

Dimensionality Reduction and Unsupervised Learning

Greg Grudic

1

Outline

- Principle Component Analysis (PCA)
- Independent Component Analysis (ICA)
- Principle Curves
- Data with outputs
- Partial Least Squares (PLS)
- K Means
- Spectral Clustering
- Locally Linear Embedding (LLE)

2

Principle Component Analysis (PCA)

- Assume data $D = \{(\mathbf{x}_1), \dots, (\mathbf{x}_N)\}$, $\mathbf{x}_i \in \mathbb{R}^d$
- Find k projections in input space

$$z_1 = \sum_{i=1}^d a_{1i} x_i$$

$$z_2 = \sum_{i=1}^d a_{2i} x_i$$

\vdots

$$z_k = \sum_{i=1}^d a_{ki} x_i$$

3

PCA II

- The k projections are called the k principle components
 - The principle components are uncorrelated
- The k projections are the eigenvectors of the data correlation matrix

4

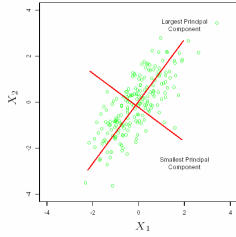


Figure 3.8: *Principal components of some input data points. The largest principal component is the direction that maximizes the variance of the projected data, and the smallest principal component minimizes that variance. Ridge regression projects \mathbf{y} onto these components, and then shrinks the coefficients of the low-variance components more than the high-variance components.*

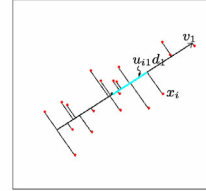


Figure 14.20: *The first linear principal component of a set of data. The line minimizes the total squared distance from each point to its orthogonal projection onto the line.*

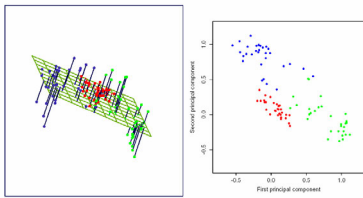


Figure 14.21: *The best rank-two linear approximation to the half-sphere data. The right panel shows the projected points with coordinates given by $\mathbf{U}_2\mathbf{D}_2$, the first two principal components of the data.*

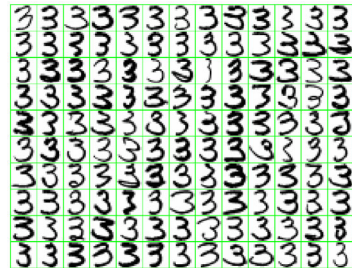


Figure 14.22: *A sample of 130 handwritten threes shows a variety of writing styles.*

$$\begin{aligned}\hat{f}(\lambda) &= \bar{x} + \lambda_1 v_1 + \lambda_2 v_2 \\ &= \boxed{3} + \lambda_1 \cdot \boxed{3} + \lambda_2 \cdot \boxed{3}.\end{aligned}$$

Here we have displayed the first two principal component directions, v_1 and v_2 , as images.

9

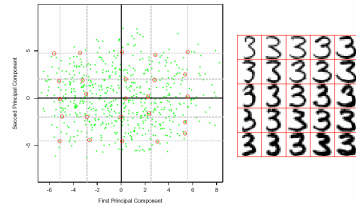


Figure 14.23: *Left plot: the first two principal components of the handwritten threes. The circled points are the closest projected images to the vertices of a grid, defined by the marginal quantiles of the principal components. Right plot: the images corresponding to the red-circled points. These show the nature of the first two principal components.*

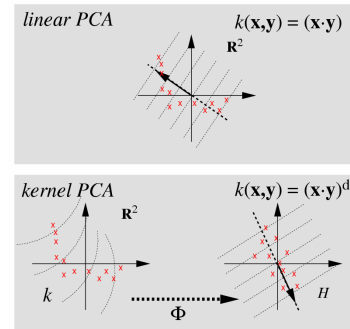
10

Kernel PCA

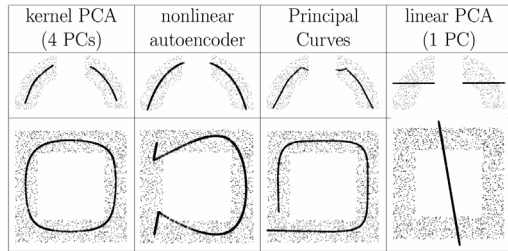
- The data is projected into a kernel matrix
- The kernel matrix is centered and the top k eigenvectors are obtained
- This gives the following nonlinear projections

$$z_1 = \sum_{i=1}^N b_{1i} K(\mathbf{x}_i, \mathbf{x}), z_2 = \sum_{i=1}^N b_{2i} K(\mathbf{x}_i, \mathbf{x}), \dots, z_k = \sum_{i=1}^N b_{ki} K(\mathbf{x}_i, \mathbf{x})$$

11



12



13

Principle Curves

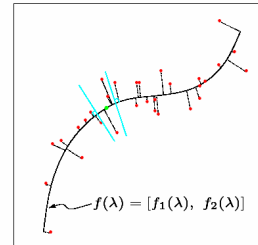


Figure 14.25: *The principal curve of a set of data. Each point on the curve is the average of all data points that project there.*

14

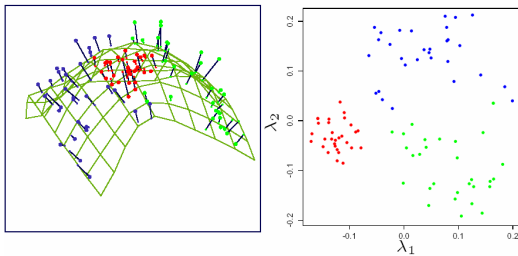


Figure 14.26: *Principal surface fit to half-sphere data. Left panel: fitted two-dimensional surface. Right panel: projections of data points onto the surface, resulting in coordinates $\hat{\lambda}_1, \hat{\lambda}_2$.*

Independent Component Analysis (ICA)

- Similar to PCA – find k projections

$$z_1 = \sum_{i=1}^d a_{1i} x_i, \quad z_2 = \sum_{i=1}^d a_{2i} x_i, \dots, \quad z_k = \sum_{i=1}^d a_{ki} x_i$$

- However, the independent components are now assumed to be statistically independent rather than uncorrelated
 - How this independence is defined is an open research questions (e.g. all moments have zero dependence)

16

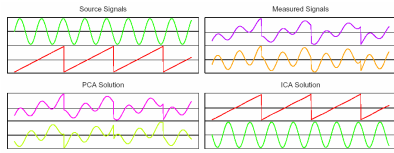


Figure 14.27: Illustration of ICA vs. PCA on artificial time-series data. The upper left panel shows the two source signals, measured at 1000 uniformly spaced time points. The upper right panel shows the observed mixed signals. The lower two panels show the principal components and independent component solutions.

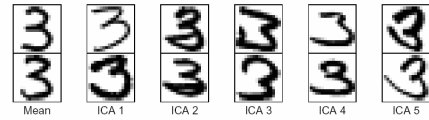


Figure 14.30: The highlighted digits from Figure 14.29. By comparing with the mean digits, we see the nature of the ICA component.

What if your data has outputs?

- Data $D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$
- Can build models in ICA, PCA or Principle Curve Space:

$$\hat{y} = \hat{f}(z_1, \dots, z_k)$$

- The model can be generated using any supervised learning algorithm
- However, the (z_1, \dots, z_k) may not be good predictors

Partial Least Squares (PLS)

- Uses the outputs to obtain the principle components.
- For the components $m = 1, \dots, k$ PLS maximizes

$$\max_{\substack{\|\alpha\|=1 \\ \varphi_j^T \alpha = 0, j=1, \dots, m-1}} \text{Corr}^2(\mathbf{y}, \mathbf{X}\alpha) \text{Var}(\mathbf{X}\alpha)$$

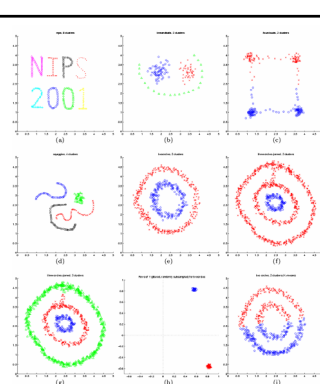
- Compare to PCA

$$\max_{\substack{\|\alpha\|=1 \\ \varphi_j^T \alpha = 0, j=1, \dots, m-1}} \text{Var}(\mathbf{X}\alpha)$$

Spectral Clustering

- Essentially K Means in the eigenvector space of the Kernel Matrix
 - Usually a Gaussian Kernel is used.

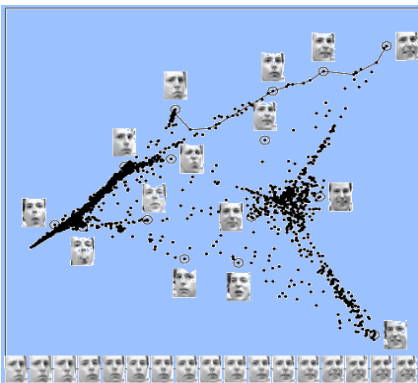
21



On spectral clustering: Analysis and an algorithm. A. Y. Ng, M. I. Jordan, and Y. Weiss. In T. Dietterich, S. Becker and Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems (NIPS) 14, 2002.

22

Locally Linear Embedding (LLE)



From:
Sam T. Roweis
roweis@cs.toronto.edu
www.cs.toronto.edu/~roweis/
and
Lawrence K. Saul
lksaul@research.ath.toronto.edu
www.research.ath.toronto.edu/~lksaul/

23