A Hilbert-Schmidt Dependence Maximization Approach to Unsupervised Structure Discovery

Matthew B. Blaschko BLASCHKO@TUEBINGEN.MPG.DE
Arthur Gretton ARTHUR@TUEBINGEN.MPG.DE
Max Planck Institute for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany

1. Introduction

In recent work by (Song et al., 2007), it has been proposed to perform clustering by maximizing a Hilbert-Schmidt independence criterion with respect to a predefined cluster structure Y, by solving for the partition matrix, Π. We extend this approach here to the case where the cluster structure Y is not fixed, but is a quantity to be optimized; and we use an independence criterion which has been shown to be more sensitive at small sample sizes (the Hilbert-Schmidt Normalized Information Criterion, or HSNIC (Fukumizu et al., 2008)). We demonstrate the use of this framework in two scenarios. In the first, we adopt a cluster structure selection approach in which the HSNIC is used to select a structure from several candidates. In the second, we consider the case where we discover structure by directly optimizing Y.

2. The normalized H-S independence criterion

Let $\mathcal{F}$ be a reproducing kernel Hilbert space of functions from $\mathcal{X}$ to $\Re$, where $\mathcal{X}$ is a separable metric space. To each point $x \in \mathcal{X}$, there corresponds an element $\phi(x) \in \mathcal{F}$ (we call ϕ the *feature map*) such that $\langle \phi(x), \phi(x') \rangle_{\mathcal{F}} = k(x, x')$, where $k : \mathcal{X} \times \mathcal{X} \to \Re$ is a unique positive definite kernel. We also define a second RKHS $\mathcal{G}$ with respect to the separable metric space $\mathcal{Y}$, with feature map ψ and kernel $\langle \psi(y), \psi(y') \rangle_{\mathcal{G}} = l(y, y')$.

Let $\Pr_{X,Y}$ be a joint measure on $(\mathcal{X} \times \mathcal{Y}, \Gamma \times \Lambda)$ (here Γ and Λ are the Borel σ-algebras on $\mathcal{X}$ and $\mathcal{Y}$), with associated marginal measures $\Pr_X$ and $\Pr_Y$ and random variables X and Y. Then following (Baker, 1973; Fukumizu et al., 2004), the covariance operator $C_{xy} : \mathcal{G} \to \mathcal{F}$ is defined such that for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$,

$$\langle f, C_{xy} g \rangle_{\mathcal{F}} = \mathbf{E}_{\mathsf{x},\mathsf{y}} \left([f(\mathsf{x}) - \mathbf{E}_{\mathsf{x}}(f(\mathsf{x}))] \, [g(\mathsf{y}) - \mathbf{E}_{\mathsf{y}}(g(\mathsf{y}))] \right).$$

In practice, we do not deal with the measure $\Pr_{\mathsf{x},\mathsf{y}}$ itself, but instead observe samples drawn independently according to it. We write an i.i.d. sample of size n from $\Pr_{X,Y}$ as $\boldsymbol{z} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and likewise $\boldsymbol{x} := \{x_1, \ldots, x_n\}$ and $\boldsymbol{y} := \{y_1, \ldots y_n\}$. Finally, we define the Gram matrices $\mathbf{K}$ and $\mathbf{L}$ of inner products in $\mathcal{F}$ and $\mathcal{G}$, respectively, between the mapped observations above: here $\mathbf{K}$ has $(i.j)$th entry $k(x_i, x_j)$ and $\mathbf{L}$ has (i, j)th entry $l(y_i, y_j)$. The Gram matrices for the variables centered in their respective feature spaces are

$$\widetilde{\mathbf{K}} := \mathbf{HKH}, \quad \widetilde{\mathbf{L}} := \mathbf{HLH},$$

where

$$\mathbf{H} = \mathbf{I} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top, \tag{1}$$

and $\mathbf{1}_n$ is an $m \times 1$ vector of ones.

We now define the normalized covariance operator, and the associated operator norm, following (Fukumizu et al., 2008). We know from (Baker, 1973) that the covariance operator can be decomposed as

$$C_{xy} = C_{xx}^{1/2} V_{xy} C_{yy}^{1/2},$$

where V_{xy} is the normalized cross-covariance operator (its maximum singular value is bounded by 1).

As discussed in (Fukumizu et al., 2008), when the kernels are characteristic, then $\|C_{xy}\|_{\mathrm{HS}}^2 = \|V_{xy}\|_{\mathrm{HS}}^2 = 0$ if and only if the random variables are independent. Universal kernels in the sense of (Steinwart, 2001) are characteristic, as are Gaussian kernels on $\Re^d$. Thus, both the covariance operator and the normalized covariance operator can be used as dependence measures between random variables. According to (Gretton et al., 2005), an empirical estimate of $\|C_{xy}\|_{\mathrm{HS}}^2$ is

$$\mathrm{HSIC}(\mathcal{F}, \mathcal{G}, \boldsymbol{x}, \boldsymbol{y}) := \mathrm{Tr}\left[\mathbf{HKHL}\right],$$

which we call the Hilbert-Schmidt Independence Criterion. Likewise, the Hilbert-Schmidt Normalized Information Criterion denotes the empirical estimate of

$\|V_{xy}\|^2_{\text{HS}}$, and is defined

$$\text{HSNIC}(\mathcal{F}, \mathcal{G}, \boldsymbol{x}, \boldsymbol{y}) := \text{Tr}\left[\widetilde{\mathbf{K}}(\widetilde{\mathbf{K}} + \varepsilon_n \mathbf{I})^{-1}\widetilde{\mathbf{L}}(\widetilde{\mathbf{L}} + \varepsilon_n \mathbf{I})^{-1}\right],$$

where ε_n is a regularization parameter which decays to zero with increasing n. It is shown (Fukumizu et al., 2008) that subject to an appropriate decay in the regularization scaling for increasing sample size, this asymptotically approaches the mean squared contingency,

$$\text{HSNIC}(\mathcal{F}, \mathcal{G}, \boldsymbol{x}, \boldsymbol{y}) \xrightarrow{P} \int\int_{\mathcal{X}\times\mathcal{Y}} \left(\frac{\Pr_{X,Y}(x,y)}{\Pr_X(x)\Pr_Y(y)} \Pr_X(x)\Pr_Y(y) d\mu(x) d\mu(y)\right). \quad (2)$$

3. Structure Selection

We now apply the HSNIC to the problem of cluster structure selection. In Figure 1, we show example images from a 9 class dataset consisting of three different faces with 3 different facial expressions each (Song et al., 2007). The first set of experiments we have performed is to use HSNIC as a measure to select out of a set of possible structures, the one that best represents the data.

For these experiments, we have constructed matrices $Y_a, \ldots, Y_h$ based on the tree structures in Figure 2. Each leaf node represents a cluster, and the entry Y_{ij} is proportional to the depth of the closest interior node that connects leaf i and leaf j. For each structure matrix, we optimize for the corresponding partition matrix, Π^*, that maximizes the HSNIC score,

$$\text{HSNIC}(\Pi) = \text{Tr}\left[M_x H \Pi Y \Pi^T H\right] \quad (3)$$

where $M_x = \widetilde{\mathbf{K}}(\widetilde{\mathbf{K}} + \varepsilon_n \mathbf{I})^{-1}$, using the algorithm described in (Song et al., 2007). In order to ensure comparability of the HSNIC scores across different structures, it is necessary to normalize Y prior to computation of the score as follows

$$\tilde{Y} = \frac{Y}{\sqrt{\text{Tr}\left[\Pi Y \Pi^T H \Pi Y \Pi^T H\right]}}. \quad (4)$$

We provide a scatter plot in Figure 3 of the resulting HSNIC scores along with the conditional entropy $H(l|c)$ of the true labels, l, given the predicted clusters, c. $H(l|c) \geq 0$ with equality only if the clusters are pure. We note that there is extremely high negative correlation between the HSNIC scores and the conditional entropy. This indicates that for these structures, HSNIC is very well able to determine which one best separates the classes.

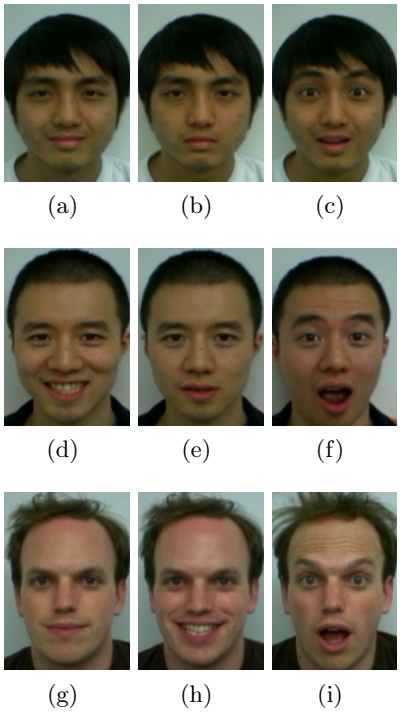

Figure 1. Face dataset

4. Structure Discovery

We have shown in the previous section that the HSNIC objective is well suited for structure prediction given a fixed set of possible structures, and we explore here how to more directly optimize the Y matrix rather than selecting from a fixed set. We describe a direct optimization strategy that constrains Y only to be positive definite. We therefore wish to find a decomposition of $M_x \approx \lambda \Pi Y \Pi^T$ subject to $Y \succeq \mathbf{0} \in \mathbb{R}^{k\times k}$, Π being a partition matrix, and $\text{Tr}\left[\Pi Y \Pi^T H \Pi Y \Pi^T H\right] = 1$, where λ is a scale factor. We have used an iterative approach to find a local optimum of this problem by solving alternately for Π and Y. First we fix Y and solve for Π using the algorithm of (Song et al., 2007). Then we fix Π and solve for Y. The second optimization can be solved in closed form by solving for the KKT conditions, yielding

$$Y^* \propto \left(\Pi^T H \Pi\right)^{-1} \Pi^T H M_x H \Pi \left(\Pi^T H \Pi\right)^{-1}. \quad (5)$$

We have repeated the experiments with initial settings of Y given by the structures in Figures 2(b) and 2(h), as well as random initializations. In each case, the learned structure converged to the same matrix, up to a permutation (Figure 4). For comparison, the learned clustering yields an entropy score of 0.7759, which is worse than the performance of the structure given in Figure 2(b) despite the higher HSNIC score. How-

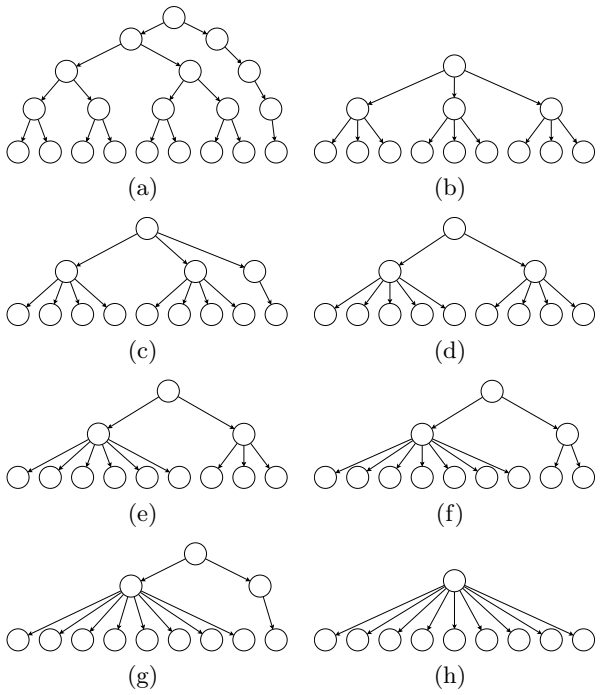

Figure 2. Structures used in the structure selection experiments

ever, the matrix in Figure 4 indeed has a plausible structure. Faces 1(a)-1(c) are conflated into the same cluster, with the exception of two singleton clusters, while the rest of the classes are perfectly clustered. Furthermore, classes 1(d) and 1(e) are given high similarity, as well as classes 1(g) and 1(h).

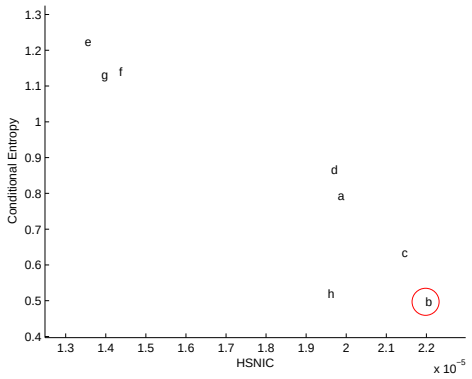

Figure 3. Scatter plot of HSNIC vs. conditional entropy. The letters correspond to the structures in Figure 2. Circled in red is the point with the best conditional entropy score, and the one that is selected by the HSNIC structure selection. HSNIC and conditional entropy scores have a correlation coefficient of −0.93.

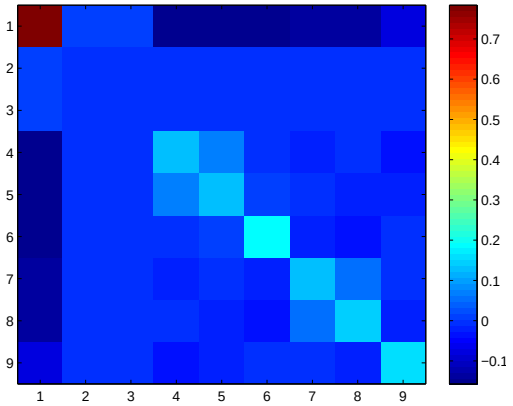

Figure 4. The learned structure for the Faces dataset.

Acknowledgements

The first author is supported by a Marie Curie fellowship under the EC funded project PerAct, EST 504321. This work is funded in part by the CLASS project, IST 027978, and the Pascal Network, IST 2002-506778.

References

Baker, C. (1973). Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society, 186*, 273–289.

Fukumizu, K., Bach, F. R., & Jordan, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J. Mach. Learn. Res.*, *5*, 73–99.

Fukumizu, K., Gretton, A., Sun, X., & Schölkopf, B. (2008). Kernel measures of conditional dependence. *NIPS 20*.

Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. *Proc. Intl. Conf. on Algorithmic Learning Theory* (pp. 63–78).

Song, L., Smola, A., Gretton, A., & Borgwardt, K. M. (2007). A dependence maximization view of clustering. *ICML* (pp. 815–822).

Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.*, *2*, 67–93.