Notation

$\mathcal{L}$	Set of classification labels
$\{(x_i, y_i)\}_{i=1}^n$	Data set, where features $x_i \in \mathbb{R}^d$ and labels $y_i \in \mathcal{L}$
n_y	Number of features labelled y
$x_i^{(y)}$	i^{th} feature labelled y
$\mathcal{H}$	Real Hilbert space
$\mathcal{H}^*$	Continuous dual space of $\mathcal{H}$
$\Phi : \mathbb{R}^d \to \mathcal{H}$	Feature map
$\gamma : \mathcal{H} \to \mathcal{L}$	Functional, where $\gamma(\Phi(x_i)) = y_i$
$\Phi_i, \Phi_i^{(\gamma)}$	Shorthand for $\Phi(x_i)$ and $\Phi(x_i^{(\gamma)})$, respectively
$\hat{p}$	Empirical distribution of the data $\{\Phi_i\}_{i=1}^n$
$\hat{p}_\gamma$	Empirical distribution conditioned on γ
$\mathbb{E}_p, \mathbb{V}_p, \mathbb{C}_p$	Expectation, variance, and covariance with respect to distribution p, respectively

Covariance

We provide a point form derivation of conditional covariance.

- The covariance with respect to $\hat{p}_\gamma$ of arbitrary functionals $\phi, \psi \in \mathcal{H}^*$ is given by

$$\mathbb{C}_{\hat{p}_\gamma}(\phi, \psi) = \frac{1}{n_\gamma} \sum_{i=1}^{n_\gamma} \left(\phi(\Phi_i^{(\gamma)}) - \frac{1}{n_\gamma} \sum_{j=1}^{n_\gamma} \phi(\Phi_j^{(\gamma)}) \right) \left(\psi(\Phi_i^{(\gamma)}) - \frac{1}{n_\gamma} \sum_{j=1}^{n_\gamma} \psi(\Phi_j^{(\gamma)}) \right) \tag{1}$$

- By the Riesz representation theorem, (1) can be written as

$$\mathbb{C}_{\hat{p}_\gamma}(\phi, \psi) = \frac{1}{n_\gamma} \sum_{i=1}^{n_\gamma} \left(\langle \Phi_i^{(\gamma)}, T\phi \rangle_{\mathcal{H}} - \frac{1}{n_\gamma} \sum_{j=1}^{n_\gamma} \langle \Phi_j^{(\gamma)}, T\phi \rangle_{\mathcal{H}} \right) \left(\langle \Phi_i^{(\gamma)}, T\psi \rangle_{\mathcal{H}} - \frac{1}{n_\gamma} \sum_{j=1}^{n_\gamma} \langle \Phi_j^{(\gamma)}, T\psi \rangle_{\mathcal{H}} \right) \tag{2}$$

  where $T\phi$ and $T\psi$ are the unique vector representations of ϕ and ψ in $\mathcal{H}$, respectively.

- Since $\mathcal{S} = \operatorname{span}\{\Phi_i\}_{i=1}^n$ is a closed subspace of $\mathcal{H}$, by the Hilbert projection theorem,

$$T\phi = \sum_{i=1}^{n} \alpha_i \Phi_i + u \quad \text{and} \quad T\psi = \sum_{i=1}^{n} \beta_i \Phi_i + v \tag{3}$$

  for some coefficients $\alpha_i, \beta_i \in \mathbb{R}$ and elements $u, v \in \mathcal{S}^\perp$.

- Substituting (3) into (2), we obtain:

$$\begin{aligned} \mathbb{C}_{\hat{p}_\gamma}(\phi, \psi) = \frac{1}{n_\gamma} \sum_{i=1}^{n_\gamma} & \left(\sum_{k=1}^{n} \alpha_k \langle \Phi_i^{(\gamma)}, \Phi_k \rangle_{\mathcal{H}} - \frac{1}{n_\gamma} \sum_{j=1}^{n_\gamma} \sum_{k=1}^{n} \alpha_k \langle \Phi_j^{(\gamma)}, \Phi_k \rangle_{\mathcal{H}} \right) \\ & \left(\sum_{k=1}^{n} \beta_k \langle \Phi_i^{(\gamma)}, \Phi_k \rangle_{\mathcal{H}} - \frac{1}{n_\gamma} \sum_{j=1}^{n_\gamma} \sum_{k=1}^{n} \beta_k \langle \Phi_j^{(\gamma)}, \Phi_k \rangle_{\mathcal{H}} \right) \end{aligned} \tag{4}$$

Note that u and v vanish in (4) as they are orthogonal to all Φ_i and thus have no influence on the covariance. From now on, we assume $u = v = 0$.

- Expressing (4) in vector/matrix notation:

$$\mathbb{C}_{\hat{p}_\gamma}(\phi,\psi) = \frac{1}{n_\gamma}\sum_{i=1}^{n_\gamma}\left(K^{(\gamma)}_{i,\cdot}\alpha - \frac{1}{n_\gamma}\mathbb{1}^T_{n_\gamma}K^{(\gamma)}\alpha\right)\left(K^{(\gamma)}_{i,\cdot}\beta - \frac{1}{n_\gamma}\mathbb{1}^T_{n_\gamma}K^{(\gamma)}\beta\right) = \frac{1}{n_\gamma}\alpha^T(K^{(\gamma)})^T C_{n_\gamma}K^{(\gamma)}\beta \tag{5}$$

where α and β are the n-dimensional vectors with components α_i and β_i, respectively, $K^{(\gamma)}$ is the $n_\gamma \times n$ matrix with entries $K^{(\gamma)}_{i,j} = \langle \Phi^{(\gamma)}_i, \Phi_j\rangle_{\mathcal{H}}$, $\mathbb{1}_{n_\gamma}$ is the vector of n_γ ones, and C_{n_γ} is the $n_\gamma \times n_\gamma$ centering matrix.

- There are three other variance/covariance formulas implied by (5) that we will need:

$$\mathbb{V}_{\hat{p}_\gamma}(\phi) = \frac{1}{n_\gamma}\alpha^T(K^{(\gamma)})^T C_{n_\gamma}K^{(\gamma)}\alpha \tag{6}$$

$$\mathbb{C}_{\hat{p}}(\phi,\psi) = \frac{1}{n}\alpha^T K^T C_n K\beta \tag{7}$$

$$\mathbb{V}_{\hat{p}}(\phi) = \frac{1}{n}\alpha^T K^T C_n K\alpha \tag{8}$$

where K is the $n \times n$ matrix with entries $K_{i,j} = \langle \Phi_i, \Phi_j\rangle_{\mathcal{H}}$.

Law of Total Variance

By the law of total variance,

$$\mathbb{V}_{\hat{p}}(\phi) = \mathbb{E}_{\hat{p}}\big[\mathbb{V}_{\hat{p}_\gamma}(\phi)\big] + \mathbb{V}_{\hat{p}}\big(\mathbb{E}_{\hat{p}_\gamma}[\phi]\big) \tag{9}$$

Hence, the ratio of "between-group" variance to "within-group" variance is given by

$$\frac{\mathbb{V}_{\hat{p}}\big(\mathbb{E}_{\hat{p}_\gamma}[\phi]\big)}{\mathbb{E}_{\hat{p}}\big[\mathbb{V}_{\hat{p}_\gamma}(\phi)\big]} = \frac{\mathbb{V}_{\hat{p}}(\phi)}{\mathbb{E}_{\hat{p}}\big[\mathbb{V}_{\hat{p}_\gamma}(\phi)\big]} - 1 = \frac{\alpha^T K^T C_n K\alpha}{\alpha^T\Big(\sum_{y\in\mathcal{L}}(K^{(y)})^T C_{n_y}K^{(y)}\Big)\alpha} - 1 \tag{10}$$

Low-Dimensional Representations

The i^{th} coordinate of a k-dimensional representation of $x \in \mathbb{R}^d$ is given by

$$\phi_i(\Phi(x)) = \langle \Phi(x), T\phi_i\rangle_{\mathcal{H}} = \sum_{j=1}^{n}\alpha_i^{(j)}\langle \Phi(x), \Phi_j\rangle_{\mathcal{H}}, \quad i = 1,\ldots,k \tag{11}$$

where $\alpha_i^{(j)}$ is the j^{th} component of the n-dimensional vector α_i which solves the optimization problem

$$\begin{aligned} \max_{\alpha\in\mathbb{R}^n} \quad & \frac{\alpha^T K^T C_n K\alpha}{\alpha^T\Big(\sum_{y\in\mathcal{L}}(K^{(y)})^T C_{n_y}K^{(y)}\Big)\alpha} \\ \text{s.t.} \quad & \alpha_\ell^T K^T C_n K\alpha = 0, \quad \ell = 1,\ldots,i-1 \end{aligned} \tag{12}$$

The constraints in (12) specify that the functionals are uncorrelated – that is, $\mathbb{C}_{\hat{p}}(\phi_i, \phi_j) = 0$ whenever $i \neq j$.

Regularization

The following optimization problem is a regularized version of (12):

$$\begin{aligned} \max_{\alpha\in\mathbb{R}^n} \quad & \frac{\alpha^T K^T C_n K\alpha}{\alpha^T \Big(\sum_{y\in\mathcal{L}} (K^{(y)})^T C_{n_y} K^{(y)} + \lambda R \Big)\alpha} \\ \text{s.t.} \quad & \alpha_\ell^T K^T C_n K\alpha = 0, \quad \ell = 1, \ldots, i-1 \end{aligned} \tag{13}$$

where λ is a nonnegative tuning parameter and R is either the $n \times n$ identity matrix I_n or K. When $R = I_n$, (13) is equivalent to adding $\lambda\|\alpha\|_2^2$ to the denominator in (12), which constrains the functional coefficients. When $R = K$, (13) is equivalent to adding $\lambda\|\phi\|_{\mathcal{H}^*}^2 = \lambda\|T\phi\|_{\mathcal{H}}^2 = \lambda\alpha^T K\alpha$ to the denominator in (12), which constrains the functional as a whole.

Generalized Rayleigh Quotients

When $i = 1$, (13) has no constraints and can be solved by solving a generalized eigenvalue problem [1]. For $i = 2, \ldots, k$, (13) is a linearly constrained generalized Rayleigh quotient and can be solved by following [2].

Data Centering

Centering the Φ_i is equivalent to centering the rows of K and $K^{(\gamma)}$ – that is, replacing every occurrence of K and $K^{(\gamma)}$ with KC_n and $K^{(\gamma)}C_n$, respectively. In this case, the k-dimensional representations of the rows of a $r \times d$ data matrix X' are given by the rows of

$$\Big(K_{X',X} - \frac{1}{n}\mathbb{1}_r\mathbb{1}_n^T K \Big) C_n A^T \tag{14}$$

where X is the $n \times d$ training data matrix and A is the $k \times n$ matrix with i^{th} row given by the functional coefficient vector α_i.

References

[1] Chen, G. (2020). Lecture 4: Rayleigh Quotients. https://www.sjsu.edu/faculty/guangliang.chen/Math253S20/lec4RayleighQuotient.pdf

[2] Cour, T. (2006). Affinely Constrained Rayleigh Quotients. https://www.cis.upenn.edu/~jshi/papers/supplement_nips2006.pdf