Statistics and Probability Letters 93 (2014) 126–133

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

A modified adaptive Lasso for identifying interactions in the Cox model with the heredity constraint

Lu Wang[a,*], Jincheng Shen[a], Peter F. Thall[b]

[a] Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, USA

[b] Department of Biostatistics, M.D. Anderson Cancer Center, Houston, TX 77030, USA

ARTICLE INFO

Article history:
Received 13 March 2014
Received in revised form 20 June 2014
Accepted 28 June 2014
Available online 5 July 2014

Keywords:
Modified adaptive Lasso
Oracle property
Penalized partial likelihood
Regularization
Variable selection

ABSTRACT

In many biomedical studies, identifying effects of covariate interactions on survival is a major goal. Important examples are treatment–subgroup interactions in clinical trials, and gene–gene or gene–environment interactions in genomic studies. A common problem when implementing a variable selection algorithm in such settings is the requirement that the model must satisfy the strong heredity constraint, wherein an interaction may be included in the model only if the interaction's component variables are included as main effects. We propose a modified Lasso method for the Cox regression model that adaptively selects important single covariates and pairwise interactions while enforcing the strong heredity constraint. The proposed method is based on a modified log partial likelihood including two adaptively weighted penalties, one for main effects and one for interactions. A two-dimensional tuning parameter for the penalties is determined by generalized cross-validation. Asymptotic properties are established, including consistency and rate of convergence, and it is shown that the proposed selection procedure has oracle properties, given proper choice of regularization parameters. Simulations illustrate that the proposed method performs reliably across a range of different scenarios.

1. Introduction

The Cox proportional hazards regression model (Cox, 1972, 1975) is the most widely used statistical model for evaluating relationships between an event time, T, and baseline covariates, $X = (X_1, \ldots, X_p)$. The Cox model is characterized by the hazard function

$$h(t|X, \beta) = h_0(t)\ \exp\{g(X, \beta)\}, \quad t > 0, \tag{1}$$

for a subject with covariates X, where $h_0(t)$ is an unspecified baseline hazard and the linear component $g(X, \beta) = \sum_{j=1}^{p} \beta_j X_j$ is characterized by a vector $\beta = (\beta_1, \ldots, \beta_p)^T$ of unknown regression coefficients. In many applications, this simple form of $g(X, \beta)$ does not adequately describe the relationship between T and X due to interactions between elements of X. For example, an anti-cancer agent tailored to attack a certain biological target typically has effects hypothesized to differ between the subgroups of patients who do and do not have the target, identified by a binary "biomarker" covariate. In a randomized trial of the targeted agent versus standard therapy, such a differential effect is characterized as a treatment–biomarker interaction which, if found to be sufficiently large, may lead to regulatory approval of the agent for patients who are biomarker

[*] Corresponding author. Tel.: +1 7346476935.
E-mail address: luwang@umich.edu (L. Wang).

http://dx.doi.org/10.1016/j.spl.2014.06.024
0167-7152/© 2014 Elsevier B.V. All rights reserved.

positive. Identifying such treatment–biomarker interactions thus is a key step in developing personalized treatments. The presence of other covariates that may or may not be associated with T, and that also may interact with treatment, complicates identification and estimation of biomarker effects. Fitting the Cox model with interactions is challenging since, even with a moderate number of covariates, the number of interaction terms may be large, and not all covariates and interactions may have meaningful effects on $h(t|X,\beta)$.

There is a large literature on variable selection methods for survival models. A family of penalized partial likelihood methods have been proposed for the Cox proportional hazard model, including the Lasso (Tibshirani et al., 1997) and the smoothly clipped absolute deviation method (Fan and Li, 2002). By shrinking some regression coefficients to zero, these methods simultaneously select important variables and estimate the regression model parameters. Zhang and Lu (2007) proposed an adaptive Lasso estimator for variable selection in the Cox model, with an adaptively weighted L_1 penalty on the regression coefficients that has a convex form. They showed that this method enjoys the oracle properties, global optima exist, and it can be implemented efficiently using standard numerical algorithms (Boyd and Vandenberghe, 2004).

All the above variable selection methods for the Cox model treat the candidate variables equally. However, when interactions are included, there is a natural hierarchical ordering among the variables in the model (Chipman, 1996; Joseph, 2006; Yuan et al., 2007). This motivates the strong heredity requirement that an interaction can be included in the model only if the interaction's component variables are included as main effects (Hamada and Wu, 1992), since models that violate this property are difficult to interpret. For linear and generalized linear regression models, Yuan et al. (2009) proposed non-negative garrote methods that naturally incorporate a general hierarchical structure among predictors. Along this line, Choi et al. (2010) extended the Lasso to identify interaction terms while obeying the strong heredity constraint, which is achieved by reparameterizing the coefficients of the interaction terms. Bien et al. (2013) investigated a Lasso for hierarchical interactions, and Radchenko and James (2010) considered a more general case with nonlinear interactions. To our knowledge, none of these papers studied the setting with time-to-event data, and none of the variable selection methods for the Cox model noted above satisfy the strong heredity constraint when interactions are included. This paper aims at filling this gap.

We propose a modified Lasso procedure for the Cox model to adaptively select covariates and interactions while automatically enforcing the strong heredity constraint. The main challenges compared to linear/generalized linear regression are that the Cox model is semiparametric and involves right-censored data. We carry out estimation and variable selection by optimizing a modified log partial likelihood that includes two adaptively weighted penalty terms, one for main effects and one for reparameterized interactions, with each penalty multiplied by a tuning parameter. The main reason that we choose the reparameterization approach is that this method can handle all p main effects simultaneously in one iteration, which has tremendous numerical advantage, especially for survival analysis when no closed form can be found. The two-dimensional tuning parameter is determined by generalized cross-validation. The proposed method is computationally convenient, has good convergence properties, and implementation is straightforward. We establish asymptotic properties, including selection consistency, rate of convergence, and the oracle property (Fan and Li, 2001; Fan and Peng, 2004) that it performs as well as if the correct underlying model were known in advance. These theoretical properties and algorithms have not been studied previously for variable selection for Cox models with interaction terms subject to the strong hierarchy constraint.

2. Adaptive lasso with strong heredity constraint using penalized partial likelihood

Let T_i denote the failure time, C_i the censoring time, and $X_i = (X_{i1}, \ldots, X_{ip})$ the covariate vector of the ith subject, for $i = 1, \ldots, n$, with $\widetilde{T}_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \le C_i)$. Suppose that $(X_1, T_1, C_1), \ldots, (X_n, T_n, C_n)$ are independent and identically distributed. We further assume non-informative censoring, $T_i \perp\!\!\!\perp C_i \mid X_i$, i.e., T_i is conditionally independent of C_i given X_i. Although our methods can be generalized to handle higher order interactions, for the ease of exposition we consider the Cox model having a linear component with all possible two-way interactions. That is, in model (1)

$$g(X,\beta,\alpha) = \sum_{j=1}^{p} \beta_j X_j + \sum_{1\le j<j'\le p} \alpha_{j,j'} X_j X_{j'}, \tag{2}$$

where $\alpha = (\alpha_{1,2}, \ldots, \alpha_{p-1,p})^T$. Our goals are to provide a method that determines which terms in $g(X,\beta,\alpha)$ have important effects on the hazard, and develop a corresponding computational algorithm and parameter estimators having desirable properties. The existing variable selection methods for the Cox model do not guarantee the strong heredity constraint, as they treat all elements of (β,α) equally and do not distinguish between elements of β and α.

We first re-parameterize the coefficients for the interaction terms in (2) as $\alpha_{j,j'} = \gamma_{j,j'}\beta_j\beta_{j'}$, so that the linear term becomes

$$g(X,\theta) = \sum_{j=1}^{p} \beta_j X_j + \sum_{1\le j<j'\le p} \gamma_{j,j'}\beta_j\beta_{j'} X_j X_{j'} \tag{3}$$

and the parameter vector is $\theta = (\beta,\gamma) = (\beta_1, \ldots \beta_p, \gamma_{1,2}, \ldots, \gamma_{p-1,p})^T$. With this reparameterization, the coefficient for an interaction term $X_jX_{j'}$ must be 0 if either of its two main effects X_j or $X_{j'}$ has coefficient 0. Conversely, if $\gamma_{j,j'}\beta_j\beta_{j'} \neq 0$, this implies that both $\beta_j \neq 0$ and $\beta_{j'} \neq 0$, which guarantees the strong heredity constraint. With the reparameterization (3), the log partial likelihood is

$$l_n(\theta) = \sum_{i=1}^{n} \delta_i \left(g(X_i, \theta) - \log \left[\sum_{r=1}^{n} I(\widetilde{T}_r \geq \widetilde{T}_i) \exp\{g(X_r, \theta)\} \right] \right), \tag{4}$$

where $I(A)$ denotes the indicator of the event A. For the variable selection problem at hand, we will minimize the adaptive penalized negative log partial likelihood

$$Q_n(\theta, \lambda_\beta, \lambda_\gamma) = -l_n(\theta) + n\lambda_\beta \sum_{j=1}^{p} w_j^\beta \left|\beta_j\right| + n\lambda_\gamma \sum_{1 \leq j < j' \leq p} w_{j,j'}^\gamma \left|\gamma_{j,j'}\right|, \tag{5}$$

where $\{w_j^\beta\}$ and $\{w_{j,j'}^\gamma\}$ are prespecified weights, and $\lambda_\beta, \lambda_\gamma$ are tuning parameters. Following Breiman (1995) and Zou (2006), we compute the weights in (5) using

$$w_j^\beta = \frac{\log(n)}{n} \left| \frac{1}{\widetilde{\beta}_j} \right|, \qquad w_{j,j'}^\gamma = \frac{\log(n)}{n} \left| \frac{\widetilde{\beta}_j \widetilde{\beta}_{j'}}{\widetilde{\alpha}_{j,j'}} \right|,$$

where $\widetilde{\beta}_j$'s and $\widetilde{\alpha}_{j,j'}$'s are the estimates from a usual, unpenalized fitted Cox model with linear component (2). The multiplier $\log(n)/n$ is included to satisfy the convergence rate and the asymptotic properties introduced in Section 3. The objective function becomes

$$Q_n(\theta, \lambda_\beta, \lambda_\gamma) = -l_n(\theta) + \log(n)\lambda_\beta \sum_{j=1}^{p} \frac{\left|\beta_j\right|}{\left|\widetilde{\beta}_j\right|} + \log(n)\lambda_\gamma \sum_{1 \leq j < j' \leq p} \left|\gamma_{j,j'}\right| \left| \frac{\widetilde{\beta}_j \widetilde{\beta}_{j'}}{\widetilde{\alpha}_{j,j'}} \right|. \tag{6}$$

The two tuning parameters in (6) control the coefficient estimates at different levels. The first tuning parameter λ_β controls main effect estimates. If β_j is shrunk to zero, all terms involving X_j, including $\beta_j X_j$ and the interactions $\gamma_{j,j'}\beta_j\beta_{j'}X_jX_{j'}$, for any j', are removed from the model. The second tuning parameter λ_γ controls the estimates only at the interaction effect level. Even if both $\beta_j \neq 0$ and $\beta_{j'} \neq 0$, it is possible that $\gamma_{j,j'} = 0$ if X_j and $X_{j'}$ do not interact. The penalty term controlled by λ_γ thus provides the flexibility of selecting only main effects of X_j and $X_{j'}$ but not their interaction.

The weights act on the objective function, $Q_n(\theta, \lambda_\beta, \lambda_\gamma)$, as follows. If the initial estimate $\widetilde{\beta}_j$ is close to 0 then w_j^β will be large and hence, as can be seen from (5), the coefficient β_j of X_j will be heavily penalized. Similarly, if $\widetilde{\alpha}_{j,j'}$ is small relative to $\widetilde{\beta}_j\widetilde{\beta}_{j'}$ then $w_{j,j'}^\gamma$ will be large and the coefficient $\gamma_{j,j'}$ of the interaction term $X_jX_{j'}$ will be heavily penalized.

3. Theoretical properties of the estimator

In this section, we study the asymptotic properties of our proposed variable selection procedure and the corresponding estimator. As $n \to \infty$, our estimator possesses the oracle property under certain regularity conditions, that is, it performs as well as if the true model were known in advance (Fan and Li, 2001). The regularity conditions that we need throughout the development are given in Appendix 1 in the web supplementary materials (see Appendix A), where we follow the notation in Andersen and Gill (1982).

Let $\theta_0 = (\beta_0^T, \gamma_0^T)^T$ denote the true parameter vector, where

$$\gamma_{0j,j'} = \begin{cases} \alpha_{0j,j'}/\beta_{0j}\beta_{0j'} & \text{if } \beta_{0j} \neq 0 \text{ and } \beta_{0j'} \neq 0 \\ 0 \quad \text{otherwise.} \end{cases} \tag{7}$$

This guarantees that the true model obeys the strong heredity constraint, that is, $\alpha_{0j,j'} = \gamma_{0j,j'} = 0$ if either $\beta_{0j} = 0$ or $\beta_{0j'} = 0$.

Define the covariate-specific tuning parameters $\lambda_{j,n}^\beta = \lambda_\beta w_j^\beta = n^{-1}\log(n)\lambda_\beta / \left|\widetilde{\beta}_j\right|$ for $j = 1, \ldots, p$ and the interaction-specific tuning parameters $\lambda_{j,j',n}^\gamma = \lambda_\gamma w_{j,j'}^\gamma = n^{-1}\log(n)\lambda_\gamma / \left|\tilde{\gamma}_{j,j'}\right|$ for $1 \leq j < j' \leq p$, where $\tilde{\gamma}_{j,j'} = \widetilde{\alpha}_{j,j'}/\widetilde{\beta}_j\widetilde{\beta}_{j'}$. Our proposed estimator is

$$\hat{\theta} = \arg\min_\theta \left\{ -l_n(\theta) + n\sum_{j=1}^{p} \lambda_{j,n}^\beta \left|\beta_j\right| + n \sum_{1 \leq j < j' \leq p} \lambda_{j,j',n}^\gamma \left|\gamma_{j,j'}\right| \right\}. \tag{8}$$

Note that the criterion in (8) is the same as $Q_n(\theta, \lambda_\beta, \lambda_\gamma)$ in (6) with different notation.

Without loss of generality, we denote $\beta_0 = (\beta_{a0}^T, \beta_{b0}^T)^T$, where β_{a0} consists of all nonzero components and β_{b0} consists of the remaining zero components of β_0. Similarly, for the true coefficient of interactions, write $\gamma_0 = (\gamma_{a0}^T, \gamma_{b0}^T, \gamma_{c0}^T)^T$, where γ_{a0} contains all nonzero components of γ_0, γ_{b0} contains the zero components of γ_0 whose corresponding main effects both are nonzero, and γ_{c0} contains the remaining zero components of γ_0 whose corresponding components of main effects have at least one zero. Correspondingly, we denote the maximizer of (8) as $\hat{\theta}_n = (\hat{\beta}_{an}^T, \hat{\beta}_{bn}^T, \hat{\gamma}_{an}^T, \hat{\gamma}_{bn}^T, \hat{\gamma}_{cn}^T)^T$, the covariate-specific tuning parameters as $\{(\lambda_{an}^{\beta})^T, (\lambda_{bn}^{\beta})^T\}$, and the interaction-specific tuning parameters as $\{(\lambda_{an}^{\gamma})^T, (\lambda_{bn}^{\gamma})^T, (\lambda_{cn}^{\gamma})^T\}$.

Let

$$\xi_n = \max\left\{(\lambda_{an}^{\beta})^T, (\lambda_{an}^{\gamma})^T\right\}, \qquad \zeta_n = \min\left\{(\lambda_{bn}^{\beta})^T, (\lambda_{bn}^{\gamma})^T\right\}.$$

Then ξ_n is the maximum among both covariate-specific and interaction-specific tuning parameters that correspond to nonzero coefficients in the model. But for ζ_n, we only consider those specific tuning parameters corresponding to zero main effects and zero interaction terms when coefficients are in γ_{b0}. That is, we consider all zero cases for the main effects, but for the interactions, we only consider those zero terms whose corresponding main effects are nonzeros. We will refer to such terms in the definition of ζ_n as non-trivial zero terms.

For the proposed variable selection procedure to perform properly, as $n \to \infty$ the covariate-specific and interaction-specific penalties for the terms whose true coefficients are nonzeros should converge to 0, and the penalties for those non-trivial zero terms should be large enough so that the estimates shrink to 0. In fact, if the tuning parameters in (8) that correspond to β_{a0} and γ_{a0} converge to 0, and those corresponding to β_{b0} and γ_{b0} are sufficiently large, then our proposed estimating procedure will have the so-called oracle property (Fan and Li, 2001). This can be guaranteed when $\sqrt{n}$-consistent estimates of θ_0 are used in the definition of $\lambda_{j,n}^{\beta}$ and $\lambda_{j,j',n}^{\gamma}$, where one can easily show that $\sqrt{n}\xi_n \to 0$ and $\sqrt{n}\zeta_n \to \infty$.

Denote the initial estimates as $\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)$, $\tilde{\gamma} = (\tilde{\gamma}_{1,2}, \ldots, \tilde{\gamma}_{p-1,p})$, and we summarize the above results in Theorem 1. A sketch of the proof of Theorem 1 is given in Appendix 2 (see Appendix A).

Theorem 1. *When $\sqrt{n}(\tilde{\beta} - \beta_0) = O_p(1)$ and $\sqrt{n}(\tilde{\gamma} - \gamma_0) = O_p(1)$ in (8) to calculate $\lambda_{j,n}^{\beta}$ and $\lambda_{j,j',n}^{\gamma}$, $\sqrt{n}\xi_n \to \infty$ and $\sqrt{n}\zeta_n \to \infty$ as $n \to \infty$. Under the regularity conditions* (1)–(3) *in Appendix* 1 *(see Appendix A), there exists a local minimizer $\hat{\theta}_n$ of $Q_n(\theta)$ such that*

(i) *(Sparsity) $P(\hat{\beta}_{bn} = 0) \to 1$, $P(\hat{\gamma}_{bn} = 0) \to 1$, and $P(\hat{\gamma}_{cn} = 0) \to 1$ as $n \to \infty$.*
(ii) *(Asymptotic normality)*

$$\sqrt{n}\left\{\begin{pmatrix}\hat{\beta}_{an}\\ \hat{\gamma}_{an}\end{pmatrix} - \begin{pmatrix}\beta_{a0}\\ \gamma_{a0}\end{pmatrix}\right\} \xrightarrow{d} N\left\{0,\ \mathcal{I}_a^{-1}(\beta_{a0}, \gamma_{a0})\right\},$$

where $\mathcal{I}_a(\beta_{a0}, \gamma_{a0})$ is the Fisher information matrix evaluated at β_{a0} and γ_{a0} assuming that $\beta_{b0} = 0$, $\gamma_{b0} = 0$, and $\gamma_{c0} = 0$ is known in advance.

Part (i) of Theorem 1 presents the sparsity property and shows that our proposed method can consistently remove the zero-effect terms with probability tending to 1. This implies that, with a sufficiently large sample, in practice our method can select the underlying true model with high probability. In part (ii) of Theorem 1, we establish that the estimates of nonzero elements of θ_0 are $\sqrt{n}$-consistent and asymptotically normal. The asymptotic distribution is the same as what would be obtained if it were known in advance which elements of θ_0 are 0 and which are not 0, the so-called oracle property.

Theorem 2 establishes the asymptotic behavior of the proposed method when p is very large, especially when $p \to \infty$ as $n \to \infty$. This is an important case in many biomedical studies, with the advance of new high throughput technologies. A proof of Theorem 2 is given in Appendix 3, provided in the web supplementary materials (see Appendix A).

When the number of predictors may increase with the sample size n, we denote p as p_n, which allows the possibility that $p_n \to \infty$ as $n \to \infty$. Including all main effects and pairwise interactions, the total number of parameters is $q_n = (p_n + 1)p_n/2$. Similarly, when appropriate we add a subscript n to other notation, and we let d_n denote the number of non-zero coefficients in the underlying true model. Then we have

Theorem 2. *Under the regularity conditions* (4)–(6) *in Appendix* 1 *(see Appendix A), if $p_n = o(n^{1/5})$ and $\sqrt{nq_n}\xi_n \to 0$, $\sqrt{n/q_n}\zeta_n \to \infty$ as $n \to \infty$, then there exists a local minimizer $\hat{\theta}_n$ of $Q_n(\theta)$ such that*

(i) *(Sparsity) $P(\hat{\beta}_{bn} = 0) \to 1$, $P(\hat{\gamma}_{bn} = 0) \to 1$, and $P(\hat{\gamma}_{cn} = 0) \to 1$ as $n \to \infty$.*
(ii) *(Asymptotic Normality)*

$$\sqrt{n}\Omega_n\mathcal{I}_{an}^{1/2}(\beta_{a0}, \gamma_{a0})\left\{\begin{pmatrix}\hat{\beta}_{an}\\ \hat{\gamma}_{an}\end{pmatrix} - \begin{pmatrix}\beta_{a0}\\ \gamma_{a0}\end{pmatrix}\right\} \xrightarrow{d} N\{0, \Sigma\},$$

where Ω_n is any arbitrary $d \times d_n$ matrix with a finite d such that $\Omega_n\Omega_n^T \to \Sigma$, Σ is a $d \times d$ semipositive definite symmetric matrix, and $\mathcal{I}_{an}(\beta_{a0}, \gamma_{a0})$ is the $d_n \times d_n$ Fisher information matrix evaluated at $(\beta_{a0}, \gamma_{a0})$ assuming that $\beta_{b0} = 0$, $\gamma_{b0} = 0$, and $\gamma_{c0} = 0$ is known in advance.

The reason we consider an arbitrary linear combination $\Omega_n(\hat{\beta}_{an}^T, \hat{\gamma}_{an}^T)^T$ in Theorem 2, instead of $(\hat{\beta}_{an}^T, \hat{\gamma}_{an}^T)^T$ as in Theorem 1, is because the latter has dimension $d_n \to \infty$ as $n \to \infty$ in the current setup, while the former has finite dimension d.

4. Computational algorithm

Expanding $l_n(\theta)$ in the objective function (6), $Q_n(\theta, \lambda_\beta, \lambda_\gamma)$ becomes

$$-\sum_{i=1}^{n}\delta_i\left[\left(\sum_{j=1}^{p}\beta_k X_{ij} + \sum_{1\le j<j'\le p}\gamma_{j,j'}\beta_j\beta_{j'}X_{ij}X_{ij'}\right) - \log\left\{\sum_{r=1}^{n} I(\widetilde{T}_r \ge \widetilde{T}_i)\times \exp\left(\sum_{j=1}^{p}\beta_j X_{rj}\right.\right.\right.$$
$$\left.\left.\left.+\sum_{1\le j<j'\le p}\gamma_{j,j'}\beta_j\beta_{j'}X_{rj}X_{rj'}\right)\right\}\right] + \log(n)\lambda_\beta\sum_{j=1}^{p}\frac{|\beta_j|}{|\widetilde{\beta}_j|} + \log(n)\lambda_\gamma\sum_{1\le j<j'\le p}\left|\gamma_{jj'}\right|\left|\frac{\widetilde{\beta}_j\widetilde{\beta}_{j'}}{\widetilde{\alpha}_{jj'}}\right|.$$

Our estimator minimizes $Q_n(\theta, \lambda_\beta, \lambda_\gamma)$, that is, $\hat{\theta}_n = \arg\min_\theta Q_n(\theta, \lambda_\beta, \lambda_\gamma)$. To carry out the optimization, we apply a modified version of the iteratively reweighted least squares algorithm (Green, 1984) with weighted L_1 penalties. Denoting the gradient vector of the partial likelihood by $\dot{l}(\theta) = -\partial l_n(\theta)/\partial\theta$ and the Hessian matrix by $\ddot{l}(\theta) = -\partial^2 l_n(\theta)/\partial\theta\partial\theta^T$, and using the Cholesky decomposition $\ddot{l}(\theta) = MM^T$, where M is an invertible lower triangular matrix, we define the pseudo response vector $Y = (M)^{-1}\{\ddot{l}(\theta)\theta - \dot{l}(\theta)\}$. By the usual second-order Taylor expansion, $-l_n(\theta)$ in (6) can be approximated by the quadratic form $(Y - M^T\theta)^T(Y - M^T\theta)/2$, and at each penalized iteratively reweighted least squares iteration we minimize

$$\frac{1}{2}\left(Y - M^T\theta\right)^T\left(Y - M^T\theta\right) + \log(n)\lambda_\beta\sum_{j=1}^{p}\frac{|\beta_j|}{|\widetilde{\beta}_j|} + \log(n)\lambda_\gamma\sum_{1\le j<j'\le p}\left|\gamma_{j,j'}\right|\left|\frac{\widetilde{\beta}_j\widetilde{\beta}_{j'}}{\widetilde{\alpha}_{j,j'}}\right|. \tag{9}$$

Because the main effect coefficients β and the interaction coefficients γ are controlled at different levels, in each step we also iterate between these two sets, first fixing β to estimate γ, then fixing γ to estimate β, and iterating until convergence. This algorithm is guaranteed to converge, since the objective function decreases at each step. When β is fixed, the optimization in γ becomes a Lasso problem, hence one can use either the LARS/Lasso algorithm (Efron et al., 2004) or quadratic programming to solve for γ efficiently. When γ is fixed, we solve for $\beta_1, \ldots, \beta_p$ sequentially. For each $j = 1, \ldots, p$, we fix γ and $\beta_{[-j]} = \left(\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p\right)$, and the optimization becomes a simple Lasso problem with only one parameter, β_j. This is similar to the shooting algorithm (Fu, 1998; Zhang and Lu, 2007; Friedman et al., 2007). The tuning parameters are selected by minimizing the generalized cross validation statistic,

$$C_{pseudo-GCV} = \frac{l(\hat{\theta})}{(1 - df/n)^2},$$

over a reasonable range of λ_β and λ_γ, where df is the number of nonzero parameters in the fitted model.

For a fixed λ_β and λ_γ, the optimization algorithm proceeds as follows:

Step 1. Center and normalize each term $X_j, X_jX_{j'}, j < j'$, and $j, j' = 1, \ldots, p$.

Step 2. Start with plausible initial values $\hat{\beta}_j^{(0)}$ and $\hat{\gamma}_{j,j'}^{(0)}, j < j'$, and $j, j' = 1, \ldots, p$, such as the conventional Cox regression parameter estimates. Set $m = 1$.

Step 3. Compute Y and M based on the current value $\hat{\theta}^{(m-1)}$. Denote

$$\tilde{l}(\theta) = -\frac{1}{2}\left(Y - M^T\theta\right)^T\left(Y - M^T\theta\right).$$

Step 4. To update $\hat{\gamma}$, let $\hat{\gamma}^{(m)} = \arg\min_\gamma \sum_{i=1}^{n}\{-\tilde{l}(\hat{\beta}^{(m-1)}, \gamma) + n\lambda_\gamma\sum_{j<j'} w_{j,j'}^{\gamma}\left|\gamma_{j,j'}\right|\}$.

Step 5. To update $\hat{\beta}$, for each $j = 1, \ldots, p$ in sequence, let

$$\hat{\beta}_j^{(m)} = \arg\min_{\beta_j}\sum_{i=1}^{n}\{-\tilde{l}(\hat{\beta}_{[-j]}^{(m-1)}, \beta_j, \hat{\gamma}^{(m)}) + \lambda_\beta w_j^{\beta}\left|\beta_j\right|\}.$$

Step 6. If the relative difference between $Q_n(\hat{\theta}^{(m-1)})$ and $Q_n(\hat{\theta}^{(m)})$,

$$\Delta^{(m)} = \frac{\left|Q_n(\hat{\theta}^{(m-1)}) - Q_n(\hat{\theta}^{(m)})\right|}{\left|Q_n(\hat{\theta}^{(m-1)})\right|},$$

is small enough, then stop. Otherwise, increment m to $m + 1$ and return to *Step* 3.

Since $\gamma_{j,j'} = \alpha_{j,j'}/\beta_j\beta_{j'}$, in *Step* 3 the minimization is actually over $\{\alpha_{j,j'}\}$ with $\beta = \hat{\beta}^{(m-1)}$. This algorithm gives exact zeros for some coefficients and guarantees that the coefficients of the corresponding interactions are set to 0 whenever the corresponding β is shrunk to 0. Based on our empirical experience, the algorithm converges quickly.

Table 1
Percentages of correctly selected models among 100 replications.

	Censoring percentage	ρ for Correlation	Method		
			Lasso	Adaptive Lasso	Proposed method
Model 1	25%	0	11%	70%	86%
		0.5	6%	66%	79%
	40%	0	3%	58%	80%
		0.5	3%	54%	82%
Model 2	25%	0	43%	49%	74%
		0.5	26%	31%	63%
	40%	0	28%	32%	74%
		0.5	22%	33%	62%

Table 2
Term-specific percentages of being selected for each main effect and interaction. (25% censoring and $\rho = 0$ for independent covariates.)

	Method	x_1	x_2	x_3	x_4	x_5	x_1x_2	x_1x_3	x_1x_4
Model 1	Lasso	22%	100%	100%	18%	25%	15%	12%	15%
	Adaptive Lasso	6%	100%	100%	3%	7%	2%	4%	1%
	Proposed method	5%	100%	100%	4%	6%	5%	5%	0%
Model 2	Lasso	6%	91%	91%	6%	5%	2%	6%	3%
	Adaptive Lasso	2%	87%	85%	4%	4%	1%	2%	1%
	Proposed method	10%	98%	99%	11%	8%	10%	10%	1%

	Method	x_1x_5	x_2x_3	x_2x_4	x_2x_5	x_3x_4	x_3x_5	x_4x_5
Model 1	Lasso	15%	100%	13%	13%	11%	13%	15%
	Adaptive Lasso	2%	100%	3%	2%	2%	2%	0%
	Proposed method	1%	100%	4%	6%	4%	6%	0%
Model 2	Lasso	3%	89%	4%	3%	9%	5%	2%
	Adaptive Lasso	2%	82%	1%	2%	1%	0%	0%
	Proposed method	1%	97%	10%	8%	11%	8%	2%

5. Numerical results

In this section, we report results of a simulation study comparing our proposed method with the Lasso and adaptive Lasso, two popular variable selection methods for the Cox model, neither of which guarantees the strong heredity constraint. We follow the simulation setup in Zhang and Lu (2007), but also include two-way interaction terms as candidates for the variable selection. We consider sample size $n = 200$, and assume that there are $p = 5$ covariates of interest in each simulated dataset, denoted by $X_1, X_2, \dots, X_5$. Thus, the number of all possible two-way interaction terms is $p \times (p-1)/2 = 10$, and there are a total of 15 candidate terms. We assume each covariate follows a standard normal distribution, and consider two scenarios: (i) the covariates are independent, and (ii) the covariates have pairwise correlations $Corr(X_j, X_{j'}) = \rho^{|j-j'|}$, with ρ = 0.5. We generate the censoring times from a Uniform distribution having support $(0, \tau)$, with τ chosen to obtain a specified censoring rate of either 25% or 40%.

Then suppose failure times are generated from a Cox model with constant baseline hazard $\lambda_0 = 0.1$, where the coefficients of X_2, X_3, and the interaction term X_2X_3 are non-zero, and the other 12 coefficients are zero. We consider the following two models:

Model 1: $\beta_2^0 = -0.8$, $\beta_3^0 = -0.8$, and $\alpha_{2,3}^0 = -0.8$, corresponding to large effects.

Model 2: $\beta_2^0 = -0.3$, $\beta_3^0 = -0.3$, and $\alpha_{2,3}^0 = -0.3$, corresponding to small effects.

We simulate each case 100 times, and apply all three methods to each simulated dataset. For all methods, the tuning parameters are selected using the generalized cross validation criterion $C_{pseudo-GCV}$.

Table 1 summarizes the percentage of fully correct variable selection among 100 replications for each method under each scenario. For all cases, the proposed method correctly selects the true model more frequently than the regular Lasso and adaptive Lasso. Specifically, for all scenarios under Model 2, where important variables and interactions have small effects, the Lasso and adaptive Lasso perform about the same, with the latter slightly better, while our proposed method performs much better than these two methods. In Model 1, where important variables and interactions have large effects, the advantage of our proposed method is still substantial when the censoring percentage is high, 40%. When censoring is moderate, 25%, the adaptive Lasso becomes competitive, but still is worse than the proposed method. The performance of the Lasso is always the worst among the three methods.

Table 2 summarizes the individual frequencies of being selected into the model for each main effect and interaction term. We only present the results here for 25% censoring under both models when the covariates are independent ($\rho = 0$) due to space limitations, but similar results are observed for the 40% censored case, and correlated covariates case. As shown in Table 2, the proposed variable selection procedure always chooses important variable and interaction terms much more

Table 3
Mean squared error, with standard errors in parentheses.

	Censoring percentage	ρ for correlation	Method		
			Lasso	Adaptive Lasso	Proposed method
Model 1	25%	0	0.213 (0.124)	0.072 (0.068)	0.063 (0.065)
		0.5	0.219 (0.136)	0.071 (0.068)	0.069 (0.065)
	40%	0	0.337 (0.209)	0.168 (0.144)	0.117 (0.219)
		0.5	0.351 (0.215)	0.162 (0.172)	0.118 (0.216)
Model 2	25%	0	0.118 (0.053)	0.105 (0.055)	0.042 (0.041)
		0.5	0.131 (0.059)	0.130 (0.059)	0.054 (0.048)
	40%	0	0.117 (0.053)	0.122 (0.058)	0.061 (0.062)
		0.5	0.137 (0.066)	0.142 (0.068)	0.082 (0.078)

often than the other two methods. Under Model 1 for large effects, the adaptive Lasso and the proposed method in general select the unimportant terms less often than the Lasso, while for Model 2 with small effects, the proposed method actually selects some of the unimportant terms more often. However, as illustrated in Table 1, the proposed method simultaneously selects all the important terms and removes the unimportant ones more frequently.

A simple, ad hoc alternative approach often employed in practice for variable selection involving interaction terms is to run either the Lasso or the adaptive Lasso, and then, depending on which terms were selected, manually add back any main terms that were not selected but that were components of any interaction term that was selected. This ensures the strong heredity constraint. As seen in our simulation, this practice does not help much in terms of how often the true model is selected correctly, while the false positive error rate is greatly increased. For example, as one can see in Table 2 for model 2, since the proposed method selected the true interaction term x_2x_3 more often than the other two methods, using the above ad hoc approach following the Lasso or the adaptive Lasso still cannot beat the proposed method. For model 1 with large effects in Table 2, the frequency of x_2, x_3, and x_2x_3 being correctly selected already achieves 100%, so this ad hoc approach does not help at all.

To measure the prediction accuracy, we average the mean squared error $C_{MSE} = (\hat{\theta} - \theta)^T \Gamma (\hat{\theta} - \theta)$ over 100 replications by following Tibshirani et al. (1997) and Zhang and Lu (2007), where Γ is the population variance–covariance matrix of the covariates. Standard errors are given in parentheses. For all scenarios, the proposed method has the smallest mean squared error (Table 3), and thus outperform the other two competitors in terms of prediction accuracy.

6. Discussion

We have extended the adaptive Lasso method to accommodate the Cox proportional hazard model including interaction terms while ensuring that the strong heredity constraint is satisfied in the selected model. Hamada and Wu (1992) consider other constraints, such as weak heredity where only one of the two main terms is required to be included when an interaction term is considered. Although this situation is not of our main interest, our methods can easily be modified to handle it by employing a different reparameterization.

Similarly, as discussed in Zhang and Lu (2007), the adaptive choice of weights may become problematic when some elements of θ are not estimable. This may occur, for example, when strong collinearity exists among covariates, or when the number of covariates p is much larger than the sample size n in high-dimensional data. In such settings, one cannot obtain the initial estimates to determine the adaptive weights. Alternatively, robust estimation of θ, such as ridge regression, may be considered.

Our work was motivated by the desire to identify key treatment–biomarker interactions for developing personalized treatments, where the number of candidate biomarkers is usually fixed or may grow slowly with n. Even with a moderate number of candidate biomarkers, however, our methodology can have an important impact on physician's actual behavior, if clinically meaningful treatment–biomarker effects are identified. The condition $p = o(n^{1/5})$ that we used may be relaxed though. There have been recent theoretical developments on high-dimensional survival models such as those in Bradic et al. (2011) and Lin and Lv (2013), but their underlying theory cannot be applied directly to our setting, which uses a reparameterization approach for selecting important interactions with heredity constraint. In order to relax our condition $p = o(n^{1/5})$, modification of the above theory to our setting would be required, and this would entail a substantial amount of work. While this is beyond the scope of the current paper, it certainly would be an interesting future research project.

Acknowledgments

The authors thank the Co-Editor-in-Chief, Associate Editor, and referee for their valuable comments.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.spl.2014.06.024.

References

Andersen, P.K., Gill, R.D., 1982. Cox's regression model for counting processes: a large sample study. Ann. Statist. 1100–1120.

Bien, J., Taylor, J., Tibshirani, R., 2013. A lasso for hierarchical interactions. Ann. Statist. 41, 1111–1141.

Boyd, S., Vandenberghe, L., 2004. Convex Optimization. Cambridge University Press, Cambridge.

Bradic, J., Fan, J., Jiang, J., 2011. Regularization for Cox's proportional hazards model with NP-dimensionality. Ann. Statist. 39 (6), 3092–3120.

Breiman, L., 1995. Better subset regression using the non-negative garrote. Technometrics 37, 373–384.

Chipman, H., 1996. Bayesian variable selection with related predictors. Canad. J. Statist. 24, 17–36.

Choi, N. H., Li, W., Zhu, J., 2010. Variable selection with the strong heredity constraint and its oracle property. J. Amer. Statist. Assoc. 105, 354–364.

Cox, D.R., 1972. Regression models and life tables (with discussion). J. R. Stat. Soc. 34, 187–220.

Cox, D.R., 1975. Partial likelihood. Biometrika 62, 269–276.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression (with Discussion). Ann. Statist. 32, 407–499.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360.

Fan, J., Li, R., 2002. Variable selection for Cox's proportional hazards model and frailty model. Ann. Statist. 30, 74–99.

Fan, J., Peng, H., 2004. Nonconcave penalized likelihood with a diverging number of parameters. Ann. Statist. 32, 928–961.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., 2007. Pathwise coordinate optimization. Ann. Appl. Stat. 1, 302–332.

Fu, W.J., 1998. Penalized regressions: the bridge versus the lasso. J. Comput. Graph. Statist. 7, 397–416.

Green, P.J., 1984. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. J. R. Stat. Soc. Ser. B Stat. Methodol. 149–192.

Hamada, M., Wu, C.J., 1992. Analysis of designed experiments with complex aliasing. J. Qual. Technol. 24, 130–137.

Joseph, V.R., 2006. A Bayesian approach to the design and analysis of fractionated experiments. Technometrics 48, 219–229.

Lin, W., Lv, J., 2013. High-dimensional sparse additive hazards regression. J. Amer. Statist. Assoc. 108, 247–264.

Radchenko, P., James, G.M., 2010. Variable selection using adaptive nonlinear interaction structures in high dimensions. J. Amer. Statist. Assoc. 105, 1541–1553.

Tibshirani, R., et al., 1997. The lasso method for variable selection in the Cox model. Stat. Med. 16, 385–395.

Yuan, M., Joseph, V.R., Lin, Y., 2007. An efficient variable selection approach for analyzing designed experiments. Technometrics 49, 430–439.

Yuan, M., Joseph, V.R., Zou, H., 2009. Structured variable selection and estimation. Ann. Appl. Stat. 1738–1757.

Zhang, H.H., Lu, W., 2007. Adaptive Lasso for Cox's proportional hazards model. Biometrika 94, 691–703.

Zou, H., 2006. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101, 1418–1429.

Web Supplementary Materials for "A Modified Adaptive Lasso for Identifying Interactions in the Cox Model with the Heredity Constraint"

Lu Wang[a,*], Jincheng Shen[a], Peter F. Thall[b]

[a] *Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, USA*
[b] *Department of Biostatistics, M.D. Anderson Cancer Center, Houston, Texas 77030, USA*

Appendix 1: Regularity conditions.

Following the notation in Andersen & Gill (1982), we consider a finite time interval $[0, \tau]$ with $\tau < \infty$. To facilitate the notation, let $N_i(t) = I\{T_i \le t, T_i \le C_i\}$ and $Y_i(t) = I\{T_i \ge t, C_i \ge t\}$. Define $H_i(\theta, t) = Y_i(t)\exp\{g(X_i, \theta)\}$, $S^{(0)}(\theta, t) = n^{-1}\sum_{i=1}^n H_i(\theta, t)$, $S^{(1)}(\theta, t) = n^{-1}\sum_{i=1}^n \nabla_\theta H_i(\theta, t)$, $S^{(2)}(\theta, t) = n^{-1}\sum_{i=1}^n \nabla_\theta^2 H_i(\theta, t)$, and $S^{(3)}(\theta, t) = n^{-1}\sum_{i=1}^n \nabla_\theta^3 H_i(\theta, t)$, where $\nabla_\theta(\cdot)$ denotes the first derivative with respect of θ, $\nabla_\theta^2(\cdot)$ and $\nabla_\theta^3(\cdot)$ denote the second and third order derivatives respectively.

We assume the following regularity conditions hold for Theorem 1:

(1) $\int_0^\tau \lambda_0(t)dt < \infty$

(2) There exists a neighbourhood Θ of θ_0 and $s^{(0)}(\theta, t)$, $s^{(1)}(\theta, t)$, $s^{(2)}(\theta, t)$ and $s^{(3)}(\theta, t)$ defined on $\Theta \times [0, \tau]$ such that for j = 0,1,2 and 3.

$$\sup_{t\in[0,\tau],\theta\in\Theta} \left\| S^{(j)}(\theta, t) - s^{(j)}(\theta, t) \right\| \longrightarrow_{\mathscr{P}} 0$$

where $\|\cdot\|$ is the L_1-norm.

(3) Let Θ, $s^{(0)}(\cdot,\cdot)$, $s^{(1)}(\cdot,\cdot)$, $s^{(2)}(\cdot,\cdot)$ and $s^{(3)}(\cdot,\cdot)$ be as in Condition (2) and define $e = s^{(1)}/s^{(0)}$ and $v = s^{(2)}/s^{(0)} - e \otimes e$. For all $\theta \in \Theta$, $t \in [0, \tau]$, $s^{(0)}(\cdot, t)$, $s^{(1)}(\cdot, t)$ and $s^{(2)}(\cdot, t)$ are continuous functions of $\theta \in \Theta$, uniformly in $t \in [0, \tau]$, $s^{(0)}(\theta, t)$, $s^{(1)}(\theta, t)$, $s^{(2)}(\theta, t)$, and $s^{(3)}(\theta, t)$ are bounded on $\Theta \times [0, \tau]$; and $s^{(0)}(\theta, t)$ is bounded away from zero on $\Theta \times [0, \tau]$. Let

[*] Corresponding author.
Email address: `luwang@umich.edu` (Lu Wang)

$u^{(2)}(\theta) = \nabla_\theta^2 g(X, \theta)$,

$$I(\theta_0) = \int_0^\tau \left\{ \nu(\theta_0, t) - u^{(2)}(\theta_0) \right\} s^{(0)}(\theta_0, t)\lambda_0(t)dt,$$

and we require the submatrix $\mathcal{I}_a(\beta_{a0}, \gamma_{a0})$ from $I(\theta_0)$ that corresponds to the non-zero $(\beta_{a0}, \gamma_{a0})$ is positive definite.

For Theorem 2, we denote $H_i(\theta_n, t) = Y_i(t) \exp\{g(\theta_n, X_{n,i})\}$. Define $\phi_{i,j}(\theta_n, t) = \{\partial H_i(\theta_n, t)/\partial\theta_{n,j}\}$ $/\{n^{-1}\sum_{i=1}^n H_i(\theta_n, t)\}$, $W_j^{(1)}(\theta_n, t) = n^{-1}\sum_{i=1}^n \phi_{i,j}(\theta_n, t)$, $W_{jk}^{(2)}(\theta_n, t) = n^{-1}\sum_{i=1}^n \partial\phi_{i,j}(\theta_n, t)/\partial\theta_{n,k}$, and $W_{jkl}^{(3)}(\theta_n, t) = n^{-1}\sum_{i=1}^n \partial\phi_{i,j}(\theta_n, t)/(\partial\theta_{n,k}\partial\theta_{n,l})$, for any $j, k, l = 1, \cdots, p_n(p_n+1)/2$. We further denote $W_{jk}^{(1,2)}(\theta_n, t) = n^{-1}\sum_{i=1}^n \{\phi_{i,j}(\theta_n, t)\phi_{i,k}(\theta_n, t)\}^2$, $W_{jk}^{(2,2)}(\theta_n, t) = n^{-1}\sum_{i=1}^n \{\partial\phi_{i,j}(\theta_n, t)/\partial\theta_{n,k}\}^2$, and $W_{jkl}^{(3,2)}(\theta_n, t) = n^{-1}\sum_{i=1}^n \{\partial^2\phi_{i,j}(\theta_n, t)/\partial\theta_{n,k}\partial\theta_{n,l}\}^2$. We assume the following regularity conditions in Theorem 2.

(4) $\int_0^\tau \lambda_0(t)dt < \infty$

(5) There exists a neighbourhood Θ_n of $\theta_{n,0}$ and $w_j^{(1)}(\theta_n, t)$, $w_{jk}^{(2)}(\theta_n, t)$, $w_{jkl}^{(3)}(\theta_n, t)$, $w_{jk}^{(1,2)}(\theta_n, t)$, $w_{jk}^{(2,2)}(\theta_n, t)$, $w_{jkl}^{(3,2)}(\theta_n, t)$, defined on $\Theta_n \times [0, \tau]$ such that for $m = 1, 2, 3$,

$$\sup_{t\in[0,\tau],\theta_n\in\Theta_n} \left\| W_{\cdot}^{(m)}(\theta_n, t) - w_{\cdot}^{(m)}(\theta_n, t) \right\| \longrightarrow_{\mathscr{P}} 0,$$

and moreover,

$$\sup_{t\in[0,\tau],\theta_n\in\Theta_n} \left\| W_{\cdot}^{(1,2)}(\theta_n, t) - w_{\cdot}^{(1,2)}(\theta_n, t) \right\| \longrightarrow_{\mathscr{P}} 0$$

$$\sup_{t\in[0,\tau],\theta_n\in\Theta_n} \left\| W_{\cdot}^{(2,2)}(\theta_n, t) - w_{\cdot}^{(2,2)}(\theta_n, t) \right\| \longrightarrow_{\mathscr{P}} 0$$

$$\sup_{t\in[0,\tau],\theta_n\in\Theta_n} \left\| W_{\cdot}^{(3,2)}(\theta_n, t) - w_{\cdot}^{(3,2)}(\theta_n, t) \right\| \longrightarrow_{\mathscr{P}} 0.$$

(6) For all $\theta_n \in \Theta_n$, $t \in [0, \tau]$, $w_{\cdot}^{(1)}(\cdot, t)$, $w_{\cdot}^{(2)}(\cdot, t)$, $w_{\cdot}^{(3)}(\cdot, t)$, $w_{\cdot}^{(1,2)}(\cdot, t)$, $w_{\cdot}^{(2,2)}(\cdot, t)$, $w_{\cdot}^{(3,2)}(\cdot, t)$ are continuous functions of $\theta_n \in \Theta_n$, uniformly in $t \in [0, \tau]$, and $w_{\cdot}^{(1)}(\theta_n, t)$, $w_{\cdot}^{(2)}(\theta_n, t)$, $w_{\cdot}^{(3)}(\theta_n, t)$ and $w_{\cdot}^{(1,2)}(\theta_n, t)$, $w_{\cdot}^{(2,2)}(\theta_n, t)$, $w_{\cdot}^{(3,2)}(\theta_n, t)$ are bounded on $\Theta_n \times [0, \tau]$. Let $u^{(2)}(\theta_n)$ denote $\nabla_{\theta_n}^2 g(X, \theta_n)$ and $w^{(2)}(\theta_n, t)$ denote the matrix with $\{w^{(2)}(\theta_n, t)\}_{jk} = w_{jk}^{(2)}(\theta_n, t)$ for all

$j, k = 1, \cdots, p_n(p_n+1)/2$. Then define

$$I(\theta_{n,0}) = \int_0^\tau \left\{ w^{(2)}(\theta_{n,0}, t) - u^{(2)}(\theta_{n,0}) \right\} s^{(0)}(\theta_{n,0}, t)\lambda_0(t)dt,$$

and let $\mathcal{I}_{an}(\beta_{a0}, \gamma_{a0})$ denote the submatrix of $I(\theta_{n,0})$ with respect to the non-zero $(\beta_{a0}, \gamma_{a0})$. It satisfies $0 < C_1 < \lambda_{\min}\{\mathcal{I}_{an}(\beta_{a0}, \gamma_{a0})\} \leq \lambda_{\max}\{\mathcal{I}_{an}(\beta_{a0}, \gamma_{a0})\} < C_2 < \infty$ for all n, where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ represent the smallest and largest eigenvalues of a matrix respectively.

Appendix 2: Proof of Theorem 1.

The log partial likelihood $l_n(\theta)$ can be written as

$$l_n(\theta) = \sum_{i=1}^n \int_0^\tau g(X_i, \theta)dN_i(s) - \int_0^\tau log\left[\sum_{i=1}^n Y_i(s)exp\left\{g(X_i, \theta)\right\}\right] d\widetilde{N}(s)$$

where $\widetilde{N}(\cdot) = \sum_{i=1}^n N_i(\cdot)$. By Theorem 4.1 and Lemma 3.1 of **?**, it follows that, for each θ in a neighbourhood of θ_0:

$$\frac{1}{n}\left\{l_n(\theta) - l_n(\theta_0)\right\} = \int_0^\tau \left[(\theta-\theta_0)^T s^{(1)}(\theta_0, t) - log\left\{\frac{s^{(0)}(\theta, t)}{s^{(0)}(\theta_0, t)}\right\} s^{(0)}(\theta_0, t)\right] \lambda_0(t)dt + O_p\left(\frac{\|\theta-\theta_0\|}{\sqrt{n}}\right).$$

Let $\eta_n = n^{-1/2} + \xi_n$, consider the C-ball $B_n(C) = \left\{\theta : \theta = \theta_0 + \eta_n\delta, \|\delta\| \leq C\right\}, C > 0$. For any $\theta \in B_n(C)$, by the second-order Taylor expansion of the log partial likelihood, and by the weak law of large numbers, we have

$$\frac{1}{n}\left\{l_n(\theta_0 + \eta_n\delta) - l_n(\theta_0)\right\} = \frac{1}{n}\nabla_\theta^T l_n(\theta_0)\eta_n\delta - \frac{1}{2}\eta_n^2\delta^T\left\{I(\theta_0) + o_p(1)\right\}\delta$$

where $\|\delta\| \leq C$. We further write $\delta = (u_1, ..., u_p, v_{12}, ..., v_{p-1,p})^T = (u^T, v^T)^T$. Then let

$$D_n(\delta) \equiv \frac{1}{n}\left\{Q_n(\theta_0 + \eta_n\delta) - Q_n(\theta_0)\right\}$$

$$= -\frac{1}{n}\left\{l_n(\theta_0 + \eta_n\delta) - l_n(\theta_0)\right\} + \sum_{j=1}^p \lambda_{j,n}^\beta\left(|\beta_{0j} + \eta_n u_j| - |\beta_{0j}|\right) + \sum_{j<j'} \lambda_{j,j',n}^\gamma\left(|\gamma_{0j,j'} + \eta_n v_{j,j'}| - |\gamma_{0j,j'}|\right)$$

$$\geq -\frac{1}{n}\left\{l_n(\theta_0 + \eta_n\delta) - l_n(\theta_0)\right\} - \eta_n^2\left(\sum_{\{j:\beta_{0j}\in\beta_{a0}\}}|u_j| + \sum_{\{(j,j'):\gamma_{0j,j'}\in\gamma_{a0}\}}|v_{j,j'}|\right)$$

$$\geq -\frac{1}{n}\left\{l_n(\theta_0+\eta_n\delta)-l_n(\theta_0)\right\}-\eta_n^2\left(|\beta_{a0}|+|\gamma_{a0}|\right)C \equiv A_1+A_2+A_3$$

where $|\cdot|$ measures the number of elements of the vector inside,

$$A_1 = -\frac{1}{n}\nabla_\theta l_n(\theta_0)\left(\eta_n\delta\right) = O_p(n^{-1/2})\left(\eta_n\delta\right)$$

$$A_2 = \frac{1}{2}\left(\eta_n\delta\right)^T\left\{I(\theta_0)+o_p(1)\right\}\left(\eta_n\delta\right) = \frac{1}{2}\left(\eta_n\delta_a\right)^T\left\{\mathcal{I}_a(\beta_{a0},\gamma_{a0})+o_p(1)\right\}\left(\eta_n\delta_a\right)$$

$$A_3 = -\eta_n^2\left(|\beta_{a0}|+|\gamma_{a0}|\right)C,$$

and δ_a is the sub-vector of δ correspond to non-zero (β_{a0},γ_{a0}). Notice that A_2 dominates A_1 and A_3 and is positive since $\mathcal{I}_a(\beta_{a0},\gamma_{a0})$ is positive definite. Therefore, for any given $\epsilon > 0$, there exists a large enough constant d such that

$$P\left\{\inf_{\theta\in B_n(d)} Q_n(\theta) > Q_n(\theta_0)\right\} \geq 1-\epsilon.$$

This implies that with probability at least $1-\epsilon$, there exists a local minimizer in the ball $B_n(C)$ such that $\left\|\hat{\theta}_n-\theta_0\right\| = O_p(\eta_n) = O_p(n^{-1/2})$.

Now for the sparsity, we first show $P(\hat{\beta}_{bn}=0)\to 1$. It is sufficient to show for any $\{j:\beta_{0j}\in\beta_{b0}\}$,

$$\frac{\partial Q_n(\hat{\theta}_n)}{\partial\beta_j} > 0 \text{ for } 0<\hat{\beta}_j<\epsilon_n \tag{1}$$

and

$$\frac{\partial Q_n(\hat{\theta}_n)}{\partial\beta_j} < 0 \text{ for } -\epsilon_n<\hat{\beta}_j<0 \tag{2}$$

with probability tending to 1, where $\epsilon_n = Cn^{-1/2}$ and $C>0$ is any constant. To show (1), notice

$$\frac{\partial Q_n(\hat{\theta}_n)}{\partial\beta_j} = -\frac{\partial l_n(\hat{\theta}_n)}{\partial\beta_j}+\lambda_{j,n}^{\beta}\text{sign}(\beta_j) = -\frac{\partial l_n(\theta_0)}{\partial\beta_j}-\sum_{k=1}^{p(p+1)/2}\frac{\partial^2 l_n(\theta_0)}{\partial\beta_j\partial\theta_k}\left(\hat{\theta}_k-\theta_{0k}\right)$$

$$-\sum_{k=1}^{\frac{p(p+1)}{2}}\sum_{l=1}^{\frac{p(p+1)}{2}}\frac{\partial^3 l_n(\tilde{\theta})}{\partial\beta_j\partial\theta_k\partial\theta_l}\left(\hat{\theta}_k-\theta_{0k}\right)\left(\hat{\theta}_l-\theta_{0l}\right)+\lambda_{j,n}^{\beta}\text{sign}(\beta_j),$$

where $\tilde{\theta}$ lies between $\hat{\theta}_n$ and θ_0. By the regularity conditions and $\left\|\hat{\theta}_n - \theta_0\right\| = O_p(n^{-1/2})$,

$$\frac{\partial Q_n(\hat{\theta}_n)}{\partial \beta_j} = \sqrt{n}\left\{O_p(1) + \sqrt{n}\lambda^{\beta}_{j,n}\mathrm{sign}(\hat{\beta}_j)\right\}.$$

As $\sqrt{n}\lambda^{\beta}_j \to \infty$ for $j \in \{j : \beta_{0j} \in \beta_{b0}\}$, the sign of $\partial Q_n(\hat{\theta}_n)/\partial \beta_j$ is dominated by $\mathrm{sign}(\hat{\beta}_j)$. Therefore,

$$P\left[\frac{\partial Q_n(\hat{\theta}_n)}{\partial \beta_j} > 0 \; for \; 0 < \hat{\beta}_j < \epsilon_n\right] \to 1 \; as \; n \to \infty$$

Similarly, we can show (2), and $P(\hat{\beta}_{bn} = 0) \to 1$ follows. We can similarly prove that $P(\hat{\gamma}_{bn} = 0) \to 1$.

For $(j, j') \in \{(j, j') : \gamma_{0j,j'} \in \gamma_{c0}\}$, without loss of generality, assume that $\beta_{0j} = 0$. Notice that $\hat{\beta}_j = 0$ implies $\hat{\gamma}_{j,j'} = 0$. Since we already have $P(\hat{\beta}_j = 0) \to 1$, we can conclude $P(\hat{\gamma}_{j,j'} = 0) \to 1$ as well, i.e. $P(\hat{\gamma}_{cn} = 0) \to 1$ as $n \to \infty$. Thus, we finish the proof for Part (i) of Theorem 1.

Next we show the asymptotic normality. Let $\widetilde{Q}_n(\theta_a)$ denote the objective function Q_n only on the nonzero component of θ, i.e. $\theta_a = (\beta_a^T, \gamma_a^T)^T$. We define $\theta_b = (\beta_b^T, \gamma_b^T, \gamma_c^T)^T$, and from the above derivation, we have $P\left(\hat{\theta}_b = 0\right) \to 1$. Thus,

$$P\left[\arg\min_{\theta_a} \widetilde{Q}_n(\theta_a) = \left(\theta_a - \text{component of } \arg\min_{\theta} Q_n(\theta)\right)\right] \to 1.$$

It means that $\hat{\theta}_a$ should satisfy

$$\frac{\partial \widetilde{Q}_n(\theta_a)}{\partial \theta_j}|_{\theta_a = \hat{\theta}_a} = 0, \quad \forall j \in \{j : \theta_j \in \theta_a\}$$

with probability tending to 1.

Let $\widetilde{l}_n(\theta_a)$ and $\widetilde{P}_\lambda(\theta_a)$ denote the log-likelihood function of θ_a and the penalty function of θ_a respectively so that we have

$$\widetilde{Q}_n(\theta_a) = -\widetilde{l}_n(\theta_a) + \widetilde{P}_\lambda(\theta_a)$$

Then

$$\nabla_{\theta_a}\widetilde{Q}_n(\hat{\theta}_a) = -\nabla_{\theta_a}\widetilde{l}_n(\hat{\theta}_a) + \nabla_{\theta_a}\widetilde{P}_\lambda(\hat{\theta}_a) = 0 \tag{3}$$

with probability tending to 1.

By Taylor expansion, it is easy to show that

$$B_1 = \nabla_{\theta_a} \widetilde{l}_n(\hat{\theta}_a) = \sqrt{n} \left[\frac{1}{\sqrt{n}} \nabla_{\theta_a} \widetilde{l}_n(\theta_{a0}) - \mathcal{I}_a(\beta_{a0}, \gamma_{a0}) \sqrt{n}(\hat{\theta}_a - \theta_{a0}) + o_p(1) \right]$$

and

$$B_2 = \nabla_{\theta_a} \widetilde{P}_\lambda(\hat{\theta}_a) = \left\{ \begin{bmatrix} \lambda^{\beta}_{j,n} \text{sign}(\beta_j) \\ \lambda^{\gamma}_{j,j',n} \text{sign}(\gamma_{j,j'}) \end{bmatrix}_{\beta_j \in \beta_a, \gamma_{j,j'} \in \gamma_a} + o_p(1)(\hat{\theta}_a - \theta_{a0}) \right\}.$$

Since we have $\left\| \hat{\theta}_a - \theta_{a0} \right\| = O_p(n^{-1/2})$, together with (3), we have

$$\sqrt{n}(\hat{\theta}_a - \theta_{a0}) = \mathcal{I}_a(\beta_{a0}, \gamma_{a0})^{-1} \frac{1}{\sqrt{n}} \nabla_{\theta_a} \widetilde{l}_n(\theta_{a0}) + o_p(1).$$

Part (ii) of Theorem 1 then follows by applying the central limit theorem.

Appendix 3: Proof of Theorem 2.

Similarly, under the regularity conditions in Appendix 1, we argue that there exists a local minimizer $\hat{\theta}_n$ of $Q_n(\theta)$ such that $\left\| \hat{\theta}_n - \theta_{n,0} \right\| = O_p(\sqrt{q_n}(n^{1/2} + \xi_n))$. Let $\eta_n = \sqrt{q_n}(n^{-1/2} + \xi_n)$, consider the C-ball $\{\theta_n = \theta_{n,0} + \eta_n \delta, \|\delta\| \le C\}, C > 0$. We define $D_n(\delta) \equiv \{Q_n(\theta_{n,0} + \eta_n \delta) - Q_n(\theta_{n,0})\}/n$, then for any $\delta = (u_1, ..., u_p, v_{12}, ..., v_{p-1,p})^T = (u^T, v^T)^T$ that satisfies $\|\delta\| \le C$, similarly as in Appendix 2, we have

$$D_n(\delta) \equiv \frac{1}{n} \{Q_n(\theta_{n,0} + \eta_n \delta) - Q_n(\theta_{n,0})\} \ge -\frac{1}{n} \{l_n(\theta_{n,0} + \eta_n \delta) - l_n(\theta_{n,0})\} - \eta_n \left(\sqrt{q_n} \xi_n \right) C$$

$$= -\frac{1}{n} \nabla^T_{\theta_n} l_n(\theta_{n,0}) \left(\eta_n \delta \right) + \frac{1}{2} \left(\eta_n \delta \right)^T \{I(\theta_{n,0}) + o_p(1)\} \left(\eta_n \delta \right) - \eta_n^2 C \equiv \widetilde{A}_1 + \widetilde{A}_2 + \widetilde{A}_3$$

where

$$\widetilde{A}_1 = -\frac{1}{n} \nabla^T_{\theta_n} l_n(\theta_{n,0}) \left(\eta_n \delta \right) \text{ and } \left| \widetilde{A}_1 \right| \le n^{-1/2} \eta_n O_p(\sqrt{q_n}) C = O_p(\eta_n^2) C,$$

$$\widetilde{A}_2 = \frac{1}{2} \left(\eta_n \delta \right)^T \{I(\theta_{n,0}) + o_p(1)\} \left(\eta_n \delta \right) = \frac{1}{2} \left(\eta_n \delta_a \right)^T \{\mathcal{I}_{an}(\beta_{a0}, \gamma_{a0}) + o_p(1)\} \left(\eta_n \delta_a \right), \quad \widetilde{A}_3 = \eta_n^2 C,$$

and δ_a is the sub-vector of δ correspond to non-zero $(\beta_{a0}, \gamma_{a0})$. Similarly, $\widetilde{A}_2$ dominates $\widetilde{A}_1$ and $\widetilde{A}_3$, and is positive since $\mathcal{I}_{an}(\beta_{a0}, \gamma_{a0})$ is positive definite. Therefore, for any given $\epsilon > 0$, there exists a large enough constant C such that

$$P\left\{\inf_{\|\delta\|\leq C} Q_n(\theta_{n,0}+\eta_n\delta) > Q_n(\theta_{n,0})\right\} \geq 1-\epsilon.$$

This implies that with probability at least $1-\epsilon$, there exists a local minimizer in the ball $B_n(C)$ such that $\left\|\hat{\theta}_n - \theta_{n,0}\right\| = O_p(\eta_n)$.

We now show $P(\hat{\beta}_{bn}=0)\to 1$. It is sufficient to show that for any $j \in \{j : \beta_{n,j} \in \beta_{bn}\}$,

$$\frac{\partial Q_n(\hat{\theta}_n)}{\partial \beta_{n,j}} > 0 \; for \; 0 < \hat{\beta}_{n,j} < \epsilon_n \tag{4}$$

$$\frac{\partial Q_n(\hat{\theta}_n)}{\partial \beta_{n,j}} < 0 \; for \; -\epsilon_n < \hat{\beta}_{n,j} < 0 \tag{5}$$

with probability tending to 1, where $\epsilon_n = Cn^{-1/2}$ and $C>0$ is any constant. Notice

$$\frac{\partial Q_n(\hat{\theta}_n)}{\partial \beta_{n,j}} = -\frac{\partial l_n(\hat{\theta}_n)}{\partial \beta_{n,j}} + \lambda_{j,n}^{\beta_n}\text{sign}(\beta_{n,j})$$

$$= -\frac{\partial l_n(\theta_{n,0})}{\partial \beta_{n,j}} - \sum_{k=1}^{q_n}\frac{\partial^2 l_n(\theta_{n,0})}{\partial \beta_{n,j}\partial\theta_{n,k}}\left(\hat{\theta}_{n,k}-\theta_{n,0k}\right)$$

$$-\sum_{k=1}^{q_n}\sum_{l=1}^{q_n}\frac{\partial^3 l_n(\tilde{\theta})}{\partial \beta_{n,j}\partial\theta_{n,k}\partial\theta_{n,l}}\left(\hat{\theta}_{n,k}-\theta_{n,0k}\right)\left(\hat{\theta}_{n,l}-\theta_{n,0l}\right) + \lambda_{j,n}^{\beta_n}\text{sign}(\beta_{n,j}),$$

where $\tilde{\theta}_n$ lies between $\hat{\theta}_n$ and $\theta_{n,0}$. By the regularity conditions, and notice $\left\|\hat{\theta}_n - \theta_{n,0}\right\| = O_p(\sqrt{q_n/n})$,

$$\frac{\partial Q_n(\hat{\theta}_n)}{\partial \beta_{n,j}} = \sqrt{nq_n}\left\{O_p(1) + \sqrt{\frac{n}{q_n}}\lambda_{j,n}^{\beta_n} sgn(\hat{\beta}_{n,j})\right\}.$$

As $\sqrt{n/q_n}\lambda_{j,n}^{\beta_n}\to\infty$ for $j \in \{j : \beta_{n,j}\in\beta_{bn}\}$, the sign of $\partial Q_n(\hat{\theta}_n)/\partial\beta_{n,j}$ is the same as $\text{sign}(\hat{\beta}_{n,j})$. Therefore,

$$P\left[\frac{\partial Q_n(\hat{\theta}_n)}{\partial \beta_{n,j}} > 0 \text{ for } 0 < \hat{\beta}_{n,j} < \epsilon_n\right] \to 1 \; as \; n\to\infty$$

and (4) holds with probability tending to 1. Parallel to this, one can show (5) holds with probability tending to 1.

Similar argument can be used to prove $P(\hat{\gamma}_{n,j,j'}=0)\to 1$ as $n\to\infty$, for $\hat{\gamma}_{n,j,j'}\in\hat{\gamma}_{bn}$, thus $P(\hat{\gamma}_{bn}=0)\to 1$.

For $\hat{\gamma}_{n,j,j'} \in \hat{\gamma}_{cn}$, without loss of generality, assume that $\beta_{n,0j} = 0$. Notice that $\hat{\beta}_{n,j} = 0$ implies $\hat{\gamma}_{n,j,j'} = 0$, because if $\hat{\gamma}_{n,j,j'} \neq 0$, then the value of the loss function does not change but the value of the penalty function increases. Therefore, $P(\hat{\gamma}_{n,j,j'} = 0) \to 1$ follows since we have already shown $P(\hat{\beta}_{n,j} = 0) \to 1$.

Thus, Part (i) of Theorem 2 is proved.

Now, we prove the asymptotic normality. Denote $\theta_{an} = (\beta_{an}^T, \gamma_{an}^T)^T$, then

$$\sqrt{n}\Omega_n I_{an}^{1/2}(\theta_{an,0}) \left(\hat{\theta}_{an} - \theta_{an,0}\right) = \sqrt{n}\Omega_n I_{an}^{-1/2}(\theta_{an,0}) I_{an}(\theta_{an,0}) \left(\hat{\theta}_{an} - \theta_{an,0}\right)$$

$$= \sqrt{n}\Omega_n \mathcal{I}_{an}^{-1/2}(\beta_{a0}, \gamma_{a0}) \left\{ \frac{1}{n} \nabla l_n(\theta_{an,0}) + o_p(n^{-1/2}) \right\}$$

$$= \frac{1}{\sqrt{n}} \Omega_n \mathcal{I}_{an}^{-1/2}(\beta_{a0}, \gamma_{a0}) \sum_{i=1}^{n} \left[\nabla l_n(\theta_{an,0})\right] + o_p\left(1\right) \equiv \sum_{i=1}^{n} Y_{ni} + o_p\left(1\right),$$

where $Y_{ni} = n^{-1/2}\Omega_n \mathcal{I}_{an}^{-1/2}(\beta_{a0}, \gamma_{a0}) \sum_{i=1}^{n} \left[\nabla l_n(\theta_{an,0})\right]$.

We now show that with probability tending to 1, $\sum_{i=1}^{n} Y_{ni} + o_p\left(1\right) \to_d N\left(0, \Sigma\right)$:

(i) We first show $\mathcal{I}_{an}(\beta_{a0}, \gamma_{a0}) \left(\hat{\theta}_{an} - \theta_{an,0}\right) = n^{-1}\nabla l_n(\theta_{an,0}) + o_p(n^{-1/2})$. With probability tending to 1,

$$0 = \nabla_{\theta_{an}} Q_n(\hat{\theta}_{an}) = -\frac{1}{n} \nabla_{\theta_{an}} l_n(\hat{\theta}_{an}) + \nabla_{\theta_{an}} \left\{ \sum_{\{j:\beta_{n,0j} \in \beta_{a0}\}} \lambda_{j,n}^{\beta_n} \hat{\beta}_{n,0j} + \sum_{\{(j,j'):\gamma_{n,0j,j'} \in \gamma_{a0}\}} \lambda_{j,j',n}^{\gamma_n} \hat{\gamma}_{n,0j,j'} \right\}.$$

Taking Taylor Expansion at $\theta_{an} = \theta_{a0}$, we have

$$0 = -\nabla_{\theta_{an}} l_n(\theta_{a0}) - \left[\nabla_{\theta_{an}}^2 l_n(\theta_{a0})\right] \left(\hat{\theta}_{an} - \theta_{a0}\right) - \frac{1}{2} \left(\hat{\theta}_{an} - \theta_{a0}\right)^T \left[\nabla_{\theta_{an}}^2 \left(\nabla_{\theta_{an}} l_n(\theta_{a0})\right)\right] \left(\hat{\theta}_{an} - \theta_{a0}\right)$$

$$+ n\nabla_{\theta_{an}} \left\{ \sum_{\{j:\beta_{n,0j} \in \beta_{a0}\}} \lambda_{j,n}^{\beta_n} \beta_{n,0j} + \sum_{\{(j,j'):\gamma_{n,0j,j'} \in \gamma_{a0}\}} \lambda_{j,j',n}^{\gamma_n} \gamma_{n,0j,j'} \right\}.$$

Thus,

$$\mathcal{I}_{an}^{-1/2}(\beta_{a0}, \gamma_{a0}) \left(\hat{\theta}_{an} - \theta_{a0}\right) = -\frac{1}{n} \nabla_{\theta_{an}}^2 l_n(\theta_{a0}) \left(\hat{\theta}_{an} - \theta_{a0}\right) + \left\{ \mathcal{I}_{an}^{-1/2}(\beta_{a0}, \gamma_{a0}) + \frac{1}{n} \nabla_{\theta_{an}}^2 l_n(\theta_{a0}) \right\} \left(\hat{\theta}_{an} - \theta_{a0}\right)$$

$$= \frac{1}{n} \nabla_{\theta_{an}} l_n(\theta_{a0}) + \frac{1}{2n} \left(\hat{\theta}_{an} - \theta_{a0}\right)^T \left[\nabla_{\theta_{an}}^2 \left(\nabla_{\theta_{an}} l_n(\theta_{a0})\right)\right] \left(\hat{\theta}_{an} - \theta_{a0}\right)$$

$$-\nabla_{\theta_{an}}\left\{\sum_{\{j:\beta_{n,0j}\in\beta_{a0}\}}\lambda_{j,n}^{\beta_n}\beta_{n,0j}+\sum_{\{(j,j'):\gamma_{n,0j,j'}\in\gamma_{a0}\}}\lambda_{j,j',n}^{\gamma_n}\gamma_{n,0j,j'}\right\}$$

$$+\left\{\mathcal{I}_{an}^{-1/2}(\beta_{a0},\gamma_{a0})+\frac{1}{n}\nabla_{\theta_{an}}^2 l_n(\theta_{a0})\right\}\left(\hat{\theta}_{an}-\theta_{a0}\right).$$

Therefore, it is sufficient to show that

$$\frac{1}{2n}\left(\hat{\theta}_{an}-\theta_{a0}\right)^T\left[\nabla_{\theta_{an}}^2\left(\nabla_{\theta_{an}}l_n(\theta_{a0})\right)\right]\left(\hat{\theta}_{an}-\theta_{a0}\right)$$

$$-\nabla_{\theta_{an}}\left\{\sum_{\{j:\beta_{n,0j}\in\beta_{a0}\}}\lambda_{j,n}^{\beta_n}\beta_{n,0j}+\sum_{\{(j,j'):\gamma_{n,0j,j'}\in\gamma_{a0}\}}\lambda_{j,j',n}^{\gamma_n}\gamma_{n,0j,j'}\right\}$$

$$+\left\{\mathcal{I}_{an}(\beta_{a0},\gamma_{a0})+\frac{1}{n}\nabla_{\theta_{an}}^2 l_n(\theta_{a0})\right\}\left(\hat{\theta}_{an}-\theta_{a0}\right)=o_p(n^{-1/2}).$$

Denote the three terms in the above equation as D_1, D_2, and D_3. First, by Cauchy-Schwarz inequality,

$$\|D_1\|^2\leq\frac{1}{4n^2}\left\|\nabla_{\theta_{an}}^2\left(\nabla_{\theta_{an}}l_n(\theta_{an,0})\right)\right\|^2\left\|\hat{\theta}_{an}-\theta_{a0}\right\|^4$$

$$=\frac{1}{4n^2}\sum_{\{(j,k,l):\theta_{n,j},\theta_{n,k},\theta_{n,l}\in\theta_{an}\}}n^2O_p(1)O_p(\frac{q_n^2}{n})=O_p(q_n^5/n^2)=o_p(1/n)$$

Secondly, because $\xi_n=o(1/\sqrt{nq_n})$,

$$\|D_2\|^2=\left\|\left(\lambda_{1,n}^{\beta_n}\text{sign}(\beta_{n,01}),\ldots,\lambda_{p_{n-1},p_n,n}^{\gamma_n}\text{sign}(\gamma_{n,0(p_{n-1},p_n)})\right)^T\right\|^2$$

$$\leq|\theta_{an}|\,\xi_n^2=|\theta_{an}|\,o(1/nq_n)=o_p(1/n)$$

Third, it can be shown that

$$\|D_3\|^2\leq\left\|\mathcal{I}_{an}(\beta_{a0},\gamma_{a0})+\frac{1}{n}\nabla_{\theta_{an}}^2 l_n(\theta_{a0})\right\|^2\left\|\hat{\theta}_{an}-\theta_{a0}\right\|^2$$

$$=o_p(1/q_n^2)O_p(q_n/n)=o_p(1/nq_n)=o_p(1/n)$$

Therefore, $D_1+D_2+D_3=o_p(n^{-1/2})$.

Next, we show $\sum_{i=1}^n Y_{ni}+o_p(1)\longrightarrow_d N(0,\Sigma)$. It is sufficient to show that Y_{ni}, $i=1,\ldots,n$ satisfies the conditions for Lindeberg-Feller central limit theorem. For any given $\epsilon>0$, by

Cauchy-Schwarz inequality,

$$\sum_{i=1}^{n} E\left[\|Y_{ni}\|^2 I\left\{\|Y_{ni}\| > \epsilon\right\}\right] = nE\left[\|Y_{ni}\|^2 I\left\{\|Y_{ni}\| > \epsilon\right\}\right] \leq nD_4^{1/2} D_5^{1/2}$$

where $D_4 = \left[E\|Y_{ni}\|^4\right]$ and $D_5 = E\left\{I\left(\|Y_{ni}\| > \epsilon\right)\right\}$. Note

$$D_4 = \frac{1}{n^2} E\left\|\Omega_n \mathcal{I}_{an}^{-1/2}(\beta_{a0}, \gamma_{a0}) \nabla_{\theta_{an}} l_n(\theta_{a0})\right\|^4$$

$$\leq \frac{1}{n^2}\left\|\Omega_n^T \Omega_n\right\|^2 \left\|I_{an}(\theta_{a0})\right\|^{-2} E\left\|\nabla_{\theta_{an}}^T l_n(\theta_{a0}) \nabla_{\theta_{an}} l_n(\theta_{a0})\right\|^2$$

$$= \frac{1}{n^2}\lambda_{max}^2(\Omega_n^T \Omega_n)\lambda_{max}^2\left\{I_{an}^{-1}(\theta_{a0})\right\} O(|\theta_{an}|^2) = O(q_n^2/n^2).$$

By Markov inequality,

$$D_5 = E\left\{I\left(\|Y_{ni}\| > \epsilon\right)\right\} = P\left(\|Y_{n1}\| > \epsilon\right) \leq \frac{E\|Y_{n1}\|^2}{\epsilon^2} = O(q_n/n).$$

Therefore,

$$\sum_{i=1}^{n} E\left[\|Y_{ni}\|^2 1\{\|Y_{ni}\| > \epsilon\}\right] \leq nO(q_n/n)O(\sqrt{q_n/n}) = o(1),$$

and part (ii) of Theorem 2 follows.