ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Feature Learning using Multi-view Kernel Partial Least Squares

Xinjie Zeng, Qinghua Tao, and Johan Suykens *

KU Leuven - ESAT-STADIUS
Kasteelpark Arenberg 10, B-3001 Heverlee - Belgium
Email: firstname.lastname@esat.kuleuven.be, correspondence to Qinghua Tao

Abstract. The multi-view learning deals with data of multiple views, aiming to explore the underlying relations between different views and use them for various tasks. In this paper, we derive a multi-view extension of kernel partial least squares for unsupervised feature learning. We establish the optimization objective in the primal as the pairwise covariance between the projection scores and derive that this model can be trained in the dual form by solving an eigenvalue problem. Experiments are also conducted to verify the effectiveness of the method with real-life multi-view datasets, where the proposed method is adopted as a feature extractor and then the clustering task is conducted for performance comparisons.

1 Introduction

The Partial Least Squares (PLS) is a well-known technique that seeks the linear combinations of two sets of variables by maximizing the covariance of the projections [1]. The kernel version of PLS (KPLS) was derived for nonlinearity [2]. The Least Squares Support Vector Machine (LSSVM) formulation of KPLS was also introduced [3]. A related method is the Kernel Canonical Correlation Analysis (KCCA), which maximizes the correlation metric [4], and can be solved by a generalized eigenvalue problem [3]. Kernel Principal Components Analysis (KPCA) is another method closely related to KPLS, where the features are learned by capturing the maximal variances of the given data [5]. It has been elucidated that KPLS can also be interpreted as learning two KPCA and meanwhile pursuing the maximal covariance of their projections [6].

The multi-view learning has proven to be successful in numerous applications [7, 8, 9, 10]. Multi-view data have different *views*, i.e., each sample can have multiple representations. Multi-view learning takes advantage of the relations between different views and looks for underlying patterns of the data with richer information. Both (kernel) CCA and PCA have been extended to multi-view cases. Tensor models have also been introduced to multi-view KCCA and KPCA for capturing high-order correlations [11, 7]. Classical KPLS deals with 2-view data in the supervised regression task, where the two views are collected as the inputs and the responses. Although PLS was also extended to

*This work is jointly supported by the European Research Council under the European Union's Horizon 2020 research and innovation program/ERC Advanced Grant E-DUALITY (787960), iBOF project Tensor Tools for Taming the Curse (3E221427), EU H2020 ICT-48 Network TAILOR, and Leuven.AI Institute.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

multi-view versions, they either still focus on supervised tasks in establishing the methodology [12, 13, 14], or only consider the linear PLS [15]. In this work, we propose a multi-view kernel extension of PLS under unsupervised settings, where the features of each view are jointly learned by aligning with the spirits in PLS, i.e., the maximization on the pairwise covariance. In such a way, a novel feature learning method for multi-view data is established.

2 Multi-view Kernel Partial Least Squares

In this section, we introduce the formulation of multi-view kernel PLS, namely MvKPLS. We first start from the LSSVM formulation of the classical KPLS. Compared to KCCA, the objective of KPLS looses the squared terms of the projection scores and only keeps the pairwise coupling [3, 6]. The objective is then extended to the multi-view version, i.e., MvKPLS. Let V be the number of views, N be the number of samples in each view. Given the dataset $\{\boldsymbol{x}_i^{[v]} \in \mathbb{R}^{d^{[v]}}\}_{i=1}^N$ with $v = 1, \ldots, V$, the primal objective is formulated as

$$\begin{aligned} \max_{\boldsymbol{w}^{[v]}, e_i^{[v]}} \quad & J := -\frac{1}{2}\sum\nolimits_{v=1}^{V} \boldsymbol{w}^{[v]\top}\boldsymbol{w}^{[v]} + \gamma \sum\nolimits_{\substack{u,v=1 \\ u>v}}^{V} \sum\nolimits_{i=1}^{N} e_i^{[u]} e_i^{[v]} \\ \text{s.t.} \quad & e_i^{[v]} = \boldsymbol{w}^{[v]\top}\varphi^{[v]}(\boldsymbol{x}_i^{[v]}),\ i = 1, \ldots, N, \quad v = 1, ..., V \end{aligned} \tag{1}$$

where $\boldsymbol{w}^{[v]} \in \mathbb{R}^{p^{[v]}}$ are the primal model variables, $e_i^{[v]}$ denote the projection scores, $\varphi^{[v]} : \mathbb{R}^{d^{[v]}} \to \mathbb{R}^{p^{[v]}}$ is the feature map for the v-th view, $d^{[v]}$ and $p^{[v]}$ are the corresponding dimensions of the input spaces and feature spaces, and γ is the regularization constant. This objective can be interpreted as to seek for the joint maximization on the pairwise couplings of the projections from different views.

The Lagrangian is obtained by introducing dual variables, i.e., the Lagrangian multipliers $\alpha_i^{[v]} \in \mathbb{R}$, such that

$$\mathcal{L} = J - \sum\nolimits_{v=1}^{V} \sum\nolimits_{i=1}^{N} \alpha_i^{[v]} (e_i^{[v]} - \boldsymbol{w}^{[v]\top}\varphi^{[v]}(\boldsymbol{x}_i^{[v]})) \tag{2}$$

By taking the partial derivatives of (2), the Karush-Kuhn-Tucker (KKT) conditions are obtained as

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \boldsymbol{w}^{[v]}} = 0 \Longrightarrow \boldsymbol{w}^{[v]} = \displaystyle\sum_{i=1}^{N} \alpha_i^{[v]} \varphi^{[v]}(\boldsymbol{x}_i^{[v]}), \quad \forall v = 1, \ldots, V, \\ \dfrac{\partial \mathcal{L}}{\partial e_i^{[v]}} = 0 \Longrightarrow \gamma \displaystyle\sum_{u \neq v} e_i^{(u)} = \alpha_i^{[v]}, \quad \forall i = 1, \ldots, N, v = 1, \ldots, V, \\ \dfrac{\partial \mathcal{L}}{\partial \alpha_i^{[v]}} = 0 \Longrightarrow e_i^{[v]} = \boldsymbol{w}^{[v]\top}\varphi^{[v]}(\boldsymbol{x}_i^{[v]}), \quad \forall i = 1, \ldots, N, v = 1, \ldots, V. \end{cases}$$

which yields $\lambda \alpha_i^{[v]} = \sum_{u \neq v} \sum_{j=1}^{N} \alpha_j^{[u]} K^{[u]}(\boldsymbol{x}_j^{[u]}, \boldsymbol{x}_i^{[u]})$ where $\lambda \triangleq 1/\gamma$, and the kernel function $K^{[v]}(\boldsymbol{x}, \boldsymbol{y}) \triangleq \varphi^{[v]}(\boldsymbol{x})^\top \varphi^{[v]}(\boldsymbol{y})$ for view v. Hence, the solution in

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

the dual is given by:

$$\underbrace{\begin{bmatrix} \mathbf{0} & \boldsymbol{K}_2 & \cdots & \boldsymbol{K}_V \\ \boldsymbol{K}_1 & \mathbf{0} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \boldsymbol{K}_V \\ \boldsymbol{K}_1 & \cdots & \boldsymbol{K}_{V-1} & \mathbf{0} \end{bmatrix}}_{\Omega} \begin{bmatrix} \boldsymbol{\alpha}^{[1]} \\ \vdots \\ \boldsymbol{\alpha}^{[V]} \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{\alpha}^{[1]} \\ \vdots \\ \boldsymbol{\alpha}^{[V]} \end{bmatrix} \tag{3}$$

with $\boldsymbol{\alpha}^{[v]} \in \mathbb{R}^N$ defined as $\boldsymbol{\alpha}^{[v]} \triangleq [\alpha_1^{[v]}, \ldots, \alpha_N^{[v]}]^\top$ and the kernel matrix $\boldsymbol{K}_v \in \mathbb{R}^{N \times N}$ defined as $\boldsymbol{K}_v = [K^{[v]}(\boldsymbol{x}_i^{[v]}, \boldsymbol{x}_j^{[v]})]$. Given a data point $\tilde{\boldsymbol{x}}^{[v]}$, the dual model representation (out-of-sample extension) for the v-th view is then obtained as

$$e^{[v]}(\tilde{\boldsymbol{x}}^{[v]}) = \sum\nolimits_{i=1}^{N} \alpha_i^{[v]} K^{[v]}(\boldsymbol{x}_i^{[v]}, \tilde{\boldsymbol{x}}^{[v]}), \tag{4}$$

with the dual variables $\alpha_i^{[v]}$ optimized from (3), where the information of all the views is taken into consideration as seen from the kernels in the matrix Ω.

The constrained optimization problem in the primal is now transformed into an eigenvalue problem in the dual. With the dual formulation, the regularization constant γ can then be automatically chosen as the reciprocal of the positive real eigenvalues and we do not need to explicitly compute the possibly high dimensional feature mappings $\varphi^{[v]}(\boldsymbol{x}_i^{[v]})$. Calculating the kernel matrices gives a computational complexity of $\mathcal{O}(VN^2\overline{d})$ where $\overline{d}$ is the average input dimensions of all views [10]. Solving the eigenvalue problem in equation (3) leads to VN eigenvalues with the corresponding N-dimensional features for each view and gives a computational complexity of $\mathcal{O}(V^3N^3)$. In practice, the dataset can have informative features in low dimensions, so we can keep only the first dimensions of features, as shown in the following experiments in this paper. Suppose k dimensions are chosen, the computational complexity can be reduced to $\mathcal{O}(V^2N^2k)$. The proposed MvKPLS is applied as a feature extractor in this work, where the features with informative data patterns can be leveraged in the downstream tasks to benefit performances.

3 Experiments

This section describes the experiments performed to examine the proposed MvKPLS method. We consider our MvKPLS as a feature extractor and compare it to other methods based on the performance of the downstream tasks using the extracted features. In this work, we proceed with clustering as the downstream task. With extracted features, two classical clustering methods are considered: k-means (KM) [16] and spectral clustering (SC) [17]. To evaluate the clustering performance, we use the Normalized Mutual Information (NMI), which measures how similar two clusterings are. We compare MvKPLS with KPCA. We also compare with the results when clustering is directly applied to the data. In the experiments, we use 3 multi-view datasets: the image-caption dataset [7],

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

the 3 Sources dataset[1], and the YouTube video dataset[2]. The views of these datasets represent colors, texture, term frequencies of words and so on[18, 19]. Each sample is annotated with a specific label. We split the datasets into 60% training data, 20% validation data and 20% test data. Based on the clustering performance on the validation set, the kernels are carefully chosen, and the kernel-specific parameters are tuned. For KPCA and MvKPLS, we set up two experiments to examine the number k of extracted features: $k = 10$ and k is the number of components capturing about 95% explained variance. We compute the NMI scores for each view and the results are presented in Table 1, Table 2. Further, we also evaluate the NMI scores for KPCA and our MvKPLS by averaging the features when setting up $k = 10$ and report the overall evaluation of the methods for each dataset as shown in Table 3, as it is more common to have a consensus for the results in multi-view learning; for the k determined by the explained variance, k can be different for each view in KPCA, and we thus omit it. For each dataset, the highest NMI score (best view) for each method is underlined, and the best performance among all methods for each dataset is shown in bold.

Table 1: The NMI results of different methods on the test sets. $k = 10$ for both KPCA and MvKPLS.

Method	Image-caption			3 Sources			YouTube video		
View	1	2	3	1	2	3	1	2	3
KM	0.502	0.297	0.174	0.263	0.158	0.052	0.335	0.190	0.037
KPCA+KM	**0.633**	0.234	0.439	0.519	0.558	0.512	0.131	0.232	0.121
MvKPLS+KM	0.574	0.572	0.475	**0.627**	0.537	0.601	**0.470**	0.443	0.342
SC	0.003	0.005	0.005	0.526	0.182	0.083	0.504	0.181	0.083
KPCA+SC	**0.594**	0.283	0.356	0.390	0.505	0.275	0.091	0.258	0.115
MvKPLS+SC	0.580	0.381	0.373	**0.535**	0.525	0.389	0.585	**0.627**	0.593

The results in Table 1 and Table 2 show that in most cases, MvKPLS gives the best NMI scores. One can also see that compared to the direct clustering on the original data, KPCA and MvKPLS both give higher scores, implying that they indeed extract useful information from the data and can thus express the samples in a compact and informative way in subspaces. When k is chosen according to the explained variance, MvKPLS has slightly worse results than KPCA in some views such as the second view of the YouTube video dataset when k-means is used. This might be caused by the extremely large value of k: in the YouTube video dataset, $k = 1236$, which can lead to significant noise. Further, Table 3 shows that when doing the clustering on the averaged features with an overall evaluation on each dataset, MvKPLS distinctively outperforms KPCA, indicating ours is better at catching the underlying relations across views.

[1]http://mlg.ucd.ie/datasets/3sources.html

[2]https://archive.ics.uci.edu/dataset/269/youtube+multiview+video+games+dataset

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Table 2: The NMI results of different methods on the test sets. k is chosen such that 95% of variance is explained.

Method	Image-caption			3 Sources			YouTube video		
View	1	2	3	1	2	3	1	2	3
KM	0.502	0.297	0.174	0.263	0.158	0.052	0.335	0.190	0.037
KPCA+KM	0.637	0.259	0.390	0.292	0.445	0.494	0.085	0.231	0.131
MvKPLS+KM	**0.733**	0.024	0.224	0.615	**0.632**	0.549	**0.400**	0.027	0.095
SC	0.003	0.005	0.005	0.544	0.180	0.082	0.504	0.181	0.083
KPCA+SC	**0.584**	0.278	0.331	0.247	0.360	0.381	0.161	0.248	0.104
MvKPLS+SC	0.580	0.544	0.336	**0.569**	0.523	0.381	**0.552**	0.265	0.108

Table 3: The NMI results of different methods on the test sets, when the extracted features are averaged. $k = 10$ for both KPCA and MvKPLS.

Method	Image-caption	3 Sources	YouTube video
KPCA+KM	0.518	0.520	0.166
MvKPLS+KM	**0.584**	**0.587**	**0.492**
KPCA+SC	0.441	0.400	0.196
MvKPLS+SC	**0.625**	**0.571**	**0.616**

These experiments together demonstrate the potential of KPLS in multi-view feature learning.

4 Conclusion

This paper introduced the multi-view kernel partial least squares method to perform KPLS on multi-view data as a feature learning technique. The model is characterized in the LSSVM setting, which maximizes the pairwise covariances between views and leads to an eigenvalue problem involving multiple kernel matrices in the dual. Numerical experiments are conducted where the method extracts information from multi-view data before downstream clustering. The results have shown the proposed model performs well on multi-view data. As a future study, it would be interesting to investigate more properties of the solutions due to the asymmetry in the eigenvalue problem. It would also be intriguing to introduce the learnable neural networks as the feature maps in the primal form (such as in Primal-Attention for Transformers [20] with kernel singular value decomposition setups [21]), so that the model becomes parametric and can be optimized by algorithms like stochastic gradient descent.

References

[1] Herman Wold. Estimation of principal components and related models by iterative least squares. *Multivariate Analysis*, 1, 1966.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

[2] Roman Rosipal and Leonard J. Trejo. Kernel partial least squares regression in reproducing kernel Hilbert space. *Journal of Machine Learning Research*, 2:97–123, 2001.

[3] Johan Suykens, Tony Van Gestel, Jos De Brabanter, Bart De Moor, and Joos Vandewalle. *Least Squares Support Vector Machine*. World Scientific, Singapore, 2002.

[4] Jon R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.

[5] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *the International Conference on Artificial Neural Networks*, pages 583–588, 1997.

[6] Luc Hoegaerts, Johan AK Suykens, Joos Vandewalle, and Bart De Moor. Primal space sparse kernel partial least squares regression for large scale problems. In *the IEEE International Joint Conference on Neural Networks*, volume 1, pages 561–563, 2004.

[7] Lynn Houthuys and Johan AK Suykens. Tensor learning in multi-view kernel PCA. In *the Artificial Neural Networks and Machine Learning*, volume 27, pages 205–215, 2018.

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

[9] Jia Chen, Gang Wang, and Georgios B. Giannakis. Graph multiview canonical correlation analysis. *IEEE Transactions on Signal Processing*, 67(11):2826–2838, 2019.

[10] Qinghua Tao, Francesco Tonin, Panagiotis Patrinos, and Johan AK Suykens. Tensor-based multi-view spectral clustering via shared latent space. *Information Fusion*, page 102405, 2024.

[11] Yong Luo, Dacheng Tao, Kotagiri Ramamohanarao, Chao Xu, and Yonggang Wen. Tensor canonical correlation analysis for multi-view dimension reduction. *IEEE Transactions on Knowledge and Data Engineering*, 27(11):3111–3124, 2015.

[12] Li Wang, Ren-Cang Li, and Wen-Wei Lin. Multiview orthonormalized partial least squares: Regularizations and deep extensions. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8):4371–4385, 2023.

[13] Flavio Camarrone and Marc M Van Hulle. Fast multiway partial least squares regression. *IEEE Transactions on Biomedical Engineering*, 66(2):433–443, 2018.

[14] Yi Mou, Long Zhou, Xinge You, Yaling Lu, Weizhen Chen, and Xu Zhao. Multiview partial least squares. *Chemometrics and Intelligent Laboratory Systems*, 160:13–21, 2017.

[15] Li Wang and Ren-Cang Li. A scalable algorithm for large-scale unsupervised multi-view partial least squares. *IEEE Transactions on Big Data*, 8(4):1073–1083, 2020.

[16] James MacQueen. Some methods for classification and analysis of multivariate observations. In *the Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.

[17] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 14, 2001.

[18] Weilong Yang and George Toderici. Discriminative tag learning on YouTube videos with latent sub-tags. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3217–3224. IEEE, 2011.

[19] Richard F. Lyon, Martin Rehn, Samy Bengio, Thomas C Walters, and Gal Chechik. Sound retrieval and ranking using sparse auditory representations. *Neural Computation*, 22(9):2390–2416, 2010.

[20] Yingyi Chen, Qinghua Tao, Francesco Tonin, and Johan AK Suykens. Primal-attention: Self-attention through asymmetric kernel svd in primal representation. *Advances in Neural Information Processing Systems*, 2023.

[21] Qinghua Tao, Francesco Tonin, Alex Lambert, Yingyi Chen, Panagiotis Patrinos, and Johan AK Suykens. Learning in feature spaces via coupled covariances: Asymmetric kernel SVD and nyström method. In *International Conference on Machine Learning*, 2024.