Covariance

Let $\mathcal{H}$ be a real Hilbert space and let $\mathbb{P}$ be a probability measure on $\mathcal{H}$. The covariance of any two bounded linear functionals ϕ, ψ is given by

$$\begin{aligned}\mathsf{Cov}_{\mathbb{P}}\left(\phi,\psi\right) &= \mathbb{E}_{\mathbb{P}}\left[(\phi-\mathbb{E}_{\mathbb{P}}\left[\phi\right])(\psi-\mathbb{E}_{\mathbb{P}}\left[\psi\right])\right] \\ &= \int_{h\in\mathcal{H}}\left(\phi(h)-\int_{h'\in\mathcal{H}}\phi\left(h'\right)d\mathbb{P}\left(h'\right)\right)\left(\psi(h)-\int_{h'\in\mathcal{H}}\psi\left(h'\right)d\mathbb{P}\left(h'\right)\right)d\mathbb{P}(h).\end{aligned} \tag{1}$$

Riesz representation theorem

For every bounded linear functional φ, there exists a unique vector $T\varphi$ such that $\varphi(h) = \langle h, T\varphi\rangle_{\mathcal{H}}$. Hence,

$$\begin{aligned}\mathsf{Cov}_{\mathbb{P}}\left(\phi,\psi\right) &= \int_{h\in\mathcal{H}}\left(\langle h,T\phi\rangle_{\mathcal{H}}-\int_{h'\in\mathcal{H}}\langle h',T\phi\rangle_{\mathcal{H}}d\mathbb{P}\left(h'\right)\right)\left(\langle h,T\psi\rangle_{\mathcal{H}}-\int_{h'\in\mathcal{H}}\langle h',T\psi\rangle_{\mathcal{H}}d\mathbb{P}\left(h'\right)\right)d\mathbb{P}(h) \\ &= \int_{h\in\mathcal{H}}\left\langle h-\int_{h'\in\mathcal{H}}h'd\mathbb{P}(h'),T\phi\right\rangle_{\mathcal{H}}\left\langle h-\int_{h'\in\mathcal{H}}h'd\mathbb{P}(h'),T\psi\right\rangle_{\mathcal{H}}d\mathbb{P}(h) \\ &= \int_{h\in\mathcal{H}}\left\langle h-\mu,T\phi\right\rangle_{\mathcal{H}}\left\langle h-\mu,T\psi\right\rangle_{\mathcal{H}}d\mathbb{P}(h)\end{aligned} \tag{2}$$

where μ is the expected $h \in \mathcal{H}$ with respect to $\mathbb{P}$ and is given by $\mu = \int_{h\in\mathcal{H}} h d\mathbb{P}(h)$.

Direct sum decomposition

Let $h_1, \ldots, h_n$ be a sample drawn from $\mathbb{P}$. Since $S = \mathsf{span}\{h_1, \ldots, h_n\}$ is a closed subspace of $\mathcal{H}$, by the Hilbert projection theorem, $\mathcal{H} = S \oplus S^{\perp}$ where $S^{\perp}$ is the orthogonal complement of S. Thus,

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

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

Empirical covariance

Substituting (3) into (2) and letting $\widehat{\mathbb{P}}$ denote the empirical distribution of the sample,

$$\begin{aligned}\mathsf{Cov}_{\widehat{\mathbb{P}}}(\phi,\psi) &= \frac{1}{n}\sum_{i=1}^{n}\left\langle h_i-\overline{h},\sum_{j=1}^{n}\alpha_j h_j+u\right\rangle_{\mathcal{H}}\left\langle h_i-\overline{h},\sum_{j=1}^{n}\beta_j h_j+v\right\rangle_{\mathcal{H}} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\left\langle h_i-\overline{h},h_j\right\rangle_{\mathcal{H}}\alpha_j\right)\left(\sum_{j=1}^{n}\left\langle h_i-\overline{h},h_j\right\rangle_{\mathcal{H}}\beta_j\right) \\ &= \frac{1}{n}\left(CG\alpha\right)^T CG\beta \\ &= \frac{1}{n}\alpha^T GCG\beta\end{aligned} \tag{4}$$

where $\overline{h}$ is the sample mean, $C = I - \frac{1}{n}11^T$ is the symmetric and idempotent (i.e., $C^2 = C$) centering matrix, G is the symmetric matrix with entries $G_{i,j} = \langle h_i, h_j\rangle_{\mathcal{H}}$, and α and β are the n-dimensional vectors with components α_j and β_j, respectively.

Maximum variance functional

Consider the problem of finding a maximum variance functional with unit norm with respect to $\widehat{\mathbb{P}}$. From (3) and (4), such a functional has the form $\phi^\star(h) = \langle h, \sum_{i=1}^n \alpha_i^\star h_i \rangle_{\mathcal{H}}$ where $\alpha^\star = (\alpha_1^\star, \dots, \alpha_n^\star)^T$ is a solution to the optimization problem

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^n} \quad & \frac{1}{n} \alpha^T G C G \alpha \\ \text{s.t.} \quad & \alpha^T G \alpha = 1 \end{aligned} \tag{5}$$

The constraint in (5) follows from the fact that $\|\phi\|_{\mathcal{H}^*} = \|T\phi\|_{\mathcal{H}}$ where $\mathcal{H}^*$ denotes the space of bounded linear functionals on $\mathcal{H}$.

Centering the data

If we first centre the h_i to obtain the sample $h_1', \dots, h_n'$ where $h_i' = h_i - \overline{h}$, then the matrix G becomes $G' = CGC$ and (5) becomes

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^n} \quad & \frac{1}{n} \alpha^T \left(G'\right)^2 \alpha \\ \text{s.t.} \quad & \alpha^T G' \alpha = 1. \end{aligned} \tag{6}$$

Lagrange multipliers

A solution $\alpha^\star$ to (6) satisfies the equation $\left(G'\right)^2 \alpha^\star = \gamma G' \alpha^\star$ for some constant γ. Hence, if G' is positive definite, then $\alpha^\star$ is an eigenvector of G' with corresponding eigenvalue γ. Since G' is symmetric (and assumed positive definite), there exists an orthonormal basis $e_1, \dots, e_n$ of eigenvectors of G' with corresponding eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n > 0$, respectively. It follows that $\alpha^\star = \frac{e_1}{\sqrt{\lambda_1}}$ is an optimal solution and, hence,

$$\phi^\star(h) = \left\langle h, \sum_{i=1}^n \alpha_i^\star h_i' \right\rangle_{\mathcal{H}} = \left\langle h, \sum_{i=1}^n \frac{e_1^{(i)}}{\sqrt{\lambda_1}} h_i' \right\rangle_{\mathcal{H}} = \frac{1}{\sqrt{\lambda_1}} \sum_{i=1}^n e_1^{(i)} \langle h, h_i' \rangle_{\mathcal{H}} \tag{7}$$

where $e_1^{(i)}$ denotes the i^{th} component of e_1.

Maximum variance uncorrelated functionals

It is straightforward to generalize to a sequence of optimization problems where the i^{th} problem is given by

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^n} \quad & \frac{1}{n} \alpha^T \left(G'\right)^2 \alpha \\ \text{s.t.} \quad & \alpha^T G' \alpha = 1 \\ & \frac{1}{n} \alpha^T \left(G'\right)^2 \alpha_j^\star = 0, \quad j = 1, \dots, i-1 \end{aligned} \tag{8}$$

and where $i = 1$ corresponds to (6). The corresponding sequence of solutions $\alpha_i^\star = \frac{e_i}{\sqrt{\lambda_i}}$ correspond to unit norm maximum variance uncorrelated functionals.

Algorithm

1. Construct the matrix G.
2. Centre G to obtain the matrix $G' = CGC$.
3. Compute the spectral decomposition of G' to obtain the matrix of eigenvectors Γ and the matrix of eigenvalues Λ.

4. (a) To obtain k-dimensional representations of the h_i, return the first k columns of $G'\Gamma\Lambda^{-\frac{1}{2}}$.
   (b) To obtain a k-dimensional representation of a new data point h, return the first k components of $\Lambda^{-\frac{1}{2}}\Gamma^T u$ where the i^{th} component of u is $\langle h - \overline{h}, h'_i \rangle_{\mathcal{H}}$.

Examples

1. PCA: Let $\mathcal{H} = \mathbb{R}^d$ and let $\mathbb{P}$ be a probability measure on $\mathcal{H}$. It follows that G has entries $G_{i,j} = x_i^T x_j$ where $x_1, \ldots, x_n$ is a sample drawn from $\mathbb{P}$.

2. Kernel PCA: Let $\mathcal{H}$ be a reproducing kernel Hilbert space with reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and let $\mathbb{P}$ be a probability measure on $\mathcal{H}$ given by

$$\mathbb{P}(E) = \mathbb{Q}(\{x \in \mathbb{R}^d \mid \Phi(x) \in E\}), \quad E \subset \mathcal{H}$$

   where $\mathbb{Q}$ is a probability measure on $\mathbb{R}^d$ and $\Phi : \mathbb{R}^d \to \mathcal{H}$ is the canonical feature map $\Phi(x) = k_x$. It follows that G has entries $G_{i,j} = \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}} = k(x_i, x_j)$ where $x_1, \ldots, x_n$ is a sample drawn from $\mathbb{Q}$.