Supervised PCA via Manifold Optimization

Alex Ritchie[1]

Joint with Clay Scott[1,2], Laura Balzano[1], Daniel Kessler[2,3], Chandra Sripada[3,4]

EECS[1], *Statistics*[2], *Psychiatry*[3], *Philosophy*[4]
University of Michigan

aritch@umich.edu

September 23, 2019

Overview

1. Introduction and Previous Work
2. Our Work
3. Experimental Results

Introduction and Previous Work

Dimensionality Reduction (DR)

$$\overbrace{\begin{bmatrix} x_{11} & x_{12} & \dots & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & \dots & x_{np} \end{bmatrix}}^{p \text{ features}} \Longrightarrow \overbrace{\begin{bmatrix} \tilde{x}_{11} & \dots & \tilde{x}_{1k} \\ \tilde{x}_{21} & \dots & \tilde{x}_{2k} \\ \vdots & \vdots & \vdots \\ \tilde{x}_{n1} & \dots & \tilde{x}_{nk} \end{bmatrix}}^{k \ll p \text{ features}}$$

Potential Benefits

- Better Generalization
- Higher Interpretability
- Lower Computational Cost

What are the new features?

- Feature Selection (choose a subset of original features)
- Feature Extraction (learn new features from originals)

Principal Component Analysis (PCA)

PCA problem statement

$$\underset{L}{\text{minimize}} \quad \|X - XL^T L\|_F^2$$
$$s.t. \quad LL^T = I_{kxk}$$

$\underset{n\times p}{X}$	Data matrix
$\underset{k\times p}{L}$	Basis for reduced X

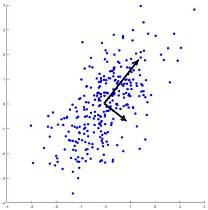


Figure: PCA of MVG data.

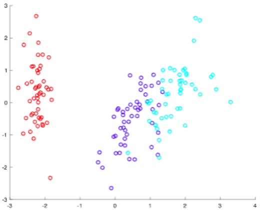


Figure: 2D PCA embedding of Iris data.

Supervised Dimensionality Reduction (SDR)

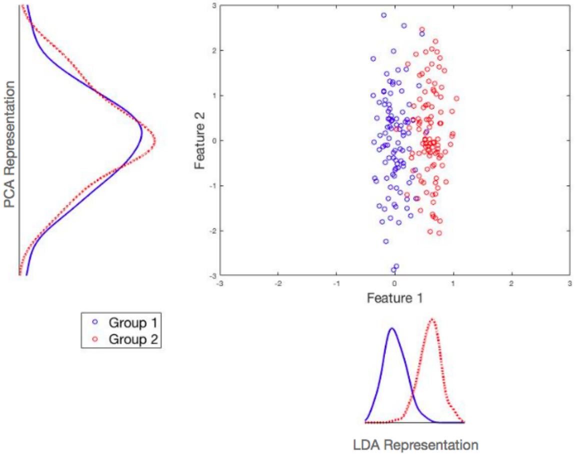


Figure: Principal Component Analysis (Unsupervised) vs. Linear Discriminant Analysis (Supervised).

A Real Example: Single Cell Data

- Significant variation attributed to housekeeping genes, or genes unrelated to feature of interest
- Important variation of minority cell populations obscured by majority cell population
- Consider Neural Stem Cell Lineage Dataset [8]

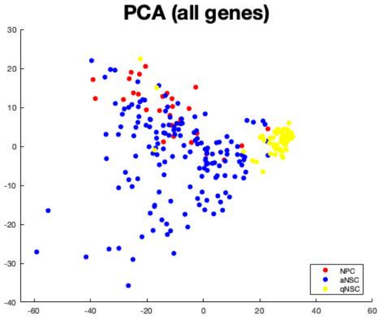
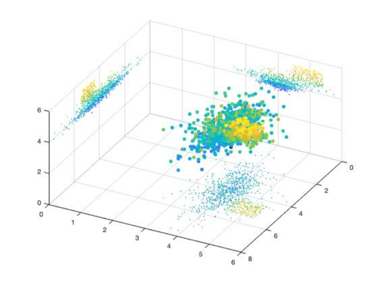


Figure: Projection of Dulken data onto first two PCs.

SDR: High Level Approach and Goals

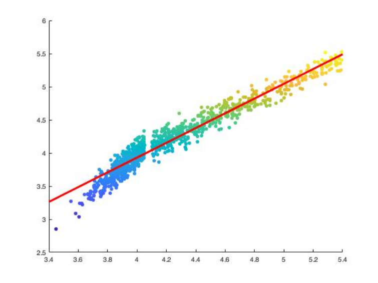
In practice, DR is often a preliminary step before prediction. It is natural to consider performing DR and prediction jointly. One set of approaches is to add supervision to PCA (SPCA).

Dimensionality Reduction



+

Prediction



Previous Work: Supervised Principal Components [1]

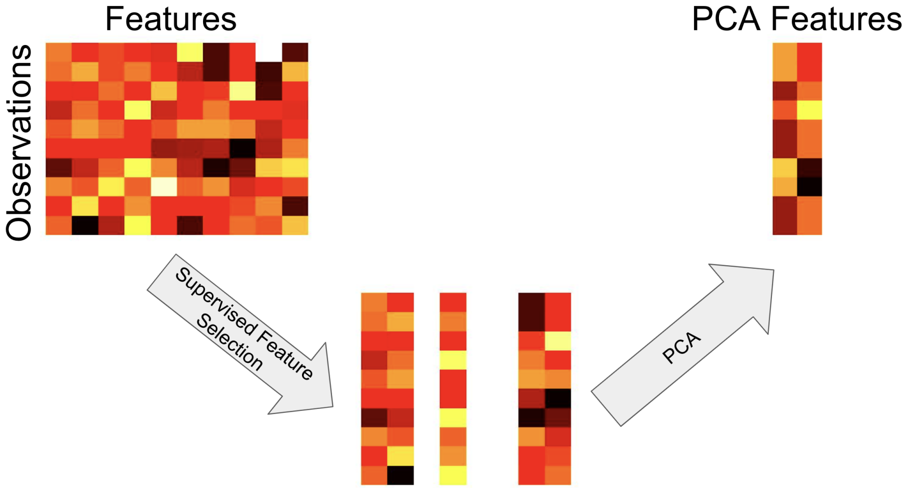


Figure: Illustrative example for Bair's method.

Previous Work: Supervised Probabilistic PCA [2]

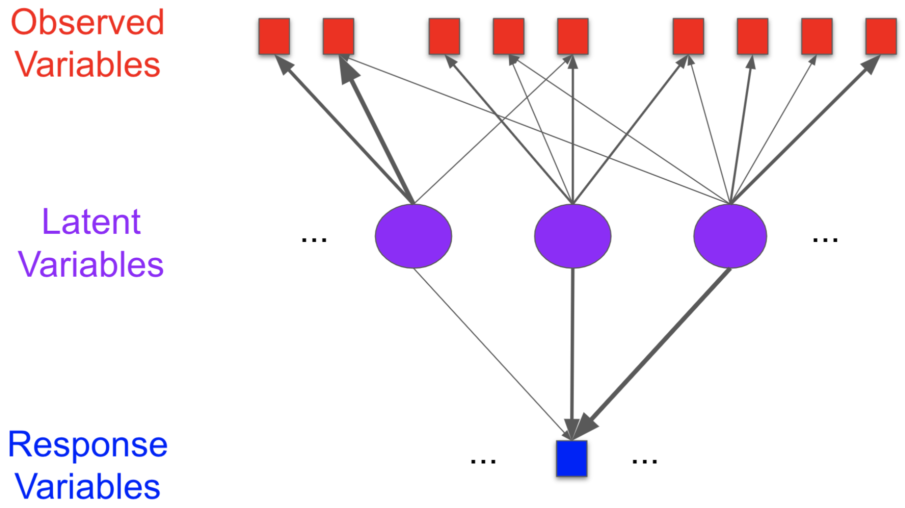


Figure: Latent variable generative model for supervised probabilistic principal component analysis.

Observations

Other Methods

- Supervised PCA via Hilbert-Schmidt independence criterion [3]
- Iterative supervised principal components [4]
- Supervised Singular Value Decomposition [6]
- Partial Least Squares
- Canonical Correlation Analysis

High Level Shortcomings:

- Do not consider objectives jointly
- No means of trade-off between prediction and variation

Our Work

Problem Statement

LSPCA Optimization Program

$$\underset{\beta,L}{\text{minimize}} \quad \mathcal{L}(Y, XL^T, \beta) + \lambda\|X - XL^TL\|_F^2$$

$$s.t. \quad LL^T = I_{kxk}$$

$\underset{n\times p}{X}$	Data matrix
$\underset{n\times q}{Y}$	Dependent variables
$\underset{k\times p}{L}$	Basis for reduced X
$\underset{k\times q}{\beta}$	Prediction Coefficients

Table: Variables

λ	regularization parameter
k	subspace dimension, $k \leq min(n, p)$

Table: Hyperparameters

Optimization on the Grassmannian

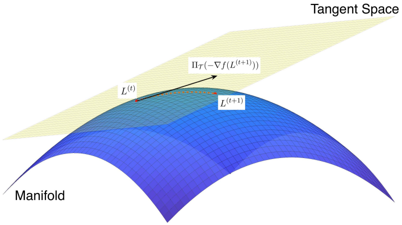


Figure: Visualization of a gradient step on the Grassmannian.

Algorithm 1 Manifold Gradient Descent for LSPCA

procedure $\mathrm{LSPCA}(X, Y, L_0, \lambda, k)$
    $t = 0$
    while Not Converged **do**
        $\nabla f(L_t) = 2(1-\lambda)L_t X^T X - 2\lambda(XL_t^T)^+ YY^T P^{\perp}_{XL^T} X$
        $H_t = -(I_{p\times p} - L_t^T L_t)\nabla f(L_t)^T$
        $U_t, \Sigma_t, V_t = \mathrm{SVD}(H_t)$
        $L_{t+1}^T = L_t^T V_t \cos(\eta_t \Sigma_t) V_t^T + U_t \sin(\eta_t \Sigma_t) V_t^T$
        $t \leftarrow t + 1$
    end while
    $Z = XL_t^T$
    return Z, L_t
end procedure

LSPCA Statistical Model

Interpreting LSPCA as MLE

$$y|x \sim N(\beta^T Lx, \sigma_y^2 I), \quad x|z \sim N(L^T z, \sigma_x^2 I), \quad z \sim N(0, \sigma_z^2 I)$$

We form the joint distribution

$$\begin{aligned} f_{x,y}(x,y) &= f_{y|x}(y|x) \int_{-\infty}^{\infty} f_{x|z}(x) f_z(z) dz \\ &\propto \exp(-\frac{1}{2\sigma_y^2}\|y - \beta' Lx\|_2^2 - \frac{\sigma_z^2}{2\sigma_x^2(\sigma_x^2+\sigma_z^2)} x'(\frac{\sigma_x^2+\sigma_z^2}{2\sigma_z^2} I - L'L)x). \end{aligned}$$

In the case $\sigma_x = \sigma_z$

$$\boxed{-\log(\ell(L,\beta)) \propto \|Y - XL^T\beta\|_F^2 + \frac{\sigma_y^2}{2\sigma_x^2}\|X - XL^T L\|_F^2.}$$

Experimental Results

Evaluation Criteria

In addition to prediction accuracy, we consider proportion of variation explained as an evaluation criteria. This is analogous to how the k is often chosen in PCA.

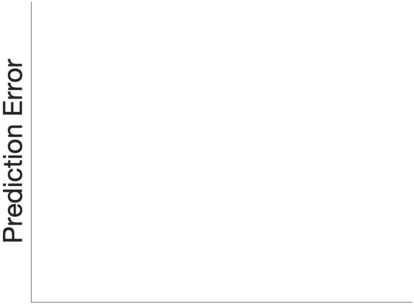


Figure: Multi-objective evaluation criteria.

Evaluation Criteria

In addition to prediction accuracy, we consider proportion of variation explained as an evaluation criteria. This is analogous to how the k is often chosen in PCA.

Figure: Multi-objective evaluation criteria.

Evaluation Criteria

In addition to prediction accuracy, we consider proportion of variation explained as an evaluation criteria. This is analogous to how the k is often chosen in PCA.

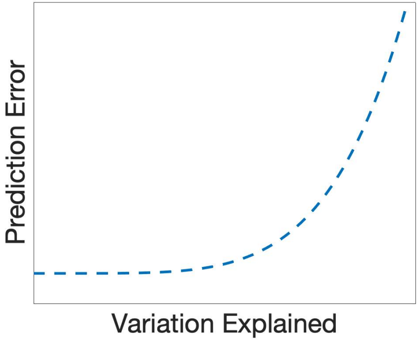


Figure: Multi-objective evaluation criteria.

Evaluation Criteria

In addition to prediction accuracy, we consider proportion of variation explained as an evaluation criteria. This is analogous to how the k is often chosen in PCA.

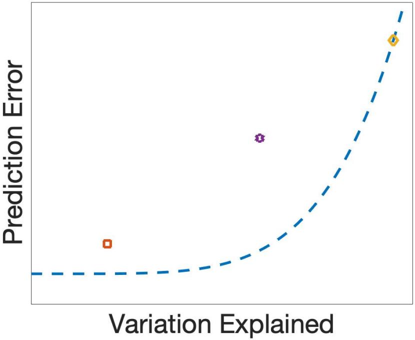


Figure: Multi-objective evaluation criteria.

Evaluation Criteria

In addition to prediction accuracy, we consider proportion of variation explained as an evaluation criteria. This is analogous to how the k is often chosen in PCA.

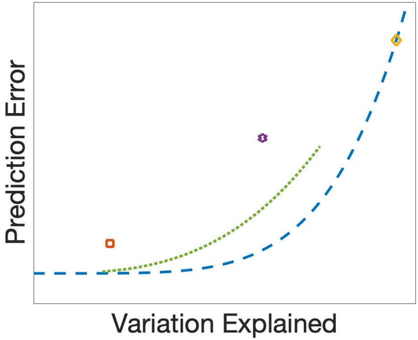


Figure: Multi-objective evaluation criteria.

Regression with Synthetic Data

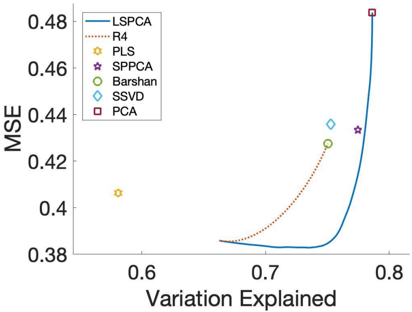


Figure: Tradeoff between prediction and variation explained on synthetic data generated according to LSPCA model.

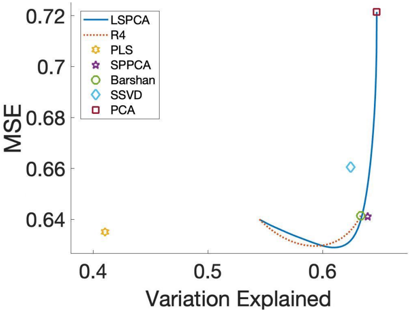


Figure: Tradeoff between prediction and variation explained on synthetic data generated according to SPPCA model.

Regression with Real Data

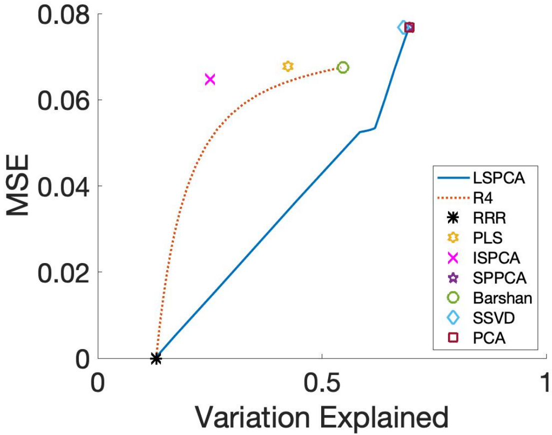


Figure: Tradeoff between prediction and variation explained on Music Dataset [5].

Classification Results

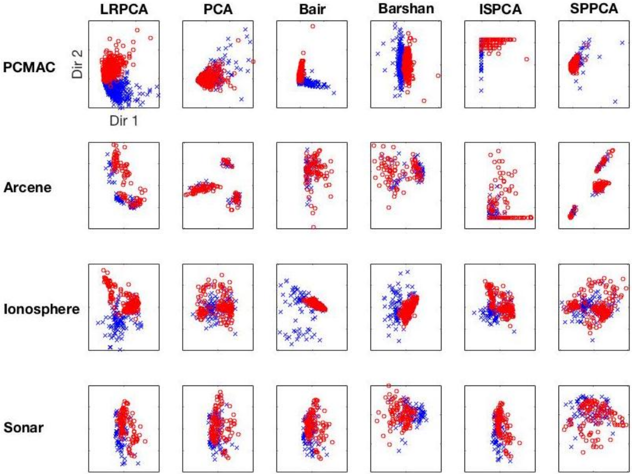


Figure: Visualization of binary classification datasets using the first two learned features of each method. Colors serve as class labels.

Classification Results

Table: Classification accuracy on binary classification datasets for $k = 2$. Results are averaged over 5 independent runs. Standard deviation of classification accuracy is given in parentheses.

Dataset	LRPCA	PCA	Bair	Barshan	ISPCA	LDA
PCMAC	0.893(0.0143)	0.520(0.075)	0.577(0.109)	0.893(0.0139)	0.727(0.039)	0.690(0.045)
Arcene	0.693(0.037)	0.660(0.029)	0.693(0.077)	0.693(0.052)	0.700(0.082)	0.823(0.040)
Ionosphere	0.868(0.036)	0.630(0.051)	0.811(0.086)	0.856(0.067)	0.797(0.069)	0.865(0.024)
Sonar	0.717(0.024)	0.552(0.090)	0.588(0.120)	0.681(0.034)	0.705(0.062)	0.693(0.065)

Table: Variation explained on binary classification datasets for $k = 2$. Results are averaged over 5 independent runs. Format: train variation explained / test variation explained.

Dataset	LRPCA	PCA	Bair	Barshan	ISPCA	LDA
PCMAC	0.229/0.095	0.274/0.092	0.199/0.067	0.071/0.052	0.103/0.039	0.008/0.015
Arcene	0.523/0.246	0.568/0.269	0.236/0.095	0.391/0.183	0.043/0.018	0.095/0.041
Ionosphere	0.586/0.582	0.624/0.650	0.452/0.458	0.395/0.390	0.477/0.486	0.177/0.177
Sonar	0.565/0.543	0.628/0.609	0.517/0.511	0.374/0.361	0.373/0.379	0.058/0.056

Interpretable (Predictive) Features

Linear Prediction

$$\underset{\beta, L}{\text{minimize}} \quad \sum_{i=1}^{N} \log(1 + e^{-y_i \mathbf{x}_i^T L^T \beta}) + \lambda \|X - XL^T L\|_F^2$$

$$s.t. \quad LL^T = I_{kxk}$$

- Form filter from basis and coefficients
- Threshold filter to find most meaningful components

Interpretable Features for MNIST

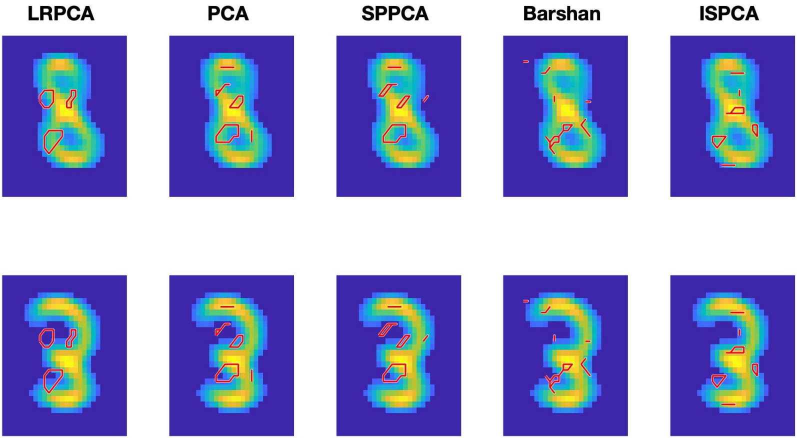


Figure: Visualization of most important features for classification of threes and eights. Obtained by thresholding

Future Work

- Extension to Domain Adaptation and Semi-Supervised Learning
- Statistical Behavior of Estimator
- Stronger Algorithmic Guarantee

References I

[1] Bair, Eric, Trevor Hastie, Debashis Paul, and Robert Tibshirani. "Prediction by supervised principal components." Journal of the American Statistical Association 101, no. 473 (2006): 119-137.

[2] Yu, Shipeng, Kai Yu, Volker Tresp, Hans-Peter Kriegel, and Mingrui Wu. "Supervised probabilistic principal component analysis." In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 464-473. ACM, 2006.

[3] Barshan, Elnaz, Ali Ghodsi, Zohreh Azimifar, and Mansoor Zolghadri Jahromi. "Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds." Pattern Recognition 44, no. 7 (2011): 1357-1371.

[4] Piironen, Juho, and Aki Vehtari. "Iterative supervised principal components." arXiv preprint arXiv:1710.06229 (2017).

[5] Zhou, Fang, Q. Claire, and Ross D. King. "Predicting the geographical origin of music." In 2014 IEEE International Conference on Data Mining, pp. 1115-1120. IEEE, 2014.

References II

[6] Li, Gen, Dan Yang, Andrew B. Nobel, and Haipeng Shen. "Supervised singular value decomposition and its asymptotic properties." Journal of Multivariate Analysis 146 (2016): 7-17.

[7] Edelman, Alan, Toms A. Arias, and Steven T. Smith. "The geometry of algorithms with orthogonality constraints." SIAM journal on Matrix Analysis and Applications 20, no. 2 (1998): 303-353.

[8] Dulken, B. W., Leeman, D. S., Boutet, S. C., Hebestreit, K., Brunet, A. (2017). Single-cell transcriptomic analysis defines heterogeneity and transcriptional dynamics in the adult neural stem cell lineage. Cell reports, 18(3), 777-790.