Semi-Supervised Laplacian Regularization of KCCA

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

Max Planck Institute for Biological Cybernetics, Tübingen, Germany

MAX-PLANCK-GESELLSCHAFT

BIOLOGISCHE KYBERNETIK

Why Semi-Supervised KCCA?

- **Data in two or more modalities,**
  - **e.g. captioned images, subtitled video, web documents with links, multi-language documents**
- **If complete correspondences between modalities known, we can use (K)CCA finds common latent factors.**
- **What if we have additional data without or with unknown correspondences?**

$$X = \{\ x_1,\ x_2,\ \ldots,\ x_n\ ,\ x_{n+1},\ \ldots,\ x_{n+p_x}\ \}$$
$$\updownarrow\quad \updownarrow\qquad \updownarrow\qquad ?\qquad\quad ?$$
$$Y = \{\ y_1,\ y_2,\ \ldots,\ y_n\ ,\ y_{n+1},\ \ldots,\ y_{n+p_y}\ \}.$$

- **Usual "solution": ignore $x_{n+1},\ldots,x_{n+p_x}, y_{n+1},\ldots,y_{n+p_y}$.**
- **Proposed: improve KCCA regularization by unpaired data**

1 (Kernel) Canonical Correlation Analysis [1]

- Given a fully paired dataset $x_1 \leftrightarrow y_1, \ldots, x_n \leftrightarrow y_n$ in $\mathcal{X}\times\mathcal{Y}$.
- Find projection directions w_x and w_y that maximize the *correlation* between the projected data, *i.e.* solve

$$\max_{w_x,w_y} \frac{\hat{E}\left[\langle x,w_x\rangle\langle y,w_y\rangle\right]}{\sqrt{\hat{E}\left[\langle x,w_x\rangle^2\right]\hat{E}\left[\langle y,w_y\rangle^2\right]}} \;\hat{=}\; \max_{w_x,w_y}\frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T C_{xx} w_x\ w_y^T C_{yy} w_y}}. \quad (1)$$

C_{xx}/C_{yy}: data covariance matrices, C_{xy}: cross-covariance matrix

- Kernelization: Apply CCA in latent Hilbert Spaces $\mathcal{H}_X$, $\mathcal{H}_Y$. Solve

$$\max_{\alpha,\beta}\frac{\alpha^T K_x K_y \beta}{\sqrt{\alpha^T K_x^2 \alpha\ \beta^T K_y^2 \beta}} \quad (2)$$

K_x/K_y: kernel matrices of X/Y, α,β: projection coefficients.

1.1 Need for Regularization

- Problem: for invertible K_x, K_y, Eq. (2) is *degenerate*. There exist α,β with perfect correlation, but non-informative.
- Use **Tikhonov regularization** to enforce *smooth* projections:

$$\max_{\alpha,\beta}\frac{\alpha^T K_x K_y \beta}{\sqrt{\alpha^T\left(K_x^2+\varepsilon_x K_x\right)\alpha\beta^T\left(K_y^2+\varepsilon_y K_y\right)\beta}}. \quad (3)$$

$\varepsilon_x, \varepsilon_y$: regularization parameters

2 Semi-Supervised Laplacian Regularization

- Paired data: $(x_1,y_1),\ldots,(x_n,y_n)\in\mathcal{H}_X\times\mathcal{H}_Y$.
- Additional unpaired data: $x_{n+1},\ldots,x_{n+p_x},\ y_{n+1},\ldots,y_{n+p_y}$.
- Data matrices: $X=(x_1,\ldots,x_n)^T$, $Y=(y_1,\ldots,y_n)^T$, $\hat{X}=(x_1,\ldots,x_{n+p_x})^T$, $\hat{Y}=\left(y_1,\ldots,y_{n+p_y}\right)^T$.
- Kernel matrices $K_{xx}=XX^T\in\mathbb{R}^{n\times n}$, $K_{yy}=YY^T\in\mathbb{R}^{n\times n}$. $K_{\hat{x}x}=\hat{X}X^T\in\mathbb{R}^{(n+p_x)\times n}$, etc.
- Graph Laplacians [2]: $\mathcal{L}_{\hat{x}} = I - D_{\hat{x}\hat{x}}^{-1/2}K_{\hat{x}\hat{x}}D_{\hat{x}\hat{x}}^{-1/2}$ for diagonal $(D_{\hat{x}\hat{x}})_{ii}=\sum_{j=1}^{n+p_x}(K_{\hat{x}\hat{x}})_{ij}$,

Semi-Supervised KCCA

- **Solve (e.g. as generalized eigenproblem):**

$$\max_{\alpha,\beta}\ \alpha^T K_{\hat{x}x}K_{y\hat{y}}\beta \quad (4)$$

$$\textbf{sb.t.}\quad \alpha^T\Big(K_{\hat{x}x}K_{x\hat{x}}+\varepsilon_x K_{\hat{x}\hat{x}}+\frac{\gamma_x}{m_x^2}K_{\hat{x}\hat{x}}\mathcal{L}_{\hat{x}}K_{\hat{x}\hat{x}}\Big)\alpha = 1, \quad (5)$$

$$\beta^T\Big(K_{\hat{y}y}K_{y\hat{y}}+\underbrace{\varepsilon_y K_{\hat{y}\hat{y}}}_{\textbf{Tikhonov}}+\underbrace{\frac{\gamma_y}{m_y^2}K_{\hat{y}\hat{y}}\mathcal{L}_{\hat{y}}K_{\hat{y}\hat{y}}}_{\textbf{Laplacian}}\Big)\beta = 1. \quad (6)$$

- **Finds projections that are smooth with respect to manifold structures of $\hat{X},\hat{Y}$ instead of ambient spaces $\mathcal{H}_X,\mathcal{H}_Y$.**

3 Model Selection

- We need to choose *regularization parameters* $\varepsilon_x,\varepsilon_y,\gamma_x,\gamma_y$.
- No external ground truth: use **dependence maximization**.

3.1 Laplacian Regularized HSNIC

- HSNIC measures dependence between random variables [3].

$$\mathrm{HSNIC}(X,Y)=\|V_{xy}\|_{HS}^2. \quad (7)$$

- V_{xy}: is normalized and regularized version of cross-covariance operator $\Sigma_{xy}:\mathcal{H}_y\to\mathcal{H}_x$:

$$V_{xy}=\big(\Sigma_{xx}+\underbrace{\varepsilon_x I}_{\text{Tikhonov}}+\underbrace{\gamma_x\Delta_{\mathcal{M}_x}}_{\text{Laplacian}}\big)^{-\frac{1}{2}}\Sigma_{xy}\big(\Sigma_{yy}+\varepsilon_y I+\gamma_y\Delta_{\mathcal{M}_y}\big)^{-\frac{1}{2}} \quad (8)$$

$$\langle f,\Sigma_{xy}g\rangle_{\mathcal{H}_x}=\mathbf{E}_{x,y}[x,y]-\mathbf{E}_x[x]\mathbf{E}_y[y]. \quad (9)$$

- We can estimate HSNIC only from kernel matrices *in closed form*.

3.2 Model Selection Criterion

- Idea: Maximize dependence that is due to *data pairing*.

$$\rho_{pair}(X,Y;\varepsilon_x,\gamma_x,\varepsilon_y,\gamma_y)=\widehat{\mathrm{HSNIC}}(\ X,Y\ ) \quad (10)$$
$$\rho_{rand}(X,Y;\varepsilon_x,\gamma_x,\varepsilon_y,\gamma_y)=\widehat{\mathrm{HSNIC}}(\ X,\Pi(Y)\ ) \quad (11)$$

for random permutation Π of samples in Y.

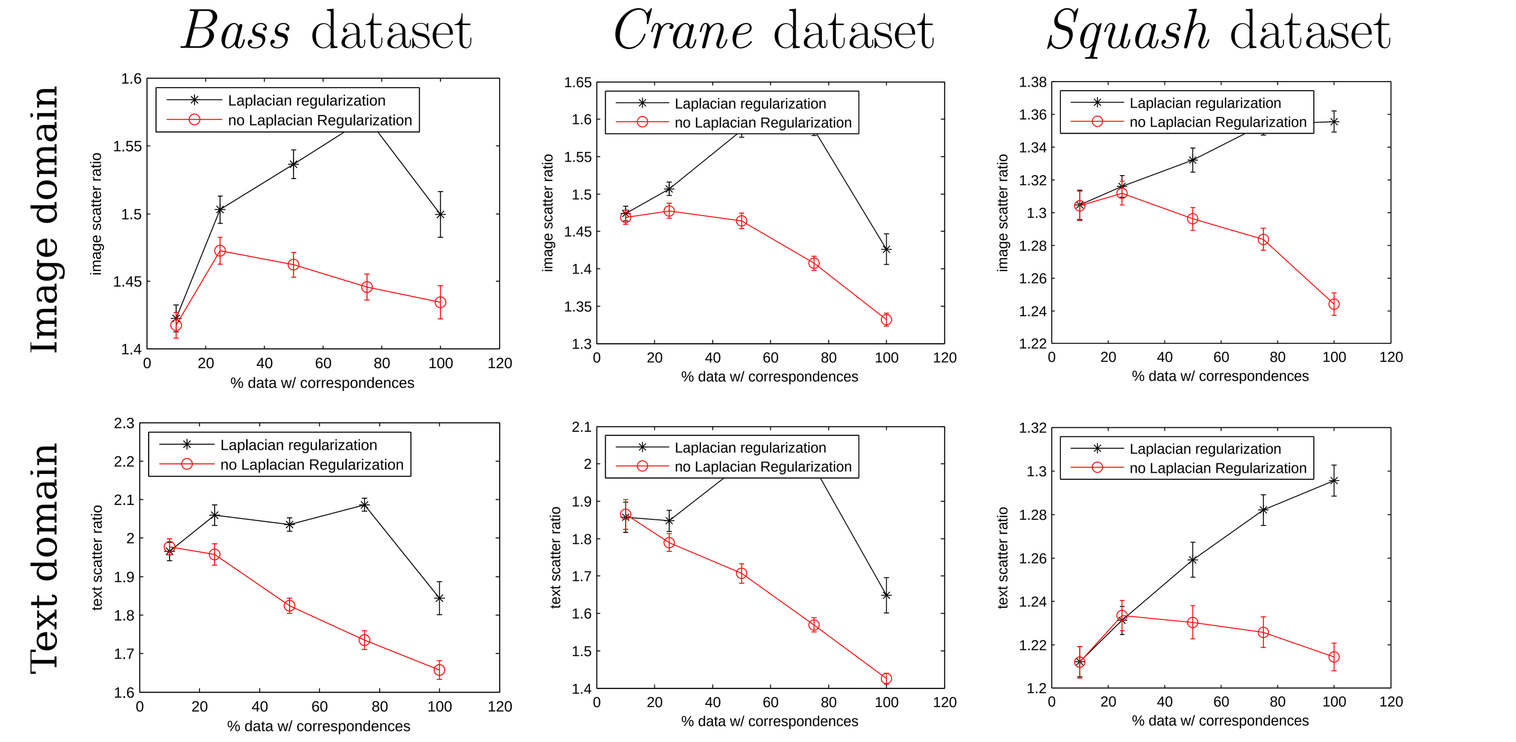
- Choose parameters $(\varepsilon_x,\gamma_x,\varepsilon_y,\gamma_y)$ that maximize

$$\rho(\varepsilon_x,\gamma_x,\varepsilon_y,\gamma_y)=\frac{\rho_{pair}(X,Y;\varepsilon_x,\gamma_x,\varepsilon_y,\gamma_y)}{\rho_{rand}(X,Y;\varepsilon_x,\gamma_x,\varepsilon_y,\gamma_y)}. \quad (12)$$

4 Experimental Results

- Datasets of images with text captions:
  - bag-of-visual-word/bag-of-word representation, χ^2-kernels
  - samples belong to different semantic classes (ground truth)
- Idea: good projection directions should retain the latent class aspect.
- Evaluation procedure:
  - Multiple splits into 50% training, 50% test
  - (Semisupervised) KCCA on training set (without class labels)
  - Measure performance on test set in terms of *scatter ratios* $\frac{|S_t|}{|\sum_{i=1}^{c}S_{C_i}|}$:
    - $S_t=\sum_{j\in test}(z_j-\mu)(z_j-\mu)^T$: data radius after projection
    - $S_{C_i}=\sum_{j\in C_i}(z_j-\mu_i)(z_j-\mu_i)^T$: data radius of class C_i samples

    μ: test data mean, C_i: test samples of class i, μ_i: mean of class i,
- Laplacian regularization improves KCCA projection directions.



References

[1] Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.R.: *Canonical Correlation Analysis: An Overview with Application to Learning Methods.* Neural Computation (2004)

[2] Belkin, M., Niyogi, P., Sindhwani, V.: *Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples.* JMLR (2006)

[3] Fukumizu, K., Gretton, A., Sun, X., Schölkopf, B.: *Kernel Measures of Conditional Dependence.* NIPS (2007)