Proceedings of Machine Learning Research vol 236:868–912, 2024 3rd Conference on Causal Learning and Reasoning

Toward the Identifiability of Comparative Deep Generative Models

Romain Lopez LOPEZ.ROMAIN@GENE.COM
Genentech Research and Early Development
Stanford University

Jan-Christian Huetter HUETTEJ1@GENE.COM
Genentech Research and Early Development

Ehsan Hajiramezanali HAJIRAMM@GENE.COM
Genentech Research and Early Development

Jonathan K. Pritchard PRITCH@STANFORD.EDU
Stanford University

Aviv Regev REGEVA@GENE.COM
Genentech Research and Early Development

Editors: Francesco Locatello and Vanessa Didelez

Abstract

Deep Generative Models (DGMs) are versatile tools for learning data representations while adequately incorporating domain knowledge such as the specification of conditional probability distributions. Recently proposed DGMs tackle the important task of comparing data sets from different sources. One such example is the setting of contrastive analysis that focuses on describing patterns that are enriched in a target data set compared to a background data set. The practical deployment of those models often assumes that DGMs naturally infer interpretable and modular latent representations, which is known to be an issue in practice. Consequently, existing methods often rely on ad-hoc regularization schemes, although without any theoretical grounding. Here, we propose a theory of identifiability for comparative DGMs by extending recent advances in the field of non-linear independent component analysis. We show that, while these models lack identifiability across a general class of mixing functions, they surprisingly become identifiable when the mixing function is piece-wise affine (e.g., parameterized by a ReLU neural network). We also investigate the impact of model misspecification, and empirically show that previously proposed regularization techniques for fitting comparative DGMs help with identifiability when the number of latent variables is not known in advance. Finally, we introduce a novel methodology for fitting comparative DGMs that improves the treatment of multiple data sources via multi-objective optimization and that helps adjust the hyperparameter for the regularization in an interpretable manner, using constrained optimization. We empirically validate our theory and new methodology using simulated data as well as a recent data set of genetic perturbations in cells profiled via single-cell RNA sequencing.

Keywords: non-linear ICA; deep generative models; variational inference; disentanglement;

1. Introduction

Since the introduction of Variational Auto-Encoders (VAEs) (Kingma and Welling, 2014; Rezende et al., 2014), these so-called Deep Generative Models (DGMs) have established themselves as a go-to tool for learning representations of heterogeneous data sets. Their applications span financial time-series analysis (Bergeron et al., 2022), speech analysis and synthesis (Girin et al., 2021), as well as biological data analysis (Lopez et al., 2020). Their natural ability to deal with multi-modal (Wu

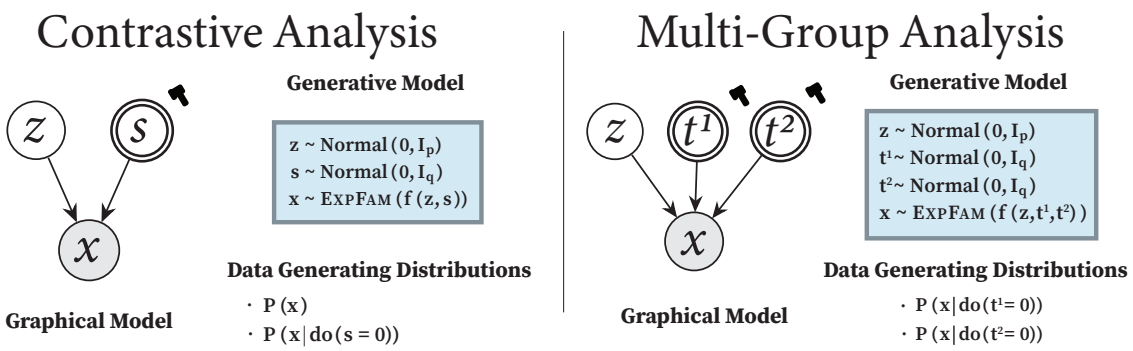

Figure 1: Presentation of the comparative deep generative models considered in this work.

and Goodman, 2018), temporal (Girin et al., 2021) and spatial data sets (Yuan et al., 2019) makes them a powerful framework for extracting informative representations of data at a massive scale. Learning informative and compact representations from data is a milestone for applications driven by goodness of fit, or specific downstream prediction tasks. For many other cases, however, learning representations that are modular and have semantic meaning is crucial for reasons of interpretability.

This paper is concerned with the problem of *comparative analysis*, which seeks to model the similarities and differences across multiple data sets. This analytical approach is often driven by scientific applications, where researchers routinely juxtapose observations from a condition of interest (e.g., a disease) with a control condition (e.g., healthy).

A particular form of comparative analysis is termed *contrastive analysis*[1]. This methodology aims to characterize how a target data set differs from a background data set (Zou et al., 2013) (Figure 1, left). In achieving this, a generative model that incorporates two sets of latent variable $(\boldsymbol{z}, \boldsymbol{s}) \in \mathbb{R}^p \times \mathbb{R}^q$ is learned from data (Abid and Zou, 2019; Jones et al., 2022). Here, the *background variables* $\boldsymbol{z}$ represent patterns inherent to the background data set, while the *salient variables* $\boldsymbol{s}$ capture nuances unique to the target data set. Contrastive analysis methods have been widely adopted in scientific research, with notable applications in omics data analysis (Boileau et al., 2020; Weinberger et al., 2023) and brain imaging studies (Louiset et al., 2023). Another approach within comparative analysis involves learning both shared and group-specific data representations using a single generative model (Davison et al., 2019; Weinberger et al., 2022b). We refer to it as the multi-group analysis setting (Figure 1, right).

The widespread success and adoption of these methods is somewhat surprising. Indeed, DGMs face important challenges in learning interpretable representations (Locatello et al., 2019), and usually require ad-hoc regularization schemes in order to yield satisfactory performance (Higgins et al., 2017; Kim and Mnih, 2018; Lopez et al., 2018b). Similarly, successful inference of comparative DGMs also requires the engineering of regularization approaches (Weinberger et al., 2022a). This brings up the important theoretical question of why such regularization strategies are necessary. A plausible hypothesis is that the model itself is not identifiable, and that regularization helps constrain the function class used to fit the model. Here, non-identifiability means that given some data and a ground truth generating process, there exists an alternative model that has equal data likelihood, but such that the subspaces recovered from it are different. We note that identifiability is also an

1. This is a distinct line of work from the field of contrastive learning, that aims at distinguishing between positive and negative pairs of data points, as used in self-supervised learning.

important question of its own, because it is a necessary condition to interpret and attribute semantic meaning to the learned representations, which is the end goal of real-world scientific applications.

We therefore explore the question of identifiability of comparative DGMs. To the best of our knowledge, such theoretical developments have been completely unexplored in the related literature. As a starting point, we highlight that data sets from different sources may be interpreted as the result of a do intervention on the graphical model (Pearl, 2009). This allows us to build upon recent advances in causal representation learning and non-linear ICA theory (Khemakhem et al., 2020; Lachapelle et al., 2022; Kivva et al., 2022) to prove the (block-wise) identifiability of comparative DGMs under the assumption of a piece-wise affine mixing function (e.g., parameterized by a ReLU / Leaky ReLU neural network). This demonstrates, for the first time, the identifiability of many recently published contrastive DGMs (Jones et al., 2022; Severson et al., 2019; Weinberger et al., 2023), and multi-group DGMs (Severson et al., 2019; Weinberger et al., 2022b). We also provide empirical evidence of this point with numerical experiments. This result is surprising, because of the practical need for regularization. To reconcile this apparent contradiction, we illustrate that identifiability guarantees are lost when the numbers of latent variables in each block are misspecified, as is often the case in real-world data analysis. In numerical experiments, we assess that existing regularization strategies considerably help mitigate this effect. Finally, motivated by this theoretical analysis, we also propose a new methodology for fitting comparative DGMs based on recent advances in multi-objective optimization (Sener and Koltun, 2018), and constrained optimization (Gallego-Posada et al., 2022).

After briefly introducing the background (Section 2), we present our novel theory of identifiability for comparative DGMs (Section 3). We then discuss limitations of the theory in the case of model misspecification in Section 4. We propose novel algorithmic methodology (Section 5), and conduct numerical experiments on simulations as well as a recent data set from a genetic screen profiled via single-cell RNA sequencing (Section 6). Discussion of related works appears in Appendix A.

2. Background

This paper is concerned with the modular recovery of latent variables of a comparative analysis model that initially generated the data (Figure 1). Therefore, we briefly introduce the field of comparative analysis, and then present recent results on the identifiability of non-linear Independent Component Analysis (ICA), one of the prominent methods for latent variable recovery.

2.1. Comparative Analysis with Deep Generative Models

Contrastive Analysis Zou et al. (2013) introduced the goal of contrastive analysis, and the first algorithmic approaches (e.g, based on mixture models). Abid et al. (2018) proposed a contrastive principal component analysis (cPCA) method that captured intriguing variations from the target data set that do not appear in the background data set. Subsequent endeavors (Li et al., 2020; Jones et al., 2022) steered towards the creation of probabilistic latent variable models tailored to contrastive analysis, with a recent focus on deep generative models (Severson et al., 2019; Abid and Zou, 2019; Ruiz et al., 2019; Weinberger et al., 2023). In this setting, a contrastive DGM has two sets of latent variables (Figure 1, left): the salient variable $\boldsymbol{s} \in \mathbb{R}^p$ and the background latent variable $\boldsymbol{z} \in \mathbb{R}^q$. The *target* data set is generated by sampling both sets of latent variables from an isotropic Gaussian prior distribution, passing them through a mixing function f and sampling the data $\boldsymbol{x}$ from the exponential family distribution ExpFam, using $f(\boldsymbol{z}, \boldsymbol{s})$ as the parameter. The *background* data set is generated

similarly, but by setting $\boldsymbol{s} = \mathbf{0}$ to make sure that $\boldsymbol{s}$ is utilized only for describing the target data set. We denote this distribution as a hard intervention $p_\theta\left(\boldsymbol{x} \mid \text{do}(\boldsymbol{s} = \mathbf{0})\right)$ (Pearl, 2009) in order to draw parallels with interventional causal representation learning (Ahuja et al., 2023).

In terms of inference procedure, all methods rely on the contrastive Variational Auto-Encoder (cVAE) framework (Abid and Zou, 2019). For each sample in the target data set (resp. the background data set), the variational distribution is $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})q_\phi(\boldsymbol{s} \mid \boldsymbol{x})$ (resp. $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ only, as $\boldsymbol{s} = \mathbf{0}$). Then, the composite evidence lower bound (ELBO) is derived as

$$\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z}, \mathbf{0})}{q_\phi(\boldsymbol{z} \mid \boldsymbol{x})} + \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})q_\phi(\boldsymbol{s}|\boldsymbol{x})} \log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{s})}{q_\phi(\boldsymbol{z} \mid \boldsymbol{x})q_\phi(\boldsymbol{s} \mid \boldsymbol{x})}. \tag{1}$$

This composite ELBO corresponds to the sum of two individual ELBOs for each of the two data sets (background and target). It may be used as an objective function for maximization, in conjunction with adequate regularization of the neural networks parameterizing the variational distribution as a function of the input data. In Abid and Zou (2019), an additional regularization term specifically promotes independence of the two sets of latent variables.

Multi-group Analysis Recently, several models have been designed to distinguish patterns that are shared by all data sets, versus the ones that are specific to each data set (Davison et al., 2019; Weinberger et al., 2022b). We provide more details about the generative model and the inference mechanism in Appendix B.

2.2. Identifiability of Non-linear Independent Component Analysis

ICA assumes that $\boldsymbol{x} \in \mathbb{R}^d$ is generated using p independent latent variables $\boldsymbol{z} = (z_1, \ldots, z_p)$, called *independent components* (Hyvärinen et al., 2002). More precisely, observations $\boldsymbol{x}$ are defined as $\boldsymbol{x} = f(\boldsymbol{z}) + \boldsymbol{\epsilon}$ with f a mixing function and $\boldsymbol{\epsilon}$ an exogenous noise variable. The ICA literature established that in the general case of a non-linear *mixing function* f, the model is unidentifiable from i.i.d. observations of $\boldsymbol{x}$ (Hyvärinen and Pajunen, 1999), and therefore the original $\boldsymbol{z}$ may not be recovered. Given this negative result, several papers introduced identifiable forms of non-linear ICA models (Harmeling et al., 2003; Sprekeler et al., 2014; Hyvärinen and Morioka, 2016, 2017), based on the observability of an *additional auxiliary* random variable. However, such auxiliary variable is not always available in practice. More recently, Kivva et al. (2022) proposed a new theory of identifiability based on the assumption that f is a piece-wise affine function. Their main result is that many previously proposed deep generative models parameterized with ReLU / Leaky ReLU neural networks and additive Gaussian observation noise, a commonly used architecture, are identifiable up to a linear transformation of the mixing function. Our work directly builds upon this line of work to assess the identifiability of comparative DGMs.

3. A Theory of Identifiability for Comparative Analysis Models

For the sake of conciseness and ease of notation, we focus on the contrastive analysis case in this section. Definitions, theorem statements, and proofs for the multi-group setting appear in Appendix C.3.

3.1. Subspace Identifiability

Identifiability is a critical property to understand whether a model's parameters can be uniquely inferred from observations (Ran and Hu, 2017). Within the context of contrastive analysis, our concern is not the recovery of latent variables at the component level. Rather, our interest lies in the retrieval of specific subspaces, namely the blocks $\mathcal{Z} = \mathbb{R}^p$ and $\mathcal{S} = \mathbb{R}^q$. To introduce this concept, we first define a general criterion for compatibility of the Cartesian product of subspaces by a map.

Definition 1 (Compatible map) *Let $(E_1, \dots, E_n)$ be n Euclidean spaces, and let $E = \prod_{i=1}^n E_n$ designate the Cartesian product space. A map $\phi : E \to E$ is said to be* compatible *with the Cartesian product $E = \prod_{i=1}^n E_n$ if there exist maps $\left(\bar{\phi}_1, \dots, \bar{\phi}_n\right)$ of the subspaces $(E_1, \dots, E_n)$ such that for all $\boldsymbol{e} = (e_1, \dots, e_n) \in E$, we have that $\phi(\boldsymbol{e}) = \left(\bar{\phi}_1(e_1), \dots, \bar{\phi}_n(e_n)\right)$.*

Now, if we denote the support of latent variables for the background data set (resp. the target data set) as $\mathcal{D}^b = \mathcal{Z} \times \{0\}$ (resp. $\mathcal{D}^t = \mathcal{Z} \times \mathcal{S}$), we define the subspace disentanglement condition as follows.

Definition 2 (Subspace Disentanglement) *Let f be the ground truth mixing function, and $\tilde{f}$ be a learned mixing function. Let us also assume that $f(\mathcal{D}^b) = \tilde{f}(\mathcal{D}^b)$, $f(\mathcal{D}^t) = \tilde{f}(\mathcal{D}^t)$, and that the map $v = f^{-1} \circ \tilde{f}$ is well-defined. $\tilde{f}$ is said to be subspace-disentangled with respect to f if v is compatible with respect to the Cartesian product $\mathcal{D}^t = \mathcal{Z} \times \mathcal{S}$.*

When this property is not verified, the learned mixing function $\tilde{f}$ and the background f provide distinct decompositions of the signal from the feature space $\mathcal{X}$ into $\mathcal{Z}$ and $\mathcal{S}$. Related definitions that appear in previous works such as Von Kügelgen et al. (2021) are discussed in Appendix A. We may now outline our definition for subspace identifiability of contrastive DGMs.

Definition 3 (Subspace Identifiability) *A contrastive analysis model with ground truth mixing function f is subspace identifiable from data if for all other mixing functions $\tilde{f}$ that yield the same background and target data distributions, we have that $\tilde{f}$ is subspace-disentangled with respect to f.*

When such a model is subspace identifiable, and we observe data $\boldsymbol{x}$, we are guaranteed that the learned representations $(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{s}}) = \tilde{f}^{-1}(\boldsymbol{x})$ are given by a transformation of each of the original spaces: $\tilde{\boldsymbol{z}} = h_{\boldsymbol{z}}(\boldsymbol{z})$, and $\tilde{s} = h_{\boldsymbol{s}}(\boldsymbol{s})$, and therefore semantic meaning can be attributed to those subspaces.

Although the fact that we observe two data sets is potentially helpful in breaking symmetry in the roles played by the shared latent variables $\boldsymbol{z}$ and $\boldsymbol{s}$, it is not true in general that all contrastive analysis models are subspace identifiable.

Example 1 (Counterexample) *For $p = 2$ and $q = 1$, let us consider the following map of $\mathbb{R}^3$:*

$$\Phi : \begin{pmatrix} z_1 \\ z_2 \\ s \end{pmatrix} \mapsto \begin{pmatrix} z_1 \cos s - z_2 \sin s \\ z_1 \sin s + z_2 \cos s \\ s \end{pmatrix}. \tag{2}$$

For any non-trivial mixing function f, we define $\tilde{f} = f \circ \Phi$. $\tilde{f}$ and f generate the same data distributions because Φ is a diffeomorphism that preserves volume, and distance to the origin. However, $\tilde{f}$ is not subspace disentangled with respect to f. The complete proof appears in Appendix C.1. Example 1 may be seen as an extension of the classical counter-example of identifiability for linear ICA (Hyvärinen et al., 2002), exploiting the rotational invariance of the Gaussian distribution but with a non-constant rotation angle.

3.2. Identifiability Result for Piece-wise Affine Mixing Functions and Noiseless Observations

The counterexample presented above suggests that we must restrict the function class for f in order to potentially obtain identifiability. We propose to build upon recent work on identifiability of DGMs with mixing functions specified as multilayer perceptrons (MLP) with ReLU / Leaky ReLU activations (Kivva et al., 2022) to obtain the first result of identifiability of comparative analysis models.

Theorem 1 (Identifiability Theorem) *Let the ground truth mixing function f and the learned mixing function $\tilde{f}$ both be continuous and injective piece-wise affine mixing functions such that $f(\boldsymbol{z}, \boldsymbol{s}) \stackrel{d}{=} \tilde{f}(\boldsymbol{z}, \boldsymbol{s})$ and $f(\boldsymbol{z}, \boldsymbol{0}) \stackrel{d}{=} \tilde{f}(\boldsymbol{z}, \boldsymbol{0})$. Then, $\tilde{f}$ is subspace disentangled with respect to f and the noiseless version of the contrastive analysis model is subspace identifiable.*

The proof appears in Appendix C.2, and consists of two steps. First, we apply the result of Kivva et al. (2022) to each of the target and background data distributions to obtain the linear identifiability of the mixing function on each domain. Then, we rely on the geometry of affine transformations to prove that the disentanglement criterion must hold on both data domains. Because this is an instance of linear disentanglement, this implies that v is a linear transformation. We also note that the assumptions of isotropic Gaussian distributions for $p(\boldsymbol{z})$ and $p(\boldsymbol{z}, \boldsymbol{s})$ for Theorem 1 could be relaxed to members of an exponential family of distributions, as long as the densities are analytic functions, and the family is closed under additive transformation (Kivva et al., 2022).

Because the problem of identifiability of non-linear ICA with additive Gaussian noise can be reduced to the noiseless case (Khemakhem et al., 2020), Theorem 1 and its multigroup variant, Theorem 3, are readily applicable to several real-world models. In the special case where f is linear injective, these results yield the identifiability of probabilistic contrastive principal component analysis (Li et al., 2020), and multi-study factor analysis (De Vito et al., 2019). The non-linear version provides the identifiability of the cross-population VAE (Davison et al., 2019), and of the contrastive VAE (cVAE) (Abid et al., 2018).

3.3. Extensions towards models with Observational Count Noise

The results from Theorem 1, and to the best of our knowledge, all previous results on identifiability of non-linear ICA models[2] only apply to noiseless measurements, or to Gaussian observation noise. However, many real-world applications of DGMs (Lopez et al., 2020), and especially comparative DGMs, have been proposed to deal with count data, such as the contrastive generalized latent variable model (CGLVM) (Jones et al., 2022), ContrastiveVI (Weinberger et al., 2023) and multi-GroupVI (Weinberger et al., 2022b). We therefore now show that the non-linear ICA identifiability problem with Poisson or negative binomial noise reduces to the noiseless one.

Theorem 2 (Reduction from observational count noise to the noiseless setting) *Let $\boldsymbol{u} \sim$ Normal $(0, I_p)$. Let $f = \sigma \circ g$ (resp. $\tilde{f} = \sigma \circ \tilde{g}$) be the composition of a scalar link function σ, valued in $\mathbb{R}_+$ (applied component-wise), with a piecewise affine function g (resp. $\tilde{g}$). Let $\boldsymbol{x} \sim p_x(f(\boldsymbol{u}))$ and $\tilde{\boldsymbol{x}} \sim p_x(\tilde{f}(\boldsymbol{u}))$ such that p_x is Poisson or negative binomial with fixed shape. If σ is a bicontinuous bijection, then,*

$$\tilde{\boldsymbol{x}} \stackrel{d}{=} \boldsymbol{x} \implies \tilde{f}(\boldsymbol{u}) \stackrel{d}{=} f(\boldsymbol{u}) \implies \tilde{g}(\boldsymbol{u}) \stackrel{d}{=} g(\boldsymbol{u}). \tag{3}$$

2. The initial version of Khemakhem et al. (2020) presented a proof of identifiability for categorical variables that has since been removed due to a mistake in the write-up.

The proof appears in Appendix D.2. Our proof appeals to calculation and identification of the Laplace transformation of the distribution of random variables $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$. We note that more general versions of this theorem were introduced in early identifiability theory (Sapatinas, 1995; Teicher, 1961). From Theorem 2, we conclude to the block-identifiability of the comparative analysis models mentioned above in the setting of observational count noise and invertible link function.

Additionally, we prove that identifiability does not hold in the case of Bernoulli observational noise without further assumptions. Explicit counterexamples appear in Appendix D.3, disproving a conjecture in Khemakhem et al. (2020). We instead hypothesize that non-identifiability holds in general, for any observational distribution with fixed finite support.

4. Impact of Misspecification

Our main results (Theorems 1 and 2) implicitly assume that the observed data have been simulated from the generative model $p_\theta(\boldsymbol{x})$. However, this may be impossible to verify in practice, as there are many assumptions that might be unknown to practitioners. Examples of such assumptions include the specification of the graphical model, a function class for the mixing function f, as well as the number of latent variables. Given any source of such model misspecification, the theory above unfortunately does not apply.

We focus in this work on a discussion of the impact of a misspecification of the number of latent variables. This is an important starting point, because it is easy to illustrate, and it is known that overestimating the number of latent variables induces severe entanglement in practice, making regularization necessary (Weinberger et al., 2022a). Beside empirical work, theoretical developments are needed to understand how this occurs in the contrastive analysis setting.

To illustrate this, let $p' \geq p$ and $q' \geq q$ be the estimated dimensions of the background space and salient space, respectively, with $p' + q' > p + q$. Further, denote by $\tilde{\boldsymbol{z}} = (\boldsymbol{z}, \boldsymbol{u})$ and $\tilde{\boldsymbol{s}} = (\boldsymbol{s}, \boldsymbol{v})$ the respective latent variables, where $\boldsymbol{u}$ and $\boldsymbol{v}$ are additional variables of dimensions $p' - p$ and $q' - q$, respectively. We consider data $\tilde{\boldsymbol{x}} = \tilde{f}(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{s}})$, generated by a learned mixing function $\tilde{f}$. Compared to our previous setting, we cannot assume injectivity of $\tilde{f}$ under equality of the data generating distributions. Indeed, if we assume that $\tilde{f}(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{s}}) \overset{d}{=} f(\boldsymbol{z}, \boldsymbol{s})$, then the support of those distributions must be equal $\tilde{f}(\mathbb{R}^{p'+q'}) = f(\mathbb{R}^{p+q})$ and have the same manifold dimension. But because of the dimension mismatch, $\tilde{f}$ cannot be injective. In particular, the lack of injectivity of $\tilde{f}$ implies that it does not have a well defined inverse, and makes theoretical analysis challenging.

We therefore first seek to characterize the case where both f and $\tilde{f}$ are linear functions. Surprisingly perhaps, we show that entanglement does not occur in this scenario.

Proposition 1 (Block-wise identifiability under misspecification for the linear case) *Let the ground truth mixing function f be injective linear, and the learned function $\tilde{f}$ be a linear function such that $f(\boldsymbol{z}, \boldsymbol{s}) \overset{d}{=} \tilde{f}(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{s}})$ and $f(\boldsymbol{z}, \boldsymbol{0}) \overset{d}{=} \tilde{f}(\tilde{\boldsymbol{z}}, \boldsymbol{0})$. Then, there exist surjective linear functions $h_{\boldsymbol{z}}$ and $h_{\boldsymbol{s}}$ such that $(\boldsymbol{z}, \boldsymbol{s}) = v(\tilde{\boldsymbol{s}}, \tilde{\boldsymbol{z}}) = (v_{\boldsymbol{s}}(\tilde{\boldsymbol{s}}), v_{\boldsymbol{z}}(\tilde{\boldsymbol{z}}))$, where $v = f^{-1} \circ \tilde{f}$.*

The proof appears in Appendix E.1 and builds upon the proof of identifiability for factor analysis. This result is interesting, as it may explain why regularization is not used for linear comparative analysis models, but was introduced with the first applications of DGMs to this setting (Abid and Zou, 2019). We introduce a broad class of examples of non-identifiable models with non-linear and non-injective mixing functions in Appendix E.2.

Although the discussion above is important to define what the lack of identifiability could imply, it ignores the impact of the variational inference procedure. This is a central point, because the regularization approaches introduced for comparative DGMs impose independence constraints for the aggregated variational posterior (Salakhutdinov and Larochelle, 2010). For the target data set, the aggregated posterior is defined as $\hat{q}_\phi^t(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{s}}) = \mathbb{E}_{p_{\text{data}}(\boldsymbol{x})}\left[q_\phi(\tilde{\boldsymbol{z}} \mid \boldsymbol{x}) q_\phi(\tilde{\boldsymbol{s}} \mid \boldsymbol{x})\right]$, and the regularizer aims to enforce the independence statement $\hat{q}_\phi^t(\tilde{\boldsymbol{z}}) \perp\!\!\!\perp \hat{q}_\phi^t(\tilde{\boldsymbol{s}})$ (Abid and Zou, 2019). Other regularization approaches are described in Appendix E.3.

Interestingly, the independence constraint may not be enough to restore identifiability in general. Our conjecture is that it does contribute to reducing entanglement by constraining the inference network, and therefore restricting the space of admissible mixing functions. We demonstrate in later sections empirical evidence that it indeed improves disentanglement, but leave theoretical analysis to future work.

5. Multi-Objective Constrained Optimization for Contrastive VAEs (MO-CO-cVAEs)

The standard routine for fitting comparative DGMs consists in casting the inference problem as an optimization problem by maximizing a lower bound on the likelihood, following the principles of variational inference (Jordan et al., 1999). By more closely inspecting the nature of the optimization problem at hand, we present here a novel method for fitting contrastive DGMs.

5.1. Maximum Likelihood Across Data Sets Using Multi-Objective Optimization

Existing methodology for fitting comparative analysis models typically derives one evidence lower bound (ELBO) for each data set, specifically, $\mathcal{L}^B(\theta, \phi)$ for the background $\mathcal{L}^T(\theta, \phi)$ for the target data set. The objective function is then defined as the sum of these ELBOs as in Equation 1. We refer to this approach as the Single Objective cVAE (SO-cVAE). In this scenario, optimizing one loss may negatively impact the optimization of the other, a common challenge in multi-task learning (Sener and Koltun, 2018). Moreover, our theoretical insights indicate that the learned parameters for the generative model should be optimal across all considered data sets. As a result, we advocate for framing this problem of inference across multiple data sets as a multi-objective optimization problem: $\min_{\theta,\phi}\left(-\mathcal{L}^B(\theta, \phi), -\mathcal{L}^T(\theta, \phi)\right)$, that we solve using the Multiple-Gradient Descent Algorithm (Désidéri, 2012), where at each step t, the direction δ_t used for the descent is a convex combination of the gradient of each ELBO:

$$\delta_t = -\alpha_t \nabla \mathcal{L}^B(\theta^t, \phi^t) - (1 - \alpha_t)\nabla \mathcal{L}^T(\theta^t, \phi^t) \tag{4}$$

$$\alpha_t = \arg\min_{\alpha \in [0,1]} \|\alpha \nabla \mathcal{L}^B(\theta^t, \phi^t) + (1 - \alpha)\nabla \mathcal{L}^T(\theta^t, \phi^t)\|_2^2, \tag{5}$$

where the quadratic optimization problem in Equation 5 admits a closed-form solution (Appendix F.1). This procedure provably converges to a Pareto-optimal design point in the batch setting (Zhou et al. (2022) discusses the stochastic setting). In our implementation, we solely rely on gradients of the last layer of the decoder to approximately calculate the optimal weight α, and have observed satisfactory performance. We refer to this approach as the Multiple Objective cVAE (MO-cVAE).

5.2. Interpretable Hyperparameter Selection via Constrained Optimization

The most common approach to regularize models in the comparative analysis literature involves adding a penalization term to the ELBO, leading to solving an unconstrained optimization problem

(U-cVAE). For example, independence constraints are typically enforced via penalization of mutual information approximations, estimated either via the density-ratio trick (Abid and Zou, 2019), or kernel-based embedding of distributions (Weinberger et al., 2022a). This approach has two significant drawbacks. First, it necessitates the calibration of the Lagrangian multiplier. Currently, the only methods available involve either using a preset value (Abid and Zou, 2019) or comparing the scales of loss functions (Weinberger et al., 2022b). Second, the primary goal of the optimization procedure is not to minimize the mutual information between the latent variables, but for the mutual information to be *sufficiently low* for practical considerations.

For these reasons, we instead explore the design of a constrained optimization problem that enhances interpretability and automation for the selection of the Lagrangian parameter:

$$\min_{\theta,\phi} \mathcal{L}(\theta,\phi) = -\mathcal{L}^B(\theta,\phi) - \mathcal{L}^T(\theta,\phi) \;\text{ such that }\; \text{CKA}\left(\hat{q}_\phi(\boldsymbol{z},\boldsymbol{s})\right) \leq \beta, \tag{6}$$

where $\beta > 0$ is a scalar , and $\text{CKA}\left(\hat{q}_\phi(\boldsymbol{z},\boldsymbol{s})\right)$ is the centered kernel alignment metric (Kornblith et al., 2019), a non-parametric measure of correlation between random vectors. Given two positive definite kernels $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, and $l : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$, we may define the cross-covariance operator $C_{\boldsymbol{z},\boldsymbol{s}}$ that embeds the joint distribution $\hat{q}_\phi(\boldsymbol{z},\boldsymbol{s})$ as a linear operator in the RKHS obtained from both kernels (Gretton et al., 2012). Similarly, we may embed each marginal distribution $\hat{q}_\phi(\boldsymbol{z})$ and $\hat{q}_\phi(\boldsymbol{s})$ as linear operators $C_{\boldsymbol{z},\boldsymbol{z}}$ and $C_{\boldsymbol{s},\boldsymbol{s}}$. Then, the CKA is defined as:

$$\text{CKA}\left(\hat{q}_\phi(\boldsymbol{z},\boldsymbol{s})\right) = \frac{\|C_{\boldsymbol{z},\boldsymbol{s}}\|^2_{\text{HS}}}{\|C_{\boldsymbol{z},\boldsymbol{z}}\|_{\text{HS}}\|C_{\boldsymbol{s},\boldsymbol{s}}\|_{\text{HS}}}, \tag{7}$$

where $\|.\|_{\text{HS}}$ designates the Hilbert-Schmidt norm of a linear operator in the RKHS. When the kernels are linear, the CKA metric becomes related to the RV-coefficient (Robert and Escoufier, 1976) as well as Tucker's congruence coefficient (Tucker, 1951), both practically used to estimate correlations between pairs of random vectors. Throughout this manuscript, we use a maximal CKA value of $\beta = 0.05$ to obtain satisfactory performance. To solve the constrained optimization problem, we consider the following equivalent Lagrangian (details appear in Appendix F.2),

$$\min_{\theta,\phi}\max_{\lambda\geq 0} \mathcal{L}^\lambda(\theta,\phi) = \mathcal{L}(\theta,\phi) + \lambda\left(\|C_{\boldsymbol{z},\boldsymbol{s}}\|^2_{\text{HS}} - \beta\|C_{\boldsymbol{z},\boldsymbol{z}}\|_{\text{HS}}\|C_{\boldsymbol{s},\boldsymbol{s}}\|_{\text{HS}}\right). \tag{8}$$

We perform simultaneous gradient descent on (θ,ϕ) and projected gradient ascent on the Lagrangian λ associated with the constraint (Lin et al., 2020):

$$[\theta^{t+1},\phi^{t+1}] = [\theta^t,\phi^t] - \eta_{\text{primal}}\nabla\mathcal{L}^\lambda(\theta^t,\phi^t) \tag{9}$$

$$\lambda^{t+1} = \left[\lambda^t + \eta_{\text{dual}}\left(\|C_{\boldsymbol{z},\boldsymbol{s}}\|^2_{\text{HS}} - \beta\|C_{\boldsymbol{z},\boldsymbol{z}}\|_{\text{HS}}\|C_{\boldsymbol{s},\boldsymbol{s}}\|_{\text{HS}}\right)\right]_+, \tag{10}$$

where $[a]_+ = \max(0,a)$. We obtain a stochastic estimate of the gradient $\nabla\mathcal{L}^\lambda(\theta^t,\phi^t)$ by subsampling data points, as well as latent variables from the variational distribution. We estimate the Hilbert-Schmidt norms using the Hilbert-Schmidt Independence Criterion (Gretton et al., 2012). More precisely, from samples $(z_i,s_i)_{i=1}^M$ from $\hat{q}_\phi(\boldsymbol{z},\boldsymbol{s})$, the kernel matrices K_z (and K_s) are defined as $K_{ij} = k(z_i,z_j)$, and the HSIC is defined as $\text{HSIC}(K,L) = (M-1)^{-2}\text{Tr}(KHLH)$, where $H = I - \frac{1}{M}\mathbf{1}\mathbf{1}^\top$ is a centering matrix. Then, $\text{HSIC}\left(K_z,K_s\right)$ is an unbiased estimator for $\|C_{\boldsymbol{z},\boldsymbol{s}}\|^2_{\text{HS}}$, and we proceed similarly for the remaining terms (Gretton et al., 2012).

In our experiments, we use the Adam optimizer (Kingma and Ba, 2015) for the gradient step described in Equation 9, with a learning rate of $\eta_{\text{primal}} = 0.001$ and gradient ascent with a learning rate of $\eta_{\text{dual}} = 1$ for updating the Lagrangian coefficient in Equation 10. We refer to this method as the COnstrained cVAE (CO-cVAE). The reader will notice that this approach may be combined with the multi-objective optimization approach described above, in which case we refer to it as MO-CO-cVAE.

6. Experiments

We present empirical evidence of our theory of identifiability with a simulation framework, where data is generated with a piece-wise linear mixing function. Then, we apply our proposed optimization framework to a real-world example from single-cell perturbation data analysis. We base our experiments upon the implementation of contrastive DGMs presented in Weinberger et al. (2023). All results are reported with mean and standard deviation over 5 random initializations. Additional experimental results and supplementary metrics appear in Appendix F.9.

6.1. Synthetic Data Experiments

In order to provide empirical evidence for our theory of identifiability, we generated synthetic data in the contrastive analysis framework, according to the generative model described in Figure 1, where f is a four-layers Leaky ReLU neural network (details in Appendix F.4).

Verification of Identifiability Our theory dictates that if the number of latent variables in each space is known, then the contrastive model is identifiable. In addition, we know that the latent variables should be linear transformations of each other, with no leakage between the distinct spaces. To verify this claim, we generated data with $p = q = 5$, and fitted a contrastive DGM with the same estimated number of latent dimensions (no regularization). We refer to those *unregularized* models as MO-cVAE and SO-cVAE, depending on whether the multi-objective optimization procedure was applied or not. We quantified the level of disentanglement using the Pearson Mean Correlation Coefficient after a linear transformation (MCC) between ground truth latent variables, and estimated latent variables (Appendix F.5). More specifically, we calculated the Pearson MCC between $\hat{\boldsymbol{z}}$ and $\boldsymbol{z}$, $\hat{\boldsymbol{s}}$ and $\boldsymbol{s}$ (higher is better), but also between $\hat{\boldsymbol{z}}$ and $\boldsymbol{s}$, and $\hat{\boldsymbol{s}}$ and $\boldsymbol{z}$ (lower is better). As an aggregated metric, we define the δ-MCC $\in [0, 1]$ as

$$\delta\text{-MCC} = \frac{1}{2}\left(\text{MCC}_{\hat{\boldsymbol{z}}\boldsymbol{z}} + \text{MCC}_{\hat{\boldsymbol{s}}\boldsymbol{s}}\right) + \frac{1}{2}\left(\text{MCC}_{\hat{\boldsymbol{s}}\boldsymbol{z}} - \text{MCC}_{\hat{\boldsymbol{z}}\boldsymbol{s}}\right). \tag{11}$$

The results highlight high conservation of each individual latent space, and remarkably low leakage between latent spaces, for both Poisson, and negative binomial observation models (Table 1). For reference, we also fitted a vanilla VAE, with the same architecture and noise model. Because the VAE only yields one set of latent variables, we ran the method with $p + q$ number of latent variables and used contrastive PCA to split the latent space into the background or the salient space (Appendix F.6). We observe poor performance of the VAE for this task, likely because it treats all samples as independent and identically distributed, and ignores additional knowledge about the background data set. This suggests, as expected, that the additional assumptions enforced by contrastive DGMs are necessary for identifiability.

Table 1: Identifiability under assumptions of known dimensions of latent spaces. Best in bold.

Model	Noise	$\text{MCC}_{\hat{\boldsymbol{z}}\boldsymbol{z}}$ (↑)	$\text{MCC}_{\hat{\boldsymbol{z}}\boldsymbol{s}}$ (↓)	$\text{MCC}_{\hat{\boldsymbol{s}}\boldsymbol{z}}$ (↓)	$\text{MCC}_{\hat{\boldsymbol{s}}\boldsymbol{s}}$ (↑)	δ-MCC (↑)
MO-cVAE	**Poisson**	0.91 ± 0.01	$\mathbf{0.08 \pm 0.01}$	$\mathbf{0.07 \pm 0.01}$	$\mathbf{0.94 \pm 0.01}$	$\mathbf{0.85 \pm 0.01}$
SO-cVAE		$\mathbf{0.93 \pm 0.01}$	0.13 ± 0.01	$\mathbf{0.07 \pm 0.02}$	0.92 ± 0.01	0.83 ± 0.01
VAE		0.87 ± 0.04	0.17 ± 0.9	0.14 ± 0.07	0.92 ± 0.04	0.74 ± 0.12
MO-cVAE	**Negative binomial**	$\mathbf{0.93 \pm 0.01}$	0.10 ± 0.01	$\mathbf{0.06 \pm 0.01}$	$\mathbf{0.94 \pm 0.01}$	$\mathbf{0.83 \pm 0.01}$
SO-cVAE		0.92 ± 0.01	$\mathbf{0.08 \pm 0.01}$	0.07 ± 0.01	$\mathbf{0.94 \pm 0.01}$	$\mathbf{0.83 \pm 0.01}$
VAE		0.81 ± 0.10	0.46 ± 0.18	0.37 ± 0.18	0.80 ± 0.10	0.39 ± 0.28

Impact of Misspecification Then, we wanted to illustrate the fact that disentanglement performance drops in the setting of misspecification of the number of latent variables. Towards this goal, we maintained p and q to 5 in the simulation framework, but augmented $\hat{q}$, the dimensionality of $\boldsymbol{s}$ in the inference method. We noticed a strong degradation in the performance, as highlighted in the drop in δ-MCC (Table 2). Careful examination of the individual MCC scores revealed a high leakage from the ground truth background variables $\boldsymbol{z}$ into the estimated salient variables $\hat{\boldsymbol{s}}$ (Table 8 and 9).

Information Constraints Improve Performance We then explored how much the independence constraints could help improve the performance in the case $\hat{q} = 10$. Towards this end, we applied the HSIC penalty with a fixed scaling factor λ, denoted as U-SO-cVAE and U-MO-cVAE (U stands for unconstrained), as well as the constrained optimization procedure with $\beta = 0.05$, denoted as CO-SO-sVAE and CO-MO-cVAE. We report values of the δ-MCC for different values of the regularization strength in Table 3. Although regularization helped partially restore the performance in some sensible range of λ, the constrained optimization approach is more practical as λ is adjusted automatically during training, and achieved competitive performance.

Comparison with other Regularizers In order to justify that the HSIC penalty is a competitive regularizer, we also assessed the performance of the regularization from both ConstrastiveVI (Weinberger et al., 2023), as well as the original cVAE (Abid and Zou, 2019) on the same benchmark in the experiment from Table 3. The ContrastiveVI regularization seems to help (δ-MCC value of 0.75), but its performance remains lower than CO-MO-cVAE. cVAE achieves a poorer result (δ-MCC value of 0.69), slightly improving over the unregularized method.

Multi-objective Optimization We also note that the approach that used multi-objective optimization systematically performed better throughout this benchmark (Tables 1, 2, and 3).

Table 2: Impact of misspecification of latent dimensionality on δ-MCC.

$\hat{q}$	SO-cVAE	MO-cVAE
5	$\mathbf{0.83 \pm 0.01}$	$\mathbf{0.85 \pm 0.01}$
7	0.73 ± 0.03	0.81 ± 0.01
10	0.66 ± 0.01	0.75 ± 0.01
15	0.58 ± 0.02	0.70 ± 0.02

Table 3: Impact of regularization on δ-MCC under misspecification.

Regularization	SO-cVAE	MO-cVAE
U ($\lambda = 0$)	0.66 ± 0.01	0.75 ± 0.01
U ($\lambda = 10$)	0.70 ± 0.02	0.78 ± 0.01
U ($\lambda = 50$)	0.76 ± 0.01	0.79 ± 0.02
U ($\lambda = 100$)	0.66 ± 0.08	0.80 ± 0.01
U ($\lambda = 200$)	0.32 ± 0.10	0.34 ± 0.13
CO	$\mathbf{0.77 \pm 0.01}$	$\mathbf{0.80 \pm 0.01}$

Table 4: Results on real data.

	ARI (↑)	NMI (↑)	ASW (↑)	cMCC-P (↓)	cMCC-S (↓)
MO-CO-cVAE	$\mathbf{0.34 \pm 0.05}$	$\mathbf{0.40 \pm 0.03}$	$\mathbf{0.10 \pm 0.01}$	$\mathbf{0.28 \pm 0.04}$	$\mathbf{0.28 \pm 0.04}$
SO-CO-cVAE	0.31 ± 0.06	0.38 ± 0.03	0.08 ± 0.01	$\mathbf{0.28 \pm 0.02}$	$\mathbf{0.28 \pm 0.02}$
MO-cVAE	0.32 ± 0.09	0.39 ± 0.06	0.07 ± 0.05	0.36 ± 0.06	0.36 ± 0.06
SO-cVAE	0.27 ± 0.02	0.31 ± 0.04	0.04 ± 0.01	0.40 ± 0.01	0.40 ± 0.01
ContrastiveVI	0.30 ± 0.06	0.38 ± 0.05	0.06 ± 0.03	0.34 ± 0.04	0.34 ± 0.05
cVAE	0.27 ± 0.10	0.31 ± 0.08	0.05 ± 0.04	0.36 ± 0.02	0.36 ± 0.02
VAE	0.28 ± 0.05	0.34 ± 0.05	0.06 ± 0.02	0.76 ± 0.10	0.74 ± 0.11

6.2. Single-cell Perturbation Analysis

As an application to real data, we present an experiment focused on characterizing the effect of genetic perturbations on single-cell gene expression levels, a central problem in modern molecular biology (Dixit et al., 2016; Norman et al., 2019). In these data sets, we observe two important sources of variation. First, cells react to the genetic perturbation they were exposed to, and modulate the expression level of their genes. Second, there is some inherent variation in gene expression levels that happens due to the cells going through biological processes such as stages of the cell cycle, or simply due to heterogeneity in the initial population of cells. An important problem therefore consists in disentangling these effects, and ContrastiveVI (Weinberger et al., 2023) was conceived with this goal in mind.

We focus on a recent data set (Norman et al., 2019) that contains expression profiles from $33,820$ erythroleukemia (cancer) cells, after interventions targeting each of 105 genes, as well as 131 pairs of those same genes. Each measurement from a single-cell combines the identity of the intervention (target genes) and a count vector where each entry is the expression level of each gene in the genome. Because of experimental limitations (Grün et al., 2014), we observe signal only for a subset of several thousand genes out of the approximately $20,000$ genes in the genome. Here we selected $d = 2,000$ genes. The goal of the experiment was to manipulate gene pairs and measure the resulting changes in cell state to gain insights into how complex phenotypes emerge and identify genes that interact to promote differentiation to a specific cell state.

To assess the performance of each method, we first evaluated how well the salient space captures the effect of perturbations. Specifically, we clustered cells based on their salient embeddings and assessed how well those clusters overlap with known biological labels attributed to each of the perturbations, following Weinberger et al. (2023). We reported the Adjusted Rand Index (ARI), the Normalized Mutual Information (NMI), as well as the Average Silhouette Width (ASW). In addition, we assessed the overlap in content between the two latent spaces by training a linear regression model from one space to the other, and reporting the MCC (cMCC-P refers to the Pearson correlation and cMCC-S to the Spearman correlation).

We applied our evaluation pipeline for the VAE, MO-CO-cVAE, SO-CO-cVAE, as well as non-regularized variants that we note as MO-cVAE and SO-cVAE (Table 4). We observed again that the multi-objective variant outperforms the single-objective method, and that constrained optimization provided an effective regularization strategy. As a point of comparison, we also reported the performance of ContrastiveVI and the original cVAE, and noticed that MO-CO-cVAE performed better. This suggests that our novel methodology is effective in practice. Qualitative comparisons appear in Appendix F.8.

7. Conclusion

This study examines the identifiability properties of recently proposed Deep Generative Models (DGMs) for comparative analysis. Our analysis highlights the block-wise identifiability of many recent contrastive and multi-group DGMs, drawing connections between data from differing sources and the broader landscape of causal representation learning. A significant contribution is the extension of non-linear ICA results to count distributions, an area previously less explored. We further assess the challenges associated with estimating the number of latent variables and provide empirical evidence that regularization is beneficial under those specific circumstances. Building on our theoretical findings, we introduce a methodology grounded in multi-objective and constrained optimization principles. As the field continues to employ DGMs in diverse scientific applications, it is crucial to emphasize the dual objectives of accurate model fit and interpretability, ensuring that the generated models are both robust and scientifically valuable.

Acknowledgments

We thank Sébastien Lachapelle for providing insights and early guidance through the conception of this work. We acknowledge Kelvin Chen, Taka Kudo for discussions about modeling single-cell perturbation data sets. We thank Jeffrey Spence, Hanchen Wang as well as Saeed Saremi for feedback on this manuscript.

Disclosures: Romain Lopez, Jan Christian Huetter, Ehsan Hajiramezanali and Aviv Regev are employees of Genentech, and / or have equity in Roche. Jonathan Pritchard acknowledges support from grant R01HG008140 from the National Human Genome Research Institute. Aviv Regev is a co-founder and equity holder of Celsius Therapeutics and an equity holder in Immunitas. She was an SAB member of ThermoFisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics, and Asimov until July 31st, 2020.

References

Abubakar Abid and James Zou. Contrastive variational autoencoder enhances salient features. *arXiv preprint arXiv:1902.04601*, 2019.

Abubakar Abid, Martin J Zhang, Vivek K Bagaria, and James Zou. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature Communications*, 9(1):1–7, 2018.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, pages 372–407, 2023.

Maxime Bergeron, Nicholas Fung, John Hull, Zissis Poulos, and Andreas Veneris. Variational autoencoders: A hands-off approach to volatility. *The Journal of Financial Data Science*, 2022.

Philippe Boileau, Nima S Hejazi, and Sandrine Dudoit. Exploring high-dimensional biological data with sparse contrastive principal component analysis. *Bioinformatics*, 36(11):3422–3430, 2020.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In *Advances in Neural Information Processing Systems*, 2023.

Joe Davison, Kristen Severson, and Soumya Ghosh. Cross-population variational autoencoders. In *4th workshop on Bayesian Deep Learning (NeurIPS)*, 2019.

Roberta De Vito, Ruggero Bellio, Lorenzo Trippa, and Giovanni Parmigiani. Multi-study factor analysis. *Biometrics*, 75(1):337–346, 2019.

Jean-Antoine Désidéri. Multiple-gradient descent algorithm (MGDA) for multi-objective optimization. *Comptes Rendus Mathematique*, 350(5-6):313–318, 2012.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7): 1853–1866, 2016.

Jose Gallego-Posada, Juan Ramirez, Akram Erraqabi, Yoshua Bengio, and Simon Lacoste-Julien. Controlled sparsity via constrained optimization or: How I learned to stop tuning penalties and love constraints. *Advances in Neural Information Processing Systems*, 35:1253–1266, 2022.

Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, Maxime Langevin, et al. A Python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, 40(2):163–166, 2022.

Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber, and Xavier Alameda-Pineda. *Dynamical Variational Autoencoders: A Comprehensive Review*, volume 15. Foundations and Trends® in Machine Learning, 2021.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Dominic Grün, Lennart Kester, and Alexander Van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature Methods*, 11(6):637–640, 2014.

Stefan Harmeling, Andreas Ziehe, Motoaki Kawanabe, and Klaus-Robert Müller. Kernel-based nonlinear blind source separation. *Neural Computation*, 15(5):1089–1124, 2003.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations*, 2017.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469, 2017.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. Independent component analysis. *Studies in Informatics and Control*, 11(2):205–207, 2002.

Yibo Jiang and Bryon Aragam. Learning nonparametric latent causal graphs with unknown interventions. In *Advances in Neural Information Processing Systems*, 2023.

Andrew Jones, F William Townes, Didong Li, and Barbara E Engelhardt. Contrastive latent variable modeling with application to case-control sequencing experiments. *Annals of Applied Statistics*, 2022.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37:183–233, 1999.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217, 2020.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *International Conference on Learning Representations*, 2014.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models under mixture priors without auxiliary information. In *Advances in Neural Information Processing Systems*, 2022.

Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial disentanglement for domain adaptation. In *International Conference on Machine Learning*, volume 162, pages 11455–11472, 2022.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529, 2019.

Sébastien Lachapelle and Simon Lacoste-Julien. Partial disentanglement via mechanism sparsity. *Conference on Uncertainty and Artificial Intelligence: Causal Representation Learning workshop*, 2022.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Conference on Causal Learning and Reasoning*, pages 428–484, 2022.

Sébastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. Additive decoders for latent variables identification and Cartesian-product extrapolation. In *Advances in Neural Information Processing Systems*, 2023.

Didong Li, Andrew Jones, and Barbara Engelhardt. Probabilistic contrastive principal component analysis. *arXiv preprint arXiv:2012.07977*, 2020.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093, 2020.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124, 2019.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018a.

Romain Lopez, Jeffrey Regier, Michael I Jordan, and Nir Yosef. Information constraints on auto-encoding variational bayes. *Advances in Neural Information Processing Systems*, 31, 2018b.

Romain Lopez, Adam Gayoso, and Nir Yosef. Enhancing scientific discoveries in molecular biology with deep generative models. *Molecular Systems Biology*, 16(9):e9198, 2020.

Romain Lopez, Natasa Tagasovska, Stephen Ra, Kyunghyun Cho, Jonathan Pritchard, and Aviv Regev. Learning causal representations of single cells via sparse mechanism shift modeling. In *Conference on Causal Learning and Reasoning*, pages 662–691, 2023.

Robin Louiset, Edouard Duchesnay, Antoine Grigis, Benoit Dufumier, and Pietro Gori. SepVAE: a contrastive VAE to separate pathological patterns from healthy ones. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*, 2023.

Thomas M Norman, Max A Horlbeck, Joseph M Replogle, Alex Y Ge, Albert Xu, Marco Jost, Luke A Gilbert, and Jonathan S Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

Zhi-Yong Ran and Bao-Gang Hu. Parameter identifiability in statistical machine learning: a review. *Neural Computation*, 29(5):1151–1203, 2017.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286, 2014.

P. Robert and Y. Escoufier. A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 25(3):257–265, 1976.

Adrià Ruiz, Oriol Martinez, Xavier Binefa, and Jakob Verbeek. Learning disentangled representations with reference-based variational autoencoders. In *ICLR workshop on Learning from Limited Labeled Data*, 2019.

Ruslan Salakhutdinov and Hugo Larochelle. Efficient learning of deep Boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pages 693–700, 2010.

Theofanis Sapatinas. Identifiability of mixtures of power-series distributions and related characterizations. *Annals of the Institute of Statistical Mathematics*, 47:447–459, 1995.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Kristen A Severson, Soumya Ghosh, and Kenney Ng. Unsupervised learning with contrastive latent variable models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4862–4869, 2019.

Alexander Shapiro. Identifiability of factor analysis: Some results and open problems. *Linear Algebra and its Applications*, 70:1–7, 1985.

Henning Sprekeler, Tiziano Zito, and Laurenz Wiskott. An extension of slow feature analysis for nonlinear blind source separation. *The Journal of Machine Learning Research*, 15(1):921–947, 2014.

Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired multi-domain causal representation learning. In *Advances in Neural Information Processing Systems*, 2023.

Henry Teicher. Identifiability of mixtures. *Annals of Mathematical Statistics*, 32:244–248, 1961.

Ledyard R Tucker. *A Method for Synthesis of Factor Analysis Studies*, volume 984. Educational Testing Service Princeton, NJ, 1951.

Cédric Villani. *Optimal Transport: Old and New*. Springer Science & Business Media, 2008.

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in Neural Information Processing Systems*, 34:16451–16467, 2021.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David M Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Advances in Neural Information Processing Systems*, 2023.

Ethan Weinberger, Nicasia Beebe-Wang, and Su-In Lee. Moment matching deep contrastive latent variable models. In *International Conference on Artificial Intelligence and Statistics*, pages 2354–2371, 2022a.

Ethan Weinberger, Romain Lopez, Jan-Christian Huetter, and Aviv Regev. Disentangling shared and group-specific variations in single-cell transcriptomics data with multiGroupVI. In *Machine Learning in Computational Biology*, volume 200, pages 16–32, 2022b.

Ethan Weinberger, Chris Lin, and Su-In Lee. Isolating salient variations of interest in single-cell transcriptomic data with contrastiveVI. *Nature Methods*, 2023.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Baichuan Yuan, Xiaowei Wang, Jianxin Ma, Chang Zhou, Andrea L Bertozzi, and Hongxia Yang. Variational autoencoders for highly multivariate spatial point processes intensities. In *International Conference on Learning Representations*, 2019.

Jiaqi Zhang, Chandler Squires, Kristjan Greenewald, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. In *Advances in Neural Information Processing Systems*, 2023.

Shiji Zhou, Wenpeng Zhang, Jiyan Jiang, Wenliang Zhong, Jinjie GU, and Wenwu Zhu. On the convergence of stochastic multi-objective gradient manipulation and beyond. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

James Yang Zou, Daniel Hsu, David C Parkes, and Ryan Prescott Adams. Contrastive learning using spectral methods. *Advances in Neural Information Processing Systems*, 2013.

Appendices

Appendix A. Related Work

Interventional Causal Representation Learning Several recent works have investigated the setting of learning from multiple data sets with interventional shifts in latent space. For example, Lachapelle et al. (2022) proposed an identifiable non-linear ICA based on the assumption that a rich set of interventional data is available, where each intervention shifts the sparse set of sufficient statistics of the latent variable prior distribution. Lopez et al. (2023) proposed an application of this theory to the setting of modeling single-cell perturbation data, and reported empirical evidence, based on simulations, that the identifiability guarantee might hold for count data. Our work is distinct from these as it considers the setting of a small number of data sets, and is mainly concerned with subspace identification, but it does provide a first line of attack towards extending the results from Lachapelle et al. (2022) for counting observational noise. Ahuja et al. (2023) recently proposed a framework for proving identifiability of noiseless auto-encoders under the assumption of a large set of interventions, and a polynomial decoder. Buchholz et al. (2023); von Kügelgen et al. (2023); Jiang and Aragam (2023) are concerned with the identifiability of non-linear ICA, under interventional data and with a general class of non-linear mixing functions (either parametric, or non-parametric). However, the assumptions made by these works are restricted to stochastic interventions, which makes them not applicable to our problem. Interestingly, Lemma 8, Appendix D.1 from Buchholz et al. (2023) points out that non-stochastic interventions create some form of unidentifiability. Consequently, they did not study it in detail (unlike our work). Also, such works require the availability of data from as many interventions as the number of the latent dimensions, while our work solely considers two separate environments.

Identifiability of Modular Representations Several recent works specifically investigated the prospect of block-wise identifiability. Lachapelle and Lacoste-Julien (2022) proved that under a relaxation of the assumptions from Lachapelle et al. (2022), we may obtain only disentanglement by block (when interventions do not dissect enough the latent space to recover the ground truth mixing function). Lachapelle et al. (2023) investigated the setting of additive decoders, where each decoder uses only a block of latent variables. In this setting, the goal is to prove the block-wise identifiability of the latent variables. The definitions of block-wise identifiability in these recent works (Lachapelle et al., 2023; Von Kügelgen et al., 2021) are essentially equivalent to the ones considered in this manuscript. Kong et al. (2022) applied non-linear ICA theory to the domain adaptation problem, and showed block-wise identifiability of the effect of the domain with the predicted outcome under their latent variable model.

Source Matching Across Domains In the classical linear ICA problem, we are interested in learning z from data generated as $x = Wz + \epsilon$. Framing the contrastive analysis problem in the paradigm of linear ICA, we would observe background data $x^b = W^b z + \epsilon^b$ as well as target data $x^t = W^t s + W^b z + \epsilon^t$. The target data set has been generated with additional sources s that we would like to identify collectively and to separate from the background sources z. In the case of genomics, the parameters W^t are also relevant, as they encode which genes are associated with which components of the novel sources s. For example, in Boileau et al. (2020) the sparse entries of the matrix W^t are used to identify genes associated with leukemia. The problem of matching sources across different ICA models is treated in Sturma et al. (2023), although in the context of the more general problem, in which the two data sets may be composed of different observable quantities (= modalities). Their solution considers an idealized scenario with a linear mixing function, and no

observation noise, but could provide a reasonable baseline derived from linear ICA. We found that the performance of the method was not competitive on the simulated data, likely because the mixing function used for generating the data is not linear.

Appendix B. Multi-group DGMs

For the sake of completeness, we describe the framework of multi-group analysis (Weinberger et al., 2022b). We note that this framework has also been referred to as cross-population deep generative modeling in Davison et al. (2019).

B.1. Generative model

When dealing with multiple data sources (focusing on two data sets for the sake of simplicity), one immediate question for data exploration is to enumerate patterns that are shared by all data sets versus the ones that appear only in one of the data sets but not the other. This is the objective of a Multi-Group analysis model (Figure 1, right). In this model, we have three blocks of latent variables. First, latent variable

$$\boldsymbol{z} \sim \text{Normal}(0, I_p), \tag{12}$$

that encodes shared variation between the two data sets. Then, latent variable

$$\boldsymbol{t^1} \sim \text{Normal}(0, I_q), \tag{13}$$

encodes variation that is unique to the first data set. Similarly,

$$\boldsymbol{t^2} \sim \text{Normal}(0, I_q), \tag{14}$$

encodes variation that is unique to the second data set. Observations x are then drawn according to an exponential family distribution, with a mixing function f:

$$\boldsymbol{x} \sim \textsc{ExpFam}\left(f(\boldsymbol{z}, \boldsymbol{t^1}, \boldsymbol{t^2})\right). \tag{15}$$

B.2. Data Set Definition

Interestingly, we have no data available from the distribution $p_\theta(\boldsymbol{x})$, because we observe data from either data set, where one of the variables is inactive ($\boldsymbol{t^1} = \mathbf{0}$ or $\boldsymbol{t^2} = \mathbf{0}$), as introduced in Weinberger et al. (2022b) as well as Davison et al. (2019). To formalize this, we model it as a hard intervention (Pearl, 2009), akin to interventional causal representation learning (Ahuja et al., 2023). For example, data set 1 is generated by sampling from the distribution $p_\theta\left(\boldsymbol{x} \mid \text{do}(\boldsymbol{t^2} = \mathbf{0})\right)$, and we operate similarly for the data set 2, sampled from $p_\theta\left(\boldsymbol{x} \mid \text{do}(\boldsymbol{t^1} = \mathbf{0})\right)$. The reader will notice that the setting of contrastive analysis (Figure 1, left) is a particular instance of this model, when the mixing function is constant with respect to one of the group-specific latent variables (e.g., $\boldsymbol{t^1}$).

B.3. Variational Inference

Both of the data likelihoods $p_\theta\left(\boldsymbol{x} \mid \text{do}(\boldsymbol{t^1} = \mathbf{0})\right)$ and $p_\theta\left(\boldsymbol{x} \mid \text{do}(\boldsymbol{t^2} = \mathbf{0})\right)$ are intractable. Therefore, Weinberger et al. (2022b) as well as Davison et al. (2019) both proceeded to posterior approximation with variational inference.

For each sample in data set 1, the variational distribution is mean-field $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})q_\phi(\boldsymbol{t^1} \mid \boldsymbol{x})$. Then, the evidence lower bound (ELBO) is written as:

$$\mathcal{L}^1(\theta, \phi) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})q_\phi(\boldsymbol{t^1}|\boldsymbol{x})} \log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{t^1}, \boldsymbol{0})}{q_\phi(\boldsymbol{z} \mid \boldsymbol{x})q_\phi(\boldsymbol{t^1} \mid \boldsymbol{x})}.$$

For each sample in data set 2, the variational distribution is $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})q_\phi(\boldsymbol{t^2} \mid \boldsymbol{x})$. Then, the evidence lower bound is written as:

$$\mathcal{L}^2(\theta, \phi) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})q_\phi(\boldsymbol{t^2}|\boldsymbol{x})} \log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{0}, \boldsymbol{t^2})}{q_\phi(\boldsymbol{z} \mid \boldsymbol{x})q_\phi(\boldsymbol{t^2} \mid \boldsymbol{x})}.$$

Both frameworks propose to optimize the following composite ELBO:

$$\mathcal{L}(\theta, \phi) = \mathcal{L}^1(\theta, \phi) + \mathcal{L}^2(\theta, \phi).$$

This composite ELBO may be used as an objective function for maximization, in conjunction with adequate regularization of the neural networks parameterizing the variational distribution as a function of the input data. In Weinberger et al. (2022a), the regularization ensures that the output of the neural network parameterizing the variational posterior for the latent variables $\boldsymbol{t^1}$ and $\boldsymbol{t^2}$ (i.e., their mean and diagonal variance vector) is close to zero for data points where the value of $\boldsymbol{t^1}$ or $\boldsymbol{t^2}$ should be zero. For example, for a point x^1 from data set 1, the regularization will penalize the sum of the square mean and the variance of $q_\phi(\boldsymbol{t^2} \mid \boldsymbol{x})$, which corresponds to the Wasserstein distance between $q_\phi(\boldsymbol{t^2} \mid \boldsymbol{x})$ and the Dirac distribution centered at $\boldsymbol{0}$.

Appendix C. A Theory of Identifiability for Noiseless Comparative Analysis DGMs

C.1. Counterexample of Identifiability for Non-linear Contrastive Analysis

Data Generating Model Let us assume we observe the target data according to the following contrastive deep generative model:

$$\boldsymbol{z} \sim \text{Normal}(0, I_p) \tag{16}$$
$$\boldsymbol{s} \sim \text{Normal}(0, I_q) \tag{17}$$
$$\boldsymbol{x} \sim \text{Normal}(f(\boldsymbol{z}, \boldsymbol{s}), \sigma^2 I_d). \tag{18}$$

We also observe a background data set from the distribution $p_\theta\left(\boldsymbol{x} \mid \text{do}(\boldsymbol{s} = \boldsymbol{0})\right)$. In this example, we consider $p = 2$ and $q = 1$ to demonstrate that there exist mixing functions f that cannot be identified from data. Towards this end, we build a function $\tilde{f}$ such that the resulting data distribution is identical to that of f, but such that $\tilde{f}$ is not subspace disentangled with respect to f.

A Key Diffeomorphism The key idea for this counterexample consists in using a diffeomorphism of $\mathbb{R}^3$ that preserves volumes, as well as Euclidean distances to the origin. We consider the following diffeomorphism:

$$\Phi : \begin{pmatrix} z_1 \\ z_2 \\ s \end{pmatrix} \mapsto \begin{pmatrix} z_1 \cos s - z_2 \sin s \\ z_1 \sin s + z_2 \cos s \\ s \end{pmatrix}. \tag{19}$$

To verify that Φ is norm-preserving, we denote by R_s the rotation operator in two dimensions and simply calculate

$$\forall z_1, z_2, s \in \mathbb{R}^3, \|\Phi(z_1, z_2, s)\|_2^2 = \|R_s((z_1, z_2)\|_2^2 + s^2 = \|(z_1, z_2, s)\|_2^2. \tag{20}$$

To verify that Φ is volume-preserving, we may calculate the Jacobian determinant:

$$|D_\Phi(z_1, z_2, s)| = \begin{vmatrix} \cos s & -\sin s & 0 \\ \sin s & \cos s & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1. \tag{21}$$

Next, we make use of the following lemma.

Lemma 1 *Let Φ be a diffeomorphism of $\mathbb{R}^d$. Let us assume that Φ preserves the Euclidean norm, that is that for all $\boldsymbol{u} \in \mathbb{R}^d$, $\|\Phi(\boldsymbol{u})\|_2 = \|\boldsymbol{u}\|_2$. Let us also assume that Φ is volume-preserving, meaning that for all $\boldsymbol{u} \in \mathbb{R}^d$, $|D_\Phi(\boldsymbol{u})| \in \{-1, 1\}$. Then, Φ leaves the isotropic Gaussian distribution invariant, that is for $\boldsymbol{x} \sim$ Normal$(0, I_d)$, $\boldsymbol{x} \overset{d}{=} \Phi(\boldsymbol{x})$.*

Proof To show that $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}} = \Phi(\boldsymbol{x})$ are equal in distribution, we will show equality of the characteristic functions.

Let $\boldsymbol{t} \in \mathbb{R}^d$. The characteristic function $\phi_{\tilde{\boldsymbol{x}}}(\boldsymbol{t})$ of random variable $\tilde{\boldsymbol{x}}$ is defined as:

$$\phi_{\tilde{\boldsymbol{x}}}(\boldsymbol{t}) = \mathbb{E}_{\tilde{\boldsymbol{x}}} e^{i\boldsymbol{t}^\top \tilde{\boldsymbol{x}}} = \int e^{i\boldsymbol{t}^\top \tilde{\boldsymbol{x}}} d\mathbb{P}_{\tilde{\boldsymbol{x}}}. \tag{22}$$

Using the change of variable formula, we have:

$$\phi_{\tilde{\boldsymbol{x}}}(\boldsymbol{t}) = \int_{\mathbb{R}^d} e^{i\boldsymbol{t}^\top \tilde{\boldsymbol{x}}} |D_{\Phi^{-1}}(\tilde{\boldsymbol{x}})| p_{\boldsymbol{x}}(\Phi^{-1}(\tilde{\boldsymbol{x}})) d\tilde{\boldsymbol{x}}. \tag{23}$$

Now, because Φ is a volume-preserving diffeomorphism, we have $|D_{\Phi^{-1}}(\tilde{\boldsymbol{x}})| = 1$ for all $\tilde{\boldsymbol{x}} \in \mathbb{R}^d$. And, because $p_{\boldsymbol{x}}$ is the density of the isotropic Gaussian distribution, we have that $p_{\boldsymbol{x}}$ depends only on the distance to the origin (i.e., the Euclidean norm). However, Φ is norm-preserving so we have $p_{\boldsymbol{x}}(\Phi^{-1}(\tilde{\boldsymbol{x}})) = p_{\boldsymbol{x}}(\tilde{\boldsymbol{x}})$ for all $\tilde{\boldsymbol{x}} \in \mathbb{R}^d$. Therefore, we have:

$$\phi_{\tilde{\boldsymbol{x}}}(\boldsymbol{t}) = \int_{\mathbb{R}^d} e^{i\boldsymbol{t}^\top \tilde{\boldsymbol{x}}} p_{\boldsymbol{x}}(\tilde{\boldsymbol{x}}) d\tilde{\boldsymbol{x}} = \phi_{\boldsymbol{x}}(\boldsymbol{t}), \tag{24}$$

which concludes the proof. □

Unidentifiability Because any diffeomorphism that preserves norm and volume leaves the isotropic Gaussian distribution invariant, $\Phi(\boldsymbol{z}, \boldsymbol{s}) \overset{d}{=} (\boldsymbol{z}, \boldsymbol{s})$. Furthermore, we notice that the restriction of Φ to the domain $\mathbb{R}^2 \times \{0\}$ satisfies the same properties, and therefore $\Phi(\boldsymbol{z}, 0) \overset{d}{=} (\boldsymbol{z}, 0)$. Now, for any non-trivial f, we define $\tilde{f} = f \circ \Phi$ so that $v = \Phi$. We have constructed a case of two functions f and $\tilde{f}$ where the data distributions are equal, but $\tilde{f}$ is not subspace disentangled with respect to f, because v is not compatible with the Cartesian product $\mathbb{R}^2 \times \mathbb{R}$. Indeed, the first component of $v(\boldsymbol{z}, \boldsymbol{s})$ depends non-trivially on $\boldsymbol{s}$.

Extension to the Multi-group setting The reader will notice that this counterexample may be easily adapted to the multi-group setting, by rotating the block of z by and angle of $\boldsymbol{t^1} + \boldsymbol{t^2}$.

C.2. Identifiability of ReLU Contrastive Analysis DGMs

We start by stating a general result from Kivva et al. (2022) that we will apply to the comparative analysis setting.

Theorem 4 (Theorem D.4 from Kivva et al. (2022)) *Let $f, g : \mathbb{R}^m \to \mathbb{R}^n$ be continuous piecewise affine functions such that f and g are both injective for almost every point in their respective images $f(\mathbb{R}^m)$ and $g(\mathbb{R}^m)$. Let $Z \sim \sum_{i=1}^{J} \lambda_i Normal(\mu_i, \Sigma_i)$ and $Z' \sim \sum_{j=1}^{J'} \lambda_j Normal(\mu'_j, \Sigma'_j)$ be a pair of variables with GMM distribution (in reduced form). Suppose that $f(Z)$ and $g(Z')$ are equally distributed. Let $\mathcal{D} \subseteq \mathbb{R}^n$ be a connected open set such that f and g are injective onto $\mathcal{D}$. Then, there exists an affine transformation $h : \mathbb{R}^m \to \mathbb{R}^m$ such that $h(Z) \stackrel{d}{=} Z'$ and $g(z) = \left(f \circ h^{-1}\right)(z)$ for every $z \in g^{-1}(\mathcal{D})$.*

We now prove our main result.

Theorem 1 (Identifiability Theorem) *Let the ground truth mixing function f and the learned mixing function $\tilde{f}$ both be continuous and injective piece-wise affine mixing functions such that $f(\boldsymbol{z}, \boldsymbol{s}) \stackrel{d}{=} \tilde{f}(\boldsymbol{z}, \boldsymbol{s})$ and $f(\boldsymbol{z}, \boldsymbol{0}) \stackrel{d}{=} \tilde{f}(\boldsymbol{z}, \boldsymbol{0})$. Then, $\tilde{f}$ is subspace disentangled with respect to f and the noiseless version of the contrastive analysis model is subspace identifiable.*

Proof Let f and $\tilde{f}$ be two continuous injective piecewise linear functions such that $f(\boldsymbol{z}, \boldsymbol{s}) \stackrel{d}{=} \tilde{f}(\boldsymbol{z}, \boldsymbol{s})$ and $f(\boldsymbol{z}, \boldsymbol{0}) \stackrel{d}{=} \tilde{f}(\boldsymbol{z}, \boldsymbol{0})$. The key idea of the proof is to first study the implications of each of the two hypotheses on equality in distribution independently. By applying the result of Kivva et al. (2022), Theorem 4, we reduce the complexity of the problem to a linear equivalence class of functions. Then, simple linear algebra allows us to conclude.

We start by applying the result of Kivva et al. (2022) to the target data set, as it is the most straightforward. f and $\tilde{f}$ are two continuous injective piecewise affine functions. Because for the Gaussian vector $(\boldsymbol{z}, \boldsymbol{s})$, we have $f(\boldsymbol{z}, \boldsymbol{s}) \stackrel{d}{=} \tilde{f}(\boldsymbol{z}, \boldsymbol{s})$, we may apply Theorem 4 and obtain that there exists an affine transformation $h_1 : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^p \times \mathbb{R}^q$ such that $h_1(\boldsymbol{z}, \boldsymbol{s}) \stackrel{d}{=} (\boldsymbol{z}, \boldsymbol{s})$, and

$$\forall (\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^p \times \mathbb{R}^q, \tilde{f}(\boldsymbol{u}, \boldsymbol{v}) = \left(f \circ h_1^{-1}\right)(\boldsymbol{u}, \boldsymbol{v}). \tag{25}$$

Now, let us consider the background data set. Let $\mathcal{D}^b = \mathbb{R}^p \times \{0\}$ denote the domain of the latent variables that generate the background data set, and let $f|_{\mathcal{D}^b}$ be the the restriction of the piecewise affine function f to the domain $\mathcal{D}^b$. Because f is piecewise affine, $f|_{\mathcal{D}^b}$ is also piecewise affine. Because the restriction of an injective function is injective, $f|_{\mathcal{D}^b}$ is injective. Given that $f(\boldsymbol{z}, \boldsymbol{0}) \stackrel{d}{=} \tilde{f}(\boldsymbol{z}, \boldsymbol{0})$, we may apply Theorem 4 and obtain that there exists an affine transformation $h_0 : \mathbb{R}^p \to \mathbb{R}^p$ such that $h_0(\boldsymbol{z}) \stackrel{d}{=} \boldsymbol{z}$, and

$$\forall \boldsymbol{u} \in \mathbb{R}^p, \tilde{f}\Big|_{\mathcal{D}^b}(\boldsymbol{u}, \boldsymbol{0}) = \left(f\big|_{\mathcal{D}^b} \circ h_0^{-1}\right)(\boldsymbol{u}, \boldsymbol{0}). \tag{26}$$

Finally, we appeal to a short linear algebraic argument to conclude. Because h_0 and h_1 are affine maps preserving the isotropic Gaussian distribution, they must be linear (i.e., the offset term is zero) in order to preserve the mean. We may write the functions as

$$h_0(\boldsymbol{u}) = R_0 \boldsymbol{u} \tag{27}$$

$$h_1(\boldsymbol{u}, \boldsymbol{v}) = \left(R_{11}\boldsymbol{u} + R_{12}\boldsymbol{v}, R_{21}\boldsymbol{u} + R_{22}\boldsymbol{v}\right). \tag{28}$$

Now, because of the injectivity of the functions f and $\tilde{f}$, we know that h_0 and h_1 are equal for all values of $\boldsymbol{u}$ whenever $\boldsymbol{v} = \mathbf{0}$. Therefore, we have that $R_{11} = R_0$, and $R_{21} = 0$. We may rewrite our functions as

$$h_0(\boldsymbol{u}) = R_0\boldsymbol{u} \tag{29}$$

$$h_1(\boldsymbol{u}, \boldsymbol{v}) = (R_0\boldsymbol{u} + R_{12}\boldsymbol{v}, R_{22}\boldsymbol{v})\,. \tag{30}$$

Now, in order to preserve the covariance of the Gaussian vector, we have

$$\begin{pmatrix} I_p & 0 \\ 0 & I_q \end{pmatrix} = \begin{pmatrix} R_0 & R_{12} \\ 0 & R_{22} \end{pmatrix} \begin{pmatrix} R_0^\top & 0 \\ R_{12}^\top & R_{22}^\top \end{pmatrix} = \begin{pmatrix} R_0R_0^\top + R_{12}R_{12}^\top & R_{12}R_{22}^\top \\ R_{22}R_{12}^\top & R_{22}R_{22}^\top \end{pmatrix}. \tag{31}$$

Therefore, we have that $R_{22}R_{22}^\top = I$, and R_{22} is an orthogonal matrix. But additionally, $R_{12}R_{22}^\top = 0$, and because R_{22} is invertible, it implies that $R_{12} = 0$. Therefore R_{12} is the null function and we have proved that $v = f^{-1} \circ \tilde{f}$ is compatible with respect to the Euclidean decomposition $\mathbb{R}^p \times \mathbb{R}^q$. □

C.3. Identifiability of ReLU Multi-Group Analysis DGMs

We start this section by introducing the notation and the definition of subspace disentanglement in this specific setting. Let us also introduce the domains for the latent variables $\mathcal{D}^1 = \mathcal{Z} \times \mathcal{T}^1 \times \{0\}$ and $\mathcal{D}^2 = \mathcal{Z} \times \{0\} \times \mathcal{T}^2$. The subspace identifiability corresponds to:

Definition 5 (Subspace Disentanglement) *A learned mixing function $\tilde{f}$ is said to be subspace-disentangled with respect to the ground truth mixing function f if $f(\mathcal{D}^1) = \tilde{f}(\mathcal{D}^1)$ and $f(\mathcal{D}^2) = \tilde{f}(\mathcal{D}^2)$ and the mapping $v = f^{-1} \circ \tilde{f}$ is a diffeomorphism such that both $v|_{\mathcal{D}^1}$ and $v|_{\mathcal{D}^2}$ are compatible with respect to the subspaces $\mathcal{D}^1$ and $\mathcal{D}^2$.*

When this property is not verified, this would imply that the learned mixing function $\tilde{f}$ does not decompose the signal in the feature space $\mathcal{X}$ into the same shared $\mathcal{Z}$ and private spaces $\mathcal{T}^1$, and $\mathcal{T}^2$. Now we proceed to a statement of the generalization of Theorem 1 to the multi-group setting.

Theorem 3 *Let the ground truth mixing function f and the learned mixing function $\tilde{f}$ both be continuous and injective piecewise affine mixing functions such that $f(\boldsymbol{z}, \boldsymbol{t^1}, \mathbf{0}) \overset{d}{=} \tilde{f}(\boldsymbol{z}, \boldsymbol{t^1}, \mathbf{0})$ and $f(\boldsymbol{z}, \mathbf{0}, \boldsymbol{t^2}) \overset{d}{=} \tilde{f}(\boldsymbol{z}, \mathbf{0}, \boldsymbol{t^2})$. Then, $\tilde{f}$ is subspace disentangled with respect to f and the noiseless version of the multi-group analysis model is subspace identifiable.*

Proof Let f and $\tilde{f}$ be two continous injective piecewise linear functions such that

$$f(\boldsymbol{z}, \boldsymbol{t^1}, \mathbf{0}) \overset{d}{=} \tilde{f}(\boldsymbol{z}, \boldsymbol{t^1}, \mathbf{0}) \quad \text{and} \quad f(\boldsymbol{z}, \mathbf{0}, \boldsymbol{t^2}) \overset{d}{=} \tilde{f}(\boldsymbol{z}, \mathbf{0}, \boldsymbol{t^2}). \tag{32}$$

We proceed very similar to the contrastive analysis case.

Let $\mathcal{D}^1 = \mathbb{R}^p \times \mathbb{R}^q \times \{0\}$ denote the domain for the latent variables that generate data set 1. Let $f|_{\mathcal{D}^1}$ be the the restriction of the piecewise affine function f to the domain $\mathcal{D}^1$. Because f is piecewise affine, $f|_{\mathcal{D}^1}$ is also piecewise affine. Because the restriction of an injective function is injective, $f|_{\mathcal{D}^1}$ is injective. Given that $f(\boldsymbol{z}, \boldsymbol{t^1}, \mathbf{0}) \overset{d}{=} \tilde{f}(\boldsymbol{z}, \boldsymbol{t^1}, \mathbf{0})$, we may apply Theorem 4 and obtain that there exists an affine transformation $h_1 : \mathcal{D}^1 \to \mathcal{D}^1$ such that $h_1(\boldsymbol{z}, \boldsymbol{t^1}, \mathbf{0}) \overset{d}{=} (\boldsymbol{z}, \boldsymbol{t^1}, \mathbf{0})$, and:

$$\forall \boldsymbol{u}, \boldsymbol{v}^1 \in \mathbb{R}^p \times \mathbb{R}^q, \tilde{f}\Big|_{\mathcal{D}^1}(\boldsymbol{u}, \boldsymbol{v}^1, \mathbf{0}) = \left(f\big|_{\mathcal{D}^B} \circ h_1^{-1}\right)(\boldsymbol{u}, \boldsymbol{v}^1, \mathbf{0}). \tag{33}$$

Applying the same argument to data set 2, we conclude that that there exists an affine transformation $h_2 : \mathcal{D}^2 \to \mathcal{D}^2$ such that $h_2(\boldsymbol{z}, \mathbf{0}, \boldsymbol{t^2}) \overset{d}{=} (\boldsymbol{z}, \mathbf{0}, \boldsymbol{t^2})$, and:

$$\forall \boldsymbol{u}, \boldsymbol{v}^2 \in \mathbb{R}^p \times \mathbb{R}^q, \tilde{f}\Big|_{\mathcal{D}^1}(\boldsymbol{u}, \mathbf{0}, \boldsymbol{v}^2) = \left(f\big|_{\mathcal{D}^B} \circ h_2^{-1}\right)(\boldsymbol{u}, \mathbf{0}, \boldsymbol{v}^2). \tag{34}$$

Finally, we make appeal to a short linear algebraic argument to conclude. Because h_1 and h_2 are affine maps preserving the isotropic Gaussian distribution, they must be linear in order to preserve the mean. We may rewrite the definitions of the functions as:

$$h_1(\boldsymbol{z}, \boldsymbol{v}^1) = \left(U_{11}\boldsymbol{z} + U_{12}\boldsymbol{v}^1, U_{21}\boldsymbol{z} + U_{22}\boldsymbol{v}^1\right) \tag{35}$$

$$h_2(\boldsymbol{z}, \boldsymbol{v}^2) = \left(V_{11}\boldsymbol{z} + V_{12}\boldsymbol{v}^2, V_{21}\boldsymbol{z} + V_{22}\boldsymbol{v}^2\right). \tag{36}$$

Due to the identifiability of the functions, we know that h_1 and h_2 overlap for all values of $\boldsymbol{z}$ when $\boldsymbol{v}^1 = \boldsymbol{v}^2 = \mathbf{0}$. Therefore, we have that $U_{11} = V_{11} = R_1$, and $U_{21} = V_{21} = 0$. Therefore, we have

$$h_1(\boldsymbol{z}, \boldsymbol{v}^1) = \left(R_1\boldsymbol{z} + U_{12}\boldsymbol{v}^1, U_{22}\boldsymbol{v}^1\right) \tag{37}$$

$$h_2(\boldsymbol{z}, \boldsymbol{v}^2) = \left(R_1\boldsymbol{z} + V_{12}\boldsymbol{v}^2, V_{22}\boldsymbol{v}^2\right). \tag{38}$$

Now, in order to preserve the covariance of the Gaussian vector $h_1(\boldsymbol{z}, \boldsymbol{t}^1)$, we have

$$\begin{pmatrix} I_p & 0 \\ 0 & I_q \end{pmatrix} = \begin{pmatrix} R_1 & U_{12} \\ 0 & U_{22} \end{pmatrix} \begin{pmatrix} R_1^\top & 0 \\ U_{12}^\top & U_{22}^\top \end{pmatrix} = \begin{pmatrix} R_1R_1^\top + U_{12}U_{12}^\top & U_{12}U_{22}^\top \\ U_{22}U_{12}^\top & U_{22}U_{22}^\top \end{pmatrix}. \tag{39}$$

Observing that U_{22} is an isometry, as in the contrastive case, we obtain that $U_{12} = 0$. Proceeding similarly for $h_2(\boldsymbol{x}, \boldsymbol{t}^1)$, we have that $V_{12} = 0$ as well. Therefore, we have proved that the restrictions of $v = f^{-1} \circ \tilde{f}$ to $\mathcal{D}^1$ and $\mathcal{D}^2$ are compatible with respected to the Euclidean decomposition $\mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}^q$. □

Appendix D. Identifiability Theory for Counting Observation Noise

D.1. Preliminary Results

The central argument for our proofs will rely on a generalization of the characteristic function that we refer to as the Laplace transform:

Definition 6 *Let $\boldsymbol{y}$ be a random vector valued in $\mathbb{R}^d$. The Laplace transform of the random vector $\boldsymbol{y}$ is defined as:*

$$\begin{aligned} \mathbb{C}^d &\to \mathbb{C} \\ \xi_{\boldsymbol{y}} : \boldsymbol{t} &\mapsto \mathbb{E}\left[e^{\boldsymbol{t}^\top \boldsymbol{y}}\right]. \end{aligned} \tag{40}$$

The restriction of $\xi_{\boldsymbol{y}}$ to the product of the imaginary lines is the characteristic function $\phi_{\boldsymbol{y}}$. Similarly, the restriction of $\xi_{\boldsymbol{y}}$ to the product of the real lines is the moment generating function.

When dealing with count distributions, it is especially interesting to consider specific results about positive random variables. In this case, the Laplace transform is defined, and even holomorphic, over the product of half-spaces:

Lemma 7 *Let $\boldsymbol{y}$ be a random vector valued in $\mathbb{R}_+^d$. Then, the function $\xi_{\boldsymbol{y}}$ is a holomorphic function of several variables on $\mathcal{H}_-^d$, where $\mathcal{H}_- = \{z \in \mathbb{C} \mid \Re(z) < 0\}$.*

Proof We first show that the function is well defined on its domain $\mathcal{H}_-^d$. Let $\boldsymbol{t} = (u_j + iv_j)_{j=1}^d$ with $u_j < 0$ for all $j \in [d]$. In this case, the integral is absolutely convergent:

$$\mathbb{E}\left|e^{\boldsymbol{t}^\top \boldsymbol{y}}\right| = \mathbb{E}\left|e^{\boldsymbol{u}^\top \boldsymbol{y} + i\boldsymbol{v}^\top \boldsymbol{y}}\right| = \mathbb{E}e^{\boldsymbol{u}^\top \boldsymbol{y}} \leq 1 \tag{41}$$

Because $\boldsymbol{y}$ is positive almost surely, and $\boldsymbol{u}$ is a vector with negative values. To show that $\xi_{\boldsymbol{y}}$ is holomorphic on $\mathcal{H}_-^d$, by Osgood's lemma, it is sufficient to prove that $\xi_{\boldsymbol{y}}$ is continuous on $\mathcal{H}_-^d$, and holomorphic in each of its variables on $\mathcal{H}_-$. To show that $\xi_{\boldsymbol{y}}$ is continuous, it is enough to notice that the integrand defining $\xi_{\boldsymbol{y}}$ is uniformly bounded by an integrable function (Equation 41), and that the function $\boldsymbol{t} \mapsto e^{\boldsymbol{t}^\top \boldsymbol{y}}$ is continuous, so continuity follows from the dominated convergence theorem. To show that $\xi_{\boldsymbol{y}}$ is holomorphic on each of its variable, it is enough to show that $t_1 \mapsto \xi_{\boldsymbol{y}}(t_1, \ldots, t_d)$ is holomorphic on $\mathcal{H}_-$. Here again, we may apply the dominated convergence theorem. This is justified because (i) the exponential function is holomorphic, (ii) for all $t_1 \in \mathcal{H}^-$ the integral exists and (iii) the integrand is uniformly bounded above by an absolutely integrable function. □

Next, we derive the Laplace transform for some specific compound variables.

Lemma 8 (Poisson noise) *Let $\boldsymbol{y}$ be a random vector valued in $\mathbb{R}_+^d$, and let us assume that we observe count data $\boldsymbol{x}$ generated as $x_j \sim$ Poisson(y_j) for $j \in [d]$. Then, for all $\boldsymbol{t} \in \mathbb{C}^d$ such that the Laplace transform of the random vector $\boldsymbol{x}$ is defined, it can be derived as:*

$$\xi_{\boldsymbol{x}}(\boldsymbol{t}) = \xi_{\boldsymbol{y}}\left(e^{\boldsymbol{t}} - \mathbf{1}\right), \tag{42}$$

where the component-wise operations are used to assemble the vector $e^{\boldsymbol{t}} - \mathbf{1} = \left(e^{t_j} - 1\right)_{j=1}^d$.

Proof This derivation simply uses the law of total expectations,

$$\xi_{\boldsymbol{x}}(\boldsymbol{t}) = \mathbb{E}_{\boldsymbol{x}}[e^{\boldsymbol{t}^\top \boldsymbol{x}}] = \mathbb{E}_{\boldsymbol{y}}\left[\mathbb{E}_{\boldsymbol{x}}[e^{\boldsymbol{t}^\top \boldsymbol{x}} \mid \boldsymbol{y}]\right], \tag{43}$$

the fact that components of $\boldsymbol{x}$ are independent conditionally on $\boldsymbol{y}$,

$$\xi_{\boldsymbol{x}}(\boldsymbol{t}) = \mathbb{E}_{\boldsymbol{y}}\left[\mathbb{E}_{\boldsymbol{x}}\left[\prod_{j=1}^d e^{t_j x_j} \mid \boldsymbol{y}\right]\right] = \mathbb{E}_{\boldsymbol{y}}\left[\prod_{j=1}^d \mathbb{E}_{x_j}\left[e^{t_j x_j} \mid \boldsymbol{y}\right]\right], \tag{44}$$

and the definition of the Laplace transform for the Poisson distribution

$$\xi_{\boldsymbol{x}}(\boldsymbol{t}) = \mathbb{E}_{\boldsymbol{y}}\left[\prod_{j=1}^d e^{y_j(e^{t_j}-1)}\right] = \xi_{\boldsymbol{y}}\left(e^{\boldsymbol{t}} - \mathbf{1}\right), \tag{45}$$

where component-wise operations are used to assemble the vector $e^{\boldsymbol{t}} - \mathbf{1} = \left(e^{t_j} - 1\right)_{j=1}^d$. □

Lemma 9 (Negative binomial noise) *Let $\boldsymbol{y}$ be a random vector valued in $\mathbb{R}_+^d$, and let us assume that we observe count data $\boldsymbol{x}$ generated as $x_j \sim$ NegativeBinomial(y_j, θ) for $j \in [d]$, where θ*

designates the shape parameter. Then, for all $\boldsymbol{t} \in \mathbb{C}^d$ *such that the Laplace transform of the random vector* $\boldsymbol{x}$ *is defined, it can be derived as:*

$$\xi_{\boldsymbol{x}}(\boldsymbol{t}) = \xi_{\boldsymbol{y}}\left(-\log(1-(e^{\boldsymbol{t}}-1)\theta)\right), \tag{46}$$

where $\log$ *is the principal branch of the complex logarithm, and component-wise operations are used to assemble the vector* $\log(1-(e^{\boldsymbol{t}}-1)\theta) = \left(\log(1-(e^{t_j}-1)\theta)\right)_{j=1}^{d}$.

Proof In this case, we use the definition of the negative binomial distribution as a Gamma-Poisson compound distribution:

$$x_j \sim \text{NegativeBinomial}(y_i, \theta) \Leftrightarrow u_i \sim \text{Gamma}(y_i, \theta), x_i \sim \text{Poisson}(u_i). \tag{47}$$

Following the Poisson case, we get

$$\xi_{\boldsymbol{x}}(\boldsymbol{t}) = \mathbb{E}_{\boldsymbol{y}}\left[\mathbb{E}_{\boldsymbol{u}}\left[\mathbb{E}_{\boldsymbol{x}}\left[e^{\boldsymbol{t}^\top \boldsymbol{x}} \mid \boldsymbol{u}\right] \mid \boldsymbol{y}\right]\right] \tag{48}$$

$$= \mathbb{E}_{\boldsymbol{y}}\left[\mathbb{E}_{\boldsymbol{u}}\left[\mathbb{E}_{\boldsymbol{x}}\left[\prod_{j=1}^{d} e^{t_j x_j} \mid \boldsymbol{u}\right] \mid \boldsymbol{y}\right]\right] \tag{49}$$

$$= \mathbb{E}_{\boldsymbol{y}}\left[\prod_{j=1}^{d} \mathbb{E}_{\boldsymbol{u}}\left[\mathbb{E}_{\boldsymbol{x}}\left[e^{t_j x_j} \mid \boldsymbol{u}\right] \mid \boldsymbol{y}\right]\right]. \tag{50}$$

Then, using the definition of the Laplace transform for the Poisson distribution and the Gamma distribution, we get

$$\xi_{\boldsymbol{x}}(\boldsymbol{t}) = \mathbb{E}_{\boldsymbol{y}}\left[\prod_{j=1}^{d} \mathbb{E}_{\boldsymbol{u}} e^{u_j(e^{t_j}-1)}\right] \tag{51}$$

$$= \mathbb{E}_{\boldsymbol{y}}\left[\prod_{j=1}^{d} \left[1-(e^{t_j}-1)\theta\right]^{-y_j}\right], \tag{52}$$

where the power with a complex base a^p is defined as $e^{p\log a}$, where $\log$ is the principal branch of the complex logarithm. Reassambling the product into a sum of terms inside the exponential distribution, we get

$$\xi_{\boldsymbol{x}}(\boldsymbol{t}) = \xi_{\boldsymbol{y}}\left(-\log(1-(e^{\boldsymbol{t}}-1)\theta)\right), \tag{53}$$

where the vector $\log(1-(e^{\boldsymbol{t}}-1)\theta) = \left(\log(1-(e^{t_j}-1)\theta)\right)_{j=1}^{d}$ is put together component-wise. □

D.2. Identifiability of mixture through observational count noise

Proposition 2 (Identifiability of mixture through observational count noise) *Let* $\boldsymbol{u}$ *be a random variable taking values in* $\mathbb{R}^p$. *Let* f *and* $\tilde{f}$ *be two functions valued in* $\mathbb{R}_+^d$, *and let* $\boldsymbol{y} = f(\boldsymbol{u})$ *and* $\tilde{\boldsymbol{y}} = \tilde{f}(\boldsymbol{u})$. *If the random variables* $\boldsymbol{x} \sim p_x(\boldsymbol{y})$ *and* $\tilde{\boldsymbol{x}} \sim p_x(\tilde{\boldsymbol{y}})$ *are equal in distributions, and* p_x *is Poisson, or negative binomial with fixed shape, then it follows that* $\boldsymbol{y} \overset{d}{=} \tilde{\boldsymbol{y}}$.

Proof Because both random vectors $\boldsymbol{y}$ and $\tilde{\boldsymbol{y}}$ are valued in $\mathbb{R}_+^d$, Lemma 7 dictates that $\xi_{\boldsymbol{y}}$ and $\xi_{\tilde{\boldsymbol{y}}}$ are holomorphic on $\mathcal{H}_-^d$. Proceeding similarly for $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$, $\xi_{\boldsymbol{x}}$ and $\xi_{\tilde{\boldsymbol{x}}}$ are holomorphic, and therefore well-defined on $\mathcal{H}_-^d$.

Under the assumption that $\boldsymbol{x} \overset{d}{=} \tilde{\boldsymbol{x}}$, we know that the characteristic functions of $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ must agree. Therefore, we have:

$$\forall \boldsymbol{t} \in (i\mathbb{R})^d, \xi_{\boldsymbol{x}}(\boldsymbol{t}) = \xi_{\tilde{\boldsymbol{x}}}(\boldsymbol{t}), \tag{54}$$

We now split the cases of Poisson and negative binomial distribution.

Poisson case Using Lemma 8, Equation 54 is equivalent to:

$$\forall \boldsymbol{t} \in (i\mathbb{R})^d, \xi_{\boldsymbol{y}}\left(e^{\boldsymbol{t}} - \mathbf{1}\right) = \xi_{\tilde{\boldsymbol{y}}}\left(e^{\boldsymbol{t}} - \mathbf{1}\right), \tag{55}$$

Therefore, the holomorphic functions $\xi_{\boldsymbol{y}}$ and $\xi_{\tilde{\boldsymbol{y}}}$ have the same value on a sequence of points $\boldsymbol{t}_n = t_n \cdot \mathbf{1}_d$ where $t_n = e^{i(\pi + \frac{1}{n})} - 1$, that converges to $-2 \cdot \mathbf{1}_d \in \mathcal{H}_-^d$.

Negative binomial case Using Lemma 9, Equation 54 is equivalent to:

$$\forall \boldsymbol{t} \in (i\mathbb{R})^d, \xi_{\boldsymbol{y}}\left(-\log(1 - (e^{\boldsymbol{t}} - 1)\theta)\right) = \xi_{\tilde{\boldsymbol{y}}}\left(-\log(1 - (e^{\boldsymbol{t}} - 1)\theta)\right), \tag{56}$$

Therefore, the holomorphic functions $\xi_{\boldsymbol{y}}$ and $\xi_{\tilde{\boldsymbol{y}}}$ have the same value on a sequence of points $\boldsymbol{t}_n = t_n \cdot \mathbf{1}_d$ where $t_n = -\log\left(1 - \theta(e^{i(\pi + \frac{1}{n})} - 1)\right)$, that converges to $-\log(1 + 2\theta) \cdot \mathbf{1}_d \in \mathcal{H}_-^d$.

End of proof In both cases, by analytic continuation, we have:

$$\forall \boldsymbol{t} \in \mathcal{H}_-^d, \xi_{\boldsymbol{y}}(\boldsymbol{t}) = \xi_{\tilde{\boldsymbol{y}}}(\boldsymbol{t}). \tag{57}$$

Now, let us notice that $\xi_{\boldsymbol{y}}(\boldsymbol{t})$ and $\xi_{\tilde{\boldsymbol{y}}}(\boldsymbol{t})$ are holomorphic, and therefore continuous on $\mathcal{H}_-^d$. For each $\boldsymbol{w} \in \mathbb{R}^d$, we denote as $\boldsymbol{t}_n^{\boldsymbol{w}} = -\frac{1}{n} \cdot \mathbf{1}_d + i\boldsymbol{w}$. We know that both functions $\xi_{\boldsymbol{y}}$ and $\xi_{\tilde{\boldsymbol{y}}}$ admit for limit the value of their respective characteristic function evaluated at $\boldsymbol{w}$ when $n \to \infty$, and then by continuous extension of the function, we must have equality of the limits, and therefore of the characteristic functions. Therefore, this is enough to guarantee $\boldsymbol{y} \overset{d}{=} \tilde{\boldsymbol{y}}$, since characteristic functions uniquely characterize a probability distribution. □

We now proceed to the proof of the theorem.

Theorem 2 (Reduction from observational count noise to the noiseless setting) *Let $\boldsymbol{u} \sim$ Normal $(0, I_p)$. Let $f = \sigma \circ g$ (resp. $\tilde{f} = \sigma \circ \tilde{g}$) be the composition of a scalar link function σ, valued in $\mathbb{R}_+$ (applied component-wise), with a piecewise affine function g (resp. $\tilde{g}$). Let $\boldsymbol{x} \sim p_x(f(\boldsymbol{u}))$ and $\tilde{\boldsymbol{x}} \sim p_x(\tilde{f}(\boldsymbol{u}))$ such that p_x is Poisson or negative binomial with fixed shape. If σ is a bicontinuous bijection, then,*

$$\tilde{\boldsymbol{x}} \overset{d}{=} \boldsymbol{x} \implies \tilde{f}(\boldsymbol{u}) \overset{d}{=} f(\boldsymbol{u}) \implies \tilde{g}(\boldsymbol{u}) \overset{d}{=} g(\boldsymbol{u}). \tag{3}$$

Proof The implication $\tilde{\boldsymbol{x}} \overset{d}{=} \boldsymbol{x} \implies \tilde{f}(\boldsymbol{u}) \overset{d}{=} f(\boldsymbol{u})$ is the result of Proposition 2.

To prove the second implication, we need to notice that because $\sigma : \mathbb{R} \to \mathbb{R}$ is a bicontinuous bijection, it must be a monotonic function. Without loss of generality, we may assume it is strictly increasing. Therefore, its inverse function σ^{-1} is also strictly increasing, and by monotonicity, for every line segment $[a, b]$ of $\mathbb{R}$, we have:

$$\forall c \in \mathbb{R}, c \in [a, b] \Leftrightarrow \sigma^{-1}(c) \in [\sigma^{-1}(a), \sigma^{-1}(b)]. \tag{58}$$

Let us consider now a hyperbox $H = \prod_{i=1}^{d}[a_i, b_i]$, and $H' = \prod_{i=1}^{d}[\sigma^{-1}(a_i), \sigma^{-1}(b_i)]$ its image component-wise by σ. We have again, for every point $\boldsymbol{c} \in \mathbb{R}^d$:

$$\boldsymbol{c} \in H \Leftrightarrow \Psi(\boldsymbol{c}) \in H', \tag{59}$$

with Ψ is defined as

$$\Psi : \boldsymbol{y} \mapsto \left(\sigma^{-1}(y_1), \ldots, \sigma^{-1}(y_d)\right). \tag{60}$$

To conclude the proof, it is enough to notice that two random variables X and Y of $\mathbb{R}^d$ are equal in distributions if they have the same generalized cumulative distribution function, that is if for every hyperbox H, we have $p(X \in H) = p(Y \in H)$.

Indeed, if we assume that $\tilde{f}(\boldsymbol{u}) \overset{d}{=} f(\boldsymbol{u})$, we have that for every hyperbox H, $p(\tilde{f}(\boldsymbol{u}) \in H) = p(f(\boldsymbol{u}) \in H)$. However, we have that $\tilde{f}(\boldsymbol{u}) \in H$ if and only if $\tilde{g}(\boldsymbol{u}) \in H'$, and similarly, $f(\boldsymbol{u}) \in H$ if and only if $g(\boldsymbol{u}) \in H'$. Therefore, we have that for every image hyperbox H', $p(\tilde{f}(\boldsymbol{u}) \in H') = p(f(\boldsymbol{u}) \in H')$. Because σ is a bicontinuous bijection, the set of images hyperboxes covers the set of all hyperboxes (because of monotonicity), and therefore, we have that $\tilde{g}(\boldsymbol{u}) \overset{d}{=} g(\boldsymbol{u})$. □

D.3. Discussion of the Bernoulli Noise Setting

The intuition is that in the one-dimensional setting, a mixture of Bernoulli distributions is still a distribution on $\{0, 1\}$ and therefore entirely determined by its first moment. To show this, we proceed with calculations similar to the ones conducted for the Poisson and negative binomial distribution, but in the case of Bernoulli noise. Let $\boldsymbol{y}$ be a random vector valued in $[0, 1]^d$, and let us assume that we observe binary data $\boldsymbol{x}$ generated as $x_j \sim \text{Bernoulli}(y_j)$ for $j \in [d]$. For $\boldsymbol{t} \in \mathbb{C}^d$, the Laplace transform of the random vector $\boldsymbol{x}$ can be derived as:

$$\xi_{\boldsymbol{x}}(\boldsymbol{t}) = \mathbb{E}_{\boldsymbol{y}}\left[\prod_{j=1}^{d} \mathbb{E}_{x_j}\left[e^{t_j x_j} \mid \boldsymbol{y}\right]\right]. \tag{61}$$

Let us notice that because $\boldsymbol{y}$ is bounded, $\xi_{\boldsymbol{x}}$ is defined on all of $\mathbb{C}^d$. Then, considering the definition of the Laplace transform of a Bernoulli distribution, we obtain

$$\xi_{\boldsymbol{x}}(\boldsymbol{t}) = \mathbb{E}_{\boldsymbol{y}}\left[\prod_{j=1}^{d}\left((1 - y_j) + y_j e^{t_j}\right)\right], \tag{62}$$

and we may expand the product, by identifying the subset $S \subset \{1, \ldots, d\}$ with a binary vector of size d

$$\xi_{\boldsymbol{x}}(\boldsymbol{t}) = \mathbb{E}_{\boldsymbol{y}}\left[\sum_{S \subset \{1,\ldots,d\}} \prod_{j=1}^{d} (1 - y_j)^{(1-S_j)} y_j^{S_j} e^{t_j S_j}\right] \tag{63}$$

$$\xi_{\boldsymbol{x}}(\boldsymbol{t}) = \sum_{S \subset \{1,\ldots,d\}} \mathbb{E}_{\boldsymbol{y}}\left[\prod_{j=1}^{d} (1 - y_j)^{(1-S_j)} y_j^{S_j}\right] e^{\boldsymbol{t}^\top S}. \tag{64}$$

It is important to notice here that each subset S in the sum above induces a unique monomial term. Give this observation, if we assume $\xi_{\boldsymbol{x}} = \xi_{\tilde{\boldsymbol{x}}}$ for another data-generating process, the equality of those two functions implies the equality of two finite complex Fourier series, and therefore their coefficients must be equal (in fact, it is an equivalence):

$$\forall \boldsymbol{t} \in \mathbb{C}^d, \xi_{\boldsymbol{x}}(\boldsymbol{t}) = \xi_{\tilde{\boldsymbol{x}}}(\boldsymbol{t}). \tag{65}$$

$$\Leftrightarrow \forall S \in \{0,1\}^d, \mathbb{E}_{\boldsymbol{y}}\left[\prod_{j=1}^{d}(1-y_j)^{(1-S_j)}y_j^{S_j}\right] = \mathbb{E}_{\tilde{\boldsymbol{y}}}\left[\prod_{j=1}^{d}(1-y_j)^{(1-S_j)}y_j^{S_j}\right]. \tag{66}$$

Interestingly, the polynomial appearing in the last equation only has term with partial degree of at most 1 (but of total degree d). Therefore, we hypothesize that it must not be true that this fully characterizes the distribution of Y, because in general, there exist several distinct functions f that could have the same moment of order 1.

We therefore focus on building a counter-example with $p = d = 2$. In this case, let us assume $\boldsymbol{z} \sim \text{Normal}(0, I_2)$ and that $\boldsymbol{y} = f(\boldsymbol{z})$, and $\tilde{\boldsymbol{y}} = \tilde{f}(\boldsymbol{z})$, and $\boldsymbol{x}$ is generated as $x_j \sim \text{Bernoulli}(y_j)$ for $j \in [d]$. $\tilde{\boldsymbol{x}}$ is generated similarly, from $\tilde{\boldsymbol{y}}$. Let us also assume $\tilde{\boldsymbol{x}} \overset{d}{=} \boldsymbol{x}$. Based on Equation 66, we have that

$$\begin{cases} \mathbb{E}_{\boldsymbol{y}}\left[(1-y_1)(1-y_2)\right] &= \mathbb{E}_{\tilde{\boldsymbol{y}}}[(1-\tilde{y}_1)(1-\tilde{y}_2)] \\ \mathbb{E}_{\boldsymbol{y}}[y_1(1-y_2)] &= \mathbb{E}_{\tilde{\boldsymbol{y}}}[\tilde{y}_1(1-\tilde{y}_2)] \\ \mathbb{E}_{\boldsymbol{y}}[(1-y_1)y_2] &= \mathbb{E}_{\tilde{\boldsymbol{y}}}[(1-\tilde{y}_1)\tilde{y}_2] \\ \mathbb{E}_{\boldsymbol{y}}[y_1 y_2] &= \mathbb{E}_{\tilde{\boldsymbol{y}}}[\tilde{y}_1\tilde{y}_2] \end{cases} \tag{67}$$

These equations are equivalent to the constraints:

$$\begin{cases} \mathbb{E}_{\boldsymbol{y}}\boldsymbol{y} &= \mathbb{E}_{\tilde{\boldsymbol{y}}}\tilde{\boldsymbol{y}} \\ \mathbb{E}_{\boldsymbol{y}}y_1 y_2 &= \mathbb{E}_{\tilde{\boldsymbol{y}}}\tilde{y}_1\tilde{y}_2 \end{cases} \tag{68}$$

It is possible to find examples of functions that will yield distributions $\boldsymbol{y}$ and $\tilde{\boldsymbol{y}}$ that satisfy the constraints in Equation 68, but have different distributions. For example, let us consider the functions f_λ for $\lambda > 0$ of the form:

$$f_\lambda : (z_1, z_2) \mapsto \left(F_{\text{Beta}}(F^{-1}_{\text{Normal}}(z_1); \lambda, \lambda), F_{\text{Beta}}(F^{-1}_{\text{Normal}}(z_2); \lambda, \lambda)\right), \tag{69}$$

where F_{Normal} denotes the cumulative distribution function (CDF) of the isotropic Gaussian distribution, and $F_{\text{Beta}}(., \lambda, \lambda)$ denotes the CDF of the Beta distribution with parameters (λ, λ). For $\boldsymbol{y} = f_1(\boldsymbol{z})$ and $\tilde{\boldsymbol{y}} = f_2(\boldsymbol{z})$, we do not have equality in distribution for $\boldsymbol{y}$ and $\tilde{\boldsymbol{y}}$, but we do have that $\boldsymbol{x} \overset{d}{=} \tilde{\boldsymbol{x}}$.

To show that this pathological case can also happen within the framework we consider in this paper, let us consider the case $f = \sigma \circ g$ where g is a piecewise affine function. Furthermore, we consider the case where f is separable:

$$f(z_1, z_2) = (\sigma(g_{a,b}(z_1)), \sigma(g_{a,b}(z_2))), \tag{70}$$

where σ denotes the sigmoid function, and

$$g_{a,b}(w) = aw\delta\{w \leq 0\} + bw\delta\{w > 0\}. \tag{71}$$

Because y_1 and y_2 are independent in this case, equality in the mean of each component of $\boldsymbol{y}$ is enough to guarantee equality in distribution of $\boldsymbol{y}$. Let us examine the function:

$$\Phi : (a, b) \mapsto \mathbb{E}\left[\sigma(g_{a,b}(u))\right] = \int_{-\infty}^{0} \sigma(aw)p(w)dw + \int_{0}^{+\infty} \sigma(bw)p(w)dw, \tag{72}$$

where u is a random variable distributed according to a standard Gaussian distribution, and $p(w)$ denotes its density evaluated at w. By continuity under the integral sign, Φ is a continuous function on its domain.

We note that Φ cannot be injective, as it otherwise could be used to define an homeomorphism from $\mathbb{R}^2$ to $\mathbb{R}$, which is impossible by the invariance of domain theorem. Therefore, we must have two distinct parameters (a, b) and (a', b') such that the induced functions f and f' generate the same distribution. Because each set of parameters designates a unique function (if two $(a, b) \neq (a', b')$, then $f_{a,b} \neq f_{a',b'}$), this is a counterexample.

Appendix E. Identifiability under Misspecification of Contrastive DGMs

E.1. Block-wise Identifiability of Misspecified Linear Model

Proposition 1 (Block-wise identifiability under misspecification for the linear case) *Let the ground truth mixing function f be injective linear, and the learned function $\tilde{f}$ be a linear function such that $f(\boldsymbol{z}, \boldsymbol{s}) \overset{d}{=} \tilde{f}(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{s}})$ and $f(\boldsymbol{z}, \boldsymbol{0}) \overset{d}{=} \tilde{f}(\tilde{\boldsymbol{z}}, \boldsymbol{0})$. Then, there exist surjective linear functions $h_{\boldsymbol{z}}$ and $h_{\boldsymbol{s}}$ such that $(\boldsymbol{z}, \boldsymbol{s}) = v(\tilde{\boldsymbol{s}}, \tilde{\boldsymbol{z}}) = (v_{\boldsymbol{s}}(\tilde{\boldsymbol{s}}), v_{\boldsymbol{z}}(\tilde{\boldsymbol{z}}))$, where $v = f^{-1} \circ \tilde{f}$.*

Proof Let us assume that the data is generated according to the contrastive analysis model, and where $f : (\boldsymbol{z}, \boldsymbol{s}) \mapsto \boldsymbol{x} = U\boldsymbol{z} + V\boldsymbol{s}$ is a linear mixing function. Let $\tilde{f} : (\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{s}}) \mapsto \tilde{\boldsymbol{x}} = \tilde{U}\tilde{\boldsymbol{z}} + \tilde{V}\tilde{\boldsymbol{s}}$ denote the learned mixing function.

We first note that f is a linear function, and $\boldsymbol{z}, \boldsymbol{s}$ is a Gaussian vector. Similarly, $\tilde{f}$ is a linear function, and $\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{s}}$ are Gaussian vectors. Therefore, we have that $f(\boldsymbol{z}, \boldsymbol{s})$ and $\tilde{f}(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{s}})$ are Gaussian vectors. Two Gaussian vectors are equal in distributions if they have the same mean and covariance matrix. Because both are centered (with mean zero), we therefore rely on the equality of their covariance matrices. We may proceed similarly for the random vectors $f(\boldsymbol{z}, \boldsymbol{0})$ and $\tilde{f}(\tilde{\boldsymbol{z}}, \boldsymbol{0})$

Because all of the random vectors $\boldsymbol{z}, \boldsymbol{s}, \tilde{\boldsymbol{z}}, \tilde{\boldsymbol{s}}$ follow an isotropic Gaussian distribution, the equality of the covariance matrices entails that

$$\begin{cases} UU^\top & = & \tilde{U}\tilde{U}^\top \\ UU^\top + VV^\top & = & \tilde{U}\tilde{U}^\top + \tilde{V}\tilde{V}^\top \end{cases}, \tag{73}$$

equivalent to

$$\begin{cases} UU^\top & = & \tilde{U}\tilde{U}^\top \\ VV^\top & = & \tilde{V}\tilde{V}^\top \end{cases}. \tag{74}$$

Based on identifiability of the factor analysis model, (Shapiro, 1985), we know that there exists two matrices O_1 and O_2 with orthogonal rows such that

$$\begin{cases} \tilde{U} & = & UO_1 \\ \tilde{V} & = & VO_2 \end{cases} \tag{75}$$

and that we have $\text{rank}(U) = \text{rank}(\tilde{U}) = p$ and $\text{rank}(V) = \text{rank}(\tilde{V}) = q$.

We now seek to compute $v^\dagger = f^{-1} \circ \tilde{f}$. Let us notice that f is an injective linear function, and therefore $f^{-1}(\boldsymbol{x}) = (W_{\boldsymbol{z}}\boldsymbol{x}, W_{\boldsymbol{s}}\boldsymbol{x})$, where $W = [W_{\boldsymbol{z}}^\top, W_{\boldsymbol{s}}^\top]^\top$ is the pseudo-inverse of $[U, V]$. Substituting this into the expression for v, we have

$$v(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{s}}) = (W_{\boldsymbol{z}}UO_1\tilde{\boldsymbol{z}} + W_{\boldsymbol{z}}VO_2\tilde{\boldsymbol{s}}, W_{\boldsymbol{s}}UO_1\tilde{\boldsymbol{z}} + W_{\boldsymbol{s}}VO_2\tilde{\boldsymbol{s}})\,. \tag{76}$$

By definition of the pseudo-inverse, we have that $W_{\boldsymbol{z}}U = I$, $W_{\boldsymbol{s}}V = I$, $W_{\boldsymbol{z}}V = 0$, and $W_{\boldsymbol{s}}U = 0$. Therefore, we may derive a simpler expression for v

$$v(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{s}}) = (O_1\tilde{\boldsymbol{z}}, O_2\tilde{\boldsymbol{s}})\,. \tag{77}$$

Identifying each of the matrices O_1 and O_2 as linear maps, we obtain the desired statement. □

E.2. Non-identifiability in the Misspecified Non-linear Case

We begin by stating the counterexample. For $p = p' = 1$, $q = 1$ and $q' = 2$, let us denote $\tilde{z} = z$ and $\tilde{\boldsymbol{s}} = (s, v)$. We define f as the identity function, and $\tilde{f}$ as the following mixing function:

$$\tilde{f} : \begin{pmatrix} z \\ s \\ v \end{pmatrix} \mapsto \begin{pmatrix} z \cdot f(s) + v \cdot g(s) \\ s \end{pmatrix}, \tag{78}$$

with $f(s) = \mathbf{1}(s \geq 0)$ and $g(s) = \mathbf{1}(s < 0)$. Below we prove the result for the piecewise constant functions, but following the argument in proof for Example 1, the counter-example holds as long as $g(0) = 0$ and $f(\boldsymbol{s}) + g(\boldsymbol{s}) = 1$ almost surely.

Equality of the target data distributions Because f is the identity function, we simply need to show that $\tilde{f}(\tilde{z}, \tilde{\boldsymbol{s}})$ follows an isotropic Gaussian distribution. We define $u = z \cdot \mathbf{1}(s \geq 0) + v \cdot \mathbf{1}(s < 0)$ and seek to assess the density of u conditionally on s. In cases where $s \geq 0$, we get:

$$\begin{aligned} p(u \mid s, \{s \geq 0\}) &= p_z(u \mid s, \{s \geq 0\}) && (79)\\ &= p_z(u). && (80) \end{aligned}$$

We proceed similarly for $s < 0$

$$\begin{aligned} p(u \mid s, \{s < 0\}) &= p_v(u \mid s, \{s < 0\}) && (81)\\ &= p_v(u). && (82) \end{aligned}$$

Therefore, we have that

$$p(u \mid s) = \frac{p_v(u)}{2} + \frac{p_z(u)}{2}. \tag{83}$$

Because $p_v(u) = p_z(u)$ are both the density of the isotropic Gaussian distribution, we conclude that u is independent from s and that u follows an isotropic Gaussian distribution.

Equality of the background data distributions Noticing that $\tilde{f}(z, 0, 0) = (z, 0)$, we get that $\tilde{f}(\tilde{z}, \mathbf{0}) \overset{d}{=} f(z, 0)$.

Entanglement If we now define $(\bar{z}, \bar{s}) = v(\tilde{z}, \tilde{\boldsymbol{s}}) = \tilde{f}(\tilde{z}, \tilde{\boldsymbol{s}})$, we notice that $\bar{z}$ depends non-trivially on $\tilde{\boldsymbol{s}}$ (via v):

$$\bar{z} = z \cdot \mathbf{1}(s \geq 0) + v \cdot \mathbf{1}(s < 0), \tag{84}$$

and therefore we have entanglement.

E.3. Review of Existing Regularization Methods for Comparative Analysis Models

We remind the reader of the definition of the aggregated posteriors, defined for each data set,

$$\hat{q}_\phi^t(\boldsymbol{z}, \boldsymbol{s}) = \mathbb{E}_{p_{\text{data}}(\boldsymbol{x})} \left[q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) q_\phi(\boldsymbol{s} \mid \boldsymbol{x}) \right] \tag{85}$$

$$\hat{q}_\phi^b(\boldsymbol{z}, \boldsymbol{s}) = \mathbb{E}_{p_{\text{data}}(\boldsymbol{x} \mid \text{do}(\boldsymbol{s}=\mathbf{0}))} \left[q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) q_\phi(\boldsymbol{s} \mid \boldsymbol{x}) \right]. \tag{86}$$

Additionally, each of those can be used to define marginal distributions over each set of latent variables,

$$\hat{q}_\phi^t(\boldsymbol{z}) = \mathbb{E}_{p_{\text{data}}(\boldsymbol{x})} \left[q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \right] \tag{87}$$

$$\hat{q}_\phi^t(\boldsymbol{s}) = \mathbb{E}_{p_{\text{data}}(\boldsymbol{x})} \left[q_\phi(\boldsymbol{s} \mid \boldsymbol{x}) \right] \tag{88}$$

$$\hat{q}_\phi^b(\boldsymbol{z}) = \mathbb{E}_{p_{\text{data}}(\boldsymbol{x} \mid \text{do}(\boldsymbol{s}=\mathbf{0}))} \left[q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \right]. \tag{89}$$

There are two prominent techniques for promoting disentanglement in the latent space. Abid and Zou (2019) initially proposed to enforce an independence constraint of the form $\hat{q}_\phi^t(\boldsymbol{z}) \perp\!\!\!\perp \hat{q}_\phi^t(\boldsymbol{s})$. Such constraint is enforced using an adversarial classifier that tries to distinguish samples from the joint $\hat{q}_\phi^t(\boldsymbol{z}, \boldsymbol{s})$ from samples from the marginals $\hat{q}_\phi^t(\boldsymbol{z}), \hat{q}_\phi^t(\boldsymbol{s})$ (obtained through random shuffling of the embeddings).

Later, Weinberger et al. (2022a) proposed to use a combination of two constraints to help with disentanglement. The first constraint $\hat{q}_\phi^t(\boldsymbol{z}) = \hat{q}_\phi^b(\boldsymbol{z})$ ensures that the distribution of z is identical between the target and the background data set. The second constraint consists in observing that the variational distribution of $\boldsymbol{s}$ for data points in the background data set is not defined, but in principle could be assessed using the amortization network parameterizing $q_\phi(\boldsymbol{s} \mid \boldsymbol{x})$:

$$\hat{q}_\phi^b(\boldsymbol{s}) = \mathbb{E}_{p_{\text{data}}(\boldsymbol{x} \mid \text{do}(\boldsymbol{s}=\mathbf{0}))} \left[q_\phi(\boldsymbol{s} \mid \boldsymbol{x}) \right]. \tag{90}$$

Then, the second constraint is $\hat{q}_\phi^b(\boldsymbol{s}) = \delta_{\mathbf{0}}$, where $\delta_{\mathbf{0}}$ denotes the Dirac distribution centered at zero. In practice, those constraints may be enforced using a maximum mean discrepancy penalty (Gretton et al., 2012) as used in Weinberger et al. (2022a). Another option for enforcing the constraint $\hat{q}_\phi^b(\boldsymbol{s}) = \delta_{\mathbf{0}}$ is to penalize with the Wasserstein-2 distance. This second penalization is available in closed-form between a Dirac and a Gaussian distribution (Villani, 2008), as introduced in Weinberger et al. (2022b) and used in Weinberger et al. (2023).

Appendix F. Experiments

F.1. Multi-Objective Optimization: the case of Two Objectives

We recall the formulation of the multi-objective optimization problem

$$\min_{\theta, \phi} \left(-\mathcal{L}^B(\theta, \phi), -\mathcal{L}^T(\theta, \phi) \right). \tag{91}$$

The first step in the Multiple-Gradient Descent Algorithm (Désidéri, 2012) consists in calculating the convex combination of the gradients from each loss used for the update of the parameters. In the case of two objective functions, it is defined as:

$$\alpha^*(\theta,\phi) = \arg\min_{\alpha\in[0,1]} \|\alpha\nabla\mathcal{L}^B(\theta,\phi) + (1-\alpha)\nabla\mathcal{L}^T(\theta,\phi)\|_2^2. \tag{92}$$

As pointed out by Désidéri (2012), this optimization problem in Equation 92 admits a closed-form solution, defined as:

$$\alpha^*(\theta,\phi) = \left[\frac{\left(\nabla\mathcal{L}^T(\theta,\phi) - \nabla\mathcal{L}^B(\theta,\phi)\right)^\top \nabla\mathcal{L}^T(\theta,\phi)}{\|\nabla\mathcal{L}^T(\theta,\phi) - \nabla\mathcal{L}^B(\theta,\phi)\|_2^2}\right]_{[0,1]}^{\tau}, \tag{93}$$

where $[a]_{[0,1]}^{\tau} = \max(0, \min(a, 1))$ designates the projection onto the compact $[0, 1]$. Taken together, the optimization procedure can be described as:

$$\alpha_t = \alpha^*(\theta^t, \phi^t) \tag{94}$$

$$\delta_t = -\alpha_t\nabla\mathcal{L}^B(\theta^t,\phi^t) - (1-\alpha_t)\nabla\mathcal{L}^T(\theta^t,\phi^t) \tag{95}$$

$$\left[\theta^{t+1}, \phi^{t+1}\right] = \left[\theta^t, \phi^t\right] - \eta\delta_t, \tag{96}$$

where we note that the gradient update in Equation 96 may be replaced with that of any first-order stochastic optimizer.

The original framework suggests that one should use the gradient with respect to all the shared parameters of the model in order to compute the alpha parameter in Equation 93. This requires collecting the gradients with respect to parameters of every layer of the decoder and the encoder. In the simulations with d_x = 150, this amounts to around 180K parameters. For ease of implementation, we calculated α using only the gradients with respect to the weights of the last layer of the decoder (around 20K parameters), and noticed improvements over the single-objective method.

F.2. Constrained Optimization

We aim at solving the following constrained optimization problem:

$$\min_{\theta,\phi} \mathcal{L}(\theta,\phi) = -\mathcal{L}^B(\theta,\phi) - \mathcal{L}^T(\theta,\phi) \;\text{ such that }\; \frac{\|C_{\boldsymbol{z},\boldsymbol{s}}\|_{\text{HS}}^2}{\|C_{\boldsymbol{z},\boldsymbol{z}}\|_{\text{HS}}\|C_{\boldsymbol{s},\boldsymbol{s}}\|_{\text{HS}}} \leq \beta. \tag{97}$$

By using an alternative constraint formulation for the ratio, the problem above is equivalent to:

$$\min_{\theta,\phi} \mathcal{L}(\theta,\phi) \;\text{ such that }\; \|C_{\boldsymbol{z},\boldsymbol{s}}\|_{\text{HS}}^2 \leq \beta\|C_{\boldsymbol{z},\boldsymbol{z}}\|_{\text{HS}}\|C_{\boldsymbol{s},\boldsymbol{s}}\|_{\text{HS}}. \tag{98}$$

Finally, utilizing the technique of Lagrange multiplier, we obtain the equivalent problem:

$$\min_{\theta,\phi}\max_{\lambda\geq 0} \mathcal{L}^\lambda(\theta,\phi) = \mathcal{L}(\theta,\phi) + \lambda\left(\|C_{\boldsymbol{z},\boldsymbol{s}}\|_{\text{HS}}^2 - \beta\|C_{\boldsymbol{z},\boldsymbol{z}}\|_{\text{HS}}\|C_{\boldsymbol{s},\boldsymbol{s}}\|_{\text{HS}}\right). \tag{99}$$

To see why this last formulation is equivalent, it is enough to see that when the constraint is not satisfied (i.e., the difference is positive), then the optimal value of the inner optimization problem is $+\infty$ (for $\lambda \to \infty$), prohibiting those values for the parameters θ, ϕ to be picked by the outer optimization problem. However, the problem is unchanged when the constraint is satified, since the optimal value of the inner optimization problem is $\mathcal{L}(\theta, \phi)$ (for $\lambda \to 0$. Similar derivations appear in Gallego-Posada et al. (2022).

F.3. Neural Network Architectures and Implementation Details

Two separate encoder neural networks were used to parameterize our approximate posterior distributions $q_\phi(\boldsymbol{s} \mid \boldsymbol{x})$ and $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$. Each network first has a single hidden layer consisting of 128 hidden units, followed by a batch normalization layer, a rectified linear unit (ReLU) activation function, and a dropout layer. Then, the output units were used as inputs to two linear layers, one parameterizing the mean and one parameterizing the log-variance of the variational distribution.

The decoder network first has two hidden layers with 128 hidden units, taking as input the concatenation of the latent variables $[\boldsymbol{z}, \boldsymbol{s}]$, and followed by batch normalization, a ReLU activation function, and a dropout layer. Then, the outputs units are fed to a linear layer with output size equal to the data dimension, with a softplus activation function.

All models were implemented using PyTorch (Paszke et al., 2019) with the scvi-tools framework (Gayoso et al., 2022). All models were trained for 500 epochs using Adam (Kingma and Ba, 2015) with a learning rate of 0.001, using the validation ELBO as an early stopping criterion.

F.4. Simulation Details

We simulate data as follows. We assume we have background measurements (resp. target measurements) from N_b samples (resp. N_t samples). We use $N_t = N_b = 1,500$ throughout the manuscript.

Target data set For sample n, background latent variable $\boldsymbol{z}_n$ is generated as:

$$\boldsymbol{z}_n \sim \text{Normal}(0, I_p), \tag{100}$$

where p is the dimension of the background space. Salient latent variable $\boldsymbol{s}_n$ is generated as:

$$\boldsymbol{s}_n \sim \text{Normal}(0, I_q), \tag{101}$$

where q is the dimension of the salient space. Measurements x_{ng} for sample n and feature $g \in \{1, \ldots, G\}$ are generated from a count distribution:

$$x_{ng} \sim \text{NegativeBinomial}\left(f^g(\boldsymbol{z}_n, \boldsymbol{s}_n), \theta_g\right), \tag{102}$$

where θ_g is the overdispersion parameter of the negative binomial. We use $G = 150$ throughout the manuscript. When the manuscript mentions Poisson noise, it means that we replace the conditional distribution above by a Poisson distribution, with mean equal to the output of the neural network f. The ground truth mixing function f is a neural network with four hidden layers of 40 units, Leaky-ReLU activations with a negative slope of 0.2, and a softmax non-linearity on the last layer to convert the outputs to counts (Lopez et al., 2018a). The weight matrices of f are sampled according to an isotropic Gaussian distribution, with orthogonal columns, to make sure f is injective (Lachapelle et al., 2022).

Background data set Proceeding similarly as above, for sample n, background latent variable $\boldsymbol{z}_n$ is generated as:

$$\boldsymbol{z}_n \sim \text{Normal}(0, I_p), \tag{103}$$

where p is the dimension of the background space. Then, measurements x_{ng} for sample n and feature g are generated from a count distribution:

$$x_{ng} \sim \text{NegativeBinomial}\left(f^g(\boldsymbol{z}_n, \boldsymbol{0}), \theta_g\right). \tag{104}$$

F.5. Evaluation Metrics

Linear Mean Correlation Coefficient (MCC) The Linear Mean Correlation Coefficient (MCC) serves as a metric to evaluate the degree of alignment between inferred and ground-truth latent factors in representation learning, particularly in the context of disentanglement (Khemakhem et al., 2020). Specifically, we utilize the mean of the variational posterior $q_\phi(\boldsymbol{z}, \boldsymbol{s} \mid \boldsymbol{x})$ as an approximation to $\hat{\boldsymbol{z}}, \hat{\boldsymbol{s}} = \tilde{f}^{-1}(\boldsymbol{x})$. To assess block-wise linear disentanglement, several linear regressions are executed. Initially, to confirm the informativeness of the latent spaces, we:

- Predict the ground-truth latent factors $\boldsymbol{s}$ using the inferred factors $\hat{\boldsymbol{s}}$ ($\text{MCC}_{\hat{\boldsymbol{s}}\boldsymbol{s}}$).
- Predict the ground-truth latent factors $\boldsymbol{z}$ using the inferred factors $\hat{\boldsymbol{z}}$ ($\text{MCC}_{\hat{\boldsymbol{z}}\boldsymbol{z}}$).

Subsequently, to ensure the absence of undesired overlaps, we:

- Predict the ground-truth latent factors $\boldsymbol{s}$ using the inferred factors $\hat{\boldsymbol{z}}$ ($\text{MCC}_{\hat{\boldsymbol{z}}\boldsymbol{s}}$).
- Predict the ground-truth latent factors $\boldsymbol{z}$ using the inferred factors $\hat{\boldsymbol{s}}$ ($\text{MCC}_{\hat{\boldsymbol{s}}\boldsymbol{z}}$).

For each set of predicted latent factors, the Pearson (and equivalently, Spearman) linear MCC is computed as the mean Pearson (or Spearman) Correlation Coefficient between predictions and the ground truth, evaluated component-wise. The MCC value ranges between -1 and 1, with values closer to 1 indicating a strong positive linear relationship, values closer to -1 indicating a strong negative linear relationship, and values around 0 indicating no linear relationship.

Average Silhouette Width (AWS) This metric assumes at disposal an Euclidean space where each data point n is associated with an embedding vector $t_n \in \mathbb{R}^d$ where n is a data point. Additionally, the AWS requires the definition of cluster assignments $y_n \in \{1, \ldots, K\}$, where K is the total number of clusters. For each data point n, the we define silhouette score SS_n of sample n as

$$\text{SS}_n = \frac{b_n - a_n}{\max(a_n, b_n)}, \tag{105}$$

where a_n is the average distance between data point n and all of other points with the same cluster label, and b_n is the average distance between n and all the points in the next nearest cluster. Then, for a data set with N samples, the AWS is defined as:

$$\text{ASW} = \frac{1}{N} \sum_{n=1}^{N} \text{SS}_n. \tag{106}$$

The value of ASW lies between -1 and 1, where a higher value indicates a better ability to distinguish the clusters in the embedding space.

Adjusted Rand Index (ARI) The Adjusted Rand Index (ARI) is a metric used to measure the similarity between two data clusterings. Consider two sets of cluster assignments: the true assignments $y_n \in \{1, \ldots, K\}$ and the predicted assignments $y'_n \in \{1, \ldots, K'\}$, where n is a data point, K is the total number of true clusters, and K' is the total number of predicted clusters. The ARI takes into

account the number of pairings of data points that are in the same or different clusters for both the true and predicted assignments. Specifically, the ARI is defined as:

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{N}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{N}{2}}, \tag{107}$$

where n_{ij} is the number of data points that are both in cluster i of the true assignments and cluster j of the predicted assignments. a_i and b_j are the total number of data points in cluster i of the true assignments and cluster j of the predicted assignments, respectively. N is the total number of data points. The value of ARI lies between -1 and 1, where a higher value indicates better clustering performance.

Normalized Mutual Information (NMI) The Normalized Mutual Information (NMI) is a metric designed to gauge the similarity between two data clusterings. Given two sets of cluster assignments: the true assignments $y_n \in \{1, \ldots, K\}$ and the predicted assignments $y'_n \in \{1, \ldots, K'\}$, where n is a data point, K is the total number of true clusters, and K' is the number of predicted clusters, the NMI is computed using the mutual information (MI) between the two assignments and their respective entropies:

$$\text{NMI}(y, y') = \frac{2 \times \text{MI}(y, y')}{H(y) + H(y')}. \tag{108}$$

Mutual information between the true and predicted assignments is given by:

$$\text{MI}(y, y') = \sum_{i=1}^{K}\sum_{j=1}^{K'} p(i,j)\log\left(\frac{p(i,j)}{p(i)p'(j)}\right), \tag{109}$$

where $p(i, j)$ is the joint probability of a data point belonging to cluster i in the true assignments and cluster j in the predicted assignments. $p(i)$ and $p'(j)$ are the probabilities of a data point belonging to cluster i and j in the true and predicted assignments, respectively.

The entropies of the true and predicted assignments are:

$$H(y) = -\sum_{i=1}^{K} p(i)\log(p(i)), \tag{110}$$

$$H(y') = -\sum_{j=1}^{K'} p'(j)\log(p'(j)). \tag{111}$$

The value of NMI lies in the range [0, 1], with a score of 1 suggesting that the two sets of cluster assignments are identical, while a score of 0 indicates no shared information between them.

F.6. Baseline Models

All baselines share the same architecture and implementations for the sake of comparison. All are modifications of the code from the ContrastiveVI package (Weinberger et al., 2023). Below we provide additional details about how each baseline was utilized.

VAE Because a VAE does not have two latent spaces, in all experiments, we double the number of latent variables to match the one from our contrastive analysis models. Then, we assign each latent variable to either the background or the salient space. Our assignment method consists in applying cPCA (Abid et al., 2018) to the learned representations from the VAE to split the latent space into a background and a salient space. As a first attempt, we used the loadings from the top cPCA eigenvalues to project the latent space into a salient space, and then used the loadings from the top PCA eigenvalue of the background data set to project the latent space into a background space. This approach had extremely poor performance (δ-MCC close to zero in most experiments, because of high cross MCC). Therefore, we proceeded as follows: We use the loadings onto the first contrastive principal component to obtain the list of latent units that contribute most to explaining the differences between the target and the background data set (using the absolute value of the eigenvector). We assign variables with the highest score to the salient space, and the ones with the lowest score to the background space.

ContrastiveVI Our ContrastiveVI implementation consists of exactly the SO-U-cVAE method with the addition of the Wasserstein penalty to the ELBO. More precisely, if $\mu_{\boldsymbol{s}}(\boldsymbol{x})$ and $\sigma_{\boldsymbol{s}}(\boldsymbol{x})$ encode the mean and standard deviation parameter of $q_\phi(\boldsymbol{s} \mid \boldsymbol{x})$, respectively, then the Wasserstein penalty $\mathcal{L}_{\mathcal{W}}$ is derived as

$$\mathcal{L}_{\mathcal{W}} = \|\mu_{\boldsymbol{s}}(\boldsymbol{x})\|_2^2 + \|\sigma_{\boldsymbol{s}}(\boldsymbol{x})\|_2^2, \tag{112}$$

with a fixed hyperparameter for the regularization strength (i.e., multiplier equal to one).

CausalDiscrepancyVAE CausalDiscrepancyVAE (Zhang et al., 2023) is generative model with a latent causal graph (DAG), a polynomial mixing function, and an interventional model for single and double-node (soft) interventions. Because Zhang et al. (2023) also apply their method to the data set from Norman et al. (2019), we discuss this work here. We note that those two models have distinct purposes. Contrastive Analysis methods aim at learning informative representation of the perturbations by removing signal from the heterogeneity of the control population. However, CausalDiscrepancyVAE aims at learning a DAG in latent space. For that, CausalDiscrepancyVAE requires the observation of many interventional regimes, as well as knowledge of the targets per intervention (ignored during our benchmark). For this reason, the models are not entirely comparable. Although CausalDiscrepancyVAE has a rigorous causal semantic, but it does not consider modeling of background latent variables, which from our experience, this is however necessary to get high-quality embeddings of the interventional data. To illustrate this point, we assessed how well the embedding from CausalDiscrepancyVAE may reflect biological information, using our benchmark (ARI, NMI, ASW of the different pathways captured by the experiment). Using the public code and available model trained by the authors (105 latent variables). We embedded all cells using the encoder network (after the DAG layer) and found that the performance was poor. Because our evaluation may be dependent on the number of latent variables, we re-tried this with a model that we fit ourselves from the available code, this time with 20 latent variables, and obtained similar results.

F.7. Real-word Data Details

Data Preprocessing We downloaded the data set using the ContrastiveVI package (Weinberger et al., 2023). The data set has measurements of the effects on gene expression levels of 284 different CRISPR-mediated perturbations on K562 cells. Each perturbation induced overexpression of a single

gene or a pair of genes. The background data set (8, 907 cells) is defined as all unperturbed cells. The target data set (24, 913 cells) is defined as all the perturbed cells whose genetic perturbation was labelled with a pathway by the authors of the original manuscript (Norman et al., 2019). The labeled pathways in the dataset are G1 Cycle, Erythroid, Pioneer Factors, Granulocyte Apoptosis, Megakaryocyte, and Pro Growth. We also filtered genes to retain the top 2, 000 highly variable genes.

Number of latent variables For our main results, we used 10 dimensional z and s .

F.8. Qualitative Comparison of Methods on Real-World Data

To understand the impact of the differences in metrics we reported in Table 4, we visualized the learned latent spaces for SO-cVAE, MO-CO-cVAE, as well as ContrastiveVI with UMAP (Figure 2).

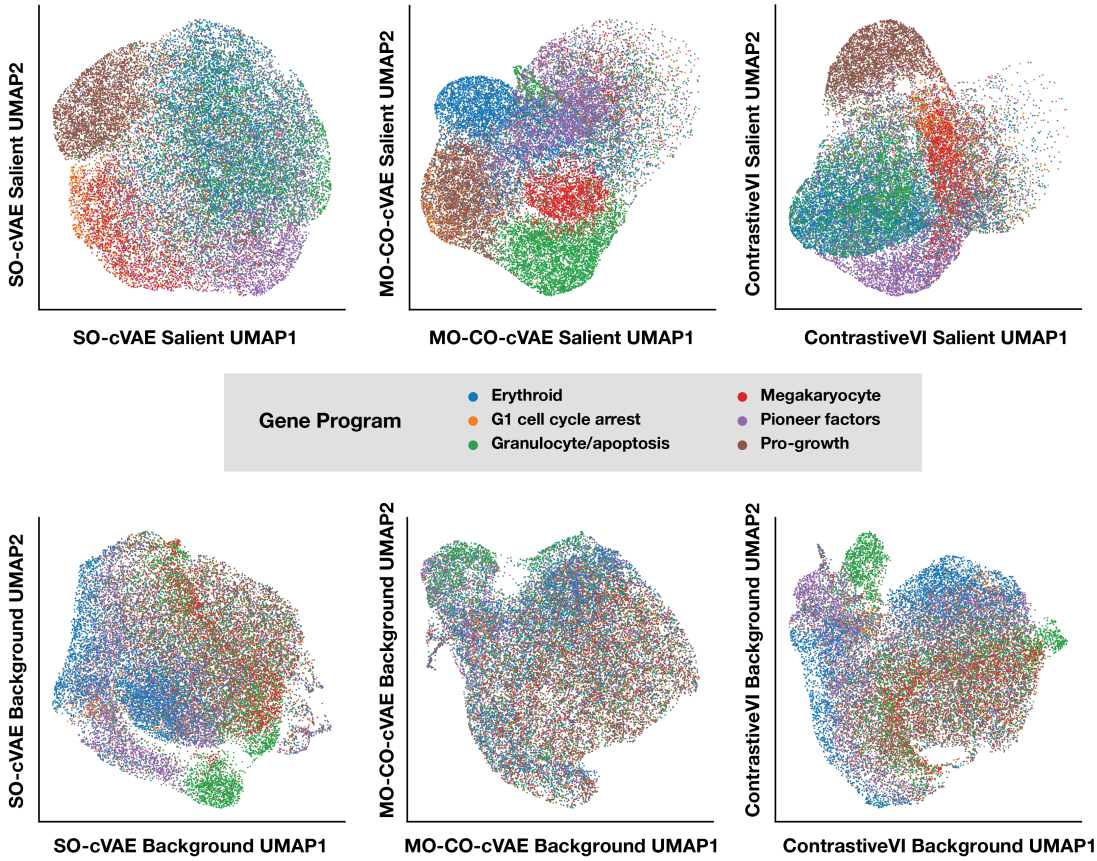

Figure 2: UMAP visualization of salient and background spaces from SO-cVAE, MO-CO-cVAE, as well as ContrastiveVI. Each point is a cell. Cells are colored by their group of genetic perturbation, where groups were assigned based on biological annotation from the authors of Norman et al. (2019).

Salient space interpretation In accordance with our desiderata, we observe for each method that the salient space recapitulates the perturbation type more accurately than the background space. However, we clearly notice that MO-CO-cVAE has best performance, as it more clearly delineates cells that underwent perturbations from the group "erythoid" versus "granulocytes/apoptosis". This was suggested by the high clustering and silhouette scores in Table 4.

Background space interpretation Additionally, we would like to make sure that the latent space contains as few information about the perturbations as possible. In this regard, it is important to note that, as expected, the un-regularized model SO-cVAE shows important separation of perturbation groups in its background latent space (as suggested by the high cross MCC in Table 4). However, the other methods show lesser leakage of that information in their background space.

Importantly, we see that the apoptosis group is still separated even in our model, which performs best in this benchmark. We attribute this to the fact that perturbations may sometimes bias cells towards expressing some gene programs that are inherently varying in the control population (such as cell death, here). Therefore, it is impossible to expect that $\boldsymbol{z}$ will be perfectly independent from the perturbation label on this data set.

F.9. Additional Results on Simulated Data

Comparison to MultiDomainCRL In order to investigate the performance of the ICA source matching method from Sturma et al. (2023), we applied their publicly available code on our simulated data, where the background data set and the target data set are used as two domains. We applied the FastICA algorithm with the known number of sources in each domain (e.g., 5 for the background data set, and 5 + 5 = 10 for the target data set). Then, we applied their individual source matching procedure, expecting it to match all the sources from the background data to one of the sources in the target data. Interestingly, the method consistently miss-identified the number of shared sources (underestimation), failing to match some of the background sources. This is already suggestive that the method may not work well on our benchmark. To report performance in a systematic fashion, we slightly modified their matching algorithm to associate each source in the background to exactly one source of the target data. The remainder of the sources for the target data set constituted the inferred salient variables, whereas the inferred salient variables for the background were set to zero. Even though we tried different metrics for matching latent variables (Wasserstein distance, as well as Smirnov Two-Sample Test) and also different strategies for normalizing the input data (raw values and logarithmic scale), the δ-MCC in the simple scenario of Table 1 was systematically between -0.1 and 0.1 for all experiments.

Misspecification of activation function We investigated how the performance changed when the data generating process was altered so that the leaky ReLU activation function is replaced by a hyperbolic tangent activation function (but the model stays the same). More precisely, we generated data according to two regimes. In the first one, the scale of the values at the hidden layer is small ($\ll 1$). This is interesting because in this case the tanh function well approximated a linear function (referred to as quasi-linear). In the second case, the scale of the values is larger ($\sim$1), and the tanh function is effectively non-linear (referred to as non-linear). For both data sets, we ran the method SO-cVAE in the ideal setting to assess disentanglement (akin to Table 1). We report those results in Table 5. As expected, the performance is high in the quasi-linear case, and indeed it surpasses the results of the manuscript. This is in agreement with the intuition that a linear model is easier

to identify from data compared to a piecewise-linear model. However, the performance drops in the non-linear case, which illustrates that identifiability is harder in this case. (It could also be explained by the mismatch between the data and the model, or by the excessive saturation from the tanh function at input values $\gg 1$).

Table 5: Identifiability under assumptions of known dimensions of latent spaces. Pearson MCC for SO-cVAE.

Data set	**MCC**$_{\hat{z}z}$ ($\uparrow$)	**MCC**$_{\hat{z}s}$ ($\downarrow$)	**MCC**$_{\hat{s}z}$ ($\downarrow$)	**MCC**$_{\hat{s}s}$ ($\uparrow$)	δ-**MCC** ($\uparrow$)
Tanh Quasi-linear	0.956 ± 0.001	0.047 ± 0.008	0.051 ± 0.005	0.963 ± 0.002	0.910 ± 0.007
Rescaled Tanh Non-linear	0.834 ± 0.011	0.122 ± 0.020	0.139 ± 0.033	0.541 ± 0.041	0.557 ± 0.043

Additional simulation results In this section, we present the following experimental results:

- The Pearson MCC scores under Gaussian noise when the number of latent variable is known (Table 6)
- The Spearman MCC scores under Poisson and negative binomial noise when the number of latent variable is known (Table 7).
- The Pearson MCC scores when the number of latent variables is unknown (Table 8).
- The Spearman MCC scores when the number of latent variables is unknown (Table 9).
- The Pearson MCC scores under regularization (Table 10).
- The Spearman MCC scores under regularization (Table 11).

Table 6: Identifiability under assumptions of known dimensions of latent spaces. Best in bold.

Model	Noise	$\mathbf{MCC}_{\hat{z}z}$ (↑)	$\mathbf{MCC}_{\hat{z}s}$ (↓)	$\mathbf{MCC}_{\hat{s}z}$ (↓)	$\mathbf{MCC}_{\hat{s}s}$ (↑)	δ-**MCC** (↑)
SO-cVAE	**Gaussian**	$\mathbf{0.96 \pm 0.01}$	$\mathbf{0.12 \pm 0.03}$	$\mathbf{0.08 \pm 0.01}$	$\mathbf{0.96 \pm 0.01}$	$\mathbf{0.86 \pm 0.01}$
MO-cVAE		$\mathbf{0.96 \pm 0.01}$	0.13 ± 0.03	$\mathbf{0.08 \pm 0.02}$	$\mathbf{0.96 \pm 0.01}$	$\mathbf{0.86 \pm 0.01}$
VAE		$\mathbf{0.96 \pm 0.01}$	0.15 ± 0.02	0.15 ± 0.01	0.95 ± 0.01	0.80 ± 0.02

Table 7: Identifiability of contrastive analysis models under assumptions of known dimensions of latent spaces (Spearman MCC).

Model	Noise	$\mathbf{MCC}_{\hat{z}z}$ (↑)	$\mathbf{MCC}_{\hat{z}s}$ (↓)	$\mathbf{MCC}_{\hat{s}z}$ (↓)	$\mathbf{MCC}_{\hat{s}s}$ (↑)	δ-**MCC** (↑)
MO-cVAE	**Poisson**	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.08 \pm 0.01}$	$\mathbf{0.07 \pm 0.02}$	$\mathbf{0.96 \pm 0.01}$	$\mathbf{0.87 \pm 0.01}$
SO-cVAE		$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.08 \pm 0.01}$	0.08 ± 0.02	$\mathbf{0.96 \pm 0.01}$	0.86 ± 0.01
VAE		0.88 ± 0.04	0.17 ± 0.09	0.14 ± 0.07	0.94 ± 0.04	0.76 ± 0.12
MO-cVAE	**Negative binomial**	$\mathbf{0.95 \pm 0.01}$	0.14 ± 0.01	$\mathbf{0.07 \pm 0.01}$	$\mathbf{0.96 \pm 0.01}$	$\mathbf{0.84 \pm 0.01}$
SO-cVAE		$\mathbf{0.95 \pm 0.01}$	$\mathbf{0.13 \pm 0.02}$	$\mathbf{0.07 \pm 0.01}$	0.95 ± 0.01	$\mathbf{0.84 \pm 0.01}$
VAE		0.82 ± 0.11	0.46 ± 0.18	0.37 ± 0.18	0.82 ± 0.10	0.41 ± 0.28

Table 8: Identifiability of contrastive analysis models under misspecification of latent dimensions (Pearson MCC).

	q	$\mathbf{MCC}_{\hat{z}z}$ (↑)	$\mathbf{MCC}_{\hat{z}s}$ (↓)	$\mathbf{MCC}_{\hat{s}z}$ (↓)	$\mathbf{MCC}_{\hat{s}s}$ (↑)	δ-**MCC** (↑)
SO-cVAE	**5**	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.08 \pm 0.01}$	$\mathbf{0.07 \pm 0.02}$	0.92 ± 0.01	0.84 ± 0.01
	7	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.08 \pm 0.01}$	0.30 ± 0.06	$\mathbf{0.94 \pm 0.02}$	0.73 ± 0.03
	10	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.08 \pm 0.01}$	0.45 ± 0.02	$\mathbf{0.94 \pm 0.02}$	0.66 ± 0.01
	15	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.08 \pm 0.02}$	0.59 ± 0.04	$\mathbf{0.94 \pm 0.02}$	0.58 ± 0.02
MO-cVAE	**5**	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.08 \pm 0.01}$	$\mathbf{0.07 \pm 0.01}$	$\mathbf{0.94 \pm 0.01}$	$\mathbf{0.85 \pm 0.01}$
	7	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.08 \pm 0.01}$	0.15 ± 0.02	$\mathbf{0.94 \pm 0.02}$	0.81 ± 0.01
	10	$\mathbf{0.91 \pm 0.01}$	0.09 ± 0.01	0.25 ± 0.03	$\mathbf{0.94 \pm 0.02}$	0.75 ± 0.01
	15	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.08 \pm 0.02}$	0.36 ± 0.04	$\mathbf{0.94 \pm 0.02}$	0.70 ± 0.02

Table 9: Identifiability of contrastive analysis models under misspecification of latent dimensions (Spearman MCC).

	q	**MCC**$_{\hat{z}z}$ (↑)	**MCC**$_{\hat{z}s}$ (↓)	**MCC**$_{\hat{s}z}$ (↓)	**MCC**$_{\hat{s}s}$ (↑)	δ-**MCC** (↑)
SO-cVAE	**5**	0.92 ± 0.01	0.08 ± 0.01	$\mathbf{0.08 \pm 0.02}$	$\mathbf{0.96 \pm 0.01}$	0.86 ± 0.01
	7	$\mathbf{0.93 \pm 0.01}$	$\mathbf{0.08 \pm 0.01}$	0.30 ± 0.06	$\mathbf{0.96 \pm 0.00}$	0.75 ± 0.03
	10	$\mathbf{0.93 \pm 0.01}$	$\mathbf{0.08 \pm 0.01}$	0.45 ± 0.03	$\mathbf{0.96 \pm 0.00}$	0.68 ± 0.02
	15	$\mathbf{0.93 \pm 0.01}$	$\mathbf{0.08 \pm 0.02}$	0.60 ± 0.03	$\mathbf{0.96 \pm 0.00}$	0.60 ± 0.02
MO-cVAE	**5**	0.92 ± 0.01	$\mathbf{0.08 \pm 0.01}$	$\mathbf{0.07 \pm 0.02}$	$\mathbf{0.96 \pm 0.01}$	$\mathbf{0.87 \pm 0.01}$
	7	$\mathbf{0.93 \pm 0.01}$	$\mathbf{0.08 \pm 0.01}$	0.15 ± 0.02	$\mathbf{0.96 \pm 0.00}$	0.83 ± 0.01
	10	$\mathbf{0.93 \pm 0.01}$	$\mathbf{0.08 \pm 0.01}$	0.26 ± 0.04	$\mathbf{0.96 \pm 0.00}$	0.77 ± 0.02
	15	$\mathbf{0.93 \pm 0.01}$	$\mathbf{0.08 \pm 0.02}$	0.37 ± 0.04	$\mathbf{0.96 \pm 0.00}$	0.72 ± 0.02

Table 10: The impact of regularization on contrastive analysis models under misspecification of latent dimensions (Pearson MCC).

		MCC$_{\hat{z}z}$ (↑)	**MCC**$_{\hat{z}s}$ (↓)	**MCC**$_{\hat{s}z}$ (↓)	**MCC**$_{\hat{s}s}$ (↑)	δ-**MCC** (↑)
SO-U-cVAE	$\lambda = 0$	$\mathbf{0.91 \pm 0.01}$	0.08 ± 0.01	0.45 ± 0.02	$\mathbf{0.94 \pm 0.00}$	0.66 ± 0.01
	$\lambda = 10$	$\mathbf{0.91 \pm 0.01}$	0.08 ± 0.01	0.37 ± 0.04	$\mathbf{0.94 \pm 0.00}$	0.70 ± 0.02
	$\lambda = 50$	$\mathbf{0.91 \pm 0.01}$	0.08 ± 0.01	0.25 ± 0.03	$\mathbf{0.94 \pm 0.00}$	0.76 ± 0.01
	$\lambda = 100$	0.80 ± 0.03	0.11 ± 0.05	0.29 ± 0.08	0.92 ± 0.03	0.66 ± 0.08
SO-CO-cVAE		$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.07 \pm 0.01}$	0.23 ± 0.03	$\mathbf{0.94 \pm 0.00}$	0.77 ± 0.01
MO-U-cVAE	$\lambda = 0$	$\mathbf{0.91 \pm 0.01}$	0.09 ± 0.01	0.25 ± 0.03	$\mathbf{0.94 \pm 0.00}$	0.75 ± 0.01
	$\lambda = 10$	$\mathbf{0.91 \pm 0.01}$	0.08 ± 0.01	0.21 ± 0.02	$\mathbf{0.94 \pm 0.00}$	0.78 ± 0.01
	$\lambda = 50$	0.88 ± 0.03	0.08 ± 0.01	0.16 ± 0.02	$\mathbf{0.94 \pm 0.00}$	0.79 ± 0.02
	$\lambda = 100$	0.79 ± 0.02	0.09 ± 0.01	0.19 ± 0.02	0.93 ± 0.00	0.73 ± 0.03
MO-CO-cVAE		$\mathbf{0.91 \pm 0.01}$	0.08 ± 0.01	$\mathbf{0.17 \pm 0.02}$	$\mathbf{0.94 \pm 0.00}$	$\mathbf{0.80 \pm 0.01}$

Table 11: The impact of regularization on contrastive analysis models under misspecification of latent dimensions (Spearman MCC).

		$\mathbf{MCC}_{\hat{z}z}$ (↑)	$\mathbf{MCC}_{\hat{z}s}$ (↓)	$\mathbf{MCC}_{\hat{s}z}$ (↓)	$\mathbf{MCC}_{\hat{s}s}$ (↑)	δ-**MCC** (↑)
SO-U-cVAE	$\lambda = 0$	$\mathbf{0.93 \pm 0.01}$	0.08 ± 0.01	0.45 ± 0.03	$\mathbf{0.96 \pm 0.00}$	0.68 ± 0.02
	$\lambda = 10$	$\mathbf{0.93 \pm 0.01}$	0.08 ± 0.01	0.38 ± 0.04	$\mathbf{0.96 \pm 0.00}$	0.71 ± 0.02
	$\lambda = 50$	0.92 ± 0.01	0.08 ± 0.01	0.25 ± 0.02	$\mathbf{0.96 \pm 0.00}$	0.78 ± 0.01
	$\lambda = 100$	0.81 ± 0.03	0.11 ± 0.06	0.29 ± 0.08	0.94 ± 0.03	0.67 ± 0.08
SO-CO-cVAE		$\mathbf{0.93 \pm 0.01}$	$\mathbf{0.07 \pm 0.01}$	0.23 ± 0.03	$\mathbf{0.96 \pm 0.00}$	0.79 ± 0.01
MO-U-cVAE	$\lambda = 0$	$\mathbf{0.93 \pm 0.01}$	0.08 ± 0.01	0.26 ± 0.04	$\mathbf{0.96 \pm 0.00}$	0.77 ± 0.02
	$\lambda = 10$	$\mathbf{0.93 \pm 0.01}$	0.08 ± 0.01	0.22 ± 0.03	$\mathbf{0.96 \pm 0.00}$	0.79 ± 0.01
	$\lambda = 50$	0.89 ± 0.03	0.08 ± 0.01	0.16 ± 0.03	$\mathbf{0.96 \pm 0.00}$	0.80 ± 0.03
	$\lambda = 100$	0.81 ± 0.02	0.08 ± 0.01	0.20 ± 0.02	$\mathbf{0.96 \pm 0.00}$	0.74 ± 0.03
MO-CO-cVAE		$\mathbf{0.93 \pm 0.01}$	0.08 ± 0.01	$\mathbf{0.18 \pm 0.02}$	$\mathbf{0.96 \pm 0.00}$	$\mathbf{0.82 \pm 0.01}$