Learning Low-Dimensional Temporal Representations Supplementary Material

Bing Su [1] Ying Wu [2]

1. Proof of Lemma 1

Lemma 1. *Objective function (#3)*[1] *is equivalent to the trace maximization problem*

$$\max tr(\mathbf{F}^T\mathbf{Z}^T\mathbf{Z}\mathbf{F}). \tag{1}$$

Proof. It turns out that the k-th column of $\mathbf{Z}\mathbf{F}\mathbf{F}^T$ is exactly $\mathbf{m}_j$ (*i.e.*, the mean of the j-th stage of class c) that the k-th column $\mathbf{z}_k$ of $\mathbf{Z}$ is aligned to. Therefore, the objective function (#3) can also be written as

$$\begin{array}{l} \min \left\| \mathbf{Z} - \mathbf{Z}\mathbf{F}\mathbf{F}^T \right\|_F^2 \\ \Leftrightarrow \min tr(\mathbf{Z}^T\mathbf{Z}) - tr(\mathbf{F}^T\mathbf{Z}^T\mathbf{Z}\mathbf{F}) \\ \Leftrightarrow \max tr(\mathbf{F}^T\mathbf{Z}^T\mathbf{Z}\mathbf{F}) \end{array} .$$

□

2. Proof of Lemma 2

Lemma 2. *Objective function (#7) is equivalent to the trace maximization problem*

$$\max tr(\mathbf{F}^T\mathbf{Z}^T\mathbf{W}(\mathbf{W}^T\mathbf{Z}\mathbf{Z}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{Z}\mathbf{F}). \tag{2}$$

Proof. If the data has zero-centered, $\sum_{i=1}^{N_t} \mathbf{z}_i/N_t = 0$, (otherwise we can remove the overall mean of all the vectors in all the training sequences), then $\mathbf{S}_t$ and $\mathbf{S}_b$ can be reformulated as

$$\mathbf{S}_t = \mathbf{Z}\mathbf{Z}^T,$$

$$\mathbf{S}_b = \sum_{i=1}^{C}\sum_{u=1}^{L} p_u^i \mathbf{m}_u^i \mathbf{m}_u^{i\,T} = \mathbf{Z}\mathbf{F}\mathbf{F}^T\mathbf{Z}.$$

Since $\mathbf{S}_b + \mathbf{S}_w = \mathbf{S}_t$, the objective (#7) is equivalent to

$$\begin{array}{l} \max tr((\mathbf{W}^T\mathbf{S}_w\mathbf{W})^{-1}\mathbf{W}^T\mathbf{S}_b\mathbf{W}) \\ \Leftrightarrow \max tr((\mathbf{W}^T\mathbf{S}_t\mathbf{W})^{-1}\mathbf{W}^T\mathbf{S}_b\mathbf{W}) \end{array} , \tag{3}$$

which can further be reformulated as

$$\begin{array}{l} tr((\mathbf{W}^T\mathbf{S}_t\mathbf{W})^{-1}\mathbf{W}^T\mathbf{S}_b\mathbf{W}) \\ = tr((\mathbf{W}^T\mathbf{Z}\mathbf{Z}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{Z}\mathbf{F}\mathbf{F}^T\mathbf{Z}^T\mathbf{W}) \\ = tr(\mathbf{F}^T\mathbf{Z}^T\mathbf{W}(\mathbf{W}^T\mathbf{Z}\mathbf{Z}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{Z}\mathbf{F}) \end{array} .$$

□

3. Proof of Theorem 1

Theorem 1. *The LT-LDA algorithm (Alg. 2) is guaranteed the converge.*

Proof. In the first stage, LT-LDA optimizes over $\mathbf{P}$ by fixing $\mathbf{W}$ using Alg. 1. According to Lemma 1, this actually learns the abstract template from the whitened and projected feature sequences by optimizing $\max tr(\mathbf{F}^T\hat{\mathbf{Z}}^T\hat{\mathbf{Z}}\mathbf{F})$. $\hat{\mathbf{Z}} = \mathbf{\Gamma}_w^{-\frac{1}{2}}\mathbf{W}^T\mathbf{Z}$ is the whitened and projected data matrix, and $\mathbf{\Gamma}_w = \mathbf{W}^T\mathbf{Z}\mathbf{Z}^T\mathbf{W}$ is the total scatter of the projected data. Thus we have

$$\begin{array}{l} tr(\mathbf{F}^T\hat{\mathbf{Z}}^T\hat{\mathbf{Z}}\mathbf{F}) \\ = tr(\mathbf{F}^T\mathbf{Z}^T\mathbf{W}(\mathbf{W}^T\mathbf{Z}\mathbf{Z}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{Z}\mathbf{F}) \end{array} ,$$

which is exactly the objective Eq. (2).

In the second stage, LT-LDA optimizes over $\mathbf{W}$ for given $\mathbf{P}$. According to Lemma 2, this also optimizes Eq. (2).

Both iterative stages decrease the objective value of Eq. (2) monotonically. Since $\mathbf{F}$ is an orthogonal matrix, the objective (2) is bounded from above. This guarantees the convergence of Alg. 2. □

4. Proof of Theorem 2

Theorem 2. *Let* $\mathbf{G} = \mathbf{Z}^T\mathbf{Z}$ *be the Gram matrix. When the dimensionality is reduced to a specific value* $d' =$

[1] Science & Technology on Integrated Information System Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing, China [2] Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA. Correspondence to: Bing Su <subingats@gmail.com>.

Proceedings of the 35^{th} *International Conference on Machine Learning*, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

[1] The numberings of equations with the mark "#" refer to the equations in the main text, while those without the mark "#" refer to the equations in this supplementary file.

$\min(CL, d, N_t)$ and a regularization term $\delta \mathbf{I}_{N_t}$ is added to the total scatter $\mathbf{S}_t$, where $\mathbf{I}_{N_t}$ is the N_t-order identity matrix, if $\mathbf{W}^*$ and $\mathbf{F}^*$ are the optimal solutions of the trace maximization problem (2):

$$\max_{\mathbf{W},\mathbf{F}} tr(\mathbf{F}^T\mathbf{Z}^T\mathbf{W}(\mathbf{W}^T(\mathbf{Z}\mathbf{Z}^T + \delta\mathbf{I}_{N_t})\mathbf{W})^{-1}\mathbf{W}^T\mathbf{Z}\mathbf{F}) \tag{4}$$

then $\mathbf{F}^*$ is also the optimal solution of the problem

$$\max_{\mathbf{F}} tr(\mathbf{F}^T(\mathbf{I}_{N_t} - (\mathbf{I}_{N_t} + \frac{1}{\delta}\mathbf{G})^{-1})\mathbf{F}) \tag{5}$$

Proof. We follow the proof in (Ye et al., 2007). According to the representer theorem (Schölkopf & Smola, 2002), the optimal projection matrix $\mathbf{W}$ has the form $\mathbf{W} = \mathbf{Z}\mathbf{A}$, where $\mathbf{A}$ is a coefficient matrix. The objective (4) is transformed to

$$\begin{aligned}&tr(\mathbf{F}^T\mathbf{Z}^T\mathbf{Z}\mathbf{A}(\mathbf{A}^T\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T + \delta\mathbf{I}_{N_t})\mathbf{Z}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Z}^T\mathbf{Z}\mathbf{F})\\ &= tr(\mathbf{F}^T\mathbf{G}\mathbf{A}(\mathbf{A}^T(\mathbf{G}\mathbf{G} + \delta\mathbf{G})\mathbf{A})^{-1}\mathbf{A}^T\mathbf{G}\mathbf{F})\\ &= tr(\mathbf{A}^T\mathbf{G}\mathbf{F}\mathbf{F}^T\mathbf{G}\mathbf{A}(\mathbf{A}^T(\mathbf{G}\mathbf{G} + \delta\mathbf{G})\mathbf{A})^{-1})\end{aligned} \tag{6}$$

By defining $\mathbf{\Gamma}_b = \mathbf{G}\mathbf{F}\mathbf{F}^T\mathbf{G}$ and $\mathbf{\Gamma}_w = \mathbf{G}\mathbf{G} + \delta\mathbf{G}$, we can find that Eq. (6) has a similar form with the generalized LDA problem (Ye, 2005), which can be solved by constructing a matrix $\mathbf{Q}$ that simultaneously diagonalizes $\mathbf{\Gamma}_b$ and $\mathbf{\Gamma}_w$. $\mathbf{Q}$ is constructed as follows.

$\mathbf{G}$ is symmetric and positive semi-definite. If the d-dimensional features are not linearly dependent (otherwise we can remove the linearly correlated dimensions), the rank is of $\mathbf{G}$ is $r = \min(d, N_t)$. The SVD of $\mathbf{G}$ has the form

$$\mathbf{G} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{U}_r^T,$$

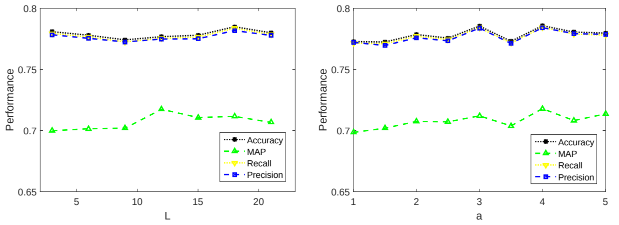
where $\mathbf{U}$ is an orthogonal and square matrix, $\mathbf{\Sigma} = diag(\lambda_1, \cdots, \lambda_r, 0, \cdots, 0)$, $\mathbf{U}_r$ is consist of the first r columns of $\mathbf{U}$, and $\mathbf{\Sigma}_r = diag(\lambda_1, \cdots, \lambda_r)$ contains only the non-zero singular values.

Define $\mathbf{V} = (\mathbf{\Sigma}_r^2 + \delta\mathbf{\Sigma}_r)^{-\frac{1}{2}}\mathbf{\Sigma}_r\mathbf{U}_r^T\mathbf{F}$. The SVD of $\mathbf{V}$ is denoted as $\mathbf{V} = \mathbf{M}\mathbf{\Sigma}_V\mathbf{N}^T$, where $\mathbf{M}$ and $\mathbf{N}$ are orthogonal matrices, $\mathbf{\Sigma}_V$ is a diagonal matrix of d'-th order, and d' is the rank of $\mathbf{V}$. $d' = \min(rank(\mathbf{S}_b), r) = \min(CL, d, N_t)$. By constructing $\mathbf{Q}$ as

$$\mathbf{Q} = \mathbf{U}diag((\mathbf{\Sigma}_r^2 + \delta\mathbf{\Sigma}_r)^{-\frac{1}{2}}\mathbf{M}, \mathbf{I}_{N_t - r}),$$

$\mathbf{\Gamma}_b$ and $\mathbf{\Gamma}_w$ are simultaneously diagonalized by $\mathbf{Q}$.

$$\mathbf{Q}^T\mathbf{\Gamma}_b\mathbf{Q} = diag(\mathbf{\Sigma}_V^2, \mathbf{0}_{N_t - r}),$$
$$\mathbf{Q}^T\mathbf{\Gamma}_w\mathbf{Q} = diag(\mathbf{I}_r, \mathbf{0}_{N_t - r}).$$



(a) (b)

Figure 1. Different performances with the rank pooling and the SVM classifier as functions of (a) the length of the abstract template L and (b) the control factor a on the Chalearn Gesture dataset.

From Theorem 3.1 in (Ye, 2005), the solution of maximizing the objective (6) over $\mathbf{A}$ is given by $\mathbf{A}^*$ consisting of the first d' columns of $\mathbf{Q}$, and the maximum of the objective (6) equals $tr((\mathbf{G}\mathbf{G} + \delta\mathbf{G})^\dagger(\mathbf{G}\mathbf{F}\mathbf{F}^T\mathbf{G}))$. Therefore,

$$\begin{aligned}&tr((\mathbf{A}^T(\mathbf{G}\mathbf{G} + \delta\mathbf{G})\mathbf{A})^{-1}\mathbf{A}^T\mathbf{G}\mathbf{F}\mathbf{F}^T\mathbf{G}\mathbf{A})\\ &\leq tr((\mathbf{G}\mathbf{G} + \delta\mathbf{G})^\dagger(\mathbf{G}\mathbf{F}\mathbf{F}^T\mathbf{G}))\\ &= tr(\mathbf{F}^T\mathbf{G}(\mathbf{G}\mathbf{G} + \delta\mathbf{G})^\dagger\mathbf{G}\mathbf{F})\\ &= tr(\mathbf{F}^T(\mathbf{I}_{N_t} - (\mathbf{I}_{N_t} + \tfrac{1}{\delta}\mathbf{G})^{-1})\mathbf{F})\end{aligned}$$

The equality holds when $\mathbf{A} = \mathbf{A}^*$, where the dimensionality is implicitly reduced to d'. □

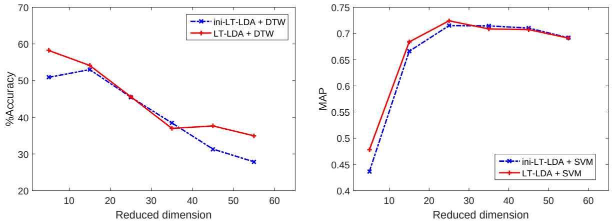
5. Experimental Setup on Another Dataset

"Spoken Arabic Digits (SAD)" dataset from the UCI Machine Learning Repository (Bache & Lichman, 2013). This dataset consists of 8,800 vector sequences from ten classes. The vectors in sequences are 13 mel-frequency cepstrum coefficients (MFCCs) corresponding to spoken Arabic digits. The ten spoken Arabic digits were repeated ten times by 44 males and 44 females Arabic native speakers and hence there are 880 sequence samples per digit class. The length of sequences varies from 4 to 93 frames. The dataset has already been split into training and test sets, where 660 samples from each class are used for training and the remaining 220 sequences are used for testing.

For the SAD dataset, each spoken digit has also been represented by a sequence of 13-dimensional frame-wide MFCC features.

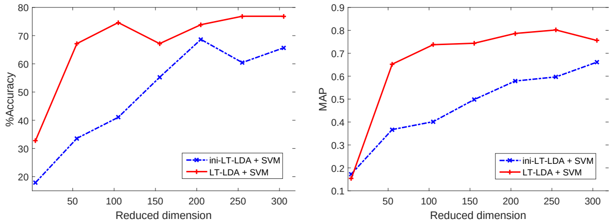
6. Influence of Parameters

In this section we evaluate the influence of the parameters on the ChaLearn gesture dataset. Similar to the MSR Action3D dataset, different performance measures including accuracy, MAP, and F-score with the rank pooling and the SVM classifier are evaluated by increasing L from 3 to 21



(a) (b)

Figure 2. Comparisons of the proposed LT-LDA without and with the joint learning of the latent alignments. (a) Accuracies with the DTW classifier and (b) MAPs with the rank pooling and the SVM classifier as functions of the dimensionality of the subspace on the Chalearn Gesture dataset.



(a) (b)

Figure 4. Comparisons of the proposed LT-LDA without and with the joint learning of the latent alignments. (a) Accuracies and (b) MAPs with the rank pooling and the SVM classifier as functions of the dimensionality of the subspace on the Olympic Sports dataset.

with an interval of 3 while fixing a to 2, and increasing a from 1 to 5 with an interval of 0.5 while fixing L to 8, respectively. The results on the dataset are shown in Fig. 1, where the reduced dimensionality is fixed to 45.

The optimal parameters are generally the same for multi-class indicators including accuracy, precision, recall and F-score, but are different for MAP. On the ChaLearn dataset, LT-LDA is insensitive to L for multi-class indicators. When $L = 12$, LT-LDA achieves the highest MAP. ATs with more stages help to discriminate the subtle differences among fine gesture actions in this dataset. LT-LDA is also not very sensitive to a on this dataset. It seems that allowing larger warping leads to better results because the subtle differences can be captured more easily by more flexible alignments.

7. Effects of the Joint Learning of the Subspace and the Latent Alignments

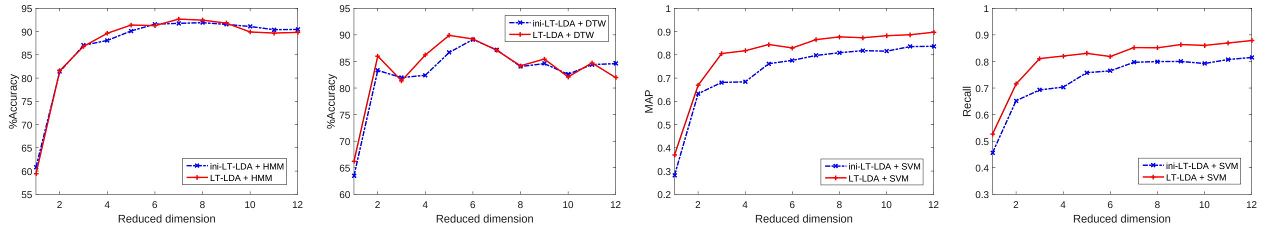
In this section we compare LT-LDA and ini-LT-LDA on more datasets. On both the ChaLearn dataset and the MSR Action3D dataset, the frame-wide features are based on the relative 3D joint positions, and we only show the comparison results on the Chalearn dataset in Fig. 2. We can find that the optimal results of LT-LDA among all dimensions with both the DTW classifer and the SVM classifer are better than those of the ini-LT-LDA.algothithm, but the improvements are quite small. The parameter C of the SVMs is fixed to 100 here, and tuning C by cross validation can further improve the performances of LT-LDA as shown in Fig.4 of the main text.

Fig. 3 shows the comparisons on the SAD dataset by different classifiers. The improvements of LT-LDA over ini-LT-LDA are still limited on accuracies by the HMM classifer and the DTW classifer, but are more significant on MAPs and multi-class recalls by the SVM classifier. This is because LT-LDA optimizes the overall separability between sequence classes, and hence sequences from different classes get better separated en bloc. However, as to a specific text sequence, it may be distributed on the boundary of a nearby class and confuse the classifier.

The comparisons of the two algorithms by the SVM classifier on the large scale Olympic Sports dataset are shown in Fig. 4. We can find LT-LDA significantly outperforms ini-LT-LDA by a much larger margin than the Kinect based datasets. This is because the depth information and the locations of human joints are available for the ChaLearn dataset. Thus the alignments in the original space are accurate and should be preserved in the optimal subspace. After the refinement of LT-LDA, the alignments remain nearly unchanged and the performance changes are small. On the Olympic dataset, only the raw videos with complex backgrounds are available. The initial alignments are quite noisy, and refining them by LTLDA improves the performances significantly.

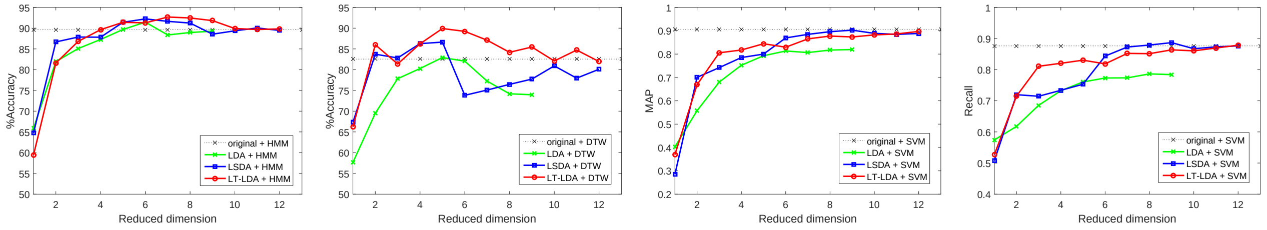
In summary, in nearly all cases, the best results among all these dimensions of LT-LDA outperform those of ini-LT-LDA. Jointly learning the latent alignments associated with the subspace does help to improve the classification performance in the subspace. This is because the temporal structures and the alignments may change from those in the original space. In the learned subspace of ini-LT-LDA, although different classes get better separated under the alignments in the original space, additional confusions may be introduced due to the change of alignments. While for LT-LDA, since the separability is maximized in the subspace under the corresponding alignments, the learned subspace gets joint optimality among all possible subspaces.

We can also observe that on some datasets with some classifiers, the performances of LT-LDA are lower that those of ini-LT-LDA in some dimensions. This is because the aligned paths of each class are learned without discriminating other classes, and the objective of LT-LDA is not directly related to a classification performance measure. For some dimensions, although our joint learning leads to bet-



(a) (b) (c) (d)

Figure 3. Comparisons of the proposed LT-LDA without and with the joint learning of the latent alignments. (a) Accuracies with the HMM classifier (b) Accuracies with the DTW classifier (c) MAPs with the rank pooling and the SVM classifier and (b) Multi-class average recalls with the rank pooling and the SVM classifier as functions of the dimensionality of the subspace on the SAD dataset.



(a) (b) (c) (d)

Figure 5. (a) Accuracies with the HMM classifier (b) accuracies with the DTW classifier and (c) MAPs with the rank pooling and the SVM classifier as functions of the dimensionality of the subspace on the SAD dataset.

ter separation of temporal structures, it may interfere the classifier, as the discrimination of the classifier may not necessarily accord to such separation but rather to other aspects of some local or global properties of sequences. Thus updating the paths may cause some fluctuates in accuracy. However, it generally leads to more discriminative subspace, because the entire sequences are more likely to get better separation if the temporal structures are better separated.

8. Comparison with Different Dimensionality Reduction Methods

In this section we compare LT-LDA with other dimensionality reduction methods on the two additional datasets. Fig. 5 depicts the performances by the three classifiers as functions of the dimensionality of the learned subspace on the SAD dataset, respectively. We can observe that the proposed LT-LDA achieves the best performances among all these dimensionality reduction methods by all the three classifiers with different evaluation measures on nearly all the datasets. On the SAD dataset, by the HMM classifier and the SVM classifier, LSDA performs comparatively with LT-LDA. This is because sufficient training samples are available on this dataset, and the dimensionality of frame-wise features is low. Therefore, LSDA can also reliably train HMMs to obtain sequence statistics.

References

Bache, K. and Lichman, M. *UCI Machine Learning Repository*. http://archive.ics.uci.edu/ml, University of California, Irvine, School of Information and Computer Sciences, 2013.

Schölkopf, B. and Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Ye, J. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *Journal of Machine Learning Research*, 6(Apr):483–502, 2005.

Ye, J., Zhao, Z., and Wu, M. Discriminative k-means for clustering. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.