Supplementary Materials: A Flexible Framework for Nonparametric Graphical Modeling that Accommodates Machine Learning

Yunhua Xiang [1] **Noah Simon** [1]

Recall that three parameters of interest in our paper respectively are expected conditional covariance $\Psi_1(P)$, expected conditional correlation $\Psi_2(P)$, and scaled expected conditional covariance$\Psi_3(P)$,

$$\Psi_1(P) = \mathrm{E}_P[\mathrm{Cov}_P(Y, Z|X)] = \int (y - \mu_{P,Y}(x))(z - \mu_{P,Z}(x))dP(x), \tag{1}$$

$$\Psi_2(P) = \mathrm{E}_P[\mathrm{Corr}_P(Y, Z|X)] = \int \frac{\mathrm{Cov}_P(Y, Z|X)}{\sqrt{\sigma^2_{P,Y}(x)\sigma^2_{P,Z}(x)}}dP(X), \tag{2}$$

$$\Psi_3(P) = \frac{\mathrm{E}_P[\mathrm{Cov}_P(Y, Z|X)]}{\sqrt{\mathrm{E}_P[\sigma^2_{P,Y}(X)]\,\mathrm{E}_P[\sigma^2_{P,Z}(X)]}} = \frac{\Psi_1(P)}{\sqrt{V_Y(P)V_Z(P)}}. \tag{3}$$

1. Proof of Theorem 1

A simple version of proof of theorem 1 is to directly calculate the discrepancy between $\hat{\Psi}_1$ and $\Psi_1(P)$, which is

$$\begin{aligned}
&\mathbb{P}_n\left[y - \hat{\mu}_Y(x)\right]\left[z - \hat{\mu}_Z(x)\right] - \Psi_1(P) \\
&= \mathbb{P}_n\left\{\left[y - \hat{\mu}_Y(x)\right]\left[z - \hat{\mu}_Z(x)\right] - \hat{\Psi}_1\right\} + \hat{\Psi}_1 - \Psi_1(P) \\
&= [\mathbb{P}_n - P]\hat{D}^{(1)} + P\hat{D}^{(1)} + \hat{\Psi}_1 - \Psi_1(P) \\
&= [\mathbb{P}_n - P]D^{(1)}_P + [\mathbb{P}_n - P][\hat{D}^{(1)} - D^{(1)}_P] + P\left[(\hat{\mu}_Y - \mu_{P,Y})(\hat{\mu}_Z - \mu_{P,Z})\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}[y_i - \mu_{P,Y}(x_i)][z_i - \mu_{P,Z}(x_i)] + o_P(n^{-1/2})
\end{aligned} \tag{4}$$

where $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(x_i)$ and $Pf = \int f dP$. The last equation in (4) holds when Assumption 1 holds. In this case, we have the asymptotic linearity and normality. As an alternative route to prove the theorem we can apply standard semi-parametric tools: We calculate the efficient influence function and then consider a first order asymptotic expansion to show that the theoretically optimal plug-in estimator of $\Psi_1(P)$ has exactly the same form as, so-called one-step estimator. Thus, it will naturally enjoy the good properties including asymptotic consistency and normality.

For $\Psi_2(P)$ however, our simple direct approach will not work, so instead we need to apply those semi-parametric tools.

2. Proof of Theorem 1 by Semi-parametric Theory

For a distribution $P \in \mathcal{M}$, let p denote the density with respect to a dominant measure ν. We define a parametric sub-model $p_\theta(u) := [1 + \theta h(u)]p(u)$ with $h(u) \in L_2(P)$, where $\mathrm{E}\{h(u)\} = 0$, $\sup_u |h(u)| < \infty$, and θ is sufficiently small, such that $p_\theta \geq 0$ and $\int p_\theta(u)d\nu(u) = 1$. Upon inspection we see that this parametric sub-model is centered at P with score

[1]Department of Biostatistics, University of Washington, Seattle, USA. Correspondence to: Yunhua Xiang <xiangyh@uw.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

$s_\theta(u)|_{\theta=0} = \frac{\partial}{\partial\theta}\log p_\theta(u)|_{\theta=0} = h(u)$. In this framework, our statistical functional $\Psi_1(P)$ is called pathwise differentiable at P with efficient influence function $D_P^{(1)}$ (Bickel et al., 1993), if

$$\frac{\partial}{\partial\theta}\Psi_1(P_\theta)(u)\bigg|_{\theta=0} = \int D_P^{(1)}(u)h(u)dP(u) \tag{5}$$

.

Consider observed data consisting of an independent and identically distributed sample of $o = (y, z, x) \in \mathcal{Y} \times \mathcal{Z} \times \mathcal{X}$ drawn from distribution P. Then, the corresponding parametric submodel p_θ can be conditionally decomposed into

$$p_\theta(o) = p_{\theta,h_y}(y|z,x)p_{\theta,h_z}(z|x)p_{\theta,h_x}(x), \tag{6}$$

with score at the origin

$$\begin{aligned} s_\theta(o)|_{\theta=0} &= s_\theta(y|z,x)|_{\theta=0} + s_\theta(z|x)|_{\theta=0} + s_\theta(x)|_{\theta=0} \\ &= h_y(y;z,x) + h_z(z;x) + h_x(x). \end{aligned} \tag{7}$$

For simplicity, we use h_y, h_z, and h_x to represent $h_y(y;z,x)$, $h_z(z;x)$, and $h_x(x)$ respectively. We first show how to obtain the efficient influence function $D_P^{(1)}$ stated in Theorem 1. Let $\psi_P(x)$, $\mu_{P,Y}(x)$, and $\sigma^2_{P,Y}(x)$ denote the conditional covariance $\mathrm{Cov}_P(Y,Z|X)$, conditional mean $\mathrm{E}_P(Y|X=x)$, and conditional variance $\mathrm{Var}_P(Y|X=x)$ evaluated under true model P, while $\psi_\theta(x)$, $\mu_{\theta,Y}(x)$, and $\sigma^2_{\theta,Y}(x)$ are evaluated under sub-model P_θ. The expected conditional covariance $\Psi_1(P)$ in (1) evaluated on $P_\theta|_{\theta=0}$ is

$$\begin{aligned} \frac{\partial}{\partial\theta}\Psi_1(P_\theta)\bigg|_{\theta=0} &= \frac{\partial}{\partial\theta}\int_{\mathcal{X}}\psi_\theta(x)dP_\theta(x)\bigg|_{\theta=0} \\ &= \int_{\mathcal{X}}\left(\frac{\partial}{\partial\theta}\psi_\theta(x)\right)dP_\theta(x)\bigg|_{\theta=0} + \int_{\mathcal{X}}\psi_\theta(x)\frac{\partial}{\partial\theta}\log p_\theta(x)dP_\theta(x)\bigg|_{\theta=0} \\ &= \int_{\mathcal{X}}\left(\frac{\partial}{\partial\theta}\psi_\theta(x)\right)dP_\theta(x)\bigg|_{\theta=0} + \int_{\mathcal{X}}\psi_1(x)h_x dP(x) \end{aligned} \tag{8}$$

Also, we have $\psi_\theta(x) = \mu_{\theta,YZ}(x) - \mu_{\theta,Y}(x)\mu_{\theta,Z}(x)$, where

$$\begin{aligned} \mu_{\theta,YZ}(x) &= \int_{\mathcal{Z}}\int_{\mathcal{Y}} yzp_\theta(y|z,x)p_\theta(z|x)dydz \\ &= \int_{\mathcal{Z}}\int_{\mathcal{Y}} yzp(y|z,x)(1+\theta h_y)p(z|x)(1+\theta h_z)dydz \\ &= \mu_{P,YZ}(x) + \theta\int_{\mathcal{Z}}\int_{\mathcal{Y}} yzp(y,z|x)(h_y+h_z)dydz \\ &\quad + \theta^2\int_{\mathcal{Z}}\int_{\mathcal{Y}} yzp(y,z|x)h_y h_z dydz \\ &= \mu_{P,YZ}(x) + \theta\,\mathrm{E}\left[YZ(h_y+h_z)|X=x\right] + \theta^2\,\mathrm{E}\left[YZh_yh_z|X=x\right]. \end{aligned} \tag{9}$$

and similarly,

$$\begin{aligned} \mu_{\theta,Y}(x) &= \mu_{P,Y}(x) + \theta\,\mathrm{E}\left[Y(h_y+h_z)|X=x\right] + \theta^2\,\mathrm{E}\left[Yh_yh_z|X=x\right], \\ \mu_{\theta,Z}(x) &= \mu_{P,Z}(x) + \theta\,\mathrm{E}\left[Z(h_y+h_z)|X=x\right] + \theta^2\,\mathrm{E}\left[Zh_yh_z|X=x\right]. \end{aligned}$$

We then get that

$$\begin{aligned} \int_{\mathcal{X}}\frac{\partial}{\partial\theta}\psi_\theta(x)dP_\theta(x)\bigg|_{\theta=0} &= \int_{\mathcal{X}}\frac{\partial}{\partial\theta}\mu_{\theta,YZ}(x) - \mu_{\theta,Y}(x)\frac{\partial}{\partial\theta}\mu_{\theta,Z}(x) - \mu_{\theta,Z}(x)\frac{\partial}{\partial\theta}\mu_{\theta,Y}(x)dP_\theta(x)\bigg|_{\theta=0} \\ &= \int_{\mathcal{X}}\mathrm{E}[YZ(h_y+h_z)|X=x] - \mu_{P,Z}(x)\,\mathrm{E}[Y(h_y+h_z)|X=x] \\ &\quad - \mu_{P,Y}(x)\,\mathrm{E}[Z(h_y+h_z)|X=x]dP(x) \\ &= \mathrm{E}\left\{\left[(Y-\mu_{P,Y}(X))(Z-\mu_{P,Z}(X)) - \Psi_1(P)\right](h_y+h_z+h_x)\right\}. \end{aligned} \tag{10}$$

Therefore,

$$\begin{aligned}\frac{\partial}{\partial\theta}\Psi_1(P_\theta)\Big|_{\theta=0} &= \int_{\mathcal{X}}\left(\frac{\partial}{\partial\theta}\psi_\theta(x)\right)dP_\theta(x)\Big|_{\theta=0} + \int_{\mathcal{X}}\Psi_1(x)h_x dP(x)\\ &= \mathrm{E}\left\{[(Y-\mu_{P,Y}(X))(Z-\mu_{P,Z}(X)) - \Psi_1(X)]\,(h_y+h_z+h_x)\right\} + \mathrm{E}[\Psi_1(X)h_x]\\ &= \mathrm{E}\left\{[(Y-\mu_{P,Y}(X))(Z-\mu_{P,Z}(X))]\,(h_y+h_z+h_x)\right\}.\end{aligned} \tag{11}$$

which gives us the efficient influence function $D_P^{(1)}(o) = (y-\mu_{P,Y}(x))(y-\mu_{P,Z}(x)) - \Psi_1(P)$ in Theorem 1. We note that, in above equation, we use the fact that $\int f(y,z)h_x dP(o) = 0$ and $\int g(x)(h_y+h_z)dP(o) = 0$ where $f(y,z)$ is an arbitrary function which does not depend on x and $g(x)$ is an arbitrary function which depends only on x.

Now, we can use the efficient influence function to obtain the so-called "one-step estimator". Consider an asymptotic von-mises expansion of Ψ_1 centered at the true P and evaluated at some $P^* \in \mathcal{M}$ (Fernholz, 2012). Then we have that

$$\begin{aligned}\Psi_1(P^*) - \Psi_1(P) &= (P^*-P)D_{P^*}^{(1)} + R_1(P^*,P)\\ &= -PD_{P^*}^{(1)} + R_1(P^*,P),\end{aligned} \tag{12}$$

where $R(P^*,P)$ is a second order remainder term and we use the fact that $PD_{P^*}^{(1)} = 0$. We can now plug in an estimated distribution $\hat{P}_n$, and use a bit of algebra to show

$$\begin{aligned}\Psi_1(\hat{P}_n) - \Psi_1(P) &= -\mathbb{P}_n\hat{D}^{(1)} + (\mathbb{P}_n - P)\hat{D}^{(1)} + R_1(\hat{P}_n,P)\\ &= -\mathbb{P}_n\hat{D}^{(1)} + (\mathbb{P}_n - P)D_P^{(1)} + (\mathbb{P}_n - P)\left(\hat{D}^{(1)} - D_P^{(1)}\right) + R_1(\hat{P}_n,P),\end{aligned} \tag{13}$$

The second term above is the linear term evaluated at the truth, with mean zero. Ideally, we can find an estimator for Ψ_1 such that this term can dominate the asymptotic performance of $\Psi_1(\hat{P}_n)$. The third and fourth one are respectively an empirical process term and second-order remainder term, which can be shown to be negligible under certain conditions on $\hat{P}_n$. That is, they both converge to 0 faster than the linear term as $n\to\infty$. However, we see that the term $\mathbb{P}_n\hat{D}^{(1)}$ is the source of the irregular behavior of $\Psi_1(\hat{P}_n)$ and can often cause non-ignorable bias. Hence, this expansion motivates us to find a proper way to cancel the effects of $\mathbb{P}_n\hat{D}^{(1)}$ and give the proposed one-step estimator for $\Psi_1(P)$

$$\begin{aligned}\hat{\Psi}_{1,onestep} &= \Psi_1(\hat{P}_n) + \mathbb{P}_n\hat{D}^{(1)}\\ &= \frac{1}{n}\sum_{i=1}^n (y_i - \hat{\mu}_Y(x_i))(y_i - \hat{\mu}_Z(x_i)),\end{aligned} \tag{14}$$

which is coincidentally the same as the proposed theoretically optimal plug-in estimator $\hat{\Psi}_1$ in our paper. Therefore, according to (13), we can also obtain the asymptotic linearity of $\hat{\Psi}_1$:

$$\hat{\Psi}_1 - \Psi_1(P) = \frac{1}{n}\sum_{i=1}^n D_P^{(1)}(o_i) + o_P(n^{-1/2}),$$

as long as the empirical process $(\mathbb{P}_n - P)\left(\hat{D}^{(1)} - D_P^{(1)}\right)$ and the second-order remainder term $R_1(\hat{P}_n,P)$ are negligible. By Assumption 1 we have that $(\mathbb{P}_n - P)\left(\hat{D}^{(1)} - D_P^{(1)}\right) = o_P(n^{-1/2})$. Thus, we only need to prove $R_1(\hat{P}_n,P) = o_P(n^{-1/2})$. For any $P^* \in \mathcal{M}$, the remainder is

$$\begin{aligned}R_1(P^*,P) &= \Psi_1(P^*) - \Psi_1(P) + PD_{P^*}^{(1)}\\ &= P\left\{(Y-\mu_{P^*,Y}(X))(Z-\mu_{P^*,Z}(X))\right\} - P\left\{(Y-\mu_{P,Y}(X))(Z-\mu_{P,Z}(X))\right\}\\ &= P\left\{(\mu_{P^*,Y}(X)-\mu_{P,Y}(X))\,(\mu_{P^*,Z}(X)-\mu_{P,Z}(X))\right\}.\end{aligned} \tag{15}$$

Hence, as long as $\mu_{P^*,Y}(X) - \mu_{P,Y}(X)$ and $\mu_{P^*,Z}(X) - \mu_{P,Z}(X)$ both converge to zero at $o_P(n^{-1/4})$, we have $R_1(P^*,P) = o_P(n^{-1/2})$. That is to say, under Assumptions 1, the asymptotic linearity of $\hat{\Psi}_1$ holds. By the central limit theorem, we can further derive the asymptotic normality of $\hat{\Psi}_1$, i.e.

$$\sqrt{n}[\hat{\Psi}_1 - \Psi_1(P)] \to_d N[0,\sigma_1^2(P)], \tag{16}$$

where $\sigma_1^2(P) = \int [D_P^{(1)}(o)]^2 dP(o)$. This completes the proof of Theorem 1

3. Proof of Theorem 3

As in (12), we can also show that the naive plug-in estimator $\hat{\Psi}_{2,naive}$ is asymptotically biased, which can be corrected by adding the irregular bias. Let $\phi_\theta(x) = \mathrm{Cor}(Y, Z|X)$ under P_θ and $\phi(x) = \phi_\theta|_{\theta=0}$. Then, $\phi_\theta(x) = \frac{\psi_\theta(x)}{g_\theta(x)}$, where $g_\theta(x) = \sqrt{\sigma^2_{\theta,Y}(x)\sigma^2_{\theta,Z}(x)}$. The conditional variance under P_θ can be expanded as follows,

$$\begin{aligned}
\sigma^2_{\theta,Y}(x) &= \mu_{\theta,Y^2}(x) - \mu^2_{\theta,Y}(x) \\
&= \mu_{Y^2}(x) + \theta\,\mathrm{E}\left[Y^2(h_y + h_z)|X = x\right] + \theta^2\,\mathrm{E}\left[Y^2 h_y h_z|X = x\right] \\
&\quad - \left\{\mu_{P,Y}(x) + \theta\,\mathrm{E}\left[Y(h_y + h_z)|X = x\right] + \theta^2\,\mathrm{E}\left[Y h_y h_z|X = x\right]\right\}^2 \\
&= \sigma^2_{P,Y}(x) + \theta\,\mathrm{Cov}(Y, Y(h_y + h_z)|X = x) + \theta^2\,\mathrm{Cov}(Y, Y h_y h_z|X = x) \\
&\quad - \mu_{P,Y}(x)\left\{\theta\,\mathrm{E}[Y(h_y + h_z)|X = x] + \theta^2\,\mathrm{E}[Y h_y h_z|X = x]\right\} \\
&\quad - \left\{\theta\,\mathrm{E}[Y(h_y + h_z)|X = x] + \theta^2\,\mathrm{E}[Y h_y h_z|X = x]\right\}^2.
\end{aligned} \tag{17}$$

$$\Rightarrow \frac{\partial}{\partial\theta}\sigma^2_{\theta,Y}(x)\bigg|_{\theta=0} = \mathrm{Cov}(Y, Y(h_y + h_z)|X = x) - \mu_{P,Y}(x)\,\mathrm{E}[Y(h_y + h_z)|X = x] \tag{18}$$

$$\Rightarrow \frac{\partial}{\partial\theta}\sigma^2_{\theta,Z}(x)\bigg|_{\theta=0} = \mathrm{Cov}(Z, Z(h_y + h_z)|X = x) - \mu_{P,Z}(x)\,\mathrm{E}[Z(h_y + h_z)|X = x]. \tag{19}$$

Thus, the derivative of $\Psi_2(P_\theta)$ with respect to θ at $\theta = 0$ is

$$\begin{aligned}
\frac{\partial}{\partial\theta}\Psi_2(P_\theta)\bigg|_{\theta=0} &= \int_{\mathcal{X}} \left(\frac{\partial}{\partial\theta}\phi_\theta(x)\right) dP_\theta(x)\bigg|_{\theta=0} + \int_{\mathcal{X}} \phi(x) h_x dP(x) \\
&= \int_{\mathcal{X}} \frac{\psi'_\theta(x) g_\theta(x) - \psi_\theta(x) g'_\theta(x)}{g^2_\theta(x)} dP_\theta(x)\bigg|_{\theta=0} + \mathrm{E}[\phi(X) h_x],
\end{aligned} \tag{20}$$

where

$$\begin{aligned}
\psi'_\theta(x)|_{\theta=0} &= \mathrm{E}[Y Z(h_y + h_z)|X = x] - \mu_{P,Z}(x)\,\mathrm{E}[Y(h_y + h_z)|X = x] \\
&\quad - \mu_{P,Y}(x)\,\mathrm{E}[Z(h_y + h_z)|X = x],
\end{aligned} \tag{21}$$

and

$$\begin{aligned}
g'_\theta(x) &= \frac{1}{2g(x)}\left\{\mathrm{E}[Y^2(h_y + h_z)|X] - 2\mu_{P,Y}(x)\,\mathrm{E}[Y(h_y + h_z)|X]\right\}\sigma^2_{P,Z}(x) \\
&\quad + \frac{1}{2g(x)}\left\{\mathrm{E}[Z^2(h_y + h_z)|X] - 2\mu_{P,Z}(x)\,\mathrm{E}[Z(h_y + h_z)|X]\right\}\sigma^2_{P,Y}(x).
\end{aligned} \tag{22}$$

Plugging in (21) and (22), (20) becomes

$$\begin{aligned}
&\int_{\mathcal{X}} \frac{\psi'_\theta(x) g_\theta(x) - \psi_\theta(x) g'_\theta(x)}{g^2_\theta(x)} dP_\theta(x)\bigg|_{\theta=0} + \mathrm{E}[\phi(X) h_x] \\
&= \mathrm{E}\left\{\left[\frac{(Y - \mu_{P,Y}(X))(Z - \mu_{P,Z}(X))}{g(X)} - \phi(X)\right](h_y + h_z + h_x)\right\} \\
&\quad - \mathrm{E}\left\{\phi(X)\left[\frac{(Z - \mu_{P,Z}(X))^2}{2\sigma^2_{P,Z}(X)} + \frac{(Y - \mu_{P,Y}(X))^2}{2\sigma^2_{P,Y}(X)} - 1\right](h_y + h_z + h_x)\right\} + \mathrm{E}[\phi(X) h_x] \\
&= \mathrm{E}\left\{\left[\frac{(Y - \mu_{P,Y}(X))(Z - \mu_{P,Z}(X))}{g(X)} - \phi(X)\left(\frac{(Z - \mu_{P,Z}(X))^2}{2\sigma^2_{P,Z}(X)} + \frac{(Y - \mu_{P,Y}(X))^2}{2\sigma^2_{P,Y}(X)} - 1\right) - \Psi_2(P)\right](h_y + h_z + h_x)\right\}.
\end{aligned} \tag{23}$$

Therefore, the efficient influence function of $\Psi_2(P)$ is

$$D^2_P(o) = \frac{(y - \mu_{P,Y}(x))(z - \mu_{P,Z}(x))}{g(x)} - \phi(x)\left(\frac{(z - \mu_{P,Z}(x))^2}{2\sigma^2_{P,Z}(x)} + \frac{(y - \mu_{P,Y}(x))^2}{2\sigma^2_{P,Y}(x)} - 1\right) - \Psi_2(P). \tag{24}$$

Thus, the one-step estimator of $\Psi_2(P)$ according to (13) is

$$\begin{aligned}\hat{\Psi}_2 &= \Psi_2(\hat{P}_n) + \mathbb{P}_n \hat{D}^{(2)} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{ \frac{(y_i - \hat{\mu}_Y(x_i))(z_i - \hat{\mu}_Z(x_i))}{\sqrt{\hat{\sigma}^2_Y(x_i)\hat{\sigma}^2_Z(x_i)}} \right. \\ &\quad \left. - \frac{\hat{\mu}_{YZ}(x_i) - \hat{\mu}_Y(x_i)\hat{\mu}_Z(x_i)}{\sqrt{\hat{\sigma}^2_Y(x_i)\hat{\sigma}^2_Z(x_i)}} \left[\frac{(y_i - \hat{\mu}_Y(x_i))^2}{2\hat{\sigma}^2_Y(x_i)} + \frac{(z_i - \hat{\mu}_Z(x_i))^2}{2\hat{\sigma}^2_Z(x_i)} - 1 \right] \right\}.\end{aligned} \tag{25}$$

The second-order remainder of $\Psi_2(P^*)$ is

$$\begin{aligned}R_2(P^*, P) &= \Psi_2(P^*) - \Psi_2(P) + PD^{(2)}_P \\ &= P\left\{ \frac{(\mu_{P^*,Y}(X) - \mu_{P,Y}(X))\,(\mu_{P^*,Z}(X) - \mu_{P,Z}(X))}{g_{P^*}(X)} \right\} \\ &\quad - P\left\{ \frac{\mathrm{Corr}_{P^*}(Y,Z|X)}{2\sigma^2_{P^*,Y}(X)} \left(\mu_{P^*,Y}(X) - \mu_{P,Y}(X)\right)^2 \right\} \\ &\quad - P\left\{ \frac{\mathrm{Corr}_{P^*}(Y,Z|X)}{2\sigma^2_{P^*,Z}(X)} \left(\mu_{P^*,Z}(X) - \mu_{P,Z}(X)\right)^2 \right\} \\ &\quad + P\left\{ \frac{\mathrm{Cov}_{P^*}(Y,Z|X) - \mathrm{Cov}_P(Y,Z|X)}{g_{P^*}(X)} \left(\frac{\sigma^2_{P^*,Y}(X) - \sigma^2_{P,Y}(X)}{2\sigma^2_{P^*,Y}(X)} + \frac{\sigma^2_{P^*,Z}(X) - \sigma^2_{P,Z}(X)}{2\sigma^2_{P^*,Z}(X)} \right) \right\} \\ &\quad - P\left\{ f_1(X)(\sigma_{P^*,Z}(X) - \sigma_{P,Z}(X))^2 - f_2(X)(\sigma_{P^*,Y}(X) - \sigma_{P,Y}(X))(\sigma_{P^*,Z}(X) - \sigma_{P,Z}(X)) \right. \\ &\quad \left. + f_3(X)(\sigma_{P^*,Y}(X) - \sigma_{P,Y}(X))^2 \right\},\end{aligned} \tag{26}$$

where $\{f_i\}_{i=1}^3$ are some functions depending only on X. Hence, to make $R_2(\hat{P}_n, P)$ converges to zero at $o_P(n^{-1/2})$, we have to guarantee that every item in (26) converges to zero at $o_P(n^{-1/2})$, which includes $\int \left(\mathrm{Cov}_{\hat{P}_n}(Y,Z|x) - \mathrm{Cov}_P(Y,Z|x)\right)\left(\sigma^2_{P^*,Y}(x) - \sigma^2_{P,Y}(x)\right) dP(x)$, $\int \left(\hat{\sigma}_Y(x) - \sigma_{P,Y}(x)\right)^2 dP(x)$, $\int \left(\hat{\sigma}_Z(x) - \sigma_{P,Z}(x)\right)^2 dP(x)$. Then, we have the asymptotic linearity of $\Psi_2(P)$ by (13),

$$\hat{\Psi}_2 - \Psi_2(P) = \frac{1}{n}\sum_{i=1}^{n} D^{(2)}_P(o_i) + o_P(n^{-1/2}),$$

and the asymptotic normality

$$\sqrt{n}[\hat{\Psi}_2 - \Psi_2(P)] \to_d N[0, \sigma^2_2(P)], \tag{27}$$

where $\sigma^2_2(P) = \int [D^{(2)}_P(o)]^2 dP(o)$. This completes the proof of Theorem 3.

3.1. Proof of Theorem 2

Before proving Theorem 2, we first show that $-1 \le \Psi_3(P) \le 1$ for all P. By applying Cauchy-Schwartz and Jensen's inequality, we get that

$$\mathrm{Cov}^2(Y,Z|X) \le \mathrm{Var}(Y|X)\,\mathrm{Var}(Z|X)$$

$$|\,\mathrm{E}[\mathrm{Cov}(Y,Z|X)]| \le \mathrm{E}\sqrt{\mathrm{Var}(Y|X)\,\mathrm{Var}(Z|X)} \le \sqrt{\mathrm{E}[\mathrm{Var}(Y|X)\,\mathrm{Var}(Z|X)]}$$

$$|\Psi_3(P)| = \left| \frac{\mathrm{E}[\mathrm{Cov}(Y,Z|X)]}{\sqrt{\mathrm{E}[\mathrm{Var}(Y|X)\,\mathrm{Var}(Z|X)]}} \right| \le 1,$$

which has the same range as the correlation.

The efficient influence function of $\Psi_3(P)$ can be easily derived from what we have developed for $\Psi_1(P)$ by delta method (Sobel, 1982). Recall that expected conditional covariance has efficient influence function $D^{(1)}_P(o)$, So the efficient influence

function for the expected conditional variance $V_Y(P)$ is $D_P^{V_Y}(o) = (y - \mu_{P,Y}(x))^2$. Let $g(u, v, w) = \frac{u}{\sqrt{vw}}$. Then we know that $\Psi_3(P) = g(\Psi_1(P), V_Y(P), V_Z(P))$ has efficient influence function

$$\begin{aligned} D_P^{(3)}(o) &= \nabla g(\Psi_1(P), V_Y(P), V_Z(P)) \times (D_P^{(1)}(o), D_P^{V_Y}(o), D_P^{V_Z}(o))^T \\ &= \frac{(y - \mu_{P,Y}(x))(z - \mu_{P,Z}(x))}{\sqrt{V_Y(P)V_Z(P)}} - \Psi_3(P)\left[\frac{(y - \mu_{P,Y}(x))^2}{2V_Y(P)} + \frac{(z - \mu_{P,Z}(x))^2}{2V_Z(P)}\right]. \end{aligned} \tag{28}$$

Thus, the one-step estimator is exactly the same as our theoretically optimal plug-in estimator, becausel of $\mathbb{P}_n\hat{D}^{(3)} = \Psi_3(\hat{P}_n) - \Psi_3(\hat{P}_n)\left[\frac{\frac{1}{n}\sum(y_i - \mu_Y(x_i))^2}{2V_Y(\hat{P}_n)} + \frac{\frac{1}{n}\sum(z_i - \mu_Z(x_i))^2}{2V_Z(\hat{P}_n)}\right] = 0$. Then the second order remainder of $\Psi_3(P^*)$ is

$$\begin{aligned} R_3(P^*, P) &= \Psi_3(P^*) - \Psi_3(P) + PD_P^{(3)} \\ &= \frac{P[(\mu_{P^*,Y}(X) - \mu_{P,Y}(X))\,(\mu_{P^*,Z}(X) - \mu_{P,Z}(X))}{\sqrt{V_Y(P^*)V_Z(P^*)}} \\ &\quad - \frac{\Psi_3(P^*)}{2}\left[\frac{P(\mu_{P^*,Y}(X) - \mu_{P,Y}(X))^2}{V_Y(P^*)} + \frac{P(\mu_{P^*,Z}(X) - \mu_{P,Z}(X))^2}{V_Z(P^*)}\right] \\ &\quad + G(P^*, P) \end{aligned} \tag{29}$$

where

$$\begin{aligned} G(P^*, P) = \frac{\Psi_1(P^*) - \Psi_1(P)}{2\sqrt{V_Y(P^*)V_Z(P^*)}}&\left[\frac{(V_Y(P^*) - V_Y(P))^2}{V_Y(P^*)} + \frac{(V_Z(P^*) - V_Z(P))^2}{V_Z(P^*)}\right] \\ - \frac{\Psi_1(P)}{\sqrt{V_Y(P^*)V_Z(P^*)}}&\left[\frac{[V_Y(P)V_Z(P) - V_Y(P^*)V_Z(P^*)]^2}{\sqrt{V_Y(P^*)V_Z(P^*)V_Y(P)V_Z(P)}}\right. \\ &+ \frac{[\sqrt{V_Z(P^*)} - \sqrt{V_Z(P)}]^2}{2V_Z(P^*)} + \frac{[\sqrt{V_Y(P^*)} - \sqrt{V_Y(P)}]^2}{2V_Y(P^*)} \\ &\left. - \frac{[\sqrt{V_Y(P^*)} - \sqrt{V_Y(P)}]\sqrt{V_Z(P^*)} - \sqrt{V_Z(P)}]}{2\sqrt{V_Y(P^*)V_Z(P^*)}}\right] \end{aligned}$$

Under Assumption 1, we have known that $\Psi_1(P^*) - \Psi_1(P) = o_P(n^{-1/2})$. Thus, $G(P^*, P) = o_P(n^{-1})$ and thus, $R_3(P^*, P) = o_P(n^{-1/2})$ is negligible. We can then obtain the asymptotical linearity and nonparametric efficiency of $\hat{\Psi}_3$ as in Theorem 2.

4. Additional experiments for asymptotic performance

For $\Psi_1(P)$, $\Psi_2(P)$, $\Psi_3(P)$, we compare the efficient estimators proposed in paper, with their corresponding naive estimators (which should theoretically not be rate optimal):

$$\hat{\Psi}_{1,naive} = \frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_{YZ}(x) - \hat{\mu}_Y(x)\hat{\mu}_Z(x)] \tag{30}$$

$$\hat{\Psi}_{2,naive} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\mu}_{YZ}(x_i) - \hat{\mu}_Y(x_i)\hat{\mu}_Z(x_i)}{\sqrt{(\hat{\mu}_{Y^2}(x_i) - \hat{\mu}_Y^2(x_i))\,(\hat{\mu}_{Z^2}(x_i) - \hat{\mu}_Z^2(x_i))}} \tag{31}$$

$$\hat{\Psi}_{3,naive}\frac{\frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_{YZ}(x) - \hat{\mu}_Y(x)\hat{\mu}_Z(x)]}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_{Y^2}(x_i) - \hat{\mu}_Y^2(x_i)) \times \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_{Z^2}(x_i) - \hat{\mu}_Z^2(x_i))}}. \tag{32}$$

4.1. Low-dimensional cases

We modify the setting of low-dimensional example in our paper slightly, by changing the underlying covariance structure of errors of Y and Z. In this case, we let

$$\vec{e}|X = (e_y, e_z)^T|X \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.5 + \frac{x}{4} \\ -0.5 + \frac{x}{4} & 1 \end{pmatrix}\right]. \tag{33}$$

The true value of $\Psi_1(P)$ is 0.25. The results are shown in Figure 1: the naive estimator again does not have a bias converging to zero at $o_P(n^{-1/2})$ and we cannot obtain a valid confidence interval by bootstrapping. In fact, we can notice that bootstrap-based methods indeed fails quite spectacularly.

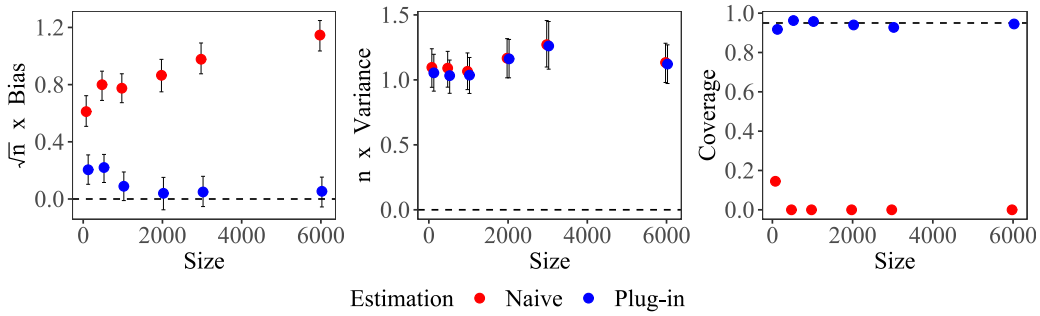

Figure 1. Low dimensional setting: Empirical $\sqrt{n}-$scaled bias (left), empirical n-scaled variance (center) and empirical coverage of 95% confidence interval (right) of the theoretically optimal plug-in estimator (blue) and the naive estimators (red) of the scaled expected conditional covariance $\Psi_1(P)$. Conditional mean is estimated by local polynomial regression.

We also use the same pattern in low-dimensional setting described in our paper to evaluate the theoretically optimal plug-in & naive estimator of $\Psi_3(P)$. Figure **??** shows the results. The empirical $\sqrt{n}$-scaled bias of our theoretically optimal estimator $\hat{\Psi}_3$ goes toward zero which this is not the case for the naive estimator. The empirical variance of both methods stabilizes when scaled by n and the confidence interval of our optimal plug-in estimators converges to the nominal 95% as sample size increases. As expected, due to excess bias, the bootstrap interval based on the "naive" estimators performs poorly (with coverage actually converging to 0)

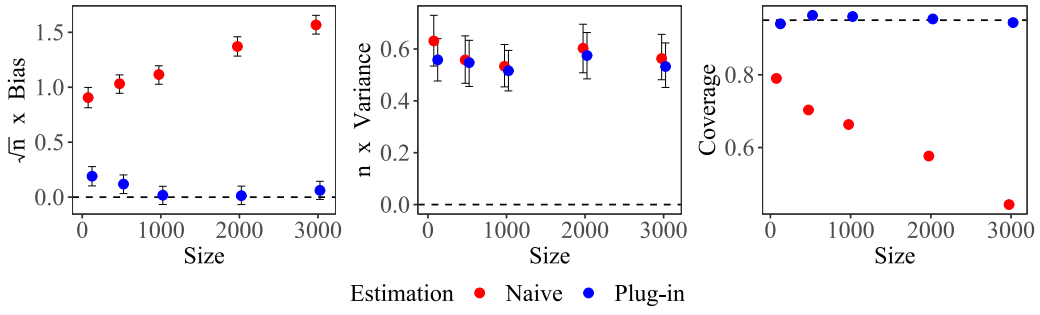

Figure 2. Low dimensional setting: Empirical $\sqrt{n}-$scaled bias (left), empirical n-scaled variance (center) and empirical coverage of 95% confidence interval (right) of the theoretically optimal estimator (blue) and the naive estimators (red) of the scaled expected conditional covariance $\Psi_3(P)$. Conditional mean is estimated by local polynomial regression.

4.2. Moderate dimensional cases

In this setting, we generate the data from following mechanism.

$$Y = f_1(x_1, ..., x_8) + e_y, \;\; Z = f_2(x_1, ..., x_8) + e_z, \tag{34}$$

where $X \sim N(0, I_8)$ and $\vec{e} = (e_y, e_z)^T \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}\right]$. Here the true value of expected conditional covariance is also $\Psi_1(P) = -0.5$. For each sample size $n \in \{300, 500, 2000, 4000, 6000, 8000, 10000\}$, we generated 400 datasets. Gradient boosting were used to estimate the conditional means $\mu_Y(x)$ and $\mu_Z(x)$ where hyper-parameters (number of trees, minimal node size and fraction of observations to sample) are tuned by a 5-fold cross validation. Since bootstrap-based approach fails to build the confidence interval and is computationally expensive. Here, we just include the Wald-type confidence interval of the optimal plug-in estimator. The results look similar to low-dimensional cases, see Figure 3. As n increases, $\sqrt{n}$-scaled bias of the optimal plug-in estimator tends to zero while that of the naive estimator diverges. The variances go to a positive constant and the empirical coverage obtained from asymptotic normality also works. In this setting, we may notice that bias of the optimal plug-in estimator converges to 0 more slowly but still gives reasonable interval estimates.

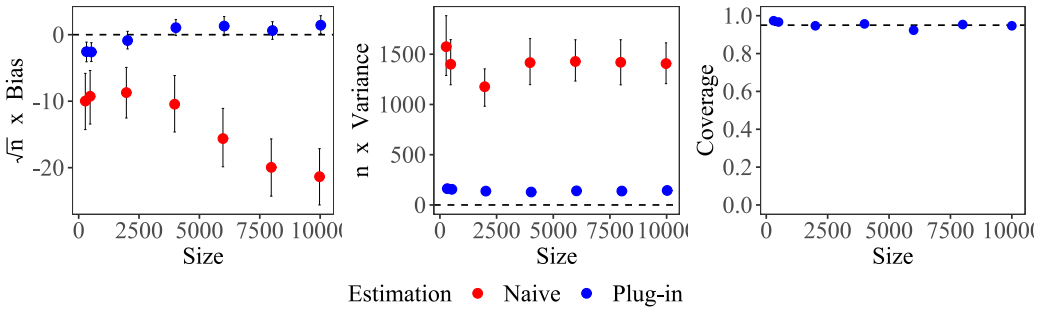

Figure 3. Moderate dimensional setting: Empirical $\sqrt{n}-$scaled bias (left), empirical n-scaled variance (center) and empirical coverage of 95% confidence interval (right) of the theoretically optimal plug-in estimator (blue) and the naive estimators (red) of the scaled expected conditional covariance $\Psi_1(P)$. Conditional mean is estimated by gradient boosting.

4.3. High-dimensional cases

We use the same setting with high-dimensional features to evaluate the performance of the scaled expected conditional covariance Ψ_3. The true parameter is $\Psi_3(P) = -0.5$. We generate random datasets of size $n \in \{500, 1000, 2000, 3000, 4000\}$ and estimate Ψ_3. The Lasso was used to estimating the conditional means $\mu_Y(x)$ and $\mu_Z(x)$ where the regularization parameter was tuned by a 5-fold cross validation. Again, the results are in-line with our theory: We see good performance for $\hat{\Psi}_3$ and poor performance for the naive estimator.

5. Real Data Analysis: Network Recovery of Boston Housing Data

We evaluate our approach on the Boston housing data (Harrison Jr & Rubinfeld, 1978) by analyzing the network structure of features that may potentially impact house price. This dataset contains information collected by the U.S Census Service concerning housing in different areas of Boston Mass. There are 506 observations and each observation is based on a single town, with information on median home value (MEDV). In addition, it provides the four types of attributes which may be potential predictors to the price of house. The first type consists of neighborhood feature: % of lower socio-economic status (LSTAT); % of residential land zoned for lots larger than 25,000 square feet (ZN); % of black residents in the population (B); per capita crime rate by town (CRIM); % of non-retail business acres per town (INDUS); the full value property tax rate (TAX); the pupil-teacher ratio by school district (PTRATIO); Charles River dummy variable (CHAS). The second type is the house structural features: the average number of rooms per dwelling (RM) and % of owner-occupied units built prior to 1940 (AGE); The third one consists of accessibility features: index of accessibility to radial highways (RAD) and the weighted distances to five Boston employment centers (DIS). The final type is about air pollution, which only includes the

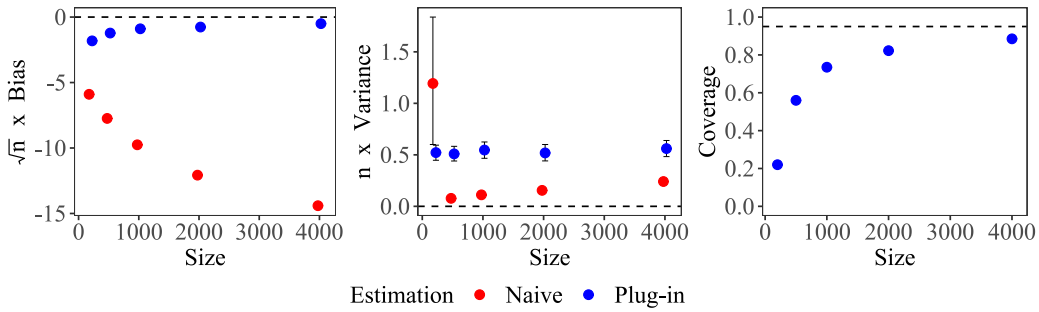

Figure 4. Empirical $\sqrt{n}-$scaled bias (left), empirical n-scaled variance (center) and empirical coverage of 95% confidence interval (right) of the plug-in estimator (blue) and the naive estimators (red) of the scaled expected conditional covariance $\Psi_3(P)$. Conditional mean is estimated by Lasso.

nitric oxides concentration (NOX).

Here, we consider Gaussian graphical model (GMM) and the scaled expected conditional covariance Ψ_3 to build a network of 14 attributes. For Ψ_3, we estimate the conditional mean using random forests and we obtain p-values according to our asymptotic Gaussian limits as discussed in Theorem 2. For GMM, we use the bootstrap to build confidence intervals. In addition, we also use the value of $\hat{\Psi}_3$ and the corresponding entry of the estimated precision matrix to represent the strength of association.

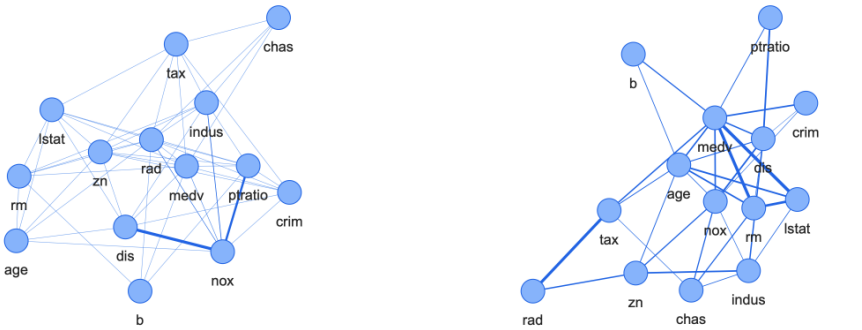

Figure 5. Network constructed using GGM (left) and Ψ_3 (right). P-values are obtained to identify edges. Width of edges represents the corresponding entry in the precision matrix (left) or the value of Ψ_3 (right).

We display the results in Figure 5. The network constructed by the scaled expected conditional covariance Ψ_3 shows that median house value (MEDV) is strongly connected with neighborhood and structural characteristics, such as the number of room (RM), weighted distances to employment centres (DIS), % of lower socio-economic status residents(LSTAT), crime rate (CRIM) and property-tax rate(TAX). This is similar to the findings of (Bi et al., 2003) and (Williamson et al., 2017) where these attributes were also marked as important. In particularly, average number of room and proportion of lower socio-economic status, which were previously found as the most important feature, also have the strongest conditional association with the price.

In this example, estimating the network using GGM gives very different results. The graph structure is much less parsimonious. This is to be expected under model-misspecification: It is likely that the *true* precision matrix derived from complicated non-Gaussian data is quite dense; it is unfortunately just a meaningless measure in such a case. In addition, edges connected to median price (MEDV) do not agree with previous published studies.

References

Bi, J., Bennett, K., Embrechts, M., Breneman, C., and Song, M. Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, 3(Mar):1229–1243, 2003.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore, 1993.

Fernholz, L. T. *Von Mises calculus for statistical functionals*, volume 19. Springer Science & Business Media, 2012.

Harrison Jr, D. and Rubinfeld, D. L. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978.

Sobel, M. E. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological methodology*, 13:290–312, 1982.

Williamson, B. D., Gilbert, P. B., Simon, N., and Carone, M. Nonparametric variable importance assessment using machine learning techniques. 2017.