Supervised Principal Component Analysis via Manifold Optimization

Alex Ritchie[1], Clayton Scott[1], Laura Balzano[1], Daniel Kessler[2,3], Chandra Sripada[3]

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

Introduction

High dimensional prediction problems are pervasive in the scientific community. In practice, dimensionality reduction (DR) is often performed as an initial step to improve prediction accuracy and interpretability. Principal component analysis (PCA) has been utilized extensively for DR, but does not take advantage of outcome variables inherent in the prediction task. Existing approaches for supervised PCA (SPCA) either take a multi-stage approach or incorporate supervision indirectly. We present a manifold optimization approach to SPCA that simultaneously solves the prediction and dimension reduction problems. The proposed framework is general enough for both regression and classification settings. Our empirical results show that the proposed approach explains nearly as much variation as PCA while outperforming existing methods in prediction accuracy.

Motivation

- From a component analysis point of view, it is natural think of learning predictive components
- When supervisory information is available, we would like to utilize it for DR
- Existing approaches to SPCA either take a multi-stage approach, or cannot trade off between prediction and representation

SPCA as Non-convex Manifold Optimization

We can write SPCA as a regularized manifold optimization problem with a general loss function:

$$\begin{aligned} &\underset{L,\beta}{\text{minimize}} \quad \sum_{i=1}^{n} l(\boldsymbol{y}_i, L\boldsymbol{x}_i, \beta) + \lambda \|X - XL^T L\|_F^2 \\ &s.t. \;\; LL^T = I_{kxk} \end{aligned}$$

Note that as $\lambda \to \infty$, the problem reduces to PCA. We choose this framework because it is general enough to handle many problem settings including classification, regression, and semi-supervised. For each, we must select the proper loss function.

Table: Variables for all problems considered.

VARIABLE	DESCRIPTION
$\underset{n\times p}{X}$	Data matrix
$\underset{n\times q}{Y}$	Dependent variables
$\underset{k\times p}{L}$	Matrix with learned components as rows
$\underset{n\times k}{XL^T}$	Dimension reduced form of X
$\underset{k\times q}{\hat{\beta}}$	Loss function parameters

Hierarchical Bayesian Interpretation

For the regression setting, we may want to use the least squares cost function in the Manifold Optimization setup (LSPCA). It can be shown that this arises naturally under a specific hierarchical Bayesian model. This model adapts Probabilistic Principal Component Analysis, similar to previous work [2], with the benefit that we obtain a closed form for the likelihood by conditioning y on x rather than z. Suppose

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

Then it can be shown that, as $\sigma_x \to \sigma_z$ the negative log likelihood corresponding to the joint distribution on x and y is

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

This is clearly a special case of the SPCA Manifold Optimization problem, with the added benefit of interpretability of, and guidance in, setting the regularization parameter.

Algorithm

It can be shown that this algorithm asymptotically converges to a first order order stationary point [1].

Algorithm 1 Manifold Gradient Descent for LSPCA

procedure LSPCA(X, Y, L_0, λ, k)
  $t = 0$
  while Not Converged **do**
    $\nabla f(L_t) = 2(1-\lambda)L_t X^T X - 2\lambda(XL_t^T)^+ YY^T P^{\perp}_{XL^T} X$
    $H_t = -(I_{p\times p} - L_t^T L_t)\nabla f(L_t)^T$
    $U_t, \Sigma_t, V_t = \text{SVD}(H_t)$
    $L_{t+1}^T = L_t^T V_t \cos(\eta_t \Sigma_t) V_t^T + U_t \sin(\eta_t \Sigma_t) V_t^T$ ▷ Where η is a step size chosen by Armijo backtracking line-search.
    $t \leftarrow t + 1$
  end while
  $Z = XL_t^T$ ▷ Generate the reduced data.
  return Z, L_t
end procedure

Experimental Results

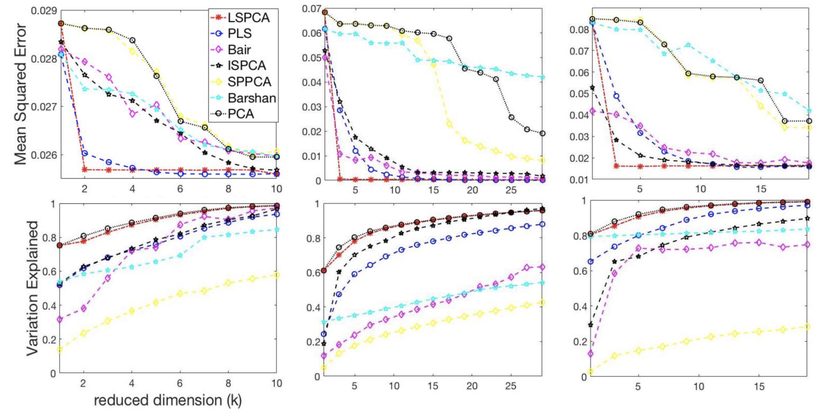

Figure: Mean squared error and variation explained vs. reduced dimension for each method on various datasets. (a) Parkinsons (b) Music (c) Residential

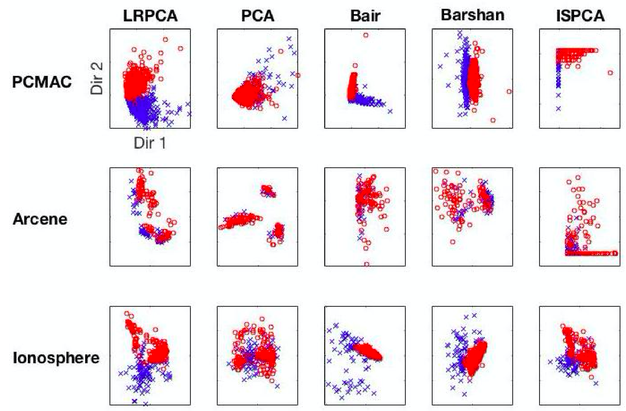

Figure: Example two dimensional visualizations using various popular methods.

Acknowledgements

ELECTRICAL & COMPUTER ENGINEERING
UNIVERSITY OF MICHIGAN

LSA STATISTICS
UNIVERSITY OF MICHIGAN

PSYCHIATRY
UNIVERSITY OF MICHIGAN
HEALTH SYSTEM

References

[1] Absil, P.A., Mahony, R. and Sepulchre, R., 2009. Optimization algorithms on matrix manifolds. Princeton University Press.

[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] Bair, E., Hastie, T., Paul, D. and Tibshirani, R., 2006. Prediction by supervised principal components. Journal of the American Statistical Association, 101(473), pp.119-137.

[5] Piironen, Juho, and Aki Vehtari. "Iterative Supervised Principal Components." In International Conference on Artificial Intelligence and Statistics, pp. 106-114. 2018.