Mining the brain with a theory of visual attention

Mads Dyrholm, Maria Nordfang, and Claus Bundesen

University of Copenhagen, Denmark

Abstract. We present a new supervised component analysis method with an application to EEG. The method detects and extracts components that are predictive of behavior relative to an expected value which is derived from a formal psychological theory of visual attention. We analyze the pre-stimulus EEG activity from a single-letter recognition task and find distinct components that each contribute to the joint prediction through separable perceptual parameters on the single-trial level.

1 Introduction

We contribute to the field of mapping from neuronal responses to behavior (e.g. Koles, 1991; Mørch et al., 1997; Parra et al., 2005; Dyrholm et al., 2007; Blankertz et al., 2008) by using a psychologically informed node that is derived from the theory of visual attention (TVA) of Bundesen (1990). TVA predicts perceptual categorization considering bottom up and top down perceptual processes, and multiple software packages exist for estimating perceptual parameters from behavioral data in purely accuracy-based paradigms with TVA (Kyllingsbæk, 2006; Dyrholm et al., 2011). In a large behavioral study, Dyrholm et al. (2011) detected marginal variation of perceptual parameters in participants performing a letter recognition task. Here we go one step further than detecting the presence and magnitude of such variation: We estimate the perceptual parameter values on a single-trial level. We do this by incorporating EEG decoders in the model such that each decoder is responsible for modulating separable perceptual TVA-parameters from trial to trial. The decoders are estimated jointly to detect and extract signals that are predictive of early categorizational perception through the TVA node.

2 Predicting perceptual categorization with TVA

TVA can be interpreted as a competition in which each item of a display is represented by a team of perceptual categories (see also Bundesen, 1990). An item wins a place in the visual short term memory (VSTM) if one or more perceptual categorizations complete before the competition is terminated and before the VSTM is full. The competition is assumed to terminate when the exposure has effectively ended. The VSTM is of limited capacity in terms of the

number of items that can be categorized, but unlimited in terms of the number of categorizations per item. Pigeonholing is sometimes used as an analogy to the encoding into VSTM, holes representing the items that are encoded, and pigeons representing potential categorizations. The result of this process is the perception that an item x belongs to a set of categories $\{i\}$. This set is only non-empty for the first K items that complete a categorization before the competition ends, K being the capacity of VSTM. The categorization rate is assumed to be constant during the effective exposure (Bundesen, 1990), i.e. the time it takes to complete a categorization follows an exponential distribution with the probability density function

$$t_{x,i} \;\sim\; v_{x,i} e^{-v_{x,i} t_{x,i}} \;, \quad t_{x,i} \geq 0 \;, \quad v_{x,i} \geq 0 \tag{1}$$

where the *categorization rate* $v_{x,i}$ is the reciprocal of the expected time to complete the categorization of item x as belonging to category i (put pigeon i in hole x). The probability that a particular categorization completes before the competition ends is derived from (1)

$$\mathrm{p}(t_{x,i} < \tau) = 1 - e^{-v_{x,i}\tau} \;, \quad \text{where} \quad \tau \equiv \begin{cases} t - t_0 & ,t >= t_0 \\ 0 & ,\text{otherwise} \end{cases} \tag{2}$$

where t_0 is a *perceptual threshold* parameter that reflects the delay from stimulus onset to the start of the competition (e.g. time for clustering the retinal image into discrete items), and τ defines the *effective exposure duration.*

In our experiment we will consider the simple task of recognizing a single letter from a display with no distractors. In that case we can equate $\mathrm{p}(t_{x,i} < \tau)$ with the probability that a participant can report the letter identity correctly without guessing (Bundesen and Harms, 1999). We will estimate the parameter $v_{x,i}$ directly, but for completeness consider equations (3) and (4) below. The categorization rate relates to bottom up and top down processes as given by

$$v_{x,i} = \eta_{x,i} \beta_i \frac{w_x}{\sum_z w_z} \tag{3}$$

where $\eta_{x,i}$ is the *sensory evidence* that category i pertains to item x, β_i is the *decision bias* associated with category i, and w_x is the attentional weight of item x (relative to all items of the display represented as $\sum_z w_z$). The attentional weights are given by

$$w_x = \sum_{j \in R} \eta_{x,j} \pi_j \tag{4}$$

where π_j is the importance of attending to elements that belong to category j, and R is the total set of perceptual categories. That is, weight will be given to items belonging to important (e.g. task relevant) perceptual categories weighted by the sensory evidence that the perceptual category does indeed pertain to the item (Bundesen, 1990; Bundesen et al., 2005).

2.1 Inserting EEG decoders

Suppose we are given the responses produced by a participant performing N trials of a single-item recognition task in brief displays. On each trial, the participant produced either a *hit* (correct identification response) or a *miss* (incorrect, or no response), and the likelihood function is then derived from (2)

$$L_n(t_0, v_{x,i}) = \begin{cases} 1 - e^{-v_{x,i}\tau} & , \text{ hit trial} \\ \text{one minus the above} & , \text{ miss trial} \end{cases} \tag{5}$$

where n is the trial number. By keeping stimulus conditions constant in the terms of equations (3) and (4) across a set of trials, but varying the exposure duration to make t_0 and $v_{x,i}$ separable, a single point estimate of t_0 and $v_{x,i}$ can be obtained to represent expected values under the conditions imposed.

But, even under identical experimental conditions, the mental state of the participant will almost surely fluctuate from trial to trial. The method that we propose here is to take trial-by-trial fluctuations into account by detecting EEG components that are predictive of behavior on the single-trial level. The EEG data is formally linked with TVA, as being potentially predictive of trial-by-trial parameter fluctuation. The parameters (in this case t_0 and $v_{x,i}$) are substituted by EEG decoder functions

$$v_{x,i} \leftarrow y_1 + \mathrm{f}(\mathbf{X}_n|\mathbf{w}_1) \quad \text{and} \quad t_0 \leftarrow y_2 + \mathrm{f}(\mathbf{X}_n|\mathbf{w}_2) \tag{6}$$

where $\mathbf{X}_n \in \mathbb{R}^{D\times T}$ is a matrix consisting of EEG measurements from D electrodes over an interval of T samples in relation to trial number n. The resulting model allows for the parameters $\mathbf{w}_1$, $\mathbf{w}_2$, y_1, and y_2 to be estimated jointly from the item recognition responses and the EEG activity by maximizing the sum of log-likelihood contributions across the trials. Once estimated, the two EEG decoders will be predictors of the perceptual parameters t_0 and $v_{x,i}$ on a single-trial level.

3 Experiment

We recorded EEG data from a participant in a single-letter recognition task which we adapted from Busch et al. (2009); Bundesen and Harms (1999); Vangkilde et al. (2012). In each trial the participant was presented with a brief display containing a single white letter either to the left or to the right of a fixation cross. The task was to correctly identify the letter. Initially, the participant was presented with a fixation cross at the center of the screen and two placeholder dots (one below and one above) each of the two possible target locations. The fixation waiting time varied on a trial basis as by a constant of 1000 ms plus a random number drawn from an exponential distribution with a mean of 500 ms (but drawn anew until the wait was less than 30 s). A letter was then presented for 0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 80, or 200 ms between either the two left or the two right placeholder dots. Then a mask came on for 500 ms and

the participant reported the letter identity, if known. The participant completed 1950 trials interleaved with short breaks at every 325 trials. Trials with an exposure duration of 0 ms were discarded from further analysis. A total of 1800 trials remained. The longest fixation waiting time was 4.7 s.

EEG was recorded at 2048Hz sampling rate using a BIOSEMI Active II system with 64 electrodes. The continuous data was inspected visually and annotated for bad segments using EEGLAB (Delorme and Makeig, 2004). The data was then lowpass filtered (anti-alias) and decimated to 64Hz sampling rate and trimmed according to the bad segment annotation. The remaining data was re-referenced to the average electrode potential and high-pass filtered to attenuate activity below 1Hz. All filters had zero group delay. Eye blink activity was subtracted using Extended ICA (Jung et al., 1998). Behavioral data from trials with missing EEG were excluded from further analysis; a total of 1795 trials remained.

3.1 Computation, results, and discussion

To analyze the pre-stimulus attentional state of the subject we used as $\mathbf{X}_n$ the 500 ms of EEG data that immediately preceded the stimulus onset on trial n. We used the following EEG decoder definition which does not focus on any particular frequency band (but see also Blankertz et al., 2008)

$$\mathrm{f}(\mathbf{X}_n|\mathbf{w}) = \lambda \,\mathrm{logistic}\left[b + \mathbf{u}^{\mathrm{T}}\mathbf{X}_n\mathbf{X}_n^{\mathrm{T}}\mathbf{u}\right] \;,\; \mathbf{w} = [\lambda; b; \mathbf{u}] \tag{7}$$

where $\mathbf{u}$ is a spatial filter (unmixing vector) such that the product $\mathbf{u}^{\mathrm{T}}\mathbf{X}_n$ produced a weighted sum of all electrode voltages in the EEG epoch of trial n (electrode j is weighted by the jth element of $\mathbf{u}$), then squared, and the sum taken over the temporal window of the epoch. This yielded a measure of *projected power* for trial n. The logistic function was sigmoidal in shape such that the decoder could limit either end or both ends of the projected power range.

We estimated $\mathbf{w}_1$ and $\mathbf{w}_2$ by maximizing the regularized Likelihood using a Gaussian prior on λ (Moody, 1992; MacKay, 1992a) while using a Gaussian Process prior on the spatial filter $\mathbf{u}$, forcing it to be smooth (Dyrholm et al., 2007). We calibrated the regularization parameters using the Evidence framework (MacKay, 1992b; Penny and Roberts, 1999). We adjusted the data scale and regularization parameters until the method started to converge in a robust manner. The EEG was in units of 0.1 mV. The regularization parameters were then calibrated by computing the Evidence on a coarse grid in 4 dimensions: prior deviation for λ_1 at values $\{0.05, .1, .2, .4, .8\}$ kHz, for λ_2 at values $\{0.5, 1, 2, 4, 8\}$ ms, for $\mathbf{u}_1$ and $\mathbf{u}_2$ at values $\{.25, .5, 1, 2\}$, spatial smoothness lengthscale for $\mathbf{u}$ at values $\{.25, .5, .75, 1.0\}$ in units of the head diameter. All in all $5^2 4^2 = 400$ calibrations were tried with the Evidence framework. We dismissed all fits for which the resulting Hessian was not positive definite or the Infinity-norm of the gradient was not below 10^{-4}. The Evidence could then be used as a Bayesian model selector. With a log-Evidence value of -581.84 (against -596.35) the EEG-informed model was better than the purely behavioral model which assumed a constant t_0 and $v_{x,i}$ across trials.

Figure-1 shows the spatial filter **u** together with a map of electrode saliencies for each EEG decoder. The saliency is a measure of how important an electrode is for the decoder to predict the response, and we computed this using the second-order method of Hassibi and Stork (1993), see also Bishop (1995). The

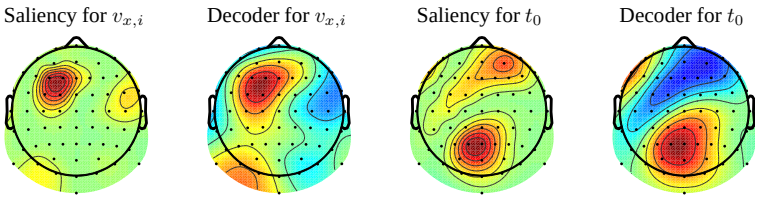

Fig. 1. Electrode saliency and spatial filter topographies of the two EEG decoders running in stereo to predict $v_{x,i}$ and t_0 from the EEG on a trial-by-trial basis. Baseline is green, positive is red, negative is blue.

figure shows that the method had produced two distinctly different decoders. Both decoders relied on frontal electrodes, but the decoder for t_0 relied on occipital electrodes as well. The frontal contribution to t_0 had reversely signed electrode weights compared to the occipital contribution which could indicate a co-modulated activity loop between the two areas with a phase shift greater than 90° — presumably a top-down modulation of the occipital cortex (e.g. Miller et al., 2012; Hillyard et al., 2004).

Figure-2 show histograms of the single-trial estimates of $v_{x,i}$ and t_0. According to this, the participant was primarily in a mode with $v_{x,i}$ values of about 50-60 Hz, but was able to boost categorization performance with increased values of $v_{x,i}$. The scaling parameter λ_1 was positive which indicated that such increase was predicted by an increase in projected EEG power. The exceedingly high bin in the histogram over $\hat{t}_0$ was due to extreme projected power values that were then suppressed by the logistic function (λ_2 was positive). However, the t_0 estimates that fell in the lower 80% quantile formed a well-behaved bump at around 15 ms with a trial-by-trial deviation of a few milliseconds. This order of deviation was in agreement with an average finding in a large population (Dyrholm et al., 2011). One possible mechanism behind the t_0 interval that follows the stimulus onset may be that attentional weights are computed according to (4) during this. It is therefore conceivable that the participant can benefit from lowering t_0 at the expense of an inaccurate weight computation. TVA predicts that an inaccurate weight computation would imply a lower $v_{x,i}$ due to distracting weights in the denominator of (3), hence we would expect $v_{x,i}$ and t_0 to be positively correlated. The single-trial estimates of $v_{x,i}$ and t_0 were indeed positively correlated (Pearson's $r = 0.44$, $p < 10^{-4}$) indicating a tradeoff between starting the competition early (low t_0) versus efficient categorization (high $v_{x,i}$). It would make sense for the participant to adjust such tradeoff based

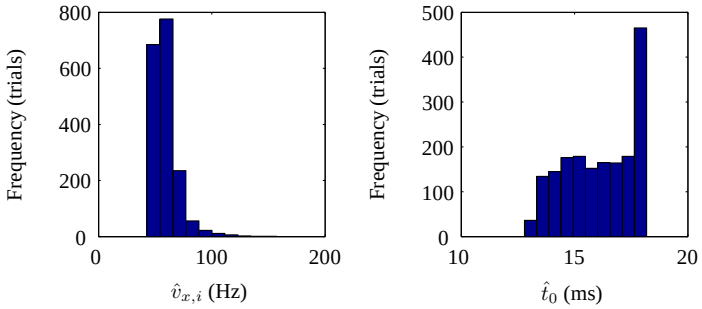

Fig. 2. Trial-by-trial estimation of $v_{x,i}$ and t_0 using the two EEG decoders.

on an expectation of whether the exposure duration of the coming stimulus is short (better with a low t_0) or long (better with a high $v_{x,i}$).

Figure-3 shows the observed behavioral performance (circles), the purely behavioral model prediction (curve), and the EEG-informed model prediction (+). The observed performance follows a more S-shaped pattern than the curve, as

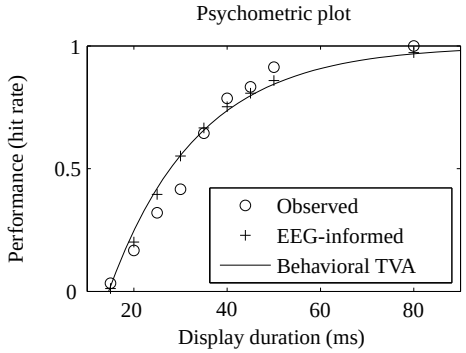

Fig. 3. Psychometric plot of behavioral performance as a function of display duration. The observations (circles) have a more S-shaped curve than the purely behavioral TVA model predicts (solid curve). The EEG-informed model (+) produces a slightly S-shaped prediction by incorporating the EEG pre-stimulus power.

has previously been shown to be a possible effect of trial-by-trial variation in t_0 (Dyrholm et al., 2011). The EEG-informed model provided a slightly closer fit to the observed performance; i.e. slightly S-shaped. The slight improvement of the EEG-informed system was realistic in the sense that we did not expect to completely predict the response variability based solely on pre-stimulus EEG activity. Future directions include the decoding of a wider time-range and using a wider range of EEG decoder functions.

Bibliography

Bishop, C. (1995). *Neural networks for pattern recognition*. Oxford university press, New York.

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., and Muller, K. (2008). Optimizing spatial filters for robust EEG single-trial analysis. *Signal Processing Magazine, IEEE*, 25(1):41–56.

Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, 97(4):523–547.

Bundesen, C., Habekost, T., and Kyllingsbæk, S. (2005). A neural theory of visual attention: Bridging cognition and neurophysiology. *Psychological Review*, 112(2):291–328.

Bundesen, C. and Harms, L. (1999). Single-letter recognition as a function of exposure duration. *Psychological Research*, 62:275–279.

Busch, N. A., Dubois, J., and VanRullen, R. (2009). The phase of ongoing EEG oscillations predicts visual perception. *J Neurosci*, 29(24):7869–76.

Delorme, A. and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1):9–21.

Dyrholm, M., Christoforou, C., and Parra, L. (2007). Bilinear discriminant component analysis. *Journal of Machine Learning Research*, 8:1097–1111.

Dyrholm, M., Kyllingsbæk, S., Espeseth, T., and Bundesen, C. (2011). Generalizing parametric models by introducing trial-by-trial parameter variability: The case of TVA. *Journal of Mathematical Psychology*, 55(6):416–429.

Hassibi, B. and Stork, D. G. (1993). Second order derivatives for network pruning: Optimal brain surgeon. *Advances in neural information processing systems*, pages 164–164.

Hillyard, S., Di Russo, F., and Martinez, A. (2004). The imaging of visual attention. *Functional neuroimaging of visual cognition (Attention and Performance XX). New York: Oxford University Press. p*, pages 381–388.

Jung, T., Humphries, C., Lee, T., Makeig, S., McKeown, M., Iragui, V., and Sejnowski, T. (1998). Extended ica removes artifacts from electroencephalographic recordings. In *Proceedings of the 1997 conference on Advances in neural information processing systems 10*, pages 894–900. MIT Press.

Koles, Z. (1991). The quantitative extraction and topographic mapping of the abnormal components in the clinical eeg. *Electroencephalography and clinical neurophysiology*, 79(6):440.

Kyllingsbæk, S. (2006). Modeling visual attention. *Behavior Research Methods*, 38:123–133.

MacKay, D. J. C. (1992a). Bayesian interpolation. *Neural Comput.*, 4:415–447.

MacKay, D. J. C. (1992b). The evidence framework applied to classification networks. *Neural Computation*, 4(5):720–736.

Miller, B. T., Vytlacil, J., Fegen, D., Pradhan, S., and DEsposito, M. (2012). The prefrontal cortex modulates category selectivity in human extrastriate cortex. *Journal of Cognitive Neuroscience*, 23(1):1–10.

Moody, J. (1992). The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In Moody, J. E. and Lippmann, R. P., editors, *Advances in Neural Information Systems*, volume 4, pages 847–854.

Mørch, N., Hansen, L. K., Strother, S. C., Svarer, C., Rottenberg, D. A., Lautrup, B., Savoy, R., and Paulson, O. B. (1997). Nonlinear versus linear models in functional neuroimaging: Learning curves and generalization crossover. In *Proceedings of the 15th international conference on information processing in medical imaging*, volume Springer Lecture Notes in Computer Science 1230, pages 259–270.

Parra, L. C., Spence, C. D., Gerson, A. D., and Sajda, P. (2005). Recipes for the linear analysis of EEG. *Neuroimage*, 28(2):326–41.

Penny, W. and Roberts, S. (1999). Bayesian neural networks for classification: how useful is the evidence framework? *Neural Networks*, 12(6):877 – 892.

Vangkilde, S., Coull, J. T., and Bundesen, C. (2012). Great expectations: Temporal expectation modulates perceptual processing speed. *Journal of Experimental Psychology: Human Perception and Performance*, Doi: 10.1037/a0026343.