Multi-View Canonical Correlation Analysis

Jan Rupnik (1), John Shawe-Taylor (2)
(1) Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
(2) University College London, Gower Street, London WC1E 6BT, United Kingdom
e-mail: jan.rupnik@ijs.si

ABSTRACT

Canonical correlation analysis (CCA) is a method for finding linear relations between two multidimensional random variables. This paper presents a generalization of the method to more than two variables. The approach is highly scalable, since it scales linearly with respect to the number of training examples and number of views (standard CCA implementations yield cubic complexity). The method is also extended to handle non-linear relations via kernel trick (this increases the complexity to quadratic complexity). The scalability is demonstrated on a large scale cross-lingual information retrieval task.

1 INTRODUCTION

Principal Component Analysis is a very popular approach to dimensionality reduction in the field of statistics and machine learning. When observations arrive from two sources that share some mutual information a related approach called the Canonical Component Analysis was developed [3].

This paper presents an efficient method that generalizes CCA to more than two views, Multi-view Canonical Correlation Analysis (MCCA). Defining a measure of cross-correlation for more than two random variables is not straightforward and many possible measures have been proposed [4]. Typical approaches define cross-correlation as a function of pairwise correlations between variables (for example the sum, product or sum of squares). Sum of correlations problem formulation, SUMCOR, was first studied in [2], where the optimization problem was formulated and a method to solve it was proposed (a generalization of the power method for standard eigenvalue problem which has been proved to converge in [1]). We will adopt and extend this approach since it is closely related to a known linear algebra problem which can be solved efficiently. The method proposed by Horst was designed to find a one-dimensional common representation.

The paper is structured in the following way: section 2 introduces the canonical correlation analysis, section 3 describes the multi-view CCA method, section 4 involves evaluation, followed by conclusions in section 5.

2 CANONICAL CORRELATION ANALYSIS

Canonical Correlation Analysis (CCA) is a dimensionality reduction technique similar to Principal Component Analysis (PCA), with an additional assumption that the data consists of feature vectors that arose from two sources (two views) that share some information. Examples include documents written in two different languages, textual information paired with images, a set of feature vectors computed from audio information and a set of feature vectors computed from the frames in a video recording, etc. Instead of looking for linear combinations of features that maximize the variance (PCA) we look for a linear combination of feature vectors from the first view and a linear combination for the second view, that are maximally correlated.

Formally, let $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be the set of n sample points (pairs of observation vectors) where $x_i \in \boldsymbol{R}^p$ and $y_i \in \boldsymbol{R}^q$ represent feature vectors from p (or q)-dimensional vector spaces. Let $\boldsymbol{X} = [x_1, \ldots, x_n]$ and let $\boldsymbol{Y} = [x_1, \ldots, x_n]$ be the matrices with observation vectors as columns, which are viewed as two samples of observations of two random vectors (X and Y). The idea is to find two linear functional (row vectors) $\alpha \in \boldsymbol{R}^p$ and $\beta \in \boldsymbol{R}^q$ so that the random variables αX and βY are maximally correlated (α and β map the random vectors to random variables, by computing weighted sums of vector components). By using the sample matrix notation $\boldsymbol{X}$ and $\boldsymbol{Y}$ this problem can be formulated as the following optimization problem:

$$\max_{\alpha \in R^p, \beta \in R^q} \alpha \boldsymbol{X}\, \boldsymbol{Y}' \beta'$$
$$s.t.$$
$$\alpha \boldsymbol{X}\boldsymbol{X}'\alpha' = 1$$
$$\beta \boldsymbol{Y}\boldsymbol{Y}'\beta' = 1$$

The optimization problem can be reduced to an eigenvalue problem and includes inverting the variance matrices $\boldsymbol{XX'}$ and $\boldsymbol{YY'}$. If they are not invertible one uses a regularization technique by replacing them with $(1-\kappa)\,\boldsymbol{XX'} + \kappa\,\boldsymbol{I}$, where $\kappa \in \boldsymbol{R}$ and $\boldsymbol{I}$ is the identity matrix.

A single canonical variable is usually inadequate in representing the original random vector, that is why one looks for $k-1$ other projection pairs $(\alpha_2, \beta_2), \ldots, (\alpha_k, \beta_k)$, so that α_i, and β_i are highly correlated and each α_i is uncorrelated to α_j for $j \neq i$ (analogously for β).

The method was extended to handle nonlinear relations between two random vectors in [6]. The approach is based on the observation that computing the canonical correlation vectors can be done by using solely the inner product information between sample vectors and that one can omit directly using any vector features. This enables the use of the dual problem formulation and application of the kernel trick [7]. For a given choice of kernel function with a corresponding feature map, this is

equivalent to first nonlinearly mapping both sets of sample vectors to a separate higher dimensional Hilbert spaces (the dimensions can be even infinite, for example when one uses a Gaussian kernel function) and look for linear relations in between the samples in those spaces. This usually makes the problem underdetermined – a high, possibly infinite, number of features and a smaller set of examples. To avoid overfitting, one needs to apply a regularization technique.

A typical regularization approach transforms the problem [7] into finding well cross-correlated projection vectors that have a high covariance as well. This enforces that the patterns discovered are not only well correlated across views but also well represented in the data.

3 MULTI-VIEW CANONICAL CORRELATION ANALYSIS

Consider a set of vectors $w_i \in \mathbb{R}^{n_i}, i = 1 \dots m$. For each random vector $\mathcal{X}_i$, with dimension n_i, we can define a univariate random variable $\mathcal{Z}_i$ as a linear combination random components: $\mathcal{Z}_i = w_i' \mathcal{X}_i$.We can now compute pairwise correlation coefficients for each pair of the variables $\mathcal{Z}_i$. The goal is to find the vectors w_i so that the sum of all pairwise correlations is the highest. One can prove that the optimization can be written as:

$$\max_{w_1,\dots,w_m} \sum_{i<j} w_i' X^i X^{j\prime} w_j$$

s.t.

$$w_i' X^i X^{i\prime} w_i = 1, \forall i,$$

where $X^i \in \mathbb{R}^{n_i \times n}$ are centered matrices of observations of random vectors X_i, containing n columns of sample vectors. Notice that every matrix X_i has the same number of columns – this corresponds to aligned sample assumption (column k of matrix X_i and column k of matrix X_j are aligned samples in two views).

We will reformulate the problem in dual form to make the problem feasible in the case of high dimensional data (e.g. text mining, where the number of features is the number of words encountered in the corpus) and with the use of the kernel trick make the solution more flexible than the linear model [9]. To express the problem in dual form we introduce new variables (we will also refer to them as dual variables), $\beta_i \in \mathbb{R}^n$, so$w_i = X^i \beta_i$. Let K_i be the kernel matrix computed on data X_i, which means that the element in the k-th row and l-th column of K_i is equal to:

$$< \phi_i(X_k^i), \phi_i(X_l^i) >$$

for some mapping $\phi_i : \mathbb{R}^{n_i} \to \mathcal{H}_i$ to some Hilbert space $\mathcal{H}_i$. The bracket denotes the inner product. The kernelized dual formulation of the problem is then:

$$\max_{\beta_1,\dots,\beta_m} \sum_{i<j} \beta_i' K^i K^{j\prime} \beta_j$$

s.t.

$$\beta_i' K^i K^{i\prime} \beta_i = 1, \forall i,$$

The constraints force the univariate random variables (linear combinations of the components of the original random vectors) to have unit variance.

If one of the kernel matrices is singular or is ill-conditioned the problem becomes numerically intractable. To remedy this problem one usually adds a low positive number on the diagonal elements of each kernel matrix in the variance equation (not the optimization criterion function).

By using Lagrangian multiplier techniques one can transform the constrained optimization problem to a generalized multivariate eigenvalue problem of the form:

$$\begin{bmatrix} \mathrm{A}_{11} & \cdots & \mathrm{A}_{1\mathrm{m}} \\ \vdots & \ddots & \vdots \\ \mathrm{A}_{\mathrm{m}1} & \cdots & \mathrm{A}_{\mathrm{mm}} \end{bmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} = \begin{pmatrix} \lambda_1 \beta_1 \\ \vdots \\ \lambda_m \beta_m \end{pmatrix},$$

Where A_{ij} are block matrices of dimension $n \times n$, β_i, are n-dimensional canonical vectors and λ_i are the generalized eigenvalues. Canonical vectors and the generalized eigenvalues are unknown and must be computed. The transformation of the kernelized dual to the multivariate eigenvalue problem can be conducted so that the λ_i become interpretable: their sum is directly proportional to the sum of correlations when one uses canonical projection vectors β_i to obtain univariate random variables from random vectors $\mathcal{X}_i$.

The solution (see [2]) to the multivariate generalized eigenvalue problem presented above can be found with a method similar to finding an eigenvector-eigenvalue pair in a square matrix by using power iteration method. The algorithm that finds the canonical projection vector requires a starting set of vectors which iteratively converge to a local optimum of the problem (several restarts with different starting vectors can prove useful). One must choose the number of iterations, denoted as *maxiter*, in advance or implement a stopping criterion.

So far we have discussed how to find a single canonical projection vector in each view. This is typically insufficient since too much information is discarded that way (In text mining for example, describing a document by a single number that represents the similarity of the document to the discovered latent vector). We denote the canonical vectors $\beta_1, \dots, \beta_m$ that we found as $\beta_1^1, \dots, \beta_m^1$ and try to find another set of concept vectors $\beta_1^2, \dots, \beta_m^2$ for which the sum of pairwise correlations is maximal with an additional constraint that they must be "different" from the first set. We can express this as a set of additional constraints:

$$\beta_i^{1\prime} K^i K^{i\prime} \beta_i^2 = 1, \forall i.$$

This forces the new set of vectors to be uncorrelated to the first. One can extend the problem to any number of sets of canonical projection vectors (each new set must be uncorrelated with all that have been discovered so far).

We can prove that the resulting optimization problem can be posed as a generalized multivariate eigenvalue problem and that it still satisfies the local convergence guarantees.

Algorithm 1: Horst algorithm

Input:matrices A_{ij}, initial vectors, α_i^0, $where\ i,j: 1 \dots m$
Output: $\alpha_1^{maxiter}, \dots, \alpha_m^{maxiter}$
$for\ i = 1\ to\ maxiter\ do$:
  $for\ j = 1\ to\ m\ do$:
    $\alpha_j^i \leftarrow \sum_k A_{j,k}\alpha_k^{i-1}$
    $\alpha_j^i \leftarrow \frac{\alpha_j^i}{\sqrt{\alpha_j^{i\prime}\alpha_j^i}}$
  $end\ for$
$end\ for$

4 EXPERIMENTS

The following section includes information retrieval experiments on the European Parliament corpus. We computed the semantic space for documents from ten different languages and compared the retrieval performance with two alternative approaches, namely Cross-lingual LSI and k-means clustering.

Subsection 4.1 details the experimental setup, subsection 4.2 describes the evaluation measure, subsection 4.3 describes the other alternative cross-lingual methods and subsection 4.4 offers an insight into the latent concept vectors discovered by MCCA.

4.1 DATA SET AND PREPROCESSING

Experiments were conducted on the EuroParl, Release v3 [8] data set and include Danish, German, English, Spanish, Italian, Dutch, Portuguese, Swedish, Finish and French language. We first removed all documents that had one translation or more missing. We split the corpus in an aligned set of documents, each representing a speech in the parliament. Cleaning the set resulted in 107.873 documents per language. We kept the first 100.000 for training and remaining 7.873 for testing or for testing. We then extracted the bag of words model for each language, where we kept all unigrams, bigrams and trigrams that occurred more than thirty times. This resulted in roughly 200.000-dimensional feature spaces for each language. Finally we computed the tf-idf weighting and normalized every document.

4.2 MATE RETRIEVAL

We used the aligned test set to measure the quality of the latent space representation. Given a test document q (view X) and its aligned document q' (view Y) and test set S' (view Y) we compute the window10 mate retrieval score in the following way: project q and S into the common semantic space, compute the similarities between projections of q and S and assign score 1 if q' is one of the top 10 most similar documents to q.

4.3 COMPARING TO CL-LSI AND K-MEANS CLUSTERING

We will compare our method with Cross-Lingual Latent Semantic Indexing and k-means clustering [11]. CL-LSI is an adaptation of LSI [10] for more than one view. The idea is to merge all document matrices into a single matrix Y by concatenating the aligned feature vectors. The matrix Y can then be used as the input for clustering or LSI. The final step when comparing to MCCA is to split the concept vectors into shorter concept vectors for each view in concordance with how the views were merged.

Language	k-means	LSI	MCCA
EN	0.7486	0.9129	0.9883
SP	0.745	0.2907	0.9855
GE	0.5927	0.8545	0.9778
IT	0.7448	0.9022	0.9836
DU	0.7136	0.9021	0.9835
DA	0.5357	0.854	0.9874
SW	0.5312	0.8623	0.988
PT	0.7511	0.9	0.9874
FR	0.7334	0.9116	0.9888
FI	0.4402	0.7737	0.983

Table 1 *Mate retrieval window 10*

We tested the performance of the three methods on mate retrieval with window10 on the 100-dimensional subspaces that the methods produced. For each source language we used all remaining nine languages, and averaged each score over all languages. Results imply that the concepts detected by MCCA in Table 1 are of higher quality than that of LSI and clustering. One way to explain this result is that MCCA takes into account that data come from several sources that share some mutual information, whereas the clustering and LSI approaches discard that information (after the views are concatenated we perform standard LSI which is "unaware" that features come from different views). LSI and CCA both find new latent features that are more informative (can detect synonyms), whereas the clustering approach uses the original features and thus performs worse than the other two methods.

4.4 CONCEPT VECTORS

The multivariate random variables from MCCA in our experiments correspond to document-vectors (in the bag of words representation) in different languages. We will now consider sets of words that are correlated between the two or more languages (sets of words that have a correlated pattern of appearance across the aligned corpus). We will assume that such sets approximate the notion of 'concepts' in each language, and that such concepts are the translation of each other. To illustrate the conceptual representation we have printed few of the most probable (most typical) words in each language for the first few components found from

the EuroParl (Figure 1). The words are sorted by their weights in the concept vectors.

5 CONCLUSIONS

We have presented an algorithm that can detect similar patterns across multiple domains. A straightforward approach would yield a cubic complexity in the number of samples whereas our implementation reduces the complexity to linear (quadratic if kernel methods are applied). We demonstrated the scalability and effectiveness on a large data set.

6 ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the IST Programme of the EC under PASCAL2 (IST-NoE-216886).

References

[1] Chu, M .T. and Watterson, L. J. (1993). On a Multivariate Eigenvalue Problem, Part I: Algebraic Theory and a Power Method. SIAM Journal on Scientific Computing, Vol.14 NO.5, 1089–1106.

[2] Horst, P. (1961). Relations among m sets of measures Psychometrika, 26, 129–149.

[3] Hotelling, H. (1935). The most predictable criterion Journal of Educational Psychology, 26, 139–142.

[4] Kettenring, J. R. (1971). Canonical analysis of several sets of variables Biometrika, 58, 433–451.

[5] Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R. A. (1990) Indexing by latent semantic analysis Journal of the Society for Information Science, 41(6), 391–407.

[6] Francis R. Bach and Michael I. Jordan (2001). Kernel Independent Component Analysis Report No. UCB/CSD-01-1166.

[7] Shawe-Taylor, J. and Cristianini, N. (2004). Kernel Methods for Pattern Analysis Cambridge University Press.

[8] Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation

[9] J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[10] D. Lewis, Y. Yang, T. Rose, F. Li, *Rcv1: A new benchmark collection for text categorization research*. Journal of Machine Learning Research (JMLR), 5:361–397, 2004.

[11] Dobsa J., Dalbelo Basic B. (2004). Comparison of Information retrieval techniques: Latent semantic indexing and concept indexing, Journal of Information and Organizational Sciences, 28(1-2), 1-15

DA	menneskerettighederne, menneskerettigheder, forretningsordenen, rusland, ndringsforslag, ã ndringsforslag
DE	menschenrechte, russland, posselt, menschenrechtsverletzungen, zusammenarbeit, nderungsantrag, verfahrensantrag
EN	amendment, amendments, russia, human rights, cooperation, resolution, of order
ES	enmienda, enmiendas, rusia, n de orden, de orden, reglamento, posselt
FI	ihmisoikeuksien, ihmisoikeuksia, tyã jã, tarkistuksen, tarkistusta, tarkistus, tarkistuksia
FR	amendement, amendements, posselt, russie, rã solution, russe, l amendement
IT	emendamenti, emendamento, risoluzione, russia, regolamento, cooperazione, bielorussia
NL	amendement, mensenrechten, amendementen, rusland, van orde, resolutie, samenwerking
PT	ponto de ordem, de ordem, alteraã, alteraã ã, direitos humanos, directiva, regimento
SV	resolutionen, ryssland, ordningsfrã, posselt, arbetsordningen, samarbete, ryska
DA	omdelt, dagsordenen, tak, er omdelt, protokollen fra, protokollen, strukturfondene
DE	tagesordnung, der tagesordnung, das protokoll der, kommissar, wurde verteilt, wurde verteilt gibt, haushalt
EN	commissioner, president commissioner, agenda, budget, commissioner the debate, commissioner the, item is
ES	comisario, distribuido, gracias, acta de la, comisaria, presupuesto, comisario el debate
FI	esityslistalla, kiitoksia, kiitos, esityslistalle, esityslistalla on, lissabonin, esityslistan
FR	merci, commissaire, jour appelle, du jour appelle, tã distribuã, ã tã distribuã, jeudi
IT	commissario, grazie, ringrazio, sono osservazioni, commissario la discussione, vi sono osservazioni, giovedã
NL	rondgedeeld, zijn rondgedeeld, de orde is, orde is, commissaris, orde is het, begroting
PT	obrigado, presidente senhor, conselho, acta, hã alguma, hã alguma observaã, comissã rio estã
SV	tack, r jag fã, rã det, kommissionsledamot, budget, har delats ut, kommissionã

Figure 1 *Two sets of latent vector*