Proof We can decompose

$$\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2} = \sum_{i=1}^{\ell} \lambda_i \mathbf{v}_i \mathbf{v}_i',$$

implying that

$$\mathbf{A} = \sum_{i=1}^{\ell} \lambda_i \mathbf{B}^{1/2}\mathbf{v}_i \left(\mathbf{B}^{1/2}\mathbf{v}_i\right)' = \sum_{i=1}^{\ell} \lambda_i \mathbf{B}\mathbf{w}_i \left(\mathbf{B}\mathbf{w}_i\right)',$$

as required. □

Definition 6.27 [Generalised deflation] The final proposition suggests how we can deflate the matrix $\mathbf{A}$ in an iterative direct solution of the generalised eigenvalue problem

$$\mathbf{A}\mathbf{w} = \lambda \mathbf{B}\mathbf{w}.$$

After finding a non-zero eigenvalue–eigenvector pair λ, $\mathbf{w}$ we deflate $\mathbf{A}$ by

$$\mathbf{A} \longleftarrow \mathbf{A} - \lambda \mathbf{B}\mathbf{w}\left(\mathbf{B}\mathbf{w}\right)' = \mathbf{A} - \lambda \mathbf{B}\mathbf{w}\mathbf{w}'\mathbf{B}',$$

leaving $\mathbf{B}$ unchanged. ■

6.5 Canonical correlation analysis

We have looked at two ways of detecting stable patterns through the use of eigen-decompositions firstly to optimise variance of the training data in kernel PCA and secondly to maximise the covariance between two views of the data typically input and output vectors. We now again consider the case in which we have two views of the data which are paired in the sense that each example as a pair of representations. This situation is sometimes referred to as a paired dataset. We will show how to find correlations between the two views.

An extreme case would be where the second view is simply the labels of the examples. In general we are interested here in cases where we have a more complex 'output' that amounts to a different representation of the same object.

Example 6.28 A set of documents containing each document in two different languages is a paired dataset. The two versions give different views of the same underlying object, in this case the semantic content of the document. Such a dataset is known as a parallel corpus. By seeking correlations between the two views, we might hope to extract features that bring out

the underlying semantic content. The fact that a pattern has been found in both views suggests that it is not related to the irrelevant representation specific aspects of one or other view, but rather to the common underlying semantic content. This example will be explored further in Chapter 10. ■

This section will develop the methodology for finding these common patterns in different views through seeking correlations between projection values from the two views. Using an appropriate regularisation technique, the methods are extended to kernel-defined feature spaces.

Recall that in Section 5.3 we defined the correlation between two zero-mean univariate random variables x and y to be

$$\rho = \operatorname{corr}(x, y) = \frac{\mathbb{E}[xy]}{\sqrt{\mathbb{E}[xx]\,\mathbb{E}[yy]}} = \frac{\operatorname{cov}(x, y)}{\sqrt{\operatorname{var}(x)}\sqrt{\operatorname{var}(y)}}.$$

Definition 6.29 [Paired dataset] A paired dataset is created when each object $\mathbf{x} \in X$ can be viewed through two distinct projections into two feature spaces

$$\phi_a : \mathbf{x} \longrightarrow F_a \quad \text{and} \quad \phi_b : \mathbf{x} \longrightarrow F_b,$$

where F_a is the feature space associated with one representation and F_b the feature space for the other. Figure 6.3 illustrates this configuration. The corresponding kernel functions are denoted κ_a and κ_b. Hence, we have a multivariate random vector $(\phi_a(\mathbf{x}), \phi_b(\mathbf{x}))$. Assume we are given a training set

$$S = \{(\phi_a(\mathbf{x}_1), \phi_b(\mathbf{x}_1)), \ldots, (\phi_a(\mathbf{x}_\ell), \phi_b(\mathbf{x}_\ell))\}$$

drawn independently at random according to the underlying distribution. We will refer to such a set as a *paired or aligned dataset* in the feature space defined by the kernels κ_a and κ_b. ■

We now seek to maximise the empirical correlation between $x_a = \mathbf{w}_a' \phi_a(\mathbf{x})$ and $x_b = \mathbf{w}_b' \phi_b(\mathbf{x})$ over the projection directions $\mathbf{w}_a$ and $\mathbf{w}_b$

$$\begin{aligned} \max \rho &= \frac{\hat{\mathbb{E}}[x_a x_b]}{\sqrt{\hat{\mathbb{E}}[x_a x_a]\,\hat{\mathbb{E}}[x_b x_b]}} \\ &= \frac{\hat{\mathbb{E}}\left[\mathbf{w}_a' \phi_a(\mathbf{x})\, \phi_b(\mathbf{x})' \mathbf{w}_b\right]}{\sqrt{\hat{\mathbb{E}}\left[\mathbf{w}_a' \phi_a(\mathbf{x})\, \phi_a(\mathbf{x})' \mathbf{w}_a\right] \hat{\mathbb{E}}\left[\mathbf{w}_b' \phi_b(\mathbf{x})\, \phi_b(\mathbf{x})' \mathbf{w}_b\right]}} \end{aligned}$$

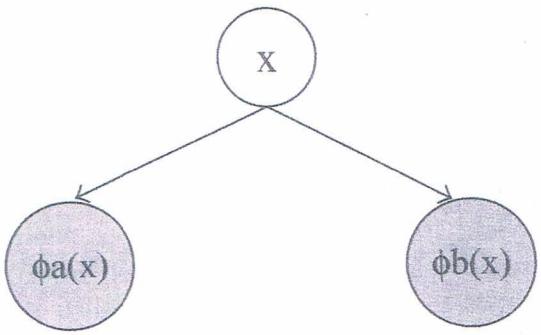

Fig. 6.3. The two embeddings of a paired dataset.

$$= \frac{\mathbf{w}_a' \mathbf{C}_{ab} \mathbf{w}_b}{\sqrt{\mathbf{w}_a' \mathbf{C}_{aa} \mathbf{w}_a \mathbf{w}_b' \mathbf{C}_{bb} \mathbf{w}_b}}, \tag{6.19}$$

where we have decomposed the empirical covariance matrix as follows

$$\begin{aligned}
\mathbf{C} &= \frac{1}{\ell} \sum_{i=1}^{\ell} \left(\phi_a(\mathbf{x}), \phi_b(\mathbf{x})\right) \left(\phi_a(\mathbf{x}), \phi_b(\mathbf{x})\right)' \\
&= \begin{pmatrix} \frac{1}{\ell} \sum_{i=1}^{\ell} \phi_a(\mathbf{x}) \phi_a(\mathbf{x})' & \frac{1}{\ell} \sum_{i=1}^{\ell} \phi_b(\mathbf{x}) \phi_a(\mathbf{x})' \\ \frac{1}{\ell} \sum_{i=1}^{\ell} \phi_a(\mathbf{x}) \phi_b(\mathbf{x})' & \frac{1}{\ell} \sum_{i=1}^{\ell} \phi_b(\mathbf{x}) \phi_b(\mathbf{x})' \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{C}_{aa} & \mathbf{C}_{ba} \\ \mathbf{C}_{ab} & \mathbf{C}_{bb} \end{pmatrix}.
\end{aligned}$$

This optimisation is very similar to that given in (6.14). The only difference is that here the denominator of the quotient measures the norm of the projection vectors differently from the covariance case. In the current optimisation the vectors $\mathbf{w}_a$ and $\mathbf{w}_b$ are again only determined up to direction since rescaling $\mathbf{w}_a$ by λ_a and $\mathbf{w}_b$ by λ_b results in the quotient

$$\begin{aligned}
\frac{\lambda_a \lambda_b \mathbf{w}_a' \mathbf{C}_{ab} \mathbf{w}_b}{\sqrt{\lambda_a^2 \mathbf{w}_a' \mathbf{C}_{aa} \mathbf{w}_a \lambda_b^2 \mathbf{w}_b' \mathbf{C}_{bb} \mathbf{w}_b}} &= \frac{\lambda_a \lambda_b \mathbf{w}_a' \mathbf{C}_{ab} \mathbf{w}_b}{\lambda_a \lambda_b \sqrt{\mathbf{w}_a' \mathbf{C}_{aa} \mathbf{w}_a \mathbf{w}_b' \mathbf{C}_{bb} \mathbf{w}_b}} \\
&= \frac{\mathbf{w}_a' \mathbf{C}_{ab} \mathbf{w}_b}{\sqrt{\mathbf{w}_a' \mathbf{C}_{aa} \mathbf{w}_a \mathbf{w}_b' \mathbf{C}_{bb} \mathbf{w}_b}}.
\end{aligned}$$

This implies that we can constrain the two terms in the denominator to individually have value 1. Hence, the problem is solved by the following optimisation problem.

Computation 6.30 [CCA] Given a paired dataset with covariance matrix

$\mathbf{C}_{ab}$, canonical correlation analysis finds the directions $\mathbf{w}_a$,$\mathbf{w}_b$ that maximise the correlation of corresponding projections by solving

$$\begin{array}{ll} \max_{\mathbf{w}_a,\mathbf{w}_b} & \mathbf{w}_a'\mathbf{C}_{ab}\mathbf{w}_b \\ \text{subject to} & \mathbf{w}_a'\mathbf{C}_{aa}\mathbf{w}_a = 1 \text{ and } \mathbf{w}_b'\mathbf{C}_{bb}\mathbf{w}_b = 1. \end{array} \tag{6.20}$$

■

Solving CCA Applying the Lagrange multiplier technique to the optimisation (6.20) gives

$$\max \ \mathbf{w}_a'\mathbf{C}_{ab}\mathbf{w}_b - \frac{\lambda_a}{2}\left(\mathbf{w}_a'\mathbf{C}_{aa}\mathbf{w}_a - 1\right) - \frac{\lambda_y}{2}\left(\mathbf{w}_b'\mathbf{C}_{bb}\mathbf{w}_b - 1\right).$$

Taking derivatives with respect to $\mathbf{w}_a$ and $\mathbf{w}_b$ we obtain the equations

$$\mathbf{C}_{ab}\mathbf{w}_b - \lambda_a\mathbf{C}_{aa}\mathbf{w}_a = \mathbf{0} \ \text{ and } \ \mathbf{C}_{ba}\mathbf{w}_a - \lambda_b\mathbf{C}_{bb}\mathbf{w}_b = \mathbf{0}. \tag{6.21}$$

Subtracting $\mathbf{w}_a'$ times the first from $\mathbf{w}_b'$ times the second we have

$$\lambda_a\mathbf{w}_a'\mathbf{C}_{aa}\mathbf{w}_a - \lambda_b\mathbf{w}_b'\mathbf{C}_{bb}\mathbf{w}_b = 0,$$

which, taking into account the two constraints, implies $\lambda_a = \lambda_b$. Using λ to denote this value we obtain the following algorithm for computing the correlations.

Algorithm 6.31 [Primal CCA] The following method finds the directions of maximal correlation:

Input	covariance matrices $\mathbf{C}_{aa}$, $\mathbf{C}_{bb}$, $\mathbf{C}_{ba}$ and $\mathbf{C}_{ab}$
Process	solve the generalised eigenvalue problem: $\begin{pmatrix} \mathbf{0} & \mathbf{C}_{ab} \\ \mathbf{C}_{ba} & \mathbf{0} \end{pmatrix}\begin{pmatrix} \mathbf{w}_a \\ \mathbf{w}_b \end{pmatrix} = \lambda\begin{pmatrix} \mathbf{C}_{aa} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{bb} \end{pmatrix}\begin{pmatrix} \mathbf{w}_a \\ \mathbf{w}_b \end{pmatrix}$
Output	eigenvectors and eigenvalues $\mathbf{w}_a^j$, $\mathbf{w}_b^j$ and $\lambda_j > 0$, $j = 1, \ldots, \ell$.

(6.22)

■

This is an example of a generalised eigenvalue problem described in the last section. Note that the value of the eigenvalue for a particular eigenvector gives the size of the correlation since $\mathbf{w}_a'$ times the top portion of (6.22) gives

$$\rho = \mathbf{w}_a'\mathbf{C}_{ab}\mathbf{w}_b = \lambda_a\mathbf{w}_a'\mathbf{C}_{aa}\mathbf{w}_a = \lambda.$$

Hence, we have all eigenvalues lying in the interval $[-1, +1]$, with each λ_i and eigenvector

$$\begin{pmatrix} \mathbf{w}_a \\ \mathbf{w}_b \end{pmatrix}$$

paired with an eigenvalue $-\lambda_i$ with eigenvector

$$\begin{pmatrix} \mathbf{w}_a \\ -\mathbf{w}_b \end{pmatrix}.$$

We are therefore only interested in half the spectrum which we can take to be the positive eigenvalues. The eigenvectors corresponding to the largest eigenvalues are those that identify the strongest correlations. Note that in this case by Proposition 6.22 the eigenvectors will be conjugate with respect to the matrix

$$\begin{pmatrix} \mathbf{C}_{aa} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{bb} \end{pmatrix},$$

so that for $i \neq j$ we have

$$0 = \begin{pmatrix} \mathbf{w}_a^j \\ \mathbf{w}_b^j \end{pmatrix}' \begin{pmatrix} \mathbf{C}_{aa} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{bb} \end{pmatrix} \begin{pmatrix} \mathbf{w}_a^i \\ \mathbf{w}_b^i \end{pmatrix} = (\mathbf{w}_a^j)' \mathbf{C}_{aa} \mathbf{w}_a^i + \left(\mathbf{w}_b^j\right)' \mathbf{C}_{bb} \mathbf{w}_b^i$$

and

$$0 = \begin{pmatrix} \mathbf{w}_a^j \\ \mathbf{w}_b^j \end{pmatrix}' \begin{pmatrix} \mathbf{C}_{aa} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{bb} \end{pmatrix} \begin{pmatrix} \mathbf{w}_a^i \\ -\mathbf{w}_b^i \end{pmatrix} = (\mathbf{w}_a^j)' \mathbf{C}_{aa} \mathbf{w}_a^i - \left(\mathbf{w}_b^j\right)' \mathbf{C}_{bb} \mathbf{w}_b^i$$

yielding

$$(\mathbf{w}_a^j)' \mathbf{C}_{aa} \mathbf{w}_a^i = 0 = \left(\mathbf{w}_b^j\right)' \mathbf{C}_{bb} \mathbf{w}_b^i.$$

This implies that, as with PCA, we obtain a diagonal covariance matrix if we project the data into the coordinate system defined by the eigenvectors, whether we project each view independently or simply the sum of the projections of the two views in the common space. The directions themselves will not, however, be orthogonal in the standard inner product of the feature space.

Dual form of CCA Naturally we wish to solve the problem in the dual formulation. Hence, we consider expressing $\mathbf{w}_a$ and $\mathbf{w}_b$ in terms of their respective parts of the training sample by creating a matrix $\mathbf{X}_a$ whose rows are the vectors $\phi_a(\mathbf{x}_i)$, $i = 1, \ldots, \ell$ and the matrix $\mathbf{X}_b$ with rows $\phi_b(\mathbf{x}_i)$

$$\mathbf{w}_a = \mathbf{X}_a' \boldsymbol{\alpha}_a \text{ and } \mathbf{w}_b = \mathbf{X}_b' \boldsymbol{\alpha}_b.$$

Substituting into (6.20) gives

$$\begin{array}{ll} \max & \boldsymbol{\alpha}_a' \mathbf{X}_a \mathbf{X}_a' \mathbf{X}_b \mathbf{X}_b' \boldsymbol{\alpha}_b \\ \text{subject to} & \boldsymbol{\alpha}_a' \mathbf{X}_a \mathbf{X}_a' \mathbf{X}_a \mathbf{X}_a' \boldsymbol{\alpha}_a = 1 \text{ and } \boldsymbol{\alpha}_b' \mathbf{X}_b \mathbf{X}_b' \mathbf{X}_b \mathbf{X}_b' \boldsymbol{\alpha}_b = 1, \end{array}$$

or equivalently the following optimisation problem.

Computation 6.32 [Kernel CCA] Given a paised dataset with respect to kernels κ_a and κ_b, kernel canonical correlation analysis finds the directions of maximal correlation by solving

$$\begin{array}{ll} \max_{\boldsymbol{\alpha}_a, \boldsymbol{\alpha}_b} & \boldsymbol{\alpha}_a' \mathbf{K}_a \mathbf{K}_b \boldsymbol{\alpha}_b \\ \text{subject to} & \boldsymbol{\alpha}_a' \mathbf{K}_a^2 \boldsymbol{\alpha}_a = 1 \text{ and } \boldsymbol{\alpha}_b' \mathbf{K}_b^2 \boldsymbol{\alpha}_b = 1, \end{array}$$

where $\mathbf{K}_a$ and $\mathbf{K}_b$ are the kernel matrices for the two representations. ■

Figure 6.4 shows the two feature spaces with the projections of 7 points. The shading corresponds to the value of the projection on the first correlation direction using a Gaussian kernel in each feature space.

Overfitting in CCA Again applying the Lagrangian techniques this leads to the equations

$$\mathbf{K}_a \mathbf{K}_b \boldsymbol{\alpha}_b - \lambda \mathbf{K}_a^2 \boldsymbol{\alpha}_a = \mathbf{0} \text{ and } \mathbf{K}_b \mathbf{K}_a \boldsymbol{\alpha}_a - \lambda \mathbf{K}_b^2 \boldsymbol{\alpha}_b = \mathbf{0}.$$

These equations highlight the potential problem of overfitting that arises in high-dimensional feature spaces. If the dimension N_a of the feature space F_a satisfies $N_a \gg \ell$, it is likely that the data will be linearly independent in the feature space. For example this is always true for a Gaussian kernel. But if the data are linearly independent in F_a the matrix $\mathbf{K}_a$ will be full rank and hence invertible. This gives

$$\boldsymbol{\alpha}_a = \frac{1}{\lambda} \mathbf{K}_a^{-1} \mathbf{K}_b \boldsymbol{\alpha}_b \tag{6.23}$$

and so

$$\mathbf{K}_b^2 \boldsymbol{\alpha}_b - \lambda^2 \mathbf{K}_b^2 \boldsymbol{\alpha}_b = \mathbf{0}.$$

This equation will hold for all vectors $\boldsymbol{\alpha}_b$ with $\lambda = 1$. Hence, we are able to find perfect correlations between arbitrary projections in F_b and an appropriate choice of the projection in F_a. Clearly these correlations are failing to distinguish spurious features from those capturing the underlying semantics. This is perhaps most clearly demonstrated if we consider a random permutation σ of the examples for the second projections to create the vectors

$$\left(\phi_a(\mathbf{x}_i), \phi_b(\mathbf{x}_{\sigma(i)})\right), i = 1, \ldots, \ell.$$

The kernel matrix $\mathbf{K}_a$ will be unchanged and hence still invertible. We are therefore still able to find perfect correlations even though the underlying semantics are no longer correlated in the two representations.

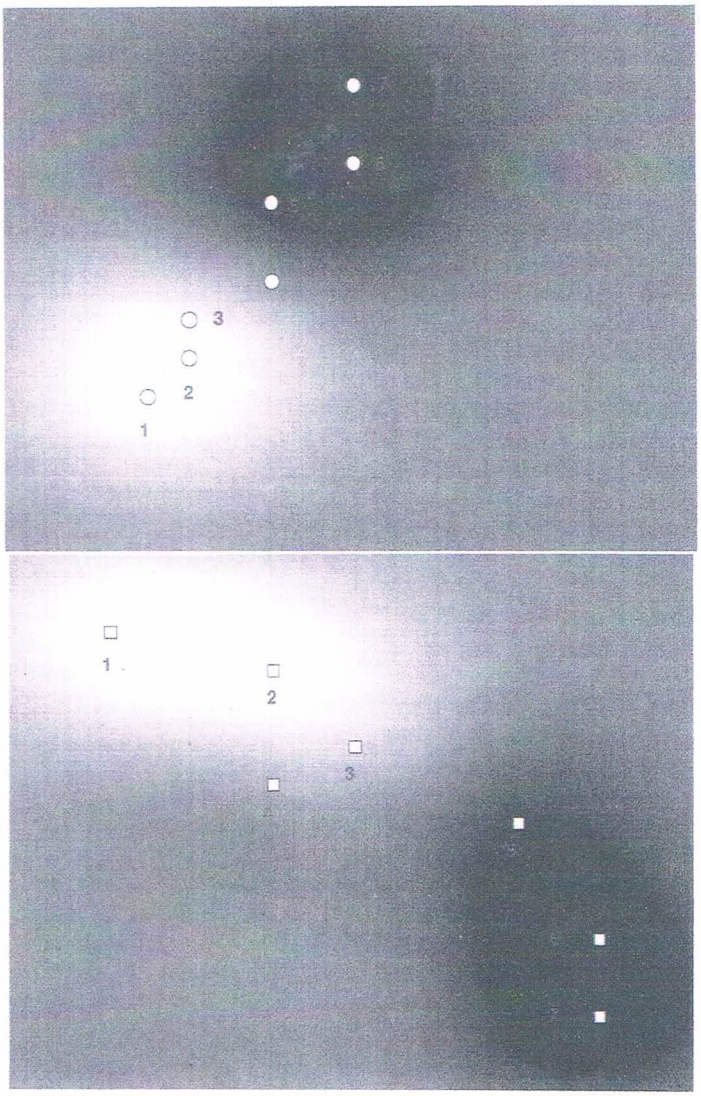

Fig. 6.4. Two feature spaces for a paired dataset with shading indicating the value of the projection onto the first correlation direction.

These observations show that the class of pattern functions we have selected are too flexible. We must introduce some regularisation to control the flexibility. We must, therefore, investigate the statistical stability of CCA, if we are to ensure that meaningful patterns are found.

Stability analysis of CCA Maximising correlation corresponds to minimising the empirical expectation of the pattern function

$$g_{\mathbf{w}_a,\mathbf{w}_b}(\mathbf{x}) = \left\|\mathbf{w}_a'\phi_a(\mathbf{x}) - \mathbf{w}_b'\phi_b(\mathbf{x})\right\|^2,$$

subject to the same conditions, since

$$\begin{aligned}\hat{\mathbb{E}}\left[\left\|\mathbf{w}_a'\phi_a(\mathbf{x}) - \mathbf{w}_b'\phi_b(\mathbf{x})\right\|^2\right] &= \hat{\mathbb{E}}\left[\left\|\mathbf{w}_a'\phi_a(\mathbf{x})\right\|^2\right] + \hat{\mathbb{E}}\left[\left\|\mathbf{w}_b'\phi_b(\mathbf{x})\right\|^2\right] - \\ &\quad 2\hat{\mathbb{E}}\left[\left\langle\mathbf{w}_a'\phi_a(\mathbf{x}), \mathbf{w}_b'\phi_b(\mathbf{x})\right\rangle\right] \\ &= 2\left(1 - \mathbf{w}_a'\mathbf{C}_{ab}\mathbf{w}_b\right).\end{aligned}$$

The function $g_{\mathbf{w}_a,\mathbf{w}_b}(\mathbf{x}) \approx 0$ captures the property of the pattern that we are seeking. It assures us that the feature $\mathbf{w}_a'\phi_a(\mathbf{x})$ that can be obtained from one view of the data is almost identical to $\mathbf{w}_b'\phi_b(\mathbf{x})$ computable from the second view. Such pairs of features are therefore able to capture underlying properties of the data that are present in both views. If our assumption is correct, that what is essential is common to both views, then these features must be capturing some important properties. We can obtain a stability analysis of the function by simply viewing $g_{\mathbf{w}_a,\mathbf{w}_b}(\mathbf{x})$ as a regression function, albeit with special structure, attempting to learn the constant 0 function. Applying the standard Rademacher bound, observe that the empirical expected value of $g_{\mathbf{w}_a,\mathbf{w}_b}(\mathbf{x})$ is simply $2\left(1 - \mathbf{w}_a'\mathbf{C}_{ab}\mathbf{w}_b\right)$. Furthermore, we can use the same technique as that described in Theorem A.3 of Appendix A.2 to represent the function as a linear function in the *feature space determined by the quadratic kernel*

$$\hat{\kappa}(\mathbf{x},\mathbf{z}) = \left(\kappa_a(\mathbf{x},\mathbf{z}) + \kappa_b(\mathbf{x},\mathbf{z})\right)^2,$$

with norm-squared

$$2\left\|\mathbf{w}_a\mathbf{w}_b'\right\|_F^2 = 2\,\mathrm{tr}\left(\mathbf{w}_b\mathbf{w}_a'\mathbf{w}_a\mathbf{w}_b'\right) = \left\|\mathbf{w}_a\right\|^2\left\|\mathbf{w}_b\right\|^2.$$

This gives the following theorem.

Theorem 6.33 *Fix A and B in $\mathbb{R}^+$. If we obtain a feature given by the pattern function $g_{\mathbf{w}_a,\mathbf{w}_b}(\mathbf{x})$ with $\|\mathbf{w}_a\| \leq A$ and $\|\mathbf{w}_b\| \leq B$, on a paired training set S of size ℓ in the feature space defined by the kernels κ_a and κ_b drawn i.i.d. according to a distribution $\mathcal{D}$, then with probability greater than $1-\delta$ over the generation of S, the expected value of $g_{\mathbf{w}_a,\mathbf{w}_b}(\mathbf{x})$ on new data is bounded by*

$$\mathbb{E}_{\mathcal{D}}\left[g_{\mathbf{w}_a,\mathbf{w}_b}(\mathbf{x})\right] \leq 2\left(1 - \mathbf{w}_a'\mathbf{C}_{ab}\mathbf{w}_b\right) +$$

$$\sqrt{\frac{2AB}{\ell}\sum_{i=1}^{\ell}\left(\kappa_a(\mathbf{x}_i,\mathbf{x}_i)+\kappa_b(\mathbf{x}_i,\mathbf{x}_i)\right)^2}+3R^2\sqrt{\frac{\ln(2/\delta)}{2\ell}},$$

where

$$R^2=\max_{\mathbf{x}\in\,\mathrm{supp}(\mathcal{D})}\left(\kappa_a(\mathbf{x},\mathbf{x})+\kappa_b(\mathbf{x},\mathbf{x})\right).$$

The theorem indicates that the empirical value of the pattern function will be close to its expectation, provided that the norms of the two direction vectors are controlled. Hence, we must trade-off between finding good correlations while not allowing the norms to become too large.

Regularisation of CCA Theorem 6.33 shows that the quality of the generalisation of the associated pattern function is controlled by the product of the norms of the weight vectors $\mathbf{w}_a$ and $\mathbf{w}_b$. We therefore introduce a penalty on the norms of these weight vectors. This gives rise to the primal optimisation problem.

Computation 6.34 [Regularised CCA] The regularised version of CCA is solved by the optimisation:

$$\begin{aligned}&\max_{\mathbf{w}_a,\mathbf{w}_b}\rho\left(\mathbf{w}_a,\mathbf{w}_b\right) \qquad (6.24)\\ =\;&\frac{\mathbf{w}_a'\mathbf{C}_{ab}\mathbf{w}_b}{\sqrt{\left((1-\tau_a)\,\mathbf{w}_a'\mathbf{C}_{aa}\mathbf{w}_a+\tau_a\left\|\mathbf{w}_a\right\|^2\right)\left((1-\tau_b)\,\mathbf{w}_b'\mathbf{C}_{bb}\mathbf{w}_b+\tau_b\left\|\mathbf{w}_b\right\|^2\right)}},\end{aligned}$$

where the two regularisation parameters τ_a and τ_b control the flexibility in the two feature spaces. ■

Notice that τ_a, τ_b interpolate smoothly between the maximisation of the correlation and the maximisation of the covariance described in Section 6.3. Dualising we arrive at the following optimisation problem.

Computation 6.35 [Kernel regularised CCA] The dual regularised CCA is solved by the optimisation

$$\begin{array}{ll}\max_{\boldsymbol{\alpha}_a,\boldsymbol{\alpha}_b} & \boldsymbol{\alpha}_a'\mathbf{K}_a\mathbf{K}_b\boldsymbol{\alpha}_b\\ \text{subject to} & (1-\tau_a)\,\boldsymbol{\alpha}_a'\mathbf{K}_a^2\boldsymbol{\alpha}_a+\tau_a\boldsymbol{\alpha}_a'\mathbf{K}_a\boldsymbol{\alpha}_a=1\\ & \text{and } (1-\tau_b)\,\boldsymbol{\alpha}_b'\mathbf{K}_b^2\boldsymbol{\alpha}_b+\tau_b\boldsymbol{\alpha}_b'\mathbf{K}_b\boldsymbol{\alpha}_b=1.\end{array}$$

■

Note that as with ridge regression we regularised by penalising the norms of the weight vectors. Nonetheless, the resulting form of the equations obtained does not in this case correspond to a simple addition to the diagonal of the kernel matrix, the so-called ridge of ridge regression.

Solving dual regularised CCA Using the Lagrangian technique, we can now obtain the equations

$$\begin{aligned} \mathbf{K}_a\mathbf{K}_b\boldsymbol{\alpha}_b - \lambda(1-\tau_a)\mathbf{K}_a^2\boldsymbol{\alpha}_a - \lambda\tau_a\mathbf{K}_a\boldsymbol{\alpha}_a &= \mathbf{0} \\ \text{and } \mathbf{K}_b\mathbf{K}_a\boldsymbol{\alpha}_a - \lambda(1-\tau_b)\mathbf{K}_b^2\boldsymbol{\alpha}_b - \lambda\tau_b\mathbf{K}_b\boldsymbol{\alpha}_b &= \mathbf{0}, \end{aligned}$$

hence forming the generalised eigenvalue problem

$$\begin{pmatrix} \mathbf{0} & \mathbf{K}_a\mathbf{K}_b \\ \mathbf{K}_b\mathbf{K}_a & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_a \\ \boldsymbol{\alpha}_b \end{pmatrix} = \lambda \begin{pmatrix} (1-\tau_a)\mathbf{K}_a^2 + \tau_a\mathbf{K}_a & \mathbf{0} \\ \mathbf{0} & (1-\tau_b)\mathbf{K}_b^2 + \tau_b\mathbf{K}_b \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_a \\ \boldsymbol{\alpha}_b \end{pmatrix}.$$

One difficulty with this approach can be the size of the resulting generalised eigenvalue problem, since it will be twice the size of the training set. A method of tackling this is to use the partial Gram–Schmidt orthonormalisation of the data in the feature space to form a lower-dimensional approximation to the feature representation of the data. As described in Section 5.2 this is equivalent to performing an incomplete Cholesky decomposition of the kernel matrices

$$\mathbf{K}_a = \mathbf{R}_a'\mathbf{R}_a \text{ and } \mathbf{K}_b = \mathbf{R}_b'\mathbf{R}_b,$$

with the columns of $\mathbf{R}_a$ and $\mathbf{R}_b$ being the new feature vectors of the training points in the orthonormal basis created by the Gram–Schmidt process. Performing an incomplete Cholesky decomposition ensures that $\mathbf{R}_a \in \mathbb{R}^{n_a \times \ell}$ has linearly independent rows so that $\mathbf{R}_a\mathbf{R}_a'$ is invertible. The same holds for $\mathbf{R}_b\mathbf{R}_b'$ with $\mathbf{R}_b \in \mathbb{R}^{n_b \times \ell}$.

We can now view our problem as a primal canonical correlation analysis with the feature vectors given by the columns of $\mathbf{R}_a$ and $\mathbf{R}_b$. This leads to the equations

$$\begin{aligned} \mathbf{R}_a\mathbf{R}_b'\mathbf{w}_b - \lambda(1-\tau_a)\mathbf{R}_a\mathbf{R}_a'\mathbf{w}_a - \lambda\tau_a\mathbf{w}_a &= \mathbf{0} \\ \text{and } \mathbf{R}_b\mathbf{R}_a'\mathbf{w}_a - \lambda(1-\tau_b)\mathbf{R}_b\mathbf{R}_b'\mathbf{w}_b - \lambda\tau_b\mathbf{w}_b &= \mathbf{0}. \end{aligned} \tag{6.25}$$

From the first equation, we can now express $\mathbf{w}_a$ as

$$\mathbf{w}_a = \frac{1}{\lambda}\left((1-\tau_a)\mathbf{R}_a\mathbf{R}_a' + \tau_a\mathbf{I}\right)^{-1}\mathbf{R}_a\mathbf{R}_b'\mathbf{w}_b,$$

which on substitution in the second gives the normal (albeit non-symmetric) eigenvalue problem

$$\left((1-\tau_b)\,\mathbf{R}_b\mathbf{R}_b' + \tau_b\mathbf{I}\right)^{-1}\mathbf{R}_b\mathbf{R}_a'\left((1-\tau_a)\,\mathbf{R}_a\mathbf{R}_a' + \tau_a\mathbf{I}\right)^{-1}\mathbf{R}_a\mathbf{R}_b'\mathbf{w}_b = \lambda^2\mathbf{w}_b$$

of dimension $n_b \times n_b$. After performing a full Cholesky decomposition

$$\mathbf{R}'\mathbf{R} = \left((1-\tau_b)\,\mathbf{R}_b\mathbf{R}_b' + \tau_b\mathbf{I}\right)$$

of the non-singular matrix on the right hand side, we then take

$$\mathbf{u}_b = \mathbf{R}\mathbf{w}_b,$$

which using the fact that the transpose and inversion operations commute leads to the equivalent symmetric eigenvalue problem

$$\left(\mathbf{R}'\right)^{-1}\mathbf{R}_b\mathbf{R}_a'\left((1-\tau_a)\,\mathbf{R}_a\mathbf{R}_a' + \tau_a\mathbf{I}\right)^{-1}\mathbf{R}_a\mathbf{R}_b'\mathbf{R}^{-1}\mathbf{u}_b = \lambda^2\mathbf{u}_b.$$

By symmetry we could have created an eigenvalue problem of dimension $n_a \times n_a$. Hence, the size of the eigenvalue problem can be reduced to the smaller of the two partial Gram–Schmidt dimensions.

We can of course recover the full unapproximated kernel canonical correlation analysis if we simply choose $n_a = \operatorname{rank}(\mathbf{K}_a)$ and $n_b = \operatorname{rank}(\mathbf{K}_b)$. Even in this case we have avoided the need to solve a generalised eigenvalue problem, while at the same time reducing the dimension of the problem by at least a factor of two since $\min(n_a, n_b) \leq \ell$. The overall algorithm is as follows.

Algorithm 6.36 [Kernel CCA] Kernel canonical correlation analysis can be solved as shown in Code Fragment 6.3. ■

This means that we can have two views of an object that together create a paired dataset S through two different representations or kernels. We use this procedure to compute correlations between the two sets that are stable in the sense that they capture properties of the underlying distribution rather than of the particular training set or view.

Remark 6.37 [Bilingual corpora] Example 6.28 has already mentioned as examples of paired datasets so-called parallel corpora in which each document appears with its translation to a second language. We can apply the kernel canonical correlation analysis to such a corpus using kernels for text that will be discussed in Chapter 10. This will provide a means of projecting documents from either language into a common semantic space. ■

Input	kernel matrices $\mathbf{K}_a$ and $\mathbf{K}_b$ with parameters τ_a and τ_b
Process	Perform (incomplete) Cholesky decompositions: $\mathbf{K}_a = \mathbf{R}_a'\mathbf{R}_a$ and $\mathbf{K}_b = \mathbf{R}_b'\mathbf{R}_b$ of dimensions n_a and n_b: perform a complete Cholesky decomposition: $(1-\tau_b)\,\mathbf{R}_b\mathbf{R}_b' + \tau_b\mathbf{I} = \mathbf{R}'\mathbf{R}$ solve the eigenvalue problem: $(\mathbf{R}')^{-1}\,\mathbf{R}_b\mathbf{R}_a'\left((1-\tau_a)\,\mathbf{R}_a\mathbf{R}_a' + \tau_a\mathbf{I}\right)^{-1}\mathbf{R}_a\mathbf{R}_b'\mathbf{R}^{-1}\mathbf{u}_b = \lambda^2\mathbf{u}_b$. to give each λ_j, $\mathbf{u}_b^j$ compute $\mathbf{w}_b^j = \mathbf{R}^{-1}\mathbf{u}_b$, $\mathbf{w}_b^j = \mathbf{w}_b^j/\|\mathbf{w}_b^j\|$ $\mathbf{w}_a^j = \frac{1}{\lambda_j}\left((1-\tau_a)\,\mathbf{R}_a\mathbf{R}_a' + \tau_a\mathbf{I}\right)^{-1}\mathbf{R}_a\mathbf{R}_b'\mathbf{w}_b^j$ $\mathbf{w}_a^j = \mathbf{w}_a^j/\|\mathbf{w}_a^j\|$
Output	eigenvectors and values $\mathbf{w}_a^j$, $\mathbf{w}_b^j$ and $\lambda_j > 0$.. $j = 1,\ldots,\min(n_a, n_b)$

Code Fragment 6.3. Pseudocode for the kernel CCA algorithm.

Remark 6.38 [More than 2 representations] Notice that a simple manipulation of equation (6.22) gives the alternative formulation

$$\begin{pmatrix} C_{aa} & C_{ab} \\ C_{ba} & C_{bb} \end{pmatrix} \begin{pmatrix} \mathbf{w}_a \\ \mathbf{w}_b \end{pmatrix} = (1+\lambda) \begin{pmatrix} C_{aa} & 0 \\ 0 & C_{bb} \end{pmatrix} \begin{pmatrix} \mathbf{w}_a \\ \mathbf{w}_b \end{pmatrix}$$

which suggests a natural generalisation, namely seeking correlations between three or more views. Given k multivariate random variables, it reduces to the generalised eigenvalue problem

$$\begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} & \cdots & \mathbf{C}_{1k} \\ \mathbf{C}_{21} & \mathbf{C}_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_{k1} & \cdots & \cdots & \mathbf{C}_{kk} \end{pmatrix} \begin{pmatrix} \mathbf{w}_1 \\ \vdots \\ \vdots \\ \mathbf{w}_k \end{pmatrix}$$

$$= \rho \begin{pmatrix} \mathbf{C}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{C}_{kk} \end{pmatrix} \begin{pmatrix} \mathbf{w}_1 \\ \vdots \\ \vdots \\ \mathbf{w}_k \end{pmatrix}.$$

where we use C_{ij} to denote the covariance matrix between the ith and jth views. Note that for $k > 2$ there is no obvious way of reducing such a generalised eigenvalue problem to a lower-dimensional eigenvalue problem as was possible using the Cholesky decomposition in the case $k = 2$. ■