Detecting the pain-evoked P300 in single trials: a comparison of template-based models

Master Thesis

Dominic Portain

Dr. Rob van der Lubbe (Department of Cognitive Psychology and Ergonomics)

Dr. ir. Jan R. Buitenweg (Department of Biomedical Signals and Systems)

Enschede, the Netherlands

University of Twente

Abstract

The enhanced Maximum Likelihood method (EML), proposed by Jaskowski and Verleger (2000), is a template-based approach for detecting signals in a continuous EEG measurement. This study compared the ability of three analysis techniques - peak picking, Woody's method and the EML - to determine amplitude and latency of a pain-evoked P3a component. A pseudo-real dataset was constructed from pain-related EEG data and deliberately contaminated with noise. There were two objectives to this research: first, to confirm the validity of our EML implementation; and second, to determine whether the EML is suitable to detect P3a amplitude habituation in single trials. 15 pseudo-simulated ERP sets were used for validation, and habituation was determined in four realistic sets of single trials. The original EML was modified twofold: first, a wavelet-transformed signal space was used instead of the fourier-transformed signal space; and second, the solver used the direct method instead of iterating. Both changes addressed issues (a lack of temporal resolution and oscillating results, respectively) that limited the original study. Comparison of the three techniques yielded results that were consistent with the prior study, validating our implementation. While the EML method proved effective to determine P3a latency in single trials, amplitude estimation was unreliable. The low accuracy of the EML results seems to be caused by the small signal-to-noise ratio. Pain habituation remained undetected in 53% of all subjects, unless single trials were binned in larger chunks. Because of the inherent limitations of template-based approaches, templates have to be tailored carefully when considering clinical applications.

Introduction

Event-related potentials (ERPs) have traditionally been studied under the assumption that each measurement contains a stationary part (stable in relation to the stimulus event) and a non-stationary part (varies independently from the stimulus event). From the assumption that ERPs remain stable over time if all environmental factors stay constant, the stationary part has been considered far more valuable than the non-stationary part. This assumption made it acceptable to sacrifice non-stationary information, in order to improve the accuracy of stationary information. Partly due to this paradigm, the widely used procedure has been to examine the average of several ERPs - instead of examining single trial ERPs. Event-related averaging has mainly been used to raise the - very low - signal-to-noise ratio (SNR) in single trials. However, continuing technological advances caused the physical noise to decline during the decades. This development is met by advancements in analysis algorithms that become more effective at separating the signal from background activity.

History of the EML

One of these algorithms, the Enhanced Maximum Likelihood (EML) method, will be the main subject of this study. Initially, Pham et al. proposed (Tuan, Möcks, Köhler and Gasser, 1987) and evaluated (Möcks, Köhler, Gasser and Tuan, 1988) a method for estimating the latency of ERP components in stimulus-locked trials. This method, initially called "iterative Fisher scoring", relies on two references: a *template* contained an idealized signal with the expected ERP component, and a *noise profile* provided an idealized representation of real-world background activity. Template-based models are considered more robust than feature-based models (which search for significant features such as lines, slopes or edges) when encountering strong sample variance. The principle of a template approach is based on a comparison between the template and a measured sample ERP. In addition to the signal, the sample also contains background noise:

$$r_j(t) = s(t) + e_j(t),$$

where t denotes the single time points of the analyzed waveshape (from 1 to T), $r_j(t)$ is the activity measured at time t, $s(t)$ is the constant event-related signal, and $e_j(t)$ is the noise signal. This simplified model assumes that the signal $s(t)$ is invariant between trials and uncorrelated to the noise set.

Woody's method. Woody's method (Woody 1967) uses statistical correlation to compare between sample and template (see figure 1). The time offset between template and sample is shifted until the correlation $\rho_{s,r}$ becomes maximal:

$$\rho_{s,r} = \frac{cov(s(t+\tau_j), r_j(t))}{\sigma_s \sigma_r}$$

Relying on a full-length template makes Woody's method very robust to noise contamination and localized artifacts (Jaskowski and Verleger, 1999). Smulders, Kenemans and Kok (1994) improved this method when finding that using covariation instead of correlation resulted in slightly increased performance.

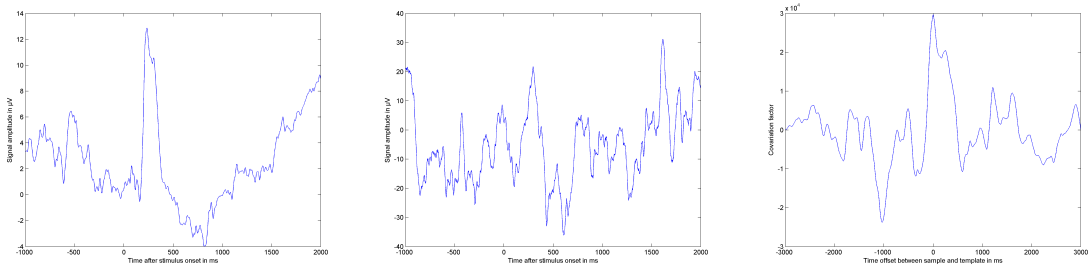

Figure 1. Illustration of Woody's method. A low-noise template (left) is cross-covariated against a noise-contaminated sample (middle). The cross-covariation is yielded as function of time offset between sample and template (right). Note that, in the central chart, the noise-induced peak at 1700ms yields a much smaller covariation value than the correct one at 280ms - despite its larger amplitude.

Maximum Likelihood method. Jaskowski and Verleger (1999) refered to the "iterative Fisher scoring" method by Tuan et al. as "Maximum Likelihood method" (ML), acknowledging the mathematical core of this method. The ML can be used to extract signal and noise components separately, as long as the simplified signal model applies and if there is

no overlap between the frequency distributions of signal and noise. Although the ML also relies on an idealized signal template, it employs a fundamental improvement over Woody's method (Jaskowski and Verleger, 2000): acknowledgment and usage of background noise as a data source. The simplified model is Fourier-transformed to fit the frequency-domain ML:

$$R_j(\omega) = S(\omega)^{-i\omega\tau_j} + E_j(\omega),$$

where $\omega = 0, (\frac{2\pi}{T}), (\frac{4\pi}{T}), ...$ describes a finite set of frequencies, and capital letters stand for Fourier transforms of time-domain data. There are two data sources available, one with the idealized signal $S(\omega)$ and one that describes the underlying noise sweep $N_j(\omega)$:

$$N_j(\omega) = (S(\omega)^{-i\omega\tau_j} - R_j(\omega))$$

The unknown latency τ_j can be estimated by maximizing[1] the log likelihood function LLF(for a complete description, see Tuan et al., 1987):

$$max_{\tau_j} \frac{N_j(\omega)}{LLF(S(\omega)^{-i\omega\tau_j} - R_j(\omega))}$$

Enhanced Maximum Likelihood method. The enhanced ML (EML) removes a limitation of the simplified signal model: the underlying signal is no longer amplitude-invariant. To allow for changing waveform amplitude between trials, Jaskowski and Verleger (1999) added an optimization variable to the previous model:

$$N_j(\omega) = (a_j S(\omega)^{-i\omega\tau_j} - R_j(\omega)),$$

[1] This maximization problem was originally solved by iterative Fisher scoring, providing the name for this procedure.

where a_j stands for the amplitude of the jth trial. Accordingly, the maximization problem uses a second parameter:

$$max_{\tau_j,a_j} \frac{N_j(\omega)}{LLF(a_j S(\omega)^{-i\omega\tau_j} - R_j(\omega))}$$

Jaskowski and Verleger (2000) reported severe problems when using the iterative Fisher solver, which allowed them to maximize both variables simultaneously. Iterations failed to converge to a stable result or converged in a local maximum. Preliminary analysis over our dataset indicated that there was no correlation between both parameters, allowing us to use a two-step direct approach:

$$max_{a_j} max_{\tau_j} \frac{N_j(\omega)}{LLF(a_j S(\omega)^{-i\omega\tau_j} - R_j(\omega))}$$

The original paper relied on Fourier-transformed data for the EML calculation, a compromise[2] between temporal resolution and frequency bandwidth (see Jaskowski and Verleger, 2000: "ML method"). This approach limited the analysis to the principal (low-frequency) components of the P3 waveform, and severely reduced the accuracy of the latency estimation results. To resolve this compromise and maximize EML accuracy and performance, we adapted the algorithm to accept wavelet-transformed data:

$$max_{a_j} max_{\tau_j} \sum_f \frac{log(I_{noise}(f))}{LLF(a_j s_f(t+\tau_j) - r_{f,j}(t))},$$

where f denotes a single frequency component in the wavelet-transformed domain. Phase information and trial dependency were removed from the noise sweep $N_j(\omega)$, resulting in a general noise intensity spectrum $I_{noise}(f)$. The logarithmic scaling of the noise profile was required to match the LLF output.

[2] Jaskowski and Verleger (2000) used a Fourier bandwidth of 10 frequency components, corresponding to a low-pass filtering with a cut-off frequency of 3.9Hz.

Previous study

Jaskowski and Verleger (2000) compared three distinct computational models in their performance to detect P3 peak latencies. All models were based on earlier studies, but had not yet been used in P3 detection. In addition to Woody's method and the EML, peak picking was also used in the performance comparison. Peak picking scans the ERP for the largest positive amplitude and returns the time position of this peak. The authors controlled P3 amplitudes, latencies and background noise levels using "pseudo-real" simulations (Gasser, Möcks and Verleger, 1983; Möcks, Gasser and Pham, 1984). Accuracy deteriorated strongly with increasing noise levels. The EML produced the most accurate results by a small margin. The authors also stress the limitations of a template-based approach, mainly due to the rigid signal model.

Study goals

The current paper is an update of the study performed by Jaskowsi and Verleger (2000). The first goal was to establish the validity of our EML implementation. While the previous study mainly focused on the determination of P3 latency, this paper extended the focus to include peak amplitude estimation (as suggested by Jaskowski and Verleger, 1999).

As result of more recent research efforts (introduced by Falkenstein, Hohnsbein and Hoormann, 1994; for an overview see Quevedo and Coghill, 2007; Polich, 2007), the P3 event was shown to consist of two independent components: the classic P3b, and the more physiologically-focused P3a. Pain stimuli are known to cause an unusually strong P3a amplitude. This effect is beneficial for the detection of small signals in single-trial analysis. In contrast to the traditional P3 production by oddball task (Donchin, Ritter and McCallum, 1987), the P3a creation is unrelated to the current task (Polich, 2003; Polich, 2007). Component onset is earlier and more intense when stimulus intensity is increased (Bromm and Scharein, 1981). Because of these features, we used pain-evoked P3a samples exclusively.

To create their pseudo-real simulations, Jaskowski and Verleger (2000) extracted the noise sweeps from a separate set of task-related ERP sweeps. Because no additional noise

measurements were required (which didn't exist in our data) we adapted this procedure to create noise sweeps.

The validation procedure consisted of two parts: the *artificial* condition and the *realistic* condition. Focus lay on the influence of noise contamination on the estimated results. The first - *artificial* - condition examined the models' performance under ideal conditions, using a pseudo-real dataset that contained only one prototypical P3a waveform (sampled from an illustration in Polich, 2007). The second part introduced ERP variation between trials. In the *realistic* condition, the pseudo-real dataset contained subject-specific grand-average ERPs.

The two traditional models (Peakpicking, Woody's method) were expected to perform similarly to those described by Jaskowski & Verleger (2000), showing a close link between noise contamination and model error rate. The two new models (ML, EML) were expected to yield more accurate results than the two traditional models, and to perform at least as well as the implementations described in the previous study.

The second goal was to test whether EML performance was sufficient to connect P3a amplitude to stimulus intensity. For this purpose, we prepared an ERP set that resulted from various amounts of painful stimulation. We wanted to see whether the EML amplitude estimations correctly detected the stimulation intensity.

Method

Data recording

The study included data from 16 subjects (students between 19 and 25 years) who produced 102 trials during each of four experimental blocks. The data for this study was created and provided by Blom and Braukmann. Participants received pain stimuli and responded according to stimulus intensity and location. Subcutaneous electrical stimulation on each forearm delivered the stimuli. Stimulus intensity was controlled by varying the number of impulses (spaced at 5ms) per pulse train. In the data for our first analysis, there were two intensity conditions, "high" and "low". We used only data from the "high" condition. The dataset for our second analysis contained responses to increasingly intense electrical stimuli. We used only trials that exceeded the subjects' pain threshold.

The measurement (BrainVision format, sampled at 500Hz) contained recordings from 63 EEG channels (electrode layout according to the 10-20 system), two EOG channels (measuring horizontal and vertical eye movements) and one ECG channel. In order to stay consistent with Jaskowski and Verleger (2000), the channel Cz was downsampled to 100Hz for the first dataset. Trials were extracted from an interval of 1s before stimulus onset and 2s after stimulus onset. Baseline-corrected trials (baseline sampling between -50ms and 0s) were then screened for artifacts by searching for blinks and values larger than $\pm 100\mu V$. Only data recorded from the Cz channel was used in this study. After screening, the first (second) dataset consisted of a total of 5176 (1809) trials.

Filter settings

Jaskowski and Verleger (2000) removed frequency components above 3.5Hz from their data. The original study supports this setting with three reasons: tradition, robustness of results (Smulders, Kenemans and Kok, 1994) and the low-frequency nature of the principal P3 portion. Peak picking and Woody's method yield the best results with pre-filtered data (e.g., Pfefferbaum, Christensen, Ford and Kopell, 1986; Pfefferbaum, Ford, Wenegrat, Roth and Kopell, 1984). Consistent with Jaskowski and Verleger (2000), data was subjected to

a zero-phase lowpass filtering (cutoff frequency of 3.5Hz) for the first analysis. However, the decision to use low-pass filtering before employing the EML was mainly caused by the resolution limitations of the Fourier transformation. Having removed this limitation, we decided to process the dataset in full temporal and frequency resolution during the second analysis.

When processed by the EML, data were wavelet-transformed into the lower 250 (50 in the first analysis) frequency components. Frequency values were assigned from a linear, evenly-spaced interval. Wavelet forms which follow the ERP peak shape provide maximum detection capabilities with minimal contrast artifacts at the peak boundaries (Mahmoodabadi, Alirezaie and Babyn, 2007). A symmetric, second-order reverse biorthogonal spline wavelet (Cohen, Daubechies and Feauveau, 1992; also used by Quiroga and Garcia, 2003) achieved the best fit[3] with a prototypical P3a event, and was used as filter for the continuous wavelet transformation.

Creation of the pseudo-real dataset

Template. The two template-based detection methods required a reference ERP sweep. For this purpose, a separate template was generated for each subject.

Template formation consisted of four steps: First, a random selection (25%) of all trials was separated from the remaining dataset. Second, the P3a latency difference $\hat{\tau}$ between each trial and a prototypical P3a waveform was estimated with Woody's method. Third, all trials were time-shifted by $-\tau$. Fourth, the event-locked samples were combined into a grand-average ERP.

Noise. The procedure to create noise sweeps (Jaskowski and Verleger, 2000) consisted of five steps. All noise-related trials were taken from a separate, random dataset selection (25% of all trials). (a) Woody's method was used to determine the P3a latency shift $\hat{\tau}$ between each sample $r(t)$ and the subject-specific grand-average ERP $s(t)$. (b) The template

[3]The similarity was measured by calculating the covariance between the wavelet shape and an idealized P3a waveform. The Matlab wavelet bior2.2 showed a covariance of 11.2% (relative to P3a autocovariance), followed by sym4 at 8.45% covariance.

was time-shifted by $-\hat{\tau}$ and multiplied by a variable factor $\hat{a}$. The factor $\hat{a}$ was varied until the sum of squared differences between sample $r(t)$ and scaled and shifted template $\hat{a}\,s(t-\hat{\tau})$ was minimal[4]. (c) A raw noise sweep $n_{raw}(t)$ was formed by subtracting $\hat{a}\,s(t-\hat{\tau})$ from $r(t)$. (d) The Fourier-transformed noise sweep $N(\omega)$ was split into phase and magnitude information. The magnitude channel determined the noise intensity spectrum and was unaltered. Randomization of the phase channel removed undesirable event-locked traces. (e) The inverse Fourier transformation of the combined channels created the final noise sweep $n(t)$.

Noise profiles were required as reference data for the EML. For this purpose, the magnitude channels from Fourier-transformed noise sweeps $N(\omega)$ were combined into a grand average. The resulting intensity spectrum $I_{noise}(f)$ was calculated for each subject, and subsequently used as noise profile.

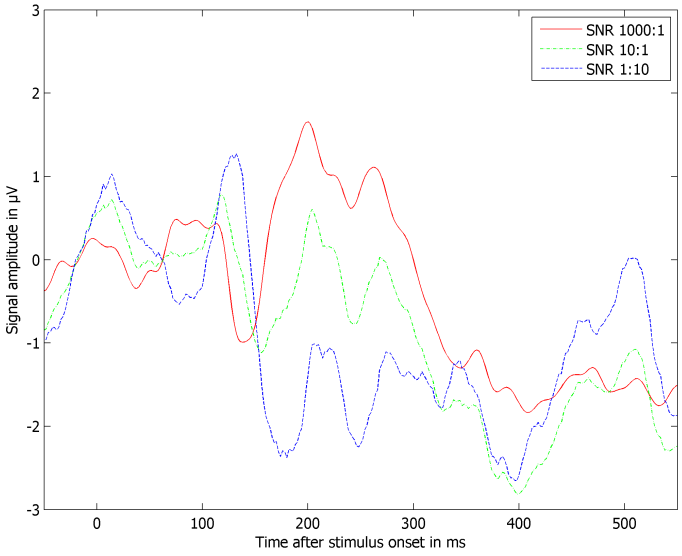

Figure 2. An examplary ERP, contaminated with three variations of the same noise sweep. The noise sweep was scaled by factors of 0.001 (red line), 0.1 (green line), and 10 (blue line). Note that the initially clear peak at 200ms shifts towards 130ms under increasing noise influence.

[4]In contrast to the original procedure, we used the interior-reflective Newton method to solve the minimization problem (Coleman, 1996). This large-scale approach, in comparison to the iterative Fisher solver used by Jaskowski and Verleger (2000), is less prone to produce oscillating results or to become trapped in local maxima.

Simulated trials. The first analysis used two distinct sample sets. The first sample "set" consisted of a grand average ERP of all trials in the dataset. The second sample set contained the 16 subject-specific templates (from the previous section, "Templates". Each sample $r_j(t)$ from both sample sets was contaminated with 13 levels of noise in two steps: (a) One noise sweep $n_j(t)$ scaled exponentially with a series of factors ν_i (i ranging between 1 and 13, ν_i ranging between -3.0 and +3.0). (b) The resulting series of noise sweeps was then combined with the sample (see figure 2) to create a series of trials $x_{ij}(t)$:

$$x_{ij}(t) = r_j(t) + 10^{\nu_i}\, n_j(t).$$

Because noise sweeps were only randomized between trials, not between noise levels, models could produce consistent results across all noise rates.

Methods for P3a estimation

Peak Picking. We used the approach described by Jaskowski and Verleger (2000) verbatim. Both sample and template are examined for the largest positive amplitude. The time difference between both peaks is defined as estimated P3a latency[5].

Woody's method. We implemented the approach by Jaskowski and Verleger (2000) unaltered. The time offset of the template is varied along the whole timerange, while the covariation between template and sample is calculated. The offset for which the covariation value is maximal was defined as estimated P3a latency.

EML. The implemented method consisted of three parts: the log likelihood function (LLF), the comparison with the noise profile, and the maximum search.

(a) The LLF (Tuan, Möcks, Köhler and Gasser, 1987) estimates which parameter is most likely to describe a given data field. A frequency component of both sample and template was selected. The two parts were time-shifted across the whole time range and

[5]Note that these methods yield a relative value. To determine the absolute P3a latency in a sample, exact knowledge about the template P3a latency is required.

subtracted from each other. The Matlab implementation MLE yielded a log likelihood [6] and a confidence interval. Only the log likelihood was used. This operation was conducted for each frequency component, resulting in a likelihood spectrum.

(b) The noise profile (from the previous subchapter, "Noise") was divided by the likelihood spectrum:

$$\sum_f \frac{log(I_{noise}(f))}{LLF(s_f(t+\tau_j) - r_f(t))}$$

Each frequency component was again processed separately. If the time offset τ_j was perfect, the signal component would be eliminated from the sample, only leaving noise. This would make the LLF minimal, yielding a maximum result for the formula. For any other time offset τ_j, a signal trace remains in the sample, rendering the LLF result larger than minimal.

(c) After establishing the conditions for finding the perfect time offset τ_j, the goal was to maximize the division between noise spectrum and LLF result. Including the two variables from before, we had to solve:

$$max_{a_j} max_{\tau_j} \sum_f \frac{log(I_{noise}(f))}{LLF(a_j s_f(t+\tau_j) - r_{f,j}(t))}$$

A direct approach was used, varying the values τ_j and a_j one after another. Range and resolution of τ_j were straightforward: the length of the whole sample, one datapoint at a time. The first step varied only τ_j, leaving a_j at a value of 1. The τ_j that yielded the maximum result was considered final. The factor a_j was then varied between $10^{-1}$and 10^{1} in 100 linear steps, while searching for a maximum result with the previous τ_j. The resulting optimal a_j was then used for a third, more accurate search around $a_j \pm 0.1$ in 39 steps. This final a_j and the previous τ_j then defined the estimated P3a latency and amplitude.

Statistics

Two sets of statistics were used to evaluate the validity and performance of the EML, respectively. (a) The three computational models compared samples from two simulated

[6] Specifically, the Matlab MLE calculates the likelihood of the sample to be non-zero.

datasets, against the respective templates. Yielded were the mean squared errors between template-template ("expected") and template-sample ("estimated") comparisons, as defined by the formula:

$$MSE = \frac{1}{N}\sum_{j}(\tau_j - \hat{\tau}_j)^2$$

where τ_j are the expected and $\hat{\tau}_j$ the estimated latencies. The average MSE over all subjects was then compared between methods and against the results of the previous study. (b) The EML compared single-trial data from all subjects to a set of subject-specific templates. Results were subjected to bivariate correlation between stimulus intensity and estimated P3a amplitude.

Results

First analysis: Validity of the improved EML

During the first analysis, accuracy of the three algorithms (peak picking, Woody's method and EML) was compared to the results of Jaskowski and Verleger (2000). Figure 3 (below) displays the effect of noise contamination on the mean squared error of latency estimations. Estimation error from using the *artificial* dataset is displayed in the left-hand chart. Estimation error from using the *realistic* dataset is presented in the right-hand chart.

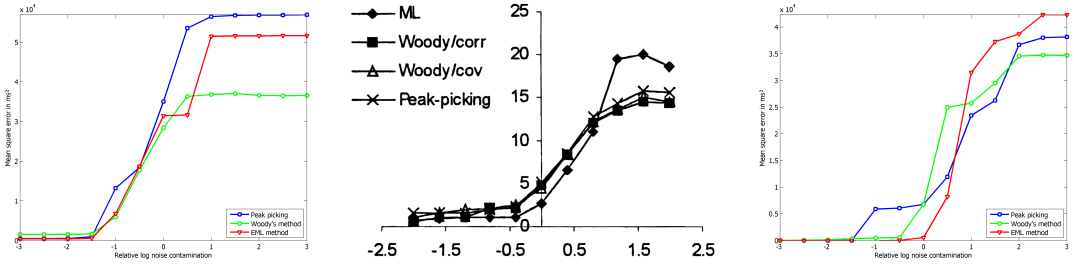

Figure 3. Error rate of P3a latency estimations from three analyses, as a function of noise contamination. Note that the X-axis describes an exponential series. Left: results from our implementations in the *artificial* condition. Middle: original results from Jaskowski and Verleger (2000) in the 70ms-condition. Right: results from our implementation in the *realistic* condition.

The *artificial* condition directly reflected on the noise-robustness of the implemented algorithms. Results were accurate in all methods during low-noise conditions. Performance started to degrade at a ν of -1.5 (equivalent to a SNR of 30). Error rates increased equally for all three models, until leveling off at ν-values between +0.5 and +1.0. In high-noise conditions, models unanimously detected a P3a-similar complex in the noise sweep. Strong differences between estimation errors in high-noise conditions mainly result from the low number of random noise sweeps and don't reflect on the performance of the algorithm.

The *realistic* condition was designed to be compared to results from Jaskowski and Verleger (2000). Results were near-perfect in the low-noise region, similar to both the previous condition and the previous study. Degradation started at ν-values between -1.0 and 0.0, which is more diverse than in the previous study (degradation begin at a ν of -0.5). Results from peak picking started to degrade first. Both peak picking and Woody's method started degrading earlier than the EML, but the rate of degradation was lower. Due to this effect, the EML provided the most accurate results up to a ν-value of +0.5. For ν-values of +1.0 and above, the EML yielded the least accurate results for all three models. This result is consistent with the findings of the previous study: the EML performed best at a ν of +0.66 and below, worst at a ν of +1.0 and above. In high-noise conditions, peak picking performs slightly worse than Woody's method but better than the EML. Except for the high-noise EML error rate - which was even higher in the previous study - our results are very similar to those of Jaskowski and Verleger (2000).

Second analysis: Connection between pain intensity and detected P3a amplitude

In this step, the EML estimated P3a peak amplitudes from the second dataset. Estimations from single trials were then compared to stimulus intensity. We removed 16.9% of the trials, for which the estimated P3a latency was outside of the time window 100-400ms (relative to stimulus onset). Because both P3a amplitude and stimulus intensities varied strongly between individuals, these variables were standardized per subject (see figure 4). Also, the mean amplitude was removed from individual P3 amplitudes.

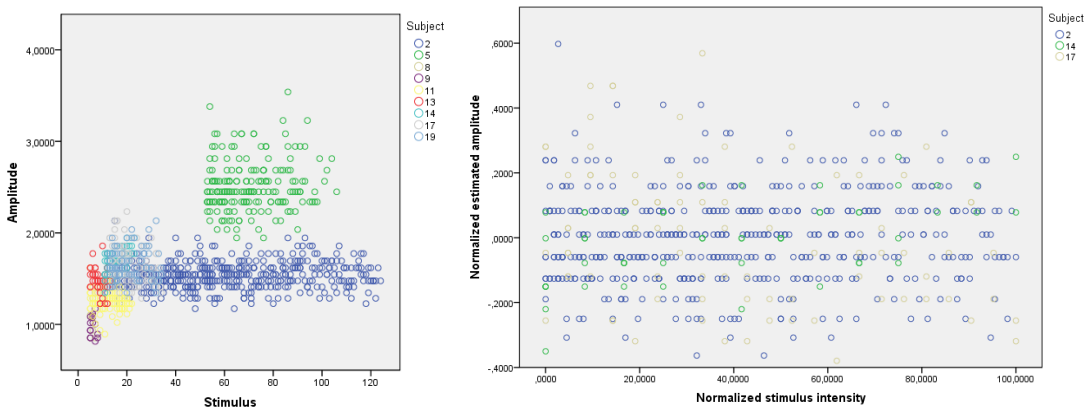

Figure 4. Comparison between stimulus intensity and EML results. Left: Single trials from 9 subjects. Estimated amplitudes are plotted against stimulus intensity. Right: Selection of subjects from the normalized data set; amplitudes are again plotted against stimulus intensity. Note: Apparent grouping along the horizontal axis is an artifact from low rounding precision. The correlation analysis is not affected.

The amplitude estimations show large variance between trials. When comparing amplitude estimations against stimulus intensity, no clear pattern or inclination emerges. The bivariate correlation between the two variables doesn't yield significant results either ($r = .038$, $p(1045) = .225$). Post-hoc analysis on a per-subject basis indicates strong individual differences: two subjects reached significance ($r = .208$, $p(118) = .024$; and $r = .194$, $p(138) = .023$) with opposite correlation coefficients.

Discussion

An algorithm was examined for its ability to estimate latency and amplitude in pain-related P3a events. The level of background noise was varied systematically using "pseudo-real" simulations. The validity of our implementation was confirmed by comparing the results against those of a previous paper. P3a amplitudes were then estimated in a real dataset, failing to detect a connection between stimulus intensity and P3a amplitude.

The EML indicated to produce correct results, when compared to the similar study by Jaskowski and Verleger (2000). Contrary to the findings during the second analysis, P3a amplitude is generally linked to stimulus intensity (Lorentz and Bromm, 1997) - even in similar settings (Bromm and Scharein, 1991; Kanda, Fujiwara, Xu et al., 1996; Houlihan, McGrath, Connolly et al, 2004). To resolve this contradiction, a number of speculations can be offered: (a) The range of stimulus intensity was too small to produce a distinct difference in P3a amplitude; (b) the amplitude estimation had been implemented inaccurately; or (c), the noise profile provided insufficient grounds for separating signal from background activity.

Subjects in the study of Blom and Braukmann received stimuli that increased from indiscernible to the limit of pain tolerance, which refutes the first speculation. To examine the second speculation, we employed a post-hoc analysis. Trials from the first dataset were organized in a series of sequential trials, where stimulus intensity varied only between two conditions. According to Bromm and Treede (1991), repeated exposure to pain stimuli decreases the connected ERP amplitudes. We extracted the original single trials from the high-intensity condition and analyzed the influence of repeated stimulation on estimated P3a amplitudes. Amplitude estimations of single trials from 15 subjects were subjected to (one-tailed) bivariate correlation against trial number. References (subject-specific templates and noise profiles) were re-used from the simulated dataset. The correlation (see figure 5) yielded a significant decrease over time ($r=-0.061$, $p < 0.001$), showing the effectiveness of the EML in principle. However, this correlation was based on the complete data set. When the data set was examined per subject - as would be the case in a clinical setting (Turk and Dworkin, 2004) - the correlation reached significance only in 7 subjects (equivalent to a detection rate

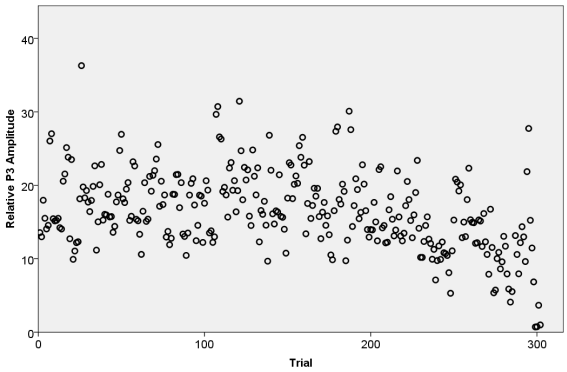

Figure 5. P3a amplitudes from a series of repeated, invariant stimulation. Chart represents standardized and binned data from all subjects.

of 47%).

This finding provides evidence for the third speculation, indicating that EML performance was limited by the high level of background noise in combination with a low amount of trials.

The vulnerability to low signal-to-noise ratios can be explained by a fundamental limitation of template-based approaches. As ERPs consist of the added effect of a multitude of neuronal clusters, the waveform will inevitably vary strongly between trials - and even more so between subjects. In order to optimize template-based estimations, the template ERP should be as personalized and as noise-free as possible. This poses the first practical limitation to the application of pain quantification - a large amount of trials is neccessary just to prepare the template and noise profile. Employing a more sophisticated de-noising algorithm (e.g., Quiroga and Garcia, 2003) for each trial could potentially decrease the amount of necessary trials. Another potential remedy to this problem is provided by binning several trials to a less noisy average before employing the amplitude estimation (in our dataset, we reached a 100% detection rate for habituation when combining 4 trials).

The major reason for the poor amplitude results could be the paradigm of a fixed ERP template. The different evoked components contain different latencies and intensities, before being combined to a ERP measurement. Optimally, a computational model would be able to separate, recognize and quantify the different components which add up to the

specific response that appears to be a single positive peak. If the template approach is desirable (either in terms of computational effort or robust algorithms), its accuracy potentially could be enhanced by performing a ICA (independent component analysis) which provides localized information to the template. With this compromise, the rigidity problem of the general template is resolved, without having to resort to more complex algorithms. And as the results from the EML estimation have shown, optimization can improve the original algorithm by orders of magnitude - even if the optimization is implemented as final refinement. From this point of view, it would seem beneficial to the final accuracy to vary the components in every possible way - not only amplitude, but also latency, length or even DC offset and frequency baseline.

Of course, each extension to current algorithms would have to be implemented very carefully: the exponentially increasing capacity demands quickly pose the need for the use of iterative solvers and search guidelines. Each decision that is made with the final result in mind will introduce a hypothesis into the analysis, and promote trials that conform to the original assumption. The risk of committing a self-fullfilling prophecy can't be prevented easily. Ideally, the analysis would be guided by a flexible algorithm that knows about the general conditions of the experiment (e.g., that two trials are always approximately 4 seconds apart) and can adapt to the individual connection between stimulus and ERP components.Neuronal networks (Gallistel, Mark, King and Latham, 2002), for example, could employ a combination of feedforward and feedback learning to excel at this task. After the initial phase of trial-and-error learning, such a model could emulate future results and adapt from the error of prediction (feedforward model). The current top candidate in the area of unsupervised learning is regularly used in computer science and uses a less limited approach: Each degree of freedom is represented as a separate dimension in a high-dimensional vector field, and every possible variation is potentially represented in it. The approach of support vector machines (*SVM*; Ranzato, Huang, Boureau and LeCun, 2007) describes an iterative process with the goal to find an optimal separation between two sets of values (e.g., signal and noise) within the vector field. Employing this paradigm

in P3 detection would combine the positive abilities of all above models (highly adaptive, maximum input granularity, blindly improving process) without many of the drawbacks (templates: rigidity; neuronal networks: scale problems). However, this luxury comes at the cost of very high needs for computational capacity - a common limiting factor in cognitive science.

References

Bromm, B., & Scharein, E. (1982). Principial component analysis of pain-related cerebral potentials to mechanical and electrical stimulation in man. *Electroencephalography and Clinical Neurophysiology*, *53*, 94-103.

Bromm, B., & Treede, R. D. (1991). Laser-evoked cerebral potentials in the assessment of cutaneous pain sensitivity in normal subjects and patients. *Revue neurologique*, *147*, 625-643.

Coleman, T. F., & Li, Y. (1996). A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables. *SIAM Journal on Optimization*, *6*, 1040-1058.

Falkenstein, M., Hohnsbein, J., & Hoormann, J. (1994). Effects of choice complexity on different subcomponents of the late positive complex of the event-related potential. *Electroencephalography and clinical neurophysiology*(2), 148-60.

Gallistel, C. R., Mark, T. A., King, A., & Latham, P. (2002). A test of gibbon's feedforward model of matching. *Learning and Motivation*, *33*, 46-62. Available from `http://bit.ly/af3Djk`

Houlihan, M. E., McGrath, P. J., Connolly, J. F., Stroink, G., Finley, G. A., Dick, B., et al. (2004). Assessing the effect of pain on demands for attentional resources using ERPs. *International Journal of Psychophysiology*, *51*, 181-187.

Jaskowski, P., & Verleger, R. (1999). Amplitudes and latencies of single-trial ERP's estimated by a maximum-likelihood method. *IEEE Transactions on Biomedical Engineering*, *46*, 987.

Jaskowski, P., & Verleger, R. (2000). An evaluation of methods for single-trial estimation of p3 latency. *Psychophysiology*, *37*, 153-162.

Mahmoodabadi, S. Z., Alirezaie, J., & Babyn, P. (2007). Bio-signal characteristics detection utilizing frequency ordered wavelet packets. In *International symposium on signal processing and information technology* (p. 748-753).

Polich, J. (2003). *Theoretical overview of p3a and p3b*. Kluwer Academic Publishers.

Polich, J. (2007). Updating p300: an integrative theory of p3a and p3b. *Clinical Neurophysiology*, *118*, 2128-2148.

Quevedo, A. S., & Coghill, R. C. (2007). Attentional modulation of spatial integration of pain: evidence for dynamic spatial tuning. *Journal of Neuroscience*, *27*, 11635.

Ranzato, M. A., Huang, F. J., Boureau, Y. L., & LeCun, Y. (2007). Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *Computer vision and pattern recognition* (p. 1-8).

Smulders, F. T. Y., Kenemans, J. L., & Kok, A. (1994). A comparison of different methods for estimating single-trial p300 latencies. *Electroencephalography and Clinical Neurophysiology*, *92*, 107-114.

Tuan, P. D., Möcks, J., Köhler, W., & Gasser, T. (1987). Variable latencies of noisy signals: estimation and testing in brain potential data. *Biometrika*, *74*, 525.

Turk, D. C., & Dworkin, R. H. (2004). What should be the core outcomes in chronic pain clinical trials? *Arthritis Research and Therapy*, *6*, 151-173.

Xu, X., Kanda, M., Shindo, K., Fujiwara, N., Nagamine, T., Ikeda, A., et al. (1995). Pain-related somatosensory evoked potentials following CO2 laser stimulation of foot in man. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, *96*, 12-23.