$ is a diagonal matrix satisfying $|\mathbf{B}| = 1$. λ and $\mathbf{B}$ determine the volume and shape of the clusters respectively. Different models are achieved by changing the volume and shape between clusters. One point to be noted is that this structure is invariant under any scaling of variables but not under linear transformations (42).

3 Model EEI ($\lambda\mathbf{DAD'}$). The common diagonal covariance corresponds to cluster with fixed volume and shape. The number of parameters to be estimated is $\alpha + r$.

4 Model EVI ($\lambda \mathbf{B_k}$). This model allows the shape to be different, and the volume is fixed. In other words, the volume parameter λ is same for all components; the diagonal matrix $\mathbf{B}$ are allowed to be different for each component. The number of parameters to be estimated is α + Kr-K+1.

5 Model VVI ($\lambda_K \mathbf{B_k}$). This model allows the volume and shape to vary, while the orientation is fixed. The number of parameters to be estimated is α + Kr.

3.2.2.3.3 General Family Variance Covariance Structure

The general family has more general structure, allowing off-diagonal elements to be different. The most unconstrained model ([VVV]) allows volume, shape and orientation to vary between clusters. Other models with fewer parameters are achieved by either fixing the volume, shape or the orientation. All structures in this family are rotationally and scale invariant.

6 Model EEE ($\lambda \mathbf{B}$). This model assumes all clusters have fixed volume, shape and orientation. which means all components have same covariance matrices with nonzero off-diagonal elements. The number of parameters need to be estimated is $\alpha + \beta$.

7 Model EEV ($\lambda \mathbf{D_k A D_k'}$). This model assumes the orientation of components can vary, while keeping shape and volume fixed. The number of parameters to be estimated is α + Kβ-(K-1)r.

8 Model EVV ($\lambda \mathbf{D_k A_k D_k'}$). This model assumes both the orientation and shape to be different among equal volume clusters. The number of parameters need to be estimated is α + Kβ-(K-1).

9 Model VVV ($\lambda_k \mathbf{D_k A_k D'_k}$). The most general, unconstrained model, allowing all geometric features to be different. It has the maximum number of unknown parameters compared to other structures. The number of parameters to be estimated is $\alpha + K\beta$.

3.2.3 Parameter estimation

In this section, we describe the procedures of parameter estimation in mixture Gaussian model. In multivariate Gaussian mixture model, we assume $\mathbf{x} = (\mathbf{x_1}, ..., \mathbf{x_n})$ consists of n i.i.d. observations from a random variable $\mathbf{x}$ of dimension r, coming from a mixture of K components. Each component follows a Gaussian distribution:

$$\phi_k(\mathbf{x}; \mathbf{u_k}, \mathbf{\Sigma_k}) = (2\pi)^{-\frac{r}{2}} |\Sigma_k|^{-\frac{1}{2}} exp\{-\frac{1}{2}(\mathbf{x} - u_k)^T \Sigma_k^{-1} (\mathbf{x} - u_k)\}, \tag{3.8}$$

In mixture Gaussian model, there is one latent variable: the indicator function of each observation's unique component membership belonging, $\mathbb{1}_i$, $\mathbb{1}_{ik} = 1$ if $\mathbf{x_i}$ belongs to the kth component mixture. After incorporating the latent variable, the complete loglikelihood will be:

$$\log L(\theta_k, \tau_k, \mathbb{1}_{ik} | \mathbf{x}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}_{ik} \log[\tau_k \phi_k(\mathbf{x_i} | \mathbf{u_k}, \mathbf{\Sigma_k})], \tag{3.9}$$

To estimate parameters of the mixture Gaussian model, we need to maximize the complete loglikelihood; however, because of the unobserved variable $\mathbb{1}_{ik}$, we can not directly take derivative of loglikelihood and numerically get the estimations. For the numerical optimization with unobserved variables, the Expectation-Maximization (EM) algorithm (61; 127) is a general method. In addition to the EM algorithm, some other methods can be used for parameters

estimation in mixture Gaussian model, such as method of moments (165; 102; 23). However, method of moments was shown to be inefficient in finite Gaussian mixture as compared to the EM algorithm (56). The EM algorithm is an iterative procedure, starting from certain initial estimators, then proceeds to successively update the loglikelihood of model until converge criteria is met. In EM algorithm, instead of maximizing the complete loglikelihood, at each iteration, the observed loglikelihood is improving. Under relatively mild conditions, the EM algorithm is known to converge to local maximum (61; 32; 234; 127).

The EM algorithm requires two steps at each iteration; at the "E-step", given the "observed" data at this iteration, the conditional expectation of complete loglikelihood is computed; at the "M-step", parameter estimation is updated to maximize the expected loglikelihood from E-step. Although in real world, the regularity conditions for EM might not always hold, EM has been widely used for mixture Gaussian model parameter estimation with good results (83).

For multivariate Gaussian mixtures, the EM algorithm procedures can be summarized as follows:

1. Start with initial guess of component membership for each observation;
2. Alternate between E-step and M-step at each iteration, at iteration t+1:
   - At E-step, given the current estimates, the posterior probability, $\hat{l_{ik}}$, of the ith observation belongs to the kth component is computed:

$$\hat{l_{ik}} = \frac{\hat{\tau_k^{(t)}}\phi_k(x_i|\hat{u_k^{(t)}}, \hat{\Sigma_k^{(t)}})}{\sum_{k=1}^{K}\hat{\tau_k^{(t)}}\phi_k(x_i|\hat{u_k^{(t)}}, \hat{\Sigma_k^{(t)}})}, \tag{3.10}$$

- At M-step, given the posterior probabilities from E-step, update the parameter estimation of τ_k, μ_k and Σ_k using the following equations:

$$\hat{\tau_k}^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \hat{l_{ik}}, \tag{3.11}$$

$$\hat{\mu_k}^{(t+1)} = \frac{1}{n\hat{\tau_k}^{(t+1)}}\sum_{i=1}^{n} x_i \hat{l_{ik}}, \tag{3.12}$$

$$\hat{\Sigma_k}^{(t+1)} = \frac{1}{n\hat{\tau_k}^{(t+1)}}\sum_{i=1}^{n} (x_i - \hat{\mu_k}^{(t+1)})'(x_i - \hat{\mu_k}^{(t+1)})\hat{l_{ik}}, \tag{3.13}$$

3 Iterate the second step until the algorithm convergences.

After we estimate model parameters, we want to infer, given a data point $\mathbf{x}$, which component it belongs to. The posterior probability of the observation belonging to the kth component is used to assign membership. If K is the number of components, $p_k(.)$ is the probability distribution of kth component; and τ_k is the proportion of members in component k, according to Bayes' theorem, the posterior probability that an observation $\mathbf{x}$ belongs to k^{th} component is:

$$p(\mathbf{x} \in k) = \frac{\tau_k f_k(\mathbf{x})}{\sum_{k=1}^{K} \tau_k f_k(\mathbf{x})}, \tag{3.14}$$

The posterior probability is calculated for each component; the final membership assignment is then done by assigning the observation to the component with highest posterior probability.

3.2.4 Initialization strategies for EM in multivariate Gaussian mixture

To implement the EM algorithm in mixture Gaussian model, two critical issues have to be addressed: choose the initial estimation for EM, and select the optimum model. In mixture Gaussian, optimum model includes determining the number of components K, and the selection of covariance matrices structure (spherical, diagonal, general) as discussed in aforementioned Section 5.3.2.

Determining the number of components is one of the essential problems of mixture Gaussian; however, there is still no unquestionable conclusion. In model-based clustering, the general procedure is to fit the model to a range of components, k = 1, 2, ..., K. By doing so, determination of the number of components can be simplified to finding the maximum number of components, K. We use model selection technique to determine the optimal number of components and select covariance structure simultaneously. Since every combination of covariance structure and number of components correspond to a probability model, we can apply model section and find the optimum fitting model. More details will be provided in the next section.

Determining initial strategy to pass on to the EM algorithm is as challenging as choosing the number of components. EM algorithm has a nice property that observed likelihood is improved at each iteration until a stationary point is achieved. However, there are some criticisms of EM: one major concern is that the stationary point in EM does not necessarily guarantee to be the global maximum, and the ability of finding global optimum depends on the initialization point of EM algorithm (127; 234; 90). Another disadvantage of EM is that it may converge slowly in

some situations, including "bad" initial point, or when the proportion of latent variable is high (101; 133).

The EM is a deterministic algorithm: given a particular initial input, it always reaches the same output (113; 142; 218). Therefore, selection of an appropriate starting value is of great importance for EM, as it can substantially determine the convergence speed and locate the global maximum. Existing initialization methods for EM can be categorized into two distinct branches. The first category is "uninformed" from data. Without using any information from observed data, it performs a grid search and uses the randomly generated values as initialization for EM, hoping to find global optimal by examining the whole landscape (116; 112). The uninformed starting strategy has a good chance of locating the optimum estimation if the number of initialization is sufficient to ensure the examination of full space. However, this strategy is time consuming, thus not always practical for high-dimensional datasets. The other initialization strategy chooses the starting values informed by the data. It directly searches the global optimal using the initial value that already incorporates the information from data. For example, Bradley (35) proposed using k-means as initialization, another common method is agglomerative hierarchical clustering, proposed by Fraley and Raftery (81; 190). This initialization strategy in general is more efficient than uninformed technique; unfortunately, this method has been criticized for using just one set of starting values, as it restricts the search for the global optimum; furthermore, certain strategy tends to favor specific shapes or patterns (12; 81).

To overcome these limitations, we develop a new initialization strategy aiming at using information from data without putting any prior distributional assumption. We would like to choose the initial values that can incorporate the information from data, but not impose any restrictions on the form of distribution; in the mean time, we would like to provide multiple initial values to have better chance of finding the global optimum. The details of our proposed method are presented in Section 3.2.4.1; the proposed method is compared with two most commonly used methods: random starting value and agglomerative hierarchical clustering, the descriptions of these two follow in Section 3.2.4.2 and 3.2.4.3.

3.2.4.1 Mahalanobis distance based method

In this section, we describe our proposed method using the Mahalanobis distance (MD) as initial value for our EM, inspired by what are described in Chapter 3.1.1. We first calculate the centroid and covariance matrix of each subgroup. For each observation, five Mahalanobis distances are computed, the Mahalanobis distance of the observation to its own subgroup, and the Mahalanobis distances of the observation to the other four subgroups. Then, we set up a threshold; for a particular threshold, we assign the "pure" membership based on the following rules:

1. Mahalanobis distance to centroid of its own group $\leq \alpha$ % of Mahalanobis distance of its own group **And**
2. Mahalanobis distance to centroid of group A $\geq \alpha$ % of Mahalanobis distance of subgroup A **And**

3. Mahalanobis distance to centroid of group B $\geq \alpha$ % of Mahalanobis distance of subgroup B **And**
4. Mahalanobis distance to centroid of group C $\geq \alpha$ % of Mahalanobis distance of subgroup C **And**
5. Mahalanobis distance to centroid of group D $\geq \alpha$ % of Mahalanobis distance of subgroup D

The membership assignment based on MD is used as an initialization value for EM. The threshold α is within a range of [0.05,0.95], we set up the threshold by grid searching the whole range, thus get multiple sets of initialization values. By doing so, we first obtain multiple initialization membership settings, thereby avoiding the search for a solution to only one possible result. Second, Mahalanobis distance for membership assignment doesn't require any arbitrary parameter set-up, in addition, it doesn't put any constraints on distribution of data. In the meantime, the MD calculation is based on the observed datasets, allowing us to initialize the EM incorporating the information from data.

3.2.4.2 Random starting method

In the context of initialization for EM in mixture Gaussian model, random starting value means classifying each observation to one of kth component randomly. We assume that each component will have an equal number of observations. For mixture Gaussian models, it is highly possible that one set of initial value may lead to a local optimum. For this reason, it is suggested to initialize a mixture model with different random values many times (204; 99) and select the best solution using the likelihood function.

It has been shown that random starting method has similar or better performance than other initialization techniques (27; 29; 235). However, the main disadvantage of this initialization is that a large number of random starts should be examined before a solution can be claimed. In addition, there is still no conclusion about the sufficient number of initialization sets to ensure a full examination of whole likelihood space (57; 105).

3.2.4.3 Agglomerative hierarchical cluster

Another common way to generate starting values is agglomerative hierarchical clustering. Hierarchical clusters are achieved by merging two clusters that provide the smallest decrease in classification likelihood function recursively (81). Given a chosen number of clusters, hierarchical cluster analysis is performed and it generates the membership assignment for each observation. More details of agglomerative hierarchical clustering can be found in Everitt (74).

Hierarchical clustering may be an accurate way to describe data that are structured hierarchically, but studies have controversy conclusions about the classifications accuracy relying on cluster analysis (69). It has been validated that hierarchical clustering classifications initialization strategy perform similarly as in a k-means cluster analysis (55; 81). Furthermore, the use of single set of initialization strategy can have negative influences in certain cases, as it restricts the search for only one possible solution.

3.2.5 Model selection

Each variance covariance structure described in Section 3.2.2.3 corresponds to the geometric features of data. Besides that, we have to determine the number of components in mixture Gaussian model. However, we don't have any prior knowledge about the true covariance struc-

ture or the number of components; therefore, model selection technique is used to determine the optimal number of components and covariance structure simultaneously, by finding the best fitting model.

Model selection, especially the optimum number of components selection, is one of the most basic problems of mixture model. In general, too many components may cause over-fitting and difficulty in interpretation; while too few components may not be adequate to approximate the true underlying structure. Many non-parametric cluster analysis techniques use subjective judgment and arbitrarily determine the number of components. In model-based cluster analysis, the procedure for model selection is to fit different models with different number of components along with certain covariance structure, and then choose the best fitting model by certain criteria. There is a large body of research which has proposed to solve this problem, such as information criteria based method, likelihood ratio test. We will start with the description of information criteria based method.

3.2.5.1 Information criteria based method

There are enormous number of methodological developments to address the issue of determination of K using the information criteria. Information criteria method is parsimony-based and it is the most widely used; a comprehensive review for information criteria applied in finite mixture model can be found in McLachlan (128).

The theoretical justification for information criteria approach is to select K that can minimize the negative log-likelihood blended with a penalty function. As the likelihood is a measure of the quality of model fitting, it tends to select more complex model with more parameters

systematically. To avoid over fitting, information criteria approach imposes a penalty function to balance the increase in model fitting against the complexity of the model. The general form for information criteria is:

$$IC_{(k,\Sigma)} = -2lnL + ds_{(k,\Sigma)} \tag{3.15}$$

where $s_{(k,\Sigma)}$ is the number of parameters associated to a solution with K components and a specific covariance structure Σ, d being the marginal cost per parameter (33).

Most conventional information criteria methods include Akaike Information Criterion (AIC) (3), Bayes Information Criterion (BIC) (188), the Integrated Classification Likelihood criterion (ICL) (26), and their various modifications, such as Minimum Information Ratio criterion (MIR) (231) and Laplace-Empirical Criterion (LEC) (128). Next, we will review three most widely used techniques: AIC, BIC and ICL.

3.2.5.1.1 Akaike information criteria

The well known information criteria used for model selection of mixture Gaussian is AIC. The AIC is calculated as:

$$AIC(K, \Sigma) = -2logp(x|K, |\Sigma) + 2d, \tag{3.16}$$

where d is the number of free parameters in the model. AIC chooses the model that asymptotically minimizes the mean Kullback-Leibler information (34) for discrimination between the proposed model and the true model.

3.2.5.1.2 Bayesian information criteria

The BIC (188) selects model with the highest posterior probability among competing models. It has been widely used for mixture models and for cluster analysis (182; 55; 128). BIC is defined as

$$\mathrm{BIC}(K, \Sigma) = 2p(x|K, |\Sigma) - d\log(n), \tag{3.17}$$

where d is the number of free parameters in the model.

The BIC uses an approximation of twice the log integrated likelihood (210), although the regularity conditions do not hold for mixture Gaussian models in general (2). It has been shown that BIC has consistent performance of choosing the optimum number of components (182) and leads to consistent estimation (182).

3.2.5.1.3 Integrated completed likelihood criterion

Biernacki (25) noted that the integrated likelihood cannot provide an evidence for a density structure, they suggested the alternative use of the Integrated Completed likelihood. The integral likelihood:

$$p(\mathbf{x}, \mathbb{1}_i|K, \Sigma) = \int p(x, \mathbb{1}|K, \Sigma, \phi)p(\phi)d\phi, \tag{3.18}$$

where $p(\phi)$is a non informative prior distribution on ϕ for this model, $\mathbb{1}$ is the indicator function for membership assignment. Rewriting (3.18), we can get:

$$p(\mathbf{x}|\mathbb{1}, K, \Sigma)p(\mathbb{1}|K, \Sigma) = \int p(x|\mathbb{1}, K, \Sigma, \phi)p(\mathbb{1}|K, \Sigma, \phi)p(\phi)d\phi, \tag{3.19}$$

Biernacki (25) suggested that the integration of $p(x|\mathbb{1}, K, \Sigma, \phi)$ over $p(\phi)$ is in closed form, as long as the prior distribution of $\mathbb{1}$ is independent of ϕ; thus he proposed the use of BIC approximation for $\log(p(x|\mathbb{1}, K, \Sigma, \phi))$ (25; 26):

$$\log p(x|\mathbb{1}, \hat{\phi}^*, K, \Sigma) - \frac{1}{2}\log(n), \tag{3.20}$$

where $\hat{\phi}^* = \text{argmax}_\phi\, p(x|\mathbb{1}, \phi, K, \Sigma)$, which is not necessarily the same as $\hat{\phi}$ from maximum likelihood estimation. Then, ICL can be defined as:

$$\text{ICL} = -2 * \log(p(x|\hat{\mathbb{1}'}, \hat{\phi}^*, K, \Sigma)) + \frac{d-(K-1)}{K} * \log(n) - 2 * \int p(\mathbb{1}|\tau, K, \Sigma)p(\tau|K, \Sigma), \tag{3.21}$$

where d is the total number of free parameters of the model, $\hat{\mathbb{1}'}$ is the MAP estimation of $\mathbb{1}$ given $\hat{\phi}^*$.

Biernacki (25) also pointed out that by dropping the $\mathbf{O(1)}$ using a Stirlings approximation (54), the approximation for $\int p(\mathbb{1}|\tau, K, \Sigma)p(\tau|K, \Sigma)$ can be written as:

$$p(\mathbb{1}|\tau, K, \Sigma)p(\tau|K, \Sigma) \approx n \sum_{k=1}^{K} \hat{\tau}_k \log(\hat{\tau}_k) - \frac{1}{2}(K-1) * \log(n), \tag{3.22}$$

Substituting BIC into (3.21), ICL can be approximated by:

$$\text{ICL} = \text{BIC} - 2 * \log(p(x, \hat{\mathbb{1}'}|\hat{\phi}^*, K, \Sigma)) + d * \log(n), \tag{3.23}$$

The ICL is the standard likelihood penalized by a measure of the quality of partition; thus compromising between the fitting of model and the mixture model classification ability.

In general, the criteria-based methods are easy to implement, but have the disadvantage of obtaining a meaningful comparison from one solution to another. Kass (108) suggested differences in BIC less than 2 as insignificant, while improvements greater than 10 are often a strong evidence. In other words, reductions in BIC score of more than ten should suggest a strong improvement in the model with increasing number of components. However, it is unclear how this score should be calibrated in different situations regarding to variations in sample sizes. This is where testing-based approaches have greater appeal, because testing-based approaches specify evidence in favor of a complex model against a simpler model with respect to the easily understood p-value.

3.2.5.2 Bootstrap LRTS approach

Another way to decide the number of components of a mixture model is to conduct successive hypothesis tests, using the likelihood ratio test statistic (LRTS). Consider the null hypothesis H_0 of k_0 classes against the alternative hypothesis H_1 of k_1 segments:

$$H_0 : K = k_o, H_1 : K = k_1 \tag{3.24}$$

where $k_1 > k_o, k_1 = k_0+1$. Under certain regularity conditions, the likelihood ratio test statistic (LRTS) provides the necessary information to choose between these two models (51; 129):

$$-2\log\lambda = -2\log[\frac{L(\widehat{(}\theta_{k_0}))}{L(\widehat{(}\theta_{k_1}))}] = 2\{\log L(\widehat{(}\theta_{k_1}) - \log L(\widehat{(}\theta_{k_0})\}, \tag{3.25}$$

Unfortunately, in the case of mixture Gaussian model, the regularity conditions do not hold for (3.25) to asymptotically follow a chi-squared distribution with degrees of freedom equal to the difference of parameters under the null and alternative hypothesis (2; 120; 117; 212). The lack of theoretical null distribution of LRTS has stimulated the development of resampling approach to produce p-value. One methodology is bootstrap method proposed by McLachlan (129) to obtain the null distribution of the LRTS; other methodologies include Monte Carlo procedure applied to mixture model (100; 1; 227).

In the resampling procedure, the LRTS from real data is compared with the test statistics that are generated from a set of bootstrapped samples, under the null hypothesis. The bootstrap procedure is:

1 Simulated a bootstrap sample x_b^* from the model under the null hypothesis with k_0 components, i.e. from the mixture Gaussian with the vector of unknown parameters replaced by MLEs obtained from the original data under H_0;

2 Compute the test statistic $LRTS_b^*$ for the bootstrap sample x_b^* from step 1 following mixture Gaussian with k_0 and k_1 number of components;

3 Repeat steps 1 and 2 many times, e.g. 1000 times, to obtain the bootstrap null distribution of LRTS*.

A p-value from bootstrap can then be approximated as:

$$p - value \approx 1 + \frac{1 + \sum_{i=1}^{M} I(LRTS_b^*) \geq LRTS_{obs}}{M + 1} \tag{3.26}$$

where $LRTS_{obs}$ denotes the test statistic from the observed sample x, I is the indicator function, I= 1 if its argument is true, M is the number of repeats.

One obvious advantage of LRTS method is that it tells us when exactly to favor H_0 over H_1, and it performs better for small sample sizes. It was advocated by McLachlan as a necessary tool for assessing p-values (128). However, the main limitation of the LRTS relates to its computational demand because of the resampling procedure (130; 227)

CHAPTER 4

HYPOTHETICAL DATA

In this chapter, we examine the performance of finite mixture model with simulated datasets. The objectives are, first, to investigate how different model selection criteria can effectively capture the characteristics of multivariate Gaussian distribution, and second, to examine the effect of initialization strategies on the performance of EM algorithm in Gaussian mixture modeling, thereby, identifying the heterogeneity existing in the data. Two examples are presented in this section. The first example is to show the ability of different model selection techniques to capture the true variance-covariance structure and the number of components. The second example is to compare the performance of different initialization strategies, i.e. Mahalanobis Distance (MD), Random variable (RV), and AHC. Working with simulation data is effective in illustrating the theoretical prospective of finite mixture models as we can generate and analyze with model characteristics already known.

4.1 The performance of model selection criteria

The first simulation is used to illustrate how model selection techniques can provide information regarding variance covariance structure and the number of components. The effects of sample size and the number of components are examined.

4.1.1 Simulation setting

Data is simulated from multivariates normal mixture model:

$$f(x;\theta) = \tau_1\phi_N(\mathbf{x};\mathbf{u_1},\mathbf{\Sigma_1}) + ... + \tau_K\phi_N(\mathbf{x};\mathbf{u_K},\mathbf{\Sigma_K}) \tag{4.1}$$

where for a fixed and given K, $\tau_1, ..., \tau_K$ are the mixing proportions for each component, $\mathbf{u_1}, ...\mathbf{u_K}$ and $\mathbf{\Sigma_1}, ..., \mathbf{\Sigma_K}$ are the mean vectors and dispersion matrices respectively. We consider two scenarios, K = 2 and K = 3. In both scenarios, for illustrative purpose, the number of features for each observation equals to 47, same as the number of features in real datasets. In order to examine the influence of sample sizes, different sample sizes are evaluated (n = 674, n = 1500, n=2000 and n=4000). We simulate the data assuming covariance structure follows a general family (VVV). The purpose of this simulation is to see whether model selection techniques can correctly select the true variance covariance structure and the true number of components, and what factors can influence the performance of model selection.

4.1.2 Simulation results

Table III and Table IV present the simulation results for BIC. When the simulated data has two components with general structure(Table III), we can see that BIC can correctly select the correct variance-covariance structure, no matter how small the sample size is. Regarding the number of components, the BIC favors K=2 over K=3 regardless of the sample size. In general, BIC scores are relatively close when comparing K=2 vs K=3. When the true number of components equals to 3 with the general structure(Table IV), BIC can still correctly identify

the correct variacne-covariance structure; however, BIC still favors K=2 over K=3, even when the sample size is large, eg, n=4000.

When examining the performance of ICL (Table V and Table VI), we notice that it has very similar results like BIC; at different scenarios, ICL can identify the right variance-covariance structure. However, it favors K=2 over K=3, when the true number of components equals to 3.

The results of LRTS are presented in Table VII and Table VIII. The LRTS is performed to assess the number of components for a specific variance covariance structure. In this simulation, our purpose is to evaluate LRTS under the general family (VVV) assumption. The first scenario when K=2: the null hypothesis is that K=1, and alternative hypothesis is that K=2. When sample size is small, e.g, n=674, we are not in a position to reject the null hypothesis, when sample size increases, the p-value becomes small, when n=4000, the model can reject the null and correctly claim the number of components. Similar performance is observed for true K=3. When sample size is large, the model is able to correctly identity the number of components.

4.2 The performance of different initialization strategies

The second simulation in this section is to compare our proposed MD-based initialization method with existing methods. It is observed that our proposed method can effectively capture the sub-populations, and thereby emphasize the disadvantages of using a single initial value or random initial value.

TABLE III

PERFORMANCES OF BIC ASSUMING GENERAL FAMILY (VVV) AND 2 COMPONENTS

K	n	Spherical Family (VII)	Diagonal Family (VVI)	General Family (VVV)
2	674	-169123.0	-165198.8	-103626.3
	1200	-300044.3	-292568.3	-174975.9
	2000	-498455.9	-484961.9	-283227.4
	4000	-999151.2	-971286.9	-551215.9
3	674	-169650.0	-165512.9	-103997.2
	1200	-280933.2	-276232.0	-182805.2
	2000	-466854.3	-458900.9	-291083.9
	4000	-937666.8	-919483.6	-560232.3

TABLE IV

PERFORMANCES OF BIC ASSUMING GENERAL FAMILY (VVV) AND 3 COMPONENTS

K	n	Spherical Family (VII)	Diagonal Family (VVI)	General Family (VVV)
2	674	-169123.0	-165198.8	-103626.3
	1200	-299940.9	-291691.1	-175281.4
	2000	-506206.5	-491789.5	-283767.5
	4000	-999533.0	-971462.5	-551520.9
3	674	-169650.0	-165512.9	-103997.2
	1200	-280933.2	-276502.2	-182335.6
	2000	-473605.0	-464914.3	-291271.8
	4000	-936515.2	-918863.3	-560365.4

TABLE V

PERFORMANCES OF ICL ASSUMING GENERAL FAMILY (VVV) AND 2 COMPONENTS

K	n	Spherical Family (VII)	Diagonal Family (VVI)	General Family (VVV)
2	674	-169655.8	-165202.5	-102837.5
	1200	-300051.1	-292404.8	-174308.6
	2000	-498475.4	-484868.9	-282686.0
	4000	-999151.2	-971286.9	-551215.9
3	674	-158112.6	-155628.0	-109012.2
	1200	-280944.7	-275861.2	-180935.7
	2000	-466876.0	-458686.7	-289876.9
	4000	-937666.8	-919483.6	-560232.3

TABLE VI

PERFORMANCES OF ICL ASSUMING GENERAL FAMILY (VVV) AND 3 COMPONENTS

K	n	Spherical Family (VII)	Diagonal Family (VVI)	General Family (VVV)
2	674	-169655.8	-165202.5	-102837.5
	1200	-299953.7	-291530.9	-174125.9
	2000	-506223.7	-491700.3	-282947.8
	4000	-999533.0	-971462.5	-551520.9
3	674	-158112.6	-155628.0	-109012.2
	1200	-281253.3	-276127.3	-180846.8
	2000	-473634.5	-464714.0	-289875.2
	4000	-936515.2	-918863.3	-560365.4

TABLE VII

LRTS FOR NUMBER OF COMPONENTS ASSUMING 2 COMPONENTS

		n=674	n=1200	n=2000	n=4000
1 vs 2	LRTS	1537.849	1731.284	2070.059	2060.015
	p-value	0.725	0.993	0.097	0.02

TABLE VIII

LRTS FOR NUMBER OF COMPONENTS ASSUMING 3 COMPONENTS

		n=674	n=1200	n=2000	n=4000
2 vs 3	LRTS	1483.865	1624.753	2074.921	2153.313
	p-value	0.639	0.682	0.008	0.001

4.2.1 Simulation setting

We simulate two parts of data sets separately. One part is to simulate multivariate Gaussian distribution, as the genome data. Another part is the survival data. In this simulation, the result from pure MD is used as ground truth. By doing so, we know the true membership assignment for each observation. To simulate the multivariate Gaussian distribution data, we use the similar setting as described in previous section. Two scenarios will be considered: K=2 and K=3, representing two components, and three components. To mimic the real data, sample size is set to 674, and the dimension of multivariate Gaussian is 47.

The survival time is generated based on exponential distribution. The median survival time is estimated based on pure MD classification using K-M method, and the dropout rate is based on real data as well.

4.2.2 Simulation results

TABLE IX

RESULTS FROM DIFFERENT INITIALIZATION STRATEGIES FOR THE SIMULATED DATASETS

K	Initialization	Ln(L)	o	ARI	π	P-value	System Time
2	MD	37114.08	21	0.5783	0.1928	0.0000468	67.63
	RV	36956.33	31	0.0716	0.4822	0.0012	102.95
	AHC	36987.86	26	0.2719	0.3397	0.0018	247.42
3	MD	37753.96	16	0.6215	0.2196	0.0034	64.73
	RV	37554.67	26	0.0258	0.3724	0.0064	192.84
	AHC	37585.22	22	0.1498	0.2953	0.19	250.64

From Table IX, we can compare our method with other existing methods. First of all, the likelihood function from the proposed method is slightly higher, suggesting a better model fitting. In addition, the number of iterations needed for our proposed method is smallest, which means our initialization methods converge faster, the total running time is the smallest compared to other methods. Further more, our proposed method has the highest ARI; it indicates that our proposed method can better identify the individual membership, and it can be validated by the OS separation, as our proposed method leads to better OS separation, i.e.

smaller p-value from log rank test. The KM curve for OS separation based on different methods is shown below.

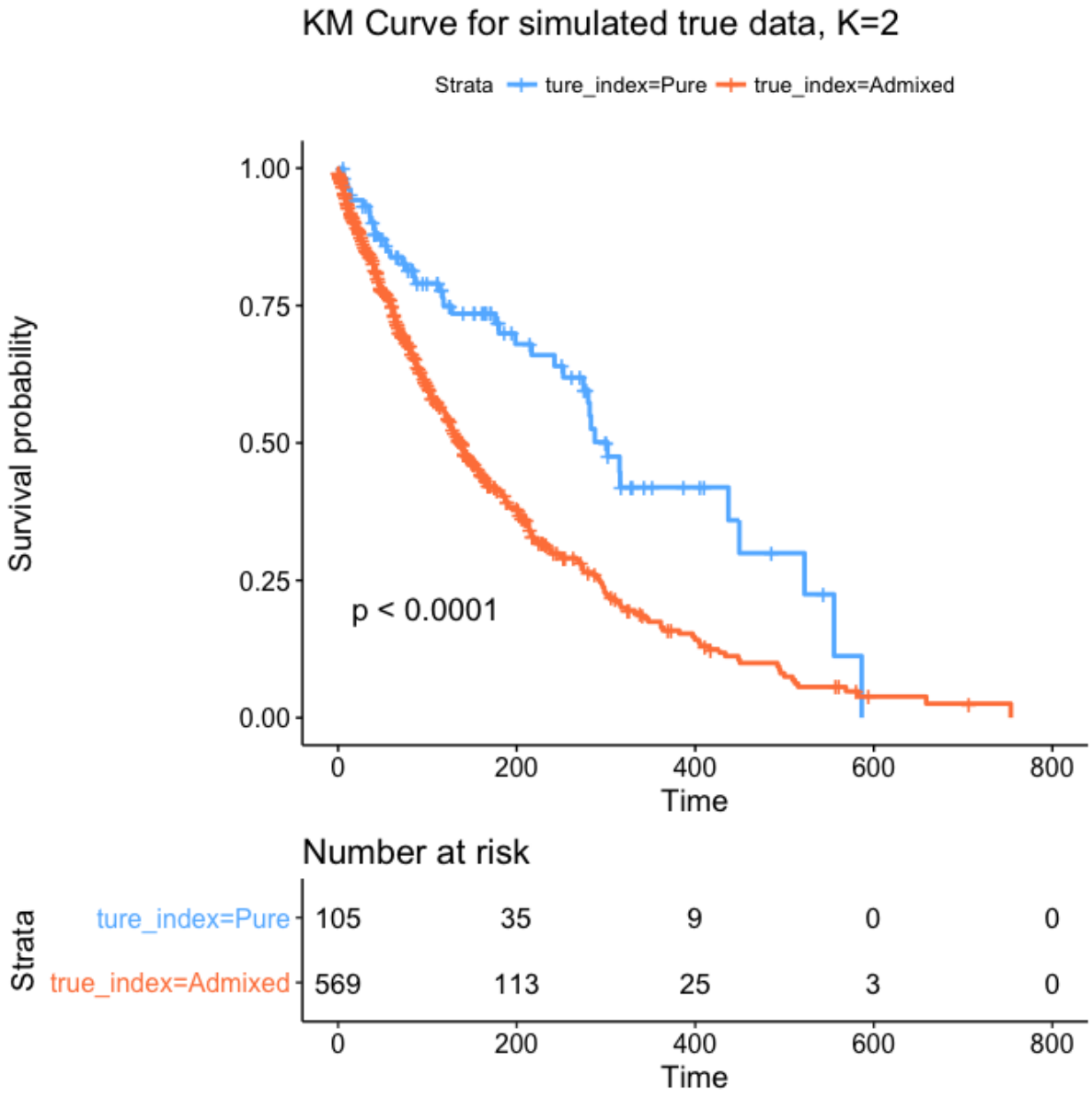

Figure 9. KM curve of true assignment assuming 2 components

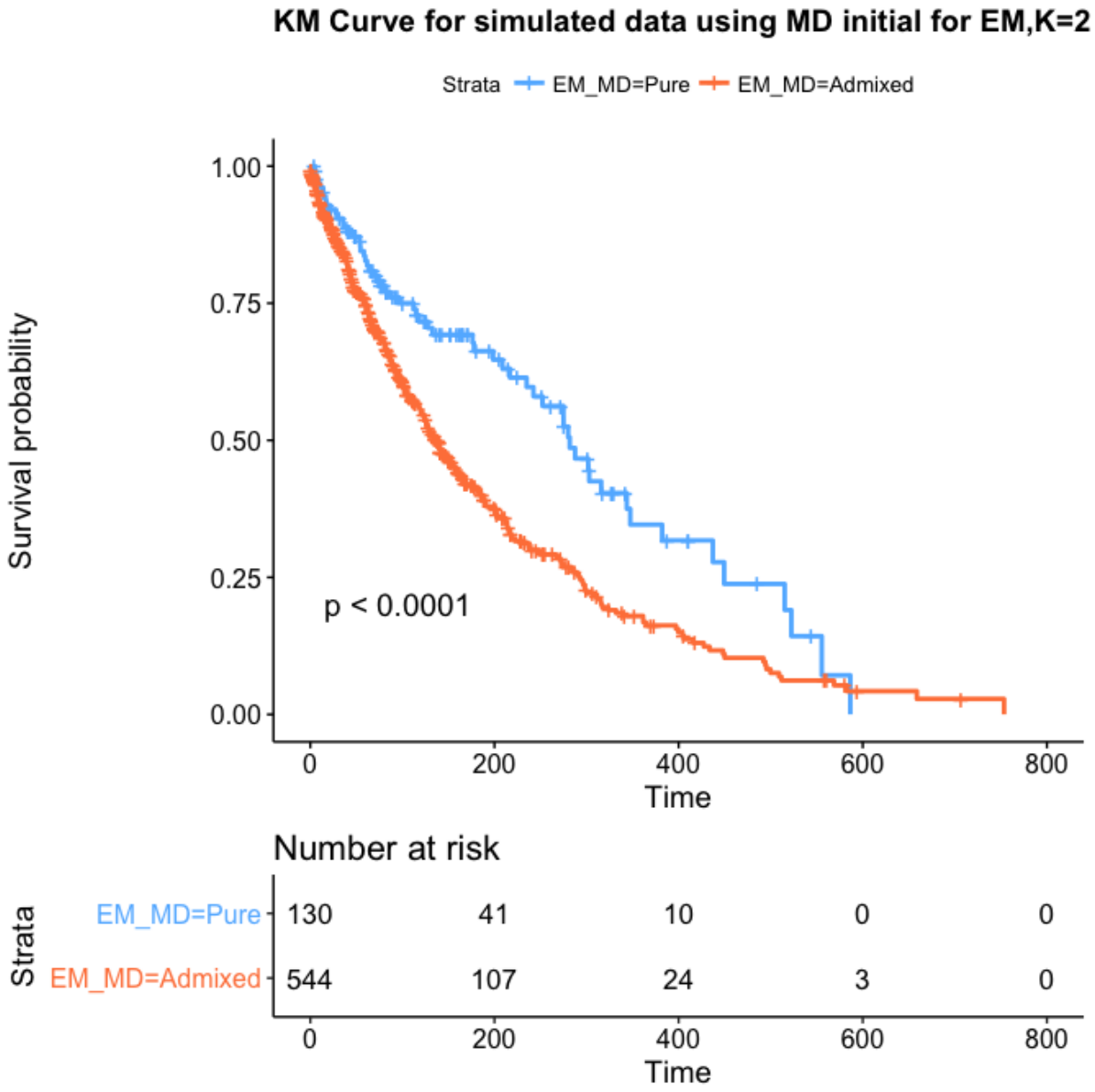

Figure 10. KM curve based on MD initialization assuming 2 components

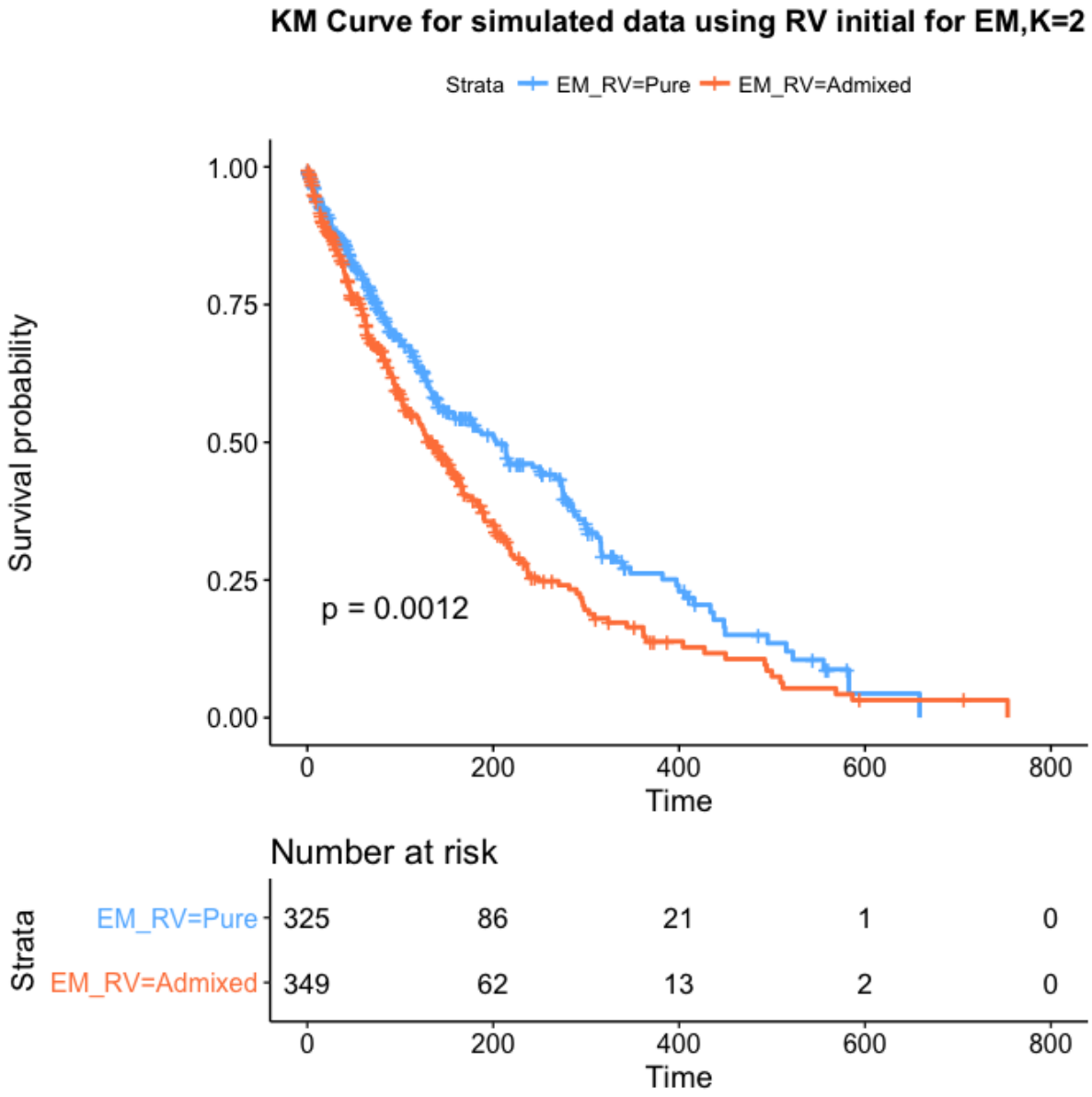

Figure 11. KM curve based on Random Variable initialization assuming 2 components

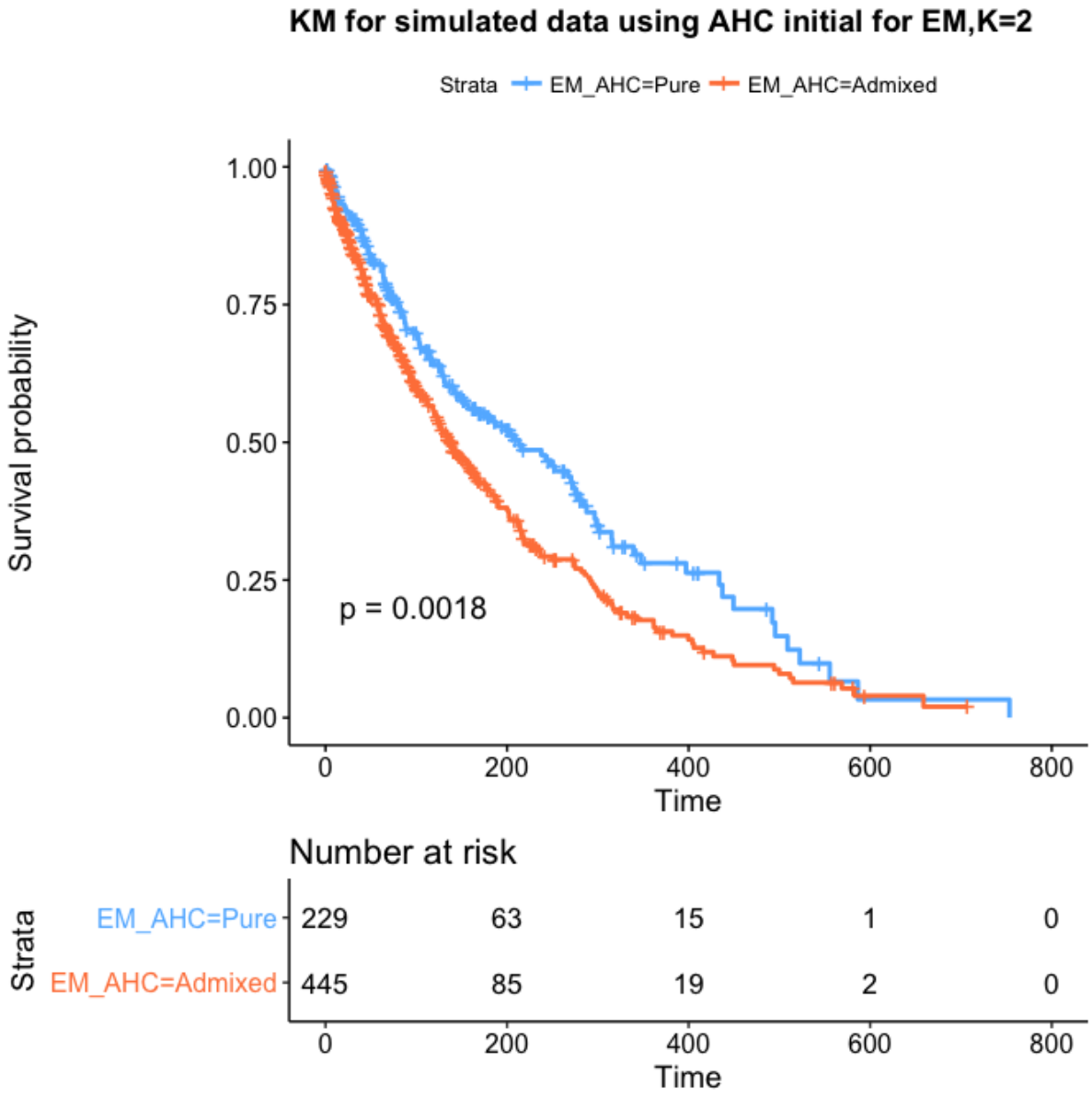

Figure 12. KM curve based on AHC initialization assuming 2 components

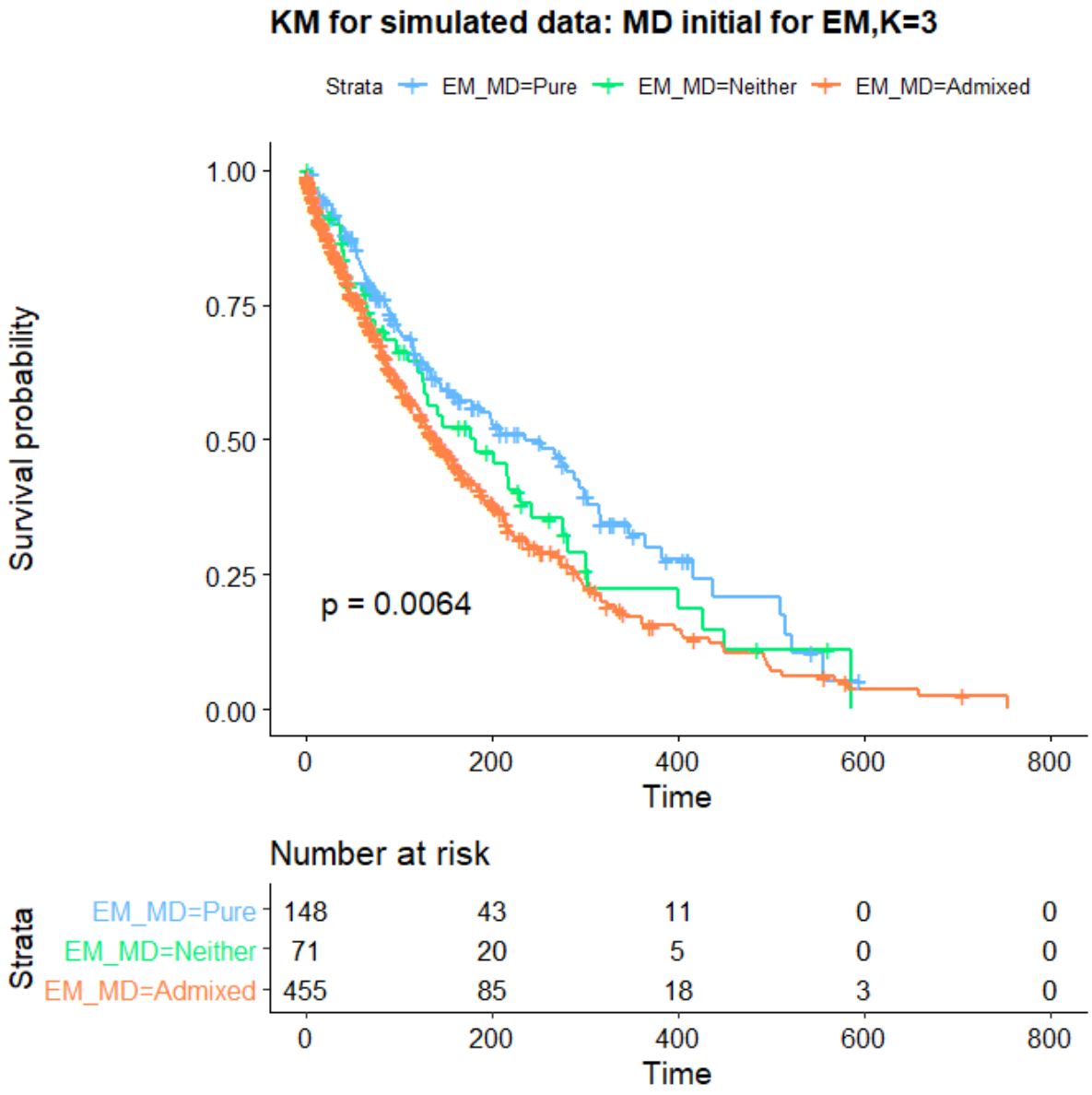

Figure 13. KM curve based on MD initialization assuming 3 components

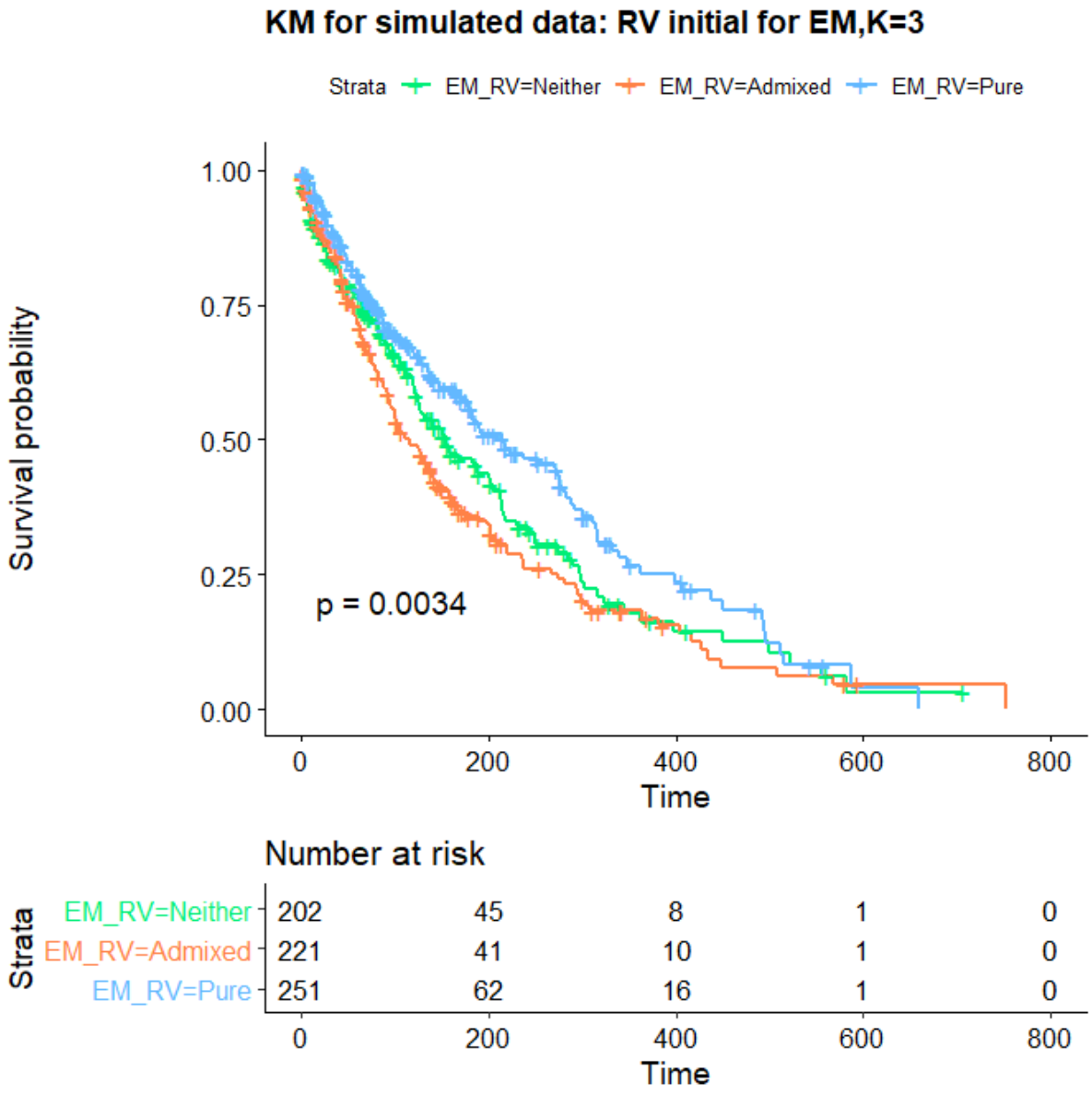

Figure 14. KM curve based on Random Variable initialization assuming 3 components

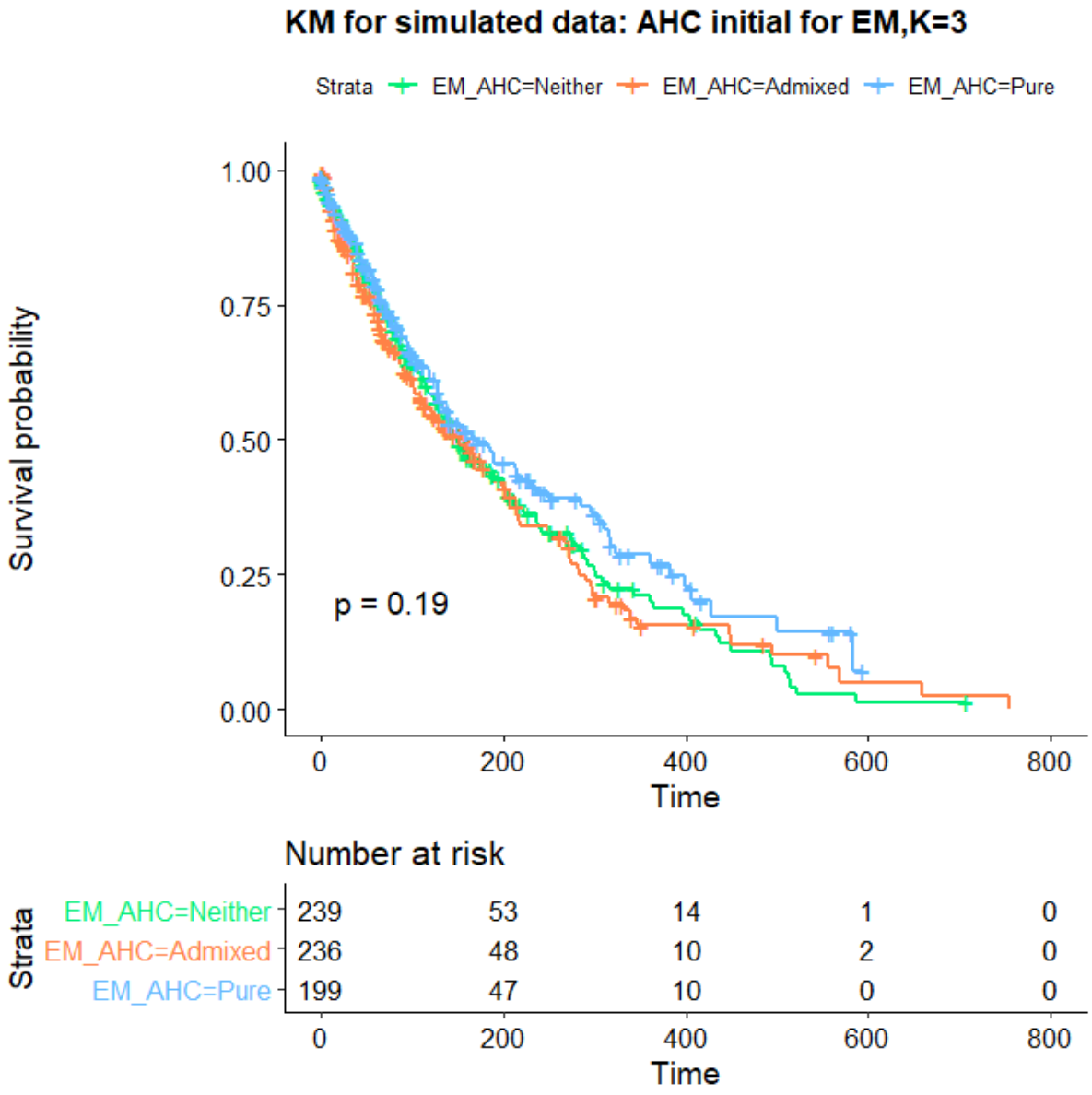

Figure 15. KM curve based on AHC initialization assuming 3 components

CHAPTER 5

EMPIRICAL DATA

5.1 Data representation

To satisfy the key assumption that $\mathbf{x}$ follows a normal distribution, the METABRIC gene expression datasets were previously median-centered and log-transformed; any missing, zero or negative gene expression values in METABRIC are replaced by real numbers randomly generated from a uniform distribution in the range of [.05,.95]. As regards TCGA, we apply the $\log_2(x+1)$ transformation for the centered TCGA gene expression data before analysis. In this thesis, METABRIC datasets are used for model development, and TCGA datasets are used for model validation, independently.

As an illustrative example for validity of the assumption, Figure 16 displays the empirical distribution of two genes from PAM50 gene expression profile. From both histogram and QQ plot, we see that $\mathbf{x}$ roughly follows a normal distribution. The Q-Q plot also confirms that the normal distribution assumption is valid.

To visualize the cluster patterns of different subtypes for high-dimensional gene expression data, we create a two-dimensional t-distributed Stochastic Neighbor Embedding (t-SNE) plots(121).

Figure 17 is the t-SNE plot using METABRICS PAM50 gene expression data; it models each subject with high dimensional features, (i.e. 47 genes) into a two-dimensional point, such

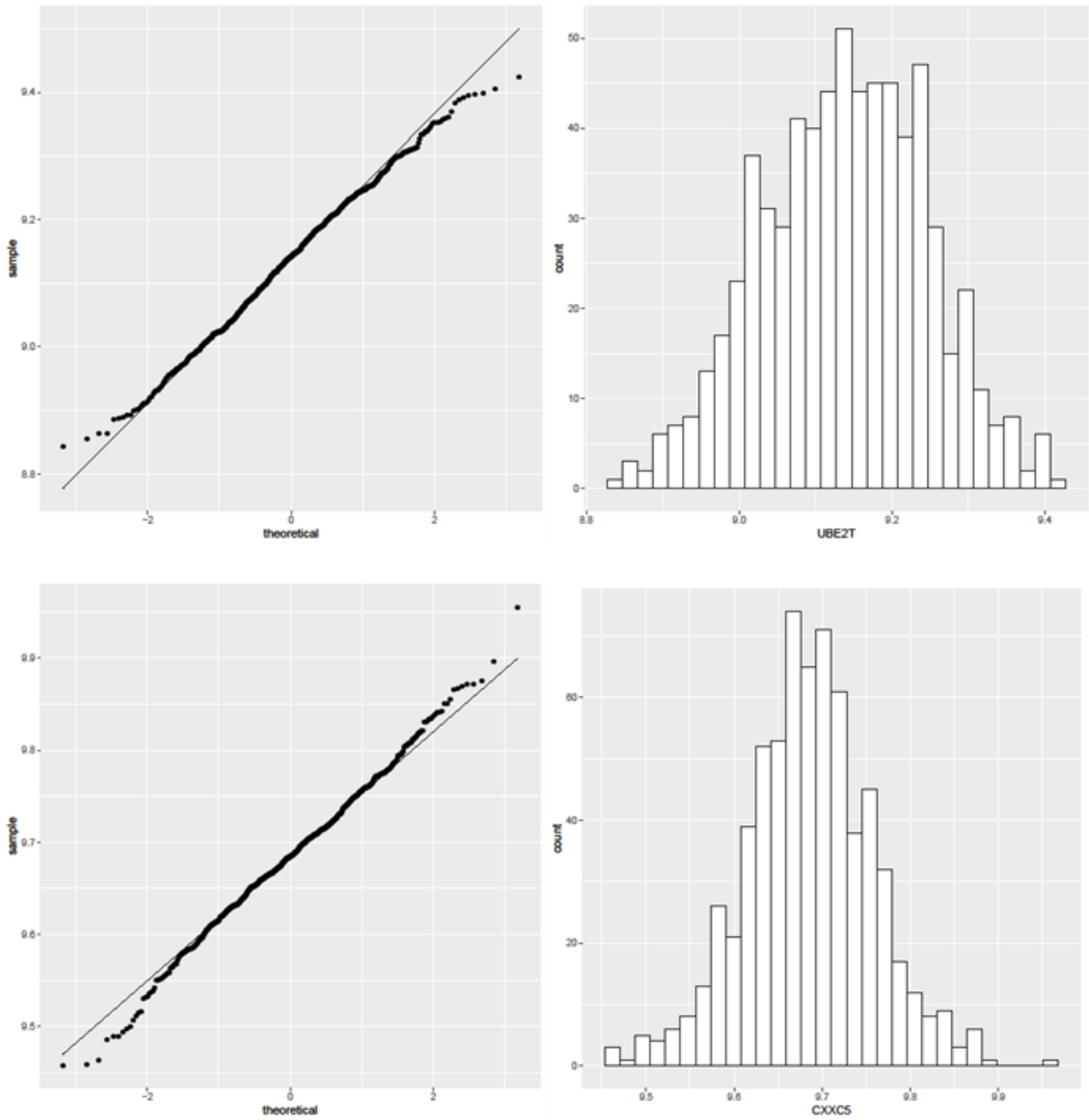

Figure 16. Histogram and QQ plot for PAM50 gene expression (UBE2T, CXXCS)

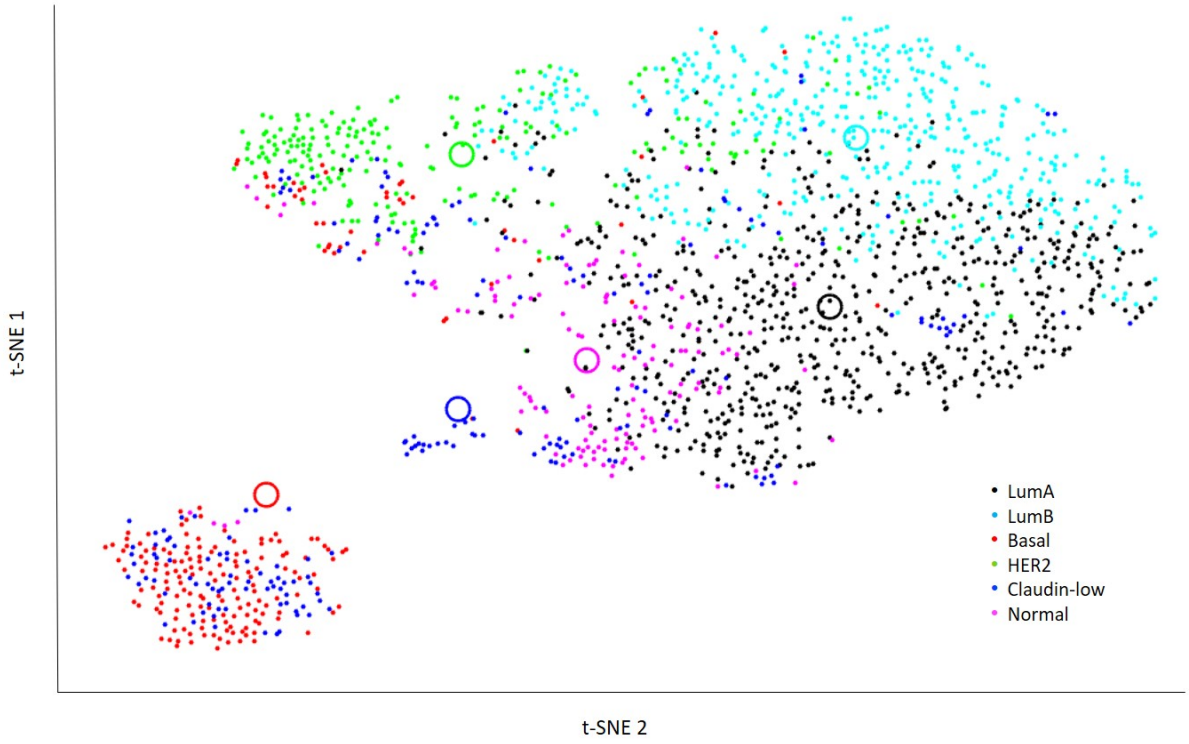

Figure 17. t-SNE plot for all subtypes in METABRIC cohort

that similar subjects are modeled as nearby points and distinct subjects are modeled as far away points with high probability. From Figure 17, we can see significant overlap between subtypes, for example, LumA are closely mixed with LumB and HER2, only Basal forms a relatively distinct subtype. In general, LumA is a relatively loose subtype; it has substantial admixed pattern across different subtypes.

5.2 Model selection results

In this section, we evaluate the performance of different model selection criteria and present how model selection techniques can determine the optimum number of components (K) and identify the variance-covariance structure (spherical,diagonal and general) for multivariate Gaussian mixture. To simplify the options of variance-covariance structure, we select the most flexible structure to represent each category, i.e. VII to represent the spherical family, VVI to represent the diagonal family and VVV to represent the general family. After this, we narrow down the number of variance covariance structure from previously mentioned 9 to 3. For the number of components K, we want to compare 2 components vs 3 components. When K=2, we assume one component contains the "pure" patients, another component contains the "admixed" patients. When K=3, one component contains the "pure" patients, one component contains the "neither" patients and the last component contains the "admixed" patients.

The results of AIC, BIC, ILC and log-likelihood for METABRIC datasets are shown in Table X, the results of Bootstrap LRTS for the number of mixture components for METABRIC are shown in Table XI.

TABLE X

COMPARISON OF MODEL SELECTION CRITERIA FOR METABRICAS DATA

Criteria	K	VII	VVI	VVV
AIC	2	37243.88	50444.25	66708.35
	3	37479.48	50562.86	64741.42
BIC	2	36803.09	49591.25	56097.749
	3	36820.55	53068.22	56823.259
ICL	2	41354.02	49813.25	56861.57
	3	34935.62	54916.93	57980.37
In(L)	2	21018	28175	37047
	3	21007	28208	37072

TABLE XI

LRTS FOR 2 COMPONENTS VS 3 COMPONENTS FOR METABRICS DATA

		VII	VVI	VVV
2 vs 3	LRTS	2900.748	2391.25	1472.073
	p-value	0.001	0.001	0.001

Table X shows the model section scores under different number of components and covariance structures. AIC, BIC and ICL all select the unconstrained structure (model VVV) over spherical (model VII) or diagonal (VVI), for both the cases of K=2 and K=3. The log-likelihoods based on unconstrained structure are the largest compared to other two structures, suggesting unconstrained structure can better represent the METABRIC data structure.

When the number of components are compared, i.e. K=2 vs. K=3, AIC favors K=2, while BIC and ICL favor K=3. However, when we compare the model selection scores of 2 and 3, AIC, BIC and ICL scores are all very close. Comparing the log-likelihood of 2 components and 3 components, log-likelihood of 3 components are slightly larger than 2 components. Unlike information criteria methods, the results of LRTS consistently favor 3 components, with p-value=0.001 under different variance-covariance structure assumptions. Model selection are conducted for TCGA datasets as well, the results are shown in Appendix A, Table XVI and Table XVII.

The results of model selection suggests that AIC, BIC and ICL all support the assumption of unconstrained structure (model VVV) in different scenarios. Regarding the number of components, information criteria based methods and LRTS have inconsistent results. However for both METABRIC and TCGA, the information criteria scores of K=2 compared to the score of K=3 are relatively close to each.

5.3 Comparison of the performances of different initialization strategies

Next, we compare the performance of different initialization strategies for EM in finite mixture Gaussian model. Three different initialization strategies are evaluated here: (a) Herein

proposed Mahalanobis distance (MD) based method for initialization; (b) 100 sets of random assignment for initialization, assuming the mixing proportions are equal for each component, and each individual is equally likely to be assigned to each component; (c) Agglomerative hierarchical clustering (AHC) initialization (84). The AHC provides only one set of initialization assignment. The threshold for assessing the convergence of log-likelihood during at EM algorithm is set to be 10^{-5}.

To examine the performance of different initialization strategies, the following criteria are used: (a) The log likelihood function (Ln(L)): higher log-likelihood means the stationary point of EM algorithm is more likely to be global optimum; (b) The number of iterations until EM algorithm converges (o): it can be seen as a measurement for good initialization, usually good initialization point requires less number of iterations; (c) total system time(seconds): includes the time for the techniques to provide the initial assignment and the time EM algorithm needs to converge, total system time is an indication of the computational efficiency.

Further more, we would like to compare abilities of different initialization strategies to cluster the experimental subjects. Overall survival (OS) separation among different components is used as validation for the clustering results. Our assumption is that a better initializing strategies can better identify the membership belonging of each individual, and thus has better OS separation. Here are the steps we used for OS comparison:

1. Assign Initial membership: apply different initializing strategies (MD, randome variable, AHC);

2. Parameter Estimation using EM:

- Use current values for parameters to evaluate posterior probabilities, for each data point;
- Use these probabilities to re-estimate means, covariance, and mixing coefficients;
- Repeat the algorithm until relative change in the likelihood is less than the threshold.

3. Final membership assignment: using the parameters estimated from EM, assign each subject to a cluster with highest posterior probability.

The following three criteria are used to evaluate the OS comparison: (a) The p-value from unstratified log-rank test for comparing OS of different components (124; 169); (b) The adjusted Rand index (ARI) (103). The ARI is a measure of membership assignment agreement between two methods. We would like to compare the membership agreement between different initial strategies with the classifications based on pure MD method proposed in Chapter 4. The ARI ranges between [-1,1]; when the assignment is completely random, ARI has zero expected value; when two assignments are identical, ARI equals to 1 (203). Note that the true membership assignment status is unknown; therefore, we use the membership assignment based on pure MD in Section 3.1.1 as reference; (c) The proportion of pure objects (τ).

Results of different EM initialization strategies for METABRICA are shown in Table XII. From model fitting perspective, the final models from different initialization strategies have a very similar log-likelihood. The highest log-likelihood function comes from AHC initialization, which makes sense, since the AHC initializes EM with the assignment that can provide the smallest decrease in the classification likelihood. In general, the likelihood results are comparable across different initialization strategies. For computational efficiency point of view, AHC

TABLE XII

RESULTS FROM DIFFERENT INITIALIZATION STRATEGIES FOR THE METABRICS

Initialization	K	Ln(L)	o	ARI	π	P-value	System time (s)
MD	2	36831	22	0.2784	0.2418	0.0000479	73.08
RV	2	36910	29	-0.0101	0.6157	0.0001834	134.72
AHC	2	36822	20	0.0499	0.4896	0.0015	273.99
MD	3	37723	15	0.4521	0.1884	0.0000113	55.13
RV	3	37463	24	0.0011	0.3249	0.0000181	115.79
AHC	3	38004	19	0.0539	0.2300	0.0012	255.036

takes much longer time to find the starting assignment and run EM until converges. Because AHC is a parametric method, at the initialization step, every time to make the decision of merging individual, the likelihood function has to be evaluated, it is a huge disadvantage for large sample size datasets. When judging the algorithm efficiency for random variable, it takes more iterations for the EM algorithm to converge, however the total system is shorter than AHC. Because random variable initialization uses uninformed starting assignment, the EM algorithm needs more iterations to find the optimum, but it takes less time to choose the initial value. Our proposed method, MD based EM, has the shortest system running time (73.08s vs 134.72s or 273.99s). It also requires less number of iterations compared to random variable. Two factors can contribute to this: first, MD based initialization is a non-parametric method, time can be saved significantly since no need to do model fitting at initialization step; second, MD based initialization utilizes the information from data to assign individual experimental units into a meaningful cluster; provides a good initialization point and thus needs less number of iterations to converge.

In addition, when OS separation is compared, the unstratified log rank test from MD based EM has the smallest p-value; it means better separation for two components with respect to the survival time. The MD based initialization has relatively smaller proportion of pure components, random variable based initialization has a larger proportion of pure component. The ARI for random variable initialization is considered low, indicating the classifications agree poorly between the random variable initialization and pure MD. The AHC initialization has low ARI as well, indicates a poor agreement with pure MD method. In contrast, MD based initialization has relatively better agreement with pure MD method; for the present data set, assuming three, the ARI of MD based method is 0.4521.

The results of different EM initialization strategies for TCGA is shown in the Appendex A, Table XVIII.

5.4 Clinical and molecular characteristics result

In this section, we present clinical characteristic results based on our proposed MD initialization, including: hazard ratio for the overall mortality of METABRIC, the Kaplan-Meier plots of overall survival, and clinical characteristics comparison for "pure" vs "admixed", including age, tumor size and tumor stage.

Results of unadjusted Cox model are presented in Table XIII and Table XIV. We also present adjusted Cox model results, adjusting for age, grade, stage and tumor size, both with 2-sided 95% CIs.

In the unadjusted model with 3 components, the hazard ratio for admixed cases relative to pure cases is 2.2461; it means the risk of death event for admixed subjects is 2.25 times higher

TABLE XIII

HR AND 95 % CI FOR OVERALL MORTALITY FOR METABRICS K=2

Criteria	Subgroups	HR (unadjusted)	95% CI (unadjusted)	HR (adjusted)	95% CI (adjusted)
EM-MD	pure	1	-	1	-
	admixed	1.7580	(1.335,2.316)	1.2880	(0.9755,1.701)
EM-RV	pure	1	-	1	-
	admixed	1.5007	(1.218 ,1.849)	1.195621	(0.969 ,1.475)
EM-AHC	pure	1	-	1	-
	admixed	1.5701	(1.274,1.935)	0.9609	(0.7724,1.195)

TABLE XIV

HR AND 95 % CI FOR OVERALL MORTALITY FOR METABRICS K=3

Criteria	Subgroups	HR (unadjusted)	95% CI (unadjusted)	HR (adjusted)	95% CI (adjusted)
EM-MD	pure	1	-	1	-
	neither	1.5959	(1.162,2.192)	1.1597	(0.8414,1.598)
	admixed	2.2461	(1.593,3.167)	1.4027	(0.9873,1.993)
EM-RV	pure	1	-	1	-
	neither	1.5317	(1.174,1.998)	1.1597	(0.8414,1.598)
	admixed	1.8051	(1.400,2.327)	1.4027	(0.9873,1.993)
EM-AHC	pure	1	-	1	-
	neither	1.3148	(1.031,1.677)	1.0811	(0.8455,1.328)
	admixed	1.6197	(1.246,2.106)	1.1195	(0.8567,1.463)

than the pure subjects, with 95% CI: (1.1593,3.167), and that for neither to pure was 1.5959, with 95% CI: 1.162,2.192). After adjusting for age, grade, stage and tumor size, the hazard ratio for admixed was 1.4027, but not statistically significant.

TABLE XV

CLINICAL CHARACTERISTIC OF LUMA PATIENTS IN METABRICS, CLASSIFIED BY MD INITIALIZATION

Variables	Pure	Neither	Admixed
Age (years)	58.5	62.8	65.8
Pre-Menopausal(%)	31%	16%	16%
post-Menopausal(%)	69%%	84%	84%
ER+(%)	78%	77%	74%
PR+(%)	54%	53%	55%
HER2+(%)	9%	15%	13%
Tumor Size(mm)	21.2	24.7	23.8
Grade(score)	1.5	1.7	1.4
Node positive (%)	40%	41%	46%
Tumor Stage I(%)	38%	31%	27%
Proliferation Score	8.99	9.05	9.05
Recurrence Score	28.8	58.7	62.7
Mutational load	5.29	5.36	5.64

From Table XV, admixed subjects are on an average 7.3 years older and more likely to be post-menopausal (84% vs 69%). While no differences are observed for clinically ER, PR or HER2 status, admixed subjects are more likely to have larger tumors (23.8 vs. 21.1), and a higher prevalence of node positivity (46% vs. 40%). Finally, pure subjects are more likely to

be at stage I at diagnosis (38% vs.27%). In general, neither subjects tend to have intermediate results for these clinical features.

Next, we present the KM plots for comparing different initialization methods assuming K=2, and K=3. The KM curves based on our proposed MD-based method display obvious survival separation from the very beginning, in addition, the pure subjects have much higher survival probability when they are followed for a long period of time. When K=2, the median survival time for pure subjects is 220 months compared to 169 months for admixed subjects. When K=3, the median survival time for pure subjects is 254 months, and 151 months for admixed subjects.

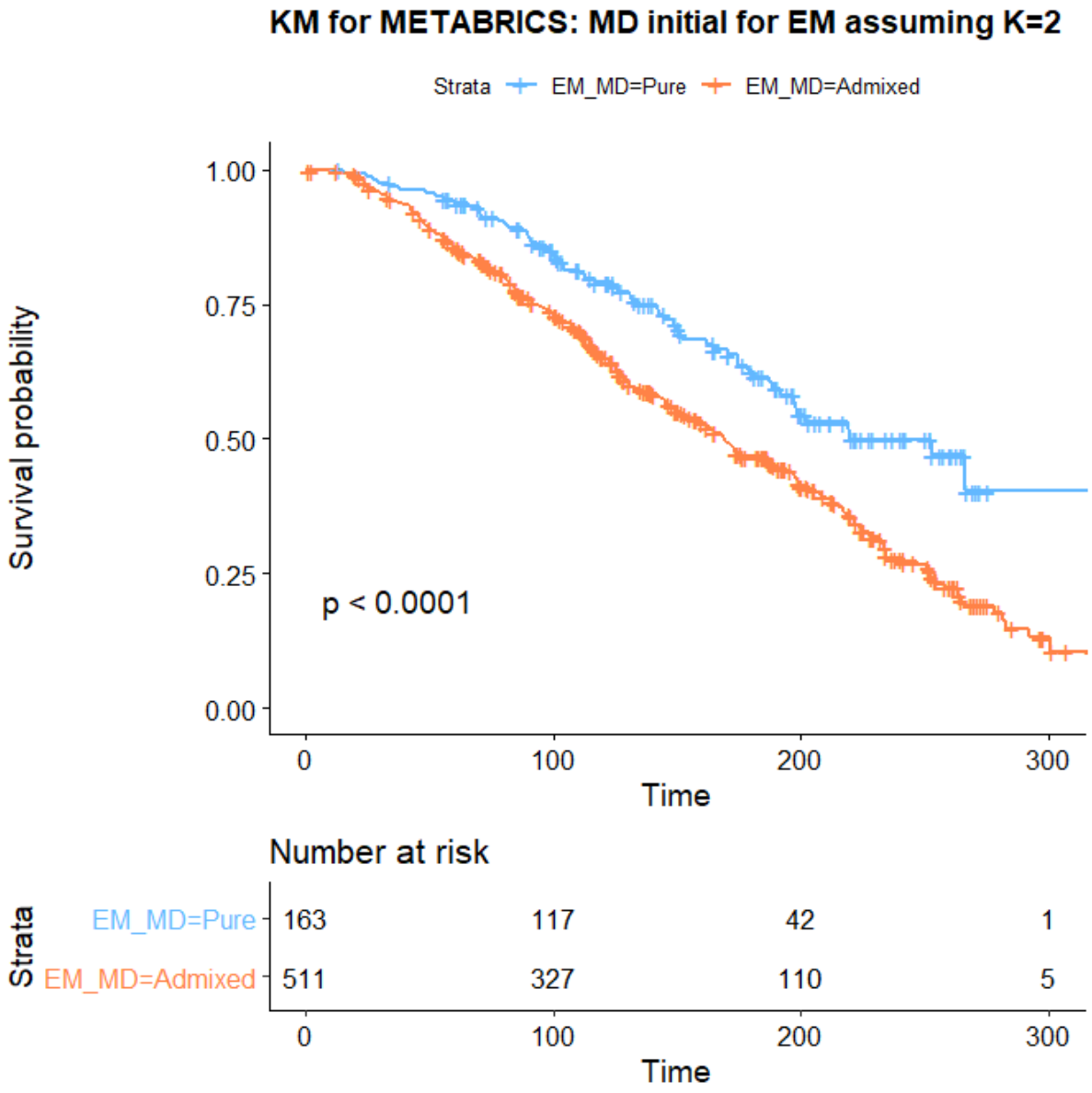

Figure 18. KM curve based on MD initialization assuming 2 components

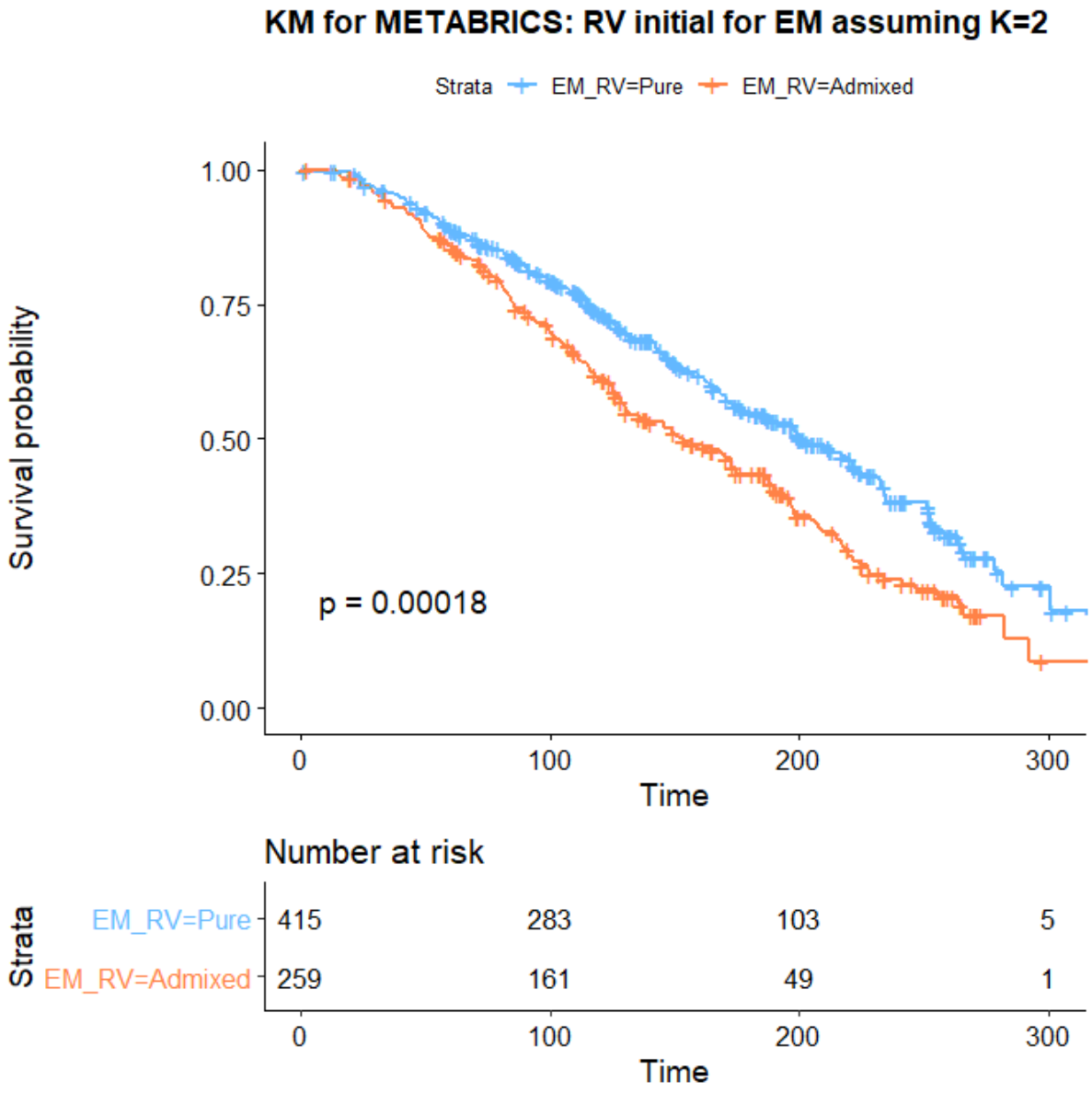

Figure 19. KM curve based on Random Variable initialization assuming 2 components

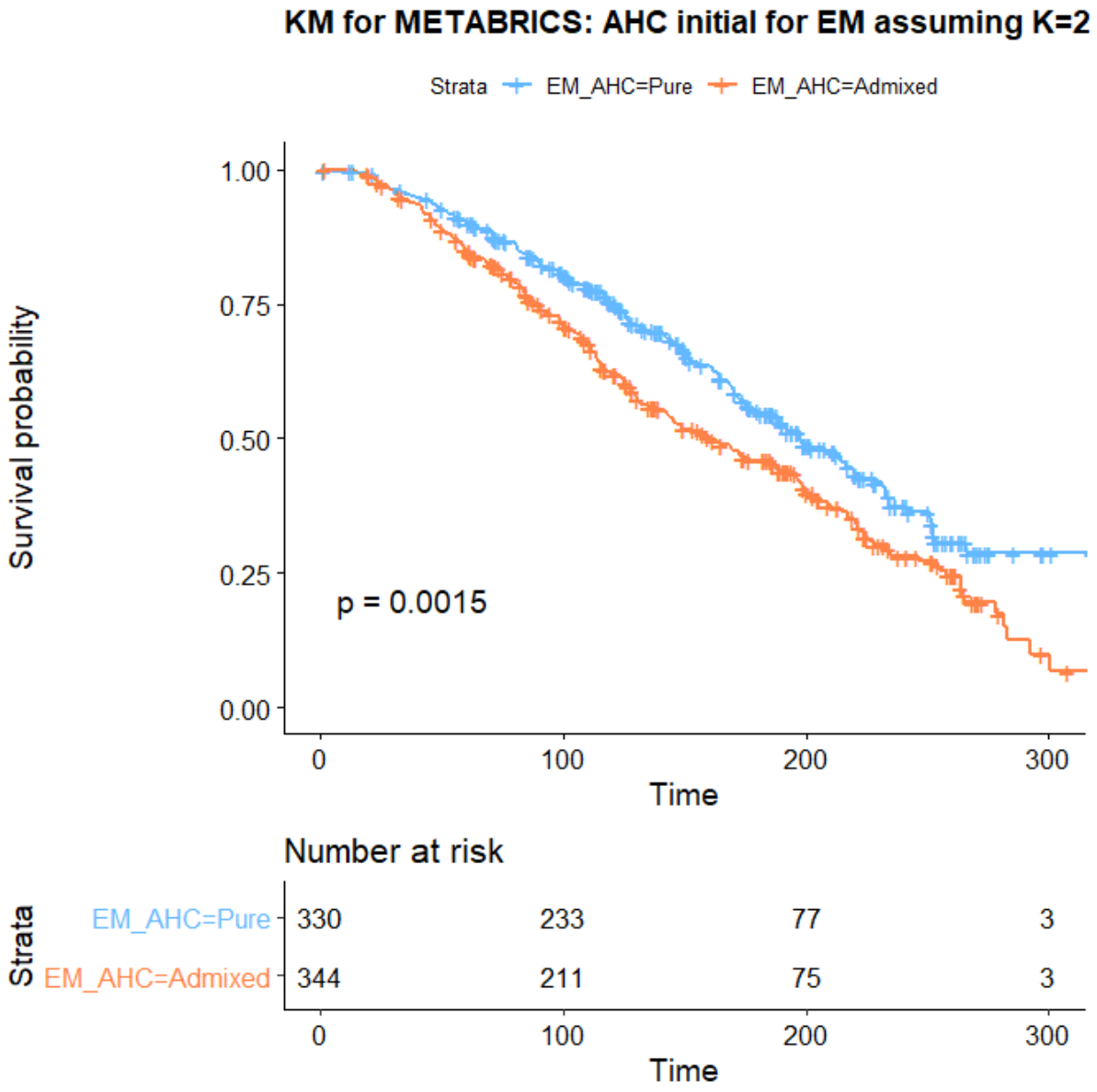

Figure 20. KM curve based on AHC initialization assuming 2 components

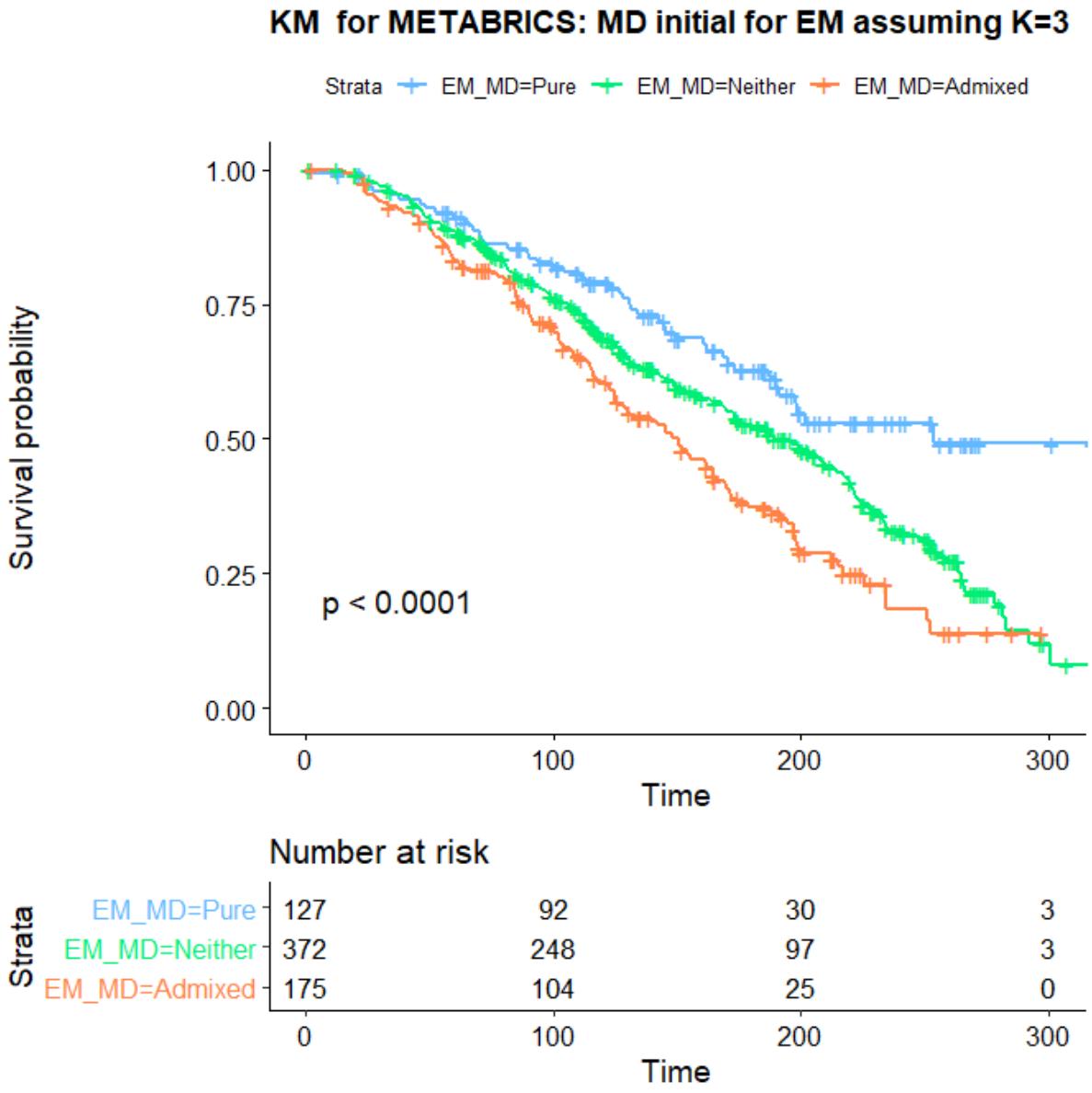

Figure 21. KM curve based on MD initialization assuming 3 components

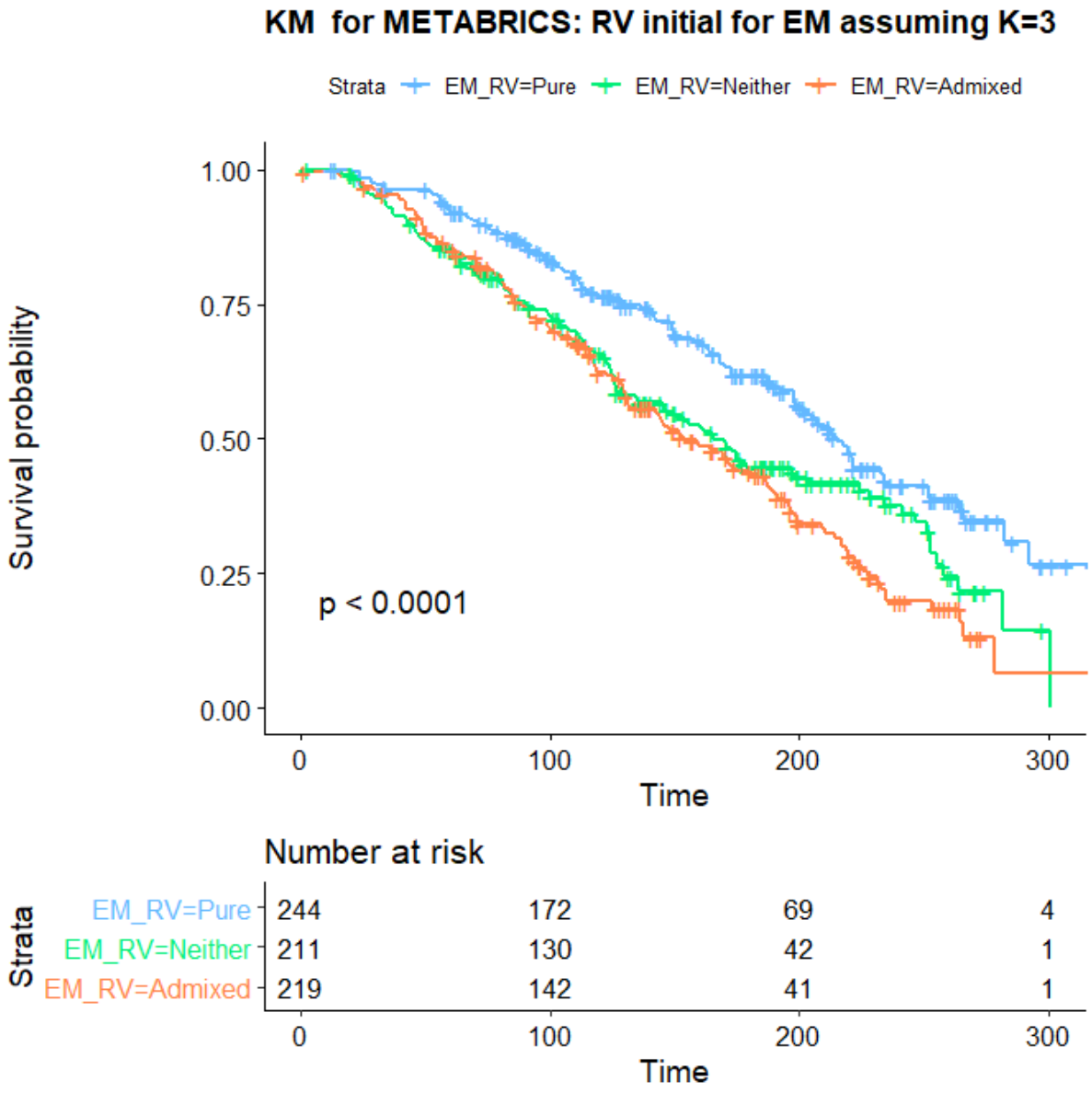

Figure 22. KM curve based on Random Variable initialization assuming 3 components

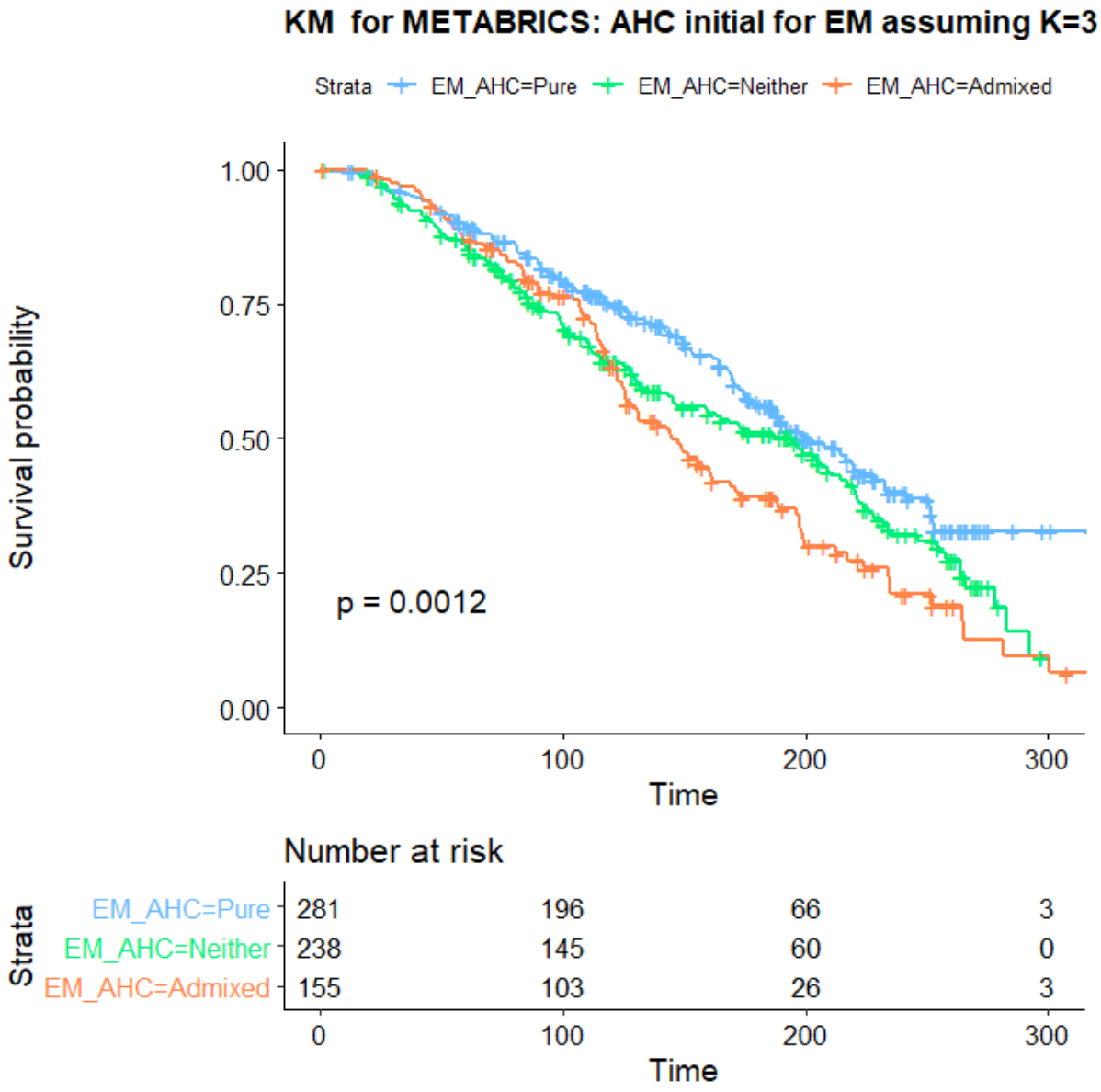

Figure 23. KM curve based on AHC initialization assuming 3 components

CHAPTER 6

CONCLUSION AND FUTURE WORK

Breast cancer poses a major threat to public health. Ample evidences suggest intratumor heterogeneity of breast cancer impedes our ability to predict response of targeted therapy for individual patients. A detailed understanding of how to capture intratumor heterogeneity of breast cancer is fundamental to the development and application of treatments for these conditions.

A critically important objective in intratumor heterogeneity research is the development of quantitative measures for intratumor heterogeneity in order to get optimal therapeutic benefit. As one of the modern gene sequencing techniques, gene expression profile, particularly, PAM50 has gained immense popularity in clinical practices due to its capacity to provide clinical information regarding prognosis and treatment response.

However, PAM50 assumes that each patient belongs to a single discrete subtype and provides no further information regarding intratumor heterogeneity. Statistical method for analysis of intratumor heterogeneity based on PAM50 is underdeveloped. There are three major methodological challenges in intratumor heterogeneity studies. First, as the gene expression data is high dimensional, appropriate methods need to be developed to address the issues of curse of dimensionality. Second, since the the finite mixture Gaussian model has latent variable, to implement EM for parameter estimations, an appropriate initialization strategy is needed. Last but not the least, the optimum model being unknown, model selection methodology should

be developed to determine the best fitting model. In this work, we proposed an innovative approach including data analysis, statistical inference and model selection for high dimensional gene expression data and applied it to real datasets.

In Chapter 5, we compared our proposed method with other two major initialization strategies in real datasets. Our results indicate that the proposed method can incorporate the information from data, but does not impose any restrictions on the form of the distribution of data; in the mean time, our model is able to provide multiple initial values, thus the model is more likely to locate the global optimum. Our proposed method improves the computation efficacy, as it does no need to do model fitting or grid search at initialization step; in addition, our method utilizes the information from data to assign each individual patient into a meaningful cluster, thus need less iterations to converge; more importantly, our method is more likely to locate the global optimum. A number of initialization strategies have been proposed in the past two decades. It is still challenging to choose the one that suits multivariate Gaussian mixture model.

Regarding the optimum model determination, the model selection techniques are able to identify the variance covariance structure. However, with respect to the number of components, there is no conclusive suggestion to make. In the simulated study, we showed that when the sample size is large enough, LRTS method is able to correctly identity the number of components.

Upon completion of this thesis, there remains a number of potential areas for future research. Indeed, the development of model-based methods for clustering analysis presents many

possibilities for future work. For example, the optimum number of components, in the future, we may incorporate prior information using Bayesian method and select the optimum model using Bayesian criteria. In addition, a larger and sufficiently informative dataset would be necessary to identify the optimum number of components and parameter estimation. Finally, our method requires the sample size to be large enough to avoid the curse of dimension, however, if we want to extend our proposed method to very-high-dimension data with relative small sample size, this cannot be easily done. Therefore, it will be very helpful to inspect the use of other optimization methods that can be implemented in very-high-dimension data.

APPENDICES

Appendix A

TCGA RESULTS

TABLE XVI

COMPARISON OF MODEL SELECTION CRITERIA FOR TCGA DATA

Criteria	K	VII	VVI	VVV
AIC	2	-106540.7	-95710.55	-72948.35
	3	-102253.4	-93595.71	-73604.22
BIC	2	-106999.1	-96605.09	-84746.55
	3	-102943.2	-94939.75	-91303.74
ICL	2	-107004.707	-96213.531	-82672.697
	3	-102952.9	-94025.67	-88314.917
In(L)	2	-132782.7	-124158.6	-109126.4
	3	-138760.0	-126098.6	-107812.9

TABLE XVII

LRT FOR 2 COMPONENTS VS 3 COMPONENTS FOR TCGA DATA

		VII	VVI	VVV
2 vs 3	LRTS	4392.305	2839.264	2912.216
	p-value	0.001	0.001	0.001

Appendix A (Continued)

TABLE XVIII

RESULTS FROM DIFFERENT INITIALIZATION STRATEGIES FOR THE TCGA

Initialization	K	Ln(L)	o	ARI	π	P-value	System time (s)
MD	2	-25722.94	21	0.0202	0.5603	0.0034	48.83
RV	2	-28417.74	31	-0.0345	0.6990	0.0014	100.76
AHC	2	-27745.32	23	0.01766	0.4396	0.0480	104.03
MD	3	-26908.45	16	0.4554	0.4139	0.01283	39.89
RV	3	-26580.13	26	0.0479	0.4554	0.01236	115.79
AHC	3	NA	NA	NA	NA	NA	120.60

Appendix B

R CODE

```
plot.em.run <- function(run, x)
{

        z <- apply(run$estep$z, 1, function (x) which.max(x))
        plot.mixture(locs=t(run$mstep$parameters$mean), z=z, obs=x)
}

plot.mixture <- function(locs, z, obs)
{
        #stopifnot(dim(obs)[2]==2)
        z <- as.factor(z)
        df1 <- data.frame(x=obs[,1], y=obs[,2], z=z)
        df2 <- data.frame(x=locs[,1], y=locs[,2])
        p <- ggplot()
        p <- p + geom_point(data=df1, aes(x=x, y=y, colour=z), shape=16, size=2, alpha=0.75)
        p <- p + geom_point(data=df2, aes(x=x, y=y), shape=16, size=3)
        #p <- p + opts(legend.position="none")
        p
}

run.em.VII <- function(k, x, init)
{
        # compute the number of data points
        n <- dim(x)[1]
        # initialize each data point to a random cluster
```

Appendix B (Continued)

```
        init.z<-init
        #init.z <- unmap(sample(1:k, n, replace=T))
        # compute the first "m step" with those posteriors
        mstep <- mstep(modelName="VII", data = x, z = init.z)
        estep <- estep(modelName="VII", data=x, parameters=mstep$parameters)
        iter <- 1

        lhood <- data.frame(iter=iter, lhood=estep$loglik)
        repeat
        {
                iter <- iter + 1
                mstep <- mstep(modelName="VII", data=x, z=estep$z)
                estep <- estep(modelName="VII", data=x, parameters=mstep$parameters)
                lhood <- rbind(lhood, c(iter=iter, lhood=estep$loglik))
                conv <- abs((lhood[iter,"lhood"] - lhood[iter-1,"lhood"]) / lhood[iter-1,"lhood"])
                # cat(sprintf("%03d : %02.3g\n", iter, conv))
                if (conv < 1e-5) break
        }
        list(estep=estep, mstep=mstep, lhood=lhood)
}

run.em.VVI <- function(k, x,init)
{
        # compute the number of data points
        n <- dim(x)[1]
        # initialize each data point to a random cluster
        init.z<-init
        #init.z <- unmap(sample(1:k, n, replace=T))
        # compute the first "m step" with those posteriors
        mstep <- mstep(modelName="VVI", data = x, z = init.z)
```

Appendix B (Continued)

```
        estep <- estep(modelName="VVI", data=x, parameters=mstep$parameters)
        iter <- 1

        lhood <- data.frame(iter=iter, lhood=estep$loglik)
        repeat
        {
                iter <- iter + 1
                mstep <- mstep(modelName="VVI", data=x, z=estep$z)
                estep <- estep(modelName="VVI", data=x, parameters=mstep$parameters)
                lhood <- rbind(lhood, c(iter=iter, lhood=estep$loglik))
                conv <- abs((lhood[iter,"lhood"] - lhood[iter-1,"lhood"]) / lhood[iter-1,"lhood"])
                # cat(sprintf("%03d : %02.3g\n", iter, conv))
                if (conv < 1e-5) break
        }
        list(estep=estep, mstep=mstep, lhood=lhood)
}

run.em.VVV <- function(k, x,init)
{
        # compute the number of data points
        n <- dim(x)[1]
        # initialize each data point to a random cluster
        init.z<-init
        #init.z <- unmap(sample(1:k, n, replace=T))
        # compute the first "m step" with those posteriors
        mstep <- mstep(modelName="VVV", data = x, z = init.z)
        estep <- estep(modelName="VVV", data=x, parameters=mstep$parameters)
        iter <- 1

        lhood <- data.frame(iter=iter, lhood=estep$loglik)
```

Appendix B (Continued)

```
        repeat
        {
                iter <- iter + 1
                mstep <- mstep(modelName="VVV", data=x, z=estep$z)
                estep <- estep(modelName="VVV", data=x, parameters=mstep$parameters)
                lhood <- rbind(lhood, c(iter=iter, lhood=estep$loglik))
                conv <- abs((lhood[iter,"lhood"] - lhood[iter-1,"lhood"]) / lhood[iter-1,"lhood"])
                # cat(sprintf("%03d : %02.3g\n", iter, conv))
                if (conv < 1e-5) break
        }
        list(estep=estep, mstep=mstep, lhood=lhood,iter=iter)
}




# compare KM curve, get P value from pure MD
threshold=seq(0.01,0.95,length=95)
luma_thresh<-matrix(NA,nrow=95,ncol=4)
for (i in 1:95){

        thresh = threshold[i]
        thresh_basal = quantile(Basal_MD$Basal, thresh)
        thresh_her2 = quantile(Her2_MD$Her2, thresh)
        thresh_la = quantile(LumA_all$LumA, thresh)
        thresh_lb = quantile(LumB_MD$LumB, thresh)
        thresh_normal = quantile(Normal_MD$Normal, thresh)
        pidx <- which(LumA_all$LumA<= thresh_la
        & LumA_all$Basal >= thresh_basal
        & LumA_all$Her2 >= thresh_her2
        & LumA_all$LumB >= thresh_lb
```

Appendix B (Continued)

```
        & LumA_all$Normal >= thresh_normal)

        LumA_all$pureMD_index<-"admixed"
        LumA_all$pureMD_index[pidx] <- "pure"



        fit <- survfit(Surv(time = LumA_all$OS_MONTHS, event = LumA_all$OS_STATUS == "DECEASED") ~
LumA_all$pureMD_index,
        data = LumA_all)

        luma_thresh[i,1] <- thresh
        luma_thresh[i,2] = surv_pvalue(fit, LumA_all)$pval
        luma_thresh[i,3] <- length(pidx)
        luma_thresh[i,4] <- dim(LumA_all)[1]-length(pidx)
        i=i+1
}



lumaBIC_1 <- mclustBIC(LumA, prior = priorControl(),c("VII", "VVI","VVV"),G=seq(from=2,to=5,by=1))
lumaBIC_1
plot(lumaBIC_1)

#AIC
IC <- Mclust(data=LumA,c("VII"), prior = priorControl(),G=2)
IC1 <- Mclust(data=LumA, c("VII"),prior = priorControl(),G=3)
aic_VII <- 2*IC$df - 2*IC$loglik
aic_VII

aic1_VII <- 2*IC1$df - 2*IC1$loglik
aic1_VII
```

Appendix B (Continued)

```
IC_VVI <- Mclust(data=LumA,c("VVI"), prior = priorControl(),G=2)
IC1_VVI <- Mclust(data=LumA, c("VVI"),prior = priorControl(),G=3)
aic_VVI <- 2*IC_VVI$df - 2*IC_VVI$loglik
aic_VVI

aic1_VVI <- 2*IC1_VVI$df - 2*IC1_VVI$loglik
aic1_VVI

IC_VVV <- Mclust(data=LumA,c("VVV"), prior = priorControl(),G=2)
IC1_VVV <- Mclust(data=LumA, c("VVV"),prior = priorControl(),G=3)
aic_VVV <- 2*IC_VVV$df - 2*IC_VVV$loglik
aic_VVV

aic1_VVV <- 2*IC1_VVV$df - 2*IC1_VVV$loglik
aic1_VVV

# LRT

VII_boot = mclustBootstrapLRT(LumA, model = "VII",maxG=5)
VII_boot
plot(EII_boot, G = 2)
plot(EEV_boot , G = 2)
plot(EEV_boot , G = 3)


VVI_boot = mclustBootstrapLRT(LumA, model = "VVI",maxG=5)
VVI_boot
```

Appendix B (Continued)

```
VVV_boot = mclustBootstrapLRT(LumA, model = "VVV", maxG=5)
VVV_boot

ICL_VII <- mclustICL(LumA,modelNames=c("VII"))
summary(ICL_VII)

ICL_VVI <- mclustICL(LumA,modelNames=c("VVI"))
summary(ICL_VVI)

ICL_VVV <- mclustICL(LumA,modelNames=c("VVV"))
summary(ICL_VVV)
plot(ICL_VVV)



thresh = 0.74*100
temp<- LumA_thresh_2[,thresh+1]
temp_neg<-1-temp
init<-as.matrix(cbind(temp,temp_neg))
VVV <- run.em.VVV(k=2, x=LumA,init=init)
VVV$iter

z.VVV <- apply(VVV$estep$z, 1, function (x) which.max(x))
lls.VVV <- VVV$estep$loglik
lls.VVV
LumA_all$EM_MD_2=as.factor(z.VVV)



# EM results virilization
plot.em.run(VVV, LumA)
plot.mixture(locs=t(VVV$mstep$parameters$mean), z=z.VVV, obs=x)
```

Appendix B (Continued)

```
#the best model's KM plot for EM_MD assuming 2 components, size 672*672
MD_KM<-Surv(time = LumA_all$OS_MONTHS, event =   LumA_all$OS_STATUS=="DECEASED")
MD_kmfit = survfit(MD_KM ~ LumA_all$EM_MD_2)
summary(MD_kmfit, times = c(seq(0, 150, by = 10)))

fit_MD <- survfit(Surv(time =  LumA_all$OS_MONTHS, event = LumA_all$OS_STATUS == "DECEASED") ~ LumA_all
data =LumA_all)
ggsurvplot(fit_MD, data = LumA_all, risk.table = TRUE, pval = TRUE, palette = c( "sienna1","steelblue1")



x<-LumA
la_mixture_thresh<-matrix(NA,nrow=95,ncol=3)
ptm<-proc.time()
for (i in 1:95)
{
        n <- dim(x)[1]
        set.seed(i+3000)
        # initialize each data point to a random cluster
        init<- unmap(sample(1:2, n, replace=T))

        VVV <- run.em.VVV(k=2, x=LumA, init=init)
        z.VVV <- apply(VVV$estep$z, 1, function (x) which.max(x))
        lls.VVV <- VVV$estep$loglik
        LumA_all$EM_RV_2=z.VVV
        fit.VVV <- survfit(Surv(time =  LumA_all$OS_MONTHS, event = LumA_all$OS_STATUS == "DECEASED") ~
```

Appendix B (Continued)

```
        #la_mixture_thresh[i,1] <-  thresh
        la_mixture_thresh[i,1] <-  surv_pvalue(fit.VVV, LumA_all)$pval
        la_mixture_thresh[i,2] <-  sum(z.VVV==2)
        la_mixture_thresh[i,3] <-  lls.VVV
        i=i+1
}
RV_time<-proc.time()-ptm


#simulation
pure<-subset(LumA_combo_new,LumA_combo_new$MD_dist=="0")
admixed<-subset(LumA_combo_new,LumA_combo_new$MD_dist=="1")
pure_gene<-pure[,2:48]
admixed_gene<-admixed[,2:48]
#mu_pure <- colMeans(pure_gene)
sigma_pure<- cov(pure_gene)
is.positive.definite(sigma_pure)

#mu_admixed <- colMeans(admixed_gene)
sigma_admixed<- cov(admixed_gene)
is.positive.definite(sigma_admixed)

redist.fun <- function(x){(x-min(x))/diff(range(x))}
pure_scaled<-apply(pure_gene,2,redist.fun)
pure_mu_scaled<-colMeans(pure_scaled)
admixed_scaled<-apply(admixed_gene,2,redist.fun)
admixed_mu_scaled<-colMeans(admixed_scaled)

hist(pure_mu_scaled)
```

Appendix B (Continued)

```
hist(admixed_mu_scaled)


mu_2<-cbind(pure_mu_scaled,admixed_mu_scaled)



# set up mu for 3 components
pure_3<-subset(LumA_combo_new3,LumA_combo_new3$MD_dist=="0")
neither_3<-subset(LumA_combo_new3,LumA_combo_new3$MD_dist=="1")
admixed_3<-subset(LumA_combo_new3,LumA_combo_new3$MD_dist=="2")
pure_gene_3<-pure_3[,2:48]
neither_gene_3<-neither_3[,2:48]
admixed_gene_3<-admixed_3[,2:48]

#reschale between 0 and 1
redist.fun <- function(x){(x-min(x))/diff(range(x))}
pure_scaled3<-apply(pure_gene_3,2,redist.fun)
pure_mu_scaled3<-colMeans(pure_scaled3)
neither_scaled3<-apply(neither_gene_3,2,redist.fun)
neither_mu_scaled3<-colMeans(neither_scaled3)
admixed_scaled3<-apply(admixed_gene_3,2,redist.fun)
admixed_mu_scaled3<-colMeans(admixed_scaled3)


set.seed(7888)
d <- 47
G <- 2
cova_VVI_2<-mclustVariance(modelName="VVI", d = 47, G = 2)
```

CITED LITERATURE

1. Aitkin, M., Anderson, D., and Hinde, J.: Statistical modelling of data on teaching styles. Journal of the Royal Statistical Society: Series A (General), 144(4):419–448, 1981.

2. Aitkin, M. and Rubin, D. B.: Estimation and hypothesis testing in finite mixture models. Journal of the Royal Statistical Society: Series B (Methodological), 47(1):67–75, 1985.

3. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In Selected papers of hirotugu akaike, pages 199–213. Springer, 1998.

4. Al-Hajj, M., Becker, M. W., Wicha, M., Weissman, I., and Clarke, M. F.: Therapeutic implications of cancer stem cells. Current opinion in genetics & development, 14(1):43–47, 2004.

5. Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J., and Clarke, M. F.: Prospective identification of tumorigenic breast cancer cells. Proceedings of the National Academy of Sciences, 100(7):3983–3988, 2003.

6. Allott, E. H., Geradts, J., Sun, X., Cohen, S. M., Zirpoli, G. R., Khoury, T., Bshara, W., Chen, M., Sherman, M. E., Palmer, J. R., et al.: Intratumoral heterogeneity as a source of discordance in breast cancer biomarker classification. Breast Cancer Research, 18(1):68, 2016.

7. Amir, E., Miller, N., Geddie, W., Freedman, O., Kassam, F., Simmons, C., Oldfield, M., Dranitsaris, G., Tomlinson, G., Laupacis, A., et al.: Prospective study evaluating the impact of tissue confirmation of metastatic disease in patients with breast cancer. Journal of clinical oncology, 30(6):587, 2012.

8. Andor, N., Graham, T. A., Jansen, M., Xia, L. C., Aktipis, C. A., Petritsch, C., Ji, H. P., and Maley, C. C.: Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. Nature medicine, 22(1):105, 2016.

9. Aparicio, S. and Caldas, C.: The implications of clonal genome evolution for cancer medicine. New England journal of medicine, 368(9):842–851, 2013.

10. Asare, A. L., Gao, Z., Carey, V. J., Wang, R., and Seyfert-Margolis, V.: Power enhancement via multivariate outlier testing with gene expression arrays. Bioinformatics, 25(1):48–53, 2008.

11. Banerji, S., Cibulskis, K., Rangel-Escareno, C., Brown, K. K., Carter, S. L., Frederick, A. M., Lawrence, M. S., Sivachenko, A. Y., Sougnez, C., Zou, L., et al.: Sequence analysis of mutations and translocations across breast cancer subtypes. Nature, 486(7403):405, 2012.

12. Banfield, J. D. and Raftery, A. E.: Model-based gaussian and non-gaussian clustering. Biometrics, pages 803–821, 1993.

13. Barbosa-Morais, N. L., Dunning, M. J., Samarajiwa, S. A., Darot, J. F., Ritchie, M. E., Lynch, A. G., and Tavaré, S.: A re-annotation pipeline for illumina beadarrays: improving the interpretation of gene expression data. Nucleic acids research, 38(3):e17–e17, 2009.

14. Barnett, V.: The study of outliers: purpose and model. Journal of the Royal Statistical Society: Series C (Applied Statistics), 27(3):242–250, 1978.

15. Barnett, V. and Lewis, T.: Outliers in statistical data. Wiley, 1974.

16. Barton, V. N., DAmato, N. C., Gordon, M. A., Christenson, J. L., Elias, A., and Richer, J. K.: Androgen receptor biology in triple negative breast cancer: a case for classification as ar+ or quadruple negative disease. Hormones and Cancer, 6(5-6):206–213, 2015.

17. Basu, A., Shioya, H., and Park, C.: Statistical inference: the minimum distance approach. Chapman and Hall/CRC, 2011.

18. Beca, F. and Polyak, K.: Intratumor heterogeneity in breast cancer. In Novel Biomarkers in the Continuum of Breast Cancer, pages 169–189. Springer, 2016.

19. Bedard, P. L., Hansen, A. R., Ratain, M. J., and Siu, L. L.: Tumour heterogeneity in the clinic. Nature, 501(7467):355, 2013.

20. Bengtsson, H., Wirapati, P., and Speed, T. P.: A single-array preprocessing method for estimating full-resolution raw copy numbers from all affymetrix genotyping arrays including genomewidesnp 5 & 6. Bioinformatics, 25(17):2149–2156, 2009.

21. Bensmail, H.: Modeles de regularisation en discrimination et classification Bayesienne. Doctoral dissertation, Paris 6, 1995.

22. Bensmail, H., Celeux, G., Raftery, A. E., and Robert, C. P.: Inference in model-based cluster analysis. statistics and Computing, 7(1):1–10, 1997.

23. Bhattacharya, C.: A simple method of resolution of a distribution into gaussian components. Biometrics, pages 115–135, 1967.

24. Bidard, F.-C., Fehm, T., Ignatiadis, M., Smerage, J. B., Alix-Panabières, C., Janni, W., Messina, C., Paoletti, C., Müller, V., Hayes, D. F., et al.: Clinical application of circulating tumor cells in breast cancer: overview of the current interventional trials. Cancer and Metastasis Reviews, 32(1-2):179–188, 2013.

25. Biernacki, C., Celeux, G., and Govaert, G.: Assessing a mixture model for clustering eith integrated classification likelihood. Rapport de Recherche, (3521), 1998.

26. Biernacki, C., Celeux, G., and Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE transactions on pattern analysis and machine intelligence, 22(7):719–725, 2000.

27. Biernacki, C., Celeux, G., and Govaert, G.: Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. Computational Statistics & Data Analysis, 41(3-4):561–575, 2003.

28. Bloom, H. and Richardson, W.: Histological grading and prognosis in breast cancer: a study of 1409 cases of which 359 have been followed for 15 years. British journal of cancer, 11(3):359, 1957.

29. Böhning, D. and Seidel, W.: Recent developments in mixture models. Computational Statistics & Data Analysis, 41(3-4):349–357, 2003.

30. Bolli, N., Avet-Loiseau, H., Wedge, D. C., Van Loo, P., Alexandrov, L. B., Martincorena, I., Dawson, K. J., Iorio, F., Nik-Zainal, S., Bignell, G. R., et al.: Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. Nature communications, 5:2997, 2014.

31. Bonsing, B. A., Corver, W. E., Fleuren, G. J., Cleton-Jansen, A.-M., Devilee, P., and Cornelisse, C. J.: Allelotype analysis of flow-sorted breast cancer cells demonstrates

genetically related diploid and aneuploid subpopulations in primary tumors and lymph node metastases. Genes, Chromosomes and Cancer, 28(2):173–183, 2000.

32. Boyles, R. A.: On the convergence of the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 45(1):47–50, 1983.

33. Bozdogan, H.: Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. Psychometrika, 52(3):345–370, 1987.

34. Bozdogan, H.: Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. In Proceedings of the first US/Japan conference on the frontiers of statistical modeling: An informational approach, pages 69–113. Springer, 1994.

35. Bradley, P. S. and Fayyad, U. M.: Refining initial points for k-means clustering. In ICML, volume 98, pages 91–99. Citeseer, 1998.

36. Brown, T. M. and Fee, E.: Rudolf carl virchow: medical scientist, social reformer, role model. American Journal of Public Health, 96(12):2104–2105, 2006.

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

38. Burstein, M. D., Tsimelzon, A., Poage, G. M., Covington, K. R., Contreras, A., Fuqua, S. A., Savage, M. I., Osborne, C. K., Hilsenbeck, S. G., Chang, J. C., et al.: Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. Clinical Cancer Research, 21(7):1688–1698, 2015.

39. Cacoullos, T.: Distance, discrimination and error. In Discriminant analysis and applications, pages 61–75. Elsevier, 1973.

40. Cairns, J. M., Dunning, M. J., Ritchie, M. E., Russell, R., and Lynch, A. G.: Bash: a tool for managing beadarray spatial artefacts. Bioinformatics, 24(24):2921–2922, 2008.

41. Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., et al.: Absolute quantification of somatic dna alterations in human cancer. Nature biotechnology, 30(5):413, 2012.

42. Celeux, G. and Govaert, G.: Gaussian parsimonious clustering models. Pattern recognition, 28(5):781–793, 1995.

43. Celià-Terrassa, T., Meca-Cortés, Ó., Mateo, F., De Paz, A. M., Rubio, N., Arnal-Estapé, A., Ell, B. J., Bermudo, R., Díaz, A., Guerra-Rebollo, M., et al.: Epithelial-mesenchymal transition can suppress major attributes of human epithelial tumor-initiating cells. The Journal of clinical investigation, 122(5):1849–1868, 2012.

44. Chaddha, R. and Marcus, L.: An empirical comparison of distance statistics for populations with unequal covariance matrices. Biometrics, pages 683–694, 1968.

45. Chaffer, C. L., Marjanovic, N. D., Lee, T., Bell, G., Kleer, C. G., Reinhardt, F., DAlessio, A. C., Young, R. A., and Weinberg, R. A.: Poised chromatin at the zeb1 promoter enables breast cancer cell plasticity and enhances tumorigenicity. Cell, 154(1):61–74, 2013.

46. Chia, S. K., Bramwell, V. H., Tu, D., Shepherd, L. E., Jiang, S., Vickery, T., Mardis, E., Leung, S., Ung, K., Pritchard, K. I., et al.: A 50-gene intrinsic subtype classifier for prognosis and prediction of benefit from adjuvant tamoxifen. Clinical cancer research, 18(16):4465–4472, 2012.

47. Ciriello, G., Sinha, R., Hoadley, K. A., Jacobsen, A. S., Reva, B., Perou, C. M., Sander, C., and Schultz, N.: The molecular diversity of luminal a breast tumors. Breast cancer research and treatment, 141(3):409–420, 2013.

48. Colleoni, M., Russo, L., and Dellapasqua, S.: Adjuvant therapies for special types of breast cancer. The Breast, 20:S153–S157, 2011.

49. Consortium, I. C. G. et al.: International network of cancer genome projects. Nature, 464(7291):993, 2010.

50. Cortazar, P., Zhang, L., Untch, M., Mehta, K., Costantino, J. P., Wolmark, N., Bonnefoi, H., Cameron, D., Gianni, L., Valagussa, P., et al.: Pathological complete response and long-term clinical benefit in breast cancer: the ctneobc pooled analysis. The Lancet, 384(9938):164–172, 2014.

51. Cramir, H.: Mathematical methods of statistics. Princeton U. Press, Princeton, page 500, 1946.

52. Cronin, M., Sangli, C., Liu, M.-L., Pho, M., Dutta, D., Nguyen, A., Jeong, J., Wu, J., Langone, K. C., and Watson, D.: Analytical validation of the oncotype dx genomic diagnostic test for recurrence prognosis and therapeutic response prediction in node-negative, estrogen receptor–positive breast cancer. Clinical chemistry, 53(6):1084–1091, 2007.

53. Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., et al.: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature, 486(7403):346, 2012.

54. Daniels, H. E.: Saddlepoint approximations in statistics. The Annals of Mathematical Statistics, pages 631–650, 1954.

55. Dasgupta, A. and Raftery, A. E.: Detecting features in spatial point processes with clutter via model-based clustering. Journal of the American statistical Association, 93(441):294–302, 1998.

56. Day, N. E.: Estimating the components of a mixture of normal distributions. Biometrika, 56(3):463–474, 1969.

57. De Boer, P.-T., Kroese, D. P., Mannor, S., and Rubinstein, R. Y.: A tutorial on the cross-entropy method. Annals of operations research, 134(1):19–67, 2005.

58. Dean, C.: Modified pseudo-likelihood estimator of the overdispersion parameter in poisson mixture models. Journal of Applied Statistics, 21(6):523–532, 1994.

59. Dean-Colomb, W. and Esteva, F. J.: Her2-positive breast cancer: herceptin and beyond. European Journal of Cancer, 44(18):2806–2812, 2008.

60. Delmar, P., Robin, S., Tronik-Le Roux, D., and Daudin, J. J.: Mixture model on the variance for the differential analysis of gene expression data. Journal of the Royal Statistical Society: Series C (Applied Statistics), 54(1):31–50, 2005.

61. Dempster, A. P., Laird, N. M., and Rubin, D. B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22, 1977.

62. Ding, L., Ellis, M. J., Li, S., Larson, D. E., Chen, K., Wallis, J. W., Harris, C. C., McLellan, M. D., Fulton, R. S., Fulton, L. L., et al.: Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature, 464(7291):999, 2010.

63. Donoho, D. L., Liu, R. C., et al.: Pathologies of some minimum distance estimators. The Annals of Statistics, 16(2):587–608, 1988.

64. Dubsky, P., Brase, J., Jakesz, R., Rudas, M., Singer, C., Greil, R., Dietze, O., Luisser, I., Klug, E., Sedivy, R., et al.: The endopredict score provides prognostic information on late distant metastases in er+/her2- breast cancer patients. British journal of cancer, 109(12):2959, 2013.

65. Dumenci, L. and Windle, M.: Cluster analysis as a method of recovering types of intraindividual growth trajectories: A monte carlo study. Multivariate Behavioral Research, 36(4):501–522, 2001.

66. Dunning, M. J., Curtis, C., Barbosa-Morais, N. L., Caldas, C., Tavaré, S., and Lynch, A. G.: The importance of platform annotation in interpreting microarray data. The lancet oncology, 11(8):717, 2010.

67. Dunning, M. J., Smith, M. L., Ritchie, M. E., and Tavaré, S.: beadarray: R classes and methods for illumina bead-based data. Bioinformatics, 23(16):2183–2184, 2007.

68. Easwaran, H., Tsai, H.-C., and Baylin, S. B.: Cancer epigenetics: tumor heterogeneity, plasticity of stem-like states, and drug resistance. Molecular cell, 54(5):716–727, 2014.

69. Edelbrock, C.: Mixture model tests of hierarchical clustering algorithms: The problem of classifying everybody. Multivariate Behavioral Research, 14(3):367–384, 1979.

70. Edge, S. B. and Compton, C. C.: The american joint committee on cancer: the 7th edition of the ajcc cancer staging manual and the future of tnm. Annals of surgical oncology, 17(6):1471–1474, 2010.

71. Ellis, M. J., Ding, L., Shen, D., Luo, J., Suman, V. J., Wallis, J. W., Van Tine, B. A., Hoog, J., Goiffon, R. J., Goldstein, T. C., et al.: Whole-genome analysis informs breast cancer response to aromatase inhibition. Nature, 486(7403):353, 2012.

72. Elston, C. W. and Ellis, I. O.: Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with

long-term follow-up. cw elston & io ellis. histopathology 1991; 19; 403–410: Author commentary. Histopathology, 41(3a):151–151, 2002.

73. Elston, C., Ellis, I., and Pinder, S.: Prognostic factors in invasive carcinoma of the breast. Clinical Oncology, 10(1):14–17, 1998.

74. Everitt, B. S., Landau, S., Leese, M., and Stahl, D.: Hierarchical clustering. Cluster analysis, 5, 2011.

75. Farmer, P., Bonnefoi, H., Becette, V., Tubiana-Hulin, M., Fumoleau, P., Larsimont, D., MacGrogan, G., Bergh, J., Cameron, D., Goldstein, D., et al.: Identification of molecular apocrine breast tumours by microarray analysis. Breast Cancer Research, 7(2):P2–11, 2005.

76. Feller, W.: On a general class of contagious distributions. In Selected Papers I, pages 643–654. Springer, 2015.

77. Fiegl, M., Tueni, C., Schenk, T., Jakesz, R., Gnant, M., Reiner, A., Rudas, M., Pirc-Danoewinata, H., Marosi, C., Huber, H., et al.: Interphase cytogenetics reveals a high incidence of aneuploidy and intra-tumour heterogeneity in breast cancer. British journal of cancer, 72(1):51, 1995.

78. Findlay, J. M., Castro-Giner, F., Makino, S., Rayner, E., Kartsonaki, C., Cross, W., Kovac, M., Ulahannan, D., Palles, C., Gillies, R. S., et al.: Differential clonal evolution in oesophageal cancers in response to neo-adjuvant chemotherapy. Nature communications, 7:11111, 2016.

79. Fitzgerald, P. J.: Homogeneity and heterogeneity in pancreas cancer: presence of predominant and minor morphological types and implications. International Journal of Gastrointestinal Cancer, 1(2):91–94, 1986.

80. Fitzmaurice, C., Allen, C., Barber, R. M., Barregard, L., Bhutta, Z. A., Brenner, H., Dicker, D. J., Chimed-Orchir, O., Dandona, R., Dandona, L., et al.: Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study. JAMA oncology, 3(4):524–548, 2017.

81. Fraley, C.: Algorithms for model-based gaussian hierarchical clustering. SIAM Journal on Scientific Computing, 20(1):270–281, 1998.

82. Fraley, C. and Raftery, A. E.: How many clusters? which clustering method? answers via model-based cluster analysis. The computer journal, 41(8):578–588, 1998.

83. Fraley, C. and Raftery, A. E.: Model-based clustering, discriminant analysis, and density estimation. Journal of the American statistical Association, 97(458):611–631, 2002.

84. Fraley, C., Raftery, A. E., Murphy, T. B., and Scrucca, L.: mclust version 4 for r: normal mixture modeling for model-based clustering, classification, and density estimation. Technical report, Technical report, 2012.

85. Geary, D.: Mixture models: Inference and applications to clustering. Journal of the Royal Statistical Society Series A, 152(1):126–127, 1989.

86. Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al.: Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. New England journal of medicine, 366(10):883–892, 2012.

87. Gerlinger, M. and Swanton, C.: How darwinian models inform therapeutic failure initiated by clonal heterogeneity in cancer medicine. British journal of cancer, 103(8):1139, 2010.

88. Geyer, F. C., Weigelt, B., Natrajan, R., Lambros, M. B., de Biase, D., Vatcheva, R., Savage, K., Mackay, A., Ashworth, A., and Reis-Filho, J. S.: Molecular analysis reveals a genetic basis for the phenotypic diversity of metaplastic breast carcinomas. The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland, 220(5):562–573, 2010.

89. Greaves, M. and Maley, C. C.: Clonal evolution in cancer. Nature, 481(7381):306, 2012.

90. Green, P. J.: On use of the em algorithm for penalized likelihood estimation. Journal of the Royal Statistical Society: Series B (Methodological), 52(3):443–452, 1990.

91. Group, E. B. C. T. C. et al.: Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. The Lancet, 365(9472):1687–1717, 2005.

92. Gucalp, A., Tolaney, S., Isakoff, S. J., Ingle, J. N., Liu, M. C., Carey, L. A., Blackwell, K., Rugo, H., Nabell, L., Forero, A., et al.: Phase ii trial of bicalutamide in

patients with androgen receptor–positive, estrogen receptor–negative metastatic breast cancer. Clinical cancer research, 19(19):5505–5512, 2013.

93. Gurzu, S., Turdean, S., Kovecsi, A., Contac, A. O., and Jung, I.: Epithelial-mesenchymal, mesenchymal-epithelial, and endothelial-mesenchymal transitions in malignant tumors: An update. World Journal of Clinical Cases: WJCC, 3(5):393, 2015.

94. Hammond, M. E. H., Hayes, D. F., Dowsett, M., Allred, D. C., Hagerty, K. L., Badve, S., Fitzgibbons, P. L., Francis, G., Goldstein, N. S., Hayes, M., et al.: American society of clinical oncology/college of american pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer (unabridged version). Archives of pathology & laboratory medicine, 134(7):e48–e72, 2010.

95. Hammond, M., Hayes, D., Dowsett, M., of Clinical Oncology, A. S., of American Pathologists, C., et al.: Guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer (unabridged version). Arch Pathol Lab Med, 134(7):e48–e72, 2010.

96. Hansen, P. and Jaumard, B.: Cluster analysis and mathematical programming. Mathematical programming, 79(1-3):191–215, 1997.

97. Harris, L. N., Ismaila, N., McShane, L. M., Andre, F., Collyar, D. E., Gonzalez-Angulo, A. M., Hammond, E. H., Kuderer, N. M., Liu, M. C., Mennel, R. G., et al.: Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: American society of clinical oncology clinical practice guideline. Journal of Clinical Oncology, 34(10):1134, 2016.

98. Hartigan, J. A.: Clustering Algorithms. New York, NY, USA, John Wiley & Sons, Inc., 99th edition, 1975.

99. Hipp, J. R. and Bauer, D. J.: Local solutions in the estimation of growth mixture models. Psychological methods, 11(1):36, 2006.

100. Hope, A. C.: A simplified monte carlo significance test procedure. Journal of the Royal Statistical Society: Series B (Methodological), 30(3):582–598, 1968.

101. Horng, S.: Examples of sublinear convergence of the em algorithm. In Proceeding of the Statistical Computing Section, American Statistical Association, pages 266–271, 1987.

102. Hosmer Jr, D. W.: A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. Biometrics, pages 761–770, 1973.

103. Hubert, L. and Arabie, P.: Comparing partitions. Journal of classification, 2(1):193–218, 1985.

104. Jain, A. K. and Dubes, R. C.: Algorithms for clustering data. Englewood Cliffs: Prentice Hall, 1988, 1988.

105. Jank, W.: The em algorithm, its randomized implementation and global optimization: Some challenges and opportunities for operations research. In Perspectives in operations research, pages 367–392. Springer, 2006.

106. Kaklamani, V.: A genetic signature can predict prognosis and response to therapy in breast cancer: Onco type dx. Expert review of molecular diagnostics, 6(6):803–809, 2006.

107. Kalluri, R. and Weinberg, R. A.: The basics of epithelial-mesenchymal transition. The Journal of clinical investigation, 119(6):1420–1428, 2009.

108. Kass, R. E. and Wasserman, L.: A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. Journal of the american statistical association, 90(431):928–934, 1995.

109. Koren, S. and Bentires-Alj, M.: Breast tumor heterogeneity: source of fitness, hurdle for therapy. Molecular cell, 60(4):537–546, 2015.

110. Kurita, T., Otsu, N., and Abdelmalek, N.: Maximum likelihood thresholding based on population mixture models. Pattern recognition, 25(10):1231–1240, 1992.

111. Kuukasjärvi, T., Karhu, R., Tanner, M., Kähkönen, M., Schäffer, A., Nupponen, N., Pennanen, S., Kallioniemi, A., Kallioniemi, O.-P., and Isola, J.: Genetic heterogeneity and clonal evolution underlying development of asynchronous metastasis in human breast cancer. Cancer research, 57(8):1597–1604, 1997.

112. Laird, N.: Nonparametric maximum likelihood estimation of a mixing distribution. Journal of the American Statistical Association, 73(364):805–811, 1978.

113. Lange, K.: A gradient algorithm locally equivalent to the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 57(2):425–437, 1995.

114. Laurikkala, J., Juhola, M., Kentala, E., Lavrac, N., Miksch, S., and Kavsek, B.: Informal identification of outliers in medical data. In Fifth international workshop on intelligent data analysis in medicine and pharmacology, volume 1, pages 20–24, 2000.

115. Lee, H. J., Kim, J. Y., Park, S. Y., Park, I. A., Song, I. H., Yu, J. H., Ahn, J.-H., and Gong, G.: Clinicopathologic significance of the intratumoral heterogeneity of her2 gene amplification in her2-positive breast cancer patients treated with adjuvant trastuzumab. American journal of clinical pathology, 144(4):570–578, 2015.

116. Leroux, B. G.: Consistent estimation of a mixing distribution. The Annals of Statistics, pages 1350–1360, 1992.

117. Li, L., Sedransk, N., et al.: Mixtures of distributions: A topological approach. The annals of Statistics, 16(4):1623–1634, 1988.

118. Liang, Z., Jaszczak, R. J., and Coleman, R. E.: Parameter estimation of finite mixtures using the em algorithm and information criteria with application to medical image processing. IEEE Transactions on Nuclear Science, 39(4):1126–1133, 1992.

119. Lindsay, B. G.: Mixture models: theory, geometry and applications. In NSF-CBMS regional conference series in probability and statistics, pages i–163. JSTOR, 1995.

120. Lo, Y., Mendell, N. R., and Rubin, D. B.: Testing the number of components in a normal mixture. Biometrika, 88(3):767–778, 2001.

121. Maaten, L. v. d. and Hinton, G.: Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579–2605, 2008.

122. Mahalanobis, P.: On tests and measures of group divergence. J. Asiat. Soc. Bengal, 26:541–588, 1930.

123. Mani, S. A., Guo, W., Liao, M.-J., Eaton, E. N., Ayyanan, A., Zhou, A. Y., Brooks, M., Reinhard, F., Zhang, C. C., Shipitsin, M., et al.: The epithelial-mesenchymal transition generates cells with properties of stem cells. Cell, 133(4):704–715, 2008.

124. Mantel, N.: Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep, 50:163–170, 1966.

125. Markou, M. and Singh, S.: Novelty detection: a reviewpart 1: statistical approaches. Signal processing, 83(12):2481–2497, 2003.

126. Martelotto, L. G., Ng, C. K., Piscuoglio, S., Weigelt, B., and Reis-Filho, J. S.: Breast cancer intra-tumor heterogeneity. Breast Cancer Research, 16(3):210, 2014.

127. McLachlan, G. and Krishnan, T.: The EM algorithm and extensions, volume 382. John Wiley & Sons, 2007.

128. McLachlan, G. and Peel, D.: Finite mixture models. John Wiley & Sons, 2004.

129. McLachlan, G. J.: On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. Journal of the Royal Statistical Society: Series C (Applied Statistics), 36(3):318–324, 1987.

130. McLachlan, G. J. and Basford, K. E.: Mixture models: Inference and applications to clustering, volume 84. M. Dekker New York, 1988.

131. Meacham, C. E. and Morrison, S. J.: Tumour heterogeneity and cancer cell plasticity. Nature, 501(7467):328, 2013.

132. Medema, J. P.: Cancer stem cells: the challenges ahead. Nature cell biology, 15(4):338, 2013.

133. Meng, X.-L. and Rubin, D. B.: Using em to obtain asymptotic variance-covariance matrices: The sem algorithm. Journal of the American Statistical Association, 86(416):899–909, 1991.

134. Mengersen, K. L., Robert, C., and Titterington, M.: Mixtures: estimation and applications, volume 896. John Wiley & Sons, 2011.

135. Merlo, L. M., Shah, N. A., Li, X., Blount, P. L., Vaughan, T. L., Reid, B. J., and Maley, C. C.: A comprehensive survey of clonal diversity measures in barrett's esophagus as biomarkers of progression to esophageal adenocarcinoma. Cancer prevention research, 3(11):1388–1397, 2010.

136. Metzger-Filho, O., Tutt, A., De Azambuja, E., Saini, K. S., Viale, G., Loi, S., Bradbury, I., Bliss, J. M., Azim Jr, H. A., Ellis, P., et al.: Dissecting the heterogeneity of triple-negative breast cancer. Journal of clinical oncology, 30(15):1879–1887, 2012.

137. Meyerson, M., Gabriel, S., and Getz, G.: Advances in understanding cancer genomes through second-generation sequencing. Nature Reviews Genetics, 11(10):685, 2010.

138. Michor, F. and Polyak, K.: The origins and implications of intratumor heterogeneity. Cancer prevention research, 3(11):1361–1364, 2010.

139. Miller, C. A., Gindin, Y., Lu, C., Griffith, O. L., Griffith, M., Shen, D., Hoog, J., Li, T., Larson, D. E., Watson, M., et al.: Aromatase inhibition remodels the clonal architecture of estrogen-receptor-positive breast cancers. Nature communications, 7:12498, 2016.

140. Miller, W. R. and Larionov, A.: Changes in expression of oestrogen regulated and proliferation genes with neoadjuvant treatment highlight heterogeneity of clinical resistance to the aromatase inhibitor, letrozole. Breast cancer research, 12(4):R52, 2010.

141. Morel, A.-P., Lièvre, M., Thomas, C., Hinkal, G., Ansieau, S., and Puisieux, A.: Generation of breast cancer stem cells through epithelial-mesenchymal transition. PloS one, 3(8):e2888, 2008.

142. Muthén, B. and Shedden, K.: Finite mixture modeling with mixture outcomes using the em algorithm. Biometrics, 55(2):463–469, 1999.

143. Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al.: Tumour evolution inferred by single-cell sequencing. Nature, 472(7341):90, 2011.

144. Navin, N., Krasnitz, A., Rodgers, L., Cook, K., Meth, J., Kendall, J., Riggs, M., Eberling, Y., Troge, J., Grubor, V., et al.: Inferring tumor progression from genomic heterogeneity. Genome research, 20(1):68–80, 2010.

145. Network, C. G. A. et al.: Comprehensive molecular portraits of human breast tumours. Nature, 490(7418):61, 2012.

146. Network, C. G. A. R. et al.: Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature, 455(7216):1061, 2008.

147. Newburger, D. E., Kashef-Haghighi, D., Weng, Z., Salari, R., Sweeney, R. T., Brunner, A. L., Zhu, S. X., Guo, X., Varma, S., Troxell, M. L., et al.: Genome evolution during progression to breast cancer. Genome research, 23(7):1097–1108, 2013.

148. Newcomb, S.: A generalized theory of the combination of observations so as to obtain the best result. American journal of Mathematics, pages 343–366, 1886.

149. Ng, C. K., Martelotto, L. G., Gauthier, A., Wen, H.-C., Piscuoglio, S., Lim, R. S., Cowell, C. F., Wilkerson, P. M., Wai, P., Rodrigues, D. N., et al.: Intra-tumor genetic heterogeneity and alternative driver genetic alterations in breast cancers with heterogeneous her2 gene amplification. Genome biology, 16(1):107, 2015.

150. Nielsen, T., Wallden, B., Schaper, C., Ferree, S., Liu, S., Gao, D., Barry, G., Dowidar, N., Maysuria, M., and Storhoff, J.: Analytical validation of the pam50-based prosigna breast cancer prognostic gene signature assay and ncounter analysis system using formalin-fixed paraffin-embedded breast tumor specimens. BMC cancer, 14(1):177, 2014.

151. Nielsen, T. O., Hsu, F. D., Jensen, K., Cheang, M., Karaca, G., Hu, Z., Hernandez-Boussard, T., Livasy, C., Cowan, D., Dressler, L., et al.: Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. Clinical cancer research, 10(16):5367–5374, 2004.

152. Nielsen, T. O., Parker, J. S., Leung, S., Voduc, D., Ebbert, M., Vickery, T., Davies, S. R., Snider, J., Stijleman, I. J., Reed, J., et al.: A comparison of pam50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor–positive breast cancer. Clinical cancer research, 16(21):5222–5232, 2010.

153. Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L. A., et al.: Mutational processes molding the genomes of 21 breast cancers. Cell, 149(5):979–993, 2012.

154. Nowell, P. C.: The clonal evolution of tumor cell populations. Science, 194(4260):23–28, 1976.

155. Oakman, C., Santarpia, L., and Di Leo, A.: Breast cancer assessment tools and optimizing adjuvant therapy. Nature reviews Clinical oncology, 7(12):725, 2010.

156. Page, D. L.: Special types of invasive breast cancer, with clinical implications. The American journal of surgical pathology, 27(6):832–835, 2003.

157. Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D., Park, T., et al.: A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. New England Journal of Medicine, 351(27):2817–2826, 2004.

158. Pandis, N., Teixeira, M. R., Adeyinka, A., Rizou, H., Bardi, G., Mertens, F., Andersen, J. A., Bondeson, L., Sfikas, K., Qvist, H., et al.: Cytogenetic comparison of primary tumors and lymph node metastases in breast cancer patients. Genes, Chromosomes and Cancer, 22(2):122–129, 1998.

159. Park, S., Koo, J., Kim, M., Park, H., Lee, J., Lee, J., Kim, S., Park, B.-W., and Lee, K.: Androgen receptor expression is significantly associated with better outcomes in estrogen receptor-positive breast cancers. Annals of Oncology, 22(8):1755–1762, 2011.

160. Park, S. Y., Lee, H. E., Li, H., Shipitsin, M., Gelman, R., and Polyak, K.: Heterogeneity for stem cell–related markers according to tumor subtype and histologic stage in breast cancer. Clinical Cancer Research, 16(3):876–887, 2010.

161. Park, Y., Lim, S., Nam, J.-W., and Kim, S.: Measuring intratumor heterogeneity by network entropy using rna-seq data. Scientific reports, 6:37767, 2016.

162. Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al.: Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology, 27(8):1160, 2009.

163. Parker, J. S. and Perou, C. M.: Tumor heterogeneity: focus on the leaves, the trees, or the forest? Cancer Cell, 28(2):149–150, 2015.

164. Patani, N., Barbashina, V., Lambros, M. B., Gauthier, A., Mansour, M., Mackay, A., and Reis-Filho, J. S.: Direct evidence for concurrent morphological and genetic heterogeneity in an invasive ductal carcinoma of triple-negative phenotype. Journal of clinical pathology, 64(9):822–828, 2011.

165. Pearson, K.: Contributions to the mathematical theory of evolution. Philosophical Transactions of the Royal Society of London. A, 185:71–110, 1894.

166. Perez, E. A.: Breast cancer management: opportunities and barriers to an individualized approach. The oncologist, 16(Supplement 1):20–22, 2011.

167. Perou, C. M., Parker, J. S., Prat, A., Ellis, M. J., and Bernard, P. S.: Clinical implementation of the intrinsic subtypes of breast cancer. The lancet oncology, 11(8):718–719, 2010.

168. Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., et al.: Molecular portraits of human breast tumours. nature, 406(6797):747, 2000.

169. Peto, R. and Peto, J.: Asymptotically efficient rank invariant test procedures. Journal of the Royal Statistical Society: Series A (General), 135(2):185–198, 1972.

170. Pinto, C. A., Widodo, E., Waltham, M., and Thompson, E. W.: Breast cancer stem cells and epithelial mesenchymal plasticity–implications for chemoresistance. Cancer letters, 341(1):56–62, 2013.

171. Plaks, V., Kong, N., and Werb, Z.: The cancer stem cell niche: how essential is the niche in regulating stemness of tumor cells? Cell stem cell, 16(3):225–238, 2015.

172. Polyak, K.: Heterogeneity in breast cancer. The Journal of clinical investigation, 121(10):3786–3788, 2011.

173. Prat, A., Parker, J. S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., He, X., and Perou, C. M.: Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast cancer research, 12(5):R68, 2010.

174. Prat, A. and Perou, C. M.: Deconstructing the molecular portraits of breast cancer. Molecular oncology, 5(1):5–23, 2011.

175. Qi, J.-p., Yang, Y.-l., Zhu, H., Wang, J., Jia, Y., Liu, N., Song, Y.-j., Zan, L.-k., Zhang, X., Zhou, M., et al.: Expression of the androgen receptor and its correlation with molecular subtypes in 980 chinese breast cancer patients. Breast cancer: basic and clinical research, 6:BCBCR–S8323, 2012.

176. Raftery, A. E. and Dean, N.: Variable selection for model-based clustering. Journal of the American Statistical Association, 101(473):168–178, 2006.

177. Rakha, E. A., Lee, A. H., Evans, A. J., Menon, S., Assad, N. Y., Hodi, Z., Macmillan, D., Blamey, R. W., and Ellis, I. O.: Tubular carcinoma of the breast: further evidence to support its excellent prognosis. Journal of Clinical Oncology, 28(1):99–104, 2009.

178. Rakha, E. A., Reis-Filho, J. S., Baehner, F., Dabbs, D. J., Decker, T., Eusebi, V., Fox, S. B., Ichihara, S., Jacquemier, J., Lakhani, S. R., et al.: Breast cancer prognostic classification in the molecular era: the role of histological grade. Breast Cancer Research, 12(4):207, 2010.

179. Ranjan Mukhopadhyay, A.: Multivariate attribute control chart using mahalanobis d 2 statistic. Journal of Applied Statistics, 35(4):421–429, 2008.

180. Roberts, R. J., Carneiro, M. O., and Schatz, M. C.: The advantages of smrt sequencing. Genome biology, 14(6):405, 2013.

181. Robinson, M. D., McCarthy, D. J., and Smyth, G. K.: edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1):139–140, 2010.

182. Roeder, K. and Wasserman, L.: Practical bayesian density estimation using mixtures of normals. Journal of the American Statistical Association, 92(439):894–902, 1997.

183. Rosen, P. P., Groshen, S., Kinne, D., and Norton, L.: Factors influencing prognosis in node-negative breast carcinoma: analysis of 767 t1n0m0/t2n0m0 patients with long-term follow-up. Journal of clinical oncology, 11(11):2090–2100, 1993.

184. Rousseeuw, P. J. and Van Zomeren, B. C.: Unmasking multivariate outliers and leverage points. Journal of the American Statistical association, 85(411):633–639, 1990.

185. Sankaranarayanan, R., Swaminathan, R., Brenner, H., Chen, K., Chia, K. S., Chen, J. G., Law, S. C., Ahn, Y.-O., Xiang, Y. B., Yeole, B. B., et al.: Cancer survival in africa, asia, and central america: a population-based study. The lancet oncology, 11(2):165–173, 2010.

186. Schlattmann, P.: Medical applications of finite mixture models.. Springer, 2009.

187. Schramm, A., Friedl, T. W., Schochter, F., Scholz, C., De Gregorio, N., Huober, J., Rack, B., Trapp, E., Alunni-Fabbroni, M., Müller, V., et al.: Therapeutic intervention based on circulating tumor cell phenotype in metastatic breast cancer: concept of the detect study program. Archives of gynecology and obstetrics, 293(2):271–281, 2016.

188. Schwarz, G. et al.: Estimating the dimension of a model. The annals of statistics, 6(2):461–464, 1978.

189. Sclove, S. L.: Application of the conditional population-mixture model to image segmentation. IEEE transactions on pattern analysis and machine intelligence, (4):428–433, 1983.

190. Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E.: mclust 5: clustering, classification and density estimation using gaussian finite mixture models. The R journal, 8(1):289, 2016.

191. Sgroi, D. C., Sestak, I., Cuzick, J., Zhang, Y., Schnabel, C. A., Schroeder, B., Erlander, M. G., Dunbier, A., Sidhu, K., Lopez-Knowles, E., et al.: Prediction of late distant recurrence in patients with oestrogen-receptor-positive breast cancer: a prospective comparison of the breast-cancer index (bci) assay, 21-gene recurrence score, and ihc4 in the transatac study population. The lancet oncology, 14(11):1067–1076, 2013.

192. Shah, M. and Allegrucci, C.: Keeping an open mind: highlights and controversies of the breast cancer stem cell theory. Breast cancer: targets and therapy, 4:155, 2012.

193. Shah, S. P., Morin, R. D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K., Guliany, R., Senz, J., et al.: Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature, 461(7265):809, 2009.

194. Shah, S. P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K., Haffari, G., et al.: The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature, 486(7403):395, 2012.

195. Shah, S. P., Xuan, X., DeLeeuw, R. J., Khojasteh, M., Lam, W. L., Ng, R., and Murphy, K. P.: Integrating copy number polymorphisms into array cgh analysis using a robust hmm. Bioinformatics, 22(14):e431–e439, 2006.

196. Siegel, R., Naishadham, D., and Jemal, A.: Cancer statistics, 2013. CA: a cancer journal for clinicians, 63(1):11–30, 2013.

197. Simar, L. et al.: Maximum likelihood estimation of a compound poisson process. The Annals of Statistics, 4(6):1200–1209, 1976.

198. Sinn, H.-P. and Kreipe, H.: A brief overview of the who classification of breast tumors. Breast care, 8(2):149–154, 2013.

199. Sørlie, T., Borgan, E., Myhre, S., Vollan, H. K., Russnes, H., Zhao, X., Nilsen, G., Lingjærde, O. C., Børresen-Dale, A.-L., and Rødland, E.: The importance of gene-centring microarray data. The lancet oncology, 11(8):719–720, 2010.

200. Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., et al.: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences, 98(19):10869–10874, 2001.

201. Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., et al.: Repeated observation of breast tumor subtypes in independent gene expression data sets. Proceedings of the National Academy of Sciences of the United States of America, 100(14):8418–8423, 2003.

202. Steinley, D.: Local optima in k-means clustering: what you don't know may hurt you. Psychological methods, 8(3):294, 2003.

203. Steinley, D.: Properties of the hubert-arable adjusted rand index. Psychological methods, 9(3):386, 2004.

204. Steinley, D. and Brusco, M. J.: Initializing k-means batch clustering: A critical evaluation of several techniques. Journal of Classification, 24(1):99–121, 2007.

205. Stephens, P. J., Tarpey, P. S., Davies, H., Van Loo, P., Greenman, C., Wedge, D. C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G. R., et al.: The landscape of cancer genes and mutational processes in breast cancer. Nature, 486(7403):400, 2012.

206. Swanton, C. and Caldas, C.: Molecular classification of solid tumours: towards pathway-driven therapeutics. British journal of cancer, 100(10):1517, 2009.

207. Tavassoéli, F., Devilee, P., Organization, W. H., et al.: tumours of the breast and female genital organs (who/iarc classification of tumours), 2003.

208. Teicher, H.: Identifiability of finite mixtures. The annals of Mathematical statistics, pages 1265–1269, 1963.

209. Teicher, H. et al.: On the mixture of distributions. The Annals of Mathematical Statistics, 31(1):55–73, 1960.

210. Tierney, L. and Kadane, J. B.: Accurate approximations for posterior moments and marginal densities. Journal of the american statistical association, 81(393):82–86, 1986.

211. Titterington, D. M., Smith, A. F., and Makov, U. E.: Statistical analysis of finite mixture distributions. Wiley,, 1985.

212. Titterington, D.: Some recent research in the analysis of mixture distributions. Statistics, 21(4):619–641, 1990.

213. Tonidandel, S. and Overall, J. E.: Determining the number of clusters by sampling with replacement. Psychological Methods, 9(2):238, 2004.

214. Torre, L. A., Islami, F., Siegel, R. L., Ward, E. M., and Jemal, A.: Global cancer in women: burden and trends, 2017.

215. Torres, L., Ribeiro, F. R., Pandis, N., Andersen, J. A., Heim, S., and Teixeira, M. R.: Intratumor genomic heterogeneity in breast cancer with clonal divergence between primary carcinomas and lymph node metastases. Breast cancer research and treatment, 102(2):143–155, 2007.

216. Tou, J. T. and Gonzalez, R. C.: Pattern recognition principles. 1974.

217. Turner, N. C. and Reis-Filho, J. S.: Genetic heterogeneity and cancer drug resistance. The lancet oncology, 13(4):e178–e185, 2012.

218. Ueda, N. and Nakano, R.: Deterministic annealing em algorithm. Neural networks, 11(2):271–282, 1998.

219. Van Der Kloot, W. A., Spaans, A. M., and Heiser, W. J.: Instability of hierarchical cluster analysis due to input order of the data: the permucluster solution. Psychological methods, 10(4):468, 2005.

220. Van der Poel, H., Oosterhof, G., Schaafsma, H., Debruyne, F., and Schalken, J.: Intratumoral nuclear morphologic heterogeneity in prostate cancer. Urology, 49(4):652–657, 1997.

221. Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., Van Der Kooy, K., Marton, M. J., Witteveen, A. T., et al.: Gene expression profiling predicts clinical outcome of breast cancer. nature, 415(6871):530, 2002.

222. Voet, T., Kumar, P., Van Loo, P., Cooke, S. L., Marshall, J., Lin, M.-L., Zamani Esteki, M., Van der Aa, N., Mateiu, L., McBride, D. J., et al.: Single-cell paired-end genome sequencing reveals structural variation per cell cycle. Nucleic acids research, 41(12):6119–6138, 2013.

223. Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W.: Cancer genome landscapes. science, 339(6127):1546–1558, 2013.

224. von Eye, A. and Bergman, L. R.: Research strategies in developmental psychopathology: Dimensional identity and the person-oriented approach. Development and psychopathology, 15(3):553–580, 2003.

225. Walter, M. J., Shen, D., Ding, L., Shao, J., Koboldt, D. C., Chen, K., Larson, D. E., McLellan, M. D., Dooling, D., Abbott, R., et al.: Clonal architecture of secondary acute myeloid leukemia. New England Journal of Medicine, 366(12):1090–1098, 2012.

226. Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., et al.: Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. The Lancet, 365(9460):671–679, 2005.

227. Wedel, M. and DeSarbo, W. S.: A mixture likelihood approach for generalized linear models. Journal of Classification, 12(1):21–55, 1995.

228. Weigelt, B., Horlings, H., Kreike, B., Hayes, M., Hauptmann, M., Wessels, L., De Jong, D., Van de Vijver, M., Veer, L. V., and Peterse, J.: Refinement of breast cancer classification by molecular characterization of histological special

types. The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland, 216(2):141–150, 2008.

229. Weigelt, B., Mackay, A., A'hern, R., Natrajan, R., Tan, D. S., Dowsett, M., Ashworth, A., and Reis-Filho, J. S.: Breast cancer molecular profiling with single sample predictors: a retrospective analysis. The lancet oncology, 11(4):339–349, 2010.

230. William, C. P.: Minimum distance estimation: a bibliography. Communications in Statistics-Theory and Methods, 10(12):1205–1224, 1981.

231. Windham, M. P. and Cutler, A.: Information ratios for validating mixture analyses. Journal of the American Statistical Association, 87(420):1188–1192, 1992.

232. Wirapati, P., Sotiriou, C., Kunkel, S., Farmer, P., Pradervand, S., Haibe-Kains, B., Desmedt, C., Ignatiadis, M., Sengstag, T., Schütz, F., et al.: Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Research, 10(4):R65, 2008.

233. Wolff, A. C., Hammond, M. E. H., Hicks, D. G., Dowsett, M., McShane, L. M., Allison, K. H., Allred, D. C., Bartlett, J. M., Bilous, M., Fitzgibbons, P., et al.: Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American society of clinical oncology/college of american pathologists clinical practice guideline update. Archives of Pathology and Laboratory Medicine, 138(2):241–256, 2013.

234. Wu, C. J. et al.: On the convergence properties of the em algorithm. The Annals of statistics, 11(1):95–103, 1983.

235. xian Wang, H., Luo, B., bing Zhang, Q., and Wei, S.: Estimation for the number of components in a mixture model using stepwise split-and-merge em algorithm. Pattern Recognition Letters, 25(16):1799–1809, 2004.

236. Yakowitz, S. J. and Spragins, J. D.: On the identifiability of finite mixtures. The Annals of Mathematical Statistics, pages 209–214, 1968.

237. Yap, T. A., Gerlinger, M., Futreal, P. A., Pusztai, L., and Swanton, C.: Intratumor heterogeneity: seeing the wood for the trees. Science translational medicine, 4(127):127ps10–127ps10, 2012.

238. Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L.: Model-based clustering and data transformations for gene expression data. Bioinformatics, 17(10):977–987, 2001.

VITA

NAME:	Dan Zhao
EDUCATION:	Bachelor of Science, Biotechnology and Economics, Sun Yat-sen University, 2011 Master of Science, Biostatistics, Yale University, 2014 Doctor of Philosophy, Biostatistics, University of Illinois at Chicago, 2019
PUBLICATIONS:	Kumar, N., Zhao, D., Bhaumik, D., Sethi, A. and Gann, P.H.: Quantification of intrinsic subtype ambiguity in Luminal A breast cancer and its relationship to clinical outcomes. BMC cancer. 19(1), p.215, 2019.