Kernel PCA Feature Extraction of Event-Related Potentials for Human Signal Detection Performance

Roman Rosipal, Mark Girolami
Department of Computing and Information Systems, University of Paisley
Paisley, PA1 2BE, Scotland

Leonard J. Trejo
Human Information Processing Research Branch, NASA Ames Research Center
Moffett Field, CA

Abstract

In this paper, we propose the application of the Kernel PCA technique for feature selection in high-dimensional feature space where input variables are mapped by a Gaussian kernel. The extracted features are employed in the regression problem of estimating human signal detection performance from brain event-related potentials elicited by task relevant signals. We report the superiority of Kernel PCA for feature extraction over linear PCA.

1 Introduction

In many safety-critical applications (e.g., air traffic control, power plant operation, military applications) control is based on the ability of human operators to detect and evaluate task-relevant signals in the presented visual data. Performance quality of operators varies over time, often falling below acceptable limits, and may result in errors with potentially serious consequences. The likelihood of such errors could be reduced if physiological methods for assessment of human performance were available.

A fundamental part in the development of such a method is to construct a model reflecting the dependence between selected physiological metrics of mental workload (e.g., Event-Related Potentials (ERPs)) and the performance characteristics of a human operator (reaction time, accuracy, and confidence). Prior research has demonstrated that linear regression and nonlinear neural networks can model the relationships between ERPs and performance (see [1, 2, 3] and ref. therein). However, when we attempt to develop such a model we are confronted with the curse of dimensionality, which arises from the complexity of physiological data. For example, 225 data samples (dimensions) are required to describe a single 1.5s segment of ERP data from three electrodes. To address this problem, we can accept two general assumptions about the real world data sets. First, there exist some correlations among input variables; thus dimensionality reduction or so-called *feature extraction* allows us to restrict the entire input space to a sub-space of lower dimensionality. Second, in many practical problems we can assume a smooth mapping from input to output space; thus we

can infer the values of the output for points where no input data are available. This can be done by an appropriate regularization technique.

In this study, we have used the recently proposed Kernel PCA [4] method for feature selection in kernel space. This allows us to obtain features (nonlinear principal components) with higher-order correlations between input variables, and second, we can extract more components if the number of data points is higher than their dimensionality [4]. The idea behind Kernel PCA [4] is based on computation of the standard linear PCA in a high dimensional feature space $\mathcal{F}$ (with dimension $\leq \infty$), into which the input data $\mathbf{x} \in R^{\mathcal{N}}$ are mapped via some nonlinear function $\Phi(\mathbf{x})$. To this end, we compute a dot product in space $\mathcal{F}$ using a kernel function, i.e. $K(\mathbf{x},\mathbf{y}) = (\Phi(\mathbf{x}).\Phi(\mathbf{y}))$. This 'kernel trick' allows us to carry out any algorithm, e.g Support Vector Regression (SVR) [5], that can be expressed in the terms of dot products in space $\mathcal{F}$. Next, we used selected features to train SVR and Kernel Principal Component Regression to estimate the dependency between ERPs and the performance of the individual subjects. The results suggest the superiority of (nonlinear) Kernel PCA for feature extraction over linear PCA in some cases.

2 Methods

2.1 Kernel PCA, Multi-Layer SVR and Kernel PCR

The PCA problem in high-dimensional feature space $\mathcal{F}$ can be formulated as the diagonalization of a n-sample estimate of the covariance matrix $\hat{C} = \frac{1}{n}\sum_{i=1}^{n}\Phi(\mathbf{x}_i)\Phi(\mathbf{x}_i)^T$, where $\Phi(\mathbf{x}_i)$ are centered nonlinear mappings of the input variables $\mathbf{x}_i \in \mathcal{R}^{\mathcal{N}}$, $i = 1,...,n$. The diagonalization represents a transformation of the original data to new coordinates defined by orthogonal eigenvectors $\mathbf{V}$. We have to find eigenvalues $\lambda \geq 0$ and non-zero eigenvectors $\mathbf{V} \in \mathcal{F}$ satisfying the eigenvalue equation $\lambda\mathbf{V} = \hat{C}\mathbf{V}$. Realizing, that all solutions $\mathbf{V}$ with $\lambda \neq 0$ lie in the span of mappings $\Phi(\mathbf{x}_1),...,\Phi(\mathbf{x}_n)$, Schölkopf et. al. [4] derived the equivalent eigenvalue problem $n\lambda\mathbf{v} = \mathbf{K}\mathbf{v}$, where $\mathbf{v}$ denotes the column vector with coefficients $v_1,\ldots,v_n$ such that $\mathbf{V} = \sum_{i=1}^{n} v_i\Phi(\mathbf{x}_i)$ and $\mathbf{K}$ is a symmetric $n \times n$ matrix with $\mathrm{K}_{ij} = (\Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j)) := K(\mathbf{x}_i,\mathbf{x}_j)$. Normalizing the solutions $\mathbf{V}^k$ corresponding to the non-zero eigenvalues λ_k of the matrix $\mathbf{K}$, translates into the condition $\lambda_k(\mathbf{v}^k.\mathbf{v}^k) = 1$ [4]. Finally, we can compute the projection of $\Phi(\mathbf{x})$ onto the k-th nonlinear principal component by

$$q(\mathbf{x})_k := (\mathbf{V}^k.\Phi(\mathbf{x})) = \sum_{i=1}^{n} v_i^k K(\mathbf{x}_i,\mathbf{x}). \tag{1}$$

We then select the first r nonlinear principal components, e.g. the directions which describe a desired percentage of data variance, and thus work in an r-dimensional sub-space of feature space $\mathcal{F}$. This allows us to construct multi-layer support vector machines [4], where a preprocessing layer extracts features for the next regression or classification task. In our study we focus on the regression problem.

Generally, the SVR problem (see e.g.[5]) can be defined as the determination of function $f(\mathbf{x},\mathbf{w})$ which approximates an unknown desired function and has the form $f(\mathbf{x},\mathbf{w}) = (\mathbf{w}.\Phi(\mathbf{x})) + b$, where b is an unknown bias term, $\mathbf{w} \in \mathcal{F}$ is a vector of

unknown coefficients and $(\mathbf{w}.\Phi(\mathbf{x}))$ is a dot product in space $\mathcal{F}$. In [7] the following regularized risk functional is shown to compute the unknown coefficients b and $\mathbf{w}$:

$$R_{reg}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}|Err|_{\varepsilon} + \frac{\eta}{2}\|\mathbf{w}\|^2, \tag{2}$$

where $Err = y_i - f(\mathbf{x}_i, \mathbf{w})$, $\eta \geq 0$ is a regularization constant to control the trade-off between complexity and accuracy of the regression model and $|Err|_{\varepsilon}$ is Vapnik's ε-insensitive loss-function [7]. In [7] it is shown that the regression estimate that minimizes the functional (2) has the form: $f(\mathbf{x}, \mathbf{a}, \mathbf{a}^*) = \sum_{i=1}^{n}(a_i^* - a_i)K_1(\mathbf{x}_i, \mathbf{x}) + b$, where $\{a_i, a_i^*\}_{i=1}^{n}$ are Lagrange multipliers [5].

Combining the Kernel PCA preprocessing step with SVR yields a multi-layer SVR in the following form [4]: $f(\mathbf{x}, \mathbf{a}, \mathbf{a}^*) = \sum_{i=1}^{n}(a_i - a_i^*)K_1(\mathbf{q}(\mathbf{x}_i), \mathbf{q}(\mathbf{x})) + b$, where components of vectors $\mathbf{q}(.)$ are defined by (1). However, in practice the choice of appropriate kernel function $K_1(.,.)$ can be difficult. In this study, a polynomial kernel of first order $K_1(\mathbf{x}, \mathbf{y}) = (\mathbf{x}.\mathbf{y})$ is employed. We are thus performing a linear SVR on the r-dimensional sub-space of $\mathcal{F}$. The advantage of linear SVR over ordinary linear regression is the possibility of using a large variety of loss functions to suit different noise models [5], e.g. the proposed Vapnik's ε-insensitive function is more robust for noise distributions close to uniform. However, in the case of Gaussian noise the best approximation to the regression provides a loss function of the form $L(y_i, f(\mathbf{x}_i, \mathbf{c})) = [y_i - f(\mathbf{x}_i, \mathbf{c})]^2$. Therefore, we used a Kernel Principal Component Regression[1] technique which minimizes the following risk functional

$$R_{pcr}(\mathbf{c}) = \frac{1}{n}\sum_{i=1}^{n}[y_i - f(\mathbf{x}_i, \mathbf{c})]^2.$$

The solution $f(\mathbf{x}, \mathbf{c})$ has the form

$$f(\mathbf{x}, \mathbf{c}) = \sum_{k=1}^{r} b_k q(\mathbf{x})_k + b_0 = \sum_{k=1}^{r} b_k \sum_{i=1}^{n} v_i^k K(\mathbf{x}_i, \mathbf{x}) + b_0 = \sum_{i=1}^{n} c_i K(\mathbf{x}_i, \mathbf{x}) + b_0,$$

where $\{c_i = \sum_{k=1}^{r} b_k v_i^k\}_{i=1}^{n}$ and $\{q(\mathbf{x})_k\}_{k=1}^{r}$ are again defined by (1). The coefficients $\{b_k\}_{k=0}^{r}$ can be found by solving the *normal equations* for least squares estimation.

2.2 Data Sample Construction

We have used ERPs and performance data from an earlier study [2]. Eight male Navy technicians experienced in the operation of display systems performed a signal detection task. Each technician was trained to a stable level of performance and tested in multiple blocks of 50–72 trials each on two separate days. Blocks were separated by 1-minute rest intervals. A set of 1000 trials were performed by each subject. Inter-trial intervals were of random duration with a mean of 3s and a range of 2.5–3.5s. The entire experiment was computer-controlled and performed with a 19-inch color CRT display. Triangular symbols subtending 42 minutes of arc and of three different luminance contrasts (0.17, 0.43, or 0.53) were presented parafoveally at a constant

[1] A more detailed description of a Principal Component Regression is given in [6].

eccentricity of 2 degrees visual angle. One symbol was designated as the target, the other as the non-target. On some blocks, targets contained a central dot whereas the non-targets did not. However, the association of symbols to targets was alternated between blocks to prevent the development of automatic processing. A single symbol was presented per trial, at a randomly selected position on a 2-degree annulus. Fixation was monitored with an infrared eye tracking device. Subjects were required to classify the symbols as targets or non-targets using button presses and then to indicate their subjective confidence on a 3-point scale using a 3-button mouse. Performance was measured as a linear composite of speed, accuracy, and confidence. A single measure, PF1, was derived using factor analysis of the performance data for all subjects, and validated within subjects. The computational formula for PF1 was

$$\text{PF1} = 0.33*\text{Accuracy} + 0.53*\text{Confidence} - 0.51*\text{Reaction Time}$$

using standard scores for accuracy, confidence, and reaction time based on the mean and variance of their distributions across all subjects. PF1 varied continuously, being high for fast, accurate, and confident responses and low for slow, inaccurate, and unconfident responses.

ERPs were recorded from midline frontal, central, and parietal electrodes (Fz, Cz, and Pz), referred to average mastoids, filtered digitally to a bandpass of 0.1 to 25 Hz, and decimated to a final sampling rate of 50 Hz. The prestimulus baseline (200 ms) was adjusted to zero to remove any DC offset. Vertical and horizontal electrooculograms (EOG) were also recorded. Epochs containing artifacts were rejected and EOG-contaminated epochs were corrected. Furthermore, any trial in which no detection response or confidence rating was made by a subject was excluded along with the corresponding ERP.

Within each block of trials, a running-mean ERP was computed for each trial. Each running-mean ERP was the average of the ERPs over a window that included the current trial plus the 9 preceding trials for a maximum of 10 trials per average. Within this 10-trial window, a minimum of 7 artifact-free ERPs were required to compute the running-mean ERP. If fewer than 7 were available, the running mean for that trial was excluded. Thus each running mean was based on at least 7 but no more than 10 artifact-free ERPs. This 10-trial window corresponds to about 30s of task time. The PF1 scores for each trial were also averaged using the same running-mean window applied to the ERPs, excluding PF1 scores for trials in which ERPs were rejected. Prior to analysis, the running-mean ERPs were clipped to extend from time zero (stimulus onset time) to 1500 ms post-stimulus, for a total of 75 time points.

3 Results

The present work was carried out with Gaussian kernels; $K(\mathbf{x},\mathbf{y}) = e^{-(\frac{\|\mathbf{x}-\mathbf{y}\|^2}{L})}$, where L is the width of the Gaussian function. The desired output PF1 was linearly normalized to have a range of 0 to 1. We trained the models on 50% of the ERPs and tested on the remaining data. The described results, for each setting of the parameters, are an average of 10 runs each on a different partition of training and testing data. The validity of the models was measured in terms of the proportion of data for which

PF1 was correctly predicted with 10% tolerance, i.e ± 0.1 in our case. The performance of a Regularized Gaussian RBF (rGRBF) network [8] and SVR trained on data pre-processed by linear PCA (LPCA) in input space was compared with the results achieved by multi-layer SVR (MLSVR) and the proposed Kernel Principal Component Regression (KPCR) on features extracted by Kernel PCA. In both cases we used features (principal components) describing 99% of data variance. We used $\varepsilon = 0.01$, $\eta = 0.01$ parameter values in the case of SVR. The results achieved on subject A (592 ERPs), B(614 ERPs) and C (417 ERPs) are depicted in Figure 1. On subjects A and B we can see consistently better results on features extracted by Kernel PCA (top and middle left graphs). These superior results achieved using Kernel PCA representation were also observed on the remaining five subjects. In addition, we can see that in all cases the SVR was superior to rGRBF on inputs extracted by linear PCA (right graphs). We have to note that on all subjects we achieved similar results with features describing 98% of data variance. Without the PCA preprocessing step in feature space $\mathcal{F}$ we did not increase the overall performance. On the contrary, on four subjects the performance was on average decreased by 0.5% on test proportion error.

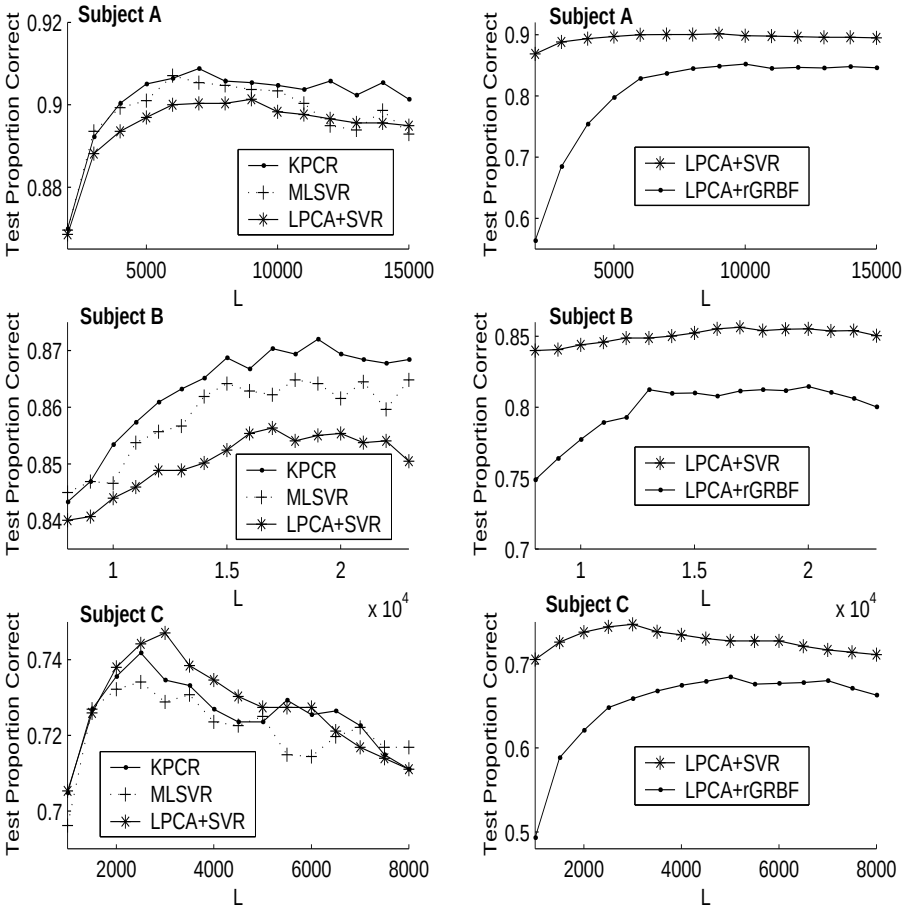

Figure 1: Comparison of the results achieved on subjects A, B and C.

4 Conclusions

The selection of appropriate features for regression has been investigated. On subjects A and B we demonstrated that (nonlinear) Kernel PCA provides a superior representation of the data set over that of linear PCA. However, on subject C the performance with features selected by linear PCA was slightly better. We have to note, that in this case the dimension of matrix $\mathbf{K}$ in feature space $\mathcal{F}$ is lower (209) than the input dimensionality (225), thus we can not exploit the advantage of Kernel PCA to improve overall performance by using more components in feature space than the number available in the input space. We used features describing 99% of the data variance that for different parameter L represents 70–90% of all nonlinear principal components and we showed that such a reduction of high-dimensional feature space $\mathcal{F}$ does not decrease the overall performance. Moreover, this can be seen as the denoising technique assuming that the noise is spread in directions with small variance. On all subjects, we demonstrated that the performance of SVR on features extracted by linear PCA was superior to Regularized Gaussian RBF.

Acknowledgments

The first author is funded by a research grant for the project "Objective Measures of Depth of Anaesthesia"; University of Paisley and Glasgow Western Infirmary NHS trust, and is partially supported by Slovak Grant Agency for Science (grants No. 2/5088/98 and No. 98/5305/468). Data were obtained under a grant from the US Navy Office of Naval Research (PE60115N), monitored by Joel Davis and Harold Hawkins. Dr. Trejo is supported by the Psychological and Physiological Stressors and Factors Project of the NASA Aerospace Operations Systems Program (RTOP 711-51-42), managed by J. V. Lebacqz.

References

[1] Trejo LJ, Shensa MJ. Feature Extraction of ERPs Using Waveletes: An Application to Human Performance Monitoring. Brain and Language 1999; 66:89–107.

[2] Trejo LJ, Kramer AF, Arnold JA. Event-related Potentials as Indices of Display-monitoring Performance. Biological Psychology 1995; 40:33–71.

[3] Koska M, Rosipal R, König A, Trejo LJ. Estimation of human signal detection performance from ERPs using feed-forward network model. In: Computer Intensive Methods in Control and Signal Processing, The Curse of Dimensionality. Birkhauser, Boston, 1997.

[4] Schölkopf B, Smola AJ, Müller KR. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Computation 1998; 10:1299–1319.

[5] Smola AJ, Schölkopf B. A Tutorial on Support Vector Regression. Technical Report NC2-TR-1998-030, NeuroColt2 Technical Report Series, 1998.

[6] Jollife IT. Principal Component Analysis. Springer-Verlag, New York, 1986.

[7] Vapnik V. The Nature of Statistical Learning Theory. Springer, New York, 1998.

[8] Chen S, Chng ES, Alkadhimi K. Regularised orthogonal least squares algorithm for constructing RBF networks. Int Journal of Control 1996; 64.