

Fast Non-linear Prediction with Large Number of Voxels with Kernel-based Partial Least Squares Regression

Human Brain Mapping Conference, June 2006

Paul Rodriguez
Department of Cognitive Science and Research Imaging Center
University of California at Irvine

UCIrvine

Introduction

Nonlinear partial least squares (PLS) regression was used to predict feature ratings in the HBM 2006 Pittsburgh Brain Activity Interpretation Competition.

Kernel based PLS has not been used in fMRI, although a linear PLS method for statistical inferencing has been described in detail (e.g. McIntosh et al, 1996).

The Gaussian kernel trick enables efficient processing, thus, this method would seem to be most effective when many voxels are useful for prediction.

This intuition was borne out by best performance for prediction of internal subject assessment, which are the least correlated among individual subjects. Excellent performance was achieved giving a 7th place showing overall (mean $r=0.432$).

Algorithm Overview

The basic PLS procedure:

Given X a mean-centered data matrix of size $T \times N$, where each row is a volume, and Y a mean centered target time course.

1. Take the cross-covariance matrix: X^*Y
2. Get the first principle component: u
3. Solve the linear regression equation: $X^*u^*B=Y$
4. Deflate X and Y with u , and repeat for upto $\text{rank}(X)$ number of components.

For k components we have the regression equation:

$$X^*[u_1 \dots u_k]^* [B_1 \dots B_k] = Y$$

Nonlinear PLS with a Gaussian kernel:

Let $K = X^*X'$ be the kernel, where $k_{ij} = \exp(-(x_i - x_j)^2 / v)$ (x_i is i -th column and v is the kernel variance parameter).

Ensure that only K is used as follows:

Find u from $X^*Y YX$, (also called the kernel trick): $X^*Y YX u = \lambda u$

Let $X'b = u$ and premultiply by X giving: $Y Y K^* b = \lambda b$

b is principle component of the $Y Y K^*$ matrix and regression equation is now $K^*[b_1 \dots b_k]^* [B_1 \dots B_k] = Y$

Importantly, K is only a $T \times T$ matrix so that algorithm efficiently process large numbers of voxels.

For prediction, replace XX' with ZX' in the linear equation where Z is the test set. (for details and pseudo code, see Shawe-Taylor & Cristianini, 2004)

Figure 1. Toy Problem Demonstration

2 voxels, 20 observations, 2 classes: '+' (1) or 'o' (-1), non-linearly separable.

Test with a grid of voxel values to produce a contour plot in original input space.

The color value is essentially a factor score. The dividing line between the classes (color value 0) bends to nearly separate the classes.

The 2nd component picks off particular observations, which may or may not improve fit.

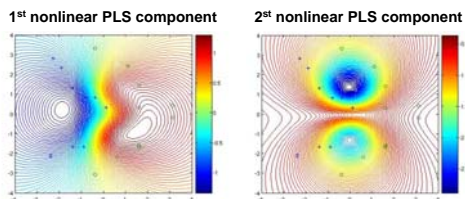


Figure 1.

Application to Brain Activity Interpretation

Pre-processing and Training.

After removing the first 4 volumes, realignment and high pass filtering (512 second period) performed with SPM2 (Wellcome Center for Neuroimaging, UC London).

1 data matrix extracted for each subject and each movie of size $T \times N$, where $N \approx 30,000$ voxels, T is the number of volumes in the scanning run.

For each subject, trained with data from movie 1 (or 2) and tested on movie 2 (or respectively, 1). This produced 2 possible sets of parameters for each subject to predict movie 3

Table 1. A Course Grid Search of Parameter Space

The PLS algorithm requires choosing a kernel variance parameter and the number of components. Using fMRI data requires choosing voxels and volumes for training. Additionally, I chose to use lagged datasets to capture HRF variations and slice timing differences.

72 tests were run for each of 13 feature rating targets and each parameter set, for a total of approximately $72 \times 6 \times 13 \approx 5600$ tests.

I merely selected the best parameters and used that to build predictions for movie 3.

Parameter to Choose	Range of Values	Mean (std)
voxels with high +/- correlation to target	threshold values of .05, .1, .15, .20, or .25	0.153 (.045) = 13551 (6910) voxels
Kernel variance factor v	mean of $(X(i,:) - X(:,j))^2$ multiplied by 1.5, 2.5, or 3.5	2.74 (.42)
Amount of lag in dataset	let $X = [X(t) \ X(t-1) \dots \ X(t-lag)]$ with $lag=1, 2, \text{ or } 3$	2.46 (.29)
Training Volumes	volumes during movie plus full rest or partial rest periods	51 % used only partial rest periods

Table 1.

Figure 2. PLS components

The first component extracted matched the target, and then each successive component varied slightly.

An example with the 1st (blue) and 2nd components (green).

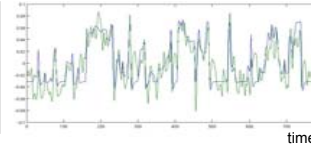


Figure 2.

Figure 3. Choosing number of components

The correlation (r) to the training data with the best prediction, $\mu=0.89$ (0.13), about 6 components, worked better than information criterion or variance plots.

An example of correlation for training (blue) and test set (red). Here the best prediction occurs with 10 components and $r=0.98$ on the training set.

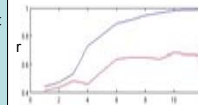


Figure 3.

Figure 4. Post-processing.

All predictions were smoothed by a moving average of +/- 1 time point window and averaged across subjects (except Arousal), and in some cases smoothed further to match the smoothness in the feature ratings.

An example of Amusement target and movie times (blue), raw prediction (green), and prediction after averaging and smoothing (red); final correlation is about 0.44.

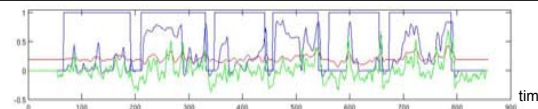


Figure 4.

Final Submissions

The mean predictions across movie 1 and 2 were averaged to produce predictions for movie 3 in the first submission. In the second submission, only movie 1 or 2 predictions (but still averaged across subjects) were tried, but this helped just a few subjects and features, including Attention and Arousal.

Figure 5A. Correlation Results

The correlation scores for training/test with movies 1&2 (blue) and the final predictions for movie 3 (red) show fairly consistent performance (* indicates best average in competition; * indicates within 5% of best for movie 3).

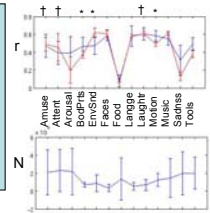


Figure 5B. Number of Voxels Used

Mean (std) of number of voxels used for movie prediction. A maximum of about 60K was used.

Figure 5.

Figure 6. Number of Voxels X Correlation

There is a negative correlation between prediction in training/test with movies 1&2, and number of voxels. Although for <30K voxels the correlation is slightly positive.

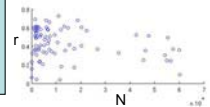


Figure 6.

Figure 7. Voxels Locations

For subject 3, train movie1/test movie 2, the following shows voxels used for Amusement (1719 total across lags 0, 1, 2, and 3, $r=0.44$) and Laughter (18765 total across lags 0, 1, 2, 3, $r=0.64$). The number of voxels decrease slightly with longer lags. It is likely that some voxels represent noise from artifacts or task, but the algorithm was allowed to sort it out.

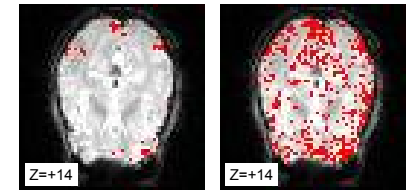


Figure 7.

Future Work

Although information criterion did not seem to pick out the best number of components, more work needs to be done here.

Due to time constraints, I only did cursory tests with cross-validation methods across both data sets and across subjects, with artifact removal in preprocessing, spatial constraints on voxels, or a finer search of parameter space.

Main Findings

The nonlinear PLS consistently outperformed the linear PLS in prediction, even though both are able to solve the regression problem.

In many cases, using lagged datasets and 10-30K voxels (over all lags) improved predictions in training/test comparisons.

Conclusion

An important advantage of kernel-based PLS that makes it particularly attractive for fMRI, in comparison say to a neural network, is that a non-linear analysis using all voxels can be processed in 2-3 minutes on a typical workstation (in Matlab™ even). An important disadvantage of this method, in comparison say to support vector machines, is that little is known about effective means of penalization to ensure sparse solutions and avoid overfitting.

References

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A. Schouten, J.L., Peitri, P. (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293, 2425-2430.
McIntosh, A.R., Bookstein, F.L., Haxby, J.L., Grady, C.L. (1996) Spatial Pattern Analysis Using Partial Least Squares.
Shawe-Taylor, J., Cristianini, N. (2004) *Kernel Methods for Pattern Analysis*. Cambridge University Press.

Acknowledgements

This work was supported by the Research Imaging Center and Department of Cognitive Science at UCIrvine. And a special thanks to Dr. Walter Schneider and the experienced based cognition group at Pittsburgh for organizing the competition.