Interactive Dimensionality Reduction for Comparative Analysis

Comparison of Dimensionality Reduction Methods

Takanori Fujiwara, Xinhai Wei, Jian Zhao, and Kwan-Liu Ma

We present this supplementary document to compare results generated by different linear dimensionality reduction (DR) methods: PCA, LDA, cPCA, and ULCA. As described in Sect. 4 and demonstrated in the case studies, ULCA's strength is in its analysis capability utilizing both discriminant analysis (e.g., LDA) and contrastive learning (e.g., cPCA) and flexibility allowing adjustment of algorithm parameters based on analysis need. Here, we apply PCA, LDA, and cPCA to the dataset used in Sect. 3 and 7, and compare the ULCA results shown in Fig. 1, 2, 3, 9, and 10. While we perform PCA for all the analysis cases, we apply LDA and cPCA only when they can (partially) fulfill the analysis purpose in each case. More specifically, LDA is designed for maximizing the separation among groups in an embedding space; thus, we only use LDA when we try to see the separation of groups during the analysis. Similarly, cPCA is to find an embedding space where a target group's variance is higher than a background group's variance; thus, we only use cPCA when we want to increase or decrease some group's variance during the analysis.

A Results for the Wine Dataset

The results corresponding to Fig. 1 are shown below. While PCA is applied to the entire dataset without label information, LDA is used with label information. Since the analysis in Fig. 1 is for separating the wine groups, cPCA is not used here. All the embedding results (Fig. A.1) with PCA, LDA, and ULCA show similar patterns; however, we can see LDA and ULCA have a better separation among groups. Also, from the axis information (Table A.1), we can see that both LDA and ULCA show `flavanoids` and `proline` have strong contribution to the embedding axes. These results demonstrate that ULCA embraces LDA's comparative analysis capabilities.

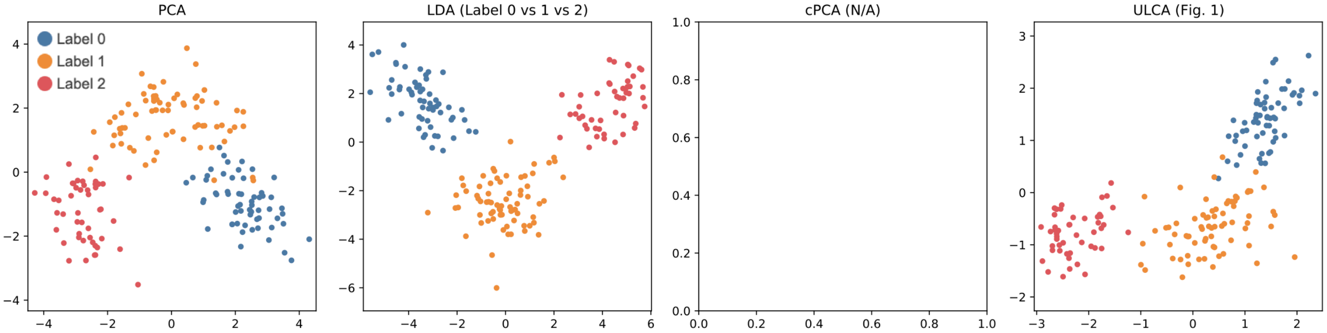

Figure A.1: The embedding results corresponding to Fig. 1.

Table A.1: The axis information of Fig. A.1.

	PCA (x)	PCA (y)	LDA (x)	LDA (y)	cPCA (x)	cPCA (y)	ULCA (x)	ULCA (y)
alcohol	0.144329	-0.483652	-0.326569	0.705754	0	0	0.047908	0.222532
malic_acid	-0.245188	-0.224931	0.184094	0.340194	0	0	-0.119098	-0.001570
ash	-0.002051	-0.316069	-0.100969	0.641759	0	0	-0.068764	0.373212
alcalinity_of_ash	-0.239320	0.010591	0.515503	-0.487473	0	0	-0.140328	-0.391263
magnesium	0.141992	-0.299634	-0.030813	-0.006591	0	0	0.020823	-0.065924
total_phenols	0.394661	-0.065040	0.385720	-0.020104	0	0	-0.052743	0.232090
flavanoids	0.422934	0.003360	-1.654628	-0.490054	0	0	0.780194	-0.385695
nonflavanoid_phenols	-0.298533	-0.028779	-0.185636	-0.202407	0	0	0.062533	-0.143723
proanthocyanins	0.313429	-0.039302	0.076533	-0.175270	0	0	0.016385	0.054407
color_intensity	-0.088617	-0.529996	0.820805	0.585410	0	0	-0.344212	-0.115462
hue	0.296715	0.279235	-0.186454	-0.345456	0	0	0.227938	-0.007595
od280/od315	0.376167	0.164496	-0.819544	0.036238	0	0	0.362277	0.386084
proline	0.286752	-0.364903	-0.845097	0.895899	0	0	0.203277	0.514844

The results corresponding to Fig. 2 are shown below. In Fig. 2, we review the factors common between Labels 1 and 2 but different from Label 0. Thus, we apply LDA to the dataset to distinguish Labels 1 and 2 from Label 0 (i.e., in LDA, we aggregate data points in Labels 1 and 2 into one class). Note that when using LDA, the number of embedding axes needs to be lower or equal to the number of classes; thus, Fig. A.2 shows 1D LDA result. Both LDA and ULCA show a clear separation between the groups and similar attributes' contribution to the embedding result, as shown in Table A.2 (e.g., `proline` has the largest contribution to *x*-axis).

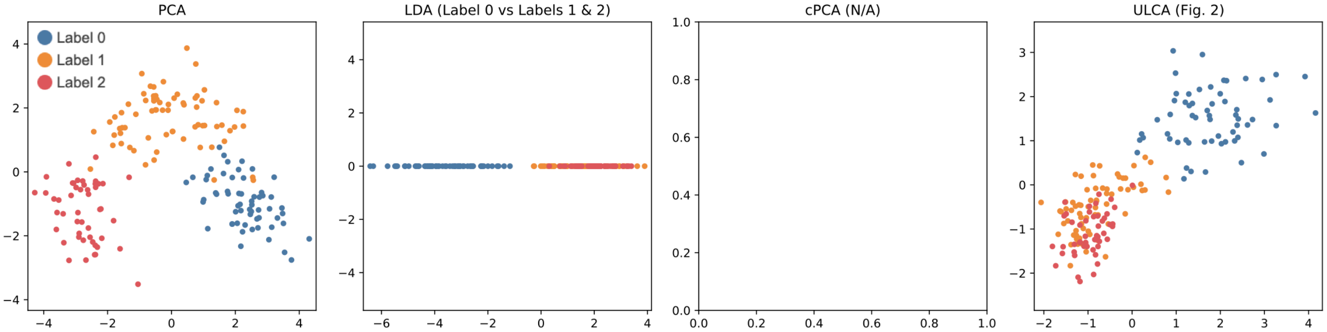

Figure A.2: The embedding results corresponding to Fig. 2.

Table A.2: The axis information of Fig. A.2.

	PCA (x)	PCA (y)	LDA (x)	LDA (y)	cPCA (x)	cPCA (y)	ULCA (x)	ULCA (y)
alcohol	0.144329	-0.483652	-0.721476	0.0	0	0	0.233373	0.437619
malic_acid	-0.245188	-0.224931	-0.104383	0.0	0	0	0.004880	-0.052372
ash	-0.002051	-0.316069	-0.516378	0.0	0	0	0.012357	0.282481
alcalinity_of_ash	-0.239320	0.010591	0.704967	0.0	0	0	-0.093801	-0.512145
magnesium	0.141992	-0.299634	-0.017396	0.0	0	0	-0.090407	0.200818
total_phenols	0.394661	-0.065040	0.288819	0.0	0	0	0.125364	0.175940
flavanoids	0.422934	0.003360	-0.839866	0.0	0	0	0.446139	-0.047882
nonflavanoid_phenols	-0.298533	-0.028779	0.007867	0.0	0	0	-0.102017	0.027326
proanthocyanins	0.313429	-0.039302	0.175919	0.0	0	0	0.051957	0.027711
color_intensity	-0.088617	-0.529996	0.179579	0.0	0	0	0.097398	0.017415
hue	0.296715	0.279235	0.106346	0.0	0	0	0.141806	0.005798
od280/od315	0.376167	0.164496	-0.609172	0.0	0	0	-0.152601	0.620854
proline	0.286752	-0.364903	-1.222698	0.0	0	0	0.804762	-0.052649

The results corresponding to Fig. 3 are shown below. In Fig. 3, we identify the factors for which wines in Label 2 have high variety than the others. Thus, for this analysis, we use cPCA but not LDA. Since cPCA only takes two groups as its inputs (i.e., target and background groups), we set data points in Label 2 as a target group and the other data points as a background group. cPCA's α value is automatically selected by using the existing implementation available online[1]. Based on Fig. A.3 and Table A.3, cPCA and ULCA have similar embedding results and axes. Throughout the analyses in this section, we can observe that ULCA embraces both LDA and cPCA's analysis capabilities.

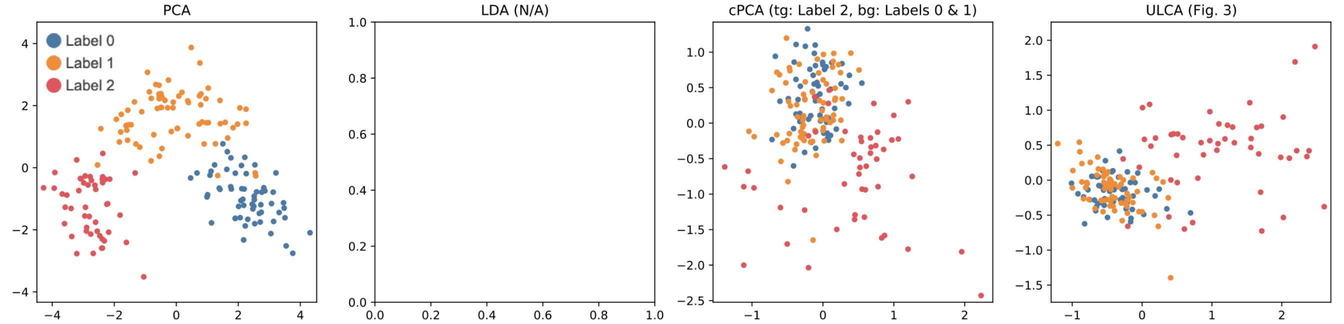

Figure A.3: The embedding results corresponding to Fig. 3.

Table A.3: The axis information of Fig. A.3.

	PCA (x)	PCA (y)	LDA (x)	LDA (y)	cPCA (x)	cPCA (y)	ULCA (x)	ULCA (y)
alcohol	0.144329	-0.483652	0	0	0.018234	-0.473961	-0.144189	0.073715
malic_acid	-0.245188	-0.224931	0	0	-0.004813	0.085529	0.092141	-0.057192
ash	-0.002051	-0.316069	0	0	0.027183	-0.238400	-0.003362	0.066421
alcalinity_of_ash	-0.239320	0.010591	0	0	0.012851	-0.045151	0.049720	0.003406
magnesium	0.141992	-0.299634	0	0	-0.163372	0.028401	0.008755	-0.088388
total_phenols	0.394661	-0.065040	0	0	0.442786	0.010927	-0.243510	0.511967
flavanoids	0.422934	0.003360	0	0	-0.742596	0.071961	-0.100889	-0.811436
nonflavanoid_phenols	-0.298533	-0.028779	0	0	0.257193	0.216337	-0.077980	0.128603
proanthocyanins	0.313429	-0.039302	0	0	0.192331	-0.135110	0.012373	0.163924
color_intensity	-0.088617	-0.529996	0	0	0.232810	-0.238384	0.903389	0.061729
hue	0.296715	0.279235	0	0	0.026516	0.005303	0.089583	-0.026993
od280/od315	0.376167	0.164496	0	0	0.250281	0.197020	0.127810	0.102718
proline	0.286752	-0.364903	0	0	0.062119	0.736438	-0.227985	-0.004630

[1] https://github.com/takanori-fujiwara/ccpca

B Results for the PPIC Statewide Survey

The results corresponding to Fig. 9a are shown below. The analysis in Fig. 9a aims to find subgroups from the Democrat supporters (Dem) by comparing them with the Republican supporters (Rep). Thus, we apply all DR methods to the dataset and cPCA's α is automatically selected. From Fig. B.1, we can see that while PCA and LDA clearly separate Dem and Rep, they do not depict any clear subgroups. On the other hand, cPCA and ULCA show subgroups of Dem. However, ULCA with manually selected α more clearly shows two distinct subgroups and the attribute with a dominant contribution to the embedding (i.e., Q30).

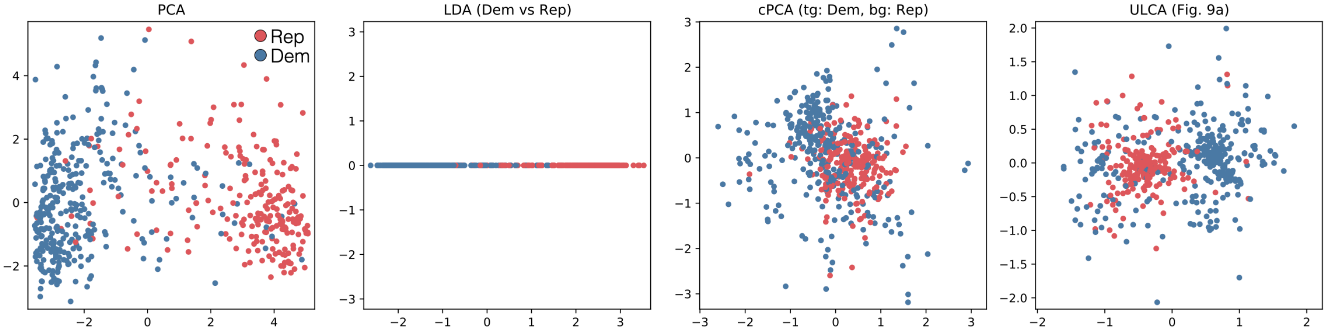

Figure B.1: The embedding results corresponding to Fig. 9a.

Table B.1: The axis information of Fig. B.1.

	PCA (x)	PCA (y)	LDA (x)	LDA (y)	cPCA (x)	cPCA (y)	ULCA (x)	ULCA (y)
q1	-0.262499	-0.029540	-0.135952	0.0	0.689553	0.162083	-0.344023	-0.050108
q7	0.061919	-0.412986	-0.097428	0.0	-0.013778	0.139085	0.036106	-0.019798
q13	-0.064540	-0.360961	0.018720	0.0	-0.161281	0.015959	-0.020793	-0.004293
q18	-0.013561	-0.076335	-0.069628	0.0	0.095442	-0.008482	0.001541	0.020085
q19	-0.078830	-0.045852	-0.178542	0.0	-0.006679	0.055236	0.057741	-0.037453
q20	0.301285	-0.056628	0.374915	0.0	0.168030	0.427085	0.183248	-0.504519
q21	0.172963	0.011879	-0.009094	0.0	0.068641	0.063169	0.046290	-0.021867
q21a	0.292849	-0.010057	0.352594	0.0	0.119974	0.193656	0.051198	0.651416
q22	0.264119	-0.079989	0.052627	0.0	0.117895	0.118933	-0.111115	0.027239
q23	0.202948	-0.056151	0.009101	0.0	-0.014212	0.194740	0.123681	-0.181099
q24	-0.242087	-0.080452	-0.101145	0.0	0.093145	-0.157120	0.064125	0.459837
q25	-0.238140	-0.035276	0.018476	0.0	0.223884	-0.015996	-0.000025	0.047087
q26	0.282375	-0.043620	-0.144249	0.0	0.293312	0.204617	-0.006155	0.053229
q27	-0.259116	0.005183	-0.087271	0.0	-0.277763	0.489935	0.053455	-0.122329
q28	-0.253654	-0.045732	-0.318113	0.0	-0.259232	0.232197	0.147868	0.002127
q29	-0.264244	-0.044850	-0.033479	0.0	-0.133985	0.002701	-0.127783	-0.110617
q30	-0.244617	-0.073278	-0.328858	0.0	-0.018518	0.143548	0.851559	0.024652
q31	0.265555	-0.000986	0.334397	0.0	-0.123369	-0.229083	-0.161481	-0.046006
q33	-0.257920	-0.076160	-0.274269	0.0	-0.038854	0.320887	-0.040543	-0.118824
q34	0.017638	-0.452132	-0.002937	0.0	-0.157850	0.156949	0.028108	0.094607
q36	0.028359	-0.385510	0.041148	0.0	-0.063469	0.196644	-0.047058	-0.067060
q37a	-0.003487	-0.078765	0.013674	0.0	0.047617	0.169870	-0.012261	-0.010002
d3a	0.023643	-0.250583	-0.011744	0.0	0.125461	-0.001833	-0.000281	0.027467
d6	-0.000859	-0.114435	-0.029078	0.0	-0.114432	0.109959	0.050348	0.017560
d7	-0.033413	-0.259972	0.123900	0.0	-0.112893	0.159850	0.067566	-0.047822
d9a	0.043446	-0.120803	0.038419	0.0	-0.116516	0.081919	0.018063	-0.056948
d1a	0.037022	-0.367580	0.024154	0.0	0.130929	-0.041311	-0.043981	0.041030

The results corresponding to Fig. 9b are shown below. The analysis in Fig. 9b is to find opinions more varied in `Dem(-)` than the other groups. Thus, only PCA and cPCA are related to this analysis. For cPCA, we set `Dem(-)` as a target group and the other two groups as an aggregated background group, and select α automatically. From Fig. B.2, when compared with the result with PCA, `Dem(-)` is more widely distributed than others when using cPCA and ULCA. Note that here we discarded `Q30` from attributes used in each DR; however, PCA has the result highly similar to Fig. B.1. This implies `Q30` has a limited influence on PCA's embedding, and the subgroup we found in Fig. B.1 with ULCA or cPCA is difficult to find with PCA. When comparing cPCA and ULCA, ULCA shows separation between `Dem(+)` and `Rep`. We can consider that this is because ULCA can individually handle each of the target groups, which is infeasible with cPCA. When computing covariance matrices, ULCA applies centering to each of the groups individually; as a result, in the embedding space, each group's variance is condensed around its centroid, unlike cPCA where the common centroid of the aggregated group is used. ULCA can provide more precise controls to embed datasets than cPCA, especially when we have more than two groups.

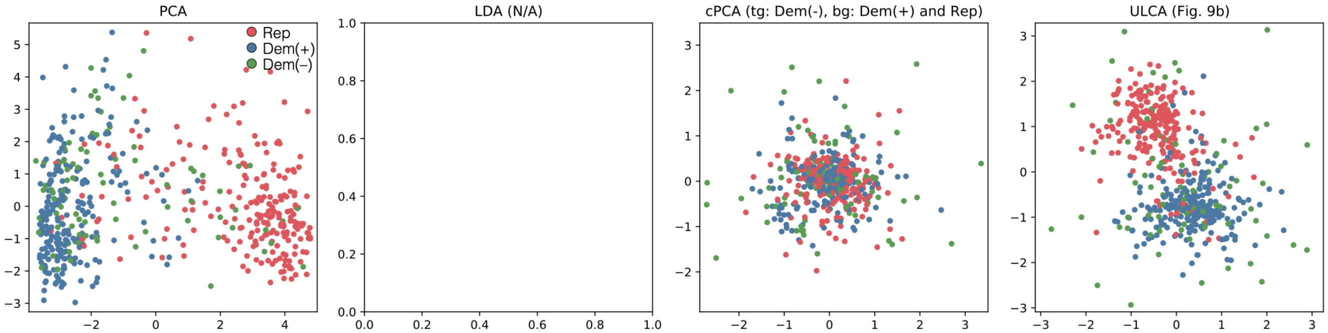

Figure B.2: The embedding results corresponding to Fig. 9b.

Table B.2: The axis information of Fig. B.2.

	PCA (x)	PCA (y)	LDA (x)	LDA (y)	cPCA (x)	cPCA (y)	ULCA (x)	ULCA (y)
q1	-0.272006	-0.025406	0	0	0.472050	0.373206	0.558180	0.159270
q7	0.051561	-0.415685	0	0	-0.085978	-0.005874	-0.183562	0.071507
q13	-0.068367	-0.347333	0	0	-0.065163	0.004781	0.015901	-0.270734
q18	-0.015867	-0.121445	0	0	0.001138	-0.024359	-0.015735	-0.003108
q19	-0.086224	-0.067838	0	0	0.055548	0.118862	0.093673	-0.043103
q20	0.307895	-0.044009	0	0	0.199424	0.494824	-0.132921	0.339701
q21	0.180770	0.015282	0	0	0.032325	0.045913	0.086195	-0.030779
q21a	0.296603	0.011459	0	0	0.197613	-0.489857	-0.222388	0.412679
q22	0.270156	-0.062986	0	0	0.194321	-0.017552	0.118428	0.050037
q23	0.203250	-0.043077	0	0	-0.073007	0.009687	-0.157588	0.100453
q24	-0.253813	-0.053852	0	0	0.271304	-0.120468	0.317152	0.080041
q25	-0.252807	-0.008591	0	0	0.386441	-0.096710	-0.013817	0.433229
q26	0.288913	-0.046929	0	0	0.373558	-0.115602	0.225641	0.314264
q27	-0.266678	0.006934	0	0	-0.153651	0.278042	0.069138	-0.276877
q28	-0.261923	-0.055755	0	0	-0.340273	0.052775	-0.192458	-0.321027
q29	-0.273050	-0.041173	0	0	-0.090726	-0.413597	-0.381342	-0.012470
q31	0.276355	-0.021581	0	0	-0.258279	0.108430	0.000249	-0.175618
q33	-0.270587	-0.074903	0	0	0.163379	-0.122102	0.000337	0.086607
q34	0.013855	-0.446806	0	0	0.118189	-0.073985	0.152876	0.019129
q36	0.024034	-0.365329	0	0	0.005342	-0.101658	-0.247427	0.155303
q37a	-0.002479	-0.094320	0	0	0.032020	0.027129	0.095491	-0.030879
d3a	0.013835	-0.259995	0	0	0.010438	0.113828	0.159899	-0.062622
d6	-0.010729	-0.079336	0	0	-0.056727	-0.018921	-0.046021	-0.178001
d7	-0.039018	-0.279497	0	0	-0.035465	0.047200	-0.087423	0.133693
d9a	0.043838	-0.115847	0	0	-0.058634	-0.100970	-0.192951	0.056562
d1a	0.032435	-0.393590	0	0	0.113705	0.035856	0.163867	0.043843

The results corresponding to Fig. 9c are shown below. The analysis in Fig. 9c is to find opinions more varied in `Dem(-)` than the other groups but at the same time clearly distinguish `Dem(+)` and `Rep`. Thus, we apply all the DR methods (LDA relates to finding the separation while cPCA related to finding more diverse opinions in `Dem(-)`). For cPCA, we use the same setting with Fig. B.2. For LDA, to focus on distinguishing `Dem(+)` and `Rep`, we only use data points in `Dem(+)` and `Rep` while learning a projection matrix, and utilize this projection matrix to plot `Dem(-)` in the same embedding space. From Fig. B.3, unlike the other methods, ULCA shows the clear separation between `Dem(+)` and `Rep` while producing high variance for `Dem(-)`.

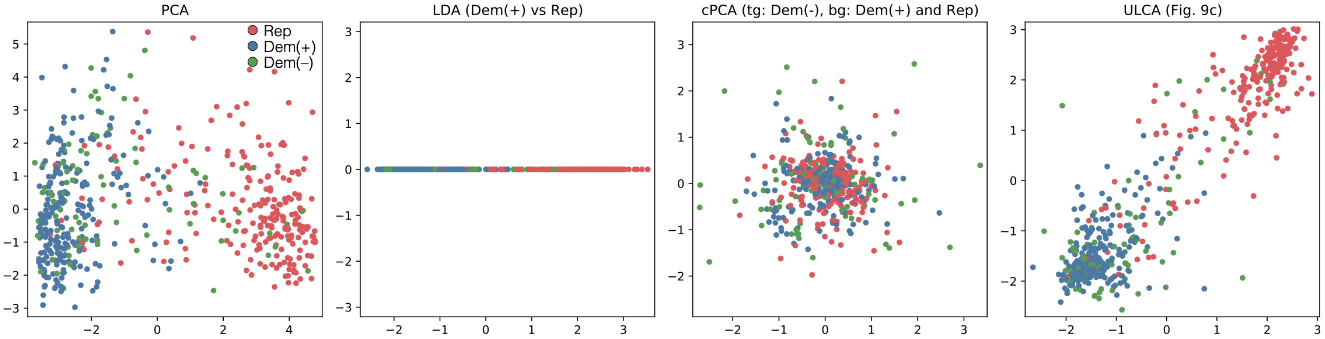

Figure B.3: The embedding results corresponding to Fig. 9c.

Table B.3: The axis information of Fig. B.3.

	PCA (x)	PCA (y)	LDA (x)	LDA (y)	cPCA (x)	cPCA (y)	ULCA (x)	ULCA (y)
q1	-0.272006	-0.025406	-0.181956	0.0	0.472050	0.373206	-0.191644	-0.124360
q7	0.051561	-0.415685	-0.113320	0.0	-0.085978	-0.005874	0.107413	0.004499
q13	-0.068367	-0.347333	-0.011961	0.0	-0.065163	0.004781	-0.075113	-0.083833
q18	-0.015867	-0.121445	-0.059593	0.0	0.001138	-0.024359	-0.062652	-0.075154
q19	-0.086224	-0.067838	-0.165659	0.0	0.055548	0.118862	-0.071863	-0.086774
q20	0.307895	-0.044009	0.449496	0.0	0.199424	0.494824	0.790036	-0.142629
q21	0.180770	0.015282	0.003009	0.0	0.032325	0.045913	0.002213	0.031159
q21a	0.296603	0.011459	0.346843	0.0	0.197613	-0.489857	0.023907	0.676308
q22	0.270156	-0.062986	0.047464	0.0	0.194321	-0.017552	0.032788	0.171582
q23	0.203250	-0.043077	0.022052	0.0	-0.073007	0.009687	0.169506	0.060299
q24	-0.253813	-0.053852	-0.159578	0.0	0.271304	-0.120468	-0.347739	0.057961
q25	-0.252807	-0.008591	0.014950	0.0	0.386441	-0.096710	0.028889	0.156015
q26	0.288913	-0.046929	-0.113532	0.0	0.373558	-0.115602	-0.144949	0.331211
q27	-0.266678	0.006934	-0.087042	0.0	-0.153651	0.278042	0.016656	-0.235233
q28	-0.261923	-0.055755	-0.341739	0.0	-0.340273	0.052775	-0.120233	-0.436055
q29	-0.273050	-0.041173	-0.080050	0.0	-0.090726	-0.413597	-0.046404	-0.123246
q31	0.276355	-0.021581	0.311197	0.0	-0.258279	0.108430	0.180075	0.136549
q33	-0.270587	-0.074903	-0.271110	0.0	0.163379	-0.122102	-0.182615	-0.109055
q34	0.013855	-0.446806	0.003489	0.0	0.118189	-0.073985	-0.134511	0.087867
q36	0.024034	-0.365329	0.035556	0.0	0.005342	-0.101658	0.086948	0.047533
q37a	-0.002479	-0.094320	0.014897	0.0	0.032020	0.027129	-0.044884	0.000106
d3a	0.013835	-0.259995	-0.020982	0.0	0.010438	0.113828	-0.011600	0.075408
d6	-0.010729	-0.079336	-0.040659	0.0	-0.056727	-0.018921	0.027291	0.061168
d7	-0.039018	-0.279497	0.137705	0.0	-0.035465	0.047200	0.149902	0.052898
d9a	0.043838	-0.115847	0.043950	0.0	-0.058634	-0.100970	0.076157	0.040297
d1a	0.032435	-0.393590	0.021741	0.0	0.113705	0.035856	-0.014708	0.009417

C Results for the MNIST Handwritten Digits Dataset

The results corresponding to Fig. 10a are shown below. In the analysis related to Fig. 10a, we review highly diverse structures in 6 and 9 but not in 0. Thus, for this analysis, we do not use LDA. For cPCA, we set 6 and 9 as a target group and 0 as a background group, and α is selected automatically. From Fig. C.1 and Fig. C.2, we can see that only ULCA clearly captures 9's straight lines mentioned in the case study.

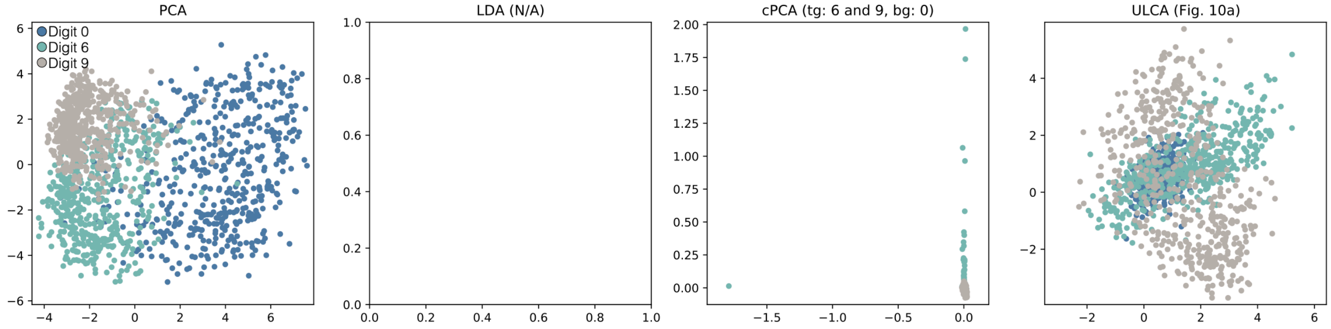

Figure C.1: The embedding results corresponding to Fig. 10a.

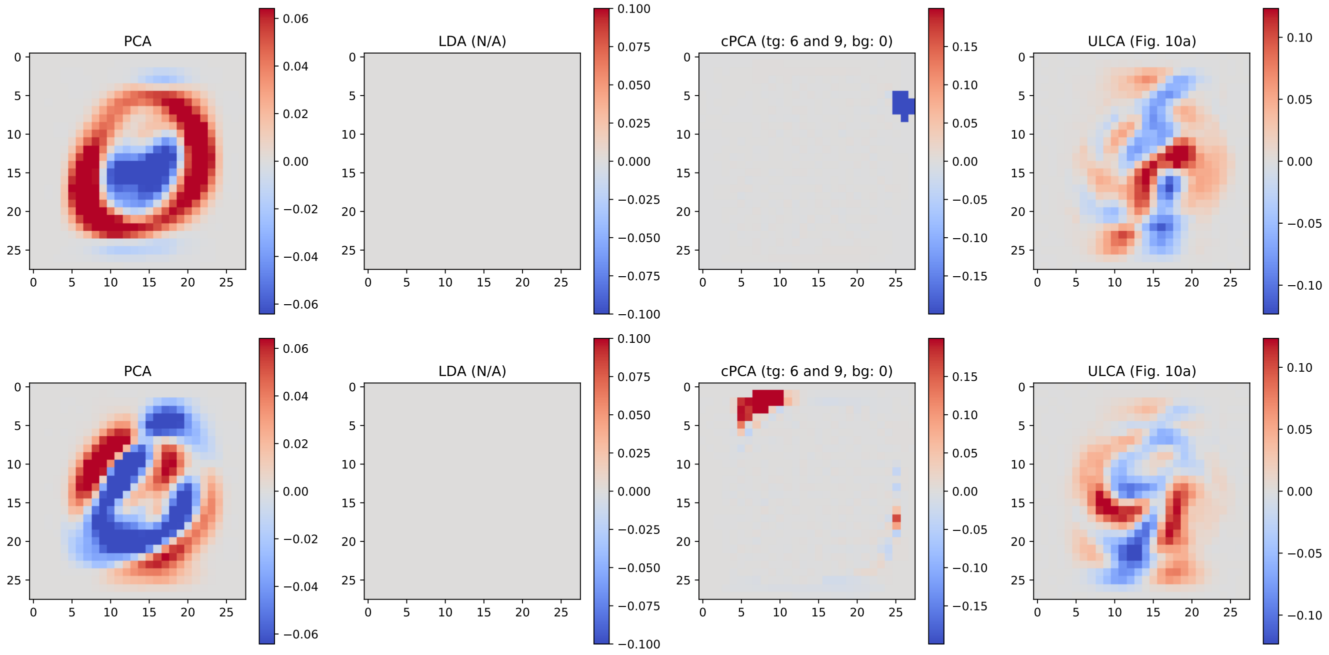

Figure C.2: The axis information of Fig. C.1. The heatmaps placed in the first and second rows correspond to x- and y-embedding axes, respectively.

The results corresponding to Fig. 10b are shown below. In Fig. 10b, we adjust the weight related to 9's variance. In Fig. C.4, we can see that ULCA captures the strokes in 6 better than Fig. C.2.

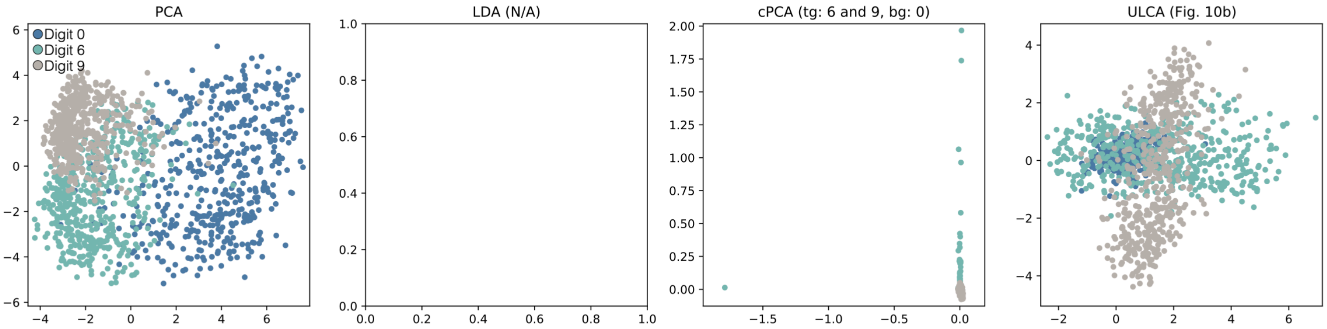

Figure C.3: The embedding results corresponding to Fig. 10b.

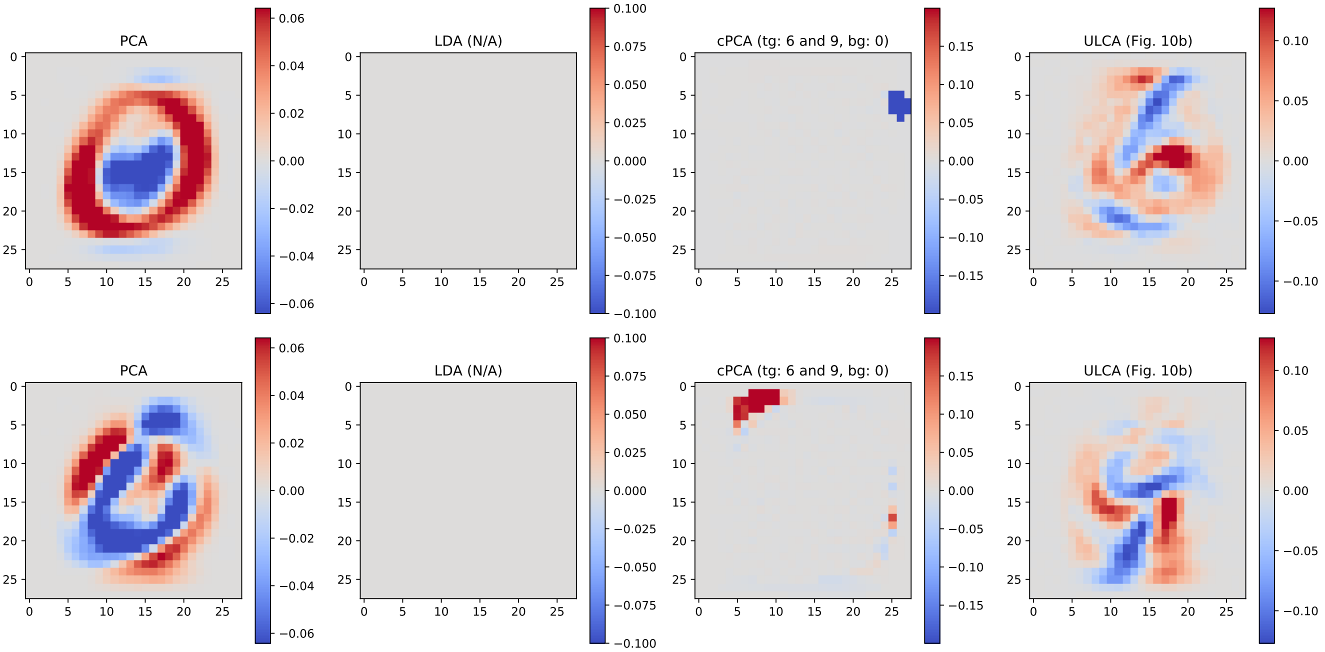

Figure C.4: The axis information of Fig. C.3.

The results corresponding to Fig. 10c are shown below. In Fig. 10c, we review the strokes that clearly differentiate 6 and 9 from 0 but are still written in various ways for 6 and 9. Thus, for this analysis, we apply LDA to distinguish an aggregated group of 6 and 9 from a group of 0. Again, as shown in Fig. C.5, only ULCA achieves to generate the embedding result that satisfies our analysis purpose. Also, from Fig. C.6, only ULCA reveals the useful information about the strokes.

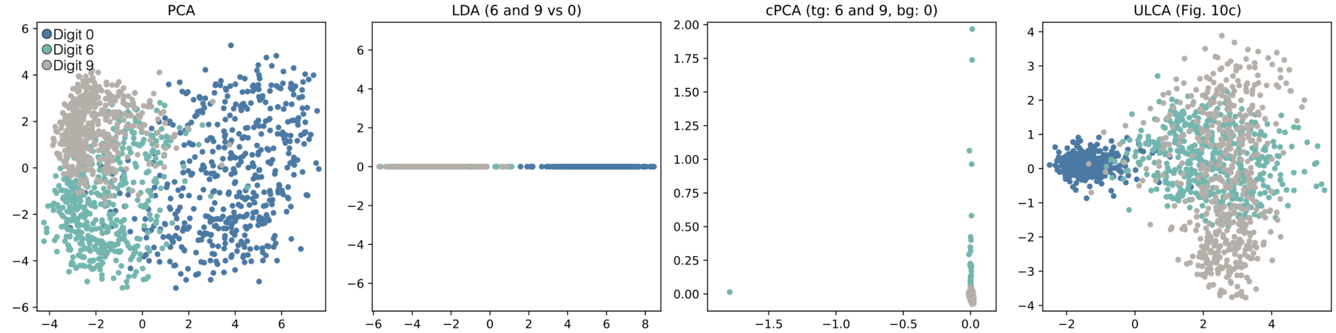

Figure C.5: The embedding results corresponding to Fig. 10c.

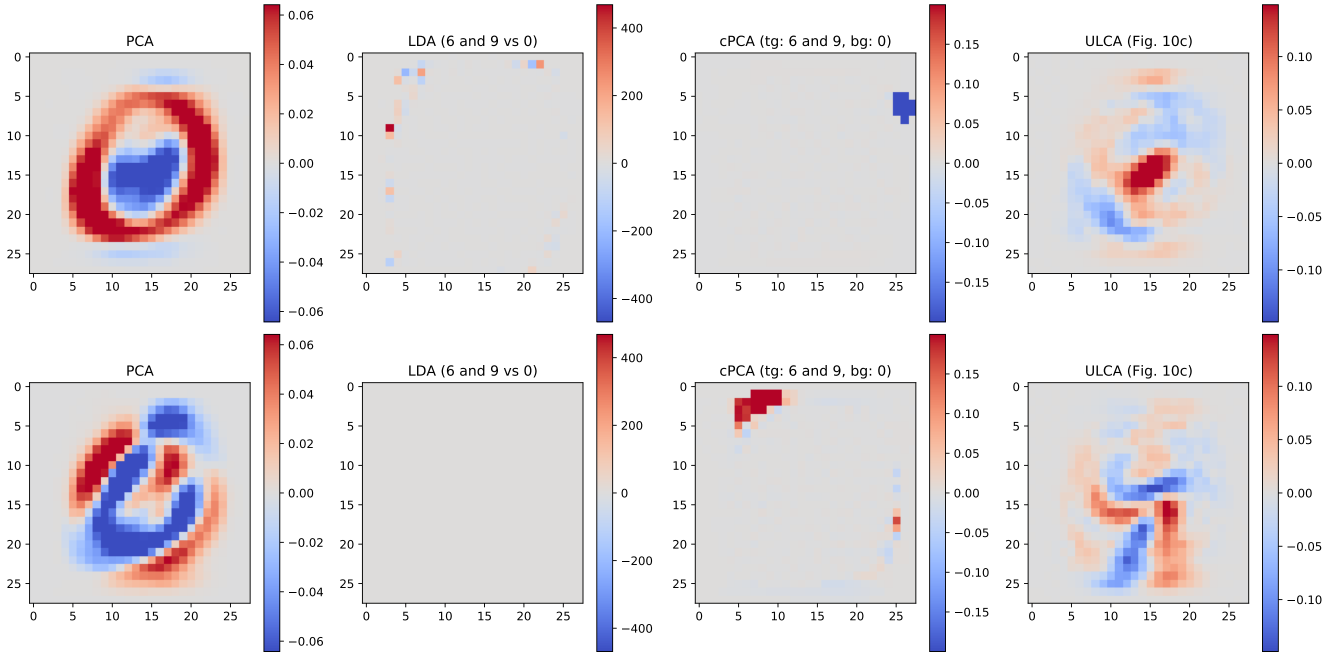

Figure C.6: The axis information of Fig. C.5.