Supplementary Material for "A Robust Consistent Information Criterion for Model Selection Based on Empirical Likelihood"

Chixiang Chen[1], Ming Wang[2], Rongling Wu[2], Runze Li[3]

[1]*Division of Biostatistics and Bioinformatics, University of Maryland School of Medicine, Baltimore, MD 21201, USA*

[2]*Division of Biostatistics and Bioinformatics, Department of Public Health Science, Pennsylvania State College of Medicine, Hershey, PA 17033, USA*

[3]*Department of Statistics and the Methodology Center, Pennsylvania State University, University Park, PA 16802, USA*

In the Supplementary Material, we will provide the technical proofs to the Theorems 1-4 and Corollary 1 in the main paper as well as the results for additional simulation studies. Note that all expectations are evaluated at the true values. For convenience, we will utilize the following simplified notations:

$E(\mathbf{g}\mathbf{g}^{\mathrm{T}}) = E(\mathbf{g}(\mathbf{D},\boldsymbol{\gamma}_0)\mathbf{g}^{\mathrm{T}}(\mathbf{D},\boldsymbol{\gamma}_0))$, $E(\partial\mathbf{g}/\partial\boldsymbol{\gamma}^{\mathrm{T}}) = E(\partial\mathbf{g}(\mathbf{D},\boldsymbol{\gamma}_0)/\partial\boldsymbol{\gamma}^{\mathrm{T}})$;

$E(\mathbf{g}_1\mathbf{g}_1^{\mathrm{T}}) = E(\mathbf{g}_1(\mathbf{D},\boldsymbol{\gamma}_0)\mathbf{g}_1^{\mathrm{T}}(\mathbf{D},\boldsymbol{\gamma}_0))$, $E(\partial\mathbf{g}_1/\partial\boldsymbol{\gamma}^{\mathrm{T}}) = E(\partial\mathbf{g}_1(\mathbf{D},\boldsymbol{\gamma}_0)/\partial\boldsymbol{\gamma}^{\mathrm{T}})$;

$E(\mathbf{g}_2\mathbf{g}_2^{\mathrm{T}}) = E(\mathbf{g}_2(\mathbf{D},\boldsymbol{\gamma}_0)\mathbf{g}_2^{\mathrm{T}}(\mathbf{D},\boldsymbol{\gamma}_0))$, $E(\partial\mathbf{g}_2/\partial\boldsymbol{\gamma}^{\mathrm{T}}) = E(\partial\mathbf{g}_2(\mathbf{D},\boldsymbol{\gamma}_0)/\partial\boldsymbol{\gamma}^{\mathrm{T}})$;

$E(\mathbf{g}_1\mathbf{g}_2^{\mathrm{T}}) = E(\mathbf{g}_2\mathbf{g}_1^{\mathrm{T}})^{\mathrm{T}} = E(\mathbf{g}_1(\mathbf{D},\boldsymbol{\gamma}_0)\mathbf{g}_2^{\mathrm{T}}(\mathbf{D},\boldsymbol{\gamma}_0))$.

S1. Technical proofs

S1.1 Proof of Theorem 1

In order to show Theorem 1 in the main paper, we need the following lemma:

Lemma 1. *Under Condition 1 and Condition 2 in the main paper, we have the following relationship:*

$$\hat{\boldsymbol{\gamma}}_{EL} = \hat{\boldsymbol{\gamma}}_{EE} + o_p(\mathbf{n}^{-\frac{1}{2}}),$$
$$\hat{\boldsymbol{\lambda}}_{EL} = \hat{\boldsymbol{\lambda}}_{EE} + o_p(\mathbf{n}^{-\frac{1}{2}}).$$

Proof. Let $\tilde{\gamma}_0 = (\boldsymbol{\gamma}_0^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$. Along the lines with the proof of Lemma 1 and Theorem 1 in Qin and Lawless (1994) under *Condition 1*, and based on asymptotic theory in generalized method of moment (Newey and McFadden, 1994), we have

$$\begin{aligned}
\hat{\boldsymbol{\gamma}}_{EL} - \boldsymbol{\gamma}_0 &= -\boldsymbol{\Gamma}^{-1} E\Big(\frac{\partial \mathbf{g}}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\Big)^{\mathrm{T}} \Big(E\mathbf{g}\mathbf{g}^{\mathrm{T}}\Big)^{-1} \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma}_0) + o_p(\mathbf{n}^{-\frac{1}{2}}), \\
\hat{\boldsymbol{\gamma}}_{EE} - \boldsymbol{\gamma}_0 &= -E\Big(\frac{\partial \mathbf{g}_1}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\Big)^{-1} \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_1(\mathbf{D}_i, \tilde{\gamma}_0) + o_p(\mathbf{n}^{-\frac{1}{2}}),
\end{aligned} \tag{S1.1}$$

with $\boldsymbol{\Gamma} = E(\partial \mathbf{g}/\partial \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}(E\mathbf{g}\mathbf{g}^{\mathrm{T}})^{-1}E(\partial \mathbf{g}/\partial \boldsymbol{\gamma}^{\mathrm{T}})$. Note that $\mathbf{g}_1(\mathbf{D}_i, \tilde{\gamma}_0) = \mathbf{g}_1(\mathbf{D}_i, \boldsymbol{\gamma}_0)$.

So we can keep using $\mathbf{g}_1(\mathbf{D}_i, \boldsymbol{\gamma}_0)$ in the following proof. Thus, we have:

$$\boldsymbol{\Gamma} = \Big(E\Big(\frac{\partial \mathbf{g}_1}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\Big)^{\mathrm{T}}, -E\Big(\frac{\partial \mathbf{g}_1}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\Big)^{\mathrm{T}} (E\mathbf{g}_1\mathbf{g}_1^{\mathrm{T}})^{-1} E(\mathbf{g}_1\mathbf{g}_2^{\mathrm{T}}) + E\Big(\frac{\partial \mathbf{g}_2}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\Big)^{\mathrm{T}} \Big) \times$$
$$\begin{pmatrix} E(\mathbf{g}_1\mathbf{g}_1^{\mathrm{T}})^{-1} & 0 \\ 0 & \mathbf{A} \end{pmatrix} \begin{pmatrix} E(\frac{\partial \mathbf{g}_1}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}) \\ -E(\mathbf{g}_2\mathbf{g}_1^{\mathrm{T}}) E(\mathbf{g}_1\mathbf{g}_1^{\mathrm{T}})^{-1} E(\frac{\partial \mathbf{g}_1}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}) + E(\frac{\partial \mathbf{g}_2}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}), \end{pmatrix},$$

with $\mathbf{A} = E(\mathbf{g}_2\mathbf{g}_2^{\mathrm{T}}) - E(\mathbf{g}_2\mathbf{g}_1^{\mathrm{T}})(E\mathbf{g}_1\mathbf{g}_1^{\mathrm{T}})^{-1}E(\mathbf{g}_1\mathbf{g}_2^{\mathrm{T}})$. By *Condition* 2, we have $E(\partial \mathbf{g}_2/\partial \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}} = E(\partial \mathbf{g}_1/\partial \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}(E\mathbf{g}_1\mathbf{g}_1^{\mathrm{T}})^{-1}E(\mathbf{g}_1\mathbf{g}_2^{\mathrm{T}})$, which leads to

$$\boldsymbol{\Gamma} = E\Big(\frac{\partial \mathbf{g}}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\Big)^{\mathrm{T}} (E\mathbf{g}\mathbf{g}^{\mathrm{T}})^{-1} E\Big(\frac{\partial \mathbf{g}}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\Big) = E\Big(\frac{\partial \mathbf{g}_1}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\Big)^{\mathrm{T}} (E\mathbf{g}_1\mathbf{g}_1^{\mathrm{T}})^{-1} E\Big(\frac{\partial \mathbf{g}_1}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\Big). \tag{S1.2}$$

Similarly, applying *Condition* 2 again, we have

$$E\Big(\frac{\partial \mathbf{g}}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\Big)^{\mathrm{T}} (E\mathbf{g}\mathbf{g}^{\mathrm{T}})^{-1} \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma}_0) = E\Big(\frac{\partial \mathbf{g}_1}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\Big)^{\mathrm{T}} (E\mathbf{g}_1\mathbf{g}_1^{\mathrm{T}})^{-1} \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_1(\mathbf{D}_i, \boldsymbol{\gamma}_0). \tag{S1.3}$$

Substituting (S1.2) and (S1.3) into (S1.1), and by noticing that $E(\partial \mathbf{g}_1/\partial \boldsymbol{\gamma}^{\mathrm{T}})$ is invertible, we have

$$\begin{aligned} \hat{\boldsymbol{\gamma}}_{EL} - \boldsymbol{\gamma}_0 =& - E\Big(\frac{\partial \mathbf{g}_1}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\Big)^{-1} \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_1(\mathbf{D}_i, \boldsymbol{\gamma}_0) + o_p(\mathbf{n}^{-\frac{1}{2}}) \\ =& \hat{\boldsymbol{\gamma}}_{EE} - \boldsymbol{\gamma}_0 + o_p(\mathbf{n}^{-\frac{1}{2}}). \end{aligned}$$

Finally, by Taylor expansion of the equation (2.4) in the main paper at $\boldsymbol{\gamma}_0$ and $\boldsymbol{\lambda}_0 = \mathbf{0}$, we can derive the following equation

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{D}_i,\boldsymbol{\gamma}_0) + E\Big(\frac{\partial \mathbf{g}}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\Big)(\hat{\boldsymbol{\gamma}}_{EE}-\boldsymbol{\gamma}_0) - E(\mathbf{g}\mathbf{g}^{\mathrm{T}})\hat{\boldsymbol{\lambda}}_{EE} + o_p(\mathbf{n}^{-\frac{1}{2}}) = \mathbf{0},$$

leading to $\hat{\boldsymbol{\lambda}}_{EL} = \hat{\boldsymbol{\lambda}}_{EE} + o_p(\mathbf{n}^{-1/2})$ by applying $\hat{\boldsymbol{\gamma}}_{EL} = \hat{\boldsymbol{\gamma}}_{EE} + o_p(\mathbf{n}^{-1/2})$. □

Now we are ready to show Theorem 1, and the proof is shown next.

Proof. Now consider any random variables $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ satisfying $\boldsymbol{\lambda}-\boldsymbol{\lambda}_0 = O_P(\mathbf{n}^{-1/2})$ and $\boldsymbol{\gamma} - \boldsymbol{\gamma}_0 = O_P(\mathbf{n}^{-1/2})$ and expand $\log R^F(\boldsymbol{\lambda},\boldsymbol{\gamma})$ at $\boldsymbol{\lambda}^*$ and $\boldsymbol{\gamma}^*$ such that $\log R^F(\boldsymbol{\lambda}^*,\boldsymbol{\gamma}^*)$ is maximized. Indeed, these two $\boldsymbol{\lambda}^*$ and $\boldsymbol{\gamma}^*$ are the maximum empirical likelihood estimates $\hat{\boldsymbol{\lambda}}_{EL}$ and $\hat{\boldsymbol{\gamma}}_{EL}$, respectively; given $l = -\log R^F(\boldsymbol{\lambda},\boldsymbol{\gamma})$, we have

$$\begin{aligned}
\log R^F(\boldsymbol{\lambda},\boldsymbol{\gamma}) = \log R^F(\hat{\boldsymbol{\lambda}}_{EL},\hat{\boldsymbol{\gamma}}_{EL}) &- \frac{\partial l}{\partial \boldsymbol{\lambda}}\Big|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}_{EL},\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}_{EL}}(\boldsymbol{\lambda}-\hat{\boldsymbol{\lambda}}_{EL}) \\
&- \frac{\partial l}{\partial \boldsymbol{\gamma}}\Big|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}_{EL},\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}_{EL}}(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}_{EL}) \\
&- \frac{1}{2}(\boldsymbol{\lambda}-\hat{\boldsymbol{\lambda}}_{EL})^{\mathrm{T}}\frac{\partial^2 l}{\partial \boldsymbol{\lambda}\partial \boldsymbol{\lambda}^{\mathrm{T}}}\Big|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}_{EL},\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}_{EL}}(\boldsymbol{\lambda}-\hat{\boldsymbol{\lambda}}_{EL}) \\
&- (\boldsymbol{\lambda}-\hat{\boldsymbol{\lambda}}_{EL})^{\mathrm{T}}\frac{\partial^2 l}{\partial \boldsymbol{\lambda}\partial \boldsymbol{\gamma}^{\mathrm{T}}}\Big|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}_{EL},\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}_{EL}}(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}_{EL}) \\
&- \frac{1}{2}(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}_{EL})^{\mathrm{T}}\frac{\partial^2 l}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}^{\mathrm{T}}}\Big|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}_{EL},\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}_{EL}}(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}_{EL}) + o_p(1).
\end{aligned}$$

To be noted that $(\partial l/\partial \boldsymbol{\lambda})\ |_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}_{EL},\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}_{EL}} = (\partial l/\partial \boldsymbol{\gamma})\ |_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}_{EL},\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}_{EL}} = \mathbf{0}$ by defi-

nition of $\hat{\boldsymbol{\lambda}}_{EL}$ and $\hat{\boldsymbol{\gamma}}_{EL}$. Furthermore, applying the weak law of large number, we have $(1/n)\partial^2 l/(\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\mathrm{T}}) \xrightarrow{P} \mathbf{0}$, $(1/n)\partial^2 l/(\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}^{\mathrm{T}}) \xrightarrow{P} -E\{\mathbf{g}(\mathbf{D},\boldsymbol{\gamma}_0)\mathbf{g}^{\mathrm{T}}(\mathbf{D},\boldsymbol{\gamma}_0)\}$, and $(1/n)\partial^2 l/(\partial\boldsymbol{\lambda}\partial\boldsymbol{\gamma}^{\mathrm{T}}) \xrightarrow{P} E\{\partial\mathbf{g}(\mathbf{D},\boldsymbol{\gamma}_0)/\partial\boldsymbol{\gamma}^{\mathrm{T}}\}$. By utilizing all the derived results, we rewrite the logarithm of empirical likelihood ratio as

$$\log R^F(\boldsymbol{\lambda},\boldsymbol{\gamma}) = -\frac{1}{2}\Big((\boldsymbol{\lambda}-\hat{\boldsymbol{\lambda}}_{EL})^{\mathrm{T}},(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}_{EL})^{\mathrm{T}}\Big)^{\mathrm{T}} \begin{pmatrix} -n\boldsymbol{\Sigma}_{11} & n\boldsymbol{\Sigma}_{12} \\ n\boldsymbol{\Sigma}_{21} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}-\hat{\boldsymbol{\lambda}}_{EL} \\ \boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}_{EL} \end{pmatrix} + \log R^F(\hat{\boldsymbol{\lambda}}_{EL},\hat{\boldsymbol{\gamma}}_{EL}) + o_p(1),$$

where $\boldsymbol{\Sigma}_{11} = E\{\mathbf{g}(\mathbf{D},\boldsymbol{\gamma}_0)\mathbf{g}^{\mathrm{T}}(\mathbf{D},\boldsymbol{\gamma}_0)\}$, $\boldsymbol{\Sigma}_{12} = E\{\partial\mathbf{g}(\mathbf{D},\boldsymbol{\gamma}_0)/\partial\boldsymbol{\gamma}^{\mathrm{T}}\}$, $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^{\mathrm{T}}$. By Lemma 1 and some algebra, we can have $\log R^F(\boldsymbol{\lambda},\boldsymbol{\gamma}) = \log R^F(\hat{\boldsymbol{\lambda}}_{EE},\hat{\boldsymbol{\gamma}}_{EE}) - (1/2)\boldsymbol{\delta}^{\mathrm{T}}(n\boldsymbol{\Sigma})\boldsymbol{\delta} + o_p(1)$ with

$$\boldsymbol{\Sigma} = \begin{pmatrix} -\boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \end{pmatrix} \text{ and } \boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \end{pmatrix},$$

where $\boldsymbol{\delta}_1 = \boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}_{EE} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}_{EE})$ and $\boldsymbol{\delta}_2 = \boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}_{EE}$.

In the end, by integrating out the random variables $\boldsymbol{\delta}$, the marginal probabil-

ity will become

$$\begin{aligned} P(\mathbf{Y}|M) =& R^F(\hat{\boldsymbol{\lambda}}_{EE}, \hat{\boldsymbol{\gamma}}_{EE}) \int \exp\Big\{ -\frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}(n\boldsymbol{\Sigma})\boldsymbol{\delta}\Big\}\rho_{\boldsymbol{\delta}}(\boldsymbol{\delta})d\boldsymbol{\delta} + o_p(1) \\ =& R^F(\hat{\boldsymbol{\lambda}}_{EE}, \hat{\boldsymbol{\gamma}}_{EE}) \int \exp\Big\{\frac{1}{2}\boldsymbol{\delta}_1^{\mathrm{T}}(n\boldsymbol{\Sigma}_{11})\boldsymbol{\delta}_1\Big\}\rho_{\boldsymbol{\delta}_1}(\boldsymbol{\delta}_1)d\boldsymbol{\delta}_1 \\ & \cdot \int \exp\Big\{ -\frac{1}{2}\boldsymbol{\delta}_2^{\mathrm{T}}(n\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})\boldsymbol{\delta}_2\Big\}\rho_{\boldsymbol{\delta}_2}(\boldsymbol{\delta}_2)d\boldsymbol{\delta}_2 + o_p(1) \end{aligned}$$

Thus, by applying non-informative prior to $\boldsymbol{\delta}_2$, i.e., $\rho_{\boldsymbol{\delta}_2}(\boldsymbol{\delta}_2) = 1$, and applying the Laplace approximation, we can have $P(\mathbf{Y}|M) = R^F(\hat{\boldsymbol{\lambda}}_{EE}, \hat{\boldsymbol{\gamma}}_{EE})(2\pi)^{p/2}\big|n\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\big|^{-1/2}\tilde{A}+ o_p(1)$, with $\tilde{A} = \int \exp\Big\{\frac{1}{2}\boldsymbol{\delta}_1^{\mathrm{T}}(n\boldsymbol{\Sigma}_{11})\boldsymbol{\delta}_1\Big\}\rho_{\boldsymbol{\delta}_1}(\boldsymbol{\delta}_1)d\boldsymbol{\delta}_1$, which finally leads to the conclusion by taking negative two logarithm of marginal probability: $-2\log P(\mathbf{Y}|M) = -2\log R^F(\hat{\boldsymbol{\lambda}}_{EE}, \hat{\boldsymbol{\gamma}}_{EE}) + p\log n + \log|\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}| - p\log(2\pi) - 2\log(\tilde{A}) + o_p(1)$. □

S1.2 Proof of Theorem 2

Proof. Given *Condition* 1 and along the lines of the proofs in Owen (2001), for any $\boldsymbol{\gamma}$ satisfying $\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| \leq Cn^{-1/3}$ with a large enough constant $C > 0$, we have $\hat{\boldsymbol{\lambda}}(\boldsymbol{\gamma}) = o_p(\mathbf{n}^{-1/3})$ and $\max_{1\leq i\leq n}\|\mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma})\| = o_p(n^{-1/3})$. Accordingly, $\max_{1\leq i\leq n}\hat{\boldsymbol{\lambda}}^{\mathrm{T}}(\boldsymbol{\gamma})\mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma}) = o_p(1)$ uniformly for $\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| \leq Cn^{-1/3}$. On the other hand, by the definition of $\hat{\boldsymbol{\lambda}}_{EE}$ and taking first order Taylor expansion at

$\boldsymbol{\gamma}_0$ and $\boldsymbol{\lambda}_0 = \mathbf{0}$, we have the following equation:

$$\begin{aligned}\mathbf{0} =& \frac{1}{n}\frac{\partial l}{\partial \boldsymbol{\lambda}}\Big|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}_{EE},\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}_{EE}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{g}(\mathbf{D}_i,\hat{\boldsymbol{\gamma}}_{EE})}{1+\hat{\boldsymbol{\lambda}}_{EE}\mathbf{g}(\mathbf{D}_i,\hat{\boldsymbol{\gamma}}_{EE})} \\ =& \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{D}_i,\boldsymbol{\gamma}_0) + \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{g}(\mathbf{D}_i,\boldsymbol{\gamma}_0)}{\partial\boldsymbol{\gamma}^{\mathrm{T}}}(\hat{\boldsymbol{\gamma}}_{EE}-\boldsymbol{\gamma}_0) \\ & -\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{D}_i,\boldsymbol{\gamma}_0)\mathbf{g}^{\mathrm{T}}(\mathbf{D}_i,\boldsymbol{\gamma}_0)(\hat{\boldsymbol{\lambda}}_{EE}-\mathbf{0}) + o_p(\boldsymbol{\varepsilon}_n),\end{aligned}$$

where $\boldsymbol{\varepsilon}_n = \|\hat{\boldsymbol{\gamma}}_{EE}-\boldsymbol{\gamma}_0\|+\|\hat{\boldsymbol{\lambda}}_{EE}\|$. By solving $\hat{\boldsymbol{\lambda}}_{EE}$ from the above formula, we have

$$\hat{\boldsymbol{\lambda}}_{EE} = \mathbf{S}_n^{-1}\Big\{\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{D}_i,\boldsymbol{\gamma}_0)+\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{g}(\mathbf{D}_i,\boldsymbol{\gamma}_0)}{\partial\boldsymbol{\gamma}^{\mathrm{T}}}(\hat{\boldsymbol{\gamma}}_{EE}-\boldsymbol{\gamma}_0)+o_p(\boldsymbol{\varepsilon}_n)\Big\}, \quad \text{(S1.4)}$$

where $\mathbf{S}_n = (1/n)\sum_{i=1}^{n}\mathbf{g}(\mathbf{D}_i,\boldsymbol{\gamma}_0)\mathbf{g}^{\mathrm{T}}(\mathbf{D}_i,\boldsymbol{\gamma}_0)$. Also, $\hat{\boldsymbol{\gamma}}_{EE}-\boldsymbol{\gamma}_0 = O_p(\mathbf{n}^{-1/2})$ and $(1/n)\sum_{i=1}^{n}\mathbf{g}(\mathbf{D}_i,\boldsymbol{\gamma}_0) = O_p(\mathbf{n}^{-1/2})$, we conclude that $\boldsymbol{\varepsilon}_n = O_p(\mathbf{n}^{-1/2})$. Thus, by *Condition* 1 and the weak law of large number, $\hat{\boldsymbol{\lambda}}_{EE}$ is rewritten as

$$\hat{\boldsymbol{\lambda}}_{EE} = \boldsymbol{\Sigma}_{11}^{-1}\Big\{\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{D}_i,\boldsymbol{\gamma}_0) + \boldsymbol{\Sigma}_{12}\big(\hat{\boldsymbol{\gamma}}_{EE}-\boldsymbol{\gamma}_0\big)\Big\} + o_p(\mathbf{n}^{-\frac{1}{2}}). \quad \text{(S1.5)}$$

Now let us expand l in the following manner:

$$\begin{aligned} l =& \sum_{i=1}^{n} \log\Big\{1 + \hat{\boldsymbol{\lambda}}_{EE}\mathbf{g}(\mathbf{D}_i, \hat{\boldsymbol{\gamma}}_{EE})\Big\} \\ =& \hat{\boldsymbol{\lambda}}_{EE}^{\mathrm{T}} \sum_{i=1}^{n} \mathbf{g}(\mathbf{D}_i, \hat{\boldsymbol{\gamma}}_{EE}) - \frac{1}{2}\hat{\boldsymbol{\lambda}}_{EE}^{\mathrm{T}}\Big\{\sum_{i=1}^{n} \mathbf{g}(\mathbf{D}_i, \hat{\boldsymbol{\gamma}}_{EE})\mathbf{g}^{\mathrm{T}}(\mathbf{D}_i, \hat{\boldsymbol{\gamma}}_{EE})\Big\}\hat{\boldsymbol{\lambda}}_{EE}^{\mathrm{T}} \\ & + O_p\Big(\sum_{i=1}^{n}\big\{\hat{\boldsymbol{\lambda}}_{EE}^{\mathrm{T}}\mathbf{g}(\mathbf{D}_i, \hat{\boldsymbol{\gamma}}_{EE})\big\}^3\Big). \end{aligned}$$

It is noted that

$$\sum_{i=1}^{n}\big\{\hat{\boldsymbol{\lambda}}_{EE}^{\mathrm{T}}\mathbf{g}(\mathbf{D}_i, \hat{\boldsymbol{\gamma}}_{EE})\big\}^3 \leq \sum_{i=1}^{n}\big\{\hat{\boldsymbol{\lambda}}_{EE}^{\mathrm{T}}\mathbf{g}(\mathbf{D}_i, \hat{\boldsymbol{\gamma}}_{EE})\big\}^2 \max_{1\leq i\leq n} \hat{\boldsymbol{\lambda}}_{EE}^{\mathrm{T}}\mathbf{g}(\mathbf{D}_i, \hat{\boldsymbol{\gamma}}_{EE}).$$

Thus, by realizing that $\max_{1\leq i\leq n} \hat{\boldsymbol{\lambda}}_{EE}^{\mathrm{T}}\mathbf{g}(\mathbf{D}_i, \hat{\boldsymbol{\gamma}}_{EE}) = o_p(1)$ and $\sum_{i=1}^{n}\big\{\hat{\boldsymbol{\lambda}}_{EE}^{\mathrm{T}}\mathbf{g}(\mathbf{D}_i, \hat{\boldsymbol{\gamma}}_{EE})\big\}^2 = O_p(1)$ by *Condition* 1, uniformly hold for $\|\hat{\boldsymbol{\gamma}}_{EE} - \boldsymbol{\gamma}_0\| \leq Cn^{-1/3}$, we have

$$l = \hat{\boldsymbol{\lambda}}_{EE}^{\mathrm{T}} \sum_{i=1}^{n} \mathbf{g}(\mathbf{D}_i, \hat{\boldsymbol{\gamma}}_{EE}) - \frac{1}{2}\hat{\boldsymbol{\lambda}}_{EE}^{\mathrm{T}}\Big\{\sum_{i=1}^{n} \mathbf{g}(\mathbf{D}_i, \hat{\boldsymbol{\gamma}}_{EE})\mathbf{g}^{\mathrm{T}}(\mathbf{D}_i, \hat{\boldsymbol{\gamma}}_{EE})\Big\}\hat{\boldsymbol{\lambda}}_{EE} + o_p(1).$$

Furthermore, by applying the Taylor expansion for $n^{-1/2}\mathbf{g}(\mathbf{D}_i, \hat{\boldsymbol{\gamma}}_{EE})$, we get

$$\begin{aligned} & \hat{\boldsymbol{\lambda}}_{EE}^{\mathrm{T}} \sum_{i=1}^{n} \mathbf{g}(\mathbf{D}_i, \hat{\boldsymbol{\gamma}}_{EE}) \\ =& n^{\frac{1}{2}}\hat{\boldsymbol{\lambda}}_{EE}^{\mathrm{T}}\Big\{n^{-\frac{1}{2}}\sum_{i=1}^{n} \mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma}_0) + n^{-\frac{1}{2}}\sum_{i=1}^{n} \frac{\partial \mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}(\hat{\boldsymbol{\gamma}}_{EE} - \boldsymbol{\gamma}_0) + o_p(\mathbf{1})\Big\} \\ =& \big(n^{\frac{1}{2}}\hat{\boldsymbol{\lambda}}_{EE}^{\mathrm{T}}\big)\mathbf{S}_n\big(n^{\frac{1}{2}}\hat{\boldsymbol{\lambda}}_{EE}\big) + o_p(1). \end{aligned}$$

The last equation is derived from (S1.4). Therefore, by *Condition* 1 again, we have $l = (1/2)(n^{\frac{1}{2}}\hat{\boldsymbol{\lambda}}_{EE}^{\mathrm{T}})\boldsymbol{\Sigma}_{11}(n^{\frac{1}{2}}\hat{\boldsymbol{\lambda}}_{EE}) + o_p(1)$. Substituting $\hat{\boldsymbol{\lambda}}_{EE}$ by (S1.5), $l = (1/2)(n^{\frac{1}{2}}\mathbf{Q}_n^{\mathrm{T}})\boldsymbol{\Sigma}_{11}^{-1}(n^{\frac{1}{2}}\mathbf{Q}_n) + o_p(1)$ with $\mathbf{Q}_n = (1/n)\sum_{i=1}^{n}\mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma}_0) + \boldsymbol{\Sigma}_{12}(\hat{\boldsymbol{\gamma}}_{EE} - \boldsymbol{\gamma}_0)$, which completes the proof of Theorem 2. □

S1.3 Proof of Corollary 1

Proof. Let us rewrite $\mathbf{Q}_n$ in the following manner:

$$\begin{aligned}\mathbf{Q}_n =& \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma}_0) - \boldsymbol{\Sigma}_{12}\Big(E\frac{\partial \mathbf{g}_1}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\Big)^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_1(\mathbf{D}_i, \boldsymbol{\gamma}_0) + o_p(\mathbf{n}^{-\frac{1}{2}}) \\ =& \boldsymbol{\Sigma}_*\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma}_0) + o_p(\mathbf{n}^{-\frac{1}{2}}),\end{aligned}$$

with $\boldsymbol{\Sigma}_* = \mathbf{I}_{L\times L} - \Big(\boldsymbol{\Sigma}_{12}\{E(\partial \mathbf{g}_1/\partial \boldsymbol{\gamma}^{\mathrm{T}})\}^{-1}, \mathbf{0}_{L\times(L-p)}\Big)$. Here $\mathbf{I}_{L\times L}$ is an identity matrix. Accordingly, we rewrite l as

$$l = \frac{1}{2}\left\{n^{-\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma}_0)\right\}^{\mathrm{T}}\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\boldsymbol{\Sigma}_*^{\mathrm{T}}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_*\boldsymbol{\Sigma}_{11}^{\frac{1}{2}}\left\{n^{-\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma}_0)\right\} + o_p(1).$$

Furthermore, according to the square-root decomposition, we have $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_{11}^{1/2}\boldsymbol{\Sigma}_*^{\mathrm{T}}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_*\boldsymbol{\Sigma}_{11}^{1/2} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{\mathrm{T}}$ with an orthogonal matrix $\mathbf{P}_{L\times\tilde{L}}$ and a diagonal matrix $\boldsymbol{\Lambda}_{L\times\tilde{L}}$ containing positive eigenvalues $\Lambda_1,\ldots,\Lambda_{\tilde{L}}$ of $\boldsymbol{\Omega}$. Here $\tilde{L}$ represents the rank of the matrix $\boldsymbol{\Omega}$ and L is the length of $\mathbf{g}(\mathbf{D}, \boldsymbol{\gamma})$. Therefore, l can be

expressed as

$$l = \frac{1}{2}\sum_{j=1}^{\tilde{L}} \Lambda_j \Big[n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{\Sigma}_{11}^{-\frac{1}{2}} \mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma}_0) \Big]_j^2 + o_p(1),$$

where $[\mathbf{J}]_j$ indicates the j^{th} element in vector $\mathbf{J}$. Together with the fact of $\big[n^{-1/2}\sum_{i=1}^{n} \mathbf{\Sigma}_{11}^{-1/2}\mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma}_0)\big]_j$ asymptotically followed by an independent standard normal distribution for $j = 1, \ldots, \tilde{L}$, we conclude that $2l$ converges in distribution to $\sum_{j=1}^{\tilde{L}} \Lambda_j \chi_1^2$, which is a weighted sum of standard χ^2 distributions.

□

S1.4 Proof of Theorem 3

Proof. For any $\boldsymbol{\gamma}$ in the neighbourhood of $\boldsymbol{\gamma}_* \neq \boldsymbol{\gamma}_0$, we define $\tilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma}) = n^{-c}(\log n)\bar{\mathbf{g}}_n$, with $\bar{\mathbf{g}}_n = (1/n)\sum_{i=1}^{n} \mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma})$ and $1/2 < c < 1$.

First by Markov inequality and *Condition* 3 with some $\delta > 0$, we have

$$\sum_{i=1}^{\infty} P\Big(\|\mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma})\|^2 > i \Big) \leq \sum_{i=1}^{\infty} \frac{E\|\mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma})\|^{2+\delta}}{i^{1+\delta/2}} < \infty.$$

Applying the Borel-Cantelli Lemma, we conclude that we can always find a large enough N such that for any $i > N$, we have$\|\mathbf{g}(\mathbf{D}_i, \boldsymbol{\gamma})\| \leq i^{-1/2}$ holds with

probability one, which further implies $\max_{1\leq i\leq n}\|\mathbf{g}(\mathbf{D}_i,\boldsymbol{\gamma})\| = o_p(n^{1/2})$. Thus,

$$\begin{aligned}&\max_{1\leq i\leq n}\|\tilde{\boldsymbol{\lambda}}^{\mathrm{T}}(\boldsymbol{\gamma})\mathbf{g}(\mathbf{D}_i,\boldsymbol{\gamma})\|\\ \leq&\|\tilde{\boldsymbol{\lambda}}(\boldsymbol{\gamma})\|\max_{1\leq i\leq n}\|\mathbf{g}(\mathbf{D}_i,\boldsymbol{\gamma})\| = n^{\frac{1}{2}-c}\log(n)\|\bar{\mathbf{g}}_n\| = o_p(1),\end{aligned}$$

where the the first inequality holds by Cauchy-Schwartz inequality and the last equality holds by *Condition* 4. Therefore, with probability approaching to 1, we have for all $1 \leq i \leq n$, $1 + \tilde{\boldsymbol{\lambda}}^{\mathrm{T}}(\boldsymbol{\gamma})\mathbf{g}(\mathbf{D}_i,\boldsymbol{\gamma}) > 0$. Finally,

$$\begin{aligned}l &= \sup_{\boldsymbol{\lambda}}\sum_{i=1}^{n}\log\left\{1+\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}(\mathbf{D}_i,\boldsymbol{\gamma})\right\} \geq \sum_{i=1}^{n}\log\left\{1+\tilde{\boldsymbol{\lambda}}^{\mathrm{T}}(\boldsymbol{\gamma})\mathbf{g}(\mathbf{D}_i,\boldsymbol{\gamma})\right\}\\ &= \sum_{i=1}^{n}\tilde{\boldsymbol{\lambda}}^{\mathrm{T}}(\boldsymbol{\gamma})\mathbf{g}(\mathbf{D}_i,\boldsymbol{\gamma}) + o_p(1) = n^{1-c}\|\bar{\mathbf{g}}_n\|^2\log(n) + o_p(1),\end{aligned}$$

where the first equality holds by the property of the dual problem, the second equality holds by the first-order Taylor expansion of the function $\log(1+x)$ at 0, and the final result holds under *Condition* 4. □

S1.5 Proof of Theorem 4

Proof. Given the fixed number of parameters in the full model, and p and p_0 as the cardinalities of candidate model M and the true model M_0, respectively, we show Theorem 4 in the following manner, which can be easily proved by the results from Theorems 2 and 3. **(I)** $\mathrm{ELCIC}(M) - \mathrm{ELCIC}(M_0) > 0$ with

probability tending to 1 for $M_0 \nsubseteq M$.

(II) ELCIC(M) − ELCIC$(M_0) > 0$ with probability tending to 1 for $M_0 \subseteq M$ and $p_0 < p$.

First, for $M_0 \nsubseteq M$, applying Theorem 2 to ELCIC(M_0) and Theorem 3 to ELCIC(M), we derive

$$\begin{aligned}\text{ELCIC}(M) - \text{ELCIC}(M_0) &= 2l(M) + p\log(n) - \big\{2l(M_0) + p_0\log(n)\big\} \\ &= 2n^{1-c}\|\bar{\mathbf{g}}_n\|^2\log(n) - (n^{\frac{1}{2}}\mathbf{Q}_{n0}^{\mathrm{T}})\boldsymbol{\Sigma}_{110}^{-1}(n^{\frac{1}{2}}\mathbf{Q}_{n0}) \\ &\quad + (p - p_0)\log(n) + o_p(1),\end{aligned}$$

where $\mathbf{Q}_{n0}$ and $\boldsymbol{\Sigma}_{110}$ denote $\mathbf{Q}_n$ and $\boldsymbol{\Sigma}_{11}$ under the true model M_0. Notice that

$$P\Big[(n^{\frac{1}{2}}\mathbf{Q}_{n0}^{\mathrm{T}})\boldsymbol{\Sigma}_{110}^{-1}(n^{\frac{1}{2}}\mathbf{Q}_{n0}) \geq \log n\Big] \leq \frac{E\big(n\mathbf{Q}_{n0}^{\mathrm{T}}\boldsymbol{\Sigma}_{110}^{-1}\mathbf{Q}_{n0}\big)}{\log(n)} = \frac{tr(\boldsymbol{\Sigma}_{110}^{-1}\mathbf{V})}{\log(n)}.$$

Applying *Condition* 5, we have $(n^{1/2}\mathbf{Q}_{n0}^{\mathrm{T}})\boldsymbol{\Sigma}_{110}^{-1}(n^{1/2}\mathbf{Q}_{n0}) = o_p(\log n)$, which further indicates that $n^{1-c}\|\bar{\mathbf{g}}_n\|^2\log(n)$ is the dominant term going to infinity under *Condition* 4. Therefore, we have ELCIC(M) − ELCIC$(M_0) > 0$ with probability tending to 1 for $M_0 \nsubseteq M$.

Second, for $M_0 \subseteq M$ and $p_0 < p$, applying Theorem 2 to ELCIC(M) and

ELCIC(M_0), we can derive

$$\begin{aligned}\text{ELCIC}(M) - \text{ELCIC}(M_0) =&(n^{\frac{1}{2}}\mathbf{Q}_n^{\mathrm{T}})\boldsymbol{\Sigma}_{11}^{-1}(n^{\frac{1}{2}}\mathbf{Q}_n) - (n^{\frac{1}{2}}\mathbf{Q}_{n0}^{\mathrm{T}})\boldsymbol{\Sigma}_{110}^{-1}(n^{\frac{1}{2}}\mathbf{Q}_{n0})\\ &+ (p - p_0)\log(n) + o_p(1).\end{aligned}$$

Since $(n^{1/2}\mathbf{Q}_{n0}^{\mathrm{T}})\boldsymbol{\Sigma}_{110}^{-1}(n^{1/2}\mathbf{Q}_{n0})$ and $(n^{1/2}\mathbf{Q}_n^{\mathrm{T}})\boldsymbol{\Sigma}_{11}^{-1}(n^{1/2}\mathbf{Q}_n)$ have the same order $o_p(\log n)$ by the same argument above, we conclude that, for $M_0 \subseteq M$ and $p_0 < p$, $\lim_{n\to\infty} P(\text{ELCIC}(M) - \text{ELCIC}(M_0) > 0) = 1$. □

S2. Additional Simulation Studies

S2.1 Variable selection in Case 2 under the GEE framework

We consider the same setups in Case 2 in the main paper to only implement the variable selection by using the first part estimating equations in (3.11) in the main paper as our full estimating equations, and thus treating correlation coefficients as nuisance parameters. The selection rates by ELCIC and QIC are summarized in table 1, which further confirm the outperformance of ELCIC.

S2.2 Variable selection for the augmented inverse probability weighting method

To show unique and more general applications of our proposed criteria compared to the existing approaches, we provide an example for illustration, and under such context, the current existing criteria are not applicable. Here, we consider the augmented inverse probability weighted (AIPW) models with main focus on variable selection in the mean structure. Note the AIPW method has been popularly used to deal with missing data (Robins et al., 1994) with extensive work in longitudinal data, survival analysis and causal inference (Bang and Robins, 2005; Seaman and Copas, 2009; Scharfstein et al., 1999; Long et al., 2011) because of efficiency improvement and double robustness. For simplicity, here we only consider a simple linear regression with missing outcomes under the assumption of missing at random (MAR), but the extension to more complicated scenarios should be doable and straightforward.

Suppose, for $i = 1, \ldots, n$, we have the data where the outcomes Y_i potentially missing, with R_i as an observation indicator, i.e., $R_i = 1$ if Y_i is observed and $R_i = 0$ otherwise. The covariates include $\mathbf{X}_i$ and $\mathbf{S}_i$. Also, we denote the observing probability of Y_i as $\pi(\mathbf{X}_i, \mathbf{S}_i) = E(R_i|\mathbf{X}_i, \mathbf{S}_i)$ parameterized by $\boldsymbol{\gamma}$. The AIPW estimators are obtained by solving the following estimating equa-

tions

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{R_i}{\hat{\pi}(\mathbf{X}_i,\mathbf{S}_i)}\mathbf{U}(Y_i,\mathbf{X}_i,\boldsymbol{\beta})-\frac{R_i-\hat{\pi}(\mathbf{X}_i,\mathbf{S}_i)}{\hat{\pi}(\mathbf{X}_i,\mathbf{S}_i)}\tilde{\mathbf{U}}(\mathbf{X}_i,\mathbf{S}_i,\boldsymbol{\beta})\right\}=\mathbf{0} \quad \text{(S2.6)}$$

where $\mathbf{U}(Y_i,\mathbf{X}_i,\boldsymbol{\beta}) = \mathbf{X}_i(Y_i-\mu_i(\boldsymbol{\beta}))$ and $\tilde{\mathbf{U}}(\mathbf{X}_i,\mathbf{S}_i,\boldsymbol{\beta}) = \mathbf{X}_i(\hat{a}(\mathbf{X}_i,\mathbf{S}_i)-\mu_i(\boldsymbol{\beta}))$, where $a(\cdot)$ is a function of $\mathbf{X}_i$ and $\mathbf{S}_i$ parameterized by $\boldsymbol{\alpha}$ with $\hat{a}$ as some estimate of $E(Y_i|\mathbf{X}_i,\mathbf{S}_i)$.

As indicated in the literature, in addition to possible efficiency gains, one advantage of the AIPW estimator is that it is doubly robust, in the sense that it yields consistent results if either the missingness mechanism or the outcome regression model is correctly specified (Scharfstein et al., 1999). However, it is challenging to correct specify $\hat{\pi}(\mathbf{X}_i,\mathbf{S}_i)$ or $\hat{a}(\mathbf{X}_i,\mathbf{S}_i)$ in practice due to limited prior knowledge, and thus the methods based on expectation of weighted quadratic mean square loss may not work (Shen and Chen, 2012, 2018). Also, likelihood based criteria are not applicable since semi-parametric approach is implemented here. In addition, the estimated quantities $\hat{\pi}(\mathbf{X}_i,\mathbf{S}_i)$ and $\hat{a}(\mathbf{X}_i,\mathbf{S}_i)$ make the model selection harder. However, our proposed ELCIC has great potential to deal with these issues, as we indicated in our method section, and the implementation is easy and straightforward. In particular, we take the formula on the left hand side of (S2.6) as the full estimating equations in (2.5) in the

main text, and also the nuisance parameters involved in $\hat{\pi}(\mathbf{X}_i, \mathbf{S}_i)$ and $\hat{a}(\mathbf{X}_i, \mathbf{S}_i)$ can be estimated from plug-in estimators. The consistency property still holds as long as these parameter estimates satisfy Condition 5 in the main text.

We conduct extensive simulation studies to empirically evaluate the performance of ELCIC under the AIPW framework. Along with similar data structure and generation procedure in Han (2014), we first generate four mutually independent covariates $x_{1i} \sim$N$(5, 1)$, $x_{2i} \sim$Bernoulli(0.5), $x_{3i} \sim$N$(0, 1)$, $x_{4i} \sim$N$(0, 1)$ and four auxiliary variables $s_{1i} = 1 + x_{1i} + 2x_{2i} + \epsilon_{2i}$, $s_{2i} = I\{(s_{1i} + 0.3\epsilon_{3i}) > 5.8\}$, $s_{3i} = \epsilon_{4i}$,$s_{4i} = x_2 + \epsilon_{5i}$, where $\boldsymbol{\epsilon}_i = (\epsilon_{1i}, \ldots, \epsilon_{5i})^T$ follow a multivariate normal distribution with mean zeros and the covariance matrix $\boldsymbol{\Sigma} =$ with the diagonal elements valued by 1, the $(1, 2)$ and $(2, 1)$ entries as 0.5 and all others as 0. The outcomes are generated from the linear model $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \epsilon_{1i}$ with $\boldsymbol{\beta} = (1, 1, 2, 1, 1)^T$. The true observing probability model is set to be logit$(\pi(\mathbf{X}_i, \mathbf{S}_i)) = \gamma_0 + \gamma_1 s_{1i} + \gamma_2 s_{2i}$ with $\boldsymbol{\gamma} = (5, -1, 3)^T$, leading to the observing probability around 0.65. We can also easily learn that the true imputation model should be $a(\mathbf{X}_i, \mathbf{S}_i) = \alpha_0 + \alpha_i x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \alpha_4 x_{4i} + \alpha_5 s_{1i}$. To further evaluate the effect of misspecification of either the model for missingness or the outcome regression model on our proposal's performance, we consider the following misspecified models: logit$(\pi^m(\mathbf{X}_i, \mathbf{S}_i)) = \gamma_0^m + \gamma_1^m x_{1i} + \gamma_2^m x_{2i} + \gamma_3^m x_{3i} + \gamma_4^m x_{4i} + \gamma_5^m s_{1i}$

and $a^m(\mathbf{X}_i, \mathbf{S}_i) = \alpha_0^m + \alpha_1^m s_{1i} + \alpha_2^m s_{2i} + \alpha_3^m s_{3i}$.

Thereafter, the variable selection is implemented based on our proposed EL-CIC under four combinations evaluated: correct-specified $\pi(\mathbf{X}_i, \mathbf{S}_i)$ and correct-specified $a(\mathbf{X}_i, \mathbf{S}_i)$, correct-specified $\pi(\mathbf{X}_i, \mathbf{S}_i)$ and misspecified $a^m(\mathbf{X}_i, \mathbf{S}_i)$, misspecified $\pi^m(\mathbf{X}_i, \mathbf{S}_i)$ and correct-specified $a(\mathbf{X}_i, \mathbf{S}_i)$, denoted by PC_IC, PC_IM, and PM_IC, respectively. Note that we do not consider the case with both misspecified $\pi^m(\mathbf{X}_i, \mathbf{S}_i)$ and $a^m(\mathbf{X}_i, \mathbf{S}_i)$ because the estimates will not be consistent,and the results based upon these inconsistent estimates are not reliable any more. For each scenario, we generate 500 Monte carlo data with sample size $n = 250, 500$. Seven candidate models are considered for variable selection with selection rates recorded. The results are summarized in Table 2 in the Appendix. Overall, the selection rates for the correct mean structure are satisfactory with a high level (i.e., $> 90\%$) when either of the models for missingness and outcome regression is correctly-specified, and increase as sample size becomes larger. In particular, when sample size $n = 250$, the selection rate is up to 93.2% when both models are correctly specified. For larger sample size (i.e., $n = 500$), the results for PC_IC, PC_IM, PM_IC are comparable, indicating our proposal is workable in the APIW framework.

S2.3 Variable Selection in Ultra-high-dimensional Cases

As suggested by the editor and one of the reviewers of the present paper, we conducted additional investigations via simulation studies into a potential extension of ELCIC to in ultra-high-dimensional cases. The theoretical proofs for (ultra-)high-dimensional cases with empirical likelihood will differ completely from those with our current strategy and will not be trivial to extend without expending more effort, which we will pursue separately and fully in future work. Here, we provide some empirical studies to show the challenges and numerical performance via simulation.

Note that this extension involves two main issues. One is that the full model in ultra-high-dimensional cases could be substantially large, which may prevent the use of the classic empirical likelihood (Chen et al., 2009). A tentative strategy is to consider the following modified criterion:

$$\begin{aligned} \text{ELCIC}^* = & -2\log R^R(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\gamma}}) - n\sum_{j=1}^{L} P_{2,\nu}(|\hat{\lambda}_j|) \\ & + n\sum_{k=1}^{p} P_{1,\pi}(|\hat{\gamma}_k|) + C_n \log(n)\text{df}_\pi, \end{aligned} \tag{S2.7}$$

where $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\gamma}}$ are solved using penalized empirical likelihood (Chang et al., 2018). $P_{1,\pi}(\cdot)$ and $P_{2,\nu}(\cdot)$ are two penalty functions regulating $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ with tuning parameters π and ν, respectively. C_n is a scaling factor that diverges to

infinity at a slow rate (Tang and Leng, 2010). df_π is the number of nonzero coefficients in $\boldsymbol{\gamma}$. Suggested by Tang and Leng (2010) and Chang et al. (2018), the *ad hoc* criterion (S2.7) works well numerically, but its theoretical properties would be difficult to investigate because of the penalty functions with two extra tuning parameters. The second issue is the complicated and unstable computational algorithm with high computational burden, which may lead to unsatisfactory results especially when using generalized estimating equations embedded with empirical likelihood (Chang et al., 2018). Therefore, it would be neither straightforward nor feasible to apply (S2.7) in ultra-high-dimensional cases.

Here, we propose an alternative two-stage selection procedure to resolve these issues but still allow our proposed criterion to be easily implemented and perform satisfactorily. In the first stage, we apply a nonparametric screening method, such as SIRS (Zhu et al., 2011), to reduce the ultra-high dimension to a relatively low one. Then, in the second stage, we use the reduced model as the full model for ELCIC to capture the true one. Note that our proposed strategy has two advantages. The first is consistency: both the SIRS procedure and ELCIC are proven to be consistent, thereby the combined two-stage selection in consistent. The second is easy and flexible implementation in practice: below, we provide some numerical evaluations to show the utility and advantage of ELCIC over other popular criteria.

We consider the mean structure $\log(\mu_i) = \mathbf{X}_i^T\boldsymbol{\beta}$ for $i = 1, \ldots, n$, where $\mathbf{X}_i$ is a covariate vector from a $p_n = 1500-$dimensional multivariate normal distribution **MVN**(0,$\mathbf{V}$) with the covariance matrix $\mathbf{V}$ as an AR1 matrix with unit variance and a correlation coefficient of 0.5. We generate outcomes from a Poisson distribution and a negative binomial distribution with $k = 2, 4$, or 8 failures. First, we apply SIRS to reduce the ultra-high dimension (i.e., $p_n = 1500$) to a relatively low one ($p_n = 20$). Then, we apply the SCAD learning procedure to implement variable selection to the reduced candidate pool, where the tuning parameter is selected by cross-validation, BIC, and ELCIC. Note that BIC is always set under the assumption of a Poisson distribution, so that we can observe how misspecification of distribution affects the performance of variable selection. We generate 200 Monte Carlo data with sample size $n = 250$ or 500, and we record several important measurements, such as consistency, prediction error, and false-negative, false-positive, and exact-selection rates. The results are summarized in Table 3 herein.

As shown, when the outcomes are generated from a Poisson distribution, BIC performs best in terms of its low false-positive rate and high exact-selection rate. However, when the true distribution is a negative binomial, BIC tends to select an over-fitted model, and the performance deteriorates as the data have higher over-dispersion. By contrast, ELCIC is robust in handling distribution

misspecification and having more-stable selection performance. The performance of cross-validation is between those of ELCIC and BIC but more time consuming. In addition, as the sample size increases up to 750, ELCIC tends to have a better selection rate and fewer false positives while BIC has negligibly improved performance, and cross-validation does not always lead to fewer false positives. Therefore, these numerical results support the potential feasibility of applying ELCIC to ultra-high-dimensional cases, having a much-reduced computational burden and highly stable performance in the meantime without sacrificing the distribution-free advantage.

S3. Further Discussion

Note that in the main context, we assume that the full model is specified correctly. However, with reference to Conditions 1 and 4, note also that the identifiability of ELCIC indeed does not require the full model to be specified correctly. We require only the existence of parameters that make the full estimating equations equal to zero. Consequently, if the full model that we use is misspecified, ELCIC will still work but only by locating the "true" values that make the current estimating equations equal to zero. However, the existence of such "true" values is theoretically nontrivial and also not always guaranteed, which may require more investigation in future work. Thus, in practice, we recommend inputting

all potentially important variables so that the selected full model fits the data well if a sufficient sample size is provided.

Extensive numerical studies show that ELCIC outperforms other alternatives. GIC is somewhat sensitive to distribution misspecification, even under a large sample size, for two possible reasons. 1) A relatively complicated bias-correction term must be estimated based upon the data structure and candidate models, thereby leading to more variability to the selection. 2) The underlying measurement of GIC is defined as the Kullback–Leibler distance evaluated at the misspecified distribution, which could introduce a systematic discrepancy to the truth, no matter how accurate is the bias correction. Therefore, GIC would intrinsically lose some power to capture the true model, particularly when the specified distribution deviates farther from the truth. The same issue also pertains to QIC, regarding its misspecification of quasi-likelihood by considering the independent correlation structure (Pan, 2001).

References

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61,** 962–973.

Chang, J., Tang, C., and Wu, T. (2018). A new scope of penalized empirical like-

lihood with high-dimensional estimating equations. *The Annals of Statistics* **46(6B),** 3185–3216.

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

Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association* **109,** 1159–1173.

Long, Q., Zhang, X., and Johnson, B. A. (2011). Robust estimation of area under roc curve using auxiliary variables in the presence of missing biomarker values. *Biometrics* **67,** 559–567.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics* **4,** 2111–2245.

Owen, A. B. (2001). *Empirical likelihood*. CRC press.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics* **57(1),** 120–125.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics* **22,** 300–325.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression

coefficients when some regressors are not always observed. *Journal of the American statistical Association* **89,** 846–866.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for non-ignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94,** 1096–1120.

Seaman, S. and Copas, A. (2009). Doubly robust generalized estimating equations for longitudinal data. *Statistics in medicine* **28,** 937–955.

Shen, C. W. and Chen, Y. H. (2012). Model selection for generalized estimating equations accommodating dropout missingness. *Biometrics* **68,** 1046–1054.

Shen, C. W. and Chen, Y. H. (2018). Joint model selection of marginal mean regression and correlation structure for longitudinal data with missing outcome and covariates. *Biometrical Journal* **60,** 20–33.

Tang, C. Y. and Leng, C. (2010). Penalized high-dimensional empirical likelihood. *Biometrika* **97(4),** 905–920.

Zhu, L. P., Li, L., Li, R., and Zhu, L. X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association* **106,** 1464–1475.

Table 1: Performance of ELCIC compared with QIC for the scenarios under longitudinal count data. 500 Monte Carlo datasets are generated with sample size $n = 100, 300$ and the number of observations within-subject $T = 3, 5$. The true mean structure is $\{x_1, x_2\}$ and an exchangeable (EXC) correlation structure with the correlation coefficient $\rho = 0.5$ is the true model. Only variable selection is considered.

Cluster Size	n	Criteria	Candidate Models					
			x_1, x_2, x_3	$\boldsymbol{x_1, x_2}$	x_1, x_3	x_2, x_3	x_1	x_3
$T = 3$	100	ELCIC	0.042	0.956	0	0.002	0	0
		QIC	0.14	0.86	0	0	0	0
	300	ELCIC	0.022	0.978	0	0	0	0
		QIC	0.116	0.884	0	0	0	0
$T = 5$	100	ELCIC	0.034	0.966	0	0	0	0
		QIC	0.11	0.89	0	0	0	0
	300	ELCIC	0.018	0.982	0	0	0	0
		QIC	0.1	0.9	0	0	0	0

Table 2: Performance of ELCIC for variable selection in the mean structure under the AIPW framework. 500 Monte Carlo data are generated with sample size $n = 250, 500$. The model with $\{x_1, x_2, x_3, x_4\}$ is the true model. PC_IC: correct-specified $\pi(\mathbf{X}_i, \mathbf{S}_i)$ and correct-specified $a(\mathbf{X}_i, \mathbf{S}_i)$; PC_IM: correct-specified $\pi(\mathbf{X}_i, \mathbf{S}_i)$ and misspecified $a^m(\mathbf{X}_i, \mathbf{S}_i)$; PM_IC: misspecified $\pi^m(\mathbf{X}_i, \mathbf{S}_i)$ and correct-specified $a(\mathbf{X}_i, \mathbf{S}_i)$.

Model	n	x_1, x_2	x_1, x_2, x_3	x_1, x_2 x_3, s_3	$\boldsymbol{x_1, x_2}$ $\boldsymbol{x_3, x_4}$	x_1, x_2, x_3 x_4, s_3	x_1, x_2, x_3 x_4, s_4	x_1, x_2, x_3 x_4, s_3, s_4
PC_IC	250	0	0	0	0.932	0.028	0.040	0
PC_IC	500	0	0	0	0.950	0.026	0.024	0
PC_IM	250	0	0	0	0.916	0.036	0.046	0.002
PC_IM	500	0	0	0	0.952	0.028	0.020	0
PM_IC	250	0	0	0	0.928	0.03	0.042	0
PM_IC	500	0	0	0	0.958	0.022	0.020	0

Table 3: Performance of ELCIC compared with cross-validation (CV) and BIC (under Poisson distribution specification without over-dispersion) for two stage ultra-high variable selection in the mean structure under Poisson distribution with potential over-dispersed outcomes. 500 Monte Carlo data are generated with sample size $n = 250, 500, 750$. NB: negative binomial; MS: consistency $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$; PS: prediction $\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}_0\|^2$; FN: false negative; FP: false positive; ES: exact selection rate; OS: over selection rate; US: under selection rate

Scenario	n	Criterion	MS	PS	FN	FP	ES	OS	US
POISSON	250	CV	0.22	0.15	0.34	1.32	0.52	0.24	0.10
		BIC	0.17	0.10	0.23	1.27	0.74	0.06	0.03
		ELCIC	0.17	0.11	0.27	1.05	0.59	0.20	0.12
	500	CV	0.07	0.06	0.09	0.70	0.71	0.25	0.04
		BIC	0.04	0.02	0.02	0.09	0.96	0.03	0.01
		ELCIC	0.03	0.03	0.04	0.26	0.89	0.10	0.02
	750	CV	0.05	0.04	0.04	0.68	0.72	0.27	0.02
		BIC	0.02	0.01	0.00	0.02	0.98	0.02	0.00
		ELCIC	0.02	0.01	0.01	0.13	0.92	0.08	0.01
NB $k = 2$	250	CV	0.61	0.43	0.82	4.71	0.05	0.42	0.08
		BIC	0.63	0.46	0.57	9.35	0.00	0.54	0.01
		ELCIC	0.53	0.38	0.69	4.08	0.07	0.43	0.08
	500	CV	0.19	0.14	0.13	4.06	0.17	0.74	0.05
		BIC	0.20	0.17	0.02	10.19	0.00	0.99	0.00
		ELCIC	0.15	0.12	0.07	2.68	0.27	0.69	0.04
	750	CV	0.12	0.09	0.05	4.00	0.24	0.74	0.03
		BIC	0.15	0.13	0.00	10.94	0.00	1.00	0.00
		ELCIC	0.08	0.07	0.00	1.91	0.39	0.61	0.00
NB k=4	250	CV	0.32	0.20	0.45	3.44	0.13	0.52	0.05
		BIC	0.32	0.22	0.33	7.28	0.03	0.67	0.00
		ELCIC	0.28	0.18	0.37	2.66	0.23	0.46	0.07
	500	CV	0.07	0.05	0.02	3.40	0.34	0.64	0.02
		BIC	0.09	0.08	0.00	7.98	0.01	1.00	0.00
		ELCIC	0.05	0.04	0.00	1.83	0.51	0.50	0.00
	750	CV	0.05	0.04	0.00	3.05	0.47	0.54	0.00
		BIC	0.07	0.06	0.00	8.65	0.01	0.99	0.00
		ELCIC	0.04	0.03	0.00	1.34	0.59	0.42	0.00
NB $k = 6$	250	CV	0.27	0.17	0.40	3.36	0.24	0.45	0.06
		BIC	0.29	0.19	0.33	6.27	0.08	0.63	0.00
		ELCIC	0.26	0.16	0.37	2.55	0.28	0.42	0.09
	500	CV	0.05	0.05	0.03	2.42	0.47	0.52	0.01
		BIC	0.06	0.05	0.01	6.43	0.08	0.92	0.00
		ELCIC	0.05	0.04	0.02	1.72	0.54	0.46	0.01
	750	CV	0.03	0.03	0.00	2.99	0.49	0.51	0.00
		BIC	0.05	0.04	0.00	7.41	0.02	0.98	0.00
		ELCIC	0.03	0.02	0.00	1.33	0.61	0.39	0.00

Table 4: Analysis of the ARIC study based on six candidate marginal mean and potential correlation structures EXC: exchangeability; AR1: auto-correlation 1; IND: independence

Variables		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Time		-0.025	-0.025	-0.025	-0.026	-0.026	-0.025
		(0.002)*	(0.002)*	(0.002)*	(0.002)*	(0.002)*	(0.002)*
Gender		0.167	0.174	0.171	0.171	0.162	0.167
		(0.014)*	(0.013)*	(0.013)*	(0.013)*	(0.013)*	(0.014)*
Smoke		0.017					0.016
		(0.014)					(0.014)
Age(year)		-0.032		-0.029	-0.026	-0.033	-0.033
		(0.012)*		(0.012)*	(0.012)*	(0.012)*	(0.012)*
Diabetes		-0.058			-0.054	-0.055	-0.056
		(0.020)*			(0.019)*	(0.019)*	(0.019)*
BMI		0.001					
		(0.001)					
Cholesterol		0.024				0.025	0.024
		(0.006)*				(0.006)*	(0.006)*
Triglycerides		0.002					0.002
		(0.005)					(0.005)
ELCIC	EXC	91.8	81.6	82.9	80.3	69.5	84.4
	AR1	86.7	84.5	85.4	85.4	**68.4**	80.6
	IND	963.1	893.4	949.4	959.1	956	948.8
QIC	AR1	-183571.2	-183485.4	-183508.6	-183529	**-183572**	-183571.6

Note that estimations with standard errors (in parenthesis) are reported; the criteria of ELCIC and QIC are summarized for model selection; *denotes the p-value< 0.05.