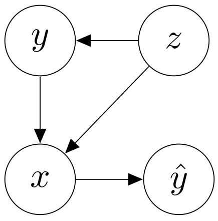

1 Causal Explanation

We assume some simplified structured causal model $S = \{X_j := f_j(PA_j, U_j)\}$ and noise distribution $U_j \sim P_{\mathsf{u}_j}$ that induces a joint distribution P_X over graph $\mathcal{G}$. Assume mechanisms lie in some parameterized family of functions $f_j^\theta \in \mathcal{F} := \left\{ f^\theta \mid \theta \in \Theta \right\}$ and that we have access to i.i.d. samples from the graph, in particular we have $\left\{ (x_j^{(i)}, pa_j^{(i)}) \right\}_{i=1}^n$. we can estimate parameters of the mechanism f_j using maximum likelihood,

$$\hat{\theta}_j = \arg\max_{\theta \in \Theta} \prod_{i=1}^n p_{\mathsf{x}_j|\mathsf{pa}_j}(x_j^{(i)} \mid pa_j^{(i)}) = \arg\max_{\theta \in \Theta} \prod_{i=1}^n \int p_{\mathsf{u}_j}(u_j) \delta \left(x_j^{(i)} - f_j^\theta(pa_j^{(i)}, u_j) \right) du_j \quad (1)$$

The form of $p_{\mathsf{x}_j|\mathsf{pa}_j}$ depends on the family of mechanisms $\mathcal{F}$. We can write out explicit forms for the conditional density if $\mathcal{F}$ is simple, for example in the case of Gaussian additive noise model, the conditional density is a translated Gaussian distribution. If $\mathcal{F}$ is parameterized by arbitrary function, i.e. a neural network, and that X_j is high dimensional, we can model the conditional density using a latent variable model, such as a conditional VAE.

$$p_{\mathsf{x}_j|\mathsf{pa}_j}(x_j|pa_j) = \int p_{\mathsf{x}_j|\mathsf{pa}_j,\mathsf{z}}(x_j|pa_j, z) p_{\mathsf{z}}(z) dz \qquad \mathsf{z} \sim \mathcal{N}(0, I) \quad (2)$$

For arbitrary function family $\mathcal{F}$, structure identifiability cannot be guaranteeed. So we have assumed that the graph structure reflects domain specific knowledge and that by learning parameters of the mechanisms, we captured the causal relationship between potential confounders z and x as input image to the classifier h. We can then use the model as a tool for evaluating if the the classifier uses unstable features. Given a particular instance x, we are interested in the presence of un-modeled confounders. We can visualize the effect of these confounders by holding values for the endogeneous noise fixed while search for soft interventions u such that the counterfactual outcome on classifier output $\hat{y}(u)$ flips sign with high probability.

$$\min_{p_{\mathsf{u}} \in \mathcal{P}_{\mathsf{u}}} D\left(p_{\mathsf{u}|\mathsf{x}}(\cdot|x), p_{\mathsf{u}}(\cdot) \right) \qquad \text{s.t. } \mathbb{E}\left[\mathbb{1}\left[\hat{y}_x(u) \neq \hat{y}_x \right] \right] > 0.5 \quad (3)$$