Multiple Environments can Reduce Indeterminacies in VAEs

Quanhan Xi, Benjamin Bloem-Reddy
{johnny.xi, benbr}@stat.ubc.ca
Department of Statistics, University of British Columbia

THE UNIVERSITY OF BRITISH COLUMBIA

NEURAL INFORMATION PROCESSING SYSTEMS

Goal: Strong Identifiability in VAEs

Additive noise deep latent variable model (**only Y is observed**):

$$X \overset{iid}{\sim} p_X^{(c)} \quad \epsilon \overset{iid}{\sim} g_\epsilon \quad g_\epsilon \perp\!\!\!\perp p_X,$$
$$Y = f(X) + \epsilon\ .$$

environment index

"decoder" "code"

Our assumptions: f is Borel-measurable and **injective**.

Strong identifiability of the decoder (strong ID)

$$p_{Y,1}^{(c)} = p_{Y,2}^{(c)} \implies f^{(1)} = f^{(2)} \quad \forall c \text{ observed}$$

Strong ID is a lot to ask for, but has important implications.

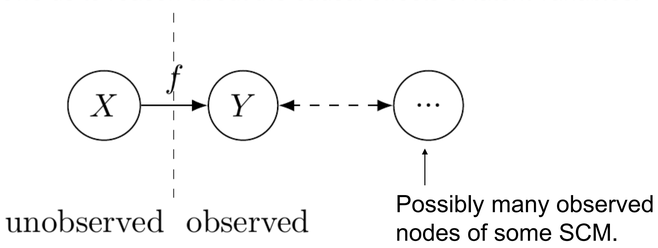
f is injective and strongly identified → X is unique to a denoised Y.

Implies an injectivity in conditional expectation:

$$x_1 \neq x_2 \implies E[Y|X = x_1] \neq E[Y|X = x_2]$$

Example: Inferring unobserved causes as VAE codes

Allows us to reason about the causal effects of latent variables.



What might we do with this?

1) Compute causal effects of X on Y, e.g., the distribution of Y when modifying (semantically meaningful) codes.

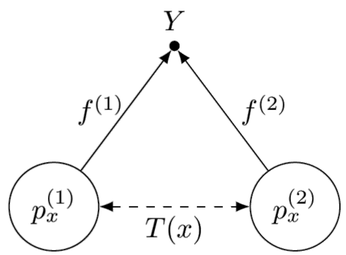
2) Can use X as an intervention variable for Y, inducing 1-to-1 soft interventions on Y that can be simulated.

Measure Automorphisms and Transports

Sources of non-ID in fully flexible models

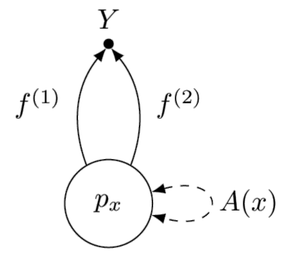
↳ i.e., simultaneous learning of decoder and prior

Suppose a single environment for illustration:



$$f^{(1)}(x) = f^{(2)}(T(x))$$

1) In fully flexible models, the indeterminacy is a prior *transport* for each environment.



$$f^{(1)}(x) = f^{(2)}(A(x))$$

2) When the prior is fixed, the indeterminacy is a prior *automorphism* for each environment.

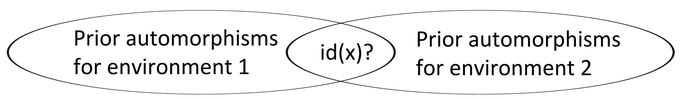
Strong ID ↔ indeterminacy is the identity map.

Strong ID is only possible if priors are fixed ahead of training the decoder.

Fully flexible models cannot have strong ID since T is never the identity map!

The identity map is always a prior automorphism, but it is not the only one in general (e.g., rotations and Gaussians).

Multiple environments can act as multiple constraints:



Recipe for strong ID: fix the prior for a subset of environments such that the only shared automorphism is the identity map.

The main takeaway: **fixed priors are the price to pay for strong ID.** Don't need smoothness, (non)-linearity, etc.

Operationalizing and Future Work

Operationally, many environments may still be fully flexible, only some priors are required to be fixed for identifiability.

One possible subset of priors: a "basis" of exponential families:

Suppose that $X \in \mathbb{R}^K$. Fix $K+1$ priors to be in **the same exponential family**, where

- m (base measure) is strictly positive, T (suff. stat) is injective on at least one dimension, and
- K of the parameters η_c, distinct for each prior, are linearly independent.

Then, the decoder f is identifiable up to equality almost everywhere.

Sketch of Proof: Only the identity function preserves a "basis" of exponential families.

In practice: learn an embedding of environment metadata into the natural parameter space as a pre-processing step.

Work in progress: evaluating different exponential families and environment embeddings empirically.

There are many theoretical questions to consider as well:

- Is there a notion of a useful "basis" of priors generalizing beyond exponential families and natural parameters?
- Can we quantify the cost in model expressiveness when priors are fixed in view of flexible parametrizations of the decoder?

Finally, we believe there are many useful applications of our results in both deep generative modeling and causal learning.

Since our analysis is not specific to latent variable models, we are particularly excited about functional identification in fully observed additive noise models alongside causal identification strategies.

We would love to hear about other possible areas of application!

References

(2020) I. Khemakhem, D. P. Kingma, R. P. Monti, and A. Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. AISTATS 2020.