A New Scope of Penalized Empirical Likelihood with High-Dimensional Estimating Equations

Jinyuan Chang

University of Melbourne

Homepage: www.jinyuanchang.com

This is a joint work with Cheng Yong Tang and Tong Tong Wu

June 14, 2016

Model setting

- $\mathbf{X}_1, \ldots, \mathbf{X}_n$: d-dimensional i.i.d. observations.
- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$: p-dimensional parameter.
- $\mathbf{g}(\mathbf{X}; \boldsymbol{\theta}) = (g_1(\mathbf{X}; \boldsymbol{\theta}), \ldots, g_r(\mathbf{X}; \boldsymbol{\theta}))^{\mathrm{T}}$: an r-dimensional estimating function.
- The true parameter $\boldsymbol{\theta}_0$ is determined by the moment condition

$$\mathbb{E}\{\mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta}_0)\} = \mathbf{0}.$$

- The goal of this paper is to construct the estimation of $\boldsymbol{\theta}_0$ when p and/or $r \gg n$.

Literature review

Both p and r are fixed:

- Hansen (1982): Generalized Methods of Moments (GMM).
- Qin and Lawless (1994): Using the idea of Empirical Likelihood (EL) proposed by Owen (1988, 1990).
- Newey and Smith (2004): Investigate higher order properties of GMM and generalized EL.

Some other advantages of EL: Wilks' theorems, Bartlett-correctable, ...

- Define an EL with estimating equations as

$$L(\boldsymbol{\theta}) = \sup\left\{ \prod_{i=1}^{n} \pi_i : \pi > 0,\ \sum_{i=1}^{n} \pi_i = 1,\ \sum_{i=1}^{n} \pi_i \mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta}) = \mathbf{0} \right\},$$

  and estimate $\boldsymbol{\theta}_0$ as $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta})$, which is equivalent to

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \max_{\boldsymbol{\lambda} \in \widehat{\Lambda}_n(\boldsymbol{\theta})} \sum_{i=1}^{n} \log\{1 + \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta})\},$$

  where $\widehat{\Lambda}_n(\boldsymbol{\theta}) = \{\boldsymbol{\lambda} \in \mathbb{R}^r : \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta}) \in \mathcal{V}, i = 1, \ldots, n\}$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\mathcal{V}$ is an open interval containing zero.
- $\widehat{\boldsymbol{\theta}}$ is called the maximum EL estimator.

When p or/and r diverging,

- Leng and Tang (2012), and Chang, Chen and Chen (2015) investigated the asymptotic behaviors of the EL estimator $\widehat{\boldsymbol{\theta}}$ in high-dimensional settings.
- Hjort, McKeague and Van Keilegom (2009), Chen, Peng and Qin (2009), and Tang and Leng (2010) considered EL for some other high-dimensional problems.
- However, EL only works when both p and r are growing at some rate slower than n. Since $|\boldsymbol{\lambda}|_2$ is required to be $o_p(1)$ in existing analyses of EL.

Such kind of restrictions on the diverging rates of p and r do not match the real practical ultra high-dimensional problems.

New scope of our work

- The relationship between r and p in the new high-dimensional paradigm with sparsity.
  - $r \geq p$ is no longer required, facilitated with penalized EL.
- Penalizing the Lagrange multiplier $\boldsymbol{\lambda}$:
  - to obtain sparse $\boldsymbol{\lambda}$ with many zero components;
  - to effectively reduce the effective dimensionality of the number of estimating functions.
- Solving the problem of EL with ultra high-dimensional parameters and estimating functions.

Why EL works?

Proposition 1

Assume that there exist uniform constants $C_1 > 0$, $C_2 > 1$ *and* $\gamma > 2$ *such that*

$$\max_{1 \le j \le r} \mathbb{E}\left\{ \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |g_j(\mathbf{X}_i; \boldsymbol{\theta})|^{\gamma} \right\} \le C_1,$$

and

$$\mathbb{P}\Bigg[C_2^{-1} \le \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \lambda_{\min}\left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta}) \mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta})^{\mathrm{T}} \right\} \le \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \lambda_{\max}\left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta}) \mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta})^{\mathrm{T}} \right\} \le C_2 \Bigg] \to 1.$$

If $r = o(n^{1/2 - 1/\gamma})$, *then the maximum EL estimator* $\widehat{\boldsymbol{\theta}}$ *satisfies* $|\bar{\mathbf{g}}(\widehat{\boldsymbol{\theta}})|_2 = O_p(r^{1/2} n^{-1/2})$ *where* $\bar{\mathbf{g}}(\widehat{\boldsymbol{\theta}}) = n^{-1} \sum_{i=1}^{n} \mathbf{g}(\mathbf{X}_i; \widehat{\boldsymbol{\theta}})$.

- Proposition 1 shows that regardless the dimensionality p of the parameter $\boldsymbol{\theta}$, with r unbiased estimating functions, $\widehat{\boldsymbol{\theta}}$ always ensures that $|\bar{\mathbf{g}}(\widehat{\boldsymbol{\theta}})|_2$ to be of $O_p(r^{1/2}n^{-1/2})$.
- If $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\bar{\mathbf{g}}(\boldsymbol{\theta}) - \mathbb{E}\{\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})\}|_\infty \to_p 0$ holds, we have $|\mathbb{E}\{\mathbf{g}(\mathbf{X}_i;\widehat{\boldsymbol{\theta}})\}|_\infty = o_p(1)$ if $r = o(n^{1/2-1/\gamma})$. With the unique identification condition, $\boldsymbol{\theta}_0$ is the unique solution of $\mathbb{E}\{\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})\} = \mathbf{0}$ (which is usually guaranteed by the condition $r \geq p$), we can obtain the consistency of $\widehat{\boldsymbol{\theta}}$.
- However, $\widehat{\boldsymbol{\theta}}$ is clearly not uniquely defined when $r < p$ rendering inapplicability of the method in practice.

- To resolve the ambiguity in the estimator $\widehat{\boldsymbol{\theta}}$, we take the approach that searches for a sparse optimizer with appropriate penalization on $\boldsymbol{\theta}$ under extra assumption that $\boldsymbol{\theta}_0$ is sparse. Write $\boldsymbol{\theta}_0 = (\theta_1^0, \ldots, \theta_p^0)^{\mathrm{T}}$ and let $\mathcal{S} = \{1 \leq k \leq p : \theta_k^0 \neq 0\}$ with $s = |\mathcal{S}|$. Sparsity means $s \ll n$. Without lose of generality, we write $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{0,(1)}^{\mathrm{T}}, \boldsymbol{\theta}_{0,(2)}^{\mathrm{T}})^{\mathrm{T}}$ where $\boldsymbol{\theta}_{0,(1)} \in \mathbb{R}^s$ being nonzero components and $\boldsymbol{\theta}_{0,(2)} = \mathbf{0} \in \mathbb{R}^{p-s}$.
- Let

$$\widetilde{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \max_{\boldsymbol{\lambda} \in \widehat{\Lambda}_n(\boldsymbol{\theta})} \left[\sum_{i=1}^{n} \log\{1 + \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta})\} + n \sum_{k=1}^{p} p_{1,\pi}(|\theta_k|) \right],$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_r)^{\mathrm{T}}$, and $p_{1,\pi}(\cdot)$ is a penalty function with tuning parameter π.

Proposition 2

Let $a_n = \sum_{k=1}^{p} p_{1,\pi}(|\theta_k^0|)$ *and* $b_n = \max\{rn^{-1}, a_n\}$. *Let* $\boldsymbol{\Theta}_* = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : |\boldsymbol{\theta} - \boldsymbol{\theta}_0|_\infty \leq \epsilon\}$ *for some sufficiently small* $\epsilon > 0$. *Assume that*

$$\max_{1 \leq j \leq r} \mathbb{E}\left\{ \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_*} |g_j(\mathbf{X}_i; \boldsymbol{\theta})|^\gamma \right\} \leq C_1,$$

$$\mathbb{P}\Bigg[C_2^{-1} \leq \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_*} \lambda_{\min}\left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta}) \mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta})^{\mathrm{T}} \right\} \leq \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_*} \lambda_{\max}\left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta}) \mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta})^{\mathrm{T}} \right\} \leq C_2 \Bigg] \to 1.$$

and

$$\inf_{\boldsymbol{\theta} \in \{\boldsymbol{\theta} \in \boldsymbol{\Theta}_* : |\boldsymbol{\theta} - \boldsymbol{\theta}_0|_\infty > \varepsilon\}} |\mathbb{E}\{\mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta})\}|_\infty \geq \Delta(\varepsilon)$$

for any $0 < \varepsilon < \epsilon$, *where* $\Delta(\cdot)$ *is a positive function satisfying* $\liminf_{\varepsilon \to 0^+} \varepsilon^{-\beta} \Delta(\varepsilon) \geq C_3$ *for some uniform constants* $C_3 > 0$ *and* $\beta > 0$. *If* $r = o(n^{1/2 - 1/\gamma})$ *and* $b_n = o(n^{-2/\gamma})$, *then there exists a local solution* $\widetilde{\boldsymbol{\theta}}_n \in \boldsymbol{\Theta}_*$ *such that* $|\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0|_\infty = O_p\{b_n^{1/(2\beta)}\}$.

- Proposition 2 implies that if facilitated with sparsity of the model parameter, the consistency of a p-dimensional estimator does not necessarily require $r \geq p$. In particular, under the local identification condition near the truth, a consistent and sparse estimator of $\boldsymbol{\theta}_0$ is feasible with no explicit requirement on the relationship between r and p, as long as the r unbiased estimating functions provide information for the zero and nonzero components in $\boldsymbol{\theta}_0$.
- This result ensures the important flexibility in the number of estimating functions for obtaining a sparse estimator using penalized EL. That is, to obtain a sparse estimator in penalized EL whose nonzero components are consistent to the truth, one can actually opt to use smaller number of estimating equations. Further, we can show that $\widetilde{\boldsymbol{\theta}}_n$ can estimate zero components of $\boldsymbol{\theta}_0$ as zero with a high probability as in the following proposition.

Proposition 3

Under the same conditions as in Proposition 2, if we further assume $g_j(\mathbf{X};\boldsymbol{\theta})$ *to be continuously differentiable with respect to* $\boldsymbol{\theta} \in \boldsymbol{\Theta}_*$ *for any* $\mathbf{X}$ *and* $j = 1, \ldots, r$ *satisfying the condition*

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_*} \max_{1\le j\le r} \max_{k\notin\mathcal{S}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left| \frac{\partial g_j(\mathbf{X}_i;\boldsymbol{\theta})}{\partial\theta_k} \right| \right\} = O_p(\varphi_n)$$

for some $\varphi_n > 0$*, which may diverge with* n*. Suppose there exist a constant* $0 < c < 1$ *and* $\chi_n \to 0$ *such that*

$$\max_{k\in\mathcal{S}} \sup_{c|\theta_k^0|<t<c^{-1}|\theta_k^0|} p'_{1,\pi}(t) = O(\chi_n).$$

In addition to the restrictions imposed on r *and* b_n *for the consistency of* $\widetilde{\boldsymbol{\theta}}_n$ *above, if* r *and* b_n *also satisfy the restriction conditions such that* $b_n = o(\min_{k\in\mathcal{S}} |\theta_k^0|^{2\beta})$*,* $s\chi_n b_n^{1/(2\beta)} = O(rn^{-1})$ *and* $rn^{-1/2}\varphi_n = o(\pi)$*, then* $\mathbb{P}\{\widetilde{\boldsymbol{\theta}}_{n,(2)} = \mathbf{0}\} \to 1$ *where* $\widetilde{\boldsymbol{\theta}}_{n,(2)}$ *is the corresponding estimation for the zero components of* $\boldsymbol{\theta}_0$ *in* $\widetilde{\boldsymbol{\theta}}_n$*.*

- Though Propositions 2 and 3 are general, there are still restrictions because conditions are violated when r is large. With Proposition 1 ensuring the behavior of the estimating equations at the maximum EL estimate, and Propositions 2 and 3 breaking the requirement of $r \geq p$ for a consistent and sparse estimator $\widetilde{\boldsymbol{\theta}}_n$.
- Define

$$\widehat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \max_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\boldsymbol{\theta})} \Bigg[\sum_{i=1}^{n} \log\{1 + \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta})\} - n \sum_{j=1}^{r} p_{2,\nu}(|\lambda_j|) + n \sum_{k=1}^{p} p_{1,\pi}(|\theta_k|) \Bigg],$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_r)^{\mathrm{T}}$, and $p_{1,\pi}(\cdot)$ and $p_{2,\nu}(\cdot)$ are two penalty functions with tuning parameters π and ν, respectively.

Main results – Consistency

Condition 1

There exist some $K_1 > 0$ *and* $\gamma > 4$ *such that*

$$\max_{1 \le j \le r} \mathbb{E}\left\{ \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |g_j(\mathbf{X}_i; \boldsymbol{\theta})|^{\gamma} \right\} \le K_1.$$

Condition 2

There exists a positive function $\Delta(\cdot)$ *such that*

$$\inf_{\boldsymbol{\theta} \in \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : |\boldsymbol{\theta} - \boldsymbol{\theta}_0|_\infty > \varepsilon\}} |\mathbb{E}\{\mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta})\}|_\infty \ge \Delta(\varepsilon)$$

for any $\varepsilon > 0$, *and there exist* $K_2 > 0$ *and* $\beta > 0$ *such that* $\liminf_{\varepsilon \to 0+} \varepsilon^{-\beta} \Delta(\varepsilon) \ge K_2$.

- For $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\boldsymbol{\lambda} \in \widehat{\Lambda}_n(\boldsymbol{\theta})$, we define

$$f(\boldsymbol{\lambda}; \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \log\{1 + \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta})\} - \sum_{j=1}^{r} p_{2,\nu}(|\lambda_j|),$$

$$S_n(\boldsymbol{\theta}) = \max_{\boldsymbol{\lambda} \in \widehat{\Lambda}_n(\boldsymbol{\theta})} f(\boldsymbol{\lambda}; \boldsymbol{\theta}) + \sum_{k=1}^{p} p_{1,\pi}(|\theta_k|).$$

- Then $\widehat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} S_n(\boldsymbol{\theta})$.
- Let $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\lambda} \in \widehat{\Lambda}_n(\boldsymbol{\theta})} f(\boldsymbol{\lambda}; \boldsymbol{\theta})$ be the Lagrange multiplier defined at $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. For any subset $\mathcal{A} \subset \{1, \ldots, r\}$, we denote by $\mathbf{g}_{\mathcal{A}}(\mathbf{X}_i; \boldsymbol{\theta})$ the subvector of $\mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta})$ with components indexed by $\mathcal{A}$. We write $\bar{\mathbf{g}}_{\mathcal{A}}(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \mathbf{g}_{\mathcal{A}}(\mathbf{X}_i; \boldsymbol{\theta})$, $\widehat{V}_{\mathcal{A}}(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \mathbf{g}_{\mathcal{A}}(\mathbf{X}_i; \boldsymbol{\theta}) \mathbf{g}_{\mathcal{A}}(\mathbf{X}_i; \boldsymbol{\theta})^{\mathrm{T}}$ and $V_{\mathcal{A}}(\boldsymbol{\theta}_0) = \mathbb{E}\{\mathbf{g}_{\mathcal{A}}(\mathbf{X}_i; \boldsymbol{\theta}_0) \mathbf{g}_{\mathcal{A}}(\mathbf{X}_i; \boldsymbol{\theta}_0)^{\mathrm{T}}\}$.

Proposition 4

Suppose that Condition 1 *hold, and the penalty function* $p_{2,\nu}(\cdot)$ *is a convex function. Assume that* $\max_{1\leq j\leq r} n^{-1}\sum_{i=1}^{n}|g_j(\mathbf{X}_i;\boldsymbol{\theta}_0)|^2 = O_p(\varpi_n)$ *for some* $\varpi_n > 0$ *that may diverge with* n*, and* $\min_{1\leq j\leq r}\mathbb{E}\{|g_j(\mathbf{X}_i;\boldsymbol{\theta}_0)|^2\}$ *is uniformly bounded away from zero. If* $(n^{-1}\varpi_n \log r)^{1/2}/\nu \to 0$ *and* $\log r = o(\min\{n^{1/3}, n^{1-2/\gamma}\varpi_n^{-1}\})$*, then* $\mathbb{P}\{S_n(\boldsymbol{\theta}_0) = \sum_{k=1}^{p} p_{1,\pi}(|\theta_k^0|)\} \to 1$ *as* $n \to \infty$ *and* $\boldsymbol{\lambda} = \mathbf{0}$ *is a local maximizer of* $f(\boldsymbol{\lambda};\boldsymbol{\theta}_0)$ *with probability approaching one.*

- Based on the convexity of $p_{2,\nu}(\cdot)$, $f(\boldsymbol{\lambda};\boldsymbol{\theta})$ is concave w.r.t. $\boldsymbol{\lambda}$. To compute $S_n(\boldsymbol{\theta})$, we only need to find a local maximizer for $f(\boldsymbol{\lambda};\boldsymbol{\theta})$.
- Moreover, for any $\boldsymbol{\theta}$ such that $|\boldsymbol{\theta}-\boldsymbol{\theta}_0|_\infty > \varepsilon_n$, Condition 2 indicates $|\mathbb{E}\{\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})\}|_\infty \geq \Delta(\varepsilon_n)$. We show in our proof that when $\boldsymbol{\theta}$ takes value departing from the truth $\boldsymbol{\theta}_0$, i.e., $\Delta(\varepsilon_n)$ decays to zero at some slow enough rate, $S_n(\boldsymbol{\theta})$ takes a value larger than $\varrho S_n(\boldsymbol{\theta}_0)$ for some diverging ϱ with probability tending to 1; see also Chang, Tang and Wu (2013, 2016) for such a phenomenon in marginal EL. This feature, together with the Proposition 4, leads to the consistency of the penalized EL estimator $\widehat{\boldsymbol{\theta}}_n$.

Theorem 1

Let $a_n = \sum_{k=1}^p p_{1,\pi}(|\theta_k^0|)$. *Assume that conditions in Proposition* 4 *and Condition* 2 *hold. Define* $b_n = \max\{n^{-1}, a_n, \nu^2\}$. *If* $b_n = o(n^{-2/\gamma})$, *then* $|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0|_\infty = O_p\{b_n^{1/(2\beta)}\}$.

Condition 3

Assume that

$$\sup_{\boldsymbol{\theta}\in\{\boldsymbol{\theta}\in\boldsymbol{\Theta}:|\boldsymbol{\theta}-\boldsymbol{\theta}_0|_\infty<\epsilon_n b_n^{1/(2\beta)}\}} |\bar{\mathbf{g}}(\boldsymbol{\theta}) - \mathbb{E}\{\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})\}|_\infty = O_p(\zeta_n)$$

for some $\zeta_n \to 0$ *and* $\epsilon_n \to \infty$, *where* b_n *is defined in Theorem 1 and* β *is specified in Condition 2. Define* $\boldsymbol{\Theta}_n = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : |\boldsymbol{\theta} - \boldsymbol{\theta}_0|_\infty < \epsilon_n b_n^{1/(2\beta)}\}$. *We also assume that*

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_n} \max_{1\le j\le r} \left\{ \frac{1}{n}\sum_{i=1}^n |g_j(\mathbf{X}_i;\boldsymbol{\theta})|^2 \right\} = O_p(\varpi_n)$$

for some $\varpi_n > 0$ *that may diverge with* n.

Define $\mathcal{M}_{\boldsymbol{\theta}} = \{1 \leq j \leq r : |\mathbb{E}\{g_j(\mathbf{X}_i; \boldsymbol{\theta})\}| \geq \nu\rho_2'(0^+)/2\}$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Proposition 5 below shows that for any $\boldsymbol{\theta}$ in a small neighborhood of $\boldsymbol{\theta}_0$, the support of the Lagrange multiplier $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ is a subset of $\mathcal{M}_{\boldsymbol{\theta}}$ with probability approaching one.

Proposition 5

Let $\{\boldsymbol{\theta}_n\}$ be a sequence of points in $\boldsymbol{\Theta}_n$ with $\boldsymbol{\Theta}_n$ being defined in Condition 3, and the penalty function $p_{2,\nu}(\cdot)$ is a convex function. Assume Conditions 1 and 3 hold, $\mathbb{P}[\lambda_{\min}\{\widehat{V}_{\mathcal{M}_{\boldsymbol{\theta}_n}}(\boldsymbol{\theta}_n)\} \geq \xi_n] \to 1$ for some $\xi_n > 0$ (ξ_n may decay to zero as $n \to \infty$), and $|\bar{\mathbf{g}}_{\mathcal{M}_{\boldsymbol{\theta}_n}}(\boldsymbol{\theta}_n)|_\infty - \nu\rho_2'(0^+) = O_p(u_n)$ for some $u_n \to 0$. If $\max\{\zeta_n, m_n u_n \varpi_n \xi_n^{-1}\} = o(\nu)$ and $m_n^{\gamma+1} u_n^\gamma n = o(\xi_n^\gamma)$ where $m_n = |\mathcal{M}_{\boldsymbol{\theta}_n}|$, then with probability approaching one there exists a sparse local maximizer $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_n)$ for $f(\boldsymbol{\lambda}; \boldsymbol{\theta}_n)$ satisfying $|\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_n)|_1 = O_p(m_n u_n \xi_n^{-1})$ and $\mathbb{P}\{\mathrm{supp}(\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_n)) \subset \mathcal{M}_{\boldsymbol{\theta}_n}\} \to 1$ as $n \to \infty$.

- Proposition 5 implies that when $\boldsymbol{\theta}$ is around $\boldsymbol{\theta}_0$, the penalty on $\boldsymbol{\lambda}$ effectively conducts a moments selection by choosing the estimating functions in a way that $\mathbb{E}\{g_j(\mathbf{X}_i;\boldsymbol{\theta})\}$ has large absolute deviation from 0 when $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.
- By taking $\boldsymbol{\theta}_n$ as $\widehat{\boldsymbol{\theta}}_n$, Proposition 5 shows that with probability approaching one, there exists a sparse Lagrange multiplier $\widehat{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_n)$ which is a local maximizer for $f(\boldsymbol{\lambda};\widehat{\boldsymbol{\theta}}_n)$.
- We also note here that $\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}$ is a random set depending on the observations $\{\mathbf{X}_1,\ldots,\mathbf{X}_n\}$. To clearly characterize the property of $\widehat{\boldsymbol{\theta}}_n$, we define $\mathcal{R}_n \subset \{1,\ldots,r\}$ to be the support of $\widehat{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_n)$, i.e., the set of effective estimating functions that ultimately contribute to estimating the nonzero components of $\widehat{\boldsymbol{\theta}}_n$, and $q_n = |\mathcal{R}_n|$.

- By definition, $\mathcal{R}_n$ is a random set depending on multiple sources including the penalty function, the tuning parameter ν, and also the realization of the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Clearly, the asymptotic variance of $\widehat{\boldsymbol{\theta}}_n$ depends on the set $\mathcal{R}_n$ collecting the effective estimating functions for the ultimate estimations.
- For the sake of simplicity in presenting the theoretical development, we assume that the set $\mathcal{R}_n$ remains invariant when $\boldsymbol{\theta}$ takes a value in a neighborhood of $\widehat{\boldsymbol{\theta}}_n$. In other words, the set of estimating functions does not vary in the final steps of optimizing $S_n(\boldsymbol{\theta})$. Actually, such a requirement is not restrictive and is easy to fulfill. In practice, if the set of estimating functions varies substantially in optimizing $S_n(\boldsymbol{\theta})$, one can always drop the penalty on some chosen estimating functions and retain a stable $\mathcal{R}_n$, especially when the iterations are closing to convergence.

Condition 4

There exist uniform constants $0 < K_3 < K_4$ *such that*

$$\lim_{n\to\infty} \mathbb{P}\left[K_3 < \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_n} \lambda_{\min}\{\widehat{V}_{\mathcal{R}_n}(\boldsymbol{\theta})\} \leq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_n} \lambda_{\max}\{\widehat{V}_{\mathcal{R}_n}(\boldsymbol{\theta})\} < K_4\right] = 1$$

where $\boldsymbol{\Theta}_n$ *is defined in Condition* 3.

Condition 5

(i) There exists a positive constant $c < 1$ *such that*

$$\max_{k\in\mathcal{S}} \sup_{c|\theta_k^0|<t<c^{-1}|\theta_k^0|} p'_{1,\pi}(t) = O(\min\{s^{-1}b_n^{-1/(2\beta)}n^{-1}, sq_n^{1/2}\nu^2, q_n^{1/\gamma+1}n^{-1/2+1/\gamma}\nu\}).$$

(ii) The penalty function $p_{2,\nu}(\cdot)$ *is convex, and there exits a uniform constant* $K_5 > 0$ *such that* $|\rho_2'(t;\nu) - \rho_2'(0^+)| \leq K_5 t$ *for any* $t \in (0,\epsilon)$ *and* $\tau > 0$ *where* $\epsilon > 0$ *is a sufficiently small constant.*

Condition 6

For each $j = 1, \ldots, p$, $g_j(\mathbf{X}; \boldsymbol{\theta})$ *is continuously differentiable with respect to* $\boldsymbol{\theta}$ *in* $\boldsymbol{\Theta}_n$ *for any* $\mathbf{X}$, *where* $\boldsymbol{\Theta}_n$ *is defined in Condition 3, and there exist a function* $B_{n,jk}$ *with* $\mathbb{E}\{B_{n,jk}^2(\mathbf{X}_i)\} \leq K_6$ *for some uniform constant* $K_6 > 0$ *and* $|\partial g_j(\mathbf{X}; \boldsymbol{\theta})/\partial \theta_k| \leq B_{n,jk}(\mathbf{X})$ *for any* $\boldsymbol{\theta} \in \boldsymbol{\Theta}_n$. *In addition, it holds that*

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_n} \max_{j \in \mathcal{R}_n} \max_{k \notin \mathcal{S}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\partial g_j(\mathbf{X}_i; \boldsymbol{\theta})}{\partial \theta_k} \right| \right\} = O_p(\varphi_n)$$

for some $\varphi_n > 0$ *that may diverge with* n.

Condition 7

For each $j = 1, \ldots, p$, $g_j(\mathbf{X}; \boldsymbol{\theta})$ *is twice continuously differentiable with respect to* $\boldsymbol{\theta}$ *in* $\boldsymbol{\Theta}_n$ *for any* $\mathbf{X}$, *where* $\boldsymbol{\Theta}_n$ *is defined in Condition 3, and there exist some functions* $B_{n,jkl}$ *with* $\mathbb{E}\{B_{n,jkl}^2(\mathbf{X}_i)\} \leq K_7$ *for a uniform constant* $K_7 > 0$ *and* $|\partial^2 g_j(\mathbf{X}; \boldsymbol{\theta})/\partial\theta_k \partial\theta_l| \leq B_{n,jkl}(\mathbf{X})$ *for any* $\boldsymbol{\theta} \in \boldsymbol{\Theta}_n$. *In addition, for* $\mathcal{R}_n$ *given in Condition 4, it holds that*

$$\lim_{n\to\infty} \mathbb{P}\left[\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_n} \lambda_{\min}([\nabla_{\boldsymbol{\theta}_{(1)}} \bar{\mathbf{g}}_{\mathcal{R}_n}(\boldsymbol{\theta})]^{\mathrm{T}} [\nabla_{\boldsymbol{\theta}_{(1)}} \bar{\mathbf{g}}_{\mathcal{R}_n}(\boldsymbol{\theta})]) > K_8 \right] = 1$$

where $\boldsymbol{\theta}_{(1)}$ *is an* s*-dimensional subvector of* $\boldsymbol{\theta}$ *with components indexed by* $\mathcal{S} = \{1 \leq k \leq p : \theta_k^0 \neq 0\}$ *and* $K_8 > 0$ *is a uniform constant.*

We define

$$\mathbf{J} = [\mathbb{E}\{\nabla_{\boldsymbol{\theta}_{(1)}} \mathbf{g}_{\mathcal{R}_n}(\mathbf{X}_i; \boldsymbol{\theta}_0)\}]^{\mathrm{T}} V_{\mathcal{R}_n}^{-1}(\boldsymbol{\theta}_0)[\mathbb{E}\{\nabla_{\boldsymbol{\theta}_{(1)}} \mathbf{g}_{\mathcal{R}_n}(\mathbf{X}_i; \boldsymbol{\theta}_0)\}],$$

$$\widehat{\boldsymbol{\psi}}_{\mathcal{R}_n} = \widehat{\mathbf{J}}^{-1}[\nabla_{\boldsymbol{\theta}_{(1)}} \bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)]^{\mathrm{T}} \widehat{V}_{\mathcal{R}_n}^{-1}(\widehat{\boldsymbol{\theta}}_n) \left[\frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{g}_{\mathcal{R}_n}(\mathbf{X}_i; \widehat{\boldsymbol{\theta}}_n)}{1 + \widehat{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_n)^{\mathrm{T}} \mathbf{g}(\mathbf{X}_i; \widehat{\boldsymbol{\theta}}_n)}\right]$$

with $\widehat{\mathbf{J}} = [\nabla_{\boldsymbol{\theta}_{(1)}} \bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)]^{\mathrm{T}} \widehat{V}_{\mathcal{R}_n}^{-1}(\widehat{\boldsymbol{\theta}}_n)[\nabla_{\boldsymbol{\theta}_{(1)}} \bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)]$. We have the following asymptotic properties for $\widehat{\boldsymbol{\theta}}_n$.

Theorem 2

Let $\widehat{\boldsymbol{\theta}}_{n,(1)}$ *and* $\widehat{\boldsymbol{\theta}}_{n,(2)}$ *be the corresponding estimations of* $\boldsymbol{\theta}_{0,(1)}$ *and* $\boldsymbol{\theta}_{0,(2)}$, *respectively. Then the following properties hold.*

- *Under Conditions* 1–6, *if* $\log r = o(\min\{n^{1/3}, n^{1-2/\gamma}\varpi_n^{-1}\})$, $q_n = o(n^{(\gamma-2)/(2\gamma+2)})$ $b_n = o(\min_{k\in\mathcal{S}} |\theta_k^0|^{2\beta})$, *and the tuning parameters* π *and* ν *satisfy the conditions that* $q_n n^{-1/2}\varphi_n = o(\pi)$ *and* $\max\{\zeta_n, q_n n^{-1/2}\varpi_n, (n^{-1}\varpi_n \log r)^{1/2}\} = o(\nu)$, *then* $\mathbb{P}\{\widehat{\boldsymbol{\theta}}_{n,(2)} = \mathbf{0}\} \to 1$ *as* $n \to \infty$.
- *Assume the eigenvalues of* $[\mathbb{E}\{\nabla_{\boldsymbol{\theta}_{(1)}} \mathbf{g}_{\mathcal{R}_n}(\mathbf{X}_i; \boldsymbol{\theta}_0)\}]^{\mathrm{T}}[\mathbb{E}\{\nabla_{\boldsymbol{\theta}_{(1)}} \mathbf{g}_{\mathcal{R}_n}(\mathbf{X}_i; \boldsymbol{\theta}_0)\}]$ *and* $V_{\mathcal{R}_n}(\boldsymbol{\theta}_0)$ *are uniformly bounded away from zero and infinity. Under Conditions* 1–7, *for any* $\boldsymbol{\alpha} \in \mathbb{R}^s$ *with unit* L_2*-norm, if* $\log r = o(\min\{n^{1/3}, n^{1-2/\gamma}\varpi$ $q_n = o(n^{(\gamma-2)/(2\gamma+2)})$, $b_n = o(\min_{k\in\mathcal{S}} |\theta_k^0|^{2\beta})$, *and* π *and* ν *satisfy the conditions that* $q_n n^{-1/2}\varphi_n = o(\pi)$, $\max\{\zeta_n, q_n n^{-1/2}\varpi_n, (n^{-1}\varpi_n \log r)^{1/2}\} = o(\nu)$ *and* $\nu = o(\min\{s^{-1/2} q_n^{-1/2} n^{-1/4}, q_n^{-1/\gamma-3/2} n^{-1/\gamma}\})$, *then*

$$n^{1/2}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{J}^{1/2}\{\widehat{\boldsymbol{\theta}}_{n,(1)} - \boldsymbol{\theta}_{0,(1)} - \widehat{\boldsymbol{\psi}}_{\mathcal{R}_n}\} \xrightarrow{d} N(0,1)$$

as $n \to \infty$.

- The first part of Theorem 2 indicates that the zero components of $\boldsymbol{\theta}_0$ can be exactly estimated as zero by $\widehat{\boldsymbol{\theta}}_n$ with probability approaching one. The bias term $\widehat{\boldsymbol{\psi}}_{\mathcal{R}_n}$ is due to the use of the penalty function $p_{2,\nu}(\cdot)$. It can be shown that $\widehat{\boldsymbol{\psi}}_{\mathcal{R}_n} = \widehat{\mathbf{J}}^{-1}[\nabla_{\boldsymbol{\theta}_{(1)}} \bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)]^{\mathrm{T}} \widehat{V}_{\mathcal{R}_n}^{-1}(\widehat{\boldsymbol{\theta}}_n)\widehat{\boldsymbol{\eta}}_{\mathcal{R}_n}$ where $\widehat{\boldsymbol{\eta}} = (\widehat{\eta}_1, \ldots, \widehat{\eta}_r)^{\mathrm{T}}$ with $\widehat{\eta}_j = \nu\rho_2'(|\widehat{\lambda}_j|;\nu)\mathrm{sgn}(\widehat{\lambda}_j)$ for $\widehat{\lambda}_j \neq 0$ and $\widehat{\eta}_j \in [-\nu\rho_2'(0^+), \nu\rho_2'(0^+)]$ for $\widehat{\lambda}_j = 0$.
- The second part suggests that the the rate of the convergence is $n^{1/2}$ for each nonzero component of $\boldsymbol{\theta}_0$.

Algorithm

- For ease and stability in implementations, we calculate the penalized EL estimator $\widehat{\boldsymbol{\theta}}_n$ by minimizing the following slightly modified objective function:

$$\widehat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \max_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\boldsymbol{\theta})} \Bigg[\underbrace{\sum_{i=1}^{n} \log_\star\{1+\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})\} - n\sum_{j=1}^{r} p_{2,\nu}(|\lambda_j|)}_{f(\boldsymbol{\lambda};\boldsymbol{\theta})} + n\sum_{k=1}^{p} p_{1,\pi}(|\theta_k|) \Bigg],$$

where $\log_\star(z)$ is a twice differentiable pseudo-logarithm function with bounded support adopted from Owen (2001):

$$\log_\star(z) = \begin{cases} \log(z) & \text{if } z \geq \epsilon; \\ \log(\epsilon) - 1.5 + 2z/\epsilon - z^2/(2\epsilon^2) & \text{if } z \leq \epsilon; \end{cases}$$

where ϵ is chosen as $1/n$ in our implementations.

- In the optimization, we apply the quadratic approximation (Fan and Li, 2001) to the penalty functions $p_{1,\pi}(\cdot)$ and $p_{2,\nu}(\cdot)$. More specifically, for a penalty function $p_\tau(\cdot)$, the quadratic approximation states

$$p_\tau(|t|) \approx p_\tau(|t_0|) + \frac{1}{2}\frac{p'_\tau(|t_0|)}{|t_0|}(t^2 - t_0^2) \tag{1}$$

for t being in a small neighborhood of t_0. The first and second derivatives are approximated by

$$p'_\tau(|t|) \approx \frac{p'_\tau(|t_0|)}{|t_0|} \cdot t \quad \text{and} \quad p''_\tau(|t|) \approx \frac{p'_\tau(|t_0|)}{|t_0|}.$$

- The computation of EL is a challenging aspect, especially with high-dimensional p and r. To compute the penalized EL estimator $\widehat{\boldsymbol{\theta}}_n$, we propose to apply a modified two-layer coordinate decent algorithm. The inner layer of the algorithm solves for $\boldsymbol{\lambda}$ with given $\boldsymbol{\theta}$ by maximizing $f(\boldsymbol{\lambda}; \boldsymbol{\theta})$. This layer only involves maximizing a concave function, and hence is stable. The outer layer of the algorithm searches for the optimizer $\widehat{\boldsymbol{\theta}}_n$. Both layers can be solved using coordinate descent by cycling through and updating each of the coordinates; see Tang and Wu (2014).

- In the inner layer, $\boldsymbol{\lambda}$ is solved at a given $\boldsymbol{\theta}$, which can be done by optimizing $f(\boldsymbol{\lambda};\boldsymbol{\theta})$ with respect to $\boldsymbol{\lambda}$ using coordinate descent. Suppose that $\boldsymbol{\lambda}$ starts at an initial value $\widehat{\boldsymbol{\lambda}}^{(0)}$. With the other coordinates fixed, the $(m+1)$th Newton's update for λ_j $(j = 1, \ldots, r)$, the jth component of $\boldsymbol{\lambda}$, is given by

$$\widehat{\lambda}_j^{(m+1)} = \widehat{\lambda}_j^{(m)} - \frac{\sum_{i=1}^n \log_\star'(t_i^{(m)}) g_j(\mathbf{X}_i;\boldsymbol{\theta}) - n p_{2,\nu}'(|\widehat{\lambda}_j^{(m)}|)}{\sum_{i=1}^n \log_\star''(t_i^{(m)}) \{g_j(\mathbf{X}_i;\boldsymbol{\theta})\}^2 - n p_{2,\nu}''(|\widehat{\lambda}_j^{(m)}|)},$$

  where $t_i^{(m)} = 1 + \mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})^{\mathrm{T}} \widehat{\boldsymbol{\lambda}}^{(m)}$ with $\widehat{\boldsymbol{\lambda}}^{(m)} = (\widehat{\lambda}_1^{(m)}, \ldots, \widehat{\lambda}_r^{(m)})^{\mathrm{T}}$. The procedure cycles through all the r components of $\boldsymbol{\lambda}$ and is repeated until convergence.

- During this process, the objective function needs to be checked to ensure it gets optimized in each step. If not, the step size continues to be halved until the objective function gets driven in the right direction.

- The outer layer of the algorithm is to optimize the objective function with respect to the parameter $\boldsymbol{\theta}$, the main interest of the penalized EL, using the coordinate descent algorithm. At a given $\boldsymbol{\lambda}$, the algorithm updates θ_k $(k = 1, \ldots, p)$, by minimizing $S_n(\boldsymbol{\theta})$ with respect to θ_k with other θ_l $(l \neq k)$ fixed. Suppose that $\boldsymbol{\theta}$ starts at an initial value $\widehat{\boldsymbol{\theta}}^{(0)}$. The $(m+1)$th update for θ_k is given by

$$\widehat{\theta}_k^{(m+1)} = \widehat{\theta}_k^{(m)} - \frac{\sum_{i=1}^n \log_\star'(s_i^{(m)}) w_{ik}^{(m)} + n p_{1,\tau}'(|\widehat{\theta}_k^{(m)}|)}{\sum_{i=1}^n [\log_\star''(s_i^{(m)})\{w_{ik}^{(m)}\}^2 + \log_\star'(s_i^{(m)}) z_{ik}^{(m)}] + n p_{1,\tau}''(|\widehat{\theta}_k^{(m)}|)},$$

where $s_i^{(m)} = 1 + \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{g}(\mathbf{X}_i; \widehat{\boldsymbol{\theta}}^{(m)})$, $w_{ik}^{(m)} = \boldsymbol{\lambda}^{\mathrm{T}} \partial \mathbf{g}(\mathbf{X}_i; \widehat{\boldsymbol{\theta}}^{(m)})/\partial \theta_k$ and $z_{ik}^{(m)} = \boldsymbol{\lambda}^{\mathrm{T}} \partial^2 \mathbf{g}(\mathbf{X}_i; \widehat{\boldsymbol{\theta}}^{(m)})/\partial \theta_k^2$ with $\widehat{\boldsymbol{\theta}}^{(m)} = (\widehat{\theta}_1^{(m)}, \ldots, \widehat{\theta}_p^{(m)})^{\mathrm{T}}$.

- Since quadratic approximations are applied in the algorithms, we follow Fan and Li (2001) and set a component $\widehat{\lambda}_j^{(m)}$ or $\widehat{\theta}_k^{(m)}$ as zero when it is less than a threshold level say 10^{-3} in an iteration.

- We summarize the computation procedure for $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ in the following pseudo-code. Suppose ξ is a pre-defined small number, say, $\xi = 10^{-4}$.

1. Set the iteration counter $m = 0$, and initialize $\widehat{\boldsymbol{\theta}}^{(0)}$ and $\widehat{\boldsymbol{\lambda}}^{(0)}$;
2. Define the $\mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta})$ function;
3. (Outer layer) For $k = 1, \ldots, p$,
   - 3.1 Calculate $\widehat{\theta}_k^{(m+1)}$;
   - 3.2 (Inner layer) For $j = 1, \ldots, r$, update $\widehat{\lambda}_j^{(m)}$ as $\widehat{\lambda}_j^{(m+1)}$;
4. If $\max_{1 \le k \le p} |\widehat{\theta}_k^{(m+1)} - \widehat{\theta}_k^{(m)}| < \xi$, then stop;
5. Otherwise repeat steps 3 through 4.

Numerical results

- Linear regression model $Y_i = \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\theta}_0 + \varepsilon_i$, where $\boldsymbol{\theta}_0 = (3, 1.5, 0, 0, 2, 0, \ldots, 0)^{\mathrm{T}}$, $\mathbf{Z}_i \in \mathbb{R}^p$ are generated from $N(\mathbf{0}, \boldsymbol{\Sigma})$ with $\sigma_{kk} = 1$ for any $k = 1, \ldots, p$ and $\sigma_{kl} = 0.5$ for any $k \neq l$, where $\boldsymbol{\Sigma} = (\sigma_{kl})_{p\times p}$, and ε_i is a standard normal distributed random variable. Write $\mathbf{X}_i = (Y_i, \mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}}$. The estimating function is $\mathbf{g}(\mathbf{X}; \boldsymbol{\theta}) = \mathbf{Z}(Y - \mathbf{Z}^{\mathrm{T}}\boldsymbol{\theta})$ with $p = r$.
- The proposed penalized EL with two penalties (namely, PEL2) is compared to the single penalty approach (PEL) discussed in Tang and Leng (2010). Three information criteria for choosing the tuning parameters π and ν in the penalty function – BIC (Schwarz, 1978), BICC (Wang, Li and Leng, 2009), and EBIC (Chen and Chen, 2008) – are used.

- The results from MLE for the three true variables (i.e., MLE-Oracle) is also reported. The column of $\boldsymbol{\theta}_{\text{nonzero}}$ reports the average number of selected predictors. The column of $\boldsymbol{\theta}_{\text{true}}$ reports the average number of true predictors that are selected. The difference is the average number of false predictors that get selected. The next column reports the model error (ME), which is defined by $\text{ME} = (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\text{T}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ for a given estimator $\widehat{\boldsymbol{\theta}}$.

(n, p, r)	Method	$\boldsymbol{\theta}_{\text{nonzeros}}$	$\boldsymbol{\theta}_{\text{true}}$	ME	No. EE's
$n = 50$	MLE-Oracle	3 (0)	NA	0.104 (0.009)	NA
$p = 100$	PEL-BIC	0 (0)	0 (0)	15.25 (0)	100 (0)
$r = 100$	PEL-BIC	0 (0)	0 (0)	15.25 (0)	100 (0)
	PEL-EBIC	0 (0)	0 (0)	15.25 (0)	100 (0)
	PEL2-BIC	4.94 (0.53)	2.92 (0.04)	0.988 (0.160)	5.73 (0.21)
	PEL2-BIC	4.73 (0.48)	2.92 (0.04)	1.025 (0.165)	5.69 (0.22)
	PEL2-EBIC	4.42 (0.43)	2.90 (0.04)	1.239 (0.211)	5.63 (0.23)
$n = 100$	MLE-Oracle	3 (0)	NA	0.047 (0.005)	NA
$p = 200$	PEL-BIC	0 (0)	0 (0)	15.25 (0)	200 (0)
$r = 200$	PEL-BIC	0 (0)	0 (0)	15.25 (0)	200 (0)
	PEL-EBIC	0 (0)	0 (0)	15.25 (0)	200 (0)
	PEL2-BIC	9.22 (1.27)	3 (0)	1.070 (0.225)	5.38 (0.17)
	PEL2-BIC	9.28 (1.28)	3 (0)	1.079 (0.227)	5.39 (0.17)
	PEL2-EBIC	8.38 (1.03)	3 (0)	1.056 (0.228)	5.34 (0.17)
$n = 100$	MLE-Oracle	3 (0)	NA	0.039 (0.003)	NA
$p = 500$	PEL-BIC	0 (0)	0 (0)	15.25 (0)	500 (0)
$r = 500$	PEL-BIC	0 (0)	0 (0)	15.25 (0)	500 (0)
	PEL-EBIC	0 (0)	0 (0)	15.25 (0)	500 (0)
	PEL2-BIC	6.28 (1.31)	3 (0)	0.946 (0.153)	5.48 (0.16)
	PEL2-BIC	5.96 (1.31)	3 (0)	0.930 (0.155)	5.38 (0.17)
	PEL2-EBIC	6.04 (1.32)	3 (0)	0.944 (0.157)	5.41 (0.16)

Discussion

- We study a new penalized EL approach with two penalties, with one encouraging sparsity of the estimator and the other encouraging sparsity of the Lagrange multiplier in the optimizations associated with the EL. Such an approach utilizes sparsity in the target parameters and effectively achieves a moment selection procedure for estimating the sparse parameter. Both theory and numerical examples confirm the merits of the new penalized EL.
- One interesting extension of the approach is to explore inferences with estimating equations after the variable selection procedure. Such a direction is a suitable stage for EL method with estimating equations who takes advantage of adaptivity to various moment conditions with less stringent distributional assumptions. (Done)
- The other interesting and challenging problem is to explore the optimality of the sparse estimator using estimating equations with high data dimensionality. (Working on)

Thank you!