3D Gaussian DisCo Trials

Attempts to reproduce the first DisCo example

May 11th, 2023

*P. Chang, L. Giannini, **J. Guiang**, Y. Xiang, E. Zenhom*

Overview

- **Goal:** repeat the first example in the PRL paper (3D gaussian variables)
- **(1)** and **(2)** define the 3D gaussians
- **(3)** and **(4)** give the rest:
  - Input: X_1, X_2 (DisCo target: X_0)
  - NN architecture: 3 hidden layers; 128 nodes per layer; ReLU between layers; sigmoid output
  - λ = 1000, Adam optimizer
  - 2M sig, 2M bkg (batch size = 40K)

IV. APPLICATIONS

This section explores the efficacy of single and double DisCo in some applications of the ABCD method.

A. Simple example: Three-dimensional Gaussian random variables

We begin with a simple example to build some intuition and validate our methods. Consider a three-dimensional space (X_0, X_1, X_2), where the signal and background are both multivariate Gaussian distributions. We choose the means $\vec{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$ for background and signal as

1

$$\vec{\mu}_b = (0,0,0), \qquad \boldsymbol{\Sigma}_b = \sigma_b^2 \begin{pmatrix} 1 & \rho_b & 0 \\ \rho_b & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\sigma_b = 1.5, \qquad \rho_b = -0.8, \tag{4.1}$$

and

2

$$\vec{\mu}_s = (2.5, 2.5, 2), \qquad \boldsymbol{\Sigma}_s = \sigma_s^2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \sigma_s = 1.5. \tag{4.2}$$

3

So for the background, all three features are centered at the origin and features X_0 and X_1 are correlated with each other but independent of X_2. For the signal, all three features are independent but are centered away from the origin. The first feature X_0 will play the role of the known feature for single DisCo in Sec. III.

All of the neural networks presented in this section use three hidden layers with 128 nodes per layer. The rectified linear unit (ReLU) activation function is used for the intermediate layers and the output is a sigmoid function. A hyperparameter of $\lambda = 1000$ is used for both single and double DisCo to ensure total decorrelation. The single DisCo training converged after 100 epochs while the double DisCo training required 200 epochs. Other networks only needed ten epochs. The double DisCo networks

4

were trained using a single neural network with a two-dimensional output. All models were trained using Tensorflow [89] through Keras [90] with Adam [91] for optimization. Two million examples were generated with 15% used for testing. A batch size of 1% of the total was used for all networks to ensure an accurate calculation of the DisCo term in the relevant loss functions.

We first consider two classifiers: a baseline classifier $f_{\mathrm{BL}}(X_1, X_2)$ trained only on X_1 and X_2 and a single DisCo classifier $f_{\mathrm{SD}}(X_1, X_2)$ which includes a penalty for correlations between f_{SD} and X_0. The values of these classifiers for events drawn from the distributions are plotted in Fig. 3 against the X_0, X_1, or X_2 values of these events. We see that even though X_0 was not used in the training of the baseline, the classifier output is still correlated with X_0 because of the correlations between X_0 and X_1. In contrast to the baseline classifier, the single DisCo classifier is independent of both X_0 and X_1 and is simply a function of X_2. Intuitively, it makes sense that a classifier that must be independent of X_0 must also be independent of X_1. This is justified rigorously in Appendix B.

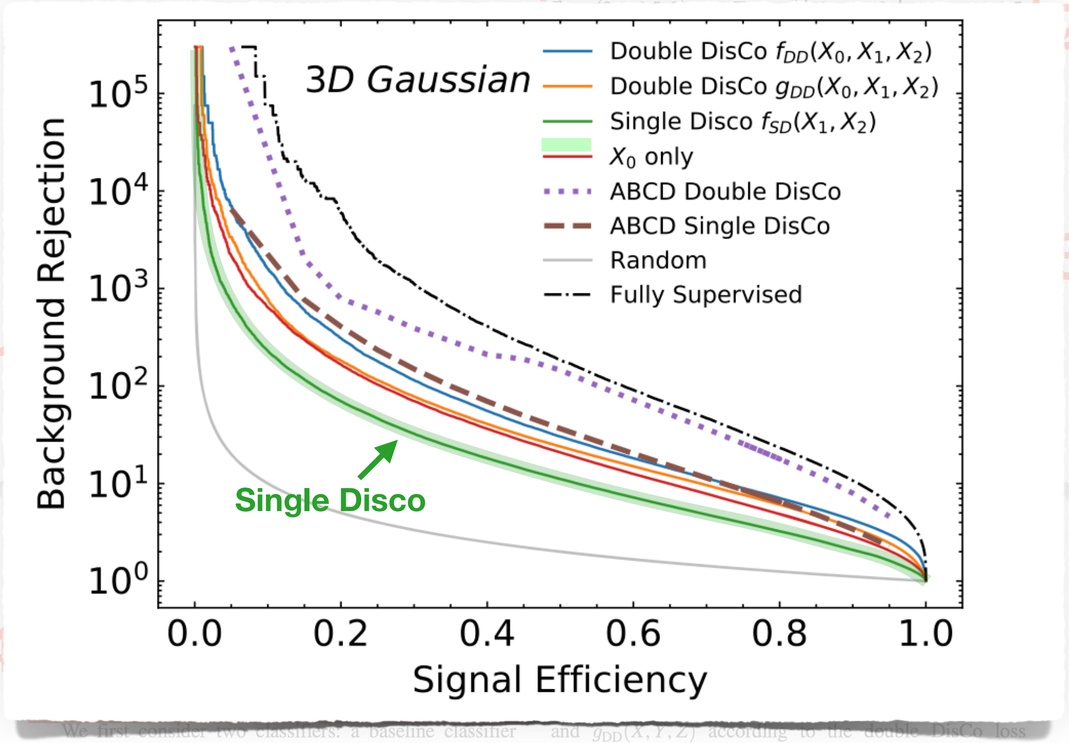
For double DisCo, we train two classifiers $f_{\mathrm{DD}}(X, Y, Z)$ and $g_{\mathrm{DD}}(X, Y, Z)$ according to the double DisCo loss function. The results are illustrated in Fig. 4. The first classifier depends mostly on Z and the second classifier depends mostly on X and Y. However, the residual dependence on all three observables is not a deficit of the training procedure: even though the three random variables are separable into two independent subsets (X, Y) and Z, the two classifiers learned by double DisCo

UC San Diego

Overview

- **Goal:** repeat the first example in the PRL paper (3D gaussian variables)
- **(1)** and **(2)** define the 3D gaussians
- **(3)** and **(4)** give the rest:
  - Input: X_1, X_2 (DisCo target: X_0)
  - NN architecture: 3 hidden layers; 128 nodes per layer; ReLU between layers; sigmoid output
  - λ = 1000, Adam optimizer
  - 2M sig, 2M bkg (batch size = 40K)

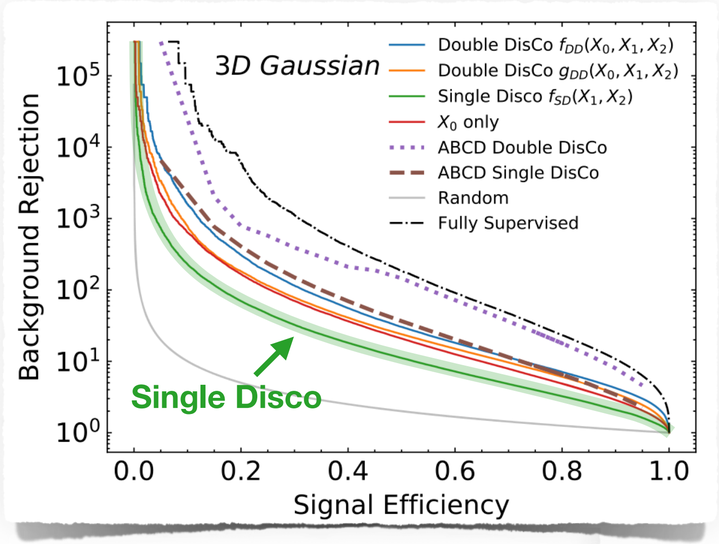
Target: recreate their plots, e.g.



TPR = TP/P = (true positives)/(positives)
FPR = FP/N = (false positives)/(negatives)

3D Gaussians: λ = 1000 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 1000 \times \text{dCorr}^2_{y=0}(f_{SD}(X_1, X_2), X_0)$$



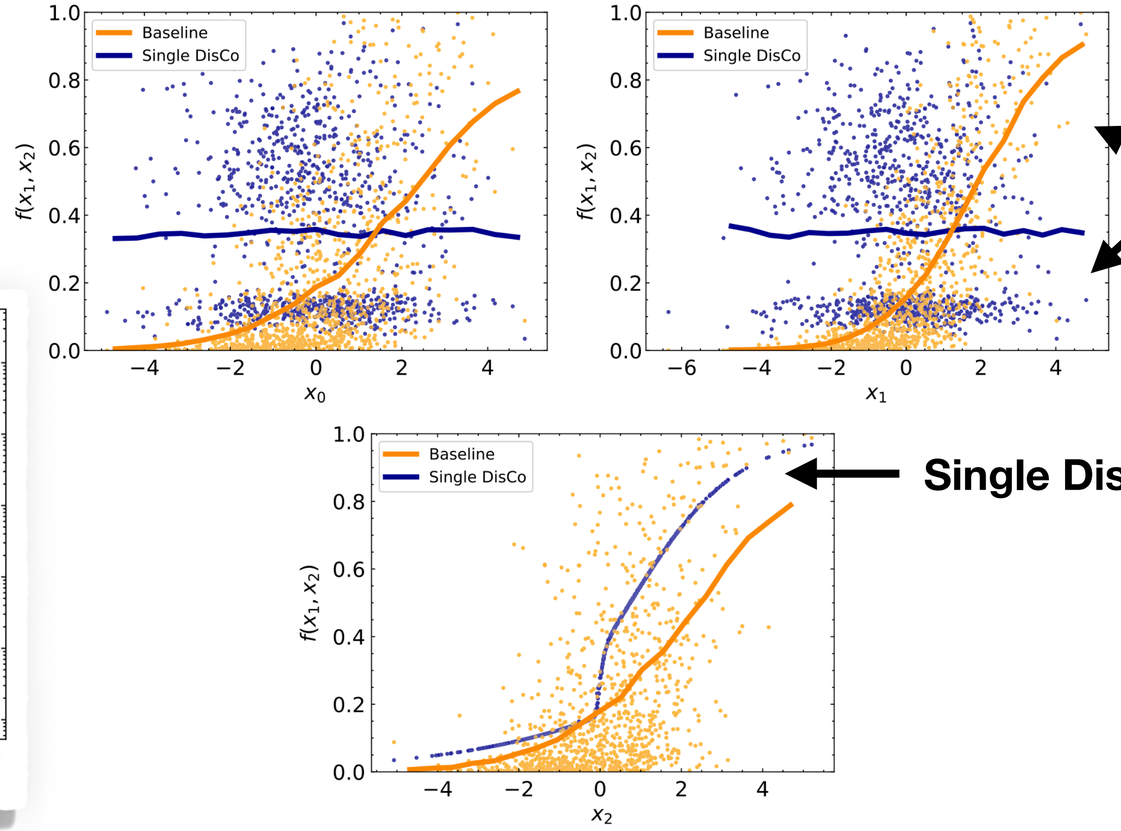
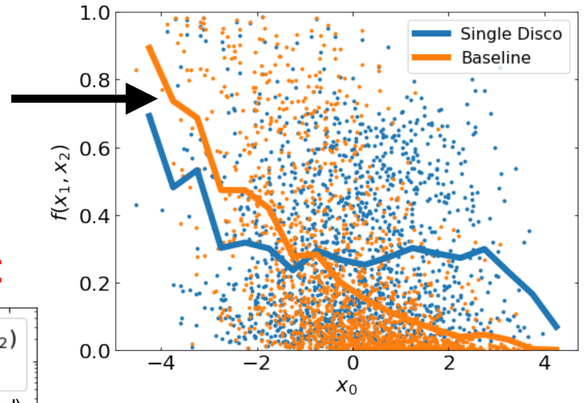


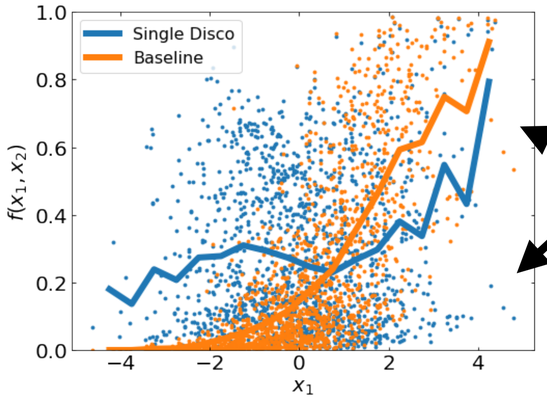
FIG. 3. Scatter plots showing the relationship (or lack thereof) between the three random variables X_0, X_1, and X_2 and (1) a baseline classifier $f_{\text{BL}}(X_1, X_2)$ trained on X_1 and X_2 with no regularization, and (2) a classifier $f_{\text{SD}}(X_1, X_2)$ trained with the single DisCo loss function that penalizes correlations with X_0. Only the background events are shown in these plots. The solid lines are the averages of the classifiers over events with the same value of X_0, X_1, or X_2. In the third panel, the scatter of the single DisCo classifier is already a line, so no average is needed.

3D Gaussians: λ = 1000 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 1000 \times \text{dCorr}^2_{y=0}(f_{SD}(X_1, X_2), X_0)$$

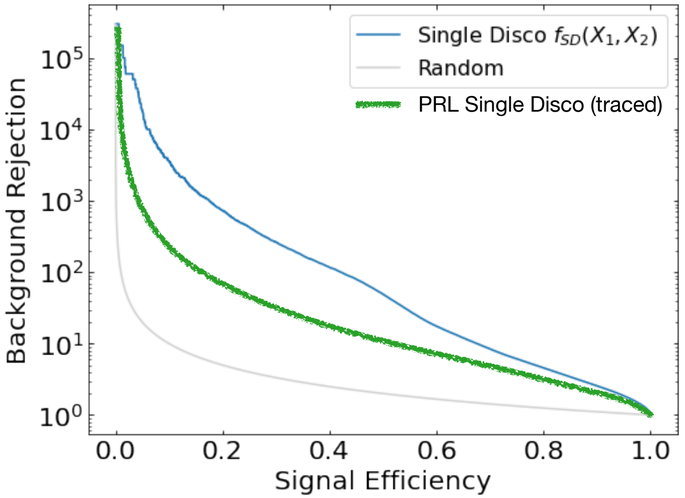
Baseline avg. opposite of that in the PRL paper

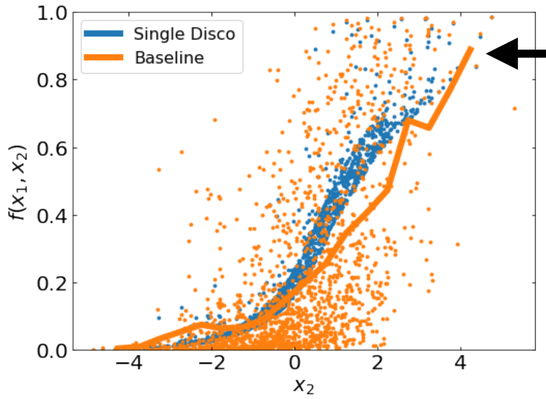




No distinct clusters

My ROC curve is different





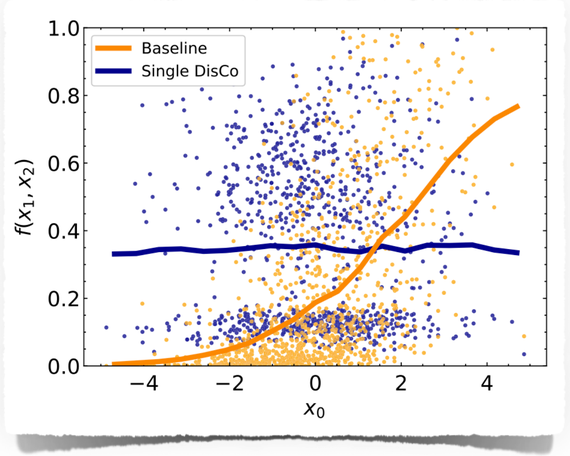
Single DisCo not a perfect line (more similar to Baseline avg.)

My plots do not match those in the PRL paper!

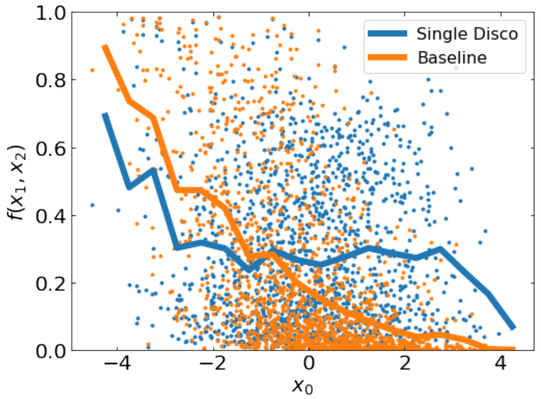
3D Gaussians: λ = 1000 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 1000 \times \text{dCorr}^2_{y=0}(f_{SD}(X_1, X_2), X_0)$$

Fig. 3

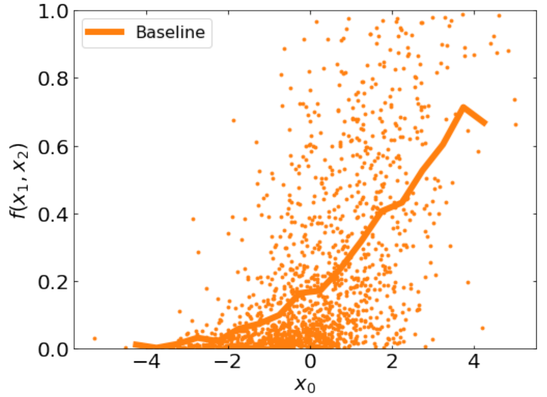


VS.



Baseline does not match!

Set ρ_b = +0.8



Baseline matches ✓

Eq. 4.1

$$\vec{\mu}_b = (0,0,0), \qquad \boldsymbol{\Sigma}_b = \sigma_b^2 \begin{pmatrix} 1 & \rho_b & 0 \\ \rho_b & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -0.8 & 0 \\ -0.8 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\sigma_b = 1.5, \qquad \rho_b = -0.8,$$

$$\Sigma_b = \sigma_b^2 \begin{pmatrix} 1 & 0.8 & 0 \\ 0.8 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

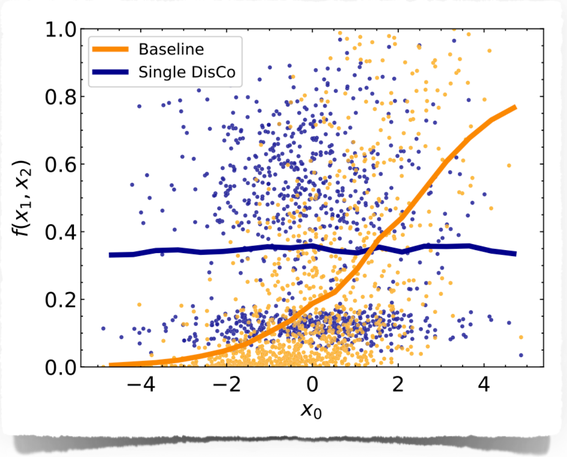
Minor note: Baseline (if not also DisCo) seems to have been trained on bkg. with ρ_b = +0.8 (not -0.8) in the covariance matrix

UC San Diego

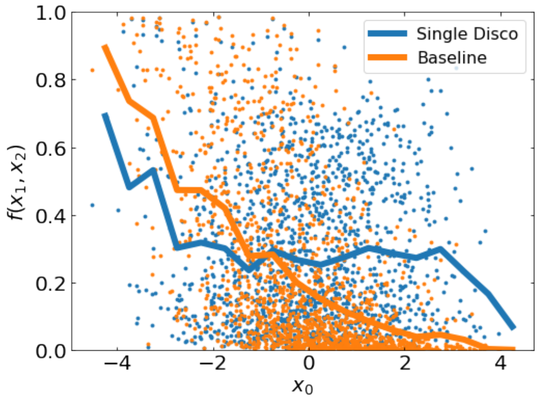
3D Gaussians: λ = 1000 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 1000 \times \mathrm{dCorr}^2_{y=0}(f_{SD}(X_1, X_2), X_0)$$

Fig. 3

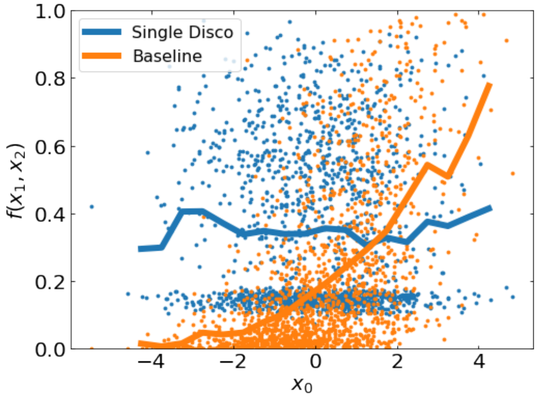


vs.



Does not match PRL

dCorr² → dCorr



Matches PRL ✓

Eq. 3.1

$$\mathcal{L}[f(X)] = \mathcal{L}_{\text{classifier}}[f(X), y] + \lambda \mathrm{dCorr}^2_{y=0}[f(X), X_0],$$

$$\lambda \mathrm{dCorr}_{y=0}[f(X), X_0]$$

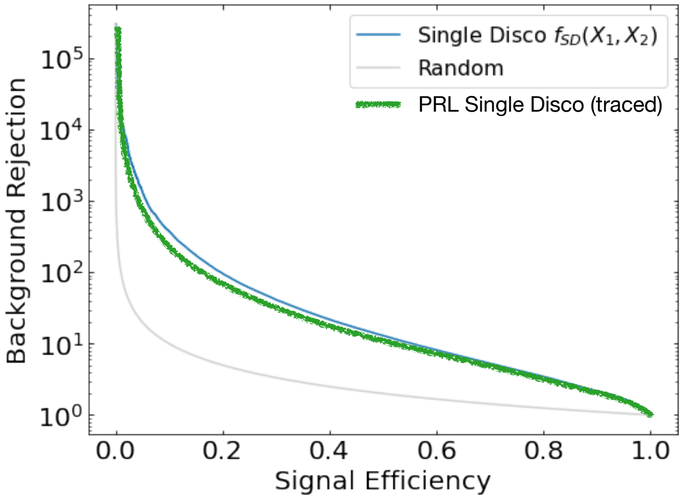
Major note: It seems that dCorr is used in the loss, not dCorr²

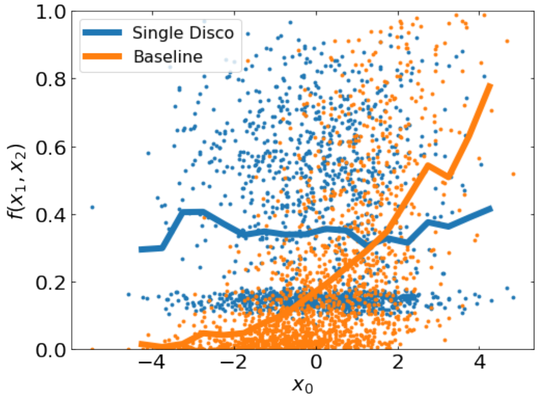
UC San Diego

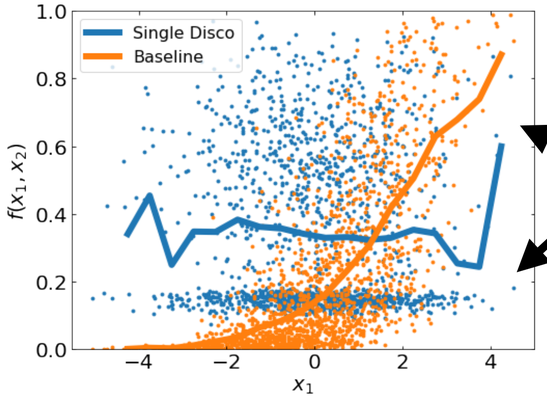
3D Gaussians: λ = 1000 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 1000 \times \text{dCorr}^1_{y=0}(f_{SD}(X_1, X_2), X_0)$$

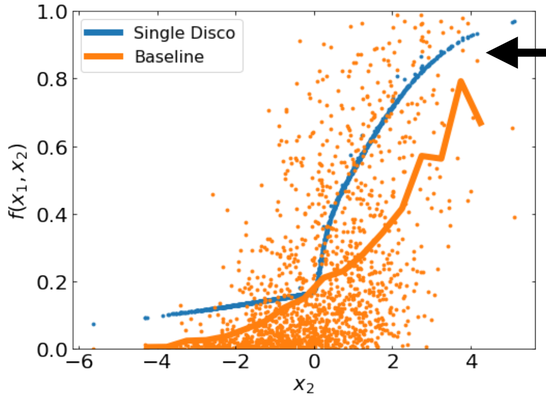
My ROC curve matches ✓







Two distinct clusters ✓



Single DisCo is a perfect line ✓

My plots match those in the PRL paper for dCorr² → dCorr

UC San Diego

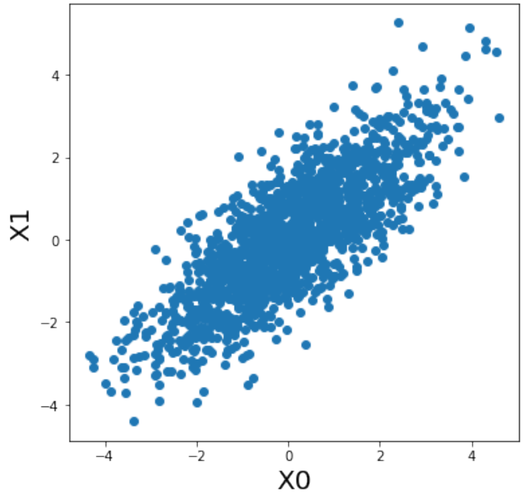
Summary

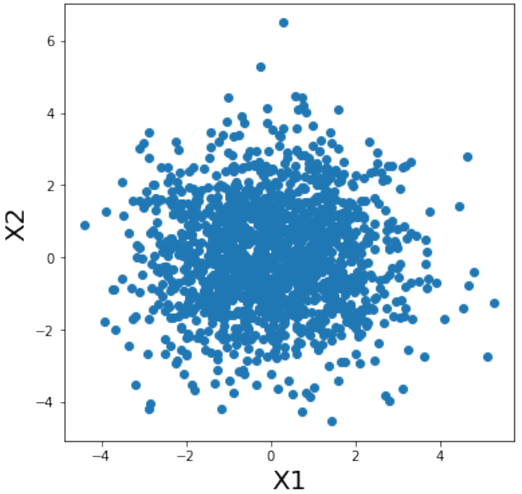
- It seems that the DisCo term in the loss should be dCorr (not $dCorr^2$)
  - From the Gaussian and top-tagging code, it seems this was indeed what was done
  - **Is this a typo, or is it $dCorr^2$ in the paper for another reason?**
- Also, there appears to be a minor typo: $\rho = 0.8$ (not -0.8) for the background Gaussian covariance matrix
- The example from the PRL paper is indeed *exactly* reproducible with the above corrections

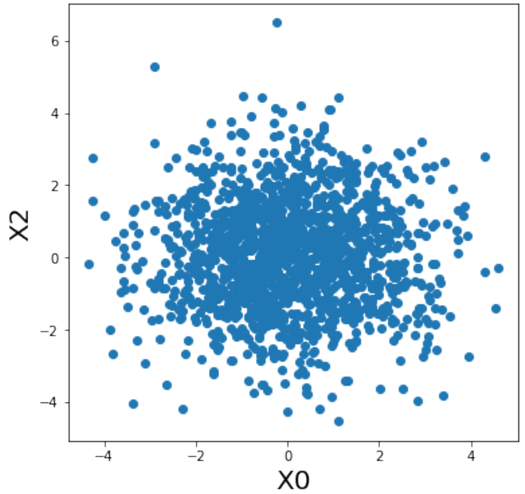
Backup

3D Gaussians

Background

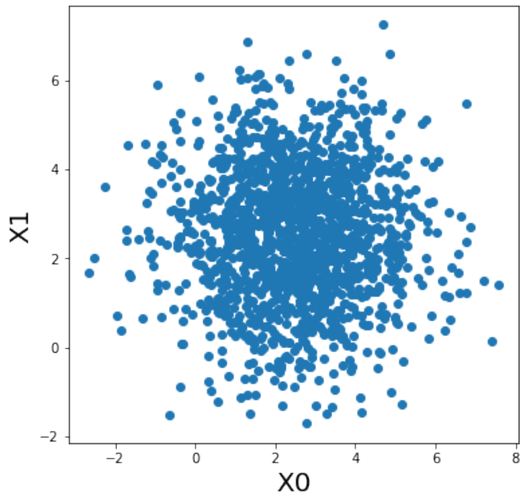


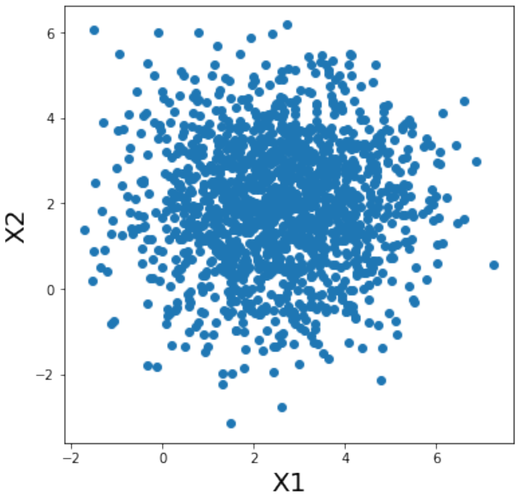


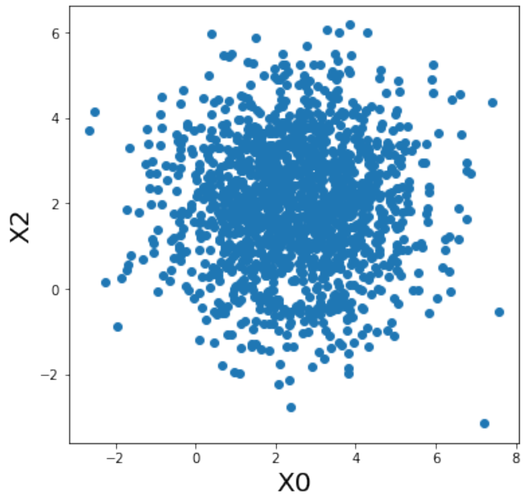


Background

Signal





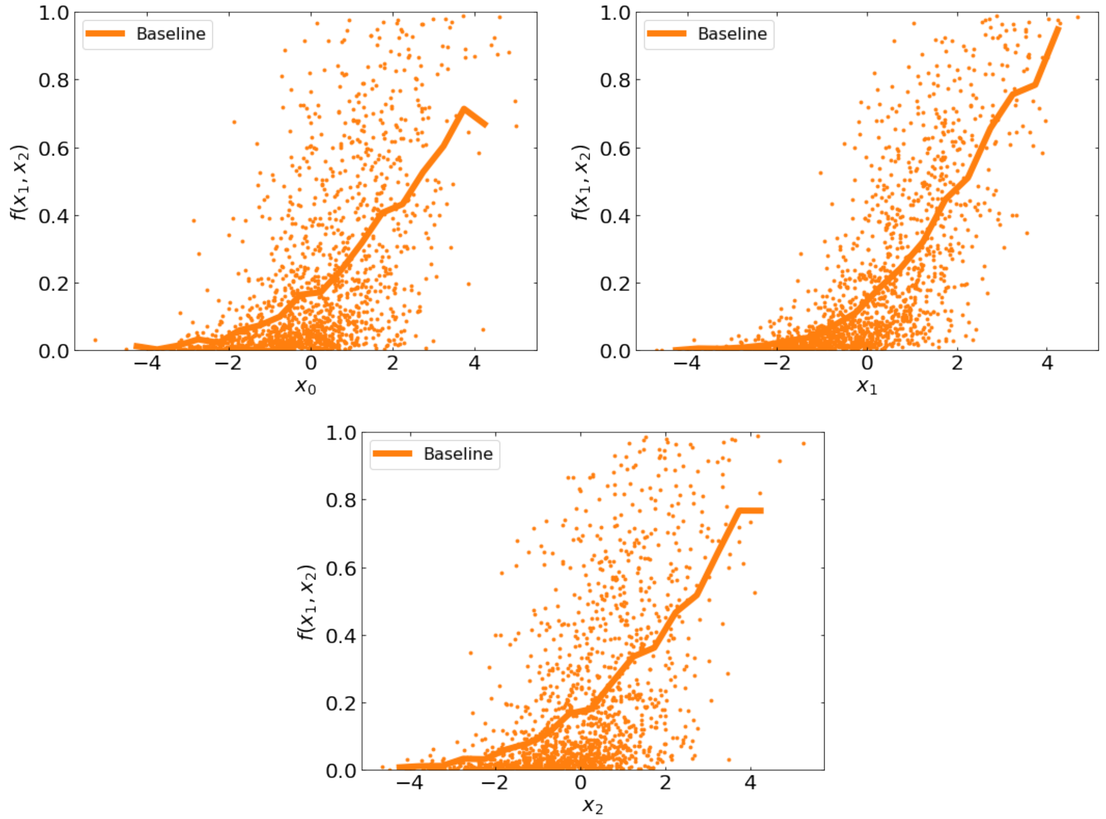


Signal

UC San Diego

3D Gaussians: Baseline

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 0 \times \mathrm{dCorr}^2_{y=0}(f_{SD}(X_1, X_2), X_0)$$



Baseline plots only match PRL plots for ρ_b = +0.8