ℓ_0-based Sparse Canonical Correlation Analysis

Ofir Lindenbaum [1] Moshe Salhov [2] Amir Averbuch [2]
Yuval Kluger[1†]

[1]Yale University, USA; [2] Tel Aviv University, Israel;

[†]Corresponding author. E-mail: yuval.kluger@yale.edu

Address: 333 Cedar St, New Haven, CT 06510, USA

June 9, 2021

Abstract

Canonical Correlation Analysis (CCA) models are powerful for studying the associations between two sets of variables. The canonically correlated representations, termed *canonical variates* are widely used in unsupervised learning to analyze unlabeled multi-modal registered datasets. Despite their success, CCA models may break (or overfit) if the number of variables in either of the modalities exceeds the number of samples. Moreover, often a significant fraction of the variables measures modality-specific information, and thus removing them is beneficial for identifying the *canonically correlated variates*. Here, we propose ℓ_0-CCA, a method for learning correlated representations based on sparse subsets of variables from two observed modalities. Sparsity is obtained by multiplying the input variables by stochastic gates, whose parameters are learned together with the CCA weights via an ℓ_0-regularized correlation loss. We further propose ℓ_0-Deep CCA for solving the problem of non-linear sparse CCA by modeling the correlated representations using deep nets. We demonstrate the efficacy of the method using several synthetic and real examples. Most notably, by gating nuisance input variables, our approach improves the extracted representations compared to other linear, non-linear and sparse CCA-based models.

1 Introduction

Canonical Correlation Analysis (CCA) [1, 2], is a classic statistical method for finding the maximally correlated linear transformations of two modalities (or views). Using modalities $\boldsymbol{X} \in \mathbb{R}^{D^x \times N}$ and $\boldsymbol{Y} \in \mathbb{R}^{D^y \times N}$, which are centered and have N samples with D^x and D^y features, respectively. CCA seeks canonical vectors $\boldsymbol{a} \in \mathbb{R}^{D^x}$, and $\boldsymbol{b} \in \mathbb{R}^{D^y}$, such that , $\boldsymbol{u} = \boldsymbol{a}^T\boldsymbol{X}$, and $\boldsymbol{v} = \boldsymbol{b}^T\boldsymbol{Y}$, will maximize the sample correlations between the *canonical variates*, i.e.

$$\max_{\boldsymbol{a},\boldsymbol{b}\neq 0} \quad \rho(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{b}^T\boldsymbol{Y}) = \frac{\boldsymbol{a}^T\boldsymbol{X}\boldsymbol{Y}^T\boldsymbol{b}}{\|\boldsymbol{a}^T\boldsymbol{X}\|_2\|\boldsymbol{b}^T\boldsymbol{Y}\|_2}. \tag{1}$$

These canonical vectors could be identified by solving a generalized eigen pair problem.

In order to identify non-linear relations between input variables, several extensions of CCA have been proposed. Kernel methods such as Kernel CCA [3], Non-parametric CCA [4] or Multi-view Diffusion maps [5, 6] learn the non-linear relations in reproducing Hilbert spaces. These

methods have several shortcomings: they are limited to a designed kernel, they require $\mathcal{O}(N^2)$ computations for training, and they have poor interpolation and extrapolation capabilities. Deep CCA [7] overcomes these limitations by learning two non-linear transformations parametrized using neural networks.

Canonical correlation models have been widely used in biology [8], neuroscience [9], medicine [10], and engineering [11], for unsupervised or semi-supervised learning. By extracting meaningful dimensionality reduced representations, CCA improves downstream tasks such as clustering, classification, or manifold learning. One key limitation of these models is that they typically require more samples than features, i.e., $N > D^x, D^y$. However, if we have more variables than samples, the estimation based on the closed-form solution of the CCA problem (in Eq. 1) breaks [12]. Moreover, in high dimensional data, often some of the variables do not measure the phenomenon that is common to both modalities (therefore are not correlated) and thus should be omitted from the transformations. For these reasons, there has been a growing interest in studying sparse CCA models.

Sparse CCA (SCCA) models seek for linear transformations which are based on a sparse subset of variables from the input modalities $\boldsymbol{X}$ and $\boldsymbol{Y}$. Sparsifying the feature space not only removes degeneracy's inherent to $N < D^x, D^y$, but also improves interpretability and prevents overfitting. To encourage sparsity of the canonical vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, several authors [13, 14] propose an ℓ_0 regularized variant of Eq. 1. However, the schemes in [13, 14] are greedy and therefore may lead to suboptimal solutions. As demonstrated by [15, 16, 17, 18, 12] replacing the ℓ_0 norm by ℓ_1 is differentiable and leads to a sparse solution to Eq. 1. However, these schemes are limited to linear transformations and may lead to shrinkage of the canonical vectors due to the ℓ_1 regularizer. There has been limited work on extending these models to sparse non-linear CCA. Specifically, there are two kernel-based extensions: two-stage kernel CCA (TSKCCA) by [19] and SCCA based on Hilbert-Schmidt Independence Criterion (SCCA-HSIC) by [20]. However, these models suffer from the same limitations as KCCA and do not scale to a high dimensional regime.

This paper presents ℓ_0-CCA, a simple yet effective method for learning correlated representation based on sparse subsets of the input variables by minimizing an ℓ_0 regularized loss. Our ℓ_0 regularization relies on a recently proposed Gaussian-based continuous relaxation of Bernoulli random variables, termed gates [21]. The gates are applied to the input features to sparsify the canonical vectors. The parameters of the gates and of the model are trained jointly via gradient descent to maximize the correlation between the representations of $\boldsymbol{X}$ and $\boldsymbol{Y}$ while simultaneously selecting only the subsets of most correlated input features. By modeling the transformations using two neural networks, our method provides a natural solution to sparse non-linear correlation analysis. We apply the proposed method to synthetic data and demonstrate that our approach can improve the estimation of the canonical vectors compared with existing sparse CCA models. Then, we use the proposed scheme on several real datasets and demonstrate that it leads to more reliable and interpretable representations than other linear and non-linear data fusion schemes.

2 Sparse CCA

The problem in Eq. 1 has a closed-form solution based on the eigendecomposition of $\boldsymbol{C}_x^{-1}\boldsymbol{C}_{xy}\boldsymbol{C}_y^{-1}\boldsymbol{C}_{yx}$ and $\boldsymbol{C}_y^{-1}\boldsymbol{C}_{yx}\boldsymbol{C}_x^{-1}\boldsymbol{C}_{xy}$, where $\boldsymbol{C}_x, \boldsymbol{C}_y$ are within view sample covariance matrices and $\boldsymbol{C}_{xy}, \boldsymbol{C}_{yx}$ are cross-view sample covariance matrices. However, if N is smaller than the number of input variables (D^x or D^y), the sample covariance may be rank deficient, and the closed-form solution becomes meaningless. To overcome this limitation, we consider the problem of sparse CCA.

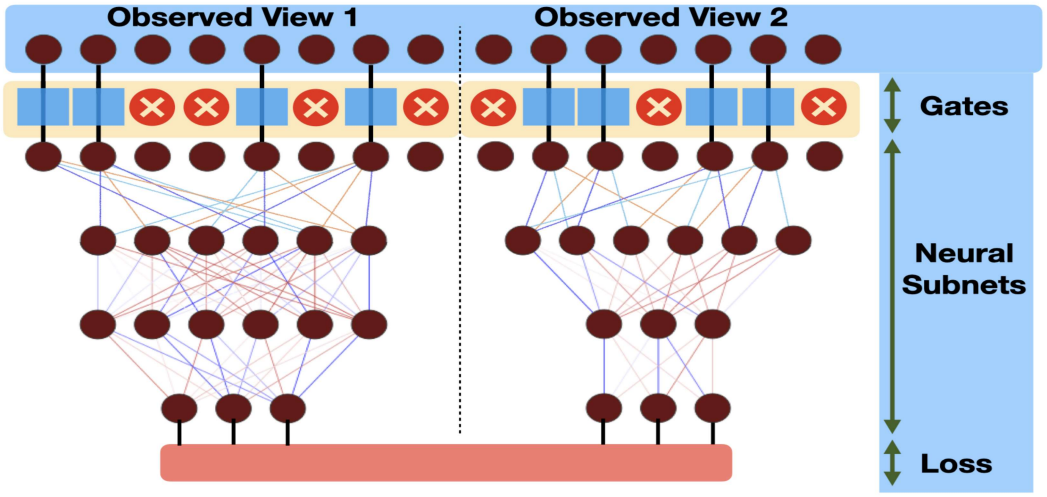

Figure 1: Illustration of ℓ_0-DCCA. Data from two views propagate through stochastic gates (defined in Eq. 4). The gates output is fed into two neural sub-nets that have a shared loss (see Eq. 5). We compute this shared loss based on the neural sub-nets outputs (with dimension $d = 3$ in this example). Our shared loss combines a total correlation term with a differentiable regularization term which induces sparsity in the input variables.

Sparse CCA deals with identifying a maximally correlated representation which is restricted to a linear combination of a sparse subset of the input variables in $\boldsymbol{X}$ and $\boldsymbol{Y}$. The problem can be formulated as the following regularized minimization

$$\min_{\boldsymbol{a},\boldsymbol{b}} \quad -\rho(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{b}^T\boldsymbol{Y}) + \lambda^x\|\boldsymbol{a}\|_0 + \lambda^y\|\boldsymbol{b}\|_0, \tag{2}$$

where λ^x and λ^y are regularization parameters which control the sparsify the input variables. If $\|\boldsymbol{a}\|_0$ and $\|\boldsymbol{b}\|_0$ are smaller than N, we can remove the degeneracy inherent to Eq. 1, and identify meaningful correlated representations based on a sparse subset of input variables. Nonetheless, for a large number of variables enumerating all the possible sparse solutions for $\boldsymbol{a}$ and $\boldsymbol{b}$ makes the bruit-force solution of the problem stated in Eq. 2 intractable.

3 Probabilistic Reformulation of Sparse CCA

The sparse CCA problem formulated in Eq. 2 becomes intractable for large D^x or D^y, due to the ℓ_0 constraint. Moreover, due to the discrete nature of the ℓ_0 norm, the problem is not amenable to gradient-based optimization schemes. Fortunately, as demonstrated in sparse regression, probabilistic models such as the spike-and-slab [22, 23, 24, 25] provide a compelling alternative. More recently, differentiable probabilistic models such as [26, 21] were proposed for sparse supervised learning. Here, we adopt these ideas by rewriting the canonical vectors as $\boldsymbol{\alpha} = \boldsymbol{\theta}^x \odot \boldsymbol{s}^x$ and $\boldsymbol{\beta} = \boldsymbol{\theta}^y \odot \boldsymbol{s}^y$, where $\odot$ denotes the Hadamard product (element wise multiplication), and $\boldsymbol{\theta}^x \in \mathbb{R}^{D^x}, \boldsymbol{\theta}^y \in \mathbb{R}^{D^y}$. The vectors $\boldsymbol{s}^x \in \{0,1\}^{D^x}$, $\boldsymbol{s}^y \in \{0,1\}^{D^y}$ are Bernoulli random vectors with parameters $\boldsymbol{\pi}^x = (\pi_1^x, ..., \pi_{D^x}^x)$ and $\boldsymbol{\pi}^y = (\pi_1^y, ..., \pi_{D^y}^y)$. These Bernoulli variables, act as gates and sparsify the coefficients of the canonical vectors. Now, the problem in Eq. 2 can be reformulated as an expectation minimization, which is parameterized by $\boldsymbol{\pi}^x$ and $\boldsymbol{\pi}^y$. Specifically, based on the following theorem, we can reformulate the problem in Eq. 2.

Theorem 3.1 *The solution of the sparse CCA problem in Eq. 2 is equivalent to the solution of the following probabilistic problem.*

$$\min_{\boldsymbol{\pi}^x, \boldsymbol{\pi}^y, \boldsymbol{\theta}^x, \boldsymbol{\theta}^y} \mathbb{E}\Big[- \rho(\boldsymbol{\alpha}^T \boldsymbol{X}, \boldsymbol{\beta}^T \boldsymbol{Y}) + \lambda^x \|\boldsymbol{s}^x\|_0 + \lambda^y \|\boldsymbol{s}^y\|_0 \Big], \tag{3}$$

where $\boldsymbol{\alpha} = \boldsymbol{\theta}^x \odot \boldsymbol{s}^x$ *and* $\boldsymbol{\beta} = \boldsymbol{\theta}^y \odot \boldsymbol{s}^y$ *and the expectation is taken with respect to the random Bernoulli variables* $\boldsymbol{s}^x$ *and* $\boldsymbol{s}^y$ *(which are parameterized by* $\boldsymbol{\pi}^x$ *and* $\boldsymbol{\pi}^y$*).*

Note that the expected values of the ℓ_0 norms can be expressed using the Bernoulli parameters as $\mathbb{E}\|\mathbf{s}^x\|_0 = \sum \pi_i^x$, and $\mathbb{E}\|\mathbf{s}^y\|_0 = \sum \pi_i^y$. The proof relies on the fact that the optimal solution to Eq. 2 is a valid solution to Eq. 3, and vice versa. The proof is presented in the Appendix, Section D, and follows the same construction as the proof of Theorem 1 in [27].

3.1 Continuous relaxation for Sparse CCA

Due to the discrete nature of $\mathbf{s}^x$ and $\mathbf{s}^y$, differentiating the leading term in Eq. 3 is not straightforward. Although solutions such as REINFORCE [28] or the straight-through estimator [29] enable differentiating through discrete random variables, they still suffer from high variance. Furthermore, they require many Monte Carlo samples for effective training [30]. Recently, several authors [31, 32, 26] have demonstrated that using a continuous reparametrization of discrete random variables can reduce the variance of the gradient estimates. Here, following the reparametrization presented in [21, 33], we use Gaussian-based relaxation for the Bernoulli random variables.

Each relaxed Bernoulli variables z_i is defined by drawing from a centered Gaussian $\epsilon_i \sim N(0, \sigma^2)$, then shifting it by μ_i and truncating its values using the following hard Sigmoid function

$$\mathrm{z}_i = \max(0, \min(1, \mu_i + \epsilon_i)). \tag{4}$$

Using these relaxed Bernoulli variables, we can define the gated canonical vectors as $\boldsymbol{\alpha} = \boldsymbol{\theta}^x \odot \mathbf{z}^x$ and $\boldsymbol{\beta} = \boldsymbol{\theta}^y \odot \mathbf{z}^y$. We incorporate these vectors into the objective defined in Eq. 3, in which the ℓ_0 regularization terms can be expressed as

$$\mathbb{E}\|\mathbf{z}^x\|_0 = \sum_{i=1}^{D^x} \mathbb{P}(\mathrm{z}_i^x \geq 0) = \sum_{i=1}^{D^x} \left(\frac{1}{2} - \frac{1}{2} \operatorname{erf}\left(-\frac{\mu_i^x}{\sqrt{2}\sigma} \right) \right),$$

where erf() is the Gaussian error function, and is defined similarly for $\mathbb{E}\|\mathbf{z}^y\|_0$.

To learn to model parameters $\boldsymbol{\theta}^x, \boldsymbol{\theta}^y$ and gate parameters $\boldsymbol{\mu}^x, \boldsymbol{\mu}^y$ we first draw realizations for the gates, then we update the parameters by applying gradient descent to minimize

$$\mathbb{E}\Big[- \rho(\boldsymbol{\alpha}^T \boldsymbol{X}, \boldsymbol{\beta}^T \boldsymbol{Y}) + \lambda^x \|\mathbf{z}^x\|_0 + \lambda^y \|\mathbf{z}^y\|_0 \Big].$$

Post training, we remove the stochasicity from the gates and use all variables such that $z_i^x = \max(0, \min(1, \mu_i^x)) > 0$ (defined similarly for $\mathbf{z}^y$).

3.2 Gate Initialization

If the gates are initialized with $\mu_i = 0.5$, they will approximate "fair" Bernoulli variables. This is a reasonable choice if no prior knowledge about the solution is available; however, we can utilize the closed-form solution of the CCA problem to derive a more suitable parameter initialization

for the gate. Specifically, given the empirical covariance matrix $\boldsymbol{C}_{xy} = \frac{\boldsymbol{X}\boldsymbol{Y}^{\boldsymbol{T}}}{(N-1)}$, we denote the thresholded covariance matrix by $\bar{\boldsymbol{C}}_{xy}$, with values defined as follows

$$(\bar{C}_{xy})_{ij} = \begin{cases} (C_{xy})_{i,j}, & \text{if } |(C_{xy})_{i,j}| > \delta \\ 0, & \text{otherwise,} \end{cases}$$

where δ is the selected threshold value based on the desired sparsity for $\boldsymbol{X}$ and $\boldsymbol{Y}$, specifically, if we assume that r percent of the values should be zeroed, then δ is set to be the r-th percentile of $|(\boldsymbol{C}_{xy})|$. Then, we compute the leading singular vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ of $\bar{\boldsymbol{C}}_{xy}$. We further threshold the absolute values of these vectors using the same percentile used for $\bar{\boldsymbol{C}}_{xy}$. The initial values of the parameters of the gates are then defined by $\boldsymbol{\mu}^x = \bar{\boldsymbol{u}} + 0.5$, and $\boldsymbol{\mu}^y = \bar{\boldsymbol{v}} + 0.5$, where $\bar{\boldsymbol{u}}$ and $\bar{\boldsymbol{v}}$ are the thresholded versions of the absolute value of the singular vectors. Using this procedure, we increase the initial probability for all gates in the support of $\bar{\boldsymbol{u}}$ and $\bar{\boldsymbol{v}}$ based on the singular vectors of $\bar{\boldsymbol{C}}_{xy}$.

3.3 ℓ_0-Deep CCA

To extend our model to non-linear function estimation, we can formulate the problem of sparse non-linear CCA by modeling the transformations using deep nets as in [7, 34]. We introduce two random Bernoulli relaxed vectors into the input layers of two neural networks trained in tandem to maximize the total correlation. Denoting the random gating vectors $\mathbf{z}^x$ and $\mathbf{z}^y$ for view $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, our ℓ_0-Deep CCA (ℓ_0-DCCA) loss is defined by

$$\mathbb{E}\Big[- \bar{\rho}(\boldsymbol{f}(\hat{\boldsymbol{X}}), \boldsymbol{g}(\hat{\boldsymbol{Y}})) + \lambda^x \|\mathbf{z}^x\|_0 + \lambda^y \|\mathbf{z}^y\|_0 \Big], \tag{5}$$

where $\boldsymbol{f}(\hat{\boldsymbol{X}}) = \boldsymbol{f}(\mathbf{z}^{\boldsymbol{x}} \odot \boldsymbol{X} | \boldsymbol{\theta}^{\boldsymbol{x}}) \in \mathbb{R}^{d \times N}$, and $\boldsymbol{g}(\hat{\boldsymbol{Y}}) = \boldsymbol{g}(\mathbf{z}^{\boldsymbol{y}} \odot \boldsymbol{Y} | \boldsymbol{\theta}^{\boldsymbol{y}}) \in \mathbb{R}^{d \times N}$ are modeled as deep networks with model parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}^{\boldsymbol{x}}, \boldsymbol{\theta}^{\boldsymbol{y}})$, and gate parameters $\boldsymbol{\mu} = (\boldsymbol{\mu}^{\boldsymbol{x}}, \boldsymbol{\mu}^{\boldsymbol{y}})$. The functions $\boldsymbol{f}$ and $\boldsymbol{g}$ embed the data into a d-dimensional space. The functional $\bar{\rho}(\boldsymbol{f}(\hat{\boldsymbol{X}}), \boldsymbol{g}(\hat{\boldsymbol{Y}}))$ measures the total correlation between the two d dimensional outputs of the deep nets, this is the sum over d correlation values computed between pairs of coordinates. Exact details on the computation of this term appear in the following subsection. The regularization parameters λ^x, λ^y control the sparsity of the input variables. The vectors $\mathbf{z}^x$ and $\mathbf{z}^y$ are Bernoulli relaxed vectors, with elements defined based on Eq. 4.

Figure. 1 highlights the proposed architecture. We first pass both observed modalities through the gates. Then, we feed these into view-specific neural sub-nets. Finally, we minimize the shared loss term in Eq. 5 by optimizing the parameters of the gates and the neural sub-nets.

3.4 Algorithm details

Denoting the centered representations for $\boldsymbol{X}, \boldsymbol{Y}$ by $\boldsymbol{\Psi}^x, \boldsymbol{\Psi}^y \in \mathbb{R}^{d \times N}$ (computed using the coupled neural sub-nets), respectively, the empirical covariance matrix between these representations can be expressed as $\widehat{\boldsymbol{C}}_{xy} = \frac{1}{N-1} \boldsymbol{\Psi}^x (\boldsymbol{\Psi}^y)^T$. Using a similar notations, we express the regularized empirical covariance matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$ as $\widehat{\boldsymbol{C}}_x = \frac{1}{N-1} \boldsymbol{\Psi}^x (\boldsymbol{\Psi}^x)^T + \gamma \boldsymbol{I}$ and $\widehat{\boldsymbol{C}}_y = \frac{1}{N-1} \boldsymbol{\Psi}^y (\boldsymbol{\Psi}^y)^T + \gamma \boldsymbol{I}$, where the matrix $\gamma \boldsymbol{I}$ ($\gamma > 0$) is added to stabilize the invertibility of $\widehat{\boldsymbol{C}}_x$ and $\widehat{\boldsymbol{C}}_y$. The total correlation in Eq. 5 (i.e. $\bar{\rho}(\boldsymbol{f}(\hat{\boldsymbol{X}}), \boldsymbol{g}(\hat{\boldsymbol{Y}}))$) can be expressed using the trace of $\widehat{\boldsymbol{C}}_y^{-1/2} \widehat{\boldsymbol{C}}_{yx} \widehat{\boldsymbol{C}}_x^{-1} \widehat{\boldsymbol{C}}_{xy} \widehat{\boldsymbol{C}}_y^{-1/2}$.

To learn the parameters of the gates $\boldsymbol{\mu}$ and of the representations $\boldsymbol{\theta}$ we apply full batch gradient decent to the loss in Eq. 5. Specifically, we use Monte Carlo sampling to estimate

Algorithm 1 ℓ_0-DCCA

Input: Coupled data, $\{\boldsymbol{X}, \boldsymbol{Y}\}$, regularization parameters λ^x, λ^y, number of epochs T.
Initialize the gate parameters: $\mu_i^x = 0.5$ for $i = 1, \ldots, D^x$, and $\mu_i^y = 0.5$ for $i = 1, \ldots, D^y$.
for $t = 1$ **to** T **do**
  Sample a stochastic gate (STG) vectors $\mathbf{z}^x, \mathbf{z}^y$ defined based on equation 4.
  Apply the STG to the data $\hat{\boldsymbol{X}} = \mathbf{z}^x \odot X$ and $\hat{\boldsymbol{Y}} = \mathbf{z}^y \odot Y$.
  Compute the loss $L = \mathbb{E}\Big[- \bar{\rho}(\boldsymbol{f}(\hat{\boldsymbol{X}}), \boldsymbol{g}(\hat{\boldsymbol{Y}})) + \lambda^x \|\mathbf{z}^x\|_0 + \lambda^y \|\mathbf{z}^y\|_0 \Big]$, where the calculation of the total correlation $\bar{\rho}$ is based on Section 3.4.
  Update $\boldsymbol{\theta}^x = \boldsymbol{\theta}^x - \gamma \nabla_{\boldsymbol{\theta}^x} L, \quad \boldsymbol{\theta}^y = \boldsymbol{\theta}^y - \gamma \nabla_{\boldsymbol{\theta}^y} L,$
    $\boldsymbol{\pi}^x = \boldsymbol{\pi}^x - \gamma \nabla_{\boldsymbol{\pi}^x} L, \quad \boldsymbol{\pi}^y = \boldsymbol{\pi}^y - \gamma \nabla_{\boldsymbol{\pi}^y} L$
end for
Return s features with largest μ_i.

the left part of Eq. 5. This is repeated for several steps (epochs), using one Monte Carlo sample between each gradient step as suggested by [26] and [21], and it worked well in our experiments. After training, we remove the stochastic part of the gates and use only the variables $i^x \in \{1, ..., D^x\}$ and $i^y \in \{1, ..., D^y\}$ such that $\mathrm{z}_{i^x}^x > 0$ and $\mathrm{z}_{i^y}^y > 0$. Empirically, we observe that our method also works well with stochastic gradient descent, as long as the batch size is not too small. Alternatively, for small batches we can use a variant of the total correlation loss as was presented in [35] or [36]. In the Appendix section C, we present a pseudo-code of ℓ_0-DCCA , and extend our formulation for a multi-modal setting (more than two views).

In Algorithm 1 we provide a pseudocode description of the proposed approach.

4 Related Work

The problem of of ℓ_0 sparse CCA was studied in [13, 14]. However, both rely on a greedy heuristic which iteratively adds non zero elements to the support of $\boldsymbol{a}$ and $\boldsymbol{b}$. We have implemented [13] and observed that it does not converge to the correct canonical vectors across our synthetic evaluation in Section 5.1. Alternatively, as proposed by [12], and ℓ_1-regularized problem can be described as

$$\begin{aligned} \boldsymbol{a}, \boldsymbol{b} = \operatorname{argmin} \Big[& - \operatorname{Cov}(\boldsymbol{a}^T \boldsymbol{X}, \boldsymbol{b}^T \boldsymbol{Y}) + \tau_1 \|\boldsymbol{a}\|_1 + \tau_2 \|\boldsymbol{b}\|_1 \Big], \\ \text{subject to} \quad & \|\boldsymbol{a}^T \boldsymbol{X}\|_2 \leq 1, \quad \|\boldsymbol{b}^T \boldsymbol{Y}\|_2 \leq 1, \end{aligned}$$

where τ_1 and τ_2 are regularization parameters. A number of variants have been proposed to this problem [15, 16, 17, 18], but they all suffer from parameter shrinkage and therefore lead to a solution which is less consistent with the correct canonical vectors (see Table 1).

5 Experimental Results

We validate the effectiveness of the proposed approach on a wide range of tasks. First, using synthetic data, we demonstrate that ℓ_0-CCA correctly identifies the canonical vectors in a challenging regime of $N \ll D^x, D^y$. Next, using a coupled video dataset, we demonstrate that ℓ_0-CCA can identify the common information from high dimensional data, and embed it into correlated, low-dimensional representations. Then, we use noisy images from MNIST and multi-channel seismic data to demonstrate that ℓ_0-DCCA finds meaningful representations of the data even in a noisy regime. Finally, we use ℓ_0-DCCA to improve cancer sub-type classification

using high dimensional genetic measurements. We use NA to denote simulations which did not converge. In all experiments validation sets are used for early stopping of the training procedure and for tuning $\lambda = \lambda^x = \lambda^y$. We refer the reader to the Appendix for a complete description of the baselines, training procedure and parameters choice for all methods.

5.1 Synthetic Example

To generate samples from $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{D \times N}$, we follow the procedure described in [12], considering data sampling from the following distribution

$$\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix} \sim N(\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_y \end{pmatrix}),$$

where $\boldsymbol{\Sigma}_{xy} = \rho_0 \boldsymbol{\Sigma}_x (\boldsymbol{\phi}\boldsymbol{\eta}^T) \boldsymbol{\Sigma}_y$. We study three cases for the covariance matrices $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_x = \boldsymbol{\Sigma}_y$.
Model I. Identity: $\boldsymbol{\Sigma} = \boldsymbol{I}_D$.
Model II. Toeplitz: $\Sigma_{i,j} = \rho_0^{|i-j|}, i, j = 1, ..., D$.
Model III. Sparse inverse matrices: $\Sigma_{i,j} = \frac{\bar{\Sigma}_{i,j}}{\sqrt{\bar{\Sigma}_{i,i}\bar{\Sigma}_{j,j}}}$,
where $\bar{\boldsymbol{\Sigma}} = \boldsymbol{\Gamma}^{-1}$, $\Gamma_{i,j} = \mathbb{1}_{i=j} + 0.5\mathbb{1}_{i=j} + 0.4\mathbb{1}_{i=j}$.

For all three cases, the vectors $\boldsymbol{\phi}, \boldsymbol{\eta} \in \mathbb{R}^D$, are sparse with 5 nonzero elements and $\rho_0 = 0.9$. The indices of the active elements are chosen randomly with values equal to $1/\sqrt{5}$. In this setting, the canonical vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ maximizing the correlation objective in Eq. 1 are $\boldsymbol{\phi}$ and $\boldsymbol{\eta}$, respectively (see Proposition 1 in [12]).

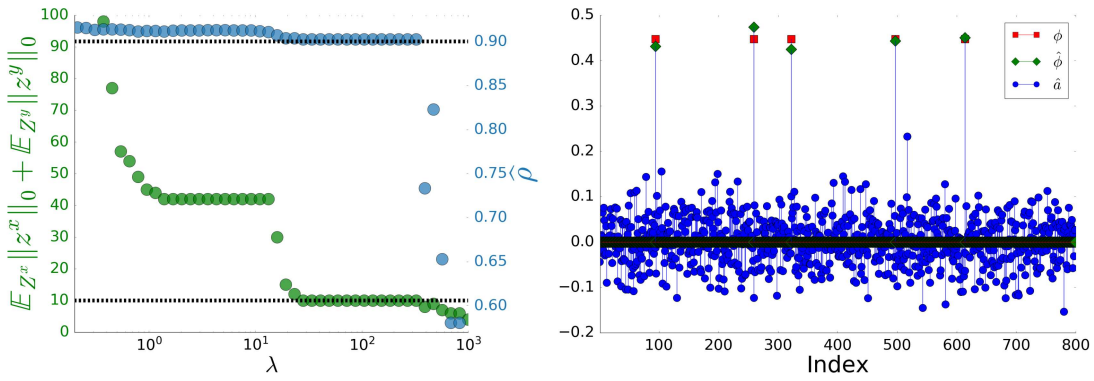

Figure 2: Left: Regularization path of ℓ_0-CCA on data generated from the linear model I (described in Section 5.1). Values on the left y-axis (green) represent the sum of active gates (by expectation). Values on the right y-axis (blue) represent the empirical correlation between the estimated representations, i.e. $\hat{\rho} = \hat{\boldsymbol{\phi}}^T \boldsymbol{X}^T \boldsymbol{Y} \hat{\boldsymbol{\eta}}$, where $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\eta}}$ are the estimated canonical vectors. Dashed lines indicate the correct number of active coefficients (10) and true correlation ρ (0.9). Note that for small values of $\lambda = \lambda^x = \lambda^y$, the model selects many variables and attains a higher correlation value; in this case, ℓ_0-CCA suffers from overfitting. Right: True canonical vector $\boldsymbol{\phi}$ along with the estimated vectors using ℓ_0-CCA ($\hat{\boldsymbol{\phi}}$) and CCA ($\hat{\boldsymbol{a}}$). Due to the small sample size, CCA overfits and fails to identify the correct canonical vectors.

Using Model I, we first generate $N = 400$ samples, with $D = 800$, and estimate the canonical vectors based on CCA and ℓ_0-CCA. In Fig. 2, we present a regularization path of the proposed scheme. Specifically, we apply ℓ_0-CCA to the data described above using various values of $\lambda = \lambda^x = \lambda^y$. We present the ℓ_0 of active gates (by expectation) along with the empirical

correlation between the extracted representations, defined by $\hat{\rho} = \rho(\hat{\boldsymbol{\phi}}^T \boldsymbol{X}, \hat{\boldsymbol{\eta}}^T \boldsymbol{Y})$. As evident from the left panel, a wide range of λ values leads to the correct number of active coefficients (10) and the correct correlation value ($\rho_0 = 0.9$). This is an indication of the robustness of the method to a wide range of λ values. Next, in the right panel we present the values of $\boldsymbol{\phi}$, the ℓ_0-CCA estimate (using $\lambda = 30$) of the canonical vector $\hat{\boldsymbol{\phi}}$, and the CCA spectral based estimate of the canonical vector $\hat{\boldsymbol{a}}$. Due to the low number of samples, the solution by CCA is wrong and not sparse, while the ℓ_0-CCA solution correctly identifies the support of $\boldsymbol{\phi}$.

	Model I- Identity covariance		
(N, D^x, D^y)	(400,800,800)	(500,600,600)	(700,1200,1200)
PMA [17]	(1.170,1.170)	(0.850,0.850)	(1.090,1.090)
IP-SCCA [37]	(1.658,1.647)	(1.051,1.051)	(1.544,1.542)
SCCA-I [18]	(1.602,1.140)	(1.143,0.282)	(1.160,0.181)
SCCA-II [38]	(0.060,0.066)	(0.053,0.057)	(0.045,0.043)
mod-SCCA [12]	(0.056,0.062)	(0.05,0.056)	(0.045,0.043)
ℓ_0-CCA	(**0.003,0.009**)	(**0.002,0.002**)	(**0.001,0.002**)
	Model II- Toeplitz covariance		
PMA [17]	(1.038,1.067)	(1.115,0.943)	(1.098,0.890)
IP-SCCA [37]	(NA,NA)	(NA,NA)	(NA,NA)
SCCA-I [18]	(1.382,1.357)	(1.351,1.299)	(1.219,1.186)
SCCA-II [38]	(0.213,0.296)	(0.145,0.109)	(0.110,0.088)
mod-SCCA [12]	(0.173,0.218)	(0.136,0.098)	(0.109,0.086)
ℓ_0-CCA	(**0.101,0.079**)	(**0.098,0.072**)	(**0.026,0.039**)
	Model III- Sparse Inverse covariance		
PMA [17]	(0.930,1.050)	(0.670,0.450)	(0.760,0.580)
IP-SCCA [37]	(0.654,0.653)	(0.092,0.091)	(0.282,0.285)
SCCA-I [18]	(1.375,0.966)	(1.041,0.502)	(0.985,0.364)
SCCA-II [38]	(0.129,0.190)	(0.069,0.062)	(0.051,0.047)
mod-SCCA [12]	(**0.092**,0.149)	(0.068,0.059)	(0.050,0.044)
ℓ_0-CCA	(0.108,**0.103**)	(**0.026,0.036**)	(**0.009,0.005**)

Table 1: Evaluating the estimation quality of the canonical vectors $\boldsymbol{\psi}$ and $\boldsymbol{\eta}$. Each pair indicates $(e_{\boldsymbol{\phi}}, e_{\boldsymbol{\eta}})$, which are the estimation errors of $\boldsymbol{\phi}$ and $\boldsymbol{\eta}$ respectively. We compare the proposed ℓ_0-CCA to other sparse CCA schemes considering three types of covariance matrices for generating $\boldsymbol{X}$ and $\boldsymbol{Y}$, and different dimensions (N, D^x, D^y). The description of all three covariances appears in Section 5.1. In each example, we highlight the smallest error obtained across all methods using boldface.

Next, we evaluate the estimation error of $\boldsymbol{\phi}$ using $e_{\boldsymbol{\phi}} = 2(1 - |\boldsymbol{\phi}^T \hat{\boldsymbol{\phi}}|)$, and $e_{\boldsymbol{\eta}}$ is defined similarly. In Table 1 we present the estimation errors of $\boldsymbol{\phi}$ and $\boldsymbol{\rho}$ (averaged over 100 simulations) for Models I, II and III (identity, Toeplitz and sparse inverse covariance matrices). As baselines, we compare the performance to 5 leading sparse CCA models. As evident from these experiments, ℓ_0-CCA significantly outperforms all baselines in its ability to learn the correct canonical vectors. In the Appendix, section B.1 we provide a runtime evaluation of the method for different values of N and D.

5.2 Multi View of Spinning Puppets

As an illustrative example, we use a dataset collected by [39] for multiview learning. The authors have generated two videos capturing rotations of 3 desk puppets. One camera captures two puppets, while the other captures another two, where one puppet is shared across cameras. A snapshot from both cameras appears in the top row of Fig. 3. All puppets are placed on spinning devices that rotate the dolls at different frequencies. In both videos, there is a shared underlying parameter, namely the rotation of the common bulldog. We use a subset of the spinning puppets dataset, with 400 images from each camera. Each image has $240 \times 320 = 76800$ pixels (using a grayscaled version of the colored image); therefore, there are more features than samples, and direct application of CCA would fail. We apply the proposed scheme using $\lambda^y = \lambda^x = 50$, a linear activation and embedding dimension $d = 2$. ℓ_0-CCA converges to embedding with a total correlation of 1.99 using 372 and 403 pixels from views $\boldsymbol{X}$ and $\boldsymbol{Y}$. The active gates are presented in the right panels of Fig. 3. In this example, the active gates highlight subsets of pixels that overlap with the common spinning Bulldog. This indicates the ability of ℓ_0-CCA to identify correlated variables in the regime of $N < D^x, D^y$.

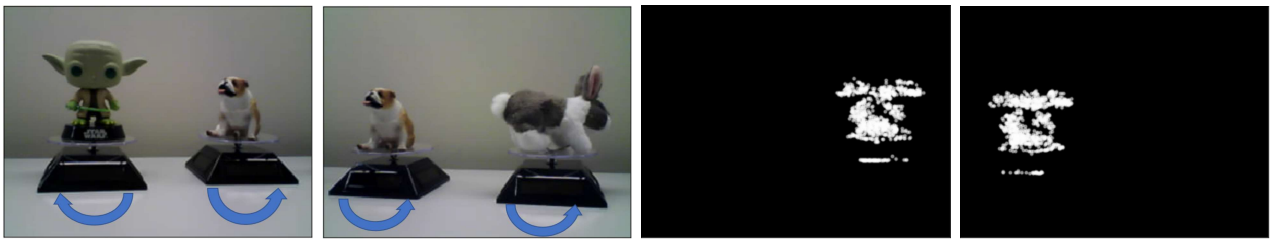

Figure 3: Left: two samples from the spinning puppets videos. The videos capture the rotation of 3 desk puppets. Arrows indicate the spinning direction of each puppet. We use ℓ_0-CCA to identify a sparse subset of pixels that are correlated from both videos. Right: the values of the gates for each video $\mathbf{z}^x$ and $\mathbf{z}^y$. After training, the values of the gates are binary (i.e., $\{0, 1\}$), with 372 and 403 active gates for the left and right videos, respectively. ℓ_0-CCA correctly select correlated subsets of pixels that highlight the common puppet (Bulldog) in this example.

In Fig. 4, we present the coupled two-dimensional embedding of both videos. Both embeddings are correlated with each other and reveal the angular orientation of the Bulldog (which is the common latent parameter in this experiment). Note that adjacent images in the embedding are not necessarily contiguous in the original ambient space because the Bunny and the Yoda puppets are correctly gated since they can not contribute to the correlated embedding.

5.3 Noisy MNIST

MNIST [40] consists of 28×28 gray-scale digit images. We use a two noisy variants of MNIST as our coupled views. The first view is created by adding noise drawn uniformly from $[0, 1]$ to all pixels. The second view is created by placing a random patch from a natural image in the background of the handwritten digits. Random samples from these modalities are presented in Fig. 5. Each view consists of $62,000$ samples, of which we use $40,000$ for training, $12,000$ for testing and $10,000$ are used as a validation set.

We train ℓ_0-DCCA to embed the data into a correlated 10 dimensional space while selecting subsets of input pixels. Our model selects 277, and 258 pixels from both modalities respectively (see bottom right corner of Fig. 5). Next, we evaluate the quality of learned embedding by

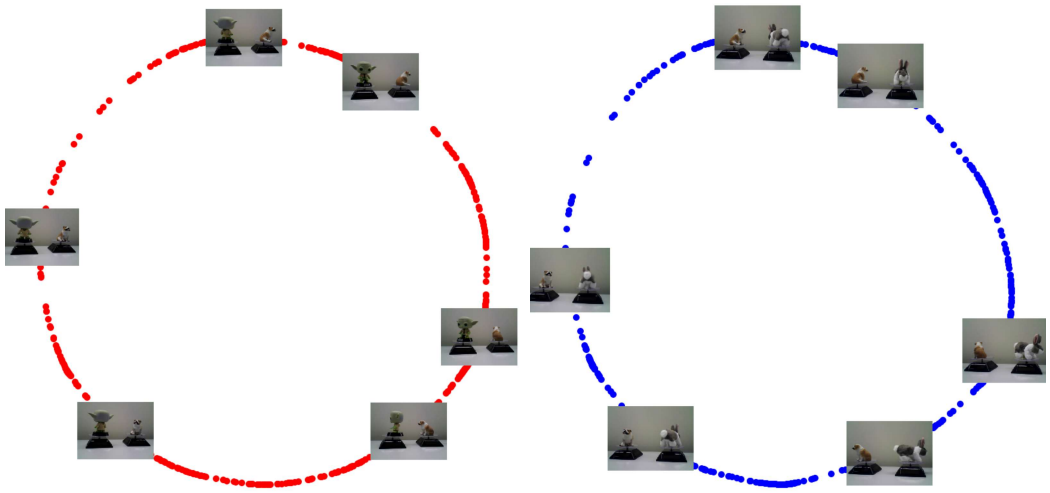

Figure 4: The correlated ℓ_0-CCA representations (visualized using [5]) of the Yoda+Bulldog video (left) and Bulldog+Bunny (right). We superimpose each embedding with 6 images corresponding to 6 points in the embedding spaces. The transformations are based on the information described by pixels whose gates are active (presented in Fig. 3). The resulting embeddings are correlated with each other, with a total correlation of $\bar{\rho} = 1.99$. The structure captured by the embeddings correctly represent the angular rotation of the Bulldog, which is the common latent parameter in this experiment.

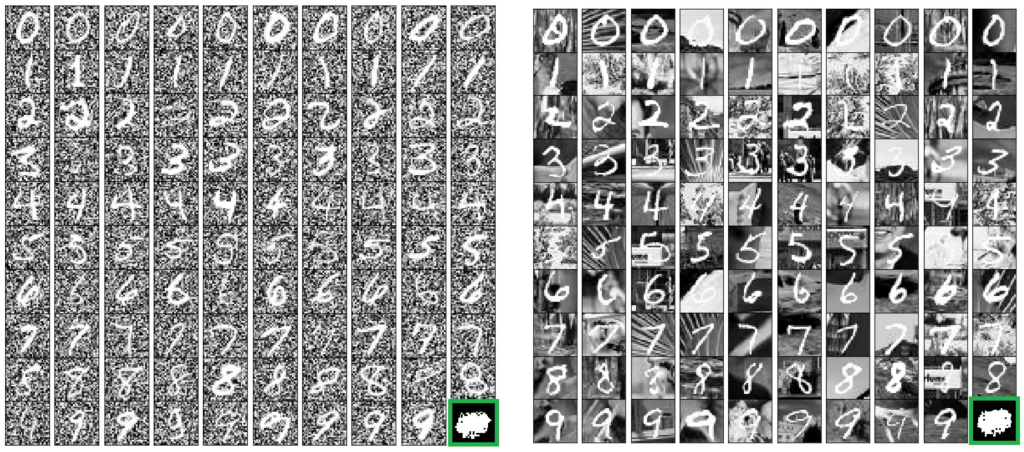

Figure 5: Images from the coupled noisy MNIST dataset. In the bottom right of both panels, we presents the active gates (white values within a green frame). There are 277 and 258 active gates for view I and II, respectively.

applying k-means to the stacked embedding of both views. We run k-means (with $k = 10$) using 20 random initializations and record the run with the smallest sum of square distances from centroids. Given the cluster assignment, k-means clustering accuracy (KM) and mutual information (MI) are measured using the true labels. Additionally, we train a Linear-SVM (SVM) model on our train and validation datasets. SVM classification accuracy is measured on the remaining test set. The embedding provided by ℓ_0-DCCA leads to higher classification and clustering results compared with several linear and non-linear modality fusion models appear

in Table 2. In the Appendix, we provide the implementation details and present experimental results demonstrating the performance of ℓ_0-DCCA for various values of $\lambda = \lambda^x = \lambda^y$.

5.4 Seismic Event Classification

Next, we evaluate the method using a dataset of seismic events studied by [41, 42]. Here, we focus on 537 explosions which are categorized into 3 quarries. Each event is recorded using two directional channels facing east (E) and north (N); these comprise the coupled views for the correlation analysis. Following the analysis by [41], the input features are sonogram representations of the seismic signal. Sonograms are time-frequency representations with bins equally tempered on a logarithmic scale. Each sonogram $\boldsymbol{z} \in \mathbb{R}^{1157}$ with 89 time bins and 13 frequency bins. An example of sonograms from both channels appears in the top row of Fig. 6.

We create the noisy seismic data by adding sonograms computed based on vehicle noise from [1]. Examples of noisy sonograms appear in the middle row of Fig. 6. We hold out 20% of the data as a validation set, and train ℓ_0-DCCA to embed the data in 3 dimensions. In Table 2 we present the MI, k-means and SVM accuracies computed based on ℓ_0-DCCA embedding. Furthermore, we compare the performance with several other baselines. Here, the proposed scheme improves performance in all 3 metrics while identifying a subset of 17 and 16 features from channel E and N, respectively. The active gates are presented in the bottom row of Fig. 6. Our results indicate that even in the presence of strong noise, ℓ_0-DCCA correctly activates the gates in frequency bins that coincide with the energy stamps of the primary and secondary waves (P and S in the top left of Fig. 6).

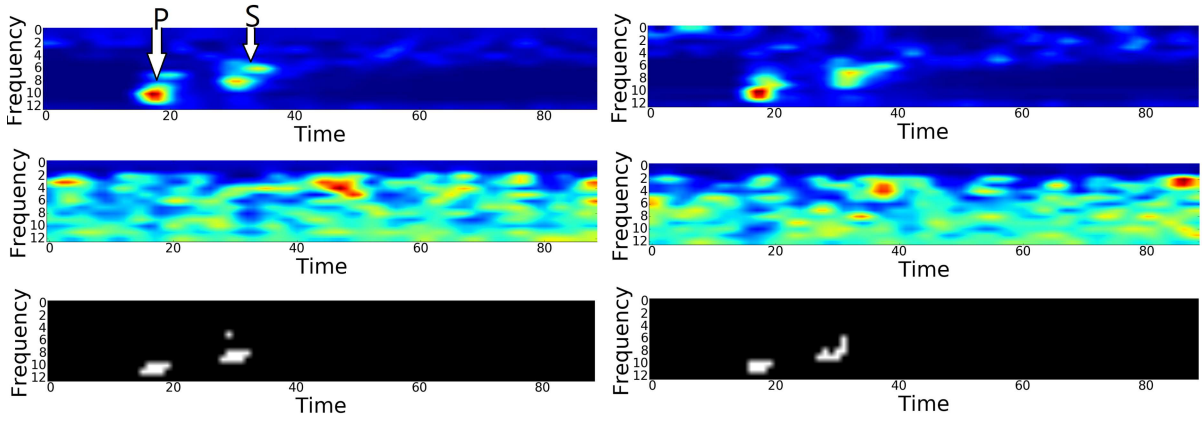

Figure 6: Top: Clean sample sonograms of an explosion based on the E and N channels (left and right, respectively). Arrows highlight the Primary (P) and Secondary (S) waves caused by the explosion. Middle: Noisy sonograms generated by adding sonograms of vehicle recordings. Bottom: the active gates for both channels. Note that the gates are active at time-frequency bins which correspond to the P and S waves (see top left figure).

5.5 Cancer Sub-Type classification

Accurate classification of cancer sub-type can be vital for extending the life span of patients by personalized treatments [43]. This task is challenging since the number of measured genes (features) is typically much larger than the number of observations. Here, we use multi-modal

[1] https://bigsoundbank.com/search?q=car

Method	Noisy MNIST [40]			Seismic [41]			METABRIC [44]		
	MI	KM (%)	SVM (%)	MI	KM (%)	SVM (%)	MI	KM (%)	SVM (%)
Raw Data	0.130	16.6	86.6	0.001	35.7	41.3	0.58	36.5	63.8
PCA [46]	0.130	16.6	89.3	0.002	38.8	41.3	0.08	19.2	23.6
CCA [47]	1.290	66.4	75.8	0.003	38.1	40.4	0.20	20.7	24.1
mod-SCCA [12]	0.342	23.9	63.1	0.610	71.7	86.9	0.12	21.0	26.0
SCCA-HSIC [20]	NA	NA	NA	0.003	38.7	49.5	NA	NA	NA
KCCA [3]	0.943	50.2	85.3	0.006	38.4	92.5	0.35	30.8	61.3
grad-KCCA [48]	NA	NA	NA	0.005	40.9	41.4	0.50	32.6	47.8
multiview-ICA [49]	1.750	88.0	90.0	0.748	90.1	94.2	0.74	44.7	62.8
NCCA [4]	1.030	47.5	77.2	0.700	86.8	91.4	0.72	48.7	63.7
DCCA [7]	1.970	93.2	93.2	0.830	94.9	94.6	0.79	45.2	72.1
DCCAE [34]	1.940	91.8	94.0	0.92	97.0	97.0	0.68	42.9	69.0
ℓ_0-DCCA	**2.05**	**95.4**	**95.5**	**0.97**	**98.1**	**97.2**	**0.88**	**50.3**	**74.1**

Table 2: Evaluation of correlated embedding extracted from the Noisy MNIST, Seismic, and METABRIC (cancer type) datasets. The representation extracted by ℓ_0-DCCA leads to higher clustering and classification accuracy compared with several baselines.

observations from the METABRIC data [44] and attempt to find correlated representations to improve cancer sub-type classification. The data consists of $1,112$ breast cancer patients which are annotated by 10 subtypes based on InClust [45]. We observe tow modalities, namely the RNA gene expression data, and Copy Number Alteration (CNA) data. The dimensions of these modalities are $15,709$ and $47,127$, respectively. We compute the ℓ_0-DCCA 10 dimensional embedding (and all baseline embeddings) and demonstrate using k-means and SVM that the representation identified using ℓ_0-DCCA can lead to more accurate cancer sub-type classification (see Table 2).

6 Conclusion

This paper presents a method for learning sparse non-linear transformations that maximize the canonical correlations between two modalities. Our approach is realized by gating the input layers of two neural networks, which are trained to optimize their output's total correlations. Input variables are gated using a regularization term which encourages sparsity. We further propose a novel scheme to initialize the gates based on a thresholded cross-covariance matrix. Our method can learn informative correlated representations even when the number of variables far exceeds the number of samples. Finally, we demonstrate that the proposed scheme outperforms existing algorithms for linear and non-linear canonical correlation analysis.

References

[1] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[2] Bruce Thompson. Canonical correlation analysis. *Encyclopedia of statistics in behavioral science*, 2005.

[3] Francis R Bach and Michael I Jordan. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48, 2002.

[4] Tomer Michaeli, Weiran Wang, and Karen Livescu. Nonparametric canonical correlation analysis. In *International Conference on Machine Learning*, pages 1967–1976, 2016.

[5] Ofir Lindenbaum, Arie Yeredor, Moshe Salhov, and Amir Averbuch. Multi-view diffusion maps. *Information Fusion*, 55:127–149, 2020.

[6] Moshe Salhov, Ofir Lindenbaum, Yariv Aizenbud, Avi Silberschatz, Yoel Shkolnisky, and Amir Averbuch. Multi-view kernel consensus for data analysis. *Applied and Computational Harmonic Analysis*, 49(1):208–228, 2020.

[7] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255, 2013.

[8] Harold Pimentel, Zhiyue Hu, and Haiyan Huang. Biclustering by sparse canonical correlation analysis. *Quantitative Biology*, 6(1):56–67, 2018.

[9] Fares Al-Shargie, Tong Boon Tang, and Masashi Kiguchi. Assessment of mental stress effects on prefrontal cortical activities using canonical correlation analysis: an fnirs-eeg study. *Biomedical optics express*, 8(5):2583–2598, 2017.

[10] Yu Zhang, Guoxu Zhou, Jing Jin, Yangsong Zhang, Xingyu Wang, and Andrzej Cichocki. Sparse bayesian multiway canonical correlation analysis for eeg pattern recognition. *Neurocomputing*, 225:103–110, 2017.

[11] Zhiwen Chen, Steven X Ding, Tao Peng, Chunhua Yang, and Weihua Gui. Fault detection for non-gaussian processes using generalized canonical correlation analysis and randomized algorithms. *IEEE Transactions on Industrial Electronics*, 65(2):1559–1567, 2017.

[12] Xiaotong Suo, Victor Minden, Bradley Nelson, Robert Tibshirani, and Michael Saunders. Sparse canonical correlation analysis. *arXiv preprint arXiv:1705.10865*, 2017.

[13] Ami Wiesel, Mark Kliger, and Alfred O Hero III. A greedy approach to sparse canonical correlation analysis. *arXiv preprint arXiv:0801.2748*, 2008.

[14] Jia Cai, Wei Dan, and Xiaowei Zhang. l0-based sparse canonical correlation analysis with application to cross-language document retrieval. *Neurocomputing*, 329:32–45, 2019.

[15] Sandra Waaijenborg, Philip C Verselewel de Witt Hamer, and Aeilko H Zwinderman. Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis. *Statistical applications in genetics and molecular biology*, 7(1), 2008.

[16] Elena Parkhomenko, David Tritchler, and Joseph Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical applications in genetics and molecular biology*, 8(1), 2009.

[17] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

[18] David R Hardoon and John Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning*, 83(3):331–353, 2011.

[19] Kosuke Yoshida, Junichiro Yoshimoto, and Kenji Doya. Sparse kernel canonical correlation analysis for discovery of nonlinear interactions in high-dimensional data. *BMC bioinformatics*, 18(1):1–11, 2017.

[20] Viivi Uurtio, Sahely Bhadra, and Juho Rousu. Sparse non-linear cca through hilbert-schmidt independence criterion. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1278–1283. IEEE, 2018.

[21] Yutaro Yamada, Ofir Lindenbaum, Sahand Negahban, and Yuval Kluger. Feature selection using stochastic gates. In *Proceedings of Machine Learning and Systems 2020*, pages 8952–8963. 2020.

[22] Edward I George and Robert E McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[23] Lynn Kuo and Bani Mallick. Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81, 1998.

[24] Mingyuan Zhou, Haojun Chen, John W Paisley, Lu Ren, Guillermo Sapiro, and Lawrence Carin. Non-parametric bayesian dictionary learning for sparse image representations. In *NIPS*, volume 9, pages 2295–2303. Citeseer, 2009.

[25] Nicholas G Polson and Lei Sun. Bayesian l 0-regularized least squares. *Applied Stochastic Models in Business and Industry*, 35(3):717–731, 2019.

[26] Christos Louizos, Max Welling, and Diederik P Kingma. Learning sparse neural networks through l_0 regularization. *arXiv preprint arXiv:1712.01312*, 2017.

[27] Mingzhang Yin, Nhat Ho, Bowei Yan, Xiaoning Qian, and Mingyuan Zhou. Probabilistic best subset selection via gradient-based optimization. *arXiv e-prints*, pages arXiv–2006, 2020.

[28] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[29] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[30] George Tucker, Andriy Mnih, Chris J Maddison, Dieterich Lawson, and Jascha Sohl-Dickstein. Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models. *arXiv preprint arXiv:1703.07370*, 2017.

[31] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

[32] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[33] Ofir Lindenbaum, Uri Shaham, Jonathan Svirsky, Erez Peterfreund, and Yuval Kluger. Let the data choose its features: Differentiable unsupervised feature selection. *arXiv preprint arXiv:2007.04728*, 2020.

[34] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International conference on machine learning*, pages 1083–1092. PMLR, 2015.

[35] Weiran Wang, Raman Arora, Karen Livescu, and Nathan Srebro. Stochastic optimization for deep cca via nonlinear orthogonal iterations. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 688–695. IEEE, 2015.

[36] Xiaobin Chang, Tao Xiang, and Timothy M Hospedales. Scalable and effective deep cca via soft decorrelation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1488–1497, 2018.

[37] Qing Mai and Xin Zhang. An iterative penalized least squares approach to sparse canonical correlation analysis. *Biometrics*, 75(3):734–744, 2019.

[38] Chao Gao, Zongming Ma, Harrison H Zhou, et al. Sparse cca: Adaptive estimation and computational barriers. *Annals of Statistics*, 45(5):2074–2101, 2017.

[39] Roy R Lederman and Ronen Talmon. Learning the geometry of common latent variables using alternating-diffusion. *Applied and Computational Harmonic Analysis*, 44(3):509–536, 2018.

[40] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[41] Ofir Lindenbaum, Yuri Bregman, Neta Rabin, and Amir Averbuch. Multiview kernels for low-dimensional modeling of seismic events. *IEEE Transactions on Geoscience and Remote Sensing*, 56(6):3300–3310, 2018.

[42] Ofir Lindenbaum, Neta Rabin, Yuri Bregman, and Amir Averbuch. Multi-channel fusion for seismic event detection and classification. In *2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, pages 1–5. IEEE, 2016.

[43] Wan Zhu, Longxiang Xie, Jianye Han, and Xiangqian Guo. The application of deep learning in cancer prognosis prediction. *Cancers*, 12(3):603, 2020.

[44] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346, 2012.

[45] Sarah-Jane Dawson, Oscar M Rueda, Samuel Aparicio, and Carlos Caldas. A new genome-driven integrated classification of breast cancer and its implications. *The EMBO journal*, 32(5):617–628, 2013.

[46] Karl Pearson. Principal components analysis. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 6(2):559, 1901.

[47] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th annual international conference on machine learning*, pages 129–136. ACM, 2009.

[48] Viivi Uurtio, Sahely Bhadra, and Juho Rousu. Large-scale sparse kernel canonical correlation analysis. In *International Conference on Machine Learning*, pages 6383–6391. PMLR, 2019.

[49] Hugo Richard, Luigi Gresele, Aapo Hyvärinen, Bertrand Thirion, Alexandre Gramfort, and Pierre Ablin. Modeling shared responses in neuroimaging studies through multiview ica. *arXiv preprint arXiv:2006.06635*, 2020.

[50] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1083–1092. JMLR.org, 2015.

[51] Paul Horst. Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, 17(4):331–347, 1961.

[52] James John McKeon. *Canonical analysis: Some relations between canonical correlation, factor analysis, discriminant function analysis, and scaling theory.* Number 13. Psychometric Society, 1966.

[53] Roderick P McDonald. A unified treatment of the weighting problem. *Psychometrika*, 33(3):351–381, 1968.

[54] J. D. Carroll. Generalization of canonical correlation analysis to three or more sets of variables. *Proceedings of the 76th annual convention of the American Psychological Association*, 3:227–228, 1968.

[55] J. R. Kettering. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 12 1971.

[56] John P Van de Geer. Linear relations amongk sets of variables. *Psychometrika*, 49(1):79–94, 1984.

[57] Arthur Tenenhaus and Michel Tenenhaus. Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2):257, 2011.

[58] Arthur Tenenhaus, Cathy Philippe, and Vincent Frouin. Kernel generalized canonical correlation analysis. *Computational Statistics and Data Analysis*, 90:114–131, 2015.

[59] Arthur Tenenhaus, Cathy Philippe, Vincent Guillemot, Kim-Anh Le Cao, Jacques Grill, and Vincent Frouin. Variable selection for generalized canonical correlation analysis. *Biostatistics*, 15(3):569–583, 02 2014.

[60] Adrian Benton, Huda Khayrallah, Biman Gujral, Dee Ann Reisinger, Sheng Zhang, and Raman Arora. Deep generalized canonical correlation analysis. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 1–6, 2019.

Appendix

A Additional Experimental Details

In the following sections we provide additional experimental details required for reproduction of the experiments provided in the main text. All the experiments were conducted using Intel(R) Xeon(R) CPU E5-2620 v3 @2.4Ghz x2 (12 cores total).

A.1 Synthetic Example

For the linear model we use a learning rate of 0.005 with $10,000$ epochs. The values of λ^x and λ^y are both set to 30. These values were obtained using a cross validation procedure. The standard deviation σ of the injected noise was set to 0.5 by [21]. They have selected this value as it maximized the gradient of the regularization term at initialization. Empirically, we observe that for ℓ^0-CCA smaller values of σ translated to improved convergence. Specifically, we used $\sigma = 0.25$, which worked well in our experiments.

For each covaraince model, we run the method 100 times with different realizations of $\boldsymbol{X}$ and $\boldsymbol{Y}$. In Table 1, we compare the average results of ℓ^0-CCA to PMA [17], IP-SCCA [37], SCCA-I [18],SCCA-II [38], and mod-SCCA [12]. Results of SCCA-II, PMA, and mod-SCCA were simulated by [12].

A.2 Noisy MNIST

In this subsection we provide additional details regarding the noisy MNIST experiment. In Fig. 7, we present the performance as a function of the number of active gates (pixels) controlled by $\lambda^x = \lambda^y$. The Mutual Information (MI) score, k-means, and linear Support Vector Machine (LSVM) accuracies were computed based on ℓ^0-DCCA embedding with learning rate of 0.01. Furthermore, the number of epochs (~ 4000) was tuned by early stopping using a validation set of size 10000. To learn 10 dimensional correlated embedding, we use the same architecture as suggested by [50] consisting of three hidden layers with 1000 neurons each. The number of dimensions in the embedding was selected based on the number of classes in MNIST. This architecture is used for ℓ^0-DCCA, DCCA [7] and DCCAE [34]. Note that for ℓ^0-DCCA using small values of the regularization parameters λ^x and λ^y, increases the number of selected features and degrades the performance. This is duo to the fact that as more features are selected more noise is introduced into the extracted representation (of size 10). It is interesting to note that the k-means was more robust to the introduced noise than the LSVM.

The regularization parameters λ^x and λ^y balance between the correlation loss and the amount of sparsification performed by the gates. These hyper parameters are tuned using the validation set by maximizing the total correlation value. We compare ℓ^0-DCCA to CCA [47][2], mod-SCCA [12], SCCA-HSIC [20], KCCA [3][3], NCCA [4] [4], multiview-ICA [49], DCCA [7] [5] and DCCAE [34]. For all methods we use an embedding with dimension 10, and evaluate performance with k-means using 20 random initilizations, and using LSVM by performing training on the training samples and testing on the remaining samples (split defined in the main text). The hyperparameters of all methods are tuned to maximize the total correlation on a validation

[2]https://scikit-learn.org/stable/modules/generated/sklearn.cross_decomposition.CCA.html
[3]https://gist.github.com/yuyay/16ce4914683da30f87d0
[4]https://tomer.net.technion.ac.il/files/2017/08/NCCAcode_v3.zip
[5]https://github.com/adrianna1211/DeepCCA_tensorflow

set. In this experiment we tried to train SCCA-HSIC [20] [6] for over two days, but it did not converge. Furthermore, we believe that the poor performance of the kernel methods stems from the high level of noise in the input images. We note that grad-KCCA [48] did not converge in this experiment, therefore we did not report its performance on MNIST.

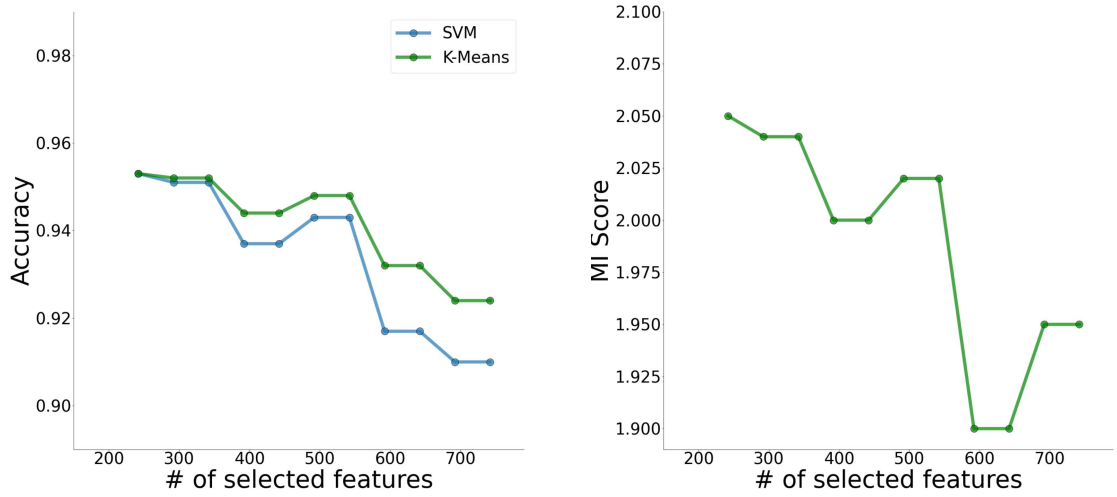

Figure 7: k-means and SVM classification accuracy (left) and mutual information score (right) vs. the number of selected features.

A.3 Seismic Event Classification

Using the seismic data, we compare the performance of ℓ^0-DCCA with a linear and non-linear activation. In this example, we use a learning rate of 0.01 with 2000 epochs. The number of neurons for the five hidden layer are: $300, 200, 100, 50$, and 40 respectively, with a tanh activation after each layer. The number of dimensions in the embedding ($d = 3$) was selected based on the number of classes in the data. Parameters are optimized manually to maximize the correlation on a validation set. In Fig. 8, we present SVM accuracy for different levels of sparsity. The presented number of features is the average over both modalities, and SVM performance is evaluated using 5-folds cross validation. We compare ℓ^0-DCCA to CCA [47], mod-SCCA [12], SCCA-HSIC [20], KCCA [3], NCCA [4], multiview-ICA [49], DCCA [7], and DCCAE [34]. For all methods we use an embedding with dimension $d = 3$, and evaluate performance with k-means using 20 random initilizations, and using linear SVM by performing a 5-folds cross validation. For the kernel methods we evaluated performance by constructing a kernel using $k = 5, 10, ..., 50$, nearest neighbors and selected the value which maximized performance in terms of total correlation.

B Cancer Sub-type Classification

In the main text, we demonstrated that the embedding extracted with ℓ_0-DCCA leads to more accurate cancer sub-type classification. Here, we provide additional details on this experiment.

[6] https://github.com/aalto-ics-kepaco/scca-hsic

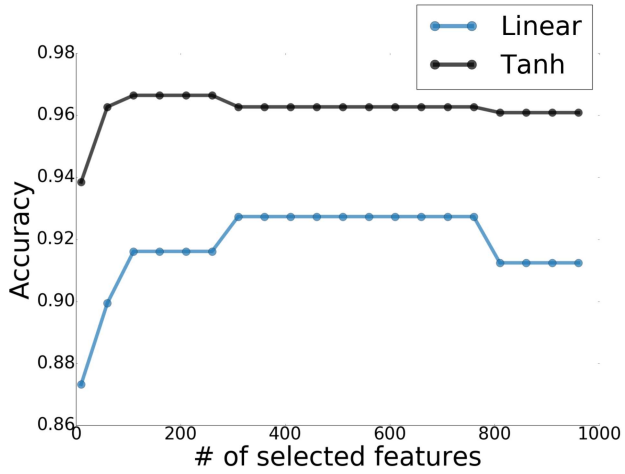

Figure 8: Classification accuracy on the noisy seismic data. Performance is evaluated using linear SVM in the 3 dimensional embedding. Comparing performance of ℓ^0-DCCA for different levels of sparsity, and using linear and nonlinear activation (tanh).

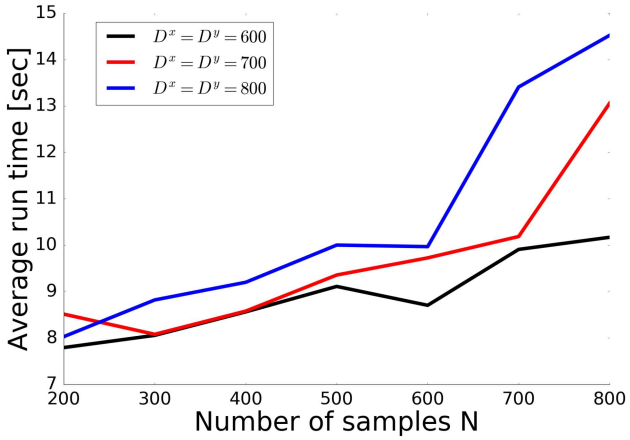

Figure 9: Run time evaluation of the proposed approach. We measure the average runtime (over 100 runs) for different values of N and D^x, D^y.

In this example, we use a learning rate of 0.5 with 2000 epochs. The number of neurons for the 3 hidden layers are: $500, 300, 100$, with a tanh activation after each layer. The number of dimensions in the embedding ($d = 10$) was selected based on the number of classes in the data. Parameters are optimized manually to maximize the correlation on a validation set. We compare ℓ^0-DCCA to CCA [47], mod-SCCA [12], SCCA-HSIC [20], KCCA [3], NCCA [4], multiview-ICA [49], DCCA [7], and DCCAE [34]. We use the same SVM and k-means scheme as described in Section A.3.

B.1 Run Time Analysis

To demonstrate the computational efficiency of our method, we run ℓ_0-CCA for different values of N and D and evaluate the empirical computational complexity of the method. In Fig. 9 we present the average runtime over 100 runs, the data is generates following Model I from Section 5.1.

C Generalized ℓ_0-based Sparse CCA

The proposed ℓ_0-CCA algorithm was designed and demonstrated on coupled datasets. However, in many applications, one may have access to multiple (>2) datasets. There are many real-life cases where more than two modalities are available such as hyper-spectral imaging, medical imaging and more. The problem of CCA was generalized decades ago to include multiple data sets by [51, 52, 53] and [54] to name some. These approaches were later summarized in [55]. Since then, many extensions have been proposed such as [56]. Some of the modern extensions of multi-view CCA comprise its regularised [57], kernelised [58] sparse [59] and deep [60] variants. In this section, we extend the proposed ℓ_0-DCCA to multi-view datasets.

Let K be the number of views and $\boldsymbol{X}^k$ the corresponding k-th view $1 \leq k \leq K$. The Deep Generalized CCA that was proposed in [60] aims to solve

$$\min_{\boldsymbol{G}\in\mathcal{R}^{n\times d},\boldsymbol{U}_k\in\mathcal{R}^{d\times d}} \sum_{k=1}^{K} \|\boldsymbol{G}-\boldsymbol{U}_k^T f(\boldsymbol{X}^k)\|_F \quad s.t. \quad \boldsymbol{G}\boldsymbol{G}^T=\boldsymbol{I}, \tag{6}$$

where $\boldsymbol{G}$ is the desired shared representation, and $\boldsymbol{U}_k$ is a linear transformation from the k-th deep network output, $f(\boldsymbol{X}^k)$, to the shared representation space.

While the proposed method in Section 3.3 describes a solution for coupled views, this solution can be naturally extended to include multiple views by introducing a shared common space into the optimization problem in Eq 6. Denoting the random gating vectors $\mathbf{z}^k$ for view $\boldsymbol{X}^k$. The generalized ℓ_0-DCCA loss is defined by

$$\min_{\boldsymbol{G}\in\mathcal{R}^{n\times d},\boldsymbol{U}_k\in\mathcal{R}^{d\times d},\boldsymbol{z}^k\in\mathcal{R}^{d_k}} \sum_{k=1}^{K} \|\boldsymbol{G}-\boldsymbol{U}_k^T f(\hat{\boldsymbol{X}}^k)\|_F + \lambda_k\|\boldsymbol{z}^k\|_0 \quad s.t. \quad \boldsymbol{G}\boldsymbol{G}^T=\boldsymbol{I}, \tag{7}$$

where $\boldsymbol{f}(\hat{\boldsymbol{X}}^{\boldsymbol{k}}) = \boldsymbol{f}(\mathbf{z}^{\boldsymbol{k}} \odot \boldsymbol{X}^{\boldsymbol{k}}|\boldsymbol{\theta}^{\boldsymbol{x}}) \in \mathbb{R}^{d\times N}$ and λ_k is the regularization parameter that control the sparsity of k-th selected gate vector $\boldsymbol{z}^k$. The optimization problem in Eq 7 can be solved using standard optimizers such as gradient descent. The initialization of $\boldsymbol{G}$ and $\boldsymbol{U}_k$ plus the analysis of the above suggested generalized gated canonical correlation are left for further research.

D Proof of Theorem 1

We first note that the equivalence of the solutions means that if we denote the solution to Eq. 2 by $(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}})$, then there is a corresponding set of values $\boldsymbol{\pi}^x, \boldsymbol{\pi}^y, \boldsymbol{\theta}^x, \boldsymbol{\theta}^y$, such that $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}})$, where $\boldsymbol{\alpha} = \boldsymbol{\theta}^x \odot \boldsymbol{s}^x$ and $\boldsymbol{\beta} = \boldsymbol{\theta}^y \odot \boldsymbol{s}^y$. The proof follows the same construction as the proof of Theorem 1 in [27]. To prove the Theorem, we will show that the optimal solution to the deterministic sparse CCA problem (in Eq. 2) is a valid solution to the probabilistic sparse CCA problem (in Eq. 3) and vice versa.

First, we denote by $(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}})$ the optimal solution of the problem defined in Eq. 2, then by setting $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{a}} \odot \hat{\boldsymbol{s}}^x$ and $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{b}} \odot \hat{\boldsymbol{s}}^y$, where $\hat{\pi}_i^x = 1$ if $\hat{a}_i \neq 0$ and $\pi_i^x = 0$ otherwise, we get a feasible solution to Eq. 3 which leads to the same objective value.

Now, we want to show that the optimal solution to Eq. 3 is also a feasible solution to Eq. 2. Denoting the optimal solution to Eq. 3 by $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, with $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\theta}}^x \odot \hat{\boldsymbol{s}}^x$ and $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\theta}}^y \odot \hat{\boldsymbol{s}}^y$, if $p(\boldsymbol{s}^x|\boldsymbol{\pi}^x)$ or $p(\boldsymbol{s}^y|\boldsymbol{\pi}^y)$ are point mass densities (i.e., $\boldsymbol{\pi}^x, \boldsymbol{\pi}^y \in \{0,1\}^D$), then the solution is effectively deterministic and $\hat{\boldsymbol{a}} = \hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{b}} = \hat{\boldsymbol{\beta}}$ would be a valid solution to problem Eq. 2. If $p(\boldsymbol{s}^x|\boldsymbol{\pi}^x)$ and $p(\boldsymbol{s}^y|\boldsymbol{\pi}^y)$ are not point mass densities, we will now show by contradiction that all the points in

the support of $p(\boldsymbol{s}^x|\boldsymbol{\pi}^x)$ (or $p(\boldsymbol{s}^y|\boldsymbol{\pi}^y)$) would lead to the same objective values. Assume without loss of generality that $\{\boldsymbol{s}_1^x, ..., \boldsymbol{s}_L^x\} \in \text{supp}[p(\boldsymbol{s}^x|\hat{\boldsymbol{\pi}}^x)]$, if not all values in the support lead to the same objective value, this means that there exist $\boldsymbol{s}_l^x, \boldsymbol{s}_m^x$ such that $L(\boldsymbol{s}_l^x) < L(\boldsymbol{s}_m^x)$, where

$$L(\boldsymbol{s}^x) = \Big[-\rho((\hat{\boldsymbol{\theta}}^{\boldsymbol{x}} \odot \boldsymbol{s}^{\boldsymbol{x}})^T \boldsymbol{X}, (\hat{\boldsymbol{\theta}}^{\boldsymbol{y}} \odot \hat{\boldsymbol{s}}^{\boldsymbol{y}})^T \boldsymbol{Y}) + \lambda^x \|\boldsymbol{s}^x\|_0 + \lambda^y \|\hat{\boldsymbol{s}}^y\|_0 \Big],$$

which is the objective from Eq. 3 with optimal values of $\hat{\boldsymbol{\theta}}^x, \hat{\boldsymbol{\theta}}^y, \hat{\boldsymbol{s}}^y$. Now if we set $\bar{\boldsymbol{\pi}}^x = \boldsymbol{s}_l^x$ we would obtain $\mathbb{E}_{\boldsymbol{s}^x \sim p(\boldsymbol{s}^x|\bar{\boldsymbol{\pi}}^x)} L(\boldsymbol{s}^x) < \mathbb{E}_{\boldsymbol{s}^x \sim p(\boldsymbol{s}^x|\hat{\boldsymbol{\pi}}^x)} L(\boldsymbol{s}^x)$ which contradicts the assumption that $\hat{\boldsymbol{\pi}}^x$ is optimal. This means that all points in $\text{supp}[p(\boldsymbol{s}^x|\hat{\boldsymbol{\pi}}^x)]$ lead to the same objective value, therefore, they are also feasible solutions to Eq. 2. Now, showing that the solution to Eq. 3 is a feasible solution to Eq. 2 and vice versa completes our proof.

E Strengths and Limitations

The proposed method provides an effective solution to the problems of sparse linear and nonlinear CCA. One advantage of the suggested method compared with existing sparse CCA solutions, is that it can embed data into a $d > 1$ dimensional space and learn the sparsity pattern simultaneously. The proposed ℓ_0-CCA problem albeit not being convex leads to an empirically stable solution across different settings. Nonetheless, the method has some limitations, specifically, tuning λ requires a cross validation procedure, which could be costly in high dimensional regime. Another caveat of the existing approach is that it lacks guarantees when trained on small batches. In the future, we aim to extend the method to enable compatibility with small batch training.