Statistics and Its Interface Volume 14 (2021) 83–94

Penalized empirical likelihood for high-dimensional generalized linear models

Xia Chen* and Liyue Mao

We develop penalized empirical likelihood for parameter estimation and variable selection in high-dimensional generalized linear models. By using adaptive lasso penalty function, we show that the proposed estimator has the oracle property. Also, we consider the problem of testing hypothesis, and show that the nonparametric profiled empirical likelihood ratio statistic has asymptotic chi-square distribution. Some simulations and an application are given to illustrate the performance of the proposed method.

AMS 2000 subject classifications: Primary 62F12; secondary 62J12.
Keywords and phrases: Penalized empirical likelihood, High-dimensional data, Variable selection, Generalized linear models.

1. INTRODUCTION

Empirical likelihood (EL), introduced by Owen [23], is a nonparametric statistical method. Its properties have been examined by Owen [25], Qin and Lawless [27], Diciccio and Hall [9], Chen and Cui [6] and references therein. After proposed by Owen [23], EL has been applied to many statistical models such as linear regression models (e.g., Owen [24]), partially linear models (e.g., Wang and Jing [35]) and generalized linear models (e.g., Kolaczyk [17], Chen and Cui [5]). Generalized linear models (GLMs) were introduced by Nelder and Wedderburn [22], where the response vector are assumed to have an exponential family distribution. Wedderburn [36] proposed the quasi-likelihood method and showed that the distributional assumption on response can be replaced by a weaker one on the form of their mean and variance. When the number of covariates is fixed, standard EL for generalized linear models has been considered by Kolaczyk [17]. The advantage of the EL in this case is in its construction of confidence regions of natural shape and orientation, rather than in parameter estimation. To fully use the information in variance structure, Chen and Cui [5] proposed an extended EL for generalized linear models.

Recently, high-dimensional data, whose dimensionality p tends to infinity as the sample size $n \to \infty$, are becoming prevalent in many areas, such as hyperspectral imagery, internet portals, finance data, especially datasets in genomics and other areas of computational biology. The emergence of high-dimensional data brings challenges to many traditional statistical methods and theory (e.g., Bai and Saranadasa [1], Fan and Lv [13]). Thereby, studying the performance of traditional statistical methods in high-dimensional settings and establishing new approaches are necessary and exigent. The changing landscape of dimensionality from low to high brings new challenges to EL method (e.g., Chen et al. [7], Hjort et al. [16]).

When dimensionality diverges, variable selection through regularization has proven to be effective. As argued in Fan and Lv [12] and Hastie et al. [15], penalized likelihood can properly adjust the biasvariance trade-off so that the performance improvement can be achieved, see Tibshirani [30], Fan and Li [11], Candes and Tao [3], Zou [39] and Fan and Lv [12] for penalized likelihood approaches and discussions. A new and efficient variable selection approach, penalized empirical likelihood (PEL) introduced by Tang and Leng [29], was applied to analyze mean vector in multivariate analysis and regression coefficients in linear models with diverging number of parameters. As demonstrated in Tang and leng [29], the PEL has merits in both efficiency and adaptivity stemming from a nonparametric likelihood method. Also, the PEL method possesses the same merit of the EL which only uses the data to determine the shape and orientation of confidence regions and without estimating the complex covariance. There are many studies in the literatures concerning the PEL approach of different models. Ren and Zhang [28] studied PEL method in conditional moment restriction model. Leng and Tang [19] applied the PEL approach to parametric estimation and variable selection for general estimating equation. Lahiri and Mukhopadhyay [18] studied high-dimensional PEL. Wu et al. [37] used the PEL method to study linear regression model with right censored data. Fan et al. [10] applied PEL method to partially linear varying coefficient model. Chang et al. [4] proposed a new PEL by applying two penalty functions respectively regularizing the model parameters and the associated Lagrange multiplier in the optimizations of EL. Wang et al. [32] discussed the PEL for the sparse Cox regression model.

Nonetheless, as far as we know PEL for high-dimensional GLMs is less studied, especially in the case that there exists a disperse parameter σ^2 between the relationship of mean and variance. In the existing literatures, such as Park and Trevor [26] and Liang et al. [20], σ^2 set to be known or it is

*Corresponding author.

not considered usually, but that may not set up in modeling count variable, which make the discussion of σ^2 significance.

Taking these issues into account, in this paper, we propose PEL with adaptive lasso in high-dimensional GLMs based on three estimating equations. Our contribution of the paper is two folded. First, we employ the PEL approach to build parsimonious and robust models and obtain the oracle property of the PEL estimator. Second, we also investigate the PEL ratio for hypothesis testing and constructing confidence region of unknown parameters. Simulation studies and real data analysis indicate that the efficiency of the proposed PEL estimator is encouraging. Although we focus on the adaptive Lasso in this paper, the penalty function could be replaced by other penalties such as SCAD (see Fan Li [11]), or a family of penalties proposed by Lv and Fan [21].

This paper formulates an empirical likelihood for GLMs that incorporates extra constraints which explore the provided variance structure proposed by Chen and Cui [5], but they only considered the EL for GLMs when the number of covariates is fixed. Our results extend the results of Chen and Cui [5] to the case of high-dimensional variable selection. Also, we generalize the results of Tang and Leng [29] to the high-dimensional GLMs and improve the rate of dimensionality.

The rest of the article is organized as follows. In Section 2, we introduce the methodology and main results. Algorithm is presented in Section 3. Section 4 states some simulations to show the finite sample performance of the proposed method. A real data example is given in Section 5. Technical proofs are given in the Appendix.

2. PENALIZED EMPIRICAL LIKELIHOOD

Let a scalar random variable Y be the response of a random vector $X \in R^p$ such that

$$\mathrm{E}(Y|X) = \mu(\beta) = G(X^T\beta)$$

and

$$\mathrm{Var}(Y|X) = \sigma^2 V\{G(X^T\beta)\}$$

where β is a $p \times 1$ vector of unknown parameter, $G(\cdot)$ is a known smooth link function, $V(\cdot)$ is a known variance function and σ^2 is the dispersion parameter. This is the framework of generalized linear models under which the quasi-likelihood has been a popular tool. The log quasi-likelihood ratio of β is defined as

$$Q\{Y;\mu(\beta)\} = \int_Y^{\mu(\beta)} \frac{Y-u}{V(u)}du. \tag{1}$$

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an independent and identically distributed sample, and $\mu_i(\beta) = G(X_i^T\beta)$. Differentiating (1) with respect to β, the quasi-score function can be written as

$$\frac{\partial Q(Y_i, \mu_i(\beta))}{\partial\beta} = \frac{\{Y_i - \mu_i(\beta)\}G'(X_i^T\beta)X_i}{V\{\mu_i(\beta)\}}.$$

From the variance structure of the model, we get

$$\mathrm{E}\{[Y_i - G(X_i^T\beta)]^2\} = \sigma^2 V\{G(X_i^T\beta)\}.$$

For $1 \le i \le n$, define

$$Z_i^{(1)}(\sigma^2,\beta) = \frac{[Y_i - G(X_i^T\beta)]^2}{\sigma^4 V\{G(X_i^T\beta)\}} - \frac{1}{\sigma^2},$$
$$Z_i^{(2)}(\beta) = \frac{[Y_i - G(X_i^T\beta)]G'(X_i^T\beta)X_i}{V\{G(X_i^T\beta)\}}.$$

The advantage of the empirical likelihood under $Z_i^{(j)}(j = 1, 2)$ is in its construction of confidence regions of natural shape and orientation, rather than in parameter estimation because $Z_i^{(2)}$ doesn't fully use information on variance function. When $V' = V'(G(X_i^T\beta)) \neq 0$, there may contain useful information on variance structure. To utilize this information, Chen and Cui [5] considered the following extra constraints

$$Z_i^{(3)}(\sigma^2,\beta) = \left(\frac{[Y_i - G(X_i^T\beta)]^2}{\sigma^4 V\{G(X_i^T\beta)\}} - \frac{1}{\sigma^2}\right)\omega(X_i^T\beta, X_i),$$

where ω is a r-dimensional weight function and $1 \le r \le p$. An optimal ω was given in [5]. Numerical studies show that this extra constraints leads to error reduction in parameter estimation. In this article, we use

$$Z_i(\sigma^2,\beta) = (Z_i^{(1)T}(\sigma^2,\beta), Z_i^{(2)T}(\beta), Z_i^{(3)T}(\sigma^2,\beta))^T$$

as the auxiliary random vector.

If (σ_0^2, β_0) is the true value of the parameters, $\mathrm{E}(Z_i(\sigma_0^2, \beta_0)) = 0$. Following Owen [23], an empirical log-likelihood ratio for (σ^2, β) is defined as

$$\ell(\sigma^2,\beta) = -\max\left\{\sum_{i=1}^n \log(np_i) : \ p_i \ge 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i Z_i(\sigma^2,\beta) = 0\right\}.$$

By the Lagrange multiplier method, we can obtain

$$p_i = \frac{1}{n}\frac{1}{1 + \lambda^T Z_i(\sigma^2,\beta)},$$

and $\ell(\sigma^2, \beta)$ can be expressed as

$$\ell(\sigma^2,\beta) = \sum_{i=1}^n \log\{1 + \lambda^T Z_i(\sigma^2,\beta)\},$$

where $\lambda = \lambda(\sigma^2, \beta)$ staisfies

$$\sum_{i=1}^{n} \frac{1}{n} \frac{Z_i(\sigma^2, \beta)}{1 + \lambda^T Z_i(\sigma^2, \beta)} = 0. \tag{2}$$

Let the non-penalized least-squares estimator $\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^{\mathrm{T}}$ be the initial estimator, which satisfies $\|\tilde{\beta} - \beta_0\| = O_p((p/n)^{\frac{1}{2}})$ (see Wang et al. [34]). Zou [39] suggested that the adaptive weights are given by $\tilde{w}_j = |\tilde{\beta}_j|^{-1}, j = 1, \ldots, p$. The penalized empirical likelihood estimator $(\hat{\sigma}^2, \hat{\beta})$ with adaptive lasso is defined as the minimizer of

$$\ell_p(\sigma^2, \beta) = \sum_{i=1}^{n} \log\{1 + \lambda^T Z_i(\sigma^2, \beta)\} + n\tau \sum_{j=1}^{p} \tilde{w}_j |\beta_j|, \tag{3}$$

where τ is a tuning parameter.

Let $\mathcal{A} = \{j : \beta_{0j} \neq 0\}$ and denote the cardinality of $\mathcal{A}$ as $|\mathcal{A}| = d$ which is unknown. Here we allow d to grow when $n \to \infty$. Without loss of generality, denote $\theta = (\theta_1^T, \theta_2^T)^T$, where $\theta_1 = (\sigma^2, \beta_1^T)^T \in \mathbb{R}^{d+1}$, $\theta_2 = \beta_2 \in \mathbb{R}^{p-d}$ correspond to the nonzero and zero components. Suppose that θ_0 is the true value of θ, thus $\theta_0 = (\sigma_0^2, \beta_{10}^T, 0^T)^T$. Similarly, denote $\hat{\theta} = (\hat{\theta}_1^T, \hat{\theta}_2^T)^T$, and $\hat{\theta}_1$, $\hat{\theta}_2$ are the PEL estimators of θ_1, θ_2 respectively. Let $\Omega = \mathrm{E}(\partial Z_i(\theta_0)/\partial\theta^T), \Sigma = \mathrm{E}(Z_i(\theta_0)Z_i^T(\theta_0))$ and $V^{-1} = \Omega^T\Sigma^{-1}\Omega$ where $Z_i(\theta) \in \mathbb{R}^{(2p+1)}, \Sigma \in \mathbb{R}^{(2p+1)\times(2p+1)}, \Omega \in \mathbb{R}^{(2p+1)\times(p+1)}$. Correspondingly we decompose V as a block matrix consisting $V_{ij} (i = 1, 2; j = 1, 2)$, where $V_{11} \in \mathbb{R}^{(d+1)\times(d+1)}$. We need the following assumptions.

- **(C1)** Let $\epsilon_i = Y_i - G(X_i^T\beta), \{\epsilon_i\}_{i=1}^n$ are i.i.d random variables independent of X_i. For $\alpha > 4$, $\mathrm{E}\|X_i\|^{3\alpha} < +\infty, \mathrm{E}\|\epsilon_i\|^{2\alpha} < +\infty$, and $\mathrm{E}\|\partial w_i/\partial\theta^T\|^{\alpha} < +\infty$.
- **(C2)** The eigenvalues of Σ satisfy $0 < C_1 \leq \gamma_1(\Sigma) \leq \cdots \leq \gamma_{p+r+1}(\Sigma) \leq C_2$, for some $C_2 > C_1 > 0$.
- **(C3)** $G(\cdot)$ is three times continuously differentiable and $V(\cdot)$ is twice continuously differentiable.
- **(C4)** As $n \to \infty, p \to \infty$ and $p^2/n \to 0$.
- **(C5)** There exists a positive constant M such that $\min_{j\in\mathcal{A}} |\beta_{0j}| \geq M$.
- **(C6)** As $n \to \infty$, the tuning parameter τ satisfies $\sqrt{nd}\tau \to 0$.

Remark 2.1. Conditions (C1)–(C3) ensure the existence and consistency of the PEL estimator. Both p and d allow to diverge as long as condition (C4) is satisfied. In particular, $p = o(n^{1/2})$ improves the conditions of [19] and [29]. It was pointed out by [7] that $p = o(n^{1/2})$ is likely the best rate for p such that the $\log EL$ ratio is asymptotically normal. Condition (C5) assures that all important coefficients are included in the final model.

Remark 2.2. It is worth noting that we require $\tau = O(^{-1/2})$ to get the consistency and sparsity of the adaptive Lasso estimator. However, condition (C6) with smaller τ is needed to obtain the asymptotic normality of the estimator, because larger τ can complete the variable selection but produce large biases. For convenience, we all use the condition (C6) to get our results.

First we show the existence, the consistency and the rate of convergence of PEL.

Theorem 2.1. *Under conditions* (C1)–(C6), *as* $n \to \infty$ *and with probability tending to* 1, $\ell_p(\theta)$ *in* (3) *has a minimum* $\hat{\theta}$ *such that* $\|\hat{\theta} - \theta_0\| = O_p((p/n)^{1/2})$.

We now present the following oracle property of PEL.

Theorem 2.2. *Let* $\hat{\theta} = (\hat{\theta}_1^T, \hat{\theta}_2^T)^T$ *be the minimizer of* (3). *Under conditions* (C1)–(C6), *as* $n \to \infty$, *we have*

(i) (Sparsity) *with probability tending to* 1, $\hat{\theta}_2 = 0$;

(ii) (Asymptotic normality) $n^{1/2}W_nB^{-1/2}(\hat{\theta}_1 - \theta_{10})^T \to N(0, G)$ *in distribution where* $W_n \in \mathbb{R}^{q\times(d+1)}$ *such that* $W_nW_n^T \to G$ *for* $G \in \mathbb{R}^{q\times q}$ *with fixed* q *and* $B = V_{11} - V_{12}V_{22}^{-1}V_{21}$.

Remark 2.3. Theorem 2.2 shows the oracle property of PEL estimator in the sense of [11]. That is, PEL estimator is consistent in model selection and is as efficient as the EL estimate assuming the true spare model was known.

Next we consider testing statistical hypothesis and constructing confidence regions for θ. Consider the null hypothesis $H_0 : L_n\theta_0 = 0$ versus $H_1 : L_n\theta_0 \neq 0$, where $L_n \in \mathbb{R}^{q\times(p+1)}$ is a matrix such that $L_nL_n^T = I_q$ for a fixed q and I_q is the q-dimensional identity matrix. This type of hypotheses covers linear functions of θ and includes individual and multiple components as special cases. Given the penalized empirical likelihood formulation, a nonparametric profiled likelihood ratio statistic is constructed as

$$\tilde{\ell}(L_n) = -2\left\{\ell_p(\hat{\theta}) - \min_{\theta: L_n\theta=0} \ell_p(\theta)\right\}.$$

In the following theorem, we will show a key property of the penalized empirical likelihood ratio.

Theorem 2.3. *Under the null hypothesis and conditions* (C1)–(C5), *as* $n \to \infty$, $\tilde{\ell}(L_n) \to \chi_q^2$ *in distribution.*

Therefore, a $(1-\alpha)$-level confidence region for $L_n\theta$ can be constructed as

$$V_\alpha = \left[\upsilon : -2\left\{\ell_p(\theta) - \min_{\theta: L_n\theta=\upsilon} \ell_p(\theta)\right\} \leq \chi_{q,1-\alpha}^2\right] \tag{4}$$

where $\chi_{q,1-\alpha}^2$ is the $(1-\alpha)$-level quantile of χ_q^2 distribution. As a direct result of Theorem 2.3, we have that $\mathrm{P}(L_n\theta_0 \in V_\alpha) \to 1 - \alpha$ as $n \to \infty$.

3. COMPUTATION

In this paper, we use an iterative nonlinear optimization algorithm together with the local quadratic approximation proposed by [11] to obtain the PEL estimator. Also, we give the choice of tuning parameter.

3.1 Algorithm of the PEL

Let $p_\tau(\beta_i) = \tau|\beta_i|/|\tilde{\beta}_i|$, we approximate $p_\tau(\beta_i)$ by

$$p_\tau(|\beta_i^{(k)}|) + \frac{1}{2}\{p'_\tau(|\beta_i^{(k)}|)/|\beta_i^{(k)}|\}(\beta_i^2 - \beta_i^{(k)2}),$$

where $\beta_i^{(k)}$ is the kth step estimate of β_i. Thus the minimization problem can be reduced to a quadratic minimization problem and the Newton-Raphson algorithm can be used, and yield the solution

$$\theta^{(k+1)} = \theta^{(k)} - \{\nabla^2\ell(\theta^{(k)}) + n\Sigma_\tau(\beta^{(k)})\}^{-1}\{\nabla\ell(\theta^{(k)}) + nU_\tau(\beta^{(k)})\},$$

where $U_\tau(\beta^{(k)}) = \Sigma_\tau(\beta^{(k)})\beta^{(k)}$,

$$\nabla\ell(\theta^{(k)}) = \frac{\partial\ell(\theta^{(k)})}{\partial\theta}, \quad \nabla^2\ell(\theta^{(k)}) = \frac{\partial^2\ell(\theta^{(k)})}{\partial\theta\partial\theta^T}.$$

Similarly $\theta^{(k)}$ is the kth step estimate of θ. When $\max|\theta^{(k+1)} - \theta^{(k)}| \le 10^{-3}$, we say the algorithm converges. We iterate between solving λ and θ, and [25] gave the method to solve the equation (2).

3.2 Selection of the tuning parameter

It is very critical to choose a proper tuning parameter τ since it determines the sparsity of the selected model. An optimal tuning parameter can result in a parsimonious model with good prediction performance. [33] showed that Bayesian information criterion (BIC) is consistent in model selection. So we employ the BIC-type criterion to choose the tuning parameter. For a given τ, we can obtain an estimate $\hat{\theta}_\tau$. The BIC-type criterion is defined by

$$\mathrm{BIC}_\tau = 2\ell_p(\hat{\theta}_\tau) + C_n\log(n)df_\tau,$$

where df_τ is the number of nonzero coefficients in $\hat{\theta}_\tau$ and C_n is the scaling factor diverging to infinity at a slow rate for $p \to \infty$. Here, we choose $C_n = \log\log p$.

4. SIMULATION STUDIES

In this section, we conduct simulation studies to evaluate the performance of penalized empirical likelihood. Different dimensionality p and sample size n are adopted in each simulation, and each designed case are repeated 1000 times. During the iteration, we follow the strategy in [11] of setting $\theta_i^{(k)}$ as zero when the i-th component of $\theta^{(k)}$ is very close to zero.

Firstly, we compare the performance of oracle empirical likelihood (OEL) given by the empirical likelihood estimates knowing the true sparsity of the model beforehand, empirical likelihood (EL), and penalized empirical likelihood (PEL) under different estimating equation, i.e. $Z_i^{(2)}$, $(Z_i^{(1)}, Z_i^{(2)})$, and $(Z_i^{(1)}, Z_i^{(2)}, Z_i^{(3)})$, the number of which are p, $p+1$ and $2p+1$ respectively.

For each simulation replication, we calculate L_2 distance: $\|\hat{\theta}-\theta_0\|_2 = \{(\hat{\theta}-\theta_0)^T(\hat{\theta}-\theta_0)\}^{1/2}$, $\|\hat{\beta}-\beta_0\|_2 = \{(\hat{\beta}-\beta_0)^T(\hat{\beta}-\beta_0)\}^{1/2}$ and $\|\hat{\sigma}^2-\sigma_0^2\|_2 = |\hat{\sigma}^2-\sigma_0^2|$. Reporting medians of the L_2 distance for all approaches. The L_2 distance illustrate the estimation accuracy. The average number of zero coefficients correctly set to zero (T) and average number of nonzero coefficients incorrectly set to zero (F) explain the variable selection performance.

Consider the following generalized linear models

$$Y_i = G(X_i^T\beta) + \sigma V^{1/2}\{G(X_i^T\beta)\}\epsilon_i.$$

Example 1. In this case, we choose $G(t) = t^2, V(t) = t$, and $\varepsilon_i \sim U[-\sqrt{3},\sqrt{3}]$. The parameter values are $\beta = (3, 1.5, 0, 0, 2, 0, \ldots)^T$ and $\sigma = 1.2$. $X_{ij} \sim U[0,1], i = 1\ldots n, j = 1\ldots p$, where X_{ij} is the j-th element of X_i.

The medians of L_2 distance for $\hat{\theta}$, $\hat{\beta}$, $\hat{\sigma}^2$ and the model selection performances are reported in Table 1. Form Table 1, we conclude that the estimation accuracy of OEL, EL and PEL increase as the number of estimating equation increase. The PEL estimators have smaller L_2 distance than EL, and the performance of the PEL is close to relevant oracle empirical likelihood, especially for large sample sizes no matter what kinds of estimating equation version used. As n increases, the average number of zero components correctly is close to $p-3$. Thus the model selection result is satisfying in three kinds of PEL. This confirms the results of Theorem 2.2. Also, we can conclude that the results of zero coefficients correctly are better with more estimating equations.

Example 2. Setting $G(t) = e^t$, $V(t) = t^2$, $\epsilon_i \sim U[-\sqrt{3},\sqrt{3}]$, $\beta = (1, 2, \ldots)^T$, $X_{ij} \sim U[0,2]$ and $\sigma = 0.5$.

The results are presented in Table 2. Form Table 2, we can get the same information about different methods and different estimating equations in terms of estimating error as Example 1. All the PEL methods have the selectivity about zero.

Example 3. In this example, we consider the binary X and set $P(X_{ij} = 1) = q, P(X_{ij} = 0) = 1-q$ with $q = 0.2, 0.5, 0.8$. The other settings are the same as those in Example 2.

Since overall pattern for the estimation accuracy and the model selection performances of the three methods with different estimating equations is similar, to save space, Table 3 depicts the simulated results of OEL, EL and PEL with three estimating equations only. From Table 3, we can see that the PEL performs satisfactorily in terms of estimation accuracy and variable selection.

Finally, the performance of penalized empirical likelihood confidence region is also evaluated in the following example.

Example 4. Following the situation: $G(t) = \sin(t), V(t) = t^2, \epsilon_i \sim N(0,1), \beta = (1, 1, 0, 0, 1, 0, \ldots)^T$, and $\sigma = 0.5$. X_i have the same distribution as Example 1.

Table 1. Medians of L_2 distances ($\times 10^{-1}$) and model selection performances for Example 1.

n	p	L_2	OEL_1	EL_1	PEL_1	OEL_2	EL_2	PEL_2	OEL_3	EL_3	PEL_3
100	10	D	-	-	-	0.343	1.385	0.530	0.338	1.327	0.377
		D_1	0.169	1.083	0.244	0.166	1.046	0.215	0.157	1.061	0.174
		D_2	-	-	-	0.093	0.232	0.211	0.089	0.151	0.126
		T	-	-	6.801	-	-	6.845	-	-	6.896
		F	-	-	0	-	-	0	-	-	0
200	15	D	-	-	-	0.165	0.964	0.257	0.157	0.912	0.180
		D_1	0.081	0.795	0.103	0.081	0.780	0.101	0.077	0.777	0.082
		D_2	-	-	-	0.041	0.114	0.101	0.040	0.058	0.056
		T	-	-	11.906	-	-	11.906	-	-	11.948
		F	-	-	0	-	-	0	-	-	0
400	20	D	-	-	-	0.079	0.600	0.117	0.070	0.562	0.076
		D_1	0.040	0.529	0.049	0.039	0.520	0.048	0.036	0.509	0.036
		D_2	-	-	-	0.019	0.049	0.045	0.017	0.023	0.021
		T	-	-	16.926	-	-	16.939	-	-	16.979
		F	-	-	0	-	-	0	-	-	0

The small Angle mark of estimators 1, 2, 3 represent the estimating equations $Z_i^{(2)}, (Z_i^{(1)}, Z_i^{(2)}), (Z_i^{(1)}, Z_i^{(2)}, Z_i^{(3)})$; D, D_1 and D_2 represent the L_2 error of θ, β, and σ^2.

Table 2. Medians of L_2 distances ($\times 10^{-1}$) and model selection performances for Example 2.

n	p	L_2	OEL_1	EL_1	PEL_1	OEL_2	EL_2	PEL_2	OEL_3	EL_3	PEL_3
100	10	D	-	-	-	0.060	0.768	0.082	0.038	0.571	0.055
		D_1	0.049	0.756	0.058	0.049	0.753	0.061	0.0275	0.551	0.036
		T	-	-	7.505	-	-	7.691	-	-	7.836
		F	-	-	0	-	-	0	-	-	0
200	15	D	-	-	-	0.030	0.597	0.042	0.018	0.364	0.025
		D_1	0.025	0.579	0.028	0.025	0.591	0.031	0.014	0.357	0.016
		T	-	-	12.564	-	-	12.703	-	-	12.899
		F	-	-	0	-	-	0	-	-	0
400	20	D	-	-	-	0.014	0.375	0.019	0.009	0.206	0.011
		D_1	0.013	0.384	0.014	0.012	0.372	0.014	0.007	0.203	0.007
		T	-	-	17.716	-	-	17.871	-	-	17.951
		F	-	-	0	-	-	0	-	-	0

Table 3. Medians of L_2 distances ($\times 10^{-1}$) and model selection performances for Example 3.

			$q = 0.2$			$q = 0.5$			$q = 0.8$		
n	p	L_2	OEL_3	EL_3	PEL_3	OEL_3	EL_3	PEL_3	OEL_3	EL_3	PEL_3
100	10	D	0.099	1.789	0.177	0.056	0.776	0.098	0.069	1.178	0.103
		D_1	0.096	1.785	0.158	0.048	0.773	0.073	0.057	1.147	0.074
		T	-	-	7.22	-	-	7.88	-	-	7.96
		F	-	-	0	-	-	0	-	-	0
200	15	D	0.059	0.768	0.077	0.018	0.519	0.033	0.026	0.735	0.039
		D_1	0.058	0.761	0.075	0.014	0.518	0.025	0.022	0.729	0.029
		T	-	-	12.7	-	-	12.89	-	-	12.93
		F	-	-	0	-	-	0	-	-	0
400	20	D	0.025	0.463	0.032	0.013	0.271	0.018	0.018	0.429	0.017
		D_1	0.023	0.461	0.029	0.011	0.271	0.016	0.013	0.422	0.013
		T	-	-	17.84	-	-	17.91	-	-	17.95
		F	-	-	0	-	-	0	-	-	0

Table 4. The empirical frequency (%) that a given value of θ_2 does not fall in the 95% confidence interval constructed by (4).

p	n	0.8	0.9	1.0(size)	1.1	1.2
6	50	18.6	15	8	15.9	23.7
10	100	40.5	27.3	6.7	31.7	40.1
15	200	76.4	61.6	6.2	67.3	83
20	400	93.7	86	5.9	91.4	96.4
26	800	99.3	95.8	6	97.1	99.3

Setting $L_n = (0, 1, 0, \ldots, 0)$ in (4) leads to a confidence set for θ_2, the second component of θ, at the $(1-\alpha)$ level. For the nominal level $\alpha = 0.05$, we report the empirical frequencies of $\theta_2 \notin V_\alpha$ for a sequence of θ_2 values in Table 4. At the true value $\theta_2 = 1$, the empirical frequency in rejecting the null hypothesis is close to the nominal level $\alpha = 0.05$, which maintain the size and confirms the result of Theorem 2.3. When the discrepancy between θ_2 and the the true value is larger, the rejection frequency increases. Particularly, when n is large and the difference between θ_2 and true value is 0.2, the rejection rate is close to 1. This shows that the proposed penalized empirical likelihood has a good power for testing the null hypothesis.

5. REAL APPLICATION

To illustrate the usefulness of proposed penalized empirical likelihood, we consider the data from the National Medical Expenditure Survey (NMES) which was conducted in 1987 and 1988 to provide a comprehensive of how Americans use and pay for health services, using a subset of individuals ages 66 and over (a total of 4406 observations). The same data set has been used to model the demand for medical care, e.g., [8] and a subset of this dataset has been analyzed in [38].

The major objective is to illustrate the number of physician office visits (OFP) while accounting for the following predictor variables: EXCLHLTH (=1 if self-perceived health is excellent), POORHLTH (=1 if self-perceived health is poor), NUMCHRON (of chronic conditions), ADLDIFF (=1 if the person has a condition that limits activities of daily living), NOREAST (=1 if the person lives in northeastern US), MIDWEST (=1 if the person lives in midwestern US), WEST(=1 if the person lives in western US), AGE (age in years divided by 10), BLACK (=1 if the person is African American), MALE (=1 if person is male), MARRIED (=1 if the person is married), SCHOOL (of years of education), FAMINC (family income in 10000), EMPLOYED (=1 if the person is employed), PRIVINS (=1 if the person is covered by private health insurance) and MEDICAID (=1 if the person is covered by Medicaid). Mean of physician office visits is 5.77 with variance 45.69. The raw data display with a high degree of overdispersed. The OFP variable also have excess number of zero, frequency histogram of it is present in Figure 1. As noted in [38], consider the following model:

$$E(Y|X) = G(X^T\beta) \quad \text{and} \quad \text{Var}(Y|X) = \sigma^2 G(X^T\beta),$$

where $G(t) = e^t$. We calculate the quasi-Poisson likelihood estimator, the best subset estimator comparing with the proposed PEL. The results are presented in Table 5. The standard errors of the proposed PEL should divided by ten. From Table 5, we can see that the PEL estimator achieves the simplest model, and this technique is close to the best subset.

Both the number of chronic conditions and self-perceived health are important determinants of utilization. Individuals with greater schooling seek care in office settings more often that may due to the reason that education makes individuals more informed consumers of medical care services. Person with supplementary private insurance seek care from physicians more often. That is because private insurance typically covers physical therapy, check ups, etc., with small deductibles and coinsurance rates while Medicare does not. Medical coverage is a significant determinant of the number of visit, that [8] give a explanation of price. Moreover, age and the region of residence appears to have some influence. The impact of a number of conditions that lead to disability can be ameliorated or postponed through therapy or rehabilitation. Men seek less care in office settings. An explanation for this phenomenon is the anecdotal fact that men tend to wait longer before seeking medical care. While, the other variables don't make contribution to the number of physician office visits.

APPENDIX A. PROOFS OF THEOREMS

We begin by collecting technical lemmas that will be used in the proof of the main theorems.

Let $a_n = (p/n)^{1/2}$, and let $D_n = \{\theta : \|\theta - \theta_0\| \le ca_n\}$ be a neighbourhood of θ_0 for some constant $c > 0$. In Lemma 1, we show that $\|\lambda_\theta\| = O_p(a_n)$ where θ is in a large enough neighbourhood of θ_0, assuring the asymptotic expansion of $\ell(\theta)$. For notational purposes, we define $I_{p+1} = (H_1^T, H_2^T)$ where $H_1 \in \mathbb{R}^{(d+1)\times(p+1)}$ and $H_2 \in \mathbb{R}^{(p-d)\times(p+1)}$. Throughout the proof, we use the Frobenius norm of a matrix A, that is $\|A\| = \{\text{tr}(A^TA)\}^{1/2}$.

Table 5. The fitted coefficients and standard errors in NMES data.

Variable	Likelihood	PEL	Best Subset
OFP	-	-	-
INTERCEPT	1.296(0.228)***	1.277(0.013)	1.229(0.082)
EXCLHLTH	-0.386(0.079)***	-0.388(0.010)	-0.378(0.030)
POORHLTH	0.287(0.047)***	0.284(0.012)	0.286(0.018)
NUMCHRON	0.163(0.011)***	0.184(0.002)	0.164(0.004)
ADLDIFF	0.093(0.043)*	0.081(0.009)	0.094(0.016)
NOREAST	0.107(0.046)*	0.110(0.009)	0.113(0.016)
MIDWEST	-0.010(0.043)	0	0
WEST	0.123(0.047)**	0.121(0.007)	0.129(0.016)
AGE	-0.055(0.028)	-0.078(0.003)	-0.051(0.010)
BLACK	-0.065(0.058)	0	0
MALE	-0.071(0.037)	-0.065(0.009)	-0.084(0.013)
MARRIED	-0.040(0.038)	0	0
SCHOOL	0.025(0.005)***	0.047(0.001)	0.026(0.001)
FAMINC	-0.002(0.006)	0	0
EMPLOYED	0.052(0.057)	0	0
PRIVINS	0.321(0.051)***	0.317(0.001)	0.325(0.019)
MEDICAID	0.288(0.065)***	0.276(0.003)	0.286(0.024)
σ^2	6.842	6.839	

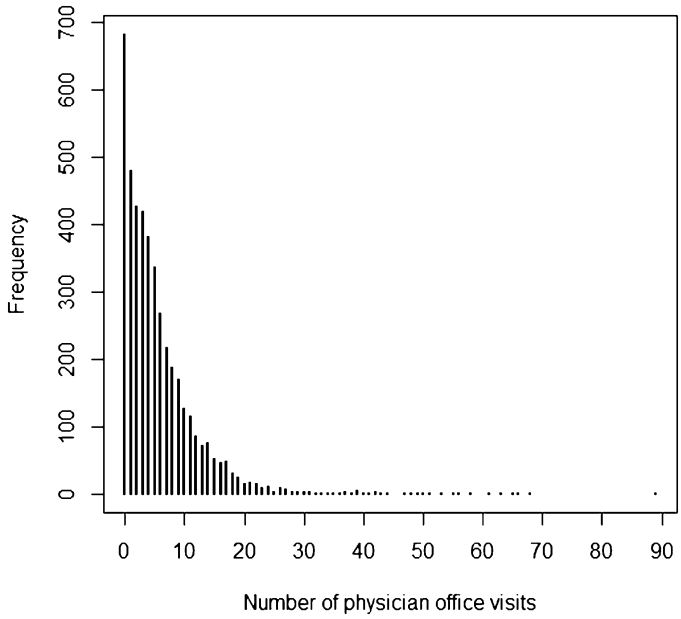

Figure 1. The frequency of physician office visits.

Lemma 1. *If Conditions* (C1)–(C4) *hold, then* $\|\lambda_\theta\| = O_p(a_n)$ *for* $\theta \in D_n$.

Proof. For $\theta \in D_n, \lambda = \rho u$ where $\|u\| = 1$ is a unit vector. Following [25], we have

$$\rho\left\{u^T S_\theta u - \max_i \|Z_i(\theta)\| \frac{1}{n}\left|\sum_{i=1}^n u^T Z_i(\theta)\right|\right\} \le \frac{1}{n}\left|\sum_{i=1}^n u^T Z_i(\theta)\right|,$$

where $S_\theta = n^{-1}\sum_{i=1}^n Z_i(\theta)Z_i^T(\theta)$. By the fact,

$$\left\|\frac{1}{n}\sum_{i=1}^n Z_i(\theta_0)\right\| = O_p((p/n)^{1/2}),$$

we have

$$\frac{1}{n}\left|\sum_{i=1}^n u^T Z_i(\theta_0)\right| = O_p((p/n)^{1/2}).$$

As

$$S_{\theta_0} = \frac{1}{n}\sum_{i=1}^n Z_i(\theta_0)Z_i^T(\theta_0) \xrightarrow{p} \Sigma,$$

so $u^T S_{\theta_0} u = O_p(1)$. Because,

$$\max_i \|Z_i(\theta_0)\| \le \max_i \|Z_i^{(1)}(\theta_0)\| + \max_i \|Z_i^{(2)}(\theta_0)\| + \max_i \|Z_i^{(3)}(\theta_0)\|,$$

under condition (C1), we have $\max_i \|Z_i(\theta_0)\| = O_p(n^{1/\alpha})$. Therefore

$$\max_i \|Z_i(\theta_0)\| \frac{1}{n}\left|\sum_{i=1}^n u^T Z_i(\theta_0)\right| = O_p(n^{1/\alpha})O_p((p/n)^{1/2}) = o_p(1),$$

and we conclude that $\|\lambda_{\theta_0}\| = O_p((p/n)^{1/2})$. By Taylor expansion, we have

$$\frac{1}{n}\left|\sum_{i=1}^n u^T Z_i(\theta)\right|$$

$$= \frac{1}{n}\left|\sum_{i=1}^{n} u^T Z_i(\theta_0) + \frac{\partial Z_i(\theta_0)}{\partial \theta}(\theta-\theta_0) + o_p(1)\right|$$
$$\le \frac{1}{n}\left|\sum_{i=1}^{n} u^T Z_i(\theta_0)\right| + \max_i \left\|\frac{\partial Z_i(\theta_0)}{\partial \theta}(\theta-\theta_0)\right\| + o_p(1).$$

From condition (C1), we have

$$\max_i \|\partial Z_i(\theta_0)/\partial\theta^T(\theta-\theta_0)\| = o_p(1),$$

then

$$\frac{1}{n}\left|\sum_{i=1}^{n} u^T Z_i(\theta)\right| = O_p((p/n)^{1/2}).$$

This together with

$$\max_i \|Z_i(\theta)\| \le \max_i \|Z_i(\theta_0)\| + \max_i \|\partial Z_i(\theta_0)/\partial\theta^T(\theta-\theta_0)\|,$$

we have

$$\max_i \|Z_i(\theta)\| \frac{1}{n}\left|\sum_{i=1}^{n} u^T Z_i(\theta)\right| = o_p(1).$$

Since

$$u^T S_\theta u = u^T S_{\theta_0} u + o_p(1) = O_p(1),$$

Lemma 1 follows. □

Proof of Theorem 2.1. It suffices to show that for any given $\epsilon > 0$, there exists a large constant C such that

$$\mathrm{P}\left\{\inf_{\theta\in\partial D_n} \ell_p(\theta) > \ell_p(\theta_0)\right\} \ge 1-\epsilon, \tag{5}$$

where ∂D_n is the boundary of D_n. This implies that, with probability at least $1-\epsilon$, there exists a local minimum in the ball $\{\theta_0 + a_n u : \|u\| \le C\}$. That is to say, there exists a local minimizer $\hat\theta$ of $\ell_p(\theta)$ such that $\|\hat\theta - \theta_0\| = O_p(a_n)$.

For $\theta \in D_n$, by the definition of EL,

$$Q_{1n}(\theta,\lambda) = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i(\theta)}{[1+\lambda^T Z_i(\theta)]} = 0.$$

From Lemma 1, we have that $\lambda^T Z_i(\theta)$ and $\|\lambda\|$ are stochastically small for $\theta \in D_n$. Applying Taylor expansion on $Q_{1n}(\theta,\lambda)$, we have

$$0 = \bar Z(\theta) - S_\theta \lambda + r_n,$$

where $\bar Z(\theta) = \frac{1}{n}\sum_{i=1}^{n} Z_i(\theta)$ and

$$r_n = \frac{1}{n}\sum_{i=1}^{n} Z_i(\theta)[\lambda^T Z_i(\theta)]^2(1+\xi_i)^{-3}$$

is the reminder term and $|\xi_i| \le |\lambda^T Z_i(\theta)|$. Inverting the expansion, we have $\lambda = S_\theta^{-1}[\bar Z(\theta) + r_n]$. Applying Taylor's expansion on $\ell(\theta)$, we have

$$2\ell(\theta,\lambda) = 2\sum_{i=1}^{n} \lambda^T Z_i(\theta) - \sum_{i=1}^{n}[\lambda^T Z_i(\theta)]^2 + \frac{2}{3}\sum_{i=1}^{n}\frac{[\lambda^T Z_i(\theta)]^3}{(1+\xi_i)^3},$$

where $|\xi_i| \le |\lambda^T Z_i(\theta)|$. Substituting λ into $\ell(\theta)$, we obtain that

$$2\ell(\theta) = n\bar Z(\theta)^T S_\theta^{-1}\bar Z(\theta) - n r_n^T S_\theta^{-1} r_n + \frac{2}{3}\sum_{i=1}^{n}\frac{[\lambda^T Z_i(\theta)]^3}{(1+\xi_i)^3}.$$

We have a decomposition as

$$2\ell(\theta) = T_0 + T_1 + T_2,$$

where

$$T_0 = n\bar Z(\theta_0)^T S_{\theta_0}^{-1}\bar Z(\theta_0),$$
$$T_1 = n\{\bar Z(\theta) - \bar Z(\theta_0)\}^T S_\theta^{-1}\{\bar Z(\theta) - \bar Z(\theta_0)\}$$

and

$$\begin{aligned} T_2 =& n\{\bar Z(\theta_0)^T[S_\theta^{-1} - S_{\theta_0}^{-1}]\bar Z(\theta_0) \\ &+ 2\bar Z(\theta_0)^T S_\theta^{-1}[\bar Z(\theta) - \bar Z(\theta_0)]\} - n r_n^T S_\theta^{-1} r_n \\ &+ \frac{2}{3}\sum_{i=1}^{n}[\lambda^T Z_i(\theta)]^3(1+\xi_i)^{-3}. \end{aligned}$$

By Taylor expansion, the definition of Ω and condition (C2), we have

$$\begin{aligned} T_1 &> n\|\bar Z(\theta) - \bar Z(\theta_0)\|^2 r_{p+r+1}^{-1}(\Sigma) \\ &= nO_p(\|\Omega\|^2\|\theta-\theta_0\|^2) \\ &= na_n^2\|u\|^2 O_p(1). \end{aligned}$$

As $T_2/T_1 \to 0$ and $2\ell(\theta_0) - T_0 \to 0$, $2\ell(\theta) - 2\ell(\theta_0)$ is dominated by T_1.

It can be easily seen that

$$\begin{aligned} &\ell_p(\theta) - \ell_p(\theta_0) \\ &\quad= \ell(\theta) - \ell(\theta_0) + n\tau\sum_{j=1}^{p}\tilde w_j(|\beta_{0j} + a_n u_j| - |\beta_{0j}|) \\ &\quad\ge \ell(\theta) - \ell(\theta_0) - n\tau a_n \sum_{j\in\mathcal{A}} \tilde w_j |u_j| \end{aligned}$$

Note that

$$\sum_{j\in\mathcal{A}} \tilde w_j^2 \le d\cdot\Big(\min_{j\in\mathcal{A}}|\tilde\beta_j|\Big)^{-2} = d\left(\frac{\min_{j\in\mathcal{A}}|\tilde\beta_j|}{\min_{j\in\mathcal{A}}|\beta_{0j}|}\right)^{-2}\cdot\Big(\min_{j\in\mathcal{A}}|\beta_{0j}|\Big)^{-2}$$

By conditions (C4) and (C5), we get

$$\left|\frac{\min_{j\in\mathcal{A}}|\tilde{\beta}_j|}{\min_{j\in\mathcal{A}}|\beta_{0j}|}-1\right|\le\Big(\min_{j\in\mathcal{A}}|\beta_{0j}|\Big)^{-1}\|\tilde{\beta}-\beta_0\|=o_p(1).$$

That is

$$\min_{j\in\mathcal{A}}|\tilde{\beta}_j|/\min_{j\in\mathcal{A}}|\beta_{0j}|=1+o_p(1). \tag{6}$$

Thus, $\sum_{j\in\mathcal{A}}\tilde{w}_j^2=O_p(d)$. This, together with condition (C6) and the Cauchy-Schwarz inequality yields that

$$\begin{aligned}n\tau a_n\sum_{j\in\mathcal{A}}\tilde{w}_j|u_j|&\le n\tau a_n\|u\|\left(\sum_{j\in\mathcal{A}}\tilde{w}_j^2\right)^{1/2}\\&=nd^{1/2}\tau a_n\|u\|O_p(1)\\&=na_n^2\|u\|O_p(1)\end{aligned}$$

Hence, for a sufficient large constant C, the sign of $\ell_p(\theta)-\ell_p(\theta_0)$ is dominated by T_1 which is nonnegative. This complete the proof. □

Proof of Theorem 2.2. Firstly, in order to prove the part (i) of this Theorem, it is sufficient to show that

$$P\Big\{\min_{j\in\mathcal{A}}|\hat{\theta}_j|>0\Big\}\to1.$$

Note that $\hat{\theta}$ is a root-(n/p)-consistent estimator of θ_0, by conditions (C4) and (C5), we have

$$\min_{j\in\mathcal{A}}|\hat{\theta}_j|\ge\min_{j\in\mathcal{A}}|\theta_{0j}|-\|\hat{\theta}-\theta_0\|\ge M-o_p(1)$$

This ends the proof of part (i).

Next we prove the estimation efficiency part (ii). By the result of part (i) and the definition of PEL, the PEL estimator $\hat{\theta}$ is the constrained minimizer of (3) subject to $H_2\theta_0=0$. According to [27], by the Lagrange multiplier method, obtaining the estimates is equivalent to minimizing a new objective function

$$\tilde{\ell}(\theta,\lambda,\nu)=n^{-1}\sum_{i=1}^{n}\log\{1+\lambda^TZ_i(\theta)\}+\sum_{j=1}^{p}p_\tau(|\beta_j|)+\nu^TH_2\theta,$$

where $\nu\in\mathbb{R}^{p-d}$ is the vector of extra Lagrange multiplier.

Define

$$\tilde{Q}_{1n}(\theta,\lambda,\nu)=\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i(\theta)}{[1+\lambda^TZ_i(\theta)]},$$

$$\begin{aligned}\tilde{Q}_{2n}(\theta,\lambda,\nu)=&\left(\frac{1}{n}\left\{\sum_{i=1}^{n}\frac{[\partial Z_i(\theta)^T/\partial\sigma^2]\lambda}{[1+\lambda^TZ_i(\theta)]}\right\}^T,\right.\\&\left.\frac{1}{n}\left\{\sum_{i=1}^{n}\frac{[\partial Z_i(\theta)^T/\partial\beta]\lambda}{[1+\lambda^TZ_i(\theta)]}+b(\beta)\right\}^T\right)^T+H_2^T\nu,\end{aligned}$$

and

$$\tilde{Q}_{3n}(\theta,\lambda,\nu)=H_2\theta,$$

where $b(\beta)=(\tau\mathrm{sgn}(\beta_1)/|\tilde{\beta}_1|,\ldots,\tau\mathrm{sgn}(\beta_p)/|\tilde{\beta}_p|)^T$.

The minimizer $(\hat{\theta},\hat{\lambda},\hat{\nu})$ satisfies $\tilde{Q}_{jn}(\hat{\theta},\hat{\lambda},\hat{\nu})=0(j=1,2,3)$. It follows Lemma 1 that $\|\hat{\lambda}\|=O_p(a_n)$ is stochastically small. Therefore, similar to the argument of [27], from $\tilde{Q}_{2n}(\hat{\theta},\hat{\lambda},\hat{\nu})=0$ we conclude that $\|\hat{\nu}\|=O_p(a_n)$. Hence, we can use stochastic expansions of $\tilde{Q}_{jn}$ around the value $(\theta_0,0,0)$ $(j=1,2,3)$.

It is straightforward to verify that

$$\begin{aligned}&\begin{pmatrix}-\tilde{Q}_{1n}(\theta_0,0,0)\\0\\0\end{pmatrix}\\&=\begin{pmatrix}-\Sigma&\Omega&0\\\Omega^T&0&H_2^T\\0&H_2&0\end{pmatrix}\begin{pmatrix}\hat{\lambda}\\\hat{\theta}-\theta_0\\\hat{\nu}\end{pmatrix}\\&\quad+R_n^{(1)}+R_n^{(2)}+R_n^{(3)}+R_n^{(4)},\end{aligned} \tag{7}$$

where $R_n^{(1)}=(R_{1n}^{(1)T},R_{2n}^{(1)T},0)^T,R_{1n}^{(1)}\in\mathbb{R}^{r+p+1},R_{2n}^{(1)}\in\mathbb{R}^{p+1}$ and the k-th component of $R_{jn}^{(1)}$ $(j=1,2)$ is given by

$$R_{jn,k}^{(1)}=\frac{1}{2}(\hat{\eta}-\eta_0)^T\frac{\partial^2\tilde{Q}_{jn,k}(\eta^*)}{\partial\eta\partial\eta^T}(\hat{\eta}-\eta_0),$$

$\eta=(\theta^T,\lambda^T)^T$, $\eta^*=(\theta^{*T},\lambda^{*T})^T$ satisfying $\|\theta^*-\theta_0\|\le\|\hat{\theta}-\theta_0\|$ and $\|\lambda^*\|\le\|\hat{\lambda}\|$, so $\|R_n^{(1)}\|=o_p(n^{-\frac{1}{2}})$. The other three terms are given by $R_n^{(2)}=\{0,\tilde{b}(\beta_0)^T,0\}^T$ with $\tilde{b}(\beta_0)=(0,b(\beta_0)^T)^T\in\mathbb{R}^{p+1}$,

$$\begin{aligned}R_n^{(3)}=&\left\{\left\{\left[\Sigma-\frac{1}{n}\sum_{i=1}^{n}Z_i(\theta_0)Z_i^T(\theta_0)\right]\hat{\lambda}\right\}^T\right.\\&\left.+\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\partial Z_i(\theta_0)/\partial\theta^T-\Omega\right)(\hat{\theta}-\theta_0)\right\}^T,0,0\right\}^T\end{aligned}$$

and

$$R_n^{(4)}=\left\{0,\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\partial Z_i(\theta_0)^T/\partial\theta-\Omega^T\right)\hat{\lambda}\right\}^T,0\right\}^T.$$

Note that the assertion (6), by conditions (C5) and (C6), it can be concluded that

$$\|R_n^{(2)}\|=\|\tilde{b}(\beta_0)\|=O_p(\sqrt{d}\tau)=o_p(n^{-1/2})$$

From Lemma 1, we can establish that $\|R_n^{(3)}\|=\|R_n^{(4)}\|=o_p(n^{-1/2})$. Therefore we get $\|R_n\|=o_p(n^{-1/2})$.

Define $K_{11}=-\Sigma$, $K_{12}=(\Omega,0)$ and $K_{21}=K_{12}^T$, where

$$K_{22}=\begin{pmatrix}0&H_2^T\\H_2&0\end{pmatrix},\quad K=\begin{pmatrix}K_{11}&K_{12}\\K_{21}&K_{22}\end{pmatrix}.$$

Let $\vartheta = (\theta^T, \nu^T)^T$, by inverting (6.1), we have

$$\begin{pmatrix} \hat{\lambda} - 0 \\ \hat{\vartheta} - \vartheta_0 \end{pmatrix} = K^{-1} \left\{ \begin{pmatrix} -\tilde{Q}_{1n}(\theta_0, 0, 0) \\ 0 \\ 0 \end{pmatrix} + R_n \right\}.$$

Applying matrix inversion by blocks, we have

$$K^{-1} = \begin{pmatrix} K_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} -K_{11}^{-1}K_{12} \\ I \end{pmatrix} A^{-1} \begin{pmatrix} -K_{21}K_{11}^{-1} & I \end{pmatrix},$$

where $A = K_{22} - K_{21}K_{11}^{-1}K_{12}$. Then

$$\hat{\vartheta} - \vartheta_0 = A^{-1}K_{21}K_{11}^{-1}\tilde{Q}_{1n}(\theta_0, 0, 0) + o_p(n^{-1/2}).$$

Another matrix inversion by blocks gives

$$A^{-1} = \begin{pmatrix} V - VH_2^T(H_2VH_2^T)^{-1}H_2^TV & VH_2^T(H_2VH_2^T)^{-1} \\ (H_2VH_2^T)^{-1}H_2V & -(H_2VH_2^T)^{-1} \end{pmatrix}.$$

This implies that

$$\begin{aligned} &\hat{\theta} - \theta_0 \\ &= -\{V - VH_2^T(H_2VH_2^T)^{-1}H_2V\}\Omega^T\Sigma^{-1}\tilde{Q}_{1n}(\theta_0, 0, 0) + R_{1n}, \end{aligned}$$

where R_{1n} is the corresponding component in vector $K^{-1}R_n$ and $\|R_{1n}\| = o_p(n^{-1/2})$.

It is clear that the expansion of the nonzero component θ_1 is

$$\begin{aligned} \hat{\theta}_1 - \theta_{10} = &-\{H_1V - H_1VH_2^T(H_2VH_2^T)^{-1}H_2V\} \\ &\times \Omega^T\Sigma^{-1}\tilde{Q}_{1n}(\theta_0, 0, 0) + o_p(n^{-\frac{1}{2}}). \end{aligned}$$

Let $B = V_{11} - V_{12}V_{22}^{-1}V_{21}$, $Y_{ni} = n^{-1/2}Z_{ni}$ and

$$\begin{aligned} Z_{ni} = &-W_nB^{-1/2}\{H_1V - H_1VH_2^T(H_2VH_2^T)^{-1}H_2V\} \\ &\times \Omega^T\Sigma^{-1}Z_i(\theta_0). \end{aligned}$$

Next we verify the Lindeberg-Feller condition ([31]). It is easy to verify that $\mathrm{E}(Z_{ni}) = 0$, $\mathrm{Var}(Z_{ni}) = W_nW_n^T$, $E(\|Y_{ni}\|^4) = O(1/n^2)$ and

$$\mathrm{P}(\|Y_{ni}\| > \epsilon) \le \frac{1}{n\epsilon^2}\mathrm{E}(\|Z_{ni}\|^2) = \frac{\mathrm{trVar}(Z_{ni})}{n\epsilon^2} = O(1/n).$$

Hence,

$$\begin{aligned} &\sum_{i=1}^n E(\|Y_{ni}\|^2)I(\|Y_{ni}\| > \epsilon) \\ &\le n\{E(\|Y_{n1}\|^4)\}^{1/2}\{\mathrm{P}(\|Y_{n1}\| > \epsilon)\}^{1/2} \\ &\xrightarrow{p} 0. \end{aligned}$$

As

$$\sum_{i=1}^n \mathrm{Var}(Y_{ni}) = W_nW_n^T \xrightarrow{p} G,$$

we have

$$n^{1/2}W_nB^{-1/2}(\hat{\theta}_1 - \theta_{10}) \longrightarrow N(0, G)$$

in distribution. Finally, by noting that

$$\|n^{1/2}W_nB^{-1/2}R_{1n}\|^2 = o_p(1),$$

the proof of part (ii) is complete. □

Proof of Theorem 2.3. First, we present the asymptotic expansion of $\ell(\hat{\theta})$ where $\hat{\theta}$ is the minimizer of $\ell(\theta)$. Let $h_i = \hat{\lambda}^T Z_i(\hat{\theta})$, as $\max_i |\hat{\lambda}^T Z_i(\hat{\theta})| = o_p(1)$ implied by Lemma 1, by Taylor expansion, we have

$$\ell(\hat{\theta}) = \sum_{i=1}^n h_i - \sum_{i=1}^n \frac{h_i^2}{2} + \sum_{i=1}^n \frac{h_i^3}{\{3(1+\xi_i)^3\}},$$

where $|\xi_i| < |\hat{\lambda}^T Z_i(\hat{\theta})|$. In the proof of Theorem 2.1, we have shown the expansion for $\theta \in D_n$ to be $\lambda = S_\theta^{-1}[\bar{Z}(\theta) + r_n]$, where $r_n = n^{-1}\sum_{i=1}^n \{Z_i(\theta)[\lambda^T Z_i(\theta)]^2(1+\xi_i)^{-3}\}$ and $|\xi_i| \le |\lambda^T Z_i(\theta)|$. Substituting the expansion of $\hat{\lambda}$ into h_i, we show that

$$\begin{aligned} (8) \qquad 2\ell(\hat{\theta}) = &n\bar{Z}(\theta_0)^T\Sigma^{-1}\Omega VH_2^T(H_2VH_2^T)^{-1} \\ &\times H_2V\Omega^T\Sigma^{-1}\bar{Z}(\theta_0) + o_p(1). \end{aligned}$$

Under the null hypothesis, because $L_nL_n^T = I_q$, there exists $\tilde{H}$ such that $\tilde{H}_2\theta = 0$ and $\tilde{H}_2\tilde{H}_2^T = I_{p-d+q}$. Now by repeating the proof of Theorem 2.2, we establish that under the null hypothesis, the estimation of θ can be obtained by minimizing

$$\begin{aligned} (9) \qquad \tilde{\ell}_p(\theta, \lambda, \nu) = &\sum_{i=1}^n \log\{1 + \lambda^T Z_i(\theta)\} \\ &+ n\sum_{j=1}^p p_\tau(|\beta_j|) + \nu^T\tilde{H}_2\theta. \end{aligned}$$

Denote the minimizer of (9) by $(\breve{\theta}, \breve{\lambda}, \breve{\nu})$, from the proof of Part (i) in Theorem 2.2, $\breve{\theta}_2 = 0$ with probability tending to 1. Similar to (8), we establish that

$$\begin{aligned} (10) \qquad 2\ell(\theta)|_{L_n\theta=0} = &n\bar{Z}(\theta_0)^T\Sigma^{-1}\Omega V\tilde{H}_2^T(\tilde{H}_2V\tilde{H}_2^T)^{-1} \\ &\times \tilde{H}_2V\Omega^T\Sigma^{-1}\bar{Z}(\theta_0) + o_p(1). \end{aligned}$$

Combining equations (8) and (9), we obtain

$$\tilde{\ell}(L_n) = n\bar{Z}(\theta_0)^T\Sigma^{-1/2}\{P_1 - P_2\}\Sigma^{-1/2}\bar{Z}(\theta_0) + o_p(1),$$

where

$$P_1 = \Sigma^{-1/2}\Omega V\tilde{H}_2^T(\tilde{H}_2V\tilde{H}_2^T)^{-1}\tilde{H}_2V\Omega^T\Sigma^{-1/2},$$

and

$$P_2 = \Sigma^{-1/2}\Omega V H_2^T (H_2 V H_2^T)^{-1} H_2 V \Omega^T \Sigma^{-1/2}.$$

As $P_1 - P_2$ is an idempotent matrix of rank q, $P_1 - P_2$ can be written as $\Xi_n^T \Xi_n$, where Ξ_n is a $q \times (p + r + 1)$ matrix such that $\Xi_n \Xi_n^T = I_q$. Further, we see that

$$n^{1/2} \Xi_n \Sigma^{-1/2} \bar{Z}(\theta_0) \longrightarrow N(0, I_q)$$

in distribution can be easily established. Then

$$n \bar{Z}(\theta_0)^T \Sigma^{-1/2} (P_1 - P_2) \Sigma^{-1/2} \bar{Z}(\theta_0)^T \longrightarrow \chi_q^2$$

in distribution and Theorem 2.3 follows. □

ACKNOWLEDGEMENTS

The authors are grateful to the editor and a referee for their valuable suggestions that greatly improved the paper. This research is supported by the Youth Fund for Humanities and Social Sciences Research of Ministry of Education (No. 18YJC910003), the Natural Science Basic Research Plan in Shaanxi Province of China (No. 2020JM-276) and the Fundamental Research Funds for the Central Universities (No. GK201901008).

Received 9 August 2019

REFERENCES

[1] Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statist. Sinica* **6**, 311–329. MR1399305

[2] Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis of Count Data*. Cambridge University Press, New York. MR1648274

[3] Candes, E. J. and Tao, T. (2005). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313–2351. MR2382644

[4] Chang, J., Tang, C. Y. and Wu, T. T. (2018). A new scope of penalized empirical likelihood with high-Dimensional estimating equations. *Ann. Statist.* **46**, 3185–3216. MR3852649

[5] Chen, S. X. and Cui, H. J. (2003). An extended empiricial likelihood for generalized linear models. *Statist. Sinica* **13**, 69–81. MR1963920

[6] Chen, S. X. and Cui, H. J. (2006). On Bartlett correction of empirical likelihood in the presence of nuisance parameters. *Biometrika* **16**, 1101–1115. MR2277752

[7] Chen, S. X., Peng, L. and Qin, Y. L. (2009). Effects of data dimension on empirical likelihood. *Biometrika* **96**, 711–722. MR2538767

[8] Deb, P. and Trivedi, P. K. (1997). Demand for medical care by the elderly: a finite mixture approach. *J. Appl. Econ.* **12**, 313–336.

[9] Diciccio, T. J., Hall, P. and Romano, J. P. (1991). Empirical likelihood is Bartlett-correctable. *Ann. Statist.* **19**, 1053–1061. MR1105861

[10] Fan, G., Liang, H. and Shen, Y. (2016). Penalized empirical likelihood for high-dimensional partially linear varying coefficient model with measurement errors. *J. Multivar. Anal.* **147**, 183–201. MR3484177

[11] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360. MR1946581

[12] Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. *J. Roy. Statist. Soc. Ser. B* **70**, 849–911. MR2530322

[13] Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Stat. Sin.* **20**, 101–148. MR2640659

[14] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928–961. MR2065194

[15] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed. Springer, New York. MR2722294

[16] Hjort, N. L., McKeague, I. and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *Ann. Statist.* **37**, 1079–1111. MR2509068

[17] Kolaczyk, E. D. (1994). Empirical likelihood for generalized linear models. *Statist. Sinica* **4**, 199–218. MR1282871

[18] Lahiri, S. N. and Mukhopadhyay, S. (2013). A penalized empirical likelihood method in high dimensions. *Ann. Statist.* **40**, 2511–2540. MR3097611

[19] Leng, C. and Tang, C. Y. (2012). Penalized empirical likelihood and growing dimensional general estimating equations. *Biometrika* **19**, 703–716. MR2966779

[20] Liang, F., Song, Q. and Yu, K. (2016). Bayesian subset modeling for high-dimensional generalized linear models. *J. Amer. Statist. Assoc.* **108**, 589–606. MR3174644

[21] Lv, J. and Fan, Y. Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37**, 3498–3528. MR2549567

[22] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135**, 370–384. MR0375592

[23] Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single function. *Biometrika* **75**, 237–249. MR0946049

[24] Owen, A. B. (1991). Empirical likelihood for linear models. *Ann. Statist.* **19**, 1725–1747. MR1135146

[25] Owen, A. B. (2001). *Empirical Likelihood*. Chapman and Hall-CRC, New York.

[26] Park, M. Y. and Trevor, H. (2007). L1-regularization path algorithm for generalized linear models. *J. Roy. Statist. Soc. Ser. B* **69**, 659–677. MR2370074

[27] Qin, J. and Lawless, J. (1994). Empirical likelihood and generalized estimating equations. *Ann. Statist.* **22**, 300–325. MR1272085

[28] Ren, Y. and Zhang, X. (2011). Variable selection using penalized empirical likelihood. *Sci. China Math.* **54**, 1829–1845. MR2827018

[29] Tang, C. Y. and Leng, C. (2010). Penalized high-dimensional empirical likelihood. *Biometrika* **97**, 905–920. MR2746160

[30] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267–288. MR1379242

[31] Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge. MR1652247

[32] Wang, D. L., Wu, T. T. and Zhao, Y. C. (2019). Penalized empirical likelihood for the sparse Cox regression model. *J. Statist. Plann. Inference* **201**, 71–85. MR3913441

[33] Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameter. *J. Roy. Statist. Soc. Ser. B* **71**, 671–683. MR2749913

[34] Wang, M., Song, L. and Wang, X. (2010). Bridge estimation for generalized linear models with a diverging number of parameters. *Statist. Probab. Lett.* **80**, 1584–1596. MR2684004

[35] Wang, Q. H. and Jing, B. Y. (1999). Empirical likelihood for partially linear models with fixed designs. *Statist. Probab. Lett.* **41**, 425–433. MR1666112

[36] Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439–447. MR0375592

[37] Wu, T. T., Li, G. and Tang, C. (2015). Empirical likelihood for censored linear regression and variable selection. *Scand. J. Statist.*

42, 798–812. MR3391693
[38] Zeileis, A., Kleiber, C. and Jackman, S. (2008). Regression models for count data in R. *J. Stat. Softw.* **27**, 1–25.
[39] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418–1429. MR2279469

Xia Chen
School of Mathematics and Information Science
Shaanxi Normal University
Xi'an 710119
China
E-mail address: xchen80@snnu.edu.cn

Liyue Mao
School of Finance and Economics Management
Chongqing College of Electronic Engineering
Chongqing 401331
China
E-mail address: liyuemao@snnu.edu.cn