Lots of high-dimensional noisy data...

Zambian President Levy Mwanawasa has won a second term in office in an election his challenger Michael Sata accused him of rigging, official results showed on Monday.

According to media reports, a pair of hackers said on Saturday that the Firefox Web browser, commonly perceived as the safer and more customizable alternative to market leader Internet Explorer, is critically flawed. A presentation on the flaw was shown during the ToorCon hacker conference in San Diego.
Lots of high-dimensional noisy data...

Zambian President Levy Mwanawasa has won a second term in office in an election his challenger Michael Sata accused him of rigging, official results showed on Monday.

According to media reports, a pair of hackers said on Saturday that the Firefox Web browser, commonly perceived as the safer and more customizable alternative to market leader Internet Explorer, is critically flawed. A presentation on the flaw was shown during the ToorCon hacker conference in San Diego.
Lots of high-dimensional noisy data...

Zambian President Levy Mwanawasa has won a second term in office in an election his challenger Michael Sata accused him of rigging, official results showed on Monday.

According to media reports, a pair of hackers said on Saturday that the Firefox Web browser, commonly perceived as the safer and more customizable alternative to market leader Internet Explorer, is critically flawed. A presentation on the flaw was shown during the ToorCon hacker conference in San Diego.
Lots of high-dimensional noisy data...

Zambian President Levy Mwanawasa has won a second term in office in an election his challenger Michael Sata accused him of rigging, official results showed on Monday.

According to media reports, a pair of hackers said on Saturday that the Firefox Web browser, commonly perceived as the safer and more customizable alternative to market leader Internet Explorer, is critically flawed. A presentation on the flaw was shown during the ToorCon hacker conference in San Diego.
Lots of high-dimensional noisy data...

Zambian President Levy Mwanawasa has won a second term in office in an election his challenger Michael Sata accused him of rigging, official results showed on Monday. According to media reports, a pair of hackers said on Saturday that the Firefox Web browser, commonly perceived as the safer and more customizable alternative to market leader Internet Explorer, is critically flawed. A presentation on the flaw was shown during the ToorCon hacker conference in San Diego.
Lots of high-dimensional noisy data...

Zambian President Levy Mwanawasa has won a second term in office in an election his challenger Michael Sata accused him of rigging, official results showed on Monday.

According to media reports, a pair of hackers said on Saturday that the Firefox Web browser, commonly perceived as the safer and more customizable alternative to market leader Internet Explorer, is critically flawed. A presentation on the flaw was shown during the ToorCon hacker conference in San Diego.

Goal: find a useful representation of data
Basic idea of linear dimensionality reduction

Represent each face as a high-dimensional vector $x \in \mathbb{R}^{361}$
Basic idea of linear dimensionality reduction

Represent each face as a high-dimensional vector \( \mathbf{x} \in \mathbb{R}^{361} \)

\[
\mathbf{x} \in \mathbb{R}^{361} \\
\mathbf{z} = \mathbf{U}^T \mathbf{x} \\
\mathbf{z} \in \mathbb{R}^{10}
\]
Basic idea of linear dimensionality reduction

Represent each face as a high-dimensional vector $\mathbf{x} \in \mathbb{R}^{361}$

This setup is the same for all methods we will talk about today; the criteria for choosing $\mathbf{U}$ determines the particular algorithm
Motivation and context

Why do dimensionality reduction? \[ Z = U^T X \]
Motivation and context

Why do dimensionality reduction? $Z = U^T X$

- Scientific: understand structure of data (visualization)
Motivation and context

Why do dimensionality reduction? $Z = U^TX$

- Scientific: understand structure of data (visualization)
- Statistical: fewer dimensions allows better generalization
Motivation and context

Why do dimensionality reduction? $Z = U^T X$

- Scientific: understand structure of data (visualization)
- Statistical: fewer dimensions allows better generalization
- Computational: compress data for efficiency (both time/space)
Motivation and context

Why do dimensionality reduction? \( Z = U^T X \)

- Scientific: understand structure of data (visualization)
- Statistical: fewer dimensions allows better generalization
- Computational: compress data for efficiency (both time/space)
- Direct: use as a model for anomaly detection
Motivation and context

Why do dimensionality reduction? \[ Z = U^T X \]

- Scientific: understand structure of data (visualization)
- Statistical: fewer dimensions allows better generalization
- Computational: compress data for efficiency (both time/space)
- Direct: use as a model for anomaly detection

In the context of this class...
Motivation and context

Why do dimensionality reduction? $Z = U^TX$

- Scientific: understand structure of data (visualization)
- Statistical: fewer dimensions allows better generalization
- Computational: compress data for efficiency (both time/space)
- Direct: use as a model for anomaly detection

In the context of this class...

- Feature selection (three weeks ago)
Motivation and context

Why do dimensionality reduction? \( Z = U^T X \)

- Scientific: understand structure of data (visualization)
- Statistical: fewer dimensions allows better generalization
- Computational: compress data for efficiency (both time/space)
- Direct: use as a model for anomaly detection

In the context of this class...

- Feature selection (three weeks ago)
- Clustering (last week)
Motivation and context

Why do dimensionality reduction? $Z = U^T X$

- Scientific: understand structure of data (visualization)
- Statistical: fewer dimensions allows better generalization
- Computational: compress data for efficiency (both time/space)
- Direct: use as a model for anomaly detection

In the context of this class...

- Feature selection (three weeks ago)
- Clustering (last week)
- Nonlinear dimensionality reduction (in 4 weeks)
Motivation and context

Why do dimensionality reduction?

\[ Z = U^T X \]

- Scientific: understand structure of data (visualization)
- Statistical: fewer dimensions allows better generalization
- Computational: compress data for efficiency (both time/space)
- Direct: use as a model for anomaly detection

In the context of this class...

- Feature selection (three weeks ago)
- Clustering (last week)
- Nonlinear dimensionality reduction (in 4 weeks)

These are mostly unsupervised methods: use only \( X \)

Contrast with supervised methods

(classification, regression), where \((X, Y)\) are given
Outline

• Introduction

• Methods
  – Principal component analysis (PCA)
  – Canonical correlation analysis (CCA)
  – Linear discriminant analysis (LDA)
  – Non-negative matrix factorization (NMF)
  – Independent component analysis (ICA)

• Case studies
  – Network anomaly detection
  – Multi-task learning
  – Part-of-speech tagging
  – Brain imaging

• Extensions, related methods, summary
Outline

- Introduction
- Methods
  - Principal component analysis (PCA)
  - Canonical correlation analysis (CCA)
  - Linear discriminant analysis (LDA)
  - Non-negative matrix factorization (NMF)
  - Independent component analysis (ICA)
- Case studies
  - Network anomaly detection
  - Multi-task learning
  - Part-of-speech tagging
  - Brain imaging
- Extensions, related methods, summary
PCA: first principal component
PCA: first principal component

Objective: maximize variance of projected data
PCA: first principal component

Objective: maximize variance of projected data

\[
= \max_{||u||=1} \sum_{i=1}^{n} \left( u^T x_i \right)^2
\]

length of projection
PCA: first principal component

\[
X = \begin{pmatrix}
    x_1 & \cdots & x_n \\
\end{pmatrix}
\]

(assume data is centered at 0)

Objective: maximize variance of projected data

\[
= \max_{||u||=1} \sum_{i=1}^{n} (u^T x_i)^2
\]

\[
= \max_{||u||=1} ||u^T X||^2
\]
PCA: first principal component

Objective: maximize variance of projected data

\[
X = (x_1 \ldots x_n)
\]

(assume data is centered at 0)

\[
\sum_{i=1}^{n} (u^T x_i)^2\]

length of projection

\[
= \max_{||u||=1} \sum_{i=1}^{n} (u^T x_i)^2
\]

\[
= \max_{||u||=1} ||u^T X||^2
\]

largest eigenvalue of \(XX^T\)

(covariance matrix)
PCA: first principal component

Objective: maximize variance of projected data

\[ X = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix} \]

(assume data is centered at 0)

\[ \max_{||u||=1} \sum_{i=1}^{n} (u^T x_i)^2 \]

= largest eigenvalue of \(XX^T\) (covariance matrix)

Another perspective:

minimize reconstruction error

\[ \sum_{i=1}^{n} ||x_i - uu^T x_i||^2 \]

(similar to least-squares regression?)
All principal components

\[ X_{d \times n} = U_{d \times d} Z_{d \times n} \]

\[
\begin{pmatrix}
  x_1 & \ldots & x_n
\end{pmatrix}
= 
\begin{pmatrix}
  u_1 & \ldots & u_d
\end{pmatrix} 
\begin{pmatrix}
  z_1 & \ldots & z_n
\end{pmatrix}
\]

- \( X \): data in original representation
- \( U \): principal components
- \( Z \): data in new representation
All principal components

\[
X_{d \times n} = U_{d \times d} Z_{d \times n}
\]

\[
\begin{pmatrix}
  x_1 & \ldots & x_n \\
\end{pmatrix} = 
\begin{pmatrix}
  u_1 & \ldots & u_d \\
\end{pmatrix} 
\begin{pmatrix}
  z_1 & \ldots & z_n \\
\end{pmatrix}
\]

- \textbf{X}: data in original representation
- \textbf{U}: principal components
- \textbf{Z}: data in new representation

- Each \( x_i \) can be expressed by a linear combination of principal components: \( x_i = \sum_{j=1}^{d} z_i^j u_j \)
- Components of projected data are uncorrelated
$r$ principal components

\[
\begin{align*}
X_{d \times n} & \cong U_{d \times r} \quad Z_{r \times n} \\
\begin{pmatrix}
x_1 & \ldots & x_n \\
\end{pmatrix} & \cong \\
\begin{pmatrix}
u_1 & \ldots & u_r \\
\end{pmatrix} \begin{pmatrix}
z_1 & \ldots & z_n \\
\end{pmatrix}
\end{align*}
\]

$X$: data in original representation
$U$: principal components
$Z$: data in new representation

Dimensionality reduction:
keep only the largest $r$ of $d$ eigenvectors

\[
x_i \approx \sum_{j=1}^{r} z_i^j u_j
\]
Eigen-faces [Turk, 1991]

Each \( \mathbf{x}_i \) is a face image, which is a vector in \( \mathbb{R}^d \)
\( d \) is the number of pixels

Each component \( \mathbf{x}_i^j \) is the intensity of the \( j \)-th pixel

\[
\mathbf{X}_{d \times n} \cong \mathbf{U}_{d \times r} \mathbf{Z}_{r \times n}
\]

These faces are from Yale face dataset.
Eigen-faces [Turk, 1991]

Each $x_i$ is a face image, which is a vector in $\mathbb{R}^d$

$d$ is the number of pixels

Each component $x_i^j$ is the intensity of the $j$-th pixel

$$X_{d \times n} \approx U_{d \times r} Z_{r \times n}$$

Used in image classification.

Individual entries in $z_i$’s are more meaningful than those in $x_i$’s.

These faces are from Yale face dataset.
Latent Semantic Analysis [Deerwater, 1990]

Each $x_i$ is a bag of words, which is a vector in $\mathbb{R}^d$.

$d$ is the number of words in the vocabulary.

Each component $x_i^j$ is the number of times word $j$ appears in document $i$.

$$X_{d \times n} \approx U_{d \times r} Z_{r \times n}$$

$$\begin{pmatrix}
\text{stocks: 2 \cdots 0} \\
\text{chairman: 4 \cdots 1} \\
\text{the: 8 \cdots 7} \\
\vdots \\
\text{wins: 0 \cdots 2} \\
\text{game: 1 \cdots 3}
\end{pmatrix} \approx 
\begin{pmatrix}
0.4 \cdots -0.001 \\
0.8 \cdots 0.03 \\
0.01 \cdots 0.04 \\
\vdots \\
0.002 \cdots 2.3 \\
0.003 \cdots 1.9
\end{pmatrix} 
\begin{pmatrix}
z_1 & \cdots & z_n
\end{pmatrix}$$
Latent Semantic Analysis [Deerwater, 1990]

Each $x_i$ is a bag of words, which is a vector in $\mathbb{R}^d$
$d$ is the number of words in the vocabulary

Each component $x^j_i$ is
the number of times word $j$ appears in document $i$

$$
\begin{bmatrix}
\text{stocks:} & 2 & \cdots & 0 \\
\text{chairman:} & 4 & \cdots & 1 \\
\text{the:} & 8 & \cdots & 7 \\
\vdots & \vdots & \ddots & \vdots \\
\text{wins:} & 0 & \cdots & 2 \\
\text{game:} & 1 & \cdots & 3
\end{bmatrix}
\approx
\begin{bmatrix}
0.4 & \cdots & -0.001 \\
0.8 & \cdots & 0.03 \\
0.01 & \cdots & 0.04 \\
\vdots & \vdots & \vdots \\
0.002 & \cdots & 2.3 \\
0.003 & \cdots & 1.9
\end{bmatrix}
\begin{bmatrix}
Z_1 & \cdots & Z_n
\end{bmatrix}
$$

Useful in information retrieval.
Eigen-documents gets at notion of semantics.
How to measure similarity between two documents?

$x_1, x_2$ versus $z_1, z_2$
Computing PCA

- Two ways of generating principal components:
  - Eigendecomposition: $XX^T = U\Lambda U^T$
  - Singular value decomposition: $X = U\Sigma V^T$

- Algorithm:
  - Center data so that $\sum_{i=1}^n x_i = 0$
  - Run SVD (which is one line in R):
    
    ```
    decomp <- svd(X, r)
    ```
    decomp$u are principal components
    decomp$d**2 are eigenvalues
How many principal components?

- Similar to question of “How many clusters?”
- Magnitude of eigenvalues indicate percentage of variance captured.
How many principal components?

- Similar to question of “How many clusters?”
- Magnitude of eigenvalues indicate percentage of variance captured.
- Eigenvalues on a face image dataset:
How many principal components?

- Similar to question of “How many clusters?”
- Magnitude of eigenvalues indicate percentage of variance captured.
- Eigenvalues on a face image dataset:

  ![Graph showing eigenvalues](image)

  - Eigenvalues drop off sharply, so don’t need that many.
  - But variance isn’t everything...
What if the data doesn’t live in a subspace?

• Ideal case: data lies in low-dimensional subspace plus Gaussian noise
What if the data doesn’t live in a subspace?

• Ideal case: data lies in low-dimensional subspace plus Gaussian noise

• A hypothetical example:
  – Original data is 100-dimensional
  – True manifold of data is 5-dimensional but lives in a 8-dimensional subspace
  – PCA can just find the 8-dimensional subspace, which still reduces redundancy
What if the data doesn’t live in a subspace?

- Ideal case: data lies in low-dimensional subspace plus Gaussian noise
- A hypothetical example:
  - Original data is 100-dimensional
  - True manifold of data is 5-dimensional but lives in a 8-dimensional subspace
  - PCA can just find the 8-dimensional subspace, which still reduces redundancy
- A cool technique: random projections
  - Randomly project data onto $O(\log n)$ dimensions
  - Pairwise distances preserved with high probability
  - Much more efficient than PCA
PCA summary

- Intuition: Capture variance of data
  Minimize reconstruction error
- Algorithm: eigenvalue problem
- Simple to use
- Applications: eigen-faces, eigen-documents, eigen-genes, etc.
Outline

• Introduction

• Methods
  – Principal component analysis (PCA)
  – Canonical correlation analysis (CCA)
  – Linear discriminant analysis (LDA)
  – Non-negative matrix factorization (NMF)
  – Independent component analysis (ICA)

• Case studies
  – Network anomaly detection
  – Multi-task learning
  – Part-of-speech tagging
  – Brain imaging

• Extensions, related methods, summary
Motivation for CCA [Hotelling, 1936]

Often, each data point actually consists of many views...

- Image retrieval: for each image, have the following:
  - Pixels (or other visual features)
  - Text around the image
Motivation for CCA [Hotelling, 1936]

Often, each data point actually consists of many views...

- **Image retrieval**: for each image, have the following:
  - Pixels (or other visual features)
  - Text around the image
- **Genomics**: for each gene, have the following:
  - Gene expression in DNA microarray
  - Position on genome
  - Chemical reactions catalyzed in metabolic pathways
Motivation for CCA [Hotelling, 1936]

Often, each data point actually consists of many views. . .

- **Image retrieval:** for each image, have the following:
  - Pixels (or other visual features)
  - Text around the image
- **Genomics:** for each gene, have the following:
  - Gene expression in DNA microarray
  - Position on genome
  - Chemical reactions catalyzed in metabolic pathways

Goal: reduce the dimensionality of the views jointly
From variance to correlation

PCA: find $u$ to maximize variance $\hat{E}(u^T x)^2$

CCA: find $(u, v)$ to maximize correlation $\hat{\text{corr}}(u^T x)(v^T y)$
From variance to correlation

PCA: find $u$ to maximize variance $\hat{E}(u^T x)^2$

CCA: find $(u, v)$ to maximize correlation $\hat{\text{corr}}(u^T x)(v^T y)$

CCA directions (green)
From variance to correlation

PCA: find $\mathbf{u}$ to maximize variance $\hat{\mathbb{E}}(\mathbf{u}^T \mathbf{x})^2$

CCA: find $(\mathbf{u}, \mathbf{v})$ to maximize correlation $\hat{\text{corr}}(\mathbf{u}^T \mathbf{x})(\mathbf{v}^T \mathbf{y})$

CCA directions (green)  PCA directions (black)
From variance to correlation

PCA: find \( u \) to maximize variance \( \hat{E}(u^T x)^2 \)

CCA: find \((u, v)\) to maximize correlation \( \text{corr}(u^T x)(v^T y) \)

CCA directions (green)  PCA directions (black)

Doing PCA separately on each view does not take advantage of relationship between two views.
CCA objective function

Objective: maximize correlation between projected views
CCA objective function

Objective: maximize correlation between projected views

\[ = \max_{u,v} \hat{\text{corr}}(u^T x, v^T y) = \max_{u,v} \frac{\hat{\text{cov}}(u^T x, v^T y)}{\sqrt{\hat{\text{var}}(u^T x)} \sqrt{\hat{\text{var}}(v^T y)}} \]
Objective: maximize correlation between projected views

$$\text{maximize } \text{corr}(u^T x, v^T y) = \max_{u,v} \frac{\text{cov}(u^T x, v^T y)}{\sqrt{\text{var}(u^T x)} \sqrt{\text{var}(v^T y)}}$$

$$= \max \text{cov}(u^T x, v^T y)$$

$$\text{var}(u^T x) = \text{var}(v^T y) = 1$$
CCA objective function

Objective: maximize correlation between projected views

\[
\begin{align*}
&= \max_{u,v} \hat{\text{corr}}(u^T x, v^T y) \\
&= \max_{u,v} \frac{\hat{\text{cov}}(u^T x, v^T y)}{\sqrt{\hat{\text{var}}(u^T x)} \sqrt{\hat{\text{var}}(v^T y)}} \\
&= \max_{\hat{\text{var}}(u^T x) = \hat{\text{var}}(v^T y) = 1} \hat{\text{cov}}(u^T x, v^T y) \\
&= \max_{||u^T X|| = ||v^T Y|| = 1} \sum_{i=1}^{n} (u^T x_i)(v^T y_i)
\end{align*}
\]
CCA objective function

Objective: maximize correlation between projected views

\[
= \max_{u,v} \hat{\text{corr}}(u^T x, v^T y) = \max_{u,v} \frac{\hat{\text{cov}}(u^T x, v^T y)}{\sqrt{\hat{\text{var}}(u^T x)} \sqrt{\hat{\text{var}}(v^T y)}}
\]

\[
= \max_{\hat{\text{var}}(u^T x)=\hat{\text{var}}(v^T y)=1} \hat{\text{cov}}(u^T x, v^T y)
\]

\[
= \max_{||u^T X||=||v^T Y||=1} \sum_{i=1}^{n} (u^T x_i)(v^T y_i)
\]

\[
= \max_{||u^T X||=||v^T Y||=1} u^T X Y^T v
\]
Objective: maximize correlation between projected views

\[
\max_{u,v} \hat{\text{corr}}(u^T x, v^T y) = \max_{u,v} \frac{\hat{\text{cov}}(u^T x, v^T y)}{\sqrt{\hat{\text{var}}(u^T x)} \sqrt{\hat{\text{var}}(v^T y)}}
\]

\[
= \max_{\hat{\text{var}}(u^T x) = \hat{\text{var}}(v^T y) = 1} \hat{\text{cov}}(u^T x, v^T y)
\]

\[
= \max_{||u^T X|| = ||v^T Y|| = 1} \sum_{i=1}^{n} (u^T x_i)(v^T y_i)
\]

\[
= \max_{||u^T X|| = ||v^T Y|| = 1} u^T X Y^T v
\]

\[
= \text{largest generalized eigenvalue } \lambda \text{ given by}
\]

\[
\begin{pmatrix}
0 & XX^T \\
YX^T & 0
\end{pmatrix}
\begin{pmatrix}
u \\
v
\end{pmatrix}
= \lambda
\begin{pmatrix}
XX^T & 0 \\
0 & YY^T
\end{pmatrix}
\begin{pmatrix}
u \\
v
\end{pmatrix},
\]

which reduces to an ordinary eigenvalue problem.
CCA objective function

Objective: maximize correlation between projected views

\[
\begin{align*}
&= \max_{u,v} \hat{\text{corr}}(u^T x, v^T y) = \max_{u,v} \frac{\hat{\text{cov}}(u^T x, v^T y)}{\sqrt{\hat{\text{var}}(u^T x)} \sqrt{\hat{\text{var}}(v^T y)}} \\
&= \max_{\hat{\text{var}}(u^T x)=\hat{\text{var}}(v^T y)=1} \hat{\text{cov}}(u^T x, v^T y) \\
&= \max_{||u^T X||=||v^T Y||=1} \sum_{i=1}^{n} (u^T x_i)(v^T y_i) \\
&= \max_{||u^T X||=||v^T Y||=1} u^T X Y^T v \\
&= \text{largest generalized eigenvalue } \lambda \text{ given by} \\
\begin{pmatrix}
0 & X Y^T \\
Y X^T & 0
\end{pmatrix}
\begin{pmatrix}
u \\
v
\end{pmatrix}
= \lambda
\begin{pmatrix}
XX^T & 0 \\
0 & YY^T
\end{pmatrix}
\begin{pmatrix}
u \\
v
\end{pmatrix},
\end{align*}
\]

which reduces to an ordinary eigenvalue problem.

Note: canonical components \(u, v\) are invariant to affine transformation of \(X, Y\).
Outline

• Introduction

• Methods
  – Principal component analysis (PCA)
  – Canonical correlation analysis (CCA)
  – Linear discriminant analysis (LDA)
  – Non-negative matrix factorization (NMF)
  – Independent component analysis (ICA)

• Case studies
  – Network anomaly detection
  – Multi-task learning
  – Part-of-speech tagging
  – Brain imaging

• Extensions, related methods, summary
Motivation for LDA [Fisher, 1936]

What is the best linear projection?
Motivation for LDA [Fisher, 1936]

What is the best linear projection?

PCA solution
Motivation for LDA [Fisher, 1936]

What is the best linear projection with these labels?

PCA solution
Motivation for LDA [Fisher, 1936]

What is the best linear projection with these labels?

PCA solution

LDA solution
Motivation for LDA [Fisher, 1936]

What is the best linear projection with these labels?

Goal: reduce the dimensionality given labels

Idea: want projection to maximize overall interclass variance relative to intraclass variance

PCA solution

LDA solution
LDA objective function

Global mean: $\mu = \sum_i x_i$ \hspace{1cm} $X_g = (x_1 - \mu, \ldots, x_n - \mu)$

Class mean: $\mu_y = \sum_{i:y_i=y} x_i$ \hspace{1cm} $X_c = (x_1 - \mu_y^1, \ldots, x_n - \mu_y^n)$
LDA objective function

Global mean: $\mu = \sum_i x_i \quad X_g = (x_1 - \mu, \ldots, x_n - \mu)$

Class mean: $\mu_y = \sum_{i:y_i=y} x_i \quad X_c = (x_1 - \mu_y, \ldots, x_n - \mu_y)$

Objective: maximize $\frac{\text{total variance}}{\text{intra-class variance}} = \frac{\text{interclass variance}}{\text{intra-class variance}} + 1$
LDA objective function

Global mean: $\mu = \sum_i x_i \quad X_g = (x_1 - \mu, \ldots, x_n - \mu)$

Class mean: $\mu_y = \sum_{i:y_i=y} x_i \quad X_c = (x_1 - \mu y_1, \ldots, x_n - \mu y_n)$

Objective: maximize $\frac{\text{total variance}}{\text{intraclss variance}} = \frac{\text{interclass variance}}{\text{intraclss variance}} + 1$

$= \max_u \frac{\sum_{i=1}^n (u^T(x_i - \mu))^2}{\sum_{i=1}^n (u^T(x_i - \mu y_i))^2}$
LDA objective function

Global mean: $\mu = \sum_i x_i \quad X_g = (x_1 - \mu, \ldots, x_n - \mu)$

Class mean: $\mu_y = \sum_{i:y_i=y} x_i \quad X_c = (x_1 - \mu y_1, \ldots, x_n - \mu y_n)$

Objective: maximize

$$\frac{\text{total variance}}{\text{intraclass variance}} = \frac{\text{interclass variance}}{\text{intraclass variance}} + 1$$

$$= \max_u \frac{\sum_{i=1}^n (u^T(x_i - \mu))^2}{\sum_{i=1}^n (u^T(x_i - \mu y_i))^2}$$

$$= \max ||u^T X_c|| = 1 \sum_{i=1}^n (u^T(x_i - \mu))^2$$
LDA objective function

Global mean: $\mu = \sum_i x_i \quad X_g = (x_1 - \mu, \ldots, x_n - \mu)$

Class mean: $\mu_y = \sum_{i:y_i=y} x_i \quad X_c = (x_1 - \mu y_1, \ldots, x_n - \mu y_n)$

Objective: maximize $\frac{\text{total variance}}{\text{intra-class variance}} = \frac{\text{inter-class variance}}{\text{intra-class variance}} + 1$

$$= \max_u \frac{\sum_{i=1}^n (u^T (x_i - \mu))^2}{\sum_{i=1}^n (u^T (x_i - \mu y_i))^2}$$

$$= \max \sum_{i=1}^n (u^T (x_i - \mu))^2 \quad \|u^T X_c\| = 1$$

$$= \max \|u^T X_g X_g^T u\| \quad \|u^T X_c\| = 1$$
LDA objective function

Global mean: \( \mu = \sum_i x_i \)
\[ X_g = (x_1 - \mu, \ldots, x_n - \mu) \]

Class mean: \( \mu_y = \sum_{i:y_i=y} x_i \)
\[ X_c = (x_1 - \mu y_1, \ldots, x_n - \mu y_n) \]

Objective: maximize \[
\frac{\text{total variance}}{\text{intraclass variance}} = \frac{\text{interclass variance}}{\text{intraclass variance}} + 1
\]

\[
\begin{align*}
\quad & = \max_u \frac{\sum_{i=1}^n (u^T (x_i - \mu))^2}{\sum_{i=1}^n (u^T (x_i - \mu y_i))^2} \\
\quad & = \max_{\|u^T X_c\|=1} \sum_{i=1}^n (u^T (x_i - \mu))^2 \\
\quad & = \max_{\|u^T X_c\|=1} u^T X_g X_g^T u \\
\quad & = \text{largest generalized eigenvalue } \lambda \text{ given by} \\
\quad & \quad (X_g X_g^T)u = \lambda (X_c X_c^T)u.
\end{align*}
\]
Summary so far

• Recall $Z \approx U^T X$; criteria for $U$:
  – PCA: maximize variance
  – CCA: maximize correlation
  – LDA: maximize $\frac{\text{interclass variance}}{\text{intraclass variance}}$

• All these methods reduce to solving generalized eigenvalue problems

• Next (NMF, ICA):
  more complex criteria for $U$
Outline

• Introduction
• Methods
  – Principal component analysis (PCA)
  – Canonical correlation analysis (CCA)
  – Linear discriminant analysis (LDA)
  – Non-negative matrix factorization (NMF)
  – Independent component analysis (ICA)
• Case studies
  – Network anomaly detection
  – Multi-task learning
  – Part-of-speech tagging
  – Brain imaging
• Extensions, related methods, summary
Motivation for NMF [Paatero, ’94; Lee, ’99]

Back to basic PCA setting (single view, no labels)

\[
\begin{align*}
X_{d \times n} & \cong U_{d \times r} Z_{r \times n} \\
\begin{pmatrix}
  x_1 & \cdots & x_n
\end{pmatrix} & \cong \\
\begin{pmatrix}
  u_1 & \cdots & u_r
\end{pmatrix} \begin{pmatrix}
  z_1 & \cdots & z_n
\end{pmatrix}
\end{align*}
\]

\(X\): data in original representation  
\(U\): principal components  
\(Z\): data in new representation
Motivation for NMF [Paatero, ’94; Lee, ’99]

Back to basic PCA setting (single view, no labels)

\[ X_{d \times n} \cong U_{d \times r} Z_{r \times n} \]

\[ \left( \begin{array}{c} x_1 \\ \vdots \\ x_n \end{array} \right) \cong \left( \begin{array}{c} u_1 \\ \vdots \\ u_r \end{array} \right) \left( \begin{array}{c} z_1 \\ \vdots \\ z_n \end{array} \right) \]

- Data is not just any arbitrary real vector:
  - Text modeling: each document is a vector of term frequencies
  - Gene expression: each gene is a vector of expression profiles
  - Collaborative filtering: each user is a vector of movie ratings
Motivation for NMF [Paatero, ’94; Lee, ’99]

Back to basic PCA setting (single view, no labels)

\[ \mathbf{X}_{d \times n} \approx \mathbf{U}_{d \times r} \mathbf{Z}_{r \times n} \]

\[ \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix} \approx \begin{pmatrix} u_1 & \cdots & u_r \end{pmatrix} \begin{pmatrix} z_1 & \cdots & z_n \end{pmatrix} \]

• Data is not just any arbitrary real vector:
  – Text modeling: each document is a vector of term frequencies
  – Gene expression: each gene is a vector of expression profiles
  – Collaborative filtering: each user is a vector of movie ratings

• Each basis vector \( \mathbf{u}_i \) is an “eigen-document/eigen-gene/eigen-user”

• Would like \( \mathbf{U} \) and \( \mathbf{Z} \) to have only non-negative entries so that we can interpret each point as combination of prototypes
Motivation for NMF [Paatero, ’94; Lee, ’99]

Back to basic PCA setting (single view, no labels)

\[ \mathbf{X}_{d \times n} \cong \mathbf{U}_{d \times r} \mathbf{Z}_{r \times n} \]

\[
\begin{pmatrix}
\mathbf{x}_1 & \ldots & \mathbf{x}_n
\end{pmatrix} \cong \begin{pmatrix}
\mathbf{u}_1 & \ldots & \mathbf{u}_r
\end{pmatrix} \begin{pmatrix}
\mathbf{z}_1 & \ldots & \mathbf{z}_n
\end{pmatrix}
\]

- Data is not just any arbitrary real vector:
  - Text modeling: each document is a vector of term frequencies
  - Gene expression: each gene is a vector of expression profiles
  - Collaborative filtering: each user is a vector of movie ratings
- Each basis vector \( \mathbf{u}_i \) is an “eigen-document/eigen-gene/eigen-user”
- Would like \( \mathbf{U} \) and \( \mathbf{Z} \) to have only non-negative entries
  so that we can interpret each point as combination of prototypes

Goal: reduce the dimensionality given non-negativity constraints
Qualitative difference between NMF and PCA

\[ x \approx \sum_{j=1}^{r} z_j u_j \]

- Sum of basis vectors must be (positively) additive \((z_j \geq 0)\)
- The basis vectors \(u_i\)'s tend to be sparse
- NMF recovers a parts-based representation of \(x\) whereas PCA recovers a holistic representations
Qualitative difference between NMF and PCA

\[ x \approx \sum_{j=1}^{r} z_j u_j \]

- Sum of basis vectors must be (positively) additive \((z_j \geq 0)\)
- The basis vectors \(u_i\)'s tend to be sparse
- NMF recovers a parts-based representation of \(x\) whereas PCA recovers a holistic representations
- Caveat for images: sparsity depends on proper alignment (remember, representation is still a bag of pixels)
NMF machinery

• Objectives to minimize (all entries in $X$, $U$, $Z$ non-negative)
  
  – Frobenius norm (same as PCA but with non-negativity constraints):
    \[
    ||X - UZ||_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{r} (X_{ji} - (UZ)_{ji})^2
    \]
  
  – KL divergence:
    \[
    KL(X||UZ) = \sum_{i=1}^{n} \sum_{j=1}^{r} X_{ji} \log \frac{X_{ji}}{(UZ)_{ji}} - X_{ji} + (UZ)_{ji}
    \]
NMF machinery

• Objectives to minimize (all entries in $X, U, Z$ non-negative)
  – Frobenius norm (same as PCA but with non-negativity constraints):
    \[
    \|X - UZ\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{r} (X_{ji} - (UZ)_{ji})^2
    \]
  – KL divergence:
    \[
    KL(X||UZ) = \sum_{i=1}^{n} \sum_{j=1}^{r} X_{ji} \log \frac{X_{ji}}{(UZ)_{ji}} - X_{ji} + (UZ)_{ji}
    \]

• Algorithm
  – Hard non-convex optimization problem:
    could get stuck in local minima, need to worry about initialization
  – Simple/fast multiplicative update rule [Lee & Seung ’99, ’01]
NMF machinery

• Objectives to minimize (all entries in $X, U, Z$ non-negative)
  – Frobenius norm (same as PCA but with non-negativity constraints):
    $$\|X - UZ\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{r} (X_{ji} - (UZ)_{ji})^2$$
  – KL divergence:
    $$\text{KL}(X \| UZ) = \sum_{i=1}^{n} \sum_{j=1}^{r} X_{ji} \log \frac{X_{ji}}{(UZ)_{ji}} - X_{ji} + (UZ)_{ji}$$

• Algorithm
  – Hard non-convex optimization problem:
    could get stuck in local minima, need to worry about initialization
  – Simple/fast multiplicative update rule [Lee & Seung '99, '01]

• Relationship to other methods
  – Vector quantization: $z_j$ is 1 in exactly one component $j$
  – Probabilistic latent semantic analysis: equivalent to 2nd objective
  – Latent Dirichlet Allocation: more Bayesian version of pLSI
Outline

• Introduction

• Methods
  – Principal component analysis (PCA)
  – Canonical correlation analysis (CCA)
  – Linear discriminant analysis (LDA)
  – Non-negative matrix factorization (NMF)
  – Independent component analysis (ICA)

• Case studies
  – Network anomaly detection
  – Multi-task learning
  – Part-of-speech tagging
  – Brain imaging

• Extensions, related methods, summary
Motivation for ICA [Herault & Jutten, ’86]

Cocktail party problem:
\[ d \text{ people, } d \text{ microphones, } n \text{ time steps} \]
Assume: people are speaking independently (z)
acoustics mix linearly through an invertible \( \mathbf{U} \)

\[ \mathbf{x} = \mathbf{Uz} \]
Motivation for ICA [Herault & Jutten, ’86]

Cocktail party problem:
- $d$ people, $d$ microphones, $n$ time steps

Assume:
- people are speaking independently ($z$)
- acoustics mix linearly through an invertible $U$

\[ x = Uz \]
Motivation for ICA [Herault & Jutten, ’86]

Cocktail party problem:

$d$ people, $d$ microphones, $n$ time steps

Assume: people are speaking independently ($z$)

acoustics mix linearly through an invertible $U$

\[ x = Uz \]

Goal: find transformation that makes components of $z$ as independent as possible
PCA versus ICA
PCA versus ICA

PCA solution
PCA versus ICA

PCA solution

ICA solution
PCA versus ICA

PCA solution

ICA solution

Original signal
PCA versus ICA

PCA solution  ICA solution  Original signal

ICA finds independent components; doesn’t work if data is Gaussian:
PCA versus ICA

PCA solution  ICA solution  Original signal

ICA finds independent components; doesn’t work if data is Gaussian:
ICA algorithm

\[ x = Uz \]

- Preprocessing: whiten data \( X \) with PCA so that components are uncorrelated
ICA algorithm

\[ x = Uz \]

- Preprocessing: whiten data \( X \) with PCA so that components are uncorrelated
- Find \( U^{-1} \) to maximize independence of \( z = U^{-1}x \)
- How to measure independence? mutual information, negentropy, non-Gaussianity (e.g., kurtosis)
ICA algorithm

\[ x = Uz \]

- Preprocessing: whiten data \( X \) with PCA so that components are uncorrelated.
- Find \( U^{-1} \) to maximize independence of \( z = U^{-1}x \).
- How to measure independence? mutual information, negentropy, non-Gaussianity (e.g., kurtosis).
- Hard non-convex optimization.
- Methods for solving: fastICA, kernelICA, ProDenICA.
Outline

• Introduction
• Methods
  – Principal component analysis (PCA)
  – Canonical correlation analysis (CCA)
  – Linear discriminant analysis (LDA)
  – Non-negative matrix factorization (NMF)
  – Independent component analysis (ICA)
• Case studies
  – Network anomaly detection
  – Multi-task learning
  – Part-of-speech tagging
  – Brain imaging
• Extensions, related methods, summary
Network anomaly detection [Lakhina, ’05]

Raw data: traffic flow on each link in the network during each time interval
Network anomaly detection [Lakhina, ’05]

Raw data: traffic flow on each link in the network during each time interval

Model assumption: traffic is sum of flows along a few paths
Apply PCA: principal component intuitively represents a path
Network anomaly detection [Lakhina, ’05]

Raw data: traffic flow on each link in the network during each time interval

Model assumption: traffic is sum of flows along a few paths
Apply PCA: principal component intuitively represents a path
Anomaly: when traffic deviates from first few principal components
Multi-task learning [Ando & Zhang, ’05]

Setup:

• Have a set of related tasks (classify documents for various users)
• Each task has a classifier (weights of a linear classifier)
• Want to share structure between classifiers
Multi-task learning [Ando & Zhang, ’05]

Setup:

• Have a set of related tasks (classify documents for various users)
• Each task has a classifier (weights of a linear classifier)
• Want to share structure between classifiers

One step of their procedure:

given a set of classifiers $x_1, \ldots, x_n$,
run PCA to identify shared structure:

$$X = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix} \approx UZ$$

Each data point is a linear classifier
Each principal component is a eigen-classifier
Unsupervised POS tagging [Schütze, ’95]

Part-of-speech (POS) tagging task:

Input: I like reducing the dimensionality of data.
Output: NOUN VERB VERB(-ING) DET NOUN PREP NOUN.
Unsupervised POS tagging [Schütze, ’95]

Part-of-speech (POS) tagging task:

- Input: I like reducing the dimensionality of data.
- Output: NOUN VERB VERB(-ING) DET NOUN PREP NOUN.

Key idea: words appearing in similar contexts should have the same POS tags.

Problem: contexts are too sparse.
Unsupervised POS tagging [Schütze, ’95]

Part-of-speech (POS) tagging task:

Input: I like reducing the dimensionality of data.
Output: NOUN VERB VERB(-ING) DET NOUN PREP NOUN.

Key idea: words appearing in similar contexts should have the same POS tags.

Problem: contexts are too sparse.

Solution: run PCA first,
then cluster using new representation.

Each data point is (the context of) a word.
Brain imaging
Brain imaging

Data: EEG/MEG/fMRI readings
Brain imaging

Data: EEG/MEG/fMRI readings
Goal: separate signals into sources

\[ S = \]

Data: EEG/MEG/fMRI readings
Goal: separate signals into sources
Brain imaging

Data: EEG/MEG/fMRI readings

Goal: separate signals into sources

One solution: ICA
Another solution: CCA [Borga, ’02]
Brain imaging

One solution: ICA
Another solution: CCA [Borga, ’02]

The two views are the signals $s$ at adjacent time steps:

$(x_1, y_1) = (s(1), s(2))$
$(x_2, y_2) = (s(2), s(3))$
$(x_3, y_3) = (s(3), s(4))$

... More robust and faster than ICA

Data: EEG/MEG/fMRI readings
Goal: separate signals into sources
Outline

• Introduction

• Methods
  – Principal component analysis (PCA)
  – Canonical correlation analysis (CCA)
  – Linear discriminant analysis (LDA)
  – Non-negative matrix factorization (NMF)
  – Independent component analysis (ICA)

• Case studies
  – Network anomaly detection
  – Multi-task learning
  – Part-of-speech tagging
  – Brain imaging

• Extensions, related methods, summary
Extensions

• Kernel trick:
  – Find non-linear subspaces with same machinery
• Produce sparse solutions
• Ensure robustness:
  – Be insensitive to outliers
• Make probabilistic (e.g., factor analysis):
  – Handle missing data
  – Estimate uncertainty
  – Natural way to incorporate in a larger model
• Automatically choose number of dimensions
Curtain call

PCA: find subspace that captures most variance in data; eigenvalue problem
Curtain call

PCA: find subspace that captures most variance in data; eigenvalue problem

CCA: find pair of subspaces that captures most correlation; generalized eigenvalue problem
Curtain call

PCA: find subspace that captures most variance in data; eigenvalue problem

CCA: find pair of subspaces that captures most correlation; generalized eigenvalue problem

LDA: find subspace that maximizes \(\frac{\text{intra-class variance}}{\text{inter-class variance}}\); generalized eigenvalue problem
Curtain call

PCA: find subspace that captures most variance in data; eigenvalue problem

CCA: find pair of subspaces that captures most correlation; generalized eigenvalue problem

LDA: find subspace that maximizes $\frac{\text{intraclass variance}}{\text{interclass variance}}$, generalized eigenvalue problem

NMF: find subspace that minimizes reconstruction error for non-negative data; non-trivial optimization problem
Curtain call

PCA: find subspace that captures most variance in data; eigenvalue problem

CCA: find pair of subspaces that captures most correlation; generalized eigenvalue problem

LDA: find subspace that maximizes $\frac{\text{intra class variance}}{\text{inter class variance}}$; generalized eigenvalue problem

NMF: find subspace that minimizes reconstruction error for non-negative data; non-trivial optimization problem

ICA: find subspace where sources are independent; non-trivial optimization problem