Available online at www.sciencedirect.com

ScienceDirect

NEUROCOMPUTING

ELSEVIER

Neurocomputing 70 (2006) 373–383

www.elsevier.com/locate/neucom

Interpretation of perceptron weights as constructed time series for EEG classification

Dik Kin Wong*, E. Timothy Uy, Marcos Perreau Guimaraes, Wayne Yang, Patrick Suppes

Center for Study of Language and Information, Stanford University, Stanford, CA, USA

Received 15 April 2005; received in revised form 4 January 2006; accepted 5 January 2006
Communicated by K.J. Cios
Available online 16 June 2006

Abstract

The interpretation of weights in a neural network is seldom straightforward. Recently, we have shown perceptron-based learning to yield better brain-wave classification rates than learning based on averaging and optimal filtering. By virtue of our implementation, we are able to interpret the weights as a time series and to relate them to prototypes generated by averaging. In this paper, some results of four closely related linear models are shown. They are based on averaging, averaging with filtering, Tikhonov regularization, and a single-layer neural network. We then introduce this interpretation for a Tikhonov-regularized linear model and a single-layer neural network with a linear-transfer function. We show, using Tikhonov regularization as an example, how such an interpretation can be used to gain insight into the mechanisms of various perceptron-based methods.
© 2006 Elsevier B.V. All rights reserved.

Keywords: EEG; Individual trials; Tikhonov regularization; Neural network

1. Introduction

The event-related potential (ERP) is a widely used representation of brain activity associated with events such as visual or auditory stimuli. An ERP is the average, over a substantial number of electroencephalography (EEG) trials, of ongoing brain activity synchronized to the onset of an event. Extensive reviews of ERP can be found in [7,4]. These ERPs formed a set of time-series waveforms that we considered as prototypes, one for each stimulus-associated class (since each stimulus is associated with a class of trials).

For methods beyond ERP, prototypes refer to the sequence of weights computed for each class after some parametric learning. As with the ERP, the filtered average (FA) method [9–11] utilizes averages of EEG time-series data from the same class to form a set of prototypes. After optimizing a bandpass filter (characterized by the low and high frequency L and H), the individual trials are classified by a least-squares criterion. An optional temporal window, characterized by its start and end, was used previously [9–11].

In addition to these two least-squares classifiers (ERP and FA), two perceptron-based models, the Tikhonov-regularized linear model (TLM) and the single-layer neural network (NN), were considered. In our implementation of these two perceptron-based models, we replaced the sum-of-squares distances of each trial to the prototypes with an inner product between the trial and the weight vectors. This inner product acts as an inverse distance measure, so that the trial is assigned to the class whose weights produce the greatest inner product. In other words, while the least-squares classifiers selected a class for test trial $\mathbf{t}$ among the prototypes $i \in \{1 \ldots I\}$ by $i = \arg\min_i \|\mathbf{p}_i - \mathbf{t}\|^2$, the TLM and NN classifiers selected a class by $i = \arg\max_i(\mathbf{p}_i \cdot \mathbf{t})$, essentially minimizing the angle. Here, the inputs $\mathbf{t}$ remain the time-ordered sequence of observations. Thus, we can physically interpret the resulting sequences of weights for a given class as time-series data, each sequence capturing, in some sense, the essence of the class. With this interpretation, we can quantitatively compare "prototypes" generated by our perceptron-based methods with

[*]Corresponding author. Tel.: +1 65067046353.
E-mail address: dkwong@stanford.edu (D.K. Wong).

0925-2312/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2006.01.020

ERPs. A similar interpretation would become more difficult in the case of multi-layer neural networks [3,15].

In the following section, we describe the underlying data and experiment used to demonstrate our interpretation. After presenting a number of tools, we will compare prototypes from the four classification methods for individual trials mentioned above (ERP, FA, TLM and NN). Finally, we will show how this way of looking at weights can be used to interpret the underlying mechanism in both single channel and multichannel implementations of Tikhonov regularization.

2. A comparison of classification schemes

2.1. Data

EEG recordings were made in our laboratory using 22 Model-12 Grass amplifiers and Neuroscan's Scan 4 software. Sensors were attached to the scalp of a subject according to the standard 10-20 EEG system as bipolar pairs with the recorded measurement in microvolts being the potential difference between each pair of sensors. The placements of the 11 pairs were C3–T5, Cz–T5, Cz–Pz, Cz–P4, C4–T6, P4–T6, C3–T3, C3–Cz, C4–Cz, C4–T4 and C4–F8. These are hereafter referred to as channels. The recording bandwidth was from 0.3 to 100 Hz with a sampling rate of 1 kHz.

A computer was used to present the stimuli under three conditions: visual image, visual word and auditory word, with the words being in Chinese for monolingual subjects. For each condition, there were eight stimuli consisting of four colors: blue, green, red, and yellow and four shapes: circle, square, triangle and line (at 45° angle measured counterclockwise).

Every experimental trial contained a pair of stimuli, selected from those just described, and presented in temporal sequence. Each stimulus lasted for 200 ms in the non-auditory cases and ranged from 394 to 617 ms for auditory stimuli. The second stimulus of the pair was shown 1500 ms after the onset of the first stimulus. After presentation of the second member of a pair, the subject used the numeric pad on the computer to respond "1" if the two stimuli in the pair were the same and "2" if they were different. (Subjects were instructed that the same–different distinction was obvious and did not require a subtle perceptual discrimination.) A new pair of stimuli were then shown 2500 ms after the previous stimulus' onset. Therefore, the length of each trial of the pair was 4 s in total, with an interstimulus interval of 1500 ms between the onset of the first and the onset of the second stimulus of the pair, and 2500 ms between the onset of the second stimulus of the first trial and the onset of the first stimulus of the next trial. The stimulus pairs chosen from the eight possible stimuli available in each condition were always matched either color-to-color or shape-to-shape. Four subjects participated in this experiment. The duration of the visual stimuli was 200 ms for the two visual conditions. The duration of the auditory stimuli ranged from 394 to 617 ms.

There were three sessions. In the first session, visual images of colors and shapes were shown. There were four categories with four pairs of stimuli each. Stimuli pairs could be categorized as (i) same color (blue–blue, green–green, red–red, yellow–yellow), (ii) different colors (blue–yellow, green–blue, red–green, yellow–red), (iii) same shape (circle–circle, line–line, square–square, triangle–triangle), or (iv) different shapes (circle–line, line–square, square–triangle, triangle–circle). In the second session, the pairs were the same as those in session one, except that the colors were presented as spoken words while the shapes were presented as visual words. For the last session, the pairs were again the same, except that the colors were presented as visual words while the shapes as spoken words. Subjects were instructed that the same–different distinction was obvious and did not require a subtle perceptual discrimination. The data from these linguistic parts of the experiment were not used in the analyses reported in this paper.

Prior to downsampling, a lowpass filter was applied at 31.25 Hz to prevent aliasing. Data were then downsampled by 16 to approximately 62.5 Hz, resulting in 88 data points per channel. For each subject, each of the 8 classes in each of the three conditions was presented 100 times, resulting in a total of 800 trials. We do not attempt to present all the detailed results of this experiment here. Instead, we chose one of the best results, that of a subject under the visual-image condition, to illustrate how we might interpret the representations derived from various classification methods.

In most cases, data were randomly partitioned into 560 training trials and 240 test trials. Within each trial, observations were scaled by $f(x)$ to have a range between -1 to $+1$, where

$$f(x) = \frac{2(x - \min(x))}{\max(x) - \min(x)} - 1.$$

In Fig. 1, scaling of the data was only applied to TLM and NN, and not to ERP and FA. While such scaling is recommended for TLM and NN, the ERP prototypes resulting from averaging are very similar with or without the scaling (the mean correlation for the eight prototypes was 0.997 with a standard deviation of 0.0012).

2.2. Analysis

2.2.1. Average model (ERP)

To generate the ERP prototypes we averaged together the training trials by class. In the visual-image conditions of the experiment with 560 training trials, we built 8 prototypes by averaging 70 trials from each class. For the ERP model, the least-squares criterion is used without any parameter estimation. Although the question of which method yields the best ERP is somewhat controversial, our averaged prototypes can undoubtedly be

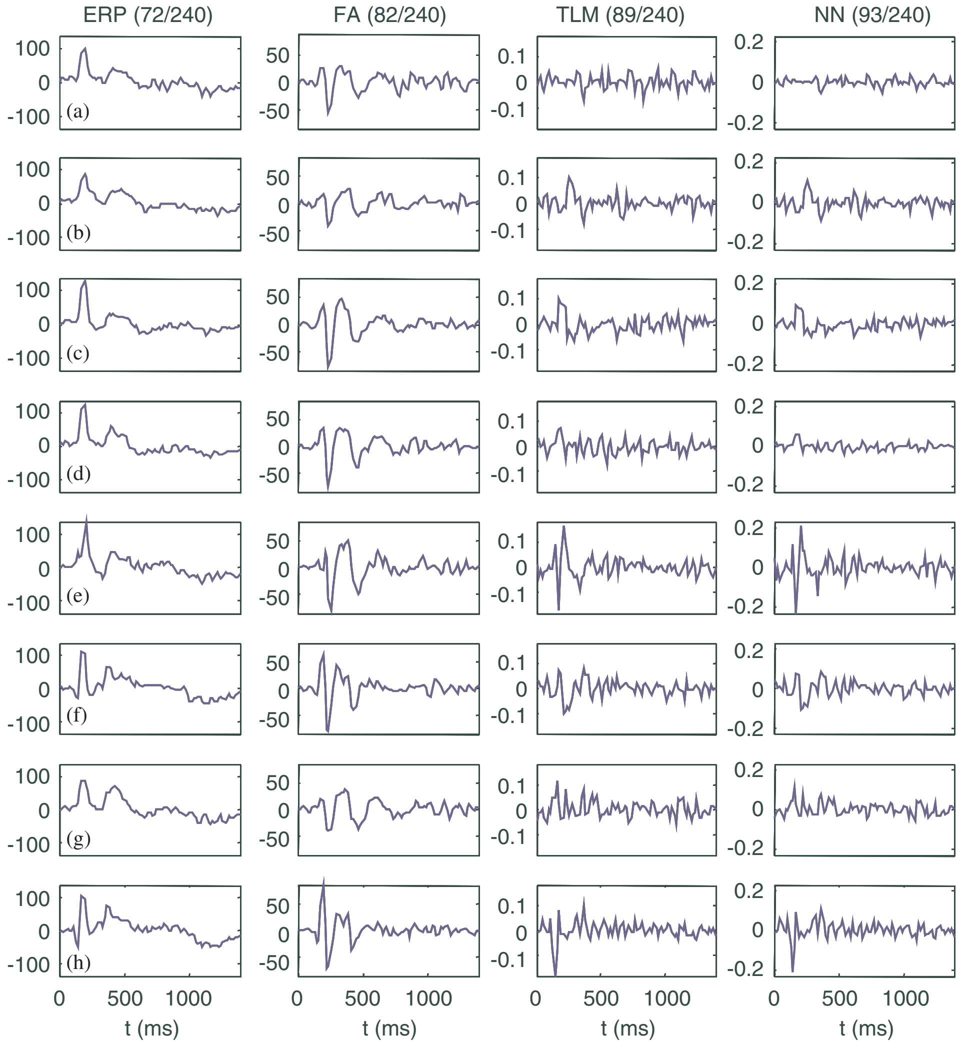


Fig. 1. Comparison of different prototypes and weights. ERP: average, FA: filtered average, TLM: Tikhonov-regularized linear model and NN: single-layer neural network. The x-axis is the time in millisecond and the y-axis is the normalized amplitude for ERP and FA and normalized weights for TLM and NN. Stimulus labels in the figures were denoted as (a) for circle, (b) for line, (c) for triangle, (d) for square, (e) for red, (f) for green, (g) for blue and (h) for yellow. The bandpass filter used for the FA models has a low cutoff at 3 Hz and high cutoff at 21 Hz.

regarded as ERPs [7,4]. By relating the other kinds of prototypes (FA, TLM, NN) to ERPs which have a natural physical interpretation, we further develop our intuition for interpreting such prototypes as constructed time series.

2.2.2. Filtered average (FA) model

To generate the FA prototypes, we first split the 560 training trials into 320 trials for prototype and 240 for the validation set. Prototype trials were then averaged together by class and used to classify the validation set. Both the averaged prototypes and single-trial test samples were bandpass filtered prior to classification. We repeated the classification over a number of different bandpass filters, each characterized by (L, H), L being the low cutoff frequency and H the high. The filter yielding the maximum classification rate, i.e., the optimal filter, was then applied to the ERP prototypes (averages over all the 560 trials in the training set) to produce the FA prototypes.

2.2.3. Tikhonov-regularized linear model (TLM)

Weights corresponding to each class were generated using a TLM [12]:

$$\mathbf{w} = [(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda^2\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}]\hat{\mathbf{y}}, \tag{1}$$

where $\mathbf{X}$ is the data matrix in which each row is a trial of the time-series data. Vector $\hat{\mathbf{y}}$ is the corresponding target vector which denotes the desired outputs of the trials. The dimension of $\mathbf{X}$ is $M \times (N+1)$, where M is the number of trials included for parameter evaluations and N is the length of the time-series data (a bias term is added). And $\hat{\mathbf{y}}$ is a $M \times 1$ vector.

In this paper, the Tikhonov regularization parameter λ^2 was chosen to be 40 based on our previous experience [13].

2.2.4. Single-layer neural network (NN)

A single-layer NN is defined by the relation $\mathbf{y} = f(\mathbf{Wx})$. The matrix $\mathbf{W}$ contains the weights, $\mathbf{x}$ is the input of the network and $\mathbf{y}$ the output. The weights can be optimized by backpropagation where $\mathbf{W}_{\mathrm{opt}}$ is obtained iteratively by a gradient method:

$$w_{ij}(l+1) = w_{ij}(l) - \eta \frac{\partial J(\mathbf{W}(l))}{\partial w_{ij}},$$

where l is the iteration, η is the learning rate and J is the error function. In a single-layer NN with only one output, the weight matrix $\mathbf{W}$ reduces to a weight vector $\mathbf{w}$ and the output vector $\mathbf{y}$ reduces to a scalar y.

For linear f, weights can simply be found by computing the pseudoinverse of the matrix $\mathbf{X}$, defined earlier with a size $M \times (N+1)$. In the case of a non-linear f which is differentiable, e.g., a sigmoid function, simple gradient descent can be used. When there is a difference between the output and target values, each component of a current $\mathbf{w}$ is recomputed in a direction that reduces the difference [1].

Forsee et al. [2] implemented a Gauss–Newton iterative method to optimize the regularization parameter with a kind Bayesian regularization of which the weights are optimized to maximize the evidence of the data [5]. In a single-layer NN, the objective function is given by

$$\alpha\|\mathbf{w}\|^2 + \beta\|f(\mathbf{w}^{\mathrm{T}}\mathbf{x}) - \hat{y}\|^2, \tag{2}$$

where α, β are the regularization parameters and f the transfer function. In our case, we chose a typical transfer function, the log sigmoid. Given the optimal weight $\mathbf{w}_{\mathrm{opt}}$, corresponding to α_{opt} and β_{opt} which were optimized by a Gauss–Newton iterative method [2], the classification of a test trial is performed by $y = f(\mathbf{w}_{\mathrm{opt}}^{\mathrm{T}}\mathbf{x})$.

For our classification purposes, a single-layer NN with multiple outputs is constructed, having one output per class. A classification of a test trial is then made by selecting the class with the maximum output.

2.3. Methods of comparison

By relating the representations of the other models to ERP, more natural interpretations of these typically somewhat unintuitive models may become possible. Indeed, ERP, FA TLM and NN are very much related. FA is related to ERP by filtering (the most common signal processing technique), which is here used to improve EEG classification. Meanwhile, Tikhonov regularization in TLM provides a way to solve the single-layer NN model with linear-transfer functions [14], where the regularization parameter λ in (1) plays a similar role to that of α and β in (2). By recognizing the time-series nature of the weights, simple correlation methods can be used to compare the prototypes.

2.3.1. Visual comparison of time series

We show, for the best performing channel, prototypes generated by each of our methods in Fig. 1 along with their associated classification rates. While the ERP and FA prototypes (columns 1 and 2) have some common features, and the TLM and NN prototypes (columns 3 and 4) are very much alike, there are few discernible similarities between the perceptron-based prototypes (TLM, NN) and the averaged prototypes (ERP, FA). It is this dissimilarity which often evokes the comment that weights from perceptron-based techniques are uninterpretable. Visually this is clearly the case, as evidenced by comparing columns 1 and 2 with columns 3 and 4 in Fig. 1.

2.3.2. Correlation maps

To quantify our visual comparisons of the ERP, FA, TLM and NN prototypes, we used the correlation coefficient ρ [6] to compare pairs of prototypes. We computed the correlation of all ERP, FA, TLM and NN prototypes with the ERP prototypes. An example of such correlation maps is shown in Fig. 2.

2.3.3. Measures of similarity based on correlation maps

In addition to computing the average correlation $\bar{\rho}$ of the different types of prototypes, we attempted to quantify further our notion of similarity. Several methods were explored. The first used each correlation map as a distance matrix for classification. We computed a classification rate based on maximum correlation, that is,

$$\kappa = \frac{1}{2N}\left[\sum_m \left(\rho_{m,m} > \max_n \rho_{m,n\neq m}\right) + \sum_n \left(\rho_{n,n} > \max_m \rho_{m\neq n,n}\right)\right],$$

where the inequalities are 1 when true, and 0 otherwise. In this equation, N is the total number of classes. The first term of the measure κ gives the average classification rate

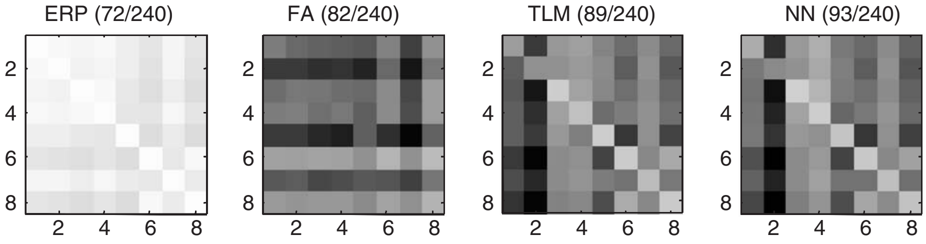


Fig. 2. Correlation maps of different sets of eight prototypes for ERP, FA, TLM, NN compared to ERP. White corresponds to a correlation coefficient of +1, while black corresponds to a correlation coefficient of −1. These results are for the visual-image experiment with eight stimuli. The number on the x and y-axes are the stimulus labels. Stimulus labels in the figures were denoted as (a) for circle, (b) for line, (c) for triangle, (d) for square, (e) for red, (f) for green, (g) for blue and (h) for yellow. The bandpass filter used for the FA models has a low cutoff at 3 Hz and high cutoff at 21 Hz.

between the prototypes and the test trials, defined by the rows and columns, respectively. The rate is computed based on maximization of the correlation of each trial against all the possible prototypes. The symmetrical condition is taken into account by the second term, of which the columns are regarded as prototypes and the rows as test trials.

We also devised a score based on the difference between the mean of the diagonal elements and the mean of the absolute value of the off-diagonal elements:

$$\gamma = \frac{1}{N}\sum_{m} \rho_{m,m} - \frac{1}{N^2 - N}\sum_{m}\sum_{n\neq m} |\rho_{m,n}|.$$

Finally, we investigated a variation on the score γ, by considering only the off-diagonal element with the greatest absolute correlation,

$$\gamma_w = \frac{1}{N}\sum_{m}\left(\rho_{m,m} - \max_{n}(|\rho_{m,n\neq m}|)\right).$$

As pointed out by a reviewer, simply shifting the data can result in zero correlation. Indeed, this is a restriction of using simple correlation to achieve the notion of similarity. Other ways of considering similarity by finding invariant features (e.g., perceptual features), rather than correlated invariance, would make a good project but are not dealt with in this article.

2.3.4. Results

In Fig. 2, the correlation maps corresponding to the prototypes in Fig. 1 with respect to the ERP are shown. Here, white corresponds to a correlation coefficient of $+1$ while black corresponds to a correlation coefficient of -1. From the presence of strong diagonals (of whiteness relative to the off-diagonals) for TLM and NN, it should be clear that perceptron-based prototypes can be interpreted as being ERP-like.

The resulting measures for Fig. 2 are presented in Table 1. However, while the quantitative measures were interesting, we found visualizing the correlation map to be sufficiently (if not more) informative. Note the high scores for TLM vs. ERP, NN vs. ERP but not FA vs. ERP.

Several additional points are worth mentioning. In a previous experiment [14], we demonstrated the use of perceptron-based methods to classify individual trials. Since training for such methods is based on individual trials, the fact that the weights can be related to averaged prototypes is significant, suggesting that features important to classification are to be found in the prototypes based on averaged trials. But perceptron-based methods (TLM and NN) are able to classify individual trials with higher rates, 93 out of 240 vs. 72 out of 240 for the averaged-trial prototype (ERP). With the chance level of 30 out of 240, both of these rates are highly significant, with $p < 10^{-12}$ and $p < 10^{-22}$.

Table 1
Different measures of correlation maps in Fig. 2

	ERP vs. ERP	FA vs. ERP	TLM vs. ERP	NN vs. ERP
$\bar{\rho}$	1.000	0.0700	0.3946	0.4013
κ	1.000	0.125	0.875	0.875
γ	0.097	0.025	0.183	0.181
γ_w	0.033	−0.161	0.052	0.023

Compared to perceptron-based methods, the diagonal for FA prototypes (Fig. 2) is barely perceptible. This is not unexpected and is related to how the filtering method is applied. In FA, test trials, in addition to the averaged prototypes, are filtered using the optimal filter prior to classification. The other prototypes are optimized to classify trials which are unfiltered except for the anti-aliasing filter used for downsampling.

For Fig. 3, we computed the correlation between the prototypes for each method, i.e., ERP vs. ERP, FA vs. FA, TLM vs. TLM, and FA vs. FA. Naturally, the correlation coefficients in the diagonal entries were one. As can be seen in Fig. 3, the correlation between different classes was much greater for the ERP and FA than for the TLM and NN. From this we concluded that the perceptron-based prototypes are more distinctive, and usually result in better classification rates. We report in Table 2 the scores corresponding to Table 1.

The four models we explore here can be used for both single-channel and multichannel data (referred to as single-channel classifier (SCC) and multichannel classifier (MCC) methods, respectively). In the following section, we show how the correlation map can be used to investigate important aspects of perceptron-based methods. Two examples are given, the first addressing regularization for a SCC, and the second addressing the extension from SCCs to a MCC. For these examples, we chose to concentrate on TLM by virtue of its straightforward regularization and computational efficiency.

3. An interpretation of Tikhonov-regularization

3.1. Theory

Weights of each class can be computed using (1) by assigning the appropriate $\hat{y}$ (each component $\hat{y}$ of vector $\hat{\mathbf{y}}$). To compute the weight for class i, all the training trials with class label i would have $\hat{y} = 1$, otherwise $\hat{y} = 0$. In other words, the negative trial would have a target output of 0 and the positive trial a target output of 1. We denote this assignment rule as $\hat{y} = (0, 1)$, and we will show in the later section that the actual rule of assigning $\hat{y}$ does not affect the final classification.

3.1.1. A mixture of least squares and averages

From (1), it is clear $\mathbf{w} \to \mathbf{X}^{\mathrm{T}}\hat{\mathbf{y}}$ as $\lambda \to \infty$. For $\hat{y} = (0, 1)$, $\mathbf{w}$ is simply an average of the positive trials for a certain class. In fact, for finite data, this asymptotic condition is even stronger, with each $\mathbf{w}$ equal to the average prototype for a large enough λ. Similarly, as $\lambda \to 0$, the

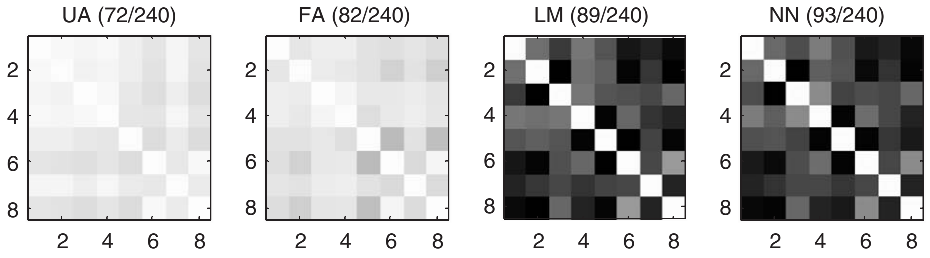


Fig. 3. Self-correlation map of different prototypes. White corresponds to a correlation coefficient of +1, while black corresponds to a correlation coefficient of −1. These results are for the best subject of the visual-image experiment. The x and y-axes are the labels of the stimuli. Stimulus labels in the figures were denoted as (a) for circle, (b) for line, (c) for triangle, (d) for square, (e) for red, (f) for green, (g) for blue and (h) for yellow. The bandpass filter used for the FA models has a low cutoff at 3 Hz and high cutoff at 21 Hz.

Table 2
Different measures of correlation maps in Fig. 3

	ERP vs. ERP	FA vs. FA	TLM vs. TLM	NN vs. NN
$\bar{\rho}$	1.000	1.000	1.000	1.000
κ	1.000	1.000	1.000	1.000
γ	0.097	0.0141	0.563	0.562
γ_w	0.033	0.054	0.405	0.442

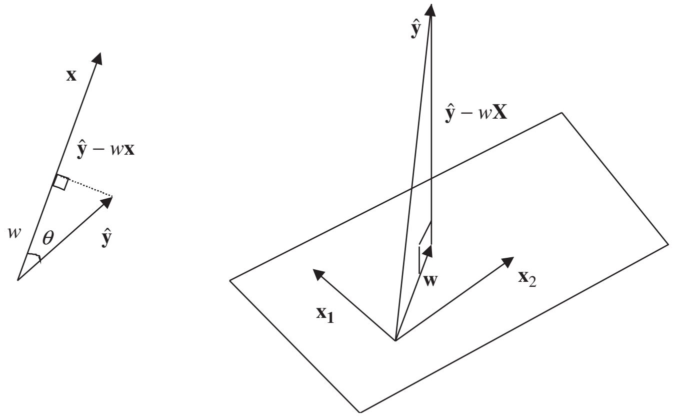


Fig. 4. The diagram on the left-hand side is the specific case for the example where $\mathbf{x} = [x_1 x_2]^{\mathrm{T}} = [1\ 3]^{\mathrm{T}}$ and $\mathbf{y} = [\hat{y}_1 \hat{y}_2]^{\mathrm{T}} = [1\ 1]^{\mathrm{T}}$. The diagram on the right-hand side is the general case.

solution $\mathbf{w} \to (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\hat{\mathbf{y}}$ based on the pseudoinverse $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$, requires a small enough λ, but not necessarily $\lambda = 0$. With assignment rule $\hat{y} = (1, 0)$, $\mathbf{w}$ is an average of the negative trials as $\lambda \to \infty$.

In addition to the geometrical interpretation behind the averaged trials, $\mathbf{w}$ also has a simple geometrical interpretation based on least squares when $\lambda = 0$, as described by Strang [8]. This geometrical description is illustrated in Fig. 4. In order to make the optimal use of $\mathbf{w}$ when $\lambda = 0$, one could classify based on the least-square distance to the target value 1, i.e., arg $\min_i \|\mathbf{w}_i\mathbf{x} - 1\|^2$ instead of simple maximization arg $\max_i \|\mathbf{w}_i\mathbf{x}\|$. But in practice, given the scaled input data with x between −1 and +1, both the classification rate and the characteristic of regularization of the two maximizations are similar.

3.1.2. A geometrical illustration of least squares

Consider the one-dimensional case and assume we observed only two trials: $(x_1, \hat{y}_1) = (1, 1)$ and $(x_2, \hat{y}_2) = (3, 1)$. A diagram illustrating the geometry, a specific case (left) and a general case (right) are shown in Fig. 4.

We would like to find the $\mathbf{w}$ which reduces the error $\|\hat{\mathbf{y}} - \mathbf{X}\mathbf{w}\|^2$. Here, $\mathbf{X}$ reduces to a vector and $\mathbf{w}$ a scalar, so we seek w such that $\|\hat{\mathbf{y}} - w\mathbf{x}\|^2$ is minimized with $\mathbf{x} = [x_1 x_2]^{\mathrm{T}} = [1\ 3]^{\mathrm{T}}$ and $\hat{\mathbf{y}} = [\hat{y}_1 \hat{y}_2]^{\mathrm{T}} = [1\ 1]^{\mathrm{T}}$. There are three ways to obtain the solution w. First, by solving the unconstrained optimization problem directly:

$$w = \arg\min_{w} f(w) \text{ where } f(w) = (1 - w)^2 + (1 - 3w)^2$$

$$\text{Set } \frac{\partial f}{\partial w} = 0,$$

$$w = 0.4.$$

Alternatively, by geometry as shown in Fig. 4:

$$\theta = \tan^{-1}(3) - \tan^{-1}(1),$$

$$w = \frac{\sqrt{2} \cdot \cos(\theta)}{\sqrt{(1^2 + 3^2)}} = 0.4.$$

Lastly, as pointed out earlier, w is the pseudoinverse solution corresponding to the Tikhonov-regularized solution without regularization ($\lambda = 0$). For the one-dimensional case:

$$\begin{aligned} w &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \\ &= \left(\sum x_i^2\right)^{-1}\left(\sum x_i \hat{y}_i\right) \\ &= \frac{\sum_{i \in y_i = 1} x_i}{\sum x_i^2} \text{ where } \hat{y}_i = (0, 1) \\ &= 0.4. \end{aligned}$$

These generalize for higher dimensions with less restricted $\hat{y}$ [8].

3.1.3. Target vector

Since the classification is based on the inner product, the classification results are the same for a family of target vectors in which α and β are constant, and α is positive. The equation for the weight $\mathbf{w}$ obtained by the

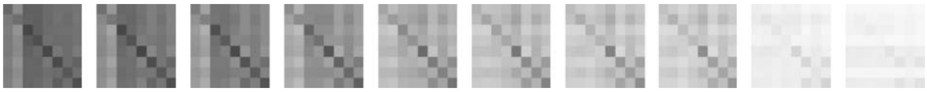


Fig. 5. Correlation maps to show the effect of regularization with $\hat{y} = (1,0)$. White corresponds to a correlation coefficient of $+1$, while black corresponds to a correlation coefficient of -1. The regularization parameter λ increases from left to right. These results are for the best subject of the visual-image experiment.

linear model is

$$\mathbf{w} = \left[(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda^2\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\right](\alpha\hat{\mathbf{y}} + \beta), \tag{3}$$

where $\mathbf{X}$ is a $M \times (N+1)$ matrix, $\mathbf{I}$, is an $(N+1) \times (N+1)$ identity matrix, while $\hat{\mathbf{y}}$ is a $M \times 1$ vector. This computation of $\mathbf{w}$ for class i addresses the problems of assignment for $\hat{y}$. For example, assigning $\hat{y}^{(1)} = (0,1)$ (0 for negative and 1 for positive trials) yields weights with the same classifications as $\hat{y}^{(2)} = (-1,1)$ due to the affine relation $\hat{y}^{(2)} = 2\hat{y}^{(1)} - 1$.

For negative α, the same classifications can be achieved by changing the maximization into a minimization. For example, the minimization of assignment $\hat{y}^{(3)} = (1,0)$ is the same as the maximization of $\hat{y}^{(1)}$. A diagram of this is shown in Fig. 5. The "inverted" diagonal (of black) should be easy to identify.

3.2. Relations with ERPs

Based on (1) with $\hat{y} = (0,1)$, the ERP vector $\mathbf{a}$ can be written as $\mathbf{X}^{\mathrm{T}}\hat{\mathbf{y}}$ for a class i, where $\hat{\mathbf{y}}$ depends on i while $\mathbf{X}$ does not . Consequently, the Tikhonov-regularized weight vector $\mathbf{w}$ is simply a linear transformation of the corresponding ERP vector $\mathbf{a}$, i.e., (1) becomes

$$\mathbf{w} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda^2\mathbf{I})^{-1}\mathbf{a}.$$

Depending on the actual bandpass filter used for creating the prototypes $\mathbf{f}$ of FA, the prototypes could also be expressed as linear transformations of the ERPs. One example is to approximate the ideal bandpass filter with a finite-impulse-response filter characterized by an impulse response $h(n)$. Writing the filter as a Toeplitz matrix $\mathbf{H}$, $\mathbf{f}$ can then be computed by a linear transformation:

$$\mathbf{f} = \mathbf{H}\mathbf{a}.$$

The fact that the weights and the FA prototypes are linear transformations of the averages implies that these different concepts have, in their representations, a linear relationship. This is a strong restriction, but makes the interpretation of each concept more intuitive.

3.3. Regularization impact

3.3.1. SCC

For EEG classifications where data are generally sparse in comparison to the complexity of the model, regularization often helps achieve good classification rates [13]. Some typical results demonstrating the importance of optimizing the regularization parameter λ are shown in Fig. 6. Similar results are found in other experiments involving visual stimuli of geographical sentences [14]. In Fig. 7, we show the correlation maps corresponding to a series of regularization parameters λ^2 with $\hat{y} = (0,1)$.

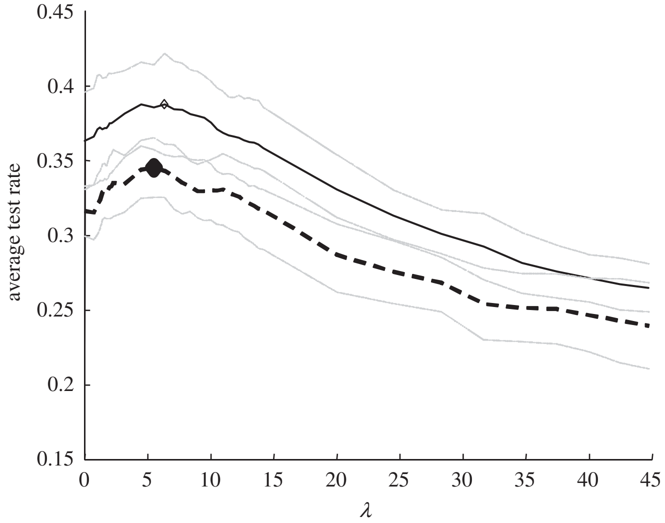


Fig. 6. Average classification rates against λ (on the x-axis), based on 10 permutations of 240 test trials of the visual-image experiment, are shown. The solid line is the test rate achieved by using 560 training trials, and the thick dashed line the rate for 320 training trials. The thin dashed lines show the standard deviations of the rates. The ◇ symbols show the maximum rates for the two conditions.

3.3.2. MCC

One might expect a similar kind of interpretation is plausible for MCC. However, there is a general problem if we carry through a simple thought experiment. Assume two identical channels $\mathbf{x}$, $\mathbf{x}'$ are recorded, and a MCC is built with weights $\mathbf{w}_{xx'}^{\mathrm{T}} = [\mathbf{w}_x^{\mathrm{T}}\mathbf{w}_{x'}^{\mathrm{T}}]$ using the concatenated training $[\mathbf{x}\mathbf{x}']$ as the training trials to compute an output. Assume the SCC obtains a unique optimal solution of $\mathbf{w}_{\mathrm{opt}}^{\mathrm{T}}$ based on training trials $\mathbf{x}$, then the set of weights $\mathbf{w}_{xx'}^{\mathrm{T}} = [\mathbf{w}_x^{\mathrm{T}}\mathbf{w}_{x'}^{\mathrm{T}}]$ which satisfies

$$\mathbf{w}_x + \mathbf{w}_{x'} = \mathbf{w}_{\mathrm{opt}},$$

are all valid solutions. The solution of MCC weights $\mathbf{w}_x$ and $\mathbf{w}_{x'}$ corresponding to the two channels can therefore be rewritten in terms of diagonal matrices $\mathbf{D}_{\boldsymbol{\alpha}}$ and $\mathbf{D}_{\mathbf{1}-\boldsymbol{\alpha}}$, with diagonals defined by vectors $\boldsymbol{\alpha}$ and $\mathbf{1} - \boldsymbol{\alpha}$:

$$\mathbf{w}_x = \mathbf{D}_{\boldsymbol{\alpha}}\mathbf{w}_{\mathrm{opt}},$$
$$\mathbf{w}_{x'} = \mathbf{D}_{\mathbf{1}-\boldsymbol{\alpha}}\mathbf{w}_{\mathrm{opt}}.$$

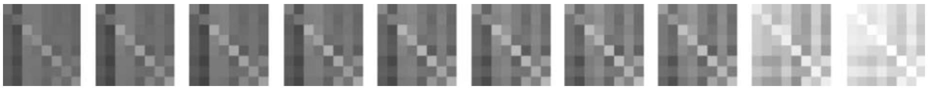


Fig. 7. Correlation maps to show the effect of regularization with $\hat{y} = (0, 1)$. White corresponds to a correlation coefficient of $+1$, while black corresponds to a correlation coefficient of -1. The regularization parameter λ increases from left to right. These results are for the best subject of the visual-image experiment.

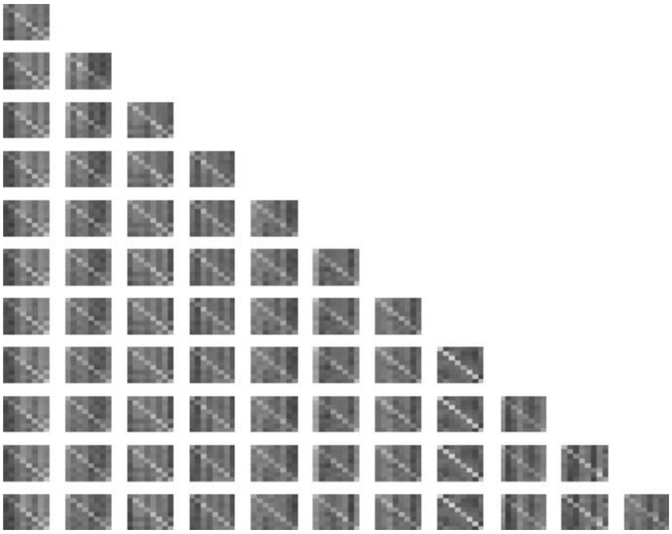


Fig. 8. Correlation maps for the regularized multichannel classifier's weights (TLM) with average prototypes (ERP). The maps are arranged in a matrix form in which map C_{jk} compares $p(k)$ and $q(k)$, where $p(k)$ is the average prototype of single channel m_k and $q(k)$ is the weight vector of same channel m_k extracted from MCC training when j channels are used. Channels are ranked by average test rates and m_k is the kth best channel. It is important to note the presence of the strong diagonals.

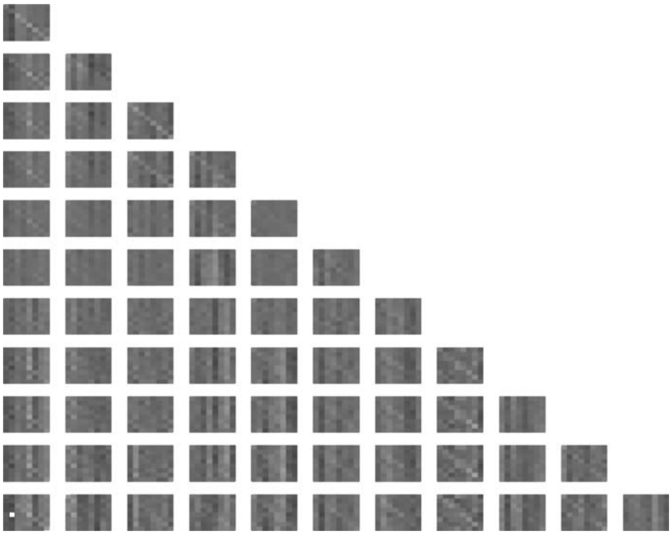


Fig. 9. Correlation maps for the non-regularized multichannel classifier's weights with the ERPs. The maps are arranged in a matrix form in which map C_{jk} compares $p(k)$ and $q(k)$, where $p(k)$ is the average prototype of single channel m_k and $q(k)$ is the weight vector of same channel m_k extracted from MCC training when j channels are used. Channels are ranked by average test rates and m_k is the kth best channel. It is important to note the absence of the strong diagonals.

Therefore, a unique solution of SCC implies an infinite number of solutions for MCC if the channels are identical. In general, the solutions are parameterized by $\boldsymbol{\alpha}$. This poses a problem for interpretation of MCC weights. An interpretation is only possible if the parameter vector is a constant for all time i (i.e., $\boldsymbol{\alpha} = \alpha$, in which the weights are scaled in a way that is invariant to correlation). In fact, this is an example of an inversion problem for (1) which results in the degeneration of the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. It is worth noting that a similar problem could also occur for SCC when the matrix $(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda^2\mathbf{I})$ is non-invertible, as is sometimes the case for small values of λ.

In reality, the recordings of different channels at any time i are not the same across trials. Even though the problem of MCC (or even SCC) may still be non-convex, we found that interpretations are still possible. In particular, we see the impact of regularization on MCC classification in our previous work.

For all MCC interpretations, channels are ranked based on average classification rates over a few permutations, and m_k is the kth best channel according to this rank. The grayscale maps (Figs. 8–13) are arranged in a matrix form, in which map C_{jk} computes the correlation between the average prototype $p(k)$ (for single channel m_k) and $q(k)$ (the

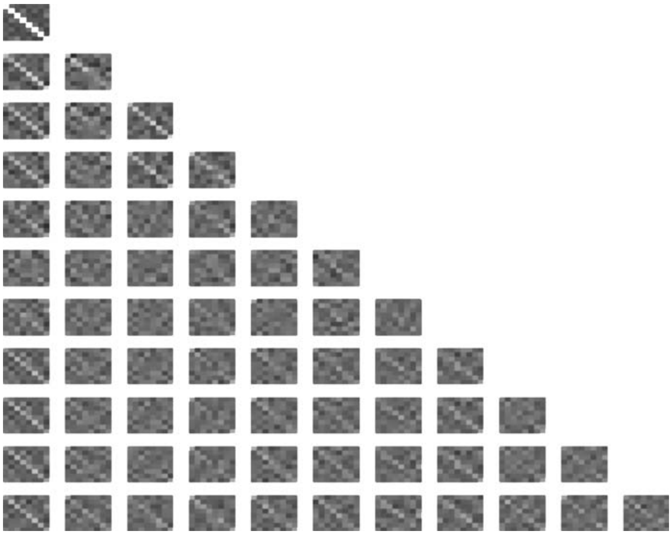


Fig. 10. Correlation maps for the non-regularized multichannel classifier's weights with non-regularized SCC weights. The maps are arranged in a matrix form in which map C_{jk} compares $p(k)$ and $q(k)$, where $p(k)$ is the average prototype of single channel m_k and $q(k)$ is the weight vector of same channel m_k extracted from MCC training when j channels are used. Channels are ranked by average test rates and m_k is the kth best channel. It is important to note the absence of the strong diagonals, excepting C_{11}.

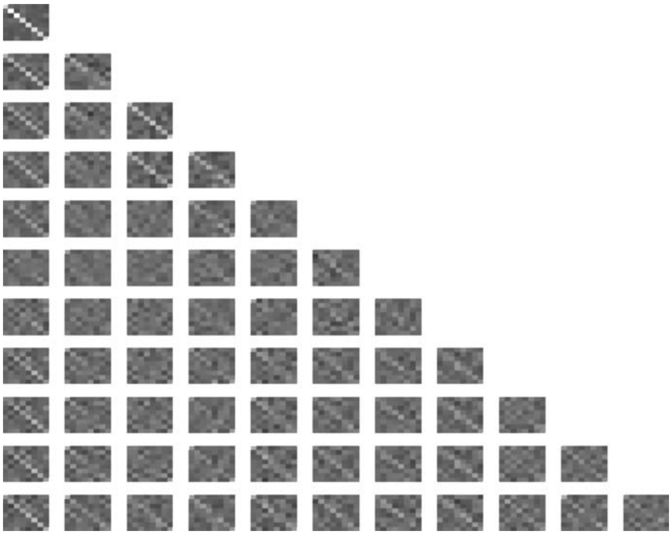


Fig. 11. Correlation maps for the non-regularized multichannel classifier's weights with regularized SCC weights. The maps are arranged in a matrix form in which map C_{jk} compares $p(k)$ and $q(k)$, where $p(k)$ is the average prototype of single channel m_k and $q(k)$ is the weight vector of same channel m_k extracted from MCC training when j channels are used. Channels are ranked by average test rates and m_i is the ith best channel. It is important to note the absence of the strong diagonals, excepting C_{11}.

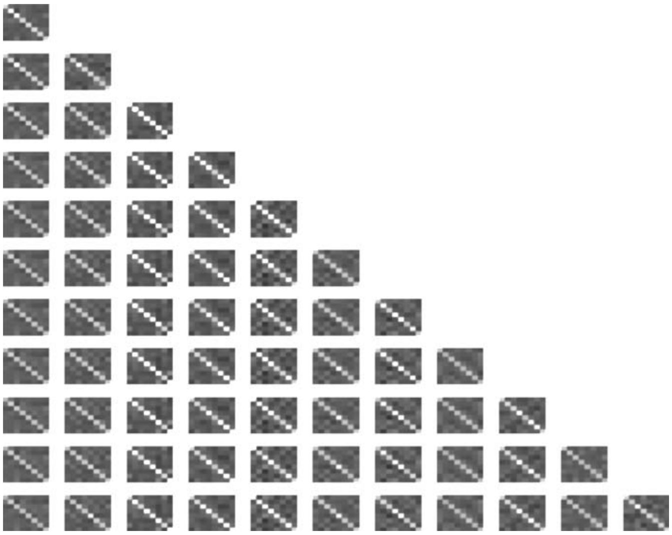


Fig. 12. Correlation maps for the regularized multichannel classifier's weights with non-regularized SCC weights. The maps are arranged in a matrix form in which map C_{jk} compares $p(k)$ and $q(k)$, where $p(k)$ is the average prototype of single channel m_k and $q(k)$ is the weight vector of same channel m_k extracted from MCC training when j channels are used. Channels are ranked by average test rates and m_k is the kth best channel. It is important to note the presence of the strong diagonals.

weights of the same channel m_k,extracted from MCC training when j channels are used). The weights are extracted from the optimal weights estimated for MCC, which is a concatenation of the channels [13,14].

Given the previous examples, one can see that in order to specify the set of maps, all we need to do is to define the two signals, $p(k)$ based on SCC data and $q(k)$ based on MCC data, for correlations. Table 3 is a summary of the maps we discuss for the remainder of this article. Our main goal is to identify the presence of strong diagonals. As stated before, the classification rates are good when TLM is regularized, especially when the number of channels used is large.

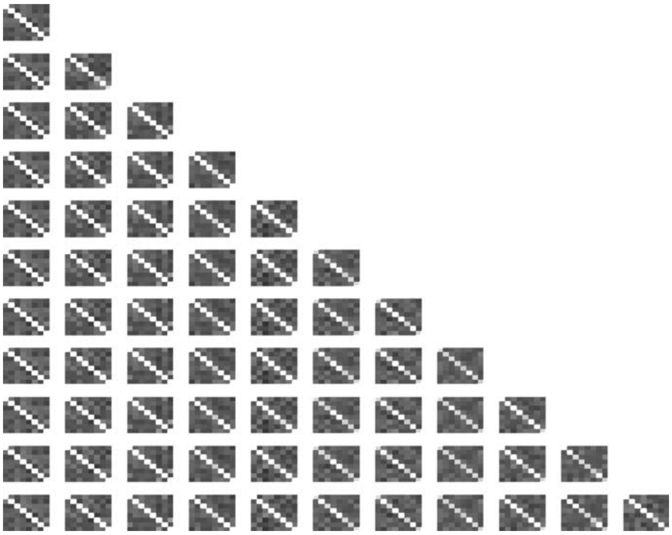


Fig. 13. Correlation maps for the regularized multichannel classifier's weights with regularized SCC weights. The maps are arranged in a matrix form in which map C_{jk} compares $p(k)$ and $q(k)$, where $p(k)$ is the average prototype of single channel m_k and $q(k)$ is the weight vector of same channel m_k extracted from MCC training when j channels are used. Channels are ranked by average test rates and m_k is the kth best channel. It is important to note the presence of the strong diagonals.

Table 3
A summary of the gray-scale maps

Fig.	$p(k)$ based on SCC	$q(k)$ based on MCC	Presence of strong diagonals
8	ERP	Regularized TLM	Yes
9	ERP	Non-regularized TLM	No
10	Non-regularized TLM	Non-regularized TLM	No
11	Regularized TLM	Non-regularized TLM	No
12	Non-regularized TLM	Regularized TLM	Yes
13	Regularized TLM	Regularized TLM	Yes

Strong diagonals are present when the weights are obtained from a regularized TLM model with good classification rates for both single and multichannel settings.

In Fig. 8, the $p(k)$ are the ERPs for single channel m_k, and the $q(k)$ are the weight vectors extracted from the regularized MCC solution for channel m_k. By observing the diagonal, or using any of the metrics above, we see that the regularized solution of a linear model yields prototypes which resemble the corresponding ERPs. All these prototypes are in fact similar to those of the SCC case shown above. Such similarity seems to be characteristic of weights that yield good classification rates.

When no regularization is used, not only do the classification rates deteriorate [13], but the resemblances among the prototypes also disappear. Fig. 9 is similar to Fig. 8, except that the $q(k)$ are extracted from the non-regularized MCC weights. Not only does the diagonal disappear in this case, but classification rates are also poor, yielding rates close to the chance level of 30 out of 240. As the number of channels increases, the diagonals become less distinct, paralleling poorer classification rates.

The same strategy can be used to see the multichannel effect of TLM by setting $p(k)$ to be the weights computed for each channel, instead of the averaged prototype (with or without regularization). In Fig. 10, the $p(k)$ are the non-regularized SCC weights and the $q(k)$ are the extracted non-regularized MCC weights. In Fig. 11, $p(k)$ are the regularized SCC solutions and $q(k)$ the non-regularized MCC weights. In both figures, diagonals are not distinct when k gets large, showing how similarity among corresponding prototypes disappears with non-regularized $q(k)$.

Figs. 12 and 13 are the same as Figs. 10 and 11, except that the $q(k)$ are regularized. In these cases, similarity exists even with a large k, regardless of whether the $p(k)$ are regularized or not. Regularization of MCC yields weights that preserve similarity to the SCC, and at the same time, achieves better classification rates than those without regularization. Given large perceptrons whose inputs are training trials of dimension $kN + 1$ (where k is the number of channels and N is the length of a downsampled training trial), channel concatenation remains a good scheme both in terms of obtaining classification rates.

4. Conclusion

By linking a Tikhonov-regularized linear model (TLM) to averaging (ERP) and observing their equivalence in the asymptotic case, TLM weights become more interpretable. In the single-channel case, the weight vector can be viewed as a non-linear mixture of the average and the pseudoinverse solution.

We naturally first stressed the importance of the linear model that yielded the best classification rate. But it is also desirable to have a physical interpretation of what the weights of the best model resemble. So our answer is that the temporal sequence of weights of the best model resembles the time series of average amplitudes of the ERP model. But of equal interest are the differences between this ERP time series and the weights of the best model. The most reasonable conjecture, it seems to us, is that the best model has eliminated some of the extraneous noise still present in the ERP time series. It has also taken explicit account of differences among stimulus-associated classes. These differences are the foundation of linear discriminant statistical analysis.

References

[1] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Wiley, New York, 2001.

[2] F.D. Foresee, M.T. Hagan, Gauss–Newton Approximation to Bayesian Regularization, 1997, pp. 1930–1935.

[3] O. Intrator, N. Intrator, Interpreting neural-network results: a simulation study, Comput. Stat. Data Anal. 37 (2001).

[4] J. Intriligator, J. Polich, On the relationship between EEG and ERP variability, Int. J. Psychophysiology 2 (1995) 59–74.

[5] D.J.C. MacKay, Bayesian interpolation, Neural Comput. 4 (3) (1992) 415–447.

[6] J.L. Rodgers, W.A. Nicewander, Thirteen ways to look at the correlation coefficient, Am. Stat. 42 (1988) 59–66.

[7] M.D. Rugg, M.G.H. Coles, Electrophysiology of Mind, Oxford Psychology Series, Oxford, 1995.

[8] G. Strang, Linear Algebra and Its Applications, Saunders, Philadelphia, 1988.

[9] P. Suppes, B. Han, Brain-wave representation of words by superposition of a few sine waves, Proc. Natl. Acad. Soc. 97 (15) (2000) 8738–8743.

[10] P. Suppes, B. Han, J. Epelboim, Z.-L. Lu, Invariance between subjects of brain wave representations of language, Proc. Natl. Acad. Soc. 96 (22) (1999) 12953–12958.

[11] P. Suppes, B. Han, J. Epelboim, Z.-L. Lu, Invariance of brain-wave representations of simple visual images and their names, Proc. Natl. Acad. Soc. 96 (25) (1999) 14658–14663.

[12] A.N. Tikhonov, V.Y. Arsenin, Solutions of Ill-posed Problems, V H Winston and sons, Washington, DC, 1977.

[13] D.K. Wong, Multichannel Classification of Brain-wave Representations of Language by Perceptron-based Models and Independent Component Analysis, PhD Dissertation, Stanford University, Stanford, 2004.

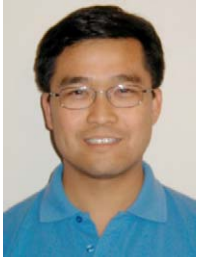
[14] D.K. Wong, M.P. Guimaraes, E.T. Uy, P. Suppes, Classification of individual trials based on the best independent component of EEG-recorded sentences, Neurocomputing 61 (2004) 479–484.

[15] Z. Zhou, S. Chen, Z. Chen, Mining typhoon knowledge with neural networks, IEEE International Conference on Tools with Artificial Intelligence, 1999, pp. 325–326.

Dik Kin Wong completed his BS in Electrical Engineering at the University of Michigan at Ann Arbor in 1997. He received his MS and Ph.D. in Electrical Engineering with a minor in Computer Science at Stanford University in 2004. He is currently a research associate at the Center for Study of Language and Information at Stanford University. He is interested in statistical learning and signal processing of brain data.

E. Timothy Uy received his BS in Applied Physics in 1997 at the California Institute of Technology. He completed his MS and Ph.D. in Applied Physics in 2005 at Stanford University focusing on dry electroencephalography and brain-wave classification. He is currently a principal at Gallivan, Gallivan and O'Melia, an electronic discovery firm. His interest is in (1) the convergence of exotic data processing schemes (e.g., holographic image processing), brain data, and data mining, and (2) interdisciplinary collaboration.

Marcos Perreau Guimaraes received his Ph.D. in computer sciences at Universitè Renè Descartes - Paris V in France, where he taught CS and applied math for two years. Since 2000, he has worked on statistical and computer methods for brain signal classification at the Center for Study of Language and Information at Stanford University.



Wayne Yang obtained his BS degree in nuclear physics from Beijing University in 1987, and he finished his graduate study in nuclear medicine in 1990. In 1998, he obtained his Ph.D. degree in experimental psychology from New York University, and he received postdoctoral training at the University of Pennsylvania. In 2000, he worked as a research associate at the Center for Study of Language and Information at Stanford University. In 2001, he worked at Vitria technology as a human factor engineer. He founded Mintel Learning Technology Inc. in October 2001.

Patrick Suppes is the Lucie Stern Professor Emeritus of Philosophy at Stanford University. He has published widely in philosophy and the social sciences, especially psychology. He is doing research on the brain, with emphasis on language and visual images. His last book appeared in 2002, Representation and Invariance of Scientific Structures. He is a member of the U.S. National Academy of Sciences.