Lesson 6

Bridging Input and Output

In the previous chapter it was shown that (one) thing plaguing PCA is its exclusive emphasis on the input variables. The next step to take is then to connect the output variables in the analysis. But, indeed, there are various ways to combine the inputs and outputs. In this chapter, two strategies from the other ends of the scientific community are studied — the first of them, *Partial Least Squares,* seems to be very popular today among chemical engineers. This approach is pragmatic, usually presented in an algorithmic form[1]. The second one, *Canonical Correlation Analysis,* has been extensively studied among statisticians, but it seems to be almost unknown among practicing engineers. However, both of these methods share very similar ideas and structure — even though the properties of the resulting models can be very different.

6.1 Partial least squares

The *Partial Least Squares (PLS)*[2] regression method has been used a lot lately, specially for *calibration* tasks in chemometrics [31],[38]. In this section, a *different* approach to PLS is taken as compared to usual practices, only honoring the very basic ideas. The reason for this is to keep the discussion better comprehensible, sticking to the already familiar eigenproblem-oriented framework.

6.1.1 Maximizing correlation

The problem with PCA approach is that it concentrates exclusively on the input data X, not taking into account the output data Y. It is not actually the data variance one wants to capture, it is the *correlation between* X *and* Y that should be maximized.

The derivation of the PLS basis vectors can be carried out as in the PCA case,

[1] PLS is sometimes characterized as being one of those "try and pray" methods; the reason for this is — it can be claimed — that a practicing engineer simply cannot grasp the unpenetrable algorithmic presentation of the PLS ideas. He/she can just use the available toolboxes and hope for the best

[2] Sometimes called also *Projection onto Latent Structure*

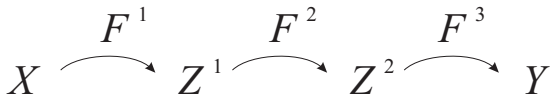

Figure 6.1: The dependency model $y = f(x)$ refined

now concentrating on correlation rather than variance. The procedure becomes slightly more complex than in the PCA case: It is not only the input X block that is restructured, but the internal structure of the output Y block is also searched for. The regression procedure becomes such that the X data is first projected onto a lower dimensional X oriented subspace spanned by the basis vectors θ_i; after that, data is projected onto the Y oriented subspace spanned by the basis vectors ϕ_i, and only after that, the final projection onto the Y space is carried out.

The objective now is to find the basis vectors θ_i and ϕ_i so that the correlation between the projected data vectors $X\theta_i$ and $Y\phi_i$ is maximized while the lengths of the basis vectors are kept constant. This objective results in the constrained optimization problem (1.27) where

$$\left\{ \begin{array}{rcll} f(\theta_i, \phi_i) & = & \frac{1}{k} \cdot \theta_i^T X^T \cdot Y\phi_i, & \text{when} \\ g_1(\theta_i) & = & 1 - \theta_i^T \theta_i & \text{and} \\ g_2(\phi_i) & = & 1 - \phi_i^T \phi_i. & \end{array} \right. \tag{6.1}$$

There are now two separate constraints, g_1 and g_2; defining the corresponding Lagrange multipliers η_i and μ_i gives the Hamiltonian

$$\frac{1}{k} \cdot \theta_i^T X^T \cdot Y\phi_i - \eta_i \left(1 - \theta_i^T \theta_i\right) - \mu_i \left(1 - \phi_i^T \phi_i\right), \tag{6.2}$$

and differentiation gives

$$\left\{ \begin{array}{l} \frac{d}{d\theta_i} \left(\frac{1}{k} \cdot \theta_i^T X^T \cdot Y\phi_i - \eta_i(1 - \theta_i^T \theta_i) - \mu_i(1 - \phi_i^T \phi_i) \right) = 0 \\ \frac{d}{d\phi_i} \left(\frac{1}{k} \cdot \theta_i^T X^T \cdot Y\phi_i - \eta_i(1 - \theta_i^T \theta_i) - \mu_i(1 - \phi_i^T \phi_i) \right) = 0, \end{array} \right. \tag{6.3}$$

resulting in a pair of equations

$$\left\{ \begin{array}{l} \frac{1}{k} \cdot X^T Y\phi_i - 2\eta_i \theta_i = 0 \\ \frac{1}{k} \cdot Y^T X\theta_i - 2\mu_i \phi_i = 0. \end{array} \right. \tag{6.4}$$

Solving the first of these for θ_i and the second for ϕ_i, the following equations can be written:

$$\left\{ \begin{array}{l} \frac{1}{k^2} \cdot X^T Y Y^T X\theta_i = 4\eta_i \mu_i \cdot \theta_i \\ \frac{1}{k^2} \cdot Y^T X X^T Y\phi_i = 4\eta_i \mu_i \cdot \phi_i. \end{array} \right. \tag{6.5}$$

This means that, again, as in Sec. 5.2, the best basis vectors are given as solutions to eigenvalue problems; the significance of the vectors θ_i (for the X block) and ϕ_i (for the Y block) is revealed by the corresponding eigenvalues $\lambda_i = 4\eta_i \mu_i$.

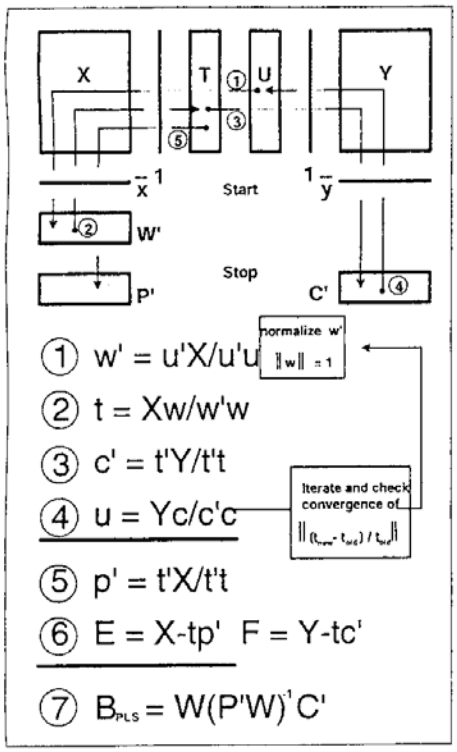

Figure 6.2: What is usually meant by "PLS": The *Algorithm*

Because the matrices X^TYY^TX and Y^TXX^TY are symmetric, the orthogonality properties again apply to their eigenvectors. The expression (5.36) can directly be utilized; the internal basis θ_{PLS} consists of a subset of eigenvectors, selection of these basis vectors being again based on the numeric values of the corresponding eigenvalues. In practice, the basis vectors ϕ_i are redundant and they need not be explicitly calculated (see Sec. 6.3.3). Because the rank of a product of matrices cannot exceed the ranks of the multiplied matrices, there will be only $\min\{n, m\}$ non-zero eigenvalues; that is why, the PCR approach may give higher dimensional models than PLS (when applying this eigenproblem oriented approach).

It should be recognized that the PLS model is usually constructed in another way (for example, see [31]); this "other way" may sometimes result in better models, but it is extremely uninstructive and implicit, being defined through an iterative algorithm (see Fig. 6.2). It can be shown that the two approaches exactly coincide only what comes to the *most significant* basis vector; other basis vectors can differ. For example, applying the approach based on the eigenvectors, the number of non-zero eigenvalues cannot exceed the number of variables in either input or the output — this means that the latent basis dimension is restricted so that $N \leq m$. Such constraint does not apply to the iterative PLS approach.

Let us study the example that was presented in the previous chapter, now in the PLS framework. The output is scalar; it is assumed that it is linearly dependent of the second input variable, so that $y(\kappa) = f \cdot x_2(\kappa)$, where f is a constant.

The matrix in (5.45) becomes

$$\begin{aligned}
&\tfrac{1}{k^2} \cdot X^T Y Y^T X \\
&= \tfrac{1}{k^2} \cdot \begin{pmatrix} \sum_\kappa x_1(\kappa)y(\kappa) \sum_\kappa x_1(\kappa)y(\kappa) & \sum_\kappa x_1(\kappa)y(\kappa) \sum_\kappa x_2(\kappa)y(\kappa) \\ \sum_\kappa x_1(\kappa)y(\kappa) \sum_\kappa x_2(\kappa)y(\kappa) & \sum_\kappa x_2(\kappa)y(\kappa) \sum_\kappa x_2(\kappa)y(\kappa) \end{pmatrix} \\
&\approx \begin{pmatrix} \mathrm{E}^2\{x_1(\kappa)y(\kappa)\} & \mathrm{E}\{x_1(\kappa)y(\kappa)\} \cdot \mathrm{E}\{x_2(\kappa)y(\kappa)\} \\ \mathrm{E}\{x_1(\kappa)y(\kappa)\} \cdot \mathrm{E}\{x_2(\kappa)y(\kappa)\} & \mathrm{E}^2\{x_2(\kappa)y(\kappa)\} \end{pmatrix} \\
&= \begin{pmatrix} 0 & 0 \\ 0 & f^2 \cdot E^2\{x_2^2(\kappa)\} \end{pmatrix},
\end{aligned}$$

because x_1 and y are not assumed to correlate. This result reveals that the maximum eigenvalue is $f^2 \cdot \mathrm{E}^2\{x_2^2(\kappa)\}$ corresponding to the *second* input variable — no matter what is the ratio between the variances of x_1 and x_2. This means that the basis always includes the vector $(\ 0 \quad 1\ )^T$ — and according to the assumed dependency structure, this is exactly what is needed to construct a working regression model. As a matter of fact, it can be seen that the eigenvalue corresponding to the first input variable is zero, reflecting the fact that x_1 has no effect on y whatsoever.

6.2 Continuum regression

6.2.1 On the correlation structure

Let us study the correlation between input and output from yet another point of view. The correlation structure is captured by the (unnormalized) cross-correlation matrix

$$X^T Y. \tag{6.6}$$

The eigenvalues and eigenvectors are already familiar to us, and it has been shown how useful they are in the analysis of matrix structures. Perhaps one could use the same approaches to analysis of this correlation matrix? However, this matrix is generally not square and the eigenstructure cannot be determined; but the *singular value decomposition,* the geralization of the eigenvalue decomposition exists (see Sec. 1.2.2)

$$X^T Y = \Theta_{XY} \Sigma_{XY} \Phi_{XY}^T. \tag{6.7}$$

Here Θ_{XY} and Φ_{XY} are orthogonal matrices, the first being compatible with X and the other being compatible with Y; Σ_{XY} is a diagonal matrix, but if the input and output dimensions do not match, it is not square. Multiplying (6.7) by its transpose either from left or right, the orthonormality of Θ_{XY} and Φ_{XY} (so that $\Theta_{XY}^T = \Theta_{XY}^{-1}$ and $\Phi_{XY}^T = \Phi_{XY}^{-1}$) means that there holds

$$X^T Y Y^T X = \Theta_{XY} \Sigma_{XY} \Sigma_{XY}^T \Theta_{XY}^{-1} \tag{6.8}$$

and

$$Y^T X X^T Y = \Phi_{XY} \Sigma_{XY}^T \Sigma_{XY} \Phi_{XY}^{-1}. \tag{6.9}$$

Because $\Sigma_{XY}\Sigma_{XY}^T$ and $\Sigma_{XY}^T\Sigma_{XY}$ are diagonal square matrices, these two expressions are eigenvalue decompositions (1.5) of the matrices X^TYY^TX and Y^TXX^TY, respectively. This means that there is a connection between the singular value decomposition and the above PLS basis vectors: The matrices Θ_{XY} and Φ_{XY} consist of the (full sets) of PLS basis vectors θ_i and ϕ_i. The diagonal elements of Σ_{XY}, the singular values, are related to the PLS eigenvalues in such a way that $\sigma_i = k \cdot \sqrt{\lambda_i}$.

What is more, one can see that the SVD of the input data block X alone is similarly closely related to the PCA constructs:

$$X^T = \Theta_X \Sigma_X \Phi_X^T, \tag{6.10}$$

so that

$$X^TX = \Theta_X \Sigma_X \Sigma_X^T \Theta_X^{-1}, \tag{6.11}$$

meaning that, again, the singular value decomposition does the trick, principal components being collected in Θ_X and singular values being related to the eigenvalues through $\sigma_i = \sqrt{k \cdot \lambda_i}$.

6.2.2 Filling the gaps

What if one defines the matrix[3]

$$(X^T)^{\alpha_X}(Y)^{\alpha_Y}, \tag{6.12}$$

so that both of the analysis methods, PCA and PLS, would be received by selecting the parameters α_X and α_Y appropriately (for PCA, select $\alpha_X = 1$ and $\alpha_Y = 0$, and for PLS, select $\alpha_X = 1$ and $\alpha_Y = 1$), and applying SVD? And, further, why not try other values for α_X and α_Y for emphasizing the input and output data in different ways in the model construction? Indeed, there is a continuum between PCA and PLS — and this is not the whole story: Letting the ratio α_X/α_Y go towards zero, we go beyond PLS, towards models where the role of the output is emphasized more and more as compared to the input, finally constructing an singular value decomposition for Y alone (or eigenvalue decomposition for Y^TY).

It is only the ratio between α_X and α_Y that is relevant; we can eliminate the other of them, for example, by fixing $\alpha_X = 1$. Representing the problem in the familiar eigenproblem framework, multiplying (6.12) from left by its transpose and compensating the number of samples appropriately one has the eigenproblem formulation for the *Continuum Regression (CR)* basis vectors defined as[4]

$$\frac{1}{k^{1+\alpha}} \cdot X^T \left(YY^T\right)^{\alpha} X \cdot \theta_i = \lambda_i \cdot \theta_i. \tag{6.13}$$

[3]The powers of non-square matrices being defined as shown in Sec. 1.2.2

[4]These eigenproblems should not be solved directly in this form: The matrix XXT has dimension $k \times k$, even though there are only n non-zero eigenvalues (or singular values)

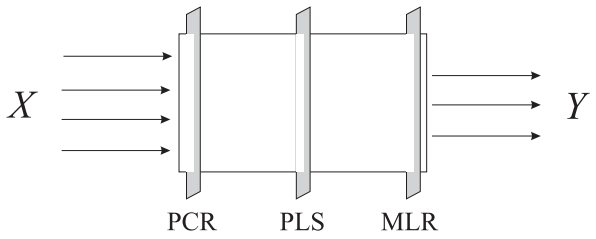

Figure 6.3: Schematic illustration of the relation between regression approaches

Correspondingly, the "dual" formulation becomes

$$\frac{1}{k^{1+1/\alpha}} \cdot Y^T \left(XX^T\right)^{\frac{1}{\alpha}} Y \cdot \phi_i = \lambda_i' \cdot \phi_i. \tag{6.14}$$

When α grows from 0 towards ∞, the modeling emphasis is first put exclusively on the input data, and finally exclusively on the output data (see Fig. 6.3); some special values of α do have familiar interpretations:

- If $\alpha = 0$, the PCA model results, only input being emphasized.
- If $\alpha = 1$, the PLS model results, input and output being in balance.
- If $\alpha \to \infty$, an "MLR type" model results, only output being emphasized[5].

Which of the regression approaches, MLR, PCR, or PLS, is the best, cannot be determined beforehand; it depends on the application and available data. All of these methods have only mathematical justification; from the physical point of view, none of them can be said to always outperform the others. It may even be so that the ultimate optimum model lies somewhere on the continuum between PCR, PLS, and MLR (it may also lie somewhere else outside the continuum).

In Figs. 6.4 and 6.5, the CR performance is visualized: There were 30 machine-generated data samples with 20 input and 20 output variables; the number of independent input variables was 10 and the "correct" dimension of the output was 5; relatively high level of noise was added. And, indeed, it seems that when the cross-validation error is plotted as the function of latent variables N and continuum parameter α as a two-dimensional map, interesting behavior is revealed: Starting from $\alpha = 0$, the minimum error is reached for about $N = 12$ whereas the overall optimum is found near MLR with $N = 6$.

6.2.3 Further explorations*

It needs to be emphasized again that there are typically no absolutely correct methods for determininf physically optimal latent basis vectors. As in the whole report, the goal here is to show that there is plenty of room for experimenting

[5]Note that MLR is not based on basis vectors; that is why, the correspondence is somewhat artificial (the first basis vector of the CR model explaining the first principal component of the output data, thus explaining maximum amount of the output variance)

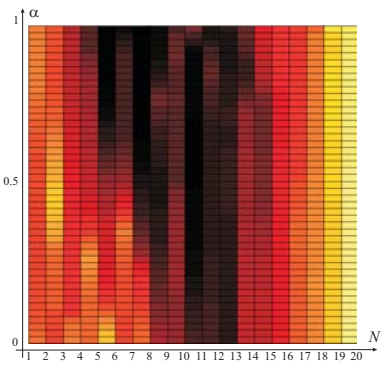

Figure 6.4: Continuum regression performance for different parameter values N and α

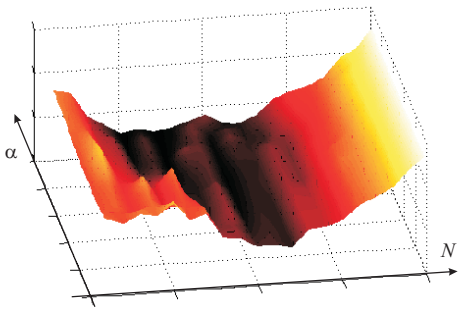

Figure 6.5: Continuum regression performance as a "mountain view"

and research (after all, the history of CR is less than ten years long; by no means one should assume that the final word has been said). For example, a whole class of methods can be defined that share the idea of continuum regression. Let us study a slightly different approach.

MLR can be interpreted as modeling the covariance structure of the estimated output Y. The problem that emerges is that the output space usually does not have the same dimension as the input space has; that is why, the output variations need to be somehow presented in the input space to make this approach compatible with the other ones, PCR and PLS. The outputs can be projected into the input space by applying MLR in the "inverse direction", that is, $\hat{X} = Y \cdot (Y^T Y)^{-1} Y^T X$, so that the covariance to be modeled has the form

$$\begin{aligned} \frac{1}{k} \cdot \hat{X}^T \hat{X} &= \frac{1}{k} \cdot X^T Y (Y^T Y)^{-1} \cdot Y^T Y \cdot (Y^T Y)^{-1} Y^T X \\ &= \frac{1}{k} \cdot X^T Y (Y^T Y)^{-1} Y^T X. \end{aligned} \tag{6.15}$$

Actually, this formulation gives a new "latent structure" oriented view of MLR. Assuming that all eigenvectors are utilized, the normal MLR results (of course, this is true for all latent variables based methods if all latent variables are employed), but if a lower dimensional internal model is constructed, the output properties are preserved based on their "visibility" in Y. It turns ot that if one defines the latent vectors θ_i as

$$\frac{1}{k^{1+\alpha_1(1+\alpha_2)}} \cdot X^T \left(Y \left(Y^T Y \right)^{\alpha_2} Y^T \right)^{\alpha_1} X \cdot \theta_i = \lambda_i \cdot \theta_i, \tag{6.16}$$

all of the above regression methods can be simulated by appropriately selecting the parameters α_1 and α_2:

- PCR is given by $\alpha_1 = 0$, whereas parameter α_2 can have any value;
- PLS results if $\alpha_1 = 1$ and $\alpha_2 = 0$; and
- MLR is found if $\alpha_1 = 1$ and $\alpha_2 = -1$.

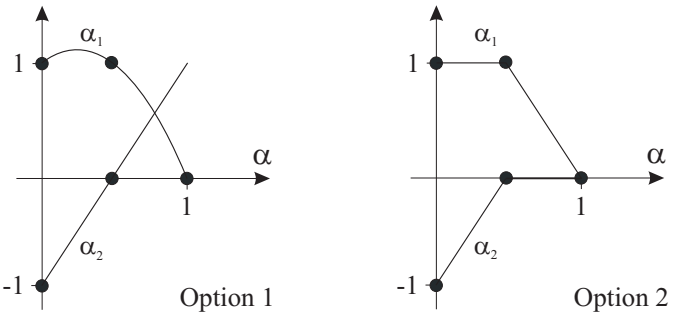

Figure 6.6: Two alternative feasible function forms (see text)

We would like to have a single parameter α spanning the continuum between the approaches, so that $\alpha = 0$ would give MLR, $\alpha = 1/2$ would give PLS, and $\alpha = 1$ would give PCR (note that the range of α is now from 0 to 1). There is an infinity of alternative options — for example, the following definitions fulfill our needs:

1. $\alpha_1 = -2\alpha^2 + \alpha + 1$ and $\alpha_2 = 2\alpha - 1$, or
2. $\alpha_1 = \frac{3}{2} - \alpha - |\alpha - \frac{1}{2}|$ and $\alpha_2 = -\frac{1}{2} + \alpha - |\alpha - \frac{1}{2}|$.

The outlooks of these functions are presented in Fig. 6.6. As an example, selecting the option 1 above, the latent vectors of CR can be calculated as solutions to the following eigenproblem:

$$\frac{1}{k^\beta} \cdot X^T \left(Y \left(Y^T Y \right)^{2\alpha - 1} Y^T \right)^{-2\alpha^2 + \alpha + 1} X \cdot \theta = \lambda \cdot \theta. \tag{6.17}$$

Here, the parameter β can be selected as $\beta = -4\alpha^3 + 2\alpha^2 + 2\alpha + 1$ to compensate for the changes in the number of samples. It needs to be noted that the outer matrix that one has to calculate the power function of may be very large (dimension being $k \times k$); however, there are only m eigenvalues different from zero, meaning that (in principle) only m power functions have to be calculated. The matrix power is best to calculate using the singular value decomposition.

The basis θ_{CR} is again constructed from the selected eigenvectors; because of the symmetricity of the matrix in (6.17), the basis is orthonormal.

6.3 Canonical correlations

Another approach to modeling the dependency structure between the input and the output is offered by *Canonical Correlation Analysis (CCA)* [32].

6.3.1 Problem formulation

Again, one would like to find the latent basis vectors θ_i and ϕ_i so that the correlation between the input and output blocks would be maximized. The

criterion to be maximized is again

$$f(\theta_i, \phi_i) = \frac{1}{k} \cdot {\theta_i}^T X^T Y \phi_i, \tag{6.18}$$

but the constraints are modified slightly:

$$\begin{cases} g_1(\theta_i) = \frac{1}{k} \cdot {\theta_i}^T X^T X \theta_i = 1 \\ g_2(\phi_i) = \frac{1}{k} \cdot {\phi_i}^T Y^T Y \phi_i = 1. \end{cases} \tag{6.19}$$

Note the difference as compared to the PLS derivation: It is not the basis vector θ_i itself that is kept constant size; it is the projected data vector size $Z_i = X\theta_i$ that is regulated, ${\theta_i}^T X^T \cdot X\theta_i$ being kept constant. The same applies also in the output block: The size of ${\phi_i}^T Y^T \cdot Y\phi_i$ is limited.

Again using the Lagrangian technique the following expression is to be maximized:

$$\frac{1}{k} \cdot {\theta_i}^T X^T Y \phi_i + \eta_i \cdot (1 - \frac{1}{k} \cdot {\theta_i}^T X^T X \theta_i) + \mu_i \cdot (1 - \frac{1}{k} \cdot {\phi_i}^T Y^T Y \phi_i). \tag{6.20}$$

This expression can be minimized with respect to both θ_i and ϕ_i separately:

$$\begin{cases} \frac{1}{k} \cdot \frac{d}{d\theta_i} \left(\theta_i^T X^T Y \phi_i - \eta_i (1 - \theta_i^T X^T X \theta_i) - \mu_i (1 - \phi_i^T Y^T Y \phi_i) \right) = 0 \\ \frac{1}{k} \cdot \frac{d}{d\phi_i} \left(\theta_i^T X^T Y \phi_i - \eta_i (1 - \theta_i^T X^T X \theta_i) - \mu_i (1 - \phi_i^T Y^T Y \phi_i) \right) = 0, \end{cases}$$

resulting in a pair of equations

$$\begin{cases} X^T Y \phi_i - 2\eta_i X^T X \theta_i = 0 \\ Y^T X \theta_i - 2\mu_i Y^T Y \phi_i = 0. \end{cases} \tag{6.21}$$

Solving the first of these for θ_i and the second for ϕ_i, the following equations can be written (assuming invertibility of the matrices):

$$\begin{cases} X^T Y (Y^T Y)^{-1} Y^T X \theta_i = 4\eta_i \mu_i \cdot X^T X \theta_i \\ Y^T X (X^T X)^{-1} X^T Y \phi_i = 4\eta_i \mu_i \cdot Y^T Y \phi_i, \end{cases} \tag{6.22}$$

or

$$\begin{cases} (X^T X)^{-1} X^T Y (Y^T Y)^{-1} Y^T X \cdot \theta_i = 4\eta_i \mu_i \cdot \theta_i \\ (Y^T Y)^{-1} Y^T X (X^T X)^{-1} X^T Y \cdot \phi_i = 4\eta_i \mu_i \cdot \phi_i. \end{cases} \tag{6.23}$$

This means that, again, the best basis vectors are given as solutions to eigenvalue problems; the significance of the vectors θ_i (for the X block) and ϕ_i (for the Y block) is revealed by the corresponding eigenvalues $\lambda_i = 4\eta_i \mu_i$ (note the equivalences of the corresponding eigenvalues in different blocks). If either $X^T X$ or $Y^T Y$ is not invertible, either one of the *generalized eigenvalue problems* in (6.22) can directly be solved.

It needs to be recognized that data must be explicitly scaled in the CCA case[6]: The property $\frac{1}{k} \cdot \theta^T X^T X \theta = I$ is not automatically guaranteed by the eigen-

[6]This kind of extra scaling is not needed in the above PCA and PLS approaches: By construction, the eigenvectors were assumed to be normalized to unit length

problem formulation. The matrix is diagonal (see the next section), but the diagonal elements are ones only after appropriate scalings:

$$\theta_i \leftarrow \theta_i / \sqrt{\frac{1}{k} \cdot \theta_i^T X^T X \theta_i}. \tag{6.24}$$

6.3.2 Analysis of CCA

If the former equation in (6.22) is multiplied from left by θ_j^T, one has

$$\theta_j^T X^T \cdot Y(Y^T Y)^{-1} Y^T \cdot X\theta_i - \lambda_i \cdot \theta_j^T X^T \cdot X\theta_i = 0. \tag{6.25}$$

When rearranged in the above way, one can see that the matrix $Y(Y^T Y)^{-1} Y^T$ is symmetric — meaning that (as in Chapter 5) the eigenproblem can be read in the "inverse" direction, and the following must hold

$$\begin{aligned} &\left(\theta_j^T X^T \cdot Y(Y^T Y)^{-1} Y^T\right) \cdot X\theta_i - \lambda_i \cdot \theta_j^T X^T \cdot X\theta_i \\ &\quad = \lambda_j \cdot \theta_j^T X^T X\theta_i - \lambda_i \cdot \theta_j^T X^T X\theta_i \\ &\quad = (\lambda_j - \lambda_i) \cdot \theta_j^T X^T \cdot X\theta_i \\ &\quad = 0, \end{aligned} \tag{6.26}$$

meaning that $X\theta_i$ and $X\theta_j$ must be orthogonal if $i \neq j$ so that $\theta_i^T X^T X\theta_j = 0$ (remember that for $i = j$ it was assumed that $\theta_i^T X^T X\theta_j = 1$). The same result can be derived for the output block: The projected variables are mutually uncorrelated. Further, if the equations in (6.21) are multiplied from left by θ_j^T and ϕ_j^T, respectively, one has

$$\begin{cases} \theta_j^T X^T Y\phi_i = 2\eta_i \cdot \theta_j^T X^T X\theta_i \\ \phi_j^T Y^T X\theta_i = 2\mu_i \cdot \phi_j^T Y^T Y\phi_i. \end{cases} \tag{6.27}$$

Observing the above uncorrelatedness result, one can conclude that also for the cross-correlations between the projected input and output blocks the same structure has emerged: Only for $j = i$ there is correlation, otherwise not; this correlation coefficient is $2\eta_i = 2\mu_i = \sqrt{\lambda_i}$. The above results can be summarized by showing the correlation structure between the latent input and output bases:

$$\begin{aligned} &\left(\begin{array}{c|c} X\Theta & Y\Phi \end{array}\right)^T \left(\begin{array}{c|c} X\Theta & Y\Phi \end{array}\right) \\ &\quad = \left(\begin{array}{ccc|ccc} 1 & & & \sqrt{\lambda_1} & & \\ & \ddots & & & \ddots & \\ & & 1 & & & \sqrt{\lambda_n} \\ \hline \sqrt{\lambda_1} & & & 1 & & \\ & \ddots & & & \ddots & \\ & & \sqrt{\lambda_n} & & & 1 \end{array}\right). \end{aligned} \tag{6.28}$$

For notational simplicity, it is assumed here that $n = m$ (otherwise, the non-diagonal blocks are padded with zeros). The basis vectors θ_i and ϕ_i are called *canonical variates* corresponding to the *canonical correlations* $\sqrt{\lambda_i}$. The very

elegant structure of (6.28) suggests that there must be going on something more important — the dependencies between the input and output blocks are channelled exclusively through these variates. Indeed, it has been recognized that the canonical variates typically reveal some kind of real physical structure underlying the observations, and they have been used for "exploratory data analysis" already in the 1960's. The underlying *real structure* will be concentrated on more in the next chapter.

Note that, because of the non-symmetricity of the eigenproblem matrices, the bases are now generally *not orthogonal!* This is one concrete difference between CCA and PCA/PLS. It can be claimed that whereas PCA and PLS are mathematically better conditioned, CCA is often physically better motivated — the underlying real structures seldom represent orthogonality.

Despite the very similar starting points, PLS and CCA bases are truly very different. For example, if Y is substituted with X in the formulas, it turns out that PLS equals PCA (because the eigenvectors of X^X are the same as those of $\left(X^T X\right)^2$, and the eigenvalues become squared), whereas CCA cannot at all distinguish between directions in the data space — check this by substituting Y with X in (6.23).

6.3.3 Regression based on PLS and CCA

In Fig. 6.1, it was explained that regression is a three-step procedure with two latent bases. However, it needs to be noted that this cumulating complexity is only illusion, presented in this form only to reach conceptual comprehensibility. In practice, it is only the first mapping from X to Z^1 where the data compression takes place, the step between Z^1 to Z^2 introducing no additional information loss — thus, the same functionality as in the "stepwise" procedure is reached if one maps the data directly from Z^1 to Y, discarding the level Z^2. With PLS, the structure of the regression model reduces into the same expression as with PCR (see page 80):

$$F_{\text{PLS}} = \theta_{\text{PLS}} \left(\theta_{\text{PLS}}^T X^T X \theta_{\text{PLS}}\right)^{-1} \theta_{\text{PLS}}^T X^T Y. \tag{6.29}$$

With CCR, however, the basis vectors are not orthogonal but the projected data score vectors are — see (6.28). That is why, there is again reduction to the algorithm presented on page 80, but the result looks very different[7]:

$$F_{\text{CCR}} = \theta_{\text{CCA}} \theta_{\text{CCA}}^T X^T Y. \tag{6.30}$$

6.3.4 Further ideas*

There are various benefits when all methods are presented in the same eigenproblem-oriented framework — one of the advantages being that one can fluently

[7]Note the similarity between these regression formulas and the expressions (4.19) and (5.36): It is always the correlation between X and Y, or $X^T Y$, being the basis for the mapping between input and output; how this basic structure is modified by the additional matrix multiplier is only dependent of the method

combine different approaches. For example, it turns out that if one defines

$$R_{\text{CR2}} = \frac{1}{k^\beta} \cdot \left(X^T X\right)^{2\alpha-1} \left(X^T Y \left(Y^T Y\right)^{2\alpha-1} Y^T X\right)^{1-\alpha}, \tag{6.31}$$

the methods from CCR to PLS and PCR are found for $\alpha = 0$, $\alpha = \frac{1}{2}$, and $\alpha = 1$, respectively![8] Parameter β can be selected as $\beta = 2\alpha - 1 + (1 - \alpha)(2\alpha - 1) = -2\alpha^2 + 5\alpha - 2$. MLR could also easily be included somewhere along the continuum when using another choice of expressions for the exponents There is one drawback, though — only for the distinct values $\alpha = \frac{1}{2}$ and $\alpha = 1$ the eigenvectors are orthogonal, as compared with the standard continuum regression.

Study yet another idea: Observe the combination of matrices in the CCA solution

$$(X^T X)^{-1} X^T Y (Y^T Y)^{-1} Y^T X. \tag{6.32}$$

Note that this can be divided in two parts: The first part can be interpreted as a mapping X from input to $\hat{Y}$, and the second part maps $\hat{Y}$ to $\hat{X}$:

$$\hat{X} = X \cdot F^1 F^2, \tag{6.33}$$

where

$$\begin{array}{rcl} F^1 & = & (X^T X)^{-1} X^T Y, \quad \text{and} \\ F^2 & = & (Y^T Y)^{-1} Y^T X. \end{array} \tag{6.34}$$

That is, CCA can be interpreted as modeling the behaviors of the mappings when data X is first projected onto output Y and from there back to input. This introduces yet another (CCA oriented) way of constructing the latent basis: One can study what are the statistical properties of this "twice projected" data in the PCA way, that is, the *orthogonal* basis vectors can be defined trough the eigenproblem

$$\hat{X}^T \hat{X} \cdot \theta_i = \lambda_i \cdot \theta_i, \tag{6.35}$$

or

$$X^T Y (Y^T Y)^{-1} Y^T X (X^T X)^{-1} X^T Y (Y^T Y)^{-1} Y^T X \cdot \theta_i = \lambda_i \cdot \theta_i. \tag{6.36}$$

[8] In this case, all the matrices that are involved are low-dimensional and the powers are easily calculated; also note that in the PLS case the square root of the nominal formulation is used for notational simplicity — the eigenvectors, however, remain invariant in both cases

Computer exercises

1. Study the robustness of the different regression methods trying different values for parameter k (number of samples):

```
k = 20;
[X,Y] = dataXY(k,5,4,3,2,0.001,1.0);

E = regrCrossVal(X,Y,'mlr(X,Y)');
errorMLR = sum(sum(E.*E))/(k*4)
E = regrCrossVal(X,Y,'mlr(X,Y,pca(X,3))');     % Try different
errorPCR = sum(sum(E.*E))/(k*4)
E = regrCrossVal(X,Y,'mlr(X,Y,pls(X,Y,2))');   % Try different
errorPLS = sum(sum(E.*E))/(k*4)
E = regrCrossVal(X,Y,'mlr(X,Y,cca(X,Y,2))');   % Try different
errorCCR = sum(sum(E.*E))/(k*4)
```

2. If installed on your computer, get acquainted with the `Chemometrics Toolbox` for `Matlab`, and `PLS Toolbox`. Try the following demos:

```
plsdemo;
crdemo;
```