Contrastive Mixture of Posteriors for Counterfactual Inference, Data Integration and Fairness

Lucas Cosier

ETH Zürich

lcosier@ethz.ch

Presentation for the Topics in Medical Machine Learning Seminar
November 15, 2022

Presentation Overview

Introduction

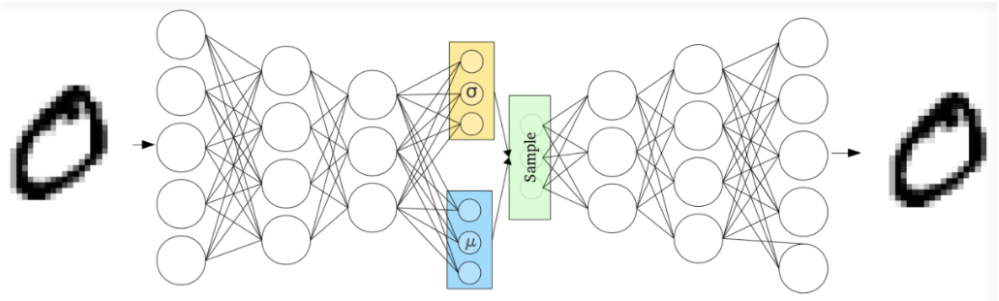
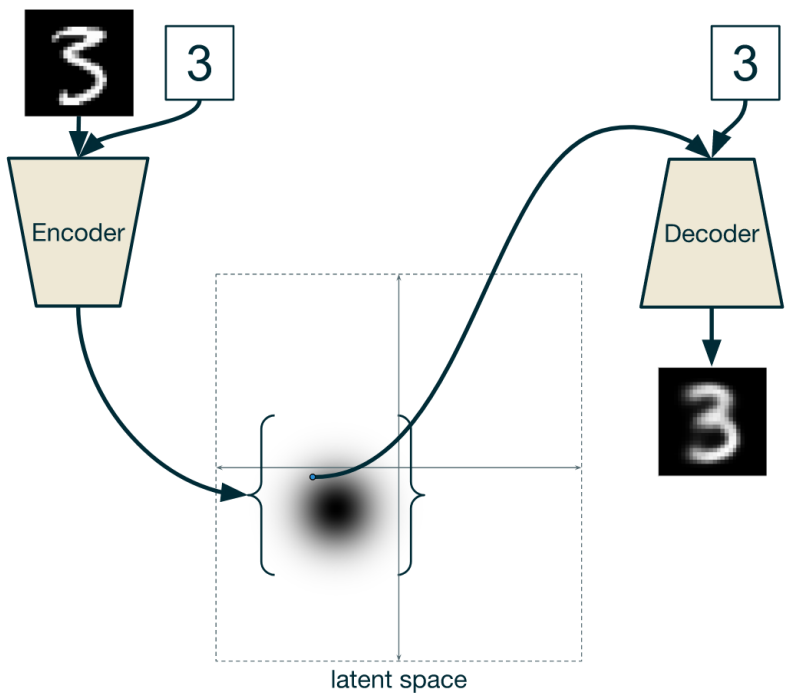


Figure: the VAE: the encoder generates a latent representation of the input $\boldsymbol{x}$ which the decoder samples from to generate new patterns $\boldsymbol{x}'$

- Generative models which explain high-dimensional features $\boldsymbol{x} \in \mathcal{X}$ using low-dimensional latent variables $\boldsymbol{z} \in \mathcal{Z}$
- use an approximate posterior $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \approx p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$ to compute an approximation to the marginal likelihood $p_\theta(\boldsymbol{x})$
- **important**: training done with isotropic priors $p(\boldsymbol{z}) = \mathcal{N}(0, I)$

Conditional Variational Auto-Encoders

Problem: VAEs cannot generate a specific type of observation $\boldsymbol{x}$ on demand.
Solution: Augment data by considering pairs $\{(\boldsymbol{x}_i, c_i)\}_{i=1}^n$, where c_is are categorical variables (representing some condition).



Training CVAEs

- Training procedure changes very little:

$$\log p_\theta(\boldsymbol{x} \mid c) \geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x},c)}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z},c)\right] - KL(q_\phi(\boldsymbol{z} \mid \boldsymbol{x},c)||p_\theta(\boldsymbol{z} \mid c)) \quad (1)$$

- For $c \neq \phi \rightarrow$ CVAE, when $c \equiv \phi$ the original VAE is retrieved
- Other CVAE (trVAE[1], VFAE[2]) variants introduce additional terms to the ELBO to penalize overlap (like the MMD kernel)

[1] Lotfollahi et al. (2019)
[2] Christos et al. (2015)

Pearl's Causal Hierarchy[3]

Layer	Typical Activity	Typical Question
$\mathbf{L}_1$ Association $p(y \mid x)$	Seeing	How would seeing x change my belief in y?
$\mathbf{L}_2$ Intervention $p(y \mid do(x), z)$	Doing	What if I do x?
$\mathbf{L}_3$ **Counterfactuals** $p(y_{x \neq x'} \mid x', y')$	Imagining	How would the observation have changed if x' had been replaced by x?

[3] For more details see Pearl (2009)

The Structural Equation Model

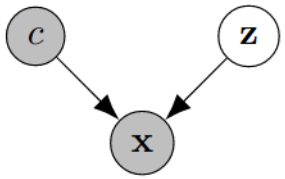


Figure: The Structural Equation Model considered. z and c are independent in the prior.

- Counterfactual questions are difficult to answer because they refer to unobservable data
- A principled approach to such questions is to adopt the framework of Structural Equation Models

The Structural Equation Model

Counterfactual inference can be then performed by

1. *abduction*: inferring the latent z from x and c using $p(z \mid x, c)$
2. *action*: swap c for c'
3. *prediction*: use $p(x \mid z, c')$ to obtain a predictive distribution for the counterfactual

Under the assumption that $z \perp\!\!\!\perp c$ counterfactual distribution of x_i can be written as

$$p(x_{c=c'} \mid x_i, c_i) = \int \underbrace{p(z \mid x_i, c_i)}_{\text{approx. by } q_\phi} \underbrace{p(x \mid z, c')}_{\text{approx by } p_\theta} dz \tag{2}$$

Contrastive Mixture of Posteriors

Summary: Learning a CVAE where $\boldsymbol{z} \perp\!\!\!\perp c$ under encoder q_ϕ by penalising misalignment between different conditions as part of the variational framework

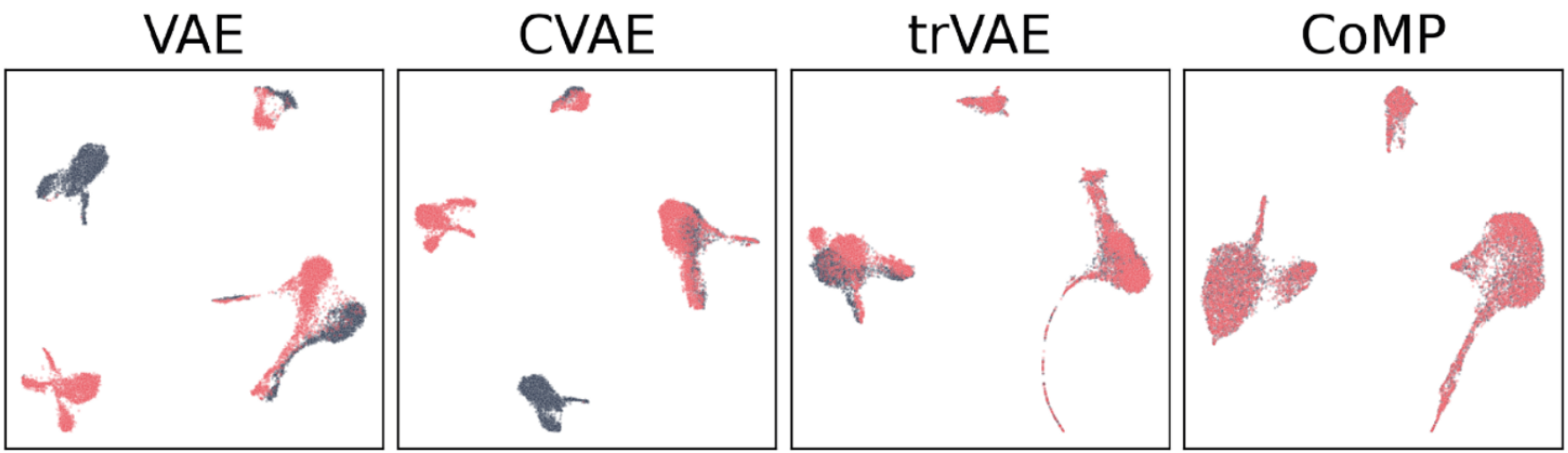


Figure: Latent representations of a single-cell gene expression under $c = red$ and $\neg c = black$

Ties to Counterfactual Inference, Fairness and Data Integration

- **Counterfactual Inference.** Non-trivial since CVAEs may choose to maximize $p_\theta(\boldsymbol{x}|c)$ and ignore the constraint, which violates the SEM. Also requires additional assumptions.
- **Data Integration.** $\boldsymbol{x}$ may suffer from noise injected from experimental conditions. $\boldsymbol{z}$ can be used instead for downstream tasks
- **Fairness.** Since $\boldsymbol{z}$ contains information about $\boldsymbol{x}$ can use it as proxy to make predictive rules about $\boldsymbol{x}$

CoMP Penalty

- Include a penalty term $\underbrace{\log q(\boldsymbol{z}_i \mid c_i)}_{\text{increase entropy}} - \underbrace{\log q(\boldsymbol{z}_i \mid \neg c_i)}_{\text{increase overlap}}$
- For some batch $\{(\boldsymbol{x}_i, c_i)\}_{i=1}^B$ let $I_c = \{j : c_j = c\}$ and $I_{\neg c}$ be its complement.
- Then, $\log q(\boldsymbol{z}_i \mid \neg c_i) \approx \log \left(\frac{1}{|I_{\neg c_i}|} \sum_{j \in I_{\neg c_i}} q(\boldsymbol{z}_i \mid \boldsymbol{x}_j, c_j) \right)$

$$\Rightarrow \begin{matrix}\text{CoMP} \\ \text{Penalty}\end{matrix} = \frac{1}{B} \sum_{i=1}^{B} \log \left(\frac{\frac{1}{|I_{c_i}|} \sum_{j \in I_{c_i}} q(\boldsymbol{z}_i \mid \boldsymbol{x}_j, c_i)}{\frac{1}{|I_{\neg c_i}|} \sum_{j \in I_{\neg c_i}} q(\boldsymbol{z}_i \mid \boldsymbol{x}_j, c_j)} \right) \tag{3}$$

CoMP Penalty

- The penalty is actually an upper bound to a sum of weighted KL terms $KL(q(\boldsymbol{z} \mid c) || q(\boldsymbol{z} \mid \neg c))$
- Bound becomes tight as $B \to \infty$
- Adding the CoMP penalty to the familiar CVAE objective results in the complete training objective for a batch of size B:

$$\mathcal{L}(\theta, \phi) = \frac{1}{B} \sum_{i=1}^{B} \left[\log \left(\frac{p_\theta(\boldsymbol{x}_i \mid \boldsymbol{z}_i, c_i) p(\boldsymbol{z}_i)}{q_\phi(\boldsymbol{z}_i \mid \boldsymbol{x}_i, c_i)} \right) \right] - \gamma \times \text{CoMP Penalty}$$

Towards Identifiable Counterfactuals

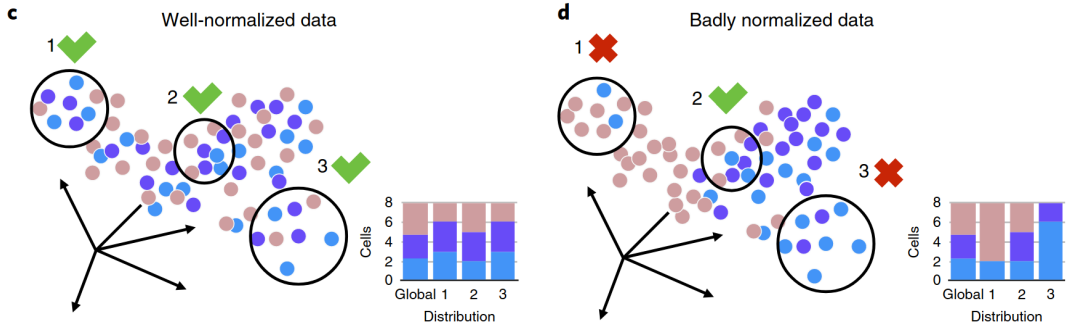
Theoretical results

- if $z \sim \mathcal{N}(0, I)$ then **identifiability breaks down** in CVAEs
- if $z \sim r(z)$, $r \overset{\text{d}}{\neq} \mathcal{N}$, necessitates additional assumptions to ensure counterfactual identifiability:
  1. linear decoders for each condition e.g. $x = A_c z$
  2. z must be decomposable as Bs, s_i are non-Gaussian of unit variance
  3. for every permutation matrix P and negation matrix N for which $PN \neq I$, $PNs \overset{\text{d}}{\neq} s$
- For CoMP particularly if $KL(r(z)|||p(z)) \leq K_1$ this **ensures consistency** and **identifiability**.

Evaluation Metrics

Two goals

1. Testing the extent of $z \perp\!\!\!\perp c$
   - **k-nearest Batch Effect Test**[4] $\text{kBET}_{k,\alpha}$



   - **Local Silhouette Coefficient**[5] $s_{k,c}$.
2. Quantify useful information retained in z
   - **mean Silhouette Coefficient** $\tilde{s}_{k,c}$ and **mean kBET**: s and kBET are calculated on the d_i subpopulations

[4] Büttner et al. (2019)
[5] Rousseauw (1987)

Alignment of tumour and cell-line samples

Tumour/Cell line dataset. Warren et al. (2021)
Task: dataset integration and batch effect correction

- consists of bulk expression profiles for tumours $(n \approx 12k)$ and cancer cell-lines $(n = 1.2k)$ across 39 different cancer types (the d_i's)

	Accuracy	s	kBET	$\tilde{s}$	m-kBET
VAE	0.209	0.658	0.974	0.803	0.581
CVAE	0.328	0.554	0.931	0.684	0.571
VFAE	**0.585**	0.168	0.258	0.198	0.188
trVAE	**0.585**	0.096	0.163	0.138	0.123
Celligner	0.578	0.082	0.525	0.568	0.226
CoMP	0.579	**0.023**	**0.160**	**0.094**	**0.101**

Figure: Tumour / Cell Line experiment results, with $k = 100$, $c =$ Cell Line, and parameter $\alpha = 0.01$ for the kBET and m-kBET metrics. $s_{k,c}$ and $\tilde{s}_{k,c}$ are the two Silhouette Coefficient variants introduced earlier. Top scores are in **bold**.

Alignment of tumour and cell-line samples

Results

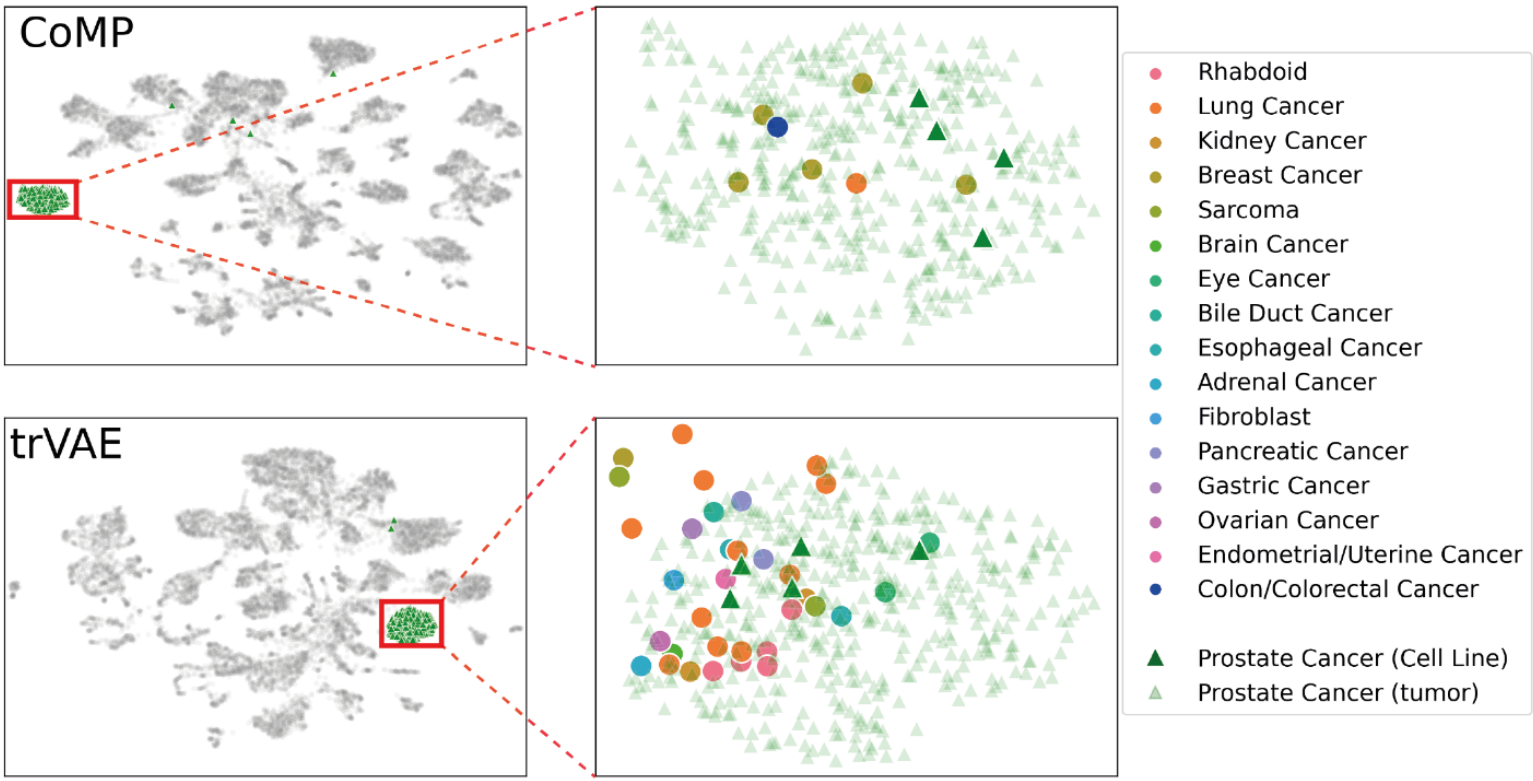


Figure: 2D UMAP projection of the CoMP and trVAE posterior means of z_i from Tumour/Cell Line data and the detailed Prostate Cancer tumour sample clusters.

Interventions

Dataset details and data processing

stimulated/untreated single-cell PBMCs expression dataset[6]
Task: Counterfactual inference.

- $\approx 14k$ single-cell expression profiles for peripheral blood mononuclear cells (PBMCs), various immune cell types
- $7k$ cells stimulated with interferon (IFN)-β, $6k$ left untreated

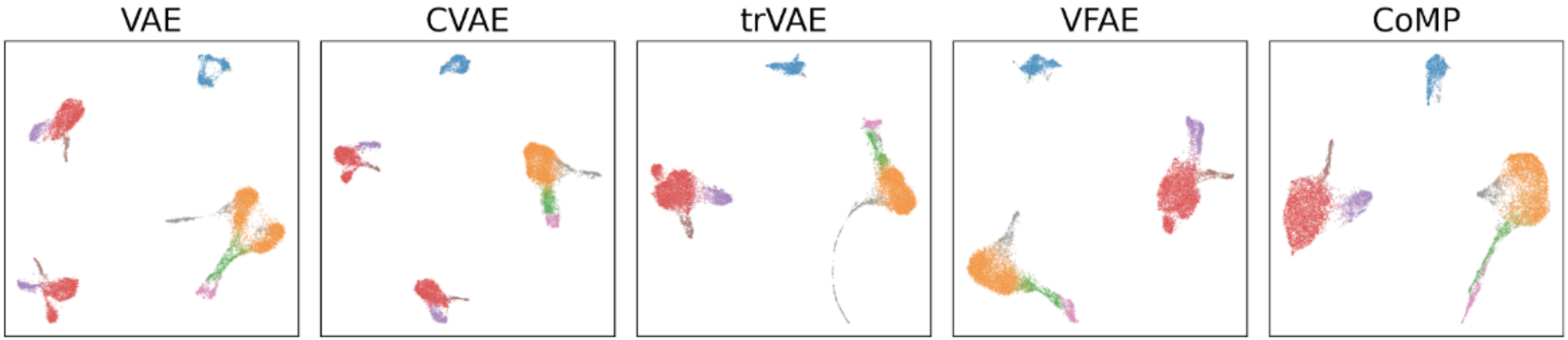


Figure: Stimulated and control PBMC scRNA-seq data with colours highlighting immune cell types

Interventions

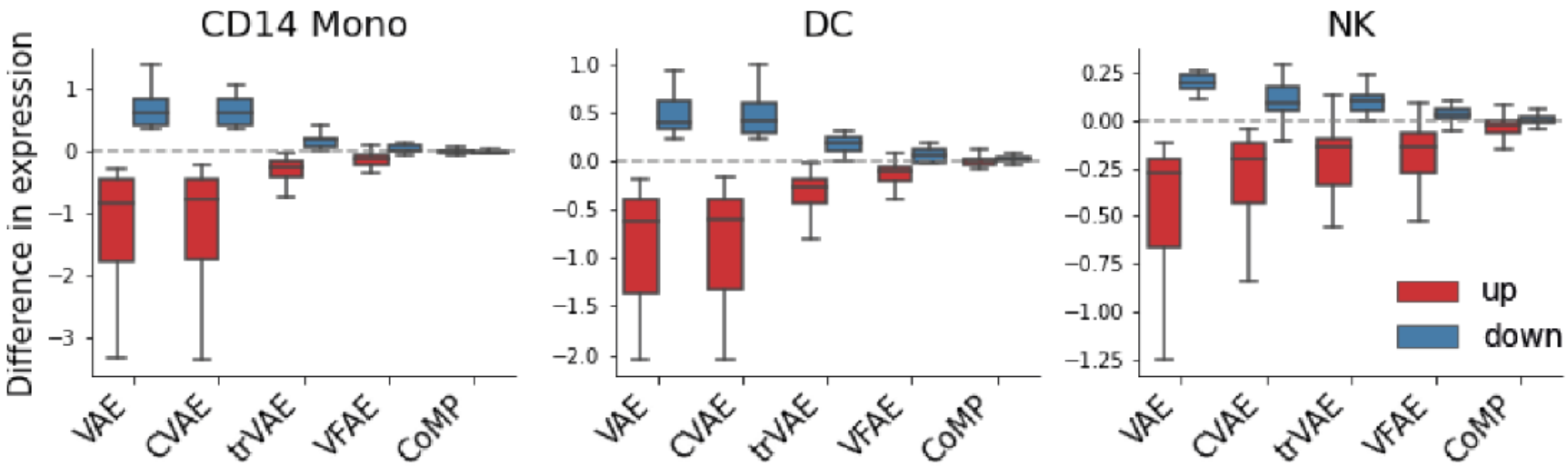


Figure: The difference in gene expression values for the top 50 differentially expressed genes (up-regulated: red, down-regulated: blue) between IFN-β stimulated cells and counterfactually stimulated control cells for CD14 monocytes, dendritic cells (DC) and natural killer (NK) cells.

Fair classification

Dataset details and results

UCI Adult Income dataset

- contains information relating to education, marriage status, ethnicity, self reported gender of census participants and a binary high/low income label

	Gender Acc	Income Acc	s	kBET
Original data	0.796	**0.849**	0.067	0.786
VAE	0.764	0.812	0.054	0.748
CVAE	0.778	0.819	0.054	0.724
VFAE-s	0.680	0.815	-	-
VFAE-m	0.789	0.805	0.046	0.571
trVAE	0.698	0.808	0.066	0.731
CoMP	**0.679**	0.805	**0.011**	**0.451**

Figure: Experiment results with $k = 1000$, $c =$ Male for silhouette score s, and $k = 100$, $\alpha = 0.01$ for kBET. A lower gender prediction accuracy is better; 0.675 is the lowest achievable

Conclusion

CoMP

- is a novel method that **enforces** latent alignment $z \perp\!\!\!\perp c$ as part of the variational framework
- introduces **identifiability** and **consistency** results to show that alignment is not always sufficient to perform valid counterfactual inference with a CVAE
- performs very well in areas of counterfactual inference, fairness and data integration

The gap between theory (which assumes linear decoders) and practice (non-linear decoders) still needs to be addressed

References I


Allison Warren et al. (2021)
Global computational alignment of tumor and cell line transcriptional profiles
Nature communications, 12(1), 1-12


Hyun Min Kang et al. (2018)
Multiplexed droplet single-cell RNA-sequencing using natural genetic variation
Nature biotechnology 36(1), 89-94.


Judea Pearl (2009)
Causality
Cambridge university press 2009.


Louisos Christos et al. (2015)
The variational fair autoencoder
arXiv preprint arXiv:1511.00830

References II

Maren Büttner et al. (2019)
A test metric for assessing single-cell RNA-seq batch correction
Nature methods, 16(1), 43-49.

Mohammad Lotfollahi et al. (2019)
Conditional out-of-sample generation for unpaired data using trVAE
arXiv preprint arXiv:1910.01791

Peter J. Rousseauw (1987)
Silhouettes: a graphical aid to the interpretation and validation of cluster analysis
Journal of computational and applied mathematics, 20, 53-65.

The End

Questions? Comments?