Bayesian models of mentalizing

GRAVE DE PERALTA MENENDEZ, Rolando, *et al*.

Abstract

Surprisingly effortless is the human capacity known as "mentalizing", i.e., the ability to explain and predict the behavior of others by attributing to them independent mental states, such as beliefs, desires, emotions or intentions. This capacity is, among other factors, dependent on the correct anticipation of the dynamics of facially expressed emotions based on our beliefs and experience. Important information about the neural processes involved in mentalizing can be derived from dynamic recordings of neural activity such as the EEG. We here exemplify how the so-called Bayesian probabilistic models can help us to model the neural dynamic involved in the perception of clips that evolve from neutral to emotionally laden faces. Contrasting with conventional models, in Bayesian models, probabilities can be used to dynamically update beliefs based on new incoming information. Our results show that a reproducible model of the neural dynamic involved in the appraisal of facial expression can be derived from the grand mean ERP over five subjects. One of the two models used to predict the individual subject dynamic yield [...]

Reference

GRAVE DE PERALTA MENENDEZ, Rolando, *et al*. Bayesian models of mentalizing. *Brain topography*, 2008, vol. 20, no. 4, p. 278-83

DOI : 10.1007/s10548-008-0047-4
PMID : 18351451

Brain Topogr (2008) 20:278–283
DOI 10.1007/s10548-008-0047-4

ORIGINAL PAPER

Bayesian Models of Mentalizing

Rolando Grave de Peralta Menendez · Amal Achaïbou · Pierre Bessière · Patrik Vuilleumier · Sara Gonzalez Andino

Accepted: 11 February 2008 / Published online: 20 March 2008

Abstract Surprisingly effortless is the human capacity known as "mentalizing", i.e., the ability to explain and predict the behavior of others by attributing to them independent mental states, such as beliefs, desires, emotions or intentions. This capacity is, *among other factors*, dependent on the correct anticipation of the dynamics of facially expressed emotions based on our beliefs and experience. Important information about the neural processes involved in mentalizing can be derived from dynamic recordings of neural activity such as the EEG. We here exemplify how the so-called Bayesian probabilistic models can help us to model the neural dynamic involved in the perception of clips that evolve from neutral to emotionally laden faces. Contrasting with conventional models, in Bayesian models, probabilities can be used to dynamically update beliefs based on new incoming information. Our results show that a reproducible model of the neural dynamic involved in the appraisal of facial expression can be derived from the grand mean ERP over five subjects. One of the two models used to predict the individual subject dynamic yield correct estimates for four of the five subjects analyzed. These results encourage the future use of Bayesian formalism to build more detailed models able to describe the single trial dynamic.

Keywords EEG · Bayesian · Mentalizing · Emotional faces · Modeling

Introduction

The ability to infer the mental states (beliefs, thoughts, and intentions) of others in order to predict and explain their behavior depends on cues extracted from many modalities. The tone of the voice, the gestures or the direction of others gaze are examples of the cues that needs to be combined to correctly infer the feelings of the person that we are interacting with. The accurate detection of others feelings is therefore a first step to anticipate others reactions. From a Bayesian perspective, these two problems, i.e., detection and prediction are not independent. Errors made in the prediction step can be used to evaluate the accuracy of the detection and refine it in a dynamical process. Also, depending on our present inferences (part of our internal beliefs) we might decide to differently weight the cues to better update our estimate of the others mental states. Consequently, the idea that the Bayesian formalism can be used to appropriately model the process of "mentalizing" or the way in which the brain combines multisensory cues for mental state detection is starting to emerge [7, 9].

R. Grave de Peralta Menendez (✉) · S. Gonzalez Andino
Electrical Neuroimaging Group, Department of Clinical Neuroscience, Geneva University Hospital, Geneva, Switzerland
e-mail: Rolando.Grave@hcuge.ch

R. Grave de Peralta Menendez · A. Achaïbou · P. Vuilleumier · S. Gonzalez Andino
Department of Neuroscience and Clinic of Neurology, University Medical Centre, Geneva, Switzerland

R. Grave de Peralta Menendez
Neurodynamics Laboratory, Department of Psychiatry and Clinical Psychobiology, University of Barcelona, 08035 Barcelona, Catalonia, Spain

A. Achaïbou · P. Vuilleumier
Laboratory for Neurology and Imaging of Cognition, Geneva, Switzerland

P. Bessière
CNRS - IMAG/GRAVIR, Grenoble, France

Some authors have proposed that the brain's mirror-neuron system [12] is at the core of the mentalizing process. The idea that there is a mirror system in the brain arises from the observation that the same brain areas are activated when we observe another person experiencing an emotion as when we experience the same emotion ourselves (see e.g. [19]). The brain's mirror system might help to explain how the first inference about others states is created. Emotions are contagious, i.e., the observers tend to imitate and feel the emotions of the persons they are interacting with. For example, through the observation of others facial expressions we can experience the emotional states of another person and help to shape and update our beliefs about his/her feelings [15].

How can then Bayesian decision theory frame the mentalizing process? A hidden state which is not directly observable, the agent's feelings, has to be inferred by the observer on the basis of new upcoming evidence extracted from multisensory cues. The new cues are combined with the present estimate of the state (that depends upon a combination of our experience and the incoming information) to update the beliefs in a sort of iterative procedure. As such Bayesian formalism provides an elegant solution to the problem of mentalizing by predicting that we are constantly assigning probabilities to our internally stored belief about others mental states and updating such beliefs in the measure that new information is obtained. This will give to the neural system the capacity to assign a given probability to transitions in others emotional states according to a combination of experience in social communication and incoming information.

While the model is sound, the challenge is to demonstrate that the brain actually relies on a Bayesian framework to perform the process of mentalizing. Such evidence can only arise from trying to model neural data using this framework and evaluating each of the many multiple models of the same process that can be accommodated within the broad Bayesian formalism. For instances, the amount of multisensory cues that can be added as variables to model the mentalizing problem is so large that might quickly lead to mathematically untreatable problems. All these factors explain why the Bayesian formalism remains as an appealing theory to model mentalizing without formal experimental support.

The widely distributed character of the mirror system, not tied to any particular brain region, suggest that evidences in favor of the Bayesian framework should rely on global rather than local measures of neural activity. One global measure of neural activity able to capture the full dynamic of neural processes at millisecond resolution is the scalp EEG. Therefore, to evaluate the adequacy of Bayesian formalism we carried out the analysis of EEG data recorded from five healthy volunteers during passive viewing of clips showing dynamic expression sequences. In the clips an initially neutral face was gradually transformed into an emotional face either portraying a happy (towards to happy) or an angry (towards to angry) expression. This is a particularly interesting experiment since contagion is considered a first step in mentalizing [7]. EMG activity recorded on the dataset analyzed here demonstrated that subjects covertly imitated the facial expressions they were observing [2].

Here, two concrete Bayesian models are proposed to model one specific component of the mentalizing process, i.e., inferring other feelings from facial expressions. It is assumed that the EEG dynamics contains information about how the observer updates his/her beliefs about the agent's feelings. The models are built from the grand average data and its capability to reproduce the dynamic is evaluated by their possibilities to predict/identify the class of facial expressions observed by the individual subjects. Note that this goal is totally different from most neuroimaging studies that have addressed the neural substrates of mentalizing (see e.g., [1, 14] for reviews).

Material and Methods

Subjects and Recordings

Five healthy volunteers (2 males, mean age = 26.1 years, age range 22–35 years) were selected for the analysis presented here. For a more detailed description of the experiment, the grand-mean ERPs and the EMG see [2].

Movie clips of dynamic facial expressions (anger or happy) were generated in E-Prime using morphed pictures [4]. The selected faces correspond to 10 different identities. Subjects passively viewed 5 blocks of 50 movies each. Each block contained 25 clips that evolved from neutral to happy faces and 25 clips that evolved from neutral to angry faces. Each clip lasted 1,460 ms. Clips were separated by an average 4 s ITI.

EEG and EMG data were simultaneously recorded during the task on a darkened room. EEG data were sampled at 2,048 Hz using a 64-channel Biosemi ActiveTwo system with sintered Ag–AgCl active electrodes. The analysis of the data was carried out using custom built in software programmed in Matlab. The data was downsampled to 512 Hz before posterior analysis.

Bayesian Models

In this section we formulate two Bayesian models for the EEG sequence maps recorded during the observation of the movie-clips displaying either happy or angry facial expressions. The models consider the scalp maps topography as a whole rather than the independent dynamic of each electrode. The idea is to build the models based on the grand mean over subjects and to evaluate the goodness of the models by testing how well they describe the dynamics on the mean ERP of the single subjects. Importantly, the primary goal of these models is the description of the dynamic of the process allowing, as a secondary goal, the identification of the class (angry or happy face) generating the data. This is to be contrasted with pattern recognition methods that focus mainly in the classification of unseen data.

In the following we will denote the potential map by V and the state variable by X. Subscripts will be used to denote the value at a single time point (V_t) or set of time points ($V_{0:t}$).

The first model (B1) aims at describing the temporal dynamics of state variable and can be applied to the EEG maps as follows. Given the EEG maps sequence V_t, t = 0,1,2,...T, the labeling of its corresponding facial expression class can be represented as a temporally accumulated posterior probability at time t, $p(X_t|V_{0:t})$, where the state variable X_t represents the class label of a map (towards happy or towards angry). Assuming that the measurement V_t is completely determined by current state X_t, the estimation of the class $X_t \subset \{2\}$ corresponding to the two classes of movie clips (i.e. happy and angry) can be obtained from a Bayesian perspective in the following way:

$$p(X_t/V_{0:t}) = \frac{p(V_t/X_t)p(X_t/V_{0:t-1})}{p(V_t/V_{0:t-1})} \quad (1)$$

Assuming conditional independence of the state variable X_t with respect to past measurements V_t, t = 0,1,2,...T−1 given X_{t-1} is equivalent to say that the state X_t is complete [16]. In other words, completeness entails that knowledge of past states or measurements carry no additional information that would help us to predict the future more accurately. We would note that this definition of completeness does not require the future to be a deterministic function of the state but just that no variables prior to X_t may influence the stochastic evolution of future states, unless this dependence is mediated through the state X_t. In our particular case, this means, the perception of the class (clip) at time *t* depends on the class at *t−1* but not on the EEG maps previous to V_t. In mathematical parlance it is expressed by the following equation:

$$p(X_t/X_{t-1}, V_{0:t-1}) = p(X_t/X_{t-1}) \quad (2)$$

Then the accumulated prior probability of X_t given the past measurements can be computed from:

$$p(X_t/V_{0:t-1}) = \int p(X_t/X_{t-1})p(X_{t-1}/V_{0:t-1})dX_{t-1} \quad (3)$$

That allows rewriting Eq. 1 as a Bayes filter algorithm without control data:

$$p(X_t/V_{0:t}) = \int p(X_{t-1}/V_{0:t-1})\frac{p(V_t/X_t)p(X_t/X_{t-1})}{p(V_t/V_{0:t-1})}dX_{t-1} \quad (4)$$

We can define the initial probability $p(X_0|V_0) = p(X_0) = 1/N$, where N = 2 is the number of classes and estimate the $p(V_t|X_t)$ and the $p(X_t|X_{t-1})$ from a given data set (learning set). Then the model given by Eq. 4 can be applied recursively to compute the state variable at each time point.

We would note that in this model, the only (and weak) connection between successive scalp maps and/or states is the one induced by Eq. 2 and thus this model ignores potential temporal dependencies between scalp maps.

The second Bayesian model proposed here (B2), is a combination of a Bayesian classifier with a law describing temporal dependencies within the class. That is, model B2 uses the output of previous time point to estimate the state at current time point. As for previous model we start from Bayes equation relating the state X_t and the observation V_t at current time point:

$$p(X_t/V_t) = \frac{p(V_t/X_t)p(X_t)}{p(V_t)} \quad (5)$$

Now defining $p(X_t) = p(X_{t-1}|V_{t-1})$ we can rewrite (5) as

$$p(X_t/V_t) = \frac{p(V_t/X_t)p(X_{t-1}/V_{t-1})}{p(V_t)} \quad (6)$$

where $p(X_0|V_0) = p(X_0) = 1/N$, with N = 2 is the number of classes and $p(V_t|X_t)$ can be determined from a given data set (learning set).

Equation 6 defines a recursive process to compute the posterior probability of the class label at each time point using as prior probability the output of previous time point. This could be interpreted as accumulation of evidences where the new decisions in favor or against one of the classes are proportional to the past experiences.

Implementation

In this paper, the learning set is defined as a temporal window selected from the two grand means over subjects for the happy and angry classes. The temporal window used for the learning set was the interval between 500 and 1,500 ms. From 500 ms onwards the EMG responses to the analyzed facial expressions was significantly different for both muscles: the zygomaticus major (ZM) that elevates the lips during a smile, and the corrugator supercillii (CS) that knits the eyebrows during a frown [2]. Consequently, at period selected for the learning set the subjects already identified and mimicked the observed facial expression.

The probability $p(V_t|X_t)$ used in both models was defined as a Gaussian distribution with means and covariance matrix computed from the learning set , that is,

$$p(V_t/X_t = k) = \eta * \exp(-(V_t - \mu_k)'\Sigma_k(V_t - \mu_k) \quad (7)$$

where k = 1,2 denotes the class and η is a normalization constant. Note also that the terms in the denominator of Eqs. 4 and 6 do not need to be computed explicitly because they do not involve the state variable. Mean and covariance matrices for each class were computed as the maximum likelihood estimators. Transition probability in Eqs. 3 and 4 was computed from the output of Eq. 5 assuming $p(X_t) = 1/N$ and the model (7) and using the grand means.

Grand mean for each movie clip class was computed as the mean of the average evoked potential from the 5 subjects. This EEG data was then used as training set to compute the parameters of models B1 and B2.

The output of the method was defined as the class to which the procedure converged after the whole sequence, i.e., at 1,500 ms. Thus, we consider a correct identification if the class identified at the end of the sequence is the correct class, i.e., the probability observed at 1,500 ms is larger than 0.5.

Results

Figure 1 depicts for each subject (row) the probability of each class as a function of time using the Bayesian filter algorithm of model B1. As can be seen the model fitting of B1 to the individual subjects suggest a dynamic behavior. The probabilities of correctly identifying the evolving to happy clip when this was the actually shown clip substantially vary over time. For subject 1, the probabilities of identifying the current class are close to one over the whole sequence and for both classes. The situation is different for subject 3 (third row). For the case of evolving to happy clips (left column) the probabilities remain close to 0.5 for nearly the whole period and are close to zero at the end of the sequence. This indicates that the model failed to identify the correct class in this case. The model also fails for subjects two and five during visualization of the evolving to angry clips. According to our definition of correct identification that considers the convergence at the end of the sequence, this model B1 fails in capturing the correct perception of the subjects in *three* out of ten situations.

Figure 2 depicts for each subject (row) the probability of each class as a function of time using the recursive Bayesian classifier of model B2. In this case, the model assumes that the probability of observing a given class at time *t* depends on the class observed at time *t−1*. This assumption can be considered as more realistic given the known similarities of maps on grand mean data.

The dynamics of this model on the single subject basis is more rigid, i.e., exhibit less temporal variability. The model is however able to identify the correct class of clips in 9 out of 10 situations studied. The model only fails to converge to the correct class for the evolving to happy clip of subject number three.

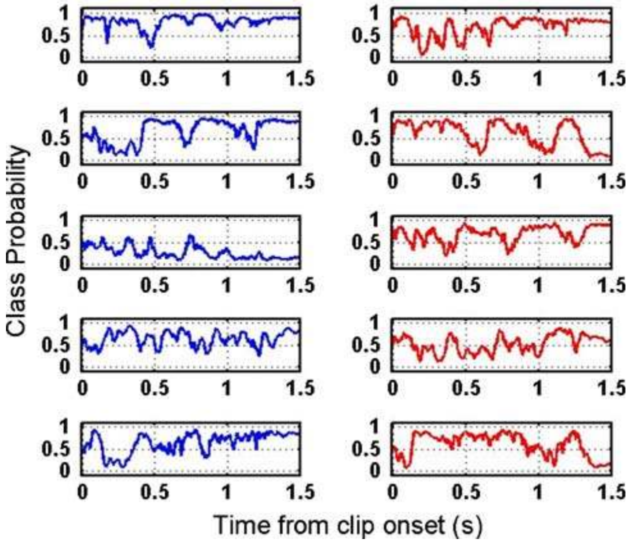

Fig. 1 Probability of the model B1 of predicting that class was happy when the actual clip evolves to happy (left plot) or to predict that clip was angry when the actual clip evolves towards angry (right). Probabilities are displayed as a function of time (seconds) from clip onset. One subject is depicted on each row. Probability values near to one indicate that the correct class was identified for the subject

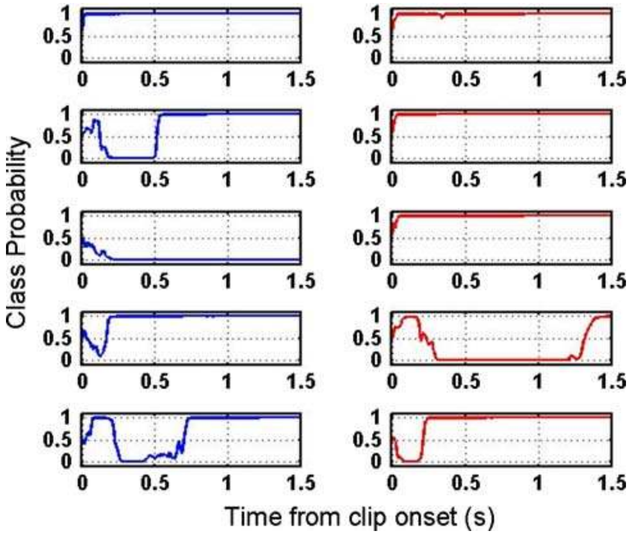

Fig. 2 Probability of the model B2 of predicting that class was happy when the actual clip evolves to happy (left plot) or to predict that clip was angry when the actual clip evolves towards angry (right). Probability values near to one indicate that the correct class was identified for the subject

Discussion

We here present two different Bayesian temporal models that aim to describe from EEG data how subjects dynamically perceive or make inferences about others mental states from their facial expressions. The second model B2 yielded a high recognition rate (9/10) of the facial expression towards which the clips evolved. This suggests that the mean ERP used to train the model contains information which is consistent at a single subject level. While model B2 assumes that the present state of the class (i.e. the type of facial expression being observed) depends explicitly only on previous state, this induces a strong temporal dependency between consecutive scalp maps. The model also show a rigid temporal dynamic that contradicts our expectancies about the existence of a sequence of perception states which transits from uncertainty at the beginning of the sequence towards a clearly defined percept at intermediate and final stages. On the contrary, model B1 shows a more flexible dynamic but fails to show good generalization properties since the recognition rates at the single subject level are lower (7/10). None of the models is however able to provide a fully satisfactory description of the dynamics of subjects perception. The temporal evolution fails to show the transition from uncertainty (probabilities near 0.5) about the facial expression towards which the clip will evolve to complete certainty information around 500 ms (where the EMG data revealed significant differences).

The differences between both models to describe the individual temporal dynamic are due to their different underlying assumptions. Model B2 assumes an explicit dependency between the classes observed at consecutive time points. This means, the probability that subjects perception is a "towards to happy clip" depends not only on current map but also on the class observed at previous instant. On the other hand, model B1 includes temporal dependencies during the computation of the transition matrix. Consequently, model B1 assumes that state transitions detected from the grand mean should arise also at the single subject level. The explicit incorporation of a dynamic law inherited from the grand mean might explain its variability and the slower temporal convergence towards the correct class. In contrast model B2 yields faster but rigid temporal behavior. Each model integrates the temporal information differently. While B1 is mainly dominated by previous time point, model B2 takes into account all precedent results into the a priori probability, then apparently little increases in the a priori probability might yield strong posterior probabilities and thus faster convergence.

We should mention, however, that the high recognition rates achieved are encouraging to continue with the refinement of Bayesian models in the study of neural activity. The idea behind Bayesian probabilistic models that accept uncertainty as a natural component of neural processes is appealing. It might for example be used to incorporate the influence of non-measurable internal states, in perception which should ultimately lead to descriptions at the single trial level.

The use of Bayesian/probabilistic models for EEG analysis and synthesis is at its infancy. Hitherto, most applications of the Bayesian framework are oriented to the solution of the electromagnetic inverse problem [3, 5, 8, 17] and the introduction of temporal constraints to model sources dynamic [6]. Bayesian models have been also applied to the problem of classifying single trials within one or more classes within the framework of Brain Computer Interfaces [10, 13]. The goal of the models described here is, in a given sense, more ambitious than the simple identification of the correct class as in Brain Computer Interface or the segmentation of neural data into stable mental states [11, 18]. The ultimate goal is to identify and characterize the dynamic of a non-observable variable (subject's internal perception and prediction) from the measured EEG. To achieve such ambitious goals, the models could be refined to explicitly incorporate the temporal dependencies between scalp maps and their transitions using the Bayesian formalism.

One interesting question to be explored in the future is if the incorporation of multimodal cues could lead to better but still mathematically treatable models of the mentalizing process. The models developed here are restricted to the problem of making inferences about others mental states on the sole basis of facial expression. Therefore, future work will focus on developing more sophisticated models that consider a more complex dynamic and probably employ auxiliary channels, e.g., EMG as control variables.

Acknowledgments This work has been supported by the European Program FP6-IST-027140 (BACS). This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

References

1. Abu-Akel A. A neurobiological mapping of theory of mind. Brain Res Rev 2003;43:29–40.
2. Achaibou A, Pourtois G, Schwartz S, Vuilleumier P. Simultaneous recording of EEG and facial muscle reactions during spontaneous emotional mimicry. Neuropsychologia 2007 doi: 10.1016/j.neuropsychologia.2007.10.019.
3. Baillet S, Garnero L. A Bayesian approach to introducing anatomo-functional priors in the EEG/MEG inverse problem. IEEE Trans Biomed Eng 1997;44:374–85.
4. Ekman P, Friesen WV. Measuring facial movement. Env Psychol Nonverbal Behav 1976;1:56–75.
5. Friston K, Harrison L, Daunizeau J, Kiebel S, Phillips C, Trujillo-Barreto N, Henson R, Flandin G, Mattout J. Multiple sparse priors for the M/EEG inverse problem. Neuroimage 2007;39:1104–20.
6. Friston K, Henson R, Phillips C, Mattout J. Bayesian estimation of evoked and induced responses. Hum Brain Mapp 2006;27:722–35.
7. Frith CD. The social brain? Philos Trans R Soc Lond B Biol Sci 2007;362:671–8.

8. Galka A, Yamashita O, Ozaki T, Biscay R, Valdes-Sosa P. A solution to the dynamical inverse problem of EEG generation using spatiotemporal Kalman filtering. Neuroimage 2004;23:435–53.
9. Kilner JM, Friston KJ, Frith CD. The mirror-neuron system: a Bayesian perspective. Neuroreport 2007;18:619–23.
10. Kohlmorgen J, Blankertz B. Bayesian classification of single-trial event-related potentials in EEG. Int J Bif Chaos 2004;14: 719–26.
11. Michel CM, Henggeler B, Lehmann D. 42-Channel potential map series to visual contrast and stereo stimuli: perceptual and cognitive event-related segments. Int J Psychophysiol 1992;12:133–45.
12. Rizzolatti G, Craighero L. The mirror-neuron system. Annu Rev Neurosci 2004;27:169–92.
13. Roberts SJ, Penny WD. Real-time brain-computer interfacing: a preliminary study using Bayesian learning. Med Biol Eng Comput 2000;38:56–61.
14. Siegal M, Varley R. Neural systems involved in "theory of mind". Nat Rev Neurosci 2002;3:463–71.
15. Singer T. The neuronal basis and ontogeny of empathy and mind reading: review of literature and implications for future research. Neurosci Biobehav Rev 2006;30:855–63.
16. Thrun S, Burgard W, Fox D. Probabilistic robotics. MIT Press; 2005.
17. Trujillo-Barreto NJ, Aubert-Vazquez E, Penny WD. Bayesian M/EEG source reconstruction with spatio-temporal priors. Neuroimage 2008;39:318–35.
18. Wackermann J, Lehmann D, Michel CM, Strik WK. Adaptive segmentation of spontaneous EEG map series into spatially defined microstates. Int J Psychophysiol 1993;14:269–83.
19. Wicker B, Keysers C, Plailly J, Royet JP, Gallese V, Rizzolatti G. Both of us disgusted in My insula: the common neural basis of seeing and feeling disgust. Neuron 2003;40:655–64.