Penalized Generalized Quasi-Likelihood Based Variable Selection for Longitudinal Data

Tharshanna Nadarajah, Asokan Mulayath Variyath, and J. Concepción Loredo-Osti

Abstract High-dimensional longitudinal data with a large number of covariates, have become increasingly common in many bio-medical applications. The identification of a sub-model that adequately represents the data is necessary for easy interpretation. Also, the inclusion of redundant variables may hinder the accuracy and efficiency of estimation and inference. The joint likelihood function for longitudinal data is challenging, particularly in correlated discrete data. To overcome this problem Wang et al. (Biometrics 68:353–360, 2012) introduced penalized GEEs (PGEEs) with a non-convex penalty function which requires only the first two marginal moments and a working correlation matrix. This method works reasonably well in high-dimensional problems; however, there is a risk of model mis-specification such as variance function and correlation structure and in such situations, we propose variable selection based on penalized generalized quasi-likelihood (PGQL). Simulation studies show that when model assumptions are true, the PGQL method has performance comparable with that of PGEEs. However, when the model is mis-specified, the PGQL method has clear advantages over the PGEEs method. We have implemented the proposed method in a real case example.

Keywords GEEs • Generalized quasi-likelihood • Longitudinal data • Variable selection

1 Introduction

Variable selection is an important topic in statistical modeling, especially for longitudinal data with high-dimensional covariates, which often arise in large-scale bio-medical studies. In practice, a large number of covariates, $(X_1, X_2, \ldots, X_p)$, are believed to have an influence on the response variable y of interest. However, some covariates have no influence or a weak influence, and a regression model that includes all the covariates is not advisable. Excluding the unimportant covariates

T. Nadarajah (✉) • A.M. Variyath • J. C. Loredo-Osti
Memorial University, St. John's, NL, Canada A1C5S7
e-mail: nadarajah.tharshana@mun.ca; variyath@mun.ca; jcloredoosti@mun.ca

B.C. Sutradhar (ed.), *Advances and Challenges in Parametric and Semi-parametric Analysis for Correlated Data*, Lecture Notes in Statistics 218,
DOI 10.1007/978-3-319-31260-6_8

results in a simpler model with better interpretive and predictive value. The problem of identifying a sub-model that adequately represents the response is generally referred to as the variable selection problem. For example, in generalized linear models, the sub-model that relates to the random variable y with the mean denoted by μ to a subset of components of X in the form

$$g(x;\mu) = X(s)\beta(s),$$

where $g(*)$ is the link function, $X(s)$ is a subset of the components of X, $\beta(s)$ is a vector of the corresponding regression parameters, and $s \subseteq (1, 2, \ldots, p)$. The variable selection problem is to find the best subset s such that the sub-model is optimal according to some criteria that give a good description of the data-generating mechanism. Statistically speaking, variable selection is a way to reduce the complexity of the model, in some cases by accepting a small amount of bias to improve the precision.

Traditionally, variable selection is achieved by evaluating all possible sub-models via information criteria such as Akaike's information criterion (AIC; Akaike 1973, 1974), Bayesian information criterion (BIC; Schwarz 1978), and Empirical-likelihood-based information-theoretic approaches (EAIC, EBIC; Variyath et al. 2010). The sub-model that minimizes the information criteria is then selected together with the corresponding covariates. For high dimensional data, which are often encountered in modern applications, the computational burden makes the direct application of these information criteria infeasible. To overcome the computational difficulties as well as to achieve some selection stability, regularization methods have drawn substantial attention. There is a large volume of literature on the penalized likelihood approach for building such models; for example, least absolute shrinkage and selection operator (LASSO; Tibshirani 1996) and the smoothly clipped absolute deviation (SCAD; Fan and Li 2001). Both approaches have many desirable properties. Other related variable selection methods include penalized empirical likelihood-based variable selection (Nadarajah 2011; Variyath 2006), adaptive LASSO (Zhang and Lu 2007; Zou 2006), least-square approximation (Wang and Leng 2007) and the folded concave penalty method (Lv and Fan 2009). However, the aforementioned variable selection methods are only applicable to generalized linear regression models.

Variable selection for longitudinal data is quite challenging due to the high dimensionality of covariates and the correlation within subject. Pan (2001) developed a quasi-likelihood information criterion (QIC) under the working independence model and the naive and robust covariance estimates of estimated regression coefficients. Cantoni et al. (2005) proposed a generalized version of Mallows's C_p suitable for use with both parametric and nonparametric models. This model selection avoids a stepwise procedure and is based on a measure of predictive error rather than on significance testing. Wang and Qu (2009) introduced a novel Bayesian information criterion type model selection procedure based on the quadratic inference function, which does not require the full likelihood. The implementation of best

subset type model selection procedures call for the evaluation of all possible sub-models, which becomes computationally intensive when the number of covariates is moderately large.

The idea of penalization is very useful in longitudinal modeling, particularly in high dimensional variable selection. Fan and Li (2004) proposed an innovative class of variable selection procedures to select significant variables in the semi-parametric models for continuous responses. Wang et al. (2008) studied regularized estimation procedures for nonparametric varying coefficient models for continuous responses that can simultaneously perform variable selection and the estimation of smooth coefficient functions. Xiao et al. (2009) recently investigated a double-penalized likelihood approach for selecting important parametric fixed effects in semiparametric mixed models for continuous responses. Dziak et al. (2009) discussed the applications of the SCAD-penalized quadratic inference function. Xu et al. (2010) investigated a GEEs-based shrinkage estimator with an artificial objective function. Xue et al. (2010) considered the model selection of a generalized additive model when responses from the same cluster are correlated. However, the aforementioned methods assume that the dimension of predictors are relatively small and some of these works are only applicable to continuous responses. The joint likelihood function for correlated discrete responses does not have a closed form when the correlations among the repeated measures are taken into account. To avoid specifying the full joint likelihood for correlated data, Wang et al. (2012) proposed penalized GEEs with a non-convex penalty function, which requires only the specification of the first two marginal moments and a working correlation matrix. This penalized GEEs method is superior to traditional methods because of their computational efficiency and stability. The true regression coefficients that are zero are automatically shrunk to zero, and the remaining coefficients are simultaneously estimated. These methods work reasonably well in high-dimensional problem; however, the GEEs estimate of β is not necessarily consistent in some situations, as discussed by Crowder (1995) and Sutradhar and Das (1999). To overcome this problem, Sutradhar (2003) has proposed a generalization of the quasi-likelihood (GQL) approach to improve the efficiency of the parameter estimates. Utilizing this, to avoid the risk of model mis-specifications such as in variance function and correlation structure, we propose penalized generalized quasi-likelihood (PGQL) based on stationary lag correlation structure. Our simulation studies show that the proposed method works well compared to PGEEs.

The remaining part of the paper is organized as follows. In Sect. 2, we discussed the importance of the GQL approach and compared its performance with the GEEs approach. In Sect. 3, we introduced PGQL-based variable selection for longitudinal data. Its theoretical properties and numerical algorithm are also explored in this section. In Sect. 4, the performance analysis of the proposed method is assessed based on Monte Carlo simulations. The proposed method is applied to health care utilization count data in Sect. 5 and our conclusions are given in Sect. 6.

2 Generalized Quasi-Likelihood

The structure of the longitudinal data-set consists of an outcome random variable y_{it} and p-dimensional vector of covariates x_{it} that are observed for subjects $i = 1, \ldots, k$ at a time point t, $t = 1, \ldots, m_i$. For the ith subject, let $y_i = (y_{i1}, \ldots, y_{im_i})^T$ be the response vector and let $X_i = (x_{i1}, x_{i2}, \ldots, x_{im_i})^T$ be the $m_i \times p$ matrix of covariates. We assume that the k subjects are independent while the repeated measurements y_{it} taken on each subject are correlated. The marginal density of y_{it} is assumed to follow a canonical exponential family (Liang and Zeger 1986) of the form

$$f(y_{it}) = \exp\left[(y_{it}\theta_{it} - a(\theta_{it}))\phi + b(y_{it}, \phi)\right], \tag{1}$$

where $\theta_{it} = g(\eta_{it})$, g is the known injective function with $\eta_{it} = x_{it}\beta$, β is a $p \times 1$ vector of the regression effects of x_{it} on y_{it}, $a(*)$, and $b(*)$ are known functional forms. The mean and the variance of y_{it} as $\mathrm{E}(y_{it}|x_{it}) = a'(\theta_{it}) = \mu_{it}$, and $\mathrm{Var}(y_{it}) = a''(\theta_{it}) = v(\mu_{it})\phi$, where ϕ is the unknown over-dispersion parameter for simplicity we assume $\phi = 1$ in this study and $v(*)$ is a known variance function.

Note that when there is a functional relationship between the mean and variance of the response, Wedderburn (1974) proposed a quasilikelihood (QL) approach for independence data which utilize both the mean and the variance in estimating the regression effects. When there is insufficient information about the data for us to specify a parametric model, quasi-likelihood is often used. The QL optimal estimating equation for β is given by

$$\sum_{i=1}^{k}\sum_{t=1}^{m_i}\left[\frac{\partial a'(\theta_{it})}{\partial\beta}\frac{(y_{it} - a'(\theta_{it}))}{\mathrm{Var}(y_{it})}\right] = 0.$$

In a longitudinal setup, the components of the response vector y_i are repeated, which are likely to be correlated. Let $C_i(\rho)$ be the $m_i \times m_i$ true correlation matrix of y_i, $i = 1, \ldots, k$, which is unknown in practice. Our primary interest is to estimate β after taking the longitudinal correlation $C_i(\rho)$ into account. For known $C_i(\rho)$, the QL estimator of β under (1) is the solution of the score equation

$$g(y;\beta) = \sum_{i=1}^{k} X_i^T A_i \Sigma_i^{-1}(\rho)(y_i - \mu_i) = 0, \tag{2}$$

where $A_i = \mathrm{diag}\left[a''(\theta_{i1}), \ldots, a''(\theta_{it}), \ldots, a''(\theta_{im_i})\right]$, and $\Sigma_i(\rho) = A_i^{1/2}C_i(\rho)A_i^{1/2}$ is the true covariance of y_i. In real applications the true correlation structure is often unknown. Ignoring the correlation among the same individual could lead to an inefficient estimation of the regression coefficients and underestimation of standard errors.

To overcome these problems, in a seminal paper Liang and Zeger (1986) proposed the generalized estimating equations (GEEs) approach. The GEEs approach

for estimating the parameter vector of the marginal regression model (1) allows the user to specify any working correlation structure for the correlation matrix of a subject's outcomes y_i. They developed the joint probability model by introducing a "working" correlation structure based on a generalized estimating equations approach to obtain consistent and efficient estimators for the regression parameter β, given by

$$g(\beta, \hat{\alpha}(\beta)) = \sum_{i=1}^{k} X_i^T A_i^{1/2} R_i^{-1}(\hat{\alpha}) A_i^{-1/2}(y_i - \mu_i) = 0, \tag{3}$$

where $A_i = m_i \times m_i$ diagonal matrix with $\text{Var}(\mu_{it})$ as the *ith* diagonal element, $R_i(\hat{\alpha})$ is the "working" correlation matrix of the m_i repeated measures used for $C_i(\rho)$ in equation (2). We can choose the form of the $m_i \times m_i$ "working" correlation matrix R_i for each y_i, defined by the (j, j') element of R_i is the known, hypothesized, or estimated correlation between y_{ij} and $y_{ij'}$. The working correlation structure can depend on an unknown $s \times 1$ correlation parameter vector α. The observation times and correlation matrix can differ from subject to subject, but the correlation matrix $R_i(\alpha)$ of the ith subject is fully specified by α. For a given working correlation structure, α can be estimated using a residual-based method of moments. The GEEs estimate of β is not necessarily consistent in some situations as discussed by Crowder (1995) and Sutradhar and Das (1999). Crowder (1995) demonstrated that there may not be any solutions for $\hat{\alpha}$, which misleads the estimation of regression parameters. Also the GEEs approach gives a consistent estimator of β, but this estimator in some situations is less efficient than the independence estimating equation approach under an arbitrary working correlation structure as shown by Sutradhar and Das (1999).

In such situations, Sutradhar (2003) has proposed a generalization of the quasi-likelihood (GQL) approach to improve the efficiency of the parameter estimates. The estimation for β is obtained by solving the GQL estimating equations given by

$$g(\beta, \rho) = \sum_{i=1}^{k} X_i^T A_i \Sigma_i^{-1}(\hat{\rho})(y_i - \mu_i) = 0, \tag{4}$$

where $\Sigma_i(\hat{\rho}) = A_i^{1/2} C_i^*(\rho) A_i^{1/2}$, with $C_i^*(\rho)$ as the stationary lag-correlation structure for any of the AR(1), MA(1), or EQC models, and

$$C_i^*(\rho) = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{m-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{m-2} \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \rho_{m-1} & \rho_{m-2} & \rho_{m-3} & \dots & 1 \end{bmatrix}. \tag{5}$$

Table 1 A class of stationary correlation models for longitudinal count data from Sutradhar (2011, Sect. 6.3)

Model	Dynamic relationship	Mean, variance and correlations
AR(1)	$y_{it} = \rho * y_{i,t-1} + d_{it},\ t = 2, 3, \ldots, m$ $y_{i1} \sim \text{Poi}(\mu_i = \exp[\tilde{X}_i\beta])$ $d_{it} \sim \text{Poi}(\mu_i(1-\rho)),\ t = 2, 3, \ldots, m$	$\text{E}[y_{it}] = \mu_i$ $\text{var}[y_{it}] = \mu_i$ $\text{corr}[y_{it}, y_{i,t+l}] = \rho_l = \rho^l$
MA(1)	$y_{it} = \rho * d_{i,t-1} + d_{it},\ t = 1, 2, \ldots, m$ $d_{i0} \sim \text{Poi}(\mu_i/(1+\rho))$ $d_{it} \sim \text{Poi}(\mu_i/(1+\rho))\ t = 1, 2, \ldots, m$	$\text{E}[y_{it}] = \mu_i$ $\text{var}[y_{it}] = \mu_i$ $\text{corr}[y_{it}, y_{i,t+l}] = \rho_l$ $= \frac{\rho}{(1+\rho)}$ for l=1
EQC	$y_{it} = \rho * y_{i1} + d_{it},\ t = 2, 3, \ldots, m$ $y_{i1} \sim \text{Poi}(\mu_i)$ $d_{it} \sim \text{Poi}(\mu_i(1-\rho)),\ t = 2, 3, \ldots, m$	$\text{E}[y_{it}] = \mu_i$ $\text{var}[y_{it}] = \mu_i$ $\text{cor}[y_{it}, y_{i,t+l}] = \rho_l = \rho$

The stationary lag-correlations are estimated by the method of moments introduced by Sutradhar and Kovacevic (2000) and given by

$$\hat{\rho}_l = \frac{\sum_{i=1}^{k}\sum_{t=1}^{m-l}\tilde{y}_{it}\tilde{y}_{i,t+l}/k(m-l)}{\sum_{i=1}^{k}\sum_{t=1}^{m}\tilde{y}_{it}^2/km}, \tag{6}$$

where $l = |t - t'|,\ t \neq t',\ t, t' = 1, \ldots, m$ and $\tilde{y}_{it}$ is the standardized residual, defined as $\tilde{y}_{it} = \{y_{it} - \mu_{it}\}/\{a''(\theta_{it})\}^{1/2}$. The stationary lag correlation approach produces consistent as well as more efficient regression estimates as compared to the independence assumption-based estimating equation approaches (Sutradhar and Das 1999). We conducted a small simulation study to illustrate the comparison of GEEs with GQL under a mis-specified correlation structure. A class of stationary correlation AR(1) model for longitudinal count data are generated as per the dynamic relationship given in Table 1, which are discussed by McKenzie (1988), and Sutradhar (2011). The stationary covariates $\tilde{x}_i = (\tilde{x}_{i1}, \tilde{x}_{i2})$ are generated from the normal distribution with mean 0, variance 1, and $\beta = (0.3, 0.2)^T$. For a given $y_{i,t-1}$, $\rho * y_{i,t-1}$ denotes the commonly called binomial thinning operation discussed by McKenzie (1988). That is, $\rho * y_{i,t-1} = \sum_{j=1}^{y_{i,t-1}} b_j(\rho)$ with $\Pr[b_j(\rho) = 1] = \rho$, $\Pr[b_j(\rho) = 0] = 1-\rho$. For the simulation's purpose, we consider the number of time points $m = 5,\ 10$ and number of subjects $k = 100$. We simulated 1000 data sets with $\rho = 0.49, 0.70$ from the above AR(1) stationary dynamic model to generate the data. Under the working exchangeable and MA(1) correlation structure, the correlation parameter $\hat{\alpha}$ can be estimated by using Eqs. (2.5) and (2.7) from Sutradhar and Das (1999). The mean of the estimated values of the regression coefficients and the

Table 2 Coverage probabilities of regression estimates for a true AR(1) correlation model data under different "working" correlation models (m = 5)

True model	Method	Parameters	Estimates	Coverage probabilities	
				95 % level	99 % level
AR(1) $\rho = 0.70$	GEEs (AR(1))	β_1	0.3000 (0.070)	0.952 (0.279)	0.987 (0.367)
		β_2	0.2009 (0.073)	0.950 (0.286)	0.988 (0.375)
	GEEs (EQC)	β_1	0.2997 (0.073)	0.911 (0.247)	0.973 (0.325)
		β_2	0.1956 (0.076)	0.902 (0.252)	0.963 (0.332)
	GQL	β_1	0.3003 (0.070)	0.952 (0.278)	0.986 (0.366)
		β_2	0.2007 (0.073)	0.950 (0.284)	0.988 (0.374)
AR(1) $\rho = 0.49$	GEEs (AR(1))	β_1	0.2989 (0.062)	0.938 (0.237)	0.988 (0.319)
		β_2	0.1956 (0.062)	0.940 (0.243)	0.981 (0.319)
	GEEs (EQC)	β_1	0.2992 (0.061)	0.899 (0.206)	0.968 (0.272)
		β_2	0.1986 (0.062)	0.908 (0.211)	0.980 (0.278)
	GEEs (MA(1))	β_1	0.2991 (0.062)	0.897 (0.205)	0.968 (0.270)
		β_2	0.1985 (0.062)	0.905 (0.210)	0.981 (0.276)
	GQL	β_1	0.2989 (0.061)	0.931 (0.235)	0.990 (0.309)
		β_2	0.1955 (0.061)	0.936 (0.241)	0.992 (0.317)

corresponding simulated standard errors in parentheses are reported in Tables 2 and 3 for different $m = 5,\ 10$ respectively. We also report the coverage probabilities as well as the width of the confidence interval for β_1 and β_2 in parentheses for confidence levels 0.95 and 0.99. In this study, we generated the data using the AR(1) correlation structure even though we used all three working correlation structures: AR(1), EQC, and MA(1) for parameter estimation under GEEs and the results are compared with the GQL approach.

We see from Tables 2 and 3 that when we use the true working correlation structure, coverage probabilities based on the GEEs and GQL approaches are

Table 3 Coverage probabilities of regression estimates for a true AR(1) correlation model data under different "working" correlation models (m = 10)

True model	Method	Parameters	Estimates	Coverage probabilities	
				95 % level	99 % level
AR(1) $\rho = 0.70$	GEEs (AR(1))	β_1	0.2961 (0.057)	0.956 (0.228)	0.992 (0.299)
		β_2	0.1978 (0.059)	0.951 (0.232)	0.988 (0.306)
	GEEs (EQC)	β_1	0.3020 (0.059)	0.811 (0.159)	0.922 (0.209)
		β_2	0.1993 (0.062)	0.802 (0.163)	0.919 (0.214)
	GQL	β_1	0.2960 (0.056)	0.955 (0.226)	0.991 (0.231)
		β_2	0.1977 (0.057)	0.944 (0.231)	0.986 (0.303)
AR(1) $\rho = 0.49$	GEEs (AR(1))	β_1	0.2977 (0.047)	0.945 (0.181)	0.990 (0.238)
		β_2	0.1994 (0.048)	0.937 (0.185)	0.987 (0.243)
	GEEs (EQC)	β_1	0.2985 (0.046)	0.848 (0.135)	0.952 (0.178)
		β_2	0.1975 (0.050)	0.826 (0.138)	0.930 (0.181)
	GEEs (MA(1))	β_1	0.3011 (0.047)	0.842 (0.134)	0.945 (0.177)
		β_2	0.1985 (0.048)	0.847 (0.137)	0.943 (0.180)
	GQL	β_1	0.2981 (0.044)	0.954 (0.178)	0.990 (0.233)
		β_2	0.2000 (0.046)	0.949 (0.182)	0.988 (0.239)

almost the same. However, under an arbitrary working correlation structure, the GQL approach performs better than the GEEs approach. This result shows a loss of efficiency of the GEEs estimators due to mis-specification of the correlation structures. We also notice that when m increases GQL has better coverage compared to GEEs based confidence interval. Rather than using any arbitrary "working correlation", it seems much better to define a lag-correlation structure for the longitudinal responses to estimate the parameters. The correlation structure (5) is quite robust, and it accommodates all three correlation structures: AR(1), EQC, and MA(1). Note, however, that the correlation structure is unknown in practice and it makes more sense to use a stationary lag-correlation structure to represent all the

three correlation structures in a unique way. We did not consider other cases, for instance in a true EQC and MA(1) correlation models, since under different working correlation structure the correlation parameter $\hat{\alpha}$ does not exist.

3 Penalized Generalized Quasi-Likelihood (PGQL)

We use the GQL approach discussed in Sect. 2 and the SCAD penalty function (Fan and Li 2001), to develop the penalized generalized quasi-likelihood for variable selection in the context of a longitudinal data analysis. The regression parameters are estimated by solving the penalized generalized estimating functions

$$\mathscr{U}(\beta) = g(\beta, \hat{\rho}(\beta)) - kp'_{\delta}(|\beta|)\text{sign}(\beta) \tag{7}$$

where $g(\beta, \hat{\rho}(\beta)) = \sum_{i=1}^{k} X_i^T A_i^{1/2} C_i^{*-1}(\rho) A_i^{-1/2}(y_i - \mu_i) = 0$ be the GQL estimating equation given in (4), $p'_{\delta}(*)$ is the first derivative of penalty function, $\text{sign}(\beta) = (\text{sign}(\beta_1), \ldots, \text{sign}(\beta_p))^T$ with $\text{sign}(t) = \text{I}(t > 0) - \text{I}(t < 0)$, and δ is the tuning parameter. Different penalty functions can be potentially adopted in this modeling. The HARD thresholding penalty proposed by Fan (1997) and Antoniadis (1997) is defined as $p_{\delta}(|\theta|) = \delta^2 - (|\theta| - \delta)^2 I(|\theta| < \delta)$. For a large value of $|\theta|$, the HARD thresholding penalty does not overpenalize. The LASSO penalty function is the L_1-penalty, $p_{\delta}(|\theta|) = \delta|\theta|$, proposed by Donoho and Johnstone (1994) in the wavelet setting and extended by Tibshirani (1996) to general likelihood settings. The penalty function used in ridge regression is the L_2 penalty, $p_{\delta}(|\theta|) = \delta|\theta|^2$. According to Fan and Li (2001), a good penalty function should result in an estimator with the following three oracle properties:

1. Unbiasedness: To avoid unnecessary modeling bias, the estimator is nearly unbiased when the true unknown parameter is large.
2. Sparsity: This is a thresholding rule that automatically sets small estimated coefficients to zero to reduce the model complexity.
3. Continuity: This property eliminates unnecessary variation in the model prediction.

However, the penalty functions L_1, L_2, and HARD do not satisfy all three oracle properties. A simple penalty function satisfying all three is the SCAD penalty proposed by Fan (1997) where the first derivative is

$$p'_{\delta}(\theta) = \delta \left\{ \text{I}(\theta \leq \delta) + \frac{(a\delta - \theta)_+}{(a-1)\delta} \text{I}(\theta > \delta) \right\} \quad \text{for some } a > 2 \text{ and } \theta > 0. \tag{8}$$

Necessary conditions for the unbiasedness, sparsity, and continuity of the SCAD penalty have been given by Antoniadis and Fan (2001). This penalty function involves two unknown parameters, a and δ. As shown in Fig. 1, all the penalty functions are singular at the origin, satisfying $p_{\delta}(0+) > 0$. This is the necessary

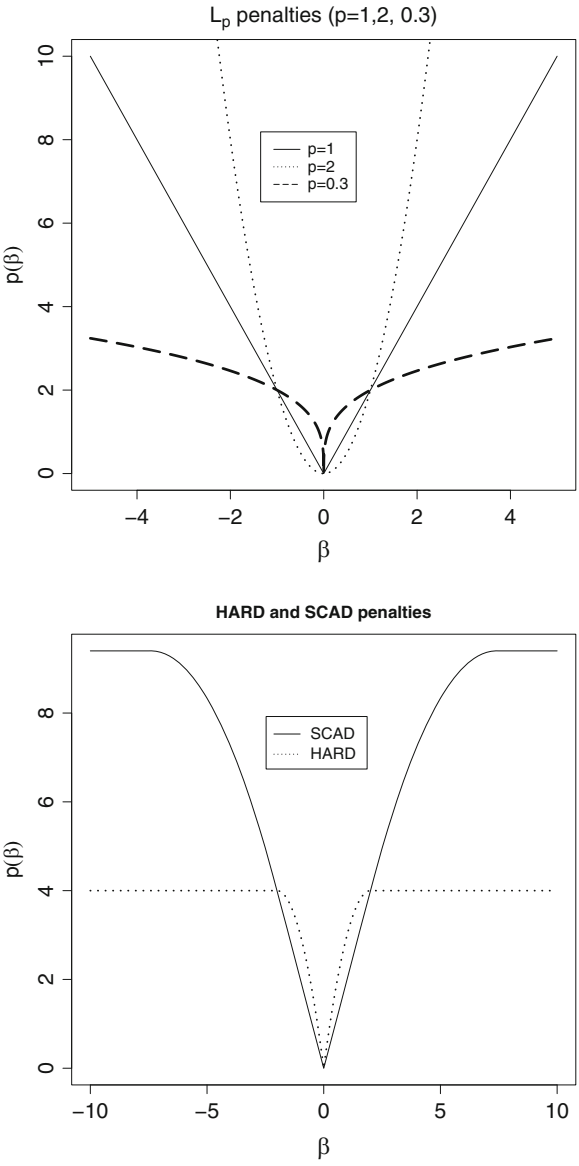

Fig. 1 L_p, SCAD, and HARD penalty functions and their quadratic approximation

condition for sparsity in variable selection. As shown in Fig. 1, the HARD and SCAD penalties are constant when β is large, indicating that there is no excessive penalization for large regression coefficients. However, SCAD is smoother than HARD and hence yields a continuous estimator.

By following Fan and Li (2001), we solve (7) to carry out estimation and variable selection simultaneously and arrive the penalized GQL estimates of the regression parameters, $\hat{\beta}_{PGQL}$, which has properties similar to $\hat{\beta}_{PGEEs}$.

3.1 Numerical Algorithm

To implement our method, we need an efficient numerical algorithm. The SCAD penalty function involves two unknown parameters, δ and a. From a Bayesian point of view, Fan and Li (2001) suggested setting $a = 3.7$ and using generalized cross-validation (GCV; Craven and Wahba 1979) to select the best value of δ, which is implemented in our simulation studies. We maximize the PGQL with respect to β given in (7). We used the modified Newton-Raphson algorithm proposed by Fan and Li (2001), which is numerically stable. At each iteration, we compute the stationary lag-correlation structure $C^*(\rho)$ given in (5) for an updated value of β. A step-by-step numerical algorithm for estimating $\hat{\beta}_{PGQL}$ for a given value of the tuning parameter δ is given below.

(a) Set $\beta = \beta^0$, and $\epsilon = 1e-08$.
(b) Let $C^*(\rho)$ be the estimated value of $C^*(\rho)$.
(c) The parameter $\hat{\beta}_{PGQL}$ is computed iteratively and the solution at the $(h+1)$th iteration is given by

$$\hat{\beta}^{(h+1)} = \hat{\beta}^{(h)} + \left\{H(\hat{\beta}^h) + kS_\delta(\hat{\beta}^h)\right\}^{-1} \left\{U(\hat{\beta}^h) - kS_\delta(\hat{\beta}^h)\hat{\beta}^h\right\} \tag{9}$$

where

$$H(\hat{\beta}^h) = \sum_{i=1}^{k} X_i^T A_i \Sigma_i^{-1}(\hat{\rho}) A_i X_i, \;\; \Sigma_i \text{ is defined in (4)},$$
$$S_\delta(\hat{\beta}^h) = \text{diag}\left[\frac{p'_\delta(|\hat{\beta}_1^h|)}{|\hat{\beta}_1^h|}, \ldots, \frac{p'_\delta(|\hat{\beta}_p^h|)}{|\hat{\beta}_p^h|}\right].$$

(d) If $\min\left|\hat{\beta}^{(h+1)} - \hat{\beta}^{(h)}\right| < \epsilon$ stop the algorithm and report $\hat{\beta}^{(h+1)}$; otherwise, $h = h+1$ and go to Step 3.

4 Performance Analysis

4.1 Stationary Count Data

A class of stationary correlation models for longitudinal count data are considered for our simulation studies, which are given in Table 1. For the purpose of the

simulation, we consider five covariates $\tilde{X}_i = (\tilde{x}_{i_1}, \ldots, \tilde{x}_{i_5})$, assumed to have distributions as $\tilde{x}_{i1} \sim$ Bernoulli(0.5), $\tilde{x}_{i2}$ to $\tilde{x}_{i5}$ from the standard normal with $\text{cor}(x_i, x_j) = 0.5^{|i-j|}$, and $\beta = (0.5, 0.5, 0.6, 0, 0)^T$. The performance analysis is carried out based on two measures, viz (a) the median of the relative model error (MRME) and (b) the average number of correct zero and non-zero coefficients. The estimated values of the nonzero coefficients and the corresponding simulated standard errors are also reported. The model error (ME) is defined as $\text{ME}(\hat{\beta}) = E_x \left\{\mu(X\beta) - \mu(X\hat{\beta})\right\}^2$, where $\mu(X\beta) = E(y|X)$ and computed the relative model error as $\text{RME} = \text{ME}/\text{ME}_{\text{full}}$ where ME_{full} is the model error calculated by fitting the data with the full model and ME is the model error of the selected model. For the purpose of this simulation, we consider the number of time points $m = 5, 10$ and number of subjects $k = 100$. The entire simulation study was repeated 1000 times for all the three true models and a summary of the performance measures is given in Tables 4 and 5. From Tables 4 and 5 we see that, when we use the true working correlation structure, the MRME of PGEEs is very close to the MRME arrived based on PGQL. The average number of zero coefficients for PGEEs and PGQL are closer to the target of two and the nonzero regression parameter estimates are close to the true values in all cases. This shows, the bias of the non-zero parameter estimates are approximately zero. However, the proposed PGQL approach has smaller MRME compared to PGEEs under an arbitrary working correlation structure. Also, we noticed the average number of zero coefficients for PGQL is closer to the target of two in all cases but not for PGEEs with a wrong correlation structure. We repeated the simulation with smaller sample sizes, and the overall conclusions were similar, so they have not been provided here. These simulations studies clearly shows that PGQL is performing better compared to PGEEs with mis-specified working approaches. Note that, in practice, it is very difficult to know the true correlation structure, so the PGQL approach is preferred for the variable selection as well as estimation of regression parameters.

4.2 Over-Dispersed Stationary Count Data

In this section, we consider the performance of our method when the model is misspecified in the context of stationary count data. We generate over-dispersed stationary count data y_{it} using $\tilde{\mu}_i = u_i \exp(\tilde{x}_i\beta)$ on the models which are discussed in Table 1 with u_i a random sample such that $\text{E}(u_i) = 1$ and $\text{Var}(u_i) = \omega$. Marginally, we have $\text{E}(y_{it}) = \tilde{\mu}_i$ and $\text{Var}(y_{it}) = \tilde{\mu}_i(1 + \tilde{\mu}_i\omega)$. The distribution of u is chosen to be gamma with parameters $(\omega, 1/\omega)$ with ω being the over-dispersion parameter. For the purpose of the simulation, we consider five covariates $\tilde{X}_i = (\tilde{x}_{i_1}, \ldots, \tilde{x}_{i_5})$, assumed to have distributions as $\tilde{x}_{i1} \sim$ Bernoulli(0.5), $\tilde{x}_{i2}$ to $\tilde{x}_{i5}$ are generated from a multivariate normal distribution with a mean of zero, and the correlation between x_i and x_j is $0.5^{|i-j|}$, $\omega = 0.25$, and $\beta = (0.5, 0.5, 0.6, 0, 0)^T$. In each simulation, we generated $m = 5$ repeated over-dispersed count data for a sample of

Table 4 Performance measures for count data with stationary covariates (m = 5)

True model	Method	MRME%	Avg. no. of zero coefficients		Estimates of nonzero coefficients		
			Correct	Incorrect	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
AR(1) $\rho = 0.70$	PGEEs(IND)	86.86	1.25	0.0	0.5002 (0.068)	0.5023 (0.075)	0.5909 (0.076)
	PGEEs(AR(1))	65.60	1.85	0.0	0.5029 (0.066)	0.5058 (0.073)	0.5930 (0.071)
	PGEEs(EQC)	69.76	1.84	0.0	0.5030 (0.066)	0.5047 (0.073)	0.5935 (0.072)
	PGQL	66.90	1.85	0.0	0.5034 (0.066)	0.5052 (0.073)	0.5930 (0.071)
AR(1) $\rho = 0.49$	PGEEs(AR(1))	63.60	1.80	0.0	0.5003 (0.056)	0.5025 (0.056)	0.5968 (0.057)
	PGEEs(MA(1))	70.57	1.65	0.0	0.5011 (0.060)	0.5021 (0.061)	0.5986 (0.062)
	PGQL	67.64	1.79	0.0	0.5014 (0.059)	0.5029 (0.060)	0.5973 (0.061)
EQC $\rho = 0.70$	PGEEs(IND)	77.95	1.23	0.0	0.5059 (0.074)	0.5053 (0.076)	0.5913 (0.080)
	PGEEs(EQC)	61.43	1.87	0.0	0.5022 (0.073)	0.5066 (0.076)	0.5921 (0.074)
	PGEEs(AR(1))	63.37	1.70	0.0	0.5025 (0.074)	0.5066 (0.076)	0.5914 (0.076)
	PGQL	62.61	1.87	0.0	0.5026 (0.073)	0.5069 (0.076)	0.5916 (0.073)
EQC $\rho = 0.49$	PGEEs(EQC)	65.39	1.82	0.0	0.5023 (0.062)	0.5057 (0.065)	0.5922 (0.064)
	PGEEs(MA(1))	75.50	1.59	0.0	0.4996 (0.065)	0.5022 (0.068)	0.5980 (0.073)
	PGQL	66.40	1.82	0.0	0.5017 (0.064)	0.5046 (0.068)	0.5938 (0.069)
MA(1) $\rho = 0.67$	PGEEs(IND)	70.29	1.54	0.0	0.5002 (0.052)	0.5006 (0.054)	0.5994 (0.054)
	PGEEs(MA(1))	63.56	1.72	0.0	0.5004 (0.052)	0.4993 (0.053)	0.5981 (0.052)
	PGEEs(AR(1))	69.39	1.78	0.0	0.5018 (0.052)	0.5013 (0.056)	0.5963 (0.054)
	PGEEs(EQC)	71.37	1.71	0.0	0.5006 (0.052)	0.5008 (0.056)	0.5974 (0.058)
	PGQL	65.20	1.75	0.0	0.5005 (0.051)	0.5004 (0.053)	0.5970 (0.053)

Table 5 Performance measures for count data with stationary covariates (m = 10)

True model	Method	MRME%	Avg. no. of zero coefficients		Estimates of nonzero coefficients		
			Correct	Incorrect	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
AR(1) $\rho = 0.70$	PGEEs (IND)	84.21	0.93	0.0	0.5007 (0.052)	0.5017 (0.056)	0.5963 (0.063)
	PGEEs (AR(1))	60.96	1.76	0.0	0.5000 (0.051)	0.5022 (0.053)	0.5967 (0.054)
	PGEEs (EQC)	60.56	1.79	0.0	0.5000 (0.052)	0.5020 (0.055)	0.5981 (0.055)
	PGQL	60.32	1.76	0.0	0.5003 (0.051)	0.5005 (0.053)	0.5988 (0.053)
AR(1) $\rho = 0.49$	PGEEs (AR(1))	71.54	1.69	0.0	0.5010 (0.042)	0.5002 (0.043)	0.5996 (0.043)
	PGEEs (MA(1))	81.85	1.27	0.0	0.5012 (0.042)	0.4989 (0.044)	0.6004 (0.046)
	PGQL	77.21	1.67	0.0	0.5010 (0.041)	0.5013 (0.043)	0.5980 (0.044)
EQC $\rho = 0.70$	PGEEs (IND)	99.04	0.743	0.0	0.5009 (0.074)	0.5041 (0.078)	0.5935 (0.086)
	PGEEs (EQC)	63.46	1.87	0.0	0.5014 (0.072)	0.5058 (0.073)	0.5928 (0.075)
	PGEEs (AR(1))	69.42	1.67	0.0	0.5026 (0.072)	0.5032 (0.076)	0.5943 (0.076)
	PGQL	64.15	1.86	0.0	0.5010 (0.071)	0.5050 (0.074)	0.5945 (0.075)
MA(1) $\rho = 0.67$	PGEEs(IND)	83.44	1.34	0.0	0.5015 (0.034)	0.4991 (0.037)	0.6005 (0.038)
	PGEEs (EQC)	71.85	1.56	0.0	0.5008 (0.033)	0.5010 (0.036)	0.5976 (0.038)
	PGEEs (AR(1))	72.39	1.62	0.0	0.4998 (0.033)	0.5005 (0.036)	0.5974 (0.038)
	PGQL	71.12	1.59	0.0	0.5001 (0.032)	0.5000 (0.035)	0.5983 (0.037)

size $k = 100$ individuals with three different correlation structures and arrived the parameter estimates for each method. The MRME, the average number of zero and nonzero coefficients, the estimated values of the nonzero regression coefficients and the corresponding standard errors over 1000 simulated data sets are summarized in Table 6. When there is an over-dispersion, as we noticed from Table 6, the proposed PGQL approach has smaller MRME compared to PGEEs but the average number of zero coefficients for PGEEs and PGQL are close to each other and the nonzero regression parameter estimates are close to the true values in all cases. This shows,

Table 6 Performance measures for over-dispersion count data with stationary covariates (m = 5)

True model	Method	MRME%	Avg. no. of zero coefficients		Estimates of nonzero coefficients		
			Correct	Incorrect	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
AR(1) $\rho = 0.50$	PGEEs(IND)	95.62	0.89	0.0	0.5009 (0.089)	0.5077 (0.106)	0.5897 (0.118)
	PGEEs(AR(1))	73.07	1.47	0.0	0.5018 (0.088)	0.5121 (0.102)	0.5864 (0.108)
	PGQL	69.04	1.52	0.0	0.5037 (0.086)	0.5110 (0.102)	0.5874 (0.108)
EQC $\rho = 0.50$	PGEEs(IND)	95.91	0.86	0.0	0.4991 (0.096)	0.5059 (0.109)	0.5910 (0.120)
	PGEEs(EQC)	82.80	1.53	0.0	0.5117 (0.092)	0.5075 (0.106)	0.5832 (0.116)
	PGQL	81.79	1.51	0.0	0.5104 (0.091)	0.5082 (0.106)	0.5850 (0.116)
MA(1) $\rho = 0.50$	PGEEs(IND)	85.43	0.95	0.0	0.5004 (0.086)	0.5029 (0.098)	0.5921 (0.114)
	PGEEs(MA(1))	73.78	1.31	0.0	0.5059 (0.086)	0.5071 (0.099)	0.5892 (0.111)
	PGQL	68.37	1.46	0.0	0.5054 (0.086)	0.5058 (0.097)	0.5892 (0.106)

the bias of the non-zero parameter estimates are approximately zero. Again, these simulation studies clearly show that PGQL is performing better compared to PGEEs with over-dispersion data. Note that, in practice, it is very difficult to know the true correlation structure; however, it is reasonable to assume that the correlation structure remains the same for all individuals.

5 Health Care Utilization Data Study

We applied our proposed methodology to a real life data-set on the health care utilization problem, studied by Sutradhar (2003). This data-set was collected by the General Hospital of St. Johns, Newfoundland, Canada. These longitudinal count data contain the complete records for $k = 144$ individuals for 4 years ($m = 4$) from 1985 to 1988. The number of visits to a physician by each individual during a given year was recorded as the response, and this was repeated for 4 years. Also, the information on four covariates: gender, number of chronic conditions, education level, and age were also recorded for each individual. We are also interested in examining whether there are any interaction effects between the parametric covariates, so we included some of these interactions in our model. In

Table 7 Estimates of the regression parameters under PGQL and PGEEs approaches in fitting health care utilization count data

	Penalized estimates			
Variable	PGEEs (AR(1))	PGEEs (EQC)	PGEEs (MA(1))	PGQL
GENDER	0.000	0.000	0.000	0.000
CHRONIC	0.105	0.103	0.104	0.104
EDUCATION	−0.492	−0.432	−0.489	−0.443
AGE	0.033	0.032	0.033	0.033
GENDER*CHRONIC	0.143	0.144	0.144	0.143
GENDER*EDUCATION	0.053	0.000	0.050	0.000
GENDER*AGE	−0.009	−0.009	−0.009	−0.009

view of the background information, it is appropriate to assume that the response variable, marginally, follows the Poisson distribution, and the repeated counts recorded for 4 years will be longitudinally correlated. We are interested in taking the longitudinal correlations into account and examining the effects of the above covariates and their interaction effects on the physician visits. We used PGEEs under different correlation structure for variable selection and compared the results with our proposed method, PGQL. A summary of the results is given in Table 7. From Table 7, we see that all methods identified CHRONIC, EDUCATION, AGE, GENDER*CHRONIC, and GENDER*AGE are the significant variables and the covariate GENDER as unimportant. Under PGEEs method with AR(1) & MA(1) working correlation structure, results indicates that GENDER*EDUCATION also significant where as it is not significant for other two methods. Parameter estimates and identified variables are almost similar for PGEEs (EQC) and PGQL. This clearly indicate that PGEEs based variable selection procedure is sensitive to the choice of covariance structure, leading to different results for different covariance structures. Since in practical situations the true correlation structure is often unknown, the PGQL approach is more appropriate since it can accommodate all three correlation structures in a unique way.

6 Concluding Remarks

We propose a penalized GQL approach for variable selection in longitudinal data analysis where both estimation and variable selection are carried out simultaneously. We used SCAD penalty to achieve oracle properties. Our performance analysis shows that the PGQL approach produces consistent as well as more efficient regression estimates as compared to the independence assumption-based PGEEs approach. The proposed PGQL approach assumes a known longitudinal lag-correlation structure with unknown correlation parameters. When the correlation structure is known the PGQL method has similar performance compared to the

PGEEs approach. However, when the model is mis-specified such as variance function and correlation structure our proposed PGQL approach outperforms the PGEEs method. The main advantage of this PGQL approach is that there is unique way to specify the correlation structure compared to the PGEEs method.

Acknowledgements The authors' are grateful for the opportunity to present their work at the 2015 International Symposium in Statistics (ISS) on Advances in Parametric and Semiparametric Analysis of Multivariate, Time Series, Spatial-temporal, and Familial-longitudinal Data. Special thanks go to Professor Brajendra Sutradhar for organizing the conference, to members of the symposium audience for insightful discussion of our presentation, and to two anonymous referees for their thoughtful comments on our manuscript. The authors' research was partially supported by grants from Natural Sciences & Engineering Research Council of Canada and Canadian Institute of Health Research.

References

Akaike, H: Information theory as a extension of maximum likelihood principle. In: Petrove, B.N., Csaki, F. (eds.) Second Symposium of Information Theory, pp. 267–282. Akademiai Kiado, Budapest (1973)

Akaike, H: A new look at the statistical model identification. IEEE Trans. Autom. Control **19**, 716–723 (1974)

Antoniadis, A.: Wavelets in statistics: a review (with discussion). J. Italian Stat. Assoc. **6**, 97–144 (1997)

Antoniadis, A., Fan, J.: Regularization of wavelets approximations. J. Am. Stat. Assoc. **96**, 939–967 (2001)

Cantoni, E., Flemming, J.M., Ronchetti, E.: Variable selection for marginal longitudinal generalized linear models. Biometrika **61**, 507–514 (2005)

Craven, P., Wahba, G.: Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numer. Math. **31**, 377–403 (1979)

Crowder, M.J.: On use of a working correlation matrix in using generalized linear models for repeated measures. Biometrika **82**, 407–410 (1995)

Donoho, D.L., Johnstone, I.M.: Ideal spatial adaptation by wavelet shrinkage. Biometrika **81**, 425–455 (1994)

Dziak, J.J., Li, R., Qu, A.: An overview on quadratic inference function approaches for longitudinal data. In Frontiers of Statistics, Volume 1: New Developments in Biostatistics and Bioinformatics, J. Fan, J.S. Liu, and X. Lin (eds), Chapter 3, 49–72. 5 Toh Tuch Link, Singapore: World Scientific Publishing, (2009)

Fan, J.: Comments on "Wavelets in Statistics: A Review" by A. Antoniadis. J. Italian Stat. Assoc. **6**, 131–138 (1997)

Fan, J., Li, R.: Variable selection via non concave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**, 1348–1360 (2001)

Fan, J., Li, R.: New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. J. Am. Stat. Assoc. **99**, 710–723 (2004)

Liang, K.Y., Zeger, S.L.: Longitudinal data analysis using generalized linear models. Biometrika **73**, 13–22 (1986)

Lv, J., Fan, Y.: A unified approach to model selection and sparse recovery using regularized least squares. Ann. Stat. **37**, 3498–3528 (2009)

McKenzie, E.: Some ARMA models for dependent sequences of Poisson counts. Adv. Appl. Probab. **20**, 822–835 (1988)

Nadarajah, T.: Penalized empirical likelihood based variable selection. M.Sc. thesis, Memorial University of Newfoundland, St. John's (2011)

Pan, W.: Akaike's information criterion in generalized estimating equations. Biometrics **57**, 120–125 (2001)

Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)

Sutradhar, B.C.: An overview on regression models for discrete longitudinal responses. Stat. Sci. **18**, 377–393 (2003)

Sutradhar, B. C. Dynamic Mixed Models for Familial Longitudinal Data. New York: Springer (2011)

Sutradhar, B.C., Das, K.: On the efficiency of regression estimators in generalized linear models for longitudinal data. Biometrika **86**, 459–465 (1999)

Sutradhar, B.C., Kovacevic, M.: Analysing ordinal longitudinal survey data: generalized estimating equations approach. Biometrika **87**, 837–848 (2000)

Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B **58**, 267–288 (1996)

Variyath, A.M.: Variable selection in generalized linear models by empirical likelihood. Ph.D. thesis, University of Waterloo, Waterloo (2006)

Variyath, A.M., Chen, J., Abraham, B.: Empirical likelihood based variable selection. J. Stat. Plan. Infer. **140**, 971–981 (2010)

Wang, H., Leng, C.: Unified LASSO estimation via least squares approximation. J. Am. Stat. Assoc. **102**, 1039–1048 (2007)

Wang, L., Qu, A.: Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. J. R. Stat. Soc. Ser. B **71**, 177–190 (2009)

Wang, L., Li, H., Huang, J.: Variable selection in nonparametric varying- coefficient models for analysis of repeated measurements. J. Am. Stat. Assoc. **103**, 1556–1569 (2008)

Wang, L., Zhou, J., Qu, A.: Penalized generalized estimating equations for high-dimensional longitudinal data analysis. Biometrics **68**, 353–360 (2012)

Wedderburn, R.W.M.: Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. Biometrika **61**(3), 439–444 (1974)

Xu, P., Wu, P., Wang, Y., Zhu, L.X.: A GEE based shrinkage estimation for the generalized linear model in longitudinal data analysis. Technical report, Department of Mathematics, Hong Kong Baptist University, Hong Kong (2010)

Xue, L., Qu, A., Zhou, J.: Consistent model selection for marginal generalized additive model for correlated data. J. Am. Stat. Assoc. **105**, 1518–1530 (2010)

Xiao, N., Zhang, D., Zhang, H.H.: Variable selection for semiparametric mixed models in longitudinal studies. Biometrics **66**, 79–88 (2009)

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

Zou, H.: The adaptive Lasso and its oracle properties. J. Am. Stat. Assoc. **101**(476), 1418–1429 (2006)