Variational Graph Autoencoders for Multiview Canonical Correlation Analysis

Yacouba Kaloga[§], Pierre Borgnat[§], Sundeep Prabhakar Chepuri[⋆], Patrice Abry[§] and Amaury Habrard[†].

[§] *Univ Lyon, Ens de Lyon, Univ. Claude Bernard, CNRS, Laboratoire de Physique, Lyon, France*
[⋆] *Department of Electrical and Communication Engineering, Indian Institute of Science, Bangalore, India*
[†] *University of Lyon, UJM-Saint-Etienne, CNRS, Laboratoire Hubert Curien, UMR 5516, France*

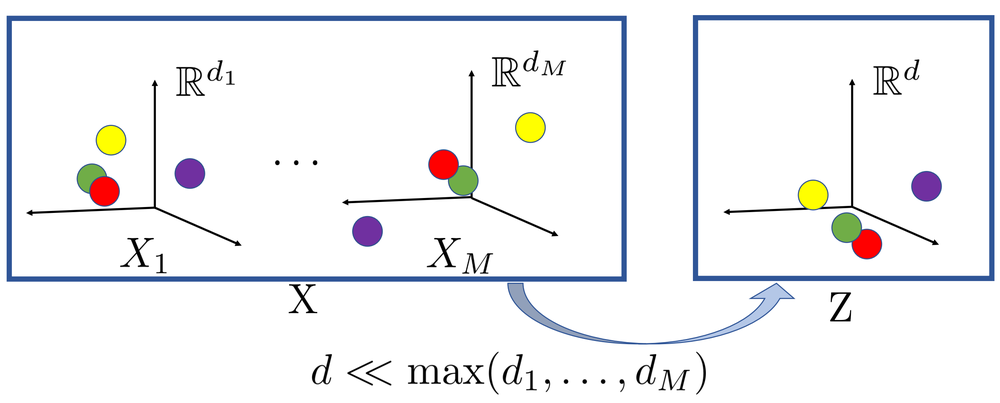
Supported by the IFCAM project MA/IFCAM/19/56, the ACADEMICS Grant of IDEXLYON, Univ. Lyon, PIA ANR-16-IDEX-0005

Introduction

We present a novel approach for multiview canonical correlation analysis based on a variational graph neural network model. For that, we propose a **nonlinear probabilistic model** which takes into account available **graph-based geometric constraints** while being **scalable to large scale datasets** with multiple views. The algorithm is competitive with state-of-the-art **multiview representation learning** techniques for classification, clustering, and recommendation tasks, in addition to being scalable and **robust to missing views data**.

Goals

- Canonical Correlation Analysis (CCA) can be used for multiview representation learning, by seeking latent low-dimensional representations that are common to all the different views.
- This common representation that encodes information from different datasets can be leveraged to improve the performance of machine learning tasks, e.g., clustering.



Background

There are two general approaches to CCA: *algebraic* or *probabilistic*.

Algebraic

- The *algebraic* approaches to CCA[1] were initially proposed for two-view data following and they obtain a latent low-dimensional manifold by **maximizing correlations** between the **projections** on the different views onto it.
- These approaches are powerful and versatile, there are many different extensions: **non-linear** (DCCA,KCCA,etc.), **graph-aware** (GMCCA) etc. But the multiview extension is a **hard problem** and **does not scale** well to large datasets.

Original CCA seeks, for two views dataset the best low dimensional projectors $U_1 \in \mathbb{R}^{d_1 \times d}$ and $U_2 \in \mathbb{R}^{d_1 \times d}$:

$$\min_{U_1,U_2} \|U_1^T X_1 - U_2^T X_2\|_F^2 \quad \text{s.t.} \quad U_i^T (X_i^T X_i) U_i = I_{d_i}.$$

Probabilistic

- Alternatively, *probabilistic* CCA solve a **Bayesian inference problem**[2].
- As recent advances in variational autoencoders[3] made Bayesian inference scalable, the probabilistic CCA approaches gained popularity because of their potential (e.g., inference task, **natural multiview extension** and **scalability**, e.g. see VCCA(p), or VPCCA).
- Currently there is no *probabilistic* method that takes into account potential graph structure of multiview data while it was shown that incorporating the available graph-induced knowledge about the common source into multiview CCA improves performance of various machine learning tasks.

$(X_1(:,i), \cdots, X_M(:,i))$ $\qquad$ $(X_1(:,j), \cdots, X_M(:,j))$

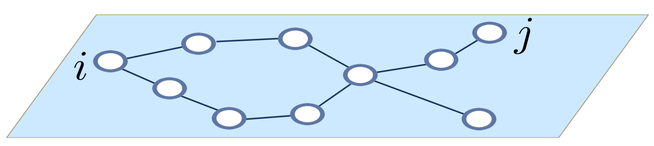


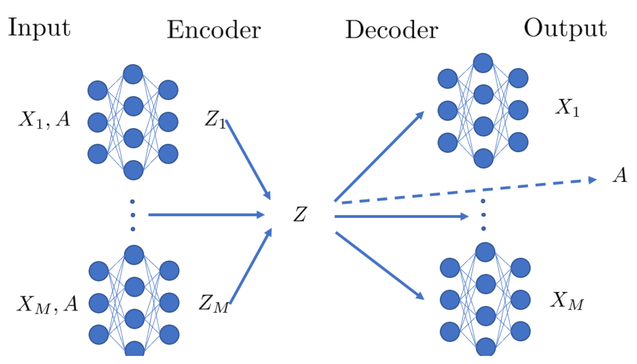
Figure 1: Multi-view data can rely on a graph (with adjency matrix A) and taking this structure into account can improve results.

Method	Complexity	Non Linear	>2 views	Graph	Robustness
CCA	$O(n)$	✗	✗	✗	✗
GMCCA	$O(n^2)$	✗	✓	✓	✗
VCCA(p)	$O(n)$	✓	✗	✗	✗
VPCCA	$O(n)$	✗	✓	✗	✗
MVGCCA	$O(n)$	✓	✓	✓	✓

Figure 2: The different approaches to CCA can be characterized by these properties.

MVGCCA

We now present our contribution that consists in proposing a **graph-aware probabilistic multiview CCA** model scalable and able to deal with missing views based on a variational autoencoders[3].



$$Z_m(:,i) \sim \mathcal{N}(\mu_m^{\text{enc}}(:,i), \text{diag}(\sigma(:,i)))$$
$$\mu_m = W_m^{\mu\,\text{enc}} \mathbf{Krylof}_m(X_m, A)$$
$$\sigma_m = W_m^{\sigma\,\text{enc}} \mathbf{Krylof}_m(X_m, A)$$
$$Z(:,i) \sim K \prod_{m=1}^{M} \mathcal{N}(\mu_m^{\text{enc}}(:,i), \text{diag}(\sigma(:,i)))$$

$$X_m(:,i) \sim \mathcal{N}(\mu_m^{dec}(:,i), \Psi_m)$$
$$\mu_m = W_m^{\mu\,\text{dec}} \mathbf{MLP}_m(X_m, A)$$
$$A_{i,j} \sim \mathcal{B}(1, \text{sigmoid}(Z(:,i)^\dagger Z(:,j)))$$

Figure 3: Representation of MVGCCA. All the views are encoded to their own latent space Z_m using the common graph. They are merged to form a common view Z. Finally, Z is tailored to decode all the views and original graph.

- We introduce decoders $p_\theta(A, X|Z)$ which is parametrized by a multilayer perceptron with weights θ.
- We introduce encoders $q_\eta(Z|X, A)$ which is parametrized by a Krylov graph convolutional neural network with weights η.

ELBO

The log-likelihood is intractable but low bounded by Evidence Lower Bound (ELBO):

$$\log p(X, A) \geq \mathbb{E}_{Z \sim q_\eta(Z|X,A)} \log p_\theta(X, A|Z) - D_{KL}(q_\eta(Z|X, A) \| p(Z)).$$

This low dimensional common latent representation is the barycenter (using precision as weights) of view's latent representations.

$$\mathbb{E}_{z \sim q_\eta(z|(x_m^i)_{m=1}^M)} = \left[\sum_{m=1}^{M} \frac{\mu_m^{\text{enc}}(k,i)}{\sigma_m^{\text{enc}2}(k,i)} \Big/ \sum_{m=1}^{M} \frac{1}{\sigma_m^{\text{enc}2}(k,i)} \right]_{k=1}^{d}.$$

Even when some views are missing given enough views, we could still compute an approximation of the common latent representation.

Experiments

UCI

UCI Handwritten Digits Dataset is a 6-view dataset of 2000 samples images representing the 9 digits (200 instances for each). These views correspond to specific transformations of the original image. Clustering, classification and reconstruction tasks are performed on this whole dataset. The prior graph over data is built from data.

- Fourier coefficients of the character shapes $X_1 \in \mathbb{R}^{76\times2000}$
- 240 pixel averages in 2 x 3 windows $X_4 \in \mathbb{R}^{240\times2000}$
- Karhunen-Loeve coefficients $X_3 \in \mathbb{R}^{64\times2000}$
- Profile correlations $X_2 \in \mathbb{R}^{216\times2000}$
- Zernike moments $X_5 \in \mathbb{R}^{47\times2000}$
- 6 morphological features $X_6 \in \mathbb{R}^{6\times2000}$

Twitter

A multiview datasetbased on post from Twitter. It consists of multiview representations of messages of users. They took 1% of the publicly available users data in April 2015. They removed all tweets that are not in english, and those from users who did not post between January and February 2015. Finally they only kept the last 100 tweets from the remaining users, yielding $n = 102327$ users. These data for each user are in the form of 6 1000-dimensional views: EgoTweets, MentionTweets, FriendTweets, FollowersTweets, FriendNetwork, and FollowerNetwork. A task of friend recommendation is performed as follows. The followed accounts are known for each user; given a highly followed account and a part of their followers, the goal is to determine, for each other users, whether or not he will follow this account after March 2015. For this task, a graph based on the Twitter dataset is built with the views Egoweets, FollowersTweets, and FriendNetwork.

Clustering and recommandation

Dataset	uci7			uci10			Recommendation		
Metric	Acc.	ARI	ARI2	Acc.	ARI	ARI2	Prec.	Recall	MRR
PCA	0.84	0.55	-	0.69	0.42	-	0.1511	0.0795	0.3450
GPCA	0.93	0.71	0.77	0.87	0.63	0.62	0.1578	0.0831	0.3649
MCCA	0.86	0.66	-	0.76	0.59	-	0.0815	0.0429	0.2225
GMCCA	**0.95**	**0.83**	0.84	0.90	0.69	0.71	**0.2290**	**0.1206**	**0.4471**
MVGCCA	**0.95**	0.82	**0.85**	**0.94**	**0.74**	**0.77**	0.1753	0.0583	0.4432

Table 2: Results of experiments on the different datasets and tasks; see text for the detailed discussion. Acc. stands for accuracy in classification; ARI for adjusted rank index in clustering tasks: ARI1 if using K-means and ARI2 if using spectral clustering. For the Recommendation task, Prec. is precision and MRR is the mean reciprocal rank.

Visualisation

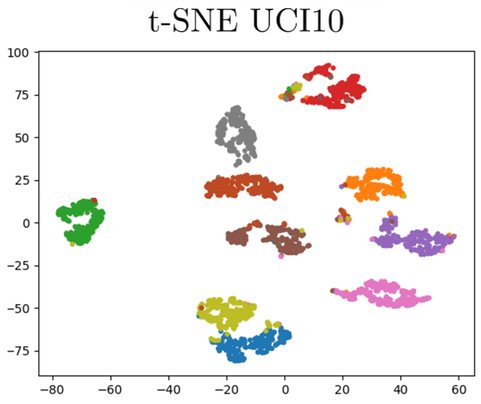


Figure 4: t-SNE visualisation in 2D of the latent space ($d = 3$) for uci dataset. Each color represents a different class.

Views reconstruction

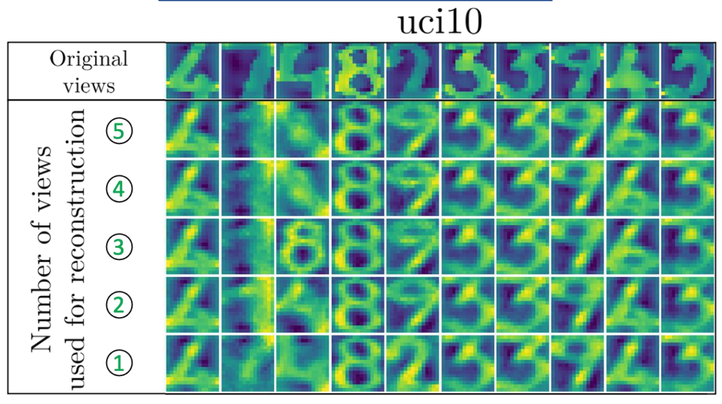
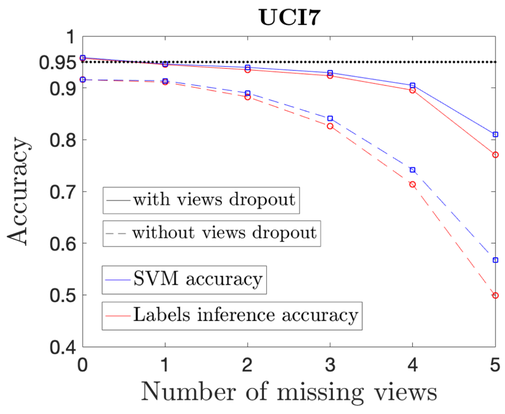


Figure 5: **UCI 4th view reconstruction.** We see here the result of the reconstruction experiment of the 4th view. The first line corresponds to the original view. Every figure in line $l > 2$ corresponds to a reconstruction from $7 - l$ views.

Robustness & reconstruction



In this experiment the quality of the reconstruction is evaluated more quantitatively and this will show the robustness of the proposed model to missing views. To do so, we add to the training set of UCI a 7th view that corresponds to a one hot encoding of the instance label. Then, we train MVGCCA on this extended train set. Finally, we evaluate accuracy on test set in two manners:

- We train a SVM-RBF on train set embeddings from trained model. Then we evaluate this SVM on test set embeddings where some views have been removed.
- We regenerate the 7th views of test set instances. This directly gives us an estimation of their labels. Once again, some views are removed.

For a given number of available views $v \in [1, 5]$, we consider any possible combination of views to form novel datasets, and we average the obtained accuracies (over 10 experiments). As we can see, the model can deal with a small number of missing views; still the performance decreases with the number of these missing views and that is not surprising.

Conclusion

We proposed **MVGCCA**, a novel **multiview** and **non linear extension of CCA** based on a **Bayesian inference model**. The proposed model is **scalable**, and it can take into account the available **graph structural information** from data. Also, it is the only **robust model** when confronted to incomplete data (with missing views). We have proposed also a robustification method by applying "*views dropout*" during training which is an original idea. The probabilistic graphical nature of the model can be used for other tasks, such as addressing link prediction tasks for multiview datasets, and that is a perspective for future work.

References

[1] H. Hotelling, Relations between two sets of variates, Biometrika 28462(1936) 321–377
[2] F.R. Bach, M. I.Jordan, A probabilistic interpretation of canonical correlation analysis, Technical Report,Department of Statictics, Uniersity of California, Berkeley, 2005
[3] T.N. Kipf, M. Welling, Variational graph auto-encoders, arXiv:stat.ML1611.07308(2016)