Structure-Preserving Method for Dimension Reduction

Ewa Nowakowska

Institute of Computer Science, Polish Academy of Sciences

Joint Statistical Meetings, San Diego, CA
29 July 2012
Outline

1. Introduction
   - Model and notation
   - Basic definitions
   - Concept

2. The method
   - Isotropic transformation
   - Weighting
   - Algorithm

3. Summary
Outline

1. Introduction
   - Model and notation
   - Basic definitions
   - Concept

2. The method
   - Isotropic transformation
   - Weighting
   - Algorithm

3. Summary
Model and notation

- data: \( X = (x_1, \ldots, x_n)^T, \ X \in \mathbb{R}^{n \times d} \)

- model: \( f(x) = \pi_1 f_1(\mu_1, \Sigma_1)(x) + \ldots + \pi_k f_k(\mu_k, \Sigma_k)(x) \), where

\[
f_l(\mu_l, \Sigma_l)(x) = \frac{1}{(\sqrt{2\pi})^d \sqrt{\det \Sigma_l}} e^{-\frac{1}{2}(x-\mu_l)^T \Sigma_l^{-1}(x-\mu_l)}.
\]

- additional assumptions:
  - equal mixing factors \( \pi_1 = \cdots = \pi_k = \frac{1}{k} \)
  - heterogeneity \( \Sigma_{l_1} \neq \Sigma_{l_2} \)
  - large space dimension \( d > k - 1 \)
  - large sample size \( n \gg d \)
  - number of components \( k \) known
Model and notation

- **data:** \( X = (x_1, \ldots, x_n)^T, \; X \in \mathbb{R}^{n \times d} \)

- **model:** \( f(x) = \pi_1 f_1(\mu_1, \Sigma_1)(x) + \ldots + \pi_k f_k(\mu_k, \Sigma_k)(x), \)

where

\[
f_l(\mu_l, \Sigma_l)(x) = \frac{1}{(\sqrt{2\pi})^d \sqrt{\det \Sigma_l}} e^{-\frac{1}{2}(x-\mu_l)^T \Sigma_l^{-1}(x-\mu_l)}.\]

- **additional assumptions:**
  - equal mixing factors \( \pi_1 = \cdots = \pi_k = \frac{1}{k} \)
  - heterogeneity \( \Sigma_{l_1} \neq \Sigma_{l_2} \)
  - large space dimension \( d > k - 1 \)
  - large sample size \( n \gg d \)
  - number of components \( k \) known
Model and notation

- **data:** \( X = (x_1, \ldots, x_n)^T, \ X \in \mathbb{R}^{n \times d} \)

- **model:** \( f(x) = \pi_1 f_1(\mu_1, \Sigma_1)(x) + \ldots + \pi_k f_k(\mu_k, \Sigma_k)(x) \),

where

\[
f_l(\mu_l, \Sigma_l)(x) = \frac{1}{(\sqrt{2\pi})^d \sqrt{\det \Sigma_l}} e^{-\frac{1}{2} (x-\mu_l)^T \Sigma_l^{-1} (x-\mu_l)}.
\]

- **additional assumptions:**
  - equal mixing factors \( \pi_1 = \cdots = \pi_k = \frac{1}{k} \)
  - heterogeneity \( \Sigma_{l_1} \neq \Sigma_{l_2} \)
  - large space dimension \( d > k - 1 \)
  - large sample size \( n \gg d \)
  - number of components \( k \) known
Basic facts and definitions

Let $\mu_X$ and $\Sigma_X$ be the empirical estimates of $\mu$ and $\Sigma$.

Definition (Scatter decomposition)
Let $T_X = n\Sigma_X$. Then $T_X = W_X + B_X$ constitutes the total scatter decomposition to its between and within cluster component.

Definition (Isotropic position)
We say that data is in isotropic position if $\mu_X = 0$ and $T_X = I$.

Definition (Principal component subspace $PC(k-1)$)
By principal component subspace $PC(k-1)$ we understand the subspace spanned by the first $k-1$ principal components.
Basic facts and definitions

Let $\mu_X$ and $\Sigma_X$ be the empirical estimates of $\mu$ and $\Sigma$.

**Definition (Scatter decomposition)**

Let $T_X = n\Sigma_X$. Then $T_X = W_X + B_X$ constitutes the total scatter decomposition to its between and within cluster component.

**Definition (Isotropic position)**

We say that data is in isotropic position if $\mu_X = 0$ and $T_X = I$.

**Definition (Principal component subspace $PC(k-1)$)**

By principal component subspace $PC(k-1)$ we understand the subspace spanned by the first $k-1$ principal components.
Basic facts and definitions

Let \( \mu_X \) and \( \Sigma_X \) be the empirical estimates of \( \mu \) and \( \Sigma \).

**Definition (Scatter decomposition)**

Let \( T_X = n\Sigma_X \). Then \( T_X = W_X + B_X \) constitutes the total scatter decomposition to its between and within cluster component.

**Definition (Isotropic position)**

We say that data is in isotropic position if \( \mu_X = 0 \) and \( T_X = I \).

**Definition (Principal component subspace \( PC(k - 1) \))**

By principal component subspace \( PC(k - 1) \) we understand the subspace spanned by the first \( k - 1 \) principal components.
Basic facts and definitions

Let $\mu_X$ and $\Sigma_X$ be the empirical estimates of $\mu$ and $\Sigma$.

**Definition (Scatter decomposition)**

Let $T_X = n\Sigma_X$. Then $T_X = W_X + B_X$ constitutes the total scatter decomposition to its between and within cluster component.

**Definition (Isotropic position)**

We say that data is in isotropic position if $\mu_X = 0$ and $T_X = I$.

**Definition (Principal component subspace $PC(k - 1)$)**

By principal component subspace $PC(k - 1)$ we understand the subspace spanned by the first $k - 1$ principal components.
Basic facts and definitions

Definition (Fisher’s subspace $S^*$)

We define Fisher’s subspace (Fisher’s discriminant) as

$$S^* = \arg\max_{S \subset \mathbb{R}^d} \frac{\sum_{j=1}^{k-1} v_j^T B_X v_j}{\sum_{j=1}^{k-1} v_j^T T_X v_j},$$

where $v_1, \ldots, v_{k-1}$ is the orthonormal basis for $S$.

Equivalently, $S^*$ – solution to an eigenproblem with $T_X^{-1} B_X$.

Definition (Structure distinctness coefficient $\bar{\lambda}^X$)

$$\bar{\lambda}^X = \frac{1}{k-1} \sum_{j=1}^{k-1} \lambda_j T_X^{-1} B_X.$$

Ewa Nowakowska

Institute of Computer Science, Polish Academy of Sciences

Structure-Preserving Method for Dimension Reduction
Basic facts and definitions

**Definition (Fisher’s subspace $S^*$)**

We define Fisher’s subspace (Fisher’s discriminant) as

$$S^* = \arg\max_{S \subseteq \mathbb{R}^d} \frac{\sum_{j=1}^{k-1} v_j^T B_X v_j}{\sum_{j=1}^{k-1} v_j^T T_X v_j},$$

where $v_1, \ldots, v_{k-1}$ is the orthonormal basis for $S$.

**Equivalently**, $S^*$ – solution to an eigenproblem with $T_X^{-1}B_X$.

**Definition (Structure distinctness coefficient $\bar{\lambda}^X$)**

$$\bar{\lambda}^X = \frac{1}{k-1} \sum_{j=1}^{k-1} \lambda_j^{-1} T_X^{-1} B_X,$$
Basic facts and definitions

**Definition (Fisher’s subspace $S^*$)**

We define Fisher’s subspace (Fisher’s discriminant) as

$$S^* = \arg\max_{S \subset \mathbb{R}^d} \frac{\sum_{j=1}^{k-1} v_j^T B X v_j}{\sum_{j=1}^{k-1} v_j^T T X v_j},$$

where $v_1, \ldots, v_{k-1}$ is the orthonormal basis for $S$.

**Equivalently**, $S^*$ – solution to an eigenproblem with $T_X^{-1} B X$.

**Definition (Structure distinctness coefficient $\bar{\lambda}^X$)**

$$\bar{\lambda}^X = \frac{1}{k-1} \sum_{j=1}^{k-1} \lambda_j T_X^{-1} B X$$
Concept

Inspired by:

original data $\rightarrow$ isotropic data $\rightarrow$ weighted data
Outline

1. Introduction
   - Model and notation
   - Basic definitions
   - Concept

2. The method
   - Isotropic transformation
   - Weighting
   - Algorithm

3. Summary
Isotropic transformation (IT)

- centering:
  - mean subtraction: \( X_0 = (x_1 - \mu_X, \ldots, x_n - \mu_X)^T \)

- decorrelation:
  - spectral decomposition: \( T_{X_0} = A_{T_{X_0}} L_{T_{X_0}} A_{T_{X_0}}^T \)
  - simple manipulation: \( (X_0 A_{T_{X_0}} L_{T_{X_0}}^{-1/2})^T (X_0 A_{T_{X_0}} L_{T_{X_0}}^{-1/2}) = I \)
  - isotropic transformation: \( Y = X_0 A_{T_{X_0}} L_{T_{X_0}}^{-1/2} \)

**Lemma (Eigenvalues preservation)**
IT does not affect eigenvalues for the Fisher’s task \( \lambda^X_j = \lambda^Y_j \).

**Corollary (Distinctness preservation)**
IT does not change structure distinctness \( \bar{\lambda}^X = \bar{\lambda}^Y \).
Isotropic transformation (IT)

- centering:
  
  mean subtraction: \( X_0 = (x_1 - \mu_X, \ldots, x_n - \mu_X)^T \)

- decorrelation:

  spectral decomposition: \( T_{X_0} = A_{T_{X_0}} L_{T_{X_0}} A_{T_{X_0}}^T \)

  simple manipulation: \( \left( X_0 A_{T_{X_0}} L_{T_{X_0}}^{-\frac{1}{2}} \right)^T \left( X_0 A_{T_{X_0}} L_{T_{X_0}}^{-\frac{1}{2}} \right) = I \)

  isotropic transformation: \( Y = X_0 A_{T_{X_0}} L_{T_{X_0}}^{-\frac{1}{2}} \)

Lemma (Eigenvalues preservation)

IT does not affect eigenvalues for the Fisher’s task \( \lambda_j^X = \lambda_j^Y \).

Corollary (Distinctness preservation)

IT does not change structure distinctness \( \bar{\lambda}^X = \bar{\lambda}^Y \).
Isotropic transformation (IT)

- centering:
  
  mean subtraction: \( X_0 = (x_1 - \mu_X, \ldots, x_n - \mu_X)^T \)

- decorrelation:
  
  spectral decomposition: \( T_{X_0} = A_{T_{X_0}} L_{T_{X_0}} A_{T_{X_0}}^T \)

  simple manipulation: \( \left( X_0 A_{T_{X_0}} L_{T_{X_0}}^{-\frac{1}{2}} \right)^T \left( X_0 A_{T_{X_0}} L_{T_{X_0}}^{-\frac{1}{2}} \right) = I \)

  isotropic transformation: \( Y = X_0 A_{T_{X_0}} L_{T_{X_0}}^{-\frac{1}{2}} \)

Lemma (Eigenvalues preservation)

IT does not affect eigenvalues for the Fisher’s task \( \lambda_j^X = \lambda_j^Y \).

Corollary (Distinctness preservation)

IT does not change structure distinctness \( \bar{\lambda}^X = \bar{\lambda}^Y \).
Isotropic transformation (IT)

- **Centering:**
  
  Mean subtraction: \( X_0 = (x_1 - \mu_X, \ldots, x_n - \mu_X)^T \)

- **Decorrelation:**
  
  Spectral decomposition: \( T_{X_0} = A_{T_{X_0}} L_{T_{X_0}} A_{T_{X_0}}^T \)
  
  Simple manipulation: \( \left( X_0 A_{T_{X_0}} L_{T_{X_0}}^{-\frac{1}{2}} \right)^T \left( X_0 A_{T_{X_0}} L_{T_{X_0}}^{-\frac{1}{2}} \right) = I \)
  
  Isotropic transformation: \( Y = X_0 A_{T_{X_0}} L_{T_{X_0}}^{-\frac{1}{2}} \)

**Lemma (Eigenvalues preservation)**

IT does not affect eigenvalues for the Fisher’s task \( \lambda_j^X = \lambda_j^Y \).

**Corollary (Distinctness preservation)**

IT does not change structure distinctness \( \bar{\lambda}^X = \bar{\lambda}^Y \).
Weighting - requirements and the function

- Differentiate variability across the directions
- Reduce variability in all the directions but the ones determined by the cluster centers
- Bring principal components close to the directions of best between-cluster discrimination
- Introduce only little distortion to the structure
- Relocate the extreme observations only leaving the core of the structure almost untouched

Weighting: \( \omega_i = \sqrt{\frac{1}{1 + \frac{1}{\alpha} \|y_i\|^2}} \) and \( Z = \text{diag}(\omega_1, \ldots, \omega_n) Y \).
**Weighting - requirements and the function**

- Differentiate variability across the directions
- Reduce variability in all the directions but the ones determined by the cluster centers
- Bring principal components close to the directions of best between-cluster discrimination
- Introduce only little distortion to the structure
- Relocate the extreme observations only leaving the core of the structure almost untouched

Weighting: \( \omega_i = \sqrt{\frac{1}{1 + \frac{1}{\alpha} \|y_i\|^2}} \) and \( Z = \text{diag}(\omega_1, \ldots, \omega_n) Y \).
Weighting - requirements and the function

- Differentiate variability across the directions
- Reduce variability in all the directions but the ones determined by the cluster centers
- Bring principal components close to the directions of best between-cluster discrimination
- Introduce only little distortion to the structure
- Relocate the extreme observations only leaving the core of the structure almost untouched

Weighting: \( \omega_i = \sqrt{\frac{1}{1 + \frac{1}{\alpha} \|y_i\|^2}} \) and \( Z = \text{diag}(\omega_1, \ldots, \omega_n)Y \).
Weighting – structure distinctness

Theorem (Structure distinctness preservation)
In agreement with the previous notation and assumptions

\[ |\bar{\lambda}^Z - \bar{\lambda}^X| \leq \frac{1}{\sqrt{n}} \left( \frac{d}{\alpha} \left( \lambda^X + \sqrt{k} \right) \right) + r_1 \left( \frac{1}{n} \right), \]

where \( r_1 \left( \frac{1}{n} \right) \) denotes a remainder of the first order of \( 1/n \).

Proof (Idea).
show smallness of weights’ variance \( \rightarrow \) translate it into small perturbation of structure distinctness
Weighting – structure distinctness

Theorem (Structure distinctness preservation)
In agreement with the previous notation and assumptions

\[ |\bar{\lambda}^Z - \bar{\lambda}^X| \leq \frac{1}{\sqrt{n}} \left( \frac{d}{\alpha} (\bar{\lambda}^X + \sqrt{k}) \right) + r_1 \left( \frac{1}{n} \right), \]

where \( r_1 \left( \frac{1}{n} \right) \) denotes a remainder of the first order of \( 1/n \).

Proof (Idea).
show smallness of weights’ variance \( \rightarrow \) translate it into small perturbation of structure distinctness
Weighting – dissimilarity between the subspaces (ssd)

\[
\text{ssd}(\text{PC}(k-1), S^*) = \frac{1}{k-1} \sum_{l=1}^{k-1} L^2(l, l),
\]

\(L\) a matrix of canonical correlations between \(\text{PC}(k-1)\) and \(S^*\).

**Figure:** Average canonical correlations for original and transformed data, \(d = 7\).
Dimension reduction algorithm

**Algorithm 2.1:** \texttt{DISTPRESERVINGDIMREDUCTION}(X)

**Step 1:** Isotropic transformation

\[ X_0 \leftarrow FX; \quad T_{X_0} \leftarrow A_{T_{X_0}} L_{T_{X_0}} A_{T_{X_0}}^T; \quad Y \leftarrow X_0 A_{T_{X_0}} L_{T_{X_0}}^{-\frac{1}{2}} \]

**Step 2:** Weighting

\[ \alpha \leftarrow 0.5; \quad \omega_i = \sqrt{\frac{1}{1 + \frac{1}{\alpha} \|y_i\|^2}}; \quad Z \leftarrow \text{diag}(\omega)Y; \quad Z_0 \leftarrow FZ \]

**Step 3:** Dimension reduction

\[ \frac{1}{n} T_{Z_0} \leftarrow A_{T_{Z_0}} G_{T_{Z_0}} A_{T_{Z_0}}^T; \quad R \leftarrow (A_{T_{Z_0}}^{(k-1)})^T Z_0^T \]

return \( R \)
Method’s performance - simulation example

<table>
<thead>
<tr>
<th></th>
<th>original</th>
<th>weighted</th>
</tr>
</thead>
<tbody>
<tr>
<td>dist</td>
<td>0.01</td>
<td>1.00</td>
</tr>
<tr>
<td>diss</td>
<td>0.48</td>
<td>0.51</td>
</tr>
</tbody>
</table>
Outline

1. Introduction
   - Model and notation
   - Basic definitions
   - Concept

2. The method
   - Isotropic transformation
   - Weighting
   - Algorithm

3. Summary
Summary

- Data transformation consists of two steps - isotropic transformation and weighting.
- It preserves distinctness of the original structure (defined in terms of variance in the Fisher’s subspace) with only negligible error.
- It brings principal component subspace $PC(k - 1)$ close to the Fisher’s subspace $S^*$ if sample is large enough.
- For transformed data $Z$, projection on $PC(k - 1)$ is similar to projection on $S^*$ but does not require the knowledge of classes.
- Facilitates further analysis of the unknown structure in the subspace of reduced dimension.
Summary

- Data transformation consists of two steps - isotropic transformation and weighting.
- It preserves distinctness of the original structure (defined in terms of variance in the Fisher’s subspace) with only negligible error.
- It brings principal component subspace $PC(k - 1)$ close to the Fisher’s subspace $S^*$ if sample is large enough.
- For transformed data $Z$, projection on $PC(k - 1)$ is similar to projection on $S^*$ but does not require the knowledge of classes.
- Facilitates further analysis of the unknown structure in the subspace of reduced dimension.
Summary

- Data transformation consists of two steps - isotropic transformation and weighting.

- It preserves distinctness of the original structure (defined in terms of variance in the Fisher’s subspace) with only negligible error.

- It brings principal component subspace $PC(k - 1)$ close to the Fisher’s subspace $S^*$ if sample is large enough.

- For transformed data $Z$, projection on $PC(k - 1)$ is similar to projection on $S^*$ but does not require the knowledge of classes.

- Facilitates further analysis of the unknown structure in the subspace of reduced dimension.
Data transformation consists of two steps - isotropic transformation and weighting. It preserves distinctness of the original structure (defined in terms of variance in the Fisher’s subspace) with only negligible error. It brings principal component subspace $PC(k - 1)$ close to the Fisher’s subspace $S^*$ if sample is large enough. For transformed data $Z$, projection on $PC(k - 1)$ is similar to projection on $S^*$ but does not require the knowledge of classes. Facilitates further analysis of the unknown structure in the subspace of reduced dimension.
Summary

- Data transformation consists of two steps - isotropic transformation and weighting.
- It preserves distinctness of the original structure (defined in terms of variance in the Fisher’s subspace) with only negligible error.
- It brings principal component subspace \( PC(k - 1) \) close to the Fisher’s subspace \( S^* \) if sample is large enough.
- For transformed data \( Z \), projection on \( PC(k - 1) \) is similar to projection on \( S^* \) but does not require the knowledge of classes.
- Facilitates further analysis of the unknown structure in the subspace of reduced dimension.
Thank you for your attention!

ewa.nowakowska@ipipan.waw.pl

Research funded by National Science Center of Poland DEC-2011/01/N/ST6/04174