A low-rank based estimation-testing procedure for matrix-covariate regression: Supplementary Materials

Hung Hung and Zhi-Yu Jou

Institute of Epidemiology and Preventive Medicine

National Taiwan University, Taiwan

1 Asymptotic Property of $\widehat{\beta}$

The asymptotic property of $\widehat{\beta}$ is parallel in spirit to the asymptotic property of the over-parameterized minimum discrepancy estimator (Shapiro, 1986), where we use KL-divergence as the discrepancy function. It is different from Shapiro (1986) in that the KL-divergence is a function of $\{X_i\}_{i=1}^n$ which is a random discrepancy function.

Proof of Theorem 1. Since θ is over-parameterized, there exists a (locally) one-to-one function $\theta = h(\tau, \bar{\tau}) : \mathbb{R}^{s_r} \times \mathbb{R}^{1+m+pq-s_r} \to \mathbb{R}^{1+m+pq}$ such that $\beta(h(\tau, \bar{\tau}))$ depends on τ only (Shapiro, 1986). Here τ can be treated as the minimal effective parameter for the rank-r GLM (8). Define $\beta^*(\tau) = \beta(h(\tau, \mathbf{0}))$ as the parameterization of β via the effective parameter τ, and define τ_0 as the unique true value of τ such that $\theta_0 = h(\tau_0, \mathbf{0})$ and, hence, $\beta_0 = \beta^*(\tau_0)$. Let $\widehat{\tau}$ be the MLE of τ_0, which satisfies $\|\widehat{\tau} - \tau_0\| = O_p(n^{-1/2})$ by conventional MLE argument. Let also $\widehat{\beta}^* = \beta^*(\widehat{\tau})$ be the corresponding MLE of β_0. By the invariance property of MLE, $\widehat{\beta}$ and $\widehat{\beta}^*$ share the same asymptotic property, and it suffices to work on $\widehat{\beta}^*$ to complete the proof. Moreover, since $\lambda = o_p(n^{-1/2})$, we can ignore the effect of penalty during the derivations.

Let $\widetilde{\beta}$ be the conventional MLE of β_0 under model (3). From the connection between MLE and KL-divergence, $\widehat{\tau}$ can be characterized as

$$\widehat{\tau} = \underset{\tau}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} D_i(\widetilde{\beta}, \beta^*(\tau))$$

with $D_i(\beta_1, \beta_2) = \int \ln \frac{f(y|X_i;\beta_1)}{f(y|X_i;\beta_2)} \cdot f(y|X_i;\beta_1) dy$ being the KL-divergence between $f(y|X_i;\beta_1)$ and $f(y|X_i;\beta_2)$, where $f(y|x;\beta)$ is the conditional distribution function of Y given $X = x$

under model (3). Let $D_{i,j}$ be the partial derivative of D_i with respect to its j-th argument, and let $D_{i,jk}$ be the partial derivative of $D_{i,j}$ with respect to its k-th argument. Direct calculation gives $\widehat{\tau}$ to be the solution of the estimating equation

$$\mathbf{0} \;=\; \frac{1}{n}\sum_{i=1}^{n} D_{i,2}(\widetilde{\beta}, \beta^*(\widehat{\tau})) \cdot \mathbf{\Delta}^*(\widehat{\tau}), \tag{1}$$

where

$$\mathbf{\Delta}^*(\tau) = \frac{\partial \beta^*(\tau)}{\partial \tau} = \mathbf{\Delta}(\theta)|_{\theta=h(\tau,\mathbf{0})} \cdot \frac{\partial h(\tau,\mathbf{0})}{\partial \tau}. \tag{2}$$

Since $\beta_0 = \beta^*(\tau_0)$, $D_i(\beta_0, \beta^*(\tau_0))$ attains the minimum value 0 and, hence, $D_{i,2}(\beta_0, \beta^*(\tau_0)) = 0$. This fact together with taking Taylor's expansion of (1) around $(\widetilde{\beta}, \widehat{\tau}) = (\beta_0, \tau_0)$ give

$$\begin{aligned}
\mathbf{0} &= \mathbf{\Delta}_0^{*\top}\left[\frac{1}{n}\sum_{i=1}^{n} D_{i,21}(\beta_0,\beta_0)\right](\widetilde{\beta}-\beta_0) + \mathbf{\Delta}_0^{*\top}\left[\frac{1}{n}\sum_{i=1}^{n} D_{i,22}(\beta_0,\beta_0)\right]\mathbf{\Delta}_0^*(\widehat{\tau}-\tau_0) + o_p(n^{-\frac{1}{2}}) \\
&= \mathbf{\Delta}_0^{*\top} D_{21}(\widetilde{\beta}-\beta_0) + \mathbf{\Delta}_0^{*\top} D_{22}\mathbf{\Delta}_0^*(\widehat{\tau}-\tau_0) + o_p(n^{-\frac{1}{2}}),
\end{aligned} \tag{3}$$

where $\mathbf{\Delta}_0^* = \mathbf{\Delta}^*(\tau_0)$, $D_{21} = E[D_{i,21}(\beta_0,\beta_0)]$, and $D_{22} = E[D_{i,22}(\beta_0,\beta_0)]$. Note that $D_{21} = -\boldsymbol{V}_0$ and $D_{22} = \boldsymbol{V}_0$ from direct calculations, where $\boldsymbol{V}_0$ is defined in Theorem 1.

To proceed the proof, we deduce from the definitions of h and (2) that

$$\left[\mathbf{\Delta}_0^*, \mathbf{0}\right] = \left[\frac{\partial \beta(h(\tau,\bar{\tau}))}{\partial(\tau,\bar{\tau})}\right]_{(\tau,\bar{\tau})=(\tau_0,\mathbf{0})} = \mathbf{\Delta}_0 \cdot \left[\frac{\partial h(\tau,\bar{\tau})}{\partial(\tau,\bar{\tau})}\right]_{(\tau,\bar{\tau})=(\tau_0,\mathbf{0})}. \tag{4}$$

Since h is one-to-one, (4) implies that

$$\mathrm{span}(\mathbf{\Delta}_0^*) = \mathrm{span}(\mathbf{\Delta}_0). \tag{5}$$

It further implies that $\mathbf{\Delta}_0^*$ is of full column rank by the assumption $\mathrm{rank}(\mathbf{\Delta}_0) = s_r$. Combining the above discussions, we conclude from (3) that

$$\sqrt{n}(\widehat{\tau}-\tau_0) = \left(\mathbf{\Delta}_0^{*\top}\boldsymbol{V}_0\mathbf{\Delta}_0^*\right)^{-1}\mathbf{\Delta}_0^{*\top}\boldsymbol{V}_0 \cdot \sqrt{n}(\widetilde{\beta}-\beta_0) + o_p(1). \tag{6}$$

To complete the proof, first note that standard argument gives the asymptotic normality of the conventional MLE $\widetilde{\beta}$ to be $\sqrt{n}(\widetilde{\beta}-\beta_0) \xrightarrow{d} N(0, \boldsymbol{V}_0^{-1})$. From (6) and applying the delta method to the transformation $\widehat{\beta}^* = \beta^*(\widehat{\tau})$, we have

$$\begin{aligned}\sqrt{n}(\widehat{\beta}^* - \beta_0) &= \boldsymbol{\Delta}_0^* \cdot \sqrt{n}(\widehat{\tau} - \tau_0) + o_p(1) \\ &\xrightarrow{d} \boldsymbol{P}_{\boldsymbol{\Delta}_0^*, \boldsymbol{V}_0} \cdot N(\mathbf{0}, \boldsymbol{V}_0^{-1}),\end{aligned}$$

where $\boldsymbol{P}_{\boldsymbol{\Delta}_0^*, \boldsymbol{V}_0} = \boldsymbol{\Delta}_0^* \left(\boldsymbol{\Delta}_0^{*\top} \boldsymbol{V}_0 \boldsymbol{\Delta}_0^*\right)^+ \boldsymbol{\Delta}_0^{*\top} \boldsymbol{V}_0$ is the projection matrix onto span$(\boldsymbol{\Delta}_0^*)$ with respect to the $\boldsymbol{V}_0$ inner product. Since $\boldsymbol{P}_{\boldsymbol{\Delta}_0^*, \boldsymbol{V}_0} = \boldsymbol{P}_{\boldsymbol{\Delta}_0, \boldsymbol{V}_0}$ due to (5), we have

$$\sqrt{n}(\widehat{\beta}^* - \beta_0) \xrightarrow{d} \boldsymbol{P}_{\boldsymbol{\Delta}_0, \boldsymbol{V}_0} \cdot N(\mathbf{0}, \boldsymbol{V}_0^{-1}).$$

The proof is completed by noting that $\boldsymbol{P}_{\boldsymbol{\Delta}_0, \boldsymbol{V}_0} \cdot \boldsymbol{V}_0^{-1} \cdot \boldsymbol{P}_{\boldsymbol{\Delta}_0, \boldsymbol{V}_0}^\top = \boldsymbol{\Delta}_0(\boldsymbol{\Delta}_0^\top \boldsymbol{V}_0 \boldsymbol{\Delta}_0)^+ \boldsymbol{\Delta}_0^\top$. $\square$