STATISTICAL METHODS IN CANCER GENOMICS

by

Ronglai Shen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2007

Doctoral Committee:

Associate Professor Debashis Ghosh, Co-Chair
Professor Jeremy M. G. Taylor, Co-Chair
Professor Arul M. Chinnaiyan
Professor Roderick J. A. Little

TO MY PARENTS AND MY HUSBAND CONG

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Drs. Debashis Ghosh and Jeremy Taylor, my doctoral advisors, for their infinite patience. This dissertation would not have come together without their constant guidance and encouragement. To the members of my doctoral committee: I thank Dr. Arul Chinnaiyan for sharing the cancer genomic data sets that instigated the studies of this dissertation, for engaging me in the various scientific projects that I have greatly enjoyed, and above all for being a perpetual source of confidence and inspiration; I also thank Professor Roderick Little for his knowledge and insights that he generously shared with me to improve this dissertation.

To my friends at Michigan, I thank Laila, Hyungwon, and Yun for the many hours we happily 'wasted' on talking about food, travel, and random gossips; Jun and Jincao for sharing our passion in football; and many friends with whom I have shared laughter and tears. My sincere thanks to Dr. Sinae Kim for her friendship in my final year at Michigan.

To my parents, I am deeply thankful of their unwavering faith in me. They always believed in my abilities every step of my life. My loving memories to my Mom who died of cancer when I was fourteen. To Mom, I dedicate my career to continue a journey she was not able to complete.

Last but hardly the least, I thank my husband Cong for sharing this life journey with me. I can not imagine going through everything without him being on my side.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

CHAPTER I

INTRODUCTION

1.1 Genomic biomarkers for cancer diagnostics and prognostics

The increasing availability of DNA microarray technology has spawn a large number of genome-scale gene expression profiling studies in cancer. These cancer microarray studies have shown potential of identifying genomic biomarkers that outperform standard clinical parameters as diagnosis and prognosis targets. For instance, in prostate cancer, screening for elevated serum prostate-specific antigen (PSA) level has become a standard clinical test for early detection, but is known to result in high percentage of false positives. Only about 30% of men with a "positive" PSA have a positive biopsy. New biomarkers are needed to improve early detection of prostate cancer. In this respect, studies have explored the utility of microarrays in identifying gene expression "signatures" with activated or repressed expression profiles in the disease status as potential diagnostic targets (Dhanasekaran et al., 2001, Luo et al., 2001, Welsh et al., 2001). Another example pertains to genomic studies in breast cancer. Estrogen Receptor (ER) positive status is a well-known predictor of patient response to hormonal therapy. In contrast, ER negative breast carcinomas generally lack effective treatment options and are correlated with higher risk of developing

disease recurrences. Much effort has been dedicated to finding gene expression signatures that can provide treatment guidance and predict patient recurrence outcome above and beyond standard clinical parameters such as ER status, lymph node status, stage of the disease (Huang et al., 2003, Sorlie et al., 2001, Sotiriou et al., 2003, van't Veer et al., 2002, Wang et al., 2005). For a review, see Van de Vijver (2005). Overall, genome-scale expression profiling has been a prolific approach in identifying novel molecular targets that delineate cancer subtypes and survival outcome.

1.2 Integrative analysis of DNA microarrays

A caveat in cancer genomic studies is that predictive genes identified in one study often can not be validated in another. When findings from independent studies are cross-examined, the individual gene signature sets tend to have little overlap in terms of gene identities. Such lack of concordance may be attributed to the differences in the study cohorts and analysis strategies. But a prominent source of variation comes from the use of different array platforms. Some of the commonly used microarray platforms include two-color spotted cDNA arrays, Affymetrix GeneChip arrays, and two-color long oligonucleotide arrays. Differences among these technologies include one- or two-channel formats, cDNA or oligonucleotide, in-house spotted or commercially developed. Expression profiling data generated from distinct array platforms can vary significantly in measurement scale and variance structure. In addition, the large p small n nature of microarray data also contributes to the problem. A search in a space of thousands of genes with a handful of samples does not lead to the gene selection stability one desires (Ein-Dor et al., 2005, 2006).

Integrative analysis of multiple studies, however, has shown great promise in compiling common gene expression patterns across data sets and even over distinct cancer

types (Rhodes et al., 2004). Various methods have been proposed for combining results across studies. Among these, Rhodes et al. (2002) proposed methods to summarize across studies the P-values from a two-sample test of gene expression differences between cancer and normal tissues. Choi et al. (2003) suggested combining effect size using a hierarchical model, where the estimated effect size in individual studies follows a normal distribution with mean zero and between study variance τ^2. From a Bayesian perspective, Wang et al. (2004) used data from one study to generate a prior distribution and subsequent microarray studies to update the parameter values of the prior.

A recent application of Bayesian mixture modeling to Microarray classification problems by Parmigiani et al. (2002) has given new insights into integrating different studies. The basic idea is to estimate the probability of over-, under- or normal expression for gene sample combinations given the observed expression measurements. As a result, *poe* (i.e., probability of expression) was introduced as a new scale and used in the context of molecular classification. The platform-free property of this scale, however, has motivated its potential use as a data transformation technique to facilitate data integration. In Chapter II, I propose an approach to meta-analyses of microarrays that is based on *poe*.

1.3 Analyzing protein expression data from Tissue Microarrays

DNA microarray studies often yield a few hundred candidate cancer genes displaying differential expression that is associated with a phenotype. Only a small portion of which will be eventually validated for the corresponding expression changes at the protein level. Translating these discovery-type findings into clinical relevance is

an important task. The advent of Tissue Microarray (TMA) technology (Kononen et al., 1998) has provided a proteomic platform for validation studies of those target discoveries. It has quickly become an integral part of cancer biomarker development (Figure 1.1). The main statistical issue in TMA data anlaysis is repeated measurements in each tumor. Immunohistochemical (IHC) staining assays are performed on multiple biopsy tissue elements of 0.6 mm in diameter and 4-8 μm in thickness to assess protein expression. The resulting staining pattern is traditionally evaluated by a pathologist and given an integer score on the scale of 0-3 to indicate no, weak, moderate, and strong staining. A primary goal of interest is to summarize such score across the multiple tissue samples and then associate with clinical outcomes of that tumor.

In Liu et al. (2004), the authors are concerned with various pooling methods (such as using the mean, median, minimum and maximum of the repeated measurements), and subsequent dichotomization of the scores. They propose to use a deviance-based survival tree (LeBlanc and Crowley, 1992) and a bump hunting (Friedman and Fisher, 1999) method for choosing the best predictor score for patient survival outcome analysis. However, TMA core-level repeated expression data harbor substantial biological and experimental variability. Those summary scores can yield large variability without explicitly adjusting for the intra-tumor expression variation.

To deal with repeated measurements, I propose a measurement error approach for analyzing quantitative protein expression data from TMA experiments. Our main interest is parameter estimation in proportional hazards models to associate the repeated core-level expression measures with patient survival outcome. In a two-stage method, I introduce a Latent Expression Index (LEI) to adjust for 1) the intra-tumor

variation, 2) the number of repeated measures, and 2) clinical covariates. A joint model is further established for simultaneous inference on the expression data and survival. When the quantitative intensity measure from the Chromavision system (Chromavision, San Juan Capistrano, CA) is concerned, a normality assumption is used on the logarithm transformed intensity measure.

The work in the final Chapter of this dissertation involves a generalization of the error model in Chapter III. One extension is to incorporate both the proportion and intensity measure of staining to summarize the protein expression profile of a tumor. For data based on a pathologist's evaluation, several empirical methods have been used. For example, a product score takes the product of the staining intensity level (0,1,2,3) and a crude proportion measure (0-100%). Another scoring system divides the staining proportion into six categories and then adds up with the intensity level (Allred et al., 1998).

Etzioni et al. (2005) pointed out that constructing such summary scores led to loss of information. The authors considered a compositional data analysis. For each tumor i, the authors define the observation vector as $(X_{i1}, X_{i2}, ...X_{iK}); i = 1, \cdots, n$. Here X_{ij} denotes the proportion of staining at intensity level $j = 1, \cdots, K$, and subjects to the constraint that $\sum_{j=1}^{K} X_{ij} = 1$. An additive log-ratio transformation is taken on X_{ij} and the transformed vector is then modeled as a multivariate normal distribution. In addition, a cumulative logit model is proposed to incorporate the order of the intensity levels. A major limitation of this method is that the staining proportion and intensity levels are considered error-free measures while in fact they represent a coarse evaluation at best from a pathologist's manual reading. In addition, this method is designed for traditional immunohistochemical experiments, which differ from quantitative TMA expeiments in two aspects: subjective (cate-

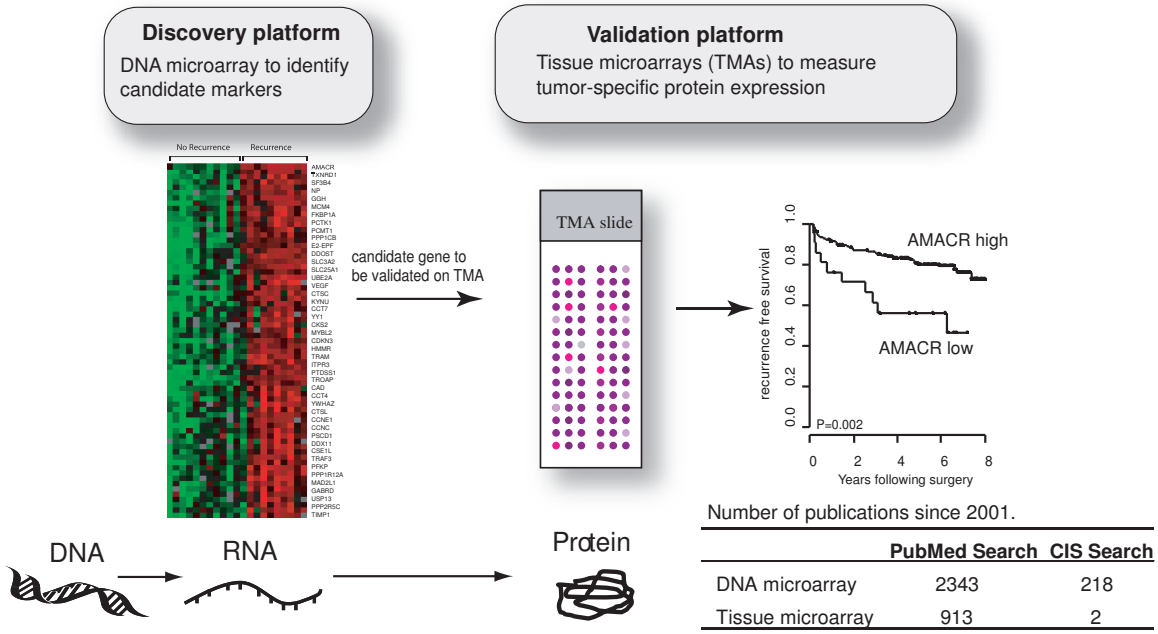
gorical) versus automated (continuous) data; one large section of the tumor versus multiple small sections. The proposed methods by Etzioni et al. (2005) is not directly applicable to quantitative TMA data.

In Chapter IV, I introduce a Cell Mixture Model (CMM) to incorporate both staining proportion and intensity measures from quantitative TMA data adjusting for measurement variability. Specifically, the protein expression profile measured from an individual core is modeled as a mixture distribution of a point-mass at zero to account for the proportion of non-staining and a continuous distribution for the intensity measure of positive staining. The whole-tumor expression profile is then reconstructed by aggregating over the individual mixture distributions. Here we deal with Chromavision data with quantitative measures that are considered substantially more accurate than those from a pathologist's scoring. However, measurement error still exists due to various reasons. A major source is the scarcity of the measurements taken per tumor. In TMA studies, the challenge is to estimate the whole-tumor expression characteristics with an average of three tissue cores each of 0.6 mm in diameters from a tumor that can be 100 times larger. An analogy is to estimate the characteristics of the population in the United States with data collected in three representative cities. In survey sampling problems, small area estimation often involves parameter estimation for small sub-population of interest. Hierarchical Bayes (HB) and Empirical Bayes (EB) approaches have been effective with continuous data. For a thorough review of various methods, see Ghosh (1994), Pfreffermann (2002), Rao (1999). For a unified analysis of discrete and continuous data, Ghosh et al. (1998) present hierarchical Bayes generalized linear models. The idea of Bayesian predictive inference and Markov Chain Monte Carlo integration technique is particularly useful for our problem at hand. In this study we extend the implementation to a zero-point

mass mixture distribution under the CMM model.

1.4 An outline of the dissertation

This dissertation is organized as follows. In Chapter II, a Bayesian mixture model based data transformation is introduced for the meta-analysis of DNA microarrays. An application of the meta-analysis approach to assimilate and analyze four independent breast cancer microarray studies is discussed. Chapter III presents the use of measurement error models for the analysis of tissue microarrays. I focus on the parameter estimation and associated inferences in censored failure time regression in the presence of measurement errors. Both a two-stage plug-in approach and a joint model of the TMA core-level repeated measures and survival are introduced. Chapter IV presents a Cell Mixture model as a generalized modeling framework for the reconstruction of complex staining patterns from TMA experiments.



Number of publications since 2001.

	PubMed Search	CIS Search
DNA microarray	2343	218
Tissue microarray	913	2

Figure 1.1: A Paradigm for Genomic Biomarker Development.

CHAPTER II

A TWO-STAGE MIXTURE MODEL FOR META-ANALYSIS OF MICROARRAY DATA

2.1 introduction

DNA microarray analysis has been shown to be a powerful tool in various aspects of cancer research. With the increasing availability of published microarray data sets, there is a tremendous need to develop approaches for validating and integrating results across multiple studies. A major concern in the meta-analysis of DNA microarrays is the lack of a single standard experimental platform for data generation. Expression profiling data based on different technologies can vary significantly in measurement scale and variation structure. It poses a great challenge to compare and integrate results across independent microarray studies. In a recent study of diffuse large B cell lymphoma (DLBCL), Wright et al. (2003) sought to bridge two different microarray platforms by validating findings from a cDNA lymphochip microarray using an independent dataset generated using Affymetrix oligonucleotide arrays. Although the idea of training and testing classifiers is frequently used for discriminant analysis, this application to distinct expression array platforms is less common.

More systematic approaches have been proposed for integration of findings from

multiple studies using different array technologies. Rhodes et al. (2002) have proposed methods to summarize significance levels of a gene in discriminating cancer versus normal samples across multiple gene profiling studies. By ranking the q-values (Storey, 2002) from sets of combinations, a cohort of genes from the four studies was identified to be abnormally expressed in prostate cancer. Choi et al. (2003) suggested combining effect size using a hierarchical model, where the estimated effect size in individual studies follows a normal distribution with mean zero and between study variance τ^2. The effect size was defined to be the difference between the tumor and normal sample means divided by pooled standard deviation. From a Bayesian perspective, Wang et al. (2004) used data from one study to generate a prior distribution of the differences in logarithm of gene expression between diseased and normal groups, and subsequent microarray studies updated the parameter values of the prior. Assuming a normal error distribution, the differences were then combined to form a posterior mean. Although phrased using different model frameworks, these methods are similar in the spirit of combining the standardized differences between two sample means across multiple studies. It has been shown, however, that the overlap between significant gene detection on different array platforms is only moderate due to low comparability of independent data sets (Mah et al., 2004). The large variability brought in by microarray datasets using different platforms is expected to affect the sensitivity and specificity of summary statistics constructed in various ways across studies. Given the inherent differences of the microarray techniques, heterogeneity of the sample populations, and low comparability of the independently generated data sets, meta-analysis of microarrays remains a difficult task.

A recent study proposed a Bayesian mixture model based transformation of DNA microarray data with potential features applicable to meta-analysis of microarray

studies (Parmigiani et al., 2002). The basic idea is to estimate the probability of over-, under- or baseline expression for gene sample combinations given the observed expression measurements. With data-driven estimation of these quantities, one can translate the raw expression measurement into a probability of differential expression. As a result, *poe* (i.e., probability of expression) was introduced as a new scale and used in the context of molecular classification (Parmigiani et al., 2002). The platform-free property of this scale, however, motivated us to incorporate *poe* in a framework to meta-analyze microarray data. Several desirable features of using *poe* as a new expression scale include the following: 1. *poe* provides a scaleless measure and thereby facilitates data integration across microarray platforms; 2. *poe* is a model-based transformation with direct biological implications in the context of gene expression data, as it is estimated based on a method that adopts an underlying mixture distribution that accommodates over-, under-, and unchanged expression categories; 3. *poe* unmasks differential expression patterns in microarray data by offsetting the influence of extreme expression values (Scharpf et al., 2003); 4. Data integration based on *poe* allows merging of samples on the unified scale rather than using gene-specific summaries.

In recent publications of breast cancer microarray studies, several groups have explored the hypothesis that the capacity to metastasize is intrinsic to the tumor and therefore can be revealed by gene expression pattern. Four independent studies have correlated gene expression profiles generated from distinct DNA microarray platforms to breast cancer prognosis (Huang et al., 2003, Sorlie et al., 2001, Sotiriou et al., 2003, van't Veer et al., 2002). Among the four, Sorlie et al. (2001) and Sotiriou et al. (2003), both cDNA microarray studies, applied unsupervised clustering and identified several breast cancer subtypes characterized by differential expression of

a cohort of genes. Further, they correlated the tumor subtypes derived from the expression profile with survival outcome and in both cases found that, as expected, the ERBB2+ subtype correlated with shorter survival times. On the other hand, van't Veer et al. (2002), an inkjet oligonucleotide array study, and Huang et al. (2003), an Affymetrix GeneChip study, have built classification models based on gene expression profiles to predict 5-year or 3-year recurrence status. In all four studies, however, the authors explored a common hypothesis that molecular profiles were able to provide a more accurate prediction of patient survival compared with clinical/pathological parameters. These studies therefore provided an excellent basis for developing a meta-analysis of microarrays with regard to disease prognosis.

This Chapter is organized as follows. In section 2.2 and 2.3, we propose a two-stage meta-analysis of microarrays with a focus on cancer prognosis prediction. In section 2.4, we apply our method to the aforementioned breast cancer DNA microarray data sets. We demonstrated the advantage of a mixture model based transformation for data integration and the gains of integrated data analysis over single analysis. Such two-stage meta-analysis approach allows an inter-study validated meta-signature based on gene expression to be developed for more robust and reliable cancer prognosis prediction across heterogeneous tumor samples.

2.2 Model based data transformation

Let x_{ijk} denote the preprocessed gene expression measurement for gene i from the jth sample in the kth study, transformed using the base two logarithm, $i = 1, \ldots, N$, $j = 1, \ldots, M_k$, $k = 1, \ldots, K$. We assume that data have been preprocessed, either by a lowess normalization for two-channel microarray data (Yang et al., 2002) or a robust analysis for Affymetrix data (Irizarry et al., 2003b). Then the available data

can be summarized by $\{X_k\}_{k=1}^K$, where X_k is a M_k by G matrix whose (i,j)th entry is x_{ijk}. Note that the value and interpretation of x_{ijk} is inherently different across array platforms and is not necessarily comparable if they are measured from independent studies. Corresponding to x_{ijk}, let e_{ijk} be a variable that takes one of three values $\{1, 0, -1\}$, indicating over-, baseline- or under- expression respectively for gene j in sample i for the kth study. If e_{ijk} were known, then this is a variable that would provide a platform-free scale which could be combined across multiple studies. We approach this problem by treating e_{ijk} as a latent variable that is inferred from the data using a mixture model.

2.2.1 A Normal-Uniform mixture distribution

We assume that x_{ijk} are realizations of the following mixture model:

$$\begin{aligned} x_{ijk} \overset{\text{iid}}{\sim} \quad & \pi_{jk}^{+} U(\alpha_{ik} + \mu_{jk}, \alpha_{ik} + \mu_{jk} + \kappa_{jk}^{+}) \quad + (1 - \pi_{jk}^{+} - \pi_{jk}^{-}) N(\alpha_{ik} + \mu_{jk}, \sigma_{jk}^2) \\ + \quad & \pi_{jk}^{-} U(\alpha_{ik} + \mu_{jk} - \kappa_{jk}^{-}, \alpha_{ik} + \mu_{jk}) \quad , \end{aligned} \tag{2.1}$$

where $\alpha_{ik} + \mu_{jk}$is both the mean of the normal distribution and the boundary to the two uniform distributions; α_{ik} is the sample effect with the constraint that $\sum_{i=1}^{M_k} \alpha_{ik} = 0$; κ_{jk}^{+} and κ_{jk}^{-} provide limits to the uniform distribution in the mixture, and are set to be at least $3\sigma_j$. The parameters $\pi_{jk}^{+} \equiv P(e_{ijk} = 1)$ and $\pi_{jk}^{-} \equiv P(e_{ijk} = -1)$ are the multinomial probabilities for the latent variable e_{ijk}. Conceptually, we can think of gene expression arising from three populations of genes in model (2.1). The first component in the model is the population of expression levels for genes that are overexpressed in the cancer samples relative to the normal samples, the second corresponds to genes that do not change between cancer and normal samples, and the third is for genes that are underexpressed in cancer samples relative to normal.

2.2.2 Posterior probability of differential expression

Let $p^+_{ijk} \equiv P(e_{ijk} = 1|x_{ijk})$ and $p^-_{ijk} \equiv P(e_{ijk} = -1|x_{ijk})$ be the conditional probabilities of over and underexpression for gene j in sample i (over- and underexpression respectively) given the microarray measurements. Then by Bayes' rule,

$$p^+_{ijk} = \frac{\pi^+_{jk} f_{1jk}(x_{ijk})}{\pi^+_{jk} f_{1jk}(x_{ijk}) + \pi^-_{jk} f_{-1jk}(x_{ijk}) + (1 - \pi^+_{jk} - \pi^-_{jk}) f_{0jk}(x_{ijk})} \tag{2.2}$$

and

$$p^-_{ijk} = \frac{\pi^-_{jk} f_{-1jk}(x_{ijk})}{\pi^+_{jk} f_{1jk}(x_{ijk}) + \pi^-_{jk} f_{-1jk}(x_{ijk}) + (1 - \pi^+_{jk} - \pi^-_{jk}) f_{0jk}(x_{ijk})}, \tag{2.3}$$

where f_{0jk} is the normal density function, and f_{1jk}, f_{-1jk} are the corresponding uniform densities for the differential expression categories for the jth gene in the kth study. In the numerator of (2.2), $f_{1jk} = 1/\kappa^+_{jk}$ if $x_{ijk} \in [\alpha_{ik} + \mu_{jk}, \alpha_{ik} + \mu_{jk} + \kappa^+_{jk}]$ and 0 otherwise; whereas in the numerator of (2.3), $f_{-1jk} = 1/\kappa^-_{jk}$ if $x_{ijk} \in [-\kappa^-_{jk} + \alpha_{ik} + \mu_{jk}, \alpha_{ik} + \mu_{jk}]$ and 0 otherwise.

Note that the supports of the two uniform distributions are disjoint. As a result, the probabilities of differential expression are mutually exclusive with the following forms:

$$(p^+, p^-) = \left(\frac{\pi^+/\kappa^+}{\pi^+/\kappa^+ + (1 - \pi^+ - \pi^-) f_0}, \quad 0 \right)$$

or

$$(p^+, p^-) = \left(0, \quad \frac{\pi^-/\kappa^-}{\pi^-/\kappa^- + (1 - \pi^+ - \pi^-) f_0} \right).$$

We then construct the following measure: $p^d_{ijk} = p^+_{ijk} - p^-_{ijk}$, ranging from -1 to 1. It can be interpreted as the signed conditional probability of differential expression of gene j in sample i in study k. The interpretation and scale of the measure is portable across array platforms and independent study data sets.

2.2.3 Estimation using MCMC algorithms

In this situation, we assume that there is only $k = 1$ study. Let $\{\Theta_j\}_{j=1}^N$ generically denote the parameter in model (2.1) for gene $j = 1, ...N$. In the microarray data setting, the total number of genes N can be a few thousands, leading to large amount of gene-specific parameters. It is sensible to adopt a hierarchical Bayesian mixture model setting for parameter estimation, where the variation of the gene-specific parameter estimates can be described by assuming prior distributions $f(\Theta_j|\psi)$ with hyperparameter space ψ. In particular, let

$$\mu_j \sim N(\xi, \tau^2), \qquad \kappa_j^+ \sim \text{Exp}(\lambda_\kappa^+), \quad logit(\pi_j^+) \sim N(\nu^+, \omega^+),$$
$$\sigma_j^{-2} \sim \text{Gamma}(\gamma, \lambda)\ , \quad \kappa_j^- \sim \text{Exp}(\lambda_\kappa^-), \quad logit(\pi_j^-) \sim N(\nu^-, \omega^-).$$

In terms of prior choice, we follow the recommendations of Parmigiani et al. (2002). To sample from the posterior distributions of the parameters, a Metropolis-Hastings MCMC algorithm was then implemented where the gene-specific parameters were repeatedly sampled from the corresponding full conditional distributions. These are given in Appendix A. We thus fit the Bayesian algorithm to each microarray dataset separately.

2.2.4 Linear rescaling

An alternative approach to integrating data across multiple datasets is to perform a study-specific global normalization. For the kth study, let $x_{ij}^k \equiv (x_{ij} - \bar{x})/s.d.(x_{ij})$ be the globally scaled expression value for gene$j$ in sample i. Each study dataset is then standardized to have zero mean and unit standard deviation. This yields a data matrix, say X_k^l for the kth study. The linearly rescaled values can also be used for data integration purposes in that expression values generated from different array platforms are standardized to a common scale.

Such an approach is much less computationally challenging compared to the mixture model-based rescaling described in the previous section. However, there are several advantages to the mixture model-based transformation. First, the method incorporates biological information into estimating the posterior probabilities of expression. The transformed values carry meaningful interpretations as signed probabilities of differential expression of a gene in a particular sample. Second, the underlying normal and uniform mixture distributions give equal density in the tails and is effective in reducing the influence of extreme expression values. And third, the Bayesian hierarchical modeling approach borrows strength across genes resulting in shrinkage-type estimators for a large correlated gene-specific parameter vector. This is a method in which the high dimensional gene expression data are denoised. We compare the performances of the two methods (mixture model-based and global standardization) in Section 2.4.

2.3 Data integration and meta-analysis

2.3.1 Integration of transformed data

Let X_k^* be the study-wise transformed expression data for the kth study. For the mixture model-based transformation, $X_k^* = P_k^d$, where P_k^d is a probability matrix with entries $p_{ij}^* \equiv p_{ij}^+ - p_{ij}^-$ as described earlier; and for the global standardization method, $X_k^* = X_k^l$, where X_k^l is the globally standardized logarithm expression matrix for the kth study whose (i, j)th entry is x_{ij}^k. For a common set of N genes that are profiled in each of the study of interest, data integration is subsequently based on the rescaled values X_k^*, and results in a combined data matrix of dimension $N \times \sum_{i=1}^{K} M_k$.

2.3.2 Classification methods for assessing meta-analysis procedures

We will assess the performance of the genes found using the meta-analysis methods based on classification accuracy. A complication is that while most methods of classification deal with data from two populations, the response with which we wish to build classifiers to predict is time to breast cancer recurrence. While the ideal data would be have information on time to recurrence on all subjects (potentially censored), not all studies have the time to recurrence information available and instead provide data on recurrence within a certain time interval (e.g., recurrence within three years versus no recurrence within three years).

To deal with this issue, we will utilize a dichotomization. Let T_i be the event time for subject i, C_i be the censoring time for subject i, and $\delta_i = 1\{T_i < C_i\}$ be the censoring indicator. Define a new variable,

$$y_i = \begin{cases} 1 & \delta_i = 1 \\ 0 & \delta_i = 0 \text{ and } C_i \geq t^*, \end{cases}$$

where t^* can be specified with clinical knowledge. The low risk group $y_i = 0$ has to satisfy the additional constraint $C_i \geq t^*$ to reduce potential bias introduced by insufficient length of follow-up in certain cohort. This is particularly relevant in cross-study analysis, given the potential heterogeneity in patient recruit criteria and study designs. In this paper, we have chosen $t^* = 3$ years. We then consider constructing classifiers using y; note that $y = 1$ corresponds to the poor outcome group and $y = 0$ to the good outcome group. Across the 305 samples from the four studies, 51.1% had $y = 1$.

Logistic regression was used to build a classifier for prognosis. For each gene j,

we fit the following univariate logistic regression model using data from all studies:

$$\text{logit}\{Pr(y_i = 1|x^*_{ij})\} = \eta_j + \beta_j x^*_{ij},$$

where x^* is the rescaled value that allows data integration across multiple studies. The estimated values of β_j, $\hat{\beta}_j$, are then used to form a risk score using a variation of the compound covariate predictor method (Radmacher et al., 2002, Tukey, 1993); for a given set of covariate values $x_1, \ldots, x_N$, the risk index is given by $RS = \sum_{j=1}^{N} \hat{\beta}_j x_j$.

If we want to assess the performance of the classifier, we must deal with the issue of training and testing the model using the same data. An "honest" estimate of the prediction error rate is obtained using leave-one-out cross-validation. Define a risk index $RI_i = \sum_{j=1}^{p} \hat{\beta}_{j,-i} x^*_j$, where $i = 1, \ldots, \sum_{i=1}^{K} M_k$, and $\hat{\beta}_{j,-i}$ is the effect estimate for gene j in the combined meta-cohort without the i^{th} sample. The risk index for sample i is a weighted linear combination of the expression profiles of the top p genes, where the ranking of the genes is based on their corresponding significance in the univariate logistic model fit. As a result, large positive values of RI indicate high risk of failure, whereas large negative values of RI indicate low risk of failure. Classification of sample i to the risk groups is then based on the i^{th} leave-one-out risk index. The classifier is $\mathcal{C}(X^*) = I\{RI_i > c\}$, with c being the empirical quantiles of the $RI's$. The number of genes p in a classifier is also treated as a parameter and optimized to minimize the prediction error rates.

2.4 Application to breast cancer data sets

Figure 2.1 depicts the workflow of applying the mixture model based meta-analysis. In the following sections, details involved in each step of the data application will be discussed.

2.4.1 Data collection and preprocessing

The four breast cancer microarray datasets mentioned in the Introduction were obtained at the author's websites from four recently published studies. Numerical descriptions of the studies are provided in Table 2.1. To perform the meta-analysis, we focused on a common set of $N = 2555$ genes compiled across array platforms by Unigene Cluster IDs that were present in all four studies. There are issues in attempting to match genes from multiple studies with different platforms (Ghosh et al., 2003), but we will ignore them in this paper. Because we are using genes only if they are present in all four studies, we exclude many genes from the analysis. While this leads to a loss of potential predictive features, it is not unreasonable to assume that the common set across studies represents the most relevant genes of interest for breast cancer prognosis.

Each data matrix of genes was then base-two log-transformed and normalized by median centering and dividing by the standard deviation for each gene. The mixture-model based approach requires complete data, missing expression values were imputed by the k-nearest neighbors imputation algorithm (Troyanskaya et al., 2001), with $k = 10$. As stated earlier, the goal of the analysis was to find a meta-signature which represents genes that discriminate samples that are recurrence-free for at least three years after surgery versus those that have recurrence within three years.

The first stage of the analysis involves data-driven estimation of the signed probability of differential expression, namely $p^d = p^+ - p^-$. The resulting values of p^d represent signed probability of differential expression for gene j in sample i, and thus provide a unified measure across studies. In the second stage, the expression profiles of tumor samples from multiple studies were combined on the p^d scale to generate

what we term a meta-cohort of genes. Class prediction for disease recurrence was then assessed based on the combined data. We define the meta-signature to be the optimal gene expression based classifier constructed in the combined data.

2.4.2 Identification of a 90-gene meta-signature

By minimizing the misclassification error in the meta-cohort via a leave-one-out cross-validation, we obtained a 90 gene meta-signature that reliably predicts outcome. This meta-signature classified 122 patients into a high risk group, where 84 (69%) of them had a recurrence. On the other hand, the signature classified 183 patients into a low risk group, where 118 (64%) of them did not recur by the end of the followup. By cross-tabulating the risk groups predicted by the meta-signature and the actual recurrence status, we obtained an estimated odds ratio of 4.0 (95% CI: 2.5-6.5, $P < 0.0001$).

A heat map representation of the *poe* profile for the 90 gene meta-signature revealed two distinct patterns of differential expression (Figure 2.2 top panel). Genes display consistent differential expression probabilities (yellow indicate over-expression and blue indicate under-expression) in the recurrent samples (R). By contrast, an example of the individual signature (bottom panel) shows a cohort-specific expression pattern that clearly can not be reproduced in independent data sets. In Figure 2.3, functional annotation revealed genes involved in many important biological processes such as cell cycle regulation (e.g., CDC28 protein kinase regulator subunit 2), cell adhesion (e.g., chemokine C-X3-C motif receptor 1), and apoptosis (e.g., secreted frizzled-related protein 4).

2.4.3 Comparison of the meta-signature with study-specific signatures.

To comprehend the potential gains of a two-stage meta-analysis over analysis of a single dataset, we compared the performance of the meta-signature to that of the individual signatures.

By minimizing the prediction errors, we obtained a set of individual signatures consisting of 10, 60, 100, and 130 genes for the Sorlie, van't Veer, Sotiriou, and Huang studies, respectively. The results of the classifiers are summarized in Table 2.2. Not only did the sizes of the study-specific signatures vary significantly, but the elements of the signatures had very little overlap. At most two genes appeared in more than one signature among the four. In addition, signatures identified in one study tended to have poor prediction in other studies. These results are presented in Table 2.3. Except for two cases (the Sorlie study signature in Huang study cohort and the Sotiriou study signature in the van't Veer study cohort), there was an increase in classification error of approximately $20 - 60\%$ in the testing sets relative to training sets.

The gene signature found by meta-analysis improves on the individual study-specific signatures in two ways. First, its overlap with the study-specific signatures ranged from $3 - 40\%$ (Table 2.2). The excluded genes are likely to be cohort-specific findings that can not be replicated. By contrast, the meta-analysis is able to detect genes that have slight signals in the individual analyses based on combining the data. Second, the meta-signature recruited 41 genes not previously picked by any of the single cohort signature, likely representing predictive features with small but consistent effects previously masked in single studies. When comparing the performances of the gene signatures, the meta-signature performed, on average, similarly to the

individually optimized signatures in differentiating patients at low risk of recurrence from those at high risk of recurrence in each single study cohort (Table 2.3, comparing bottom row with the diagonals). This shows that the meta-signature can serve as a common breast cancer recurrence index that is able to predict patient survival in heterogeneous sample populations. When a gene signature built in one study cohort performs differently in another, such meta analysis provides a solution to identify a cross-study validated expression signature that holds across independent sample cohorts.

2.4.4 Comparison with simple linear rescaling.

To study the potential benefit of data integration based on p^d compared to that based on x^l, described in Section 2.2.4. We applied the same classifier to data combined on global standardization and compared the model performances based on data integrated by these two transformation strategies. Figure 2.4A shows that with the p^d transformation, misclassification rates steadily decreases as more genes are used in the classifier. Performance based on the linearly rescaled data (Figure 2.4B), however, is unpredictable. Figure 2.4C and 2.4D use a 90-gene meta-signature based on the mixture model transformation and global standardization, respectively, for predicting recurrence. The signature based on the signed probability of differential expression (p^d) is noticeably better than the signature based on the global standardization (x^l), in differentiating patients at low risk of recurrence from those at high risk of recurrence. Taken together, the mixture model based transformation outperforms the linear rescaling method in combining multiple microarray data sets. The meta-signature identified based on p^d measures therefore offers more reliable prediction of recurrence-free survival in the meta-cohort of breast cancer patients.

2.4.5 Independent validation of the meta-signature

Independent validation of a gene signature is essential in assessing the true predictive value of the finding. Two data sets are considered for the validation of the meta-signature. One includes 295 consecutive patients with primary breast carcinomas from the Netherlands Cancer Institute published in van de Vijver et al. (2002). The second one published by Wang et al. (2005) consists of frozen tumor samples from 286 patients with lymph-node-negative breast cancer who were treated at the Erasmus Medical Center (Rotterdam, Netherlands) during 1980 to 1995. Figure 2.5 shows the validation Kaplan-Meier curves generated for the meta-signature and each of the individual study signatures. It is clear that the meta-signature stands as the most consistent performer of all. One of the individual signatures — the van't Veer 70-gene — performs better than the other individual signatures in this validation analysis, especially in Figure 2.5 (b). It should be pointed out that the van de Vijver validation cohort is not strictly independent for the van't Veer 70-gene signature as part of the 295 samples were used to generate that particular signature (this also affects the meta-signature, but to a less extent as other data sets are mixed in the meta-cohort).

2.5 Discussion

Several important issues to consider when integrating microarray studies include use of different gene expression measurement scales, varying analytical power and reliability of the results for individual studies. To address these issues in a meta-analysis framework, we proposed a two-stage mixture modeling strategy. The goal of the mixture model-based transformation is to transform the preprocessed data to the probability scale ($p^d = p^+ - p^-$), which is then integrated across datasets. In

particular, the signed probability of differential expression p^d is easily interpretable and is platform-independent. The Normal-Uniform mixture distribution under a Bayesian hierarchical model setting has several desirable properties such as reducing the influence of extreme tail elements; borrowing strength across genes for parameter estimation; and shrinkage for the estimation of the correlated vector of the gene-specific parameters.

At the second stage of the analysis, combining samples on the probability scale mitigates the influence of potential artifacts from a single study. The effect is reflected on two counts. One, integrated sample cohorts improve the reliability of the findings by guarding against false positive results from a single study. Two, it increases the statistical power to detect small consistent effects that can be otherwise masked by inadequacy of the sample size of an individual data set. By implementing this modeling approach, we were able to combine information from four microarray studies to build an inter-study validated meta-signature for predicting recurrence in breast cancer patients.

As described earlier, a common set of 2555 genes was used in this meta-analysis, as it is important to provide the same context for data-driven estimation of the posterior probabilities. Although we assume the common set comprises the most biologically relevant genes, the loss of potential predictive genes, however, may offset the statistical power of the analysis. Alternative approaches to allow genes profiled in some studies but not others is a topic for future research.

A distinction of the analysis presented here relative to those by other authors (Rhodes et al., 2002; Wang et al., 2004) is that we sought to find genes that were predictive of recurrence rather than predictive of diseased versus nondiseased status. Given the heterogeneity of the tumors with respect to treatment response and

survival outcome, a prognostic prediction analysis is generally more difficult because it is a more complicated phenotype. Further, a prognostic signature (classifier) of failure risk trained in one cohort is often times difficult to validate in independent cohorts. The meta-analysis method presented here may potentially provide more powerful gene signatures that are predictive of prognosis because they are validated across multiple studies.

Table 2.1: Description of the breast cancer gene expression datasets used in the meta-analysis.

Authors	Array Platform	No. of array elements	sample size	good outcome	poor outcome
Sorlie *et al.*	Spotted cDNA	8102	58	23	35
van't Veer *et al.*	Inkjet oligonucleotide	25000	78	44	34
Sotiriou *et al.*	Spotted cDNA	7650	99	54	45
Huang *et al.*	Affymetrix chip	12625	89	54	35

Table 2.2: Comparisons of the signatures. Table lists the number of genes (Size), the number of genes overlap with the meta-signature (overlap), and the prediction error rate for the classifiers identified in individual study cohort and in the meta-cohort.

	Signature				
	Sorlie	van't Veer	Sotiriou	Huang	**Meta**
Size	10	60	90	140	**90**
Overlap	4	14	19	6	-
Prediction error rate	0.28	0.29	0.35	0.18	**0.33**

Table 2.3: Comparison of the performances of the individual signatures and the meta-signature. Table lists odds ratios (95% confidence interval) comparing the odds of actual recurrence for those being classified as high risk to the odds of recurrence for those being classified as low risk of recurrence by each signature.

		Cohort		
Signature(D)	Sorlie (n=58)	van't Veer (n=78)	Sotiriou (n=98)	Huang (n=71)
Sorlie (10)	18.6 (5.0, 69.5)	2.1(0.8, 5.4)	2.3 (1.0, 5.3)	10.87 (3.5, 33.8)
van't Veer (60)	3.1 (1.1, 9.2)	10.6(3.3,33.9)	4.1(1.7,9.7)	1.3(0.5,3.4)
Sotiriou (100)	1.7(0.6,5.0)	3.5 (1.4,8.9)	7.8(3.0,20.1)	1.5(0.6,3.7)
Huang (130)	5.1(1.6,15.7)	2.3(0.9,5.6)	0.9(0.4,2.0)	184.9(30.1,1137.2)
Meta (90)	25.0(4.2,149.0)	4.1(1.6,10.6)	6.0(2.5,14.5)	5.8(2.1,16.5)

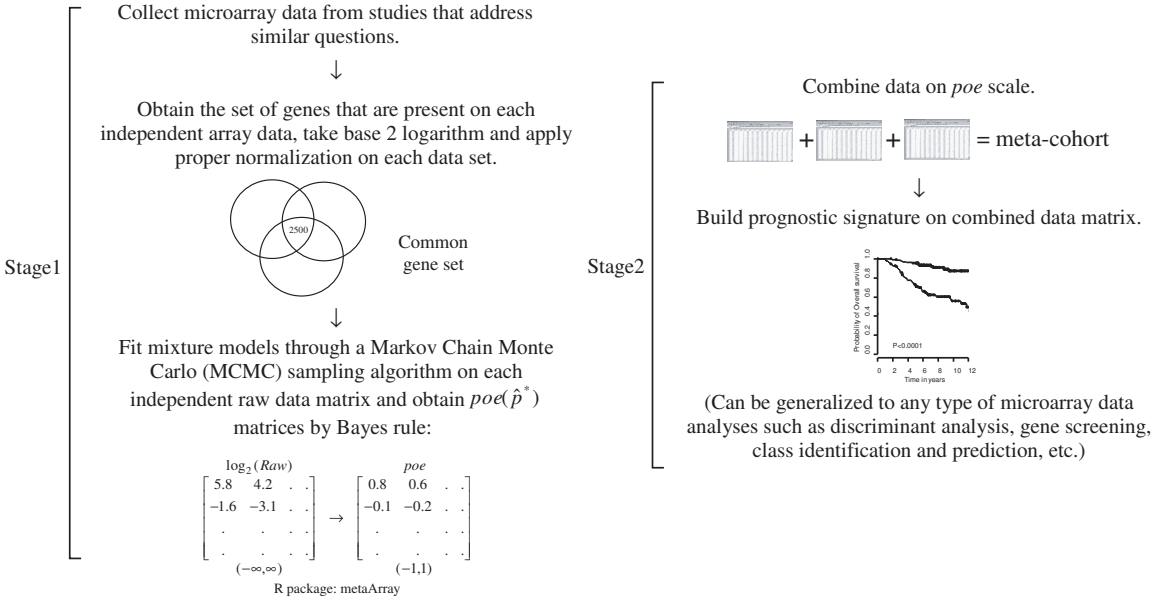


Figure 2.1: Diagram of the meta-analysis.

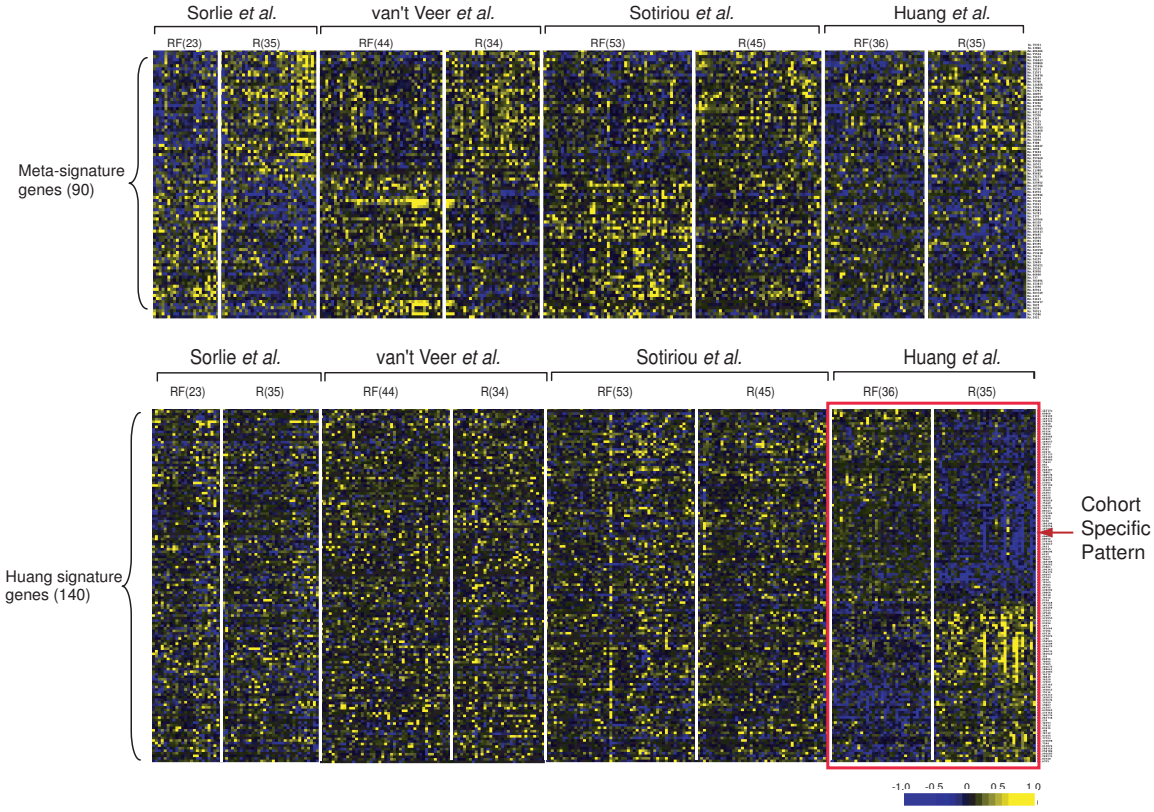
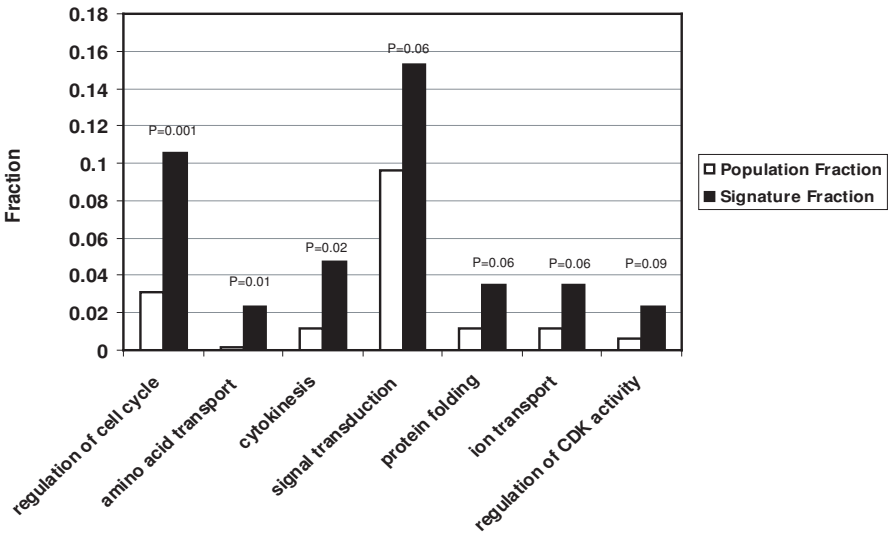


Figure 2.2: Heatmap representation of the 90 gene meta-signature expression pattern (top panel) and the Huang signature expression pattern (bottom panel).



Regulation of cell cycle	Amino acid transport	Cytokinesis	Signal transduction	Protein folding	Ion transport	Regulation of CDK activity
TFDP1	SLC3A2	PPP1CB	TXNRD1	FKBP1A	GABRD	CKS2
PCTK1	SLC12A5	CKS2	VEGF	CCT7	SLC21A3	CDKN3
VEGF		CCNE1	TRAF3	CCT4	SLC12A5	
CCT7		CCNC	GABRD			
MYBL2			PPP2R5C			
CCT4			SFRP4			
CCNE1			INSR			
CCNC			ADRA2A			
BCL2			SCAP1			
			GUCA1A			

Figure 2.3: Top seven over-represented functional classes in the meta-signature.

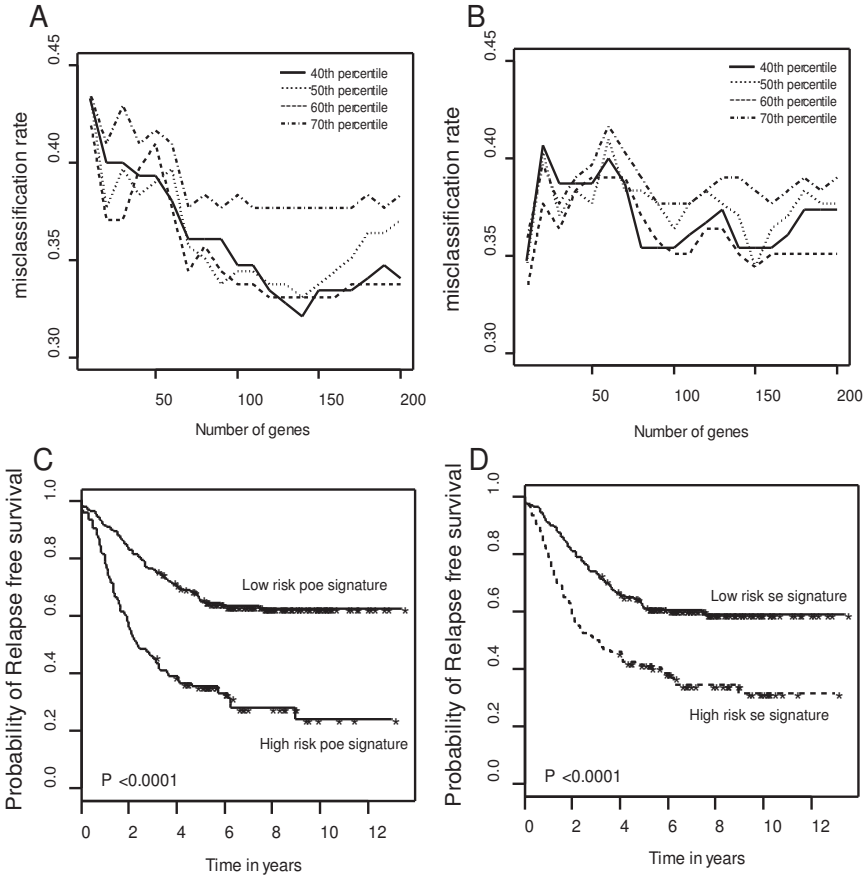
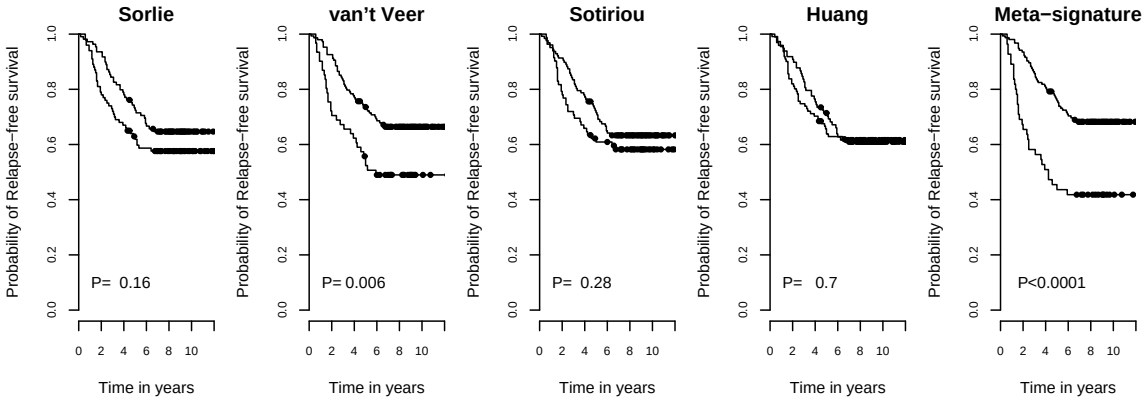
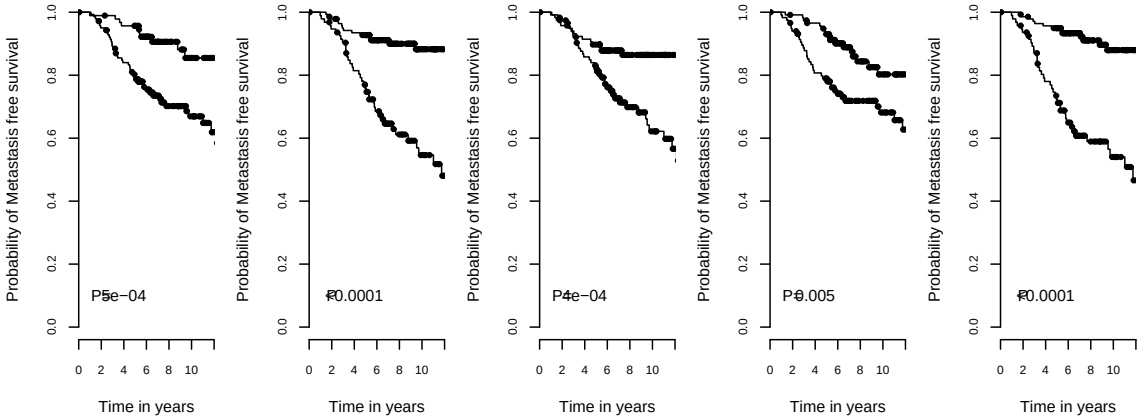


Figure 2.4: Comparison of model performances based on data integrated by the *poe* transformation (A and C) and global standardization (B and D). A. Misclassification rates based on poe transformation and B. based on global standardization. C. Performance of the 90-gene signature built on poe and D. built on global standardized data in differentiating patients at low risk of recurrence from those at high risk of recurrence.



(a) Wang *et al.* data set (n=286)



(b) van't Vijver *et al.* data set (n=295).

Figure 2.5: Validation of the signatures in two independent data sets.

CHAPTER III

MODELING INTRA-TUMOR PROTEIN EXPRESSION HETEROGENEITY IN TISSUE MICROARRAY EXPERIMENTS

3.1 Introduction

DNA microarray technology has enabled expression measurement of thousands of genes simultaneously, providing a platform for rapid screening of genomic biomarkers in cancer. Translating these discovery-type findings into clinical relevance is a more challenging task. Gene expression profiling studies using spotted cDNA arrays or Affymetrix GeneChip arrays often yield a few hundred candidate cancer genes that are associated with a phenotype. Only a small portion of these will be eventually validated for the corresponding expression changes at the protein level. The advent of Tissue Microarray (TMA) technology has provided a proteomic platform for such validation studies to find clinically useful biomarkers. TMA experiments measure tumor-specific protein expression via high-density immunohistochemical staining assays, allowing simultaneous evaluation of hundreds of patient samples in a single experiment (Kononen et al., 1998). Since their initial development, TMA-based expression studies have quickly become an integral part of cancer biomarker development (Divito et al., 2004, Rubin et al., 2005, Seligson et al., 2005).

A typical tissue array slide comprises up to 1000 tiny biopsy tissue elements, which we will refer to as cores, with multiple cores corresponding to repeated sampling from the same tumor. Expression measures on these replicate cores constitute the TMA core-level data. These can display substantial within-subject variability for both biological and experimental reasons. Biologically, for tumors that are highly infiltrative and heterogeneous in nature (e.g., prostate tumors), protein expression pattern can be quite variable. For example, cell proliferation genes often exhibit localized high expression within a tumor, indicating elevated aggressiveness and metastatic potential in the corresponding areas. Replicate sampling from various regions of the tumor is therefore important in capturing the underlying heterogeneity within a tumor. Experimental sources of the variability can come from a combination of probe affinity, measurement imprecision, and further missing data due to insufficient sampling. Without appropriately accounting for these variabilities, the noise-prone measurements tend to attenuate the prognostic value of a potential biomarker in predicting disease outcome. The lack of a model-based approach for TMA core-level expression data to effectively model the intra-tumor variation has motivated us to carry out a full investigation.

A good analogy for understanding TMA data structure is from probe-level data generated by the Affymetrix GeneChip arrays. GeneChip arrays measure gene expression at the mRNA transcripts level. The probe-level data refer to the replicate expression measures on a set of 16-20 small oligonucleotide probe sets derived for a target gene. The biological variation comes primarily from these oligonucleotide probe sequence variants, while experimental variations arise during the process of slide printing, hybridization and optical reading. Li and Wang (2001) reported that the variation of a specific probe across multiple arrays could be considerably smaller

than the variance across probes within a probe set. Modeling Affymetrix probe-level data has generated much attention (Irizarry et al., 2003a, Li and Wang, 2001) as the technology has become more mature and widely used.

Similarly in TMA experiments, modeling within-tumor protein expression heterogeneity is an important problem. In these tissue-based experiments, the variation across core samples within a tumor can be substantially larger than the variation observed across subjects. Etzioni et al. (2005) used a compositional analysis to model such heterogeneity, and compared the proportion of cells stained at different intensity levels between normal and tumorous tissues. In this study, we focus on the effect of modeling intra-tumor variation in the context of predicting patient survival outcome. In a latent variable modeling framework, we assume that an underlying 'true' expression value predicts survival. In real experiments, this 'true' expression can not be precisely measured due to sampling variabilities and measurement imprecision. Instead, one observes the core-level expression measurements that are subject to these measurement errors. In a two-stage method, we adapt ideas from measurement error modeling and propose a latent expression index (LEI) to approximate the underlying true value, and focus on its behavior in proportional hazards models. Specifically, we adapt an empirical Bayes estimator (Tsiatis et al., 1995) to 1) incorporate important clinical parameters such as Gleason score and pathological stage of the tumor and 2) adjust for the varying number of cores. We further establish a joint model for TMA core-level data and survival outcome via a shared random effect. There is a large literature on joint modeling of longitudinal data and survival (Brown and Ibrahim, 2003, Faucett and Thomas, 1996, Guo and Carlin, 2004, Henderson et al., 2000, Tadesse et al., 2005, Wang and Taylor, 2001, Wulfsohn and Tsiatis, 1997, Xu and Zeger, 2001). These methods have been developed predominantly for modeling sur-

vival and CD4 counts in AIDS patients; here their application to Tissue Microarray data in cancer biomarker studies is novel. Using both simulations and two published TMA data sets, the performances of the naive, two-stage LEI, and the joint model approach are compared in terms of the parameter estimates and associated inference.

The chapter is organized as follows. Section 3.2 specifies notation and models for the TMA core-level expression data and for the patient survival data. Section 3.3 introduces LEI and its use in a two-stage method. Section 3.4 presents the joint modeling approach and the Bayesian estimation framework. Simulation results to compare the performances of these methods are then discussed in Section 3.5. Case studies using two prostate cancer TMA data sets are presented in Section 3.6. Further discussion can be found in Section 3.7.

3.2 Model specification

Measurement model for the TMA core-level data.

Let X_i^* be the latent expression value for a biomarker in tumor $i, i = 1, \cdots, n$. Assuming the observed TMA core-level measurement for the jth core in the ith tumor is

$$X_{ij} = X_i^* + U_{ij}, \quad j = 1, ..., r_i; \quad i = 1, ..., n, \tag{3.1}$$

where we assume $X_i^* \sim N(\mu_{x^*}, \sigma_{x^*}^2)$. The mean μ_{x^*} is a linear function of clinical covariates: $\mu_{x^*} = \theta_0 + \theta \mathbf{Z}_i'$, where $\mathbf{Z}_i = (Z_{1i}, Z_{2i}, ..., Z_{pi})$ constitutes a row vector of p clinical parameters characterizing histologic and pathologic features of the tumor; and $\theta = (\theta_1, \ldots, \theta_p)$ is the associated p-dimensional row vector of effect sizes. In this model, U_{ij} represents the variation of the r_i core-level expression measurements. We assume U_{ij} is i.i.d. $N(0, \sigma_u^2)$ and independent of X_i^*.

Survival model

Let the observed survival time for patient i $(i = 1, ..., n)$ be $T_i = \min(Y_i, C_i)$, where Y_i is the time from diagnosis to disease recurrence; C_i is the time to censoring which is independent of Y_i, and $\delta_i = 1\{Y_i < C_i\}$ is the censoring indicator. Under the Cox proportional hazards model, the hazard rate for patient i is

$$\lambda(t) = \lambda_0(t)e^{\beta^* X_i^*}, \tag{3.2}$$

where $\lambda_0(t)$ is the baseline hazard function and β^* is the true regression coefficient. We also consider a parametric Weibull regression model with the following form for the hazard function:

$$\lambda(t) = \gamma t^{\gamma-1} e^{\beta_0 + \beta^* X_i^*}. \tag{3.3}$$

3.3 Two-stage plug-in method

Given the basic assumption that the measurement error U_{ij} has no predictive value, i.e., $\lambda(t|\mathbf{X}_i, X_i^*) = \lambda(t|X_i^*)$, Prentice (1982) introduced the induced hazard rate

$$\lambda(t|\mathbf{X}_i) = \lambda_0(t)e^{\beta^* E[X_i^*|\mathbf{X}_i]}, \tag{3.4}$$

and proposed to estimate β^* by maximizing the corresponding partial likelihood. Note that (3.4) is an approximation to (3.2). Define the Latent Expression Index (LEI) to be an estimate of the conditional mean, $\text{LEI}_i = \hat{E}[X_i^*|\mathbf{X}_i]$, for each subject i. A two-stage plug-in method can be described by the following algorithm: 1) Compute LEI_i $(i = 1, ..., n)$ as a surrogate expression estimate that adjusts for measurement error; and 2) Apply the Cox or Weibull regression model using LEI_i to obtain an estimate of β^* and the associated standard error.

In the next section, we will describe methods for computing LEI for tissue microarray data. These include an empirical Bayes estimator conditional on clinical covariates, a full Bayes approach and a Varying Replicate Number (VRN) method as an extension to adjust for the number of cores per tumor.

3.3.1 Methods for computing LEI

The Empirical Bayes and full Bayes estimator.

Express (3.1) as a mixed effects model

$$X_{ij} = \theta_0 + \theta \mathbf{Z}_i' + \nu_i + U_{ij}, \tag{3.5}$$

where $\nu_i \sim N(0, \sigma^2_{x^*})$. The empirical Bayes estimator can then be derived as

$$LEI_i^{eb} = \hat{\gamma}_i \bar{X}_i + (1 - \hat{\gamma}_i)(\hat{\theta}_0 + \hat{\theta}\mathbf{Z}_i'), \tag{3.6}$$

where $\hat{\gamma}_i \equiv \hat{\sigma}^2_{x^*}(\hat{\sigma}^2_{x^*} + \hat{\sigma}^2_u r_i^{-1})^{-1}$ is the attenuation factor (Carroll et al., 1995). Parameter estimates $\{\hat{\theta}_0, \hat{\theta}, \hat{\sigma}^2_u, \hat{\sigma}^2_{x^*}\}$ can be obtained by fitting a mixed effects model as described in (3.5), using a restricted maximum likelihood (REML) approach (Harville, 1977, Laird and Ware, 1982).

The empirical Bayes estimator conditions on the set of parameter estimates derived from the data. The uncertainty of these estimates are not accounted for in LEI^{eb}. For this reason, a full Bayesian estimator, LEI_i^{fb}, is also considered. Hyperprior distributions are adopted as follows: $\sigma_u^{-2}, \sigma_{x^*}^{-2} \sim \Gamma(r_0, \gamma_0)$. The full Bayes estimator

$$LEI_i^{fb} = \tilde{\theta}_0 + \tilde{\theta}\mathbf{Z}_i' + \tilde{\nu}_i,$$

is then based on the posterior inference from model (3.5) where $\{\tilde{\theta}_0, \tilde{\theta}, \tilde{\nu}_i\}$ are the posterior means given the data.

The Varying Replicate Number (VRN) method.

In a typical TMA construction, K_i cores are placed on the array for each tumor i. However, not all of the measurements $\{X_{i1}, \cdots, X_{iK_i}\}$ are available. Several reasons contribute to a varying number of replicate measure. These include heterogeneous tissue composition and technical defects such as image corruption. The expression measurement from non-tumorous tissue types or a corrupted image is typically considered unsuitable for an outcome analysis and excluded. Let $M_{ij} = 0, j = 1, \cdots, K_i$ indicate that the jth core from the ith tumor is lost due to the aforementioned reasons and $M_{ij} = 1$ if it is available. Expression measures are retained for $r_i \equiv \sum_{j=1}^{K_i} M_{ij}$ cores, where r_i varies across tumor samples and possibly depends on covariate $\mathbf{Z}_i$. We assume r_i to follow a Binomial distribution given K_i and $P(M_{ij} = 1)$ with possible over-dispersion. The following logistic mixed effects model is adopted:

$$\text{logit}P(M_{ij} = 1) = \psi_{0i} + \psi \mathbf{Z}_i', \tag{3.7}$$

where $\psi_{0i} \sim N(\psi_0, \sigma_\psi^2)$, $\mathbf{Z}_i = (Z_{1i}, Z_{2i}, ..., Z_{gi})$ is the vector of g clinical covariates that can be the same or different from those in (3.5), and $\psi = (\psi_1, \psi_2, ..., \psi_g)$ is the associated vector of coefficients. Therefore

$$r_i \sim \text{Binomial}\left(K_i, p_i = \frac{e^{\psi_{0i} + \psi \mathbf{Z}_i'}}{1 + e^{\psi_{0i} + \psi \mathbf{Z}_i'}}\right). \tag{3.8}$$

The expression index under the VRN model is then derived by averaging over all the possible values of (r_i, K_i). In particular,

$$\begin{aligned} LEI_i^{vrn} &= E_{(r_i, K_i)} E[X_i^* | \mathbf{X}_i, \mathbf{Z}_i, r_i, K_i] \\ &= \sum_{s=1}^{R} \sum_{m=0}^{s} \left\{ \frac{\hat{\sigma}_{x^*}^2 m}{\hat{\sigma}_{x^*}^2 m + \hat{\sigma}_u^2} \bar{X}_i + \frac{\hat{\sigma}_u^2}{\hat{\sigma}_{x^*}^2 m + \hat{\sigma}_u^2} (\hat{\theta}_0 + \hat{\theta} \mathbf{Z}_i') \right\} \\ &\quad \times \begin{pmatrix} s \\ m \end{pmatrix} \left(\frac{e^{\hat{\psi}_{0i} + \hat{\psi} \mathbf{Z}_i'}}{1 + e^{\hat{\psi}_{0i} + \hat{\psi} \mathbf{Z}_i'}} \right)^m \left(\frac{1}{1 + e^{\hat{\psi}_{0i} + \hat{\psi} \mathbf{Z}_i'}} \right)^{s-m} \hat{P}(K_i = s) \end{aligned} \tag{3.9}$$

An additional assumption for the above is that the expression measures do not correlate with r_i or K_i. Parameter estimates $\{\hat{\theta}_0, \hat{\theta}, \hat{\sigma}_u^2, \hat{\sigma}_{x^*}^2\}$ can be obtained by fitting a mixed effects model as described in (3.5). A logistic mixed effects model in the form of (3.7) was fitted to obtain $\{\hat{\psi}_{0i}, \hat{\psi}\}$. Estimation is via methods described in Lindstrom and Bates (1990) and McCulloch (1994). The empirical proportions were used for $\hat{P}(K_i = s)$.

In a relatively balanced TMA array where the number of replicate measures r_i does not vary much across subjects, $\gamma_i \equiv \sigma_{x^*}^2(\sigma_{x^*}^2 + \sigma_u^2 r_i^{-1})^{-1}$ is an approximately constant adjustment factor. The amount of shrinkage in LEI^{eb} toward the overall mean depends primarily on the ratio of the within- to between-subject variation in that particular data set. In our example, however, r_i is a highly variable quantity. It exerts a larger role in determining how much weight LEI_i^{eb} gives to a particular subject's data relative to the estimated population mean. The motivation for LEI^{vrn} is to provide a replicate number-averaged expression estimate that alleviates the variability induced by r_i in the empirical Bayes estimator.

3.4 Joint Modeling of survival and TMA core-level data

The two-stage approaches described above are attractive for their simplicity and straightforward interpretation. They require minimal computation and can be easily implemented using existing statistical packages. However, there are major limitations for the two-stage method (Tsiatis and Davidian, 2004). First, the two-stage method involves a first order approximation and ignores the second-order term $\beta^{*2}\sigma^2(X_i^*|\mathbf{X})$ in the induced hazard rate function (3.4). As will be illustrated in the simulation study, such approximation works well when β^* is close to zero, but otherwise lead to

sizeable bias in $\hat{\beta}^*$. Second, parameter estimates in the second stage do not account for the uncertainty in estimating LEI in the first stage. The associated standard error for $\hat{\beta}^*$ will be over-optimistic. Given these considerations, it is desirable to make inference based on the joint likelihood of the failure time and TMA expression data. In this study, a shared random effect model is adopted to induce correlation between the TMA data and the survival outcome.

Given the measurement model specified in (3.5) for the TMA data, we write the proportional hazards model for the survival outcome as

$$\lambda(t) = \lambda_0(t)\exp(\mathbf{b}\mathbf{Z}_i' + \beta^*\nu_i). \tag{3.10}$$

The parameter ν_i constitutes a shared random effect that connects the measurement model (3.5) and the survival outcome model (3.10). The expression data and survival times are then assumed to be independent given ν_i. The joint likelihood for $\{T_i, \delta_i, \mathbf{X}_i, \nu_i\}$ is therefore

$$L_{Joint} = \prod_{i=1}^{n} f(t_i, \delta_i|\nu_i, \mathbf{z}_i) \times \prod_{i=1}^{n} f(\mathbf{x}_i|\nu_i, \mathbf{z}_i) f(\nu_i) = L_{SURV} \times L_{ME}, \tag{3.11}$$

where

$$\begin{aligned} L_{ME} &= \prod_{i=1}^{n}\left(\prod_{j=1}^{r_i}\frac{1}{\sqrt{2\pi\sigma_u^2}}\exp\Big\{-\frac{(x_{ij}-\theta_0-\theta\mathbf{Z}_i'-\nu_i)^2}{2\sigma_u^2}\Big\}\right)\frac{1}{\sqrt{2\pi\sigma_{x^*}^2}}\exp\Big\{-\frac{\nu_i^2}{2\sigma_{x^*}^2}\Big\}; \\ L_{SURV} &= \prod_{i=1}^{n}\prod_{l=1}^{L}\lambda_l^{d_l}\exp(\sum_{i\in D_l}\mathbf{b}\mathbf{Z}_i' + \beta^*\nu_i)\exp(-\lambda_l\sum_{i\in R_l}\Delta_{il}e^{\mathbf{b}\mathbf{Z}_i'+\beta^*\nu_i}). \end{aligned} \tag{3.12}$$

We used a piecewise constant hazards model in which the time axis is partitioned into L disjoint intervals, $I_1, ..., I_L$, where $I_l = [a_{l-1}, a_l)$ with $a_0 < t_i$ and $a_L > t_i$ for all $i = 1, ..., n$. Assume a constant baseline hazard in the lth interval, $\lambda_0(t) = \lambda_l$ for $t \in I_l$. R_l is the set at risk at the beginning of interval l; d_l is the number of failures in interval l; and $\Delta_{il} = min(t_i, a_l) - a_{l-1}$.

Alternatively, a parametric Weibull model can be assumed for the survival outcome using the following hazard function:

$$\lambda(t) = \gamma t^{\gamma-1} \exp(b_0 + \mathbf{b}\mathbf{Z}_i' + \beta^* \nu_i). \tag{3.13}$$

When $\gamma = 1$, the above reduces to exponential distribution with constant failure rate $\exp(b_0 + \mathbf{b}\mathbf{Z}_i' + \beta^* \nu_i)$. The survival time component of the joint likelihood in (3.12) is then replaced by

$$L_{SURV} = \prod_{i=1}^{n} \left\{ \gamma t_i^{\gamma-1} e^{\gamma(b_0 + \mathbf{b}\mathbf{Z}_i' + \beta^* \nu_i)} \right\}^{\delta_i} \exp(-e^{\gamma(b_0 + \mathbf{b}\mathbf{Z}_i' + \beta^* \nu_i)} t_i^{\gamma}). \tag{3.14}$$

In a Bayesian estimation framework, the following prior distributions are specified for the model parameters:

$$\begin{aligned} &(\beta^*, \theta_0, \theta, b_0, b) \sim N(\mu_0, \sigma_0^2); \\ &(\sigma_u^{-2}, \sigma_{x^*}^{-2}) \sim \Gamma(r_0, \gamma_0); \\ &(\gamma, \lambda_l,\ l = 1, ..., L) \sim \Gamma(r_0, \gamma_0). \end{aligned} \tag{3.15}$$

Relatively noninformative hyperparameters are chosen, in particular, $\mu_0 = 0, \sigma_0^2 = 10000, r_0 = 0.001, \gamma_0 = 0.001$. Samples from the posterior distribution are obtained using Markov Chain Monte Carlo (MCMC) methods.

3.5 Simulation

3.5.1 Simulation Setup

The additive measurement error model in (3.5) with one covariate Z_{1i} is used to simulate the expression measure $X_{ij}, i = 1, ..., n$, and $j = 1, ..., r_i$. In this model, $\theta_0 = 0$ and $\theta_1 = 1$. Furthermore, $\nu_i \sim N(0, 1)$, $U_{ij} \sim N(0, 0.5)$. The covariate Z_{1i} is simulated from a $N(0, 1)$ distribution. The total number of cores sampled, K_i, takes values in $\{1, 2, ..., 12\}$ with $P(K_i = 6) = 0.4, P(K_i = 1) = ... = P(K_i = 5) = 0.1$,

and $P(K_i = 7) = ... = P(K_i = 12) = 0.017$, mimicking the proportions from the actual tissue array data set used in this study. The number of repeated measures $r_i \equiv \sum_{j=1}^{K_i} M_{ij}$ is simulated from a Binomial(K_i, p_i), where $p_i = 1 - \pi^{1/K_i}$ such that the missing proportion equals π. The survival time T_i is simulated from a proportional hazards model in the form of (3.2) with $\lambda_0(t) \equiv 1$ and $\beta^* = 1$ or 2. An additional covariate Z_{1i} is further assumed to associate with T_i with the coefficient being one. The censoring time is simulated from an independent exponential distribution that results in a 30% censoring proportion. Results are summarized over 100 such simulated data sets each of a sample size $n = 200$. In general, parameter values are assigned in the simulation to mimic those for the real data sets.

Computation of LEI^{eb}, LEI^{vrn} were carried out using the PROC MIXED and the IML procedure in SAS (SAS Institute, Cary, NC). LEI^{fb} and the joint models were implemented using OpenBUGS via the R interface BRugs (Spiegelhalter et al., 2003, Thomas, 2004). We ran two chains with 1000 burn-in and 1000 updates per chain for the MCMC convergence.

3.5.2 Simulation Results

The simulation results are summarized in Table 3.1. For $\beta^* = 1$ in the survival models, the naive approach (using $\bar{X}_i$ as a surrogate expression) attenuates the true effect size by around 25%. The coverage probability of a nominal 95% confidence interval of $\hat{\beta}^*$ is 0.10 at best. The two stage methods (LEI) achieved a considerable bias correction by adjusting for the measurement error in the LEI imputation. The joint modeling approach gives the best estimate $\hat{\beta}^* = 1.03$ and a coverage probability of 95% compared to the truth.

Next a larger effect size is simulated ($\beta^* = 2$). The bias in the two-stage ap-

proaches due to the first-order approximation is evident. Overall the two-stage methods generate less biased $\hat{\beta}^*$'s compared to the naive estimate. Nevertheless, these are 15-25% smaller than the true β^*. The coverages are poor for the two-stage approaches. The joint modeling approach should be advocated in this scenario for inference. Notice that the joint model estimates of β^* are slightly bigger than 2, especially under the Weibull distribution. This may be driven by the prior distributions adopted for the parameters under the Bayesian estimation scheme. We have observed that such a difference disappears when the sample size gets larger.

3.6 Case study in prostate cancer

3.6.1 Data description

In this study, we consider two prostate tumor tissue microarray data sets. The α-Methylacyl CoA racemase (AMACR) is a peroxisomal and mitochondrial enzyme that plays an important role in fatty acid metabolism. AMACR has been shown to consistently overexpress in prostate tumors (Rhodes et al., 2002). Rubin et al. (2005) profiled AMACR protein expression using a TMA constructed on 203 prostate tumors from a surgical cohort who underwent radical prostatectomy at the University of Michigan as a primary therapy for clinically localized prostate cancer diagnosed between 1994 and 1998. They found AMACR is a significant predictor of the Prostate Specific Antigen (PSA) failure in these 203 patients.

The second biomarker evaluated in this study is BM28. BM28 encodes a highly conserved mini-chromosome maintenance protein (MCM) that is involved in the initiation of genome replication. Bismar et al. (2006) profiled a total of 41 genes (including BM28) in a TMA-based proteomic study. They identified a 12-gene model showing the expression combination of the twelve genes significantly associates with

tumor progression and PSA failure in a set of 79 men following surgery for clinically localized prostate cancer. The expression level of BM28 however did not show significant prognostic value in their analysis. We chose BM28 to evaluate the possibility of its being a false negative biomarker due to measurement error.

For the AMACR data, an average of $K_i = 5.5$ (range: 2 to 12) tissue core specimens were taken from each tumor sample and put on a tissue array for immunohistochemical staining. After initial diagnostic evaluation of each core, an average of 29% (range: 0-86%) of the cores were excluded due to reasons discussed earlier, leading to 5.5% missing subjects. The BM28 data has a similar tissue array design. A quantitative imaging analysis of the staining intensity was obtained using the ACIS II (Chromavision, San Juan Capistrano, CA) system. The intensity level ranges from 0 to 255 chromogen intensity units, and is transformed using the natural logarithm (one unit added to avoid taking logarithm of 0) and normalized to have mean zero and standard deviation one. Disease recurrence is defined as a serum PSA increase >0.2ng/mL after radical prostatectomy. Censored observations are those free of the recurrence at the time of last follow-up.

3.6.2 Measurement error and regression attenuation

Figure 3.1 illustrates a considerable amount of measurement error in the core-level expression data from the two TMA experiments. In the AMACR data, methods-of-moments estimates of the variance components are $\hat{\sigma}_u^2 = 0.46$ and $\hat{\sigma}_{x^*}^2 = 0.54$. In the BM28 data, the methods-of-moments estimates of the variance components are $\hat{\sigma}_u^2 = 0.62$ and $\hat{\sigma}_{x^*}^2 = 0.21$. The within-subject variation is almost three times the between-subject variation.

In a Cox proportional hazards model context, we did a simple simulation where

$X_i^* \sim N(0,1)$, $\beta^* = 1$. A naive estimator $\bar{X}_i = r_i^{-1} \sum_{j=1}^{r_i} X_{ij}$ —the average core-level expression for tumor i— is used as the surrogate expression to replace X_i^* in (3.2). We simulate situations where measurement error is small (σ_u^2=0.1), moderate (σ_u^2=0.5), and large ($\sigma_{x^*}^2 = 1$). Figure 3.2 shows various degrees of regression attenuation in the estimate of β^* as a function of replicate number r_i and the amount of error σ_u^2. When the parameter values are set to resemble the AMACR data, the naive estimate of β^* is approximately 30% smaller than the true value. With the current TMA construction protocol specifying three cores per subject due to economic and tissue-preservation reasons, and a great amount of within-subject variability routinely observed in the core-level expression data, Figure 3.2 effectively conveys the importance of modeling measurement error in TMA data. In the following two sections, we implement the measurement error models to demonstrate how statistical inference differs from the previous results.

3.6.3 AMACR expression and biochemical recurrence in prostate cancer

In prostate cancer, Gleason score, pathologic stage and tumor size are among the most important clinical parameters. We include these as clinical covariates $\mathbf{Z}_i$ to adjust in the measurement model (3.5), the replicate number model (3.7), and the survival outcome model (3.10). In the measurement model, $\hat{\theta}_{TumorSize} = 0.32$ with an associated standard error of 0.13, indicating a marginal association of tumor size with AMACR expression level. In the replicate number model, $\hat{\psi}_{TumorSize} = 0.72$ with an associated standard error of 0.16, which is consistent with our expectation that a larger tumor sample provides more abundant number of cores.

Table 3.2 lists the estimates and associated standard errors (posterior standard deviation) of β^* in the outcome model. The measurement error adjustment has

significantly improved upon the naive estimate. The error-adjusted $\hat{\beta}^*$ is around 0.75 ($\hat{se}(\hat{\beta}^*) = 0.26$), approximately 31% larger in absolute value than the naive estimate which is 0.57 ($\hat{se}(\hat{\beta}^*) = 0.20$). The amount of attenuation in β^* is quite consistent with what we conclude from the simulated datasets in the previous section. In this dataset, the two-stage methods (LEI^{eb}, LEI^{fb}, LEI^{vrn}) perform equally well as the joint modeling approach. The simplicity and computational efficiency of LEI serves as a satisfactory core-level expression index for AMACR. However as mentioned earlier, the two-stage methods are based on a first-order approximation, the accuracy of which is largely driven by the size of β^* and the ratio of within- and between-subject variation. As will be shown in the other data example, the two-stage methods will not be always a suitable approach.

Kaplan-Meier curves are useful as a graphical representation of the prognostic value of a biomarker. We examined these plots by dividing the subjects into different risk groups based on the values of AMACR expression estimates derived under each method. In Figure 3.3(a), subjects with AMACR high, median, and low expression groups based on LEI^{vrn} and the joint model estimates (C and D respectively) are significantly better separated in terms of probability of recurrence-free survival, when compared to that using the naive mean estimates (A).

3.6.4 BM28 expression and biochemical recurrence in prostate cancer

The measurement model indicates a marginal association of pathologic stage of the tumor with BM28 expression: $\hat{\theta}_{PathStage} = 0.59$ ($\hat{se}(\hat{\theta}) = 0.22$). The replicate number model again suggests a strong dependence on the size of the tumor: $\hat{\psi}_{Tumorsize} = 1.12$ ($\hat{se}(\hat{\psi}) = 0.43$).

In Table 3.2, the differences under various models are more discernable in this

datasets. First, the Weibull and piecewise exponential model overall generate slightly different results given a small sample size (n=52). Second, the empirical Bayes estimate differs substantially from the full Bayes LEI estimate. It is likely due to the large uncertainties in the parameter estimates that determine LEI^{eb}. Finally, the bias introduced by the first-order approximation is prominent here. Both the coefficient β^* and the noise ratio in this dataset are much larger in magnitude compared to the AMACR data example. In this case, the two-stage methods alleviate regression attenuation, only to a limited extent. The joint model should be used for parameter estimation and associated inference.

Figure 3.3(b) plots the Kaplan-Meier curves using different expression estimates. The 5-year PSA recurrence-free survival probability is 0.95 ($\hat{se} = 0.05$) versus 0.58 ($\hat{se} = 0.10$) for low and high BM28 expression estimated by the joint model. Adjusting for measurement error in this dataset has made a dramatic change in the conclusion about the prognostic value of BM28, compared to the naive method.

3.6.5 Improved expression estimates

Figure 3.4 compares the naive and the joint model expression estimates, plotted against the survival time on the x-axis. The mean expression $\pm$ two standard deviations is plotted for each individual. Two improvements under the joint model are clear: 1) the noise, represented by the error bars, is greatly reduced via the joint modeling, and 2) the mean expression levels are distributed more tightly around a regression line, accentuating the relationship of the TMA expression data and survival time.

3.7 Discussion

In TMA data analysis, statistical methods often focus on downstream models in predicting disease outcome assuming $\bar{X}_i$ is a sufficient expression summary measure. Relatively little attention has been given to the modeling of within-tumor variation in these TMA experiments. As we have shown in this paper with real data examples, analysis ignoring intra-tumor variation can lead to false negative results which are tremendous wastes of valuable tissue resource and experimental costs. In this study, we proposed both two-stage and joint analysis methods to analyze tissue microarray data for bias correction. Adjusting for covariates Z_i and the number of repeated measures (K_i, r_i) can further improve the efficiency of the expression estimates. Both simulation and the case studies show that our methods outperform the common approach in estimating the prognostic value of a biomarker.

The proposed error model assumes constant variance across all subjects. To test the validity of this assumption, we performed the Levene's test for homogeneity of variance. The resulting test P-value is 0.003 and 0.04 for AMACR and BM28 respectively. There is some evidence suggesting a violation of the constant variance assumption especially for the AMACR data set. It may be of interest to consider a heteroscedastic model.

Since the initial development of TMAs, there have been many technical improvements over the years. Recent advances in quantitative assessment of the immunohistochemical staining provide precise, objective, and reproducible protein expression measurements. Compared to the conventional pathologist scoring on an ordinal scale, the Chromavision system used in our data examples enables quantification of the antigen level on a continuous scale, free of the subjectivity associated with

pathologist-based visual scoring system. AQUA (Camp et al., 2002), which stands for Automated Quantitative Analysis, is an academic system that measures fluorescence signals, leading to higher sensitivity to very low antibody concentrations. In addition, it allows the separation of tumor from stromal elements and the sub-cellular localization of signals for a co-localization of the antigens in different cell compartments. As the technology is becoming widely applied for cancer biomarker studies, robust statistical analysis methods underpinning both biological and experimental issues need to be established.

Table 3.1: Simulation study. Results are summarized over 100 simulated datasets each of $n = 200$.

		$\hat{\beta}^*$	$\hat{se}(\beta^*)$	$sd(\hat{\beta}^*)$	coverage	$\hat{\beta}^*$	$\hat{se}(\beta^*)$	$sd(\hat{\beta}^*)$	coverage
		Weibull				Proportional Hazards			
$\beta^* = 1$	X^*	1.01	0.08	0.08	0.96	1.01	0.08	0.08	0.97
	Naive	0.75	0.07	0.08	0.10	0.75	0.07	0.08	0.09
	LEI^{eb}	0.93	0.08	0.09	0.85	0.93	0.08	0.09	0.87
	LEI^{fb}	0.93	0.08	0.09	0.85	0.93	0.08	0.09	0.88
	LEI^{vrn}	0.97	0.09	0.10	0.89	0.96	0.09	0.10	0.87
	Joint Model	1.03	0.11	0.11	0.95	1.03	0.11	0.11	0.95
$\beta^* = 2$	X^*	2.05	0.12	0.12	0.95	2.03	0.11	0.10	0.94
	Naive	1.13	0.08	0.10	0	1.18	0.08	0.11	0
	LEI^{eb}	1.51	0.11	0.13	0.03	1.58	0.11	0.13	0.06
	LEI^{fb}	1.48	0.11	0.12	0	1.55	0.10	0.12	0.04
	LEI^{vrn}	1.62	0.11	0.14	0.19	1.70	0.11	0.15	0.30
	Joint Model	2.16	0.27	0.30	0.89	2.07	0.19	0.20	0.93

Table 3.2: A case study using prostate cancer TMA datasets.

		$\hat{\beta}^*$	$\hat{se}(\beta^*)$	$\hat{\beta}^*$	$\hat{se}(\beta^*)$
		Weibull		Proportional hazards	
	Naive	-0.573	0.201	-0.571	0.198
AMACR	LEI^{eb}	-0.761	0.266	-0.751	0.266
(n=203)	LEI^{fb†}	-0.753	0.262	-0.752	0.263
	LEI^{vrn}	-0.735	0.260	-0.737	0.258
	Joint modeling†	-0.742	0.268	-0.745$^{\flat}$	0.268
		$\hat{\beta}^*$	$\hat{se}(\beta^*)$	$\hat{\beta}^*$	$\hat{se}(\beta^*)$
		Weibull		Proportional Hazards	
	Naive	0.828	0.414	0.900	0.399
BM28	LEI^{eb}	1.051	0.519	1.146	0.494
(n=52)	LEI^{fb†}	1.457	0.564	1.493	0.504
	LEI^{vrn}	1.034	0.552	1.151	0.537
	Joint Modeling†	1.753	0.766	1.592$^{\flat}$	0.600

† - Estimation is based on MCMC methods, where we sampled 2 chains each with 10000 burn-in and 10000 updates.
$^{\flat}$We used $J = 5$ and $J = 3$ intervals for the piecewise exponential distribution in the AMACR and BM28 dataset respectively. The estimate of the Weibull shape parameter is $\hat{\gamma} = 0.71$, and $\hat{\gamma} = 0.97$ from the joint modeling for the AMACR and BM28 dataset respectively.

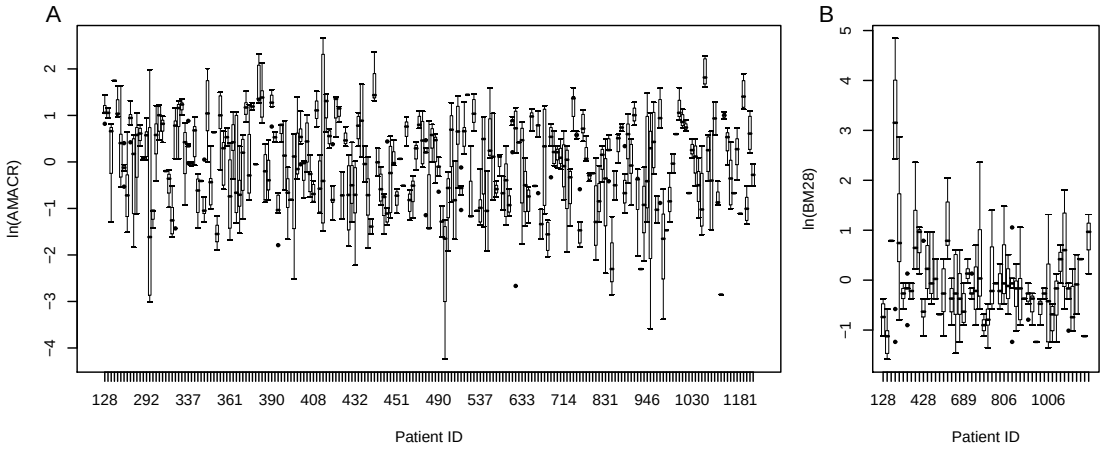


Figure 3.1: Variance plots to represent the within-subject variation in the TMA core-level expression data. A) The AMACR data. Estimates of the variance components are: $\hat{\sigma}_u^2 = 0.46$ and $\hat{\sigma}_{x^*}^2 = 0.54$. B) The BM28 data. Estimates of the variance components are: $\hat{\sigma}_u^2 = 0.62$ and $\hat{\sigma}_{x^*}^2 = 0.21$.

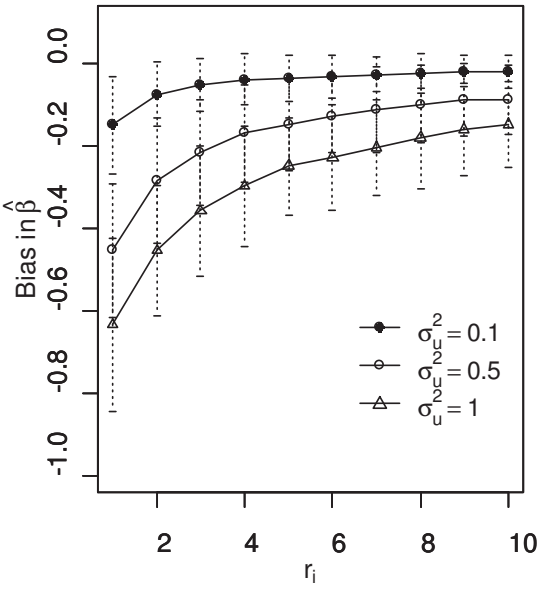
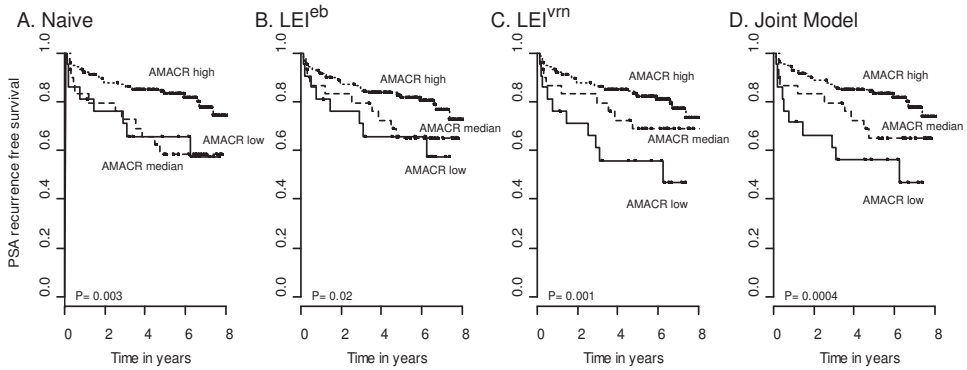
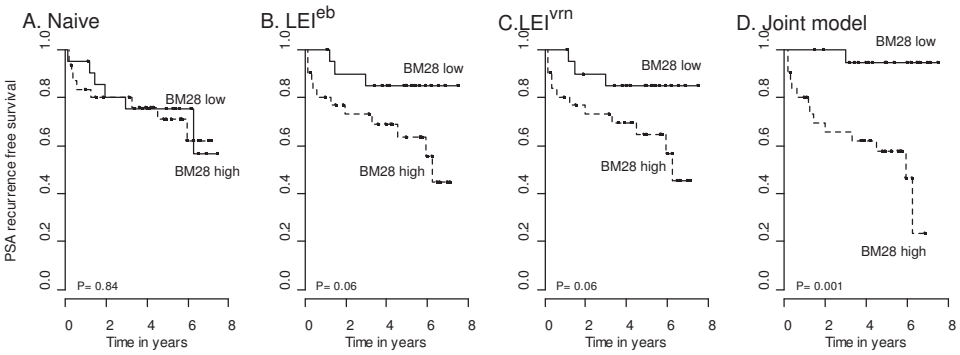


Figure 3.2: A simulation demonstration of the bias in Cox regression coefficient estimate as a function of the number of repeated measures r_i. The average bias with a 95% CI over 100 simulated datasets of sample size n=200 is plotted.



(a) AMACR. The 10th and 25th expression quantile was used to divide the 203 patient into three risk groups.



(b) BM28. The median of the expression estimates was used to divide the 52 patients into two risk groups.

Figure 3.3: Kaplan-Meier plots of prostate cancer recurrence. Patients are categorized into risk groups based on the protein expression level of (a) AMACR and (b) BM28 profiled using TMAs. The expression estimates are based on the A. Naive B. LEI^{eb} C. LEI^{vrn} and D. Joint model.

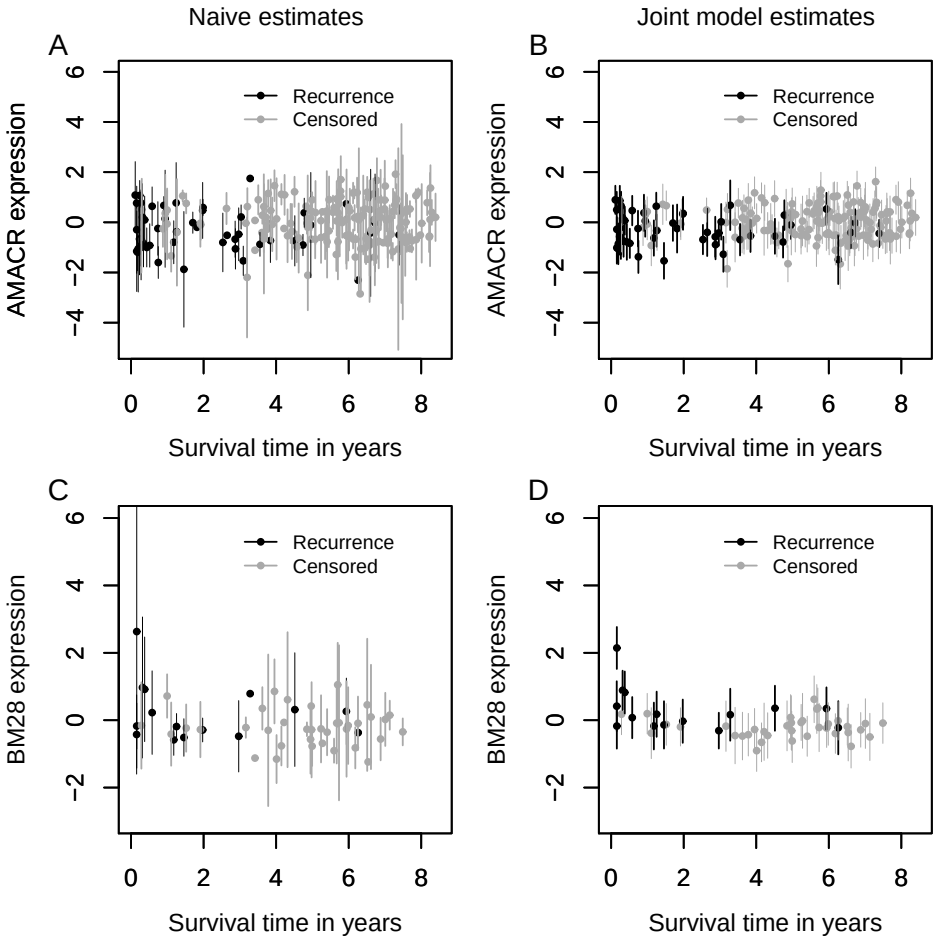


Figure 3.4: Comparison of the naive expression estimates (A, C) and the joint model expression estimates (B, D). The top panel depicts the comparison in the AMACR data, the bottom panel depicts the comparison in the BM28 data. The survival times are plotted on the x-axis.

CHAPTER IV

RECONSTRUCTING TUMOR-WISE PROTEIN EXPRESSION IN TISSUE MICROARRAY STUDIES USING A CELL MIXTURE MODEL

4.1 Introduction

In Chapter III, I have presented methods for analyzing protein expression data generated from Tissue Microarrays (TMAs). I focus on estimating the tumor means of a biomarker adjusting for intra-tumor variability. In particular, a measurement error model assuming normality is used to model the TMA core-level data with repeated measurements. The number of repeated measures per tumor and clinical/pathological covariates are incorporated to improve the precision of the expression estimates. A joint model relating the error model with patient survival information is used to estimate recurrence risks. Through the study, the intensity of the staining has been considered the relevant expression measure and specifically modeled using measurement error approaches. A normal distribution assumption on the log-transformed intensity measures is found to be sufficient for analysis purposes. Nevertheless when a more heterogeneous staining pattern is encountered — a mixture pattern composed of areas of non-staining (intensity equals zero) and areas of positive staining, it is hard to model the protein expression using any standard

distributional assumptions.

The motivation of this chapter is therefore to generalize the expression model for reconstructing complex staining patterns. For this purpose, I introduce the concept of a Cell Mixture Model (CMM). As illustrated in Figure 4.1, the basic idea can be decomposed into the following aspects. 1) A tumor is represented by a population of R_i cores (the total sampling capacity of a tumor); 2) The expression values in each individual core is a mixture distribution with a point mass at zero (the non-staining area); 3) The whole-tumor expression can be recapitulated by adding up (e.g., weighted summation) the distributions of the expression values in all the cores. The mathematical description will be put forward in Section 4.2.

There are difficulties of implementing the CMM model in TMA expression data. First, the experimental data are only collected on a small number (r_i out of R_i) of random samples of cores. Generally speaking, the number of measured cores r_i often averages from 3-5 whereas R_i can be in the hundreds, though both may vary proportionate to the size of the tumor. Second, each core is a very small sub-area measured in millimeters compare to the whole tumor averaging around 1-2 centimeters (prostate tumors). When our interest is to obtain accurate estimates for tumor- and core-level expression characteristics, sample-based methods will not be satisfactory. An analogy is estimating the characteristics of the population in the United States with data collected in three representative cities. In survey sampling problems, small area estimation often involves parameter estimation for small sub-population of interest. Hierarchical Bayes (HB) and Empirical Bayes (EB) approaches have been effective with continuous data. For a thorough review of various methods, see Ghosh (1994), Pfreffermann (2002), Rao (1999). For a unified analysis of discrete and continuous data, Ghosh et al. (1998) present hierarchical Bayes generalized lin-

ear models. The idea of Bayesian predictive inference and Markov Chain Monte Carlo integration technique is particularly useful for our problem at hand. In this study we extend the implementation to a zero-point mass mixture distribution under the CMM model. Details of constructing the CMM expression estimators will be discussed in Section 4.5.

Associating tumor-wise expression features with patient survival information is of scientific interest in TMA studies. The prognostic value of a potential biomarker is tested. Therefore accurate estimation of the disease risk associated with a biomarker is essential. To achieve this, a joint modeling approach would be most effective where the expression data and the survival data are simultaneously modeled. Markov Chain Monte Carlo methods offer a convenient framework for complex problems where analytic solutions are often unavailable or cumbersome. As will be discussed in detail in Section 4.6, linking the CMM model on the expression data with survival requires an imputation step within each MCMC iteration where draws are obtained from posterior predictive distributions.

This chapter is organized as follows. Section 4.2 and 4.3 introduce the concept along with the basic notation of the CMM model. In Section 4.4, a hierarchical Zero-Augmented Gamma model is imposed for the quantitative expression measures from tissue microarray experiments. Section 4.5 describes the construction of CMM estimators based on a Bayesian imputation strategy and Monte Carlo integrations. Section 4.6 extends the CMM model to jointly analyze TMA expression data and patient survival outcome. Simulation studies are carried out in Section 4.7, and case studies using two prostate cancer TMA data sets follow in Section 4.8. Conclusions and further discussion can be found in Section 4.9.

4.2 Notation and the Model

Figure 4.1 describes the concept of the cell mixture model. The cartoon illustrates a tumor being dissected into a population of R_i tissue core samples. Each core j ($j = 1, \cdots, R_i$) captures a sample of cells stained at different intensities. Let $a_{ij}(x)$ denote the number of cells measured at staining intensity $x, x \in [0, M]$ in core j of tumor i. Thus the density function can be expressed as $g_{ij}(x) = a_{ij}(x)/n_{ij}$, where n_{ij} is the total number of cells in core j of tumor i. The total number of cells measured is $N_i \equiv \sum_{l=1}^{R_i} n_{ij}$. In Figure 4.1, each histogram is informative of g_{ij}, which is assumed to be a mixture density with a point mass at zero for the non-staining area and some density function $f(\cdot)$ for the positively stained area. In particular,

$$g_{ij}(x) = (1 - \pi_{ij})I(x = 0) + \pi_{ij} f(x|\mu_{ij}, \sigma_{ij}^2)I(x > 0), \tag{4.1}$$

where π_{ij} denotes the proportion of staining; μ_{ij}, σ_{ij} are mean and variance parameters associated with the density f. Subsequently, the tumor-wise density function $g_i(x)$ is aggregated over all the $g_{ij}(x)'s$:

$$g_i(x) = \sum_{j=1}^{R_i} \omega_{ij} g_{ij}(x), \tag{4.2}$$

where $\omega_{ij} = n_{ij}/N_i$ and $\sum_{l=1}^{R_i} \omega_{ij} = 1$.

4.3 Description of the data

The tumor sampling scheme in TMA experiments has a 'geographical' clustered sampling structure. Consider each tumor is a population of cells. Small areas of 0.6mm (cores) are taken from the tumor where cells within each area are measured for protein expression. Let X_{ijk} be the resulting intensity measure in tumor i ($i = 1, \cdots, m$), core j ($j = 1, \cdots, r_i$), and cell k ($k = 1, \cdots, n_{ij}$). It needs to be pointed

out that X_{ijk} is an idealized measure where measurements can be taken per cell. The current technology instead provides a crude mean intensity measure for cells that have non-zero intensity

$$Y_{ij} = \sum_{k=1}^{n_{ij}} X_{ijk} I(X_{ijk} > 0)/n_{1ij}$$

per core. As illustrated in Figure 4.2, Y_{ij} is the actual observed data whereas the cell-level data are latent. The empirical estimate of μ_{ij} is y_{ij}. For the zero-mass part, we observe the number of positively staining cells and the number of non-staining cells which are

$$n_{1ij} = \sum_k I(X_{ijk} > 0), \quad n_{0ij} = \sum_k I(X_{ijk} = 0)$$

respectively. And $n_{ij} = n_{1ij} + n_{0ij}$ will be the total number of cells measured in tumor i core j. The empirical estimate of π_{ij} is n_{1ij}/n_{ij}.

4.4 A hierarchical Zero-Augmented Gamma model

In this section, we introduce a Zero-Augmented Gamma (hZAG) model for the observed data.

4.4.1 Modeling the positive staining intensity

We start by assuming $X_{ijk}|X_{ijk} > 0$ follow a Gamma distribution $G\big(1/\delta, \delta\mu_{ij}\big)$ with mean μ_{ij}, variance $\delta\mu_{ij}^2$, and the coefficient of variation being $1/\sqrt{\delta}$. In our application, we set $\delta = 0.2$. The observed Y_{ij} subsequently adopts a Gamma distribution with standardized shape and scale parameters. A Gamma-Inverse Gamma-Normal hierarchical model is set up as follows:

$$
\begin{aligned}
Y_{ij} &\overset{\text{ind}}{\sim} \text{Gamma}\Big(\frac{n_{1ij}}{\delta}, \frac{\delta}{n_{1ij}}\mu_{ij}\Big),\ i = 1, ..., n; j = 1, ..., r_i, \\
\mu_{ij} &\overset{\text{iid}}{\sim} \text{Inverse Gamma}\Big(\frac{1}{\nu} + 2, \frac{\nu+1}{\nu} e^{a_{0i} + \mathbf{a}\mathbf{z}_i'}\Big), \\
a_{0i} &\overset{\text{iid}}{\sim} \text{Normal}(0, \tau_a^2).
\end{aligned} \tag{4.3}
$$

In this model, $\{\mu_{i1}, \cdots, \mu_{ir_i}\}$ denotes the vector of core-level random effects for subject i and $\{a_{01}, \cdots, a_{0n}\}$ denotes the vector of subject-level random effects. Given the Gamma-Inverse Gamma conjugacy, the marginal densities integrated over μ_{ij} has the following analytic form:

$$
f(y_{ij}|a_{0i}, \mathbf{a}, \mathbf{z}_i) = \frac{\Gamma(\frac{n_{1ij}}{\delta} + \frac{1}{\nu} + 2)}{\Gamma(\frac{n_{1ij}}{\delta})\Gamma(\frac{1}{\nu} + 2)} \times \frac{(\frac{\nu+1}{\nu} e^{a_{0i} + a\mathbf{z}_i'})^{\frac{1}{\nu}+2} y_{ij}^{\frac{n_{1ij}}{\delta}-1}}{(\frac{\delta}{n_{1ij}})^{\frac{n_{1ij}}{\delta}} (\frac{n_{1ij}}{\delta} y_{ij} + \frac{\nu+1}{\nu} e^{a_{0i} + a\mathbf{z}_i'})^{\frac{n_{1ij}}{\delta} + \frac{1}{\nu} + 2}}, \tag{4.4}
$$

where $\mathbf{z}_i$ is a vector of tumor-level covariates and $\mathbf{a}$ is the associated coefficients.

4.4.2 Modeling the point mass at Zero

To model the point mass at zero in the mixture density of (4.1), we assume the following hierarchical structure:

$$
\begin{aligned}
n_{1ij} &\sim Bin(n_{ij}, \pi_{ij}), \\
logit(\pi_{ij}) &= b_{0i} + \mathbf{b}\mathbf{z}_i' + \epsilon_{ij},
\end{aligned} \tag{4.5}
$$

where $b_{0i} \sim N(0, \tau_b^2)$, $\epsilon_{ij} \sim N(0, \sigma_b^2)$, and $\mathbf{z}_i$ can be the same or different than those included in (4.3). Let $b_{0ij} = logit(\pi_{ij})$ such that $\pi_{ij} = \exp(b_{0ij})/(1 + \exp(b_{0ij}))$.

The core- and subject-level parameter spaces are

$$
\boldsymbol{\Theta}_{ij} = \{\mu_{ij}, b_{0ij}\}, \qquad \boldsymbol{\Theta}_i = \{a_{0i}, \mathbf{a}, \tau_a^2, \nu, b_{0i}, \mathbf{b}, \sigma_b^2, \tau_b^2\}
$$

respectively (as illustrated in Figure 4.2). The likelihood function treating the latent

quantities as parameters can be written as:

$$
\begin{aligned}
L_{cmm} \propto & \left\{ \prod_{i=1}^{n} \prod_{j=1}^{r_i} \left(\frac{1}{1+e^{b_{0ij}}} \right)^{n_{0ij}} \left(\frac{e^{b_{0ij}}}{1+e^{b_{0ij}}} \right)^{n_{1ij}} \right\} \\
& \times \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{r_i} \left(\frac{b_{0ij} - b_{0i} - \mathbf{b}\mathbf{z}_i'}{\sigma_b} \right)^2 \right\} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \left(\frac{b_{0i}}{\tau_b} \right)^2 \right\} \\
& \times \left\{ \prod_{i=1}^{n} \prod_{j=1}^{r_i} IG_{\mu_{ij}} \left(\frac{n_{1ij}}{\delta} + \frac{1}{\nu} + 2, \frac{n_{1ij}}{\delta} y_{ij} + (\frac{1}{\nu} + 1) e^{a_{0i} + \mathbf{a}\mathbf{z}_i'} \right) \right\} e^{-\frac{1}{2}\sum_{i=1}^{n} \left(\frac{a_{0i}}{\tau_a} \right)^2}.
\end{aligned} \tag{4.6}
$$

To complete the hierarchy for the Bayesian model, the following prior distributions are specified as:

$$
\begin{aligned}
& a_k \sim N(\mu_{a_k}, \sigma_{a_k}^2), \quad \tau_a^{-2} \sim Gamma(r_{\tau_a^2}, \gamma_{\tau_a^2}), \quad \nu \sim Gamma(r_\nu, \gamma_\nu); \\
& b_k \sim N(\mu_{b_k}, \sigma_{b_k}^2), \quad \sigma_b^{-2} \sim Gamma(r_{\sigma_b^2}, \gamma_{\sigma_b^2}), \quad \tau_b^{-2} \sim Gamma(r_{\tau_b^2}, \gamma_{\tau_b^2}).
\end{aligned} \tag{4.7}
$$

Posterior inference will then be based on the joint posterior distribution $f(\mathbf{\Theta}_{ij}, \mathbf{\Theta}_i | \mathbf{D})$. Gibbs sampling is used to iteratively sample from the full conditionals of each parameter given the rest of the parameters and the data.

4.5 Estimation of tumor-wise expression characteristics

In this section, I focus on estimating the tumor-wise protein expression characteristics. Three quantities are of interest: the tumor-wise proportion of staining (π_i), mean intensity of staining (μ_i^+), and a composite intensity (μ_i). Under the proposed cell mixture model assumptions, these quantities are defined as

$$
\pi_i = \sum_{j=1}^{R_i} \omega_{ij} \pi_{ij}, \quad \mu_i^+ = \sum_{j=1}^{R_i} \omega_{ij} \mu_{ij}, \quad \mu_i = \sum_{j=1}^{R_i} \omega_{ij} \pi_{ij} \mu_{ij}, \tag{4.8}
$$

respectively. Here $\pi_{ij} = \exp(b_{0ij})/(1+\exp(b_{0ij}))$. For the rest of the Chapter, I use η_i as a general notation for the above expression characteristics.

Assume independence among the cores and, without loss of generality, assume the first r_i cores from the ith tumor are observed and the rest of the cores are not

observed, we decompose η_i as

$$\begin{aligned}
\pi_i &= \sum_{j=1}^{r_i} \omega_{ij}\pi_{ij} + \sum_{j=r_i+1}^{R_i} \omega_{ij}\pi_{ij}^m, \\
\mu_i^+ &= \sum_{j=1}^{r_i} \omega_{ij}\mu_{ij} + \sum_{j=r_i+1}^{R_i} \omega_{ij}\mu_{ij}^m, \\
\mu_i &= \sum_{j=1}^{r_i} \omega_{ij}\pi_{ij}\mu_{ij} + \sum_{j=r_i+1}^{R_i} \omega_{ij}\pi_{ij}^m\mu_{ij}^m,
\end{aligned} \tag{4.9}$$

where the first components of the expansion are estimable given the data $\mathbf{D} = \big(y_{ij}, n_{1ij}, n_{ij} : i = 1, \cdots, n; j = 1, \cdots, r_i\big)$, and the second components involve latent quantities Θ_{ij}^m where data are not observed for core j $(j = r_i + 1, \cdots, R_i)$.

4.5.1 The CMM model-based estimator

To obtain a CMM model-based estimate of η_i, I propose the following in a Bayesian framework. (1) The first component of (4.9) is computed based on a set of draws $\mathbf{\Theta}_{ij}^{(g)} = \{b_{0ij}^{(g)}, \mu_{ij}^{(g)} : g = 1, \cdots, G\}$ from the posterior density $f(\mathbf{\Theta}_{ij}|\mathbf{\Theta}_i, \mathbf{D})$ for $j = 1, \cdots, r_i$. The posterior means $\tilde{\pi}_{ij} = G^{-1}\sum_g \exp(b_{0ij}^{(g)})/(1 + \exp(b_{0ij}^{(g)}))$; $\tilde{\mu}_{ij}^+ = G^{-1}\sum_g \mu_{ij}^{(g)}$, and $\tilde{\mu}_{ij} = G^{-1}\sum_g \exp(b_{0ij}^{(g)})/(1 + \exp(b_{0ij}^{(g)}))\mu_{ij}^{(g)}$ are then readily obtained from the posterior samples. (2) Let $\mathbf{\Theta}_{ij}^m \equiv (b_{0ij}^m, \mu_{ij}^m)$— the parameter vector involved in the second component of (4.9). In the absence of knowledge about $\mathbf{\Theta}_{ij}^m$, we replace the latent quantities with their expectation $E[\mathbf{\Theta}_{ij}^m|\mathbf{D}]$. To calculate this, we need the posterior predictive density function

$$p(\mathbf{\Theta}_{ij}^m|\mathbf{D}) = \int p(\mathbf{\Theta}_{ij}^m|\mathbf{\Theta}_i, \mathbf{D}) f(\mathbf{\Theta}_i|\mathbf{D}) d\mathbf{\Theta}_i.$$

Using Monte Carlo integration technique, we first draw $\mathbf{\Theta}_i$ from their joint posterior distribution $f(\mathbf{\Theta}_i|\mathbf{D})$ and then simulate $\mathbf{\Theta}_{ij}^m$ according to (4.3) and (4.5). Let $\{\mathbf{\Theta}_{ij}^{(p)} : p = 1, \cdots, P\}$ be the set of predictive draws at each of the G MCMC iterations. The

following quantities can then be computed:

$$\tilde{E}[\pi_{ij}^m|\mathbf{D}] = \frac{1}{G}\sum_{g=1}^{G}\frac{1}{P}\sum_{p=1}^{P}\left[\frac{\exp(b_{0ij}^{(p)})}{1+\exp(b_{0ij}^{(p)})}\Big|\boldsymbol{\Theta}_i^{(g)}\right]. \tag{4.10}$$

Similarly, we simulate a set of $\{\mu_{ij}^{(p)}, m = 1, \cdots, P\}$, given $\tilde{\boldsymbol{\Theta}}_i^{(g)}$, for $g = 1, \cdots, G$, using (4.3) and obtain

$$\tilde{E}[\mu_{ij}^m|\mathbf{D}] = \frac{1}{G}\sum_{g=1}^{G}\frac{1}{P}\sum_{p=1}^{P}\left[\mu_{ij}^{(p)}\Big|\boldsymbol{\Theta}_i^{(g)}\right]. \tag{4.11}$$

Finally, the composite mean is computed as

$$\tilde{E}[\pi_{ij}^m\mu_{ij}^m|\mathbf{D}] = \frac{1}{G}\sum_{g=1}^{G}\frac{1}{P}\sum_{p=1}^{P}\left[\frac{\exp(b_{0ij}^{(p)})}{1+\exp(b_{0ij}^{(p)})}\mu_{ij}^{(p)}\Big|\boldsymbol{\Theta}_i^{(g)}\right]. \tag{4.12}$$

These are essentially imputation steps within each MCMC iteration. Assuming equal weights $\omega_{ij} \equiv 1/R_i$, the CMM estimates of are

$$\begin{aligned}
\tilde{\pi}_i^{cmm} &= \frac{1}{R_i}\left\{\sum_{j=1}^{r_i}\tilde{\pi}_{ij} + \sum_{j=r_i+1}^{R_i}\tilde{E}[\pi_{ij}^m|\mathbf{D}]\right\}, \\
\tilde{\mu}_i^{+cmm} &= \frac{1}{R_i}\left\{\sum_{j=1}^{r_i}\tilde{\mu}_{ij} + \sum_{j=r_i+1}^{R_i}\tilde{E}[\mu_{ij}^m|\mathbf{D}]\right\}, \\
\tilde{\mu}_i^{cmm} &= \frac{1}{R_i}\left\{\sum_{j=1}^{r_i}\tilde{\pi}_{ij}\tilde{\mu}_{ij} + \sum_{j=r_i+1}^{R_i}\tilde{E}[\pi_{ij}^m\mu_{ij}^m|\mathbf{D}]\right\}.
\end{aligned} \tag{4.13}$$

Since $R_i \gg r_i$, (4.13) is dominated by the second component.

4.5.2 Sample-based estimators

The sample-based estimates are derived as:

$$\hat{\pi}_i^s = \frac{\sum_{j=1}^{r_i} n_{1ij}}{\sum_{j=1}^{r_i} n_{ij}}, \qquad \hat{\mu}_i^{+s} = \frac{\sum_{j=1}^{r_i} n_{ij}y_{ij}}{\sum_{j=1}^{r_i} n_{ij}}, \qquad \hat{\mu}_i^{s} = \frac{\sum_{j=1}^{r_i} n_{1ij}y_{ij}}{\sum_{j=1}^{r_i} n_{ij}}. \tag{4.14}$$

These sample-based estimates are implied by the proposed model by setting $\sigma_b^2 = 0$ in (4.5) and $\nu = 0$ in (4.3) such that homogeneity is assumed across cores within a tumor. These estimates are unbiased when the sample cores have the same characteristics as the tumor.

4.6 Joint analysis with patient survival outcome

Associating the expression characteristics to patient survival data is of major interest in many TMA studies. A joint modeling approach would be the most effective way to obtain accurate estimates of disease risks associated with a biomarker. To extend the CMM model into a joint model with censored failure time data, we use a piecewise constant hazards model in which the time axis is partitioned into L disjoint intervals, $I_1, ..., I_L$, where $I_l = [a_{l-1}, a_l)$ with $a_0 < t_i$ and $a_L > t_i$ for all $i = 1, ..., n$. Assume a constant baseline hazard in the lth interval, $\lambda_0(t) = \lambda_l$ for $t \in I_l$. R_l is the set at risk at the beginning of interval l; d_l is the number of failures in interval l; and $\Delta_{il} = min(t_i, a_l) - a_{l-1}$. Further by treating the latent variables b_{0ij}, μ_{ij} as a set of parameters in a Bayesian framework, the joint likelihood function is given as

$$
\begin{aligned}
L_{Joint} \propto & \left\{ \prod_{i=1}^{n} \prod_{j=1}^{r_i} \left(\frac{1}{1+e^{b_{0ij}}} \right)^{n_{0ij}} \left(\frac{e^{b_{0ij}}}{1+e^{b_{0ij}}} \right)^{n_{1ij}} \right\} \\
& \times \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{r_i} \left(\frac{b_{0ij} - b_{0i} - \mathbf{b}\mathbf{z}_i'}{\sigma_b} \right)^2 \right\} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \left(\frac{b_{0i}}{\tau_b} \right)^2 \right\} \\
& \times \left\{ \prod_{i=1}^{n} \prod_{j=1}^{r_i} IG_{\mu_{ij}} \Big(\frac{n_{1ij}}{\delta} + \frac{1}{\nu} + 2, \frac{n_{1ij}}{\delta} y_{ij} + (\frac{1}{\nu} + 1) \exp\{a_{0i} + \mathbf{a}\mathbf{z}_i'\} \Big) \right\} \\
& \times \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \left(\frac{a_{0i}}{\tau_a} \right)^2 \right\} \\
& \times \prod_{l=1}^{L} \lambda_l^{d_l} \exp\Big(\sum_{i \in D_l} \beta\eta_i + \kappa \mathbf{z}_i' \Big) \exp\Big(-\lambda_l \sum_{i \in R_l} \Delta_{il} e^{\beta\eta_i + \kappa \mathbf{z}_i'} \Big),
\end{aligned}
\tag{4.15}
$$

where $\mathbf{\Theta}_{ij} = (b_{0ij}, \mu_{ij})$. The following priors in addition to those specified in (4.7) are chosen:

$$
\lambda_l \sim Gamma(r_{\lambda_l}, \gamma_{\lambda_l}), \beta \sim N(\mu_\beta, \sigma_\beta^2), \kappa_j \sim N(\mu_{\kappa_j}.\sigma_{\kappa_j}^2). \tag{4.16}
$$

The parameter spaces are expanded to:

$$
\mathbf{\Theta}_{ij} = \{\mu_{ij}, b_{0ij}\}, \quad \mathbf{\Theta}_i = \{a_{0i}, a, \tau_a^2, \nu, b_{0i}, b, \sigma_b^2, \tau_b^2\}, \quad \mathbf{\Omega}_i = \{\lambda_l : l = 1, ..., L, \beta, \kappa\},
$$

The full conditional of β is given by

(4.17)
$$\beta|\cdot \propto \exp\left\{\beta\sum_{i\in D_l}\eta_i + \kappa\mathbf{z}_i' - \sum_{l=1}^{L}\lambda_l\sum_{i\in R_l}\Delta_{il}\exp(\beta\eta_i+\kappa\mathbf{z}_i')\right\}\exp\left\{\frac{1}{2}\left(\frac{\beta-\mu_\beta}{\sigma_\beta}\right)^2\right\},$$

where at the gth MCMC iteration, computation of η_i involves predictive draws and Monte Carlo integration as discussed in the previous section. The details of the MCMC implementation can be found in Appendix B.

4.7 Simulation study

4.7.1 Simulation setup

In the simulation study, we assign parameter values in the simulation to mimic those for the real data sets. In particular, the parameter values under the hZAG model are specified as follows: $\tau_a^2 = 0.01, \sigma_b^2 = 1, \tau_b^2 = 1$. The model has one covariate Z_{1i} simulated from $N(0,1)$ with associated model coefficient $a_1 = 0.5, b_1 = 0.5$. For each tumor, r_i is simulated from Binomial$(10, 0.5)$. Simulation of R_i, the total sampling capacity of a tumor, is relatively subjective as no information is available. We simulate R_i from a Binomial$(200, p_i)$ where p_i is allowed to vary with covariates such as tumor size. The survival time T_i is simulated from a proportional hazards model in the following form

(4.18)
$$\lambda(t) = \lambda_0(t)e^{\beta\eta_i+\kappa_1 z_{1i}},$$

with $\lambda_0(t) \equiv 1$. The censoring time is simulated from an independent exponential distribution that results in a 30% censoring proportion.

Parameter initialization is set up as follows. For $\mathbf{\Theta}_i^0$, crude estimates from fitting a generalized linear mixed model using a Penalized Quasi-Likelihood approach (Breslow and Clayton, 1993) are used. For the glmm fit, we use log- and logit- link

respectively for the intensity and zero-mass model with Gamma and Binomial family distribution. Next, to initialize the core-level parameters $\mathbf{\Theta}_{ij}$, we use the empirical estimates. Specifically, we set $\mu_{ij}^0 = y_{ij}, b_{0ij}^0 = \exp(n_{1ij}/n_{ij})/(1 + \exp(n_{1ij}/n_{ij}))$. For the piecewise exponential model, a total of $L = 4$ intervals were chosen such that each interval contains approximately equal number of events. We set λ_l^0 to be the empirical estimates of the event rate within each interval. Samples from the joint posterior distribution is obtained by Gibbs sampling (Gelfand and Smith, 1990, Geman and Geman, 1984). The full conditional density functions are specified in Appendix B. Noninformative proper priors are chosen. All programming is done using the R programming language. For simulations, we discard the first 1000 samples as the burn-in period. Every 10th sample is then retained to achieve a total of 1000 samples. Convergence is monitored using traceplots. Each simulation consisted of 100 replicate data, each of $n = 100$ subjects. Results are summarized over replicated data sets.

Finally, we point out that in the intensity model, the marginal mean and variance of Y_{ij} has analytic forms. A combination of penalized quasi-likelihood estimation and BLUP estimation (termed PQL-BLUP) is therefore applicable. The details of this estimation procedure can be found in Appendix C. We found the PQL-BLUP estimates were similar to those obtained from the Bayesian estimation.

4.7.2 Simulation results

Model misspecification

Figure 4.3 illustrates the sensitivity to misspecified δ. We fit the CMM model with $\delta = 0.2$ whereas the true simulated value is 0.5. A, B, and C draws the fitted (dotted line) and true (solid line) density of X_{ijk} over the simulated latent data. Although

the shape of distribution is somewhat sensitive to the value of δ, the density of the observed measure $f(y_{ij}|\boldsymbol{\Theta}_i)$ is not (D). Next, we simulate Y_{ij} from a log-normal density function. Figure 4.4 reveals no serious model misspecification problems.

Accuracy of the expression estimates

To evaluate the accuracy of the sample-based estimates and the proposed CMM model-based estimates in approximating the true expression quantity, the Mean Squared Error (MSE), Absolute Relative Error (ARE), and Relative Difference (RD) are computed as follows

$$\begin{aligned} \text{MSE} &= \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{\eta}_i - \eta_i\Big)^2, \\ \text{ARE} &= \frac{1}{n}\sum_{i=1}^{n}\Big|\frac{\hat{\eta}_i - \eta_i}{\eta_i}\Big|, \\ \text{RD} &= \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{\hat{\eta}_i - \eta_i}{\eta_i}\Big). \end{aligned} \tag{4.19}$$

Table 4.1 lists the mean of these statistics (taking median results in similar comparison) over 100 simulated data sets. We chose different values for ν and σ_b^2 to control the amount of within-subject variation observed in Y_{ij} and n_{1ij} respectively. When the within-subject variance is relatively small, the difference between the sample-based and the CMM estimates is not apparent. However, the amount of decrease in MSE and ARE by the CMM estimator is incremental as the within-subject variation gets larger. No significant bias is observed by examining the RDs in Table 4.1.

Joint modeling with survival

The interest in this section is to estimate the Cox regression coefficient β in (4.18). Three approaches are compared. A naive method where the sample-based expression estimates are plugged in a Cox model; a two-stage CMM method where the CMM

estimates are plugged in the Cox model; and the joint modeling approach based on the joint likelihood (4.15). The first two methods are considered two-stage methods as compared to the joint model. The two-stage methods have several major limitations. First, the survival information is not used in the CMM model to reconstruct tumor expression, which can cause bias and efficiency loss in estimating β in the second stage. Second, the uncertainty of estimating the expression quantity is not assimilated in the second stage, leading to over-optimistic standard error estimates of $\hat{\beta}$. The joint modeling approach concurrently updates the CMM model and the survival model by iteratively sampling through the joint posterior distribution of the combined parameter space. We therefore expect more accurate inference from the joint model. In Table 4.2, the top panel simulates $\beta_{\pi_i} = 2, \beta_{\mu_i^+} = 0, \beta_{\mu_i} = 0$, the middle panel assumes $\beta_{\pi_i} = 0, \beta_{\mu_i^+} = 2.5, \beta_{\mu_i} = 0$, and the bottom panel assumes $\beta_{\pi_i} = 0, \beta_{\mu_i^+} = 0, \beta_{\mu_i} = 1.8$. It is evident that the joint model performs best in terms of the estimates and coverage probabilities for $\hat{\beta}$. Furthermore, we examined the effect of misspecified value for δ on the joint model estimates of β (results are shown under Joint model* in Table 4.2). Specifically, δ is fixed to be 0.2 in the estimation procedure while the true value in the simulated data set is 0.5. Such misspecification has led to only small differences in estimating β in the joint model compared to the correctly specified model. A possible explanation lies in the fact that the influence of δ diminishes because of the standardization by n_{ij} (which is a large number) in (4.3).

4.8 Case study using prostate Cancer Tissue Microarray Experiments

4.8.1 Data description

The same prostate cancer TMA data sets used in Chapter III apply here. Details of the data description can be found in the previous Chapter. Gleason score and pathologic stage are included as the clinical covariates $\mathbf{Z}_i$. A batch effect is added to the AMACR data set, as evident in Figure 4.5, the staining intensity distribution is bimodal. In Rubin et al. (2005), an array-wise normalization was performed to eliminate the batch effect resulting from experiment-to-experiment variation of immunohistochemical staining. For the MCMC convergence of the joint model, we use the first 10,000 draws as burn-in, and retain every 20th draw till 1000 samples are collected for inference. The approximate computing time to fit the AMACR data set is 10 hours.

4.8.2 BM28 expression characteristics and patient survival

Figure 4.5 suggests that BM28 is a homogeneously stained marker. All of the 52 tumors showed over 94% staining. We therefore focus on analyzing the intensity of the staining of this gene biomarker.

The top panel of Table 4.3 describes the performance of Cox regression models relating the estimated mean intensity of BM28 to PSA-recurrence adjusting for Gleason score and Pathological stage of the tumor. Among the two stage estimation procedures of β, the CMM estimator of μ_i^+ does not perform better than the sample-based estimator. It is likely that the CMM estimates in the data set does not approximate the true expression quantity significantly better than would the sample-based estimates when ν is small ($\hat{\nu} = 0.006$). As we have shown in the simulation study, the

MSE and ARE differences are not discernable when the within-subject variation is not too large. The joint model estimate is however more than two times larger than those under the two-stage estimation. The estimated hazard ratio under the joint model is 4.4 (95% CI:1.6-11.7) compared to 1.9 (95% CI: 1.2-3.0) estimated under two-stage methods. However, a hypothesis test of $H_0 : \beta = 0$ would give similar conclusions as the estimated standard error from the joint model is also substantially larger than those from the two-stage estimation. After controlling for Gleason and Pathological stage of the disease, the mean intensity of BM28 staining in the tumor is a significant predictor of prostate cancer PSA-recurrence. A further notion is that these results are consistent with those observed under the measurement error model in Chapter III. The underlying Gamma-Inverse-Gamma assumption on the intensity measure versus the log-normal assumption adopted previously does not seem to have large influence on estimating the Cox regression coefficient β in the joint model.

4.8.3 AMACR expression characteristics and patient survival

Table 4.3 summarizes the results from analyzing the AMACR data set. A distinct feature is the relationship between the expression characteristics. The composite mean, μ_i, resembles (though not strictly equivalent to) an interaction term of the proportion (π_i) and the mean intensity (μ_i) of the staining fitted in the same model adjusting for the Gleason score of the tumor and the stage of the disease. The associated coefficient $\hat{\beta}_{\mu_i}$ is significant in both the two-stage CMM model and the joint model. Another evidence of interaction is that when fitted individually, $\hat{\beta}_{\pi_i}$ and $\hat{\beta}_{\mu_i}$ are close to zero (data not presented).

A second observation is that when comparing the naive model and the joint model fitted in the AMACR data set, the regression coefficients are biased in different

directions. Results in Table 4.3 suggest that β_{π_i} and β_{μ_i} are underestimated while $\beta_{\mu_i^+}$ is overestimate. It should be pointed out that in this data set, we observe a positive correlation between the proportion of staining and the intensity of staining. The correlation coefficient between the sample-based estimates, $\hat{\pi}_i^s$ and $\hat{\mu}_i^s$, is 0.80 and 0.53 for batch 1 and 2 respectively. When multiple error-prone covariates are concerned in a regression model, the direction and magnitude of the bias can also depend on the correlation between the predictor variables (Carroll et al., 1995).

Figure 4.9 reveals the complexity of AMACR protein expression as a predictor of PSA recurrence outcome. Each of the three expression estimates are dichotomized into two risk groups using the lower quartile as cutoff, resulting in a total of eight combinations (though one group has 0 observations). Overall, the joint model (plot B) demonstrates better differentiation of recurrence risks than the naive model (plot A). In both figures, tumors demonstrating low staining proportion, low intensity, and low composite intensity (curve 1) has the highest recurrence risk of all. One significant difference between A and B lies in curves 3 and 4. The joint model has generated substantially different estimates of the recurrence risks for these two groups compared to sample-based methods.

4.9 Discussion

A cell mixture model is proposed to reconstruct tumor expression characteristics from tissue microarray data. The concept is to assemble the whole-tumor expression pattern from the subpopulation of tissue cores. We let each individual core density adopt a zero-augmented Gamma density function to describe the proportion of non-staining and the intensity of the positive staining respectively. A main difficulty is model estimation. One wishes to obtain accurate estimates of both core- and

subject-level parameters with an average of three cores per tumor. A hierarchical Bayes model is therefore imposed to borrow strength across cores and across tumors. We find that the reconstructed expression features are relatively robust under model misspecification. Expression estimates under the CMM model have better accuracy than the sample-based estimates. A joint model is presented to link the CMM expression model with a survival model for censored failure time observations. The implementation involves imputation steps within each MCMC iteration and Monte Carlo integration technique. With the advent of modern computing power, complex models are feasible. Simulation studies show that the joint model can effectively reduce the attenuation of the disease risk estimates evident in two-stage methods. In addition, when interactions among the expression features exist, relating noise-inflated expression estimates to survival can lead to misleading results. Applying the joint model effectively avoids erroneous interpretations of the risk estimates.

In this study, we estimated the percentage of staining, the mean intensity of staining, and a composite mean staining of a tumor from its reconstructed expression distribution. These expression characteristics are further associated with censored survival time to estimate recurrence risks in prostate tumors. In fact, exploring other expression characteristics is possible given the reconstructed distribution under the CMM model. For example, in addition to the mean, lower (e.g. 10^{th}) or upper (e.g. 90^{th}) percentile of expression may be a relevant quantity to summarize the tumor expression for biological reasons. In many TMA studies, with an average of three cores observed for a tumor, the sample minimum ($Y_{i(r_i)} = \min_{1\leq j\leq r_i} y_{ij}$) and the sample maximum ($Y_{i(1)} = \min_{1\leq j\leq r_i} y_{ij}$) of staining are often used in place of a specific quantile of expression. These sample-based statistics apparently target at a tail quantity of the distribution of Y_{ij}, but their behavior is not clear. The sample maximum of

one tumor may map to the 90^{th} percentile of $f(Y_{ij})$, while the sample maximum of another tumor may map to an entirely different percentile. In this respect, the CMM model will allow more comparable and precise estimates of the expression quantiles, while the Bayesian framework will provide a straightforward implementation where any percentiles of the posterior samples can be readily obtained.

In our current model, we assume the proportion of staining and the intensity of the staining are independently distributed given covariates $\mathbf{Z}_i$. To relax such conditional independence assumption made for the CMM model, a random effect can be added to induce correlation between the zero-mass and the intensity model. Such extension is useful when a biomarker has inherently correlated expression pattern whereas the nature of such correlation may be unknown and therefore difficult to adjust for with the known covariates.

Table 4.1: Accuracy of the expression estimates.

	$\nu = 0.01$						$\nu = 0.1$					
	$\sigma_b^2 = 0.5$			$\sigma_b^2 = 1$			$\sigma_b^2 = 0.5$			$\sigma_b^2 = 1$		
	MSE	ARE	RD	MSE	ARE	RD	MSE	ARE	RD	MSE	ARE	RD
$^*\hat{\pi}_i^s$	0.004	0.079	-0.0002	0.018	0.17	-0.018	0.0041	0.079	0.0025	0.017	0.17	-0.007
$^\dagger\tilde{\pi}_i^{cmm}$	0.004	0.086	0.019	0.011	0.15	0.019	0.0038	0.085	0.017	0.010	0.14	0.035
$\hat{\mu}_i^s$	11.9	0.042	0.0031	9.9	0.038	0.001	101.6	0.12	-0.016	115.6	0.12	0.003
$\tilde{\mu}_i^{cmm}$	9.4	0.036	0.0039	7.1	0.034	0.001	35.1	0.07	-0.0067	31.3	0.069	0.007
$\hat{\mu}_i^s$	25.3	0.088	0.0035	91.3	0.18	-0.015	66.4	0.14	-0.016	114.7	0.21	-0.005
$\tilde{\mu}_i^{cmm}$	25.2	0.094	0.0232	51.6	0.15	0.025	36.4	0.11	0.0099	61.9	0.16	0.04

*Sample-based estimates, †CMM model-based estimates.
MSE — Mean Squared Error.
ARE — Absolute Relative Error.
RD — Relative Difference.

Table 4.2: Cox regression. Results are summarized over 100 simulated data sets each of $n = 100$. The CMM model parameter values are simulated to be the same as in Table 4.1.

	true β	$\hat{\beta}$	$sd(\hat{\beta})$	$\hat{se}(\beta)$	coverage
π_i	2	2.06	0.24	0.23	0.97
$\hat{\pi}_i^s$		1.48	0.23	0.18	0.27
$\tilde{\pi}_i^{cmm}$(2stg)		1.60	0.22	0.22	0.53
Joint model		2.06	0.32	0.40	0.97
Joint model*		2.07	0.30	0.36	0.97
μ_i^+	2.5	2.50	0.30	0.26	0.93
$\hat{\mu}_i^{+s}$		1.43	0.25	0.23	0.39
$\tilde{\mu}_i^{+cmm}$(2stg)		2.07	0.27	0.23	0.44
Joint model		2.48	0.55	0.49	0.94
Joint model*		2.55	0.51	0.52	0.94
μ_i	1.8	1.82	0.21	0.20	0.95
$\hat{\mu}_i^s$		1.40	0.18	0.16	0.48
$\tilde{\mu}_i^{cmm}$ (2stg)		1.68	0.16	0.19	0.79
Joint model		1.75	0.41	0.47	0.95
Joint model*		1.73	0.40	0.44	0.95

*Joint model under misspecified δ.

Table 4.3: Case study using prostate cancer TMA data sets. Prediction of patient PSA-recurrence using tumor-wise protein expression estimates.

BM28 (n=52)						
	Sample-based		CMM (2stg)		Joint model	
	$\hat{\beta}$	$\hat{se}(\beta)$	$\hat{\beta}$	$\hat{se}(\beta)$	$\hat{\beta}$	$\hat{se}(\beta)$
μ_i^+	0.668	0.232	0.630	0.236	1.481	0.501
Gleason	0.666	0.601	0.683	0.561	0.592	0.558
Stage	0.938	0.507	0.837	0.535	0.822	0.501

AMACR (n=203)						
	Sample-based		CMM (2stg)		Joint model	
	$\hat{\beta}$	$\hat{se}(\beta)$	$\hat{\beta}$	$\hat{se}(\beta)$	$\hat{\beta}$	$\hat{se}(\beta)$
π_i	0.827	0.358	1.284	0.539	1.778	0.586
μ_i^+	-1.132	0.464	-0.554	0.402	-0.488	0.389
μ_i	-0.736	0.457	-1.008	0.458	-2.372	0.728
Gleason	1.237	0.418	1.177	0.42	1.025	0.513
Stage	1.345	0.298	1.254	0.298	1.276	0.293

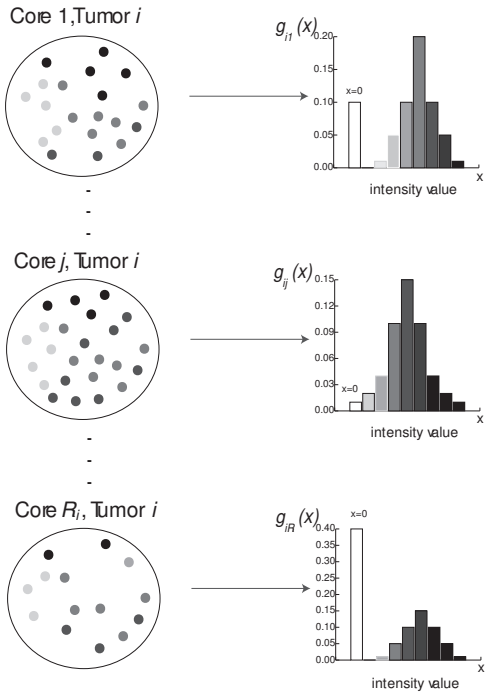


Figure 4.1: A conceptual model for the whole tumor. Each tumor i represented by a population of R_i cores.

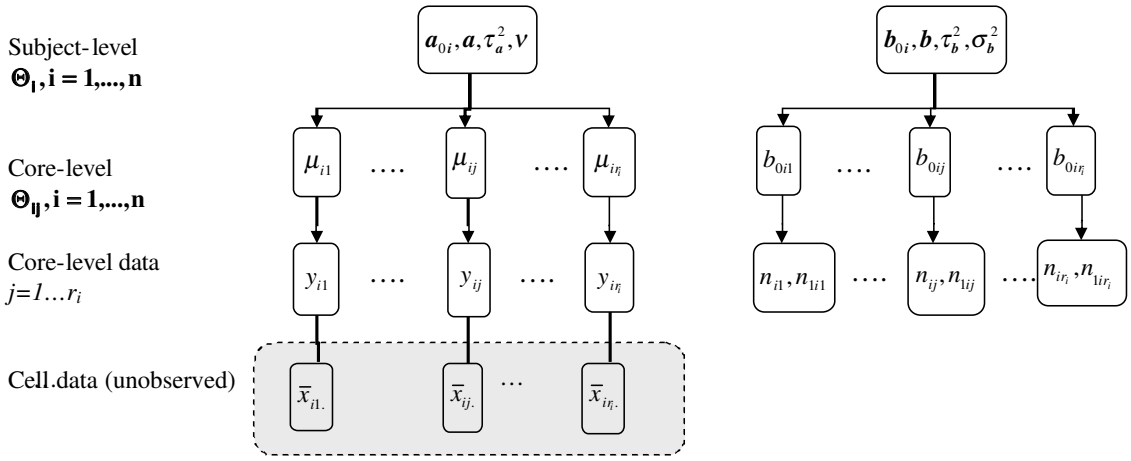


Figure 4.2: Model structure

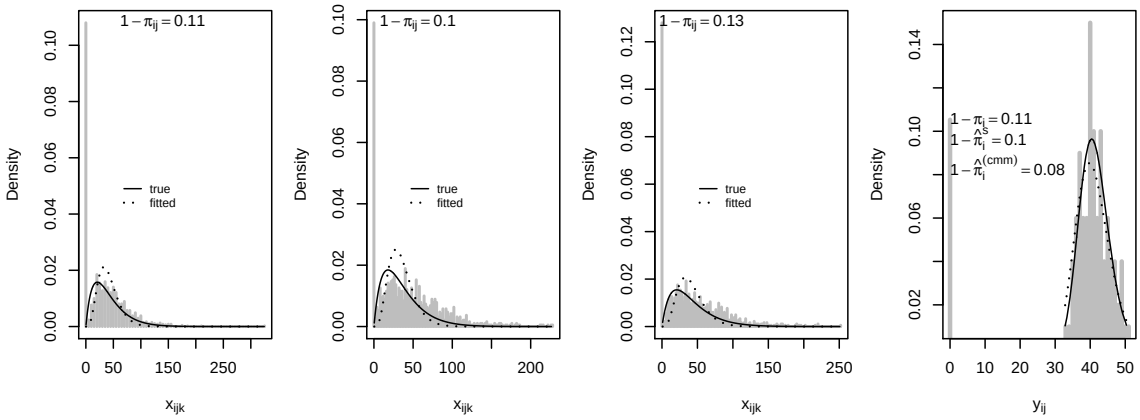


Figure 4.3: Sensitivity to misspecification of δ. (A, B, C) plots the fitted (dotted line, assuming $\delta = 0.2$) and the true (solid line, $\delta = 0.5$) density of X_{ijk} over the histogram of the latent data. (D) shows the fitted and true density curves of the observed data over the histogram of $Y_{ij}, j = 1, ..., R_i$.

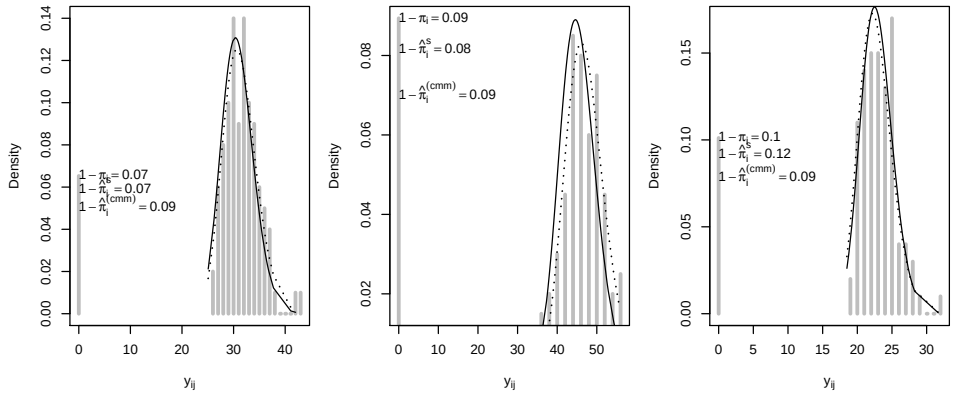


Figure 4.4: Model misspecification. Here Y_{ij} is simulated from a log-normal (LNN) distribution with mean $\mu_i = a_{0i} + az_i$ and variance $\sigma^2 = 0.01$. Solid lines are the LNN density curves. Dotted lines are the CMM fitted density curves.

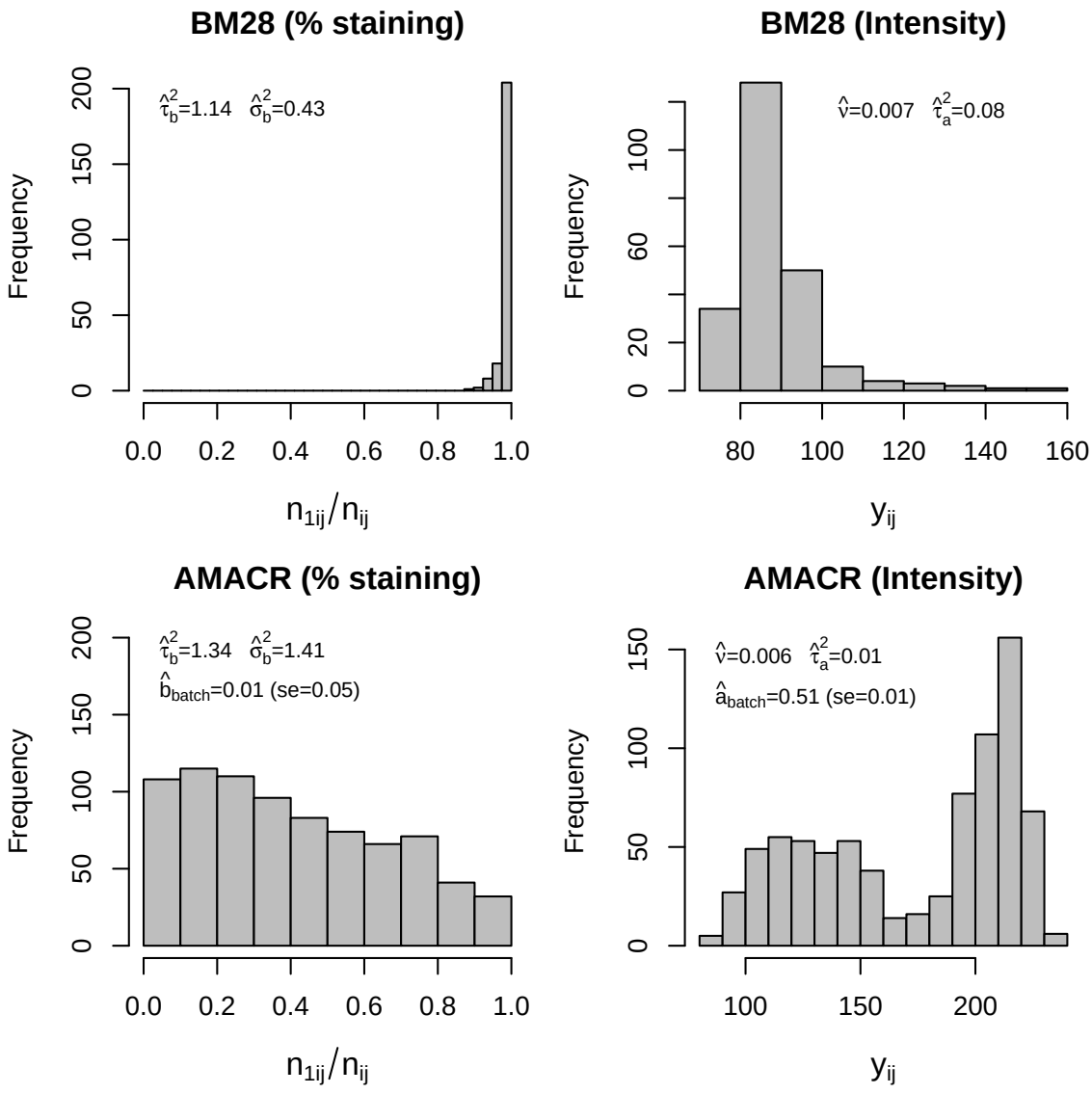
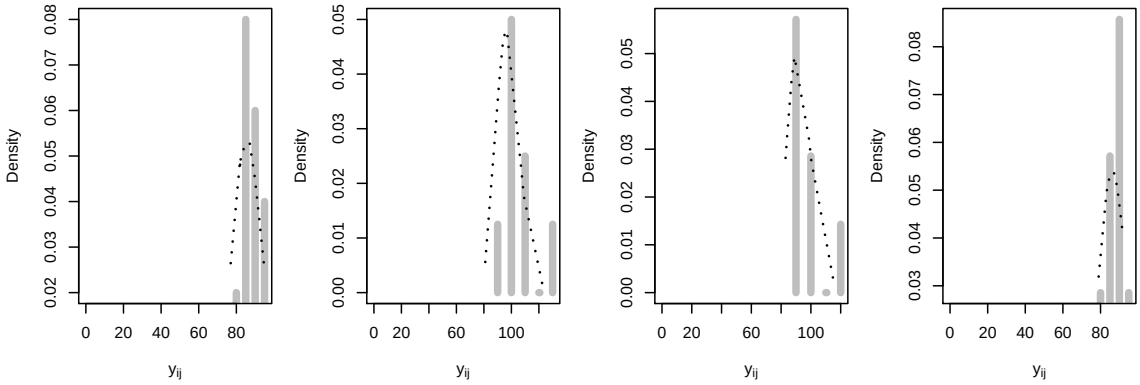
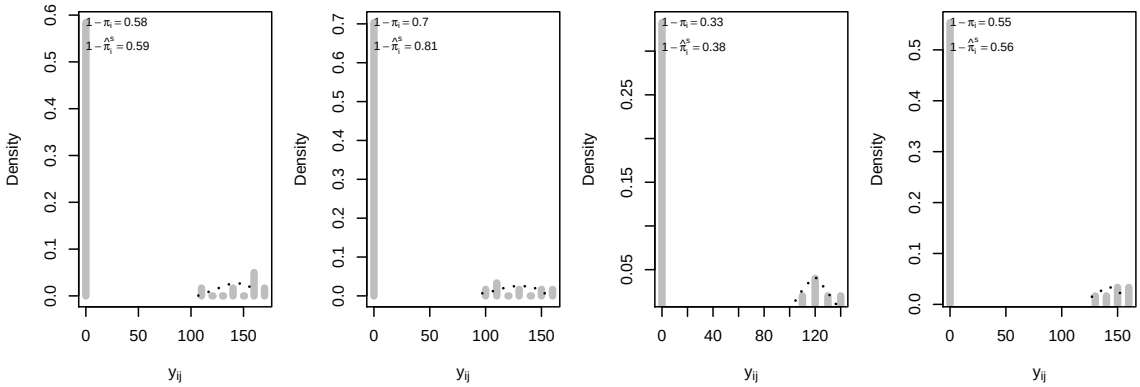


Figure 4.5: Histograms of the percentage of staining and the intensity of staining. The estimated variance parameters in the CMM model are indicated in the plots. For the AMACR data, the batch effect for the Gamma-Inverse-Gamma model is listed.



(a) BM28 protein expression in four different tumors



(b) AMACR protein expression in four different tumors

Figure 4.6: Reconstructed tumor expression under the CMM model.

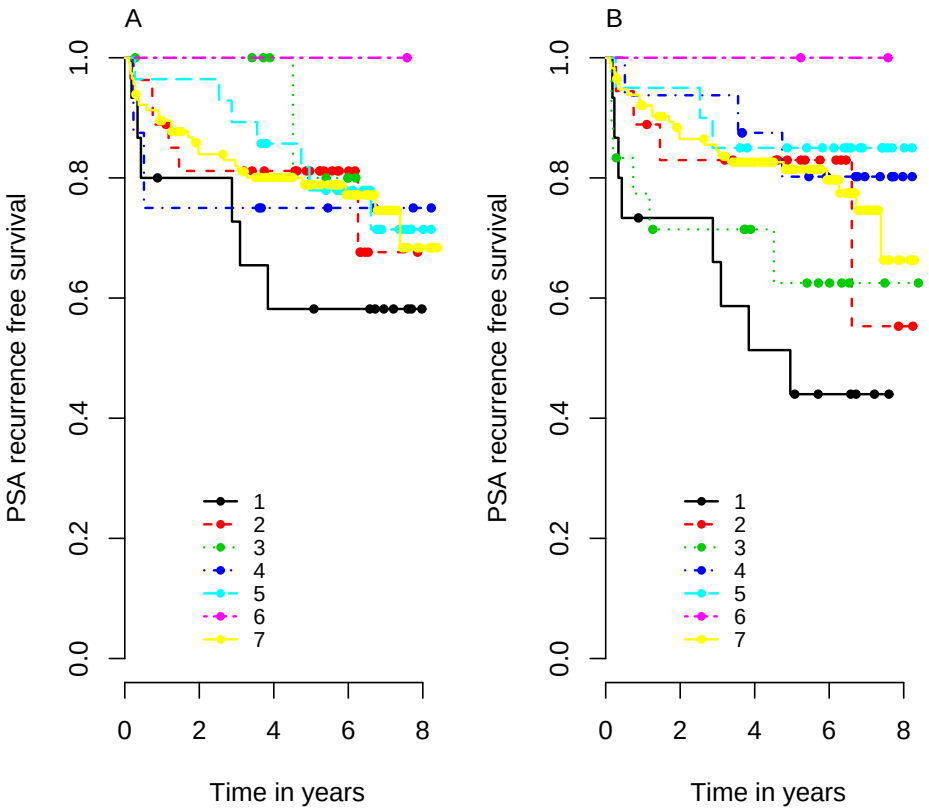


Figure 4.7: Kaplan-Meier plots. Patients are categorized into risk groups based on the protein expression estimates (A: Sample-based, B: Joint model). The lower quartiles are used for dichotomization. 1. low π_i, low μ_i^+, low μ_i, 2. low π_i, high μ_i^+, low μ_i, 3. low π_i, high μ_i^+, high μ_i, 4. high π_i, low μ_i^+, low μ_i, 5. high π_i, low μ_i^+, high μ_i, 6. high π_i, high μ_i^+, low μ_i, 7. high π_i, high μ_i^+, high μ_i.

CHAPTER V

CONCLUSION

The use of genomic and proteomic approaches has revolutionized cancer research in the past decade. Cutting-edge technologies such as DNA microarrays and Tissue Microarrays (TMAs) have provided high-throughput platforms for identifying and validating genome biomarkers for cancer diagnosis and prognosis. Developing meaningful and robust statistical methods is a key element for reliable and reproducible findings. In this dissertation, I address some of the statistical issues related to biomarker studies using microarray data.

In Chapter II, I presented a Bayesian mixture model approach to the meta-analysis of DNA microarrays. The estimated probability of differential expression, *poe*, is used as a unified scale to eliminate array platform differences. Data integration based on *poe* has several advantages in a meta-analysis context. One, integrated sample cohorts improve the reliability of the findings by guarding against false positive results from a single study. A meta-signature is therefore more likely to be validated in independent data sets. Two, data integration increases the statistical power to detect small consistent effects that can be otherwise masked by inadequacy of the sample size of an individual data set. The utility of such meta-analysis framework is broad with the increasing amount of publicly accessible microarray data.

Choi et al. (2007) (submitted manuscript) extended the application in prostate cancer to compare metastatic and localized disease across multiple microarray studies. Two softwares, POE (Parmigiani et al., 2002) and MetaArray (Choi et al., 2007), that implement MCMC methods are available to generate *poe* values from raw expression data. They can be downloaded from the R Bioconductor project (http://www.bioconductor.org).

Although the *poe* transformation eliminates the measurement scale differences across array platforms, additional steps can be taken to improve the reproducibility of the gene expression profile from experiment to experiment. For example, in the proposed meta-analysis strategy, compiling common genes across array platforms is an important step. Matching probes across cDNA arrays and Affymetrix arrays to ascertain they target at the same full-length mRNA transcript is hardly a straightforward task. Unigene ID is a common choice for cross-platform mapping, but several studies have found that Unigene ID alone is insufficient for matching and often lead to poor correlation between gene expression across platforms. Sequence matching based on RefSeq database can significantly improves the quality of matches and subsequently increase the cross-platform consistency and reproducibility (Ji et al., 2006, Mecham et al., 2005).

In Chapter III, I have addressed statistical issues in analyzing protein expression data from tissue microarray experiments. A Latent Expression Index (LEI) is introduced to adjust for 1) the intra-tumor variability, 2) the number of repeated measure per tumor, and 3) clinical covariates. As a validation tool, accurate estimation of the disease risk associated with a biomarker from TMA data is essential. A joint model is proposed for simultaneous inference on the expression data and patient survival information. Both simulation studies and data application have shown that

the joint model is an effective approach to eliminate the attenuation in the coefficient estimates caused by measurement error. In this study, our primary interest is parameter estimation in proportional hazards models with variables measured with error. The proposed joint model is useful in eliminating bias in estimating the Cox regression coefficient. However, it should be pointed out that such error model is not necessary when prediction of the outcome is concerned which is beyond the scope of this dissertation.

In Chapter IV, a Cell Mixture Model (CMM) is proposed to reconstruct complex tissue staining patterns in TMA experiments. The concept is to assemble the whole-tumor expression pattern by aggregating over the subpopulation of tissue cores. Each individual core is assumed to be a zero-augmented distribution to assimilate the non-staining areas and the staining areas. A hierarchical Bayes model is imposed to borrow strength across cores and across tumors. A joint model is presented to link the CMM expression model with a survival model for censored failure time observations. The implementation involves imputation steps within each MCMC iteration and Monte Carlo integration technique. Possible future work includes correlating the proportion of staining and the intensity of the staining in the CMM model.

In summary, the existing methods in TMA studies have two major limitations: 1) they generally treat the expression measures as error-free quantities and 2) there is a lack of a unified modeling approach to incorporate both staining proportion and intensity measure. The main contribution of Chapter III and IV is to provide methods for quantitative TMA data analysis that effectively deal with these statistical issues.

APPENDICES

APPENDIX A

FULL CONDITIONAL DISTRIBUTIONS FOR CHAPTER II

In the meta-analysis, the following gene-specific parameters were repeatedly drawn from the full conditional distributions as shown below

$$\kappa_j^{+(t+1)}|x_{ij} \sim Gamma(\sum_{i=1}^{M_k} p_{ij}^{+(t)} + 1, \theta_\kappa^+),$$

$$\kappa_j^{-(t+1)}|x_{ij} \sim Gamma(\sum_{i=1}^{M_k} p_{ij}^{-(t)} + 1, \theta_\kappa^-),$$

$$\sigma_j^{-2^{(t+1)}}|x_{ij} \sim Gamma\left(\gamma + \frac{r_j^{(t)}}{2}, \frac{1}{2}[s_j^{2^{(t)}} + \frac{M_k r_j^{(t)}}{M_k + r_j^{(t)}}(\xi - \bar{x}_j^{(t)})^2 + 2\lambda]\right),$$

$$\mu_j^{(t+1)}|x_{ij}, \sigma_j^{-2^{(t+1)}} \sim N\left(\frac{r_j^{(t)}\bar{x}_j^{(t)} + M_k\xi}{M_k + r_j^{(t)}}, \frac{\sigma_j^{2^{(t+1)}}}{M_k + r_j^{(t)}}\right),$$

$$(\pi_j^{+(t+1)}, \pi_j^{-(t+1)}, 1 - \pi_j^{+(t+1)} - \pi_j^{-(t+1)})$$

$$\sim Dirichlet\left(\sum_{i=1}^{M_k} p_{ij}^{+(t)} + 1, \sum_{i=1}^{M_k} p_{ij}^{-(t)} + 1, \sum_{i=1}^{M_k}(1 - p_{ij}^{+(t)} - p_{ij}^{-(t)}) + 1\right),$$

$$p_{ij}^{+(t)} = \frac{\pi_j^{+(t)} f_1(x_{ij}; \mu_j^{(t)}, \kappa_j^{+(t)})}{f(x_{ij}; \Theta_j^{(t)})} \qquad p_{ij}^{-(t)} = \frac{\pi_j^{-(t)} f_{-1}(x_{ij}; \mu_j^{(t)}, \kappa_j^{-(t)})}{f(x_{ij}; \Theta_j^{(t)})},$$

where $r_j^{(t)} = \sum_{i=1}^{M_k}(1 - p_{ij}^{+(t)} - p_{ij}^{-(t)})$, $\bar{x}_j^{(t)} = \sum_{i=1}^{M_k}(1 - p_{ij}^{+(t)} - p_{ij}^{-(t)})x_{ij}/r_j^{(t)}$, $s_j^{2^{(t)}} = \sum_{i=1}^{M_k}(1 - p_{ij}^{+(t)} - p_{ij}^{-(t)})(x_{ij} - \bar{x}_j^{(t)})^2$, and M_k is the sample size for study k; .

The derivation of these conditionals is fairly standard; see Diebolt and Robert (1994).

APPENDIX B

FULL CONDITIONAL DISTRIBUTIONS FOR CHAPTER IV

At each MCMC iteration, samples are successively drawn from the following full conditionals:

$$\text{(i)}\quad a_k\big|\cdot \propto \left\{\prod_{i=1}^{n}\prod_{j=1}^{r_i} IG_{\mu_{ij}}\Big(\frac{n_{1ij}}{\delta}+\frac{1}{\nu}+2, \frac{n_{1ij}}{\delta}y_{ij}+(\frac{1}{\nu}+1)e^{a_{0i}+\mathbf{az}_i'}\Big)\right\}$$
$$\times \exp\left\{-\frac{1}{2}\Big(\frac{a_k-\mu_{a_k}}{\sigma_{a_k}}\Big)^2\right\};$$
$$b_k\big|\cdot \sim N\left(V^{-1}\Big(\sigma^{-2}\mu_{b_k}+\sum_{i=1}^{n}\sum_{j=1}^{r_i}(b_{0ij}-b_{0i}-\mathbf{b}_{-k}\mathbf{z}'_{-k,i})z_{ki}\Big),\ \ V^{-1}\right)$$
$$\text{where}\quad V=\sigma_{b_k}^{-2}+\sigma_b^{-2}\sum_{i=1}^{n} r_i z_{ki}^2;$$

$$\text{(ii)}\quad a_{0i}\big|\cdot \propto \left\{\prod_{i=1}^{n}\prod_{j=1}^{r_i} IG_{\mu_{ij}}\Big(\frac{n_{1ij}}{\delta}+\frac{1}{\nu}+2, \frac{n_{1ij}}{\delta}y_{ij}+(\frac{1}{\nu}+1)e^{a_{0i}+\mathbf{az}_i'}\Big)\right\}$$
$$\times \exp\Big\{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{a_{0i}-a_0}{\tau_a}\right)^2\Big\};$$
$$b_{0i}\big|\cdot \propto N\left((\tau_b^{-2}+\sigma^{-2}r_i)^{-1}\Big(\tau_b^{-2}+\sigma^{-2}\sum_{j=1}^{r_i}(b_{0ij}-\mathbf{bz}_i')\Big),\ \ (\tau_b^{-2}+\sigma^{-2}r_i)^{-1}\right);$$

$$\text{(iii)}\quad \tau_a^{-2}\big|\cdot \propto G\left(\frac{1}{2}(r_{\tau_a^2}+n),\ \ \frac{1}{2}\Big(\sum_{i=1}^{n}(a_{0i}-a_0)^2+\gamma_{\tau_a^2}\Big)\right);$$
$$\tau_b^{-2}\big|\cdot \propto G\left(\frac{1}{2}(r_{\tau_b^2}+n),\ \ \frac{1}{2}\Big(\sum_{i=1}^{n}(b_{0i}-b_0)^2+\gamma_{\tau_b^2}\Big)\right);$$
$$\sigma_b^{-2}\big|\cdot \propto G\left(\frac{1}{2}(r_{\sigma_b^2}+n),\ \ \frac{1}{2}\Big(\sum_{i=1}^{n}\sum_{j=1}^{r_i}(b_{0ij}-b_{0i}-\mathbf{bz}_i')^2+\gamma_{\sigma_b^2}\Big)\right);$$

(iv) $\mu_{ij}\big|\cdot \propto \prod_{i=1}^{n}\prod_{j=1}^{r_i} IG_{\mu_{ij}}\Big(\frac{n_{1ij}}{\delta}+\frac{1}{\nu}+2, \frac{n_{1ij}}{\delta}y_{ij}+(\frac{1}{\nu}+1)e^{a_{0i}+\mathbf{a}\mathbf{z}_i'}\Big);$

$$b_{0ij}\big|\cdot \propto \left\{\prod_{i=1}^{n}\prod_{j=1}^{r_i}\left(\frac{1}{1+e^{b_{0ij}}}\right)^{n_{0ij}}\left(\frac{e^{b_{0ij}}}{1+e^{b_{0ij}}}\right)^{n_{1ij}}\right\}$$

$$\times \exp\left\{\sum_{i=1}^{n}\sum_{j=1}^{r_i}\left(\frac{b_{0ij}-b_{0i}-\mathbf{b}\mathbf{z}_i'}{\tau_{b_1}}\right)^2\right\};$$

(vi) $$\mu_{ij}^{p}\big|\cdot \sim IG\left(\frac{1}{\nu}+2, \frac{\nu+1}{\nu}\exp(a_{0i}+\mathbf{a}\mathbf{z}_i')\right);$$

$$b_{0ij}^{p}\big|\cdot \sim N\left(b_{0i}+\mathbf{b}\mathbf{z}_i', \tau_b^2\right);$$

(vii) $$\eta_i = \begin{cases} \frac{1}{P}\sum_{p=1}^{P}\frac{\exp(b_{0ij}^p)}{1+\exp(b_{0ij}^p)} & \text{Proportion of staining;} \\ \frac{1}{P}\sum_{p=1}^{P}\mu_{ij}^p & \text{Mean intensity;} \\ \frac{1}{P}\sum_{p=1}^{P}\frac{\exp(b_{0ij}^p)}{1+\exp(b_{0ij}^p)}\mu_{ij}^p & \text{Composite mean;} \end{cases}$$

(viii) $$\beta\big|\cdot \propto \exp\left\{\beta\sum_{i\in D_l}\eta_i+\kappa\mathbf{z}_i'-\sum_{l=1}^{L}\lambda_l\sum_{i\in R_l}\Delta_{il}\exp(\beta\eta_i+\kappa\mathbf{z}_i')\right\}$$

$$\times \exp\left\{\frac{1}{2}(\frac{\beta-\mu_\beta}{\sigma_\beta^2})^2\right\};$$

$$\lambda_l\big|\cdot \propto G\left(r_\lambda+d_l,\ \ \gamma_\lambda+\sum_{i\in R_l}\Delta_{il}\exp\left\{\beta\eta_i+\kappa\mathbf{z}_i'\right\}\right).$$

In the above G denotes a Gamma distribution; IG denotes an Inverse-Gamma distribution; $\mathbf{b}$ is a $1\times K$ row vector of coefficients.

APPENDIX C

PQL-BLUP ESTIMATION FOR THE INTENSITY MODEL IN CHAPTER IV

Given the conditional mean and variance

$$E[y_{ij}|a_{0i},\mathbf{a},\mathbf{z}] = e^{a_{0i}+a\mathbf{z}_i'}$$
$$\mathrm{Var}(y_{ij}|a_{0i},\mathbf{a},\mathbf{z}) = \left(\frac{\delta}{n_{1ij}}\nu+\nu+1\right)e^{2(a_{0i}+\mathbf{a}\mathbf{z}_i')},$$

we estimate $(\alpha_{0i},\mathbf{a})$ by a PQL approach Breslow and Clayton (1993) via maximizing the Laplace approximation of the penalized quasi-likelihood

$$\mathrm{pql} = \sum_{i=1}^{n}\sum_{l=1}^{r_i}\frac{(y_{ij}-e^{a_{0i}+\mathbf{a}\mathbf{z}_i})}{(\frac{\delta}{n_{1ij}}\nu+\nu+1)e^{2a_{0i}+2\mathbf{a}\mathbf{z}_i}} - \sum_{i=1}^{n}\frac{1}{\tau^2}a_{0i}^2 = 0.$$

Next, given $(\hat{\alpha}_{0i},\hat{\mathbf{a}})$, we obtain the BLUP estimates of μ_{ij} in the form

$$\hat{\mu}_{ij} = \frac{\frac{n_{1ij}}{\delta}y_{ij}+(\frac{1}{\nu}+1)e^{\hat{a}_{0i}+\hat{a}z_i}}{\frac{n_{1ij}}{\delta}+\frac{1}{\nu}+1}.$$

In the final step, we propose to estimate ν as

$$\hat{\nu} \;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{l=1}^{r_i}(\hat{\mu}_{ij}-e^{\hat{a}_{0i}+\hat{a}z_i})^2}{r_i e^{2\hat{a}_{0i}+2\hat{a}z_i}},$$

conditional on $(\hat{a}_{0i},\hat{\mathbf{a}},\hat{\mu}_{ij})$. In this estimation procedure, δ is fixed to be 0.2 as part of the model assumption.

BIBLIOGRAPHY

BIBLIOGRAPHY

D. C. Allred, J. M. Harvey, M. Berardo, and G. M. Clark. Prognostic and predictive factors in breast cancer by immunohistochemical analysis. *Modern Pathology*, 11:155–68, 1998.

T. A. Bismar, F. Demichelis, A. Riva, R. Kim, S. Varambally, and L. He. Defining aggressive prostate cancer using a 12-gene model. *Neoplasia*, 8:59–68, 2006.

N. Breslow and D. Clayton. Approximate inference in generalized linear mixed models. *Journal of American Statistical Association*, 88:9–25, 1993.

E. R. Brown and J. G. Ibrahim. A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics*, 59:221–28, 2003.

R. L. Camp, G. G. Chung, and D. L. Rimm. Automated subcellular localization and quantification of protein expression in tissue microarrays. *Nature Medicine*, 8:1323–28, 2002.

R. J. Carroll, D. Ruppert, and L. A. Stefanski. *Measurement error in nonlinear models.* Chapman and Hall, Boca Raton, Florida, 1995.

H. Choi, R. Shen, A. M. Chinnaiyan, and D. Ghosh. A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. *Manuscript submitted for publication.*, 2007.

J. K. Choi, U. Yu, S. Kim, and O. J. Yoo. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19:i84–i90, 2003.

S. M. Dhanasekaran, T. R. Barrette, D. Ghosh, R. Shah, S. Varambally, K. Kurachi, K. J. Pienta, M. A. Rubin, and A. M. Chinnaiyan. Delineation of prognostic biomarkers in prostate cancer. *Nature*, 412(6849):822–6, 2001.

J. Diebolt and C. P. Robert. Estimation of finite mixture distributions through Bayesian sampling. *Journal of Royal Statistical Society: Series B*, 56:363–375, 1994.

K. A. Divito, A. J. Berger, R. L. Camp, M. Dolled-Filhart, D. L. Rimm, and H. M. Kluger. Automated quantitative analysis of tissue microarrays reveals an association between high bcl-2 expression and improved outcome in melanoma. *Cancer Research*, 64:8773–7, 2004.

L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–8, 2005.

L. Ein-Dor, O. Zuk, and E. Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences*, 103(15): 5923–8, 2006.

R. Etzioni, S. Hawley, D. Billheimer, L. D. True, and B. Knudsen. Analyzing patterns of staining in immunohistochemical studies: application to a study of prostate cancer recurrence. *Cancer Epidemiology and Biomarkers Prevention*, 14:1040–6, 2005.

C. J. Faucett and D. C. Thomas. Simultaneously modeling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine*, 15:1663–85, 1996.

J. Friedman and N. Fisher. Bump hunting in high dimensional data. *Statistical Computing*, 9: 123–43, 1999.

A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of American Statistical Association*, 85:398–409, 1990.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images.. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–41, 1984.

D. Ghosh, T. Barrette, D. Rhodes, and A. M. Chinnaiyan. Statistical issues and procedures for meta-analysis of microarray data: a case study in prostate cancer. *Journal of Functional and Integrative Genomics*, 3, 2003.

J. N. K. Ghosh, M. Rao. Small area estimation: An appraisal. *Statistical Science*, 9:55–93, 1994.

M. Ghosh, K. Natarajan, T. W. Stroud, and B. P. Carlin. Generalized linear models for small-area estimation. *Journal of American Statistical Association*, 93:273–32, 1998.

X. Guo and B. Carlin. Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*, 58:1–9, 2004.

D. A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72:320–38, 1977.

R. Henderson, P. Diggle, and A. Dobson. Joint modeling of longitudinal measurements and event time data. *Biostatistics*, 1:465–80, 2000.

E. Huang, S. H. Cheng, H. Dressman, J. Pittman, M. H. Tsou, C. F. Horng, et al. Gene expression predictors of breast cancer outcomes. *Lancet*, 361:1590–6, 2003.

R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. Summaries of Affymetrix genechip probe level data. *Nucleic Acids Research*, 31:1–8, 2003a.

R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, , and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4, 2003b.

Y. Ji, K. Coombes, J. Zhang, S. Wen, J. Mitchell, L. Pusztai, W. F. Symmans, and J. Wang. Refseq refinements of unigene-based gene matching improve the correlation of expression measurements between two microarray platforms. *Applied Bioinformatics*, 5:89–98, 2006.

J. Kononen, L. Bubendorf, A. Kallioniemi, M. Barlund, P. Schraml, et al. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature Medicine*, 4:844–7, 1998.

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38:963–74, 1982.

M. LeBlanc and J. Crowley. Relative risk regression trees for censored survival data. *Biometrics*, 48:411–25, 1992.

C. Li and W. H. Wang. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, 98:31–6, 2001.

M. J. Lindstrom and D. M. Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46:673–87, 1990.

X. Liu, V. Minin, Y. Huang, D. B. Seligson, and S. Horvath. Statistical methods for analyzing tissue microarray data. *J Biopharm Stat.*, 14:671–85, 2004.

J. Luo, D. J. Duggan, Y. Chen, J. Sauvageot, C. M. Ewing, M. L. Bittner, J. M. Trent, and W. B. Isaacs. Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Research*, 61(12):4683–8, 2001.

N. Mah, A. Thelin, T. Lu, S. Nikolaus, Kühbacher T., and Y. Gurbuz. A comparison of oligonucleotide and cDNA-based microarray systems. *Physiological Genomics*, 16:361–70, 2004.

C. E. McCulloch. Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, 89:330–35, 1994.

B. H. Mecham, G. T. Klus, J. Strovel, M. Augustus, D. Byrne, P. Bozso, D. Z. Wetmore, T. J. Mariani, I. S. Kohane, and Z. Szallasi. Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acid Research*, 32:e74, 2005.

G. Parmigiani, E. S. Garrett, R. Anbazhagan, and E. Gabrielson. A statistical framework for expression-based molecular classification in cancer. *Journal of Royal Statistical Society: Series B*, 64:717–36, 2002.

D. Pfreffermann. Small area estimation-new developments and directions. *International Statistical Review*, 70:125–43, 2002.

M. D. Radmacher, L. M. McShane, and R. Simon. A paradigm for class prediction using gene expression profiles. *Journal of Computational Biology*, 9, 2002.

J. N. K. Rao. Some recent advances in model based small area estimation. *Survey Methodology*, 25:175–186, 1999.

D. R. Rhodes, T. R. Barrette, M. A. Rubin, D. Ghosh, and A. M. Chinnaiyan. Meta-analysis of microarrays: Interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Research*, 62:4427–33, 2002.

D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A. M. Chinnaiyan. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of National Academy of Sciences*, 101(25):9309–9314, 2004.

M. A. Rubin, T. A. Bismar, O. Andrén, L. Mucci, R. Kim, R. Shen, D. Ghosh, J.T. Wei, A. M. Chinnaiyan, H. Adami, P.W. Kantoff, and Johansson J. Decreased α-Methylacyl CoA racemase expression in localized prostate cancer is associated with an increased rate of biochemical recurrence and cancer-specific death. *Cancer Epidemiology and Biomarkers Prevention*, 14:1424–31, 2005.

R. Scharpf, E. S. Garrett, J. Hu, and G. Parmigiani. Statistical modeling and visualization of molecular profiles in cancer. *BioTechniques*, 34:S22–S29, 2003.

D. B. Seligson, S. Horvath, T. Shi, H. Yu, S. Tze, M. Grunstein, and S. K. Kurdistani. Global histone modification patterns predict risk of prostate cancer recurrence. *Nature*, 435:1262–6, 2005.

T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of National Academy of Sciences*, 98:10869–74, 2001.

C. Sotiriou, S. Y. Neo, L. M. McShane, E. L. Korn, P. M. Long, A. Jazaeri, et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of National Academy of Sciences*, 100:10393–8, 2003.

D. Spiegelhalter, A. Thomas, N. Best, , and D. Lunn. *WinBUGS 1.4 Manual.*, 2003. http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml.

J. D. Storey. Statistical significance for genomewide studies. *Journal of Royal Statistical Society: Series B*, 64:479–98, 2002.

M. G. Tadesse, J. G. Ibrahim, R. Gentleman, S. Chiaretti, J. Ritz, and R. Foa. Bayesian error-in-variable survival model for the analysis of genechip arrays. *Biometrics*, 61:488–97, 2005.

A. Thomas. *BRugs: An R interface to OpenBUGS. Version 1.0 User's Manual.*, 2004. http://mathstat.helsinki.fi/openbugs/.

O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17, 2001.

A. A. Tsiatis and M. Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14:809–34, 2004.

A. A. Tsiatis, V. DeGruttola, and M. S. Wulfsohn. Modeling the relationship of survival to longitudinal data measured with error. applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, 90:429, 1995.

J. W. Tukey. Tightening the clinical trial. *Controlled Clinical Trials*, 14, 1993.

M. Van de Vijver. Gene-expression profiling and the future of adjuvant therapy. *Oncologist*, 10 (Suppl 2):30–4, 2005.

M. J. van de Vijver, Y. D. He, L. J. van 't Veer, H. Dai, A. A. M. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347:1999–2009, 2002.

L. J. van't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–6, 2002.

J. Wang, K. R. Coombes, W. E. Highsmith, M. J. Keating, and L. V. Abruzzo. Differences in gene expression between b-cell chronic lymphocytic leukemia and normal b cells: a meta-analysis of three microarray studies. *Bioinformatics*, 20:3166–78, 2004.

Y. Wang and J. M. G. Taylor. Joint modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association*, 96: 895–905, 2001.

Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365:671–9, 2005.

J. B. Welsh, L. M. Sapinoso, A. I. Su, S. G. Kern, J. Wang-Rodriguez, C. A. Moskaluk, H. F. Jr Frierson, and G. M. Hampton. Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Research*, 61:5974–8, 2001.

G. Wright, B. Tan, A. Rosenwald, E. H. Hurt, A. Wiestner, and L. M. Staudt. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large b cell lymphoma. *Proceedings of National Academy of Sciences*, 100:9991–6, 2003.

M. S. Wulfsohn and A. A. Tsiatis. A joint modeling for survival and longitudinal data measured with error. *Biometrics*, 53:330–39, 1997.

J. Xu and S. L. Zeger. Joint analysis of longitudinal data comprising repeated measures and times to events. *Applied Statistics*, 50:375–87, 2001.

Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, , and T. P. Speed. Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30, 2002.