APPENDIX

Parameters Optimisation

For the proposed method where Gaussian radial basis function (RBF) kernels are used on all three random variables of interest, there are two groups of hyper-parameters. The first one includes the kernel bandwidths for the kernel matrices K, L and M. The bandwidth for K and L are fixed by using median heuristic. Although we also fixed the bandwidth for M by median heuristic in the main text, here we include a discussion on the optimisation of such parameter when one has a different kernel bandwidth for the regression on the feature space of X and on that of Y. Let's call them σ^2_{xz} and σ^2_{yz} respectively. The second group contains the regularisation parameters $\lambda^{\mathbf{x}}$ and $\lambda^{\mathbf{y}}$ for the kernel ridge regressions. These parameters can be optimised through grid search or gradient descent on the total cross-validation error as follows.

For a K-fold cross-validation, we aim to minimise the sum of the square of the residuals of each unseen fold given the model trained on the other $K-1$ folds over some (λ, σ^2) values. Here, we present detailed derivation for X and that of Y follows similarly. For a pair of fixed $(\lambda^x, \sigma^2_{xz})$, the cross-validation error of the k^{th} fold is:

$$CVerror^k_\lambda = \frac{1}{u}\sum_{i\in tst} \|\phi(x^{tst}_i) - \Phi^T_{tr}\beta^x(z^{tst}_i)\|^2 \quad (23)$$

where $\{x^{tr}_i, z^{tr}_i\}_i$ and $\{x^{tst}_i, z^{tst}_i\}_i$ denote respectively the training and testing sets for the k^{th} fold, $\Phi_{tr} = (\phi(x^{tr}_1), \phi(x^{tr}_2), ...)^\top$ and u is the size of the testing set. Then, the overall cross-validation error is

$$CVerror_\lambda = \frac{1}{K}\sum_{k=1}^{K} CVerror^k_\lambda \,. \quad (24)$$

To simplify (23), we note that

$$\begin{aligned}
&\sum_{i\in tst} \|\phi(x^{tst}_i) - \Phi^T_{tr}\beta^x(z^{tst}_i)\|^2 && (25)\\
=&\sum_{i\in tst} \|\phi(x^{tst}_i) - \Phi^T_{tr}(M_{tr,tr}+\lambda^x I)^{-1} \\
&(m(z^{tst}_i, z^{tr}_1), m(z^{tst}_i, z^{tr}_2), ..., m(z^{tst}_i, z^{tr}_m))^T\|^2 && (26)\\
=&Tr([\Phi_{tst} - \mathcal{M}_z][\Phi_{tst} - \mathcal{M}_z]^T) && (27)\\
=&Tr(\Phi_{tst}\Phi^T_{tst} - \Phi_{tst}\mathcal{M}^T_z - \mathcal{M}_z\Phi^T_{tst} + \mathcal{M}_z\mathcal{M}^T_z) && (28)
\end{aligned}$$

where $\mathcal{M}_z = M_{tst,tr}(M_{tr,tr}+\lambda^x I)^{-1}\Phi_{tr}$. Such prediction error for the k^{th} fold (i.e. (28)) can be explicitly formulated in terms of $(\lambda^x, \sigma^2_{xz})$ by writing the gram matrices in terms of square distance matrices:

$$M_{*,\cdot} = \exp(-0.5\, D_{*,\cdot}/\sigma^2_{xz}) \quad (29)$$

where $D_{*,\cdot}$ is the square distance matrix computed on the respective sets, i.e. $(D_{*,\cdot})_{ij} = \|z_i - z_j\|^2_2$. Note, no explicit feature map representation is required.

Since both parameters are constrained to be positive, we can rewrite $\lambda^x = \exp(\eta)$ and $\sigma^2_{xz} = \exp(\gamma)$ and take first order derivate with respect to η and γ to obtain a gradient descent algorithm. Given the expression in (24), taking the first order derivative with respect to η (Appendix D of Dattorro [2016]):

$$\begin{aligned}
&\frac{d}{d\eta}Error_{(\lambda^x, \sigma^2_{xz})}\\
&\quad= 2\,Tr\{K_{tst,tr}\mathcal{M}_\eta M_{tr,tst}\\
&\qquad - M_{tst,tr}\mathcal{M}_\eta K_{tr,tr}\mathcal{M}_{inv}M_{tr,tst}\}
\end{aligned}$$

where $\mathcal{M}_\eta := (M_{tr,tr}+\lambda^x I)^{-1}[\exp(\eta)I](M_{tr,tr}+\lambda^x I)^{-1}$ and $\mathcal{M}_{inv} := (M_{tr,tr}+\lambda^x I)^{-1}$. Similarly, take the first order derivative with respect to γ:

$$\begin{aligned}
&\frac{d}{d\gamma}Error_{(\lambda^x, \sigma^2_{xz})}\\
&\quad= Tr\{2\,K_{tst,tr}\mathcal{M}_{inv}\mathcal{D}_{tr,tr}\mathcal{M}_{inv}M_{tr,tst}\\
&\qquad - 2\,K_{tst,tr}\mathcal{M}_{inv}\mathcal{D}_{tr,tst}\\
&\qquad + \mathcal{D}_{tst,tr}\mathcal{M}_{inv}K_{tr,tr}\mathcal{M}_{inv}M_{tr,tst}\\
&\qquad - M_{tst,tr}\mathcal{M}_{inv}\mathcal{D}_{tr,tr}\mathcal{M}_{inv}K_{tr,tr}\mathcal{M}_{inv}M_{tr,tst}\\
&\qquad - M_{tst,tr}\mathcal{M}_{inv}K_{tr,tr}\mathcal{M}_{inv}\mathcal{D}_{tr,tr}\mathcal{M}_{inv}M_{tr,tst}\\
&\qquad + M_{tst,tr}\mathcal{M}_{inv}K_{tr,tr}\mathcal{M}_{inv}\mathcal{D}_{tr,tst}\} \qquad (30)
\end{aligned}$$

where

$$\mathcal{D}_{tr,\cdot} := [0.5D_{tr,\cdot} \times \exp(-0.5D_{tr,\cdot}\exp(-\gamma))\exp(-\gamma)I].$$

At iteration i, we update η and γ according to

$$\eta_i = \eta_{i-1} - \alpha\frac{d}{d\eta}Error_{(\lambda^x, \sigma^2_{xz})}\Big|_{\eta_{i-1},\gamma_{i-1}} \quad (31)$$

$$\gamma_i = \gamma_{i-1} - \alpha\frac{d}{d\gamma}Error_{(\lambda^x, \sigma^2_{xz})}\Big|_{\eta_{i-1},\gamma_{i-1}} \quad (32)$$

with learning rate α until convergence. Various adaptive learning rate schemes can be used for faster convergence.

Boston Housing Data

The main variable of interest is the median value of occupied homes (medv). Based on the result using KRESIT (Figure 6), the proportion of lower status people (lstat) and (medv) are conditionally dependent given the rest of the variables, so is average number of rooms per unit (rm) and (medv). However, the estimated CPDAG cannot be more specific on the direction of the causation. Additionally, (medv) has been identified to cause the percentage of

black population (b). We observe that such link is indeed plausible. In the US around the time of the 1970 census, residential segregation was common place, and neighbourhoods with a higher percentage of blacks were regarded as less desirable to live in. Banks and the federal government itself refused to give loans to blacks in general, and specifically refused to give them loans when they wanted to buy homes in neighbourhoods that had historically been white.

The variable (medv) has also been identified to cause whether the house is next to the Charles River (chas). However, this edge seems to be misdirected. The result of Flaxman et al. [2015] (using RESIT) has shown the opposite (chas) $\rightarrow$ (medv) which agrees with authors' belief in the original paper [Harrison and Rubinfeld, 1978]. Besides this, they also showed that (nox), (lstat) and (rm) are the causes for (medv). While the strong independence test of Zhang et al. [2011] also identified (lstat) and (rm) as the causes, it additionally concludes crime rates (crim) and proportion of unit built before 1940 (age) as causes. The strong CI test of Fukumizu et al. [2008] agrees with Zhang et al. [2011] on (lstat) and (age).

Ozone Data

We consider the ozone dataset used in Breiman and Friedman [1985] and pre-whitened each variable using time with a GP regression [Flaxman et al., 2015] as the dataset consists of 330 daily observations with clear temporal autocorrelation. We refer to [Breiman and Friedman, 1985] for the details of each variable and the main variable of interest is the ozone concentration (Ozone). We set the significance level to be $\alpha = 0.05$ as the number of variables is only 9.

Figure 7 illustrates that though some edges are present in both our result and the one obtained by Flaxman et al. [2015], the two graphs are quite different. The CPDAG obtained using KRESIT shows that there is no cause of (Ozone) among the variables under consideration, but (Ozone) causes temperature (Temp), humidity (Hum) and inversion base height (InvHt). Flaxman et al. [2015] also concludes that there is no cause of (Ozone) among the variables under consideration, but (Ozone) is one of the causes for visibility (Vis).

The correlations between the concentration of ozone, temperature, inversion base height and humidity in the atmosphere have been shown to exist in many research on meteorology, public health and environmental science. However, the causal directions are less clear as the interaction between them are complex and there are additional factors that plays a part which have not been considered in this dataset. For example, the ratio of volatile organic compounds (VOC) and nitrogen oxides (Nox), the concentration of other green house gases, if heavy cloud is present, etc. We regard our method as promising as it discovers the links between these variables.

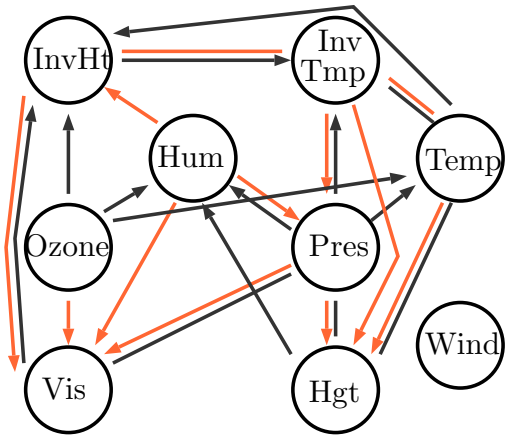

Figure 7: Results of PC algorithm with KRESIT (black) and RESIT [Flaxman et al., 2015] (orange) using pre-whitened Ozone dataset.