Canonical Ridge Analysis with Ridge Parameter Optimization

Finn Årup Nielsen[1],
Lars Kai Hansen[1],
Stephen C. Strother[2]

[1] Section for Digital Signal Processing,
Department of Mathematical Modelling,
Technical University of Denmark,
Lyngby, Denmark

[2] PET Imaging Service
Medical Center,
Veterans Administration,
Minneapolis, Minnesota, USA

The Human Brain Project P20 MP57180
The Danish Research Councils

Abstract

We present a general model — the so-called *Canonical Ridge Analysis* model — that encompasses several multivariate linear analysis methods: Canonical Correlation Analysis, Partial Least Square, Orthonormalized Partial Least Square, Fisher Linear Discriminant.

By the introduction of a *ridge* parameter we are able to interpolate between Canonical Correlation Analysis and Partial Least Square. We will optimize our Canonical Ridge model so that the result of the analysis yield better generalization than the previous multivariate analyses.

We recommend Canonical Ridge Analysis should be favored over Canonical Correlation Analysis and Partial Least Square.

Canonical Ridge Analysis

- Canonical Ridge Analysis combine Canonical Correlation Analysis and the "Ridge" technique.
- Discovered by Vinod in 1976 [13] (referenced by Mardia [8]). The word "Canonical Ridge" is Vinod's invention).
- Rediscovered later by N. A. Campbell [2] in 1980 (reviewed in Ripley [12]). More general approach: Shrinkage methods.
- Related to Hoerl and Kennard's Ridge Regression [5].
- Related to *weight decay* in artificial neural networks. [1] [12].
- Compatible with the general idea of *regularization*: "Adding a penalty term", "Smoothing constraints", ...
- The model is an instance of a general class of multivariate linear models that investigates the covariance/correlation between a data matrix and a design matrix

Terminology ...

- All the multivariate linear models described here model the covariance/correlation between two matrices. In our case a design matrix $\mathbf{G}$ and a data matrix $\mathbf{X}$
- The data matrix $\mathbf{X}$ is a scan $\times$ voxel matrix

$$\begin{array}{cc} & \text{Voxels} \\ \mathbf{X} = & \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & \vdots & & \\ x_{i1} & \cdots & x_{ij} & & \\ \vdots & & & & \\ x_{n1} & & & & x_{np} \end{bmatrix} & \text{Scans} \end{array} \quad (1)$$

- The design matrix is a scan $\times$ design variables. Design variables could be, e.g., group indicators or basis function for modeling the hemodynamic response (in fMRI)

$$\begin{array}{cc} & \text{Design variables} \\ \mathbf{G} = & \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1q} \\ \vdots & & \vdots & & \\ x_{i1} & \cdots & x_{ij} & & \\ \vdots & & & & \\ x_{n1} & & & & x_{nq} \end{bmatrix} & \text{Scans} \end{array} \quad (2)$$

... Terminology

By singular value decomposition of a data matrix:

$$\mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}} = \mathbf{X} \quad (3)$$

... we get three matrices:

- $\mathbf{\Lambda}$ contains the eigenvalues.
- The vectors in $\mathbf{V}$ are eigenvectors and in neuroimaging analysis they have been called *eigenimages* by Friston.
- We call the columns in $\mathbf{U}$ the *eigensequences*.

Here we analyze the matrices from a covariance / correlation analysis such as Partial Least Square: [9]

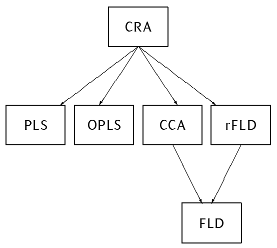
$$\mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}} = \mathbf{G}^{\mathsf{T}}\mathbf{X} \quad (4)$$

... to be called:

- *Canonical images* for the vectors in $\mathbf{V}$.
- *Canonical sequences* for the vectors in $\mathbf{U}$

$\mathbf{V}$ and $\mathbf{U}$ are also called the canonical correlation vectors.
The square root of the diagonal elements of $\mathbf{\Lambda}$ are called the canonical correlation coefficients.

Hierarchy of Multivariate Analyses



CRA — Canonical Ridge Analysis is a generalization of Partial Least Square, Orthonormalized Partial Least Square and Canonical Correlation Analysis.
FLD — Fisher's Linear Discriminant — is concerned with finding *one* canonical vector, — contrary to the others multivariate analyses, and this canonical vector is similar to the first CCA canonical vector.
rFLD — Ridge FLD [6] — regularizes the FLD the same way CRA regularizes CCA.

Overview (From [14])

- SVD, PCA, SSM. Singular Value Decomposition, Principal Components Analysis, Subprofile Scaling Model. [4] [10]

$$\mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}} = \mathbf{X} \quad (5)$$

- PLS. Partial Least Square. [9]

$$\mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}} = \mathbf{G}^{\mathsf{T}}\mathbf{X} \quad (6)$$

- OPLS. Orthonormalized Partial Least Square. [14]

$$\mathbf{U}'\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}} = \left(\mathbf{G}^{\mathsf{T}}\mathbf{G}\right)^{-1/2}\mathbf{G}^{\mathsf{T}}\mathbf{X} \quad (7)$$

$$\mathbf{U} = \left(\mathbf{G}^{\mathsf{T}}\mathbf{G}\right)^{-1/2}\mathbf{U}' \quad (8)$$

- CCA. Canonical Correlation (or: CVA. Canonical Variate Analysis). [3] [8]

$$\mathbf{U}'\mathbf{\Lambda}\mathbf{V}'^{\mathsf{T}} = \left(\mathbf{G}^{\mathsf{T}}\mathbf{G}\right)^{-1/2}\mathbf{G}^{\mathsf{T}}\mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1/2} \quad (9)$$

$$\mathbf{U} = (\mathbf{G}^{\mathsf{T}}\mathbf{G})^{-1/2}\mathbf{U}' \quad (10)$$

$$\mathbf{V} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1/2}\mathbf{V}' \quad (11)$$

CRA — Canonical Ridge Analysis

We can formulate a normalization that can contain both the CCA and PLS. Here for the data matrix:

$$\left[(1-k_X)\,\mathbf{X}^{\mathsf{T}}\mathbf{X} + k_X\mathbf{I}\right] \quad (12)$$

With $k_X = 0$ we get CCA and with $k_X = 1$ we end up with PLS and we can interpolate between PLS and CVA with $0 < k_X < 1$
The full equation being:

$$\mathbf{U}'\mathbf{\Lambda}\mathbf{V}'^{\mathsf{T}} = \left[(1-k_G)\,\mathbf{G}^{\mathsf{T}}\mathbf{G} + k_G\mathbf{I}\right]^{-1/2}\mathbf{G}^{\mathsf{T}} \times \mathbf{X}\left[(1-k_X)\,\mathbf{X}^{\mathsf{T}}\mathbf{X} + k_X\mathbf{I}\right]^{-1/2} \quad (13)$$

To get back to the canonical images and sequences we have to apply the normalization again:

$$\mathbf{U} = \left[(1-k_G)\,\mathbf{G}^{\mathsf{T}}\mathbf{G} + k_G\mathbf{I}\right]^{-1/2}\mathbf{U}' \quad (14)$$

$$\mathbf{V} = \left[(1-k_X)\,\mathbf{X}^{\mathsf{T}}\mathbf{X} + k_X\mathbf{I}\right]^{-1/2}\mathbf{V}' \quad (15)$$

SVD-CVA
Canonical Variate Analysis with Singular Value Decomposition

With rank-deficient data matrix — as typical in neuroimaging data — we first decompose $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}^*\mathbf{\Lambda}^*\mathbf{V}^{*\mathsf{T}} \quad (16)$$

$$\mathbf{X}^* = \mathbf{U}^*\mathbf{\Lambda}^* = \mathbf{X}\mathbf{V}^* \quad (17)$$

$\mathbf{X}^*$ contains the projected data matrix and has full rank. We singular value decompose the covariance matrix:

$$\mathbf{U}'\mathbf{\Lambda}\mathbf{V}'^{\mathsf{T}} = \left(\mathbf{G}^{\mathsf{T}}\mathbf{G}\right)^{-1/2}\mathbf{G}^{\mathsf{T}}\mathbf{X}^*\left(\mathbf{X}^{*\mathsf{T}}\mathbf{X}^*\right)^{-1/2} \quad (18)$$

Apply the normalization to get to the canonical correlation vectors:

$$\mathbf{U} = \left(\mathbf{G}^{\mathsf{T}}\mathbf{G}\right)^{-1/2}\mathbf{U}' \quad (19)$$

$$\mathbf{V}^{\#} = \left(\mathbf{X}^{*\mathsf{T}}\mathbf{X}^*\right)^{-1/2}\mathbf{V}' \quad (20)$$

To get back to voxel space we rotate using the eigenimages:

$$\mathbf{V} = \mathbf{V}^*\mathbf{V}^{\#} \quad (21)$$

However, $\mathbf{X}^*\left(\mathbf{X}^{*\mathsf{T}}\mathbf{X}^*\right)^{-1/2}$ is spherical if all components are maintained! This means that there is no principal axis in the covariance matrix, and we are likely to end up with canonical images containing noise.

SVD-CRA
Canonical Ridge Analysis with Singular Value Decomposition

Due to rank-deficiency in typical neuroimaging data we choose to use a modified form of canonical ridge analysis using singular value decomposition.

$$\mathbf{X} = \mathbf{U}^*\mathbf{\Lambda}^*\mathbf{V}^{*\mathsf{T}} \quad (22)$$

$$\mathbf{X}^* = \mathbf{U}^*\mathbf{\Lambda}^* = \mathbf{X}\mathbf{V}^* \quad (23)$$

$\mathbf{X}^*$ contains the projected data matrix and has full rank. However, it is not overdetermined, and to get principal axes we regularize the normalization as in CRA without SVD:

$$\mathbf{U}'\mathbf{\Lambda}\mathbf{V}'^{\mathsf{T}} = \left[(1-k_G)\,\mathbf{G}^{\mathsf{T}}\mathbf{G} + k_G\mathbf{I}\right]^{-1/2}\mathbf{G}^{\mathsf{T}}\ldots \mathbf{X}^*\left[(1-k_X)\,\mathbf{X}^{*\mathsf{T}}\mathbf{X}^* + k_X\mathbf{I}\right]^{-1/2} \quad (24)$$

Apply the normalization to get to the canonical correlation vectors:

$$\mathbf{U} = \left[(1-k_G)\,\mathbf{G}^{\mathsf{T}}\mathbf{G} + k_G\mathbf{I}\right]^{-1/2}\mathbf{U}' \quad (25)$$

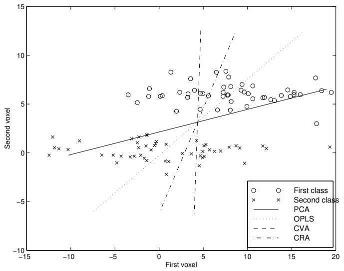
$$\mathbf{V}^{\#} = \left[(1-k_X)\,\mathbf{X}^{*\mathsf{T}}\mathbf{X}^* + k_X\mathbf{I}\right]^{-1/2}\mathbf{V}' \quad (26)$$

And, finally, back to voxel space using the eigenimages:

$$\mathbf{V} = \mathbf{V}^*\mathbf{V}^{\#} \quad (27)$$

Illustration

It is hard to illustrate a rank-deficient problem so here we consider a well-posed toy problem: A "two voxel brain" with 50 activation scans and 50 base line scans. As there are only two groups there will be only one canonical component.



CRA interpolates between the CCA canonical component (image) and the PLS Canonical component (image).

Optimization of the ridge parameter

- The ridge parameter should not be set ad hoc.
- For ridge FLD Loh [6] [7] has suggested a direct method for computing the ridge parameter
- For canonical ridge analysis we must use some kind of cross-validation/bootstrap.

Generalization

- The model should be optimized (i.e., the ridge parameter should be estimated) so it has the best *generalization*. I.e., the model should perform well when evaluated on a independent test set [11].
- The model should *not* be optimized for best *reproducibility*: The most conservative model (the most regularized model) would the most reproducible. I.e., the partial least square model would be the best.

Performance measure

- Canonical Correlation Analysis finds the maximal correlation between two groups of variables (the design matrix and the data matrix)
- The performance measure for our model should be the same, however it should be evaluated on a independent test set
- We compute the canonical images $V_{\mathcal{D}}$ using a training set $\mathcal{D}$. The images are used to project a test set $\mathcal{T}$ onto new canonical sequences:

$$\mathbf{U}_{\mathcal{T}} = \mathbf{X}_{\mathcal{T}} \mathbf{V}_{\mathcal{D}} \tag{28}$$

This $\mathbf{U}_{\mathcal{T}}$ matrix is not normalized and can be exposed to a simple canonical correlation analysis where the objective would be the size of the canonical correlation coefficients.

If the design matrix only contains group indicators we can use a sort of Wilk's Λ ratio computing the between $\mathbf{B}$, total $\mathbf{T}$ and within variance $\mathbf{W}$ of $\mathbf{U}_{\mathcal{T}}$. The ratio (matrix) becoming:

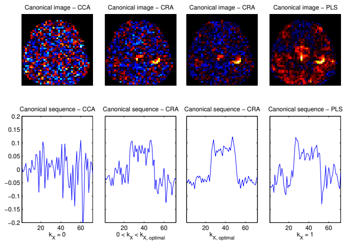
$$\mathbf{\Lambda} = \mathbf{W}^{-1}\mathbf{T} \tag{29}$$

If we are only concerned with the first canonical image this ratio matrix becomes a scalar value.

Example: fMRI study ...

Single subject sequential left hand finger opposition task (SFO), R. Savoy, Massachusetts General Hospital: 9 runs with 72 scans each.

We construct a design matrix with 72 groups, use the 4 first runs to find the canonical images and the 4 last for validation

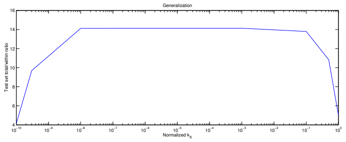


As expected the CCA canonical image and sequence are completely dominated by noise. The PLS on the other hand seems to pick up areas that have no importance for SFO. The CRA with the optimal k_X yields the "cleanest" canonical image and sequence.

... Example: fMRI study

The optimal value of k_X depends on the performance measure we use.

In this study we have restricted ourselves to only look at the first canonical sequence. Thus the total variance / within variance ratio becomes a scalar value we can plot:



We have normalized the k_X parameter:

$$\left[(1-k_X)\,\mathbf{X}^{*\mathsf{T}}\mathbf{X}^{*} + \beta k_X \mathbf{I}\right] \tag{30}$$

with a constant β proportional to the trace of $\mathbf{X}^{*\mathsf{T}}\mathbf{X}^{*}$

The plot clearly shows that the optimal k_X is neither the k_X for PLS ($k_X = 1$) nor CCA ($k_X = 0$, only shown in the limit)

The setting of the ridge parameter is not critical for this particular data set — as long as one does not set it "near" the PLS or CCA value

Conclusion

- Partial Least Square
  - + Stable determination of normalizing dispersion matrix
  - – Not optimal discrimination
  - – Sensitive to structured noise
- Canonical Correlation Analysis
  - + Potentially best discrimination
  - – Sensitive to noise
- Canonical Ridge Analysis
  - + Most generalizable discrimination
  - – Estimation of the ridge parameter needed: More computation time is needed.

Remark:
Whether one sees the presented multivariate analyses as "final" analyses or exploratory analyses, will depend on the basic assumption: Should we regard the neuroimages as images of functional segregation or are the neuroimages representing the mental state as a vector in a hyperspace.

Future work

- Comparison to other forms of regularization of the normalizing dispersion matrix, e.g., discarding SVD-components.
- Method for determining the dimension of the interesting subspace: "How many canonical components?"
- Guidelines for efficient determination of the ridge parameter: Optimization procedures, validation issues.
- Proposal for a convenient normalization of the regularization term
- Learning curve: "How many scans?", How many subjects?"
- Investigation of the effect of mean estimate: "Is the sample mean the best?"
- Investigate if it is possible to do Canonical Ridge Analysis without the initial singular value decomposition.

Bibliography

[1] Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press Oxford, 1995.

[2] Norm A. Campbell. Shrunken estimators in discriminant and canonical variate analysis. *Applied Statistics*, 29(1):5–14 (24), 1980.

[3] K. J. Friston, C. D. Frith, R. S. J. Frackowiak, and R. Turner. Characterizing dynamic brain responses with fMRI: A multivariate approach. *Neuroimage*, 2:166–172, 1995.

[4] K. J. Friston, C. D. Frith, P. F. Liddle, and R. S. J. Frackowiak. Functional connectivity: The principal-component analysis of large (PET) data sets. *J. Cereb. Blood Flow Metab.*, 13:5–14, 1993.

[5] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[6] Wei-Liem Loh. On linear discriminat analysis with adaptive ridge classification rules. *Journal of Multivariate Analysis*, 53:264–278, 1995.

[7] Wei-Liem Loh. Linear discrimination with adaptive ridge classification rules. *Journal of Multivariate Analysis*, 62(2):169–180, 1997.

[8] Kantilal Varichand Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

C. L. Grady. Spatial pattern analysis of functional brain images using Partial Least Square. *NeuroImage*, 3:143–157, 1996.

[10] J.R. Moeller, S. C. Strother, J.J.Sidtis, and D. A. Rottenberg. Scaled subprofile model: A statistical approach to the analysis of functional patterns in positron emission tomographic data. *Journal of Cerebral Blood Flow and Metabolism*, (7):649–658, 1987.

[11] Niels Mørch, Lars K. Hansen, Stephen C. Strother, Claus Svarer, David A. Rottenberg, Benny Lautrup, Robert Savoy, and Olaf B. Paulson. Nonlinear versus linear models in functional neuroimaging: Learning curves and generalization crossover. Lecture Notes in Computer Science 1230, pages 259–270. 15th International Conference, IPMI'97, Poultney, Vermont, USA, Springer, June 1997.

[12] Brain D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[13] H. D. Vinod. Canonical ridge and econometrics of joint production. *J. Econometrics*, 4:147–166, 1976.

[14] Keith J. Worsley. An overview and some new developments in the statistical analysis of PET and fMRI data. *Human Brain Mapping*, 5(4):254–258, 1997.