DEPARTMENT OF BIOSTATISTICS AND MEDICAL INFORMATICS
1300 University Avenue
University of Wisconsin-Madison
Madison, WI 53706

TECHNICAL REPORT NO. 227
July 1, 2012

Discrete mixture regression models for heterogenous time-to-event data: Cox assisted clustering

Kevin H. Eng
Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison

Bret M. Hanlon
Department of Statistics
University of Wisconsin-Madison

Abstract

Despite mounting evidence that genetic subgroups for the same clinical disease exist in cancer genomic studies, little attention has been given to exploring how this heterogeneity affects statistical model building. We describe a discrete mixture model that is able to identify and model unforseen heterogeneity in regression problems for time-to-event data. Our illustrations highlight the variety of effects specific to challenges in molecular signature development in ovarian cancer. Additionally, we propose and inspect an expectation maximization algorithm for obtaining maximum likelihood estimates in these models. Analyses exploring complex genetic spaces should find this class of models and algorithms useful.

Keywords: Cancer genomics, Heterogeneity, Mixture Model, Survival Analysis.

Discrete mixture regression models for heterogenous time-to-event data: Cox Assisted Clustering

Kevin H. Eng[1] and Bret M. Hanlon[2]

Department of Biostatistics and Medical Informatics[1]
Department of Statistics[2]
University of Wisconsin-Madison
1300 University Avenue
Madison, WI 53706
U.S.A.
email: eng@stat.wisc.edu

Abstract

Despite mounting evidence that genetic subgroups for the same clinical disease exist in cancer genomic studies, little attention has been given to exploring how this heterogeneity affects statistical model building. We describe a discrete mixture model that is able to identify and model unforseen heterogeneity in regression problems for time-to-event data. Our illustrations highlight the variety of effects specific to challenges in molecular signature development in ovarian cancer. Additionally, we propose and inspect an expectation maximization algorithm for obtaining maximum likelihood estimates in these models. Analyses exploring complex genetic spaces should find this class of models and algorithms useful.

Keywords: Cancer genomics, Heterogeneity, Mixture Model, Survival Analysis.

1 Introduction

In high-dimensional cancer genomic studies, unobserved heterogeneity obfuscates the effort to build accurate descriptive and predictive models of risk stratification. In ovarian cancer, for example, Vaughan et al. (2011) note that the set of patients with the same clinical disease really have distinct molecular diseases. With respect to inference, this implies that the same regression models may not be valid for every patient and further that it is unclear which patients should be considered together. Therefore, a major statistical task is to organize patients into previously unknown classes while simultaneously fitting their time-to-event models.

The examples throughout the article are taken from our ongoing analysis of the ovarian cancer project of The Cancer Genome Atlas (TCGA), which has the goal of cataloging all of the genomic alterations in cancer. For each patient, there is a tremendous amount and variety of data: 12,000 genes in expression arrays, 1 million SNP genotypes, exome and whole genome sequence, methylation of thousands of CpG islands, and the expression of microRNA. To paraphrase TCGA investigator Dr. Douglas Levine, "with current technology, we can survey all the changes that exist in the genome. The question, then, is which changes are the important ones" (Levine, 2012). We anticipate that in answering Dr. Levine's question, our statistical task will serve in an exploratory role to extract and characterize genetic subgroups.

Some statistical algorithms do attempt to address heterogeneity. When such factors are known, standard stratified Cox regression models are indicated. Recently, Lostritto et al. (2011) have proposed an iterative conditional classification tree model that assumes a common baseline hazard to learn the partitions. Continuous mixtures, that is, frailty models (Aalen, 1988) or, more coarsely, random effects models (O'Quigley and Stare, 2002) also model non-homogenous variation in the relative hazard. In all of these cases, strong restrictions are placed on the form of the baseline hazard.

In Section 6, we will show that treating heterogeneity naturally leads to interpretable risk stratification. For summarizing the impact of an expression signature, one often classifies patients and tries to show "well-separated survival curves." Unfortunately, as Na et al. (2009) review, the methods in use either awkwardly dichotomize continuous predictors (i.e., model selection in a Cox regression setting) or disconnectedly rely on unsupervised clustering. In that sense, our attempt to find genetic subgroups supervised by their survival times directly addresses these problems.

In this paper, we propose a discrete mixture regression model for handling potential heterogeneity in time-to-event data. Concretely, we assume that observations belong to unlabeled classes with class-specific

proportional hazards regression models relating their genetic covariates to survival time outcomes (Section 2). This conditional semi-parametric model leads to a surprising variety of model effects which we illustrate in Section 3. In Section 4 we describe an algorithm and the considerations for fitting the model. Simulations highlight the handling of censored data (Section 5) and a data analysis shows how our models can handle modeling a single pathway (Section 6). Our discussion (Section 7) again emphasizes the exploratory role that this analysis may adopt.

2 Methods

Let (Y_i, δ_i, x_i), $i = 1, \ldots, n$ be an independent, right-censored sample with regression covariates $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$. We will denote survival times, censoring indicators and covariate vectors as $Y = (Y_1, \ldots, Y_n)$, $\Delta = (\delta_1, \ldots, \delta_n)$, and $\boldsymbol{x} = (x_1, \ldots, x_n)$.

To account for heterogeneity, we propose that each patient arises from one of K latent classes with probability $\pi_k, k = 1, \ldots, K$, $\sum_k \pi_k = 1$. We assume Cox's proportional hazards (PH) model within each class k (Cox, 1972) so that the covariate vector x enters the model log-linearly via a class specific hazard: $\log h_k(t|x) = \log h_{0k}(t) + x'\beta_k$. In particular, recall that a right-censored observation following a PH model has the density

$$f_k(y, \delta|x) = \left[h_{0k}(y) \exp(x'\beta_k)\right]^{\delta} \exp\left[-H_{0k}(y) \exp(x'\beta_k)\right], \tag{1}$$

where $h_0(t)$ and $H_0(t)$ are baseline hazard and baseline cumulative hazard for the kth class. The mixture density may be written as

$$f(Y, \Delta|\boldsymbol{x}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(Y_i, \delta_i|x_i). \tag{2}$$

If we also observe the latent class $U = (U_1, U_2, \ldots, U_n)$, where $U_i \sim \text{Multinomial}(\pi)$, $U_i \in \{1, 2, \ldots, K\}$ and $u_{ik} = 1_{\{U_i = k\}}$, we may write the density of the complete data as

$$f(Y, \Delta|\boldsymbol{x}, U) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\pi_k f_k(Y_i, \delta_i|x_i)\right]^{u_{ik}} \tag{3}$$

To estimate the regression coefficients and baseline hazard parameters, we propose maximizing this likelihood via the EM procedure described in Section 4.

Because this discrete mixture incorporates latent random variables, it is related to continuous frailty models (Aalen, 1988) and random effects models (O'Quigley and Stare, 2002). Along with stratification, these models still make inflexible assumptions on the regression function: heterogenous variation might be absorbed in a log-linear shift in baseline hazard but the regression function remains the same.

Additionally, the discrete treatment leads to the model's interpretation as organizing observations into clusters. This type of supervised clustering should not be confused with "clustered survival data" which typically refers to the case where class labels are known. Instead, observations are gathered according to their best-fitting regression model.

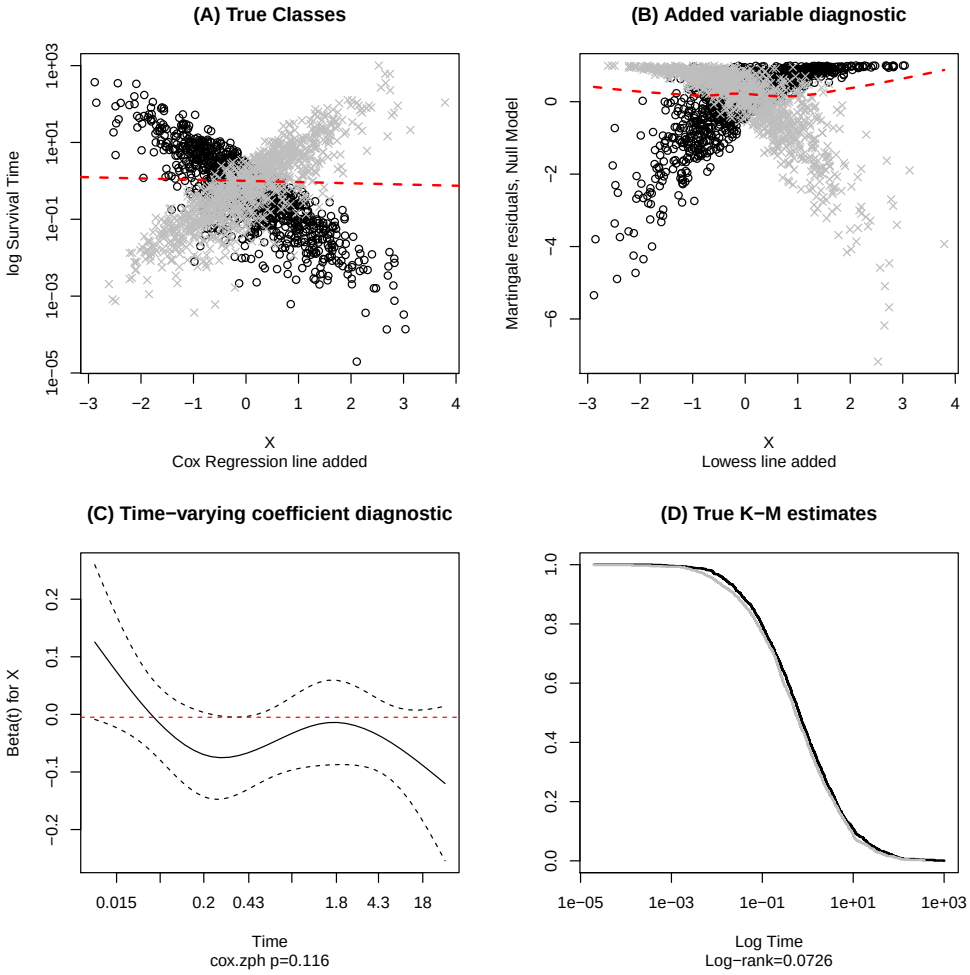
Our mixture here allows some relaxation of the PH assumption; we only need to assume that hazards are proportional within their given clusters. The interpretation of this property is highlighted in the next section.

3 Illustrations

3.1 Genetic Heterogeneity

As we mentioned, there is evidence that distinct molecular subgroups lead to the same clinical presentation of ovarian cancer (Vaughan et al., 2011). This form of genetic heterogeneity may arise because the commonalities leading to cancer may aggregate in pathways and not on the level of genes (Jones et al., 2008).

Figure 1: (A) Log simulated survival times by covariate; underlying heterogeneity is represented by black and blue classes. The Cox model estimated relationship with covariate X is indicated by the dashed line. (B). An added variable plot with the estimated functional relationship between survival time and X drawn in red. (C) A check for non-proportional hazards finds no significant deviation. (D) Even though the generating models are quite different, the true, marginal survival estimates show no significant difference.



In the following case study, we highlight the ability of the mixture to produce unusual associations between covariates and survival and how these unusual associations may augment our understanding of subgroup discovery.

Suppose that $X \sim \mathcal{N}(0,1)$ represents a single, typical, normalized gene expression measurement and that patients do indeed have survival times arising from two distinct hazard models, $h_1(t|x) = h_0(t)\exp(-2x)$ and $h_2(t|x) = h_0(t)\exp(+2x)$. These hazards represent an extreme version of heterogeneity in expression; in one class, the gene has a protective effect and in the other it is equally deleterious.

Assuming the baseline hazard is exponential ($h_0(t) = 1$), we generate 1000 complete survival times, Y, under each of these hazards and plot them on the log scale with their randomly generated expression in Figure 1A. Without knowledge of the true classes, fitting a standard Cox regression to these data finds no significant relationship between Y and X ($\hat{\beta} = -0.0314$,$p = 0.1$). This effect is strong enough that the relationship is easily identified if the true classes are known ($\hat{\beta}_1 = 1.93$, $p < 0.001$ and $\hat{\beta}_2 = -2.12$, $p < 0.001$).

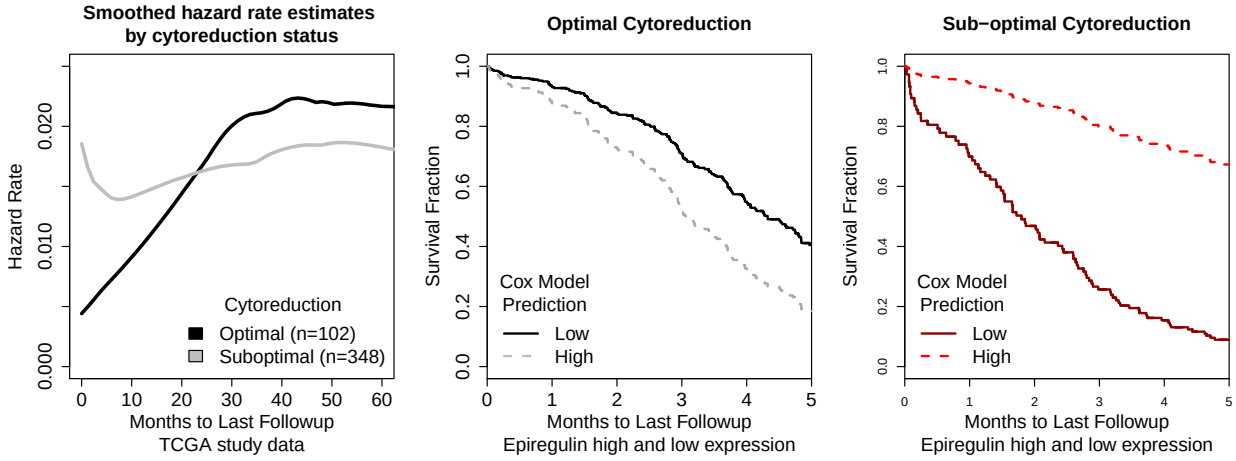
A standard diagnostic technique to estimate non-linear relationships between Y and X is to use a smoothing estimate on the added variable plot (Figure 1B), but it does not identify any important effect. Estimating

a time-varying effect is another diagnostic for assessing PH (Grambsch and Therneau, 1994). Again it does not distinguish any non-proportional effect ($p = 0.116$) or time-varying effect (Figure 1C). So, by the standard analyses, this important gene would not be identified for further study.

With respect to gene expression analyses, this is a case where differential expression (DE) models that look for mean difference will not work: there is no true underlying survival difference between classes attributable to X (Figure 1D). This means that models that try to estimate a rule $1_{\{X>c\}}$ that can classify patients will not work here. So, the mixture reflects a different way to model risk in gene expression. The question remains about whether this kind of example exists in real data. We present a gene that does just this in the following illustration.

3.2 Cytoreduction and Epiregulin

Figure 2: Smoothed hazard rate estimates stratified on optimal and suboptimal cytoreduction classes show a non-proportional relationship. We identify gene epiregulin whose relationship to survival inverts across these underlying classes.



We draw known class labels from the surgical covariate cytoreduction. It is a measure of success of debulking surgery, which is a component of primary therapy in ovarian cancer. Patients who are sub-optimally cytoreduced have a clinically significant amount of residual tumor that will seed future recurrent disease and progression (Bhoola and Hoskins, 2006).

Using the TCGA study to be introduced in Section 6, we separate the patients into optimal and sub-optimal categories and provide a kernel smoothing estimate of their hazards (Figure 2). The estimated hazards are clearly non-proportional ($p = 0.007$) reflecting the early protective effect of optimal cytoreduction.

Fitting separate models in each subgroup, we searched for genes whose relationship with survival inverts over classes finding epiregulin, which has been recently highlighted as a progression marker (Amsterdam et al., 2011). In optimally cytoreduced patients, increased expression is detrimental to survival ($\hat{\beta} = 0.156, p = 0.014$); in sub-optimal patients, increased expression is protective ($\hat{\beta} = -0.452, p = 0.012$). In Figure 2, we have plotted the estimated survival for high and low epiregulin expression for optimal and sub-optimal patients. So these effects do exist and, in genomic surveys, this effect is an ideal target for functional studies.

Given that we want to identify genetic subgroups with different prognoses, we should favor a model that admits unknown and possibly dramatically different survival experience. The mixture model should let us estimate labels and should be able to resolve non-proportional hazards.

4 Algorithm: Cox Assisted Clustering

We refer to the following algorithm for maximizing the mixture likelihood as Cox Assisted Clustering (CAC). For convenience, we write the parameters to be estimated as $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$, the mixing proportions, $\mathbf{h} = \{h_{01}(t), \ldots, h_{0K}(t)\}$, the set of baseline hazard functions, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$, the coefficient vectors. We further abbreviate the hazards at their evaluation points: $h_{0ki} = h_{0k}(Y_i)$ and $H_{0ki} = H_{0k}(Y_i)$.

The complete data likelihood with mixing parameters $\boldsymbol{\pi}$ and class specific parameters $\mathbf{h}$ and $\boldsymbol{\beta}$ may be separated into a mixing distribution part and a component distribution part: $\log L(\boldsymbol{\pi}, \mathbf{h}, \boldsymbol{\beta}; Y, \Delta, U|\boldsymbol{x}) = \log L_1(\boldsymbol{\pi}; U) + \log L_2(\mathbf{h}, \boldsymbol{\beta}; Y, \Delta|U, \boldsymbol{x})$. The first is simply $\log L_1(\boldsymbol{\pi}; U) = \sum_{k=1}^{K} \left(\sum_{i=1}^{n} u_{ik}\right) \log \pi_k$. The likelihood associated with the component distributions is

$$\log L_2(\mathbf{h}, \boldsymbol{\beta}; Y, \Delta|U, \boldsymbol{x}) = \sum_{k=1}^{K} \sum_{i=1}^{n} \delta_i u_{ik} \log h_{0ki} + \delta_i u_{ik} x_i' \beta_k - u_{ik} H_{0ki} \exp\left(x_i' \beta_k\right). \tag{4}$$

To compute the maximum likelihood estimate, we follow an Expectation Maximization (EM) approach (Dempster et al., 1977) that estimates and optimizes the observed data log likelihood by plugging $\widehat{u}_{ik} = E(u_{ik}|Y_i, \delta_i)$ into the complete data likelihood. Supposing that the current values of the parameters at the mth iteration are $\pi_k^{(m)}$, $h_{0ki}^{(m)}$, $H_{0ki}^{(m)}$, and $\beta_k^{(m)}$, the algorithm proceeds as follows.

In the E-step, conditional mean is

$$\widehat{u}_{ik} = \frac{\pi_k \left[h_{0ki}^{(m)} \exp\left(x_i' \beta_k^{(m)}\right)\right]^{\delta_i} \exp\left[-H_{0ki}^{(m)} \exp(x_i' \beta_k^{(m)})\right]}{\sum_{k'} \pi_{k'} \left[h_{0k'i}^{(m)} \exp\left(x_i' \beta_{k'}^{(m)}\right)\right]^{\delta_i} \exp\left[-H_{0k'i}^{(m)} \exp(x_i' \beta_{k'}^{(m)})\right]} \tag{5}$$

after the application of Bayes rule.

In the M-step, the update for mixing proportions π_k is straightforward:

$$\pi_k^{(m+1)} = \frac{\sum_{i=1}^{n} \widehat{u}_{ik}}{n}. \tag{6}$$

To update $\mathbf{h}$, we make a profile likelihood argument that leads to a partial likelihood (Johansen, 1983). Suppose we hold β_k constant. Maximizing over h_{0k}, we obtain profile estimates of the hazards as a function of the $\beta_k^{(m+1)}$ that are similar to Breslow (1974):

$$h_{0k}^{(m+1)}(Y_i) = \frac{\widehat{u}_{ik}}{\sum_{j:Y_j \geq Y_i} \widehat{u}_{jk} \exp(x_j' \beta_k^{(m+1)})} \tag{7}$$

$$H_{0k}^{(m+1)}(Y_i) = \sum_{l:Y_l \leq Y_i} \frac{\widehat{u}_{lk}}{\sum_{j:Y_j \geq Y_l} \widehat{u}_{jk} \exp(x_j' \beta_k^{(m+1)})}. \tag{8}$$

The profiled M-step objective is is a partial likelihood weighted by the $\hat{u}_{ik}$:

$$\log L_2\left(\mathbf{h}(\boldsymbol{\beta}), \boldsymbol{\beta}; Y, \delta, \widehat{U} \mid \boldsymbol{x}\right) = \sum_{k=1}^{K} \sum_{i=1}^{n} \delta_i \left\{ \widehat{u}_{ik} x_i' \beta_k - \log \sum_{j:Y_j \geq Y_i} \exp\left[\widehat{u}_{jk} x_j' \beta_k\right] \right\} \tag{9}$$

Each component indexed by k, may be maximized separately to obtain the $\beta_k^{(m+1)}$ update using standard statistical software. The M-step is operationally equivalent to fitting K weighted Cox models. Finally, one iterates between the E and M step until the increment in log-likelihood is small.

4.1 Starting Conditions and Number of classes

An EM algorithm is typically seeded by a choice of parameters $\boldsymbol{\beta}^{(0)}, \boldsymbol{\pi}^{(0)}$. For analyses that begin with strong biological hypotheses, the corresponding parameters may be set directly.

An alternative is to choose starting parameters by assigning observations to specific classes and estimating the initial $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\pi}^{(0)}$. This is equivalent to setting an initial value for every $\hat{u}_{ik}$ and running the algorithm

forward. This assignment may be random; one may set a randomly selected $\hat{u}_{ik}$ to 0.8 and divide the remaining weight among the other classes. In practice, we use multiple random starts and pick the best by the fitted log-likelihood.

As in other clustering problems (Fraley and Raftery, 1998), we select the number of classes using the Bayesian Information criterion (BIC). Let $L(K) = L(\mathbf{h}_K, \boldsymbol{\beta}_K, \boldsymbol{\pi}_K; Y, \delta, U|\boldsymbol{x})$, where we have added the K subscript to emphasize the dependence. The BIC criterion is expressed as

$$BIC(K) = -2\log L(K) + pK\log(n). \tag{10}$$

5 Simulation Studies

While there are several properties of the model and algorithm to highlight, we focus on its treatment of censored data and a demonstration of its estimation ability.

5.1 Censored data case studies

We study the effect of censoring on uncertainty and clustering in two artificial case study settings. Let true class indicator $U_i \in \{1, 2\}$ be evenly split among $2n$ observations with a single covariate $(X_1, \ldots, X_{2n}) \sim \mathcal{N}(\mu 1_{2n}, I_{2n})$ independent normal with mean $\mu \geq 0$ and variance 1. As is common in gene expression studies, we will work with scaled and centered X, so μ reflects the sensitivity of the analysis to this standardization.

The relationship between survival and X is controlled by $\beta \geq 0$ where the first class has $\beta_1 = \beta$ and the second class has $\beta_2 = -\beta$. The survival time for the ith patient is then $T_i = \frac{\epsilon_i}{\exp(X_i\beta_{(U_i)})}$ where $\epsilon_i \sim$ Exponential(1). The censoring time is generated from $C_i \sim$ Uniform$(0, \lambda)$ where λ depends on the choice of μ and β and a target censoring rate. Finally, the observed survival time is $Y_i = \min(T_i, C_i)$.

We set $n = 200$ patients in each class and set $\beta = 3$ so that $\beta_1 = +3$ and $\beta_2 = -3$. We target 40% censoring by setting $\lambda = \exp(0.99)$ for $\mu = 0$ and $\lambda = \exp(12.83)$ for $\mu = 5$. For this simulation, we run our algorithm at the true number of clusters $K = 2$.

In Figure 3, we plot the log observed survival time Y_i by the true covariate values X_i: two classes are indicated by symbols (+ and x) and colors (red/pink versus black/grey). The lighter shades indicate censored observations.

In the $\mu = 0$ case, we observe a set of points with similar X values and similar survival response regardless of class suggesting that it may be futile to try to classify these cases. The scenario where $\mu = 5$ is a clear example of a generating model where X has some relationship that cannot be detected marginally: one cannot find a rule $1_{\{X>c\}}$ that classifies the data. However, in the time dimension, there are clearly two modes representing early and late failures.

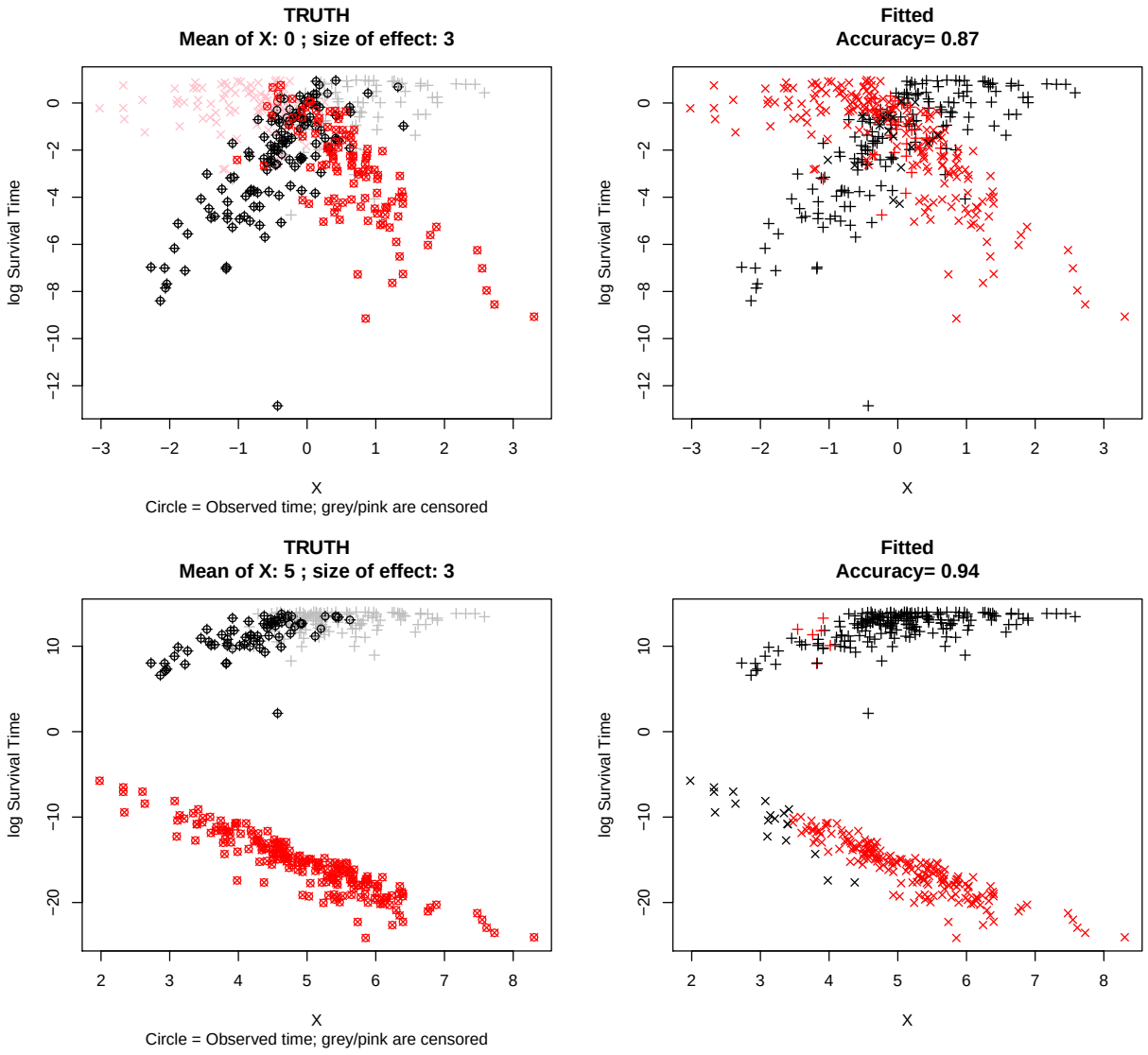
The right hand panels show the CAC fitted models for $K = 2$ run for likelihood tolerance < 0.001. We report the estimated β_k (choosing $\hat{\beta}_1 \leq \hat{\beta}_2$ for identifiability) in Table 1 alongside the oracle estimator that knows the true classes. Intuitively, if the data are perfectly classified, the oracle estimate will have properties consistent with the well-studied Cox model estimates. Thus the accuracy of the CAC estimate is an ideal measure of loss of performance due to uncertainty. The realized censoring rate is also reported.

Additionally, we study the same scenarios through 1000 simulations. While the results imply that the clustering algorithm works well in the face of heavy censoring and mean mis-specification (common to the statistical treatment of expression studies), there are two important points. First, we see a "greedy bias" as the algorithm takes observations whose true class may be hard to identify and uses them to reinforce what it has already learned. Second, we see a censoring bias for the $\mu = 5$ study: the negative β group is more likely to be censored so it has a lower effective sample size. As the figure shows, the higher values are also more likely to be censored so the estimate ends up being biased towards zero relative to the oracle

Our understanding of the standard assumptions on censored data need to be augmented in this mixture model. Typically we assume that censoring times are uninformative which means that while we use the fact that $T_i > C_i$ when $\delta_i = 0$, the time C_i itself is not meaningful. However, under the mixture framework, groups are encouraged to have different average survival times and we note that a group with long average survival is more likely to be censored.

For example on the top row of Figure 3, censoring appears in both groups equally (76/200 and 89/200). In contrast, in the bottom row example, censoring appears in one group preferentially (158/200 and 0/200).

Figure 3: Two simulated data sets ($\mu = 0$ and $\mu = 5$) with censoring and true classes indicated.



Intuitively, this makes sense: because our goal is to identify groups which have distinct survival models, meaningful ones may have drastically different means and interact differently with the censoring process.

6 Data Analysis

Because of its frequency among gynecological cancers, its high lethality, and poor options for treatment (Vaughan et al., 2011), serous ovarian cancer was a pilot target for molecular characterization in The Cancer Genome Atlas (TCGA) (The Cancer Genome Atlas Research Network, 2011). The study collected banked surgical samples from $n = 503$ patients with highly-annotated clinical followup whose cancers had been surgically debulked and who had been treated with platinum-based chemotherapy (Bhoola and Hoskins, 2006).

Platinum resistance is an important concept in the treatment of ovarian cancer because these cancers respond poorly to any type of chemotherapy (Bookman, 2005). While resistance is not an ideal predictive marker, because it is defined through treatment, the development of an independently-queried molecular model is precisely the promise of a large repository study like TCGA.

Table 1: Case Study results for the data presented in Figure 3. Below are results from multiple repeated simulations.

Case Study	$\mu = 0$		$\mu = 5$	
$\beta_{CAC,1}$	3.45		3.26	
$\beta_{CAC,2}$	-2.94		-2.05	
$\beta_{oracle,1}$	2.90		3.05	
$\beta_{oracle,2}$	-2.80		-2.92	
accuracy	0.87		0.92	
Censoring rate	0.39		0.40	
1000 Simulations	$\mu = 0$		$\mu = 5$	
$\beta_{CAC,1}$ (sd)	3.45	(0.53)	3.24	(0.58)
$\beta_{CAC,2}$ (sd)	-3.46	(0.52)	-2.14	(0.55)
$\beta_{oracle,1}$ (sd)	3.05	(0.34)	3.03	(0.28)
$\beta_{oracle,2}$ (sd)	-3.03	(0.33)	-3.12	(0.57)
accuracy (range)	0.87	(0.78-0.94)	0.91	(0.63-1.00)
censoring rate (range)	0.39	(0.31-0.47)	0.39	(0.36-0.44)

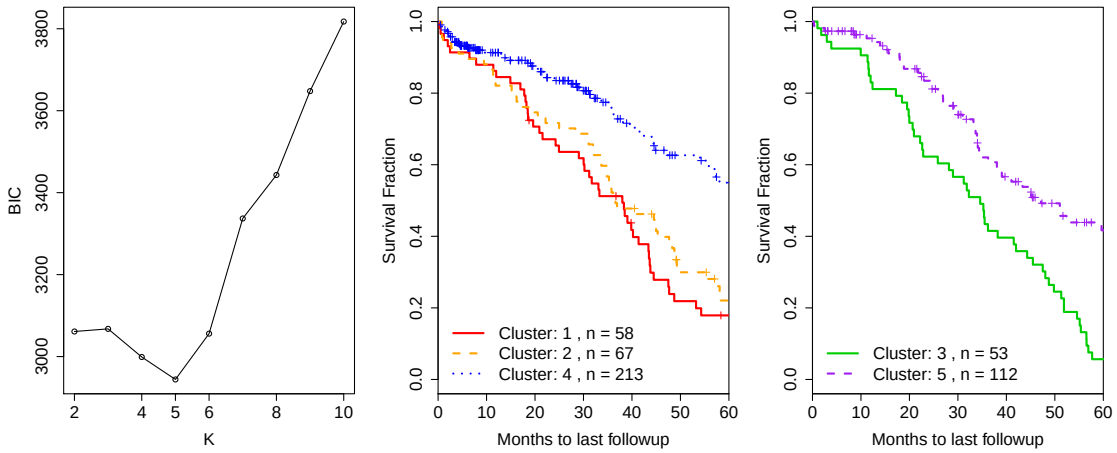
As we have mentioned, one complication is the expectation of genetic heterogeneity (Konstantinopoulos et al., 2008): patients with similar survival outcomes may have dissimilar molecular profiles. If this heterogeneity appears to take the form of subgroups and mixtures (as in the illustrations), we anticipate that our model and algorithm will be able to address it.

Therefore, we demonstrate the use of of our model to explore possibly heterogenous data by modeling a potential mechanism of platinum resistance in TCGA patients. Because recent reviews of resistance highlight the homologous repair pathway for repairing DNA damage (Martin et al., 2008; Cooke and Brenton, 2011), we focus on modeling the function of this set of genes. The homologous repair pathway is defined by KEGG annotation (hsa:03440) (Kanehisa et al., 2010) and corresponds to 27 unique gene symbols.

We fit our model for $K = 2, \ldots, 10$ using 100 random starts for each K and selecting the best fit by log likelihood. Survival times are truncated at 5 years of observation to reduce the influence of 76 patients who are observed beyond the time of interest. In total, 186/503 (37%) patients are censored before 5 years.

We plot the BIC trace in Figure 4, select $K = 5$ clusters. In addition, we plot the Kaplan-Meier (KM) estimates corresponding to the assignment of each patient to their maximum posterior probability class. We have separated these into two panels for visual clarity; we note that the crossing survival estimates (a standard sign of non-proportional hazards) are entirely acceptable in this mixture framework.

Figure 4: BIC model selection and fitted CAC model KM curves for TCGA data, homologous repair pathway.



The quality of the cluster fits may be described by the relative weight of each cluster ($\sum_i \hat{u}_{ik}$), the number of patients assigned (n), and the mean posterior probability for patients in their assigned clusters. The number of events in each cluster and the restricted mean (up to 60 months) is noted.

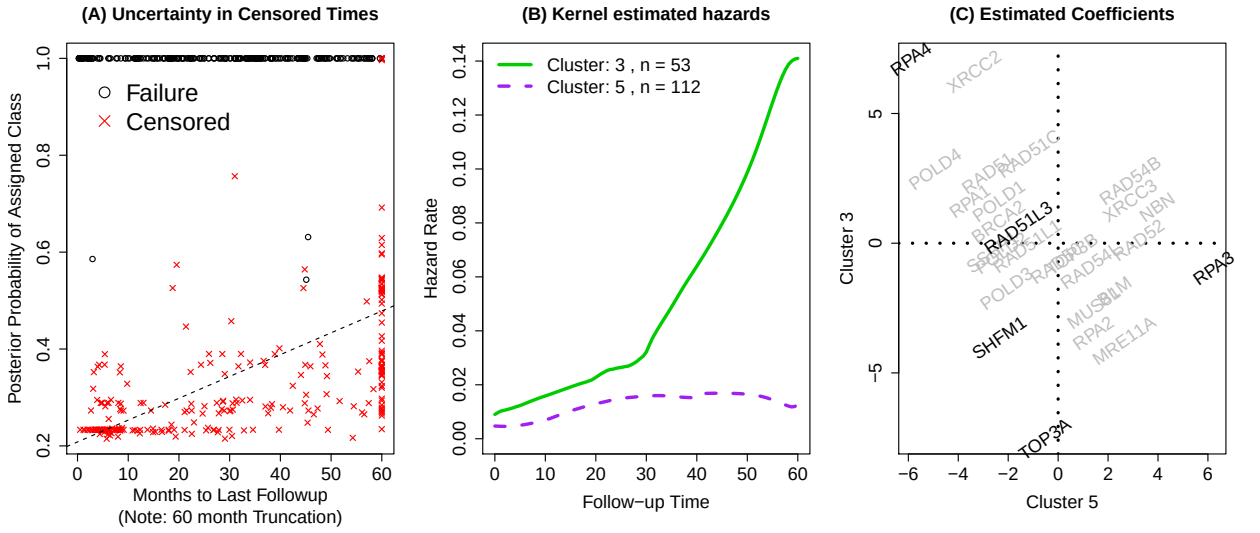
Table 2: Fitted cluster diagnostics for $K = 5$, Homologous repair model.

Cluster	**4**	**5**	**2**	**1**	**3**
restricted mean (months)	34.16	26.54	15.21	12.69	5.90
standard error (mean)	3.04	3.40	3.11	3.07	1.83
n	213.00	112.00	67.005	58.00	53.00
Events	50.00	46.00	50.00	45.00	50.00
Weight	103.22	69.07	57.555	51.47	51.05
mean $\hat{u}$	0.48	0.62	0.86	0.89	0.96

We observe that clusters 4 and 5 have the largest number of patients assigned, but the smallest mean posterior probabilities implying that their members are less similar internally. The clustering appears to be driven by the poor prognosis patients in clusters 1,2 and 3. Note that it is inappropriate to report log-rank statistics for these curves (viz. Figure 5B).

With respect to low posterior probabilities, we observe that the algorithm makes intuitive use of censored observations. We plot the maximum posterior probability for each patient by their survival time (censoring times are indicated in red) in Figure 5A. Because censored patients do not have definitive events, the algorithm is less certain about to which cluster to assign these observations. Importantly, as it observes more time, the algorithm's certainty rises. To wit, the hardest to classify patients are the censored patients with little followup time.

Figure 5: Uncertainty in censored observations with regression line for the censored points added. Hazards and estimated coefficients for clusters 3 and 5. Genes highlighted in the text are bolded.



Within each cluster, we checked model diagnostics and look for influential points. We note that the global test for non-proportional hazards (Grambsch and Therneau, 1994) using all the data was uncomfortably close to significant ($p = 0.0592$); after fitting our model, the within-cluster tests are all strongly insignificant. All DFBETAS for all genes in each cluster are smaller than 1 standard deviation implying no leverage points.

Because the mixture allows different clusters to have different baseline hazard functions, in Figure 5(B), we used a kernel smoothing algorithm to visualize their estimates (Müller and Wang, 1994). We emphasize the non-proportionality of the hazards for clusters 3 and 5: patients in cluster 3 have have a sudden acceleration in their hazard after 30 months which may be consistent with the loss of effect in platinum treatments.

In Figure 5(C), we have plotted the estimated coefficients for clusters 3 and 5. Keeping in mind that the linear predictor in the Cox model scales hazard relative to the cluster specific baseline hazard, we highlight three genes. RPA4's coefficients $(\beta_3, \beta_5) = (+7.13, -5.90)$ imply that it has strong deleterious effect in cluster 3 (exacerbating the jump in hazard) while it has a strong protective effect in cluster 5. Compare to RPA3 $(-0.96, +6.24)$ which only increases risk in cluster 5 and TOP3A $(-7.54, -0.54)$ which is protective in cluster 3 only. At this point, these genes are good candidates for followup studies: we have identified their effect specific to a subgroup of patients.

Further, the clustering model may still recover a sense of differential expression (DE) for survival data. Because we have learned risk classes, we may consider differential expression across clusters. We focus on DE across clusters 3 and 5: SHFM1 (bonferroni p=0.011), RPA3 (p=0.011), RAD51L3 (p=0.020) all have significant shifts in expression. Notably, RPA4 and TOP3A do not have significant DE implying that they fit into the class of variables whose effects are class specific.

Based on our analysis, we conjecture that we are able to identify a subgroup of patients (cluster 3) who experience a significant increase in hazard around month 30. We are able to identify genes whose expression leads to increased risk specific to a subgroup or whose relationship inverts across clusters. We note that there is a tremendous amount of untapped information remaining in the fitted model. For example, every pairwise comparison between clusters should be informative as well as their holistic interpretation, foreshadowing the utility of this methodology for exploratory data analysis.

7 Discussion

In this article, we have presented a model for heterogeneity in time-to-event data. While its actual formulation is straightforward (continuous mixtures and mixtures with known classifications have been previously addressed, for example Lawless (1982)), the treatment of unknown classification, a consideration of the implications for censoring, the effect on genomic predictors and diagnostic analysis have not been previously considered. Finally, we have presented a novel and informative analysis in Section 6 which begins to answer Dr. Levine's question: the set of alterations in expression that are important are subgroup dependent.

With respect to developing ovarian cancer biomarkers, our data analysis has shown an example where class identification leads to risk stratification. We might further identify high and low risk classes within the assigned clusters as is standard practice, but this is no longer a necessary part of the expression analysis. The CAC algorithm has also given us its posterior weights allowing a concrete measure of uncertainty for downstream analyses.

Admirably, this model relaxes the whole model PH assumption to conditional PH given cluster membership. In an exploratory framework, the utility of this flexibility cannot be underestimated. Additionally, the ability to seed the algorithm from an existing set of classifications will be particularly useful: imagine starting with unsupervised k-means clustering and running this algorithm to compare a supervised model fit. Both our simulated and applied analysis highlight that our understanding of censoring has been augmented and the use of information in the model is intuitively simple.

A complication of this flexibility is that the interpretation of effects and subgroups has to account for several few moving pieces: the covariates themselves may have different distributions across groups (differential expression), the effect on the relative hazard may be different (different regression models) and the baseline hazards may be different. This is not non-identifiability per se, but end users should be careful about across cluster inferences made on the predictors.

Acknowledgements

This work was funded by a grant from the National Library of Medicine, NIH (LM007359) (KE).

References

O.O. Aalen. Heterogeneity in survival analysis. *Statistics in Medicine*, 7(11):1121–1137, 1988.

A. Amsterdam, E. Shezen, C. Raanan, Y. Slilat, A. Ben-Arie, D. Prus, L. Schreiber, et al. Epiregulin as a marker for the initial steps of ovarian cancer development. *International Journal of Oncology*, 39(5):1165, 2011.

S. Bhoola and W.J. Hoskins. Diagnosis and management of epithelial ovarian cancer. *Obstetrics & Gynecology*, 107(6):1399, 2006.

MA Bookman. Standard treatment in advanced ovarian cancer in 2005: the state of the art. *International Journal of Gynecological Cancer*, 15:212–220, 2005.

N. Breslow. Covariance analysis of censored survival data. *Biometrics*, pages 89–99, 1974.

S.L. Cooke and J.D. Brenton. Evolution of platinum resistance in high-grade serous ovarian cancer. *The Lancet Oncology*, 2011.

D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

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

P.M. Grambsch and T.M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994.

S. Johansen. An extension of cox's regression model. *International Statistical Review/Revue Internationale de Statistique*, pages 165–174, 1983.

S. Jones, X. Zhang, D. W. Parsons, J. C.-H. Lin, R. J. Leary, P. Angenendt, and et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, 321:1801, 2008.

M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa. Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 38:D355–D360, 2010.

P.A. Konstantinopoulos, D. Spentzos, and S.A. Cannistra. Gene-expression profiling in epithelial ovarian cancer. *Nature Clinical Practice Oncology*, 5(10):577–587, 2008.

J.F. Lawless. *Statistical Models and Methods for Lifetime Data*. Wiley New York, 1982.

Douglas A. Levine. Interview at International Gynecologic Cancer Society, December 2012. http://ecancer.org/tv/pubdate/710.

K. Lostritto, R. Strawderman, and A. Molinaro. A partitioning deletion/substitution/addition algorithm for creating survival risk groups. *Arxiv preprint arXiv:1101.4331*, 2011.

L.P. Martin, T.C. Hamilton, and R.J. Schilder. Platinum resistance: the role of dna repair pathways. *Clinical Cancer Research*, 14(5):1291–1295, 2008.

HG Müller and JL Wang. Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics*, 50(1):61, 1994.

Y.J. Na, J. Farley, A. Zeh, M. del Carmen, R. Penson, and M.J. Birrer. Ovarian cancer: markers of response. *International Journal of Gynecological Cancer*, 19(11):S21, 2009.

J. O'Quigley and J. Stare. Proportional hazards models with frailties and random effects. *Statistics in Medicine*, 21(21):3219–3233, 2002.

The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474:609–615, 2011.

S. Vaughan, J.I. Coward, R.C. Bast, A. Berchuck, J.S. Berek, J.D. Brenton, G. Coukos, C.C. Crum, R. Drapkin, D. Etemadmoghadam, et al. Rethinking ovarian cancer: recommendations for improving outcomes. *Nature Reviews Cancer*, 11(10):719–725, 2011.