Fitting a Computational Model of Perceptual Inference to Principal Component Weights of ERP Responses

Lukas Vogelsang[1−2⋆] **(lukasvog@student.ethz.ch)**
Lilian Weber[2⋆] **(weber@biomed.ee.ethz.ch)**
Sara Tomiello[2] **(sara.tomiello@biomed.ee.ethz.ch)**
Dario Schöbi[2] **(dschoebi@biomed.ee.ethz.ch)**
Katharina V. Wellstein[2] **(wellstein@biomed.ee.ethz.ch)**
Sandra Iglesias[2] **(iglesias@biomed.ee.ethz.ch)**
Klaas Enno Stephan[2−4] **(stephan@biomed.ee.ethz.ch)**

[1]Institute of Neuroinformatics, University of Zurich & ETH Zurich, Zurich, Switzerland
[2]Translational Neuromodeling Unit, University of Zurich & ETH Zurich, Zurich, Switzerland
[3]Wellcome Centre for Human Neuroimaging, University College London, London, United Kingdom
[4]Max Planck Institute for Metabolism Research, Cologne, Germany
*These authors contributed equally to this work

Abstract

The mismatch negativity (MMN), a well-studied electrophysiological response to irregularities in the sensory input stream, has often been used to examine how the brain learns the statistics of its environment. This response has also been found to be systematically altered in clinical populations such as patients with schizophrenia. These deviations in electrophysiology, however, cannot easily be linked to inter-individual differences in cognitive processing style due to the lack of direct behavioral readouts, which limits the paradigm's usefulness for cognitive science and computational psychiatry. To bridge this gap, we present a pipeline for inferring parameters of a generative model of learning and inference, the Hierarchical Gaussian Filter (HGF), given EEG recordings obtained as part of the auditory MMN paradigm. Our pipeline includes a data-driven feature selection step as well as a proposal for mapping belief updates to the EEG features.

Keywords: mismatch negativity; perceptual inference; hierarchical gaussian filter; computational psychiatry

Introduction

As part of our daily lives, we are constantly immersed in noisy sensory information from the outside world and are required to infer upon the hidden state of the environment. The study of perceptual inference and the integration of sensory information has a long and rich history. On the cognitive level, there is a variety of models that, given experimentally acquired readouts of behavior, allow for inference upon inter-individual differences in information processing. On the physiological level, neuroimaging studies have proven useful for identifying neural signatures of perceptual inference.

A striking example is the mismatch negativity (MMN), an electrophysiological response typically observed as part of an oddball paradigm, where a sequence of identical stimuli ('standards') is eventually interrupted by a stimulus differing in one of the stimulus dimensions ('deviant'). These stimuli are usually either auditory or visual. Here, we focused on the auditory MMN. The observed deviant-induced increase in negativity in the event-related potential (ERP) between 100 and 250 ms after stimulus onset constitutes the auditory MMN and is typically assessed by subtracting the average ERPs to standard stimuli from the average ERPs to deviant stimuli.

Since its discovery by Näätänen et al. (1978), the MMN has often been used to illustrate that humans learn the statistics of their environment (Winkler, 2007) and that deviations in the MMN response can be indicative of certain pathologies. For example, patients with schizophrenia are characterized by weaker MMN amplitudes (Avissar et al., 2018; Erickson et al., 2016). Schizophrenia has also been associated with aberrant NMDA-receptor function (Stephan et al., 2009; Friston et al., 2016) which, in turn, has been proposed to relate to prediction error signaling in predictive coding frameworks related to Bayesian inference (Friston, 2005).

While the MMN is well-studied, robust, and appears to be clinically relevant, inter-individual differences in the expression of this feature cannot readily be related to inter-individual differences in cognitive processing style in the absence of direct behavioral readouts. To bridge this gap, we here seek to relate EEG recordings acquired as part of an auditory MMN paradigm to a generative model of learning and inference, the Hierarchical Gaussian Filter (HGF) (Mathys et al., 2011). Put simply, we propose an approach that allows for fitting the HGF directly to electrophysiological (as opposed to behavioral) data. The key challenges in this approach are the high level of noise in the electrophysiological data and its high dimensionality, rendering a direct mapping from beliefs of our cognitive model to observable EEG responses challenging. We therefore first engaged in data-driven feature selection based on multilinear principal component analysis (MPCA) and subsequently constructed a response model for the HGF that generates trial-wise predictions of principal component weights for recorded electrophysiological responses.

Methods & Results

Subjects

We draw on data from a total of 81 healthy volunteers from a previous pharmacological EEG study which examined the effect of dopaminergic and cholinergic antagonism on MMN expression (Weber et al., in prep.). Here, participants received either amisulpride, biperiden or placebo in a between-subjects, double-blind design.

Experimental paradigm

Participants engaged in a new variant of the roving auditory MMN paradigm, introducing different levels of volatility to the underlying statistical structure of the tone sequence. In the typical auditory MMN paradigm, following a sequence of repeating sounds, a deviant sound is presented that differs in a certain stimulus dimension (typically in pitch). In the roving MMN paradigm, this initially odd sound is being repeated and, over time, becomes the new standard. In the volatile variant employed here, the probability governing those sound flips changes over time such that, over the $n = 1800$ trials, some periods are more stable than others (see Figure 1).

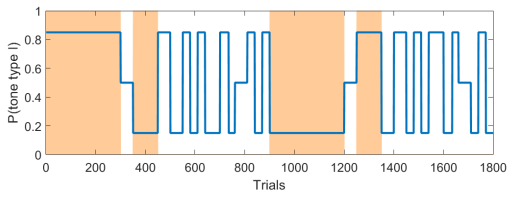

Figure 1: The MMN stimulus sequence visualized. The blue line indicates the probability of a high-pitched vs. low-pitched tone occurring. The orange area indicates stable phases, where for at least 100 trials no rule change occurred.

Data acquisition and preprocessing

Electrophysiological data were recorded using a 64-channel EEG cap (EASYCAP, BrainProducts). EEG analysis was performed using SPM12 and MATLAB. The continuous EEG signal was re-referenced to the average and filtered as follows: First, a high-pass Butterworth filter with a cutoff frequency of 0.1 Hz was applied. Next, the data were down-sampled to 250 Hz. Finally, a low-pass Butterworth filter with a cutoff frequency of 30 Hz was employed. The EEG data were epoched into 550 ms long segments, beginning 100 ms prior to stimulus onset. Epochs were baseline-corrected. Following the rejection of subjects with poor signal quality, EEG data of 72 subjects were further analyzed.

Feature selection

An MPCA-based approach To reduce data dimensionality while retaining as much of the rich signal as possible, we engaged in data-driven feature selection based on multilinear principal component analysis (MPCA) (Kroonenberg & De Leeuw, 1980), an extension of PCA that allows processing matrices with more than two dimensions.

We defined the data of a single subject as containing $n = 1800$ measurements (trials) of $(m \times k)$ 'two-dimensional' samples, with $m = 139$ time points and $n = 63$ EEG sensors. MPCA was carried out using the implementation of Lu et al. (2008). Given a desired number of linearly uncorrelated temporal $(l_1 < m)$ and spatial $(l_2 < k)$ components, MPCA decomposes the data into a weight matrix, W (of size $n \times l_1 \times l_2$), and two coefficient matrices: T, of size $l_1 \times m$, for the temporal components, and S, of size $l_2 \times k$, for the spatial components. As W is three-dimensional, a joint reconstruction in space and time can be carried out.

After having successfully validated the consistency of principal components across subjects, we carried out a single MPCA on a three-dimensional data matrix, where subjects' recordings, each of size $n \times m \times k$, are stacked along the first dimension to form an input matrix $X \in \mathbb{R}^{(n \times s) \times m \times k}$, where s represents the number of subjects. We therefore interpreted data from different subjects as additional observations of the same variables. This leads to all subjects sharing the same components and differing only in their component weights.

The first ten temporal and spatial principal components obtained through this MPCA are shown in Figures 2 and 3.

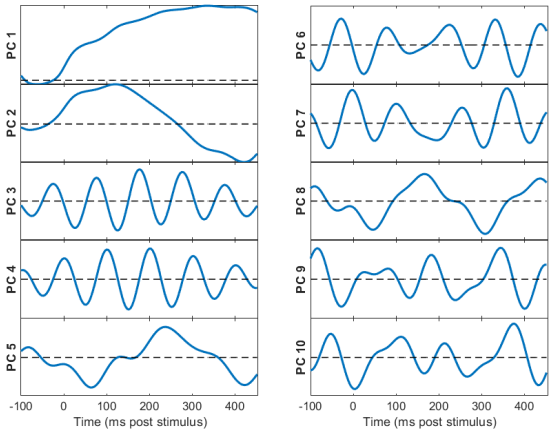

Figure 2: Top-10 temporal components for stacked MPCA

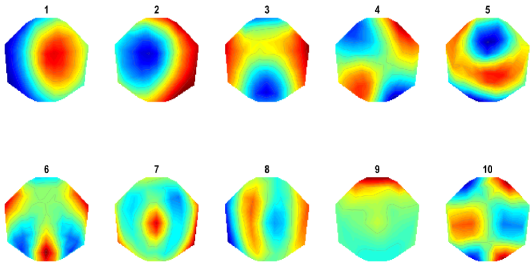

Figure 3: Top-10 spatial components for stacked MPCA

Component selection In order to select the principal components for the subsequent modelling step, we assessed how much of the relevant signal inherent in the original data is maintained when projecting the MPCA weights back to their original space using only a small subset of spatial and temporal components. We defined the relevant 'signal' as the amplitude of the negativity in the typical MMN time window at a sensor which typically shows strong MMN effects. To quantify the signal-to-noise ratio (SNR) both in the original and in the MPCA-reduced data, we assessed the maximum negative voltage within a time window of +100 to +248 ms (relative to the average voltage across all other time points of the measurement) at sensor Fz for each trial, and compared the distribution of measures for all standard vs. deviant trials using Welch's t-test statistic for unequal variances.

Beginning with the frontal electrode Fz and using the approach outlined above, we assessed the SNR of our signal when using only a single spatial and temporal component, relative to that of the original EEG data. We found that, for most subjects, the MPCA reconstruction using the combination of spatial component 1 and temporal component 8 resulted in higher scores than the raw data (see Figure 4), whereas for all other combinations, the opposite was the case.

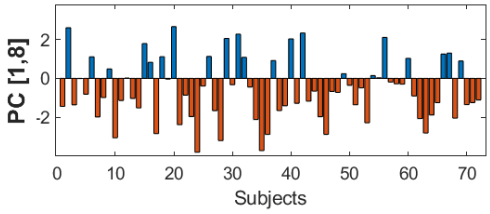

Figure 4: Difference in t-values expressing SNR of partially reconstructed data using spatial component 1 and temporal component 8, compared to the original EEG data, per participant. Relative negative scores (red) represent an improvement in SNR, positive scores (blue) a reduction.

This combination of spatial and temporal component also turned out to be the winning feature when (i) instead of the maximal negative voltage within the specified time window, the area under the curve was measured; when (ii) data from different sensors (Fz, FCz, and Oz) were assessed; and when (iii) the component was chosen with the strongest correlation between trial-by-trial component weights and the trial-by-trial precision-weighted prediction error trajectories of a Bayes-optimal agent under the HGF.

Modeling

The perceptual model We hypothesized that subjects exposed to the MMN tone sequence would infer upon two quantities: (i) the probability of hearing a high-pitched vs. low-pitched tone and (ii) the volatility of this probability, or how quickly this quantity is currently changing. As perceptual model, we therefore chose the 3-level HGF for binary inputs (Mathys et al., 2011), where the second level describes the belief about the current tendency towards a tone category (high-pitch vs. low-pitch), i.e., the current regularity in the tone sequence, and the third level captures the agent's belief about environmental volatility.

In the HGF, the agent's beliefs about both of these quantities are updated on each trial k in response to the tone input, where the magnitude of the update of the posterior mean is proportional to a precision-weighted prediction error:

$$\Delta\mu_i^{(k)} \propto \frac{\hat{\pi}_{i-1}^{(k)}}{\pi_i^{(k)}}\delta_{i-1}^{(k)}, \tag{1}$$

where $\Delta\mu_i^{(k)}$ denotes the change in the posterior mean at level i of the HGF hierarchy, and where $\delta_{i-1}^{(k)}$ represents the prediction error measuring the difference between the prediction made for, and the actual input observed at, the level below the current one. This prediction error is then weighed by a ratio of precisions, such that more precise predictions about the level below the current one $(\hat{\pi}_{i-1}^{(k)})$ lead to stronger updates, while more precise beliefs at the current level $(\pi_i^{(k)})$ are less impacted by PE-driven updates.

On the lower hierarchical level, the mean μ_2 of the belief about the regularity is updated according to Equation 2.

$$\mu_2^{(k)} = \hat{\mu}_2^{(k)} + \sigma_2^{(k)} * \delta_1^{(k)}, \tag{2}$$

where δ_1 represents the prediction error about the stimulus and σ_2 is the uncertainty (inverse precision) about the current estimate of μ_2. Together, they form the lower-level precision-weighted PE ε_2.

The HGF has a number of perceptual parameters that describe the individual cognitive processing style of an agent and determine the exact belief updates. For the full model specification including all parameters, please refer to Mathys et al. (2011). Here, we were particularly interested in the tonic learning rate ω_2, which quantifies an agent's general willingness to update her beliefs irrespective of (or in addition to) her current estimate of volatility. We thus focused on inferring this parameter from the EEG data. In the following, we always fixed all other perceptual parameters to their Bayes optimal value, which is the value resulting in the least surprise over the whole input sequence.

The response model Given the MMN input sequence and a choice of parameter values for the perceptual model, the HGF generates trial-by-trial trajectories of beliefs. Based on those trajectories, we designed a response model to generate trial-by-trial EEG responses $y^{(k)}$ – here, corresponding to the trial-by-trial weights of the chosen principal components. As the MMN has frequently been viewed as the manifestation of a precision-weighted prediction error (Lieder et al., 2013), we sought to generate our responses (see Equation 3) based on the lower-level prediction error that is irrespective of a particular stimulus category (the absolute of δ_1), weighted by its precision weight (σ_2), as shown in Equation 3.

$$y^{(k)} = \beta_0 + \beta_1\ logit(|\delta_1^{(k)}|)\sigma_2^{(k)} + \eta, \quad \text{where } \eta \sim \mathcal{N}(0,\zeta), \tag{3}$$

Here, the response model parameter β_0 can be understood as the offset, β_1 as the effect of our logit-transformed and precision-weighted prediction error, and η models Gaussian noise with zero mean. The logit transform (see Equation 4) was applied to map the (absolute) prediction errors, which, due to the binary nature of the inputs, lie between 0 to 1, to a range from minus to plus infinity.

$$\text{logit}(x) = \log(\frac{x}{1-x}) \tag{4}$$

Parameter recovery Examining the practical identifiability of parameters with our model, we found that the recovery of a single parameter in isolation is unproblematic but that β_1 and ω_2 are highly correlated and that high levels of noise ($\zeta \geq 2000$) pose problems for a successful parameter recovery. In particular, in those noise regimes, only parameter sets including reasonably high values of ω_2 and β_1 can be recovered reliably.

Fitting the data Given the difficulty of simultaneously estimating β_1 and ω_2, we started by fixing β_1 and estimating β_0, ω_2 as well as ζ. Figure 5 shows the results for fitting ω_2, β_0, and ζ while fixing $\beta_1 = -5$. We see that, for all subjects, the fitting procedure succeeded and the posterior means differed, for all three parameters (β_0, ω_2, and ζ), from the corresponding priors. We next systematically varied β_1 and found internal consistency in the ω_2 estimates across the range of β_1-values that, in a previous parameter recovery, were found to be identifiable.

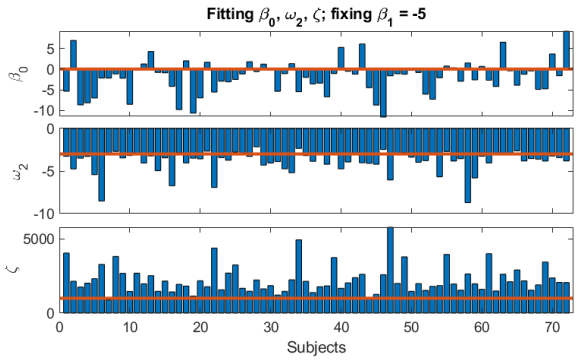

Figure 5: Fitting β_1, ω_2, and ζ, with β_1 = -5. The prior means for (β_0, ω_2, ζ) were set to (0, -3, 500); their corresponding variances to ($4, 4^4$, 4). The red lines correspond to the prior means; the y-axes indicate the parameter values.

Discussion

Overall, while our results represent a first step towards successfully estimating perceptual parameters of the HGF based on EEG recordings, the exceedingly high estimates for our noise parameter as well as the high parameter correlation between β_1 and ω_2 currently still pose restrictions on the interpretation of our results. For future applications of this pipeline, it will be important to solve these issues and potentially explore other mappings from HGF quantities to EEG features. In the future, we hope that our approach will increase the utility of the MMN paradigm for cognitive neuroscience and computational psychiatry by helping to reveal the nature and time course of perceptual inference in schizophrenia and other illnesses.

Acknowledgments

This study was supported by the University of Zurich and the René and Susanne Braginsky Foundation.

References

Avissar, M., Xie, S., Vail, B., Lopez-Calderon, J., Wang, Y., & Javitt, D. C. (2018). Meta-analysis of mismatch negativity to simple versus complex deviants in schizophrenia. *Schizophrenia research*, *191*, 25–34.

Erickson, M. A., Ruffle, A., & Gold, J. M. (2016). A meta-analysis of mismatch negativity in schizophrenia: from clinical risk to disease specificity and progression. *Biological psychiatry*, *79*(12), 980–987.

Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, *360*(1456), 815–836.

Friston, K., Brown, H. R., Siemerkus, J., & Stephan, K. E. (2016). The dysconnection hypothesis (2016). *Schizophrenia research*, *176*(2-3), 83–94.

Kroonenberg, P. M., & De Leeuw, J. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, *45*(1), 69–97.

Lieder, F., Stephan, K. E., Daunizeau, J., Garrido, M. I., & Friston, K. J. (2013). A neurocomputational model of the mismatch negativity. *PLoS computational biology*, *9*(11), e1003288.

Lu, H., Plataniotis, K. N., & Venetsanopoulos, A. N. (2008). Mpca: Multilinear principal component analysis of tensor objects. *IEEE transactions on Neural Networks*, *19*(1), 18–39.

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A bayesian foundation for individual learning under uncertainty. *Frontiers in human neuroscience*, *5*, 39.

Näätänen, R., Gaillard, A. W., & Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta psychologica*, *42*(4), 313–329.

Stephan, K. E., Friston, K. J., & Frith, C. D. (2009). Dysconnection in schizophrenia: from abnormal synaptic plasticity to failures of self-monitoring. *Schizophrenia bulletin*, *35*(3), 509–527.

Winkler, I. (2007). Interpreting the mismatch negativity. *Journal of Psychophysiology*, *21*(3-4), 147–163.