3.17.2 Nonlinear modeling in functional neuroimaging

There are a number of ways in which nonlinear models can be applied in functional neuroimaging:

- Nonlinear regression for parametric activation designs. Consider a saccadic eye movement study, where subjects are moving their eyes toward an external visual stimuli at varying frequency. There might exist a nonlinear relationship between the frequency of the eye movement and the measured brain signal, e.g., the rCBF. The task of the nonlinear regression is to model this relationship.

  This kind of modeling can be made within the general linear model be augmenting the design matrix with nonlinear terms, e.g., in connection with a saccadic eye movement the frequency would be in one column of the design matrix and in another column the squared frequency which should model the nonlinear effect. Such an approach is taken in Büchel et al. (1996) and nonlinear effects are demonstrated on PET data sets with second order polynomial expansion and cosine basis functions. Büchel et al. (1998) describes the similar method on an fMRI study with a block design (34 seconds activations — 34 seconds baseline): The design matrix has three columns with the zeroth, first and second order polynomial expansion, and the values in the design matrix corresponding to the baseline are zero, while the values corresponding to activation scans are non-zero.

  This kind of nonlinear modeling can also be perform with linear models if "agnostic labeling" is used for each different parameter value. Then there has to be one column in the design matrix for each parameter value, which gives a large DOF with less power in a statistical test, but is more flexible.

- Nonlinear time effects for efMRI. In event-related fMRI (efMRI) it has been shown that rapidly presented events produce a nonlinear BOLD response, see section 2.3.4.

  Friston, Josephs, Rees, and Turner (1998b) and Friston, Mechelli, Turner, and Price (2000b) applied Volterra-series expansion based on the Gamma mixture density hemodynamic response function. ANN model as represented in equation 3.102 was described shortly by (Nielsen, Goutte, and Hansen 1999). ANN models were also used by (Pedersen 1997; Lange, Hansen, Pedersen, Savoy, and Strother 1996; Knowles, Manton, Issa, and Turnbull 1998a; Knowles, Manton, and Turnbull 1998b). The *Balloon model* (Buxton, Wong, and Frank 1998) represents a plausible physiological model and the parameters in this model are non-linear.

- Design variables can couple nonlinearly to individual voxel activations. This is interesting when the cognitive components are not additive and can be modeled within the GLM by including an interaction term (Friston, Holmes, Worsley, Poline, Frith, and Frackowiak 1995d), e.g., by multiplying two design variables to form a third variable.

- Voxels activations couple nonlinearly to a design variable. This cannot be handle with the ordinary GLM. The unsupervised nonlinear PCA method (Friston, Philips, Chawla, and Büchel 1999b; Friston, Philips, Chawla, and Büchel 2000c) is able to model nonlinearities with second order interactions. An advantage of this model is that linear and nonlinear effects is modeled in different parameters, enabling an identification of interaction effects. Another method of dealing with this kind of nonlinearity is by using nonlinear regression from the brain scans data to the design variables. Using neural network this has been developed by the group of Lars Kai Hansen, see, e.g., (Mørch et al. 1997; Lundsager and Kristensen 1996; Mørch et al. 1996b). Contrary to the nonlinear PCA method it does not separate linear and interaction effects.

3.18 Canonical analyses

"Canonical analyses" (as it will be termed here) concern the description of the relationship between (a set of) two multivariate variables $\mathbf{X}$ and $\mathbf{Y}$, e.g., to *summarize their relationship in a subspace* or to *predict one from the other*. The *linear analyses* of this type can be said to include *canonical correlation analysis*, *partial least squares*, *canonical ridge analysis* and *principal component regression* among others. Some of these can be regarded as "bottleneck models" where the relation between the two sets of variables is through a low dimensional subspace, hence the name "low-dimensional linear models" (Borga 1998). In other case the $\mathbf{X}$ and $\mathbf{Y}$ are regarded as being related to (generated from) a hidden subspace or *latent variables*, hence "latent variable models", see, e.g., (Burnham, MacGregor, and Viveros 1999). The type of models can also be considered as "heteroassociative models" (Diamantaras and Kung 1996, section 6). Other names are bilinear models (BLM, Martens and Næs 1989) and reduced rank models.

Usually the variables are considered Gaussian distributed, but models for multinomial data via a softmax layer have also been described (Yee and Hastie 2000).

3.18.1 Canonical correlation analysis

Canonical correlation analysis (CCA) was developed in (Hotelling 1935; Hotelling 1936). Introductions to CCA are available in Mardia et al. (1979, chapter 10) and (Anderson 1984, chapter 12). The estimation problem in CCA is to find two matrices $\mathbf{A}(P \times K)$ and $\mathbf{B}(Q \times K)$ so that the correlations between to linear projects $\mathbf{\Phi} = \mathbf{XA}$ and $\mathbf{\Psi} = \mathbf{YB}$ is the largest. To identify this uniquely the solution should be constrained, and the typical constraints are that the projections $\mathbf{\Phi}(N \times K)$ and $\mathbf{\Psi}(N \times K)$ should be orthogonal $\mathbf{\Phi}^\mathsf{T}\mathbf{\Phi} = \mathbf{\Psi}^\mathsf{T}\mathbf{\Psi} = \mathbf{I}_K$ and the elements in $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are to be arranged so that the first column pair $\boldsymbol{\eta}_1$ and $\boldsymbol{\psi}_1$ has the highest correlation, the second column pair the second highest correlation, and so forth.

There are different ways of estimating the parameters $\mathbf{A}$ and $\mathbf{B}$ with Mardia et al. (1979) relying on SVD:

$$\mathbf{K} = \mathbf{S}_{\mathbf{xx}}^{-1/2}\mathbf{S}_{\mathbf{xy}}\mathbf{S}_{\mathbf{yy}}^{-1/2}, \tag{3.106}$$

where covariances are set as $\mathbf{S}_{\mathbf{xx}} = N^{-1}\mathbf{X}^\mathsf{T}\mathbf{H}\mathbf{X}$, $\mathbf{S}_{\mathbf{yy}} = N^{-1}\mathbf{Y}^\mathsf{T}\mathbf{H}\mathbf{Y}$ and $\mathbf{S}_{\mathbf{xy}} = N^{-1}\mathbf{X}^\mathsf{T}\mathbf{H}\mathbf{Y}$ with $\mathbf{H}$ being the centering matrix.

$$\mathbf{U}\mathbf{L}\mathbf{V}^\mathsf{T} = \mathsf{svd}\left[\mathbf{K}\right] \tag{3.107}$$

$$\mathbf{A} = \mathbf{S}_{\mathbf{xx}}^{-1/2}\mathbf{U} \tag{3.108}$$

$$\mathbf{B} = \mathbf{S}_{\mathbf{yy}}^{-1/2}\mathbf{V}, \tag{3.109}$$

and with a perhaps abusive notation:

$$\mathbf{A}\mathbf{L}\mathbf{B}^\mathsf{T} = \mathbf{S}_{\mathbf{xx}}^{-1/2}\mathsf{svd}\left[\mathbf{S}_{\mathbf{xx}}^{-1/2}\mathbf{S}_{\mathbf{xy}}\mathbf{S}_{\mathbf{yy}}^{-1/2}\right]\mathbf{S}_{\mathbf{yy}}^{-1/2} \tag{3.110}$$

The diagonal elements of $\mathbf{L}$ are called the *canonical correlation coefficients*, and Mardia et al. (1979) call the columns of $\mathbf{A}$ and $\mathbf{B}$ the *canonical correlation vectors*. Furthermore, $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are called the *canonical correlation variables*. If one of the matrices, e.g., $\mathbf{Y}(N \times Q) = \mathbf{Y}\,(N \times (K-1))$ is an indicator matrix indicating group membership in K groups for the N objects in $\mathbf{X}$

$$y_{nk} = \begin{cases} 1 & \text{if } \mathbf{x}_n \text{ in the } k\text{'th group,} \\ 0 & \text{otherwise,} \end{cases} \tag{3.111}$$

then (the first column of) $\mathbf{H}$ are also called *Fisher's linear discriminant function* (Fisher 1936) or the *first canonical variate* (Mardia et al. 1979, exercise 11.5.4), — with the terminology from discriminant analysis. The first column of $\mathbf{A}$ ($\mathbf{a}_1$) is often derived as the maximum eigenvector from a (generalized) eigenvalue decomposition:

$$\mathbf{a}_1 = \mathsf{max\ eig}\left[\mathbf{W}^{-1}\tilde{\mathbf{B}}\right], \tag{3.112}$$

where $\mathbf{W}$ and $\tilde{\mathbf{B}}$ are sum of squares and product (SSP) matrices:

$$\mathbf{T} = \mathbf{W} + \tilde{\mathbf{B}} \qquad \text{Total SSP matrix,} \tag{3.113}$$

$$\mathbf{W} = N_{k=1}\mathbf{S}_{\mathbf{xx},k=1} + \ldots + N_{k=K}\mathbf{S}_{\mathbf{xx},k=K} \qquad \text{Within group SSP matrix,} \tag{3.114}$$

$$\tilde{\mathbf{B}} = \mathbf{T} - \mathbf{W} \qquad \text{Between group SSP matrix.} \tag{3.115}$$

Such an analysis are often referred to as Fisher's linear discriminant analysis (FLD/LDA), canonical variate analysis (CVA) or canonical discriminant analysis (CDA).

This ordinary canonical correlation analysis will not work if any of the two matrices $\mathbf{X}$ and $\mathbf{Y}$ are singular, since the covariance matrices $\mathbf{S}_{\mathbf{xx}}$ and $\mathbf{S}_{\mathbf{yy}}$ will not be invertable. The canonical correlation coefficients are not affected by non-singular affine transformations and the canonical correlation vectors are related through a linear transformation (Mardia et al. 1979, theorem 10.2.4). Furthermore the canonical correlation vectors are not (necessarily) orthogonal.

3.18.2 Partial least squares

"Partial least squares" (PLS) constitutes a group of algorithms and model, see table 3.17 for some of them. Canonical correlation analysis can be regarded as a *mode B PLS* algorithm (Martens and Næs 1989, page 165). The one considered here is where the parameters of the model are determined through singular value decomposition of the covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$ (McIntosh et al. 1996; Worsley et al. 1997) — a decomposition associated with the SIMPLS algorithm (de Jong 1993)

$$\mathbf{U}\mathbf{L}\mathbf{V}^\mathsf{T} = \mathsf{svd}\left(\mathbf{X}^\mathsf{T}\mathbf{Y}\right) \tag{3.116}$$

The "canonical vectors" $\mathbf{U}$ and $\mathbf{V}$ can of course also be found through eigenvalue decomposition of the inner and outer product matrix (equations 3.22 and 3.23). It should be immediately apparent that if one of the matrices is the identity matrix then PLS will correspond to ordinary SVD on the other matrix. Unlike canonical correlation analysis the PLS estimate depends on the centering and scaling of the data in $\mathbf{X}$ and $\mathbf{Y}$ matrices and if no prior knowledge is available about the data it is recommend to scale and center the variables before the singular value decomposition (Wold, Ruhe,

Abbr.	Description	Reference
NIPALS	"Nonlinear iterative PLS"	Wold (1966)
N-PLS	"N-Way PLS", $\mathbf{X}$ is of size $N \times P \times Q \times \ldots$	
PLS1	One-dimensional "response" matrix: $\mathbf{Y}(N \times 1) = \mathbf{y}$	
PLS2	Multidimensional "response" matrix: $\mathbf{Y}(N \times Q),\ Q > 1$	e.g., Martens and Næs (1989, page 146+)
PLS A mode	Regression from latent variables to measured	e.g., Martens and Næs (1989, page 188)
PLS B mode	Regression from measured variables to latent	e.g., Martens and Næs (1989, page 188)
PLS C mode	Combination of A and B mode	e.g., Martens and Næs (1989, page 188)
SIMPLS	Eigendecomposition of covariance matrix	de Jong (1993)

Table 3.17: Some PLS algorithms and models. See also (de Jong and Phatak 1997) for a description of some of the PLS types.

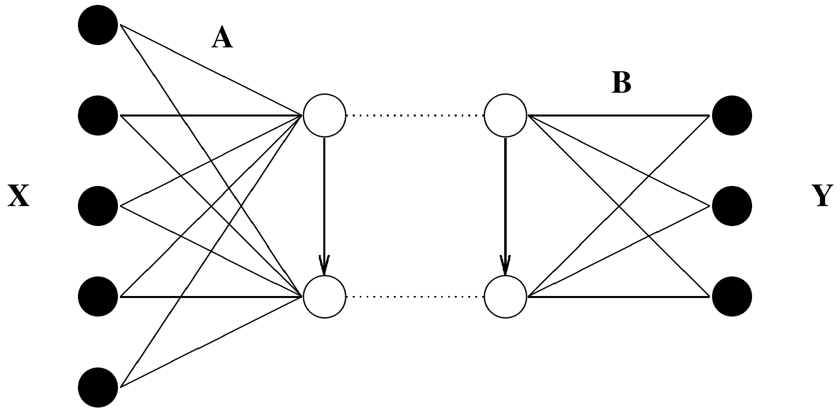

Figure 3.9: Cross-correlation APCA network between $\mathbf{X}(N \times P)$ and $\mathbf{Y}(N \times Q)$ with $P = 5$ and $Q = 3$ with lateral orthogonalization between the $K = 2$ hidden units. Black circles denote directly observed data. Comparable to (Diamantaras and Kung 1996, figure 6.5(a))

Wold, and Dunn III 1984, section 5), e.g., here with a $N \times N$ centering matrix $\mathbf{H}$ and diagonal matrices $\mathbf{D}_x$ and $\mathbf{D}_y$ containing the square root of the diagonal elements in the covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$

$$\mathbf{ULV}^{\mathsf{T}} = \mathsf{svd}\left(\mathbf{D}_x \mathbf{X}^{\mathsf{T}} \mathbf{H}_N \mathbf{Y} \mathbf{D}_y\right). \tag{3.117}$$

The PLS method is somewhat similar to the *cross-correlation asymmetric principal component analysis* (cross-correlation APCA) (Diamantaras and Kung 1994a) and (Diamantaras and Kung 1996, section 6.5)[5]. Diamantaras and Kung (1996, section 6.5) developed the estimation in the model shown by the connectionistic network in figure 3.9 and the maximization of the following cost function

$$E_{\text{APCA}} = \mathsf{trace}\left(\mathbf{A}\mathbf{S}_{\mathbf{xy}}\mathbf{B}^{\mathsf{T}}\right), \tag{3.118}$$

with the orthogonality constraint on the parameter matrices $\mathbf{A}(K \times P)$ and $\mathbf{B}(K \times Q)$

$$\mathbf{A}^{\mathsf{T}}\mathbf{A} = \mathbf{B}^{\mathsf{T}}\mathbf{B} = \mathbf{I}. \tag{3.119}$$

As cross-correlation APCA does not put constraints on the "ordering" of the vectors in $\mathbf{A}$ and $\mathbf{B}$ any solution can be rotated with an orthogonal matrix.

The parameters in the cross-correlation APCA model can be found through an iterative scheme called "cross-coupled Hebbian rule".

[5] Diamantaras and Kung (1996, equation 6.27) approximately defines "Cross-correlation" as the inner product of the two matrices: $N^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{Y}$ (not considering centering of the matrices). "Correlation" is often used differently in statistics, digital signal processing and the neural network literature depending on whether it is normalized with the standard deviation of the individual variables or not and whether it is centered. $N^{-1}\mathbf{D}^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{HXD}^{-1}$ is denoted the (sample) "correlation matrix" by Mardia, Kent, and Bibby (1979, equation 1.4.13) and Conradsen (1984b, page 2.6) while $N^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{X}$ is denoted the "correlation matrix" by (Hertz, Krogh, and Palmer 1991, equation 8.4) and Diamantaras and Kung (1996, equation 6.27). In digital signal processing cross-correlation is often regarded as between two signal in time where the cross-correlation in lag zero would be defined as $\mathbf{x}^{\mathsf{T}}\mathbf{y}$ or $N^{-1}\mathbf{x}^{\mathsf{T}}\mathbf{y}$ depending on the signal, see, e.g., (Mitra 2001, equations 2.103 and 2.144). See also note by Borga (1998, page 25).

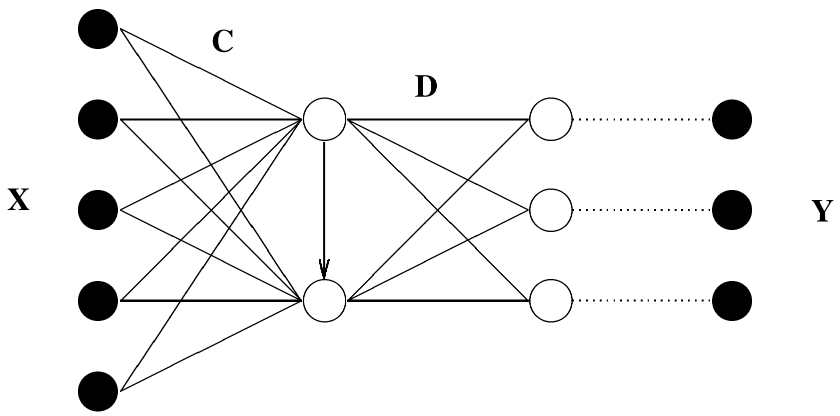

Figure 3.10: Linear approximation APCA network between input $\mathbf{X}(N \times P)$ and target $\mathbf{Y}(N \times Q)$ with $P = 5$ and $Q = 3$ lateral orthogonalization between the $K = 2$ hidden units. The black circles are variables directly observed while the white circles are hidden variables. $\mathbf{B}$ from equation 3.121 is $\mathbf{B} = \mathbf{CD}$. The network structure is different from (Diamantaras and Kung 1996, figure 6.1): An extra layer has been added to incorporate the "noise" between $\mathbf{XB}$ and $\mathbf{Y}$.

3.18.3 Orthonormalized PLS, redundancy analysis, etc.

"Orthonormalized partial least squares" (OPLS) is proposed by Worsley et al. (1997) where one of the matrices is normalized by the square root of the full covariance matrix.

$$\tilde{\mathbf{U}}\mathbf{L}\mathbf{V}^{\mathsf{T}} = \mathsf{svd}\left(\mathbf{S}_{\mathbf{xx}}^{-1/2}\mathbf{X}^{\mathsf{T}}\mathbf{Y}\right). \tag{3.120}$$

In (Worsley 1997; Worsley et al. 1997) CCA, PLS and OPLS are called *canonical variables analyses*. The *linear heteroencoder with bottleneck* (LHB) as described by Roweis and Brody (1999, section 3) uses the same subspace as OPLS for its singular value decomposition. The bottleneck is shown to be optimal with the following cost function — the *reconstruction error*

$$E_{\mathrm{LHB}} = \|\mathbf{XB} - \mathbf{Y}\|_F^2 = \mathsf{trace}\left[(\mathbf{XB} - \mathbf{Y})^{\mathsf{T}}(\mathbf{XB} - \mathbf{Y})\right], \tag{3.121}$$

and restricting the rank of the parameters $K = \mathsf{rank}(\mathbf{B})$ to a number less than $\mathsf{min}\left[\mathsf{rank}(\mathbf{X}), \mathsf{rank}(\mathbf{Y})\right]$

$$\hat{\mathbf{B}}_{\mathrm{LHB},K} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1/2} \mathsf{svd}_K\left[\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1/2}\mathbf{X}^{\mathsf{T}}\mathbf{Y}\right] \tag{3.122}$$

where svd_K is the truncated singular value decomposition to rank K.

Similar models have been studied by several authors: According to Jackson (1991, section 12.6.4) Rao (1964) proposed it as *principal components of instrumental variables* and Fortier (1966) introduced the same model. Van den Wollenberg (1977) called it redundancy analysis (RDA). Diamantaras and Kung (1996, chapter 6) account that Scharf (1991) introduced it as *reduced-rank Wiener filtering* and that Baldi and Hornik (1989) also studied it. Diamantaras and Kung (1996) and Diamantaras and Kung (1994b) themselves called it *linear approximation asymmetric principal component analysis*, the network structure is depicted in figure 3.10. Burnham, MacGregor, and Viveros (1999) referred to it as *reduced-rank regression* (RRR). Furthermore, Borga (1998) has denoted it *multivariate linear regression* (MLR).

Diamantaras and Kung (1996) show that parameter estimation can also be obtained by generalized singular value decomposition (GSVD) of a matrix pair $\mathbf{X}^{\mathsf{T}}\mathbf{Y}$ and $\mathbf{X}$, and (Borga 1998, section 4.5) describes a solution based on generalized eigenvalue decomposition.

RRR is one of the end points — with PCR on the other end — in the so-called *joint continuum regression* that is able to interpolate between the two methods on a continuous scale by adjusting a single parameter (Brooks and Stone 1994).

As a regression model RRR is optimal and can perform better than PCR, PLS and CCA (Jackson 1991, table 12.8, Linnerud data) if the performance is measured as the trace of the covariance to the residuals. If the performance is measured by the determinant of the residual covariance the CCA is the best.

3.18.4 Canonical ridge analysis

Canonical ridge analysis (CRA) (Vinod 1976) can be viewed as canonical correlation analysis with regularization, and as with ordinary ridge regression small non-zero regularization parameters helps to stabilize the solution (Mardia et al. 1979, exercise 10.2.14). Canonical ridge analysis adds a diagonal term to the $\mathbf{S}_{\mathbf{xx}}$ and $\mathbf{S}_{\mathbf{yy}}$ covariance matrices

$$\tilde{\mathbf{S}}_{\mathbf{xx}} = \mathbf{S}_{\mathbf{xx}} + k_X\mathbf{I} \tag{3.123}$$

$$\tilde{\mathbf{S}}_{\mathbf{yy}} = \mathbf{S}_{\mathbf{yy}} + k_Y\mathbf{I}, \tag{3.124}$$

and then CRA proceeds as ordinary CCA with these regularized covariance matrices. The regularization parameters k_X and k_Y should be non-negative, and if they are positive $k_X, k_Y > 0$ then the regularized covariance $\tilde{\mathbf{S}}$ will be invertable even if $\mathbf{S}$ are singular.

With a slight different application of the regularization parameters the regularized covariance can be written as

$$\tilde{\mathbf{S}}_{\mathbf{xx}} = (1 - k_X)\mathbf{S}_{\mathbf{xx}} + k_X \mathbf{I} \tag{3.125}$$

$$\tilde{\mathbf{S}}_{\mathbf{yy}} = (1 - k_Y)\mathbf{S}_{\mathbf{yy}} + k_Y \mathbf{I}, \tag{3.126}$$

with $0 \leq k_X, k_Y \leq 1$. Ordinary CCA are obtained with $k_X = k_Y = 0$ and the PLS of the previous section can be obtained with $k_X = k_Y = 1$. Thus CRA forms a continuum between PLS and CCA as extremes (Nielsen, Hansen, and Strother 1998). It can be noted that if a matrix is orthonormal implying $\mathbf{S} = \mathbf{I}$ then the regularization for that matrix has no effect. This could be the case in connection with a design matrix.

This type of regularization has also been used in discriminant analysis and termed penalized discriminant analysis (PDA), where it is usually formulated as a regularization on the within-group SSP matrix (Ripley 1996, equation 3.8)

$$\tilde{\mathbf{W}} = \mathbf{W} + k_W \mathbf{I} \tag{3.127}$$

or the total SSP matrix (Kustra and Strother 2000). Krzanowski, Jonathan, McCarthy, and Thomas (1995, section 3.1) referred to this as *generalized ridge discrimination* (GRD).

3.18.5 Canonical correlation analysis with singular matrices

Ordinary canonical correlation analysis is not possible if any of the two matrices $\mathbf{X}$ and $\mathbf{Y}$ are singular because it will not be possible to invert the covariances $\mathbf{S}_{\mathbf{xx}}$ and $\mathbf{S}_{\mathbf{yy}}$. It has been suggested to use generalized inverses (Mardia, Kent, and Bibby 1979, page 287), e.g., the Moore-Penrose inverse. Krzanowski, Jonathan, McCarthy, and Thomas (1995) term a related method in discriminant analysis based on the generalized inverse for *modified canonical analysis* (MCA). However, this will not identify the canonical vectors uniquely for canonical correlation analysis with $Q > 1$ as all structure is taken out of the matrices so that $\mathbf{K} = \mathbf{S}_{\mathbf{xx}}^{-1/2}\mathbf{S}_{\mathbf{xy}}\mathbf{S}_{\mathbf{yy}}^{-1/2}$ will be spheric (all non-zero singular values are unity), and it will not be possible to identify a "first" canonical component due to the rotational ambiguity. A subspace can be identified however.

The use of SVD or PCA (so-called preliminary principal component analysis) will correspond to using the Moore-Penrose inverse if all components are used, but if only some of them are used it will be possible to identify the singular vectors uniquely. First the $\mathbf{X}$ and $\mathbf{Y}$ matrix are decomposed via SVD

$$\mathbf{U}\mathbf{L}\mathbf{V}^{\mathsf{T}} = \mathsf{svd}\left[\mathbf{X}\right] \tag{3.128}$$

$$\mathbf{C}\mathbf{D}\mathbf{E}^{\mathsf{T}} = \mathsf{svd}\left[\mathbf{Y}\right]. \tag{3.129}$$

Then the principal components $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ are obtained with L and M as the number of components maintained in the analysis

$$\tilde{\mathbf{X}}_L = \mathbf{U}_L \mathbf{L}_L \tag{3.130}$$

$$\tilde{\mathbf{Y}}_M = \mathbf{C}_M \mathbf{D}_M. \tag{3.131}$$

The principal component are used as "input" to CCA and intermediate vectors are obtained

$$\mathbf{P}\mathbf{R}\mathbf{Q}^{\mathsf{T}} = \mathsf{svd}\left[\tilde{\mathbf{K}}\right] = \mathsf{svd}\left[\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1/2}\mathbf{S}_{\tilde{\mathbf{x}}\tilde{\mathbf{y}}}\mathbf{S}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}}^{-1/2}\right] \tag{3.132}$$

$$= \mathsf{svd}\left[\mathbf{U}_L^{\mathsf{T}}\mathbf{C}_M\right]. \tag{3.133}$$

Finally the canonical correlation vectors $\mathbf{A}(P \times K)$ and $\mathbf{B}(Q \times K)$ in the space of $\mathbf{X}$ and $\mathbf{Y}$ is found

$$\mathbf{A} = \mathbf{V}_L \mathbf{L}_L^{-1} \mathbf{P} \tag{3.134}$$

$$\mathbf{B} = \mathbf{E}_M \mathbf{D}_M^{-1} \mathbf{Q} \tag{3.135}$$

Another possibility is to use CRA to make the covariance matrices $\mathbf{S}_{\mathbf{xx}}$ and $\mathbf{S}_{\mathbf{yy}}$ positive definite. The ridge regularization can be applied in the in the full space or in the space spanned by the principal components from a preliminary principal component analysis (PPCA), — the result does not differ, provided that all PPCA components are maintained and that the principal component maintain the variance, i.e., not sphered The canonical correlation vectors can then be found through a low-dimensional eigenvalue decomposition, see theorem B.2.1 on page 136

$$\mathbf{A} = \mathbf{V}\breve{\mathbf{\Lambda}}_x \mathbf{U}^{\mathsf{T}} \breve{\mathbf{P}} \tag{3.136}$$

$$\mathbf{B} = \mathbf{E}\breve{\mathbf{\Lambda}}_y \mathbf{C}^{\mathsf{T}} \breve{\mathbf{Q}} \tag{3.137}$$

$$\mathbf{R} = \mathbf{\Psi}^{1/2} = \mathbf{\Upsilon}^{1/2}, \tag{3.138}$$

where $\breve{\mathbf{P}}(N \times K)$ and $\breve{\mathbf{Q}}(N \times K)$ are scaled right eigenvectors and the diagonal matrices $\mathbf{\Psi}(K \times K)$ and $\mathbf{\Upsilon}(K \times K)$ contain K eigenvalues each from a symmetric $K \times K$ eigenvalue decomposition

$$\mathbf{P\Psi P}^{-} = \mathsf{reig}\left(\mathbf{C}\mathbf{\Lambda}_y^2\mathbf{C}^\mathsf{T}\mathbf{U}\mathbf{\Lambda}_x^2\mathbf{U}^\mathsf{T}\right) \tag{3.139}$$

$$\mathbf{Q\Upsilon Q}^{-} = \mathsf{reig}\left(\mathbf{U}\mathbf{\Lambda}_x^2\mathbf{U}^\mathsf{T}\mathbf{C}\mathbf{\Lambda}_y^2\mathbf{C}^\mathsf{T}\right) \tag{3.140}$$

$$\breve{\mathbf{P}} = \mathbf{P}\ \left\{\mathsf{diag}\left[\mathsf{diag}\left(\mathbf{P}^\mathsf{T}\mathbf{U}\mathbf{\Lambda}_x^2\mathbf{U}^\mathsf{T}\mathbf{P}\right)\right]\right\}^{-1/2} \tag{3.141}$$

$$\breve{\mathbf{Q}} = \mathbf{Q}\ \left\{\mathsf{diag}\left[\mathsf{diag}\left(\mathbf{Q}^\mathsf{T}\mathbf{C}\mathbf{\Lambda}_y^2\mathbf{C}^\mathsf{T}\mathbf{Q}\right)\right]\right\}^{-1/2} \tag{3.142}$$

$$\breve{\mathbf{\Lambda}}_x = \left(\mathbf{L}^2 + k_x\mathbf{I}_K\right)^{-1}\mathbf{L} \tag{3.143}$$

$$\breve{\mathbf{\Lambda}}_y = \left(\mathbf{D}^2 + k_y\mathbf{I}_K\right)^{-1}\mathbf{D} \tag{3.144}$$

$$\mathbf{\Lambda}_x = \left(\mathbf{L}^2 + k_x\mathbf{I}_K\right)^{-1/2}\mathbf{L} \tag{3.145}$$

$$\mathbf{\Lambda}_y = \left(\mathbf{D}^2 + k_y\mathbf{I}_K\right)^{-1/2}\mathbf{D}, \tag{3.146}$$

where $\mathbf{B}$ may also be computed as

$$\mathbf{B} = \mathbf{E}\breve{\mathbf{\Lambda}}_y\mathbf{C}^\mathsf{T}\mathbf{U}\mathbf{\Lambda}_x^2\mathbf{U}^\mathsf{T}\breve{\mathbf{P}} \tag{3.147}$$

3.18.6 Bilinear modeling

Martens and Næs (1989, section 3.3.3) and Burnham, MacGregor, and Viveros (1999) regard the so-called bilinear model (BLM) particularly used in chemometrics

$$\mathbf{X} = \mathbf{TP} + \mathbf{E} \tag{3.148}$$

$$\mathbf{Y} = \mathbf{TQ} + \mathbf{F}, \tag{3.149}$$

where $\mathbf{X}(N \times P)$ and $\mathbf{Y}(N \times Q)$ are measured variables, $\mathbf{T}(N \times K)$ contains latent variables, $\mathbf{P}(K \times P)$ and $\mathbf{Q}(N \times Q)$ are parameter matrices ("regression coefficients") and $\mathbf{E}(N \times P)$ and $\mathbf{F}(N \times Q)$ are residuals (random errors, noise, ...). This model can be regarded as a PLS A mode model (see table 3.17) with the observed variables as a function of the latent variables. It is often presented in a regression framework where the task is to predict $\mathbf{Y}$ from $\mathbf{X}$ where it aims to substitute the multivariate regression model $\mathbf{Y} = \mathbf{XB} + \mathbf{U}$ and its maximum likelihood (ordinary least squares) estimate $\hat{\mathbf{B}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y}$ with an estimate that is better for rank-deficient and ill-posed problems. This is done by making K smaller than the rank of $\mathbf{X}^\mathsf{T}\mathbf{X}$.

One way to estimate the parameters in this model is to estimate the latent variables $\mathbf{T}$ from the observed variables $\mathbf{X}$, (Martens and Næs 1989, equation 3.34), (Burnham, MacGregor, and Viveros 1999, equation 3)

$$\hat{\mathbf{T}} = \mathbf{X\Phi}. \tag{3.150}$$

This model is not uniquely determined, and a common constraint for the latent variables is an orthonormality constraint

$$\hat{\mathbf{T}}^\mathsf{T}\hat{\mathbf{T}} = \mathbf{I}_K. \tag{3.151}$$

If the noise matrices are independent Gaussian distributed $\mathbf{e} \sim \mathcal{N}_P\left(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{ee}}\right)$ and $\mathbf{f} \sim \mathcal{N}_Q\left(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{ff}}\right)$ then the joint log-likelihood can be written as (Burnham, MacGregor, and Viveros 1999, equation 6)

$$\ell = -\frac{1}{2}\mathsf{trace}\left[\left(\mathbf{X} - \mathbf{X\Phi P}\right)\mathbf{\Sigma}_{\mathbf{ee}}^{-1}\left(\mathbf{X} - \mathbf{X\Phi P}\right)^\mathsf{T}\right] - \frac{1}{2}\mathsf{trace}\left[\left(\mathbf{Y} - \mathbf{X\Phi Q}\right)\mathbf{\Sigma}_{\mathbf{ff}}^{-1}\left(\mathbf{Y} - \mathbf{X\Phi Q}\right)^\mathsf{T}\right] \tag{3.152}$$

The profile log-likelihood for $\mathbf{\Phi}$ can be found when maximum likelihood estimates for $\mathbf{P}$ and $\mathbf{Q}$ with fixed $\mathbf{\Phi}$ are plugged into equation 3.152 (Burnham, MacGregor, and Viveros 1999, equation 9)

$$\ell(\mathbf{\Phi}) = \frac{1}{2}\mathsf{trace}\left[\mathbf{\Phi}^\mathsf{T}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{\Sigma}_{\mathbf{ee}}^{-1}\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{X}^\mathsf{T}\mathbf{Y}\mathbf{\Sigma}_{\mathbf{ff}}^{-1}\mathbf{Y}^\mathsf{T}\mathbf{X}\right)\mathbf{\Phi}\right] \tag{3.153}$$

The maximum likelihood estimate of the parameter $\mathbf{\Phi}$ is found as the eigenvectors to a generalized eigenvalue problem

$$\left(\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{\Sigma}_{\mathbf{ee}}^{-1}\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{X}^\mathsf{T}\mathbf{Y}\mathbf{\Sigma}_{\mathbf{ff}}^{-1}\mathbf{Y}^\mathsf{T}\mathbf{X}\right)\mathbf{\Phi} = \mathbf{X}^\mathsf{T}\mathbf{X\Phi D} \tag{3.154}$$

Burnham, MacGregor, and Viveros (1999) show that by varying the covariance of the residuals the estimate of $\mathbf{\Phi}$ would vary between an estimate associated with principal component regression (PCR) and another associated with redundancy analysis (RDA). This is done by assuming uncorrelated noise $\mathbf{\Sigma}_{\mathbf{ee}} = \sigma_{\mathbf{e}}^2\mathbf{I}$ and $\mathbf{\Sigma}_{\mathbf{ff}} = \sigma_{\mathbf{f}}^2\mathbf{I}$

$$\left(\sigma_{\mathbf{e}}^2\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{X} + \sigma_{\mathbf{f}}^2\mathbf{X}^\mathsf{T}\mathbf{Y}\mathbf{Y}^\mathsf{T}\mathbf{X}\right)\mathbf{\Phi} = \mathbf{X}^\mathsf{T}\mathbf{X\Phi D} \tag{3.155}$$

By multiplication with $\sigma_{\mathbf{e}}\sigma_{\mathbf{f}}$, setting $\phi = \sigma_{\mathbf{e}}/\sigma_{\mathbf{f}}$ and $\breve{\mathbf{D}} = \sigma_{\mathbf{e}}\sigma_{\mathbf{f}}\mathbf{D}$

$$\left(\frac{1}{\phi}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{X} + \phi\mathbf{X}^\mathsf{T}\mathbf{Y}\mathbf{Y}^\mathsf{T}\mathbf{X}\right)\mathbf{\Phi} = \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{\Phi}\breve{\mathbf{D}}. \tag{3.156}$$

With $\phi \to \infty$ corresponding to the errors in $\mathbf{X}$ being much larger than those of $\mathbf{Y}$ the estimate associated with redundancy analysis (RDA, section 3.18.3) is obtained

$$\left(\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{Y}\mathbf{Y}^\mathsf{T}\mathbf{X}\right)\mathbf{\Phi} = \mathbf{\Phi}\mathbf{D}. \tag{3.157}$$

$$\mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{S}_{\mathbf{xy}}\mathbf{S}_{\mathbf{yx}}\mathbf{\Phi} = \mathbf{\Phi}\mathbf{D}, \tag{3.158}$$

Here the mean is disregarded in the calculation of the covariances. With $\phi \to 0$ corresponding to errors in $\mathbf{Y}$ being much larger than those of $\mathbf{X}$ the estimate is associated with principal component regression (PCR)

$$\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)\mathbf{\Phi} = \mathbf{\Phi}\mathbf{D}. \tag{3.159}$$

$$\mathbf{S}_{\mathbf{xx}}\mathbf{\Phi} = \mathbf{\Phi}\mathbf{D}. \tag{3.160}$$

Similarly, the left canonical correlation vector can be obtained by setting the covariances to $\mathbf{\Sigma}_{\mathbf{ee}} = \mathbf{X}^\mathsf{T}\mathbf{X}$ and $\mathbf{\Sigma}_{\mathbf{ff}} = \mathbf{Y}^\mathsf{T}\mathbf{Y}$, ignoring the normalization term N^{-1}

$$\left(\mathbf{S}_{\mathbf{xx}}\mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{S}_{\mathbf{xx}} + \mathbf{S}_{\mathbf{xy}}\mathbf{S}_{\mathbf{yy}}^{-1}\mathbf{S}_{\mathbf{yx}}\right)\mathbf{\Phi} = \mathbf{S}_{\mathbf{xx}}\mathbf{\Phi}\mathbf{D} \tag{3.161}$$

With an invertable $\mathbf{S}_{\mathbf{xx}}$ and with $\tilde{\mathbf{D}} = \mathbf{D} - \mathbf{I}$, cf. derivations in theorem B.4.1 on page 139

$$\mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{S}_{\mathbf{xy}}\mathbf{S}_{\mathbf{yy}}^{-1}\mathbf{S}_{\mathbf{yx}}\mathbf{\Phi} = \mathbf{\Phi}\tilde{\mathbf{D}} \tag{3.162}$$

The square root of the eigenvalue $\tilde{\mathbf{D}}$ are the canonical correlation coefficients.

Once the latent variables are determine the $\mathbf{Q}$ matrix can be estimated. This can be done by a least square estimate:

$$\hat{\mathbf{Q}}_{\text{OLS}} = \left(\mathbf{T}^\mathsf{T}\mathbf{T}\right)^{-1}\mathbf{T}^\mathsf{T}\mathbf{Y}. \tag{3.163}$$

If the latent variables are orthonormal the expression can be simplified

$$\hat{\mathbf{Q}}_{\text{OLS}} = \mathbf{T}^\mathsf{T}\mathbf{Y}. \tag{3.164}$$

The estimate is complicated by that there still might be noise in the latent variables, — especially for large K. This has led Høy, Westad, and Martens (2001) to introduce (generalized) ridge regression in the estimation of $\mathbf{Q}$

$$\hat{\mathbf{Q}}_{\text{GRR}} = \left(\mathbf{T}^\mathsf{T}\mathbf{T} + \mathbf{\Lambda}\right)^{-1}\mathbf{T}^\mathsf{T}\mathbf{Y}, \tag{3.165}$$

where $\mathbf{\Lambda}$ is a diagonal matrix with ridge regularization hyperparameters.

Bilinear model as factor analysis

The bilinear model can be expressed as a factor analysis model (section 3.9.2), if the variances of the noise represented in the matrices $\mathbf{E}$ and $\mathbf{F}$ are uncorrelated. First is a matrix $\mathbf{Z}$ constructed from $\mathbf{X}$ with $\mathbf{Y}$ appended

$$\mathbf{Z} = (\mathbf{X}, \mathbf{Y}) \tag{3.166}$$

Disregarding any mean for clarity, and applying equations 3.148 and 3.149, under the assumption of orthonormal latent variables $N^{-1}\mathbf{T}^\mathsf{T}\mathbf{T} = \mathbf{I}$ and that there is no correlation between the latent variables and the noise the covariance of $\mathbf{z}$ is

$$\mathsf{V}[\mathbf{z}] = N^{-1}\mathbf{Z}^T\mathbf{Z} = N^{-1}\begin{bmatrix} \mathbf{P}^\mathsf{T}\mathbf{P} + \mathbf{E}^\mathsf{T}\mathbf{E} & \mathbf{P}^\mathsf{T}\mathbf{Q} \\ \mathbf{Q}^\mathsf{T}\mathbf{P} & \mathbf{Q}^\mathsf{T}\mathbf{Q} + \mathbf{F}^\mathsf{T}\mathbf{F} \end{bmatrix} \tag{3.167}$$

Now the matrix of factor loading can be constructed as

$$\mathbf{\Lambda} = (\mathbf{P}, \mathbf{Q}) \tag{3.168}$$

and the covariance of the *specific factors* is

$$\boldsymbol{\Upsilon} = \begin{bmatrix} \boldsymbol{\Upsilon}_{\mathbf{e}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Upsilon}_{\mathbf{f}} \end{bmatrix}, \tag{3.169}$$

where the covariance of the noise are diagonal matrices

$$\mathsf{V}[\mathbf{e}] = \boldsymbol{\Upsilon}_{\mathbf{e}} = \mathsf{diag}(\upsilon_{\mathbf{e},11}, \ldots, \upsilon_{\mathbf{e},PP}) \tag{3.170}$$
$$\mathsf{V}[\mathbf{f}] = \boldsymbol{\Upsilon}_{\mathbf{f}} = \mathsf{diag}(\upsilon_{\mathbf{f},11}, \ldots, \upsilon_{\mathbf{f},QQ}) \tag{3.171}$$
$$\mathsf{C}[\mathbf{e}, \mathbf{f}] = \mathbf{0}. \tag{3.172}$$

Thus the condition on the covariance of $\mathbf{z}$ is, cf. (Mardia et al. 1979, equation 9.2.5)

$$\boldsymbol{\Sigma}_{\mathbf{zz}} = \boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{\Lambda} + \boldsymbol{\Upsilon}. \tag{3.173}$$

3.18.7 Optimization of the ridge parameter and subspace dimension

A test for the subspace dimension in ordinary canonical correlation analysis can be developed (Mardia et al. 1979, section 10.2.3). For the test of no subspace $\boldsymbol{\Sigma}_{\mathbf{xy}} = \mathbf{0}$ the likelihood ratio statistics has a Wilk's lambda (Andersen's U) distribution when $\mathbf{x}$ and $\mathbf{y}$ is Gaussian distributed

$$\Lambda(Q, N-1-P, P) \sim \left|\mathbf{I} - \mathbf{S}_{\mathbf{yy}}^{-1}\mathbf{S}_{\mathbf{yx}}\mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{S}_{\mathbf{xy}}\right| = \prod_{k=1}^{K}(1 - r_k^2), \tag{3.174}$$

where r_k^2 is the square canonical correlation coefficients. Wilk's lambda is related to the mutual information between $\mathbf{x}$ and $\mathbf{y}$, see derivation on page 138:

$$I(\mathbf{x}; \mathbf{y}) = -\frac{1}{2}\ln\Lambda. \tag{3.175}$$

As Wilk's lambda is not tabulated this is often replaced by χ^2 $(N \to \infty)$-asymptotic approximation.

The asymptotic approximation for the test of only S of the canonical correlations coefficients are non-zero is (Mardia, Kent, and Bibby 1979, equation 10.2.14)

$$\chi^2_{PQ} \sim -\left(N - \frac{1}{2}(P+Q+3)\right)\ln\left[\prod_{k=S+1}^{K}(1 - r_k^2)\right]. \tag{3.176}$$

This approximation might not be precise when N is close to P or Q which is sometimes the case in functional neuroimaging. Other tests are proposed by (Marriott 1952) and (Gunderson and Muirhead 1997).

These tests do not directly work in conjunction with ridge regularization and singular data. Loh (1995) develops a formula for penalized discriminant analysis for the determination of what corresponds to our k_X ridge parameter. In PLS a common procedure is to estimate the dimension of the subspace with cross-validation (Wold 1978).

Bullmore et al. (1996b) and Mencl et al. (2000) used permutation test in connection with PLS to determine the significant components (subspace dimensions): The columns of the $\mathbf{Y}$ matrix containing design variables are randomly shuffled and corresponding to a null-hypothesis and the PLS analysis is repeated, e.g., 1000 times.

Frutiger et al. (2000) used the resampling technique of (Strother et al. 1997a; Strother et al. 1997b), where the data is split and two sets of canonical correlation vectors are computed. The correlation between corresponding canonical correlations vectors are reported as a reproducibility index.

3.18.8 Generalization of canonical analysis

There exist three types of extensions to canonical analysis:

- Multiway analyses on matrices with more that two dimensions, e.g, $\mathbf{X}(N \times P \times Q \times \ldots)$. Development has occurred both for unsupervised algorithms and for canonical analysis with two data sets (Bro 1998). This is of interest in connection with functional neuroimaging since data set are often from balanced experiments, e.g., $\mathbf{X}$(subjects $\times$ sessions $\times$ runs $\times$ conditions). Unsupervised analysis with a "SVD-$k$modes" has been applied to brain data, see, e.g., (Leibovici 2000).
- Multiset canonical analysis with more that two matrices, e.g., $\mathbf{X}(N \times P)$, $\mathbf{Y}(N \times P)$, $\mathbf{Z}(N \times Q)$. Kettenring (1971) and Sen Gupta (1982) describe multiset canonical correlation analysis.

- Nonlinear canonical analysis or "canonical manifold analysis" where the data is regarded as related through or generated from a low-dimension non-linear sub-manifold. Tibshirani (1992) describes "principal curves" which is one-dimensional sub-manifold estimated from a single set of data. A nonlinear sub-manifold of two sets of data (two matrices) can be identified with two artificial neural networks with the output in a latent space. Auto-associative models appear in (Diamantaras and Kung 1996, section 6.6). Other references are (Becker 1992; Becker and Hinton 1992; Asoh and Takechi 1994; DeMers and Cottrell 1993; Lai and Fyfe 1999).

3.18.9 Canonical analyses in functional neuroimaging

Friston, Frith, Frackowiak, and Turner (1995b) used canonical correlation analysis with $\mathbf{X}(N \times P)$ containing $N = 120$ scans represented as $P = 35$ eigensequences (from a preliminary principal component analysis of a preprocessed data matrix, P being determined as the number of eigenvalues above the mean of the non-zero eigenvalues) and $\mathbf{Y}(N \times Q)$ being a design matrix containing four sinusoid basis function (as a hemodynamic response function model) for each of three conditions: $Q = 4 \times 3 = 12$ design variables. As there are more scans than principal components and design variables the covariance matrices can be inverted. A type of χ^2-test with correction for correlation among the scans was used to determine a significant subspace.

In the case that the task is repeated *agnostic labeling* can be used which was done by Strother, Lange, Savoy, Anderson, Sidtis, Hansen, Bandettini, O'Craven, Rezza, Rosen, and Rottenberg (1996a): The fMRI data set consisted of 8 runs each with 72 scans (24 baseline + 24 activation + 24 baseline scans). The agnostic labeling groups the first scan of the first run in the same group as first scans of the 7 other runs. Likewise with the second and following scans, so the design matrix is be $\mathbf{Y}(N \times Q) = \mathbf{Y}\left((8 \times 72) \times (72 - 1)\right)$. Strother et al. (1996a) also depicted the canonical variate trajectory in the first few canonical subspaces termed *state-space*. Friston, Poline, Holmes, Frith, and Frackowiak (1996b) and (Fletcher et al. 1996) also used agnostic labeling on a PET data set for canonical variate analysis with preliminary principal component analysis.

Friman, Cedefamn, Lundberg, Borga, and Knutsson (2001b), Friman, Borga, Lundberg, and Knutsson (2001a), Friman, Lundberg, Cedefamn, Borga, and Knutsson (2001c) used canonical correlation on a *local* area of the volume from fMRI: A 3×3 voxel size intraslice region is taken around each voxel. This 9 dimensional vector is then reduced by spatially symmetric filters to 5 dimensions, thus for the i'th voxel and with 180 scans the data matrix is $\mathbf{X}_i(N \times P) = \mathbf{X}_i(180 \times 5)$. The other matrix is constructed with 6 sinusoid basis function $\mathbf{Y}(N \times Q) = \mathbf{Y}(180 \times 6)$, — this matrix is common for all voxels. As the summary image is reported the canonical correlation coefficient for each voxel and in (Friman et al. 2001a) they suggest using a test to determine a threshold.

Ledberg (1998) related two sets of volume data with a kind of canonical analysis: Splitting the data from a PET study across subjects into two new data matrices these underwent their own individual preliminary principal component analysis maintaining a number of the largest PCs. Once these are found the "closet subspace" are determined by minimizing a distance measure between the subspaces from different groups, — a method based on (Krzanowski 1979).

Bullmore, Rabe-Hesketh, Morris, Williams, Gregory, Gray, and Brammer (1996b) extracted 170 voxels from a mass-univariate analysis of 100 fMRI scans and performed singular value decomposition both directly on the $\mathbf{X}_1(N \times P) = \mathbf{X}_1(100 \times 170)$ matrix as well as on another matrix $\mathbf{X}_2(N \times P) = \mathbf{X}_2(6 \times 170)$ constructed from the regression coefficients associated with 6 sinusoidal design variables from the mass-univariate analysis. A further independent canonical variate analysis was performed on 101 voxels that where grouped according to the anatomical region they occupied, i.e., the canonical variate analysis was performed on the transposed data matrix. These voxels were represented with the 6 regression coefficients: $\mathbf{X}(N \times P) = \mathbf{X}(101 \times 6)$. The other matrix contains group indication for one of the six regions of interest (ROI) $\mathbf{Y}(N \times Q) = \mathbf{Y}(101 \times 5)$.

Frutiger, Strother, Anderson, Sidtis, Arnold, and Rottenberg (2000) used preliminary principal component analysis on 26130 voxel PET data set with 140 scans choosing heuristically the first 20 principal component for a further 10-group CVA corresponding to 10 conditions: $\mathbf{X}(140 \times 20)$ and $\mathbf{Y}(140 \times 9)$. In this study CVA was also used on behavioral data collected during the scanning.

McIntosh, Bookstein, Haxby, and Grady (1996) used partial least squares (PLS) on face processing PET data represented in the $\mathbf{X}$ matrix and with two different design matrices as $\mathbf{Y}$: one with condition, the other with reaction time behavioral score. A permutation test was used to access the significance. Partial least square was also used by (Mencl et al. 2000) on fMRI scans. PLS with PET data was also investigated by McIntosh, Nyberg, Bookstein, and Tulving (1997) in a type of analysis inspired by the "seed voxel" correlation analysis of (Horwitz, Grady, Haxby, Schapiro, Rapoport, Ungerleider, and Mishkin 1992): The design matrix $\mathbf{Y}$ is constructed from 3 representative voxels selected from 3 activated regions involved in episodic memory retrieval. Since the design consists of 4 conditions the design matrix is 12-dimensional (3 voxels $\times$ 4 conditions). The data matrix $\mathbf{X}$ contains the full voxel data.

Other uses of canonical analysis appear in (Busatto, Howard, Ha., Brammer, Wright, Woodruff, Simmons, Williams, David, and Bullmore 1997) (CVA on fMRI), (McCrory and Ford 1991) (CCA on SPECT), (Lobaugh, West, and McIntosh 2001) (PLS on ERP).

Friston et al. (1996b) called the K canonical correlation vectors $\mathbf{A}(P \times K)$ *canonical images* as they where associated with P voxels. In (Frutiger et al. 2000) they are called *canonical eigenimages*. Similarly, if the K canonical correlation

N	P	Q	Type	Reference
120 scans	35 PCs	3 conditions and 6 sinusoid	SVD/CCA, fMRI	Friston et al. (1995b)
60 scans	14 PCs	12 conditions	SVD/CVA, PET	Friston et al. (1996b)
Scans	Voxels	Conditions	PLS, PET	McIntosh et al. (1996)
Scans	Voxels	Reaction time	PLS, PET	McIntosh et al. (1996)
Scans	$\approx$ 60000 voxels	4 conditions $\times$ 3 voxels	PLS, PET	McIntosh et al. (1997)
101 voxels	6 regression coefficients	5 ROI indicators	CCA, fMRI	Bullmore et al. (1996b)
140 scans	20 PCs	10 conditions	SVD/CVA	Frutiger et al. (2000)
180 scans	5 filter outputs	6 Sinusoid	CCA, fMRI	Friman et al. (2001b)

Table 3.18: Canonical analyses in functional neuroimaging. N, P and Q refer to the dimensions of the matrices $\mathbf{X}(N \times P)$ and $\mathbf{Y}(N \times Q)$. “PCs” is principal components. “SVD” denotes preliminary principal component analysis.

vectors $\mathbf{B}(Q \times K)$ are associated with Q time samples (e.g., fMRI) they can be call *canonical sequences*
Canonical analysis has not been used as much a univariate analysis in functional neuroimaging.

3.18.10 Example on canonical analysis

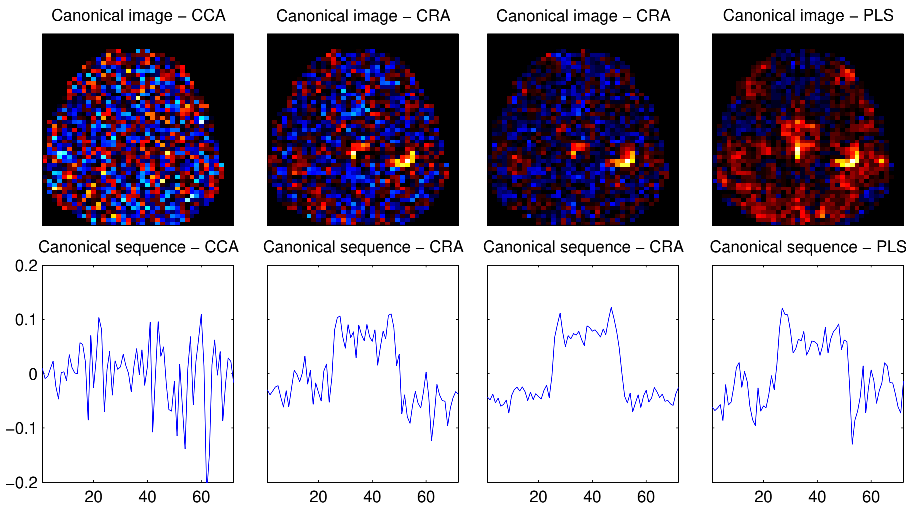

Figure 3.11: Canonical ridge analysis of an fMRI data set. The top row are the canonical images while the bottom row is the eigensequences.

Figure 3.11 shows an example of canonical ridge analysis as the ridge parameter is varied. The data is a single subject multiple runs fMRI block design where each run contains 72 scans: 24 baseline, 24 activation, 24 baseline with the activation consisting of left-handed finger-to-thumb opposition, TR $= 2.5$s and voxel size $3.1 \times 3.1 \times 8$mm. The data was acquired on the Massachusett General Hospital by Robert Savoy, masked and motion realigned before analysis. A $\mathbf{Y}(N \times Q)$ matrix is constructed as an indicator with data matrix with $Q = 72$ as agnostic labeling of each of the scan within the run, i.e., first scan of the first run is in the same group as the first scan of the second run. The $\mathbf{X}$ matrix contains the fMRI image data. The top row of figure 3.11 displays the canonical images in a transversal slice through the primary motor cortex and the supplementary motor area (SMA), and the lower run the canonical sequences. The different columns represent different settings of the ridge parameter k_X for the data matrix $\mathbf{X}$. The left plot has $k_X = 0$ corresponding to canonical correlation analysis with preliminary SVD, and the right plot has $k_X = 1$ (c.f. equation 3.125) corresponding to the “PLS”. To two middle plot are for $0 < k_X < 1$ with the third from left being the optimal.

3.19 From image to points

When the summary image has been formed from the statistical analysis it is often summarized with a set of points or *foci*. A “foci”, “hot spot” or “location” can be defined and found in number of different ways:

- Manually, where the researcher look at the summary images and selects an appropriate voxel, e.g, the center of an activated region.