

Penalized Empirical Likelihood and Growing Dimensional General Estimating Equations

BY CHENLEI LENG AND CHENG YONG TANG

*Department of Statistics and Applied Probability, National University of Singapore, Singapore
117546, Republic of Singapore*

stalc@nus.edu.sg statc@nus.edu.sg

SUMMARY

When a parametric likelihood function is not specified for a model, estimating equations provide an instrument for statistical inference. Qin & Lawless (1994) illustrated that empirical likelihood makes optimal use of these equations in inferences for fixed (low) dimensional unknown parameters. In this paper, we study empirical likelihood for general estimating equations with growing (high) dimensionality and propose a penalized empirical likelihood approach for parameter estimation and variable selection. We quantify the asymptotic properties of empirical likelihood and its penalized version, and show that penalized empirical likelihood has the oracle property. The performance of the proposed method is illustrated via several simulated applications and a data analysis.

Some key words: Empirical likelihood; General estimating equations; High dimensional data analysis; Penalized likelihood; Variable selection.

1. INTRODUCTION

Empirical likelihood is a computationally intensive nonparametric approach for deriving estimates and confidence sets for unknown parameters. Detailed in Owen (2001), empirical likelihood shares some merits of parametric likelihood approach, such as limiting chi-square distributed likelihood ratio and Bartlett correctability (DiCiccio et al., 1991; Chen & Cui, 2006). On the other hand, as a data driven nonparametric approach, it is attractive in robustness and flexibility in incorporating auxiliary information (Qin & Lawless, 1994). We refer to Owen (2001) for a comprehensive overview, and Chen & Van Keilegom (2009) for a survey of recent development in various areas.

Let Z_1, \dots, Z_n be independent and identically distributed random vectors from some distribution, and $\theta \in \mathbb{R}^p$ be a vector of unknown parameters. Suppose that data information is available in the form of an unbiased estimating function $g(z; \theta) = \{g_1(z; \theta), \dots, g_r(z; \theta)\}^T$ ($r \geq p$) such that $E\{g(Z_i; \theta_0)\} = 0$. Besides the score equations derived from a likelihood, the choice of $g(z; \theta)$ is more flexible and accommodates a wider range of applications, for example, the pseudo-likelihood approach (Godambe & Heyde, 1987), the instrumental variables method in measurement error models (Fuller, 1987) and survey sampling (Fuller, 2009), the generalized method of moments (Hansen, 1982; Hansen & Singleton, 1982) and the generalized estimating equations approach in longitudinal data analysis (Liang & Zeger, 1986).

When $r = p$, θ can be estimated by solving the estimating equations $0 = n^{-1} \sum_{i=1}^n g(Z_i; \theta)$. Allowing $r > p$ provides a useful device to combine available information for improved efficiency, but then directly solving $0 = n^{-1} \sum_{i=1}^n g(Z_i; \theta)$ may not be feasible. Hansen (1982) and

49 Godambe & Heyde (1987) discussed optimal ways to combine these equations for fixed p . They
 50 showed that the optimal estimator $\tilde{\theta}$ satisfies $\sqrt{n}(\tilde{\theta} - \theta_0) \rightarrow N\{0, V(\theta_0)\}$ in distribution with

$$51 \quad V(\theta) = \left(E \left\{ \frac{\partial g(Z_i; \theta)}{\partial \theta^T} \right\} [E\{g(Z_i; \theta)g^T(Z_i; \theta)\}]^{-1} E \left\{ \frac{\partial g(Z_i; \theta)}{\partial \theta} \right\} \right)^{-1}. \quad (1)$$

54 Qin & Lawless (1994) showed that empirical likelihood optimally combines information. More
 55 specifically, the maximizer $\tilde{\theta}$ of

$$56 \quad L(\theta) = \sup \left\{ \prod_{i=1}^n n w_i : w_i \geq 0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i g(Z_i; \theta) = 0 \right\} \quad (2)$$

60 is optimal in the sense of Godambe & Heyde (1987). Define the empirical likelihood ratio
 61 as $\ell(\theta) = -[\log\{L(\theta)\} - n \log(n)]$. Qin & Lawless (1994) further showed that $-2\{\ell(\tilde{\theta}) -$
 62 $\ell(\theta_0)\} \rightarrow \chi_p^2$ in distribution as $n \rightarrow \infty$. This device is useful in testing hypotheses and obtaining
 63 confidence regions for θ . Compared to the Wald type confidence region, this approach respects
 64 the range of θ and imposes no shape constraint (Qin & Lawless, 1994).

65 Our motivations for this paper are multiple-fold. Contemporary statistics often deals with
 66 datasets with diverging dimensionality. Sparse models can help interpretation and improve pre-
 67 diction accuracy. There is a large literature on the penalized likelihood approach for building such
 68 models; for example lasso (Tibshirani, 1996), the smoothly clipped absolute deviation method
 69 (Fan & Li, 2001), adaptive lasso (Zou, 2006; Zhang & Lu, 2007), least squares approximation
 70 (Wang & Leng, 2007), the folded concave penalty (Lv & Fan, 2009). Despite these develop-
 71 ments, it is not clear how existing methods can be applied to general estimating equations with
 72 diverging dimensionality. When likelihood is not available, estimating equations can be more
 73 flexible and information from additional estimating equations can improve the estimation effi-
 74 ciency (Hansen, 1982). Reducing the effective dimension of the unknown parameter θ may
 75 lead to extra efficiency gain. From this perspective, sparse models in the estimating equations
 76 framework provide additional insights.

77 The importance of high dimensional statistical inference using empirical likelihood was only
 78 recently recognized by Hjort et al. (2009) and Chen et al. (2009). Neither paper explored model
 79 selection. Tang & Leng (2010) studied variable selection using penalty in the empirical likeli-
 80 hood framework, which is limited to mean vector estimation and linear regression models. When
 81 dimension grows, variable selection using more general estimating equations is thus of greater
 82 interest. Empirical likelihood for general estimating equations with growing dimensionality is
 83 challenging, theoretically and computationally. First, the number of Lagrange multipliers which
 84 are used to characterize the solution and to derive asymptotic results, increases with the sample
 85 size. It is not clear how appropriate bounds can be obtained. Second, empirical likelihood usu-
 86 ally involves solving nonconvex optimization and any generalization of it to address the issue of
 87 variable selection is nontrivial. The main contributions of this work are summarized as follows:

- 89 1. We show that empirical likelihood gives efficient estimates by combining high dimensional
 90 estimating equations. This generalizes the results in Qin & Lawless (1994) derived for fixed
 91 dimension, which may be of independent interest;
- 92 2. For building sparse models, we propose an estimating equation-based penalized empirical
 93 likelihood, a unified framework for variable selection in optimally combining estimating
 94 equations. With a proper penalty function, the resulting estimator retains the advantages
 95 of both empirical likelihood and the penalized likelihood approach. More specifically, this
 96 method has the oracle property (Fan & Li, 2001; Fan & Peng, 2004) by identifying the true

sparse model with probability tending to one and with optimal efficiency. Moreover, Wilks' theorem continues to apply and serves as a robust method for testing hypothesis and constructing confidence regions.

The oracle property of the proposed method does not require strict distributional assumptions, thus is robust against model misspecification. The proposed method is widely applicable as long as unbiased estimating equations can be formed, even when a likelihood is unavailable. Later we outline four such applications in which estimating equations are more natural than the usual likelihood function, and the efficiency of the estimates is improved by having more estimating equations than the parameters. To our best knowledge, variable selection for these examples in a high dimensional setup has not been investigated.

2. EMPIRICAL LIKELIHOOD FOR HIGH DIMENSIONAL ESTIMATING EQUATIONS

We first extend the fixed dimensional results in Qin & Lawless (1994) to cases with diverging dimensionality, i.e., $r, p \rightarrow \infty$ as $n \rightarrow \infty$. Via Lagrange multipliers, the weights $\{w_i\}_{i=1}^n$ in (2) are given by $w_i = n^{-1}\{1 + \lambda_\theta^\top g(Z_i; \theta)\}^{-1}$ where λ_θ satisfies $n^{-1} \sum_{i=1}^n g(Z_i; \theta)\{1 + \lambda_\theta^\top g(Z_i; \theta)\}^{-1} = 0$. By noting that the global maximum of (2) is achieved at $w_i = n^{-1}$, the empirical likelihood ratio is given by

$$\ell(\theta) = -[\log\{L(\theta)\} - n \log(n)] = \sum_{i=1}^n \log\{1 + \lambda_\theta^\top g(Z_i; \theta)\}. \quad (3)$$

Thus maximizing (2) is equivalent to minimizing (3). In high dimensional empirical likelihood, the magnitude of $\|\lambda_\theta\|$ is no longer $O_p(n^{-1/2})$ as in the fixed dimensional case (Hjort et al., 2009; Chen et al., 2009). To develop an asymptotic expansion for (3), $\lambda_\theta^\top g(Z_i; \theta)$ needs to be stochastically small uniformly, which is ensured by Lemma 1 in the Appendix. Let $a_n = (p/n)^{1/2}$, and $D_n = \{\theta : \|\theta - \theta_0\| \leq C a_n\}$ be a neighborhood of θ_0 for some constant $C > 0$. Let $g_i(\theta) = g(Z_i; \theta)$ and $g(\theta) = E\{g_i(\theta)\}$. The following regularity conditions are assumed.

- A.1 The support of θ denoted by Θ is a compact set in \mathbb{R}^p , $\theta_0 \in \Theta$ is the unique solution to $E\{g_i(\theta)\} = 0$.
- A.2 $E\{\sup_{\theta \in \Theta} (\|g_i(\theta)\| r^{-1/2})^\alpha\} < \infty$ for some $\alpha > 10/3$ when n is large.
- A.3 Let $\Sigma(\theta) = E[\{g_i(\theta) - g(\theta)\}\{g_i(\theta) - g(\theta)\}^\top]$. There exists b and B such that the eigenvalues of $\Sigma(\theta)$ satisfy $0 < b \leq \gamma_1\{\Sigma(\theta)\} \leq \dots \leq \gamma_r\{\Sigma(\theta)\} \leq B < \infty$ for all $\theta \in D_n$ when n is large.
- A.4 As $n \rightarrow \infty$, $p^5/n \rightarrow 0$ and $p/r \rightarrow y$ for some y such that $C_0 < y < 1$ where $C_0 > 0$.
- A.5 There exist $C_1 < \infty$ and $K_{ij}(z)$ such that for all $i = 1, \dots, r$ and $j = 1, \dots, p$

$$\frac{\partial g_i(z; \theta)}{\partial \theta_j} \leq K_{ij}(z), \quad E\{K_{ij}^2(Z)\} \leq C_1 < \infty, \quad (i = 1, \dots, r, j = 1, \dots, p).$$

There exist C_2 and $H_{ijk}(z)$ such that for the i th estimating equation

$$\frac{\partial^2 g_i(z; \theta)}{\partial \theta_j \partial \theta_k} \leq H_{ijk}(z), \quad E\{H_{ijk}^2(Z)\} \leq C_2 < \infty.$$

Conditions A.1 and A.2 are from Newey & Smith (2004) to ensure the existence and consistency of the minimizer of (3) and to control the tail probability behavior of the estimating function. Condition A.4 requires that $p = o(n^{1/5})$ where the rate on p should not be taken as

restrictive because empirical likelihood is studied in a broad framework based on general estimating equations. Since no particular structural information is available on $g(z; \theta)$, establishing the theoretical result is very challenging so that strong regularity conditions are needed and the bounds in the stochastic analysis are conservative. This is also the case in Fan & Peng (2004) in studying the penalized likelihood approach in high dimension. When specific model structure is available, the restriction on dimensionality p can be relaxed. Here $r/p \rightarrow y$ in A.4 is for simplicity in presenting the theoretical results. There are also situations when p is fixed and r is diverging (Xie & Yang, 2003), in which our framework also applies. We emphasize that the dimensionality r effectively can not exceed n because the convex hull of $\{g(Z_i; \theta)\}_{i=1}^n$ is at most a subset in \mathbb{R}^n as seen from the definition (2).

We now show the consistency of the empirical likelihood estimate and its rate of convergence.

THEOREM 1. *Under Conditions A.1-A.5, as $n \rightarrow \infty$ and with probability tending to 1, the minimizer $\hat{\theta}_E$ of (3) satisfies a) $\hat{\theta}_E \rightarrow \theta_0$ in probability, and b) $\|\hat{\theta}_E - \theta_0\| = O_p(a_n)$.*

We now present the theoretical property of the high dimensional empirical likelihood.

THEOREM 2. *Under Conditions A.1-A.5, $\sqrt{n}A_n V^{-1/2}(\theta_0)(\hat{\theta}_E - \theta_0) \rightarrow N(0, G)$ in distribution where $A_n \in \mathbb{R}^{q \times p}$ such that $A_n A_n^T \rightarrow G$ and G is a $q \times q$ matrix with fixed q and $V(\theta_0)$ is given by (1).*

From Theorem 2, the asymptotic variance $V(\theta_0)$ of $\hat{\theta}_E$ remains optimal as in fixed dimensional cases (Hansen, 1982; Godambe & Heyde, 1987). Theorem 2 implies that for the high dimensional estimating equation, empirical likelihood based estimate achieves the optimal efficiency.

We remark that the framework presented in this paper is applicable only to the case where the sample size is larger than the dimension of the parameter. When that is violated, preliminary methods such as sure independence screening (Fan & Lv, 2008) may be used to reduce the dimensionality. This condition cannot be improved because empirical likelihood does not have a solution due to the fact that there are more constraints than the observations.

3. PENALIZED EMPIRICAL LIKELIHOOD

In high dimensional data analysis, it is reasonable to expect that only a subset of the covariates are relevant. To identify the subset of influential covariates, we propose to use the penalized empirical likelihood by complementing (2) with a penalty functional. Using Lagrange multipliers, we consider equivalently minimizing the penalized empirical likelihood ratio defined as

$$\ell_p(\theta) = \sum_{i=1}^n \log\{1 + \lambda^T g(Z_i; \theta)\} + n \sum_{j=1}^p p_\tau(|\theta_j|), \quad (4)$$

where $p_\tau(|\theta_j|)$ is some penalty function with tuning parameter τ controlling the trade-off between bias and model complexity (Fan & Li, 2001).

Write $\mathcal{A} = \{j : \theta_{0j} \neq 0\}$ and its cardinality as $d = |\mathcal{A}|$. Without loss of generality, let $\theta = (\theta_1^T, \theta_2^T)^T$ where $\theta_1 \in \mathbb{R}^d$ and $\theta_2 \in \mathbb{R}^{p-d}$ correspond to the nonzero and zero components respectively. This implies $\theta_0 = (\theta_{10}^T, 0)^T$. We correspondingly decompose $V(\theta_0)$ in (1) as

$$V(\theta_0) = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}.$$

The following regularity conditions on the penalty function are assumed.

A.6 As $n \rightarrow \infty$, $\tau(n/p)^{1/2} \rightarrow \infty$ and $\min_{j \in \mathcal{A}} \theta_{0j}/\tau \rightarrow 0$.

A.7 Assume $\max_{j \in \mathcal{A}} p'_\tau(|\theta_{0j}|) = o\{(np)^{-1/2}\}$ and $\max_{j \in \mathcal{A}} p''_\tau(|\theta_{0j}|) = o(p^{-1/2})$.

Condition A.6 states that the nonzero parameters can not converge to zero too fast. This is reasonable because otherwise the noise is too strong. Condition A.7 holds by many penalty functions such as the penalty in Fan & Li (2001) and the minimax concave penalty (Zhang, 2010). The penalized empirical likelihood has the following oracle property.

THEOREM 3. *Let $\hat{\theta} = (\hat{\theta}_1^\top, \hat{\theta}_2^\top)^\top$ be the minimizer of (4). Under Conditions A.1-A.7, as $n \rightarrow \infty$, we have the following results.*

1. *With probability tending to one, $\hat{\theta}_2 = 0$.*
2. *Let $V_p(\theta_0) = V_{11} - V_{12}V_{22}^{-1}V_{21}$. Then $\sqrt{n}B_nV_p^{-1}(\theta_0)(\hat{\theta}_1 - \theta_{10}) \rightarrow N(0, G)$ in distribution, where $B_n \in \mathbb{R}^{q \times d}$, q is fixed and $B_nB_n^\top \rightarrow G$ as $n \rightarrow \infty$.*

Theorem 3 implies that the zero components in θ_0 are estimated as zero with probability tending to one. Comparing Theorem 3 to Theorem 2, penalized empirical likelihood gives more efficient estimates of the nonzero components. As shown in the proof of Theorem 3, the efficiency gain is due to the reduction of the effective dimension of θ via penalization. It can be shown further that the penalized empirical likelihood estimate $\hat{\theta}_1$ is optimal in the sense of Heyde & Morton (1993) as if empirical likelihood were applied to the true model. We show in the later simulations that the improvement can be very large, sometimes substantial.

Next we consider testing statistical hypotheses and constructing confidence regions for θ . Consider the null hypothesis of fixed dimensionality in the following form

$$H_0 : L_n\theta_0 = 0, \quad H_1 : L_n\theta_0 \neq 0,$$

where $L_n \in \mathbb{R}^{q \times d}$ such that $L_nL_n^\top = I_q$ for a fixed q , and I_q is the q -dimensional identity matrix. Such hypotheses include testing for individual and multiple components of θ_0 as special cases, and can be easily extended to linear functions of θ_0 . A similar type of hypothesis testing was considered in Fan & Peng (2004) under a parametric likelihood framework. Based on the empirical likelihood formulation, a penalized empirical likelihood ratio test statistic is constructed as

$$\tilde{\ell}(L_n) = -2 \left\{ \ell_p(\hat{\theta}) - \min_{\theta, L_n\theta=0} \ell_p(\theta) \right\}. \quad (5)$$

We show the asymptotic property of this ratio in the following theorem.

THEOREM 4. *Under the null hypothesis and Conditions A.1-A.7, as $n \rightarrow \infty$, $\tilde{\ell}(L_n) \rightarrow \chi_q^2$.*

As a consequence, a $(1 - \alpha)$ -level confidence set for $L_n\theta$ can be constructed as

$$V_\alpha = \left[v : -2 \left\{ \ell_p(\hat{\theta}) - \min_{\theta, L_n\theta=v} \ell_p(\theta) \right\} \leq \chi_{q,1-\alpha}^2 \right] \quad (6)$$

where $\chi_{q,1-\alpha}^2$ is the $1 - \alpha$ level quantile of χ_q^2 distribution.

Theorem 4 extends the results in Qin & Lawless (1994) to growing dimensionality. For the full parametric likelihood approach, Fan & Peng (2004) showed that the likelihood ratio statistic has similar properties given in Theorem 4.

The attractiveness of empirical likelihood and its penalized version comes at the expense of computation. Due to the nonconvexity, computing empirical likelihood is nontrivial (Owen,

2001). Penalized empirical likelihood computation involving a non-differentiable penalty is obviously more involved. We propose a nested optimization procedure in minimizing (4). Due to the non-quadratic nature of the loss function, we iterate between solving for λ and θ . When λ is fixed, we use the local quadratic approximation in Fan & Li (2001) by approximating $p_\tau(|\theta_j|)$ as $p_\tau(|\theta_j^{(k)}|) + \frac{1}{2}\{p'_\tau(|\theta_j^{(k)}|)/|\theta_j^{(k)}|\}\{\theta_j^2 - (\theta_j^{(k)})^2\}$, where $\theta_j^{(k)}$ is the k th step estimate of θ_j . We then make use of the algorithm discussed in Owen (2001) Chapter 12 to obtain the minimizer of (4) through nonlinear optimization. The procedure is repeated until convergence by using the resulting minimizer as the next initial value. Our experience suggests that this algorithm converges quickly, usually in fewer than ten iterations given a good initial value.

To choose the penalty parameter τ , we use the following BIC type function proposed by Wang et al. (2009)

$$\text{BIC}(\tau) = -2\ell(\theta_\tau) + C_n \cdot \log(n) \cdot \text{df}_\tau$$

where θ_τ is the estimate of θ with τ being the tuning parameter; df_τ is the number of nonzero coefficient in θ_τ ; C_n is a scaling factor diverging to infinity at a slow rate as $p \rightarrow \infty$. When p is fixed, we can simply take $C_n = 1$ as for the usual BIC. Otherwise, $C_n = \max\{\log \log p, 1\}$ seems to be a good choice. The growing C_n is used to offset the effect of a growing dimension. However, a rigorous justification is nontrivial and will be studied in future work.

4. SIMULATION AND DATA ANALYSIS

We present extensive simulation studies to illustrate the usefulness of penalized empirical likelihood. We choose examples from cases where $r > p$ such that the number of estimating equations is greater than the number of parameters. The proposed method is also applicable for $r = p$ when likelihood score functions or the first derivatives of a loss function are used. We compare the penalized empirical likelihood estimates with competing methods whenever appropriate in terms of estimation accuracy. We also give variable selection results for the simulation studies, as well as hypothesis testing results in terms of the size and power. In our implementation, we use the penalty in Fan & Li (2001) although other penalties can also be used. Specifically, the first derivative of the penalty function is defined as

$$p'_\tau(\theta) = \tau \left\{ I(\theta \leq \tau) + \frac{(a\tau - \theta)_+}{(a-1)\tau} I(\theta > \tau) \right\},$$

for $\theta > 0$, where $a = 3.7$, and $(s)_+ = s$ for $s > 0$ and 0 otherwise.

Example 1. Longitudinal data arise commonly in biomedical research with repeated measurements from the same subject or within the same cluster. Let Y_{it} and X_{it} be the response and covariate of the i th subject measured at time t . Here, $i \in \{1, \dots, n\}$ and $t \in \{1, \dots, m_i\}$ index the subject and measurement respectively. The estimating equations utilize the marginal moment conditions without resorting to the likelihood, which is complicated especially for categorical responses. Let $E(Y_{it}) = \mu(X_{it}^T \beta) = \mu_{it}$ where $\beta \in \mathbb{R}^p$ is the parameter of interest. Incorporating the dependence among the repeated measurements is essential for efficient inference. Liang & Zeger (1986) proposed to estimate β by solving $0 = \sum_{i=1}^n \dot{\mu}_i^T W_i^{-1} (Y_i - \mu_i)$. Here for the i th subject, $Y_i = (Y_{i1}, \dots, Y_{im_i})^T$, $\mu_i = (\mu_{i1}, \dots, \mu_{im_i})^T$, $\dot{\mu}_i = \partial \mu_i / \partial \beta$ and $W_i = v_i^{1/2} R v_i^{1/2}$ where v_i is a diagonal matrix of the conditional variances of subject i and $R = R(\alpha)$ is a working correlation matrix indexed by α . This is the estimating equations method with $g(Z_i; \beta) = \dot{\mu}_i^T W_i^{-1} (Y_i - \mu_i)$ where $Z_i = (Z_{i1}^T, \dots, Z_{im_i}^T)^T$, $Z_{it} = (Y_{it}, X_{it}^T)^T$ and $r = p$. Liang & Zeger (1986) proposed to estimate α and the dispersion parameter by the method of moments.

More recently, Qu et al. (2000) proposed to model R^{-1} by $\sum_{i=1}^m a_i M_i$ where M_1, \dots, M_m are known matrices and a_1, \dots, a_m are unknown constants. Then β can be estimated by the quadratic inference functions approach (Qu et al., 2000) that uses

$$g(Z_i; \beta) = \begin{pmatrix} \mu_i^T v_i^{-1/2} M_1 v_i^{-1/2} (Y_i - \mu_i) \\ \vdots \\ \mu_i^T v_i^{-1/2} M_m v_i^{-1/2} (Y_i - \mu_i) \end{pmatrix}, \quad (i = 1, \dots, n). \quad (7)$$

This falls into our framework with $r > p$ when $m > 1$, and with $r = p$ if $m = 1$.

In this simulation study, we consider the model

$$y_{ij} = x_{ij}^T \beta + \varepsilon_{ij}, \quad (i = 1, \dots, n; j = 1, 2, 3),$$

where $\beta = (3, 1.5, 0, 0, 2, 0, \dots, 0)^T \in \mathbb{R}^p$, x_{ij} are generated from multivariate normal distribution $N(0, \Sigma)$ with $\Sigma_{kl} = 0.5^{|k-l|}$. The random error $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3})^T$ is generated from a three-dimensional normal distribution with mean zero, marginal variance 1. The correlation we simulate for the random error is either compound symmetry or AR(1) with parameter 0.7. We use two sets of basis matrices in fitting the model. We take $M_1 = I_3$ as the identity matrix. The second basis matrix M_2 is either a matrix with 0 on the diagonal and 1 elsewhere, or a matrix with two main off-diagonals being 1 and 0 elsewhere. Note that these two sets of basis matrices are referred to as the working structures and are called compound symmetry and AR(1) working assumptions respectively (Qu et al., 2000). In our setup, there are $r = 2p$ estimating equations to estimate p parameters. For each simulation, we repeat the experiment 1000 times. We try different sample sizes $n = 50, 100, 200, 400$ and we take p as the integer part of $10(3n)^{1/5.1} - 20$, which enables us to study the asymptotic properties of empirical likelihood. We compare the usual least-squares estimate, the Oracle least-squares estimator, empirical likelihood estimator, the oracle empirical likelihood estimator and the proposed penalized empirical likelihood estimator, in terms of the mean squared error $\text{MSE} = E\{(\hat{\beta} - \beta)^T(\hat{\beta} - \beta)\}$. For the oracle estimates, only the covariates corresponding to the nonzero coefficients are used in estimation. We report the Monte Carlo estimate of MSE and its sample standard error in 1000 simulations.

The results are summarized in Table 1. Empirical likelihood is more efficient than least-squares because more estimating equations are used. Similar phenomenon happens for their oracle versions. These agree with the general conclusion in Qu et al. (2000). The proposed method has smaller MSE than empirical likelihood and oracle least squares, indicating the gain in accuracy by having more estimating equations and using the penalized method for variable selection. Furthermore, the MSE of the proposed method is close to that of oracle empirical likelihood, especially so for larger sample sizes and larger models. This confirms the efficiency results in Theorem 3 empirically. Finally, using the correct working structure gives more efficient estimates, which can be seen by the smaller MSE's when the true correlation is used in Table 1. This agrees with Qu et al. (2000).

In addition, we record the average correctly estimated zero coefficients and the average numbers of incorrectly estimated zero coefficients for penalized empirical likelihood. The results are summarized in Table 1. The model selection result is satisfactory. As n increases, the average correctly estimated zero coefficients is approaching $p - 3$, while the average numbers of incorrectly estimated zero coefficients is 0 throughout. This confirms the selection consistency in Theorem 3.

To verify the penalized empirical likelihood ratio result in Theorem 4, we test the null hypothesis $H_0 : \beta_1 = a$ for $a = 2.8, 2.9, 3.0, 3.1, 3.2$ respectively, where β_1 is the first component of β . Using a nominal level $\alpha = 0.05$, we document the empirical size and power results in Table

Table 1. Mean square errors ($\times 10^{-2}$) for estimating equations in longitudinal data analysis. The largest standard error over the mean is 2.35

n	p	True	Working	LS	O-LS	EL	O-EL	PEL	C	IC
50	6	CS	CS	6.66	2.47	5.52	1.71	1.98	2.75	0
		CS	AR(1)	-	-	5.38	1.86	2.37	2.70	0
		AR(1)	CS	6.55	2.44	5.34	1.70	2.38	2.72	0
		AR(1)	AR(1)	-	-	5.35	1.80	2.38	2.74	0
100	10	CS	CS	5.54	1.25	4.21	0.63	1.13	6.59	0
		CS	AR(1)	-	-	4.22	0.76	1.44	6.52	0
		AR(1)	CS	5.44	1.25	4.07	0.75	1.38	6.59	0
		AR(1)	AR(1)	-	-	3.92	0.76	1.28	6.61	0
200	15	CS	CS	4.16	0.61	2.85	0.28	0.53	11.69	0
		CS	AR(1)	-	-	3.03	0.41	0.67	11.63	0
		AR(1)	CS	4.14	0.63	2.95	0.35	0.64	11.68	0
		AR(1)	AR(1)	-	-	2.96	0.34	0.61	11.67	0
400	20	CS	CS	2.74	0.31	1.95	0.19	0.19	16.91	0
		CS	AR(1)	-	-	2.08	0.21	0.25	16.86	0
		AR(1)	CS	2.74	0.31	2.05	0.16	0.25	16.85	0
		AR(1)	AR(1)	-	-	2.02	0.18	0.23	16.86	0

LS, least-squares; O-LS, oracle least-squares; EL, empirical likelihood; O-EL, oracle empirical likelihood; PEL, penalized empirical likelihood; C, the average of correctly estimated zeros; IC, the average of incorrectly estimated zeros; CS, compound symmetry

2. We can see clearly that the size of the test is close to 0.05 as the sample size increases and the power goes to 1 as either the sample size increases or a deviates more from the true $\beta_1 = 3$. These results show that the proposed test statistic performs satisfactorily.

Example 2. Consider a multivariate extension of Example 1 in Qin & Lawless (1994). Let the j th variable be

$$X_j \sim N(\theta_j, \theta_j^2 + 0.1), \quad (j = 1, \dots, p)$$

where $\theta = (\theta_1, \dots, \theta_p)^T = (1, -1, 0, 0, 1, 0, \dots, 0)^T$. We consider the following estimating equations (Qin & Lawless, 1994)

$$g_1(X, \theta) = \begin{pmatrix} X_1 - \theta_1 \\ \vdots \\ X_p - \theta_p \end{pmatrix}, \quad g_2(X, \theta) = \begin{pmatrix} X_1^2 - 2\theta_1^2 - 0.1 \\ \vdots \\ X_p^2 - 2\theta_p^2 - 0.1 \end{pmatrix}.$$

We generate $x_i \in \mathbb{R}^p$ ($i = 1, \dots, n$) from p -dimensional normal distribution with mean θ and the AR(1) correlation matrix with parameter 0.5. The marginal variance matrix is a diagonal matrix with entries $\theta_j^2 + 0.1$. To consider the scenario of a diverging dimensionality, we let p be the integer part of $20n^{1/5.1} - 36$ and consider several sample sizes. To make a comparison, we compute mean square errors of the usual sample mean, the oracle sample mean assuming that the zero entries in θ were known, empirical likelihood estimate without penalization, the oracle empirical likelihood estimator by using the estimating equations only for the nonzero entries, and finally the proposed penalized empirical likelihood. The sample mean estimator can be seen as using g_1 only in the estimating equation. Note that the oracle empirical likelihood estimate is suboptimal because the estimating equations for the zero entries can be exploited to

Table 2. Size and power for testing $H_0 : \beta_1 = 3$. The nominal level is 0.05

n	p	True	Working	2.8	2.9	3.0	3.1	3.2
50	6	CS	CS	0.87	0.43	0.12	0.43	0.87
		CS	AR(1)	0.83	0.38	0.11	0.39	0.84
		AR(1)	CS	0.83	0.38	0.12	0.33	0.79
		AR(1)	AR(1)	0.85	0.41	0.11	0.35	0.83
100	10	CS	CS	0.98	0.58	0.09	0.58	0.98
		CS	AR(1)	0.96	0.55	0.10	0.53	0.96
		AR(1)	CS	0.96	0.52	0.09	0.50	0.96
		AR(1)	AR(1)	0.96	0.55	0.09	0.54	0.96
200	15	CS	CS	1.00	0.83	0.09	0.83	1.00
		CS	AR(1)	1.00	0.78	0.08	0.77	1.00
		AR(1)	CS	1.00	0.76	0.07	0.75	1.00
		AR(1)	AR(1)	1.00	0.80	0.08	0.77	1.00
400	20	CS	CS	1.00	0.99	0.07	0.99	1.00
		CS	AR(1)	1.00	0.97	0.07	0.97	1.00
		AR(1)	CS	1.00	0.97	0.07	0.96	1.00
		AR(1)	AR(1)	1.00	0.98	0.08	0.97	1.00

LS, least-squares; O-LS, oracle least-squares; EL, empirical likelihood; O-EL, oracle empirical likelihood; PEL, penalized empirical likelihood; C, the average of correctly estimated zeros; IC, the average of incorrectly estimated zeros; CS, compound symmetry

improve the efficiency of nonzero entries. This phenomenon was also noted in Tang & Leng (2010). The results from 1000 replications for each sample size are summarized in Table 3. We see that empirical likelihood is more accurate than sample mean, because g_2 is incorporated. The penalized empirical likelihood has the smallest MSE's, because $X_j - \theta_j$ from the zero entries can be exploited to improve the efficiency of the estimates for the nonzero entries. The model selection results are provided in Table 4. We see that variable selection is satisfactory as the average number of correctly estimated zeros is close to $p - 3$.

To verify the result in Theorem 4, we test the null hypothesis $H_0 : \theta_1 = a$ for $a = 0.8, 0.9, 1.0, 1.1, 1.2$ respectively. Using a nominal level $\alpha = 0.05$, we document the empirical size and power results in Table 5. We can see clearly that the size of the test is close to 0.05 as the sample size increases and the power goes to 1 as either the sample size increases or a deviates more from the true $\theta_1 = 3$, especially when the hypothesized value is less than the true value. These results show that the proposed empirical likelihood test statistic performs satisfactorily.

Example 3. The instrumental variable method is widely used in measurement error models (Fuller, 1987), survey sampling (Fuller, 2009) and econometrics (Hansen & Singleton, 1982). Briefly speaking, this approach starts from conditional moment condition $E\{h(X_i; \theta)|\mathcal{F}\} = 0$ where $h(X_i; \theta) \in \mathbb{R}^q$ and \mathcal{F} is information generated by data. Hence, for any \mathcal{F} -measurable variable $U_i \in \mathbb{R}^{r \times q}$, so-called the instrument variables, $E\{U_i h(Z_i; \theta)\} = 0$. Then θ can be estimated using $g(Z_i; \theta) = U_i h(X_i; \theta)$ as estimating equations. Since the dimensionality of U_i is not restricted, this approach is an estimating equations method with $r \geq p$.

We consider the model $y_i = x_i^T \beta + \varepsilon_i$, where two noisy copies of x_i denoted by u_i and v_i instead of x_i , and y_i are observed. We follow the classical measurement error model assumption Fuller (1987) by assuming $u_i = x_i + e_{1i}$ and $v_i = x_i + e_{2i}$, where e_{1i} and e_{2i} are p -dimensional

Table 3. *Mean squares errors* ($\times 10^{-2}$)

Heterogeneity in Variance Example						
n	p	SM	O-SM	EL	O-EL	PEL
50	7	5.85	5.07	4.97	3.63	3.12
100	13	3.53	2.54	2.96	1.30	1.23
200	20	2.17	1.33	1.78	0.58	0.50
400	28	1.29	0.66	1.11	0.27	0.23
Instrumental Variable Example						
n	p	LS	O-LS	EL	O-EL	PEL
50	8	44.4	15.4	41.8	8.97	21.9
100	16	43.2	9.66	41.5	3.96	11.6
200	25	34.1	6.96	30.1	1.81	5.01
400	35	24.7	5.53	20.4	0.83	1.90
Two-sample Example						
n	p	SM	O-SM	EL	O-EL	PEL
50	8	16.1	6.14	12.5	3.51	6.25
100	16	16.0	2.95	12.9	1.55	3.98
200	25	12.5	1.51	9.87	0.75	1.61
400	35	8.76	0.75	6.97	0.39	0.61

EL, empirical likelihood; O-EL, oracle empirical likelihood; PEL, penalized empirical likelihood; SM, sample mean; O-SM, oracle sample mean; LS, least squares; O-LS, oracle least squares

Table 4. *Model selection results for examples*

n	Example 2			Example 3			Example 4		
	p	C	IC	p	C	IC	p	C	IC
50	7	3.57	0	8	3.99	0	8	4.04	0
100	13	9.63	0	16	11.4	0	16	11.5	0
200	20	16.9	0	25	20.8	0	25	21.1	0
400	28	25.0	0	35	31.4	0	35	31.6	0

C, the average of correctly estimated zeros; IC, the average of incorrectly estimated zeros

Table 5. *Size and power for testing $H_0 : \theta_1 = 1$ in Example 2. The nominal level is 0.05*

n	p	0.8	0.9	1.0	1.1	1.2
50	7	0.70	0.30	0.12	0.32	0.62
100	13	0.82	0.33	0.11	0.29	0.54
200	20	0.92	0.52	0.09	0.34	0.51
400	28	0.97	0.63	0.08	0.36	0.53

mean zero random vector independent of each other and independent of ε_i . Via instrumental variables, we formulate two sets of estimating equations as

$$g_1(U, V, Y, \beta) = U^T(Y - V^T\beta), \quad g_2(U, V, Y, \beta) = V^T(Y - U^T\beta).$$

It is known that the ordinary least squares estimates are usually biased (Fuller, 1987). We generate β and x according to Example 1, whereas e_1 and e_2 are generated from a multivariate normal with mean zero and exchangeable correlation matrix with parameter 0.5. Each component of e_1 and e_2 has marginal variance 0.04. Furthermore, we generate ε from $N(0, 0.25)$. To make comparisons, we compute the mean square errors for the ordinary least squares estimate using U

481 as X , the oracle least squares estimate using the sub-vector of U corresponding to the nonzero
 482 component of X , the empirical likelihood estimator, the oracle empirical likelihood estimator
 483 and the penalized empirical likelihood. Note that least squares uses $g(U, Y, \beta) = U^T(Y - U^T\beta)$
 484 as the estimating equation and that both least squares and oracle least squares give biased esti-
 485 mates due to the measurement error. The results on MSE are summarized in Table 3. We see that
 486 penalized empirical likelihood is much more accurate than empirical likelihood. For large n , the
 487 MSEs of empirical likelihood is closer to those of oracle empirical likelihood, indicating that the
 488 proposed method is closer to the oracle for large sample sizes. In addition, our method performs
 489 satisfactorily in variable selection, as can be seen from Table 4. We also conducted hypothesis
 490 testing using the null hypothesis in Example 1. The results are similar to that in Example 1 and
 491 are omitted to save space.

492 **Example 4.** We consider the two sample problem with common means in Qin & Lawless (1994).
 493 In particular, we have a pair of random variables (X_j, Y_j) such that $E(X_j) = E(Y_j) = \theta_j$ ($j =$
 494 $1, \dots, p$). We set $\theta = (\theta_1, \dots, \theta_p)^T = (1, -1, 0, 0, 0.5, 0, \dots, 0)^T$. We generate x_i and y_i inde-
 495 pendently from p -dimensional multivariate normal distribution with mean θ and an AR(1) co-
 496 variance matrix with parameter 0.5. We take p as the integer part of $25n^{1/5.1} - 45$. We compare
 497 the following estimators: the sample mean, the oracle sample mean, the empirical likelihood
 498 estimate, the oracle empirical likelihood estimate and the penalized empirical likelihood. Once
 499 again, we see from Table 3 that the proposed method gives MSEs close to the oracle estimator, es-
 500 pecially when the sample size becomes large. In addition, penalized empirical likelihood is much
 501 more accurate than the usual empirical likelihood. This indicates that variable selection can en-
 502 hance estimation accuracy if the underlying model is sparse. The penalized empirical likelihood
 503 performs well in variable selection, as can be seen from Table 4.

504 **Higher dimensionality.** Since the proposed method is based on empirical likelihood, it is not
 505 possible to allow p or r greater than n . Otherwise, empirical likelihood can not be applied. To
 506 explore higher dimensionality problems, we fix the sample size to be 100 and investigate the
 507 performance of the method for Example 2 to 4 with p ranging from 10 to 25 (r ranging from 20
 508 to 50). The results are presented in Figure 1. Clearly, with higher dimensions, the performance
 509 of the proposed method deteriorates especially when $p > 15$. However, the proposed method
 510 always outperform the empirical likelihood method with no penalization. We note additionally
 511 that with $r = 2p$ estimating equations when $p \geq 30$, the optimization of empirical likelihood
 512 can be unstable and sometimes may fail, a phenomenon observed by Tsao (2004) and Grendár
 513 & Judge (2009). Therefore penalized empirical likelihood still performed reasonably well with
 514 larger p while caution needs to be taken when the number of estimating equations is too large
 515 comparing to the sample size.

516 **Example 5.** To illustrate the usefulness of penalized empirical likelihood, we consider the CD4
 517 data (Diggle et al., 2002) where there are 2,376 observations for 369 subjects ranging from
 518 3 years to 6 years after seroconversion. The major objective is to characterize the population
 519 average time course of CD4 decay while accounting for the following predictor variables: age
 520 (in years), smoking (packs per day), recreational drug use (yes or no), number of sexual partners,
 521 and depression symptom score (larger values indicate more severe depression symptoms). As in
 522 Diggle et al. (2002), we consider the square-root-transformed CD4 numbers whose distribution
 523 is more near Gaussian. We parametrize the variable time by using a piecewise polynomial

$$524 \quad f(t) = a_1 t + a_2 t^2 + a_3 (t - t_1)_+^2 + \dots + a_8 (t - t_6)_+^2$$

526 where $t_0 = \min(t_{ij}) < t_1 < \dots < t_6 < t_7 = \max(t_{ij})$ are equally spaced points and $(t -$
 527 $t_j)_+^2 = (t - t_j)^2$ if $t \geq t_j$ and $(t - t_j)_+^2 = 0$ otherwise. This spline representation is motivated
 528 by the data analysis in Fan & Peng (2004). We normalize all the covariates such that their sam-

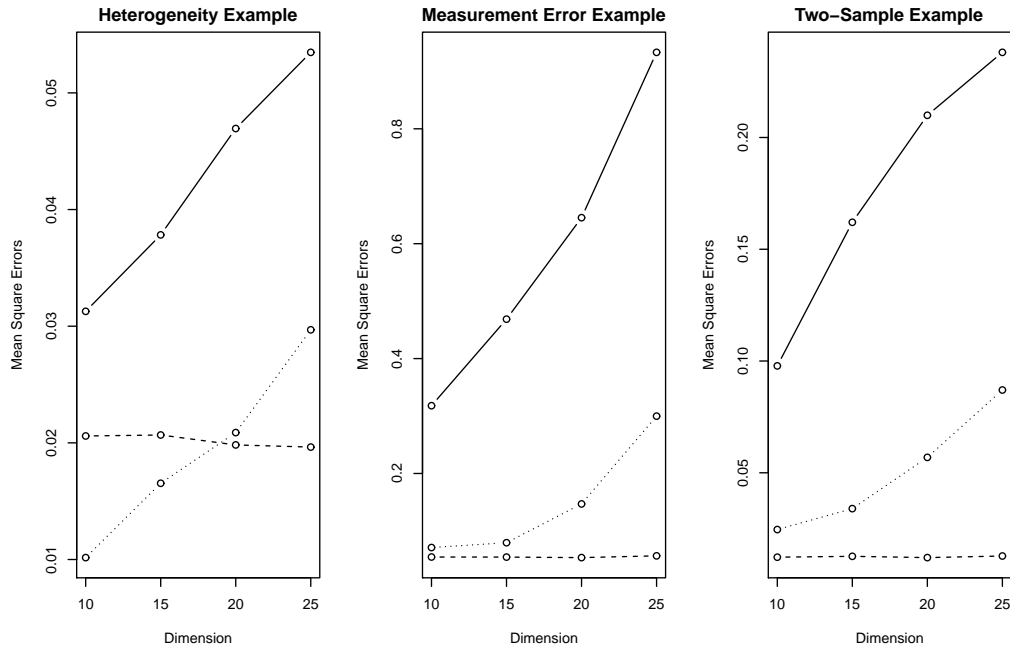


Fig. 1. Comparison of the mean squared errors using the empirical likelihood method (solid), the oracle empirical likelihood method (dashed) and the penalized empirical likelihood method (dotted).

ple means are zero and sample variance is one, which is routinely done in variable selection (Tibshirani, 1996).

We use the quadratic inference function method by using the compound symmetry and AR(1) matrices, respectively. In total there are 14 variables in the model and 28 estimating equations. The intercept is not penalized. We also combine the estimating equations which use the compound symmetry and AR(1) working structure. This gives a model with an additional 14 estimating equation. In total, there are 42 estimating equations for this estimator. The detail of the quadratic inference function modeling approach can be found in Example 1 and Qu et al. (2000). The fitted time curves of the square root of CD4 trajectory against time via the three penalized empirical likelihood, together with the unpenalized fits using independent, compound symmetry and AR(1) working correlation structures, are plotted in Figure 2. These curves are plotted when all the other covariates are fixed at zero. These curves show close agreement with the data points and with each other. The only exception is that if the working correlation is assumed to be independent, the fitted trajectory differs from other fitted curves for large time.

Table 6 gives the generalized estimating equation estimates using various working correlation matrices and the three penalized empirical likelihood estimates for the five variables. It is noted that all the estimates identify smoking as the important variable.

ACKNOWLEDGMENT

We are grateful to Professor Anthony Davison, an associate editor and a referee for constructive comments. Research supports from National University of Singapore research grants and National University of Singapore Risk Management Institute grants are gratefully acknowledged.

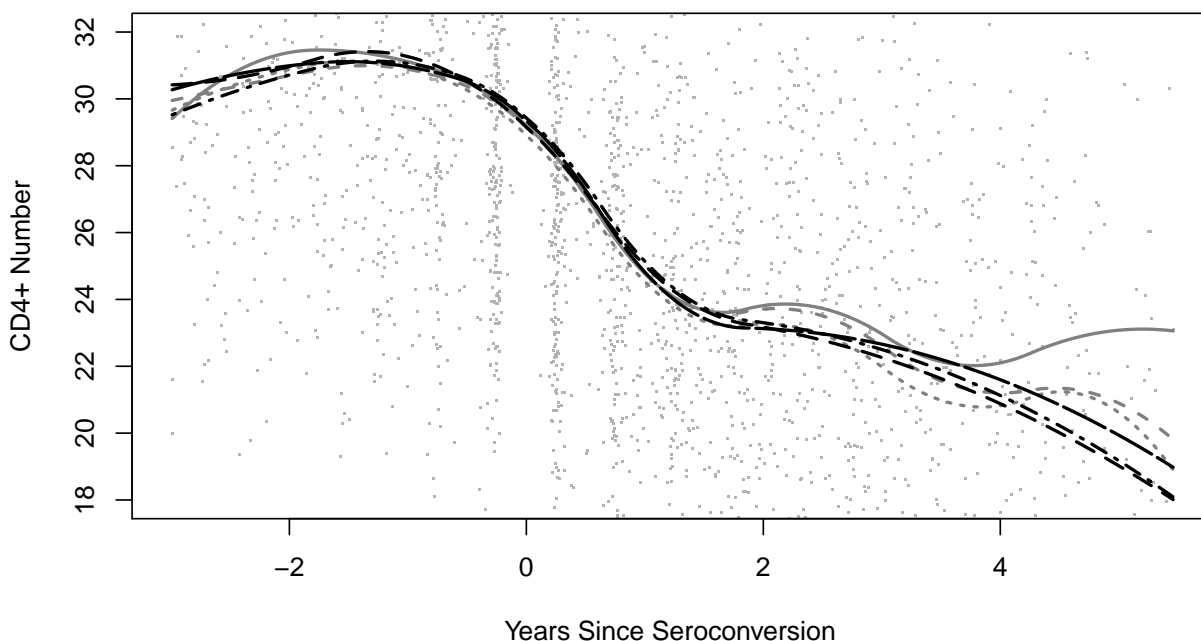


Fig. 2. The fits and the CD4 data: independent (gray solid), general estimating equations using compound symmetry correlations (gray long dash), general estimating equations using AR(1) correlations (gray short dash), penalized empirical likelihood using compound symmetry correlations (long dash), penalized empirical likelihood using AR(1) correlations (dash), penalized empirical likelihood using compound symmetry and AR(1) correlations (dot-dash).

Table 6. The fitted coefficients and their standard errors

Variable	Independence	CS	AR(1)	PEL-CS	PEL-AR1	PEL-CM
age	0.014 _(0.035)	0.002 _(0.032)	0.014 _(0.033)	0	0	0
smoking	0.981 _(0.184)	0.608 _(0.136)	0.281 _(0.190)	0.806	0.641	0.756
drug	1.064 _(0.529)	0.463 _(0.361)	0.414 _(0.356)	0	0	0
partner	-0.065 _(0.059)	0.059 _(0.042)	0.052 _(0.041)	0	0	0
depression	-0.032 _(0.021)	-0.048 _(0.015)	-0.047 _(0.015)	0	0	0

CS, compound symmetry; PEL-CS, penalized empirical likelihood using compound symmetry correlations; PEL-AR1, penalized empirical likelihood using AR(1) correlations; PEL-CM, penalized empirical likelihood using compound symmetry and AR(1) correlations

SUPPLEMENTARY MATERIAL

Supplementary Material available at *Biometrika* online includes the proofs of Lemmas 1-4 and Theorems 1-4, as well as quantile-quantile plots for demonstrating the empirical distributions of the estimated parameters in simulations.

APPENDIX

The Appendix sketches the main idea in the proofs of Theorems 1-4, and the important lemmas for the proofs.

Let $\ell(\theta, \lambda) = n^{-1} \sum_{i=1}^n \log\{1 + \lambda^T g_i(\theta)\}$, $\bar{g}(\theta) = n^{-1} \sum_{i=1}^n g_i(\theta)$. We present Lemmas 1-3 following the approach in Newey & Smith (2004), which is used in proving Theorem 1. The proofs of the lemmas are given in the Supplementary Material.

625 LEMMA 1. Under Conditions A.1, A.2 and A.4, for any ξ with $(1/\alpha + 1/10) \leq \xi < 2/5$ and
 626 as $n \rightarrow \infty$, then $\max_{1 \leq i \leq n} \sup_{\theta \in \Theta} |\lambda^T g(Z_i; \theta)| = o_p(1)$ for all $\lambda \in \Lambda_n = \{\lambda : \|\lambda\| \leq n^{-\xi}\}$,
 627 and $\Lambda_n \subseteq \hat{\Lambda}_n(\theta)$ for all $\theta \in \Theta$ where $\hat{\Lambda}_n(\theta) = \{\lambda : \lambda^T g_i(\theta) > -1, i = 1, \dots, n\}$.
 628

629 LEMMA 2. Under Conditions A.1-A.4, with probability tending to 1, $\lambda_{\theta_0} =$
 630 $\arg \max_{\lambda \in \hat{\Lambda}_n(\theta_0)} \ell(\lambda, \theta_0)$ exists, $\|\lambda_{\theta_0}\| = O_p(a_n)$, and $\sup_{\lambda \in \hat{\Lambda}_n(\theta_0)} \ell(\lambda, \theta_0) \leq O_p(a_n^2)$.
 631

632 LEMMA 3. Under Conditions A.1-A.4, $\|\bar{g}(\hat{\theta}_E)\|^2 = O_p(n^{-3/5})$.
 633

634 The proof of part a) of Theorem 1 follows the arguments in Newey & Smith (2004) by applying
 635 Lemmas 1-3, generalizing the results in Newey & Smith (2004) to allow diverging r and p . Upon
 636 establishing the consistent result in part a), the proof given in the Supplementary Material for part
 637 b) of Theorem 1 for the rate of convergence follows the arguments in Huang et al. (2008). The
 638 following Lemma 4 is used in proving Theorem 2:

639 LEMMA 4. Under Conditions A.1-A.5, $\|\lambda_{\hat{\theta}_E}\| = O_p(a_n)$.
 640

641 Given Theorem 1 and Lemma 4, stochastic expansions for $\hat{\theta}_E$ and the empirical likelihood
 642 ratio (3) can be developed, which facilitates the proof of Theorems 2-4. The proofs of Theorems
 643 2-4 are available in the Supplementary Material.
 644

645 REFERENCES

- 646 CHEN, S. X. & CUI, H. J. (2006). On Bartlett correction of empirical likelihood in the presence of nuisance param-
 647 eters. *Biometrika* **93**, 215-220.
 648 CHEN, S. X., PENG, L. & QIN, Y. L. (2009). Effects of data dimension on empirical likelihood. *Biometrika* **96**,
 649 711-722.
 650 CHEN, S. X. & VAN KEILEGOM, I. (2009). A review on empirical likelihood methods for regression (with discus-
 651 sion). *Test* **18**, 415-447.
 652 DICICCIO, T. J., HALL, P. & ROMANO, J. P. (1991). Empirical likelihood is Bartlett-correctable. *Ann. Statist.* **19**,
 653 1053-1061.
 654 DIGGLE, P. J., HEAGERTY, P., LIANG, K. Y. & ZEGER, S. L. (2002). *Analysis of Longitudinal Data*. 2nd Edition.
 655 New York: Oxford University Press.
 656 FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am.*
 657 *Statist. Assoc.* **96**, 1348-1360.
 658 FAN, J. & LV, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussions). *J.*
 659 *R. Statist. Soc. B* **70**, 849-911.
 660 FAN, J. & PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.*
 661 **32**, 928-961.
 662 FULLER, W. A. (1987). *Measurement Error Models*. New York: Wiley.
 663 FULLER, W. A. (2009). *Sampling Statistics*. New York: Wiley.
 664 GODAMBE, V. P. & HEYDE, C. C. (1987). Quasi-likelihood and optimal estimation. *Int. Statist. Rev.* **55**, 231-244.
 665 GRENDÁR, M. & JUDGE, G. (2009). Empty set problem of maximum empirical likelihood methods. *Elect. J. Statist.*
 666 **3**, 1542-1555.
 667 HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**,
 668 1029-1054.
 669 HANSEN, L. P. & SINGLETON, K. J. (1982). Generalized instrumental variables estimation of nonlinear rational
 670 expectation models. *Econometrica* **50**, 1269-1285.
 671 HJORT, N. L., MCKEAGUE, I. & VAN KEILEGOM, I. (2009). Extending the scope of empirical likelihood. *Ann.*
 672 *Statist.* **37**, 1079-1111.
 HUANG, J., HOROWITZ, J. L. & MA, S. (2008). Asymptotic properties of bridge estimators in sparse high-
 dimensional regression models. *Ann. Statist.* **36**, 587-613.
 HEYDE, C. C. & MORTON, R. (1993). On constrained quasi-likelihood estimation. *Biometrika* **80**, 755-761.
 LIANG, K. Y. & ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**,
 13-22.
 LV, J. & FAN, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares.
Ann. Statist. **37**, 3498-3528.

- 673 NEWWEY, W. K. & SMITH, R. J. (2004). Higher order properties of gmm and generalized empirical likelihood
674 estimators. *Econometrica* **72**, 219–255.
- 675 OWEN, A. B. (2001). *Empirical Likelihood*. New York: Chapman and Hall-CRC.
- 676 QIN, J. & LAWLESS, J. (1994). Empirical likelihood and generalized estimating equations. *Ann. Statist.* **22**, 300–325.
- 677 QU, A., LINDSAY, B. G. & LI, B. (2000). Improving estimating equations using quadratic inference functions.
678 *Biometrika* **87**, 823–836.
- 679 TANG, C. Y. & LENG, C. (2010). Penalized high dimensional empirical likelihood. *Biometrika* **97**, 905–920.
- 680 TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* **58**, 267–288.
- 681 TSAO, M. (2004). Bounds on coverage probabilities of the empirical likelihood ratio confidence regions. *Ann. Statist.*
682 **32**, 1215–1221.
- 683 WANG, H. & LENG, C. (2007). Unified lasso estimation via least squares approximation. *J. Am. Statist. Assoc.* **101**,
684 1039–1048.
- 685 WANG, H., LI, B. & LENG, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters.
686 *J. R. Statist. Soc. B* **71**, 671–683.
- 687 XIE, M. & YANG, Y. (2003). Asymptotics for generalized estimating equations with large cluster sizes. *Ann. Statist.*
688 **31**, 310–347.
- 689 ZHANG, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894–942.
- 690 ZHANG, H. H. & LU, W. (2007). Adaptive lasso for Cox’s proportional hazard model. *Biometrika* **94**, 691–703.
- 691 ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418–1429.

[Received XXXX 0000. Revised XXXX 0000]

692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720