Sufficient Identification Conditions and Semiparametric Estimation under Missing Not at Random Mechanisms
(Supplementary Material)

Anna Guo[1] **Jiwei Zhao**[2] **Razieh Nabi**[1]

[1]Dept. of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, USA
[2]Dept. of Biostatistics & Medical Informatics, University of Wisconsin, Madison, Wisconsin, USA

The appendix is organized as follows. In Appendix A, we provide a counterexample for lack of target law identification in the criss-cross MNAR model using continuous variables under normal distributions. Appendix B contains our identification proofs in the exponential family distribution: target law with univariate X (B.1), target law with multivariate X (B.2) and full law (B.3). In Appendix C, we include several examples on parametric identification of popular distributions in the exponential family distributions. Appendix D contains our proofs regarding asymptotic behaviors of our suggested estimators for conditional likelihood with order statistics (D.1) and generalized method of moments (D.2). In Appendix E, we provide additional discussions on (non)parametric estimation approaches. Appendix F contains additional experiments.

A COUNTEREXAMPLE FOR LACK OF TARGET LAW IDENTIFICATION

Consider two distinct distributions p_1 and p_2 defined over variables in $\{X, Y, R_x, R_y\}$ as follows:

Model 1: $Y \sim \mathbb{N}(1,1)$, $X \mid Y \sim \mathbb{N}(y,1)$, $p_1(R_x = 1 \mid y) = \frac{\sqrt{5/6}}{\sqrt{5/6}+\exp\left[-\frac{1}{12}(y-1)^2\right]}$, and

$$p_1(R_y = 1 \mid x, R_x) = \begin{cases} \phi(x), \text{ when } R_x = 1 \\ \phi(\frac{x-5}{\sqrt{5}}), \text{ when } R_x = 0 \end{cases}$$

Model 2: $Y \sim \mathbb{N}(1, \frac{6}{5})$, $X \mid Y \sim \mathbb{N}(y,1)$, $p_2(R_x = 1 \mid y) = \frac{\exp\left[-\frac{1}{12}(y-1)^2\right]}{\sqrt{5/6}+\exp\left[-\frac{1}{12}(y-1)^2\right]}$, and

$$p_2(R_y = 1 \mid x, R_x) = \begin{cases} \phi(x), \text{ when } R_x = 1 \\ \exp(-\frac{8}{9}) * \sqrt{\frac{2}{5}}\phi(x - \frac{7}{3}), \text{ when } R_x = 0. \end{cases}$$

Here $\phi(.)$ denotes the standard normal CDF, and $p_i(x, y, R_x, R_y) = p_i(y)\, p(x \mid y)\, p_i(R_x \mid y)\, p_i(R_y \mid x, R_x)$, $i = 1, 2$. Note that $p_1 \neq p_2$. In what follows, we analyze the four missingness patterns one by one and show that the above two models map to the exact same observed data distribution and thus the target law is not identifiable as a unique function of the observed data law.

1. Missingness pattern $(R_x = 1, R_y = 1)$. We need to prove

$$p_1(x, y, R_x = 1, R_y = 1) = p_2(x, y, R_x = 1, R_y = 1).$$

   This holds since

$$\begin{aligned} & p_1(y)\, p(x \mid y)\, p_1(R_x = 1 \mid y)\, p_1\left(R_y = 1 \mid x, R_x = 1\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y-1)^2\right\} \times p(x \mid y) \times \frac{\sqrt{\frac{5}{6}}}{\sqrt{\frac{5}{6}} + \exp\left[-\frac{1}{12}(y-1)^2\right]} \times \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\} \end{aligned}$$

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

$$
\begin{aligned}
&= \frac{1}{\sqrt{2\pi}\sqrt{\frac{6}{5}}} \exp\left\{-\frac{1}{2\times\frac{6}{5}}(y-1)^2\right\} \times p(x\mid y) \times \frac{\exp\left[-\frac{1}{12}(y-1)^2\right]}{\sqrt{\frac{5}{6}}+\exp\left[-\frac{1}{12}(y-1)^2\right]} \times \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}x^2\right\} \\
&= p_2(y)\, p(x\mid y)\, p_2(R_x=1\mid y)\, p_2\left(R_y=1\mid x, R_x=1\right).
\end{aligned}
$$

2. <u>Missingness pattern $(R_x=1, R_y=0)$.</u> We need to prove

$$
\int p_1(x,y,R_x=1,R_y=0)dy = \int p_2(x,y,R_x=1,R_y=0)dy.
$$

That is,

$$
\begin{aligned}
&\int p_1(y)p(x\mid y)p_1\left(R_x=1\mid y\right)p_1\left(R_y=0\mid x, R_x=1\right)dy \\
&\qquad = \int p_2(y)p(x\mid y)p_2\left(R_x=1\mid y\right)p_2\left(R_y=0\mid x, R_x=1\right)dy.
\end{aligned}
$$

Or in other words:

$$
\begin{aligned}
&\int p_1(y)p(x\mid y)p_1\left(R_x=1\mid y\right)dy - \int p_1(y)p(x\mid y)p_1\left(R_x=1\mid y\right)p_1(R_y=1\mid x,R_x=1)dy \\
&\quad = \int p_2(y)p(x\mid y)p_2\left(R_x=1\mid y\right)dy - \int p_2(y)p(x\mid y)p_2\left(R_x=1\mid y\right)p_2(R_y=1\mid x,R_x=1)dy.
\end{aligned}
$$

Since $\int p_1(y)p(x\mid y)p_1\left(R_x=1\mid y\right)p_1(R_y=1\mid x,R_x=1)dy = \int p_2(y)p(x\mid y)p_2\left(R_x=1\mid y\right)p_2(R_y=1\mid x,R_x=1)dy$ holds by the missingness pattern $(R_x=1,R_y=1)$, we only need to show

$$
\int p_1(y)p(x\mid y)p_1\left(R_x=1\mid y\right)dy = \int p_2(y)p(x\mid y)p_2\left(R_x=1\mid y\right)dy.
$$

We have:

$$
\begin{aligned}
&p_1(y)\, p(x\mid y)\, p_1\left(R_x=1\mid y\right) \\
&\qquad = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}(y-1)^2\right\} \times p(x\mid y) \times \frac{\sqrt{\frac{5}{6}}}{\sqrt{\frac{5}{6}}+\exp\left[-\frac{1}{12}(y-1)^2\right]} \\
&\qquad = \frac{1}{\sqrt{2\pi}\sqrt{\frac{6}{5}}}\exp\left\{-\frac{1}{2\times\frac{6}{5}}(y-1)^2\right\} \times p(x\mid y) \times \frac{\exp\left[-\frac{1}{12}(y-1)^2\right]}{\sqrt{\frac{5}{6}}+\exp\left[-\frac{1}{12}(y-1)^2\right]} \\
&\qquad = p_2(y)\, p(x\mid y)\, p_2\left(R_x=1\mid y\right).
\end{aligned}
$$

3. <u>Missingness pattern $(R_x=0,\ R_y=1)$.</u> We need to prove

$$
\int p_1(x,y,R_x=0,R_y=1)dx = \int p_2(x,y,R_x=0,R_y=1)dx.
$$

For any μ and $\sigma>0$, it is true that

$$
\begin{aligned}
&\int \phi(x-y)\times\phi(\frac{x-\mu}{\sigma})dx \\
&= \int \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}(x-y)^2\right\} \times \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}dx \\
&= \frac{1}{\sqrt{2\pi}}\times\frac{1}{\sqrt{2\pi}\sigma}\int\exp\left\{-\frac{1}{2}x^2+xy-\frac{1}{2}y^2-\frac{1}{2\sigma^2}x^2+\frac{1}{\sigma^2}x\mu-\frac{1}{2\sigma^2}\mu^2\right\}dx \\
&= \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{2\pi}\sigma}\times\int\exp\left\{-\frac{1}{2\times\frac{\sigma^2}{\sigma^2+1}}\left[x^2-2x\left(y+\frac{\mu}{\sigma^2}\right)\frac{\sigma^2}{\sigma^2+1}+\left(y+\frac{\mu}{\sigma^2}\right)^2\left(\frac{\sigma^2}{\sigma^2+1}\right)^2\right]\right\}\exp\left[-\frac{1}{2}y^2-\frac{1}{2\sigma^2}\mu^2+\frac{1}{2\frac{\sigma^2}{\sigma^2+1}}\times\left(y+\frac{\mu}{\sigma^2}\right)^2\left(\frac{\sigma^2}{\sigma^2+1}\right)^2\right]dx \\
&= \frac{1}{\sqrt{2\pi}}\times\sqrt{\frac{1}{1+\sigma^2}}\times\exp\left[-\frac{1}{2}\frac{1}{1+\sigma^2}y^2+\frac{1}{1+\sigma^2}\mu y-\frac{1}{2}\frac{\mu^2}{1+\sigma^2}\right].
\end{aligned}
$$

Thus, we have:

$$
\begin{aligned}
&p_1(y)p_1\left(R_x=0 \mid y\right) \int p(x \mid y) p_1\left(R_y=1 \mid x, R_x=0\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y-1)^2\right\} \times \frac{\exp\left[-\frac{1}{12}(y-1)^2\right]}{\sqrt{\frac{5}{6}+\exp\left[-\frac{1}{12}(y-1)^2\right]}} \times \frac{1}{\sqrt{2\pi}}\sqrt{\frac{1}{6}} \exp\left[-\frac{1}{12}y^2+\frac{5}{6}y-\frac{1}{2}\times\frac{25}{6}\right] \\
&= \frac{1}{2\pi}\sqrt{\frac{1}{6}} \frac{1}{\sqrt{\frac{5}{6}+\exp\left[-\frac{1}{12}(y-1)^2\right]}} \times \exp\left\{-\frac{7}{12}(y-1)^2-\frac{1}{12}y^2+\frac{5}{6}y-\frac{1}{2}\times\frac{25}{6}\right\} \\
&= \frac{1}{2\pi}\sqrt{\frac{1}{6}} \frac{1}{\sqrt{\frac{5}{6}+\exp\left[-\frac{1}{12}(y-1)^2\right]}} \times \exp\left\{-\frac{2}{3}y^2+2y-\frac{8}{3}\right\} \\
&= p_2(y)p_2\left(R_x=0 \mid y\right) \int p(x \mid y) p_2\left(R_y=1 \mid x, R_x=0\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2\times\frac{6}{5}}(y-1)^2\right\} \times \frac{\sqrt{\frac{5}{6}}}{\sqrt{\frac{5}{6}+\exp\left[-\frac{1}{12}(y-1)^2\right]}} \times \exp(-\frac{8}{9})\sqrt{\frac{2}{5}}\frac{1}{\sqrt{2\pi}}\sqrt{\frac{1}{2}} \exp\left[-\frac{1}{4}y^2+\frac{7}{6}y-\frac{49}{36}\right] \\
&= \frac{1}{2\pi}\sqrt{\frac{1}{6}} \exp(-\frac{8}{9}) \frac{1}{\sqrt{\frac{5}{6}+\exp\left[-\frac{1}{12}(y-1)^2\right]}} \exp\left\{-\frac{5}{12}(y-1)^2-\frac{1}{4}y^2+\frac{7}{6}y-\frac{49}{36}\right\} \\
&= \frac{1}{2\pi}\sqrt{\frac{1}{6}} \frac{1}{\sqrt{\frac{5}{6}+\exp\left[-\frac{1}{12}(y-1)^2\right]}} \exp\left\{-\frac{2}{3}y^2+2y-\frac{8}{3}\right\}.
\end{aligned}
$$

4. Missingness pattern ($R_x = 0,\ R_y = 0$). We need to prove

$$\int p_1(x, y, R_x = 0, R_y = 0) dxdy = \int p_2(x, y, R_x = 0, R_y = 0) dxdy,$$

which is guaranteed to hold since the previous three missingness patterns yield the same observed data law and the fact that probabilities should integrate to one.

This concludes the claim that the target law is not identified in the criss-cross MNAR model.

B IDENTIFICATION PROOFS

B.1 THEOREM 1 (TARGET LAW PARAMETRIC IDENTIFICATION: UNIVARIATE X)

We have

$$X \sim \exp\left\{\frac{x\eta_x - b_x(\eta_x)}{\Phi_x} + c_x(x;\ \Phi_x)\right\}$$
$$Y \mid X \sim \exp\left\{\frac{y\eta - b(\eta)}{\Phi} + c(y;\ \Phi)\right\}, \quad g(\mu(\eta)) = \alpha + \beta x.$$

The parameters of interest are $\theta = (\alpha, \beta, \Phi, \eta_x, \Phi_x)$. Since $p(x \mid y)$ is nonparametrically (np)-identified, we can select two distinct points of X, say x_1 and x_0 and write

$$\begin{aligned}
\frac{p(x_1 \mid y)}{p(x_0 \mid y)} &= \frac{p(y \mid x_1)p(x_1)}{p(y)} \div \frac{p(y \mid x_0)p(x_0)}{p(y)} = \frac{p(y \mid x_1)}{p(y \mid x_0)} \times \frac{p(x_1)}{p(x_0)} \\
&= \exp\left\{\frac{y(\eta_1 - \eta_0) - [b(\eta_1) - b(\eta_0)]}{\Phi}\right\} \times \exp\left\{\frac{\eta_x(x_1 - x_0)}{\Phi_x} + c(x_1;\ \Phi_x) - c(x_0;\ \Phi_x)\right\}.
\end{aligned}$$

We take a *log* on both sides. The left-hand side is only a function of y. Suppose the coefficient of y on the left-hand side is ϕ_1 and the intercept is ζ_1. For the ease of notation, define $\varphi = [g \circ \mu]^{-1}$ and $\zeta = b([g \circ \mu]^{-1})$. We can then write the following:

$$\begin{aligned}
\phi_1(\theta) &= \frac{\eta_1 - \eta_0}{\Phi} = \frac{[g \circ \mu]^{-1}(\alpha + x_1\beta) - [g \circ \mu]^{-1}(\alpha + x_0\beta)}{\Phi} = \frac{\varphi(\alpha + x_1\beta) - \varphi(\alpha + x_0\beta)}{\Phi} \\
\zeta_1(\theta) &= \left\{\frac{-[b(\eta_1) - b(\eta_0)]}{\Phi} + \frac{\eta_x(x_1 - x_0)}{\Phi_x} + c(x_1;\ \Phi_x) - c(x_0;\ \Phi_x)\right\} \\
&= \left\{\frac{-\left[b\left([g \circ \mu]^{-1}(\alpha + x_1\beta)\right) - b\left([g \circ \mu]^{-1}(\alpha + x_0\beta)\right)\right]}{\Phi} + \frac{\eta_x(x_1 - x_0)}{\Phi_x} + c(x_1;\ \Phi_x) - c(x_0;\ \Phi_x)\right\} \\
&= \left\{\frac{-\zeta(\alpha + x_1\beta) + \zeta(\alpha + x_0\beta)}{\Phi} + \frac{\eta_x(x_1 - x_0)}{\Phi_x} + c(x_1;\ \Phi_x) - c(x_0;\ \Phi_x)\right\}.
\end{aligned}$$

Suppose we have $k + 1$ distinct values of x. We can then create $2k$ equations like above, say ϕ_i and ζ_i with $i = 1, \ldots, k$. The core of our identification proof relies on the *implicit function theorem*. In order to use this theorem, the above equations need to satisfy the followings:

1. There exists at least one solution θ_0 that satisfies the above equations,

2. $\phi_i(\theta)$ and $\zeta_i(\theta)$ are continuous in Θ, i.e., the parameter space with θ_0 as an inner point,

3. $\phi_i(\theta)$ and $\zeta_i(\theta)$ are first order partially differentiable in Θ,

4. Let $\Phi = \{\phi_1, \ldots, \phi_k\}$ and $Z = \{\zeta_1, \ldots, \zeta_k\}$. Define the Jacobian matrix J as $J = \frac{\partial(\Phi, Z)}{\partial(\theta)}$, which is described below:

$$\text{J=}\begin{bmatrix}
\varphi'(\alpha + x_1\beta) - \varphi'(\alpha + x_0\beta) & \varphi'(\alpha + x_1\beta)\,x_1 - \varphi'(\alpha + x_0\beta)\,x_0 & \varphi(\alpha + x_1\beta) - \varphi(\alpha + x_0\beta) & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\varphi'(\alpha + x_k\beta) - \varphi'(\alpha + x_0\beta) & \varphi'(\alpha + x_k\beta)\,x_k - \varphi'(\alpha + x_0\beta)\,x_0 & \varphi(\alpha + x_k\beta) - \varphi(\alpha + x_0\beta) & 0 & 0 \\
\zeta'(\alpha + x_1\beta) - \zeta'(\alpha + x_0\beta) & \zeta'(\alpha + x_1\beta)\,x_1 - \zeta'(\alpha + x_0\beta)\,x_0 & \zeta(\alpha + x_1\beta) - \zeta(\alpha + x_0\beta) & x_1 - x_0 & -\frac{\eta_x(x_1 - x_0)}{\Phi_x^2} + \frac{\partial c(x_1, \Phi_x)}{\partial \Phi_x} - \frac{\partial c(x_0, \Phi_x)}{\partial \Phi_x} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\zeta'(\alpha + x_k\beta) - \zeta'(\alpha + x_0\beta) & \zeta'(\alpha + x_k\beta)\,x_k - \zeta'(\alpha + x_0\beta)\,x_0 & \zeta(\alpha + x_k\beta) - \zeta(\alpha + x_0\beta) & x_k - x_0 & -\frac{\eta_x(x_k - x_0)}{\Phi_x^2} + \frac{\partial c(x_k, \Phi_x)}{\partial \Phi_x} - \frac{\partial c(x_0, \Phi_x)}{\partial \Phi_x}
\end{bmatrix}$$

   J must be of full rank under $(\theta_0, \phi_i(\theta_0), \zeta_i(\theta_0))$,

5. The number of equations must be greater or equal to the number of unknown parameters, i.e., $2k \geq dim(\theta)$.

Under the above conditions, there exists neighborhood U around the true parameters θ_0 as $U = B(\theta_0, \epsilon) \subset \Theta$, and the neighborhood V around $(\phi_i(\theta_0), \zeta_i(\theta_0))$ as $V = B\left((\phi_1(\theta_0), \ldots, \phi_k(\theta_0), \zeta_1(\theta_0), \ldots, \zeta_k(\theta_0)), \eta\right) \subset R^{2k}$ with $\epsilon, \eta > 0$, and uniquely defined functions $g = (g_1, \ldots, g_{2k})$ on V that each g_i is first-order continuously differentiable. We have

$$\theta = g\left(\phi_1(\theta), \ldots, \phi_k(\theta), \zeta_1(\theta), \ldots, \zeta_k(\theta)\right),$$

where $(\phi_1(\theta), \ldots, \phi_k(\theta), \zeta_1(\theta), \ldots, \zeta_k(\theta)) \in V$, with $\theta \in U$. Given that the $(\phi_1, \ldots, \phi_k, \zeta_1, \ldots, \zeta_k)$ we observed is generated under the true value θ_0, which is observed $(\phi_1, \ldots, \phi_k, \zeta_1, \ldots, \zeta_k) = (\phi_1(\theta_0), \ldots, \phi_k(\theta_0), \zeta_1(\theta_0), \ldots, \zeta_k(\theta_0))$, by applying g, we can uniquely find $\theta_0 = g\left(\phi_1(\theta_0), \ldots, \phi_k(\theta_0), \zeta_1(\theta_0), \ldots, \zeta_k(\theta_0)\right)$.

B.2 TARGET LAW PARAMETRIC IDENTIFICATION: MULTIVARIATE X

B.2.1 Multivariate normal X

Suppose

$$\begin{aligned} &X \sim \mathbb{N}_d(\mu, \Sigma) \\ &Y \mid X \sim \exp\left\{\frac{y\eta - b(\eta)}{\Phi} + c(y;\, \Phi)\right\}, \quad g(\mu(\eta)) = \alpha + x^T\beta. \end{aligned}$$

Assume the nuisance parameter Σ is known and $\theta = (\alpha, \beta, \Phi, \mu)$. We can write down the following equation:

$$\begin{aligned} \frac{p(x_1 \mid y)}{p(x_0 \mid y)} &= \frac{p(y \mid x_1)}{p(y \mid x_0)} \times \frac{p(x_1)}{p(x_0)} \\ &= \exp\left\{\frac{y(\eta_1 - \eta_0) - [b(\eta_1) - b(\eta_0)]}{\Phi}\right\} \exp\left\{-\frac{1}{2}(x_1 - \mu)^T \Sigma^{-1}(x_1 - \mu) + \frac{1}{2}(x_0 - \mu)^T \Sigma^{-1}(x_0 - \mu)\right\}. \end{aligned}$$

Taking a log on both sides yields the following equation:

$$\log[p(x_1 \mid y)] - \log[p(x_0 \mid y)] = y \times \frac{\eta_1 - \eta_0}{\Phi} - \frac{b(\eta_1) - b(\eta_0)}{\Phi} - \frac{1}{2}(x_1 - \mu)^T \Sigma^{-1}(x_1 - \mu) + \frac{1}{2}(x_0 - \mu)^T \Sigma^{-1}(x_0 - \mu).$$

The left-hand side is only a function of y. Suppose the coefficient of y is ϕ_1 and the intercept is ζ_1. For the ease of notation, define $\varphi = [g \circ \mu]^{-1}$ and $\zeta = b([g \circ \mu]^{-1})$. Then, we obtain the following equation:

$$\begin{aligned} \phi_1(\theta) &= \frac{\eta_1 - \eta_0}{\Phi} = \frac{[g \circ \mu]^{-1}(\alpha + x_1^T\beta) - [g \circ \mu]^{-1}(\alpha + x_0^T\beta)}{\Phi} = \frac{\varphi(\alpha + x_1^T\beta) - \varphi(\alpha + x_0^T\beta)}{\Phi} \\ \zeta_1(\theta) &= -\frac{b(\eta_1) - b(\eta_0)}{\Phi} - \frac{1}{2}(x_1 - \mu)^T \Sigma^{-1}(x_1 - \mu) + \frac{1}{2}(x_0 - \mu)^T \Sigma^{-1}(x_0 - \mu) \\ &= -\frac{\zeta(\alpha + x_1^T\beta) - \zeta(\alpha + x_0^T\beta)}{\Phi} - \frac{1}{2}(x_1 - \mu)^T \Sigma^{-1}(x_1 - \mu) + \frac{1}{2}(x_0 - \mu)^T \Sigma^{-1}(x_0 - \mu). \end{aligned}$$

Suppose we have $k+1$ distinct values of x. Thus, we can construct $2k$ equations, ϕ_i and ζ_i with $i = 1, \ldots, k$. In order to use this theorem, the above equations need to satisfy the followings:

1. There exists at least one solution θ_0 that satisfies the above equations,
2. $\phi_i(\theta)$ and $\zeta_i(\theta)$ are continuous on Θ, i.e., the parameter space with θ_0 as an inner point,
3. $\phi_i(\theta)$ and $\zeta_i(\theta)$ are first order partially differentiable on Θ,
4. Let $\Phi = \{\phi_1, \ldots, \phi_k\}$ and $Z = \{\zeta_1, \ldots, \zeta_k\}$. Define then Jacobian matrix J as $J = \frac{\partial(\Phi, Z)}{\partial(\theta)}$, described below:

$$\text{J=}\begin{bmatrix} \varphi'(\alpha + x_1^T\beta) - \varphi'(\alpha + x_0^T\beta) & \varphi'(\alpha + x_1^T\beta)x_1^T - \varphi'(\alpha + x_0^T\beta)x_0^T & \varphi(\alpha + x_1^T\beta) - \varphi(\alpha + x_0^T\beta) & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \varphi'(\alpha + x_k^T\beta) - \varphi'(\alpha + x_0^T\beta) & \varphi'(\alpha + x_k^T\beta)x_k^T - \varphi'(\alpha + x_0^T\beta)x_0^T & \varphi(\alpha + x_k^T\beta) - \varphi(\alpha + x_0^T\beta) & 0 \\ \zeta'(\alpha + x_1^T\beta) - \zeta'(\alpha + x_0^T\beta) & \zeta'(\alpha + x_1^T\beta)x_1^T - \zeta'(\alpha + x_0^T\beta)x_0^T & \zeta(\alpha + x_1^T\beta) - \zeta(\alpha + x_0^T\beta) & (x_1 - x_0)^T \Sigma^{-1} \\ \vdots & \vdots & \vdots & \vdots \\ \zeta'(\alpha + x_k^T\beta) - \zeta'(\alpha + x_0^T\beta) & \zeta'(\alpha + x_k^T\beta)x_k^T - \zeta'(\alpha + x_0^T\beta)x_0^T & \zeta(\alpha + x_k^T\beta) - \zeta(\alpha + x_0^T\beta) & (x_k - x_0)^T \Sigma^{-1} \end{bmatrix}$$

J must be of full rank under $(\theta_0, \phi_i(\theta_0), \zeta_i(\theta_0))$,

5. The number of equations must be greater or equal to the number of unknown parameters, i.e., $2k \geq dim(\theta)$.

Under the special case where $Y \mid X \sim \mathbb{N}\left(\alpha + x^T\beta, \Phi\right)$, we have:

$$\begin{aligned}
\phi_i(\theta) &= \frac{\left(x_i - x_0\right)^T \beta}{\Phi} \\
\zeta_i(\theta) &= -\frac{\left(\alpha + x_i^T\beta\right)^2 - \left(\alpha + x_0^T\beta\right)^2}{2\Phi} - \frac{1}{2}\left(x_i - \mu\right)^T \Sigma^{-1}\left(x_i - \mu\right) + \frac{1}{2}\left(x_0 - \mu\right)^T \Sigma^{-1}\left(x_0 - \mu\right),
\end{aligned}$$

where $i \in (1, \ldots, k),$ and

$$\mathbf{J}=\begin{bmatrix}
0 & \frac{(x_1-x_0)^T}{\Phi} & -\frac{(x_1-x_0)^T\beta}{\Phi^2} & 0 \\
\vdots & \vdots & \vdots & \vdots \\
0 & \frac{(x_k-x_0)^T}{\Phi} & -\frac{(x_k-x_0)^T\beta}{\Phi^2} & 0 \\
-\frac{(x_1-x_0)^T\beta}{\Phi} & -\frac{\alpha(x_1-x_0)^T+\beta^T(x_1x_1^T-x_0x_0^T)}{\Phi} & \frac{(\alpha+x_1^T\beta)^2-(\alpha+x_0^T\beta)^2}{2\Phi^2} & \left(x_1 - x_0\right)^T \Sigma^{-1} \\
\vdots & \vdots & \vdots & \vdots \\
-\frac{(x_k-x_0)^T\beta}{\Phi} & -\frac{\alpha(x_k-x_0)^T+\beta^T(x_kx_k^T-x_0x_0^T)}{\Phi} & \frac{(\alpha+x_k^T\beta)^2-(\alpha+x_0^T\beta)^2}{2\Phi^2} & \left(x_k - x_0\right)^T \Sigma^{-1}
\end{bmatrix}$$

After performing some rank-preserving modifications to this matrix, we have

$$\mathbf{J}=\begin{bmatrix}
0 & (x_1 - x_0)^T & -(x_1 - x_0)^T\beta & 0 \\
\vdots & \vdots & \vdots & \vdots \\
0 & (x_1 - x_0)^T & -(x_1 - x_0)^T\beta & 0 \\
(x_1 - x_0)^T\beta & -\left[\alpha(x_1 - x_0)^T + \beta^T(x_1x_1^T - x_0x_0^T)\right] & \frac{(\alpha+x_1^T\beta)^2-(\alpha+x_0^T\beta)^2}{2} & \left(x_1 - x_0\right)^T \Sigma^{-1} \\
\vdots & \vdots & \vdots & \vdots \\
(x_k - x_0)^T\beta & -\left[\alpha(x_k - x_0)^T + \beta^T(x_kx_1^T - x_0x_0^T)\right] & \frac{(\alpha+x_k^T\beta)^2-(\alpha+x_0^T\beta)^2}{2} & \left(x_k - x_0\right)^T \Sigma^{-1}
\end{bmatrix}$$

The dimension of J is $dim(J) = 2k \times (2 + 2d)$. Assume $2k \geq (2 + 2d)$. A sufficient condition to make J full rank is knowing at least α.

Note that in this example $p(X \mid Y)$ **is in the exponential family**, since:

$$\begin{aligned}
p(x \mid y) &= \frac{p(y \mid x)p(x)}{p(y)} \\
&= \exp\left\{-\frac{\left[y - \left(\alpha + x^T\beta\right)\right]^2}{2\Phi} + \log\frac{1}{\sqrt{2\pi\Phi}} - \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) + \log\frac{1}{\sqrt{(2\pi)^d|\Sigma|}} - \log(y)\right\} \\
&= \exp\left\{ -\frac{(y-\alpha)^2}{2\Phi} + \frac{(y\beta - \alpha\beta)^T}{\Phi}x - \frac{\operatorname{tr}\left(\beta\beta^T xx^T\right)}{2\Phi} + \log\frac{1}{\sqrt{2\pi\Phi}} + \mu^T\Sigma^{-1}x - \frac{1}{2}x^T\Sigma^{-1}x - \frac{1}{2}\mu^T\Sigma^{-1}\mu + \log\frac{1}{\sqrt{(2\pi)^d|\Sigma|}} - \log(y)\right\} \\
&= \exp\left\{\left[\frac{(y\beta - \alpha\beta)^T}{\Phi} + \mu^T\Sigma^{-1}, -\frac{\operatorname{vec}\left(\beta\beta^T\right)^T}{2\Phi}\right]\begin{pmatrix} x \\ \operatorname{vec}\left(xx^T\right)\end{pmatrix} - \frac{(y-\alpha)^2}{2\Phi} + \log\frac{1}{\sqrt{2\pi\Phi}} - \frac{1}{2}x^T\Sigma^{-1}x - \frac{1}{2}\mu^T\Sigma^{-1}\mu + \log\frac{1}{\sqrt{(2\pi)^d|\Sigma|}} - \log(y)\right\}.
\end{aligned}$$

Here *tr*(.) denotes the trace of the input matrix and *vec*(.) refers to the vectorization operation applied to the input matrix, e.g., $A_{n\times m}$, as stacking the rows of the matrix one by one to form a long column vector with size nm, i.e.,

$$\operatorname{vec}[A] = \operatorname{vec}\left[\begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm}\end{pmatrix}\right] = \begin{pmatrix} a_{11} \\ \vdots \\ a_{1m} \\ \vdots \\ a_{nm}\end{pmatrix}.$$

B.2.2 Multinomial X

Suppose

$$\begin{aligned}
&X \sim \text{Multinomial}_d(n, p),\\
&Y \mid X \sim \exp\left\{\frac{y\eta - b(\eta)}{\Phi} + c(y;\ \Phi)\right\}, \quad g(\mu(\eta)) = \alpha + x^T\beta,
\end{aligned}$$

where $p = (p_1, \ldots, p_d)$ is the vector of event probabilities, and n is the number of trials. We can write $p(x) = \exp[x^T\eta + c(x)]$ with $\eta = \left(\log p_1, \ldots, \log_{p_d}\right), c(x) = \log \frac{n!}{x_1! \cdots x_d!}$. Assume the nuisance parameter n is known and $\theta = (\alpha, \beta, \Phi, \eta)$. We can write down the following:

$$\begin{aligned}
\frac{p\left(x_1 \mid y\right)}{p\left(x_0 \mid y\right)} &= \frac{p\left(y \mid x_1\right)}{p\left(y \mid x_0\right)} \times \frac{p\left(x_1\right)}{p\left(x_0\right)}\\
&= \exp\left\{\frac{y\left(\eta_1 - \eta_0\right) - \left[b\left(\eta_1\right) - b\left(\eta_0\right)\right]}{\Phi}\right\} \exp\left\{\left(x_1 - x_0\right)^T\eta + c\left(x_1\right) - c\left(x_0\right)\right\}.
\end{aligned}$$

Taking a log on both sides yields the following:

$$\log\left[p\left(x_1 \mid y\right)\right] - \log\left[p\left(x_0 \mid y\right)\right] = y\frac{\eta_1 - \eta_0}{\Phi} - \frac{b\left(\eta_1\right) - b\left(\eta_0\right)}{\Phi} + \left(x_1 - x_0\right)^T\eta + c\left(x_1\right) - c\left(x_0\right)$$

The left-hand side is only a function of y. Suppose the coefficient of y is ϕ_1 and the intercept is ζ_1. For the ease of notation, define $\varphi = [g \circ \mu]^{-1}$ and $\zeta = b([g \circ \mu]^{-1})$. Thus, we obtain the following:

$$\begin{aligned}
\phi_1(\theta) &= \frac{\eta_1 - \eta_0}{\Phi} = \frac{[g \circ \mu]^{-1}\left(\alpha + x_1^T\beta\right) - [g \circ \mu]^{-1}\left(\alpha + x_0^T\beta\right)}{\Phi} = \frac{\varphi\left(\alpha + x_1^T\beta\right) - \varphi\left(\alpha + x_0^T\beta\right)}{\Phi}\\
\zeta_1(\theta) &= -\frac{b\left(\eta_1\right) - b\left(\eta_0\right)}{\Phi} + \left(x_1 - x_0\right)^T\eta + c\left(x_1\right) - c\left(x_0\right)\\
&= -\frac{\zeta(\alpha + x_1^T\beta) - \zeta(\alpha + x_0^T\beta)}{\Phi} + \left(x_1 - x_0\right)^T\eta + c\left(x_1\right) - c\left(x_0\right).
\end{aligned}$$

Suppose we have $k+1$ distinct values of x. Thus, we can construct $2k$ equations, ϕ_i and ζ_i with $i = 1, \ldots, k$. To apply the implicit function theorem, the equations need to satisfy the following conditions:

1. There exists at least one solution θ_0 that satisfies the above equations,

2. $\phi_i(\theta)$ and $\zeta_i(\theta)$ are continuous on Θ, i.e., the parameter space with θ_0 as an inner point,

3. $\phi_i(\theta)$ and $\zeta_i(\theta)$ are first order partially differentiable on Θ,

4. Let $\Phi = \{\phi_1, \ldots, \phi_k\}$ and $Z = \{\zeta_1, \ldots, \zeta_k\}$. Define then Jacobian matrix J as $J = \frac{\partial(\Phi, Z)}{\partial(\theta)}$, described below:

$$\text{J=}\begin{bmatrix}
\varphi'\left(\alpha + x_1^T\beta\right) - \varphi'\left(\alpha + x_0^T\beta\right) & \varphi'\left(\alpha + x_1^T\beta\right)x_1^T - \varphi'\left(\alpha + x_0^T\beta\right)x_0^T & \varphi\left(\alpha + x_1^T\beta\right) - \varphi\left(\alpha + x_0^T\beta\right) & 0\\
\vdots & \vdots & \vdots & \vdots\\
\varphi'\left(\alpha + x_k^T\beta\right) - \varphi'\left(\alpha + x_0^T\beta\right) & \varphi'\left(\alpha + x_k^T\beta\right)x_k^T - \varphi'\left(\alpha + x_0^T\beta\right)x_0^T & \varphi\left(\alpha + x_k^T\beta\right) - \varphi\left(\alpha + x_0^T\beta\right) & 0\\
\zeta'\left(\alpha + x_1^T\beta\right) - \zeta'\left(\alpha + x_0^T\beta\right) & \zeta'\left(\alpha + x_1^T\beta\right)x_1^T - \zeta'\left(\alpha + x_0^T\beta\right)x_0^T & \zeta\left(\alpha + x_1^T\beta\right) - \zeta\left(\alpha + x_0^T\beta\right) & \left(x_1 - x_0\right)^T M\\
\vdots & \vdots & \vdots & \vdots\\
\zeta'\left(\alpha + x_k^T\beta\right) - \zeta'\left(\alpha + x_0^T\beta\right) & \zeta'\left(\alpha + x_k^T\beta\right)x_k^T - \zeta'\left(\alpha + x_0^T\beta\right)x_0^T & \zeta\left(\alpha + x_k^T\beta\right) - \zeta\left(\alpha + x_0^T\beta\right) & \left(x_k - x_0\right)^T M
\end{bmatrix}$$

   where $M_{d \times d-1} = \begin{bmatrix} \mathbb{I}_{d-1 \times d-1} \\ (-1, -1, \cdots, -1)_{1 \times d-1} \end{bmatrix}$, $\mathbb{I}$ is the identity matrix.

   The Jacobian matrix J must be of full rank under $(\theta_0, \phi_i(\theta_0), \zeta_i(\theta_0))$.

5. The number of equations must be greater or equal to the number of unknown parameters, i.e., $2k \geq dim(\theta)$.

Under the special case where $Y \mid X \sim \mathbb{N}\left(\alpha + x^T\beta, \Phi\right)$, we have:

$$\begin{aligned}
\phi_i(\theta) &= \frac{\left(x_i - x_0\right)^T \beta}{\Phi} \\
\zeta_i(\theta) &= -\frac{\left(\alpha + x_i^T\beta\right)^2 - \left(\alpha + x_0^T\beta\right)^2}{2\Phi} + \left(x_i - x_0\right)^T \eta + c\left(x_i\right) - c\left(x_0\right), \quad i \in (1, 2, \cdots, k),
\end{aligned}$$

$$\text{J=}\begin{bmatrix}
0 & \frac{(x_1-x_0)^T}{\Phi} & -\frac{(x_1-x_0)^T\beta}{\Phi^2} & 0 \\
\vdots & \vdots & \vdots & \vdots \\
0 & \frac{(x_k-x_0)^T}{\Phi} & -\frac{(x_k-x_0)^T\beta}{\Phi^2} & 0 \\
-\frac{(x_1-x_0)^T\beta}{\Phi} & -\frac{\alpha(x_1-x_0)^T+\beta^T(x_1x_1^T-x_0x_0^T)}{\Phi} & \frac{(\alpha+x_1^T\beta)^2-(\alpha+x_0^T\beta)^2}{2\Phi^2} & \left(x_1 - x_0\right)^T M \\
\vdots & \vdots & \vdots & \vdots \\
-\frac{(x_k-x_0)^T\beta}{\Phi} & -\frac{\alpha(x_k-x_0)^T+\beta^T(x_kx_k^T-x_0x_0^T)}{\Phi} & \frac{(\alpha+x_k^T\beta)^2-(\alpha+x_0^T\beta)^2}{2\Phi^2} & \left(x_k - x_0\right)^T M
\end{bmatrix}$$

After performing some rank-preserving modifications to this matrix, we get:

$$\text{J=}\begin{bmatrix}
0 & (x_1 - x_0)^T & -(x_1 - x_0)^T\beta & 0 \\
\vdots & \vdots & \vdots & \vdots \\
0 & (x_k - x_0)^T & -(x_k - x_0)^T\beta & 0 \\
(x_1 - x_0)^T\beta & -\left[\alpha(x_1 - x_0)^T + \beta^T(x_1x_1^T - x_0x_0^T)\right] & \frac{(\alpha+x_1^T\beta)^2-(\alpha+x_0^T\beta)^2}{2} & \left(x_1 - x_0\right)^T M \\
\vdots & \vdots & \vdots & \vdots \\
(x_k - x_0)^T\beta & -\left[\alpha(x_k - x_0)^T + \beta^T(x_kx_k^T - x_0x_0^T)\right] & \frac{(\alpha+x_k^T\beta)^2-(\alpha+x_0^T\beta)^2}{2} & \left(x_k - x_0\right)^T M
\end{bmatrix}$$

The dimension of J is $dim(J) = 2k \times (1 + 2d)$. Assume $2k \geq (1 + 2d)$. A sufficient condition to make J full rank is knowing α or at least one element of η.

Note that in this example, $p(X \mid Y)$ **is in the exponential family**, since:

$$\begin{aligned}
p(x \mid y) &= \frac{p(y \mid x)p(x)}{p(y)} \\
&= \exp\left\{-\frac{\left[y - \left(\alpha + x^T\beta\right)\right]^2}{2\Phi} + \log\frac{1}{\sqrt{2\pi\Phi}} + x^T\eta + c(x) - \log p(y)\right\} \\
&= \exp\left\{\left[\frac{(y\beta - \alpha\beta)^T}{\Phi} + \eta^T, -\frac{\operatorname{vec}\left(\beta\beta^T\right)^T}{2\Phi}\right]\begin{pmatrix} x \\ \operatorname{vec}\left(xx^T\right)\end{pmatrix} - \frac{(y - \alpha)^2}{2\Phi} + c(x) - \log p(y)\right\}.
\end{aligned}$$

B.3 LEMMA 1 (FULL LAW IDENTIFICATION)

Using the DAG factorization we have

$$p\left(X, Y, R_x = 1, R_y = 1\right) = p(X, Y)\, p(R_x = 1 \mid Y)\, p(R_y = 1 \mid X, R_x = 1).$$

Given the above relation and the fact that the target law $p(X, Y)$ is identified, it is straightforward to conclude that $p(R_x \mid Y)$ is also identified. We now prove under the completeness condition, $p(R_y \mid X, R_x)$ is also identified. Therefore the full law is identified. The full observed data law can be written down as follows:

$$\mathcal{L}_{\text{full}}(Z_{\text{obs}}, R; \theta, \psi) = \prod_{R_x=1, R_y=1} p(X, Y, R_x = 1, R_y = 1) \times \prod_{R_x=1, R_y=0} \int p(X, Y, R_x = 1, R_y = 0) dy$$
$$\times \prod_{R_x=0, R_y=1} \int p(X, Y, R_x = 0, R_y = 1) dx \times \prod_{R_x=0, R_y=0} \int p(X, Y, R_x = 0, R_y = 0) dx dy.$$

Given the fact that $p(X, Y)$, $p(R_x = 1 \mid Y)$, and $p(R_x = 0, R_y = 0)$ are all identified, the following would stay the same across different models:

$$\prod_{R_x=1, R_y=1} p(X, Y, R_x = 1, R_y = 1) \times \prod_{R_x=1, R_y=0} \int p(X, Y, R_x = 1, R_y = 0) dy \times \prod_{R_x=0, R_y=0} \int p(X, Y, R_x = 0, R_y = 0) dx dy.$$

Suppose there exist $p_1(Ry \mid X, R_x)$ and $p_2(Ry \mid X, R_x)$ such that

$$\int p(X, Y) p(R_x = 0 \mid Y) p_1(R_y = 1 \mid R_x = 0, X) dx = \int p(X, Y) p(R_x = 0 \mid Y) p_2(R_y = 1 \mid R_x = 0, X) dx$$

Let $g(X) = p_1(R_y = 1 \mid R_x = 0, X) - p_2(R_y = 1 \mid R_x = 0, X)$, we have

$$p(R_x = 0 \mid Y = y)\, p(Y = y) \int p(x \mid Y = y)\, g(x)\, dx = 0,\ \forall y$$

This must mean that $E[g(X) \mid y] = 0,\ \forall y$. In our case, $g(X)$ is bounded, thus is with finite mean. Based on the completeness condition, $g(X) = 0$ almost surely, which implies $p_1(R_y \mid X, R_x) = p_2(R_y \mid X, R_x)$ almost surely. This concludes that the full law is indeed identified.

C EXAMPLES FROM THE EXPONENTIAL FAMILY DISTRIBUTIONS

In order to better illustrate the implications of Theorem 1, we provide explicit sufficient identification conditions in a variety of examples in the class of exponential family distributions. In all subsequent examples, we assume that if X is continuous, a sufficient number of unique X values have been observed such that the first condition in Theorem 1, namely that $k \geq dim(\theta)$, is satisfied. If X is discrete, it is assumed that every category of X is observed in the sample.

C.1 X AND Y ARE BIVARIATE NORMAL

Suppose

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim \mathbb{N}\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right].$$

According to Theorem 1, $p(X, Y)$ is identifiable if at least μ_1 or μ_2 is known, in addition to knowing at least one more parameter in $\{\sigma_1, \sigma_2, \rho\}$. As special cases, when either the marginal distribution of X or Y is known, we can identify $p(X, Y)$.

The above claim can be proven as follows. First, we note that $p(X \mid Y)$ also follows a normal distribution:

$$X \mid Y \sim \mathbb{N}\left[\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(y - \mu_1), \left(1 - \rho^2\right)\sigma_2^2\right].$$

Since $p(X \mid Y)$ is nonparametrically identified, it means the mean and variance are both identifiable, i.e., $\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(y - \mu_1)$ and $\left(1 - \rho^2\right)\sigma_2^2$. Thus the following three parameters are identified:

$$\mu_2 - \rho\frac{\sigma_2}{\sigma_1}\mu_1, \quad \rho\frac{\sigma_2}{\sigma_1}, \quad \left(1 - \rho^2\right)\sigma_2^2$$

Let $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. By taking derivative with respect to θ, we obtain the following Jacobian matrix:

$$J = \begin{bmatrix} -\rho\frac{\sigma_2}{\sigma_1} & 1 & \rho\frac{\sigma_2}{\sigma_1^2}\mu_1 & -\rho\frac{1}{\sigma_1}\mu_1 & -\frac{\sigma_2}{\sigma_1}\mu_1 \\ 0 & 0 & -\rho\frac{\sigma_2}{\sigma_1^2}\mu_1 & \rho\frac{1}{\sigma_1} & \frac{\sigma_2}{\sigma_1} \\ 0 & 0 & 0 & 2\left(1 - \rho^2\right)\sigma_2 & -2\rho\sigma_2^2 \end{bmatrix}$$

The number of unknown parameters is greater than the number of equations. To establish target law identification, we need to assume two of the five parameters are known. However, not every pair of parameters will be useful in establishing identification. We go over different options one by one: ($|J|$ denotes the determinant of matrix J.)

1. Assume μ_1, μ_2 are known, then $|J| \neq 0 \Longrightarrow$ target law is identified

$$J = \begin{bmatrix} \rho\frac{\sigma_2}{\sigma_1^2}\mu_1 & -\rho\frac{1}{\sigma_1}\mu_1 & -\frac{\sigma_2}{\sigma_1}\mu_1 \\ -\rho\frac{\sigma_2}{\sigma_1^2}\mu_1 & \rho\frac{1}{\sigma_1} & \frac{\sigma_2}{\sigma_1} \\ 0 & 2\left(1 - \rho^2\right)\sigma_2 & -2\rho\sigma_2^2 \end{bmatrix}$$

2. Assume μ_1, σ_1 are known, then $|J| \neq 0 \Longrightarrow$ target law is identified

$$J = \begin{bmatrix} 1 & -\rho\frac{1}{\sigma_1}\mu_1 & -\frac{\sigma_2}{\sigma_1}\mu_1 \\ 0 & \rho\frac{1}{\sigma_1} & \frac{\sigma_2}{\sigma_1} \\ 0 & 2\left(1 - \rho^2\right)\sigma_2 & -2\rho\sigma_2^2 \end{bmatrix}$$

3. Assume μ_1, σ_2 are known, then $|J| \neq 0 \Longrightarrow$ target law is identified

$$J = \begin{bmatrix} 1 & \rho\frac{\sigma_2}{\sigma_1^2}\mu_1 & -\frac{\sigma_2}{\sigma_1}\mu_1 \\ 0 & -\rho\frac{\sigma_2}{\sigma_1^2}\mu_1 & \frac{\sigma_2}{\sigma_1} \\ 0 & 0 & -2\rho\sigma_2^2 \end{bmatrix}$$

4. Assume μ_1, ρ are known, then $|J| \neq 0 \implies$ target law is identified

$$J = \begin{bmatrix} 1 & \rho\frac{\sigma_2}{\sigma_1^2}\mu_1 & -\rho\frac{1}{\sigma_1}\mu_1 \\ 0 & -\rho\frac{\sigma_2}{\sigma_1^2}\mu_1 & \rho\frac{1}{\sigma_1} \\ 0 & 0 & 2\left(1-\rho^2\right)\sigma_2 \end{bmatrix}$$

5. Assume μ_2, σ_1 are known, then $|J| \neq 0 \implies$ target law is identified

$$J = \begin{bmatrix} -\rho\frac{\sigma_2}{\sigma_1} & -\rho\frac{1}{\sigma_1}\mu_1 & -\frac{\sigma_2}{\sigma_1}\mu_1 \\ 0 & \rho\frac{1}{\sigma_1} & \frac{\sigma_2}{\sigma_1} \\ 0 & 2\left(1-\rho^2\right)\sigma_2 & -2\rho\sigma_2^2 \end{bmatrix}$$

6. Assume μ_2, σ_2 are known, then $|J| \neq 0 \implies$ target law is identified

$$J = \begin{bmatrix} -\rho\frac{\sigma_2}{\sigma_1} & \rho\frac{\sigma_2}{\sigma_1^2}\mu_1 & -\frac{\sigma_2}{\sigma_1}\mu_1 \\ 0 & -\rho\frac{\sigma_2}{\sigma_1^2}\mu_1 & \frac{\sigma_2}{\sigma_1} \\ 0 & 0 & -2\rho\sigma_2^2 \end{bmatrix}$$

   This recovers the case studied in Zhao and Shao [2015].

7. Assume μ_2, ρ are known, then $|J| \neq 0 \implies$ target law is identified

$$J = \begin{bmatrix} -\rho\frac{\sigma_2}{\sigma_1} & \rho\frac{\sigma_2}{\sigma_1^2}\mu_1 & -\rho\frac{1}{\sigma_1}\mu_1 \\ 0 & -\rho\frac{\sigma_2}{\sigma_1^2}\mu_1 & \rho\frac{1}{\sigma_1} \\ 0 & 0 & 2\left(1-\rho^2\right)\sigma_2 \end{bmatrix}$$

8. Assume σ_1, σ_2 are known, then $|J| = 0 \implies$ target law is **not** identified

$$J = \begin{bmatrix} -\rho\frac{\sigma_2}{\sigma_1} & 1 & -\frac{\sigma_2}{\sigma_1}\mu_1 \\ 0 & 0 & \frac{\sigma_2}{\sigma_1} \\ 0 & 0 & -2\rho\sigma_2^2 \end{bmatrix}$$

9. Assume σ_1, ρ are known, then $|J| = 0 \implies$ target law **is not** identified

$$J = \begin{bmatrix} -\rho\frac{\sigma_2}{\sigma_1} & 1 & -\rho\frac{1}{\sigma_1}\mu_1 \\ 0 & 0 & \rho\frac{1}{\sigma_1} \\ 0 & 0 & 2\left(1-\rho^2\right)\sigma_2 \end{bmatrix}$$

10. Assume σ_2, ρ are known, then $|J| = 0 \implies$ target law **is not** identified

$$J = \begin{bmatrix} -\rho\frac{\sigma_2}{\sigma_1} & 1 & \rho\frac{\sigma_2}{\sigma_1^2}\mu_1 \\ 0 & 0 & -\rho\frac{\sigma_2}{\sigma_1^2}\mu_1 \\ 0 & 0 & 0 \end{bmatrix}$$

This concludes that under the bivariate normal distribution, the target law is identified if either μ_1 or μ_2 is known, in addition to knowing at least one more parameter in $\{\sigma_1, \sigma_2, \rho\}$.

It is straightforward to show that $p(X \mid Y)$ **lies in the exponential family**.

C.2 X AND $Y \mid X$ ARE NORMAL UNDER INVERSE LINK

Suppose

$$X \sim \mathbb{N}\left(\mu, \phi_x\right), \qquad Y \mid X \sim \mathbb{N}\left((\alpha + \beta x)^{-1}, \phi\right).$$

According to Theorem 1, $p(X, Y)$ is identifiable without any additional assumptions on the unknown parameter vector $\theta = (\alpha, \beta, \phi, \mu, \phi_x)$. This can be proven as follows: based on Theorem 1, we have the following equations,

$$\begin{aligned}
\phi_i(\theta) &= \frac{\left(\alpha + \beta x_i\right)^{-1} - \left(\alpha + \beta x_0\right)^{-1}}{\phi} \\
\zeta_i(\theta) &= \left\{ -\frac{b\left[\left(\alpha + \beta x_i\right)^{-1}\right] - b\left[\left(\alpha + \beta x_0\right)^{-1}\right]}{\phi} + \frac{\mu\left(x_i - x_0\right)}{\phi_x} + c\left(x_i, \phi_x\right) - c\left(x_0, \phi_x\right) \right\} \\
&= -\frac{\left(\alpha + \beta x_i\right)^{-2} - \left(\alpha + \beta x_0\right)^{-2}}{2\phi} + \frac{\mu\left(x_i - x_0\right)}{\phi_x} - \frac{x_i^2 - x_0^2}{2\phi_x}, \quad \text{where } i \in (1, \ldots, k).
\end{aligned}$$

The Jacobian matrix is as follows:

$$\begin{bmatrix}
-\frac{(\alpha+\beta x_1)^{-2}-(\alpha+\beta x_0)^{-2}}{\phi} & -\frac{(\alpha+\beta x_1)^{-2}x_1-(\alpha+\beta x_0)^{-2}x_0}{\phi} & -\frac{(\alpha+\beta x_1)^{-1}-(\alpha+\beta x_0)^{-1}}{\phi^2} & 0 & 0 \\
\vdots & & & & \vdots \\
-\frac{(\alpha+\beta x_k)^{-2}-(\alpha+\beta x_0)^{-2}}{\phi} & -\frac{(\alpha+\beta x_k)^{-2}x_k-(\alpha+\beta x_0)^{-2}x_0}{\phi} & -\frac{(\alpha+\beta x_k)^{-1}-(\alpha+\beta x_0)^{-1}}{\phi^2} & 0 & 0 \\
2\frac{(\alpha+\beta x_1)^{-3}-(\alpha+\beta x_0)^{-3}}{2\phi} & 2\frac{(\alpha+\beta x_1)^{-3}x_1-(\alpha+\beta x_0)^{-3}x_0}{2\phi} & \frac{(\alpha+\beta x_1)^{-2}-(\alpha+\beta x_0)^{2}}{2\phi^2} & \frac{x_1-x_0}{\phi_x} & \frac{(x_1-x_0)(x_1+x_0-2\mu)}{2\phi_x^2} \\
\vdots & & & & \vdots \\
2\frac{(\alpha+\beta x_k)^{-3}-(\alpha+\beta x_0)^{-3}}{2\phi} & 2\frac{(\alpha+\beta x_k)^{-3}x_k-(\alpha+\beta x_0)^{-3}x_0}{2\phi} & \frac{(\alpha+\beta x_k)^{-2}-(\alpha+\beta x_0)^{2}}{2\phi^2} & \frac{x_k-x_0}{\phi_x} & \frac{(x_k-x_0)(x_k+x_0-2u)}{2\phi_x^2}
\end{bmatrix}$$

After performing some rank-preserving modifications to this matrix, we get:

$$\begin{bmatrix}
\left(\alpha + \beta x_1\right)^{-2} - \left(\alpha + \beta x_0\right)^{-2} & \left(\alpha + \beta x_1\right)^{-2} x_1 - \left(\alpha + \beta x_0\right)^{-2} x_0 & \left(\alpha + \beta x_1\right)^{-1} - \left(\alpha + \beta x_0\right)^{-1} & 0 & 0 \\
\vdots & & & & \vdots \\
\left(\alpha + \beta x_k\right)^{-2} - \left(\alpha + \beta x_0\right)^{-2} & \left(\alpha + \beta x_k\right)^{-2} x_k - \left(\alpha + \beta x_0\right)^{-2} x_0 & \left(\alpha + \beta x_k\right)^{-1} - \left(\alpha + \beta x_0\right)^{-1} & 0 & 0 \\
\left(\alpha + \beta x_1\right)^{-3} - \left(\alpha + \beta x_0\right)^{-3} & \left(\alpha + \beta x_1\right)^{-3} x_1 - \left(\alpha + \beta x_0\right)^{-3} x_0 & \frac{1}{2}\left[\left(\alpha + \beta x_1\right)^{-2} - \left(\alpha + \beta x_0\right)^{-2}\right] & x_1 - x_0 & \left(x_1 - x_0\right)\left(x_1 + x_0 - 2\mu\right) \\
\vdots & & & & \vdots \\
\left(\alpha + \beta x_k\right)^{-3} - \left(\alpha + \beta x_0\right)^{-3} & \left(\alpha + \beta x_k\right)^{-3} x_k - \left(\alpha + \beta x_0\right)^{-3} x_0 & \frac{1}{2}\left[\left(\alpha + \beta x_k\right)^{-2} - \left(\alpha + \beta x_0\right)^{-2}\right] & x_k - x_0 & \left(x_k - x_0\right)\left(x_k + x_0 - 2\mu\right)
\end{bmatrix}$$

which is of full rank.

It is worth pointing out that unlike the example in (C.1), $p(X \mid Y)$ **in this example is not in the exponential family**, since:

$$\begin{aligned}
p(x \mid y) &= \frac{p(y \mid x)p(x)}{p(y)} = \frac{\mathbb{N}\left((a + bx)^{-1}, \sigma_y^2\right)\mathbb{N}\left(\mu, \sigma_x^2\right)}{p(y)} \\
&= \exp\left\{ -\frac{\left(y - \frac{1}{a+bx}\right)^2}{2\sigma_y^2} + \log\frac{1}{\sqrt{2\pi}\sigma_y} - \frac{(x-\mu)^2}{2\sigma_x^2} + \log\frac{1}{\sqrt{2\pi}\sigma_x} - \log p(y) \right\} \\
&= \exp\left\{ -\frac{\frac{1}{(a+bx)^2} - \frac{2y}{a+bx} + y^2}{2\sigma_y^2} + \log\frac{1}{\sqrt{2\pi}\sigma_y} - \frac{(x-\mu)^2}{2\sigma_x^2} + \log\frac{1}{\sqrt{2\pi}\sigma_x} - \log p(y) \right\}.
\end{aligned}$$

C.3 X AND Y ARE BINARY

Suppose $p(X = 0, Y = 1) = p_1$, $p(X = 1, Y = 0) = p_2$, $p(X = 0, Y = 0) = p_3$, and $p(X = 1, Y = 1) = p_4$, where $\sum_{i=1}^4 p_i = 1, p_i \neq 0$. The unknown parameters of interest are $\theta = (p_1, p_2, p_3, p_4)$.

In this binary case, there are at most two distinct values of X as 0 or 1. According to Theorem 1, $p(X, Y)$ is identifiable if any one of p_i is known or marginal distribution of either X or Y is known.

In order to prove the above claim, we look at two distinct parameterizations of $p(X, Y)$.

C.3.1 Parameterization 1

Suppose $p_1 = p(X = 0, Y = 1)$, $p_2 = p(X = 1, Y = 0)$, $p_3 = p(X = 0, Y = 0)$, $p_4(X = 1, Y = 1)$, $p_i \neq 0$, $i = 1, \ldots, 4$.

Since $p(X \mid Y)$ is nonparametrically identified, we obtain the following three equations with four unknowns:

$$p(X = 1 \mid Y = 1) = \frac{p_4}{p_1 + p_4}, \qquad p(X = 1 \mid Y = 0) = \frac{p_2}{p_2 + p_3}, \qquad \sum_{i=1}^{4} p_i = 1$$

In order to possibly achieve identification, we need to assume one parameter is known. We consider the four different scenarios one by one.

1. Assume p_1 is known, then $|J| \neq 0 \implies$ target law is identified

$$J = \begin{bmatrix} 0 & 0 & \frac{p_1}{(p_1+p_4)^2} \\ \frac{p_3}{(p_2+p_3)^2} & \frac{-p_2}{(p_2+p_3)^2} & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

2. Assume p_2 is known, then $|J| \neq 0 \implies$ target law is identified

$$J = \begin{bmatrix} \frac{-p_4}{(p_1+p_4)^2} & 0 & \frac{p_1}{(p_1+p_4)^2} \\ 0 & \frac{p_3}{(p_2+p_3)^2} & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

3. Assume p_3 is known, then $|J| \neq 0 \implies$ target law is identified

$$J = \begin{bmatrix} \frac{-p_4}{(p_1+p_4)^2} & 0 & \frac{p_1}{(p_1+p_4)^2} \\ 0 & \frac{p_3}{(p_2+p_3)^2} & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

4. Assume p_4 is known, then $|J| \neq 0 \implies$ target law is identified

$$J = \begin{bmatrix} \frac{-p_4}{(p_1+p_4)^2} & 0 & 0 \\ 0 & \frac{p_3}{(p_2+p_3)^2} & \frac{-p_2}{(p_2+p_3)^2} \\ 1 & 1 & 1 \end{bmatrix}$$

In the binary case, it is also useful to assume

1. Assume $p(Y = 1) = p_1 + p_4$ is known, then $|J| \neq 0 \implies$ target law is identified

$$J = \begin{bmatrix} -\frac{p_4}{(p_1+p_4)^2} & 0 & 0 & \frac{p_1}{(p_1+p_4)^2} \\ 0 & \frac{p_3}{(p_2+p_3)^2} & -\frac{p_2}{(p_2+p_3)^2} & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

2. Assume $p(X = 1) = p_2 + p_4$ is known, then $|J| \neq 0 \implies$ target law is identified

$$J = \begin{bmatrix} -\frac{p_4}{(p_1+p_4)^2} & 0 & 0 & \frac{p_1}{(p_1+p_4)^2} \\ 0 & \frac{p_3}{(p_2+p_3)^2} & -\frac{p_2}{(p_2+p_3)^2} & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

C.3.2 Parameterization 2

We can also adopt another parameterization. Suppose

$$X \sim \mathrm{Bern}(p), \qquad Y \mid X \sim \mathrm{Bern}(a + bX)$$

More specifically,

$$\begin{aligned} p(x) &= \exp\left\{x \log \frac{p}{1-p} + \log(1-p)\right\} = \exp\left\{x \cdot \eta_x - \log\left(1 + e^{\eta_x}\right)\right\} \quad \text{where } \eta_x = \log \frac{p}{1-p} \\ p(y \mid x) &= (a+bx)^y (1-a-bx)^{1-y} \\ &= \exp\left\{y \log \frac{a+bx}{1-(a+bx)} + \log[1-(a+bx)]\right\} \end{aligned}$$

The parameter vector of interest is $\theta = (a, b, \eta_x)$. Based on Theorem 1, we have the following equations. Note that since X is binary, there are at most two distinct values of X. Therefore, we have the following two equations:

$$\begin{aligned} \phi_1(\theta) &= \log \frac{a+bx_1}{1-(a+bx_1)} - \log \frac{a+bx_0}{1-(a+bx_0)} \\ \zeta_1(\sigma) &= \log\left[1-(a+bx_1)\right] - \log\left[1-(a+bx_0)\right] + (x_1 - x_0)\,\eta_x, \text{ where } x_1 = 1, x_0 = 0. \end{aligned}$$

The resulted Jacobian matrix is:

$$J = \begin{bmatrix} \frac{1}{(a+b)[1-(a+b)]} - \frac{1}{a(1-a)} & \frac{1}{(a+b)[1-(a+b)]} & 0 \\ \frac{-1}{1-(a+b)} + \frac{1}{1-a} & \frac{-1}{1-(a+b)} & x_1 - x_0 \end{bmatrix}$$

This concludes that in order to establish target law identification, we need to know at least one parameter in $\{a, b, \eta_x\}$.

It is straightforward to show that $p(X \mid Y)$ **lies in the exponential family**.

C.4 X IS BINARY AND $Y \mid X$ IS NORMAL UNDER CANONICAL LINK

Suppose

$$X \sim \mathrm{Bern}(p), \qquad Y \mid X \sim \mathbb{N}\left(a + bX, \sigma_y^2\right).$$

More specifically,

$$\begin{aligned} p(x) &= p^x (1-p)^{1-x} = \exp\left\{x \cdot \log \frac{p}{1-p} + \log(1-p)\right\} = \exp\left\{x \cdot \eta - \log\left(1+e^{\eta}\right)\right\}, \text{ where } \eta = \log \frac{p}{1-p} \\ p(y \mid x) &= \exp\left\{\frac{y(a+bx) - \frac{1}{2}(a+bx)^2}{\phi} + \left[-\frac{y^2}{2\phi} - \frac{1}{2}\log\left(2\pi\phi\right)\right]\right\}, \text{ where } \phi = \sigma_y^2. \end{aligned}$$

The unknown parameter vector of interest is $\theta = (a, b, \phi, \eta)$. According to Theorem 1, $p(X, Y)$ is identifiable if at either a or η is known, in addition to knowing one extra parameter in θ. Knowing η is equivalent to knowing p.

In order to prove the above claim, we can construct the following equations: (note that when X is binary, we only have at most two distinct values)

$$\begin{aligned} \phi_1(\theta) &= \frac{(a+bx_1) - (a+bx_0)}{\phi} = \frac{b\left(x_i - x_0\right)}{\phi} \\ \zeta_1(\theta) &= -\frac{\left(a+bx_1\right)^2 - \left(a+bx_0\right)^2}{2\phi} + \eta\left(x_1 - x_0\right), \text{ where } x_1 = 1, x_0 = 0. \end{aligned}$$

The Jacobian matrix is:

$$J = \begin{bmatrix} 0 & \frac{x_1 - x_0}{\phi} & -\frac{b(x_1-x_0)}{\phi^2} & 0 \\ -\frac{b(x_1-x_0)}{\phi} & -\frac{a(x_1-x_0)+b\left(x_1^2-x_0^2\right)}{\phi} & \frac{(a+bx_1)^2-(a+bx_0)^2}{2\phi^2} & x_1 - x_0 \end{bmatrix}.$$

After some rank-preserving operations, we get:

$$\begin{bmatrix} 0 & x_1 - x_0 & x_1 - x_0 & 0 \\ 1 & a\left(x_1 - x_0\right) + b\left(x_1^2 - x_0^2\right) & a\left(x_1 - x_0\right) + \frac{b}{2}\left(x_1^2 - x_0^2\right) & 1 \end{bmatrix}.$$

This concludes the claim that a sufficient set of assumptions for target law identification is knowing either a or η, in addition to knowing one more parameter in θ.

Note that in this example, $p(X \mid Y)$ **is in exponential family** since:

$$\begin{aligned} p(x \mid y) &= \frac{p(y \mid x)p(x)}{p(y)} = \frac{N_y\left(a + bx, \sigma_y^2\right) p^x (1-p)^{1-x}}{p(y)} \\ &= \exp\left\{-\frac{1}{2}\left(\frac{y - (a + bx)}{\sigma_y}\right)^2 + \log\frac{1}{\sqrt{2\pi}\sigma_y} + x \log p + (1 - x)\log(1 - p) - \log\left[p(y)\right]\right\} \\ &= \exp\left\{-\frac{1}{2}\frac{\left(x, x^2\right)\left(2ab - 2by, b^2\right)^T + (a - y)^2}{\sigma_y^2} + \log\frac{1}{\sqrt{2\pi}\sigma_y} + x\log p + (1 - x)\log(1 - p) - \log\left[p(y)\right]\right\} \\ &= \exp\left\{\left(x, x^2\right)\left(-\frac{ab - by}{\sigma_y^2} + log(\frac{p}{1-p}), -\frac{b^2}{2\sigma_y^2}\right)^T - \frac{(a - y)^2}{2\sigma_y^2} + \log\frac{1}{\sqrt{2\pi}\sigma_y} + log(1 - p) - \log\left[p(y)\right]\right\}. \end{aligned}$$

C.5 X IS POISSON AND $Y \mid X$ IS NORMAL UNDER CANONICAL LINK

Suppose

$$X \sim \text{Poisson}(\lambda), \qquad Y \mid X \sim \mathbb{N}\left(a + bx, \sigma_y^2\right).$$

More specifically,

$$\begin{aligned} p(y \mid x) &= \exp\left\{\frac{y(a + bx) - \frac{1}{2}(a + bx)^2}{\phi} + \left[-\frac{y^2}{2\phi} - \frac{1}{2}\log\left(2\pi\phi\right)\right]\right\}, \text{ where } \phi = \sigma_y^2 \\ p(x = k) &= \frac{\lambda^k e^{-\lambda}}{k!} = \exp\{k \log\lambda - \lambda - \log k!\} = \exp\left\{k\eta_x - e^{\eta_x} - \log k!\right\}, \text{ where } \eta_x = \log\lambda \end{aligned}$$

The unknown parameter vector of interest is $\theta = \left(a, b, \sigma_y^2, \lambda\right)$. According to Theorem 1, $p(X, Y)$ is identifiable if either a or λ is known.

In order to prove the above claim, we can construct the following equations:

$$\begin{aligned} \phi_i(\theta) &= \frac{(a + bx_i) - (a + bx_0)}{\phi} = \frac{b\left(x_i - x_0\right)}{\phi} \\ \zeta_i(\theta) &= -\frac{\left(a + bx_i\right)^2 - \left(a + bx_0\right)^2}{2\phi} + \eta_x\left(x_i - x_0\right) + \left(-\log x_i! + \log x_0!\right), \quad \text{where } i \in (1, \ldots, k) \end{aligned}$$

The Jacobian matrix is then as follows:

$$J = \begin{bmatrix} 0 & \frac{x_1 - x_0}{\phi} & -\frac{(bx_1 - x_0)}{\phi^2} & 0 \\ 0 & \frac{x_2 - x_0}{\phi} & -\frac{(bx_2 - x_0)}{\phi^2} & 0 \\ \vdots & & & \vdots \\ 0 & \frac{x_k - x_0}{\phi} & -\frac{(bx_k - x_0)}{\phi^2} & 0 \\ -\frac{b(x_1 - x_0)}{\phi} & -\frac{a(x_1 - x_0) + b\left(x_1^2 - x_0^2\right)}{\phi} & \frac{(a + bx_1)^2 - (a + bx_0)^2}{2\phi^2} & x_1 - x_0 \\ -\frac{b(x_2 - x_0)}{\phi} & -\frac{a(x_2 - x_0) + b\left(x_2^2 - x_0^2\right)}{\phi} & \frac{(a + bx_2)^2 - (a + bx_0)^2}{2\phi^2} & x_2 - x_0 \\ \vdots & & & \vdots \\ -\frac{b(x_k - x_0)}{\phi} & -\frac{a(x_k - x_0) + b\left(x_k^2 - x_0^2\right)}{\phi} & \frac{(a + bx_k)^2 - (a + bx_0)^2}{2\phi^2} & x_k - x_0 \end{bmatrix}.$$

After some rank-preserving operations, we get:

$$\begin{bmatrix} 0 & x_1 - x_0 & x_1 - x_0 & 0 \\ 0 & x_2 - x_0 & x_2 - x_0 & 0 \\ \vdots & & & \vdots \\ 0 & x_k - x_0 & x_k - x_0 & 0 \\ x_1 - x_0 & a\left(x_1 - x_0\right) + b\left(x_1^2 - x_0^2\right) & a\left(x_1 - x_0\right) + \frac{b}{2}\left(x_1^2 - x_0^2\right) & x_1 - x_0 \\ x_2 - x_0 & a\left(x_2 - x_0\right) + b\left(x_2^2 - x_0^2\right) & a\left(x_2 - x_0\right) + \frac{b}{2}\left(x_2^2 - x_0^2\right) & x_2 - x_0 \\ \vdots & & & \vdots \\ x_k - x_0 & a\left(x_k - x_0\right) + b\left(x_k^2 - x_0^2\right) & a\left(x_k - x_0\right) + \frac{b}{2}\left(x_k^2 - x_0^2\right) & x_k - x_0 \end{bmatrix}.$$

We need to know either a or η_x to establish identifiability.

Note that in this example, $p(X \mid Y)$ **is in the exponential family** since:

$$\begin{aligned} p(x \mid y) &= \frac{p(y \mid x)p(x)}{p(y)} = \frac{N_y\left(a + bx, \sigma_y^2\right) \frac{\lambda^x e^{-\lambda}}{x!}}{p(y)} \\ &= \exp\left\{ -\frac{1}{2}\left(\frac{y - (a + bx)}{\sigma_y}\right)^2 + \log\frac{1}{\sqrt{2\pi}\sigma_y} + x\log\lambda - \lambda - \log x! - \log p(y) \right\} \\ &= \frac{1}{x!}\exp\left\{ -\frac{1}{2}\frac{\left(x, x^2\right)\left(2ab - 2by - 2\sigma_y^2\log\lambda, b^2\right)^T + (a - y)^2}{\sigma_y^2} + \log\frac{1}{\sqrt{2\pi}\sigma_y} - \lambda - \log p(y) \right\} \\ &= \frac{1}{x!}\exp\left\{ \left(x, x^2\right)\left(-\frac{ab - by - \sigma_y^2\log\lambda}{\sigma_y^2}, -\frac{b^2}{2\sigma_y^2}\right)^T - \frac{(a - y)^2}{2\sigma_y^2} + \log\frac{1}{\sqrt{2\pi}\sigma_y} - \lambda - \log p(y) \right\}. \end{aligned}$$

C.6 X IS EXPONENTIAL AND $Y \mid X$ IS NORMAL UNDER CANONICAL LINK

Suppose

$$X \sim \text{exponential}(\lambda), \qquad Y \mid X \sim \mathbb{N}\left(a + bx, \sigma_y^2\right).$$

More specifically,

$$\begin{aligned} p(x) &= \lambda e^{-\lambda x} = \exp\{-\lambda x + \log\lambda\} \\ p(y \mid x) &= \exp\left\{ \frac{y(a + bx) - \frac{1}{2}(a + bx)^2}{\phi} + \left[-\frac{y^2}{2\phi} - \frac{1}{2}\log\left(2\pi\phi\right) \right] \right\} \quad \text{where } \phi = \sigma_y^2 \end{aligned}$$

The unknown vector of parameters is $\theta = (a, b, \phi, \lambda)$. According to Theorem 1, $p(X, Y)$ is identifiable if either a or λ is known.

In order to prove the above claim, we can construct the following equations:

$$\begin{aligned} \phi_i(\theta) &= \frac{b\left(x_i - x_0\right)}{\phi} \\ \zeta_i(\theta) &= -\frac{\left(a + bx_i\right)^2 - \left(a + bx_0\right)^2}{2\phi} - \lambda\left(x_i - x_0\right), \quad \text{where } i \in (1, \ldots, k) \end{aligned}$$

The Jacobian matrix is

$$J = \begin{bmatrix} 0 & \frac{x_1 - x_0}{\phi} & -\frac{b(x_1 - x_0)}{\phi^2} & 0 \\ \vdots & & & \vdots \\ 0 & \frac{x_k - x_0}{\phi} & -\frac{b(x_k - x_0)}{\phi^2} & 0 \\ -\frac{b(x_1 - x_0)}{\phi} & -\frac{a(x_1 - x_0) + b\left(x_1^2 - x_0^2\right)}{\phi} & \frac{(a + bx_1)^2 - (a + bx_0)^2}{2\phi^2} & -(x_1 - x_0) \\ \vdots & & & \vdots \\ -\frac{b(x_k - x_0)}{\phi} & -\frac{a(x_k - x_0) + b\left(x_k^2 - x_0^2\right)}{\phi} & \frac{(a + bx_k)^2 - (a + bx_0)^2}{2\phi^2} & -(x_k - x_0) \end{bmatrix}.$$

After some rank-preserving operations, we get:

$$\begin{bmatrix} 0 & x_1 - x_0 & x_1 - x_0 & 0 \\ \vdots & & & \vdots \\ 0 & x_k - x_0 & x_k - x_0 & 0 \\ x_1 - x_0 & -\left[a\left(x_1 - x_0\right) + b\left(x_1^2 - x_0^2\right)\right] & -\left[a\left(x_1 - x_0\right) + \frac{b}{2}\left(x_1^2 - x_0^2\right)\right] & x_1 - x_0 \\ \vdots & & & \vdots \\ x_k - x_0 & -\left[a\left(x_k - x_0\right) + b\left(x_k^2 - x_0^2\right)\right] & -\left[a\left(x_k - x_0\right) + \frac{b}{2}\left(x_k^2 - x_0^2\right)\right] & x_k - x_0 \end{bmatrix}.$$

This concludes the initial claim.

Note that in this example, $p(X \mid Y)$ **is in the exponential family** since:

$$\begin{aligned} p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)} &= \frac{N\left((a+bx), \sigma_y^2\right)\lambda e^{-\lambda x}}{p(y)} \\ &= \exp\left\{-\frac{1}{2}\left(\frac{y-(a+bx)}{\sigma_y}\right)^2 + \log\frac{1}{\sqrt{2\pi}\sigma_y} + \log\lambda - \lambda x - \log p(y)\right\} \\ &= \exp\left\{-\frac{1}{2}\frac{\left(x, x^2\right)\left(2ab - 2by - 2\sigma_y^2\lambda, b^2\right)^T + (a-y)^2}{\sigma_y^2} + \log\frac{1}{\sqrt{2\pi}\sigma_y} + \log\lambda - \log p(y)\right\} \\ &= \exp\left\{\left(x, x^2\right)\left(-\frac{ab - by - \sigma_y^2\lambda}{\sigma_y^2}, -\frac{b^2}{2\sigma_y^2}\right)^T - \frac{(a-y)^2}{2\sigma_y^2} + \log\frac{1}{\sqrt{2\pi}\sigma_y} + \log\lambda - \log p(y)\right\}. \end{aligned}$$

C.7 X IS EXPONENTIAL AND $Y \mid X$ IS EXPONENTIAL UNDER CANONICAL LINK

Suppose

$$\begin{aligned} X &\sim \text{exponential}(\lambda_x) \\ Y \mid X &\sim \text{exponential}(\lambda) = \exp\{y(-\lambda) + \log\lambda\} = \exp\{y(a+bx) + \log[-(a+bx)]\}. \end{aligned}$$

The unknown parameter vector is $\theta = (a, b, \lambda_x)$. According to Theorem 1 and without any further assumptions on θ, $p(X, Y)$ is identifiable.

In order to prove the above claim, we can construct the following equations:

$$\begin{aligned} &\phi_i(\theta) = b\left(x_i - x_0\right) \\ &\zeta_i(\theta) = \log\left[-(a + bx_i)\right] - \log\left[-(a + bx_0)\right] - \lambda_x\left(x_i - x_0\right), \quad i \in (1, \ldots, k) \end{aligned}$$

The Jacobian matrix is

$$J = \begin{bmatrix} 0 & x_1 - x_0 & 0 \\ \vdots & & \vdots \\ 0 & x_k - x_0 & 0 \\ \frac{1}{a+bx_1} - \frac{1}{a+bx_0} & \frac{x_1}{a+bx_1} - \frac{x_0}{a+bx_0} & -\left(x_1 - x_0\right) \\ \vdots & & \vdots \\ \frac{1}{a+bx_k} - \frac{1}{a+bx_0} & \frac{x_k}{a+bx_k} - \frac{x_0}{a+bx_0} & -\left(x_k - x_0\right) \end{bmatrix}.$$

After some rank-preserving operations, we get:

$$\begin{bmatrix} 0 & x_1 - x_0 & 0 \\ 0 & x_2 - x_0 & 0 \\ \vdots & & \vdots \\ 0 & x_k - x_0 & 0 \\ \frac{1}{(a+bx_1)(a+bx_0)} & \frac{1}{(a+bx_1)(a+bx_0)} & 1 \\ \frac{1}{(a+bx_2)(a+bx_0)} & \frac{1}{(a+bx_2)(a+bx_0)} & 1 \\ \vdots & & \vdots \\ \frac{1}{(a+bx_k)(a+bx_0)} & \frac{1}{(a+bx_k)(a+bx_0)} & 1 \end{bmatrix}.$$

This matrix is full rank and thus it concludes the initial claim.

Note that in this example, $p(X \mid Y)$ **is not in exponential family (unless a and b are known)**, since:

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)} = \frac{\exp\{y(a+bx) + \log[-(a+bx)] + x(-\lambda x) + \log \lambda_x\}}{p(y)}.$$

The main difficulty is with the term $\log[-(a+bx)]$.

D ESTIMATION PROOFS

D.1 THEOREM 2 (CONDITIONAL LIKELIHOOD WITH ORDER STATISTICS)

Proof. Denote $l(\theta) = -\frac{2}{N(N-1)} \sum_{1 \le i < k \le N} R_{x_i} R_{y_i} R_{x_k} R_{y_k} \log\{1 + Q_{ik}(\theta)\}$. Following the Taylor expansion, we have

$$0 = \frac{\partial l(\widetilde{\theta})}{\partial \theta} = \frac{\partial l(\theta_0)}{\partial \theta} + (\widetilde{\theta} - \theta_0)\frac{\partial^2 l(\theta_0)}{\partial \theta^2} + o_p(N^{-1/2}).$$

Therefore,

$$\sqrt{N}(\widetilde{\theta} - \theta_0) = -\left\{\frac{\partial^2 l(\theta_0)}{\partial \theta^2}\right\}^{-1} \sqrt{N}\frac{\partial l(\theta_0)}{\partial \theta} + o_p(1).$$

Since both $\frac{\partial^2 l(\theta_0)}{\partial \theta^2}$ and $\frac{\partial l(\theta_0)}{\partial \theta}$ are U-statistics, from the theory of U-statistics, we have

$$\frac{\partial^2 l(\theta_0)}{\partial \theta^2} \xrightarrow{p} A, \text{ and } \sqrt{N}\frac{\partial l(\theta_0)}{\partial \theta} \xrightarrow{d} \mathbb{N}(0, B),$$

which completes the proof. □

D.2 THEOREM 3 (GENERALIZED ESTIMATING EQUATIONS)

Proof. The proof of (a) is straightforward following the standard argument of generalized estimating equations, so omitted here. In order to find the optimal choice for $f(Y)$, we can compute

$$\begin{aligned}
C &= \mathbb{E}\left\{-\Psi'\left(X, Y, R_x, R_y; \theta_0\right)\right\} \\
&= \mathbb{E}\left[\frac{R_x R_y}{p\left(R_y = 1 \mid R_x = 1, X\right)} \left.\frac{\partial \mathbb{E}(X \mid Y)}{\partial \theta}\right|_{\theta=\theta_0} f(Y)^T\right] \\
&= \mathbb{E}\left[R_x \left.\frac{\partial \mathbb{E}(X \mid Y)}{\partial \theta}\right|_{\theta=\theta_0} f(Y)^T\right] \\
&= \mathbb{E}\left\{w(Y) a(Y) f(Y)^T\right\},
\end{aligned}$$

and

$$\begin{aligned}
D &= \mathbb{E}\left[\frac{R_x R_y}{p^2\left(R_y = 1 \mid R_x = 1, X\right)} (X - \mathbb{E}(X \mid Y))^2 f(Y) f(Y)^T\right] \\
&= \mathbb{E}\left[R_x \frac{(X - \mathbb{E}(X \mid Y))^2}{\pi(X)} f(Y) f(Y)^T\right] \\
&= \mathbb{E}\left[w(Y) \frac{(X - \mathbb{E}(X \mid Y))^2}{\pi(X)} f(Y) f(Y)^T\right] \\
&= \mathbb{E}\left[w(Y) E\left[\frac{(X - \mathbb{E}(X \mid Y))^2}{\pi(X)} \mid Y\right] f(Y) f(Y)^T\right] \\
&= \mathbb{E}\left[w(Y) b(Y) f(Y) f(Y)^T\right],
\end{aligned}$$

where $b(Y) = \mathbb{E}\left[\frac{(X - \mathbb{E}(X \mid Y))^2}{\pi(X)} \mid Y\right]$ and $w(Y) = p(R_x = 1 \mid Y)$. Based on Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left(uv^T\right)\left\{\mathbb{E}\left(vv^T\right)\right\}^{-1} \mathbb{E}\left(vu^T\right) \lesssim \mathbb{E}\left(uu^T\right)$$

with equality hold at $u = v$. Here $M \lesssim N$ simply means $M - N$ is negative semi-definite.

Define $v = \sqrt{w(Y)}\sqrt{b(Y)} f(Y)$ and $u = \sqrt{\frac{w(Y)}{b(Y)}} a(Y)$, then we have

$$\mathbb{E}\{w(Y) f(Y) a(Y)^T\}\left[\mathbb{E}\{w(Y) b(Y) f(Y) f(Y)^T\}\right]^{-1} \mathbb{E}\{w(Y) a(Y) f(Y)^T\} \lesssim \mathbb{E}\left\{\frac{w(Y)}{b(Y)} a(Y) a(Y)^T\right\}, i.e.,$$

$$\mathbb{E}\left\{\frac{w(Y)}{b(Y)}a(Y)a(Y)^T\right\}^{-1}\mathbb{E}\{w(Y)b(Y)f(Y)f(Y)^T\}\mathbb{E}\{w(Y)f(Y)a(Y)^T\}^{-1} \gtrsim \mathbb{E}\left\{\frac{w(Y)}{b(Y)}a(Y)a(Y)^T\right\}^{-1}.$$

Note that the right-hand side is irrespective of $f(Y)$. Thus, when $f(Y) = f_{opt}(Y) = \frac{a(Y)}{b(Y)}$, the equality holds, and we have the optimal variance $\left\{\frac{w(Y)}{b(Y)}a(Y)a(Y)^T\right\}^{-1}$. □

E ADDITIONAL DISCUSSIONS ON ESTIMATION

E.1 NONPARAMETRIC ESTIMATION UNDER ADDITIONAL ASSUMPTIONS

In addition to independence restrictions in display (3), we assume $p(R_y = 1 \mid R_x, X)$ is not a function of X when $R_x = 0$. This additional assumptions moves us from the criss-cross MNAR model to the permutation model considered by Robins [1997]. In the permutation model, one can proceed with estimation of arbitrary functions of X and Y as follows.

Let our parameter of interest be $\beta_h = \mathbb{E}[h(X,Y)]$, which can be identified via the following function of the observed data:

$$\beta_h = \mathbb{E}\left[\frac{R_x \; R_y \; h(X,Y)}{p(R_x = 1 \mid Y) \; p(R_y = 1 \mid R_x = 1, X^*)}\right].$$

The core idea of deriving the efficient influence function (EIF) for β_h is to use an intermediate variable that first takes care of the missingness of X, and then Y in a sequential manner. Intuitively, this is due to the fact that we can rewrite β_h via an intermediate variable $\widetilde{\beta}_h(X, R_x, Y)$ as follows:

$$\widetilde{\beta}_h(X, R_x, Y) = \frac{R_x}{p\left(R_x = 1 \mid Y\right)} \; h\left(X, Y\right), \qquad \beta_h = \mathbb{E}\left[\frac{R_y}{p\left(R_y = 1 \mid R_x, X^*\right)} \; \widetilde{\beta}_h(X, R_x, Y)\right].$$

The claim made by Robins [1997] is that EIF for β_h is equal to the EIF for $\mathbb{E}\left[\dfrac{R_y}{p\left(R_y = 1 \mid R_x, X^*\right)} \; \phi(\widetilde{\beta}_h)\right]$, where $\phi(\widetilde{\beta}_h) = \text{EIF}_{\widetilde{\beta}_h} + \mathbb{E}[\widetilde{\beta}_h]$ and $\text{EIF}_{\widetilde{\beta}_h}$ denotes the efficient influence function for $\mathbb{E}\big[\widetilde{\beta}_h(X, R_x, Y)\big]$. Therefore, we first need to derive the EIF for $\mathbb{E}\big[\widetilde{\beta}_h(X, R_x, Y)\big]$.

$$\begin{aligned}
\left.\frac{\partial \mathbb{E}[\widetilde{\beta}_h\left(p_\varepsilon\right)]}{\partial \varepsilon}\right|_{\varepsilon=0} &= \left.\frac{\partial}{\partial \varepsilon} \int \frac{R_x h\left(X,Y\right)}{p\left(R_x = 1 \mid Y\right)} dp_\varepsilon\left(X,Y,R_x\right)\right|_{\varepsilon=0} \\
&= -\int \frac{R_x h\left(X,Y\right)}{p\left(R_x = 1 \mid Y\right)} S\left(R_x \mid Y\right) dp\left(X,Y,R_x\right) + \int \frac{R_x h\left(X,Y\right)}{p\left(R_x = 1 \mid Y\right)} S\left(X,Y,R_x\right) dp\left(X,Y,R_x\right) \\
&= -\int \frac{R_x \mathbb{E}\left[h\left(X,Y\right) \mid R_x = 1, Y\right]}{p\left(R_x = 1 \mid Y\right)} S\left(R_x \mid Y\right) dp\left(R_x, Y\right) + \int \left\{\frac{R_x h\left(X,Y\right)}{p\left(R_x = 1 \mid Y\right)} - \mathbb{E}\left[h\left(X,Y\right)\right]\right\} S\left(X,Y,R_x\right) dp\left(X,Y,R_x\right) \\
&= -\int \left\{\frac{R_x \mathbb{E}\left[h\left(X,Y\right) \mid R_x = 1, Y\right]}{p\left(R_x = 1 \mid Y\right)} - \mathbb{E}\left[h\left(X,Y\right) \mid R_x = 1, Y\right]\right\} S\left(R_x, Y\right) dp\left(R_x, Y\right) \\
&+ \int \left\{\frac{R_x h\left(X,Y\right)}{p\left(R_x = 1 \mid Y\right)} - \mathbb{E}\left[h\left(X,Y\right)\right]\right\} S\left(X,Y,R_x\right) dp\left(X,Y,R_x\right) \\
&= -\int \left\{\frac{R_x \mathbb{E}\left[h\left(X,Y\right) \mid R_x = 1, X\right]}{p\left(R_x = 1 \mid Y\right)} - \mathbb{E}\left[h\left(X,Y\right) \mid R_x = 1, Y\right]\right\} S\left(Y, R_x, X\right) dp\left(R_x, X, Y\right) \\
&+ \int \left\{\frac{R_x h\left(X,Y\right)}{p\left(R_x = 1 \mid Y\right)} - \mathbb{E}\left[h\left(X,Y\right)\right]\right\} S\left(X,Y,R_x\right) dp\left(X,Y,R_x\right).
\end{aligned}$$

Therefore, the efficient influence function for $\mathbb{E}[\widetilde{\beta}_h]$, denoted by $\text{EIF}_{\widetilde{\beta}_h}$, is as follows

$$\text{EIF}_{\widetilde{\beta}_h} = \frac{R_x}{p\left(R_x = 1 \mid Y\right)} \Big\{h\left(X,Y\right) - \mathbb{E}\left[h\left(X,Y\right) \mid R_x = 1, Y\right]\Big\} + \Big\{\mathbb{E}[h(X,Y) \mid R_x = 1, Y] - \mathbb{E}[h(X,Y)]\Big\}.$$

Thus we get:

$$\phi(\widetilde{\beta}_h) = \frac{R_x}{p\left(R_x = 1 \mid Y\right)} \Big\{h\left(X,Y\right) - \mathbb{E}\left[h\left(X,Y\right) \mid R_x = 1, Y\right]\Big\} + \mathbb{E}\big[h(X,Y) \mid R_x = 1, Y\big].$$

Following a similar procedure, we can easily obtain the EIF for $\mathbb{E}\left[\dfrac{R_y}{p\left(R_y = 1 \mid R_x, X^*\right)} \; \phi(\widetilde{\beta}_h)\right]$, which yields the EIF for β_h as follows:

$$\text{EIF}_{\beta_h} = \frac{R_y}{p\left(R_y = 1 \mid R_x, X^*\right)} \Big\{\phi(\widetilde{\beta}_h) \; - \; \mathbb{E}\big[\phi(\widetilde{\beta}_h) \mid R_y, R_x, X^*\big]\Big\} + \Big\{ \; \mathbb{E}\big[\phi(\widetilde{\beta}_h) \mid R_y = 1, R_x, X^*\big] - \beta_h\Big\}.$$

E.2 MAXIMUM LIKELIHOOD ESTIMATION

In the criss-cross MNAR model, the *observed full data likelihood*, denoted by $\mathcal{L}_{\text{obs}}(Z;\theta)$, can be written down as follows:

$$\begin{aligned}\mathcal{L}_{\text{obs}}(X,Y,R;\theta,\psi) = &\prod_{R_x=1,R_y=1} p(X,Y,R_x=1,R_y=1) \times \prod_{R_x=1,R_y=0} \int p(X,Y,R_x=1,R_y=0)dy \\ &\times \prod_{R_x=0,R_y=1} \int p(X,Y,R_x=0,R_y=1)dx \times \prod_{R_x=0,R_y=0} \int p(X,Y,R_x=0,R_y=0)dxdy\end{aligned}$$

Under the conditions of Theorem 1 and Condition 1, one can simply estimate the entire parameter vector of the full law, assuming the parametric forms of the propensity scores in the missingness mechanism are known.

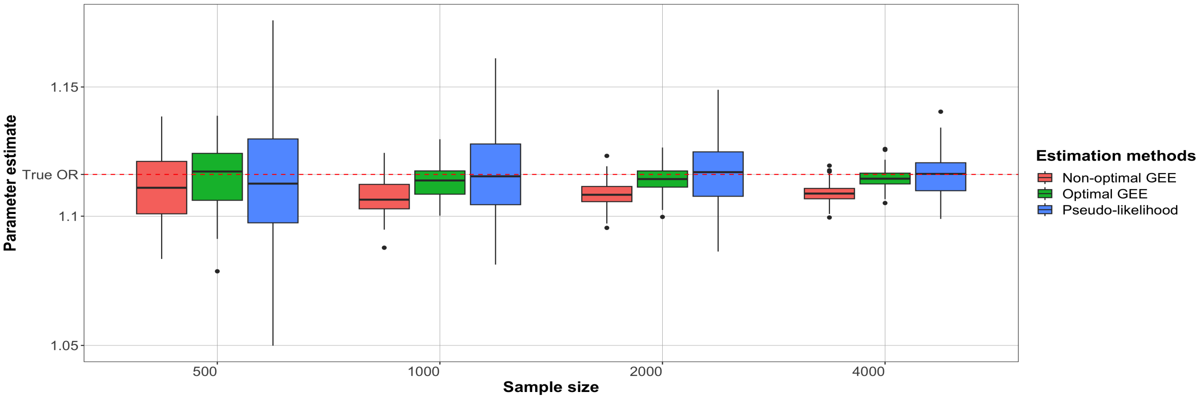

Figure 1: OR estimation under model misspecification.

F ADDITIONAL EXPERIMENTAL RESULTS

F.1 SIMULATION RESULTS

Varying ρ. We examine the effect of changing the correlation coefficient on the efficiency of the estimators by varying ρ across the range of values from -0.9 to 0.9, with increments of 0.2. The sample size used is $N = 1000$. Table 1 displays the standard deviation (SD) of the three suggested estimators for different values of ρ. To avoid distorting the SD patterns after applying the Delta method, we summarize the SD of the direct estimates of each method instead of converting it to OR. The results indicate that both GEE methods provide more efficient estimators when X and Y are highly correlated, but exhibit more estimation uncertainty when the correlation is low. In contrast, the conditional likelihood estimator has less variability when the correlation is low.

Table 1: Standard deviation of estimators with varying correlation between X and Y

ρ	β **(non-optimal GEE)**	β **(optimal GEE)**	**logOR(conditional likelihood)**
-0.9	0.0468	0.0354	0.1272
-0.7	0.0678	0.0622	0.0470
-0.5	0.0847	0.1033	0.0268
-0.3	0.108	0.1319	0.0206
-0.1	0.127	0.1023	0.0201
0.1	0.118	0.0979	0.0179
0.3	0.154	0.0783	0.0189
0.5	0.0877	0.0535	0.0267
0.7	0.0628	0.0432	0.0413
0.9	0.0296	0.0211	0.0917

Model misspecification. To understand the behavior of the proposed estimators under model misspecification, we generate data under missing mechanism for Y as $p(R_y = 1 \mid X, R_x) = \text{expit}(2 - R_x + 0.7X + 0.2X^2)$. While estimation with GEE is carried out, the relations between R_y and $\{X, R_x\}$ is assumed to be linear. Under model misspecification, Figure 1 illustrates that both GEE methods fail to provide an unbiased estimate of the OR despite an increasing sample size. The conditional likelihood still yields unbiased estimates especially with large sample size. Same observation is made in the estimation of α and β as shown in Table 2. Bias and high MSE persist for both methods even with large sample size whereas SD shrinks as sample size increases.

The simulation results indicate all three methods yield unbiased estimators when the model is correctly specified. GEE methods are more efficient than the conditional likelihood. As expected, the optimal GEE is consistently more efficient than the non-optimal GEE regardless of the sample size. On the other hand, for OR estimation, the conditional likelihood

Table 2: Estimation under model misspecification

		Non-optimal GEE		**Optimal GEE**	
N	Statistics	α	β	α	β
500	bias	-0.3435	0.1260	-0.3352	0.1224
	MSE	0.1180	0.0159	0.1124	0.0150
	SD	0.4557	0.1966	0.4483	0.1930
1000	bias	-0.4667	0.1607	-0.4606	0.1578
	MSE	0.2178	0.0258	0.2122	0.0249
	SD	0.3254	0.1397	0.3160	0.1346
2000	bias	-0.4859	0.1737	-0.4747	0.1689
	MSE	0.2361	0.0302	0.2253	0.0285
	SD	0.2343	0.1041	0.2358	0.1042
4000	bias	-0.4497	0.1616	-0.4387	0.1568
	MSE	0.2022	0.0261	0.1924	0.0246
	SD	0.1524	0.0689	0.1487	0.0673

method is more robust under model misspecification meaning that it yields unbiased estimators even when $p(R_y \mid X, R_x)$ is misspecified. In the presence of a strong correlation between X and Y, the GEE estimators exhibit higher efficiency. Conversely, under conditions of weak correlation, the conditional likelihood estimator displays higher efficiency.

F.2 REAL DATA RESULTS

We also applied our proposed methods to analyze data from the KLIPS dataset, which includes information on monthly income for 2511 regular wage earners in 2005 and 2006. The combined monthly income for these two years has approximately 40% missing data. Our objective was to investigate whether past income has a lasting effect on future income. We defined X as the logarithm of monthly income in 2005 and Y as the logarithm of monthly income in 2006. Based on empirical data distributions, we assumed that X, Y, and $X|Y$ are normally distributed. Specifically, we modeled $X|Y$ as $\mathbb{N}(\alpha + \beta Y, \sigma^2)$, where σ^2 was empirically estimated.

Using our nonparametric identification results, we were able to determine α and β without making any additional assumptions. For estimating these parameters, we employed generalized estimating equations (GEEs). Additionally, we used all three methods to estimate $\log(OR)$, where OR represents the odds ratio between the income of the two years. The parameter estimates obtained are summarized in Table 3.

Table 3: Parameter estimates for KLIPS data

	α	β	**log(OR)**
Non-optimal GEE	0.25 (0.289)	0.923 (0.055)	12.621 (0.706)
Optimal GEE	0.348 (0.153)	0.905 (0.029)	12.364 (0.376)
Pseudo-likelihood			10.467 (0.025)

The findings presented above indicate a significant and persistent effect of income. Specifically, high income in the past is strongly predictive of high income in the future, and conversely, low income in the past is predictive of low income in the future. These results provide confirmation that the optimal GEE approach outperforms the non-optimal GEE, particularly in terms of higher efficiency when dealing with continuous variable distributions.

References

James M. Robins. Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in Medicine*, 16:21–37, 1997.

Jiwei Zhao and Jun Shao. Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association*, 110(512):1577–1590, 2015.