scientific reports

OPEN

Improved data quality and statistical power of trial-level event-related potentials with Bayesian random-shift Gaussian processes

Dustin Pluta[1,5], Beniamino Hadj-Amar[2,5], Meng Li[2], Yongxiang Zhao[3], Francesco Versace[4] & Marina Vannucci[2]✉

Studies of cognitive processes via electroencephalogram (EEG) recordings often analyze group-level event-related potentials (ERPs) averaged over multiple subjects and trials. This averaging procedure can obscure scientifically relevant variability across subjects and trials, but has been necessary due to the difficulties posed by inference of trial-level ERPs. We introduce the Bayesian Random Phase-Amplitude Gaussian Process (RPAGP) model, for inference of trial-level amplitude, latency, and ERP waveforms. We apply RPAGP to data from a study of ERP responses to emotionally arousing images. The model estimates of trial-specific signals are shown to greatly improve statistical power in detecting significant differences in experimental conditions compared to existing methods. Our results suggest that replacing the observed data with the de-noised RPAGP predictions can potentially improve the sensitivity and accuracy of many of the existing ERP analysis pipelines.

In studying the association of electroencephalogram (EEG) recordings and cognitive and behavioral outcomes in humans, researchers often focus on the study of event-related potentials (ERP), i.e., specific EEG segments time-locked to cognitive events or experimental stimuli[1]. Due to the substantial noise present in EEG recordings and the relatively weak signal, ERP analyses often focus on group-level effects that can be estimated from average ERPs[2]. For example, in the case study we consider in this paper, the goal is to detect differences in response to images with differing levels of emotional arousal, e.g. highly arousing, mildly arousing, or unarousing (Fig. 1). A commonly used analysis technique for this data summarises a specified time-window in the form of average amplitudes (i.e., voltage) to be analysed, for example, via one-way repeated measures ANOVAs across groups of subjects and/or conditions[3], or with an empirical bootstrap procedure[4,5]. Consideration of the average ERP in statistical analyses is often necessary due to a lack of available tools to model trial-level ERP features. Consequently, the potential impact of trial-level features on the estimation of the average ERP is often ignored or mitigated through pre-processing procedures[6,7]. Moreover, while there has been some study of within-subject variability of ERP characteristics, such as latency and amplitude, most results regarding the association of these features with cognitive and behavioral outcomes have been established only at the group level[8–11].

Early approaches for the analysis of trial-specific latencies of ERPs were presented in[12] and[13], and later extended to include estimation of trial-specific amplitudes by[14]. This initial work on trial-specific latency and amplitude estimation lead to the development of a class of so-called Variable Signal plus Ongoing Activity (VSPOA) models, a general framework for the analysis of brain potentials that assumes that the observed signal for each trial is a linear combination of a fixed number of components, each of which may be shifted by some trial-specific latency, plus trial-specific ongoing background activity. The differentially variable component analysis (dVCA) introduced by[15] is a Bayesian VSPOA that provides maximum a posteriori estimates of trial-specific latencies, amplitudes and component waveforms. In an effort to improve upon dVCA[16], proposed the analysis

[1]Department of Biostatistics and Data Science, Augusta University, Augusta, GA 30912, USA. [2]Department of Statistics, Rice University, Houston, TX 77005, USA. [3]Department of Statistics and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA. [4]Department of Behavioral Science, MD Anderson Cancer Center, Houston, TX 77030, USA. [5]These authors contributed equally: Dustin Pluta and Beniamino Hadj-Amar. ✉email: marina@rice.edu

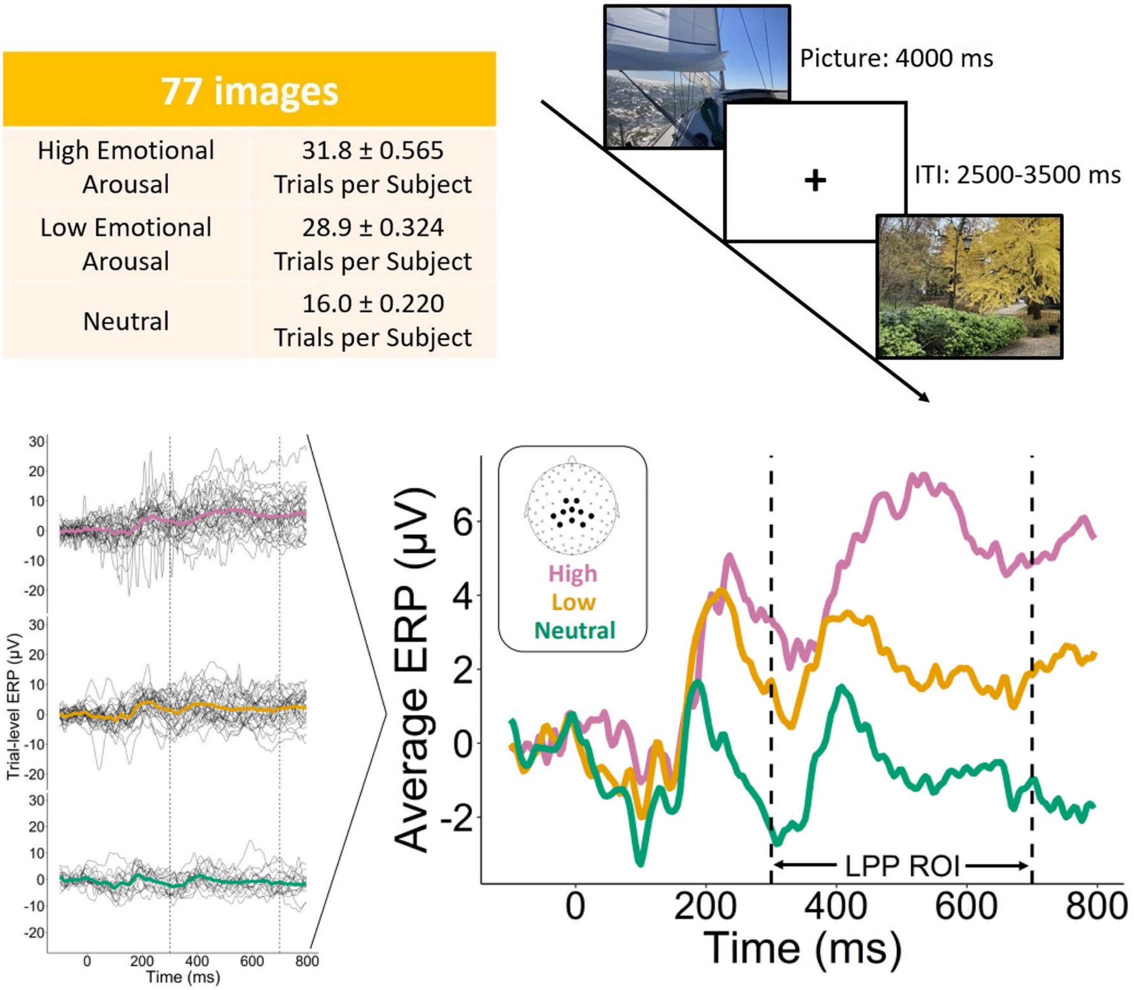

Figure 1. Trial-level ERP data modeling. In each experimental trial, a subject is shown an image drawn from a database consisting of images which are classified as high emotional arousal, low emotional arousal, and emotionally neutral. Trial-level data from ERP experiments is often summarized into condition-specific average curves for further analysis. The late positive potential (LPP) window is defined as ranging from 300 to 700 ms after stimulus presentation.

of single-trial ERP and ongoing activity (ASEO) method, which assumes autoregressive ongoing brain activity, rather than white noise, as in dVCA, and which allows for estimation of latency on a continuous scale, whereas dVCA restricts latencies to be integer multiples of the sampling interval. However, both dVCA and ASEO suffer from the need to initialize the number of components and component waveforms. In practice, it is common to inspect the average ERP of the data and initialize the components as segments of the average ERP[17,18]. However this data-driven initialization may impact the resulting inference and bias the estimated components toward the average ERP. In addition to VSPOA models, a number of other methods for trial-level ERP analysis have been proposed, including spatial and wavelet filtering[19,20], graph-based variability[21,22], and linear mixed models[23,24]. While these methods are effective in obtaining estimates of trial-specific parameters and component waveforms, they are often not suitable for inference of arbitrary single-trial characteristics, and may suffer from restrictive assumptions regarding waveform structure and initialization, the single-trial noise model, and incorporation of prior scientific beliefs.

Methods described above have been mainly developed for the analysis of trial-level ERP signal, that is, single trial EEG data collected and processed according to time-locked stimuli. In recent years, there has been renewed interest in trial-by-trial analyses of raw EEG signals, i.e., not time-locked, using, e.g., support vector machines[25,26], spatial network discriminant analysis[27], topological data analysis[28], and deep neural networks[29–31]. In contrast to model-based trial-level ERP analysis, these methods are primarily for the prediction of trial conditions from single-trial EEG waveforms, rather than inference of ERP components and trial-specific quantities of interest. A comparison of the performance of logistic regression, support vector machines, and neural networks in EEG signal classification and ERP detection in different signal-to-noise ratio (SNR) settings is given in[32]. While these predictive algorithms have found success in application, particularly for training of brain-computer interface systems[33–35], it is often difficult to identify and interpret the discriminating features.

Motivated by the need for rigorous, flexible, and interpretable methods for trial-level analysis of ERP data, we developed the Random Phase-Amplitude Gaussian Process (RPAGP) modeling framework, which assumes that individual trials are generated as a common structural signal transformed by a trial-specific amplitude and phase shift plus ongoing brain activity. In the proposed RPAGP framework, a form of VSPOA model, the unknown signal is modeled via a Gaussian Process (GP) prior and an autoregressive process is assumed for the ongoing brain activity. We set priors on the trial-specific model parameters and design an efficient algorithm for posterior inference. In contrast to methods for group-level analysis of ERP data, our method produces estimates and uncertainty quantification of the structural signal and trial-specific characteristics, along with posterior distributions for trial-specific signals. This feature makes it possible for RPAGP to improve the sensitivity and accuracy of existing ERP analysis pipelines by replacing the observed data with the de-noised model predictions. Here we use the data from[36] to illustrate the use of RPAGP for testing of trial-level characteristics. Previous group-level analysis of this data has shown significance evidence that the amplitude of the late positive potential (LPP, see Fig. 1) is proportional to the level of arousal of the image, a result which is corroborated by the existing literature[3,37–41]. We reproduce such result with a trial-level analysis of single subject data. We also show how data quality and statistical power can potentially be greatly improved via our model-based de-noising.

Our proposed model falls within the general class of VSPOA models, as discussed above. Even though motivated by neuroscientific beliefs about the structure and characteristics of ERP signals, it has been difficult to develop a practical statistical implementation of these general models, due to the challenge of specifying the form of the component curves a priori in a manner that admits valid statistical inference. The closest construction to the proposed RPAGP model is the ASEO method of[16] which assumes a VSPOA model with autoregressive ongoing activity and uses an iterative numerical algorithm for inference that depends on extensive use of the Fourier transform. To the authors' knowledge, the most readily available implementation of ASEO is in the FieldTrip MATLAB toolbox as part of the "event timing analysis" routine[42]. However, running ASEO via FieldTrip requires a priori specification of the number of components, and either the initial latencies of each component or initial estimates of the component waveforms through a decomposition of the average ERP, which in turn requires choosing the time window associated with each component. More importantly, while providing accurate point estimates of model parameters and reconstruction of component waveforms, the ASEO algorithm does not provide uncertainty quantification of these estimates, and so it is limited for statistical inference. On the contrary, RPAGP allows a flexible construction via the use of Gaussian process priors, which provides full inference on trial-level latency, amplitude, and the structural signals associated with each experimental condition. We demonstrate the proposed model is practical for application to data produced by most current ERP experiments, and does not require subjective initialization of component curves. In the Supplementary Material A1 we use simulated data to further investigate performance of the proposed RPAGP model to reconstruct the structural signal.

Results

We consider the data analysed in[36] to study emotional processes. In each experimental trial, a subject is shown an image drawn from a database of images[43–45]. For the analyses of this paper we included 32 images rated as highly arousing, 29 rated as low arousing, and 16 rated as emotionally neutral (Fig. 1, see "Methods"). Scientific interest is primarily regarding the mean of the task-related signal over the LPP window, which is defined as ranging from 300 to 700 ms after stimulus presentation. Previous analyses have shown that the magnitude of the LPP mean averaged over trials and subjects is positively associated with the level of arousal of the stimuli[3]. We reanalysed the data by fitting the proposed RPAGP model (see "Methods") to the trial-level data in the LPP window, standardized to $[0, 1]$ with increments $\Delta t = 1/224$, and assuming a common structural signal across all trials.

Single subject analysis

In order to illustrate the application of the RPAGP model we first present results from the analysis of one subject. Figure 2 shows the data (left panel), together with empirical means, averaged by condition, and 90% bootstrap

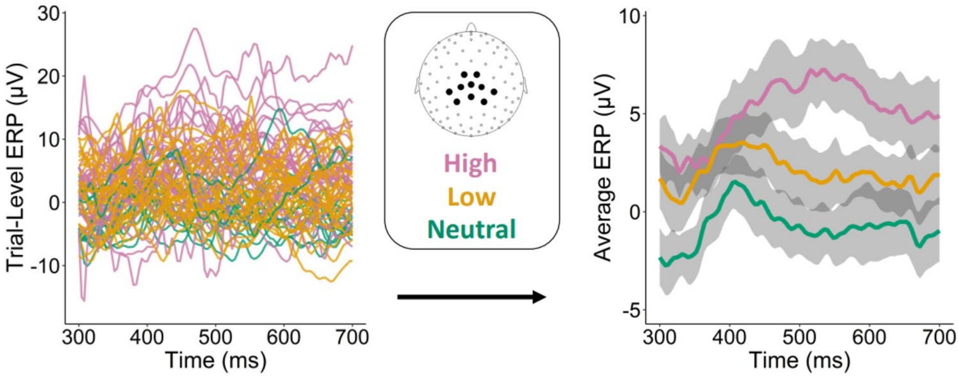

Figure 2. Single-subject analysis of trial-level ERP signals. (left) Raw trial data from a single subject. (right) Means by condition show the expected relationship, but the associated 90% confidence bands, obtained by the empirical bootstrap, are wide and do not clearly suggest a statistically significant difference between the low and neutral categories.

confidence intervals (right panel) computed by a stratified bootstrap procedure (see "Methods"). The 90% confidence bands are used here for better visualization of the condition estimates, as the 95% bands would obscure the differences. The empirical estimates suggest differences in the LPP means by condition, but it is not clear if these differences are statistically significant. As is evident from the data, there is substantial noise present in single-trial ERP data, due to on-going brain activity and idiosyncratic fluctuations. These noise components are not accounted for in the empirical bootstrap method, resulting in wide confidence bands. On the contrary, Fig. 3 shows the temporally-aligned RPAGP trial predictions $\tilde{y}_i(t)$, estimated via the posterior median, obtained by fitting the proposed PGP model (left panel) and the RPAGP model means obtained by averaging the aligned trial estimates by condition, together with the 90% credible intervals (right panel). We note that the posterior estimates of the means by condition calculated from the RPAGP model exhibit similar shape and scale to the empirical estimates, but with much less uncertainty relative to the empirical bootstrap method; the separation of estimated response by condition is clear from the model fits, but less apparent when considering the bootstrap estimates. Application of RPAGP allows removal of the noise and latency components, which results in narrower credible bands and a more powerful detection of differences among conditions.

One may notice that the empirical means and posterior mean ERPs in Figs. 2 and 3 look markedly different, particularly in the neutral condition. This is the result of the RPAGP modeling assumptions and differences in the methods of estimation. In the empirical approach, each condition is considered individually, and the respective trials are averaged to obtain a condition-specific empirical ERP estimate. With this approach, structural characteristics of the background activity may be present in the condition averages. Furthermore, subsetting the trials by condition can result in larger variance estimates. In comparison, the model-based RPAGP estimates are obtained under the assumption that the structural shape of the ERP is common across all conditions. Therefore, the higher voltage from 300 to 500 ms observed in the Neutral condition, for example, is likely not captured by the model as it is not consistently present in the other two conditions. Furthermore, since amplitudes of the Neutral condition trials are close to zero, the posterior mean curve will be close to the zero curve, and could obscure features in the Neutral condition even if it were significantly present in all conditions. The zeroing of the Neutral condition curve is not unexpected, as the assumption of the Neutral condition images is that they produce a small or no LPP response. This also explains why we see little difference in the High condition curves, as the signal is strongest in this condition and structures from background activity are less prominent in the empirical mean relative to the Neutral condition. Plots of raw trial data and means by condition, together with the temporally-aligned RPAGP trial posterior estimates and means by condition, for 4 additional subjects, are shown in the Supplementary Material A1.

Posterior densities of the trial LPP means (Fig. 4, see "Methods") show that trials from the selected subject follow the expected trend on average, with a large amount of variability in the LPP means of individual trials, resulting in distributions by condition with substantial overlap. Despite the wide range of trial LPP means, the posterior distributions of the difference of LPP means among conditions are positive and tightly concentrated away from zero, clearly supporting the LPP hypotheses $\mu_{LPP}^{High} > \mu_{LPP}^{Low} > \mu_{LPP}^{Neutral}$. The largest difference in mean is exhibited by the comparison of the High and Neutral conditions, with an estimated difference in means (90% CI) of 1.37 (1.33, 1.42); the High–Low and Low–Neutral differences were similar, with means and CIs of 0.78 (0.75, 0.81) and 0.60 (0.55, 0.64) respectively. For comparison, the empirical bootstrap procedure also finds the largest difference between the High and Neutral categories with an estimated difference in means (90% confidence interval) of 1.41 (0.97, 1.84). The Low–Neutral and High–Low differences are estimated as 0.61 (0.22, 0.98) and 0.70 (0.24, 1.15) respectively. While each of the three comparisons are found to be significantly greater than zero and with similar point estimates for both methods, we observe that the bootstrap confidence intervals are wider than the model credible intervals.

Finally, Fig. 5 shows the RPAGP estimates of trial LPP means ordered sequentially over the course of the experiment. These plots can be useful to understand the variability of the characteristic of interest over the course of the experiment. We note that both the Low and High trials tend to have larger mean than the Neutral trials

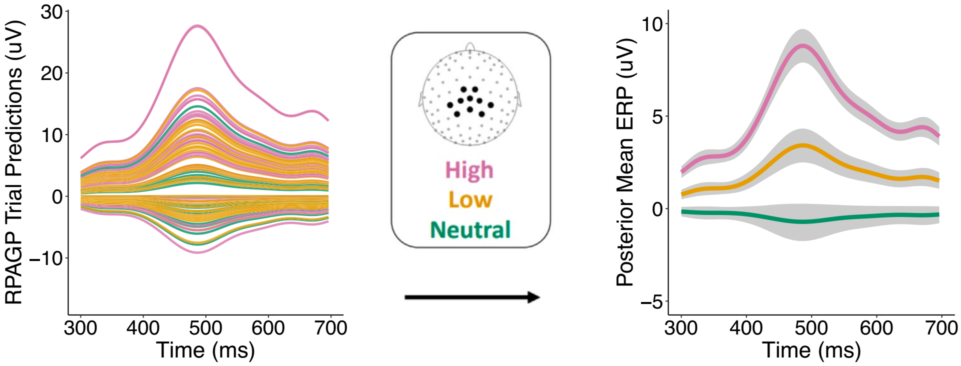

Figure 3. Inference from RPAGP on single-subject trial-level ERP signals. (left) The temporally-aligned RPAGP trial predictions, estimated via the posterior median, extract the task-related signal from the trial-specific noise. (right) Means by condition obtained by averaging the trial estimates. The 90% credible bands provide strong evidence of a difference in means among conditions.

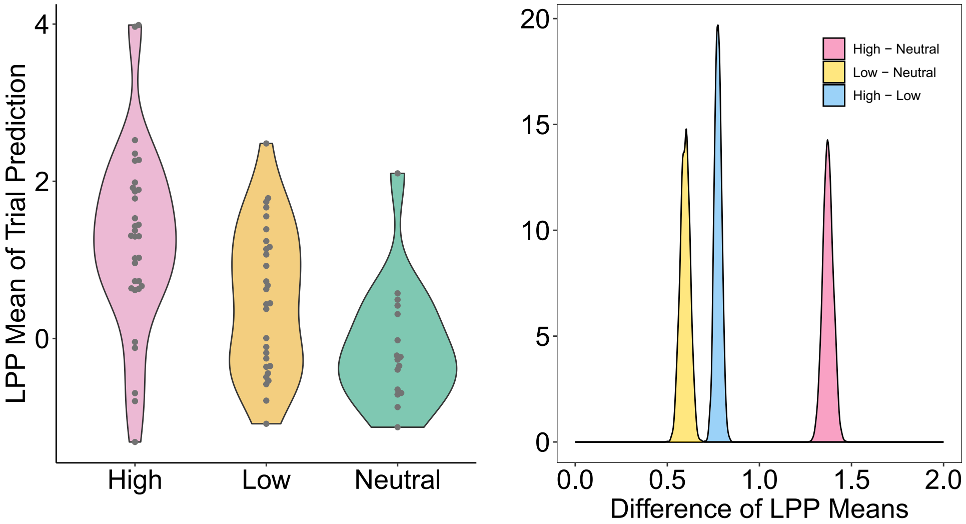

Figure 4. Inference on LPP means. (left) The posterior densities of the RPAGP estimates of trial LPP means by condition suggest that the selected subject exhibits the expected relations among the LPP means on average. There is also substantial variability apparent in the LPP means between trials within the same condition. (right) The posterior densities of the differences in LPP means are tightly concentrated and significantly greater than zero, in accordance with the expected relationships among conditions.

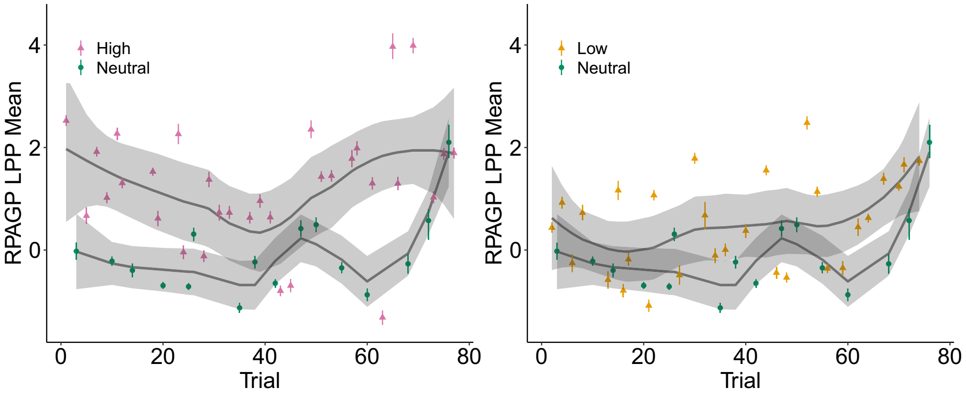

Figure 5. Trend analysis of the LPP mean estimates. Longitudinal plots of trial-specific LPP means (points), posterior 90% credible intervals (error bars), loess smoothed curves (solid lines), and loess standard error bands (shaded regions). These plots illustrate trial-level variability and can reveal trends over the course of the experiment.

and that there is greater variability present in both emotionally arousing categories compared to neutral images, with the High condition images showing the greatest variability overall.

Impact on data quality

Next, we present results from the application of the RPAGP model to ERP data recorded from 161 subjects participating in the LPP emotional response image experiment of[36]. Table 1 shows the proportion of subjects satisfying the expected LPP mean relationships among conditions for the RPAGP method, the empirical bootstrap (EMP) method and a standard ANOVA test performed on the empirical means from the original trial data, averaged by condition. We also report results we obtained by applying the ASEO algorithm via the Fieldtrip MATLAB toolbox[42] to obtain estimates of the trial-specific signals, from which the ASEO trial-specific LPP means were computed. Confidence intervals of the differences in LPP means by condition were constructed by following the empirical bootstrap procedure but replacing the observed data with the ASEO trial-specific estimates. Significance of difference in LPP means was determined from these confidence intervals as with the empirical bootstrap procedure. On all three tests, RPAGP finds a much higher proportion of subjects satisfying the expected mean relationships compared to ASEO, EMP, and ANOVA, with RPAGP detecting 86.3%, 75.8%, and 70.2% of subjects showing significant differences for the High–Neutral, High–Low, and Low - Neutral tests, respectively. This is compared to 45.3%, 26.7%, 18.6% for ASEO; 44.1%, 29.2%, and 22.4% for EMP; 42.9%,

	High–neu.	High–low	Low–neu.	All 3 tests
RPAGP	0.863	0.758	0.702	0.497
ASEO	0.453	0.267	0.186	0.031
EMP	0.441	0.292	0.224	0.068
ANOVA	0.429	0.288	0.168	0.038
All 4 methods	0.429	0.267	0.168	0.031

Table 1. Impact on data quality. Proportion of subjects giving expected results. RPAGP detects a substantially higher proportion of subjects following the expected mean relationship among categories compared to ASEO, EMP and ANOVA. This is primarily due to the de-noising process of the model, which reduces uncertainty about the estimates for the differences among conditions, compared to the other methods that do not account for the structured and idiosyncratic noise present in ERP data (see Figs. 6 and 7 and text).

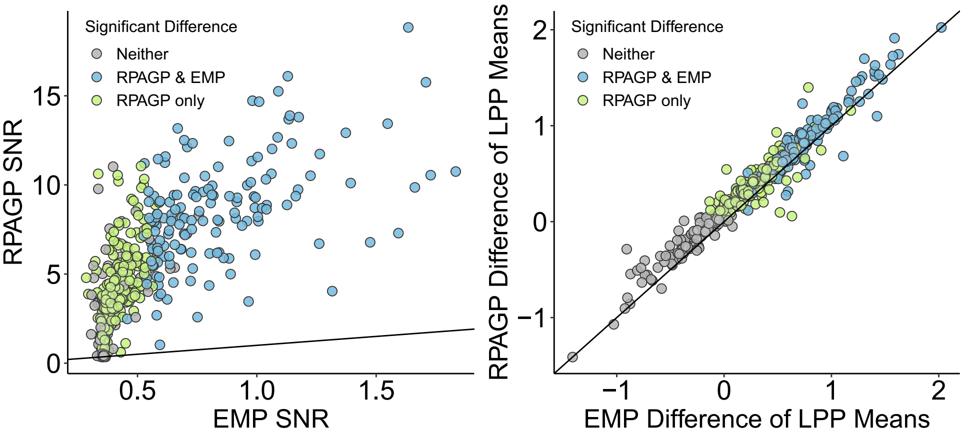

Figure 6. Improvement on data quality of the RPAGP denoising. (left) Comparison of the signal-to-noise ratio (SNR) of the model predictions (RPAGP) against the empirical means (EMP). The de-noising process of the model improves the SNR for all but one trial, with a 15-fold improvement on average. (right) The RPAGP and empirical estimates of the differences in LPP means by condition essentially agree for all trials, thus the improvements in statistical power for RPAGP primarily result from the removal of noise.

28.8%, and 16.8% for ANOVA, on the same tests, respectively. We note that RPAGP shows joint significance on all three comparisons for 52.2% of subjects, compared to only 6.8% of subjects detected by EMP. For a further examination of test performance, we focus on a comparison of RPAGP and EMP, since ASEO and ANOVA show lower proportions of significant subjects on all three tests.

We also tested whether our method increases the risk of detecting false positives. For each subject, we randomly divided the trials of each condition into two groups and performed an ANOVA test for the difference in means. We repeated this random split 1000 times for each condition and calculated the percentages of significant tests. With the modeled data, the average of significant results was 4.3%, 4.5% and 4.6% for the High, Low and Neutral conditions, respectively. For comparison, applying the same procedure to the raw ERP data resulted in 4.5%, 4.5% and 4.4% significant differences for the High, Low and Neutral conditions, respectively. We conclude that our method substantially increases the chance of detecting a true positive while not increasing the probability of false positives.

To quantify the practical effect of the model de-noising provided by RPAGP in the LPP experiment, Fig. 6 illustrates the change in signal-to-noise ratio (SNR, see "Methods") and a comparison of the point estimates of difference in the means by condition, for the RPAGP and EMP methods, and Fig. 7 shows a comparison of the CI widths for these estimates. The RPAGP model greatly improves the SNR compared to the EMP method, with a 15-fold increase in SNR on average. This improvement in SNR could result from increases in the estimated differences, decreases in the widths of the corresponding CIs, or a combination thereof. Examination of the point estimates for differences in means produced by RPAGP and EMP shows that they are essentially equivalent (Fig. 6), but, as seen in Fig. 7, the CIs produced by EMP are much wider than those produced by RPAGP. Thus the improvement in SNR is entirely due to the narrower confidence intervals associated with the RPAGP estimates, rather than increased point estimates. The additional cases detected by RPAGP but not EMP (bright green) are primarily those with small point estimates, but also include a number of cases with relatively large point estimates. RPAGP detects a significant relationship for all but one test, with a point estimate of differences among the means by condition above 0.135, and shows consistent CI widths across all subjects and tests. In contrast, EMP fails to identify a significant relationship even for some comparisons with estimated differences; the EMP CIs are substantially larger than the RPAGP CIs across all tests and subjects, and much greater variability of CI

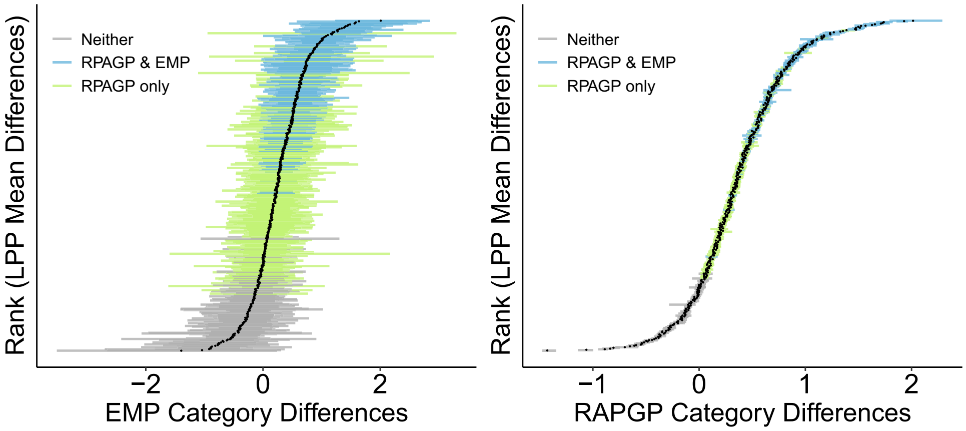

Figure 7. Differences in LPP means between categories. The 90% credible intervals for differences in LPP means between categories produced by RPAGP (right) are substantially narrower than the 90% confidence intervals obtained via empirical bootstrap (left).

width is evidence in the EMP CIs. This stark difference in CIs is due to the separation of signal and noise in the RPAGP model, making the resulting error estimates less susceptible to subject- and trial-specific variations that are not related to the stimuli.

For further evidence, we calculated the Standardized Measurement Error (SME, see "Methods"), a measure of (subject-level) data quality proposed by[46] for averaged ERPs, using both the raw data and the RPAGP trial predictions. Results showed that the RPAGP de-noising decreases the SME compared to the raw data, but also that the resulting change in SME is relatively small, with the RPAGP trial predictions showing an average decrease (standard error) of 0.039 ± 0.062 relative to an average SME of 0.555 computed over the raw data and model predictions. This close correspondence in SME from the raw data and trial predictions indicates that the model does not substantially reduce inter-trial variability relative to the raw data. In conjunction with the results on improvement in SNR from application of RPAGP, the correspondence of SME suggests that the improvement of RPAGP over EMP (Table 1) comes primarily from a reduction in noise of the estimated parameters, rather than from a boosting of the LPP signal (Fig. 6), nor from a reduction in the inter-trial variability.

In the Supplementary Material A1, we report additional investigations into the performances of RPAGP that were performed based on simulated data. Simulated settings, in particular, allow us to calculate Mean Squared Errors (MSEs) between the estimated and the true structural signal. We simulated data for two categories, following the general structure suggested by the real ERP data of our application. Details of data generation are provided in the Supplementary Material A1. We investigated performances for different values of the model parameters $\sigma_\beta, \sigma_\tau$ and σ_ε (see "Methods") and by averaging results over 50 replicated datasets. Figure A.1 shows the true structural signal f, the simulated trial-level data and the RPAGP estimate $\tilde{f} = \widehat{\bar{\beta} f}$, computed as the posterior median of the distribution of the temporally-aligned trial estimates, for one dataset simulated with $\sigma_\beta = 0.1, \sigma_\tau = 0.01$ and $\sigma_\varepsilon = 0.1$ and $n = 30$. By adjusting for the trial-specific latencies, RPAGP is able to provide a more accurate estimate of the true structural signals relative to the empirical means. MSEs obtained by estimating the true structural signal by RPAGP and by the empirical mean of the raw data, and averaged over the 50 replicates, are reported in Table A.1. RPAGP shows substantially smaller MSE compared to EMP in the presence of latency, and is at least as accurate as EMP when no latency is present.

Improvement in statistical power

Results above have indicated that the denoised signals produced by RPAGP are potentially less noisy and of better quality than the observed trial-level data. Our next question is to whether the availability of the denoised estimates can also result in a reduction of the number of trials needed to power a particular study, with respect to the standard practice of using average curves obtained from raw data. We therefore assessed the power of the RPAGP method for testing the three hypotheses implied by $\mu_{LPP}^{High} > \mu_{LPP}^{Low} > \mu_{LPP}^{Neutral}$ based on the LPP emotional responses data from[36] and varying number of trials (see "Methods"), for the subject considered in "Single subject analysis". Results of this analysis for total number of trials $n = 15; 25$, are reported in Table 2 and show that RPAGP has power at least as great as the empirical bootstrap test for all tests and sample sizes considered. Furthermore, RPAGP substantially outperforms the empirical method in testing for differences between the Low and Neutral conditions, which is the test with lowest signal-to-noise ratio.

For further investigation, we performed a two-group power analysis on simulated data with varying numbers of trials and by generating data with different signal-to-noise ratios. Details of the signals generation are given in the Supplementary Materials A1. For each setting, results were averaged over 50 replicated datasets. The estimated powers are given in Table A.2. The results show that the power of the RPAGP test is substantially greater than that of the empirical method in the majority of simulation settings considered, and is universally better than the empirical method for the largest sample size considered ($n = 30$). In the easier settings, with low SNR, RPAGP has estimated power near or equal to 1, compared to 0.35 for the empirical method.

	$n = 15$	$n = 25$
Empirical H-L	0.65	0.9
RPAGP H-L	0.85	1.0
Empirical H-N	0.95	1
RPAGP H-N	1	1
Empirical L-N	0.50	0.50
RPAGP L-N	0.75	0.90

Table 2. Statistical power analysis. Estimated power in testing for differences in mean amplitudes across conditions for empirical bootstrap and RPAGP on ERP data. RPAGP consistently shows power at least as great as the empirical testing power, with the greatest increase in testing the difference of Low and Neutral conditions.

Discussion

We have developed the Random Phase-Amplitude Gaussian Process (RPAGP) modeling framework, as a Bayesian implementation of a VSPOA model for trial-level analysis of ERP data. The approach assumes ERP trial data are generated by trial-specific phase shifts and amplitude scalings of a common structural signal that is flexibly modeled with a Gaussian process prior. In contrast to methods for group-level analysis of ERP data, our method produces estimates and uncertainty quantification of the trial-specific parameters and the structural signal, in addition to posterior distributions of trial-specific signals. Trial-level models for ERP data offer markedly advantages over average-based approaches. Firstly, adjusting for variations in trial-level characteristics may result in improved ERP signal estimates, as we have demonstrated in the analyses of this paper. This, in turn, leads to increased statistical power for condition-level inferences compared to empirical methods. Furthermore, analysis of the distribution of trial-level ERP features may reveal scientifically relevant patterns across trials within a subject, or across subjects.

The availability of trial-level signal estimates makes it possible for RPAGP to supplement any existing ERP analysis method by replacing the observed data with de-noised model predictions. De-noised trial predictions may be used, without requiring additional processing, to improve the sensitivity and accuracy of existing ERP analysis pipelines, such as averaging the extracted signals over a scientifically relevant time window, perform spectral analysis and/or assist in detecting outlying or erroneous trials that may not have been removed during pre-processing. We have applied RPAGP to data from a study where the interest is in detecting differences in ERP responses to images with differing levels of emotional arousal. Results, in particular, have shown how data quality and statistical power can potentially be greatly improved via our model-based de-noising.

Possible extensions of the RPAGP model presented here could incorporate alternative assumptions regarding the structure of the data, and further expand its range of applications. For example, for clarity of presentation and application to LPP experimental data, we have assumed a common structural signal for all trials, with potential differences among conditions determined solely by latencies or amplitudes. A direct and simple extension of the proposed model is to allow for different structural signals across experimental categories, each modeled as an independent GP prior, for comparison and testing of trial-level features across categories when there is little information regarding the structure of the condition-specific signals. This form of the model can be implemented with minor changes to the Algorithm 1. Furthermore, the most general class of VSPOA models assume that observed data result from a linear combination of multiple component waveforms, each with its own trial-specific phase and amplitude. While identifying the approximate time intervals and shapes of these components for algorithm initialization can facilitate estimation of trial-specific features, as in[16], building a flexible inferential procedure incorporating these structural assumptions has not yet been developed, to our knowledge. With the proposed RPAGP model as a starting point, a multiple-component trial-level model similar to the framework in[16] may be possible through further refining of the model priors. Finally, we did consider some analysis and interpretation of trends in the amplitude and latencies over the course of the experiment, given some of the apparent condition differences that can be seen in Fig. 5. However, given the characteristics of the experimental design, these patterns in a single subject are likely spurious. A formal analysis of trends in ERP amplitudes over the course of the experiment using the proposed model is an interesting possible direction for future work.

Methods

Study participants

For the analyses of this paper, we used data from 161 individuals enrolled in a previous study in which ERPs were collected on brain reactivity to an array of emotionally arousing images[36]. Participants in the original study lived at a stable address within Harris county, TX, did not report any psychiatric disorder during a screening interview, were between the ages of 18 and 55 years, were able to speak English, had access to a telephone, were negative to a urine drug panel and to a urine pregnancy test. The mean age of the sample was 35 years (SD = 9), 75% of the sample were females and the racial distribution included 37% Blacks, 28% Whites, 14% Asian, 21% others (including not reported).

Picture viewing task

The images used in the study were selected from the International Affective Picture System (IAPS)[43] and from a collection used in[44,45] created by downloading images from the internet and from another libraries of images.

The images used for the study belonged to 8 categories (Erotica, Romantic, Food, Neutral, Neutral Objects, Pollution, Attack, Mutilations) with 16 pictures in each condition. To conduct the analyses described in this paper, we excluded images depicting objects (i.e., food, pollution, and neutral objects) and re-classified erotic and mutilation contents as "High arousing" and romantic and attack contents as "Low arousing". The images were presented for 4 s in pseudo-random sequences (no more than two consecutive pictures of the same condition). A 3–5 s random intertrial interval, showing a white fixation cross against a black background, followed each image. The slideshow lasted approximately 20 min and included four 30 s intervals, during which the participant was instructed to relax. The images were presented on a plasma screen placed approximately 1.5 m from the participant's eyes. The images subtended approximately a 24° horizontal viewing angle.

Data collection procedures

All study procedures for the original study were approved by the UT MDAnderson IRB and were conducted in accordance with the relevant guidelines and regulations. Written informed consent was collected from each participant before starting any procedure. Details about the study are provided in the original publication[36]. Briefly, the study included one in-person laboratory visit when, after obtaining informed consent, eligibility was confirmed and the electroencephalogram (EEG) data were recorded while participants passively looked at a slideshow that included a wide array of images selected from a standardized database. During the slideshow, EEG was continuously recorded using a 129-channel Geodesic Sensor Net, amplified with an AC-coupled high input impedance (200 MΩ) amplifier (Geodesic EEG System 200; Electrical Geodesics Inc., Eugene, OR), and referenced to Cz. The sampling rate was 250 Hz, and data were filtered online by using 0.1 Hz high-pass and 100 Hz low-pass filters. Scalp impedance of each sensor was kept below 50 KΩ, as suggested by the manufacturer. At the conclusion of the visit, participants were debriefed and compensated for their time.

Data reduction procedures

Eyeblink artifacts were corrected using a spatial filtering method as implemented in the BESA software (BESA GmbH, Grafelfing, Germany), data were re-referenced to the average reference and eyeblinks were corrected. Data were imported into BrainVision Analyzer 2.1 (Brain Products GmbH, Gilching, Germany) and filtered with a high-pass filter of 0.1 Hz (12 dB/octave), a low-pass filter of 30 Hz (12 dB/octave), and a notch filter of 60 Hz. The data were then segmented into 900-ms segments, starting 100 ms before stimulus presentation. The 100-ms interval before stimulus presentation was defined as the baseline and subtracted from every data point in the segments. Artifacts were identified in the segmented data and channels contaminated by artifacts in more than 40% of the segments were interpolated using six neighboring channels. Voltage data averaged from 10 centroparietal sensors (EGI electrodes 7, 31, 37, 54, 55, 79, 80, 87, 106, 129) were used in the analyses, as these channels had shown the highest LPP differences between experimental conditions[36].

Empirical bootstrap estimates

For $B = 10,000$ bootstrap iterations, trials were resampled with replacement, stratified by condition, so that each bootstrap sample contained the same number of trials in each condition as the original data. For each bootstrap iteration, mean curves by condition were then computed over the resampled trials. The estimated means by condition and associated confidence intervals were then calculated as the mean and empirical quantiles of the bootstrapped distribution of the means by condition. Empirical bootstrap estimates of the differences in the LPP means by condition were obtained similarly by calculating the average over the LPP window for each of the bootstrapped means by condition. For each pair of conditions, we conclude that the LPP means are significantly different if the 5% empirical quantile of the bootstrap distribution of the difference in the means is positive.

Random phase-amplitude Gaussian process model

Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be n realizations of a stochastic process on $[a, b] \to \mathbb{R}$ observed at T evenly spaced points; without loss of generality, we will assume $a = 0, b = 1$ in the sequel. We first consider a general form of a curve registration model, to motivate the construction of the proposed RPAGP model. In curve registration, the aim is to match two or more functions that might exhibit a common shape distorted by function-specific variations referred to as *warpings*[47–50]. The problem of curve registration is typically stated as follows: given two functions, f_1, f_2, find the warping function γ such that f_1 and $f_2 \circ \gamma$ are optimally matched according to some criteria. The general class of warping functions consists of all orientation preserving diffeomorphisms of [0, 1], but in practice the space of warping functions considered is often restricted to some subset of parameterized functions, e.g. linear transformations[51]. To accommodate transformations along both the horizontal and vertical axes, an additional warping function is introduced; in this setting, it is common to consider linear transformations γ_1, γ_2, resulting in the phase-amplitude curve registration problem[52–55]. Accordingly, we assume that the observed data $\mathbf{y}_i$ result from a *structural* signal function f, composed with trial-specific left- and right-transformations g_i, h_i respectively, plus structured noise ν_i, as

$$\begin{aligned} y_i(t) = (g_i \circ f \circ h_i)(t) + \nu_i(t), \quad i = 1, \ldots, n; \quad t = k\Delta t, \\ \text{for } k = 0, \ldots, T-1. \end{aligned} \tag{1}$$

Specification of this model requires selecting the form of the transformations g_i, h_i, a class of signal functions, and the noise structure. This general form of curve registration model has been explored in the literature for different classes of transformations and signal curves, at various levels of generality[47,51,56].

In ERP studies, there is often scientific interest in estimating differences in signal amplitude and latency across trials, as these differences have been shown to be associated with cognitive and behavioral outcomes[2,3,17]. This suggests parameterizing the class of transformations as $g_i(x) = \beta_i x$, $h_i(x) = x - \tau_i$, where β_i, τ_i are the trial-specific amplitude and latency respectively. We assume f is an element of the reproducing kernel Hilbert space (RKHS) determined by a positive-definite kernel function $\kappa(t', t)$. This allows for the class of signal functions to be tailored to a specific application through the choice of an appropriate kernel and prior distribution over the RKHS. In the present work, we assume Gaussian process priors with radial basis kernel for f, which is both sufficiently flexible and computationally tractable for the analysis of ERP data[11]. Lastly, following empirical results in the ERP literature[16,57], the ongoing background activity is assumed to follow an autoregressive structure. Formally, our proposed RPAGP model can be written as

$$\begin{aligned} y_i(t) &= \beta_i f(t - \tau_i) + v_i(t), i = 1, \ldots, n;\ t = k\Delta t, \\ &\text{for } k = 0, \ldots, T-1 \end{aligned} \tag{2}$$

$$\begin{aligned} f &\sim \mathcal{GP}(0, \kappa) \\ v_i(t) &= \sum_{j=1}^{p} \phi_i v_i(t - j\Delta t) + \varepsilon_i(t) \\ \varepsilon_i(t) &\overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \end{aligned} \tag{3}$$

with $\Delta t = 1/(T-1)$. Similar to the terminology of[53], we refer to f as the *structural signal.*

Priors on model parameters

To complete the specification of the RPAGP model, we propose the following priors for the model parameters. Letting $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)$ the vector of latency parameters, the prior for $\boldsymbol{\tau}$ is constructed assuming a normal distribution

$$\boldsymbol{\tau} \sim \mathcal{N}\left(0, \sigma_\tau^2 \left(I_n - \frac{1}{n+1}\mathbf{1}\mathbf{1}'\right)\right), \tag{4}$$

where the marginal variance σ_τ^2 should be chosen according to the plausible range of latency for the given scientific application. If the RPAGP model is implemented with time coded on the unit interval (as in the present work), then it will be necessary to translate the scale of the real latency (in, e.g. milliseconds) to the corresponding value in [0, 1].

When the trial-specific amplitudes β_i can be assumed to be unrestricted on $\mathbb{R}$, it is convenient to adopt the prior $\beta_i \overset{ind}{\sim} \mathcal{N}(\mu_\beta, \sigma_\beta^2)$, which allows for efficient sampling. Choice of the prior parameters is discussed below. In some applications, it may be necessary to restrict $\beta_i \geq 0$, in which case a truncated normal may be used. As for the GP kernel κ and associated priors, these should be selected according to prior beliefs about the structural signal f. Throughout this article, we adopt the commonly used squared exponential kernel given by $\kappa_{SE}(t', t) = \exp\left\{-\frac{\rho^2}{2}(t' - t)^2\right\}$. The length scale parameter $\rho > 0$ governs the smoothness of the functions in the associated RKHS; we use the prior $\rho \sim Gamma(a_\rho, b_\rho)$, with a_ρ and b_ρ chosen according to the plausible range of values for the signal length scale. Further details are provided in[58]; an extensive investigation of GP covariance structures in the context of EEG data is given in[11]. As for the AR coefficients $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$, these are given a prior $\mathcal{N}(0, \Sigma_\phi)$, where $\Sigma_\phi = \text{diag}\{\sigma_{\phi_1}^2, \ldots, \sigma_{\phi_p}^2\}$, with variance values chosen to concentrate the prior mass in the interval $[-1, 1]$. Finally, the prior for the white noise variance σ^2 is $Gamma(a_\sigma, b_\sigma)$, with hyperparameters chosen to place diffuse mass over reasonable values for the error variance for the observed data.

We note that, as formulated thus far, the scales of the β_i and f are not identifiable due to invariance of the likelihood under transformations of the form $\beta_i \to r\beta_i, f \to \frac{1}{r}f$ for any constant $r \neq 0$ applied over all $i = 1, \ldots, n$. To resolve this, one may fix a nonzero value for β_i for one trial, or alternatively, fix a nonzero value of $f(t^*) = a$ at one point t^*. When exploring these options, we found that the former solution may introduce errors or affect the model fit if the selected trial is an outlier or especially noisy relative to the other trials, and therefore generally advocate for the latter solution. In fixing a value of f, it is beneficial to choose a time point for which f is likely nonzero, and to fix the value close to the value of the expected signal. If the scale of f and the observed signals are substantially different, priors for the β_i's must be calibrated accordingly. In practice, it is useful to examine the empirical mean to identify a reasonable time point and scale for f. In this case, it is suggested to set the prior parameters μ_β and σ_β so that there is prior mass over the plausible range of observed scaling factors relative to the mean curve. A sensitivity analysis of RPAGP, conducted on simulated data, indicated robust performances of RPAGP in estimating trial-specific amplitude and latency parameters (see Supplementary Material, Table A.3), suggesting that hyperparameter tuning can be reasonably guided by prior scientific knowledge.

Posterior inference via MCMC

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau}, \rho, \boldsymbol{\phi}, \sigma^2)$ represent the vector of model parameters, with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n) \in \mathbb{R}^n$ the vector of trial amplitudes, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n) \in \mathbb{R}^n$ the vector of trial latencies, $\rho > 0$ the length scale parameter of the GP squared exponential kernel for the GP prior, $\phi = (\phi_1, \ldots, \phi_p)$ the vector of autoregressive coefficients, and σ^2 the error variance. Given f and $\boldsymbol{\theta}$, the conditional distribution of y can be computed from the multivariate normal conditioning formula as

$$\boldsymbol{y}_i|f \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{y}_i|f}, \Sigma_{\boldsymbol{y}_i|f}), \tag{5}$$

with $\boldsymbol{\mu}_{\boldsymbol{y}_i|f} = K_{\boldsymbol{y}_i f} K_f^{-1} f$ and $\Sigma_{\boldsymbol{y}_i|f} = \Sigma_{\boldsymbol{y}_i} - K'_{\boldsymbol{y}_i f} K_f^{-1} K_{\boldsymbol{y}_i f}$, and where $K_{\boldsymbol{y}_i f}(t', t) = \beta_i \exp\{-\frac{\rho^2}{2}(t' - t - \tau_i)^2\}$ and $\Sigma_{\boldsymbol{y}_i} = \beta_i^2 K_f + \Sigma_{\boldsymbol{v}}$ for $K_f(t', t) = \exp\{-\frac{\rho^2}{2}(t' - t)^2\}$ and $\Sigma_{\boldsymbol{v}}$ the covariance matrix of an $AR(p)$ process. The log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{n} \left[\left(\boldsymbol{y}_i - \boldsymbol{\mu}_{\boldsymbol{y}_i|f}\right)' \Sigma_{\boldsymbol{y}_i|f}^{-1} \left(y_i - \boldsymbol{\mu}_{\boldsymbol{y}_i|f}\right) + \log(|\Sigma_{\boldsymbol{y}_i|f}|) \right] + nT \log(2\pi).$$

Given that $(f, \boldsymbol{y})|\boldsymbol{\theta}$ are jointly normal, the closed-form posterior for $f|\boldsymbol{y}, \boldsymbol{\theta}$ is calculated as

$$p(f|\boldsymbol{y}, \theta) \propto p(\boldsymbol{y}|f, \boldsymbol{\theta}) p(f|\theta) = \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu}_{\boldsymbol{y}_i|f})' \Sigma_{\boldsymbol{y}_i|f}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_{\boldsymbol{y}_i|f})\right\} \exp\left\{-\frac{1}{2} f' K_f^{-1} f\right\},$$

that is, $f|\boldsymbol{y}, \theta \sim \mathcal{N}(\boldsymbol{\mu}_{f|\boldsymbol{y}}, \Sigma_{f|\boldsymbol{y}})$, with

$$\boldsymbol{\mu}_{f|\boldsymbol{y}} = \Sigma_{f|\boldsymbol{y}} \sum_{i=1}^{n} \boldsymbol{y}_i' \Sigma_{\boldsymbol{v}}^{-1} K_{\boldsymbol{y}_i f} K_f^{-1} \text{ and } \Sigma_{f|\boldsymbol{y}}^{-1} = K_f^{-1} \left(\sum_{i=1}^{n} K_{\boldsymbol{y}_i f} \Sigma_{\boldsymbol{v}}^{-1} K_{\boldsymbol{y}_i f} \right) K_f^{-1} + K_f^{-1}.$$

The procedure in Algorithm 1 produces a chain of B draws from the joint posterior of $(\boldsymbol{\theta}, f)$. Given the current parameter values $\boldsymbol{\theta}^{b-1}$, values for $\boldsymbol{\tau}^b$, $\boldsymbol{\beta}^b$, and ρ^b are generated as a series of Metropolis-within-Gibbs steps; alternatively, if $\boldsymbol{\beta}$ is unconstrained, $\boldsymbol{\beta}^b$ may be sampled from the closed-form posterior for Bayesian linear regression (BLR). Then, f^b is sampled from the closed-form posterior conditional on $\boldsymbol{\tau}^b, \boldsymbol{\beta}^b, \rho^b$. The residual activity is calculated by removing the trial-specific mean $\boldsymbol{v}_i^b = \boldsymbol{y}_i - \mathbb{E}[\boldsymbol{y}_i|\boldsymbol{\theta}^b]$, from which draws of $\boldsymbol{\phi}^b, \sigma^{2,b}$ are made using closed-form posterior updates. We note that in general this procedure requires $\mathcal{O}(T^3)$ operations for inverting K_f and $\Sigma_{\boldsymbol{v}}$, but scales linearly with n, and is thus usually computationally feasible for real data from ERP experiments.

```
Result: MCMC samples from the posterior of $(\boldsymbol{\theta}, f)|\boldsymbol{y}$.
Initialize $\boldsymbol{\theta}^0, f^0$;
for $b$ in $1 : B$ do
    Update $\boldsymbol{\beta}^b|f^{b-1}, \boldsymbol{\tau}^{b-1}, \rho^{b-1}, \phi^{b-1}, \sigma^{2,b-1}$ using MH (alt. BLR);
    Update $f^b|\boldsymbol{\beta}^b, \boldsymbol{\tau}^{b-1}, \rho^{b-1}, \phi^{b-1}, \sigma^{2,b-1}$ from the closed-form posterior;
    Update $\boldsymbol{\tau}^b|f^b, \boldsymbol{\beta}^b, \rho^{b-1}, \phi^{b-1}, \sigma^{2,b-1}$ using MH;
    Update $\rho^b|f^b, \boldsymbol{\beta}^b, \boldsymbol{\tau}^b, \phi^{b-1}, \sigma^{2,b-1}$ using MH;
    Update $\phi^b, \sigma^{2,b}|f^b, \boldsymbol{\beta}^b, \boldsymbol{\tau}^b, \rho^b$;
end
```

Algorithm 1. Sampling algorithm.

Given the output of Algorithm 1, inference may be conducted through analysis of the posterior samples $\{\boldsymbol{\theta}^b\}_{b=1}^B$. In particular, testing for differences in mean latency and amplitude across groups of trials (i.e., experimental conditions) can be carried out directly from the posterior samples of $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$, as we show in the simulations below. Furthermore, the posterior draws of the structural signal f can be used to provide visualization of the estimated signal shape and posterior variability, as described below.

Inference on trial-level ERP signals

An important feature of the output from our model is the availability of samples on the structural signal f, which in turn allows us to produce posterior samples of trial-specific signals. Let $\boldsymbol{\tau}^b$, $\boldsymbol{\beta}^b$ and f^b be the bth MCMC draws of the respective parameters. Model predictions $\hat{y}_i(t)$ of the individual trials can then be obtained as, e.g., the pointwise-median of the posterior samples of the signal for trial i calculated as $\hat{y}_i^b(t) = \beta_i^b f^b(t - \tau_i^b)$. Here, however, we focus on the *temporally-aligned* trial estimates $\tilde{y}_i(t) = \widehat{\beta_i f}(t)$, calculated as the median of the posterior samples $\tilde{y}_i^b(t) = \beta_i^b f^b(t)$. This effectively treats the estimated latencies $\hat{\tau}_i$ as nuisance parameters, allowing analysis of features of interest adjusting for trial-specific latency. In our application, preference for the aligned estimates follows from the scientific assumption that the LPP window (see Fig. 1) should be adjusted for each trial according to the trial-specific latency. The aligned estimates may also be averaged to compute $\tilde{f} = \widehat{\bar{\beta} f}$, the structural signal scaled by the mean amplitudes of the trials (possibly within each condition), which is the most "natural" scaling of f relative to the data.

Inference on LPP means

Next, we present inference on differences in LPP means across categories using the RPAGP fit. Let us denote the LPP mean as μ_{LPP}. When referring to the parameter for a specific stimulus condition, e.g. Highly arousing, we will denote this as μ_{LPP}^{High}. To estimate the LPP means, we produce the aligned trial estimates $\tilde{\boldsymbol{y}}_i^b$ from the bth MCMC draw and compute the corresponding LPP means by averaging the $\tilde{\boldsymbol{y}}_i^b$ within each condition, and then averaging over the LPP window. That is, for condition $c \in \{High, Low, Neutral\}$, trial conditions c_i, n_c the number of trials in condition c, and LPP window $[t_{lwr} = 300, t_{upr} = 700]$ with length $T_{LPP} = t_{upr} - t_{lwr}$, the bth posterior draw of the LPP mean is

$$\hat{\mu}_{LPP}^{c,b} = \frac{1}{n_c T_{LPP}} \sum_{t=t_{lwr}}^{t_{upr}} \sum_{i:c_i=c} \tilde{y}_i^b(k\Delta t)$$

and the posterior distribution of the differences of means can then be calculated as $\hat{\mu}_{LPP}^{c_1,b} - \hat{\mu}_{LPP}^{c_2,b}$, for conditions c_1, c_2. Similarly, an *LPP* mean for a single trial i, is calculated as $\hat{\mu}_{LPP}^{i,b} = \frac{1}{T_{LPP}} \sum_{t=t_{lwr}}^{t_{upr}} \tilde{y}_i^b(k\Delta t)$.

Testing of the LPP hypotheses for differences in conditions is conducted by computing the 5% lower credible bound for the mean differences, and concluding significance if this value is positive. In comparing performance of RPAGP and frequentist methods, we note that while the bootstrap confidence intervals and Bayesian credible intervals have different theoretical foundations, it has been shown that Bayesian CIs approximately satisfy frequentist coverage properties under fairly general conditions. Further discussion is provided in[59].

Parameter settings

Considering the choice of priors and hyperpriors, our model has a total of $6 + p$ free parameters, for autoregressive order p, which reduces to 7 parameters if all autoregressive coefficient priors are chosen to be the same. The parameters and hyperparameters that must be set to initialize the model are: (i) the variance of the latency paramaters; (ii) the variance of the amplitudes; (iii) the hyperprior scale and shape for the length scale of the structural signal; (iv) the prior for the autoregressive noise parameters, including the choice of the AR order; and (v) the hyperprior scale and shape for the variance of the white noise. In all analyses of this paper, prior hyperparameters were chosen to specify weakly informative priors consistent with the assumptions and available prior information from existing ERP studies. In order to ensure identifiability of the $\boldsymbol{\beta}$ parameters, we chose to fix the value of the structural signal 400 ms after stimulus presentation to be approximately equal to the value of the grand average at this time point. We consequently chose the prior $\beta_i \overset{ind}{\sim} \mathcal{N}(1, 0.5)$ to place the majority of the prior mass on (0, 2), $\sigma_\tau^2 = 0.02$ to put approximately 95% of marginal prior mass for latencies equal to ±9 time points, or shifts of ±4% of the LPP window, and $a_\rho = 12, b_\rho = 1$ to give a wide range of plausible GP length scales. The autoregressive order for v was selected as $p = 2$ after examination of the autocorrelation and partial autocorrelation functions of data residuals resulting from subtraction of the subject-specific average ERPs and the residuals from RPAGP fit with white noise error; prior variance for ϕ_1, ϕ_2 was set at 0.5. The prior for σ^2 was set as $Gamma(1, 1)$.

The RPAGP model was fit via Algorithm 1 for $B = 8000$ MCMC draws. On a MacBook Pro computer with 2 GHz Quad-Core Intel Core i5 and 16 GB RAM, this took about 10 hours, for one subject. Analyses on multiple subjects were run in parallel on a cluster computer. For Metropolis-Hastings sampling, Gaussian proposal distributions centered at the previous parameter draw were used, with variances set at 0.005 and 1 for $\boldsymbol{\tau}$ and ρ, respectively. Convergence of all parameters was assessed by the Gelman-Rubin diagnostic measure $\hat{R}$, with sampling terminated when all parameters yield $\hat{R} < 1.1$ as recommended in[60]. Effective sample sizes for all parameter posterior samples were also calculated following termination and evaluated according to the stopping rule in[61] as an additional check for any issues with sampling.

Data quality measures

To quantify the practical effect of the model de-noising provided by RPAGP in the LPP experiment, we consider the signal-to-noise ratio (SNR), here defined as the ratio of the point estimate LPP mean for condition c to the 90% credible interval width for the estimate, that is $SNR = \frac{\hat{\mu}_{LPP}^c}{CI_{upr} - CI_{lwr}}$[62]. The SNR quantifies the magnitude of the signal relative to the variability of the noise in the data, i.e., a larger SNR indicates a more easily detected signal. When computing SNR for difference in LPP means between different experimental condition, we evaluated the SNR as the ratio of the absolute value of the difference LPP means for condition c_1 and c_2 to the corresponding 90% credible interval width, namely SNR = $\frac{\hat{\mu}_{LPP}^{c1,c2}}{CI_{upr}^{c1,c2} - CI_{lwr}^{c1,c2}}$, with $\hat{\mu}_{LPP}^{c1,c2} = |\hat{\mu}_{LPP}^{c_1} - \hat{\mu}_{LPP}^{c_2}|$

An alternative measure of (subject-level) data quality is the standardized measurement error (SME) proposed by[46]. The SME of a sample y relative to a statistic of interest $R(y)$ is defined as the sampling standard deviation, which can be estimated via a bootstrap resampling procedure. In our context, with the LPP mean $\hat{\mu}_{LPP}$ as the statistic of interest, the SME is $SD(\hat{\mu}_{LPP}^{c_1} - \hat{\mu}_{LPP}^{c_2})$, where c_1, c_2 refer to two experimental conditions, e.g., High and Low arousing images. A lower SME score indicates lower variability of the test statistic across trials.

Power analysis

We assessed the power of the RPAGP method for testing the three hypotheses implied by $\mu_{LPP}^{High} > \mu_{LPP}^{Low} > \mu_{LPP}^{Neutral}$, for varying number of trials, as follows. In order to mimic the relative sizes of the categories in the original sample, for total number of trials $n = 15; 25$, we resampled with replacement $0.4n$ trials from each of the High and

Low conditions, and $0.2n$ trials from the Neutral condition, to create 20 synthetic data sets for each sample size. We then conducted the test for difference of LPP means, concluding significance if the 5% quantile of the posterior distribution of the difference of the means is positive. The estimated power is computed as the proportion of significant tests. Similarly, for comparison, we conducted the empirical bootstrap test for differences in means by conditions. For each synthetic data set, we resample with replacement to create $B = 10{,}000$ bootstrap data sets. On each of these data sets, we compute the *LPP* means by condition and pairwise differences in means, concluding a significant difference when the 5% quantile of the bootstrap distribution of difference of means is positive.

Data availability

ERP data supporting the findings of this study are available from the corresponding author upon reasonable request.

Code availability

Code implementing the Bayesian RPAGP model described in the "Methods" section of this paper can be downloaded from GitHub at https://github.com/Beniamino92/BayesRPAGP/.

Received: 10 October 2023; Accepted: 12 April 2024
Published online: 17 April 2024

References

1. Gasser, T. & Molinari, L. The analysis of the EEG. *Stat. Methods Med. Res.* **5**(1), 67–99 (1996).
2. Luck SJ. 27. In: Event-related Potentials, vol. 1 American Psychological Association, 523-546 (2012).
3. Gibney, K. D. *et al.* Estimating statistical power for event-related potential studies using the late positive potential. *Psychophysiology* **57**(2), e13482 (2020).
4. Tibshirani, R. J., & Efron, B. An introduction to the bootstrap. *Monogr. Stat. Appl. Probab.* **57**(1) (1993).
5. Schupp, H. T. & Kirmse, U. M. Case-by-case: Emotional stimulus significance and the modulation of the EPN and LPP. *Psychophysiology* **58**(4), e13766 (2021).
6. Aunon, J. I., McGillem, C. D. & Childers, D. G. Signal processing in evoked potential research: Averaging and modeling. *Crit. Rev. Bioeng.* **5**(4), 323–367 (1981).
7. Kalcher, J. & Pfurtscheller, G. Discrimination between phase-locked and non-phase-locked event-related EEG activity. *Electroencephalogr. Clin. Neurophysiol.* **94**(5), 381–384 (1995).
8. Stolarova, M., Keil, A. & Moratti, S. Modulation of the C1 visual event-related component by conditioned stimuli: Evidence for sensory plasticity in early affective perception. *Cereb. Cortex* **16**(6), 876–887 (2006).
9. Toscano, J. C., McMurray, B., Dennhardt, J. & Luck, S. J. Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychol. Sci.* **21**(10), 1532–1540 (2010).
10. Jeste, S. S. *et al.* Electrophysiological evidence of heterogeneity in visual statistical learning in young children with ASD. *Dev. Sci.* **18**(1), 90–105 (2015).
11. Yu, C. H., Li, M., Noe, C., Fischer-Baum, S. & Vannucci, M. Bayesian inference for stationary points in Gaussian process regression models for event-related potentials analysis. *Biometrics* **79**(2), 629–641 (2023).
12. Woody, C. D. Characterization of an adaptive filter for the analysis of variable latency neuroelectric signals. *Med. Biol. Eng.* **5**(6), 539–554 (1967).
13. Tuan, P. D., Möcks, J., Köhler, W. & Gasser, T. Variable latencies of noisy signals: Estimation and testing in brain potential data. *Biometrika* **74**(3), 525–533 (1987).
14. Jaskowski, P. & Verleger, R. Amplitudes and latencies of single-trial ERP's estimated by a maximum-likelihood method. *IEEE Trans. Biomed. Eng.* **46**(8), 987–993 (1999).
15. Truccolo, W. *et al.* Estimation of single-trial multicomponent ERPs: Differentially variable component analysis (dVCA). *Biol. Cybern.* **89**(6), 426–438 (2003).
16. Xu, L. *et al.* ASEO: A method for the simultaneous estimation of single-trial event-related potentials and ongoing brain activities. *IEEE Trans. Biomed. Eng.* **56**(1), 111–121 (2008).
17. Liu, Y., Keil, A. & Ding, M. Effects of emotional conditioning on early visual processing: Temporal dynamics revealed by ERP single-trial analysis. *Hum. Brain Mapp.* **33**(4), 909–919 (2012).
18. Liu, Y., Huang, H., McGinnis-Deweese, M., Keil, A. & Ding, M. Neural substrate of the late positive potential in emotional processing. *J. Neurosci.* **32**(42), 14563–14572 (2012).
19. Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M. & Muller, K. R. Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Process. Mag.* **25**(1), 41–56 (2007).
20. Hu, L. *et al.* Taking into account latency, amplitude, and morphology: Improved estimation of single-trial ERPs by wavelet filtering and multiple linear regression. *J. Neurophysiol.* **106**(6), 3216–3229 (2011).
21. Gramfort, A., Keriven, R. & Clerc, M. Graph-based variability estimation in single-trial event-related neural responses. *IEEE Trans. Biomed. Eng.* **57**(5), 1051–1061 (2010).
22. Dimitriadis, S. I., Brindley, L., Evans, L. H., Linden, D. E. & Singh, K. D. A novel, fast, reliable, and data-driven method for simultaneous single-trial mining and amplitude-latency estimation based on proximity graphs and network analysis. *Front. Neuroinform.* **12**, 59 (2018).
23. Hu, L., Zhang, Z., Mouraux, A. & Iannetti, G. D. Multiple linear regression to estimate time-frequency electrophysiological responses in single trials. *Neuroimage* **111**, 442–453 (2015).
24. Volpert-Esmond, H. I., Merkle, E. C., Levsen, M. P., Ito, T. A. & Bartholow, B. D. Using trial-level data and multilevel modeling to investigate within-task change in event-related potentials. *Psychophysiology* **55**(5), e13044 (2018).
25. Noh, E., Herzmann, G., Curran, T. & de Sa, V. R. Using single-trial EEG to predict and analyze subsequent memory. *Neuroimage* **84**, 712–723 (2014).
26. Zheng, Q., Zhu, F. & Heng, P. A. Robust support matrix machine for single trial EEG classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* **26**(3), 551–562 (2018).
27. Si, Y. *et al.* Predicting individual decision-making responses based on single-trial EEG. *Neuroimage* **206**, 116333 (2020).
28. Wang, Y., Ombao, H. & Chung, M. K. Topological data analysis of single-trial electroencephalographic signals. *Ann. Appl. Stat.* **12**(3), 1506 (2018).
29. Sturm, I., Lapuschkin, S., Samek, W. & Müller, K. R. Interpretable deep neural networks for single-trial EEG classification. *J. Neurosci. Methods* **274**, 141–145 (2016).

30. Tang, Z., Li, C. & Sun, S. Single-trial EEG classification of motor imagery using deep convolutional neural networks. *Optik* **130**, 11–18 (2017).
31. Bagchi, S. & Bathula, D. R. EEG-ConvTransformer for single-trial EEG-based visual stimulus classification. *Pattern Recogn.* **129**, 108757 (2022).
32. Zang, S., Ding, X., Wu, M., & Zhou, C. An EEG Classification-Based Method for Single-Trial N170 Latency Detection and Estimation. *Comput. Math. Methods Med.* (2022).
33. Razzak, I., Blumenstein, M. & Xu, G. Multiclass support matrix machines by maximizing the inter-class margin for single trial EEG classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* **27**(6), 1117–1127 (2019).
34. Raza, H., Chowdhury, A., Bhattacharyya, S., & Samothrakis, S. Single-trial EEG classification with EEGNet and neural structured learning for improving BCI performance. In *2020 International Joint Conference on Neural Networks (IJCNN) IEEE*. 1–8 (2020).
35. Leoni, J., Strada, S. C., Tanelli, M., Brusa, A. & Proverbio, A. M. Single-trial stimuli classification from detected P300 for augmented Brain-Computer Interface: A deep learning approach. *Mach. Learn. Appl.* **9**, 100393 (2022).
36. Versace, F., Kypriotakis, G., Basen-Engquist, K. & Schembre, S. M. Heterogeneity in brain reactivity to pleasant and food cues: Evidence of sign-tracking in humans. *Soc. Cogn. Affect. Neurosci.* **11**(4), 604–611 (2016).
37. Bradley, M. M. Natural selective attention: Orienting and emotion. *Psychophysiology* **46**(1), 1–11 (2009).
38. Hajcak, G., & Foti, D. Significance?... Significance! Empirical, methodological, and theoretical connections between the late positive potential and P300 as neural responses to stimulus significance: An integrative review. *Psychophysiology* **57**(7), e13570 (2020).
39. Lang, P. J. & Bradley, M. M. Emotion and the motivational brain. *Biol. Psychol.* **84**(3), 437–450 (2010).
40. Schupp, H. T. *et al.* Affective picture processing: the late positive potential is modulated by motivational relevance. *Psychophysiology* **37**(2), 257–261 (2000).
41. Weinberg, A. & Hajcak, G. Beyond good and evil: The time-course of neural activity elicited by specific picture content. *Emotion* **10**(6), 767 (2010).
42. Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* (2011).
43. Lang, P. J. *et al. International affective picture system (IAPS): Affective ratings of pictures and instruction manual* (Center for the Study of Emotion & Attention Gainesville, FL, NIMH, 2005).
44. Miccoli, L. *et al.* Affective pictures and the Open Library of Affective Foods (OLAF): tools to investigate emotions toward food in adults. *PLoS ONE* **11**(8), e0158991 (2016).
45. Versace, F., Sambuco, N., Deweese, M. M. & Cinciripini, P. M. Electrophysiological normative responses to emotional, neutral, and cigarette-related images. *Psychophysiology* **1**, 1 (2023).
46. Luck, S. J., Stewart, A. X., Simmons, A. M. & Rhemtulla, M. Standardized measurement error: A universal metric of data quality for averaged event-related potentials. *Psychophysiology* **58**(6), e13793 (2021).
47. Telesca, D. & Inoue, L. Y. T. Bayesian hierarchical curve registration. *J. Am. Stat. Assoc.* **103**(481), 328–339 (2008).
48. Lu, Y., Herbei, R. & Kurtek, S. Bayesian registration of functions with a Gaussian process prior. *J. Comput. Graph. Stat.* **26**(4), 894–904 (2017).
49. Fu, E. & Heckman, N. Model-based curve registration via stochastic approximation EM algorithm. *Comput. Stat. Data Anal.* **131**, 159–175 (2019).
50. Roy, A., Borg, J. S. & Dunson, D. B. Bayesian time-aligned factor analysis of paired multivariate time series. *J. Mach. Learn. Res.* **22**, 250–1 (2021).
51. Rønn, B. B. Nonparametric maximum likelihood estimation for shifted curves. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **63**(2), 243–259 (2001).
52. Brumback, L. C. & Lindstrom, M. J. Self modeling with flexible, random time transformations. *Biometrics* **60**(2), 461–470 (2004).
53. Gervini, D. & Gasser, T. Self-modelling warping functions. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **66**(4), 959–971 (2004).
54. Tucker, J. D., Wu, W. & Srivastava, A. Generative models for functional data using phase and amplitude separation. *Comput. Stat. Data Anal.* **61**, 50–66 (2013).
55. Marron, J. S., Ramsay, J. O., Sangalli, L. M. & Srivastava, A. Functional data analysis of amplitude and phase variation. *Stat. Sci.* **30**(4), 468–484 (2015).
56. Ahn, K., Tucker, J. D., Wu, W. & Srivastava, A. Regression models using shapes of functions as predictors. *Comput. Stat. Data Anal.* **151**, 107017 (2020).
57. Brovelli, A. *et al.* Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by Granger causality. *Proc. Natl. Acad. Sci.* **101**(26), 9849–9854 (2004).
58. Williams, C. K. & Rasmussen, C. E. *Gaussian processes for machine learning* (MIT press Cambridge, MA, 2006).
59. Rousseau, J. On the frequentist properties of Bayesian nonparametric methods. *Annu. Rev. Stat. Appl.* **3**, 211–231 (2016).
60. Gelman, A. *et al. Bayesian Data Analysis* (Chapman & Hall/CRC, Boca Raton, FL, 2014).
61. Roy, V. Convergence diagnostics for markov chain monte carlo. *Annu. Rev. Stat. Appl.* **7**, 387–412 (2020).
62. Im CH. Computational EEG Analysis: Methods and Applications (Springer, 2018).

Acknowledgements

We sincerely thank the Editor and the Referees for their valuable and insightful comments, which led to a significantly improved manuscript.

Author contributions

D.P., M.L. and M.V. designed the research and conceived the analysis plan; D.P. performed the initial model implementation; D.P. and Y.Z. performed the data analyses for the original submission; BHA developed an enhanced version of the code and produced all the results included in the revised version of the paper; F.V. provided general feedback and counsel on data analysis and interpretation; M.V. provided overall supervision and direction of the work. D.P., F.V. and M.V. wrote the original paper.

Funding

D.P. was supported by grants R01-MH124115/R01-MH122948, M.V. and BHA were supported by grant R01-MH124115, National Institute of Health, M.L. by grant DMS-2015569, National Science Foundation, and FV by the National Institute on Drug Abuse award number 2R01DA032581 and the NIH/NCI under award number P30CA016672. The content is solely the responsibility of the authors and does not necessarily represent official views of the NIH.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-59579-2.

Correspondence and requests for materials should be addressed to M.V.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.