The Pennsylvania State University
The Graduate School

MODEL SELECTION AND SURVIVAL ANALYSIS WITH APPLICATION TO LARGE TIME-VARYING NETWORKS

A Dissertation in
Statistics,
by
Xizhen Cai

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2014

The dissertation of Xizhen Cai was reviewed and approved* by the following:

David R. Hunter
Head and Professor of Statistics
Dissertation Co-Advisor and Co-Chair of Committee

Runze Li
Distinguished Professor of Statistics and Public Health Sciences
Dissertation Co-Advisor and Co-Chair of Committee

Debashis Ghosh
Professor of Statistics and Public Health Sciences

Marcel Salathé
Assistant Professor of Biology
Adjunct Faculty in Computer Science and Engineering

Aleksandra Slavkovic
Associate Professor of Statistics and Associate Head for Graduate Studies

*Signatures are on file in the Graduate School.

Abstract

Survival models have been applied to time-to-event data for a long time, and usually a number of covariates are assumed to influence the distribution of the time to event through the model. The Cox proportional hazard model is commonly used in this context. To have a parsimonious model without losing consistency in estimation, several authors have extended the variable selection techniques of Fan and Li (2001) to survival settings. For example, the variable selection problem for the Cox model is studied in Fan and Li (2002). Recently, survival models like the Cox model are also extended to apply to dynamic network data (Vu et al., 2011b; Perry and Wolfe, 2013), where the observations are dependent. In this dissertation, we study the variable selection problem for a survival model other than the Cox model. In addition, we extend the variable selection work to the dynamic network model setting.

We first discuss the problem of variable selection for the proportional odds model, an alternative to Cox's model, and show how to maximize the penalized profile likelihood to estimate parameters and select variables simultaneously. Using a novel application of the semiparametric theory developed by Murphy and Van der Vaart (2000), we derive asymptotic properties of the resulting estimators, including consistency results and the oracle property. In addition, we propose algorithms to maximize the penalized likelihood estimator based on a majorization-minimization (MM) algorithm. Tests on simulated and real data sets demonstrate that the newly proposed algorithm performs well in practice.

Next, we extend the penalization idea to the Cox model in an egocentric approach to dynamic networks, and select covariates by maximizing the penalized partial likelihood function. Asymptotic properties of both the unpenalized and penalized partial likelihood estimates are developed under certain regularity conditions. We also implement the estimation and test the prediction performance of these estimates in a citation network. Since the covariates are time-varying, the

computation cost is high. After variable selection, the model is reduced, which simplifies the calculation for future predictions. Another method to reduce the computational complexity is to use the case-control approximation, in which instead of using all the at-risk nodes in the network, only a subset is sampled to evaluate the partial likelihood function. By using this approximation, the computation time is shortened dramatically, while the prediction performance is still satisfactory in the citation network.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First of all, I am sincerely grateful to my dissertation advisors, Dr. David Hunter, and Dr. Runze Li for their valuable guidance, inspirational encouragement and helpful comments on my academic career. Without their supervision and support, it would be impossible for me to get progress in the academic research area and complete this dissertation. Second, I wish to express my propound appreciation to my committee members, Dr. Ghosh, and Dr. Salathé, for their precious time and comments to improve this dissertation.

Further, I am highly grateful to every member of my family, especially my mother and my husband. Without their love, support and understanding, I could not be here and complete the dissertation.

Finally, I would like to say thank you to all who helped and supported me during the completion of the dissertation, to all who ever helped and guided me in my life.

Chapter 1

Introduction and Literature Reviews

1.1 Introduction

Systematic collection of a group of explanatory variables to predict and summarize the pattern of response variables has been used very commonly in various area of science. Especially in recent years, the development of technology brings the possibility of obtaining massive data and data with complicated structures to reduce potential model bias. However, the tradeoff is that models can become extremely complicated by introducing too many covariates, or by giving too much freedom on coefficients by assuming a nonparametric form. This motivates researchers to find proper approaches simplifying the model while keeping the accuracy of model estimation or prediction. One approach is to select a smaller number of explanatory variables using variable selection techniques.

Variable selection techniques have been well-developed for the linear regression and general linear model settings. Properly selecting the set of variables can reduce the prediction error while only introducing a tolerable bias. Best subset selection through classic variable selection criteria, including Adjusted R^2, Mallow's C_P, AIC, and BIC, might result in the best model. However, it is not workable for models with a large number of covariates because it is computationally intensive. Stepwise selection is too computationally simple but has some significant drawbacks. In recent years, approaches using penalization have developed quickly. Tibshirani (1996) introduced the idea of shrinkage by proposing a restricted least squares approach. The resulting estimates shrink to zero as compared with the

original least squares estimates, and some of them will be set to zero. This approach is equivalent to penalized least squares using the L_1 penalty. Fan and Li (2001) studied properties of penalized least squares and penalized likelihood estimates using different penalty functions, and proposed the SCAD penalty. They showed that under certain conditions and properly chosen penalty functions, penalized likelihood estimates possess an oracle property, i.e. the zero coefficients can be identified correctly with probability approaching to 1, and those nonzero coefficients could be estimated as if the zero coefficients were known to be zero in advance, as sample size goes to infinity. Sparsity of the estimate depends on singularity of the penalty function at the origin, which creates computational issues for maximization of the penalized likelihood function. Fan and Li (2001) proposed using local quadratic approximation (LQA) of the penalty term. Hunter and Li (2005) showed this algorithm is a type of minorization-maximization algorithm and proposed a slightly perturbed version of LQA. Zou and Li (2008) proposed using local linear approximation (LLA) for the penalty term, which is known to be the best convex approximation for penalty terms. Details of this literature will be discussed in Section 2 of this chapter.

Besides linear regression and general linear models, survival models have also been used widely. The response variable in a survival model setting is time to the event of interest, and researchers are interested in how this time to event variable is affected by a set of covariates. Typical approaches for estimation in a survival analysis are likelihood-based. These likelihood-based approaches usually maximize some type of "likelihood" such as partial likelihood, profile likelihood, or even marginal likelihood to estimate the coefficients of the covariates. The penalized version of these "likelihoods" can be used for model selection. Fan and Li (2002) studied the variable selection problem for Cox's proportional hazard model and derived the oracle properties for the estimator. The Cox proportional hazards model is widely used because it has a partial likelihood which excludes the nonparametric baseline hazard parameter and hence facilitates the estimation. Another semi-parametric survival model is the proportional odds model, which can be used as an alternative to the Cox model but which does not have a partial likelihood function. This brings many challenges for variable selection in the proportional odds model. In Chapter 2 of this dissertation, we discuss these chal-

lenges and study the variable selection problem for the proportional odds model through the penalized profile likelihood approach. In addition, efficient algorithms are developed based on an MM algorithm philosophy. In Section 3, more details about comparisons between Cox's model and the proportional odds model will be discussed.

Survival models like the Cox model and the Aalen additive model are also used in settings when one does not have independent observations, for example, in analysis of network data. Usually the counting process approach is adopted to relax the assumption of independent observations. Especially in recent years, the study of network data has attracted a lot of attention and the methodology has developed quickly in a variety of fields. Researchers have developed more interest in large and time-varying networks. Some approaches model the intensity process for nodes or edges using the Cox or Aalen model (Vu et al., 2011a,b; Perry and Wolfe, 2013). In the Section 4 of this chapter, some related models will be reviewed and discussed. These models encompass the dynamic property of time-varying networks and also include time-varying covariates, which can include numerous aspects to characterize various dynamic properties of the networks. We can obtain "snapshots" of such networks even though sometimes they change and grow quickly. However, the tradeoff is that the estimation for the model can become extremely complicated if we introduce a large number of covariates. This motivates us to find proper approaches simplifying the model while keeping the accuracy of model estimation or prediction. In other words, we want to build suitable survival models for large, time-varying networks and apply model selection techniques on these models. In this disseration, we apply the idea of penalization to select variables in a network setting using survival models. Chapter 3 will discuss theoretical properties of the maximizer of the penalized likelihood function and Chapter 4 provides some algorithms to implement the optimization.

1.2 Variable Selection Techniques

Many settings in statistical modeling require including a large number of covariates to reduce model bias in the initial stage of model fitting. Also, a revolution in data collection technology has made it possible to obtain observations for a huge

number of variables at the same time. Although including more variables results in smaller bias for the model, it also has problems associated with it. One problem is lack of freedom in estimation: in some statistical settings, the number of observations one can obtain is far fewer than the number of variables in the model. For regression problems, this can cause collinearity and hence explosion of variance. Another problem is the curse of dimensionality, since in high-dimensional settings, independency among variables can become very tricky. Finally large numbers of variables will increase the model complexity as well as the computational burden. All these call for model and variable selection methods that balance between statistical accuracy and model and computational complexity.

The rest of this section is organized as follows: the first subsection reviews literature for classic variable selection criteria, and the next two subsections introduce variable selection approaches through penalized least squares or likelihood.

1.2.1 Subset Selection and Classic Model Selection Criterion

1.2.1.1 Linear Regression Settings

Variable selection techniques are first developed for linear regression. For a simple linear regression model, we assume

$$y = \boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{x} + \epsilon, \epsilon \sim N(0,1)$$

where $\mathbf{x} = (x_1, x_2, ..., x_p)$ is a p-dimensional vector, independent of $\boldsymbol{\epsilon}$. Observations are n i.i.d pairs $(\mathbf{x}^{(i)}, y^{(i)})$. Using matrix notation, let $\mathbf{X} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(n)})$ and $\mathbf{Y} = (y^{(1)}, y^{(2)}, ..., y^{(n)})$. If $n > p$, the least squares estimator(LSE) of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{1.1}$$

If $\boldsymbol{h}$ is a q dimension vector, and C is a $q \times p$ matrix, the major inference problem for $\boldsymbol{\beta}$ is to test

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{h} \text{ versus } H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{h}. \tag{1.2}$$

If $\hat{\boldsymbol{\beta}}$ is the least squares estimate, then the constrained least squares estimate

$\hat{\boldsymbol{\beta}}_0$ under the null hypothesis is

$$\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\{\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\}^{-1}(\mathbf{C}\boldsymbol{\beta} - \mathbf{h}). \tag{1.3}$$

One may use the corresponding F-test to solve this testing problem:

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/q}{\mathrm{RSS}_1/(n-p)} \sim F_{q,n-d}, \tag{1.4}$$

where RSS_0 and RSS_1 are the residual sum of squares under the null and alternative hypotheses, defined as:

$$\mathrm{RSS}_0 = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_0\|^2, \mathrm{RSS}_1 = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2. \tag{1.5}$$

It can be shown that the F-test is equivalent to the likelihood ratio test under the normal assumption on the error term.

To improve model predictability, one can select significant variables by comparing models with different subsets of predictors, which is also necessary when $n < p$. The selection criterion should have good properties so that the model selected enjoys desired advantages.

1.2.1.2 Classic Model Selection Criteria

There is a lot of literature about criteria for variable subset selection. The most commonly used ones include C_p, AIC and BIC. Suppose for each candidate model, we can calculate least square estimates and likelihoods, and all the following criteria are given for models with d predictors.

The most intuitive criterion would be the residual sum of squares (RSS). The smaller the RSS is, the better the fit it might be. People use

$$R_d^2 = 1 - \mathrm{RSS}_d/\mathrm{RSS}_0 \tag{1.6}$$

as one criterion due to this intuition. However, including more variables in a model will definitely improve the model accuracy, hence, decrease the RSS_d, and increase the R_d^2. So R_d^2 cannot serve as a variable selection criterion. The Adjusted R_d^2 is an improved version of R_d^2, and unlike R_d^2, it will not necessarily increase when more variables are added into the model. The Adjusted R_d^2 is also referred to as

Fisher's A-statistic, and it adjusts the degrees of freedom as follows:

$$A_d = 1 - (1 - R_d^2)\frac{n-1}{n-d}. \tag{1.7}$$

While comparing candidate models, one wants to choose the model with greatest A_d.

Another criterion, which has the same intuition, is Mallow's C_P, which tries to minimize the scaled sum of squared errors,

$$J_d = \frac{1}{\sigma^2}\|\mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{X}\boldsymbol{\beta}\|^2, \tag{1.8}$$

where $\mathbf{X}_1$ represents the design matrix with only d variables in the candidate model and $\boldsymbol{\beta}_1$ are the corresponding coefficients. Then C_P is an unbiased estimator of $E(J_d|\mathbf{X})$ and it is defined as

$$\mathrm{C}_P = \frac{\mathrm{RSS}_d}{\sigma^2} - (n - 2d). \tag{1.9}$$

The model with the minimum value of C_P is preferable.

Prediction is always a very important perspective for linear regression, so models with smaller prediction errors might be more favorable. Based on this perspective, Allen (1974) proposed a variable selection criterion called the prediction sum of squares (PRESS) statistic,

$$\mathrm{PRESS}_d = \sum_{i=1}^{n}(y_i - \hat{y}_{id})^2, \tag{1.10}$$

where $\hat{y}_{id}$ is the predicted value by a model with d variables, which is estimated by all but the ith observation. Similarly, the Cross Validation (CV) approach also puts a small subset of data aside, and uses the rest to calculate the prediction error. CV can be one-fold or K-fold, depending on how many subsets the observations are divided into. One disadvantage of this approach is that one needs to fit several regression models to evaluate one criterion value. This might be computationally intensive when n is large. To avoid this, Golub et al. (1979) found that if n is much larger than d and under some other mild conditions, the PRESS_d statistic can be

approximated by

$$\text{PRESS}_d \approx \frac{RSS_d}{(1 - d/n)^2}. \tag{1.11}$$

The right hand side of (1.11) is actually n times the so-called Generalized Cross Validation (GCV) criterion, which is computationally easier and suitable for broader types of problems.

The remaining two criteria depend on log-likelihood. One is the AIC, or Akaike's Information Criterion, (Akaike, 1973),

$$\text{AIC}_d = -2\ell(\hat{\boldsymbol{\beta}}_{\boldsymbol{d}}) + 2d. \tag{1.12}$$

It is motivated by estimating the Kullback-Leibler divergence between maximized likelihood values evaluated at the true and candidate models. It can be shown that AIC is equivalent to C_P asymptotically. The other criterion is the Bayesian Information Criterion, or BIC (Schwarz, 1978),

$$\text{BIC}_d = -2\ell(\hat{\boldsymbol{\beta}}_{\boldsymbol{d}}) + \log(n)d. \tag{1.13}$$

As variable selection criteria, AIC and BIC use different penalties on the size of the model. If one assumes that the true model is in the candidate class of models, BIC will choose the true model consistently as sample size increases. On the other hand, without this assumption, the model selected by AIC is asymptotically loss-efficient.

1.2.1.3 Algorithms for Variable Selection

For linear regression with a finite number of variables, it is possible to list all sub-models, and pick the one with the best variable selection criterion. However, exhaustive search over all subsets for moderate p will be very computationally intensive and may not be feasible for large p. Instead, we may use forward selection, backward elimination, or stepwise selection as algorithms for variable selection.

Forward selection starts from the null model, then adds one variable at a time. The variable chosen to be added is the one that increases the selected criterion the most. Backwards elimination begins with the full model with all variables, and deletes the one that increases the selected criterion the most at each step. Both

procedures stop if the variable selection criterion reaches its optimum value. The drawback of these procedures is that once you add or delete a variable, you can not delete or add it ever again. So they do not necessarily result in the same subset that would be chosen by best subset selection. On the other hand, stepwise selection can be viewed as a combination of forward and backward selection. Unlike them, stepwise selection allows both adding a variable and deleting a variable in each step, and it will not stop until no variables can be added to or deleted from the current subset.

1.2.2 Penalized Least Square Approach for Linear Regression

Some of the variable selection criteria above can be written as scaled RSS plus an additional term which penalizes variables with large coefficients. These criteria are all under the framework of Penalized Least Square (PLS) approaches, which have attracted a lot of attention in recent years. One advantage is that one can conduct variable selection and parameter estimation simultaneously. A penalized least squares function could be written as

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + n\sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{1.14}$$

where $p_\lambda(\cdot)$ is the penalty function and λ is a tuning parameter which controls model complexity. If the penalty properly is chosen properly, minimization of this function can shrink some coefficients to zero, hence resulting in a sparse estimator. That is the reason this approach is suitable for variable selection.

1.2.2.1 Types of Penalty Functions

In regression settings, best subset regression with C_P, AIC, and BIC are equivalent to minimizing the following functions, respectively:

$$\frac{1}{2}\|\mathbf{y} \quad - \quad \mathbf{X}\boldsymbol{\beta}\|^2 + \sigma^2\sum_{j=1}^{p} I(\beta_i \neq 0), \tag{1.15}$$

$$\frac{1}{2}\|\mathbf{y} \quad - \quad \mathbf{X}\boldsymbol{\beta}\|^2 + n\frac{(\sigma^2\sqrt{2/n})}{2}\sum_{j=1}^{p} I(\beta_i \neq 0), \tag{1.16}$$

$$\frac{1}{2}\|\mathbf{y} \quad - \quad \mathbf{X}\boldsymbol{\beta}\|^2 + n\frac{(\sigma^2\sqrt{\log(n)/n})}{2}\sum_{j=1}^{p} I(\beta_i \neq 0). \tag{1.17}$$

Compared with the Penalized Least Squares problem (1.14), the common penalty function

$$p_\lambda(\boldsymbol{\beta}) = \lambda\sum_{i=1}^{p} I(\beta_i \neq 0) \tag{1.18}$$

is the L_0 penalty, which is also called the entropy penalty. It penalizes the number of variables in the model, hence it will result in a sparse model. However, it is difficult to optimize due to the discontinuity of the penalty function at the origin. Each of (1.15), (1.16) and (1.17) is a special case of (1.14) using the L_0 penalty.

In order to make the penalized least squares easy to minimize, the penalty term may be relaxed. Hoerl and Kennard (1970) introduced penalized least squares with an L_2 penalty:

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \frac{n\lambda}{2}\|\boldsymbol{\beta}\|^2. \tag{1.19}$$

For a fixed λ, the solution has a closed form:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + n\lambda I_d)^{-1}\mathbf{X}^T\mathbf{y}, \tag{1.20}$$

which is also referred to as a ridge regression estimator and which was originally used as an estimator of β when $\mathbf{X}$ has collinearity among its columns. It has been used widely due to its computational simplicity and can give more accurate predictions than subset regression if the true model is not sparse. In this setting, the estimate from ridge regression is better than ordinary least squares because it shrinks the estimates selectively, hence increasing the accuracy of the estimates (Breiman, 1995). On the other hand, for sparse models, subset selection can beat ridge regression in terms of giving a sparse model. In this sense, ridge regression is not suitable for variable selection.

As argued by Breiman (1995), prediction accuracy may be improved by shrinking some coefficients. If the goal is variable selection, it may also be helpful to set some of the variables to zero. Breiman (1995) introduced the non-negative garrote

based on constrained least squares. It gives each coefficient a shrinkage scale, then constrains the magnitude of the sum of these scales. This method can eliminate some variables and shrink others, and it is relatively stable. Also, the optimization problem can be solved by quadratic programming, hence relatively simply. However, the sign and magnitude of the estimates depend on ordinary least squares estimates, which incurs difficulty in some settings, for example, collinearity or overfit. Tibshirani (1996) introduced LASSO, which is also based on constrained least squares estimates, and which is equivalent to penalized least squares with L_1 penalty:

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + n\lambda \sum_{j=1}^{p} |\beta_j|. \tag{1.21}$$

LASSO avoids using ordinary least squares estimates directly, and it can give different signs from the least square estimates. Efron et al. (2004) introduced Least Angle Regression (LARS) to find least squares estimates stepwise, and its modified version can be used to generate the entire path of LASSO solutions. This algorithm is stable and computationally economical, hence is used widely as an algorithm to solve least squares with L_1 penalties. In general, penalized L_q $(0 \leq q \leq 2)$ least squares estimates are called bridge regression estimators since they bridge L_0 and L_2.

1.2.2.2 Choice of Penalty Functions and SCAD Penalty

As discussed above, there are various penalty functions, and some of them will shrink estimates selectively and set some estimates to zero. So properly choosing a penalty function can facilitate variable selection as well as improving estimation accuracy. Some of the penalties are displayed below.

One may ask what kind of penalty function is the best in the setting of variable selection. To understand this, Fan and Li (2001) started from the simplest case, where X is orthonormal such that $\frac{1}{n}\mathbf{X}^T\mathbf{X} = I_p$, then they considered the simplest penalized least squares:

$$\frac{1}{2}(z - \theta)^2 + p_\lambda(|\theta|). \tag{1.22}$$

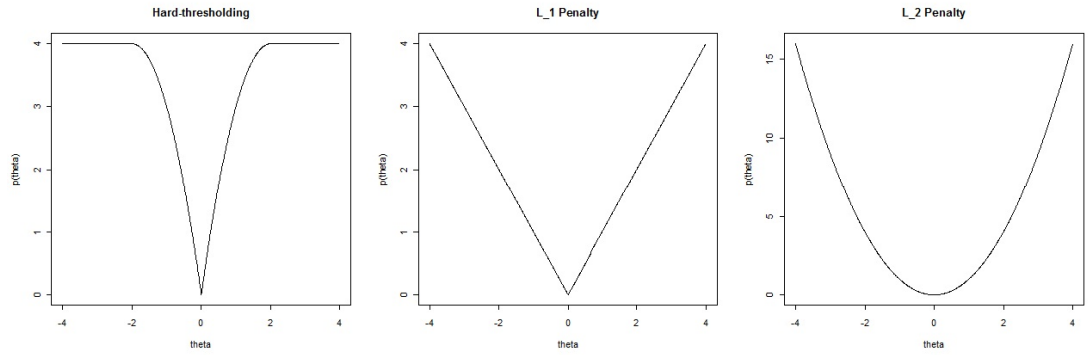

Figure 1.1. Plots of different penalty functions, $\lambda = 2$

Under the L_0 penalty, the minimizer is the hard-thresholding estimator,

$$\hat{\theta} = zI(|z| \geq \lambda), \tag{1.23}$$

which sets very small coefficients to zero directly. Penalized least squares with hard-thresholding penalty,

$$p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda), \tag{1.24}$$

also results in the hard-threshold estimator.

The penalized least squares estimator resulting from L_1 has the closed from solution

$$\hat{\theta} = \text{sgn}(z)(|z| - \lambda)_+. \tag{1.25}$$

Figure 1.2 displays the relation between ordinary least squares estimates and penalized least squares estimates under different penalty functions. The x-axis is the penalized least squares estimate, and the y-axis is the ordinary least squares estimate. The dotted line in each plot indicates the case when the two estimates are equal. We can examine the properties of penalized least squares estimates with different penalty functions from this plot. Best subset selection can set small coefficients to zero and keep large estimates the same as ordinary least squares estimates, hence unbiased. However, it is not stable since there is a jump point around the value of λ. For LASSO (or L_1), the estimate is continuous as the ordinary least squares estimate changes, hence more stable. But it sets small coefficients to zero

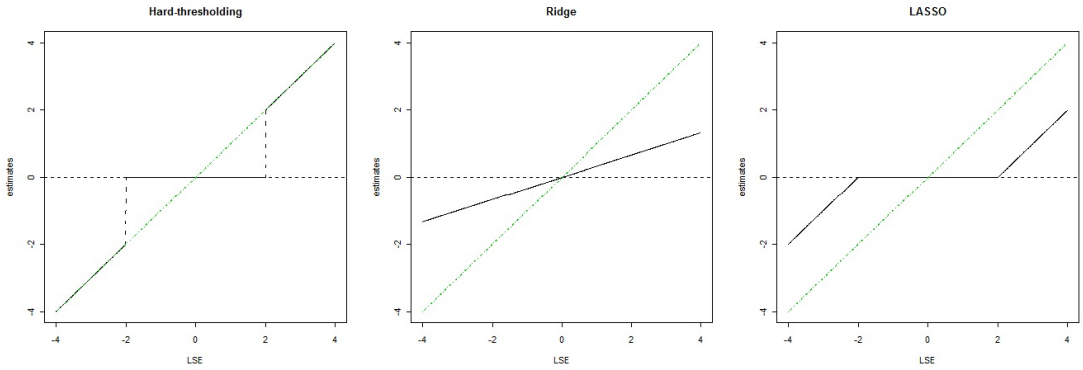

Figure 1.2. Estimates from different penalty functions, and in all cases, λ is set to be 2

and also shrinks the estimates when they are large. The purpose of shrinking is to reduce the prediction error, so the shrinkage has to be selective: shrink small coefficients to zero while keeping large estimates the same as ordinary least square estimates. Ridge regression shrinks estimates while never setting to zero any coefficients, hence it does not give sparse estimates. To summarize, none of the three penalty functions fulfills the requirements of variable selection and reduction of prediction error: hard-thresholding is not continuous, LASSO is biased, and L_2 is not sparse.

Based on these observations, Fan and Li (2001) outline three properties to be a good penalty function for variable selection: (1) **Unbiasedness**: avoiding introducing modeling bias, (2) **Sparsity**: reducing model complexity by eliminating unimportant variables, and (3) **Continuity**: avoiding instability in model prediction. They also give sufficient conditions for penalty functions to satisfy these conditions:

- Unbiasedness: If and only if $p'_\lambda(|\theta|) = 0$ for large $|\theta|$.
- Sparsity: If $\min_\theta\{|\theta| + p'_\lambda(|\theta|)\} > 0$.
- Continuity: If and only if $\arg\min_\theta\{|\theta| + p'_\lambda(|\theta|)\} = 0$.

They also proposed a new penalty function, called Smoothly Clipped Absolute Deviation (SCAD) penalty, which satisfies all three desired properties, defined by

$$p'_\lambda(|\theta|) = I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(|\theta| - \lambda) \text{ for some } a > 2 \text{ and } \theta > 0. \qquad (1.26)$$

Usually, a is set to be 3.7, and λ is chosen by a data driven method. Figure 1.3 gives the shape of the SCAD penalty and the relation between ordinary least squares estimates and SCAD estimates. It is clear that the estimate is continuous and shrinks small estimates to zero while keeping large estimates unbiased.

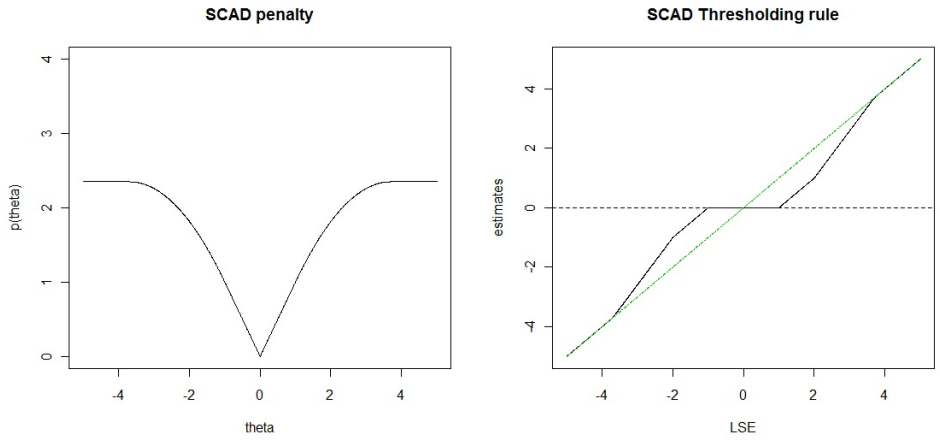

Figure 1.3. SCAD penalty, $a = 3.7, \lambda = 1$

1.2.3 Penalized Likelihood Approach for Generalized Linear Model

In the last section, we discussed variable selection through penalized least squares. This approach can be extended to more general settings, like binary response or count responses, by replacing the least squares part by a log-likelihood function. This is more general since penalized least squares is the log-likelihhod for normal likelihood. In this section, we review the settings of generalized linear models first, then present the theoretical results and literature for algorithms for variable selection through penalized likelihood functions. In the end, we briefly mention the existing methods of tuning parameter selection.

1.2.3.1 Generalized Linear Model and Penalized Likelihood functions

There are three components for linear models: (a) a normal random components $y|\mathbf{x}_i$ with mean μ_i, (b) a linear predictor $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$ as the systematic component and (c) the link relation $\mu_i = \eta_i$. Generalized linear models allow extensions of two

components: the random components can be non-normal but in an exponential family, and the link relation can be a function $\mu_i = g(\eta_i)$ other than the identity function. The likelihood function is usually the product of density functions, denoted by $L(\mathbf{x}^T\boldsymbol{\beta}, \mathbf{y})$, with log-likelihood function $\ell(\mathbf{x}^T\boldsymbol{\beta}, \mathbf{y})$. Corresponding to ordinary least squares estimators in generalized likelihood settings are Maximum Likelihood Estimators (MLE).

In order to guarantee the existence and asymptotic normality of the MLE, Fan and Li (2001) proposed regularity conditions for the likelihood function. The three regularity conditions put requirements on the smoothness of the likelihood function, prepare conditions to use the Dominated Convergence Theorem, and guarantee that one can use Taylor expansion for the log-likelihood in a neighborhood of the true parameter.

If the likelihood function satisfies all the regularity conditions, then the penalized likelihood function is defined as

$$Q(\boldsymbol{\beta}) = \ell(\mathbf{x}^T\boldsymbol{\beta}, \mathbf{y}) - n\sum_{j=1}^{p} p_\lambda(|\beta_j|). \tag{1.27}$$

The penalty function $p_\lambda(|\beta_j|)$ can take any of the forms presented in the previous section. Under certain conditions and with properly chosen λ, maximizing this function will result in a sparse estimate, hence this approach can perform variable selection and estimation simultaneously.

1.2.3.2 Theoretical Result of Variable Selection Through Penalized Likelihood

As discussed in the last section, a good estimate should be sparse and have small prediction error. In other words, if the true model is sparse, the estimates should be zero or extremely small for parameters that are truly zero, and close to the true parameters for those that are nonzero for the true model. This is called the oracle property. Fan and Li (2001) showed that in general likelihood settings, under certain conditions, the maximizer of equation (1.27) exists, and this maximizer possesses the oracle property.

Assume in a general linear model that observations $\mathbf{v_i} = (\mathbf{x_i}, y_i)$ are i.i.d and

the model parameters are identifiable. Denote the density function by $f(\mathbf{v}, \boldsymbol{\beta})$. Fan and Li (2001) gave the following regularity conditions for the density function:

(A) The observations $\mathbf{v}_i$ are independent and identically distributed with probability density $f(\mathbf{v}, \boldsymbol{\beta})$ with respect to some measure μ. $f(\mathbf{v}, \boldsymbol{\beta})$ does not depend on $\boldsymbol{\beta}$ and $\boldsymbol{\beta}$ is uniquely identifiable if $f(\mathbf{v}, \boldsymbol{\beta})$ is known. Furthermore, the first and second logarithmic derivatives of f satisfy the equations

$$E_\beta \left[\frac{\partial \log f(\mathbf{V}, \boldsymbol{\beta})}{\partial \beta_j} \right] = 0 \text{ for } j = 1, ..., p \tag{1.28}$$

and

$$\begin{aligned} I_{jk}(\boldsymbol{\beta}) &= E_\beta \left[\frac{\partial}{\partial \beta_j} \log f(\mathbf{V}, \boldsymbol{\beta}) \frac{\partial}{\partial \beta_k} \log f(\mathbf{V}, \boldsymbol{\beta}) \right] \\ &= E_\beta \left[-\frac{\partial^2}{\partial \beta_j \partial \beta_k} \log f(\mathbf{V}, \boldsymbol{\beta}) \right]. \end{aligned} \tag{1.29}$$

(B) The Fisher information matrix

$$I(\boldsymbol{\beta}) = E \left\{ \left[\frac{\partial}{\partial \boldsymbol{\beta}} \log f(\mathbf{V}, \boldsymbol{\beta}) \right] \left[\frac{\partial}{\partial \boldsymbol{\beta}} \log f(\mathbf{V}, \boldsymbol{\beta}) \right]^T \right\} \tag{1.30}$$

is finite and positive definite at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

(C) There exists an open subset ω of Ω that contains the true parameter point $\boldsymbol{\beta}_0$ such that for almost all $\mathbf{v}$ the density $f(\mathbf{v}, \boldsymbol{\beta})$ admits all third derivatives $(\frac{\partial^3 f(\mathbf{v}, \boldsymbol{\beta})}{\partial \beta_j \partial \beta_k \partial \beta_l})$ for all $\boldsymbol{\beta} \in \omega$. Furthermore, there exist functions M_{jkl} such that

$$\left| \frac{\partial^3}{\partial \beta_j \partial \beta_k \partial \beta_l} \log f(\mathbf{v}, \boldsymbol{\beta}) \right| \leq M_{jkl}(\mathbf{v}) \text{ for all } \boldsymbol{\beta} \in \omega \tag{1.31}$$

where $m_{jkl} = E_{\boldsymbol{\beta}_0}[M_{jkl}(\mathbf{v})] < \infty$ for j, k, l.

If the density satisfies all these regularity condtions and also the true model is sparse, i.e.,

$$\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{20}), \boldsymbol{\beta}_{10} \in \mathbb{R}^s, \boldsymbol{\beta}_{20} \in \mathbb{R}^r, r + s = p, \text{and } \boldsymbol{\beta}_{20} = \mathbf{0}, \tag{1.32}$$

then if we let $a_n = \max_{1\le j\le p}\{p'_{\lambda_n}(|\beta_{j0}|) : \beta_{j0} \neq 0\}$, and $b_n = \max_{1\le j\le p}\{p''_{\lambda_n}(|\beta_{j0}|) : \beta_{j0} \neq 0\}$, Fan and Li (2001) showed existence of the penalized likelihood estimator through the following theorem:

Theorem 1 *Under the regularity conditions, if $b_n \to 0$, then there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}\| = O_p(n^{-1/2} + a_n)$.*

This theorem gives existence and the convergence rate for penalized likelihood estimators. It states that there exists a local maximum $\hat{\boldsymbol{\beta}}$ for $Q(\boldsymbol{\beta})$ such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}\| = O_p(n^{-1/2} + a_n)$. If $a_n \to 0$ faster than $\frac{1}{\sqrt{n}}$, then $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent. For a specific penalty function, by choosing λ_n properly, one may be able to get a $\sqrt{n}$-consistent estimator. For example, for the hard-threshold or SCAD penalty, put $\lambda_n \to 0$, and we will have a $\sqrt{n}$-consistent estimator.

Fan and Li (2001) then showed that the maximized penalized likelihood estimator as derived in Theorem 1 is indeed zero for parameters which are zeros in the true model. This idea is summarized by the following lemma:

Lemma 1 *Assume that*

$$\liminf_{n\to+\infty} \liminf_{\theta\to 0^+} \lambda_n^{-1} p_{\lambda_n}(\theta) > 0.$$

Then if $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to +\infty$ as $n \to \infty$, with probability tending to 1, for any given $\boldsymbol{\beta_1}$ satisfying $\|\boldsymbol{\beta_1} - \boldsymbol{\beta_{10}}\| = O_p(n^{-1/2})$ and any constant C,

$$Q\{(\boldsymbol{\beta_1}, \mathbf{0})\} = \max_{\|\boldsymbol{\beta_2}\| \le Cn^{-1/2}} Q\{(\boldsymbol{\beta_1}, \boldsymbol{\beta_2})\}.$$

The result of Lemma 1 is actually part of Theorem 2, which is the oracle property:

Theorem 2 *(Oracle Property) Assume that the penalty function satisfies the conditions in Lemma 1. Then if $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to +\infty$ as $n \to +\infty$, with probability tending to 1, the $\sqrt{n}$-consistent local maximizer $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_{\mathbf{1}}, \hat{\boldsymbol{\beta}}_{\mathbf{2}})^T$ in Theorem 1 must satisfy:*

(a) Sparisty: $\hat{\boldsymbol{\beta}}_{\mathbf{2}} = \mathbf{0}$;

(b) Asymptotic normality:

$$\sqrt{n}(I_{10}+\Sigma_n)\{\hat{\boldsymbol{\beta}}_{\mathbf{1}}-\boldsymbol{\beta}_{\mathbf{10}}+(I_{10}+\Sigma_n)^{-1}\boldsymbol{b}_{\boldsymbol{n}}\}\xrightarrow{\mathcal{L}}N(0,I_{10}),$$

where $I_{10}=I_1(\boldsymbol{\beta}_{\mathbf{10}})$*, the Fisher information knowing* $\boldsymbol{\beta}_{\mathbf{2}}=\mathbf{0}$*, and*

$$\begin{aligned}\Sigma_n &= diag\{p''_{\lambda_n}(|\beta_{10}|),p''_{\lambda_n}(|\beta_{20}|),...,p''_{\lambda_n}(|\beta_{s0}|)\},\\ \mathbf{b} &= (p'_{\lambda_n}(|\beta_{10}|)sgn(\beta_{10}),p'_{\lambda_n}(|\beta_{20}|)sgn(\beta_{20}),...,p'_{\lambda_n}(|\beta_{s0}|)sgn(\beta_{s0}))^T.\end{aligned}$$

Theorem 2 gives the asymptotic bias and variance of the nonzero part of the estimators. If $\Sigma_n\to 0$ and $\mathbf{b}\to 0$, then it follows that

$$\sqrt{n}\{\hat{\boldsymbol{\beta}}_{\mathbf{1}}-\boldsymbol{\beta}_{\mathbf{10}}\}\xrightarrow{\mathcal{L}}N(0,I_{10}),\text{ and }\hat{\boldsymbol{\beta}}_{\mathbf{2}}=0,$$

which has the same asymptotic distribution as the MLE for the nonzero part $\boldsymbol{\beta}_{\mathbf{1}}$, and zero variance for the zero part $\boldsymbol{\beta}_{\mathbf{2}}$. So in this case, the penalized likelihood estimator is more efficient than the regular MLE. In particular, for hard threshold and SCAD penalties, this is satisfied when $\lambda_n\to 0$. But for L_1, if $a_n=\lambda_n\to 0$ as $n\to\infty$, $\sqrt{n}$-consistency requires $\lambda_n=O_p(\sqrt{n})$ while the oracle property requires $\sqrt{n}\lambda_n\to\infty$. Hence, the oracle property does not hold for LASSO. Another merit of this asymptotic normality result is that it provides a standard error formula for the estimated parameters. By a sandwich formula, the covariance of $\hat{\beta}$ can be estimated by

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}_1)=\{\nabla^2\ell(\hat{\boldsymbol{\beta}}_{\mathbf{1}})+n\Sigma_\lambda(\hat{\boldsymbol{\beta}}_{\mathbf{1}})\}^{-1}\widehat{\text{cov}}\{\nabla\ell(\hat{\boldsymbol{\beta}}_{\mathbf{1}})\}\times\{\nabla^2\ell(\hat{\boldsymbol{\beta}}_{\mathbf{1}})+n\Sigma_\lambda(\hat{\boldsymbol{\beta}}_{\mathbf{1}})\}^{-1}.\quad(1.33)$$

1.2.3.3 Algorithms

The sum of squared residuals is a quadratic function, hence convex and twice differentiable. And in exponential families, the negative logarithm of the likelihood function is also convex and twice differentiable. In this case, if the penalty function is convex with proper smoothness everywhere except for the origin, any method suitable for minimizing a convex smooth function could be employed to find the penalized likelihood estimator. However, neither the SCAD penalty function nor the L_p penalty for $p<1$ is convex, which may make the whole penalized likelihood

function non-concave. Also, some penalty functions are not sufficiently smooth. For instance, the L_0 penalty is not differentiable at the value of λ. To solve these problems, Fan and Li (2001) studied variable selection via the nonconcave SCAD penalty using a local quadratic approximation (LQA) of the penalty term. For each j, if the true coefficient $\theta^{(0)}$ is close to zero, then $\hat{\theta}$ is set to zero directly. Otherwise, the derivative of the penalty function can be approximated by

$$p'_\lambda(|\theta|) = p'_\lambda(|\theta|)\text{sgn}(\theta) \approx \{p'_\lambda(|\theta|)/|\theta^{(0)}|\}\theta. \tag{1.34}$$

Therefore, the penalty function is approximated locally in the following way,

$$p_\lambda(|\theta|) \approx p_\lambda(|\theta^{(0)}|) + \frac{1}{2}\{p'_\lambda(|\theta^{(0)}|)/|\theta^{(0)}|\}(\theta^2 - (\theta^{(0)})^2), \text{ for } \theta \approx \theta^{(0)}. \tag{1.35}$$

But this approach shares the same drawback with the backward stepwise variable selection: when a variable is deleted at any iteration, it will be excluded from the final model. Hunter and Li (2005) showed that LQA is a type of MM algorithm and introduced a method for remedying the flaw of LQA; however, this approach required introducing a slightly perturbed penalty function, thus slightly weakening the theoretical properties of the resulting estimators. Zou and Li (2008) introduced a different class of MM algorithm for variable selection called local linear approximation (LLA) of the penalty function, which overcomes both drawbacks. In LLA, the penalty function is approximated by

$$p_\lambda(|\theta|) \approx p_\lambda(|\theta^{(0)}|) + p'_\lambda(|\theta^{(0)}|)(|\theta| - |\theta^{(0)}|), \text{ for } \theta \approx \theta^{(0)}. \tag{1.36}$$

In this way, the penalty function is replaced by a convex function, and the penalized likelihood is concave (though the concavity is strict only if the log-likelihood is strictly concave). From the MM algorithm point of view, LLA actually finds a so-called minorizing function for the penalty function, and we can maximize a concave function instead of the original one. So a method of maximizing concave functions can apply now. The following Figure 1.4 (Zou and Li, 2008) demonstrates the LLA and LQA for $L_{0.5}$ and SCAD penalty. It shows that the LLA is the "best" convex approximation of the penalty functions.

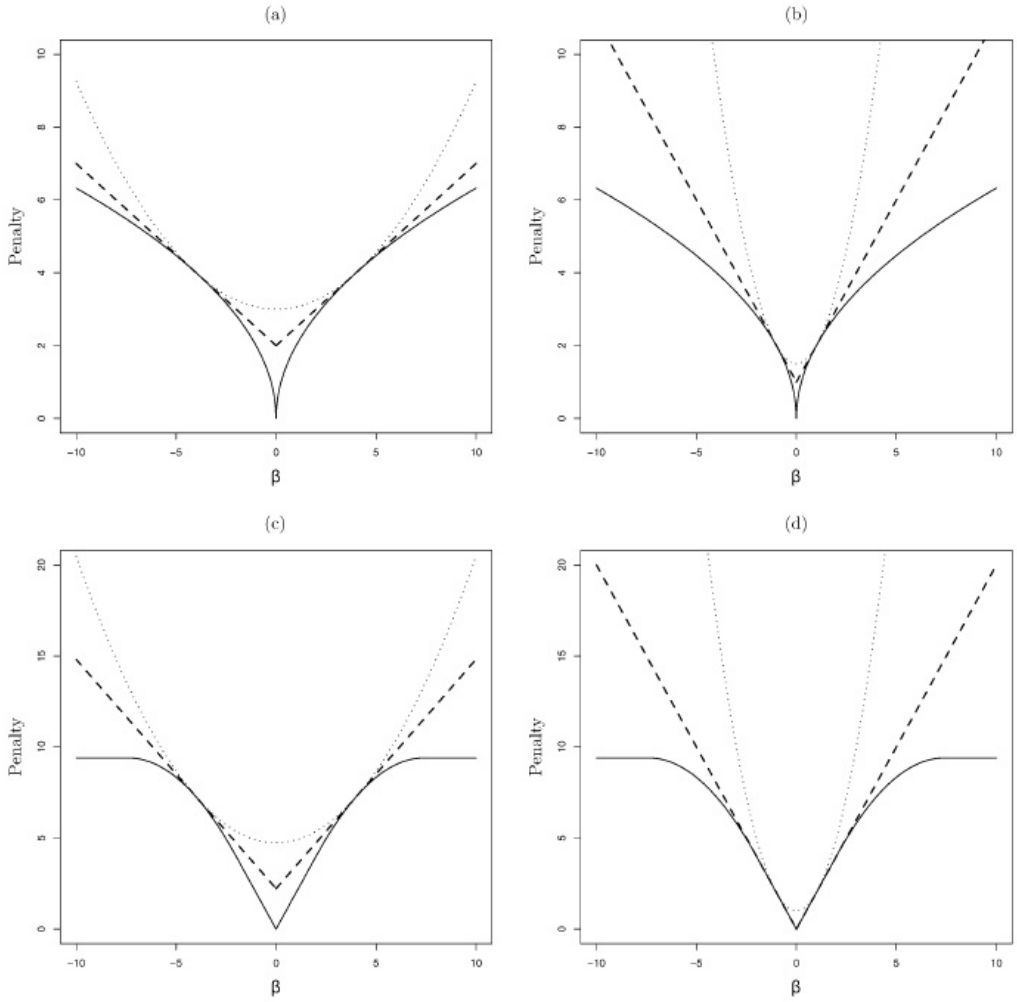

Figure 1.4. Local Quadratic Approximation (thick broken lines) and Local Linear Approximation (thin dotted lines) for different penalty functions, (a) and (b) are for $L_{0.5}$ penalty with $\lambda = 2$ and (c)(d) are for SCAD with $\lambda = 2$. (Zou and Li, 2008)

1.2.3.4 Tuning Parameter Selection Criterion

Both penalized least squares and penalized likelihood require setting the value of λ before maximizing/minimizing the objective function. According to the theoretical yield the previous section, a properly chosen λ is necessary to result in oracle estimates. Usually, people use data-driven methods to choose this tuning parameter; we will review some of the criteria in the literature. In practice, we first define a set of grid points that cover a certain interval (which is decided by the type of penalty function), then for each λ in the grid, we calculate the score of a certain

criterion. The best λ is taken to be the one which optimizes the criterion.

Fan and Li (2001) discussed tuning parameter selection procedures through five-fold cross validation and generalized cross validation. In five-fold cross validation, the whole data set is divided into five parts, $T^1, ...T^5$. One at a time, we let one subset be the test set and the remaining four be training sets. The estimates based only on the training sets and are denoted by $\hat{\boldsymbol{\beta}}^{(\nu)}(\lambda), \nu = 1, 2, ..., 5$. Then the cross-validation score for this λ is

$$CV(\lambda) = \sum_{\nu=1}^{5} \sum_{(y_k, \mathbf{x_k}) \in T^\nu} \{y_k - \mathbf{x_k^T} \hat{\boldsymbol{\beta}}^{(\nu)}(\lambda)\}^2, \tag{1.37}$$

where estimates of β depend on λ, and the λ which minimizes the cross-validation score is the optimum one. The second criterion is generalized cross validation, which is defined by

$$GCV(\lambda) = \frac{1}{n} \frac{\|\mathbf{y} - \mathbf{X}\beta(\lambda)\|^2}{\{1 - \mathrm{df}(\lambda)/n\}^2}, \tag{1.38}$$

where $df(\lambda)$ is the degrees of freedom of the model, or the trace of the projection matrix $\mathbf{X}\{\mathbf{X^T X} + n\Sigma_\lambda(\hat{\boldsymbol{\beta}})\}^{-1}\mathbf{X^T}$. In practice, df is taken to be the number of variables in the model, since asymptotically they are equal (Zhang et al., 2010). Similarly to cross validation, the minimizer is the best λ.

Wang, Li, and Tsai (2007) found that GCV is similar to the classic variable selection criterion AIC. Furthermore, the model selected by GCV in linear regression settings tends to include unnecessary variables in the model. This motivated them to propose a new tuning parameter selector for penalized least square estimates using the SCAD penalty, called BIC:

$$BIC_\lambda = \log \hat{\sigma}_\lambda^2 + \mathrm{df}(\lambda) \log(n)/n. \tag{1.39}$$

They showed that if the true model is among the candidate models (is a linear regression model), the model selected by BIC_λ will identify the true model consistently as the sample size increases, while AIC and GCV will overfit the model.

To generalize this to settings other than penalized least square with SCAD, Zhang, Li, and Tsai (2010) proposed a new tuning parameter selector called generalized information criterion. It is a class of criteria, including GCV not only and

BIC as special cases, but also its modified version can be used as classic variable selection criteria. GIC is defined as

$$GIC_{\kappa_n}(\lambda) = \frac{1}{n}\{G(\mathbf{y}, \hat{\boldsymbol{\beta}}_\lambda) + \kappa_n \mathrm{df}_\lambda\}, \tag{1.40}$$

where $G(\mathbf{y}, \hat{\boldsymbol{\beta}}_\lambda)$ measures the goodness of fit of the model estimated by fixing λ. For example, $G(\mathbf{y}, \hat{\boldsymbol{\beta}}_\lambda)$ could be a log-likelihood. The number κ_n depends on n and df_λ is the degrees of freedom of the model. When $\kappa_n \to 2$, (1.40) is called an AIC-type selector and when $\kappa_n \to \infty$ and $\kappa_n/\sqrt{n} \to 0$, it becomes a BIC-type selector. When the true model is among the candidate models, a BIC-type selector will choose the true model consistently while an AIC-type selector will overfit the model. However, when the true model can only be approximated by candidate models, an AIC-type selector is asymptotically loss-efficient while a BIC-type selector is not. This agrees with the result by Wang et al. (2007), and again verifies that GCV will overfit the model. So which tuning parameter criterion to use is really depends on whether the true model is in the candidate model class or can be approximated by a candidate model.

1.3 Survival Model and Variable Selection

Survival models have been used widely in many application areas, where the main interest is to investigate the distribution of the time T to some event. This event is usually related to death or cessation of a certain disease. Typically, the existence of censoring makes it impossible to observe all events of interest for every individual. We only know whether the event occurs prior to, after or within a certain period of time, depending on which censoring scheme the data uses. So modeling survival data requires novel models and techniques. In this section, the basic setting of survival analysis will be introduced, followed by introduction and comparison of two survival models and the corresponding variable selection literature.

1.3.1 Survival Models

1.3.1.1 Data structure and Basic settings

Denoting by T the time to an event of interest, the distribution of T can be characterized by any of the three functions, the cumulative distribution function, the hazard function, or the survival function. The hazard function $h(t)$ is the instantaneous risk of event at time t given that the event does not happen before time t; the survivor function $S(t)$ defines the probability of surviving longer than time t. The above three functions can be converted from one to another, so modeling any one of them is appropriate in survival analysis.

Observations in survival analysis may be censored or truncated. Possible censoring schemes are: right censoring, in which event time is known to be after some censoring time; left censoring, where what is known is only that the event occurs prior to the study; and interval censoring, when events occur within some time intervals (Klein and Moeschberger, 2003). The scheme considered in this dissertation is right-censored data, where the observed data for each individual are $Y = \min(T, C)$, the censoring indicator $\delta = I_{\{T<C\}}$, and the covariate vector $\mathbf{X}$, where C indicates the censoring time and we usually assume that given the covariate, T and C are independent, an assumption called non-informative censoring. Some of the many types of survival models used in the literature will be discussed in the following section.

1.3.1.2 Proportional Hazard and Proportional Odds Models

In survival analysis, individuals' survival behaviors might be affected by their own characteristics. Motivated by this, Cox (1972) introduced the proportional hazard model, which is a type of multiplicative hazard rate model. It models the hazard function as a function of explanatory variables and corresponding coefficients multiplied by some unknown baseline hazard function. Denote by h_0 the baseline hazard function, and suppose that we have n i.i.d observations of $(Y_i, \delta_i, \mathbf{x}_{(i)}), \mathbf{x}_{(i)} \in R^p, i = 1, 2, 3..., n$. Then Cox's proportional odds model is

$$h(t) = h_0(t) \exp\left(\sum_{j=1}^{p} \beta_j x_j\right). \tag{1.41}$$

An important feature of this model is that for any two given sets of covariate values $\mathbf{x}$ and $\mathbf{x}^*$, the associated hazards are proportional for all time t, i.e.,

$$\frac{h(t|\mathbf{x})}{h(t|\mathbf{x}^*)} = \frac{h_0(t)e^{\sum_{i=1}^{p}\beta_i x_i}}{h_0(t)e^{\sum_{i=1}^{p}\beta_i x_i^*}} = \exp\left[\sum_{i=1}^{p}\beta_i(x_i - x_i^*)\right]. \tag{1.42}$$

Unlike for other models, construction of likelihood functions in survival analysis should consider censored data. Assuming noninformative censoring and right censored data, the likelihood function is of the form

$$\prod_{i\in U} f(Y_i)\prod_{i\in C}\{1 - F(Y_i)\}, \tag{1.43}$$

where U is the set of uncensored observations and C is the set of censored observations.

Model (1.41) is a semi-parametric model, in which $\boldsymbol{\beta}$ is the parametric part and $h_0(t)$ is the nonparameteric part. Usually, the goal of investigation is to make an inference about $\boldsymbol{\beta}$ in the global sense, or, more often than not, to make an inference about a subset of $\boldsymbol{\beta}$ (Klein and Moeschberger, 2003). So, $h_0(t)$ is generally considered a nuisance parameter. Hence, the model estimation or inference will usually be based on so-called partial likelihood, which is only a function of $\boldsymbol{\beta}$. The partial likelihood is the product of the conditional probabilities of individual death at t_i given there is only one death at each t_i. If $t_1 < t_2 < ... < t_J$ are the ordered observed event times, and $R(t_j)$ is the "at risk" set (set of individuals who have not experienced an event) at time t_j, then the partial likelihood has the following expression

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{J}\frac{e^{\boldsymbol{\beta}^T\mathbf{x}_j}}{\sum_{i\in R(t_j)} e^{\boldsymbol{\beta}^T\mathbf{x}_j}} \tag{1.44}$$

It can be shown that the partial likelihood is the same as the profile likelihood for Cox's regression (Klein and Moeschberger, 2003). Therefore, maximizing the partial likelihood will result in the same estimator of $\boldsymbol{\beta}$ as maximizing the full likelihood function.

The Cox proportional hazard model has a closed form of the partial likelihood, hence it is computationally easy to find estimates of $\boldsymbol{\beta}$. However, it may not be

suitable for modeling the case when the ratio of hazard function is changing over time. Bennett (1983) proposed a proportional odds model in which the odds of two groups are proportional. The usual odds ratio of two events is defined as

$$r = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}. \tag{1.45}$$

A generalization of the definition of odds to distribution functions gives the following odds ratio for two groups of subjects:

$$r = \frac{F_1(t)/\{1-F_1(t)\}}{F_2(t)/\{1-F_2(t)\}}. \tag{1.46}$$

For more than two groups of individuals, each with covariate x_i, Bennett (1983) gave the proportional odds model as

$$\frac{F(t;r_i)}{1-F(t;r_i)} = \frac{F_0(t)}{1-F_0(t)} r_i, \tag{1.47}$$

where $\dfrac{F_0(t)}{1-F_0(t)} \doteq H(t)$ is called the baseline odds and $r_i = \boldsymbol{\beta}^T \mathbf{x}_j$. This is also a semi-parametric model, with $\boldsymbol{\beta}$ the parametric part and $F_0(t)$ the nonparametric part. Bennett (1983) also suggested in the two-group case that the ratio of the hazard functions converges monotonically from r to 1, while the ratio of survival functions diverges from 1 to r, as time tends to infinity. Hence, the model is appropriate to show an effective cure.

Murphy, Rossini, and Van der Vaart (1997) reparameterized the model and studied the maximum likelihood estimator. But unlike for the proportional hazard model, there exists no partial likelihood function for the proportional odds model. For the full likelihood function,

$$\prod_{i=1}^{n} \left(\frac{e^{-\boldsymbol{\beta}^T \mathbf{x}_j}}{H(Y_i) + e^{-\boldsymbol{\beta}^T \mathbf{x}_j}} \right) \left(\frac{\Delta H(Y_i)}{H(Y_i) + e^{-\boldsymbol{\beta}^T \mathbf{x}_j}} \right)^{\delta_i}, \tag{1.48}$$

the existence of the MLE has been verified and its consistency and asymptotic normality are known. Another important result is that the MLE of H is a step function with jumps $\Delta H(Y_i)$ only at each uncensored time. Hence the maximum

likelihood problem becomes a parametric model with H replaced by the jumps $h_i = \Delta H(Y_i)$ at the uncensored event times. So the number of parameters in the model is the number of covariates plus the number of the different uncensored event times. Although it is a parametric model, the number of parameters is large.

Hunter and Lange (2002) reparametrized the likelihood function as in (1.48), in order that the log-likelihood ibe concave. Suppose m different event times are observed, which are sorted as $U_1 < U_2 < .... < U_m$. Let $\gamma_j = \ln(h_i), 1 < j < m$, and $\omega_i = \max\{j : U_j \leq Y_j\}$. Now $\boldsymbol{\theta} = (\boldsymbol{\beta}^t, \boldsymbol{\gamma}^t) \in \mathbb{R}^{p+m}$, and the log-likelihood function becomes

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} -\mathbf{z}_i^t\boldsymbol{\beta} - \ln D_i(\boldsymbol{\theta}) + \delta_i\{\gamma_{\omega_i} - \ln[D_i(\boldsymbol{\theta}) - e^{\gamma_{\omega_i}}]\}, \tag{1.49}$$

where $D_i(\boldsymbol{\theta}) = e^{-\mathbf{z}_i^t\boldsymbol{\beta}} + \sum_{j=1}^{\omega_i} e^{\gamma_j}$. They also proposed a minorization-maximization (MM) algorithm to estimate $\boldsymbol{\beta}$, which guarantees convergence to the maximum likelihood estimator whenever it exists.

1.3.2 Variable Selection for Survival Models

The two models described in Section 1.3.1.2 depend on covariates through a linear predictor. In the initial stage of the study, one would include many covariates to reduce bias, then select the best subset among the large number of covariates. Both models are semiparametric, but the model selection methodology discussed in Section 2 is only applied for a parametric model or likelihood. Possible adjustments here could be using partial likelihood or profile likelihood, as shown in the following two papers.

The problem of variable selection for Cox's model is considered by Fan and Li (2002). Since there exists a partial likelihood for the Cox model, which is a parametric function only depending on $\boldsymbol{\beta}$, they used penalized partial likelihood approaches to select and estimate variables simultaneously. Under several conditions which guarantee the asymptotic normality of the maximum partial likelihood estimates, they first proved existence of penalized partial likelihood estimates which converge at rate $O_p(n^{-1/2}+a_n)$, where a_n is as defined in Fan and Li (2001). Hence

by properly choosing a penalty function, there exists a $\sqrt{n}-$consistent estimator. Based upon this result, they showed further that these estimates possess oracle properties, exactly as in Theorem 2 in Section 1.2.3.2 for parametric models. The only difference is the modified proofs use the information corresponding to the partial likelihood. These results imply if a penalty function and tuning parameter are chosen properly, the penalized partial likelihood estimates will perform as well as we know in advance $\boldsymbol{\beta_2} = 0$. From this asymptotic normality result, one can derive a formula for estimating the covariance matrix directly via a sandwich formula.

To compute the penalized partial likelihood estimates, Fan and Li (2002) suggest using a Newton-Raphson algorithm. Starting from $k = 1$, suppose we already obtained the estimate for kth step, $\boldsymbol{\beta}_k$. Then the $(k+1)$th step estimate can be obtained through the formula

$$\boldsymbol{\beta_{k+1}} = \boldsymbol{\beta_k} - \{\nabla^2 \ell(\boldsymbol{\beta_k}) - n\Sigma_\lambda(\boldsymbol{\beta_k})\}^{-1}\{\nabla \ell(\boldsymbol{\beta_k}) - n\boldsymbol{\beta_k}\Sigma_\lambda(\boldsymbol{\beta_k})\}, \tag{1.50}$$

where $\ell(\boldsymbol{\beta_k})$ is the logarithm of partial likelihood and Σ_λ as defined in the first section in this chapter. Since partial likelihood is the same as profile likelihood for Cox's model, all above results are actually properties of penalized likelihood estimates.

As mentioned in previous section, no partial likelihood exists for proportional odds model, and the profile likelihood for β does not have a closed form. Instead, Lu and Zhang (2007) considered marginal likelihood, which is the integral of the full likelihood function with respect to the nonparametric baseline odds function. If we denote $V_{(j)} = H(T_{(j)}), k = 1, ...J$, where $H(\cdot)$ is the baseline odds function and J is the number of distinct observed event times, the marginal likelihood is

$$L_{n,M}(\beta) = \int \cdots \int_{V_{(1)}<\cdots<V_{(J)}} \prod_{i=1}^{n} \{h(V_{(k_i)}) + \mathbf{x}_i^T \boldsymbol{\beta}\}^{\delta_i} e^{-\Lambda(\mathbf{x}_{(k_i)} + \mathbf{z}_i \boldsymbol{\beta})} \prod_{j=1}^{J} dV_{(j)}. \tag{1.51}$$

Since this integration does not have an analytical form, Lu and Zhang (2007) use importance sampling to approximate it. Then they conduct variable selection by maximizing the penalized marginal likelihood. Motivated by the idea of shrinkage, they use LASSO penalties and to reduce bias. To reduce the bias brought by LASSO, they further used Adaptive LASSO, which is a scaled version

of LASSO and will be discussed more in Chapter 3. The estimation is performed by an iterative computation algorithm, iterating between maximized marginal likelihood and penalized least squares until convergence. The simulation results show this approach performs well in identifying nonzero and zero coefficients separately. However, use of this marginal likelihood function may lose information in the full likelihood function and stochastic approximation of the marginal likelihood function may also influence the selection of variables. In addition, there is no theoretical justification for properties of the resulting estimates.

Recently, Liu and Zeng (2013) studied the problem of variable selection for linear transformation models, which include the proportional odds model as a special case. In a linear transformation model setting, a proportional odds model with time-varying covariates $Z(\cdot)$ can be written as

$$\Lambda(t|Z(\cdot)) = G\left[\int_0^t \exp\{\boldsymbol{\beta}^T \boldsymbol{Z}(\boldsymbol{s})\} d\Lambda(s)\right], \tag{1.52}$$

where $\Lambda(\cdot)$ is the cumulate hazard function and $G(x) = \log(1+x)$. Zeng and Lin (2007) treat ζ as a nuisance variable with exponential(1) distribution and show that the model (1.52) is equivalent to

$$\Lambda(t|Z(\cdot), \zeta) = \zeta \int_0^t \exp\{\boldsymbol{\beta}^T \boldsymbol{Z}(\boldsymbol{s})\} d\Lambda(s). \tag{1.53}$$

So given $Z(\cdot)$ and ζ, (1.52) can be treated as a proportional hazard model with ζ missing. Thus one can work with the complete data including ζ and use an EM algorithm to obtain the MLE. In terms of variable selection, Liu and Zeng (2013) first profiled out ζ from the optimized likelihood function, then add penalty function to find the sparse estimates. The resulting estimates are sparse, consistent and has the oracle properties. Their approach avoids the ambiguity in concavity of the full likelihood function, instead, maximize a surrogate function of the original penalized likelihood function.

Different from the two approaches above, we propose a variable selection method based on maximizing the penalized profile likelihood function. The profile likelihood function is more general than the partial likelihood function, but it is still a parametric function of the regression coefficients only. The major challenge in

deriving some of the desired properties is that there is no closed form of the profile likelihood function, hence one never knows whether it is differentiable or not. Therefore, the proofs using the Taylor expansion cannot be used here. Instead, we use the expansion derived for any profile likelihood function by Murphy and Van der Vaart (2000). We also propose some combined algorithms for maximizing the penalized profile likelihood function, which are under the framework of MM algorithms. Hunter and Lange (2002) introduce an algorithm for calculating maximum likelihood estimators for the proportional odds model by iteratively updating and maximizing a surrogate function of the likelihood function. We construct a new surrogate function for the penalized likelihood function by plugging in this surrogate function, thus creating a parametric problem. Some existing algorithms are then applied for maximizing this new function: ICM by Zhang and Li (2009) , MIST by Schifano et al. (2010) and Coordinate Descent by Wu and Lange (2008). We also compare our algorithms and approaches with those of Lu and Zhang (2007) and Liu and Zeng (2013) in the same settings in their papers. The newly proposed algorithms perform well and computationally fast. Detailed results can be found in Chapter 2.

1.4 Social Network Models

Network data have attracted a lot of attention recently because of their wide applications in a variety of disciplines, including biology, engineering, and social science. In general and informally defined, networks may be thought of as "a collection of interactive things" (Kolaczyk, 2009), and they can be represented by a collection of nodes and directed or indirected edges between nodes. This graphical representation is used in a formal manner for all kinds of networks. In recent years, due to the development of data collection techniques, large and time-varying networks have been studied more often. One area of research is using survival model in dynamic network analysis. But unlike previous sections, the data structure in a network setting is often dependent. Thus, regular survival models cannot be applied directly. In this section, literature on network modeling-related survival models will be reviewed. And the following Chapters 3 and 4 will be based on some of these network models.

1.4.1 Relational Events Model

The essence of the modeling approach introduced by Butts (2008) is the relational event, which is defined as a directed action from a social actor (sender) to another one (receiver). Each action can be characterized by

$$a = (i, j, k, t), i \in \mathcal{S}, j \in \mathcal{R}, k \in \mathcal{C} \text{ and } t \in \mathbb{R}, \tag{1.54}$$

where $\mathcal{S}$ is set of senders, $\mathcal{R}$ is set of receivers, k represents the action type, and t is the time of the action. Butts's approach is to model $\mathbf{A}_t$, which is the set of all actions occurring before time t. And he assumes that given this past history $\mathbf{A}_t$, all actions at t will arise independently. Therefore $\mathbf{A}_t$ is a stochastic process with each event happening independently conditional on the past realization of events. This assumption is used for modeling a Poisson process as a stochastic counting process.

Suppose by time t, we have observed M events, as $a_1, a_2, ..., a_M$, separately. Further assume that they are sorted in ascending order of time. For each event a_i, conditional on its sender $s(a_i)$, receiver $r(a_i)$, action type $c(a_i)$, covariates $\mathbf{X}_{a_i}$, and past history for previous events $\mathbf{A}_{\tau(a_{i-1})}$, their conditional joint distribution is assumed to be independent, and these conditional hazard functions and survival function are denoted by $h(t|\cdot)$, and $S(t|\cdot)$ respectively. If the support set of $\mathbf{A}_t$ is defined as $\mathbb{A}(\mathbf{A}_t) \subset \mathcal{S} \times \mathcal{R} \times \mathcal{C}$, all possible combinations of sender-receiver-type given the realized history A_t, then the likelihood function of the relational event history has the following form:

$$\begin{aligned} p(\mathbf{A}_t) \quad = \quad & \left[\prod_{i=1}^{M} \left[\begin{array}{c} h(\tau(a_i)|s(a_i), r(a_i), c(a_i), \mathbf{X}_{a_i}, \mathbf{A}_{\tau(a_{i-1})}) \times \\ \prod_{a' \in \mathbb{A}(\mathbf{A}_{\tau(a_i)})} S(\tau(a_i) - \tau(a_{i-1})|s(a'), r(a'), c(a'), \mathbf{X}_{a'}, \mathbf{A}_{\tau(a_{i-1})}) \end{array} \right] \right] \\ & \times \left[\prod_{a' \in \mathbb{A}(\mathbf{A}_t)} S(t - \tau(a_M)|s(a'), r(a'), c(a'), \mathbf{X}_{a'}, \mathbf{A}_t) \right], \end{aligned} \tag{1.55}$$

where the part in the first line counts all probability associated with the events happening prior to t, the second line indicates that between any two consecutive events, no event has occurred, and the last line gives the probability that no event

has happened yet since the most recent event until time t.

If the hazard function is specified, we may obtain a parametric form of this likelihood function. Butts suggested using a piecewise constant hazard function, which depends on some unknown parameter $\boldsymbol{\theta}$ as follows:

$$\lambda_{a,\mathbf{A}_t,\theta} = \lambda(s(a), r(a), c(a), \mathbf{X}_a, \mathbf{A}_t, \theta). \tag{1.56}$$

He also gives a specification of a hazard function in an exponential form as follows:

$$\lambda(s(a), r(a), c(a), \mathbf{X}_a, \mathbf{A}_t, \theta) = \exp[\lambda_0 + \boldsymbol{\theta}^T \mathbf{u}(s(a), r(a), c(a), \mathbf{X}_a, \mathbf{A}_t)], \tag{1.57}$$

where the function $u(\cdot)$ is a vector of covariates for the model and λ_0 is some constant. The choice of covariates decides dependency among system components, which heavily rely on the properties of the network. Butts (2008) discusses several covariates. One example could be persistence covariates, which measure the tendency of past contacts to become future contacts. If $d(i, j, A_k)$ is the accumulated volume of communication from actor i to actor j by time k, this statistic is defined as

$$u_P(a, A_t, X) = d(s(a), r(a), A_t) / \sum_{j=1}^{|\mathcal{R}|} d(i, j, A_k).$$

Other examples of covariates are the preferential attachment covariates, which capture the phenomenon that previous contacts are more likely to be the target of communication. Hence, they are the fraction of receiver's communication volume out of all communication for the actor, i.e.,

$$u_{PA}(a, A_t, X) = \frac{d^+(r(a), A_t) + d^-(r(a), A_t)}{\sum_{j=1}^{|\mathcal{S}|} d^+(j, A_k) + \sum_{j=1}^{|\mathcal{S}|} d^-(j, A_k)},$$

where $d^+(j, A_k) = \sum_{j=1}^{|\mathcal{R}|} d(i, j, A_k)$ and $d^-(i, A_k) = \sum_{j=1}^{|\mathcal{S}|} d(j, i, A_k)$. In the specific setting of radio communication in disasters, Butts (2008) discusses several participation shift covariates, which are all indicator functions. Based on this model, Butts (2008) provides an application for relational event modeling on responder radio communication during the early hours of the World Trade Center Disaster. He

specifies covariates in this situation and estimates the parameter θ by maximizing the likelihood function through a Newton-Raphson algorithm.

1.4.2 Survival Models in Networks

The approach of Butts (2008) in the previous section utilizes the idea of stochastic processes to model large dynamic networks. Another approach using this idea is to apply the survival models mentioned in section 1.3 under the counting process framework, which do not assume independence among observations. However, it is not trivial to use the theoretical results from the counting process directly in network settings. In this section, several network models in the literature using the counting process approach will be reviewed, including some of the theoretical justifications. First, the general counting process framework will also be briefly reviewed.

1.4.2.1 Counting Processes in Survival Settings

In a survival model setting with multiple observations, $(Y_i, \boldsymbol{x}_i, \delta_i), i = 1, 2, \ldots, n$, the counting process for individual i is defined as

$$N_i(t) = I(T_i \leq t, \delta_i = 1), \tag{1.58}$$

which turns to one from zero at the moment the ith individual dies. The sum of all n counting processes is a new counting process,

$$N(t) = \sum_{i=1}^{n} N_i(t). \tag{1.59}$$

Some prior knowledge before time t, including information about status (died or censored) and characteristics of each individual, might be obtained. This knowledge increases with time and the cumulative prior knowledge is called the history or filtration of the counting process at time t, denoted by $\mathcal{F}_t$. Mathematically, the filtration at time t is the σ-algebra generated by $N(s), 0 \leq s \leq t$. If $Y_i(t)$ is defined as the indicator for the ith individual dying at or after time t, then $Y(t) = \sum_{i=1}^{n} Y_i(t)$

number of individuals dying at or after time t. It can be verified that

$$E[dN(t)|\mathcal{F}_{t-}] = Y(t)h(t)dt, \tag{1.60}$$

where $dN(t)$ is defined as the change of $N(t)$ in the next dt time, and $h(t)$ is the hazard function in the survival models. The process $\lambda(t) = Y(t)h(t)$ is called the intensity process. Further, if we define the cumulative intensity process as $\Lambda(t) = \int_0^t \lambda(s)ds$, the counting process $N(t)$ can be decomposed by Doob-Meyer decomposition (Klein and Moeschberger, 2003)

$$N(t) = \Lambda(t) + M(t), \tag{1.61}$$

where $M(t)$ is a martingale, which has the property that the best guess of $M(t+s)$ for all $s > 0$ when given $\mathcal{F}_t$ should be $M(t)$, i.e.,

$$E[M(t+s)|\mathcal{F}_t] = M(t). \tag{1.62}$$

$\Lambda(t)$ is also called a compensator of the counting process. The decomposition (1.61) can be understood to mean that the counting process consists two parts, the smoothly varying compensator and random noise $M(t)$ with expectation 0 (Klein and Moeschberger, 2003).

Large-sample theory is well developed for martingales. The counting process approach can be used on estimation problems in nonparametric models for survival analysis. For example, one can estimate coefficients in Cox's proportional hazard model with time-varying covarites, as well as asymptotic properties of these estimates. Furthermore, there is no assumption about independence among observations. Some generalized versions can be used for network modeling, and this will be elaborated in the next section.

1.4.2.2 Counting Process for Network Models

The idea of using survival models is to apply counting processes on nodes or on edges in a dynamic network, depending on the feature of greatest interest in the networks. Recently, several papers using the counting process approach for

continuous-time longitudinal network data have been proposed. Depending on where the counting process is defined, they have been termed as the relational approach or the egocentric approach.

The relational approach defines counting processes on pairs of nodes, or more simply, edges. There are two recent papers using this approach. The first paper (Vu et al., 2011a) introduces a continuous regression modeling framework for network event data and incorporation of time-varying regression coefficients. In a dynamic network, nodes can enter the network over time and then edges are generated between the existing nodes and the newly entered node at different event times. For each pair of nodes (i, j), let $N_{ij}(t)$ be the counting process recording the number of edges from i to j before time t. By the theory for counting processes, the multivariate counting process $\mathbf{N}(t) = (N_{ij}(t), i, j \in \{1, 2, ..., n\}, i \neq j)$ can be decomposed into the sum of a cumulative intensive process and a martingale residual,

$$\mathbf{N}(t) = \int_0^t \boldsymbol{\lambda}(s)ds + \mathbf{M}(t). \tag{1.63}$$

Vu et al. (2011a) models the intensity process $\lambda_{ij}(t)$ given the event history $\mathcal{F}_{t-}$ just before time t by two different models. One is Cox's proportional hazard model while the other is Aalen's additive model. Both new models involve time-varying covariates, and the additive model also includes time-varying coefficients. The models are as follows:

$$\lambda_{ij}(t|\mathcal{F}_{t^-}) = Y_{ij}(t)\alpha_0(t)\exp[\boldsymbol{\beta}^T\mathbf{s}_{ij}(t)], \tag{1.64}$$

$$\lambda_{ij}(t|\mathcal{F}_{t^-}) = Y_{ij}(t)[\beta_0(t) + \boldsymbol{\beta}(t)^T\mathbf{s}_{ij}(t)]. \tag{1.65}$$

Recall from a previous section that $Y_{ij}(t)$ is the at-risk indicator. The $\mathbf{s}_{ij}(t)$ vector consists of various covariates on the network. In the first model, $\alpha_0(t)$ is considered to be a nuisance parameter and we can obtain the estimates of $\boldsymbol{\beta}$ through maximizing the partial likelihood of the Cox proportional hazard model. The second model can be considered as a time-varying coefficient model; spline or kernel methods for estimation for Aalen models can be used to estimate $\beta(t)$. Both models include parts which are time dependent, and a main inference problem is to test if these parts are changing over time. By properties which are decided empiri-

cally, Vu et al. (2011a) discussed 8 covariates, including out-degree and in-degree of both the senders and the receivers, reciprocity and transitivity, shared contactees and contacters, and triangle closure. For example, the out-degree of sender i is defined to be $s_{ij}^{\text{out}}(t) = \sum_{h \in V, h \neq i} N_{ih}(t^-)$ and the in-degree of sender i would be $s_{ij}^{\text{in}}(t) = \sum_{h \in V, h \neq i} N_{hi}(t^-)$.

The other paper, by Perry and Wolfe (2013) investigates interaction networks and also models the counting process $N_{ij}(t)$, the number of directed edges from node i to j. This paper also uses the Cox's proportional hazard model (1.64) for the intensity process. Perry and Wolfe first consider estimations in a simpler situation with only one sender i_m and a single receiver j_m at each event time t_m. If $\mathcal{J}_{t_m}$ denotes all the nodes toward which node i_m could create an edge at time t_m, then the partial likelihood function can be written as

$$\log PL_t(\boldsymbol{\beta}) = \sum_{t_m < t} \left\{ \boldsymbol{\beta}^\top \boldsymbol{s}_{i_m j_m}(t) - \log \left[\sum_{j \in \mathcal{J}_{t_m}(i_m)} \exp\{\boldsymbol{\beta}^\top \boldsymbol{s}_{i_m j}(t)\} \right] \right\}. \tag{1.66}$$

Estimation for $\boldsymbol{\beta}$ can be done by maximizing this partial likelihood function. Under several regularity conditions, Perry and Wolfe (2013) prove that the maximum partial likelihood estimator is consistent and has an asymptotic normal distribution. These asymptotic results cannot be derived directly from the work of Andersen and Gill (1982), because the data structure in a network is usually different from that in a regular counting process setting. This is because the number of observations as well as the range of observations will go to infinity at the same time. In proofs for a fixed n, Perry and Wolfe (2013) rescale the observations to a finite interval and then use a discretized version of the original score function to derive the final results. Since an interaction network may have multiple events occurring at the same time, Perry and Wolfe further study the case when there are one sender and multiple receivers. The observations now are $(i_m, J_m, t_m), m = 1, 2, \ldots, n$, where J_m is the set of receivers at time t_m. The partial likelihood function becomes more

complicated:

$$\log PL_t(\boldsymbol{\beta}) = \sum_{t_m<t} \left\{ \sum_{j\in J_m} \boldsymbol{\beta}^\top \boldsymbol{x}_{i_m j}(t_m) - \log \left[\sum_{\substack{J \subseteq \mathcal{J}_{t_m}(i_m) \\ |J| = |J_m|}} \exp\left\{ \sum_{j\in J} \boldsymbol{\beta}^\top \boldsymbol{x}_{i_m j}(t_m) \right\} \right] \right\}. \tag{1.67}$$

Using this likelihood will increase the computational burden quickly as the size of the network increases. Instead of maximizing this original likelihood function, Perry and Wolfe (2013) use an approximation,

$$\log \widetilde{PL}_t(\boldsymbol{\beta}) = \sum_{t_m<t} \left\{ \sum_{j\in J_m} \boldsymbol{\beta}^\top \boldsymbol{x}_{i_m j}(t_m) - |J_m| \log \left[\sum_{j\in \mathcal{J}_{t_m}(i_m)} \exp\left\{ \boldsymbol{\beta}^\top \boldsymbol{x}_{i_m j}(t_m) \right\} \right] \right\} \tag{1.68}$$

In addition, they show the difference in the first two derivatives between the original (1.67) and the approximation one (1.68) will be bounded by how fast the receiver set growth sequence G_n grows. In particular

$$\left\| \bigtriangledown[\log PL_{t_n}(\boldsymbol{\beta})] - \bigtriangledown[\log \widetilde{PL}_{t_n}(\boldsymbol{\beta})] \right\| = O_P(G_n). \tag{1.69}$$

$$\left\| \bigtriangledown^2[\log PL_{t_n}(\boldsymbol{\beta})] - \bigtriangledown^2[\log \widetilde{PL}_{t_n}(\boldsymbol{\beta})] \right\| = O_P(G_n), \tag{1.70}$$

where the receiver set growth sequence is defined as

$$G_n = \sum_{t_m \le t_n} \frac{1\{|J_m| > 1\}}{|\mathcal{J}_{t_m}(i_m)|}.$$

Therefore, if G_n is bounded by $O_p(\sqrt{n})$, then the maximizer of the approximated partial likelihood function will also be consistent and its proper scaled version converges to a normal distribution. And since the partial likelihood function is twice differentiable, algorithms like Newton-Raphson can be used to find the maximizer.

On the other hand, the above relational modeling procedure for edges is inappropriate in some network settings, for example, a citation network. Instead, an egocentric framework can be used, which models the nodes by counting processes. Vu et al. (2011b) focuses on citation network analysis, where each paper is treated as a node and citations among papers are directed edges. Similarly to the previous

paper, the approach depends on the decomposition of a multivariate counting process $\mathbf{N}(t) = (N_i(t), i \in \{1, 2, ..., n\})$ and model the corresponding intensity process $\lambda_i(t)$ by

$$\lambda_i(t|H_{t^-}) = Y_i(t)\alpha_0(t)\exp[\beta^T \mathbf{s}_i(t)]. \tag{1.71}$$

Since the model for the intensity process is the Cox proportional hazard model, the coefficient β can be estimated by maximizing the partial likelihood function. To be realistic, the authors considered the situation of multiple receivers as in Perry and Wolfe (2013). A similar approximated partial likelihood function is maximized to obtain the estimates of $\boldsymbol{\beta}$. However, there are no theoretical properties derived for the resulting maximizer. Since our variable selection will be based on this model, in Chapter 3, we first derive some theory similar to that of Perry and Wolfe (2013) in a egocentric framework.

Vu et al. (2011b) also discussed choice of the covariates including 8 predictors for network structures and a fifty-dimensional LDA predictor vector. The 8 network structure covariates include three preferential attachment covariates, three triangle covariates, and two out-path covariates. As discussed by Butts (2008), preferential attachments measures the tendency that past contacts lead to future contacts, where the predictive contacts could be direct (first order) or indirect (second-order). If Y_{ij} records the number of edges from node i to node j at time t, the first-order and second-order preferential attachments covariates are defined respectively as

$$s_j^{PA1}(t) = \sum_{i=1}^{N} Y_{ij}(t), \tag{1.72}$$

$$s_j^{PA2}(t) = \sum_{i \neq k} Y_{ki}(t)Y_{ij}(t). \tag{1.73}$$

Also, they consider a Recency-based first-order PA

$$s_j^{Rec-PA1}(t) = \sum_{i=1}^{N} Y_{ij}(t)I(t - t_i^{\text{arr}} < T_w), \tag{1.74}$$

where T_w is a specific time window. The triangle covariates are based on the only possible triangle configuration in citation networks: paper B joins the network and

cites paper A, and then paper C joins the network and cites A and B. In this situation, A is called the "seller", B is called the "broker" and C is called the "buyer" by Vu et al. (2011b). The three covariates associated with this triangle relation are:

$$\begin{aligned} s_j^{\text{seller}}(t) &= \sum_{i \neq k} Y_{ki}(t) Y_{ij}(t) Y_{kj}(t), && (1.75) \\ s_j^{\text{broker}}(t) &= \sum_{i \neq k} Y_{kj}(t) Y_{ji}(t) Y_{ki}(t), && (1.76) \\ s_j^{\text{buyer}}(t) &= \sum_{i \neq k} Y_{jk}(t) Y_{ki}(t) Y_{ji}(t). && (1.77) \end{aligned}$$

Finally, the out-path covariates count the number of out-going citations:

$$\begin{aligned} s_j^{OD1}(t) &= \sum_{i=1}^{N} Y_{ji}(t), \\ s_j^{OD2}(t) &= \sum_{i \neq k} Y_{jk}(t) Y_{ki}(t). && (1.78) \end{aligned}$$

The fifty-dimensional statistic vector is based on a Latent Dirichlet Allocation (LDA) model (Blei et al., 2003), which is a three-level hierarchical Bayesian model. We assume that there is a pool of topics, then words in a document are generated from this pool of topics. Each topic is characterized by a distribution over words. In this model, documents are represented by random mixtures over latent topics, yet the number of topics must be set a priori in a likelihood modeling framework. So in a citation network, each paper is associated with a topic vector. The LDA covariates are the coordinate-wise product of the two topic vectors of the two papers, and they measure the similarities on topics between two papers. Presumably, the more similar the two papers are, the more likely that there will be a citation event between them. In an egocentric framework, if the topic vector for a node is denoted by $\boldsymbol{\theta}$, and the arrival time by t^{arr}, then when node j arrives at the network, for all the at-risk nodes i, the vector of LDA covariates can be calculated by

$$\boldsymbol{s}_i^{LDA}(t_j^{arr}) = \boldsymbol{\theta}_i \odot \boldsymbol{\theta}_j \qquad (1.79)$$

In general, the total number of covariates can be large, especially for the LDA

covariates, since the number of topics is often taken to be quite large but arbitrary. To reduce model complexity and determine which of the covariates are not useful in predicting network behavior, variable selection techniques on survival models can be generalized to this network model setting. In Chapter 3, we study the variable selection problem by maximizing an approximated penalized likelihood function. If the true $\boldsymbol{\beta}$ is sparse, then the estimator is shown to be consistent with an asymptotic normal distribution. The so-called oracle property is also established. Finally Chapter 4 will focus on the computational implementation of this approach.

1.5 Organization of This Dissertation

In summary, the organization of this dissertation is as follows. In Chapter 2, we study a novel penalized profile likelihood approach to variable selection for the proportional odds model and derive asymptotic properties of the resulting estimators. Additionally, some novel algorithms are proposed under the framework of MM algorithms. Chapters 3 and 4 study the variable selection problem in a dynamic model setting using a penalized partial likelihood approach. Specifically, Chapter 3 focuses on the asymptotic properties of the estimates while Chapter 4 is about computational challenges and algorithms. Finally, chapter 5 describe some future work arising from this dissertation.

Chapter 2

Variable Selection for the Proportional Odds model

2.1 Introduction

The proportional odds model for survival data is a popular alternative to the well-known Cox proportional hazards model. Each of these models allows the distribution of the survival time to depend on covariates in a prescribed way. A key modeling task, particularly when the number of such covariates is large, is to find a subset of variables that parsimoniously describes the survival distribution; in particular, variables without important predictive power should be excluded if possible.

Variable selection for Cox's model has been extensively studied in the literature. For example, Tibshirani et al. (1997) use a variation of the LASSO method to shrink some coefficient estimates to zero. Fan and Li (2002) extend the SCAD penalty, introduced by Fan and Li (2001) as an alternative to the LASSO approach, to the Cox model. They demonstrate that, under certain regularity conditions, the resulting estimates have the oracle property, which essentially means that estimation behaves asymptotically as though the unimportant variables are known a priori. Zhang and Lu (2007) improve the original LASSO estimator by the adaptive LASSO for Cox's model, so that the resulting estimates also have the oracle property.

In contrast, there is only a small literature about the theoretical properties of variable selection techniques for the proportional odds model. One reason for this difference is the lack of a closed-form partial likelihood function for this model. In one of the few articles on this topic, Lu and Zhang (2007) propose to maximize the penalized marginal likelihood function, which has no closed form, to select variables. However, the marginal likelihood function can only be approximated by stochastic algorithms and the theoretical properties of this method are unknown. More recently, Liu and Zeng (2013) study the problem of variable selection for linear transformation models, which include both the Cox proportional hazards model and the proportional odds model as special cases. Their method introduces a latent variable that leads, using familiar ideas related to EM algorithms, to a replacement for the log-likelihood function. They derive asymptotic properties of the penalized version of this new objective function, including consistency and oracle properties, and also propose a new algorithm for the optimization.

Distinct from the two above approaches, this article proposes a variable selection method based on maximizing the penalized profile likelihood function. The profile likelihood function is more general than the partial likelihood function, but it is still a parametric function of the regression coefficients only. We prove consistency, asymptotic normality, and an oracle property for the resulting estimators under certain regularity conditions. We also discuss the similarities and differences between our method and the methods above, particularly that of Liu and Zeng (2013). Unlike the work of Fan and Li (2002) or Liu and Zeng (2013), our proofs cannot rely on Taylor expansions/derivatics because the differentiability of the profile likelihood cannot be established. Instead, we extend the work of Murphy and Van der Vaart (2000) to prove our results. Because this work applies to semiparametric likelihoods more generally, our results should have broad applicability beyond the case of proportional odds models. Finally, we develop several novel algorithms to maximize the penalized likelihood function iteratively and discuss their relative merits. Certain of these algorithms are shown to perform well in simulations and in practice.

The remainder of this article is organized as follows: Section 2.2 describes the proportional odds model, Section 2.3 demonstrates the application of penalization methods to the proportional odds loglikelihood, Section 2.4 proves two

theorems establishing results similar to those in the literature on penalized parametric likelihoods, Section 2.5 discusses algorithms for implementing our theory, and Sections 2.6 and 2.7 test our algorithms on simulated and real datasets.

2.2 The Proportional Odds Model

In survival analysis, we are interested in modeling the distribution of event times T, often as a function of covariates measured on a sample of individuals and based on censored observations of T for those individuals. With right-censored data, the observed data for each individual are $Y = \min(T, C)$, the censoring indicator $\delta = I_{\{T<C\}}$, and the covariate vector $\mathbf{x}$, where C indicates the censoring time. We usually assume that given the covariate, T and C are independent, an assumption called non-informative censoring. The distribution of T can be completely characterized by either of two functions, the survival function $S(t) = P(T > t)$ and the hazard function $h(t) = -(d/dt) \log S(t)$.

A popular model in survival analysis is Cox's proportional hazards model (Cox, 1972), so-called because its hazard function

$$h(t) = h_0(t) \exp\{\boldsymbol{\beta}^\top \mathbf{x}\} \tag{2.1}$$

implies that the hazards for two different individuals are always in the same proportion for all t:

$$\frac{h_1(t)}{h_2(t)} = \exp\{\boldsymbol{\beta}^\top (\mathbf{x}_1 - \mathbf{x}_2)\}. \tag{2.2}$$

The function $h_0(t)$ in Equation (2.1) is called the baseline hazard function and it may be interpreted as the hazard function of an individual for whom $\mathbf{x} = \mathbf{0}$.

In contrast to the proportional hazard assumption, the proportional odds model assumes time-invariant ratios of odds between different individuals. It was first introduced by McCullagh (1980) to analyze categorical data, then later Bennett (1983) generalized the model to a medical context using the language of survival analysis. The proportional odds model is defined as

$$\frac{F(t; \mathbf{x}_i)}{1 - F(t; \mathbf{x}_i)} = \frac{F_0(t)}{1 - F_0(t)} \exp\{\boldsymbol{\beta}^\top \mathbf{x}_i\},$$

where $H_0(t) \doteq F_0(t)/[1 - F_0(t)]$ is called the baseline odds.

Unlike Cox's model, the proportional odds model has the property that the ratio of hazards converges to 1 as t tends to ∞. This property makes the proportional odds model an alternative to Cox's model when the proportional hazards assumption does not hold. Examples include situations when initial effects disappear over time, as when a cure or treatment is effective (Bennett, 1983).

Both Cox's model and the proportional odds model are semi-parametric models, with $\boldsymbol{\beta}$ the parametric part and the baseline hazard/odds the nonparametric part. Interest usually centers on how the covariates will influence the distribution of time-to-event via $\boldsymbol{\beta}$, so the baseline hazard/odds is treated as a nuisance parameter (Oakes, 1981). Thus, modifications of the regular likelihood function are needed to reduce the dimensionality of the "nuisance" nonparametric baseline function (Cox, 1975).

One such modification is the use of partial likelihoods (Cox, 1975). The technical definition of a partial likelihood is somewhat complicated, but any partial likelihood is based on conditioning, and in fact any conditional likelihood is a special case of a partial likelihood. In survival model settings, a partial likelihood function is a product of conditional probabilities, each conditioning term encompassing all previous events (Oakes, 1981). When the data on which we condition are the event times and censoring indicators, Cox's proportional hazards model results in the simple partial likelihood function

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\exp\{\boldsymbol{\beta}^\top \boldsymbol{x}_i\}}{\sum_{j \in R_i} \exp\{\boldsymbol{\beta}^\top \boldsymbol{x}_j\}}, \tag{2.3}$$

where R_i is the "at-risk" set at event time i. This partial likelihood function is completely free of the parameter h_0 and thus leads to straightforward estimation. In addition, the partial likelihood in the Cox model coincides with the profile likelihood, $\max_{h_0} L_{\text{full}}(\boldsymbol{\beta}, h_0)$, whose maximizer $\hat{\boldsymbol{\beta}}$ coincides with that of the full likelihood. It is the convenience of these facts from a computational perspective that is partly responsible for the widespread popularity of the proportional hazards model in survival analysis.

In contrast, the proportional odds model admits no known simple partial likelihood form, and even profile likelihood is somewhat complicated. However, Murphy

et al. (1997) showed that the profile likelihood function may be studied because in $\max_{H_0} L_{\text{full}}(\boldsymbol{\beta}, H_0)$, the maximization may be taken over a smaller class than all possible baseline odds functions. They verified the existence of the MLE, derived its consistency and asymptotic normality, and showed that the MLE of H_0, the baseline odds function, is a step function with jumps $\Delta H_0(Y_i)$ only at each observed (i.e., uncensored) time. Therefore, they re-parameterized the nonparametric parameter H_0 by jumps $\Delta H_0(Y_i)$ at the uncensored event times. We will discuss this reparameterization in the next section after first describing penalization methods.

2.3 Variable Selection via Penalization of the Profile Likelihood

Many settings in statistical modeling require selecting from a potentially large number of covariates those which are important. In recent years, variable selection via penalized likelihood-based functions has attracted much attention. One advantage of this approach is that it allows one to conduct variable selection and parameter estimation simultaneously. A penalized likelihood function can be written as

$$\ell(\beta) - n\sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{2.4}$$

where $\ell(\beta)$ is the (log-)likelihood function and $p_\lambda(|\cdot|)$ is the penalty function. Commonly used penalty functions include the L_1 or LASSO penalty (Tibshirani, 1996),

$$p_\lambda(\theta) = \lambda|\theta|, \tag{2.5}$$

and the smoothly clipped absolute deviation (SCAD) penalty (Fan, 1997), whose derivative for $\theta > 0$ is defined by

$$p_\lambda'(\theta) = \lambda\left\{I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\lambda > \theta)\right\} \text{ for some } a > 2. \tag{2.6}$$

If the penalty is chosen properly, maximization of the objective function (2.4) can shrink some coefficients to zero, which is the reason this approach is suitable for variable selection. If the likelihood function is concave and the penalty function

convex and smooth, any method suitable for maximizing a concave smooth function can be employed to find the penalized likelihood estimator. However, neither the SCAD penalty function nor the L_p penalty for $p \leq 1$ is convex and smooth, which can make the whole penalized likelihood function non-concave. To solve this problem, Fan and Li (2001) studied variable selection via the nonconcave SCAD penalty using local quadratic approximation (LQA) of the penalty term. But this approach shares a drawback with stepwise variable selection: When a variable is deleted at any iteration, it will be excluded from the final model. Hunter and Li (2005) showed that LQA is a type of MM algorithm, as discussed in Section 2.5.1, and introduced a method for remedying the flaw of LQA. However, this approach requires introducing a slightly perturbed penalty function, thus slightly weakening the theoretical properties of the resulting estimators. Zou and Li (2008) introduce a different class of MM algorithm for variable selection, called local linear approximation (LLA) of the penalty function, which overcomes both drawbacks. In LLA, the penalty function is approximated for $\theta \approx \theta^{(0)}$ as

$$p_\lambda(|\theta|) \approx p_\lambda(|\theta^{(0)}|) + p_\lambda^{'}(|\theta^{(0)}|)(|\theta| - |\theta^{(0)}|). \tag{2.7}$$

Therefore, the penalty function is approximated by a convex function, and the penalized likelihood is concave (though the concavity is strict only if the log-likelihood is strictly concave). From the MM algorithm point of view, LLA actually finds a so-called minorizing function for the penalty function, a fact that we exploit in this article.

In variable selection for the proportional odds model, we only want to penalize the coefficients of the covariates, not the baseline odds function. Murphy et al. (1997) show that the full likelihood may be written as

$$\prod_{i=1}^{n} \left(\frac{e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}}{H_0(Y_i) + e^{-\mathbf{x}_i^\top \beta}} \right) \left(\frac{\Delta H_0(Y_i)}{H_0(Y_i) - \Delta H_0(Y_i) + e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}} \right)^{\delta_i} \tag{2.8}$$

without affecting its maximizer, since the best $H_0(t)$ is always a step function with jumps $\Delta H_0(Y_i)$. Hunter and Lange (2002) prove the log-likelihood is concave under a further reparameterization: If m is the number of distinct uncensored event times, then for $1 \leq j \leq m$ we define γ_j to be the log of the jump in the baseline

odds for the jth smallest uncensored event time. In other words, if the distinct uncensored times are $U_1 < U_2 < .... < U_m$, and if we let $\omega_i = \max\{j : U_j \leq Y_i\}$ for all $1 \leq i \leq n$, then $\gamma_{\omega_i} = \log \Delta H_0(Y_i)$. Under this reparameterization, the log-likelihood function becomes

$$\ell(\boldsymbol{\theta}) = \log L_{\text{full}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} -\mathbf{z}_i^\top \boldsymbol{\beta} - \ln D_i(\boldsymbol{\theta}) + \delta_i\{\gamma_{\omega_i} - \ln[D_i(\boldsymbol{\theta}) - e^{\gamma_{\omega_i}}]\}, \quad (2.9)$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathbb{R}^{p+m}$ and $D_i(\boldsymbol{\theta}) = e^{-\mathbf{z}_i^\top \boldsymbol{\beta}} + \sum_{j=1}^{\omega_i} e^{\gamma_j}$. Substituting Equation (2.9) into Equation (2.4), the objective function to be maximized becomes

$$\xi^{\text{full}}(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - n \sum_{i=1}^{p} p(|\beta_i|, \lambda_n^i), \lambda_n^i > 0, \quad (2.10)$$

where the notation $\xi^{\text{full}}(\boldsymbol{\theta})$ indicates that full log-likelihood function is penalized. In equation (2.10), we take $\lambda_n^i = \lambda_n$ for all i for the sake of simplicity. The tuning parameter λ_n is usually determined by a data-driven method that depends on the sample size n.

Since a primary goal is to select variables, the baseline odds can be treated as a nuisance parameter. One approach to handling semi-parametric models with functional nuisance parameters is to use partial likelihood (Cox, 1975), though as we point out in Section 2.2, this approach is intractable in the proportional odds model. Lu and Zhang (2007) address this problem by penalizing the marginal likelihood function, which is baseline-free but has no analytical form. They use importance sampling to approximate the penalized marginal likelihood function. Liu and Zeng (2013) use an alternative approach in which they apply an EM algorithm to the unpenalized log-likelihood after embedding the proportional odds model in a larger class of models that also includes Cox's model, then penalize a function related to the log-likelihood that is constructed as part of this EM algorithm. We discuss their approach in more detail in Section 2.5.1.

Here, we adopt a profiling approach, where the profile log-likelihood is defined as

$$p\ell(\boldsymbol{\beta}) = \max_{\boldsymbol{\gamma} \in \mathbb{R}^m} \ell\left[(\boldsymbol{\beta}, \boldsymbol{\gamma})\right]. \quad (2.11)$$

The profile likelihood function retains more complete information of the full likelihood function than the marginal likelihood function does, but generally it has no closed form and it need not be differentiable. Murphy and Van der Vaart (2000) study profile likelihood functions and derived an expansion of the profile log-likelihood, akin to a Taylor expansion, using the so-called effective score and effective in place of the usual derivatives. For any $\tilde{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$, they show that

$$\begin{aligned} p\ell(\tilde{\boldsymbol{\beta}}) &= p\ell(\boldsymbol{\beta}_0) + (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \sum_{i=1}^{n} \tilde{\ell}_0(\mathbf{x}_i) \\ &\quad -\frac{1}{2} n (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \tilde{I}_0 (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_{P_{\boldsymbol{\beta}_0, h_0}}(\sqrt{n}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| + 1)^2, \end{aligned}$$

where the effective score function $\tilde{\ell}_0$ can be considered as the score of the full log-likelihood function in the direction of the so-called "least favorable sub-models", and the effective information $\tilde{I}_0$ is the covariance of the effective score function. Similar to estimation in parametric models with nuisance parameters, the effective score can be regarded as the orthogonal projection of the score function of $\boldsymbol{\beta}$ onto the space spanned by the score of the nuisance parameter. Therefore, the effective information is the maximum attainable information when estimating the parameter of interest in the presence of the nuisance parameters (Severini and Wong, 1992; van der Vaart, 2000; Murphy and Van Der Vaart, 1999). In addition, Murphy and Van der Vaart (2000) give the form of the effective score and information for several specific semi-parametric models, including the proportional odds model. This motivates us to replace the penalized likelihood by the penalized profile likelihood function for the proportional odds model,

$$\xi^{\text{prof}}(\boldsymbol{\beta}) = p\ell(\boldsymbol{\beta}) - n \sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|), \tag{2.12}$$

as the maximizers of the two functions are identical. Since the penalty term does not involve the nuisance parameter that is being profiled out, $\xi^{\text{prof}}(\boldsymbol{\beta})$ may be viewed as either the penalized profile likelihood or the profile penalized likelihood.

2.4 Asymptotic Results

In this section, we show that the maximizer of $\xi^{\text{prof}}(\boldsymbol{\beta})$ is consistent and has an oracle property under certain regularity conditions. Let us decompose the true parameter vector $\boldsymbol{\beta}^0$ into a nonzero part and a zero part, i.e., $\boldsymbol{\beta}^0 = (\boldsymbol{\beta}^{10}, \boldsymbol{\beta}^{20}), \boldsymbol{\beta}^{10} \in \mathbb{R}^s, \boldsymbol{\beta}^{20} \in \mathbb{R}^r, r + s = p$, and $\boldsymbol{\beta}^{20} = \mathbf{0}$. We first show that under certain regularity conditions, there exists a penalized likelihood estimator that is consistent for $\boldsymbol{\beta}^0$. Furthermore, the rate of convergence may be shown to be $n^{-1/2}$ for certain choices of penalty function, which parallels the consistency result for parametric models of Fan and Li (2001).

Before stating the theorems, we introduce the following regularity conditions:

1. As $n \to \infty$,
$$a_n \stackrel{\text{def}}{=} \max_{1 \le j \le s} \{p'_{\lambda_n}(|\beta_j^0|)\} \to 0$$
and
$$b_n \stackrel{\text{def}}{=} \max_{1 \le j \le s} \{p''_{\lambda_n}(|\beta_j^0|)\} \to 0.$$

2. The sequence λ_n tends to zero in such a way that there exists $0 < \delta < \frac{1}{2}$ such that $n^\delta \lambda_n \asymp 1$, i.e., there exist $0 < m < M$ and N such that $m < |n^\delta \lambda_n| < M$ for all $n > N$.

3. The penalty function $p_{\lambda_n}(x)$ is twice differentiable at all $x > 0$ and there exist positive constants K and K' such that for all $n = 1, 2, ...$ and for all $x > 0$, $p''_{\lambda_n}(x)$ exists, $|p''_{\lambda_n}(x)| < K'$, and $p'_{\lambda_n}(0+)/\lambda_n > K$.

4. There exists $0 < \delta^* \le 1$ such that for any random sequence
$$\tilde{\boldsymbol{\theta}}^n = (\tilde{\boldsymbol{\theta}}^{1n}, \tilde{\boldsymbol{\theta}}^{2n}) \to \boldsymbol{\theta}^0 = (\boldsymbol{\theta}^{10}, \mathbf{0}),$$
we have
$$\begin{aligned} p\ell(\tilde{\boldsymbol{\theta}}^n) &= p\ell(\boldsymbol{\theta}^0) + (\tilde{\boldsymbol{\theta}}^n - \boldsymbol{\theta}^0)^\top \sum_{i=1}^n \tilde{\ell}_0(\mathbf{x}_i) - \frac{1}{2} n (\tilde{\boldsymbol{\theta}}^n - \boldsymbol{\theta}^0)^\top \tilde{I}_0 (\tilde{\boldsymbol{\theta}}^n - \boldsymbol{\theta}^0) \\ &\quad + o_P(n\|\tilde{\boldsymbol{\theta}}^{1n} - \boldsymbol{\theta}^{10}\|^2) + o_P(n\|\tilde{\boldsymbol{\theta}}^{1n} - \boldsymbol{\theta}^{10}\| \cdot \|\tilde{\boldsymbol{\theta}}^{2n}\|) \\ &\quad + o_P(n\|\tilde{\boldsymbol{\theta}}^{2n}\|^{1+\delta^*}), \end{aligned} \tag{2.13}$$

in which the $o_P(n\|\tilde{\boldsymbol{\theta}}^{1n} - \boldsymbol{\theta}^{10}\|^2)$ remainder term does not depend on $\tilde{\boldsymbol{\theta}}^{2n}$.

Theorem 1. Assume $(\boldsymbol{Y}_i, \boldsymbol{\delta}_i, \boldsymbol{x}_i), i = 1, 2, ..., n$, are independent and identically distributed from a proportional odds model and $p\ell(\boldsymbol{\beta})$ denotes the profile likelihood function for $\boldsymbol{\beta}$. Define $\xi^{\text{prof}}(\boldsymbol{\beta})$ as in Equation (2.12).

(a) ($\sqrt{n}$ consistency) Assume Condition 1 holds. Then there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of $\xi^{\text{prof}}(\boldsymbol{\beta})$, and hence a local maximizer of $\xi^{\text{full}}(\boldsymbol{\beta})$, such that

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| = O_p(n^{-1/2} + a_n).$$

(b) In addition, suppose Conditions 2 and 3 hold. Let $\hat{\boldsymbol{\beta}}^{1n}$ denote the first s components of the $\sqrt{n}-$consistent estimator $\hat{\boldsymbol{\beta}}$ in (a). Then for $c > 0$,

$$\hat{\boldsymbol{\beta}}^{2n} \stackrel{\text{def}}{=} \arg \max_{\|\boldsymbol{\beta}^{2n}\| \le cn^{-1/2}} \xi^{\text{prof}}\{(\hat{\boldsymbol{\beta}}^{1n}, \boldsymbol{\beta}^{2n})\} = o_P(n^{-1+\delta}).$$

(c) (sparsity) Furthermore, if we assume that Condition 4 holds, λ_n is a sequence satisfying Condition 2, and $\delta = \delta^*/(1 + \delta^*)$, then for any $\boldsymbol{\beta}^{1n}$ satisfying $\|\boldsymbol{\beta}^{1n} - \boldsymbol{\beta}^{10}\| = O_P(n^{-1/2})$ and any constant $c > 0$,

$$P\left(\xi^{\text{prof}}\{(\boldsymbol{\beta}^{1n}, \mathbf{0})\} = \max_{\|\boldsymbol{\beta}^{2n}\| \le cn^{-1/2}} \xi^{\text{prof}}\{(\boldsymbol{\beta}^{1n}, \boldsymbol{\beta}^{2n})\}\right) \to 1 \text{ as } n \to +\infty.$$

That is, $\hat{\boldsymbol{\beta}}^{2n} = \mathbf{0}$ with probability tending to one.

Theorem 1, whose proof is in the Appendix, has three parts, and each part gives more advanced results than the previous one. Part (a) establishes the consistency of the estimates, with the convergence rate depending on the form of penalty functions. For the hard thresholding and SCAD penalty functions, if $\lambda_n \to 0$ as $n \to \infty$, then $a_n = 0$ for n large enough. Therefore, in this case part (a) proves that the estimator is $\sqrt{n}$-consistent. Part (b) shows that in fact $\hat{\boldsymbol{\beta}}^{2n}$ approaches its true value of $\mathbf{0}$ faster than $n^{-1/2}$, though this is still not as strong as the sparsity condition, adopting the terminology of Fan and Li (2001), established by part (c). Establishing these results is challenging in the current context because, unlike in Fan and Li (2001), we cannot assume the differentiability of the profile log-likelihood. Although our sufficient conditions in equation (2.13) are stronger than

those of Murphy and Van der Vaart (2000), they are weaker than differentiability and they are the weakest conditions for which we are currently able to establish sparsity.

Theorem 2 establishes the asymptotic normality of the nonzero component $\hat{\boldsymbol{\beta}}^{1n}$, which can be proved even without sparsity condition of Theorem 1(c).

Theorem 2 (asymptotic normality). Suppose $\boldsymbol{\beta}^{10} = \{\beta_1^0, \beta_2^0, ..., \beta_s^0\}$ and let

$$\boldsymbol{d}_n = \begin{pmatrix} p'_{\lambda_n}(|\beta_1^0|)\text{sgn}(\beta_1^0) \\ \vdots \\ p'_{\lambda_n}(|\beta_s^0|)\text{sgn}(\beta_s^0) \end{pmatrix}, \Sigma_n = \begin{pmatrix} p''_{\lambda_n}(|\beta_1^0|)\text{sgn}(\beta_1^0) & & 0 \\ & \ddots & \\ 0 & & p''_{\lambda_n}(|\beta_s^0|)\text{sgn}(\beta_s^0) \end{pmatrix}.$$

Assume $\sqrt{n}\boldsymbol{d}_n \to 0$ and Conditions 1, 2, and 3 hold. Then the $\sqrt{n}$-consistent local maximizer of Theorem 1(a) must satisfy

$$\sqrt{n}(\tilde{I}_{11}^0 + \Sigma_n))\{\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10} + (\tilde{I}_{11}^0 + \Sigma_n)^{-1}\boldsymbol{d}_n\} \xrightarrow{\mathcal{L}} N(0, \tilde{I}_{11}^0),$$

where $\tilde{I}_{11}^0$ is the upper left $s \times s$ submatrix of $\tilde{I}_0$.

The above theorems establish results analogous to those of Fan and Li (2001): Under certain regularity conditions and with properly chosen penalty functions, the maximizer of the penalized profile likelihood function will have desirable properties like consistency, sparsity and asymptotic normality. As a consequence, the asymptotic variance is

$$\frac{1}{n}(\tilde{I}_{11}^0 + \Sigma_n)^{-1}\tilde{I}_{11}^0(\tilde{I}_{11}^0 + \Sigma_n)^{-1},$$

which approaches $(\tilde{I}_{11}^0)^{-1}$ when $\lambda_n \to 0$. As discussed at the end of Section 2.3, the effective information is an upper bound on the information in a semiparametric model. Therefore, the maximizer of the penalized profile likelihood function is asymptotically efficient.

Our results, which are very similar to those derived for the estimator of Liu and Zeng (2013), rely nevertheless on a different method of proof, since differentiability of the penalized profile log-likelihood may not hold. In addition, Theorems 1 and 2 should extend to any semi-parametric model that allows for the expansion of the profile log-likelihood function seen in Equation (2.13).

2.5 Maximizing the Penalized Likelihood

Since there is no closed form of the profile likelihood function, optimization of the penalized profile likelihood function cannot be done in a straightforward way. The algorithm introduced in this section falls under the general framework of MM algorithms. We will first describe this framework, then introduce the algorithm.

2.5.1 MM Algorithms

MM is not a specific algorithm, but a general principle for deriving algorithms to solve optimization problems. It can often separate variables, making it suitable for high-dimensional problems. Convexity is often crucial to establish the properties of an MM algorithm, which is the reason for the reparametrization in the previous section.

In maximization problems, a standard MM algorithm consists of alternations between two steps. The first M stands for minorization and the second M is for maximization. For a fixed $\boldsymbol{\theta}^{(k)}$, a function $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ is said to minorize another function $f(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^{(k)}$ if

$$g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) \leq f(\boldsymbol{\theta}) \quad \text{for all } \boldsymbol{\theta}; \tag{2.14}$$

$$g(\boldsymbol{\theta}^{(k)}|\boldsymbol{\theta}^{(k)}) = f(\boldsymbol{\theta}^{(k)}). \tag{2.15}$$

At each iteration, given the value of the parameter $\boldsymbol{\theta}^{(k)}$, we will maximize the minorizing function $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ in $\boldsymbol{\theta}$. Then the maximum point $\boldsymbol{\theta}^{(k+1)}$ will force the value of the objective function $f(\boldsymbol{\theta})$ uphill in the sense that $f(\boldsymbol{\theta}^{(k+1)}) \geq f(\boldsymbol{\theta}^{(k)})$ is guaranteed. This is called the ascent property of an MM algorithm, which protects the algorithm from unpredictable behavior. The well-known class of EM algorithms is a subset of the MM algorithms; in fact, the E-step of any EM algorithm is actually a minorization step (Hunter and Lange, 2004). We shall revisit EM algorithms below.

The key to an effective MM algorithm is to find a good minorizing function. Hunter and Lange (2002) provide a minorizing function for the log-likelihood func-

tion (2.9) of the proportional odds model:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = \sum_{i=1}^{n} Q_i^{\beta}(\boldsymbol{\beta}|\boldsymbol{\theta}^{(k)}) + \sum_{j=1}^{m} Q_j^{\gamma}(\gamma_j|\boldsymbol{\theta}^{(k)}), \tag{2.16}$$

where

$$Q_i^{\beta}(\boldsymbol{\beta}|\boldsymbol{\theta}^{(k)}) = -\mathbf{z}_i^{\top}\boldsymbol{\beta} - e^{-\mathbf{z}_i^{\top}\boldsymbol{\beta}}\left[\frac{1}{D_i(\boldsymbol{\theta}^{(k)})} + \frac{\delta_i}{D_i(\boldsymbol{\theta}^{(k)}) - e^{\gamma_{\omega_i}^{(k)}}}\right], \tag{2.17}$$

$$Q_j^{\gamma}(\gamma_j|\boldsymbol{\theta}^{(k)}) = u_j v_j - e^{\gamma_j}\left[\sum_{i:\omega_i \geq j} \frac{1}{D_i(\boldsymbol{\theta}^{(k)})} + \sum_{i:\omega_i > j} \frac{\delta_i}{D_i(\boldsymbol{\theta}^{(k)}) - e^{\gamma_{\omega_i}^{(k)}}}\right]. \tag{2.18}$$

The most important features of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ are that it separates the parameters into γ_j and $\boldsymbol{\beta}$, and both parts are twice differentiable. Also, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ itself is a concave function of $\boldsymbol{\theta}$, with negative definite Hessian matrix with respect to $\boldsymbol{\beta}$. Thus, we can use Newton-Raphson in the maximization step for updating $\boldsymbol{\beta}$. Also, for each fixed $\boldsymbol{\theta}^{(k)}$, there is a closed form maximizer for $\boldsymbol{\gamma}$, as we explain in Section 2.5.2.

Adding the penalty term, we obtain

$$Q^{\text{pen}}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) - n\sum_{i=1}^{p} p_{\lambda}(|\beta_i|)$$

as a minorizing function of $\xi^{\text{full}}(\boldsymbol{\theta})$ at the point $\boldsymbol{\theta}^{(k)}$. This minorizing function is much easier to maximize than $\xi^{\text{full}}(\boldsymbol{\theta})$ itself, and this fact leads to the iterative algorithms of Section 2.5.2.

The idea of Liu and Zeng (2013) is similar but with a key difference. Their modeling framework makes clever use of a latent (unobserved) variable and establishes the proportional odds model as a special case of a broader class of models that also includes Cox's model. To find the unpenalized MLE for the proportional odds case, Liu and Zeng (2013) exploit the missing-data structure of their modeling framework and construct a standard EM algorithm. At each iteration of this algorithm, the E-step constructs a minorizing function. When the maximizer is achieved, their idea is to utilize the minorizing function at the MLE as an approximation to the log-likelihood, and apply a penalty to this minorizer. They then maximize this penalized approximate log-likelihood. By contrast, our method is

to apply a penalty directly to the log-likelihood—though as explained above, we also use a (different) MM algorithm to achieve the maximization. The distinction is illustrated by Figure 2.1. The asymptotic results suggest no clear theoretical advantage of one method over the other, though our simulation studies in Section 2.6 suggest a small advantage for our approach for larger sample sizes.

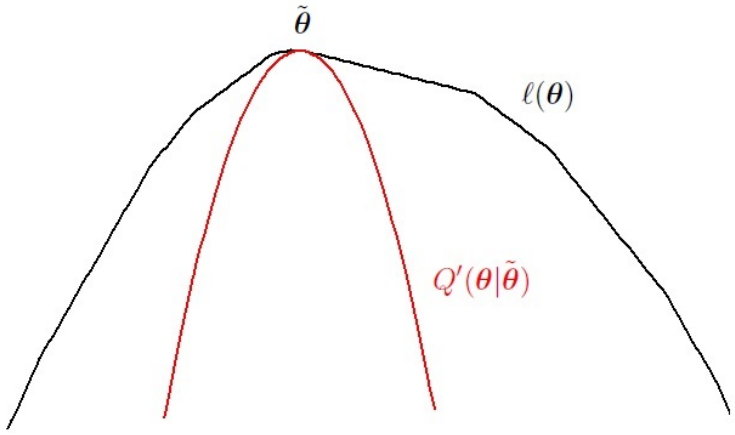

Figure 2.1. Whereas our method applies a penalty function directly to the log-likelihood $\ell(\boldsymbol{\theta})$, Liu and Zeng (2013) penalize a minorizer of $\ell(\boldsymbol{\theta})$ at the MLE $\tilde{\boldsymbol{\theta}}$, depicted here as $Q'(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$.

2.5.2 Iterative Conditional Maximization

In order to maximize $Q^{\text{pen}}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$, we still face problems of non-concavity and lack of smoothness because of the penalty term. For penalized parametric likelihoods, Zhang and Li (2009) propose an iterative conditional maximization (ICM) algorithm to solve such problems. Throughout their paper, they assume that the likelihood function follows the regularity conditions in Fan and Li (2001), in order to guarantee that the log-likelihood function is at least twice differentiable and locally concave at the true unknown parameter $\boldsymbol{\beta}_0$. This algorithm is simple and enjoys a fast convergence rate.

For the proportional odds model, since $Q(\boldsymbol{\theta})$ can be separated into a function of $\boldsymbol{\beta}$ and a function of $\boldsymbol{\gamma}$, and we only penalize $\boldsymbol{\beta}$, maximizing $Q^{\text{pen}}(\boldsymbol{\theta})$ can be done

separately for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. The optimal γ_j is the maximizer of $Q_j^{\gamma}(\gamma_j|\boldsymbol{\theta}^{(k)})$, which has the closed form expression

$$\boldsymbol{\gamma}^{(k+1)} = \log u_j - \log\left[\sum_{i:\omega_i>j} \frac{1}{D(\boldsymbol{\theta}^{(k)})} + \sum_{i:\omega_i>j} \frac{\delta_i}{D(\boldsymbol{\theta}^{(k)}) - e^{\gamma_{\omega_i}^{(k)}}}\right] \tag{2.19}$$

given by Hunter and Lange (2002). For updating $\boldsymbol{\beta}$, we need to maximize

$$Q^{\beta-\text{pen}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} Q_i^{\beta}(\boldsymbol{\beta}|\boldsymbol{\theta}^{(k)}) - n\sum_{j=1}^{p} p_\lambda(|\beta_j|) = Q^{\beta}(\boldsymbol{\beta}|\boldsymbol{\theta}^{k}) - n\sum_{j=1}^{p} p_\lambda(|\beta_j|). \tag{2.20}$$

Although $Q^{\beta-\text{pen}}(\boldsymbol{\beta})$ is not a penalized likelihood function, it may be treated like one for the purpose of maximization by the ICM algorithm. Only cosmetic changes to the proofs in Zhang and Li (2009) are necessary to validate all the theoretical results in that paper as they apply to maximizing $Q^{\beta-\text{pen}}(\boldsymbol{\beta})$. Now we can combine the MM and ICM algorithms to get a new algorithm to solve the variable selection problem for the proportional odds model.

The algorithm consists of two loops: the outer loop encompasses the MM algorithm and the inner loop maximizes the $Q^{\beta-\text{pen}}(\boldsymbol{\beta})$ as a "penalized likelihood function" using the ICM algorithm. The steps of the algorithm are as follows:

1. Initialize $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\gamma}^{(0)})$; set $k = 0$.

2. Update $\boldsymbol{\gamma}^{(k)}$ to $\boldsymbol{\gamma}^{(k+1)}$ using equation (2.19).

3. Update $\boldsymbol{\beta}$ by maximizing $Q^{\beta-\text{pen}}(\boldsymbol{\beta})$, as follows:

$$\begin{aligned}
\Delta\boldsymbol{b}^{(k)} &= -d^2Q^{\beta}(\boldsymbol{\beta}^{(k)}|\boldsymbol{\theta}^{(k)})^{-1}dQ^{\beta}(\boldsymbol{\beta}^{(k)}|\boldsymbol{\theta}^{(k)})^{\top},\\
\alpha^{(k)} &= \arg\max_{\alpha\in[0,1]} Q^{\beta}(\boldsymbol{\beta}^{(k)} + \alpha\Delta\boldsymbol{b}^{(k)}|\boldsymbol{\theta}^{(k)}),\\
\boldsymbol{b}^{(k)} &= \boldsymbol{\beta}^{(k)} + \alpha^{(k)}\Delta\boldsymbol{b}^{(k)}.
\end{aligned}$$

   Then use ICM to get $\boldsymbol{\beta}^{(k+1)}$:

   (a) Set $\widetilde{\boldsymbol{\beta}}^{(0)} = \boldsymbol{b}^{(k)}$,
   compute $\mathbf{m} = (m_1, m_2..., m_p)$, where $m_j = \min_{\theta>0}\{\theta + p_\lambda'(\theta)/\hat{I}_{jj}\}$, and $\hat{I}_{jj} = -d^2 f(\boldsymbol{b}^{(k)}|\boldsymbol{\theta}^{(k)})/n$.

(b) Start from $t = 1$. Update

$$\widetilde{\beta}_j^{*(t)} = \widetilde{\beta}_j^{(t-1)} - \sum_{k=1}^{p} \frac{\hat{I}_{jk}}{\hat{I}_{jj}} \left(\widetilde{\beta}_k^{(t-1)} - \widetilde{\beta}_k^{(0)} \right)$$

and

$$\widetilde{\beta}_j^{(t)} = \left[\widetilde{\beta}_j^{*(t)} - \frac{1}{\hat{I}_{jj}} p_\lambda'(|\widetilde{\beta}_j^{(t-1)}|) \mathrm{sgn}(\widetilde{\beta}_j^{(t-1)}) \right] \; I(|\widetilde{\beta}_j^{*(t)}| > m_j)$$

for $j = 1, 2, ..., p$. Then $t = t + 1$.

(c) Repeat (b) until convergence. Then

$$\boldsymbol{\beta}^{(k+1)} = \widetilde{\boldsymbol{\beta}}_{\text{converge}}$$

4. Replace k by $k + 1$ and return to step 2 until some convergence criterion has been satisfied.

As in Zhang and Li (2009), one may update $\boldsymbol{\beta}$ one coordinate at a time in step (b) above, that is, instead of $(\widetilde{\beta}_1^{(t-1)}, ..., \widetilde{\beta}_j^{(t-1)}, ..., \widetilde{\beta}_p^{(t-1)})$, update $\widetilde{\beta}_j^{(t)}$ using $(\widetilde{\beta}_1^{(t)}, ..., \widetilde{\beta}_{j-1}^{(t)}, \widetilde{\beta}_j^{(t-1)}, ..., \widetilde{\beta}_p^{(t-1)})$. Such coordinate updates typically speed convergence, so we use them here. The stopping criterion is based on the squared Euclidean distance between $\boldsymbol{\beta}^k$ and $\boldsymbol{\beta}^{k+1}$.

2.5.3 Coordinate Descent

An alternative way to maximize $Q^{\beta-\text{pen}}(\boldsymbol{\beta})$ in step 3 of our MM algorithm is the coordinate descent algorithm, proposed by Wu and Lange (2008) for solving LASSO penalized regression. This method calculates the directional derivatives for each coordinate at the current estimate, and the objective function is maximized along the coordinate with the "most negative" directional derivative. This procedure is repeated until there are no negative directional derivatives. Although the LASSO penalty is not differentiable at the origin, its directional derivative is always a constant. Therefore, this algorithm is attractive in terms of computational simplicity.

For a unit vector $\boldsymbol{v}$, we define the directional derivative along $\boldsymbol{v}$ as

$$d_{\boldsymbol{v}}Q^{\beta-\text{pen}}(\boldsymbol{\beta}) = \lim_{\tau\downarrow 0}\frac{Q^{\beta-\text{pen}}(\boldsymbol{\beta}+\tau\boldsymbol{v}) - Q^{\beta-\text{pen}}(\boldsymbol{\beta})}{\tau}.$$

Letting $\boldsymbol{e_k}$ denote the standard basis vector with 1 in the kth place, the two directional derivatives along this coordinate direction are

$$d_{\boldsymbol{e_k}}Q^{\beta-\text{pen}}(\boldsymbol{\beta}) = \frac{\partial}{\partial\beta_k}Q^{\beta}(\boldsymbol{\beta}) + \begin{cases} p'_\lambda(|\beta_k|) & \beta_k > 0, \\ p'_\lambda(0+) & \beta_k = 0, \\ -p'_\lambda(|\beta_k|) & \beta_k < 0 \end{cases}$$

and

$$d_{-\boldsymbol{e_k}}Q^{\beta-\text{pen}}(\boldsymbol{\beta}) = -\frac{\partial}{\partial\beta_k}Q^{\beta}(\boldsymbol{\beta}) + \begin{cases} -p'_\lambda(|\beta_k|), & \beta_k > 0, \\ p'_\lambda(0+), & \beta_k = 0, \\ p'_\lambda(|\beta_k|), & \beta_k < 0. \end{cases}$$

Because here we don't have a simple regression model, maximizing the surrogate function along a certain coordinate does not have a closed form. However, this one-dimensional optimization problem can be easily solved by the golden section or bisection algorithms.

2.5.4 Minimization by Iterative Soft Thresholding

Yet another alternative to ICM in step 3 of our MM algorithm is the Minimization by Iterative Soft Thresholding (MIST) method proposed by Schifano et al. (2010) for optimizing penalized likelihood-based functions of the general form

$$\xi(\beta) = g(\beta) + \lambda\varepsilon\|\beta\|^2 + \sum_{i=1}^{p} p(|\beta_i|, \boldsymbol{\lambda_i}), \tag{2.21}$$

where $\varepsilon \geq 0$ and each $\boldsymbol{\lambda}_i$ is possibly vector-valued with first component equal to $\lambda > 0$. Because of the MIST requirement that $p(|\beta_i|, \boldsymbol{\lambda_i})$ be a concave function, MIST cannot always be applied directly for certain types of penalty functions, such as SCAD; however, it is possible to majorize such a penalty function using a concave function via the LLA idea of Zou and Li (2008).

The MIST algorithm is developed from solving optimization problems for the sum of two functions and it has two stages. The first stage updates the unpenalized likelihood-based function part and the second stage uses a soft-thresholding operator to handle the penalty part. The algorithm iterates between these stages until convergence. It is computationally simple since it avoids high-dimensional matrix inversion. We find that the behavior of this algorithm is good in simulated data sets, even outperforming the ICM-MM algorithm in some cases. However, an important challenge of MIST is that one needs to identify a tuning parameter, ω, for running the algorithm. Usually, if the objective function $g(\boldsymbol{\theta})$ is Lipschitz continuous of order $1/L$, then ω is chosen to be between 0 and $L/2$. This parameter is hard to identify if one doesn't know the true underlying model (likelihood function), resulting in difficulties of implementation in real data analysis, especially for the case of proportional odds models. So we omit the detailed algorithm and corresponding results in the tests that follow. A related algorithm that avoids choosing the tuning parameter is the fast iterative-thresholding algorithm (FISTA) proposed by Beck and Teboulle (2009). The version that incorporates a backtracking stepsize rule could be applied for maximizing the objective function here, but we do not discuss this algorithm in detail.

2.6 Simulation Studies

In this section, we evaluate the performance of different algorithms on simulated data sets using the setting of Lu and Zhang (2007). The eight covariates $(Z_1, Z_2, ..., Z_8)$ are generated from a multivariate normal distribution with $\text{Corr}(Z_i, Z_j) = \rho^{|i-j|}$ for $i \neq j$ and $\rho = 0.2$. The corresponding regression coefficients are $\boldsymbol{\beta} = (-0.7, 0, 0, -0.7, 0, 0, -0.7, 0)$. We choose $H(t) = 3\log(t)$ as the baseline odds function. Two censoring settings, 25 and 40 percent, and two sample size settings, n=100 and 400, are considered. We test both ICM and coordinate descent as the inner loop algorithm in step 3 of the MM algorithm, and both adaptive LASSO and SCAD penalty functions are implemented. Since two algorithms with the same penalty function produce very similar simulation results, Table 2.1 reports only the average results of the ICM and coordinate descent algorithms. Table 1 also summarizes results produced by the algorithm of Liu and Zeng (2013)

using code provided by the authors and reproduces results reported by Lu and Zhang (2007) in the case $n = 100$. A more detailed version of the results is found in the appendix.

There are several possible methods of choosing the tuning parameter λ. Liu and Zeng (2013) and Lu and Zhang (2007) use generalized cross-validation (GCV) for this purpose. Since the form of the true model is known, GCV is comparatively conservative (Zhang et al., 2010). Instead, we use the consistent tuning parameter selector BIC (Zhang et al., 2010). Specifically, the optimal λ is taken to be the point in an equally spaced grid of points that minimizes

$$BIC(\lambda) = \frac{-2\ell(\hat{\boldsymbol{\theta}}) + \mathrm{df}_\lambda \log(n)}{n},$$

where $\ell(\hat{\boldsymbol{\theta}})$ is the log-likelihood function evaluated at the estimates, df_λ is the degrees of freedom of the model, always taken to be the number of nonzero components for $\boldsymbol{\beta}$ for a fixed λ, and n is the number of observations.

There are several observations from the simulation results of Table 2.1. At the smaller sample size ($n = 100$), we find that no algorithm is uniformly better than the others at identifying the true zeros and non-zeros nor at minimizing MSE. However, for the larger sample size ($n = 400$), the two MM algorithms, based on penalizing the true log-likelihood, appear to outperform the algorithm of Liu and Zeng (2013) that penalizes a minorizer of the log-likelihood at the MLE, both in terms of variable selection and MSE. Comparing the SCAD penalty with the adaptive LASSO penalty, it appears that SCAD usually (but not always) enjoys an advantage in terms of MSE and that SCAD tends to declare more coefficients to be zero, both correctly and incorrectly. However, we wish to emphasize that in our experience, all of these results appear to be quite dependent on the choice of tuning parameter.

2.7 Application to Real Data

We apply the algorithm to the Veteran's Administration Lung Cancer Trial data, a dataset on 97 males having lung cancer and no prior therapy that is used by many authors as an example of data following the proportional odds model. The

Table 2.1. Results for 1000 repetitions of the simulation study, where LZ 2013 is the method of Liu and Zeng (2013) and LZ 2007 is the marginal likelihood method of Lu and Zhang (2007). The "Correct" and "Incorrect" columns give the number of parameters correctly and incorrectly set to zero. The mean squared error is given by $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \Sigma (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

censoring	n	Method	Correct(5)	Incorrect(0)	Median MSE
25%	100	MM-aLASSO	4.1	0.1	0.207
		MM-SCAD	4.7	0.3	0.173
		LZ 2013	3.9	0.1	0.211
		LZ 2007	4.6	0.1	0.229
	400	MM-aLASSO	4.9	0	0.0342
		MM-SCAD	4.9	0	0.0259
		LZ 2013	4.3	0	0.0399
40%	100	MM-aLASSO	4.0	0.1	0.249
		MM-SCAD	4.7	0.3	0.280
		LZ 2013	3.8	0.1	0.241
		LZ 2007	4.4	0.2	0.303
	400	MM-aLASSO	4.9	0	0.0399
		MM-SCAD	4.9	0	0.0290
		LZ 2013	4.2	0	0.0443

subjects are randomly assigned to receive a standard treatment or chemotherapy. Six covariates are measured: treatment (1=standard, 2=test), celltype (1=squamous, 2=smallcell, 3=adeno, 4=large), karno (Karnofsky performance score, where 100=good), diagtime (months from diagnosis to randomization), age (in years), and prior (prior therapy, where 0=no, 10=yes). Table 2.2 lists the estimates calculated using various methods, including our penalized profile log-likelihood method using both SCAD and adaptive LASSO penalties, the adaptive LASSO-penalized marginal likelihood method of Lu and Zhang (2007), the method of Liu and Zeng (2013) that applies adaptive LASSO to the minorizer of the log-likelihood at the MLE, and the unpenalized maximum likelihood estimator.

Table 2.2. Coefficient estimates for different methods of fitting the proportional odds model to the Veteran's Administration dataset, where LZ 2013 is the method of Liu and Zeng (2013) and LZ 2007 is the marginal likelihood method of Lu and Zhang (2007).

Covariate	MM-SCAD	MM-aLASSO	LZ 2007	LZ 2013	MLE
trt	0.000	0.000	0.000	0.000	0.177
cell.ad	1.445	1.322	0.841	0.864	1.463
cell.sm	1.281	1.179	0.706	0.821	1.347
cell.sq	0.000	0.000	0.000	0.000	0.0012
prior	0.000	0.000	0.000	0.000	0.0195
diagtime	0.0005	0.000	0.000	0.000	0.0012
age	0.000	0.000	0.000	0.000	-0.0025
karno	-0.056	-0.055	-0.053	-0.055	-0.0566

As pointed out at the end of Section 2.6, the choice of tuning parameter is highly influential. For our methods, we use a BIC-type selector, a special case of the generalized information criterion (GIC) in Zhang et al. (2010), and select tuning parameters of 0.12 for MM-SCAD and 0.016 for MM-aLASSO. Comparing the four variable selection methods, we find that the three methods that use adaptive LASSO select the same set of variables, though the estimates are slightly different, with those that use our method tending to be closer to the unpenalized MLE values. The first method, using SCAD, allows for one additional nonzero coefficient.

2.8 Discussion

This article studies the problem of variable selection for the proportional odds model through direct penalization of the (profile) log-likelihood function. We show that the maximizer of this function is consistent and has an oracle property under some regularity conditions. In particular, the estimates of the true zero coefficients converge to 0 at a faster rate than $\sqrt{n}$ and (under a slightly stronger regularity condition) are exactly zero with probability approaching one, while the estimates of the nonzero components are efficient and asymptotically normally distributed. Proofs of these results rely on an expansion of the profile log-likelihood function provided by Murphy and Van der Vaart (2000) together with some mild modified

conditions that we introduce. Our results are therefore novel in the sense that they do not depend on Taylor expansions, unlike most work on penalized likelihood functions in the literature. Furthermore, our results should be generalizable to other semi-parametric models that satisfy these regularity conditions, allowing for the possibility of penalization in other situations when goodness-of-fit measures are non-differentiable.

Ordinarily, working directly with the likelihood or profile likelihood in the proportional odds model is difficult, but we overcome this difficulty by using MM algorithms for purposes of maximization. MM algorithms operate by alternately constructing a surrogate function for the penalized likelihood function given the current estimates, then maximizing it. For the maximization step in our MM algorithms, we test several different numerical methods, including an ICM algorithm (Zhang and Li, 2009), a coordinate descent algorithm (Wu and Lange, 2008), and a Minimization by Iterative Soft Thresholding, or MIST, algorithm (Schifano et al., 2010). Our MM algorithms enjoy a fast convergence rate while maintaining satisfactory performance. The simulation results demonstrate the efficacy of our algorithms, which outperform existing algorithms particularly for larger samples. On the other hand, we find that results are quite sensitive to tuning parameter selection, which is therefore a topic for further study.

Our exploration of various optimization algorithms in the maximization step of our MM algorithms did not identify a clear-cut winner, suggesting another topic for further study. Complicating this question is the fact that different types of penalty functions may require different optimization algorithms; for instance, SCAD penalties make the profile likelihood possibly nonconcave, whereas LASSO and adaptive LASSO do not share this challenge.

Yet another topic for further research is the question of improving the asymptotic results, for instance, to allow for the number of parameters to grow with the sample size. In recent years, problems with large numbers of parameters have attracted more attention due to the development of technology. For example, Cai et al. (2005) proposed to maximize the penalized pseudo-partial likelihood function to select variables for multivariate failure time data. They assumed that the number of parameters in the model grows in a rate slower than sample size, and obtained estimates which are consistent, and asymptotically normally distributed

and also possess an oracle property in the sense of Fan and Li (2001). Yet the rate of convergence in $\boldsymbol{\beta}$ is no longer $\sqrt{n}$, but involves the increasing rate of the number of parameters in the model. Such issues might also be of interest in our setting of the proportional odds model, yet for the present, it appears that adapting the asymptotic framework we use here to the case of increasing numbers of parameters is a nontrivial modification.

All in all, we believe that this article's combination of theoretical results for penalization of a non-differentiable objective function, together with techniques like MM for aiding the numerical optimization of complicated objective functions, has the potential to open new avenues for the use of penalization in the context of model selection.

Appendix A: Detailed Simulation Output

Table 2.3 gives detailed results of our simulation studies in Section 2.6. Unlike in Table 2.1, the results using the coordinate descent (CD) algorithm of Wu and Lange (2008) are separated from those that use the ICM algorithm of Zhang and Li (2009). Furthermore, one additional sample size ($n = 200$) is included. All results of Zhang and Li (2009) are based on code provided by the authors. Table 2.4 lists the proportion of times that each of the eight variables is included in the final model, though this information is not available for the method of Lu and Zhang (2007).

Table 2.3. Results for 1000 repetitions of the simulation study, where LZ 2013 is the method of Liu and Zeng (2013). The "Correct" and "Incorrect" columns give the number of parameters correctly and incorrectly set to zero. The mean squared error is given by $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \Sigma (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

censoring	n	Method	Correct(5)	Incorrect(0)	Median MSE
25%	100	CD-aLASSO	4.098	0.067	0.2087
		ICM-aLASSO	4.108	0.069	0.2062
		CD-SCAD	4.733	0.265	0.1738
		ICM-SCAD	4.738	0.264	0.1728
		LZ 2013	3.956	0.052	0.2113
	200	CD-aLASSO	4.691	0.005	0.08389
		ICM-aLASSO	4.698	0.005	0.08452
		CD-SCAD	4.881	0.012	0.05939
		ICM-SCAD	4.883	0.011	0.06008
		LZ 2013	4.157	0	0.08541
	400	CD-aLASSO	4.875	0	0.03437
		ICM-aLASSO	4.879	0	0.03405
		CD-SCAD	4.931	0	0.02584
		ICM-SCAD	4.931	0	0.02596
		LZ 2013	4.252	0	0.03994
40%	100	CD-aLASSO	4.016	0.102	0.2495
		ICM-aLASSO	4.039	0.106	0.2493
		CD-SCAD	4.707	0.363	0.2850
		ICM-SCAD	4.702	0.353	0.2759
		LZ 2013	3.839	0.075	0.2414
	200	CD-aLASSO	4.677	0.014	0.09656
		ICM-aLASSO	4.686	0.014	0.09490
		CD-SCAD	4.888	0.030	0.06808
		ICM-SCAD	4.874	0.028	0.06914
		LZ 2013	4.041	0.002	0.1022
	400	CD-aLASSO	4.854	0	0.04008
		ICM-aLASSO	4.856	0	0.03974
		CD-SCAD	4.933	0	0.02917
		ICM-SCAD	4.933	0	0.02873
		LZ 2013	4.162	0	0.04428

Table 2.4. Results for 1000 repetitions of the simulation study, where LZ 2013 is the method of Liu and Zeng (2013). The proportions give the fraction of time that each of the eight variables was included in the model; variables 1, 4, and 7, shown in bold, are in the true model, whereas the others are absent.

censoring	n	Method	Proportions
25%	100	CD-aLASSO	**0.974** 0.181 0.180 **0.984** 0.184 0.174 **0.975** 0.183
		ICM-aLASSO	**0.974** 0.182 0.176 **0.983** 0.180 0.171 **0.974** 0.183
		CD-SCAD	**0.914** 0.056 0.050 **0.918** 0.050 0.054 **0.903** 0.057
		ICM-SCAD	**0.911** 0.059 0.046 **0.918** 0.048 0.049 **0.907** 0.060
		LZ 2013	**0.980** 0.215 0.206 **0.989** 0.206 0.202 **0.979** 0.215
	200	CD-aLASSO	**0.996** 0.058 0.046 **1.000** 0.071 0.072 **0.999** 0.062
		ICM-aLASSO	**0.996** 0.054 0.045 **1.000** 0.069 0.068 **0.999** 0.066
		CD-SCAD	**0.995** 0.020 0.012 **0.997** 0.031 0.032 **0.996** 0.024
		ICM-SCAD	**0.995** 0.021 0.012 **0.997** 0.031 0.034 **0.996** 0.026
		LZ 2013	**1.000** 0.168 0.162 **1.000** 0.170 0.174 **1.000** 0.169
	400	CD-aLASSO	**1.000** 0.022 0.025 **1.000** 0.021 0.031 **1.000** 0.026
		ICM-aLASSO	**1.000** 0.022 0.025 **1.000** 0.018 0.031 **1.000** 0.025
		CD-SCAD	**1.000** 0.013 0.018 **1.000** 0.011 0.017 **1.000** 0.010
		ICM-SCAD	**1.000** 0.013 0.018 **1.000** 0.011 0.017 **1.000** 0.010
		LZ 2013	**1.000** 0.023 0.029 **1.000** 0.026 0.028 **1.000** 0.025
40%	100	CD-aLASSO	**0.963** 0.211 0.182 **0.972** 0.195 0.189 **0.963** 0.207
		ICM-aLASSO	**0.963** 0.208 0.177 **0.971** 0.190 0.187 **0.960** 0.199
		CD-SCAD	**0.871** 0.065 0.063 **0.890** 0.052 0.059 **0.876** 0.054
		ICM-SCAD	**0.876** 0.064 0.064 **0.893** 0.052 0.062 **0.878** 0.056
		LZ 2013	**0.972** 0.249 0.220 **0.981** 0.232 0.221 **0.972** 0.239
	200	CD-aLASSO	**0.990** 0.061 0.043 **0.998** 0.070 0.075 **0.998** 0.065
		ICM-aLASSO	**0.990** 0.062 0.045 **0.998** 0.070 0.078 **0.998** 0.068
		CD-SCAD	**0.986** 0.024 0.014 **0.995** 0.023 0.032 **0.989** 0.019
		ICM-SCAD	**0.985** 0.023 0.017 **0.995** 0.023 0.031 **0.990** 0.018
		LZ 2013	**0.998** 0.182 0.185 **1.000** 0.193 0.210 **1.000** 0.189
	400	CD-aLASSO	**1.000** 0.030 0.029 **1.000** 0.024 0.039 **1.000** 0.024
		ICM-aLASSO	**1.000** 0.027 0.027 **1.000** 0.024 0.035 **1.000** 0.022
		CD-SCAD	**1.000** 0.014 0.015 **1.000** 0.008 0.018 **1.000** 0.012
		ICM-SCAD	**1.000** 0.015 0.017 **1.000** 0.007 0.017 **1.000** 0.011
		LZ 2013	**1.000** 0.167 0.169 **1.000** 0.155 0.175 **1.000** 0.172

Appendix B: Proofs of Theoretical Results

Recall that the penalized profile likelihood function is

$$\xi^{\text{prof}}(\boldsymbol{\beta}) = p\ell(\boldsymbol{\beta}, x) - n\sum_{j=1}^{n} p_{\lambda_n}(|\beta_j|)$$

and

$$a_n = \max_{1\le j\le p}\{p'_{\lambda_n}(|\beta_j^0|) : \beta_j^0 \neq 0\},$$

$$b_n = \max_{1\le j\le p}\{p''_{\lambda_n}(|\beta_j^0|) : \beta_j^0 \neq 0\}.$$

Proof of theorem 1: Let $\alpha_n = n^{-1/2} + a_n$. It is sufficient to show that for any $\varepsilon > 0$, there exist positive constants C and N such that

$$P\left(\sup_{\|\mathbf{u}\|=C} \xi^{\text{prof}}(\alpha_n\mathbf{u} + \boldsymbol{\beta}^0) < \xi^{\text{prof}}(\boldsymbol{\beta}^0)\right) \ge 1-\varepsilon \text{ for all } n > N.$$

Let $D_n(\boldsymbol{u}) = \xi^{\text{prof}}(\alpha_n\boldsymbol{u} + \boldsymbol{\beta}^0) - \xi^{\text{prof}}(\boldsymbol{\beta}^0)$. Then

$$\begin{aligned} D_n(\boldsymbol{u}) &= p\ell(\boldsymbol{\beta}^0 + \alpha_n\mathbf{u}) - p\ell(\boldsymbol{\beta}^0) - n\sum_{j=1}^{p} p_{\lambda_n}(|\beta_j^0 + \alpha_n u_j|) + n\sum_{j=1}^{p} p_{\lambda_n}(|\beta_j^0|) \\ &\le \underbrace{p\ell(\boldsymbol{\beta}^0 + \alpha_n\mathbf{u}) - p\ell(\boldsymbol{\beta}^0)}_{\text{(I)}} \underbrace{- n\sum_{j=1}^{s}\left[p_{\lambda_n}(|\beta_j^0 + \alpha_n u_j|) - p_{\lambda_n}(|\beta_j^0|)\right]}_{\text{(II)}}. \end{aligned}$$

We consider (I) and (II) separately. Theorem 1 in Murphy and Van der Vaart (2000) shows that for any sequence $\tilde{\boldsymbol{\beta}}_n \xrightarrow{P} \boldsymbol{\beta}^0$, we have the expansion

$$\begin{aligned} p\ell(\tilde{\boldsymbol{\beta}}) &= p\ell(\boldsymbol{\beta}^0) + (\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0)^\top \sum_{i=1}^{n} \tilde{\ell}_0(\mathbf{x}_i) - \frac{n}{2}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0)^\top \tilde{I}_0(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0) \\ &\quad + o_P(\sqrt{n}\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0\| + 1)^2, \end{aligned} \tag{2.22}$$

where $\tilde{\ell}_0(x)$ and $\tilde{I}_0$ are called the effective (or efficient) score vector and information matrix, respectively. The effective information is the covariance matrix of the effective score; more details on the effective score may be found in Murphy and

Van Der Vaart (1999).

Writing $\tilde{\boldsymbol{\beta}}_n = \boldsymbol{\beta}^0 + \alpha_n \boldsymbol{u}$, we obtain

$$\begin{aligned}
\text{(I)} &= \alpha_n \boldsymbol{u}^\top \sum_{i=1}^{n} \tilde{\ell}_0(\mathbf{x}_i) - \frac{n\alpha_n^2}{2} \boldsymbol{u}^\top \tilde{I}_0 \boldsymbol{u} + o_P(\sqrt{n}\alpha_n \|\boldsymbol{u}\| + 1)^2. \\
&\leq \kappa_n \left[\frac{1}{\sqrt{n}} \frac{\boldsymbol{u}^\top}{\|\boldsymbol{u}\|} \sum_{i=1}^{n} \tilde{\ell}_0(\mathbf{x}_i) - \frac{\kappa_n \lambda_{\min}}{2} + o_P\left(\kappa_n + \frac{1}{\kappa_n}\right) \right], \qquad (2.23)
\end{aligned}$$

where $\kappa_n = \sqrt{n}\alpha_n \|\boldsymbol{u}\|$ and $\lambda_{\min}$ is the smallest eigenvalue of $\tilde{I}_0$.

We now argue that C may be chosen large enough so that for any $\|\boldsymbol{u}\| = C$, the right side of (3.24) is strictly negative with probability close to one. First, the $\tilde{\ell}_0(\mathbf{x}_i)$ are independent and identically distributed with mean 0, so the Central Limit Theorem implies that

$$\frac{1}{\sqrt{n}} \frac{\boldsymbol{u}^\top}{\|\boldsymbol{u}\|} \sum_{i=1}^{n} \tilde{\ell}_0(\mathbf{x}_i)$$

is bounded in probability. Furthermore, since $\sqrt{n}\alpha_n > 1$, we get $\kappa_n > 1$ whenever $C > 1$. Therefore, $C > 1$ implies that $o_P(\kappa_n + \kappa_n^{-1}) = o_P(\kappa_n)$. Since $\tilde{I}_0$ is positive definite, $\lambda_{\min} > 0$, and we conclude that $-\kappa_n \lambda_{\min}/2 + o_P(\kappa_n)$ can be made strictly negative for large n with probability arbitrarily close to one by choosing C large enough.

Next, consider

$$\begin{aligned}
\text{(II)} &= n \sum_{j=1}^{s} \left| \alpha_n p'_{\lambda_n}(|\beta_j^0|) u_j + \alpha_n^2 p''_{\lambda_n}(|\beta_j^0|) u_j^2 (1 + o_p(1)) \right| \\
&\leq \sqrt{s} n \alpha_n a_n \|\mathbf{u}\| + n\alpha_n^2 b_n \|\mathbf{u}\|^2 (1 + o_p(1)) \\
&= n\alpha_n^2 (\sqrt{s} \cdot \frac{a_n}{\alpha_n} \|\mathbf{u}\|) + n\alpha_n^2 \|\mathbf{u}\|^2 b_n (1 + o_p(1)) \\
&= n\alpha_n^2 \left[\sqrt{s} \cdot \frac{a_n}{\alpha_n} \|\mathbf{u}\| + \|\mathbf{u}\|^2 b_n + o_p(1) \right].
\end{aligned}$$

Since $a_n \to 0$ and $b_n \to 0$, (II) is also dominated by the second term in (I). So $D_n(\boldsymbol{u}) < 0$, which concludes the proof. ■

Lemma A Assume that regularity conditions 1, 2, and 3 are satisfied and that $\hat{\boldsymbol{\beta}}^{1n}$ denotes the first s components of the $\sqrt{n}-$consistent estimator obtained in

Theorem 1. Then with probability tending to 1,

$$\hat{\boldsymbol{\beta}}^{2n} = \arg \max_{\|\boldsymbol{\beta}^{2n}\| \leq cn^{-1/2}} \xi^{\text{prof}}\{(\hat{\boldsymbol{\beta}}^{1n}, \boldsymbol{\beta}^{2n})\} = o_P(n^{-1+\delta}). \tag{2.24}$$

Proof of Lemma A: To prove the above result, we only need to prove that for any $\boldsymbol{\beta}^{1n}$ satisfying

$$\|\boldsymbol{\beta}^{1n} - \boldsymbol{\beta}^{10}\| = O_P(n^{-1/2}),$$

with probability tending to 1 as $n \to \infty$, for any $c > 0$,

$$\xi^{\text{prof}}\{(\boldsymbol{\beta}^{1n}, \mathbf{0})\} > \sup_{cn^{-1+\delta} \leq \|\boldsymbol{\beta}^{2n}\| \leq cn^{-1/2}} \xi^{\text{prof}}\{(\boldsymbol{\beta}^{1n}, \boldsymbol{\beta}^{2n})\}. \tag{2.25}$$

To see why inequality (2.25) implies equation (3.37): $\hat{\boldsymbol{\beta}}^{1n}$ is $\sqrt{n}-$consistent, so if it satisfies equation (2.25) for any $c > 0$,

$$\begin{aligned}
& P\left(\xi^{\text{prof}}\{(\hat{\boldsymbol{\beta}}^{1n}, \mathbf{0})\} > \sup_{cn^{-1+\delta} \leq \|\boldsymbol{\beta}^{2n}\| \leq cn^{-1/2}} \xi^{\text{prof}}\{(\hat{\boldsymbol{\beta}}^{1n}, \boldsymbol{\beta}^{2n})\}\right) \to 1 \\
\Rightarrow \quad & P(\|\arg \max_{\|\boldsymbol{\beta}^{2n}\| \leq cn^{-1/2}} \xi^{\text{prof}}\{(\hat{\boldsymbol{\beta}}^{1n}, \boldsymbol{\beta}^{2n})\}\| < cn^{-1+\delta}) \to 1 \\
\Rightarrow \quad & P\left(\|\hat{\boldsymbol{\beta}}^{2n}\| < cn^{-1+\delta}\right) \to 1 \\
\Rightarrow \quad & P\left(n^{1-\delta}\|\hat{\boldsymbol{\beta}}^{2n}\| < c\right) \to 1 \\
\Rightarrow \quad & \hat{\boldsymbol{\beta}}^{2n} = o_P(n^{-1+\delta}).
\end{aligned}$$

Now let us prove (2.25). Let $\tilde{\ell}_{10}(\boldsymbol{X})$ and $\tilde{\ell}_{20}(\boldsymbol{X})$ denote the first s and last r dimensions of $\tilde{\ell}_0(\boldsymbol{X})$, respectively. Furthermore, let

$$\tilde{I}_0 = \begin{pmatrix} \tilde{I}^0_{11} & \tilde{I}^0_{12} \\ \tilde{I}^0_{21} & \tilde{I}^0_{22} \end{pmatrix}.$$

The above inequality is equivalent to

$$\sup_{cn^{-1+\delta} \leq \|\boldsymbol{\beta}^{2n}\| \leq cn^{-1/2}} \xi^{\text{prof}}\{(\boldsymbol{\beta}^{1n}, \boldsymbol{\beta}^{2n})\} - \xi^{\text{prof}}\{(\boldsymbol{\beta}^{1n}, \mathbf{0})\} < 0. \tag{2.26}$$

According to the expansion in equation (2.22)

$$\begin{aligned}
& \xi^{\text{prof}}\{(\boldsymbol{\beta}^{1n}, \boldsymbol{\beta}^{2n})\} - \xi^{\text{prof}}\{(\boldsymbol{\beta}^{1n}, \mathbf{0})\} \\
= \;& p\ell(\boldsymbol{\beta}^{1n}, \boldsymbol{\beta}^{2n}) - n\sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|) - p\ell(\boldsymbol{\beta}^{1n}, \mathbf{0}) + n\sum_{j=1}^{s} p_{\lambda_n}(|\beta_j|) \\
= \;& p\ell(\boldsymbol{\beta}^{1n}, \boldsymbol{\beta}^{2n}) - p\ell(\boldsymbol{\beta}^{1n}, 0) - n\sum_{j=s+1}^{p} p_{\lambda_n}(|\beta_j|) \\
= \;& (\boldsymbol{\beta}^{2n})^\top \sum_{i=1}^{n} \tilde{\ell}_{20}(\mathbf{x}_i) - \frac{1}{2} n(\boldsymbol{\beta}^{2n})^\top \tilde{I}_{22}^0 \boldsymbol{\beta}^{2n} + o_P(\sqrt{n}\|\boldsymbol{\beta}^{2n}\| + 1)^2 - n\sum_{j=s+1}^{p} p_{\lambda_n}(|\beta_j|).
\end{aligned} \tag{2.27}$$

Since $\boldsymbol{\beta}^{2n} = O_P(\frac{1}{\sqrt{n}})$ by Theorem 1 and $\sum_{j=1}^{n} \tilde{\ell}_{20}(\mathbf{x}_i) = O_P(\sqrt{n})$,

$$\begin{aligned}
& (\boldsymbol{\beta}^{2n})^\top \sum_{i=1}^{n} \tilde{\ell}_{20}(\mathbf{x}_i) + o_P(\sqrt{n}\|\boldsymbol{\beta}^{2n}\| + 1)^2 \\
= \;& n\|\boldsymbol{\beta}^{2n}\| O_P(\frac{1}{\sqrt{n}}) + o_P(1) \\
= \;& n\|\boldsymbol{\beta}^{2n}\| o_P(\frac{1}{n^\delta}).
\end{aligned} \tag{2.28}$$

The last equation is because $\|\boldsymbol{\beta}^{2n}\| \geq cn^{-1+\delta} \Rightarrow \frac{1}{n\|\boldsymbol{\beta}^{2n}\|} \leq \frac{1}{cn^\delta}$, so

$$o_P(1) = n\|\boldsymbol{\beta}^{2n}\| o_P\left(\frac{1}{n\|\boldsymbol{\beta}^{2n}\|}\right) = n\|\boldsymbol{\beta}^{2n}\| o_P\left(\frac{1}{cn^\delta}\right) = n\|\boldsymbol{\beta}^{2n}\| o_P\left(\frac{1}{n^\delta}\right).$$

So the left hand side of equation (2.26) becomes

$$LHS \leq \sup_{cn^{-1+\delta} \leq \|\boldsymbol{\beta}^{2n}\| \leq cn^{-1/2}} n\|\boldsymbol{\beta}^{2n}\| o_P(\frac{1}{n^\delta}) - n\sum_{j=s+1}^{p} p_{\lambda_n}(|\beta_j|). \tag{2.29}$$

For the penalty term, by condition 2, for any $x > 0, \lambda_n > 0$, there exists $0 < x^* < x$ such that

$$p_{\lambda_n}(x) \;=\; x p'_{\lambda_n}(0+) + \frac{x^2}{2} p''_{\lambda_n}(x^*)$$

$$\geq \quad xp'_{\lambda_n}(0+) + \frac{x^2}{2}M. \tag{2.30}$$

Therefore,

$$\begin{aligned}
\sup_{cn^{-1+\delta}\leq\|\boldsymbol{\beta}^{2n}\|\leq cn^{-1/2}} n\sum_{j=s+1}^{p} p_{\lambda_n}(|\beta_j|) &\geq \sup_{cn^{-1+\delta}\leq\|\boldsymbol{\beta}^{2n}\|\leq cn^{-1/2}} n\sum_{j=s+1}^{p}\left\{p'_{\lambda_n}(0+)|\beta_j| + \|\beta_j\|^2 M\right\} \\
&= \sup_{cn^{-1+\delta}\leq\|\boldsymbol{\beta}^{2n}\|\leq cn^{-1/2}} np'_{\lambda_n}(0+)\sum_{j=s+1}^{p}|\beta_j| + n\|\boldsymbol{\beta}^{2n}\|^2 M \\
&\geq \sup_{cn^{-1+\delta}\leq\|\boldsymbol{\beta}^{2n}\|\leq cn^{-1/2}} np'_{\lambda_n}(0+)\|\boldsymbol{\beta}^{2n}\|^2
\end{aligned}$$

So now as with equation (8), we get

$$\begin{aligned}
&\sup_{cn^{-1+\delta}\leq\|\boldsymbol{\beta}^{2n}\|\leq cn^{-1/2}} n\|\boldsymbol{\beta}^{2n}\| o_P(\frac{1}{n^\delta}) - n\sum_{j=s+1}^{p} p_{\lambda_n}(|\beta_j|) \\
\leq &\sup_{cn^{-1+\delta}\leq\|\boldsymbol{\beta}^{2n}\|\leq cn^{-1/2}} n\|\boldsymbol{\beta}^{2n}\|\left[o_P(\frac{1}{n^\delta}) - p'_{\lambda_n}(0+)\right] \\
= &\sup_{cn^{-1+\delta}\leq\|\boldsymbol{\beta}^{2n}\|\leq cn^{-1/2}} n\lambda_n\|\boldsymbol{\beta}^{2n}\|\left[o_P(\frac{1}{n^\delta\lambda_n}) - p'_{\lambda_n}(0+)/\lambda_n\right] < 0
\end{aligned}$$

with probability going to 1. ■

Lemma B Assume regularity condition 2 holds. In addition, we assume that there exits $0 < \delta^* \leq 1$ such that for any random sequence $\tilde{\boldsymbol{\theta}}^n = (\tilde{\boldsymbol{\theta}}^{1n}, \tilde{\boldsymbol{\theta}}^{2n}) \to \boldsymbol{\theta}^0 = (\boldsymbol{\theta}^{10}, \mathbf{0})$,

$$\begin{aligned}
p\ell_n(\tilde{\boldsymbol{\theta}}^n) &= p\ell_n(\boldsymbol{\theta}^0) + (\tilde{\boldsymbol{\theta}}^n - \boldsymbol{\theta}^0)^\top \sum_{i=1}^{n}\tilde{\ell}_0(\mathbf{x}_i) - \frac{1}{2}n(\tilde{\boldsymbol{\theta}}^n - \boldsymbol{\theta}^0)^\top \tilde{I}_0(\tilde{\boldsymbol{\theta}}^n - \boldsymbol{\theta}^0) \\
&\quad + o_P(n\|\tilde{\boldsymbol{\theta}}^{1n} - \boldsymbol{\theta}^{10}\|^2) + o_P(n\|\tilde{\boldsymbol{\theta}}^{1n} - \boldsymbol{\theta}^{10}\| \cdot \|\tilde{\boldsymbol{\theta}}^{2n}\|) + o_P(n\|\tilde{\boldsymbol{\theta}}^{2n}\|^{1+\delta^*}),
\end{aligned} \tag{2.31}$$

in which the first reminder term only depends on $\|\tilde{\boldsymbol{\theta}}^{1n} - \boldsymbol{\theta}^{10}\|$. Now take λ_n to be a sequence satisfying regularity condition 2 with $\delta = \delta^*/(1+\delta^*)$. Then for any

$\boldsymbol{\beta}^{1n}$ satisfying $\|\boldsymbol{\beta}^{1n} - \boldsymbol{\beta}^{10}\| = O_P(n^{-1/2})$ and any constant $c > 0$,

$$P\left(\xi^{\text{prof}}\{(\boldsymbol{\beta}^{1n}, \mathbf{0})\} = \max_{\|\boldsymbol{\beta}^{2n}\| \leq cn^{-1/2}} \xi^{\text{prof}}\{(\boldsymbol{\beta}^{1n}, \boldsymbol{\beta}^{2n})\}\right) \to 1$$

as $n \to +\infty$.

Proof of Lemma B: We only need to prove that for any $(\boldsymbol{\beta}^{1n}, \boldsymbol{\beta}^{2n})$ satisfying

$$\boldsymbol{\beta}^{1n} - \boldsymbol{\beta}^{10} = O_P(n^{-1/2}), \|\boldsymbol{\beta}^{2n}\| \leq cn^{-1/2}, \boldsymbol{\beta}^{2n} \neq 0$$

with probability tending to 1 as $n \to \infty$,

$$\xi^{\text{prof}}\{(\boldsymbol{\beta}^{1n}, \boldsymbol{\beta}^{2n})\} - \xi^{\text{prof}}\{(\boldsymbol{\beta}^{1n}, \mathbf{0})\} < 0. \tag{2.32}$$

Replace (2.22) with our slightly stronger condition (2.31). The same argument as in Lemma A shows that the left hand side of (2.32) is bounded above by

$$\begin{aligned}
&(\boldsymbol{\beta}^{2n})^\top \sum_{i=1}^{n} \tilde{\ell}_{20}(\mathbf{x}_i) + o_P(n\|\boldsymbol{\beta}^{2n}\|^{1+\delta^*}) + o_P(n\|\boldsymbol{\beta}^{1n} - \boldsymbol{\beta}^{10}\| \cdot \|\boldsymbol{\beta}^{2n}\|) \\
&-n \sum_{j=s+1}^{p} p_{\lambda_n}(|\beta_j|).
\end{aligned} \tag{2.33}$$

We may apply the same discussion for the first and last terms as in Lemma A, and also for any $\sqrt{n}-$consistent $\tilde{\boldsymbol{\beta}}^{1n}$, to obtain

$$\begin{aligned}
& o_P(n\|\boldsymbol{\beta}^{2n}\|^{1+\delta^*}) + o_P(n\|\boldsymbol{\beta}^{1n} - \boldsymbol{\beta}^{10}\| \cdot \|\boldsymbol{\beta}^{2n}\|) \\
= \;& n\|\boldsymbol{\beta}^{2n}\| o_P(\|\boldsymbol{\beta}^{2n}\|^{\delta/(1-\delta)}) + n\|\boldsymbol{\beta}^{2n}\| o_P(1/\sqrt{n}).
\end{aligned} \tag{2.34}$$

So now expression (2.33) becomes

$$\begin{aligned}
& n\|\boldsymbol{\beta}^{2n}\| \left[O_P(\frac{1}{\sqrt{n}}) + o_P(\|\boldsymbol{\beta}^{2n}\|^{\delta/(1-\delta)}) - p'_{\lambda_n}(0+)\right] \\
= \;& n\lambda_n\|\boldsymbol{\beta}^{2n}\| \left[o_P(\frac{1}{n^\delta \lambda_n}) + o_P(\|\boldsymbol{\beta}^{2n}\|^{\delta/(1-\delta)}/\lambda_n) - p'_{\lambda_n}(0+)/\lambda_n\right].
\end{aligned} \tag{2.35}$$

In Lemma A we already proved that $\hat{\boldsymbol{\beta}}^{2n} = o_P(n^{-1+\delta})$; here we can limit $\|\boldsymbol{\beta}^{2n}\| = o_P(n^{-1+\delta})$. Therefore, it can be easily verified that the last term still dominates

the remaining terms. So with probability going to 1, the desired inequality holds. ■

Proof of Theorem 2:

We first derive the following result for penalty function:

$$\begin{aligned}\xi^{\text{prof}}\{\hat{\boldsymbol{\beta}}\} &= p\ell(\hat{\boldsymbol{\beta}}) - n\sum_{j=1}^{p} p(|\hat{\beta}_j|) \\ &= p\ell(\hat{\boldsymbol{\beta}}) - n\sum_{j=1}^{s} p(|\hat{\beta}_j|) + o_P(1).\end{aligned}$$

The second equality holds because for δ specified for λ, we already proved that $\hat{\boldsymbol{\beta}}^{2n} = o_P(n^{-1+\delta})$ in Lemma A. Since the SCAD penalty function is a linear function with slope λ_n at a neighborhood of zero,

$$\begin{aligned}n\sum_{j=s+1}^{p} p(|\hat{\beta}_j|) &\leq n\lambda_n \max_{s+1\leq j\leq p} |\hat{\beta}_j| \cdot r \\ &\leq n\lambda_n o_P(n^{-1+\delta}) \\ &= o_P(n^{\delta}\lambda_n) = o_P(1).\end{aligned}$$

Therefore, the expansion of $\xi^{\text{prof}}\{\hat{\boldsymbol{\beta}}\}$ is as follows:

$$\begin{aligned}\xi^{\text{prof}}\{\hat{\boldsymbol{\beta}}\} &= p\ell(\boldsymbol{\beta}^0) \\ &\quad +(\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10})^\top \left[\sum_{i=1}^{n} \tilde{\ell}_{10}(\mathbf{x}_i) - n\boldsymbol{d}_n\right] + (\hat{\boldsymbol{\beta}}^{2n} - \boldsymbol{\beta}^{20})^\top \sum_{i=1}^{n} \tilde{\ell}_{20}(\mathbf{x}_i) \\ &\quad -\frac{n}{2}(\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10})^\top \left[\tilde{I}^0_{11} + \Sigma_n\right](\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10}) - \frac{n}{2}(\hat{\boldsymbol{\beta}}^{2n} - \boldsymbol{\beta}^{20})^\top \tilde{I}^0_{22}(\hat{\boldsymbol{\beta}}^{2n} - \boldsymbol{\beta}^{20}) \\ &\quad -\frac{n}{2}(\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10})^\top \tilde{I}^0_{12}(\hat{\boldsymbol{\beta}}^{2n} - \boldsymbol{\beta}^{20}) - \frac{n}{2}(\hat{\boldsymbol{\beta}}^{2n} - \boldsymbol{\beta}^{20})^\top \tilde{I}^0_{21}(\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10}) \\ &\quad +o_P(\sqrt{n}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| + 1)^2 + o_P(\sqrt{n}\|\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^0_1\|)^2, \qquad (2.36)\end{aligned}$$

where we recall that $\tilde{\ell}_{10}(\boldsymbol{X})$ and $\tilde{\ell}_{20}(\boldsymbol{X})$ denote the first s and last r dimensions of $\tilde{\ell}_0(\boldsymbol{X})$ separately and

$$\tilde{I}_0 = \begin{pmatrix} \tilde{I}^0_{11} & \tilde{I}^0_{12} \\ \tilde{I}^0_{21} & \tilde{I}^0_{22} \end{pmatrix}.$$

Since $\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10} = O_P(n^{-1/2})$ and $\hat{\boldsymbol{\beta}}^{2n} - \boldsymbol{\beta}^{20} = \hat{\boldsymbol{\beta}}^{2n} = o_P(n^{-1+\delta})$ for some $0 < \delta < 1/2$, we can rewrite Equation (2.36) only in terms of $\hat{\boldsymbol{\beta}}^{1n}$ and $\boldsymbol{\beta}^{10}$.

- Since $\sum_{j=1}^{n} \tilde{\ell}_{20}(\mathbf{x}_i) = O_P(\sqrt{n})$, we get

$$(\hat{\boldsymbol{\beta}}^{2n} - \boldsymbol{\beta}^{20})^\top \sum_{i=1}^{n} \tilde{\ell}_{20}(\mathbf{x}_i) = o_P(n^{-1+\delta+1/2}) = o_P(1).$$

- Since $\tilde{I}^0_{22} = O_P(1)$,

$$-\frac{n}{2}(\hat{\boldsymbol{\beta}}^{2n} - \boldsymbol{\beta}^{20})^\top \tilde{I}^0_{22}(\hat{\boldsymbol{\beta}}^{2n} - \boldsymbol{\beta}^{20}) = o_P(n^{-2+2\delta+1}) = o_P(n^{-1+2\delta}) = o_P(1).$$

- Since $\tilde{I}^0_{12} = (\tilde{I}^0_{21})^\top = O_P(1)$,

$$\begin{aligned} -\frac{n}{2}(\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10})^\top \tilde{I}^0_{12}(\hat{\boldsymbol{\beta}}^{2n} - \boldsymbol{\beta}^{20}) &= -\frac{n}{2}(\hat{\boldsymbol{\beta}}^{2n} - \boldsymbol{\beta}^{20})^\top \tilde{I}^0_{21}(\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10}) \\ &= o_P(n^{1-1+\delta-1/2}) \\ &= o_P(n^{-1/2+\delta}) = o_P(1). \end{aligned} \tag{2.37}$$

Therefore,

$$\begin{aligned} \xi^{\text{prof}}\{\hat{\boldsymbol{\beta}}\} &= p\ell(\boldsymbol{\beta}^0) + (\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10})^\top \left[\sum_{i=1}^{n} \tilde{\ell}_{10}(\mathbf{x}_i) - n\boldsymbol{d}_n\right] \\ &\quad -\frac{n}{2}(\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10})^\top \left[\tilde{I}^0_{11} + \Sigma_n\right](\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10}) \\ &\quad +o_P(\sqrt{n}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| + 1)^2 + o_P(\sqrt{n}\|\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^0_1\|)^2. \end{aligned}$$

Actually, the term $o_P(\sqrt{n}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| + 1)^2 = o_P(\sqrt{n}\|\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^0_1\| + 1)^2$. This is because $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|$ is the same order as $\|\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^0_1\| + \|\hat{\boldsymbol{\beta}}^{2n} - \boldsymbol{\beta}^0_2\|$ (which can be verified using the inequality $(a+b)/\sqrt{2} \leq \sqrt{a^2+b^2} \leq a+b$ for all positive a, b) and $\|\hat{\boldsymbol{\beta}}^{2n} - \boldsymbol{\beta}^0_2\| = o_P(n^{-1+\delta})$. After these simplifications, we get

$$\begin{aligned} \xi^{\text{prof}}\{\hat{\boldsymbol{\beta}}\} &= p\ell(\boldsymbol{\beta}^0) + (\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10})^\top \left[\sum_{i=1}^{n} \tilde{\ell}_{10}(\mathbf{x}_i) - n\boldsymbol{d}_n\right] \\ &\quad -\frac{n}{2}(\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10})^\top \left[\tilde{I}^0_{11} + \Sigma_n\right](\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10}) \end{aligned}$$

$$+o_P(\sqrt{n}\|\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}_1^0\| + 1)^2. \tag{2.38}$$

On the other hand, let $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}^{1n}, \mathbf{0})$, giving

$$\tilde{\boldsymbol{\beta}}^{1n} = \boldsymbol{\beta}^{10} + \frac{1}{n}(\tilde{I}_{11}^0 + \Sigma_n)^{-1} \left(\sum_{i=1}^n \tilde{\ell}_{10}(\mathbf{x}_i) - n\boldsymbol{d}_n \right).$$

Since $\tilde{\boldsymbol{\beta}}^{1n} \xrightarrow{P} \boldsymbol{\beta}^{10}$, by the same procedure as above, we get

$$\begin{aligned}
\xi^{\text{prof}}\{\tilde{\boldsymbol{\beta}}\} &= p\ell(\boldsymbol{\beta}^{\mathbf{0}}) + \frac{1}{n}\left[\sum_{i=1}^n \tilde{\ell}_{10}(\mathbf{x}_i) - n\boldsymbol{d}_n\right]^\top (\tilde{I}_{11}^0 + \Sigma_n)^{-1} \left[\sum_{i=1}^n \tilde{\ell}_{10}(\mathbf{x}_i) - n\boldsymbol{d}_n\right] \\
&\quad -\frac{1}{2n}\left[\sum_{i=1}^n \tilde{\ell}_{10}(\mathbf{x}_i) - n\boldsymbol{d}_n\right]^\top (\tilde{I}_{11}^0 + \Sigma_n)^{-1} \left[\sum_{i=1}^n \tilde{\ell}_{10}(\mathbf{x}_i) - n\boldsymbol{d}_n\right] \\
&\quad +o_P(\sqrt{n}\|\tilde{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10}\| + 1)^2 \\
&= p\ell(\boldsymbol{\beta}^{\mathbf{0}}) + \frac{1}{2n}\left[\sum_{i=1}^n \tilde{\ell}_{10}(\mathbf{x}_i) - n\boldsymbol{d}_n\right]^\top (\tilde{I}_{11}^0 + \Sigma_n)^{-1} \left[\sum_{i=1}^n \tilde{\ell}_{10}(\mathbf{x}_i) - n\boldsymbol{d}_n\right] \\
&\quad +o_P(\sqrt{n}\|\tilde{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10}\| + 1)^2.
\end{aligned} \tag{2.39}$$

Since

$$\sqrt{n}(\tilde{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10}) = (\tilde{I}_{11}^0 + \Sigma_n)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\ell}_{10}(\mathbf{x}_i) - \sqrt{n}\boldsymbol{d}_n \right),$$

$n^{-1/2} \sum_{i=1}^n \tilde{\ell}_{10}(\mathbf{x}_i) = O_P(1)$, and $\sqrt{n}\boldsymbol{d}_n \to 0$ as $n \to \infty$,

$$\begin{aligned}
\xi^{\text{prof}}\{\tilde{\boldsymbol{\beta}}\} &= p\ell(\boldsymbol{\beta}^{\mathbf{0}}) \\
&\quad +\frac{1}{2n}\left[\sum_{i=1}^n \tilde{\ell}_{10}(\mathbf{x}_i) - n\boldsymbol{d}_n\right]^\top (\tilde{I}_{11}^0 + \Sigma_n)^{-1} \left[\sum_{i=1}^n \tilde{\ell}_{10}(\mathbf{x}_i) - n\boldsymbol{d}_n\right] \\
&\quad +o_P(1).
\end{aligned}$$

To simply further, denote

$$\Delta_n = \frac{1}{\sqrt{n}} \left(\sum_{i=1}^n \tilde{\ell}_{10}(\mathbf{x}_i) - n\boldsymbol{d}_n \right), \quad \hat{h} = \sqrt{n}(\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10}).$$

We get

$$\begin{aligned}\xi^{\text{prof}}\{\hat{\boldsymbol{\beta}}\} &= p\ell(\boldsymbol{\beta}^0) + \hat{h}^\top \Delta_n - \frac{1}{2}\hat{h}^\top \left[\tilde{I}_{11}^0 + \Sigma_n\right] \hat{h} + o_P(\|\hat{h}\| + 1)^2, \\ \xi^{\text{prof}}\{\tilde{\boldsymbol{\beta}}\} &= p\ell(\boldsymbol{\beta}^0) + \frac{1}{2}\Delta_n^\top (\tilde{I}_{11}^0 + \Sigma_n)^{-1}\Delta_n^\top + o_P(1).\end{aligned}$$

We proved in Theorem 1 that $\hat{\boldsymbol{\beta}}$ is the maximizer of ξ^{prof} and $\tilde{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10} = O_P(n^{-1/2})$, so

$$\hat{h}^\top \Delta_n - \frac{1}{2}\hat{h}^\top \left[\tilde{I}_{11}^0 + \Sigma_n\right] \hat{h} - \frac{1}{2}\Delta_n^\top (\tilde{I}_{11}^0 + \Sigma_n)^{-1}\Delta_n^\top \geq -o_P(\|\hat{h}\| + 1)^2.$$

On the other hand, the left hand side can be rewritten as

$$\begin{aligned}& -\frac{1}{2}\left[\hat{h} - (\tilde{I}_{11}^0 + \Sigma_n)^{-1}\Delta_n\right]^\top (\tilde{I}_{11}^0 + \Sigma_n)\left[\hat{h} - (\tilde{I}_{11}^0 + \Sigma_n)^{-1}\Delta_n\right] \\ \leq\ & -c\|\hat{h} - (\tilde{I}_{11}^0 + \Sigma_n)^{-1}\Delta_n\|^2\end{aligned}$$

for a positive constant c, by the non-singularity of $\tilde{I}_{11}^0 + \Sigma_n$. Theorem 1 implies that $\|\hat{h}\| = O_P(1)$ and therefore $\|\hat{h} - (\tilde{I}_{11}^0 + \Sigma_n)^{-1}\Delta_n\| = o_P(1)$. Together, these imply

$$\hat{h} = (\tilde{I}_{11}^0 + \Sigma_n)^{-1}\Delta_n + o_p(1).$$

Equivalently,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10}) = (\tilde{I}_{11}^0 + \Sigma_n)^{-1}\frac{1}{\sqrt{n}}\left[\sum_{i=1}^n \tilde{\ell}_{10}(\mathbf{x}_i) - n\boldsymbol{d}_n\right] + o_p(1),$$

which yields

$$\sqrt{n}(\tilde{I}_{11}^0 + \Sigma_n)\left\{\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10} + (\tilde{I}_{11}^0 + \Sigma_n)^{-1}\boldsymbol{d}_n\right\} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\ell}_{10}(\mathbf{x}_i) + o_p(1) \xrightarrow{\mathcal{L}} N(0, \tilde{I}_{11}^0)$$

by the Central Limit Theorem and Slutsky's Theorem. ■

Chapter 3

Variable Selection for Dynamic Networks

3.1 Introduction

Section 1.4.2.2 reviewed several counting process approaches for dynamic networks. In these models, the occurrences of edges depend on various network covariates through some survival models. These covariates help characterize the tendency of having future edges in the network, so more of them can lead to better characterizations. However, the tradeoff is that estimation in these models can become complicated and computationally intensive if too many covariates are included in the model. This motivates us to find proper approaches to simplify the model while keeping the accuracy of the model estimation or prediction.

There exists some literature on variable selection for network models. For example, the work of Fan et al. (2009) about variable selection in graphical models is a bridge between network data analysis and model selection methodology. However, these authors only consider static graphic models. In Chapters 3 and 4, we aim to build suitable survival models for large time-varying networks and apply model selection techniques on these models.

In Chapter 1, two approaches using counting processes are introduced, namely, the egocentric approach and the relational approach. The edges in these approaches are assumed to be instantaneous, and the main interests are on the number of times

a certain edge appears. But if the interest also involves how long the edges last, then the current approaches are not suffice. One needs to consider more complicated process structures, for example, pairs of counting processes for beginning and ending separately. But in this dissertation, we only consider instantaneous edges. Both of the egocentric approach and the relational approach model the intensity process of the corresponding counting process by the Cox model. The partial likelihood function or its approximation is then maximized to estimate the coefficients of the covariates. The maximum partial likelihood estimates are usually not sparse, therefore, to select variables, penalty functions need to be added. As in Fan and Li (2001), if the true model is sparse, good estimates will be consistent and have the oracle property. In this chapter, we extend the work of Fan and Li (2001) to a network setting and show that under certain regularity conditions, the maximum of the penalized partial likelihood function will have these desired properties.

The theory presented in this section demonstrates that the maximizer of the penalized partial likelihood function is sparse and has the oracle property. Therefore it can be used to select sufficient statistic in network models. Although the statements and the proofs of the theorems are similar to those in Fan and Li (2001, 2002), they are derived here for different data structures. In dynamic networks, the observations are not independent and may not be restricted to a bounded time interval. The counting process approach allows observations to be dependent and the corresponding partial likelihood function coincides with the one in the independent survival time setting. Moreover, the properties studied in Perry and Wolfe (2013) for maximum partial likelihood estimators enlighten us in deriving similar results for the egocentric situations, which further provides a bridge to connect Fan and Li (2001, 2002) to sufficient statistic selection for networks. This work represents the first time when covariates are selected via penalization for dynamic network models of the type studied here.

The majority of the following sections focus on an egocentric approach, though, it can be generalized to a relational approach setting very easily, as discussed in Section 3.4. In Section 3.2, we first provide some properties of the partial likelihood function and of its approximations, which prepares the following Section 3.3 for penalized partial likelihood functions. The consistency and oracle properties

are established for a directed network under both single and multiple receivers scenarios.

3.2 Properties of the Partial Likelihood Function and its Approximation

The theorems presented in this section can be derived in a similar way to those of Perry and Wolfe (2013), so we omit certain details of the proofs. We present these theorems here as theoretical justifications for Vu et al. (2011b). In addition, some of these results will be used directly to derive properties of our estimator for variable selection in the egocentric model approach.

3.2.1 Properties of the Partial Likelihood Function

As described in Chapter 1, the general setting for an egocentric approach (Vu et al., 2011b) in a dynamic network is to construct a counting process $N_i(t)$ for each node i, counting the number of edges directed to it or from it, depending on the context. Denote all nodes that are in the network during the observation period as the receiver set $\mathcal{J}$ (or sender set $\mathcal{I}$ if the counting process is counting the edges from a node). Nodes are allowed to enter and leave the network during the observation time. The multivariate counting process $\boldsymbol{N}(t) = (N_j(t), j \in \mathcal{J})$ can be decomposed into a cumulative hazard $\boldsymbol{\Lambda}(t)$ and martingale $\boldsymbol{M}(t)$ as

$$\boldsymbol{N}(t) = \int_0^t \boldsymbol{\lambda}(s)ds + \boldsymbol{M}(t). \tag{3.1}$$

Then the Cox model is used to model each intensity process as

$$\lambda_j(t) = Y_i(t)\lambda_0(t)\exp\{\boldsymbol{\beta}^\top \boldsymbol{s}_j(t)\}, \tag{3.2}$$

where $Y_j(t)$ is the at-risk indicator and $\boldsymbol{s}_j(t)$ is the vector of the covariates just prior to time t.

Suppose for a directed dynamic network that one can observe all the time stamps when edges are established, as well as their corresponding senders and

receivers (although in an egocentric approach, knowing the receiver is sufficient, in most networks, one can observe both senders and receivers). Assume that only one edge can be established at each time point, and the covariates for all the at-risk nodes can be evaluated just prior to the event time. Denote observations by $(i_m, j_m, t_m), m = 1, 2, ..., n$, where i_m is the sender, j_m is the receiver, and t_m is the time stamp. The covariates for all possible receivers $j \in \mathcal{J}_{t_m}$, calculated just prior to time t_m, are denoted by $\boldsymbol{s}_i(t)$. Then the partial likelihood process can be written as

$$\log PL_t(\boldsymbol{\beta}) = \sum_{t_m<t} \left\{ \boldsymbol{\beta}^\top \boldsymbol{s}_{j_m}(t_m) - \log \left[\sum_{j \in \mathcal{J}_{t_m}} \exp\{\boldsymbol{\beta}^\top \boldsymbol{s}_j(t_m)\} \right] \right\}, \tag{3.3}$$

where $\mathcal{J}_t = \{j \in \mathcal{J} | \text{ s.t. } Y_j(t) = 1\}$ is called the at-risk set at time t. These notations are similar those of Perry and Wolfe (2013), where the covariates and at-risk set depend on both sender and receiver. Here, in an egocentric approach, we only define counting process on the receivers.

Perry and Wolfe (2013) study the properties of the partial likelihood function $\log PL_t(\boldsymbol{\beta})$ (see equation (1.66)) as well as the maximum partial likelihood estimator in the relational approach. They also verify that the maximum partial likelihood estimator is consistent and has an asymptotic normal distribution with covariance matrix $[\Sigma_1(\beta_0)]^{-1}$, where the inverse Hessian of $-\log PL_t(\cdot)$ is a good estimate of $[\Sigma_1(\beta_0)]^{-1}$. These results are not a direct application of the asymptotic results found in Andersen and Gill (1982), since the data structure of a network is different from that in the regular counting process approach for survival models. To increase the number of observations, one has to expand the observation time interval, so the asymptotic theory cannot be derived for a bounded time interval. Thus, Perry and Wolfe (2013) rescale the time interval for each fixed n and derive properties of the maximum partial likelihood estimator through a discretized version of the score function.

In an egocentric approach, we can derive parallel results following very similar arguments to those of Perry and Wolfe (2013). In the rest of this subsection, we will present these results in theorems with outlines of proofs. To simplify some of

the regularity conditions, we define the notations

$$
\begin{aligned}
w_t(\boldsymbol{\beta}, j) &= \exp\left\{\boldsymbol{\beta}^\top \boldsymbol{s}_j(t)\right\} Y_j(t), && (3.4)\\
W_t(\boldsymbol{\beta}) &= \sum_{j \in \mathcal{J}_t} w_t(\boldsymbol{\beta}, j), && (3.5)\\
E_t(\boldsymbol{\beta}) &= \frac{1}{W_t(\boldsymbol{\beta})} \sum_{j \in \mathcal{J}_t} w_t(\boldsymbol{\beta}, j) \boldsymbol{s}_j(t), && (3.6)\\
V_t(\boldsymbol{\beta}) &= \frac{1}{W_t(\boldsymbol{\beta})} \sum_{j \in \mathcal{J}_t} w_t(\boldsymbol{\beta}, j) \{\boldsymbol{s}_j(t) - E_t(\boldsymbol{\beta})\}^{\otimes 2}, && (3.7)
\end{aligned}
$$

where $\boldsymbol{a}^{\otimes 2}$ means $\boldsymbol{a}\boldsymbol{a}^\top$ for a column vector $\boldsymbol{a}$. Using this notation, the log-partial likelihood function (3.3) given data up to time t can be written as

$$
\log PL_t(\boldsymbol{\beta}) = \sum_{t_m < t} \left\{\boldsymbol{\beta}^\top \boldsymbol{s}_{j_m}(t_m) - \log W_{t_m}(\boldsymbol{\beta})\right\} \tag{3.8}
$$

and the corresponding gradient and negative hessian are

$$
\begin{aligned}
U_t(\boldsymbol{\beta}) &= \nabla \log PL_t(\boldsymbol{\beta}) = \sum_{t_m < t} \{\boldsymbol{s}_{j_m}(t_m) - E_{t_m}(\boldsymbol{\beta})\}, && (3.9)\\
I_t(\boldsymbol{\beta}) &= -\nabla^2 \log PL_t(\boldsymbol{\beta}) = \sum_{t_m < t} V_{t_m}(\boldsymbol{\beta}). && (3.10)
\end{aligned}
$$

Following Perry and Wolfe (2013), we establish the following regularity conditions for proving the asymptotic properties.

(C1) The covariates are uniformly square-integrable, i.e.,

$$
E\left[\sup_{j,t} ||\boldsymbol{s}_j(t)||^2\right] < \infty
$$

(C2) For any $\boldsymbol{\beta}$ in a neighborhood of the true parameter value $\boldsymbol{\beta}_0$, and any $\alpha \in [0, 1]$, as $n \to +\infty$, then there exists a positive semi-definite matrix $\Sigma_\alpha(\boldsymbol{\beta})$ such that,

$$
\frac{1}{n} \int_0^{t_{\lfloor \alpha n \rfloor}} V_s(\boldsymbol{\beta}) W_s(\boldsymbol{\beta}) \lambda_0(s) ds \xrightarrow{P} \Sigma_\alpha(\boldsymbol{\beta}).
$$

(C3) For each n, $P(t_n < \infty) = 1$

(C4) $\{V_{t_n}(\cdot)\}$ is an equicontinuous family of functions.

These conditions are egocentric version of conditions (A1) to (A4) in Perry and Wolfe (2013). They allow us to extend Theorems 3.1 and 3.2 of Perry and Wolfe (2013) to the case of egocentric network models, as follows.

Theorem 3.2.1. *Suppose the multivariate counting process $\boldsymbol{N}$ for nodes has an intensity process (3.2). If the above (C1) to (C4) hold, then as $n \to +\infty$,*

(a) $\frac{1}{\sqrt{n}} U_{t_{\lfloor \alpha n \rfloor}}(\boldsymbol{\beta}_0)$ *converges weakly to a Gaussian process on* $[0,1]$ *with covariance function* $\Sigma_\alpha(\boldsymbol{\beta}_0)$.

(b) For any consistent estimator $\hat{\boldsymbol{\beta}}_n$ *of* $\boldsymbol{\beta}_0$, *then*

$$\sup_{\alpha \in [0,1]} \left\| \frac{1}{n} I_{t_{\lfloor \alpha n \rfloor}}(\hat{\boldsymbol{\beta}}_n) - \Sigma_\alpha(\boldsymbol{\beta}_0) \right\| \xrightarrow{P} 0.$$

A sketch of the proofs is as follows:
According to the Doob-Meyer decomposition, $N_j(t) = \Lambda_j(t) + M_j(t)$, where $\Lambda_j(t) = \int_0^s \lambda_j(s) ds$ is the compensator. So the score function (3.9) at $\boldsymbol{\beta}_0$ can be rewritten as

$$U_t(\boldsymbol{\beta}_0) = \sum_{j \in \mathcal{J}} \int_0^t \{\boldsymbol{s}_j(s) - E_s(\boldsymbol{\beta}_0)\} \, dN_j(s) = \sum_{j \in \mathcal{J}} \int_0^t \{\boldsymbol{s}_j(s) - E_s(\boldsymbol{\beta}_0)\} \, dM_j(s),$$

where the second equality is because $\sum_{j \in J} \int_0^t \{\boldsymbol{s}_j(s) - E_s(\boldsymbol{\beta}_0)\} \, d\Lambda_j(s) = 0$. It can be verified that $U_t(\boldsymbol{\beta}_0)$ is locally square integrable with predictable variation $\int_0^t V_s(\boldsymbol{\beta}_0) d\Lambda(s)$. For each n, define a rescaled score for $\alpha \in [0,1]$ by

$$\tilde{U}_\alpha^{(n)}(\boldsymbol{\beta}_0) = U_{t_{\lfloor \alpha n \rfloor}}(\boldsymbol{\beta}_0). \tag{3.11}$$

This rescaled score is a square-integrable martingale adapted to $\mathcal{F}_{t_{\lfloor \alpha n \rfloor}}$. Therefore by assumption (A.2) and the Martingale Central Limit Theorem in Rellbolledo (1980) (the Lindeberg condition can be verified), part (a) is proved.

For part (b),

$$\begin{aligned} I_{t_{\lfloor \alpha n \rfloor}}(\boldsymbol{\beta}_0) &= \int_0^{t_{\lfloor \alpha n \rfloor}} V_s(\boldsymbol{\beta}_0) dN(s) \\ &= \int_0^{t_{\lfloor \alpha n \rfloor}} V_s(\boldsymbol{\beta}_0) d\Lambda(s) + \int_0^{t_{\lfloor \alpha n \rfloor}} V_s(\boldsymbol{\beta}_0) dM(s), \end{aligned} \tag{3.12}$$

which implies

$$\begin{aligned} \left\| \frac{1}{n} I_{t_{\lfloor \alpha n \rfloor}}(\hat{\boldsymbol{\beta}}_n) - \Sigma_\alpha(\boldsymbol{\beta}_0) \right\| \leq & \left\| \frac{1}{n} \int_0^{t_{\lfloor \alpha n \rfloor}} V_s(\boldsymbol{\beta}_0) dM(s) \right\| + \left\| \frac{1}{n} \int_0^{t_{\lfloor \alpha n \rfloor}} V_s(\boldsymbol{\beta}_0) d\Lambda(s) - \Sigma_\alpha(\boldsymbol{\beta}_0) \right\| \\ & + \left\| \frac{1}{n} \int_0^{t_{\lfloor \alpha n \rfloor}} \left[V_s(\hat{\boldsymbol{\beta}}_n) - V_s(\boldsymbol{\beta}_0) \right] d\Lambda(s) \right\|. \end{aligned} \tag{3.13}$$

The second and third terms converge to 0 in probability by assumptions (C4) and (C2). By condition (C3), Lenglart's Inequality (Corollary 3.4.1 in Fleming and Harrington (2011)) can be applied. So the first term follows the following inequality, for any positive ϵ and δ,

$$P\left\{ \sup_{t \in [0, t_n]} \left\| \frac{1}{n} \int_0^t V_s(\boldsymbol{\beta}_0) dM(s) \right\| > \epsilon \right\} \leq \frac{\delta}{\epsilon^2} + P\left\{ \frac{1}{n^2} \int_0^{t_n} \left\| V_s(\boldsymbol{\beta}_0) \right\|^2 d\Lambda(s) \geq \delta \right\}$$

The second term on right hand side is bounded by $\frac{1}{n^2} 16 (\sup_{j,t} \boldsymbol{s}_j(t))^4 \Lambda(t_n)$. By assumption (C1), $(\sup_{j,t} \boldsymbol{s}_j(t))^4 / n \xrightarrow{P} 0$. Also $E[\Lambda(t_n)] = n$ and ϵ is arbitrary, so the left hand side also converges to zero. □

The result of (b) is weaker than convergence for all time points. However, we will only need to use the results when $\alpha = 1$ in the next theorem as well as some later theorems in Section 3.3. The following theorem establishes consistency and asymptotic normality of the maximum partial likelihood estimator.

Theorem 3.2.2. *Suppose the conditions in Theorem 1 hold, and* $I_{t_n}(\boldsymbol{\beta}) \xrightarrow{P} \Sigma_1(\boldsymbol{\beta})$ *for* $\boldsymbol{\beta}$ *in a neighborhood of* $\boldsymbol{\beta}_0$*, where* $\Sigma_1(\cdot)$ *is locally Lipschitz and the smallest eigenvalue is bounded away from zero in that neighborhood. As* $n \to +\infty$*, if the maximum partial likelihood estimator is* $\hat{\boldsymbol{\beta}}_n$*, then*

(a) $\hat{\boldsymbol{\beta}}_n$ *is consistent.*

(b) $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{L}} N(\mathbf{0}, [\Sigma_1(\boldsymbol{\beta}_0)]^{-1})$

A sketch of the proof is as follows:
Since $I_t(\boldsymbol{\beta})$ is positive semi-definite, $\log PL_t(\cdot)$ is concave. Also by assumption that the smallest eigenvalue of $\Sigma_1(\cdot)$ is bounded away from zero in a neighborhood of true parameter value $\boldsymbol{\beta}_0$, there exists a local maximum for $\log PL_t(\cdot)$ in that neighborhood when n is sufficiently large. Denote the local maximum as $\hat{\boldsymbol{\beta}}_n$, which should also be the global maximum.

Define $Z_n = \left[\frac{1}{n} I_{t_n}(\boldsymbol{\beta}_0)\right]^{-1} \left[\frac{1}{n} U_{t_n}(\boldsymbol{\beta}_0)\right]$, By Theorem 3.2.1 and Slutsky's theorem, $\sqrt{n} Z_n \xrightarrow{\mathcal{L}} N(0, [\Sigma_1(\boldsymbol{\beta}_0)]^{-1})$ and $Z_n \xrightarrow{P} 0$. Therefore, the one-step Newton estimator

$$\boldsymbol{\beta}_{1,n} = \boldsymbol{\beta}_0 - Z_n$$

will have the properties in (a) and (b). In addition, in a neighborhood of $\boldsymbol{\beta}_0$, by assumption $\Sigma_1(\cdot)$ is locally Lipschitz, $\left\| \frac{1}{n} [I_{t_n}(\boldsymbol{\beta}_0)]^{-1} \right\|$ is bounded, and $Z_n = O_P(n^{-1/2})$, which implies that for sufficiently large n,

$$\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| \leq 2\|Z_n\| \xrightarrow{P} 0$$

by the weaker version of the Kantorovich Theorem (Lemma B.4 in Perry and Wolfe (2013)). Also, by the same theorem, for large enough n, there exists a large constant K such that,

$$\sqrt{n}\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{1,n}\| \leq 2\sqrt{n} K \|Z_n\|^2 \xrightarrow{P} 0.$$

which concludes the proof. □

The above two theorems are derived under the assumptions that only a single receiver and a single receiver are allowed at each event time. However, this may not always be the case, especially for interaction networks or citation networks. In the next section, estimates for networks with a single sender and multiple receivers will be considered, and similar results can be established under further regularity conditions.

3.2.2 Approximations of the Partial Likelihood Function

In some dynamic networks, at each event time a single sender may send edges towards multiple receivers. In this scenario, the partial likelihood function will have a more complicated form than (3.3), namely,

$$\log PL_t(\boldsymbol{\beta}) = \sum_{t_m<t} \left\{ \sum_{j\in J_m} \boldsymbol{\beta}^\top \boldsymbol{s}_j(t_m) - \log \left[\sum_{J\subseteq \mathcal{J}_{t_m}, |J|=|J_m|} \exp\left\{ \sum_{j\in J} \boldsymbol{\beta}^\top \boldsymbol{s}_j(t_m) \right\} \right] \right\}, \tag{3.14}$$

where J_m is the receiver set for the sender i_m at time t_m and $\mathcal{J}_{t_m}$ is still the at-risk set at time t_m. To calculate this partial likelihood function, one has to do calculation over all the subsets of the at-risk set that have the same size as the receiver set. The computational cost increases rapidly when $\mathcal{J}_{t_m}$ grows fast with t_m. To avoid this computational complexity, both Vu et al. (2011b) and Perry and Wolfe (2013) considered the so-called Breslow Approximation, by which (3.14) can be approximated by

$$\log \widetilde{PL_t}(\boldsymbol{\beta}) = \sum_{t_m<t} \left\{ \sum_{j\in J_m} \boldsymbol{\beta}^\top \boldsymbol{s}_j(t_m) - |J_m| \log \left[\sum_{j\in \mathcal{J}_{t_m}} \exp\{\boldsymbol{\beta}^\top \boldsymbol{s}_j(t_m)\} \right] \right\}. \tag{3.15}$$

How good the approximation is relates to the growth rates of the at-risk set $\mathcal{J}_{t_m}$ and J_m. As in Perry and Wolfe (2013), we define the receiver set growth sequence as

$$G_n = \sum_{t_m\le t_n} \frac{1\{|J_m|>1\}}{|\mathcal{J}_{t_m}|}. \tag{3.16}$$

Then the approximation errors in the gradient and Hessian are bounded by G_n as in the following theorem.

Theorem 3.2.3. *Assume that regularity condition (C1) holds and* $\sup_m |J_m|$ *is bounded in probability. Then for* $\boldsymbol{\beta}$ *in a neighborhood of the true parameter value* $\boldsymbol{\beta}_0$*,*

$$\| \bigtriangledown [\log PL_{t_n}(\boldsymbol{\beta})] - \bigtriangledown[\log \widetilde{PL}_{t_n}(\boldsymbol{\beta})] \| = O_P(G_n), \tag{3.17}$$

$$\| \bigtriangledown^2 [\log PL_{t_n}(\boldsymbol{\beta})] - \bigtriangledown^2[\log \widetilde{PL}_{t_n}(\boldsymbol{\beta})] \| = O_P(G_n). \tag{3.18}$$

A sketch of the proof is as follows:
Equations (3.14) and (3.15) differ only in the summation over the at-risk set. In particular,

$$\log PL_{t_n}(\boldsymbol{\beta}) - \log \widetilde{PL}_{t_n}(\boldsymbol{\beta}) = \sum_{t_m<t} \left\{ \tilde{W}_{t_m}(\boldsymbol{\beta}, |J_m|) - \widetilde{W}_{t_m}(\boldsymbol{\beta}, |J_m|) \right\},$$

where

$$\begin{aligned} W_{t_m}(\boldsymbol{\beta}, L) &= \log \sum_{J \subseteq \mathcal{J}_{t_m}, |J|=L} \left\{ \prod_{j \in J} w_{t_m}(\boldsymbol{\beta}, j) \right\}, \\ \widetilde{W}_{t_m}(\boldsymbol{\beta}, L) &= L \log \left\{ \sum_{j \in \mathcal{J}_{t_m}} w_{t_m}(\boldsymbol{\beta}, j) \right\}. \end{aligned}$$

Differentiating with respect to $\boldsymbol{\beta}$ gives

$$\begin{aligned} \bigtriangledown W_{t_m}(\boldsymbol{\beta}, L) &= \frac{\sum_{J \subseteq \mathcal{J}_{t_m}, |J|=L} \left\{ \prod_{j \in J} w_{t_m}(\boldsymbol{\beta}, j) \sum_{j \in J} \boldsymbol{s}_j(t_m) \right\}}{\sum_{J \subseteq \mathcal{J}_{t_m}, |J|=L} \left\{ \prod_{j \in J} w_{t_m}(\boldsymbol{\beta}, j) \right\}}, \\ \bigtriangledown \widetilde{W}_{t_m}(\boldsymbol{\beta}, L) &= L \frac{\sum_{j \in \mathcal{J}_{t_m}} w_{t_m}(\boldsymbol{\beta}, j) \boldsymbol{s}_j(t_m)}{\sum_{j \in \mathcal{J}_{t_m}} w_{t_m}(\boldsymbol{\beta}, j)}. \end{aligned}$$

Consider experiments of randomly drawing L nodes $\{j_1, j_2, \ldots, j_L\}$ using weights $w_j(\boldsymbol{\beta}, t_m)$ at each t_m. The above two equations can be considered as the expected sum of covariates values for these L nodes. The second one is drawing i.i.d samples, and the first one is the second one conditional on the L nodes are all different. Suppose the two probability laws are denote by $\widetilde{P}_{t_m,\boldsymbol{\beta};L}$ and $P_{t_m,\boldsymbol{\beta};L}$. For any joint distribution $P^*_{t_m,\boldsymbol{\beta};L}$ of drawing $\{j_1, j_2, \ldots, j_L\}$ and $\{\widetilde{j}_1, \widetilde{j}_2, \ldots, \widetilde{j}_L\}$, which has associate marginally distribution $P_{t_m,\boldsymbol{\beta};L}$ and $\widetilde{P}_{t_m,\boldsymbol{\beta};L}$, the following inequality holds.

$$\begin{aligned} & \| \bigtriangledown W_{t_m}(\boldsymbol{\beta}, L) - \bigtriangledown \widetilde{W}_{t_m}(\boldsymbol{\beta}, L) \| \\ \leq\ & E^*_{t_m,\boldsymbol{\beta};L} \left[2L \left\{ \sup_{j,t} \|s_j(t)\| \right\} I \left\{ (j_1, j_2, \ldots, j_L) \neq (\widetilde{j}_1, \widetilde{j}_2, \ldots, \widetilde{j}_L) \right\} \right] \\ =\ & 2L \left[\sup_{j,t} \|s_j(t)\| \right] P^*_{t_m,\boldsymbol{\beta};L} \left\{ (j_1, j_2, \ldots, j_L) \neq (\widetilde{j}_1, \widetilde{j}_2, \ldots, \widetilde{j}_L) \right\} \end{aligned} \tag{3.19}$$

Perry and Wolfe (2013) construct a joint distribution $P^*_{t_m,\boldsymbol{\beta};L}$ satisfying the condition, and also the probability on the right hand side is the smallest. The sampling is as follows,

- Draw $(\widetilde{j}_1, \widetilde{j}_2, \ldots, \widetilde{j}_L)$ according to $\widetilde{P}_{t_m,\boldsymbol{\beta};L}$
- if $(\widetilde{j}_1, \widetilde{j}_2, \ldots, \widetilde{j}_L)$ are all different, set $(j_1, j_2, \ldots, j_L) = (\widetilde{j}_1, \widetilde{j}_2, \ldots, \widetilde{j}_L)$. Otherwise, draw $(j_1, j_2, \ldots, j_L)$ by $P_{t_m,\boldsymbol{\beta};L}$

In this sampling scheme, the probability in the last equation is bounded by the probability that $(\widetilde{j}_1, \widetilde{j}_2, \ldots, \widetilde{j}_L)$ is not all different, which is then bounded by

$$\begin{aligned}\sum_{k<l} \widetilde{P}_{t_m,\boldsymbol{\beta};L}\left\{\widetilde{j}_l = \widetilde{j}_k\right\} &= \binom{L}{2} \sum_{j\in\mathcal{J}_{t_m}} \left[\frac{w_{t_m}(\boldsymbol{\beta}, j)}{\sum_{j\in\mathcal{J}_{t_m}} w_{t_m}(\boldsymbol{\beta}, j)}\right]^2 \\ &\leq \binom{L}{2} \frac{\exp\{4K\|\boldsymbol{\beta}\|\}}{|\mathcal{J}_{t_m}|} \end{aligned} \tag{3.20}$$

where $K = \sup_{j\in\mathcal{J}_t, t} \|\boldsymbol{s}_j(t)\|$. Therefore, the difference in equation (3.19) is bounded by $K\exp\{4K\|\boldsymbol{\beta}\|\}\dfrac{L^2(L-1)}{|\mathcal{J}_{t_m}|}$. Since K is bounded in probability, so the difference in the gradient is bounded in probability by sum of these orders, i.e.

$$\sum_{t_m<t_n} \frac{|J_m|^2(|J_m|-1)}{|\mathcal{J}_{t_m}|}$$

Since $\sup_m |J_m|$ is bounded, it is of the same order of G_n. Similar proofs can be used for the Hessian. □

These boundaries can help to evaluate the differences between the maximum $\tilde{\boldsymbol{\beta}}_n$ of the approximated partial likelihood function (3.15) and the real MLE $\hat{\boldsymbol{\beta}}_n$:

Theorem 3.2.4. *Under the assumption of Theorem 3.2.3, denote the maximizer of* $\log \widetilde{PL}_{t_n}(\cdot)$ *as* $\tilde{\boldsymbol{\beta}}_n$ *and the maximizer of* $\log PL_{t_n}(\cdot)$ *as* $\hat{\boldsymbol{\beta}}_n$. *Assume for all n that* $\dfrac{1}{n}\bigtriangledown^2\left[\log \widetilde{PL}_{t_n}(\cdot)\right]$ *is uniformly locally Lipschitz and the smallest eigenvalue is bounded away from zero in a neighborhood of* $\hat{\boldsymbol{\beta}}_n$. *Then if* $G_n/n \xrightarrow{\mathcal{P}} 0$.

$$\|\tilde{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_n\| = O_P(G_n/n).$$

A sketch of the proof is as follows:
Let $\hat{\boldsymbol{\beta}}_n$ be the initial value to maximize $\log \widetilde{PL}_{t_n}(\cdot)$. Then as in the proofs of Theorem 3.2.2, by the Kantorovich Theorem, $\|\tilde{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_n\|$ is bounded above by

$$\left\{ \bigtriangledown^2 \left[\frac{1}{n} \log \widetilde{PL}_{t_n}(\hat{\boldsymbol{\beta}}) \right] \right\}^{-1} \bigtriangledown \left[\frac{1}{n} \log \widetilde{PL}_{t_n}(\hat{\boldsymbol{\beta}}) \right] = O_P(G_n/n)$$

□

Therefore, if $\mathcal{J}_{t_m}$ grows fast enough and $G_n = O(\sqrt{n})$, then $\tilde{\boldsymbol{\beta}}_n$ is still consistent and $\sqrt{n}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ still has an asymptotic normal distribution with the variance $[\Sigma_1(\boldsymbol{\beta}_0)]^{-1}$. For example, if $\mathcal{J}_{t_m} = O(m)$, then the maximum approximated partial likelihood estimate still enjoys the nice properties of the maximum partial likelihood estimate.

3.3 Variable Selection via Penalized Partial Likelihood

In this section, we will consider the problem of sufficient statistic selection for dynamic networks with the egocentric approach described in previous section. Using the idea of penalization, we will maximize the penalized partial likelihood function to get a sparse estimate of $\boldsymbol{\beta}$, thus attain the goal of sufficient statistic selection. As in other literature for variable selection via penalization, we will derive asymptotic properties of the resulting estimates including consistency and the oracle property. Since we want to consider a broader range of dynamic networks, we will introduce results for both single-receiver and multiple-receivers scenarios in two consecutive subsections.

3.3.1 Network with Single Sender and Single Receiver

Recall with a single receiver, the observations are (i_m, j_m, t_m), and under the regularity conditions (C1)-(C4), we have

$$n^{-1/2} \bigtriangledown [\log PL_{t_n}(\boldsymbol{\beta}^0)] \quad \xrightarrow{\mathcal{L}} \quad N(\mathbf{0}, \Sigma_1(\boldsymbol{\beta}^0) \text{ as } n \to +\infty, \tag{3.21}$$

$$-n^{-1} \bigtriangledown^2 [\log PL_{t_n}(\boldsymbol{\beta}^0))] \quad \xrightarrow{P} \quad \Sigma_1(\boldsymbol{\beta}^0) \qquad \text{as } n \to +\infty, \tag{3.22}$$

where $\boldsymbol{\beta}^0$ is the true value of parameter $\boldsymbol{\beta}$, and $\Sigma_1(\boldsymbol{\beta}^0)$ is the covariance matrix of $\boldsymbol{\beta}$ evaluated at $\boldsymbol{\beta}^0$.

Suppose the true model is sparse, which means the true parameter vector can be decomposed into a nonzero part and a zero part, i.e., $\boldsymbol{\beta}^0 = (\boldsymbol{\beta}^{10}, \boldsymbol{\beta}^{20}), \boldsymbol{\beta}^{10} \in \mathbb{R}^s, \boldsymbol{\beta}^{20} \in \mathbb{R}^r, r + s = p$, and $\boldsymbol{\beta}^{20} = \mathbf{0}$. We first show that under further regularity conditions, there exists a penalized likelihood estimator that is consistent for $\boldsymbol{\beta}^0$. Furthermore, the rate of convergence may be shown to be $n^{-1/2}$ for certain choices of penalty function, which parallels the consistency result for any parametric model (Fan and Li, 2001). The proofs follow the theorems.

We need further regularity conditions in addition to (C1)-(C4):

(C5) As $n \to \infty$,

$$a_n \overset{\text{def}}{=} \max_{1 \le j \le s} \{p'_{\lambda_n}(|\beta_j^0|)\} \to 0$$

and

$$b_n \overset{\text{def}}{=} \max_{1 \le j \le s} \{p''_{\lambda_n}(|\beta_j^0|)\} \to 0.$$

(C6) The penalty function $p_{\lambda_n}(x)$ is twice differentiable at all $x > 0$ and there exist positive constants K and M such that for all $\lambda_n > 0$ and $x > 0$, $p''_{\lambda_n}(x)$ exists, $|p''_{\lambda_n}(x)| < M$, and $p'_{\lambda_n}(0+)/\lambda_n > K$.

(C7) As $n \to +\infty$, $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to +\infty$.

(C8) There exist an open subset ω which contains the true parameter point $\boldsymbol{\beta}^0$ such that the partial likelihood function admits all third derivatives. Furthermore, the third-order derivatives are bounded in probability.

$$\frac{1}{n} \left| \frac{\partial^3 \log PL_{t_n}(\boldsymbol{\beta})}{\partial\beta_i \partial\beta_j \partial\beta_k} \right| = O_p(1), i, j, k \in \{1, 2, ..., p\}.$$

We must first derive a form of expansion of the log partial likelihood function. For $\boldsymbol{\beta}$ in a neighborhood of true parameter $\boldsymbol{\beta}^0$, the Taylor expansion of the log partial likelihood function $\log PL_{t_n}(\boldsymbol{\beta})$ around $\boldsymbol{\beta}^0$ is

$$\log PL_{t_n}(\boldsymbol{\beta}) \quad = \quad \log PL_{t_n}(\boldsymbol{\beta}^0) + (\boldsymbol{\beta} - \boldsymbol{\beta}^0)^T \bigtriangledown [\log PL_{t_n}(\boldsymbol{\beta}^0)]$$

$$
\begin{aligned}
&+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^0)^T \bigtriangledown^2 [\log PL_{t_n}(\boldsymbol{\beta}^0)](\boldsymbol{\beta}-\boldsymbol{\beta}^0) \\
&+\sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\frac{\partial^3 \log PL_{t_n}(\boldsymbol{\beta}^*)}{\partial\beta_i\partial\beta_j\partial\beta_k}(\beta_i^*-\beta_i^0)(\beta_j^*-\beta_j^0)(\beta_k^*-\beta_k^0) \\
= \;& \log PL_{t_n}(\boldsymbol{\beta}^0)+(\boldsymbol{\beta}-\boldsymbol{\beta}^0)^T \bigtriangledown [\log PL_{t_n}(\boldsymbol{\beta}^0)] \\
&+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^0)^T \bigtriangledown^2 [\log PL_{t_n}(\boldsymbol{\beta}^0)](\boldsymbol{\beta}-\boldsymbol{\beta}^0) \\
&+O_P(n\|\boldsymbol{\beta}^*-\boldsymbol{\beta}^0\|^3).
\end{aligned}
$$

where $\boldsymbol{\beta}^*$ is between $\boldsymbol{\beta}^0$ and $\boldsymbol{\beta}$.

Theorem 3.3.1. ***(Consistency)*** *Assume we have observed n interactions* $\{(i_m, j_m, t_m), m = 1, ..., n.\}$ *from a network. And* $PL_{t_n}(\boldsymbol{\beta})$ *denotes the partial likelihood function given the observations. Define*

$$\xi(\boldsymbol{\beta}) = \log PL_{t_n}(\boldsymbol{\beta}) - n\sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|) \tag{3.23}$$

as the penalized partial likelihood function. Let a_n *and* b_n *be defined in the regularity conditions and assume regularity condition (C1)-(C8) hold, and the minimum eigenvalue for* $\Sigma_1(\boldsymbol{\beta})$, λ_{min}, *is bounded away from zero. Then there exists a local maximizer* $\hat{\boldsymbol{\beta}}$ *of* $\xi(\boldsymbol{\beta})$, *such that* $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\| = O_p(n^{-1/2}+a_n)$.

Proof. Let $\alpha_n = n^{-1/2}+a_n$. It is sufficient to show that for any $\varepsilon > 0$, there exists a large constant C such that

$$\lim_{n\to\infty} P\left(\sup_{\|\mathbf{u}\|=C} \xi(\alpha_n\mathbf{u}+\boldsymbol{\beta}^0) < \xi(\boldsymbol{\beta}^0)\right) \geq 1-\varepsilon.$$

Let $D_n(\boldsymbol{u}) = \xi(\alpha_n\boldsymbol{u}+\boldsymbol{\beta}^0) - \xi(\boldsymbol{\beta}^0)$. Then

$$
\begin{aligned}
D_n(\boldsymbol{u}) &= \log PL_{t_n}(\boldsymbol{\beta}^0+\alpha_n\mathbf{u}) - \log PL_{t_n}(\boldsymbol{\beta}^0) - n\sum_{j=1}^{p} p_{\lambda_n}(|\beta_j^0+\alpha_n u_j|) + n\sum_{j=1}^{p} p_{\lambda_n}(|\beta_j^0|) \\
&\leq \underbrace{\log PL_{t_n}(\boldsymbol{\beta}^0+\alpha_n\mathbf{u}) - \log PL_{t_n}(\boldsymbol{\beta}^0)}_{\text{A}} \underbrace{- n\sum_{j=1}^{s}\left[p_{\lambda_n}(|\beta_j^0+\alpha_n u_j|) - p_{\lambda_n}(|\beta_j^0|)\right]}_{\text{B}}.
\end{aligned}
$$

We first consider A and B separately. For any sequence $\tilde{\boldsymbol{\beta}}_n \xrightarrow{P} \boldsymbol{\beta}^0$, we have the Taylor expansion ($\tilde{\boldsymbol{\beta}}_n^*$ is between $\tilde{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}^0$),

$$\begin{aligned}\log PL_{t_n}(\tilde{\boldsymbol{\beta}}) &= \log PL_{t_n}(\boldsymbol{\beta}^0) + (\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0)^T \bigtriangledown [\log PL_{t_n}(\boldsymbol{\beta}^0)] \\ &\quad + \frac{1}{2}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0)^T \bigtriangledown^2 [\log PL_{t_n}(\boldsymbol{\beta}^0)](\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0) + O_P(n\|\tilde{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}^0\|^3)\end{aligned}$$

Then for $\tilde{\boldsymbol{\beta}}_n = \boldsymbol{\beta}^0 + \alpha_n \boldsymbol{u}$, we obtain

$$\begin{aligned}\text{A} &= \alpha_n \boldsymbol{u}^T \bigtriangledown [\log PL_{t_n}(\boldsymbol{\beta}^0)] + \frac{\alpha_n^2}{2} \boldsymbol{u}^T \bigtriangledown^2 [\log PL_{t_n}(\boldsymbol{\beta}^0)] \boldsymbol{u} + O_P(n\alpha_n^3 \|\boldsymbol{u}\|^3) \\ &= \sqrt{n}\alpha_n \boldsymbol{u}^T \left\{ \frac{1}{\sqrt{n}} \bigtriangledown [\log PL_{t_n}(\boldsymbol{\beta}^0)] \right\} - \frac{n\alpha_n^2}{2} \boldsymbol{u}^T \left\{ -\frac{1}{n} \bigtriangledown^2 [\log PL_{t_n}(\boldsymbol{\beta}^0)] \right\} \boldsymbol{u} \\ &\quad + O_P(n\alpha_n^3) \\ &= \sqrt{n}\alpha_n \boldsymbol{u}^T \left\{ \frac{1}{\sqrt{n}} \bigtriangledown [\log PL_{t_n}(\boldsymbol{\beta}^0)] \right\} - \frac{n\alpha_n^2}{2} \boldsymbol{u}^T \left\{ \Sigma_1(\boldsymbol{\beta}^0) + o_P(1) \right\} \boldsymbol{u} \\ &\quad + O_P(n\alpha_n^3) \\ &\leq \sqrt{n}\alpha_n \|\boldsymbol{u}^T\| Z_n - \frac{n\alpha_n^2}{2} \|\boldsymbol{u}\|^2 \left\{ \lambda_{\min} + o_P(1) \right\} + O_P(n\alpha_n^3) \qquad (3.24)\end{aligned}$$

where $Z_n = \|\frac{1}{\sqrt{n}} \bigtriangledown [\log PL_{t_n}(\boldsymbol{\beta}^0)]\|$ The last equation is because of (3.22).

Next, consider

$$\begin{aligned}\text{B} &= n \sum_{j=1}^{s} \left| \alpha_n p'_{\lambda_n}(|\beta_j^0|) u_j + \alpha_n^2 p''_{\lambda_n}(|\beta_j^0|) u_j^2 (1 + o_p(1)) \right| \\ &\leq \sqrt{s} n \alpha_n a_n \|\mathbf{u}\| + n\alpha_n^2 b_n \|\mathbf{u}\|^2 (1 + o_p(1)) \qquad (3.25)\end{aligned}$$

Therefore,

$$\begin{aligned}\sup_{\|\mathbf{u}\|=C} D_n(\boldsymbol{u}) &= \sup_{\|\mathbf{u}\|=C} \{A - B\} \leq \sup_{\|\mathbf{u}\|=C} A + \sup_{\|\mathbf{u}\|=C} |B| \\ &= \sqrt{n}\alpha_n C Z_n - \frac{n\alpha_n^2}{2} C^2 \left\{ \lambda_{\min} + o_P(1) \right\} + O_P(n\alpha_n^3) \\ &\quad + \sqrt{s} n \alpha_n a_n C + n\alpha_n^2 b_n C^2 (1 + o_p(1)) \\ &= n\alpha_n^2 \{ C \frac{Z_n}{\sqrt{n}\alpha_n} - \frac{C^2}{2} \left[\lambda_{\min} + o_P(1) \right] + O_P(\alpha_n) \\ &\quad + \sqrt{s} C \frac{a_n}{\alpha_n} + b_n C^2 \left[1 + o_P(1) \right] \}\end{aligned}$$

$$\overset{b_n=o_P(1)}{=} \quad n\alpha_n^2 \left\{ C\left[\frac{Z_n}{\sqrt{n}\alpha_n} + \sqrt{s}\frac{a_n}{\alpha_n}\right] - \frac{C^2}{2}\left[\lambda_{\min} + o_P(1)\right] + o_P(1)\right\} \tag{3.26}$$

If denote $D_n = \dfrac{Z_n}{\sqrt{n}\alpha_n} + \sqrt{s}\dfrac{a_n}{\alpha_n}$, then

$$\begin{aligned} & CD_n - \frac{C^2}{2}\left[\lambda_{\min} + o_P(1)\right] + o_P(1) < 0 \\ \Leftrightarrow \quad & D_n + o_P < \frac{C}{2}\lambda_{\min} \end{aligned} \tag{3.27}$$

According to (3.21),

$$\frac{1}{\sqrt{n}} \bigtriangledown \left[\log PL_{t_n}(\boldsymbol{\beta}^0)\right] \quad \text{is bounded in probability.} \tag{3.28}$$

So $Z_n = O_P(1)$, and $D_n + o_P(1) = O_P(1)$ Therefore, for any ε, there exists a large number C, such that

$$\lim_{n\to\infty} P\left(|D_n + o_P(1)| < \frac{C}{2}\lambda_{\min}\right) \geq 1 - \varepsilon$$

so,

$$\lim_{n\to\infty} P\left(\sup_{\|\mathbf{u}\|=C} D_n(\boldsymbol{u}) < 0\right) \geq 1 - \varepsilon$$

□

Remark: If $a_n \to 0$ at least as fast as $n^{-1/2}$, then $\hat{\boldsymbol{\beta}}$ in Theorem 1 is $\sqrt{n}$-consistent. For the hard thresholding and SCAD penalty functions, if $\lambda_n \to 0$ as $n \to \infty$, then $a_n = 0$ for n large enough. Therefore, Theorem 1 implies that the penalized likelihood estimator is $\sqrt{n}$-consistent in this case.

Before we derive the oracle property, we will first derive the following lemmas, which give the consistency and convergence rate of the estimator of $\boldsymbol{\beta}$ under different assumptions.

Lemma 3.3.2. ***(Sparsity)*** *Assume that regularity conditions (C1)-(C8) are satisfied and $\hat{\boldsymbol{\beta}}_{1n}$ is the first s-components of the $\sqrt{n}-$consistent estimator obtained*

in Theorem 1. Then with probability tending to 1,

$$\hat{\boldsymbol{\beta}}^{2n} = \arg \max_{\|\boldsymbol{\beta}^{2n}\| \leq cn^{-1/2}} \xi\{(\hat{\boldsymbol{\beta}}^{1n}, \boldsymbol{\beta}^{2n})\} = 0. \tag{3.29}$$

Proof. It is sufficient to show that with probability tending to 1 as $n \to +\infty$, for any $\boldsymbol{\beta}^{1n}$ satisfying $\boldsymbol{\beta}^{1n} - \boldsymbol{\beta}^{10} = O_P(n^{-1/2})$ and for some small $\epsilon_n = Cn^{-1/2}$ and $j = s+1, ..., p$,

$$\begin{aligned} \frac{\partial \xi(\boldsymbol{\beta})}{\partial \beta_j} < 0 \quad \text{for} \qquad 0 \quad < \beta_j < \quad \epsilon_n, \\ \frac{\partial \xi(\boldsymbol{\beta})}{\partial \beta_j} > 0 \quad \text{for} \quad -\epsilon_n \quad < \beta_j < \quad 0. \end{aligned} \tag{3.30}$$

Take $j \in \{1, \ldots, s\}$. $\beta_j \in (-\varepsilon_j, \varepsilon_j)$, $\beta_j ! = 0$, Taylor expansion gives

$$\begin{aligned} \frac{\partial \xi(\boldsymbol{\beta})}{\partial \beta_j} &= \frac{\partial \log LP_{t_n}(\boldsymbol{\beta})}{\partial \beta_j} - np'_{\lambda_n}(\beta_j)\text{sgn}(\beta_j) \\ &= \bigtriangledown [\log PL_{t_n}(\boldsymbol{\beta}^0)]_j + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^\top \bigtriangledown^2 [\log PL_{t_n}(\boldsymbol{\beta}^0)]_{\cdot j} \\ &\quad + \sum_{l=1}^{p} \sum_{k=1}^{p} \frac{\partial^3 log PL_{t_n}(\boldsymbol{\beta}^*)}{\partial \beta_j \partial \beta_j \partial \beta_k} \times (\beta_l^* - \beta_l^0)(\beta_k^* - \beta_l^k) \\ &\quad - np'_{\lambda_n}(\beta_j)\text{sgn}(\beta_j), \end{aligned} \tag{3.31}$$

where $\boldsymbol{\beta}^*$ is between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^0$, $[\log PL_{t_n}(\boldsymbol{\beta}^0)]_j$ is the jth element of $\log PL_{t_n}(\boldsymbol{\beta}^0)$, and $[\log PL_{t_n}(\boldsymbol{\beta}^0)]_{\cdot j}$ is the jth column of $\log PL_{t_n}(\boldsymbol{\beta}^0)$. According to (3.21), the first term is of order $O_P(\sqrt{n})$. Similarly, by (3.22) and regularity condition (C4),

$$\begin{aligned} (\boldsymbol{\beta} - \boldsymbol{\beta}^0)^\top \bigtriangledown^2 [\log PL_{t_n}(\boldsymbol{\beta}^0)] = (\boldsymbol{\beta} - \boldsymbol{\beta}^0)^\top n \left\{\Sigma(\boldsymbol{\beta}^0) + o_P(1)\right\} &= O_P(\sqrt{n}) \\ \sum_{l=1}^{p} \sum_{k=1}^{p} \frac{\partial^3 \log PL_{t_n}(\boldsymbol{\beta}^*)}{\partial \beta_j \partial \beta_j \partial \beta_k} \times (\beta_l^* - \beta_l^0)(\beta_k^* - \beta_l^k) = O_P(n\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^0\|^2) &= O_P(1) \end{aligned}$$

So combining the first three terms, we can rewrite (3.31) as

$$O_P(\sqrt{n}) - np'_{\lambda_n}(\beta_j)\text{sgn}(\beta_j) = n\lambda_n \left\{ O_P(\frac{1}{\sqrt{n}\lambda_n}) - \frac{1}{\lambda_n} p'_{\lambda_n}(\beta_j)\text{sgn}(\beta_j) \right\} \tag{3.32}$$

Since $\sqrt{n}\lambda_n \to +\infty$ as $n \to +\infty$, and $p'_{\lambda_n}(0+)/\lambda_n > 0$, the second term will domi-

nate the first term. Therefore $\dfrac{\partial \xi(\boldsymbol{\beta})}{\partial \beta_j}$ is of opposite sign from β_j, which concludes the proof. □

Theorem 3.3.3. ***(Oracle Property)*** *Suppose* $\boldsymbol{\beta}^{10} = \{\beta_1^0, \beta_2^0, ..., \beta_s^0\}$ *and let*

$$\boldsymbol{d}_n = \begin{pmatrix} p'_{\lambda_n}(|\beta_1^0|)\,\mathrm{sgn}(\beta_1^0) \\ \vdots \\ p'_{\lambda_n}(|\beta_s^0|)\,\mathrm{sgn}(\beta_s^0) \end{pmatrix}, \Sigma_n^P = \begin{pmatrix} p''_{\lambda_n}(|\beta_1^0|)\,\mathrm{sgn}(\beta_1^0) & & 0 \\ & \ddots & \\ 0 & & p''_{\lambda_n}(|\beta_s^0|)\,\mathrm{sgn}(\beta_s^0) \end{pmatrix}.$$

Assume $\sqrt{n}\boldsymbol{d}_n \to \mathbf{0}$, *each element of* Σ_n^P *is* O_1 *and regularity conditions (C1)-(C8) hold. Then the* $\sqrt{n}$ *consistent local maximizer* $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}^{1n}, \hat{\boldsymbol{\beta}}^{2n})^T$ *in Theorem 1 must satisfy:*

(a) Sparisty: $\hat{\boldsymbol{\beta}}^{2n} = \mathbf{0}$ *with probability tending to one*

(b) Asymptotic normality:

$$\sqrt{n}(\Sigma_1(\boldsymbol{\beta}^0) + \Sigma_n^P))\{\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10} + (\Sigma_1(\boldsymbol{\beta}^0) + \Sigma_n^P)^{-1}\boldsymbol{d}_n\} \xrightarrow{\mathcal{L}} N(0, \Sigma_1(\boldsymbol{\beta}^0)),$$

where $\Sigma_1(\boldsymbol{\beta}^0)$ *is the upper* $s \times s$ *submatrix of* $\Sigma_1(\boldsymbol{\beta}^0)$.

Proof. Part (a) follows directly from the Lemma 3.3.2. By Theorem 3.3.1, there exists a $\sqrt{n}$ consistent local maximizer $\hat{\boldsymbol{\beta}}^{1n}$ of $\xi\{(\boldsymbol{\beta}^1, \mathbf{0})^\top\}$, which will satisfy:

$$\left.\frac{\partial \xi(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}=(\hat{\boldsymbol{\beta}}^{1n}, \mathbf{0})^\top} = 0, \text{ for } j = 1, ..., s. \tag{3.33}$$

As in Equation (3.31),

$$\begin{aligned}
\left.\frac{\partial \xi(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}=(\hat{\boldsymbol{\beta}}^{1n}, \mathbf{0})^\top} &= \left.\frac{\partial \log LP_{t_n}(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}=(\hat{\boldsymbol{\beta}}^{1n}, \mathbf{0})^\top} - np'_{\lambda_n}(\beta_j)\mathrm{sgn}(\beta_j) \\
&= \bigtriangledown[\log PL_{t_n}(\boldsymbol{\beta}^0)]_j + (\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10})^\top \bigtriangledown^2 [\log PL_{t_n}(\boldsymbol{\beta}^0)]_{\cdot j} \\
&\quad + \sum_{l=1}^{p}\sum_{k=1}^{p} \frac{\partial^3 \log PL_{t_n}(\boldsymbol{\beta}^*)}{\partial\beta_j\partial\beta_j\partial\beta_k} \times (\beta_l^* - \beta_l^0)(\beta_k^* - \beta_k^0) \\
&\quad - \left[np'_{\lambda_n}(\beta_j^0)\mathrm{sgn}(\beta_j) + \{p''_{\lambda_n}(\beta_j^0) + o_P(1)\}(\hat{\beta}_j - \beta_j^0)\right] \\
&= \sqrt{n}\left\{\frac{1}{\sqrt{n}} \bigtriangledown [\log PL_{t_n}(\boldsymbol{\beta}^0)]_j\right\} + (\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10})^\top n \left\{[\Sigma_1(\boldsymbol{\beta}^0)]_{\cdot j} + o_P(1)\right\}
\end{aligned}$$

$$- \left[np'_{\lambda_n}(\beta_j^0)\text{sgn}(\beta_j) + \{p''_{\lambda_n}(\beta_j^0) + o_P(1)\}(\hat{\beta}_j - \beta_j^0) \right]. \quad (3.34)$$

By (3.21) and Slutsky's theorem, we obtain

$$\sqrt{n}(\Sigma_1(\boldsymbol{\beta}^0) + \Sigma_n^P))\{\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10} + (\Sigma_1(\boldsymbol{\beta}^0) + \Sigma_n^P)^{-1}\boldsymbol{d}_n\} \xrightarrow{\mathcal{L}} N(0, \Sigma_1(\boldsymbol{\beta}^0)).$$

□

3.3.2 Network with Single Sender and Multiple Receivers

Recall that if there is a single sender and multiple receivers at each event time, the observations will be $\{(i_m, t_m, J_m), m = 1, ..., n.\}$, where J_m is the receiver set. As discussed in previous sections, the original partial likelihood function will be difficult to implement. Therefore, the approximated version will be used. For sufficient statistic selection, a penalized approximated partial likelihood function is maximized. We will use the results of Section 3.2.2 to get the desired properties of the estimator obtained in this way.

Suppose all the assumptions and regularity conditions are the same as in single receiver situation. Assume additionally

(C9) There exists an open subset ω which contains the true parameter point $\boldsymbol{\beta}^0$ such that the partial likelihood function admits all third derivatives. Furthermore, the third-order derivative is bounded in probability:

$$\frac{1}{n}\left|\frac{\partial^3 \log \widetilde{PL}_{t_n}(\boldsymbol{\beta})}{\partial\beta_i\partial\beta_j\partial\beta_k}\right| = O_p(1), i, j, k \in \{1, 2, ..., p\}$$

Similarly, we first derive a form of expansion of the approximated log partial likelihood function. For $\boldsymbol{\beta}$ in a neighborhood of the true parameter $\boldsymbol{\beta}^0$, the Taylor expansion of the log partial likelihood function $\log \widetilde{PL}_{t_n}(\boldsymbol{\beta})$ around $\boldsymbol{\beta}^0$ is

$$\begin{aligned}
\log \widetilde{PL}_{t_n}(\boldsymbol{\beta}) &= \log \widetilde{PL}_{t_n}(\boldsymbol{\beta}^0) + (\boldsymbol{\beta} - \boldsymbol{\beta}^0)^T \bigtriangledown [\log \widetilde{PL}_{t_n}(\boldsymbol{\beta}^0)] \\
&\quad + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^T \bigtriangledown^2 [\log \widetilde{PL}_{t_n}(\boldsymbol{\beta}^0)](\boldsymbol{\beta} - \boldsymbol{\beta}^0) \\
&\quad + \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p} \frac{\partial^3 \log \widetilde{PL}_{t_n}(\boldsymbol{\beta}^*)}{\partial\beta_i\partial\beta_j\partial\beta_k}(\beta_i^* - \beta_i^0)(\beta_j^* - \beta_j^0)(\beta_k^* - \beta_k^0)
\end{aligned}$$

$$
\begin{aligned}
= \quad & \log \widetilde{PL}_{t_n}(\boldsymbol{\beta}^0) + (\boldsymbol{\beta} - \boldsymbol{\beta}^0)^T \bigtriangledown [\log \widetilde{PL}_{t_n}(\boldsymbol{\beta}^0)] \\
& + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^T \bigtriangledown^2 [\log \widetilde{PL}_{t_n}(\boldsymbol{\beta}^0)](\boldsymbol{\beta} - \boldsymbol{\beta}^0) \\
& + O_P(n\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^0\|^3),
\end{aligned}
$$

where $\boldsymbol{\beta}^*$ is between $\boldsymbol{\beta}^0$ and $\boldsymbol{\beta}$. The following two theorems are extensions of Theorem 3.3.1 and Theorem 3.3.3 to the multiple receivers scenario.

Theorem 3.3.4. ***(Consistency)*** *Assume we have observed n interactions* $\{(i_m, t_m, J_m), m = 1, ..., n.\}$ *from a network and* $\widetilde{PL}_{t_n}(\boldsymbol{\beta})$ *denotes the approximated partial likelihood function given the observations. Define*

$$
\xi(\boldsymbol{\beta}) = \log \widetilde{PL}_{t_n}(\boldsymbol{\beta}) - n \sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|) \tag{3.35}
$$

as the penalized partial likelihood function. Let a_n *and* b_n *be defined as in condition (C5). Assume regularity conditions (C1)-(C9) hold and the minimum eigenvalue for* $\Sigma_1(\boldsymbol{\beta})$, λ_{min} *is bounded away from zero. Finally, assume the receiver growth sequence defined in equation (3.16) satisfies* $G_n = O(\sqrt{n})$. *Then there exists a local maximizer* $\hat{\boldsymbol{\beta}}$ *of* $\xi(\boldsymbol{\beta})$ *such that* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| = O_p(n^{-1/2} + a_n)$.

Proof. Let $\alpha_n = n^{-1/2} + a_n$. It is sufficient to show that for any $\varepsilon > 0$, there exist a large constants C such that

$$
\lim_{n\to\infty} P\left(\sup_{\|\mathbf{u}\|=C} \xi(\alpha_n \mathbf{u} + \boldsymbol{\beta}^0) < \xi(\boldsymbol{\beta}^0) \right) \geq 1 - \varepsilon \text{ for all } n > N.
$$

Let $D_n(\boldsymbol{u}) = \xi(\alpha_n \boldsymbol{u} + \boldsymbol{\beta}^0) - \xi(\boldsymbol{\beta}^0)$. Then

$$
\begin{aligned}
D_n(\boldsymbol{u}) \quad &= \quad \log \widetilde{PL}_{t_n}(\boldsymbol{\beta}^0 + \alpha_n \mathbf{u}) - \log \widetilde{PL}_{t_n}(\boldsymbol{\beta}^0) - n \sum_{j=1}^{p} p_{\lambda_n}(|\beta_j^0 + \alpha_n u_j|) + n \sum_{j=1}^{p} p_{\lambda_n}(|\beta_j^0|) \\
&\leq \quad \underbrace{\log \widetilde{PL}_{t_n}(\boldsymbol{\beta}^0 + \alpha_n \mathbf{u}) - \log \widetilde{PL}_{t_n}(\boldsymbol{\beta}^0)}_{\text{A}} \underbrace{- n \sum_{j=1}^{s} \left[p_{\lambda_n}(|\beta_j^0 + \alpha_n u_j|) - p_{\lambda_n}(|\beta_j^0|) \right]}_{\text{B}}.
\end{aligned}
$$

We consider A and B separately. For any sequence $\tilde{\boldsymbol{\beta}}_n \xrightarrow{P} \boldsymbol{\beta}^0$, we have the

Taylor expansion ($\tilde{\boldsymbol{\beta}}_n^*$ is between $\tilde{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}^0$),

$$\begin{aligned}\log \widetilde{PL}_{t_n}(\tilde{\boldsymbol{\beta}}) &= \log \widetilde{PL}_{t_n}(\boldsymbol{\beta}^0) + (\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0)^T \bigtriangledown [\log \widetilde{PL}_{t_n}(\boldsymbol{\beta}^0)] \\ &\quad + \frac{1}{2}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0)^T \bigtriangledown^2 [\log \widetilde{PL}_{t_n}(\boldsymbol{\beta}^0)](\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^0) + O_P(n\|\tilde{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}^0\|^3)\end{aligned}$$

Then for $\tilde{\boldsymbol{\beta}}_n = \boldsymbol{\beta}^0 + \alpha_n \boldsymbol{u}$, we obtain

$$\begin{aligned}\mathrm{A} &= \alpha_n \boldsymbol{u}^T \bigtriangledown [\log \widetilde{PL}_{t_n}(\boldsymbol{\beta}^0)] + \frac{\alpha_n^2}{2} \boldsymbol{u}^T \bigtriangledown^2 [\log \widetilde{PL}_{t_n}(\boldsymbol{\beta}^0)] \boldsymbol{u} + O_P(n\alpha_n^3 \|\boldsymbol{u}\|^3) \\ &= \alpha_n \boldsymbol{u}^T [\bigtriangledown \log PL_{t_n}(\boldsymbol{\beta}^0) + O_P(\sqrt{n})] + \frac{\alpha_n^2}{2} \boldsymbol{u}^T [\bigtriangledown^2 \log PL_{t_n}(\boldsymbol{\beta}^0) + O_P(\sqrt{n})] \boldsymbol{u} + O_P(n\alpha_n^3 \|\boldsymbol{u}\|^3) \\ &= \sqrt{n}\alpha_n \boldsymbol{u}^T \left\{ \frac{1}{\sqrt{n}} \bigtriangledown [\log PL_{t_n}(\boldsymbol{\beta}^0)] + O_P(1) \right\} - \frac{n\alpha_n^2}{2} \boldsymbol{u}^T \left\{ -\frac{1}{n} \bigtriangledown^2 [\log PL_{t_n}(\boldsymbol{\beta}^0)] + o_P(1) \right\} \boldsymbol{u} \\ &\quad + O_P(n\alpha_n^3) \\ &= \sqrt{n}\alpha_n \boldsymbol{u}^T \left\{ \frac{1}{\sqrt{n}} \bigtriangledown [\log PL_{t_n}(\boldsymbol{\beta}^0)] + O_P(1) \right\} - \frac{n\alpha_n^2}{2} \boldsymbol{u}^T \left\{ \Sigma(\boldsymbol{\beta}^0) + o_P(1) \right\} \boldsymbol{u} \\ &\quad + O_P(n\alpha_n^3) \\ &\leq \sqrt{n}\alpha_n \|\boldsymbol{u}\| Z_n - \frac{n\alpha_n^2}{2} \|\boldsymbol{u}\|^2 \left\{ \lambda_{\min} + o_P(1) \right\} + O_P(n\alpha_n^3) \qquad (3.36)\end{aligned}$$

where $Z_n = \| \bigtriangledown [\log PL_{t_n}(\boldsymbol{\beta}^0)] + O_P(1)\| = O_P(1)$. The Part B is the same as in (3.25). Therefore, similarly as in (3.26)

$$\begin{aligned}\sup_{\|\mathbf{u}\|=C} D_n(\boldsymbol{u}) &= \sup_{\|\mathbf{u}\|=C} \{A - B\} \leq \sup_{\|\mathbf{u}\|=C} A + \sup_{\|\mathbf{u}\|=C} |B| \\ &= n\alpha_n^2 \left\{ C \left[\frac{Z_n}{\sqrt{n}\alpha_n} + \sqrt{s}\frac{a_n}{\alpha_n} \right] - \frac{C^2}{2} \left[\lambda_{\min} + o_P(1) \right] + o_P(1) \right\}\end{aligned}$$

Similarly, as proved in Theorem 3.3.1, $D_n + o_P(1) = O_P(1)$ will leads to the conclusion that there exists a large number C, such that

$$\lim_{n\to\infty} P(\sup_{\|\mathbf{u}\|=C} D_n(\boldsymbol{u}) < 0) \geq 1 - \varepsilon$$

□

Lemma 3.3.5. ***(Sparsity)*** *Assume that regularity conditions (C1)-(C9) are satisfied and $\hat{\boldsymbol{\beta}}_{1n}$ is the first s components of the $\sqrt{n}-$consistent estimator obtained*

in Theorem 1. Then with probability tending to 1,

$$\hat{\boldsymbol{\beta}}^{2n} = \arg \max_{\|\boldsymbol{\beta}^{2n}\| \leq cn^{-1/2}} \xi\{(\hat{\boldsymbol{\beta}}^{1n}, \boldsymbol{\beta}^{2n})\} = 0. \tag{3.37}$$

Proof. It is sufficient to show that with probability tending to 1 as $n \to +\infty$, for any $\boldsymbol{\beta}^{1n}$ satisfying $\boldsymbol{\beta}^{1n} - \boldsymbol{\beta}^{10} = O_P(n^{-1/2})$ and for some small $\epsilon_n = Cn^{-1/2}$ and $j = s+1, ..., p$

$$\begin{aligned} \frac{\partial \xi(\boldsymbol{\beta})}{\partial \beta_j} < 0 \quad \text{for} \qquad 0 \quad < \beta_j < \quad \epsilon_n, \\ \frac{\partial \xi(\boldsymbol{\beta})}{\partial \beta_j} > 0 \quad \text{for} \quad -\epsilon_n \quad < \beta_j < \quad 0. \end{aligned} \tag{3.38}$$

Take $j \in \{1, \ldots, s\}$. For $\beta_j \in (-\varepsilon_n, \varepsilon_n)$, $\beta_j ! = 0$, Taylor expansion gives

$$\begin{aligned} \frac{\partial \xi(\boldsymbol{\beta})}{\partial \beta_j} &= \frac{\partial \log \widetilde{LP}_{t_n}(\boldsymbol{\beta})}{\partial \beta_j} - n p'_{\lambda_n}(\beta_j) \operatorname{sgn}(\beta_j) \\ &= \bigtriangledown [\log \widetilde{LP}_{t_n}(\boldsymbol{\beta}^0)]_j + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^\top \bigtriangledown^2 [\log \widetilde{LP}_{t_n}(\boldsymbol{\beta}^0)]_{\cdot j} \\ &\quad + \sum_{l=1}^{p} \sum_{k=1}^{p} \frac{\partial^3 \log \widetilde{LP}_{t_n}(\boldsymbol{\beta}^*)}{\partial \beta_j \partial \beta_j \partial \beta_k} \times (\beta_l^* - \beta_l^0)(\beta_k^* - \beta_l^k) \\ &\quad - n p'_{\lambda_n}(\beta_j) \operatorname{sgn}(\beta_j) \\ &= \{\bigtriangledown [\log LP_{t_n}(\boldsymbol{\beta}^0)]_j + O_P(\sqrt{n})\} + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^\top \{\bigtriangledown^2 [\log LP_{t_n}(\boldsymbol{\beta}^0)]_{\cdot j} + O_P(\sqrt{n})\} \\ &\quad + \sum_{l=1}^{p} \sum_{k=1}^{p} \frac{\partial^3 \log \widetilde{LP}_{t_n}(\boldsymbol{\beta}^*)}{\partial \beta_j \partial \beta_j \partial \beta_k} \times (\beta_l^* - \beta_l^0)(\beta_k^* - \beta_l^k) \\ &\quad - n p'_{\lambda_n}(\beta_j) \operatorname{sgn}(\beta_j), \end{aligned} \tag{3.39}$$

where $\boldsymbol{\beta}^*$ is between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^0$, $[\log PL_{t_n}(\boldsymbol{\beta}^0)]_j$ is the jth element of $\log PL_{t_n}(\boldsymbol{\beta}^0)$, and $[\log PL_{t_n}(\boldsymbol{\beta}^0)]_{\cdot j}$ is the jth column of $\log PL_{t_n}(\boldsymbol{\beta}^0)$. According to the convergence of the score function, the first term is of order $O_P(\sqrt{n})$. Similarly, by regularity condition (C4),

$$\begin{aligned} (\boldsymbol{\beta} - \boldsymbol{\beta}^0)^\top \{\bigtriangledown^2 [\log PL_{t_n}(\boldsymbol{\beta}^0)] + O_P(\sqrt{n})\} = (\boldsymbol{\beta} - \boldsymbol{\beta}^0)^\top n \left\{\Sigma(\boldsymbol{\beta}^0) + o_P(1)\right\} &= O_P(\sqrt{n}), \\ \sum_{l=1}^{p} \sum_{k=1}^{p} \frac{\partial^3 log \widetilde{PL}_{t_n}(\boldsymbol{\beta}^*)}{\partial \beta_j \partial \beta_j \partial \beta_k} \times (\beta_l^* - \beta_l^0)(\beta_k^* - \beta_l^k) = O_P(n \|\boldsymbol{\beta}^* - \boldsymbol{\beta}^0\|^2) &= O_P(1). \end{aligned}$$

Combining the first three terms, we can rewrite (3.39) as

$$O_P(\sqrt{n}) - np'_{\lambda_n}(\beta_j)\text{sgn}(\beta_j) = n\lambda_n \left\{ O_P(\frac{1}{\sqrt{n}\lambda_n}) - \frac{1}{\lambda_n} p'_{\lambda_n}(\beta_j)\text{sgn}(\beta_j) \right\} \quad (3.40)$$

Since $\sqrt{n}\lambda_n \to +\infty$ as $n \to +\infty$ and $p'_{\lambda_n}(0+)/\lambda_n > 0$, the second term will dominate the first term. Therefore $\dfrac{\partial \xi(\boldsymbol{\beta})}{\partial \beta_j}$ is of opposite sign from β_j, which concludes the proof. □

Theorem 3.3.6. ***(Oracle Property)*** *Suppose* $\boldsymbol{\beta}^{10} = \{\beta_1^0, \beta_2^0, ..., \beta_s^0\}$ *and let*

$$\boldsymbol{d}_n = \begin{pmatrix} p'_{\lambda_n}(|\beta_1^0|)\,\text{sgn}(\beta_1^0) \\ \vdots \\ p'_{\lambda_n}(|\beta_s^0|)\,\text{sgn}(\beta_s^0) \end{pmatrix}, \Sigma_n^P = \begin{pmatrix} p''_{\lambda_n}(|\beta_1^0|)\,\text{sgn}(\beta_1^0) & & 0 \\ & \ddots & \\ 0 & & p''_{\lambda_n}(|\beta_s^0|)\,\text{sgn}(\beta_s^0) \end{pmatrix}.$$

Assume $\sqrt{n}\boldsymbol{d}_n \to \mathbf{0}$, *each element of* $\Sigma_n^P =$ *is O(1), regularity conditions (C1)-(C9) hold, and*

$$G_n = O_P(\sqrt{n})$$

Then the $\sqrt{n}$ *consistent local maximizer* $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}^{1n}, \hat{\boldsymbol{\beta}}^{2n})^T$ *in Theorem 1 must satisfy:*

(a) Sparisty: $\hat{\boldsymbol{\beta}}^{2n} = \mathbf{0}$ *with probability tending to one*

(b) Asymptotic normality:

$$\sqrt{n}(\Sigma_1(\boldsymbol{\beta}^0) + \Sigma_n^P))\{\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10} + (\Sigma_1(\boldsymbol{\beta}^0) + \Sigma_n^P)^{-1}\boldsymbol{d}_n\} \xrightarrow{\mathcal{L}} N(0, \Sigma_1(\boldsymbol{\beta}^0)),$$

where $\Sigma_1(\boldsymbol{\beta}^0)$ *is the upper* $s \times s$ *submatrix of* $\Sigma_1(\boldsymbol{\beta}^0)$.

Proof. Part (a) follows directly from the Lemma 3.3.5. Now we start to prove part (b). By Theorem 3.3.4, there exist a $\sqrt{n}$ consistent local maximizer $\hat{\boldsymbol{\beta}}^{1n}$ of $\xi\{(\boldsymbol{\beta}^1, \mathbf{0})^\top\}$, which will satisfy:

$$\left.\frac{\partial \xi(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}=(\hat{\boldsymbol{\beta}}^{1n}, \mathbf{0})^\top} = 0, \text{ for } j = 1, ..., s. \quad (3.41)$$

If $G_n = O_P(n^{1/2})$, then

$$\frac{1}{\sqrt{n}} \bigtriangledown [\log \widetilde{LP}_{t_n}(\boldsymbol{\beta}^0)] = \frac{1}{\sqrt{n}} \bigtriangledown [\log LP_{t_n}(\boldsymbol{\beta}^0)] + o_P(1) \tag{3.42}$$

Again, by (3.42) and Taylor expansion,

$$\begin{aligned}
\left.\frac{\partial \xi(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}=(\hat{\boldsymbol{\beta}}^{1n},\mathbf{0})^\top} &= \left.\frac{\partial \log \widetilde{LP}_{t_n}(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}=(\hat{\boldsymbol{\beta}}^{1n},\mathbf{0})^\top} - np'_{\lambda_n}(\beta_j)\,\mathrm{sgn}(\beta_j) \\
&= \bigtriangledown[\log \widetilde{LP}_{t_n}(\boldsymbol{\beta}^0)]_j + (\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10})^\top \bigtriangledown^2 [\log \widetilde{LP}_{t_n}(\boldsymbol{\beta}^0)]_{\cdot j} \\
&\quad + \sum_{l=1}^{p}\sum_{k=1}^{p} \frac{\partial^3 \log \widetilde{LP}_{t_n}(\boldsymbol{\beta}^*)}{\partial\beta_j\partial\beta_j\partial\beta_k} \times (\beta_l^* - \beta_l^0)(\beta_k^* - \beta_k^0) \\
&\quad - \left[np'_{\lambda_n}(\beta_j^0)\,\mathrm{sgn}(\beta_j) + \{p''_{\lambda_n}(\beta_j^0) + o_P(1)\}(\hat{\beta}_j - \beta_j^0)\right] \\
&= \sqrt{n}\left\{\frac{1}{\sqrt{n}} \bigtriangledown [\log PL_{t_n}(\boldsymbol{\beta}^0)]_j + o_P(1)\right\} + (\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10})^\top n\left\{[\Sigma_1(\boldsymbol{\beta}^0)]_{\cdot j} + o_P(1)\right\} \\
&\quad - \left[np'_{\lambda_n}(\beta_j^0)\,\mathrm{sgn}(\beta_j) + \{p''_{\lambda_n}(\beta_j^0) + o_P(1)\}(\hat{\beta}_j - \beta_j^0)\right]
\end{aligned} \tag{3.43}$$

Finally, by Slutsky's theorem, we obtain

$$\sqrt{n}(\Sigma_1(\boldsymbol{\beta}^0) + \Sigma_n^P))\{\hat{\boldsymbol{\beta}}^{1n} - \boldsymbol{\beta}^{10} + (\Sigma_1(\boldsymbol{\beta}^0) + \Sigma_n^P)^{-1}\boldsymbol{d}_n\} \xrightarrow{\mathcal{L}} N(0, \Sigma_1(\boldsymbol{\beta}^0)).$$

□

3.4 Discussion and Extension

In this chapter, we first discuss properties of the maximum partial likelihood estimators in an egocentric approach of Vu et al. (2011b) for networks with no tied event. We also consider approximation in the likelihood function to derive estimates in networks with tied events. These results are not limited to egocentric setting only. One generalization of the covariates selection work would be to the relational model in Perry and Wolfe (2013) since the partial likelihood and its approximation have been well-studied in Perry and Wolfe (2013) and our proofs for the penalized version do not rely on anything specific to the egocentric approach. The major differences will lie in the computational implementation part. There are

also many possibilities to extend the proofs here since they do not use the specific form of the partial likelihood function. However, there might be different computational challenges associated with different models. Some of these challenges will be discussed later in Chapter 5 when we consider future works. In addition, regularity conditions for theories of penalized estimates includes convergence rate for the tuning parameter λ. And ones used in this chapter are suitable for SCAD and other similar penalty functions, but not the adaptive LASSO penalty. However, slightly modification of these regularity conditions may yield the same properties of the penalized estimates. And this won't impact the computation implementation in the next chapter for these theories.

In the next chapter, we work on computational implementations of the theories described here in a specific citation network (Vu et al., 2011b). The covariates include 7 network structure covariates and a 50-dimensional LDA covariates (see Chapter 1). The theory requires that the covariates are uniformly square integrable (Condition (C1)). We now argue that it actually impose the sparsity on the network. Since the LDA structure covariates are bounded, so the restriction only applies on the network structure covariates. We assume that each paper can only cite a limit number of other papers. Therefore, all the out-degrees are bounded. Assuming in-degree related covariates square integrable restricts the probability that it goes to infinity, and can be think as a way to assume that the network is sparse. In Chapter 4, the effectiveness of the case control approximation also depends on this sparsity. In addition, we also investigate the sparsity in the coefficients of the covariates. So, in next Chapter, we will study the computation implementation of these two types of sparsities, and use them to reduce the computational complexity in finite sample predictions.

Chapter 4

Implementation of Variable Selection for Dynamic Network Models

4.1 Introduction

In this chapter, we will discuss several issues related to implementing computational algorithms for the egocentric dynamic network models described in Chapter 3, including parameter estimation and variable selection. We will focus a data set for a citation network, which is called "arXiv-TH" in Vu et al. (2011b). The data set records the citation network of arXiv high energy physics theory articles spanning from January 1993 to April 2003. It contains 29,557 papers (nodes) and 352,807 citation events (edges). The time stamps when papers join the network are also recorded: these time stamps are also treated as the times when the citation events happen. In addition, a fifty-dimensional topic vector is extracted from each paper using the Latent Dirichlet Model (Blei et al., 2003), and these vectors are used to construct the LDA covariates.

In a citation network, multiple citation events (edges) may be established simultaneously. To relieve the computational burden of working with the exact partial likelihood when multiple tied event times are present, in all of the following estimations, we will use approximation (3.15) of the partial likelihood function. Vu et al. (2011b) have developed a C++ program to calculate the maximum approximated partial likelihood estimator. Here, we first create an R package wrapper

for the C++ program to improve its usability. Then similar calculations are re-implemented in R to prepare further extension of these ideas as in Chapter 3.

In this framework, we are interested not only in parameter estimation but also in how well our fitted model predicts new edges in the network. The dynamic setting makes it possible to test prediction by splitting data into separate training and testing phases, as we describe below in more detail. The covariates considered in this chapter include not only several network structure variables, but also the LDA-based topic-similarity variables, which can improve the prediction performance effectively. However, the computational cost for the model including the LDA topic covariates increases fast as the size of the network grows. Therefore, we further consider a case-control approximation of the partial likelihood function to reduce the computation complexity. In the second half of this chapter,, we apply variable selection algorithms to select covariables for this network and evaluate the prediction performance of models before and after variable selection.

4.2 Partial Likelihood Approximation and Estimation

This section describes the computational aspects of approximating and maximizing the partial likelihood (3.15) corresponding to model for a dataset with multiple tied edge times.

4.2.1 R Package "ego" for the Maximum Partial Likelihood Estimates

In Vu et al. (2011b), the maximum partial likelihood estimator is calculated by a C++ program using part or all of the citation network. The inputs include two time points, in between which are the citation network observations used for estimation. The outputs are the maximum partial likelihood estimates for the coefficients of the time-varying covariates and their estimated covariance matrix. To avoid repeatedly calculating parts of the partial likelihood function, Vu et al. propose the idea of "caching", which will be discussed later in this section. The Newton-Raphson algorithm is used for the optimization. To make this program

available for R users, we first incorporate this codes in an R package named "ego", which can call the C++ program and return the output to R.

The main challenge is that the C++ program calls an external "boost" library for matrix calculation, and we need to indicate where to locate this library in the configure file when we install the R package. With the uncompiled source code, one can run the following to install the package "ego.tar.gz":

```
R CMD INSTALL
--configure-args="--with-boost-include=/where/is/the/boost/library"
ego.tar.gz
```

Alternatively, from within R one may use

```
install.packages("ego",configure.args="--with-boost-include
=/where/is/your/boost/library")
```

to obtain parameter estimates in R, one needs the arXiv-TH dataset. This dataset is saved in a text file in the following format

```
...
540 62
548 392 417
549 392 477 519
...
```

The first number in each row is the label of the sender (citer), and each label following the first is cited by the sender. The senders are labeled in increasing order of the time they join the network; thus, in principle all receivers of citations should have smaller numerical labels than the senders. However, this rule is sometimes violated in practice, though in the arXiv-TH dataset the only violation is a single article that cites only itself (which we omit from the analysis). Besides, another file associated with the network records the time when each paper are published. The part related to above segment of the network is as follows,

```
...
540 8203.99042824074 708824773.000 17-Jun-1992 19:46:13
541 8204.406226851852 708860698.000 18-Jun-1992 05:44:58
```

```
542 8204.766354166666 708891813.000 18-Jun-1992 14:23:33
...
```

The first column is the node labels. The rest columns are time in different format, the second is the time in days and the third in seconds. In the C++ program, only the second column is used as the time.

In R, from the directory where the dataset is located, parameter estimates may obtained by typing

```
library(ego)
egonet(type-of-model, start-time, end-time, signature)
```

The "type-of-model" can be 0 to 5, indicating which covariates to include in the model. For example, 0 uses the preferential attachments covariates $s^{PA1}(t)$ (see Chapter 1 equation (1.72)) only; 1 uses all the network structure covariates except for the Recency-based first-order PA $s^{Rec-PA1}(t)$ and LDA (see Chapter 1 equation (1.74)); 2 uses all the network structure covariates except for the LDA; 3 uses the LDA and all other covariates except for $s^{Rec-PA1}(t)$; 4 uses LDA covariates only; 5 uses the LDA and all other covariates. The "start-time" and "end-time" will locate the citation events to use in the model estimation, as in the file above. And "signature" gives a signature for the output file in the original C++ program, which will not be useful in R. For example,

```
egonet(0, 11464, 11585, 1000)
```

will construct a network model with only the preferential attachment covariates, and use citation events between time 11464 and 11585. The output in R looks like

```
----------------FINAL RESULT----------------
Time of Newton-Raphson with LBF: 0.014326 seconds
Computing training log-likelihood
Time of training log-likelihood computation with LBF: 0.001656 seconds
training logLLH = -178738.260085
Beta Estimates:
 0.00380700474465828741
```

```
Beta Covariance:
0.00000000051143505007
```

Despite the multiple models implemented by the C++ code and its efficient implementation resulting in fast computational speed, we find that it is easier to re-implement these models in R in order to allow for flexibility we need in order to extend the capabilities of the code. Therefore we describe implementation of the same algorithm using R in the next subsection.

4.2.2 Calculating the Maximum Partial Likelihood Estimates in R

Here, we discuss computation implementation in R of the models with network structure covariates and the LDA covariates. We first consider the network structure covariates. Network structure covariates take only integer values and accumulate with time, so they are nondecreasing. If we assume the receiver is bounded, they are still bounded, though. Yet, this assumption is not very practical for citation networks, and we will discuss more about it in Chapter 5.

We fist discuss how to store the network efficiently. One possibility is to store the network as a lower triangular matrix, in which (i, j)-th element equal to one means that there is an edge from node i to node j when node i enter the network $(i > j)$. However, this usually requires a very large memory space and it is not affordable for large networks. To track change of the time-varying covariates at each unique event time, another possibility is to read the network file line by line and store the updates of the covariates values at each unique event time. This idea is workable, yet to update some of the network structure covariates it is not efficient because it may involves previous events as well. For example, to update the second order out-degree $s_k^{PA2}(t)$ for a node k, we need to go back and track the number of receivers (citations) for all the receivers of k (papers cited by k). This requires going back and forth between previous rows and the current row of node k. To avoid this inefficiency, we can store the network in some edgelist structure, then use these edgelists to update the values of the network structure covariates.

The edgelist structure consists of four parts, which help to quickly locate a certain node in the original network list. The four parts are “out.edgelist”, “out.node”,

"in.edgelist", and "in.node". The "out.edgelist" stacks the senders and receivers of all the citation events, where the first column are the senders and the second column are the receivers. If there are multiple receivers at the same unique event time, they will be listed in increasing order of the receivers' labels. The "out.node" tells in which row a certain sender starts in the out.edgelist. For example, the following are the corresponding parts of "out.edgelist" and "out.node" for the part of network in the previous section. Node 540 starts from row 40 in the "out.edgelist", so in the "out.node" list, the number in the second column for 540 is 40.

```
$out.edgelist
...
 [40,]  540   62
 [41,]  548  392
 [42,]  548  417
 [43,]  549  392
 [44,]  549  477
 [45,]  549  519
 ...
```

```
$out.node
...
 [33,]  540   40
 [34,]  548   41
 [35,]  549   43
 ...
```

Similarly, the "in.edgelist" and the "in.node" store the same information for receivers. The first column in "in.edgelist" are receivers, and the second column are the corresponding senders. A node may be a receiver for different senders at different event times, and these events are listed in the increasing order of the senders' labels. For example, node 62 is a receiver for nodes 526, 540, and 559, so they are listed by the sender order in the "in.edgelist". Similarly to "out.node", the "in.node" helps to locate row indices of the receivers in the "in.edgelist". If using the network example in the previous section, part of the "in.edgelist" and "in.node" are as follows:

```
$in.edgelist
 ...
 [51,]   62  526
 [52,]   62  540
 [53,]   62  559
 ...
```

```
$in.node
  ...
[19,]   62   51
...
[57,]  392  121
...
```

```
[121,]  392  477
[122,]  392  519
[123,]  392  538
[124,]  392  548
...
```

To work with some part of the network, one only needs to construct these edgelist structures, and all the later estimation can be based on these edgelists solely. They also benefit the process of tracking the updates of the network structure covariates. For example, suppose the current node is 540. If we want to know the update of the second-order out-degree for node 540 after it comes in, we can use the "out.edgelist" and "out.node". From the out.edgelist, we know that 540 has an edge to node 62. Then we can locate node 62 in the out.node list, and quickly figure how many edges it sends, which is the increase of the second-order out-degree for node 540. Similarly, the "in.edgelist" and the "in.node" are useful for updating the in-degree covariates. And all of them are used to update the triangular effect covariates.

To optimize the approximated partial log-likelihood function using a Newton-Raphson algorithm, we need to evaluate the partial log-likelihood function (3.15), along with its gradient and Hessian matrix. These calculations, according to Equations (3.8), (3.9) and (3.10), require that the following terms be evaluated at each unique event time t_m:

$$S_0(\boldsymbol{\beta}, t_m) = \sum_{j \in \mathcal{J}_{t_m}} \exp\{\boldsymbol{\beta}^\top \boldsymbol{s}_j(t_m)\}, \tag{4.1}$$

$$S_1(\boldsymbol{\beta}, t_m) = \sum_{j \in \mathcal{J}_{t_m}} \exp\{\boldsymbol{\beta}^\top \boldsymbol{s}_j(t_m)\} \boldsymbol{s}_j(t_m), \tag{4.2}$$

$$S_2(\boldsymbol{\beta}, t_m) = \sum_{j \in \mathcal{J}_{t_m}} \exp\{\boldsymbol{\beta}^\top \boldsymbol{s}_j(t_m)\} \boldsymbol{s}_j^{\otimes 2}(t_m). \tag{4.3}$$

There are two potential problems when these terms are calculated. First, the at-risk set $\mathcal{J}_{t_m}$ can become large as t_m increases, since we assume all nodes entered prior to t_m are at risk at t_m. Therefore, the number of calculations for the summation over $\mathcal{J}_{t_m}$ can also be large. However, at each unique event time, the network structure covariates will change only for a small number of nodes. So adopting

the "caching" technique in Vu et al. (2011b) can save a lot of computational time compared to direct calculation. To implement caching, we first generate a list called "stat.update", which stacks all the changes on the covariates and their corresponding nodes' labels at each event time. The following is an example of such a list. The first column is the time, second column tells which node has changed, the third column tells which covariates has a change, and the last columns gives the amount of the change.

```
time node stat change
...
 33  540    3      1
 33   62    1      1
 34  548    3      2
 34  392    1      1
 34  417    1      1
...
```

Using these technique, we only need to update the terms which have a change in the covariates in $S_0(\boldsymbol{\beta}, t_m)$, $S_1(\boldsymbol{\beta}, t_m)$, and $S_2(\boldsymbol{\beta}, t_m)$ at a given event time t_m. Taking the update of $S_0(\boldsymbol{\beta}, t_m)$ as an example, the formula is

$$\begin{aligned} S_0(\boldsymbol{\beta}, t_{m+1}) &= \sum_{j \in \mathcal{J}_{t_m}} \exp\{\boldsymbol{\beta}^\top \boldsymbol{s}_j(t_m)\} + \sum_{j \in \mathcal{J}_{t_{m+1}} / \mathcal{J}_{t_m}} \exp\{\boldsymbol{\beta}^\top \boldsymbol{s}_j(t_{m+1})\} \\ &+ \sum_{j \in \mathcal{C}_{t_m}} \left[\exp\{\boldsymbol{\beta}^\top \boldsymbol{s}_j(t_{m+1})\} - \exp\{\boldsymbol{\beta}^\top \boldsymbol{s}_j(t_m)\}\right] \end{aligned} \tag{4.4}$$

Vu et al. (2011b), where $\mathcal{C}_{t_m}$ includes all nodes whose network covariates have changes in the interval $[t_m, t_{m+1})$. Similar update equations can be used for $S_1(\boldsymbol{\beta}, t_m)$ and $S_2(\boldsymbol{\beta}, t_m)$.

The second potential problem arises when the covariates are accumulated for a long time and yield a large value. Then the $\exp\{\boldsymbol{\beta}^\top \boldsymbol{s}_j(t_m)\}$ values may exceed the upper (or lower) bound that R can handle, so the computer will recognize then as infinity (or zero) during calculation. This can be resolved by adding one step after obtaining all the covariates update. We search the maximum value of covariates at each unique event time and then re-scale all the covariates to $\boldsymbol{s}_j(t_m) - \boldsymbol{s}_{\max}(t_m)$.

This gives non-positive values for the covariates and makes the exponent bounded by 1. After all these preconditioning and simplifying techniques, we implement the Newton-Raphson algorithm to obtain the maximum partial likelihood estimate in R. To summarize, the general procedure is **"create edgelist structure → find covariates update → find maximum covariates → calculate the MLE"**.

To evaluate the prediction performance of these estimates, we divide the data into three parts: the building period, the training period, and the testing period. The building period is for accumulating the network structure covariates and avoiding the period of dramatic changes in the network covariates. The citation events in the training period are used for estimating the coefficients of the covariates, and then the predictive performance of the fitted model will be evaluated in the testing period. The number of unique event times for building is 2000, for training is 100, and for testing is 1000. We find that 2000 is sufficient for building since different training sizes produce very similar results. The evaluation during the testing period is based on the ranks of the partial likelihood function values of the actual citation events among all possible citation events using the estimated coefficient values. If there are multiple citation events in the actual network, we calculate the average of the ranks for these citations. The normalized ranks are ranks normalized to $[0, 1]$, so the closer to 1, the better the performances are. Figure 4.1 shows the normalized ranks for the first 100 nodes. The red reference line

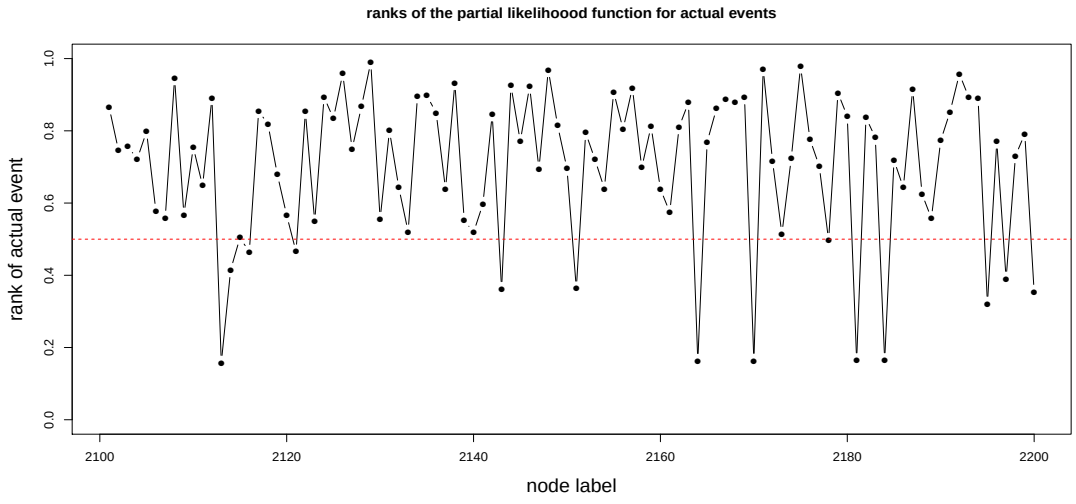

Figure 4.1. First 100 ranks for actual events for netstat Model

is 0.50, which is the average normalized rank for random guessing. It can be seen from the Figure 4.1 that there is a lot of variation among the ranks, but 90% of the ranks are above the line, which means the network structure covariates help predicting the citation events, but there might be other characteristics which also help explain which edge to be established in the network.

The average ranks for the first 100 nodes is 0.712. If we average the ranks for every 100 nodes, we can get the following Figure 4.2. There is a slightly decreasing

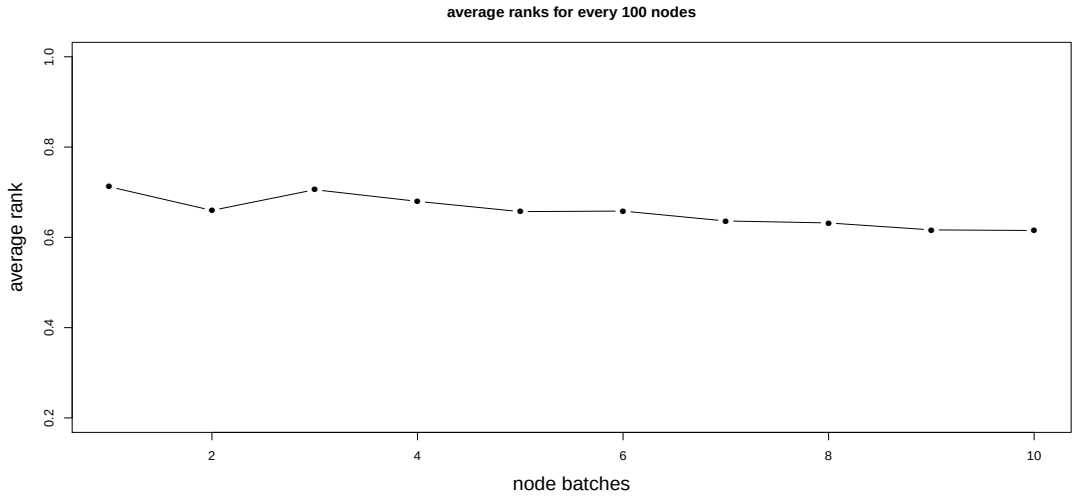

Figure 4.2. Average ranks over different node batches for netstat model

trend. In Vu et al. (2011b), a similar decreasing pattern is also observed for a longer building period. This fact indicates that the predictive power of the model degrades over time if the coefficients are held constant.

The implementations introduced above are mostly designed for models with network structure covariates, like $s^{PA1}(t)$, $s^{PA2}(t)$,.... Yet, in a citation network, similarity on topic between papers is also an important feature to be considered for citation events. One way to quantify similarity is by using the LDA covariates calculated on the arXiv-TH paper abstracts using topic vectors of length 50. Recall that the LDA covariates for node i when node j joins the network are defined through

$$\boldsymbol{s}_i^{LDA}(t_j^{arr}) = \boldsymbol{\theta}_i \odot \boldsymbol{\theta}_j, \tag{4.5}$$

which is the coordinate-wise product of the topic vector of node i and that of

the node that joins the network most recently. Therefore, the LDA covariates for all the at-risk nodes must completely recalculated whenever a new node joins the network, and they do not depend on their previous values. So the caching technique is not helpful and we need to consider a direct implementations. However, there are potential computational issues for direct implementation. The LDA covariate is 50-dimensional and, as pointed out earlier, calculating S_0, S_1 and S_2 involves summation over the entire at-risk set. So if the network has a large number of nodes, then the number of at-risk nodes can be large at each event time. The computation speed will be very slow. In the next section, we consider a case-control approximation of the partial likelihood function to reduce the computation cost. But in the rest of this subsection, we will consider direct implementation.

Using the same set of nodes for training and testing, the first 100 ranks using the model with LDA covariates only are compared with that using the model with network structure covariates only in Figures 4.3. It can be seen that the model with the LDA covariates has higher ranks (the average is 0.891) for the actual citation events, as well as less variation (standard deviation is 0.13, comparing to 0.22 for network covariates). Only 3 points fall below 0.5 and most of them are very close to 1. The average ranks for 10 different node batches are also compared in Figure 4.4. The average ranks for model with LDA is better than those for

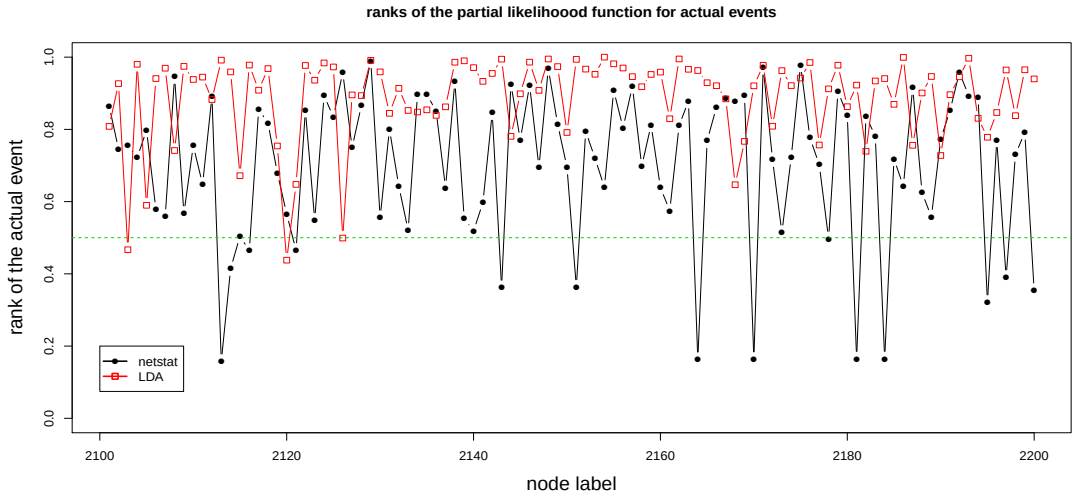

Figure 4.3. Comparing first 100 ranks between netstat model and LDA model

model with the network covariates. The curve for LDA in Figure 4.4 does not have

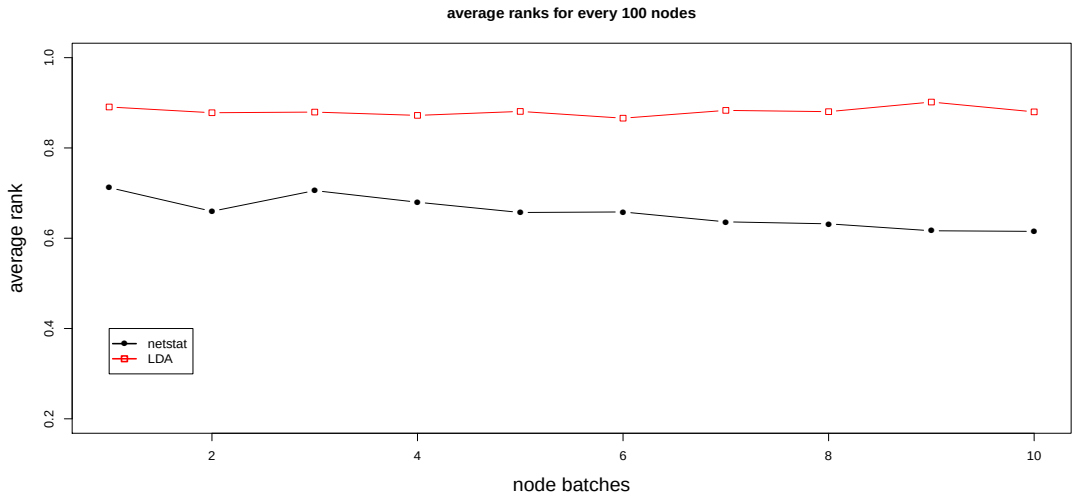

Figure 4.4. Comparing average ranks over different paper batches between netstat model and LDA model

the decreasing trend, which validates the constant effect for the LDA covariates over time. Both curves are above 0.5. This suggests that both types of covariates might be important in predicting future citation events. Therefore, we will consider a model to include both of them. Since the "cache" is not useable for LDA, we need to use direct implementation for both network covariates and the LDA covariates. This will substantially slow down the computation speed. We combine the maximum partial likelihood estimates found for LDA and networks structure covariates and treat it as the initial value for the optimization. The following Figure 4.5 compares the average ranks among all three models. And Figure 4.6 compares the percentage of the true citation events included in the top-K elements of the sorted partial likelihood list among the three models. And the percentages are averaged over all the 1000 testing set. From these two figures, we can see that the model with both LDA and network structure covariates behaviors similarly to the model with only the LDA covariates. And both of them out-perform the model with only the network structure covariates.

To further compare prediction the performance between the model with only LDA and the model with both LDA and network covariates, we compare the 1000 ranks of the testing nodes for two models using the paired Wilcoxon signed-rank test. The p-value of the test is 0.07393, suggesting that there is no significant

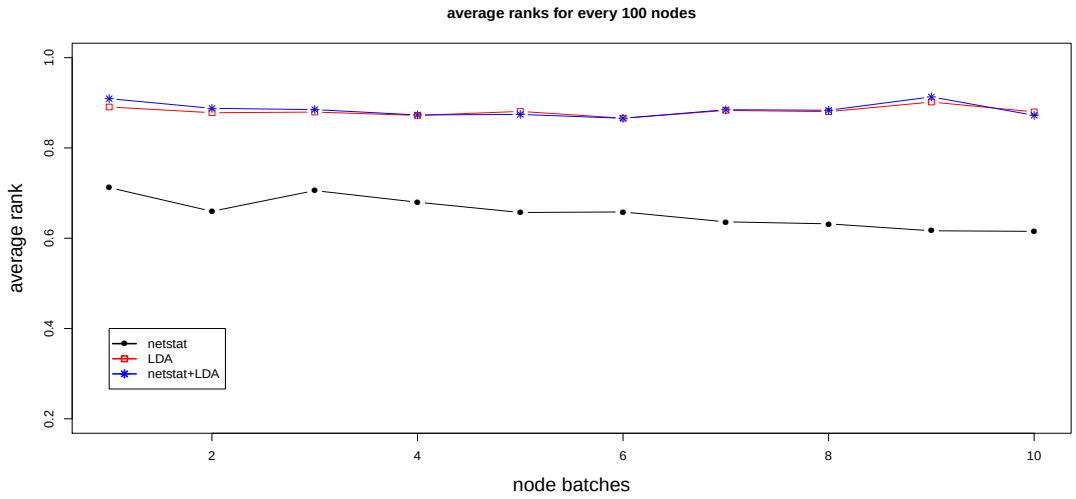

Figure 4.5. Comparing average ranks over different paper batches among LDA model, netstat model and LDA+netstat model

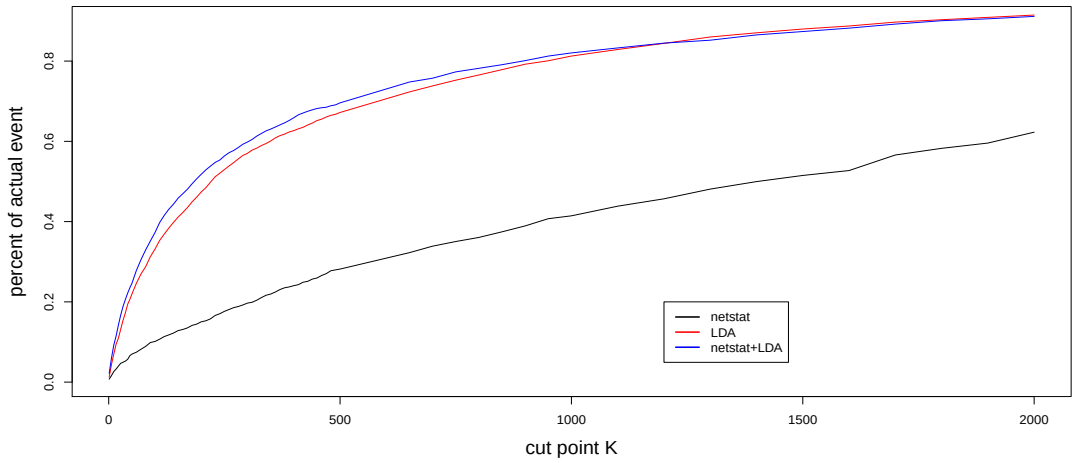

Figure 4.6. Comparing percent of actual event in the top-K recommendation list among LDA model, netstat model and LDA+netstat model

difference between the two rank vectors. In addition, if we treat the model with only LDA covariates as a reduced model, and the model with both LDA and network structure covariates as the full model, we may compare the BIC criterion (see equation (1.13)) for them. The BIC for the reduced model is 168.3 and for the full model is 202.8913. Therefore the reduced model is preferred as a simpler model.

The computation speed for models discussed above is slow, so it is not efficient to be used for large networks. This is primly because of the computation complexity brought by LDA covariates. So in the next section, we get back to the model with LDA covariates only and propose an approximation method to reduce the computational cost.

4.2.3 Case Control Approximation with LDA Covariates

The idea is to "shrink" the at-risk set by case-control sampling. The nested case-control is usually used in survival analysis when the event of interest is rare or it is expensive to collect covariates information for the entire cohort (Langholz, 2005). There are different types of case-control sampling methods, and the common feature of them is that only a smaller proportion of the none-event group (or the at-risk set excluding the event group) is sampled to be used in estimations. And this smaller set is called the "control" set. The simplest way is to pick "controls" by simple random sampling in the "control pool". An improved version could be dividing the full cohort into several strata based on some criterion, then collect random samples in each stratum. In this way, the likelihood function can be approximated even better. Raftery et al. (2012) consider applying the idea of case-control to likelihood approximation for latent space network models. In their work, observations are dyadic, indicating whether there is directed edges between pair of nodes, so presence of an edge is regarded as having the event of interests, or the "case". To use the case control approximation, they sample a few pairs of nodes among all pairs which do not have an edge connecting them, and use them as the "controls". The likelihood function is calculated by using the "case" and the sampled "controls". It is also properly scaled so that it becomes an unbiased estimate of the original likelihood function. This approximation saves a lot of computation time when the total number of nodes in the network is large and the performance of correctly predicting edges is also satisfactory.

As discussed in the end of Chapter 3, the theory applies on the sparse networks, where the occurrences of edges are rare comparing to the number of nodes in the network. The Citation networks we work on can be considered as a sparse network, since the number of citations at each event time is much smaller than the number

of papers in the network. Figure 4.7 gives the histogram of numbers of citations among papers in this network. So we will also apply this case-control idea to the

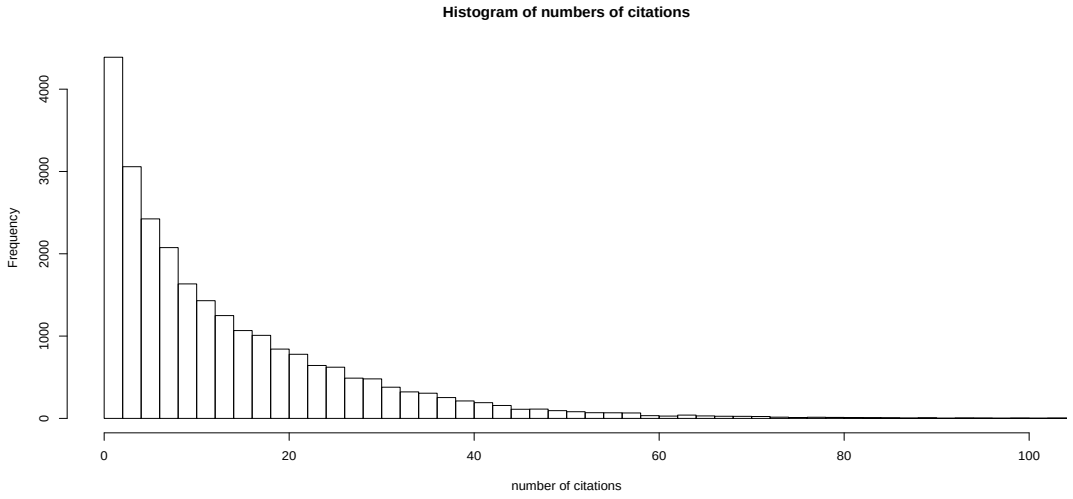

Figure 4.7. Histogram of numbers of citations for papers in the arXiv-TH network

approximated partial likelihood function (3.15) in our model setting. Similarly as in Raftery et al. (2012), the event of interest relates the presences of edges. Specifically in an egocentric approach, node i has the event means that there is an edges directed towards it. Therefore, at each event time, all nodes which receive an edge is the "cases", and all nodes which do not receive any edges are the "control pool", where the "controls" are going to sampled. We use simple random sampling to sample "controls" at each event time. If denote the sampled "control" at time t_m as C_m, then the case-control partial likelihood function for this sampling scheme is

$$\log PL_t^{\text{CC}}(\boldsymbol{\beta}) = \sum_{t_m<t} \left\{ \sum_{j\in J_m} \boldsymbol{\beta}^\top \boldsymbol{s}_j(t_m) - |J_m| \log \left[\sum_{j\in J_m \bigcup C_m} \exp\left\{\boldsymbol{\beta}^\top \boldsymbol{s}_j(t_m)\right\} \right] \right\}. \tag{4.6}$$

The only difference between this partial likelihood function and the one in (3.15) is the second logarithm term: this likelihood (4.6) sums fewer terms, and the covariates only need to be evaluated for this smaller set. So the computation is simplified in this way.

We consider different sizes of "control" set, specifically, different proportions

of the full control pool to be sampled. For a fixed $\boldsymbol{\beta}$, there is no doubt that the time to evaluate the partial likelihood function, its gradient and hessian is less if using a smaller proportion of the data. We compare the computation time to get the maximum case-control partial likelihood estimates using different control sizes. Specifically, we compare estimations using different proportions of the control-pool, namely, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90%. For each proportion, we calculate the maximum case-control partial likelihood estimates 10 times using different sampled controls. Figure 4.8 summarizes the computing times for these calculations. The left plot gives the actual computing times for all samples in each proportion and the right one describes the trend of the average computing times when the control proportion increases, where the blue line is the computing time for estimation using the full control pool. It can be seen from these figures that the

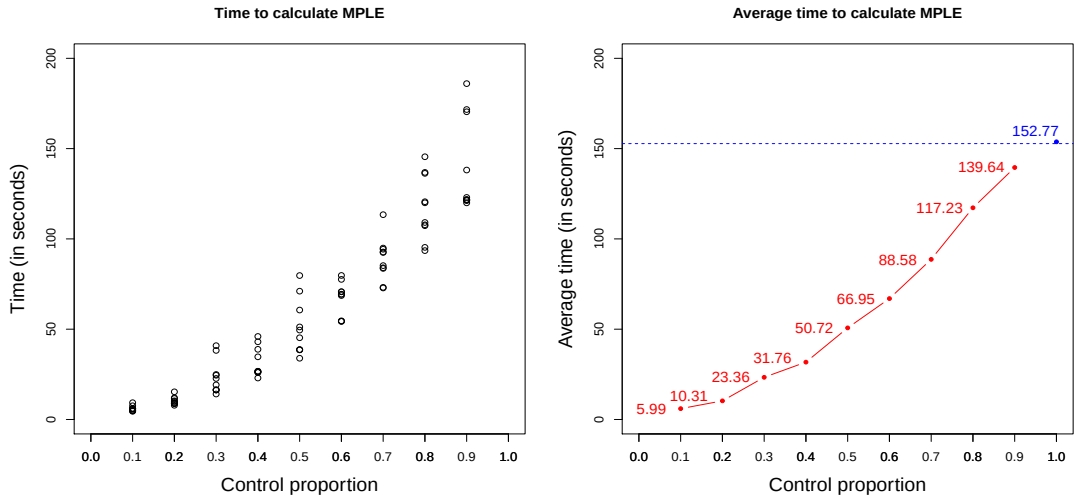

Figure 4.8. Computation times for calculating the maximum partial likelihood estimates in LDA model using case-control approximations with different sampled control proportions

less "control" one use in estimation, the less computation time it costs to maximize the case-control partial likelihood function. And the time reduction is dramatic for smaller proportions.

By applying the case-control approximation, the computation speed is fast. On the other hand, we don't want to sacrifice the good prediction performance. To see how much difference there is in the prediction among models using different

control sizes, we compare three different aspects of the ranks for the actual event produced by models using the 10 different control proportions. It turns out that all comparisons suggest that the ranks in these case-control approximation models are all very close to the original model using all the controls. First, Figure 4.9 describe the relation between the ranks using estimates from the case-control approximation and ranks using the original model, for control proportion 10%, 50% and 90%. It can be seen from the figure that the ranks from model using smaller control proportions vary more around the ranks from the original model than that with higher proportions. But the correlations are all very high for all three comparisons. Second, we can also compare the effects along with time by plotting the average

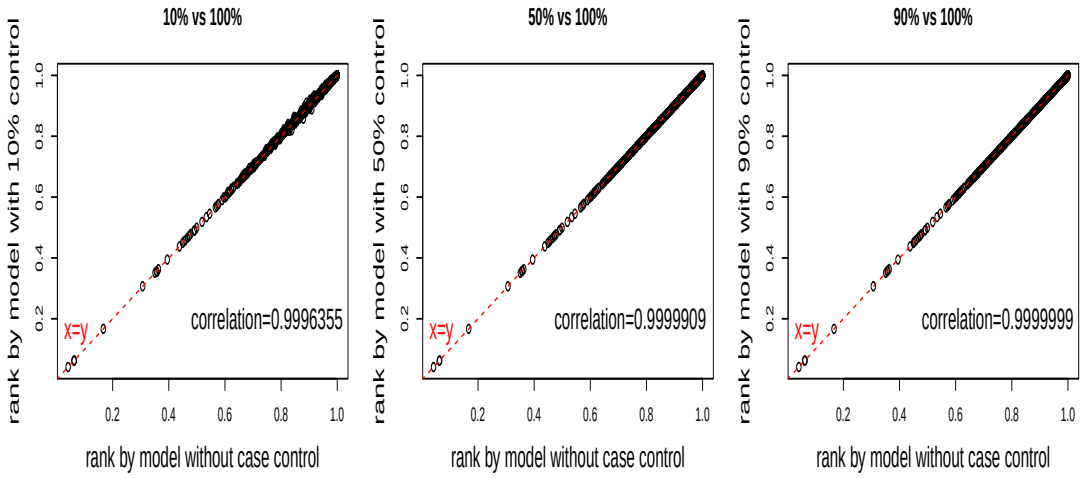

Figure 4.9. Relations of ranks between LDA model using all data and LDA models using different case-control approximations. Compared with different control proportion: 10%, 50% and 90%

ranks curves produced by these models. The comparisons are shown in Figure 4.10. To make the differences clearer, the curves are shown with a zoomed-in y axis, so they appear to change dramatically over time, which is not true in a regular plot as in Figure 4.4. But even in this zoom-in plot, the differences in the average ranks are not very significant. Last, the curves in Figure 4.11 compares the proportion of the actual events included in the top-K list of sorted partial likelihood values among these models. The left plot is for cut point up to 2000, and the right one is the zoomed-in for cut point up to 200 to show the differences. The differences

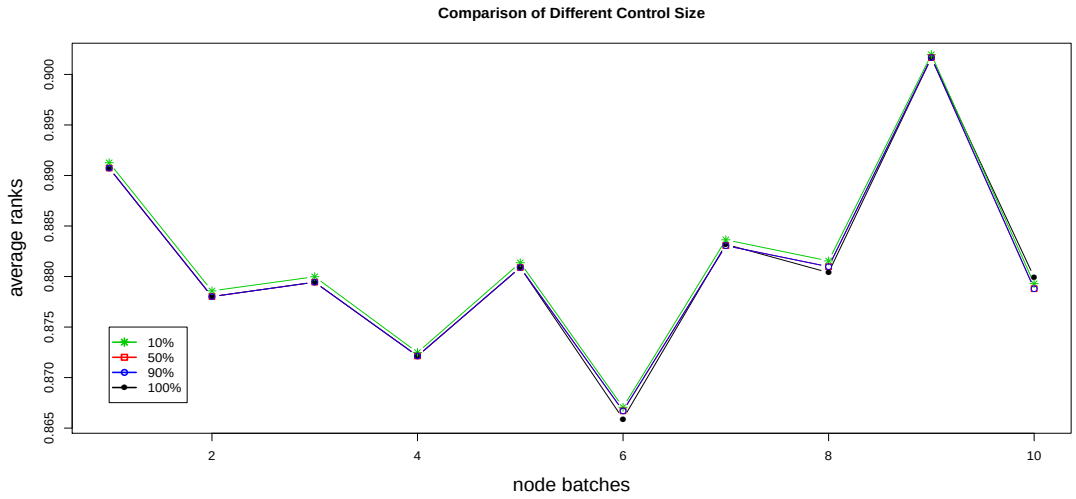

Figure 4.10. Comparisons of average ranks among LDA model using all data and LDA models using different case-control approximations. Compared with different control proportion: 10%, 50% and 90%

in patterns among these curves are small. From these three perspectives, we show that using the case-control approximation will not sacrifice much prediction power of the model, while it indeed gains efficiency in the computation speed.

Since there is randomness in the sampled control set, we also compare the results among models with the same control proportion and different control samples. For each control proportion, we consider 100 different samples, and estimate the coefficients by the case-control approximation. Figure 4.12 describes the variation among the estimated coefficients for 3 different components of LDA covariates (the results are omitted for the rest, since the patterns are similar). The red reference line is the corresponding estimate using the full control pool. These plots show that there are a lot of variation among the estimates from different samples of the same proportion, and the variation decreases when more controls are used in the model. Since the smaller proportion of the control pool is sampled, the more variation there is in the sampled control set, and more variation in the estimates. We also compare the prediction performance of models using 10 different case control samples all with sampled control proportion 10%. The ranks on the test set are a lot similar among these 10 models. Figure 4.13 and Figure 4.14 summarized the comparison among these samples. We calculate the standard deviation of the

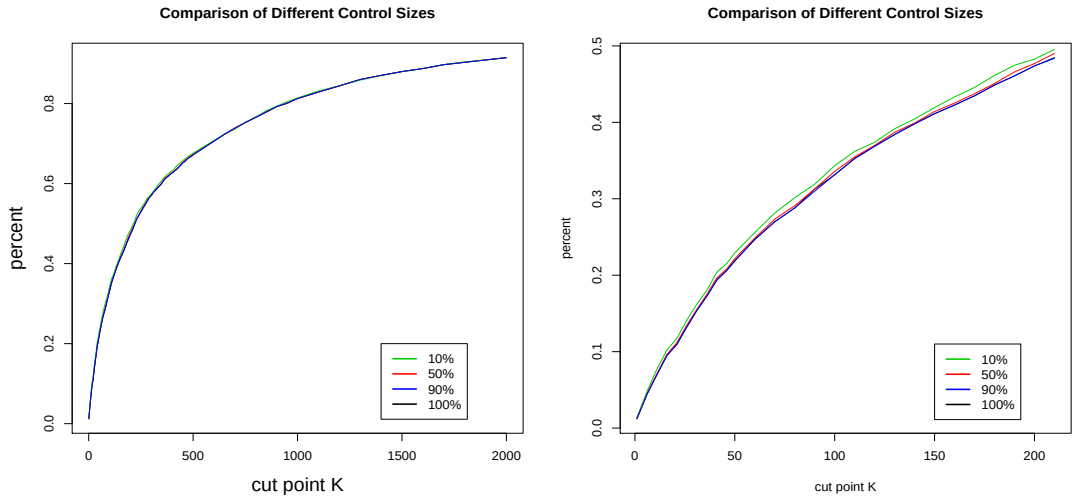

Figure 4.11. Comparisons of percent of the actual events included in the sorted partial likelihood list among LDA model using all data and LDA models using different case-control approximations. Compared with different control proportion: 10%, 50% and 90%

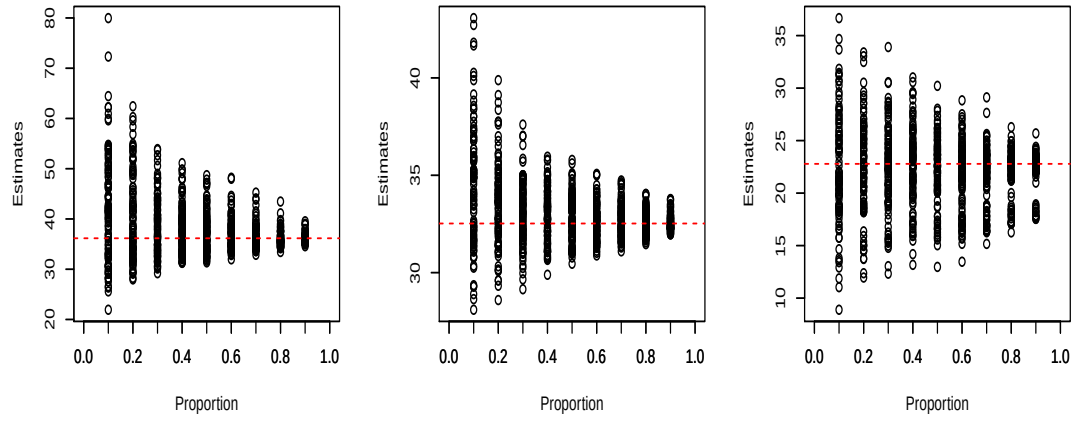

Figure 4.12. Variation in estimates by using different control samples

ranks among 10 models using different control samples for every testing node. And Figure 4.13 shows the plot and histogram for the 1000 standard deviations for all 1000 testing nodes. From these plots, we can see that most of the standard deviations are very small and do not show a pattern by time, suggesting that the variation among ranks from the 10 models is small. The Figure 4.16 shows the

the trend of average ranks over node batches, in which different colored curves represents models using different control samples. The average rank curves are also almost identical. Therefore, although using different control samples pro-

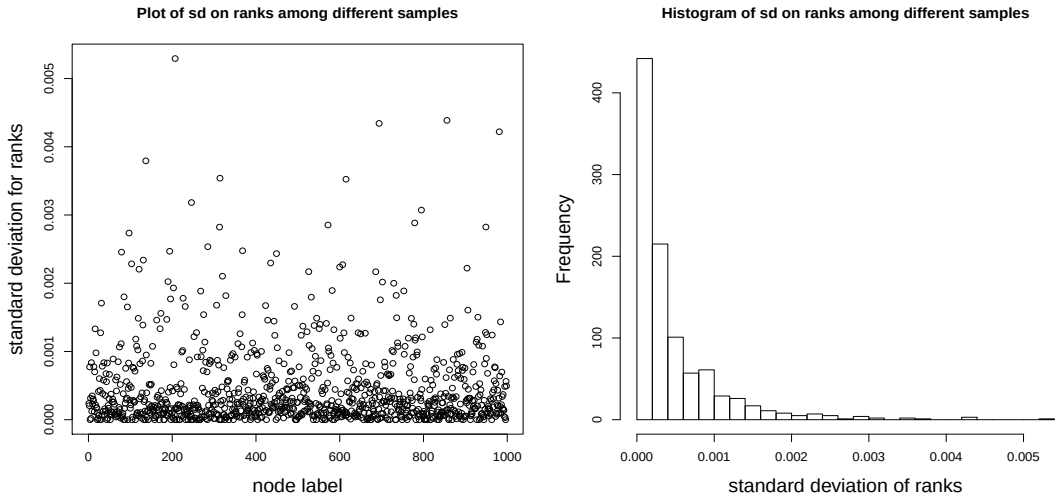

Figure 4.13. Plot and histogram of standard deviations of ranks among 10 LDA models with different control samples and 10% control proportion

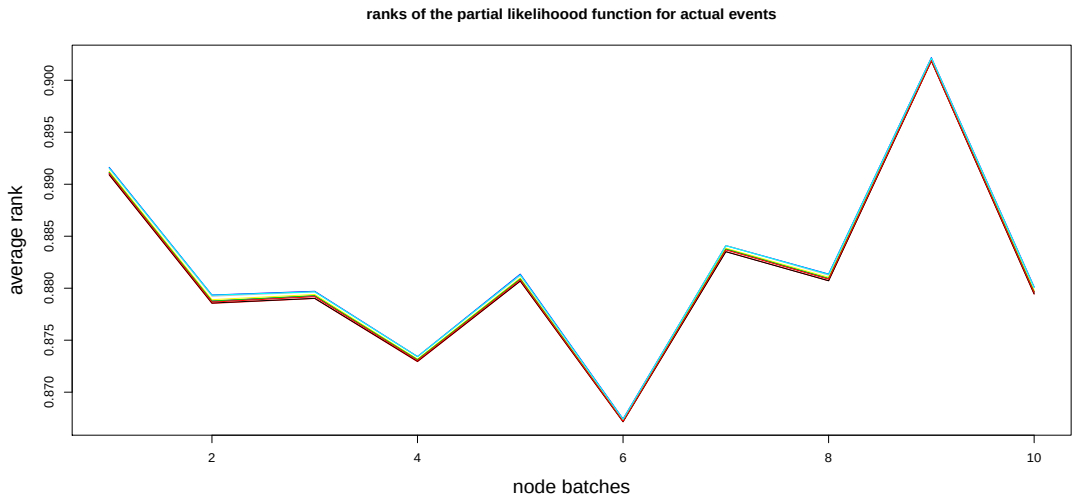

Figure 4.14. Comparisons of average ranks among 10 LDA models with different control samples and 10% control proportion

duces very different estimates in the coefficients, the prediction performances are very similar for models using these estimates. Overall, the case-control approxi-

mation method reduces the computation cost greatly and preserves the good and consistent performance in prediction.

We can apply the case control approximation to models with both the network covariates and LDA covariates. However, the prediction performance of case-control approximation on this bigger model is not as good as the one with only the LDA covariates. The average ranks using models with difference control proportions are summarized in Figure 4.15. Comparing to Figure 4.10, the pattern that using a smaller control size results in worse prediction performance is clearer. Similarly, in order to capture the variation among models with the same control

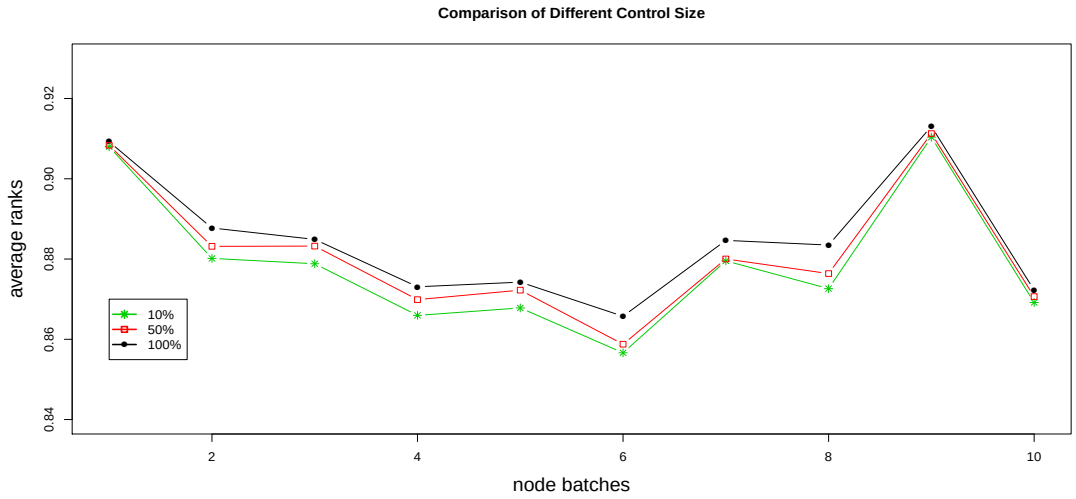

Figure 4.15. Comparisons of average ranks among models using different control proportions for LDA+netstat model

proportion but different control samples, Figure 4.16 show plots of average ranks for 10 models using 10% controls but different control samples. The black dashed line is the average ranks for model without case control approximation, and rest 10 colored lines represent average ranks for 10 models with different control samples. From these two figures, we observe that using the case-control sampling with the network structure covariates can magnify the inconsistency and variation in results brought by approximation. Combing this with the observations in Figure 4.5 and 4.6, we will consider model with only the LDA covariates in the rest of the Chapter.

To summarize, using case-control approximation on the partial likelihood and

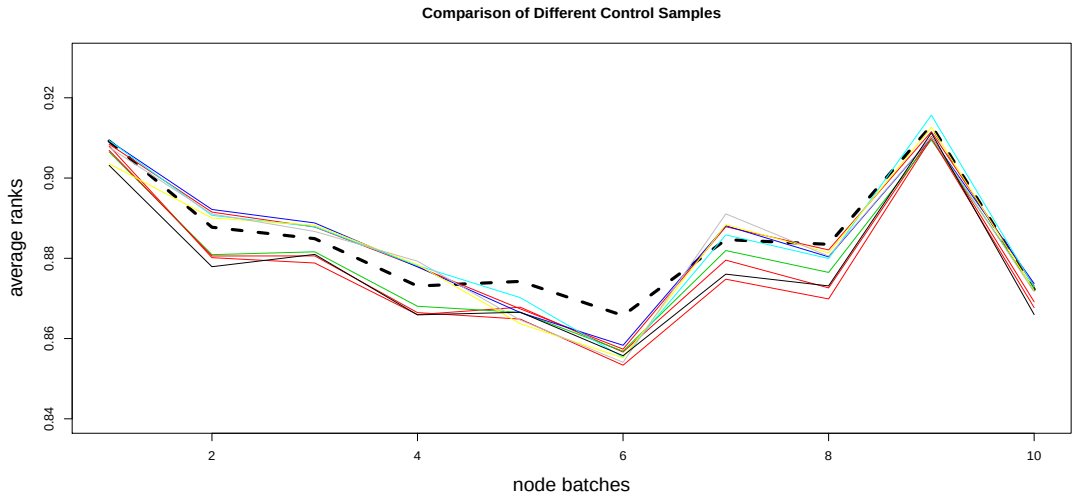

Figure 4.16. Comparisons of average ranks among 10 models with different control samples and 10% proportion for LDA+netstat model. The black dashed line is the one with model using all controls.

maximize it to estimate the coefficients of covariates will gain a lot of efficiency in computation speed. And it still reserve good performance in prediction future edges. This is one way to reduce the intensive computation for model using LDA covariates on large network data. Another way to speed up the estimation and simplify the calculation is to reduce the number of LDA covariates used in the model. We will implement this by variable selection for LDA covariates in the next section.

4.3 Algorithms for Variable Selection via Penalization

4.3.1 Introduction

In previous section, we study estimations of coefficients of LDA covariates using direct implementation and case-control approximation. As discussed in last chapter, we also want to select LDA covariates in order to have a parsimonious model without loosing consistency in estimation. In Chapter 3, we have showed that asymptotically maximizing the penalized approximated partial likelihood function

yields a sparse and efficient estimator. In this section, we will apply a algorithm to implement this optimization and study the finite sample performance of it. Recall the objective function is

$$-\log \widetilde{PL_{t_n}}(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \tag{4.7}$$

And we will consider the adaptive LASSO as the penalty function with weight $\tilde{\beta}$

$$p_\lambda(|\beta|) = \lambda \frac{|\beta|}{|\tilde{\beta}|} \tag{4.8}$$

The goal is to minimize equation (4.7).

There are many literatures on computing algorithms for optimizing the penalized likelihood function in the generalize linear model setting and the survival model setting. To use for variable selection for networks, the algorithm should be able to deal with time-varying covariates and may also take the advantages of available un-penalized estimates. Considering these aspects, we can use the computation routine proposed by Zhang and Lu (2007), which is also used by Liu and Zeng (2013). The idea is to initialize the algorithm by the maximum partial likelihood estimates, and transform the original partial likelihood function into a pseudo least square, then apply the modified shooting algorithm in Fu (1998) to get the sparse estimates. In our setting, suppose $\tilde{\boldsymbol{\beta}}$ is the maximizer of the approximated partial likelihood function $\log \widetilde{PL_t}(\boldsymbol{\beta})$. For any $\boldsymbol{\beta}$ in a neighborhood of $\tilde{\boldsymbol{\beta}}$, we have

$$\nabla \log \widetilde{PL_t}(\boldsymbol{\beta}) \approx -\left\{\nabla^2 \log \widetilde{PL_t}(\boldsymbol{\beta})\right\}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^\top \tag{4.9}$$

And also, the second order Taylor expansion of $\log \widetilde{PL_t}(\boldsymbol{\beta})$ gives,

$$\log \widetilde{PL_t}(\tilde{\boldsymbol{\beta}}) \approx \log \widetilde{PL_t}(\boldsymbol{\beta}) + (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^\top \nabla \log \widetilde{PL_t}(\boldsymbol{\beta}) + \frac{1}{2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^\top \nabla^2 \log \widetilde{PL_t}(\boldsymbol{\beta})(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) \tag{4.10}$$

Substituting equation (4.9) into (4.10) gives,

$$\log \widetilde{PL_t}(\tilde{\boldsymbol{\beta}}) \approx \log \widetilde{PL_t}(\boldsymbol{\beta}) - \frac{1}{2}\nabla \log \widetilde{PL_t}(\boldsymbol{\beta})^\top \left\{\nabla^2 \log \widetilde{PL_t}(\boldsymbol{\beta})\right\}^{-1} \nabla \log \widetilde{PL_t}(\boldsymbol{\beta}) \tag{4.11}$$

Therefore, maximizing the (4.7) is equivalent as maximizing

$$
\begin{aligned}
& -\frac{1}{2}\nabla \log \widetilde{PL_t}(\boldsymbol{\beta})^\top \left\{\nabla^2 \log \widetilde{PL_t}(\boldsymbol{\beta})\right\}^{-1} \nabla \log \widetilde{PL_t}(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \\
= \; & \frac{1}{2}(Y - X\boldsymbol{\beta})^\top (Y - X\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \qquad (4.12)
\end{aligned}
$$

where

$$
\begin{aligned}
X \quad s.t. \quad & -\nabla^2 \log \widetilde{PL_t}(\boldsymbol{\beta}) = X^\top X \qquad (4.13) \\
Y \quad = \quad & (X^\top)^{-1}\{-\nabla^2 \log \widetilde{PL_t}(\boldsymbol{\beta})\boldsymbol{\beta} + \nabla \log \widetilde{PL_t}(\boldsymbol{\beta})\} \qquad (4.14)
\end{aligned}
$$

And the modified shooting algorithm can be used to minimize (4.12)

The complete algorithm is as follows (Zhang and Lu, 2007; Liu and Zeng, 2013),

1. Set initial value as the maximizer $\tilde{\boldsymbol{\beta}}$ of the approximated partial likelihood function.
2. Calculate $\nabla^2 \log \widetilde{PL_t}(\boldsymbol{\beta})$, $\nabla \log \widetilde{PL_t}(\boldsymbol{\beta})$ at current estimates. Also get X, Y by (4.13) and (4.14)
3. Using the modified shooting algorithm to optimize (4.12)
4. Iterate between 2 and 3 until converges.

The tuning parameter λ can be selected by minimizing the GCV criterion (Zhang and Lu, 2007; Liu and Zeng, 2013).

Using this algorithm, we select LDA covariates in the next subsection. And in the subsection after, we study the maximizer of the penalized case-control partial likelihood function, in order to reduce the computational complexity.

4.3.2 Variable Selection for LDA Covariates

Using the same sizes of building, training and testing as before, the algorithm in the last subsection is applied for model with all 50 LDA covariates. We select 9 components out of 50. The estimates are listed in the following Table 4.1. By theory in Chapter 3, asymptotically, the estimates are consistent when the

Table 4.1. Nonzero estimates for coefficients of LDA covariates

β_5	β_7	β_{12}	β_{17}	β_{24}	β_{29}	β_{36}	β_{44}	β_{49}
4.02	13.50	13.12	4.04	27.99	8.73	14.09	18.32	11.12

true model is sparse. Since we don't know the true model, we will compare the prediction performance between the penalized and unpenalized LDA models using the ranks for the 1000 testing nodes calculated by these two models. The two rank vectors are very similar to each other from Figure 4.17. The left one is the plot of ranks from the penalized model versus ranks from the unpenalized model. The right one shows the histogram of the differences between these ranks (ranks by the unpenalized model subtrates ranks by the penalized model). It is slightly right-skewed, suggesting that the unpenalized model provides slightly higher ranks for the actual citation events. But the maximum difference is only 0.005, and the skewness is not severer. We still consider the ranks for the penalized model is quite close to the ranks for the unpenalized model. The average ranks over paper batches

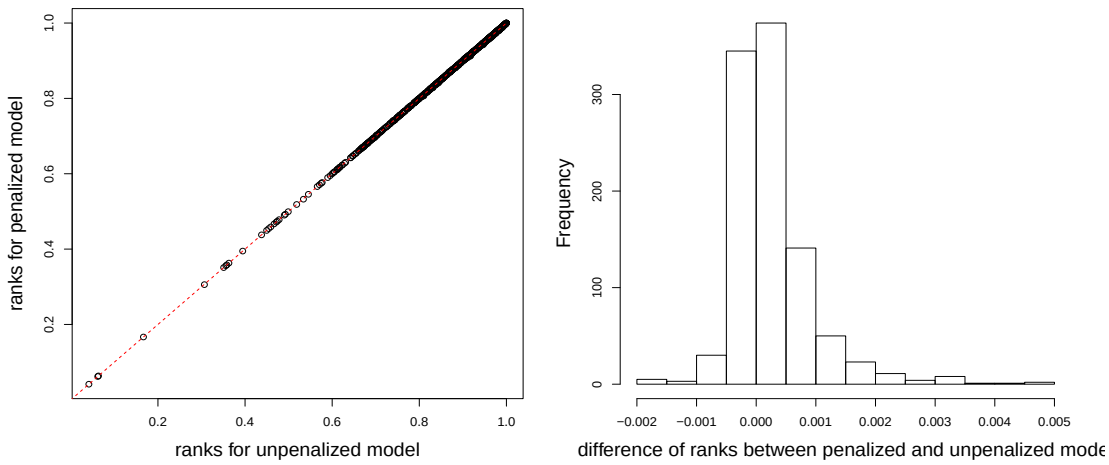

Figure 4.17. Relation of ranks using penalized and unpenalized LDA models. The difference in the right plot is ranks of the unpenalized model minus the ranks of the penalized model

are also compared in Figure 4.18. From the Figure, the average ranks for models before and after variable selection are quite similar. Figure 4.19 further compares the percents of actual citations in the sorted partial likelihood list between the penalized and unpenalized models. The left one considers up to top 2000 of the

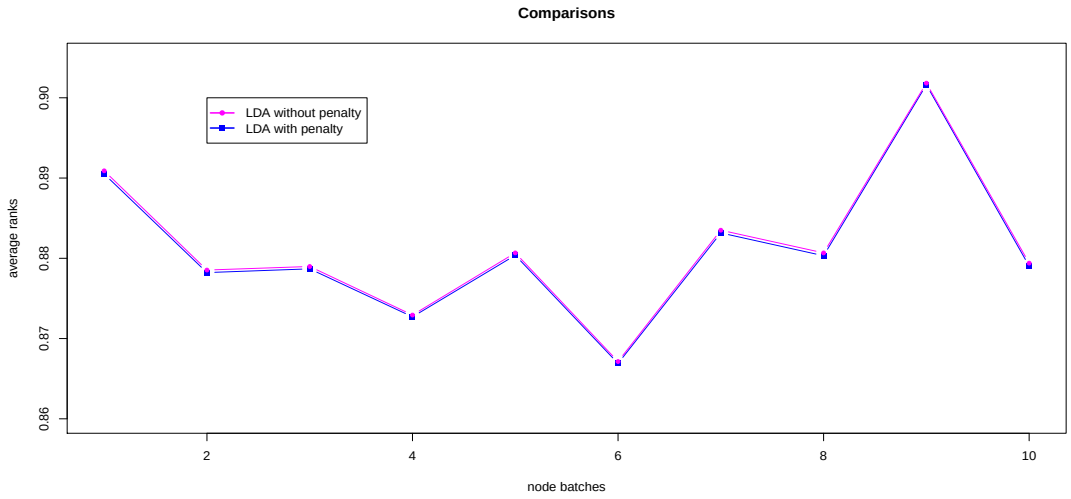

Figure 4.18. Comparisons of average ranks between penalized and unpenalized LDA models

list, which does not suggests any significant differences between the two curves. The right one is for the top 200 list, and this zoom-in plot shows that after variable selection, the percent is just a slightly worse than the original model. This suggests

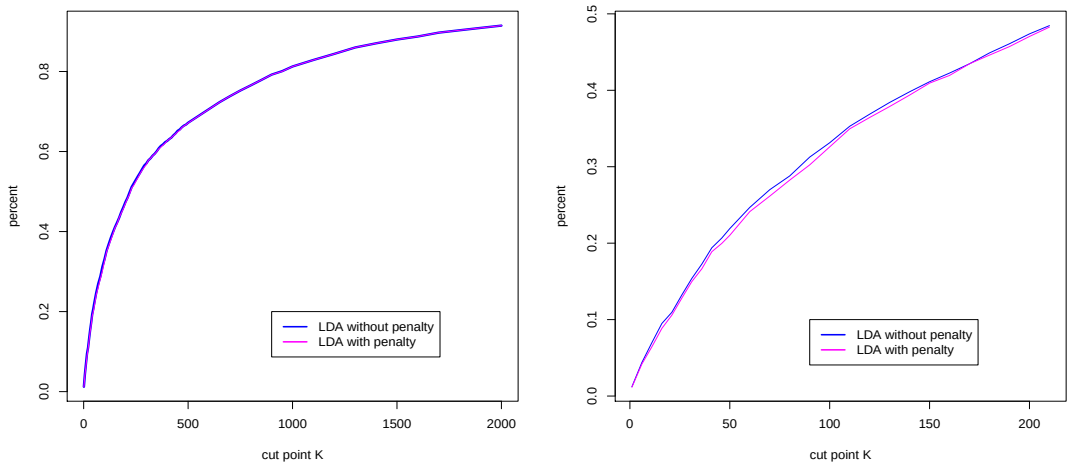

Figure 4.19. Comparisons of percents of the actual event in the top K sorted partial likelihood list between penalize and unpenalized LDA models

that the model after variable selection does not perform worse than the original model. But on the other hand, it indeed simplifies and reduces the computation with a much parsimonious model.

As most of other algorithms for maximizing the penalized likelihood function, the above algorithms works the best when the computational costs for evaluating the likelihood function, its gradient and Hessian are cheap. However, in a model with a 50-dimensional LDA covariates, the computational costs to evaluate them directly are really high, so it is very slow to maximize the penalized partial likelihood function in this setting. In the next section, we consider using case-control approximations of the penalized partial likelihood function to reduce computational cost.

4.3.3 Variable Selection Using Case Control

The implementation of case-control approximation for variable selection follows directly by replacing $\log \widetilde{PL_t}(\boldsymbol{\beta})$ in (4.7) by the case-control partial likelihood $\log PL_t^{\text{CC}}(\boldsymbol{\beta})$ in (4.6). The algorithm is the same as described in Section 4.3.1. The LDA covariates selected using models with different proportions of control samples are not significantly different from each other. For each control proportion, we maximize the penalized partial likelihood function using 10 different samples. Table 4.2 summarizes aspects of the estimated nonzero coefficients by these 10 different samples in each control proportion. The covariates selected among these samples restrict to a set contains only 10 variables: the β_5, β_7, β_{12}, β_{17}, β_{24}, β_{29}, β_{36}, β_{42}, β_{44} and β_{49} components of the LDA covariates. For each proportion, the results contains 3 rows. The first row is the average of the estimated coefficients, the second row is the corresponding standard deviation among the 10 samples, and the third row is the proportion of nonzero estimates among these samples. From the table, we can observe several things. First, using the case-control approximation is more likely to miss important variables instead of including unimportant variables in the model, assuming the penalized estimates without case control approximation identifies the true model. Second, if the penalized estimates is large, then the corresponding penalized case-control estimates are usually nonzero, and the variation among estimates is also small. For example, β_7, β_{12} and β_{44}. Also, the variation among estimates generally decreases when more controls are used in the estimation. In general, the average estimates are close to the penalized estimates without using case control samples. And comparing to the unpenalized

case-control estimates in section 4.2.3, the variation among estimates using the same proportion but different samples is smaller.

Table 4.2. Penalized Estimates from different case control samples

p	β_5	β_7	β_{12}	β_{17}	β_{24}	β_{29}	β_{36}	β_{42}	β_{44}	β_{49}
10%	**6.87**	**18.58**	**9.17**	**0.96**	**19.55**	**8.34**	**13.43**	**0**	**13.86**	**3.78**
	(4.70)	(5.96)	(0.42)	(2.21)	(8.37)	(3.31)	(1.17)	0	(0.84)	(1.28)
	[0.8]	[1]	[1]	[0.2]	[1]	[1]	[1]	[0]	[1]	[1]
20%	**5.72**	**15.24**	**10.55**	**0.57**	**19.22**	**6.67**	**12.36**	**0**	**15.74**	**6.06**
	(3.58)	(5.27)	(0.37)	(1.31)	(9.08)	(3.59)	(3.09)	0	(0.74)	(0.96)
	[0.9]	[1]	[1]	[0.2]	[1]	[0.9]	[1]	[0]	[1]	[1]
30%	**6.68**	**13.53**	**11.24**	**2.32**	**20.57**	**7.57**	**13.94**	**0.052**	**16.56**	**7.81**
	(4.40)	(3.24)	(0.44)	(3.76)	(7.42)	(2.73)	(0.57)	(0.17)	(0.67)	(0.65)
	[0.8]	[1]	[1]	[0.4]	[1]	[1]	[1]	[0.1]	[1]	[1]
40%	**4.97**	**13.47**	**11.83**	**1.31**	**24.23**	**8.03**	**14.00**	**0**	**17.05**	**8.72**
	(4.95)	(2.04)	(0.39)	(2.26)	(1.67)	(1.33)	(0.55)	(0)	(0.38)	(0.20)
	[0.6]	[1]	[1]	[0.4]	[1]	[1]	[1]	[0]	[1]	[1]
50%	**4.58**	**12.08**	**11.97**	**0.86**	**21.00**	**6.82**	**13.58**	**0**	**17.25**	**8.78**
	(4.52)	(2.37)	(0.27)	(1.71)	(8.86)	(2.92)	(0.71)	(0)	(0.42)	(1.11)
	[0.6]	[1]	[1]	[0.4]	[1]	[0.9]	[1]	[0]	[1]	[1]
60%	**4.25**	**12.49**	**12.27**	**1.70**	**20.49**	**6.37**	**13.51**	**0**	**17.51**	**9.10**
	(4.41)	(1.94)	(0.37)	(2.77)	(10.02)	(3.73)	(1.11)	(0)	(0.55)	(1.54)
	[0.7]	[1]	[1]	[0.4]	[1]	[0.8]	[1]	[0]	[1]	[1]
70%	**2.28**	**10.65**	**12.35**	**1.28**	**16.18**	**4.64**	**13.14**	**0**	**17.53**	**9.03**
	(2.85)	(2.74)	(0.19)	(2.16)	(10.99)	(3.49)	(1.08)	(0)	(0.59)	(1.46)
	[0.6]	[1]	[1]	[0.3]	[0.9]	[0.7]	[1]	[0]	[1]	[1]
80%	**3.17**	**11.83**	**12.68**	**1.70**	**20.80**	**6.32**	**13.67**	**0**	**17.89**	**10.00**
	(2.81)	(2.31)	(0.26)	(1.89)	(8.46)	(2.69)	(0.81)	(0)	(0.55)	(0.86)
	[0.7]	[1]	[1]	[0.5]	[1]	[0.9]	[1]	[0]	[1]	[1]
90%	**3.03**	**12.49**	**12.89**	**2.92**	**24.59**	**7.45**	**13.82**	**0**	**18.12**	**10.57**
	(1.67)	(1.93)	(0.14)	(1.61)	(7.11)	(2.64)	(0.61)	(0)	(0.39)	(0.76)
	[0.8]	[1]	[1]	[0.8]	[1]	[0.9]	[1]	[0]	[1]	[1]
all	**4.02**	**13.50**	**13.12**	**4.04**	**27.99**	**8.73**	**14.09**	**0**	**18.32**	**11.12**

We also compare the prediction performance between the penalized case-control estimates and the original penalized estimates. Figure 4.20 compares the ranks of the 1000 testing nodes using the original penalized estimates and ranks using case-control estimates. Control proportions 10%, 50% and 90% are compared separately. All points are very close to the 45-degree line in each plot, and there is no apparent difference among the three plots. In addition, the average rank curves and the plots of the percent of actual citations included in the top-K recommendation lists are indistinguishable among penalized models using different proportion of the controls, according to Figure 4.21.

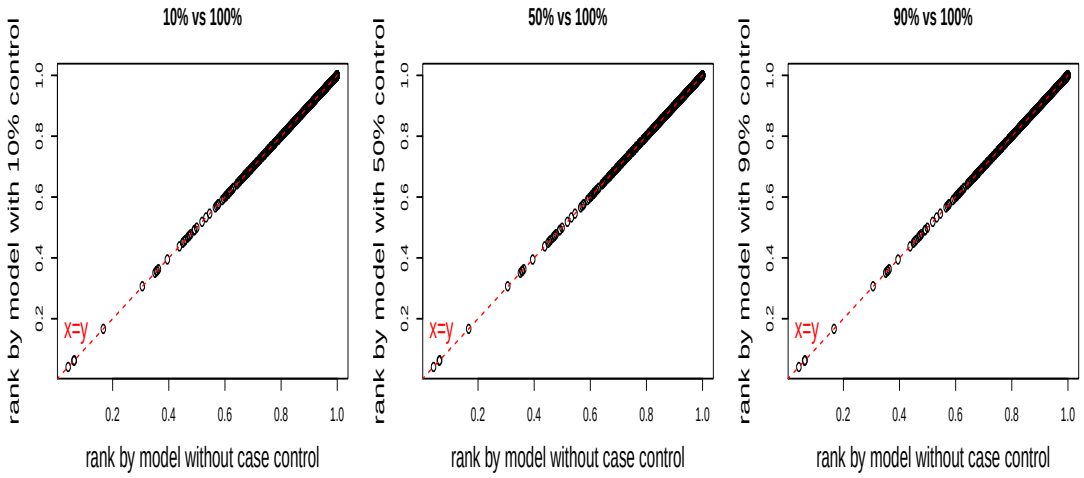

Figure 4.20. Relations of ranks between the penalized LDA model using all data and penalized LDA models using different case-control approximations. Compared with different control proportion: 10%, 50% and 90%

Similarly as in Section 4.2.3, we also want to evaluate the variation of the ranks among penalized models using the same control proportion but different control samples. For each of the control proportion 10%, 50%, and 90%, we calculate ranks of the testing nodes for 10 models using different control samples. Then for each of the testing node, we compare the 10 ranks using models with 10 different samples, and calculate the sample standard deviation of the 10 ranks. The following Figure 4.22, Figure 4.23 and 4.24 show the plots and histograms of these standard deviations for control proportion 10%, 50% and 90% separately. The first plots in these figures suggests that there is no pattern in variation among ranks by different samples along with time for all three control proportions. Though the histograms show that the range of the standard deviations shrinks when the control proportion increases, for all three control proportions, the standard deviations of the ranks are very small. This suggests that there is only very small variations among the ranks using models with different control samples.

To conclude all the observations above, using the case control approximation to derive estimates in variable selection may reduce the computation time, while still maintains very similar and stable prediction performance as the estimates from model without using case control approximation.

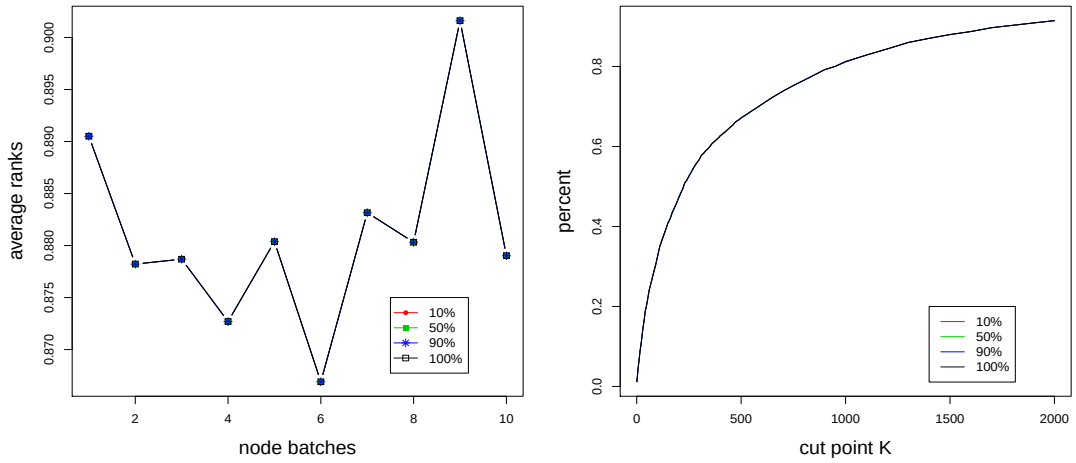

Figure 4.21. Left: Comparisons of average ranks among the penalized LDA model using all data and penalized LDA models using different case-control approximations. Right: Comparisons of percent of average ranks included in the top K elements of the sorted partial likelihood lists among the LDA model using all data and penalized LDA models using different case-control approximations. Compared with different control proportion: 10%, 50% and 90%

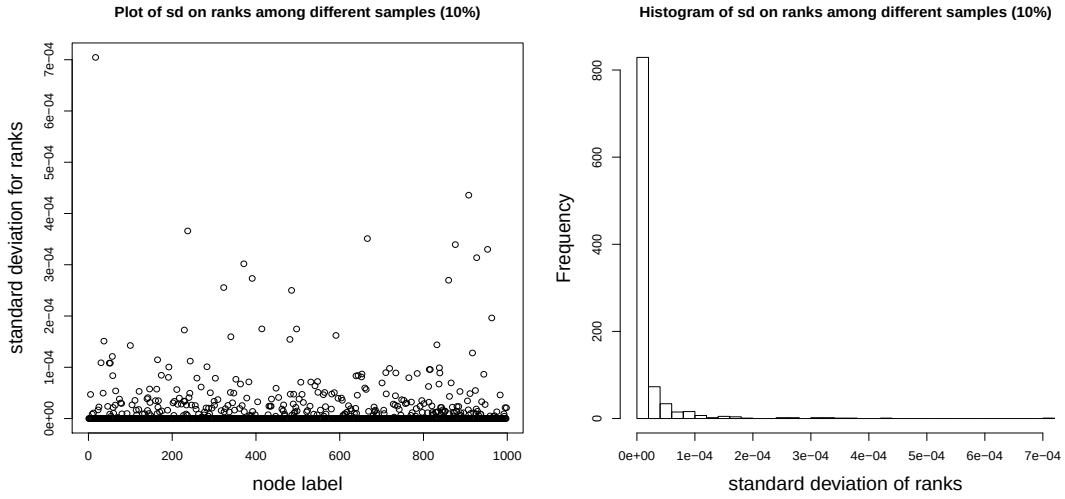

Figure 4.22. Plot and Histogram of standard deviations of ranks among 10 LDA models using different case control samples (10% control proportion)

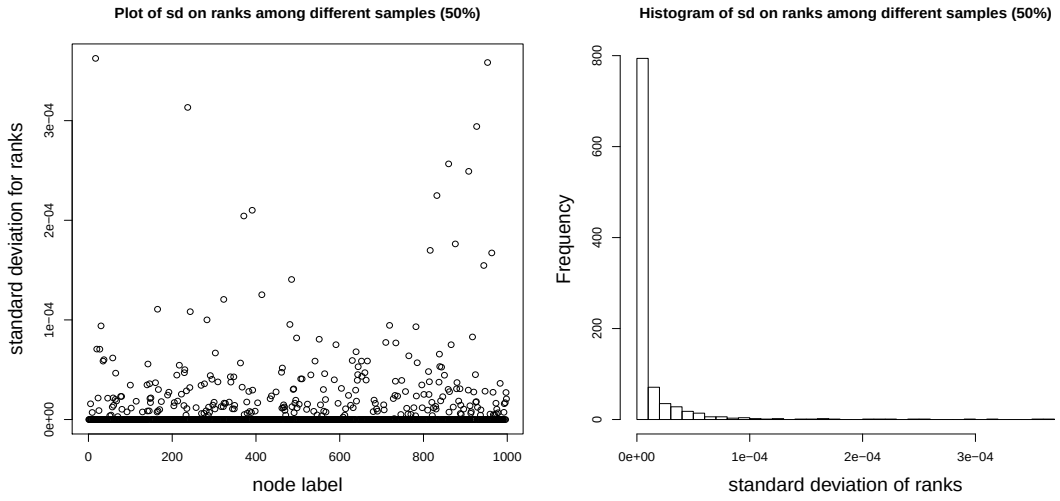

Figure 4.23. Plot and Histogram of standard deviations of ranks among 10 LDA models using different case control samples (50% control proportion)

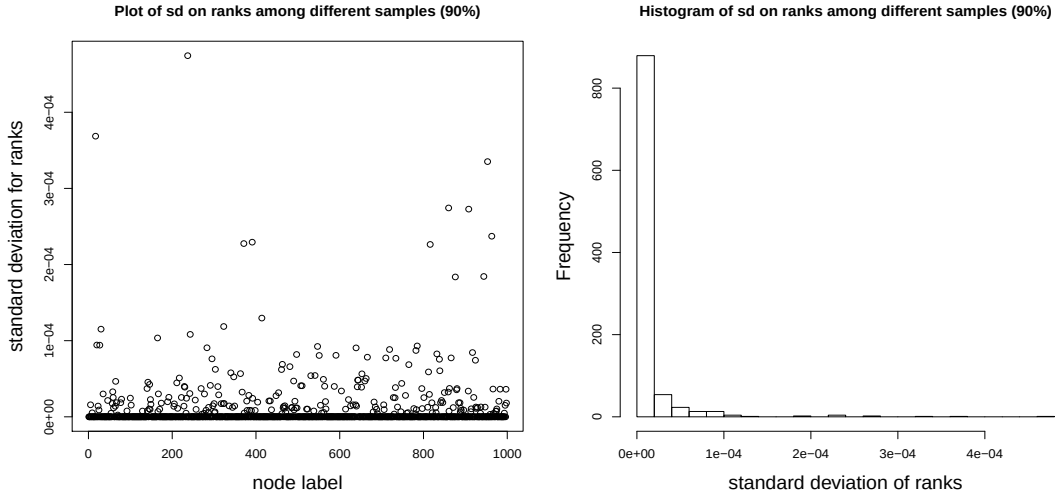

Figure 4.24. Plot and Histogram of standard deviations of ranks among 10 LDA models using different case control samples (90% control proportion)

Chapter 5

Future Work

In this chapter, we will discuss several possible future extensions of the work in this dissertation.

5.1 Variable Selection for Both Network Structure and LDA Covariates

In Chapter 4, we apply penalization to select the LDA covariates. The models before and after variable selection have good and similar prediction performance. We then apply the case-control approximation on the penalized partial likelihood to simplify the computation. The resulting sparse model with only 9 LDA covariates has a similar prediction performance to the model with both 7 network structure covariates and 50 LDA covariates. Therefore, the model after LDA covariates selection is sufficient to predict future edges.

Variable selection can also be applied on the model with only the network structure covariates. Suppose the LDA information for nodes is not available for some networks. Applied on the citation network data, the only network structure covariate selected is $\boldsymbol{s}^{PA1}(\cdot)$. This agrees with the results reported by Vu et al. (2011b), in which the performance of the model with only $\boldsymbol{s}^{PA1}(\cdot)$ and that of the model with all network structure covariates are shown to be very similar. Figure 5.1 compares the average predicted ranks over paper batches for models before and after variable selection. The two curves have similar locations and trends. The

difference in the standardized ranks of 1000 testing events has mean 0.0008596 and standard deviation 0.1053. So the two models perform similarly. Yet the model

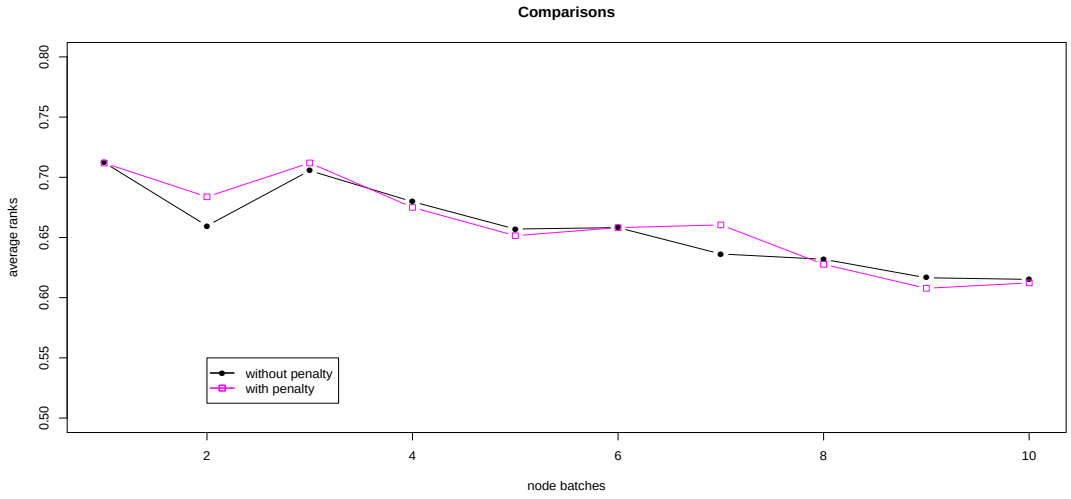

Figure 5.1. Comparisons of average ranks for model before and after selection

after variable selection only uses $\boldsymbol{s}^{PA1}(\cdot)$, which is easy to obtain. Therefore, future prediction using this reduced model is sufficient and fast.

We also apply the variable selection for the model with both the network structure covariates and LDA covariates. The covariates selected are $\boldsymbol{s}^{PA1}(\cdot)$ and $\boldsymbol{s}^{PA2}(\cdot)$ for the network structure covariates , and components 7, 12, 24, 29, 36, 42 and 49 for the LDA covariates. The performance of this estimate is not as good as the one with only the selected LDA covariates. Figure 5.2 compares the average ranks among penalized models with network structure covariates only, with LDA covariates only, and with both. The ranks for the penalized model with both types of covariates are significantly smaller than for the model before penalization, and also demonstrate a decreasing trend with time. This deceasing trend is similar to that of the model with only the network structure covariates. And it seems that the ranks are much smaller than the ranks by the model with only the LDA and network structure covariates. This suggests that the reduced model (with both types of covariates after penalization) is not compatible with the full model (with both types of covariates) in prediction performance, or some of the important sufficient covariates (especially for the LDA covariates) are not being selected in the reduced model. This might be caused by selecting the tuning parameter

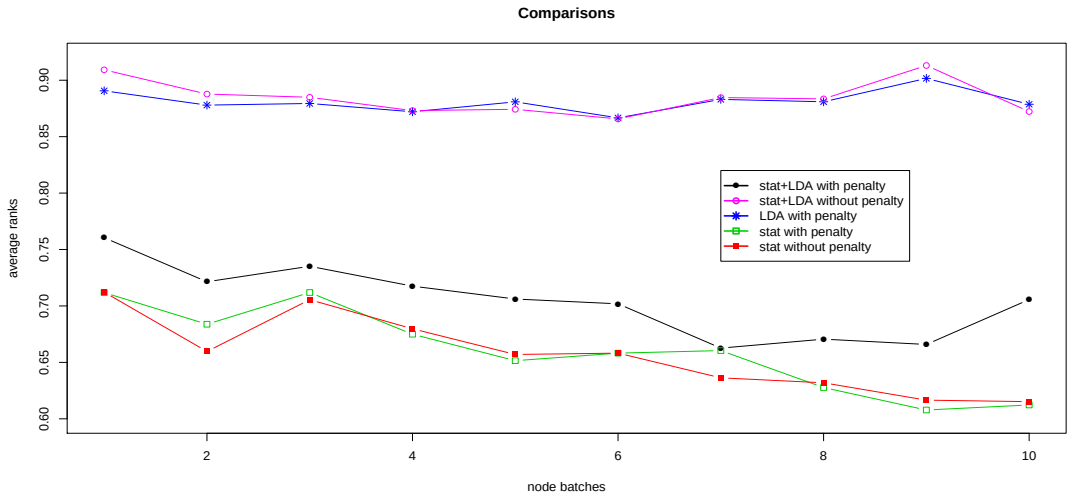

Figure 5.2. Comparisons of average ranks for penalized model with different covariates

improperly. According to the definition, network structure covariates and LDA covariates are not on the same scale. Network structure covariates are generally large positive integer values and LDA covariates are bounded by 1. This might result in dramatic differences in the Hessian matrix elements, which further affects the approximated degree of freedom $p(\lambda)$ in GCV. (This doesn't affect our variable selection for model with only LDA though, since all the covariates are of the same scale). Literature using GCV as a tuning parameter selection criterion usually standardizes the covariates. However, in a dynamic network with time-varying covariates, standardizing the covariates needs to be done over all event times for all nodes. This is not attractive computationally. Thus, developing/searching for a new tuning parameter selection techniques and/or variable selection criterion suitable for models with variables on different scales is of interest.

5.2 Extension on Theory and Implementations for the Citation Network

In Chapters 3 and 4, we discuss the theory for the unpenalized/penalized approximated partial likelihood estimators and implement variable selection for a citation network. Several aspects of this theory and implementations can be studied fur-

ther.

First, in Chapter 4, the prediction performance is evaluated by the rank in partial likelihood of the actual citation events among all possible citations using the estimated coefficients. Since we can obtain the negative Hessian matrix at each estimate, the inverse of it can be used an estimate of covariance matrix for the estimators. If standard errors of these ranks can be written as a function of standard errors of the estimators, then we can provide interval estimates for the predicted ranks.

Second, in Chapter 4, a case-control approximation for the partial likelihood function is introduced for the LDA covariates and the performance of the approximation is evaluated on the prediction performance of future edges. The case-control approximation is further implemented in variable selection to reduce the computational complexity. However, none of the asymptotic properties for these estimators are considered. Goldstein et al. (1992) study the asymptotic theory for nested case-control sampling in Cox's model. They prove the consistency and asymptotic normality of the maximizer of the case-control partial likelihood function with i.i.d. observations within a bounded observation time interval. We might extend the theory to network settings using the same rescaling technique in Perry and Wolfe (2013). Then we want to justify the order of the difference between the original log-partial likelihood function and the case-control log partial likelihood function. In Chapter 4, Figure 4.12 shows the convergence pattern for case-control estimates for different control sizes. The plots suggest a pattern similar to $\sqrt{n}$. If the theoretical justification indicates an order smaller than $\sqrt{n}$, then the theory for variable selection also holds for models using case-control partial likelihood function. If not, we may apply the similar idea of counter-matching for Cox's model, in which we divide the control pool into several strata using some of the covariate information or other nodal properties, then sample controls from each stratum. In this way, the approximation in the log-partial likelihood can be improved.

Third, the implementation in Chapter 4 is based on the LDA vectors for each paper provided by Vu et al. (2011b). If we can implement the LDA model to extract topic vectors from papers, we can apply our implementations to other citation networks, or networks whose nodes have text and that is important for the establishment of the edges. Further, a joint model with both LDA and the

network structure covariates can be available.

Fourth, in this dissertation, we obtain the learned LDA topic-vectors provided by Vu et al. (2011b), and use them directly to construct the LDA covariates. In this way, the measurement errors for the LDA covariates are not accounted for, which may cause attenuation in the estimated coefficients. One way to avoid this is to use a joint model for LDA and the egocentric models. The idea was first proposed by Vu and Hunter in a MURI project (2011). The model is more complicated than the current egocentric model and has hieratical structure. The estimation might be achieved by using a variational EM algorithm. We may also consider estimation involving the penalty functions.

Last but not the least, we observe a decreasing trend in the prediction performance for the model with only the network structure covariates, but not with the LDA model. One possible reason is that there might be some time-varying effects in the network structure covariates. We can test for the time-varying effects or build models with both time-varying covariates and time-varying coefficients. The next section considers one such model.

5.3 Aalen's Additive model

In addition to Cox's model with time-varying covarites, an alternative to the proportional hazard assumption is Aalen's additive hazard model, which allows time-varying covariates and time-varying coefficients. The model is defined as

$$h(t) = \beta_0(t) + \sum_{j=1}^{p} \beta_j(t) x_j(t). \tag{5.1}$$

Two other versions assume that $\beta_j(t), j = 1, 2, ..., p$, do not depend on t (Lin and Ying, 1994) or that $\beta_j(t) = \alpha(t), j = 0, 1, 2, ..., p$ (Aalen, 1980). Vu et al. (2011b) also use model (5.1) as an alternative for the relational approach for networks. The model can be employed in an egocentric model framework. One can also apply variable selection techniques to models with time-varying covariates.

Bibliography

Aalen, OO (1980), "A model for nonparametric regression analysis of counting processes." *Lecture Notes in Statistics*, 2, 1–25.

Akaike, H. (1973), "Maximum likelihood identification of gaussian autoregressive moving average models." *Biometrika*, 60, 255.

Allen, David M. (1974), "The relationship between variable selection and data agumentation and a method for prediction." *Technometrics*, 16, pp. 125–127.

Andersen, Per Kragh and Richard D Gill (1982), "Cox's regression model for counting processes: a large sample study." *The annals of statistics*, 1100–1120.

Beck, A. and M. Teboulle (2009), "A fast iterative shrinkage-thresholding algorithm for linear inverse problems." *SIAM Journal on Imaging Sciences*, 2, 183–202.

Bennett, S. (1983), "Analysis of survival data by the proportional odds model." *Statistics in Medicine*, 2, 273–277.

Blei, D.M., A.Y. Ng, and M.I. Jordan (2003), "Latent dirichlet allocation." *The Journal of Machine Learning Research*, 3, 993–1022.

Breiman, L. (1995), "Better subset regression using the nonnegative garrote." *Technometrics*, 37, 373–384.

Butts, C.T. (2008), "A relational event framework for social action." *Sociological Methodology*, 38, 155–200.

Cai, Jianwen, Jianqing Fan, Runze Li, and Haibo Zhou (2005), "Variable selection for multivariate failure time data." *Biometrika*, 92, 303–316.

Cox, D.R. (1972), "Regression models and life-tables." *Journal of the Royal Statistical Society. Series B*, 34, 187–220.

Cox, D.R. (1975), "Partial likelihood." *Biometrika*, 62, 269–275.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004), "Least angle regression." *The Annals of Statistics*, 32, 407–499.

Fan, J., Y. Feng, and Y. Wu (2009), "Network exploration via the adaptive lasso and scad penalties." *The Annals of Applied Statistics*, 3, 521–541.

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

Fan, Jianqing (1997), "Comments on "Wavelets in statistics: A review", by A. Antoniadis." *Journal of Italian Statisitcal Society*, 6, 131–138.

Fan, Jianqing and Runze Li (2002), "Variable selection for Cox's proportional hazards model and frailty model." *The Annals of Statistics*, 30, 74–99.

Fleming, Thomas R and David P Harrington (2011), *Counting processes and survival analysis*, volume 169. John Wiley & Sons.

Fu, Wenjiang J (1998), "Penalized regressions: the bridge versus the lasso." *Journal of computational and graphical statistics*, 7, 397–416.

Goldstein, Larry, Bryan Langholz, et al. (1992), "Asymptotic theory for nested case-control sampling in the cox regression model." *The Annals of Statistics*, 20, 1903–1928.

Golub, G.H., M. Heath, and G. Wahba (1979), "Generalized cross-validation as a method for choosing a good ridge parameter." *Technometrics*, 21, 215–223.

Hoerl, A.E. and R.W. Kennard (1970), "Ridge regression: Biased estimation for nonorthogonal problems." *Technometrics*, 12, 55–67.

Hunter, D.R. and K. Lange (2002), "Computing estimates in the proportional odds model." *Annals of the Institute of Statistical Mathematics*, 54, 155–168.

Hunter, D.R. and K. Lange (2004), "A tutorial on MM algorithms." *The American Statistician*, 58, 30–38.

Hunter, D.R. and R. Li (2005), "Variable selection using MM algorithms." *Annals of Statistics*, 33, 1617–1642.

Klein, J.P. and M.L. Moeschberger (2003), *Survival Analysis: Techniques for Censored and Truncated Data*. Springer Verlag, New York.

Kolaczyk, E.D. (2009), *Statistical Analysis of Network Data: Methods and Models*. Springer Verlag, New York.

Langholz, B (2005), *Encyclopedia of Biostatistics*, second edition, volume 1. Springer Verlag, New York. Edited by Armitage, P and Colton, T.

Lin, DY and Z. Ying (1994), "Semiparametric analysis of the additive risk model." *Biometrika*, 81, 61.

Liu, X. and D Zeng (2013), "Variable selection in semiparametric transformation models for right-censored data." *Biometrika*, 100, 1–18.

Lu, W. and H.H. Zhang (2007), "Variable selection for proportional odds model." *Statistics in medicine*, 26, 3771–3781.

McCullagh, Peter (1980), "Regression models for ordinal data." *Journal of the Royal Statistical Society.Series B (Methodological)*, 42, 109–142.

Murphy, SA, AJ Rossini, and AW Van der Vaart (1997), "Maximum likelihood estimation in the proportional odds model." *Journal of the American Statistical Association*, 92, 968–976.

Murphy, S.A. and AW Van der Vaart (2000), "On profile likelihood." *Journal of the American Statistical Association*, 95, 449–465.

Murphy, Susan A and Aad W Van Der Vaart (1999), "Observed information in semi-parametric models." *Bernoulli*, 5, 381–412.

Oakes, David (1981), "Survival times: Aspects of partial likelihood." *International Statistical Review / Revue Internationale de Statistique*, 49, 235–252.

Perry, Patrick O and Patrick J Wolfe (2013), "Point process modelling for directed interaction networks." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75, 821–849.

Raftery, Adrian E, Xiaoyue Niu, Peter D Hoff, and Ka Yee Yeung (2012), "Fast inference for the latent space network model using a case-control approximate likelihood." *Journal of Computational and Graphical Statistics*, 21, 901–919.

Schifano, E.D., R.L. Strawderman, and M.T. Wells (2010), "MM algorithms for minimizing nonsmoothly penalized objective functions." *Electronic Journal of Statistics*, 4, 1258–1299.

Schwarz, G. (1978), "Estimating the dimension of a model." *The Annals of Statistics*, 6, 461–464.

Severini, Thomas A and Wing Hung Wong (1992), "Profile likelihood and conditionally parametric models." *The Annals of Statistics*, 1768–1802.

Tibshirani, R. (1996), "Regression shrinkage and selection via the LASSO." *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

Tibshirani, Robert et al. (1997), "The lasso method for variable selection in the cox model." *Statistics in medicine*, 16, 385–395.

van der Vaart, A.W. (2000), *Asymptotic Statistics*. Cambridge University Press.

Vu, D.Q., A. U. Asuncion, D.R. Hunter, and S Padhraic (2011a), "Continuous-timeregression models for longitudinal networks." *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, to appear.

Vu, D.Q., A. U. Asuncion, D.R. Hunter, and S Padhraic (2011b), "Dynamic egocentric models for citation networks." *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, 857–864.

Wang, H., R. Li, and C.L. Tsai (2007), "Tuning parameter selector for scad." *Biometrika*, 94, 553–568.

Wu, T.T. and K. Lange (2008), "Coordinate descent algorithms for lasso penalized regression." *The Annals of Applied Statistics*, 224–244.

Zeng, D and DY Lin (2007), "Maximum likelihood estimation in semiparametric regression models with censored data." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 507–564.

Zhang, Hao Helen and Wenbin Lu (2007), "Adaptive lasso for cox's proportional hazards model." *Biometrika*, 94, 691–703.

Zhang, Y. and R. Li (2009), "Iterative conditional maximization algorithm for nonconcave penalized likelihood." *Nonparametric Statistics and Mixture Models*, 336–351.

Zhang, Y., R. Li, and C.L. Tsai (2010), "Regularization parameter selections via generalized information criterion." *Journal of the American Statistical Association*, 105, 312–323.

Zou, H. and R. Li (2008), "One-step sparse estimates in nonconcave penalized likelihood models." *Annals of Statistics*, 36, 1509.

Xizhen Cai

325 Thomas Building, University Park, PA 16802 Mobile: (814)206-4058 Email: xzc103@psu.edu

Education

Doctorate in Statistics,
Department of Statistics, the Pennsylvania State University (PSU) August, 2014
- Dissertation: Model Selection and Survival Analysis with Application to Large Time-Varying Networks

B.S. in Mathematics and Applied Mathematics,
Department of Mathematics, Zhejiang University, China June, 2008
- Thesis Title: Mathematical Models for Weather Forecast

Research Experience

Research Assistant Spring - Fall 2012, Fall 2013
Department of Statistics, PSU

Graduate Consultant Spring 2010, Fall 2010
Statistical Consulting Center, PSU

Teaching Experience

Graduate Instructor, Department of Statistics, PSU

STAT/MATH 418 (Probability Theory)	Spring 2011
STAT/MATH 414 (Intro to Probability)	Summer 2011,Fall 2011
STAT 462 (Applied Regression Analysis)	Summer 2013

Teaching Assistant, Department of Statistics, PSU

STAT 100 (Stat Concepts and Reasoning)	Fall 2008 - Spring 2010
STAT 503 (Design of Experiments)	Fall 2010
STAT 513 (Theory of Statistics I)	Fall 2010
STAT 200 (Elementary Statistics)	Spring 2013

Manuscript

Cai, X. and Hunter, D.R., Theory and Algorithms for Penalized Proportional Odds Models.

Awards

- **1st place in Ph.D candidacy exam**, 2010
  Department of Statistics, PSU.
- **The William L. Harkness Graduate Teaching Award**, 2012
  Department of Statistics, PSU.
- Student Travel Grant for Joint Statistical Meetings, 2012 and 2013
  Department of Statistics, PSU.

Conference Participation

Presentations
- August, 2012. Joint Statistical Meetings, San Diego. *Variable Selection Algorithms for the Proportional Odds Model.*
- August, 2013. Joint Statistical Meetings, Montréal, Canada. *Sufficient Statistic Selection for Dynamic Networks.*
- October, 2013. The Rao Prize Conference, the Pennsylvania State University. *Sufficient Statistic Selection for Dynamic Networks.*