Pair-Wise Cluster Analysis

David R. Hardoon[*]
Department of Computer Science
University College London
London
davidrh@me.com

Kristiaan Pelcksmn[†]
Division of Systems and Control
Dept. of Information Technology Uppsala University, Sweden
kristiaan.pelckmans@it.uu.se

Abstract

This paper studies the problem of learning clusters which are consistently present in different (continuously valued) representations of observed data. Our setup differs slightly from the standard approach of (co-) clustering as we use the fact that some form of 'labeling' becomes available in this setup: a cluster is only interesting if it has a counterpart in the alternative representation. The contribution of this paper is twofold: (i) the problem setting is explored and an analysis in terms of the PAC-Bayesian theorem is presented, (ii) a practical kernel-based algorithm is derived exploiting the inherent relation to Canonical Correlation Analysis (CCA), as well as its extension to multiple views. A content based information retrieval (CBIR) case study is presented on the multi-lingual aligned Europal document dataset which supports the above findings.

1 Introduction

Consider the setup where individual observations come in two different representations (x, y). This paper focuses on the questions: 'If we observe a new x, what can be said about the corresponding y, and vice versa?' While this abstract problem has obvious relations to classical supervised learning, its inherent symmetry relates it to unsupervised learning as well. This paper studies the above problem, specifying the properties to be predicted in terms of pre-specified membership functions. Figure (1) differentiates the above problem - termed PairWise Cluster Analysis (PWCA) - from the supervised, unsupervised, semi-supervised, transfer- and multiple-task learning [1] and self-taught learning [2]. The present learning strategy has direct relations to co-occurrence analysis, co-clustering [3], kernel Canonical Correlation Analysis (kCCA) [4] and has been motivated by the previous works of Pelckmans et al. [5] and Sim et al. [6] which explore an application in relating text corpus - microarray expression and multi-attribute co-clustering respectively.

The analysis given in Section 2 phrases the learning problem in terms of the PAC-Bayesian theorem, much in the spirit of the recent work of Seldin & Tishby [7]. Although, while the latter concerns density estimation for discrete variables, the presented ideas cover a spectrum of unsupervised learning (clustering). The analysis presented in [7] concerns, essentially, the same quantity $E_Q[\mathcal{R}(h)]$ as in subsection 2.1, equation (6), which characterizes how well some hypotheses Q aligns with the distribution underlying the data. Our extension to pairwise clustering is fundamentally different -

[*] www.davidroihardoon.com

[†] http://www.it.uu.se/katalog/kripe367

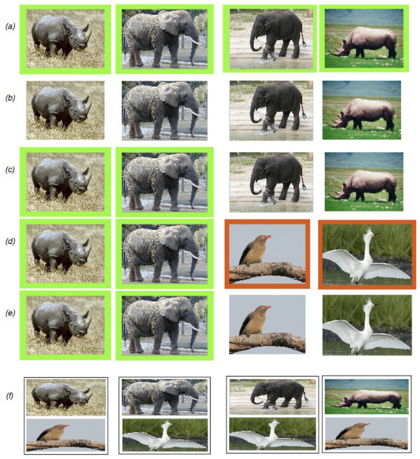

Figure 1: Pictorial representation of different learning paradigms, extending the picture in [2]. Suppose the aim is to discriminate elephants from rhinos. When a picture appears in a frame, a corresponding class-label is available. In cases: (a) supervised classification. (b) unsupervised learning. (c) semi-supervised learning. (d) transfer learning (the two different colors indicate two different learning tasks). (e) selftaught learning, and (f) pairwise cluster analysis (PWCA). Note that in the latter we try not to find the class labels themselves, but to recover the symbiotic relation between elephant-egret, and rhino-oxpeckers. Specifically, the presence of oxpeckers might help us in predicting the presence of a rhino, and vice versa.

incorporating a notion of prediction 'loss' - while the relation of Kuller-Leibler (KL) divergence and the norm of an hypotheses establishes a relation with the learning algorithm.

Section 3 (i) derives an effective learning algorithm, boiling down to a quadratic (or a generalized) eigenvalue problem. This learning machine is closely related to kernel Canonical Correlation Analysis (see e.g. [8, 4] and references therein). Empirical (ii) evidence for this learning paradigm, and the proposed algorithm is then presented. We proceed to demonstrated the benefit of learning structure within the data on a multi-lingual text-corpora [9]. Section 4 indicates a number of open questions.

2 A Generic Analysis using the PAC-Bayes Theorem

Consider a function $h_r : \{x\} \to [0,1]$ that verifies, for a given problem setting, how good a certain 'rule' r performs on a sample x. The goal of a learning algorithm is to find the best rule r in a given set of plausible rules (the hypothesis set). Then, learning proceeds by collecting a dataset $\{X_i\}_{i=1}^n$ of n observations assumed to be sampled independently from identical distributions (i.i.d)[1]. The empirical risk $\mathcal{R}_n(h_r)$ and the actual risk $\mathcal{R}(h_r)$ of an 'hypothesis' $h_r \in \mathcal{H}$ is defined as

$$\begin{cases} \mathcal{R}_n(h_r) = \frac{1}{n}\sum_{i=1}^n h_r(X_i) \\ \mathcal{R}(h_r) = \mathbb{E}[h_r(X)], \end{cases} \tag{1}$$

where the expectation $\mathbb{E}[\cdot]$ concerns the fixed, unknown distribution underlying the n i.i.d observations. For supervised learning problems, (informally) an observation x consists typically of a couple (z, y) with a covariate z and an 'output' y. Then h_r is often rephrased as $h_r(x) = \ell(y - r(z))$, where $\ell : \mathbb{R} \to [0,1]$ is the 'prediction loss' between the actual observation y and its prediction $r(z)$. In a Bayesian context, we assume that the hypothesis $h_r \in \mathcal{H}$ are also 'stochastic' elements[2], possessing some notion of likelihood, say $Q : \mathcal{H} \to [0,1]$ such that $\int_{\mathcal{H}} Q(h_r)dh = 1$. Consider at first the case where $\mathcal{H}$ is finite, we are interested in what happens on functions $E_Q[h_r(x)]$, which is defined as

$$E_Q\left[h_r(x)\right] = \sum_{h_r \in \mathcal{H}} h_r(x) Q(h_r). \tag{2}$$

If $|\mathcal{H}|$ is infinite, then the sum can be replaced by an integral as usual, or $E_Q\left[h_r(x)\right] = \int_{\mathcal{H}} h_r(x)Q(h_r)dh_r$. In the analysis we will assume $|\mathcal{H}| < \infty$ in order to avoid technical issues. Note that this is not quite a regular (well-known) expectation $\mathbb{E}[\cdot]$ as before. Now let the Kullback-Leibler distance be defined for each $0 < p, q < 1$ be defined as $\mathrm{KL}(q,p) = q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}$,

[1] We will use the convention to denote stochastic variables as capital letters, e.g. $X, Y, \ldots$, while deterministic quantities are denoted in lower case, e.g. $h, f, i, x, y, n, \ldots$.

[2] In a PAC-Bayesian context, we will merely consider weighted sums of the elements in $\mathcal{H}$, rather than assuming a truly Bayesian setup.

where $\log(\cdot)$ denote the natural logarithm. Let the function $P : \mathcal{H} \to [0, 1]$ be the prior weighting function over $\mathcal{H}$. If $Q : \mathcal{H} \to [0, 1]$ and $P : \mathcal{H} \to [0, 1]$ are two functions, we extend the definition as

$$\mathrm{KL}(Q, P) = \sum_{h_r \in \mathcal{H}} Q(h_r) \log \frac{Q(h_r)}{P(h_r)}. \tag{3}$$

We state the PAC-Bayes theorem as in [10]:

Theorem 1 *For $\delta > 0$ and for $n \geq 8$, we have that with probability exceeding $1 - \delta$ we have that for all $Q : \mathcal{H} \to [0, 1]$ the following inequality holds:*

$$\mathrm{KL}\left(E_Q[\mathcal{R}_n(h_r)], E_Q[\mathcal{R}(h_r)]\right) \leq \frac{\mathrm{KL}(Q, P) + \log \frac{1}{\delta} + \log(2\sqrt{n})}{n}. \tag{4}$$

Specifically, this holds for a Q_n found by an algorithm based on the n i.i.d. observations. Note that this result is currently the most tight inequality, refining the ideas presented in [11]. While till date most applications are found in the context of supervised learning, we will argue in the following that this theorem finds a 'natural' application towards unsupervised learning.

2.1 An Application of PAC-Bayes Towards Clustering

In what follows, assume that the n i.i.d. samples $\{X_i\}_{i=1}^n$ take values in a bounded set in $S \subset \mathbb{R}^d$ for a given $d \in \mathbb{N}$. In order to use the PAC-Bayes result to the generic application of clustering, we need to specify the loss function $\ell : \mathbb{R}^d \to [0, 1]$ of interest. A 'cluster', represented as an indicator function $h : \mathbb{R}^d \to \{0, 1\}$, is understood here as a member of a user-specified set of indicator functions $\mathcal{H} = \left\{h : \mathbb{R}^d \to \{0, 1\}\right\}$. Formally, one defines for a set $c \subset \mathbb{R}^d$

$$h_c(x) = I(x \in c) = \begin{cases} 1 & x \in c \\ 0 & x \notin c. \end{cases} \tag{5}$$

Now, we look a bit closer at what the term $E_Q[\mathcal{R}(h_c)]$ represents in this context.

$$E_Q[\mathcal{R}(h_c)] = \sum_{h_c \in \mathcal{H}} \mathbb{P}(X \in c) Q(h_c) = \mathbb{E}\left[\sum_{h_c \in \mathcal{H}} h_c(X) Q(h_c)\right], \tag{6}$$

where the second equality holds by linearity of the expectation, and where $\mathbb{P}$ denotes the probability rules underlying the data. Consequently, the term $E_Q[\mathcal{R}(h)]$ characterizes how well Q *aligns* with the distribution underlying the data. Assume that the $\mathcal{H}$ is designed such that all sets c corresponding to a $h_c \in \mathcal{H}$ (i) cover the space S and (ii) are disjunct.

The function $P : \mathcal{H} \to [0, 1]$ is the prior weighting function (think of it as a 'prior distribution' over $\mathcal{H}$). In general, it is up to the user in a specific application to decide how to design $(\mathcal{H}, P)$: it is good practice to make it equally likely for each hypothesis $h \in \mathcal{H}$ to explain the data by itself, -suggesting a uniform prior P over this set $\mathcal{H}$- while the result should be useful for the application in mind. Assume for example that all probability mass (underlying the samples) concentrates in the set corresponding with a single h_c, and $Q(h_c) = I(i = j)$, then this measure equals 1. On the other hand, if all samples are equally distributed over the $|\mathcal{H}|$ sets $h_c \in \mathcal{H}$, the measure equals $\frac{1}{|\mathcal{H}|}$. This motivates the naming of $E_Q[\mathcal{R}(h)]$ as the *explanatory power* of $(\mathcal{H}, Q)$. Specifically, if $\mathcal{H} = \{I(x \in [-1, 1]^d)\}$, the explanatory power of $(\mathcal{H}, Q)$ is 1, but it however is not very useful, surprising nor *falsifiable*.

We argue that this PAC-Bayesian interpretation to clustering is often 'natural' because of three reasons. (i) The present analysis does not need to recover the density function underlying the data, a feature which is highly desirable if working with high-dimensional data. (ii) The set of 'underlying' clusters is not recovered exactly, nor assumed to exists in reality. The actual stochastic rules underlying the observed data only say how well the hypothesis clustering 'explains' the data. When dealing with data arising from complex processes the assumption of a 'true clustering' is often an oversimplification. (iii) The characterization of performance of the found rule Q_n in terms of its deviation from the prior P is desirable if clustering is meant for looking for 'consistent' irregularities.

Specifically, if the result Q_n is not what we (more or less) expected before seeing the data, substantial empirical evidence should be presented motivating this property. Those reasons differentiate the approach substantially from approaches based on density estimation, or on mixtures of distributions. Remark that this description of explanatory power is strongly related to the ideas presented in [12]. The following clustering algorithm is then motivated by application of the PAC-Bayesian theory:

$$Q_n = \arg\min_{Q} E_Q[\mathcal{R}_n(h)] \text{ s.t. } \ \mathrm{KL}(Q, P) \leq \omega_n, \tag{7}$$

where $\omega_n > 0$. This objective is also motivated from an information theoretical approach to clustering, as e.g. in [3].

2.2 An Application of PAC-Bayes Towards Pairwise Clustering

Now we explain how the above insights lead to an analysis of the pairwise clustering setup. Let again $\mathbb{Z}$ and $\mathbb{Y}$ denote respectively the two domains of interest in which pairwise observations (x, y) are made. A first approach would be to rephrase the pairwise clustering problem as a standard clustering approach, where instead of the class of indicator functions $\mathcal{H}_f \subset \{f : \mathbb{Z} \to [0, 1]\}$ in the first domain, one studies the cross-product of this class with the class of indicator functions in the other domain $\mathcal{H}^{f,g} = \mathcal{H}_f \times \mathcal{H}_g$, or

$$\mathcal{H}^{f,g} \subset \Big\{ h = (f_h, g_h) \ \Big| \ f_h : \mathbb{Z} \to [0, 1], g_h : \mathbb{Y} \to [0, 1] \Big\}. \tag{8}$$

However, the reasoning in the introduction suggests another route. To see this, we formalize the intuition of the pairwise observation (x, y) being a target for prediction: (i) let $z \in \mathbb{Z}$ represent the part of a sample $x = (z, y)$ which might be used to predict (a property) of the (unobserved) $y \in \mathbb{Y}$; and/or (ii) given $y \in \mathbb{Y}$, predict (a property) of the corresponding (unobserved) $z \in \mathbb{Z}$. Given a set $\mathcal{H}^{f,g}$: the knowledge of the 'cluster' to which X belongs, will be used to predict the cluster memberships of the corresponding y.

We will say that f_h *explains* $z \in \mathbb{Z}$ if $f_h(z) = 1$, and similarly that g_h *explains* $y \in \mathbb{Y}$ if $g_h(y) = 1$. In an ideal case, one would be able to associate exactly one distinct $f_h \in \mathcal{H}_f$ to every $g_h \in \mathcal{H}_g$ (i.e. describe a permutation). As such, one could predict the cluster g_h containing y corresponding to a given z. In the worst case, the choice of g that explains y is independent of z being explained by f. The pairwise clustering setup however differs from such a multi-class classification (structured output prediction) task as it is essentially symmetric: a given z is used to predict (cluster membership of) the corresponding y, and a given y is used to predict (cluster memberships of) the corresponding x. Now, a pairwise cluster $h = (f, g) \in \mathcal{H}^{f,g}$ was useful for a sample $(z, y) \in \mathbb{Z} \times \mathbb{Y}$ if $f(z) = g(y)$. Alternatively, a pairwise cluster $c = (f, g)$ *contradicts* a sample if $f(z) \neq g(y)$. This motivates the following risk function

$$\begin{cases} \mathcal{R}_n(h) = \frac{1}{n} \sum_{i=1}^n I(f_h(Z_i) \neq g_h(Y_i)) \\ \mathcal{R}(h) = \mathbb{P}\left(f_h(Z) \neq g_h(Y)\right), \end{cases} \tag{9}$$

defined again in an 'empirical' and an 'actual' flavor. This definition measures how many (for how large a probability mass) datapoints are contradicted by a pairwise cluster $h = (f_h, g_h)$. Now the term $E_Q[\mathcal{R}(h)]$ becomes

$$E_Q[\mathcal{R}(h)] = \sum_{h \in \mathcal{H}^{f,g}} \mathbb{P}\left(f_h(Z) \neq g_h(Y)\right) Q(h), \tag{10}$$

which basically captures how many mistakes are made when focussing on the subset of $\mathcal{H}^{f,g}$ as directed by Q. This motivates the following practical approach: (i) given a dataset $\{X_i = (Z_i, Y_i)\}_{i=1}^n$, with the elements taking values in $\mathbb{Z} \times \mathbb{Y}$, and (ii) a a set $\mathcal{H}^{f,g}$ of pairwise clusters represented as $h = (f, g)$, and a 'prior' weighting function $P : \mathcal{H}^{f,g} \to [0, 1]$, then we aim to find a new weighting function $Q_n : \mathcal{H}^{f,g} \to [0, 1]$ which is not too different from P, and which aligns well with the probability rules underlying the data as

$$Q_* = \arg\min_{Q} E_Q(\mathcal{R}(h_c)) \text{ s.t. } \ \mathrm{KL}(Q, P) \leq \omega, \tag{11}$$

where $\omega > 0$. The PAC-Bayes theorem now guarantees that this problem is approximatively solved based on the data as

$$Q'_n = \arg\min_{Q} E_Q(\mathcal{R}_n(h_c)) \text{ s.t. } \ \mathrm{KL}(Q, P) \leq \omega, \tag{12}$$

where $\omega > 0$. The resulting Q'_n will emphasize the pairwise clusters which are most often consistent with the data. Here we have a natural trade-off between specificity and accuracy, regulated by ω_n. If ω_n were small, the solution Q'_n cannot deviate from the uniform distributions over all pairwise clusters in $\mathcal{H}^{f,g}$, but then many different pairwise clusters will contradict on different samples, leading in turn to low explanatory power. On the other hand, allowing for arbitrary Q'_n will explain the individual samples fairly well (allowing a single pairwise cluster per sample), but the PAC-Bayesian result will not guarantee accuracy of the result anymore.

We now express the 'regularization term' $\mathrm{KL}(Q, P)$ in a more convenient form.

Proposition 1 (Bound to K.-L. Divergence) *Assume $|\mathcal{H}| < \infty$ and $P(h) = \frac{1}{|\mathcal{H}|}$ for all $h \in \mathcal{H}$, then*

$$\mathrm{KL}(Q, P) \leq \log \sum_{h \in \mathcal{H}} Q^2(h) + \log(|\mathcal{H}|). \tag{13}$$

This is a consequence of the following inequality on the entropy of a vector $p \in]0, 1[^d$ with $1^T p = 1$

$$\mathbf{h}(p) = \sum_{i=1} p_d \log(p_d) \leq \log\left(\sum_{i=1}^{d} p_i^2\right), \tag{14}$$

by application of Jensen's inequality. Let $s^Q \in [0, 1]^{|\mathcal{H}|}$ be a vector representing the function Q where $s_i^Q = Q(h_i)$ (enumerating the different elements $h_i \in \mathcal{H}$), then

$$s_n^Q = \underset{s^Q \geq 0_n \sum_i s_i^Q = 1}{\arg\min} \|s^Q\|_2 \text{ s.t. } E_Q[\mathcal{R}_n(h)] = 0. \tag{15}$$

implementing the socalled *realizable* case (as in the theory of Support Vector Machines). The optimal solution Q_n will try to find as many pairwise clusters as possible which are not contradicting the given data. We illustrate this notion in figure 2. In the ideal case, all observations are explained. In more realistic cases, merely a few pairwise clusters are found (i.e., the set $\{h \in \mathcal{H} : Q(h) > 0\}$ contains only a few elements).

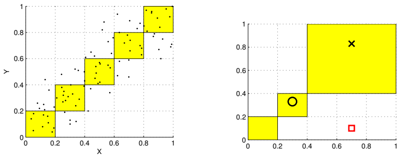

Figure 2: Schematic representation of all pairwise clusters in a hypothesis space $\mathcal{H}$ based on the 5 disjunct intervals $d + [0, 0.2]$ in either domain (dotted lines). The dots $(X, Y) \in \mathbb{R} \times \mathbb{R}$ represent samples from an underlying distribution. Suppose the different hypothesis can be factorized as $h_c = (f, g)$, where $f : \mathbb{R} \to [0, 1]$ and $g : \mathbb{R} \to [0, 1]$, being the corresponding indicator functions in either domain. This means that there are 25 possible different pairwise clusters h_c (dotted squares), or $|\mathcal{H}^{f,g}| = 25$, (a) about 70% of the observations (dots) do not contradict the 5 pairwise clusters (yellow squares) simultaneously; (b) Only one sample ('□') contradicts the shown pairwise cluster h_c (yellow squares), while the other two ('$\circ$' and '$\times$') are consistent with h_c.

We extend this model to account for infinite $\mathcal{H}$, defined as $h = (\delta_z, \delta_y)$ for each $(z, y) \in \mathbb{Z} \times \mathbb{Y}$, and where δ_x denotes the Dirac delta. When extending the formulation in order to deal with infinite hypothesis spaces $\mathcal{H}^{f,g}$, we replace vectors s_Q by functions $Q : \mathcal{H} \to \mathbb{R}^+$, which (for convenience) are assumed to be elements of a Hilbert spaces $\mathbf{H}$. This space is equipped with a corresponding inner-product (reproducing kernel) $k : \mathbf{H} \times \mathbf{H} \to \mathbb{R}$, implicitly defining $\mathcal{H}$ and P. Note that $Q(h) \geq 0$ for all $h \in \mathcal{H}$, and $\int_{\mathcal{H}} Q(h)dh = 1$. This motivates the replacement of the term $\mathrm{KL}(Q, P)$ by $\|Q\|_{\mathbb{H}}$. As such (12) is equivalent (up to normalization) to

$$Q''_n = \arg\min_{Q} \|Q\|_{\mathbf{H}} \text{ s.t. } E_Q[\mathcal{R}_n(h)] = 0. \tag{16}$$

where $Q''_n(h) \geq 0$ for all $h \in \mathcal{H}$, and $\int_{\mathcal{H}} Q''_n(h)dh = 1$. Note that for the majority of pairwise clusters no data is sampled contradicting the cluster, and a smooth transition of Q inbetween the sample becomes possible. In the remainder we will assume the relevant Hilbert space $\mathbf{H}$ can be decomposed additively uniquely as $\mathbf{H}_{\mathbb{Z}} \otimes \mathbf{H}_{\mathbb{Y}}$, and the norm of a function Q can then be written as $\|Q\|_{\mathbf{H}}^2 = \|F\|_{\mathbf{H}_{\mathbb{Z}}}^2 + \|G\|_{\mathbf{H}_{\mathbb{Y}}}^2$. Assume $\mathcal{H}^{f,g}$ contains all pairwise clusters $h = (\delta_z, \delta_y)$ for all $(z, y) \in \mathbb{Z} \times \mathbb{Y}$ and δ the Dirac delta. Under the assumtion no ties occur in the data, problem (17) is

$$(F_n, G_n) = \underset{F,G}{\arg\min} \|F\|_{\mathbf{H}_{\mathbb{Z}}}^2 + \|G\|_{\mathbf{H}_{\mathbb{Y}}}^2 \text{ s.t. } F_i = G_i, \ \forall i = 1, \dots, n. \tag{17}$$

enforcing that $F(h) = G(h)$ for all $h \in \mathcal{H}^{f,g}$, and enforcing again that $F(h) \geq 0$ for all $h \in \mathcal{H}^{f,g}$ as well as that $\int_{\mathcal{H}^{f,g}} F(h)dh = 1$. Here $F_i = F(\delta_{Z_i})$ and $G_i(\delta_{Y_i}) = Q((\delta_{Z_i}, \delta_{Y_i}))$ for all $i = 1, \dots, n$. The next section shows how to solve this problem, relaxing the (in)equality constraints.

3 Kernel PairWise Component Analysis

3.1 PWCA for paired Observations

This section studies how the learning problem (17) is solved (approximatively) by an efficient algorithm. Let $X^a = (X_1^T, \dots, X_n^T)^T \in \mathbb{R}^{\ell \times m}$ and $Y^b = (Y_1^T, \dots, Y_n^T)^T \in \mathbb{R}^{\ell \times n}$ be matrices where ℓ is the number of samples and m, n are the number of attributes/features for the first and second representation respectively. The functions Q are parametrised as $F_{\mathbf{v}_c}(z) = \mathbf{v}_c^T z$ and $G_{\mathbf{w}_c}(y) = \mathbf{w}_c^T y$. The inequalities $Q(h) \geq 0$ are enforced by representing this as $Q(h) = F^2(f) = G^2(h)$ for all $h \in \mathcal{H}^{f,g}$. This is imposed by enforcing $c_i = \sqrt{Q((\delta_{Z_i}, \delta_{Y_i}))} = F(\delta_{Y_i}) = G(\delta_{Y_i})$, then $\int_{\mathcal{H}} Q(h)dh = 1$ is enforced by imposing the constraint $\mathbf{c}'\mathbf{c} = 1$ (similarly, maximizing $\mathbf{c}'\mathbf{c}$). As such (12) becomes

$$\max_{\mathbf{c} \in \mathbb{R}^\ell, \mathbf{v}_c \in \mathbb{R}^m, \mathbf{w}_c \in \mathbb{R}^n} \mathbf{c}'\mathbf{c} - \gamma(\mathbf{w}_c'\mathbf{w}_c + \mathbf{v}_c'\mathbf{v}_c), \tag{18}$$

where A' is the transpose of matrix, or vector, A and such that $\mathbf{c}_i = X_{a,i}\mathbf{w}_c = Y_{b,i}\mathbf{v}_c$, for $i = 1, \dots, \ell$. Associating Lagrange multipliers α_i, β_i to each of the ℓ constraints gives the following Lagrangian

$$\mathcal{L} \;=\; \frac{1}{2}\mathbf{c}'\mathbf{c} - \frac{\gamma}{2}(\mathbf{w}_c'\mathbf{w}_c + \mathbf{v}_c'\mathbf{v}_c) - \boldsymbol{\alpha}'(\mathbf{c} - X_a\mathbf{w}_c) - \boldsymbol{\beta}'(\mathbf{c} - Y_b\mathbf{v}_c). \tag{19}$$

Taking derivatives of equation (19) with respect to $\mathbf{w}_c, \mathbf{v}_c, \mathbf{c}$ and setting to zero give the following conditions for optimality as

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}_c} = \mathbf{0} \rightarrow \mathbf{w}_c = \frac{1}{\gamma}X_a'\boldsymbol{\alpha}, \ \frac{\partial \mathcal{L}}{\partial \mathbf{v}_c} = \mathbf{0} \rightarrow \mathbf{v}_c = \frac{1}{\gamma}Y_b'\boldsymbol{\beta}, \ \frac{\partial \mathcal{L}}{\partial \mathbf{c}} = \mathbf{0} \rightarrow \mathbf{c} = (\boldsymbol{\alpha} + \boldsymbol{\beta}).$$

Setting back into the optimisation in equation (18) gives the following dual problem

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^\ell, \boldsymbol{\beta} \in \mathbb{R}^\ell} \mathcal{J} = \frac{1}{2}(\boldsymbol{\alpha} + \boldsymbol{\beta})'(\boldsymbol{\alpha} + \boldsymbol{\beta}) - \frac{1}{2\gamma}(\boldsymbol{\alpha}'K_a\boldsymbol{\alpha} + \boldsymbol{\beta}'K_b\boldsymbol{\beta}),$$

where $K_a = X_aX_a'$ and $K_b = Y_bY_b'$ are the kernel matrices. Taking derivatives and setting to zero shows that $\mathcal{J}$ achieves a (local) optimum when

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial \boldsymbol{\alpha}} = \mathbf{0} \quad &\rightarrow \quad \gamma(\boldsymbol{\alpha} + \boldsymbol{\beta}) = K_a\boldsymbol{\alpha} \\ \frac{\partial \mathcal{J}}{\partial \boldsymbol{\beta}} = \mathbf{0} \quad &\rightarrow \quad \gamma(\boldsymbol{\alpha} + \boldsymbol{\beta}) = K_b\boldsymbol{\beta}. \end{aligned} \tag{20}$$

We are able to observe that at optimum $K_a\boldsymbol{\alpha} = K_b\boldsymbol{\beta}$, which illustrates a direct relationship to KCCA condition. Due to limited space we do not explore the relationship to KCCA within the scope of this manuscript. Equation (20) can be rewritten as

$$\begin{bmatrix} K_a & 0_\ell \\ 0_\ell & K_b \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} = \gamma \begin{bmatrix} I_\ell & I_\ell \\ I_\ell & I_\ell \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix}, \tag{21}$$

where I_ℓ is the identity matrix and 0_ℓ is a matrix of zeros, both of size $\ell \times \ell$. This equation may be solved as a generalized eigenvalue problem in the form of $A\mathbf{x} = \lambda B\mathbf{x}$. Alternatively, we observe that by setting $\boldsymbol{\beta} = \left(\frac{1}{\gamma}K_a - I\right)\boldsymbol{\alpha}$, we can express $\frac{1}{\gamma}K_a\boldsymbol{\alpha} = \frac{1}{\gamma^2}K_bK_a\boldsymbol{\alpha} - \frac{1}{\gamma}K_b\boldsymbol{\alpha}$, which results in the following generalized eigenvalue problem for $\boldsymbol{\alpha}$

$$K_bK_a\boldsymbol{\alpha} = \gamma\left(K_a + K_b\right)\boldsymbol{\alpha}, \tag{22}$$

and by setting R to be the Cholesky decomposition of K_bK_a such that $K_bK_a = RR'$ we obtain the following symmetric eigenvalue problem

$$I_\ell\boldsymbol{\alpha} = \gamma R^{-1}\left(K_a + K_b\right)R^{-1'}\boldsymbol{\alpha}.$$

It may be necessary to regularize equation (21) with some small value τ on the diagonal. This will result in our optimisation being rewritten as

$$\begin{bmatrix} K_a & 0_\ell \\ 0_\ell & K_b \end{bmatrix}\begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} = \gamma\begin{bmatrix} I_\ell(1+\tau) & I_\ell \\ I_\ell & I_\ell(1+\tau) \end{bmatrix}\begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix}.$$

Furthermore, the above eigenvalue problem can be written as $\boldsymbol{\beta} = \left(\frac{1}{\gamma}K_a - \tau I_\ell\right)\boldsymbol{\alpha}$ and

$$K_bK_a\boldsymbol{\alpha} = \gamma^2(I_\ell - \tau^2 I_\ell)\boldsymbol{\alpha} + \gamma(\tau I_\ell K_a + \tau I_\ell K_b)\boldsymbol{\alpha},$$

which can be solved as a quadratic eigenvalue problem. It follows from the conditions for optimality that a new sample $(\bar{\mathbf{x}}_a, \bar{\mathbf{y}}_b)$ can be projected in the learnt semantic space by the functions

$$\begin{cases} F(\bar{\mathbf{x}}_a) = \mathbf{w}_c'\bar{\mathbf{x}}_a = \frac{1}{\gamma}\boldsymbol{\alpha}'K_a(\mathbf{x}_a, \bar{\mathbf{x}}_a), \\ G(\bar{\mathbf{y}}_b) = \mathbf{v}_c'\bar{\mathbf{y}}_b = \frac{1}{\gamma}\boldsymbol{\beta}'K_b(\mathbf{y}_b, \bar{\mathbf{y}}_b). \end{cases}$$

Then it is also reasonable to assign the sample $(\bar{\mathbf{x}}_a, \bar{\mathbf{y}}_b)$ to the cluster $(1, \ldots, \ell)$ which has highest (absolute) factors $|F(\bar{\mathbf{x}}_a)|_1^\ell$ and $|G(\bar{\mathbf{y}}_b)|_1^\ell$ respectively.

3.2 PWCA for Multiview Observations

In this section we generalize our methodology to multiple views. Expressing optimization in equation (18) for three sources gives

$$\max_{\mathbf{c}\in\mathbb{R}^\ell, \mathbf{w}_c\in\mathbb{R}^m, \mathbf{v}_c\in\mathbb{R}^n, \mathbf{z}_c\in\mathbb{R}^s} \frac{1}{2}\mathbf{c}'\mathbf{c} - \frac{\gamma}{2}(\mathbf{w}_c'\mathbf{w}_c + \mathbf{v}_c'\mathbf{v}_c + \mathbf{z}_c'\mathbf{z}_c), \tag{23}$$

such that $c_i = X_{a,i}\mathbf{w}_c = X_{b,i}\mathbf{v}_c = X_{c,i}\mathbf{z}_c$, for $i = 1, \ldots, \ell$. Taking derivatives of equation (23) with respect to $\mathbf{w}_c, \mathbf{v}_c, \mathbf{z}_c, \mathbf{c}$ and setting to zero will give the conditions for optimality. Substituting these conditions back into equation (23) gives the following dual problem

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^\ell, \boldsymbol{\beta}\in\mathbb{R}^\ell, \boldsymbol{\nu}\in\mathbb{R}^\ell} \mathcal{J} = \frac{1}{2}(\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\nu})'(\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\nu}) - \frac{1}{2\gamma}(\boldsymbol{\alpha}'K_a\boldsymbol{\alpha} + \boldsymbol{\beta}'K_b\boldsymbol{\beta} + \boldsymbol{\nu}'K_c\boldsymbol{\nu}),$$

where $K_a = X_aX_a'$, $K_b = X_bX_b'$ and $K_c = X_cX_c'$ are the kernel matrices. Taking derivatives and setting to zero shows that $\mathcal{J}$ achieves a (local) optimum when

$$\frac{\partial\mathcal{J}}{\partial\boldsymbol{\alpha}} = \mathbf{0} \rightarrow \gamma(\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\nu}) = K_a\boldsymbol{\alpha},\ \frac{\partial\mathcal{J}}{\partial\boldsymbol{\beta}} = \mathbf{0} \rightarrow \gamma(\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\nu}) = K_b\boldsymbol{\beta},\ \frac{\partial\mathcal{J}}{\partial\boldsymbol{\nu}} = \mathbf{0} \rightarrow \gamma(\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\nu}) = K_c\boldsymbol{\nu}.$$

which can be rewritten as

$$\begin{bmatrix} K_a & 0_\ell & 0_\ell \\ 0_\ell & K_b & 0_\ell \\ 0_\ell & 0_\ell & K_c \end{bmatrix}\begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \\ \boldsymbol{\nu} \end{bmatrix} = \gamma\begin{bmatrix} I_\ell & I_\ell & I_\ell \\ I_\ell & I_\ell & I_\ell \\ I_\ell & I_\ell & I_\ell \end{bmatrix}\begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \\ \boldsymbol{\nu} \end{bmatrix},$$

where again I_ℓ is the identity matrix and 0_ℓ is a matrix of zeros, both of size $\ell\times\ell$. Therefore, without loss of generality, we can extend this to multiple $\mathbf{i} = 1, \ldots, s$ views, where $s \geq 2$, similarly to the previously proposed multi-view extension for CCA by [8], such that

$$\begin{bmatrix} K_1 & \ldots & 0_\ell \\ \vdots & \ddots & \vdots \\ 0_\ell & \ldots & K_s \end{bmatrix}\begin{bmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_s \end{bmatrix} = \gamma\begin{bmatrix} I_\ell & \ldots & I_\ell \\ \vdots & \ddots & \vdots \\ I_\ell & \ldots & I_\ell \end{bmatrix}\begin{bmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_s \end{bmatrix}.$$

This equation may be solved as a generalized eigenvalue problem in the form of $A\mathbf{x} = \lambda B\mathbf{x}$.

4 Experiments of PWCA on Europal

We proceed to compare PWCA to KCCA for a mate-retrieval task [13, 14, 15, 16], i.e. given a document query $\mathbf{q}_i$ in language x to retrieve the (exact) matching document in the paired language y. For this purpose we use the multi-lingual Europal dataset [9], which has a total of 11968 aligned documents. We use the following eight languages with the number of features/words in brackets; da - Danish (78720), de - German (153499), en - English (60369), es - Spanish (171821), it - Italian (66548), nl - Dutch (105318), pt - Portuguese (66922) and sv - Swedish (51116). We use linear kernels throughout and arbitrarily set the regularization parameter to $\tau = 0.01$ for both methods. Finally, the performance is evaluated using Average Precision (AP) [17] which is computed as $AP = \frac{1}{\ell}\sum_{i=1}^{\ell}\frac{1}{I_i}$ where I_i is the rank location of the exact paired document for query document $\mathbf{q}_i$. Therefore $AP = 0.5$ indicates that the paired document is on average situated at location $I = 2$. We select the rank by sorting the, absolute, inner products values of $F(\mathbf{q}_i)'G(\mathbf{y}_j)$ (as well as for $F(\mathbf{x}_i)'G(\mathbf{q}_j)$) for all possible paired test documents, i.e. we rank the retrieved documents according to their similarity (in the learnt space) with our query. In our experiments we use the CCA formulation as proposed by [8] for both pair- and multi-view.

In the first of our two experiments, for each pairing combination of languages, we randomly select 500 paired-documents for training and 5000 for testing. The analysis has been repeated 10 times and averaged across. The results given in table 1 are the AP averaged across of all possible language-pair combinations for the language indicated in the column (i.e. column *da* is the average of all the language pairing with *da - xx*). We are able to observe that PWCA is able to perform, on average, on a par with KCCA. The mean AP across all languages for KCCA is 0.4435 whereas for PWCA it is 0.4459.

Table 1: We compare KCCA and PWCA on a bilingual mate-retrieval task (see text for language abbreviation). The reported results are the AP for retrieving the exact paired document in another language, averaged across all possible language-pair combination for the language indicated in the column. The results are averaged over 10 repeats of the analysis.

	da	de	en	es	it	nl	pt	sv
KCCA	0.4174	0.3839	**0.4979**	0.4243	**0.4572**	0.4023	**0.4939**	0.4714
PWCA	**0.4294**	**0.4416**	0.4747	**0.4344**	0.4368	**0.4111**	0.4679	**0.4716**

In the second experiment we extend the previous analysis to a trilingual mate-retrieval task, i.e. we train on an aligned document corpus from three languages whereas during testing we compute the mean average precision of all the individual pair-wise mate-retrieval tasks (of the three languages). In other words, we train on the trilingual alignment of *da-de-en* while we test the query retrieval on the bilingual task of *da-de, da-en, de-en*. In this experiment we randomly select 500 tripartite-documents for training and 2000 for testing. Due to increased complexity we only repeat the analysis, for each 3 language combination, once. The results given in table 2, as in the previous table, are the mean average precision for the language stated in the column and all its possible tripartite combinations (without repetition, i.e. for example; *da-da-en* is not be allowed). We are clearly able to see the improvement gained by PWCA over KCCA despite increasing the training alignment complexity. Furthermore, not only did the added aligned language not hinder the mate retrieval task, it improved performance as visible when comparing table 1 with table 2.

Table 2: We compare KCCA and PWCA on a trilingual mate-retrieval task (see text for language abbreviation). The reported results are the mean average precision for retrieving the exact paired document in another language for all possible tripartite combinations of the language stated in the column (without repetition) for training.

	da	de	en	es	it	nl	pt	sv
KCCA	0.3687	0.3290	0.3930	0.3742	0.3792	0.3501	0.3917	0.3909
PWCA	**0.5407**	**0.5155**	**0.5427**	**0.5394**	**0.5310**	**0.5246**	**0.5406**	**0.5504**

CCA (and KCCA) does not seek to maintain any pre-existing structure within the views while seeking to maximise correlation across the views. This aspect that may lead to over-fitting when having multiple views, PWCA addresses this by directly seeking to maintain internal structure by trying to find as many pairwise (or n-wise) clusters as possible which do not contradict the given data. We hypothesis that the PWCA performance improvement is a direct result of the clustering condition.

5 Discussion

This study presented a novel learning paradigm and corresponding algorithm that aims at finding structure (pairwise clusters) in paired (multi-view) observations. A case study on bilingual and trilingual mate-retrieval task, and a motivation using the PAC-Bayesian results are given. While this paper described a theoretical as well as applied proof of concept, many issues including efficiency, out-of-sample extensions and relations to other techniques remain.

References

[1] S. Ben-David and R. Schuller. Exploiting task relatedness for multiple task learning. In *Proceedings of The Conference on Learning Theory*, 2003.

[2] R. Raina, A. Battle, H. Lee, B. Packer, and A.Y. Ng. Self-taught learning: Transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, page 766. ACM, 2007.

[3] A. Banerjee, S. Merugu, I.S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *The Journal of Machine Learning Research*, 6:1705–1749, 2005.

[4] D.R. Hardoon, S.Szedmak, and J.Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[5] K. Pelckmans, S. Van Vooren, B. Coessens, J.A.K. Suykens, and B. De Moor. Mutual spectral clustering: Microarray experiments versus text corpus. In *Proc. of the workshop on Probabilistic Modeling and Machine Learning in Structural and Systems Biology*, pages 55–58. Helsinki University Printing House, Helsinki, Finland, 2006.

[6] K. Sim, V. Gopalkrishnan, H. N. Chua, and S. K. Ng. MACs: Multi-attribute co-clusters with high correlation information. In *Proceedings of The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Part II*, pages 398–413, 2009.

[7] Y. Seldin and N. Tishby. PAC-Bayesian Generalization Bound for Density Estimation with Application to Co-clustering. In *JMLR Workshop and Conference Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5, pages 472–479, 2009.

[8] F. Bach and M. Jordan. Kernel independent component analysis. *Journal of Machine Leaning Research*, 3:1–48, 2002.

[9] P. Koehn. Europarl: A multilingual corpus for evaluation of machine translation. *http://people.csail.mit.edu/~koehn/publications/europarl.ps*, unpublished.

[10] A. Maurer. A note on the PAC-Bayesian theorem. *Arxiv preprint cs/0411099*, 2004.

[11] D.A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170. ACM New York, NY, USA, 1999.

[12] J. Shawe-Taylor and A. Dolla. A framework for probability density estimation. In *Proceedings of the International Workshop on Artificial Intelligence and Statistics (AISTATS)*, pages 468–475, 2007.

[13] A. Vinokourov, J. Shawe-Taylor, and N. Christianini. Inferring a semantic representation of text via cross-language correlation analysis. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *Advances of Neural Information Processing Systems 15*, 2003.

[14] Y. Li and J. Shawe-Taylor. Using kcca for Japanese-English cross-language information retrieval and document classification. *Journal of Intelligent Information Systems*, 27(2):117–133, 2006.

[15] B. Fortuna, J. Rupnik, B. Pajntar, M. Grobelnik, and D. Mladenic. Cross-lingual search over 22 european languages. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, page 883, 2008.

[16] L. Guezouli and H. Essafi. CASIT: Content based identification of textual information in a large database. In *IEEE 24th International Conference on Advanced Information Networking and Applications Workshops*, 2010.

[17] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *Proceedings of the 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval*, pages 11–18, 2006.