Multi-Output Regularized Projection

Kai Yu[1], Shipeng Yu[12], Volker Tresp[1]

kai.yu@siemens.com
spyu@dbs.informatik.uni-muenchen.de
volker.tresp@siemens.com

SIEMENS

[1]Siemens Corporate Technology
Department of Neural Computation

Ludwig Maximilians Universität München LMU

[2]University of Munich
Institute for Informatics

Outline

- Motivation: **A supervised algorithm for dimensionality reduction**
- PCA: Unsupervised Projection
- Multi-Output Regularized Projection (MORP)
  - Idea: **Minimize reconstruction errors of both input and output**
  - Primal form: Linear mappings
  - Dual form: Non-linear mappings
  - How to choose between primal and dual form
- Connections to Related Work
- Experimental Results
  - Visualization, User preference prediction, multi-label classification
- Conclusion

Motivation

We are dealing with high-dimensional data in pattern recognition.

- What are the problems?
  - Noisy dimensions: Only a small number of dimensions suffice
  - Learnability: "curse of dimensionality"
  - Inefficiency: Computational cost is too high
- How to solve these problems? Dimensionality Reduction
  - Feature selection: Select part of the dimensions
  - Feature transformation/projection: Learn a mapping that maps from the high-dimensional input space into a low-dimensional latent space
- Some notations: We have N documents
  - Document i is denoted as $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^M$, with output $\mathbf{y}_i \in \mathcal{Y} \subset \mathbb{R}^L$
  - $\mathbf{X} \in \mathbb{R}^{N \times M} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T$, $\mathbf{Y} \in \mathbb{R}^{N \times L} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^T$
  - We aim to derive a mapping $\Psi : \mathcal{X} \mapsto \mathcal{V}$ such that $\mathcal{V} \subset \mathbb{R}^K$, $K < M$

Principal Component Analysis

A well-known unsupervised feature transformation method.

- Some formulations
  - Mapping directions with largest data covariance
  - Best rank K approximation to the data matrix $\mathbf{X}$
- An optimization problem to minimize the reconstruction error:

$$\begin{aligned} \min_{\mathbf{A},\mathbf{V}} \quad & \|\mathbf{X} - \mathbf{V}\mathbf{A}\|_F^2 \\ \text{s.t.} \quad & \mathbf{V}^\top \mathbf{V} = \mathbf{I}, \end{aligned}$$

  with $\mathbf{V} \in \mathbb{R}^{N \times K}$ the latent semantics, and $\mathbf{A} \in \mathbb{R}^{K \times M}$ the factor loadings.
- Drawbacks of PCA:
  - PCA is unsupervised and may not be beneficial to supervised learning
  - No inter-correlation between $\mathbf{X}$ and $\mathbf{Y}$ is considered in the mapping
  - No intra-correlation between dimensions of $\mathbf{Y}$ (if multiple outputs) is considered in the mapping

MORP

The optimization problem solved by MORP (with $0 \leq \beta \leq 1$):

$$\begin{aligned} \min_{\mathbf{A},\mathbf{B},\mathbf{V}} \quad & (1-\beta)\|\mathbf{X}-\mathbf{V}\mathbf{A}\|_F^2 + \beta\|\mathbf{Y}-\mathbf{V}\mathbf{B}\|_F^2 \\ \text{s.t.} \quad & \mathbf{V}^\top\mathbf{V} = \mathbf{I}, \mathbf{V} = \mathbf{X}\mathbf{W}. \end{aligned}$$

- We are minimizing the reconstruction errors of both $\mathbf{X}$ and $\mathbf{Y}$
- We are constraining the mappings to be linear in $\mathbf{X}$

Denote $\mathbf{K} = (1-\beta)\mathbf{X}\mathbf{X}^\top + \beta\mathbf{Y}\mathbf{Y}^\top$. Let $[\mathbf{v}_1, \ldots, \mathbf{v}_N]$ be its eigenvectors with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_N$. We obtain at the optimum,

- $\mathbf{A} = \mathbf{V}^\top\mathbf{X}$, $\mathbf{B} = \mathbf{V}^\top\mathbf{Y}$;
- $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_K]\mathbf{R}$ where $\mathbf{R}$ is an arbitrary $K \times K$ orthogonal matrix;
- The optimum of the cost function is $\sum_{i=K+1}^{N} \lambda_i$;
- Denote $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$, each $\mathbf{w}$ solves the optimization problem:

$$\max_{\mathbf{w}\in\mathbb{R}^M} \quad \mathbf{w}^\top\mathbf{X}^\top\mathbf{K}\mathbf{X}\mathbf{w} \qquad \text{s.t.} \quad \mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} = 1.$$

MORP: Primal Form

The optimization problem for $\mathbf{w}$ is ill-posed when $\mathrm{rank}(\mathbf{X}) < M$.

One way to deal with this problem is to introduce Tickhonov regularizer:

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^M} \quad & \mathbf{w}^\top \mathbf{X}^\top \mathbf{K}^{-1} \mathbf{X}\mathbf{w} + \gamma \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} = 1. \end{aligned}$$

We summarize the primal form of the MORP solution:

- Calculate $\mathbf{K} = (1-\beta)\mathbf{X}\mathbf{X}^\top + \beta \mathbf{Y}\mathbf{Y}^\top$;
- Solve a generalized eigenvalue problem

$$[\mathbf{X}^\top \mathbf{K}^{-1} \mathbf{X} + \gamma \mathbf{I}]\mathbf{w} = \lambda \mathbf{X}^\top \mathbf{X}\mathbf{w},$$

  obtain generalized eigenvectors $\mathbf{w}_1, \ldots, \mathbf{w}_K$ with smallest K eigenvalues $\lambda_1 \leq \ldots \leq \lambda_K$;
- Form mapping functions $\psi_j(\mathbf{x}) = \sqrt{\lambda_j}\mathbf{w}_j^\top \mathbf{x}, j = 1, \ldots, K$, and finally $\Psi(\mathbf{x}) = [\psi_1(\mathbf{x}), \ldots, \psi_K(\mathbf{x})]^\top$ defines the mapping Ψ.

MORP: Dual Form

Non-linear mappings are obtained by applying representer theorem and defining dual variable $\boldsymbol{\alpha}$ as

$$\mathbf{w} = \mathbf{X}^\top \boldsymbol{\alpha}.$$

We summarize the dual form of the MORP solution:

- Calculate $\mathbf{K}_x$, $\mathbf{K}_y$ using kernel functions κ_x, κ_y, and $\mathbf{K} = (1-\beta)\mathbf{K}_x + \beta\mathbf{K}_y$;
- Solve a generalized eigenvalue problem

$$[\mathbf{K}_x\mathbf{K}^{-1}\mathbf{K}_x + \gamma\mathbf{K}_x]\boldsymbol{\alpha} = \lambda\mathbf{K}_x^2\boldsymbol{\alpha},$$

  obtain generalized eigenvectors $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K$ with smallest K eigenvalues $\lambda_1 \leq \ldots \leq \lambda_K$;
- Form mapping functions $\psi_j(\mathbf{x}) = \sqrt{\lambda_j}\sum_{i=1}^{N}(\boldsymbol{\alpha}_j)_i\kappa_x(\mathbf{x}_i, \mathbf{x})$, and finally $\Psi(\mathbf{x}) = [\psi_1(\mathbf{x}), \ldots, \psi_K(\mathbf{x})]^\top$ defines the mapping Ψ.

Discussion

Which form to choose in real world applications?

- Primal MORP solves an $M \times M$ generalized eigenvalue problem
  - is more efficient when $M < N$ and only learns a linear mapping for $\mathbf{X}$
- Dual MORP solves an $N \times N$ generalized eigenvalue problem
  - is more efficient when $N < M$ for linear mappings
  - can learn non-linear mappings with carefully chosen kernel function κ_x

Two extreme cases of MORP:

- When $\beta = 0$, MORP is identical to PCA (primal) and kernel PCA (dual)
- When $\beta = 1$, MORP shows similar spirit with kernel dependency estimation (KDE), but is better since MORP has one unified optimization framework

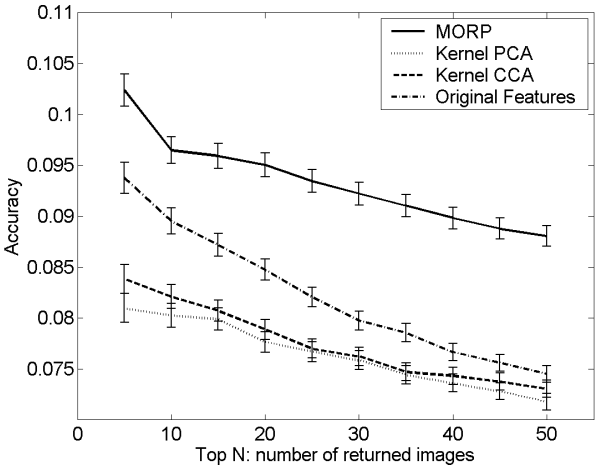
Other supervised projection methods

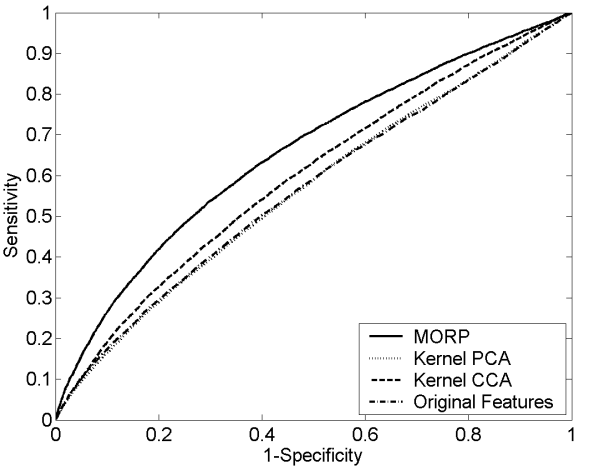
- FDA: only focuses on single output with binary classification
- CCA: minimizes inter-correlation but ignores self-correlations
- PLS: is a penalized CCA and focuses on the regression of known outputs

Experiment 1: User Preference Prediction

The Goal: Evaluate projection methods with prediction performance.

We extract 642 paintings from 47 artists and collect 190 user preference data from an online survey. We select some training users and make predictions for test users based on low-level image features and ratings of other users. For projection methods, a linear SVM classifier is trained on the 50-dimensional latent space.





MORP is consistently better than other methods

Experiment 2: Multi-label Classification

The Goal: Evaluate projection methods in terms of classification.

Data set: 1021 images from Corel, with 491 features and 37 categories.

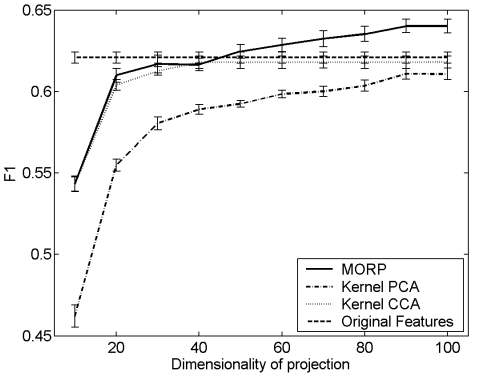
We manually labeled the data and it has a multi-label setting, i.e., each document can belong to multiple categories.

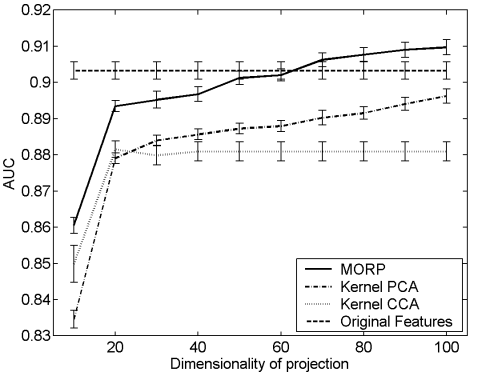
We test the following two settings:

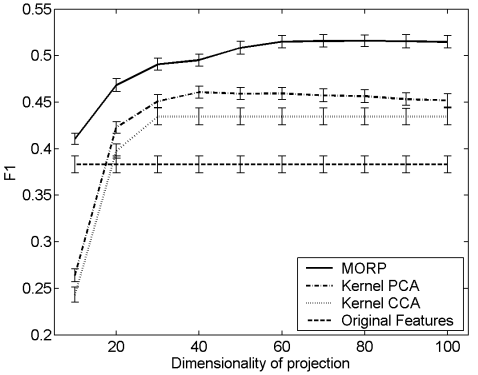
- Setting (A): We pick up 70% categories for classification and employ 5-fold cross-validation with one fold training and 4 folds testing
- Setting (B): Evaluate the classification performance on the rest 30% categories for previously unseen data with newly derived features

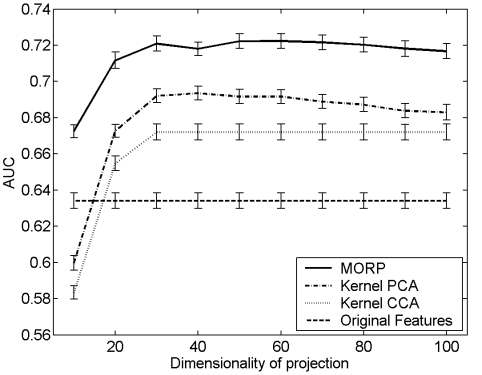
For projection methods, linear SVMs are trained on the (non-linearly) projected feature space. For "Original Features" an SVM with RBF kernel is trained.

Results: (setting A: top; setting B: bottom)









- MORP achieves the best performance
- CCA can only obtain effective dimensions less than the number of categories
- Only MORP can obtain significantly better performance than Original Features

Conclusion

MORP has the following advantages:

- It is supervised and takes PCA as a special case (when $\beta = 0$)
- It considers both the inter-correlation between $\mathbf{X}$ and $\mathbf{Y}$, and the intra-correlation of $\mathbf{Y}$
- Both linear and non-linear mappings are easy to derive
- It handles multiple outputs simultaneously

Experimental results are very encouraging.