Semi-Supervised Laplacian Regularization of Kernel Canonical Correlation Analysis

Matthew B. Blaschko, Christoph H. Lampert, and Arthur Gretton

Max Planck Institute for Biological Cybernetics
Department of Empirical Inference
Spemannstr. 38, 72076 Tübingen, Germany
firstname.lastname@tuebingen.mpg.de

Abstract. Kernel canonical correlation analysis (KCCA) is a dimensionality reduction technique for paired data. By finding directions that maximize correlation, KCCA learns representations that are more closely tied to the underlying semantics of the data rather than noise. However, meaningful directions are not only those that have high correlation to another modality, but also those that capture the manifold structure of the data. We propose a method that is simultaneously able to find highly correlated directions that are also located on high variance directions along the data manifold. This is achieved by the use of semi-supervised Laplacian regularization of KCCA. We show experimentally that Laplacian regularized training improves class separation over KCCA with only Tikhonov regularization, while causing no degradation in the correlation between modalities. We propose a model selection criterion based on the Hilbert-Schmidt norm of the semi-supervised Laplacian regularized cross-covariance operator, which we compute in closed form.

1 Introduction

Kernel canonical correlation analysis (KCCA) is a fundamental technique for dimensionality reduction that relies on paired data to learn directions that maximize correlation between the projected representations in each space [1, 2]. Techniques based on only one space are susceptible to failure in the event that there are high-variance, semantically meaningless noise directions. KCCA overcomes this weakness by requiring that the projected data be correlated to a projection of the other modality, and has been shown to increase class separability when compared to single modality dimensionality reduction [3]. While KCCA often gives superior results to single modality dimensionality reduction techniques, correlation with some output modality may not be the only criterion of interest. We wish to find directions that not only relate the two modalities, but also lie along the data manifold, in order to better represent the structure of the data and improve class separability. In this work, we describe a method to incorporate these two goals into a common optimization by employing semi-supervised Laplacian regularization. This method gives an embedding of the data that makes use of the information between modalities, as well as the information within each single

modality. By using Laplacian regularization, we are able to learn directions that tend to lie along the data manifold estimated from a much larger set of data [4]. This gives us greater confidence that the learned directions represent the underlying statistical structure of the data and that we have not been misled by small sample effects. We show experimentally that learning along the manifold results in increased performance, even in the fully supervised setting, in that the learned embeddings give better class separability on a variety of datasets.

One way to evaluate the performance of KCCA is to take the sum of the squared correlations that it reveals. This quantity turns out to be the Hilbert-Schmidt norm of the normalized covariance operator between the feature representations of each modality, and is referred to as the Hilbert-Schmidt normalized independence criterion [5]. The underlying concept of semi-supervised Laplacian regularization of KCCA can also be applied to an empirical estimate of this operator, and therefore also to the independence criterion. Here, we make use of this Laplacian regularized estimate to define a model selection criterion for the regularization parameters that can be computed in closed form from the kernel matrices and Laplacian.

The rest of the paper is organized as follows. We discuss related work in Section 2 and give a review of KCCA in Section 3. In Section 4 we present the semi-supervised Laplacian regularization of KCCA. In Section 4.3 we discuss the relationship between the proposed algorithm and a recently introduced semi-supervised Fisher linear discriminant analysis algorithm. We describe our model selection criterion in Section 5 and also introduce the semi-supervised Laplacian regularized estimate of the HSNIC. Experimental results are presented in Section 6. Finally, we conclude in Section 7.

2 Related Work

Although KCCA has been applied in many situations, including cross media information retrieval [2, 6], multi-modal data clustering [3], analysis of fMRI data [7], extraction of gene clusters [8], testing for independence [9, 10], and ICA [11], to our knowledge there have been no semi-supervised extensions of the algorithm. Laplacian regularization is a common technique for semi-supervised learning [4, 12]. [13] have recently proposed a semi-supervised Fisher linear discriminant analysis algorithm based on Laplacian regularization, which we show in Section 4.3 to be a special case of the algorithm proposed here.

In our experiments, we will perform model selection by making use of various statistics computed on the correlation operator spectrum (see Section 5): we therefore provide a brief overview of methods used to evaluate and summarize this spectrum. A variety of statistics on the correlation operator spectrum are presented in [9] (for the spline kernel RKHS), where these are used for independence testing. Statistics on the correlation operator used for independent component analysis in [11] include the maximum singular value and kernel generalized variance, where the latter is an upper bound near independence on the mutual information [14]. Finally, a closed form expression for the Hilbert-

Schmidt norm of the correlation operator is provided in [5], where it is shown that this norm is an estimate of the mean squared contingency. Finally, the spectrum of two correlation operators can be compared directly for model selection, as in [2].

3 A Review of Kernel Canonical Correlation Analysis

3.1 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) seeks to utilize paired datasets to simultaneously find projections from each feature space that maximize the correlation between the projected representations [1]. Given a sample from a paired dataset[1] $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ we would like to simultaneously find directions w_x and w_y that maximize the correlation of the projections of x onto w_x with the projections of y onto w_y. This is expressed as

$$\max_{w_x, w_y} \frac{\hat{E}\left[\langle x, w_x\rangle \langle y, w_y\rangle\right]}{\sqrt{\hat{E}\left[\langle x, w_x\rangle^2\right] \hat{E}\left[\langle y, w_y\rangle^2\right]}}, \tag{1}$$

where $\hat{E}$ denotes the empirical expectation. We denote the covariance matrix of (x, y) by C and use the notation C_{xy} (C_{xx}) to denote the cross (auto) covariance matrices between x and y. Equation (1) is equivalent to

$$\max_{w_x, w_y} \frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T C_{xx} w_x \; w_y^T C_{yy} w_y}}. \tag{2}$$

This Rayleigh quotient can be optimized as a generalized eigenvalue problem, or by decomposing the problem using the Schur complement as described in [2].

There is a natural extension of CCA in the event where there are more than two modalities. This can be written as a generalized eigenvector problem that subsumes two-way CCA as a special case

$$\begin{pmatrix} C_{11} & \ldots & C_{1k} \\ \vdots & \ddots & \vdots \\ C_{k1} & \ldots & C_{kk} \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_k \end{pmatrix} = \lambda \begin{pmatrix} C_{11} & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & C_{kk} \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_k \end{pmatrix}. \tag{3}$$

3.2 Kernel Canonical Correlation Analysis

We can extend CCA, *e.g.* to non-vectorial domains by defining kernels over x and y: $k_x(x_i, x_j) = \langle \phi_x(x_i), \phi_x(x_j)\rangle$ and $k_y(y_i, y_j) = \langle \phi_y(y_i), \phi_y(y_j)\rangle$, and searching for solutions that lie in the span of $\phi_x(x)$ and $\phi_y(y)$: $w_x = \sum_i \alpha_i \phi_x(x_i)$ and $w_y = \sum_i \beta_i \phi_y(y_i)$. In this setting we use an empirical estimator for C:

$$\hat{C}_{xy} = \frac{1}{n} \sum_{i=1}^{n} \phi_x(x_i) \cdot \phi_y(y_i)^T, \tag{4}$$

[1] We assume the samples have zero mean for notational convenience.

where n is the sample size, and $\phi_x(x_i)$ and $\phi_y(y_i)$ are assumed to have 0 mean. $\hat{C}_{xx}$ and $\hat{C}_{yy}$ are defined similarly. Denoting the kernel matrices defined by our sample as K_x and K_y, the solution of Equation (2) is equivalent to maximizing the following with respect to coefficient vectors, α and β

$$\frac{\alpha^T \frac{1}{n} K_x K_y \beta}{\sqrt{\alpha^T \frac{1}{n} K_x^2 \alpha \beta^T \frac{1}{n} K_y^2 \beta}} = \frac{\alpha^T K_x K_y \beta}{\sqrt{\alpha^T K_x^2 \alpha \beta^T K_y^2 \beta}}. \tag{5}$$

As discussed in [2] this optimization leads to degenerate solutions in the case that either K_x or K_y is invertible so we maximize the following regularized expression

$$\frac{\alpha^T K_x K_y \beta}{\sqrt{\alpha^T \left(K_x^2 + \varepsilon_x K_x\right) \alpha \beta^T \left(K_y^2 + \varepsilon_y K_y\right) \beta}}, \tag{6}$$

which is equivalent to Tikhonov regularization of the norms of w_x and w_y in the denominator of Equation (2). In the limit case that $\varepsilon_x \to \infty$ and $\varepsilon_y \to \infty$, the algorithm maximizes covariance instead of correlation.

The formulation of CCA in Equation (3) is also readily regularized and kernelized, and allows one to take advantage of more than two modalities at a time.

4 Semi-Supervised Kernel Canonical Correlation Analysis

If we have additional data available that do not have correspondences to the other modality, we can search for solutions that lie in the span of the larger set of training points, and regularize using the additional data. We propose Laplacian regularization, which tends to find solutions that lie along an empirical estimate of the data manifold [4]. This gives increased robustness to the algorithm, and increases class separability in the absence of label information.

4.1 The Two-Modality Case

We have training data $\{x_1, \ldots, x_n\}$ with corresponding data $\{y_1, \ldots, y_n\}$ as well as additional training data $\{x_{n+1}, \ldots, x_{n+p_x}\}$ and $\{y_{n+1}, \ldots, y_{n+p_y}\}$ that do not have correspondences. We use the variables $m_x = n + p_x$ ($m_y = n + p_y$) to denote the total number of samples in modality x (y). We denote the $d \times n$ data matrix $X = (x_1, \ldots, x_n)$, and the matrix including all data with and without correspondences $\hat{X} = (x_1, \ldots, x_n, x_{n+1}, \ldots, x_{n+p_x})$, and similarly for Y and $\hat{Y}$. Furthermore we denote kernel matrices between the various sets of data as follows: $\Phi_x(X)^T \Phi_x(X) = K_{xx}$, $\Phi_x(\hat{X})^T \Phi_x(X) = K_{\hat{x}x}$, $\Phi_x(\hat{X})^T \Phi_x(\hat{X}) = K_{\hat{x}\hat{x}}$, *etc.*. Kernel matrices for Y are defined analogously. We wish to optimize the following generalization of Equation (6)

$$\frac{\alpha^T K_{\hat{x}x} K_{y\hat{y}} \beta}{\sqrt{\alpha^T \left(K_{\hat{x}x} K_{x\hat{x}} + R_{\hat{x}}\right) \alpha \beta^T \left(K_{\hat{y}y} K_{y\hat{y}} + R_{\hat{y}}\right) \beta}}, \tag{7}$$

where $R_{\hat{x}} = \varepsilon_x K_{\hat{x}\hat{x}} + \frac{\gamma_x}{m_{\hat{x}}^2} K_{\hat{x}\hat{x}} \mathcal{L}_{\hat{x}} K_{\hat{x}\hat{x}}$ and $\mathcal{L}_{\hat{x}}$ is the empirical graph Laplacian estimated from the m_x samples of labeled and unlabeled data.

4.2 The General Case

In the general case, we have more than two modalities. As a result, the data that has correspondences between modalities 1 and 2 can be different than the data that has correspondences between modalities 2 and 3, *etc.*. We abuse the notation $K_{\hat{i}j}$ to denote the kernel matrix computed between all the data for modality i and the data for modality i that also has correspondences to the data in modality j. This matrix has dimensionality $m_i \times n_{ij}$, where m_i is the total number of training examples (with or without correspondences) for modality i, and n_{ij} is the number of correspondences between modalities i and j.

The following generalizes Equations (3) and (7)

$$\begin{pmatrix} \mathbf{0} & \dots & \frac{1}{n_{1k}} K_{\hat{1}k} K_{1\hat{k}} \\ \vdots & \ddots & \vdots \\ \frac{1}{n_{1k}} K_{\hat{k}1} K_{k\hat{1}} & \dots & \mathbf{0} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} = \tag{8}$$
$$\lambda \begin{pmatrix} \frac{1}{m_1} K_{\hat{1}1} K_{1\hat{1}} + R_{\hat{1}} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \frac{1}{m_k} K_{\hat{k}k} K_{k\hat{k}} + R_{\hat{k}} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}.$$

4.3 Fisher Linear Discriminant Analysis

There is an intimate relationship between CCA and Fisher linear discriminant analysis (LDA) [15]. LDA is a special case of CCA where the second modality is the labels [16, 17], consequently, any semi-supervised algorithm for CCA implies a semi-supervised LDA algorithm as well. Recently [13] have proposed a semi-supervised LDA approach. If we use the identity kernel on the labels, set the label regularization parameters to 0, and set $\varepsilon_x = 0$, the directions learned from Equation (7) are the same as those found using the method of [13].

Similarly, when one of the spaces is one-dimensional (*i.e.* the kernel matrix is rank 1), Laplacian regularized KCCA gives a generalization of kernel ridge regression.

5 Model Selection

We propose a model selection algorithm based on the Hilbert-Schmidt normalized information criterion (HSNIC). The HSNIC is closely related to KCCA and is equivalent to the squared ℓ_2 norm of the spectrum of the normalized cross-covariance operator, which in the limit is independent of the kernel[2] used in

[2] Assuming that the kernel comes from the class of *characteristic* kernels, as defined in [10]. A Gaussian kernel is sufficient.

its estimation [5, 10]. Because the spectrum of the normalized cross-covariance operator is identical to the spectrum of the solutions to KCCA, this provides us with a useful statistic upon which we can base our model selection. HSNIC gives us access to the ℓ_2 norm of the spectrum, which is dominated by the first KCCA directions for kernels with quickly decaying spectra, such as the Gaussian kernel [11, 18].

We first derive the semi-supervised empirical HSNIC estimate in Section 5.1 and then use this result to define our model selection criterion in Section 5.2.

5.1 Semi-Supervised Empirical HSNIC Estimate

The HSNIC is the Hilbert-Schmidt norm of the normalized cross-covariance operator, V_{xy}, which we define to be regularized using the Laplace-Beltrami operators on the manifolds of the data [4], $\Delta_{\mathcal{M}_x}$ and $\Delta_{\mathcal{M}_y}$,

$$V_{xy} = \left(\Sigma_{xx} + \varepsilon_x I + \gamma_x \Delta_{\mathcal{M}_x}\right)^{-\frac{1}{2}} \Sigma_{xy} \left(\Sigma_{yy} + \varepsilon_y I + \gamma_y \Delta_{\mathcal{M}_y}\right)^{-\frac{1}{2}}. \tag{9}$$

We estimate the normalized cross-covariance operator empirically using a finite sample of data, yielding

$$\hat{V}_{xy} = \left(\frac{1}{n} X X^T + \varepsilon_x I + \frac{\gamma_x}{m_x^2} \hat{X} \mathcal{L}_{\hat{x}} \hat{X}^T\right)^{-\frac{1}{2}} \frac{1}{n} X Y^T \cdot \left(\frac{1}{n} Y Y^T + \varepsilon_y I + \frac{\gamma_y}{m_y^2} \hat{Y} \mathcal{L}_{\hat{y}} \hat{Y}^T\right)^{-\frac{1}{2}}. \tag{10}$$

The semi-supervised Laplacian regularized empirical estimate of the HSNIC is therefore

$$\|\hat{V}_{xy}\|_{HS}^2 = \operatorname{Tr}\left[\hat{V}_{xy} \hat{V}_{xy}^T\right] = \operatorname{Tr}\left[M_x M_y\right], \tag{11}$$

where

$$M_x = I - n\left(nI + \frac{1}{\varepsilon_x} K_{xx} - \frac{1}{\varepsilon_x} K_{x\hat{x}} \left(\frac{m_x^2 \varepsilon_x}{\gamma_x} I + \mathcal{L}_{\hat{x}} K_{\hat{x}\hat{x}}\right)^{-1} \mathcal{L}_{\hat{x}} K_{\hat{x}x}\right)^{-1}, \tag{12}$$

and M_y is defined analogously. See Section A for the derivation.

5.2 HSNIC Model Selection Criterion

HSNIC is an interesting model selection criterion for many problems as it provides an estimate of the dependence between X and Y [10]. As discussed earlier, KCCA in high dimensional feature spaces requires regularization to return non-trivial projection directions: in the event that all regularization is set to 0, HSNIC estimates perfect correlation if the kernel matrices, K_{xx} and K_{yy}, are invertible. Since choosing the parameters that maximize HSNIC risks overfitting, it is more meaningful to consider the amount by which the dependence

witnessed by HSNIC increases over its value at independence (i.e., in the absence of correlations between X and Y). We can simulate the latter quantity by randomly permuting the labels relative to the data: if we were to average several such permutations, we would obtain an estimate of HSNIC at independence. The averaging procedure is computationally expensive, however: thus, we use a single data permutation to approximate the HSNIC value at independence. We observed on our data that the values of HSNIC for different permutations were highly concentrated about their mean, which makes this a reasonable approximation. The model selection criterion consists of the ratio between the non-permuted and the permuted HSNIC values. If this ratio is high, we are confident that the correlation found is genuine and is not a result of overfitting. The HSNIC estimate for the permuted dataset is easily computed using a random permutation matrix, P,

$$\|\hat{V}_{xy_R}\|^2_{HS} = \mathrm{Tr}\left[M_x P^T M_y P\right], \tag{13}$$

where $\hat{V}_{xy_R}$ is the empirical estimate computed using $Y_R = YP$ in place of Y in Equation (10). This can be verified with an analogous derivation to that in Section A. We denote the model selection criterion

$$\rho(\varepsilon_x, \gamma_x, \varepsilon_y, \gamma_y) = \frac{\|\hat{V}_{xy}\|^2_{HS}}{\|\hat{V}_{xy_R}\|^2_{HS}}, \tag{14}$$

and maximize with respect to its parameters. The cost of computing Equation (14) is only marginally higher than computing Equation (11) as we can reuse the computation of M_x and M_y in the permuted version.

6 Experimental Results

6.1 Data

We have performed experiments on a number of datasets of images with associated text. We have used the three datasets included in the UIUC-ISD collection [19]. These consist of images collected from search engines using ambiguous search terms, "bass," "crane," and "squash," the webpages in which the images originally appeared, and an annotation of which sense of the word the image represents, *e.g.* fish vs. musical instrument. There are 2881 images in the Bass dataset which have been grouped into 6 categories, 2650 in the Crane dataset grouped into 9 categories, and 1948 images in the Squash dataset grouped into 6 categories. For all three datasets, we extracted 128 dimensional SURF descriptors without rotation invariance and with the keypoint threshold set to 0 [20] and constructed a codebook of 1000 visual words using k-means with 50000 sampled descriptors. Images were represented by a normalized histogram of these visual words. For the text representation, we used term frequency histograms extracted from the webpage title, removing special characters and stop words using the list from [21]. Both image and text similarities were computed using a χ^2 kernel,

$$k(x, x') = e^{-\frac{1}{2A}\sum_{i=1}^{d}\frac{(x_i - x'_i)^2}{x_i + x'_i}}, \tag{15}$$

with normalization parameter A set to the median of the χ^2 distances in the training set.

Additionally, we have used the multimedia image-text web database used in [2, 22] which consists of samples from three classes: sports, aviation, and paintball, with 400 image-text pairs each. Images were represented using HSV color and Gabor textures as in [2, 22]. Text was represented using term frequencies. As in [2] we have used a Gaussian kernel for the image space, and a linear kernel for text.

6.2 Evaluation Methodology

To evaluate the performance of the algorithm, the following evaluation is performed. We randomly split the data into equally sized train and test portions. The train portion is further split into data with and without correspondences between the different modalities. Semi-supervised Laplacian regularized KCCA is trained using data with and without correspondences, using parameters learned with grid search on the objective described in Section 5.2. Test data are embedded using the learned parameters, and correlations are computed between the embeddings of the two modalities. We repeat this procedure 40 times, and evaluate the performance using two metrics: the mean ℓ_2 norm of the cross-correlation coefficients, to determine how well the projected data are correlated; and the ratio of the determinant of the total scatter matrix and the determinant of sum of within class scatter matrices for each modality to determine how well the within-class variation along the data manifold is captured:

$$\frac{|S_t|}{|\sum_{i=1}^{c} S_{c_i}|}, \tag{16}$$

where μ denotes the mean of the embedded test data, c_i denotes the test data that are in class i, μ_i denotes the mean of class i,

$$S_t = \sum_{j \in test} (x_j - \mu)(x_j - \mu)^T, \tag{17}$$

and

$$S_{c_i} = \sum_{j \in c_i} (x_j - \mu_i)(x_j - \mu_i)^T. \tag{18}$$

Although class labels are available for the dataset, we only use them at test time during this evaluation in order to measure the separation of semantic classes achieved by the embeddings. We compute the embeddings without the semi-supervised Laplacian regularization and also visualize the norm of the resulting correlation coefficients, and scatter ratios.

In all experiments except for the "Sports Aviation Paintball" dataset, we have defined the Laplacian using the similarity matrix, W, defined by the kernel described in Section 6.1. In the "Sports Aviation Paintball" dataset, we use a linear kernel on the text modality for consistency with previous publications [2,

3]. Insted, we have used a Gaussian kernel to compute the Laplacian matrix for the text modality. In all cases, we use the symmetric normalized Laplacian, $\mathcal{L} = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}}$, where D is the diagonal matrix whose entries are the row sums of W.

6.3 Results

Figure 1 gives results for the four datasets described in Section 6.1. The plots have been computed by varying the percentage of training data for which correspondences between images and text have been provided to the algorithms. For more pairs of data, we see that correlations are better represented for KCCA with and without Laplacian regularization, as expected. The advantage of Laplacian regularization is shown by improved class separability (as measured by scatter ratios) for three of the four datasets. This indicates that the manifold structure of these datasets is important for class separability, and that this is captured without sacrificing performance on correlation.

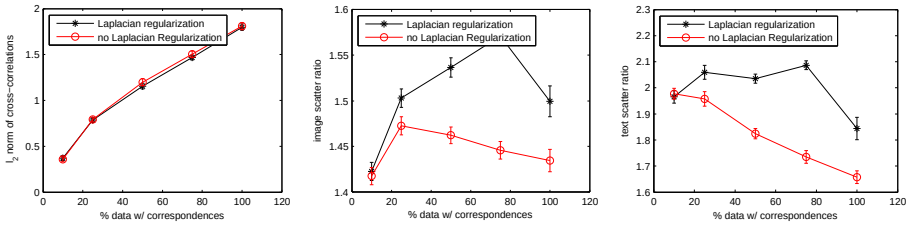
Laplacian regularization slightly decreases cross-correlation and scatter ratios in the "Sport, Aviation, Paintball" dataset (first column). This can be described in part by the relatively simple structure of the "Sport, Aviation, Paintball" dataset. PCA, kernel-PCA, and KCCA all give similar embeddings for this dataset, with the majority of variance contained in only two dimensions [3]. The use of Laplacian regularization is therefore unnecessary as there is little non-linearity in the data manifold; manifold structure is effectively captured by linear high variance directions and non-parametric Laplacian regularization degrades performance.

For the "Bass," "Crane," and "Squash" datasets, Laplacian regularization is able to capture relevant discriminative structure that is available in each modality without sacrificing performance in finding directions that show correlation between the modalities.

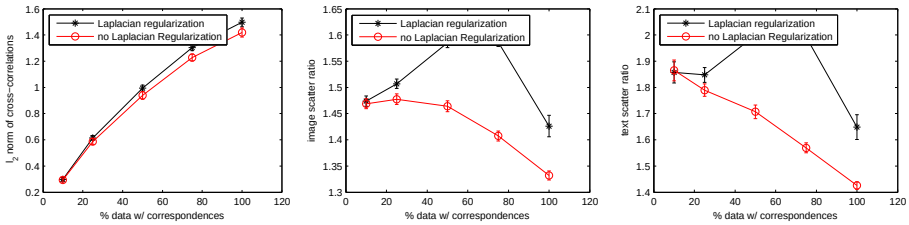
7 Conclusions and Future Work

In this work we have proposed the use of Laplacian regularized kernel canonical correlation analysis as a dimensionality reduction technique. Experimental results show increased performance in class separation for datasets that have sufficient nonlinear structure. We have proposed a model selection criterion based on the Hilbert-Schmidt norm of the Laplacian regularized normalized cross covariance operator and have derived its solution in closed form (Equation (11)).

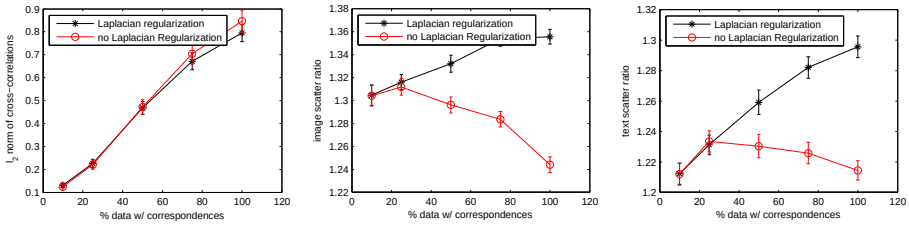
The Hilbert-Schmidt normalized information criterion is an important statistical object that can be used to test for independence of sets of variables, which gives rise to many applications in machine learning. A promising area for future work is to experimentally validate the benefit of using the Laplacian regularized empirical estimate in applications where only Tikhonov regularization has been previously applied. Examples include causality inference [10] and ICA [11].



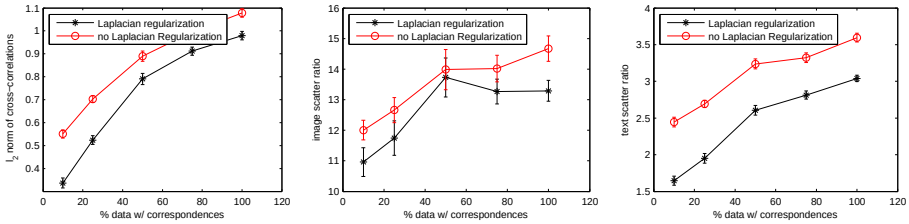
(a) Bass



(b) Crane



(c) Squash



(d) Sports, Aviation, Paintball

Fig. 1. Experimental results for four different datasets. The first column is the ℓ_2 norm of the cross correlations between the modalities of the held out data. The second and third columns are the scatter ratios for images and text, respectively.

All experiments here have been performed using only two modalities. Laplacian regularization of KCCA for multiple modalities, as described in Equation (8) warrants further experimental evaluation. This is particularly relevant, *e.g.*, in multi-language text corpora for which correspondences for some but not all documents are known. Laplacian regularization would allow better modeling of the characteristics of each of the individual languages.

Finally, depending on the structure of a dataset, iterated Laplacian regularization may be appropriate in some cases [23]. This gives stronger conditions on the structure of the manifold which may help in avoiding overfitting.

Acknowledgments

The first author is supported by a Marie Curie fellowship under the EU funded project PerAct, EST 504321. This work is funded in part by the CLASS project, IST 027978.

References

1. Hotelling, H.: Relations Between Two Sets of Variates. Biometrika **28** (1936) 321–377
2. Hardoon, D.R., Szedmák, S., Shawe-Taylor, J.R.: Canonical Correlation Analysis: An Overview with Application to Learning Methods. Neural Computation **16** (2004) 2639–2664
3. Blaschko, M.B., Lampert, C.H.: Correlational Spectral Clustering. In: CVPR. (2008)
4. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. JMLR **7** (2006) 2399–2434
5. Fukumizu, K., Bach, F.R., Gretton, A.: Statistical Consistency of Kernel Canonical Correlation Analysis. JMLR **8** (2007) 361–383
6. Li, Y., Shawe-Taylor, J.: Using kcca for japanese—english cross-language information retrieval and document classification. J. Intell. Inf. Syst. **27** (2006) 117–133
7. Hardoon, D.R., Mourão-Miranda, J., Brammer, M., Shawe-Taylor, J.: Unsupervised Analysis of fMRI Data Using Kernel Canonical Correlation. NeuroImage **37** (2007) 1250–1259
8. Yamanishi, Y., Vert, J.P., Nakaya, A., Kanehisa, M.: Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. Bioinformatics **19** (2003) i323–330
9. Dauxois, J., Nkiet, G.M.: Nonlinear canonical analysis and independence tests. Ann. Statist. **26** (1998) 1254–1278
10. Fukumizu, K., Gretton, A., Sun, X., Schölkopf, B.: Kernel Measures of Conditional Dependence. In: NIPS. (2007)
11. Bach, F.R., Jordan, M.I.: Kernel Independent Component Analysis. JMLR **3** (2002) 1–48
12. Chapelle, O., Schölkopf, B., Zien, A., eds.: Semi-Supervised Learning. MIT Press, Cambridge, MA (2006)
13. Cai, D., He, X., Han, J.: Semi-supervised discriminant analysis. In: ICCV. (2007)

14. Gretton, A., Herbrich, R., Smola, A., Bousquet, O., Schölkopf, B.: Kernel methods for measuring independence. J. Mach. Learn. Res. **6** (2005) 2075–2129
15. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of Eugenics **7** (1936) 179–188
16. De Bie, T.: Semi-supervised learning based on kernel methods and graph cut algorithms. Phd thesis, K.U.Leuven (Leuven, Belgium), Faculty of Engineering (2005)
17. Bach, F.R., Jordan, M.I.: A Probabilistic Interpretation of Canonical Correlation Analysis. Technical Report 688, Department of Statistics, University of California, Berkeley (2005)
18. Braun, M.L.: Accurate error bounds for the eigenvalues of the kernel matrix. JMLR **7** (2006) 2303–2328
19. Loeff, N., Alm, C.O., Forsyth, D.A.: Discriminating Image Senses by Clustering with Multimodal Features. In: ACL. (2006)
20. Bay, H., Tuytelaars, T., Gool, L.J.V.: SURF: Speeded Up Robust Features. In: ECCV. (2006) 404–417
21. van Rijsbergen, C.J.: Information Retrieval. Butterworths (1975)
22. Kolenda, T., Hansen, L.K., Larsen, J., Winther, O.: Independent Component Analysis for Understanding Multimedia Content. In: IEEE Workshop on Neural Networks for Signal Processing. (2002) 757–766
23. Zhou, D., Schölkopf, B.: Discrete regularization. In Chapelle, O., Schölkopf, B., Zien, A., eds.: Semi-supervised learning. Adaptive computation and machine learning. MIT Press, Cambridge, Mass., USA (2006) 221–232

A Derivation of Semi-Supervised Empirical HSNIC Estimate

The measure of interest is the Hilbert-Schmidt norm of the semi-supervised empirical estimate of the normalized cross-covariance operator

$$\|\hat{V}_{xy}\|_{HS}^2 = \operatorname{Tr}\left[\hat{V}_{xy}\cdot\hat{V}_{xy}^T\right] \tag{19}$$

$$= \frac{1}{n^2}\operatorname{Tr}\left[X^T\left(\frac{1}{n}XX^T+\varepsilon_x I+\frac{\gamma_x}{m_x^2}\hat{X}\mathcal{L}_{\hat{x}}\hat{X}^T\right)^{-1}X\right.$$
$$\left.Y^T\left(\frac{1}{n}YY^T+\varepsilon_y I+\frac{\gamma_y}{m_y^2}\hat{Y}\mathcal{L}_{\hat{y}}\hat{Y}^T\right)^{-1}Y\right]. \tag{20}$$

Using the Woodbury matrix identity, $(A+BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}$, we substitute $A = \varepsilon_x I + \frac{1}{n}XX^T$, $B = \hat{X}$, $C = \frac{\gamma_x}{m_x^2}\mathcal{L}_{\hat{x}}$, and

$D = \hat{X}^T$. The following holds

$$\left(\frac{1}{n}XX^T + \varepsilon_x I + \frac{\gamma_x}{m_x^2}\hat{X}\mathcal{L}_{\hat{x}}\hat{X}^T\right)^{-1} = \left(\varepsilon_x I + \frac{1}{n}XX^T\right)^{-1} - \left(\varepsilon_x I + \frac{1}{n}XX^T\right)^{-1}\hat{X}\cdot \left(\frac{m_x^2}{\gamma_x}\mathcal{L}_{\hat{x}}^{-1} + \hat{X}^T\left(\varepsilon_x I + \frac{1}{n}XX^T\right)^{-1}\hat{X}\right)^{-1}\cdot \hat{X}^T\left(\varepsilon_x I + \frac{1}{n}XX^T\right)^{-1}. \quad (21)$$

We apply the same identity again with the substitution $A = \varepsilon_x I$, $B = X$, $C = \frac{1}{n}I$, and $D = X^T$ to achieve the result

$$\left(\varepsilon_x I + \frac{1}{n}XX^T\right)^{-1} = \frac{1}{\varepsilon_x}I - \frac{1}{\varepsilon_x^2}X\left(nI + \frac{1}{\varepsilon_x}X^TX\right)^{-1}. \quad (22)$$

Plugging in the results of Equations (21) and (22), along with the analogous term for Y, into Equation (20), we achieve the result

$$\|\hat{V}_{xy}\|_{HS}^2 = \mathrm{Tr}\left[M_x M_y\right], \quad (23)$$

where

$$\begin{aligned} M_x = \frac{1}{n}\Bigg(&\frac{1}{\varepsilon_x}K_{xx} - \frac{1}{\varepsilon_x^2}K_{xx}\left(nI + \frac{1}{\varepsilon_x}K_{xx}\right)^{-1}K_{xx} - \\ &\left(\frac{1}{\varepsilon_x}K_{x\hat{x}} - \frac{1}{\varepsilon_x^2}K_{xx}\left(nI + \frac{1}{\varepsilon_x}K_{xx}\right)^{-1}K_{x\hat{x}}\right)\cdot \\ &\Bigg(\frac{m_x^2}{\gamma_x}\mathcal{L}_{\hat{x}}^{-1} + \frac{1}{\varepsilon_x}K_{\hat{x}\hat{x}} - \\ &\qquad\frac{1}{\varepsilon_x^2}K_{\hat{x}x}\left(nI + \frac{1}{\varepsilon_x}K_{xx}\right)^{-1}K_{x\hat{x}}\Bigg)^{-1}\cdot \\ &\left(\frac{1}{\varepsilon_x}K_{\hat{x}x} - \frac{1}{\varepsilon_x^2}K_{\hat{x}x}\left(nI + \frac{1}{\varepsilon_x}K_{xx}\right)^{-1}K_{xx}\right)\Bigg), \end{aligned} \quad (24)$$

and M_y is defined analogously. We can further simplify this expression by applying the Woodbury matrix identity in reverse twice, which results in

$$M_x = I - n\left(nI + \frac{1}{\varepsilon_x}K_{xx} - \frac{1}{\varepsilon_x}K_{x\hat{x}}\left(\frac{m_x^2\varepsilon_x}{\gamma_x}I + \mathcal{L}_{\hat{x}}K_{\hat{x}\hat{x}}\right)^{-1}\mathcal{L}_{\hat{x}}K_{\hat{x}x}\right)^{-1}. \quad (25)$$