Kernel Methods for fMRI Pattern Prediction

Yizhao Ni, Carlton Chu, Craig J Saunders and John Ashburner

***Abstract*— In this paper, we present an effective computational approach for learning patterns of brain activity from the fMRI data. The procedure involved correcting motion artifacts, spatial smoothing, removing low frequency drifts and applying multivariate linear and non-linear kernel methods. Two novel techniques are applied: one utilizes the Cosine Transform to remove low-frequency drifts over time and the other involves using prior knowledge about the spatial contribution of different brain regions for the various tasks. Our experiment results on the PBAIC2007 competition data set show a great improvement for brain activity prediction, especially on some sensory experience such as hearing and vision.**

I. Introduction

Functional magnetic resonance imaging (fMRI) is a technique for measuring human brain activities, by detecting changes in blood oxygenation and flow that it elicits. This method is usually used to produce activation maps showing which parts of the brain are involved in particular mental processes. The time series at each voxel is modeled as a linear combination of the experimental conditions, and statistical tests are applied to the regression coefficients to infer where particular stimuli have significant effects on the pattern of brain activity.

An alternative challenge is to predict the experimental conditions from the fMRI data. Typically, a pattern-recognition procedure is trained with fMRI data and the known mental states (e.g., whether the subject is looking at a face) during the scanning. The objective is then to predict unknown mental states given only the fMRI data. In the last two years, the *Pittsburgh Brain Activity Interpretation Competition*[1] (PBAIC) has been held, challenging multiple groups (motivated by a $\$10,000$ prize) to use state-of-the-art techniques to infer subject-driven actions and sensory experience from a rigorously collected fMRI data set. Several machine learning techniques, such as neural network and kernel methods, have been applied by the entrants. E. Olivetti et al. (2006), the winner of PBAIC2006, sought the mutual information between each feature (voxel) and each task function and selected those features with highest mutual information as the input. They then used a neural network to learn the mapping from the feature space to the target functions. Because they used an explicit expression of the feature space, computation expense limited them to using about 100 features.

Yizhao Ni & Craig J Saunders are with ISIS group, Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK (e-mail: Yizhao.NI@googlemail.com, cjs@ecs.soton.ac.uk).

Carlton Chu & John Ashburner are the researchers in Wellcome Trust Centre for Neuroimaging, 12 Queen Square, London, WC1N 3BG, UK (email: carltonchu1@gmail.com, john@fil.ion.ucl.ac.uk).

[1]http://www.ebc.pitt.edu/2007/competition.html

Recently, kernel based techniques have shown promising results when applied to fMRI analysis (D. Hardoon and L. M. Manevitz, 2005). For PBAIC2007, D. Chigirev, et al. (2007) also utilized radial basis function (RBF) kernels to measure the similarity between different brain states, and used ridge regression and a general SVM to learn the brain activity associated with a specific motor activity. To our knowledge, most common approaches to applying kernel methods for the fMRI analysis are purely data-driven, exploratory techniques. For example, D. Hardoon and L. M. Manevitz (2005) used all voxels to construct their kernel, no matter whether they are related to the task or not. D. Chigirev, et al. (2007) filtered out voxels with low mutual information and constructed RBF kernels with the remainder. In our experiments, we observed that such purely data-driven exploratory techniques may be too general to explore some types of brain activity, such as auditory experience, and thus do not produce the best performance.

Typically, a training dataset would consist of a series of several hundred volumetric fMRI scans (images), where each scan is a volume of around $64 \times 64 \times 34$ voxels. Kernels are essentially square, symmetric and positive definite matrices that encode measures of similarity between each pair of scans. *Ugly Duckling Theorem* (S. Watanabe, 1969) tells us that prior knowledge is essential for quantifying the similarity between things, so in this work we made use of several pieces of knowledge in order to achieve more informative similarity measures:

- Brain activity occurs in the gray matter of the brain, so signal from other regions of the scans can be ignored.
- The variance associated with brain activity is more spatially smooth than the noise in the images.
- Several years of brain imaging experiments tell us which brain regions are most likely to be important for particular tasks.
- Subjects move slightly in the scanner, and this movement should not be considered informative about the task.
- Series of fMRI scans contain low frequency drifts over time. This drift should not be considered informative.

Our winning entry to the PBAIC 2007 competition used kernel methods, but incorporated the above prior knowledge (our "hypothesis-driven" technique) in the kernel generation. Our goal was to demonstrate simple and efficient methods which are easy to implement as well as computationally fast. We also show that with specifically designed kernels, which utilize prior knowledge (hypothesis), we can achieve better predictions for certain brain states. In addition, we confirmed that better image pre-processing improves prediction accu-

racy.

The rest of this paper is structured as follows: In Section II, we begin by describing our data pre-processing, including the detrending and feature selection. Then, we present our learning scheme and kernel design in Section III. Afterwards, in Section IV we evaluate the performance of our model and kernels using the PBAIC2007 competition data set and compare the results with other state-of-the-art techniques. Finally, we draw conclusions and mention areas for future work in Section V.

II. Data Pre-processing

Figure 1 illustrates our pre-processing of the fMRI data. We follow the key processing steps (black boxes) used in fMRI analysis, while introducing a unique noise filter and the hypothesis–driven feature extraction technique (blue boxes). Generally speaking, the fMRI sequence was re-aligned to reduce variance in the data that arises through subject motion, then spatially smoothed by convolving with a Gaussian smooth function and detrended by high–pass filtering. Finally, tissue segmentation was done directly on an EPI scan of each subject, using the algorithm in the SPM5 software (SPM5, 2005). A specified mask was generated to remove irrelevant tissue classes and the remaining voxels were used as the input features. The reader is refered to R. S. J. Frackowiak, et al., (2003) for further details. Below we present the two unique steps: the noise filter and the feature extraction.

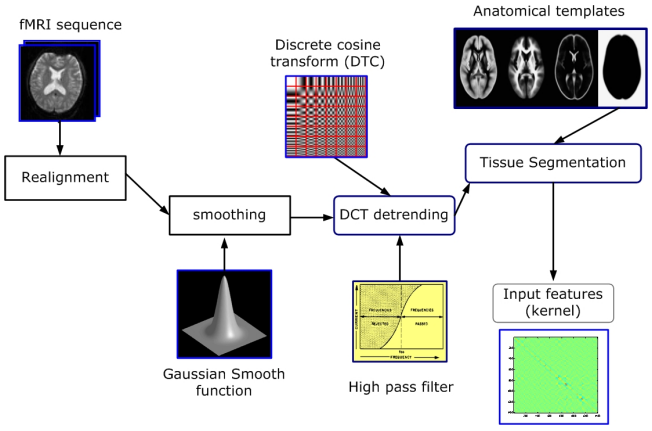


Fig. 1. The pre-processing of the fMRI data. The black boxes are the key processing steps used in fMRI analysis, while the blue boxes are our noise filter and hypothesis–driven feature extraction technique.

A. Discrete cosine transform (DCT) noise filter

Low frequency drift has often been reported in time series fMRI data. This drift has often been attributed to physiological noise or subject motion, but few studies have been done to test this assumption (Smith A.M. et,al., 1999). In our preliminary experiment, we observed that there was large amount of low frequency (0.0 - 0.015 Hz) drift in the linear detrend fMRI data provided by the PBAIC committee. Hence, we utilized a discrete cosine transform (DCT) to removed additional low frequency noise. Mathematically, for each voxel v, the time sequence $\mathbf{v} = \{v_k\}_{k=0}^{K-1}$ is collected from K time points and transformed into a frequency sequence $\{f_l\}_{l=0}^{K-1}$

$$f_l = \sqrt{\frac{2}{K}} \sum_{k=0}^{K-1} v_k \cos\left[\frac{\pi}{K}(k+\frac{1}{2})l\right] \quad l = 0, \ldots, K-1. \tag{1}$$

After pruning the low frequency noise (basis functions) $\mathbb{L}$, the detrend sequence $\bar{\mathbf{v}} = \{\bar{v}_k\}_{k=0}^{K-1}$ is obtained by the inverse transforms

$$\bar{v}_k = \sqrt{\frac{2}{K}} \sum_{l \notin \mathbb{L}} f_l \cos\left[\frac{\pi}{K} l(k+\frac{1}{2})\right] \quad k = 0, \ldots, K-1. \tag{2}$$

Observe that the DCT can be represented as a matrix operation. Let $\mathbf{D}$ be the $K \times L$ DCT matrix with $D_{k,l} = \sqrt{\frac{2}{K}} \cos\left[\frac{\pi}{K}(k+\frac{1}{2})l\right]$ and L denoting the number of the basis functions, it is easy to prove that the detrend sequence is

$$\bar{\mathbf{v}} = (\mathbf{I} - \mathbf{D}\mathbf{D}^T)\mathbf{v} \tag{3}$$

Where the matrix $\mathbf{R} = (\mathbf{I} - \mathbf{D}\mathbf{D}^T)$ is called the residual forming matrix.

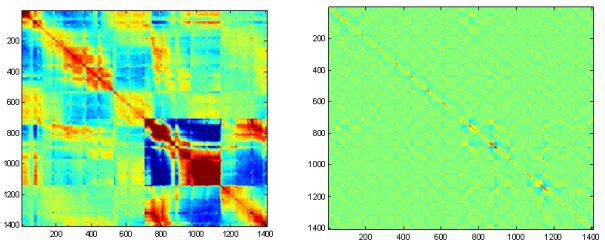
Using this matrix operation, we can apply the detrending directly on the fMRI kernel, which is time efficient. Suppose we defined the input $\mathbf{X}$ as a $P \times M$ matrix, which contains M input points $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_M]$ and each vector $\mathbf{x}_i$ contains P voxels $\mathbf{x}_i = [v_{i,1}, \ldots, v_{i,P}]^T$. Then, given $\mathbf{R}$, the detrended kernel can be expressed as

$$\bar{\mathbf{K}}_{DCT} = \langle \mathbf{X}\mathbf{R}, \mathbf{X}\mathbf{R} \rangle = \mathbf{R}^T\mathbf{X}^T\mathbf{X}\mathbf{R} = \mathbf{R}^T\mathbf{K}\mathbf{R} \tag{4}$$

As applying improper transfer function or potential aperiodic basis functions in DCT may cause unwanted distortion–Gibbs phenomena and aliasing, we used the cross–validation technique to decide the optimal number of basis functions L. In our experiments, we used $L = 8$ basis functions (including the constant term); this procedure is equivalent to a high pass filter with a cut–off around $1/176$ Hz. After detrending, much of the variance in the fMRI kernel is removed, which shows the effect of filtering out the low-frequency noise (see Figure 2).

B. Hypothesis–driven Feature Extraction

Blood oxygen level-dependent (BOLD) signal changes arise largely in the gray matter of the brain. Other signal in the scans can be considered as noise, and was therefore excluded from the kernels. A tissue segmentation procedure was used to identify gray matter directly on the functional images (Echo Planar Imaging, EPI), using the procedure in the SPM5 software (SPM5, 2005). Years of brain imaging studies provide prior knowledge about which functional brain regions are likely to be involved in processing different stimuli. It is therefore possible to exclude gray matter voxels



Linear detrend kernel DCT detrend kernel

Fig. 2. The fMRI kernel matrix generated from linear detrend data provided by the PBAIC committee (left) and the fMRI kernel matrix generated from data after DCT detrending (right).

that are believed to be uninformative for particular tasks, by masking out such regions in the scans.

Signal changes in fMRI that are due to brain activity tend to be slightly lower frequency (over space) than the noise. From a Wiener filtering perspective, the signal to noise ratio can be increased by spatially smoothing the scans. We found that accuracy could be increased by convolving the scans with a 6mm full width at half maximum (FWHM) Gaussian kernel.

III. Kernel methods and kernel applications

As mentioned in most manuscripts about kernel methods, it is possible to benefit from two useful properties of these methods

- The kernel trick reduces the computational complexity for high dimensional data as the parameter evaluation domain is reduced from the explicit feature space into the kernel space.
- With an appropriate kernel function one can map the input feature space into higher dimensions. This allows non-linear approaches in the original feature space to be achieved by linear approaches in the higher dimensional space.

In our experiments we use two kernel methods, Kernel Ridge Regression (KRR) and Relevance Vector Regression (RVR), for predicting continuous brain states. An alternative would be to use classification, which predicts categorical states.

Mathematically, we denote the fMRI scan images as $\{\mathbf{x}_i\}_{i=1}^M$ which are embed in a voxel feature space $\mathbf{x} \in \mathcal{R}^P$ and the outputs are values of N different brain activities (task functions) $\left\{\{y_{i,n}\}_{i=1}^M\right\}_{n=1}^N$. Since we deal with the tasks individually, we remove the task index n and abbreviate the output as $\{y_i\}_{i=1}^M$ for clarity.

A. Kernel Ridge Regression

For each task, ridge regression learns a linear operator $\mathbf{w}$ to minimize the squared difference between the predictions $\{\bar{y}_i|\bar{y}_i = \mathbf{w}^T\mathbf{x}_i,\ i = 1,\ldots,m\}$ and the real values $\{y_i\}_{i=1}^M$

$$\mathbf{w} = \arg\min_{\bar{\mathbf{w}}} \sum_{i=1}^{M} (y_i - \bar{\mathbf{w}}^T\mathbf{x}_i)^2 + \lambda\|\bar{\mathbf{w}}\|^2 \tag{5}$$

where λ is the so called regularization parameter. Let $\mathbf{X} = [\mathbf{x}_1,\ldots,\mathbf{x}_M]^T$ and $\mathbf{y} = [y_1,\ldots,y_M]^T$, an analytic solution of $\mathbf{w}$ is $\mathbf{w} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$ with $\mathbf{I}$ denoting the identity matrix.

Note that the input fMRI scans are represented by tens of thousands of active voxels, which makes the direct derivation of $\mathbf{w}$ computationally expensive. Alternatively, if we define the dual variables $\boldsymbol{\theta} = \{\theta_i\}_{i=1}^M$ and apply the Lagrange multiplier technique, we obtain the dual form solution of the ridge regression which simplify the problem into M-parameters estimation

$$\boldsymbol{\theta} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y} \tag{6}$$

where $\mathbf{K} = \mathbf{X}^T\mathbf{X}$ is the well-known kernel matrix. The above dual form formulation is called kernel ridge regression (C. Saunders, et al., 1998) and provides the exact solution to (5) as the primal form does.

In the dual form, the primal weight $\mathbf{w}$ in (5) is expressed as

$$\mathbf{w} = \sum_{i=1}^{M} \theta_i\mathbf{x}_i \tag{7}$$

and the prediction is

$$\bar{y}_j = \mathbf{w}^T\mathbf{x}_j = \sum_{i=1}^{M} \theta_i K(\mathbf{x}_i, \mathbf{x}_j) \tag{8}$$

To predict the output value of a particular fMRI scan, the similarity measures between this scan and all the training fMRI scans are required.

B. Relevance Vector Regression

Relevance Vector Regression (RVR) is formulated in a Bayesian framework while the general expression takes the SVM-like form

$$\bar{y}_j = \sum_{i=1}^{M} w_i K(\mathbf{x}_j, \mathbf{x}_i) + b = \sum_{i=1}^{M+1} w_i\phi_{j,i} \tag{9}$$

where $\boldsymbol{\phi}$ is the $M \times (M+1)$ 'design' matrix $\boldsymbol{\phi} = [\mathbf{K}, \mathbf{1}]$ with $\mathbf{K}$ denoting the kernel matrix and $\mathbf{1}$ denoting a column of ones.

The prior of the weight $\mathbf{w}$ is then modeled as a Gaussian $p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{M+1}\mathcal{N}(w_i|0,\alpha_i^{-1})$ and the solution involves optimizing the following marginal likelihood (type-II maximum likelihood) with respect to the vector of hyper-parameters $\boldsymbol{\alpha}$ and a noise variance σ^2

$$\begin{aligned} P(\mathbf{y}|\boldsymbol{\alpha},\sigma^2) = & \int p(\mathbf{y}|\mathbf{w},\sigma^2)p(\mathbf{w}|\boldsymbol{\alpha})dw \\ & = (2\pi)^{-\frac{N}{2}}|\sigma^2\mathbf{I} + \boldsymbol{\phi}\mathbf{A}^{-1}\boldsymbol{\phi}^T|^{-\frac{1}{2}} \\ & \exp\left\{-\tfrac{1}{2}\mathbf{y}^T(\sigma^2\mathbf{I} + \boldsymbol{\phi}\mathbf{A}^{-1}\boldsymbol{\phi}^T)^{-1}\mathbf{y}\right\} \end{aligned} \tag{10}$$

Where $\mathbf{A} = diag(\alpha_1,\ldots,\alpha_{M+1})$. The best $\boldsymbol{\alpha}$ and noise variance σ^2 can be determined by a EM style approach and we refer the readers to (M.E. Tipping, 2001) or (C. Bishop, 2006) for details.

Predictions through RVR are given by

$$\bar{y}_j = \sum_{i=1}^{M+1} \mu_i \phi_{j,i} \tag{11}$$

With $\boldsymbol{\mu} = \sigma^{-2}\left(\sigma^{-2}\boldsymbol{\phi}^T\boldsymbol{\phi} + \mathbf{A}\right)^{-1}\boldsymbol{\phi}^T\mathbf{y}$ is the posterior mean of the parameter $\mathbf{w}$.

Formulation (11) is similar to the KRR, but for RVR the result is a sparse representation, which implies some of the training fMRI scans don't contribute to the prediction.

The underlying advantage and rationale for applying the two kernel methods relys on the high-dimensional and small sample sized characteristics of fMRI. Compared with other approaches using explicit feature expressions, the above methods are more computationally efficient since the amount of parameters to be learnt (O(M)) is much less than that for the approaches with explicit feature expression (O(P)).

C. kernel applications

Before creating the kernels, we assume that all tasks are independent. That is, for example, what the subject heard won't influence his visual brain activity and vice verse. Based on this assumption, we designed different fMRI kernels for different tasks using the hypothesis–driven feature extraction technique. For the PBAIC2007 competition data set, we create three types of kernels

1) Gray matter kernel (GMK) – As majority of BOLD signal comes from the gray matter, we masked out any non-gray matter voxels (Figure 3 (a)) to reduce the background noise and create the gray matter kernel. In most tasks, this kernel provides good performance and robust results.
2) Auditory cortex kernel (ACK) – The auditory cortex is the region of the brain that is responsible for processing of auditory information and this kernel is created for some tasks that are strongly related to sound (e.g., the "dog barking" task in the PBAIC2007 competition data set). Technically, anatomical templates[2] of the auditory cortex (Figure 3 (b)) were non-linear warp into individual subjects, then the auditory cortex kernels are generated from the masked voxels.
3) Visual cortex kernel (VCK) – The primary visual cortex is the best studied visual area in the brain. It is highly specialized for processing information about static and moving objects and is excellent in pattern recognition. In our experiment, this kernel is created for some tasks that are strongly related to the human's visual experience (e.g., the "interior and exterior" task in PBAIC2007 data set). The same technique is applied as that for creating the auditory cortex kernel.

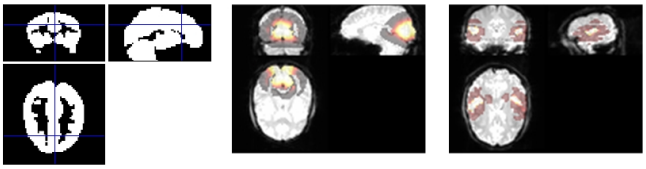
Observed that some brain activities (e.g., emotional experience) may not have a linear pattern of the voxels, we also projected the fMRI data into higher dimensional feature space where the detection of pattern may benefit from the new representations. Technically, We can create such new feature space using non-linear kernels, by applying the "kernel trick" (John Shawe-Taylor and Nello Cristianini, 2004). In our experiment, we proposed two common non-linear kernels, the radial basis function (RBF)

$$\begin{aligned} K_{RBF}(\mathbf{x}_i, \mathbf{x}_j) &= \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2) \\ &= \exp\{-\gamma(K(\mathbf{x}_i, \mathbf{x}_i) - 2K(\mathbf{x}_i, \mathbf{x}_j) \\ &\quad + K(\mathbf{x}_j, \mathbf{x}_j))\} \end{aligned}$$

and the polynomial kernel

$$K_{poly}(\mathbf{x}_i, \mathbf{x}_j) = (\beta + \mathbf{x}_i^T\mathbf{x}_j)^d = (\beta + K(\mathbf{x}_i, \mathbf{x}_j))^d$$

where γ, β, d are functional parameters and are usually learnt through cross-validation. Note that these non-linear kernels can be directly generated from the linear kernel $\mathbf{K}$ and thus are computational efficient.



(a) Gray Matter mask (b) Visual Cortex mask (c) Auditory Cortex mask

Fig. 3. The feature masks generated by SPM5.

IV. Experiments

The experimental data we used is the PBAIC2007 competition data set[3]. In this competition, there are three subjects, each of them played a virtual reality game inside an MRI scanner for roughly 1 hour. The goal is to predict 13 different tasks (feature ratings) for each subject, which are derived from the virtual reality environment. The fMRI scans are of these subjects executing tasks or experiencing the environment while 34 slices were obtained through the head at each time point[4]. The individual fMRI scans are of size 64×64 and thus one training point $\mathbf{x}_i$ contains $64 \times 64 \times 34$ voxels. The scans and ratings are then divided into three sessions, roughly 20 minutes each, in which the first two are for training and the third is for evaluation. Overall, each subject provides 1408 training points each of which has 13 feature ratings (has been quantified into continues or discrete values) and the fMRI scans for the third session (704 time points) are given to predict what the subject did or experienced (feature ratings) in this session, based on the training with the previous two sessions. The performance is measured by correlating the predicted task values with the actual "reference" feature time series data. Details of

[2]The templates from the McConnell Brain Imaging Center can be download freely on http://www.bic.mni.mcgill.ca/cytoarchitectonics/

[3]The data can be downloaded freely for the participants on http://www.ebc.pitt.edu/2007/competition.html.

[4]Actually, each slice is not derived at the same time but with a time delay. However, after pre-processing we treat them as if they were derived at the same time.

the tasks description and how the score is calculated can be found in (PBAIC, 2007).

During the experiments, we concentrated our analysis on three aspects:

1) The effect of the hypothesis–driven feature extraction. We compared our kernel methods plus hypothesis kernels (KM+GMK/ACK/VCK) with neural network plus mutual information features (NN+MIF) (Rajan Patel, 2007) and SVM (and ridge regression) plus RBF kernel (SVM+RBF) (D. Chigirev, et al., 2007). Especially, we focus the comparison on some particular brain activities[5]: auditory experience ("dog barking", "instruction"), visual experience ("interior and exterior") and emotion experience ("valance").
2) The effect of detrending.
3) Comparison between KRR and RVR. We compare several task results produced by KRR and RVR, aiming to explore the impact of the sparse representation used in RVR.

A. Hypothesis Driven versus Data Driven

Table I illustrates the average task scores across the three subjects. In our model, the optimal hypothesis kernels for the tasks are selected by cross-validation.

TABLE I

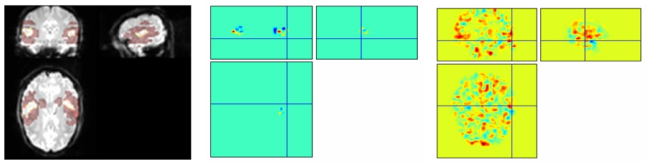
COMPARISON RESULTS WITH DIFFERENT METHODS

Method	KM + GMK/ACK/VCK	NN + MIF	SVM + RBF
Dog	**0.57** (ACK)	0.34	0.26
In & Ex	**0.46** (VCK)	0.28	0.36
Instruction	0.99 (GMK)	0.99	0.99
Valance	0.2 (GMK+RBF)	**0.38**	0.11

In Table I, we are able to observe that the hypothesis–based kernel usually perform the best, especially on tasks "dog" and "interior and exterior". We believe this is because the lower signal to noise ratio in the mask-out region of the brain for these tasks, which were relatively less well predicted. Figure 4 demonstrates the weight volume $\mathbf{w}$ for task "dog" using the auditory cortex kernel and the gray matter kernel respectively. It is clear that with GMK a few non-auditory cortex voxels also contribute to the prediction, which is a potential cause in decreasing the prediction accuracy.

These promising results validate our expectation: with prior knowledge of the task, one can select specified features to further reduce the noise. However, the exception is also observed in the experiment: the emotional experience "valance". Due to the lack of prior knowledge, purely data driven techniques, such as mutual information feature selection, work better. Moreover, we observed that in our model the RBF kernel works best in this task, which implies it is hard to find a linear feature space (i.e. a mask) to characterize this brain activity. We believe this is also the reason why the neural network performed well in this task.

[5]All other competition results are available on http://www.ebc.pitt.edu/2007/2007.html or upon request.



(a) Auditory Cortex (b) weight volume (ACK) (c) weight volume (GMK)

Fig. 4. The weight volumes generated by ACK and GMK respectively. Images from left to right are (a) auditory cortex, (b) weight volume $\mathbf{w}$ generated by ACK and (c) weight volume $\mathbf{w}$ generated by GMK. The train data is collected from subject14 in PBAIC2007 competition data set and the weight volume $\mathbf{w}$ is computed by equation (7).

B. Detrending effect

In Figure 5 we demonstrated 4 task scores of a particular subject with different levels of detrending. The more basis function it uses, the more low frequency signal is removed. In the competition data set, we observed that high level detrending usually produced better results, which implies large drift noise in the raw fMRI scans.

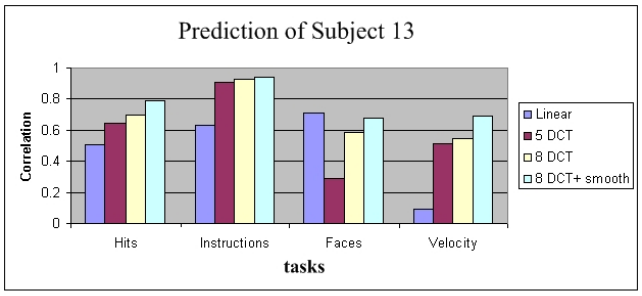


Fig. 5. Detrending effect on different tasks. The training and test data are collected from Subject13 in PBAIC2007 data set.

The other trend we observed is that different tasks require different level of detrending (e.g., the task "face" prefers linear detrending rather than a DCT detrending). This inspires us to learn the optimal number of basis functions by maximizing the marginal likelihood in RVR, or in KRR if we treat KRR as a Gaussian Processor Regression. Since the detrending can be applied to the input kernel directly, the learning procedure can be time efficient.

C. Kernel Ridge Regression versus Relevance Vector Regression

In Table II we compared KRR and RVR with four different tasks: two emotion experience, "arousal" and "valance", and two sensory experience, "hits" (hearing) and "weaponsTools" (vision). In addition, we presented the sparseness of RVR (percentage of the training fMRI scans contributed to the prediction) in Table III. The training and the test data are collected from subject14 in the PBAIC2007 competition data set.

As we observed, if the fMRI pattern for a brain activity is consistent, that is, in most of the cases, the same cognitive and sensory state will active the same fMRI pattern then the sparse representation would incur the loss of information for

TABLE II

COMPARISON RESULTS WITH KRR AND RVR

Method	Emotion Experience		Sensory Experience	
	Arousal	Valence	Hits	WeaponsTools
KRR	0.24	0.28	**0.78**	**0.55**
RVR	**0.32**	**0.38**	0.75	0.50

TABLE III

PERCENTAGE OF THE TRAINING FMRI SCANS CONTRIBUTE TO THE RVR PREDICTION

Method	Emotion Experience		Sensory Experience	
	Arousal	Valence	Hits	WeaponsTools
RVR	2.3%	1.2%	24.4%	22.8%

further prediction. For example, the fMRI pattern elicited by some sensory experience, such as hearing, is stable and easy to capture. In this case using sparse representation (RVR) would lose some useful samples to capture this pattern.

In contrast, we observed that some cognitive states, such as emotion, are very unpredictable. We suppose this is because such brain activity would cause variation for fMRI patterns in different time or environments. In this case using all training samples would result in an estimation of the average fMRI pattern and might not be good for further prediction. Alternatively, the sparse representation would estimate the commonest fMRI pattern, which might result in a potential better prediction.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a general framework to process the fMRI data and predict human brain activity. The overall performance is dominated by the joint performance of the explanatory hypothesis and corresponding feature extraction based on the prior knowledge of the task or specific features. We found that with reliable prior knowledge, the hypothesis driven prediction is better than a purely data driven prediction. Furthermore, we confirm that the pre-processing, especially the reduction of the low frequency noise, is essential for fMRI based prediction. Some limitations of our approach have also been pointed out, although our competition results showed that this procedure can significantly improve the prediction accuracy.

The potential contribution of this paper is in three aspects: first is the particularly effective computational technique we applied – kernel methods. Second is that we utilize the hypothesis driven technique in feature extraction and confirm its effect in our experiments. Finally is our unique application in low frequency noise detrending.

For future work, we would further investigate the detrending technique as it might be more powerful if we are possible to learn the detrending parameter automatically. Rather than apply DCT to remove the low frequency drift after choosing an arbitrary cut off, it is better to determine the amount of filtering by maximizing the marginal likelihood in RVR or in KRR if we treat KRR as a Gaussian Processor Regression. Furthermore, the work presented here was single subject learning. As all subjects are doing similar tasks, it is also possible to utilize all subjects' information to predict one subject's brain state. We intend to further this idea since it might be fruitful, especially for human lie detection.

ACKNOWLEDGMENT

The data collection was supported by **Experience Based Cognition Project** (Walter Schneider, PI, University of Pittsburgh).

REFERENCES

[1] A. M. Smith, B. K. Lewis, U. E. Ruttimann, F. Q. Ye, Y. Yang, J. H. Duyn and J. A. Frank, *Investigation of Low Frequency Drift in fMRI Signal*, NeuroImage, Volume 9, Number 5, May 1999, pp. 526–533.

[2] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006, pp. 293, pp. 345–356.

[3] C. Saunders, A. Gammerman and V. Vovk *Ridge Regression Learning Algorithm in Dual Variables*, In: Proceedings, 15th International Conference on Machine Learning , Madison, WI, 1998, pp. 515–521.

[4] David Hardoon and Larry M. Manevitz, *fMRI Analysis via One-class Machine Learning Techniques*, 19th International Joint Conference on Artifical Intelligence, Edinburgh, Scotland, UK, 2005.

[5] Denis Chigirev and The Princeton EBC Team, *One Size Does Not Fit All: Regressor and Subject Specific Techniques for Predicting Behavior in a Structured Environment*, PBAIC report, 2007, http://www.ebc.pitt.edu/2007/chigirev.html.

[6] E. Olivetti, D. Sona, S. Veeramachaneni, *Gaussian process regression and recurrent neural networks for fMRI image classification*, PBAIC report, 2006, http://www.ebc.pitt.edu/2007/2006.html.

[7] James P. Morris, Kevin A. Pelphrey and Gregory MCCarthy, *Regional Brain Activation Evoked When Approaching A Virtual Human on a Virtual Walk*, Journal of Cognitive Neuroscience 17:11, pp. 1744-1752.

[8] John Shawe-Taylor and Nello Cristianini, *Kernel Methods for pattern Analysis*, Cambridge University Press, 2004, pp. 80–82, pp. 232–233, pp. 290–293.

[9] K. Amunts, A. Malikovic, H. Mohlberg, T. Schormann, K. Zilles, *Brodmann's areas 17 and 18 brought into stereotaxic space – where and how variable?*, Neuroimage, Volume 11, 2000, pp. 66–84.

[10] Matthew FS Rushworth, Michael Krams and Richard E Passingham, *The Attentional Role of the Left Parietal Cortex: The Distinct Lateralization and Localization of Motor Attention in the Human Brain*, Journal of Cognitive Neuroscience 13:5, pp. 698-710.

[11] M. E. Tipping, *Sparse Baysian Learning and the Relevance Vector Machine*, Journal of Machine Learning Research (2001), 1, pp. 211–244.

[12] MP. Deiber, V. Ibanez, N. Sadato and M. Hallett, *Cerebral Structures Participating in Motor Preparation in Humans: A Positron Emission Tomography Study*, Journal of Neurophysiology, Vol 75, No. 1 Januray 1996.

[13] PBAIC, *2007 Competition Guide book*, 2007, http://www.ebc.pitt.edu/2007/docs/CompetitionGuideBook2007v7.pdf

[14] P. Morosan, J. Rademacher, A. Schleicher, T. Schormann, K. Zilles, *Human primary auditory cortex: Cytoarchitectonic subdivisions and mapping into a spatial reference system*, Neuroimage, Volume 13, 2001, pp. 684–701.

[15] Rajan Patel, *Prediction of dynamic experiences with neural networks and fMRI*, PBAIC report, 2007, http://www.ebc.pitt.edu/2007/patel.html.

[16] R. Cunnington, C. Windischberger, Deecke and E. Moser, *The preparation and Execution of Self–Initiated and Externally–Triggered Movement: AStudy of Event–Related fMRI*. NeuroImage 15, 2002, pp. 373-385.

[17] R. S. J. Frackowiak, K. J. Friston, C. D. Frith, R. J. Dolan, C. J. Price, S. Zeki, J. Ashburner and W.D. Penny, *Human Brain Fuction*, Academic Press, 2nd edition, 2003.

[18] SPM5, The Welcome Trust Centre for Neuroimaging at UCL, 2005, http://www.fil.ion.ucl.ac.uk/spm/software/spm5/.

[19] S. Watanabe, *Review of 'Knowing and Guessing, A Quantitative Study of Inference and Information*, Journal of Information Theory, Vol 16, Issue: 3, 1970, pp. 361–362.