Published in final edited form as:

Comput Stat. 2019 March ; 34(1): 395–414. doi:10.1007/s00180-018-0827-6.

Fusion Learning Algorithm to Combine Partially Heterogeneous Cox Models

Lu Tang, Ling Zhou, and Peter X.K. Song

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

lutang@umich.edu, zholing@umich.edu, pxsong@umich.edu

Abstract

We propose a fusion learning procedure to perform regression coefficients clustering in the Cox proportional hazards model when parameters are partially heterogeneous across certain predefined subgroups, such as age groups. One major issue pertains to the fact that the same covariate may have different influence on the survival time across different subgroups. Learning differences in covariate effects is of critical importance to understand the model heterogeneity resulted from the between-group heterogeneity, especially when the number of subgroups is large. We establish a computationally efficient procedure to learn the heterogeneous patterns of regression coefficients across the subgroups in Cox proportional hazards model. Utilizing a fusion learning algorithm coupled with the estimated parameter ordering, the proposed method mitigates greatly computational burden with little loss of statistical power. Extensive simulation studies are conducted to evaluate the performance of our method. Finally with a comparison to some popular conventional methods, we illustrate the proposed method by a vehicle leasing contract renewal analysis.

Keywords

Fused lasso; Regression coefficient clustering; Extended BIC; Cox proportional hazards model

1 Introduction

This article concerns the integrative data analysis of partially heterogeneous subgroups where the outcome of interest is time to event. As the most popular statistical model for survival analysis, the Cox proportional hazards model [7] offers a flexible form to study the impacts of covariates on survival probability via the following hazard function specification:

$$\lambda(t|x) = \lambda_0(t) \exp(x^T \beta), \quad (1)$$

where $\lambda(\cdot)$ is a hazard function of time T , x and β are $p \times 1$ vectors of covariates and corresponding regression coefficients, respectively, and $\lambda_0(t)$ is an unspecified baseline hazard function. To handle data with potential parameter heterogeneity across sample of K subgroups, the most straightforward way is to fit an interaction-based Cox model [4] of the form

$$\lambda(t|x) = \lambda_0(t) \exp \left(\sum_{k=1}^K x^T \beta_k I\{g(x) = k\} \right), \quad (2)$$

with $g(x)$ being the subgroup index of the subject x , which accounts for the parameter heterogeneity across subgroups via distinct sets of regression coefficients $\{\beta_k\}_{k=1}^K$. Model (2) is similar to analyzing individual subgroup dataset separately, assuming no common parametric structure between any pair of subgroups. Despite its simplicity, this model fails to achieve high statistical power since almost no information is shared across subgroups [20]. An alternative strategy to model (2) is the homogeneous model which estimates a single coefficient vector β , assuming $\beta = \beta_1 = \dots = \beta_K$. However, such strong assumption is hard to be justified and may bias the parameter estimates if violated [9, 16], especially when the number of subgroups K is large. When the underlying parameter structure is partially heterogeneous, say, regression coefficients of the same effect are clustered between some subgroups, it is necessary to strike a balance between using too few and using too many parameters, which is the primary objective of this paper. In here, the membership of individuals belonging to the K subgroups is fixed *a priori*.

This paper is motivated by an analysis of vehicle contract renewal data, in which lease contracts on 5 different vehicle models, A, B, C, D, and E, made by a foreign motor company, are recorded with sample sizes ranged between 2,000 and 6,000. Customers typically renew their lease contracts for new vehicles at the end of their current lease contracts, but some customers may opt to do so at an earlier time due to various reasons. Thus, it is of great interest to understand the timing that current customers decide to lease new vehicles. The event time or survival time of a current contract is defined as the duration from the previous contract time to the time of signing a new contract. Predictors of interest include vehicle value, allowed mileage, region of sale, and contract terms. The overfitted model (2) estimates distinct sets of parameters, $\beta_A, \beta_B, \beta_C, \beta_D$ and β_E , each for one vehicle model. Instead, we assume the regression coefficients are partially heterogeneous, and we aim to achieve two goals: (a) to identify predictors ($x_A, \dots, x_E$) that are associated with the distribution of contract renewal events; and (b) to assess whether any of the identified predictor effects in $\beta_A, \dots, \beta_E$ exhibit similar impacts on the probability of contract renewals across these five vehicle models.

In current literature, effort has been devoted to models accounting for intersubgroup parameter heterogeneity. Two types of the most popular methods include interaction-based Cox models [7, 15] and Cox models with random effects [17, 19]. However, the method using interaction terms to account for parameter heterogeneity may easily lead to overfitting, and hence falls short of statistical power. In this setting, one may circumvent the overfitting problem by identifying homogeneous parameter groups through exhaustive hypothesis tests, which clearly suffers from unduly computational burden as well as the curse of dimensionality in testing hypotheses. In addition, different or even conflicting conclusions may be encountered from hypotheses on arbitrary pairs of study-specific coefficients. On the other hand, random effect models assume fully heterogeneous covariate effects, in which

model simplification is difficult to be carried out, and such full heterogeneity is hard to be interpreted in practice.

It is reasonable to believe that the inter-subgroup parameter structure lies in between the space of complete heterogeneity (i.e., all covariate effects are different) and complete homogeneity (i.e., all covariate effects are exactly the same), which we refer to as partial heterogeneity. Recently, some researchers (e.g., [6], [22] and [20]) developed fusion learning approaches to cluster regression coefficients of cross-sectional or longitudinal data from multiple heterogeneous data sources. Along the lines of clustering similar regression coefficients, in this paper we focus on the development of fusion learning in the Cox proportional hazards model for time-to-event data from partially heterogeneous subgroups. Our proposed fusion learning method enjoys the following advantages: (i) flexibility. It identifies homogeneity, partial or complete heterogeneity in a flexible way to facilitate model specification and parameter interpretation; (ii) efficiency. It detects homogeneous groups and merges them together so as to share the same parameterization, resulting in simpler model specification which leads to improvement in estimation efficiency and statistical power; and (iii) computation simplicity. Assessing heterogeneity and homogeneity is automatic and carried out by the regularized fused lasso [21, 11] during the time of parameter estimation. In this way, it overcomes the exhaustive amount of computational burden incurred in the simultaneous hypothesis testing methods and thus greatly reduces computation time.

The rest of the paper is organized as follows. Section 2 describes the proposed fusion learning method to cluster and combine regression coefficients in the Cox proportional hazards model. Section 3 presents simulation studies to evaluate the performance of the proposed method. A real data analysis is given in Section 4 to illustrate the use of the fused coefficient estimates. In Section 5, we introduce the R package metafuse which implements the proposed method. We conclude the paper with a brief discussion in Section 6.

2 Fusion Learning Methodology

Fusion learning methodology is built upon a fused lasso approach that enables clustering of regression coefficients to simplify parameter specification.

2.1 Regularized estimation

We start with necessary notations for the formulation of Cox model. Through-out the paper, we use i , j and k to index subject, covariate and subgroup, respectively. The total number of subgroups is K ; the number of covariates is p ; the sample size of the k th subgroup is n_k , $k = 1, \dots, K$; and the combined sample size is $N = \sum_{k=1}^K n_k$. The dataset of the k th subgroup is denoted as $\{y_{k,i}, x_{k,i}, \delta_{k,i}\}$, with individual's covariates $x_{k,i} = (x_{1,k,i}, \dots, x_{p,k,i})^T$, observed time $y_{k,i} \triangleq \min(T_{k,i}, c_{k,i})$ and event indicator $\delta_{k,i} \triangleq \mathbb{I}(T_{k,i} < c_{k,i})$, where $T_{k,i}$ and $c_{k,i}$ are the failure time and the censoring time, respectively. The vector of all regression coefficients is denoted as $\beta = (\beta_1^T, \dots, \beta_p^T)^T$, a Kp -element vector, where the subgroup-specific vector of coefficients corresponding to the j th covariate is $\beta_{j\cdot} = (\beta_{j,1}, \dots, \beta_{j,K})^T$. Let $c = (c_1, \dots, c_p)^T$ be an indicator vector of flag values, where $c_j = 1$ indicates that the j th covariate is likely to be

heterogeneous (i.e., coefficients $\beta_{j,1}, \dots, \beta_{j,K}$ are not all the same), or otherwise $c_j = 0$ if homogeneous ($\beta_{j,1} = \dots = \beta_{j,K}$, a common coefficient across K subgroups). In the latter case, the coefficient vector $\beta_{j\cdot}$ reduces to a common scalar vector $\beta_{j,1_K} = (\beta_{j,1}, \dots, \beta_{j,1})^T$ if $c_j = 0$, where 1_K is the p -element vector of ones.

We assume that $T_{k,i}$ and $C_{k,i}$ are conditionally independent given $x_{k,i}$ and that the censoring mechanism is noninformative. Additionally, we assume that all subgroups have a common baseline hazard $\lambda_0(t)$. We start from the complete heterogeneous Cox proportional hazards model specification that takes the form of model (2). The partial likelihood for model (2) can be written as the product of partial likelihoods from all subgroups,

$$L(\beta) = \prod_{k=1}^K L_k(\beta) = \prod_{k=1}^K \prod_{i=1}^{n_k} \left[\frac{\exp(x_{k,i}^T \beta_k)}{\sum_{k'=1}^K \sum_{i'=1}^{n_{k'}} I(y_{k',i'} \geq y_{k,i}) \exp(x_{k',i'}^T \beta_{k'})} \right]^{\delta_{k,i}}, \quad (3)$$

where $L_k(\beta)$ is the likelihood piece corresponding to the k th subgroup. We specify the risk set (i.e., the denominator) in (3) based on all subjects to assess the overall risk. Clearly, model (2) is an overfitted model and not ideal for interpretation when some coefficients are similar across some of the subgroups. To detect potentially partial homogeneity in the model specification, we invoke the fused lasso penalty to identify homogeneous parameters and then fuse them as common parameters to achieve model parsimony. In particular, we utilize the sparse fused lasso penalty with parameter ordering [22, 20] on the partial likelihood (3), which leads to the following optimization problem:

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \left[-\frac{1}{K} \sum_{k=1}^K \frac{1}{n_k} \ell_k(\beta) + P_{\lambda, \alpha}(\beta) \right], \quad (4)$$

where $\ell_k(\beta)$ is the log partial likelihood of the k th subgroup given by

$$\ell_k(\beta) = \log L_k(\beta) = \sum_{i=1}^{n_k} \delta_{k,i} \left[x_{k,i}^T \beta_k - \log \left[\sum_{k'=1}^K \sum_{i'=1}^{n_{k'}} I(y_{k',i'} \geq y_{k,i}) \exp(x_{k',i'}^T \beta_{k'}) \right] \right], \quad (5)$$

and $P_{\lambda, \alpha}(\beta)$ is the sparse fused lasso penalty [20] defined by

$$P_{\lambda, \alpha}(\beta) = \lambda \sum_{j=1}^p c_j \sum_{k=2}^K \omega_{j,k} |\beta_{j,(k)} - \beta_{j,(k-1)}| + \alpha \lambda \sum_{j=1}^p \omega_{j,1} |\beta_{j,[1]}|, \quad (6)$$

with tuning parameters $\lambda \geq 0$ and $\alpha \geq 0$, adaptive weights $\omega_{j,k}$, $j = 1, \dots, p$, $k = 1, \dots, K$, $\beta_{j,(1)}, \dots, \beta_{j,(K)}$ the ordered $\beta_j = (\beta_{j,1}, \dots, \beta_{j,K})^T$ from the smallest to largest (ties resolved), and $\beta_{j,[1]}$ the smallest parameter in $|\beta_{j,\cdot}| = (|\beta_{j,1}|, \dots, |\beta_{j,K}|)^T$ (i.e., the statistic closest to zero).

The first part of $P_{\lambda,\alpha}(\beta)$ is the l_1 -norm of parameter differences of adjacently ordered pairs which enforces fusion between regression coefficients across subgroups, and the second part of $P_{\lambda,\alpha}(\beta)$ is the sparsity penalty which induces shrinking on the parameter closest to zero. The force of sparsity penalty is a times the force of fusion penalty. As a result of both parts of $P_{\lambda,\alpha}(\beta)$, the penalty function is capable to shrink the entire group of coefficients β_j to zero. The flag variable C_j is predefined so that fusion is only applied to those covariates β_j that are suspected as being heterogeneous, i.e., $j \in \{j: C_j = 1\}$. In this way, the computational burden may be greatly reduced. As a special case of (6), we can turn off variable selection by setting $\alpha = 0$. Following [23], we use the adaptive weights $\omega_{j,k} = 1/|\beta_{j,(k)} - \beta_{j,(k-1)}|^r$, $j = 2, \dots, p$, and $\omega_{j,1} = 1/|\beta_{j,[1]}|^r$.

Since the true parameter ordering and weights in penalty (6) are unknown, we propose to estimate the ordering from the estimated ranks of a certain root- n consistent estimates $\hat{\beta}$, which may be, for example obtained by solving equation (4) with $\lambda = 0$; and such initial estimates $\hat{\beta}$ can be used to calculate the weights $\omega_{j,k}$'s. For both simulation experiments and data application in this paper, we set $r = 1$ in the adaptive weight $\omega_{j,k}$'s.

2.2 Implementation

To carry out the optimization in (4), following [20], we can turn it into a lasso regression via the trick of reparameterization. For the vector of coefficients β_j , with respect to the j th covariates, we transform the vector of parameters into a new vector of parameters $\theta_{j,\cdot} = (\theta_{j,1}, \dots, \theta_{j,K})^T$, $j = 1, \dots, p$, according to the following scheme:

$$\begin{cases} \theta_{j,1} = \beta_{j,[1]}, \\ \theta_{j,k} = \beta_{j,(k)} - \beta_{j,(k-1)}, \quad \text{for } k = 2, \dots, K, \end{cases}$$

where $\beta_{j,(k)}$ denotes the k th largest value in the vector of β_j . Then the sparse fused lasso penalty $P_{\lambda,\alpha}(\beta)$ in (6) may be rewritten as follows:

$$P_{\lambda,\alpha}(\theta) = \lambda \sum_{j=1}^p \left(\omega_{j,1} |\theta_{j,1}| + \sum_{k=2}^K \omega_{j,k} |\theta_{j,k}| \right) = \lambda \sum_{j=1}^p \sum_{k=1}^K \omega_{j,k} |\theta_{j,k}|,$$

with $\omega_{j,1} = \alpha/|\theta_{j,1}|^r$ and $\omega_{j,k} = c_j/|\theta_{j,k}|^r$ for $k = 2, \dots, K$. Therefore, a solution to the sparse fused lasso problem can be obtained equivalently by performing a traditional lasso problem with respect to coefficients θ . Conveniently, the lasso optimization for the Cox proportional hazards model is already implemented in R package `glmnet` (version 2.0–2) [12], which is the primary numerical toolbox used in our proposed fusion learning procedure.

2.3 Selection and interpretation of tuning parameter

As usual, the tuning parameter λ can be selected according to certain model selection criteria, for example the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the generalized cross-validation method (GCV). As the number of subgroups K becomes large, the extended BIC (EBIC) [5, 13] appears to perform better at selecting clustering patterns due to the fact that EBIC encourages more sparseness (or

smaller number of heterogeneous groups of parameters) when compared to the conventional BIC, as shown in [20]. Thus, here we advocate the use of EBIC for the Cox model; EBIC is given by

$$EBIC_{\lambda} = -2 \sum_{k=1}^K \frac{\bar{n}}{n_k} \ell_k(\hat{\beta}(\lambda)) + df(\hat{\beta}(\lambda)) \log(N) + 2\gamma \log \sum_{j=1}^p \binom{K}{df(\hat{\beta}_{j\cdot}(\lambda))}, \quad (7)$$

where $\hat{\beta}(\lambda)$ is the regularized estimation of β at a given λ , $\bar{n} = N/K$, $df(\hat{\beta}(\lambda))$ is the number of distinct coefficients in $\hat{\beta}(\lambda)$, and $\ell_k(\hat{\beta}(\lambda))$ is the log partial likelihood in (5) evaluated at $\hat{\beta}(\lambda)$. As suggested by [5] and [13], the third term in (7) attempts to penalize the total number of possible models with given degrees of freedom in $\hat{\beta}_{j\cdot}$, $j = 1, \dots, p$, at a given λ . For example, when $K = 10$, $\binom{K}{5}$ is much larger than $\binom{K}{2}$ so larger models with 5 parameters tend more likely to be selected than smaller models with 2 parameters due to the popularity of model space not due to the importance. EBIC regularizes such unbalanced model space and encourages a smaller $df(\hat{\beta}_{j\cdot}(\lambda))$, or a larger λ . The tuning parameter $\gamma \in [0, 1]$ controls the weight of this additional penalty. When $\gamma = 0$, EBIC reduces to BIC. For more details, we refer to [5] and [13]. In our numerical experiments and data example, we fix $\gamma = 1$ to impose penalty on the model space as similar to [5]. Simulation studies in Section 3 provide a useful comparison evidence in favor of EBIC when the performances of BIC and EBIC are compared in coefficient fusion. The selection of α is dependent on the ratio of magnitudes between sparse penalty and fusion penalties, and we suggest setting $\alpha = 1$ for equal penalties between fusion and sparseness when variable selection is of interest.

We now introduce a metric to characterize parameter heterogeneity across data subgroups based on the solution path and values of λ . Since most parameter fusion occurs at small values of λ , especially for those homogeneous covariates, for the ease of visualization, we use log-scaled λ with base 10 such that $\tilde{\lambda} = \log_{10}(\lambda + 1)$ in this paper. In the effort of fusing the coefficients in $\beta_{j\cdot}$, $\tilde{\lambda}_{opt} = 0$ suggests no fusion, namely a completely heterogeneous $\beta_{j\cdot}$. On the other hand, $\tilde{\lambda}_{opt} = \tilde{\lambda}_{fuse, j}$, the smallest value of $\tilde{\lambda}$ to achieve the complete fusion, suggests a full homogeneity, namely all coefficients in $\beta_{j\cdot}$ are fused to a common value. Conceptually, the relative magnitude of $\tilde{\lambda}_{opt}$ over $\tilde{\lambda}_{fuse, j}$ reflects the degree of heterogeneity among elements of $\beta_{j\cdot}$. Thus we propose a metric of heterogeneity as follow:

$$\tilde{f}_j = 1 - \frac{\min(\tilde{\lambda}_{opt}, \tilde{\lambda}_{fuse, j})}{\tilde{\lambda}_{fuse, j}}, \quad \text{for } j = 1, \dots, p.$$

This metric $\tilde{f}_j$ is termed as the *friction of fusion*, which ranges between 0 and 1. Intuitively, higher $\tilde{f}_j$ corresponds to a $\tilde{\lambda}_{opt}$ closer to 0 than $\tilde{\lambda}_{fuse}$, leading to higher heterogeneity, and *vice versa*.

3 Simulation Study

In this section, we conduct two simulation studies to evaluate the numerical performance of proposed fusion learning method.

3.1 Simulation I

In the first simulation study, we consider a dataset with $K=10$ subgroups. We consider both equal sample sizes where $n_1 = \dots = n_{10} = 100$ and unequal sample sizes $\{n_1, \dots, n_{10}\} = \{30, 30, 50, 50, 70, 70, 100, 100, 150, 150\}$. No sparsity penalty is applied on the covariates (i.e., $\alpha = 0$), so we want to exclusively examine the performance of the fused lasso method. We consider the following Cox model with three baseline covariates (x_1, x_2, x_3) :

$$\Lambda(T_{k,i}) = \Lambda_0(T_{k,i}) \exp(\beta_{1,k} x_{1,k,i} + \beta_{2,k} x_{2,k,i} + \beta_{3,k} x_{3,k,i}), \quad i = 1, \dots, 100, k = 1, \dots, 10,$$

where $\Lambda(\cdot)$ and $\Lambda_0(\cdot)$ are the cumulative and baseline cumulative hazard functions. The true coefficient vectors across the 10 subgroups take the following grouping patterns:

$$\beta_{1,1} = \dots = \beta_{1,10} = 0,$$

$$\beta_{2,1} = \dots = \beta_{2,5} = 0, \beta_{2,6} = \dots = \beta_{2,10} = 1,$$

$$\beta_{3,1} = \dots = \beta_{3,3} = -1, \beta_{3,4} = \dots = \beta_{3,7} = 0, \beta_{3,8} = \dots = \beta_{3,10} = 1.$$

Here the two vectors $\beta_{2,\cdot}$ and $\beta_{3,\cdot}$ for covariates x_2 and x_3 are heterogeneous, and vector $\beta_{1,\cdot}$ for covariate x_1 is homogeneous. The survival time $T_{k,i}$ is generated by $t_{k,i} = \epsilon_{k,i} - \beta_k^T x_{k,i}$ where $\epsilon_{k,i}$ is realized from a standard extreme value distribution, for $k = 1, \dots, K$, $i = 1, \dots, n$. To ensure that event times are nonnegative, we add $|\min_{k,i} t_{k,i}| + 0.5$ to the simulated values, which will not change the coefficient estimates and the baseline hazard evaluated at sample time points. The censoring time C is generated from a uniform distribution on the interval $[1/3, 1] * \max_{k,i} t_{k,i}$ to obtain a censoring proportion of about 40%. The three covariates are generated from a multivariate normal distribution, with mean zero, marginal variance 1 and exchangeable correlation 0.3. We allow all covariate effects to be heterogeneous across 10 subgroups with no prior knowledge of the clustering structure. Thus, all coefficients are fused simultaneously.

In the simulation results, for each $\hat{\beta}_j, j = 1, \dots, p$, we report the average number of distinct regression coefficient estimates across all replications, denoted by $df(\hat{\beta}_j)$. Given any coefficient vector $\hat{\beta}_j$, the total number of pairs is $\binom{K}{2}$ and the pairs can be split into the group of equal pairs or the group of distinct pairs. To measure the clustering performance of $\hat{\beta}_j$

with respect to the true β_j , we report both sensitivity and specificity, where sensitivity measures the proportion of correctly identified equal coefficient pairs and specificity measures the proportion of correctly identified unequal coefficient pairs. Also, we compare the mean square error (MSE) between $\hat{\beta}_j$ and β_j defined by

$MSE(\hat{\beta}_j) = \frac{1}{K} \sum_{k=1}^K (\hat{\beta}_{j,k} - \beta_{j,k})^2$, in the yielded model by our method ($\lambda = \lambda_{opt}$), the homogeneous model ($\lambda = \lambda_{fuse} := \max_j \lambda_{fuse,j}$) and the heterogeneous model ($\lambda = 0$). Additionally, to compare the overall estimation of the baseline cumulative hazard function, we report the MSE between $\log \hat{\Lambda}_0(t)$ and $\log \Lambda_0(t)$ at the sample time points, or

$\frac{1}{N} \sum_{k=1}^K \sum_{i=1}^{n_k} \{\log \hat{\Lambda}_0(t_{ki}) - \log \Lambda_0(t_{ki})\}^2$ where $N = \sum_k n_k$. The cumulative baseline is estimated by

$$\hat{\Lambda}_0(t) = \sum_{k=1}^K \sum_{i=1}^{n_k} \int_0^t \frac{dN_{i,j}(u)}{\sum_{l=1}^K \sum_{j=1}^{n_l} y_{l,j}(u) \exp(x_{l,j}^T \hat{\beta}_l)}$$

where t and u are only assessed at the observed sample points for simplicity [2]. The true cumulative baseline hazard function is $\Lambda_0(t) = \exp(t)$.

Table 1 lists the results of the first simulation study over 1000 round of simulations, where the optimal tuning parameter λ_{opt} is selected by either BIC or EBIC. We list the results from both the equal and unequal sample size settings. From Table 1, we can see the proposed method works well in terms of identifying the correct clustering, with high sensitivity and near perfect specificity. Specificity is not available for $\hat{\beta}_1$ since it contains no truly unequal pairs. The estimated frictions of fusion $\tilde{f}_1$, $\tilde{f}_2$ and $\tilde{f}_3$ for β_1 , β_2 and β_3 , respectively, are very effective to describing the heterogeneity level of the covariate effects across subgroups. The most heterogeneous vector β_3 exhibits the highest friction while the homogeneous vector β_1 accompanies a friction score close to zero. MSE results show that our method provides preferable estimation of β and Λ_0 than the other two naive models. It can be seen that the method also works very well under the unequal sample size setting. In this simulation study, model selection by BIC or EBIC yields very comparable results because $K = 10$ and $p = 3$ are relatively small in the study. However, the difference becomes larger in the next simulation study when we choose larger K and p .

3.2 Simulation II

The second simulation study designs a more challenging scenario with $K = 100$, $n_1 = \dots = n_K = 100$, and $p = 8$. In terms of data generation, we used the same mechanism as those used in the first simulation study except the number of covariates being increased to 8, namely $x_1, x_2, \dots, x_8$. To be specific, we set the first four vectors of coefficients, $\beta_1, \dots, \beta_4$, to be nonzero with the following true clustered effect patterns:

$$\beta_{1,1} = \dots = \beta_{1,50} = 0, \beta_{1,51} = \dots = \beta_{1,100} = 1,$$

$$\beta_{2,1} = \dots = \beta_{2,30} = -1, \beta_{2,31} = \dots = \beta_{2,70} = 0, \beta_{2,71} = \dots = \beta_{2,100} = 1,$$

$$\beta_{3,1} = \dots = \beta_{3,25} = -1, \beta_{3,26} = \dots = \beta_{3,50} = 0, \beta_{3,51} = \dots = \beta_{3,75} = 1, \beta_{3,76} = \dots = \beta_{3,100} = 2,$$

$$\beta_{4,1} = \dots = \beta_{4,20} = -2, \beta_{4,21} = \dots = \beta_{4,40} = -1, \beta_{4,41} = \dots = \beta_{4,60} = 0, \beta_{4,61} = \dots = \beta_{4,80} = 1, \beta_{4,81} = \dots = \beta_{4,100} = 2.$$

and the others set as $\beta_{5,\cdot} = \dots = \beta_{8,\cdot} = 0$. In this study, we impose the sparsity penalty by first setting $\alpha = 1$, and we measure the performance of variable selection by reporting the proportion of zeros in the estimates. The results are reported in the top part of Table 2. All metrics are reported based on 1000 replications. Similar to Simulation I, the proposed method with $\lambda = \lambda_{opt}$ has the smallest MSE. Since K is large, the model selected by EBIC clearly outperforms the one selected by BIC by achieving higher sensitivity, indicating that the proposed method based on the EBIC criterion encourages stronger fusion and improves the ability to detect equal coefficient pairs. Furthermore, EBIC demonstrates better sparsity detection among all 8 covariates, including both important ($\beta_1, \dots, \beta_4$) and nonimportant ones ($\beta_5, \dots, \beta_8$). As suggested from Table 2, EBIC appears robust and provides better model selection than the conventional BIC for the proposed fusion learning. Additionally, in the bottom part of Table 2, we show results at different values of α with models selected based on EBIC. Although we see a slight improvement in sensitivity and sparsity when α is large, results are not very sensitive to the choice of α . Thus for computational simplicity during practice, we suggest setting $\alpha = 1$ or tuning based on a sequence of rough grids. Additional simulation results for Simulation II are presented in the Appendix to study the change in efficiency of the proposed method with smaller and unequal sample sizes.

4 Application

In this section, we apply the proposed method to the analysis of vehicle contract renewal data that has been introduced in Section 1. As aforementioned, the goal of this data analysis is two-fold: (i) to identify important predictors for the time to contract renewal, (ii) to assess potential parameter heterogeneity among five different car models ($K = 5$). We fit the data by the following Cox model :

$$\lambda(T_{k,i}) = \lambda_0(T_{k,i}) \exp \left(\sum_{j=1}^4 \beta_{j,k} x_{j,k,i} \right), \quad k = 1, \dots, 5, i = 1, \dots, n_k,$$

where $x_{1,k,i}$ is the value of a vehicle, $x_{2,k,i}$ represents the allowed mileage limit, $x_{3,k,i}$ is a region indicator with “0” being the East and “1” being the West, and $x_{4,k,i}$ is another indicator for contract term with “0” representing short-term and “1” for long-term. The number of renewal contracts n_k and the descriptive statistics of the covariates for each of the

five car models is listed in Table 3. Event time $T_{k,i}$ is the number of months from the previous contract starting time to the renewal time, and the censoring time $C_{k,i}$ is the end of the study (June 30, 2015). Thus, the observed time to event is $y_{k,i} = \min(T_{k,i}, C_{k,i})$ with 84.8% of censored contracts. In order to achieve the study goals, we use the sparse fused lasso penalty function in (6) with $\alpha = 1$.

Table 4 lists the estimated coefficients from three methods: (I) the integrated analysis that fits all data regardless of vehicle models by the fully homogeneous Cox model, (II) the fully heterogeneous analysis that fits vehicle-specific covariate effects by the Cox model, and (III) our proposed fusion learning approach that fits the entire data by partially heterogeneous Cox model using the EBIC as model selection criterion. Since p -values are not available for our method, in (III), we report the variable selection consistency from 1000 bootstrap resampling. Figure 1 shows the solution paths of the proposed regularized estimates, with the vertical dotted line denoting the λ_{opt} selected by EBIC. The results for analysis III suggest that the western region has longer contract renewal time than the eastern region, and such difference is homogeneous across all models. This is probably due to milder weather in the West where vehicles are more durable. Interestingly, we also find that renewal time for model A, C, D and E is affected by contract type, with a fused coefficient equal to -0.106 , while renewal time for model B is independent of contract type. This implies that longer contract term for models A, C, D and E tends to the occurrence of a longer duration of the current contract. Vehicle value and mileage do not appear to be important on the probabilities of contract renewal across 5 vehicle models. Even though the effect of vehicle value is significant for some models but not all in the stratified analysis (II), as shown in Figure 1, the relative effects of vehicle value are smaller compared to those of region and term. Thus due to the dominance of other more predictive covariates, the entire group of coefficients for vehicle value are shrunk to zero. Our method provides more interpretable results as compared to the integrated homogeneous model based analysis and the heterogeneous model based analysis.

In addition, we report extra analysis results in Table 5 that lists the coefficient estimates from the Cox interaction model, including interactions between all covariates and vehicle model type. The parameter homogeneity or heterogeneity identified by our proposed model is largely in agreement with the hypothesis testing results drawn from the interaction-effects model. For example, the interaction between contract term and model B (with model A as baseline) is significant but not for models C, D and E. Overall, our method is conceptually more appealing and computationally easier, and more importantly, the results from our proposed method are more informative and interpretable than the interaction-effects model. It is worth mentioning that our method is scalable to a large-scale problem with big K as illustrated in the second simulation study, whereas the interaction-effects model will become extremely complicated and involves multiple comparison dilemma that impairs statistical power.

5 R Package

In this section, we give a brief introduction to the R package `metafuse`, which has been developed to provide researchers, teachers and statisticians a convenient open source tool for

fusion learning analysis. The R package can be downloaded from the Comprehensive R Archive Network at <https://cran.r-project.org/web/packages/metafuse>. The package supports fusion learning for linear, logistic, Poisson and Cox proportional hazards models. Here, we give an example of analyzing a time-to-event dataset with partially heterogeneous subgroup effects.

First we install and load the R package metafuse to the default directory:

```
> install.packages("metafuse")
> library(metafuse)
```

Next we simulate a dataset with $p = 3$ covariates, $K = 10$ subgroups, and each with a sample size of $n_k = 100$, $k = 1, \dots, 10$:

```
> n <- 100 # sample size in each subgroup
> K <- 10 # number of subgroups
> p <- 4 # number of covariates + 1
> beta0 <- matrix(c(0,0,0,0,0,0,0,0,0,0, # beta0 (ignored in Cox)
0,0,0,0,0,0,0,0,0,0, # beta1
0,0,0,0,0,1,1,1,1,1, # beta2
-1,-1,-1,0,0,0,0,1,1,1), # beta3
K,p)
> data <- datagenerator(n=n, beta0=beta0, family="cox", seed=123)
> dim(data)
[1] 1000 6
> colnames(data)
[1] "time" "status" "x1" "x2" "x3" "group"
```

Note that the dataset is simulated with partially heterogeneous regression coefficients across subgroups with clustering structures same as Simulation 1. The resulting dataset has $\sum_{k=1}^K n_k = 1000$ rows and six columns, where time and status are the observed time and event status indicator, respectively. When status=1, time is the event time, otherwise, time is the censoring time. The column group is a list of integers taking values from 1,..., 10, indicating the subgroup membership of each subject. Columns x1, x2 and x3 are the three covariates generated from $N(0,1)$ with exchangeable correlation of 0.3. We then prepare the inputs y, X and sid (subgroup indicator) and fit a fusion learning model using function metafuse():

```
> y <- data[c("time", "status")]
> X <- data[c("x1", "x2", "x3")]
> sid <- data[, "group"] # an integer array
> fit <- metafuse(X=X, y=y, sid=sid, fuse.which=c(1,2,3),
family="cox", intercept=FALSE, alpha=0,
criterion="EBIC", verbose=FALSE)
```

Finally, we can view the fitted coefficients, degree of freedom, tuning parameters $\tilde{\lambda}_{opt}, \tilde{\lambda}_{fuse, j}$ and friction of fusion $\tilde{f}_j$ for each covariate:

```
fit$betahat
x1 x2 x3
study1 0.05726975 -0.07776176 -0.8937963
study2 0.05726975 -0.07776176 -0.8937963
study3 0.05726975 -0.07776176 -0.8937963
study4 0.05726975 -0.07776176 -0.0396067
study5 0.05726975 -0.07776176 -0.0396067
study6 0.05726975 0.98017624 -0.0396067
study7 0.05726975 0.98017624 -0.0396067
study8 0.05726975 0.98017624 0.8088566
study9 0.05726975 0.98017624 0.8088566
study10 0.05726975 0.98017624 0.8088566
> fit$betainfo$log10_scale
x1 x2 x3
DF 1.00000000 2.00000000 3.00000000
lambda_opt 0.07108185 0.07108185 0.07108185
lambda_fuse 0.07108185 0.58187948 0.58187948
friction 0.00000000 0.87784094 0.87784094
```

An additional function `metafuse.l()` is provided to fit a fusion learning model when the λ_{opt} is given, which avoids fitting the entire solution path.

6 Concluding Remarks

This paper develops a computationally effective method to deal with partial data heterogeneity in the application of the Cox proportional hazards model for the survival data analysis. The proposed method allows the flexibility to explore the heterogeneity pattern in the regression coefficients as well as to identify the homogeneous clusters among large number of subgroups. By fusing parameters in a homogeneous group into one parameter we achieve parsimony of model, which specifically leads to higher statistical power and better interpretation. Furthermore, instead of using conventional BIC as the selection criterion, we find EBIC more suitable due to its better performance in encouraging sparsity. Based on the relative position of the optimal tuning parameter along solution paths, we propose a metric of friction coefficient to quantify the level of heterogeneity, which serves as a useful summary statistic in the fusion learning. Through applying our method in the real data from the vehicle leasing contract study, we find that longer contract term for the most of vehicle models appears to be associated with the probability of choosing a later lease contract renewal, and the western region has longer contract renewal time than the eastern region regardless of vehicle model.

It is worth pointing out that our proposed method is based on the assumption of a common baseline hazard function in the Cox model across different subgroups. When baseline

heterogeneity is present, a more flexible modeling framework is needed. One possible approach is to consider shared frailty across subgroup-specific baselines $\lambda_0^k(t)$ along with the idea in [14]. In addition, the method of confidence distribution, originally proposed by Fisher [10] and later formally formulated by Efron [8], may be utilized to achieve scalability of the proposed method for very large dataset based on fusion of summary statistics within the framework of divide-and-conquer algorithms. Furthermore, it is of interest to extend the proposed method to the situation where certain clustering structure is present across the subgroups, but the membership of individuals among the subgroups is not known *a priori*. This may be achieved by following ideas of either EM algorithm based approaches [1, 3] or fusion penalty based approaches [18]. Finally, it requires further improvement of the proposed method in the case of highly unbalanced settings where subgroups with small sample sizes are dominant. And post-selection inference is one of the directions to extend the current work in order to conduct accurate inference after variable-selection.

Acknowledgements

We are grateful to three anonymous reviewers for their valuable comments that have led to an improvement of this paper. This research is partially supported by the National Science Foundation DMS 1513595 and the National Institutes of Health R01 ES024732.

Appendix

Appendix: Additional Results for Simulation II

To better explore the change of efficiency due to finite and unbalanced sample sizes. We conduct a sensitivity analysis for Simulation II from Section 3. The parameter setting are the same as Simulation II, with a fixed $\alpha = 1$. The only difference is in the sample sizes across the $K = 100$ subgroups. We consider the case when the sample sizes are unequal, with a proportion c of randomly selected subgroups having smaller sample size $n_k = 50$. Others remain $n_k = 100$. Specifically, we consider $c = 0\%$, $c = 25\%$ and $c = 50\%$. Note that $c = 0\%$ is the scenario identical to the original simulation setting as in Table 2.

The simulation results for fusion learning based on EBIC are displayed in Table 6. Results are averaged across 100 replications. As the percentage of small subgroups c increases, we see a decline in clustering performance as reflected by $df(\hat{\beta})$, sensitivity and sparsity. The decrease in efficiency may be due to the higher variability in the initial estimated ordering. However, this decline is not as steep when c changes from 0% to 25% than from 25% to 50%. This implies that as long as the proportion of small subgroups is not too large, the performance of clustering is still guaranteed. Additionally, estimation performance as captured by MSE remains small for $\hat{\beta}$ and $\hat{\lambda}(t)$, despite a small increase with respect to the increase of c . In summary, this simulation demonstrate that our method is robust against mixture of large and small subgroups, as long as the proportion of the smaller subgroups is not very large.

References

1. Aitkin M (1996) A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and computing* 6(3):251–262
2. Andersen PK, Borgan O, Gill RD, Keiding N (2012) *Statistical models based on counting processes*. Springer Science & Business Media
3. Azzimonti L, Ieva F, Paganoni AM (2013) Nonlinear nonparametric mixed-effects models for unsupervised classification. *Computational Statistics* 28(4):1549–1570
4. Bonetti M, Gelber RD, et al. (2000) A graphical method to assess treatment-covariate interactions using the Cox model on subsets of the data. *Statistics in Medicine* 19(19):2595–2609 [PubMed: 10986536]
5. Chen J, Chen Z (2008) Extended bayesian information criteria for model selection with large model spaces. *Biometrika* 95(3):759–771
6. Cheng X, Lu W, Liu M (2015) Identification of homogeneous and heterogeneous variables in pooled cohort studies. *Biometrics* 71(2):397–403 [PubMed: 25732747]
7. Cox DR (1972) Regression models and life tables (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 34(2):187–220
8. Efron B (1993) Bayes and likelihood calculations from confidence intervals. *Biometrika* 80(1):3–26
9. Egger M, Smith GD, Schneider M, Minder C (1997) Bias in metaanalysis detected by a simple, graphical test. *BMJ: British Medical Journal* 315(7109):629–634 [PubMed: 9310563]
10. Fisher RA (1956) *Statistical methods and scientific inference*.
11. Friedman J, Hastie T, Hofling H, Tibshirani R (2007) Pathwise coordinate optimization. *The Annals of Applied Statistics* 1(2):302–332
12. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1):1 [PubMed: 20808728]
13. Gao X, Song PXX (2010) Composite likelihood bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association* 105(492):1531–1540
14. Gasperoni F, Ieva F, Paganoni A, Jackson C, Sharples L (2017) Nonpara-metric shared frailty model for classification of survival data In: *Proceedings of the conference “Statistics and Data Science: new challenges, new generations”*, pp 451–456
15. Hakulinen T, Tenkanen L (1987) Regression analysis of relative survival rates. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 36(3):309–317
16. Higgins J, Thompson SG (2002) Quantifying heterogeneity in a metaanalysis. *Statistics in Medicine* 21(11):1539–1558 [PubMed: 12111919]
17. Klein JP (1992) Semiparametric estimation of random effects using the cox model based on the EM algorithm. *Biometrics* 48(3):795–806 [PubMed: 1420842]
18. Ma S, Huang J (2017) A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association* 112(517):410–423
19. Nielsen GG, Gill RD, Andersen PK, Sørensen TI (1992) A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics* 19(1):25–43
20. Tang L, Song PXX (2016) Fused lasso approach in regression coefficients clustering-learning parameter heterogeneity in data integration. *Journal of Machine Learning Research* 17(113):1–23
21. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K (2005) Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(1):91–108
22. Wang F, Wang L, Song PXX (2016) Fused lasso with the adaptation of parameter ordering in combining multiple studies with repeated measurements. *Biometrics* 72(4):1184–1193 [PubMed: 26909642]
23. Zou H (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101(476):1418–1429

Fig. 1: Solution paths of the proposed regularized estimator of regression coefficients. The dotted vertical line denotes the optimal tuning parameter λ_{ebic} by EBIC.

Table 1:

Summary result of the first simulation study for the proposed fusion learning method with BIC and EBIC as model selection criteria, respectively.

Criterion ($\tilde{\lambda}_{opt}$)		$df(\hat{\beta})$	$\tilde{\lambda}_{fuse, j}$	$\tilde{f}_j$	Sensitivity	Specificity	$MSE \text{ when } \lambda =$		
							λ_{opt}	λ_{fuse}	0
Equal sample sizes									
BIC (0.113)	β_1	1.119	0.086	0.059	0.95	-	0.003	0.002	0.027
	β_2	2.188	0.758	0.870	0.96	1.00	0.007	0.283	0.034
	β_3	3.229	0.853	0.886	0.95	1.00	0.011	0.603	0.035
	$\log \Lambda_0(t)$	-	0.021	0.184	0.026				
EBIC (0.116)	β_1	1.099	0.086	0.051	0.96	-	0.003	0.002	0.027
	β_2	2.167	0.757	0.866	0.96	1.00	0.007	0.283	0.034
	β_3	3.206	0.853	0.883	0.95	1.00	0.011	0.603	0.035
	$\log \Lambda_0(t)$	-	0.021	0.184	0.026				
Unequal sample sizes									
BIC (0.128)	β_1	1.269	0.113	0.124	0.91	-	0.008	0.003	0.060
	β_2	2.409	0.645	0.819	0.89	0.99	0.023	0.255	0.068
	β_3	3.513	0.773	0.854	0.88	0.99	0.029	0.647	0.086
	$\log \Lambda_0(t)$	-	0.026	0.199	0.039				
EBIC (0.132)	β_1	1.238	0.113	0.112	0.92	-	0.008	0.003	0.060
	β_2	2.378	0.645	0.812	0.90	0.99	0.023	0.255	0.068
	β_3	3.474	0.773	0.849	0.89	0.99	0.029	0.647	0.086
	$\log \Lambda_0(t)$	-	0.026	0.199	0.039				

Table 2: Summary results of the second simulation study with BIC and EBIC as model selection criteria, respectively.

Criterion ($\tilde{\lambda}_{opt}$)	$df(\hat{\beta})$	$\tilde{\lambda}_{fuse,j}$	$\tilde{f}_j$	Sensitivity	Specificity	Sparsity	$MSE \text{ when } \lambda =$			
							λ_{opt}	λ_{fuse}	0	
BIC (0.183)	Equal fusion and sparsity ($\alpha = 1$)									
	β_1	5.990	1.921	0.911	0.59	1.00	0.30	0.004	0.363	0.040
	β_2	8.089	1.866	0.909	0.61	1.00	0.24	0.004	0.601	0.042
	β_3	9.843	2.135	0.920	0.64	1.00	0.16	0.005	1.374	0.060
	β_4	11.451	2.120	0.919	0.67	1.00	0.13	0.005	2.001	0.070
	β_5	3.637	0.255	0.375	0.52	-	0.54	0.003	0.000	0.030
	β_6	3.627	0.255	0.373	0.53	-	0.54	0.003	0.000	0.030
	β_7	3.632	0.253	0.372	0.53	-	0.54	0.003	0.000	0.030
EBIC (0.256)	β_8	3.635	0.254	0.377	0.52	-	0.54	0.003	0.000	0.030
	$\log \Lambda_1(\ell)$	-	-	-	-	-	-	0.007	3.081	0.142
	β_1	4.167	1.920	0.874	0.77	1.00	0.39	0.003	0.363	0.040
	β_2	5.801	1.866	0.871	0.78	1.00	0.31	0.003	0.601	0.042
	β_3	7.109	2.135	0.886	0.80	1.00	0.20	0.004	1.374	0.061
	β_4	8.347	2.120	0.886	0.82	1.00	0.16	0.004	2.002	0.070
	β_5	2.356	0.254	0.208	0.73	-	0.74	0.002	0.000	0.030
	β_6	2.391	0.256	0.207	0.73	-	0.74	0.002	0.000	0.030
EBIC (0.256)	β_7	2.418	0.254	0.210	0.72	-	0.74	0.002	0.000	0.030
	β_8	2.379	0.255	0.208	0.73	-	0.74	0.002	0.000	0.030
	$\log \Lambda_1(\ell)$	-	-	-	-	-	-	0.008	3.081	0.142

Criterion ($\tilde{\lambda}_{opt}$)	$df(\hat{\beta})$	$\tilde{\lambda}_{fuse, j}$	$\tilde{f}_j$	Sensitivity	Specificity	Sparsity	MSE when $\lambda =$		
							λ_{opt}	λ_{fuse}	0
β_6	2.377	0.248	0.208	0.73	-	0.74	0.002	0.000	0.030
β_7	2.411	0.246	0.207	0.73	-	0.74	0.002	0.000	0.030
β_8	2.365	0.247	0.208	0.73	-	0.75	0.002	0.000	0.030
$\log \wedge_1(\theta)$	-	-	-	-	-	-	0.008	3.081	0.142
Less fusion more sparsity ($\alpha = 4$)									
β_1	4.052	1.945	0.869	0.78	1.00	0.40	0.003	0.363	0.040
β_2	5.627	1.888	0.866	0.79	1.00	0.31	0.003	0.601	0.042
β_3	6.934	2.156	0.882	0.82	1.00	0.20	0.004	1.374	0.060
β_4	8.161	2.142	0.881	0.83	1.00	0.16	0.004	2.001	0.070
β_5	2.281	0.262	0.194	0.74	-	0.75	0.002	0.000	0.030
β_6	2.299	0.264	0.193	0.75	-	0.76	0.002	0.000	0.030
β_7	2.295	0.261	0.195	0.74	-	0.76	0.002	0.000	0.030
β_8	2.291	0.263	0.195	0.75	-	0.76	0.002	0.000	0.030
$\log \wedge_1(\theta)$	-	-	-	-	-	-	0.008	3.081	0.142

Table 3:

Descriptive statistics of the covariates by car models. Means and standard deviations are reported for continuous covariates; counts are reported for binary covariates.

Model	Sample size n_k	Event	Value (\$1,000)	Mileage (mpg)	Region (0/1)	Term
A	2197	221	22.2 ± 3.1	27.8 ± 8.0	1076/1121	1462/735
B	5566	997	28.5 ± 3.6	26.8 ± 7.4	3772/1794	4220/1346
C	5997	945	30.4 ± 3.3	27.3 ± 7.6	4210/1787	4286/1711
D	5288	840	38.4 ± 4.0	27.7 ± 8.1	2765/2523	3840/1448
E	2432	261	40.9 ± 5.5	31.6 ± 9.0	1341/1091	1107/1325

Table 4:

Coefficient estimates from three different methods, the integrated analysis without differentiating the car models, the car model-specific effect analysis, and our proposed fusion approach using EBIC, respectively. For Model I and II, numbers in brackets denote p-values. For Model III, numbers in brackets denote selection stability (percentage of zero estimates from bootstrap).

Model	$\hat{\beta}_{value}$	$\hat{\beta}_{mileage}$	$\hat{\beta}_{region}$	$\hat{\beta}_{term}$
(I) Integrated Analysis (full homogeneity)				
A-E	0.015 (0.422)	0.007 (0.813)	-0.663 (< 0.001)	-0.160 (0.011)
(II) Car Model-Specific Analysis (full heterogeneity)				
A	0.324 (0.016)	0.101 (0.331)	-0.735 (< 0.001)	-1.981
B	0.183 (0.004)	-0.054 (0.325)	-0.554 (< 0.001)	0.178 (0.121)
C	0.057 (0.432)	0.039 (0.476)	-0.786 (< 0.001)	-0.028 (0.814)
D	-0.063 (0.288)	-0.017 (0.763)	-0.557 (< 0.001)	-0.193 (0.141)
E	-0.003 (0.971)	0.083 (0.320)	-0.997 (< 0.001)	-3.974
(III) Fusion Learning Approach (partial heterogeneity)				
A	0.000 (0.937)	-0.506 (0.972)	-0.106 (0.009)	(0.651)
B	0.000 (0.936)	-0.506 (0.981)	0.000 (0.009)	(0.957)
C	0.000 (0.959)	-0.506 (0.968)	-0.106 (0.008)	(0.706)
D	0.000 (0.995)	-0.506 (0.979)	-0.106 (0.008)	(0.736)
E	0.000 (0.985)	-0.506 (0.971)	-0.106 (0.008)	(0.498)
$\tilde{f}_j$	0.000	0.348		

Table 5:

Coefficient estimates obtained from the interaction Cox model.

Effect	$\hat{\beta}$	$\widehat{sd}(\hat{\beta})$	p-value
Value	0.320	0.134	0.017
Mileage	0.106	0.104	0.307
Region	-0.725	0.143	0.000
Term	-0.401	0.233	0.085
Model B	-0.298	0.239	0.211
Model C	-0.449	0.238	0.059
Model D	-0.248	0.245	0.311
Model E	-0.336	0.275	0.223
Value X Model B	-0.126	0.148	0.395
Value X Model C	-0.259	0.152	0.089
Value X Model D	-0.383	0.146	0.009
Value X Model E	-0.320	0.156	0.040
Mileage X Model B	-0.162	0.117	0.168
Mileage X Model C	-0.067	0.118	0.571
Mileage X Model D	-0.126	0.119	0.289
Mileage X Model E	-0.014	0.133	0.918
Region X Model B	0.182	0.163	0.263
Region X Model C	-0.059	0.169	0.729
Region X Model D	0.174	0.160	0.279
Region X Model E	-0.205	0.203	0.313
Term X Model B	0.571	0.260	0.028
Term X Model C	0.362	0.260	0.164
Term X Model D	0.166	0.267	0.533
Term X Model E	-0.202	0.304	0.507

Table 6:

Summary results of the second simulation study with composition of small subgroups with sample size $nk = 50$ and large subgroups with $nk = 100$. Model selection is based on EBIC.

	% groups with $n_k = 50$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\log \Lambda_1(\hat{\theta})$
$df(\hat{\beta})$	0%	4.319	6.000	7.681	8.638	2.319	2.425	2.511	2.468	-
	25%	5.558	7.465	8.954	10.442	3.093	3.279	3.326	3.326	-
	50%	8.593	11.093	12.759	14.574	5.056	5.241	4.981	5.037	-
Sensitivity	0%	0.741	0.767	0.779	0.807	0.725	0.716	0.689	0.683	-
	25%	0.655	0.681	0.727	0.747	0.666	0.605	0.631	0.625	-
	50%	0.413	0.454	0.517	0.531	0.392	0.377	0.380	0.381	-
Specificity	0%	1.000	1.000	1.000	1.000	-	-	-	-	-
	25%	0.998	0.999	0.999	0.999	-	-	-	-	-
	50%	0.998	0.997	0.998	0.999	-	-	-	-	-
Sparsity	0%	0.377	0.315	0.198	0.154	0.749	0.725	0.681	0.686	-
	25%	0.345	0.281	0.182	0.146	0.702	0.653	0.677	0.662	-
	50%	0.227	0.185	0.128	0.106	0.419	0.418	0.411	0.420	-
MSE	0%	0.003	0.003	0.004	0.004	0.002	0.002	0.002	0.002	0.007
	25%	0.008	0.008	0.009	0.009	0.004	0.004	0.003	0.003	0.009
	50%	0.017	0.020	0.020	0.022	0.008	0.010	0.008	0.008	0.014