Bayesian Classification of Single-Trial Event-Related Potentials in EEG*

Jens Kohlmorgen Benjamin Blankertz

Fraunhofer FIRST.IDA
Kekuléstr. 7, D-12489 Berlin, Germany
E-mail: {jek, blanker}@first.fraunhofer.de

Abstract

We present a systematic and straightforward approach to the problem of single-trial classification of event-related potentials (ERP) in EEG. Instead of using a generic classifier off-the-shelf, like a neural network or support vector machine, our classifier design is guided by prior knowledge about the problem and statistical properties found in the data. In particular, we exploit the well-known fact that event-related drifts in EEG potentials, albeit hard to detect in a single trial, can well be observed if averaged over a sufficiently large number of trials. We propose to use the average signal and its variance as a generative model for each event class and use Bayes' decision rule for the classification of new, unlabeled data. The method is successfully applied to a data set from the NIPS*2001 Brain-Computer Interface post-workshop competition. Our result turns out to be competitive with the best result of the competition.

1 Introduction

The analysis of EEG (electro-encephalogram) is one of the most challenging problems in signal processing and machine learning research. A particularly difficult task is the analysis of event-related potentials (ERP) from individual events ('single-trial'), which recently gained increasing attention for building brain-computer interfaces. The problem is in the high inter-trial variability of the EEG signal, where the interesting quantity, e.g. a slow shift of the cortical potential, is largely hidden in the 'background' activity and only becomes evident by averaging over a large number of trials.

Birbaumer et al. investigate slow cortical potentials (SCP) and how they can be self-regulated in a feedback scenario. In their *thought translation device*

*A preliminary version of this article appeared in the Proceedings of the International Conference on Artificial Neural Networks (ICANN), August 2002.

(Birbaumer et al. [1999]), patients learn to produce cortical negativity or positivity at a central scalp location at will, which is fed back to the user. After some training, patients are able to transmit binary decisions in a 4 sec periodicity with accuracy levels up to 85% and therewith control a language support program or an Internet browser.

Pfurtscheller et al. built a BCI system based on event-related (de-)synchronization (ERD/ERS), typically of the μ and central β rhythm, which are electrical oscillations originating from the motor areas of the brain in the 8–13 Hz resp. 15–30 Hz frequency range. It is used for on-line classification of imaginations or preparations of, for example, left/right index finger, feet, and tongue movement. Typical pre-processing techniques use adaptive autoregressive parameters, common spatial patterns (after band-pass filtering) and band power in subject-specific frequency bands. Classification is done by Fisher discriminant analysis or multi-layer neural networks. In classification of exogenous movement preparations, rates of 98%, 96% and 75% (for three subjects, respectively) are obtained before movement onset[1] in a 3-classes task and trials of 8 sec (Peters et al. [2001]). Only *selected*, artifact free trials (less than 40%) were used. A tetraplegic patient (i.e. paralyzed in all limbs resulting from injury to the spinal cord) controls his hand orthosis using the Graz BCI system.

Wolpaw et al. study EEG-based cursor control (Wolpaw et al. [2000]), translating the power in subject-specific frequency bands, or autoregressive parameters, from two spatially filtered scalp locations over sensorimotor cortex into vertical cursor movement. Users initially gain control by various kinds of motor imagery (the setting favors >movement< vs. >no movement< in contrast to >left< vs. >right<), which they report to use less and less as feedback training continues. In cursor control trials of at least 4 sec duration, *trained* subjects reach accuracies of over 90%. Some subjects acquired also considerable control in a two-dimensional set-up.

To approach the problem of single-trial ERP classification, we use an EEG data set from the NIPS*2001 Brain-Computer Interface (BCI) post-workshop competition.[2] The data set consists of 516 single trials of pressing a key on a computer keyboard with fingers of either the left or right hand in a self-chosen order and timing ('self-paced key typing'). A detailed description of the experiment can be found in Blankertz et al. [2002]. For each trial, the measurements from 27 Ag/AgCl electrodes are given in the interval from 1620 ms to 120 ms *before* the actual key press. The sampling rate of the chosen data set is 100 Hz, so each trial consists of a sequence of $N = 151$ data points. The task is to predict if the upcoming key press is from the left or right hand, given only the respective EEG sequence. A total of 416 trials are labeled (219 'left' events, 194 'right' events, and 3 rejected trials due to artifacts) and can be used for building a binary classifier. One hundred trials are unlabeled and make up the evaluation test set for the competition. We construct our classifier under the conditions of the competition, i.e. without using the test set for building the model, but since

[1] more precisely: before the *mean* EMG onset time. For some trials this is before, for others after EMG onset.

[2] publicly available at http://newton.bme.columbia.edu/competition.htm

the true test set labels are publicly available now, we can report the test set error of our classifier at the intermediate steps of our model design and finally compare it with the result of the competition.

2 Designing a Bayesian Classifier

As outlined in Blankertz et al. [2002], the experimental set-up used for obtaining the competition dataset aims at detecting lateralized slow negative shifts of cortical potential, known as 'Bereitschaftspotential' (BP), which have been found to precede the initiation of the movement (Lang et al. [1989]; Cui et al. [1999]). These shifts are typically most prominent at the lateral scalp positions C3 and C4 of the international 10-20 system, which are located over the left and right hemispherical primary motor cortex.

Fig. 1 illustrates this for the given training data set. The left panel in Fig. 1 shows the measurements from each of the two channels, C3 and C4, *averaged* over all trials for *left* finger movements, and the right panel depicts the respective averages for *right* finger movements. Respective plots are also shown for channel C2, which is located next to C4. It can be seen that, on the average, a right finger movement clearly corresponds to a preceding negative shift of the potential over the left motor cortex (C3), and a left finger movement corresponds to a negative shift of the potential over the right motor cortex (C2, C4), which in this case is even more prominent in C2 than in C4 (left panel). The crux is that this effect is largely obscured in the individual trials due to the large variance of the signal, which makes the classification of individual trials so difficult. Therefore, instead of training a generic classifier on the individual trials, we propose to exploit the above (prior) knowledge straightaway and use the averages directly as the underlying model for left and right movements.

It can be seen from Fig. 1 that the difference between the average signals C4 and C3, and likewise between C2 and C3, is decreasing for left events, but is increasing for right events. We can therefore merge the relevant information from both hemispheres into only one scalar signal by using the difference of either C4 and C3 or C2 and C3. In fact, it turned out that the best performance (in terms of the leave-one-out cross-validation error, see below) can be achieved when subtracting C3 from the mean of C4 and C2. That is, as a first step of pre-processing/variable selection, we just use the scalar EEG signal, $y = (C2 + C4)/2 - C3$, for our further analysis.

The respective averages, $\mu_L(t)$ and $\mu_R(t)$, of the signal $y(t)$ for all left and right events in the training set, together with the standard deviations at each time step, $\sigma_L(t)$ and $\sigma_R(t)$, are shown in Fig. 2. A scatter plot of all the training data points underlies the graphs to illustrate the large variance of the data in comparison to the feature of interest: the drift of the mean.

The idea is then to use the left and right averages and the corresponding standard deviations directly as generative models for the left or right trials. Under the assumption of a Gaussian distribution, the probability of observing y at time t given the left model, $M_L = (\mu_L, \sigma_L)$, can be expressed by the density

function

$$p(y(t) \mid M_L) = \frac{1}{\sqrt{2\pi}\,\sigma_L(t)} \exp\left(-\frac{(y(t)-\mu_L(t))^2}{2\sigma_L(t)^2}\right). \quad (1)$$

The probability density $p(y(t) \mid M_R)$ for the right model can be expressed accordingly. Assuming a Gaussian distribution is indeed justified for this data set: we estimated the density of the data at each time step with a kernel density estimator and consistently found a distribution very close to a Gaussian (see Fig. 3). Because of the small sample size, it might then be justified to assume that also the joint density of the sequence of observations, $y = (y(1), \ldots, y(N))^T$, is Gaussian, i.e. for the left events we assume

$$p(y \mid (\mu_L, \Sigma_L)) = \frac{1}{(2\pi)^{N/2}\,|\Sigma_L|^{1/2}} \exp\left(-\frac{1}{2}(y-\mu_L)^T \Sigma_L^{-1}(y-\mu_L)\right), \quad (2)$$

where Σ_L is the $N \times N$ covariance matrix. For the right events, we accordingly get $p(y \mid (\mu_R, \Sigma_R))$. In principle, one could use Eq. (2) – with the full covariance matrix – as a generative model for classification. For the competition data set however, there are good reasons not to do that. First of all, the off-diagonal elements in the covariance matrices estimated from the left and right training data are small in comparison to the elements on the diagonal. This suggests to prefer a simpler model with a diagonal matrix, since the deviation of each estimated matrix from a diagonal matrix could just be due to the small sample size. Second, we actually applied the full covariance model to classify the data and found that it was outperformed by the simplified models (see section 3). Moreover, the model is very susceptible to overfit the training data. This is not surprising, if one considers the many parameters that are to be estimated in the covariance matrix.

We therefore choose a diagonal matrix for our model. The joint density of y in Eq. (2) can then be written as the product of the densities of the individual observations given in Eq. (1),

$$p(y \mid M_L) = \prod_{t=1}^{N} p(y(t) \mid M_L). \quad (3)$$

Vice versa, the probability of the model M_L – given y – can be expressed by Bayes' rule,

$$p(M_L|y) = \frac{p(y|M_L)\; p(M_L)}{p(y)}, \quad (4)$$

where $p(y)$ is the unconditional probability density of y, called evidence, and $p(M_L)$ is the unconditional (prior) probability of M_L. We get $p(M_R|y)$ accordingly and then use Bayes' decision rule, $p(M_L|y) > p(M_R|y)$, to decide which model to choose for a given y. According to Eq. (4), this can be written as

$$p(y|M_L)\; p(M_L) \;>\; p(y|M_R)\; p(M_R). \quad (5)$$

By applying the negative logarithm we get

$$-\log p(y|M_L) - \log p(M_L) \;<\; -\log p(y|M_R) - \log p(M_R) \quad (6)$$

and by inserting Eqs. (3) and (1)

$$\sum_{t=1}^{N} \log \sigma_L(t) \; + \; \sum_{t=1}^{N} \frac{(y(t) - \mu_L(t))^2}{2\sigma_L(t)^2} \; - \; \log p(M_L)$$
$$< \quad \sum_{t=1}^{N} \log \sigma_R(t) \; + \; \sum_{t=1}^{N} \frac{(y(t) - \mu_R(t))^2}{2\sigma_R(t)^2} \; - \; \log p(M_R). \qquad (7)$$

We are left with determining the prior probabilities of the models, $p(M_L)$ and $p(M_R)$. Since there is no a priori preference for left or right finger movements in the key typing task, we can set $p(M_L) = p(M_R)$, which cancels out the respective terms in Eq. (7). Furthermore, – for the given data set – the standard deviations for left and right trials, $\sigma_L(t)$ and $\sigma_R(t)$, are very similar to each other (cf. Fig. 2), i.e. $\sigma_L(t) \approx \sigma_R(t)$, and also their variation in time can be neglected, such that we can replace $\sigma_L(t)$ and $\sigma_R(t)$ in our model by a single constant σ. It turns out that this does not diminish the classification performance. In fact, this simplification actually *improves* the (leave-one-out) performance of our model on the training set. For the competition data, the decision rule can thus be written as

$$\sum_{t=1}^{N} (y(t) - \mu_L(t))^2 \quad < \quad \sum_{t=1}^{N} (y(t) - \mu_R(t))^2. \qquad (8)$$

The terms on both sides are now simply the squared Euclidean distances of a given input sequence y to the left and right average signal. The statistical properties of the data ultimately allowed us to derive this very simple classification rule from a rather general Bayesian approach.

3 Results and Improvements

The obtained decision rule (Eq. (8)) can readily be applied to our selected quantity y from the competition data set. The result without any further preprocessing of the signal is 19.13% misclassifications (errors) on the training set, 19.85% leave-one-out (LOO) cross-validation error (on the training set), and 17% error on the test set. The leave-one-out cross-validation error is obtained by leaving out one sample from the training set, computing the models on the remaining samples, and testing the resulting models on the left-out sample. This procedure is performed for all samples in the training set, such that each sample is held out and tested once. The leave-one-out error is then the error over all held-out samples. It is the LOO error that is to be minimized in order to obtain a good generalization performance.

To improve the result, we normalized the data of each trial to zero-mean, which improved the errors to 15.74%/17.19%/6% training/LOO/test set error. In particular the test set error is already remarkably small at this point. Since

this quantity can not be used to optimize the classifier (under competition conditions it would not be available at all), we must rely solely on the minimization of the LOO error, which is still very large.[3]

The normalization of the data to unit-variance further enhances the LOO performance (15.25%/16.46%/6% training/LOO/test set error). A more substantial improvement can easily be understood from Fig. 2. Clearly, the data points at the end of the sequence have much more discriminatory power than the points at the beginning. Moreover, we presume that the points at the beginning mainly introduce undesirable noise into the decision rule. We therefore successively reduced the length of the sequence that enters into the decision rule via a new parameter D,

$$\sum_{t=D}^{N}(y(t)-\mu_L(t))^2 \quad < \quad \sum_{t=D}^{N}(y(t)-\mu_R(t))^2. \tag{9}$$

Fig. 4 shows the classification results for $D = 1,\ldots,N$, ($N = 151$), on the normalized data. Surprisingly, using only the last 12 data points of the EEG sequence yields the best LOO performance: the LOO error minimum of 8.96% is at $D = 140$, which is a considerable improvement against 16.46% for $D = 1$.

A further, somehow related improvement can be achieved by excluding a number of data points from the end of the sequence when computing the mean for the zero-mean normalization. This effectively improves the alignment of the sequences with a common (zero) baseline. The normalization is then performed as follows: For each sequence, just the partial average $\bar{y} = \sum_{t=1}^{M} y(t)/M$, with $M < N$, is subtracted from each element in the sequence, which results in pre-processed sequences that have a zero mean only with respect to the first M data points. Fig. 5 depicts the classification results for $M = 1,\ldots,N$, using $D = 140$ and unit-variance normalization (still performed with respect to all data points in each sequence). The LOO minimum is 7.99% at $M = 125$. We found that this is indeed the optimal LOO error with respect to all possible combinations of M and D. At this optimum, we obtain a test set error of 5%. Fig. 6 shows the respective distances (cf. Eq. (9)) of all trials to the left and right model. For the majority of trials there is a clear difference between the distance to the left and the distance to the right model. Note that the test set error in Fig. 5 even reaches a minimum of 2% at $M = 35$. However, this solution can not be found given only the training set and must be considered as a fluke for this particular test set.

We considered other types of pre-processing or feature selection, like unit-variance normalization with respect to a certain window, other choices of windows for zero-mean normalization, or using the bivariate C3/C4 signal instead of the difference signal. However, these variants did not result in better classification performance. The best leave-one-out result that we could obtain for the full covariance model was 9.93% LOO error with a corresponding test set error of

[3]The unusual result that the test set error is just about half as large as the error on the training set was consistently found throughout our experiments and is apparently due to a larger fraction of easy trials in the test set.

8%. The best result that we obtained in our preliminary investigation reported in Kohlmorgen and Blankertz [2002], was by using a model with a diagonal covariance matrix. The leave-one-out error was 8.96% and the corresponding test set error was between 4 and 5%.

4 Summary and Discussion

We presented a generative model approach to the problem of single-trial classification of event-related potentials in EEG. The finally obtained classification scheme requires only 2 or 3 EEG channels and the classification process is easily interpretable as a comparison with the average signal of each class. The application to a data set from the NIPS*2001 BCI competition led to further improvements of the classification rule, which finally resulted in 95% correct classifications on the test set (without using the test set for improving the model). The best result of the competition was 96% correct classifications. This is not a significant difference to our result, since it means that the difference is in a single test trial. However, the method used is very different. The result was achieved with a recurrent neural network with six fully connected neurons. It was trained with a procedure called dynamic noise annealing (Sottas and Gerstner [2002]). Compared to our approach, this method is much more complicated, but it has the advantage that it is not necessarily restricted to classes with Gaussian distributions. The other participants of the competition obtained results between 54% and 95% correct classifications.[4]

In contrast to the approach of using a rather generic method, like, e.g., a recurrent neural network, we demonstrated how prior knowledge about the problem as well as statistical properties of the data can be used to design a problem-specific classifier. As a result, we were able to obtain a very simple but competitive classification scheme. Due to its simplicity, it is well suited as a performance reference for evaluating more sophisticated, future approaches to the problem of EEG classification, in particular for building brain-computer interfaces. Furthermore, we expect that our approach is also useful for other time series classification tasks.

Acknowledgements: We thank S. Lemm, P. Laskov, and K.-R. Müller for fruitful discussions. This work was supported by grant 01IBB02A from the BMBF.

[4] see http://newton.bme.columbia.edu/competitionresults.htm

References

N. Birbaumer, N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kübler, J. Perelmouter, E. Taub, and H. Flor. A spelling device for the paralysed. *Nature*, 398:297–298, 1999.

B. Blankertz, G. Curio, and K.-R. Müller. Classifying single trial EEG: Towards brain computer interfacing. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14 (NIPS*01)*, pages 157–164, Cambridge, MA, 2002. MIT Press.

R. Q. Cui, D. Huter, W. Lang, and L. Deecke. Neuroimage of voluntary movement: topography of the Bereitschaftspotential, a 64-channel DC current source density study. *Neuroimage*, 9(1):124–134, 1999.

J. Kohlmorgen and B. Blankertz. A simple generative model for single-trial EEG classification. In J. R. Dorronsoro, editor, *Artificial Neural Networks – ICANN 2002*, pages 1156–1161. Springer, 2002.

W. Lang, O. Zilch, C. Koska, G. Lindinger, and L. Deecke. Negative cortical DC shifts preceding and accompanying simple and complex sequential movements. *Exp. Brain Res.*, 74(1):99–104, 1989.

B. O. Peters, G. Pfurtscheller, and H. Flyvbjerg. Automatic differentiation of multichannel EEG signals. *IEEE Trans. Biomed. Eng.*, 48(1):111–116, 2001.

P.-E. Sottas and W. Gerstner. Dynamic noise annealing for learning temporal sequences with recurrent neural networks. In J. R. Dorronsoro, editor, *Artificial Neural Networks – ICANN 2002*, pages 1144–1149. Springer, 2002.

J. R. Wolpaw, D. J. McFarland, and T. M. Vaughan. Brain-computer interface research at the Wadsworth Center. *IEEE Trans. Rehab. Eng.*, 8(2):222–226, 2000.

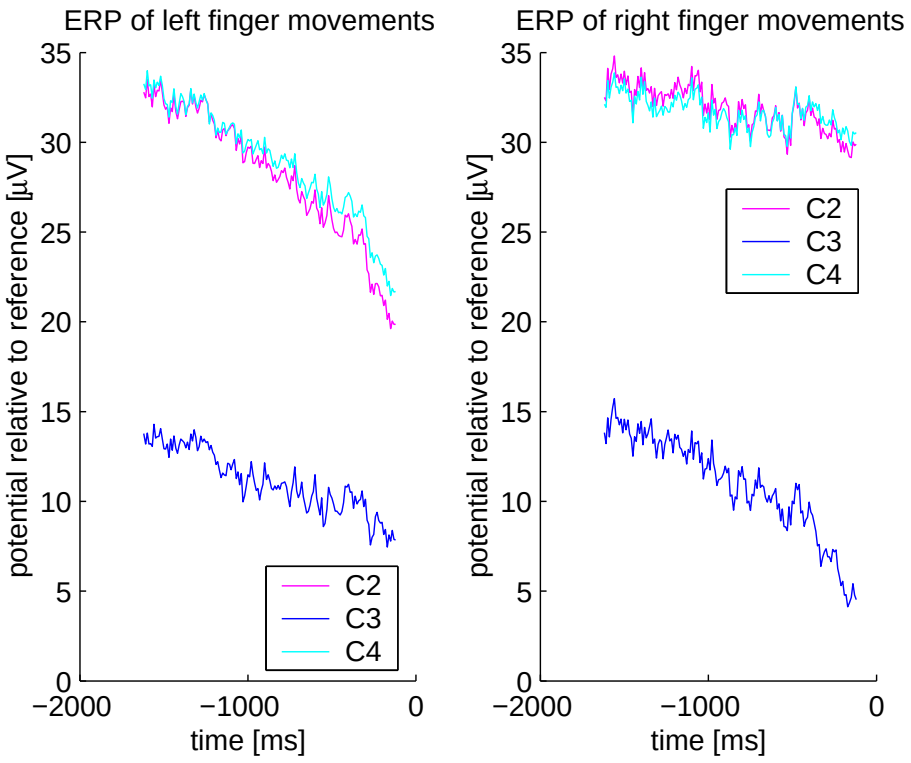

Figure 1: Averaged EEG recordings at positions C2, C3, and C4, separately for left and right finger movements. The averaging was done over all training set trials of the BCI competition data set.

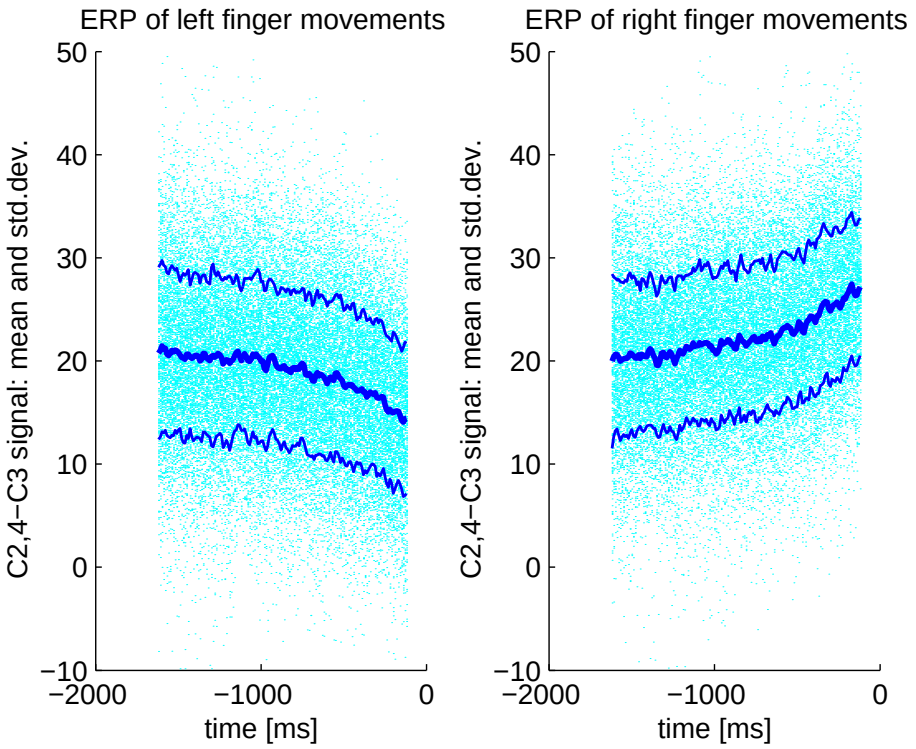

Figure 2: Mean and standard deviation of the difference signal, $y = (C2 + C4)/2 - C3$, over a scatter plot of all data points. Clearly, there is a large variance in comparison to the drift of the mean.

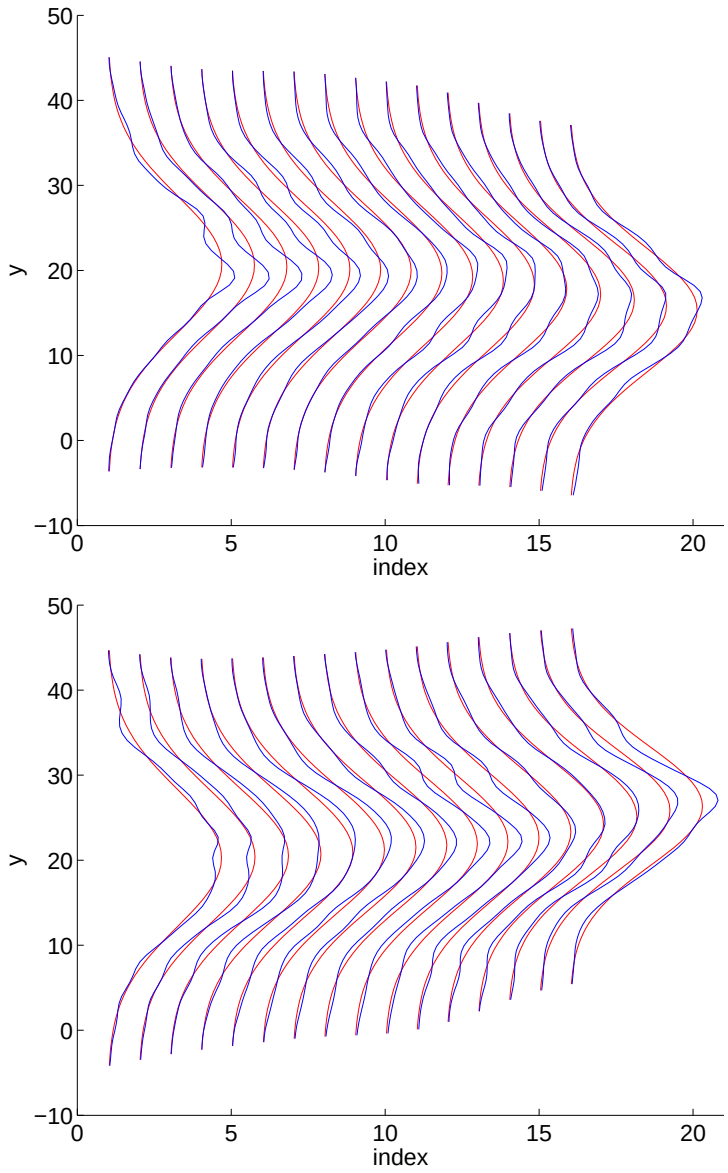

Figure 3: Top: A sequence of 16 density functions estimated from the left trials $y(t)$ in the training set at individual time steps. At every tenth time step ($t = 1, 11, 21, ..., 151$), a non-parametric kernel density function (blue lines) is estimated and plotted over a Gaussian estimate (red lines). Obviously, the non-parametric estimate is very close to a Gaussian distribution. A qualitatively similar result was obtained for the right trials (bottom).

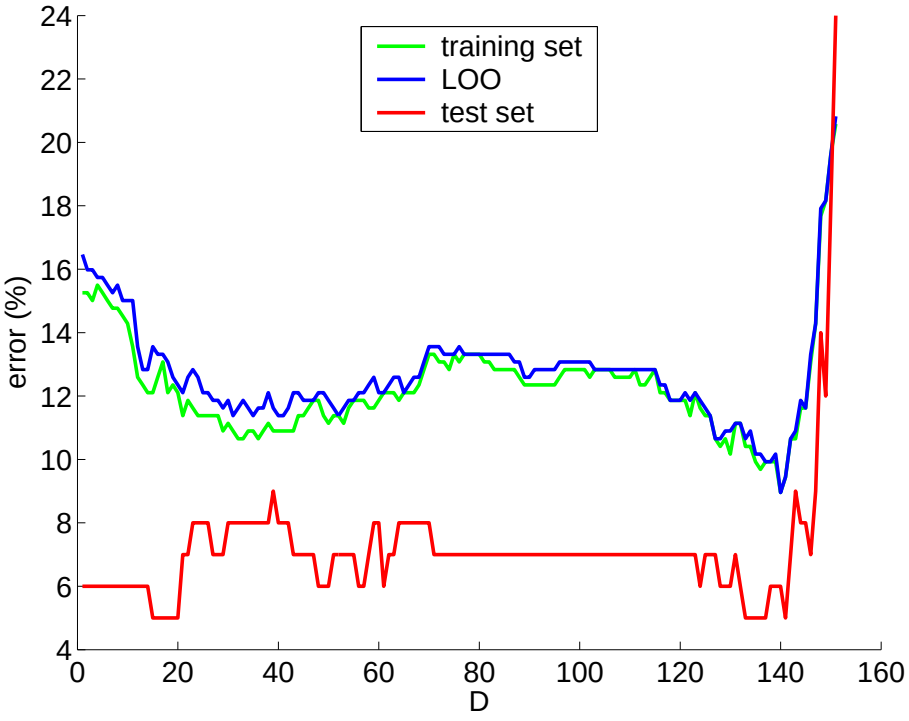

Figure 4: Training, leave-one-out (LOO), and test set error in dependence of the starting point D of the observation window.

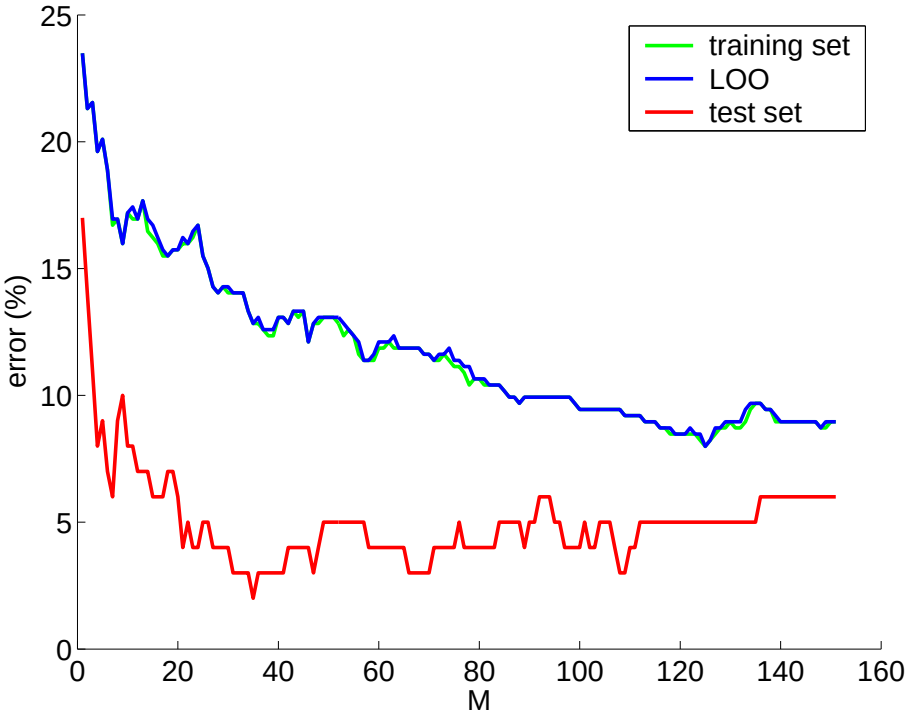

Figure 5: Training, leave-one-out (LOO), and test set error in dependence of the size M of the zero-mean window (results for $D = 140$).

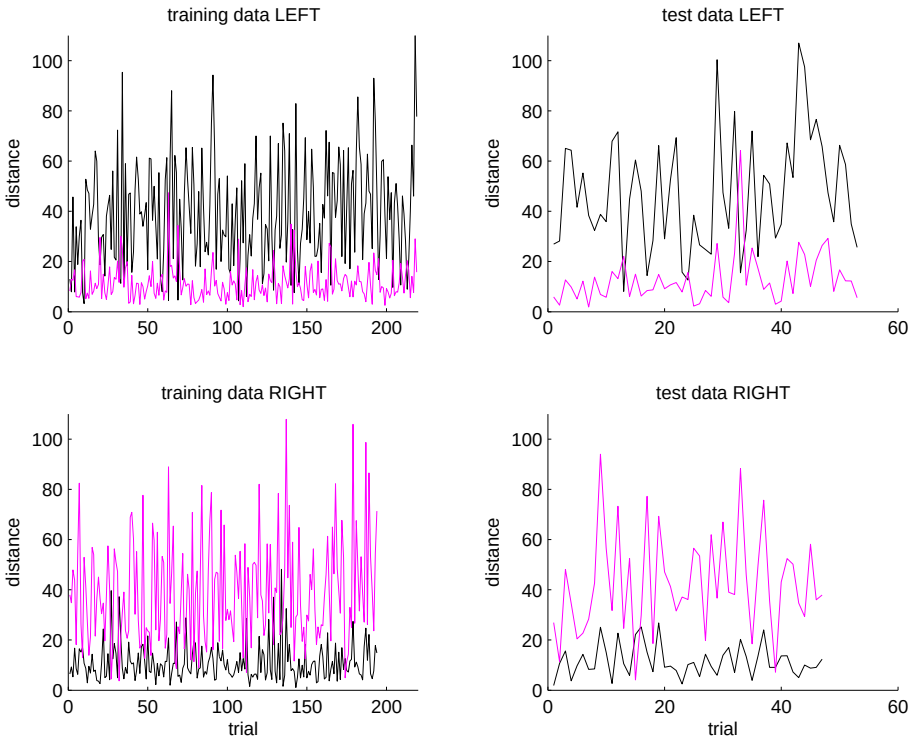

Figure 6: Distances (cf. Eq.(9)) from all trials to the finally chosen models for left and right event-related potentials ($D = 140$, $M = 125$). In most cases of misclassifications both models exhibit a small distance to the input, which indicates that there is no clear drift of the potential in these trials. (left model: magenta, right model: black)