Component Analysis Methods for Computer Vision and Pattern Recognition

Fernando De la Torre

Component Analysis for CV & PR

- Computer Vision & Image Processing
  - Structure from motion.
  - Spectral graph methods for segmentation.
  - Appearance and shape models.
  - Fundamental matrix estimation and calibration.
  - Compression.
  - Classification.
  - Dimensionality reduction and visualization.
- Signal Processing
  - Spectral estimation, system identification (e.g. Kalman filter), sensor array processing (e.g. cocktail problem, eco cancellation), blind source separation, …
- Computer Graphics
  - Compression (BRDF), synthesis,...
- Speech, bioinformatics, combinatorial problems.

Component Analysis for CV & PR

- Computer Vision & Image Processing
  - Structure from motion.
  - Spectral graph methods for segmentation.
  - Appearance and shape models.
  - Fundamental matrix estimation and calibration.
  - Compression.
  - Classification.
  - Dimensionality reduction and visualization.
- Signal Processing
  - Spectral estimation, system identification (e.g. Kalman filter), sensor array processing (e.g. cocktail problem, eco cancellation), blind source separation, …
- Computer Graphics
  - Compression (BRDF), synthesis,...
- Speech, bioinformatics, combinatorial problems.
Component Analysis for CV & PR

Computer Vision & Image Processing
- Structure from motion.
- Spectral graph methods for segmentation.
- Appearance and shape models.
- Funda
- Comp
- Classi
- Dimer
- Signal F
  - Spectral estimation, sensor array processing (e.g. cocktail problem, eco cancellation), blind source separation.
- Computer Graphics
  - Compression (BRDF), synthesis, ...
- Speech

Component Analysis for CV & PR

Computer Vision & Image Processing
- Structure from motion.
- Spectral graph methods for segmentation.
- Appearance and shape models.
- Fundamental matrix estimation and calibration.
- Compression.
- Classification.
- Dimensionality reduction and visualization.
- Signal Processing
  - Spectral estimation, system identification (e.g. Kalman filter), sensor array processing (e.g. cocktail problem), eco cancellation, blind source separation, ...
- Computer Graphics
  - Compression (BRDF), synthesis,...
- Speech, bioinformatics, combinatorial problems.

Independent Component Analysis (ICA)
Component Analysis for CV & PR

Computer Vision & Image Processing
- Structure from motion.
- Spectral graph methods for segmentation.
- Appearance and shape models.
- Fundamental matrix estimation and calibration.
- Compression.
- Classification.
- Dimensionality reduction and visualization.

Signal Processing
- Spectral estimation, system identification (e.g. Kalman filter), sensor array processing (e.g. cocktail problem, eco cancellation), blind source separation, ...

Computer Graphics
- Compression (BRDF), synthesis, ...

Speech, bioinformatics, combinatorial problems.

Why CA for CV & PR?

- Learn from high dimensional data and few samples.
  - Useful for dimensionality reduction.

- Easy to incorporate
  - Robustness to noise, missing data, outliers (del la Torre & Black, 2003a)
  - Invariance to geometric transformations (Frey et al. 99, de la Torre & Black, 2003b, Cox et al. 2008)
  - Non-linearities (Kernel methods) (Scholkopf & Smola, 2002; Shawe-Taylor & Cristianini, 2004)
  - Probabilistic (latent variable models) (Everitt, 1984)
  - Multi-factorial (tensors) (Paatero & Tapper, 1994; O'Leary & Peleg, 1983; Vasilescu & Terzopoulos, 2002; Vasilescu & Terzopoulos, 2003)
  - Exponential family PCA (Gordon, 2002; Collins et al. 01)

- Efficient methods \( O(d < n^2) \)
  - Latent variable models.
  - Tensor factorization

Are CA Methods Popular/Useful/Used?

- About 28% of CVPR-07 papers use CA.

- Google:
  - Results 1-10 of about 1,870,000 for "principal component analysis".
  - Results 1-10 of about 506,000 for "independent component analysis".
  - Results 1-10 of about 273,000 for "linear discriminant analysis".
  - Results 1-10 of about 46,100 for "negative matrix factorization".
  - Results 1-10 of about 491,000 for "kernel methods".

- Still work to do
  - Results 1-10 of about 65,300,000 for "Britney Spears".

Outline

- Introduction
- Generative models
  - Principal Component Analysis (PCA) and extensions
  - K-means, spectral clustering and extensions
  - Non-negative Matrix Factorization (NMF)
  - Independent Component Analysis (ICA)
- Discriminative models
  - Linear Discriminant Analysis (LDA) and extensions
  - Oriented Component Analysis (OCA)
  - Canonical Correlation Analysis (CCA) and extensions
- A unifying view of CA
- Standard extensions of linear models
  - Latent variable models.
  - Tensor factorization
Generative Models

$$D \approx BC$$

- Principal Component Analysis/Singular Value Decomposition
  1) Robust PCA/SVD, PCA with uncertainty and missing data.
  2) Parameterized PCA
  3) Filtered Component Analysis
  4) Subspace regression
  5) Kernel PCA
- K-means and spectral clustering
- Non-Negative Matrix Factorization
- Independent Component Analysis.

Principal Component Analysis (PCA)

(Pearson, 1901; Hotelling, 1933; Mardia et al., 1979; Jolliffe, 1986; Diamantaras, 1996)

- PCA finds the directions of maximum variation of the data based on linear correlation.
- PCA decorrelates the original variables.

Snap-shot Method & SVD

- If $d \gg n$ (e.g. images 100*100 vs. 300 samples) no $DD^T$.
- $DD^T$ and $D^TD$ have the same eigenvalues (energy) and related eigenvectors (by $D$).
- $B$ is a linear combination of the data! (Sirovich, 1987)
  $$DD^T B = BA \quad B = Da \quad DD^T Da = DDDa = D^2 a \Lambda$$
- $[a, L] = \text{eig}(D^TD)$
  $$B = D a (\text{diag}(\text{diag}(L)))^{-0.5}$$
- SVD factorizes the data matrix $D$ as: $DD^T = UAU^T$
  (Beltrami, 1873; Schmidt, 1907; Golub & Loan, 1989)

$$D = BC$$
$$B \in \mathbb{R}^{d \times k} \quad C \in \mathbb{R}^{k \times k} \quad U \in \mathbb{R}^{k \times k} \quad \Sigma \in \mathbb{R}^{k \times k} \quad V \in \mathbb{R}^{n \times n}$$
$$B^T B = I \quad CC^T = \Lambda \quad U^T U = I \quad V^T V = I \quad \Sigma \text{ diagonal}$$

PCA

SVD
Error Function for PCA

- PCA minimizes the following function:
  \[ E(B, C) = \sum_{j=1}^{n} \| d_j - Bc_j \|^2 = \| D - BC \|^2 \]

- Not unique solution: \( R^{-1}C = BC \)
  \( R \in \mathbb{R}^{k \times d} \)
- To obtain same PCA solution \( R \) has to satisfy:
  \[ \hat{B} = BR \quad \hat{C} = R^{-1}C \]
  \[ \hat{B}^T\hat{B} = I \quad \hat{C}\hat{C}^T = \Lambda \]

- \( R \) is computed as a generalized \( k \times k \) eigenvalue problem.

\[ (CC^T)^{-1}R = B^TBR\Lambda^{-1} \quad \text{(de la Torre, 2006)} \]

### 1-Robust PCA

- Two types of outliers:
  - Sample outliers
  - Intra-sample outliers
  \( \text{(Xu & Yuille, 1995)} \)
  \( \text{(de la Torre & Black, 2001b; Skocija & Leonardis, 2003)} \)

- Standard PCA solution (noisy data):

### Robust PCA

- Using robust statistics:

\[ E_{\text{robust}}(B, C, \mu) = \sum_{i=1}^{n} \sum_{p=1}^{d} \rho(b_{ip} - \mu_p) \sum_{j=1}^{r} c_{jp}^T c_p \]

- Baseline (B) & Coefficients (C)
Numerical Problems

- No closed form solution in terms of an eigen-equation.
- Deflation approaches do not hold.

\[ A' = A - \lambda_1 u_1 u_1^T \]

First eigenvector with highest eigenvalue.

\[ A'' = A' - \lambda_2 u_2 u_2^T \]

Second eigenvector with highest eigenvalue.

- In the robust case all the basis have to be computed simultaneously (including the mean).

How to Optimize it?

\[ E_{\text{pca}}(B, C, \mu) = \sum_{i=1}^{n} \sum_{j=1}^{d} p(d_{ji} - \mu_p - \sum_{j=1}^{k} b_{pj} c_{ji}, \sigma_p) \]

- Normalized Gradient descent

\[ B^{n+1} = B^n - [H_b]^{-1} \frac{\partial E_{\text{pca}}}{\partial B} \quad H_b = \max \text{ diag } \left( \frac{\partial^2 E_{\text{pca}}}{\partial b \partial b^T} \right) \]

\[ C^{n+1} = C^n - [H_c]^{-1} \frac{\partial E_{\text{pca}}}{\partial C} \quad H_c = \max \text{ diag } \left( \frac{\partial^2 E_{\text{pca}}}{\partial c \partial c^T} \right) \]

- Deterministic annealing methods to avoid local minima.

(Blake & Zisserman, 1987)

Example

- Small region
- Short amount of time

Robust PCA

Original  PCA  RPCA  Outliers
Structure from Motion

More work on Robust PCA

• Robust estimation of coefficients (Black & Jepson, 1998; Ke & Kanade, 2000)
• Robust estimation of basis and coefficients (Gabriel & Odoro, 1984; Croux & Filzmoser, 1981; Skočaj et al., 2002; Skočaj & Leonardis, 2003; de la Torre & Black, 2001b; de la Torre & Black, 2003a)
• Other Robust PCA techniques (sample outliers) (Campbell, 1980; Ruymagaart, 1981; Xu & Yuille, 1995)

1- PCA with Uncertainty and Missing Data

• Adding uncertainty $E_1(B,C) = \sum_{i} w_i (d_i - B - \sum_k h_{ik})^2$

• If weights are separable $w = w_1 w_2 \ldots$ closed-form solution.

General Case

• For arbitrary weights no closed-form solution.

$E_1(B,C) = \sum_{i} w_i (d_i - B - \sum_k h_{ik})^2$

• Alternated least squares algorithms
  - Slow convergence, easy implementation.
  - Damped Newton Algorithm
  - Fast convergence, (Buchanan & Fitzgibbon, 2005)

$E_1(B,C) = \sum_{i} w_i (d_i - B - \sum_k h_{ik})^2$

$\nabla E_1 = \nabla F_1 + \lambda \left[ \lambda I + 2 \right]$

$H = \frac{\nabla^2 E_1}{\nabla v \nabla v^T}$

$g = \frac{\nabla E_1}{\nabla v}$

$\lambda = 10 \lambda$

repeat

$y = y - \alpha \cdot g$

until $F(y) < F(x)$

$x = y; \lambda = \lambda / 10$

until convergence

Related work

• Iterative (Wiberg, 1976; Shum et al., 1995; Morris & Kanade, 1998; Aans et al., 2002; Guerreiro & Aguiar, 2002)
• Closed-form (Aguiar & Moura, 1999; Irani & Anandan, 2000)
• Power factorization (Hartley & Schaffitzky, 2003)
• Bayesian estimation (L.Torresani & Bregler, 2004)
2- Parameterized Component Analysis (PaCA) (de la Torre & Black, 2003b)

- Learn a subspace invariant to geometric transformations?
- Data has to be geometrically normalized
  - Tedious manual cropping.
  - Inaccuracies due to matching ambiguities.
  - Hard to achieve sub-pixel accuracy.

Error function for PaCA

\[ E(B, C, a) = \sum_{i=1}^{T} d(f(x, a_i), Bc_i)^2 + p_2(a) + p_3(c) \]

Motion (warping) + Basis (B) & coefficients (c) + Regularization

\[ \sum_{t=1}^{T} \sum_{l=1}^{L} \lambda_t \| e_t - \Gamma c_{l-1} \|_w^2 + \lambda_2 \| a_t - \Gamma a_{t-1} \|_w^3 \]

EigenEye Learning

More examples

- UPS dataset.
  - Random selection of 100 images (16x16 pixels).
  - Incrementally update until preserve 80% of the energy.

Original  PaK-PCA  Congealing
Improving facial landmark labeling

• Hand label (red dots), PaK-PCA label (yellow)

More on Parameterized CA

• Probabilistic model
  – Search scales exponentially with the number of motion parameters (Frey & Jojic, 1999a; Frey & Jojic, 1999b; Williams & Titsias, 2004)

• Other continuous approaches.
  (Schewitzer, 1999; Rao, 1999; Shashua et al., 2002; Cox et al. 2008)

• Invariant clustering
  (Fitzgibbon & Zisserman, 2003)

• Non-rigid motion
  (Baker et al., 2004)

• Invariant recognition
  (Black & Jepson, 1998)

• Invariant support vector machines (Avidan, 2001)

• Parameterized Kernel Component Analysis (De la Torre, 2008)

Multi-band representation


• Filters (Gabor, Wavelets, Volterra, Fourier transform, …)

3- Filtered Component Analysis

(de la Torre et al., 2007b)

(a) Normalized correlation
   (ideal template)

1) No local minimum in the expected place.

2) Many local minima.
Multi-band representation

1) Global minimum in the expected place.

2) Distance between global and other minima is larger.

Filtered Component Analysis (FCA)

\[ E(F_{1}, \ldots, F_{r}) = \sum_{j=1}^{n} \sum_{i=1}^{2} (d_{ij} - \mu) \odot F_{ij} \|^{2}_{2} \]

Robustness of FCA

Training: 100 images

Correct global minimum

Gray  FCA (4)  Gabor(4)
41 %  74 %  62%

Correct to 2nd minimum distance

Average number of local minima

14.59  26  19.68
3.28  1.4  1.92

Other work

- Incremental PCA (de la Torre et al., 1998b; Ross et al., 2004; Brand, 2002; Skocaj & Leonardis, 2003; Champagne & Liu., 1998; A. Levy, 2000)

- Mixture of subspaces (Vidal et al., 2003; Leonardis et al., 2002)

- Changing the margin in SVM (Ashraf and Lucey 2010)

- Exponential family PCA (Collins et al. 01)
4- Subspace Regression: From a Single Image to a Subspace

- Traditional subspace methods
- Subspace Regression (Kim et al. 2010)

Subspace Regression (II)

- Generated samples for each pose
- Optimization problem

Experiment I

<table>
<thead>
<tr>
<th>Error Measure</th>
<th>Baseline I (img -&gt; img)</th>
<th>Baseline II (img -&gt; subsp)</th>
<th>Subspace Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matlab's subspace()</td>
<td>1.3507 (1.2312)</td>
<td>1.4088 (1.1645)</td>
<td>1.0860 (1.0651)</td>
</tr>
</tbody>
</table>
Experiment II

- Predicting a Subspace for Illumination
  - CMU PIE data set
  - 60 aligned subjects
  - 19 different illuminations

Subspace tracking

- (Template Matching: 42.99)
- IVT-SS: 38.41
- Subspace Regression: 37.98

5-Kernel PCA

- Suppose $\phi(.)$ is given as follows
  \[
  \phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)
  \]
- An inner product in the feature space is
  \[
  \langle \phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right), \phi\left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}\right) \rangle = (1 + x_1y_1 + x_2y_2)^2
  \]
- So, if we define the kernel function as follows, there is no need to carry out $\phi(.)$ explicitly
  \[
  K(x, y) = (1 + x_1y_1 + x_2y_2)^2
  \]
- This use of kernel function to avoid carrying out $\phi(.)$ explicitly is known as the kernel trick. In any linear algorithm that can be expressed by inner products can be made nonlinear by going to the feature space

Kernel Methods

- Computation in the feature space can be costly because it is (usually) high dimensional
  - The feature space is typically infinite-dimensional!
Kernel PCA
(Scholkopf et al., 1998)

Generative Models

\[ D \approx BC \]

- Principal Component Analysis/Singular Value Decomposition
  1) Robust PCA/SVD, PCA with uncertainty and missing data.
  2) Parameterized PCA
  3) Filtered Component Analysis
  4) Subspace regression
  5) Kernel PCA
- K-means and spectral clustering
  6) Aligned Cluster Analysis (ACA)
- Non-Negative Matrix Factorization
- Independent Component Analysis.

The Clustering Problem

- Partition the data set in c-disjoint "clusters" of data points.
- Number of possible partitions
  \[ S(n, c) = \frac{1}{c} \sum_{i=1}^{c} (-1)^i \binom{c}{i} \approx 10^{12} \]
- NP-hard and approximate algorithms (k-means, hierarchical clustering, mog, …)

K-means

\[ E_0(M, G) = \| D - MG^T \|_F \]

\[ D = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 & 10 \end{bmatrix} \]

\[ M = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 & 10 \end{bmatrix} \]

\[ G = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \]

\[ MG = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \]
Spectral Clustering

\[ E_0 (M, C) = \| \Gamma MC^T \|_F \]

\[ \Gamma = [\varphi(d_1), \varphi(d_2), \ldots, \varphi(d_n)] = \varphi(D) \]

Normalized Cuts (Shi & Malik '00)
Ratio-cuts (Hagen & Kahng '02)

6- Aligned Cluster Analysis (ACA)

- Mining facial expression

Problem

- Mining facial expression for one subject

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
</table>

- Summarization

Looking up  | Sleeping  | Waking up | Looking forward | Smiling
---|---|---|---|---

Problem

- Mining facial expression for one subject

- Summarization
Problem

• Mining facial expression of one subject

Summarization
• Embedding

Component Analysis for CV & PR    F. De la Torre
CVPR Easter School-2011

Problem

• Mining facial expression for one subject

Summarization
 Embedding
• Indexing

Component Analysis for CV & PR    F. De la Torre
CVPR Easter School-2011

k-means and kernel k-means
(MacQueen 67, Ding et al. 02, Dhillon et al. 04, Zass and Shashua 05, De la Torre 06)

\[ J(M, G) = \| \phi(X) - MG \|_F^2 \]

\[ G = \{ s_1, s_2, \ldots, s_k \} \]

\[ M = \{ x_1, x_2, \ldots, x_n \} \]

\[ K = \phi(x)^T \phi(x) \]

Component Analysis for CV & PR    F. De la Torre
CVPR Easter School-2011

Problem formulation for ACA (I)

\[ J_{aca}(M, G, S) = \| \phi(x_{s_1-1}) - \phi(x_{s_1-2}) - \phi(x_{s_1-3}) - \phi(x_{s_1-4}) \|_F^2 \]

\[ 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10 \]

Labels (G)

Start and end of the segments (s)

Component Analysis for CV & PR    F. De la Torre
CVPR Easter School-2011
Problem formulation for ACA (II)

\[ J_{\text{aca}}(M,G,S) = \| \varphi(X_{i_{1},i_{2}},X_{i_{2},i_{3}}, \ldots, X_{i_{n-1},i_{n}}) - MG \|_F^2 \]

\[ = \sum_{i=1}^{n} \sum_{c=1}^{m} g_{ci} \| \varphi(X_{i_S_i, S_{i}}, m_c) - m_c \|_2^2 \]

Dynamic Time Alignment Kernel (Shimodaira et al. 01)

\[ X_{i_{S_i}, S_{i}}, m_c \]

Optimizing ACA (forward step)

- Efficient Dynamic Programming

\[ \begin{array}{c|c|c|c|}
  \nu & \nu & \nu & \\
  29 & 29 & 29 & \\
  \end{array} \]

\[ w_{\text{max}} \]

Optimizing ACA (backward step)

\[ J_{i_{\text{im}}} = \text{tr}(KL) \quad \text{with} \quad L = I_n - G^T (GG^T)^{-1}G \]

\[ J_{\text{aca}} = \text{tr}(K (L \circ W)) \quad \text{with} \quad L = I_n - HG^T (GG^T)^{-1}G \]

\[ G \in [0,1]^n \]

Optimizing ACA (forward step)

\[ v \quad \nu \quad w \quad \nu \quad w \]

\[ 29 \quad 29 \quad 29 \quad \\
  \]

Optimizing ACA (backward step)

\[ O(n^2 w_{\text{max}}) \]

\[ \begin{array}{c|c|c|c|}
  \nu & \nu & \nu & \\
  29 & 29 & 29 & \\
  \end{array} \]

Component Analysis for CV & PR

F. De la Torre

CVPR Easter School-2011
Honey bee dance data (Oh et al. 08)

Three behaviors:
1-waggle, 2-left turn, 3-right turn

<table>
<thead>
<tr>
<th></th>
<th>Seq 1</th>
<th>Seq 2</th>
<th>Seq 3</th>
<th>Seq 4</th>
<th>Seq 5</th>
<th>Seq 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACA</td>
<td>0.845</td>
<td>0.925</td>
<td>0.600</td>
<td>0.922</td>
<td>0.878</td>
<td>0.928</td>
</tr>
<tr>
<td>PS-SLDS (Oh et al 08)</td>
<td>0.759</td>
<td>0.924</td>
<td>0.831</td>
<td>0.934</td>
<td>0.904</td>
<td>0.910</td>
</tr>
<tr>
<td>HDP-VAR(1)-HMM (Fox et al 08)</td>
<td>0.465</td>
<td>0.441</td>
<td>0.456</td>
<td>0.832</td>
<td>0.932</td>
<td>0.887</td>
</tr>
<tr>
<td>Spectral Clustering</td>
<td>0.698</td>
<td>0.631</td>
<td>0.509</td>
<td>0.671</td>
<td>0.577</td>
<td>0.649</td>
</tr>
</tbody>
</table>

Facial image features

- Active Appearance Models (Baker and Matthews '04)
- Image features
  - Upper face
  - Lower face

Facial event discovery across subjects

- Cohn-Kanade: 30 people and five different expressions (surprise, joy, sadness, fear, anger)

Facial event discovery across subjects

- Cohn-Kanade: 30 people and five different expressions (surprise, joy, sadness, fear, anger)

- 10 sets of 30 people

<table>
<thead>
<tr>
<th></th>
<th>ACA</th>
<th>Spectral Clustering (SC)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.87 (0.05)</td>
<td>0.56 (0.04)</td>
</tr>
</tbody>
</table>

Facial image features

- Active Appearance Models (Baker and Matthews '04)
- Image features
  - Upper face
  - Lower face

Facial event discovery across subjects

- Cohn-Kanade: 30 people and five different expressions (surprise, joy, sadness, fear, anger)

Facial event discovery across subjects

- Cohn-Kanade: 30 people and five different expressions (surprise, joy, sadness, fear, anger)

- 10 sets of 30 people

<table>
<thead>
<tr>
<th></th>
<th>ACA</th>
<th>Spectral Clustering (SC)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.87 (0.05)</td>
<td>0.56 (0.04)</td>
</tr>
</tbody>
</table>
Unsupervised facial event discovery

Clustering human motion

clustering of human motion II

Generative Models

\[ D \approx BC \]

- Principal Component Analysis/Singular Value Decomposition
  1) Robust PCA/SVD, PCA with uncertainty and missing data.
  2) Parameterized PCA
  3) Filtered Component Analysis
  4) Kernel PCA
- K-means and spectral clustering
  5) Aligned Cluster Analysis (ACA)
- Non-Negative Matrix Factorization
- Independent Component Analysis.
“Intercorrelations among variables are the bane of the multivariate researcher’s struggle for meaning”

Cooley and Lohnes, 1971

Part-based Representation

- The firing rates of neurons are never negative.
- Independent representations.

NMF & ICA

Non-negative Matrix Factorization

- Positive factorization.
  \[ E(\mathbf{B}, \mathbf{C}) = \| \mathbf{D} - \mathbf{BC} \|_F \quad \mathbf{B}, \mathbf{C} \geq 0 \]
- Leads to part-based representation.

Nonnegative Factorization

(Lee & Seung, 1999; Lee & Seung, 2000)

\[
\min_{\mathbf{B} \geq 0, \mathbf{C} \geq 0} F = \sum_y |d_y - (\mathbf{BC})_y|^2
\]

Inference:

\[
C_y \leftarrow C_y \frac{(\mathbf{B}^T \mathbf{D})_y}{(\mathbf{B}^T \mathbf{B} \mathbf{V})_y}
\]

Derivatives:

\[
\frac{\partial F}{\partial C_y} = (\mathbf{B}^T \mathbf{B} \mathbf{C})_y - (\mathbf{B}^T \mathbf{C})_y
\]

Learning:

\[
B_y \leftarrow B_y \frac{(\mathbf{D}^T C^T)_y}{(\mathbf{B}^T \mathbf{C}^T)_y}
\]

- Multiplicative algorithm can be interpreted as diagonally rescaled gradient descent.
Generative Models

\[ D \approx BC \]

- Principal Component Analysis/Singular Value Decomposition
  1) Robust PCA/SVD, PCA with uncertainty and missing data.
  2) Parameterized PCA
  3) Filtered Component Analysis
  4) Kernel PCA
- K-means and spectral clustering
  5) Aligned Cluster Analysis (ACA)
- Non-Negative Matrix Factorization
- Independent Component Analysis.

ICA

(Hyvärinen et al., 2001)

\[ D = BC \quad C \approx S = WD \quad W \approx B^{-1} \]

- Look for \( s_i \) that are independent.
- PCA finds uncorrelated variables, the independent components have non Gaussian distributions.
- Uncorrelated \( E(s_is_j) = E(s_i)E(s_j) \)
- Independent \( E(g(s_i)f(s_j)) = E(g(s_i))E(f(s_j)) \) for any non-linear \( f, g \)

ICA vs PCA
Many optimization criteria

- Minimize high order moments: e.g. kurtosis
  \[ \text{kurt}(W) = E(s^4) - 3(E(s^2))^2 \]
- Many other information criteria.
- Also an error function: (Olhausen & Field, 1996)
  \[ \sum_{i=1}^{n} |d_i - Bc_i| + \sum_{i=1}^{n} S(c_i) \text{ Sparseness (e.g. } S=|\cdot|) \]
- Other sparse PCA.
  (Chennubhotla & Jepson, 2001b; Zou et al., 2005; dAspremont et al., 2004;)

Outline

- Introduction
- Generative models
  - Principal Component Analysis (PCA) and extensions
  - K-means, spectral clustering and extensions
  - Non-negative Matrix Factorization (NMF)
  - Independent Component Analysis (ICA)
- Discriminative models
  - Linear Discriminant Analysis (LDA) and extensions
  - Oriented Component Analysis (OCA)
  - Canonical Correlation Analysis (CCA) and extensions
- A unifying view of CA
- Standard extensions of linear models
  - Latent extensions models.
  - Tensor factorization

Basis of natural images
Discriminative Models

- Linear Discriminant Analysis (LDA)
  7) Discriminative Cluster Analysis
- Multimodal Oriented Discriminant Analysis
- Oriented Component Analysis (OCA)
- Canonical Correlation Analysis (CCA)
- Dynamical Coupled Component Analysis
  9) Dynamical Coupled Component Analysis
  10) Canonical Time Warping

Linear Discriminant Analysis (LDA)
(Fisher, 1936; Mardia et al., 1979; Bishop, 1995)

\[
S_b = \sum_{i=1}^{C} \sum_{j=1}^{C} (\mu_i - \mu_j)(\mu_i - \mu_j)^T
\]

\[
S_w = \sum_{i=1}^{C} \sum_{j=1}^{C} (d_i - \mu_j)(d_i - \mu_j)^T
\]

- Optimal linear dimensionality reduction if classes are Gaussian with equal covariance matrix.

Error function for LDA
(de la Torre & Kanade, 2006)

\[
E_{LDA}(A, B) = \frac{1}{2} \left( G^T - B \Lambda^T D \right)^2
\]

Equations n x c
Unknowns d x c

- d>>n an UNDETERMINED system of equations! (over-fitting)

7-Discriminative Cluster Analysis (DCA)
(de la Torre & Kanade, 2006)

\[
E_{DCA}(A, B, G) = \frac{1}{2} (G^T - BA^T D)^2
\]

PCA+k-means
DCA
8- Multimodal Oriented Component Analysis (MODA)

(de la Torre & Kanade, 2005a)

- How to extend LDA to deal with:
  - Model class covariances.
  - Multimodal classes.
  - Deal efficiently with huge covariance matrices (e.g. 100*100).
MODA

\[ \text{MAXIMIZES the Kullback-Leibler divergence between clusters among classes.} \]

\[ \sum_{i=1}^{C} \sum_{j=1}^{r} \sum_{k,l \in C_{r}} \text{tr}(B^T \Sigma_i^{-1} + \Sigma_j^{-1} + (\mu_{ik} - \mu_{lj})^T (\Sigma_i^{-1} + \Sigma_j^{-1}) (\mu_{ik} - \mu_{lj})^T B) \]

\[ \text{1 mode per class and equal covariances equivalent to LDA.} \]

Optimization

\[ J(B) = - \sum_{i=1}^{C} \text{tr}((B^T \Sigma_i B)^{-1} (B^T A, B)) \]

\[ \text{Iterative Majorization (Kiers, 1995; Leeuw, 1994)} \]

Related LDA work

- Face recognition (Belhumeur et al., 1997; Zhao, 2000; Martinez & Kak, 2003)
- Small sample problem (Chen et al., 2000; Yu & Yang, 2001)
- Mixture (Hastie et al., 1995; Zhu & Martinez, 2006)
- Neural approaches (Gallinari et al., 1991; Lowe & Webb, 1991)
- Heteroscedastic discriminant analysis (Kumar & Andreou, 1998; Fukunaga, 1990; Martinet al., 1979; Saon et al., 2000)

Discriminative Models

- Linear Discriminant Analysis (LDA)
  7) Discriminative Cluster Analysis
- Multimodal Oriented Discriminant Analysis
- Oriented Component Analysis (OCA)
- Canonical Correlation Analysis (CCA)
  8) Dynamical Coupled Component Analysis
  9) Canonical Time Warping
Oriented Component Analysis (OCA)

\[ \mathbf{b}^T_{OCA} \Sigma_{signal} \mathbf{b}_{OCA} \]

- Generalized eigenvalue problem: \( \Sigma_b \mathbf{b}_k = \Sigma_e \mathbf{b}_k \lambda \)
- \( \mathbf{b}_{OCA} \) is steered by the distribution of noise

Representational Oriented Component Analysis (ROCA)

\[ \mathbf{b}^T_k \Sigma_b \mathbf{b}_k \]

Discriminative Models

- Linear Discriminant Analysis (LDA)
- Discriminative Cluster Analysis
- Multimodal Oriented Discriminant Analysis
- Oriented Component Analysis (OCA)
- Canonical Correlation Analysis (CCA)
- Dynamical Coupled Component Analysis
- Canonical Time Warping

Canonical Correlation Analysis (CCA)

(Mardia et al., 1979; Borga 98)

- Learn relations between multiple data sets? (e.g. find features in one set related to another data set)
- Given two sets \( \mathbf{X} \in \mathbb{R}^{n \times d_x} \) and \( \mathbf{Y} \in \mathbb{R}^{n \times d_y} \), CCA finds the pair of directions \( \mathbf{w}_x \) and \( \mathbf{w}_y \) that maximize the correlation between the projections (assume zero mean data)

\[ \rho = \frac{\mathbf{w}_x^T \mathbf{X} \mathbf{Y} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \mathbf{X} \mathbf{X} \mathbf{w}_x \mathbf{w}_y^T \mathbf{Y} \mathbf{Y} \mathbf{w}_y}} \]

- Several ways of optimizing it:

\[ \mathbf{A} = \begin{bmatrix} 0 & \mathbf{X}^T \mathbf{Y} \\ \mathbf{X}^T \mathbf{Y} & 0 \end{bmatrix} \in \mathbb{R}^{(d_x+d_y) \times (d_x+d_y)}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{X}^T \mathbf{X} & 0 \\ 0 & \mathbf{Y}^T \mathbf{Y} \end{bmatrix} \in \mathbb{R}^{(d_x+d_y) \times (d_x+d_y)}, \quad \mathbf{w} = \begin{bmatrix} \mathbf{w}_x^T \\ \mathbf{w}_y^T \end{bmatrix} \]

- An stationary point of \( \mathbf{r} \) is the solution to CCA.

\[ \mathbf{A} \mathbf{w} = \lambda \mathbf{B} \mathbf{w} \]
9- Dynamic Coupled Component Analysis (DCCA) (de la Torre & Black, 2001a)

- Learning the coupling.
- High dimensional data.
- Limited training data.

Solutions?

- PCA independently and general mapping
- Signals dependent signals with small energy can be lost.

DCCA

\[ E_{cda}(W, \hat{B}, A, C, \mu, \hat{\mu}) = \sum_{i=1}^{n} \left( d_i - \hat{d}_i \right)^2 + \lambda_1 \sum_{i=1}^{n} \left( c_i - \hat{c}_i \right)^2 + \lambda_2 \sum_{j=1}^{m} \left( B_j - \hat{B}_j \right)^2 \]

Projection

Reconstruction

Dynamic Coupled Component Analysis
10- Canonical Time Warping (CTW)

Canonical Correlation Analysis (CCA)

(Hotelling 1936)

- CCA minimizes:

$$J_{\text{cca}}(V_x, V_y) = \| V_x^T X - V_y^T Y \|^2$$

subject to:

$$V_x^T X = V_y^T Y$$

Different #rows, same #columns

$$X \in \mathbb{R}^{d \times n_x}, Y \in \mathbb{R}^{d \times n_y}$$

Different #rows, different #columns

$$X \in \mathbb{R}^{d_x \times n_x}, Y \in \mathbb{R}^{d_y \times n_y}$$

Spatial transformation

Canonical Time Warping (CTW)

Reminder

$$J_{\text{ctw}}(W_x, W_y) = \| W_x^T X - W_y^T Y \|^2$$

Different #rows, different #columns

$$X \in \mathbb{R}^{d_x \times n_x}, Y \in \mathbb{R}^{d_y \times n_y}$$

Spatial transformation

$$J_{\text{ctw}}(W_x, W_y, V_x, V_y) = \| V_x^T W_x^T X - V_y^T Y W_y^T Y \|^2$$

Subject to:

$$V_x^T W_x^T X = V_y^T Y W_y^T Y$$

Temporal alignment

$$s.t. V_x^T W_x^T X V_x = I_{d_x}$$

$$V_y^T Y W_y^T Y V_y = I_{d_y}$$

A least-square formulation for DTW

Spatial transformation

Different #rows, different #columns
Facial expression alignment

Aligning human motion

Boxing
Opening a cabinet

Aligning motion capture and video
Outline

• Introduction
• Generative models
  – Principal Component Analysis (PCA) and extensions
  – K-means, spectral clustering and extensions
  – Non-negative Matrix Factorization (NMF)
  – Independent Component Analysis (ICA)
• Discriminative models
  – Linear Discriminant Analysis (LDA) and extensions
  – Oriented Component Analysis (OCA)
  – Canonical Correlation Analysis (CCA) and extensions
• A unifying view of CA
• Standard extensions of linear models
  – Latent variable models.
  – Tensor factorization

The fundamental equation of CA

Given two datasets \( D \in \mathbb{R}^{d \times n} \) and \( X \in \mathbb{R}^{x \times n} \)

\[
E_0(A, B) = \| C \|_F^2
\]

Regression matrices

Weights for rows

Weights for columns

Properties of the cost function

• \( E_0(A, B) \) has a unique global minimum (Baldi and Hornik-89).

• Closed form solutions for \( A \) and \( B \) are:

\[
E_0(A) = tr\left( (A^T A)^{-1} (A^T A) \right)
\]

\[
E_0(B) = tr\left( (B^T B)^{-1} (B^T B) \right)
\]

Principal Component Analysis (PCA)

(Pearson, 1901; Hotelling, 1933; Mardia et al., 1979; Jolliffe, 1986; Diamantaras, 1996)

• PCA finds the directions of maximum variation of the data based on linear correlation.

• Kernel PCA finds the directions of maximum variation of the data in the feature space.

\[
(x_1, x_2) \rightarrow (x_1, x_2, x_1^2 + x_2^2) = (z_1, z_2, z_3)
\]
**PCA-Kernel PCA**

- Error function for KPCA:  
  \[ E_0(A, B) = || \Phi_r (\Gamma - BA^T \Phi) \Phi ||_F \]
  \[ E_{kpcA}(A, B) = || (D - BA^T) \Phi \Phi ||_F \]

- The primal problem:
  \[ E_{kpcA}(B) = \text{tr}\left((B^T B)^{-1} (B^T D D^T B)\right) \]

- The dual problem:
  \[ E_{kpcA}(A) = \text{tr}\left((A^T A)^{-1} (A^T \Phi(D)^T \Phi(D)A)\right) \]

---

**Linear Discriminant Analysis (LDA)**

(Fisher, 1936; Mardia et al., 1979; Bishop, 1995)

- Objective function:
  \[ J(B) = \text{tr}\left((B^T S_b^{-1} B - B^T S_i^{-1} B)\right) \]

- Optimal linear dimensionality reduction if classes are Gaussian with equal covariance matrix.

---

**Canonical Correlation Analysis (CCA)**

(Fisher, 1936; Mardia et al., 1979)

- Given two sets \( X \in \mathbb{R}^{n \times s} \) and \( D \in \mathbb{R}^{c \times s} \), CCA finds the pair of directions \( w_x \) and \( w_y \) that maximize the correlation between the projections (assume zero mean data)

\[
\rho = \frac{w_x^T X^T D w_d}{\sqrt{w_x^T X^T X w_x^T D^T D w_d}}
\]

---

**Canonical Correlation Analysis (CCA)**

- Objective function:
  \[ E_0(A, B) = || W_r (\Gamma - BA^T \Psi') \Phi ||_F \]

- Objective function:
  \[ E_{CCA}(A, B) = || (D^T D)^{-1/2} (D - BA^T X) ||_F \]

- CCA is the same as LDA changing the label matrix by a new set \( X \)
K-means

\[ E_0(A, B) = \| (D - BA^T) \|_F \]

\[ D = x \]
\[ B = y \]
\[ A = BA^T = x \]

(Ding et al., ’02, Torre et al ’06)

Normalized cuts

\[ E_0(A, B) = \| (D - BA^T) \|_F \]

\[ \Gamma = [\varphi(d_1), \varphi(d_2), \ldots, \varphi(d_d)] \]

(Dhillon et al., ’04, Zass & Shashua, 2005; Ding et al., 2005, De la Torre et al ’06)

Other Connections

• The LS-KRRR (\( E_0 \)) is also the generative model for:
  – Laplacian Eigenmaps, Locality Preserving projections, MDS, Partial least-squares, …

• Benefits of LS framework:
  – Common framework to understand difference and communalities between different CA methods (e.g. KPCA, KLDA, KCCA, Ncuts)
  – Better understanding of normalization factors and generalizations
  – Efficient numerical optimization less than \( \Theta(n^3) \) or \( \Theta(d^3) \), where \( n \) is number of samples and \( d \) dimensions

Outline

• Introduction
• Generative models
  – Principal Component Analysis (PCA).
  – Non-negative Matrix Factorization (NMF).
  – Independent Component Analysis (ICA).
• Discriminative models
  – Linear Discriminant Analysis (LDA).
  – Oriented Component Analysis (OCA).
  – Canonical Correlation Analysis (CCA).
• A unifying view of CA
  – Standard extensions of linear models
    – Latent variable models.
    – Tensor factorization
Factor Analysis

- A Gaussian distribution on the coefficients and noise is added to PCA: Factor Analysis. (Mardia et al., 1979)
  \[ d = \mu + Bc + \eta \]
  \[ p(c) = N(c \mid 0, I) \]
  \[ p(d \mid c, B) = N(d \mid \mu + Bc, \Psi) \]
  \[ p(\eta) = N(\eta \mid 0, \Psi) \]
  \[ \Psi = \text{diag}(\eta_1, \eta_2, \ldots, \eta_p) \]
  \[ \text{cov}(d) = E((d - \mu)(d - \mu)^T) = BB^T + \Psi \]

- Inference (Roweis & Ghahramani, 1999; Tipping & Bishop, 1999a)
  \[ p(c, d) \]
  \[ p(c \mid d) = N(c \mid m, V) \]
  \[ m = B^T (BB^T + \Psi)^{-1} (d - \mu) \]
  \[ V = (I + B^T \Psi^{-1} B)^{-1} \]

PCA reconstruction low error.
FA high reconstruction error (low likelihood).

More on PPCA

- Extension to mixtures of PPCA (mixture of subspaces).
  (Tipping & Bishop, 1999b; Black et al., 1998; Jebara et al., 1998)
- Tracking (Yang et al., 1999; Yang et al., 2000a; Lee et al., 2005; de la Torre et al., 2000b)
- Recognition/Detection (Moghaddam et al., 2000; Shakhnarovich & Moghaddam, 2004; Everingham & Zisserman, 2006)
- PCA for the exponential family (Collins et al., 2001)

Pppca

- If \( \Psi = E(\eta \eta^T) = \eta \eta^T \) PPCA.
- If \( \epsilon \rightarrow 0 \) is equivalent to PCA: \( \epsilon \rightarrow 0 \) \( B^T (BB^T + \Psi)^{-1} = (B^T B)^{-1}B^T \)
- Probabilistic visual learning (Moghaddam & Pentland, 1997;)

\[
p(d) = \int p(d \mid c)p(c)dc = \int \frac{e^{-\frac{1}{2}d^T(d - \mu)^T \Psi^{-1}(d - \mu)}}{(2\pi)^{p/2} |\Psi|^{1/2}} = \frac{1}{(2\pi)^{p/2} |\Psi|^{1/2}} \left( \frac{1}{\det(\Psi)} \prod_{i=1}^{p} \left| e^{\frac{1}{2}(d_i^T \mu_i - \frac{1}{2}d_i^T \mu_i)} \right| \right)
\]

\[ c_i = B^T d_i \]
Tensor Factorization

Tensor faces
(Vasilescu & Terzopoulos, 2002; Vasilescu & Terzopoulos, 2003)

Tensor faces

Views

Illuminations

Expressions

Data Organization

Eigenfaces

• Facial images (identity change)

• Eigenfaces bases vectors capture the variability in facial appearance (do not decouple pose, illumination, …)

Linear/PCA: Data Matrix

- \( R \) pixels x images
- a matrix of image vectors

Multilinear: Data Tensor \( D \)

- \( R \) people x views x ilums x express x pixels
- N-dimensional matrix
- 28 people, 45 images/person
- 5 views, 3 illuminations,
  3 expressions per person

Images

People

Expressions

Illuminations
N-Mode SVD Algorithm

\[ D = Z \times U \]

- \( N = 3 \)
- pixels x views x people

Strategic Data Compression = Perceptual Quality

- TensorFaces data reduction in illumination space primarily degrades illumination effects (cast shadows, highlights)
- PCA has lower mean square error but higher perceptual error

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean Sq. Err.</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>TensorFaces</td>
<td>409.15</td>
<td>6 illum + 11 people param.</td>
</tr>
<tr>
<td>TensorFaces</td>
<td>85.75</td>
<td>3 illum + 11 people param.</td>
</tr>
<tr>
<td>PCA</td>
<td></td>
<td>33 parameters</td>
</tr>
</tbody>
</table>

Acknowledgments

- The content of some of the slides has been taken from previous presentations/papers of:
  - Ales Leonardis
  - Horst Bischof
  - Michael Black
  - Rene Vidal
  - Anat Levin
  - Alex Martinez
  - Juha Karhunen
  - Andrew Fitzgibbon
  - Daniel Lee
  - Chris Ding
  - M. Alex Vasilescu
  - Sam Roweis
  - Daoqiang Zhang
  - Ammon Shashua
Bibliography

BibNet

Bibliography

BibNet

Bibliography

BibNet