Appendix:
Causal Modeling with Stochastic Confounders

Thanh Vinh Vo [1] **Pengfei Wei** [1] **Wicher Bergsma** [2] **Tze-Yun Leong** [1]
[1]School of Computing, National University of Singapore
[2]Department of Statistics, London School of Economics and Political Science
votv@comp.nus.edu.sg, dcsweip@nus.edu.sg, w.p.bergsma@lse.ac.uk, leongty@comp.nus.edu.sg

1 PROOF OF LEMMA 1

We restate that $f_y\colon \mathcal{Z} \times \mathcal{W} \times \mathcal{S} \mapsto \mathcal{F}_y$, $f_w\colon \mathcal{Z} \times \mathcal{S} \mapsto \mathcal{F}_w$, $f_z\colon \mathcal{Z} \mapsto \mathcal{F}_z$, $f_s\colon \mathcal{S} \mapsto \mathcal{F}_s$, $f_x\colon \mathcal{Z} \times \mathcal{S} \mapsto \mathcal{F}_x$, and $f_q\colon \mathcal{Y} \times \mathcal{W} \times \mathcal{S} \times \mathcal{X} \to \mathcal{F}_z$. In this work, $\mathcal{F}_y = \mathbb{R}$, $\mathcal{F}_w = \mathbb{R}$, $\mathcal{F}_s = \mathbb{R}^{d_s}$, $\mathcal{F}_x = \mathbb{R}^{d_x}$ and $\mathcal{F}_z = \mathbb{R}^{d_z}$.

We also note that $\boldsymbol{\mathcal{A}} = \{y, w, x, s, z, q\}$.

Lemma 1. *Let κ_i be kernels and $\mathcal{H}_i$ their associated reproducing kernel Hilbert space (RKHS), where $i \in \boldsymbol{\mathcal{A}}$. Let the empirical risk obtained from the negative ELBO be $\widehat{\mathcal{L}}$. Consider minimizing the following objective function*

$$J = \widehat{\mathcal{L}}\left(\bigcup\nolimits_{i\in\boldsymbol{\mathcal{A}}} f_i\right) + \sum_{i\in\boldsymbol{\mathcal{A}}} \lambda_i \|f_i\|^2_{\mathcal{H}_i} \tag{1}$$

with respect to functions f_i $(i \in \boldsymbol{\mathcal{A}})$, where $\lambda_i \in \mathbb{R}^+$. Then, the minimizer of (1) has the following form $f_i = \sum_{l=1}^{T\times L} \kappa_i(\,\cdot\,, \boldsymbol{\nu}_l^i)\boldsymbol{\beta}_l^i$ $(\forall i \in \boldsymbol{\mathcal{A}})$, where $\boldsymbol{\nu}_l^i$ is the l^{th} input to function f_i, i.e., it is a subset of the l^{th} tuple of $\mathcal{D}$, and the coefficients $\boldsymbol{\beta}_l^i$ are vectors in the Hilbert space $\mathcal{F}_i$. This minimizer further emits the following solution: $\boldsymbol{\beta}^i = [\boldsymbol{\beta}_l^i,...,\boldsymbol{\beta}_{TL}^i]^\top = \left(\sum_{l=1}^{L} \mathbf{K}_i^{l\,\top}\mathbf{K}_i^l + \lambda_i L\mathbf{K}_i\right)^{-1}\sum_{l=1}^{L}\mathbf{K}_i^{l\,\top}\boldsymbol{\psi}^i$ for $i \in \boldsymbol{\mathcal{A}} \setminus \{w, q\}$ and $\boldsymbol{\psi}^y = \mathbf{y}$, $\boldsymbol{\psi}^x = \mathbf{x}$, $\boldsymbol{\psi}^s = \mathbf{s}$, $\boldsymbol{\psi}^z = \mathbf{K}_q\boldsymbol{\beta}^q$.

Proof. We divide the proof of Lemma 1 into two parts **(1)** and **(2)** as follows.

(1) The solution form of f_i $(i \in \boldsymbol{\mathcal{A}})$

We further define $f_x = [f_{x,1},..., f_{x,d_x}]$ with $f_{x,d}\colon \mathcal{Z} \times \mathcal{S} \mapsto \mathbb{R}$ $(d = 1,\ldots,d_x)$. Similarly, $f_s = [f_{s,1},..., f_{s,d_s}]$ with $f_{s,d}\colon \mathcal{S} \mapsto \mathbb{R}$ $(d = 1,\ldots,d_s)$, $f_z = [f_{z,1},..., f_{z,d_z}]$ with $f_{z,d}\colon \mathcal{Z} \mapsto \mathbb{R}$ $(d = 1,\ldots,d_z)$, and $f_q = [f_{q,1},..., f_{q,d_z}]$ with $f_{q,d}\colon \mathcal{Y} \times \mathcal{W} \times \mathcal{S} \times \mathcal{X} \mapsto \mathbb{R}$ $(d = 1,\ldots,d_z)$.

Consider the subspaces $\mathcal{B}_i \subset \mathcal{H}_i$, $(i \in \boldsymbol{\mathcal{A}})$ defined as follows:

$$\begin{aligned}
\mathcal{B}_y &= \mathbf{span}\big\{\kappa_y(\cdot, [w_t, \mathbf{s}_t, \mathbf{z}_t^l]) \,:\, t = 1,...,T; l = 1,...,L\big\},\\
\mathcal{B}_w &= \mathbf{span}\left\{\kappa_w(\cdot, [\mathbf{s}_t, \mathbf{z}_t^l]) \,:\, t = 1,...,T; l = 1,...,L\right\},\\
\mathcal{B}_x &= \mathbf{span}\left\{\kappa_x(\cdot, [\mathbf{s}_t, \mathbf{z}_t^l]) \,:\, t = 1,...,T; l = 1,...,L\right\},\\
\mathcal{B}_s &= \mathbf{span}\left\{\kappa_s(\cdot, \mathbf{s}_t) \,:\, t = 1,...,T\right\},\\
\mathcal{B}_z &= \mathbf{span}\left\{\kappa_z(\cdot, \mathbf{z}_t^l) \,:\, t = 1,...,T; l = 1,...,L\right\},\\
\mathcal{B}_q &= \mathbf{span}\left\{\kappa_q(\cdot, [y_t, \mathbf{x}_t, w_t, \mathbf{s}_t]) \,:\, t = 1,...,T\right\}.
\end{aligned}$$

We project f_y, f_w, $f_{x,d}$ $(d = 1,...,d_x)$, $f_{s,d}$ $(d = 1,...,d_s)$, $f_{z,d}$ $(d = 1,...,d_z)$, and $f_{q,d}$ $(d = 1,...,d_z)$ onto the subspaces $\mathcal{B}_y$, $\mathcal{B}_w$, $\mathcal{B}_x$, $\mathcal{B}_s$, $\mathcal{B}_z$, $\mathcal{B}_q$, respectively, to obtain f_y^{sp}, f_w^{sp}, $f_{x,d}^{\text{sp}}$, $f_{s,d}^{\text{sp}}$, $f_{z,d}^{\text{sp}}$ and $f_{q,d}^{\text{sp}}$, and also project them onto the perpendicular spaces of the subspaces to obtain $f_y^{\perp}$, $f_w^{\perp}$, $f_{x,d}^{\perp}$, $f_{s,d}^{\perp}$, $f_{s,d}^{\perp}$ and $f_{q,d}^{\perp}$.

Note that $f^{\text{sp}}_{(\cdot)} + f^{\perp}_{(\cdot)} = f_{(\cdot)}$. Thus, $\|f_{(\cdot)}\|^2_{\mathcal{H}_{(\cdot)}} = \|f^{\text{sp}}_{(\cdot)}\|^2_{\mathcal{H}_{(\cdot)}} + \|f^{\perp}_{(\cdot)}\|^2_{\mathcal{H}_{(\cdot)}} \geq \|f^{\text{sp}}_{(\cdot)}\|^2_{\mathcal{H}_{(\cdot)}}$, which implies that $\lambda_{(\cdot)}\|f_{(\cdot)}\|^2_{\mathcal{H}_{(\cdot)}}$ is minimized if $f_{(\cdot)}$ is in its subspace $\mathcal{B}_{(\cdot)}$. (a)

Moreover, from the reproducing property, we have that

$$\begin{aligned} f_y(w_t, \mathbf{s}_t, \mathbf{z}^l_t) &= \left\langle f_y, \kappa_y(\cdot, [w_t, \mathbf{s}_t, \mathbf{z}^l_t])\right\rangle_{\mathcal{H}_y} = \left\langle f^{\text{sp}}_y + f^{\perp}_y, \kappa_y(\cdot, [w_t, \mathbf{s}_t, \mathbf{z}^l_t])\right\rangle_{\mathcal{H}_y} \\ &= \left\langle f^{\text{sp}}_y, \kappa_y(\cdot, [w_t, \mathbf{s}_t, \mathbf{z}^l_t])\right\rangle_{\mathcal{H}_y} + \left\langle f^{\perp}_y, \kappa_y(\cdot, [w_t, \mathbf{s}_t, \mathbf{z}^l_t])\right\rangle_{\mathcal{H}_y} \\ &= \left\langle f^{\text{sp}}_y, \kappa_y(\cdot, [w_t, \mathbf{s}_t, \mathbf{z}^l_t])\right\rangle_{\mathcal{H}_y} = f^{\text{sp}}_y(w_t, \mathbf{s}_t, \mathbf{z}^l_t), \end{aligned}$$

and similarly $f_w(\mathbf{z}^l_t, \mathbf{s}_t) = f^{\text{sp}}_w(\mathbf{z}^l_t, \mathbf{s}_t)$, $f_{x,d}(\mathbf{z}^l_t, \mathbf{s}_t) = f^{\text{sp}}_{x,d}(\mathbf{z}^l_t, \mathbf{s}_t)$, $f_{s,d}(\mathbf{s}_t) = f^{\text{sp}}_{s,d}(\mathbf{s}_t)$, $f_{z,d}(\mathbf{z}_t) = f^{\text{sp}}_{z,d}(\mathbf{z}_t)$, and $f_{q,d}(y_t, w_t, \mathbf{s}_t, \mathbf{x}_t) = f^{\text{sp}}_{q,d}(y_t, w_t, \mathbf{s}_t, \mathbf{x}_t)$. Hence, we have

$$\widehat{\mathcal{L}}\left(\bigcup\nolimits_{i\in\boldsymbol{\mathcal{A}}} f_i\right) = \widehat{\mathcal{L}}\left(\bigcup\nolimits_{i\in\boldsymbol{\mathcal{A}}} f^{\text{sp}}_i\right).$$

The last equation implies that $\widehat{\mathcal{L}}(\cdot)$ depends only on the component of $f_y, f_w, f_{x,d}, f_{s,d}, f_{z,d}, f_{q,d}$ lying in the subspaces $\mathcal{B}_y$, $\mathcal{B}_w$, $\mathcal{B}_x$, $\mathcal{B}_s$, $\mathcal{B}_z$, $\mathcal{B}_q$, respectively. (b)

From (a) and (b), we have

$$\begin{aligned} f_y &= \textstyle\sum_{l=1}^{TL} \kappa_y(\,\cdot\,, [w_l, \mathbf{s}_l, \mathbf{z}_l])\boldsymbol{\beta}^y_l, & f_w &= \textstyle\sum_{l=1}^{TL} \kappa_w(\,\cdot\,, [\mathbf{s}_l, \mathbf{z}_l])\boldsymbol{\beta}^w_l, \\ f_x &= \textstyle\sum_{l=1}^{TL} \kappa_x(\,\cdot\,, [\mathbf{s}_l, \mathbf{z}_l])\boldsymbol{\beta}^x_l, & f_s &= \textstyle\sum_{l=1}^{TL} \kappa_s(\,\cdot\,, \mathbf{s}_l)\boldsymbol{\beta}^s_l, \\ f_z &= \textstyle\sum_{l=1}^{TL} \kappa_z(\,\cdot\,, \mathbf{z}_l)\boldsymbol{\beta}^z_l, & f_q &= \textstyle\sum_{l=1}^{TL} \kappa_q(\,\cdot\,, [y_l, \mathbf{x}_l, w_l, \mathbf{s}_l])\boldsymbol{\beta}^q_l, \end{aligned}$$

where $\boldsymbol{\beta}^i_l$ is in the Hilbert space $\mathcal{F}_i$ $(i \in \boldsymbol{\mathcal{A}})$. This completes the proof on the solution form of f_i $(i \in \boldsymbol{\mathcal{A}})$. We further derive the solution of $\boldsymbol{\beta}^i_l$ $(i \in \boldsymbol{\mathcal{A}})$ in the next subsection.

(2) The solution of $\boldsymbol{\beta}^i$ $(i \in \boldsymbol{\mathcal{A}})$

Let $\mathbf{x}_{:,d}$, $\mathbf{s}_{:,d}$, $\mathbf{z}_{:,d}$ be the d-th column of $\mathbf{x}$, $\mathbf{s}$ and $\mathbf{z}$, respectively, i.e., each element of $\mathbf{x}_{:,d}$, $\mathbf{s}_{:,d}$, $\mathbf{z}_{:,d}$ is the d-th dimension of $\mathbf{x}_t$, $\mathbf{s}_t$ and $\mathbf{z}_t$.

We further denote $\mathbf{K}_s, \mathbf{K}_q \in \mathbb{R}^{T\times T}$ as kernel matrices computed with the kernel functions $\kappa_s(\mathbf{s}_i, \mathbf{s}_j)$ and $\kappa_q([\mathbf{x}_i, y_i, w_i, \mathbf{s}_i], [\mathbf{x}_j, y_j, w_j, \mathbf{s}_j])$, respectively.

The following form of matrices are applied to $\mathbf{K}_x$, $\mathbf{K}_w$, $\mathbf{K}_z$, $\mathbf{K}_y$, $\mathbf{K}^l_x$, $\mathbf{K}^l_w$, $\mathbf{K}^l_z$, $\mathbf{K}^l_y$:

$$\mathbf{K}_{(\cdot)} = \begin{bmatrix} \kappa_{(\cdot)}(\Phi^1_1, \Phi^1_1) \dots \kappa_{(\cdot)}(\Phi^1_1, \Phi^1_T) \dots \kappa_{(\cdot)}(\Phi^1_1, \Phi^L_1) \dots \kappa_{(\cdot)}(\Phi^1_1, \Phi^L_T) \\ \vdots \quad \ddots \quad \vdots \quad\quad\quad \vdots \quad \ddots \quad \vdots \\ \kappa_{(\cdot)}(\Phi^1_T, \Phi^1_1) \dots \kappa_{(\cdot)}(\Phi^1_T, \Phi^1_T) \dots \kappa_{(\cdot)}(\Phi^1_T, \Phi^L_1) \dots \kappa_{(\cdot)}(\Phi^1_T, \Phi^L_T) \\ \vdots \quad\quad\quad \vdots \quad\quad\quad \vdots \quad\quad\quad \vdots \\ \kappa_{(\cdot)}(\Phi^L_1, \Phi^1_1) \dots \kappa_{(\cdot)}(\Phi^L_1, \Phi^1_T) \dots \kappa_{(\cdot)}(\Phi^L_1, \Phi^L_1) \dots \kappa_{(\cdot)}(\Phi^L_1, \Phi^L_T) \\ \vdots \quad \ddots \quad \vdots \quad\quad\quad \vdots \quad \ddots \quad \vdots \\ \kappa_{(\cdot)}(\Phi^L_T, \Phi^1_1) \dots \kappa_{(\cdot)}(\Phi^L_T, \Phi^1_T) \dots \kappa_{(\cdot)}(\Phi^L_T, \Phi^L_1) \dots \kappa_{(\cdot)}(\Phi^L_T, \Phi^L_T) \end{bmatrix} \in \mathbb{R}^{TL\times TL},$$

$$\mathbf{K}^l_{(\cdot)} = \begin{bmatrix} \kappa_{(\cdot)}(\Phi^l_1, \Phi^1_1) \dots \kappa_{(\cdot)}(\Phi^l_1, \Phi^1_T) \dots \kappa_{(\cdot)}(\Phi^l_1, \Phi^L_1) \dots \kappa_{(\cdot)}(\Phi^l_1, \Phi^L_T) \\ \vdots \quad \ddots \quad \vdots \quad\quad\quad \vdots \quad \ddots \quad \vdots \\ \kappa_{(\cdot)}(\Phi^l_T, \Phi^1_1) \dots \kappa_{(\cdot)}(\Phi^l_T, \Phi^1_T) \dots \kappa_{(\cdot)}(\Phi^l_T, \Phi^L_1) \dots \kappa_{(\cdot)}(\Phi^l_T, \Phi^L_T) \end{bmatrix} \in \mathbb{R}^{T\times TL},$$

where

- For $\mathbf{K}_x$ and $\mathbf{K}^l_x$, $\kappa_{(\cdot)}(\Phi^a_i, \Phi^b_j) = \kappa_x([\mathbf{s}_i, \mathbf{z}^a_i], [\mathbf{s}_j, \mathbf{z}^b_j])$,
- For $\mathbf{K}_w$ and $\mathbf{K}^l_w$, $\kappa_{(\cdot)}(\Phi^a_i, \Phi^b_j) = \kappa_w([\mathbf{s}_i, \mathbf{z}^a_i], [\mathbf{s}_j, \mathbf{z}^b_j])$,

- For $\mathbf{K}_z$ and $\mathbf{K}_z^l$, $\kappa_{(\cdot)}(\Phi_i^a, \Phi_j^b) = \kappa_z(\mathbf{z}_i^a, \mathbf{z}_j^b)$,
- For $\mathbf{K}_y$ and $\mathbf{K}_y^l$, $\kappa_{(\cdot)}(\Phi_i^a, \Phi_j^b) = \kappa_y([w_i, \mathbf{s}_i, \mathbf{z}_i^a], [w_i, \mathbf{s}_j, \mathbf{z}_j^b])$.

The regularized empirical loss function can be expanded as follows:

$$
\begin{aligned}
J = \widehat{\mathcal{L}}(f_y, f_w, f_x, f_s, f_z, f_q) &+ \lambda_y \|f_y\|^2_{\mathcal{H}_y} + \lambda_w \|f_w\|^2_{\mathcal{H}_w} + \sum_{d=1}^{d_x} \lambda_x \|f_{x,d}\|^2_{\mathcal{H}_x} \\
&+ \sum_{d=1}^{d_s} \lambda_s \|f_{s,d}\|^2_{\mathcal{H}_s} + \sum_{d=1}^{d_z} \lambda_z \|f_{z,d}\|^2_{\mathcal{H}_z} + \sum_{d=1}^{d_z} \lambda_q \|f_{q,d}\|^2_{\mathcal{H}_q} \\
= \frac{1}{L} \sum_{l=1}^{L} \Bigg(&\frac{1}{2\sigma_y^2} ((\boldsymbol{\beta}^y)^\top {\mathbf{K}_y^l}^\top \mathbf{K}_y^l \boldsymbol{\beta}^y - 2\mathbf{y}^\top \mathbf{K}_y^l \boldsymbol{\beta}^y + \mathbf{y}^\top \mathbf{y}) + T \log \sigma_y \\
&- \mathbf{w}^\top \log \varphi(\mathbf{K}_w^l \boldsymbol{\beta}^w) - (\mathbf{1} - \mathbf{w})^\top \log \varphi(-\mathbf{K}_w^l \boldsymbol{\beta}^w) \\
&+ \frac{1}{2\sigma_s^2} \sum_{d=1}^{d_s} ((\boldsymbol{\beta}_d^s)^\top \mathbf{K}_s^\top \mathbf{K}_s \boldsymbol{\beta}_d^s - 2\mathbf{s}_{:,d}^\top \mathbf{K}_s \boldsymbol{\beta}_d^s + \mathbf{s}_{:,d}^\top \mathbf{s}_{:,d}) + T d_s \log \sigma_s \\
&+ \frac{1}{2\sigma_x^2} \sum_{d=1}^{d_x} ((\boldsymbol{\beta}_d^x)^\top {\mathbf{K}_x^l}^\top \mathbf{K}_x^l \boldsymbol{\beta}_d^x - 2\mathbf{x}_{:,d}^\top \mathbf{K}_x^l \boldsymbol{\beta}_d^x + \mathbf{x}_{:,d}^\top \mathbf{x}_{:,d}) + T d_x \log \sigma_x \\
&+ \frac{1}{2\sigma_z^2} \sum_{d=1}^{d_z} ((\boldsymbol{\beta}_d^z)^\top {\mathbf{K}_z^l}^\top \mathbf{K}_z^l \boldsymbol{\beta}_d^z - 2(\boldsymbol{\beta}_d^q)^\top \mathbf{K}_q^\top \mathbf{K}_z^l \boldsymbol{\beta}_d^z + (\boldsymbol{\beta}_d^q)^\top \mathbf{K}_q^\top \mathbf{K}_q \boldsymbol{\beta}_d^q) + T d_z \log(\sigma_z/\sigma_q) + \frac{1}{2} T d_z \sigma_q^2/\sigma_z^2 \Bigg) \\
&+ \lambda_y (\boldsymbol{\beta}^y)^\top \mathbf{K}_y \boldsymbol{\beta}^y + \lambda_w (\boldsymbol{\beta}^w)^\top \mathbf{K}_w \boldsymbol{\beta}^w + \sum_{d=1}^{d_x} \lambda_x (\boldsymbol{\beta}_d^x)^\top \mathbf{K}_x \boldsymbol{\beta}_d^x \\
&+ \sum_{d=1}^{d_s} \lambda_s (\boldsymbol{\beta}_d^s)^\top \mathbf{K}_s \boldsymbol{\beta}_d^s + \sum_{d=1}^{d_z} \lambda_z (\boldsymbol{\beta}_d^z)^\top \mathbf{K}_z \boldsymbol{\beta}_d^z + \sum_{d=1}^{d_z} \lambda_q (\boldsymbol{\beta}_d^q)^\top \mathbf{K}_q \boldsymbol{\beta}_d^q
\end{aligned}
$$

Taking derivative of J with respect to $\boldsymbol{\beta}^y$ and equates to 0, we obtain

$$
\nabla_{\boldsymbol{\beta}^y} J = \frac{1}{L} \sum_{l=1}^{L} \left(\frac{1}{2\sigma_y^2} (2{\mathbf{K}_y^l}^\top \mathbf{K}_y^l \boldsymbol{\beta}^y - 2{\mathbf{K}_y^l}^\top \mathbf{y}) \right) + 2\lambda_y \mathbf{K}_y \boldsymbol{\beta}^y = 0,
$$

which implies that

$$
\begin{aligned}
\boldsymbol{\beta}^y &= \left(\frac{1}{\sigma_y^2} \frac{1}{L} \sum_{l=1}^{L} {\mathbf{K}_y^l}^\top \mathbf{K}_y^l + 2\lambda_y \mathbf{K}_y \right)^{-1} \frac{1}{\sigma_y^2} \left(\frac{1}{L} \sum_{l=1}^{L} \mathbf{K}_y^l \right)^\top \mathbf{y} \\
&= \left(\sum_{l=1}^{L} {\mathbf{K}_y^l}^\top \mathbf{K}_y^l + 2\lambda_y L \sigma_y^2 \mathbf{K}_y \right)^{-1} \left(\sum_{l=1}^{L} \mathbf{K}_y^l \right)^\top \mathbf{y}.
\end{aligned}
$$

Similarly, taking derivative of J w.r.t $\boldsymbol{\beta}_d^x, \boldsymbol{\beta}_d^s, \boldsymbol{\beta}_d^z$ and set to 0, we obtain the following update rules

$$
\begin{aligned}
\boldsymbol{\beta}_d^x &= \left(\sum_{l=1}^{L} {\mathbf{K}_x^l}^\top \mathbf{K}_x^l + 2\lambda_x L \sigma_x^2 \mathbf{K}_x \right)^{-1} \left(\sum_{l=1}^{L} \mathbf{K}_x^l \right)^\top \mathbf{x}_{:,d}, \\
\boldsymbol{\beta}_d^s &= \left(\mathbf{K}_s + 2\lambda_s \sigma_s^2 \mathbf{I} \right)^{-1} \mathbf{s}_{:,d}, \\
\boldsymbol{\beta}_d^z &= \left(\sum_{l=1}^{L} {\mathbf{K}_z^l}^\top \mathbf{K}_z^l + 2\lambda_z L \sigma_z^2 \mathbf{K}_z \right)^{-1} \sum_{l=1}^{L} \left({\mathbf{K}_z^l}^\top \mathbf{K}_q \boldsymbol{\beta}_d^q \right).
\end{aligned}
$$

Absorbing the all the constants into $\lambda_{(\cdot)}$ and denoting $\boldsymbol{\beta}^i = [\boldsymbol{\beta}^i_1, \ldots, \boldsymbol{\beta}^i_{d_i}]$ $(i \in \{x, s, z\})$, we obtain

$$\boxed{\begin{aligned}
\boldsymbol{\beta}^y &= \left(\sum_{l=1}^{L} {\mathbf{K}^l_y}^\top \mathbf{K}^l_y + \lambda_y L \mathbf{K}_y\right)^{-1} \left(\sum_{l=1}^{L} \mathbf{K}^l_y\right)^\top \mathbf{y}, \\
\boldsymbol{\beta}^x &= \left(\sum_{l=1}^{L} {\mathbf{K}^l_x}^\top \mathbf{K}^l_x + \lambda_x L \mathbf{K}_x\right)^{-1} \left(\sum_{l=1}^{L} \mathbf{K}^l_x\right)^\top \mathbf{x}, \\
\boldsymbol{\beta}^s &= (\mathbf{K}_s + \lambda_s \mathbf{I})^{-1} \mathbf{s}, \\
\boldsymbol{\beta}^z &= \left(\sum_{l=1}^{L} {\mathbf{K}^l_z}^\top \mathbf{K}^l_z + \lambda_z L \mathbf{K}_z\right)^{-1} \sum_{l=1}^{L} \left({\mathbf{K}^l_z}^\top \mathbf{K}_q \boldsymbol{\beta}_q\right).
\end{aligned}}$$

We remark that $\boldsymbol{\beta}^s = (\mathbf{K}_s + \lambda_s \mathbf{I})^{-1} \mathbf{s} = \left(\sum_{l=1}^{L} \mathbf{K}_s \mathbf{K}_s + \lambda_s L \mathbf{K}_s\right)^{-1} (\sum_{l=1}^{L} \mathbf{K}_s)^\top \mathbf{s}$, thus we can present a general form of the solution as in Lemma 1. □

2 PROOF OF LEMMA 2

We repeat Lemma 2 here for convenience:

Lemma 2. *For any fixed $\boldsymbol{\beta}^q$, the objective function J in Eq. (1) is convex with respect to $\boldsymbol{\beta}^i$ for all $i \in \boldsymbol{\mathcal{A}} \setminus \{q\}$.*

Proof. From Lemma 1, we see that the objective function J is a combination of several components including $(\boldsymbol{\beta}^i)^\top \mathbf{C} \boldsymbol{\beta}^i$, $\mathbf{c}^\top \boldsymbol{\beta}^i$, $-\mathbf{w}^\top \log \varphi(\mathbf{K}^l_w \boldsymbol{\beta}^w)$ and $-(\mathbf{1} - \mathbf{w})^\top \log \varphi(-\mathbf{K}^l_w \boldsymbol{\beta}^w)$, where $i \in \{y, s, x, z\}$, $\mathbf{C}$ is a positive semi-definite matrix and $\mathbf{c}$ is a vector.

For the first and second term, we have

$$\begin{aligned}
\nabla^2_{\boldsymbol{\beta}^i} \left\{ (\boldsymbol{\beta}^i)^\top \mathbf{C} \boldsymbol{\beta}^i \right\} &= \mathbf{C} + \mathbf{C}^\top = 2\mathbf{C} \succeq 0, \\
\nabla^2_{\boldsymbol{\beta}^i} \left\{ \mathbf{c}^\top \boldsymbol{\beta}^i \right\} &= \mathbf{0} \succeq 0.
\end{aligned}$$

For the third term, we have

$$\begin{aligned}
\nabla_{\boldsymbol{\beta}^w} \left\{ -\mathbf{w}^\top \log \varphi(\mathbf{K}^l_w \boldsymbol{\beta}^w) \right\} &= -(\nabla_{\boldsymbol{\beta}^w} \log \varphi(\mathbf{K}^l_w \boldsymbol{\beta}^w)) \mathbf{w} \\
&= -(\mathbf{K}^l_w)^\top \mathrm{diag}(\mathbf{w}) \varphi(-\mathbf{K}^l_w \boldsymbol{\beta}^w),
\end{aligned}$$

and thus

$$\begin{aligned}
\nabla^2_{\boldsymbol{\beta}^w} \left\{ -\mathbf{w}^\top \log \varphi(\mathbf{K}^l_w \boldsymbol{\beta}^w) \right\} &= -(\nabla_{\boldsymbol{\beta}^w} \varphi(-\mathbf{K}^l_w \boldsymbol{\beta}^w))((\mathbf{K}^l_w)^\top \mathrm{diag}(\mathbf{w}))^\top \\
&= -(\nabla_{\boldsymbol{\beta}^w} \varphi(-\mathbf{K}^l_w \boldsymbol{\beta}^w)) \mathrm{diag}(\mathbf{w}) \mathbf{K}^l_w \\
&= (\mathbf{K}^l_w)^\top \mathrm{diag}(\varphi(-\mathbf{K}^l_w \boldsymbol{\beta}^w) \odot \varphi(\mathbf{K}^l_w \boldsymbol{\beta}^w) \odot \mathbf{w}) \mathbf{K}^l_w \succeq 0.
\end{aligned}$$

Similarly, for the last term, we have

$$\nabla_{\boldsymbol{\beta}^w} \left\{ -(\mathbf{1} - \mathbf{w})^\top \log \varphi(-\mathbf{K}^l_w \boldsymbol{\beta}^w) \right\} = (\mathbf{K}^l_w)^\top \mathrm{diag}(\mathbf{1} - \mathbf{w}) \varphi(\mathbf{K}^l_w \boldsymbol{\beta}^w),$$

and

$$\nabla^2_{\boldsymbol{\beta}^w} \left\{ -(\mathbf{1} - \mathbf{w})^\top \log \varphi(-\mathbf{K}^l_w \boldsymbol{\beta}^w) \right\} = (\mathbf{K}^l_w)^\top \mathrm{diag}(\varphi(\mathbf{K}^l_w \boldsymbol{\beta}^w) \odot \varphi(-\mathbf{K}^l_w \boldsymbol{\beta}^w) \odot (\mathbf{1} - \mathbf{w})) \mathbf{K}^l_w \succeq 0.$$

Moreover, the second-order derivative of J with respect to $\sigma_{(\cdot)}$s are also positive. Consequently, J is convex because it is a linear combination of convex functions. □

3 AUXILIARY DISTRIBUTION $\widetilde{p}(\mathbf{y} \mid \mathbf{s}, \mathbf{x}, \mathbf{w})$, $\widetilde{p}(\mathbf{w} \mid \mathbf{s}, \mathbf{x})$ AND $\widetilde{p}(\mathbf{s} \mid \mathbf{x})$

This section outlines the approximation of $p(\mathbf{y}|\mathbf{s}, \mathbf{x}, \mathbf{w})$, $p(\mathbf{w}|\mathbf{s}, \mathbf{x})$ and $p(\mathbf{s}|\mathbf{x})$. Denoting their corresponding approximation as $\widetilde{p}(\mathbf{y}|\mathbf{s}, \mathbf{x}, \mathbf{w})$, $\widetilde{p}(\mathbf{w}|\mathbf{s}, \mathbf{x})$, $\widetilde{p}(\mathbf{s}|\mathbf{x})$, we estimate parameters of those distribution using classical representer theorem. Learning parameters of $\widetilde{p}(\mathbf{w} \mid \mathbf{s}, \mathbf{x})$ is presented in the main text. In the following, we present $\widetilde{p}(\mathbf{y} \mid \mathbf{s}, \mathbf{x}, \mathbf{w})$ and $\widetilde{p}(\mathbf{s} \mid \mathbf{x})$.

3.1 Learning $\widetilde{p}(\mathbf{y} \mid \mathbf{s}, \mathbf{x}, \mathbf{w})$

$$\widetilde{p}(\mathbf{y} \mid \mathbf{s}, \mathbf{x}, \mathbf{w}) = \prod_{t=1}^{T} \widetilde{p}(y_t \mid y_{t-1}, \mathbf{s}_t, \mathbf{x}_t, w_t) = \prod_{t=1}^{T} \mathsf{N}(y_t \mid g_y(y_{t-1}, \mathbf{s}_t, \mathbf{x}_t, w_t), \omega_y^2).$$

The regularized empirical risk is as follows:

$$\begin{aligned} J_y &= \frac{1}{2\omega_y^2} \sum_{t=1}^{T} (y_t - g_y(y_{t-1}, \mathbf{s}_t, \mathbf{x}_t, w_t))^2 + T \log \omega_y + \delta_y \|g_y\|_{\mathcal{V}_y}^2 \\ &= \frac{1}{2\omega_y^2} (\mathbf{y} - \mathbf{K}_{\widetilde{y}} \boldsymbol{\alpha}^y)^\top (\mathbf{y} - \mathbf{K}_{\widetilde{y}} \boldsymbol{\alpha}^y) + T \log \omega_y + \delta_y (\boldsymbol{\alpha}^y)^\top \mathbf{K}_{\widetilde{y}} \boldsymbol{\alpha}^y, \end{aligned}$$

where $\boldsymbol{\alpha}^y = [\alpha_1^y, \ldots, \alpha_T^y]^\top$ is the parameter to be learned and $\mathbf{K}_{\widetilde{y}}$ is the matrix computed with kernel function $\tau_y([y_{i-1}, \mathbf{s}_i, \mathbf{x}_i, w_i], [y_{j-1}, \mathbf{s}_j, \mathbf{x}_j, w_j])$. Taking derivative of J_y w.r.t $\boldsymbol{\alpha}^y$ and set to 0, we obtain

$$\boxed{\boldsymbol{\alpha}^y = (\mathbf{K}_{\widetilde{y}} + 2\delta_y \omega_y^2 \mathbf{I}_T)^{-1} \mathbf{y}.}$$

3.2 Learning $\widetilde{p}(\mathbf{s} \mid \mathbf{x})$

$$\widetilde{p}(\mathbf{s} \mid \mathbf{x}) = \prod_{t=1}^{T} \widetilde{p}(\mathbf{s}_t \mid \mathbf{s}_{t-1}, \mathbf{x}_t) = \prod_{t=1}^{T} \mathsf{N}(\mathbf{s}_t \mid g_s(\mathbf{s}_{t-1}, \mathbf{x}_t), \omega_s^2 \mathbf{I}_{d_s}).$$

The regularized empirical risk is as follows:

$$\begin{aligned} J_s &= \frac{1}{2\omega_s^2} \sum_{t=1}^{T} \|\mathbf{s}_t - g_s(\mathbf{s}_{t-1}, \mathbf{x}_t)\|_2^2 + T d_s \log \omega_s + \sum_{d=1}^{d_s} \delta_s \|g_{s,d}\|_{\mathcal{V}_s}^2 \\ &= \frac{1}{2\omega_s^2} \sum_{d=1}^{d_s} (\mathbf{s}_{:,d} - \mathbf{K}_{\widetilde{s}} \boldsymbol{\alpha}_d^s)^\top (\mathbf{s}_{:,d} - \mathbf{K}_{\widetilde{s}} \boldsymbol{\alpha}_d^s) + T d_s \log \omega_s + \sum_{d=1}^{d_s} \delta_s (\boldsymbol{\alpha}_d^s)^\top \mathbf{K}_{\widetilde{s}} \boldsymbol{\alpha}_d^s, \end{aligned}$$

where $\mathbf{s}_{:,d}$ denotes the d-th column of $\mathbf{s}$, $\boldsymbol{\alpha}_d^s = [\alpha_{1,d}^s, \ldots, \alpha_{T,d}^s]^\top$ is the parameter to be learned and $\mathbf{K}_{\widetilde{s}}$ is the matrix computed with kernel function $\tau_s([\mathbf{s}_{i-1}, \mathbf{x}_i], [\mathbf{s}_{j-1}, \mathbf{x}_j])$. Taking derivative of J_s w.r.t $\boldsymbol{\alpha}_d^s$ and set to 0, we obtain $\boldsymbol{\alpha}_d^s = (\mathbf{K}_{\widetilde{s}} + 2\delta_s \omega_s^2 \mathbf{I}_T)^{-1} \mathbf{s}_{:,d}$ where $d = 1, 2, \ldots, d_s$. Thus we have

$$\boxed{\boldsymbol{\alpha}^s = (\mathbf{K}_{\widetilde{s}} + 2\delta_s \omega_s^2 \mathbf{I}_T)^{-1} \mathbf{s}.}$$

4 USING ADDITIVE KERNELS

This section presents the update rules for some additional parameters when using an additive kernel function. For example, we define an additive kernel function as follows

$$\kappa_y([w, s, z], [w', s', z']) = \gamma_0^y + \gamma_w^y \, \phi_w(w, w') + \gamma_s^y \, \phi_s(s, s') + \gamma_z^y \, \phi_z(z, z')$$

where each function $\phi_w(\cdot, \cdot)$, $\phi_s(\cdot, \cdot)$, $\phi_z(\cdot, \cdot)$ is a typical kernel function such as Matérn kernel, RBF kernel, or RQ kernel. Then, the kernel matrix is as follows

$$\mathbf{K}_y = \gamma_0^y \mathbf{J}_{TL} + \gamma_w^y \mathbf{M}_w + \gamma_s^y \mathbf{M}_s + \gamma_z^y \mathbf{M}_z, \qquad \mathbf{K}_y^l = \gamma_0^y \mathbf{J}_{T \times TL} + \gamma_w^y \mathbf{M}_w^l + \gamma_s^y \mathbf{M}_s^l + \gamma_z^y \mathbf{M}_z^l.$$

where $\mathbf{J}_a \in \mathbb{R}^{a\times a}$ and $\mathbf{J}_{a\times b} \in \mathbb{R}^{a\times b}$ are matrices of ones, $\mathbf{M}_w = \mathbf{J}_L \otimes \mathbf{\Lambda}_w$ and $\mathbf{M}_w^l = \mathbf{1}_L^\top \otimes \mathbf{\Lambda}_w$ with $\mathbf{\Lambda}_w \in \mathbb{R}^{T\times T}$ is a kernel matrix computed with kernel function $\phi_w(w_i, w_j)$ and $\mathbf{1}_d$ denotes the d-dimensional column vector of ones, $\otimes$ is the Kronecker product; similarly, $\mathbf{M}_s = \mathbf{J}_L \otimes \mathbf{\Lambda}_s$ and $\mathbf{M}_s^l = \mathbf{1}_L^\top \otimes \mathbf{\Lambda}_s$ with $\mathbf{\Lambda}_s \in \mathbb{R}^{T\times T}$ is a kernel matrix computed with kernel function $\phi_s(\mathbf{s}_i, \mathbf{s}_j)$. We remark that $\mathbf{M}_w, \mathbf{M}_s \in \mathbb{R}^{TL\times TL}$ and $\mathbf{M}_w^l, \mathbf{M}_s^l \in \mathbb{R}^{T\times TL}$. Lastly,

$$\mathbf{M}_z = \begin{bmatrix} \phi_z(\mathbf{z}_1^1, \mathbf{z}_1^1) \dots \phi_z(\mathbf{z}_1^1, \mathbf{z}_T^1) \dots \phi_z(\mathbf{z}_1^1, \mathbf{z}_1^L) \dots \phi_z(\mathbf{z}_1^1, \mathbf{z}_T^L) \\ \vdots \quad \ddots \quad \vdots \quad\quad \vdots \quad \ddots \quad \vdots \\ \phi_z(\mathbf{z}_T^1, \mathbf{z}_1^1) \dots \phi_z(\mathbf{z}_T^1, \mathbf{z}_T^1) \dots \phi_z(\mathbf{z}_T^1, \mathbf{z}_1^L) \dots \phi_z(\mathbf{z}_T^1, \mathbf{z}_T^L) \\ \vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \\ \phi_z(\mathbf{z}_1^L, \mathbf{z}_1^1) \dots \phi_z(\mathbf{z}_1^L, \mathbf{z}_T^1) \dots \phi_z(\mathbf{z}_1^L, \mathbf{z}_1^L) \dots \phi_z(\mathbf{z}_1^L, \mathbf{z}_T^L) \\ \vdots \quad \ddots \quad \vdots \quad\quad \vdots \quad \ddots \quad \vdots \\ \phi_z(\mathbf{z}_T^L, \mathbf{z}_1^1) \dots \phi_z(\mathbf{z}_T^L, \mathbf{z}_T^1) \dots \phi_z(\mathbf{z}_T^L, \mathbf{z}_1^L) \dots \phi_z(\mathbf{z}_T^L, \mathbf{z}_T^L) \end{bmatrix} \in \mathbb{R}^{TL\times TL},$$

$$\mathbf{M}_z^l = \begin{bmatrix} \phi_z(\mathbf{z}_1^l, \mathbf{z}_1^1) \dots \phi_z(\mathbf{z}_1^l, \mathbf{z}_T^1) \dots \phi_z(\mathbf{z}_1^l, \mathbf{z}_1^L) \dots \phi_z(\mathbf{z}_1^l, \mathbf{z}_T^L) \\ \vdots \quad \ddots \quad \vdots \quad\quad \vdots \quad \ddots \quad \vdots \\ \phi_z(\mathbf{z}_T^l, \mathbf{z}_1^1) \dots \phi_z(\mathbf{z}_T^l, \mathbf{z}_T^1) \dots \phi_z(\mathbf{z}_T^l, \mathbf{z}_1^L) \dots \phi_z(\mathbf{z}_T^l, \mathbf{z}_T^L) \end{bmatrix} \in \mathbb{R}^{T\times TL}.$$

Similarly, we have

$$\begin{aligned} \mathbf{K}_w &= \gamma_0^w \mathbf{J}_{TL} + \gamma_s^w \mathbf{M}_s + \gamma_z^w \mathbf{M}_z, & \mathbf{K}_w^l &= \gamma_0^w \mathbf{J}_{T\times TL} + \gamma_s^w \mathbf{M}_s^l + \gamma_z^w \mathbf{M}_z^l \\ \mathbf{K}_x &= \gamma_0^x \mathbf{J}_{TL} + \gamma_s^x \mathbf{M}_s + \gamma_z^x \mathbf{M}_z, & \mathbf{K}_x^l &= \gamma_0^x \mathbf{J}_{T\times TL} + \gamma_s^x \mathbf{M}_s^l + \gamma_z^x \mathbf{M}_z^l \\ \mathbf{K}_z &= \gamma_0^z \mathbf{J}_{TL} + \gamma_z^z \mathbf{M}_z, & \mathbf{K}_z^l &= \gamma_0^z \mathbf{J}_{T\times TL} + \gamma_z^z \mathbf{M}_z^l \\ \mathbf{K}_s &= \gamma_0^s \mathbf{J}_T + \gamma_s^s \mathbf{\Lambda}_s, & \mathbf{K}_q &= \gamma_0^q \mathbf{J}_T + \gamma_y^q \mathbf{\Lambda}_y + \gamma_w^q \mathbf{\Lambda}_w + \gamma_s^q \mathbf{\Lambda}_s + \gamma_x^q \mathbf{\Lambda}_x. \end{aligned}$$

where $\mathbf{M}_s$, $\mathbf{M}_z$, $\mathbf{M}_w$, $\mathbf{M}_s^l$, $\mathbf{M}_z^l$, $\mathbf{M}_w^l$, $\mathbf{\Lambda}_w$, $\mathbf{\Lambda}_s$ are defined as above, and $\mathbf{\Lambda}_x, \mathbf{\Lambda}_y \in \mathbb{R}^{T\times T}$ are kernel matrices computed with kernel functions $\phi_x(\mathbf{x}_i, \mathbf{x}_j)$ and $\phi_y(y_i, y_j)$, respectively.

Substitute the above kernel matrices to the objective function in Section 1, we obtain

$$\begin{aligned} J = \frac{1}{L}\sum_{l=1}^{L} \Bigg(& \frac{1}{2\sigma_y^2}((\boldsymbol{\beta}^y)^\top \left[\gamma_0^y \mathbf{J}_{T\times TL} + \gamma_w^y \mathbf{M}_w^l + \gamma_s^y \mathbf{M}_s^l + \gamma_z^y \mathbf{M}_z^l\right]^\top [\gamma_0^y \mathbf{J}_{T\times TL} + \gamma_w^y \mathbf{M}_w^l \\ & + \gamma_s^y \mathbf{M}_s^l + \gamma_z^y \mathbf{M}_z^l]\boldsymbol{\beta}^y - 2\mathbf{y}^\top [\gamma_0^y \mathbf{J}_{T\times TL} + \gamma_w^y \mathbf{M}_w^l + \gamma_s^y \mathbf{M}_s^l + \gamma_z^y \mathbf{M}_z^l]\boldsymbol{\beta}^y + \mathbf{y}^\top \mathbf{y}) + T\log\sigma_y \\ & - \mathbf{w}^\top \log\varphi([\gamma_0^w \mathbf{J}_{T\times TL} + \gamma_s^w \mathbf{M}_s^l + \gamma_z^w \mathbf{M}_z^l]\boldsymbol{\beta}^w) \\ & - (\mathbf{1} - \mathbf{w})^\top \log\varphi(-[\gamma_0^w \mathbf{J}_{T\times TL} + \gamma_s^w \mathbf{M}_s^l + \gamma_z^w \mathbf{M}_z^l]\boldsymbol{\beta}^w) \\ & + \frac{1}{2\sigma_s^2}\sum_{d=1}^{d_s}((\boldsymbol{\beta}_d^s)^\top [\gamma_0^s \mathbf{J}_{T\times TL} + \gamma_s^s \mathbf{\Lambda}_s]^\top [\gamma_0^s \mathbf{J}_{T\times TL} + \gamma_s^s \mathbf{\Lambda}_s]\boldsymbol{\beta}_d^s \\ & - 2\mathbf{s}_{:,d}^\top [\gamma_0^s \mathbf{J}_{T\times TL} + \gamma_s^s \mathbf{\Lambda}_s]\boldsymbol{\beta}_d^s + \mathbf{s}_{:,d}^\top \mathbf{s}_{:,d}) + Td_s \log\sigma_s \\ & + \frac{1}{2\sigma_x^2}\sum_{d=1}^{d_x}((\boldsymbol{\beta}_d^x)^\top [\gamma_0^x \mathbf{J}_{T\times TL} + \gamma_s^x \mathbf{M}_s^l + \gamma_z^x \mathbf{M}_z^l]^\top [\gamma_0^x \mathbf{J}_{T\times TL} + \gamma_s^x \mathbf{M}_s^l + \gamma_z^x \mathbf{M}_z^l]\boldsymbol{\beta}_d^x \\ & - 2\mathbf{x}_{:,d}^\top [\gamma_0^x \mathbf{J}_{T\times TL} + \gamma_s^x \mathbf{M}_s^l + \gamma_z^x \mathbf{M}_z^l]\boldsymbol{\beta}_d^x + \mathbf{x}_{:,d}^\top \mathbf{x}_{:,d}) + Td_x \log\sigma_x \\ & + \frac{1}{2\sigma_z^2}\sum_{d=1}^{d_z}((\boldsymbol{\beta}_d^z)^\top [\gamma_0^z \mathbf{J}_{T\times TL} + \gamma_z^z \mathbf{M}_z^l]^\top [\gamma_0^z \mathbf{J}_{T\times TL} + \gamma_z^z \mathbf{M}_z^l]\boldsymbol{\beta}_d^z \\ & - 2(\boldsymbol{\beta}_d^q)^\top [\gamma_0^q \mathbf{J}_T + \gamma_y^q \mathbf{\Lambda}_y + \gamma_w^q \mathbf{\Lambda}_w + \gamma_s^q \mathbf{\Lambda}_s + \gamma_x^q \mathbf{\Lambda}_x]^\top [\gamma_0^z \mathbf{J}_{T\times TL} + \gamma_z^z \mathbf{M}_z^l]\boldsymbol{\beta}_d^z \\ & + (\boldsymbol{\beta}_d^q)^\top [\gamma_0^q \mathbf{J}_T + \gamma_y^q \mathbf{\Lambda}_y + \gamma_w^q \mathbf{\Lambda}_w + \gamma_s^q \mathbf{\Lambda}_s + \gamma_x^q \mathbf{\Lambda}_x]^\top [\gamma_0^q \mathbf{J}_T + \gamma_y^q \mathbf{\Lambda}_y \\ & + \gamma_w^q \mathbf{\Lambda}_w + \gamma_s^q \mathbf{\Lambda}_s + \gamma_x^q \mathbf{\Lambda}_x]\boldsymbol{\beta}_d^q) + Td_z \log(\sigma_z/\sigma_q) + \frac{1}{2}Td_z\sigma_q^2/\sigma_z^2 \Bigg) \end{aligned}$$

$$+ \lambda_y(\boldsymbol{\beta}^y)^\top[\gamma_0^y\mathbf{J}_{TL} + \gamma_w^y\mathbf{M}_w + \gamma_s^y\mathbf{M}_s + \gamma_z^y\mathbf{M}_z]\boldsymbol{\beta}^y + \lambda_w(\boldsymbol{\beta}^w)^\top[\gamma_0^w\mathbf{J}_{TL} + \gamma_s^w\mathbf{M}_s + \gamma_z^w\mathbf{M}_z]\boldsymbol{\beta}^w$$
$$+ \sum_{d=1}^{d_x}\lambda_x(\boldsymbol{\beta}_d^x)^\top[\gamma_0^x\mathbf{J}_{TL} + \gamma_s^x\mathbf{M}_s + \gamma_z^x\mathbf{M}_z]\boldsymbol{\beta}_d^x + \sum_{d=1}^{d_s}\lambda_s(\boldsymbol{\beta}_d^s)^\top[\gamma_0^s\mathbf{J}_T + \gamma_s^s\boldsymbol{\Lambda}_s]\boldsymbol{\beta}_d^s$$
$$+ \sum_{d=1}^{d_z}\lambda_z(\boldsymbol{\beta}_d^z)^\top[\gamma_0^z\mathbf{J}_{TL} + \gamma_z^z\mathbf{M}_z]\boldsymbol{\beta}_d^z + \sum_{d=1}^{d_z}\lambda_q(\boldsymbol{\beta}_d^q)^\top[\gamma_0^q\mathbf{J}_T + \gamma_y^q\boldsymbol{\Lambda}_y + \gamma_w^q\boldsymbol{\Lambda}_w + \gamma_s^q\boldsymbol{\Lambda}_s + \gamma_x^q\boldsymbol{\Lambda}_x]\boldsymbol{\beta}_d^q.$$

Taking derivative of J with respect to γ_z^y, γ_w^y, γ_s^y, γ_0^y γ_z^x, γ_s^x, γ_0^x, γ_z^z, γ_0^z, γ_s^s, γ_0^s and equates to 0 to solve for the solution of these parameters.

Remark 1. *By using the above additive kernel functions, the form of latent confounders is as follows*

$$\mathbf{z}_t = \gamma_0^z(\boldsymbol{\beta}^z)^\top\mathbf{1}_{TL} + \sum_{j=1}^{T}\sum_{l=1}^{L}\gamma_z^z((\boldsymbol{\beta}^z)^\top)_{:,jl}\phi(\mathbf{z}_{t-1}, \mathbf{z}_j^l) + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_{d_z}),$$

where $((\boldsymbol{\beta}^z)^\top)_{:,jl}$ denotes the jl-th column of $(\boldsymbol{\beta}^z)^\top$. Thus the form of $\boldsymbol{\alpha}_0$, $\boldsymbol{\alpha}_{jl}$ and $k_z(\cdot,\cdot)$ stated in the synthetic experiments (Section 5.2 in the main text) are $\boldsymbol{\alpha}_0 = \gamma_0^z(\boldsymbol{\beta}^z)^\top\mathbf{1}_{TL}$, $\boldsymbol{\alpha}_{jl} = \gamma_z^z((\boldsymbol{\beta}^z)^\top)_{:,jl}$, and $k_z(\cdot,\cdot) = \phi(\mathbf{z}_{t-1}, \mathbf{z}_j^l)$.

5 SIMULATING SYNTHETIC DATA

In this section, we present how the synthetic datasets were generated in Section 5.2 of the main text. To simulate the data, we use the following Eqs. (2)-(6).

$$\mathbf{z}_t = f_z(\mathbf{z}_{t-1}) + \boldsymbol{e}_t, \qquad \boldsymbol{e}_t \sim \mathsf{N}(\mathbf{0}, \sigma_z^2\mathbf{I}_{d_z}), \tag{2}$$
$$\mathbf{s}_t = f_s(\mathbf{s}_{t-1}) + \boldsymbol{o}_t, \qquad \boldsymbol{o}_t \sim \mathsf{N}(\mathbf{0}, \sigma_s^2\mathbf{I}_{d_s}), \tag{3}$$
$$\mathbf{x}_t = f_x(\mathbf{z}_t, \mathbf{s}_t) + \boldsymbol{r}_t, \qquad \boldsymbol{r}_t \sim \mathsf{N}(\mathbf{0}, \sigma_x^2\mathbf{I}_{d_x}), \tag{4}$$
$$y_t = f_y(w_t, \mathbf{z}_t, \mathbf{s}_t) + v_t, \qquad v_t \sim \mathsf{N}(0, \sigma_y^2), \tag{5}$$
$$w_t = \mathbb{1}(\varphi(f_w(\mathbf{z}_t, \mathbf{s}_t)) \geq u_t), \qquad u_t \sim \mathsf{U}[0,1], \tag{6}$$

where we set the ground truth hyperparameters $\sigma_z, \sigma_s, \sigma_x, \sigma_y$ to 1 and the dimension of $\mathbf{s}_t, \mathbf{z}_t, \mathbf{x}_t$ are 2. We specify the ground truth for each of the functions f_z, f_s, f_x, f_y, f_w as neural networks whose specification are as follows:

- For $f_z(\mathbf{z}_{t-1})$: Input (2 dims) $\to \{$Hidden layer i (10 dims) $\to$ LeakyReLU$\}_{i=1}^{j} \to$ Output (2 dim),
- For $f_s(\mathbf{s}_{t-1})$: Input (2 dims) $\to \{$Hidden layer i (10 dims) $\to$ LeakyReLU$\}_{i=1}^{j} \to$ Output (2 dim),
- For $f_x(\mathbf{z}_t, \mathbf{s}_t)$: Input (4 dims) $\to \{$Hidden layer i (10 dims) $\to$ LeakyReLU$\}_{i=1}^{j} \to$ Output (2 dim),
- For $f_w(\mathbf{z}_t, \mathbf{s}_t)$: Input (4 dims) $\to \{$Hidden layer i (10 dims) $\to$ LeakyReLU$\}_{i=1}^{j} \to$ Output (1 dims),
- For $f_y(\mathbf{z}_t, w_t, \mathbf{s}_t)$: We set the ground truth $f_y(\mathbf{z}_t, w_t, \mathbf{s}_t) = (1 - w_t)f_{y_0}(\mathbf{z}_t, \mathbf{s}_t) + w_t f_{y_1}(\mathbf{z}_t, \mathbf{s}_t)$, where $f_{y_0}(\mathbf{z}_t, \mathbf{s}_t)$ and $f_{y_1}(\mathbf{z}_t, \mathbf{s}_t)$ are two different neural networks
  - $f_{y_0}(\mathbf{z}_t, \mathbf{s}_t)$: Input (4 dims) $\to \{$Hidden layer i (10 dims) $\to$ LeakyReLU$\}_{i=1}^{j} \to$ Output (1 dims),
  - $f_{y_1}(\mathbf{z}_t, \mathbf{s}_t)$: Input (4 dims) $\to \{$Hidden layer i (10 dims) $\to$ LeakyReLU$\}_{i=1}^{j} \to$ Output (1 dims),

where j in the above specification is the number of hidden layers. Using different numbers of the hidden layers, i.e., $j = 2, 4, 6$, we construct three synthetic datasets, namely TD2L, TD4L, and TD6L. For these three datasets, we sample the latent confounder variable $\mathbf{z}$ satisfying Eq. (2).

We also construct another dataset, TD6L-iid, that uses six hidden layers but with the *iid* latent confounder variable, i.e., $\mathbf{z}_a \perp\!\!\!\perp \mathbf{z}_b$, and $\mathbf{s}_a \perp\!\!\!\perp \mathbf{s}_b$, $\forall a, b \in \{1, 2, ..., T\}$. In particular, the simulation of $\mathbf{z}_t$ and $\mathbf{s}_t$ in TD6L-iid was done using the following Eqs. (7) and (8)

$$p(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathsf{N}(\mathbf{z}_t|\mu_z, \sigma_z^2\mathbf{I}_{d_z}), \tag{7}$$
$$p(\mathbf{s}_t|\mathbf{s}_{t-1}) = \mathsf{N}(\mathbf{s}_t|\mu_s, \sigma_s^2\mathbf{I}_{d_s}), \tag{8}$$

where we set the ground truth $\mu_z = \mu_s = 1$. Simulation of the other variables remain the same, i.e, we use Eqs. (4)-(6) for simulation of $\mathbf{x}_t$, y_t and w_t.

Finally, for all the simulation, we discard $\mathbf{z}_t$ and only keep y_t, w_t, $\mathbf{x}_t$ and $\mathbf{s}_t$ as observed data for training the models. We summarize the steps to simulate the datasets in Algorithm 1. In each step of Algorithm 1, we also save $f_{y_0}(\mathbf{z}_t, \mathbf{s}_t)$ and $f_{y_1}(\mathbf{z}_t, \mathbf{s}_t)$ for analysing the performance of each method.

Algorithm 1: Simulate synthetic dataset

begin
  $\mathbf{z}_0 \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_2)$;
  $\mathbf{s}_0 \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_2)$;
  for $t := 1$ *to* T **do**
    Sample $\mathbf{z}_t$ using Eq. (2) for `DT`j`L` or Eq. (7) for `DT6L-iid`;
    Sample $\mathbf{s}_t$ using using Eq. (3) for `DT`j`L` or Eq. (8) for `DT6L-iid`;
    Sample $\mathbf{x}_t$ using Eq. (4);
    Sample w_t using Eq. (6);
    Sample y_t using Eq. (5);
  return $\mathbf{w}, \mathbf{y}, \mathbf{x}, \mathbf{s}$;

— End —