Supplement to "On the Identifiability and Estimation of Functional Causal Models in the Presence of Outcome-Dependent Selection"

This supplementary material provides the proofs and some details which are omitted in the submitted paper. The equation numbers in this material are consistent with those in the paper.

S1. An Illustration on the Effect of Output-Dependent Selection Bias on the Estimation of Functional Causal Models

If the selection depends solely on the effect, as depicted in Figure 1(c), then $p_{Y|X,S=1} \neq p_{Y|X}$, and the selection bias, if not corrected, will mislead inference. Consider, for example, a standard assumption in functional causal modeling that the effect Y is a function of the cause variable X and an noise variable E that is independent of X. Suppose this assumption holds in the population. With the outcome-dependent selection, X and E are typically not independent in the selected sample, as they are typically not independent conditional on S (which is a descendant of a collider between X and E, i.e., Y). Furthermore, even if one fits a regression model on selected sample, the estimated residual (which is usually different from the true noise term in the causal process) is usually not independent from X; we will get back to this issue in Section 4.1. An illustration is given in Figure 6, where data are generated from a linear additive noise model $Y = X + E$, with selection on Y. As one can see from the chart on the right, without correction, the estimated noise and the cause are not independent.

Theoretical results on this issue are given in Corollary 3, which, roughly speaking, states that if the OSB is not corrected, one cannot fit a restricted functional causal model with independent noise on the data.

S2. Proof of Theorem 1

Proof 1 *This proof is an adaptation of the proof of Theorems 1 and 8 in (Zhang & Hyvärinen, 2009). From (5) one can see* $\frac{\partial^2 \log p_{Z\tilde{E}}}{\partial z \partial \tilde{e}} = \tilde{\eta}_1'' \frac{\partial t}{\partial z} - \eta_2'' h' \frac{\partial e}{\partial z} - \eta_2' h'' \frac{\partial t}{\partial z} = \tilde{\eta}_1'' h_1' - \eta_2'' h' + \eta_2'' h'^2 h_1' - \eta_2' h'' h_1'$. *Eq. 6 then implies* $\tilde{\eta}_1'' h_1' - \eta_2'' h' + \eta_2'' h'^2 h_1' - \eta_2' h'' h_1' = 0$. *From this equation one can see that* $h_1' = 0$ *implies* $\eta_2'' h' = 0$. *Consequently, the points which satisfy* $\eta_2'' h' \neq 0$ *also make* $h_1' \neq 0$. *For such points, dividing both sides of this equation by* $h_1' \eta_2'' h'$ *finally leads to (8). Furthermore, since* h_1 *is a functions of* z_2 *and does not depend on* e_1, *we have* $\partial\Big(\frac{1}{h_1'}\Big)\Big/\partial e_1 = 0$. *According to (8), we have* $\partial\Big(\frac{\tilde{\eta}_1'' + \eta_2'' h'^2 - \eta_2' h''}{\eta_2'' h'}\Big)/\partial e_1 = 0$, *which gives* $2\eta_2''^2 h'^2 h'' - \eta_2' \eta_2'' h' h''' + \eta_2'' \tilde{\eta}_1''' h' - \eta_2' \eta_2''' h'^2 h'' + \eta_2' \eta_2'' h''^2 + \eta_2''' \tilde{\eta}_1'' h'^2 - \eta_2'' \tilde{\eta}_1'' h'' = 0$. *For the points satisfying* $\eta_2'' h' \neq 0$, *we divide both sides of the above equation by* $\eta_2'' h'$. *After some simplifications, (7) is obtained.*

The next step is to solve the partial differential equation (7). Compare this equation with (4) in (Zhang & Hyvärinen, 2009), one can see that the former is obtained by substituting $\tilde{\eta}_1$ *for* η_1 *in the latter, whose solution was given in Theorem 8 in (Zhang & Hyvärinen, 2009); see Table 1 there. The solution to (7), given in Table 1 directly follows from Table 1 in (Zhang & Hyvärinen, 2009). The only difference is that here we omit the constraint on* $\tilde{\eta}_1$ *and that on* h_1. *Q.E.D.*

S3. Proof of Theorem 2

Proof 2 *Because* $y = f_1(x) + e_1 = f_2(x) + e_2$. *We have* $e_1 = f_2(x) - f_1(x) + e_2$. *As seen from the LHS of (13),* J_{AN} *is determined by functions* $\beta_r(y)$, $p_X^{(1)}(x)$, *and* $p_{E_1}(e_1)$; *by the change of variables, it can be repented as a function of* x *and* e_2, *and we can find the partial derivative:*

$$\begin{aligned}
\frac{\partial^2 J_{AN}}{\partial x \partial e_2} &= \frac{\partial^2}{\partial e_2 \partial x}\Big(- \log \beta_r(y) + \eta_{X_1}(x) + \eta_{E_1}(e_1)\Big) \\
&= \frac{\partial}{\partial e_2}\Big(- l'_\beta(y)\frac{\partial y}{\partial x} + \eta'_{X_1}(x) + \eta'_{E_1}(e_1)\frac{\partial e_1}{\partial x} \\
&= \frac{\partial}{\partial e_2}\Big(- l'_\beta(y) f^{(2)\prime} + \eta'_{X_1}(x) + \eta'_{E_1}(e_1)\cdot \\
&\qquad\qquad (f^{(2)\prime} - f^{(1)\prime})\Big) \\
&= -l''_\beta(y) f^{(2)\prime} + \eta''_{E_1}(f^{(2)\prime} - f^{(1)\prime}).
\end{aligned}$$

As assumed, $\eta''_{E_1}(e_1) f^{(1)'} = 0$ *only at finite points. In the range where* $f^{(2)\prime} \neq 0$, *setting* $\frac{\partial^2 J_{AN}}{\partial x \partial e_2}$ *to zero gives*

$$-l''_\beta(y) = \eta''_{E_1} \cdot \Big(\frac{f^{(1)\prime}}{f^{(2)\prime}} - 1\Big).$$

That is,

$$-l''_\beta(f_1(x) + e_1) = \eta''_{E_1} \cdot \Big(\frac{f^{(1)\prime}}{f^{(2)\prime}} - 1\Big). \qquad (24)$$

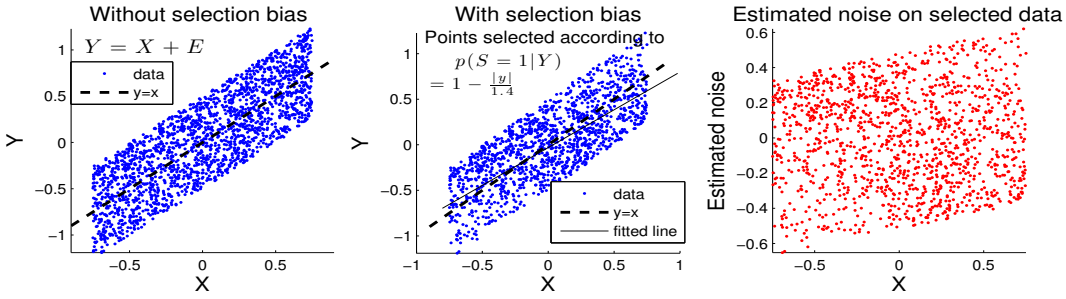


Figure 6: Illustration of the effect of outcome-dependent selection. The data were generated from a linear additive noise model $Y = X + E$ with selection on Y. Left: The data distribution on the whole population (before applying selection bias). Middle: The distribution of selected data (with selection bias). Right: The distribution of the estimated noise and cause on the selected data: they are clearly not independent.

First note that if $\eta''_{E_1} \neq 0$, $\frac{f^{(1)\prime}}{f^{(2)\prime}}$ can be written as a function of $f^{(1)}(x)$; otherwise, there will exist two points x_1 and x_2 corresponding to the same one value of $f^{(1)}(x)$ but different values of $\frac{f^{(1)\prime}}{f^{(2)\prime}}$, and then (24) cannot hold on both x_1 and x_2, leading to a contradiction.

Since $\frac{f^{(1)\prime}}{f^{(2)\prime}} - 1$ as a function of $f^{(1)}(x)$, we let $\frac{f^{(1)\prime}}{f^{(2)\prime}} - 1 = f_c(f^{(1)}(x))$. Further note that l''_β is a function of $(f^{(1)}(x) + e_1)$ and that η''_{E_1} is a function of e_1. Eq. 24 is then the multiplicative Pexider functional equation with arguments $f^{(1)}(x)$ and e_1. According to Theorem 3.2.3 in (Castillo, 1992), we have the following three possible solutions to the above functional equation:

P_1: $l''_\beta(y) \equiv 0, \eta''_{E_1} \equiv 0$;

P_2: $l''_\beta(y) \equiv 0, \frac{f^{(1)\prime}}{f^{(2)\prime}} - 1 \equiv 0$;

P_3: $l''_\beta(y) = -abe^{cy}, \eta''_{E_1} = ae^{ce_1}, \frac{f^{(1)\prime}}{f^{(2)\prime}} - 1 = be^{cf^{(1)}(x)}$.

Here a, b, and c are some constants.

We consider the above possible solutions one by one.

1. *Solution P_1 is not valid, because the condition that $\eta''_{E_2} \equiv 0$ does not correspond to a valid distribution (Kagan et al., 1973).*

2. *If P_2 holds, we have $f^{(1)\prime} = f^{(2)\prime}$, i.e., $f^{(1)}(x) = f^{(2)}(x) + c_1$, where c_1 is a constant. Consequently $e_2 = e_1 + c_1$. (Note that the mean of the noise is not fixed; if one sets it to a constant, say, 0, c_1 will then be 0.) Moreover, the condition $l''_\beta(y) \equiv 0$ implies that $l'_\beta(y) = c_2$ or that $l_\beta(y) = c_2 y + d_1$. That is,*

$$\beta_r(y) = e^{c_2 y + d_1} = e^{d_1} \cdot e^{c_2 f_1(x)} \cdot e^{c_2 e_1}.$$

Bearing (12) in mind , we then have the following relationships between $p_X^{(1)}$ and $p_X^{(2)}$ and between p_{E_1} and p_{E_2} accordingly:

$$p_X^{(2)} \propto p_X^{(1)} \cdot e^{-c_2 f^{(1)}(x)},$$
$$p_{E_2} \propto p_{E_1} \cdot e^{-c_2 e_1} = \propto p_{E_1}(e_2 - c_1) \cdot e^{-c_2 e_2}.$$

3. *If P_3 holds, c must be zero such that E_1 has a valid distribution. Furthermore, E_1 must be Gaussian (and correspondingly a must be negative) of the form $p_{E_1} \propto e^{\frac{a}{2}e_1^2 + c_3 e_1 + d_2}$ (Kagan et al., 1973). If the noise is assumed to have a zero mean, then*

$$p_{E_1} \propto e^{\frac{a}{2}e_1^2}. \tag{25}$$

Equation $\frac{f^{(1)\prime}}{f^{(2)\prime}} - 1 = b$ implies that $f^{(2)'} = \frac{1}{1+b} f^{(1)'}$, i.e., $f^{(2)}(x) = \frac{1}{1+b} f^{(1)}(x) + d_3$. Correspondingly, $l''_\beta(y) = -ab$, leading to

$$\beta_r(y) = e^{\frac{-ab}{2}y^2 + c_4 y + d_4}, \tag{26}$$

which is a Gaussian function.

Q.E.D.

S4. Proof of Corollary 3

Proof 3 *This directly follows from Theorem 2. Here we set $\beta_2(y) \equiv 1$, i.e., $(\mathcal{F}_2, \beta_2(y))$ is an ordinary ANM $\mathcal{F}_2$. According to Theorem 1, when E_1 is non-Gaussian, if $(\mathcal{F}_2, \beta_2(y))$ and $(\mathcal{F}_1, \beta_1(y))$ produce the same distribution over (X, Y), then $\beta_r(y) = \beta_2(y)/\beta_1(y) = \beta_1^{-1}(y) \propto e^{c_2 y}$ for a constant c_2, which contradicts a). Similarly, when E_1 is Gaussian, to make (11) hold, $\beta_r(y) = \beta_1^{-1}(y) \propto e^{-\frac{ab}{2}y^2 + c_4 y}$ for some constants a, b, and c_4, contradicting b). Q.E.D.*

S5. Proof of Corollary 4

Proof 4 *According to Theorem 2(b), under the given assumptions,* $f^{(2)}(x) = f^{(1)}(x) + c_1$. *Hence,* $y - f^{(2)}(x) = y - f^{(1)}(x) - c_1$, *or* $E_2 = E_1 - c_1$. *Since* p_{E_1} *and* p_{E_2} *are both symmetric about the origin, they have zero mean, and therefore* $c_1 = 0$, *i.e.,* $E_2 = E_1$. *Moreover, since* p_{E_1} *and* p_{E_2} *are both symmetric about the origin, for any* δ *we have* $p_{E_2}(\delta) \propto p_{E_1}(\delta)e^{-c_2\delta}$ *and* $p_{E_2}(-\delta) \propto p_{E_1}(-\delta)e^{c_2\delta} = p_{E_1}(\delta)e^{c_2\delta}$. $p_{E_2}(\delta) = p_{E_2}(-\delta)$ *implies* $p_{E_1}(\delta)e^{-c_2\delta} = p_{E_1}(\delta)e^{c_2\delta}$. *As* p_{E_1} *is a valid density, there must exist a non-zero value* δ_0 *such that* $p_{E_1}(\delta_0) > 0$. *Thus,* $e^{2c_2\delta_0} = 1$, *implying* $c_2 = 0$. *Therefore,* $p_{E_1}(e_1) = p_{E_2}(e_2)$. *Q.E.D.*

S6. Parameterization of the Functions and Densities

The additive noise model for the data-generating process, (9), implies that $p^{\mathcal{F}}_{Y|X} = p_E(y - f^{AN}(x))$. We parameterize $\beta(y)$ as the exponential transformation of a nonlinear function represented by MLP's (with the tanh activation function); this automatically guarantees the nonnegativity constraint of $\beta(y)$, as required in (17). Furthermore, we represent $p^{\mathcal{F}}_X$ with a mixture of Gaussians, the nonlinear function f^{AN} with MLP's (with the tanh activation function), and p_E with another mixture of Gaussians.

$$\begin{aligned}
\beta(y) &= e^{\tilde{w}(y)}, \\
\tilde{w}(y) &= c_1 + \sum_{i=1}^{K_1} \alpha_{1i} \tanh(b_{1i}(y + d_{1i})), \\
p^{\mathcal{F}}_X(x) &= \sum_{i=1}^{K_2} \alpha_{2i} \mathcal{G}(x; \mu_{2i}, \sigma^2_{2i}), \\
p_E(e) &= \sum_{i=1}^{K_3} \alpha_{2i} \mathcal{G}(e; \mu_{3i}, \sigma^2_{3i}), \\
f^{AN}(x) &= c_2 x + \sum_{i=1}^{K_4} \alpha_{4i} \tanh(b_{4i}(y + d_{4i})),
\end{aligned} \tag{27}$$

where $\alpha_{2i} \geq 0$, $\sum_{i=1}^{K_2} \alpha_{2i} = 1$, $\alpha_{3i} \geq 0$, and $\sum_{i=1}^{K_3} \alpha_{3i} = 1$.

In our experiments, we set $K_1 = 4$, $K_2 = 5$, $K_3 = 5$, and $K_4 = 4$.

S7. More Detail on the Method Based on Score Matching

The maximum likelihood estimation involves sample-average approximation to enforce that p^{β}_{XY} is a valid density. Alternatively, we can estimate the parameters by score matching (Hyvärinen, 2005), i.e., by minimizing the expected squared distance between the gradient of the log-density given by the model and the gradient of the log-density of the observed data. This procedure aims to match the *shape* of the density given by the model and that of the empirical density of the observed data, and is invariant to the scaling factor of the model density. As a clear advantage, in the optimization procedure one does not need to guarantee that p^{β}_{XY} is a valid density.

Given any model density $p_Z(z;\theta)$ of a m-dimensional random vector Z, the score function is the gradient of the log-density w.r.t. the data vector, i.e., $\psi(z;\theta) = (\psi_1(z;\theta), ..., \psi_m(z;\theta))^\intercal = (\frac{\partial \log p_Z(z;\theta)}{\partial z_1}, ..., \frac{\partial \log p_Z(z;\theta)}{\partial z_m})^\intercal$. Note that the score function is invariant to scale transformations in $p_Z(z)$, i.e., it is invariant to the normalization constant for a valid density. One can then estimate model parameters by minimize the expected squared distance between the model score function $\psi(\cdot;\theta)$ and the data score function $\psi_Z(\cdot;\theta)$, i.e., minimize $\frac{1}{2}\int_{z\in\mathbb{R}^m} p_Z(z)||\psi(z;\theta) - \psi_Z(z)||^2 dz$. It has been shown in (Hyvärinen, 2005) that minimizing the above squared distance is equivalent to minimizing

$$J^{SM}(\theta) = \int_{z\in\mathbb{R}^m} p_Z(z) \sum_{i=1}^{m} \big[\tilde{\psi}_i(z;\theta) + \frac{1}{2}\psi_i^2(z;\theta)\big] dz,$$

where $\tilde{\psi}_i(z;\theta) = \frac{\partial \psi_i(z;\theta)}{\partial z_i}$. The sample version of $J^{SM}(\theta)$ over the sample $\mathbf{z}_1, ..., \mathbf{z}_n$ is

$$\hat{J}^{SM}(\theta) = \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{m} \big[\tilde{\psi}_i(\mathbf{z}_k;\theta) + \frac{1}{2}\psi_i^2(\mathbf{z}_k;\theta)\big]. \tag{28}$$

In particular, here we have $\psi_1 = \psi_X$ and $\psi_2 = \psi_Y$; noting that $p^{\mathcal{F}}_{Y|X} = p_E(y - f^{AN}(x))$, we can write down the involved derivatives involved in (28):

$$\begin{aligned}
\psi_X &= \frac{\partial \log p^{\beta}_{XY}}{\partial x} = \frac{\partial \log p_X}{\partial x} + \frac{\partial \log p_E(y - f^{AN}(x))}{\partial x}, \\
\psi_Y &= \frac{\partial \log p^{\beta}_{XY}}{\partial y} = \frac{\partial \log \beta(y)}{\partial y} + \frac{\partial \log p_E(y - f^{AN}(x))}{\partial y}, \\
\tilde{\psi}_X &= \frac{\partial \psi_X}{\partial x} = \frac{\partial^2 \log p_X}{\partial x^2} + \frac{\partial^2 \log p_E(y - f^{AN}(x))}{\partial x^2}, \\
\tilde{\psi}_Y &= \frac{\partial \psi_Y}{\partial y} = \frac{\partial^2 \log \beta(y)}{\partial y^2} + \frac{\partial^2 \log p_E(y - f^{AN}(x))}{\partial y^2},
\end{aligned}$$

in which the involved parameters are denoted by θ_1, θ_2, θ_3, and θ_4, respectively. We use the same regularization term on $\beta(y)$ as in (16).

We use score matching to estimate the parameters involved in (10). The model parameterization was given

in (27). We estimate the parameters by minimizing (28) with the proper constraints on α_{2i} and α_{3i} with the constrained nonlinear optimization toolbox (implemented by the function "fmincon" in MATLAB).

To do so, one has to find the derivative of (28) w.r.t. the involved parameters θ:

$$\frac{\partial \hat{J}(\theta)}{\partial \theta} = \frac{1}{n}\sum_{k=1}^{n}\Big[\frac{\partial \tilde{\psi}_X(k)}{\partial \theta} + \frac{\partial \tilde{\psi}_Y(k)}{\partial \theta} + \psi_X(k)\cdot\frac{\partial \psi_X(k)}{\partial \theta} + \psi_Y(k)\cdot\frac{\partial \psi_Y(k)}{\partial \theta}\Big].$$

More specifically,

$$\frac{\partial \hat{J}(\theta)}{\partial \theta_1} = \frac{1}{n}\sum_{k=1}^{n}\Big[\frac{\partial^3 \tilde{w}(y_k)}{\partial y^2 \partial \theta_1} + \psi_Y(k)\frac{\partial^2 \tilde{w}(y_k)}{\partial y \partial \theta_1}\Big],$$

$$\frac{\partial \hat{J}(\theta)}{\partial \theta_2} = \frac{1}{n}\sum_{k=1}^{n}\Big[\frac{\partial^3 \log p_X(x_k)}{\partial x^2 \partial \theta_2} + \psi_X(k)\frac{\partial^2 \log p_X(x_k)}{\partial x \partial \theta_2}\Big],$$

$$\begin{aligned}\frac{\partial \hat{J}(\theta)}{\partial \theta_3} = \frac{1}{n}\sum_{k=1}^{n}\Big[&\frac{\partial^3 \log p_E(y_k - f^{AN}(x_k))}{\partial x^2 \partial \theta_3} + \\ &\frac{\partial^3 \log p_E(y_k - f^{AN}(x_k))}{\partial y^2 \partial \theta_3} + \\ &\psi_X(k)\cdot\frac{\partial^2 \log p_E(y_k - f^{AN}(x_k))}{\partial x \partial \theta_3} + \\ &\psi_Y(k)\cdot\frac{\partial^2 \log p_E(y_k - f^{AN}(x_k))}{\partial y \partial \theta_3}\Big],\end{aligned}$$

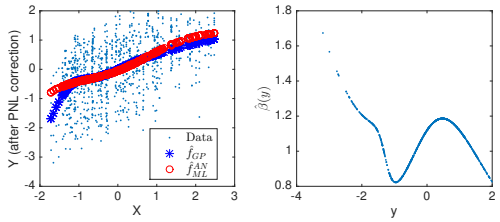
$$\begin{aligned}\frac{\partial \hat{J}(\theta)}{\partial \theta_4} = \frac{1}{n}\sum_{k=1}^{n}\Big[&\frac{\partial^3 \log p_E(y_k - f^{AN}(x_k))}{\partial x^2 \partial \theta_4} + \\ &\frac{\partial^3 \log p_E(y_k - f^{AN}(x_k))}{\partial y^2 \partial \theta_4} + \\ &\psi_X(k)\cdot\frac{\partial^2 \log p_E(y_k - f^{AN}(x_k))}{\partial x \partial \theta_3} + \\ &\psi_Y(k)\cdot\frac{\partial^2 \log p_E(y_k - f^{AN}(x_k))}{\partial y \partial \theta_4}\Big].\end{aligned}$$

The involved partial derivatives can be calculated according to the parameterization (27).

S8. More Results on Real Data

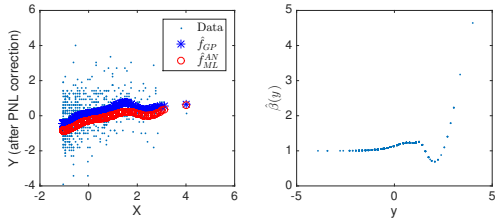
We went through the cause-effect pairs (`http://webdav.tuebingen.mpg.de/cause-effect/`) to find data sets which are likely to suffer the OSB issue according to *commonsense or background knowledge*. We select Pairs 25, 40, and 41: Pair 25 is about the relationship between the age (X) and the concrete compressive strength (Y) of different samples of concrete; Pair 40 is on the relations between the age (X) and diastolic blood pressure (Y) of different subjects; Pair 41 contains the age (X) of the subjects and their plasma glucose concentration a 2 hours in an oral glucose tolerance test (Y).

The empirical distribution of the data in Pair 25 suggests that it is very likely for the effect to suffer from a PNL distortion. We use a rough way to take into account both the PNL distortion in the causal process and the OSB. We first fit the PNL causal model (Zhang & Hyvärinen, 2009) on the data and correct the data with the estimated PNL transformation on the hypothetical effect. We then fit the ANM-OSB procedure on the corrected data. To avoid local optima, we run the algorithm presented in Section 5.1 five times with random initializations and choose the one with the highest likelihood. Figure 7 shows the result on Pair 25. As seen from $\hat{\beta}(y)$, it seems for some reason, the samples whose compressive strength is very high were not selected. The estimated function $\hat{f}_M^{GP}L$ seems to address this issue. For Pair 40, whose results are shown in Figure 8, $\hat{\beta}(y)$ suggests that people with relatively high diastolic blood pressure seem more likely to take part in the test, which seems natural. The interpretation on the results on Pair 41 (Figure 9) may require some domain expertise knowledge.



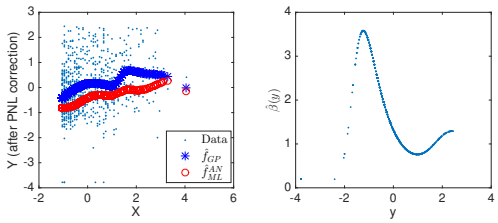
(a) Data & estimated functions. (b) $\hat{\beta}(y)$.

Figure 7: Results on pair 25 of the cause-effect pairs. (a) The scatterplot of the data (after correcting the nonlinear distortion in the hypothetical cause with the PNL causal model, the nonlinear regression function $\hat{f}_{GP}$ on the data, and the estimated function $\hat{f}^{AN}_{ML}$ by the proposed maximum likelihood approach. (b) The estimated density ratio $\beta(y)$ for the selection procedure.



(a) Data & estimated functions. (b) $\hat{\beta}(y)$.

Figure 8: Results on pair 40 (original data without PNL correction).



(a) Data & estimated functions. (b) $\hat{\beta}(y)$.

Figure 9: Results on pair 41 (original data without PNL correction).