Published in final edited form as:

Biometrics. 2015 June ; 71(2): 397–403. doi:10.1111/biom.12285.

Identification of Homogeneous and Heterogeneous Variables in Pooled Cohort Studies

Xin Cheng[1,*], **Wenbin Lu**[2,**], and **Mengling Liu**[1,***]

[1]Departments of Population Health and Environmental Medicine, New York University School of Medicine, New York, U.S.A

[2]Department of Statistics, North Carolina State University, Raleigh, North Carolina, U.S.A.

Summary

Pooled analyses integrate data from multiple studies and achieve a larger sample size for enhanced statistical power. When heterogeneity exists in variables' effects on the outcome across studies, the simple pooling strategy fails to present a fair and complete picture of the effects of heterogeneous variables. Thus, it is important to investigate the homogeneous and heterogeneous structure of variables in pooled studies. In this paper, we consider the pooled cohort studies with time-to-event outcomes and propose a penalized Cox partial likelihood approach with adaptively weighted composite penalties on variables' homogeneous and heterogeneous effects. We show that our method can characterize the variables as having heterogeneous, homogeneous, or null effects, and estimate non-zero effects. The results are readily extended to high-dimensional applications where the number of parameters is larger than the sample size. The proposed selection and estimation procedure can be implemented using the iterative shooting algorithm. We conduct extensive numerical studies to evaluate the performance of our proposed method and demonstrate it using a pooled analysis of gene expression in patients with ovarian cancer.

Keywords

Adaptive group lasso; Cox proportional hazard model; Heterogeneity; Penalized partial likelihood; Pooled analysis; Structure identification

1. Introduction

Pooled studies can achieve a large sample size and facilitate investigations on rare diseases, rare exposures, and topics not easily addressed in a single study. For example, Ganzfried et al. (2013) made a concerted effort to create a curated database consisting of clinical and microarray gene expression data on 2970 ovarian cancer patients from 23 studies using 11 gene expression measurement platforms. This pooled database empowers researchers to

[*] xc311@nyu.edu
[**] lu@stat.ncsu.edu
[***] mengling.liu@nyu.edu

Supplementary Materials

Web Appendix referenced in Section 2.3 and R code to reproduce the analyses in Section 3.2 are available with this paper at the Biometrics website on Wiley Online Library.

investigate the prognostic effect of genetic biomarkers on ovarian cancer survival in a uniform and consistent fashion. But as seen in this and many other pooled studies, inter-study heterogeneity in the association between the biomarkers and the outcome often exists and its source includes differences in study populations, sampling methods, disease ascertainments, and measurement methods. Although harmonizing the data can alleviate this issue (Ganzfried et al., 2013), heterogeneity often is inherent in pooled studies. In some studies, heterogeneity itself is important for understanding disease disparity and progression at different phases (Moreno et al., 1996). Therefore, the analysis of pooled studies needs to properly account for heterogeneity to yield meaningful results.

To estimate covariate effects in pooled studies, the two-step procedure is commonly used, in which study-specific effects are first estimated using individual study data and then combined using a fixed-effects model (Hedges and Olkin, 1985) or a random-effects model (DerSimonian and Laird, 1986). However, the two-step procedure has difficulty in handling multiple variables. If heterogeneity exists in variables' effects across studies we want to distinguish the variables with heterogeneous effects versus those without, as the effect of a homogeneous predictor should be modeled using a common parameter across studies to reduce model complexity and improve efficiency, while heterogeneous effects should be modeled by distinct parameters for different studies to build accurate models. Methods for discovering heterogeneity in variable's effects include examining its interactions with the study-membership indicator variables or the heterogeneity statistics such as Cochran's Q and I^2, but these often have low power especially when the number of studies is small and the number of predictors is large (Hedges and Olkin, 1985).

In this paper, we consider the heterogeneity issue in pooled studies with time-to-event endpoint, and formulate the problem in the framework of group variable selection. Specifically, we treat a variable's effects across studies as a group and aim to classify the variables into three categories according to their effects: homogeneous, heterogeneous, and null. Group penalty regularized methods have been proposed to select variables with pre-specified group structure (Kim et al., 2012; Ma et al., 2007), and the composite absolute penalties (CAP) of Zhao et al. (2009) could accommodate complex group structures. Recently, Liu et al. (2013) investigated the use of adaptive CAP regularized partial likelihood estimation in the context of pooled nested case-control studies with heterogeneity. These methods, however, cannot delineate variables into the desirable categories.

Inspired by some recent developments in structure identification in the partially linear model (Zhang et al., 2011) and the time-varying Cox model (Yan and Huang, 2012), we employ the adaptively weighted L_1 and L_1/L_2 penalties on variables' average effects and heterogeneous effects respectively, and propose a penalized partial likelihood approach to characterize the variables' heterogeneity and simultaneously estimate variables' effects. We establish asymptotic results for the proposed estimator when the number of parameters is fixed and also when it diverges with the sample size. The rest of the article is organized as follows. We introduce the composite $L_1 + L_1/L_2$ penalty regularized partial likelihood approach and the computation algorithm in Section 2. We also establish the theoretical properties of our proposed estimator, with the proofs provided in the Web Appendix.

Section 3 contains numerical simulations and real study applications in the pooled ovarian cancer study. Section 4 gives concluding remarks.

2. Regularized Method for Identifying Homogeneous and Heterogeneity Variables

2.1 Penalized partial likelihood function

We consider a pooled study consisting of K sub-studies, with n_k subjects in study k and $\sum_{k=1}^{K} n_k = n$. Let T_{ki}^{*} and C_{ki} be the failure time and censoring time for the ith subject in study k. Define the observed event time $T_{ki} = \min(T_{ki}^{*}, C_{ki})$, and the occurrence indicator of the failure event $\delta_{ki} = I(T_{ki}^{*} < C_{ki})$. We assume a Cox proportional hazards model for T_{ki}^{*}:

$$\lambda_k\{t|Z_{ki}\} = \lambda_{0k}(t)\ \exp\{\beta_k^{'} Z_{ki}\}, k=1,\ldots,K, \quad (1)$$

where $\lambda_{0k}(\cdot)$ is the baseline hazard function, and $\beta_k = (\beta_{k1}, \ldots, \beta_{kp})'$ is a $p \times 1$ vector characterizing the effects of covariates Z in study k. For the pooled study data under model (1), the log partial likelihood is expressed as

$$\ell(\beta_1,\ldots,\beta_K) = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\delta_{ki}\left[\beta_k^{'} Z_{ki} - \log\{\sum_{j=1}^{n_k} I(T_{kj} >= T_{ki})\ \exp(\beta_k^{'} Z_{kj})\}\right].$$

To separate out homogeneous and heterogeneous effects, we reformulate model (1) into

$$\lambda_k\{t|Z_{ki}\} = \lambda_{0k}(t)\ \exp(\mu' Z_{ki} + (\beta_k - \mu)' Z_{ki}\} \triangleq \lambda_{0k}(t)\ \exp\{\mu' Z_{ki} + \alpha_{k\cdot}^{'} Z_{ki}\},$$

where $\mu = (\mu_1, \ldots, \mu_p)'$ denotes the average effects and $\alpha_{k\cdot} = (\alpha_{k1}, \ldots, \alpha_{kp})'$ denotes the deviance of effects in study k from the average effects μ. To accommodate the constraint that, for each covariate l, $l = 1, \ldots, p$, $\sum_{k=1}^{K} \alpha_{kl} = 0$, we denote $\alpha_l = (\alpha_{2l}, \ldots, \alpha_{Kl})'$ and work with $\theta = (\mu', \alpha_1^{'}, \ldots, \alpha_p^{'})'$. We classify p predictors into three mutually exclusive categories: (1) homogeneous effects if $\mu_l \neq 0$ and the Euclidean norm of α_l: $\|\alpha_l\| = 0$; (2) heterogeneous effects if $\|\alpha_l\| \neq 0$; (3) null effects if $\mu_l = 0$ and $\|\alpha_l\| = 0$. To estimate this homogeneous and heterogeneous structure, we propose the following composite penalty regularized partial likelihood estimator

$$\hat{\theta} = \arg\ \min Q_n(\theta) = \arg\ \min\left\{-\ell(\theta) + \lambda_{1n}\sum_{l=1}^{p}\omega_{0l}|\mu_l| + \lambda_{2n}\sum_{l=1}^{p}\omega_{1l}\|\alpha_l\|\right\}, \quad (2)$$

where ω_{0l} and ω_{1l} are data-dependent weights and here chosen as $\omega_{0l} = 1/|\tilde{\mu}_l|$, $\omega_{1l} = 1/\|\tilde{\alpha}_l\|$, where $(\tilde{\mu}_l, \tilde{\alpha}_l)$ are some initial root-n consistent estimators.

2.2 Computation algorithm

We use the iterative shooting algorithm (Fu, 1998; Zhang and Lu, 2007) to minimize $Q_n(\theta)$ in (2). Let $G = -\partial\ell/\partial\theta$, $H = -\partial^2\ell/\partial\theta\partial\theta'$, and $X'X$ be the Cholesky decomposition of H. By defining a pseudo response vector $Y = (X')^{-1}\{H\theta - G\}$, we can approximate $-\ell(\theta)$ by a quadratic form $\frac{1}{2}(Y-X\theta)'(Y-X\theta)$. Furthermore, we consider the L_1 norm as a special case of the Euclidean norm with one element and rewrite the composite penalty terms in (2) as an adaptive group lasso problem with $2p$ groups

$$\frac{1}{2}(Y-X\theta)'(Y-X\theta)+\sum_{g=1}^{2p}\lambda_g\|\theta_g\|, \quad (3)$$

where $\theta_g = \mu_g$ and $\lambda_g = \lambda_{1n}\omega_{0g}$ for $g = 1, \ldots, p$ and $\theta_g = \alpha_g$ and $\lambda_g = \lambda_{2n}\omega_{1g}$ for $g = (p+1), \ldots, 2p$.

By the Karush–Kuhn–Tucker condition, Yuan and Lin (2006) showed that the necessary and sufficient condition for θ to be a solution of (3) is

$$-X_g'(Y-X\theta)+\lambda_g\theta_g/\|\theta_g\|=0, \theta_g \neq 0; \quad (4)$$

$$\|-X_g'(Y-X\theta)\| \leq \lambda_g, \theta_g=0. \quad (5)$$

Note that the condition (4) is equivalent to

$$S_g=(X_g'X_g+\lambda_g\theta_g/\|\theta_g\|I_{d_g})\theta_g, \quad (6)$$

where $S_g=X_g'(Y-X\theta_{-g})$, with $\theta_{-g}=(\theta_1',\ldots,\theta_{g-1}',0,\theta_{g+1}',\theta_{2p}')'$, I_{d_g} is the identity matrix of dimension d_g, and d_g is the number of parameters in group g. Thus, our shooting algorithm is:

1. Initialize with $\theta^{(0)}$.
2. For each $g = 1, \ldots, 2p$, if $\theta_g = 0$, then it stays at 0. Otherwise, update θ_g with

$$\theta_g=\begin{cases} (X_g'X_g+\lambda_g/\|\theta_g\|)^{-1}S_g, & \text{if}\|S_g\|>\lambda_g; \\ 0, & \text{if}\|S_g\| \leq \lambda_g. \end{cases}$$

3. Update θ with the new θ_g, and repeat until convergence.

We choose the tuning parameters λ_{1n} and λ_{2n} over a two-dimensional grid by minimizing the Bayesian information criterion (BIC),

$$\text{BIC}(\hat\theta)=-\ell(\hat\theta)+\log(n)\times df,$$

where $\hat{\theta}=(\hat{\mu}_l^{\prime},\hat{\alpha}_l^{\prime})^{\prime}$ is a minimizer of (2) under λ_{1n} and λ_{2n}, and the degree of freedom (df) is defined following Yuan and Lin (2006), $df=\sum_{l=1}^{p}I(|\hat{\mu}_l|>0)+\sum_{l=1}^{p}I(\|\hat{\alpha}_l\|>0)+(K-2)\sum_{l=l}^{p}\|\hat{\alpha}_l\|/\|\tilde{\alpha}_l\|$. The BIC-based tuning parameter selection corresponds to maximizing the posterior probability of selecting the true model and has been shown to be consistent for model selection in various settings (Wang et al., 2007; Zhang et al., 2010). Our simulation experiences also support the use of the BIC-based selection method.

Following Fan and Li (2001), we estimate the covariance of $\tilde{\theta}$ by

$$\widehat{\mathrm{COV}}(\hat{\theta})=\left\{(H+\Sigma)^{-1}H(H+\Sigma)^{-1}\right\}|_{\hat{\theta}}, \quad (7)$$

where $\Sigma = \mathrm{diag}\{\lambda_g/\|\theta_g\|I_{d_g}\}_{g=1,\ldots,2p}$.

2.3 Theoretical properties

Denote the true parameters by θ_n^{*}, where we use the general notation with the subscript n to allow the number of parameters (p_n) to go to infinity as the sample size n increases. Let $q_n = Kp_n$ be the total number of parameters. Define $\mathscr{A}_{1n} = \{l : \mu_l \neq 0, l = 1, \ldots, p_n\}$, $\mathscr{A}_{2n} = \{l : \|\alpha_l\| \neq 0, l = 1, \ldots, p_n\}$, and $\mathscr{A}_n = \mathscr{A}_{1n} \cup \mathscr{A}_{2n}$. The total number of parameters corresponding to $\mathscr{A}_n$ is $s_n = |\mathscr{A}_{1n}| + (K - 1) \times |\mathscr{A}_{2n}|$. Under the regularity conditions specified in Web Appendix, the following asymptotic properties hold.

Theorem 1 (Estimation Consistency)—*If* $\lambda_{1n}/\sqrt{n}\rightarrow 0, \lambda_{2n}/\sqrt{n}\rightarrow 0$, *and* $q_n^4/n\rightarrow 0$, *then* $\|\hat{\theta}_n-\theta_n^{*}\|=O_p\{(n/q_n)^{-1/2}\}$.

Theorem 2 (Selection Consistency)—*If* $\lambda_{1n}/q_n \rightarrow \infty$ *and* $\lambda_{2n}/q_n \rightarrow \infty$, *then* $P(\hat{\theta}_{\mathscr{A}_n^c}=0)\rightarrow 1$.

Because the dimension of $\hat{\theta}_{\mathscr{A}_n}$ may diverge as sample size goes to infinity, for asymptotic normality property below, we consider its arbitrary linear combination $B_n\hat{\theta}_{\mathscr{A}_n}$, where B_n is an arbitrary $m \times s_n$ matrix with a finite m and $B_nB_n^{\prime}\rightarrow G$ and G is positive-definite.

Theorem 3 (Asymptotic Normality)—*If* $\lambda_{1n}/\sqrt{n/g_n}\rightarrow 0, \lambda_{2n}/\sqrt{n/g_n}\rightarrow 0$, $\lambda_{1n}/q_n \rightarrow \infty$, $\lambda_{2n}/q_n \rightarrow \infty$, *and* $q_n^4/n\rightarrow 0$, *then*

$$\sqrt{n}B_nI_{\mathscr{A}_n}^{1/2}(\hat{\theta}_{\mathscr{A}_n}-\theta_{\mathscr{A}_n}^{*})\rightarrow_d N(0,G),$$

where $I_{\mathscr{A}_n}$ *is the Fisher information matrix corresponding to* $\theta_{\mathscr{A}_n}^{*}$.

Therefore, as the sample size goes to infinity, the proposed estimator $\hat{\theta}_n$ in (2) can perform as well as the correct model when that correct model is known in advance. Proofs are given in Web Appendix.

3. Numerical Studies

3.1 Simulations

We conducted simulations to evaluate the performance of our method under practical settings and compared it with four methods including the maximum likelihood estimation (MLE) method, the two-step method (two-step) with the random-effects model (DerSimonian and Laird, 1986), and the penalized partial likelihood methods with adaptive group Lasso (agLASSO) penalty (Yuan and Lin, 2006; Kim et al., 2012) and adaptive composite absolute (aCAP) penalty (Zhao et al., 2009; Liu et al., 2013). For the MLE method, we obtained the study-specific effects and checked whether the average effect and the heterogeneous effects of each variable were significantly different from 0. In the two-step method, we used Cochran's Q-test to examine the heterogeneity. The agLASSO method imposed a weighted L_2 penalty on the group of coefficients consisting of each variable's effects across studies. The aCAP approach imposed a weighted composite penalty as $\lambda_{1n}\sum_{l=1}^{p}\omega_{0l}\|(\mu_l,\alpha_l)\|+\lambda_{2n}\sum_{l=1}^{p}\omega_{1l}\|\alpha_l\|$.

3.1.1 Example 1: fixed number of covariates—

We first considered a pooled study with 3 sub-studies of size N = 150 or 300, and generated 11 covariates from a multivariate normal distribution with mean 0 and covariance cov(z_i, z_j) = $0.5^{|i-j|}$. The survival times were generated using the Cox proportional hazards model (1) with the Weibull distribution (shape= 10, scale = 1) determining the baseline hazard. The true θ* was specified as follows:

$$\theta^{*}=\begin{pmatrix}\mu\\ \alpha_{2\cdot}^{\prime}\\ \alpha_{3\cdot}^{\prime}\end{pmatrix}=\left(\begin{array}{ccccccccccc} Z_1 & Z_2 & Z_3 & Z_4 & Z_5 & Z_6 & Z_7 & Z_8 & Z_9 & Z_{10} & Z_{11}\\ 0.6 & 0.4 & 0.6 & 0 & 0 & 0 & 0.6 & 0 & 0 & 0 & 0\\ 0 & -0.4 & -0.5 & 0 & 0 & 0 & 0 & 0.4 & 0 & 0 & 0\\ 0 & -0.4 & 0.4 & 0 & 0 & 0 & 0 & -0.5 & 0 & 0 & 0 \end{array}\right).$$

Thus, covariates Z_1 and Z_7 had homogenous effects, Z_2, Z_3, and Z_8 had heterogeneous effects, and the rest were null variables. Censoring times were generated from the uniform distributions [0, 1.28] or [0, 1.77] to yield event rates around 25% or 45%.

Tables 1 – 4 report the simulation results. Table 1 reports the average numbers of correctly and incorrectly selected homogeneous and heterogeneous variables over 200 simulations. We use the square root of mean square errors $(E\|\hat{\theta}-\theta^*\|^2)^{1/2}$ (rMSE) to measure the estimation error. Our method performs very well for identifying the correct structure in all scenarios and outperforms all other methods in terms of having the smallest rMSE. Its overall performance improves with increasing sample size. When the sample size increases to 300, with an event rate of 45%, the average number of correctly identified homogeneous variables by the proposed method is 1.96, and the average number of correctly identified heterogeneous variables is 3. The agLASSO method is only capable of identifying the non-zero group of coefficients, and cannot differentiate homogenous from non-homogenous effects. The random-effects model gives a large estimation error as the method cannot estimate the study-specific effects for each variable.

Table 2 presents the detection frequencies of each variable's average and heterogeneous effects under the scenario with 45% event rate and $N = 300$ in each study. Our method identifies all variables' structure with good accuracy. The agLASSO method cannot differentiate between the average effect and heterogeneous effects, and always selects the variable as long as it has at least one non-zero effect, e.g. Z_1 and Z_7. The aCAP also shows a reasonable performance except for variable Z_8, which has zero mean effect but non-zero heterogeneous effects. The two-step method fails to identify the non-zero average effect of variable Z_2, because Z_2 has a small average effect size, and the heterogeneity of the effects of Z_2 is captured with a large variance estimated by the random-effects model. Because of the collinearity between variables, the MLE method does not perform well for variables with null effects. Similar results are observed in other scenarios.

Table 3 presents the estimated standard errors for covariates Z_1, Z_2, and Z_8 based on the sandwich formula (7), and compares them with the sample standard deviations calculated from 200 iterations. The empirical variance estimates and asymptotic estimates show some discrepancies in these finite-sample settings, but the overall performance improves with the sample size.

3.1.2 Example 2: diverging number of covariates p > N—We still considered pooling 3 studies, with each study size of $N = 200$. The number of covariates was set to be $p = 250$ or 450, and the covariate vector was generated from the multivariate normal distribution with mean 0 and covariance $\mathrm{cov}(z_i, z_j) = \rho^{|i-j|}$, $\rho = 0.5$ or 0.75. The survival times were generated from the Cox proportional hazards model with a constant baseline hazard function $\lambda_0(t) = 0.1$ and the true coefficients θ^* were specified as follows:

$$\theta^* = \begin{pmatrix} \mu \\ \alpha_{2\cdot} \\ \alpha_{3\cdot} \end{pmatrix} = \left(\begin{array}{cccc} Z_{1,\cdots,8} & Z_{9,\cdots,13} & Z_{14,\cdots,18} & Z_{19,\cdots,p} \\ 1.5\cdot\mathbf{1}_8 & 1\cdot\mathbf{1}_5 & 0.5\cdot\mathbf{1}_5 & \mathbf{0}_{p-18} \\ \mathbf{0}_8 & 1\cdot\mathbf{1}_5 & -0.5\cdot\mathbf{1}_5 & \mathbf{0}_{p-18} \\ \mathbf{0}_8 & -1\cdot\mathbf{1}_5 & 0.5\cdot\mathbf{1}_5 & \mathbf{0}_{p-18} \end{array}\right),$$

where $\mathbf{a}_m$ denotes a m-vector of a's. Therefore, covariates Z_1 to Z_8 had homogeneous effects, Z_9 to Z_{18} had heterogeneous effects, and the rest were null variables. Censoring times were generated from the uniform distribution $U(0, 2)$ to yield the event rate around 40% in each cohort. Table 4 reports the average numbers of correctly and incorrectly selected homogeneous and heterogeneous variables by our method and the rMSE over 200 replicates. Our results show that the proposed method achieves good accuracy in identifying the homogeneous and heterogeneous effects and maintains low error rates when the number of covariates is greater than the sample size.

3.2 Pooled ovarian cancer study

Using the pooled data from 1676 patients from 10 studies, which had complete information on tumor stage and debulking surgery, Ganzfried et al. (2013) found that the expression level of chemokine CXCL12 was associated with patient survival, which was not detected in individual studies due to insufficient power. We first examined whether the three variables CXCL12, tumor stage and debulking were homogeneous across studies, and applied our method and the meta-analysis method with random-effects model respectively. Both methods identified all three variables as homogeneous variables and yielded similar results on effect estimation. Figure 1 shows the forest plot of hazard ratios (HRs) of the variables.

To further study the association between other genetic variables and survival, we examined 21 candidate genes related to breast and ovarian cancer according to the reports from the National Cancer Institute (http://www.cancer.gov/cancer topics/pdq/genetics/breast-and-ovarian/HealthProfessional), which are CXCL12, CXCR4, RAD51C, RAD51, BABAM1, MLH1, MSH2, MSH6, TP53, HOXD1, CHEK2, HOXD3, CASP8, IRS1, TIPARP, PLEKHM1, BNC2, SKAP1, CERS6, BRCA1, and BRCA2. We also included three clinical variables: tumor stage, debulking, and age at initial pathologic diagnosis. The pooled analysis was conducted in 1053 patients from 4 studies in the database, with study size over 100 and complete information on the genetic and clinical variables. Table 5 reports the estimated coefficients by our method and the meta-analysis. The proposed method identified no heterogeneous variables, supporting the quality of data curation by Ganzfried et al. (2013).We found that clinical variables age, tumor stage, and debulking still remained as important risk factors associated with ovarian cancer, and identified important genetic biomarkers including BNC2, BRCA2, CASP8, CXCL12, CXCR4, HOXD1, and IRS1 (Goode et al., 2010; Welcsh and King, 2001; Ding et al., 2012; Kajiyama et al., 2008; Ma et al., 2011). The meta-analysis method also identified 6 out of these 10 variables as important factors with no heterogeneity, with the remaining 4 variables being non-significant, which resembled our findings in the simulation study that the meta-analysis can be of low power for small average effects. The remaining 14 variables were classified as null variables by our method, 11 of which were concluded the same by the meta-analysis method.

4. Discussion

In this article, we address the question of identifying variables' homogeneous and heterogeneous effects on a time-to-event outcome in pooled studies using a group variable selection approach based on penalized regression. The proposed method requires that each study has the same predictors in the pooled studies. We establish the theoretical properties and show good numerical performances of the proposed method. In practice, to better estimate the variables' effects, we could refit the data using the variables and their homogeneous/heterogeneous structure identified by our method, or we could randomly partition the data into two parts, one part of which is to detect heterogeneity and the other part for estimating effects. Also note that the proposed method can be easily extended to linear models and generalized linear models.

When the number of covariates diverges with the sample size, we establish the convergence rate of $p_n^4/n \to 0$ following Cai et al. (2005). More recently, Huang et al. (2013) established the oracle inequalities in the $p_n \gg n$ sparse Cox model setting, which may potentially be applicable to our context and needs further investigation.

For tuning parameter selection, it is well known that the generalized cross validation (GCV) and AIC-based methods may select irrelevant predictors with a non-vanishing probability as $n \to \infty$ (Wang et al., 2007). The BIC-based selection of tuning parameters to select models is consistent for model selection in various settings (Wang et al., 2007; Zhang et al., 2010). The log partial likelihood of the Cox proportional hazards model can be quadratically approximated so that the optimization is conducted in a similar fashion to the least-square setting, which supports the use of BIC for our proposed method. The cross-validation score approximating the Kullback-Leibler divergence can also be used to select the tuning parameter (Du et al., 2010), but needs further investigation when both *n* and *p* tend to infinity.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the editor Jeremy M. G. Taylor, the Associate Editor, and two referees for their careful review and constructive comments that substantially improved the presentation of the paper. This work was partially supported by the National Cancer Institute (R01 CA140632, R21 CA116585).

References

Cai J, Fan J, Li R, Zhou H. Variable selection for multivariate failure time data. Biometrika. 2005; 92:303–316. [PubMed: 19458784]

DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials. 1986; 7:177–188. [PubMed: 3802833]

Ding YC, McGuffog L, Healey S, Friedman E, Laitman Y, Paluch-Shimon S, Kaufman B, Liljegren A, Lindblom A, Olsson H, Kristoffersson U, Stenmark-Askmalm M, Melin B, et al. A nonsynonymous polymorphism in IRS1 modifies risk of developing breast and ovarian cancers in BRCA1 and ovarian cancer in BRCA2 mutation carriers. Cancer Epidemiology Biomarkers & Prevention. 2012; 21:1362–1370.

Du P, Ma S, Liang H. Penalized variable selection procedure for Cox models with semiparametric relative risk. The Annals of Statistics. 2010; 38:2092–2117. [PubMed: 20802853]

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association. 2001; 96:1348–1360.

Fu WJ. Penalized regressions: the bridge versus the lasso. Journal of Computational and Graphical Statistics. 1998; 7:397–416.

Ganzfried BF, Riester M, Haibe-Kains B, Risch T, Tyekucheva S, Jazic I, Wang XV, Ahmadifar M, Birrer MJ, Parmigiani G, Huttenhower C, Waldron L. CuratedOvarianData: clinically annotated data for the ovarian cancer transcriptome. Database: The Journal of Biological Databases and Curation. 2013; 2013 bat013.

Goode EL, Chenevix-Trench G, Song H, Ramus SJ, Notaridou M, Lawrenson K, Widschwendter M, Vierkant RA, Larson MC, Kjaer SK, Birrer MJ, et al. A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. Nature Genetics. 2010; 42:874–879. [PubMed: 20852632]

Hedges, LV.; Olkin, I. Statistical methods for meta-analysis. Orlando: Academic Press; 1985.

Huang J, Sun T, Ying Z, Yu Y, Zhang C-H. Oracle inequalities for the lasso in the Cox model. The Annals of Statistics. 2013; 41:1142–1165. [PubMed: 24086091]

Kajiyama H, Shibata K, Terauchi M, Ino K, Nawa A, Kikkawa F. Involvement of SDF-1α/CXCR4 axis in the enhanced peritoneal metastasis of epithelial ovarian carcinoma. International Journal of Cancer. 2008; 122:91–99.

Kim J, Sohn I, Jung S-H, Kim S, Park C. Analysis of survival data with group lasso. Communications in Statistics - Simulation and Computation. 2012; 41:1593–1605.

Liu M, Lu W, Krogh V, Hallmans G, Clendenen TV, Zeleniuch-Jacquotte A. Estimation and selection of complex covariate effects in pooled nested case-control studies with heterogeneity. Biostatistics. 2013; 14:682–694. [PubMed: 23632625]

Ma S, Song X, Song X, Huang J. Supervised group lasso with applications to microarray data analysis. BMC Bioinformatics. 2007; 8:60–76. [PubMed: 17316436]

Ma X, Zhang J, Liu S, Huang Y, Chen B, Wang D. Polymorphisms in the CASP8 gene and the risk of epithelial ovarian cancer. Gynecologic Oncology. 2011; 122:554–559. [PubMed: 21714991]

Moreno V, Martin ML, Bosch FX, de Sanjosé S, Torres F, Muñoz N. Combined analysis of matched and unmatched case-control studies: comparison of risk estimates from different studies. American Journal of Epidemiology. 1996; 143:293–300. [PubMed: 8561164]

Wang H, Li R, Tsai C-L. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika. 2007; 94:553–568. [PubMed: 19343105]

Welcsh PL, King MC. BRCA1 and BRCA2 and the genetics of breast and ovarian cancer. Human Molecular Genetics. 2001; 10:705–713. [PubMed: 11257103]

Yan J, Huang J. Model selection for Cox models with time-varying coefficients. Biometrics. 2012; 68:419–428. [PubMed: 22506825]

Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B. 2006; 68:49–67.

Zhang HH, Cheng G, Liu Y. Linear or nonlinear? Automatic structure discovery for partially linear models. Journal of the American Statistical Association. 2011; 106:1099–1112. [PubMed: 22121305]

Zhang HH, Lu W. Adaptive lasso for Cox's proportional hazards model. Biometrika. 2007; 94:691–703.

Zhang Y, Li R, Tsai C-L. Regularization parameter selections via generalized information criterion. Journal of the American Statistical Association. 2010; 105:312–323. [PubMed: 20676354]

Zhao P, Rocha G, Yu B. The composite absolute penalties family for grouped and hierarchical variable selection. The Annals of Statistics. 2009; 37:3468–3497.

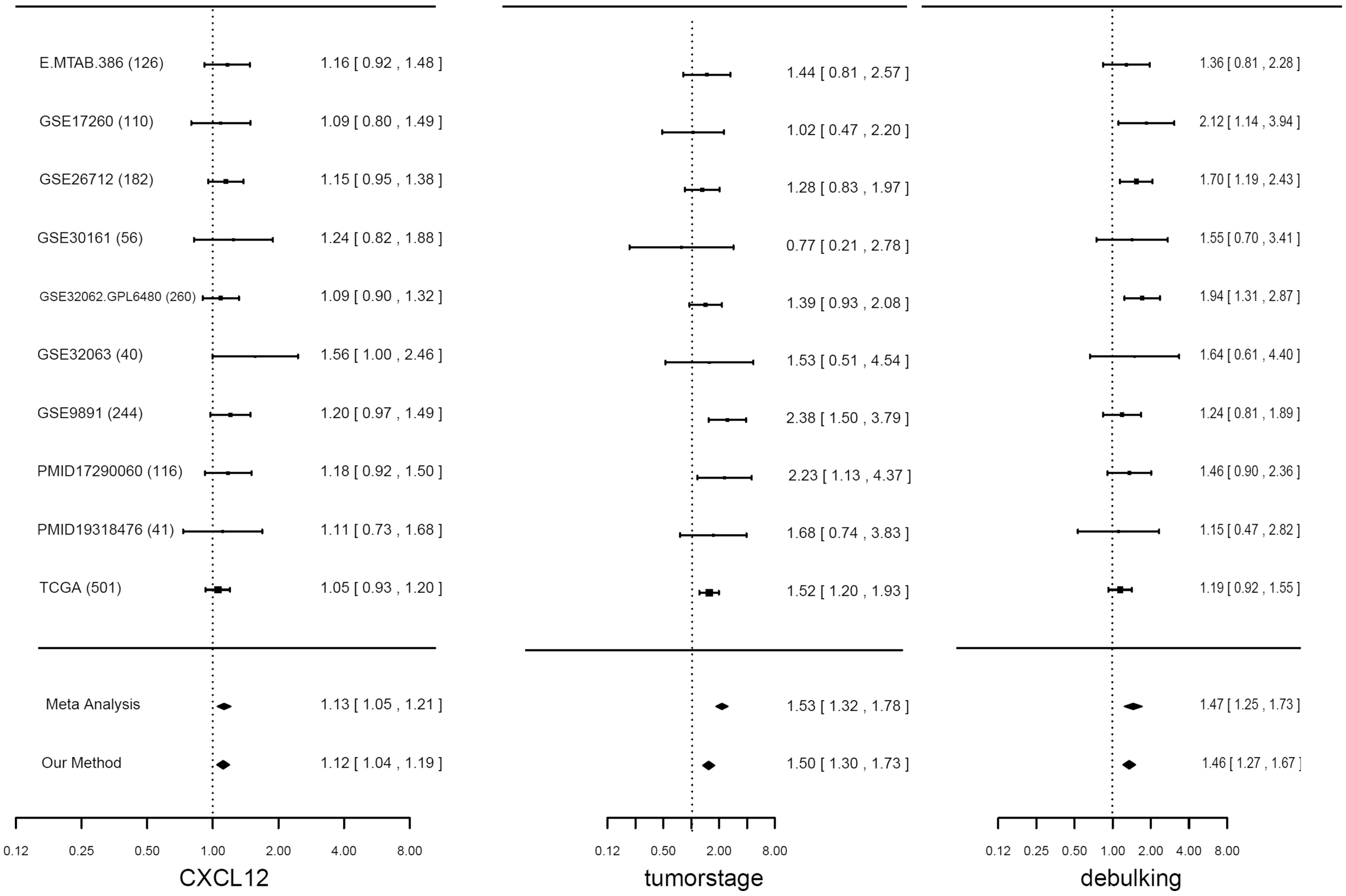

Figure 1.

Forest plot of the hazard ratio estimates of variables CXCL12, tumorstage, debulking from the pooled ovarian cancer study with 10 sub-studies. The study names are listed at the left and study sizes are given in the parentheses. Three vertical dash lines are reference lines of the hazard ratio being 1.

Table 1

Simulation results on variable selection and root mean square errors in Example 1.

		25% event rate					45% event rate				
		Homo	Homo	Hetero	Hetero		Homo	Homo	Hetero	Hetero	
N	Method	Corr (2)	Incorr (0)	Corr (3)	Incorr (0)	rMSE	Corr (2)	Incorr (0)	Corr (3)	Incorr (0)	rMSE
150	our method	1.9	0.69	2.37	0.31	0.83	1.93	0.3	2.84	0.26	0.57
	agLASSO	0	0	2.63	2.25	0.87	0	0	2.93	2.21	0.62
	aCAP	1.76	0.55	1.96	0.35	0.98	1.76	0.44	2.41	0.35	0.79
	two-step	1.78	0.60	2.41	0.74	1.39	1.84	0.41	2.89	0.61	1.30
	MLE	1.67	0.68	2.63	1.24	1.42	1.72	0.43	2.94	1.02	0.93
300	our method	1.94	0.14	2.95	0.27	0.48	1.96	0.06	3.00	0.16	0.35
	agLASSO	0	0	2.98	2.16	0.54	0	0	3	2.14	0.40
	aCAP	1.77	0.28	2.55	0.3	0.70	1.84	0.29	2.57	0.21	0.64
	two-step	1.85	0.25	2.97	0.57	1.27	1.87	0.31	3	0.51	1.25
	MLE	1.75	0.30	3	0.87	0.79	1.80	0.35	3.00	0.89	0.55

Homo/Hetero Corr: average numbers of correct homogeneous/heterogeneous effects; Homo/Hetero Incorr: average numbers of incorrect homogeneous/heterogeneous effects; rMSE: root mean square errors.

Table 2

Simulation results on the frequency of identifying nonzero mean effect and heterogeneous effects for each variable in Example 1 with cohort size of 300 and 45% event rate.

		Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}	Z_{11}
		200 **0**	**200** **200**	**200** **200**	**0** **0**	**0** **0**	**0** **0**	**200** **0**	**0** **200**	**0** **0**	**0** **0**	**0** **0**
Method	Structure											
our method	μ ≠ 0	200	200	200	2	3	2	200	1	1	0	4
	α ≠ 0	2	200	200	3	5	3	6	200	2	7	3
agLASSO	μ ≠ 0	200	200	200	1	9	5	200	200	4	6	3
	α ≠ 0	200	200	200	1	9	5	200	200	4	6	3
aCAP	μ ≠ 0	200	199	200	15	18	11	200	114	10	8	6
	α ≠ 0	9	199	200	2	4	1	24	114	1	1	1
two-step	μ ≠ 0	200	0	119	14	16	7	200	0	5	13	7
	α ≠ 0	7	200	200	14	15	14	20	200	8	13	11
MLE	μ ≠ 0	200	200	200	15	21	11	200	10	8	15	14
	α ≠ 0	19	200	200	27	25	18	22	200	15	25	25

Table 3

Simulation results on the estimated standard error and sample standard deviation (in the parenthesis) of estimators for selected variables in Example 1 with cohort size of 300 and 45% event rate.

		variable 1	**variable 2**			**variable 8**	
N	**Method**	**μ_1**	**μ_2**	**α_{22}**	**α_{32}**	**α_{28}**	**α_{38}**
150	our method	0.083(0.109)	0.086(0.131)	0.099(0.135)	0.099(0.130)	0.088(0.140)	0.084(0.124)
	agLASSO	0.073(0.119)	0.090(0.118)	0.122(0.139)	0.122(0.146)	0.078(0.144)	0.077(0.130)
	aCAP	0.082(0.121)	0.093(0.114)	0.093(0.115)	0.093(0.105)	0.090(0.138)	0.088(0.122)
	MLE	0.107(0.126)	0.124(0.144)	0.165(0.199)	0.165(0.191)	0.161(0.187)	0.161(0.162)
300	our method	0.059(0.068)	0.063(0.077)	0.077(0.082)	0.077(0.086)	0.068(0.088)	0.065(0.083)
	agLASSO	0.056(0.072)	0.067(0.072)	0.090(0.083)	0.090(0.086)	0.065(0.099)	0.064(0.095)
	aCAP	0.058(0.069)	0.067(0.079)	0.073(0.072)	0.073(0.072)	0.065(0.093)	0.064(0.083)
	MLE	0.068(0.072)	0.079(0.080)	0.106(0.104)	0.106(0.101)	0.103(0.105)	0.103(0.103)

Table 4

Simulation results on variable selection and root mean square errors in Example 2 with cohort size of 200 and 40% event rate.

		Homo	**Homo**	**Hetero**	**Hetero**	
ρ	**p**	**Corr (8)**	**Incorr (0)**	**Corr (10)**	**Incorr (0)**	**rMSE**
0.5	250	6.92	0.96	9.81	1.58	1.86
	450	6.87	1.03	9.75	1.56	1.90
0.75	250	6.52	0.38	9.61	2.02	2.02
	450	6.31	0.41	9.62	2.31	2.05

Homo/Hetero Corr: average numbers of correct homogeneous/heterogeneous effects; Homo/Hetero Incorr: average numbers of incorrect homogeneous/heterogeneous effects; rMSE: root mean square errors

Table 5

Estimates of the log hazard ratios of genetic and clinical variables in the pooled ovarian cancer study.

Variable	Our method	Two-step
age	0.022(0.004)	0.026(0.006)[*]
tumor stage	0.455(0.085)	0.536(0.100)[*]
debulking	0.248(0.071)	0.297(0.157)
BNC2	0.076(0.030)	0.145(0.058)[*]
BRCA2	0.140(0.036)	0.183(0.049)[*]
CASP8	−0.061(0.030)	−0.138(0.074)
CXCL12	0.042(0.026)	0.089(0.055)
CXCR4	−0.098(0.034)	−0.118(0.051)[*]
HOXD1	−0.011(0.015)	−0.111(0.137)
IRS1	0.072(0.029)	0.158(0.057) [aaaaaab]
BABAM1	0(−)	−0.075(0.131) [+]
BRCA1	0(−)	0.024(0.054)
CERS6	0(−)	−0.014(0.100)
CHEK2	0(−)	−0.023(0.102)
HOXD3	0(−)	0.072(0.072)
MSH2	0(−)	0.126(0.083)
MSH6	0(−)	−0.159(0.072)[*]
MLH1	0(−)	−0.018(0.051)
PLEKHM1	0(−)	0.034(0.055)
RAD51	0(−)	0.108(0.067)
RAD51C	0(−)	0.017(0.099)[+]
SKAP1	0(−)	0.020(0.052)
TIPARP	0(−)	0.017(0.086)
TP53	0(−)	−0.014(0.058)

[*] denotes nonzero average effects;

[+] denotes nonzero heterogeneous effects with random-effects model; Standard errors are reported in the parentheses.