Penalized Jackknife Empirical Likelihood in High Dimensions

Zhoupung Li\textsuperscript{1}, Jinfeng Xu\textsuperscript{2}, Na Zhao\textsuperscript{2} and Wang Zhou\textsuperscript{3}

\textsuperscript{1}Lanzhou University, \textsuperscript{2}The University of Hong Kong and \textsuperscript{3}National University of Singapore

Abstract: Jackknife empirical likelihood (JEL) proposed by Jing et al. (2009) is an attractive approach for statistical inferences with nonlinear statistics such as $U$-statistics. However, most contemporary problems involve high dimensional model selection and its feasibility in theory and practice remains largely unexplored in the situations where the number of parameters diverges to infinity. In this paper, we propose a penalized JEL method which preserves the main advantages of JEL and leads to reliable variable selection based on the estimating equations with $U$-statistic structure in the high-dimensional setting. Under certain regularity conditions, we establish the asymptotic theory and oracle property for the JEL and its penalized version when the number of estimating equations and parameters increases along with the sample size. Simulation studies and real data analysis were carried out to examine the performance of the proposed methods and illustrate its practical utilities.

Key words and phrases: Estimating equations, high-dimensional data analysis, jackknife empirical likelihood, penalized likelihood, $U$-statistics, variable selection.
1. Introduction

Statistical inference based on estimating equations with $U$-statistics structure ($U$-type estimating equations) is common in nonparametric and semiparametric situations such as quantile regression and rank regression (Jin et al., 2003). Suppose that observations $X_1, \ldots, X_n$ are independent and identically distributed random vectors and the unknown parameters $\theta = (\theta_1, \ldots, \theta_p)^T$ can be estimated by solving the following $r$ ($r \geq p$) estimating equations

$$U_n(\theta) = \binom{n}{k}^{-1} \sum_{1 \leq i_1 < \ldots < i_k \leq n} h(X_{i_1}, \ldots, X_{i_k}; \theta) = 0, \quad (1.1)$$

where $h(\cdot) = (h_1(\cdot), \ldots, h_r(\cdot))^T$ are symmetric in $X = (X_1, \ldots, X_k)^T$'s, and satisfy $Eh(X_1, \ldots, X_k; \theta_0) = 0$ with $\theta_0$ which denotes the true value of $\theta$. Here $\Sigma$ denotes the summation over subsets of $k$ integers $\{i_1, \ldots, i_k\}$ from $\{1, \ldots, n\}$. For many estimation procedures in the literature, the estimator $\hat{\theta}$ is formally defined as the solution of the above $U$-type estimating equations (1.1), which is known as a $U$-type estimating problem. See Jin et al. (2001), Song and Ma (2010), Li et al. (2016), among others.

In this paper, we are interested in this problem under the high-dimensional setting, that is, $p$ and $r$ diverge with the sample size $n$. Hence we use $p_n$ and $r_n$ throughout the paper to emphasize the dependence of $p$ and $r$ on $n$.

High-dimensional data have now become ubiquitous in many applications such as
microarray data analysis, neuroimaging, and portfolio selection, where the number of parameters or variables, $p_n$, is very large, usually in thousands or beyond. Penalized methods are effective in analyzing such data, and various penalty functions have been proposed including the lasso (Tibshirani, 1996), the SCAD (Fan and Li, 2001), the adaptive lasso (Zou, 2006), least-squares approximation (Wang and Leng, 2007), and so on. Although these methodologies significantly improved the inference and variable selection procedures in the high-dimensional settings such as linear regression and generalized linear models, to our best knowledge, its feasibility in jackknife empirical likelihood inference with $U$-type estimating equations remains largely unexplored.

The empirical likelihood (EL) method introduced by Owen (1988, 1990), has been extensively studied and widely used to construct confidence regions and to test hypotheses in the literature. One nice feature of this method is that the confidence intervals and $p$-values of a test can be easily obtained without estimating the covariance matrix. More details can be found in Owen (2001), Chen and Van Keilegom (2009). Notice that the standard EL method works well in dealing with linear constraints, and an effective way of formulating the EL ratio statistic is via estimating equations as in Qin and Lawless (1994), while for nonlinear constraints, the EL method is extremely complicated in computation. To overcome the computational difficulty, Jing et al. (2009) proposed jackknife empirical likelihood (JEL) with
particular attention on nonlinear statistics involving $U$-statistics. Subsequently, Li et al. (2011), and Peng (2012) extended JEL to general estimating equations. Recently, high-dimensional data analysis incorporating with EL method attracted more attentions, and was investigated by Hjort et al. (2009), Chen et al. (2009), Tang and Leng (2010), Lahiri and Mukhopadhyay (2012), Leng and Tang (2012), Peng et al. (2014), Chang et al. (2015), Chen et al. (2015), Li et al. (2017), Chang et al. (2018), Wang et al. (2019), Tang et al. (2020), Chang et al. (2021) and references therein. Motivated by these developments, when dimension grows with the sample size $n$, simultaneous estimation of parameters and variable selection using $U$-type estimating equations (1.1) is of great interest and challenging, both in theory and computation. Our theoretical results can be summarized as two-fold.

1. We prove that JEL method is efficient when dealing with high-dimensional $U$-type estimating equations, and provide the corresponding algorithms in this scenario, which extends the scope of the JEL methods for $U$-type estimating equations from fixed dimensions (Li et al., 2016) to the case of diverging dimensions.

2. We propose a novel penalized JEL (PJEL) approach based on the $U$-type estimating equations. By choosing the proper penalty function, the approach preserves the main advantages of the JEL and the penalized method and the resulting estimator possesses good properties such as the oracle property,
rectly select the true sparse model with probability tending to one and with optimal efficiency. Also, Wilks’ theorem continues to hold for constructing confidence regions and testing hypotheses.

The rest of the paper is organized as follows. Section 2 presents the JEL method in high-dimensional settings. In Section 3, we describe the PJEL methodology and obtain its asymptotic theories. In Sections 4 and 5, we further report simulation results as well as a real data application to assess its finite-sample performance and illustrate its practical utility. All proofs are relegated to the Appendix.

2. Jackknife empirical likelihood with a diverging number of parameters

For the $U$-type estimating problem, the estimator $\hat{\theta}$ of $\theta_0$ is the solution to $U_n^T(\theta) = (U_{n,1}(\theta), ..., U_{n,r_n}(\theta))^T = 0$, where

$$U_{n,l}(\theta) = \left(\begin{array}{c} n \\ k \end{array}\right)^{-1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} h_l(X_{i_1}, ..., X_{i_k}; \theta), \quad l = 1, \ldots, r_n.$$ 

As the straightforward application of the standard EL method involves many non-linear constraints, which causes heavy computational burden, and thus is not favorable. To overcome this difficulty, Jing et al. (2009) proposed to first obtain $n$ jackknife pseudo values and then apply the standard EL to the nonlinear constraints in $U$-statistics. They demonstrated that this procedure is particular efficient in this
situation. However, their discussion is restricted to the estimation of the mean of one-sample and two-sample statistics, and the dimension is fixed, in this paper we propose a general JEL procedure for inferences with this $U$-type estimation problem which avoids the nonlinear constraints and permits the high-dimensional estimating equations. The method is constructed based on the fact that $U_n(\theta)$ is an unbiased and consistent estimator of $Eh(X_1, ..., X_k; \theta)$ and has mean zero at the true parameter value $\theta_0$. The details are as follows.

Define $T_n = U_n(\theta)$ and $T_{n-1}^{(-i)} = T(X_1, ..., X_{i-1}, X_{i+1}, ..., X_n; \theta)$, the statistic computed on the original data set with the $i$th observation removed. The jackknife pseudo-values

$$\hat{V}_i(\theta) = nT_n - (n - 1)T_{n-1}^{(-i)}$$

(2.2)

can be shown to be asymptotically independent under mild conditions. As they also unbiasedly and consistently estimate $Eh(X_1, ..., X_k; \theta)$, a standard empirical likelihood can then be constructed on $\hat{V}_i(\theta), i = 1, ..., n$, instead of the original observations $X_1, ..., X_n$ as follows. Specifically, the JEL function is defined as

$$\mathcal{L}(\theta) = \max \left\{ \prod_{i=1}^{n} p_i : \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \hat{V}_i(\theta) = 0 \right\},$$

(2.3)
with the corresponding likelihood ratio

\[ \mathcal{R}(\theta) = \max \left\{ \prod_{i=1}^{n} (np_i) : \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \hat{V}_i(\theta) = 0 \right\} \]  \hspace{1cm} (2.4)

Throughout the paper, \( \| \cdot \| \) always refers to the \( l_2 \)-norm \( \| \cdot \|_2 \), \( q \) is any fixed integer. In the Appendix, we prove the following theorems.

**Theorem 1.** Let \( \hat{\theta}_E \) be the minimizer of (2.4). Under Conditions C1-C5 in the Appendix, as \( n \to \infty \) and with probability tending to 1, we have

(i) \( \hat{\theta}_E \overset{P}{\to} \theta_0 \), (ii) \( \| \hat{\theta}_E - \theta_0 \| = O_p(a_n) \), where \( a_n \) is given in the Appendix.

**Theorem 2.** Under Conditions C1-C5 in the Appendix, we have

\[ A_n \Omega_{22.1}^{1/2} \sqrt{n}(\hat{\theta}_E - \theta_0) \overset{d}{\to} N(0, \Delta), \]

where \( A_n \in \mathbb{R}^{q \times p_n} \) such that \( A_n A_n^T \to G \) and \( G \) is a \( q \times q \) nonnegative matrix, \( \Omega_{22.1} \) is given in the Appendix.

### 3. Penalized jackknife empirical likelihood

In this section, we establish the asymptotic theory for the PJEL estimator. Suppose that \( \theta \) can be partitioned as \( \theta = (\theta_1^T, \theta_2^T)^T \), where \( \theta_1 \in \mathbb{R}^d \) and \( \theta_2 \in \mathbb{R}^{p_n-d} \) corre-
sponding to the nonzero and zero components, respectively. Then the true parameter value \( \theta_0 = (\theta_{10}^T, 0)^T \). We estimate the unknown parameter vector \( \theta_0 \) by minimizing

\[
l_p(\theta) = \sum_{i=1}^{n} \log \{1 + \lambda^T \hat{V}_i(\theta)\} + n \sum_{j=1}^{p_n} p_r(|\theta_j|),
\]

where \( p_r(|\theta_j|) \) is some penalty function, \( \tau \) is a tuning parameter that controls the trade-off between bias and model complexity (see Fan and Li, 2001). The PJEL has the following oracle property.

**Theorem 3.** Let \( \hat{\theta} = (\hat{\theta}_1^T, \hat{\theta}_2^T)^T \) be the minimizer of (3.5). Under Conditions C1-C7 in the Appendix, as \( n \to \infty \), we have

(i). \( P(\hat{\theta}_2 = 0) \to 1; \)

(ii). \( n^{1/2} B_n \Delta_{11}^{-1/2} (\hat{\theta}_1 - \theta_{10}) \xrightarrow{d} N(0, G^*) \), where \( B_n \) is a \( q \times d \) matrix such that \( B_n B_n^T \to G^* \), \( G^* \) is a \( q \times q \) nonnegative symmetric matrix, \( \Delta_{11} \) is given in the Appendix.

As the number of parameters may diverge to infinity, we next study the problem of testing the following linear hypothesis:

\[
H_0 : L_n \theta_{10} = 0, \quad v.s. \quad H_1 : L_n \theta_{10} \neq 0,
\]

where \( L_n \) is a \( q \times d \) matrix such that \( L_n L_n^T = I_q \) for fixed \( q \). Noting that this type of
hypothesis includes simultaneously testing whether a few variables are statistically significant. In the penalized likelihood context, Fan and Peng (2004) investigated this type of hypothesis testing in the parametric likelihood framework, and Leng and Tang (2012) lately considered this problem in a standard empirical likelihood settings for general estimating equations. Our results further generalized those existing results. Specifically, a PJEL ratio test statistic is defined as

\[ \tilde{l}(L_n) = -2\{l_p(\hat{\theta}) - \min_{\theta_1, L_n\theta_1 = 0} l_p(\theta)\}. \] (3.6)

The following theorem derives the asymptotic null distribution of the above test statistic.

**Theorem 4.** Under the null hypothesis and Conditions C1-C7 in the Appendix, we have

\[ \tilde{l}(L_n) \xrightarrow{d} \chi^2_q, \quad \text{as} \ n \to \infty. \]

From Theorem 4, a \((1 - \alpha)\)-level confidence region for \(L_n\theta_1\) can be constructed as

\[ C_\alpha = \left\{ v : -2\{l_p(\hat{\theta}) - \min_{\theta_1, L_n\theta_1 = v} l_p(\theta)\} \leq \chi^2_{q, 1-\alpha} \right\}, \]

where \(\chi^2_{q, 1-\alpha}\) is the \((1 - \alpha)\)th quantile of the \(\chi^2_q\) distribution.

Finally, the algorithms for implementing the proposed PJEL method and con-
structing the related confidence region are summarized as follows.

**Algorithm 1 Algorithm of PJEL**

Input \{X_i, Y_i\}; suppose \(\gamma, \epsilon_0\) are two pre-defined small numbers, for example, \(\gamma = 10^{-4}\).

For fixed \(\theta\), define \(\lambda(\theta)\) that minimizes \(l_p(\theta; \lambda)\);

Let \(k = 0\), initialize \(\theta^{(0)}\) by a general PIM estimator (e.g., probabilistic index model in the simulations, etc.).

repeat

Let \(\Theta_k = \{j : \theta_{j}^{(k)} \neq 0\}\), find \(\theta\) such that

\[
\theta = \arg\min_{\theta, \theta_{\Theta_k} = 0} l_p(\theta; \lambda(\theta)),
\]

where \(\Theta_k^c\) is the complimentary set of \(\Theta_k\) and \(\theta_A = \{\theta_j, j \in A\}\).

for \(j \in \Theta_k\)

if \(|\theta_j| < \gamma\) then

Take \(\theta_{j}^{(k+1)} = 0\),

else

Take \(\theta_{j}^{(k+1)} = \theta_j\);

end if

end for

Calculate \(L_{k+1} = l_p(\theta^{(k+1)}; \lambda(\theta^{(k+1)}))\);

\(k = k + 1;\)

until \(\max_{j \in \Theta_{k-1}} |\theta_{j}^{(k-1)} - \theta_{j}^{(k)}| < \epsilon_0\).
Algorithm 2 Construct confidence region by PJEL

Input \{X_i, Y_i\}, matrix \(L_n \in \mathbb{R}^{q \times d}\) for the hypothesis testing.
Let \(\hat{\theta}\) be the optimal estimator of \(\theta_0\) by Algorithm 1 and calculate the corresponding \(l_p(\hat{\theta}; \lambda(\hat{\theta}))\). \(C\) is a set of grid points \(\{c_j, j = 1, \cdots, K\}\) with some pre-defined constant \(K > 0\), where \(c_j\) are evenly spaced at interval \(C_0 = [L_n\hat{\theta} - c_0, L_n\hat{\theta} + c_0]\).
Here \(c_0\) is some pre-defined positive constant that ensures \(C_0\) can cover the \((1-\alpha)\)th level confidence interval.

for \(j\) from 1 to \(K\) do
  Fix \(\theta_{10}\) such that \(L_n\theta_{10} = c_j\), obtain the optimal estimator \(\tilde{\theta}^T = (\theta_{10}, \theta_{20})^T\) that minimizes \(l_p(\tilde{\theta}; \lambda(\tilde{\theta}))\) by Algorithm 1, obtain \(l_p(\tilde{\theta}; \lambda(\tilde{\theta}))\);
  if \(-2(l_p(\hat{\theta}; \lambda(\hat{\theta})) - l_p(\tilde{\theta}; \lambda(\tilde{\theta}))) \leq \chi^2_{q,1-\alpha}\) then
  Add \(c_j\) into \(\Theta_0\);
  end if
end for

Construct the \((1-\alpha)\)th level confidence region for \(L_n\theta_{10}\): \(C_\alpha = [c_l, c_u]\), where \(c_l\) and \(c_u\) are the minimum and maximum of \(C\) respectively.

4. Simulation studies

In this section, we use the probabilistic index model (PIM) (Thas et al, 2012) as an example to illustrate the proposed method for \(U\)-structured problems. The effect of the covariates \(X\) on the response \(Y\) is evaluated through the probabilistic index, which is defined as the probability \(P(Y_i \leq Y_j) := P(Y_i < Y_j) + 0.5P(Y_i = Y_j)\), where \(Y_i\) and \(Y_j\) are independent response variables with identical distribution function \(F\).

Data consist of i.i.d. observations \((Y_i, X_i), i = 1, \ldots, n\). A PIM is defined as:

\[
P(Y_i \leq Y_j | X_i, X_j) = m(X_i, X_j; \beta) = g^{-1}(Z_{ij}^T \beta),
\]
where \( g(\cdot) \) is a link function and \( Z_{ij} \) depends on \( X_i \) and \( X_j \). Following Thas et al (2012), let \( Z_{i,j} = X_j - X_i \) and for the probit and logit link function, the model is termed as the linear PIM and Cox PIM, respectively. De Neve and Thas (2015) proposed the \( U \)-type estimating equations:

\[
\sum_{(i,j)} U_{ij}(\beta) = 0, \quad U_{ij}(\beta) := Z_{ij} \left[ I(Y_i \leq Y_j) - g^{-1}(Z_{ij}^T \beta) \right].
\]

As in Thas et al (2012), we consider two scenarios as follows.

(a) **Normal linear model**: \( Y_i | X_i \) are i.i.d. \( N(\alpha_1 X_{1i} + \alpha_2 X_{2i}, 1) \). In this setting, the corresponding PIM is given by \( \Phi^{-1} \{ P(Y_i \leq Y_j | X_i, X_j) \} = \beta_1 (X_{1j} - X_{1i}) + \beta_2 (X_{2j} - X_{2i}) \), where \( \beta_i = \alpha_i / \sqrt{2} \) and \( \Phi \) is the distribution function of a standard normal distribution.

(b) **Exponential model**: \( Y_i | X_i \) are i.i.d. Exponential \( \{ \exp(\alpha_1 X_{1i} + \alpha_2 X_{2i}) \} \). In this setting, the corresponding PIM is given by logit \( \{ P(Y_i \leq Y_j | X_i, X_j) \} = \beta_1 (X_{1j} - X_{1i}) + \beta_2 (X_{2j} - X_{2i}) \), where \( \beta_i = -\alpha_i \).

In the above settings, the covariate \( X_1 \) is a Bernoulli random variable with success probability 0.5 and, \( X_2 \) follows \( U[0, 10] \), and \( \alpha_1 = 1, \alpha_2 = 1 \). For various combinations of \( p \) and \( n \), simulations were conducted with 1000 repetitions.

Tuning parameters \( \tau \) is taken from a fine grid and chosen by the BIC-type crite-
BIC(τ) = 2ℓ_p(β_τ) + C_n log ndf_τ,

where β_τ is the PJEL with tuning parameter τ and df_τ is the number of nonzero coefficient in β_τ. For fixed p, C_n = 1 and for diverging p, C_n = max{log log p, 1}. In the numerical study, we find that the selected tuning parameter for BIC decreases as sample size increases, while the remaining parameters like covariates dimension and value of β keep the same. Also, the ratio of λ to the square root of p/n increases along with the sample size (not shown in the tables). The results are consistent with the regularized condition in our assumptions, which validates the BIC-type criterion here.

We also compare the proposed method with penalized empirical likelihood (PEL) introduced by Tang and Leng (2010), and penalized maximum smoother rank correlation (PMSRC) introduced by Lin and Peng (2013), using the same datasets and tuning criteria. Since the method in Tang and Leng (2010) only considered the linear constraints case, we first use the sequential linearization method in Wood et al. (1996) to linearize the nonlinear constraints, and then applied the PEL to it. The latter method achieved the estimation through a penalized smoothing objective function for maximum rank correlation, which is a typical U-statistics. The simulation
results are summarized in Table 1 and Table 2, which report the performance of the proposed method in terms of the median of $L_2$ distance (MD) of the estimation, the average correctly selected zero coefficients (C) and the average incorrectly estimated zero coefficients (IC) for variable selection. It can be seen that in both scenarios, the PJEL method performs well in both the estimation and variable selection. Moreover, it identified most of the true zero-coefficients to be zero, while for PEL, the performance is not good enough for the variable selection. For PMSRC, although the estimation is quite good in linear PIM model, for Cox model, the performance is not as good as PJEL. Moreover, the variable selection capability is not comparable to the PJEL. Also, in order to illustrate the computational improvement of introducing jackknife method to EL model, we report the average computational time of PJEL and PEL in the above setting for per simulation in Table 3. From the table, we can find that, in both linear and Cox PIM models, the time PEL spent is much larger than that of PJEL, and when the sample size or covariates dimension grows, the ratio of the time expenditure between these two methods is growing larger. This indicates that when we are dealing with large sample size dataset via $U$-statistic structure estimating equations, with high dimension covariates, PEL is not applicable due to the computational burden. This reveals the advantage of the PJEL method compared to the PEL method in terms of the computational efficiency.
Table 1: Simulation results for linear PIM model with PJEL, PEL and PMSRC methods.

<table>
<thead>
<tr>
<th>n</th>
<th>p</th>
<th>PJEL</th>
<th></th>
<th></th>
<th>PEL</th>
<th></th>
<th></th>
<th>PMSRC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MD</td>
<td>C</td>
<td>IC</td>
<td>MD</td>
<td>C</td>
<td>IC</td>
<td>MD</td>
</tr>
<tr>
<td>200</td>
<td>5</td>
<td>0.0548</td>
<td>2.936</td>
<td>0.000</td>
<td>0.0527</td>
<td>2.346</td>
<td>0.000</td>
<td>0.0354</td>
</tr>
<tr>
<td>300</td>
<td>10</td>
<td>0.0418</td>
<td>7.934</td>
<td>0.000</td>
<td>0.0392</td>
<td>5.878</td>
<td>0.000</td>
<td>0.0295</td>
</tr>
<tr>
<td>400</td>
<td>15</td>
<td>0.0378</td>
<td>12.918</td>
<td>0.000</td>
<td>0.0366</td>
<td>9.395</td>
<td>0.000</td>
<td>0.0207</td>
</tr>
</tbody>
</table>

Table 2: Simulation results for Cox PIM model with PJEL, PEL and PMSRC methods.

<table>
<thead>
<tr>
<th>n</th>
<th>p</th>
<th>PJEL</th>
<th></th>
<th></th>
<th>PEL</th>
<th></th>
<th></th>
<th>PMSRC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MD</td>
<td>C</td>
<td>IC</td>
<td>MD</td>
<td>C</td>
<td>IC</td>
<td>MD</td>
</tr>
<tr>
<td>200</td>
<td>5</td>
<td>0.0647</td>
<td>2.930</td>
<td>0.053</td>
<td>0.0850</td>
<td>2.278</td>
<td>0.000</td>
<td>0.1252</td>
</tr>
<tr>
<td>300</td>
<td>10</td>
<td>0.0551</td>
<td>7.905</td>
<td>0.000</td>
<td>0.0548</td>
<td>6.598</td>
<td>0.000</td>
<td>0.1058</td>
</tr>
<tr>
<td>400</td>
<td>15</td>
<td>0.0446</td>
<td>12.904</td>
<td>0.000</td>
<td>0.0460</td>
<td>9.732</td>
<td>0.000</td>
<td>0.0919</td>
</tr>
</tbody>
</table>

Table 3: The average consumption time (in seconds) for PJEL and PEL with deterministic sample size $n$ and covariates $p$.

<table>
<thead>
<tr>
<th>n</th>
<th>p</th>
<th>Linear PIM</th>
<th></th>
<th></th>
<th>Cox PIM</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PJEL</td>
<td>PEL</td>
<td></td>
<td>PJEL</td>
<td>PEL</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>5</td>
<td>4.140</td>
<td>16.394</td>
<td>5.236</td>
<td>34.895</td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>10</td>
<td>18.093</td>
<td>115.586</td>
<td>40.399</td>
<td>344.050</td>
<td></td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>15</td>
<td>58.542</td>
<td>451.907</td>
<td>191.416</td>
<td>1692.695</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Next, we illustrate the performance of PJEL in constructing confidence region. We set $L_n$ in (3.6) to be $(1, 0, \cdots, 0)$, and take the null hypothesis testing $H_0 : \beta_{10} = a$ with $a = \beta_1 - 0.2, \beta_1 - 0.1, \beta_1, \beta_1 + 0.1, \beta_1 + 0.2$ separately, where $\beta_{10}$ denotes the first component of $\beta$. Under the nominal level $\alpha_0 = 0.05$, we summarize the empirical size for deterministic value of $a$ in Table 4. From the table, we can see the size of
our test is close to the nominal level, and increases when the null value differs from
the true value $\beta_1$. The results validate the likelihood ratio test under PJEL, and
indicates the feasibility to construct confidence region by taking hypothesis at finer
grid points.

Table 4: The empirical percentages that a given value does not fall in the 95%
confidence interval

<table>
<thead>
<tr>
<th>PIM model</th>
<th>$n$</th>
<th>$p$</th>
<th>$\beta_{10} - 0.2$</th>
<th>$\beta_{10} - 0.1$</th>
<th>$\beta_{10}$</th>
<th>$\beta_{10} + 0.1$</th>
<th>$\beta_{10} + 0.2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal linear</td>
<td>200</td>
<td>5</td>
<td>24.40</td>
<td>11.20</td>
<td>4.60</td>
<td>10.00</td>
<td>26.10</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>10</td>
<td>40.40</td>
<td>15.80</td>
<td>6.40</td>
<td>12.00</td>
<td>33.80</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>15</td>
<td>54.70</td>
<td>22.50</td>
<td>4.70</td>
<td>12.20</td>
<td>40.70</td>
</tr>
<tr>
<td>Exponential</td>
<td>200</td>
<td>5</td>
<td>26.90</td>
<td>14.10</td>
<td>6.50</td>
<td>9.10</td>
<td>16.80</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>10</td>
<td>31.70</td>
<td>15.60</td>
<td>6.30</td>
<td>10.10</td>
<td>21.40</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>15</td>
<td>41.60</td>
<td>20.00</td>
<td>6.90</td>
<td>10.60</td>
<td>26.70</td>
</tr>
</tbody>
</table>

5. An application to the air pollution study

Adverse effects of air pollution on human health have been explored in many scientific
studies. In particular, it is of great interest to evaluate the effect of air pollution on
human mortality. Here, we apply the proposed method to the air pollution study
(McDonald et al., 1973) to identify the factors associated with air that affect the
mortality the most. The dataset consists of 60 observations and 15 features. The
response variable is “Mortality”, representing the total age-adjusted mortality rate
per 100,000. A detailed description of 15 variables in the air pollution study is given
McDonald et al. (1973) proposed to use multiple linear regression to assess the covariate effects on the mortality rate. Here, apply the proposed method to fit a normal linear PIM, to identify the factors which are relevant to the mortality. The BIC-type criterion is used to choose the parameter $\lambda$. Then using the penalized jackknife empirical likelihood method, six covariates are identified to be relevant ones and reported in Table 5 along with the constructed 95% confidence intervals by the penalized JEL procedures for inference. The identified covariates are: \textit{average annual precipitation} (PREC), \textit{average January temperature} (JANT), \textit{median school years completed over 22 years} (EDU), \textit{percent non-white population} (NONW), \textit{hydrocarbon pollution potential} (HC) and \textit{nitrous oxide pollution potential} (NOX). The variables
selected here are similar to the ones by McDonald et al. (1973). They used two different criteria to choose the factors. For the result of using ridge regression, out of 7 features they chose, we have 4 of them in common (annual precipitation, January temperature, education, and non-white percentage). In addition, we find that the hydrocarbon as well as nitric oxides pollution potential also have adverse effects on the human mortality rate.

Table 6: 95% confidence intervals of the estimated non-zero coefficients for air pollution study.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Estimated coefficient</th>
<th>Confidence interval by PJEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREC</td>
<td>0.731</td>
<td>[0.361, 1.203]</td>
</tr>
<tr>
<td>JANT</td>
<td>-1.113</td>
<td>[-1.472, -0.822]</td>
</tr>
<tr>
<td>EDUC</td>
<td>-0.390</td>
<td>[-0.712, -0.161]</td>
</tr>
<tr>
<td>NONW</td>
<td>1.743</td>
<td>[1.36, 2.23]</td>
</tr>
<tr>
<td>HC</td>
<td>-3.542</td>
<td>[-5.612, -2.65]</td>
</tr>
<tr>
<td>NOX</td>
<td>3.956</td>
<td>[2.096, 5.89]</td>
</tr>
</tbody>
</table>

Supplementary Material

The Supplementary Material includes detailed proofs of the main theorems.

Acknowledgements

The authors sincerely thank the Editor, an Associate Editor, and two anonymous reviewers for their very helpful and constructive comments that have led to significant improvements in the paper. Zhouping Li’s research was supported by the National...
REFERENCES

Natural Science Foundation of China (11571154), Jinfeng Xu and Na Zhao’s research was supported by National Natural Science Foundation of China (72033002), the University of Hong Kong-Zhejiang Institute of Research and Innovation Seed Fund, and General Research Fund (17306619, 17308018 and 17308820) of Hong Kong, Wang Zhou’s research was partially supported by a grant R-155-000-192-114 at the National University of Singapore.

References


School of Mathematics and Statistics, Center for Data Science, Lanzhou University, Lanzhou 730000, China.

E-mail: lizhp@lzu.edu.cn

Zhejiang Institute of Research and Innovation, The University of Hong Kong, Hangzhou, China

Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, China.

E-mail: xujf@hku.hk

Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, China.

E-mail: zhaona@connect.hku.hk

Department of Statistics and Applied Probability, National University of Singapore, Singapore
E-mail: stazw@nus.edu.sg