A Robust Nonlinear Projection Method

Nicolas Donckers
John Aldo Lee
Michel Verleysen
Amaury Lendasse
Nonlinear dependencies: not PCA

Linear dependencies: PCA

HOW?

- Time-series prediction?
- Classification?
- Data mining?

WHY & HOW to project databases?
Which criterion for $M_?$

\[ \{ d \mathcal{R} \ni \frac{\partial}{\partial d} x \} = dX \]

\[ (d < u) \]

\[ \leftarrow \mathcal{W} \rightarrow \]

\[ \{ u \mathcal{R} \ni \frac{\partial}{\partial u} x \} = uX \]

Problem Statement

Projection

(unknown)

Mapping

Database
Drawbacks

• Only linear projection (on a hyperplane).
• Only one parameter to set (loss of total variance).
• Easily programmable & quickly computed;

Advantages

• Well known method (strong theoretical background);
• Linear projection with maximal variance.

Criterion

Linear Solution: PCA
1\textsuperscript{st} Nonlinear Solution: CCA

Criterion

\[
d_{i,j}^n = \|x_i^n - x_j^n\| \quad \leftarrow \text{preservation} \quad \rightarrow \quad d_{i,j}^p = \|x_i^p - x_j^p\|
\]

Error function

\[
E_{CCA} = \frac{1}{2} \sum_{i \neq j} (d_{i,j}^n - d_{i,j}^p)^2 F(d_{i,j}^p)
\]

\[
\downarrow \quad \text{minimization} \quad \text{(specific gradient descent)}
\]

\[
X^p = \{x_i^p \in \mathbb{R}^p\}
\]

Parameters

\[
\alpha(t) = \text{learning rate for gradient descent;}
\]

\[
\lambda(t) = \text{neighborhood factor for weighting function } F.
\]
Hard nonlinear database ⇐ slow convergence.

• Difficult to parameterize (neighborhood factor y(t));

• No guarantee on result (local minimum, slow convergence);

• Computationally demanding (partially solved by VQ);

Drawbacks

• Easily programmable.

• More powerful than Sammon's mapping and nonlinear MDS;

• Nonlinear capabilities;

Advantages

CCA: Advantages and Drawbacks
\[ \{ d \Xi \in \frac{d}{dx} \} = dX \]

Error function

Curvilinear distance \( d \) preservation.

Nonlinear Solution: CD\_2

Parameters
Curvilinear Distance

Theory

The distance between 2 vectors along an object.

Example

- a - two points in a spiral,
- b - Euclidian distance between them,
- c - curvilinear distance.
\[(f_{i_1}^G)_{i_2} (f_{i_1}^G - f_{i_1}^P) \triangleq \frac{Z}{i} = \mathbb{D} \mathbb{C} \mathbb{A} \mathbb{B} \]

\[\uparrow \]

\[
f_{i_1}^G(t) + f_{i_1}^P((t)\circ - 1) = f_{i_1}^G \mathbb{D}
\]

To keep fully generality, replace \(f_{i_1}^G \) by a generalized distance \( D \).

**Note**

1. Vector Quantization on the database \( X_u \).
2. Link the prototypes.
3. Shortest path between prototypes and \( x \).

**Curvilinear Distance**
• Complex implementation (VQ, links, shortest path, etc.).
• No guarantee on result (local minimum).
• Computationally demanding (partially solved by VQ).

Drawbacks

• Convergence less affected by hard nonlinear databases.
• Easy to parameterize.
• Very good nonlinear capabilities.

Advantages

CD4: Advantages and Drawbacks
Analyses of distortions between $d_{ij}$ and $g_{ij}$.

Analyses of VA results (prototypes, links, LPCA, etc.):

Specific Vector Quantization methods:

How?

MAKE CDFA AS EASY-TO-USE AS PCA.

$\uparrow$

Automation Feasible

$\uparrow$

Easy Parameter setting for CDFA

CDFA: Automatic Parameter Setting
Local dimension of structure $X^n$

\[ \uparrow \]

PCA on each Voronoi region

**Local PCA**

- connected components
- links per prototypes
- prototypes

... of ...

**CDA: Analyses of VG Results**
Distance distortions for a spiral.

CD4: Distance Distortions
Example 1: CD4 projection of a noisy trefoil knot (3d → 1d).
CD4 projection of the 'broken bicyle' (3D $\rightarrow$ 2D).

CD4 Example 2
CDA projection of a sphere (3D $\rightarrow$ 2D).

CDA: Example 3
CDA: Example 4

CDA projection of a ‘rolled red carpet’ ($3D \rightarrow 2D$).
CDAA projection of Santa Fe A time-series (10d - 30).

Regressor projection

Regressor length = 10

(Re)error > 5%

Bifurcation mapping

← CDAA →

Example 5
Possible automation of the whole CDA process.

Easy parameter setting for CDA.

Consequences

<table>
<thead>
<tr>
<th>Robustness</th>
<th>Simplicity converged of CCA/CDA error function</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Curvilinear distances less distorted than Euclidean ones</td>
</tr>
</tbody>
</table>

Key idea

Conclusion