Deep Learning Project:
Causal Representation Learning via WAE

노요한* 이현종* 정승필†

November 11, 2022

1 Proof of Theorem 1

Theorem 1. *Let $d(x, y) = \|x - y\|_2$ for $x, y \in \mathcal{X}$. If P_X has a density with respect to the Lebesgue measure, and the measurable function $g : \mathcal{Z} \to \mathcal{X}$ is injective, then*

$$W_2^2(P_X, g_\sharp P_Z) = \inf_{f \in \mathcal{Q}} \mathbb{E}_{P_X} d^2(X, g(f(X))), \tag{1}$$

where $\mathcal{Q}$ is the set of all measurable functions from $\mathcal{X}$ to $\mathcal{Z}$ such that $f_\sharp P_X = P_Z$.

Proof. Let $P_G = g_\sharp P_Z$. Under the conditions of the theorem, the Monge-Kantorovich equivalence holds:

$$W_2^2(P_X, P_G) = \inf_{T:\mathcal{X}\to\mathcal{X}:T_\sharp P_X = P_G} \mathbb{E}_{P_X} d^2(X, T(X)).$$

Hence it suffices to show that

$$\inf_{f:\mathcal{X}\to\mathcal{Z}:f_\sharp P_X = P_Z} \int_{\mathcal{X}} d^2(x, g(f(x))) dP_X = \inf_{T:\mathcal{X}\to\mathcal{X}:T_\sharp P_X = P_G} \int_{\mathcal{X}} d^2(x, T(x)) dP_X$$

or equivalently

$$\{g \circ f : f : \mathcal{X} \to \mathcal{Z}, f_\sharp P_X = P_Z\} = \{T : \mathcal{X} \to \mathcal{X} : T_\sharp P_X = P_G\}.$$

First,

$$\{g \circ f : f : \mathcal{X} \to \mathcal{Z}, f_\sharp P_X = P_Z\} \subset \{T : \mathcal{X} \to \mathcal{X} : T_\sharp P_X = P_G\}$$

since for any measurable $f : \mathcal{X} \to \mathcal{Z}$ such that $f_\sharp P_X = P_X f^{-1} = P_Z$ we have $g \circ f : \mathcal{X} \to \mathcal{X}$ measurable and for any Borel set $E \subset \mathcal{X}$

$$(g \circ f)_\sharp P_X(E) = P_X(g \circ f)^{-1}(E) = P_X(f^{-1} \circ g^{-1})(E) = P_Z(g^{-1}(E)) = g_\sharp P_Z(E) = P_G(E).$$

Second, suppose $T : \mathcal{X} \to \mathcal{X}$ is measurable and satisfies $T_\sharp P_X = P_G$. There exists a set $A \subset \mathcal{X}$ with $P_X(A) = 1$ such that $g : \mathcal{Z} \to T(A)$ is surjective. Otherwise, there exists $B \subset \mathcal{X}$ with $P_X(B) > 0$ and $\tilde{g}^{-1}(T(B)) = \emptyset$, and hence $0 = \tilde{g}_\sharp P_Z(T(B)) = T_\sharp P_X(T(B)) = P_X(B) > 0$ that is a contradiction. In addition, since $g : \mathcal{Z} \to \mathcal{X}$ is injective, there exists a left inverse $g^\dagger : \mathcal{X} \to \mathcal{Z}$. Note that $g^\dagger|_{T(A)} : T(A) \to \mathcal{Z}$ is an inverse function of $g : \mathcal{Z} \to T(A)$. Let $f = g^\dagger \circ T$. Then $g \circ f = g \circ g^\dagger \circ T = T$ almost surely in P_X and also for any Borel set $F \subset \mathcal{Z}$

$$\begin{aligned} f_\sharp P_X(F) &= P_X(g^\dagger \circ T)^{-1}(F) \\ &= P_X(T^{-1}(g^\dagger)^{-1}(F)) \\ &= P_G((g^\dagger)^{-1}(F)) \\ &= P_Z(g^{-1}((g^\dagger)^{-1}(F))) \\ &= P_Z((g^\dagger \circ g)^{-1}(F)) = P_Z(F). \end{aligned}$$

*Department of Statistics, SNU
†Graduate School of Public Health, SNU

Therefore,

$$\{g \circ f : f : \mathcal{X} \to \mathcal{Z}, f_\sharp P_X = P_Z\} \supset \{T : \mathcal{X} \to \mathcal{X} : T_\sharp P_X = P_G\}$$

□

2 Decoder of CausalVAE

[1]의 decoder 구조를 살펴보면, p개의 특성을 나타내는 causal representation $\mathbf{z} \in \mathbb{R}^p$를 생각하자. [1]에서는 다음과 같은 linear Structured Causal Model(SCM)을 고려하고 있다.

$$\mathbf{z} = \mathbf{A}^T\mathbf{z} + \boldsymbol{\epsilon} = (\mathbf{I} - \mathbf{A}^T)^{-1}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{2}$$

여기서, $\mathbf{A} \in \mathbb{R}^{p\times p}$는 특성 사이의 directed acyclic graph(DAG) 구조를 나타내는 가중치 인접행렬을 나타낸다. 행렬 $\mathbf{A}$는 모른다고 가정하고 학습되는 파라미터이다. SCM을 고려하여 Causal representation, $\mathbf{z}$를 학습하기 위해 Figure 1과 같은 generative model을 제시하였다.

$$p_\theta(\mathbf{x}, \mathbf{z}, \boldsymbol{\epsilon}|\mathbf{u}) = p_\theta(\mathbf{x}|\mathbf{z}, \boldsymbol{\epsilon}, \mathbf{u})p_\theta(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{u})$$
$$p_\theta(\mathbf{x}|\mathbf{z}, \boldsymbol{\epsilon}, \mathbf{u}) = p_\theta(\mathbf{x}|\mathbf{z}), \quad p_\theta(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{u}) = p(\boldsymbol{\epsilon})p_\theta(\mathbf{z}|\mathbf{u})$$

이를 정리하면,

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad z_{1i}|u_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\lambda_1(u_i), \lambda_2^2(u_i)),\ i = 1, \dots, p$$
$$\mathbf{z}_2 = g(\mathbf{A}^T\mathbf{z}_1) + \boldsymbol{\epsilon} \approx \mathbf{z}_1, \quad \mathbf{x} \sim p(\mathbf{x}|\mathbf{z}_2)$$

g는 Mask layer [2]를 의미하고, $\mathbf{u}$는 실제 얼굴 사진의 attribute을 나타내는 범주형 변수이며 모델의 identifiability를 보장하기 위해 도입된다([1, Theorem 1]).

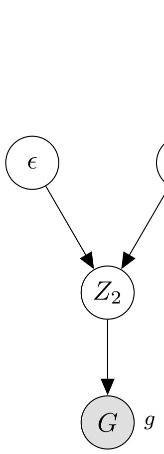

Figure 1: CausalVAE Decoder Structure

앞서 특성 $\mathbf{z}$ 사이의 관계를 나타내는 DAG의 인접행렬 $\mathbf{A}$는 학습되는 파라미터라고 했는데, $\mathbf{A}$는 다음과 같은 제약 조건을 만족하도록 학습된다.

$$\mathbb{E}_{P_X}\|\mathbf{u} - \sigma(\mathbf{A}^T\mathbf{u})\|_2^2 \le \kappa \tag{3}$$
$$tr\left[\left(\mathbf{I} + \frac{c}{p}\mathbf{A} \circ \mathbf{A}\right)^p\right] - p = 0 \tag{4}$$

σ는 logit 함수를 나타내고, κ는 충분히 작은 양수, c는 임의의 양수를 나타낸다. Equation 3는 identifiability를 보장하고, 특성 사이의 관계를 $\mathbf{A}$가 나타내도록 하는 조건이고, equation 4는 DAG의 가중치 인접행렬이 가져야하는 조건을 의미한다.

3 Applying Theorem 1 to the CausalVAE

Injective한 decoder를 생각하기 위해 $\mathbf{u}$를 도입하자. Figure 1으로부터 decoder $g : \mathcal{Z} \times \mathcal{U} \to \mathcal{X}$를 다음과 같이 정의하자.

$$g(\boldsymbol{\epsilon}, \mathbf{u}) = g_3(g_2(\mathbf{A}^T g_1(\mathbf{u})) + \boldsymbol{\epsilon})$$

여기서, $g_1 : \mathcal{U} \to \mathcal{Z}$, $g_3 : \mathcal{Z} \to \mathcal{X}$는 각각 neural network를 나타내고, $g_2 : \mathcal{Z} \to \mathcal{Z}$는 mask layer[2]이다. 특히 g_3는 leakyReLU 또는 sigmoid와 같은 injective activation을 사용한 neural network를 가정한다. Decoder g를 확장시켜 $\tilde{g} : \mathcal{Z} \times \mathcal{U} \to \mathcal{X} \times \mathcal{U}$, $\tilde{g}(\boldsymbol{\epsilon}, \mathbf{u}) = (g(\boldsymbol{\epsilon}, \mathbf{u}), \mathbf{u})$를 생각하면, $\tilde{g}$는 단사 함수(injective function)이다.

$\tilde{g}$에 대해 Theorem 1을 적용하기 위해 $\mathcal{U}$ 공간에 거리 d'을 생각하자. 이때, $\mathcal{X} \times \mathcal{U}$ 공간에서의 거리를 $\tilde{d} = \sqrt{d^2 + d'^2}$로 정의하자. 이제 P_{XU}와 $P_{\boldsymbol{\epsilon}} \otimes P_U$, $\tilde{g}$에 대해 Theorem 1을 적용하면,

$$\begin{aligned} W_2^2(P_{XU}, \tilde{g}_\sharp(P_{\boldsymbol{\epsilon}} \otimes P_U)) &= \inf_{\tilde{f} \in \tilde{\mathcal{F}}} \mathbb{E}_{P_{XU}} \tilde{d}^2 \left(\begin{pmatrix} X \\ U \end{pmatrix}, \tilde{g}(\tilde{f}(X, U)) \right) \\ &= \inf_{f \in \mathcal{F}} \mathbb{E}_{P_{XU}} \tilde{d}^2 \left(\begin{pmatrix} X \\ U \end{pmatrix}, \begin{pmatrix} g_3(g_2(\mathbf{A}^T g_1(U)) + f(X, U)) \\ U \end{pmatrix} \right) \\ &= \inf_{f \in \mathcal{F}} \mathbb{E}_{P_{XU}} d^2(X, g_3(g_2(\mathbf{A}^T g_1(U)) + f(X, U))), \end{aligned}$$

$\tilde{\mathcal{F}} = \{\tilde{f} : \mathcal{X} \times \mathcal{U} \to \mathcal{Z} \times \mathcal{U} | \tilde{f}_\sharp P_{XU} = P_{\boldsymbol{\epsilon}} \otimes P_U)\}$, $\mathcal{F} = \{f : \mathcal{X} \times \mathcal{U} \to \mathcal{Z} | (f, \Pi_U)_\sharp P_{XU} = P_{\boldsymbol{\epsilon}} \otimes P_U\}$, $\Pi_U(X, U) = U$를 의미한다. 두 번째 등호는 $\tilde{f}(X, U) = (f(X, U), \Pi_U(X, U))$를 생각하면 성립함을 알 수 있다. $\mathcal{F}$의 조건은

$$f(X, U) \overset{d}{=} \boldsymbol{\epsilon}, \quad f(X, U) \perp\!\!\!\perp U,$$

2가지 조건으로 생각할 수 있다.

두 분포가 얼마나 다른지 측정하는 $\mathcal{D}$와 독립성을 측정하는 $\mathcal{H}$에 대해 위의 2가지 조건을 패널티 항으로 추가하여 WAE objective를 정리하면,

$$\min_g \min_f \mathbb{E}_{P_{XU}} d^2(X, g_3(g_2(\mathbf{A}^T g_1(U)) + f(X, U))) + \lambda_1 \mathcal{D}(f_\sharp P_{XU} \| P_{\boldsymbol{\epsilon}})) + \lambda_2 \mathcal{H}(f(X, U), U). \tag{5}$$

예를 들어, $\mathcal{D}$는 [3]에서처럼 MMD 또는 GAN loss를 사용할 수 있고, $\mathcal{H}$는 HSIC(Hilbert-Schmidt Independence Criterion) [4]를 사용할 수 있다.

References

[1] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9593–9602, 2021.

[2] Ignavier Ng, Shengyu Zhu, Zhuangyan Fang, Haoyang Li, Zhitang Chen, and Jun Wang. Masked gradient-based causal structure learning. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 424–432. SIAM, 2022.

[3] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

[4] Romain Lopez, Jeffrey Regier, Michael I Jordan, and Nir Yosef. Information constraints on auto-encoding variational bayes. *Advances in neural information processing systems*, 31, 2018.