Score-Based Interaction Testing in Pairwise Experiments

Jana Osea[†*]
Valence Labs
Montreal, Canada
jana.osea@stat.ubc.ca

Zuheng Xu[*†]
University of British Columbia
Vancouver, Canada
zuheng.xu@stat.ubc.ca

Cian Eastwood
Valence Labs
London, UK
cian@valencelabs.com

Jason Hartford
Valence Labs & University of Manchester
London, UK
jason@valencelabs.com

Abstract

Interaction tests are crucial in the sciences, particularly in pairwise perturbation experiments where they can be used to reveal causal relationships in a system. Recently, Zuheng et al. [2024] proposed a framework and statistical tests for detecting pairwise interactions from unstructured data like images. While effective, these tests can be prohibitively expensive due to training costs that are quadratic in the number of perturbations. To address this, we explore alternative *score-based interaction tests* that can be linear in the number of perturbations. In particular, we propose using the aggregated Kernelized Stein Discrepancy (KSD, Schrab et al. 2023) as a formal hypothesis test. In our experiments, we compare to the Fisher Divergence (FD)—a score-based test that scales quadratically in the number of experiments—and show that: (i) with low-dimensional inputs, both methods perform well; (ii) with high-dimensional inputs like images, KSD's sensitivity to kernel choice hurts performance; and (iii) projecting high-dimensional data into lower-dimensional spaces solves this issue for KSD, resulting in an effective and computationally-efficient interaction test.

1 Introduction

Detecting interactions between experimental perturbations is a common and significant challenge in many scientific fields. This is particularly true for pairwise perturbation experiments which seek to understand the effect of combined/paired interventions on system in order to reveal interactions that would be missed by single interventions [Lehner, 2011]. For example, *synthetic lethality* [Nijman, 2011] relationships between genes occur when knocking out an individual gene results in no effect on a cell, but knocking out pairs of genes results in the cell dying. To detect these relationships, the dominant approach is to pre-select an outcome variable of interest—such as cell viability in the synthetic lethality example. However, this approach is limited as it can only detect interactions if one has selected the correct outcome variable. Recently, Zuheng et al. [2024] showed that it is possible to detect these interactions from unstructured data (such as the pixels in an image) using a Kullback–Leibler (KL) divergence-based interaction test. While effective, this method requires fitting a quadratic number of models in order to estimate these KL divergences. For large screening experiments, this can be computationally prohibitive.

[*]Work done during an internship at Valence Labs
[†]Equal contribution

Accepted to the NeurIPS 2024 Workshop on Causal Representation Learning.

In this work, we propose an alternative approach based on the (Stein) score function, $\nabla_x \log p(x)$, which is estimated by many of the popular diffusion modelling techniques [Ho et al., 2020]. We show that it is possible to test for the same dependencies using the Kernelized Stein Discrepancy (KSD, Liu et al. 2016, Chwialkowski et al. 2016, Gretton et al. 2012). This approach allows us to estimate the condition score function from single-perturbation experiments, and test for dependence using samples come from the double-perturbation experiments. As a result, *we only need to fit an estimator for the score across a linear number of perturbations*.

In our experiments, we compare the KSD to two alternative methods on synthetic data: (1) the exact Fisher Divergence (FD), which also uses score functions but involves fitting the score for all pairwise perturbations (i.e. a quadratic number of condition); and (2) the KL-based test originally proposed by Zuheng et al. [2024]. We find that the KSD is effective in low-dimensional settings but struggles with high-dimensional images. However, by employing an appropriate low-dimensional representation, we are able to achieve similar performance to that seen in the low-dimensional experiments. In particular, we find that the final hidden layer of a classifier is the most effective feature space for testing interactions.

2 Related Work

While we do not attempt to disentangle latent variables, our approach to detecting interactions builds on the modelling assumptions from causal representation learning [e.g. Hyvarinen et al., 2019, Schölkopf, 2021, Lachapelle et al., 2022, Lippe et al., 2023, Buchholz et al., 2023, Ahuja et al., 2023, Zhang et al., 2024], by assuming the observation of a nonlinear mixing function of latent variables. Like Varici et al. [2023], we leverage the score function, but do not attempt to disentangle latent variables. If successful, disentangling latent variables would be sufficient to test for independence directly, but disentanglement is extremely challenging in practice because most methods rely on untestable assumptions that are difficult to verify in scientific applications. Our tests of separability is based on the framework introduced by Zuheng et al. [2024], which tests the notion of independence by leveraging assumptions about separable concepts [Wang et al., 2024]. Like us, Wang et al. uses the score functions, but they assume separability holds rather than attempting to test for it. Our test is built on the kernelized Stein Discrepancy which has been explored in [Liu et al., 2016, Chwialkowski et al., 2016, Gretton et al., 2012] which is based on Stein's method [Stein, 1972]; for an accessible introduction to the KSD, see Liu [2016].

3 Background

3.1 Latent separability

We follow a similar setup as introduced by Zuheng et al. [2024] with a slightly different presentation. We assume that observation X obeys the following generative process involving latent random variable Z and noise variable U:

$$X^t = f(Z^t, U), \text{ where } Z^t \sim p_Z(z|T=t), \quad U \sim p_U, \quad T \perp\!\!\!\perp U, \quad T \not\perp\!\!\!\perp Z, \quad U \perp\!\!\!\perp Z. \tag{1}$$

Here T denotes the perturbation variable and t indexes the experimental perturbations. Throughout, we assume that there are n single perturbations, denoted as $\delta_i := \{T = i\}$ ($\forall i \in [n]$), and $\binom{n}{2}$ pairwise perturbations, denoted as $\delta_{ij} := \delta_i \cap \delta_j$ ($i \neq j$). And we take $\delta_0 = \{T = 0\}$ to represent the unperturbed environment. $T \not\perp\!\!\!\perp Z, T \perp\!\!\!\perp U$ yield that the perturbation only intervenes the latent variable Z but not the noise U, and the structural equation f does not get intervened as well. Hence, Eq. (1) ensures that $X \perp\!\!\!\perp (T_1, \ldots, T_n)|Z$. Throughout we assume that the latent variable Z admits a causal factorization, such that,

$$p_Z(z) = \prod_{l=1}^{L} p_{Z_l}(z_l|\text{Pa}(z_l)), \tag{2}$$

where $\forall l \in [L], \text{Pa}(Z_l) \subseteq \{Z_1, \ldots, Z_L\}$ denotes the parent nodes of z_l, the set of latent variables that causally influence Z_l. Each perturbation targets a subset of latents, inducing a soft intervention, which changes the corresponding conditional distributions (note that hard interventions which remove the depencence on causal parents are a special case). For example, suppose that the latent variable Z_i

is targeted by the intervention δ_i, then $p_{Z_i}(z_i|\text{Pa}(z_i))$ gets changed into $p^{\dagger}_{Z_i}(z_i|\text{Pa}(z_i))$, hence

$$p_Z(z|\delta_i) = p^{\dagger}_{Z_i}(z_i|\text{Pa}(z_i)) \cdot \prod_{l \neq i} p_{Z_l}(z_l|\text{Pa}(z_l)).$$

This model leads to a natural interpretation of the interaction between two perturbations: if two perturbations are non-interacting, they should target distinct latent factors. Zuheng et al. [2024] formally define this relationship as *separability*.

Definition 3.1. *Denote $\mathcal{I}(t)$ the index of latent variables that are targeted by the perturbation t. Perturbations δ_i, δ_j are separable if $\mathcal{I}(i) \cap \mathcal{I}(j) = \emptyset$.*

Zuheng et al. showed that separability has a testable implication: if δ_i and δ_j are separable, then,

$$\log \frac{p(x|\delta_i)}{p(x|\delta_0)} + \log \frac{p(x|\delta_j)}{p(x|\delta_0)} = \log \frac{p(x|\delta_{ij})}{p(x|\delta_0)}.$$

In Theorem 3.2, we show a similar relationship based on score functions (the gradients of the log-densities) instead of log-densities, given an injectivity condition on the structural equation f. Notably, despite the similar implications, Theorem 3.2 operates in a more general setting compared to Zuheng et al. [2024, Theorem 3.6]. Specifically, we do not explicitly assume a diffeomorphism between the latent variable Z and the observation X.

Theorem 3.2. *In the model described by Eq.* (1)*, further assume that the structural equation f is injective. Then, if perturbations δ_i and δ_j are separable, we have*

$$s(x|\delta_{ij}) = s(x|\delta_i) + s(x|\delta_j) - s(x|\delta_0), \tag{3}$$

where $s(x|\delta_i) = \nabla_x \log p(x|\delta_i)$ for all $i \in [n]$.

On the right-hand side of Eq. (3), $s(x|\delta_i) + s(x|\delta_j) - s(x|\delta_0)$ should be considered as the intended score function of the double perturbation group when δ_i and δ_j are separable. In contrast, $s(x|\delta_{ij})$ represents the actual score function.

3.2 Exact Fisher divergence and kernelized Stein discrepancy

In this section, we review two statistical divergences involving score functions of distributions, the Fisher divergence and the kernelized Stein discrepancy, which will be used to quantify the violation of the score additivity Eq. (3) between a pair of perturbations.

Fisher divergence (FD) FD [Hyvärinen et al., 2009, Lyu, 2009] measures the discrepancy between two distributions p, q by comparing their score function. It is defined as:

$$\mathrm{D}_F(p||q) := \mathbb{E}_{x\sim p}\left[\|\nabla \log p(x) - \nabla \log q(x)\|^2\right].$$

Notice that $\mathrm{D}_F(p||q) = 0$ if and only if $p = q$, making it a valid statistical divergence. In practice, if the score functions $\nabla \log p(x)$ and $\nabla \log q(x)$ (or their estimates) are available, FD can be estimated using Monte Carlo samples from p.

Kernelized stein discrepancy (KSD) KSD [Liu et al., 2016, Chwialkowski et al., 2016, Gorham and Mackey, 2015] is a nonparametric measure that assesses the goodness-of-fit between a target distribution p and a model distribution q by leveraging Stein's method [Stein, 1972] and reproducing kernel Hilbert spaces (RKHS). It is defined as:

$$\mathrm{D}_s(p||q) := \max_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_{x\sim p}\left[f(x)s_q(x) + \nabla_x f(x)\right] \tag{4}$$

where $s_q(x) := \nabla_x \log q(x)$, $\mathcal{H}$ is a RKHS associated to some positive definite kernel $k(\cdot,\cdot)$. When both p and q have smooth densities, Eq. (4) has a more explicit expression provided with the choice of kernel k:

$$\mathrm{D}_s(p||q) := \mathbb{E}_{x,x'\sim p}\left[u_q\left(x, x'\right)\right],$$

where $u_q\left(x, x'\right)$ is the Steinized kernel expressed as follows:

$$s_q(x)^T k\left(x, x'\right) s_q\left(x'\right) + s_q(x)^T \nabla_{x'} k\left(x, x'\right) + \nabla_x k\left(x, x'\right)^T s_q\left(x'\right) + \operatorname{tr}\left(\nabla_{x,x'} k\left(x, x'\right)\right).$$

In fact, KSD can also be viewed as a specific maximum mean discrepancy (MMD) [Gretton et al., 2012] with kernel u_q. For most choice of k, $\mathrm{D}_s(p||q) = 0$ if and only if $p = q$.

In practice, one can estimate $\mathrm{D}_s(p||q)$ unbiasedly via i.i.d. samples $\{x_i\}$ from q as follows:

$$\hat{\mathrm{D}}_s(p||q) = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} u_q\left(x_i, x_j\right). \tag{5}$$

An important property of this estimator is that it only requires the score function s_q and samples from p. Moreover, the statistical properties of this estimate are well-understood for both cases when $p = q$ and $p \neq q$ [Liu et al., 2016], enabling the use of KSD for nonparametric goodness-of-fit hypothesis testing without the need for explicit density estimation [Liu et al., 2016, Chwialkowski et al., 2016, Schrab et al., 2023].

4 Score-based pairwise interaction test

Now we are ready to present our methodology of pairwise interaction test. To test the separability for a given pair of perturbations δ_i, δ_j, we aim to measure how well the composition of single perturbation scores characterize the double perturbation score—the relationship described in Eq. (3). Specifically, we need a principled metric to quantify the violation of Eq. (3).

A natural option is to measure the FD between the left side and right side of Eq. (3) using the estimated scores of all perturbation groups, i.e.,

$$\int \left\| \hat{s}(x|\delta_{ij}) - \hat{s}(x|\delta_i) - \hat{s}(x|\delta_j) + \hat{s}(x|\delta_0) \right\|^2 p(x|\delta_{ij}) \mathrm{d}x, \tag{6}$$

where $\hat{s}$ denotes the estimated score functions obtained from the data of corresponding perturbation group. In our experiments, we estimate the scores using the denoising diffusion probabilistic model (DDPM) [Ho et al., 2020], and compute the expectation using samples from double perturbation group $p(x|\delta_{ij})$. As previously explained, Eq. (6) should be interpreted as the FD between the true data distribution under the double perturbation and the separability model.

However, this still requires us to learn the score functions for all pairwise combinations of perturbations ($\hat{s}_{ij}$), resulting in a quadratic computational cost with respect to the total number of single perturbations. Moreover, while this method provides a measure of the violation of the score additivity relationship, it lacks a rigorous statistical framework for determining whether the separability assumption for δ_i, δ_j is indeed violated.

We can address both problems using KSD. Specifically, we first estimate $\hat{s}_{ij}^{\text{sep}} := \hat{s}(x|\delta_i) + \hat{s}(x|\delta_j) - \hat{s}(x|\delta_0)$, which can be interpreted as the estimated score function for the double perturbation group under the separability model. Then, we obtain the KSD estimator Eq. (5) using data from $p(x|\delta_{ij})$ and $\hat{s}_{ij}^{\text{sep}}$. This amounts to measure the goodness-of-fit of the data distribution $p(x|\delta_{ij})$ to the separability model. Notice that the KSD estimates only involve learning scores for the *single* perturbation groups, avoiding the quadratic training cost as using the FD estimates (which require scores from the double perturbation group as well).

To address the second issue, we use a KSD based non-parametric goodness-of-fit testing procedure called the aggregated KSD test [Schrab et al., 2023], which is a variant of KSD-based goodness-of-fit test that allows users to combine multiple KSD test results evaluated on different choice of kernels. Then, it uses the bootstrap method to estimate the p-value for the hypothesis test

$$H_0 : s_{ij}^{\text{sep}}(x) = \nabla \log p(x|\delta_{ij}), \quad H_1 : s_{ij}^{\text{sep}}(x) \neq \nabla \log p(x|\delta_{ij}).$$

This allows us to draw justified statistical conclusion whether the separability relationship holds (null hypothesis).

4.1 KSD test on learned embeddings

In our experiments, both the FD and KSD scores, as well as the aggregated KSD test, work effectively for low-dimensional observations. However, we find that in our image examples, the KSD-based method fails to perform reliably in the raw pixel space; the estimated KSD fails to reveal the true interaction relationships, and statistical conclusions from the aggregated KSD test are unreliable.

The fundamental reason is the lack of appropriate kernel choices when data do not follow Euclidean geometry. Typically, for kernel-based tests, we choose Gaussian kernels, Matérn kernels, or inverse quadratic kernels, all of which implicitly assume the data is defined in a Euclidean space. However, the pixel space of image data is clearly not Euclidean. Furthermore, the power of kernel-based tests decays with increasing dimensionality [Gao and Shao, 2023].

Therefore, for high-dimensional data, we first learn a low-dimensional embedding and then apply the KSD framework to these embeddings. The question then arises: How should we learn this embedding? Heuristically, one can obtain the embedding using the encoder from a trained Variational Autoencoder (VAE) or by extracting features from the last layer of a classifier. Interestingly, we can show that in our model assumptions Eq. (1), the representation layer from a classifier can be a preferred choice.

Proposition 4.1 states that the propensity score $p(T|X)$, which is an estimatable quantity using classifiers, captures *all* relevant information about perturbation T with respect to the latent variable Z. Indeed, in our experiments, we find the embedding learned via a classifier significantly outperforms those learned from VAEs.

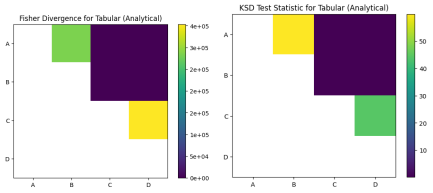
Proposition 4.1. *Denote the propensity score* $\pi(z) := p(T|Z = z) \in [0, 1]^{\#perturbations}$. *Under the same assumption of Theorem 3.2, we have that propensity scores given the observation* X *equals to the one given the latent* Z*, i.e.,* $\pi(Z) = \pi(X)$. *Furthermore,* $T \perp\!\!\!\perp Z|\pi(X)$.

We remark that this result was previously established in multimodality matching literature [Xi and Hartford, 2024]. It's also worth noting that the Fisher Divergence works relatively well on both low- and high-dimensional data. The general precedure for estimating FD and KSD in high-dim is depicted in Fig. 4.

5 Experiments

Setup All experiments are performed on generated synthetic data. We consider two main types of data, low dimensional which are 3 dimensional real numbers and high dimensional which are 32x32x3 images. For both cases, we assume 4 perturbation classes labelled $A - D$ and 3 latent variables. The DAG structure as well as the data generation process can be found in Appendix A.2.1. We evaluate how well the Fisher Divergence and KSD detects the separability of the perturbation classes using three main experiment settings: (1) low dimensional with analytical score functions, (2) low dimensional with estimated score functions, and (3) high dimensional with estimated score functions. We estimate the scores using a Denoising Diffusion Probabilistic Model (DDPM) framework [Ho et al., 2020]. For setting (3), we also evaluate the KSD using the estimated score function of a lower dimensional embedding representation of the images. As baseline for setting (2) and (3), we compare the results with the log density estimates separability score estimates from the [Zuheng et al., 2024]. Further details on the data generation, model training, and evaluation are found in Appendix A.

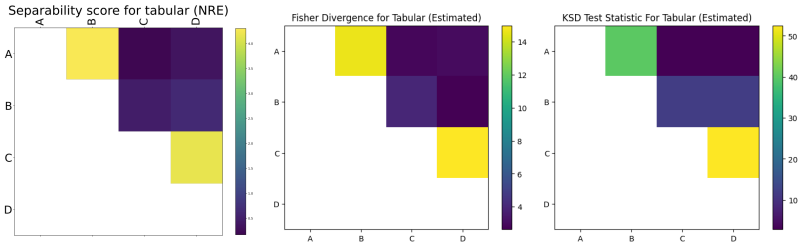
Low Dimension For all the following results, it is important to note that the most top left and most bottom right quadrants should have the highest interaction value e.g. most bright since interventions A & B and C & D are interacting. Figure 1 evaluates the interaction using analytical scores. Both subfigures 1a and 1b show that Fisher Divergence and KSD correctly identifies interacting perturbations.



(a) Fisher Div. (b) KSD

Figure 1: Separability of synthetic low dimensional (tabular) data evaluated using Fisher Divergence and KSD on exact analytical score functions.

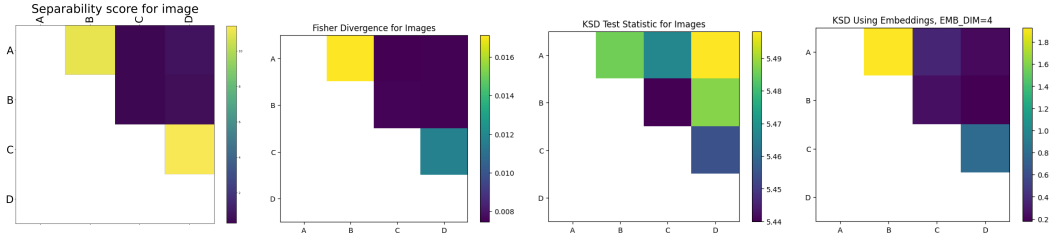
Figure 2 presents the results of using estimated scores to evaluate the separability of interventions with 3 dimensional real number observations. In both subfigures 2a and 2b, the KL-divergence of log densities and the Fisher Divergence of scores are nearly identical in successfully identifying the correct interacting pairs, while also keeping the non-interacting pairwise values low. Subfigure 2c demonstrates that while KSD is effective at detecting the correct interacting pairs with the correct p-value accepts and reject, it assigns relatively higher values to the non-interacting pairs compared to the previous methods.



(a) KL Div. & Log Density (b) Fisher Div. (c) KSD

Figure 2: Separability of synthetic low dimensional (tabular) data evaluated using Fisher Divergence and KSD on estimated score functions.

High-Dimension Figure 3 presents the results of using estimated scores to evaluate the separability of interventions with high dimensional 32x32x3 images. The KL Divergence between log densities shows that it correctly identifies the interacting interventions with equal magnitude of separability value. For the KSD evaluation, two different score estimation methods are used where we estimate score functions from the images directly and we also estimate score functions from a learned embedding space. In 3d, we use a 4 dimensional embedding space obtained from the last layer of a classifier. The results of the KSD using the embedding space is very similar to the Fisher Divergence that identifies the correct interacting interventions. Contrary to this, the KSD using the score function trained on the image directly performs very poorly with incorrectly identified interacting interventions.



(a) KL Div. & Log Density (b) Fisher Div. (c) KSD (Images) (d) KSD (Embeddings, dim=4)

Figure 3: Separability of synthetic high dimensional (image) data evaluated using Fisher Divergence and KSD on estimated score functions in images and embedding space.

6 Discussion

We have shown that score-based methods are effective for detecting interactions in pairwise perturbation experiments, performing similarly to the KL divergence between log-density ratios under comparable conditions. While the Fisher Divergence method still has quadratic training costs in the number of perturbations,mirroring the computational challenges of using the KL divergence, KSD has linear training costs in the number of perturbations since it *only requires score estimators for single perturbations* (and data from pairwise experiments).

Limitations One of the main limitations of this paper lies in the evaluation, which was conducted solely on synthetic tabular and image data with relatively-simple data-generating processes. While

this serves as a promising first step in validating the use of score functions for interaction detection, it falls short of capturing the complexity found in real-world pairwise perturbation experiments, such as gene knockout studies, drug combinations, or material science applications, where interactions are often more nuanced and challenging to detect. Another main limitation is that unlike our synthetic examples where we have access to ground truth, in real-world examples we do know at which noise time step is best to evaluate the separability. Correctly evaluating separability heavily relies on knowing the correct noising timesteps is appropriate.

Future work In our experiments, we focused solely on a single score estimation method—the DDPM. Exploring alternative methods, such as Flow Matching with interpolants with characterizable score functions, could provide valuable insights into how Fisher Divergence and KSD evaluations might change. We also did not assess the sensitivity of these score-based methods to noise, which is a crucial factor in real-world data. It would be important to evaluate how robust these methods are when applied to noisy or incomplete data.

References

K. Ahuja, D. Mahajan, Y. Wang, and Y. Bengio. Interventional causal representation learning. In *International conference on machine learning*, pages 372–407. PMLR, 2023.

S. Buchholz, G. Rajendran, E. Rosenfeld, B. Aragam, B. Schölkopf, and P. K. Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=q131tA7HCT.

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *International conference on machine learning*, pages 2606–2615. PMLR, 2016.

H. Gao and X. Shao. Two sample testing in high dimension via maximum mean discrepancy. *Journal of Machine Learning Research*, 24(304):1–33, 2023.

J. Gorham and L. Mackey. Measuring sample quality with stein's method. *Advances in neural information processing systems*, 28, 2015.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

A. Hyvärinen, J. Hurri, P. O. Hoyer, A. Hyvärinen, J. Hurri, and P. O. Hoyer. Estimation of non-normalized statistical models. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*, pages 419–426, 2009.

A. Hyvarinen, H. Sasaki, and R. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

A. Kotelnikov, D. Baranchuk, I. Rubachev, and A. Babenko. Tabddpm: Modelling tabular data with diffusion models, 2022. URL https://arxiv.org/abs/2209.15421.

S. G. Krantz and H. R. Parks. *Geometric integration theory*. Springer Science & Business Media, 2008.

S. Lachapelle, P. Rodriguez, Y. Sharma, K. E. Everett, R. L. PRIOL, A. Lacoste, and S. Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In B. Schölkopf, C. Uhler, and K. Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 428–484. PMLR, 11–13 Apr 2022. URL https://proceedings.mlr.press/v177/lachapelle22a.html.

B. Lehner. Genetic interactions. *Current Biology*, 21(20):R811–R815, 2011.

P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and E. Gavves. BISCUIT: Causal representation learning from binary interactions. In R. J. Evans and I. Shpitser, editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 1263–1273. PMLR, 31 Jul–04 Aug 2023. URL https://proceedings.mlr.press/v216/lippe23a.html.

Q. Liu. A short introduction to kernelized stein discrepancy. 2016. URL https://api.semanticscholar.org/CorpusID:16209224.

Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284. PMLR, 2016.

S. Lyu. Interpretation and generalization of score matching. In *Uncertainty in Artificial Intelligence*, 2009.

T. maintainers and contributors. Torchvision: Pytorch's computer vision library. https://github.com/pytorch/vision, 2016.

S. M. Nijman. Synthetic lethality: general principles, utility and detection using genetic screens in human cells. *FEBS letters*, 585(1):1–6, 2011.

B. Schölkopf. Toward causal representation learning. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2021. URL https://proceedings.mlr.press/v139/scholkopf21a.html.

A. Schrab, B. Guedj, and A. Gretton. Ksd aggregated goodness-of-fit test, 2023. URL https://arxiv.org/abs/2202.00824.

C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*, volume 6, pages 583–603. University of California Press, 1972.

B. Varici, E. Acarturk, K. Shanmugam, A. Kumar, and A. Tajer. Score-based causal representation learning with interventions, 2023. URL https://arxiv.org/abs/2301.08230.

P. von Platen, S. Patil, A. Lozhkov, P. Cuenca, N. Lambert, K. Rasul, M. Davaadorj, D. Nair, S. Paul, S. Liu, W. Berman, Y. Xu, and T. Wolf. Diffusers: State-of-the-art diffusion models, 2024. URL https://github.com/huggingface/diffusers.

Z. Wang, L. Gui, J. Negrea, and V. Veitch. Concept algebra for (score-based) text-controlled generative models, 2024. URL https://arxiv.org/abs/2302.03693.

J. Xi and J. Hartford. Propensity score alignment of unpaired multimodal data. *arXiv preprint arXiv:2404.01595*, 2024.

J. Zhang, K. Greenewald, C. Squires, A. Srivastava, K. Shanmugam, and C. Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 36, 2024.

Zuheng, Xu, M. Jain, A. Denton, S. Whitfield, A. Didolkar, B. Earnshaw, and J. Hartford. Automated discovery of pairwise interactions from unstructured data, 2024. URL https://arxiv.org/abs/2409.07594.

A Appendix

A.1 Proofs

Proof of Theorem 3.2. By the injectivity of f, we can apply the generalized change of variable formula [Krantz and Parks, 2008, Lemma 5.1.4] to express $\log p(x|T)$ by:

$$\begin{aligned}\log p(x|T) &= \log p_{Z,U}(f^{-1}(x)|T) + \log|\left(J_{f^{-1}}(x)\right)| \\ &= \log p_Z([f^{-1}(x)]_Z|T) + \log p_U([f^{-1}(x)]_U|T) + \log|\left(J_{f^{-1}}(x)\right)| \\ &= \log p_Z([f^{-1}(x)]_Z|T) + \log p_U([f^{-1}(x)]_U) + \log|\left(J_{f^{-1}}(x)\right)|,\end{aligned}$$

where

$$J_{f^{-1}}(x) = \sqrt{\det\left(\left[\frac{\partial}{\partial x}f^{-1}(x)\right]^T\left[\frac{\partial}{\partial x}f^{-1}(x)\right]\right)}, \quad \text{if } dim(X) > dim(Z) + dim(U),$$

$$J_{f^{-1}}(x) = \sqrt{\det\left(\left[\frac{\partial}{\partial x}f^{-1}(x)\right]\left[\frac{\partial}{\partial x}f^{-1}(x)\right]^T\right)}, \quad \text{if } dim(X) < dim(Z) + dim(U),$$

$$J_{f^{-1}}(x) = \det\left(\left[\frac{\partial}{\partial x}f^{-1}(x)\right]\right), \quad \text{if } dim(X) = dim(Z) + dim(U),$$

the second equality is by $Z \perp\!\!\!\perp U$, and the last equality is by $T \perp\!\!\!\perp Z$.

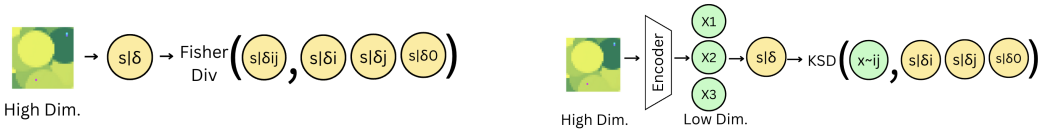
Given the separability assumption and Eq. (2), following the identical derivation as in the proof of Zuheng et al. [2024, Theorem 3.6.] yields that

$$\log p(x|\delta_{ij}) = \log p(x|\delta_i) + \log p(x|\delta_j) - \log p(x|\delta_0).$$

Taking gradient with respect to x on both side completes the proof. □

A.2 Experiment Details

Below is a summary of the methodology comparison between the Fisher Divergence and KSD method using high dimensional data.



(a) Fisher Div. (b) KSD

Figure 4: Interaction tests. (a) Fisher Divergence can be used with low- and high-dimensional data but needs to be trained on both single and pairwise/double perturbations. (b) KSD requires low-dimensional data (for the score estimator), but can be trained using only single perturbations. To use KSD on high-dimensional data, we learn low-dimensional embedding of the data.

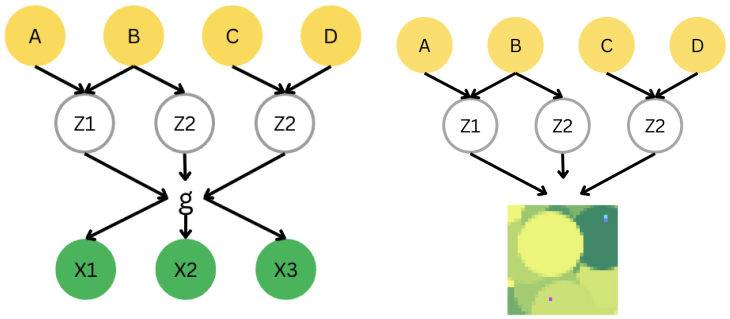
A.2.1 Data Generating Process

As mentioned in Section 5, we consider 4 perturbation classes labelled $A - D$ and their corresponding pairwise combinations which intervene on 3 latent variables labelled $Z1 - Z3$ as visualized in Figure 5. We consider a binary intervention, that is the latent variables have the following distributions

$$P_1 \sim \begin{cases}\text{Normal}(0,1) & \text{if unperturbed} \\ \text{Normal}(3,1) & \text{if perturbed by (A) or/and (B)}\end{cases}$$

$$P_2 \sim \begin{cases}\text{Normal}(0,1) & \text{if unperturbed} \\ \text{Normal}(3,1) & \text{if perturbed by (B)}\end{cases}$$

$$P_3 \sim \begin{cases}\text{Normal}(0,1) & \text{if unperturbed} \\ \text{Normal}(3,1) & \text{if perturbed by (C) or/and (D)}\end{cases}$$



(a) Low Dimension (b) High Dimension

Figure 5: DAG Structure between interventions ($A - D$) and latent space ($Z1 - Z3$). Note that interventions A & B and C & D are interacting since they have overlapping latent variables.

and so the joint probability distribution of the latent Z is

$$P_Z(z_1, z_2, z_3) = P(z_1) \cdot P(z_2) \cdot P(z_3)$$

We generate observations for each perturbation by first sampling latent variables using the joint probability distribution of Z. To generate low-dimensional observations (3 dimensional real numbers), we use a deterministic g which consists a 5-layer Multi layer perceptron (MLP) with LeakyReLU activations. To generate high-dimensional observations (images), we use the latent variables as values for the x and y coordinates for the small foreground balls which provides the perturbation and random background balls which provides the noise. Figure 6 bottom row shows examples of the generated images.

A.2.2 Model Architecture & Training

The main score estimation method used is the Denoising Diffusion Probabilistic model [Ho et al., 2020]. For the low-dimensional score function estimation, we use an conditional MLP that is loosely based on the Tab-DDPM [Kotelnikov et al., 2022] that contains several fully connected layer with ReLU activations that incorporates perturbation information as an input embedded with the timestep. The diffusion process encodes the time using the standard cos and sin encoding with 1000 timesteps. For the high-dimensional score function estimation, we use a conditional UNet architecture from Hugging Face [von Platen et al., 2024] composed of standard down-sample blocks, bottle-neck, and up-sample blocks. Perturbation information is incorporated as label projected onto an embedding space and used as input along with the pixel information. We ensure model quality by comparing the generated images with actual images to determine overall quality of the model as shown in Figure 6.

For the high-dimensional experiments, extra processing was done in order to perform KSD evaluation using a lower dimensional embedding space. The embeddings were obtained using the last layer of a classification model that uses a standard ResNet architecture from torchvision [maintainers and contributors, 2016]. In order to choose the dimensionality of the embedding space, we evaluate the KSD of score functions obtained from $\{4, 8, 16, 32\}$ embedding space. Figure 7 show that the best embedding space dimension is 4.

We also consider higher dimensions like $\{192, 384\}$ by using a standard convolutional autoencoder He et al. [2015] with the corresponding dimensions. As shown in Figure 7, the KSD does not perform well in detecting the correct interacting interventions.

For both the low- and high-dimensional cases, the models were trained for 10,000 epochs with an AdamW optimizer and learning rate 1e-3 and an MSE-loss and were trained on NVIDIA H100 GPUs.

A.2.3 Evaluation Details

To evaluate the Fisher Divergence and KSD test statistic, we used score function estimators derived from a DDPM framework. This required selecting a specific noising time step for evaluation. Given

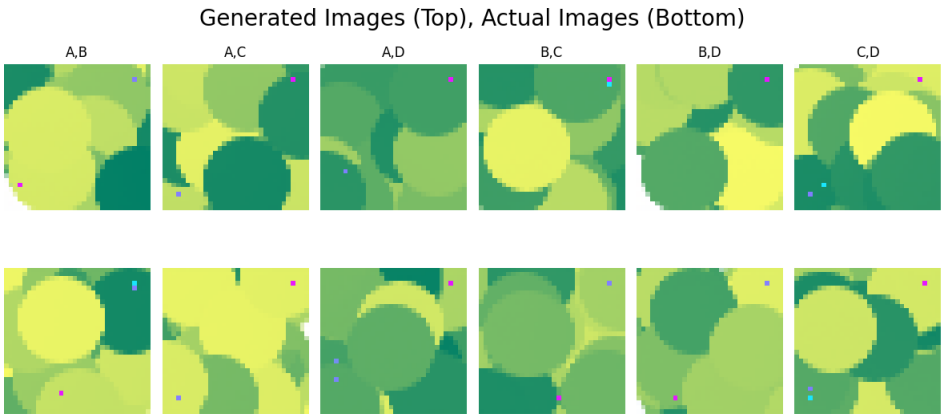
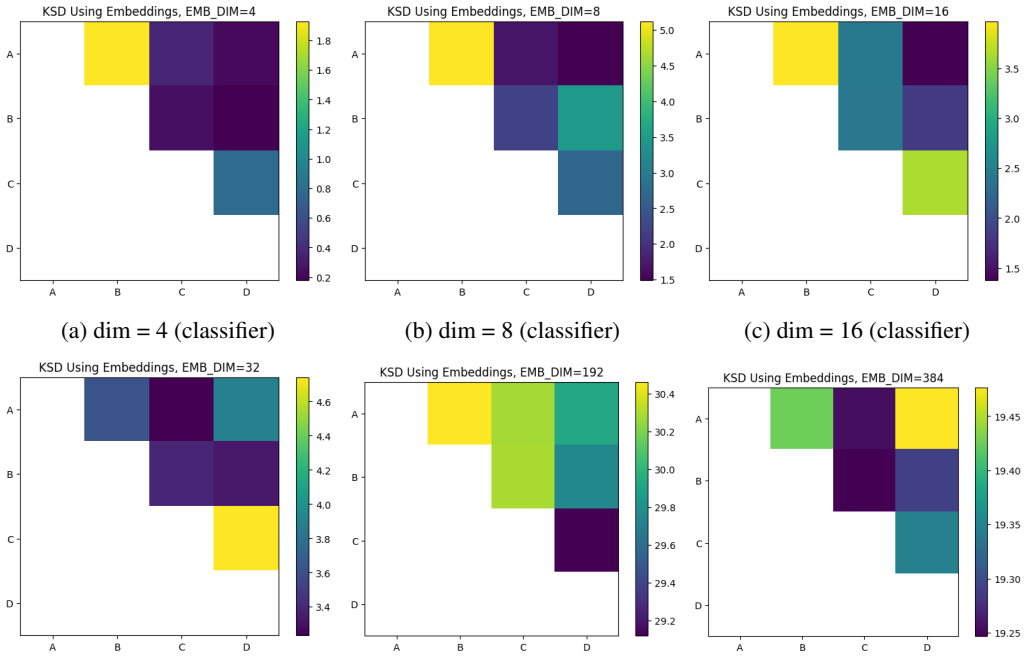


Figure 6: Generated images from Denoising Diffusion Probabilistic model trained on synthetic images. Top row contains generated images and bottom row contains actual images.

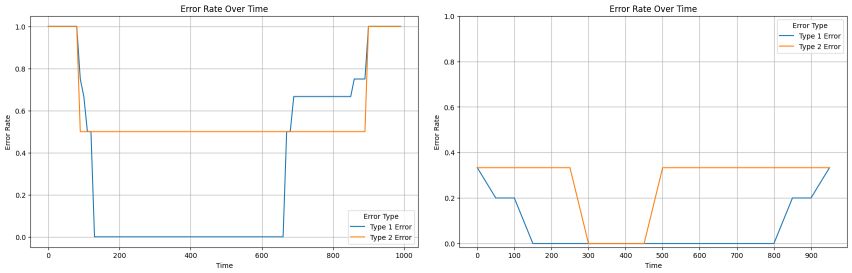


(a) dim = 4 (classifier) (b) dim = 8 (classifier) (c) dim = 16 (classifier)

(d) dim = 32 (classifier) (e) dim = 192 (autoencoder) (f) dim = 384 (autoencoder)

Figure 7: Separability values of KSD using score functions trained on various n-dimensional embedding space.

access to ground truth data, we chose the time step that empirically minimized both Type I and Type II error rates. As shown in Figure 8, the optimal time step was around $t = 400 - 500$, which we utilized in our results.



(a) Fisher Div. (b) KSD

Figure 8: Error rate of Fisher Divergence and KSD test over diffusion noise timesteps.