Semi-Supervised Laplacian Regularization of Kernel Canonical Correlation Analysis

Matthew B. Blaschko, Christoph H. Lampert, & Arthur Gretton



Max Planck Institute for Biological Cybernetics
Tübingen, Germany

ECML KPDD, Antwerp, Belgium, September 16, 2008

MAX-PLANCK-GESELLSCHAFT

BIOLOGISCHE KYBERNETIK

Overview

- **Paired Data**
- Kernel Canonical Correlation Analysis
- Semi-Supervised Laplacian Regularization
- Model Selection
- Results
- Summary and Outlook

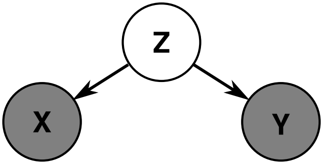
Paired Datasets

Realistic data comes in many **different modalities**: text, images...

We call data **paired**, if the samples come in more than one such representation at the same time, e.g.

- images + captions,
- audio + transscript,
- multi-language text documents
- MRI + CT scans

We assume a **latent aspect** that relates the representations.



Paired Datasets

A **fully paired dataset** has correspondences for all samples:

$$
\begin{array}{rcccccl}
X = \{ & x_1, & x_2, & \ldots, & x_n & \}, \\
 & \updownarrow & \updownarrow & & \updownarrow & \\
Y = \{ & y_1, & y_2, & \ldots, & y_n & \}.
\end{array}
$$

"Miss Summers?"

"Good call."

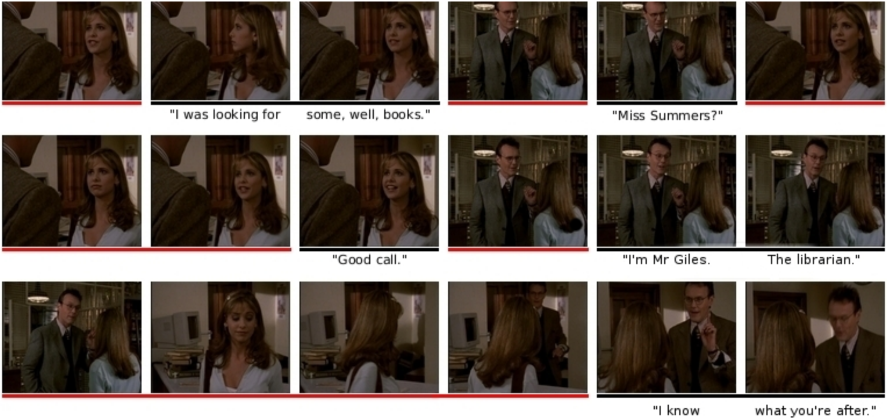
I'm Mr Giles

The librarian.

Paired Datasets

However, data is often only **partially paired**:

$$\begin{array}{cccccccccc} \hat{X} = \{ & x_1, & x_2, & \dots, & x_n, & x_{n+1}, & \dots, & x_{n+p_x} & \}, \\ & \updownarrow & \updownarrow & & \updownarrow & ? & & ? & \\ \hat{Y} = \{ & y_1, & y_2, & \dots, & y_n, & y_{n+1}, & \dots, & y_{n+p_y} & \}. \end{array}$$



Overview

- Paired Data
- **Kernel Canonical Correlation Analysis**
- Semi-Supervised Laplacian Regularization
- Model Selection
- Results
- Summary and Outlook

Review of Canonical Correlation Analysis

Principal Component Analysis (PCA)

- Single dataset $x_1, \ldots, x_n$.
- Find projections that maximize the **variance** of the projected data.
- simple eigenvalue problem.

Canonical Correlation Analysis (CCA)

- Fully paired data $x_1 \leftrightarrow y_1, \ldots, x_n \leftrightarrow y_n$.
- Find projection directions w_x and w_y that maximize the **correlation** between the projected data.

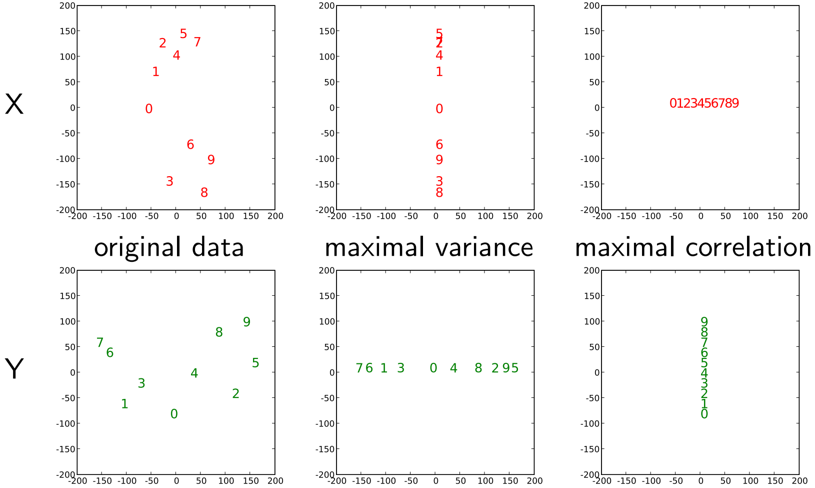
$$\max_{w_x, w_y} \frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T C_{xx} w_x \; w_y^T C_{yy} w_y}}$$

$C_{xy}/C_{xx}/C_{yy}$:
(cross) correlation matrices

- generalized eigenvalue problem.
- supervised situation: $y \in \{-1, 1\}$ ground truth labels $\rightarrow$ CCA $\equiv$ LDA!

Why is Correlation better than Variance?

- Toy dataset: 1 signal direction, 1 (high variance) noise direction.



- **CCA** can ignore noise that is uncorrelated between X and Y.

Kernelization

Kernel Canonical Correlation Analysis

- Kernelize CCA, to use it for arbitrary input domains, and (latent) data embeddings $\phi_x : \mathcal{X} \to \mathcal{H}_X$, $\phi_y : \mathcal{Y} \to \mathcal{H}_Y$.
- $k_x(x_i, x_j) = \langle \phi_x(x_i), \phi_x(x_j) \rangle, \quad k_y(y_i, y_j) = \langle \phi_y(y_i), \phi_y(y_j) \rangle$
- $w_x = \sum_i \alpha_i \phi_x(x_i), \quad w_y = \sum_i \beta_i \phi_y(y_i)$
- KCCA $\equiv$ CCA in $\mathcal{H}_X / \mathcal{H}_Y$: Solve

$$\max_{\alpha,\beta} \frac{\alpha^T K_x K_y \beta}{\sqrt{\alpha^T K_x^2 \alpha \; \beta^T K_y^2 \beta}} \qquad K_x, K_y\text{: kernel matrices of } X, Y,$$

equivalent to the constrained optimization problem

$$\max_{\alpha,\beta} \quad \alpha^T K_x K_y \beta \qquad \text{sb.t.} \quad \alpha^T K_x^2 \alpha = 1 \quad \text{and} \quad \beta^T K_y^2 \beta = 1.$$

Need for Regularization

- Problem: Maximization is **degenerate** for invertible K_x or K_y.

$$\max_{\alpha,\beta} \quad \alpha^T K_x K_y \beta \quad \text{sb.t.} \quad \alpha^T K_x^2 \alpha = 1 \quad \text{and} \quad \beta^T K_y^2 \beta = 1$$

  allows α, β with perfect correlation but without learning anything!

- We need to **regularize**!

Need for Regularization

- Problem: Maximization is **degenerate** for invertible K_x or K_y.

$$\max_{\alpha,\beta} \quad \alpha^T K_x K_y \beta \quad \text{sb.t.} \quad \alpha^T K_x^2 \alpha = 1 \quad \text{and} \quad \beta^T K_y^2 \beta = 1$$

allows α, β with perfect correlation but without learning anything!

- **Tikhonov regularization:**

$$\max_{w_x, w_y} \frac{w_x^T C_{xy} w_y}{\sqrt{\left(w_x^T C_{xx} w_x + \varepsilon_x \|w_x\|^2\right)\left(w_y^T C_{yy} w_y + \varepsilon_y \|w_y\|^2\right)}}$$

$$= \max_{\alpha,\beta} \frac{\alpha^T K_x K_y \beta}{\sqrt{\alpha^T \left(K_x^2 + \varepsilon_x K_x\right) \alpha \; \beta^T \left(K_y^2 + \varepsilon_y K_y\right) \beta}}$$

- Regularization parameters ε_x, ε_y need to be model selected.

Overview

- Paired Data
- Kernel Canonical Correlation Analysis
- **Semi-Supervised Laplacian Regularization**
- Model Selection
- Results
- Summary and Outlook

Proposed Laplacian Regularization

Manifold assumption:

- The data lie on a lower-dimensional manifold $\mathcal{M} \subset \mathcal{H}$.

- Use the Laplace operator $\Delta_{\mathcal{M}}$ to measure smoothness along $\mathcal{M}$.

Laplacian regularization:

- Approximate $\Delta_{\mathcal{M}}$ by Graph Laplacian $\mathcal{L} = D^{-1/2}(D - W)D^{-1/2}$ (W similarity matrix, D diagonal matrix with $D_{ii} = \sum_j W_{ij}$).

$$\begin{aligned}
\max_{\alpha,\beta} \quad & \alpha^T K_x K_y \beta \\
\text{sb.t.} \quad & \alpha^T \Big(K_x K_x + \varepsilon_x K_x + \gamma_x K_x \mathcal{L}_x K_x \Big) \alpha = 1, \\
& \beta^T \Big(K_y K_y + \underbrace{\varepsilon_y K_y}_{\text{Tikhonov}} + \underbrace{\gamma_y K_y \mathcal{L}_y K_y}_{\text{Laplacian}} \Big) \beta = 1
\end{aligned}$$

- Favour projections that vary **smoothly** wrt. the manifold structure.

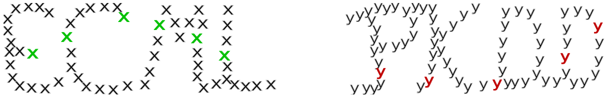
Making use of unpaired data:

- We don't need correspondences to compute Laplacian.
- More data allows a better estimate of the manifold structure.

Partially paired data revisited

Making use of unpaired data:

- We don't need correspondences to compute Laplacian.
- More data allows a better estimate of the manifold structure.



Notation:

- Paired training data $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$,
- Additional unpaired data $\{x_{n+1}, \ldots, x_{n+p_x}\}$ and $\{y_{n+1}, \ldots, y_{n+p_y}\}$.

- Data matrix $X = (x_1, \ldots, x_n)^T \quad \in \mathcal{H}_X^n$,
- Extended data matrix $\hat{X} = (x_1, \ldots, x_{n+p_x})^T \quad \in \mathcal{H}_X^{n+p_x}$,
- Kernel matrix $K_{xx} = XX^T \quad \in \mathbb{R}^{n \times n}$,
- Extended kernel matrices $K_{\hat{x}x} = \hat{X}X^T \quad \in \mathbb{R}^{(n+p_x)\times n}$,
- $K_{\hat{x}\hat{x}} = \hat{X}\hat{X}^T \quad \in \mathbb{R}^{(n+p_x)\times(n+p_x)}$,
- etc.

Semi-Supervised Laplacian Regularization

- **Semi-Supervised KCCA with Laplacian Regularization**:

$$\max_{\alpha,\beta} \quad \alpha^T K_{\hat{x}x} K_{y\hat{y}} \beta$$

$$\text{sb.t.} \quad \alpha^T \Big(K_{\hat{x}x} K_{x\hat{x}} + \varepsilon_x K_{\hat{x}\hat{x}} + \frac{\gamma_x}{(n+p_x)^2} K_{\hat{x}\hat{x}} \mathcal{L}_{\hat{x}} K_{\hat{x}\hat{x}} \Big) \alpha = 1$$

$$\beta^T \Big(K_{\hat{y}y} K_{y\hat{y}} + \underbrace{\varepsilon_y K_{\hat{y}\hat{y}}}_{\text{Tikhonov}} + \underbrace{\frac{\gamma_y}{(n+p_y)^2} K_{\hat{y}\hat{y}} \mathcal{L}_{\hat{y}} K_{\hat{y}\hat{y}}}_{\text{“semi-supervised” Laplacian}} \Big) \beta = 1$$

- Finds data projections that
  - achieve high correlation when applied to $X \leftrightarrow Y$
  - vary smoothly on manifolds defined by $\hat{X}, \hat{Y}$.

- Four regularization parameters: $\varepsilon_x, \varepsilon_y, \gamma_x, \gamma_y \quad \rightarrow$ model selection

Overview

- Paired Data
- Kernel Canonical Correlation Analysis
- Semi-Supervised Laplacian Regularization
- **Model Selection**
- Results
- Summary and Outlook

Model Selection

- Supervised szenario: cross-validation or similar.
- Unsupervised szenario: **dependence maximization HSNIC**
- HSNIC: kernel measure of dependence between random variables (Fukumizu, Bach, & Gretton 2007)
  - Hilbert-Schmidt norm of normalized cross-covariance operator $\mathcal{V}_{xy}$

$$\mathrm{HSNIC}(\mathrm{x}, \mathrm{y}) \equiv \|\mathcal{V}_{xy}\|_{HS}^2$$

  - Closely related to KCCA
- Regularization with Laplace-Beltrami operators on data manifolds

$$\mathcal{V}_{xy} = (\Sigma_{xx} + \underbrace{\varepsilon_x I}_{\text{Tikhonov}} + \underbrace{\gamma_x \Delta_{\mathcal{M}_x}}_{\text{Laplacian}})^{-\frac{1}{2}} \underbrace{\Sigma_{xy}}_{\substack{\text{covariance} \\ \text{operator}}} \left(\Sigma_{yy} + \varepsilon_y I + \gamma_y \Delta_{\mathcal{M}_y}\right)^{-\frac{1}{2}}$$

- Finite sample set: **closed form expression** $\hat{V}_{xy}$ (see paper or poster)

Model Selection: Concluded

- $\|\hat{V}_{xy}\|^2_{HS}$ estimates correlation achievable by KCCA.
- But beware: overfitting! trivial maximum at $\varepsilon_x = \varepsilon_y = \gamma_x = \gamma_y = 0$.
- Better model selection criterion:
  Maximize **increase in correlation only due to the data pairing**

$$\rho(\varepsilon_x, \varepsilon_y, \gamma_x, \gamma_y) = \frac{\|\hat{V}_{xy}\|^2_{HS}}{\|\hat{V}_{x\Pi(y)}\|^2_{HS}}$$

with $\Pi(y)$ randomly reshuffled version(s) of y.

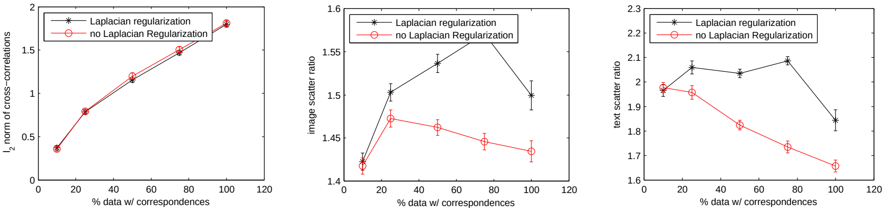
Overview

- Paired Data
- Kernel Canonical Correlation Analysis
- Semi-Supervised Laplacian Regularization
- Model Selection
- **Results**
- Summary and Outlook

Example Results

- Four datasets of *images + associated text*
  - Multiple *train / test* splits
  - Vary percentage of correspondences
- Example results: **UIUC-ISD Bass**, 6 semantic classes
  - Perform (semi-supervised) KCCA on training set
  - Measure correlations on test set
  - Measure scatter ratio on test set (larger is better)

Cross correlation | Image scatter ratio | Text scatter ratio



- More results in paper.

Summary & Future Work

Summary:

- Laplacian regularization for KCCA
- Allows straight-forward semi-supervised extension
- Model selection: closed form expression for Laplacian regularized kernel independence measure (HSNIC)
- Experimental results: projections better respect latent aspect

Future work:

- Improved model selection criterion
  - ratio used somewhat heuristic
- Other application of Laplacian regularized HSNIC
  - Causality inference
  - ICA

Summary & Future Work

Summary:

- Laplacian regularization for KCCA
- Allows straight-forward semi-supervised extension
- Model selection: closed form expression for Laplacian regularized kernel independence measure (HSNIC)
- Experimental results: projections better respect latent aspect

Future work:

- Improved model selection criterion
  - ratio used somewhat heuristic
- Other application of Laplacian regularized HSNIC
  - Causality inference
  - ICA

Thank you.

More about $\mathcal{V}_{xy}$

- Covariance operator: $\Sigma_{xy} : \mathcal{H}_Y \to \mathcal{H}_X$ defined by

$$\langle f, \Sigma_{xy} g \rangle_{H_X} = \mathbf{E}_{\mathrm{x,y}}[f(\mathrm{x})g(\mathrm{y})] - \mathbf{E}_{\mathrm{x}}[f(\mathrm{x})]\mathbf{E}_{\mathrm{y}}[g(\mathrm{y})]$$

for all $f \in \mathcal{H}_X$, $g \in \mathcal{H}_g$.

More about $\mathcal{V}_{xy}$

- Covariance operator: $\Sigma_{xy} : \mathcal{H}_Y \rightarrow \mathcal{H}_X$ defined by

$$\langle f, \Sigma_{xy} g \rangle_{H_X} = \mathbf{E}_{\mathrm{x,y}}[f(\mathrm{x})g(\mathrm{y})] - \mathbf{E}_{\mathrm{x}}[f(\mathrm{x})]\mathbf{E}_{\mathrm{y}}[g(\mathrm{y})]$$

  for all $f \in \mathcal{H}_X$, $g \in \mathcal{H}_g$.

- Normalized Cross-covariance operator, V_{xy}, such that

$$\Sigma_{xy} = \Sigma_{xx}^{\frac{1}{2}} V_{xy} \Sigma_{yy}^{\frac{1}{2}}$$

Model Selection: Laplacian Regularized HSNIC

- Empirical estimate of $\mathcal{V}_{xy}$:

$$\hat{V}_{xy} = \left(\frac{1}{n}X^TX + \varepsilon_x I + \frac{\gamma_x}{m_x^2}\hat{X}^T\mathcal{L}_{\hat{x}}\hat{X}\right)^{-\frac{1}{2}}\frac{1}{n}X^TY \cdot \left(\frac{1}{n}Y^TY + \varepsilon_y I + \frac{\gamma_y}{m_y^2}\hat{Y}^T\mathcal{L}_{\hat{y}}\hat{Y}\right)^{-\frac{1}{2}}$$

- Closed form solution

$$\|\hat{V}_{xy}\|_{HS}^2 = \mathrm{Tr}\left[\hat{V}_{xy}\hat{V}_{xy}^T\right] = \mathrm{Tr}\left[M_x M_y\right]$$

with

$$M_x = I - n\left(nI + \frac{1}{\varepsilon_x}K_{xx} - \frac{1}{\varepsilon_x}K_{x\hat{x}}\left(\frac{m_x^2\varepsilon_x}{\gamma_x}I + \mathcal{L}_{\hat{x}}K_{\hat{x}\hat{x}}\right)^{-1}\mathcal{L}_{\hat{x}}K_{\hat{x}x}\right)^{-1}$$