Journal of Statistical Research *https://doi.org/10.47302/jsr.2021550104*
2021, Vol. 55, No. 1, pp. 21-41 *ISSN 0256 - 422 X*

VARIABLE SELECTION FOR LONGITUDINAL SURVEY DATA

LAURA DUMITRESCU

School of Mathematics and Statistics,
Victoria University of Wellington, Wellington 6140, New Zealand.
Email: laura.dumitrescu@vuw.ac.nz

WEI QIAN

Statistics Canada,
100 Tunney's Pasture Driveway, Ottawa, K1A 0T6, Canada.
Email: wei.qian@canada.ca

J. N. K. RAO*

School of Mathematics and Statistics,
Carleton University, Ottawa, K1S 5B6, Canada.
Email: jrao@math.carleton.ca

SUMMARY

In this article we propose a new variable selection method for analyzing data collected from longitudinal sample surveys. The procedure is based on the survey-weighted quadratic inference function, which was recently introduced as an alternative to the survey-weighted generalized estimating function. Under the joint model-design framework, we introduce the penalized survey-weighted quadratic inference estimator and obtain sufficient conditions for the existence, weak consistency, sparsity and asymptotic normality. To illustrate the finite sample performance of the model selection procedure, we include a limited simulation study.

Keywords and phrases: Complex sampling design; longitudinal data; model selection; oracle property; quadratic inference functions; SCAD; super-population model.

AMS Classification: Primary 62D05; secondary 62F12.

1 Introduction

In economics, social and health sciences, longitudinal sample surveys often exhibit complex sampling design features such as unequal selection probabilities, stratification and clustering of individuals. For data collected from some large-scale surveys, or from surveys which have been linked to administrative data files, to explore relationships between the outcome variables and the covariates, special methods are required for variable selection.

* Corresponding author

The problem of variable selection is important since missing important covariates leads to underfitting, inducing estimation bias and poor prediction performance, whereas the inclusion of too many factors in the model generates overfitting, making the model unnecessarily complex and difficult to interpret, as well as producing unstable estimates. Traditional approaches such as stepwise deletion and subset selection are usually used in practice, ignoring the stochastic errors acquired in the variable selection process.

To improve the prediction accuracy and interpretability of regression models with a larger number of covariates, various penalized methods that force some coefficients to zero, have been proposed in the literature. For example, the least absolute shrinkage and selection operator (LASSO) was introduced in Tibshirani (1996), but even though it has several attractive properties, for large regression coefficients, the method introduces a significant bias towards zero. Alternative methods have been considered for bias reduction, such as e.g. the adaptive LASSO (ALASSO) (see Zou, 2006) via a weighted penalty approach, and the smoothly clipped absolute deviation (SCAD) penalty, introduced in Fan and Li (2001). The latter was shown to satisfy desirable theoretical and empirical properties; for a parameter which is close to zero, it preserves the penalization rate of the LASSO, but, as the absolute value of the parameter increases, the rate is continuously relaxed. However, the SCAD penalty is non-convex, which generates numerical challenges in obtaining the solution, and, in practice, additional considerations are needed for the selection of the regularization parameter.

The traditional approach to analyze sample survey data is to make a design-based inference on the finite population parameters with respect to the distribution induced by the probability sampling design. A penalized method for variable selection, based on an empirical likelihood approach to model univariate responses from complex sampling surveys, was recently proposed in Zhao et al. (forthcoming). Nevertheless, to draw conclusions which are valid beyond the reference population, a stochastic model on the population elements is often needed and in such case, the large sample properties of the estimators are obtained within a joint model-design framework, formally established in Rubin-Bleuer and Schiopu Kratina (2005). The corresponding results specify an average behaviour of estimators which would have been obtained from taking potential samples from all possible finite populations.

The topic of variable selection in models for complex longitudinal sample survey data using penalty functions is rather scarce in the literature and a first procedure, based on the survey-weighted generalized estimating equation (GEE), was introduced in Wang et al. (2014). The motivating example was the Canadian National Population Health Survey, where the binary variable of the loss of independence among seniors is modelled using the logistic regression, as a function of eleven other variables from the data set: sex, age, BMI, chronic conditions, smoking status, residence area, education level, income level, living in company, active status and alcohol consumption. To select significant variables and simultaneously estimate coefficients, using the SCAD penalty, the authors used the survey-weighted GEE, introduced by Rao (1998) and further discussed in Roberts et al. (2009) and Carrillo et al. (2010).

Under the semiparametric marginal modeling approach for longitudinal observations, the correlations between measurements taken at different occasions of the survey are unknown and an alternative to the survey-weighted GEE was proposed in Dumitrescu et al. (2021). Its main advantage is that it yields estimators that are more efficient, under misspecification of the working correlation matrix and are as efficient as the GEE counterpart, when the correlation matrix is correctly specified. Moreover, the procedure based on the survey-weighted quadratic inference function (QIF) avoids the additional step of estimating the nuisance correlation parameters and, more importantly, provides an inference function for model diagnostics, as well as for goodness-of-fit tests.

To automatically and simultaneously select variables, we propose the penalized survey-weighted quadratic inference criterion, yielding an estimator with attractive large sample properties. The rates of convergence of the penalized estimator depend on the regularization parameter and, under certain conditions, we show that it retains the oracle property for selecting the correct model: the null components of the estimator are estimated as zero, with probability converging to one, whereas, the nonzero components are estimated as well as the correct submodel is known.

The paper is organized as follows. Our framework and notations are presented in Section 2. In Section 3 we review the survey-weighted GEE and the survey-weighted QIF methods and show how a result in Dumitrescu et al. (2021) can be used to obtain the limiting distribution of the survey-weighted GEE estimator. The penalized survey-weighted QIF is introduced in Section 4 and we obtain sufficient conditions for weak consistency, sparsity, as well as the asymptotic normality of its nonzero components. In Section 5 we illustrate the finite sample performance of the proposed model selection procedure and obtain numerical results on the estimated coefficients.

2 The Model: Assumptions and Notations

Longitudinal data comprise several observations, made at different time points on a set of individuals or units and recorded measurements consist of a sequence of n size m vectors, denoted as $\{y_{i1}, \ldots, y_{im}\}$, $i = 1, \ldots, n$. The usual assumption is that there is a correlation within the measurements from each unit i but observations from different units are independent. Furthermore, corresponding to y_{ij} there is a set of d non-stochastic covariates, denoted as a d-dimensional vector $\mathbf{x}_{ij}$. Hence $\mathbf{y}_1, \ldots, \mathbf{y}_n$ is a sample of independent m dimensional random vectors, defined on a probability space $(\Theta, \mathcal{A}, P_\beta)$, $\boldsymbol{\beta} \in \Omega$ and the objective is to estimate the parameter $\boldsymbol{\beta}$.

There are several estimating approaches which have been used in the literature and one of the most popular assumes a marginal model for the response y_{ij} which depends on the parameter $\boldsymbol{\beta}$ through the value $\theta_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}$,

$$\mathrm{E}_{\boldsymbol{\beta}}(y_{ij}) = \mu(\theta_{ij}) := \mu_{ij}(\boldsymbol{\beta}) \text{ and } \mathrm{Var}_{\boldsymbol{\beta}}(y_{ij}) = \phi\mu'(\theta_{ij}) := \phi\sigma_{ij}^2(\boldsymbol{\beta}),\ \phi \neq 0.$$

Here, μ denotes the (canonical) link function, assumed to be a continuously differentiable function, with $\mu' > 0$ and ϕ is an over-dispersion parameter. In the marginal modelling approach, the true correlation within cluster i is not specified and the GEE method involves the use of a working correlation matrix instead, denoted as $\mathbf{R}_i(\boldsymbol{\alpha})$, which depends on a nuisance parameter $\boldsymbol{\alpha}$. An estimator of $\boldsymbol{\beta}$ is taken to be the solution of the GEE, defined as

$$\mathbf{g}_n(\boldsymbol{\beta}) := \sum_{i=1}^{n} \left[\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \right]^T \mathbf{A}_i(\boldsymbol{\beta})^{-1/2} \mathbf{R}_i(\boldsymbol{\alpha})^{-1} \mathbf{A}_i(\boldsymbol{\beta})^{-1/2} \big[\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}) \big] = \mathbf{0}, \qquad (2.1)$$

where $\boldsymbol{\mu}_i(\boldsymbol{\beta}) = (\mu_{i1}(\boldsymbol{\beta}), \ldots, \mu_{im}(\boldsymbol{\beta}))^T$ and $\mathbf{A}_i(\boldsymbol{\beta}) = \mathrm{diag}\{\phi\sigma_{i1}^2(\boldsymbol{\beta}), \ldots, \phi\sigma_{im}^2(\boldsymbol{\beta})\}$. Assuming that there exists a "true" value of the regression parameter, denoted as $\boldsymbol{\beta}_0$, which is an interior point of Ω, the following identifiability assumption is imposed

$$\mathrm{E}_{\boldsymbol{\beta}}\big[\mathbf{g}_n(\boldsymbol{\beta})\big] = \mathbf{0} \text{ if and only if } \boldsymbol{\beta} = \boldsymbol{\beta}_0. \tag{2.2}$$

When the sample is obtained through a complex sampling survey from a finite population U of size N, using for example stratification or unequal cluster selection probabilities, the inference needs to account for the sampling features. Namely, from the set of labels $U = \{1, \ldots, N\}$, a subset $s \subset U$ of indices is selected according to a probability sampling design, P_π such that $P(i \in s) = \pi_i$, $1 = 1, \ldots, N$ denotes the first-order inclusion probability. Then, to draw conclusions which are valid beyond the reference population, we assume the stochastic model (2.1) on the population elements in $\mathcal{F} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$. This leads to a joint model-design inference, where the large sample properties of the estimator are obtained according to the probabilities induced by the model and survey design. The results then specify an average behaviour of the estimator, which would have been obtained from taking potential samples from all possible finite populations. Within the framework of a product probability space of the super-population and the design space (as introduced in Rubin-Bleuer and Schiopu Kratina, 2005), an important assumption is that, given the design variables, the sample selection and the model characteristics are independent.

In what follows, the following matrix notations will be used. For a $d \times 1$ vector, $\boldsymbol{\lambda}$, we use the notation $\|\boldsymbol{\lambda}\|$ for its Euclidean norm, whereas if $\mathbf{A}$ is a $d \times d$ matrix, then $\|\mathbf{A}\| = \sup_{\|\boldsymbol{\lambda}\|=1} \|\mathbf{A}\boldsymbol{\lambda}\|$ is used for its operator norm. If $\mathbf{A}$ is symmetric, we denote by $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ its minimum and maximum eigenvalues, respectively. In addition, for any matrix $\mathbf{A}$, we have $\|\mathbf{A}\| = [\lambda_{\max}(\mathbf{A}^T\mathbf{A})]^{1/2}$.

3 The Survey-weighted Quadratic Inference Function

A standard assumption in the literature of longitudinal data is that the within-individual correlations are equal and typical choices include

(a) $\mathbf{R}(\alpha) = \mathbf{I}_m$, where $\mathbf{I}_m$ denotes the identity matrix of order m,

(b) $\mathbf{R}(\alpha) = \{\rho_{lr}(\alpha)\}_{1 \le l,r \le m}$, where $\rho_{lr}(\alpha) = \alpha,\ l \neq r,\ 0 \le \alpha \le 1$, $\rho_{ll}(\alpha) = 1$,

(c) $\mathbf{R}(\alpha) = \{\rho_{lr}(\alpha)\}_{1 \le l,r \le m}$, where $\rho_{lr}(\alpha) = \alpha^{|l-r|}$, $0 \le \alpha \le 1$,

(d) $\mathbf{R}(\alpha) = \{\rho_{lr}(\alpha)\}_{1 \le l,r \le m}$, where $0 \le \rho_{lr}(\alpha) \le 1$, $l, r = 1, \ldots, m$.

The above forms correspond to (a) *independence*, (b) *exchangeable*, (c) *first-order autoregressive* and (d) *unspecified* correlation structure, respectively.

Remark 1. As noted in Qu et al. (2000), the inverses of the matrices in (a) - (c) can be written as a linear combinations of a small number of simple basis matrices:

(a) $\mathbf{M}_1 = \mathbf{I}_m$,

(b) $\mathbf{M}_1 = \mathbf{I}_m$ and $\mathbf{M}_2 = \{\gamma_{lr}\}_{1 \le l,r \le m}$, $\gamma_{lr} = 1$, for $1 \le l \neq r \le m$ and $\gamma_{ll} = 0, l = 1, \ldots, m$,

(c) $\mathbf{M}_1 = \mathbf{I}_m$, $\mathbf{M}_2 = \{\gamma_{lr}\}_{1 \le l,r \le m}$, $\gamma_{l,l-1} = 1$, $l = 2, \ldots, m$ and $\gamma_{l,r} = 0$, for $r \neq l-1$, $1 \le l, r \le m$ and $\mathbf{M}_3 = \{\gamma_{lr}\}_{1 \le l,r \le m}$, with $\gamma_{11} = \gamma_{mm} = 1$ and 0 elsewhere.

Hence, we assume that the working covariance matrix $\mathbf{R}^{-1}(\boldsymbol{\alpha})$ can be written as

$$\mathbf{R}^{-1}(\boldsymbol{\alpha}) = \sum_{l=1}^{L} c_l(\boldsymbol{\alpha})\mathbf{M}_l, \tag{3.1}$$

with given $\mathbf{M}_1, \ldots, \mathbf{M}_L$. The "census extended quasi-score" vector $\bar{\mathbf{q}}_N(\boldsymbol{\beta})$ is defined as

$$\bar{\mathbf{q}}_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{q}_i(\boldsymbol{\beta}) = \frac{1}{N}\begin{pmatrix} \sum_{i=1}^{N}\left[\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right]^T \mathbf{A}_i(\boldsymbol{\beta})^{-1/2}\mathbf{M}_1\mathbf{A}_i(\boldsymbol{\beta})^{-1/2}\big[\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\big] \\ \vdots \\ \sum_{i=1}^{N}\left[\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right]^T \mathbf{A}_i(\boldsymbol{\beta})^{-1/2}\mathbf{M}_L\mathbf{A}_i(\boldsymbol{\beta})^{-1/2}\big[\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\big] \end{pmatrix}.$$

and the survey-weighted extended quasi-score is obtained using the survey design weights $w_i = \pi_i^{-1}, i \in s$

$$\mathbf{q}_n(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i\in s} w_i \mathbf{q}_i(\boldsymbol{\beta}).$$

If the census extended quasi-score satisfies a central limit theorem (CLT) under the model probability (assumption (N_1)) and a CLT holds for the survey-weighted QIF, under the design probability (assumption (N_2)), the limiting distribution of the latter was obtained in Theorem 2 of Dumitrescu et al. (2021), under the joint model-design probability. The notation $\xrightarrow{\mathcal{L}}$ is used for the convergence in distribution of random variables.

Theorem 1. *Assume that the following conditions hold.*

(N_0) $n/N \longrightarrow f,$ *as* $n \to \infty$, *with* $0 \le f < 1$.

(N_1) $N^{1/2}\bar{\mathbf{q}}_N(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\beta_0}),$ *as* $N \to \infty,$ *under* $P_{\beta_0},$ *where* $\boldsymbol{\Sigma}_{\beta_0} > 0$.

(N_2) *Given the sequence of finite populations,* $\mathcal{F}_N = (\mathbf{y}_{1N}, \ldots, \mathbf{y}_{NN})$, *we have*

$$n^{1/2}\big[\mathbf{q}_n(\boldsymbol{\beta}_0) - \bar{\mathbf{q}}_N(\boldsymbol{\beta}_0)\big] \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_d), \text{ as } n \to \infty, \text{ under } P_\pi,$$

where $\boldsymbol{\Sigma}_d > 0$ *is non-stochastic*.

Then, as $n \to \infty$,

$$(Q) \;\; n^{1/2}\mathbf{q}_n(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_0), \text{ under } P_{\beta_0,\pi}, \text{ where } \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_d + f\boldsymbol{\Sigma}_{\boldsymbol{\beta}_0}. \tag{3.2}$$

The statement of Theorem 1 is very general and it can be used to obtain the limiting distribution of a sequence of estimators $\bar{\boldsymbol{\beta}}_n$, obtained as a solution of the survey-weighted GEE, $\mathbf{g}_n^{WGEE}(\boldsymbol{\beta}) = \mathbf{0}$. The pseudo-GEE estimating function

$$\mathbf{g}_n^{WGEE}(\boldsymbol{\beta}) := \sum_{i=1}^{n} w_i \left[\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right]^T \mathbf{A}_i(\boldsymbol{\beta})^{-1/2}\mathbf{R}_i(\boldsymbol{\alpha})^{-1}\mathbf{A}_i(\boldsymbol{\beta})^{-1/2}\big[\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\big], \tag{3.3}$$

was first proposed in Rao (1998) and Roberts et al. (2009) to analyze longitudinal survey data.

Sufficient conditions for the pseudo-GEE estimator $\bar{\boldsymbol{\beta}}_n$ to be weakly consistent, under the joint model and design probability, were given in Carrillo et al. (2010). Assuming that (Q) holds for $\mathbf{g}_n^{WGEE}$, we next show a CLT result for this estimator. Denote $\mathcal{B}_n = \{\boldsymbol{\beta} \in \Omega : \ n^{1/2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq r\},$ with $r > 0$ and let

$$\nabla \mathbf{g}_n^{WGEE}(\boldsymbol{\beta}) = -\frac{\partial \mathbf{g}_n^{WGEE}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T},$$

$$\bar{\mathbf{g}}_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right]^T \mathbf{A}_i(\boldsymbol{\beta})^{-1/2}\mathbf{R}_i(\boldsymbol{\alpha})^{-1}\mathbf{A}_i(\boldsymbol{\beta})^{-1/2}\big[\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\big].$$

Proposition 3.1. *Assume that (N_0) holds, together with the following conditions:*

(G_1) $N^{1/2}\bar{\mathbf{g}}_N(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^g_{\beta_0}),$ *as $N \to \infty$, under P_{β_0}, where $\boldsymbol{\Sigma}^g_{\beta_0} > 0$.*

(G_2) *Given the sequence of finite populations, $\mathcal{F}_N = (\mathbf{y}_{1N}, \ldots, \mathbf{y}_{NN})$, we have*

$$n^{1/2}\big[\mathbf{g}_n^{WGEE}(\boldsymbol{\beta}_0) - \bar{\mathbf{g}}_N(\boldsymbol{\beta}_0)\big] \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^g_d), \text{ as } n \to \infty, \text{ under } P_\pi, \text{ where } \boldsymbol{\Sigma}^g_d > 0 \text{ is non-stochastic}.$$

(G_3) *there exists an invertible non-stochastic $d \times d$ matrix $\mathbf{D}_0$, such that, as $n \to \infty$, we have*

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}_n}\left\|\nabla \mathbf{g}_n^{WGEE}(\boldsymbol{\beta}) - \mathbf{D}_0\right\| \stackrel{P_{\boldsymbol{\beta}_0,\pi}}{\longrightarrow} 0.$$

Then, as $n \to \infty$,

$$n^{1/2}(\bar{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \mathbf{D}_0^{-1}\boldsymbol{\Sigma}^g_0\mathbf{D}_0^{-1}), \text{ under } P_{\boldsymbol{\beta}_0,\pi}, \text{ where } \boldsymbol{\Sigma}^g_0 = \boldsymbol{\Sigma}^g_d + f\boldsymbol{\Sigma}^g_{\beta_0}. \tag{3.4}$$

Proof. The Mean Value theorem applied to $\mathbf{g}_n^{WGEE}$, on the set $\{\mathbf{g}_n^{WGEE}(\bar{\boldsymbol{\beta}}_n) = 0, \bar{\boldsymbol{\beta}}_n \in \mathcal{B}_n\}$ gives

$$\mathbf{g}_n^{WGEE}(\bar{\boldsymbol{\beta}}_n) = \mathbf{g}_n^{WGEE}(\boldsymbol{\beta}_0) - \nabla \mathbf{g}_n^{WGEE}(\boldsymbol{\beta}_n^*)(\bar{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0),$$

where $\boldsymbol{\beta}_n^* \in \mathcal{B}_n$ and we write

$$\begin{aligned} n^{1/2}\mathbf{g}_n^{WGEE}(\boldsymbol{\beta}_0) = {} & n^{1/2}\mathbf{D}_0^{1/2}\Big[\mathbf{D}_0^{-1/2}\nabla \mathbf{g}_n^{WGEE}(\boldsymbol{\beta}_n^*)\mathbf{D}_0^{-1/2} - \mathbf{I}_d\Big]\mathbf{D}_0^{1/2}(\bar{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \\ & + n^{1/2}\mathbf{D}_0(\bar{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0). \end{aligned} \tag{3.5}$$

By (G_3), the first term in (3.5) is $o_{P_{\boldsymbol{\beta}_0\pi}}(1)$ so that the asymptotic distribution of $n^{1/2}(\bar{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ is equal to the asymptotic distribution of $n^{1/2}\mathbf{D}_0^{-1}\mathbf{g}_n^{WGEE}(\boldsymbol{\beta}_0)$ and an application of Theorem 1 concludes the proof. □

We now turn to the extended quasi-score vector. Let

$$\mathbf{C}_n(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i \in s} w_i \mathbf{q}_i(\boldsymbol{\beta})\mathbf{q}_i(\boldsymbol{\beta})^T,$$

and assume that it is $P_{\beta_0,\pi}$- a.s. invertible and that there exists a constant $K > 0$ such that for any n, we have $\inf_{\boldsymbol{\beta}\in\Omega} \lambda_{\min}[\mathbf{C}_n(\boldsymbol{\beta})] > K,\ P_{\beta_0,\pi}$- a.s.

Furthermore, suppose that the link function μ is three times continuously differentiable on a neighbourhood of $\boldsymbol{\beta}_0, \mathcal{U}_\delta = \{\boldsymbol{\beta} \in \Omega;\ \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < \delta\}$, with $\delta > 0$ so that $\mathbf{q}_n$ is twice continuously differentiable on $\mathcal{U}_\delta,\ P_{\boldsymbol{\beta}_0,\pi}$- a.s. Let $\mathcal{D}_n(\boldsymbol{\beta}) = \dfrac{\partial \mathbf{q}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}$, whose k-th column is given by $\mathbf{d}_n^{(k)}(\boldsymbol{\beta})$ and, for each $k = 1, \ldots, d$, we denote $\mathbf{G}_n^{(k)}(\boldsymbol{\beta}) = \dfrac{\partial \mathbf{C}_n(\boldsymbol{\beta})}{\partial \beta_k}$, with uniformly continuous entries on $\mathcal{U}_\delta$.

The survey-weighted quadratic inference function is defined as

$$Q_n(\boldsymbol{\beta}) = n\mathbf{q}_n(\boldsymbol{\beta})^T \mathbf{C}_n(\boldsymbol{\beta})^{-1} \mathbf{q}_n(\boldsymbol{\beta})$$

and its first and second order partial derivatives, with $1 \leq k, l \leq d$, satisfy

$$\frac{\partial n^{-1} Q_n(\boldsymbol{\beta})}{\partial \beta_k} = 2\mathbf{d}_n^{(k)}(\boldsymbol{\beta})^T \mathbf{C}_n(\boldsymbol{\beta})^{-1} \mathbf{q}_n(\boldsymbol{\beta}) - \mathbf{q}_n(\boldsymbol{\beta})^T \mathbf{C}_n(\boldsymbol{\beta})^{-1} \mathbf{G}_n^{(k)}(\boldsymbol{\beta}) \mathbf{C}_n(\boldsymbol{\beta})^{-1} \mathbf{q}_n(\boldsymbol{\beta}),$$
$$\frac{\partial^2 n^{-1} Q_n(\boldsymbol{\beta})}{\partial \beta_k \partial \beta_l} = 2\mathbf{d}_n^{(k)}(\boldsymbol{\beta})^T \mathbf{C}_n(\boldsymbol{\beta})^{-1} \mathbf{d}_n^{(l)}(\boldsymbol{\beta}) + r_n^{(k,l)}(\boldsymbol{\beta}),$$

where

$$\begin{aligned} r_n^{(k,l)}(\boldsymbol{\beta}) = 2\left[\frac{\partial \mathbf{d}_n^{(k)}(\boldsymbol{\beta})}{\partial \beta_l}\right]^T & \mathbf{C}_n(\boldsymbol{\beta})^{-1} \mathbf{q}_n(\boldsymbol{\beta}) - 2\mathbf{d}_n^{(k)}(\boldsymbol{\beta})^T \mathbf{C}_n(\boldsymbol{\beta})^{-1} \mathbf{G}_n^{(l)}(\boldsymbol{\beta}) \mathbf{C}_n(\boldsymbol{\beta})^{-1} \mathbf{q}_n(\boldsymbol{\beta}) \\ & - 2\mathbf{d}_n^{(l)}(\boldsymbol{\beta}) \mathbf{C}_n(\boldsymbol{\beta})^{-1} \mathbf{G}_n^{(k)}(\boldsymbol{\beta}) \mathbf{C}_n(\boldsymbol{\beta})^{-1} \mathbf{q}_n(\boldsymbol{\beta}) \\ & + 2\mathbf{q}_n(\boldsymbol{\beta}) \mathbf{C}_n(\boldsymbol{\beta})^{-1} \mathbf{G}_n^{(l)}(\boldsymbol{\beta}) \mathbf{C}_n(\boldsymbol{\beta})^{-1} \mathbf{G}_n^{(k)}(\boldsymbol{\beta}) \mathbf{C}_n(\boldsymbol{\beta})^{-1} \mathbf{q}_n(\boldsymbol{\beta}) \\ & - \mathbf{q}_n(\boldsymbol{\beta})^T \mathbf{C}_n(\boldsymbol{\beta})^{-1} \frac{\partial \mathbf{G}_n^{(k)}(\boldsymbol{\beta})}{\partial \beta_l} \mathbf{C}_n(\boldsymbol{\beta})^{-1} \mathbf{q}_n(\boldsymbol{\beta}). \end{aligned}$$

In Dumitrescu et al. (2021), the following assumptions were used to obtain the limiting distribution of the survey-weighted QIF estimator

(S_1) there exists a non-stochastic and positive definite $Ld \times Ld$ matrix, $\mathcal{W}_0(\boldsymbol{\beta})$ such that $\sup_{\boldsymbol{\beta}\in\mathcal{U}_\delta} \|\mathbf{C}_n(\boldsymbol{\beta})^{-1} - \mathcal{W}_0(\boldsymbol{\beta})\| \overset{P_{\beta_0,\pi}}{\longrightarrow} 0$, as $n \to \infty$;

(S_2) there exists a non-stochastic matrix $\mathcal{D}_0(\boldsymbol{\beta})$, of size $Ld \times d$, such that $\sup_{\boldsymbol{\beta}\in\mathcal{U}_\delta} \|\mathcal{D}_n(\boldsymbol{\beta}) - \mathcal{D}_0(\boldsymbol{\beta})\| \overset{P_{\boldsymbol{\beta}_0,\pi}}{\longrightarrow} 0$, as $n \to \infty$;

(S_3) the matrix $\mathcal{J}_0(\boldsymbol{\beta}) := \mathcal{D}_0(\boldsymbol{\beta})^T \mathcal{W}_0(\boldsymbol{\beta}) \mathcal{D}_0(\boldsymbol{\beta})$ is non-singular on $\mathcal{U}_\delta$.

By Theorem 1 we have

$$\sqrt{n}\mathbf{q}_n(\boldsymbol{\beta}_0) = O_{P_{\boldsymbol{\beta}_0,\pi}}(1)$$

and, due to (S_1) and (S_2), for every $k = 1, \ldots, d$, we obtain

$$\frac{\partial n^{-1/2} Q_n(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} = n^{1/2} \mathcal{D}_n(\boldsymbol{\beta}_0)^T \mathbf{C}_n(\boldsymbol{\beta}_0)^{-1} \mathbf{q}_n(\boldsymbol{\beta}) + \mathcal{R}_n^1(\boldsymbol{\beta}_0), \ \|\mathcal{R}_n^1(\boldsymbol{\beta}_0)\| = o_{P_{\boldsymbol{\beta}_0 \pi}}(1). \tag{3.6}$$

In addition, since $r_n^{(k,l)}(\boldsymbol{\beta}_0) = o_{P_{\boldsymbol{\beta}_0,\pi}}(1)$, for any $1 \le k, l \le d$ we have

$$\frac{\partial^2 n^{-1} Q_n(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = 2\mathcal{D}_n(\boldsymbol{\beta}_0)^T \mathbf{C}_n(\boldsymbol{\beta}_0)^{-1} \mathcal{D}_n(\boldsymbol{\beta}_0) + \mathcal{R}_n^2(\boldsymbol{\beta}_0), \text{ with } \|\mathcal{R}_n^2(\boldsymbol{\beta}_0)\| = o_{P_{\boldsymbol{\beta}_0 \pi}}(1), \tag{3.7}$$

which implies the element-wise convergence in $P_{\boldsymbol{\beta}_0 \pi}$ of $\dfrac{\partial^2 n^{-1} Q_n(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}$ to $2\mathcal{J}_0(\boldsymbol{\beta}_0)$, as $n \to \infty$.

Definition 3.1. The pseudo-QIF estimator is defined as

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta} \in \Omega} Q_n(\boldsymbol{\beta}). \tag{3.8}$$

Under the joint randomization framework, the expected value of the survey-weighted extended quasi-score is equal to

$$\mathrm{E}_{\boldsymbol{\beta}_0,\pi}\big[\mathbf{q}_n(\boldsymbol{\beta})\big] = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\Delta}_i(\boldsymbol{\beta}),$$

where

$$\boldsymbol{\Delta}_i(\boldsymbol{\beta}) = \mathrm{E}_{\boldsymbol{\beta}_0}\big[\mathbf{q}_i(\boldsymbol{\beta})\big] = \begin{pmatrix} \left[\dfrac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right]^T \mathbf{A}_i(\boldsymbol{\beta})^{-1/2} \mathbf{M}_1 \mathbf{A}_i(\boldsymbol{\beta})^{-1/2} \big[\boldsymbol{\mu}_i(\boldsymbol{\beta}_0) - \boldsymbol{\mu}_i(\boldsymbol{\beta})\big] \\ \vdots \\ \left[\dfrac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right]^T \mathbf{A}_i(\boldsymbol{\beta})^{-1/2} \mathbf{M}_L \mathbf{A}_i(\boldsymbol{\beta})^{-1/2} \big[\boldsymbol{\mu}_i(\boldsymbol{\beta}_0) - \boldsymbol{\mu}_i(\boldsymbol{\beta})\big] \end{pmatrix}.$$

In Dumitrescu et al. (2021), the weak consistency of the pseudo-QIF estimator, with respect to the joint model-design probability, was obtained (see their Theorem 1). Assumptions (A_1) and (A_2) guarantee that the objective function approaches, uniformly, a census-type function, which, due (A_3), has a unique minimum at the true value of the parameter.

Theorem 2. *Assume that the following conditions are satisfied.*

(A_1) $\sup_{\boldsymbol{\beta} \in \Omega} \|\mathbf{C}_n(\boldsymbol{\beta})^{-1} - \mathbf{W}_N(\boldsymbol{\beta})\| \overset{P_{\beta_0,\pi}}{\longrightarrow} 0$, *as* $n \to \infty$, *for some positive definite* $Ld \times Ld$ *matrix* $\mathbf{W}_N(\boldsymbol{\beta})$.

(A_2) $\sup_{\boldsymbol{\beta} \in \Omega} \left\| \mathbf{q}_n(\boldsymbol{\beta}) - \dfrac{1}{N} \sum_{i=1}^{N} \boldsymbol{\Delta}_i(\boldsymbol{\beta}) \right\| \overset{P_{\boldsymbol{\beta}_0,\pi}}{\longrightarrow} 0$, *as* $n \to \infty$.

(A_3) *For every* $N \ge 1$, *the equation* $\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\Delta}_i(\boldsymbol{\beta}) = \mathbf{0}$ *has a unique solution,* $\boldsymbol{\beta}_0$.

Then, as $n \to \infty$, we have

$$(C) \;\; \hat{\boldsymbol{\beta}}_n \xrightarrow{P_{\beta_0,\pi}} \boldsymbol{\beta}_0.$$

The limiting distribution of the pseudo-QIF estimator follows from the convergence of the survey-weighted extended quasi-score (as in Theorem 3.2), by showing that this estimator is asymptotically linear, as defined in e.g. (3.3) of Newey and McFadden (1986). When terms of the functions of $\boldsymbol{\beta}$ are evaluated at $\boldsymbol{\beta}_0$, we suppress $\boldsymbol{\beta}_0$ and denote $\mathcal{D}_0 = \mathcal{D}_0(\boldsymbol{\beta}_0)$, $\mathcal{J}_0 = \mathcal{J}_0(\boldsymbol{\beta}_0)$ and $\mathcal{W}_0 = \mathcal{W}_0(\boldsymbol{\beta}_0)$. The next result was shown in Theorem 3 of Dumitrescu et al. (2021).

Theorem 3. *Assume that (C) and (Q) are satisfied, as well as (S_1), (S_2) and (S_3). Then, as $n \to \infty$,*

$$(U) \;\; \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(\mathbf{0}, \mathcal{J}_0^{-1}\big[\mathcal{D}_0^T \mathcal{W}_0 \boldsymbol{\Sigma}_0 \mathcal{W}_0 \mathcal{D}_0\big]\mathcal{J}_0^{-1}\Big), \text{ under } P_{\boldsymbol{\beta}_0,\pi}. \quad (3.9)$$

In addition to yielding an estimator which is at least as efficient as the one obtained from the survey-weighted GEE, the survey-weighted QIF can be used to construct a pseudolikelihood ratio type statistics for testing composite hypotheses on model parameters, and a statistic for testing the goodness-of-fit of the marginal model. Their limiting distributions are weighted sums of independent chi-squared random variables, each with one degree of freedom.

4 The Penalized Survey-weighted Quadratic Inference Function

Based on the survey-weighted QIF, we introduce a new approach to variable selection for longitudinal survey data which can incorporate the within cluster correlation, as well as the survey design features, through penalization:

$$Q_n^P(\boldsymbol{\beta}) = Q_n(\boldsymbol{\beta}) + n\sum_{k=1}^{d} p_\lambda(|\beta_k|),$$

where p_λ is a penalty function which depends on a regularization parameter λ. The main advantage of penalized methods versus other procedures, such as stepwise deletion and subset selection, is that they can automatically and simultaneously select variables, hence, avoiding the corresponding stochastic errors. There are several functions which have been used as penalties, such as the L^2 penalty leading to the ridge regression, or the L^1 penalty, used in LASSO. Due to its properties, in our simulations, we use the SCAD, introduced in Fan and Li (2001), which is a nonconcave penalty function on $(0,\infty)$, defined as a quadratic spline function with knots at λ and $a\lambda$ (for some constant $a > 0$)

$$p_\lambda(|\theta|) = \begin{cases} \lambda|\theta|, & |\theta| \le \lambda, \\ -\frac{|\theta|^2 - 2a\lambda|\theta| + \lambda^2}{2(a-1)}, & \lambda < |\theta| \le a\lambda, \\ \frac{(a+1)\lambda^2}{2}, & |\theta| > a\lambda. \end{cases} \quad (4.1)$$

This penalty is singular at origin (yielding sparse solutions), bounded (so that, for large coefficients, the resulting estimators are nearly unbiased) and continuous (leading to a stable model selection procedure). The function is continuously differentiable, with

$$p'_\lambda(|\theta|) = \lambda I(|\theta| \le \lambda) + \frac{a\lambda - |\theta|}{a-1} I(\lambda < |\theta| \le a\lambda),\ a > 2.$$

In Fan and Li (2001), the value $a = 3.7$ was shown to give good practical performance for various selection problems and it was shown that results were similar to those obtained by using the generalized cross-validation method.

Definition 4.1. The penalized pseudo-QIF estimator is defined as

$$\tilde{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}\in\Omega} Q_n^P(\boldsymbol{\beta}).$$

For any d-dimensional vector $\boldsymbol{\beta}$, we consider the partition $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$ into subvectors of size d_1 and $d - d_1$, respectively, and use the corresponding notation

$$\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T,$$

assuming, without loss of generality that $\boldsymbol{\beta}_{20} = \mathbf{0}$.

In our simulations, to calculate the penalized pseudo-QIF estimator, we use the iterative method, based on a local quadratic approximation, as proposed in Fan and Li (2001). The method approximates the nonconvex SCAD penalty term by

$$p_\lambda\big(|\beta_k^{(t)}|\big) + \frac{1}{2}\frac{p'_\lambda\big(|\beta_k^{(t)}|\big)}{|\beta_k^{(t)}|}\big[(\beta_k^{(t)})^2 - \beta_k^2\big], \quad \beta_k^{(t)} \neq 0,$$

where $\boldsymbol{\beta}^{(t)} = (\beta_1^{(t)}, \dots, \beta_d^{(t)})^T$ is the estimator obtained at step t. If $\beta_k^{(t)}$ is such that $|\beta_k^{(t)}| < 0.001$, we set $\beta_k^{(t+1)} = 0$ and write $\boldsymbol{\beta}^{(t)} = (\boldsymbol{\beta}_1^{(t)}, \boldsymbol{\beta}_2^{(t)})^T$, where $\beta_k^{(t)} \neq 0$, for $k = 1, \dots, d_1$ and $\beta_k^{(t)} = 0$, for $k = d_1 + 1, \dots, d$. The algorithm is initialized with the value of the pseudo-QIF estimator $\hat{\boldsymbol{\beta}}_n$ and, based on a previous value $\boldsymbol{\beta}^{(t)}$, the objective function is approximated by

$$Q_n(\boldsymbol{\beta}^{(t)}) + \left[\frac{\partial Q_n(\boldsymbol{\beta}^{(t)})}{\partial \boldsymbol{\beta}_1}\right]^T (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^{(t)}) + \frac{1}{2}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^{(t)})^T \left[\frac{\partial^2 Q_n(\boldsymbol{\beta}^{(t)})}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T}\right]^T (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^{(t)}) + \frac{1}{2} n \boldsymbol{\beta}_1^T \boldsymbol{\Gamma}(\boldsymbol{\beta}^{(t)})\boldsymbol{\beta}_1,$$

where $\boldsymbol{\beta}_1$ is a vector with d_1 non-zero entries and $\boldsymbol{\Gamma}(\boldsymbol{\beta}^{(t)}) = \mathrm{diag}\left\{\frac{p'_\lambda(|\beta_1^{(t)}|)}{|\beta_1^{(t)}|}, \dots, \frac{p'_\lambda(|\beta_{d_1}^{(t)}|)}{|\beta_{d_1}^{(t)}|}\right\}$. Newton-Raphson algorithm gives the minimizer as

$$\boldsymbol{\beta}_1^{(t+1)} = \boldsymbol{\beta}_1^{(t)} - \left[\frac{\partial^2 Q_n(\boldsymbol{\beta}^{(t)})}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T} + n\boldsymbol{\Gamma}(\boldsymbol{\beta}^{(t)})\right]^{-1} \left[\frac{\partial Q_n(\boldsymbol{\beta}^{(t)})}{\partial \boldsymbol{\beta}_1} + n\boldsymbol{\Gamma}(\boldsymbol{\beta}^{(t)})\boldsymbol{\beta}_1^{(t)}\right]. \qquad (4.2)$$

The performance of the variable selection procedure depends on the choice of the tuning parameter and, following Wang and Qu (2009) and Cho and Qu (2013), in our simulations, we use a Bayesian information criterion, based on the pseudo-QIF and take λ_n to be the minimizer of

$$WBIC(\lambda) = \big[Q_n(\tilde{\boldsymbol{\beta}}_\lambda) + \log(n)\mathrm{df}(\tilde{\boldsymbol{\beta}}_\lambda)\big].$$

Here, $\tilde{\boldsymbol{\beta}}_\lambda$ denotes the penalized pseudo-QIF estimator and $\mathrm{df}(\tilde{\boldsymbol{\beta}}_\lambda)$ corresponds to its number of non-zero entries. It can be shown that, with probability tending to 1, this criterion selects the tuning parameter that identifies the true model (we refer to Qian, 2018, for a proof).

4.1 Asymptotic properties

In this section we investigate the asymptotic properties of the penalized pseudo-QIF estimator $\tilde{\boldsymbol{\beta}}_n$. Results are formulated for a general penalty function, whose derivative is continuous and we use the techniques in Fan and Li (2001) to obtain the desired properties. The first theorem shows that, under assumption (P_1), with probability converging to 1, the penalized pseudo-QIF estimator $\tilde{\boldsymbol{\beta}}_n$ exists within a ball $\mathcal{B}_n(r) = \{\boldsymbol{\beta} \in \Omega;\ n^{1/2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq r\}$, $r > 0$ and hence, it is $\sqrt{n}-$ weakly consistent. Assumption (P_2) ensures that the penalty function does not have much more influence than the pseudo-QIF function on the penalized estimator.

Theorem 4 (Existence and weak consistency)**.** *Assume that* (Q), (S_1), (S_2), (S_3) *are satisfied, together with the following conditions.*

(P_1) $n^{1/2}\max\big\{p'_{\lambda_n}(|\beta_{0k}|), k = 1, \dots d_1\big\} = O(1)$.

(P_2) $\max\big\{|p''_{\lambda_n}(|\beta_{0k}|)|, k = 1, \dots d_1\big\} = o(1)$.

Then, there exists a sequence $\tilde{\boldsymbol{\beta}}_n$ *of random variables satisfying*

(a) $P_{\boldsymbol{\beta}_0,\pi}\big(\tilde{\boldsymbol{\beta}}_n$ *is a local minimizer of* $Q_n^P(\boldsymbol{\beta})$ *on* $\mathcal{B}_n(r)\big) \overset{n\to\infty}{\longrightarrow} 1$ *and*

(b) $\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_{P_{\boldsymbol{\beta}_0,\pi}}(n^{-1/2})$.

Proof. (a) We show that, for any $\varepsilon > 0$, there exist $r > 0$ and $n_{\varepsilon,r}$ such that the event $E_n = \big\{Q_n^P(\boldsymbol{\beta}_0) < \inf_{\boldsymbol{\beta}\in\partial\mathcal{B}_n(r)} Q_n^P(\boldsymbol{\beta})\big\}$ has the property

$$P_{\boldsymbol{\beta}_0,\pi}(E_n) \geq 1 - \varepsilon, \text{ for all } n \geq n_{\varepsilon,r}, \tag{4.3}$$

where $\partial\mathcal{B}_n(r) = \{\boldsymbol{\beta} \in \Omega;\ n^{1/2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = r\}$. This implies that, with probability at least $1 - \varepsilon$, there is a minimum in the ball $\mathcal{B}_n(r)$, i.e. there exists a local minimizer $\tilde{\boldsymbol{\beta}}_n$ such that $\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_{P_{\boldsymbol{\beta}_0,\pi}}(n^{-1/2})$.

Let $\boldsymbol{\beta} \in \partial\mathcal{B}_n(r)$ be arbitrarily fixed. By Taylor's expansion of $Q_n(\boldsymbol{\beta})$, using $p_{\lambda_n}(0) = 0$

$$\begin{aligned} Q_n^P(\boldsymbol{\beta}) - Q_n^P(\boldsymbol{\beta}_0) &= Q_n(\boldsymbol{\beta}) - Q_n(\boldsymbol{\beta}_0) + n\sum_{k=1}^{d}\Big[p_{\lambda_n}\big(|\beta_k|\big) - p_{\lambda_n}\big(|\beta_{0k}|\big)\Big] \\ &\geq \left[\frac{\partial Q_n(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right]^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\frac{\partial^2 Q_n(\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) \\ &+ n\sum_{k=1}^{d_1}\left[p'_{\lambda_n}\big(|\beta_{0k}|\big)\frac{|\beta_{0k}|}{\beta_{0k}}\big(\beta_k-\beta_{0k}\big) + \frac{1}{2}p''_{\lambda_n}\big(|\beta_k^{**}|\big)\big(\beta_k-\beta_{0k}\big)^2\right] \\ &:= T_1 + T_2 + T_3, \end{aligned} \tag{4.4}$$

where $\boldsymbol{\beta}^*$ and $\{\beta_k^{**}\}_{1\leq k\leq d_1}$ are such that $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\| \leq \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| = n^{-1/2}r$ and $|\beta_k^{**}-\beta_{0k}| \leq |\beta_k - \beta_{0k}| \leq n^{-1/2}r$, for every $k = 1,\dots,d_1$.

The Cauchy-Schwarz inequality gives $T_1 \geq -r\left\|\dfrac{\partial n^{-1/2}Q_n(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}\right\|$.

We evaluate

$$\begin{aligned} T_2 &= \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\left[\frac{\partial^2 Q_n(\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} - \frac{\partial^2 Q_n(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right](\boldsymbol{\beta}-\boldsymbol{\beta}_0) + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\frac{\partial^2 Q_n(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) \\ &= o_{P_{\boldsymbol{\beta}_0,\pi}}(1) + n(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\mathcal{I}_0(\boldsymbol{\beta}-\boldsymbol{\beta}_0), \end{aligned}$$

where by (S_3), $\mathcal{I}_0 = \mathcal{D}_0^T\mathcal{W}_0\mathcal{D}_0$ is positive definite.

From the Cauchy-Schwarz inequality, using (P_1) and (P_2), we obtain

$$\begin{aligned} T_3 &\geq -r\sqrt{d_1}\sqrt{n}\max\Big\{p'_{\lambda_n}(|\beta_{0k}|),\ k=1,\dots,d_1\Big\} \\ &- \frac{r^2}{2}\max\Big\{\big|p''_{\lambda_n}(|\beta_{0k}|)\big|,\ k=1,\dots,d_1\Big\} + o(1) \end{aligned}$$

which is dominated by the leading term of T_2. Hence, using (P_2), we have

$$\begin{aligned} Q_n^P(\boldsymbol{\beta}) - Q_n^P(\boldsymbol{\beta}_0) &\geq -r\left\|\frac{\partial n^{-1/2}Q_n(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}\right\| + r^2\lambda_{\min}(\mathcal{I}_0) \\ &- r\sqrt{d_1}\sqrt{n}\max\Big\{p'_{\lambda_n}(|\beta_{0k}|),\ k=1,\dots,d_1\Big\} - o_{P_{\boldsymbol{\beta}_0,\pi}}(1) \end{aligned}$$

and for $\varepsilon > 0$ arbitrarily fixed, choosing r_ε such that, for sufficiently large n, the right hand side of the above inequality is strictly positive, with probability of at least $1-\varepsilon$, concludes the proof. Part (b) now follows from (a). □

For the SCAD penalty, if $\lambda_n \to 0$, then, with n large enough, $\max\{p'_{\lambda_n}(|\beta_{0k}|), k = 1,\dots d_1\} = 0$ and $\max\{p''_{\lambda_n}(|\beta_{0k}|), k=1,\dots d_1\} = 0$ so that (P_1) and (P_2) are satisfied. We now show that, with a regularization parameter chosen such that $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$, the penalized pseudo-QIF estimator, using the SCAD penalty performs as well as the oracle procedure. Firstly, it identifies the non-zero components correctly, with a probability converging to 1 and secondly, these estimators are as efficient as the estimator obtained if $\boldsymbol{\beta}_{20} = \mathbf{0}$ were known. In the general case, the assumption (P_3), below assures that the penalty function singular at the origin so that the penalized pseudo-QIF estimator possess the sparsity property.

Theorem 5 (Sparsity). *Assume that (Q), (S_1), (S_2), (S_3) hold, together with*

$$(P_3)\ \liminf_{n\to\infty} \liminf_{\theta\to 0^+} p'_{\lambda_n}(\theta)\lambda_n^{-1} > 0.$$

If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$, as $n \to \infty$, then, with probability converging to 1, for any sequence $\boldsymbol{\beta}_{1n}$ such that $n^{1/2}\|\boldsymbol{\beta}_{1n} - \boldsymbol{\beta}_{10}\| = O_{P_{\boldsymbol{\beta}_0,\pi}}$ and $C > 0$, we have

$$\mathbf{0} = \arg \min_{n^{1/2}\|\boldsymbol{\beta}_2\|\le C} Q_n\big[(\boldsymbol{\beta}_{1n}^T, \boldsymbol{\beta}_2^T)^T\big].$$

Proof. Let $\boldsymbol{\beta}_{1n}$ be such that $n^{1/2}\|\boldsymbol{\beta}_{1n} - \boldsymbol{\beta}_{10}\| = O_{P_{\boldsymbol{\beta}_0,\pi}}(1)$ and let $C > 0$ be arbitrary. With $k = d_1 + 1, \ldots, d$, using a Taylor's expansion of the function $\dfrac{\partial Q_n^P[(\boldsymbol{\beta}_{1n}^T, \boldsymbol{\beta}_2^T)^T]}{\partial \beta_k}$ (as a function of $\boldsymbol{\beta}_2$), around $\boldsymbol{\beta}_{20} = \mathbf{0}$, we have

$$\begin{aligned}
\frac{\partial Q_n^P\big[(\boldsymbol{\beta}_{1n}^T, \boldsymbol{\beta}_2^T)^T\big]}{\partial \beta_k} &= \frac{\partial Q_n\big[(\boldsymbol{\beta}_{1n}^T, \mathbf{0}^T)^T\big]}{\partial \beta_k} + \sum_{l=d_1+1}^{d} \frac{\partial^2 Q_n\big[(\boldsymbol{\beta}_{1n}^T, \boldsymbol{\beta}_2^{*T})^T\big]}{\partial \beta_k \partial \beta_l}\beta_l + np'_{\lambda_n}\big(|\beta_k|\big)\text{sign}(\beta_k), \\
&= \frac{\partial Q_n\big[(\boldsymbol{\beta}_{10}^T, \mathbf{0}^T)^T\big]}{\partial \beta_k} + o_{P_{\boldsymbol{\beta}_0\pi}}(1) + \sum_{l=d_1+1}^{d} \frac{\partial^2 Q_n\big[(\boldsymbol{\beta}_{10}^T, \mathbf{0}^T)^T\big]}{\partial \beta_k \partial \beta_l}\beta_l + o_{P_{\boldsymbol{\beta}_0\pi}}(1) \\
&\quad + np'_{\lambda_n}\big(|\beta_k|\big)\text{sign}(\beta_k), \qquad (4.5)
\end{aligned}$$

where $\boldsymbol{\beta}_2^*$ is such that $\|\boldsymbol{\beta}_2^*\| \le \|\boldsymbol{\beta}_2\| \le Cn^{-1/2}$ and we used the continuity of the first and second order partial derivatives of $Q_n(\boldsymbol{\beta})$, with respect to β_k, $k = d_1 + 1, \ldots, d$ around $\boldsymbol{\beta}_0$. From (3.6), the first term in (4.5) is $O_{P_{\boldsymbol{\beta}_0\pi}}(n^{1/2})$, whereas (3.7), together with (S_1) and (S_2), implies that the third term is $O_{P_{\boldsymbol{\beta}_0\pi}}(n^{1/2})$.

Since $n^{-1/2}\lambda_n^{-1} \to 0$, we obtain

$$\begin{aligned}
\frac{\partial Q_n^P\big[(\boldsymbol{\beta}_{1n}^T, \boldsymbol{\beta}_2^T)\big]}{\partial \beta_k} &= np'_{\lambda_n}\big(|\beta_k|\big)\text{sign}(\beta_k) + O_{P_{\boldsymbol{\beta}_0\pi}}(n^{1/2}) + o_{P_{\boldsymbol{\beta}_0\pi}}(1) \\
&= n\lambda_n\Big[\lambda_n^{-1}p'_{\lambda_n}\big(|\beta_k|\big)\text{sign}(\beta_k) + o_{P_{\boldsymbol{\beta}_0\pi}}(1) + o_{P_{\boldsymbol{\beta}_0\pi}}(n^{-1}\lambda_n^{-1})\Big],
\end{aligned}$$

which, due to (P_3), for large enough n, we have $\text{sign}\left\{\dfrac{\partial Q_n^P[(\boldsymbol{\beta}_{1n}^T, \boldsymbol{\beta}_2^T)]}{\partial \beta_k}\right\} = \text{sign}(\beta_k)$ for any $k = d_1 + 1, \ldots, d$ and $C > 0$. Hence, $Q_n^P[(\boldsymbol{\beta}_{1n}^T, \boldsymbol{\beta}_2^T)]$ has a local minimum within the ball $\{\boldsymbol{\beta}_2 : n^{1/2}\|\boldsymbol{\beta}_2\| \le C\}$, at $\boldsymbol{\beta}_2 = \mathbf{0}$. □

Under the assumptions of Theorem 5, with probability converging to 1, if the penalized pseudo-QIF estimator $\tilde{\boldsymbol{\beta}}_n = (\tilde{\boldsymbol{\beta}}_{1n}^T, \tilde{\boldsymbol{\beta}}_{2n}^T)^T$ is $\sqrt{n}$-consistent, then it must satisfy $\tilde{\boldsymbol{\beta}}_{2n} = \mathbf{0}$.

Let $\mathcal{D}_{10}$ be the $Ld_1 \times d_1$ matrix obtained from $\mathcal{D}_0\left\{\binom{\boldsymbol{\beta}_{10}}{\mathbf{0}}\right\}$ by selecting the rows and columns corresponding to $\boldsymbol{\beta}_{10}$, i.e. include the first d_1 columns and rows numbered by $\eta d+1,\ldots,\eta d+d_1$, where $\eta = 0,\ldots,(L-1)$. Similarly, let $\mathcal{W}_{10}$ be the $Ld_1 \times Ld_1$ matrix obtained from $\mathcal{W}_0\left\{\binom{\boldsymbol{\beta}_{10}}{\mathbf{0}}\right\}$ by selecting the rows and columns which are numbered as $\eta d+1,\ldots,\eta d+d_1$, where $\eta = 0,\ldots,(L-1)$. Finally, let $\mathcal{I}_{10}$ denote the upper $d_1 \times d_1$ corner of the matrix $\mathcal{I}_0\left\{\binom{\boldsymbol{\beta}_{10}}{\mathbf{0}}\right\}$ and

$$\mathbf{b}_n = \Big(p'_{\lambda_n}\big(|\beta_{01}|\big)\operatorname{sign}(\beta_{01}),\ldots,p'_{\lambda_n}\big(|\beta_{0d_1}|\big)\operatorname{sign}(\beta_{0d_1})\Big)^T,$$
$$\mathbf{B}_n = \operatorname{diag}\Big\{p''_{\lambda_n}\big(|\beta_{01}|\big),\ldots,p''_{\lambda_n}\big(|\beta_{0d_1}|\big)\Big\}.$$

Theorem 6 shows that the limiting distribution of $\tilde{\boldsymbol{\beta}}_{1n}$ is normal and that, for certain penalties, including the SCAD, the penalized pseudo-QIF of $\boldsymbol{\beta}_{10}$ is asymptotically as efficient as the estimator obtained if $\boldsymbol{\beta}_{20} = \mathbf{0}$ were known.

Theorem 6 (Limiting distribution)**.** *Assume that (Q), (S_1), (S_2), (S_3) and (P_3) hold. If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$, as $n \to \infty$, then, with probability converging to 1, the $\sqrt{n}$-consistent local minimizer of $Q_n^P(\boldsymbol{\beta})$, $\tilde{\boldsymbol{\beta}}_n = (\tilde{\boldsymbol{\beta}}_{1n}^T, \mathbf{0}^T)^T$ satisfies*

$$\sqrt{n}\big[2\mathcal{I}_{10} + \mathbf{B}_n\big]\Big\{\tilde{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10} + [2\mathcal{I}_{10} + \mathbf{B}_n]^{-1}\mathbf{b}_n\Big\} \xrightarrow{\mathcal{L}} \mathcal{N}\Big(\mathbf{0}, 4\mathcal{D}_{10}^T\mathcal{W}_{10}\boldsymbol{\Sigma}_0^{(1)}\mathcal{W}_{10}\mathcal{D}_{10}\Big),$$

under $P_{\boldsymbol{\beta}_0\pi}$, where $\boldsymbol{\Sigma}_0^{(1)}$ denotes the $Ld_1 \times Ld_1$ matrix obtained from $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_d + f\boldsymbol{\Sigma}_{(\boldsymbol{\beta}_{10}^T,\mathbf{0}^T)^T}$ by selecting the rows and columns which are numbered as $\eta d+1,\ldots,\eta d+d_1$ and $\eta = 0,\ldots,(L-1)$.

Proof. Since the local minimizer of $Q_n^P(\boldsymbol{\beta}_1, \mathbf{0})$, $\tilde{\boldsymbol{\beta}}_{1n}$, is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}_{10}$, a Taylor's expansion gives

$$\frac{\partial Q_n^P}{\partial \boldsymbol{\beta}_1}\left\{\begin{pmatrix}\tilde{\boldsymbol{\beta}}_{1n}\\ \mathbf{0}\end{pmatrix}\right\} = \frac{\partial Q_n}{\partial \boldsymbol{\beta}_1}\left\{\begin{pmatrix}\boldsymbol{\beta}_{10}\\ \mathbf{0}\end{pmatrix}\right\} + \frac{\partial^2 Q_n}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T}\left\{\begin{pmatrix}\boldsymbol{\beta}_1^*\\ \mathbf{0}\end{pmatrix}\right\}(\tilde{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) + n\mathbf{b}_n$$
$$+\, n\operatorname{diag}\Big\{p''_{\lambda_n}(|\beta_1^{**}|),\ldots,p''_{\lambda_n}(|\beta_{d_1}^{**}|)\Big\}(\tilde{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}),$$

where $\boldsymbol{\beta}_1^*$ and β_k^{**} are such that $\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_{10}\| \le \|\tilde{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}\|$ and $|\beta_k^{**} - \beta_{0k}| \le |\tilde{\beta}_{nk} - \beta_{0k}|$, with $k = 1\ldots,d_1$. After rearranging, we obtain

$$-\sqrt{n}\frac{\partial n^{-1}Q_n}{\partial \boldsymbol{\beta}_1}\left\{\begin{pmatrix}\boldsymbol{\beta}_{10}\\ \mathbf{0}\end{pmatrix}\right\} = \frac{\partial^2 n^{-1}Q_n}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T}\left\{\begin{pmatrix}\boldsymbol{\beta}_{10}\\ \mathbf{0}\end{pmatrix}\right\}\sqrt{n}(\tilde{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) + \sqrt{n}\mathbf{b}_n + \mathbf{B}_n\sqrt{n}(\tilde{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10})$$
$$+ \left[\frac{\partial^2 n^{-1}Q_n}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T}\left\{\begin{pmatrix}\boldsymbol{\beta}_1^*\\ \mathbf{0}\end{pmatrix}\right\} - \frac{\partial^2 n^{-1}Q_n}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T}\left\{\begin{pmatrix}\boldsymbol{\beta}_{10}\\ \mathbf{0}\end{pmatrix}\right\}\right]\sqrt{n}(\tilde{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10})$$
$$+ \operatorname{diag}\Big\{p''_{\lambda_n}(|\beta_1^{**}|) - p''_{\lambda_n}(|\beta_{01}|),\ldots,p''_{\lambda_n}(|\beta_{d_1}^{**}|) - p''_{\lambda_n}(|\beta_{0d_1}|)\Big\}\sqrt{n}(\tilde{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}).$$

Due to (3.7), $\frac{\partial^2 n^{-1}Q_n}{\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_1^T}\Big\{\Big(\begin{smallmatrix}\boldsymbol{\beta}_{10}\\ \mathbf{0}\end{smallmatrix}\Big)\Big\} = 2\mathcal{I}_{10} + o_{P_{\boldsymbol{\beta}_0\pi}(1)}$ which, together with the fact that the entries of $\frac{\partial^2 n^{-1}Q_n}{\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_1^T}\Big\{\Big(\begin{smallmatrix}\boldsymbol{\beta}_{1}\\ \mathbf{0}\end{smallmatrix}\Big)\Big\}$ are continuous at $(\boldsymbol{\beta}_{10}^T, \mathbf{0}^T)^T$, the asymptotic distribution of

$$\sqrt{n}\big[2\mathcal{I}_{10} + \mathbf{B}_n\big]\Big\{\tilde{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10} + [2\mathcal{I}_{10} + \mathbf{B}_n]^{-1}\mathbf{b}_n\Big\}$$

is equivalent to that of $-\sqrt{n}\frac{\partial n^{-1}Q_n}{\partial\boldsymbol{\beta}_1}\Big\{\Big(\begin{smallmatrix}\boldsymbol{\beta}_{10}\\ \mathbf{0}\end{smallmatrix}\Big)\Big\}$. The latter is equivalent to the limiting distribution of $-2\frac{\partial\mathbf{q}_n}{\partial\boldsymbol{\beta}_1^T}\Big\{\Big(\begin{smallmatrix}\boldsymbol{\beta}_{10}\\ \mathbf{0}\end{smallmatrix}\Big)\Big\}\Big\{\mathbf{C}_n\Big\{\Big(\begin{smallmatrix}\boldsymbol{\beta}_{10}\\ \mathbf{0}\end{smallmatrix}\Big)\Big\}\Big\}^{-1}\sqrt{n}\mathbf{q}_n\Big\{\Big(\begin{smallmatrix}\boldsymbol{\beta}_{10}\\ \mathbf{0}\end{smallmatrix}\Big)\Big\}$, using (3.6), which, in turn, is given by $\mathcal{N}\big(\mathbf{0}, 4\mathcal{D}_{10}^T\mathcal{W}_{10}\boldsymbol{\Sigma}_0^{(1)}\mathcal{W}_{10}\mathcal{D}_{10}\big)$, under the joint model-design probability, due to Theorem 1 and assumptions (S_1) and (S_2). □

Remark 2. Under the assumptions of Theorem 6, if $\|\mathbf{b}_n\| \to 0$ and $\|\mathbf{B}_n\| \to 0$ then

$$\sqrt{n}\big(\tilde{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}\big) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(\mathbf{0}, \mathcal{I}_{10}^{-1}[\mathcal{D}_{10}^T\mathcal{W}_{10}\boldsymbol{\Sigma}_0^{(1)}\mathcal{W}_{10}\mathcal{D}_{10}]\mathcal{I}_{10}^{-1}\Big),$$

under $P_{\boldsymbol{\beta}_0\pi}$, showing that the penalized pseudo-QIF estimator is as efficient as the oracle estimator that assumes the true model, with $\boldsymbol{\beta}_{20} = \mathbf{0}$ is known.

By Theorem 6, an estimator of the asymptotic variance of $\tilde{\boldsymbol{\beta}}_{1n}$ is given by

$$n^{-1}\left[\frac{\partial^2 n^{-1}Q_n}{\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_1^T}\left\{\begin{pmatrix}\tilde{\boldsymbol{\beta}}_{1n}\\ \mathbf{0}\end{pmatrix}\right\} + \mathbf{B}_n\right]^{-1}\hat{\mathcal{V}}_0^{(1)}\left[\frac{\partial^2 n^{-1}Q_n}{\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_1^T}\left\{\begin{pmatrix}\tilde{\boldsymbol{\beta}}_{1n}\\ \mathbf{0}\end{pmatrix}\right\} + \mathbf{B}_n\right]^{-1},$$

where $\hat{\mathcal{V}}_0^{(1)} = 4\mathcal{D}_{1n}(\tilde{\boldsymbol{\beta}}_{1n})^T\big[\mathbf{C}_{1n}(\tilde{\boldsymbol{\beta}}_{1n})\big]^{-1}\hat{\boldsymbol{\Sigma}}_0^{(1)}\big[\mathbf{C}_{1n}(\tilde{\boldsymbol{\beta}}_{1n})\big]^{-1}\mathcal{D}_{1n}(\tilde{\boldsymbol{\beta}}_{1n})$, with $\hat{\boldsymbol{\Sigma}}_0^{(1)} = \hat{\boldsymbol{\Sigma}}_d + f\hat{\boldsymbol{\Sigma}}_{(\tilde{\boldsymbol{\beta}}_{1n}^T,\mathbf{0})^T}$. In addition, $\mathcal{D}_{1n}(\tilde{\boldsymbol{\beta}}_{1n})$ denotes the $Ld_1 \times d_1$ matrix obtained from $\mathcal{D}_n\Big\{\Big(\begin{smallmatrix}\tilde{\boldsymbol{\beta}}_{1n}\\ \mathbf{0}\end{smallmatrix}\Big)\Big\}$ by selecting the first d_1 columns and rows numbered by $\eta d + 1, \ldots, \eta d + d_1$ and $\mathbf{C}_{1n}(\tilde{\boldsymbol{\beta}}_{1n})$ is the $Ld_1 \times Ld_1$ matrix obtained from $\mathbf{C}_n\Big\{\Big(\begin{smallmatrix}\tilde{\boldsymbol{\beta}}_{1n}\\ \mathbf{0}\end{smallmatrix}\Big)\Big\}$ by selecting the rows and columns which are numbered as $\eta d + 1, \ldots, \eta d + d_1$, where $\eta = 0, \ldots, (L-1)$.

In complex sample surveys the asymptotic variance of an estimator may have a complicated form and resampling methods have to be used. Wang et al. (2014) used the estimating function bootstrap method for variance estimation under the penalized pseudo-GEE method for variable selection. A bootstrap method can be employed to obtain a variance estimator of $\tilde{\boldsymbol{\beta}}_{1n}$ by generating bootstrap weights using the Rao-Wu rescaling method (see Rao and Wu, 1988; Rao et al., 1992) and taking a one-step bootstrap. Then, as in (4.2), for each set of bootstrap weights, the value of the estimator of the non-zero components can be updated as follows

$$\tilde{\boldsymbol{\beta}}_{1n}^{(b)} = \tilde{\boldsymbol{\beta}}_{1n} - \left[\frac{\partial^2 Q_n^{(b)}(\tilde{\boldsymbol{\beta}}_n)}{\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_1^T} + n\boldsymbol{\Gamma}(\tilde{\boldsymbol{\beta}}_n)\right]^{-1}\left[\frac{\partial Q_n^{(b)}(\tilde{\boldsymbol{\beta}}_n)}{\partial\boldsymbol{\beta}_1} + n\boldsymbol{\Gamma}(\tilde{\boldsymbol{\beta}}_n)(\tilde{\boldsymbol{\beta}}_{1n})\right], \tag{4.6}$$

where $Q_n^{(b)}(\tilde{\boldsymbol{\beta}}_{1n})$ is the bootstrap weighted QIF, calculated using the b-th set of bootstrap weights. Consequently, a bootstrap variance estimator is given by

$$\hat{\mathbf{V}}^B_{\tilde{\boldsymbol{\beta}}_{1n}} := \frac{1}{B}\sum_{b=1}^{B}\left(\tilde{\boldsymbol{\beta}}^{(b)}_{1n} - \tilde{\boldsymbol{\beta}}_{1n}\right)\left(\tilde{\boldsymbol{\beta}}^{(b)}_{1n} - \tilde{\boldsymbol{\beta}}_{1n}\right)^T. \tag{4.7}$$

5 Numerical Results

We generate a finite population with correlated binary responses from the following marginal logistic model

$$\text{logit}\mu_{ij}(\boldsymbol{\beta}_0) = \mathbf{x}_{ij}^T\boldsymbol{\beta}_0,\ x_{ij}^k \overset{\text{indep}}{\sim} \mathcal{U}(0, 0.8),\ i = 1, \ldots N,\ j = 1, \ldots, m,\ k = 1, \ldots d,$$

$N = 30000$, $m = 5$, $d = 10$, $\boldsymbol{\beta}_0 = (0.8, -0.7, -0.6, 0, 0, 0, 0, 0, 0, 0)^T$, choosing the exchangeable correlation matrix with parameter $\alpha = 0.4$ as the true correlation. In this case, the basis matrices are

$$M_1 = \mathbf{I}_5 \quad \text{and} \quad \mathbf{M}_2 = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

and $d_1 = 3$.

From each generated finite population we obtain a sample of n clusters, using informative sampling: clusters are selected with probability proportional to the size measures $z_i = \sum_{j=1}^{5} y_{ij} + 1$, with replacement. We consider sample sizes of $n = 300$ and $n = 500$ and repeat the procedure of generating a finite population and then selecting the sample $H = 500$ times.

For each sample, we apply each the following methods: unweighted QIF with SCAD penalty (UNWGT), weighted QIF with SCAD penalty (PQIF), weighted GEE with SCAD penalty (PGEE) and ORACLE, which is the weighted QIF under the true model (with three nonzero coefficients and seven zero coefficients).

To evaluate the performance of the proposed method, two working correlations: exchangeable (EX) and fist-order autoregressive (AR1) are considered for each procedure and assessed in Table 1, as follows. The columns labeled as "(True)", "(Over)" and "(Under)" give, respectively, the percentage of times the true model (only first three components of the estimator are non-zero) is selected, the percentage of times the variables are over-selected (more than the first three components of the estimator are non-zero) and the percentage of times the variables are under-selected (at least one of the first three components of the estimator is zero). The results in Table 1 show that the unweighted PQIF is not capable of variable selection and only less than 10% of the time the true model selected. Furthermore, PQIF and PGEE both perform well in terms of selecting the true model (True), with PQIF leading to slightly larger values.

Table 1: Correlated binary responses in surveys: comparisons of the unweighted QIF, penalized pseudo-QIF, penalized pseudo-GEE and oracle QIF under exchangeable and AR(1) working correlation matrices. The columns labeled as "(True)" indicate the percentage of times the true model is selected (i.e. only first three components of the estimator are non-zero). The columns labeled as "(Over)" indicate the percentage of times the variables are over-selected (i.e. more than the first three components of the estimator are non-zero). The columns labeled as "(Under)" give the percentage of times the variables are under-selected (i.e. at least one of the first three components of the estimator is zero).

		Exchangeable				AR(1)			
Sample size	Method	(True)	(Over)	(Under)	MSE	(True)	(Over)	(Under)	MSE
$n = 300$	UNWGT	5.6	34.2	60.2	0.538	2.4	16.0	71.6	0.610
	PQIF	80.8	11.2	8.0	0.129	75.0	11.4	13.6	0.163
	PGEE	73.0	10.6	16.4	0.172	72.2	10.4	17.4	0.181
	ORACLE	100.0	-	-	0.072	100.0	-	-	0.080
$n = 500$	UNWGT	7.0	66.6	26.4	0.384	5.4	59.6	35.0	0.435
	PQIF	91.2	8.6	0.2	0.049	87.8	11.0	1.2	0.065
	PGEE	88.2	10.0	1.8	0.054	87.0	10.2	1.8	0.062
	ORACLE	100.0	-	-	0.039	100.0	-	-	0.046

In addition, we compute the Monte Carlo average MSE of the estimators, reported in Table 1 as MSE

$$MSE = \frac{1}{H}\sum_{h=1}^{H}\left(\tilde{\boldsymbol{\beta}}_n^{(h)} - \boldsymbol{\beta}_0\right)^T\left(\tilde{\boldsymbol{\beta}}_n^{(h)} - \boldsymbol{\beta}_0\right),$$

where $\tilde{\boldsymbol{\beta}}_n^{(h)}$ is the survey-weighted penalized estimator, calculated at iteration h, with $h = 1, \ldots, H$ and $H = 500$. Table 1 shows that UNWGT leads to larger MSE, while PQIF exhibits smaller MSE than PGEE. As expected, ORACLE performs the best in terms of correctly selecting the true model and MSE.

In Table 2 we report the absolute relative bias of the non-zero coefficients, calculated as

$$ARB(\tilde{\beta}_{nk}) = \frac{\left|H^{-1}\sum_{h=1}^{H}\tilde{\beta}_{nk}^{(h)} - \beta_{0k}\right|}{|\beta_{0k}|} \times 100,\ k = 1, 2, 3,$$

where $\tilde{\beta}_{nk}^{(h)}$ is the k-th component of $\tilde{\boldsymbol{\beta}}_n^{(h)}$ and β_{0k} is the k-th entry of $\boldsymbol{\beta}_0$. The unweighted PQIF yields biased results, whereas the relative biases of PQIF estimates are less that those of PGEE. In addition, in contrast to the unweighted methods, the values the relative bias decrease in case of the weighted methods.

Table 2: Correlated binary responses in surveys: comparisons of percent absolute relative bias (ARB) of regression coefficients, obtained from the unweighted QIF, penalized pseudo-QIF, penalized pseudo-GEE and oracle QIF, under exchangeable and AR(1) working correlation matrices.

		Exchangeable			AR(1)		
Sample size	Method	$\tilde{\beta}_{n1}$	$\tilde{\beta}_{n2}$	$\tilde{\beta}_{n3}$	$\tilde{\beta}_{n1}$	$\tilde{\beta}_{n2}$	$\tilde{\beta}_{n3}$
$n = 300$	UNWGT	28.8	30	55	28.8	37.1	61.7
	PQIF	1.3	1.4	6.7	2.5	0	8.3
	PGEE	7.5	1.4	13.3	6.3	1.4	13.3
	ORACLE	1.3	1.4	3.3	0	0	3.3
$n = 500$	UNWGT	28.4	22.9	41.4	28.6	26.6	46.5
	PQIF	1.7	2.3	4.0	1.4	2.0	3.9
	PGEE	3.9	1.2	5.8	3.5	0.9	4.5
	ORACLE	1.8	2.4	3.9	1.5	2.0	3.2

Next, we evaluate the performance of the bootstrap variance estimator given by (4.7). To obtain the bootstrap standard error, we draw 500 bootstrap samples of size $n-1$, each selected with replacement and equal probabilities from the n sampled units. For each b-th bootstrap sample, the bootstrap weights are calculated using the rescaling formula

$$w_i^{(b)} = w_i \frac{n}{n-1} t_i^{(b)},$$

where $t_i^{(b)}$ is the number of repetitions of unit i in the b-th bootstrap sample.

As in Fan and Li (2001), to illustrate the performance of the proposed variance estimator, we first compute SD (for each $k = 1, 2, 3$), defined as the ratio between the median absolute deviation of the estimator $\tilde{\beta}_{nk}$

$$\text{median}\left\{ \left| \tilde{\beta}_{nk}^{(h)} - \text{median}\{\tilde{\beta}_{nk}^{(h)},\ h = 1, \ldots, H\} \right|,\ h = 1, \ldots, H \right\},$$

and 0.6745. Due to the normality of the limiting distribution of $\tilde{\beta}_{nk}$, this value is an estimate of its true standard error and in Table 3, we compare it with the one obtained from the bootstrap method, denoted as SD_m. The latter is obtained as the median of the H bootstrap estimated standard deviations

$$\text{median}\left\{ \sqrt{\hat{\mathbf{v}}_{\tilde{\beta}_{nk}}^{B,h}},\ h = 1, \ldots, H \right\},$$

where, at each iteration h, $\hat{\mathbf{v}}_{\tilde{\beta}_{nk}}^{B,h} = B^{-1} \sum_{b=1}^{B} \left(\tilde{\beta}_{nk}^{(b,h)} - \tilde{\beta}_{nk}^{(h)} \right)^2$ is the bootstrap variance estimate of $\tilde{\beta}_{nk}^{(h)}$. Here, $\tilde{\beta}_{nk}^{(b,h)}$ denotes the k-th component of $\tilde{\boldsymbol{\beta}}_{1n}^{(b,h)}$ whose form is given in (4.6), $h = 1, \ldots, H$.

Table 3: Correlated binary responses in surveys: accuracy of the bootstrap approximation of the variance of the penalized pseudo-QIF estimator, under exchangeable and AR(1) working correlation matrices.

Sample size	$\tilde{\beta}_{n1}$ SD	$\tilde{\beta}_{n1}$ SD_m (SD_{mad})	$\tilde{\beta}_{n2}$ SD	$\tilde{\beta}_{n2}$ SD_m (SD_{mad})	$\tilde{\beta}_{n3}$ SD	$\tilde{\beta}_{n3}$ SD_m (SD_{mad})
			Exchangeable			
$n = 300$	0.172	0.163 (0.016)	0.152	0.161 (0.015)	0.142	0.156 (0.015)
$n = 500$	0.112	0.127 (0.011)	0.125	0.125 (0.010)	0.099	0.123 (0.010)
			AR(1)			
$n = 300$	0.181	0.172 (0.018)	0.170	0.170 (0.016)	0.162	0.165 (0.016)
$n = 500$	0.124	0.134 (0.011)	0.137	0.133 (0.011)	0.125	0.131 (0.011)

Furthermore, for each $k = 1, 2, 3$, the median absolute deviation error of the 500 bootstrap estimated standard errors, denoted as SD_{mad} is evaluated as the ratio between

$$\text{median}\left\{\left|\sqrt{\hat{\mathbf{v}}_{\tilde{\beta}_{nk}}^{B,h}} - SD_k\right|,\ h = 1, \ldots, H\right\},$$

and 0.6745, where SD_k is the estimate of the true standard error of $\tilde{\beta}_{nk}$. Results in Table 3 compare estimates obtained under the two working correlations and different sample sizes. They show that in each case, the one-step bootstrap method for variance estimation performs well in tracking the value of the true standard error of $\tilde{\beta}_{nk}$, $k = 1, 2, 3$.

6 Conclusions

In this article, based on a penalized survey-weighted quadratic inference function, we proposed a new approach to model selection for longitudinal survey data. Under the joint model-design framework, the procedure performs parameter estimation and model selection simultaneously and takes into account the within-cluster correlation and the complex sampling survey-design features. The method is directly applicable to discrete or continuous data and has the additional advantage of incorporating the within-cluster correlation without specifying the full likelihood function, or estimating the correlation parameters. The penalized survey-weighted approach was shown to be consistent for model selection and to satisfy the oracle property. Our simulation study illustrated the finite sample performance of the proposed procedure in terms of true model selection, under informative sampling.

In complex sample surveys the asymptotic variance of an estimator may have a complicated form and, in practice, to obtain a variance estimator, the first stage sampling fractions under a multistage design are assumed to be small and treated as if the first stage sampling units are drawn with replacement. Following the lines of Section 4.3 in Dumitrescu et al. (2021), a simplified variance estimator of the proposed penalized pseudo-QIF can be derived under this set up and it would be of interest to compare its performance to the bootstrap variance estimator given in (4.7).

As we advocate throughout the paper, the survey-weighted quadratic inference function approach for modelling longitudinal survey data is a flexible and convenient tool. It has the appealing feature of yielding efficient estimators, but it also provides a pseudolikelihood ratio type statistics to test composite hypotheses on model parameters, and a statistic for testing the goodness-of-fit of the marginal model. Furthermore, due to modern challenges in the analysis of contemporary survey data, in some model selection problems the number of parameters is large. In the case of non-survey responses, in Fan and Peng (2004) and Cho and Qu (2013) it was shown that the rate of divergence of the number of parameters influences the convergence rate of penalized estimators. However, for complex sampling surveys, to obtain accurate results, the survey design features should be taken into consideration and we anticipate that the survey-weighted quadratic inference function can be employed to analyze high-dimensional survey data. We intend to formalize these ideas in a future project.

Acknowledgements

The authors would like to thank the anonymous referee for the provided comments which have improved the presentation of the article. J. N. K. Rao wishes to thank the Natural Sciences and Engineering Research Council of Canada for supporting this research.

Disclaimer: The contents of this report reflect the views of the author and not necessarily the official views or opinions of Statistics Canada.

References

Carrillo, I. A., Chen, J. and Wu, C. (2010), “The pseudo-GEE approach to the analysis of longitudinal surveys,” *Canadian Journal of Statistics,* 38, 540–554.

Cho, H. and Qu, A. (2013), “Model selection for correlated data with diverging number of parameters,” *Statistica Sinica*, 23, 901–927.

Dumitrescu, L., Qian, W. and Rao, J. N. K. (2021), “Inference for longitudinal data from complex sampling surveys: an approach based on quadratic inference functions,” *Scandinavian Journal of Statistics*, 48, 246–274.

Fan, J. and Li, R. (2001), “Variables selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J. and Peng, H. (2004), “Nonconcave penalized likelihood with a diverging number of parameters,” *The Annals of Statistics*, 32, 928–961.

Newey, W. K. and McFadden, D. (1986), "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, ed1, 4, chapter 36. Elsevier.

Qian, W. (2018), *Analysing Correlated Data from Survey with Complex Design*. Carleton University dissertation.

Qu, A., Lindsay, B. G. and Li, B. (2000), "Improving generalised estimating equations using quadratic inference functions," *Biometrika*, 87, 823–836.

Rao, J. N. K. (1998), "Marginal models for repeated observation: Inference with survey data," *American Statistical Association - Proceedings of the Survey Research Methods Section*, 76–82.

Rao, J. N. K. and Wu, C. F. (1988), "Resampling inferences with complex survey data," *Journal of the American Statistical Association*, 83, 231–241.

Rao, J. N. K., Wu, C. F. and Yue, K. (1992), "Some recent work on re-sampling methods for complex surveys," *Survey Methodology,* 18, 209–217.

Roberts, G., Ren, Q. and Rao, J. N. K. (2009), "Using marginal mean models for data from longitudinal surveys with a complex design: some advances in methods," In *Methodology of Longitudinal Surveys*, R. Lynn (Ed.), John Wiley & Sons, New York, 351–366.

Rubin-Bleuer, S. and Schiopu Kratina, I. (2005), "On the two-phase framework for joint model and design-based inference," *The Annals of Statistics*, 33, 2789–2810.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, 58 267–288.

Wang, L. and Qu, A. (2009), "Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach," *Journal of the Royal Statistical Society, Series B*, 71, 177–190.

Wang, L., Wang, S. and Wang, G. (2014), "Variable selection and estimation for longitudinal survey data," *Journal of Multivariate Analysis*, 130, 409–424.

Zhao, P., Haziza, D. and Wu, C. (forthcoming), "Sample empirical likelihood and the design-based oracle variable selection theory," *Statistica Sinica*, forthcoming.

Zou, H. (2006), "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Received: April 2, 2021

Accepted: April 21, 2021